Methodological advances in the Network Scale-up Method

by

Sergio Díaz Aranda

A dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in

MATHEMATICAL ENGINEERING

Universidad Carlos III de Madrid

Advisors:

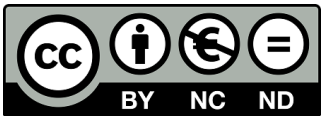
Antonio Fernández Anta
Rosa E. Lillo

Tutor:

Rosa E. Lillo

October 2026

This thesis is distributed under license “Creative Commons **Atributtion - Non Commercial - Non Derivatives**”.



Acknowledgements

First, I would like to express my deepest gratitude to my supervisors, Antonio and Rosa, for believing in me and giving me the opportunity to be part of this incredible project. I will always be grateful for your guidance, encouragement, and the genuine care you have shown for me throughout these years. I would also like to thank Jose and Juan Marcos for their continuous support. This work would not have been the same without your generosity, dedication, and willingness to help whenever it was needed.

I would like to thank IMDEA Networks Institute for funding my research and creating an excellent environment. I am also grateful to the administrative staff at IMDEA Networks for their assistance with administrative matters, financial procedures, travel arrangements, and grant management.

I am especially grateful to Professor Tyler H. McCormick for hosting me in Seattle. It was a wonderful experience, and I will always appreciate your support and hospitality.

Finally, I want to acknowledge the support of the Comunidad de Madrid through the grant PIPF-2022/COM-24467.

This work has been supported by:



Published and submitted content

The materials from the following sources are included in the thesis. Their inclusion is not indicated by typographical means or references since they are fully embedded in each of the chapters indicated below.

1. **Díaz-Aranda, S.**, Aguilar, J., Ramírez, J. M., Rabanedo, D., Fernández Anta, A., and Lillo, R. E. (2025). "Performance Analysis of NSUM Estimators in Social-Network Topologies". The American Statistician, 79(2), 247-264.
   - https://doi.org/10.1080/00031305.2024.2421361
   - Fully included in Chapter 2.

2. **Díaz-Aranda, S.**, Ramirez, J. M., Daga, M., Champati, J. P., Aguilar, J., Lillo, R., and Fernández Anta, A. (2025, August). "Error bounds for the network scale-up method". In Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2 (pp. 498-509).
   - https://doi.org/10.1145/3711896.3736940
   - Fully included in Chapter 3.

3. Srivastava, A., Ramirez, J. M., **Díaz-Aranda, S.**, Aguilar, J., Fernández Anta, A., Ortega, A., and Lillo, R. E. (2024, March). "Nowcasting temporal trends using indirect surveys". In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 38, No. 20, pp. 22359-22367).
   - https://doi.org/10.1609/aaai.v38i20.30242
   - Fully included in Chapter 4.

4. **Díaz-Aranda, S.**, Ramírez, J. M., Aguilar, J., Lillo, R. E., and Fernández Anta, A. (2026). "Robust network scale-up method estimators". Social Networks, 84, 46-61.
   - https://doi.org/10.1016/j.socnet.2025.08.002
   - Fully included in Chapter 5.

5. Arevalillo, J. M., Ramírez, J. M., **Díaz-Aranda, S.**, Aguilar, J., Fernández Anta, A., and Lillo, R. E. (2026) "Network Scale-up Methods on Aggregated Relational Data to Estimate the Outcome of Elections". Social Networks, 86, 312-324.
   - https://doi.org/10.1016/j.socnet.2026.04.006
   - Fully included in Chapter 6.

6. Ramírez, J. M., **Díaz-Aranda, S.**, Balduf, L., and Fernández Anta, A. (2026). "An Alternative to Crawlers: Scalable P2P Network Monitoring Using the Network Scale-Up Method".
   - Submitted for publication
   - Fully included in Chapter 7.

Other research merits

Research stays:

1. April 2025 - June 2025. Department of Statistics, University of Washington. Research tutor: Tyler H. McCormick.

Awards:

1. Honorable Mention for the AI for Social Impact track at AAAI-2024 for the research article "Nowcasting temporal trends using indirect surveys".

Competitive projects:

1. Predoctoral grant PIPF-2022/COM-24467, funded by Comunidad de Madrid.
2. Member of the project *CuidaNSUM*, funded by Instituto de las Mujeres.
3. Member of the project *SocialProbing. Scalable and Cost-Competitive Data Collection and Analysis Techniques for Social Probing*, grant TED2021-131264B-I00, funded by MCIN/AEI/10.13039/501100011033/.
4. Member of the project *DRONAC Distributed Reliable Objects for Networked Applications Coordination*, grant PID2022-140560OB-I00, funded by Ministry of Science and Innovation MCIN/AEI/10.13039/501100011033/.

Invited talks and conference presentations:

1. *Error Bounds for the Network Scale-Up Method.* Contributed conference paper (Oral and Poster) - 2025 ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Toronto, Canada, August 2025.
2. *Error Bounds for the Network Scale-Up Method.* Contributed talk - 2025 IMS International Conference on Statistics and Data Science (ICSDS), Seville, Spain, December 15-18, 2025.
3. *Evaluating the Sensitivity of NSUM Estimators within Social Networks.* Contributed conference poster - 3rd Meeting of the Spanish Society of Complex Systems, Madrid, Spain, February 2025.
4. *Sensitivity Analysis of NSUM Estimators in the context of Social-Networks.*Invited talk - RSME's 7th Congress of Young Researchers, Bilbao, Spain, January 2025.

Delivered Seminars:

1. *Error Bounds for the Network Scale-Up Method* - Invited Research Seminar - Department of Computer Science, University of Victoria (May 2025).

2. *Simulation-based Comparison of Network Scale-Up Methods* - Invited Research Seminar - IMDEA Networks Institute (July 2023).

Master Thesis co-directed as a result of this thesis research:

1. *Community Detection from Aggregated Relational Data via Graph Estimation and Graph Neural Networks*, Ángela Caniego González (2025). Master's Degree in Big Data Analytics - Universidad Carlos III de Madrid. Supervisors: Rosa E. Lillo, Jose Aguilar, **Sergio Díaz-Aranda**.

Bachelor Thesis co-directed as a result of this thesis research:

1. *Graph Neural Networks: Decoding Insights and Real-World Impact*, Helena Martínez Río (2024). Bachelor's Degree in Data Science and Engineering - Universidad Carlos III de Madrid. Supervisors: Rosa E. Lillo, Jose Aguilar, **Sergio Díaz-Aranda**.

Co-supervision of Research Interns:

1. Giuliana Galassi, BSc student in Data Science and Engineering, Universidad Carlos III de Madrid, October 2025 - Present.

2. Lucía Soria González, BSc student in Data and Business Analytics, Universidad Carlos III de Madrid, September 2025 - Present.

3. Milene Catalina Gutiérrez, IMDEA Software Institute, August 2025 - January 2026.

4. Amisha Gupta, BSc student in Computer Science and Data Science, UC Berkeley, June 2025 - July 2025.

5. Victoria Durán Valero, BSc student in Applied Mathematics, Yale University, February 2025 - May 2025.

6. Julio Álvarez Villaescusa, BSc student in Computer Engineering, Universidad de Alcalá, June 2024 - September 2024.

Other research papers not included in this thesis:

1. Fernández-Anta, A., Aguilar, J., Ramírez, J. M., Lillo, R. E., and **Díaz-Aranda, S.** (2024). "Coronasurveys: encuestas indirectas en línea para monitorizar la evolución del COVID-19". Revista Española de Comunicación en Salud, 15-23.

2. **Díaz-Aranda, S.**, Fernández Anta, A., Gutiérrez, M., Jauregui, B., and Ramirez, J. M., "Melting Membership with the Neighbors: Network Scale-Up Methods with Self-Reporting". Submitted for publication.

3. Soria González, L., Gutiérrez, M., **Díaz-Aranda, S.**, Ramírez, J. M., Lillo, R. E., Fernández Anta, A., and Aguilar, J. "Use of LLMs for Generating Synthetic ARDs to Make Estimates in Electoral Processes". Submitted for publication.

Contents

1. INTRODUCTION 1
1.1. Aggregated Relational Data (ARD) 1
1.2. Network Scale-up Methods (NSUM) 2
1.2.1. Biases of the NSUM 3
1.2.2. NSUM estimators. 4
1.3. Network models 7
1.3.1. Social network models 7
1.3.2. Peer-to-peer networks 9
1.4. Main contributions and structure of the thesis 10
2. PERFORMANCE ANALYSIS OF NSUM ESTIMATORS IN SOCIAL-NETWORK TOPOLOGIES. 21
2.1. Introduction 22
2.2. Overview of the Network Scale-Up Methods 23
2.3. NSUM estimators 24
2.3.1. Frequentist estimators 24
2.3.2. Bayesian estimators 25
2.3.3. Generalized Network Scale-Up Method 26
2.4. Simulation Environment and Parameters. 27
2.4.1. Main scenarios 27
2.4.2. Bias modeling 28
2.4.3. Motivation for the other parameters 29
2.5. Simulation Study 30
2.5.1. Simulations results 31
2.5.2. Discussion 40
2.6. Case Study 45
2.6.1. Real network with subpopulations distributed uniformly 45
2.6.2. Real network with SIR for the unknown subpopulation 46
2.6.3. Real network with real subpopulations 47

2.6.4. Discussion . . . 48
2.7. Conclusion . . . 49
Supplementary Material . . . 54
2.S.1. NSUM estimators . . . 54
2.S.1.1. Basic estimators . . . 54
2.S.1.2. Generalized Network Scale-Up Method . . . 56
2.S.1.3. Overdispersed estimator . . . 57
2.S.1.4. Maltiel's Estimators . . . 57
2.S.1.5. TPC Estimator . . . 58
2.S.2. Networks . . . 58
2.S.3. Recall error modeling . . . 59
2.S.4. Results . . . 60
2.S.5. Case study . . . 102
2.S.6. Comparison between estimators based on known subpopulations and based on direct degree estimation . . . 104
3. ERROR BOUNDS FOR THE NETWORK SCALE-UP METHOD . . . 105
3.1. Introduction . . . 106
3.2. Model and Definitions . . . 107
3.3. Upper Bounds on the Error of MoR and RoS in Random Networks . . . 111
3.3.1. Upper Bound on the Error of the MoR Estimator . . . 111
3.3.2. Upper Bound on the Error with the RoS Estimator . . . 112
3.3.3. Sample Size . . . 112
3.4. Upper Bounds for Erdős–Rényi and Scale-Free Networks . . . 113
3.4.1. Erdős–Rényi Network . . . 113
3.4.2. Scale-Free Network . . . 114
3.5. Simulations . . . 115
3.5.1. Erdős–Rényi . . . 115
3.5.2. Scale-Free . . . 117
3.6. Results with Real Networks . . . 119
3.7. Conclusions and Future Work . . . 120
3.8. Acknowledgments . . . 122

Appendix. 126
3.S.A. Concentration Bounds for Negatively Correlated Random Variables . . 126
3.S.B. Proofs of Lemma 3.1 and Theorem 3.2. 128
3.S.C. Proof of Theorem 3.3 131
3.S.D. Deferred Proofs from Section 3.4 133
3.S.D.1. Proof of Theorem 3.6 133
3.S.D.2. Proof of Theorem 3.7 135
4. NOWCASTING TEMPORAL TRENDS USING INDIRECT SURVEYS . . 137
4.1. Introduction 138
4.1.1. Related Work 138
4.1.2. Contributions 139
4.2. Methodology 140
4.2.1. Latent Graph Formulation 140
4.2.2. Leveraging Smoothness 143
4.3. Results and Analysis 146
4.3.1. Synthetic Experiments. 146
4.3.2. Real Dataset. 148
4.4. Discussion 152
4.5. Conclusions 153
Appendix. 157
4.S.A. Proofs 157
4.S.A.1. Extension to Biased Reporting 163
4.S.B. Additional Results 164
4.S.B.1. Results on UMD surveys 165
4.S.C. Additional Methodology Details. 171
4.S.C.1. Code 171
4.S.C.2. Outlier Filtering 171
4.S.C.3. Synthetic Experiments 171
4.S.D. Algorithm in Practice 172
5. ROBUST NETWORK SCALE-UP METHOD ESTIMATORS 174
5.1. Introduction 175

5.2. Robustness and NSUM 176
5.2.1. The NSUM in the Robust Statistics framework 177
5.3. Robust NSUM estimators 178
5.3.1. Location estimators 179
5.3.2. Robustification of MoR and RoS 182
5.4. Simulation Study 183
5.4.1. Simulation environments 184
5.4.2. Simulations Results 188
5.4.3. Discussion 199
5.5. Case Study 200
5.5.1. The United Kingdom COVID-19 dataset 200
5.5.2. Spanish Elections Dataset 202
5.5.3. Discussion 203
5.6. Conclusions 206
Supplementary Material 212
5.S.A. Formal definition of robustness concepts 212
5.S.B. Location measures and trimmed means 213
5.S.C. Erdős–Rényi Network Model 216
5.S.C.1. Without Contamination 216
5.S.C.2. With Contamination 220
5.S.C.3. Transmission Error 224
5.S.C.4. Recall Error 225
5.S.D. Scale-Free Model 226
5.S.D.1. Without Contamination 226
5.S.D.2. With Contamination 228
5.S.D.3. Transmission Error 229
5.S.D.4. Recall Error 230
5.S.E. Stochastic Block Model 231
5.S.E.1. Without Contamination 231
5.S.E.2. Performance versus the Connectivity Factor Without Contamination 232

5.S.E.3. With Contamination 233
6. NETWORK SCALE-UP METHODS ON AGGREGATED RELATIONAL DATA TO ESTIMATE THE OUTCOME OF ELECTIONS 234
6.1. Introduction 235
6.2. Related Work 237
6.3. Methodology 239
6.3.1. Survey Design and Data Collection 240
6.3.2. Data Processing 240
6.3.3. Estimation of Voting Intention 241
6.4. Application of the Methodology 244
6.4.1. Survey Design and Data Collection 244
6.4.2. Data Processing 246
6.4.3. Sampling Stratification for NSUM-based Estimation Procedures . . . 247
6.5. Results 248
6.5.1. Results for the General Elections (Madrid) 248
6.5.2. Results for the General Elections (Andalusia) 251
6.5.3. Results for the General Elections (Valencia) 252
6.6. Discussion on the Use of the Methodology in the Case Study 255
6.6.1. Survey Design and Data Collection 255
6.6.2. Data Processing 256
6.6.3. NSUM Estimation 258
6.7. Conclusions 258
Supplementary Data 266
6.S.A. Supplementary material 266
6.S.A.1. Questionnaire for the General Election in Madrid, Andalusia and Valencia 266
6.S.A.2. Questionnaire for the Regional Election in Madrid 267
6.S.A.3. Results of the Regional Election in Madrid 269
6.S.A.4. Tables with NSUM-based Confidence Intervals 272
6.S.B. D'Hondt Method 280
6.S.C. Sensitivity Analysis under Homophilous Mixing 281

6.S.D. Robustness of Vote Share and Seat Allocation to Trimming and Filtering Choices. 283
7. AN ALTERNATIVE TO CRAWLERS: SCALABLE P2P NETWORK MONITORING USING THE NETWORK SCALE-UP METHOD 287
7.1. Introduction 288
7.2. Background 289
7.2.1. The Network Scale-up Method 289
7.2.2. Peer-to-Peer Networks and IPFS 290
7.3. Methodology and Datasets 291
7.3.1. NSUM IPFS Characterization 291
7.3.2. Datasets and Experiments 292
7.4. Results 294
7.4.1. Estimator Benchmark 294
7.4.2. Implementation Results 297
7.5. Discussion 298
7.6. Related Work 299
7.7. Conclusions and Future Work 300
Appendix. 305
7.S.A. Nodes with and without NAT 305
8. CONCLUSION 308

List of Figures

2.1 Results obtained by the standard methods varying the visibility factor in the W-U scenario. (a)-(e) Boxplots of the unknown subpopulation estimates across methods. (f)-(j) Mean of unknown subpopulation estimates. (k)-(o) MAE of the unknown subpopulation estimates. 32

2.2 Results obtained by the GNSUM and the adjusted methods varying the visibility factor in the W-U scenario. The adjusted methods have been corrected by the inverse of the estimation of the transmission rate using the same sample of the unknown subpopulation as the GNSUM. (a)-(e) Boxplots of the unknown subpopulation estimates across methods. (f)-(j) Mean of unknown subpopulation estimates. (k)-(o) MAE of the unknown subpopulation estimates. 33

2.3 Results obtained varying the memory factor in the W-U scenario. (a)-(e) Boxplots of the unknown subpopulation estimates across methods. (f)-(j) Mean of unknown subpopulation estimates. (k)-(o) MAE of the unknown subpopulation estimates. 34

2.4 Results obtained varying the memory factor in the W-S scenario. (a)-(e) Boxplots of the unknown subpopulation estimates across methods. (f)-(j) Mean of unknown subpopulation estimates. (k)-(o) MAE of the unknown subpopulation estimates. 35

2.5 Results obtained varying the number of known subpopulations in the W-U scenario. Overlapping of subpopulations is included. (a)-(e) Boxplots of the unknown subpopulation estimates across methods. (f)-(j) Mean of unknown subpopulation estimates. (k)-(o) MAE of the unknown subpopulation estimates. 36

2.6 Results obtained for different subpopulations in the W-U scenario. (a)-(e) Boxplots of the unknown subpopulation estimates across methods. (f)-(j) Mean of unknown subpopulation estimates. (k)-(o) MAE of the unknown subpopulation estimates. 37

2.7 Results obtained for different subpopulations in the P-S scenario. (a)-(e) Boxplots of the unknown subpopulation estimates across methods. (f)-(j) Mean of unknown subpopulation estimates. (k)-(o) MAE of the unknown subpopulation estimates. 38

2.8 Results obtained varying the sample size in the W-U scenario. (a)-(e) Boxplots of the unknown subpopulation estimates across methods. (f)-(j) Mean of unknown subpopulation estimates. (k)-(o) MAE of the unknown subpopulation estimates. 39

2.9 Results obtained varying the sample size in the P-U scenario. (a)-(e) Boxplots of the unknown subpopulation estimates across methods. (f)-(j) Mean of unknown subpopulation estimates. (k)-(o) MAE of the unknown subpopulation estimates. 40

2.10 Histogram of the ratio of responses y_{iu}/d_i for the four scenarios. . 42

2.11 Results obtained from a real network with a real unknown subpopulation (Real), a synthetic unknown subpopulation uniformly distributed (Synthetic Uniform), and a synthetic unknown subpopulation using a SIR model (Synthetic SIR). (a)-(c) Boxplots of the estimates. (d)-(f) Mean of estimates. (g)-(i) RMAE estimates. 46

2.12 Results obtained from a real network with real subpopulations. (a)-(c) Boxplots of the unknown subpopulation estimates. (d)-(f) Mean of unknown subpopulation estimates. (g)-(i) RMAE of the unknown subpopulation estimates. 47

2.1 Two networks with similar average degree. Figure (a) has more triadic closures and uniform degree distribution, while Figure (b) has nodes with many links and few triangles. 59

3.1 Example instances I_1 and I_2 with $k = 4$ and $n = 9$. All the edges are bidirectional, and the hidden nodes are in orange. 110

3.2 Erdős–Rényi ($n = 10^6$). (a) Boxplots $\mathcal{E}_{\mathcal{M}}$ versus $|S|$ for MoR and RoS for $\rho = 0.05$ and $p(n-1) = 30$. (b) Average $\mathcal{E}_{\text{MoR}}$ versus $|S|$ for $\rho = 0.05$ and different $p(n-1)$. (c) Average $\mathcal{E}_{\text{RoS}}$ versus $|S|$ for $\rho = 0.05$ and different $p(n-1)$. 115

3.3 Erdős–Rényi ($n = 10^6$). Simulation curves and bounds for (—) $p(n-1) = 15$, (—) $p(n-1) = 30$, and (—) $p(n-1) = 50$. (a) $\Pr[\mathcal{E}_{\text{MoR}} > 1 + \epsilon]$ versus $|S|$ for $\epsilon = 0.05$ and ρ. (b) Sample size versus ϵ for $\rho = 0.05$. (c) Sample size versus ρ for $\epsilon = 0.05$. (d) $\Pr[\mathcal{E}_{\text{RoS}} > 1 + \epsilon]$ versus $|S|$ for $\epsilon = 0.05$ and $\rho = 0.05$. (e) Sample size versus ϵ for $\rho = 0.05$. (f) Sample size versus ρ for $\epsilon = 0.05$. . . 116

3.4 Scale Free ($n = 10^6$). (a) Boxplots of $\mathcal{E}_{\mathcal{M}}$ versus $|S|$ for MoR and RoS. (b) Average $\mathcal{E}_{\text{MoR}}$ versus $|S|$ for different ρ. (c) Average $\mathcal{E}_{\text{RoS}}$ versus $|S|$ for different ρ. 117

3.5 Scale Free ($n = 10^6$). Simulation curves and bounds for (—) $\rho = 0.02$, (—) $\rho = 0.05$, and (—) $\rho = 0.10$. (a) $\Pr[\mathcal{E}_{\text{MoR}} > 1+\epsilon]$ versus $|S|$ for $\epsilon = 0.05$. (b) Sample size versus ϵ. (c) Sample size versus ρ for $\epsilon = 0.05$. (d) $\Pr[\mathcal{E}_{\text{RoS}} > 1 + \epsilon]$ versus $|S|$ for $\epsilon = 0.05$. (e) Sample size versus ϵ. (f) Sample size versus ρ for $\epsilon = 0.05$. . . 118

3.6 Deezer Music Dataset. Histograms of the node degree for the networks obtained from Croatia, Hungary, and Romania. 119

3.7 Gemsec Deezer dataset. Hungary. ($n = 47,538$). $\Pr[\mathcal{E}_{\mathcal{M}} > 1 + \epsilon]$ versus $|S|$ obtained using (- -) MoR, (- -) RoS,(—) Thm 3.2, (—) Thm. 3.3, and (-·-·-) Thm. 3.4 for $\epsilon = 0.10$ and music genres: (a) Soundtracks, (b) Comedy, (c) Singer & Songwriter, (d) Film-Games, (e) Alternative, and (f) Dance. 121

4.1 The first and second differences for f_t and σ_t are small – derived from denoised COVID-19 cases in California. t is the number of days since January 23, 2020. 144

4.2 γ_f and γ_{σ^2} calculated over various window sizes for COVID-19 reported cases in California. The values are small as desired, particularly for small w. 147

4.3 1st row: MAE vs n, the number of respondents. Indirect with moving average is the best for low n. All moving averages converge as n increases. 2nd row: MAE vs *period* in the survey question. Smoothed results converge as *period* increases. 3rd row: MAE vs *accum*, the accumulation width. Errors reduce quickly as *accum* increases. Larger values make the moving averages slightly worse. 4th row: MAE vs w, the smoothing window. Errors reduce as w increases and increase slightly for large w. 150

4.4 Result of one of the survey simulations. 151

4.5 Normalized COVID-19 incidence estimated from the US CMU-CTIS data for California. The parameters of the proposed approach were set to $accum = 7$ and $w = 1$. 151

4.6 Household size obtained from CMU-CTIS data for California, Texas, New York, and Pennsylvania between September 2020 and March 2022. 165

4.7 Normalized COVID-19 incidence curves and the 95% confidence interval using 10% of the daily samples for the four states and for $accum = 7$ and $w = 2$. 166

4.8 Normalized COVID-19 incidence curves and the 95% confidence interval using 3% of the daily samples for the four states and for $accum = 7$ and $w = 2$. 167

4.9 MAE of the normalized COVID-19 incidence curves using 10% of the daily samples for the four states and for $accum = 7$ and $w = 2$. 168

4.10 MAE vs d, approximate average degree of the respondents. Choice of d does not significantly affect the relative performance of the methods 168

4.11 MAE vs n_d, the subpopulation potentially covered by the sampled respondents. Choice of n_d does not significantly affect the relative performance of the methods 169

4.12 Comparison of estimated incidences from UMD-CTIS surveys. Our approach (IndirectCLI) provides a better estimate than DirectCLI, and even better than asking directly about positive tests (as opposed to indirect on CLI). 170

4.13 MAE of the normalized COVID-19 incidence curves using 3% of the daily samples for the four states and for $accum = 7$ and $w = 2$. 172

5.1 Erdős–Rényi ($\mu = 50$). Histogram of ARD responses y_i for (a) $\rho = 1\%$, (b) $\rho = 5\%$, and (c) $\rho = 10\%$. Histogram of ratios y_i/d_i for (d) $\rho = 1\%$, (e) $\rho = 5\%$, and (f) $\rho = 10\%$. 185

5.2 Scale-Free network model ($\gamma = 2.50$). Histogram of ARD responses y_i for (a) $\rho = 1\%$, (b) $\rho = 5\%$, and (c) $\rho = 10\%$. Histogram of ratios y_i/d_i for (d) $\rho = 1\%$, (e) $\rho = 5\%$, and () $\rho = 10\%$. 186

5.3 Stochastic Block model ($a = e^4$). Histogram of ARD responses y_i for (a) $\rho = 1\%$, (b) $\rho = 5\%$, and (c) $\rho = 10\%$. Histogram of ratios y_i/d_i for (d) $\rho = 1\%$, (e) $\rho = 5\%$, and (f) $\rho = 10\%$. 187

5.4 Erdős–Rényi. Histogram of the degree for (a) $\mu = 15$, (b) $\mu = 50$, and (c) $\mu = 150$. 188

5.5 Erdős–Rényi ($\mu = 50$). RMSRE versus the prevalence without contamination for the groups of robust estimators: (a) MoR-type and (b) RoS-type. 189

5.6 Erdős–Rényi ($\mu = 50$). |NB| versus the prevalence without contamination for the groups of robust estimators: (a) MoR-type and (b) RoS-type. 191

5.7 Erdős–Rényi ($\mu = 50$). RMSRE versus the contamination level for MoR-type methods and prevalences: (a) 1%, (b) 5%, (c) 10%; RoS-type methods and prevalences: (d) 1%, (e) 5%, (f) 10%. . . . 192

5.8 Scale-Free ($\mu = 50$). RMSRE versus the prevalence without contamination for the groups of robust estimators: (a) MoR-type and (b) RoS-type. 194

5.9 Scale-Free network model ($\gamma = 2.50$). RMSRE versus the contamination level for MoR-type methods and prevalences: (a) 1%, (b) 5%, (c) 10%; RoS-type methods and prevalences: (d) 1%, (e) 5%, (f) 10%. 195

5.10 Stochastic Block Model. RMSRE versus the prevalence level for different values of the connectivity factor a. MoR-type methods: (a) $a = e^{-4}$, (b) $a = 1$, (c) $a = e^{4}$; RoS-type methods: (d) $a = e^{-4}$, (e) $a = 1$, (f) $a = e^{4}$. 196

5.11 Stochastic Block Model ($\gamma = 2.50$). RMSRE versus the contamination level for MoR-type methods, prevalence $\rho = 5\%$ and connectivity factors: (a) $a = e^{-4}$, (b) $a = 1$, (c) $a = e^{4}$; RoS-type methods, prevalence $\rho = 5\%$ and connectivity factors: (d) $a = e^{-4}$, (e) $a = 1$, (f) $a = e^{4}$. 198

5.12 Histograms of ARD responses for (a) Sick Month, (b) Hospitalization Month, and (c) Death Month (d) Sick Today, (e) Sick 7 Days, and (c) Vaccines. 200

5.13 Histograms of ratios (a) PP, (b) PSOE, (c) Vox, and (d) Sumar. . 204

5.S.1 Erdős–Rényi ($\mu = 50$). Boxplots yielded by robust MoR-type methods in scenarios without contamination for different prevalence levels. 216

5.S.2 Erdős–Rényi ($\mu = 50$). Boxplots yielded by robust RoS-type methods in scenarios without contamination for different prevalence levels. 216

5.S.3 Erdős–Rényi ($\mu = 15$). RMSRE versus the prevalence without contamination for robust estimators: (a) MoR-type and (b) RoS-type. 217

5.S.4 Erdős–Rényi ($\mu = 15$). The absolute value of the NB versus the prevalence without contamination for the robust estimators: (a) MoR-type and (b) RoS-type. 217

5.S.5 Erdős–Rényi ($\mu = 15$). Boxplots yielded by robust MoR-type methods in scenarios without contamination for different prevalence levels. 217

5.S.6 Erdős–Rényi ($\mu = 15$). Boxplots yielded by robust RoS-type methods in scenarios without contamination for different prevalence levels. 218

5.S.7 Erdős–Rényi ($\mu = 150$). RMSRE versus the prevalence without contamination for the robust estimators: (a) MoR-type and (b) RoS-type. 218

5.S.8 Erdős–Rényi ($\mu = 150$). The absolute value of the NB across various prevalences without contamination for the robust estimators: (a) MoR-type and (b) RoS-type. 218

5.S.9 Erdős–Rényi ($\mu = 150$). Boxplots yielded by robust MoR-type methods in scenarios without contamination for different prevalence levels. 219

5.S.10 Erdős–Rényi ($\mu = 150$). Boxplots yielded by robust RoS-type methods in scenarios without contamination for different prevalence levels. 219

5.S.11 Erdős–Rényi ($\mu = 50$, $\rho = 1\%$). Boxplots yielded by robust MoR-type methods for different contamination levels ϵ. 220

5.S.12 Erdős–Rényi ($\mu = 50$, $\rho = 1\%$). Boxplots yielded by robust RoS-type methods for different prevalence levels ϵ. 220

5.S.13 Erdős–Rényi ($\mu = 50$, $\rho = 5\%$). Boxplots yielded by robust MoR-type methods for different contamination levels ϵ. 221

5.S.14 Erdős–Rényi ($\mu = 50$, $\rho = 5\%$). Boxplots yielded by robust RoS-type methods for different prevalence levels ϵ. 221

5.S.15 Erdős–Rényi ($\mu = 50$, $\rho = 10\%$). Boxplots yielded by robust MoR-type methods for different contamination levels ϵ. 222

5.S.16 Erdős–Rényi ($\mu = 50$, $\rho = 10\%$). Boxplots yielded by robust RoS-type methods for different prevalence levels ϵ. 222

5.S.17 Erdős–Rényi ($\mu = 50$). |NB| versus the contamination level for MoR-type methods and prevalences: (a) 1%, (b) 5%, (c) 10%; RoS-type methods and prevalences: (d) 1%, (e) 5%, (f) 10%. . . . 223

5.S.18 Erdős–Rényi ($\mu = 50$ and $\rho = 10\%$). RMSRE versus the transmission error level for the groups of robust estimators: (a) MoR-type and (b) RoS-type. 224

5.S.19 Erdős–Rényi ($\mu = 50$ and $\rho = 10\%$). |NB| versus the transmission error level for the groups of robust estimators: (a) MoR-type and (b) RoS-type. 224

5.S.20 Erdős–Rényi ($\mu = 50$ and $\rho = 10\%$). RMSRE versus the number of misreported nodes (recall error) for the groups of robust estimators: (a) MoR-type and (b) RoS-type. 225

5.S.21 Erdős–Rényi ($\mu = 50$ and $\rho = 10\%$). |NB| versus the number of misreported nodes (recall error) for the groups of robust estimators: (a) MoR-type and (b) RoS-type. 225

5.S.22 Scale-Free ($\mu = 50$). Boxplots yielded by robust MoR-type methods in scenarios without contamination for different prevalence levels. 226

5.S.23 Scale-Free ($\mu = 50$). Boxplots yielded by robust RoS-type methods in scenarios without contamination for different prevalence levels. 226

5.S.24 Scale-Free ($\mu = 50$). |NB| versus the prevalence without contamination for the groups of robust estimators: (a) MoR-type and (b) RoS-type. 227

5.S.25 Scale-Free ($\gamma = 2.50$). |NB| versus the contamination level for MoR-type methods and prevalences: (a) 1%, (b) 5%, (c) 10%; RoS-type methods and prevalences: (d) 1%, (e) 5%, (f) 10%. . . . 228

5.S.26 Scale-Free ($\mu = 50$ and $\rho = 10\%$). RMSRE versus the transmission error level for the groups of robust estimators: (a) MoR-type and (b) RoS-type. 229

5.S.27 Scale-Free ($\mu = 50$ and $\rho = 10\%$). |NB| versus the transmission error level for the groups of robust estimators: (a) MoR-type and (b) RoS-type. 229

5.S.28 Scale-Free ($\mu = 50$ and $\rho = 10\%$). RMSRE versus the number of misreported nodes (recall error) for the groups of robust estimators: (a) MoR-type and (b) RoS-type. 230

5.S.29 Scale-Free ($\mu = 50$ and $\rho = 10\%$). |NB| versus the number of misreported nodes (recall error) for the groups of robust estimators: (a) MoR-type and (b) RoS-type. 230

5.S.30 Stochastic Block Model. |NB| versus the prevalence level for different values of the connectivity factor a. MoR-type methods: (a) $a = e^{-4}$, (b) $a = 1$, (c) $a = e^{4}$; RoS-type methods: (d) $a = e^{-4}$, (e) $a = 1$, (f) $a = e^{4}$. 231

5.S.31 SBM. RMSRE versus the factor a for three different prevalence levels. MoR-type methods: (a) $\rho = 1\%$, (b) $\rho = 5\%$, (c) $\rho = 10\%$; RoS-type methods: (d) $\rho = 1\%$, (e) $\rho = 5\%$, (f) $\rho = 10\%$. 232

5.S.32 SBM ($\gamma = 2.50$). |NB| versus the contamination level for MoR-type methods, prevalence $\rho = 5\%$ and connectivity factors: (a) $a = e^{-4}$, (b) $a = 1$, (c) $a = e^{4}$; RoS-type methods, prevalence $\rho = 5\%$ and connectivity factors: (d) $a = e^{-4}$, (e) $a = 1$, (f) $a = e^{4}$. 233

6.1 Age declaration from the network of respondents against the age of the respondent (left). Voting declaration from the network of respondents against the age of respondent (right) (Numbers in percentages). The parties' acronyms are UP-Unidas Podemos, Vox, MM-Más Madrid, PS-Partido Socialista de Madrid, and PP-Partido Popular. 247

6.2 Flowchart of the methodology 256

6.3 Boxplots of declared network size with (left) and without (right) outliers for Madrid surveys. 257

6.S.1 Madrid regional elections, May 2023. Indirect survey collected in the Community of Madrid. Demographics extracted from raw data. (a) Age groups, (b) Education, (c) Gender, and (d) Income. 272

6.S.2 Madrid regional elections, May 2023. Indirect survey collected in the Community of Madrid. Demographics of the survey respondents after filtering out using the criteria of Section 6.3.2. (a) Age groups, (b) Education, (c) Gender, and (d) Income. 276

6.S.3 Bootstrap distributions of seat allocations by party and estimation method. Community of Madrid. Each panel shows the empirical distribution of the number of seats a given party obtains under a specific estimation method. 277

6.S.4 Bootstrap distributions of seat allocations by party and estimation method. Andalusia. Each panel shows the empirical distribution of the number of seats a given party obtains under a specific estimation method. 278

6.S.5 Bootstrap distributions of seat allocations by party and estimation method. Valencian Community. Each panel shows the empirical distribution of the number of seats a given party obtains under a specific estimation method. 279

6.S.6 Estimated prevalence $\hat{\rho}$ in % against the homophily factor α obtained by the different NSUM estimators under evaluation for each party. 282

6.S.7 Mean Relative Error versus the homophily factor α obtained by the different NSUM estimators under evaluation for each party. . 283

7.1 IPFS-Benchmark Dataset. Normalized bias (NB) of prevalence estimates as a function of sample size for six IPFS node attributes. The top row shows results are obtained with the MoR estimator, and the bottom row with the RoS estimator. Solid lines represent the median NB across the 360 snapshots, whereas the shaded regions indicate the $p25$-$p75$ and $p10$-$p90$ percentile intervals. . . . 295

7.2 IPFS-Benchmark Dataset. Comparison of crawl-derived attribute prevalences with the Network Scale-up Method (NSUM) estimates (MoR and RoS) from 10 randomly sampled nodes at a snapshot captured on April 5, 2026. (a) country and (b) cloud-provider. (c)–(h) Temporal evolution of selected attributes. Evolution of (i) Non-cloud and (j) Cloud. 296

7.3 Bucket Dataset. Comparison of crawl-derived attribute prevalences with NSUM estimates (MoR and RoS) using the proposed measurement technique with 10 local nodes at a single snapshot. (a) country (b) cloud-provider. (c)–(h) Temporal evolution of selected attributes. Evolution of (i) Non-cloud and (j) Cloud. . . 297

7.S.1 Bucket Dataset. Comparison of crawl-derived attribute prevalences with NSUM estimates (MoR and RoS) using the proposed measurement technique with 5 NATed nodes at a single snapshot. (a) country (b) cloud-provider. (c)–(h) Temporal evolution of selected attributes. 306

7.S.2 Bucket Dataset. Comparison of crawl-derived attribute prevalences with NSUM estimates (MoR and RoS) using the proposed measurement technique with 5 NoNATed nodes at a single snapshot. (a) country (b) cloud-provider. (c)–(h) Temporal evolution of selected attributes. 307

List of Tables

2.1 Basic NSUM estimators. The rows represent the way to obtain the degree and the columns the size 25

2.2 Main scenarios to be compared. 27

2.3 Characteristics for each basic scenario. 30

2.4 Location of the tables in the article and supplementary material. 31

2.5 Relative subpopulation sizes employed in the simulations. 36

2.6 Best method for each scenario. For each metric, it shows the average of the values corresponding to each parameter. AoR and RoA are not included in the simulations about the number and sizes of known subpopulations because they do not use the known subpopulations. 41

2.7 Popular musical genres and the percentage of users for the Deezer dataset. 45

2.S.1 Basic NSUM estimators. The rows represent the way to obtain the degree, and the columns the size. 56

2.S.2 Metrics for the simulations varying the visibility factor in the W-U scenario. 62

2.S.3 Metrics for the simulations of corrected estimators varying visibility factor in the W-U scenario. 63

2.S.4 Metrics for the simulations varying visibility factor in the W-S scenario. 64

2.S.5 Metrics of the corrected methods and the GNSUM for the simulations varying the visibility factor in the W-S scenario. 65

2.S.6 Metrics for the simulations varying the visibility factor in the P-U scenario. 66

2.S.7 Metrics of the corrected methods and the GNSUM for the simulations varying the visibility factor in the P-U scenario. 67

2.S.8 Metrics for the simulations varying the visibility factor in the P-S scenario. 68

2.S.9 Metrics of the corrected methods and the GNSUM for the simulations varying the visibility factor in the P-S scenario. 69

2.S.10 Metrics for the simulations varying the memory factor in the W-U scenario. 70
2.S.11 Metrics for the simulations varying the memory factor in the W-S scenario. 71
2.S.12 Metrics for the simulations varying the memory factor in the P-U scenario. 72
2.S.13 Metrics for the simulations varying the recall error in the P-S scenario. 73
2.S.14 Metrics for the simulations varying the number of subpopulations in the W-U scenario. 74
2.S.15 Metrics for the simulations varying the number of subpopulations in the W-U scenario with disjoint subpopulations. 75
2.S.16 Metrics for the simulations varying the number of subpopulations in the W-U scenario with small subpopulations. 76
2.S.17 Metrics for the simulations varying the number of subpopulations in the W-U scenario with disjoint and small subpopulations. . . . 77
2.S.18 Metrics for the simulations varying the number of subpopulations in the W-S scenario 78
2.S.19 Metrics for the simulations varying the number of subpopulations in the W-S scenario with disjoint subpopulations. 79
2.S.20 Metrics for the simulations varying the number of subpopulations in the W-S scenario with small subpopulations 80
2.S.21 Metrics for the simulations varying the number of disjoint subpopulations in the W-S scenario with disjoint and small subpopulations. 81
2.S.22 Metrics for the simulations varying the number of subpopulations in the P-U scenario. 82
2.S.23 Metrics for the simulations varying the number of subpopulations in the P-U scenario with disjoint subpopulations. 83
2.S.24 Metrics for the simulations varying the number of subpopulations in the P-U scenario with small subpopulations. 84
2.S.25 Metrics for the simulations varying the number of subpopulations in the P-U scenario with disjoint and small subpopulations. . . . 85
2.S.26 Metrics for the simulations varying the number of subpopulations in the P-S scenario. 86
2.S.27 Metrics for the simulations varying the number of subpopulations in the P-S scenario with disjoint subpopulations. 87

2.S.28 Metrics for the simulations varying the number of subpopulations in the P-S scenario with small subpopulations. 88

2.S.29 Metrics for the simulations varying the number of subpopulations in the P-S scenario with disjoint and small subpopulations. 89

2.S.30 Metrics for the simulations varying the sizes of the subpopulations in the W-U scenario. 90

2.S.31 Metrics for the simulations varying the sizes of the subpopulations in the W-U scenario with disjoint subpopulations. 91

2.S.32 Metrics for the simulations varying the sizes of the subpopulations in the W-S scenario. 92

2.S.33 Metrics for the simulations varying the sizes of subpopulations in the W-S scenario with disjoint subpopulations. 93

2.S.34 Metrics for the simulations varying the sizes of the subpopulations in the P-U scenario. 94

2.S.35 Metrics for the simulations varying the sizes of the subpopulations under the P-U scenario with disjoint subpopulations. 95

2.S.36 Metrics for the simulations varying the sizes of the subpopulations under the P-S scenario. 96

2.S.37 Metrics for the simulations varying the sizes of the subpopulations in the P-S scenario. 97

2.S.38 Metrics for the simulations of the sample size under the W-U scenario. 98

2.S.39 Metrics for the simulations varying the sample size in the W-S scenario. 99

2.S.40 Metrics for the simulations of the sample size under P-U scenario. 100

2.S.41 Metrics for the simulations varying the sample size in the P-S scenario. 101

2.S.42 Metrics of the estimates of several unknown subpopulations, with the subpopulations simulated uniformly over the Deezer network. The numbers in parentheses represent the sizes of the subpopulations. 102

2.S.43 Metrics of the estimates of several unknown subpopulations, with the known subpopulations simulated uniformly and the unknown subpopulation modeled via an SIR process over the Deezer network. The numbers in parentheses represent the sizes of the subpopulations. 103

2.S.44 Metrics of the estimates of several unknown subpopulations, using real subpopulations of the Deezer dataset. The numbers in parentheses represent the sizes of the subpopulations. 103

2.S.45 Best method for each scenario. For each metric, it shows the average of the values corresponding to each parameter. *AoR* and *RoA* are not included in the simulations. 104

2.S.46 Best method for each scenario. For each metric, it shows the average of the values corresponding to each parameter. All estimators are included in the simulations. 104

3.1 Summary of characteristics of friendship networks from Deezer dataset (Rozemberczki et al., 2020). 119

3.2 Summary of Analytical Error Bounds for NSUM Hidden Population Rate Estimators. Sample $S \subseteq V$ is random uniform with $m = |S|$, $R_S = \sum_{v \in S} R_v$, and any $\beta = 1 + \epsilon > 1$, any $\alpha > 0$, and any $\delta \in (0, 1)$. 122

4.1 MAE of the normalized COVID-19 incidence curves estimated from the US CMU-CTIS data for California, Texas, New York, and Pennsylvania and for different values of *accum* and w. 152

4.2 MAE on country-level comparison from UMD-CTIS surveys. . . . 169

5.1 Robust NSUM estimators for each type of classical NSUM. . . . 183

5.2 Estimated percentages of COVID-19 incidence and mortality for UK. 201

5.3 Estimated percentages of vote intention for the different parties in Spain. 203

6.1 Standard NSUM approaches based on sums and averages. 243

6.2 Weights by age for the stratified sampling scheme by region . . . 248

6.3 Spanish general elections, July 2023. Indirect survey collected in the Community of Madrid. Estimation of the percentage of votes by political group and blank votes after filtering out samples that contained a significant number of blank votes. Estimated percentage for the ideological blocs (conservative: PP+Vox and progressive: PSOE+Sumar). 249

6.4 Spanish general elections, July 2023. Indirect survey collected in the Community of Madrid. Estimated percentages of votes by political group using the NSUM approaches after filtering out samples that contained a significant number of blank votes. Extrapolation of proportion estimates to representative seats using the D'Hondt method and comparison with other studies. Official results obtained from the Spanish general elections in Madrid, including (CERA included) and excluding (CERA excluded) the votes of Spanish nationals who reside in foreign countries. 250

6.5 Spanish general elections, July 2023. Indirect survey collected in Andalusia. Estimation of the percentage of votes by political group and blank votes after filtering out responses by using all data cleaning criteria. Estimated percentage for the ideological blocs (conservative: PP+Vox and progressive: PSOE+Sumar). . . 252

6.6 Spanish general elections, July 2023. Indirect survey collected in Andalusia. Estimated percentages of votes by political group using the NSUM approaches after filtering out responses using all data cleaning criteria. Extrapolation of proportion estimates to representative seats using the D'Hondt method and comparison with other studies. Official results obtained from the Spanish general elections in Andalusia, including (CERA included) and excluding (CERA excluded) the votes of Spanish nationals who reside in foreign countries. 253

6.7 Spanish general elections, July 2023. Indirect survey collected in the Valencian Community. Estimation of the percentage of votes by political group and blank votes after filtering out respondents at the top 20% network size who also declare a network voting intention for a specific party above 95%. Estimated percentage for the ideological blocs (conservative: PP+Vox and progressive: PSOE+Sumar). 254

6.8 Spanish general elections, July 2023. Indirect survey collected in the Valencian Community. Estimated percentages of votes by political group using the NSUM approaches after filtering out respondents at the top 20% network size who also declare a network voting intention for a specific party above 95%. Extrapolation of proportion estimates to representative seats using the D'Hondt method and comparison with other studies. Official results obtained from the Spanish general elections in the Valencian Community, including (CERA included) and excluding (CERA excluded) the votes of Spanish nationals who reside in foreign countries. 254

6.S.1 Spanish general elections, July 2023. Indirect survey collected in the Community of Madrid. Estimation of the percentage of votes by political group and blank votes from the raw dataset. Estimated percentage for the ideological blocs (left-wing: PSOE+Sumar and right-wing: PP+Vox). 267

6.S.2 Regional elections: Indirect survey collected in the Community of Madrid. Estimation of the percentage of votes by political group and blank votes from the raw dataset. Estimated percentage for the ideological blocs (left-wing: PSM+MM+UP and right-wing: PP+Vox). 269

6.S.3 Estimation of the percentage of votes by political group for the regional elections. 273

6.S.4 Spanish general elections, July 2023. Indirect survey collected in the Community of Madrid. Estimation of the percentage of votes by political group and blank votes and its 95% confidence interval after filtering out samples that contained a significant number of blank votes. Estimated percentage for the ideological blocs (conservative: PP+Vox and progressive: PSOE+Sumar). . . 273

6.S.5 Spanish general elections, July 2023. Indirect survey in the Community of Madrid. Bootstrap voting intention shares by political group and blank vote, and its 95% confidence interval. 274

6.S.6 Spanish general elections, July 2023. Indirect survey collected in Andalusia. Estimation of the percentage of votes by political group and blank votes and its 95% confidence interval after filtering out samples that contained a significant number of blank votes. Estimated percentage for the ideological blocs (conservative: PP+Vox and progressive: PSOE+Sumar). 274

6.S.7 Spanish general elections, July 2023. Indirect survey in Andalusia. Bootstrap voting intention shares by political group and blank vote, and its 95% confidence interval. 275

6.S.8 Spanish general elections, July 2023. Indirect survey collected in the Valencian Community. Estimation of the percentage of votes by political group and blank votes and its 95% confidence interval after filtering out samples that contained a significant number of blank votes. Estimated percentage for the ideological blocs (conservative: PP+Vox and progressive: PSOE+Sumar). 275

6.S.9 Spanish general elections, July 2023. Indirect survey in the Valencian Community. Bootstrap voting intention shares by political group and blank vote, and its 95% confidence interval. 276

6.S.10 Voting Shares and Seat Allocation for Political Groups in the July 2023 Spanish General Elections: Results from an Indirect Survey in Andalusia with Various NSUM Approaches and Filtering Configurations (with the 95% rule). 285

6.S.11 Voting Shares and Seat Allocation for Political Groups in the July 2023 Spanish General Elections: Results from an Indirect Survey in Andalusia with Various NSUM Approaches and Filtering Configurations (without the 95% rule). 286

Chapter 1

Introduction

This thesis consists of a compendium of six research articles, including both published and under-review manuscripts, that contribute to the development of techniques based on indirect reporting. This introductory chapter presents the fundamental concepts and background required for the remainder of the thesis and provides an overview of its structure. Its main objective is to familiarize the reader with the notation and estimators used throughout the document, while also providing a concise introduction to key concepts. In this way, the chapter aims to make the thesis as self-contained and accessible as possible. Section 1.1 introduces the types of data collected through indirect reporting, together with their main applications and properties. Section 1.2 describes the methods of size estimation based on indirect reporting. Section 1.3 presents several network models that capture social and communication network structures. Finally, Section 1.4 outlines the structure of the thesis and connects it to the main research challenges addressed throughout this work.

1.1. Aggregated Relational Data (ARD)

The problem of estimating the size of a group or population in a social network is long-standing and remains very common. In the absence of data, researchers rely on surveys conducted on a sample of individuals. However, in sensitive groups, the participants may not be willing to participate or respond honestly, leading to unreliable estimates. An important example is hard-to-reach populations, that is, populations whose sampling frame does not exist, and for whom public acknowledgment is potentially threatening (Heckathorn, 1997).

To estimate hard-to-reach populations, researchers commonly use alternatives to traditional survey methods, such as snowball sampling (Goodman, 1961) and respondent-driven sampling (Heckathorn, 1997). In these approaches, participants recruit future subjects from within their social networks. However, because these sampling procedures depend heavily on the social network, these procedures present some limitations, including recruitment bias caused by homophily in social ties (Gile and Handcock, 2010), inability to reach the populations due to insufficient network connectivity (Gile et al., 2015), and declining participation rates (Crawford et al., 2018).

Given these limitations, an alternative is the collection of *Aggregated Relational*

Data (ARD), which consists of the number of contacts that a person has in certain groups. This was proposed by Bernard et al. (1989) to estimate hard-to-reach populations following the 1985 Mexico earthquake. ARD leverages the underlying social network structure without observing any link of the network, making it feasible to implement through survey instruments, where ARD is collected through questions of the type "How many people do you know who have characteristic X?".

ARD-based surveys do not require respondents to disclose their own membership in the groups, thereby enhancing privacy and encouraging participation. In addition, each participant provides information about the surrounding neighborhood, effectively increasing the fraction of the network for which data are available. Particularly, this approach has been effective in reducing data collection costs by approximately $70 - 80\%$ in socioeconomic networks while yielding the same conclusion that the entire network (Breza et al., 2020).

The advantages of ARD have motivated a wide range of applications. ARD has been primarily used to estimate the size of populations in social networks, including applications in public health (Teo et al., 2019; Garcia-Agundez et al., 2021), stigmatized populations (Ezoe et al., 2012; Sully et al., 2020), and impact of disasters Bernard et al. (2001). ARD has also been employed to estimate the personal network sizes of a subset of individuals (McCormick et al., 2010; Lubbers et al., 2019; Hofstra et al., 2021), and to characterize demographic features of hard-to-reach populations (McCormick and Zheng, 2012, 2015). Another application is the estimation of network properties Breza et al. (2020) and model parameters Breza et al. (2023).

1.2. Network Scale-up Methods (NSUM)

The *Network Scale-up Methods* (NSUM) comprises a series of techniques for estimating the size or proportion of a group or population within a social network using ARD. To that purpose, the NSUM techniques use different scaling procedures based on the personal network size of the participants, which are denoted as *degrees*, following the terminology from graph theory. The main idea behind NSUM is that the surrounding networks of the participants possess representative information of the total population. Concretely, if y_{ij} is the ARD of individual i about the group j, d_i the degree of i, N_j is the size of group j, and N the size of the population, NSUM methods assume that

$$\frac{y_{ij}}{d_j} \approx \frac{N_j}{N}. \tag{1.1}$$

Since its early development, NSUM research has been primarily application-driven, focusing on the estimation of sensitive hard-to-reach populations of social relevance. These domains include drug users (Salganik et al., 2010), members of the LGTBQ+ community (Ezoe et al., 2012), cancer patients (Carletti et al., 2017), religious populations (Yang and Yang, 2017), university students' high-risk behaviors (Sajjadi

et al., 2018), sex workers (Jing et al., 2018), abortions (Sully et al., 2020), victims of massive terrorist attacks (Bernard et al., 2001), and HIV infections (Teo et al., 2019), currently used by The Joint United Nations Programme on HIV/AIDS (UNAIDS) (UNAIDS, 2010). A more detailed list of many applications is included in (Ocagli et al., 2021).

1.2.1. Biases of the NSUM

Response behavior and lack of random mixing may violate the assumption of representativeness of the ARD given in (1.1), leading to several sources of bias for the NSUM (Laga et al., 2021). Quantifying the impact of these biases is challenging since these biases can manifest in multiple ways. Specifically, the biases are four (Laga et al., 2021): transmission error, recall error, barrier effects, and response bias.

Transmission error Transmission error occurs when the participants are unaware whether their contacts belong to the specified groups (Killworth et al., 2006). This violates the assumption that respondents have complete knowledge of the groups being asked. Although this bias may affect many groups, its magnitude can vary considerably across groups (Killworth et al., 2006). Transmission error is commonly quantified through the transmission rate or visibility factor, which measures the proportion of contacts who are aware whether the respondent belongs to the groups. Estimating the transmission rate is relevant for correcting some NSUM estimators (Feehan and Salganik, 2016). However, in practice, obtaining such estimates involves a sample from the targeted group (Feehan and Salganik, 2016), eliciting expert opinion (Rastegari et al., 2014), or using game-like questions to estimate unknown group membership (Salganik et al., 2011).

Recall error Recall error is a bias related to reporting behavior. It is produced when participants do not remember their personal network composition. This error can be exacerbated as the survey progresses, and can be amplified by time constraints for each question imposed in data collection (Killworth et al., 2006). The respondents may both undercount and overcount the number of contacts that they know in some groups. Previous research has shown that participants tend to overcount the number of contacts in small groups while undercounting those in large groups (McCarty et al., 2001). Additional factors contributing to recall error include rounding of responses to multiples of five and guessing the number of contacts rather than counting them in larger populations (McCarty et al., 2001). Several studies have examined the relationship between the ARD of a group and group sizes, and have proposed correction methods based on these exploratory analyses (Killworth et al., 2003; Zheng et al., 2006; McCormick et al., 2007; Maltiel et al., 2015).

Barrier effect Barrier effect arises when the propensity to know someone in a group depends on the characteristics of the respondent, thereby violating Eq. (1.1). These effects may result from homophily, social distance, and other sociodemographic factors (Killworth et al., 2006). Preventing barrier effects a priori is difficult because of their multiple underlying causes. A study involving more than 3000 responses found substantial differences in the number of known contacts across subpopulations for nearly all sociodemographic variables, including questions about the number of contacts with a specific name (Killworth et al., 2006). To address this issue, Feehan and Salganik (2016) proposed a correction for one NSUM estimator based on the *degree ratio*, defined as the ratio between the average number of connections within a group and the average number of connections in the overall population. However, implementing this correction requires a sample of the target group, which may be difficult to obtain in practice.

Response bias Response bias refers to deliberate misreporting of the number of contacts that belong to the specified groups. Although this bias is not specific to the NSUM, it may produce substantial deviations of the estimates (McCarty et al., 2001). In some situations, participants are not willing to recognize knowing members of stigmatized groups, such as female sex workers (Jing et al., 2018). In other cases, they provide malicious responses or random responses, often involving substantially unrealistically large values. Methods to mitigate malicious responses are very limited and rely on ad hoc cut-offs (Zheng et al., 2006; McCarty et al., 2001) and standard filtering procedures (Garcia-Agundez et al., 2021). From a survey design perspective, the randomized response technique has also been applied (Jing et al., 2018). In this approach, participants are split into two groups, with one group answering an unrelated question instead of the sensitive one, allowing researchers to estimate the mean response using the proportion of responses about the sensitive and the unrelated question.

1.2.2. NSUM estimators

Consider a network of size N; an *unknown group* is a subset of nodes of size N_u, where N_u is not known. The ARD of the unknown group for the individual i is denoted as y_{iu}, and d_i is the degree of i. NSUM methods estimate the size of the group N_u or its prevalence $\rho_u = N_u/N$ using the ARD from a sample of size n. Note that an estimate of the size induces an estimate of the prevalence and vice versa using the following relation $N\hat{\rho} = \hat{N}_u$. Simple NSUM estimators assume that the surrounding networks are representative (on average) of the prevalence of the unknown group. Concretely, the simple NSUM estimators are defined as follows.

- The *Mean of Ratios* (MoR) estimator is the average of the individual ratios of the

participants, defined as

$$N_u^{\text{MoR}} = N\frac{1}{n}\sum_{i=1}^{n}\frac{y_{iu}}{d_i}. \tag{1.2}$$

- The *Ratio of Sums* (RoS) estimator is the ratio of the sum of the number of reported neighbors in the unknown group over the sum of the degrees of the participants, defined as

$$N_u^{\text{RoS}} = N\frac{\sum_{i=1}^{n} y_{iu}}{\sum_{i=1}^{n} d_i}. \tag{1.3}$$

MoR was proposed assuming that Eq (1.1) is representative on average. Subsequently, Killworth et al. (1998b) introduced the RoS as the maximum likelihood estimator of the Binomial model $y_i \sim Bin(d_i, \rho_u)$, which has become the most widely used NSUM estimator in the literature (see, for example, Ocagli et al. (2021)).

The degree of the sample nodes may not be known, which is usually the case in survey data. A possible solution is to ask the participants about their personal network size. In this case, the degree corresponds to the reported ARD about the total number of nodes in the social network, a group with size $N - 1$. Where necessary, the terms *Average of Ratios* (AoR) and *Ratio of Averages* (RoA) will be used throughout this document to refer to the MoR and RoS estimators, in cases where the degree is reported directly by the participants.

Known Population Method

Obtaining the degree is crucial for the NSUM. However, respondents often struggle to recall the number of their contacts (Killworth et al., 1990), particularly when the definition of contacts becomes broader (Feehan et al., 2016). The estimation of the degrees is very frequent with hard-to-reach populations, whose sizes are very small and tend to be connected to the rest of the population through weak ties, i.e., acquaintances rather than close friends, making it challenging for the participants to recall their acquaintance networks.

To mitigate this issue, the *Known Population Method* (KPM) has been proposed, which estimates the degree of the sample nodes using the ARD of groups whose sizes are known. Specifically, the ARD response of individual i for group k is denoted by y_{ik}, for $k = 1, \ldots, L$, while N_k represents the known size of group k. These groups are commonly referred to as *known groups* or *known subpopulations* because their sizes are available. Examples include physicians and self-employed workers whose numbers can often be obtained from public administrative records. In contrast to earlier approaches, KPM estimators have been developed under both frequentist and Bayesian frameworks.

Frequentist KPM estimators Frequentist estimators use a first-step procedure, estimating the degree with the ARD of the known groups, and then plug in these

estimates to produce a size estimation for N_u. There are two proposals for degree estimation. The first strategy assumes the model $y_{ik} \sim Bin(d_i, N_k/N)$ for $k = 1, \ldots, L$, and then estimates the degrees via a maximum likelihood estimation:

$$\hat{d}_i^{(1)} = N \cdot \frac{\sum_{k=1}^{L} y_{ik}}{\sum_{k=1}^{L} N_k}. \tag{1.4}$$

Another option uses Eq (1.1) to obtain L estimates of d_i, $d_i^k = y_{ik}N/N_k$, and then takes the average:

$$\hat{d}_i^{(2)} = \frac{1}{L} \sum_{k=1}^{L} d_i^k. \tag{1.5}$$

The degree estimators are plugged into the MoR and RoS formulas to obtain several size estimators. *Plug-in Maximum Likelihood Estimator* (PiMLE) (Killworth et al., 1998a) combines the MoR with the fist degree estimator

$$N_u^{PiMLE} = N \frac{1}{n} \sum_{i=1}^{n} \frac{y_{iu}}{d_i^{(1)}}. \tag{1.6}$$

The introduction of $d_i^{(1)}$ into the RoS gives to the "*Maximum Likelihood Estimator*" (MLE), which is the most employed NSUM (Ocagli et al., 2021)

$$N_u^{MLE} = N \frac{\sum_{i=1}^{n} y_{iu}}{\sum_{i=1}^{n} d_i^{(1)}}. = \sum_{i=1}^{n} y_{iu} \frac{\sum_{k=1}^{L} N_k}{\sum_{i=1}^{n} \sum_{k=1}^{L} y_{ik}}. \tag{1.7}$$

The *Means of Sums* (MoS) (Habecker et al., 2015) uses the estimated degrees $d_i^{(2)}$ in the formula of MoR

$$N_u^{MoS} = N \frac{1}{n} \sum_{i=1}^{n} \frac{y_{iu}}{d_i^{(2)}}. \tag{1.8}$$

The last combination involves the degree estimation based on averages, using $d_i^{(2)}$ and RoS (Kunke et al., 2024)

$$N_u^{KLN} = N \frac{\sum_{i=1}^{n} y_{iu}}{\sum_{i=1}^{n} d_i^{(2)}} \tag{1.9}$$

Bayesian KPM estimators Several Bayesian NSUM estimators were proposed to incorporate bias corrections and variability of ties and network size. These estimators are based on hierarchical models, with posteriors calculated with MCMC algorithms. The first model was proposed by Zheng et al. (2006) to handle the overdispersion produced by the variation of propensities of the groups to form connections. Concretely, the model is

$$y_{ij} \sim NegativeBinomial\left(mean = e^{\alpha_i + \beta_j}, overdispersion = \omega_j\right), \tag{1.10}$$

where ω_j is the parameter measuring the overdispersion, e^{α_i} is the expected degree of i, and e^{β_j} is the proportion of links of group j, where the group may be known or unknown. In addition, they assume Normal prior distributions for α_i and β_j, a

uniform prior in the interval $(0, 1)$ to $1/\omega_j$, and propose a Gibbs sampling. Zheng et al. (2006) include further details about the prior distributions and algorithm. The estimator based on this model is denoted as the *Overdispersed* estimator.

Teo et al. (2019) proposed an estimator based on the following model, which is denoted as the TPC estimator

$$y_{ij} \sim Poisson\left(\lambda \alpha_i N_j\right), \tag{1.11}$$

where λ is a scale-up parameter such that $y_{ij} = \lambda N_j$, and α_i characterizes the variability in the network size. This model is complemented with uniform prior distributions with large support for the N_k, $N(0, \tau)$ for α_i, and uniform priors between 0 and 10 for τ and λ. Teo et al. (2019) also provide improvements, introducing demographic characteristics and Likert scales.

Generalized Network Scale-up Method The *Generalized Network Scale-up Method* (GNSUM) is an improvement proposed by Feehan and Salganik (2016). This estimator is based on additional data from a sample of the unknown group about the number of contacts within several groups that know the participant's membership in the unknown group. These groups may differ from the known groups of the KNP. The GNSUM estimator is

$$\hat{N}_u^{GNSUM} = \frac{\sum_{i \in s_F} y_{iu}/\pi_i}{\hat{\tilde{v}}_{u,F}}, \tag{1.12}$$

where s_F is a sample from the frame population, π_i are the inclusion probabilities, and $\hat{\tilde{v}}_{u,F}$ is the estimated relative visibility, which is defined as

$$\hat{\tilde{v}}_{u,F} = \frac{N_F}{N_{\mathcal{A}}} \frac{\sum_{i \in s_U} \sum_j \tilde{v}_{i,\mathcal{A}_j}/c\pi_i}{\sum_{i \in s_U} 1/(c\pi_i)}, \tag{1.13}$$

where s_U is a sample from the unknown groups, $\mathcal{A}_j$ is the group j, $N_{\mathcal{A}}$ is the sum of the sizes of the groups, N_F the number of individuals of the frame, $\tilde{v}_{i,\mathcal{A}_j}$ the number of people in the group j who knows that participant i belongs to the unknown population, and c a constant of proportionality. GNSUM is unbiased under the assumption that the number of in-reports (number of reported contacts) is equal to the number of out-reports (number of respondents that report an individual), and other conditions related to the visibility of the unknown group, which are explained in detail in (Feehan and Salganik, 2016).

1.3. Network models

Methods based on ARD aim to infer properties of node attributes from partial network information, specifically from the number of connections with nodes with specific features. These inferences can be affected by the complex structure of network

topology, including node popularity, clustering, node distance in the network, and the propensity to form ties due to group membership. To capture these characteristics, a variety of network models have been used throughout this thesis.

1.3.1. Social network models

Social interactions within real-world populations give rise to complex network structures. Over the years, numerous network models have been proposed to reproduce key characteristics commonly observed in empirical networks, such as degree distribution, path length, and community structure. Relevant examples are Erdős–Rényi networks, small-world networks, scale-free networks, and stochastic block models.

Erdős–Rényi

The Erdős–Rényi model is one of the classical random graph models, which has been widely used to study network properties. In this model, every pair of edges has the same probability of forming a link $p = \mu/(N-1)$, where μ represents the mean degree and N is the network size (Cheng et al., 2019). In this model, the edges are formed independently, producing uniform degree distributions with low clustering. The distribution of the ARD y_{ij} conditional on the degree d_i follows a Hypergeometric distribution. When the network and the group are large, this distribution approximates the Binomial distribution $y_{ij} \sim Bin(d_i, \rho_j)$, which is the distribution that many NSUM estimators assume, including the most used NSUM RoS and MLE.

Small-world networks

Small-world networks contain a family of graphs whose average path length is short, and therefore, every part of the network can be accessed with a chain of a few neighbors. This property was found for the first time in the experiment of the social psychologist Stanley Milgram (Milgram et al., 1967). In these experiments, the researchers asked random participants to forward letters to a certain target. They found that roughly a third of the letters eventually arrived at the target, in a median of six steps (Easley et al., 2012).

The Watts-Strogatz model is a type of small-world network that extends the Erdős–Rényi model. In this model, the nodes are labeled and connected to their nearest neighbors, then, with a fixed probability, an edge is removed and placed between two nodes chosen at random. This ensures the small-world properties as well as a high clustering coefficient. Similar to the Erdős–Rényi model, the Watts-Strogatz model networks have unrealistic degree distributions, with distributions that have pronounced peaks where many nodes have the same degree.

Scale-free networks

The degree of the nodes can have a lot of variation in several social networks. In some of them, a small subset of individuals is more popular and has more ties with other people, while most of the individuals have low degrees, such as collaboration networks (Newman, 2018) and financial networks (Soramäki et al., 2007). Scale-free networks aim to model this effect. These networks are characterized by their degree distribution, which follows a power law distribution, i.e., the probability $p(d)$ that a node in the network has d connections to other nodes is expressed as $d^{-\gamma}$.

There are several models for scale-free networks. The model of Goh et al. (2001) adds links based on weights to the nodes. The nodes are indexed with integer numbers, and the ith node receives a weight proportional to $i^{-\alpha}$, for $\alpha \in [0, 1)$. the degree distribution follows a power law with $\gamma = (\alpha + 1)/\alpha$. This model can have nodes with very low degrees or even isolated nodes. An alternative to solve this issue is a preferential attachment network (Newman, 2018). In this case, a fixed number of m edges are added successively to every node so that this node forms m connections with the rest of the nodes. These m nodes are selected with probabilities proportional to the current degrees.

Stochastic Block Model

In some social networks, individuals exhibit different propensities to form connections depending on their membership in particular groups, which constitutes a form of barrier effect. Stochastic block models are designed to capture these community patterns between K disjoint groups. In this case, a node belongs to one of these groups; the probability of an edge between two nodes depends solely on the groups to which the nodes belong. More specifically, the probabilities of connection are determined by a symmetric matrix $P = (P_{lm})$, where P_{lm} denotes the probability of an edge between a node in group l and a node in group m (Holland et al., 1983). This framework allows the model to represent homophily when intra-group connection probabilities are higher than inter-group probabilities, and heterophily in the opposite case.

1.3.2. Peer-to-peer networks

Peer-to-peer (P2P) networks are a type of virtual network in which the devices, referred to as peers, communicate directly with their neighbors to share resources without relying on centralized servers. These networks have been widely applied for file sharing (Pouwelse et al.; Cohen, 2003; Benet, 2014), and decentralized payment systems (Buterin, 2014; Tschorsch and Scheuermann, 2016). Representative examples include BigTorrent, Gnutella (Ripeanu et al., 2002), the the InterPlanetary File

System (IPFS) (Benet, 2014), Bitcoin (Nakamoto, 2008), and Ethereum (Wood et al., 2014).

A subclass of P2P networks, known as structured P2P networks, imposes logical structures upon the connections based on distance metrics. Each node maintains a routing table containing information about its neighbors, such as their network identifiers, IP addresses, and additional metadata, which allow efficient routing through the network towards a given target node. Processes known as *crawling* exploit the routing table and logical structure to enumerate all nodes and their routing tables. Starting from an initial small set of nodes (*e.g.*, new nodes joining the network), crawlers iteratively issue targeted routing queries to progressively uncover the network topology and routing tables.

Kademlia

Kademlia (Maymounkov and Mazières, 2002) is an structured P2P of many modern peer-to-peer (P2P) systems (*e.g.*, BitTorrent, IPFS). Every device that joins the Kademlia network is randomly assigned an ID in a large key space of l bits, typically 256 bits. Distances are measured based on the XOR distance, which is defined as the integer value resulting from the binary number from the exclusive OR of the node IDs (Maymounkov and Mazières, 2002). A node v maintains up to k connections, typically $k = 20$, uniformly at random for every range distance $[2^i, 2^{i+1})$ for $i = 0, \ldots, l-1$. These nodes are listed in the routing tables, which are known as the k-buckets. The first bucket contains peers that share no common prefix with v, with each consecutive bucket containing peers that share one more bit of common prefix.

1.4. Main contributions and structure of the thesis

The contributions of this thesis are divided into different chapters. These chapters address issues identified during the process of this thesis. The identified research challenges are the following.

(1) *Biases.* NSUM estimation techniques suffer from several biases arising from specific reporting behavior in indirect surveys and deviations from the random-mixing assumptions. Quantifying the impact of these biases is challenging since these biases can manifest in multiple ways. Among them, transmission error and barrier effect are the only biases that have been analyzed and corrected in the traditional Killworth et al. (1998b) estimator (MLE) (Feehan and Salganik, 2016). However, to mitigate these biases, the social visibility of the unknown group is required, which substantially increases data collection efforts. In practice, this may involve obtaining a sample from the unknown group (Feehan

and Salganik, 2016), eliciting expert opinion (Rastegari et al., 2014), or using game-like questions to estimate the unknown group membership (Salganik et al., 2011).

(2) *Known Population Method.* An accurate estimate of the participants' network sizes is essential for the NSUM. However, recalling the number of contacts is challenging for the participants (Killworth et al., 1990), particularly when the definition of "known" is broad enough to include weak ties. This issue occurs in many studies that define contacts as individuals who interact in person, by phone, or online (Sully et al., 2020). A common approach to mitigate this problem is the Known Population Method, which estimates the degrees with the questions about the known groups, which in turn are easier for the respondents to recall. While accurate degree estimation is crucial for NSUM, there is limited understanding of how to select these known groups, such as the number and sizes of the groups, the effect of the intersection of the groups, and their impact on the final estimates.

(3) *Underlying network.* ARD gives local information about the personal network of an individual. Surveys with small sample sizes are likely to have independent ARD variables. However, when the sample size increases, dependencies naturally arise since multiple nodes report the same contacts. This issue is crucial since the structure of these dependencies is influenced by the underlying network topology and the link formation dynamics. This raises important questions about which types of networks and NSUM estimators produce more reliable estimates and how estimation accuracy varies across different network structures. Despite its importance, the impact of different network structures and the distribution of the population across them, for example, in the context of disease spread in epidemiology, has received limited attention.

(4) *Sampling.* Despite the wide applicability of the estimation methods using ARD, there are no clear guidelines for selecting the sample size. The randomness of the estimated degrees may be a problem since the plug-in procedures may underestimate uncertainty and produce overly confident inference. Additionally, NSUM estimators assume that the ARD observations are independent, and when the sample size increases, dependencies arise, and some nodes are counted multiple times. This can affect the performance of the NSUM estimator, even when the entire population is sampled. To the best of our knowledge, only a recent study (Josephs et al., 2024) proposes a sample size heuristic for only one of the NSUM estimators using a Central Limit Theorem when the personal network sizes are known.

(5) *Analytical guarantees* NSUM has been widely used to estimate the size of sensitive groups in epidemiology and public health. However, there are no analytical guarantees about the error that the NSUM estimators can have, or

whether this error is lower than that of traditional methods using direct surveys. One reason is the presence of dependencies that arise as the sample size increases, violating the assumption of independence of ARD observations. Another challenge is that the same observations may correspond to different underlying social network structures. Additionally, some estimators involve ratios of random variables due to the estimated or reported degrees, complicating theoretical analysis. A recent study (Josephs et al., 2024) proposes a heuristic for one estimator, but it assumes perfect knowledge of the degrees.

(6) *Temporal dimension.* Online surveys enable the collection of massive amounts of social data in short periods of time. One of these cases is online indirect surveys collecting ARD, such as the COVID-19 Trends and Impact Surveys (Astley et al., 2021; Salomon et al., 2021) and the Coronasurveys project (Baquero et al., 2021). These works produce point estimates rather than estimating the temporal trends. The ability to capture temporal evolution with ARD is relevant to enhance the applicability, such as the study of seasonality, forecasting future values, and identification of turning points, which is crucial for taking informed actions to address the potential challenges or threats of these subpopulations.

(7) *Robustness.* Many NSUM methods rely on sample means, which are highly sensitive to deviations, so that a small number of outliers can substantially distort the results. This issue is particularly relevant in the context of NSUM because the data is collected through surveys that may include sensitive or stigmatizing questions, making responses prone to misreporting due to social desirability or sympathy toward the target population. Respondents may provide inaccurate responses intentionally by reporting extreme values or simply by responding randomly. Despite these challenges, there is limited work on methods to mitigate the impact of outliers.

(8) *Effectiveness beyond hard-to-reach populations.* NSUM have been applied in several cases in epidemiology and public health. Due to their simplicity and privacy-preserving properties, NSUM estimators have been used primarily to estimate hard-to-reach populations in social networks. However, the potential advantages in other scenarios beyond hard-to-reach populations are unexplored, particularly as an alternative to traditional survey methods. NSUM techniques assume that the membership of the groups that respondents should report is well defined. The performance of NSUM under perceived or uncertain memberships, however, remains unclear. Another scenario that has not been analyzed is online networks, whose network structures rapidly change, and monitoring the information is demanding, like peer-to-peer networks.

This thesis advances the previous research gaps in four directions: empirical analysis, theoretical analysis, new proposals, and new domains of applications. The

empirical analysis is included in Chapter 2, which addresses the research questions (1)-(4). The theoretical analysis focuses on the questions (3)-(6), and is contained in Chapter 3 and Chapter 4. The new proposed methods in Chapter 5 cover the research problems (1), (3), and (7). Finally, the new application of Chapter 6 relate to the topics (1) and (8)

Chapter 2, published in *The American Statistician*, analyzes the performance of most of the NSUM estimators in the literature through an extensive simulation environment. The aim is to assess the sensitivity of the estimators to their assumptions and enhance the survey design process. The simulation environment incorporates well-established social network models, such as Watts-Strogatz and preferential attachment networks, alongside different configurations of known groups. Unknown groups are distributed uniformly and with barrier effects produced via SIR processes on the network. Within this setting, estimator performance is analyzed by varying the key parameters: number, sizes, and intersection of subpopulations, transmission error, recall error, and sample size. In addition, the simulation results are validated using real networks with real and synthetically generated groups.

Chapter 3, published in *ACM SIGKDD Conference on Knowledge Discovery and Data Mining,* develops a theoretical framework for analyzing the error of the NSUM estimators, concretely, the multiplicative relative error. This work explores the configuration of groups and network structures that produce identical ARD observations, as well as the dependencies induced in the ARD variables. As a consequence, this paper derives analytical error bounds in two scenarios. In the first case, an adversarial setting is considered, where both the network structure and the distribution of the unknown group are chosen to yield the worst possible estimation. In the second case, the article focuses on error bounds when the underlying network has been generated through certain random structures.

Chapter 4, published in *AAAI Conference on Artificial Intelligence*, analyzes the use of indirect reporting techniques for estimating temporal trends. The increasing availability of online surveys has enabled the continuous collection of large-scale data over time, including some surveys containing ARD. This chapter proposes a framework for monitoring temporal trends using indirect reporting. Concretely, this research article develops a procedure for monitoring the temporal trends and incorporates the trend dynamics through temporal aggregations. It also provides a theoretical comparison of its advantages over traditional survey methods. The effectiveness of the indirect reporting is evaluated through simulations of dynamic processes and with real data.

Chapter 5, published in *Social Networks*, is dedicated to the robustness of the NSUM estimators within the framework of robust statistics. Classical NSUM estimators rely on sample means, which makes them highly sensitive to outliers and deviations from model assumptions. To address this limitation, this work

proposes several robust extensions of NSUM, replacing the sample means with different location estimators, including approaches such as the median, trimmed mean, adaptive methods, and M-estimators, each tailored to different types of data contamination and distributional deviations. The performance of the robust proposals together with the traditional NSUM is evaluated through a simulated scenario designed to assess both common properties of robust statistics and the biases of the NSUM. The results are validated with real-world datasets of COVID-19 and voting intention.

Chapter 6, published in *Social Networks*, studies the application of NSUM for estimating voting intention in electoral processes. The information about voting intention is a privacy-sensitive issue, in which direct surveys have not been exempt from public controversy due to poor estimates. This work proposes a methodology for estimating voting intention using NSUM based on collected ARD containing the perceived voting intention of the participants' contacts. This methodology comprises a unified framework that ranges from the design of survey questions, the sampling procedures, outlier filters, and the implementation of the NSUM methods. The methodology is evaluated in three regions of the July 2023 Spanish general elections.

Chapter 7 develops a lightweight framework for monitoring network attributes in structured P2P networks, which are very relevant for content dissemination, decentralized payment systems, and blockchain technologies. Existing measurement approaches rely on crawlers that exploit the routing mechanism to gather information about all nodes and their connections. This work introduces an NSUM-based framework that estimates network properties with a much less invasive approach using only the routing tables of a small number of nodes. The proposed methodology is evaluated using 360 IPFS snapshots collected between April 1 2026 and May 1 2026, and with an empirical implementation using data from 10 nodes.

The chapters of this thesis present methodological contributions for the techniques based on NSUM and ARD published in the fields of statistics, data mining, machine learning, and sociology, as well as implementation in network measurements, which is currently under development. These contributions extend the traditional applications beyond public health and epidemiology. Although NSUM and ARD are relatively recent, they have experienced substantial growth in recent years, with more than 150 publications appearing between 2020 and 2025 (Lubbers et al., 2026). The theoretical developments, empirical evaluations, and diverse applications presented throughout this thesis highlight the potential of ARD and NSUM as a promising and rapidly expanding interdisciplinary research area. Due to the interdisciplinary nature of this work, terminology and notation have been adapted to the conventions of the different disciplines, which may result in minor variations across chapters.

Bibliography

C. M. Astley et al. Global monitoring of the impact of the COVID-19 pandemic through online surveys sampled from the Facebook user base. Proceedings of the National Academy of Sciences, 118(51), 2021.

C. Baquero, P. Casari, A. Fernandez Anta, A. García-García, D. Frey, A. García-Agundez, C. Georgiou, B. Girault, A. Ortega, M. Goessens, et al. The coronasurveys system for covid-19 incidence data collection and processing. Frontiers in Computer Science, 3:641237, 2021.

J. Benet. IPFS - Content Addressed, Versioned, P2P File System. CoRR, abs/1407.3561, 2014. URL http://arxiv.org/abs/1407.3561.

H. R. Bernard, E. C. Johnsen, P. D. Killworth, and S. Robinson. Estimating the size of an average personal network and of an event subpopulation. In M. Kochen, editor, The Small World, pages 159–175. Ablex Publishing, Norwood, NJ, 1989.

H. R. Bernard, P. D. Killworth, E. C. Johnsen, G. A. Shelley, and C. McCarty. Estimating the ripple effect of a disaster. Connections, 24(2):18–22, 2001.

E. Breza, A. G. Chandrasekhar, T. H. McCormick, and M. Pan. Using aggregated relational data to feasibly identify network structure without network data. American Economic Review, 110(8):2454–84, 2020.

E. Breza, A. G. Chandrasekhar, S. Lubold, T. H. McCormick, and M. Pan. Consistently estimating network statistics using aggregated relational data. Proceedings of the National Academy of Sciences, 120(21):e2207185120, 2023.

V. Buterin. Ethereum: A next-generation smart contract and decentralized application platform. 2014. URL https://ethereum.org/content/whitepaper/whitepaper-pdf/Ethereum_Whitepaper_-_Buterin_2014.pdf.

G. Carletti, N. Soriani, M. Mattiazzi, and D. Gregori. Suppl-1, m6: A social network approach to the estimation of perceived quality of health care. The Open Nursing Journal, 11:219, 2017.

S. Cheng, D. J. Eck, and F. W. Crawford. Estimating the size of a hidden finite set: large-sample behavior of estimators, 2019. URL https://arxiv.org/abs/1808.04753.

B. Cohen. Incentives build robustness in bittorrent. In Workshop on Economics of Peer-to-Peer systems, volume 6, pages 68–72, 2003. URL http://jmvidal.cse.sc.edu/library/cohen03a.pdf.

F. W. Crawford, J. Wu, and R. Heimer. Hidden population size estimation from respondent-driven sampling: a network approach. Journal of the American Statistical Association, 113(522):755–766, 2018.

D. Easley, J. Kleinberg, et al. Networks, crowds, and markets. Cambridge Books, 2012.

S. Ezoe, T. Morooka, T. Noda, M. L. Sabin, and S. Koike. Population size estimation of men who have sex with men through the network scale-up method in japan. PloS one, 7(1):e31184, 2012.

D. M. Feehan and M. J. Salganik. Generalizing the network scale-up method: a new estimator for the size of hidden populations. Sociological methodology, 46(1): 153–186, 2016.

D. M. Feehan, A. Umubyeyi, M. Mahy, W. Hladik, and M. J. Salganik. Quantity versus quality: A survey experiment to improve the network scale-up method. American journal of epidemiology, 183(8):747–757, 2016.

A. Garcia-Agundez, O. Ojo, H. A. Hernández-Roig, C. Baquero, D. Frey, C. Georgiou, M. Goessens, R. E. Lillo, R. Menezes, N. Nicolaou, et al. Estimating the covid-19 prevalence in spain with indirect reporting via open surveys. Frontiers in Public Health, 9:658544, 2021.

K. J. Gile and M. S. Handcock. 7. respondent-driven sampling: An assessment of current methodology. Sociological methodology, 40(1):285–327, 2010.

K. J. Gile, L. G. Johnston, and M. J. Salganik. Diagnostics for respondent-driven sampling. Journal of the Royal Statistical Society Series A: Statistics in Society, 178(1):241–269, 2015.

K.-I. Goh, B. Kahng, and D. Kim. Universal behavior of load distribution in scale-free networks. Physical review letters, 87(27):278701, 2001.

L. A. Goodman. Snowball sampling. The annals of mathematical statistics, pages 148–170, 1961.

P. Habecker, K. Dombrowski, and B. Khan. Improving the network scale-up estimator: Incorporating means of sums, recursive back estimation, and sampling weights. PloS one, 10(12):e0143406, 2015.

D. D. Heckathorn. Respondent-driven sampling: a new approach to the study of hidden populations. Social problems, 44(2):174–199, 1997.

B. Hofstra, R. Corten, and F. van Tubergen. Beyond the core: who has larger social networks? Social Forces, 99(3):1274–1305, 2021.

P. W. Holland, K. B. Laskey, and S. Leinhardt. Stochastic blockmodels: First steps. Social networks, 5(2):109–137, 1983.

L. Jing, Q. Lu, Y. Cui, H. Yu, and T. Wang. Combining the randomized response technique and the network scale-up method to estimate the female sex worker population size: an exploratory study. Public Health, 160:81–86, 2018.

N. Josephs, D. M. Feehan, and F. W. Crawford. A sample size formula for network scale-up studies. Sociological Methods & Research, 53(3):1252–1289, 2024.

P. D. Killworth, E. C. Johnsen, H. R. Bernard, G. A. Shelley, and C. McCarty. Estimating the size of personal networks. Social Networks, 12(4):289–312, 1990.

P. D. Killworth, E. C. Johnsen, C. McCarty, G. A. Shelley, and H. R. Bernard. A social network approach to estimating seroprevalence in the united states. Social networks, 20(1):23–50, 1998a.

P. D. Killworth, C. McCarty, H. R. Bernard, G. A. Shelley, and E. C. Johnsen. Estimation of seroprevalence, rape, and homelessness in the united states using a social network approach. Evaluation review, 22(2):289–308, 1998b.

P. D. Killworth, C. McCarty, H. R. Bernard, E. C. Johnsen, J. Domini, and G. A. Shelley. Two interpretations of reports of knowledge of subpopulation sizes. Social networks, 25(2):141–160, 2003.

P. D. Killworth, C. McCarty, E. C. Johnsen, H. R. Bernard, and G. A. Shelley. Investigating the variation of personal network size under unknown error conditions. Sociological Methods & Research, 35(1):84–112, 2006.

J. P. Kunke, I. Laga, X. Niu, and T. H. McCormick. Comparing the robustness of simple network scale-up method estimators. Sociological methodology, 54(2): 385–403, 2024.

I. Laga, L. Bao, and X. Niu. Thirty years of the network scale-up method. Journal of the American Statistical Association, 116(535):1548–1559, 2021. doi: 10.1080/01621459.2021.1935267. URL https://doi.org/10.1080/01621459.2021.1935267. PMID: 37994314.

M. J. Lubbers, J. L. Molina, and H. Valenzuela-García. When networks speak volumes: Variation in the size of broader acquaintanceship networks. Social Networks, 56:55–69, 2019.

M. J. Lubbers, B. Völker, and M. Bojanowski. From hidden populations to social structure: Evolution of the network scale-up method, aggregated relational data, and their applications, 2026.

R. Maltiel, A. E. Raftery, T. H. McCormick, and A. J. Baraff. Estimating population size using the network scale up method. The annals of applied statistics, 9(3): 1247, 2015.

P. Maymounkov and D. Mazières. Kademlia: A peer-to-peer information system based on the XOR metric. In P. Druschel, M. F. Kaashoek, and A. I. T. Rowstron, editors, Peer-to-Peer Systems, First International Workshop, IPTPS 2002, Cambridge, MA, USA, March 7-8, 2002, Revised Papers, volume 2429 of Lecture Notes in Computer Science, pages 53–65. Springer, 2002.

C. McCarty, P. D. Killworth, H. R. Bernard, E. C. Johnsen, and G. A. Shelley. Comparing two methods for estimating network size. Human organization, 60(1): 28–39, 2001.

T. H. McCormick and T. Zheng. Latent demographic profile estimation in hard-to-reach groups. The annals of applied statistics, 6(4):1795, 2012.

T. H. McCormick and T. Zheng. Latent surface models for networks using aggregated relational data. Journal of the American Statistical Association, 110(512):1684–1695, 2015.

T. H. McCormick, T. Zheng, et al. Adjusting for recall bias in "how many x's do you know?" surveys. In Proceedings of the joint statistical meetings, 2007.

T. H. McCormick, M. J. Salganik, and T. Zheng. How many people do you know?: Efficiently estimating personal network size. Journal of the American Statistical Association, 105(489):59–70, 2010.

S. Milgram et al. The small world problem. Psychology today, 2(1):60–67, 1967.

S. Nakamoto. Bitcoin: A peer-to-peer electronic cash system. 2008.

M. Newman. Networks. Oxford university press, 2018.

H. Ocagli, D. Azzolina, G. Lorenzoni, S. Gallipoli, M. Martinato, A. S. Acar, P. Berchialla, D. Gregori, I. S. Group, et al. Using social networks to estimate the number of covid-19 cases: The incident (hidden covid-19 cases network estimation) study protocol. International journal of environmental research and public health, 18(11):5713, 2021.

J. Pouwelse, P. Garbacki, D. Epema, and H. Sips. The bittorrent p2p file-sharing system: Measurements and analysis. In P2P Systems IV, Lecture Notes in Computer Science, pages 205–216. Springer. ISBN 978-3-540-31906-1. doi: 10.1007/11558989_19.

A. Rastegari, M. R. Baneshi, S. Haji-Maghsoudi, N. Nakhaee, M. Eslami, H. Malekafzali, and A. A. Haghdoost. Estimating the annual incidence of abortions in iran applying a network scale-up approach. Iranian Red Crescent Medical Journal, 16(10), 2014.

M. Ripeanu, I. T. Foster, and A. Iamnitchi. Mapping the gnutella network: Properties of large-scale peer-to-peer systems and implications for system design. CoRR, cs.DC/0209028, 2002. URL https://arxiv.org/abs/cs/0209028.

H. Sajjadi, Z. Jorjoran Shushtari, M. Shati, Y. Salimi, M. Dejman, M. Vameghi, S. Karimi, and Z. Mahmoodi. An indirect estimation of the population size of students with high-risk behaviors in select universities of medical sciences: A network scale-up study. PloS one, 13(5):e0195364, 2018.

M. Salganik, D. Fazito, M. Mello, et al. Estimating the number of heavy drug users in curitiba, brazil using multiple methods. Technical report, Technical report. UNAIDS, 2010.

M. J. Salganik, M. B. Mello, A. H. Abdo, N. Bertoni, D. Fazito, and F. I. Bastos. The game of contacts: estimating the social visibility of groups. Social networks, 33(1):70–78, 2011.

J. A. Salomon et al. The us covid-19 trends and impact survey: Continuous real-time measurement of covid-19 symptoms, risks, protective behaviors, testing, and vaccination. Proceedings of the National Academy of Sciences, 118(51):e2111454118, 2021.

K. Soramäki, M. L. Bech, J. Arnold, R. J. Glass, and W. E. Beyeler. The topology of interbank payment flows. Physica A: Statistical Mechanics and its Applications, 379(1):317–333, 2007.

E. Sully, M. Giorgio, and S. Anjur-Dietrich. Estimating abortion incidence using the network scale-up method. Demographic Research, 43:1651–1684, 2020.

A. K. J. Teo, K. Prem, M. I. Chen, A. Roellin, M. L. Wong, H. H. La, and A. R. Cook. Estimating the size of key populations for hiv in singapore using the network scale-up method. Sexually transmitted infections, 95(8):602–607, 2019.

F. Tschorsch and B. Scheuermann. Bitcoin and beyond: A technical survey on decentralized digital currencies. IEEE Commun. Surveys Tuts., 18(3):2084–2123, 2016. doi: 10.1109/COMST.2016.2535718.

W. UNAIDS. Guidelines on estimating the size of populations most at risk to hiv. Geneva, Switzerland: World Health Organization, 51, 2010.

G. Wood et al. Ethereum: A secure decentralised generalised transaction ledger. Ethereum project yellow paper, 2014.

X. Y. Yang and F. Yang. Estimating religious populations with the network scale-up method: A practical alternative to self-report. Journal for the scientific study of religion, 56(4):703–719, 2017.

T. Zheng, M. J. Salganik, and A. Gelman. How many people do you know in prison? using overdispersion in count data to estimate social structure in networks. Journal of the American Statistical Association, 101(474):409–423, 2006.

Chapter 2

Performance Analysis of NSUM Estimators in Social-Network Topologies

In *The American Statistician*, 79(2), 247–264, 2025.
https://doi.org/10.1080/00031305.2024.2421361

Sergio Díaz-Aranda[1,2] José Aguilar[1,3] Juan Marcos Ramírez[1]
David Rabanedo[2] Antonio Fernández Anta[1] Rosa E. Lillo[2,4]

[1] IMDEA Networks Institute
[2] uc3m-Santander Big Data Institute, Universidad Carlos III de Madrid
[3] Universidad de Los Andes
[4] Department of Statistics, Universidad Carlos III de Madrid

Abstract

The Network Scale-up Methods (NSUM) are methods to estimate unknown populations based on indirect surveys in which the participants provide information about aggregated data of their acquaintances. This preserves the privacy and may lead to higher participation. During the last thirty years, new NSUM estimators have emerged. However, conditions related to the design of the experiments and the robustness of the estimators have not been studied in depth, especially in a realistic simulation environment. This study aims to compare nine NSUM estimators under relevant conditions in the literature through simulation experiments. We have analyzed how the NSUM is affected by the network topology, transmission and recall errors, the distribution of the unknown subpopulation, the number and sizes of subpopulations, and sample size. This article shows that some barely used NSUM estimators are better and more robust under some conditions, especially when the network is scale-free or under barrier effects. In addition, some methods are very sensitive to recall errors. In terms of the subpopulations configuration, we observe that the number of known subpopulations usually employed is quite large and that the most common NSUM is robust to the number and sizes of the subpopulations.

Keywords: Aggregated relational data, Indirect surveys, Size estimation, Ego-networks

2.1. Introduction

The Network Scale-up Methods (NSUM) are survey-based methods that use information about the acquaintances of the participants instead of their own. The methods arise from the fact that surveys with sensitive populations face reliability issues since the respondents might not be willing to participate or to be honest, resulting in sampling bias and underestimation. Over the last few decades, indirect reporting has emerged as a strategy to address these issues, which can increase participation and data collection while reducing study costs. The NSUM studies only require aggregate data of the contact network, which is obtained through the reporting of different groups known by the respondent (what is called aggregated relational data (ARD) (Feehan and Salganik, 2016)), using considerably less data and respecting the privacy of the surveyed. For these reasons, the NSUM have been employed to estimate the size of subpopulations in which no sampling frame exists and public acknowledgment of membership in the population is potentially threatening, where the term subpopulation denotes a subset of the general population, and size refers to the number of people in the subpopulation. Based on the work of Bernard et al. (1989), it has been employed in several contexts such as abortions (Sully et al., 2020), female sex workers (Jing et al., 2018), men who have sex with men (Ezoe et al., 2012), drug users (Salganik et al., 2010), and HIV issues, currently used by the UNAIDS[1] (UNAIDS, 2010).

Throughout these years, much research about NSUM has emerged (see, for example, the reviews (Laga et al., 2021) and (McCormick, 2020)) focusing mainly on the introduction of new estimators, especially in the presence of bias (see (Laga et al., 2021)). However, the performance and comparison of the methods under various conditions have not been studied in depth. These conditions include the performance under biases, sample size, the unknown subpopulation, and the selection of the known subpopulation. We will refer to an unknown population as a subpopulation whose size is not known, and to known subpopulations as subpopulations whose size is known and are used to estimate the network size of the participants. In particular, there is no guideline on the sizes of the known subpopulations chosen, the number of known subpopulations to employ, or the influence of the intersection of the known subpopulations, which we will denote as overlapping of subpopulations. Most of the studies only compare two estimators (Killworth et al., 1998b), (Salganik et al., 2011), (Snidero et al., 2004) (the last two only in one dataset), or three (Kunke et al., 2024). A comparison of many of them was done in the supplementary material of (Laga et al., 2021), but for a single dataset. Snidero et al. (2004) analyzed two NSUM estimators in Monte Carlo experiments to analyze the robustness under the size, coverage, optimal number, and overlapping of the populations. Kunke et al. (2024) compared the estimators given in (Killworth et al., 1998a) and (Killworth

[1]The Joint United Nations Programme on HIV/AIDS

et al., 1998b) under the prevalence of the unknown subpopulation, dissortative and assortative mixing, and degree ratio. The number of subpopulations required for the first NSUM estimator was studied by (Killworth et al., 1998a) in an elementary Monte Carlo simulation. Srivastava et al. (2024) proposed a latent network model obtaining analytical conditions in terms of variance and dependency with respect to the network size, in which the NSUM error is lower than classical methods relying solely on participant information. Finally, analytical error bounds for two classical NSUM estimators were provided in (Díaz-Aranda et al., 2024), with refinements for random networks.

The main contribution is the comparison of a large set of NSUM estimators, nine in total, through an extensive simulation study under the presence of the relevant conditions of the NSUM literature. Specifically, we explore the impact of the distribution of the subpopulations (number of subpopulations, sizes, presence of overlapping), the distribution of the unknown subpopulation, sample size, and the presence of the next biases: transmission error and recall error. Actually, to our knowledge, it is the first work that compares the performance of the NSUM under the transmission and recall errors. Additionally, we present a way to model the transmission error and recall error biases. Finally, we measure the NSUM performance in a realistic simulation environment, where we included social network topologies that can be found in real data, such as small-world and preferential attachment networks.

This article is organized as follows. Section 2 presents a historical overview of the NSUM and the biases faced by the estimators. Section 3 displays the NSUM estimators that will be used in the subsequent simulations. In Section 4, we present the parameters and domains that will be compared. The results of the simulations are exhibited in Section 5, along with a discussion of estimator performance. Section 6 assesses the results on a real network using both real and simulated subpopulations. Final remarks are in Section 7.

2.2. Overview of the Network Scale-Up Methods

The NSUM are the methods that employ ARD to estimate unknown subpopulations or personal network sizes (Laga et al., 2021). The ARD is the data collected from the ego-networks (networks surrounding the subjects that contain the participants and their contacts (Newman, 2018)) using questions of the type: "How many people do you know who have characteristic X?". The individual from whom the networks are generated is called ego, and the contacts are denoted as alters. The general framework of all the methods is that the ego-network contains information that is approximately representative of the entire population, at least for the ARD questions. More concretely, suppose that we have a population of size N and several ARD

questions; the main assumption is that the following equations hold:

$$\mathbb{E}\left[\frac{y_{ij}}{d_i}\right] = \frac{N_j}{N} \tag{2.1}$$

where y_{ij} represents the number of people that participant i knows in the subpopulation j, d_i is the degree, the size of the ego-network, and N_j is the total number of people belonging to the subpopulation j. We will consider L known subpopulations, with known sizes denoted as N_k, and an unknown subpopulation. We will refer to y_{iu} and N_u as the number of people that participant i knows in the unknown subpopulation, and the size of the unknown subpopulation, which is often called the "hidden population" or "hard-to-reach population". The NSUM uses the ARD provided by the respondents to obtain values for the participants' degrees. This is mainly done using known subpopulations, although other approaches ask directly for the degrees (Garcia-Agundez et al., 2021). Additional samples of the unknown population are employed in some estimators (Feehan and Salganik, 2016) and (Verdery et al., 2019).

Over the years, several biases have been discovered that affect the constant proportionality assumption, and as a consequence, the performance of the NSUM. These biases are the barrier effect, transmission error, and recall error. These errors come from survey behavior and the way the different subpopulations are distributed. The transmission error is the bias produced by the fact that the participant does not know if the acquaintances are in the subpopulations. The recall error occurs when the participants recall inaccurately the number of alters of some subpopulations. The participant may overestimate or underestimate this number. The barrier effect is the bias produced because the probability of knowing someone in a population depends on the characteristics of the ego. This means that a participant is more likely to know someone in certain subpopulations.

2.3. NSUM estimators

Laga et al. (2021) reviews the previous NSUM in the literature from which we will take most of them. For simplicity, we will exclude estimators that use demographic variables (McCormick and Zheng, 2012), and venue-based sampling (Verdery et al., 2019). We include the estimator of (Garcia-Agundez et al., 2021), and two unemployed estimators, one of them pointed out by (Kunke et al., 2024). During the development of the paper, a new NSUM has emerged (Laga et al., 2023), that takes into account the correlations between the variables. We do not analyze the correlation of the variables in our work, but the overlapping of the known subpopulations may lead to correlated ARD responses because the participants are more likely to know a person in several subpopulations. Our results do not show a significant variation in the performance of the estimators in this case, as we show in the supplementary material.

2.3.1. Frequentist estimators

The frequentist estimations usually use a two-step procedure: an estimation of the degrees of the subjects, and a posterior estimation of the unknown subpopulation using that information. In terms of the degrees, in some cases, they use the fundamental Eq. (2.1) for the unknown subpopulations to obtain L estimates of the d_i, and then take the average:

$$\hat{d}_i^A = \frac{1}{L}\sum_{k=1}^{L} d_i^k, \tag{2.2}$$

where the d_i^k is the estimate corresponding to $y_{ik}N/N_k$. Another option is to assume that the y_{ik} are L independent random variables that follow a binomial distribution, $y_{ik} \sim Bin(d_i, N_k/N)$, and then estimate the degrees via a maximum likelihood estimation:

$$\hat{d}_i^{ML} = N \cdot \frac{\sum_{k=1}^{L} y_{ik}}{\sum_{k=1}^{L} N_k}. \tag{2.3}$$

For the estimation of the size of the unknown population, there are two options. One proposal is to use Eq. (2.1) to obtain n estimates of the number of the unknown population, $N_u^i = Ny_{iu}/d_i$, and to take averages,

$$\hat{N}_u^A = \frac{1}{n}\sum_{i=1}^{n} N_u^i. \tag{2.4}$$

Another proposal for the size estimation is to assume that $y_{iu} \sim Bin(d_i, N_u/N)$ for $i = 1, \cdots, n$ are independent random variables, and to use a maximum likelihood estimation

$$\hat{N}_u^{ML} = N\frac{\sum_{i=1}^{n} y_{iu}}{\sum_{i=1}^{n} d_i}. \tag{2.5}$$

The NSUM proposal is to substitute the degrees for the estimated degrees in the equations for the size estimation, Eq. (2.4) and Eq. (2.5). An alternative is to employ the approach of (Garcia-Agundez et al., 2021), when participants provide the degree, which we denote as δ_i. These six combinations are summarized in the next Table. The MLE estimator is by far the most employed NSUM, see (Ocagli et al., 2021) for a list of NSUM studies.

Degree \ Size	Average	Max. Likelihood
Average	MoS (Means of Sums) (Killworth et al., 1998a)	KLN (Kunke et al., 2024)
Max. Likelihood	Plug-in MLE (PiMLE) (Killworth et al., 1998a)	MLE (Killworth et al., 1998b)
Direct	Average of Ratios (AoR)	Ratio of Averages (RoA) (Garcia-Agundez et al., 2021)

Table 2.1: Basic NSUM estimators. The rows represent the way to obtain the degree and the columns the size

2.3.2. Bayesian estimators

The first estimator is the Overdispersed estimator (Zheng et al., 2006). It was created to handle the overdispersion of the data produced by the different propensities of the

subpopulation to form ties. This is represented by the parameter ω_k. The model is:

$$y_{ik} \sim Negative\ Binomial\left(mean = e^{\alpha_i+\beta_k}, overdispersion = \omega_k\right). \tag{2.6}$$

where $k = 1, \cdots, L$ or u, e^{α_i} is the expected degree of i, and e^{β_k} is the proportion of links of group k. For further details about the prior distribution and the algorithm, see (Zheng et al., 2006).

On the other hand, Teo et al. (2019) proposed a new estimator that we denote by the TPC estimator. They presented the following model:

$$y_{ik} \sim Poisson\left(\lambda\alpha_i N_k\right), \tag{2.7}$$

with λ be a scale-up parameter such that $y_{ik} = \lambda N_k$, for every $k = 1, \cdots, L$ or u; and α_i a parameter that characterizes the variability in the network size. Details about the algorithm can be found in (Teo et al., 2019). This work includes other improvements to the model through questions about perceptions and demographic variables. These last adjustments are not included because they exceed the scope of the study.

Maltiel et al. (2015) proposed a Bayesian model and 4 modifications to handle the biases. The Maltiel's estimators include priors whose support is bounded. This feature generates acceptance rate problems in our data. For that reason, we have not included this model in the rest of the article. For more details about our simulations with these estimators, see the supplementary material.

2.3.3. Generalized Network Scale-Up Method

The Generalized Network Scale-up Method (GNSUM) was suggested by Feehan and Salganik (2016) to reduce the transmission error. In addition, the GNSUM allows other types of sampling. However, this estimator needs an additional sample from the unknown subpopulation. This sample from the unknown subpopulation allows to collect data about the number of alters in some groups who are aware of the participants belonging to the unknown subpopulation. This data is known as enriched aggregate relational data (EARD). The purpose of EARD is to estimate the relative visibility of the unknown subpopulations. The relative visibility refers to the average of the visibility of each individual in the unknown subpopulation, which is the number of contacts that know the belonging of the individual to the unknown subpopulation. The EARD questions measure the visibility of the participants of the unknown subpopulations in some groups, denoted as alter groups. Similar to other NSUM estimators, the EARD is used to estimate the relative visibility and then estimate the unknown subpopulation. The GNSUM estimator is:

$$\hat{N}_u^{GNSUM} = \frac{\sum_{i \in s_F} y_{iu}/\pi_i}{\hat{\bar{v}}_{u,F}}, \tag{2.8}$$

where π_i are the inclusion probabilities from a sampling design and $\hat{\hat{v}}_{u,F}$ is the estimated relative visibility. In our specific case, we consider the known subpopulations as alter groups, and the sample is taken uniformly. More details about the GNSUM are in Section 1.2 in the supplementary material.

Adjusted estimators. Feehan and Salganik (2016) proved that the GNSUM estimator is equal to the unknown subpopulation estimate Eq. (2.5) corrected by some adjustment factors, one of them the inverse of the transmission rate. The transmission rate is the fraction of a respondent's network that is aware that the respondent belongs to the unknown subpopulation. However, the transmission rate estimation also needs the information of a sample of the unknown subpopulation. Feehan and Salganik (2016) proposed using the MLE estimator corrected by these factors. It is also known as GNSUM. Apart from this adjustment, we consider the adjustment of the other NSUM estimators by the inverse of an estimate of the transmission rate, which we will denote in the rest of the paper as adjusted NSUM estimators.

2.4. Simulation Environment and Parameters

We define four main scenarios characterized by the type of network and distribution of the unknown population, as shown in Table 2.2. For each main scenario, we define the belonging to the subpopulations, the ARD, and the EARD. Then, we compare, for several samples, the performance of the different methods as the other relevant parameters vary. These parameters of the other relevant scenarios are transmission error, recall error, number of known subpopulations, sizes of known subpopulations, presence of disjoint subpopulations (subpopulations with no overlapping), and sample size. The NSUM estimators are compared using the following metrics: mean, standard deviation, mean absolute error (MAE), and mean squared error (MSE).

2.4.1. Main scenarios

	Network	unknown subpopulation
W-U	Watts-Strogatz	Uniform
W-S	Watts-Strogatz	SIR
P-U	Pref. attachment	Uniform
P-S	Pref. attachment	SIR

Table 2.2: Main scenarios to be compared.

Networks

We use two types of networks with properties found in real networks. On the one hand, we use a small-world network. Therefore, it has a high clustering coefficient and

a short average path length (Newman, 2018). Concretely, we use a Watts-Strogatz network. The network is built as follows: the nodes are placed and are connected to their nearest neighbors, and with a fixed probability, an edge is removed, and two nodes are connected. On the other hand, we use a scale-free network, a network generated by a preferential attachment process. The construction of the graph is done by adding m vertices successively so that the vertices join proportionally to the node with the most connections. The degree distribution follows a power distribution, unlike the Watts-Strogatz networks, whose nodes have similar degrees. More details about the network topologies are in the supplementary material.

Unknown subpopulation distribution

The first case is when the individuals of the unknown subpopulation are randomly chosen. In this setting, the probability is the same for every node and independent of the network. On the other hand, we model the dependencies via an SIR process through the network. It is a foundational model for infections, which models some use cases from NSUM, such as drug uses (Salganik et al., 2010) and COVID-19 (Garcia-Agundez et al., 2021). In this model, there are three groups: susceptible, infected (unknown subpopulation), and recovered from the infection. A node infects a neighbor with probability β, and the nodes are recovered with probability γ, which cannot be infected anymore.

2.4.2. Bias modeling

Transmission error

We model the transmission error through a parameter $\nu \in [0, 1]$ that we call the visibility factor. This number represents the probability that a person who knows an acquaintance of the unknown subpopulation is informed of the belonging of this acquaintance in the unknown subpopulation. We assume that the transmission error uniformly affects the nodes in the unknown subpopulation, and then the expected value of y_{iu} is:

$$E[y_{iu}] = \nu d_i \frac{N_u}{N} = d_i \frac{N_u}{N/\nu}. \tag{2.9}$$

In our setting, we can correct the estimators by multiplying them by $1/\nu$.

A measure related to the transmission error is the transmission rate (Laga et al., 2021) (Feehan and Salganik, 2016). This is one of the factors proposed by Feehan and Salganik (2016) to correct the MLE. The expected value of the transmission rate is ν. We estimate this factor using the coming-out-rate (Ezoe et al., 2012). This method requires a sample of the unknown subpopulation. It divides the average number of acquaintances who are aware of belonging to the unknown subpopulation by their average personal network size.

Recall error

The recall error is the bias due to memory inaccuracies. We model the recall error through a non-negative number ρ that we call the memory factor. There is no barrier error when ρ is zero, and the bias is more pronounced as ρ increases. We create the report of the different variables by rounding (to two decimals) a truncated normal distribution whose mean is the real value, and the standard deviation is the product of the real value times ρ. The truncation of the normal distributions is because they may produce unrealistic data. Additionally, we assume symmetry for simplicity. We only include the details for the modeling of the unknown population responses. The rest is in the supplementary material.

Recall that y_{iu} is the response of the participant i about the ARD question of the unknown subpopulation. There are two natural bounds for the responses of the unknown population: $y_{iu} \geq 0$ and $y_{iu} \leq \delta_i$, where δ_i is the response of the size of the ego-network of i. But, we add some constraints to get a distribution as symmetrical as possible, $a = max(-0.5, 2x_{iu} - \delta_i + 0.5)$ and $b = min(2x_{iu} + 0.5, \delta_i + 0.5)$. The truncated normal distribution has mean $\mu = x_{iu}$, standard deviation $\sigma = x_{iu}\rho$, lower bound a, and upper bound b. The differences of the intervals from the mean are both equal to $min(x_{iu}, \delta_i - x_{iu})$. We add and subtract 0.5 so that the boundaries of the normal can be reached. In addition, the mean and bounds are also selected to have the expected value of the real personal size. Finally, later, y_{iu} is obtained by taking a value of the distribution and rounding it.

Barrier effect

This bias is not compared systematically in this paper. However, the SIR produces barrier effects. Additionally, the degree of variation is not considered in the design phase and may produce this bias.

2.4.3. Motivation for the other parameters

Subpopulations

More than 20 subpopulations are usually employed because the PiMLE needed more than 20 subpopulations to be essentially unbiased in a simple Monte Carlo simulation. However, it is barely employed, and the simulations of Snidero et al. (2004) showed in their simulation study that the number did not affect the performance of the MLE and PiMLE. Concerning the sizes of the subpopulations, some estimators are more affected if the subpopulations have unequal sizes, (Habecker et al., 2015). They also suggested using homogeneous subpopulations for the MoS to outperform the estimates, and because the variance is higher when the subpopulations are small.

Snidero et al. (2004) also shows that homogeneous subpopulations improve the estimates, and the bias of the PIMLE depends on the size of the subpopulations. The overlapping of the subpopulations is another source of error. Snidero et al. (2004) showed that the bias of the PiMLE increased impressively when a significant overlapping arose, whereas the effect in the MLE is limited.

Sample size

Most of the NSUM estimators are unbiased if the degree estimation is perfect. However, this is not possible in most cases. Moreover, the NSUM estimators assume that the information of the ARD is independent, and therefore, the participants are not connected. In practice, this condition may also not hold, affecting the estimates. Another issue is that there is a lack of guidelines about the sample size that the researchers should choose for their studies.

2.5. Simulation Study

We construct the four basic scenarios considering networks of $N = 10000$ nodes to make the population as big as possible but computationally feasible. We chose a Watts-Strogatz network with 25 initial nodes connected to each side and the probability of changing the connection equal to 0.1. For the case of the preferential attachment network, we pick a network with a linear preferential attachment (power equal to one) and a number of edges added each time equal to 25 to get a similar average degree (49.935 compared to 50 of the Watts-Strogatz). For the distribution of the unknown subpopulation, we set 100 nodes at random in every case, and we implement an SIR process in each network. The first one is a SIR with $\beta = 0.046$ and $\gamma = 0.01$, where we choose the process at iteration 5. The other process, the SIR of the preferential attachment, has $\beta = 0.039$ and $\gamma = 0.010$, and the SIR is chosen at iteration 2. We select parameters of the SIR model such that the size of the unknown subpopulations is 100 in order to compare with the uniform case.

Simulation goal	Visibility factor (ν)	Memory factor (ρ)	Number known subp.	Sizes known subp.	Presence of disjoint subp.	Sample size	Section in the paper
Trans. error	[0,1]	0	20	Baseline	No	50	2.5.1
Barrier error	1	[0,1]	20	Baseline	No	50	2.5.1
Subpop. number	1	0	2-20	Even (big & small)	Yes/No	500	2.5.1
Subpop.sizes	1	0	20	All (see Table 2.5)	Yes/No	500	2.5.1
Sample size	1	0	20	Baseline	No	[25,3000]	2.5.1

Table 2.3: Characteristics for each basic scenario.

The types of simulations that we conduct are summarized in Table 2.3. We vary one parameter as the other parameters remain fixed. The rows represent the family of simulations. The parameter that is measured is shown in the first column. For the fifth column, the Baseline subpopulations refer to the configuration S_0 of Table 2.5.

These are the sizes of the subpopulations that we use, except in those simulations where we observe the effect of the subpopulations. In the case of even subpopulations in the third row, it refers to subpopulations of sizes 0.1 and subpopulations of sizes 0.01. The rest of the subpopulation sizes employed can be found in Table 2.5. The sixth column represents the presence of cases with disjoint known subpopulations.

Comparison between estimators Some NSUM estimators require more information than others. In particular, the GNSUM estimator requires a sample from the unknown subpopulation, which is a lot of information compared to other estimators. The same is true for the adjusted estimators. Similarly, the RoA and AoR use the information on the personal network size of the participants. In the cases where the recall error is not present, the RoA and AoR have perfect knowledge of the size of the ego-network, and hence, the comparison with the rest of the estimators is not fair. However, the results of the RoA and AoR help to understand the behavior of other estimators closely related to them (see Table 2.1).

2.5.1. Simulations results

We show the performance of the NSUM estimators under the scenarios of Table 2.5.1. We present figures with boxplots and radar charts for the mean and MAE for some simulation parameters. In addition, we show tables in the supplementary material whose rows represent the value of the estimators for the mean, standard deviation, MAE, and MSE, and the columns represent the parameters that it is being analyzed. The tables with all the parameters are included for a detailed comparison between the estimators. The Table 2.4 shows the location of the figures and tables. The Tables in the supplementary material appear under the reference 2.S.number.

Simulation	W-U scenario	W-S scenario	P-U scenario	P-S scenario
Trans. error	Figures 2.1 and 2.2, and Tables 2.S.2 and 2.S.3	Tables 2.S.4 and 2.S.5	Tables 2.S.6 and 2.S.7	Tables 2.S.8 and 2.S.9.
Barrier error	Figure 2.3 and Table 2.S.10	Figure 2.4 and Table 2.S.11	Table 2.S.12	Table 2.S.13
Subpop. number	Figure 2.5 and Tables 2.S.14 and 2.S.15	Tables 2.S.18 and 2.S.19	Tables 2.S.22 and 2.S.23	Tables 2.S.26 and 2.S.27
Subpop. number small	Tables 2.S.16 and 2.S.17	Tables 2.S.20 and 2.S.21	Tables 2.S.24 and 2.S.25	Tables 2.S.28 and 2.S.29
Subpop. sizes	Figure 2.6 and Tables 2.S.30 and 2.S.31	Tables 2.S.32 and 2.S.33	Tables 2.S.34 and 2.S.35	Figure 2.7 and Tables 2.S.36 and 2.S.37
Sample size	Figure 2.8 and Table 2.S.38	Table 2.S.39	Figure 2.9 and Table 2.S.40	Table 2.S.41

Table 2.4: Location of the tables in the article and supplementary material.

Transmission error simulations

We compare the NSUM estimators while the transmission error varies, which is modeled with the visibility factor ν. We provide two figures showing the performance of the nine NSUM estimators (Figure 2.1) and the adjusted estimators (Figure 2.2) in the W-U scenario. These figures do not include the Overdispersed estimator because of its bad performance.

We observe that the GNSUM is an estimator that works well for every scenario and value of the visibility factor, especially when the visibility factor is greater than

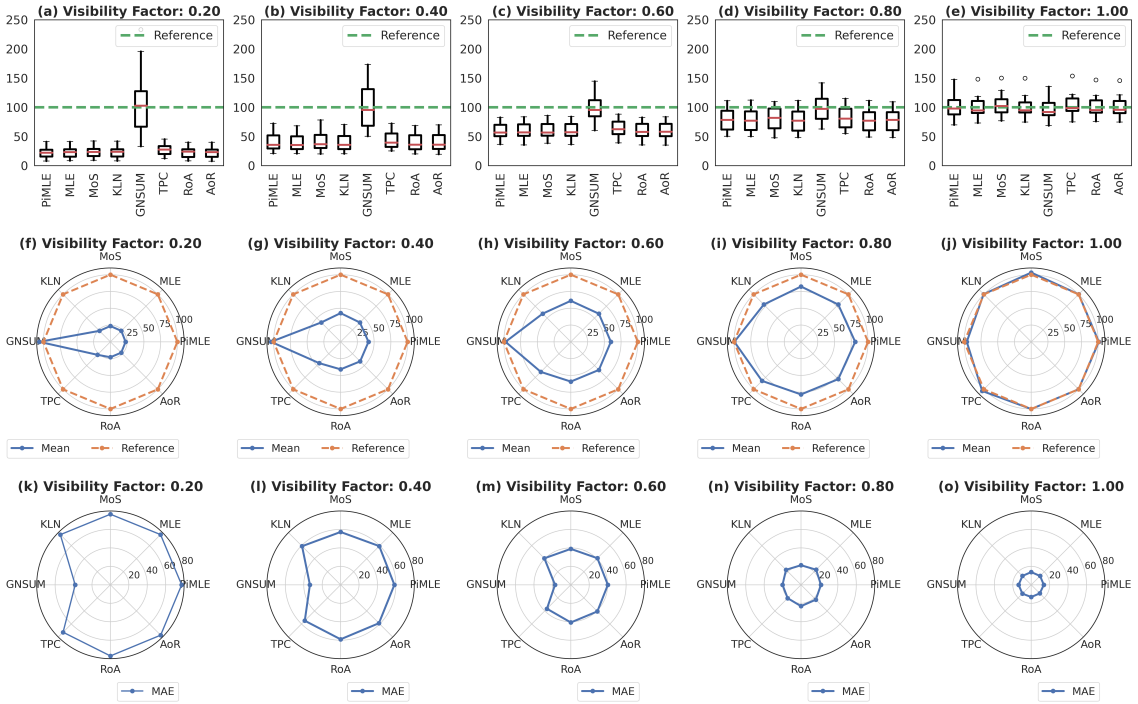


Figure 2.1: Results obtained by the standard methods varying the visibility factor in the W-U scenario. (a)-(e) Boxplots of the unknown subpopulation estimates across methods. (f)-(j) Mean of unknown subpopulation estimates. (k)-(o) MAE of the unknown subpopulation estimates.

0.2. This is observed in Figure 2.1 in the W-U scenario. In addition, we observe that the estimates of the GNUSM are good, although the standard deviation is high. The standard deviation, MAE, and MSE decrease when the visibility factor increases, and therefore, when the transmission error decreases. The Overdispersed estimator performs badly for every value of ν in each scenario. For the rest of the estimators, they have bad performance when the visibility is lower than 1. A special case is the P-S scenario, where the estimates are not good for the mean, MAE, and MSE, even when the visibility factor is 1, and therefore, there is no transmission error.

The adjusted estimators give similar values to the GNSUM, as depicted in Figure 2.2. As we see in Table 2.6, some adjusted estimators are better than the GNSUM in some scenarios. Additionally, we notice that the MAE and MSE decrease as the transmission error decreases. However, in the P-S scenario, the adjusted estimators do not produce reliable estimates.

Recall error simulations

We show the results of the NSUM estimators in the W-U scenario, while the recall error increases in Figure 2.3. This figure contains the boxplot and radar charts for the 9 NSUM estimators. We observe that the AoR is the most affected estimator by this bias. This estimator gives considerably higher estimates than the rest of the

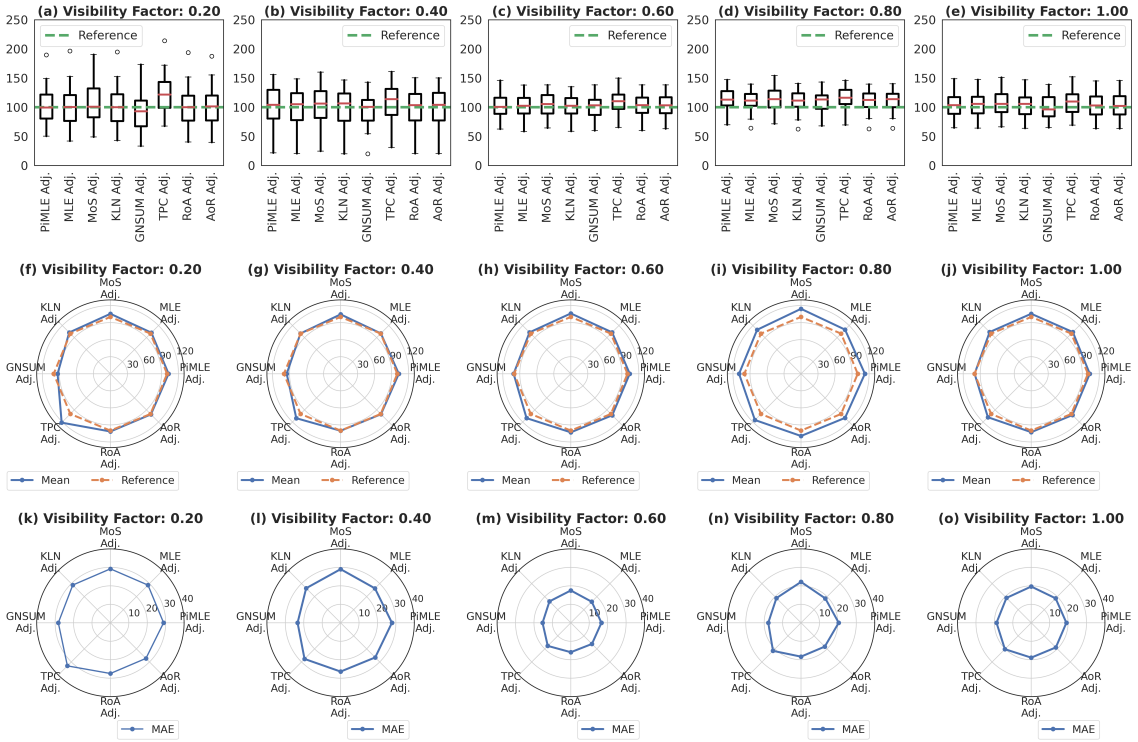


Figure 2.2: Results obtained by the GNSUM and the adjusted methods varying the visibility factor in the W-U scenario. The adjusted methods have been corrected by the inverse of the estimation of the transmission rate using the same sample of the unknown subpopulation as the GNSUM. (a)-(e) Boxplots of the unknown subpopulation estimates across methods. (f)-(j) Mean of unknown subpopulation estimates. (k)-(o) MAE of the unknown subpopulation estimates.

estimates with a low increment of the bias, around 0.3 for the mean (see Table 2.S.10), and significantly higher MAE and MSE for values of the memory factor above 0.4. For all the estimators, the MAE increases as the memory factor increases, as Figure 2.3 displays, but the errors are not as significant as the AoR and Overdispersed. Additionally, PiMLE and MoS have a moderate increase of the mean, MAE, and MSE as the memory factor increases, as we observe in Table 2.S.10 for the values 0.6 and 0.8 of the memory factor. Additionally, as in the previous Section, the Overdispersed estimator gives bad estimates.

The case of the W-S scenario is in Figure 2.4 and Table 2.S.11. We observe that the AoR has a high mean and MAE when the memory factor is greater than 0.5. Unlike the previous case, the other estimators, except the Overdispersed, are similar as the memory factor increases, being the MoS by far the best in terms of MAE and MSE, as we see in Table 2.6.

In the P-U scenario, the results are similar, although the estimators tend to underestimate. The supplementary material shows the results in Table 2.S.12. The AoR is the worst as in the other scenarios; the Overdispersed provides worse metrics for the mean, MAE, and MSE. PiMLE and MoS are comparable to the other

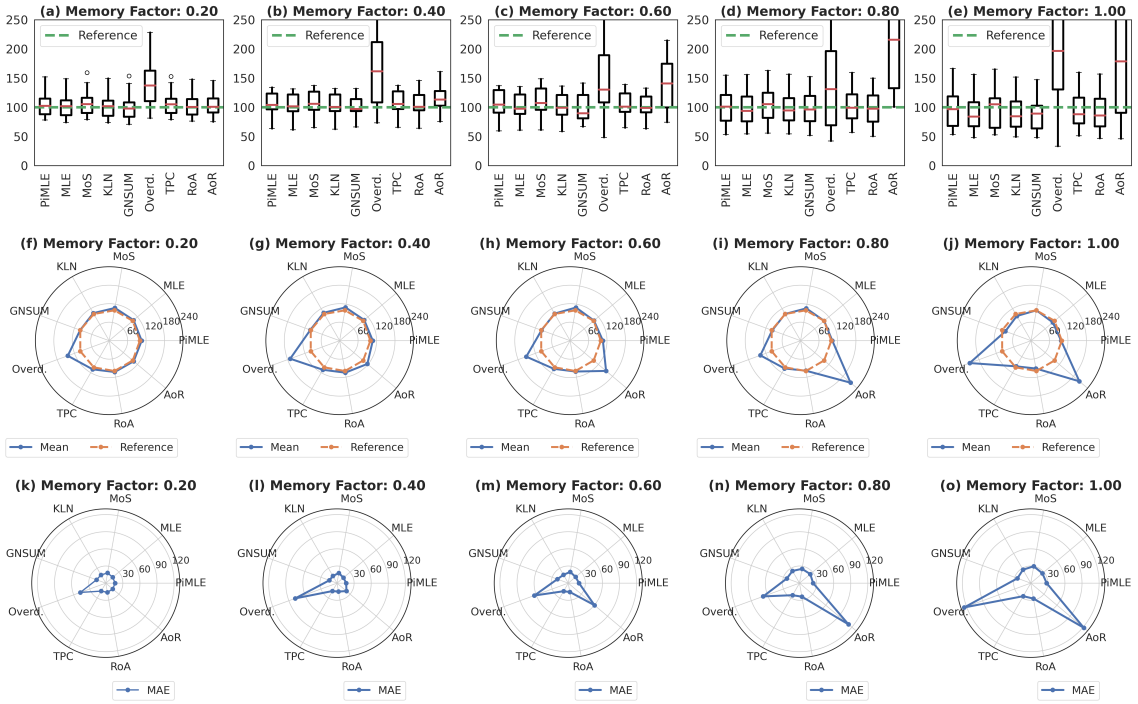


Figure 2.3: Results obtained varying the memory factor in the W-U scenario. (a)-(e) Boxplots of the unknown subpopulation estimates across methods. (f)-(j) Mean of unknown subpopulation estimates. (k)-(o) MAE of the unknown subpopulation estimates.

estimators but are more unstable when the recall error increases. Table 2.S.12 displays that PiMLE and MoS have a high mean for a memory factor equal to 0.8, and high MAE and MSE for a memory factor of 0.9. The GNSUM is slightly worse, but this behavior is constant across all the memory factor values.

In the P-S scenario, all the estimators, except the GNSUM, have a bad performance, as we see in Table 2.S.13. The AoR is again considerably worse when the memory factor increases, while the other estimators do not present significant differences.

Number of subpopulations simulations

We have calculated the NSUM estimates varying the number of known subpopulations used to estimate the degree. To illustrate this, we provide Figure 2.5 and Tables in the supplementary material (see Table 2.4). The figure includes a boxplot and radar charts for the mean and MAE in the W-U scenario with known subpopulations with overlapping. This figure does not include the Overdispersed estimator due to its significantly bad performance.

We do not observe significant differences when the number of subpopulations varies, as we can see in Figure 2.5 and the supplementary material (Tables 2.S.14, 2.S.18, 2.S.22, and 2.S.26). Most of the estimators are very close, except the

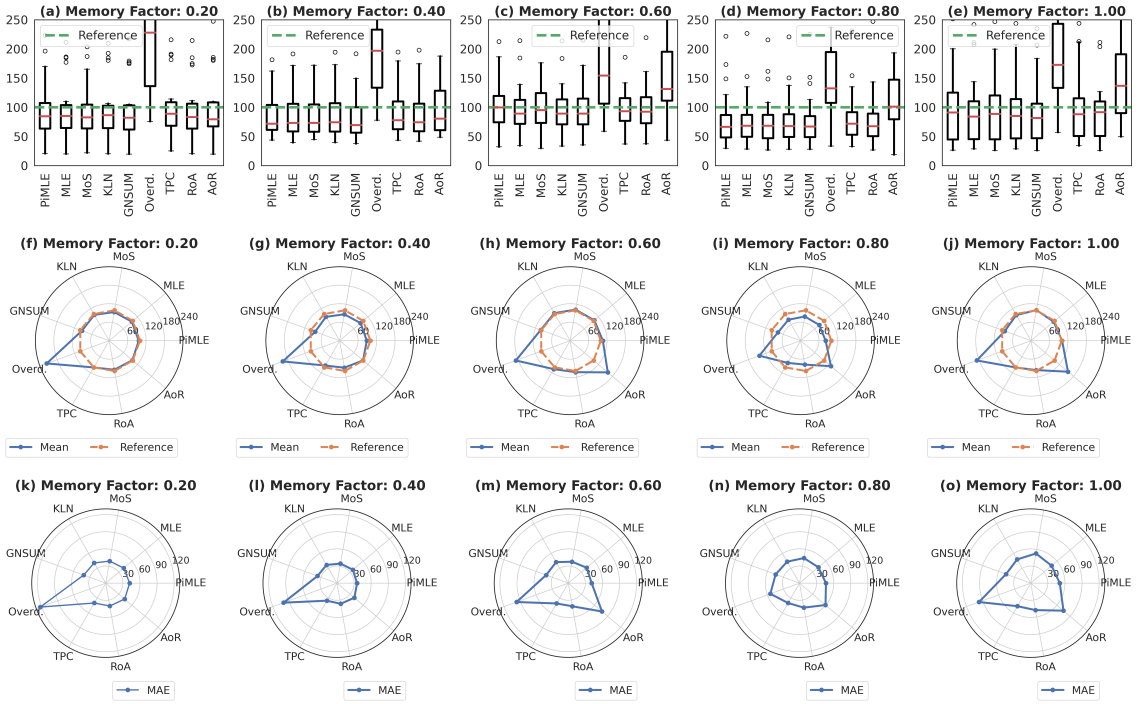


Figure 2.4: Results obtained varying the memory factor in the W-S scenario. (a)-(e) Boxplots of the unknown subpopulation estimates across methods. (f)-(j) Mean of unknown subpopulation estimates. (k)-(o) MAE of the unknown subpopulation estimates.

Overdispersed. The PiMLE, MoS, and TPC slightly differ from the rest of the estimators when the number of known subpopulations is small, as we see for example in the case of two subpopulations in the W-U scenario in Table 2.S.14. Additionally, the values are closer to the RoA and AoR, which do not use subpopulations to estimate the degree, and instead, in these simulations, they use the real personal network size. The variation of the RoA and AoR is then produced by randomness. We notice that the estimators are very similar in terms of all metrics with six or more subpopulations. The MAE and MSE decrease with the number of subpopulations, but the gains are not very significant. In addition, the MLE and KLN are closer to the RoA and AoR with four or more. Regarding the effect of the number of subpopulations when the subpopulations are small, we observe that the results are worse for the standard deviation, MAE, and MSE. The disjoint subpopulations improve the MAE and MSE in the case of medium sizes, but when the subpopulations are small, there is no significant effect.

For the PiMLE and MoS, we observe that the estimators do not give values when the subpopulation sizes are small (see Tables 2.S.16, 2.S.20, 2.S.24, 2.S.28, 2.S.17, 2.S.21, 2.S.25, and 2.S.29). Indeed, the estimators need at least 12 subpopulations to obtain results (see Table 2.S.16). The values of the metric in the few cases where they work are not good, with high mean MAE and MSE. When the subpopulations have bigger sizes, some cases do not give results with two subpopulations (see Tables

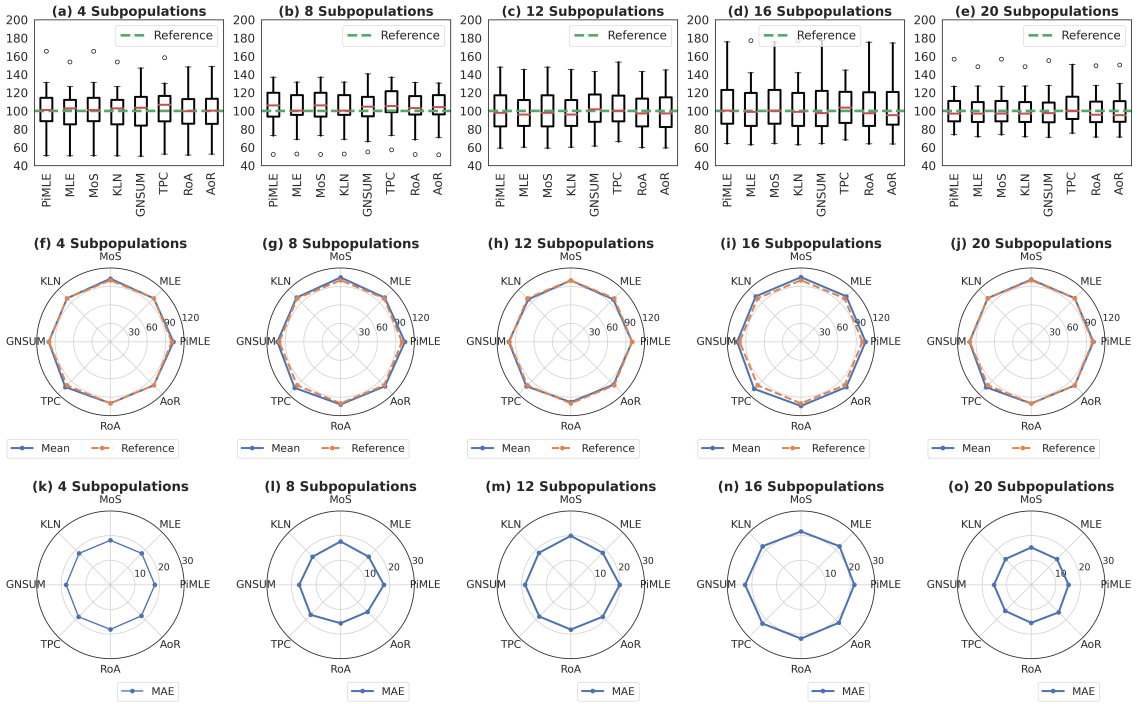


Figure 2.5: Results obtained varying the number of known subpopulations in the W-U scenario. Overlapping of subpopulations is included. (a)-(e) Boxplots of the unknown subpopulation estimates across methods. (f)-(j) Mean of unknown subpopulation estimates. (k)-(o) MAE of the unknown subpopulation estimates.

2.S.22, 2.S.23, 2.S.26, and 2.S.27). This behavior is produced because some estimates for the degree are zero. Additionally, in the case of Watts-Strogatz network and SIR (see Tables 2.S.18 and 2.S.19), with less than eight and four subpopulations (for the non-disjoint and disjoint case, respectively), the PiMLE and MoS have high MAE and MSE, and mean above the real value. For the rest of the main scenarios and parameters, the values are similar to those of the other estimators.

The Overdispersed estimator does not provide good results as in other simulations. In this case, increasing the number of subpopulations increases the bias, MAE, and MSE when the subpopulations are big. This effect does not occur in small subpopulations. Also, the values in this case are better than in other scenarios, although they are worse than the other NSUM estimators. Only the number of subpopulations equal to 6, 8, 12, and 16 gives similar values to the rest of the estimators in the P-U scenario.

Sizes of subpopulations simulations

For the simulations of the sizes of the subpopulation, we use the subpopulations of Table 2.5. This group includes our baseline subpopulations, two even subpopulations, one uneven subpopulation, and other combinations where the size of groups is small, except one population of a large size.

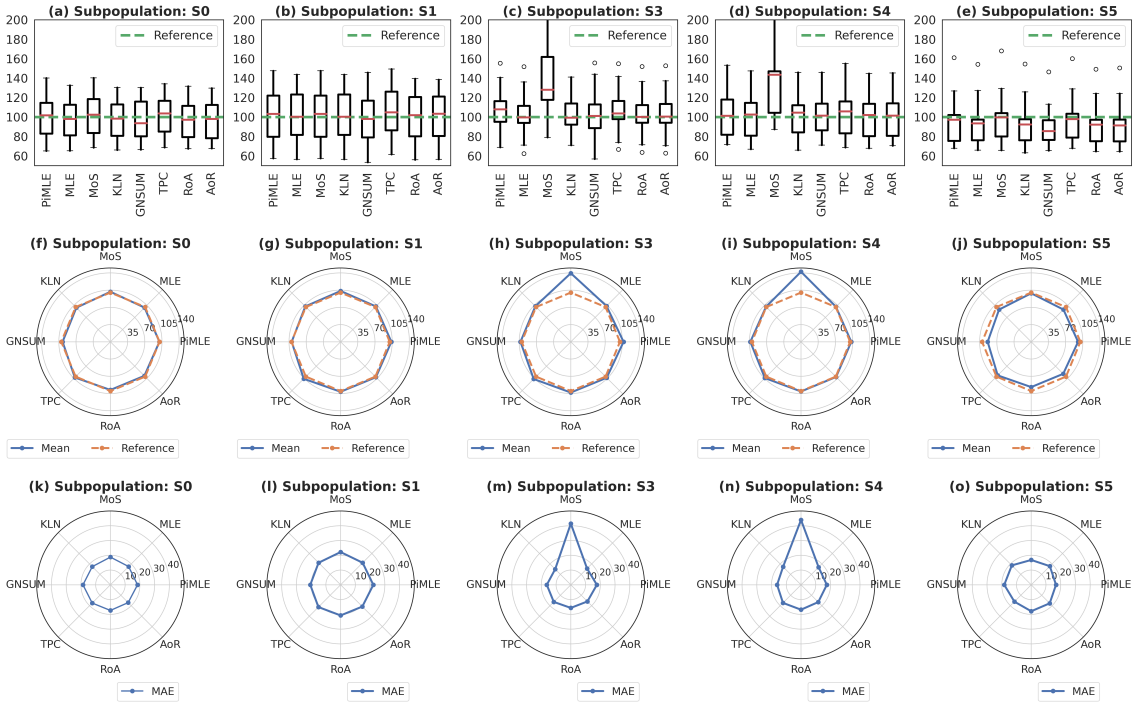


Figure 2.6: Results obtained for different subpopulations in the W-U scenario. (a)-(e) Boxplots of the unknown subpopulation estimates across methods. (f)-(j) Mean of unknown subpopulation estimates. (k)-(o) MAE of the unknown subpopulation estimates.

Subp.	Scenario	Sizes
S_0	Baseline	0.016, 0.02, 0.02, 0.02, 0.02, 0.02, 0.02, 0.02, 0.02, 0.02 0.02, 0.04, 0.04, 0.04, 0.04, 0.04, 0.08, 0.08, 0.08, 0.16
S_1	Even	0.1, 0.1, 0.1, 0.1, 0.1, 0.1, 0.1, 0.1
S_2	Even and small	0.01, 0.01, 0.01, 0.01, 0.01, 0.01, 0.01, 0.01
S_3	One large, the rest small	0.3, 0.01, 0.01, 0.01, 0.01, 0.01, 0.01, 0.01
S_4	An even bigger one, the rest small	0.5, 0.01, 0.01, 0.01, 0.01, 0.01, 0.01, 0.01
S_5	Uneven	0.2, 0.2, 0.1, 0.1, 0.05, 0.05, 0.02,0.02

Table 2.5: Relative subpopulation sizes employed in the simulations.

We show the results of the W-U and P-S scenarios in Figures 2.6 and 2.7, respectively. We excluded the subpopulations S_2 in Table 2.5 since the PiMLE and MoS do not give values, and the Overdispersed estimator due to its bad performance.

The MoS is the most affected by the different sizes, especially when the subpopulations have unequal sizes, and there is a big difference between them, subpopulations S_3 and S_4. We observe this behavior in Figures 2.6 and 2.7, where the estimates differ considerably in those cases. However, the MoS is less affected if there are intermediate sizes between the bigger and the smaller, as in S_5. As we can observe in Figures 2.6 and 2.7.

For the rest of the estimators, the values are close to each other, except in the case of a preferential attachment network and SIR, where only the GNUSM gets reliable estimates for all of the subpopulation configurations, as Figure 2.7 shows.

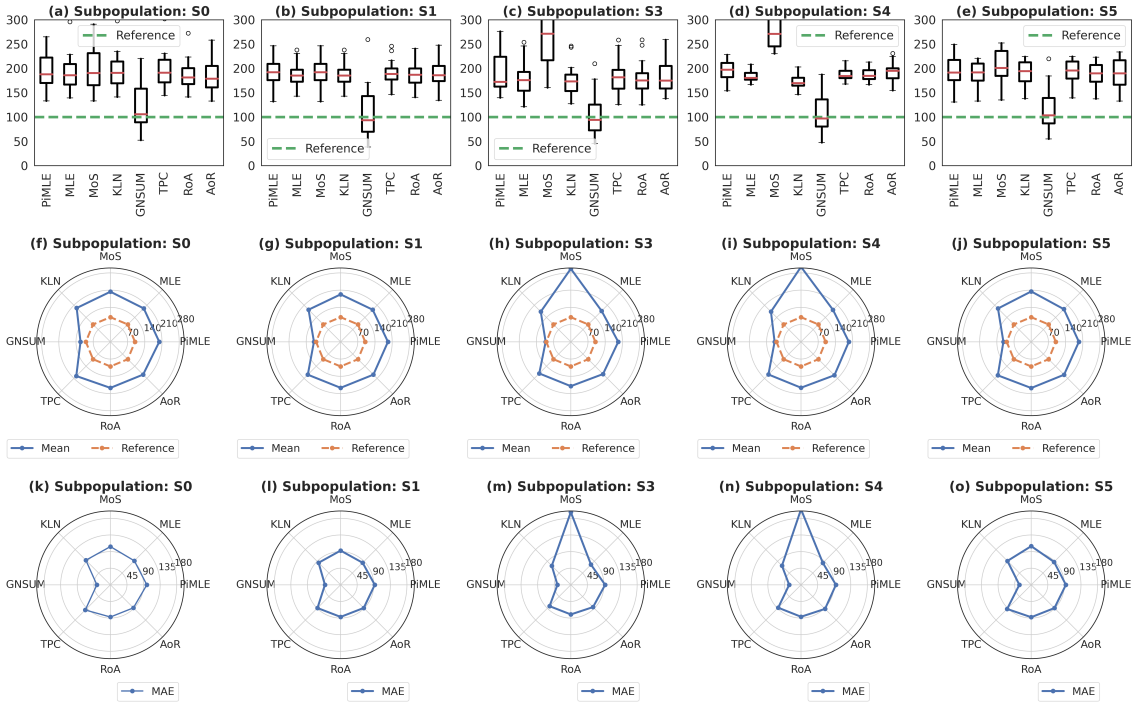


Figure 2.7: Results obtained for different subpopulations in the P-S scenario. (a)-(e) Boxplots of the unknown subpopulation estimates across methods. (f)-(j) Mean of unknown subpopulation estimates. (k)-(o) MAE of the unknown subpopulation estimates.

In addition, the values are similar to the estimators RoA and AoR, which use the real degree in this context. However, the PiMLE, TPC, and GNSUM are slightly affected by the known subpopulation configurations. PiMLE is affected by the unequal sizes, represented by subpopulations S_3 and S_5, but it is not affected by even subpopulations with one big subpopulation (S_4), unlike the MoS. Additionally, PIMLE presents issues when the known subpopulations are small, represented in case S_2 because of the zero-degree estimates. The TPC is not affected in the P-U scenario, where it is the best method (see Table 2.6) and is less affected than the PiMLE and GNSUM in the W-S scenario. The MLE and KLN show to be very robust, varying the sizes of the known subpopulations.

Sample size simulations

We evaluate the performance of the NSUM estimators as the sample size varies. We provide two figures with the boxplots and radar charts in the scenarios W-U and P-U. We include detailed tables with results in all the scenarios in the supplementary material.

We note in Figures 2.8 and 2.9 that the estimators reduce the error when the sample size increases. In the W-U scenario, all the estimators converge to the real value, as Figure 2.8 shows. The same behavior is observed when the unknown

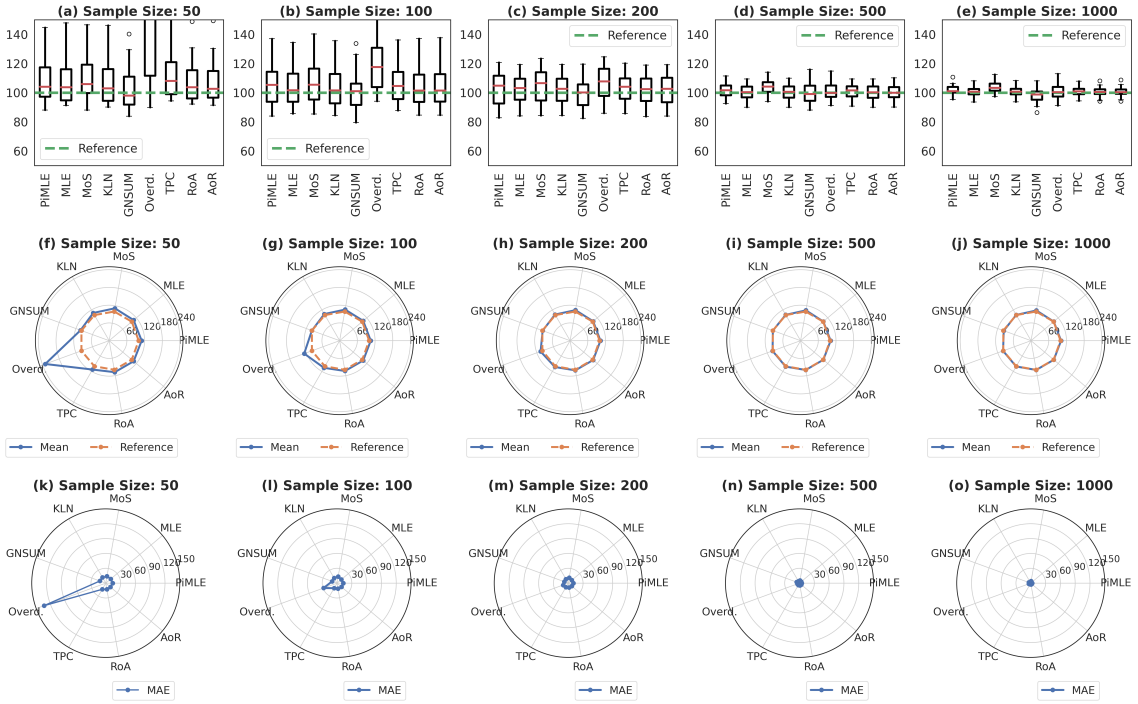


Figure 2.8: Results obtained varying the sample size in the W-U scenario. (a)-(e) Boxplots of the unknown subpopulation estimates across methods. (f)-(j) Mean of unknown subpopulation estimates. (k)-(o) MAE of the unknown subpopulation estimates.

subpopulation is replaced in the W-S scenario in Table 2.S.38. In the P-U scenario, Figure 2.9 shows that increasing the sample size improves the estimates, although PiMLE and MoS are the only estimators that converge to the real value, as noted in Table 2.S.40. The GNSUM slightly overestimates, while MLE, KLN, TPC, RoA, and AoR slightly underestimate. Additionally, Figure 2.9 and Table 2.S.40 indicate that the GNUSM has high variance and MAE and MSE higher than the other methods. In the last case, the P-S scenario, the sample size does not lead to better results, with GNSUM being the only estimator providing a reliable estimate, although with a slightly higher mean.

The MAE and MSE are better as the sample size increases. The MAE is below ten percent for a sample of $200-300$ for the W-U scenario; 1000 for the W-S scenario; and 200 for the P-U scenario. In the P-S scenario, the errors are always bigger than ten percent.

In addition, we observe that the Overdispersed estimator needs more samples than the other NSUM estimators to obtain reliable estimates for the mean, and error similar to the other estimators, as is seen in Figures 2.8 and 2.9. It needs a sample between 100 in the W-U scenario, 200 in the P-U scenario, and 500 in the W-S scenario. This does not apply to the special case of the P-S scenario, where the estimates are far from the real value for every value of the sample size, like the other estimators (except the GNSUM). In this case, the Overdispersed estimator is similar

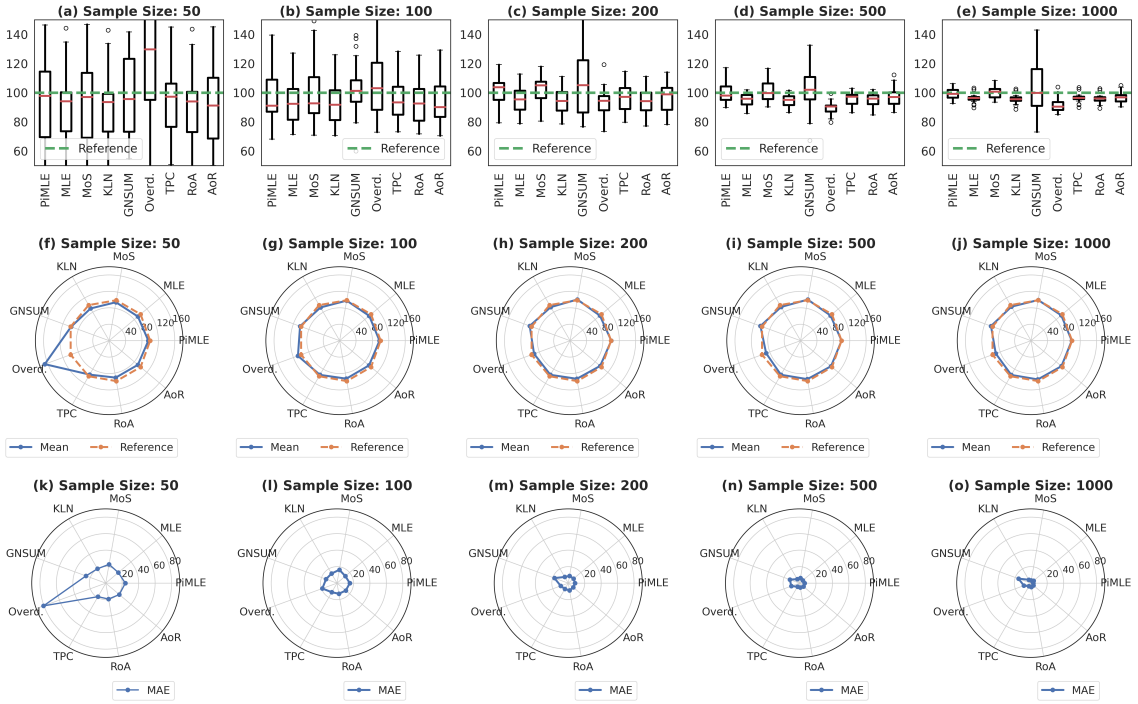


Figure 2.9: Results obtained varying the sample size in the P-U scenario. (a)-(e) Boxplots of the unknown subpopulation estimates across methods. (f)-(j) Mean of unknown subpopulation estimates. (k)-(o) MAE of the unknown subpopulation estimates.

to the other estimators for a sample of 50.

2.5.2. Discussion

The simulation shows that the network topology, the unknown subpopulation distribution, and the different parameters of the simulation affect the results. Table 2.6 displays the best method in each simulation and scenario. This table does not include the estimators RoA and AoR in the simulations varying the number and sizes of the known subpopulations because they do not use the known subpopulations. We also include Tables 2.S.45 and 2.S.46 in the supplementary material, which provide the values of the best method for the subpopulation number and subpopulation sizes, both with and without including RoA and AoR. Moreover, the rows of the table corresponding to the number of subpopulations only include the case of subpopulations of medium sizes with overlapping. Additionally, we include in Figure 2.10 how the expectation of the ratios of the ARD of the unknown subpopulation by the real degree differs from Eq. (2.1).

Effect of the network topology and unknown subpopulation distribution. In the Watts-Strogatz case with uniform distribution of the unknown subpopulation, the W-U scenario, the MLE is usually the most robust, as we observe in Table 2.6.

Simulation goal	Main scenarios	Mean	MAE	MSE
Trans. error	W-U	GNSUM/100.64	GNSUM/24.78	GNSUM/1183.1
	W-S	MLE(adj)/98.27	MoS(adj)/44.08	MoS(adj)/3198.95
	P-U	PiMLE(adj)/100.74	KLN(adj)/24.98	KLN(adj)/1072.24
	P-S	GNSUM/112.20	GNSUM/39.89	GNSUM/3214.67
Recall error	W-U	MLE/100.35	MLE/18.38	MLE/546.82
	W-S	TPC/96.14	MoS/38.8	MoS/2337.73
	P-U	PiMLE/99.65	TPC/18.95	TPC/564.44
	P-S	GNSUM/118.82	GNSUM/38.55	GNSUM/2884.27
Subpop. number	W-U	MLE-KLN/100.00	MLE-KLN/18.36	MLE-KLN/530.82
	W-S	MLE-KLN/105.38	TPC/42.39	MLE-KLN/3036.94
	P-U	TPC/99.93	TPC/15.71	TPC/382.1
	P-S	GNSUM/115.58	GNSUM/ 37.52	GNSUM/ 2976.36
Subpop. sizes	W-U	MLE-GNSUM/99.80	MLE/18.34	KLN/501.14
	W-S	KLN/105.02	KLN/42.87	KLN/2768.76
	P-U	TPC/99.53	TPC/17.31	TPC/439.21
	P-S	GNSUM/111.43	GNSUM/ 36.52	GNSUM/ 2531.72
Sample size	W-U	GNSUM/99.06	RoA/8.29	GNSUM/181.61
	W-S	TPC/98.12	MoS/20.04	MoS/797.07
	P-U	MoS/100.32	TPC/9.36	TPC/183.81
	P-S	GNSUM/110.28	GNSUM/ 32.07	GNSUM/ 1959.8

Table 2.6: Best method for each scenario. For each metric, it shows the average of the values corresponding to each parameter. AoR and RoA are not included in the simulations about the number and sizes of known subpopulations because they do not use the known subpopulations.

The Watts-Strogatz model is related to the Erdős–Rényi model, as it becomes the Erdős–Rényi model in the limit case when the probability of rewiring each edge is 1. The binomial model, in which MLE, KLN, and RoA are based, approximates the Erdős–Rényi model (Kunke et al., 2024). These models are especially good in this scenario.

In the W-S scenario, the estimates are worse due to the barrier effect since the probability of knowing someone in the unknown subpopulation depends on the position in the network. In this case, the unknown distribution is not distributed uniformly, and hence, it does not hold the conditions of the methods involving the binomial model for the size estimation. Despite this fact, these estimators, MLE, KLN, and RoA, have a good performance compared to other methods. However, the main assumption of Eq 2.1 holds in this scenario, as seen in Figure 2.10, and the methods relying only on this equation, MoS, and PiMLE, have better performance.

In the P-U scenario, the estimates are also worse than in the W-U scenario. Some higher-degree nodes are counted multiple times, slightly affecting the estimates in terms of MAE and leading to some underestimation. However, PiMLE and MoS use the ratios between the ARD responses for the unknown subpopulation and the degree estimated with the ARD of the subpopulation, which reduces the influence of higher-connected nodes. The TPC model assumes that $y_{ij} = \lambda N_k$, and introduces an additional parameter α_i to account for the variability of the network. This seems to be particularly effective in this scenario. It performs best, especially in terms of MAE and MSE, as shown in Table 2.6.

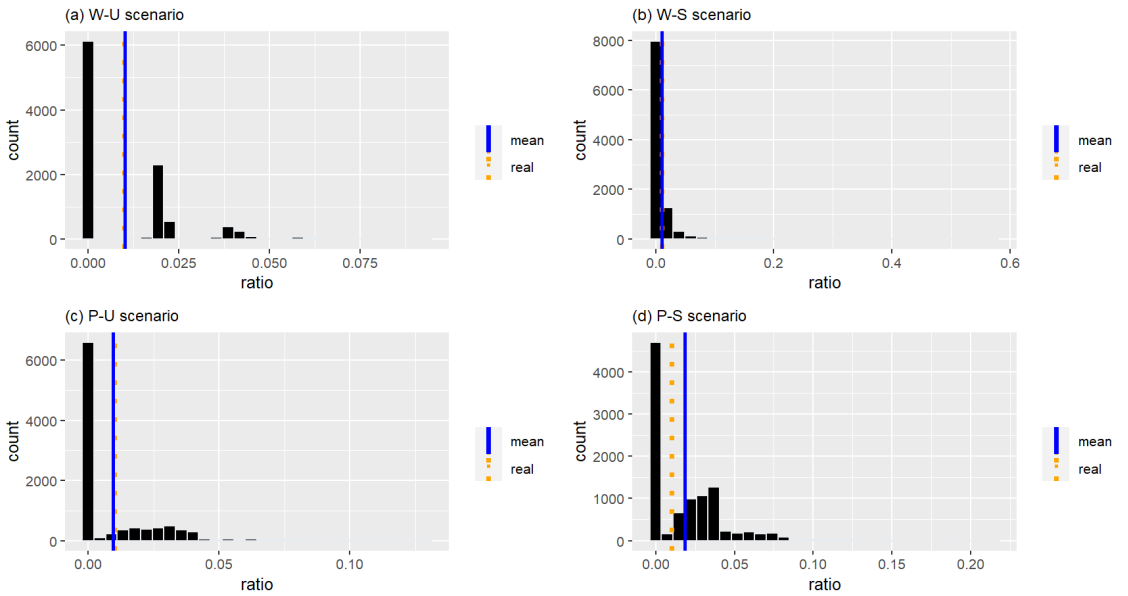


Figure 2.10: Histogram of the ratio of responses y_{iu}/d_i for the four scenarios.

In the P-S scenario, only the GNSUM produces reliable estimates. The SIR process produces that the high-degree nodes are more likely to belong to the unknown subpopulation. This means the higher degree nodes belonging to the unknown subpopulation nodes are counted multiple times. It produces higher ARD responses for the unknown subpopulations, leading to overestimation. The GNSUM corrects this issue with additional information on the links of the unknown subpopulation through the sample of the unknown subpopulation.

The GNSUM estimator does not seem to depend on the network structure and performs well in every scenario. However, it requires significantly more information, including a sample of the unknown subpopulation. The Overdispersed estimator gives a bad performance in the simulations for every scenario. The overdispersion parameters for the propensity of subpopulation links may cause some overfitting, as in our case, the propensity of the known someone in the subpopulations is similar for each group.

Effect of the transmission error. The GNSUM estimator is the best when the transmission error is present, which is expected since it utilizes significantly more information. The GNSUM corrects bias using how visible the members of the unknown subpopulation are. For this, the GNSUM needs a sample of the unknown subpopulations and a battery of questions about the contacts of the sample, which diverges from the philosophy of preserving the privacy of the participants of the NSUM. The rest of the estimators underestimate when the transmission error is present. The values are proportional to the transmission error because the expected value of y_{iu} is $\nu d_i N_u/N$ (Eq 2.9) when the transmission error is the same across all individuals in the network, as we observed in Section 2.4.2. Consequently, the

estimators adjusted by the inverse of the estimated ν are close to the ground truth in most scenarios. The correction proves ineffective in the P-S scenario, unlike the GNSUM, which makes sense since the estimators without transmission error are inadequate. Feehan and Salganik (2016) prove that the GNSUM is equivalent to the correction by some factors of the MLE and RoA, provided that the estimated average degree in the denominators is accurate. However, this adjustment includes another factor, the degree ratio, the fraction of links of the unknown subpopulation relative to the number of links in the total population.

Effect of the recall error. The AoR is the most affected. Since the recall error applies both to the ARD of y_{iu} and the reported degree δ_i, the ratios y_{iu}/δ_i may be altered substantially if the recall error is high. PiMLE and MoS, which also use the average of the ratios to estimate the size (Eq. (2.4)), are affected to a lesser extent, being closer to the real value. Unlike the AoR, PiMLE and MoS use estimated degrees, which are obtained through linear combinations of the ARD responses of the known populations y_{ik}. As a result, the errors tend to cancel out, making the estimated degrees more accurate and improving the size estimation. MLE, KLN, and RoA use ratios of averages (see Table 2.1), and the errors of the two averages tend to cancel out. The GNUSM has a similar performance to MLE, KLN, and RoA, since it is also expressed as the ratio of a linear combination of the ARD in the numerator, and a linear combination of the EARD in the denominator (see Eq. (2.8) and the supplementary material).

Effect of the number of known subpopulations The main conclusion of the variation of the number of known subpopulations is that the common number of subpopulations, 20, can be reduced. The estimators MLE and KLN are particularly stable. The randomness of choosing a person with contacts in a known subpopulation is compensated in the sample because these estimators only need to estimate the average degree (see Eq. (2.5) and Table 2.1). The GNSUM is the same, replacing the estimated degree with the estimated relative visibility (see Section 2.3.3 and supplementary material), which is obtained by a linear combination of the EARD questions of the known subpopulations. MLE and KLN are very close with more than 4 subpopulations in every scenario. GNSUM and TPC are also stable; they need around 6 subpopulations to stabilize the MAE and have a very tight estimation, as we see in the supplementary material. When the subpopulations have small sizes, PiMLE and MoS present issues since they use the average of the ratios of the ARD response by the estimated degree. When the subpopulations have small sizes, the sample nodes may not be connected with any member of the subpopulations, and then, the estimated personal sizes are 0. Additionally, PiMLE and Mos only have good performance with small sizes in the W-U scenario with 20 subpopulations. Finally, the presence of disjoint known subpopulations improves the estimates, although it is

not very significant.

Effect of the sizes of the known subpopulations The MoS is affected if the sizes of the known subpopulations are small since some degree estimates are zero (see Eq. (2.2) and 2.4). In the case of a combination of small subpopulations with a bigger one, MoS gives bad estimates. MoS estimates the participant's degree by a linear combination of y_{ik} for the L subpopulations, where the small subpopulations are weighted more (Eq.(2.2)). Since it is less likely that a participant is linked with someone in the subpopulations, some personal sizes are underestimates, which causes the final overestimations in the subpopulations S_3, S_4 of Table 2.5. Subpopulations of intermediate sizes may compensate for this issue, as is the case with subpopulation S_5. PiMLE is also critically affected when the subpopulations are all small for the same reason as the MoS. The cases of small subpopulations with one big subpopulation are considerably better, although they affect the estimates. This is because the degree of each participant is estimated by Eq. (2.3), which is the ratio of $N\sum_{k=1}^{L} y_{ik}$ by $\sum_{k=1}^{L} N_k$, and therefore, the effect of the small subpopulation is more limited. The MLE and KLN use the estimated average degree at which the errors cancel out. The same occurs for the GNSUM because it uses the relative visibility of the unknown subpopulation. The TPC is affected, although the estimates are closer to the real value. The TPC estimator assumes that the scaling is the same, with $y_{ij} = \lambda N_j$. Therefore, it is affected by unequal populations.

Effect of the sample size. We observe that increasing the sample size improves the estimation, with the estimators yielding good results, especially in terms of MAE and MSE errors. This is particularly noticeable in the Overdispersed case, which did not yield good results with small samples. The overdispersion parameters for the propensity to form links with the subpopulations may cause some overfitting, as the propensity of the known someone in the subpopulations is similar for each group. Increasing the sample can reduce this overfitting.

Additionally, we observe that the estimators require more samples in the scenarios with more barrier error, W-S and P-U scenarios, than in the W-U scenario. The values of the mean become closer when the sample size increases. However, while all the estimators converge in the W-U scenario, in the W-S and P-U scenarios, most methods slightly underestimate. We conjecture that this is because some nodes are counted multiple times, and most of them do not belong to the unknown subpopulation. In the P-U scenario, the nodes with more links are considered multiple times, and most of them do not belong to the unknown subpopulation. There are three special estimators in this case: PiMLE, MoS, and GNSUM. PiMLE and MoS use the ratios between the ARD of the unknown subpopulation and the estimated degree. The estimated degrees use the ARD of the known subpopulation, so the higher connected nodes are less taken into account, see Eq. (2.2) and (2.3).

We hypothesize that GNSUM is overestimated because the relative visibility is underestimated, as the unknown subpopulation nodes are mainly in the tail of the degree distribution.

2.6. Case Study

This section evaluates the performance of the NSUM estimators in a real network. Specifically, we use a friendship network extracted from Deezer, a popular music streaming platform, focusing on Croatian users. This dataset contains mutual friendship connections and users' preferred music genres. This dataset was collected in November 2017 by Rozemberczki et al. (2019) as a part of their study on graph embedding techniques, and it is publicly available through the Stanford Large Network Dataset Collection: https://snap.stanford.edu/data/gemsec-Deezer.html. This dataset consists of 85 musical preferences for each user, with 54573 nodes and an average degree of 18.26. We take the 21 musical preferences with more presence in the network. The genres and their relative size are in Table 2.7. We also use synthetic subpopulations with the same sizes generated with the procedures of Section 2.4.

Musical Genres		
Asian Music (0.91)	Country (4.89)	Folk (18.57)
Blues (5.37)	Dancefloor (9.27)	Hard Rock (9.12)
Chill Out (1.57)	Dancehall Ragga (1.29)	Indie Rock (5.95)
Classical (3.19)	Disco (5.55)	Jazz (4.91)
Comedy (2.22)	East Coast (2.92)	Kids (3.85)
Contemporary RB (13.21)	Film Scores (16.44)	Latin Music (12.69)
Contemporary Soul (2.50)	Films Games (17.36)	Metal (9.67)

Table 2.7: Popular musical genres and the percentage of users for the Deezer dataset.

This section is organized as follows. First, we evaluate the NSUM estimators by simulating subpopulations uniformly across the real network. Next, we present a scenario where an unknown subpopulation is distributed through an SIR process over the network. Subsequently, we estimate some musical preferences of the Deezer dataset using other musical preferences as known subpopulations. Finally, we discuss the results for all the cases.

2.6.1. Real network with subpopulations distributed uniformly

We use the Deezer dataset as the underlying network and simulate the subpopulations uniformly using the sizes of Table 2.7. We use synthetic subpopulations with sizes corresponding to Asian Music, Blues, and Chillout as the unknown subpopulations. We evaluate the results of 50 estimates, taking samples of 100 nodes and other

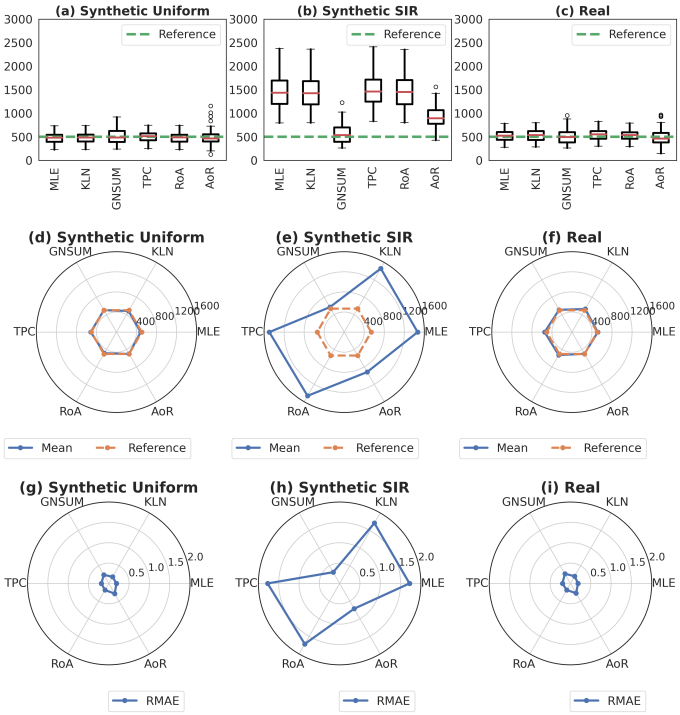


Figure 2.11: Results obtained from a real network with a real unknown subpopulation (Real), a synthetic unknown subpopulation uniformly distributed (Synthetic Uniform), and a synthetic unknown subpopulation using a SIR model (Synthetic SIR). (a)-(c) Boxplots of the estimates. (d)-(f) Mean of estimates. (g)-(i) RMAE estimates.

samples of 30 nodes for the unknown subpopulation. We provide a boxplot, mean, and relative mean absolute error (RMAE) of the estimates using Asian Music as the unknown subpopulation in Figure 2.11. In Table 2.S.42 of the supplementary material, we include the values for the next metrics: mean, standard deviation, RMAE, and mean squared relative error (MSRE).

We observe in Figure 2.11 that MLE, KLN, GNSUM, TPC, AoR, and RoA are close to the real values. However, we observe an underestimation in these estimators except GNSUM and AoR. Additionally, we observe more variance for the GNSUM and AoR. PiMLE and MoS do not give values because some individuals do not have contacts in any subpopulation, producing degree estimates with zero values. The Overdispersed has poor performance in the cases of Asian Music and Chillout, but its estimates are good in the case of Blues, as seen in Table 2.S.42.

2.6.2. Real network with SIR for the unknown subpopulation

We generate the known subpopulations as in the previous case, but we also simulate an unknown subpopulation using an SIR process. The synthetic subpopulations have sizes relatively similar to the real data. We evaluate the performance of the estimators for 50 iterations. In each iteration, we use samples of 100 from the entire network and 30 from the subset of nodes in the unknown subpopulation. We include

a boxplot and two radar charts for the mean and RMAE in Figure 2.11 for the case of synthetic Asian Music. Values of the mean, standard deviation, RMAE, and MSRE are included in Table 2.S.43.

In this case, the GNSUM is the only estimator that provides reliable estimates, as observed in Figure 2.11. The other estimators present a very significant overestimation. AoR is less affected but has a high error and is far from the real value. Since the subpopulations have small sizes, PiMLE and MoS do not produce results, since they estimate some degrees as zero. The other two cases are similar, as Table 2.S.43 shows.

2.6.3. Real network with real subpopulations

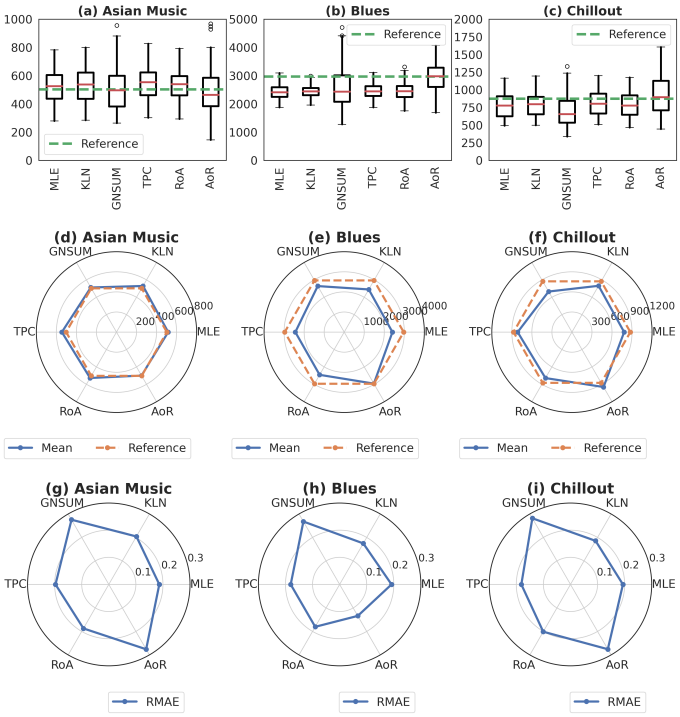


Figure 2.12: Results obtained from a real network with real subpopulations. (a)-(c) Boxplots of the unknown subpopulation estimates. (d)-(f) Mean of unknown subpopulation estimates. (g)-(i) RMAE of the unknown subpopulation estimates.

We use the 21 most popular Deezer music genres, displayed in Table 2.7. We select one of the musical genres as the unknown subpopulation and the rest as the known subpopulations. We apply this three times, using as unknown subpopulations the musical genres Asian Music, Blues, and Chill Out. We provide the results for 50 estimates of the subpopulations using samples of 30 from the unknown subpopulation and 100 from the entire network. We include in Figure 2.12 the boxplot for the three cases and radar charts for the mean and RMAE. In the supplementary material, we provide the mean, standard deviation, RMAE, and MSE values in Table 2.S.44.

We observe in Figure 2.12 that the distribution of the subpopulations affects

the estimators. While the estimates are close in the case of Asian Music, MLE, KLN, RoA, and TPC are affected in terms of mean estimating Blues. Additionally, GNSUM underestimates the Chillout subpopulations. We note that AoR is more stable in terms of the mean, but the RMAE is high. The other estimates are stable for this metric, although the GNUSM also has a high error. PiMLE and MoS present issues since some degrees estimates are zero. The Overdispersed estimator provides reliable estimates in the case of Blues, but significantly overestimates in the other two cases, as seen in Table 2.S.44.

2.6.4. Discussion

We observe that the results of the real network and synthetic subpopulations validate our conclusions of Section 2.5.2. The real network is similar to the preferential attachment network since its degree distribution is exponential. The case of subpopulations simulated with subpopulations distributed uniformly, and the case of SIR for the unknown subpopulation, are similar to the P-U and P-S scenarios, and the performance of the estimators is comparable with these scenarios. One difference with our simulations is the bad performance of the PiMLE and MoS due to the small relative size of the subpopulations, which is consistent with our analysis of the sizes of the known subpopulations. The other difference is that the AoR is slightly better. We guess that one factor is the lower average degree, decreasing the frequency with which multiple nodes are reported.

The real network with real subpopulations of the Deezer dataset shows that the distribution of the subpopulations affects the performance. Nodes with similar preferences tend to be connected with each other, increasing the barrier effect. The difference with the P-S scenario is that concentration is not produced in the higher-degree nodes for the selected real subpopulations. This is similar to the P-U scenario, underestimating whether the subpopulations are connected to the higher-degree nodes. For the preference Blues, the degree of the individuals in the subpopulation is lower than the average degree. This condition affects the NSUM as Feehan and Salganik (2016) show. However, this does not occur for the preference Chillout, in which there are deviations from the mean. In the P-U scenario, PiMLE and MoS are the best because they use the average of the ratios, Eq. (2.4), using the estimated degrees through subpopulations concentrated mainly in the tail of the distribution. In the real data, PiMLE and MoS do not give results because the subpopulations are small. The AoR, which uses Eq. (2.4) with the real degrees, is better than the other methods compared to the P-U scenario. We conjecture that it is produced because the network is bigger and the average degree is lower, reducing the impact of the higher degree nodes in the ratios. In addition, AoR and GNSUM are the most reliable in the real case, although the variance, RMAE, and MSRE are high in some cases (see Figure 2.12 and Table 2.S.44). For the AoR, the behavior is caused by

the variation in the proportion of contacts in some groups, and the effect on the GNSUM is likely to be caused by the small sample.

2.7. Conclusion

This work compares via simulation most of the NSUM estimators in networks with 10000 nodes and some properties found in real networks: a small world network (a Watts-Strogatz estimator) and a scale-free network obtained by a linear preferential attachment process. In each of them, we conduct two examples of the unknown subpopulation. In the first case, the unknown subpopulation is obtained uniformly and independently of the network. The second case is an unknown subpopulation distributed by an SIR process through the network, which is an example of the presence of the barrier effect. In these cases, the paper studies how the NSUM estimators vary as two biases (transmission and recall error) increase, how the number and sizes of the subpopulation affect the estimates, and the effect of the sample size. Additionally, we evaluate the difference in the performance of the estimators with real data. Concretely, we use a real network with real subpopulations and synthetic subpopulations simulated with the same procedures as the simulations.

The main findings comparing the NSUM estimators across the scenarios are as follows: (1) the number of subpopulations does not affect the estimators, except for the Overdispersed, PiMLE, and MoS; and for the last two, only six can be employed if the subpopulations are large, (2) the network topology has an impact on the results, the best method depends on the network, and in the preferential attachment networks, some estimators are slightly biased or give unreliable results, (3) dependent responses obtained by reporting the same node multiple times when the sample increases do not have an impact on the estimates, (4) in the presence of the recall error, the AoR is very unstable, and the PiMLE and MoS are slightly unstable in some cases when the recall error is high, (5) the Overdispersed estimator needs more sample to obtain more reliable estimates than the other estimators, and the errors increase with more subpopulations, (6) the adjustment by an estimation of the transmission error is similar than using the GNSUM in some cases, and some corrected estimators are better; but in the the case of a SIR process and preferential attachment the network, the results, with a certain transmission rate estimator, are considerably worse respect to the GNSUM, indicating the necessity for additional adjustment.

This work has some limitations. The first limitation is that the analysis is based on a simulated environment, and therefore, there may be some differences in real cases. More concretely, the known subpopulations are assumed to be independent and uniformly distributed along the network, when in most real cases this does not happen. Also, we assume that the underlying network is an undirected graph, which means that the links between the individuals are reciprocal. Another limitation is

that we use large-dimensional matrices to calculate the reported values after the transmission and barrier error. This leads to high computational cost. For that reason, we use a few examples of networks and unknown subpopulation distribution, and one instance for each parameter in every simulation.

The NSUM methods have not been tested on many datasets. This is the first study that compares many of the estimators for the first time. Additionally, due to the computational cost, few examples were added. The first further work is to compare the estimators in more examples of network topologies, the distribution of the subpopulations, and the unknown subpopulation. Since the epidemiological SIR processes usually do not occur in the contact network, further work is needed to analyze the case when the unknown subpopulation is propagated through another network. Another future work is the study of the effect of the demographic characteristics to be able to compare new estimators, such as McCormick and Zheng (2012), that take these variables into account. Because the sample size depends on the network topology and the distribution of the unknown subpopulation, the study of the effective sample size needs further development. Another future work is the correction of the NSUM estimators under other biases, since we observe that the transmission error correction is insufficient when the barrier effect is present. It is also worth analyzing another bias arising from malicious responses, which is known as response bias, and analyzing the performance of the estimators under different types of contamination. Another subsequent line of research is the analysis of the effect of the connection between the subpopulations produced in real cases. Finally, some estimators that are not frequently employed are better in some scenarios; further research may help to select the best method.

Code and Data Availability Statement

The data and code are publicly available at https://github.com/SDiazAranda/NSUM-Simulations

Declaration of interest statement

We have no conflicts of interest to disclose.

SUPPLEMENTARY MATERIAL

Supplementary Material to "Performance Analysis of NSUM Estimators in Social-Network Topologies": This supplementary material contains more exhaustive details on the estimators and network topologies. Moreover, it includes the tables with all the simulations and metrics.

References

L. S. Bastos, N. S. Paiva, F. I. Bastos, and D. A. Villela. Fast approaches for bayesian estimation of size of hard-to-reach populations using network scale-up. arXiv preprint arXiv:1804.04678, 2018.

H. R. Bernard, E. C. Johnsen, P. D. Killworth, and S. Robinson. Estimating the size of an average personal network and of an event subpopulation. In The Small World, pages 159–175, 1989.

S. Díaz-Aranda, J. M. Ramírez, M. Daga, J. P. Champati, J. Aguilar, R. E. Lillo, and A. F. Anta. Error bounds for the network scale-up method. arXiv preprint arXiv:2407.10640, 2024.

D. Easley, J. Kleinberg, et al. Networks, crowds, and markets. Cambridge Books, 2012.

S. Ezoe, T. Morooka, T. Noda, M. L. Sabin, and S. Koike. Population size estimation of men who have sex with men through the network scale-up method in japan. PloS one, 7(1):e31184, 2012.

D. M. Feehan and M. J. Salganik. Generalizing the network scale-up method: a new estimator for the size of hidden populations. Sociological methodology, 46(1): 153–186, 2016.

A. Garcia-Agundez, O. Ojo, H. A. Hernández-Roig, C. Baquero, D. Frey, C. Georgiou, M. Goessens, R. E. Lillo, R. Menezes, N. Nicolaou, et al. Estimating the covid-19 prevalence in spain with indirect reporting via open surveys. Frontiers in Public Health, 9:658544, 2021.

P. Habecker, K. Dombrowski, and B. Khan. Improving the network scale-up estimator: Incorporating means of sums, recursive back estimation, and sampling weights. PloS one, 10(12):e0143406, 2015.

L. Jing, Q. Lu, Y. Cui, H. Yu, and T. Wang. Combining the randomized response technique and the network scale-up method to estimate the female sex worker population size: an exploratory study. Public Health, 160:81–86, 2018.

P. D. Killworth, E. C. Johnsen, C. McCarty, G. A. Shelley, and H. R. Bernard. A social network approach to estimating seroprevalence in the united states. Social networks, 20(1):23–50, 1998a.

P. D. Killworth, C. McCarty, H. R. Bernard, G. A. Shelley, and E. C. Johnsen. Estimation of seroprevalence, rape, and homelessness in the united states using a social network approach. Evaluation review, 22(2):289–308, 1998b.

J. P. Kunke, I. Laga, X. Niu, and T. H. McCormick. Comparing the robustness of simple network scale-up method estimators. Sociological Methodology, page 00811750241242791, 2024.

I. Laga, L. Bao, and X. Niu. Thirty years of the network scale-up method. Journal of the American Statistical Association, 116(535):1548–1559, 2021.

I. Laga, L. Bao, and X. Niu. A correlated network scale-up model: Finding the connection between subpopulations. Journal of the American Statistical Association, pages 1–18, 2023.

R. Maltiel, A. E. Raftery, T. H. McCormick, and A. J. Baraff. Estimating population size using the network scale up method. The annals of applied statistics, 9(3): 1247, 2015.

T. H. McCormick. The network scale-up method. The Oxford Handbook of Social Networks, page 153, 2020.

T. H. McCormick and T. Zheng. Latent demographic profile estimation in hard-to-reach groups. The annals of applied statistics, 6(4):1795, 2012.

M. Newman. Networks. Oxford university press, 2018.

H. Ocagli, D. Azzolina, G. Lorenzoni, S. Gallipoli, M. Martinato, A. S. Acar, P. Berchialla, D. Gregori, I. S. Group, et al. Using social networks to estimate the number of covid-19 cases: The incident (hidden covid-19 cases network estimation) study protocol. International journal of environmental research and public health, 18(11):5713, 2021.

B. Rozemberczki, R. Davies, R. Sarkar, and C. Sutton. Gemsec: Graph embedding with self clustering. In Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2019, pages 65–72. ACM, 2019.

M. Salganik, D. Fazito, M. Mello, et al. Estimating the number of heavy drug users in curitiba, brazil using multiple methods. Technical report, Technical report. UNAIDS, 2010.

M. J. Salganik, D. Fazito, N. Bertoni, A. H. Abdo, M. B. Mello, and F. I. Bastos. Assessing network scale-up estimates for groups most at risk of hiv/aids: evidence from a multiple-method study of heavy drug users in curitiba, brazil. American journal of epidemiology, 174(10):1190–1196, 2011.

S. Snidero, R. Corradetti, and D. Gregori. The network scale-up method: A simulation study in case of overlapping sub-populations. Advances in Methodology and Statistics, 1(2):395–405, 2004.

A. Srivastava, J. M. Ramirez, S. Díaz-Aranda, J. Aguilar, A. F. Anta, A. Ortega, and R. E. Lillo. Nowcasting temporal trends using indirect surveys. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 38, pages 22359–22367, 2024.

E. Sully, M. Giorgio, and S. Anjur-Dietrich. Estimating abortion incidence using the network scale-up method. Demographic Research, 43:1651–1684, 2020.

A. K. J. Teo, K. Prem, M. I. Chen, A. Roellin, M. L. Wong, H. H. La, and A. R. Cook. Estimating the size of key populations for hiv in singapore using the network scale-up method. Sexually transmitted infections, 95(8):602–607, 2019.

W. UNAIDS. Guidelines on estimating the size of populations most at risk to hiv. Geneva, Switzerland: World Health Organization, 51, 2010.

A. M. Verdery, S. Weir, Z. Reynolds, G. Mulholland, and J. K. Edwards. Estimating hidden population sizes with venue-based sampling: extensions of the generalized network scale-up estimator. Epidemiology (Cambridge, Mass.), 30(6):901, 2019.

T. Zheng, M. J. Salganik, and A. Gelman. How many people do you know in prison? using overdispersion in count data to estimate social structure in networks. Journal of the American Statistical Association, 101(474):409–423, 2006.

Supplementary Material

2.S.1. NSUM estimators

2.S.1.1. Basic estimators

The first frequentist estimators are characterized by using a combination of two procedures: an estimation of the network size and a posterior estimation of the unknown population. In particular, there are two different strategies for each procedure, consisting of using Eq. (2.1) or using a maximum likelihood approach. The estimation is sometimes replaced by a direct question about the size of the contact network, as in Garcia-Agundez et al. (2021).

Before delving into the estimators, we will set additional notation. We will denote as N_k the size of the population k, and p_k as the frequency of the population k in the total population, N_k/N.

First, we will focus on the estimation of the unknown subpopulation. It can be done if we have the information about the personal network size, d_i, but often this is not possible. Killworth et al. (1998a) suggested using Eq. (2.1) to obtain n estimates of the number of the unknown population, scaling by N the quotient of the report of the unknown subpopulation, y_{iu}, by the degree (d_i) of the individual i

$$N_u^i = N\frac{y_{iu}}{d_i}, \tag{2.S.1}$$

and then take the average to obtain an estimation of the unknown subpopulation size

$$\hat{N}_u^A = \frac{1}{n}\sum_{i=1}^{n} N_u^i. \tag{2.S.2}$$

The second proposal for the estimation of the unknown subpopulation was also introduced in Killworth et al. (1998a). They assumed that the $y_{iu} \sim Bin(d_i, p_u)$ and they are independent, where p_u is N_u/N. Subsequently, they use a maximum likelihood estimate to obtain an estimation of N_u,

$$\hat{N}_u^{ML} = \frac{\sum_{i=1}^{n} y_{iu}}{\sum_{i=1}^{n} d_i}. \tag{2.S.3}$$

For the estimation of the personal network size, the process follows a similar process. One version employed the fundamental Eq. (2.1) and the known population to obtain L estimates of the d_i, and then take the average:

$$\hat{d}_i^A = \frac{1}{L}\sum_{k=1}^{L} d_i^k, \tag{2.S.4}$$

where the d_i^k is the estimate corresponding to $y_{ik}p_k^{-1}$. Note that the $N_k's$ are known. The second approach is similar to the case of the population estimates, and assumes

that the y_{ik} are L independent random variables that follow a binomial distribution, $y_{ik} \sim Bin(d_i, p_k)$. The maximum likelihood is:

$$\hat{d}_i^{ML} = N \cdot \frac{\sum_{k=1}^{L} y_{ik}}{\sum_{k=1}^{L} N_k} \tag{2.S.5}$$

Combining these equations leads to several plug-in estimators. The PiMLE estimator was the first NSUM estimator, it was introduced in Killworth et al. (1998a) and combines Eq. (2.S.4) and Eq. (2.S.3):

$$\hat{N}_u^{PiMLE} = \frac{N}{n} \sum_{i=1}^{n} \frac{y_{iu}}{\hat{d}_i^A}. \tag{2.S.6}$$

.

Another plug-in estimator is the MLE, which estimates the personal network size in a different manner using Eq. (2.S.5):

$$\hat{N}_u^{MLE} = N \frac{\sum_{i=1}^{n} y_{iu}}{\sum_{i=1}^{n} d_i^{ML}} = \sum_{i=1}^{n} y_{iu} \frac{\sum_{k=1}^{L} N_k}{\sum_{i=1}^{n} \sum_{k=1}^{L} y_{ik}}. \tag{2.S.7}$$

The MLE is by far the most popular estimator (see (Ocagli et al., 2021), for example), and its first appearance was in Killworth et al. (1998b). The mean of sums estimator (MoS) is an alternative method proposed in Killworth et al. (1998a) but owes its name to the article Habecker et al. (2015). This method uses Eq. (2.S.4) and Eq. (2.S.2), and hence, it does not need the binomial assumptions.

$$\hat{N}_u^{MoS} = \frac{N}{n} \sum_{i=1}^{n} \frac{y_{iu}}{\hat{d}_i^A}, \tag{2.S.8}$$

The last combination of the previous equations is the one that uses Eq. (2.S.5) and Eq. (2.4). This estimator has not been employed yet, as Kunke et al. (2024) pointed out, but we consider it appropriate to include it since it can help to analyze the effect of the different combinations. The expression is:

$$\hat{N}_u^{KLN} = \frac{N}{n} \sum_{i=1}^{n} \frac{y_{iu}}{\hat{d}_i^{ML}} \tag{2.S.9}$$

On the other hand, some estimators can be produced by combining the equations of the population size, Eq. (2.S.2) and Eq. (2.S.3), with the values of the personal network size provided by the participants (δ_i). This is the approach of the study Garcia-Agundez et al. (2021) when the elicitation of the network size is integrated with Eq. (2.5):

$$\hat{N}_u^{RoA} = N \frac{\sum_{i=1}^{n} y_{iu}}{\sum_{i=1}^{n} \delta_i}. \tag{2.S.10}$$

We employ another alternative using the means instead, the Average of Ratios (AoR, see Eq. (2.S.2)):

$$\hat{N}_u^{AoR} = \frac{N}{n} \sum_{i=1}^{n} \frac{y_{iu}}{\delta_i} \tag{2.S.11}$$

The different types of methods are summarized in the next table.

Degree \ Size	Average	Max. Likelihood
Average	MoS	KLN
Max. Likelihood	PiMLE	MLE
Direct	AoR	RoA

Table 2.S.1: Basic NSUM estimators. The rows represent the way to obtain the degree, and the columns the size.

2.S.1.2. Generalized Network Scale-Up Method

The Generalized Network Scale-up Method (GNSUM) was suggested by Feehan and Salganik (2016) to reduce the transmission error. The approach is not based on Eq. (2.1) and includes additional data from the unknown subpopulation. This data is known as enriched aggregated relational data (EARD) and is made up of questions about the number of alters within an alter group that knows the participant's belonging to the unknown subpopulation. The alter groups are known subpopulations that can differ from the known subpopulations used to collect ARD. The EARD is used instead of the ARD of the known subpopulations. Additionally, the GNSUM has other assumptions, such as the number of in-reports (number of people that a participant reports) is equal to the number of out-reports (number of participants that report an individual), and other conditions related to the visibility of the members of the unknown subpopulation within the alters groups (see (Feehan and Salganik, 2016) for more details). The GNSUM estimator is:

$$\hat{N}_u^{GNSUM} = \frac{\sum_{i \in s_F} y_{iu}/\pi_i}{\hat{\tilde{v}}_{u,F}}, \tag{2.S.12}$$

where s_F is a sample from the frame population, π_i are the probabilities of inclusion from a sampling design, and $\hat{\tilde{v}}_{u,F}$ is the estimated relative visibility defined as follows:

$$\hat{\tilde{v}}_{u,F} = \frac{N_F}{N_{\mathcal{A}}} \frac{\sum_{i \in s_U} \sum_j \tilde{v}_{i,\mathcal{A}_j}/c\pi_i}{\sum_{i \in s_U} 1/(c\pi_i)}, \tag{2.S.13}$$

where s_U is a sample form the unknown subpopulation, $\mathcal{A}_j$ is the alter group j, $N_{\mathcal{A}}$ is the sum of the sizes of the alter groups, N_F the number of individuals of the frame population, $\tilde{v}_{i,\mathcal{A}_j}$ is the number of people in the alter group who knows that participant i belongs to the unknown population, and c a constant of proportionality.

In our particular case, we assume that the subset of possible participants is the entire population, and hence $N_F = N$. In addition, we use as alters groups the known subpopulations employed for degree estimation. Then, we have that $N_{\mathcal{A}} = \sum_k N_k$ and $\tilde{v}_{i,\mathcal{A}_j}$ is the number of people of the subpopulation j that are aware of the belonging of i to the unknown subpopulation, which without transmission error matches the response y_{ij}. Finally, we sample uniformly from the total population and the unknown subpopulations, and therefore, the inclusion probabilities are n/N and $|s_h|/N_u$.

The name of GNSUM comes from the theorem in Feehan and Salganik (2016), which says that under the conditions above, the GNSUM estimator is equal to the unknown subpopulation estimate Eq. (2.S.3) corrected by some adjustment factors (Feehan and Salganik, 2016). One of these adjustment factors is the transmission rate, τ (the ratio between the numbers of in-reports and the total connection from the unknown subpopulation to the rest of the frame). Additionally, Feehan and Salganik (2016) proposed using the correction of the MLE estimator. This is also known as GNSUM in the literature.

2.S.1.3. Overdispersed estimator

The Overdispersed estimator (Zheng et al., 2006) is the first Bayesian NSUM estimator to appear in the literature. It was created to handle the overdispersion of the data, which is likely produced by the variation of the degree distribution and the barrier error. The model is:

$$y_{ik} \sim Negative\ Binomial\left(mean = e^{\alpha_i+\beta_k}, overdispersion = \omega_k\right). \qquad (2.S.14)$$

where $k = 1, \cdots, L$ or h, e^{α_i} is the expected degree of i, e^{β_k} is the proportion of links of the group k The Poisson distribution is due to the variation of the degrees in some real networks, and the gamma distribution is used to model the barrier error. The ω_k represents a shape parameter of the gamma distribution of the propensity of forming ties with the subpopulation k. Moreover, they assign a normal prior distribution for α_i and β_k, a uniform prior in the interval $(0, 1)$ to $1/\omega_k$, and propose a Gibbs sampling. For further details about the prior distribution and the algorithm, see (Zheng et al., 2006).

2.S.1.4. Maltiel's Estimators (Maltiel et al., 2015)

A decade later, Maltiel et al. (2015) proposed a Bayesian estimator and 4 modifications to handle the barrier error, the transmission error, the recall error, and combinations of them. The Maltiel et al. (2015) estimators include priors whose support is bounded. Whenever the proposed MCMC values are outside the range of the priors, a new value is proposed until a value falls in the prior range. Bastos et al. (2018) pointed out that issue; they found that the estimators are sensitive to the variance of the network.

We found that, under our setting, the estimators do not get a reasonable acceptance rate. More concretely, we tested the estimators in our four main scenarios (see Table 2.2) in 20 samples, and we found that the estimators lasted more than two hours because of the acceptance rate, and some did not finish in three hours. In the two examples that finished, neither estimator was the best. Later, we tested the four estimators separately, and three of them lasted around 20-30 minutes in two

scenarios, which is more than expected without acceptance rate issues. The rest of the combination lasted more than an hour. Additionally, in only one example, an estimator was the best. For that reason, we do not include these estimators in the rest of the paper.

2.S.1.5. TPC Estimator (Teo et al., 2019)

Teo et al. (2019) presented a model considering that the ARD follows a Poisson distribution. Let λ be a scale-up parameter such that $y_{ik} = \lambda N_k$, for every $k = 1, \cdots, L$ or h; and α_i is a parameter that characterizes the variability in the network size. The basic model of Teo et al. (2019) is:

$$y_{ik} \sim Poisson\left(\lambda \alpha_i N_j\right). \tag{2.S.15}$$

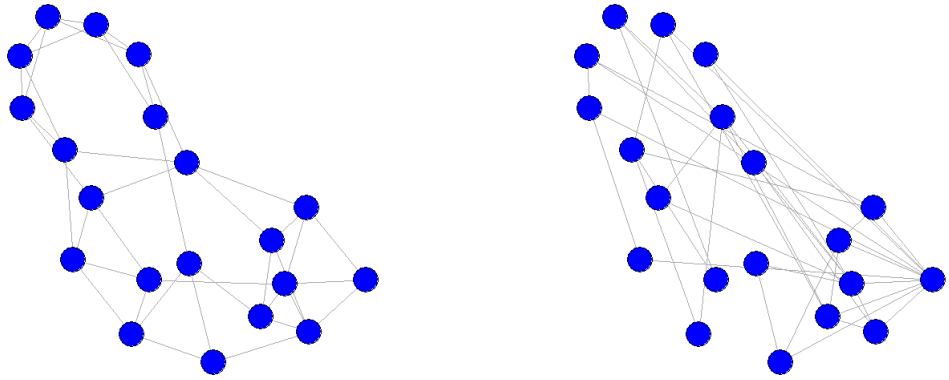
The prior distributions of the N_k are uniform distributions with large support to make them uninformative, $\alpha_i \sim N(0, \tau)$, and the prior of τ and λ are uniform distributions between 0 and 10. We include the algorithm from the supplementary material of (Zheng et al., 2006).

This work includes other model improvements, which allow transmission and barrier errors to be controlled. They introduce questions about the perceptions of the subpopulations and corrections through demographic variables. These last adjustments are not included because they exceed the scope of the study.

2.S.2. Networks

Our interest lies in topologies with real properties. We use a small-world network (a Watts-Strogatz model) because many social networks present this type of effect. On the other hand, we use a scale-free network generated by a preferential attachment process to evaluate the effect of the degree distribution since the other network selected has nodes with similar degrees.

Watts-Strogatz As a structure for social networks, the Watts–Strogatz approach is one of the best options, since it has two important small-world properties: a high clustering coefficient and a short average path length (Newman, 2018). However, it has the disadvantage of its degree distribution, since it has a pronounced peak and then decays exponentially, obtaining a network where many nodes have similar degrees. The two variables of interest in the Watts–Strogatz model are the number of neighbors that each individual has and the amount of randomness that is introduced to the connections between neighbors. The network is built as follows: the nodes are placed and connected to their nearest neighbors. Then, with probability p (the probability of randomizing connections), an edge is removed and placed between two nodes chosen at random.



(a) A Watts-Strogatz network (b) A preferential attachment network

Figure 2.1: Two networks with similar average degree. Figure (a) has more triadic closures and uniform degree distribution, while Figure (b) has nodes with many links and few triangles.

Preferential attachment The degree distribution of the Watts-Strogatz model is often unrealistic, and some networks have scale-free structures with a power law degree distribution, which may better represent the degrees of a real social network, such as the links in the World Wide Web and the citation of scientific papers (Newman, 2018). To model this possibility, we will consider the case of a Preferential Attachment. The construction of the graph is done by adding m vertices successively to every node so that the vertices join proportionally to the node with the most connections. However, its clustering coefficient is small because it does not generate either local clustering or triadic closures.

According to the type of the network, the distribution of the unknown subpopulation may differ substantially. The infection stage in an SIR process on a network with a high clustering coefficient, such as the Watts-Strogatz, is more likely to stop or not reach all the nodes. In a network with few triadic closures, opposition is hardly found (Easley et al., 2012). Furthermore, to compare the two approaches, we select the parameters of the SIR process so that the number of infected nodes is the same.

2.S.3. Recall error modeling

Modeling the personal network size recalling. Let δ_i be the personal network size. We want that $y_{iu} \leq d_i$ to get consistent responses. Additionally, we want the truncated normal to be symmetric and with mean δ_i. For that reason, the

simulated degree is generated by rounding a two-sided truncated normal distribution with mean $\mu = \delta_i$, standard deviation $\sigma_i = \delta_i\rho$, lower bound $a = y_{iu} - 0.5$ and upper bound $b = 2\delta_i - y_{iu} + 0.5$:

$$TN(\mu = \delta_i, \sigma = \delta_i\rho, a = y_{iu} - 0.5, b = 2\delta_i - y_{iu} + 0.5). \tag{2.S.16}$$

The lower limit is selected because the declared degree, δ_i, must be greater than the ARD of the unknown subpopulation. The upper limit comes from the fact that if $y_{iu} \leq \delta_i$ then $\delta_i \leq 2\delta_i - y_{iu}$. We add and subtract 0.5 so that the boundaries of the normal can be reached. In addition, the mean and bounds are also selected to have the expected value of the real personal size.

Modeling the ARD. Similarly, we obtain the ARD with barrier error, y_{ik}, by rounding a truncated normal:

$$TN(\mu = x_{ik}, \sigma = x_{ik}\rho, a = z_{ik} - 0.5, b = 2x_{ik} - z_{ik} + 0.5) \tag{2.S.17}$$

where x_{ik} is the real number of people to whom individual i is connected, z_{ik} is the real number of contacts that belong to both the unknown subpopulation and subpopulation k, and $\rho \in (0, 1)$ is the memory factor. We assume that the recall error is also present in the question associated with the unknown subpopulation. There are two natural bound: $y_{iu} \geq 0$ and $y_{iu} \leq \delta_i$. But, we add some constraints to get a distribution as symmetrical as possible, $a = max(-0.5, 2x_{iu} - \delta_i + 0.5)$ and $b = min(2x_{iu} + 0.5, \delta_i + 0.5$. The truncated normal distribution is then:

$$TN(\mu = x_{iu}, \sigma = x_{iu}\rho, a, b). \tag{2.S.18}$$

The differences of the intervals from the mean are equal to $min(x_{iu}, \delta_i - x_{iu})$. Later, we get the y_{iu} value by taking a value of the distribution and rounding it

Modeling the EARD. The philosophy in this case is the same. For simplicity, we do not consider the possible combination of barrier error in the other variables as we did before. We model the visibility report $\tilde{v}_{i,\mathcal{A}_k}$, by rounding the truncated normal

$$N(\mu = v_{i,\mathcal{A}_k}, \sigma = v_{i,\mathcal{A}_k}\rho, a = -0.5, b = 2v_{i,\mathcal{A}_k} + 0.5), \tag{2.S.19}$$

where $v_{i,\mathcal{A}_k}$ is the total number of people who know they belong to the unknown subpopulation of i.

2.S.4. Results

We show the results of the simulation. The results are presented in tables with rows representing the values for the mean, standard deviation, MAE, and MSE for

the different NSUM estimators, and columns representing the parameters of the simulation. The W-U scenario represents the case of a Watts-Strogatz network and an unknown population distributed uniformly. The W-S scenario is the case of a Watts-Strogatz network and SIR process for the unknown population. The case of the preferential attachment network and uniform distribution of the unknown population is the P-U scenario. Finally, the P-S represents the preferential attachment network and the SIR process.

For the transmission error simulation, we show the results for the W-U scenario in Tables 2.S.2 and 2.S.3, the W-S scenario in Tables 2.S.4 and 2.S.5, the P-U scenario in Tables 2.S.6 and 2.S.7, and the P-S scenario in Tables 2.S.8 and 2.S.9.

Regarding the recall error simulations, we provide the results for the W-U scenario in Table 2.S.10, the W-S scenario in Table 2.S.11, the P-U scenario in Table 2.S.12, and the P-S scenario in Table 2.S.13.

In the case of the number of subpopulations, we show the results for the W-U scenario in Table 2.S.14, the W-S scenario in Table 2.S.18, the P-U scenario in Table 2.S.22, and the P-S scenario in Table 2.S.26. Additionally, we include all the cases of disjoint subpopulations in Tables 2.S.15, 2.S.19, 2.S.23 and 2.S.27, all the cases of not disjoint small subpopulations in Tables 2.S.16, 2.S.20, 2.S.24, 2.S.28, and the cases of disjoint and small subpopulations in Tables 2.S.17, 2.S.21, 2.S.25 and 2.S.29.

For the subpopulation size simulations, we include the results of the W-U scenario in Table 2.S.30, the W-S scenario in Table 2.S.32, the P-U scenario in Table 2.S.34, and the P-S scenario in Table 2.S.36. The cases of disjoint subpopulations are in Tables 2.S.31, 2.S.33, 2.S.35 and 2.S.37.

Finally, for the sample size, we show the W-U scenario in Table 2.S.38, the W-S scenario in Table 2.S.39, the P-U scenario in Table 2.S.40, and the P-S scenario in Table 2.S.41.

Metric	Method	Visibility factor									
		0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
Mean	PiMLE	12.01	22.66	28.2	41.62	50.76	59.8	64.92	80.82	85.92	100.86
	MLE	12.19	23.05	28.2	40.69	50.3	59.22	63.57	78.59	84.35	100.18
	MoS	12.43	23.51	28.65	42.75	52.47	61.07	66.05	82.36	88.56	103.15
	KLN	12.23	23.13	28.37	40.84	50.37	59.37	63.73	78.57	84.38	100.43
	GNSUM	132.13	**106.21**	**90.41**	**103.83**	**96.76**	**97.49**	**91.53**	**99.51**	**92.23**	96.2
	Overd.	**99.01**	147.55	147.53	177.69	192.22	181.36	184.49	187.54	204.86	239.67
	TPC	16.37	27.35	32.31	45	54.17	63.35	67.55	82.29	88.26	103.54
	RoA	12.24	22.95	28.2	40.99	50.39	59.22	63.89	78.36	84.32	100.13
	AoR	12.03	22.85	28	41.19	50.13	59.24	64.18	78.6	84.68	**99.88**
SD	PiMLE	**6.67**	**8.83**	11.98	15.33	10.98	**12.41**	17.09	18.69	19.87	17.77
	MLE	6.9	9.31	12.02	14.76	**10.92**	12.52	15.78	18.58	18.28	17.49
	MoS	7	9.12	12.03	16.53	11.11	13.14	17.86	19.55	21.7	17.91
	KLN	6.98	9.44	12.2	14.91	11.14	12.78	**15.75**	18.69	18.33	17.68
	GNSUM	76.2	50.99	36.54	39.59	21.71	21.63	22.66	23.81	21.27	**16.09**
	Overd.	40.95	69.03	73.77	111.31	96.95	106.57	137.67	104.84	117.47	160.47
	TPC	6.9	9.52	12.11	**14.61**	10.92	12.87	15.77	18.24	**17.85**	17.98
	RoA	6.95	9.26	12.01	14.99	11.35	12.61	16.29	17.88	18.7	17.05
	AoR	6.94	9.16	**11.79**	15.09	11.14	12.79	16.47	**17.81**	18.75	16.89
MAE	PiMLE	87.99	77.34	71.8	58.38	49.24	40.2	35.74	21.63	20.7	13.64
	MLE	87.81	76.95	71.8	59.31	49.7	40.78	36.66	23.15	20.14	13.57
	MoS	87.57	76.49	71.35	57.25	47.53	38.93	34.65	21.22	21.09	13.8
	KLN	87.77	76.87	71.63	59.16	49.63	40.63	36.41	23	20.15	13.56
	GNSUM	62.85	**38.22**	**31.92**	**33.34**	**17.8**	**17.17**	**19.89**	**20.1**	19.7	13.7
	Overd.	**31.07**	63.62	61.52	99.06	101.21	91.56	107.01	100.36	112.21	142.9
	TPC	83.63	72.65	67.69	55	45.83	36.65	33.08	20.5	**17.73**	13.56
	RoA	87.76	77.05	71.8	59.01	49.61	40.78	36.62	23.12	20.53	13.2
	AoR	87.97	77.15	72	58.81	49.87	40.76	36.37	22.73	20.66	**13.12**
MSE	PiMLE	7783.72	6055.46	5292.1	3631.6	2539.05	1762.83	1508.16	699.75	573.13	300.68
	MLE	7755.99	6004.27	5292	3724.59	2583.59	1812.2	1564	786.07	562.31	290.65
	MoS	7715.37	5929.26	5228.2	3537.13	2376.27	1679.66	1456.08	674.16	578.26	314.63
	KLN	7749.46	5993.17	5271.79	3710.75	2580.88	1806.18	1551.25	791.13	563.33	297.05
	GNSUM	6549.21	**2508.96**	**1360.62**	**1503.47**	**458.19**	**450.91**	**559.6**	**538.94**	490.14	**260.45**
	Overd.	**1593.92**	6788.51	7429.02	17805.1	17434.88	17409.63	25144.66	18105.89	24102.97	43971.43
	TPC	7038.37	5364.51	4720.88	3228.01	2213.91	1500.85	1289	629.61	**440.41**	319.64
	RoA	7747.38	6017.36	5291.61	3695.85	2583.41	1814.32	1555.79	771.96	578.08	276.26
	AoR	7785.17	6031.17	5315.98	3675.19	2604.54	1816.62	1540.67	759.3	568.63	270.95

Table 2.S.2: Metrics for the simulations varying the visibility factor in the W-U scenario.

Metric	Method	Visibility factor									
		0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
Mean	PiMLE(adj)	121.74	102.58	104.24	102.63	105.24	103.92	**100.16**	112.66	102.56	103.26
	MLE(adj)	117.76	102.25	99.85	100.44	103.93	103.08	98.14	109.83	100.59	103.77
	MoS(adj)	123.73	105.53	104.54	104.54	105.78	106.17	102.92	114.3	102.98	105.6
	KLN(adj)	117.96	102.76	99.41	**100.24**	103.61	103.04	98.7	109.48	**99.96**	103.79
	GNSUM	**114.57**	93.16	97.72	95.06	**100.34**	**100.89**	98.4	**109.32**	97.01	**99.88**
	Overd.(adj)	1039.99	649.97	453.57	471.88	356.59	261.88	231.5	228.07	231.38	187.98
	TPC(adj)	161.28	121.61	112.62	110.57	111.93	110.42	104.2	115.05	105.43	108.11
	RoA(adj)	117.69	101.56	99.69	100.29	103.83	103.04	98.14	109.49	101.17	102.76
	AoR(adj)	115.35	**101.1**	**100.02**	100.72	103.93	103.35	98.21	109.76	101.48	102.78
SD	PiMLE(adj)	82.9	35.9	40.08	35.04	27.49	20.38	23.1	21.33	17.6	23.11
	MLE(adj)	80.28	37.43	39.76	31.97	27.51	19.65	22.81	20.42	**17.2**	22.95
	MoS(adj)	84.87	35.89	40.46	35.91	30.16	20.28	23.04	23.43	17.91	23.73
	KLN(adj)	79.94	37.33	39.35	31.97	28.43	19.63	22.78	20.95	17.43	23.36
	GNSUM	**72.9**	**35.65**	41.61	**30.97**	**26.25**	**18.45**	22.85	**19.41**	18.43	**22.41**
	Overd.(adj)	505.65	191.56	190.42	246.84	231.07	154.63	151.65	158.24	128.25	103.1
	TPC(adj)	83.33	36.8	40.13	32.48	28.63	20.35	22.78	20.22	18.04	23.73
	RoA(adj)	80.2	36.38	38.98	32.33	27.01	19.22	22.53	20.5	17.77	23.01
	AoR(adj)	79.89	35.71	**38.51**	32.68	26.93	19.21	**22.4**	19.86	18.18	23.09
MAE	PiMLE(adj)	67.01	28.86	34.07	27.85	18.24	16.62	16.99	20.44	14.81	19.08
	MLE(adj)	65.4	28.85	34.4	26.36	18.73	16.09	16.32	18.65	**14.01**	**18.72**
	MoS(adj)	70.31	29.2	33.93	29.02	20.39	17.49	17.33	22.13	14.97	19.63
	KLN(adj)	65.71	28.89	34.05	26.45	19.11	16.39	16.8	18.89	14.01	19.03
	GNSUM	**59.13**	28.27	36.3	**23.36**	**18.02**	**15.35**	16.28	**17.78**	14.51	18.84
	Overd.(adj)	939.99	549.97	359.18	371.88	260.84	164.41	137.57	133.62	137.74	93.14
	TPC(adj)	73.57	33.01	**32.71**	27.69	21.14	17.7	16.39	21.5	14.99	20.21
	RoA(adj)	65.14	27.46	33.39	26.43	18.84	15.94	**15.94**	18.34	14.4	18.87
	AoR(adj)	64.44	**27.28**	33.15	26.55	18.73	16.21	16.21	18.24	14.7	18.88
MSE	PiMLE(adj)	7001.48	1231.35	1544.1	1173.48	745.37	409.98	506.93	592.63	300.99	518.21
	MLE(adj)	6438.46	1335.67	1502.19	971.22	734.35	376.28	497.87	492.71	**281.4**	514.71
	MoS(adj)	7406.61	1254.22	1575.87	1245.84	897.66	428.92	512.78	725.9	313.74	566.37
	KLN(adj)	6394.01	1331.69	1471.41	971.01	780.92	375.47	494.5	506.66	288.69	532.68
	GNSUM	**5261.27**	1253.93	1649.83	**935.49**	**654.54**	**324.32**	498.4	**444.7**	331.52	**476.98**
	Overd.(adj)	1126481.97	337321.22	159459.1	196183.38	116565.05	48922.41	39140.37	40190.13	32885.62	17839.67
	TPC(adj)	10352.2	1753.25	1689.07	1114.18	921.15	502.07	510.82	614.61	338.64	600.91
	RoA(adj)	6422.9	1260.03	1443.29	992.81	707.46	360.24	485.55	489.1	301.5	510.39
	AoR(adj)	6299.2	**1212.54**	**1408.9**	1014.83	704.2	361.63	**479.88**	470.03	316.15	514.09

Table 2.S.3: Metrics for the simulations of corrected estimators varying visibility factor in the W-U scenario.

Metric	Method	Visibility factor									
		0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
Mean	PiMLE	8.78	16.47	30.91	34.23	41.88	55.02	51.29	78.2	93.43	93.8
	MLE	8.79	16.14	30.99	34.51	42.29	57.47	51.62	81.18	93.3	91.65
	MoS	9.05	16.37	30.43	33.76	42.8	54.2	52.08	76.95	90.83	92.59
	KLN	8.8	16.32	30.95	34.56	42.31	57.41	51.75	81.42	92.78	92.19
	GNSUM	94.39	**78**	**100.41**	**84.91**	**83.06**	**92.5**	**71.4**	**99.62**	103.07	88.38
	Overd.	**95.9**	122.66	131.56	156.02	163.42	170.8	161.55	174.34	239.82	165.36
	TPC	12.94	20.08	34.81	38.94	46.23	60.93	55.47	85.23	**97.67**	**95.68**
	RoA	8.82	16.14	30.85	34.22	41.92	57.71	51.95	81.11	93.12	91.5
	AoR	8.96	16.14	31.15	34.31	41.87	57.58	52.38	80.71	93.96	93.02
SD	PiMLE	**7.14**	12.96	17.29	17.78	18.59	28.16	27.67	35.9	40.53	55.22
	MLE	7.15	12.8	18.43	19.04	18.4	29.59	28.04	38.88	39.48	51.82
	MoS	7.37	13.28	**16.54**	**16.82**	18.7	**25.83**	**27.38**	**34.19**	**36.07**	52.04
	KLN	7.2	13.02	18.41	18.93	18.73	29.01	28.08	38.81	38.8	52.24
	GNSUM	85.38	65.28	55.45	49.75	38.11	49.02	37.05	48.08	45.54	**49.69**
	Overd.	37.43	57.05	57.52	80.81	56.68	91.48	62.81	75.47	96.82	102.25
	TPC	7.17	12.77	18.18	19.07	18.41	29.77	28.22	38.92	39.41	52.05
	RoA	7.21	**12.75**	18.13	18.96	**18.3**	29.81	28.44	38.83	39.13	52.22
	AoR	7.27	12.85	18.51	18.97	18.5	30.64	29.48	38.12	40.38	54.96
MAE	PiMLE	91.22	83.53	69.09	65.77	58.12	49.24	52.7	34.8	33.51	41.39
	MLE	91.21	83.86	69.01	65.49	57.71	47.74	52.32	35.49	33.35	41.22
	MoS	90.95	83.63	69.57	66.24	57.2	48.83	51.77	34.62	**30.43**	39.9
	KLN	91.2	83.68	69.05	65.44	57.69	47.45	52.19	35.47	32.7	41.04
	GNSUM	66.14	59.97	**38.64**	**41.42**	**34.68**	**35.67**	**40.67**	37.7	36.72	40.92
	Overd.	**29.41**	**43.19**	58.97	62.34	69.47	76.73	66.05	83.36	139.82	81.13
	TPC	87.06	79.92	65.19	61.06	53.77	44.84	49.44	**33.39**	32.9	**39.84**
	RoA	91.18	83.86	69.15	65.78	58.08	47.63	52.4	35.45	33.2	41.86
	AoR	91.04	83.86	68.85	65.69	58.13	48.09	52.68	34.99	33.99	43.06
MSE	PiMLE	8368.61	7136.11	5056.86	4626.12	3705.95	2776.79	3100.18	1699.5	1604.13	2935.13
	MLE	8368.01	7187.69	5084.66	4633.63	3652.05	2640.39	3087.48	1790.34	1525.41	2620.42
	MoS	8324.05	7161.44	5100.32	4656.47	3604.44	2732.02	3008.29	**1641.52**	**1319.8**	2627.75
	KLN	8367.46	7164.1	5089.9	4622.94	3661.93	2613.37	3077.31	1776	1481.98	2653.7
	GNSUM	6956.14	4532.48	**2921.25**	**2579.42**	**1666.85**	**2339.16**	**2121.79**	2196.23	1979.48	**2480.87**
	Overd.	**1347.51**	**3605.58**	4138.63	9341.81	7073.85	12962.25	7535.83	10936.5	28455.36	14204.42
	TPC	7627.62	6542.63	4563.63	4073.33	3213.71	2368.45	2739.63	1657.32	1480.58	2592.68
	RoA	8364.1	7186.31	5093.41	4668.73	3691.84	2632.52	3077.27	1789.14	1502.25	2663.11
	AoR	8338.17	7188.9	5065.4	4656.49	3704.28	2691.65	3093.25	1752.84	1585.47	2918.28

Table 2.S.4: Metrics for the simulations varying visibility factor in the W-S scenario.

Metric	Method	Visibility factor									
		0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
Mean	PiMLE(adj)	126.32	90.18	108.05	87.08	94.29	90.56	88.24	91.71	96.29	102.42
	MLE(adj)	125.93	89.19	111.62	89.2	94.13	94.5	89	93.39	94.59	101.19
	MoS(adj)	**121.35**	88.74	**106.06**	87.23	93.68	87.16	87.37	87.6	96.22	100.91
	KLN(adj)	125.47	89.11	110.76	89.57	94.09	94.52	89.13	92.9	94.56	101.61
	GNSUM	126.72	84.19	108.68	88.13	91.87	92.78	86.47	89.24	91.15	98.95
	Overd.(adj)	997.4	590.16	400.69	341.07	297.03	262.19	233.79	195.8	233.27	177.6
	TPC(adj)	165.36	**108.88**	125.47	**98.96**	**101.04**	**101.18**	**95.17**	**98.07**	**99.65**	104.78
	RoA(adj)	127.29	88.69	111.56	88.58	95.22	94.07	88.67	92.93	94.71	**100.38**
	AoR(adj)	130	90.12	111.49	88.19	94.84	94.38	88.55	93.58	95.4	101.9
SD	PiMLE(adj)	96.33	66.61	59.91	42.41	56.1	54.87	42.96	48.81	46.75	67.75
	MLE(adj)	96.51	65.66	63.63	42.49	56.14	61.96	45.26	47.61	45.7	66.72
	MoS(adj)	**90.77**	63.67	**58.02**	42.38	**53.77**	**50.5**	**41.02**	**43.72**	43.85	**62.24**
	KLN(adj)	95.62	65.33	62.68	42.99	55.13	61.65	44.92	47.44	44.89	67.1
	GNSUM	94.61	**61.9**	64.78	41.94	55.88	59.38	44.99	44.42	**43.14**	65.37
	Overd.(adj)	402.6	288.1	237.32	131.31	95.51	98.25	109.25	100.34	96.31	76.06
	TPC(adj)	95.03	64.94	63.61	43.01	55.74	62.19	45.59	47.6	46.02	65.91
	RoA(adj)	98.11	65.33	63.74	**41.88**	57.25	61.55	45.4	46.8	45.84	65.92
	AoR(adj)	99.86	66.55	64.44	42.55	55.82	61.84	45.3	47.95	47.12	67.74
MAE	PiMLE(adj)	75.19	56.88	49.36	34.43	39.6	44.74	35.98	39.54	36.08	50.63
	MLE(adj)	74.87	56.94	53.18	36.12	39.52	48.17	37.85	38.75	36.72	51.02
	MoS(adj)	**71.79**	54.86	**47.96**	**34.17**	37.8	**41.45**	**34.9**	37.29	**33.4**	**47.16**
	KLN(adj)	74.83	56.67	52.27	36.55	38.96	47.71	37.73	38.59	36.36	51.1
	GNSUM	73.82	**53.49**	53.67	35.35	40.82	47.15	37.79	**36.5**	35.38	49.69
	Overd.(adj)	897.4	491.01	300.69	241.15	197.03	162.69	141.74	111.79	133.27	82.51
	TPC(adj)	86.23	55.94	54.67	35.19	**36.43**	47.02	36.84	37.62	35.73	49.32
	RoA(adj)	75.96	56.92	53.56	35.32	40.27	48.14	37.48	38.37	36.24	50.38
	AoR(adj)	77.08	57.87	53.86	35.64	39.88	48.01	37.36	39.2	36.55	51.53
MSE	PiMLE(adj)	9509.11	4311.46	3474.81	1875.86	3022.32	2949.56	1891.34	2332.5	2090.3	4366.45
	MLE(adj)	9519.82	4212.42	3981.78	1831.58	3029.02	3676.96	2067.35	2196.68	2013.63	4230.35
	MoS(adj)	**8283.02**	3978.12	**3235.08**	1869.74	**2786.71**	**2587.54**	**1757.73**	**1969.74**	**1840.69**	**3681.12**
	KLN(adj)	9334.4	4172.88	3847.89	1864.14	2922.72	3640.13	2035.07	2188.3	1943.77	4279.76
	GNSUM	9216.49	**3890.11**	4062.49	1812.07	3032.95	3401.77	2106	1989.91	1846.03	4061.25
	Overd.(adj)	959318.68	319113.12	143915.57	74494.95	47487.66	35475.9	29239.74	18742.15	26573.18	11517.79
	TPC (adj)	12851.75	4085.75	4492.67	**1758.39**	2952.53	3675.23	1997.74	2156.58	2011.64	4150.21
	RoA(adj)	9888.97	4182.06	3993.66	1796.82	3136.05	3633.62	2086.1	2131.05	2024.14	4128.28
	AoR(adj)	10374.12	4304.73	4076.96	1859.66	2987.22	3664.77	2080.84	2225.18	2130.08	4363.18

Table 2.S.5: Metrics of the corrected methods and the GNSUM for the simulations varying the visibility factor in the W-S scenario.

Metric	Method	Visibility factor									
		0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
Mean	PiMLE	8.08	22.09	28.75	36.15	46.55	60.55	67.94	91.13	92.92	98.35
	MLE	8.71	21.36	25.84	38.54	46.85	58.45	64.01	84.96	88.91	92.28
	MoS	8.48	21.97	29.51	36.14	47.4	62.26	68.77	**94.15**	92.42	**99.38**
	KLN	8.67	21.39	25.62	38.41	46.4	58.14	63.72	85.12	87.92	91.69
	GNSUM	**116.27**	**114.59**	**100.29**	**100.02**	**94.16**	**118.18**	**103.75**	109.75	**101.9**	102.17
	Overd.	72.56	129.54	120.53	155.84	203.47	172.12	135.88	166.18	167.89	182.92
	TPC	12.86	25.86	29.69	42.74	51.45	62.71	67.79	89.23	92.97	96.81
	RoA	8.75	21.35	25.44	38.3	46.77	58.51	63.69	84.3	87.63	91.96
	AoR	8.1	21.45	27.14	35.86	46.95	58.54	66.31	85.61	89.99	96.52
SD	PiMLE	6.96	11.33	9.86	14.09	14.91	21.41	23.14	23.47	25.88	16.06
	MLE	7.05	**8.99**	8.6	12.72	**14.84**	**16.89**	18.69	20.61	20.63	11.86
	MoS	7.19	11.91	10.97	13.39	16.79	22.42	24.6	24.3	25.04	15.5
	KLN	7.01	9.04	8.58	**12.48**	15.16	16.92	**18.26**	20.81	**19.98**	**11.72**
	GNSUM	99.8	49.79	39.34	37.16	33.82	45.1	31.04	29.9	42.03	33.56
	Overd.	39.36	44.44	73.62	59.26	123.24	126.99	98.07	124.19	84.52	112.6
	TPC	**6.74**	9.29	8.7	12.71	15.53	17.17	18.42	21.03	20.52	11.79
	RoA	7.07	9.02	**8.31**	12.9	14.88	17.02	18.59	**20.48**	20.09	12.3
	AoR	7	11.01	9.06	13.44	15.06	19.9	22.54	21.66	22.19	15.93
MAE	PiMLE	91.92	77.91	71.25	63.85	53.45	41.42	34.04	21.15	22.63	12.46
	MLE	91.29	78.64	74.16	61.46	53.15	41.87	37.28	21.08	20.15	10.19
	MoS	91.52	78.03	70.49	63.86	52.6	40.21	34.19	21.21	21.95	11.43
	KLN	91.33	78.61	74.38	61.59	53.6	42.17	37.29	21.27	20.05	10.77
	GNSUM	82.34	**41.41**	**30.01**	**30.13**	**25.85**	**31.25**	**26.43**	23.14	29.58	23.46
	Overd.	**40.91**	45.06	59.92	62.64	118.86	95.73	62.66	82.68	74.25	90.99
	TPC	87.14	74.14	70.31	57.26	48.55	38.22	33.75	**19.24**	**18.7**	**8.76**
	RoA	91.25	78.65	74.56	61.7	53.23	41.71	37.57	21.66	20.04	10.2
	AoR	91.9	78.55	72.86	64.14	53.05	42.6	35.03	21.37	20.8	11.56
MSE	PiMLE	8495.96	6191.22	5169.45	4265.08	3067.9	1992.2	1536.19	602.06	686.48	247.84
	MLE	8380.33	6261.65	5570.12	3931.47	3033.84	1997.26	1627.54	629.6	527.37	193.24
	MoS	8424.62	6224.08	5083.61	4248.47	3034.99	1901.86	1550.55	595.28	653.23	228.56
	KLN	8387.08	6257.87	5602.71	3941.83	3091.71	2023.83	1633.04	633.09	525.11	199.59
	GNSUM	9726.38	**2567.96**	**1470.69**	**1311.96**	**1120.44**	2263.01	**929.41**	944.28	1681.48	1074.88
	Overd.	**2225.15**	2749.19	5571.06	6454.57	25135.87	20522.36	10423.45	19033.16	11395.52	18920.86
	TPC	7636.37	5578.04	5015.97	3432.7	2586.15	**1670.41**	1359.87	**536.17**	**449.44**	**142.17**
	RoA	8374.38	6263.4	5624.87	3964.62	3043.61	1996.67	1646.48	644.94	536.38	208.53
	AoR	8491.79	6285.39	5387.07	4285.43	3029.68	2095.32	1617.84	653.03	568.26	253.18

Table 2.S.6: Metrics for the simulations varying the visibility factor in the P-U scenario.

Metric	Method	Visibility factor									
		0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
Mean	PiMLE(adj)	93.98	**101.15**	105.89	102.33	97.65	93.56	101.87	103.8	95.62	111.58
	MLE(adj)	85.51	96.28	102.93	104.68	88.47	88.21	97.39	97.16	93.06	102.16
	MoS(adj)	101.72	104.95	106.96	101.85	**98.64**	94.05	103.8	103.92	**97.63**	112.49
	KLN(adj)	84.64	95.74	102.17	104.15	88.16	88.07	97.28	96.94	92.54	101.44
	GNSUM	**101.31**	105.58	**100.71**	115.94	96.15	89.36	**99.05**	98.73	107.06	117.82
	Overd.(adj)	900.18	584.27	437.98	456.85	385.11	228.96	184.22	245.01	177.04	165.46
	TPC(adj)	124.85	117.14	116.98	115.14	97.15	**95.51**	103.37	102.53	97.27	105.96
	RoA(adj)	85.81	95.74	103.14	104.27	87.66	88.1	97.05	95.94	93.06	**101.25**
	AoR(adj)	90.25	97.84	105.99	**99.71**	93.25	90.68	98.58	**99.08**	93.89	109.09
SD	PiMLE(adj)	86.67	52.78	57.14	38.29	29.68	25.82	19.4	28.97	24.17	23.3
	MLE(adj)	64.94	37.8	38.96	31.32	18.7	22.97	**18.43**	25.22	21.18	17.83
	MoS(adj)	97.11	55.23	60.17	38.44	30.56	27.04	18.66	28.94	25.52	23.61
	KLN(adj)	**64.01**	37.44	38.8	31.23	18.6	23	18.71	25.66	**21.15**	17.62
	GNSUM	87.17	60.75	45.15	54.94	46.09	32.57	23.65	41.92	39.5	34.73
	Overd.(adj)	442.37	252.31	238.52	222.76	158.23	128.4	101.12	176.73	111.12	115.97
	TPC(adj)	66.61	**37.19**	**37.62**	**30.59**	19.08	**22.57**	18.85	25.14	21.3	**17.43**
	RoA(adj)	65.66	37.43	39.73	31.61	**17.82**	23.05	18.47	**24.39**	21.28	17.85
	AoR(adj)	81.25	43.73	52.26	34.28	25.4	24.14	18.98	27.74	22.18	22.93
MAE	PiMLE(adj)	70.05	36.4	42.4	29.37	22.51	21.1	15.56	23.8	19.48	21.48
	MLE(adj)	56.09	**30.84**	29.03	27.66	17.71	20.13	15.39	21.28	16.95	15.54
	MoS(adj)	77.83	37.23	44.2	29.2	22.77	22.58	14.69	24.13	20.52	22.19
	KLN(adj)	**55.24**	30.94	**28.96**	27.48	17.52	20.3	15.68	21.52	16.92	15.26
	GNSUM	70.29	35.83	34.09	45.59	30.32	27.58	18.35	32.03	34.09	28.99
	Overd.(adj)	800.18	484.27	337.98	356.85	287.24	137.5	85.83	165.74	81.12	66.9
	TPC(adj)	59.98	31.51	29.88	29.6	**15.82**	**19.29**	**14.57**	**19.63**	**16.86**	**15.17**
	RoA(adj)	56.81	30.92	29.52	27.92	17.79	19.88	15.61	20.99	17.09	15.57
	AoR(adj)	66.64	32.86	39.56	**26.89**	21.81	20.26	16.24	23.77	17.71	20.79
MSE	PiMLE(adj)	7171.76	2647.74	3136.62	1398.4	842.57	674.6	361.05	811.49	574.13	649.73
	MLE(adj)	4216.48	1371.2	1450.7	953.63	465.34	640.27	**329.6**	612.47	474.21	306.68
	MoS(adj)	8962.58	2922.01	3487.91	1407.12	889.28	729.86	345.37	811.21	624.23	685.79
	KLN(adj)	**4128.45**	1349.53	**1434.73**	**943.58**	469.1	644.67	339.87	634.78	480.55	**297.1**
	GNSUM	7220.54	3536.79	1937.52	3121.74	2033.34	1120.85	532.25	1670.68	1532.04	1463.7
	Overd.(adj)	826190.35	294991.18	168276.37	174480.94	105073.22	32293.57	16808.22	50699.31	17666.11	17061.03
	TPC(adj)	4832.8	1608	1632.7	1118.21	**354.1**	**504.17**	348.82	606.98	**438.59**	324.25
	RoA(adj)	4297.45	**1348.82**	1509.14	967.17	454.12	646.3	332.78	**581.78**	478.19	304.26
	AoR(adj)	6366.9	1821.73	2630.89	1116.72	658.44	640.46	344.42	732.08	504.76	582.11

Table 2.S.7: Metrics of the corrected methods and the GNSUM for the simulations varying the visibility factor in the P-U scenario.

Metric	Method	Visibility factor									
	PiMLE	14.77	36.22	58.63	74.32	92.75	113.4	134.39	154.43	163.96	194.51
	MLE	15.09	36.23	56.82	71.81	91.79	106.49	130.48	143.51	156.38	183.16
	MoS	15.28	37.38	57.65	75.02	93.18	116.13	135.84	157.96	165.19	198.56
	KLN	15	36.36	56.24	71.57	91.09	105.83	129.66	143.06	155.35	182.03
Mean	GNSUM	91.99	**109.21**	**123.24**	**118.62**	120.68	122.12	**117.06**	**127.17**	**93.18**	**116.84**
	Overd.	**92.56**	134.36	140.66	181.32	183.77	204.49	177.63	231.67	176.63	192.52
	TPC	19.23	40.55	60.69	76.53	**96.06**	110.23	134.47	147.64	160.16	186.94
	RoA	15.16	36.27	56.44	71.56	91.89	**105.68**	130.63	142.11	156.63	183.34
	AoR	14.18	34.95	55.42	71.2	89.55	109.24	129.47	146.42	158.96	189.1
	PiMLE	5.18	12.84	20.25	17.54	22.09	24.77	26.04	29.18	42.07	34.68
	MLE	5.59	12.64	17.88	15.77	20.79	21.2	20.44	25.97	25.5	28.85
	MoS	5.98	14.34	18.7	17.61	21.68	24.99	25.26	31.37	41.99	36.39
	KLN	5.62	12.85	17.86	**15.32**	**20.45**	21.23	**19.71**	26.21	**25.38**	**28.09**
SD	GNSUM	47.34	55.79	70.55	69.01	45.57	41.19	53.92	51.12	34.87	54.82
	Overd.	47.8	94.96	96.31	90.27	105.16	139.95	110.77	161.39	54.59	69.62
	TPC	5.87	13.1	17.81	15.49	20.78	21.47	20.46	25.54	26.03	28.11
	RoA	5.71	12.57	**17.65**	15.72	20.45	**20.17**	20.53	**24.93**	25.83	28.67
	AoR	**4.66**	**11.88**	17.76	17.13	21.33	21.49	23.99	27.83	37.47	31.69
	PiMLE	85.23	63.78	42.3	27.03	18.88	23.88	34.64	54.43	63.96	94.51
	MLE	84.91	63.77	43.18	28.19	17.32	18.99	30.48	44	56.38	83.16
	MoS	84.72	62.62	42.43	26.65	18.5	25.42	36.6	57.96	65.19	98.56
	KLN	85	63.64	43.76	28.43	17.32	18.67	**29.66**	43.63	55.35	82.03
MAE	GNSUM	**37.09**	**45.25**	51.85	43.34	38.57	38.96	40.19	**38.99**	**26.37**	**36.6**
	Overd.	41.16	72.12	73.47	92.24	94.3	112.21	82.72	137.96	76.63	97.95
	TPC	80.77	59.45	**39.31**	**23.47**	**15.82**	20.39	34.47	47.64	60.16	86.94
	RoA	84.84	63.73	43.56	28.44	16.74	**17.57**	30.63	42.85	56.63	83.34
	AoR	85.82	65.05	44.58	30.46	19.12	19.93	29.8	46.54	58.96	89.1
	PiMLE	7290.52	4224.32	2100.73	951.63	515.91	762.6	1826.7	3771.4	5771.89	10075.53
	MLE	7238.94	4218.43	2168.42	1030.96	478.12	469.28	1326.06	2533.73	3795.9	7706.89
	MoS	7211.05	4116.06	2125.7	918.79	493.16	853.5	1890.42	4294.02	5925.03	10972.58
	KLN	7255.57	4207.13	2217.54	1031.12	476.6	462.31	**1249.15**	2506.43	3675.77	7477.94
MSE	GNSUM	**2192.91**	**3042.29**	5268.78	4870.51	2400.78	2101.29	3052.81	3220.45	**1201.47**	**3138.55**
	Overd.	2225.96	9747.4	10465.66	14354.12	17522.4	29523.73	17683.47	42081.16	8703.81	13165.08
	TPC	6556.88	3696.89	**1846.39**	**778.94**	**425.58**	542.58	1585.59	2888.96	4263.22	8309.44
	RoA	7228.33	4211.7	2193.77	1043.79	463.26	**418.83**	1338.69	**2363.62**	3840.93	7726.25
	AoR	7385.57	4365.81	2287.29	1108.26	541.51	523.87	1415.14	2891.11	4810.45	8892.46

Table 2.S.8: Metrics for the simulations varying the visibility factor in the P-S scenario.

Metric	Method	Visibility factor									
		0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
Mean	PiMLE(adj)	222.25	167.33	182.87	175.43	188.12	189.88	194.61	182.76	185.48	193.19
	MLE(adj)	222.92	185.79	173.96	171.38	189.48	189.41	187.25	182.03	175.8	184.81
	MoS(adj)	223.18	165.58	185.35	177.67	191.03	193.34	195.96	188.97	188.36	197.15
	KLN(adj)	221.92	184.6	172.58	170.37	188.01	187.9	186.2	181.44	174.92	183.53
	GNSUM	**134.01**	**129.68**	**101.63**	**91.82**	**114.55**	**114.27**	**119.23**	**106.46**	**96.85**	**113.54**
	Overd.(adj)	1175.74	680.04	527.11	556.4	384.28	239.66	226.17	202.92	189.69	192.56
	TPC(adj)	262.04	205.51	186.49	181.73	197.28	196.63	192.96	188.1	180.66	188.75
	RoA(adj)	221.62	184.77	174.53	170.68	187.41	188.69	186.29	181.4	174.83	183.89
	AoR(adj)	218.62	169.76	178.38	174.46	180.92	184.19	188.65	175	179.99	187.68
SD	PiMLE(adj)	93.39	52.13	67.07	51.97	34.8	39.78	38.27	45.53	39.27	28.27
	MLE(adj)	81.12	50.15	48.37	53.88	33.19	31.08	36.94	54.32	31.99	27.24
	MoS(adj)	91.56	54.4	68.62	53.48	39.52	41.07	40.54	48.8	37.58	27.8
	KLN(adj)	80.63	**49.81**	48.18	53.67	33.58	**29.59**	36.72	55.26	30.88	**26.82**
	GNSUM	**71.29**	85.13	**42.95**	**32.09**	50.83	38.7	72.82	**43.26**	**30.53**	56.17
	Overd.(adj)	469.81	372.84	250.74	301.39	227.96	91.54	94.4	86.16	138.15	68.05
	TPC(adj)	86.44	49.84	46.38	54.66	32.41	32.25	**36.6**	54.9	32.04	28.4
	RoA(adj)	80.58	50.71	48.04	53.38	**29.9**	31.59	37.91	53.49	31.33	27.79
	AoR(adj)	85.93	53.03	66.59	52.92	30.61	37.11	39.34	43.6	36.2	27.39
MAE	PiMLE(adj)	129.28	72.05	86.03	75.86	88.12	91.48	94.61	84.19	85.48	93.19
	MLE(adj)	124.59	87.98	75.94	72.15	89.48	89.41	87.25	82.05	75.84	84.81
	MoS(adj)	130.64	70.31	88.94	78.52	91.03	94.54	95.96	89.82	88.36	97.15
	KLN(adj)	123.64	86.83	74.81	71.49	88.01	87.9	86.2	81.65	74.92	83.53
	GNSUM	**55.4**	**61.15**	**35.8**	**26.65**	**37.34**	**30.44**	**50.83**	**34.73**	**23.88**	**42.71**
	Overd.(adj)	1075.74	580.04	427.11	456.4	284.28	139.95	130.64	107.14	95.39	99.01
	TPC(adj)	162.04	106.05	86.74	81.73	97.28	96.63	92.96	88.1	80.66	88.75
	RoA(adj)	123.03	86.7	76.83	71.95	87.41	88.69	86.29	81.49	74.83	83.89
	AoR(adj)	125.16	73.51	80.86	74.46	80.92	85.6	88.65	77.22	79.99	87.68
MSE	PiMLE(adj)	23230.6	7115.56	11141.3	8255.88	8915.41	9581.93	10341.55	8817.9	8771.03	9442.98
	MLE(adj)	21361.26	9748.7	7693.13	7852.94	9052.73	8912.43	8908.35	9532.44	6717.18	7898.02
	MoS(adj)	23137.41	**7111.98**	11757.93	8749.39	9770.65	10315.81	10768.99	10178.59	9148.22	10171.22
	KLN(adj)	21041.67	9514.64	7472.95	7688.37	8816.04	8557.65	8712.17	9533.25	6519.57	7659.96
	GNSUM	**5985.7**	7766.23	**1755.1**	**1045.25**	**2666.23**	**1626.27**	**5406.92**	**1819.29**	**895.19**	**3180.56**
	Overd.(adj)	1366912.64	468504.15	242147.59	294590.76	130184.4	27467.63	24385.16	17644.71	26176.65	12965.57
	TPC(adj)	33353.8	13492.79	9524.71	9518.5	10461.19	10325.61	9915.01	10624.34	7482.25	8642.46
	RoA(adj)	20959.08	9628.85	7746.63	7702.82	8489.81	8814.26	8811.85	9343.17	6531.82	7770.86
	AoR(adj)	21086.72	7538.17	10356.34	8204.78	7437.25	8395.78	9328.68	7429.96	7644.19	8400.26

Table 2.S.9: Metrics of the corrected methods and the GNSUM for the simulations varying the visibility factor in the P-S scenario.

Metric	Method	Memory factor									
		0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
Mean	PiMLE	102.28	105.69	95.65	107.81	107.21	106.65	110.03	103.26	107.93	97.81
	MLE	100.68	103.65	94.57	104.41	102.58	101.42	103	100.16	**100.7**	92.28
	MoS	103.77	107.81	96.82	109.67	108.64	109.32	113.29	106.48	110.23	**100**
	KLN	**100.66**	103.2	94.72	104.37	103.1	101.29	103.68	**100.15**	101.85	92.49
	GNSUM	97.98	**100.38**	90.29	**101.51**	**99.72**	**98.95**	**100.47**	98.2	97.91	88.92
	Overd.	232.05	143.29	146.58	171.16	188.05	151.75	198.75	139.2	196.96	212.31
	TPC	104.19	107.41	98.74	109.03	106.99	105.58	106.75	104.19	105.13	96.17
	RoA	100.75	103.31	95.15	105.49	104.51	101.71	105.67	98.95	99.1	91.83
	AoR	101.24	104.07	**99.11**	117.57	155.79	152.75	229.46	212.23	176.12	205.27
SD	PiMLE	16.75	20.56	18.06	18.98	29.92	22.49	27.72	30.63	27.93	33.81
	MLE	15.77	20.9	16.79	18.09	**26.82**	20.49	25.49	30.35	24.78	31.46
	MoS	16.31	22.63	17.21	20.13	31.33	23.54	28.47	31.14	29.03	36.48
	KLN	**14.99**	21.42	16.41	**17.84**	27.47	20.65	25.76	30.23	25.38	31.71
	GNSUM	17.29	22.32	**16.11**	17.91	28.8	23.77	26.17	28.83	**23.31**	**28.41**
	Overd.	158.51	46.44	81.32	81.33	111.88	75.79	93.7	77.44	92.16	115.32
	TPC	15.28	**20.55**	16.71	18.45	27.2	20.15	**25.45**	30.45	24.6	31.68
	RoA	16.23	21.39	16.54	19.5	32.25	**19.81**	28.55	**28.41**	25.46	32.2
	AoR	16.22	20.79	19.67	22.21	101.62	69.69	109.37	81.86	82.1	124.24
MAE	PiMLE	14.29	16.37	13.69	16.13	22.76	18.91	21.78	23.96	20.11	27.89
	MLE	13.44	16	12.72	14.8	**20.21**	16.83	20.44	24.65	17.98	26.71
	MoS	13.56	18.12	13.25	17.99	23.44	19.93	23.31	25.4	21.37	29.91
	KLN	**12.69**	16.58	**12.5**	14.44	20.29	16.87	**20.26**	24.65	18.66	27.24
	GNSUM	14.32	17.26	15.17	**14.37**	22.68	20.37	21.07	**23.74**	**17.84**	**24.83**
	Overd.	135.71	47.51	55.36	78.09	96.28	63.35	101.22	67.87	100.98	124.29
	TPC	13.41	16.13	12.65	16.01	20.56	16.34	20.3	24.36	18.67	26.36
	RoA	13.68	16.3	12.95	14.97	22.31	**15.71**	23.51	24.04	18.05	27.49
	AoR	13.52	**15.98**	15.08	21.59	60.59	60.34	137.84	112.23	85.94	121.7
MSE	PiMLE	271.6	433.93	328.72	403.32	902.54	524.75	830.67	902.12	804.2	1091.07
	MLE	236.66	428.15	297.23	330.27	**689.84**	400.8	**626.46**	875.19	583.74	999.89
	MoS	266.86	547.33	291.37	478.63	1007.24	613.08	946.72	963.27	905.12	1264.18
	KLN	**213.87**	446.07	283.69	321.37	726.6	406.6	644.08	867.99	615.29	1011.68
	GNSUM	287.94	473.23	340.98	**306.97**	788.15	537.94	651.06	793.05	**520.67**	**889.35**
	Overd.	41306.32	3923.59	8452.12	11348.53	19644.51	8134.91	18093.17	7232.84	17471.49	25246.33
	TPC	239.4	456.27	**266.82**	404.94	751.76	417.01	660.84	898.34	601.01	967.82
	RoA	250.87	445.68	283.42	391.55	1008.68	**375.79**	806.43	**767.79**	616.71	1051.93
	AoR	251.52	**427.29**	368.4	777.49	12923.23	7395.55	28122.71	18961.32	12197.02	25745.87

Table 2.S.10: Metrics for the simulations varying the memory factor in the W-U scenario.

Metric	Method	Memory factor									
		0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
Mean	PiMLE	107.59	96.2	92.95	88.29	91.42	106.47	81.95	81.5	88.25	101.15
	MLE	109.55	96.11	93.19	88.86	88.22	102.96	81.02	79.34	84.68	95.92
	MoS	**104.87**	94.35	92.77	87	85.08	102.29	82.62	79.68	87.58	100.71
	KLN	109.57	96.83	93.49	89.31	87.27	103.26	80.74	78.64	84.69	95.93
	GNSUM	106.44	94.36	92.38	84.65	84.87	**101.64**	78.39	78.41	82.37	90.88
	Overd.	213.72	216.31	180.06	196.88	180.97	187.9	172.41	142.59	181.35	187.71
	TPC	114.26	**100.32**	97.27	93.18	92	106.6	84.72	**83.38**	**89.34**	**100.38**
	RoA	109.01	95.94	93.85	89.1	90.11	103.81	80.88	78.84	86.56	96.94
	AoR	109.4	98.52	**100.72**	**99.56**	**105.92**	160.04	**111.86**	129.03	130.5	156.82
SD	PiMLE	56.21	53.84	35.08	39.86	51.41	60.42	31.66	50.24	53.33	69.62
	MLE	58.11	53.5	38.45	42.48	47.3	60.48	30.83	47.53	51.97	62.19
	MoS	**51.37**	**49.75**	**34.87**	**38.35**	**42.18**	**56.02**	**28.51**	47.76	49.71	69.84
	KLN	57.96	53.88	38.66	42.98	46.03	60.57	30.01	**46.64**	50.75	62.1
	GNSUM	56.56	52.95	37.76	40.49	44.77	58.31	31.12	47.63	**49.62**	**59.17**
	Overd.	85.42	104.25	47.41	82.6	93.64	106.1	87.77	58.77	79.48	82.36
	TPC	58.19	53.65	38.7	42.61	46.89	60.12	30.86	46.96	52.38	61.76
	RoA	57.26	52.57	39.66	43.17	50.09	59.57	29.93	49.88	54.47	66.82
	AoR	58.18	57.84	42.26	47.22	56.47	87.13	63.16	89.09	85.55	91.03
MAE	PiMLE	44.06	41.91	27.41	35.24	44.19	41.11	29.38	46.45	46.3	50.98
	MLE	43.82	41.1	31.39	36.43	42.64	42.13	30.15	43.22	46.27	47.58
	MoS	**41.4**	**39.22**	**26.83**	**34.78**	**38.89**	**38.05**	**27**	44.56	**44.17**	52.73
	KLN	44.19	40.81	31.19	36.77	41.81	42.52	29.85	43.42	45.81	47.91
	GNSUM	43.07	41.24	30.76	36.49	41.44	41.4	31.97	44.68	45.75	**46.21**
	Overd.	114.77	122.44	81.16	100	90.79	96.53	82.14	54.48	86.1	96.74
	TPC	43.82	40.15	30.5	35.47	41.03	41.07	28.12	**40.14**	45.11	46.59
	RoA	43.39	40.67	31.76	36.73	44.11	41.12	30.17	43.54	46.47	47.91
	AoR	44.84	43.26	33.22	39.73	47.22	77.28	48.53	59.64	67.03	74.99
MSE	PiMLE	3059.59	2768.13	1218.77	1646.38	2584.74	3510.11	1277.93	2739.95	2839.9	4606.5
	MLE	3298.72	2734.17	1450.72	1838.82	2264.24	3483.47	1262.97	2572.54	2800.81	3690.78
	MoS	**2530.31**	**2383.69**	**1207.53**	**1566.37**	**1912.95**	**2986.1**	**1074.18**	2580.33	**2501.74**	4634.09
	KLN	3283.13	2768.33	1462.27	1869.48	2174.96	3496.44	1226.52	2523.12	2681.16	3680.15
	GNSUM	3080.61	2695.76	1412.81	1793.37	2132.94	3232.41	1386.86	2621.36	2650.18	**3409.47**
	Overd.	19864.88	23853.04	8545.33	15866.32	14886.89	18420.49	12561.81	5095.13	12619.57	14137.47
	TPC	3420.64	2734.93	1430.18	1771.48	2153.07	3476.83	1138.19	**2371.56**	2720.59	3623.19
	RoA	3195.38	2642.22	1531.81	1889.43	2481.17	3385.26	1216.39	2811.67	2998.97	4250.78
	AoR	3304.41	3179.88	1697.46	2118.46	3065	10818.05	3930.23	8383.37	7884.06	11099.51

Table 2.S.11: Metrics for the simulations varying the memory factor in the W-S scenario.

Metric	Method	Memory factor									
		0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
Mean	PiMLE	**99.07**	94.13	104.98	96.77	**99.65**	95.65	101.72	96.26	108.72	99.57
	MLE	94.26	92.91	99.55	94.03	90.08	86.93	94.2	92.9	96.35	91.4
	MoS	101.28	95.26	105.92	**99.64**	99.63	**97.08**	104.73	96.32	108.94	**100.09**
	KLN	93.75	91.8	98.91	93.18	89.71	86.43	93.89	91.33	96.18	90.94
	GNSUM	111.04	91.06	**100.45**	97.15	96.74	89.28	97.36	**98.56**	102.56	106.04
	Overd.	139.06	163.3	164.06	197.53	182.53	178.37	192.98	172.72	189.3	152.36
	TPC	98.57	**96.96**	104.12	97.82	94.3	91.03	**98.32**	97.18	**100.25**	95.11
	RoA	93.75	92.38	98.82	94.27	89.86	88.42	94.06	93.96	97.97	89.63
	AoR	95.22	92.77	104.04	102.98	113.09	128.67	118.76	145.05	184.94	160.49
SD	PiMLE	18.6	22.01	29.24	24.8	40	22.63	35.4	**26.28**	28.38	38.31
	MLE	18.02	19.08	24.86	19	29.15	**17.84**	30.54	27.21	19.28	28.5
	MoS	20.02	22.05	28.24	24.78	41.21	24.44	35.49	27.27	26.89	40.1
	KLN	17.99	19.09	24.6	**18.44**	**28.6**	18.31	30.34	26.45	**19.17**	**28.29**
	GNSUM	37.7	23.14	29.5	31.96	35.2	26.94	41.73	44.7	27.2	35.31
	Overd.	65.12	71.72	108.91	107.71	89.55	121.77	89.11	77.91	72.81	73.98
	TPC	18.55	18.97	25.29	19.06	29.3	18.18	**30.15**	27.35	19.18	28.82
	RoA	**17.35**	**18.38**	**24.36**	19.83	29.66	19.1	33.52	28.21	21.17	32.97
	AoR	17.99	22.11	27.26	26.92	47.25	48.04	67.03	54.63	113.99	75.57
MAE	PiMLE	**13.5**	16.85	22.35	17.39	33.15	19.43	27.22	**20.71**	23.88	32.52
	MLE	15.09	15.74	20.43	14.75	22.26	19.11	23.51	22.62	16.37	23.56
	MoS	15.17	16.92	21.8	16.28	33.5	20.79	27.22	22.31	22.27	34.17
	KLN	14.91	16.42	20.29	14.29	22.16	19.43	23.43	21.94	**16.16**	**23.11**
	GNSUM	27.64	19.59	23.36	27.54	28.04	22.26	31.24	33.61	22.29	27.12
	Overd.	47.48	69.55	71.36	107.12	89.89	85.26	98.06	83.31	96.41	64.68
	TPC	15.58	**14**	20.93	**14.23**	**21.71**	**16.85**	**22.89**	22.98	16.83	23.46
	RoA	14.78	15.81	**19.65**	15.57	22.13	19.68	26	24.81	18.16	26.12
	AoR	14.94	17.23	20.39	20.8	37.14	41.48	41.96	56.08	88.62	75.69
MSE	PiMLE	329.43	494.62	837.16	594.76	1519.8	505.28	1193.27	**670.32**	841.42	1394.51
	MLE	341.4	396.14	587.51	378.65	905.44	472.96	919.63	753.62	366.62	845.34
	MoS	382.55	484.24	792.76	583.25	1613.47	575.96	1218.64	720.19	766.67	1527.93
	KLN	346.53	413.63	576.29	369.58	882.69	502.73	912.04	739.68	363.61	842.19
	GNSUM	1472.32	588.55	826.66	978.27	1187.72	804.46	1661.08	1900	709.25	1220.89
	Overd.	5553.6	8894.56	15372.59	20533.75	14428.89	20229.21	16188.8	11053.87	13010.14	7941.5
	TPC	328.84	**351.18**	624.54	**349.8**	**848.34**	**394.25**	**866.54**	718.61	**349.42**	**812.89**
	RoA	**325**	379.07	**565.23**	406.55	938.55	480.77	1102.62	792.51	429.96	1140.17
	AoR	330.25	516.84	722.37	697.17	2292.08	3014.88	4620.57	4864.23	19557.81	9084.39

Table 2.S.12: Metrics for the simulations varying the memory factor in the P-U scenario.

Metric	Method	Memory factor									
		0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
Mean	PiMLE	200.38	198.41	183.87	184.16	193.84	208.99	184.54	188.93	189.92	188.69
	MLE	187.15	185.69	186.13	174.12	185.72	197.34	171.46	174.41	175.24	184.13
	MoS	202.8	199.59	184.88	187.99	196.93	208.08	186.25	192.41	195.26	190.46
	KLN	185.89	183.9	185.11	174.06	183.65	196.95	171.31	174.6	174.94	183.32
	GNSUM	**109.24**	**121.2**	**129.49**	**136.1**	**115.65**	**128.93**	**100.25**	**122.08**	**102.15**	**123.11**
	Overd.	229.88	206.57	217	184.08	206.29	192.64	227.49	215.51	260.4	204.75
	TPC	191.6	189.37	190.6	178.35	189.88	201.9	174.87	178.93	179.26	188.33
	RoA	185.56	183.07	188.18	170.78	187.95	192.97	173.99	170.05	182.05	190.14
	AoR	191.11	190.96	185.13	200.12	257.22	294.86	300.35	254.59	288.47	333.94
SD	PiMLE	33.47	38.28	**33.87**	37.58	39.59	42.05	47.69	51.37	47.27	41.86
	MLE	20.47	32.15	53.92	24.54	40.93	45.17	34.38	**32.39**	33.46	39.21
	MoS	33.67	38.04	35.98	35.8	40.08	40.79	50.84	55.13	54.46	39.6
	KLN	**19.94**	31.23	52.72	23.36	**39.49**	46.7	**33.59**	32.86	**32.38**	**38.42**
	GNSUM	34.8	41.15	61.82	52.35	52.31	54.35	44.61	61.2	48.52	45.22
	Overd.	113.38	117.31	77.75	72.58	100.27	101.82	107.3	136.07	196.55	102.75
	TPC	19.95	32.29	54.19	24.67	40.86	45.22	34.5	32.72	33.37	38.89
	RoA	20.46	**30.92**	58.64	**22.25**	42.87	**37.49**	40.88	32.93	43.98	42.16
	AoR	29.4	33.82	39.18	39.72	62.23	112.49	149.06	72.38	85.67	125.79
MAE	PiMLE	100.38	98.41	83.87	84.16	93.84	108.99	84.54	88.93	89.92	88.69
	MLE	87.15	85.69	86.13	74.12	85.72	97.34	71.46	74.41	75.24	84.89
	MoS	102.8	99.59	84.88	87.99	96.93	108.08	86.42	92.41	95.26	90.46
	KLN	85.89	83.9	85.11	74.06	83.65	96.95	71.31	74.6	74.94	84.24
	GNSUM	**27.19**	**31.26**	**42.79**	**45.95**	**36.87**	**43.86**	**34.74**	**50.51**	**34.64**	**37.72**
	Overd.	129.88	112.66	117	88.14	112.16	101.36	127.49	118.71	164.6	108.87
	TPC	91.6	89.37	90.6	78.35	89.88	101.9	74.87	78.93	79.26	88.73
	RoA	85.56	83.07	88.18	70.78	87.95	92.97	73.99	70.05	82.05	90.14
	AoR	91.11	90.96	85.13	100.12	157.22	194.86	200.35	154.59	188.47	233.94
MSE	PiMLE	11141.16	11076.29	8123.6	8424.8	10295.77	13558.79	9308.43	10416.34	10209.22	9530.3
	MLE	7993.08	8325.01	10180.22	6065.39	8939.76	11413.36	6229.23	6533.19	6724.9	8538.23
	MoS	11645.57	11292.13	8435.47	8959.41	10921.32	13262.55	9895.1	11425.99	11892.85	9673.24
	KLN	7754.68	7964.86	9883.93	6003.38	8478.96	11471.06	6156.87	6590.54	6612.12	8344.43
	GNSUM	**1235.84**	**2058.41**	**4500.54**	**3906.6**	**2844.92**	**3642.67**	**1890.61**	**4045.58**	**2240.97**	**2476.53**
	Overd.	29080.73	24430.35	19431.68	12073.31	20849.67	18431.26	27192.45	30930.54	62427.75	21002.41
	TPC	8768.21	8978.24	10998	6717.58	9664.68	12326.71	6736.47	7246.4	7340.97	9238.49
	RoA	7718.12	7809.68	11041.69	5480.37	9480.66	9979.78	7061.55	5936.99	8569.7	9814.44
	AoR	9122.75	9359.96	8704.49	11522.81	28395.65	49990.85	61250.35	28875.36	42491.26	69759.95

Table 2.S.13: Metrics for the simulations varying the recall error in the P-S scenario.

Metric	Method	Number of Subpopulations									
		2	4	6	8	10	12	14	16	18	20
Mean	PiMLE	**90.6**	102.61	105.76	104.67	**99.47**	**99.65**	100.94	105.03	113.2	101.21
	MLE	85.28	**99.9**	100.98	**101.51**	96.92	97.75	101.05	104.4	111.9	100.28
	MoS	**90.6**	102.61	105.76	104.67	**99.47**	**99.65**	100.94	105.03	113.2	101.21
	KLN	85.28	**99.9**	100.98	**101.51**	96.92	97.75	101.05	104.4	111.9	100.28
	GNSUM	84.01	100.56	103.55	102.92	97.67	101.13	100.99	**103.94**	**110.39**	100.69
	Overd.	215.07	390.79	474.38	495.08	676.35	713.95	748.42	782.52	854.07	828.62
	TPC	89.49	103.96	105.03	105.86	100.65	102.18	104.73	108.13	116.31	104.21
	RoA	86.45	99.86	100.69	101.83	96.8	98.1	**100.89**	104.25	111.76	100.31
	AoR	86.22	99.89	**100.48**	102.07	97.51	98.34	100.91	104.14	111.88	**100.24**
SD	PiMLE	20.01	23.98	**24.16**	21.87	24.01	23.82	20.19	27.82	24.95	20.35
	MLE	18.81	22.93	24.33	21.2	23.17	22.59	19.81	28.49	24.87	**19.47**
	MoS	20.01	23.98	**24.16**	21.87	24.01	23.82	20.19	27.82	24.95	20.35
	KLN	18.81	22.93	24.33	21.2	23.17	22.59	19.81	28.49	24.87	**19.47**
	GNSUM	**18.16**	23.15	25.78	21.64	23.62	23.09	19.45	28.67	25.89	20.74
	Overd.	93.98	282.71	268.05	334.58	296.69	297.42	280.53	255.07	109.29	232.91
	TPC	19.01	23.18	24.91	21.27	23.25	22.75	**19.32**	28.39	24.91	19.85
	RoA	19.64	22.89	24.79	20.65	**22.83**	**22.42**	19.94	27.92	25.26	20.01
	AoR	19.32	**22.71**	24.75	**20.26**	22.87	22.59	20.09	**27.49**	**24.79**	20.34
MAE	PiMLE	**18.4**	18.09	**20.39**	17.61	17.83	19.9	17.36	**21.69**	17.26	15.2
	MLE	20.67	18.05	20.91	16.13	18.38	18.38	17.06	22.15	17.06	**14.81**
	MoS	**18.4**	18.09	**20.39**	17.61	17.83	19.9	17.36	**21.69**	17.26	15.2
	KLN	20.67	18.05	20.91	16.13	18.38	18.38	17.06	22.15	17.06	**14.81**
	GNSUM	19.82	18.02	21.64	16.85	18.05	18.62	**15.96**	22.78	17.86	15.17
	Overd.	116.37	292.32	374.38	397.64	576.35	613.95	648.42	682.52	754.07	728.62
	TPC	18.79	18.36	21.94	17.23	**17.1**	18.18	17.26	22.22	19.32	14.99
	RoA	20.93	18.11	20.91	15.62	18.07	**18.18**	16.86	21.84	17.64	15.44
	AoR	20.66	**17.83**	20.56	**15.52**	17.86	18.34	17.01	21.79	**17.01**	15.81
MSE	PiMLE	468.69	553.06	587.68	476.17	548.07	539.04	388.29	760.74	765.69	394.91
	MLE	552.75	499.72	**563.24**	429.06	519.31	490	373.86	790.64	729.31	**360.34**
	MoS	468.69	553.06	587.68	476.17	548.07	539.04	388.29	760.74	765.69	394.91
	KLN	552.75	499.72	**563.24**	429.06	519.31	490	373.86	790.64	729.31	**360.34**
	GNSUM	568.89	509.56	644.1	453.55	535.58	507.8	**360.2**	796.36	744.94	409.15
	Overd.	21632.08	160485.18	208418.28	262436.21	415805.29	460965.91	495212.28	527645.6	579972.56	582421.19
	TPC	**453.77**	525.91	614.87	463.95	513.76	496.32	376.9	831.49	855.5	391.9
	RoA	549.87	497.76	584.14	408.47	505.47	**480.99**	378.38	758.68	744.33	380.39
	AoR	544.33	**490.16**	582.24	**394.25**	**503.01**	487.63	384.07	**735.02**	**725.04**	392.89

Table 2.S.14: Metrics for the simulations varying the number of subpopulations in the W-U scenario.

Metric	Method	Number of Subpopulations				
		2	4	6	8	10
Mean	PiMLE	**99.78**	102.43	102.08	96.96	**103.67**
	MLE	91.61	100.44	**100.48**	96.5	104.22
	MoS	**99.78**	102.43	102.08	96.96	**103.67**
	KLN	91.61	100.44	**100.48**	96.5	104.22
	GNSUM	92.35	100.21	100.5	97.02	103.81
	Overd.	300.19	338.04	419.15	676.39	629.11
	TPC	96.48	104.12	104.65	**100.39**	108.36
	RoA	93.15	100.17	100.52	96.93	104.22
	AoR	93.08	**100.15**	100.64	96.69	103.67
SD	PiMLE	22.85	25.05	26.62	16.72	**26.75**
	MLE	20.51	23.13	26.1	16.17	27.04
	MoS	22.85	25.05	26.62	16.72	**26.75**
	KLN	20.51	23.13	26.1	16.17	27.04
	GNSUM	23	23.03	26.35	16.61	26.76
	Overd.	219.49	298.98	275.79	265.95	273.1
	TPC	20.55	23.45	26.05	**15.72**	27.77
	RoA	20.35	**22.7**	25.95	15.9	27.04
	AoR	**20.14**	22.88	**25.92**	16.27	26.75
MAE	PiMLE	18.74	20.94	21.44	13.78	**21.89**
	MLE	18.86	19.34	21.06	13.63	22.31
	MoS	18.74	20.94	21.44	13.78	**21.89**
	KLN	18.86	19.34	21.06	13.63	22.31
	GNSUM	19.57	19.56	21.23	13.61	22.3
	Overd.	201.38	240.9	319.15	576.39	529.11
	TPC	**16.8**	19.55	20.93	**12.07**	23.73
	RoA	18.06	**18.99**	20.96	13.33	22.31
	AoR	17.85	19.25	**20.91**	13.74	21.89
MSE	PiMLE	495.92	602.06	677.74	274.78	**693.03**
	MLE	469.81	508.22	647.35	260.59	712.5
	MoS	495.92	602.06	677.74	274.78	**693.03**
	KLN	469.81	508.22	647.35	260.59	712.5
	GNSUM	560.89	503.86	659.91	271.05	694.85
	Overd.	85842.11	141584.23	174109.27	399421.74	350814.42
	TPC	**413.61**	539.22	666.47	**234.84**	802.49
	RoA	440.44	**489.6**	639.82	249.45	712.5
	AoR	433.25	497.22	**638.77**	262.59	693.03

Table 2.S.15: Metrics for the simulations varying the number of subpopulations in the W-U scenario with disjoint subpopulations.

Metric	Method	Number of Subpopulations									
		2	4	6	8	10	12	14	16	18	20
Mean	PiMLE	-	-	-	-	-	125.84	-	-	110.37	**101.15**
	MLE	105.48	106.08	**101.3**	103.75	93.32	95.17	**101.1**	100.55	98.34	87
	MoS	-	-	-	-	-	125.84	-	-	110.37	**101.15**
	KLN	105.48	106.08	**101.3**	103.75	93.32	95.17	**101.1**	100.55	98.34	87
	GNSUM	90.26	**105.8**	103.3	107.08	91.2	106.43	95.26	**99.73**	103.42	90.03
	Overd.	165.18	122.07	131.49	122.47	185.06	112.24	123.78	121.96	90.92	132.48
	TPC	113.38	110.91	106.28	108.39	**96.6**	**99.49**	104.93	104.73	102.76	90.75
	RoA	99.58	107.15	101.89	**101.73**	93.29	97.21	101.75	100.65	**99.07**	88.14
	AoR	**99.74**	106.95	102.08	101.84	93.37	97.36	101.19	100.71	98.84	88.03
SD	PiMLE	-	-	-	-	-	37.8	-	-	24.06	22.68
	MLE	24.11	21.43	30.96	17.44	19.67	17.39	14.5	22.06	21.7	18.47
	MoS	-	-	-	-	-	37.8	-	-	24.06	22.68
	KLN	24.11	21.43	30.96	17.44	19.67	**17.39**	14.5	22.06	21.7	**18.47**
	GNSUM	23.33	30.33	30.51	18.97	**18.54**	18.61	20.34	25.72	**20.75**	21.42
	Overd.	139.85	51.47	58.84	70.94	161.67	54.25	54.93	83.03	36.8	79.89
	TPC	25.07	22.66	31.27	18.2	18.7	17.86	**14.39**	**22.05**	22.27	18.76
	RoA	**20.59**	20.41	**29.72**	16.09	20.17	18.38	16.5	23.41	21.09	19.21
	AoR	21.11	**20.18**	30.05	**15.85**	20.6	18.36	16.18	23.26	21.61	19.38
MAE	PiMLE	-	-	-	-	-	33.8	-	-	20.08	17.98
	MLE	19.97	**16.45**	25.56	13.85	17.52	14.61	**11.42**	**17.35**	17.06	17.99
	MoS	-	-	-	-	-	33.8	-	-	20.08	17.98
	KLN	19.97	**16.45**	25.56	13.85	17.52	14.61	**11.42**	**17.35**	17.06	17.99
	GNSUM	21.83	22.58	25.35	16.61	**15.48**	16.25	16.82	20.8	16.47	20.36
	Overd.	85.22	45.02	43.36	45.22	90.83	36.86	37.36	43.63	29.44	56.94
	TPC	22.63	19.66	25.41	15.47	15.69	**13.47**	11.76	17.53	17.61	**16.64**
	RoA	**16.59**	17.97	**24.03**	12.96	16.98	14.61	12.69	19.82	**16.19**	18.84
	AoR	17.07	17.52	24.23	**12.79**	17.12	14.79	12.38	19.43	16.62	18.94
MSE	PiMLE	-	-	-	-	-	2025.12	-	-	657.21	490
	MLE	582.4	473.09	912.33	303.09	412.03	310.69	**200.89**	**462.64**	450.25	492.99
	MoS	-	-	-	-	-	2025.12	-	-	657.21	490
	KLN	582.4	473.09	912.33	303.09	412.03	310.69	**200.89**	**462.64**	450.25	492.99
	GNSUM	611.88	907.37	895.24	392.07	403.91	370.37	415.54	628.3	**420.9**	535.09
	Overd.	22826.77	3004.1	4281.16	5285.25	32065.79	2945.26	3432.13	7031.41	1369.32	7117.77
	TPC	776.13	606.72	968.26	385.05	**343.61**	**303.24**	221.01	484.14	478.91	**420**
	RoA	**403.1**	446.78	**842.51**	248.78	431.59	328.77	261.6	521.21	423.25	491.44
	AoR	423.24	**435.38**	862	**241.91**	447.02	327.28	250.1	514.33	445.13	500.17

Table 2.S.16: Metrics for the simulations varying the number of subpopulations in the W-U scenario with small subpopulations.

Metric	Method	Number of Subpopulations									
		2	4	6	8	10	12	14	16	18	20
Mean	PiMLE	-	-	-	-	-	-	-	118.82	116.23	**102.15**
	MLE	106.19	107.79	102.15	**101.19**	100.85	101.77	108.9	100.42	**101.11**	93.55
	MoS	-	-	-	-	-	-	-	118.82	116.23	**102.15**
	KLN	106.19	107.79	102.15	**101.19**	100.85	101.77	108.9	100.42	**101.11**	93.55
	GNSUM	**100.7**	106.56	**98.8**	103.39	103.42	105.28	**101.66**	100.36	108.17	92.94
	Overd.	130.94	109.99	108.57	124.72	111.43	125.48	159.98	117.76	123.54	126.71
	TPC	111.47	113.74	107.78	105.87	104.97	106.59	113.28	105.16	105.68	97.5
	RoA	102.74	**105.02**	101.47	102.91	101.16	100.42	108.03	100.28	101.89	94.64
	AoR	102.98	105.23	101.37	103.23	**100.66**	**100.27**	107.87	**99.82**	102.08	94.36
SD	PiMLE	-	-	-	-	-	-	-	27.96	21.54	21.56
	MLE	27.2	23.35	19.52	21.65	**22.2**	22.57	26.58	22.95	15.76	19.81
	MoS	-	-	-	-	-	-	-	27.96	21.54	21.56
	KLN	27.2	23.35	19.52	21.65	**22.2**	22.57	26.58	22.95	15.76	19.81
	GNSUM	32.44	34.97	20.42	23.12	23.58	23	24.98	24.67	17.07	21.86
	Overd.	84	35.59	42.81	47.04	45.44	47.77	130.08	44.44	111.05	77.99
	TPC	28.32	24.25	20.84	21.93	22.93	22.73	26.75	23.36	**15.22**	**19.67**
	RoA	**20.02**	19.85	**19.27**	**20.64**	24.09	22.03	**24.81**	20.42	15.68	20.59
	AoR	20.44	**19.76**	19.5	20.65	24.11	**21.76**	25.01	**20.38**	16.5	20.48
MAE	PiMLE	-	-	-	-	-	-	-	27.78	22.09	16.2
	MLE	22.4	18.57	15.31	**17.43**	**16.24**	19.39	22.74	18.27	14.16	15.6
	MoS	-	-	-	-	-	-	-	27.78	22.09	16.2
	KLN	22.4	18.57	15.31	**17.43**	**16.24**	19.39	22.74	18.27	14.16	15.6
	GNSUM	26.54	26.01	**14.32**	18.79	19.44	**17.51**	**20.83**	19.69	15.57	17.95
	Overd.	55.98	28.57	26.64	33.25	35.62	37.35	77.97	38.61	53.67	52.68
	TPC	24.22	20.51	17.38	18.49	17.27	19.6	23.76	19.5	14.81	**15.24**
	RoA	**16.16**	**15.43**	15.24	18.04	17.77	18.75	21.01	16.28	**13.86**	15.97
	AoR	16.62	15.46	15.36	17.92	17.55	18.47	20.92	**16.21**	14.62	15.9
MSE	PiMLE	-	-	-	-	-	-	-	1096.61	704.36	446.03
	MLE	741.01	578.9	366.53	446.53	**468.79**	487.25	750.25	500.69	237.13	414.5
	MoS	-	-	-	-	-	-	-	1096.61	704.36	446.03
	KLN	741.01	578.9	366.53	446.53	**468.79**	487.25	750.25	500.69	237.13	414.5
	GNSUM	999.98	1204.6	397.54	519.48	539.75	530.61	**595.57**	578.28	343.62	503.7
	Overd.	7660.99	1302.75	1814.72	2712.77	2092.4	2816.6	19671.72	2191.23	12269.91	6491.31
	TPC	893.5	747.42	472.92	491.19	524.08	534.21	856.21	545.09	252.38	**373.82**
	RoA	**388.29**	399.65	**354.93**	**413.15**	552.84	461.4	649.02	396.04	**237.06**	431.44
	AoR	405.92	**398.32**	363.24	415.37	552.6	**450.02**	656.08	**394.73**	263.01	430.05

Table 2.S.17: Metrics for the simulations varying the number of subpopulations in the W-U scenario with disjoint and small subpopulations.

Metric	Method	Number of Subpopulations									
		2	4	6	8	10	12	14	16	18	20
Mean	PiMLE	127.87	131.75	128.8	118.91	104.24	98.47	86.18	116.7	104.92	86.04
	MLE	**104.96**	120.43	124.37	117.17	98.9	98.5	84.54	**112.73**	105.01	87.21
	MoS	127.87	131.75	128.8	118.91	104.24	98.47	86.18	116.7	104.92	86.04
	KLN	**104.96**	120.43	124.37	117.17	98.9	98.5	84.54	**112.73**	105.01	87.21
	GNSUM	111.16	125.85	124.33	**117.01**	106.5	97.13	85.09	113.93	**104.5**	85.07
	Overd.	286.15	450.02	524.51	530.21	653.77	722.68	753.97	777.54	809.88	845.64
	TPC	108.86	124.15	129.12	121.65	103.14	102.08	**88.89**	116.72	109.12	**90.96**
	RoA	105.28	**118.33**	**124.07**	117.47	98.85	98.13	84.49	113.15	105.3	87.52
	AoR	105.88	118.78	125.71	117.86	**99.98**	**98.58**	84.41	113.22	105.73	87.94
SD	PiMLE	73.52	81.81	54.21	58.82	57.22	38.84	**37.99**	81.55	**53.11**	35.53
	MLE	59.24	70.65	52.18	58.83	**49.99**	36.33	39.37	**76.04**	53.22	35.71
	MoS	73.52	81.81	54.21	58.82	57.22	38.84	**37.99**	81.55	**53.11**	35.53
	KLN	59.24	70.65	52.18	58.83	**49.99**	36.33	39.37	**76.04**	53.22	35.71
	GNSUM	65.68	73.9	54.16	**56.29**	53.22	36.75	40.7	76.8	54.41	**35.22**
	Overd.	142.51	200.68	234.49	234.02	193.03	181.28	165.76	139.42	105.64	60.66
	TPC	58.87	70.5	52.64	59.14	50.26	**35.95**	39.43	76.47	53.58	35.41
	RoA	**56.72**	**68.46**	**51.71**	58.45	50.19	36.26	39.22	76.37	53.93	35.76
	AoR	57.7	69.89	53.16	58.11	52.22	37.62	38.96	76.25	54.59	36.54
MAE	PiMLE	56.4	68.3	53.24	47.03	40.91	29.64	**32.72**	62.91	39.08	32.76
	MLE	40.8	57.77	49.6	**46.69**	38.28	28.48	33.8	58.95	38.2	32.67
	MoS	56.4	68.3	53.24	47.03	40.91	29.64	**32.72**	62.91	39.08	32.76
	KLN	40.8	57.77	49.6	**46.69**	38.28	28.48	33.8	58.95	38.2	32.67
	GNSUM	44.48	59.79	50.9	47.1	37.7	**28.05**	35.49	59.51	38.34	32.99
	Overd.	186.15	350.02	426.33	430.21	553.77	622.68	653.97	677.54	709.88	745.64
	TPC	**40.44**	57.78	51.97	47.74	**36.94**	28.2	32.74	**58.48**	**37.8**	**31.8**
	RoA	40.65	**56.24**	**49.36**	47.2	38.23	28.41	33.82	59.3	38.27	32.61
	AoR	40.91	56.87	51	47.72	39.15	28.78	33.64	59.57	38.71	33.47
MSE	PiMLE	5912.09	7366.01	3620.49	3644.07	3128.37	1435.44	**1562.11**	6596.88	**2703.76**	1394.23
	MLE	3358.29	5159.76	3180.75	3582.41	**2375.3**	1256.29	1711.29	**5655.17**	2715.46	1374.74
	MoS	5912.09	7366.01	3620.49	3644.07	3128.37	1435.44	**1562.11**	6596.88	**2703.76**	1394.23
	KLN	3358.29	5159.76	3180.75	3582.41	**2375.3**	1256.29	1711.29	**5655.17**	2715.46	1374.74
	GNSUM	4222.72	5855.92	3378.39	**3299.86**	2732.59	1291.28	1795.87	5796.63	2833.07	1400.95
	Overd.	53946.11	160773.21	232439.68	237105.88	342057.09	418947.86	453778.66	477532.97	514531.22	559470.56
	TPC	3370.83	5304.83	3480.41	3791.49	2409.6	**1232.06**	1600.04	5834.45	2810.03	**1272.67**
	RoA	**3084.57**	**4787.77**	**3119.48**	3550.6	2394.59	1252.48	1702.19	5713.86	2791.44	1370.48
	AoR	3197.16	4992.98	3345.57	3527.18	2590.36	1346.73	1684.9	5697.68	2863.78	1414.01

Table 2.S.18: Metrics for the simulations varying the number of subpopulations in the W-S scenario

Metric	Method	Number of Subpopulations				
		2	4	6	8	10
Mean	PiMLE	115.1	104.76	78.48	103.28	104.28
	MLE	97.5	**101.53**	77.62	101.37	**104.12**
	MoS	115.1	104.76	78.48	103.28	104.28
	KLN	97.5	**101.53**	77.62	101.37	**104.12**
	GNSUM	105.1	102.18	77.57	107.04	104.96
	Overd.	272.51	296.45	348.83	619.14	606.8
	TPC	101.57	105.88	**81.55**	105.96	108.09
	RoA	99.02	101.77	77.74	**101.35**	104.12
	AoR	**100.12**	101.93	78.06	101.44	104.28
SD	PiMLE	80.62	42.83	33.21	46.76	**38.1**
	MLE	61.3	42.14	31.93	45.71	38.5
	MoS	80.62	42.83	33.21	46.76	**38.1**
	KLN	61.3	42.14	31.93	45.71	38.5
	GNSUM	67.76	**39.93**	32.41	48.28	38.34
	Overd.	161.24	145.23	155.26	208.73	269.23
	TPC	**61.09**	42.88	**31.83**	**45.43**	38.76
	RoA	63.63	41.21	32.42	45.86	38.5
	AoR	63.85	41.32	32.71	45.8	38.1
MAE	PiMLE	54.77	33.03	31.6	36.07	29.21
	MLE	40.85	32.53	31.32	35.63	**29.11**
	MoS	54.77	33.03	31.6	36.07	29.21
	KLN	40.85	32.53	31.32	35.63	**29.11**
	GNSUM	46.11	32.26	32.26	36.12	29.14
	Overd.	180.84	196.45	251.43	519.14	509
	TPC	**39.91**	33.33	**29.54**	**34.23**	29.63
	RoA	41.58	**32.09**	31.98	35.85	29.11
	AoR	41.67	32.42	31.98	35.37	29.21
MSE	PiMLE	6401.76	1765.73	1510.74	2087.6	**1397.66**
	MLE	3575.92	1689.41	1469.33	**1986.46**	1424.95
	MoS	6401.76	1765.73	1510.74	2087.6	**1397.66**
	KLN	3575.92	1689.41	1469.33	**1986.46**	1424.95
	GNSUM	4387.62	**1519.53**	1501.13	2263.88	1421.24
	Overd.	54458.01	58627.98	84818.53	310894.77	325709.59
	TPC	**3547.35**	1781.56	**1303.01**	1996.2	1492.54
	RoA	3847.01	1616.32	1493.87	2000.06	1424.95
	AoR	3872.49	1625.49	1497.52	1994.53	1397.66

Table 2.S.19: Metrics for the simulations varying the number of subpopulations in the W-S scenario with disjoint subpopulations.

Metric	Method	Number of Subpopulations									
		2	4	6	8	10	12	14	16	18	20
Mean	PiMLE	-	-	-	-	-	154.35	132.45	**99.38**	111.3	131.05
	MLE	92.48	96.38	**102.14**	121.42	101.66	108.52	**105.5**	90.22	93.96	114.95
	MoS	-	-	-	-	-	154.35	132.45	**99.38**	111.3	131.05
	KLN	92.48	96.38	**102.14**	121.42	101.66	108.52	**105.5**	90.22	93.96	114.95
	GNSUM	82.34	89.42	86.89	135.08	107.85	111.95	115.85	84.14	**101.48**	**108.79**
	Overd.	153.23	134.63	134.69	171.19	155.8	125.85	144.9	120.49	149.29	154.97
	TPC	**98.63**	**101.72**	106.98	125.8	105.92	112.57	109.82	94.76	97.54	118.95
	RoA	83.19	96.79	102.25	**121.33**	98.12	105	107.22	91.74	93.78	115.44
	AoR	83.9	97.39	102.94	122.5	**98.82**	**104.8**	106.79	91.54	93.97	117.17
SD	PiMLE	-	-	-	-	-	95.23	76.05	52.71	75.66	90.21
	MLE	45.34	43.35	45.79	75.32	60.45	60.75	**54.21**	45.4	54.86	75.73
	MoS	-	-	-	-	-	95.23	76.05	52.71	75.66	90.21
	KLN	45.34	43.35	45.79	75.32	60.45	60.75	**54.21**	45.4	54.86	75.73
	GNSUM	46.15	**40.58**	**43.04**	94.47	62.85	66.31	59.72	45.14	60.24	71.82
	Overd.	76.81	56.77	52.91	95.41	71.4	59.26	56.49	51	78.81	**43.24**
	TPC	46.45	43.19	46.61	75.4	60.75	60.64	54.64	46.11	**53.8**	75.39
	RoA	**36.88**	42.63	48.3	**74.23**	**57.61**	56.46	56.41	44.91	55.97	79.31
	AoR	37.58	43.07	48.27	76.46	57.69	**56.16**	56.56	**44.46**	55.94	81.43
MAE	PiMLE	-	-	-	-	-	76.14	58.24	41.28	55.54	65.34
	MLE	39.26	34.46	35.03	57.79	44.36	49.79	41.82	37.92	45.06	56.53
	MoS	-	-	-	-	-	76.14	58.24	41.28	55.54	65.34
	KLN	39.26	34.46	35.03	57.79	44.36	49.79	41.82	37.92	45.06	56.53
	GNSUM	38.33	**29.48**	37.94	65.73	42.72	51.74	46.03	40.52	47.71	**55.69**
	Overd.	67.1	52.53	51.18	80.06	67.03	**46.52**	61.9	41.5	61.29	56.1
	TPC	39.01	32.13	**34.57**	58.12	43.18	49.6	**41.44**	37.58	**42.72**	56.1
	RoA	**34.07**	32.82	37.15	**56.22**	**42.05**	47.44	41.83	37.51	45.81	58.36
	AoR	34.42	33.08	36.45	57.2	42.41	47.34	41.85	**37.03**	45.76	59.21
MSE	PiMLE	-	-	-	-	-	11569.97	6547.67	2639.77	5565.25	8694.86
	MLE	2009.32	1798.75	1996.91	5847.97	3473.68	3578.32	**2821.65**	2054.12	2895.57	5671.38
	MoS	-	-	-	-	-	11569.97	6547.67	2639.77	5565.25	8694.86
	KLN	2009.32	1798.75	1996.91	5847.97	3473.68	3578.32	**2821.65**	2054.12	2895.57	5671.38
	GNSUM	2335.12	**1676.58**	**1931.5**	9708.05	3814.54	4320.26	3638.93	2187.58	3450.03	4977.7
	Overd.	8437.82	4260.26	3863.24	13716.66	7957.4	4003.95	5047.25	2891.08	8329.24	**4797.9**
	TPC	2051.8	1774.81	2112.97	6066.73	3541.33	3651.3	2932.43	2047.59	**2755.46**	5758.34
	RoA	**1574.48**	1736.33	2221.07	**5689.65**	**3156.73**	3053.88	3074.75	1984.46	3014.68	6213.84
	AoR	1600.65	1768.87	2221.87	6060.09	3163.06	**3019.24**	3085.19	**1949.65**	3009.65	6593.83

Table 2.S.20: Metrics for the simulations varying the number of subpopulations in the W-S scenario with small subpopulations

Metric	Method	Number of Subpopulations									
		2	4	6	8	10	12	14	16	18	20
Mean	PiMLE	-	-	-	-	-	-	-	112.48	149.09	-
	MLE	**102.8**	**98.17**	107.53	103	86.94	97.68	110.45	115.61	**123.36**	106.82
	MoS	-	-	-	-	-	-	-	112.48	149.09	-
	KLN	**102.8**	**98.17**	107.53	103	86.94	97.68	110.45	115.61	**123.36**	106.82
	GNSUM	107.35	84.22	86.42	107.11	**96**	103.26	124	**104.69**	138.83	**97.63**
	Overd.	139.66	133.84	152.94	144.11	131.96	149.64	144.72	135.76	131.62	155.76
	TPC	108.63	103.77	112.67	107.27	90.75	**101.67**	114.8	119.75	127.73	111.08
	RoA	106.22	93.03	**107.52**	**101.96**	86.9	95.96	**109.24**	116.29	123.9	103.81
	AoR	106.33	93.49	108.45	102.2	88.01	96.37	110.3	117.06	123.37	104.25
SD	PiMLE	-	-	-	-	-	-	-	49.86	103.21	-
	MLE	40.86	43.69	54.46	45.62	36.28	55.32	42.15	55.25	75.27	44.21
	MoS	-	-	-	-	-	-	-	49.86	103.21	-
	KLN	40.86	43.69	54.46	45.62	36.28	55.32	42.15	55.25	75.27	44.21
	GNSUM	45.09	**37.84**	**44.35**	45	41.72	59.11	46.18	**49.22**	87.73	**39.02**
	Overd.	60.84	76.07	78.94	66.45	37	**53.52**	57.3	53.71	**49.65**	67.86
	TPC	**38.52**	45.34	55.17	46.45	36.16	55.43	41.71	55.74	75.27	44.14
	RoA	40.85	38.8	55.49	**42.51**	**34.96**	54.49	**39.32**	53.62	75.89	42.49
	AoR	42.09	38.94	56.62	43.04	36.28	55.45	39.72	54.56	75.33	42.38
MAE	PiMLE	-	-	-	-	-	-	-	38.23	68.15	-
	MLE	34.83	36.65	41.34	37.36	31.25	47.34	31.2	43.72	49.57	36.08
	MoS	-	-	-	-	-	-	-	38.23	68.15	-
	KLN	34.83	36.65	41.34	37.36	31.25	47.34	31.2	43.72	49.57	36.08
	GNSUM	39.99	35.16	**39.75**	38.38	34.61	48.63	40.04	**37.24**	58.81	**30.65**
	Overd.	48.54	54.24	60.79	60.09	37.7	54.63	50.95	52.94	**47.01**	63.61
	TPC	**33.22**	38	41.43	37.94	**29.51**	46.7	31.43	44.61	50.21	36.63
	RoA	33.8	**32.22**	39.8	**36.19**	31.28	46.63	**30.32**	43.05	50.45	33.74
	AoR	34.21	32.36	40.66	36.54	31.83	**46.62**	30.8	43.75	50.03	33.92
MSE	PiMLE	-	-	-	-	-	-	-	2517.64	12528.68	-
	MLE	1594.25	1816.42	2874.42	1985.8	1420.77	2912.94	1796.72	3144.1	5927.29	1903.62
	MoS	-	-	-	-	-	-	-	2517.64	12528.68	-
	KLN	1594.25	1816.42	2874.42	1985.8	1420.77	2912.94	1796.72	3144.1	5927.29	1903.62
	GNSUM	1985.89	1609.23	**2052.83**	1973.99	1669.8	3330.19	2602.03	**2323.76**	8820.19	**1452.34**
	Overd.	5089.48	6642.59	8722.82	6140.5	2321.8	5184.61	5118.75	4019	**3342.09**	7484.21
	TPC	**1484.04**	1967.57	3052.24	2102.55	**1327.36**	2921.57	1871.7	3341.74	6151.49	1973.79
	RoA	1623.76	**1478.84**	2981.37	**1720.82**	1332.34	**2837.01**	**1554.07**	2996.82	6042.71	1729.92
	AoR	1722.89	1482.69	3117.44	1764.32	1394.34	2933.6	1604.49	3118.54	5937.56	1724.25

Table 2.S.21: Metrics for the simulations varying the number of disjoint subpopulations in the W-S scenario with disjoint and small subpopulations.

Metric	Method	Number of Subpopulations									
		2	4	6	8	10	12	14	16	18	20
Mean	PiMLE	-	**100.68**	89.18	101.57	**100.72**	103.8	97.39	89.59	98.84	**99.92**
	MLE	102.09	96.33	87.36	97.67	93.78	101.22	96.17	88.11	95.74	99.18
	MoS	-	**100.68**	89.18	101.57	**100.72**	103.8	97.39	89.59	98.84	**99.92**
	KLN	102.09	96.33	87.36	97.67	93.78	101.22	96.17	88.11	95.74	99.18
	GNSUM	109.85	106.56	**94.57**	105.8	101.08	102.11	105.04	**103.65**	115.96	97.1
	Overd.	301.29	428.82	516.11	591.77	643.78	685.45	822.84	799.95	794.64	830.13
	TPC	106.82	101.05	91.58	101.71	97.49	105.64	**99.94**	92.04	**99.39**	103.68
	RoA	**98.88**	94.36	88.62	98.64	92.84	**101.14**	97.25	89.04	94.85	99.91
	AoR	101.3	92.7	87.71	**100.64**	96.93	101.62	95.97	88.36	98.1	99.53
SD	PiMLE	-	24.32	26.01	28.4	22.58	23.81	23.39	20.34	24.08	19.39
	MLE	**16.45**	18.74	**22.19**	23.61	16.61	19.19	21.49	18.46	15.62	19.16
	MoS	-	24.32	26.01	28.4	22.58	23.81	23.39	20.34	24.08	19.39
	KLN	**16.45**	18.74	**22.19**	23.61	16.61	19.19	21.49	18.46	15.62	19.16
	GNSUM	34.61	25.33	34.76	39.3	34.24	26.7	30.19	45.78	41.88	24.85
	Overd.	269.91	287.55	328.88	319.02	316.34	289.58	177.85	230.66	213.98	121.15
	TPC	16.92	18.82	22.52	24.22	**15.95**	19.4	**21.23**	**18.37**	15.48	19.48
	RoA	17.29	**17.64**	22.36	**23.23**	16.47	**18.69**	22.11	18.65	**15.03**	**18.79**
	AoR	18.5	21.9	25.11	26.82	20.53	21.65	25.22	19.56	23.56	18.84
MAE	PiMLE	-	17.17	23.27	22.53	17.24	16.66	18.23	18.24	18.99	**14.49**
	MLE	**13.14**	**14.51**	21.11	19.1	14.53	14.16	17.95	19.01	13.14	15.3
	MoS	-	17.17	23.27	22.53	17.24	16.66	18.23	18.24	18.99	**14.49**
	KLN	**13.14**	**14.51**	21.11	19.1	14.53	14.16	17.95	19.01	13.14	15.3
	GNSUM	24.34	19.17	28.88	30.76	25.25	21.47	23.45	33.33	32.25	19.6
	Overd.	201.29	331.45	416.11	491.77	543.78	585.45	722.84	699.95	694.64	730.13
	TPC	14.49	14.59	**20.22**	19.19	**12.81**	14.23	**17.56**	**16.51**	**12.08**	15.41
	RoA	13.44	14.66	20.26	**18.64**	14.62	**13.99**	18.23	18.61	12.82	15.2
	AoR	13.23	18.06	22.11	21.87	15.58	15.81	19.71	17.48	18.63	14.74
MSE	PiMLE	-	562.22	759.91	768.76	484.75	552.77	526.39	501.43	552.4	357.14
	MLE	**261.48**	347.13	627.75	534.89	300.86	351.22	453.48	465.06	249.95	349.45
	MoS	-	562.22	759.91	768.76	484.75	552.77	526.39	501.43	552.4	357.14
	KLN	**261.48**	347.13	627.75	534.89	300.86	351.22	453.48	465.06	249.95	349.45
	GNSUM	1235.12	652.42	1177.61	1500.87	1114.76	681.72	891.1	2004.67	1920.99	594.81
	Overd.	109726.2	186676.67	275897.63	338524.54	390766.7	422411.25	552551.15	540478.68	526025.84	547026.24
	TPC	318.59	337.5	**552.83**	560.2	**248.08**	389.47	**428.36**	**383.93**	**228.14**	373.94
	RoA	285.16	**327.43**	604.33	**514.32**	309.02	**333.18**	472.09	450.49	241.21	**335.55**
	AoR	326.83	508.88	750.34	683.77	409.77	447.9	620.43	498.94	530.99	337.4

Table 2.S.22: Metrics for the simulations varying the number of subpopulations in the P-U scenario.

Metric	Method	Number of Subpopulations				
		2	4	6	8	10
Mean	PiMLE	-	108.87	101.74	85.16	90.75
	MLE	93.2	102.9	98.96	86.35	87.79
	MoS	-	108.87	101.74	85.16	90.75
	KLN	93.2	102.9	98.96	86.35	87.79
	GNSUM	**101.55**	113.61	95.61	**95.81**	87.37
	Overd.	269.9	292.3	382.4	598.85	622.24
	TPC	97.61	106.63	103.53	89.92	**91.71**
	RoA	89.23	**102.17**	98.77	87	87.79
	AoR	89.69	103.31	**100.42**	84.69	90.75
SD	PiMLE	-	27.7	20.73	23.31	29.06
	MLE	17.97	21.83	19.83	**14.99**	22.99
	MoS	-	27.7	20.73	23.31	29.06
	KLN	17.97	21.83	19.83	**14.99**	22.99
	GNSUM	26.96	38.95	22.86	29.35	28.17
	Overd.	185.09	240.04	269.25	277.92	305.96
	TPC	18.47	21.7	19.86	15.12	**22.72**
	RoA	**15.55**	**21.62**	**19.3**	15.38	22.99
	AoR	18.03	26.68	20.86	23.02	29.06
MAE	PiMLE	-	23.49	16.01	20.33	27.35
	MLE	14.82	**16.83**	16.29	14.88	23.06
	MoS	-	23.49	16.01	20.33	27.35
	KLN	14.82	**16.83**	16.29	14.88	23.06
	GNSUM	23.32	27.81	18.87	22.04	25.62
	Overd.	170.11	192.38	282.99	498.85	522.24
	TPC	**13.77**	17.21	15.93	**13.05**	**21.3**
	RoA	14.59	17.06	**15.59**	14.64	23.06
	AoR	15.7	21.79	16.78	20.43	27.35
MSE	PiMLE	-	807.41	411.2	736.53	887.71
	MLE	352.95	461.06	374.53	399.83	651.26
	MoS	-	807.41	411.2	736.53	887.71
	KLN	352.95	461.06	374.53	399.83	651.26
	GNSUM	692.85	1626.33	515.95	835.99	913.51
	Overd.	61412.86	91716.33	148623.92	322221.83	361664.69
	TPC	**329.91**	491.17	387.24	**318.64**	**559.23**
	RoA	345.74	**448.64**	**355.23**	393.75	651.26
	AoR	415.36	687.41	413.67	738.06	887.71

Table 2.S.23: Metrics for the simulations varying the number of subpopulations in the P-U scenario with disjoint subpopulations.

Metric	Method	Number of Subpopulations									
		2	4	6	8	10	12	14	16	18	20
Mean	PiMLE	-	-	-	-	-	-	-	-	-	-
	MLE	116.92	91.24	88.36	91.77	87.11	102.32	**100.3**	108.99	97.74	94.21
	MoS	-	-	-	-	-	-	-	-	-	-
	KLN	116.92	91.24	88.36	91.77	87.11	102.32	**100.3**	108.99	97.74	94.21
	GNSUM	120.2	124.21	111.07	89.48	89.96	111.74	109.65	120.08	112.54	**99.97**
	Overd.	137.23	137.03	130.87	131.78	139.13	124.89	142.02	**98.06**	141.17	151.28
	TPC	126.39	**96.18**	**92.83**	96.43	90.9	106.78	104.92	113.2	101.89	98.65
	RoA	**92.07**	93.34	91.67	**97.13**	92.07	**99.81**	100.43	102.44	**99.66**	96.07
	AoR	90.38	94.23	90.09	96.31	**92.78**	97.9	103.8	107.36	96.05	94.33
SD	PiMLE	-	-	-	-	-	-	-	-	-	-
	MLE	33.88	19.92	**17.11**	**18.34**	**20.56**	22.79	24.78	27.74	19.09	22.29
	MoS	-	-	-	-	-	-	-	-	-	-
	KLN	33.88	19.92	**17.11**	**18.34**	20.56	22.79	24.78	27.74	19.09	22.29
	GNSUM	54.86	76.35	38.24	25.35	25.74	37.2	39.04	28.76	30.03	31.63
	Overd.	78.32	109.75	62.17	84.18	71.02	77.73	98.95	50.69	74.56	99.45
	TPC	34.67	20.25	17.98	18.47	21.38	23.98	24.91	27.77	18.67	21.89
	RoA	**17.76**	**18.85**	19.09	19.07	21.66	**19.12**	**22.72**	**26.15**	**18.3**	22.51
	AoR	18.3	21.61	18.67	22.29	24.93	24.72	27.09	31.88	19.78	**21.41**
MAE	PiMLE	-	-	-	-	-	-	-	-	-	-
	MLE	26.36	18.42	17	15.54	20.36	15.78	18.63	22.76	16.15	18.72
	MoS	-	-	-	-	-	-	-	-	-	-
	KLN	26.36	18.42	17	15.54	20.36	15.78	18.63	22.76	16.15	18.72
	GNSUM	41.83	46.5	30.75	23.93	23.86	24.21	33.51	27.6	25.04	23.93
	Overd.	49.27	59.82	48.54	52.29	50.02	42.3	67.84	38.85	56.87	69.7
	TPC	31.66	16.13	**15.47**	**13.68**	**18.97**	15.91	18.84	23.45	16.24	**17.67**
	RoA	**15.16**	**15.97**	16.07	14.59	19.38	**14.42**	**17.96**	**20.82**	15.53	18.49
	AoR	16.89	18.24	15.94	16.85	20.86	19.31	23.24	24.43	**15.29**	18.67
MSE	PiMLE	-	-	-	-	-	-	-	-	-	-
	MLE	1376.63	453.88	413.43	387.41	567.85	498.76	583.27	812.04	351.46	505.61
	MoS	-	-	-	-	-	-	-	-	-	-
	KLN	1376.63	453.88	413.43	387.41	567.85	498.76	583.27	812.04	351.46	505.61
	GNSUM	3267.29	6123.1	1511.56	721.23	730.43	1452.55	1540.91	1188.88	1014	950.35
	Overd.	7212.86	12814.88	4624.65	7741.96	6321.89	6359.54	11067.32	2444.55	6975.36	12025.01
	TPC	1838.22	404.3	**358.69**	**336.75**	517.03	592.29	613.81	906.63	334.8	**457.1**
	RoA	**362.66**	**381.79**	415.6	353.59	**508.44**	**347.28**	**490.62**	**655.52**	**318.11**	496.73
	AoR	410.8	477.1	429.33	485.57	642.39	584.93	711.37	1019.46	387.38	467.74

Table 2.S.24: Metrics for the simulations varying the number of subpopulations in the P-U scenario with small subpopulations.

Metric	Method	Number of Subpopulations									
		2	4	6	8	10	12	14	16	18	20
Mean	PiMLE	-	-	-	-	-	-	-	-	-	-
	MLE	**103.81**	**102.41**	94.95	86.02	93.82	101.49	92.17	**99.09**	90.4	96.91
	MoS	-	-	-	-	-	-	-	-	-	-
	KLN	**103.81**	**102.41**	94.95	86.02	93.82	101.49	92.17	**99.09**	90.4	96.91
	GNSUM	117.1	123.02	115.59	89.25	94.57	115.09	119.33	105.14	110.17	104.89
	Overd.	152.7	137.54	109.31	**97.92**	144.08	105.76	174.86	146.35	**97.76**	84.55
	TPC	110.44	106.78	**100.18**	90.4	**98.2**	106.17	**96.14**	103.15	95.25	**101.02**
	RoA	88.46	104.61	95.92	93.2	96.75	**100.33**	93.03	94.74	89.96	96.59
	AoR	88.36	108.42	98.17	94.12	94.05	101.12	91.73	96.34	89.8	97.51
SD	PiMLE	-	-	-	-	-	-	-	-	-	-
	MLE	25.08	23.71	19.48	**18.76**	22.37	18.44	17.42	20.83	18.74	22.63
	MoS	-	-	-	-	-	-	-	-	-	-
	KLN	25.08	23.71	19.48	**18.76**	22.37	18.44	17.42	20.83	18.74	22.63
	GNSUM	40.55	47.66	28.32	33.86	34.74	31.68	38.46	27.44	36.7	38.63
	Overd.	128.64	138.72	59.32	22.12	113.92	33.21	120.96	124.11	37.63	35.21
	TPC	25.09	24.44	20.13	19.09	22.57	19.04	17.32	21.28	18.59	22.96
	RoA	**24.13**	**23.41**	**16.76**	20.04	22.13	**17.27**	**17.15**	**19.42**	17.24	**22.35**
	AoR	24.23	25.29	20.26	23.34	**21.4**	18.05	18.49	19.79	**15.45**	27.26
MAE	PiMLE	-	-	-	-	-	-	-	-	-	-
	MLE	21.92	**17.41**	16.31	19.79	17.77	15.4	15.47	17.96	17.51	18.33
	MoS	-	-	-	-	-	-	-	-	-	-
	KLN	21.92	**17.41**	16.31	19.79	17.77	15.4	15.47	17.96	17.51	18.33
	GNSUM	33.2	38.45	25.38	29.69	27.66	24.16	33.73	21.29	27.51	29.17
	Overd.	67.79	68.49	37.3	17.87	62.62	24.02	88.98	65.97	25.49	31.04
	TPC	23.22	18.23	16.28	17.28	18.34	16.92	**13.34**	17.93	16	**17.94**
	RoA	**20.83**	18.98	**14.06**	**16.91**	18.59	**14.01**	13.78	16.51	16.9	17.99
	AoR	21.57	20.41	15.48	20.43	**17.39**	14.98	15.68	**16.17**	**14.08**	20.23
MSE	PiMLE	-	-	-	-	-	-	-	-	-	-
	MLE	**612.17**	**539.94**	385.96	529.66	513.61	325.23	349.64	413.15	425.76	496.09
	MoS	-	-	-	-	-	-	-	-	-	-
	KLN	**612.17**	**539.94**	385.96	529.66	513.61	325.23	349.64	413.15	425.76	496.09
	GNSUM	1854.48	2687.43	1005	1205	1176.01	1181.42	1779.27	741.67	1382.85	1441.62
	Overd.	18497.89	19690.53	3429.96	469.14	14272.64	1080.64	19503.24	16781.26	1349.92	1416.13
	TPC	707.2	613.27	384.96	438.22	487.14	382.61	**299.91**	440.19	350.99	501.82
	RoA	686.4	541.89	**283.35**	**427.53**	475.76	**283.55**	328.17	385.72	383.18	**486.24**
	AoR	693.45	678.62	393.28	552.23	**470.6**	310.9	393.23	**385.51**	**330.92**	712.02

Table 2.S.25: Metrics for the simulations varying the number of subpopulations in the P-U scenario with disjoint and small subpopulations.

Metric	Method	Number of Subpopulations									
		2	4	6	8	10	12	14	16	18	20
Mean	PiMLE	-	202.11	198.1	186.16	179.36	185.46	178.14	188.3	184.68	196.19
	MLE	209.76	190.33	188.94	185.33	179.53	184.32	172.08	179.5	180.96	188.12
	MoS	-	202.11	198.1	186.16	179.36	185.46	178.14	188.3	184.68	196.19
	KLN	209.76	190.33	188.94	185.33	179.53	184.32	172.08	179.5	180.96	188.12
	GNSUM	**114.82**	**131.3**	**106.05**	**116.87**	**125.33**	**114.69**	**109.96**	**93.49**	**120.65**	**122.67**
	Overd.	327.7	328.14	492.67	531.49	582.18	571.99	668.86	809.23	737.82	864.59
	TPC	215.15	194.65	192.21	190.2	183.72	188.82	175.77	182.72	185.38	191.72
	RoA	199.71	188.09	188.7	188.8	179.71	184.93	174.9	180.58	179.62	190.64
	AoR	203.14	189.21	191.89	184.95	174.46	183.29	177.61	184.82	181.6	194.49
SD	PiMLE	-	38.71	31.97	30.44	36.06	31.33	27.1	20.31	26.19	34.46
	MLE	32.49	28.44	28.51	27.67	40.75	**29.61**	**17.77**	20.02	29.96	33.39
	MoS	-	38.71	31.97	30.44	36.06	31.33	27.1	20.31	26.19	34.46
	KLN	32.49	28.44	28.51	27.67	40.75	**29.61**	**17.77**	20.02	29.96	33.39
	GNSUM	35.17	42.41	48.06	49.7	81.11	53.29	55.18	32.01	67.94	42.06
	Overd.	178.29	189.13	296.22	238.5	294.89	303.64	255.97	205.24	202.89	143.73
	TPC	32.32	28.78	**28.01**	28.6	40.35	29.62	18.35	**19.85**	30.29	33.41
	RoA	**31.48**	**25.67**	29.56	27.04	39.93	29.86	18.09	19.89	28.25	33.29
	AoR	39.13	34.35	30.42	**26.6**	**33.45**	31.83	25.72	20.2	**23.67**	**32.45**
MAE	PiMLE	-	102.11	98.1	86.16	79.36	85.46	78.14	88.3	84.68	96.19
	MLE	109.76	90.33	88.94	85.33	79.53	84.32	72.08	79.5	80.96	88.12
	MoS	-	102.11	98.1	86.16	79.36	85.46	78.14	88.3	84.68	96.19
	KLN	109.76	90.33	88.94	85.33	79.53	84.32	72.08	79.5	80.96	88.12
	GNSUM	**25.03**	**40.78**	**33.7**	**40.48**	**46.89**	**38.67**	**37.89**	**26.94**	**48.85**	**35.95**
	Overd.	227.7	228.14	392.67	431.49	482.18	471.99	568.86	709.23	637.82	764.59
	TPC	115.15	94.65	92.21	90.2	83.72	88.82	75.77	82.72	85.38	91.72
	RoA	99.71	88.09	88.7	88.8	79.71	84.93	74.9	80.58	79.62	90.64
	AoR	103.14	89.21	91.89	84.95	74.46	83.29	77.61	84.82	81.6	94.49
MSE	PiMLE	-	11849.27	10595.17	8304.52	7534.15	8236.13	6804.05	8188.87	7822.91	10380.13
	MLE	13049.68	8927.35	8682.89	8009.02	7902.75	7943.21	5495.86	6700.77	7407.58	8823.91
	MoS	-	11849.27	10595.17	8304.52	7534.15	8236.13	6804.05	8188.87	7822.91	10380.13
	KLN	13049.68	8927.35	8682.89	8009.02	7902.75	7943.21	5495.86	6700.77	7407.58	8823.91
	GNSUM	**1394.68**	**2688.01**	**2230.96**	**2631.17**	6890.94	**2914.04**	**2992.04**	**1015.77**	**4811.63**	**2194.32**
	Overd.	82046.82	86031.51	237549.51	240219.24	315110.09	310364.3	385842.53	543020.72	445922.46	604218.56
	TPC	14251.63	9745.82	9248.18	8913.73	8555.71	8722.61	6061.14	7216.71	8162.08	9473.73
	RoA	10883.43	8385.51	8697.7	8579.21	7869.11	8060.25	5921.24	6868.96	7098.07	9268.77
	AoR	12092.25	9078.71	9323.43	7889.37	**6607.07**	7899.5	6651.27	7581.22	7190.27	9928.4

Table 2.S.26: Metrics for the simulations varying the number of subpopulations in the P-S scenario.

Metric	Method	Number of Subpopulations				
		2	4	6	8	10
Mean	PiMLE	-	204.13	198.07	193.73	203.36
	MLE	181.31	187.97	193.24	188.16	198.87
	MoS	-	204.13	198.07	193.73	203.36
	KLN	181.31	187.97	193.24	188.16	198.87
	GNSUM	**115.23**	**91.11**	**144.86**	**122.43**	**117.38**
	Overd.	290.83	360.8	465.53	626.91	531.92
	TPC	185.95	192.46	197.25	192.39	203.32
	RoA	174.4	187.37	192.5	187.26	198.87
	AoR	177.83	192.98	193.3	191.27	203.36
SD	PiMLE	-	37.55	36.04	29.63	34.46
	MLE	30.86	**25.89**	31.98	25.61	30.13
	MoS	-	37.55	36.04	29.63	34.46
	KLN	30.86	**25.89**	31.98	25.61	30.13
	GNSUM	55.75	34.26	57.09	64.6	48.13
	Overd.	164.85	219.57	219.32	319.19	301.03
	TPC	30.96	26.31	31.92	25.71	**29.75**
	RoA	**29.66**	25.92	**31.39**	**25.12**	30.13
	AoR	33.1	31.21	33.29	29.59	34.46
MAE	PiMLE	-	104.13	98.07	93.73	103.36
	MLE	81.31	87.97	93.24	88.16	98.87
	MoS	-	104.13	98.07	93.73	103.36
	KLN	81.31	87.97	93.24	88.16	98.87
	GNSUM	**39.89**	**28.39**	**56.81**	**41.94**	**34.21**
	Overd.	190.83	260.8	365.53	526.91	431.92
	TPC	85.95	92.46	97.25	92.39	103.32
	RoA	74.4	87.37	92.5	87.26	98.87
	AoR	77.83	92.98	93.3	91.27	103.36
MSE	PiMLE	-	12181.96	10852.07	9619.05	11811.4
	MLE	7515.54	8376.06	9664.95	8394.66	10637.36
	MoS	-	12181.96	10852.07	9619.05	11811.4
	KLN	7515.54	8376.06	9664.95	8394.66	10637.36
	GNSUM	**3184.38**	**1193.73**	**5109.11**	**4467.75**	**2502.69**
	Overd.	62232.41	113814.41	179302.87	374415.03	272645.25
	TPC	8298.13	9206.32	10426.09	9164.75	11515.77
	RoA	6371.36	8271.31	9492.22	8213.63	10637.36
	AoR	7097.69	9569.89	9758	9162.03	11811.4

Table 2.S.27: Metrics for the simulations varying the number of subpopulations in the P-S scenario with disjoint subpopulations.

Metric	Method	Number of Subpopulations									
		2	4	6	8	10	12	14	16	18	20
Mean	PiMLE	-	-	-	-	-	-	-	-	-	-
	MLE	215.27	179.49	200.97	190.67	190.44	170.02	187.56	184.33	176.05	188.15
	MoS	-	-	-	-	-	-	-	-	-	-
	KLN	215.27	179.49	200.97	190.67	190.44	170.02	187.56	184.33	176.05	188.15
	GNSUM	**135.09**	**98.98**	**119.68**	**118.76**	**129.26**	**103.48**	**114.01**	**129**	**91.39**	**107.63**
	Overd.	168.69	140.53	200.77	163.57	168.35	137.08	171	138.22	156.06	141.3
	TPC	223.96	184.8	206.34	196.35	195.49	174.88	192.79	188.14	180.39	192.2
	RoA	180.61	188.6	194.6	187.55	195.68	183.4	187.58	178.26	178.68	196.29
	AoR	183.6	189.29	186.23	187.2	197.77	184.12	189.64	179.23	171.31	194.89
SD	PiMLE	-	-	-	-	-	-	-	-	-	-
	MLE	38.06	33.91	40.78	38.77	35.89	**30.62**	24.31	26.04	26.65	29.74
	MoS	-	-	-	-	-	-	-	-	-	-
	KLN	38.06	33.91	40.78	38.77	35.89	**30.62**	24.31	26.04	26.65	29.74
	GNSUM	61.42	43.15	53.5	51.01	53.25	46.49	45.97	81.36	35.01	**29.62**
	Overd.	69.38	45.68	104.08	87.11	69.73	45.19	54.67	52.18	72.62	60.33
	TPC	38.62	34.82	40.17	39	35.68	31.56	25.54	25.13	26.35	30.06
	RoA	**29.35**	**30.67**	39.01	**33.21**	33.87	31.32	**21.14**	24.83	**24.06**	30.68
	AoR	31.92	34.61	**37.37**	33.63	**29.55**	32.91	25.58	**22.59**	30.62	30.58
MAE	PiMLE	-	-	-	-	-	-	-	-	-	-
	MLE	115.27	79.49	100.97	90.67	90.44	70.02	87.56	84.33	76.05	88.15
	MoS	-	-	-	-	-	-	-	-	-	-
	KLN	115.27	79.49	100.97	90.67	90.44	70.02	87.56	84.33	76.05	88.15
	GNSUM	**51.39**	**31.97**	**43.37**	**40.67**	**47.03**	**39.96**	**37.31**	**46.87**	**25.95**	**23.46**
	Overd.	74.57	52.03	103.93	76.26	75.74	46.19	71.99	50.72	72.38	56.16
	TPC	123.96	84.8	106.34	96.35	95.49	74.88	92.79	88.14	80.39	92.2
	RoA	80.61	88.6	94.6	87.55	95.68	83.4	87.58	78.26	78.68	96.29
	AoR	83.6	89.29	86.23	87.2	97.77	84.12	89.64	79.23	71.31	94.89
MSE	PiMLE	-	-	-	-	-	-	-	-	-	-
	MLE	14662.46	7411.13	11775.2	9647.8	9402.26	5793.32	8227.39	7756.19	6457.54	8610.16
	MoS	-	-	-	-	-	-	-	-	-	-
	KLN	14662.46	7411.13	11775.2	9647.8	9402.26	5793.32	8227.39	7756.19	6457.54	8610.16
	GNSUM	**4814.75**	**1770.16**	**3106.67**	**2823.8**	**3550.61**	**2065.07**	**2203.85**	7128.83	**1238.6**	**891.69**
	Overd.	9290.38	3625.08	20447.16	11248.91	9290.78	3314.87	7880.64	**4047.48**	8153.74	5163.41
	TPC	16782.25	8342.51	12841.46	10728.22	10326.92	6552.5	9230.44	8368.55	7122.67	9359.83
	RoA	7315.82	8743.17	10394.01	8713.15	10244.07	7888.16	8094.9	6710.08	6740.34	10166.6
	AoR	7957.61	9111.63	8761.39	8677.98	10389.23	8104.41	8657.03	6762.14	5976.09	9892.51

Table 2.S.28: Metrics for the simulations varying the number of subpopulations in the P-S scenario with small subpopulations.

Metric	Method	Number of Subpopulations									
		2	4	6	8	10	12	14	16	18	20
Mean	PiMLE	-	-	-	-	-	-	-	-	-	-
	MLE	232.85	178.21	194.73	186.36	185.33	165.49	189.04	197.63	183.06	172.19
	MoS	-	-	-	-	-	-	-	-	-	-
	KLN	232.85	178.21	194.73	186.36	185.33	165.49	189.04	197.63	183.06	172.19
	GNSUM	**139.41**	**124.49**	**106.07**	**116.83**	**114.98**	**103.46**	**131.77**	**121.27**	**103.53**	**97.63**
	Overd.	165.02	145.44	162.66	151.73	175.98	150.47	157.68	152.19	149.43	153.31
	TPC	243.98	185.59	200.42	191.37	189.63	169.8	194.34	202.76	188.26	177.31
	RoA	183.78	185.67	188.89	181.99	189.45	172.86	187.01	192.24	181.58	178.53
	AoR	182.33	184.98	188.32	177.36	185.72	176.13	192.65	193.94	181.88	182.33
SD	PiMLE	-	-	-	-	-	-	-	-	-	-
	MLE	46.05	**29.48**	26.12	34.19	**27.8**	**19.73**	**15.84**	29.75	27.6	**22.36**
	MoS	-	-	-	-	-	-	-	-	-	-
	KLN	46.05	**29.48**	26.12	34.19	**27.8**	**19.73**	**15.84**	29.75	27.6	22.36
	GNSUM	43.38	38.53	44.01	57.36	53.81	54.77	56.66	64.85	46.75	39.51
	Overd.	66.76	38.51	62.35	66.09	98.3	67.89	57.5	54.74	52.48	59.33
	TPC	50.5	31.86	26.99	34.97	27.84	20.22	16.39	29.84	28.19	22.4
	RoA	**26.54**	30.8	**23.63**	30.01	28.64	20.86	18.66	**26.44**	**26.3**	24.95
	AoR	30.94	34.93	24.93	**18.36**	32.61	24.8	23.83	28.63	27.79	29.48
MAE	PiMLE	-	-	-	-	-	-	-	-	-	-
	MLE	132.85	78.21	94.73	86.36	85.33	65.49	89.04	97.63	83.06	72.19
	MoS	-	-	-	-	-	-	-	-	-	-
	KLN	132.85	78.21	94.73	86.36	85.33	65.49	89.04	97.63	83.06	72.19
	GNSUM	**42.22**	**35.19**	**30.01**	**38.91**	**36.61**	**39.84**	**42.07**	**49.53**	**34.51**	**30.7**
	Overd.	73.79	46.76	70.05	60.65	76.62	60.79	65.26	61.22	59.49	65.01
	TPC	143.98	85.59	100.42	91.37	89.63	69.8	94.34	102.76	88.26	77.31
	RoA	83.78	85.67	88.89	81.99	89.45	72.86	87.01	92.24	81.58	78.53
	AoR	82.33	84.98	88.32	77.36	85.72	76.13	92.65	93.94	81.88	82.33
MSE	PiMLE	-	-	-	-	-	-	-	-	-	-
	MLE	19664.85	6941.89	9622.56	8569.28	8015.02	4658.97	8166.49	10371.5	7622.29	5686.82
	MoS	-	-	-	-	-	-	-	-	-	-
	KLN	19664.85	6941.89	9622.56	8569.28	8015.02	4658.97	8166.49	10371.5	7622.29	5686.82
	GNSUM	**3340.54**	**2009.89**	**1877.02**	**3408.81**	**2975.12**	**2861.28**	**4059.18**	**4448.4**	**2088.56**	**1488.7**
	Overd.	8461.67	3473.28	7619.21	6825.59	14951.56	6925.4	6468.2	5570.02	5059.47	6186.19
	TPC	23152.8	8289.54	10777.31	9510.99	8770.25	5260.99	9155.72	11405.71	8543.73	6452.99
	RoA	7687.49	8240.73	8431.3	7578.27	8779.83	5721.22	7900.75	9171.78	7312.48	6758.44
	AoR	7687.67	8380.63	8391.22	6304.76	8358.5	6380.46	9122.95	9604.02	7438.02	7603.33

Table 2.S.29: Metrics for the simulations varying the number of subpopulations in the P-S scenario with disjoint and small subpopulations.

Metric	Method	Subpopulations					
		S_0	S_1	S_2	S_3	S_4	S_5
Mean	PiMLE	**99.75**	103.24	-	107.33	102.25	95.35
	MLE	98.3	101.66	**102.79**	102.78	**100.61**	92.67
	MoS	101.66	103.24	-	139.37	142.33	**98.58**
	KLN	98.05	101.66	102.79	**102.11**	100.74	92.59
	GNSUM	97.35	**99.6**	107.56	102.46	103.15	88.69
	Overd.	245.87	614.21	128.65	159.51	207.6	172.65
	TPC	102.41	106.12	107.69	106.83	104.06	96.44
	RoA	97.8	101.44	103.38	103.03	100.85	91.95
	AoR	97.73	101.68	103.79	103.26	100.93	92.01
SD	PiMLE	22.73	27.09	-	21.86	21.32	22.71
	MLE	21.49	25.66	25.32	21.68	20.63	21.69
	MoS	23.08	27.09	-	32.64	45.21	23.67
	KLN	21.31	25.66	25.32	**19.79**	21.16	22.21
	GNSUM	21.32	25.46	26.89	22.35	**20.04**	**19.39**
	Overd.	162.94	331.43	40.39	111.67	128.06	115.88
	TPC	20.92	25.33	25.88	21.68	21.33	21.95
	RoA	20.83	**25.02**	24.19	21.22	20.37	20.77
	AoR	**20.46**	25.23	**24.13**	21.72	20.04	20.83
MAE	PiMLE	18.4	22.18	-	17.67	17.5	16.8
	MLE	17.49	21.14	21.09	15.59	16.87	17.86
	MoS	18.76	22.18	-	41.46	43.96	16.83
	KLN	17.41	21.14	21.09	**15**	17.23	18.61
	GNSUM	18.61	**20.58**	21.66	16.3	**16.24**	18.31
	Overd.	152.9	515.52	36.38	83.59	112.96	79.58
	TPC	17.46	21.3	22.91	16.36	17.3	**16.13**
	RoA	17.34	20.73	**20.46**	15.63	16.84	17.88
	AoR	**17.04**	20.83	20.65	15.92	16.53	17.71
MSE	PiMLE	490.81	707.87	-	507.87	436.88	511.67
	MLE	441.44	628.19	617.08	454.28	404.59	500.71
	MoS	508.7	707.87	-	2561.48	3733.7	534.31
	KLN	435.17	628.19	617.08	**376.63**	426.07	523.69
	GNSUM	438.72	615.92	744.19	480.66	391.5	485.29
	Overd.	46499.64	368771.68	2370.5	15387.1	27156.18	18034.08
	TPC	421.72	647.06	695.14	493.27	448.83	**470.18**
	RoA	417.11	**596.78**	**567.19**	436.94	395	474.63
	AoR	**403.04**	607.44	567.54	458.69	**382.47**	476

Table 2.S.30: Metrics for the simulations varying the sizes of the subpopulations in the W-U scenario.

Metric	Method	Subpopulations					
		S_0	S_1	S_2	S_3	S_4	S_5
Mean	PiMLE	96.42	103.5	-	104.71	94.69	101.99
	MLE	94.05	101.76	98.68	102.23	92.63	**100.04**
	MoS	**98.81**	103.5	-	143.65	125.56	104.06
	KLN	94.05	101.76	98.68	105.24	92.46	100.61
	GNSUM	92.55	98.68	102.99	**98.5**	93.46	98.2
	Overd.	176.83	548.04	123.12	188.26	140.55	204.17
	TPC	98.36	106.06	103.32	105.94	**96.74**	103.88
	RoA	93.62	101.3	99.21	101.89	92.93	100.39
	AoR	94.05	**101.14**	**99.28**	101.82	92.57	100.18
SD	PiMLE	17.73	**19.06**	-	20.83	19.97	22.04
	MLE	16.46	19.63	21.22	19.45	18.13	21.95
	MoS	19.28	**19.06**	-	43.35	43.27	23.15
	KLN	16.45	19.63	21.22	22.27	20.12	22.48
	GNSUM	16.76	19.35	22.17	19.77	18.47	22.8
	Overd.	93.33	300.35	50.25	147.19	69.24	110.19
	TPC	16.81	20.4	21.34	**19.43**	**17.55**	22.25
	RoA	**16.17**	19.33	19.24	19.88	17.83	21.91
	AoR	16.53	19.66	**18.91**	20.11	17.77	**21.67**
MAE	PiMLE	14.36	15.28	-	17.92	15.9	17.07
	MLE	13.31	15.01	15.09	16.2	15.1	**16.63**
	MoS	14.6	15.28	-	47.22	36.89	17.9
	KLN	**13.2**	15.01	15.09	17.84	16.72	17.1
	GNSUM	14.45	**13.65**	16	**15.58**	14.24	18.88
	Overd.	86.5	448.04	35.46	108.92	56.18	109
	TPC	13.25	17.84	13.95	16.8	**13.81**	16.79
	RoA	13.36	14.61	13.93	16.58	14.75	16.77
	AoR	13.59	14.74	**13.73**	16.78	14.92	16.79
MSE	PiMLE	311.34	357.28	-	434.47	407.14	465.27
	MLE	292.9	369.05	429.42	**364.48**	366.68	457.9
	MoS	354.46	357.28	-	3691.17	2431.71	525.57
	KLN	292.62	369.05	429.42	498.41	441.27	480.55
	GNSUM	322.2	357.62	475.93	373.74	366.69	497.08
	Overd.	14178.55	286437.64	2933.58	28371.14	6198.3	22386.66
	TPC	**271.28**	432.04	443.76	393.77	**303.2**	485.45
	RoA	288.98	**356.73**	352.14	378.94	351.96	456.02
	AoR	294.95	368.52	**340.35**	387.43	355.23	**446.25**

Table 2.S.31: Metrics for the simulations varying the sizes of the subpopulations in the W-U scenario with disjoint subpopulations.

Metric	Method	Subpopulations					
		S_0	S_1	S_2	S_3	S_4	S_5
Mean	PiMLE	115.76	113.28	-	133.59	81.51	113.33
	MLE	103.89	116.97	**104.89**	120	82.15	102.54
	MoS	118.32	113.28	-	186.97	120.3	113.04
	KLN	103.82	116.97	104.89	**119.53**	82.38	102.5
	GNSUM	114.93	**108.57**	110.88	127.58	80.78	106.06
	Overd.	197.74	611.61	150.78	160.16	**104.01**	231.04
	TPC	107.9	121.13	108.85	124.54	85.91	106.19
	RoA	**103.48**	117.3	105.64	119.57	83.02	101.1
	AoR	103.86	118.02	105.15	120.66	83.07	**101.02**
SD	PiMLE	63.59	44.73	-	61.16	51.59	50.13
	MLE	50.96	46.2	**69.96**	52.29	52.08	38.86
	MoS	66.54	44.73	-	83.68	132.97	52.15
	KLN	51.04	46.2	69.96	52	**50.95**	38.4
	GNSUM	58.61	**43.82**	75.75	55.71	53.97	39.17
	Overd.	73.43	253.7	87.9	111	58.39	124.81
	TPC	51.11	46.46	70.35	52.05	52.13	38.89
	RoA	50.91	46.71	70.66	**51.98**	53.73	**38.13**
	AoR	**50.86**	46.68	70.56	52.6	53.63	38.9
MAE	PiMLE	44.59	39.53	-	53.86	43.94	40.76
	MLE	38.04	42.02	60.05	43.88	43.45	31.92
	MoS	47.45	39.53	-	92.7	61.63	41.1
	KLN	38.24	42.02	60.05	**43.08**	42.74	**31.12**
	GNSUM	40.5	**36.79**	64.19	48.74	45.66	31.53
	Overd.	98.18	511.61	66.96	75.31	44.3	135.24
	TPC	**36.79**	43.61	59.56	44.34	**41.12**	31.87
	RoA	37.62	42.53	59.74	43.75	43.78	31.24
	AoR	37.85	43.31	**59.51**	44.46	44.03	32.01
MSE	PiMLE	4089.25	2076.81	-	4682.33	2870.55	2564.46
	MLE	2482.57	2315.47	**4673.37**	2997.33	2895.52	1441.39
	MoS	4542.2	2076.81	-	14214.79	17209.55	2753.12
	KLN	2489.76	2315.47	4673.37	2950.27	**2776.63**	1407.08
	GNSUM	3486.48	**1897.59**	5569.42	3709.25	3136.68	1494.23
	Overd.	14676.47	322896.14	9918.4	15325.21	3254.57	31970.14
	TPC	2544.39	2497.09	4780.21	3175.65	2780.24	1474.82
	RoA	2474.42	2372.35	4775.29	**2949.89**	3030.62	**1382.47**
	AoR	**2472.17**	2394.65	4756.38	3054.63	3019.62	1438.5

Table 2.S.32: Metrics for the simulations varying the sizes of the subpopulations in the W-S scenario.

Metric	Method	Subpopulations					
		S_0	S_1	S_2	S_3	S_4	S_5
Mean	PiMLE	90.33	92.87	-	100.97	92.13	116.7
	MLE	84.38	96.97	104.99	96.31	90.1	106.6
	MoS	**92.84**	92.87	-	139.79	134.94	116.61
	KLN	84.31	96.97	104.99	96.75	91.48	107.68
	GNSUM	90.64	90.69	**98.95**	101.49	87.52	112.66
	Overd.	211.22	507.04	127.36	176.64	**104.55**	238.03
	TPC	88.23	**101.58**	109.73	**100.71**	94.36	110.59
	RoA	84.11	97.01	105.39	96.99	90.73	106.54
	AoR	84.73	98.04	105.57	96.37	91.38	**106.08**
SD	PiMLE	40.93	**38.22**	-	44.68	43.52	70.6
	MLE	33.71	41.37	49.78	41.58	42.56	60.24
	MoS	42.89	**38.22**	-	72.74	81.88	69.82
	KLN	33.11	41.37	49.78	**41.26**	44.63	60.57
	GNSUM	35.31	40.36	**39.93**	43.96	42.45	63.78
	Overd.	80.27	214.13	58.43	116.72	52.36	107.42
	TPC	33.99	41.34	50.93	42.12	**42.38**	60.72
	RoA	**33.04**	42.01	49.12	42.54	43.68	60.3
	AoR	33.73	43.19	48.98	42.45	43.07	**59.46**
MAE	PiMLE	36.39	**29.77**	-	36.43	36.04	57.89
	MLE	31.1	32.74	35.58	35.46	36.18	50.72
	MoS	37.89	**29.77**	-	63.76	61.73	56.55
	KLN	30.31	32.74	35.58	35.72	37.07	50.4
	GNSUM	30.86	33.09	**31.74**	35.2	36.46	52.05
	Overd.	116.51	407.04	48.33	96.07	43.75	139.41
	TPC	**30.16**	32.44	35.97	**34.46**	**35.38**	50.4
	RoA	30.66	33.51	36.53	35.71	36.86	50.19
	AoR	31.07	33.89	36.88	35.82	36.84	**49.74**
MSE	PiMLE	1684.59	**1438.4**	-	1897.66	1860.92	5013.24
	MLE	1323.93	1635.45	2379.17	1656.31	1819.01	3490.69
	MoS	1799.19	**1438.4**	-	6609.51	7589.78	4907.43
	KLN	1287.46	1635.45	2379.17	**1627.81**	1964.57	3544.8
	GNSUM	1272.42	1634.31	**1515.52**	1837.89	1868.01	4024.25
	Overd.	18491.34	209245.6	3991.54	18816.88	2624.95	30013.52
	TPC	**1236.11**	1626.17	2558.93	1685.79	**1737.96**	3614.66
	RoA	1289.88	1685.57	2320.83	1728.45	1898.28	3496.46
	AoR	1314.3	1776.17	2310.08	1725.44	1836.52	**3395.99**

Table 2.S.33: Metrics for the simulations varying the sizes of subpopulations in the W-S scenario with disjoint subpopulations.

Metric	Method	Subpopulations					
		S_0	S_1	S_2	S_3	S_4	S_5
Mean	PiMLE	109.32	102.17	-	95.03	91.39	103.71
	MLE	101.57	95.79	95.28	91.4	91.33	97.28
	MoS	109.92	102.17	-	142.36	158.79	108.79
	KLN	101.88	95.79	95.28	90.01	94.98	97.8
	GNSUM	107.73	109.5	92.91	104.23	**100.04**	95.29
	Overd.	190.81	583.27	130.02	108.82	132.2	180.67
	TPC	105.3	**100.1**	**99.8**	**95.77**	94.76	101.45
	RoA	**98.7**	93.65	97.85	92.4	90.7	96.85
	AoR	103.65	97.22	95.07	90.69	88.63	**98.66**
SD	PiMLE	27.6	22.77	-	30.35	28.05	19.94
	MLE	20.86	20.32	21.83	23.69	20.5	19.11
	MoS	28.05	22.77	-	76.81	78.96	23.27
	KLN	20.84	20.32	21.83	24.69	22.32	20.02
	GNSUM	35.49	36.49	24.45	32.04	32.87	**18.35**
	Overd.	124.24	297.1	72.41	74.03	64.79	109.43
	TPC	20.71	20.3	22.19	24	**19.81**	19.8
	RoA	**19.96**	**19.68**	**18.9**	**23.18**	20.79	18.6
	AoR	23.8	21.3	25.13	26.91	25.97	19.42
MAE	PiMLE	23.56	18.7	-	26.06	24.59	15.99
	MLE	16.63	16.81	18.31	20.06	18.43	15.49
	MoS	24.6	18.7	-	60.34	69.05	19.19
	KLN	16.59	16.81	18.31	22.28	19.22	16.13
	GNSUM	25.75	28.57	20.55	22.18	26.79	**14.44**
	Overd.	100.43	483.27	52.4	53.19	53.5	82.88
	TPC	17.18	**16.41**	17.82	19.33	**16.95**	16.15
	RoA	**15.68**	16.62	**14.96**	**18.99**	18.58	15.46
	AoR	19.55	17.36	19.83	23.83	23.33	15.67
MSE	PiMLE	810.53	497.17	-	899.79	821.39	391.56
	MLE	415.81	409.82	474.85	607.32	474.45	354.3
	MoS	846.09	497.17	-	7399.06	9378.23	591.74
	KLN	416.01	409.82	474.85	678.88	498.67	385.57
	GNSUM	1256.6	1355.41	617.99	993.11	1026.56	342.12
	Overd.	22910.74	317400.21	5881.9	5284.62	5024.95	17883.83
	TPC	435.7	**391.53**	467.71	**565.31**	**400.39**	374.64
	RoA	**380.25**	408.36	**343.89**	568.01	496.97	**338.67**
	AoR	551.41	438.81	624.32	774.59	769.74	360.11

Table 2.S.34: Metrics for the simulations varying the sizes of the subpopulations in the P-U scenario.

Metric	Method	Subpopulations					
		S_0	S_1	S_2	S_3	S_4	S_5
Mean	PiMLE	106.51	111.2	-	93.8	103.48	**99.41**
	MLE	**99.91**	103.35	89.86	89.78	**99.52**	93.03
	MoS	109.23	111.2	-	141.37	191.01	102.9
	KLN	100.74	103.35	89.86	88.05	101.38	93.45
	GNSUM	107.98	110.53	**97.56**	93.72	111.96	102.19
	Overd.	193.06	673.1	113.15	**101.54**	157.02	254.67
	TPC	104.51	107.71	94.01	93.69	104.01	97.06
	RoA	97.11	**100.58**	90.69	91.36	98.82	92.69
	AoR	99.67	103.1	91.74	90.01	98.4	94.04
SD	PiMLE	26.42	27.22	-	23.06	26.92	20.23
	MLE	20.22	25.6	16.73	16.21	22.33	13.83
	MoS	29.32	27.22	-	71.42	86.43	24.28
	KLN	20.48	25.6	16.73	**15.24**	23.37	14.37
	GNSUM	35.66	30.37	33.17	25.57	38.09	31.95
	Overd.	133.56	326.04	51.61	63.95	129.68	157.14
	TPC	20.76	25.6	**16.72**	16.95	**22.09**	14.38
	RoA	**19.84**	24.79	18.19	15.64	22.29	**13.6**
	AoR	23.54	**24.46**	17.57	17.5	25.18	19.03
MAE	PiMLE	20.53	24.64	-	18.41	21.54	14.84
	MLE	16.12	21.89	14.66	15.86	17.42	12.71
	MoS	23.75	24.64	-	54.85	101.42	17.73
	KLN	**16.05**	21.89	14.66	15.15	18.36	13.1
	GNSUM	29.99	23.87	27.12	22.03	32.78	25.39
	Overd.	100.75	573.1	39.82	47.28	83.5	157.36
	TPC	16.25	21.85	**13.56**	14.69	17.95	**12.29**
	RoA	16.21	21.38	17	**14.43**	**16.94**	12.59
	AoR	17.71	**21.23**	16.29	15.57	18.46	15.65
MSE	PiMLE	705.56	829.22	-	543.64	700.3	389.08
	MLE	388.25	633.62	368.72	353.85	473.91	230.32
	MoS	901.94	829.22	-	6556.98	15379.35	568.5
	KLN	398.96	633.62	368.72	363.31	520.53	239.2
	GNSUM	1271.63	987.26	1051.43	660.68	1521.08	974.41
	Overd.	25605.73	429435.13	2703.91	3887.3	19226.78	47378.65
	TPC	429.68	682.13	**301.52**	312.77	479.78	**205.19**
	RoA	**382.21**	584.36	400.85	**307.05**	**473.27**	229.29
	AoR	526.37	**577.92**	361.54	390.64	604.72	379.66

Table 2.S.35: Metrics for the simulations varying the sizes of the subpopulations under the P-U scenario with disjoint subpopulations.

Metric	Method	Subpopulations					
		S_0	S_1	S_2	S_3	S_4	S_5
Mean	PiMLE	198.87	192.72	-	193.08	194.74	194.05
	MLE	191.94	184.52	177.16	177.37	183.82	187.66
	MoS	203.56	192.72	-	296.91	304.55	204.7
	KLN	194.15	184.52	177.16	173.04	173.17	191.42
	GNSUM	**121.73**	**108.58**	**115.02**	**102.49**	**107.52**	**113.26**
	Overd.	181.99	471.44	163.24	189.41	201.91	230.81
	TPC	196.42	188.86	180.97	181.71	187.63	192.48
	RoA	187.17	186.91	181.49	180.2	186.49	187.71
	AoR	188.26	188.64	180.65	185.3	192.16	189.04
SD	PiMLE	47.12	32.72	-	43	20.31	31.54
	MLE	35.27	**26.37**	32.66	33.23	**11.78**	25.29
	MoS	51.81	32.72	-	105.72	88.54	35.12
	KLN	36.04	**26.37**	32.66	**32.81**	14.5	26.24
	GNSUM	44.11	55.44	66.27	46.49	38.9	42.9
	Overd.	46.46	290.81	79.95	76.05	116.04	120.44
	TPC	36.06	26.4	32.25	33.07	12.32	25.58
	RoA	**30.82**	28.03	**31.55**	33.35	13.72	**24.98**
	AoR	41.96	31.63	38.15	36.04	20.34	30.3
MAE	PiMLE	98.87	92.72	-	93.08	94.74	94.05
	MLE	91.94	84.52	77.16	77.37	83.82	87.66
	MoS	103.56	92.72	-	196.91	204.55	104.7
	KLN	94.15	84.52	77.16	73.04	73.17	91.42
	GNSUM	**36.73**	**42.54**	**39.15**	**36.3**	**32.18**	**32.25**
	Overd.	83.96	371.44	69.49	94.1	102.59	130.81
	TPC	96.42	88.86	80.97	81.71	87.63	92.48
	RoA	87.17	86.91	81.49	80.2	86.49	87.71
	AoR	88.26	88.64	80.65	85.3	92.16	89.04
MSE	PiMLE	11884.35	9613.67	-	10420.88	9367.17	9790.26
	MLE	9634.81	7804.44	6966.66	7035.64	7158.1	8292.62
	MoS	13274.45	9613.67	-	49391.26	49288.09	12134.26
	KLN	10097.35	7804.44	6966.66	6356.65	5553.4	9012.33
	GNSUM	**2320.79**	**2993.91**	**4397.6**	**2059.43**	**1494.03**	**1924.53**
	Overd.	8773.41	218309.45	10071.31	13488.72	23177.05	30891.84
	TPC	10531.71	8557.97	7544.03	7715.48	7824.15	9173.71
	RoA	8500.24	8299.25	7586.65	7488.05	7658.93	8285.61
	AoR	9461.64	8807.48	7886.54	8509.17	8887.42	8800.86

Table 2.S.36: Metrics for the simulations varying the sizes of the subpopulations under the P-S scenario.

Metric	Method	Subpopulations					
		S_0	S_1	S_2	S_3	S_4	S_5
Mean	PiMLE	204.09	201.22	-	210.75	193.26	199.31
	MLE	198.59	189.66	175.44	188.82	188.8	188.69
	MoS	208.96	201.22	-	298.2	273.42	213.23
	KLN	200.36	189.66	175.44	179.09	178.09	192.68
	GNSUM	**112.92**	**124.96**	**131.72**	**113.71**	**117.87**	**112.42**
	Overd.	203.67	407.38	141.63	198.77	178.61	225.55
	TPC	202.91	193.29	180.63	193	192.8	193.84
	RoA	192.24	191.62	180.63	193.53	191.51	188.45
	AoR	190.21	197.76	186.06	199.18	192.14	192.8
SD	PiMLE	45.95	27.4	-	42.68	36.76	38.22
	MLE	38.84	24.59	38.84	27.71	25.91	35.84
	MoS	49.61	27.4	-	117.98	81.48	41.21
	KLN	40.17	24.59	38.84	30.85	28.15	37.16
	GNSUM	49.77	46.73	62.33	41.53	59.52	53.71
	Overd.	76.47	178.58	58.82	181.6	88.33	184.55
	TPC	39.25	**24.42**	40.33	**26.82**	26.42	36.23
	RoA	**35.38**	24.64	**28.85**	28.61	**25.21**	**35.65**
	AoR	42.84	27.66	30.07	37.86	33.94	37.97
MAE	PiMLE	104.09	101.22	-	110.75	93.26	99.31
	MLE	98.59	89.66	75.44	88.82	88.8	88.69
	MoS	108.96	101.22	-	198.2	173.42	113.23
	KLN	100.36	89.66	75.44	79.09	78.09	92.68
	GNSUM	**36.41**	**40.28**	**47**	**34.58**	**39.15**	**39.8**
	Overd.	107.5	307.38	60.43	113.92	86.13	141.46
	TPC	102.91	93.29	80.63	93	92.8	93.84
	RoA	92.24	91.62	80.63	93.53	91.51	88.45
	AoR	90.21	97.76	86.06	99.18	92.14	92.8
MSE	PiMLE	12839.85	10957.73	-	13996.52	9981.38	11250.07
	MLE	11153.3	8613.34	7123.65	8618.93	8523.19	9085.96
	MoS	14210.77	10957.73	-	52506.71	36380.32	14433.44
	KLN	11605.57	8613.34	7123.65	7159.56	6850.95	9901.77
	GNSUM	**2520.1**	**2697.46**	**4697.28**	**1826.8**	**3685.18**	**2894.57**
	Overd.	16302.9	124775.74	5019.51	41084.65	13590.52	48118.33
	TPC	12053.99	9269.47	8045.69	9332.89	9275.05	10052.01
	RoA	9696.95	8971.04	7292.8	9526.26	8977.83	9031.71
	AoR	9880.75	10283.33	8265.39	11197.46	9584.33	9982

Table 2.S.37: Metrics for the simulations varying the sizes of the subpopulations in the P-S scenario.

Metric	Method	Sample size									
		25	50	100	200	300	400	500	1000	2000	3000
Mean	PiMLE	106.52	110.04	105.37	102.82	102.88	102.14	101.87	101.72	102	101.6
	MLE	103.1	108.21	104.25	102.03	101.66	101.03	100.76	100.56	100.75	100.51
	MoS	110.3	111.08	106.78	104.83	104.74	103.82	104.04	103.65	103.9	103.44
	KLN	103.66	108.25	103.93	102.05	101.71	100.97	100.89	100.62	100.78	**100.44**
	GNSUM	97.24	**103.17**	**100.17**	**98.97**	98.55	**99.7**	**99.83**	98.45	97.93	96.61
	Overd.	203.48	230.55	127.8	106.47	102.19	105.74	101.03	100.82	102.17	101.45
	TPC	111.6	112.77	106.03	102.85	102.23	101.43	101.3	100.78	100.87	100.63
	RoA	**102.18**	108.21	104.08	101.89	**101.35**	101.1	100.65	**100.53**	**100.66**	100.49
	AoR	102.47	108.29	104.29	101.95	101.64	101.27	100.63	100.64	100.73	100.56
SD	PiMLE	36.69	17.36	15.15	11.49	**8.26**	6.87	5.52	3.52	3.97	2.65
	MLE	34.77	16.82	14.28	10.99	8.41	**6.7**	5.54	3.46	3.78	2.51
	MoS	39.27	17.29	16.24	12.23	8.88	7.21	5.67	3.8	4.03	2.7
	KLN	35.27	16.85	14.66	11.01	8.73	6.78	5.67	3.52	**3.74**	2.56
	GNSUM	**31.67**	16.49	**14.06**	**10.51**	9.38	7.71	7.88	5.51	3.86	5.72
	Overd.	129.57	161.15	38.97	11.51	10.18	8.23	6.19	5.57	3.88	2.89
	TPC	35.42	16.84	14.1	10.93	8.39	6.79	**5.44**	**3.33**	3.78	2.54
	RoA	33.84	**16.46**	14.65	10.6	8.81	6.94	5.68	3.51	3.83	**2.44**
	AoR	33.48	16.51	14.87	10.72	8.63	6.85	5.76	3.58	3.84	2.48
MAE	PiMLE	26.84	13.93	12.65	10.24	7.28	5.63	4.65	2.96	3.18	2.5
	MLE	25.76	13.21	11.38	9.45	**6.84**	**5.3**	**4.42**	2.64	2.76	1.9
	MoS	28.92	14.23	13.39	11.13	8.39	6.44	5.62	4.17	4.25	3.68
	KLN	26.26	13.26	11.79	9.43	7.1	5.31	4.5	2.73	**2.74**	1.89
	GNSUM	25.59	12.7	**10.55**	**8.71**	8.3	6.31	6.32	4.23	3.59	5.47
	Overd.	105.02	133.07	29.01	11.1	7.49	7.83	4.6	4.34	3.3	2.52
	TPC	26.09	14.42	11.71	9.47	7.06	5.42	4.47	2.68	2.78	1.98
	RoA	25.39	12.64	11.41	9.12	7.05	5.46	4.51	**2.64**	2.8	**1.86**
	AoR	**25.11**	**12.46**	11.76	9.31	7.04	5.42	4.51	2.66	2.79	1.92
MSE	PiMLE	1321.3	387.17	246.93	133.24	73.05	49.47	32.49	14.75	18.99	9.22
	MLE	1158.04	336.18	211.72	118.78	**69.87**	**43.76**	**29.7**	11.67	14.11	6.26
	MoS	1570.82	407.02	296.43	165.5	97.35	63.97	46.93	27.04	30.61	18.75
	KLN	1195.06	337.59	219.66	119.41	75.3	44.63	31.32	12.16	**13.87**	6.41
	GNSUM	**960.35**	**268.34**	**187.95**	**106.02**	85.67	56.51	59.06	31.26	18.44	42.51
	Overd.	26656.52	41712.57	2215.4	167.67	103.17	97.25	37.43	30.14	18.97	10.04
	TPC	1326.61	432.71	225.16	121.58	71.78	45.87	29.84	**11.13**	14.35	6.55
	RoA	1092.66	325.01	220.57	110.33	75.5	46.94	31.13	11.96	14.35	**5.87**
	AoR	1070.95	327.71	228.59	113.02	73.52	46.24	31.91	12.62	14.53	6.14

Table 2.S.38: Metrics for the simulations of the sample size under the W-U scenario.

Metric	Method	Sample size									
		25	50	100	200	300	400	500	1000	2000	3000
Mean	PiMLE	90.38	87.83	101.47	86.5	96.27	99.38	95.53	96.84	97.82	99.27
	MLE	93.77	89.13	101	87.12	97.77	100.38	96.67	98.21	98.46	100.39
	MoS	88.18	85.61	100.9	85.34	95.3	97.23	93.44	95.12	96.22	97.52
	KLN	93.94	89.84	101.08	86.94	97.67	100.28	96.66	98.26	98.49	100.41
	GNSUM	92.14	87.78	98.04	85.12	96.05	97.69	94.17	97.02	95.79	97.21
	Overd.	185.13	196.3	178.21	124.16	115.32	111.42	104.99	102.11	**100.53**	102.47
	TPC	**102.74**	**92.9**	102.99	**88.29**	**98.36**	101.1	**97.14**	**98.68**	98.56	100.42
	RoA	94.69	88.77	**100.44**	87.23	97.83	**100.22**	96.39	98.01	98.33	**100.23**
	AoR	94.53	88.98	100.48	87.13	97.98	100.85	96.82	98.44	98.81	100.66
SD	PiMLE	57.78	45.13	36.75	25.42	19.9	20.67	22.89	8.93	7.52	4.91
	MLE	63.24	43.49	36.16	26.08	20.15	20.92	23.1	9.98	7.77	5.04
	MoS	**52.56**	**41.12**	35.9	**23.89**	**18.89**	**19.57**	**21.83**	**8.55**	**6.96**	**4.8**
	KLN	64.15	44.64	36.25	26.02	19.92	21.01	23.14	9.98	7.78	5
	GNSUM	59.96	42.35	35.43	24.97	19.69	21.18	22.78	10.37	7.77	6.26
	Overd.	94.69	81.48	92.9	50.85	23.2	24.34	24.96	10.24	7.44	5.15
	TPC	65.08	43.66	36.21	26.09	19.94	20.96	23.21	10.11	7.7	5.07
	RoA	63.44	42.87	35.68	26.08	20.27	21	22.83	9.9	7.73	5.1
	AoR	63.03	43.81	**35.09**	26.15	20.45	21.47	23.04	9.97	7.89	5.09
MAE	PiMLE	45.41	36.57	31.52	23.04	16.03	16.89	18.85	**7.56**	6.05	3.88
	MLE	48.51	35.58	30.92	23.01	16.11	16.8	18.6	8.19	6.06	3.88
	MoS	**43.75**	35.18	30.06	22.77	**15.44**	**15.97**	18.69	7.8	6.39	4.35
	KLN	48.74	35.84	30.85	23.1	15.99	16.7	18.69	8.17	6.05	**3.83**
	GNSUM	47.59	35.77	30.28	23.25	16.15	16.59	19.08	8.59	7.18	5.33
	Overd.	92.7	98.77	82.54	39.85	22.37	21.3	19.68	8.52	**5.48**	4.45
	TPC	46.58	**34.58**	30.9	**22.67**	15.89	16.96	18.55	8.39	5.96	3.91
	RoA	48.67	35.39	30.57	22.94	16.18	16.9	18.45	8.09	6.08	3.97
	AoR	48.19	35.53	**29.94**	22.99	16.09	17.24	**18.41**	8.2	6.02	4.1
MSE	PiMLE	3264.49	2082.67	1285.37	796.11	389.98	406.41	517.8	**85.8**	58.55	**23.45**
	MLE	3837.96	1915.17	1243.26	812.25	390.71	415.84	517.99	97.9	59.66	24.25
	MoS	**2764.46**	**1813.66**	1225.21	**757.48**	**361.05**	**371.37**	**495.86**	93.35	60.21	28.08
	KLN	3946.17	1996.48	1249.3	813.66	382.55	419.34	519.94	97.63	59.84	23.93
	GNSUM	3477.35	1853.1	1196.46	813.68	383.95	431.49	526.77	111.06	75.11	44.99
	Overd.	15763.86	15580.89	14315.22	3040.34	745.89	693.06	616.62	104.1	**52.9**	31.29
	TPC	4030.53	1861.6	1254.28	783.85	380.31	418.59	520.09	98.94	58.45	24.57
	RoA	3851.48	1871.88	1209.88	809.43	395.12	418.91	508.28	97.12	59.54	24.75
	AoR	3804.06	1944.55	**1169.99**	815.34	401.39	438.65	514.37	96.83	60.57	25.05

Table 2.S.39: Metrics for the simulations varying the sample size in the W-S scenario.

Metric	Method	Sample size									
		25	50	100	200	300	400	500	1000	2000	3000
Mean	PiMLE	104.27	95.5	97.24	**100.41**	**99.9**	96.8	**99.98**	99.32	98.61	99.35
	MLE	97.19	91.47	94.11	94.9	94.95	92.9	95.35	96.15	94.87	95.32
	MoS	105.6	95.01	**99.03**	101.78	101.1	98.05	101.53	**100.5**	**100.02**	**100.62**
	KLN	96.1	90.59	93.35	94.37	94.49	92.47	94.86	95.61	94.43	94.85
	GNSUM	119.94	**98.95**	102.76	104.88	102.06	**100.28**	104.15	103.95	102.69	105.26
	Overd.	215.37	167.18	108.27	93.59	90.56	88.49	89.41	91.58	88.55	89.32
	TPC	105.02	95.42	96.11	96.03	95.92	93.34	95.87	96.36	94.98	95.37
	RoA	96.82	91.39	93.92	94.22	94.68	92.58	95.08	95.77	94.47	94.95
	AoR	**100.96**	91.48	93.56	95.69	96.91	93.36	96.93	96.76	95.11	95.94
SD	PiMLE	37.65	27.86	18.79	10.45	10.92	9.27	7.69	3.76	3.02	2.39
	MLE	24.51	25.27	15.21	9.22	9.39	6.98	4.52	**3.12**	2.67	1.92
	MoS	38.63	27.63	20.69	10.42	11.57	9.81	7.83	4.17	2.98	2.46
	KLN	24.13	25.03	15.22	**9.12**	**9.09**	**6.95**	**4.34**	3.17	2.66	1.89
	GNSUM	53.55	29.15	19.94	21.54	20.12	24.5	16.67	19.87	19.36	15.75
	Overd.	88.79	92.57	23.07	10.2	9.17	8.43	4.54	5.1	**2.44**	2.57
	TPC	24.3	24.89	15.25	9.43	9.29	7.01	4.61	3.14	2.61	1.92
	RoA	**23.3**	24.93	**14.9**	9.36	9.54	7.01	4.6	3.15	2.72	**1.88**
	AoR	34.08	**24.61**	15.73	10	10.45	8.16	7.11	3.84	2.77	2.09
MAE	PiMLE	28.95	23.77	15.52	8.23	7.93	**7.49**	6.22	**3.04**	2.62	2.01
	MLE	19.19	19.92	13.22	8.45	8	8.34	5.18	4.4	5.44	4.68
	MoS	29.74	23.12	16.42	8.87	8.58	8.06	6.39	3.31	**2.17**	**1.98**
	KLN	19.45	20.12	13.38	8.64	8.05	8.55	5.58	4.79	5.83	5.15
	GNSUM	39.81	25.62	14.2	17.89	14.17	17.9	12.54	15.86	15.74	13.16
	Overd.	122.14	80.55	18.98	9.78	10.98	12.53	10.59	8.86	11.45	10.68
	TPC	19.09	**18.95**	**12.7**	8.09	**7.6**	8.06	**4.89**	4.23	5.32	4.63
	RoA	**18.23**	19.59	13.03	8.95	8.2	8.67	5.33	4.69	5.8	5.05
	AoR	27	21.51	13.95	**7.85**	7.76	8.09	6.3	4.26	5.08	4.06
MSE	PiMLE	1364.95	757.57	343.05	103.89	113.2	91.9	56.23	**13.9**	10.6	**5.86**
	MLE	578.5	679.43	254.58	106.73	109.32	96.78	41.01	24.08	33.05	25.36
	MoS	1449.25	750.42	407.8	106.28	128.36	95.32	60.64	16.8	**8.46**	6.11
	KLN	568.55	683.95	264.47	110.68	108.73	102.56	44.34	28.77	37.74	29.98
	GNSUM	3121.47	808.16	385.32	464.54	388.69	570.21	281.28	390.6	363.27	263.29
	Overd.	20800.7	12653.04	574.06	139.95	169.02	199.82	131.66	95.51	136.86	120.41
	TPC	586.33	**609.34**	**236.08**	**100.16**	**98.69**	**91.01**	**37.29**	22.61	31.64	24.91
	RoA	**526.03**	664.53	247.93	116.65	114.82	101.71	44.27	27.3	37.63	28.81
	AoR	1104.45	648.03	276.64	113.56	113.3	107.33	57.45	24.54	31.23	20.63

Table 2.S.40: Metrics for the simulations of the sample size under P-U scenario.

Metric	Method	Sample size									
		25	50	100	200	300	400	500	1000	2000	3000
	PiMLE	179.43	188.36	185.56	191.24	189.7	190.15	192.99	187.7	192	192.67
	MLE	184.09	183.83	186.24	187.01	183.92	184.24	185.71	183.57	185.33	187.34
	MoS	179.92	192.22	187.78	194.68	193.56	192.93	195.45	190	194.97	195.48
	KLN	181.92	182.7	184.39	186.35	183.34	183.34	184.63	182.4	184.46	186.41
Mean	GNSUM	**114.95**	**105.81**	**108.16**	**88.51**	**116.21**	**117.4**	**104.84**	**123.58**	**117.15**	**106.21**
	Overd.	227.42	188.15	174.64	178.16	172.98	175.92	174.01	172.89	172.65	175.17
	TPC	191.97	188.67	188.37	188.14	184.51	184.87	186.45	183.66	185.4	187.41
	RoA	184.07	183	184.84	186.97	182.95	183.6	185.03	182.69	184.49	186.59
	AoR	174.75	180.36	177.68	185.46	182.84	182.72	185.65	181.24	185.01	185.54
	PiMLE	**46.11**	40.51	18.01	18.66	15.57	11.61	9.43	6.92	5.96	3.74
	MLE	49.72	30.42	27.1	**14.05**	12.26	10.06	7.54	7.54	4.89	3.14
	MoS	46.32	40.83	20.5	19.5	16.29	12.2	9.47	6.86	6	3.81
	KLN	48.88	30.16	27.15	14.4	11.84	10.19	7.65	7.41	4.83	**3.14**
SD	GNSUM	49.82	40.29	41.29	30.99	39.17	44.58	39.63	55.08	37.54	47.83
	Overd.	164.27	85.13	27.59	18.51	15.29	13.89	11.26	10.63	6.18	3.86
	TPC	49.87	**29.68**	27.62	14.23	12.73	9.9	**7.35**	7.46	4.77	3.31
	RoA	48.63	30.1	25.61	14.47	**11.76**	**9.77**	7.59	7.73	**4.64**	3.16
	AoR	46.29	39.01	**15.57**	16.89	13.03	9.93	9.03	**6.84**	5.34	3.29
	PiMLE	79.99	88.36	85.56	91.24	89.7	90.15	92.99	87.7	92	92.67
	MLE	84.09	83.83	86.24	87.01	83.92	84.24	85.71	83.57	85.33	87.34
	MoS	81.03	92.22	87.78	94.68	93.56	92.93	95.45	90	94.97	95.48
	KLN	81.92	82.7	84.39	86.35	83.34	83.34	84.63	82.4	84.46	86.41
MAE	GNSUM	**30.93**	**30.18**	**31.07**	**28**	**31.37**	**35.71**	**29.52**	**44.56**	**28.81**	**30.58**
	Overd.	132.87	94.34	74.64	78.16	72.98	75.92	74.01	72.89	72.65	75.17
	TPC	91.97	88.67	88.37	88.14	84.51	84.87	86.45	83.66	85.4	87.41
	RoA	84.07	83	84.84	86.97	82.95	83.6	85.03	82.69	84.49	86.59
	AoR	75.15	80.36	77.68	85.46	82.84	82.72	85.65	81.24	85.01	85.54
	PiMLE	8329.65	9367.02	7629.46	8654.48	8276.41	8254.53	8732	7737.05	8497.51	8601.07
	MLE	9419.88	7905.97	8134.63	7758.25	7184.69	7192.93	7400.63	7038.26	7303.86	7637.86
	MoS	8425.2	10088.98	8103.89	9325.5	9005.92	8776.65	9196.55	8143.86	9053.25	9129.65
	KLN	8981.4	7702.94	7821.97	7653.44	7078.17	7044.16	7217.75	6842.41	7154.97	7475.75
MSE	GNSUM	**2581.1**	**1576.19**	**1686.4**	**1044.34**	**1720.59**	**2190.76**	**1515.66**	**3438.25**	**1633.2**	**2211.54**
	Overd.	41872.61	14655.12	6294.55	6434.64	5547.58	5947.32	5597.73	5420.24	5314.36	5664.1
	TPC	10821.37	8698.37	8533.1	7961.89	7295.43	7296.05	7524.08	7051.28	7315	7650.38
	RoA	9314.68	7750.66	7821.48	7763.26	7011.28	7079.23	7284.34	6893.68	7159.68	7507.74
	AoR	7623.21	7902.53	6265.1	7574.88	7024.41	6935.68	7413.61	6643.74	7253.64	7327.13

Table 2.S.41: Metrics for the simulations varying the sample size in the P-S scenario.

2.S.5. Case study

We present the performance of the estimators using a real network from the Deezer dataset. Table 2.S.42 shows the values of the mean, standard deviation, relative mean absolute error (RMAE), and mean squared relative error (MSRE) when the subpopulations are simulated uniformly. Table 2.S.43 displays the values for the metrics using an SIR process to generate the unknown subpopulation, while the known subpopulations are simulated uniformly. The mean, standard deviation, RMAE, and MSRE using the real subpopulations are shown in Table 2.S.44.

Metric	Unknown subp.	Methods								
		PiMLE	MLE	MoS	KLN	GNSUM	Overd.	TPC	RoA	AoR
Mean	Asian (503)	-	482.92	-	484.99	504.39	1789.53	513.68	486.24	501.71
	Blues (2969)	-	2888.98	-	2909.69	3019.88	2991.59	2922.8	2908.96	2959.95
	Chillout (875)	-	864.93	-	871.35	891.85	1703.52	893.67	866.26	913.89
SD	Asian	-	133.42	-	134.1	175.88	1157.6	131.59	135.36	219.39
	Blues	-	293.97	-	304.86	766.11	407.2	296.3	299.81	435.91
	Chillout	-	135.47	-	141.99	238.95	757.15	134.12	135.55	249.62
RMAE	Asian	-	0.19	-	0.19	0.24	2.56	0.18	0.19	0.29
	Blues	-	0.07	-	0.07	0.2	0.1	0.07	0.07	0.11
	Chillout	-	0.13	-	0.13	0.2	0.97	0.12	0.13	0.23
MSRE	Asian	-	0.05	-	0.05	0.09	11.09	0.05	0.05	0.16
	Blues	-	0.01	-	0.01	0.06	0.02	0.01	0.01	0.02
	Chillout	-	0.02	-	0.02	0.04	0.94	0.02	0.02	0.05

Table 2.S.42: Metrics of the estimates of several unknown subpopulations, with the subpopulations simulated uniformly over the Deezer network. The numbers in parentheses represent the sizes of the subpopulations.

Metric	Unknown subp.	Methods								
		PiMLE	MLE	MoS	KLN	GNSUM	Overd.	TPC	RoA	AoR
Mean	Asian (539)	-	1465.97	-	1461.68	573.25	2453.03	1495.36	1465	917.78
	Blues (3495)	-	7600.44	-	7607.82	3693.12	7699.17	7636.34	7595.15	5589.19
	Chillout (1099)	-	2354.46	-	2346.53	1129.09	2859.07	2386.83	2333.42	1756.84
SD	Asian	-	373.11	-	370.01	220.08	1020.38	376.43	373.96	265.31
	Blues	-	1038.47	-	1021.66	724.8	1216.2	1031.14	1021.11	796.22
	Chillout	-	441.04	-	433.58	235.25	858.43	440.45	433.91	409.62
RMAE	Asian	-	1.72	-	1.71	0.32	3.55	1.77	1.72	0.72
	Blues	-	1.17	-	1.18	0.16	1.2	1.18	1.17	0.6
	Chillout	-	1.14	-	1.14	0.19	1.6	1.17	1.12	0.6
MSRE	Asian	-	3.43	-	3.4	0.17	16.33	3.62	3.42	0.71
	Blues	-	1.46	-	1.46	0.05	1.55	1.48	1.45	0.4
	Chillout	-	1.3	-	1.29	0.04	2.56	1.37	1.26	0.36

Table 2.S.43: Metrics of the estimates of several unknown subpopulations, with the known subpopulations simulated uniformly and the unknown subpopulation modeled via an SIR process over the Deezer network. The numbers in parentheses represent the sizes of the subpopulations.

Metric	Unknown subp.	Methods								
		PiMLE	MLE	MoS	KLN	GNSUM	Overd.	TPC	RoA	AoR
Mean	Asian (503)	-	516.1	-	530.06	512.65	1991.01	545.29	527.47	499.64
	Blues (2969)	-	2407.57	-	2451.13	2644.8	3081.55	2440.29	2455.69	2960.41
	Chillout (875)	-	781.84	-	798.08	699.18	2286.68	812.95	792.57	946.78
SD	Asian	-	122.19	-	131.4	196.58	1116.22	122.54	121.19	182.23
	Blues	-	306.87	-	286.19	801.9	744.22	307.57	335.37	468.75
	Chillout	-	172.79	-	170	233.49	1469.69	173.74	186.92	285.75
RMAE	Asian	-	0.19	-	0.2	0.28	2.96	0.2	0.19	0.28
	Blues	-	0.19	-	0.17	0.27	0.16	0.18	0.18	0.13
	Chillout	-	0.19	-	0.19	0.28	1.63	0.18	0.2	0.28
MSRE	Asian	-	0.05	-	0.06	0.11	14.24	0.06	0.05	0.12
	Blues	-	0.04	-	0.04	0.1	0.06	0.04	0.04	0.03
	Chillout	-	0.04	-	0.03	0.08	2.65	0.03	0.04	0.08

Table 2.S.44: Metrics of the estimates of several unknown subpopulations, using real subpopulations of the Deezer dataset. The numbers in parentheses represent the sizes of the subpopulations.

2.S.6. Comparison between estimators based on known subpopulations and based on direct degree estimation

This section displays the best method for each scenario in the case of the variation of the number of known subpopulations and the sizes of the known subpopulations. We include one table that considers all estimators and another table that excludes the estimators that do not use the known subpopulations, namely RoA and AoR. We observe that the best estimator using the known subpopulations is very close to the estimators that have the real information of the network size.

Simulation goal	Main scenarios	Mean	MAE	MSE
Subpop. number	W-U	MLE-KLN/100.00	MLE-KLN/18.36	MLE-KLN/530.82
	W-S	MLE-KLN/105.38	TPC/42.39	MLE-KLN/3036.94
	P-U	TPC/99.93	TPC/15.71	TPC/382.1
	P-S	GNSUM/115.58	GNSUM/ 37.52	GNSUM/ 2976.36
Subpop. sizes	W-U	MLE-GNSUM/99.80	MLE/18.34	KLN/501.14
	W-S	KLN/105.02	KLN/42.87	KLN/2768.76
	P-U	TPC/99.53	TPC/17.31	TPC/439.21
	P-S	GNSUM/111.43	GNSUM/ 36.52	GNSUM/ 2531.72

Table 2.S.45: Best method for each scenario. For each metric, it shows the average of the values corresponding to each parameter. *AoR* and *RoA* are not included in the simulations.

Simulation goal	Main scenarios	Mean	MAE	MSE
Subpop. number	W-U	MLE-KLN/100.00	AoR/18.24	MLE-KLN/523.87
	W-S	RoA/105.26	TPC/42.39	RoA/2976.75
	P-U	TPC/99.93	TPC/15.71	TPC/382.1
	P-S	GNSUM/115.58	GNSUM/ 37.52	GNSUM/ 2976.36
Subpop.sizes	W-U	AoR/99.90	AoR/18.11	AoR/481.28
	W-S	RoA/105.02	KLN/42.87	KLN/2768.76
	P-U	TPC/99.53	RoA/16.72	RoA/422.69
	P-S	GNSUM/111.43	GNSUM/ 36.52	GNSUM/ 2531.72

Table 2.S.46: Best method for each scenario. For each metric, it shows the average of the values corresponding to each parameter. All estimators are included in the simulations.

Chapter 3

Error Bounds for the Network Scale-Up Method

In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, 498-509, 2025. https://doi.org/10.1145/3711896.3736940

Sergio Díaz-Aranda[1,2] Juan Marcos Ramírez[1] Mohit Daga[3]
Jaya Prakash Champati[4] José Aguilar[1,3] Rosa E. Lillo[2]
Antonio Fernández Anta[6,1]

[1] IMDEA Networks Institute, Spain
[2] Universidad Carlos III de Madrid, Spain
[3] KTH Royal Institute of Technology, Sweden
[4] University of Victoria, Canada
[5] Universidad de Los Andes, Venezuela
[6] IMDEA Software Institute, Spain

Abstract

Epidemiologists and social scientists have used the Network Scale-Up Method (NSUM) for over thirty years to estimate the size of a hidden sub-population within a social network. This method involves querying a subset of network nodes about the number of their neighbors belonging to the hidden sub-population. In general, NSUM assumes that the social network topology and the hidden sub-population distribution are well-behaved; hence, the NSUM estimate is close to the actual value. However, bounds on NSUM estimation errors have not been analytically proven. This paper provides analytical bounds on the error incurred by the two most popular NSUM estimators. These bounds assume that the queried nodes accurately provide their degree and the number of neighbors belonging to the hidden sub-population. Our key findings are twofold. First, we show that when an adversary designs the network and places the hidden sub-population, then the estimate can be a factor of $\Omega(\sqrt{n})$ off from the real value (in a network with n nodes). Second, we also prove error bounds when the underlying network is randomly generated, showing that a small constant factor can be achieved with high probability using samples of logarithmic size $O(\log n)$. We present improved analytical bounds for Erdős–Rényi and Scale-Free networks. Our theoretical analysis is supported by an extensive set of numerical experiments designed to determine

the effect of the sample size on the accuracy of the estimates in both synthetic and real networks.

3.1. Introduction

In a society, estimating the size of hidden or hard-to-reach sub-populations (such as individuals affected by disaster, disease, or clandestine networks) is critical for developing targeted strategies to address the challenges or threats associated with these hidden groups. Direct enumeration methods for estimating the hidden sub-population size are often impractical due to these groups' dispersed, unreachable, or secretive nature. The Network Scale-Up Method (NSUM) (Bernard et al., 1988, 1991) is an innovative solution that addresses this challenge by leveraging the social network to estimate the size. Instead of directly querying individuals whether they belong to the hidden population, NSUM uses *indirect reporting*. NSUM collects data through queries of the type "How many neighbors do you know?" and "How many of those neighbors belong to the hidden sub-population?" The answers to these questions form the Aggregated Relational Data (ARD), which is used to estimate the size of the hidden sub-population. Using indirect reporting increases the privacy and the proportion of the population a survey reaches, which, combined with the simplicity of the NSUM estimator, makes it very powerful. Thus, it has been widely used in applications including estimating the number of affected people during disasters (Bernard et al., 1988, 2001) and epidemic prevalence and evolution (e.g., AIDS (UNAIDS, 2010), COVID-19 (Garcia-Agundez et al., 2021)).

There is a suite of NSUM methods that have been used in the literature (Zheng et al., 2006; McCormick et al., 2010; McCormick and Zheng, 2012; Maltiel et al., 2015), and they differ in the way they use the ARD to provide the estimates (Laga et al., 2021). The most popular NSUM estimators are the Mean of Ratios (MoR) (Killworth et al., 1998a) and Ratio of Means (RoS) (Killworth et al., 1998b) estimators. While MoR was the first estimator proposed for prevalence (Killworth et al., 1998a; Laga et al., 2021), relying on averaging the naive estimates of each individual, RoS became the more widely used estimator in the NSUM literature (Ahmadi-Gohari et al., 2019; Ezoe et al., 2012; Killworth et al., 1998b; Snidero et al., 2012; Wang et al., 2015).

Although NSUM has been widely used, to the best of our knowledge, no existing work provides analytical guarantees on the quality of the aforementioned NSUM estimators. The main challenge in analyzing the NSUM estimate is due to the indirect reporting, which may result in multiplicities in the reported hidden nodes, thereby introducing correlations in the ARD. Recently, the authors in (Chen et al., 2016; Srivastava et al., 2024) have provided analytical bounds from indirect reporting. Chen et al. (2016) estimated the size of the network and the hidden communities, providing analytical guarantees for Erdős-Rényi and Stochastic Block Model topolo-

gies. However, they assume that the number of hidden nodes in the sample is known, which is not the case in ARD. Srivastava et al. (2024) estimated the trend (not the value) of the hidden sub-population prevalence over time. They show that, if the degree of the social network has a constant mean over time and its variance is small, the trend of the hidden sub-population prevalence can be estimated using the ARD, and the error is smaller than with direct reporting.

Our Contributions This paper partially fills the void of NSUM analytical bounds by providing a theoretical performance analysis for NSUM estimators in general networks. Let ρ denote the fraction of the population that belongs to the hidden sub-population, and $\hat{\rho}$ is its NSUM estimate. The *error* of the estimator $\hat{\rho}$ is how far it is from ρ (as the ratio between them, or equivalently, the factor that transforms one into the other). These are the main results.

- A surprising result we prove is that, in adversarial settings, any NSUM estimator has an error that grows with $\sqrt{n}$, where n is the size of the network, even when the ARD is collected from all the nodes in the network.
- For general random networks with a fixed hidden population, we provide analytically probabilistic upper bounds for MoR and RoS estimators. We address the challenge of the duplicates by proving *negative correlation* between the underlying random variables. The results upper bound the probability that the error is above a fixed threshold $(1+\epsilon)$. Then, we use these bounds to show that, with both estimators, it is enough to collect ARD from a logarithmic number of nodes in a random network to have a small error of $(1+\epsilon)$ with high probability. We also show that RoS has smaller error bounds than MoR.
- As special cases of these general random networks, we study two popular classes of random networks, Scale-Free (Clauset et al., 2009) and Erdős–Rényi networks (Erdős and Rényi, 1959), and compute probabilistic error bounds for MoR and RoS estimates.
- We evaluate the goodness of the bounds using simulation. We observe that the analytical bounds are close to the actual empirical measures. We also observe that, as hinted by the analytical bounds, the estimates obtained with RoS have lower errors than those obtained with MoR, which is especially evident in the Scale-Free networks.
- Finally, we apply the bounds to real networks. We observe that the bounds are still close to the actual empirical measures, even though the networks may not exactly be the type of random networks assumed in the analysis.

3.2. Model and Definitions

Network model. We model a social network as a directed graph $G = (V, E)$[2], in which V is the population under study. A subset $H \subseteq V$ is the hidden sub-population of the network, which is the set of network members G with a given property. A directed edge $(u, v) \in E$ implies that node v knows node u (in-neighbor) and whether it belongs to the hidden sub-population H. A *bidirectional network* is a directed graph, where all edges are bidirectional, i.e., $\forall u, v \in V, (u, v) \in E$ iff $(v, u) \in E$.

Sampling. A subset $S \subseteq V$ of the population will be selected uniformly at random from V and asked to report information about their in-neighbors in G. In the special case of full sampling, we have that $S = V$. Each sampled vertex v reports:

- the number of in-neighbors v has in G, denoted by R_v,
- the number of in-neighbors that belong to the hidden population H, denoted by C_v.

To preserve its privacy, a sampled vertex v does not count itself in R_v and C_v.

Definition 3.1. *We define the following random variables. We assume the in-neighbors of v numbered from 1 to R_v for easy reference.*

- *For each vertex $v \in S$, if the j-th in-neighbor of v, for $j \in \{1, 2, \ldots, R_v\}$, belongs to H, then the indicator random variable $X_{v_j} = 1$; otherwise, $X_{v_j} = 0$. Observe that $C_v = \sum_{j=1}^{R_v} X_{v_j}$.*
- *For each vertex $v \in S$, we define $Y_v \triangleq C_v/R_v$.*

Problem definition. An instance $I = (G, H)$ is a pair of a network and a hidden sub-population. The prevalence of instance I is defined as $\rho(I) \triangleq |H|/|V|$. $I[S]$ represents the set of pairs $\{(C_v, R_v) : v \in S\}$ of the instance I obtained from the sample S.

Definition 3.2 (Problem)**.** *Given an unknown instance $I = (G, H)$ with $n = |V|$ and $h = |H|$, and a sampled vertex set S, we are interested in finding an accurate estimate $\hat{\rho}(I[S])$ of the prevalence $\rho(I) = h/n$.*

NSUM Estimators. The estimators studied in this work for estimating the prevalence $\rho(I)$ of instance $I = (G, H)$ are:

- The *Mean of Ratios* (MoR) estimate is the average of the individual ratios of the sampled vertices, defined as

$$\hat{\rho}_{\text{MoR}}(I[S]) \triangleq \frac{1}{|S|} \sum_{v \in S} \frac{C_v}{R_v} = \frac{1}{|S|} \sum_{v \in S} Y_v.$$

[2]We will indistinctly use graph and network, vertex and node, and edge and link.

- The *Ratio of Sums* (RoS) estimate is the ratio of the sum of the number of hidden neighbors over the sum of in-degrees of the sampled vertices, defined as

$$\hat{\rho}_{\text{RoS}}(I[S]) \triangleq \frac{\sum_{v\in S} C_v}{\sum_{v\in S} R_v} = \frac{\sum_{v\in S}\sum_{j=1}^{R_v} X_{v_j}}{\sum_{v\in S} R_v}.$$

Errors. We define two different types of errors: upper error and lower error. Let $\mathcal{M}$ be any estimation method and $I = (G, H)$ an instance. Let $\hat{\rho}_{\mathcal{M}}(I[S])$ be the estimate of $\rho(I)$ obtained by $\mathcal{M}$. We define $\mathcal{E}^{+}_{\mathcal{M}}(I, S) \triangleq \max\left(1, \frac{\hat{\rho}_{\mathcal{M}}(I[S])}{\rho(I)}\right)$ as the *upper error* and $\mathcal{E}^{-}_{\mathcal{M}}(I, S) \triangleq \max\left(1, \frac{\rho(I)}{\hat{\rho}_{\mathcal{M}}(I[S])}\right)$ as the *lower error.* Note that $\mathcal{E}^{+}_{\mathcal{M}}(I, S)$ quantifies the factor by which $\hat{\rho}_{\mathcal{M}}(I[S])$ is higher than $\rho(I)$, and $\mathcal{E}^{-}_{\mathcal{M}}(I, S)$ quantifies the factor by which $\hat{\rho}_{\mathcal{M}}(I[S])$ is lower than $\rho(I)$. Finally, we define the *error of method $\mathcal{M}$ applied to instance I and sample S* as

$$\begin{aligned}\mathcal{E}_{\mathcal{M}}(I, S) &\triangleq \max\left(\mathcal{E}^{+}_{\mathcal{M}}(I, S), \mathcal{E}^{-}_{\mathcal{M}}(I, S)\right) \\ &= \max\left(\frac{\rho(I)}{\hat{\rho}_{\mathcal{M}}(I[S])}, \frac{\hat{\rho}_{\mathcal{M}}(I[S])}{\rho(I)}\right).\end{aligned}$$

We will denote $\mathcal{E}^{+}_{\mathcal{M}}(I)$, $\mathcal{E}^{-}_{\mathcal{M}}(I)$, and $\mathcal{E}_{\mathcal{M}}(I)$ to the errors $\mathcal{E}^{+}_{\mathcal{M}}(I, V)$, $\mathcal{E}^{-}_{\mathcal{M}}(I, V)$, and $\mathcal{E}_{\mathcal{M}}(I, V)$ corresponding to the full sampling case.

We say that $f(n, h)$ is an upper bound of the error of $\mathcal{M}$, i.e., $\mathcal{E}_{\mathcal{M}} \leq f(n, h)$, if for all $I = (G = (V, E), H)$ such that $|V| = n$ and $|H| = h$, it holds that $\mathcal{E}_{\mathcal{M}}(I, S) \leq f(n, h)$. Conversely, we say that the error of $\mathcal{M}$ is lower bounded by $f(n, h)$, i.e., $\mathcal{E}_{\mathcal{M}} \geq f(n, h)$, if there exists $I = (G, H)$ and sample S such that $\mathcal{E}_{\mathcal{M}}(I, S) \geq f(n, h)$.

Lower Bound for Adversarial Instances. Let us consider any NSUM estimation method $\mathcal{M}$. We allow $\mathcal{M}$ to use all pairs (R_v, C_v), for all $v \in V$, in any manner the method deems fit. We show that there are instances for which the error is $\mathcal{E}_{\mathcal{M}} = \Omega(\sqrt{n})$.

Theorem 3.1. *Any deterministic estimation method $\mathcal{M}$ that only uses the set $\{(R_v, C_v), \forall v \in V\}$ to estimate the prevalence ρ has error $\mathcal{E}_{\mathcal{M}} \geq \sqrt{(n-1)/2}$.*

Proof. We construct a bidirectional network $G = (V, E)$ and pair it with two different hidden population sets $H_1 \subset V$ and $H_2 \subset V$ resulting in instances $I_1 = (G, H_1)$ and $I_2 = (G, H_2)$. These two instances have two different prevalences $\rho(I_1) = |H_1|/n$ and $\rho(I_2) = |H_2|/n$. Our construction ensures that $\{(R_v, C_v), \forall v \in V\}$ is the same for both choices of H_1 and H_2. Hence, the *deterministic* method $\mathcal{M}$ should give the same estimate for both I_1 and I_2, i.e., $\hat{\rho}_{\mathcal{M}}(I_1[V]) = \hat{\rho}_{\mathcal{M}}(I_2[V])$. We show that for all possible deterministic methods $\mathcal{M}$, this value is either far from $\rho(I_1)$ or far from $\rho(I_2)$.

The network G is the union of a set of k nodes (set V_c) forming a clique, a set of k additional nodes (set V_a), each connected to a different clique node (and only

to this node), and one single node s connected to all the clique nodes V_c; see Fig. 3.1. Observe that $n = 2k + 1$, and hence $k = \Theta(n)$. We consider the instances $I_1 = (G, \{s\})$ and $I_2 = (G, V_a)$. In I_1, the set of hidden nodes contains only node s, and hence the prevalence is $\rho(I_1) = 1/n = 1/(2k + 1)$. In I_2, all the k nodes in V_a belong to the hidden population, and hence the prevalence is $\rho(I_2) = k/(2k + 1)$.

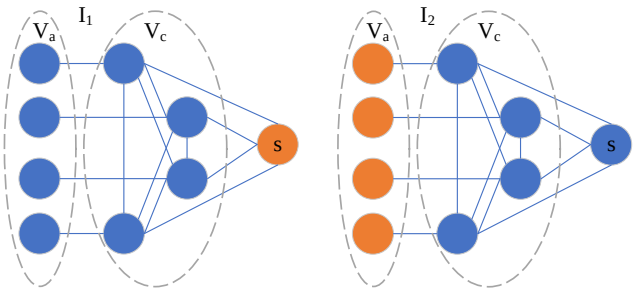


Figure 3.1: Example instances I_1 and I_2 with $k = 4$ and $n = 9$. All the edges are bidirectional, and the hidden nodes are in orange.

However, observe that the pairs (R_v, C_v) reported by all nodes are exactly the same in both I_1 and I_2: the graph G is the same, s and the nodes in V_a have no hidden neighbor (and hence return $C_v = 0$), and in both cases, the nodes in V_c have exactly one hidden neighbor (and hence return $C_v = 1$). Hence, it holds that $\hat{\rho}_{\mathcal{M}}(I_1) = \hat{\rho}_{\mathcal{M}}(I_2)$. Let us denote this estimate as $\hat{\rho}_{\mathcal{M}}$ and its error as $\mathcal{E}_{\mathcal{M}}$.

To bound the error, let us first consider $\hat{\rho}_{\mathcal{M}} \notin [\rho(I_1)), \rho(I_2)]$. There are two cases: if $\hat{\rho}_{\mathcal{M}} < \rho(I_1)$ then $\mathcal{E}_{\mathcal{M}}(I_2) = \rho(I_2)/\hat{\rho}_{\mathcal{M}} > \rho(I_2)/\rho(I_1) = k$; if $\hat{\rho}_{\mathcal{M}} > \rho(I_2)$ then $\mathcal{E}_{\mathcal{M}}(I_1) = \hat{\rho}_{\mathcal{M}}/\rho(I_1) > \rho(I_2)/\rho(I_1) = k$. On the other hand, when $\hat{\rho}_{\mathcal{M}} \in [\rho(I_1)), \rho(I_2)]$,

$$\mathcal{E}^{+}_{\mathcal{M}}(I_1) = \max\left(1, \frac{\hat{\rho}_{\mathcal{M}}}{\rho(I_1)}\right) = (2k + 1)\ \hat{\rho}_{\mathcal{M}},$$

and

$$\mathcal{E}^{-}_{\mathcal{M}}(I_2) = \max\left(1, \frac{\rho(I_2)}{\hat{\rho}_{\mathcal{M}}}\right) = \frac{k}{(2k + 1)\hat{\rho}_{\mathcal{M}}}.$$

The error incurred by method $\mathcal{M}$ is lower bounded by the maximum of $\mathcal{E}^{+}_{\mathcal{M}}(I_1)$ and $\mathcal{E}^{-}_{\mathcal{M}}(I_2)$ for both I_1 and I_2. This maximum value is minimized when $\hat{\rho}_{\mathcal{M}} = \sqrt{k}/(2k+1)$, which implies that $\mathcal{E}_{\mathcal{M}} \geq \max(\mathcal{E}^{+}_{\mathcal{M}}(I_1), \mathcal{E}^{-}_{\mathcal{M}}(I_2)) \geq \sqrt{k}$. From our construction $k = (n - 1)/2$, and the error is at least $\mathcal{E}_{\mathcal{M}} \geq \sqrt{k} = \sqrt{(n - 1)/2}$. □

The lower bound in Thm. 3.1 is surprising because it is valid for the full sampling scenario where $S = V$. This raises the question regarding the goodness of the estimates provided by MoR and RoS, which only use $\{(R_v, C_v), \forall v \in S\}$. We address this question and show that the error can be bounded for these methods for random networks, with a small number of samples. We include two popular classes of random networks, Scale-Free (Clauset et al., 2009) and Erdős–Rényi networks (Erdős and Rényi, 1959).

Random networks. In the sequel, we study instances $I = (G, H)$ in which G is a random network over a fixed population V and a given hidden sub-population $H \subseteq V$. To be precise, the in-neighbors of each vertex $v \in V$ of the network G are selected independently as follows:

- Vertex v selects a degree R_v from a given degree distribution P_{deg}.
- Given R_v, vertex v selects R_v vertices uniformly at random from $V \setminus \{v\}$ to be its set of in-neighbors.

This construction is general in that we build different graph families by choosing different degree distributions. As special cases of these general random networks, we study two popular classes of random networks, Scale-Free (Clauset et al., 2009) and Erdős–Rényi networks (Erdős and Rényi, 1959). A directed Scale-Free network is a random graph in which the fraction of nodes with in-degree k is proportional to $k^{-\gamma}$, where γ is a small constant (typically, $\gamma \in (2, 3)$). A directed random graph generated with the Erdős–Rényi model $G(n, p)$ has n nodes, and for any two nodes $u, v \in V$, the edge (u, v) is included in the edge set E with probability p, independently of any other edge.

Note that in random networks, the random variables Y_v, for all $v \in S$, are dependent. Likewise, X_{j_v}, for all $v \in S$ and $j \in \{1, 2, \ldots, R_v\}$, are also dependent. Thus, the direct application of concentration bounds to obtain upper bounds for the errors is not possible. We address this challenge by using conditional independence in the random networks and showing a negative correlation for the above variables.

3.3. Upper Bounds on the Error of MoR and RoS in Random Networks

For ease in exposition and conciseness in the expressions, in this section, we prove results for random networks in which the support of the degree distribution P_{deg} is positive integers, i.e., the degree of any node is non-zero. We state it in the following assumption:

Assumption 3.1. *$R_v \geq 1$ for all $v \in V$.*

For random networks without the above assumption, but with bounded probability of zero degree, one can modify the error bounds derived in this section to accommodate those scenarios. The omitted proofs can be found in the Appendix.

3.3.1. Upper Bound on the Error of the MoR Estimator

Recall that $\hat{\rho}_{\text{MoR}}(I[S]) = \frac{1}{|S|} \sum_{v \in S} \frac{C_v}{R_v} = \frac{1}{|S|} \sum_{v \in S} Y_v$. We first show that, for random networks, $\hat{\rho}_{\text{MoR}}(I[S])$ is an *unbiased estimate* of $\rho(I)$, i.e., $\mathbb{E}[\hat{\rho}_{\text{MoR}}(I[S])] = \mathbb{E}[Y_v] = \rho(I)$. Since a vertex v cannot select itself as an in-neighbor (i.e., there is no self-reporting), the distribution of the random variables X_{vj} depends on whether v is

in H. Using combinatorial arguments, we show that $\mathbb{E}[X_{v_j} \mid v \in H] = \frac{h-1}{n-1}$, and $\mathbb{E}[X_{v_j} \mid v \notin H] = \frac{h}{n-1}$, for all $j \in \{1, 2, \dots R_v\}$. From this result, we obtain that the expectations of X_{vj} and Y_v are equal to $\rho(I)$.

Lemma 3.1. *For all $v \in S$ and $j \in \{1, 2, \dots, R_v\}$, $\mathbb{E}[X_{v_j}] = \mathbb{E}[Y_v] = \frac{h}{n} = \rho(I)$.*

We note that the variables Y_v are not independent if we do not know the membership of v in H. Consider, for example, the case where $|H| = 1$ and two vertices $v_1, v_2 \in S$. Then,

$$\Pr[Y_{v_2} > 0 \mid Y_{v_1} > 0] < \Pr[Y_{v_2} > 0],$$

because if $Y_{v_1} > 0$, it is sure that $v_1 \notin H$. This increases the probability of $v_2 \in H$, reducing the probability that $Y_{v_2} > 0$. However, we show that variables Y_v, for all $v \in S$, are negatively correlated. Using this and the fact that $\mathbb{E}[Y_v] = \rho(I)$, we apply a concentration bound for negatively correlated random variables to obtain an upper bound for the error stated in the following theorem. Let us define the following function

$$F(x, y) = \left(\frac{e^{x-1}}{x^x}\right)^y + \left(\frac{e^{\frac{1}{x}-1}}{x^{-1/x}}\right)^y.$$

Theorem 3.2. *Consider an instance $I = (G, H)$ where G is a random network. If $S \subseteq V$ is sampled uniformly at random, with $m = |S|$, then the* MoR *estimator has the following error bound for any $\beta = 1 + \epsilon > 1$.*

$$\Pr\left[\mathcal{E}_{\text{MoR}}(I, S) > \beta\right] \leq F(\beta, m\rho(I)).$$

3.3.2. Upper Bound on the Error with the RoS Estimator

For the RoS estimator, we define the random variables $R_S = \sum_{v \in S} R_v$ and $C_S = \sum_{v \in S} C_v = \sum_{v \in S} \sum_{j=1}^{R_v} X_{vj}$. By definition, the estimator is $\hat{\rho}_{\text{RoS}}(I[S]) = C_S / R_S$. It was already shown in Lemma 3.1 that the expectation of X_{vj} is equal to $\rho(I)$. We can also show that the variables X_{vj} are negatively correlated. Then, we can use a concentration bound to prove the following theorem.

Theorem 3.3. *Consider an instance $I = (G, H)$ where G is a random network. If $S \subseteq V$ is sampled uniformly at random, with $m = |S|$, then for $R_S = \sum_{v \in S} R_v$ and any $\beta = 1 + \epsilon > 1$, the* RoS *estimator has the following error bound.*

$$\Pr\left[\mathcal{E}_{\text{RoS}}(I, S) > \beta\right] \leq \sum_{R} F(\beta, R\rho(I)) \Pr\left[R_S = R\right].$$

Corollary 3.1. *Consider an instance $I = (G, H)$ where G is a random network G. If $S \subseteq V$ is sampled uniformly at random, with $m = |S|$, then the* RoS *estimator has the following error bound, for any $\beta = 1 + \epsilon > 1$.*

$$\Pr\left[\mathcal{E}_{\text{RoS}}(I, S) > \beta\right] \leq F(\beta, m\rho(I)).$$

This bound matches the bound for MoR in Thm. 3.2. Since, in general, we expect $R_S > |S|$, the bound for RoS in Thm. 3.3 is usually better than the bound in Thm. 3.2 for MoR.

3.3.3. Sample Size

From Thm. 3.2 and Cor. 3.1, we can show that it is possible to achieve an error of at most $\beta = 1 + \epsilon > 1$ with high probability using a logarithmic sample set.

Theorem 3.4. *Given an instance $I = (G, H)$ where G is a random network. If G is sampled uniformly at random, with*

$$m = |S| \geq \frac{\ln 2 + \alpha \ln n}{\rho(I)(1 - \frac{1}{\beta}(\ln \beta + 1))},$$

then the error of MoR *and* RoS *estimators is no larger than $\beta = 1 + \epsilon > 1$ with probability at least $1 - 1/n^\alpha$, for any constant $\alpha > 0$.*

Proof. We show that $\left(\frac{e^{\beta-1}}{\beta^\beta}\right)^{m\rho(I)} \leq 1/(2n^\alpha)$ and $\left(\frac{e^{\frac{1}{\beta}-1}}{\beta^{-1/\beta}}\right)^{m\rho(I)} \leq 1/(2n^\alpha)$, which implies the claim from Theorem 3.2 and Corollary 3.1. Hence, m must simultaneously satisfy

$$\begin{aligned}
\left(\frac{e^{\beta-1}}{\beta^\beta}\right)^{m\rho(I)} &\leq 1/(2n^\alpha) \\
m\rho(I) \ln \left(\frac{e^{\beta-1}}{\beta^\beta}\right) &\leq -(\ln 2 + \alpha \ln n) \\
m \geq \frac{\ln 2 + \alpha \ln n}{\rho(I) \ln \left(\frac{\beta^\beta}{e^{\beta-1}}\right)} &= \frac{\ln 2 + \alpha \ln n}{\rho(I)(\beta \ln \beta - \beta + 1)},
\end{aligned}$$

and

$$\begin{aligned}
\left(\frac{e^{\frac{1}{\beta}-1}}{\beta^{-1/\beta}}\right)^{m\rho(I)} &\leq 1/(2n^\alpha) \\
m\rho(I) \ln \left(\frac{e^{\frac{1}{\beta}-1}}{\beta^{-1/\beta}}\right) &\leq -(\ln 2 + \alpha \ln n) \\
m \geq \frac{\ln 2 + \alpha \ln n}{\rho(I) \ln \left(\frac{\beta^{-1/\beta}}{e^{\frac{1}{\beta}-1}}\right)} &= \frac{\ln 2 + \alpha \ln n}{\rho(I)(-\frac{1}{\beta} \ln \beta - \frac{1}{\beta} + 1)},
\end{aligned}$$

It holds that $\beta \ln \beta - \beta + 1 \geq -\frac{1}{\beta} \ln \beta - \frac{1}{\beta} + 1 > 0$ for $\beta > 1$. Hence, the bound on m satisfies that $\left(\frac{e^{\beta-1}}{\beta^\beta}\right)^{m\rho(I)} \leq 1/(2n^\alpha)$ and $\left(\frac{e^{\frac{1}{\beta}-1}}{\beta^{-1/\beta}}\right)^{m\rho(I)} \leq 1/(2n^\alpha)$, and hence $\Pr[\mathcal{E}(I, S) > \beta] \leq 1/n^\alpha$. □

The bound on the sample size provided by this theorem is $O(\log n)$ as long as α, β, and $\rho(I)$ are constant. Observe that α and β can be freely chosen, but the prevalence $\rho(I)$ is not known. Instead, a lower bound on $\rho(I)$ has to be assumed and used to obtain the size m in practice.

3.4. Upper Bounds for Erdős–Rényi and Scale-Free Networks

3.4.1. Erdős–Rényi Network

Consider an instance $I = (G, H)$ where the directed network $G = (V, E)$ is generated with the Erdős–Rényi model $G(n, p)$. In this network, a directed link (u, v) between nodes $u, v \in V$ is created independently at random with probability p. This is equivalent to assigning to each vertex $v \in V$ an in-degree $R_v \sim Binomial(n-1, p)$ and selecting the in-neighbors of v uniformly at random from $V \setminus \{v\}$. Unfortunately, Assumption 3.1 does not hold for Erdős–Rényi networks since there may be vertices with in-degree zero. To prevent this, we assume that zero-degree nodes are not allowed, and distribute the probability of that event among the rest of the in-degrees proportionally.

The bound for the MoR **estimator.** We can directly apply Theorem 3.2.

Theorem 3.5. *Consider an instance $I = (G, H)$ where the directed random network $G = (V, E)$ is generated with the Erdős–Rényi model $G(n, p)$ without zero in-degree nodes. If $S \subseteq V$ is sampled uniformly at random, with $m = |S|$, then the* MoR *estimator has the following error bound, for any $\beta = 1 + \epsilon > 1$.*

$$\Pr\left[\mathcal{E}_{\text{MoR}}(I, S) > \beta\right] \leq F(\beta, m\rho(I))$$

The bound for the RoS **estimator.** Let us now consider the RoS estimator. Observe that by construction of the Erdős–Rényi model $G(n, p)$ without zero in-degree nodes, the average in-degree is no less than $p(n-1)$.

Theorem 3.6. *Consider an instance $I = (G, H)$ where the directed random network $G = (V, E)$ is generated with the Erdős–Rényi model $G(n, p)$ without zero in-degree nodes, and a set $S \subseteq V$ is sampled uniformly at random, with $m = |S|$. Then, the* RoS *estimator has the following error bound, for any $\delta \in (0, 1)$ and any $\beta = 1 + \epsilon > 1$ and where $\mu = mp(n-1)$.*

$$\Pr\left[\mathcal{E}_{\text{RoS}}(I, S) > \beta\right] \leq \left(\frac{e^{-\delta}}{(1-\delta)^{(1-\delta)}}\right)^{\mu} + F(\beta, (1-\delta)\mu\rho(I)).$$

3.4.2. Scale-Free Network

In a Scale-Free random network $G = (V, E)$, the probability that a node has in-degree $k \geq 1$, independent of other nodes, is given by a degree distribution $P_{\deg}(k) = \nu k^{-\gamma}$,

where $\nu = 1/\sum_{k=1}^{n-1} k^{-\gamma}$ is a constant that ensures that $\sum_{k=1}^{n-1} P_{\text{deg}}(k) = 1$. Note that for this network, there are no nodes with in-degree zero, and hence the Assumption 3.1 is true, and the results in Section 3.3 can be directly applied. We adopt the Scale-Free network model in which for any vertex v with in-degree k, the k in-neighbors of v are picked uniformly at random from $|V| \setminus \{v\}$. Let us consider an instance $I = (G, H)$ in which the random network G is a scale-network as described above.

The bounds for the MoR **estimator.** For the MoR estimator, the error bound we obtain is the one provided by Thm. 3.2.

The bound for the RoS **estimator.** For the RoS estimator, the error bound provided by Cor. 3.1 also applies. However, in this case, we obtain a better bound by using Thm. 3.3. Recall that Scale-Free networks usually have $\gamma \in (2, 3)$. The result below assumes $\gamma > 2$ to avoid a negative value of μ.

Theorem 3.7. *Consider an instance $I = (G, H)$ in which G is a Scale-Free random network with $\gamma > 2$. If $S \subseteq V$ is sampled uniformly at random, with $m = |S|$, then the* RoS *estimator has the following error bound, for any $\delta \in (0, 1)$, and any $\beta = 1 + \epsilon > 1$ and where $\mu \approx m \frac{1-\gamma}{2-\gamma} \frac{1-(n-1)^{2-\gamma}}{1-(n-1)^{1-\gamma}}$.*

$$\Pr\left[\mathcal{E}_{\text{RoS}}(I, S) > \beta\right] \leq \left(\frac{e^{-\delta}}{(1-\delta)^{(1-\delta)}}\right)^{\mu} + F(\beta, (1-\delta)\mu\rho(I)).$$

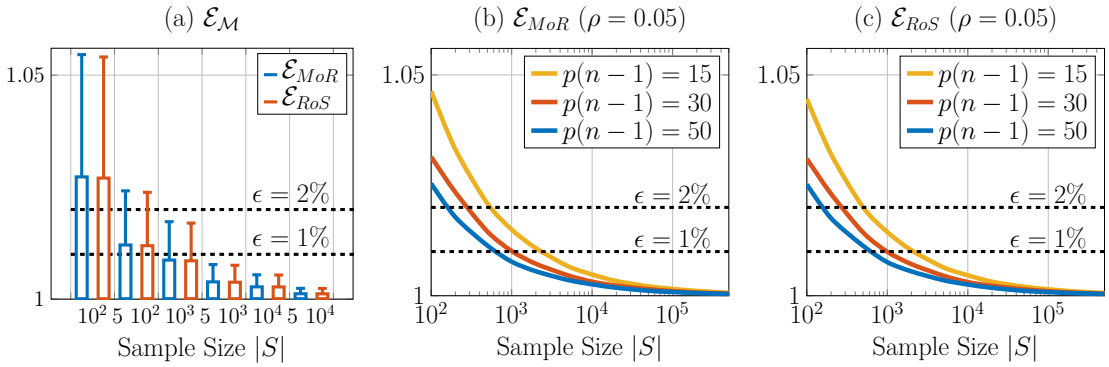


Figure 3.2: Erdős–Rényi ($n = 10^6$). (a) Boxplots $\mathcal{E}_{\mathcal{M}}$ versus $|S|$ for MoR and RoS for $\rho = 0.05$ and $p(n-1) = 30$. (b) Average $\mathcal{E}_{\text{MoR}}$ versus $|S|$ for $\rho = 0.05$ and different $p(n-1)$. (c) Average $\mathcal{E}_{\text{RoS}}$ versus $|S|$ for $\rho = 0.05$ and different $p(n-1)$.

3.5. Simulations

We evaluate the behavior of MoR and RoS estimators under different conditions using two network models: Erdős–Rényi and Scale-Free. For a given ρ, we have generated 100 instances $I = (G, H)$ with that prevalence for each network model. Then, we obtained 200 sample sets with a particular size $|S|$ from each instance.

For each graph instance and sample set, we estimated prevalence with $\hat{\rho}_{\mathrm{MoR}}(I)$ and $\hat{\rho}_{RoS}(I)$, and computed the respective errors $\mathcal{E}_{\mathrm{MoR}}(I)$ and $\mathcal{E}_{\mathrm{RoS}}(I)$. Simulation codes were implemented in MATLAB and executed on a Dell Inspiron 14 7000 laptop with Intel Core i7 CPU 2.80 GHz, RAM 16 GB, and Ubuntu 22.04 OS.

3.5.1. Erdős–Rényi

We assess the errors $\mathcal{E}_{\mathcal{M}}$ using the Erdős–Rényi network model. Each generated graph has a size of $n = 10^6$ and a probability parameter p, where the mean in-degree is given by $p(n-1)$. Specifically, we evaluate the performance of the estimators for different mean in-degrees.

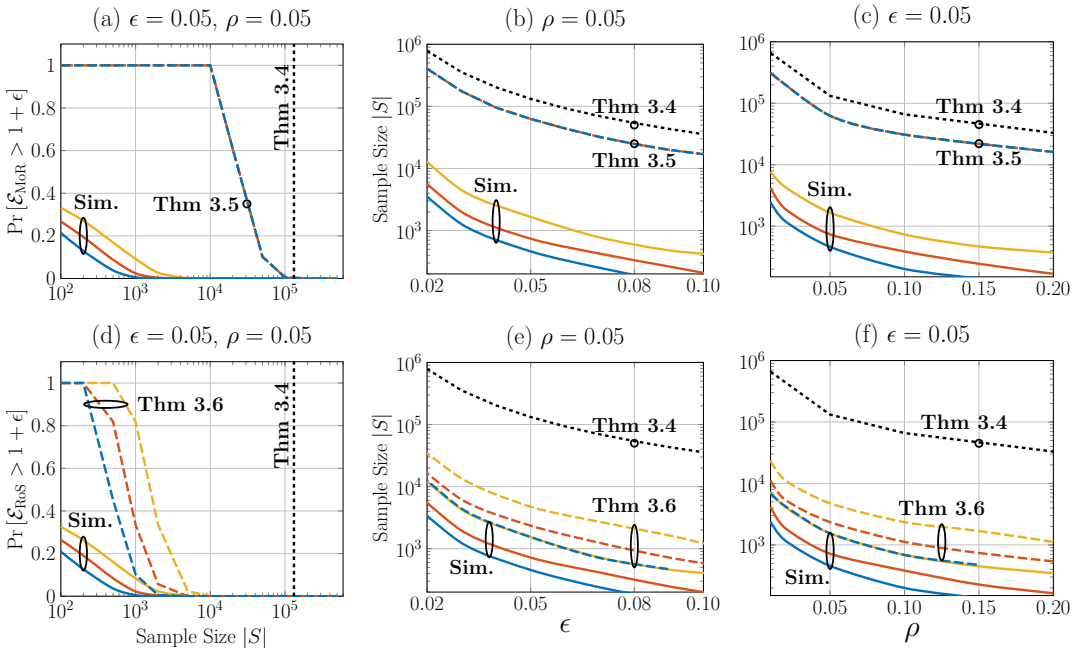


Figure 3.3: Erdős–Rényi ($n = 10^6$). Simulation curves and bounds for (—) $p(n-1) = 15$, (—) $p(n-1) = 30$, and (—) $p(n-1) = 50$. (a) $\Pr[\mathcal{E}_{\mathrm{MoR}} > 1+\epsilon]$ versus $|S|$ for $\epsilon = 0.05$ and ρ. (b) Sample size versus ϵ for $\rho = 0.05$. (c) Sample size versus ρ for $\epsilon = 0.05$. (d) $\Pr[\mathcal{E}_{\mathrm{RoS}} > 1+\epsilon]$ versus $|S|$ for $\epsilon = 0.05$ and $\rho = 0.05$. (e) Sample size versus ϵ for $\rho = 0.05$. (f) Sample size versus ρ for $\epsilon = 0.05$.

Error versus sample size. We analyze the behavior of $\mathcal{E}_{\mathcal{M}}$ in relation to the sample size $|S|$. Fig. 3.2(a) displays the boxplots of $\mathcal{E}_{\mathrm{MoR}}$ and $\mathcal{E}_{\mathrm{RoS}}$ as function of the sample size $|S|$ for a mean in-degree of $p(n-1) = 30$ and $\rho = 0.05$ (5%). These boxplots indicate that $\mathcal{E}_{\mathrm{MoR}}$ and $\mathcal{E}_{\mathrm{RoS}}$ decrease as $|S|$ increases. For $|S| > 500$, most estimated errors are below 1.05 ($\epsilon = 5\%$). Fig. 3.2(b) and 3.2(c) show the average $\mathcal{E}_{\mathrm{MoR}}$ and $\mathcal{E}_{\mathrm{RoS}}$, respectively, versus $|S|$ for different mean in-degrees $p(n-1) = 15$, 30, and 50, with $\rho = 0.05$. These figures include the threshold levels for $\epsilon = 1\%$ and 2% for easy reference. Average errors decrease exponentially with increasing sample size. In

addition, higher mean in-degrees lead to lower errors, showing that the estimation accuracy improves with the increasing number of neighbors. Notice that the average $\mathcal{E}_{\text{MoR}}$ and $\mathcal{E}_{\text{RoS}}$ perform similarly for the Erdős–Rényi graph. In summary, the findings indicate that larger sample size, higher hidden population prevalence, and higher mean in-degree improve accuracy in estimating hidden population prevalence in an Erdős–Rényi network.

Analysis of $\mathcal{E}_{\text{MoR}}$. Then, we assess the behavior of $\mathcal{E}_{\text{MoR}}$ using the Erdős–Rényi model. Fig. 3.3(a) shows the probability $\Pr[\mathcal{E}_{\text{MoR}} > 1+\epsilon]$ versus $|S|$ for $\epsilon = 0.05$, $\rho = 0.05$, and $p(n-1) = 15$, 30, and 50. This figure compares simulation results with the analytical bounds from Thm. 3.4 and 3.5. The bound derived in Thm. 3.4 is obtained using $\alpha = 1/2$. Fig. 3.3(b) illustrates the sample size required to achieve $\Pr[\mathcal{E}_{\text{MoR}} > 1+\epsilon] = 0.05$ versus ϵ for $\rho = 0.05$ and different mean in-degrees. Fig. 3.3(c) displays the sample size required to achieve $\Pr[\mathcal{E}_{\text{MoR}} > 1+\epsilon] = 0.05$ versus ρ for $\epsilon = 0.05$ and various mean in-degrees. The required sample size decreases as ρ increases, indicating that higher ρ leads to more accurate estimates. Figs 3.3(a)-(c) indicate that estimator performance improves with increasing mean in-degree. A higher mean in-degree results in better coverage of the target population; therefore, more information is available to estimate ρ. Finally, the bounds from theorems 3.4 and 3.5 are conservative concerning the simulation results.

Analysis of $\mathcal{E}_{\text{RoS}}$. We examine the behavior of $\mathcal{E}_{\text{RoS}}$ using the Erdős–Rényi model. To this end, we consider the analytical bound in Thm. 3.6. Fig. 3.3(d)-(f) illustrate the behavior of $\mathcal{E}_{\text{RoS}}$ in various scenarios. The bound derived in Thm. 3.6 depends on the mean in-degree, and the resulting bounds are closer to the simulation results than those obtained by $\mathcal{E}_{\text{MoR}}$.

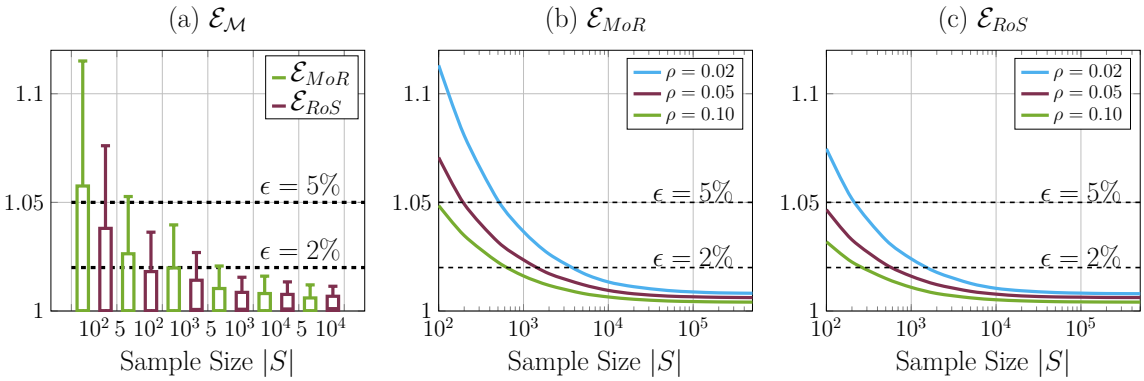


Figure 3.4: Scale Free ($n = 10^6$). (a) Boxplots of $\mathcal{E}_{\mathcal{M}}$ versus $|S|$ for MoR and RoS. (b) Average $\mathcal{E}_{\text{MoR}}$ versus $|S|$ for different ρ. (c) Average $\mathcal{E}_{\text{RoS}}$ versus $|S|$ for different ρ.

3.5.2. Scale-Free

In this section, we evaluate the accuracy of the estimators using scale-free graphs with size $n = 10^6$.

Error versus sample size. Fig. 3.4(a) shows the boxplots of $\mathcal{E}_{\text{MoR}}$ and $\mathcal{E}_{\text{RoS}}$ versus $|S|$ for $\rho = 0.05$ (5%). Errors decrease with an increase in $|S|$. Additionally, the RoS estimator yields lower errors than the MoR estimator across various sample sizes. Fig. 3.4(b) and (c) illustrate the average $\mathcal{E}_{\text{MoR}}$ and $\mathcal{E}_{\text{RoS}}$, respectively, versus $|S|$ for $\rho = 0.02$, 0.05, and 0.10. The results show that higher ρ leads to lower estimation errors, indicating improved estimator performance. Moreover, the $\hat{\rho}_{\text{RoS}}$ estimator demonstrates better accuracy than the $\hat{\rho}_{\text{MoR}}$ estimator as the sample size increases. These findings suggest that larger sample sizes and higher hidden population prevalence lead to more accurate estimations, with the RoS estimator providing superior performance.

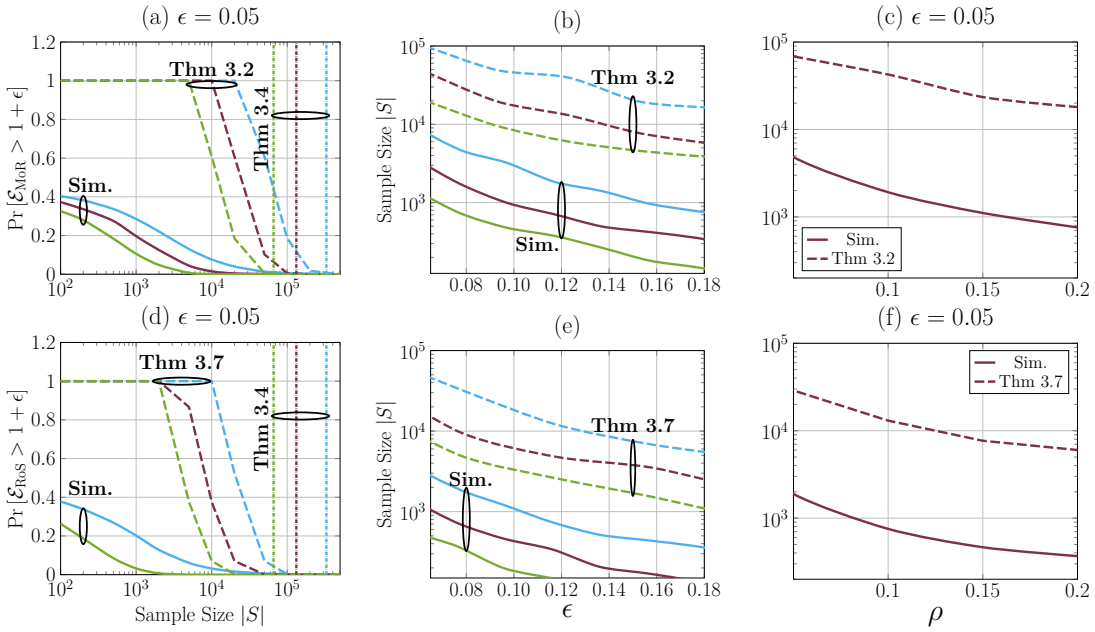


Figure 3.5: Scale Free ($n = 10^6$). Simulation curves and bounds for (—) $\rho = 0.02$, (—) $\rho = 0.05$, and (—) $\rho = 0.10$. (a) $\Pr[\mathcal{E}_{\text{MoR}} > 1 + \epsilon]$ versus $|S|$ for $\epsilon = 0.05$. (b) Sample size versus ϵ. (c) Sample size versus ρ for $\epsilon = 0.05$. (d) $\Pr[\mathcal{E}_{\text{RoS}} > 1 + \epsilon]$ versus $|S|$ for $\epsilon = 0.05$. (e) Sample size versus ϵ. (f) Sample size versus ρ for $\epsilon = 0.05$.

Analysis of $\mathcal{E}_{\text{MoR}}$. In this case, we analyze the behavior of $\mathcal{E}_{\text{MoR}}$ in the Scale-Free model under various conditions. Fig. 3.5(a) shows the probability $\Pr[\mathcal{E}_{\text{MoR}} > 1 + \epsilon]$ versus $|S|$ for $\epsilon = 0.05$ and different actual $\rho = 0.02$, 0.05, and 0.10. This figure includes the theoretical bounds from Thms 3.2 and 3.4, with Thm. 3.4 bounds obtained using $\alpha = 1/2$. The bounds from Thm. 3.4 are close to the simulation results. Fig. 3.5(b) displays the sample size required to achieve $\Pr[\mathcal{E}_{\text{MoR}} > 1+\epsilon] = 0.05$ versus ϵ for different ρ. The sample sizes required from Thm. 3.2 are approximately

one order of magnitude higher than the simulations. Fig. 3.5(c) shows the required sample size versus the actual ρ for $\Pr[\mathcal{E}_{\text{MoR}} > 1 + \epsilon] = 0.05$.

Analysis of $\mathcal{E}_{\text{RoS}}$. We evaluate the behavior of $\mathcal{E}_{\text{RoS}}$ using the Scale-Free network model, employing the bounds derived in Thm. 3.4 and 3.7. Fig. 3.5(d) shows the probability $\Pr[\mathcal{E}_{\text{RoS}} > 1 + \epsilon]$ versus sample size $|S|$ for $\epsilon = 0.05$ and $\rho = 0.02$, 0.05, and 0.10. The bounds from Thm. 3.4 and 3.7 are compared to the simulation results. Fig. 3.5(e) illustrates the required sample size to achieve $\Pr[\mathcal{E}_{\text{RoS}} > 1 + \epsilon] = 0.05$ versus ϵ for $\rho = 0.05$. The results indicate that the sample sizes predicted by Thm. 3.7 are slightly higher than those obtained from simulations. Fig. 3.5(f) presents the required sample size versus ρ for $\Pr[\mathcal{E}_{\text{RoS}} > 1 + \epsilon] = 0.05$ and $\epsilon = 0.05$.

3.6. Results with Real Networks

Dataset. This section evaluates the performance of the hidden population rate estimators by using (undirected) friendship networks extracted from Deezer, a popular music streaming platform, in November 2017. Specifically, these datasets contain friendship networks of the Deezer users from three European countries: Hungary, Romania, and Croatia. In these datasets, each node represents a Deezer user, and each edge denotes a mutual friendship between users. Moreover, based on their liked song lists, each node is labeled with a set of music genres (from a list of 84 music genres) they have shown interest in. These datasets were collected by Rozemberczki et al. (2020) as a part of their study on graph embedding techniques. Additionally, these datasets are publicly available through the Stanford Large Network Dataset Collection: `https://snap.stanford.edu/data/gemsec-Deezer.html`.

Table 3.1: Summary of characteristics of friendship networks from Deezer dataset (Rozemberczki et al., 2020).

	Network		
	Croatia	Hungary	Romania
Nodes	54,573	41,538	41,773
Edges	498,202	222,887	125,826
Avg Node Deg	18.26	9.38	6.02

Table 3.1 summarizes the characteristics of the friendship networks extracted from the Gemsec Deezer dataset. These networks comprise roughly between 41,000 and 55,000 nodes, and between 125,000 and 500,000 edges. Notice that the networks exhibit different averages of node degree. Fig. 3.6 illustrates the node degree distributions for the friendship networks obtained from the three countries.

Analysis of $\mathcal{E}_{\text{MoR}}$. Our study identified user groups interested in particular music genres. For example, we have chosen user sets with preferences for soundtracks, comedy, singer & songwriter, film games, alternative, and dance as hidden populations in Hungary. We selected the music genres for each country to encompass a range of

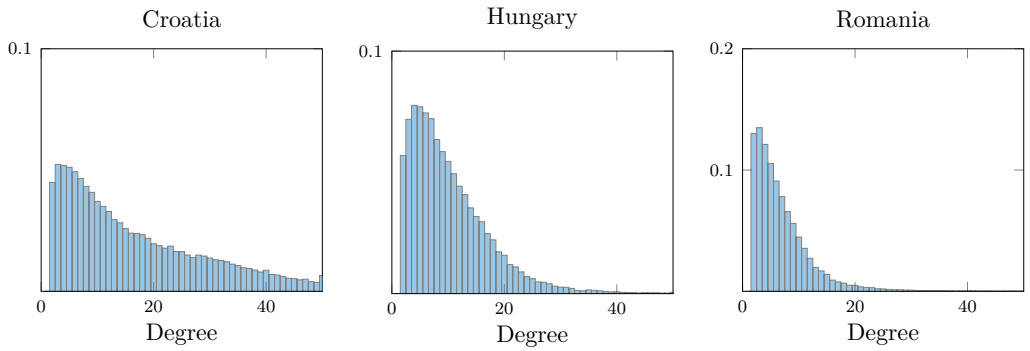


Figure 3.6: Deezer Music Dataset. Histograms of the node degree for the networks obtained from Croatia, Hungary, and Romania.

hidden population rates from 1% to 50%. Our approach involves utilizing varying sample sizes for each hidden population rate, and for each sample size, we generate 10,000 realizations of the sampling process. We calculate the hidden population rate at each trial using the MoR estimator and the corresponding $\mathcal{E}_{\text{MoR}}$. Fig. 3.7 shows the probabilities $\Pr[\mathcal{E}_{\text{MoR}} > 1+\epsilon]$ as a function of the sample size $|S|$ obtained through simulations, Thm. 3.2, and Thm. 3.4 for $\epsilon = 10\%$ and the different music genres selected from the Hungary dataset. Notice that the sample size bound in Thm. 3.4 is calculated using $\alpha = 1/2$. For illustrative purposes, the full version (Díaz-Aranda et al., 2024) displays $\Pr[\mathcal{E}_{\text{MoR}} > 1+\epsilon]$ versus $|S|$ for $\epsilon = 10\%$ for Croatia and Romania. As can be seen in these figures, bounds from Thm. 3.2 and Thm. 3.4 are conservative concerning probabilities obtained from simulations.

Analysis of $\mathcal{E}_{\text{RoS}}$. We have utilized the RoS estimator to calculate hidden population rates and determined the corresponding $\mathcal{E}_{\text{MoR}}$. In Fig. 3.7 , we have also depicted the probabilities $\Pr[\mathcal{E}_{\text{RoS}} > 1+\epsilon]$ against $|S|$ obtained from simulations and Thm. 3.3 for $\epsilon = 0.10$ and the music genres selected from the Hungary dataset. To derive the bound as defined in Theorem 3.3, we have used the probability mass distribution of the variable R_S based on 10,000 realizations of the sampling process for each sample size. Histograms of R_S for different countries and sample sizes, as well as plots of $\Pr[\mathcal{E}_{\text{RoS}} > 1+\epsilon]$ versus $|S|$ for $\epsilon = 10\%$ for Croatia and Romania, are shown in the full version (Díaz-Aranda et al., 2024). In this case, the bounds derived from Thm. 3.3 are slightly higher than those obtained from simulations.

3.7. Conclusions and Future Work

In this paper, we derived analytical error bounds for NSUM hidden population rate estimators. The summary of the derived analytical bounds is presented in Table 3.2. Notice that these analytical bounds were determined assuming that sampled nodes accurately report their in-degree and the number of neighboring nodes that belong to the hidden population. Specifically, in adversarial scenarios, the error can remain

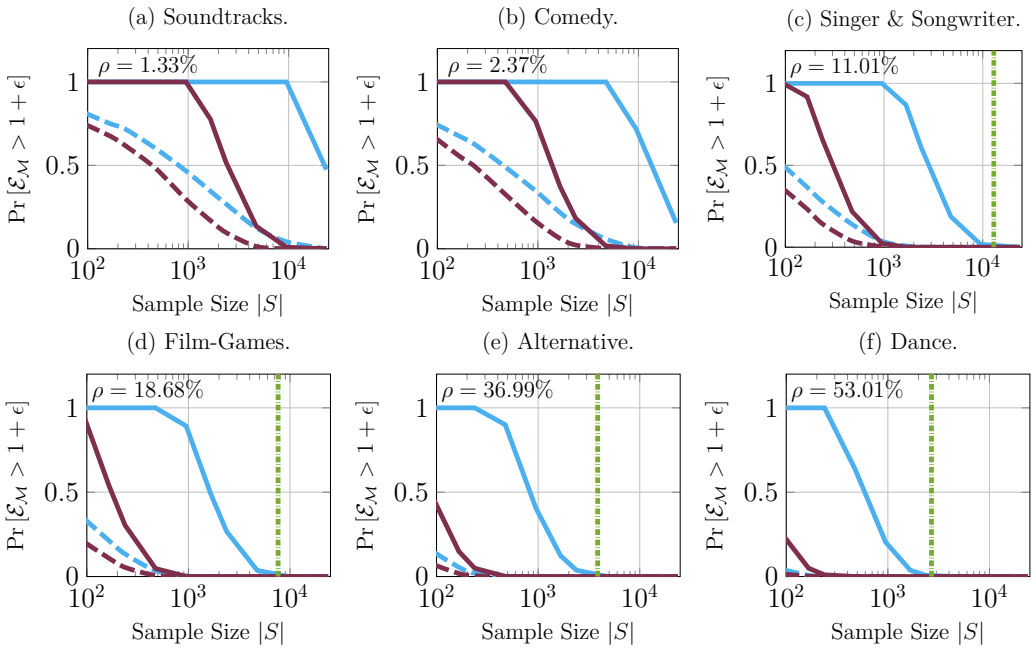


Figure 3.7: Gemsec Deezer dataset. Hungary. ($n = 47,538$). $\Pr[\mathcal{E}_{\mathcal{M}} > 1 + \epsilon]$ versus $|S|$ obtained using (▬ ▬) MoR, (▬ ▬) RoS,(▬) Thm 3.2, (▬) Thm. 3.3, and (▬ ▪ ▬) Thm. 3.4 for $\epsilon = 0.10$ and music genres: (a) Soundtracks, (b) Comedy, (c) Singer & Songwriter, (d) Film-Games, (e) Alternative, and (f) Dance.

high even when the ARD of all nodes is available. For such cases, we establish a lower bound. In addition, we have derived analytical error upper bounds for two NSUM hidden population rate estimators (the Mean of Ratios, MoR, and the Ratio of Sums, RoS) in random networks and provided the analytical error bounds for Erdős–Rényi and Scale-Free networks.

On the other hand, extensive numerical simulations have been conducted on synthetic and real networks to evaluate the behavior of the error bounds as the sample size increases. Simulations have allowed us to estimate the minimum sample size required to achieve a given error probability for different error thresholds and hidden population rates. Particularly, the results of the numerical experiments carried out support and corroborate the theoretical analysis on the error bounds.

To our knowledge, this paper is the first to present theoretical analytical bounds on the performance of NSUM methods, opening many avenues for future research. For instance, it will be of interest to study which additional information can be collected from the samples in order to overcome the worst-case results. Regarding random networks, topologies beyond Erdős–Rényi and Scale-Free can be explored. For instance, the Stochastic Block Model or Hyperbolic topologies seem of special interest. It is also worth analyzing the worst error that may arise when the topology is random, but the dissemination of the hidden sub-population is controlled by an adversary after the network has been defined. Finally, efficient methods to obtain

Table 3.2: Summary of Analytical Error Bounds for NSUM Hidden Population Rate Estimators. Sample $S \subseteq V$ is random uniform with $m = |S|$, $R_S = \sum_{v \in S} R_v$, and any $\beta = 1 + \epsilon > 1$, any $\alpha > 0$, and any $\delta \in (0, 1)$.

Case	Description	Bound	Theorem
Adversarial Instances	Lower bound, Deterministic $\mathcal{M}$	$\mathcal{E}_{\mathcal{M}} \geq \sqrt{(n-1)/2}$	3.1
Random Networks	Upper Bound MoR	$\Pr\left[\mathcal{E}_{\text{MoR}} > \beta\right] \leq \left(\frac{e^{\beta-1}}{\beta^{\beta}}\right)^{m\rho} + \left(\frac{\frac{1}{e^{\beta}}-1}{\beta^{-1/\beta}}\right)^{m\rho}$	3.2
	Upper Bound RoS	$\Pr\left[\mathcal{E}_{\text{RoS}} > \beta\right] \leq \sum_R \left(\left(\frac{e^{\beta-1}}{\beta^{\beta}}\right)^{R\rho} + \left(\frac{\frac{1}{e^{\beta}}-1}{\beta^{-1/\beta}}\right)^{R\rho}\right) \Pr\left[R_S = R\right]$	3.3
	Sample size $\mathcal{M} \in \{\text{MoR}, \text{RoS}\}$	$m = \lvert S\rvert \geq \frac{\ln 2 + \alpha \ln n}{\rho(1 - \frac{1}{\beta}(\ln \beta + 1))} \implies \Pr\left[\mathcal{E}_{\mathcal{M}} > \beta\right] \leq 1/n^{\alpha}$	3.4
Erdős–Rényi Networks	Upper Bound MoR	See above	3.5
	Upper Bound RoS $\mu = mp(n-1)$	$\Pr\left[\mathcal{E}_{\text{RoS}} > \beta\right] \leq \left(\frac{e^{-\delta}}{(1-\delta)^{(1-\delta)}}\right)^{\mu} + \left(\frac{e^{\beta-1}}{\beta^{\beta}}\right)^{(1-\delta)\mu\rho} + \left(\frac{\frac{1}{e^{\beta}}-1}{\beta^{-1/\beta}}\right)^{(1-\delta)\mu\rho}$	3.6
Scale-Free Networks	Upper Bound MoR	See above	3.2
	Upper Bound RoS, $\gamma > 2$ $\mu \approx m\frac{1-\gamma}{2-\gamma}\frac{1-(n-1)^{2-\gamma}}{1-(n-1)^{1-\gamma}}$	$\Pr\left[\mathcal{E}_{\text{RoS}} > \beta\right] \leq \left(\frac{e^{-\delta}}{(1-\delta)^{(1-\delta)}}\right)^{\mu} + \left(\frac{e^{\beta-1}}{\beta^{\beta}}\right)^{(1-\delta)\mu\rho} + \left(\frac{\frac{1}{e^{\beta}}-1}{\beta^{-1/\beta}}\right)^{(1-\delta)\mu\rho}$	3.7

information about the social network from ARD beyond the size of the hidden sub-population, e.g., the degree distribution or the topology, would be of great interest.

3.8. Acknowledgments

This paper has been funded by project PID2022-140560OB-I00 (DRONAC) funded by MICIU/AEI /10.13039/501100011033 and ERDF, EU. This research is part of the I+D+i projects PID2022-137243OB-I00 funded by MCIN/AEI/10.13039/501100011033 and European Union NextGenerationEU/PRTR and the project CuidaNSUM of the Instituto de las Mujeres. This initiative has also been partially carried out within the framework of the Recovery, Transformation and Resilience Plan funds, financed by the European Union (Next Generation) through the grant ANTICIPA (INCIBE) and the ENIA 2022 Chairs for the creation of university-industry chairs in AI-AImpulsa: UC3M-Universia. The work of Sergio Díaz-Aranda has been funded by *Comunidad de Madrid* predoctoral grant PIPF-2022/COM-24467.

References

M. Ahmadi-Gohari, F. Zolala, A. Iranpour, and M. R. Baneshi. Twelve-hour before driving prevalence of alcohol and drug use among heavy vehicle drivers in south east of iran using network scale up. Addiction & health, 11(4):256, 2019.

H. Bernard, E. Johnsen, P. Killworth, and S. Robinson. How many people died in the Mexico City earthquake. Technical report, University of Florida, 1988.

H. R. Bernard, E. C. Johnsen, P. D. Killworth, and S. Robinson. Estimating the size of an average personal network and of an event subpopulation: Some empirical results. Social Science Research, 20(2):109–121, 1991.

H. R. Bernard, P. D. Killworth, E. C. Johnsen, G. A. Shelley, and C. McCarty. Estimating the ripple effect of a disaster. Connections, 24(2):18–22, 2001.

L. Chen, A. Karbasi, and F. W. Crawford. Estimating the size of a large network and its communities from a random sample. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, Advances in Neural Information Processing Systems, volume 29, pages 3072–3080, Barcelona, Spain, 2016. Curran Associates, Inc.

A. Clauset, C. R. Shalizi, and M. E. Newman. Power-law distributions in empirical data. SIAM review, 51(4):661–703, 2009.

S. Díaz-Aranda, J. M. Ramírez, M. Daga, J. P. Champati, J. Aguilar, R. E. Lillo, and A. F. Anta. Error bounds for the network scale-up method, 2024. URL https://arxiv.org/abs/2407.10640.

P. Erdős and A. Rényi. On random graphs i. Publicationes Mathematicae, 6:290–297, 1959.

S. Ezoe, T. Morooka, T. Noda, M. L. Sabin, and S. Koike. Population size estimation of men who have sex with men through the network scale-up method in japan. PloS one, 7(1):e31184, 2012.

K. Garbe and J. Vondrak. Concentration of lipschitz functions of negatively dependent variables. arXiv preprint arXiv:1804.10084, 2018.

A. Garcia-Agundez, O. Ojo, H. A. Hernández-Roig, C. Baquero, D. Frey, C. Georgiou, M. Goessens, R. E. Lillo, R. Menezes, N. Nicolaou, et al. Estimating the covid-19 prevalence in spain with indirect reporting via open surveys. Frontiers in Public Health, 9:658544, 2021.

W. Hoeffding. Probability Inequalities for sums of Bounded Random Variables, pages 409–426. Springer New York, New York, NY, 1994. ISBN 978-1-4612-0865-5. doi: 10.1007/978-1-4612-0865-5_26.

P. D. Killworth, E. C. Johnsen, C. McCarty, G. A. Shelley, and H. R. Bernard. A social network approach to estimating seroprevalence in the united states. Social networks, 20(1):23–50, 1998a.

P. D. Killworth, C. McCarty, H. R. Bernard, G. A. Shelley, and E. C. Johnsen. Estimation of seroprevalence, rape, and homelessness in the united states using a social network approach. Evaluation review, 22(2):289–308, 1998b.

I. Laga, L. Bao, and X. Niu. Thirty years of the network scale-up method. Journal of the American Statistical Association, 116(535):1548–1559, 2021.

R. Maltiel, A. E. Raftery, T. H. McCormick, and A. J. Baraff. Estimating population size using the network scale up method. The Annals of Applied Statistics, 9(3): 1247, 2015.

T. H. McCormick and T. Zheng. Latent demographic profile estimation in hard-to-reach groups. The annals of applied statistics, 6(4):1795, 2012.

T. H. McCormick, M. J. Salganik, and T. Zheng. How many people do you know?: Efficiently estimating personal network size. Journal of the American Statistical Association, 105(489):59–70, 2010.

A. Panconesi and A. Srinivasan. Randomized distributed edge coloring via an extension of the chernoff–hoeffding bounds. SIAM Journal on Computing, 26(2): 350–368, 1997.

B. Rozemberczki, R. Davies, R. Sarkar, and C. Sutton. Gemsec: graph embedding with self clustering. In Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM '19, page 65–72, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450368681. doi: 10.1145/3341161.3342890. URL https://doi.org/10.1145/3341161.3342890.

S. Snidero, N. Soriani, I. Baldi, F. Zobec, P. Berchialla, and D. Gregori. Scale-up approach in cati surveys for estimating the number of foreign body injuries in the aero-digestive tract in children. International journal of environmental research and public health, 9(11):4056–4067, 2012.

A. Srivastava, J. M. Ramirez, S. Díaz-Aranda, J. Aguilar, A. F. Anta, A. Ortega, and R. E. Lillo. Nowcasting temporal trends using indirect surveys. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 38, pages 22359–22367, 2024.

W. UNAIDS. Guidelines on estimating the size of populations most at risk to HIV. Geneva: World Health Organization, 2010.

J. Wang, Y. Yang, W. Zhao, H. Su, Y. Zhao, Y. Chen, T. Zhang, and T. Zhang. Application of network scale up method in the estimation of population size for men who have sex with men in shanghai, china. PloS one, 10(11):e0143118, 2015.

T. Zheng, M. J. Salganik, and A. Gelman. How many people do you know in prison? using overdispersion in count data to estimate social structure in networks. Journal of the American Statistical Association, 101(474):409–423, 2006.

Appendices

3.S.A. Concentration Bounds for Negatively Correlated Random Variables

Let us define the negative correlation of random variables (rvs).

Definition 3.1. *The rvs $Z_1, Z_2, \ldots, Z_n$ are negatively correlated if for any subset $B \subseteq \{1, 2, \ldots, n\}$, it holds that*

$$\mathbb{E}\left[\prod_{i \in B} Z_i\right] \leq \prod_{i \in B} \mathbb{E}\left[Z_i\right].$$

Definition 3.2. *The rvs $Z_1, Z_2, \ldots, Z_n$ are said to exhibit negative cylinder dependence (Garbe and Vondrak, 2018) if both sets of rvs $Z_1, Z_2, \ldots, Z_n$ and $1 - Z_1, 1 - Z_2, \ldots, 1 - Z_n$ are negatively correlated.*

We use the multiplicative error Chernoff bound for bounded rvs $0 \leq Z_i \leq 1$ with negative cylinder dependence presented in the following theorem. The proof of the theorem is a modification of the result for 0-1 rvs from Panconesi and Srinivasan (1997) and is presented below for completeness.

Theorem 3.8. *Let $Z_1, Z_2, \ldots, Z_n$ be rvs with negative cylinder dependence, where $0 \leq Z_i \leq 1$. Let $Z = \sum_{i=1}^{n} Z_i$ and let $\mu = \mathbb{E}[Z]$, then for $\beta > 1$, it holds that*

$$\Pr\left[\neg\left(\frac{\mu}{\beta} \leq Z \leq \beta\mu\right)\right] \leq \left(\frac{e^{\beta-1}}{\beta^{\beta}}\right)^{\mu} + \left(\frac{e^{\frac{1}{\beta}-1}}{\beta^{-1/\beta}}\right)^{\mu} = F(\beta, \mu),$$

where $\Pr[\neg A]$ is the probability of A complement.

Proof. Using Markov's inequality, for $t \geq 0$, we have

$$\Pr[Z \geq \beta\mu] = \Pr\left[e^{tZ} \geq e^{t\beta\mu}\right] \leq \frac{\mathbb{E}\left[e^{tZ}\right]}{e^{t\beta\mu}}, \tag{3.1}$$

$$\text{where } \mathbb{E}\left[e^{tZ}\right] = \mathbb{E}\left[e^{t\sum_{i=1}^{n} Z_i}\right] = \mathbb{E}\left[\prod_{i=1}^{n} e^{tZ_i}\right].$$

We prove that $\mathbb{E}\left[\prod_{i=1}^{n} e^{tZ_i}\right] \leq \prod_{i=1}^{n} \mathbb{E}\left[e^{tZ_i}\right]$. Let $\hat{Z}_1, \ldots, \hat{Z}_n$ be independent 0-1 rvs such that $\mathbb{E}\left[\hat{Z}_i\right] = \mathbb{E}\left[Z_i\right]$ for each $i = \{1, 2 \ldots, n\}$, and define $\hat{Z} := \hat{Z}_1 + \cdots + \hat{Z}_n$. For $k > 0$, $\alpha = \{\alpha_1, \alpha_2, \ldots, \alpha_n\}$ is a decomposition such that $\sum_{i=1}^{n} \alpha_i = k$, where $\alpha_i \geq 0$. We have $\mathbb{E}\left[Z^k\right] = \sum_{\boldsymbol{\alpha}} \mathbb{E}\left[\prod_{i=1}^{n} Z_i^{\alpha_i}\right]$. We claim that for any $\boldsymbol{\alpha}$, we have $\mathbb{E}\left[\prod_{i=1}^{n} Z_i^{\alpha_i}\right] \leq \mathbb{E}\left[\prod_{i=1}^{n} \hat{Z}_i^{\alpha_i}\right]$. Without loss of generality, for any $\boldsymbol{\alpha}$, we have $\alpha_i > 0$, for all $i \leq l$, and $\alpha_i = 0$ for $i > l$, for some $l \leq n$. Therefore, we have

$$\mathbb{E}\left[\prod_{i=1}^{n} Z_i^{\alpha_i}\right] = \mathbb{E}\left[\prod_{i=1}^{l} Z_i^{\alpha_i}\right] \leq \mathbb{E}\left[\prod_{i=1}^{l} Z_i\right] \leq \prod_{i=1}^{l} \mathbb{E}\left[Z_i\right] = \prod_{i=1}^{l} \mathbb{E}\left[\hat{Z}_i\right] = \mathbb{E}\left[\prod_{i=1}^{l} \hat{Z}_i^{\alpha_i}\right] = \mathbb{E}\left[\prod_{i=1}^{n} \hat{Z}_i^{\alpha_i}\right].$$

We used $Z_i \in [0,1]$ for the second step above, and for the third step, we used the negative correlation property. For the fifth step, we used the fact that $\hat{Z}_i$ are 0-1 independent rvs. We have

$$\mathbb{E}\left[Z^k\right] \leq \sum_{\boldsymbol{\alpha}} \mathbb{E}\left[\prod_{i=1}^{n} Z_i^{\alpha_i}\right] \leq \sum_{\boldsymbol{\alpha}} \mathbb{E}\left[\prod_{i=1}^{n} \hat{Z}_i^{\alpha_i}\right] = \mathbb{E}\left[\hat{Z}^k\right].$$

Thus, for any $k \geq 1$, we have $\mathbb{E}\left[Z^k\right] \leq \mathbb{E}\left[\hat{Z}^k\right]$. Using this result in the Taylor expansion of $e^{t\hat{Z}}$ in $\mathbb{E}\left[e^{t\hat{Z}}\right]$, we obtain $\mathbb{E}\left[e^{tZ}\right] \leq \mathbb{E}\left[e^{t\hat{Z}}\right]$. We now bound $\mathbb{E}\left[e^{t\hat{Z}_i}\right]$. Since $\hat{Z}_i$ are 0-1 rvs, we have $\Pr(\hat{Z}_i = 1) = \mathbb{E}\left[\hat{Z}_i\right]$ and $\Pr(\hat{Z}_i = 0) = 1 - \mathbb{E}\left[\hat{Z}_i\right]$. We have

$$\mathbb{E}\left[e^{t\hat{Z}_i}\right] = e^t \mathbb{E}\left[\hat{Z}_i\right] + \mathbb{E}\left[\hat{Z}_i\right] = \mathbb{E}\left[Z_i\right](e^t - 1) + 1 \leq e^{\mathbb{E}[Z_i](e^t-1)}.$$

We used $1 + x \leq e^x$ for all $x \in \mathbb{R}$ in the last inequality. Hence,

$$\mathbb{E}\left[e^{tZ}\right] \leq \prod_{i=1}^{n} e^{\mathbb{E}[Z_i](e^t-1)} = e^{\mu(e^t-1)}. \tag{3.2}$$

Substituting (3.2) in (3.1) and choosing $t = \ln \beta$, we obtain

$$\Pr\left[Z \geq \beta\mu\right] \leq \left(\frac{e^{\beta-1}}{\beta^\beta}\right)^\mu. \tag{3.3}$$

Again, from Markov's inequality, we have

$$\Pr\left[Z \leq \frac{\mu}{\beta}\right] \leq e^{\frac{t\mu}{\beta}} \mathbb{E}\left[e^{-tZ}\right].$$

Given that variables $1 - Z_i$ are negatively correlated due to the negative cylinder dependence, we can apply the steps outlined in the previous case to obtain that

$$\mathbb{E}\left[e^{t\sum_{i=1}^{n}(1-Z_i)}\right] \leq \mathbb{E}\left[e^{t\sum_{i=1}^{n}(1-\hat{Z}_i)}\right],$$

and then

$$\begin{aligned}
\mathbb{E}\left[e^{-tZ}\right] &= e^{-tn}\mathbb{E}\left[e^{tn-tZ}\right] = e^{-tn}\mathbb{E}\left[e^{t\sum_{i=1}^{n}(1-Z_i)}\right] \\
&\leq e^{-tn}\mathbb{E}\left[e^{t\sum_{i=1}^{n}(1-\hat{Z}_i)}\right] = e^{-tn}\mathbb{E}\left[e^{tn-t\hat{Z}}\right] \\
&= \mathbb{E}\left[e^{-t\hat{Z}}\right].
\end{aligned}$$

Following similar steps as above, we obtain

$$\Pr\left[Z \leq \frac{\mu}{\beta}\right] \leq \left(\frac{e^{\frac{1}{\beta}-1}}{\beta^{-1/\beta}}\right)^\mu. \tag{3.4}$$

From (3.3) and (3.4) and the fact that $\Pr(\neg(A \cap B)) \leq \Pr(\neg A) + \Pr(\neg B)$. □

The following corollary uses $\beta = 1/(1-\delta)$ in (3.4), for $\delta \in (0,1)$.

Corollary 3.2. *Let $Z_1, Z_2, \ldots, Z_n$ be rvs with negative cylinder dependence, where $0 \leq Z_i \leq 1$. Let $Z = \sum_{i=1}^{n} Z_i$ and let $\mu = \mathbb{E}[Z]$, then for $\delta \in (0,1)$, it holds that*

$$\Pr\left[Z \leq (1-\delta)\mu\right] \leq \left(\frac{e^{-\delta}}{(1-\delta)^{(1-\delta)}}\right)^\mu.$$

3.S.B. Proofs of Lemma 3.1 and Theorem 3.2

Lemma 3.2. *If $v \in H$ then $\mathbb{E}[X_{vj} \mid v \in H] = \frac{h-1}{n-1}$. If $v \notin H$ then $\mathbb{E}[X_{vj} \mid v \notin H] = \frac{h}{n-1}$.*

Proof. Let us assume $v \in H$. By definition of the expectation of an indicator variable, $\mathbb{E}[X_{vj} \mid v \in H]$ is the probability that the jth in-neighbor selected by v belongs to H. This is

$$\mathbb{E}[X_{vj} \mid v \in H] = \frac{(h-1)(n-2)!}{(n-1)!} = \frac{h-1}{n-1}, \tag{3.5}$$

where the numerator is the possible permutations of the $n-1$ vertices in $V \setminus \{v\}$ in which the jth vertex is in H, and the denominator is the number of possible permutations of the $n-1$ vertices in $V \setminus \{v\}$. The same argument can be applied to the case $v \notin H$ to obtain $\mathbb{E}[X_{vj} \mid v \notin H] = \frac{h}{n-1}$. □

Lemma 3.1. *For all $v \in S$ and $j \in \{1, 2, \ldots, R_v\}$, $\mathbb{E}[X_{v_j}] = \mathbb{E}[Y_v] = \frac{h}{n} = \rho(I)$.*

Proof.

$$\begin{aligned}
\mathbb{E}[X_{vj}] &= \Pr[v \in H] \cdot \mathbb{E}[X_{vj} \mid v \in H] + \Pr[v \notin H] \cdot \mathbb{E}[X_{vj} \mid v \notin H] \\
&= \frac{h}{n}\frac{h-1}{n-1} + \frac{n-h}{n}\frac{h}{n-1} = \frac{h}{n}\left(\frac{h-1}{n-1} + \frac{n-h}{n-1}\right) = \frac{h}{n} = \rho(I)
\end{aligned}$$

$$\begin{aligned}
\mathbb{E}[Y_v] &= \sum_r \sum_c \frac{c}{r} \Pr[C_v = c, R_v = r] \\
&= \sum_r \frac{1}{r} \Pr[R_v = r] \cdot \sum_c c \Pr[C_v = c \mid R_v = r] \\
&= \sum_r \frac{1}{r} \Pr[R_v = r] \cdot \mathbb{E}[C_v \mid R_v = r] \\
&= \sum_r \frac{1}{r} \Pr[R_v = r] \cdot \sum_{j=1}^{r} \mathbb{E}[X_{vj}] \\
&= \sum_r \frac{1}{r} \Pr[R_v = r] r\rho(I) = \sum_r \Pr[R_v = r]\rho(I) = \rho(I).
\end{aligned}$$

□

Lemma 3.3. *Let $S \subseteq V$ be a set with size m chosen uniformly at random, then the rvs Y_v, for all $v \in S$, are negatively correlated, i.e., $\forall S_0 \subseteq S, \mathbb{E}[\prod_{v \in S_0} Y_v] \leq \prod_{v \in S_0} \mathbb{E}[Y_v]$.*

Proof. Let $S_0 \subseteq S$ be a set of nodes. The variables Y_v, for $v \in S_0$, only depend on whether v belongs to H, and they are mutually independent. Then, given a particular configuration $A = (S_0 \cap H, S_0 \cap \bar{H})$ of hidden-population node memberships, the expectation of the product of the variables Y_v, conditional on this configuration, is equal to

$$\mathbb{E}\left[\prod_{v \in S_0} Y_v \mid A\right] = \prod_{v \in S_0 \cap H} E[Y_v|A] \prod_{v \in S_0 \cap \bar{H}} E[Y_v|A]$$

$$= \prod_{v \in S_0 \cap H} E[Y_v | v \in H] \prod_{v \in S_0 \cap \bar{H}} E[Y_v | v \notin H]$$
$$= \left(\frac{h-1}{n-1}\right)^{|S_0 \cap H|} \left(\frac{h}{n-1}\right)^{|S_0 \cap \bar{H}|}.$$

Let $Z = |S_0 \cap H|$ and $m_0 = |S_0|$. Since we do not know if the nodes are in H, Z is a hypergeometric random variable with probability mass function $\Pr[Z = z] = \frac{\binom{h}{z}\binom{n-h}{m_0-z}}{\binom{n}{m_0}}$ and $\mathbb{E}[Z] = \frac{m_0 h}{n}$. Let $\{A_i\}$ be the partition of the sample space composed of all the individual cases in which $Z = z$. Note that $\sum_i \Pr[A_i] = \Pr[Z = z]$. We can apply the Law of total expectation to $\prod_{v \in S_0} Y_v \mid Z = z$ over the sample space restricted to $Z = z$. Then,

$$\begin{aligned}
\mathbb{E}\left[\prod_{v \in S_0} Y_v \mid Z = z\right] &= \sum_i \mathbb{E}\left[\prod_{v \in S_0} Y_v \mid A_i, Z = z\right] \Pr[A_i \mid Z = z] \\
&= \sum_i \mathbb{E}\left[\prod_{v \in S_0} Y_v \mid A_i\right] \Pr[A_i \mid Z = z] \\
&= \sum_i \left(\frac{h-1}{n-1}\right)^z \left(\frac{h}{n-1}\right)^{m_0 - z} \Pr[A_i \mid Z = z] \\
&= \sum_i \left(\frac{h-1}{n-1}\right)^z \left(\frac{h}{n-1}\right)^{m_0 - z} \frac{\Pr[A_i]}{\Pr[Z = z]} \\
&= \left(\frac{h-1}{n-1}\right)^z \left(\frac{h}{n-1}\right)^{m_0 - z}.
\end{aligned}$$

The unconditional expectation of the product is

$$\begin{aligned}
\mathbb{E}\left[\prod_{v \in S_0} Y_v\right] &= \sum_{z=0}^{m_0} \mathbb{E}\left[\prod_{v \in S_0} Y_v \mid Z = z\right] \Pr[Z = z] \\
&= \sum_{z=0}^{m_0} \left(\frac{h-1}{n-1}\right)^z \left(\frac{h}{n-1}\right)^{m_0 - z} \Pr[Z = z] \\
&= \frac{h^{m_0}}{(n-1)^{m_0}} \sum_{z=0}^{m_0} (1 - h^{-1})^z \Pr[Z = z] \\
&= \frac{h^{m_0}}{(n-1)^{m_0}} \mathbb{E}[e^{\ln(1-h^{-1})Z}]. \qquad (3.6)
\end{aligned}$$

$\mathbb{E}[e^{tZ}]$ is the moment generating function (MGF) of Z. We relate the MGF of Z with the MGF of a binomial random variable X with parameters $(m_0, \frac{h}{n})$. Note that $\mathbb{E}[X] = \mathbb{E}[Z]$. We state the claim.

Claim 3.1. *MGF of X is at least the MGF of Z, i.e., $\mathbb{E}[e^{tX}] \geq \mathbb{E}[e^{tZ}]$, for all $t \in \mathbb{R}$.*

Proof. By Hoeffding (1994), $\mathbb{E}[f(\sum Z_i)] \geq \mathbb{E}[f(\sum X_i)]$, where Z_i are random samples without replacement from a finite population C, X_i are random samples with replacement from C, and f is a convex function. The claim follows from the fact that

$X = \sum_{i=1}^{m} X_i$, with $X_i \sim Bernoulli(\frac{h}{n})$, $Z = \sum_{i=1}^{m} Z_i$, where Z_i are variables without replacement indicating the membership in H of the ith draw, and $x \mapsto exp(tx)$ is a convex function. □

The MGF of the binomial random variable X is given by

$$\mathbb{E}[e^{tX}] = \left(1 - \frac{h}{n} + \frac{he^t}{n}\right)^m.$$

From the claim, we have

$$\begin{aligned}\mathbb{E}[e^{(1-h^{-1})Z}] \leq \mathbb{E}[e^{\ln(1-h^{-1})X}] &= \left(1 - \frac{h}{n} + \frac{he^{\ln(1-h^{-1})}}{n}\right)_0^m \\ &= \left(1 - \frac{h}{n} + \frac{h}{n}(1-h^{-1})\right)_0^m = \frac{(n-1)^{m_0}}{n^{m_0}}.\end{aligned}$$

Using this in (3.6), we obtain

$$\mathbb{E}\left[\prod_{v\in S_0} Y_v\right] \leq \frac{h^{m_0}}{(n-1)^{m_0}} \frac{(n-1)^{m_0}}{n^{m_0}} = \left(\frac{h}{n}\right)^{m_0} = \prod_{v\in S_0} \mathbb{E}[Y_v].$$

□

Lemma 3.4. *Let $S \subseteq V$ be a set with size m chosen uniformly at random, then the rvs Y_v, for all $v \in S$ exhibit negative cylinder dependence, i.e., $\forall S_0 \subseteq S$*

(i) $\mathbb{E}[\prod_{v\in S_0} Y_v] \leq \prod_{v\in S_0} \mathbb{E}[Y_v]$

(ii) $\mathbb{E}[\prod_{v\in S_0} (1 - Y_v)] \leq \prod_{v\in S_0} \mathbb{E}[1 - Y_v]$

Proof. The first part is proven in Lemma 3.3. For the negative correlation of the $1 - Y_v$, we observe that the variables $1 - Y_v$ correspond to the variables Y'_v from a random network obtained by changing the hidden population to its complement. That is, an instance (G', H'), with degree distribution $P'_{\deg}$, and variables R'_v and C'_v, where $G' = G$, $H' = \bar{H}$, $P'_{\deg} = P_{\deg}$, $R'_v = R_v$ and $C'_v = R_v - C_v$. Thus, applying Lemma 3.3 to the variables of this random network, we conclude the proof. □

Theorem 3.2. *Consider an instance $I = (G, H)$ where G is a random network. If $S \subseteq V$ is sampled uniformly at random, with $m = |S|$, then the* MoR *estimator has the following error bound for any $\beta = 1 + \epsilon > 1$.*

$$\Pr\left[\mathcal{E}_{\mathrm{MoR}}(I, S) > \beta\right] \leq F(\beta, m\rho(I)).$$

Proof. Given $\mathbb{E}[Y_v] = \rho(I)$ and Lemma 3.4, we apply Theorem 3.8 to the rv Y_v, and $Y_S = \sum_{v\in S} Y_v$ with $\mu = \mathbb{E}[Y_S] = m\rho(I)$.

$$\Pr\left[\mathcal{E}_{\mathrm{MoR}}(I) > \beta\right] = \Pr\left[\max\left(\frac{\rho(I)}{\hat{\rho}_{\mathrm{MoR}}(I)}, \frac{\hat{\rho}_{\mathrm{MoR}}(I)}{\rho(I)}\right) > \beta\right]$$

$$
\begin{aligned}
&= \Pr\left[\neg\left(\frac{\rho(I)}{\beta} \le \hat{\rho}_{\mathrm{MoR}}(I) \le \rho(I)\beta\right)\right] \\
&= \Pr\left[\neg\left(\frac{m\rho(I)}{\beta} \le Y_S \le m\rho(I)\beta\right)\right] \\
&\le \left(\frac{e^{\beta-1}}{\beta^{\beta}}\right)^{m\rho(I)} + \left(\frac{e^{\frac{1}{\beta}-1}}{\beta^{-1/\beta}}\right)^{m\rho(I)} \\
&= F(\beta, m\rho(I)).
\end{aligned}
$$

□

3.S.C. Proof of Theorem 3.3

Lemma 3.5. *Let $S \subseteq V$ be a set with size m chosen uniformly at random, then for all $v \in S$ and $j \in \{1, 2, \ldots, R_v\}$, the rvs X_{vj} are negatively correlated, i.e., for all $S_0 \subseteq S$ and $B_v \subseteq \{1, \cdots, R_v\}$ with $v \in S_0$, $\mathbb{E}\left[\prod_{v\in S_0}\prod_{j=\in B_v} X_{vj}\right] \le \prod_{v\in S_0}\prod_{B_v}\mathbb{E}[X_{vj}]$.*

Proof. Let $S_0 \subseteq S$ be a set of nodes, and m_0 the size of S_0. For notational simplicity, we consider only the case in which each index set takes the form $B_v = \{1, \cdots, R_v\}$. Observe that the variables for different nodes X_{vj} may be dependent, but are independent if the membership in H is known. The conditional expectation of the product $\prod_{v\in S_0}\prod_{j=1}^{R_v} X_{vj}$ is equal to

$$
\left(\prod_{v\in S_0\cap H} \mathbb{E}\left[\prod_{j=1}^{R_v} X_{vj} \mid v \in H\right]\right)\left(\prod_{v\in S_0\cap \bar{H}} \mathbb{E}\left[\prod_{j=1}^{R_v} X_{vj} \mid v \notin H\right]\right). \tag{3.7}
$$

We also have that

$$
\mathbb{E}\left[\prod_{j=1}^{R_v} X_{vj} \mid v \in H\right] = \Pr\left[\prod_{j=1}^{R_v} X_{vj} = 1 \mid v \in H\right] = \prod_{j=1}^{R_v}\frac{h-j}{n-j} \le \rho(I)^{R_v},
$$

and

$$
\mathbb{E}\left[\prod_{j=1}^{R_v} X_{vj} \mid v \notin H\right] = \Pr\left[\prod_{j=1}^{R_v} X_{vj} = 1 \mid v \notin H\right] = \prod_{j=1}^{R_v}\frac{h-j-1}{n-j} \le \rho(I)^{R_v}.
$$

Let $\{B_i\}$ be a partition of the sample space composed of all the possible combinations of nodes in H. Then,

$$
\begin{aligned}
\mathbb{E}\left[\prod_{v\in S_0}\prod_{j=1}^{R_v} X_{vj}\right] &= \sum_i \mathbb{E}\left[\prod_{v\in S_0}\prod_{j=1}^{R_v} X_{vj} \mid B_i\right]\Pr[B_i] \\
&= \sum_i \prod_{v\in S_0}\mathbb{E}\left[\prod_{j=1}^{R_v} X_{vj} \mid B_i\right]\Pr[B_i] \\
&\le \sum_i \prod_{v\in S_0}\rho(I)^{R_v}\Pr[B_i]
\end{aligned}
$$

$$
\begin{aligned}
&= \prod_{v \in S_0} \rho(I)^{R_v} \\
&= \rho(I)^{\sum_{v \in S_0} R_v} \\
&= \prod_{v \in S_0} \prod_{j=1}^{R_v} \mathbb{E}[X_{vj}].
\end{aligned}
$$

□

Lemma 3.6. *Let $S \subseteq V$ be a set with size m chosen uniformly at random, then for all $v \in S$ and $j \in \{1, 2, \ldots, R_v\}$, the random variables X_{vj} have negative cylinder dependence, i.e., for all $S_0 \subseteq S$, and $B_v \subseteq \{1, \cdots, R_v\}$ with $v \in S_0$*

(i) $\mathbb{E}\left[\prod_{v \in S_0} \prod_{j \in B_v} X_{vj}\right] \leq \prod_{v \in S_0} \prod_{j \in B_v} \mathbb{E}[X_{vj}]$

(ii) $\mathbb{E}\left[\prod_{v \in S_0} \prod_{j \in B_v} (1 - X_{vj})\right] \leq \prod_{v \in S_0} \prod_{j \in B_v} \mathbb{E}[1 - X_{vj}]$

Proof. (i) is a consequence of Lemma 3.5. To prove (ii), we see that the variables $X'_{vj} = 1 - X_{vj}$ are the variables of a random network constructed from an instance $(G, \bar{H})$, degree distribution $P_{\deg}$, and variables $R'_v = Rv$ and $C'_v = R_v - C_v$. □

Theorem 3.3. *Consider an instance $I = (G, H)$ where G is a random network. If $S \subseteq V$ is sampled uniformly at random, with $m = |S|$, then for $R_S = \sum_{v \in S} R_v$ and any $\beta = 1 + \epsilon > 1$, the* RoS *estimator has the following error bound.*

$$
\Pr\left[\mathcal{E}_{\mathrm{RoS}}(I, S) > \beta\right] \leq \sum_{R} F(\beta, R\rho(I)) \Pr\left[R_S = R\right].
$$

Proof. By definition, $\hat{\rho}_{\mathrm{RoS}}(I) = C_S / R_S$ and $C_S = \sum_{v \in S} C_v = \sum_{v \in S} \sum_{j=1}^{R_v} X_{vj}$, and hence $\mathbb{E}[C_S] = R_S \rho(I)$. Given Lemma 3.6, we apply Theorem 3.8 to the random variables C_S.

$$
\begin{aligned}
\Pr\left[\mathcal{E}_{\mathrm{RoS}}(I, S) > \beta\right] &= \sum_{R} \Pr\left[\mathcal{E}_{\mathrm{RoS}}(I) > \beta \mid R_S = R\right] \Pr\left[R_S = R\right] \\
&= \sum_{R} \Pr\left[\neg\left(\frac{\rho(I)}{\beta} \leq \hat{\rho}_{\mathrm{RoS}}(I) \leq \rho(I)\beta\right) \mid R_S = R\right] \Pr\left[R_S = R\right] \\
&= \sum_{R} \Pr\left[\neg\left(\frac{R\rho(I)}{\beta} \leq C_S \leq R\rho(I)\beta\right) \mid R_S = R\right] \Pr\left[R_S = R\right] \\
&\leq \sum_{R} \left(\left(\frac{e^{\beta-1}}{\beta^{\beta}}\right)^{R\rho(I)} + \left(\frac{e^{\frac{1}{\beta}-1}}{\beta^{-1/\beta}}\right)^{R\rho(I)}\right) \Pr\left[R_S = R\right] \\
&= \sum_{R} F(\beta, R\rho(I)) \Pr\left[R_S = R\right].
\end{aligned}
$$

□

3.S.D. Deferred Proofs from Section 3.4

3.S.D.1. Proof of Theorem 3.6

Lemma 3.7. *Consider an instance $I = (G, H)$, where G is a random network, and a set $S \subseteq V$ sampled uniformly at random, with $m = |S|$. Let us denote by R_{vu} the Bernoulli random variable for whether $(u, v) \in E$, for all $v \in S$ and all $u \in V \setminus \{v\}$. The variables $\{R_{vu} : v \in S \wedge u \in V \setminus \{v\}\}$ are negatively correlated.*

Proof. From Definition 3.1, the variables R_{vu} are negatively correlated if for any subset $B \subseteq \{(v, u) : v \in S \wedge u \in V \setminus \{v\}\}$, it holds that

$$\mathbb{E}\left[\prod_{(v,u)\in B} R_{vu}\right] \leq \prod_{(v,u)\in B} \mathbb{E}\left[R_{vu}\right].$$

Observe that variables $R_{v_1u_1}$ and $R_{v_2u_2}$ are independent if $v_1 \neq v_2$. Let us define $B_v = \{u \in V \setminus \{v\} : (v, u) \in B\}$. Hence,

$$\prod_{(v,u)\in B} R_{vu} = \prod_{v\in S} \prod_{u\in B_v} R_{vu}$$

Hence it is enough to show that for each $v \in S$,

$$\mathbb{E}\left[\prod_{u\in B_v} R_{vu}\right] \leq \prod_{u\in B_v} \mathbb{E}\left[R_{vu}\right].$$

We prove this inequality by induction on the size of B_v. When $|B_v| = 2$, let $u_1, u_2 \in B_v$, we have that

$$\mathbb{E}[R_{vu_1}R_{vu_2}] = \Pr[R_{vu_1}R_{vu_2} = 1] = \Pr[R_{vu_1} = 1 \mid R_{vu_2} = 1]\Pr[R_{vu_2} = 1] = \Pr[R_{vu_1} = 1 \mid R_{vu_2} = 1]\mathbb{E}[R$$

since $\mathbb{E}[R_{vu_2}] = \Pr[R_{vu_2} = 1]$. Then, it suffices to prove that

$$\Pr[R_{vu_1} = 1 \mid R_{vu_2} = 1] \leq \Pr[R_{vu_1} = 1] = \mathbb{E}[R_{vu_1}]$$

We show that this holds for all possible values of R_v. I.e., we show that

$$\Pr[R_{vu_1} = 1 \mid R_{vu_2} = 1, R_v = k] \leq \Pr[R_{vu_1} = 1 \mid \mathbb{E}[R_{vu_1}] = k]$$

for all positive integer k. This inequality follows from the fact that

$$Pr[R_{vu_1} = 1 \mid R_{vu_2} = 1, R_v = k] = \frac{\binom{n-1-2}{k-2}}{\binom{n-1}{k}} = \frac{k}{n-1}\frac{k-1}{n-2},$$

which is less than $\Pr[R_{vu_1} = 1 \mid R_v = k] = k/n - 1$. Assuming that is true for subsets of sizes up to m_0, for nodes $u_1, \cdots, u_{m_0+1} \in B_v$, we have

$$\begin{aligned}
\mathbb{E}\left[\prod_{i=1}^{m_0+1} R_{vu_i}\right] &= \Pr[R_{vu_1}=1,\cdots,R_{vu_{m_0+1}}=1] \\
&= \Pr[R_{vu_1}=1 \mid R_{vu_2}=1,\cdots,R_{vu_{m0+1}}=1]\Pr[R_{vu_2}=1,\cdots,R_{vu_{m_0+1}}=1] \\
&= \Pr[R_{vu_1}=1 \mid R_{vu_2}=1,\cdots,R_{vu_{m_0+1}}=1]\mathbb{E}\left[\prod_{i=2}^{m_0+1} R_{vu_i}\right] \\
&\leq \Pr[R_{vu_1}=1 \mid R_{vu_2}=1,\cdots,R_{vu_{m_0+1}}=1]\prod_{i=2}^{m_0+1}\mathbb{E}\left[R_{vu_i}\right].
\end{aligned}$$

Where the last inequality is from the induction hypothesis. The first factor of the last equation is less than or equal to $\Pr[R_{vu_1}=1]=\mathbb{E}[R_{vu_1}]$, since for any positive k

$$\begin{aligned}
\Pr[R_{vu_1}=1 \mid R_{vu_2}=1,\cdots,R_{vu_{m_0+1}}=1,R_v=k] &= \frac{\binom{n-1-(m_0+1)}{k-(m_0+1)}}{\binom{n-1}{k}} \\
&= \frac{k}{n-1}\prod_{i=1}^{m_0}\frac{k-i}{n-(i+1)} \\
&= \Pr[R_{vu_1}=1 \mid R_v=k]\prod_{i=1}^{m_0}\frac{k-i}{n-(i+1)} \\
&\leq \Pr[R_{vu_1}=1 \mid R_v=k].
\end{aligned}$$

This completes the proof. □

Lemma 3.8. *Consider an instance $I=(G,H)$, where G is a random network, and a set $S\subseteq V$ sampled uniformly at random, with $m=|S|$. Let us denote by R_{vu} the Bernoulli random variable for whether $(u,v)\in E$, for all $v\in S$ and all $u\in V\setminus\{v\}$. The variables $\{R_{vu}: v\in S\wedge u\in V\setminus\{v\}\}$ have negative cylinder dependence*

Proof. It is sufficient to show that the variables $1-R_{vu}$ are negatively correlated. This follows from the fact that the variables $1-Rvu$ are the variables constructed analogously to R_{uv}, but in the complement graph. That is, $1-R_{vu}=1$ if $(u,v)\notin E$, and 0 otherwise. □

Theorem 3.6. *Consider an instance $I=(G,H)$ where the directed random network $G=(V,E)$ is generated with the Erdős–Rényi model $G(n,p)$ without zero in-degree nodes, and a set $S\subseteq V$ is sampled uniformly at random, with $m=|S|$. Then, the* RoS *estimator has the following error bound, for any $\delta\in(0,1)$ and any $\beta=1+\epsilon>1$ and where $\mu=mp(n-1)$.*

$$\Pr\left[\mathcal{E}_{\mathrm{RoS}}(I,S)>\beta\right] \leq \left(\frac{e^{-\delta}}{(1-\delta)^{(1-\delta)}}\right)^{\mu} + F(\beta,(1-\delta)\mu\rho(I)).$$

Proof. Let us denote by R_{vu} the Bernoulli random variable for whether $(u,v)\in E$, for all $v\in S$ and all $u\in V\setminus\{v\}$. From Lemma 3.8, these variables have negative cylinder

dependence. By definition, $R_S = \sum_{v \in S} \sum_{u \in V \setminus \{v\}} R_{vu}$, and hence $\mathbb{E}[R_S] = mp(n-1)$. Applying Corollary 3.2 to the variables R_{vu} we have that,

$$\Pr\left[R_S \leq (1-\delta)mp(n-1)\right] \leq \left(\frac{e^{-\delta}}{(1-\delta)^{(1-\delta)}}\right)^{mp(n-1)}.$$

Let us now denote $\mu = mp(n-1)$ and $\zeta = \frac{e^{-\delta}}{(1-\delta)^{(1-\delta)}}$. We have,

$$\begin{aligned}
\Pr\left[\mathcal{E}_{\text{RoS}}(I,S) > \beta\right] =& \Pr\left[R_S \leq (1-\delta)\mu\right] \Pr\left[\mathcal{E}_{\text{RoS}}(I) > \beta \mid R_S \leq (1-\delta)\mu\right] + \\
& \Pr\left[R_S > (1-\delta)\mu\right] \Pr\left[\mathcal{E}_{\text{RoS}}(I) > \beta \mid R_S > (1-\delta)\mu\right] \\
& \leq \zeta^{\mu} + \Pr\left[R_S > (1-\delta)\mu\right] \Pr\left[\mathcal{E}_{\text{RoS}}(I) > \beta \mid R_S > (1-\delta)\mu\right] \\
& \leq \zeta^{\mu} + \sum_{R > (1-\delta)\mu} \left(\left(\frac{e^{\beta-1}}{\beta^{\beta}}\right)^{R\rho(I)} + \left(\frac{e^{\frac{1}{\beta}-1}}{\beta^{-1/\beta}}\right)^{R\rho(I)}\right) \Pr\left[R_S = R\right] \\
& \leq \zeta^{\mu} + \left(\frac{e^{\beta-1}}{\beta^{\beta}}\right)^{(1-\delta)\mu\rho(I)} + \left(\frac{e^{\frac{1}{\beta}-1}}{\beta^{-1/\beta}}\right)^{(1-\delta)\mu\rho(I)} \\
& = \left(\frac{e^{-\delta}}{(1-\delta)^{(1-\delta)}}\right)^{\mu} + F(\beta, (1-\delta)\mu\rho(I)).
\end{aligned}$$

□

3.S.D.2. Proof of Theorem 3.7

Theorem 3.7. *Consider an instance $I = (G, H)$ in which G is a Scale-Free random network with $\gamma > 2$. If $S \subseteq V$ is sampled uniformly at random, with $m = |S|$, then the* RoS *estimator has the following error bound, for any $\delta \in (0,1)$, and any $\beta = 1 + \epsilon > 1$ and where $\mu \approx m\frac{1-\gamma}{2-\gamma}\frac{1-(n-1)^{2-\gamma}}{1-(n-1)^{1-\gamma}}$.*

$$\Pr\left[\mathcal{E}_{\text{RoS}}(I,S) > \beta\right] \leq \left(\frac{e^{-\delta}}{(1-\delta)^{(1-\delta)}}\right)^{\mu} + F(\beta, (1-\delta)\mu\rho(I)).$$

Proof. Let us denote by R_{vu} the Bernoulli random variable for whether $(u,v) \in E$, for all $v \in S$ and all $u \in V \setminus \{v\}$. From Lemma 3.8, these variables exhibit negative cylinder dependence. The expectation $\mathbb{E}[R_{vu}] = \Pr[R_{vu} = 1]$ can be approximated as

$$\begin{aligned}
\Pr[R_{vu} = 1] &= \sum_{k=1}^{n-1} \Pr[R_v = k] \Pr[R_{vu} = 1 \mid R_v = k] \\
&= \sum_{k=1}^{n-1} P_{\deg}(k) \frac{k}{n-1} \\
&= \sum_{k=1}^{n-1} \nu k^{-\gamma} \frac{k}{n-1} \\
&= \frac{1}{n-1} \frac{\sum_{k=1}^{n-1} k^{1-\gamma}}{\sum_{k=1}^{n-1} k^{-\gamma}}
\end{aligned}$$

$$\approx \frac{1}{n-1}\frac{\int_{k=1}^{n-1} k^{1-\gamma}}{\int_{k=1}^{n-1} k^{-\gamma}}$$
$$= \frac{1}{n-1}\frac{1-\gamma}{2-\gamma}\frac{1-(n-1)^{2-\gamma}}{1-(n-1)^{1-\gamma}}$$

By definition, $R_S = \sum_{v\in S}\sum_{u\in V\setminus\{v\}} R_{vu}$, and hence $\mathbb{E}[R_S] \approx m(n-1)\frac{1}{n-1}\frac{1-\gamma}{2-\gamma}\frac{1-(n-1)^{2-\gamma}}{1-(n-1)^{1-\gamma}} = m\frac{1-\gamma}{2-\gamma}\frac{1-(n-1)^{2-\gamma}}{1-(n-1)^{1-\gamma}}$. Applying Corollary 3.2 to the variables R_{vu} we have that,

$$\Pr\left[R_S \leq (1-\delta)\mathbb{E}[R_S]\right] \leq \left(\frac{e^{-\delta}}{(1-\delta)^{(1-\delta)}}\right)^{\mathbb{E}[R_S]}.$$

Let us now denote $\mu = m\frac{1-\gamma}{2-\gamma}\frac{1-(n-1)^{2-\gamma}}{1-(n-1)^{1-\gamma}}$ and $\zeta = \frac{e^{-\delta}}{(1-\delta)^{(1-\delta)}}$. We have,

$$\begin{aligned}
\Pr\left[\mathcal{E}_{\mathrm{RoS}}(I,S) > \beta\right] =& \Pr\left[R_S \leq (1-\delta)\mu\right]\Pr\left[\mathcal{E}_{\mathrm{RoS}}(I) > \beta \mid R_S \leq (1-\delta)\mu\right] + \\
& \Pr\left[R_S > (1-\delta)\mu\right]\Pr\left[\mathcal{E}_{\mathrm{RoS}}(I) > \beta \mid R_S > (1-\delta)\mu\right] \\
& \leq \zeta^{\mu} + \Pr\left[R_S > (1-\delta)\mu\right]\Pr\left[\mathcal{E}_{\mathrm{RoS}}(I) > \beta \mid R_S > (1-\delta)\mu\right] \\
& \leq \zeta^{\mu} + \sum_{R>(1-\delta)\mu}\left(\left(\frac{e^{\beta-1}}{\beta^{\beta}}\right)^{R\rho(I)} + \left(\frac{e^{\frac{1}{\beta}-1}}{\beta^{-1/\beta}}\right)^{R\rho(I)}\right)\Pr\left[R_S = R\right] \\
& \leq \zeta^{\mu} + \left(\frac{e^{\beta-1}}{\beta^{\beta}}\right)^{(1-\delta)\mu\rho(I)} + \left(\frac{e^{\frac{1}{\beta}-1}}{\beta^{-1/\beta}}\right)^{(1-\delta)\mu\rho(I)} \\
& = \left(\frac{e^{-\delta}}{(1-\delta)^{(1-\delta)}}\right)^{\mu} + F(\beta,(1-\delta)\mu\rho(I)).
\end{aligned}$$

□

Chapter 4

Nowcasting Temporal Trends Using Indirect Surveys

In *Proceedings of the AAAI Conference on Artificial Intelligence*, 30, 20, 22359-22367, 2024. https://doi.org/10.1609/aaai.v38i20.30242

Ajitesh Srivastava[1] Juan Marcos Ramírez[2] Sergio Díaz-Aranda[2,3]
José Aguilar[2] Antonio Fernández Anta[2] Antonio Ortega[1] Rosa E. Lillo[3]

[1] University of Southern California, USA
[2] IMDEA Networks Institute, Spain
[3] uc3m-Santander Big Data Institute, Universidad Carlos III de Madrid, Spain

Abstract

Indirect surveys, in which respondents provide information about other people they know, have been proposed for estimating (nowcasting) the size of a hidden population where privacy is important, or the hidden population is hard to reach. Examples include estimating casualties in an earthquake, conditions among female sex workers, and the prevalence of drug use and infectious diseases. The Network Scale-up Method (NSUM) is the classical approach to developing estimates from indirect surveys, but it was designed for one-shot surveys. Further, it requires certain assumptions and asking for or estimating the number of individuals in each respondent's network. In recent years, surveys have been increasingly deployed online and can collect data continuously (e.g., COVID-19 surveys on Facebook during much of the pandemic). Conventional NSUM can be applied to these scenarios by analyzing the data independently at each point in time, but this misses the opportunity of leveraging the temporal dimension. We propose to use the responses from indirect surveys collected over time and develop analytical tools (i) to prove that indirect surveys can provide better estimates for the trends of the hidden population over time, as compared to direct surveys, and (ii) to identify appropriate temporal aggregations to improve the estimates. We demonstrate through extensive simulations that our approach outperforms traditional NSUM and direct surveying methods. We also empirically demonstrate the superiority of our approach on a real indirect survey dataset of COVID-19 cases.

4.1. Introduction

Direct reporting through surveys is the most widely used method for collecting data on a given characteristic among individuals in a population. It is well known that direct surveys sometimes face reliability and efficiency problems, as respondents may refuse to participate or may choose to misreport sensitive private information. An alternative approach to overcome some of these issues is using surveys with *indirect* reporting, where respondents answer questions about people they know instead of providing information about themselves. These surveys collect what is known in the literature as *aggregated relational data (ARD)*. Their main advantages are primarily two: (1) privacy is preserved as the respondents do not have to report their own status, thus improving participation and data collection about sensitive populations (Rossier, 2010); (2) one individual response gives the researchers access to information about many different individuals, thus leading to cost reductions in the data collection (Breza et al., 2020), (Alix-Garcia et al., 2021). Indirect surveys have been employed in a variety of domains, such as estimating the number of casualties in an earthquake (Bernard et al., 1989), conditions among female sex workers (Jing et al., 2018), or the prevalence of drug use (Salganik et al., 2010), HIV (Teo et al., 2019), or COVID-19 (Garcia-Agundez et al., 2021).

While indirect surveys have a long history (Laga et al., 2021), the ubiquity of internet access among the general public has made it possible to develop *online indirect surveys*, which can be deployed rapidly and allow the continuous collection of ARD, instead of consisting of one-shot surveys. The importance of recurrent online surveys, including indirect surveys, has become apparent during the COVID-19 pandemic, where one of the most important challenges has been estimating (nowcasting) the number of cases (especially when testing was not widely available), the number of deaths, the number of people vaccinated, etc. While the best-known online surveys are the COVID-19 Trends and Impact Surveys (CTIS) (Astley et al., 2021; Salomon et al., 2021), which collected more than 100,000 responses daily, other surveys were also deployed (Geldsetzer, 2020; Oliver et al., 2020; Garcia-Agundez et al., 2021).

Our main goals are (i) to *quantify the advantages of indirect surveys over direct* ones for nowcasting, by determining under what conditions, more accurate estimates can be obtained via indirect questions; and (ii) to *develop a new method to identify a hidden temporal trend* from the continuously collected ARD from indirect surveys. To achieve these goals, we combine a detailed theoretical analysis with extensive experimental evaluations with both synthetic and real datasets. Given the availability, throughout the paper, we use the COVID-19 surveys dataset as a test case.

4.1.1. Related Work

ARD from indirect surveys are used to nowcast the size of a subpopulation or *hidden population*, i.e., the fraction of the population that has some characteristic. For example, in a COVID-19-related survey, a respondent may report on how many people they know who have tested positive recently, and this information will be used to estimate the fraction of the overall population that is infected at that time. Thus, each respondent is expected to provide information about their own *personal network* (PN) – how many people they know– *and* the number of people they know who are part of the hidden population (e.g., tested positive).

NSUM

The methods proposed in the literature for the estimation of the size of hidden populations (people infected in our example) using ARD are generally known as Network Scale-Up Methods (NSUM) (Laga et al., 2021). To estimate the size of the hidden population from the ARD, NSUM assumes that the proportion of those belonging to the hidden population within a respondent's PN is a good approximation to the same proportion in the overall population. NSUM can work well under several assumptions: (1) the respondents can accurately recall the people in their PN, (2) the respondents know, for each person in their PN, if they belong to the hidden population, and (3) all individuals have the same probability of belonging to the hidden population. Errors resulting from violations of these conditions are called recall error, transmission error, and barrier effects, respectively (see Laga et al. (2021) for more details). Multiple NSUM extensions have been proposed (Laga et al., 2021), but all of them require requesting or estimating PN sizes (Killworth et al., 1998b; Laga et al., 2021; Garcia-Agundez et al., 2021), or asking individuals in the hidden population about those who know their condition (Feehan and Salganik, 2016).

Our goal is to obtain better estimates of the evolution of the hidden population size *without a need to obtain or estimate the size of individuals' PNs* and, thus, without having to rely on the above recall assumptions. Specifically, one of the key contributions of our work is to leverage the temporal dynamics to improve estimates from continuously collected data. We provide analytical tools to unveil the advantages of appropriately aggregating continuous indirect surveys rather than simply using existing methods over fixed temporal windows. To the best of our knowledge, prior work does not leverage the temporal nature of continuous surveys.

4.1.2. Contributions

We propose a method to estimate the evolution of the size of the hidden population over a period of time using indirect surveys. Our contributions are as follows:

- We propose a latent graph formulation that allows us to prove that the expected response to the indirect survey is proportional to the size of the hidden population (Theorem 4.1). Unlike existing work, we do not assume that every individual has the same probability of reporting someone belonging to the hidden population.
- We prove that within a reasonable upper bound on the latent graph degree variance, the indirect survey provides a better estimate of the hidden population than the direct survey, given the same number of samples (Theorem 4.2).
- We leverage the smoothness of the underlying temporal dynamics to show that a weighted moving average provides better estimates than a series of individual estimates (Lemma 4.5, Theorem 4.3, and Theorem 4.4).
- We verify our claims through a simulated generation of the hidden population with a dynamic process and a simulated survey. We present the impact of various survey parameters (Section 4.3.1).
- We evaluate our approach in the estimation of COVID-19 cases in the US for a period of 18 months (Section 4.3.2).

Our analytical results can be useful for survey design. Note that our objective is nowcasting (rather than forecasting) the time series of the size of the hidden population.

4.2. Methodology

4.2.1. Latent Graph Formulation

Consider a population given by a set N. Suppose that, at time t, N contains a hidden population denoted by $H_t \subseteq N$, leading to a hidden population rate $f_t = |H_t|/|N|$. Let $G = (N, E)$ be a directed graph, where N is the set of nodes and E is the set of edges. In particular, G includes an edge (v, u) if node u possesses knowledge of node v and is willing to report whether node v belongs to the hidden population. Also, we allow self-loops (u, u). These edges may not be the same as those in the contact graph or the social network graph containing the same nodes. The edges present on this (unobserved) graph depend on the specific wording of the survey questions, e.g., "how many in your community ...", "... your household", "your immediate neighbors and coworkers", etc.

Consider a random process that selects **one node**, at time t, to report the number of its neighbors belonging to the hidden population. We denote by X_t the random variable corresponding to this response. In the surveys, at a given time t, multiple nodes are selected randomly (possibly, with replacement) that provide multiple observations for X_t. We will later use the mean and variance of X_t to identify the properties of the *sample mean* $\bar{X}_t$ obtained from these responses. Finally, D is a random variable representing the in-degree of a randomly selected node, with $\mathbb{E}(D) = \mu_D$.

Assumption 4.1. *For any node v belonging to the neighborhood of node u, the event $v \in H_t$ is independent of the in-degree of u.*

This implies that ***having a certain in-degree does not affect whether a randomly selected neighbor is part of the hidden population***. This assumption is more flexible than that used in the traditional NSUM approach, in which every node must have the same probability of finding a neighbor belonging to the hidden population (Laga et al., 2021). Under Assumption 4.1, nodes can have different probabilities of having neighbors in H_t (for example, this could depend on their respective occupations). However, if we consider the union of neighbors of all nodes with a particular indegree, the probability of a random node in this union belonging to H_t remains f_t, irrespective of the indegree considered. This assumption will allow us to eliminate the need to ask for the in-degree of each node. Based on Assumption 4.1, we have the following theorem.

Theorem 4.1. $\mathbb{E}(X_t) = \mu_D \cdot f_t$.

Therefore, the ***mean indirect response is proportional to what we wish to estimate***. Additionally, we make the following observation regarding the underlying graph G.

Observation 4.1. *The mean in-degree of all nodes, μ_D, remains constant over time.*

This observation is supported by the studies of Dunbar (Dunbar, 2010). It can also be observed in the data collected by the Carnegie Mellon University US COVID-19 Trends and Impact Survey (CMU-CTIS) (Salomon et al., 2021), where, over time, different respondents provided the household size in which they reported the number of infections (see Supplementary Material in full version (Srivastava et al., 2023)).

Recall that G is not an acquaintance or physical contact network. Instead, the connection of a node in G represents the network of people the respondent will think of when answering the question. Respondents may not report on the same people each time the survey is completed. So G **may be different at each time t, but the mean of the in-degrees remains constant**. From Observation 4.1 and Theorem 4.1, the time series $\mathbb{E}(X_t)$ is proportional to the time series f_t, representing the fraction of the hidden population, with μ_D as the constant of proportionality. Hence, we can estimate the trend of f_t without knowing μ_D. For applications in which precise f_t values are needed, if the true value of f_t is available for some $t = \tau$, then we can estimate $\mu_D = \mathbb{E}(X_\tau)/f_\tau$ and use this constant to estimate f_t at any t. For example, when f_t represents the rate of active infections, τ could correspond to those dates for which serological studies or wastewater concentration data are available.

Comparison Against Direct Reporting.

With direct reporting, each node reports whether it belongs to the hidden population. Thus, for a randomly selected node v, the response is the binary indicator function I_v, where $I_v = 1$ iff $v \in H_t$. Let Y_t be a random variable denoting the response of a randomly selected node. Observe that Y_t is a binary random variable whose samples follow a Bernoulli distribution with mean $\mathbb{E}(Y_t) = f_t$ and variance $\sigma^2_{Y_t} = f_t(1 - f_t)$. To compare direct and indirect reporting scenarios, we also need to compute the variance of X_t. Since the links in our latent graph do not represent physical contact, neighbors of node u are not necessarily dependent, and we introduce a parameter ϕ_t that controls the level of covariance.

Definition 4.1. *For a pair of nodes v_1 and v_2, with a common neighbor u, $\mathbb{E}(I_{v_1} I_{v_2} | \delta(u)) = \mathbb{E}(I_{v_1} I_{v_2}) = \phi_t f_t$, for some $0 \leq \phi_t \leq 1$.*

Here $\phi_t = f_t$ implies independence, $\phi_t < f_t$ leads to negative covariance and $\phi_t > f_t$ leads to positive covariance. Now we can find bounds on the variance of X_t.

Lemma 4.1. *If σ^2_D is the variance of the degree distribution,*

$$\sigma^2_{X_t} = f_t(\mu^2_D(\phi_t - f_t) + \mu_D(1 - \phi_t) + \sigma^2_D \phi_t)\,. \tag{4.1}$$

Further, $\mu_D f_t(1 - \mu_D f_t) \leq \sigma^2_{X_t} \leq f_t(\sigma^2_D + \mu^2_D(1 - f_t))$.

Suppose we have the same number of responses n for X_t and Y_t, and $n \gg 1$, we show that within a practical upper bound of degree variance σ^2_D, the indirect survey is a better estimator than the direct survey.

Lemma 4.2 (Central Limit Theorem, CLT)**.** *When the number of samples n is large, $\left(\frac{\bar{X}_t - \mathbb{E}(X_t)}{\sigma_{X_t}/\sqrt{n}}\right)$ and $\left(\frac{\bar{Y}_t - \mathbb{E}(Y_t)}{\sigma_{Y_t}/\sqrt{n}}\right)$ follow standard normal distribution, where $\bar{X}_t$ and $\bar{Y}_t$ are the sample means.*

Now, we can show, under Lemma 4.2, that the probability of deviating from the true fraction of hidden population f_t is lower for the estimate obtained from indirect responses $\bar{X}_t/\mu_D$ compared to direct responses $\bar{Y}_t$.

Theorem 4.2. *For any $\lambda > 0$, $P(|\bar{X}_t/\mu_D - f_t| > \lambda) \leq P(|\bar{Y}_t - f_t| > \lambda)$, if the variance of degree distribution $\sigma^2_D \leq \mu_D(\mu_D - 1)(1 - \phi_t)/\phi_t$*

This means ***within realistic bounds on degree variance, indirect surveys are better than direct surveys.*** The condition on σ^2_D is reasonable for applications where f_t at any given time is a small fraction of the population. To see this, first note that if the membership of two neighbors in the hidden population is independent, $\phi_t = f_t$. Assuming that the maximum degree in the graph is $\delta_{max} = 50$, and $\mu_D = 10$, the variance can be bounded using (Bhatia and Davis, 2000) by

$\sigma_D^2 \leq (\delta_{max} - \mu_D)(\mu_D - 1)$. To satisfy this bound, and still violate the assumption on degree variance in Theorem 4.2, would require $f_t \geq 0.2$, i.e., for 20% of the population to be in the hidden population (e.g., positive with COVID-19) simultaneously. This is unrealistically high, noting that the highest number of COVID-19 tests performed (which is much higher than reported positive cases) in a week in California was approximately 11.2% ($< 20\%$) of the population. Theorem 4.2 also motivates framing the indirect survey questions in such a way that the variance of the graph is small.

4.2.2. Leveraging Smoothness

The hidden population in many real-life processes is **driven by smooth dynamic processes**, leading to the following.

Observation 4.2. *$|\Delta f_t| \leq \epsilon_{f,1} f_t$ and $|\Delta^2 f_t| \leq \epsilon_{f,2} f_t$, for some small $\epsilon_{f,1}, \epsilon_{f,2} \geq 0$.*

This is reasonable for epidemics as the epidemiology follows smooth dynamics, which, over a large population, should produce smooth case counts. The reported data may appear to be noisy due to reporting behavior, schedule, and delayed dumps (a death occurring today may be reported 1-2 weeks from now). However, the artifacts of reporting to the state dashboards are not something we wish to capture. Instead, we wish to capture the actual incidence of cases based on surveys. An individual will know if their friends are infected independently of how and when it is reported to the state dashboards. To support this observation, we compute f_t from values for COVID-19 reported cases (after denoising and removing outliers) in California based on the number of people who were reported positive within the previous 7 days of t. Figure 4.1 shows the value of $|\Delta f_t|/f_t$ and $\Delta^2|f_t|/f_t$ over time. Note that $|\Delta^2 f_t|/f_t \ll |\Delta f_t|/f_t \ll 1$. The same is observed for a simulated epidemic (see Section 4.3.1 for simulation details). We repeat the analysis for $\sigma^2_{\bar{X}_t}$. We calculated σ_{X_t} by setting $\mu_D = 15, \sigma_D^2 = 100, \phi_t = f_t$. These smoothness properties are used to derive our results that demonstrate that a weighted smoothing of responses provides better estimates of f_t compared to unsmoothed estimation using $\bar{X}_t$.

Now, we present two results (Lemmas 4.3 and 4.4) that help ***bound the result of aggregating over smooth sequences***.

Lemma 4.3. *For any non-negative sequence g_t such that $|\Delta g_t| \leq \epsilon_{g,1} \cdot g_t, \forall t$ for some $\epsilon_{g,1} \geq 0$, let $g_{max} = \max\{g_t, g_{t+1}, \ldots, g_{t+j}\}$. Then $|g_{t+j} - g_t| \leq \frac{|j|\epsilon_{g,1}}{1-|j|\epsilon_{g,1}} g_t \, \forall j \in \mathbb{Z}$.*

This implies that a time series with bounded differences will not change rapidly in a small window of time.

Lemma 4.4. *For any non-negative sequence g_t such that $|\Delta^2 g_t| \leq \epsilon_{g,2} \cdot g_t, \forall t$ for some $\epsilon_{g,2} \geq 0$, $\left| \frac{\sum_{i=-w}^{w} g_{t+i}}{2w+1} - g_t \right| \leq g_t \cdot E_g(w)$, where, $E_g(w) = \frac{w(w+1)}{6}\epsilon_{g,2} + o(w^4 \epsilon_{g,2}^2)$.*

This implies that if we apply a moving average smoothing to a time-series g_t, the resulting time-series $g_{t,w} = \sum_{i=-w}^{w} g_{t+i}/(2w+1)$ is close to g_t for small windows.

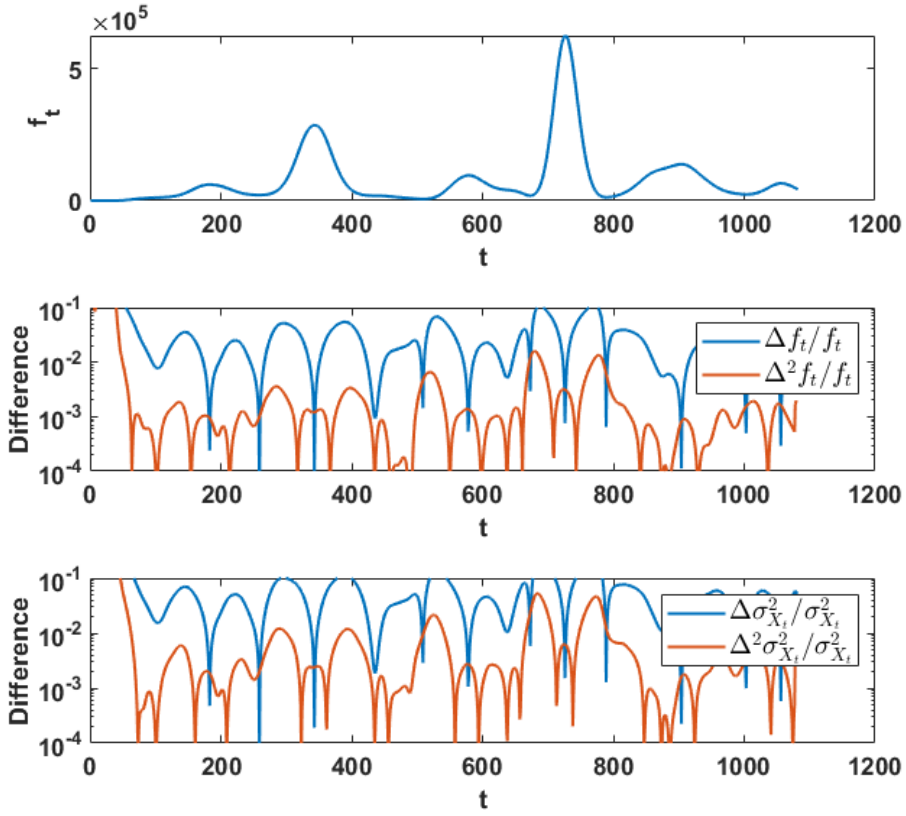


Figure 4.1: The first and second differences for f_t and σ_t are small – derived from denoised COVID-19 cases in California. t is the number of days since January 23, 2020.

Demonstrating the Advantage of Smoothing

We will use the fact that typical real-world signals to be estimated, f_t, are smooth (Observation 4.2) to find better estimates through aggregation/a weighted moving average. Instead of trying to estimate $\mathbb{E}(X_t)$ from data, we can try to estimate some aggregation over a window, $\mathbb{E}(X_{t,w})$. Therefore, we can use more responses, which may decrease sample variance but may also introduce an error as $\mathbb{E}(X_{t,w}) \neq \mathbb{E}(X_t)$. The following lemma identifies the conditions when such aggregation is better than individually estimating $\mathbb{E}(X_t)$. We will measure this by finding λ so that smoothing $(\bar{X}_{t,w}/\mu_D)$ is less likely to result in a *fractional error* (ratio of difference from f_t to f_t) greater than some λ compared to no smoothing $(\bar{X}_t/\mu_D)$.

Lemma 4.5. *Let $\bar{X}_{t,w}$ be some linear combination of $\{\bar{X}_{t-w}, \ldots, \bar{X}_{t+w}\}$ such that $|\mathbb{E}(\bar{X}_{t,w})/\mu_D - f_t| \leq \lambda' f_t$. Then, the probability of fractional error by λ is lower in the smoothed response than the unsmoothed response, $P\left(|\frac{\bar{X}_{t,w}}{\mu_D} - f_t| \geq \lambda f_t\right) \leq P\left(|\frac{\bar{X}_t}{\mu_D} - f_t| \geq \lambda f_t\right)$ if $\lambda \geq \lambda' / \left(1 - \frac{\sigma_{\bar{X}_{t,w}}}{\sigma_{X_t}/\sqrt{n_t}}\right)$.*

Lemma 4.5 suggests that ***an aggregation across the window is good if*** (i) λ' is small, i.e., $\mathbb{E}(\bar{X}_{t,w})$ does not deviate too much from $\mathbb{E}(X_t)$, and (ii) $\sigma_{\bar{X}_{t,w}} \ll \sigma_{X_t}/\sqrt{n_t}$.

Let $\bar{X}_{t,w}$ be the random variable defined as $\bar{X}_{t,w} = \sum_{i=-w}^{w} \frac{n_{t+i}}{n_w} \bar{X}_{t+i}$, where n_t is the number of responses at time t and $n_w = \sum_{i=-w}^{w} n_i$. To see why this is a good aggregation, we make the following observation.

Observation 4.3. *Over a selected window w, the responses at different t are independent of each other.*

At each time t, we randomly select individuals to respond to the survey, and therefore the responses are independent. This will be violated in some extreme cases, such as surveying a highly infectious disease where the respondents happen to be the same every day. Then the response from the same person on consecutive surveys may become dependent. However, we assume that such extreme cases do not occur. Further, this provides another guideline for designing such surveys, i.e., avoiding asking a fixed set of individuals.

Assumption 4.2. *Over a selected window w, for a pair of nodes v_1 and v_2, with a common neighbor, $\mathbb{E}(I_{v_1} I_{v_2}) = \phi_t f_t, \forall t$, for some smooth $\phi_t \in [0, 1]$, such that $|\Delta^2 \phi_t| \leq \epsilon_\phi$*

This is reasonable because the covariance of the infection state of two randomly selected nodes with a common neighbor should not vary rapidly over time. Recall that $\sigma^2_{X_t} = f_t(\mu_D^2(\phi_t - f_t) + \mu_D(1-\phi_t) + \sigma_D^2 \phi_t)$. Since all terms of $\sigma^2_{X_t}$ are product of smooth functions, for some $\epsilon_{\sigma^2}, |\Delta \sigma^2_{X_t}| \leq \epsilon_{\sigma^2}$. This is also demonstrated in Figure 4.1. For the sake of demonstration, we calculated σ_{X_t} by setting $\mu_D = 15, \sigma_D = 10, \phi_t = f_t$. With Observation 4.3 and Assumption 4.2, we are ready to prove the following theorem.

Theorem 4.3. *The probability of fractional error of λ is lower in the smoothed response compared to the unsmoothed response, $P\left(\left|\frac{\bar{X}_{t,w}}{\mu_D} - f_t\right| \geq \lambda f_t\right) \leq P\left(\left|\frac{\bar{X}_t}{\mu_D} - f_t\right| \geq \lambda f_t\right)$ if*

$$\lambda \geq w\epsilon_{f,1} / \left(1 - \left(1 + \frac{w\epsilon_{\sigma^2,1}}{1 - w\epsilon_{\sigma^2,1}}\right) \sqrt{n_t / n_w}\right). \tag{4.2}$$

The theorem suggests that **the smoothed response is less likely to deviate** by some small λ from the true value compared to the unsmoothed response. The inequality that λ needs to satisfy to justify smoothing suggests that if the first differences of the hidden time series and its variance are small, we can consider a larger window to smooth the responses. Further, $n_t \ll n_w$ is desirable.

Stronger Results when Variance of n_t is Small

Assume that the number of responses per unit time n_t does not vary drastically over a window.

Assumption 4.3. $\sigma_n / \mu_n \ll 1$.

This may not be true if we aggregate responses for each day since, within a week, weekdays may have different patterns than weekends. However, for aggregated weekly observations, n_t may not vary significantly. Suppose $\mu_n = n_w/(2w+1)$ and σ_n represent the mean and standard deviation of $\{n_{t-w}, \ldots, n_{t+w}\}$, respectively.

Lemma 4.6. *For any smoothly varying sequence g_t with bounded second difference, if $\sigma_n/\mu_n < 1$, then $\left|\sum_{i=-w}^{w} \frac{n_{t+i}}{n_w} g_{t+i} - g_t\right| \leq g_t\gamma_g$, for some small $\gamma_g \geq 0$.*

This leads to a **better error bound** from indirect surveys.

Theorem 4.4. *The probability of fractional error of λ is lower in the smoothed response compared to the unsmoothed response, $P\left(\left|\frac{\bar{X}_{t,w}}{\mu_D} - f_t\right| \geq \lambda f_t\right) \leq P\left(\left|\frac{\bar{X}_t}{\mu_D} - f_t\right| \geq \lambda f_t\right)$ if*

$$\lambda \geq \gamma_f / \left(1 - \sqrt{\frac{n_t}{n_w}(1+\gamma_{\sigma^2})}\right), \tag{4.3}$$

where $\gamma_g = E_g(w) + \epsilon_{g,1}\frac{\sigma_n}{\mu_n}\frac{w\epsilon_{g,1}}{1-w\epsilon_{g,1}}$ and $E_g(w) \approx \frac{w(w+1)}{6}\epsilon_{g,2}$.

To demonstrate that γ_f and γ_{σ^2} are indeed small, we calculate them for various window sizes over COVID-19 reported cases (Figure 4.2). Their values increase with larger w (recall that the window size is $2w+1$). For these calculations, we set $\sigma_n/\mu_n = 0.3$. For small windows, the values are small, and so smoothing is advantageous. As expected, for large windows, the signal is oversmoothed, resulting in higher values of γ_f and γ_{σ^2}, consequently higher errors.

4.3. Results and Analysis

4.3.1. Synthetic Experiments

To evaluate our claims and analyze the effect of various variables, we ran an epidemic simulation in conjunction with the simulation of surveys over randomly generated networks[3].

Epidemic simulation We use an extended SIR model to simulate an epidemic with varying infection parameters. The infection parameter starts at a value so that the reproduction number R0 (Dietz, 1993) is above 2. At random times, we introduce "interventions" that reduce R0 smoothly to a value below 1. We run several such simulations and pick one that produces multiple peaks over 600 days, to emulate complex, realistic epidemics like Influenza and COVID-19 that have multiple waves. We acknowledge that, in reality, not all infections will be detectable. However, assuming that each infection will be detected with a fixed probability only scales the

[3]Our code is available at https://github.com/GCGImdea/coronasurveys/tree/master/papers/2024-AAAI-Nowcasting-Temporal-Trends-Using-Indirect-Surveys.

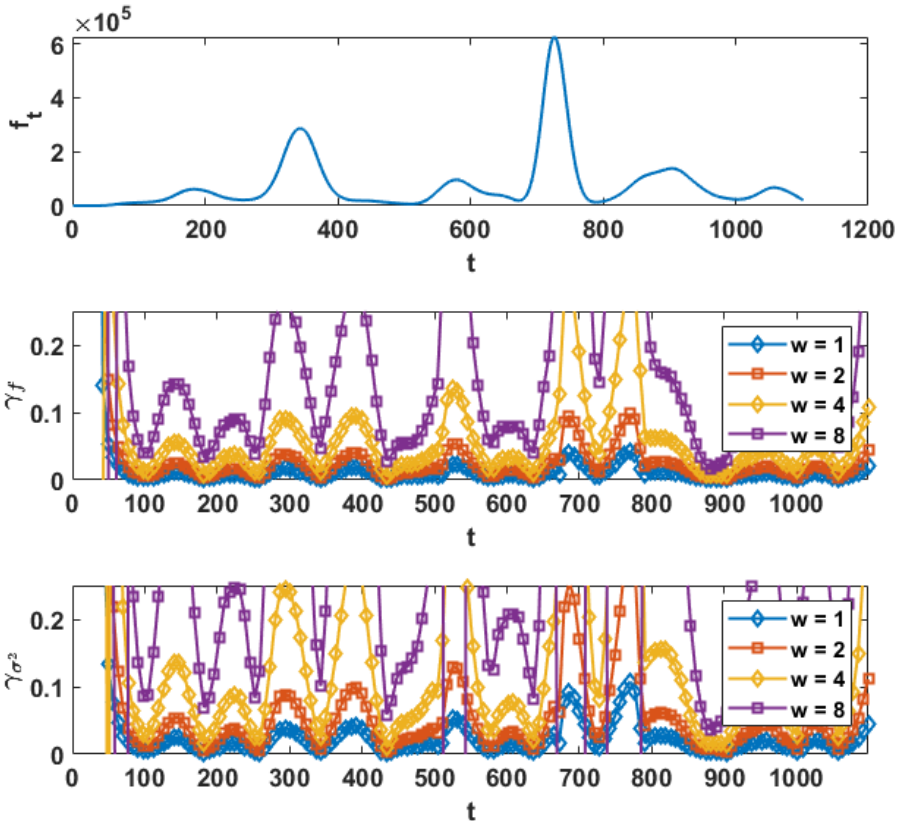


Figure 4.2: γ_f and γ_{σ^2} calculated over various window sizes for COVID-19 reported cases in California. The values are small as desired, particularly for small w.

time-series $I(t)$ by a constant. Therefore, for the purpose of this study, we directly use $I(t)$ to compute the hidden population over time.

Survey simulation We simulate sampling nodes (respondents) from a graph with a power law distribution $p_k \propto k^{-2}$, with a bounded maximum degree, such that the mean degree is approximately a parameter d. The choice of distribution was driven by the intention to introduce some skewness in the degree distribution to push the limits of our approach. The main conclusions do not change on Erdos-Renyi graphs. The simulation has the following parameters. (i) d: approximate average degree in the latent graph; (ii) n: upper limit on the number of individuals who respond on a given day. The actual number is uniformly selected from 1 to n; (iii) n_d: number of nodes that can potentially be covered by the respondents; (iv) *accum*: number of days over which the responses are accumulated ; (v) *period*: time window within which the respondents are to count the hidden population. E.g., if the question is "how many people do you know who have had COVID-like illnesses in the last 7 days," then $period = 7$. For each combination of the above parameters, we ran 16 simulations, resulting in **82,000 combinations** of parameters and simulations. For indirect surveys, we randomly infect each neighbor of the responders with the probability $\sum_{\tau=0}^{period-1} I(t-\tau)$ and obtain the indirect response from each responding

node. We then accumulate the responses obtained over *accum* number of days. For comparison, we also introduced the traditional NSUM approach (Killworth et al., 1998a), where the response from each node is normalized by its degree. For direct surveys, we infect nodes among the responders and note the number of infections produced.

Results

For each of the indirect (**Ind**), NSUM, and direct (**Dir**) survey methods, we introduced the following post-processing methods: (1) **NoS**: Average response for each time unit (defined by *accum*) without any smoothing. (2) **WA**: Moving average of NoS weighted by the number of responses over a window w. (3) **UA**: Unweighted moving average of NoS over a window w. These methods were compared against the infections $I(t)$ using MAE with time granularity redefined by the choice of *accum*. Before computing the error, we apply a range normalization, so that the maximum of each time series is set to 1 and the minimum to 0. Note that here, $I(t)$ is not the same as f_t. To construct f_t, we would count the number of infections in the last *period* number of days. Secondly, note that setting the parameter $accum > 1$ is equivalent to performing a weighted average (scaled by a constant factor). Therefore, *accum* > 1 and $w = 0$ will have a similar effect as using $w > 1$ on $accum = 1$.

Figure 4.4 shows the result of the survey simulation with parameters $d = 5, n = 10, n_d = 480, accum = 14, period = 7$, and $w = 1$. Time series obtained from the smoothed indirect survey are much more similar to the true infections $I(t)$ compared to those obtained from direct surveys. The time series of smoothed NSUM is also close to the true infections, but at times, worse than our indirect method.

Figure 4.3 shows the distribution of errors obtained by different methods. In this figure, to focus on the impact of one parameter, we fix the others ($d = 5$, $n = 20$, $n_d = 60$, $period = 7$, $accum = 7$, and $w = 2$). In general, we note that the indirect methods (Ind-*) produce lower median errors than NSUM-* the direct methods (Dir-*). Also, there is no significant difference between weighted and unweighted smoothing strategies. In terms of parameters, the choice of d and n_d does not impact the relative patterns across the 9 methods (see Supplementary Material in full version (Srivastava et al., 2023)). As expected from our analysis (Theorems 4.3 and 4.4), increasing *accum* first decreases the errors, but a high value worsens the performance for the moving averages (*-UA, *-WA). A similar observation can be made for increasing w – *-NoS, which has $w = 0$ produces a higher error than $w = 1$. All moving averages become similar as w increases to 8.

4.3.2. Real Dataset

The objective of these experiments is to evaluate the performance of the proposed approach using datasets drawn from the US COVID-19 Trends and Impact Survey (CMU-CTIS) (Salomon et al., 2021). It has data on self-reported symptoms, symptoms in the respondent's community, testing, isolation measures, vaccination acceptance, and mental health, among other factors, to assess the spread of COVID-19. Approximately 40,000 US respondents participated in this survey daily between April 6, 2020, and June 25, 2022. We chose the period of September 8, 2020, through March 1, 2022, because CMU-CTIS included a question regarding individual COVID-19 positive test results, which is used to estimate the direct survey results. A null value detection and removal, as well as an outlier filter are applied to the dataset for each state.

Figure 4.5 shows the normalized COVID-19 incidence curves from direct and indirect surveys for California from September 2020 to March 2022. For the curves derived from indirect surveys, we included both the CLI incidence reported in the household (**Indirect1**) and the CLI incidence reported within the local community (**Indirect2**). In addition, Figure 4.5 displays the curve obtained by the NSUM method (Killworth et al., 1998a) using household questions (CTIS has a question asking for the size of the household). For these curves, we set the parameters to $accum = 7$ and $w = 1$. For comparison purposes, we include the normalized incidence curves obtained from datasets provided by the Johns Hopkins Coronavirus Resource Center (Dong et al., 2020). We observe that the curves obtained from indirect surveys are very similar to the official ones. To quantitatively evaluate the proposed approach, Table 4.1 shows the MAE of the normalized incidence curves obtained from direct, NSUM, and indirect surveys conducted in California, Texas, New York, and Pennsylvania for a variety of values of $accum$ and w. The reference curves are based on official data. For each ($accum$, w) pair, a bold font and underlined values correspond to the best and the second-best values, respectively. As given in Table 4.1, the incidence curves obtained from indirect surveys exhibit lower MAE values than those obtained by both NSUM and direct approaches. Furthermore, the MAE values obtained from indirect surveys in the local community (Indirect2) are generally lower than those extracted from indirect surveys in the household (Indirect1).

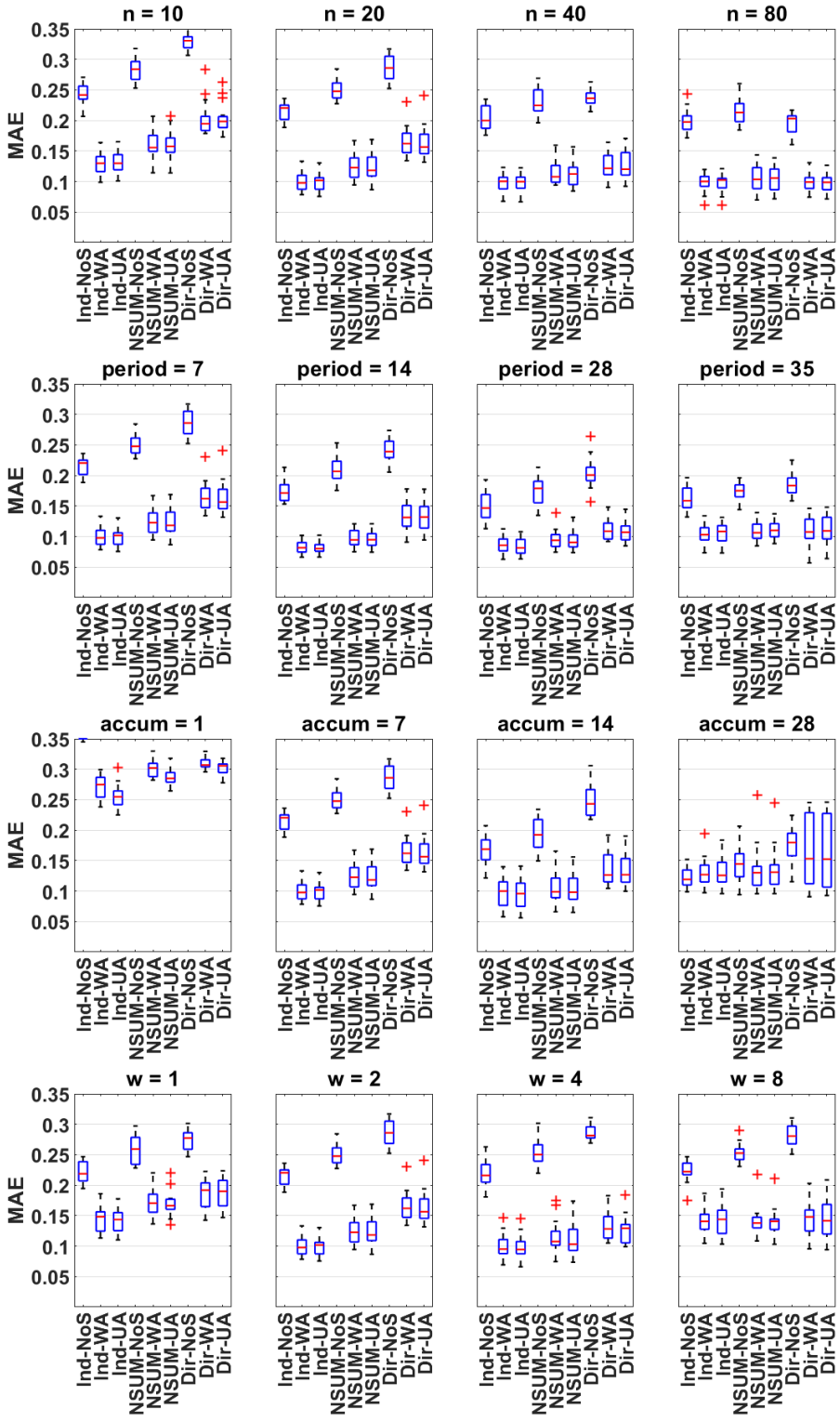


Figure 4.3: 1st row: MAE vs n, the number of respondents. Indirect with moving average is the best for low n. All moving averages converge as n increases. 2nd row: MAE vs *period* in the survey question. Smoothed results converge as *period* increases. 3rd row: MAE vs *accum*, the accumulation width. Errors reduce quickly as *accum* increases. Larger values make the moving averages slightly worse. 4th row: MAE vs w, the smoothing window. Errors reduce as w increases and increase slightly for large w.

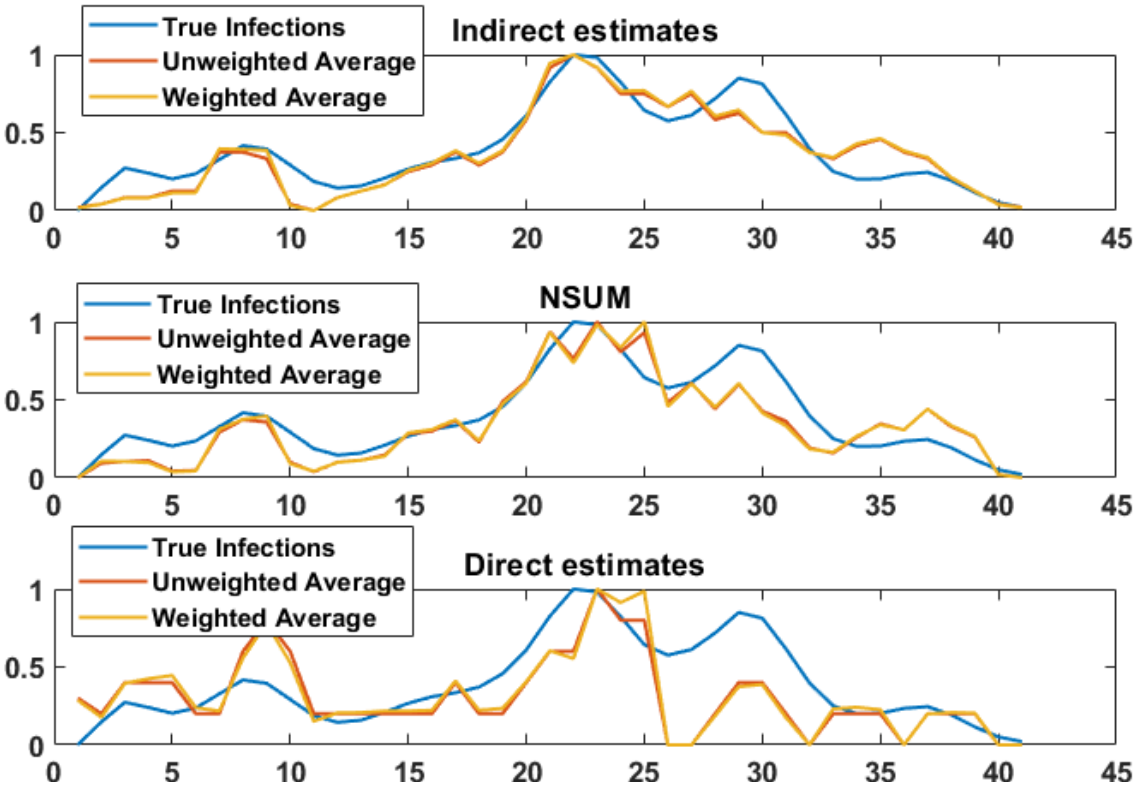


Figure 4.4: Result of one of the survey simulations.

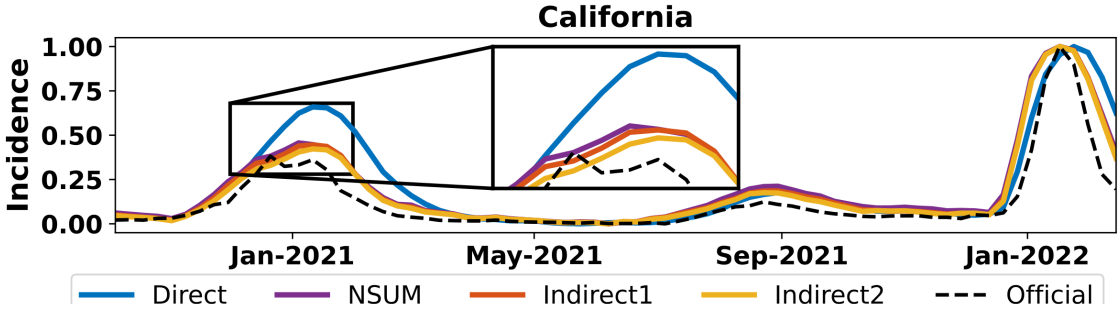


Figure 4.5: Normalized COVID-19 incidence estimated from the US CMU-CTIS data for California. The parameters of the proposed approach were set to $accum = 7$ and $w = 1$.

4.4. Discussion

Survey Design. Our assumptions and results can be used to design better indirect surveys. As lower variance in the degree distribution is desirable due to Theorem 4.2, the question can be framed in a way to keep the variance low. For instance, instead of simply asking "how many people do you know who ... ", we could restrict the set of people to be counted – "Among your and your two immediate neighboring households, how many do you know who ... ."

Targeted Surveys. Our approach allows the responses to be from a restricted subset of the population. This is advantageous for targeted surveys. For instance, healthcare workers are more likely to know of Influenza hospitalizations than the general public. Therefore, we can restrict the survey to them to estimate the number of hospitalizations over time. Further, any direct survey on an online platform can only estimate the hidden subpopulation among those using the platform. Instead, an indirect survey will cover all the neighbors of the platform users in the latent graph.

accum	w	state	Direct	NSUM	Indirect1	Indirect2
7	1	CA	0.0703	0.0443	0.0341	**0.0292**
		TX	0.0661	0.0379	0.0289	**0.0270**
		NY	0.0785	0.0315	0.0301	**0.0299**
		PN	0.0572	0.0368	0.0300	**0.0263**
	3	CA	0.1148	0.0988	0.0881	**0.0811**
		TX	0.1236	0.0890	0.0813	**0.0782**
		NY	0.1210	0.0930	0.0910	**0.0886**
		PN	0.0956	0.0907	0.0816	**0.0691**
14	1	CA	0.0836	0.0624	0.0524	**0.0477**
		TX	0.0779	0.0385	0.0343	**0.0336**
		NY	0.0929	0.0520	0.0504	**0.0500**
		PN	0.0689	**0.0389**	0.0391	0.0429
	3	CA	0.1441	0.1217	0.1116	**0.1059**
		TX	0.1349	0.1058	0.1042	**0.1027**
		NY	0.1571	0.1165	0.1126	**0.1090**
		PN	0.1349	0.1182	0.1110	**0.1005**

Table 4.1: MAE of the normalized COVID-19 incidence curves estimated from the US CMU-CTIS data for California, Texas, New York, and Pennsylvania and for different values of *accum* and w.

Extension to Biased Reporting In this paper, we have assumed that the survey responses are honest and correct. However, in reality, the reporting could be biased to exaggerate counts or undercount, or the respondents may incorrectly recall. A

detailed analysis of bias is beyond the scope of this paper. However, it can be shown that all our results apply if the respondents have different biases that underestimate or overestimate at the same rate over time – these biases scale the estimate by a constant factor. These results do not take into account respondents who are malicious, deliberately inaccurate (possibly due to the sensitivity of the questions), or exhibit other behaviors. In practice, additional steps can be taken to handle different biases (Scheers, 1992; Kazemzadeh et al., 2016; Ezoe et al., 2012; Salganik et al., 2011).

4.5. Conclusions

We have proposed a latent graph formulation to estimate the temporal trends in the size of a hidden population from indirect surveys, leading to better estimates than those achievable with direct surveys having the same number of responses. We leveraged the temporal dynamics of the underlying process and identified the conditions under which a weighted moving average of responses leads to better estimates compared to raw responses. We performed extensive simulations of a temporal process over which a simulated survey is performed, to study the impact of various parameters on the estimation error. We demonstrated that our approach outperforms traditional Network Scale-Up Methods and the direct approach with and without performing a moving average. We also demonstrated that our approach is able to better estimate the trend of COVID-19 cases on real-world surveys over time.

Ethical Statement

The Ethics Board (IRB) of IMDEA Networks Institute approved this work on 2021/07/05. IMDEA Networks has signed Data Use Agreements with Facebook and Carnegie Mellon University (CMU) to access their data. Specifically, CMU project STUDY2020_00000162 entitled ILI Community-Surveillance Study. Informed consent has been obtained from all participants in this survey by this institution. All the methods in this study have been carried out in accordance with relevant ethics and privacy guidelines and regulations.

Availability of Data and Materials

The data presented in this paper (in aggregated form) and the codes used to process it will be made publicly available at https://github.com/GCGImdea/coronasurveys/tree/master/papers/2024-AAAI-Nowcasting-Temporal-Trends-Using-Indirect-Surveys. The microdata of the CTIS survey from which the aggregated

data were obtained cannot be shared, as per the Data Use Agreements signed with Facebook and Carnegie Mellon University (CMU).

Acknowledgements

This work was partially supported by grants TED2021-131264B-I00 (SocialProbing) and PID2022-137243OB-I00, funded by MCIN/AEI/10.13039/501100011033 and the European Union "NextGenerationEU"/PRTR. The authors would like to thank Mohamed Kacem for his contribution to pre-processing the data used in this work.

References

J. M. Alix-Garcia, K. R. Sims, and L. Costica. Better to be indirect? testing the accuracy and cost-savings of indirect surveys. World Development, 142:105419, 2021.

C. M. Astley et al. Global monitoring of the impact of the COVID-19 pandemic through online surveys sampled from the Facebook user base. Proceedings of the National Academy of Sciences, 118(51), 2021.

H. R. Bernard, E. C. Johnsen, P. D. Killworth, and S. Robinson. Estimating the size of an average personal network and of an event subpopulation. In The Small World, pages 159–175, 1989.

R. Bhatia and C. Davis. A better bound on the variance. The american mathematical monthly, 107(4):353–357, 2000.

E. Breza, A. G. Chandrasekhar, T. H. McCormick, and M. Pan. Using aggregated relational data to feasibly identify network structure without network data. American Economic Review, 110(8):2454–84, 2020.

K. Dietz. The estimation of the basic reproduction number for infectious diseases. Statistical methods in medical research, 2(1):23–41, 1993.

E. Dong, H. Du, and L. Gardner. An interactive web-based dashboard to track covid-19 in real time. The Lancet infectious diseases, 20(5):533–534, 2020.

R. Dunbar. How many friends does one person need? Dunbar's number and other evolutionary quirks. Harvard University Press, 2010.

S. Ezoe, T. Morooka, T. Noda, M. L. Sabin, and S. Koike. Population size estimation of men who have sex with men through the network scale-up method in japan. PloS one, 7(1):e31184, 2012.

D. M. Feehan and M. J. Salganik. Generalizing the network scale-up method: a new estimator for the size of hidden populations. Sociological methodology, 46(1): 153–186, 2016.

A. Garcia-Agundez et al. Estimating the covid-19 prevalence in spain with indirect reporting via open surveys. Frontiers in Public Health, 9:658544, 2021.

P. Geldsetzer. Knowledge and perceptions of covid-19 among the general public in the united states and the united kingdom: a cross-sectional online survey. Annals of internal medicine, 173(2):157–160, 2020.

L. Jing, Q. Lu, Y. Cui, H. Yu, and T. Wang. Combining the randomized response technique and the network scale-up method to estimate the female sex worker population size: an exploratory study. Public health, 160:81–86, 2018.

Y. Kazemzadeh, M. Shokoohi, M. R. Baneshi, and A. A. Haghdoost. The frequency of high-risk behaviors among iranian college students using indirect methods: network scale-up and crosswise model. International journal of high risk behaviors & addiction, 5(3), 2016.

P. D. Killworth, E. C. Johnsen, C. McCarty, G. A. Shelley, and H. R. Bernard. A social network approach to estimating seroprevalence in the united states. Social networks, 20(1):23–50, 1998a.

P. D. Killworth, C. McCarty, H. R. Bernard, G. A. Shelley, and E. C. Johnsen. Estimation of seroprevalence, rape, and homelessness in the united states using a social network approach. Evaluation review, 22(2):289–308, 1998b.

I. Laga, L. Bao, and X. Niu. Thirty years of the network scale-up method. Journal of the American Statistical Association, 116(535):1548–1559, 2021.

N. Oliver et al. Assessing the impact of the covid-19 pandemic in spain: large-scale, online, self-reported population survey. Journal of medical Internet research, 22 (9):e21319, 2020.

C. Rossier. The anonymous third party reporting method. Methodologies for estimating abortion incidence and abortion-related morbidity: a review, pages 99–106, 2010.

M. Salganik et al. Estimating the number of heavy drug users in curitiba, brazil using multiple methods. Technical report, Technical report. UNAIDS, 2010.

M. J. Salganik et al. The game of contacts: estimating the social visibility of groups. Social networks, 33(1):70–78, 2011.

J. A. Salomon et al. The us covid-19 trends and impact survey: Continuous real-time measurement of covid-19 symptoms, risks, protective behaviors, testing, and vaccination. Proceedings of the National Academy of Sciences, 118(51):e2111454118, 2021.

N. Scheers. A review of randomized response techniques. Measurement and Evaluation in Counseling and Development, 1992.

A. Srivastava, J. M. Ramírez, S. Díaz-Aranda, J. Aguilar, A. Ortega, A. Fernández Anta, and R. E. Lillo. Nowcasting temporal trends using indirect surveys, 2023.

A. K. J. Teo et al. Estimating the size of key populations for hiv in singapore using the network scale-up method. Sexually transmitted infections, 95(8):602–607, 2019.

Supplementary Material

4.S.A. Proofs

Here we provide detailed proofs of all the Theorems and Lemmas used in the paper.

Proof of Theorem 4.1. Let p_k be the fraction of nodes with in-degree k in G. Consider a node partitioning graph G into sets $\{A_1, A_2, \dots\}$, where A_k represents the set of all nodes of in-degree k. By Assumption 4.1, for any v belonging to the neighborhood of some $u \in A_k$, the indicator random variable I_{vk} of the event $v \in H_t$ is independent of k, i.e., $\mathbb{E}(I_{vk}) = f_t$. Now, if we pick a random node u from A_k for the indirect survey, due to linearity of expectation,

$$\mathbb{E}(X_t|A_k) = \mathbb{E}\left(\sum_{(v,u)\in E} I_{vk}|A_k\right) \tag{4.4}$$

$$= \sum_{(v,u)\in E} \mathbb{E}(I_{vk}|A_k) = kf_t\,. \tag{4.5}$$

When we randomly select a node u in G, $P(u \in A_k) = p_k$. By the law of total expectation,

$$\mathbb{E}(X_t) = \sum_k \mathbb{E}(X_t|A_k)P(A_k) \tag{4.6}$$

$$= \sum_k kf_t \cdot p_k = (\sum_k kp_k)f_t = \mu_D f_t\,. \tag{4.7}$$

□

Proof of Lemma 4.1. We wish to compute the variance using $\sigma^2_{X_t} = \mathbb{E}(X_t^2) - (\mathbb{E}(X_t))^2$. We already have $(\mathbb{E}(X_t))^2 = \mu_D^2 f_t^2$. By the law of total expectation, $\mathbb{E}(X_t^2) = \mathbb{E}(X_t^2|A_k)P(A_k)$.

$$\mathbb{E}(X_t^2) = \sum_k \mathbb{E}(X_t^2|A_k)P(A_k) \tag{4.8}$$

$$= \sum_k \mathbb{E}\left(\left(\sum_{(v_i,u),(v_j,u)\in E} I_{v_i}I_{v_j})\right)^2 |A_k\right) p_k \tag{4.9}$$

$$= \sum_k \left(\sum_{i=1}^{k} \mathbb{E}(I^2_{v_{ik}}) + \sum_{\substack{i,j=1 \\ i\neq j}}^{k} \mathbb{E}(I_{v_i}I_{v_j})\right) p_k$$

$$= \sum_k \left(kf_t + \sum_{\substack{i,j=1 \\ i\neq j}}^{k} \mathbb{E}(I_{v_i}I_{v_j})\right) p_k \tag{4.10}$$

Since, I_{v_i} and I_{v_j} are non-negative, and $I_{v_i} I_{v_j} \leq \min\{I_{v_i}, I_{v_j}\}$,

$$0 \leq \mathbb{E}(I_{v_i} I_{v_j}) \leq f_t \,. \tag{4.11}$$

Putting $\mathbb{E}(I_{v_i} I_{v_j}) = \phi_t f_t$ in Equation 4.10, we get

$$\begin{aligned}
\mathbb{E}(X_t^2) &= \sum_k p_k \left(k f_t + (k^2 - k)\phi_t f_t \right) \\
&= \mu_D f_t + (\sigma_D^2 + \mu_D^2)\phi_t f_t - \mu_D \phi_t f_t \\
&= f_t(\mu_D + \sigma_D^2 \phi_t + \mu_D^2 \phi_t - \mu_D \phi_t) \,.
\end{aligned}$$

Therefore,

$$\begin{aligned}
\sigma_{X_t}^2 &= \mathbb{E}(X_t^2) - (\mathbb{E}(X_t))^2 \\
&= f_t(\mu_D + \sigma_D^2 \phi_t + \mu_D^2 \phi_t - \mu_D \phi_t) - \mu_D^2 f_t^2 \\
&= f_t(\mu_D^2(\phi_t - f_t) + \mu_D(1 - \phi_t) + \sigma_D^2 \phi_t) \,.
\end{aligned} \tag{4.12}$$

Since, $0 \leq \phi_t \leq 1$, we have $\mu_D f_t(1 - \mu_D f_t) \leq \sigma_{X_t}^2 \leq f_t(\sigma_D^2 + \mu_D^2(1 - f_t))$. □

Proof of Theorem 4.2. By Lemma 4.2,

$$P(|\bar{Y}_t - f_t| > \lambda) = P\left(\left| \frac{\bar{Y}_t - f_t}{\sigma_{Y_t}/\sqrt{n}} \right| > \frac{\sqrt{n}\lambda}{\sigma_{Y_t}} \right) = 1 - erf\left(\frac{\sqrt{n}\lambda}{\sigma_{Y_t}} \right) , \tag{4.13}$$

where erf is the error function, and the equation follows from the fact that $\left(\frac{\bar{Y}_t - f_t}{\sigma_{Y_t}/\sqrt{n}} \right)$ follows a standard normal distribution. Similarly,

$$\begin{aligned}
P(|\bar{X}_t/\mu_D - f_t| > \lambda) &= P\left(\left| \frac{\bar{X}_t - \mu_D f_t}{\sigma_{X_t}/\sqrt{n}} \right| > \frac{\mu_D \sqrt{n}\lambda}{\sigma_{X_t}} \right) \\
&= 1 - erf\left(\frac{\mu_D \sqrt{n}\lambda}{\sigma_{X_t}} \right) ,
\end{aligned} \tag{4.14}$$

Suppose the probability of deviation is higher from the indirect estimate. Then, from Equations 4.13 and 4.14,

$$\begin{aligned}
& P(|\bar{X}_t/\mu_D - f_t| > \lambda) > P(|\bar{Y}_t - f_t| > \lambda) \\
\implies & 1 - erf\left(\frac{\mu_D \sqrt{n}\lambda}{\sigma_{X_t}} \right) > 1 - erf\left(\frac{\sqrt{n}\lambda}{\sigma_{Y_t}} \right) \\
\implies & erf\left(\frac{\mu_D \sqrt{n}\lambda}{\sigma_{X_t}} \right) < erf\left(\frac{\sqrt{n}\lambda}{\sigma_{Y_t}} \right)
\end{aligned}$$

Since erf is an increasing function,

$$\frac{\mu_D \sqrt{n}\lambda}{\sigma_{X_t}} < \frac{\sqrt{n}\lambda}{\sigma_{Y_t}} \implies \sigma_{X_t} > \mu_D \sigma_{Y_t} \implies \sigma_{X_t}^2 > \mu_D^2 \sigma_{Y_t}^2 \,. \tag{4.15}$$

Putting the value of $\sigma^2_{X_t}$ from Lemma 4.1, combining with Equation 4.15, after some algebraic manipulations gives

$$\sigma^2_D > \mu_D(\mu_D - 1)(1 - \phi_t)/\phi_t. \tag{4.16}$$

Hence, if $\sigma^2_D \leq \mu_D(\mu_D - 1)(1 - \phi_t)/\phi_t$, then $P(|\bar{X}_t/\mu_D - f_t| > \lambda) \leq P(|\bar{Y}_t - f_t| > \lambda)$. □

Proof of Lemma 4.3. First, assume that $j > 0$.

$$\begin{aligned}
|g_{t+j} - g_t| &= |g_{t+j} - g_{t+(j-1)} + g_{t+(j-1)} + \cdots + g_{t+1} - g_t| \\
&\quad \text{[Adding and subtracting the terms } g_{t+(j-1)}, \ldots, g_{t-1}\text{]} \\
&\leq |g_{t+j} - g_{t+(j-1)}| + |g_{t+(j-1)} - g_{t+(j-2)}| + \cdots + |g_{t+1} - g_t| \\
&\leq \epsilon_{g,1} \sum_{i=1}^{j} g_{t+i} \leq \epsilon_{g,1} j g_{max}\,.
\end{aligned} \tag{4.17}$$

Since the index of g_{max} must be less than j, we can reapply the above as

$$g_{max} - g_t \leq j\epsilon_{g,1} g_{max} \implies g_{max} \leq \frac{1}{1 - j\epsilon_{g,1}} g_t \tag{4.18}$$

From Equations 4.17 and 4.18, we get

$$|g_{t+j} - g_t| \leq \frac{j\epsilon_{g,1}}{1 - j\epsilon_{g,1}} gt\,. \tag{4.19}$$

Proof for $j < 0$ is similar. □

Proof of Lemma 4.4. Let $S_j = g_{t+j} + g_{t-j}$. Since, $\Delta^2 g_t \leq \epsilon_{g,2} g_t, \forall t$, we have

$$\begin{aligned}
g_{t+1} - 2g_t + g_{t-1} &\leq \epsilon_{g,2} g_t \\
g_{t+1} + g_{t-1} &\leq 2g_t + \epsilon_{g,2} g_t \implies S_1 \leq 2g_t + \epsilon_{g,2} g_t\,.
\end{aligned}$$

And,

$$\begin{aligned}
S_2 &= g_{t+2} + g_{t-2} \\
&\leq 2g_{t+1} - g_t + \epsilon_{g,2} g_{t+1} \\
&\quad + 2g_{t-1} - g_t + \epsilon_{g,2} g_{t-1} = 2S_1 - 2g_t + \epsilon_{g,2} S_1
\end{aligned}$$

More generally,

$$\begin{aligned}
S_j &= g_{t+j} + g_{t-j} \\
&\leq 2g_{t+(j-1)} - g_{t+(j-2)} + \epsilon_{g,2} g_{t+(j-1)} \\
&\quad + 2g_{t-(j-1)} - g_{t-(j-2)} + \epsilon_{g,2} g_{t-(j-1)} \\
&= 2S_{j-1} - S_{j-2} + \epsilon_{g,2} S_{j-1} \\
\implies S_j - S_{j-1} &\leq S_{j-1} - S_{j-2} + \epsilon_{g,2} S_{j-1}
\end{aligned}$$

Telescoping the above to $j = 2$, we get

$$S_j - S_1 \leq S_{j-1} - 2g_t + \epsilon_{g,2} \sum_{i=1}^{j-1} S_i$$

Since, $S_1 \leq 2g_t + \epsilon_{g,2} g_t$, we get

$$\begin{aligned} S_j - S_{j-1} &\leq 2g_t + \epsilon_{g,2} g_t - 2g_t + \epsilon_{g,2} \sum_{i=1}^{j-1} S_i \\ S_j - S_{j-1} &\leq \epsilon_{g,2} g_t + \epsilon_{g,2} \sum_{i=1}^{j-1} S_i \\ &\leq \epsilon_{g,2} g_t + \epsilon_{g,2}(T_{j-1} - g_t) = \epsilon_{g,2} T_{j-1}\,, \end{aligned}$$

Where $T_j = \sum_{i=-j}^{j} g_{t+j}$. Telescoping again down to $j = 2$, we get

$$\begin{aligned} S_j &\leq S_1 + \epsilon_{g,2} \sum_{i=1}^{j-1} T_i \\ \implies T_j &\leq T_{j-1} + (2 + \epsilon_{g,2}) g_t + \epsilon_{g,2} \sum_{i=1}^{j-1} T_i \end{aligned} \tag{4.20}$$

for $j > 1$, and $T_1 = g_t + S_1 \leq (3 + \epsilon_{g,t}) g_t$

We can prove that

$$T_w \leq (2w+1) \left(1 + \frac{w(w+1)}{6} \epsilon_{g,2} + o(w^4 \epsilon_{g,2}^2)\right) g_t\,. \tag{4.21}$$

We omit the details for brevity. The idea is to use induction with base case $w = 1$ for which $T_1 \leq (3 + \epsilon_{g,2}) g_t \leq ((2 \cdot 1 + 1) + \epsilon_{g,2} 1 \cdot 2 \cdot 3/6)\, g_t$. Therefore,

$$\begin{aligned} &\frac{\sum_{i=-j}^{j} g_{t+j}}{2w+1} - g_t \\ &\leq \frac{(2w+1)\left(1 + \frac{w(w+1)}{6}\epsilon_{g,2} + o(w^4\epsilon_{g,2}^2)\right) g_t}{2w+1} - g_t \\ &\leq \left(\frac{w(w+1)}{6}\epsilon_{g,2} + o(w^4\epsilon_{g,2}^2)\right) g_t\,. \end{aligned}$$

Repeating the above analysis replacing $\epsilon_{g,2}$ with $-\epsilon_{g,2}$ and reversing all inequalities,

$$\frac{\sum_{i=-j}^{j} g_{t+j}}{2w+1} - g_t \geq -\left(\frac{w(w+1)}{6}\epsilon_{g,2} + o(w^4\epsilon_{g,2}^2)\right) g_t\,.$$

which completes the proof. □

Proof of Lemma 4.5. By using the triangle inequality, followed by the bound on λ', we have

$$\left|\frac{\bar{X}_{t,w}}{\mu_D} - f_t\right| \leq \left|\frac{\bar{X}_{t,w}}{\mu_D} - \frac{\mathbb{E}(\bar{X}_{t,w})}{\mu_D}\right| + \left|\frac{\mathbb{E}(\bar{X}_{t,w})}{\mu_D} - f_t\right|$$

$$\leq \left|\frac{\bar{X}_{t,w}}{\mu_D} - \frac{\mathbb{E}(\bar{X}_{t,w})}{\mu_D}\right| + \lambda' f_t$$
$$\leq \left|\frac{\bar{X}_{t,w}}{\mu_D} - \frac{\mathbb{E}(\bar{X}_{t,w})}{\mu_D}\right| + \lambda\left(1 - \frac{\sigma_{\bar{X}_{t,w}}}{\sigma_{X_t}/\sqrt{n_t}}\right) f_t$$

Using this inequality, we get

$$P\left(\left|\frac{\bar{X}_{t,w}}{\mu_D} - f_t\right| \geq \lambda f_t\right)$$
$$\leq P\left(\left|\frac{\bar{X}_{t,w}}{\mu_D} - \frac{\mathbb{E}(\bar{X}_{t,w})}{\mu_D}\right| + \lambda\left(1 - \frac{\sigma_{\bar{X}_{t,w}}}{\sigma_{X_t}/\sqrt{n_t}}\right) f_t \geq \lambda f_t\right)$$
$$\leq P\left(\left|\frac{\bar{X}_{t,w}}{\mu_D} - \frac{\mathbb{E}(\bar{X}_{t,w})}{\mu_D}\right| \geq \frac{\lambda f_t \sigma_{\bar{X}_{t,w}}}{\sigma_{X_t}/\sqrt{n_t}}\right)$$
$$\leq P\left(\left|\frac{\bar{X}_{t,w} - \mathbb{E}(X_{t,w})}{\sigma_{\bar{X}_{t,w}}}\right| \geq \frac{\lambda f_t \mu_D}{\sigma_{X_t}/\sqrt{n_t}}\right) = 1 - erf\left(\frac{\lambda f_t \mu_D}{\sigma_{X_t}/\sqrt{n_t}}\right) \tag{4.22}$$
$$= P\left(\left|\frac{\bar{X}_t - \mu_D f_t}{\sigma_{X_t}/\sqrt{n_t}}\right| \geq \frac{\lambda f_t \mu_D}{\sigma_{X_t}/\sqrt{n_t}}\right) = P\left(\left|\bar{X}_t/\mu_D - f_t\right| \geq \lambda f_t\right). \tag{4.23}$$

Equation 4.22 follows from the fact that $\bar{X}_{t,w}$, being a linear combination of normally distributed variables, also follows a normal distribution. □

Proof of Theorem 4.3. By linearity of expectation,

$$\mathbb{E}(\bar{X}_{t,w}) = \sum_{i=-w}^{w} \frac{n_{t+i}}{n_w}\mathbb{E}(X_{t+i}) = \sum_{i=-w}^{w} \frac{n_{t+i}}{n_w}\mu_D f_{t+i}\,. \tag{4.24}$$

Using Lemma 4.3,

$$\mathbb{E}(\bar{X}_{t,w})/\mu_D - f_t = \sum_{i=-w}^{w} \frac{n_{t+i}}{n_w} f_{t+i} - f_t$$
$$\leq \max_{i=-w}^{w} f_{t+i} - f_t \leq \frac{w\epsilon_{f,1}}{1 - w\epsilon_{f,1}} f_t\,.$$

Similarly, $\mathbb{E}(\bar{X}_{t,w})/\mu_D - f_t \geq -w\epsilon_f f_t$, which gives

$$\left|\mathbb{E}(\bar{X}_{t,w})/\mu_D - f_t\right| \leq \frac{w\epsilon_{f,1}}{1 - w\epsilon_{f,1}} f_t \implies \lambda' = \frac{w\epsilon_{f,1}}{1 - w\epsilon_{f,1}}\,. \tag{4.25}$$

Now,

$$\sigma^2_{\bar{X}_{t,w}} = \sum_{i=-w}^{w} \frac{n_i^2}{n_w^2}\sigma^2_{\bar{X}_{t+i}} = \sum_{i=-w}^{w} \frac{n_i^2}{n_w^2}\frac{\sigma^2_{X_{t+i}}}{n_i} \tag{4.26}$$
$$= \frac{1}{n_w}\sum_{i=-w}^{w} \frac{n_i}{n_w}\sigma^2_{X_{t+i}} \tag{4.27}$$
$$\leq \frac{1}{n_w}\max_{i=-w}^{w}\sigma^2_{X_{t+i}} \leq \frac{1 + \frac{w\epsilon_{\sigma^2,1}}{1 - w\epsilon_{\sigma^2,1}}}{n_w}\sigma^2_{X_t}$$

$$\implies \frac{\sigma_{\bar{X}_{t,w}}}{\sigma_{X_t}/\sqrt{n_t}} \leq \sqrt{\frac{n_t}{n_w}}\left(1+\frac{w\epsilon_{\sigma^2,1}}{1-w\epsilon_{\sigma^2,1}}\right)$$

$$\text{Therefore, } \lambda \geq \frac{\frac{w\epsilon_{f,1}}{1-w\epsilon_{f,1}}}{1-\left(1+\frac{w\epsilon_{\sigma^2,1}}{1-w\epsilon_{\sigma^2,1}}\right)\sqrt{n_t/n_w}}. \tag{4.28}$$

□

Proof of Lemma 4.6.

$$\frac{1}{n_w}\sum_{i=-w}^{w} n_{t+i}g_{t+i} \leq \frac{1}{n_w}\sum_{i=-w}^{w}(\mu_n+n_t-\mu_n)g_{t+i}$$

$$\leq \frac{1}{\mu_n(2w+1)}\sum_{i=-w}^{w}\mu_n g_{t+i} + \frac{1}{\mu_n(2w+1)}\sum_{i=-w}^{w}(n_{t+i}-\mu_n)g_{t+i}$$

$$\leq g_t(1+\epsilon_{g,2}) + \frac{1}{\mu_n(2w+1)}\sum_{i=-w}^{w}(n_{t+i}-\mu_n)g_{t+i}$$

[By Lemma 4.4]

$$\leq g_t(1+E_g(w)) + \frac{1}{\mu_n(2w+1)}\sum_{i=-w}^{w}(n_{t+i}-\mu_n)(g_{t+i}-g_t+g_t)$$

$$\leq g_t(1+E_g(w)) + \frac{1}{\mu_n(2w+1)}\sum_{i=-w}^{w}(n_{t+i}-\mu_n)(g_{t+i}-g_t)$$

[Since $g_t\sum_{i=-w}^{w}(n_{t+i}-\mu_n)=0$]

$$\leq g_t(1+E_g(w)) + \sqrt{\sum_{i=-w}^{w}\frac{1}{\mu_n^2}\frac{(n_{t+i}-\mu_n)^2}{2w+1}\sum_{i=-w}^{w}\frac{(g_{t+i}-g_t)^2}{2w+1}}$$

[By Cauchy-Schwarz Inequality]

$$\leq g_t(1+E_g(w)) \tag{4.29}$$

$$+ \sqrt{\sum_{i=-w}^{w}\frac{1}{\mu_n^2}\frac{(n_{t+i}-\mu_n)^2}{2w+1}}\max_{i=-w}^{w}(g_{t+i}-g_i)$$

$$\leq g_t(1+E_g(w)) + \frac{\sigma_n}{\mu_n}g_t\frac{w\epsilon_{g,1}}{1-w\epsilon_{g,1}} \qquad \text{[By Lemma 4.3]}$$

$$\leq g_t\left(1+E_g(w)+\epsilon_{g,1}\frac{\sigma_n}{\mu_n}\frac{i\epsilon_{g,1}}{1-i\epsilon_{g,1}}\right). \tag{4.30}$$

Setting $\gamma_g = E_g(w) + \epsilon_{g,1}\frac{\sigma_n}{\mu_n}\frac{w\epsilon_{g,1}}{1-w\epsilon_{g,1}}$,

$$\frac{1}{n_w}\sum_{i=-w}^{w} n_{t+i}g_{t+i} - g_t \leq \gamma_g g_t \tag{4.31}$$

Similarly, we can show

$$\frac{1}{n_w}\sum_{i=-w}^{w} n_{t+i}g_{t+i} - g_t \geq \gamma_g g_t \tag{4.32}$$

□

Proof of Theorem 4.4. Using Lemma 4.6 in Equation 4.24 and 4.27, we get

$$\mathbb{E}\left(\left|\bar{X}_{t,w}/\mu_D - f_t\right|\right) \leq \gamma_f f_t\,, \text{ and } \left|\sigma^2_{\bar{X}_{t,w}} - \sigma^2_{X_t}\right| \leq \gamma_{\sigma^2}\sigma^2_{X_t}\,. \tag{4.33}$$

Putting these values in Lemma 4.5, we must have

$$\lambda \geq \frac{\gamma_f}{1 - \sqrt{\frac{n_t}{n_w}(1+\gamma_{\sigma^2})}}$$

□

4.S.A.1. Extension to Biased Reporting

In this paper, we have assumed that the response X_t is honest and correct. However, in reality, the reporting could be biased to exaggerate counts or undercount, or the respondents may incorrectly recall. Suppose the bias follows the following pattern.

Assumption 4.4. *The respondents can be grouped, where an individual belongs to the group j with probability q_j (independently of its in-degree) and overestimates the hidden population by a factor of α_j (underestimates if $\alpha_j < 1$). Also, q_j and α_j are constants over time.*

Suppose Z_t is a randomly selected response with possible bias following Assumption 4.4 and $\bar{Z}_t$ is the sample mean. Then

Theorem 4.5. $\mathbb{E}(Z_t) = B\mu_D f_t$, $\sigma^2_{Z_t} = B^2\sigma^2_{X_t}$, *and for any* $\lambda \leq 0$

$$P\left(\left|\frac{\bar{Z}_t}{B\mu_D} - f_t\right| > \lambda\right) = P\left(\left|\frac{\bar{X}_t}{\mu_D} - f_t\right| > \lambda\right)\,, \tag{4.34}$$

where $B = \sum_j q_j\alpha_j$.

Proof. The expectation of response Z_t from a randomly selected person is given by

$$\begin{aligned}
\mathbb{E}(Z_t) &= \sum_k \mathbb{P}(\delta(u)=k)\sum_j q_j\alpha_j \mathbb{E}\left(\sum_{v\in in(u)} I_v|\delta(u)=k\right)\\
&= \sum_k \mathbb{P}(\delta(u)=k)\sum_j q_j\alpha_j \sum_{i=1}^{k}\mathbb{E}(I_{v_{ik}}|\delta(u)=k)\\
&= \sum_k p_k \sum_j q_j\alpha_j \sum_{i=1}^{k}\mathbb{E}(I_{v_{ik}})\\
&= (\sum_j q_j\alpha_j)\sum_k k p_k f_t = f_t B\mu_D
\end{aligned} \tag{4.35}$$

where $B = \sum_j q_j\alpha_j$. This suggests that consistent biases among the groups scale the time series of expected responses by a constant Similarly, we can show that $\sigma^2_{Z_t} = B^2\sigma^2_{X_t}$. Considering the probability of a certain deviation obtained by scaling of the sample mean $\bar{Z}_t$:

$$
\begin{aligned}
P\left(\left|\frac{\bar{Z}_t}{B\mu_D} - f_t\right| > \lambda\right) &= P\left(\left|\frac{\bar{X}_t - \mu_D B f_t}{\sigma_{Z_t}/\sqrt{n}}\right| > \frac{\mu_D\sqrt{n}\lambda}{\sigma_{X_t}}\right) \\
&= 1 - erf\left(\frac{\mu_D B\sqrt{n}\lambda}{B\sigma_{X_t}}\right) \\
&= 1 - erf\left(\frac{\mu_D\sqrt{n}\lambda}{\sigma_{X_t}}\right) \qquad (4.36)
\end{aligned}
$$

which is identical to the expression in Equation 4.14 where bias is absent. □

Therefore, all of our results still apply after normalizing the time-series of $\bar{Z}_t$ and $\bar{Z}_{t,w}$ even if there are biases in the responses under Assumption 4.4.

The above analysis does not take into account respondents who are malicious, deliberately inaccurate (possibly due to the sensitivity of the questions), or exhibit other behaviors that deviate from Assumption 4.4. In practice, additional steps can be taken to handle these different biases (Scheers, 1992; Kazemzadeh et al., 2016; Ezoe et al., 2012; Salganik et al., 2011).

4.S.B. Additional Results

Testing Observation 4.1 In our analysis, we indicated that the average degree of the latent graph remains roughly constant. The CMU-CTIS recorded indirect survey information on COVID-like illness (CLI) incidence both in the household and in the local community. Figure 4.6 shows the average household size obtained across time from four states (California, Texas, New York, and Pennsylvania). The average remains roughly constant, supporting Observation 4.1.

Undersampling Real Data To demonstrate that our approach works with a small number of responses as well, we performed experiments where we undersampled the US CMU-CTIS data. Figure 4.7 displays the averaged COVID-19 incidence curves and the corresponding 95% confidence interval from direct and indirect surveys using 10% of the daily samples for the four states and for $accum = 7$ and $w = 2$. Specifically, the curves are the ensemble average of 25 realizations of the experiment, where a random subset of 10% of the daily samples was selected at each realization. As can be seen in this figure, the curves from direct and indirect surveys closely follow the trend of the curve obtained from official data. Fig. 4.9 illustrates the boxplots of errors obtained by the various methods using 10% of the daily samples for the four states and for $accum = 7$ and $w = 2$. As can be seen in these boxplots, the indirect methods yield the lower median errors compared to those obtained by NSUM and direct techniques. Figures 4.8 and 4.13 show, respectively, the normalized incidence curves and the distribution of errors using 3% of the daily samples for the four states and for $accum = 7$ and $w = 2$.

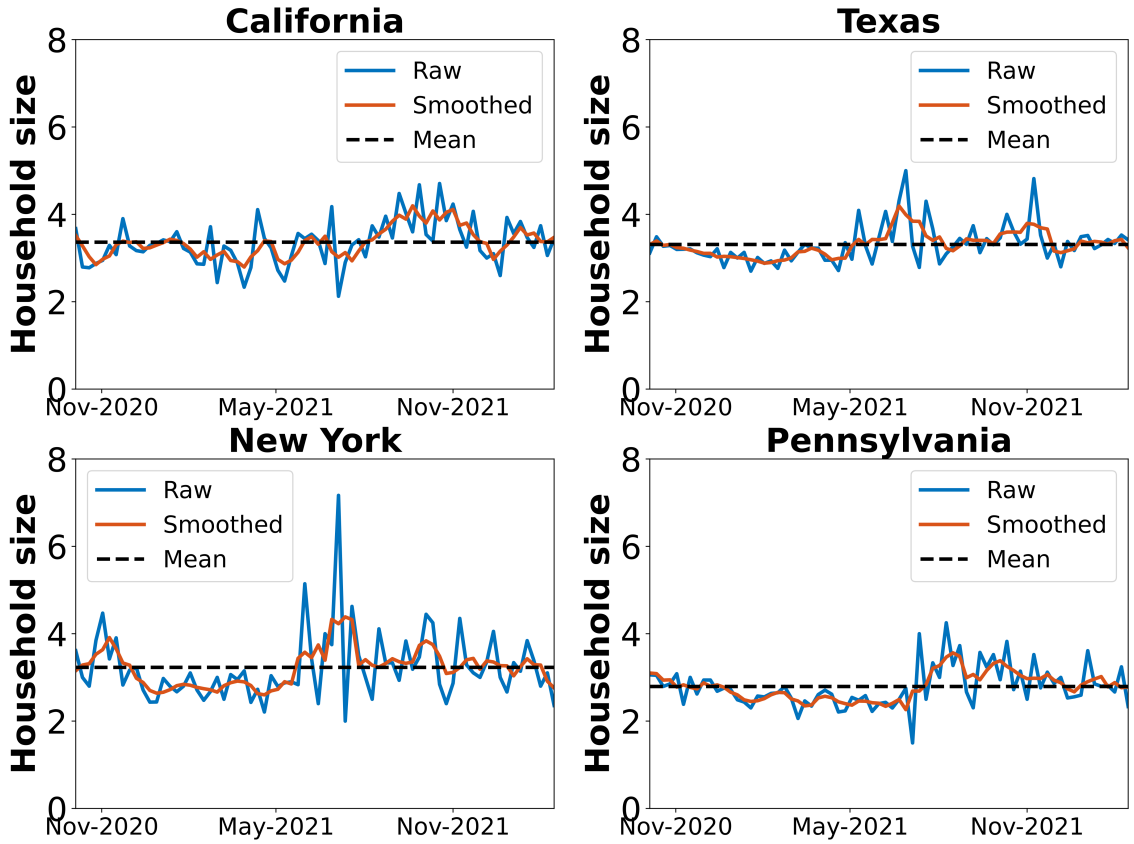


Figure 4.6: Household size obtained from CMU-CTIS data for California, Texas, New York, and Pennsylvania between September 2020 and March 2022.

More Simulation Results Here we present the boxplots omitted from the main paper showing the impact of d and n_d. These parameters had a negligible impact.

4.S.B.1. Results on UMD surveys

We also conducted experiments on UMD surveys [4] that collected survey data at the country level for various countries. In this dataset, there are responses regarding if one has tested positive recently (DirectTPR), if they have COVID-like symptoms (DirectCLI), and the "indirect" question – how many they know who have COVID-like symptoms. The last question allows us to perform estimation using our approach (Indirect). We used $w = 1$, i.e., a window of 3 to aggregate the responses. There is no information on the personal network size, which is needed for NSUM. Therefore, we used an existing technique[5] that estimates the personal network size for comparison. The ground truth was obtained from the reported case count available from Our World in Data[6]. The results are presented in Table 4.2. We note that Indirect

[4] https://covidmap.umd.edu/
[5] https://arxiv.org/abs/2108.03284
[6] https://ourworldindata.org/covid-cases

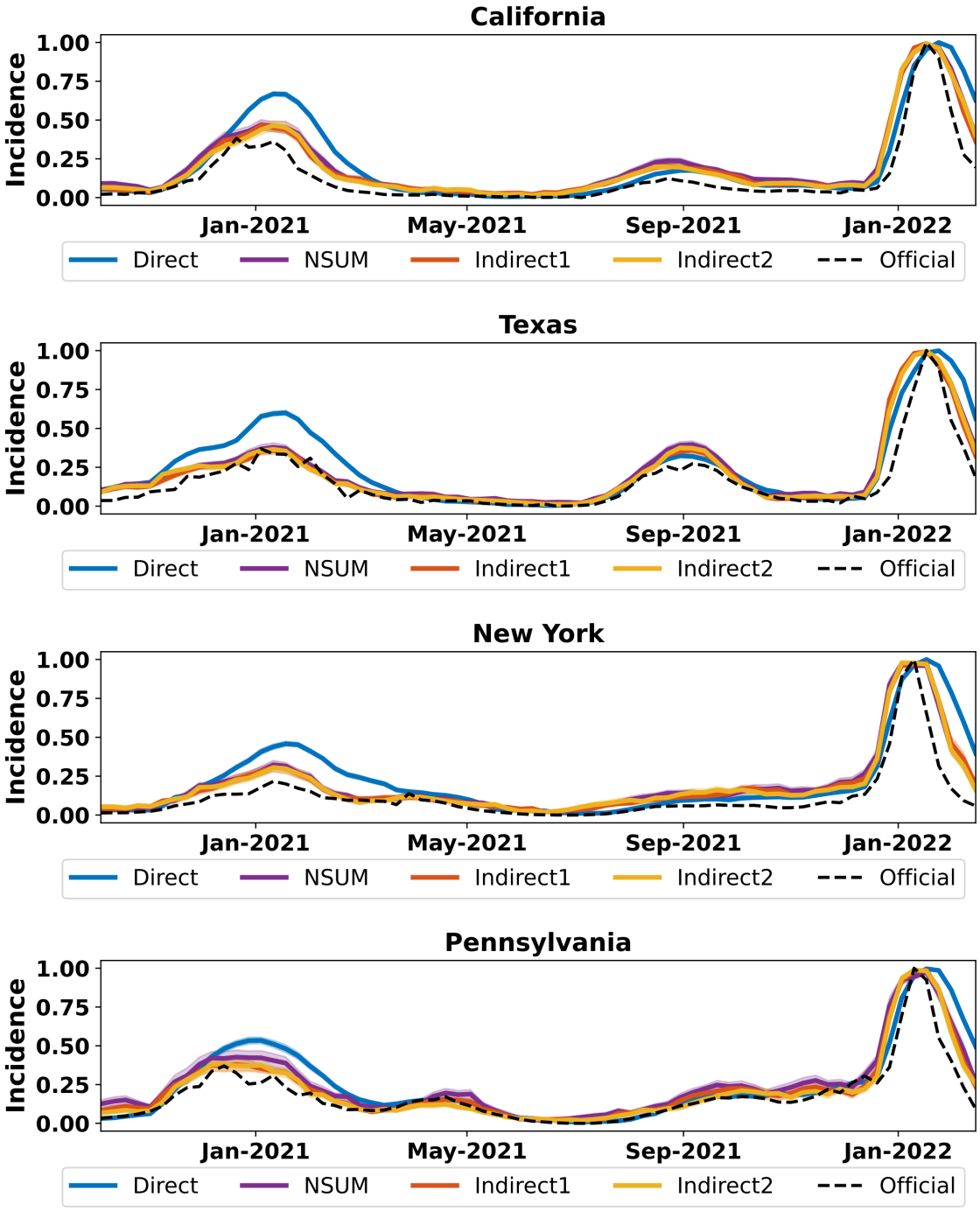


Figure 4.7: Normalized COVID-19 incidence curves and the 95% confidence interval using 10% of the daily samples for the four states and for $accum = 7$ and $w = 2$.

outperforms DirectCLI and NSUM on all countries with a significant margin in most. In fact, Indirect (regarding CLI) provides a better estimate than directly asking about positive tests in most countries. Figure 4.12 shows the comparison of estimated incidences for the United Kingdom. Our approach (IndirectCLI) most closely resembles the ground truth. Our estimates are not only better than DirectCLI but also better than asking directly about positive tests (as opposed to indirect on

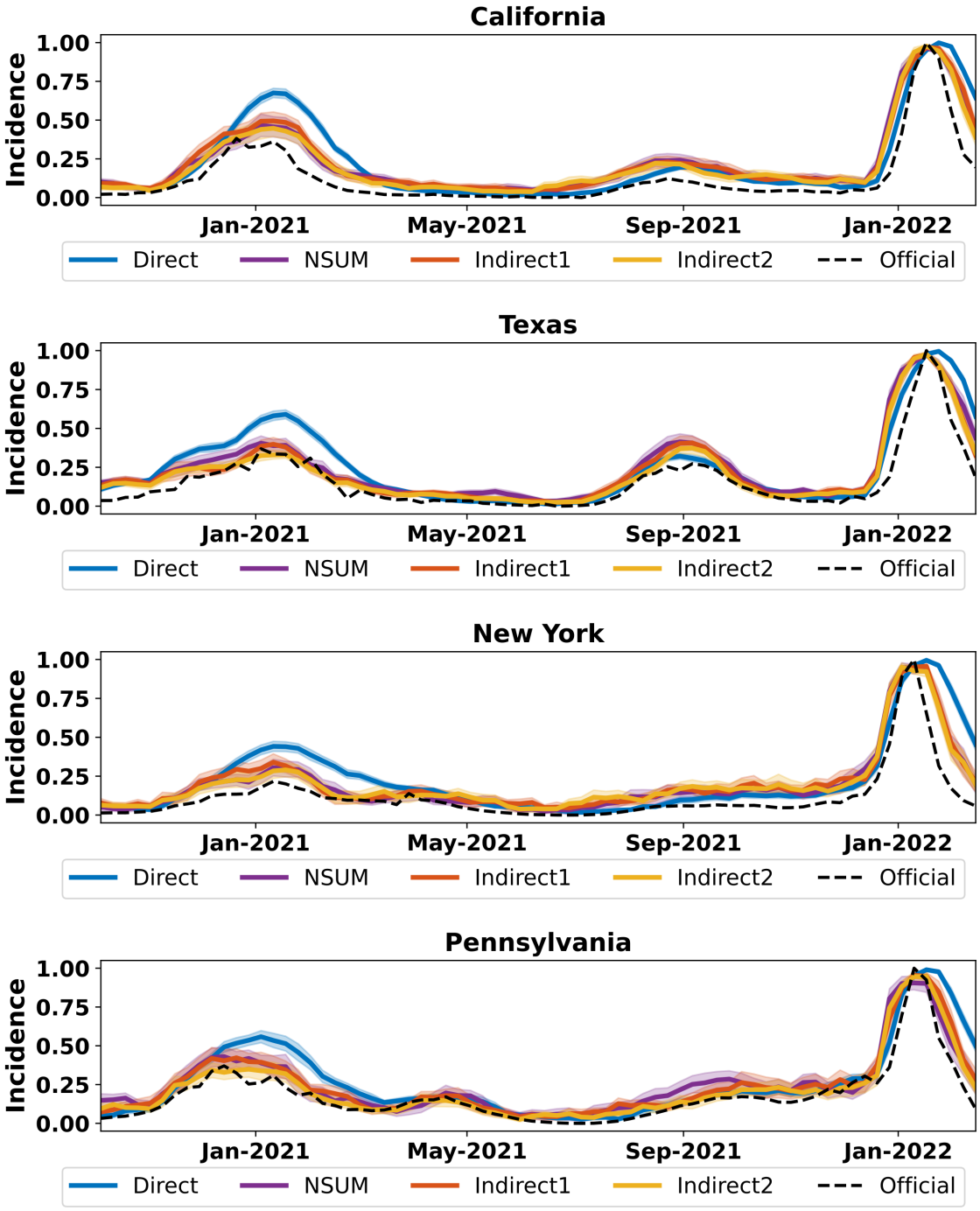


Figure 4.8: Normalized COVID-19 incidence curves and the 95% confidence interval using 3% of the daily samples for the four states and for $accum = 7$ and $w = 2$.

CLI).

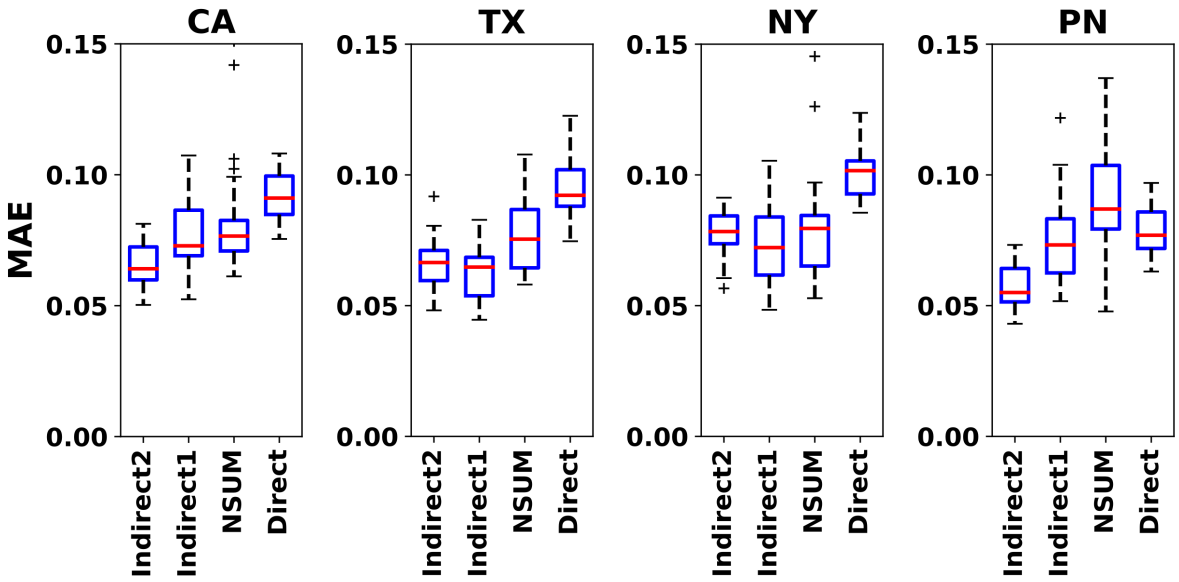


Figure 4.9: MAE of the normalized COVID-19 incidence curves using 10% of the daily samples for the four states and for $accum = 7$ and $w = 2$.

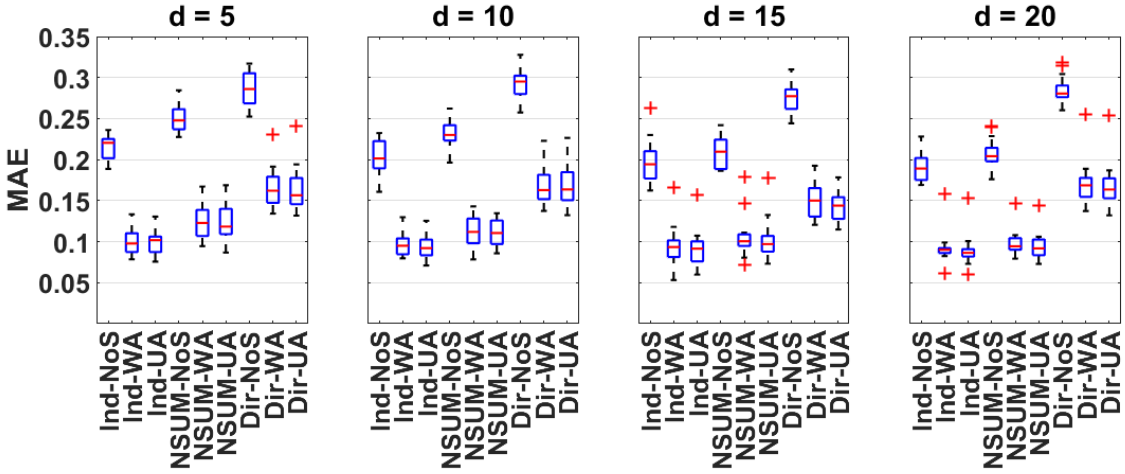


Figure 4.10: MAE vs d, approximate average degree of the respondents. Choice of d does not significantly affect the relative performance of the methods

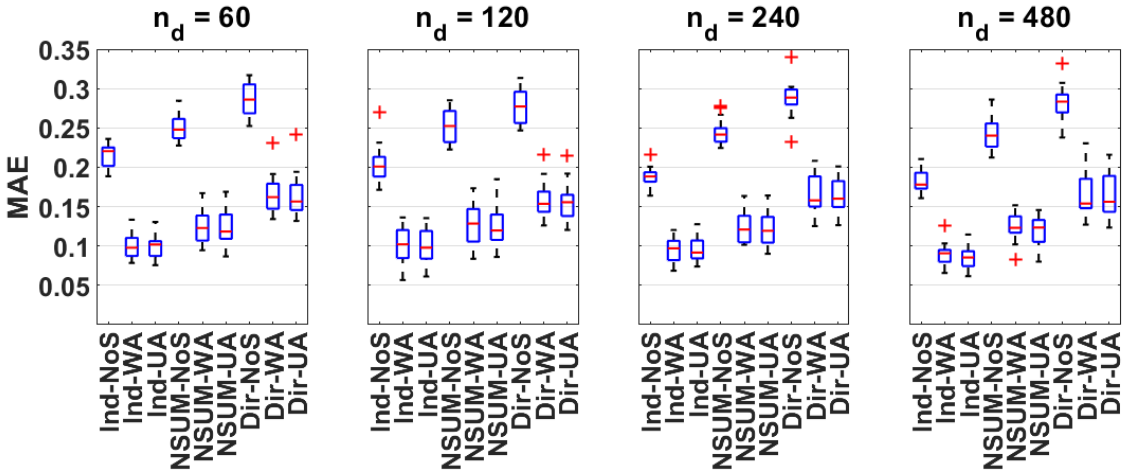


Figure 4.11: MAE vs n_d, the subpopulation potentially covered by the sampled respondents. Choice of n_d does not significantly affect the relative performance of the methods

Country	DirectTPR	NSUM	DirectCLI	Indirect
Brazil	**0.14**	0.17	0.15	0.15
Canada	0.14	0.19	0.4	**0.12**
France	0.17	0.24	0.22	**0.13**
Germany	**0.1**	0.19	0.29	0.12
Israel	0.32	0.19	0.17	**0.12**
Italy	0.17	0.13	0.05	**0.1**
Japan	0.15	0.23	0.25	**0.08**
Spain	0.31	0.2	0.16	**0.1**
Turkey	0.24	0.29	0.16	**0.12**
UK	0.24	0.26	0.23	**0.11**

Table 4.2: MAE on country-level comparison from UMD-CTIS surveys.

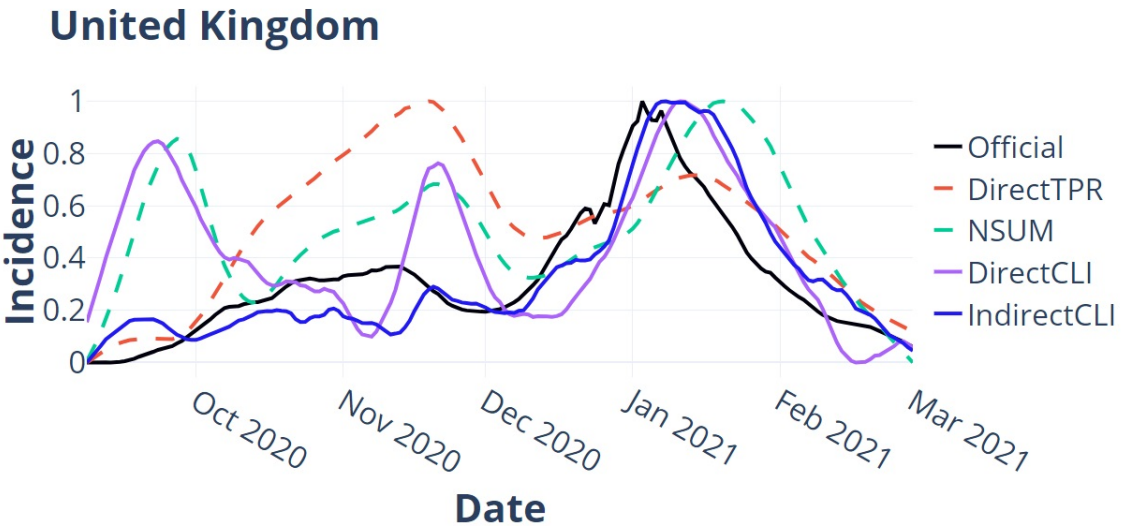


Figure 4.12: Comparison of estimated incidences from UMD-CTIS surveys. Our approach (IndirectCLI) provides a better estimate than DirectCLI, and even better than asking directly about positive tests (as opposed to indirect on CLI).

4.S.C. Additional Methodology Details

4.S.C.1. Code

All our code and aggregate datasets are publicly available[7]. The synthetic experiments were run on a machine with Intel(R) Core(TM) i8 CPU, 3GHz, 6 cores, and 16GB RAM. The code was written in MATLAB 2021b. The experiments on real-world data were conducted on a machine with Intel(R) Xeon(R) CPU, 2.60GHz, 16 cores, and 32GB RAM. The pre-processing code to generate aggregate data from the CTIS survey was written in R. The results demonstrated in the paper were generated using Python.

4.S.C.2. Outlier Filtering

In order to remove inconsistent and potentially malicious responses to the CTIS survey, we apply a simple outlier filter based on the following rules:

- We remove responses that report negative values in numeric questions that can only have non-negative responses.
- We remove responses that report more than 100 in the number of people in the household with COVID-like illness (CLI), in the number of people in the household in any age range (< 18, 18-64, and ≥ 65), or in the number of days with symptoms.
- We remove inconsistent responses that report people in the household with CLI but no symptoms, or report more people with CLI than the household size.

4.S.C.3. Synthetic Experiments

Here we present more details on the simulations.

Epidemic simulation We use a simple SIR model to simulate an epidemic with varying infection parameters whose reproduction number R0 (Dietz, 1993) is above 2. At random points in time, we introduce "interventions" that reduce R0 smoothly to a value below 1. We run several such simulations and pick one that produces multiple peaks over 600 days, to emulate realistic epidemics like Influenza and COVID-19 that have multiple waves. We acknowledge that, in reality, not all infections will be detectable. However, assuming that each infection will be detected with a fixed probability only scales the time-series $I(t)$ (the number of infectious individuals by the SIR process) by a constant. Therefore, for the purpose of this study, we directly use $I(t)$ to compute the hidden population over time.

[7]Our code is available at https://github.com/GCGImdea/coronasurveys/tree/master/papers/2024-AAAI-Nowcasting-Temporal-Trends-Using-Indirect-Surveys.

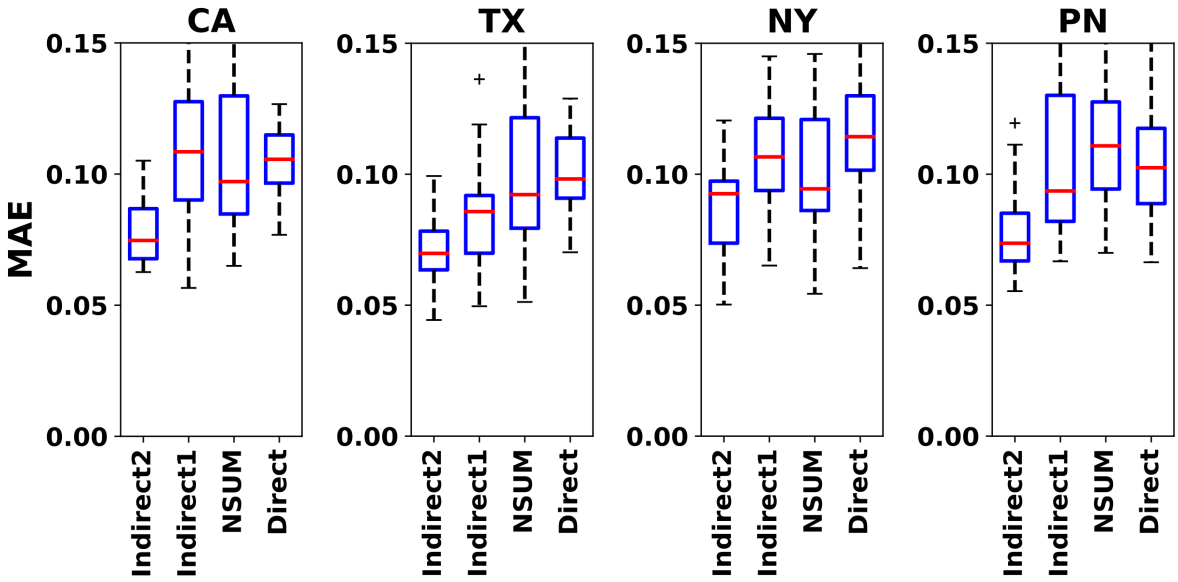


Figure 4.13: MAE of the normalized COVID-19 incidence curves using 3% of the daily samples for the four states and for $accum = 7$ and $w = 2$.

Survey simulation. We simulate, at each time t, random sets L and R of sizes n and n_d, and a bipartite digraph between them with terminal set L, which represents the surveyed individuals. To construct the digraph, we pick the degree of vertex $u \in L$, δ_u from a power law distribution $p_k \propto k^{-2}, k \in [1, n_d/2]$, such that the mean degree is approximately d. The choice of the probability distribution was driven by the intention to introduce some skewness in the degree distribution. Then, we add an edge between u and $v \in R$ with a probability δ_u/n_d.

For the indirect surveys simulation, we randomly infect each node in R with the probability $\sum_{\tau=0}^{period-1} I(t-\tau)$, where period is the time window within which the respondents are to count the hidden population. For comparison, we also introduced the traditional NSUM approach (Killworth et al., 1998a), where the response from each node is normalized by its degree. For direct surveys, we infect nodes in L itself and note the number of infections produced in L.

4.S.D. Algorithm in Practice

Suppose we have some bounds on $\epsilon_{f,1}$ and $\epsilon_{f,2}$. This could be obtained from prior experience. For example, if we are estimating the number of people who test positive for COVID-19, we can use the prior trend of COVID-19 when the reporting quality was better (before home tests). We can use these along with Theorems 4.3 and 4.4 to identify a window that is likely to improve our estimation as given in Algorithm 1. Suppose we are given a threshold λ acceptable fractional deviation, e.g., $\lambda = 0.1$. Then we can search for a w that either satisfies Inequality 4.2 or Inequality 4.3.

Algorithm 1 Indirect Survey Estimation

procedure AGGREGATEDESTIMATE($\{\bar{X}_t\}, \{n_t\}, \lambda$)
  Pick some w
  while $w > 0$ **do**
    Compute σ_n and μ_n for the window w
    $\lambda_1 \leftarrow$ RHS of Inequation 4.2
    $\lambda_2 \leftarrow$ RHS of Inequation 4.3
    if $\min\{\lambda_1, \lambda_2\} \leq \lambda$ **then**
      Break
    $w \leftarrow w - 1$
  return $\bar{X}_{t,w} \leftarrow \frac{\sum_{i=-w}^{w} n_{t+i}\bar{X}_{t+i}}{\sum_{i=-w}^{w} n_{t+i}}$
end procedure

If Inequality 4.2 is satisfied, then by Theorem 4.3, aggregating with window w must be better than no smoothing. If Inequality 4.3, then the same applies due to Theorem 4.4. We can apply this aggregation over discrete bins (daily, weekly, monthly) or in a rolling fashion (smoothing), or some combination of the two.

Chapter 5

Robust Network Scale-up Method Estimators

In *Social Networks*, 84, 46-61, 2026.
https://doi.org/10.1016/j.socnet.2025.08.002

Sergio Díaz-Aranda[1,2] Juan Marcos Ramírez[1] José Aguilar[1,3] Rosa E. Lillo[2,4]
Antonio Fernández Anta[5,1]

[1] IMDEA Networks Institute, Madrid, Spain
[2] uc3m-Santander Big Data Institute, Madrid, Spain
[3] Universidad de Los Andes, Mérida, Venezuela
[4] Department of Statistics, Universidad Carlos III de Madrid, Madrid, Spain
[5] IMDEA Software Institute, Madrid, Spain

Abstract

The Network Scale-up Method (NSUM) is a relatively recent statistical approach for estimating the prevalence of unknown populations through indirect surveys utilizing information about the respondents' social circles. The popularity of NSUM has increased in recent years due to its ability to uphold privacy and cost-effectiveness. However, the NSUM is not exempt from biases resulting from participants' behavior. In addition, the simpler and most popular NSUM estimators are based on averages, making them sensitive to deviations in the samples, which may cause significant errors. This work aims to study how robust procedures can overcome misreporting, contamination, and deviation due to conditions such as barrier effects, prevalence, skewness, and tail length. Specifically, the central objective of the article is to analyze the statistical robustness of NSUM methods, studying whether these methods are affected by outliers or unusual data. We employ eight robust proposals for each of the two classical NSUM estimators. We analyze robust estimators through simulation experiments using synthetic random networks such as Erdős–Rényi, Scale Free, and Stochastic Block Model structures to model different degree distributions and community structures with different prevalence levels in contaminated and uncontaminated scenarios. We compare the results of the simulations with real data on COVID-19 indicators in the United Kingdom and voting intention in the Spanish General Elections of 2023. This article shows that the classical NSUM estimators perform poorly in contaminated scenarios, while most of

the robust proposals are not considerably affected. However, the performance of some robust NSUM estimators decreases under barrier effects. In addition, we observe that distortions created by small prevalence play an important role in selecting the most suitable robust NSUM estimator. Particularly, the robustification of the Mean of Ratios (MoR) estimator based on the Myriad operator typically exhibits the best performance (for MoR methods) across the various social network structures for different prevalence levels, reducing the estimation error regarding the non-robust methods by up to three orders of magnitude in contaminated scenarios.

Keywords: Network Scale-up Method, Aggregated relational data, Robust estimators, Adaptive procedures, M-estimators

5.1. Introduction

The *Network Scale-up Methods* (NSUM) are methods to estimate population prevalences based on information collected from individual social networks (Laga et al., 2021). Specifically, NSUM uses the number of connections that a respondent reports within specific groups during a survey, known as *aggregated relational data* (ARD) (McCormick, 2020), to estimate the prevalence or size of these groups. Since NSUM do not require personal information from participants, they have been widely applied to estimate the size of sensitive groups, where disclosing the membership to a group may pose a potential threat, such as sex workers (Jing et al., 2018), drug users (Laga et al., 2023b), child traffickers (Nyarko-Agyei et al., 2024), people who had abortions (Sully et al., 2020), people with COVID-19 (Garcia-Agundez et al., 2021), or people with HIV (Laga et al., 2023a).

During recent years, there have been many improvements in the classical NSUM (Killworth et al., 1998a,b), as noted in the review of Laga et al. (2021). However, no method solves the lack of robustness of classical methods, which use sample means and may be drastically affected by the inclusion of outliers (Hampel et al., 2011). Some studies, such as Garcia-Agundez et al. (2021) and Ramirez et al. (2023), use data preprocessing to mitigate this issue. However, Garcia-Agundez et al. (2021) employ non-robust filtering techniques, while the filtering based on Medcouple used by Ramirez et al. (2023) is overly aggressive, removing up to half of the data. In addition, estimators addressing robustness concerns have only been proposed in two articles. However, these works only focus on unbiased estimates under specific biases that may affect the estimators. Feehan and Salganik (2016) try to improve the NSUM estimates when the respondents are unaware of their acquaintances' membership, and the propensities of forming ties differ depending on the participants. However, this method is also based on sample means and requires significantly more information–specifically, access to a sample from the unknown group. Maltiel et al.

(2015) also consider these biases, but Díaz-Aranda et al. (2024) show that it is very sensitive to deviations. Beyond biases, survey data may have issues due to outlying responses, frequently due to malicious participants.

The main contribution of this work is the proposal of several robust NSUM estimators, eight estimators for each classical method, to mitigate these problems. Each of these estimators has been evaluated for its robustness to different types of data deviations, including contamination and distributions with skew and tail length. To our knowledge, this is the first work that proposes robust NSUM procedures within the framework of robust statistics.

A second contribution is the empirical evaluation of the robustness of the estimators in a simulated environment. The simulated social networks are similar to those found in real populations. In this case, the objective is to study how NSUM techniques resist outliers or unusual data, as defined in robust statistics–an aspect that has not been studied in the literature for NSUM techniques. The aim is to study the quality of the results when the data are imperfect or unusual. To this end, we have used the notion of contamination, which allows us to analyze their behavior for various types of deviating data.

Additionally, we validate the performance in real datasets on COVID-19 and voting intention, representing cases of outlying values due to malicious responses and different propensities of forming links, respectively. In addition, we also assess the performance of the NSUM under the classic bias of the NSUM literature: barrier effect, response bias, recall error, and transmission error, as some are equivalent to a few of the cases studied in our contamination (see the Supplementary Material). In particular, to the best of our knowledge, this is the first study analyzing the effect of the response bias and proposing alternatives to mitigate it.

This article is organized as follows. Section 2 presents the NSUM estimators and their robustness properties. In Section 3, we present the robust proposal for each estimator. Section 4 displays the simulations conducted and a discussion about their performance. Section 5 assesses the results on real datasets. The final remarks are in Section 6.

5.2. Robustness and NSUM

The NSUM use the personal network size to scale the ARD of the social circles of the participants. Traditionally, NSUM utilize groups with known prevalences to first estimate the personal network size of the participants, usually denoted as the *participants' degree.* In our approach, we assume that participants directly report their degrees and that these reports are accurate. In addition, we include another ARD question about a group whose prevalence is unknown. Let d_i represent the reported degree of participant i and y_i the ARD for the unknown group, with the

prevalence of this group denoted by ρ. Additionally, we denote by I_n the set of responses obtained by a survey of size n.

The classical NSUM methods are the frequentist estimators, firstly introduced in Killworth et al. (1998a,b); Habecker et al. (2015) (see Laga et al. (2021) for a review of the NSUM estimators). The main idea of the NSUM estimators is that the ratios y_i/d_i possess representative information about the prevalence of the estimators. Under our assumptions, there are two classic estimators:

- The *Mean of Ratios* (MoR) estimator is the average of the individual ratios of the participants, defined as
$$\text{MoR}(I_n) = \frac{1}{n}\sum_{i=1}^{n}\frac{y_i}{d_i}. \tag{5.1}$$
- The *Ratio of Sums* (RoS) estimator is the ratio of the sum of the number of reported neighbors in the unknown group over the sum of degrees of the participants, defined as
$$\text{RoS}(I_n) = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} d_i}. \tag{5.2}$$

Killworth et al. (1998a) proposed the MoR assuming that the ratios are representative on average. Subsequently, Killworth et al. (1998b) proposed the RoS, the most widely used NSUM estimator in the literature (see, for example, Ocagli et al. (2021)), complementing the previous idea with a Binomial model for the ARD responses for the unknown group conditioned on the degree

$$y_i \sim Bin(d_i, \rho). \tag{5.3}$$

Killworth et al. (1998b) demonstrated that the maximum likelihood estimator corresponds to the ratio of the sums described in Eq. 5.2.

The NSUM are prone to biases due to the response behavior (Laga et al., 2021). *Response bias* corresponds to malicious responses. *Transmission error* stems from the participants' lack of knowledge of whether their acquaintances belong to the specified groups. *Recall error* arises from inaccuracies in recalling the number of contacts within these groups. Finally, another potential source of bias is the *barrier effect*. This bias arises from differences in the propensity to know someone based on the participant's characteristics, which contradicts the assumption that the ratios are representative.

5.2.1. The NSUM in the Robust Statistics framework

Robust Statistics is the field of Statistics that examines how the deviations from the main assumptions and the addition of outlying values affect the estimates. A fundamental concept in robust statistics is the *contamination*. Contamination refers to the alteration of a distribution. Concretely, the contamination of a distribution F with a distribution H is defined by the mixture distribution of F with probability

$1-\epsilon$ and H with probability ϵ, denoted as $(1-\epsilon)F + \epsilon H$. A common example is the contamination with a single value, where the variable takes a fixed value x with some probability. This is modeled mathematically as $F_{x,\epsilon} = (1-\epsilon)F + \epsilon\delta_x$, where δ_x is the point-mass distribution with support in x, assigning probability 1 to x and 0 elsewhere.

The most common estimator is the sample mean. However, the sample mean is not a robust estimator since the breakdown point is zero and the influence function is not bounded (Hampel et al., 2011). The *breakdown point* is the minimum fraction of contamination of a bad observation that can produce an arbitrary value (see the Supplementary Material). The *influence function* assesses the impact of a single outlier, concretely, the variation of the estimates when the fraction of contamination tends to zero (see Section 5.S.A of the Supplementary Material for a formal definition).

As we observe in Eqs. 5.1 and 5.2, the NSUM estimators are based on the sample mean and therefore are not robust. Additionally, the NSUM estimators have more problems because of the nature of the ARD. For instance, the variables y_i, d_i, and y_i/d_i are generally not normal. This is especially true for the ARD of the unknown group, y_i, when the group size is small, as y_i is a discrete variable and most values tend to be zero. The distribution of the y_i has been observed to exhibit heavy tails and skewness in real data (Killworth et al., 1998a) (Zheng et al., 2006). Similarly, real networks are often modeled by degree distributions that deviate from normality, such as power law distributions (Newman, 2018). Heavy-tailed distributions may lead to inflated standard errors and poor convergence of the sample mean to a Normal distribution, thereby negatively impacting inferences (Wilcox, 2011). Skewness is another source of inaccurate inferences and can be difficult to distinguish from asymmetric contamination (Wilcox, 2011). Moreover, the assumption that the mean is a reliable measure for the typical participant under study in skewed distributions is questionable (Wilcox, 2011).

The biases of the NSUM can be modeled within this framework for some cases. The response biases caused by malicious responses at the extreme values can be modeled by contamination at the extreme of the distributions, which is a common approach in robust statistics. Contamination naturally applies to the transmission error when a fraction of the respondents do not know the membership of their acquaintances, and therefore follow a different distribution. Additionally, if the unknown group is small, as in many NSUM studies, the transmission error is approximately the contamination with a point-mass distribution with support at 0. The recall error is difficult to model and, to the best of our knowledge, there is only one proposal in Díaz-Aranda et al. (2024). Some instances of undercounting and overcounting can be explained by contamination caused by distributions shifted to the sides. Finally, while the barrier effect is less natural in this framework, it can still be modeled as contamination if the participants having different propensities to know members of the unknown group are limited to a small number of individuals.

5.3. Robust NSUM estimators

In this section, we propose several robust alternatives for each NSUM estimator. A robust procedure should achieve a good performance in the assumed conditions, stability under small deviations, and avoid catastrophic outcomes under high deviations (Huber and Ronchetti, 2009). Among the various approaches and procedures in robust statistics (see Staudte and Sheather (2011)), we focus on estimators based on descriptive measures of distributions that are not necessarily parametric. A descriptive measure is a parameter associated with each distribution. Below, we present the key concepts for our work.

Definition 5.1. *A descriptive measure is a map $T : \mathcal{M} \to \mathbb{R}$, where $\mathcal{M}$ is the space of all probability distributions on the sample space (Staudte and Sheather, 2011).*

An example is the population mean, which can be expressed as $T(F) = E[X]$, where X is a random variable with distribution F. There is a natural estimator of the descriptive measure $T(F)$, $T(x_1, \cdots, x_n)$, defined as $T(F_n)$, where $x_1, \cdots, x_n$ are samples of the distribution F and $F_n = \frac{1}{n}\sum_i \delta_{x_i}$ is the empirical measure. We are particularly interested in a type of descriptive measure: *location measures.*

Definition 5.2. *Location measures are descriptive measures that satisfy the following properties for a random variable X with distribution F_X (Wilcox, 2011):*

(1) $T(F_{X+b}) = T(F_X) + b$.
(2) $T(F_{-X}) = -T(F_X)$.
(3) if $X \geq 0$, *then* $T(F_X) \geq 0$.
(4) $T(F_{aX}) = aT(F_X)$ *for* $a > 0$.

An estimator of a location measure is referred to as a location estimator.

This section is organized as follows. First, we define and present the location measures that we will employ throughout the study. Next, we propose robustifications of MoR and RoS based on these measures.

5.3.1. Location estimators

We present the estimators based on location measures that we propose to robustify the NSUM estimators. They are based on the idea of giving less weight to extreme observations. Although the robustness of the sample mean has been widely studied, there is no extensive study regarding improvements when the distribution has skewness, as seen in Wilcox (2011), since asymmetric contamination may be confused with contamination (Huber and Ronchetti, 2009). Additionally, the frequent approach of the estimator in robust statistics assumes continuous distributions (see Huber and Ronchetti (2009)). In our case, we propose estimators derived from the adaptive

estimation literature and foundational techniques in robust statistics. Concretely, we consider four types of estimators: the median, the trimmed means, the adaptive trimmed means, and the M-estimators.

Median. The median is an estimator that is very resistant to big deviations since its breakdown point is 0.5 (Maronna et al., 2019). Additionally, the median has good stability with bounded influence function (Huber and Ronchetti, 2009). The median is often selected as a location measure for skewed distributions, where the mean does not represent the typical participant under study (Wilcox, 2011). However, the median can have high bias and low efficiency in discrete distributions, due to the limited number of possible outcomes.

Trimmed Mean (TM). The trimmed mean is a type of L-estimator, an estimator that is a linear combination of order statistics $\sum_i a_i X_{(i)}$ (Huber and Ronchetti, 2009), where $X_{(i)}$ is the ith order statistic. The TM is a particular case in which the order statistics of the extremes are given a zero value. The *α-trimmed mean* (TM_α) is the estimator defined as

$$TM_\alpha(x_1, \cdots, x_n) = \frac{1}{n-2g} \sum_{i=g+1}^{n-g} x_{(i)}, \tag{5.4}$$

where $g = \lfloor n\alpha \rfloor$ is the largest integer not exceeding $n\alpha$. TM_α has breakdown point of α and a bounded influence function (Huber and Ronchetti, 2009). More information can be found in Section 5.S.A of the Supplementary Material.

Adaptive trimmed means. The TM estimator assigns a zero weight to extreme observations from both sides, which works well for symmetric distributions. However, in the presence of skewness, it is unclear whether the TM is an appropriate location measure. The asymmetric trimming arises as a solution for location measures in skewed data (Hogg, 1974). The (α_1, α_2)-trimmed mean is the estimator

$$TM_{(\alpha_1,\alpha_2)}(x_1, \cdots, x_n) = \frac{1}{n-g_1-g_2} \sum_{i=g_1+1}^{n-g_2} x_{(i)}. \tag{5.5}$$

where $g_1 = \lfloor n\alpha_1 \rfloor$ and $g_2 = \lfloor n\alpha_2 \rfloor$. Note that $TM_\alpha = TM(\alpha, \alpha)$. We prove in Section 5.S.B in the Supplementary Material that the (α_1, α_2)-trimmed mean has a bounded influence curve, and a breakdown point of $min(\alpha_1, \alpha_2)$. We also show in Section 5.S.B the location measure for the (α_1, α_2)-trimmed mean.

The primary challenge with asymmetric trimmed means is selecting the appropriate trimming levels. Some trimmings may perform poorly in symmetric cases, and the best trimming values are not known in advance. One proposal is the *adaptive trimmed means*, in which the trimming values are based on the data obtained by

the sample (Hogg, 1974). A way to adapt asymmetric trimming rates is the hinge location estimators proposed in Reed III and Stark (1996). They defined some measures of skewness and tail length

$$\begin{aligned} Q_1 &= (U_{(.2)} - L_{(.2)})/(U_{(.5)} - L_{(.5)}), \\ Q_2 &= (U_{(.05)} - T_{.25})/(T_{.25} - L_{(.5)}), \\ SK_2 &= (X_{(1)} - XMD)/(XMD - X_{(n)}), \\ SK_5 &= (X_{(1)} - XM))/(XM - X_{(n)}), \end{aligned}$$

where $L_{(\alpha)}$ is the mean of the smallest αn observations, $U_{(\alpha)}$ is the mean of the largest αn observations, $T_{.25}$ is the 0.25-trimmed mean $m_n(0.25)$, XM is the mean, and XMD the median. Calculating the adaptive trimmed means has the following steps:

(1) Set the total amount of trimming γ;

(2) Determine the corresponding trimming proportion from the left side of the distribution

$$\alpha_l = \gamma[UW_X/(UW_X + LW_X)], \tag{5.6}$$

where UW_X and LW_X are the numerator and denominator portions of the statistics above;

(3) Set the upper trimming proportion $\alpha_u = \gamma - \alpha_l$;

(4) Calculate the sample (α_l, α_u)-trimmed mean.

Note that $\alpha_l = \gamma[UW_X/(UW_x + LW_x)] = \gamma[1/(1+1/X)]$, where X represents one of the skewness and tail length measures presented above. Thus, a higher metric value implies more trimming to the lower size.

We denoted the adaptive hinge estimators as HQ_1, HQ_2, HSK_2, and HSK_5. Reed III and Stark (1996) presented more methods. However, we have selected these four because simulations revealed that HQ_1, HQ_2, HSK_2, and HSK_5 were among the top 4 estimators, and HQ_1 and HQ_2 among the top 3 (Reed III and Stark, 1996). Additionally, HSK_2 and HSK_5 were also good with symmetric distributions. In addition, Keselman et al. (2007) conducted some experiments to measure the type I error and power of the estimators testing the equality of means, showing that HQ_1 is the best estimator in terms of power and protection against type I error, and HSK_2 and HSK_5 are better with a distribution not far from a normal distribution. To the best of our knowledge, there is no research regarding the classical measures of robustness. One possible reason is the early criticism of other adaptive procedures, like the asymptotically efficient adaptive procedures (Huber and Ronchetti, 2009). A criticism of the adaptive trimmed means is the difficulty of differentiating skewness from contamination (Huber and Ronchetti, 2009). However, adaptive procedures have been shown to be empirically adequate in some families of distributions (Agostinelli et al., 2016) and for hypothesis testing of equality of means (Keselman et al., 2007).

M-estimators. M-estimators are estimators associated with the minimization problem

$$\arg\min_{\theta} \sum_{i=1}^{n} \xi(x_i; \theta), \tag{5.7}$$

or, with equation

$$\sum \psi(x_i; \theta) = 0, \tag{5.8}$$

being ψ the derivative of ξ with respect to θ. The associated measure is the solution of

$$E_F[\psi(X; T(F))] = \int \psi(x; T(F)) F(dx). \tag{5.9}$$

An M-estimator of location is the solution to Eq. 5.7 in the location problem, i.e., when $\xi(x; \theta)$ is expressed as $\xi(x - \theta)$. One example of an M-estimator is the sample mean, which is the M-estimator with $\xi(x) = x^2/2$. Well-known M-estimators are based on the MoR-Huber functions

$$\psi(x) = \max\{-k, \min(x, k)\} \tag{5.10}$$

and

$$\xi(x) = \begin{cases} 1/2x^2 & \text{if } x \leq K, \\ |x|K - 1/2K^2 & \text{if } x > K. \end{cases} \tag{5.11}$$

The idea of the Huber M-estimators is to assign less weight to the extreme values that can distort the sample mean. In particular, we will use the Huber Proposal 2 (Huber and Ronchetti, 2009). This is a Huber M-estimator that uses a scale parameter τ. Concretely, the M-estimator is a solution of

$$\sum_i \psi\left(\frac{x_i - \theta}{\tau}\right) = 0. \tag{5.12}$$

In this proposal, τ and θ are estimated jointly, using an M-estimator of scale for the scale parameter τ (Huber and Ronchetti, 2009). Specifically, we use the MoR-Huber M-estimator with $K = 1.5$. The breakdown point for Proposal 2 is 0.27, and it has a bounded influence function (Huber and Ronchetti, 2009).

An M-estimator well suited for heavy-tail distributions is the myriad (Arce, 2004). It gives less weight to the extreme values than the Huber type. Particularly, it corresponds to the maximum likelihood estimator for the Cauchy distribution (Arce, 2004). It is defined as an M-estimator with ξ function

$$\xi(x) = \log(K^2 + x^2), \tag{5.13}$$

where K is a parameter that measures the influence of the extreme values, with a larger K representing a greater effect on these values. The myriad is employed as a filter for signal processing (Ramirez and Paredes, 2017), where K represents the rejection capability to impulsive noise (Ramirez and Paredes (2016)). To calculate the solution of the myriad, we will use the fast weighted myriad computation of Arce (2004).

5.3.2. Robustification of MoR and RoS

In the previous section, various location measures resistant to contamination by outlying values have been introduced. They behave similarly to the mean under normal or symmetric short-tailed distributions but offer better performance in the presence of asymmetry and heavy tails. This flexibility allows us to change the underlying assumptions of the classical NSUM estimators.

Robustification of MoR. MoR relies on the assumption that the average of y_i/d_i is representative. This can be expressed in our context as $E[y_i/d_i] = \rho$. Our proposals to robustify the MoR consist of substituting the population mean with one of the robust location measures T, i.e., assuming that $T[y_i/d_i] = \rho$. This leads to an estimator of the prevalence using the location estimator applied to the sample ratios

$$T(I_n) = T(y_1/d_1, \cdots, y_n/d_n). \tag{5.14}$$

Under distributions closer to a Normal distribution, the estimates are similar to the sample means, but with estimates more resistant to deviations.

Table 5.1: Robust NSUM estimators for each type of classical NSUM.

Robust NSUM	
MoR**-type**	RoS**-type**
MoR-Median	RoS-Median
MoR-TM10	RoS-TM10
MoR-HQ1	RoS-HQ1
MoR-HQ2	RoS-HQ2
MoR-HSK2	RoS-HSK2
MoR-HSK5	RoS-HSK5
MoR-Huber	RoS-Huber
MoR-Myriad	RoS-Myriad

Robustification of RoS. For the RoS, we first observe that for the Binomial model

$$E[y_i] = E[E[y_i|d_i]] = \sum_k k\rho P[d_i = k] = \rho E[d_i]. \tag{5.15}$$

Similarly to MoR, we assume that there are some location measures such that $T[y_i] = \rho T[d_i]$. We can construct an estimator for the prevalence by taking the ratios of the estimators of y_i and d_i

$$RoT(I_n) = T(y_1, \cdots, y_n)/T(d_1, \cdots, d_n). \tag{5.16}$$

Table 5.1 displays the robust estimates that we use in the rest of the study. Each column corresponds to the robust proposals for each non-robust estimator.

The first column represents the robustification of MoR. We use the location estimator, which applies the ratios of y_i/d_i, and it is denoted using the name of the location estimator. MoR-TM10 denotes the 0.10-trimmed mean $TM_{0.10}$ and MoR-Huber is the name of the Huber M-estimator. The right column represents the robustification of RoS, which uses the ratios of the location estimators.

5.4. Simulation Study

This section evaluates the performance of the NSUM estimators through an extensive simulation study designed to capture the properties of real-world networks. This approach allows us to gather evidence on how these estimators perform under various conditions that reflect practical scenarios. First, we define the environment and its relation to conditions impacting the stability of the classical NSUM. Subsequently, we present the simulation results. Finally, we include a discussion and remarks on the performance of the estimators. Additional analysis of the barrier effect and other biases–transmission error and recall error–is presented in the Supplementary Material.

5.4.1. Simulation environments

This section outlines the simulation environment used to assess robust NSUM estimation methods. Specifically, we generate synthetic graphs based on diverse network models, unknown groups with different prevalence levels, and sampling schemes with varying contamination intensities. This simulation framework provides insights into the scenarios where robust estimators must be most effective. In this work, we consider three network models: Erdős–Rényi (ER), Scale-Free (SF), and Stochastic Block Model (SBM). ER networks represent models with good conditions for the MoR and RoS when the prevalence is high. The ER models with low prevalences reflect deviations due to skewness and tail length. An extreme case of these conditions is reflected in SF networks. Finally, SBM networks assess the influence of the barrier effect. For a comprehensive analysis, we describe the characteristics of each model, the type of contamination applied, and the metrics used to evaluate the quality of the estimators.

Erdős–Rényi

The Erdős–Rényi (ER) model is a classic random graph model widely used to study the properties of networks. This model involves randomly connecting nodes with a fixed probability $p = \mu/(N-1)$, where μ represents the mean degree and N is the network size (Cheng et al., 2019). We have chosen the ER model to represent favorable conditions for the classical methods when the prevalence is moderate.

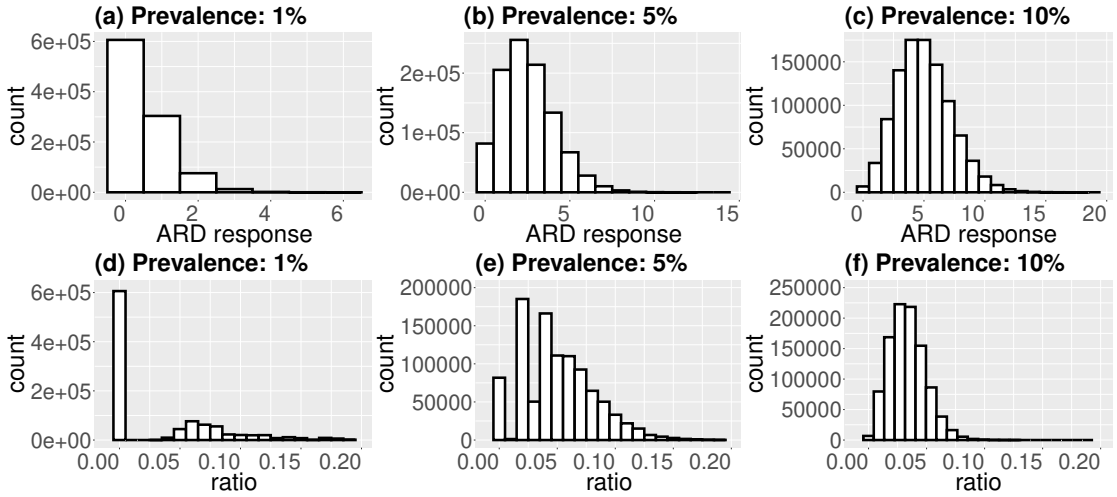


Figure 5.1: Erdős–Rényi ($\mu = 50$). Histogram of ARD responses y_i for (a) $\rho = 1\%$, (b) $\rho = 5\%$, and (c) $\rho = 10\%$. Histogram of ratios y_i/d_i for (d) $\rho = 1\%$, (e) $\rho = 5\%$, and (f) $\rho = 10\%$.

Figure 5.1(f) illustrates that, with relatively high prevalence, the distributions of the ratios are very close to a symmetric distribution with short tails, which benefits the performance of MoR. Additionally, the ER model closely matches the underlying assumptions of RoS, as the conditional distribution of y_i given d_i under an ER model follows a Hypergeometric distribution, which approximates a Binomial distribution when both the network and the unknown group are large (Kunke et al., 2024). The Binomial model corresponds to a similar network model in which nodes connect with others at a fixed probability, with replacement (Kunke et al., 2024). Thus, the ER model provides an ideal balance between the theoretical foundations of RoS and real-world applicability.

We use low prevalences to assess the stability of the NSUM estimators. We observe in Figure 5.1 that the distribution of the ARD of the unknown group and the ratios present skewness and heavy tails as the prevalence decreases. In addition, low prevalences compromise the approximation to normal distributions of the ARD and the ratios.

Scale-Free

This network model belongs to a class of random graphs with non-uniform degree distributions. The degree of SF networks follows a power law distribution, i.e., the probability $p(d)$ that a node in the network has d connections to other nodes is expressed as $d^{-\gamma}$ (usually with $\gamma > 2$). In particular, this implies a heavy tail in the degree distribution, in which most nodes in these networks have relatively low degrees (few connections to other nodes), while a few nodes, dubbed hubs, have very high degrees (many connections). The SF network model is widely adopted to

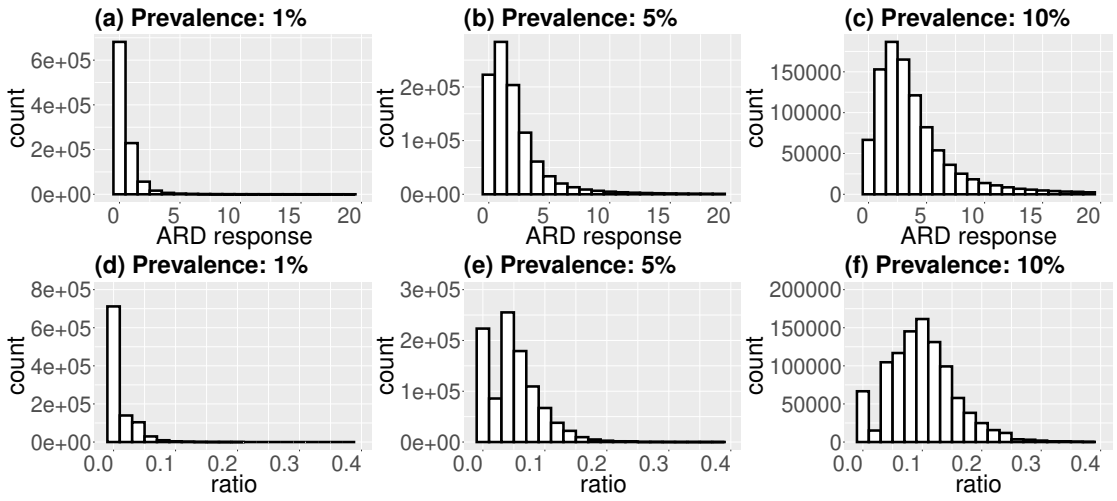


Figure 5.2: Scale-Free network model ($\gamma = 2.50$). Histogram of ARD responses y_i for (a) $\rho = 1\%$, (b) $\rho = 5\%$, and (c) $\rho = 10\%$. Histogram of ratios y_i/d_i for (d) $\rho = 1\%$, (e) $\rho = 5\%$, and () $\rho = 10\%$.

simulate many real-world networks, including social and biological systems (Newman, 2018).

Particularly, we use the model of Goh et al. (2001). This model indexes the nodes with integer numbers, and gives the ith node a weight proportional to $i^{-\alpha}$, for $\alpha \in [0, 1)$. Then, two nodes are selected to form a link with a probability proportional to their respective normalized weights. Goh et al. (2001) show that the degree distribution follows a power law with $\gamma = (\alpha + 1)/\alpha$.

This model represents a significant departure from conditions of symmetry and short tails under which the classical methods perform well. Figure 5.2 shows that the distributions exhibit increased skewness and longer tails, even at higher prevalence levels.

Stochastic Block Model

The barrier effect involves different propensities to be connected to different groups depending on the characteristics of the participants. One type of barrier effect arises when the characteristics depend on membership in some groups. Kunke et al. (2024) proposed an SBM to model the propensity of different groups to form links. Specifically, the SBM model is a widely used network model that aims at capturing community patterns in which nodes do not have an equal likelihood of forming links within any group. These networks are valuable for social network analysis, where the interactions within and between distinct groups are different. In general, nodes within the same group are more (or less) likely to connect with each other than with nodes in different groups. To be more precise, consider a network with N nodes partitioned into K disjoint groups, i.e., each node belongs to exactly one group.

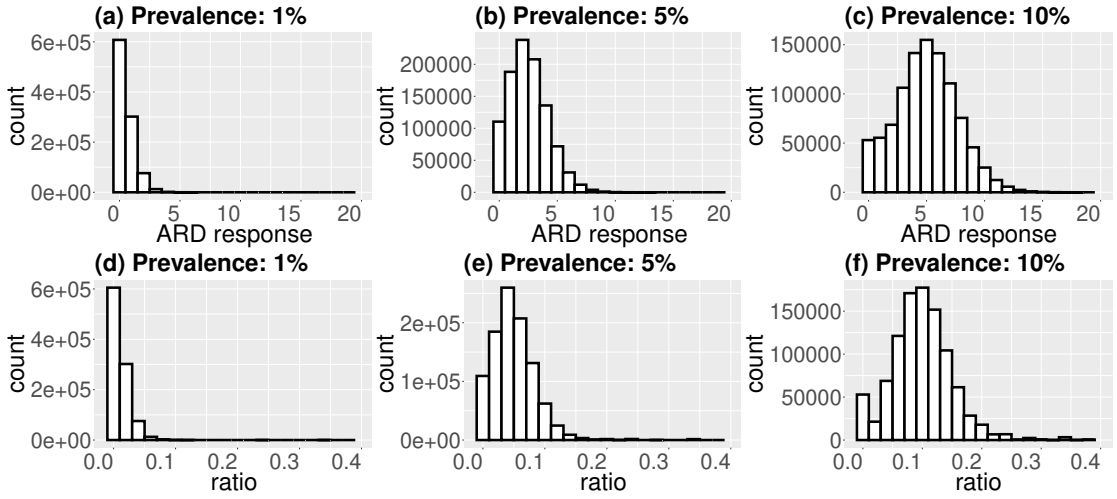


Figure 5.3: Stochastic Block model ($a = e^4$). Histogram of ARD responses y_i for (a) $\rho = 1\%$, (b) $\rho = 5\%$, and (c) $\rho = 10\%$. Histogram of ratios y_i/d_i for (d) $\rho = 1\%$, (e) $\rho = 5\%$, and (f) $\rho = 10\%$.

The likelihood of an edge (connection) between two nodes depends on the groups to which the nodes belong. This behavior is typically represented by a symmetric matrix denoted as P, where p_{ij} represents the probability of an edge between a node in group i and a node in group j.

In our case, we replicate the scenario of (Kunke et al., 2024). They consider two blocks, the unknown group and its complement. The probability of connection between the two groups is p, while the intra-group link probability is ap, where a is a non-negative real representing the network assortativity. When $a < 1$ (dissortativity), nodes in the unknown subpopulation are less likely to connect with other nodes in the same group, which is often unrealistic in practical applications. When $a = 1$, it corresponds to the Erdős–Rényi network model. In contrast, when $a > 1$ (assortativity), the nodes in the unknown subpopulation are more likely to connect with other nodes within the same population (Kunke et al., 2024).

Figure 5.3 shows the impact of prevalence on the skewness and tails of the ARD distribution in a highly assortative scenario.

Contamination model

In our simulation studies, we address the introduction of contamination to the ARD response vector to evaluate the resilience of robust estimation methods under adverse conditions. Specifically, for ER and SF network models, a predetermined fraction ϵ of the ARD response vector entries y_i is deliberately altered to introduce noise. This alteration is conducted as follows. Half of the specified proportion $\epsilon/2$ of the ARD response vector entries is set to zero. This perturbation simulates scenarios where response data might be missing or fail to be reported, a common issue in indirect

survey data collection processes. The remaining half $\epsilon/2$ of the affected entries is substituted with the respective node degree value d_i. This replacement simulates a different type of data corruption, potentially due to errors in data handling or misalignment in database entries.

The contamination scenarios defined (missing data and data corruption) constitute extreme forms of under- and over-reporting, something highly relevant to NSUM techniques. Furthermore, some of these contaminations can be classified as biases typical of ARD and NSUM techniques.

Metrics

In this study, we assess the performance of robust NSUM estimators through two statistical metrics: the root mean squared relative error (RMSRE) and normalized bias (NB). These metrics allow us to evaluate the accuracy and reliability of robust estimators under varied simulation conditions. On the one hand, the RMSRE measures the relative discrepancies between the estimated values and the actual prevalence. More precisely, the RMSRE is defined as follows

$$\text{RMSRE} = \sqrt{\frac{1}{R}\sum_{i=1}^{R}\left(\frac{\rho_i - \rho}{\rho}\right)^2} \tag{5.17}$$

where R is the number of realizations or simulation runs, ρ_i denotes the estimated prevalence obtained from the ith realization i, and ρ is the actual prevalence value. This metric effectively captures the average magnitude of errors relative to the true value, providing a measure of estimator precision. To complement the analysis provided by RMSRE, we also calculate the normalized bias (NB), which offers insights into the systematic deviation of the estimators from the actual value. The NB is expressed as

$$\text{NB} = \frac{\left(\frac{1}{R}\sum_{i=1}^{R}\rho_i\right) - \rho}{\rho} = \frac{\left(\frac{1}{R}\sum_{i=1}^{R}\rho_i\right)}{\rho} - 1. \tag{5.18}$$

Hence, the NB quantifies how closely the average estimate matches the actual parameter, normalized by the actual prevalence value. NB helps compare estimators across different prevalence levels as it scales the bias to the actual prevalence value.

5.4.2. Simulations Results

For each network model, we generated ten synthetic graphs with 10^6 nodes to simulate large-scale network structures. It is essential to mention that our analysis focuses on assessing the accuracy of robust NSUM estimators for prevalence levels ranging from 1% to 10%. In this context, we created ten unknown subpopulations for each network model and prevalence value ρ. Notice that prevalence refers to the proportion of the network nodes belonging to the unknown subpopulation. Every value in tables

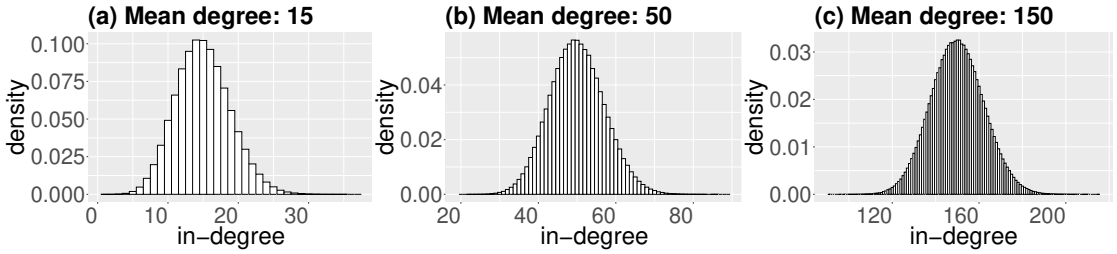


Figure 5.4: Erdős–Rényi. Histogram of the degree for (a) $\mu = 15$, (b) $\mu = 50$, and (c) $\mu = 150$.

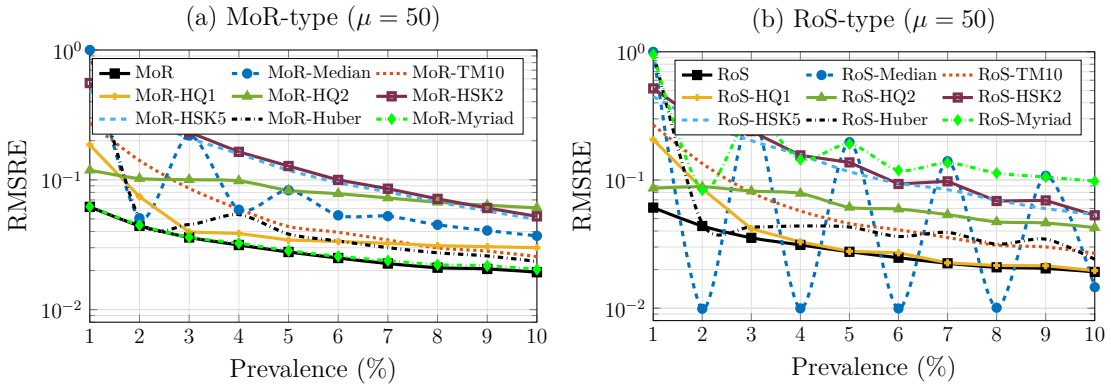


Figure 5.5: Erdős–Rényi ($\mu = 50$). RMSRE versus the prevalence without contamination for the groups of robust estimators: (a) MoR-type and (b) RoS-type.

and figures shown in the section on simulation results is obtained by averaging 1000 realizations. Specifically, the 1000 realizations consider ten networks of a particular model, ten unknown subpopulations with a given ρ, and ten trials. At each trial, a set of node degree samples and a set of ARD responses with a sample size of $n = 500$ are extracted. In each trial, samples are selected uniformly at random.

Regarding the parameter of the robust NSUM, we fix the trimming proportion to 0.20, i.e., the trimmed means give zero weight to 20% of the data. In particular, we use $TM_{0.1}$, which we will refer to as MoR-TM10 and RoS-TM10 in the rest of the study. Furthermore, the linearity parameter K assigned to the Myriad estimator is 0.2. We aim to assess the performance of the Myriad-based estimators across various forms of contamination related to NSUM while maintaining constant linearity parameters. The parameter for the MoR-Huber and RoS-Huber M-estimators is fixed to 1.5.

Erdős–Rényi

This section evaluates the performance of robust NSUM estimation methods using synthetic graphs generated according to the ER model, specifically focusing on networks with mean degrees of 15, 50, and 150. These values were selected considering the social group sizes for good friends, friends, and casual acquaintances according to

Dunbar's theory of social circles (Dunbar, 2010). Figure 5.4 shows histograms of the degree distribution for ER networks with (a) $\mu = 15$, (b) $\mu = 50$, and (c) $\mu = 150$. As expected, these histograms exhibit a bell-shaped profile centered on the average degree, similar to a binomial distribution.

First, we examine the performance of robust estimators across varying prevalence levels under conditions without contamination. This setup enables the examination of their usability, accuracy, and potential biases without external disruptions, providing insight into their reliability in ideal scenarios without contamination. Our primary focus is on networks with a mean degree of 50, as this provides a midrange complexity suitable for detailed analysis. Figures 5.1(a)-(c) display the histograms of ARD responses for unknown groups at prevalence levels of 1%, 5%, and 10%, respectively. At the lowest prevalence (1%), more than 60% of respondents report zero known members in the hidden group. As the prevalence increases to 5% and 10%, zero-valued ARD responses show a noticeable decrease. Figures 5.1(d)-(f) complement this analysis by showing histograms of the ratios for the same prevalence levels. These histograms exhibit a concentration at zero when prevalence is low, gradually diminishing as the hidden population becomes more prevalent.

Figure 5.5 illustrates the RMSRE obtained by robust estimation methods as a function of the prevalence of the hidden population without contamination. Figure 5.5(a) displays the RMSRE for the MoR-type estimators across the prevalence interval. Without contamination, the non-robust MoR estimator consistently achieves the lowest RMSRE values, outperforming all robust variants across the prevalence range. Among the robust approaches, MoR-Myriad exhibits the closest performance to the standard MoR estimator, indicating that this estimator is effective in clean environments. Other robust estimators, such as MoR-TM10, MoR-HQ1, and MoR-Huber, show progressively improved accuracy as prevalence increases, whereas MoR-Median exhibits unstable performance, particularly at low prevalence levels. Notably, MoR-HQ2, MoR-HSK2, and MoR-HSK5 perform comparatively poorly, yielding the highest RMSRE values and demonstrating limited utility in uncontaminated settings. For additional insights, Figure 5.S.1 in the Supplementary Material provides boxplots of the estimated hidden population proportions obtained by the MoR-type methods for prevalence levels ranging from 1% to 10%. These visualizations offer additional insight into estimator variability and bias under uncontaminated conditions.

Figure 5.5(b) displays the RMSRE obtained by RoS-type methods versus the prevalence in the absence of contamination. Similar to the MoR family, the standard (non-robust) RoS estimator consistently yields the lowest RMSRE values across the prevalence range, indicating superior efficiency under uncontaminated conditions. RoS-Median occasionally outperforms the standard RoS estimator at specific prevalence levels. It is worth noting that the RoS-Median estimator exhibits oscillatory behavior as prevalence increases. This phenomenon stems from the use of medi-

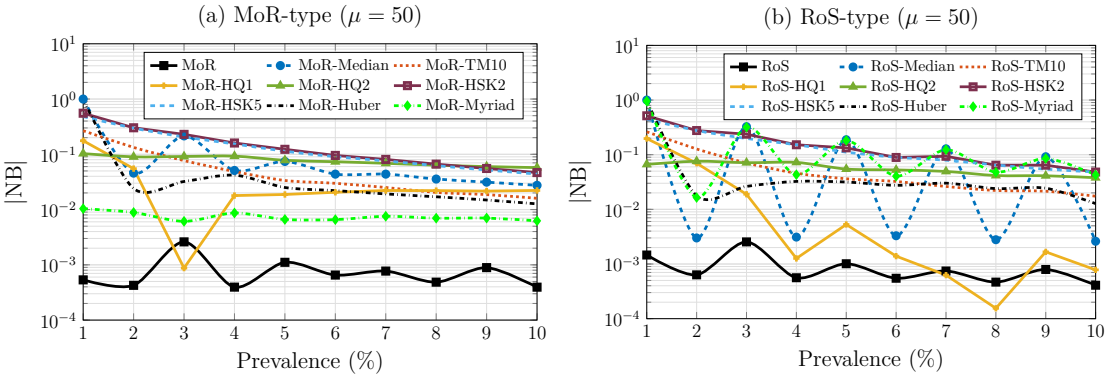


Figure 5.6: Erdős–Rényi ($\mu = 50$). |NB| versus the prevalence without contamination for the groups of robust estimators: (a) MoR-type and (b) RoS-type.

ans, which are not smoothly responsive to small changes in the data. Specifically, RoS-Median is defined as the ratio between the median of the ARD responses and the median of the reported degrees. Because the median is a discrete, order-based statistic, the ratio can remain constant over a range of prevalence values and then suddenly shift when a small change in prevalence causes the median of the ARD response to increase by one. For example, in a network with a median degree of 50, the estimator can shift from 1/50 to 2/50 as soon as the median ARD count increases by one, resulting in a stepwise behavior in estimation and an oscillatory error pattern as prevalence changes smoothly. RoS-Myriad also shows an unstable behavior. These trends are further illustrated in the boxplots of Figure 5.S.2 (Supplementary Material). The RoS-HQ1, RoS-Huber, and RoS-TM10 show improved accuracy as prevalence increases, indicating that the performance of these methods becomes more reliable with a larger unknown population. In contrast, RoS-HSK2 and RoS-HSK5 consistently underperform across all prevalence levels, producing the highest RMSRE values.

Figure 5.6(a) displays the absolute value of the normalized bias |NB| obtained by MoR-type estimators for different prevalence levels (1% to 10%) under contamination-free conditions. As expected, the standard (non-robust) MoR estimator yields the lowest bias throughout the prevalence interval. This behavior is consistent with the non-robust MoR being an unbiased maximum likelihood estimator (MLE) when ratios follow a Gaussian distribution (Laga et al., 2021). On the other hand, all MoR-type robust methods exhibit non-zero bias when there is no contamination, indicating deviations of estimations from actual prevalence values. We also observe a reduction in bias as the prevalence increases for all robust techniques, leading to more accurate estimations as the proportion of the unknown group increases. Among the robust approaches, MoR-Myriad consistently achieves the lowest bias across all prevalence levels. Moreover, MoR-HQ1 and MoR-Huber estimators also exhibit relatively low bias for prevalence levels above 1%, reinforcing their utility in moderate or high prevalence scenarios.

Figure 5.6(b) illustrates the absolute values of the normalized bias |NB| yielded by the RoS-type estimators across a range of prevalence levels under uncontaminated conditions. The non-robust RoS method exhibits the lowest bias across prevalence levels. Among the robust methods, RoS-HQ1 shows very low bias magnitudes as the prevalence increases, suggesting minimal deviation from the true prevalence. In contrast, RoS-Median shows high variability in its bias, especially at lower prevalence levels, reducing its reliability as an estimator in scenarios involving small hidden populations. The RoS-HSK2 and RoS-HSK5 estimators yield the largest bias values across the entire prevalence range, indicating poor estimation accuracy in environments without contamination.

To further examine the performance of the estimation methods in scenarios without contamination, Figure 5.S.3 in the Supplementary Material depicts the RMSRE produced by the various estimation techniques against the prevalence for ER networks with a mean degree of $\mu = 15$. Moreover, Figure 5.S.4 presents the absolute values of the NB yielded by the estimators for different prevalence values. In addition, Figures 5.S.5 and 5.S.6 display the boxplots of the estimates of the unknown population rate yielded by the various methods for different prevalence levels in the absence of contamination and mean degree of $\mu = 15$.

Similarly, Figure 5.S.7 in the Supplementary Material shows the RMSRE generated by the NSUM estimators as a function of the prevalence for ER networks with a mean degree of $\mu = 150$. Figure 5.S.8 displays the absolute values of the NB for different prevalence levels without contamination. These figures indicate that the non-robust methods outperform the robust estimators when the network follows an ER model in scenarios without contamination. Additionally, robust estimation methods exhibit lower error and bias values as the mean degree increases, indicating that their responses are more accurate as the link probabilities between network nodes increase. Figures 5.S.9 and 5.S.10 show the boxplots of the estimates of the unknown population rate produced by the various methods for different prevalence levels in the absence of contamination and mean degree $\mu = 150$.

We also evaluate the performance of NSUM estimation methods for the ER model under varying contamination levels. Figures 5.7(a)-(c) display the RMSRE produced by the MoR-type methods at contamination levels ranging from 0% (no noise) to 10%, across three prevalence values of 1%, 5%, and 10%, respectively. As illustrated in the figures, the performance of the non-robust MoR estimator significantly deteriorates as the level of contamination increases across all three prevalence scenarios. It is worth noting that the MoR-HQ2 method exhibits a marked decline in performance with increasing contamination. In particular, once the contamination level exceeds 2%, MoR-HQ2 produces estimation errors larger than those of the non-robust MoR estimator, indicating inferior performance in contaminated settings. In contrast to the MoR and MoR-HQ2 methods, which exhibit considerable sensitivity to noise, the remaining robust estimators maintain relatively stable performance across

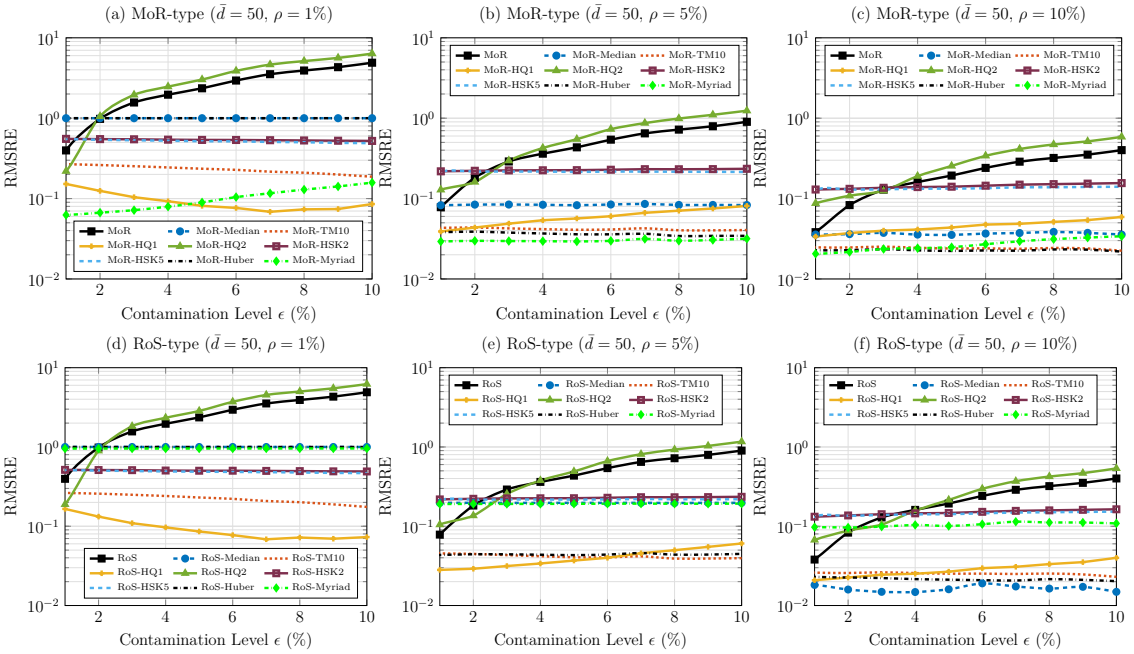


Figure 5.7: Erdős–Rényi ($\mu = 50$). RMSRE versus the contamination level for MoR-type methods and prevalences: (a) 1%, (b) 5%, (c) 10%; RoS-type methods and prevalences: (d) 1%, (e) 5%, (f) 10%.

varying contamination levels. As shown in Figure 5.7(a), for a prevalence of 1%, the robust estimator MoR-Myriad achieves RMSRE values of at least one order of magnitude lower than those produced by the non-robust MoR method. This highlights the robustness of this approach, particularly in scenarios involving small hidden populations, a setting that often represents a worst-case scenario in ER networks. Similar performance gains are observed for prevalence levels of 5% and 10%. Figures 5.S.11, 5.S.13, and 5.S.15 in the Supplementary Material depict the boxplots of the unknown population rate estimates generated by the MoR-type estimator at prevalences 1%, 5%, and 10%.

Among the robust methods, MoR-Myriad, MoR-TM10, and MoR-HQ1 estimators consistently exhibit low error values across the entire contamination interval for the three prevalence levels, highlighting their potential for application in noisy, large-scale networks. Although MoR-Huber and MoR-Median show improved performance at higher prevalence levels, they exhibit significant errors at a 1% prevalence, which can be problematic in estimating small unknown groups. To provide a deeper insight, Figures 5.S.17(a)-(c) show |NB| versus the contamination levels produced by the MoR-type methods for prevalences of 1%, 5%, and 10%, respectively.

Figures 5.7(d)-(f) show the RMSRE yielded by RoS-type methods versus contamination levels across prevalences of 1%, 5%, and 10%, respectively. In addition, Figures 5.S.12, 5.S.14, and 5.S.16 in the Supplementary Material illustrate the boxplots obtained by the RoS-type techniques against contamination levels and

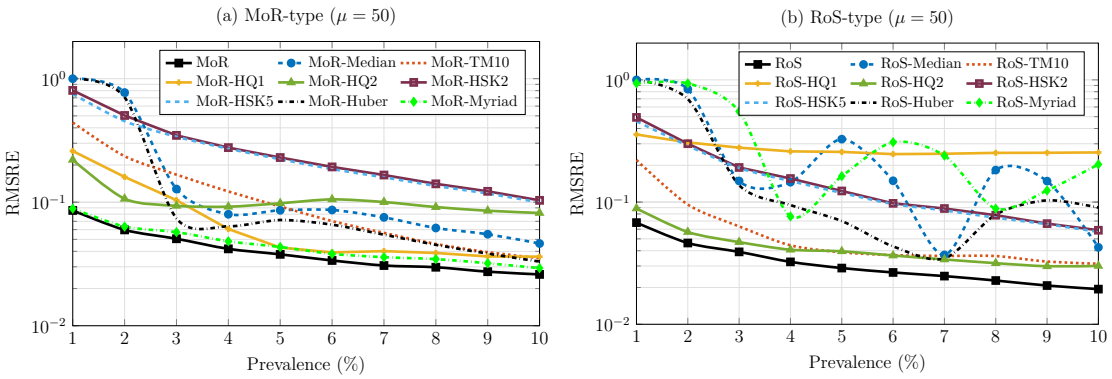


Figure 5.8: Scale-Free ($\mu = 50$). RMSRE versus the prevalence without contamination for the groups of robust estimators: (a) MoR-type and (b) RoS-type.

prevalence 1%, 5%, and 10%, respectively. Moreover, Figures 5.S.17(d)-(f) in the Supplementary Material depict the |NB| against the contamination level yielded by the RoS-type estimators for prevalences of 1%, 5%, and 10%, respectively. It is observed that both errors and bias values generated by the non-robust RoS and RoS-HQ2 methods significantly increase as the contamination levels increase, demonstrating their sensitivity to noise. The remaining RoS-type methods, however, exhibit consistent error and bias values throughout the contamination interval. Among the robust methods, the RoS-HQ1 and RoS-TM10 estimators perform best for the three prevalence values. Figures 5.S.18, 5.S.19, 5.S.20, 5.S.21 in the Supplementary Material illustrate the behavior of various NSUM estimators in response to perturbations simulating other NSUM biases (precisely, transmission error and recall error) for the ER network model.

Scale-Free

This section evaluates the performance of the NSUM estimation methods using Scale-Free networks characterized by a power-law degree distribution with a scaling exponent $\gamma = 2.50$ and mean degree $\mu = 50$. Figure 5.2 illustrates the histograms of ARD responses and ratios for three prevalence levels: 1%, 5%, and 10%. First, we focus on assessing the accuracy of the estimations without contamination. Figure 5.8 displays the RMSRE obtained by the different estimators versus the prevalence in the absence of noise. Additionally, the boxplots of the unknown population rate estimated by the various estimators for prevalence values ranging from 1% to 10% without contamination are displayed in Figures 5.S.22 and 5.S.23 in the Supplementary Material.

Figure 5.8(a) shows the RMSRE generated by MoR-type methods versus the prevalence. In addition, Figure 5.S.24(a) in the Supplementary Material illustrates the |NB| values yielded by the MoR-type methods for different prevalence levels. In SF networks with $\gamma = 2.50$ and no contamination, the non-robust MoR method gives

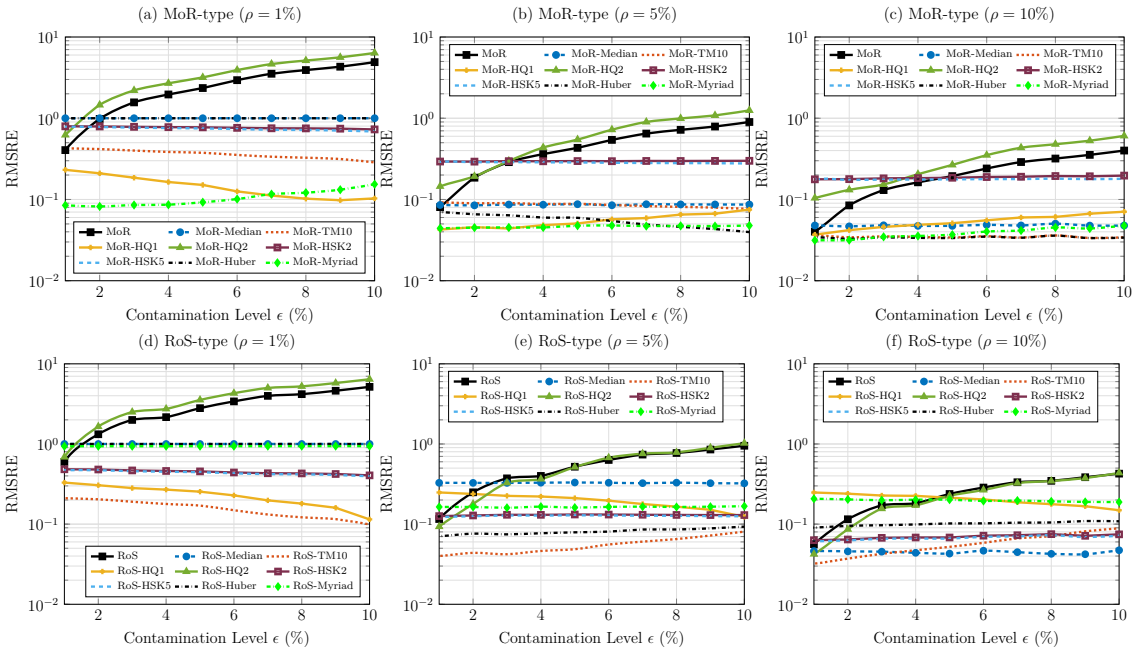


Figure 5.9: Scale-Free network model ($\gamma = 2.50$). RMSRE versus the contamination level for MoR-type methods and prevalences: (a) 1%, (b) 5%, (c) 10%; RoS-type methods and prevalences: (d) 1%, (e) 5%, (f) 10%.

the lowest error and bias values across the entire prevalence interval, similarly to the ER network results. These results indicate that this method maintains high accuracy in scenarios without contamination despite the non-robust MoR estimator not being the optimal operator when the network degree follows a power-law distribution. Among robust estimators, the MoR-Myriad method performs closest to the non-robust MoR technique. Notice also that MoR-TM10 and MoR-HQ1 slowly improve their error values as the prevalence increases. Observe also that MoR-Median and MoR-Huber yield the largest errors for low prevalence values (1% and 2%), making them unreliable for estimating the prevalence of small unknown populations. In scenarios without contamination, MoR-HSK2 and MoR-HSK5 show the poorest performances regarding RMSRE and NB.

Figure 5.8(b) illustrates the RMSRE produced by RoS-type estimation methods as a function of the prevalence. Further, Figure 5.S.24(b) in the Supplementary Material includes the |NB| values yielded by the RoS-type methods across the same prevalence levels. The non-robust RoS, RoS-HQ2, and RoS-TM10 methods exhibit the lowest RMSRE and NB across all tested prevalence levels. These results indicate that RoS, RoS-HQ2, and RoS-TM10 methods produce accurate and consistent estimates in uncontaminated environments. Conversely, the RoS-Median, RoS-Myriad, and RoS-Huber methods demonstrate variability in their performance across the prevalence axis. More precisely, these methods exhibit unstable behavior, particularly noted in their fluctuating error rates along the prevalence levels. A notable decline in effectiveness is observed at lower prevalence levels, a significant

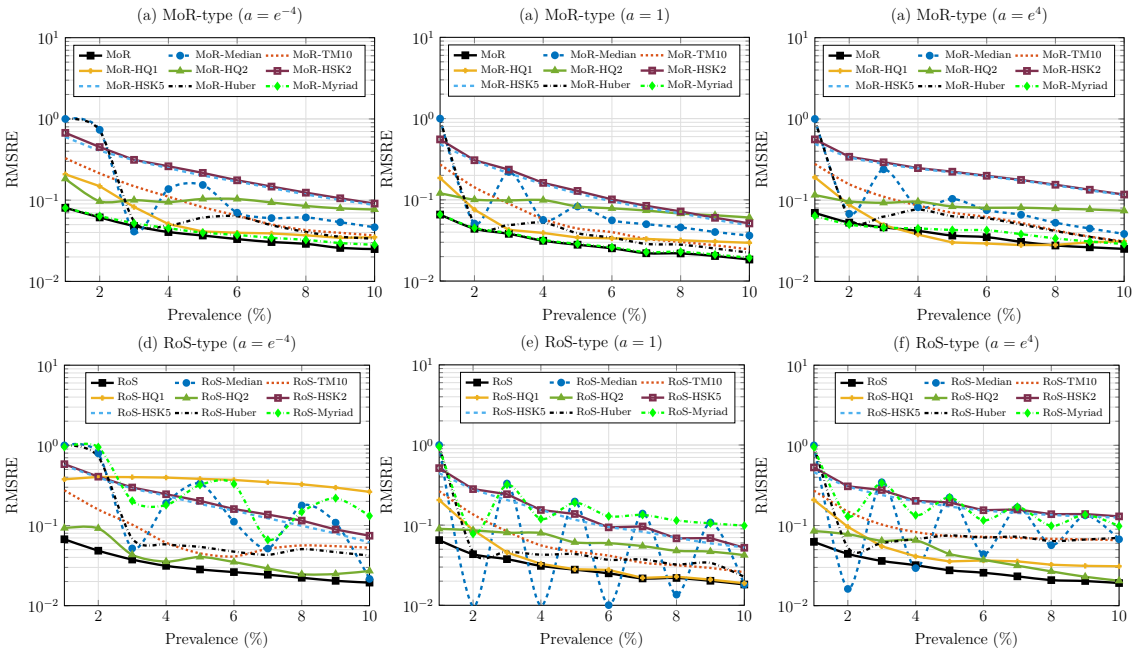


Figure 5.10: Stochastic Block Model. RMSRE versus the prevalence level for different values of the connectivity factor a. MoR-type methods: (a) $a = e^{-4}$, (b) $a = 1$, (c) $a = e^{4}$; RoS-type methods: (d) $a = e^{-4}$, (e) $a = 1$, (f) $a = e^{4}$.

limitation for applications involving small, unknown groups.

We have also analyzed the performance of NSUM estimators using Scale-Free networks under noisy conditions and varying prevalence levels. Figures 5.9(a)-(c) illustrate the RMSRE produced by the MoR-type estimators as a function of the contamination level at three different prevalence values: 1%, 5%, and 10%. It is worth noting that the error values decrease for all estimation methods as prevalence levels increase, which suggests that higher node activation within the network leads to more accurate estimations regardless of noise. For the MoR and MoR-HQ2 methods, error values increase significantly as contamination levels rise, highlighting their limitations in applications with highly contaminated environments. On the other hand, the MoR-Myriad operator and the MoR-HQ1 method consistently deliver the best performance across the contamination range for the three prevalence levels. Notably, at the 1% prevalence level, MoR-Median, MoR-Huber, MoR-HSK2, and MoR-HSK5 exhibit significant errors, underlying the weakness of these methods when dealing with small unknown groups.

Figures 5.9(d)-(f) depict the RMSRE yielded by the RoS-type estimators versus the contamination level at three different prevalence values: 1%, 5%, and 10%. In this case, the RoS-TM10 estimator consistently performs best along the contamination interval for the three prevalence levels. Finally, Figure 5.S.25 in the Supplementary Material displays the |NB| produced by robust estimators versus the contamination level for the three prevalence values. Figures 5.S.26, 5.S.27, 5.S.28, 5.S.29 in the

Supplementary Material depict the behavior of NSUM estimators in response to perturbations simulating other NSUM biases (transmission error and recall error) for the Scale-Free network model.

Stochastic Block Model

In this section, we analyze the performance of the NSUM estimators using two-group SBM graphs with varying connectivity factor a. These graphs maintain a constant mean degree of $\mu = 50$. We apply nine distinct factors $a = e^M$ with $M = -4, -3, \ldots, 3, 4$. Positive values of M represent more connections between the unknown group, while negative values of M represent more connections between the unknown group and the rest of the individuals of the population. This set of values enables us to assess the robustness of NSUM estimation methods concerning the barrier effect. All graphs have the same mean degree, ensuring that the average connectivity across the network remains consistent. Figure 5.3 displays the histograms of ARD responses and ratios for assortativity $a = e^4$ and prevalence values of 1%, 5%, and 10%. Figures 5.10(a)-(c) display the RMSRE versus the prevalence yielded by the MoR-type methods for three different connectivity factors a: e^{-4}, 1, and e^4. These factors indicate various levels of connectivity within the group. To provide a comprehensive analysis, Figures 5.S.30(a)-(c) in the Supplementary Material present the |NB| values generated by the MoR-type methods versus the prevalence for connectivity factors $a = e^{-4}$, $a = 1$, $a = e^4$, respectively. The non-robust MoR and the MoR-Myriad methods consistently show superior performance across the entire prevalence level interval for the three connectivity factors. Conversely, the MoR-HQ2 method consistently produces errors about one order of magnitude larger than those obtained by the non-robust MoR method across the prevalence interval for the three connectivity factors. In addition, the MoR-HQ1 and MoR-TM10 exhibit improved accuracy for the three connectivity factors as the prevalence increases. In other words, these estimators perform poorly at low prevalence values but demonstrate better accuracy as the unknown group rate rises. The MoR-Median and MoR-Huber methods demonstrate unstable trends when the prevalence is below 6%, indicating potential difficulties in handling low unknown group rates. The MoR-HSK2 and MoR-HSK5 techniques consistently yield the highest error values, suggesting these methods may be less suitable for networks modeled by SBM with the specified parameters.

Figures 5.10 (d)-(f) provide a visual representation of the RMSRE generated by the RoS-type estimators as a function of the prevalence under three distinct connectivity factors a: e^{-4}, 1, and e^4. Figures 5.S.30(d)-(f) in the Supplementary Material complement this analysis by displaying the |NB| values for these methods when the connectivity factor $a = e^4$. In this case, the non-robust RoS method performs best across the prevalence interval for the three connectivity factors. In

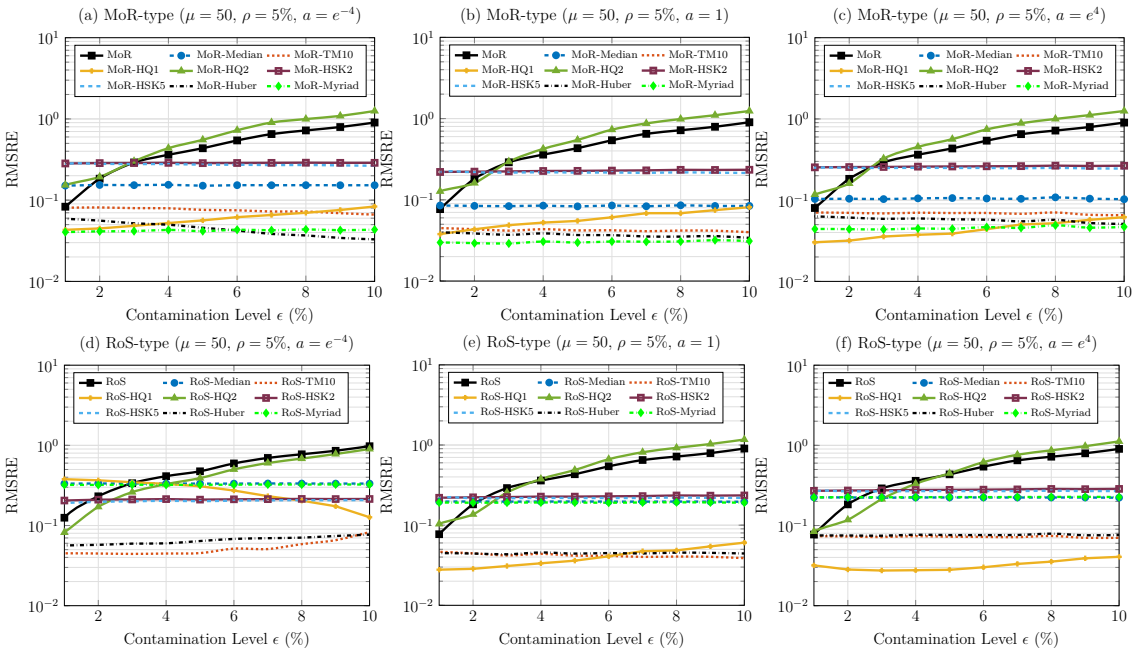


Figure 5.11: Stochastic Block Model ($\gamma = 2.50$). RMSRE versus the contamination level for MoR-type methods, prevalence $\rho = 5\%$ and connectivity factors: (a) $a = e^{-4}$, (b) $a = 1$, (c) $a = e^{4}$; RoS-type methods, prevalence $\rho = 5\%$ and connectivity factors: (d) $a = e^{-4}$, (e) $a = 1$, (f) $a = e^{4}$.

networks with a low connectivity factor (where $a = e^{-4}$), the RoS-HQ1 method shows high RMSRE values across all prevalence levels. When the connectivity factor is $a = 1$, the RoS-HQ1 method significantly improves performance, generating the lowest errors for prevalence values exceeding 3%. Similarly, in SBM networks with a connectivity factor of $a = e^{4}$, RoS-HQ1 shows a substantial decrease in errors. The RoS-HQ2 method performs well for $a = e^{-4}$ and $a = e^{4}$ connectivity factors. On the other hand, the MoR-TM10 and MoR-Huber methods gradually reduce their errors for the three connectivity factors as the prevalence increases. Finally, the RoS-Median, RoS-Myriad, RoS-HSK2, and MoR-HSK5 display the more significant error values in SBM networks.

Figures 5.11(a)–(c) present the RMSRE values obtained by the MoR-type estimators as a function of contamination levels for a prevalence rate of $\rho = 5\%$, under three different connectivity scenarios defined by the parameter a: (a) $a = e^{-4}$, (b) $a = 1$, and (c) $a = e^{4}$. In addition, Figures 5.S.32(a)-(c) in the Supplementary material depict the absolute normalized bias ($|\text{NB}|$) for the same settings, offering complementary insight into estimator accuracy. Among the robust methods, MoR-TM10, MoR-HQ1, MoR-Huber, and MoR-Myriad consistently demonstrate superior performance across all contamination levels and connectivity factors. Their ability to maintain low RMSRE values, even as contamination rises and group connectivity diverges, highlights their robustness under structurally heterogeneous conditions. These results indicate that such methods are well-suited for real-world network

settings with uneven contact probabilities between groups.

Figures 5.11(d)–(f) display the RMSRE results for the RoS-type estimators across contamination levels, again with prevalence fixed at $\rho = 5\%$ and connectivity factors set to (d) $a = e^{-4}$, (e) $a = 1$, and (f) $a = e^{4}$. Among the robust RoS variants, RoS-TM10 and RoS-Huber estimators consistently exhibit lower RMSRE values, demonstrating robustness to contamination across the different structural configurations. Finally, Figure 5.S.31 in the Supplementary Material depicts the RMSRE obtained by the robust NSUM approaches against the connectivity factor for different prevalences. In essence, this figure displays the response of robust NSUM methods to the barrier effect for SBM graphs.

5.4.3. Discussion

In this section, we provide an overview of the main findings regarding the performance of the NSUM estimation methods for different network models and contamination scenarios. First, the non-robust MoR and RoS estimators demonstrate the best performance in terms of RMSRE and NB metrics in environments free of contamination across the tested prevalence ranges for the three network models. This superior performance suggests that the non-robust MoR and RoS methods are the most effective approaches for estimating unknown group rates within networks with different characteristics in uncontaminated scenarios.

Among the MoR-type robust methods, the MoR-Myriad approach consistently exhibits superior performance across tested prevalence levels for the three network models when data is uncontaminated. Specifically, the accuracy generated by the MoR-Myriad method is comparable to that obtained by non-robust approaches, which is essential for practical applications where minimal contamination may occur. In addition, the MoR-HQ1, MoR-Huber, and MoR-TM10 methods improve performance as prevalence increases. However, these methods experience a significant decline in performance at low prevalence values, making them less suitable for estimating the size of small unknown groups. In robust RoS-type methods, the RoS-HQ1 estimator behaves similarly to non-robust approaches when prevalence exceeds 3%. This performance indicates a significant degree of reliability in scenarios without contamination but also highlights a potential limitation: RoS-type methods may not be as effective in networks with hard-to-reach unknown groups, especially in prevalence levels below 3%.

Non-robust methods severely degrade their performance in contaminated scenarios as the data noise level increases. Therefore, finding a robust alternative that enables accurate estimations when different kinds of perturbations corrupt data is necessary. For robust MoR-type methods, the MoR-Myriad approach consistently performs best for the three network models, different prevalence values, and contamination levels. Therefore, we can select the MoR-Myriad operator as a reliable option to

estimate the unknown group rate in contaminated scenarios. Furthermore, MoR-HQ1, MoR-Huber, and MoR-TM10 are also efficient methods compared to non-robust approaches with contamination, in particular, for prevalence values exceeding 5%. In the case of the RoS-type robust method, RoS-HQ1, RoS-TM10, and RoS-Huber perform best in contaminated scenarios for the ER model. However, these methods degrade their effectiveness for the SF model. Hence, in contaminated scenarios, RoS-type methods should consider both the network structure and the unknown group characteristics.

5.5. Case Study

This section evaluates the performance of the NSUM estimators in real scenarios. For that purpose, we use two datasets containing ARD from surveys about COVID-19 cases and voting intention in Spanish elections.

The dataset of COVID-19 exhibits skewed distributions and a type of contamination similar to our simulation study, with contamination at the extreme of the distribution. In contrast, the election dataset provides a particularly unfavorable scenario to validate our approach. Due to political polarization, extreme values–such as participants knowing only contacts with the same political affiliation–are more likely to occur.

This section is structured as follows. First, we introduce the COVID-19 dataset and present the results of the NSUM estimator. Next, we examine the robust NSUM through voting intention data. Finally, we conclude with a summary of key findings.

5.5.1. The United Kingdom COVID-19 dataset

We use the COVID-19 dataset of Ramirez et al. (2023), consisting of responses from an online survey in the United Kingdom during the period of January 18-26, 2023. The survey contains information about vaccinations, hospitalizations, fatalities, and known cases of acquaintances in several time slots. The main characteristic of this dataset is that the degree is fixed in advance to the 15 closest contacts, based on the good-friends support group of Dunbar's theory (Dunbar, 2010). Another characteristic of this data is the extreme skewness of the variables (Ramirez et al., 2023). We observe in Figure 5.12 that their values concentrate at one extreme of the distribution. Some of these variables, such as hospitalizations and contacts sick in the last month, which have a small prevalence in the population (see Table 5.2), show responses close to the degree of the nodes (15), suggesting malicious responses or an overrepresentation of people affected by the illness.

Table 5.2 displays the results obtained in the UK dataset. The table columns correspond to the percentage of monthly infection cases, monthly hospitalizations,

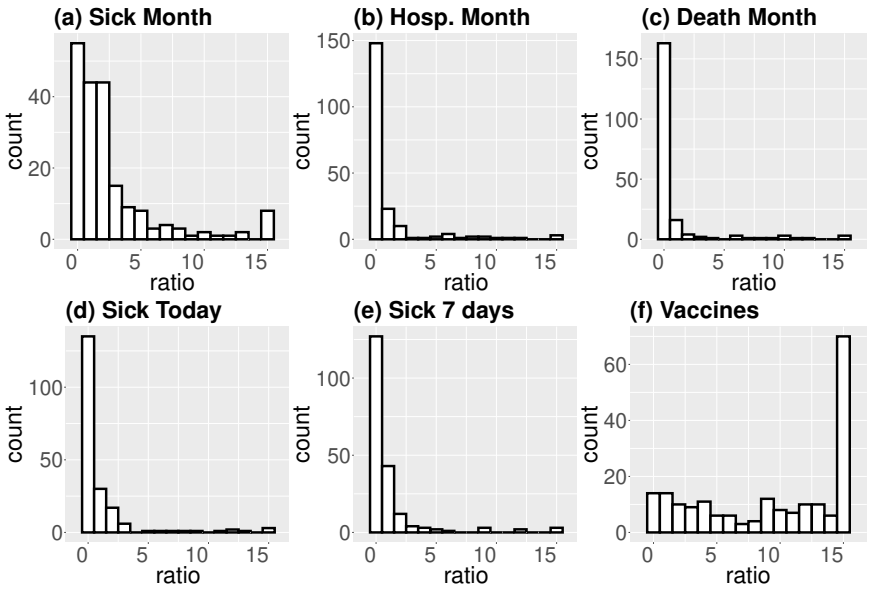


Figure 5.12: Histograms of ARD responses for (a) Sick Month, (b) Hospitalization Month, and (c) Death Month (d) Sick Today, (e) Sick 7 Days, and (c) Vaccines.

fatalities in a month, daily infections, weekly infections, and vaccinations. Specifically, the bold numbers in the table are the best performing results, and the underlined values are the second best results, with respect to the official data. The official data was obtained from the *Office for National Statistics* (ONS)[8]. Bold numbers represent the best method for each variable, and underlined values correspond to the second-best method.

The non-robust estimators, MoR and RoS, consistently overestimate small-sized groups while underestimating the larger group, vaccination. Overall, these estimators perform worse across all variables compared to the robust estimators, except for MoR-HQ2 and RoS-HQ2. MoR-Myriad is the best estimator when the size of the targeted group is not small, particularly in the case of vaccination, where it outperforms other methods. However, the performance of MoR-Myriad diminishes for smaller groups. In such cases, MoR-HSK2, RoS-HSK2, MoR-HSK5, RoS-HSK5, and RoS-Myriad yield the best results.

The adaptive trimmed means, with the exception of MoR-HQ2 and RoS-HQ2, are better than the classical estimators. While MoR-HQ1, RoS-HQ1, MoR-HQ2, and RoS-HQ2 do not perform as well as the non-adaptive MoR-TM10 and RoS-TM10; MoR-HSK2, RoS-HSK2, MoR-HSK5, and RoS-HSK5 offer clear improvements over MoR-TM10 and RoS-TM10 across all variables. Notably, MoR-HSK2, RoS-HSK2, MoR-HSK5, and RoS-HSK5 prove to be the best estimators in the low prevalence groups, such as weekly and daily infection rates, as well as monthly fatalities.

MoR-Median and RoS-Median yield relatively better results for vaccination

[8]https://www.ons.gov.uk/.

Table 5.2: Estimated percentages of COVID-19 incidence and mortality for UK.

	sick month	hosp. month	deaths month	sick today	sick 7d	vacc.
Official	9.663	0.044	0.005	1.458	1.116	88.2
MoR	17.367	6.767	5.800	7.000	6.767	62.433
RoS	17.367	6.767	5.800	7.000	6.767	62.433
MoR-Median	13.333	0.000	**0.000**	0.000	0.000	73.333
MoR-TM10	11.875	1.708	0.750	2.500	2.625	65.292
MoR-HQ1	14.161	3.478	2.153	3.395	3.602	67.246
MoR-HQ2	19.710	8.406	7.250	8.696	8.406	63.768
MoR-HSK2	8.654	0.500	**0.000**	1.042	**1.375**	70.683
MoR-HSK5	8.820	0.621	**0.000**	**1.159**	1.491	67.826
MoR-Huber	11.409	0.000	**0.000**	0.000	0.000	62.669
MoR-Myriad	**10.052**	2.158	1.399	2.862	3.086	**84.370**
RoS-Median	13.333	0.000	**0.000**	0.000	0.000	73.333
RoS-TM10	11.875	1.708	0.750	2.500	2.625	65.292
RoS-HQ1	14.161	3.478	2.153	3.395	3.602	67.246
RoS-HQ2	19.710	8.406	7.250	8.696	8.406	63.768
RoS-HSK2	8.654	0.500	**0.000**	1.042	**1.375**	70.683
RoS-HSK5	8.820	0.621	**0.000**	**1.159**	1.491	67.826
RoS-Huber	11.409	0.000	**0.000**	0.000	0.000	62.669
RoS-Myriad	6.698	**0.054**	0.033	0.081	0.108	99.900

prevalence compared to most other methods. However, when the prevalence decreases, these estimators return zero prevalence estimates. The same issue arises with MoR-Huber and RoS-Huber. In addition, we observe that MoR-type methods yield values identical to the corresponding RoS-type methods, which use the ratios of the location estimator, with the exception of MoR-Myriad and RoS-Myriad.

5.5.2. Spanish Elections Dataset

We employ a dataset of voting intention in Spanish general elections in July 2023 (Arevalillo et al., 2024). This dataset includes the ARD of the 4 main political groups and the ARD associated with the blank vote and other parties. These parties are *Partido Popular* (PP), *Partido Socialista Obrero Español* (PSOE), *Vox*, and *Sumar*. This case is an example containing barrier effects since the participants are more likely to form connections with people with the same political views. While participants were not directly asked for their degree, we derive this by summing their responses since the questions are disjoint.

Table 5.3 shows the estimated percentage for each of the four most popular political parties. Columns represent the political parties. The real data corresponds to the published official data.

Table 5.3: Estimated percentages of vote intention for the different parties in Spain.

	PP	PSOE	Vox	Sumar
Real	40.2	27.6	13.9	15.3
MoR	32.1	<u>27.0</u>	15.4	**15.2**
RoS	31.7	20.6	**14.4**	11.5
MoR-Median	25.0	25.0	8.3	8.7
MoR-TM10	28.9	24.9	11.3	11.3
MoR-HQ1	31.4	26.8	<u>13.0</u>	13.0
MoR-HQ2	<u>35.3</u>	29.3	17.2	<u>17.1</u>
MoR-HSK2	24.7	21.8	7.6	7.4
MoR-HSK5	25.9	22.1	8.1	7.7
MoR-Huber	30.2	25.8	11.5	11.4
MoR-Myriad	24.0	23.8	9.1	8.8
RoS-Median	27.8	**27.8**	5.6	11.1
RoS-TM10	24.4	23.1	9.6	9.8
RoS-HQ1	16.6	15.9	6.8	7.1
RoS-HQ2	**36.8**	22.3	16.8	12.6
RoS-HSK2	22.8	21.5	7.5	8.2
RoS-HSK5	22.8	22.4	7.5	8.3
RoS-Huber	26.2	25.9	7.7	10.3
RoS-Myriad	11.8	11.8	0.2	0.1

The results indicate that non-robust NSUM estimators generally outperform robust methods in predicting voting intentions. Specifically, MoR performs best across the four main parties, excelling in Sumar and ranking second for PSOE. Certain robust MoR-type estimators show strong performance for specific parties but fall short for others. For instance, MoR-HQ1 ranks second for Vox but diverges for Sumar, while MoR-HQ2 performs well for PP and Sumar but tends to overestimate for PSOE and Vox. Among RoS-type estimators, RoS generally surpasses its robust alternatives, except for RoS-HQ2. Notably, some robust MoR alternatives, like MoR-TM10, MoR-HQ1, and MoR-HQ2, outperform RoS in several cases. Additionally, we note that the voting intention for PP, shown in the first column, is consistently underestimated, with only RoS-HQ2 coming close to the real outcome. However, RoS-HQ2 is worse than MoR because of the deviations for other parties.

In conclusion, MoR emerges as the most effective estimator across all political parties, with no robust MoR-type method outperforming it. The closest alternative is MoR-HQ2, which provides the second-best predictions for two political parties. RoS demonstrates an inferior performance in comparison with MoR and the MoR robustifications. While RoS-HQ2 offers a performance improvement over RoS, especially being the only one good at PP, it does not surpass several robust MoR-type

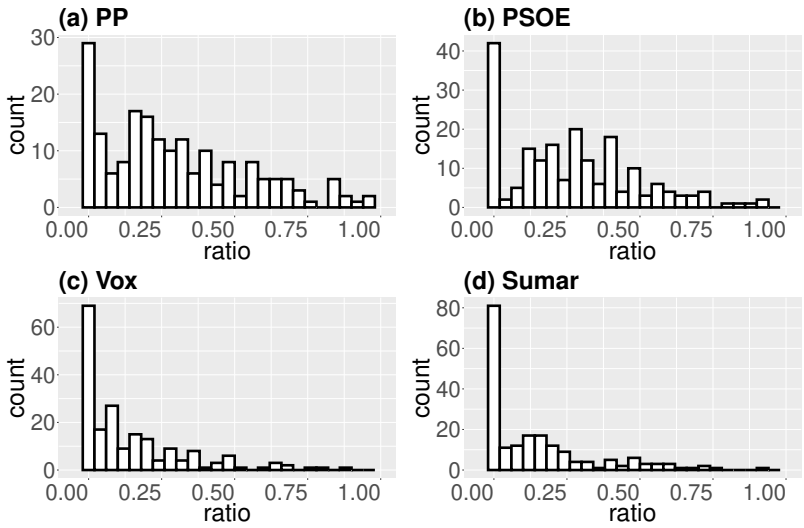


Figure 5.13: Histograms of ratios (a) PP, (b) PSOE, (c) Vox, and (d) Sumar.

estimators.

5.5.3. Discussion

The results of the COVID-19 dataset show that the robust estimators perform better in contaminated scenarios. This dataset has a marked presence of outliers (Ramirez et al., 2023), resulting in poor performance of MoR and RoS. The performance of the robust estimators is lower with respect to the simulation study. However, the results are not different if you compare them with each other, especially for groups with non-small prevalence, such as monthly, weekly, and daily infections and vaccinations. For small prevalence, the adaptive methods are better. However, one possible cause is the lack of precision of these methods since they estimate the prevalence as 0. Additionally, the performance of MoR-Median and RoS-Median provides relatively similar results. The MoR-Huber and RoS-Huber estimate all the groups as zero, indicating a possible more conservative behavior than the Myriad, focusing more on safety.

We observe that the values for MoR-type methods are equal to the corresponding RoS-type method, which uses the ratios of the location estimator, except for MoR-Myriad and RoS-Myriad. This consistency arises because the degree is fixed at 15, causing location estimators to return identical values when all input values are constant. In addition, the MoR-Myriad estimator performs well for groups without small prevalence, as in the simulations. However, in this case, MoR-Myriad suffers with small groups. Due to the small linearity parameter ($K = 0.2$), the Myriad operator approximates the mode (Arce, 2004), particularly in datasets with low variability. This may lead to a behavior similar to the RoS-Median, approximating to the closest integer. This may explain the bad performance of MoR-Myriad and

RoS-Myriad, as well as the discrepancy between their estimates.

The UK COVID-19 dataset shows that some adaptive estimators are resistant to deviations in skewness and contamination. Notably, MoR-HSK2 and MoR-HSK5 seem more adequate than the other adaptive estimators, having good performance compared to the previous simulations, presenting closer values while the simulations consistently underestimate. This contrasts with the fact that MoR-HQ1 performs poorly. The estimators MoR-HSK2 and MoR-HSK5 use the factors that involve the smallest and largest values, in addition to the mean and median. One possible factor is that the smallest and largest values are limited to 15. Another factor is that MoR-HQ1 and MoR-HQ2 use sums over a proportion of the smallest and largest. Therefore, they may consider more outlying values when calculating the trimming proportions.

The analysis of the voting intention dataset reveals that the non-robust estimators are better in the presence of the barrier effect, conclusions that align with our findings from the simulation experiments. Specifically, the non-robust MoR estimator consistently outperforms its corresponding robust MoR-type alternatives, while the robust RoS-type estimators generally perform worse than the non-robust RoS operator, except for RoS-HQ2. This observation corroborates our simulation results, where robust RoS-type methods, except RoS-HQ2, performed poorly under high assortativity and prevalence. Importantly, the results from real data suggest that the improvements seen in RoS-HQ2 during the simulations may become even more pronounced, potentially surpassing RoS in scenarios with higher prevalences. In this example, the variables have distributions with values on the extremes that are less likely to be produced by contamination. Identifying whether the extreme values are due to contamination is much harder. This may be the reason why the estimators underestimate all the groups.

Another feature in the Spanish elections is the outperforming of MoR with respect to RoS, which is consistent with our previous theoretical analysis. RoS assumes that every contact has the same possibility of knowing someone in the unknown group, but regarding political opinions, this rarely happens. In addition, the methods based on the ratios of the location estimators perform worse than the procedures based on the ratios of the responses. Indeed, MoR-HQ1 and MoR-HQ2 are better than RoS, and MoR-TM10 and MoR-Huber are very close to RoS.

Another relevant fact is that even MoR-Myriad yields bad estimates. In this dataset, people of one party are more likely to have all their acquaintances voting for the same political option. This produces distributions that differ from the SBM distributions of Figure 5.3. We observe in Figure 5.13 that, unlike high prevalences, there are numerous participants whose social circles do not include anyone voting for a particular party. Conversely, some social circles tend to vote exclusively for a single party. Our robust proposals are based on giving less weight to extreme

values. However, these extreme values are meaningful, and giving them less weight is counterproductive. The better performance of the RoS-HQ2 with respect to other robust estimators may be produced because it gives more weight to the extreme data, and this feature is the cause of the overestimation in the simulations, especially in the contaminated scenarios. MoR-HQ2 employs the same statistic for the trimming proportions, but in this case, the values of the ratios are bounded.

The real data suggest that, in the absence of a barrier effect, MoR-Myriad is the preferred estimator for groups with significant prevalence. However, MoR-Myriad struggles with small prevalence rates and networks of lower degrees, as the discrete nature of responses complicates accurate estimation. In these cases, alternatives such as MoR-HSK2, MoR-HSK5, RoS-HSK2, and RoS-HSK5 may be better options. However, due to the poor performance of these estimators in the simulation, it may be good to select other estimators, such as MoR-TM10, to safeguard against bad estimates. In the presence of barrier effects in groups with high prevalence, where participants may have all or none of their acquaintances belonging to the same group, leading to meaningful extremes of the distribution, impacting our proposals. In this case, the MoR is the preferred option, while the RoS and RoS-type robustifications suffer from bad performance since their assumptions require the absence of barrier effects. For a combination of barrier effect and contamination, the selected estimates would be a robustification of MoR such as MoR-TM10, MoR-Myriad, or MoR-Huber, balancing strong simulation results with reasonable results in real data.

5.6. Conclusions

This work focused on the area of robust statistics, analyzing the behavior of NSUM techniques in the face of outliers and unusual data. The purpose of this work was to study the reliability and accuracy of the methods, which is highly relevant, since real-world data often do not meet the ideal conditions assumed by many NSUM techniques. This analysis of statistical robustness in NSUM techniques has shown its significant impact, as illustrated with simulations. Real-world cases have been presented with COVID-19 and voting intention datasets.

This work proposes the first robustifications of the classical NSUM estimators most frequently employed in the literature. We propose robustifications substituting the sample means of the classical estimators with location estimators. To assess their effectiveness, we conduct a series of experiments in networks with real-world properties. Specifically, we employ Erdős–Rényi networks as a suitable model for classical NSUM, analyzing deviations varying prevalence levels. We include simulations in Scale-Free networks, providing insights into the effects of skewed distributions with heavy tails. We evaluate the performance of the estimators under contamination, simulating participant misreporting by replacing a fraction of unknown group ARD values with zeros or personal network sizes. We also measure the barrier effect through the

Stochastic Block Model, considering scenarios in which the members of the unknown group have different propensities to form ties with other groups. Finally, we validate our findings through a comparison of NSUM estimators on real datasets.

The main findings regarding the robust NSUM estimators across various scenarios are as follows. Firstly, non-robust estimators consistently outperform robust approaches across all network models and tested prevalence levels in uncontaminated environments. These results indicate that non-robust estimators are particularly effective in scenarios without contamination, delivering optimal performance for different network structures. In contrast, in networks with noisy data, the MoR-Myriad method stands out by outperforming other estimation methods across various prevalence values and contamination levels. Its robustness makes the MoR-Myriad method especially valuable in situations where external noise may distort data, particularly at low prevalence levels. Additionally, the techniques MoR-TM10, MoR-Huber, RoS-Huber, and RoS-TM10 show improved performance as prevalence levels increase, both in contaminated and uncontaminated environments. This trend highlights the potential limitations of these methods when estimating the sizes of hard-to-reach groups. Finally, median-based methods such as MoR-Median and RoS-Median demonstrate instability, especially at low prevalence levels.

This work has some limitations. Firstly, conclusions are based primarily on simulation environments, and although we have validated experiments with two datasets, these may not entirely capture the complexity of real-world distributions, biases, and contamination. Another limitation is the use of undirected graphs, which assume reciprocal relationships, potentially oversimplifying the asymmetric nature of some real-world social networks. Moreover, while our results show that robust estimators perform consistently well in various scenarios, we recognize that using fixed parameter values may limit their full capabilities. In this study, we utilized standard settings for key estimators (Myriad, Huber, and trimmed mean), which might not be optimal for the specific types of contamination we examined. For example, although there are parameter selection strategies available for α-stable noise models (Gonzalez and Arce, 2001), our research focuses on a different type of impulsive contamination. Therefore, a systematic investigation into the effects of these hyperparameters and the potential for tuning them to enhance performance remains an important area for future research.

To the best of our knowledge, this is the first study to introduce robust NSUM estimators and provide a comprehensive comparative evaluation of their performance. The first further work is to explore robust versions of traditional NSUM applications where participants' degrees are unknown, requiring estimation from responses related to groups with known prevalence. Another future alternative consists of the analysis of the contamination of the responses of the reported network size. Further studies could focus on additional contamination types, such as contamination in one extreme of the distributions, in a neighborhood of the data, or representing the transmission

and recall error. Given the impact of prevalence, network characteristics, and contamination on estimator choice, further development is needed to tailor robust estimators to these specific features. Another area of further work is the proposal of methods that give more weight to extreme values in scenarios with a high barrier effect. Since many NSUM studies focus on low-prevalence groups, future research on estimators resistant to very low prevalence rates is also warranted. Finally, a subsequent line of research is the analysis of the effect of the sample size and the robustness of the procedures designed for selecting it (Josephs et al., 2024).

Declaration of interest statement

The authors have no conflicts of interest to disclose.

Code availability

The source code used for the simulations can be downloaded from the following link: https://github.com/JuanMarcosRamirez/Robust_NSUM_IMDEA.

Acknowledgments

This work has been supported by Ministerio de Ciencia e Innovación, Gobierno de España, grant numbers PDC2022-133359, PID2022-137243OB-100 and TED2021-131264B-100 funded by MCIN/AEI/10.13039/501100011033, European Union NextGenerationEU/PRTR. This paper has been funded by the project PID2022-140560OB-I00 (DRONAC) funded by MICIU/AEI /10.13039/501100011033 and ERDF, EU. This initiative has also been partially carried out within the framework of Recovery, Transformation and Resilience Plan funds, financed by the European Union (Next Generation) through the grant ANTICIPA. The work of Sergio Díaz-Aranda has been funded by Comunidad de Madrid predoctoral grant PIPF-2022/COM-24467.

References

C. Agostinelli, A. Basu, P. Filzmoser, D. Mukherjee, et al. Recent advances in robust statistics: theory and applications. Springer, 2016.

G. R. Arce. Nonlinear signal processing: a statistical approach. John Wiley & Sons, 2004.

J. M. Arevalillo, J. M. Ramírez, S. Díaz-Aranda, J. Aguilar, A. Fernández Anta, and R. E. Lillo. Network scale-up methods on aggregated relational data to estimate the outcome of elections. (Submitted article), 2024.

S. Cheng, D. J. Eck, and F. W. Crawford. Estimating the size of a hidden finite set: large-sample behavior of estimators, 2019. URL https://arxiv.org/abs/1808.04753.

S. Díaz-Aranda, J. Aguilar, J. M. Ramírez, D. Rabanedo, A. Fernńdez-Anta, and R. Lillo. Performance analysis of nsum estimators in social-network topologies. The American Statistician, 0(0):1–18, 2024. doi: 10.1080/00031305.2024.2421361.

R. Dunbar. How many friends does one person need? Dunbar's number and other evolutionary quirks. Harvard University Press, 2010.

D. M. Feehan and M. J. Salganik. Generalizing the network scale-up method: a new estimator for the size of hidden populations. Sociological methodology, 46(1): 153–186, 2016.

A. Garcia-Agundez, O. Ojo, H. A. Hernández-Roig, C. Baquero, D. Frey, C. Georgiou, M. Goessens, R. E. Lillo, R. Menezes, N. Nicolaou, et al. Estimating the covid-19 prevalence in spain with indirect reporting via open surveys. Frontiers in Public Health, 9:658544, 2021.

K.-I. Goh, B. Kahng, and D. Kim. Universal behavior of load distribution in scale-free networks. Physical review letters, 87(27):278701, 2001.

J. Gonzalez and G. Arce. Optimality of the myriad filter in practical impulsive-noise environments. IEEE Transactions on Signal Processing, 49(2):438–441, 2001. doi: 10.1109/78.902126.

P. Habecker, K. Dombrowski, and B. Khan. Improving the network scale-up estimator: Incorporating means of sums, recursive back estimation, and sampling weights. PloS one, 10(12):e0143406, 2015.

F. Hampel, E. Ronchetti, P. Rousseeuw, and W. Stahel. Robust Statistics: The Approach Based on Influence Functions. Wiley Series in Probability and Statistics. Wiley, 2011. ISBN 9781118150689. URL https://books.google.es/books?id=XK3uhrVefXQC.

R. V. Hogg. Adaptive robust procedures: A partial review and some suggestions for future applications and theory. Journal of the American Statistical Association, 69(348):909–923, 1974.

P. J. Huber and E. M. Ronchetti. Robust statistics, 2009.

L. Jing, Q. Lu, Y. Cui, H. Yu, and T. Wang. Combining the randomized response technique and the network scale-up method to estimate the female sex worker population size: an exploratory study. Public Health, 160:81–86, 2018.

N. Josephs, D. M. Feehan, and F. W. Crawford. A sample size formula for network scale-up studies. Sociological Methods & Research, 53(3):1252–1289, 2024.

H. Keselman, R. R. Wilcox, L. M. Lix, J. Algina, and K. Fradette. Adaptive robust estimation and testing. British Journal of Mathematical and Statistical Psychology, 60(2):267–293, 2007.

P. D. Killworth, E. C. Johnsen, C. McCarty, G. A. Shelley, and H. R. Bernard. A social network approach to estimating seroprevalence in the united states. Social networks, 20(1):23–50, 1998a.

P. D. Killworth, C. McCarty, H. R. Bernard, G. A. Shelley, and E. C. Johnsen. Estimation of seroprevalence, rape, and homelessness in the united states using a social network approach. Evaluation review, 22(2):289–308, 1998b.

J. P. Kunke, I. Laga, X. Niu, and T. H. McCormick. Comparing the robustness of simple network scale-up method (nsum) estimators, 2024. URL https://arxiv.org/abs/2303.07490.

I. Laga, L. Bao, and X. Niu. Thirty years of the network scale-up method. Journal of the American Statistical Association, 116(535):1548–1559, 2021.

I. Laga, L. Bao, and X. Niu. A correlated network scale-up model: Finding the connection between subpopulations. Journal of the American Statistical Association, pages 1–18, 2023a.

I. Laga, J. Kunke, T. McCormick, and X. Niu. The role of scaling and estimating the degree ratio in the network scale-up method. arXiv preprint arXiv:2305.04381, 2023b.

R. Maltiel, A. E. Raftery, T. H. McCormick, and A. J. Baraff. Estimating population size using the network scale up method. The annals of applied statistics, 9(3): 1247, 2015.

R. A. Maronna, R. D. Martin, V. J. Yohai, and M. Salibián-Barrera. Robust statistics: theory and methods (with R). John Wiley & Sons, 2019.

T. H. McCormick. The network scale-up method. The Oxford Handbook of Social Networks, page 153, 2020.

M. Newman. Networks. Oxford university press, 2018.

A. Nyarko-Agyei, S. Moser, R. G. Seymour, B. Brewster, S. Li, E. Weir, T. Landman, E. Wyman, C. B. Torres, I. Fell, et al. A partially pooled nsum model: Detailed estimation of csem trafficking prevalence in philippine municipalities. arXiv preprint arXiv:2407.13267, 2024.

H. Ocagli, D. Azzolina, G. Lorenzoni, S. Gallipoli, M. Martinato, A. S. Acar, P. Berchialla, D. Gregori, and I. S. Group. Using social networks to estimate the number of covid-19 cases: The incident (hidden covid-19 cases network estimation) study protocol. International Journal of Environmental Research and Public Health, 18(11):5713, 2021.

J. M. Ramirez and J. L. Paredes. Recursive weighted myriad based filters and their optimizations. IEEE Transactions on Signal Processing, 64(15):4027–4039, 2016.

J. M. Ramirez and J. L. Paredes. Recursive myriad-mean filters: Adaptive algorithms and applications. Signal Processing, 139:12–24, 2017.

J. M. Ramirez, S. Diaz-Aranda, J. Aguilar, O. Ojo, R. E. Lillo, and A. Fernandez Anta. A Snapshot of COVID-19 Incidence, Hospitalizations, and Mortality from Indirect Survey Data in China in January 2023. medRxiv, pages 2023–02, 2023.

J. Reed III and D. Stark. Hinge estimators of location: Robust to asymmetry. Computer Methods and Programs in Biomedicine, 49(1):11–17, 1996.

R. G. Staudte and S. J. Sheather. Robust estimation and testing. John Wiley & Sons, 2011.

E. Sully, M. Giorgio, and S. Anjur-Dietrich. Estimating abortion incidence using the network scale-up method. Demographic Research, 43:1651–1684, 2020.

R. R. Wilcox. Introduction to robust estimation and hypothesis testing. Academic press, 2011.

T. Zheng, M. J. Salganik, and A. Gelman. How many people do you know in prison? using overdispersion in count data to estimate social structure in networks. Journal of the American Statistical Association, 101(474):409–423, 2006.

Supplementary Material

5.S.A. Formal definition of robustness concepts.

There are several approaches to robust statistics (Staudte and Sheather, 2011). In this work, we follow the descriptive measure approach. That is, we consider measures as a parameter associated with each distribution. The idea is choosing a descriptive measure with good performance in the assumed conditions, stability under small deviations, and avoiding catastrophic outcomes under high deviations, as Huber and Ronchetti (2009) proposed. To understand the properties of the measures and their relationships to estimators, it is useful to include the functional notation. A descriptive measure can be seen as a map from $\mathcal{M}$ to $\mathbb{R}$, where $\mathcal{M}$ is the set of all probability distributions on the sample space. The map is associated with an estimator $T(x_1, \ldots, x_n) = T(F_n)$, where $x_1, \ldots, x_n$ are samples of the distribution F and $F_n = \frac{1}{n}\sum_i \delta_{x_i}$ is the empirical measure. Classical textbooks such as Huber and Ronchetti (2009) and Hampel et al. (2011) follow a similar notation, but they consider parametric families, and the descriptive measure corresponds to the parameter of the family.

There is extensive theoretical research regarding the way to measure robustness (Huber and Ronchetti, 2009). In our approach, the robustness properties of the estimator are the properties of the descriptive measure. One concept to assess stability is the *quantitative robustness*. It consists of the analysis of the performance in a contaminated neighborhood. The contamination of a distribution F with a distribution H is modeled by the mixture distribution of F with probability $1-\epsilon$ and H with probability ϵ, denoted as $(1-\epsilon)F + \epsilon H$. A common example of such contamination is the contamination with a single value, which we define as $F_{x,\epsilon} = (1-\epsilon)F + \epsilon\delta_x$, where δ_x is the point-mass distribution with support in x.

An important concept of quantitative robustness is the *influence function*, defined as

$$IF(x) = \lim_{\epsilon \to 0} \frac{T(F_{x,\epsilon}) - T(F)}{\epsilon}. \tag{5.S.1}$$

This function assesses the impact of a single outlier when the fraction of contamination tends to zero. Many references, such as Huber and Ronchetti (2009) and Hampel et al. (2011), argue that the estimators should have a bounded influence function. Many concepts of robust statistics are based on the influence function. A reference for this approach based on influence functions is given in Hampel et al. (2011).

One of the concepts related to the resistance to high deviations is the *breakdown point*, the minimum fraction of contamination, ϵ, such that a single outlier can produce an arbitrary value. The breakdown point is formally defined in our approach as

$$\min\{\epsilon : \sup_x |T(F_{x,\epsilon})| = \infty\}. \tag{5.S.2}$$

This is equivalent to the effect of replacing a fraction ϵ of a finite sample with a value x. This measure is widely used to protect against deviations (see (Huber and Ronchetti, 2009)), as a low breakdown point means that the estimator can assume extreme values, thereby reducing researchers' confidence in the results.

5.S.B. Location measures and trimmed means

There are two definitions of trimmed means. The definition of trimmed means we use, although it is classical, does not correspond with the natural estimator of the location measure. This section relates the classical concept to the location measures. Additionally, we derive the properties for the asymmetric trimmed means.

The definition of trimmed mean employed in our study is

$$TM_\alpha = \frac{1}{n-2g}\sum_{i=g+1}^{n-g} X_{(i)}, \tag{5.S.3}$$

where $g = \lfloor n\alpha \rfloor$ is the largest integer not exceeding $n\alpha$. The trimmed mean defined by Huber and Ronchetti (2009) is the natural estimator of the location measure

$$\mu_\alpha(F) = \frac{1}{1-2\alpha}\int_\alpha^{1-\alpha} F^{-1}(s)ds = \frac{1}{1-2\alpha}\int_{x_\alpha}^{x_{1-\alpha}} xdF(x). \tag{5.S.4}$$

The trimmed mean is defined as

$$\mu_{\alpha,n} = \frac{1}{n-2n\alpha}\left(\sum_{i=g+1}^{n-g} X_{(i)} + (g-\alpha n)(X_{(g)} + X_{(n-g+1)})\right). \tag{5.S.5}$$

We observe that this trimmed mean converges to the classical trimmed mean when the sample size increases. Therefore, $TM_\alpha(x_1, \cdots, x_n)$ converges to $lim_n \mu_{\alpha,n}(x_1, \cdots, x_n) = \mu_\alpha(F)$. This implies that the robustness properties, such as the breakdown point and the influence curve, are the same.

The case of asymmetric trimmed means is equivalent. The (α_1, α_2)-trimmed mean is the estimator

$$TM_{(\alpha_1,\alpha_2)} = \frac{1}{n-g_1-g_2}\sum_{i=g_1+1}^{n-g_2} X_{(i)}. \tag{5.S.6}$$

where $g_1 = \lfloor n\alpha_1 \rfloor$ and $g_2 = \lfloor n\alpha_2 \rfloor$. Similarly, the trimmed mean of Huber and Ronchetti (2009) is associated with the measure

$$\mu_{(\gamma_1,\gamma_2)}(F) = \frac{1}{1-(\gamma_1+\gamma_2)}\int_{x_{\gamma_1}}^{x_{1-\gamma_2}} xdF(x). \tag{5.S.7}$$

The expression of the alternative (α_1, α_2)-trimmed mean is (Huber and Ronchetti, 2009)

$$\mu_{(\gamma_1,\gamma_2),n} = \frac{\sum_{i=g_1+1}^{n-g_2} X_{(i)} + (g_1-\gamma_1 n)X_{(g_1)} + (g_2-\gamma_2 n)X_{(n-g_2+1)}}{n-n(\gamma_1+\gamma_2)}. \tag{5.S.8}$$

Similarly to the previous case, the estimators converge to the same values, and hence, to the same location measure. The properties of the asymmetric trimmed means are not well-known. However, they can be proven analogously to the results of (Huber and Ronchetti, 2009) for symmetric trimmed means. We add additional results for the qualitative robustness of the asymmetric trimmed means. This indicates that a robust issue of the adaptive trimmed means should be produced by a non-robust selection of the trimming proportions.

Lemma 5.1. *The influence function of the* (α_1, α_2)*-trimmed mean is*

$$IF(x) = \begin{cases} \dfrac{F^{-1}(\alpha_1) - \bar{\mu}_W}{1 - (\alpha_1 + \alpha_2)} & \text{if } x < F^{-1}(\alpha_1), \\ \dfrac{x - \bar{\mu}_W}{1 - (\alpha_1 + \alpha_2)} & \text{if } F^{-1}(\alpha_1) < x < F^{-1}(1 - \alpha_2), \\ \dfrac{F^{-1}(1 - \alpha_2) - \bar{\mu}_W}{1 - (\alpha_1 + \alpha_2)} & \text{if } F^{-1}(1 - \alpha_2) < x, \end{cases} \tag{5.S.9}$$

where $\bar{\mu}_W$ *is the non-symmetric Winsorized mean*

$$\bar{\mu}_W = \int_{x_\alpha}^{x_{1-\alpha_1}} x dF(x) + \alpha_1 x_{\alpha_1} + \alpha_2 x_{1-\alpha_2}, \tag{5.S.10}$$

where x_{α_i} *is the* α_i *quantile.*

Proof. This is proven using the fact that trimmed means are a type of L-estimators, estimators that are a linear combination of order statistics (Huber and Ronchetti, 2009)

$$T_n = \sum a_{n,i} X_{(i)}, \tag{5.S.11}$$

where $X_{(i)}$ is the ith order statistics, and the coefficients $a_{n,i}$ are generated by a signed measure M on $(0, 1)$. Huber and Ronchetti (2009) show that the L-estimators are the natural estimators of the operator

$$T(F) = \int (F^{-1}(s) M(ds), \tag{5.S.12}$$

where $F^{-1}(s) = \inf\{x : F(x) \geq s\}$ is the sth quantile, x_s. If M has a density m, the previous equation is equivalent to

$$T(F) = \int m(F(y)) dy. \tag{5.S.13}$$

In the case of the (α_1, α_2)-trimmed mean, the measure m is

$$m(s) = \begin{cases} (1 - 2\alpha)^{-1} & \text{if } \alpha_1 < s < 1 - \alpha_2 \\ 0 & \text{otherwise.} \end{cases} \tag{5.S.14}$$

The influence function of the L-estimates is (Huber and Ronchetti, 2009)

$$IF(x) = \int_{-\infty}^{x} m(F(y)) dy - \int_{-\infty}^{\infty} (1 - F(y)) m(F(y)) dy. \tag{5.S.15}$$

The result of the influence function for the (α_1, α_2)-trimmed mean follows from the fact that the second term of the right-hand side of Eq. (5.S.15) is

$$-\int_{-\infty}^{\infty}(1-F(y))m(y)dy = \frac{F^{-1}(\alpha_1)-\bar{\mu}_W}{1-(\alpha_1+\alpha_2)}. \tag{5.S.16}$$

This is because

$$\int_{-\infty}^{\infty} m(F(y)) = \frac{F^{-1}(1-\alpha_2)-F^{-1}(\alpha_1)}{1-(\alpha_1+\alpha_2)}, \tag{5.S.17}$$

and

$$\begin{aligned}
\int_{-\infty}^{\infty} F(y)m(F(y)) &= \frac{1}{1-(\alpha_1+\alpha_2)}\int_{F^{-1}(\alpha_1)}^{F^{-1}(1-\alpha_2)} F(y)dy \\
&= \frac{1}{1-(\alpha_1+\alpha_2)}[yF(y)]_{F^{-1}(\alpha_1)}^{F^{-1}(\alpha_2)} - \int yf(y)dy \\
&= \frac{1}{1-(\alpha_1+\alpha_2)}\left[F^{-1}(1-\alpha_2)(1-\alpha_2) - F^{-1}(\alpha_1)\alpha_1 - \int_{\alpha_1}^{1-\alpha_2} F^{-1}(u)du\right]
\end{aligned}$$

□

Huber and Ronchetti (2009) show that the breakdown point for the L-measure of location is the largest number α such that $[\alpha, 1-\alpha]$ is contained in the support of M. Therefore, the breakdown point for the (α_1, α_2)-trimmed mean is $\min(\alpha_1, \alpha_2)$.

5.S.C. Erdős–Rényi Network Model

5.S.C.1. Without Contamination

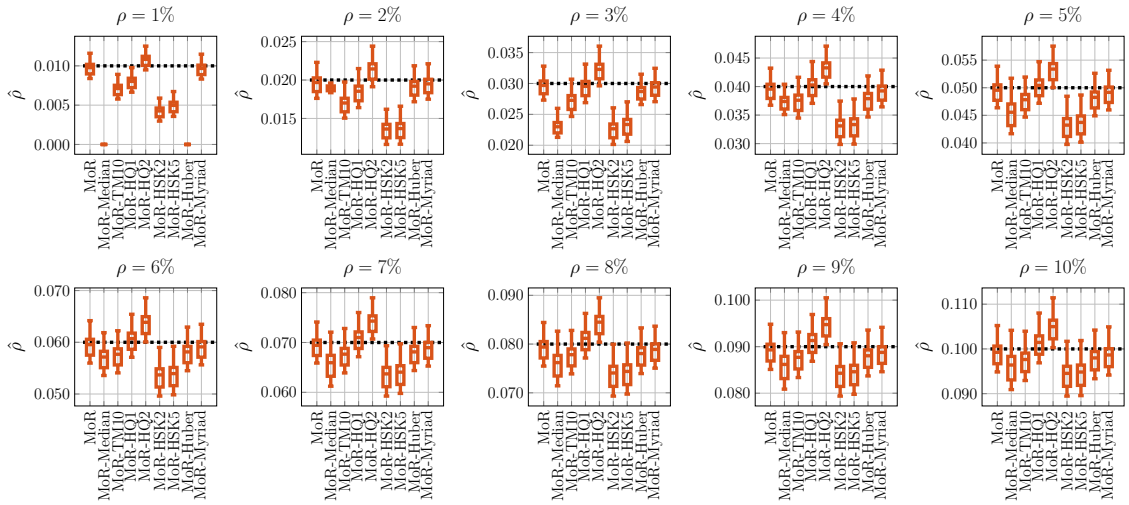


Figure 5.S.1: Erdős–Rényi ($\mu = 50$). Boxplots yielded by robust MoR-type methods in scenarios without contamination for different prevalence levels.

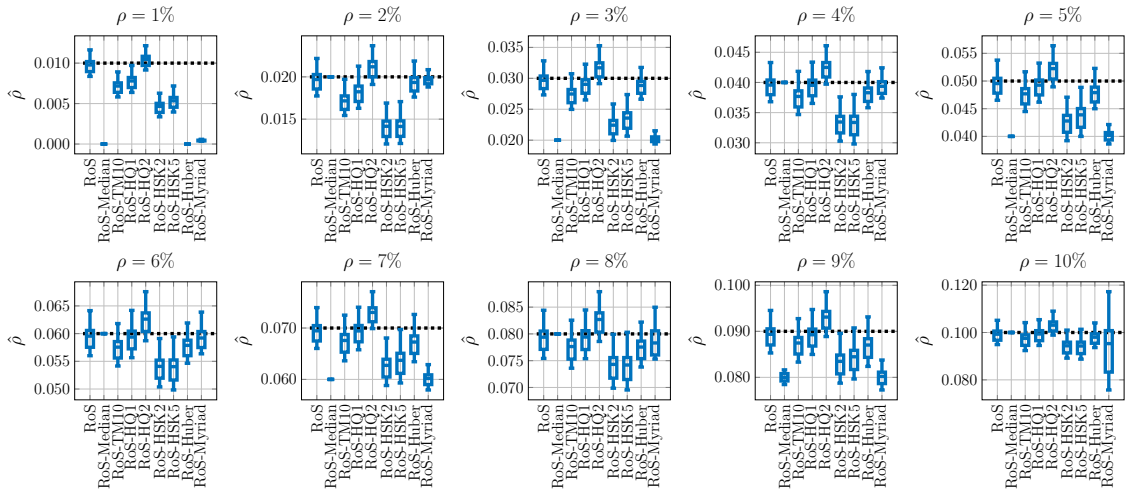


Figure 5.S.2: Erdős–Rényi ($\mu = 50$). Boxplots yielded by robust RoS-type methods in scenarios without contamination for different prevalence levels.

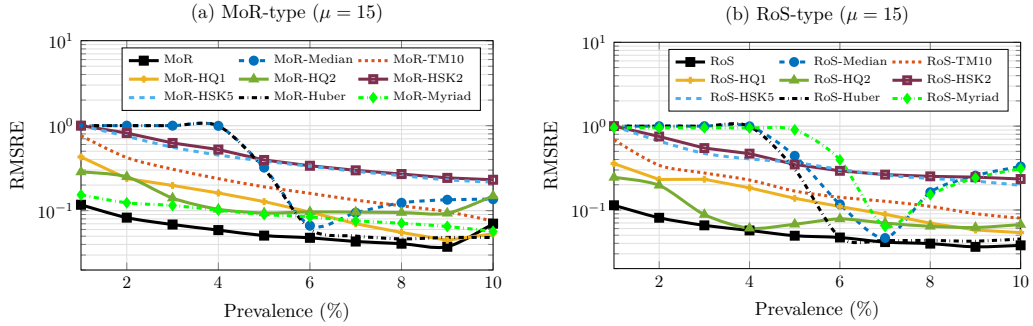


Figure 5.S.3: Erdős–Rényi ($\mu = 15$). RMSRE versus the prevalence without contamination for robust estimators: (a) MoR-type and (b) RoS-type.

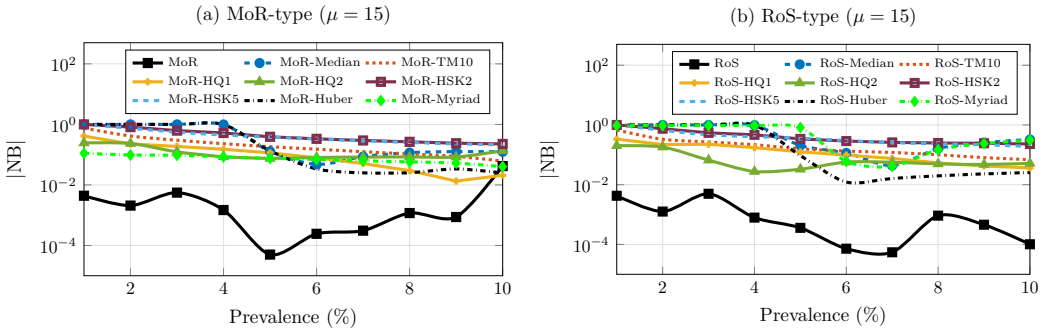


Figure 5.S.4: Erdős–Rényi ($\mu = 15$). The absolute value of the NB versus the prevalence without contamination for the robust estimators: (a) MoR-type and (b) RoS-type.

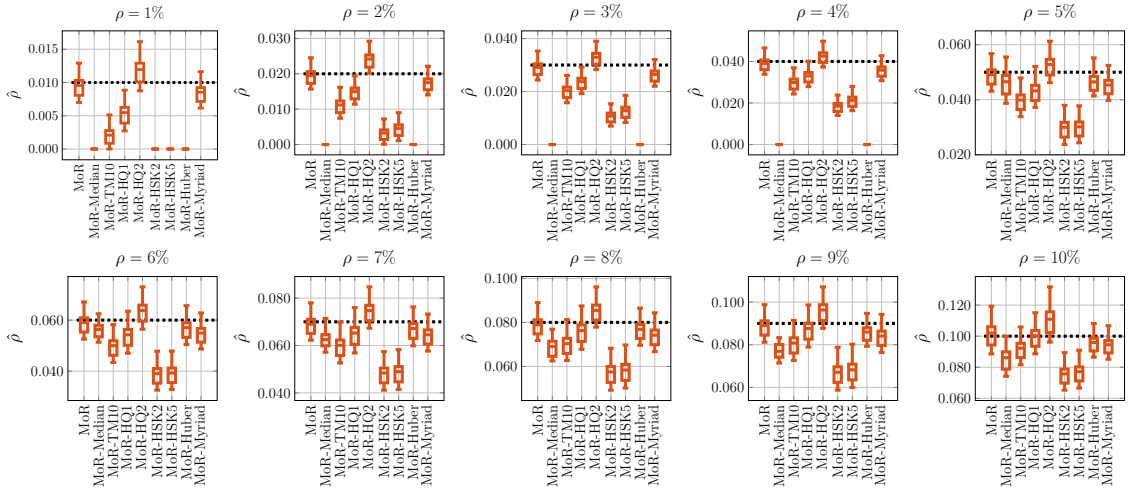


Figure 5.S.5: Erdős–Rényi ($\mu = 15$). Boxplots yielded by robust MoR-type methods in scenarios without contamination for different prevalence levels.

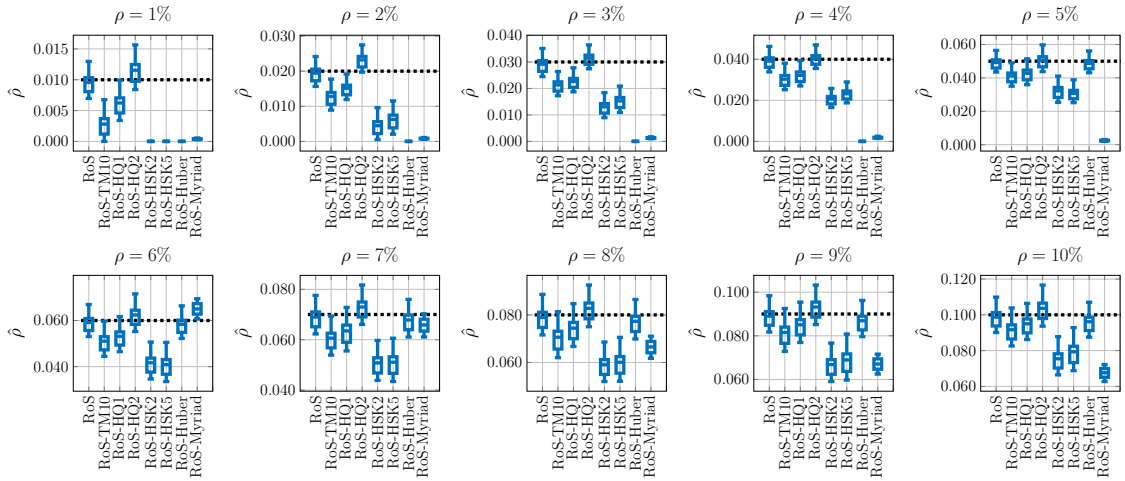


Figure 5.S.6: Erdős–Rényi ($\mu = 15$). Boxplots yielded by robust RoS-type methods in scenarios without contamination for different prevalence levels.

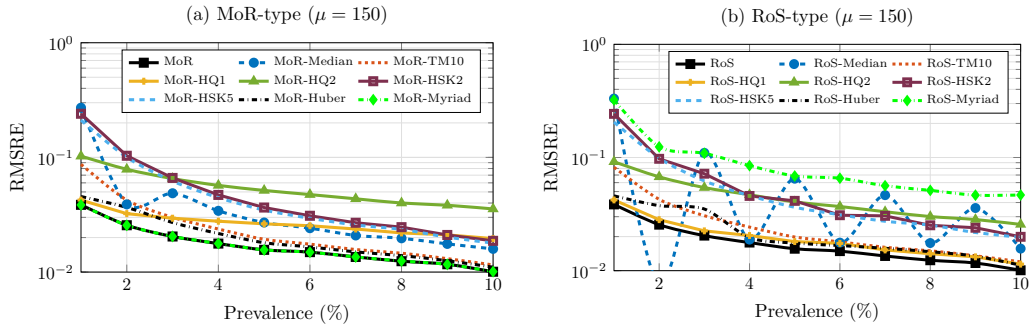


Figure 5.S.7: Erdős–Rényi ($\mu = 150$). RMSRE versus the prevalence without contamination for the robust estimators: (a) MoR-type and (b) RoS-type.

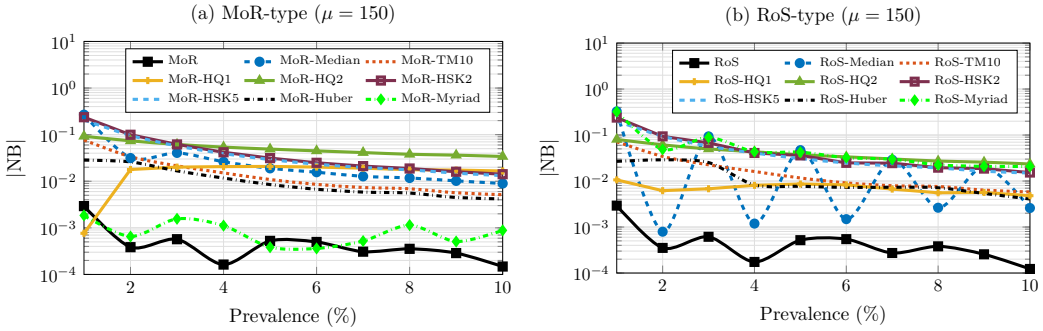


Figure 5.S.8: Erdős–Rényi ($\mu = 150$). The absolute value of the NB across various prevalences without contamination for the robust estimators: (a) MoR-type and (b) RoS-type.

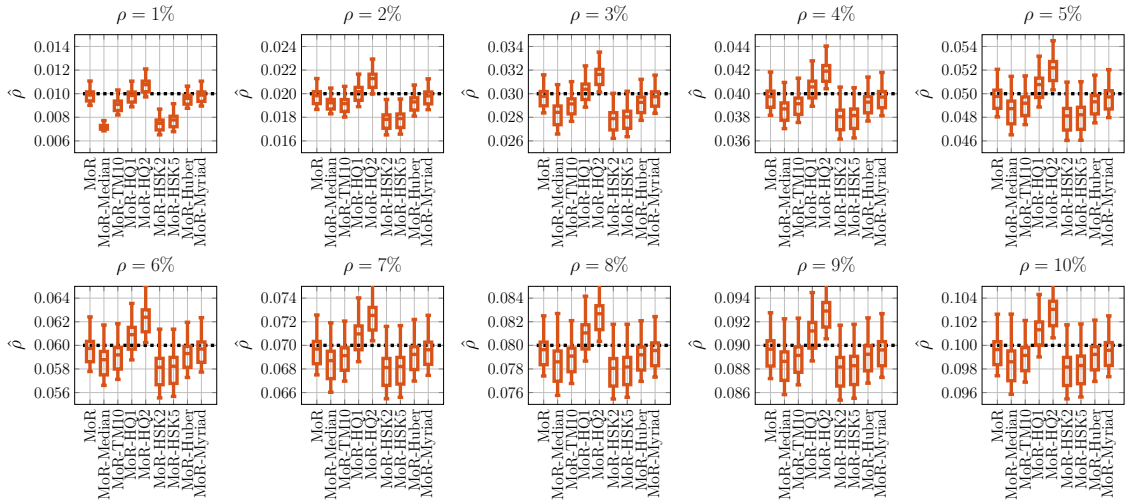


Figure 5.S.9: Erdős–Rényi ($\mu = 150$). Boxplots yielded by robust MoR-type methods in scenarios without contamination for different prevalence levels.

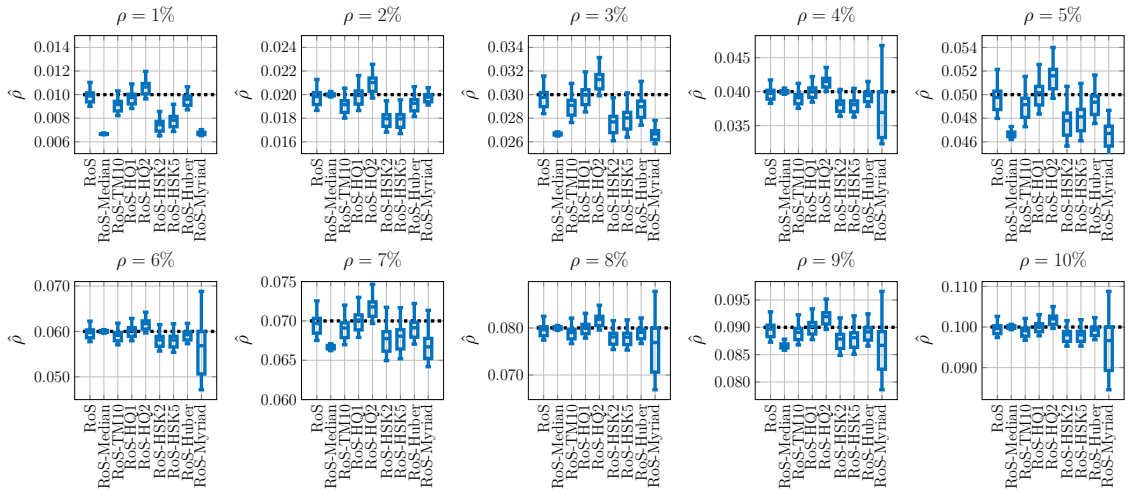


Figure 5.S.10: Erdős–Rényi ($\mu = 150$). Boxplots yielded by robust RoS-type methods in scenarios without contamination for different prevalence levels.

5.S.C.2. With Contamination

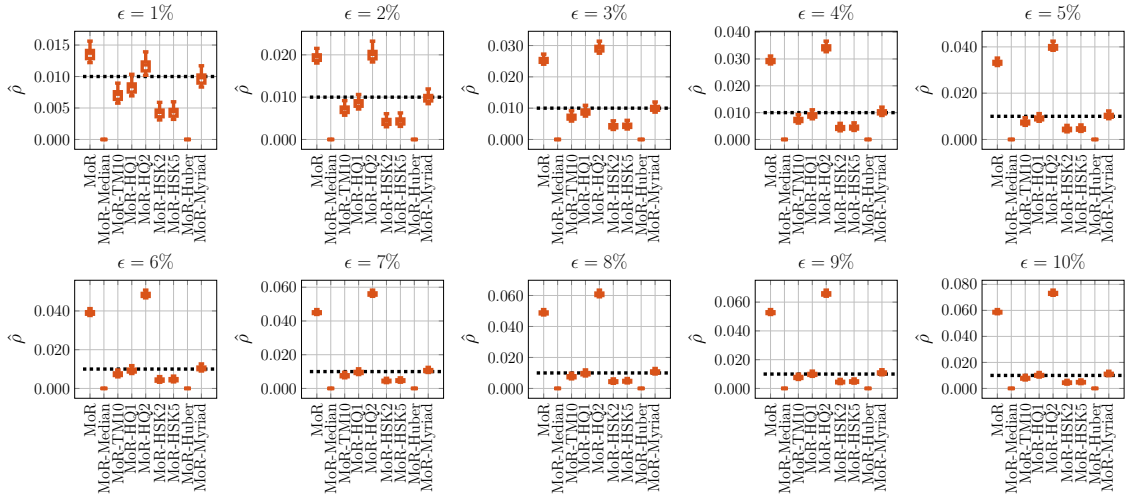


Figure 5.S.11: Erdős–Rényi ($\mu = 50$, $\rho = 1\%$). Boxplots yielded by robust MoR-type methods for different contamination levels ϵ.

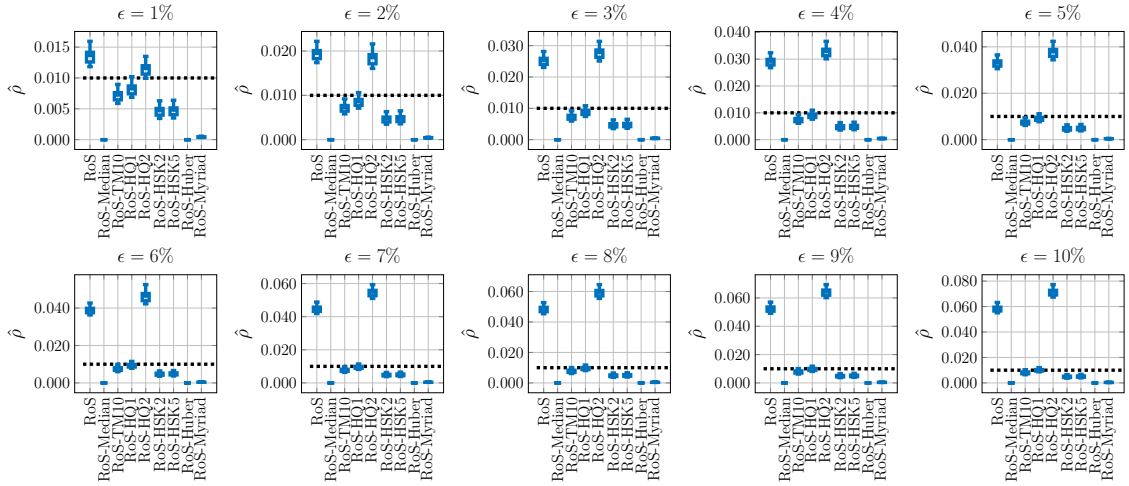


Figure 5.S.12: Erdős–Rényi ($\mu = 50$, $\rho = 1\%$). Boxplots yielded by robust RoS-type methods for different prevalence levels ϵ.

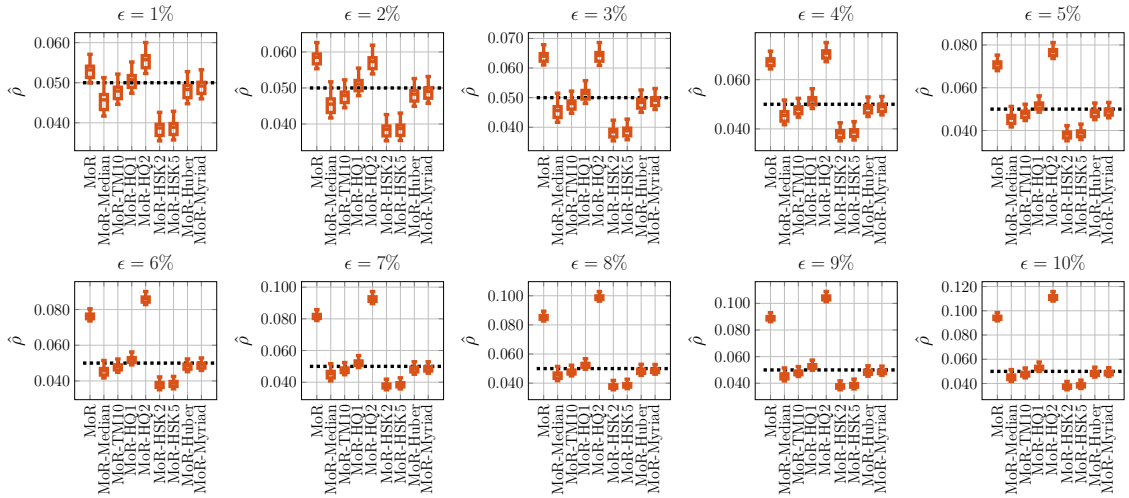


Figure 5.S.13: Erdős–Rényi ($\mu = 50$, $\rho = 5\%$). Boxplots yielded by robust MoR-type methods for different contamination levels ϵ.

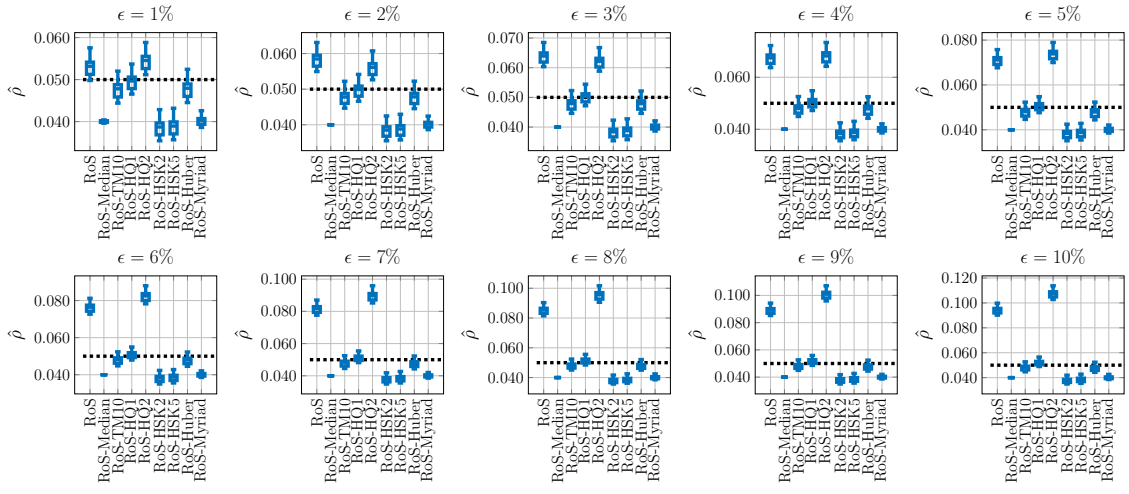


Figure 5.S.14: Erdős–Rényi ($\mu = 50$, $\rho = 5\%$). Boxplots yielded by robust RoS-type methods for different prevalence levels ϵ.

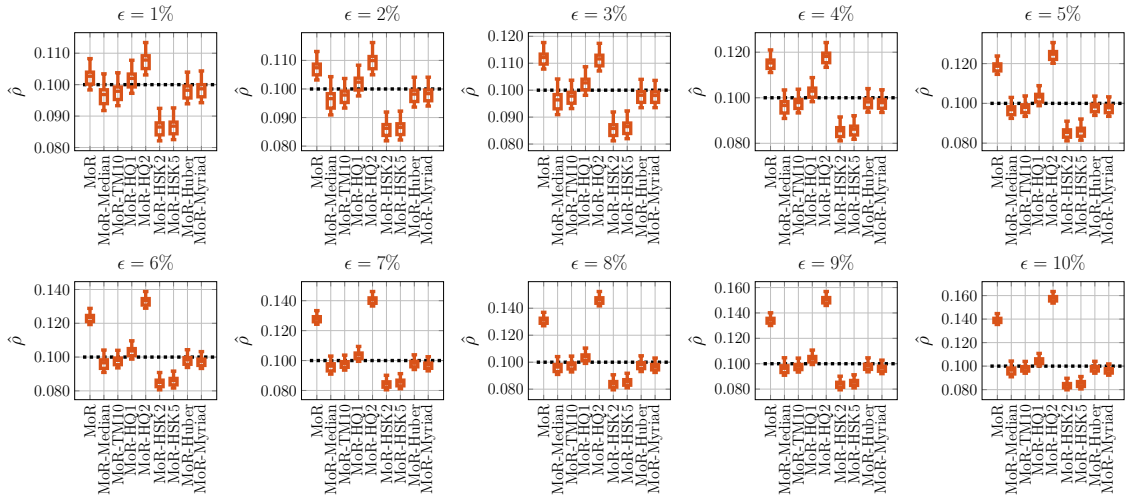


Figure 5.S.15: Erdős–Rényi ($\mu = 50$, $\rho = 10\%$). Boxplots yielded by robust MoR-type methods for different contamination levels ϵ.

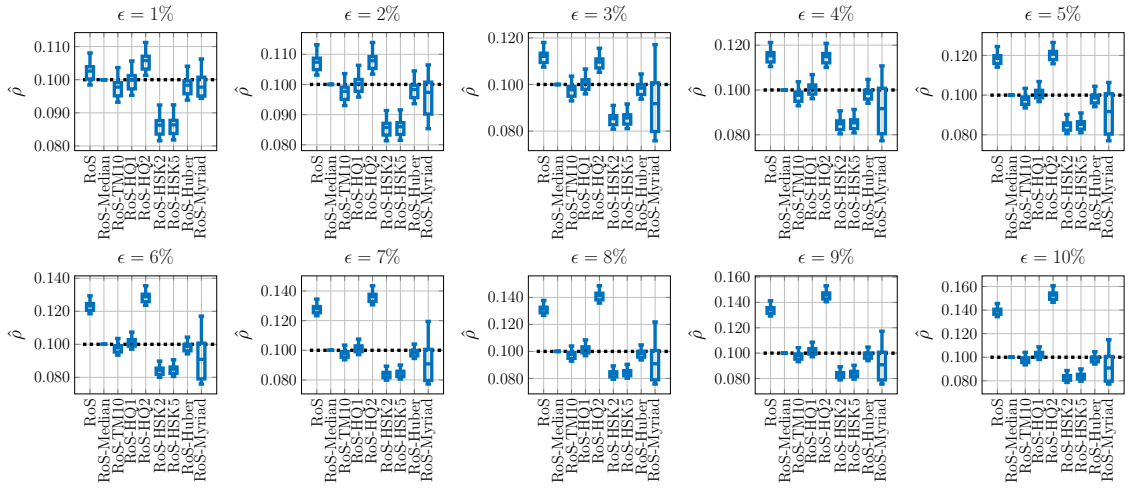


Figure 5.S.16: Erdős–Rényi ($\mu = 50$, $\rho = 10\%$). Boxplots yielded by robust RoS-type methods for different prevalence levels ϵ.

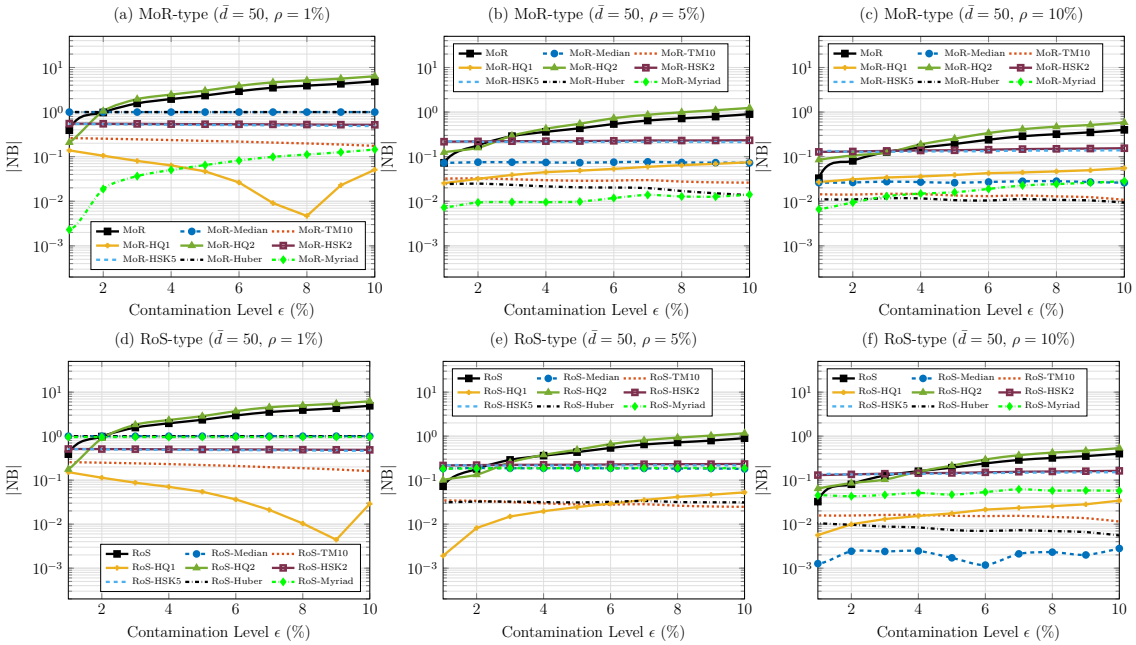


Figure 5.S.17: Erdős–Rényi ($\mu = 50$). |NB| versus the contamination level for MoR-type methods and prevalences: (a) 1%, (b) 5%, (c) 10%; RoS-type methods and prevalences: (d) 1%, (e) 5%, (f) 10%.

5.S.C.3. Transmission Error

We show the impact of the transmission error in our simulation studies. Transmission error occurs when respondents are unaware that some of their contacts belong to the hidden population, leading to under-reporting in survey responses (Laga et al., 2021). To simulate this error, we assume that a predetermined proportion of survey respondents ϵ_t has a particular visibility fraction τ. The visibility fraction, also called the transmission rate, indicates the proportion of neighbors they correctly recognize as members of the unknown population. Figures 5.S.18 and 5.S.19 display, respectively, the RMSRE and the |NB| against the visibility factor τ obtained by the various NSUM estimators for the ER network model, mean degree of $\mu = 50$, unknown group rate of $\rho = 10\%$, and a proportion of contaminated respondents of $\epsilon_t = 25\%$.

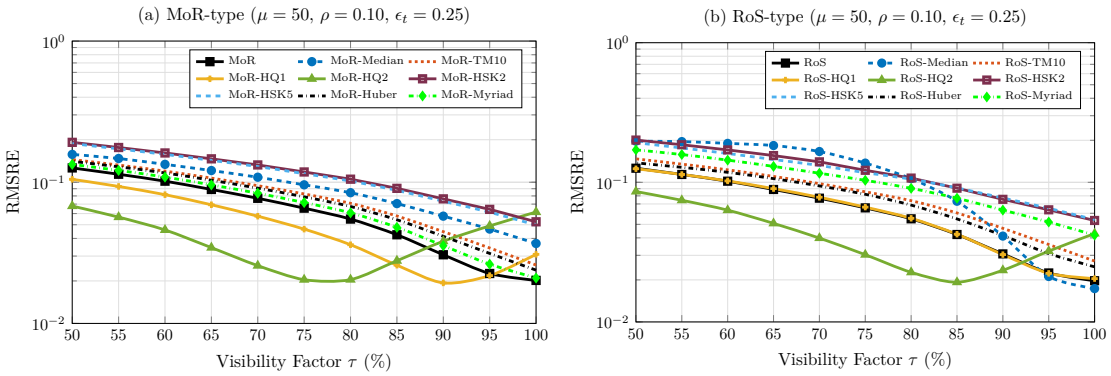


Figure 5.S.18: Erdős–Rényi ($\mu = 50$ and $\rho = 10\%$). RMSRE versus the transmission error level for the groups of robust estimators: (a) MoR-type and (b) RoS-type.

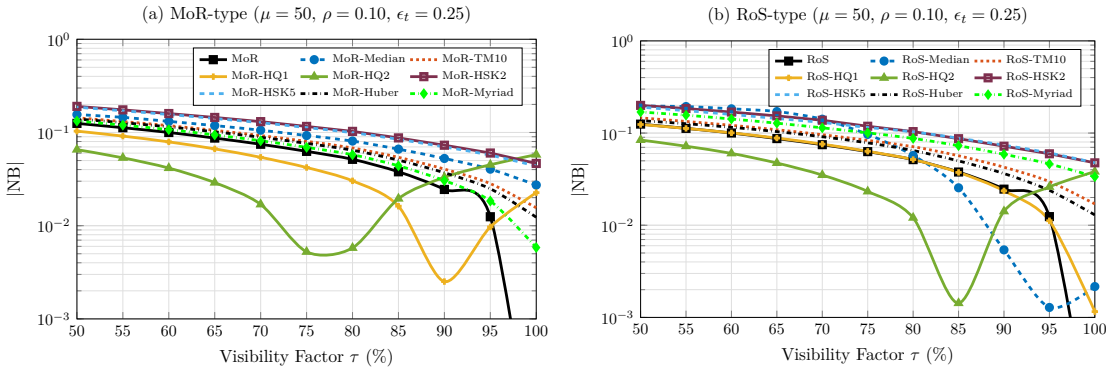


Figure 5.S.19: Erdős–Rényi ($\mu = 50$ and $\rho = 10\%$). |NB| versus the transmission error level for the groups of robust estimators: (a) MoR-type and (b) RoS-type.

5.S.C.4. Recall Error

Recall error arises when respondents inaccurately report the number of contacts belonging to the unknown group (Laga et al., 2021). Such inaccuracies may stem from several factors, including cognitive limitations or time constraints during survey administration. To simulate this type of error, we introduce a predefined proportion of respondents, denoted by ϵ_t, who either overcount or undercount a specific number of nodes belonging to the unknown group—referred to as the number of misreported nodes. For instance, when the number of misreported nodes is set to one, a proportion ϵ_t of respondents will randomly either add or subtract one from their true count, with equal probability. Figures 5.S.20 and 5.S.21 show, respectively, the RMSRE and the |NB| versus the number of misreported nodes obtained by the various NSUM estimators for the ER network model, mean degree of $\mu = 50$, unknown group rate of $\rho = 10\%$, and a proportion of contaminated respondents of $\epsilon_t = 25\%$.

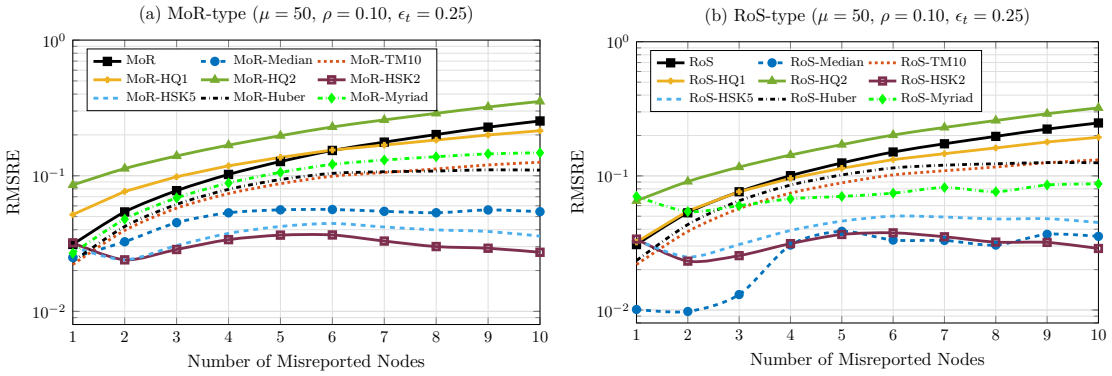


Figure 5.S.20: Erdős–Rényi ($\mu = 50$ and $\rho = 10\%$). RMSRE versus the number of misreported nodes (recall error) for the groups of robust estimators: (a) MoR-type and (b) RoS-type.

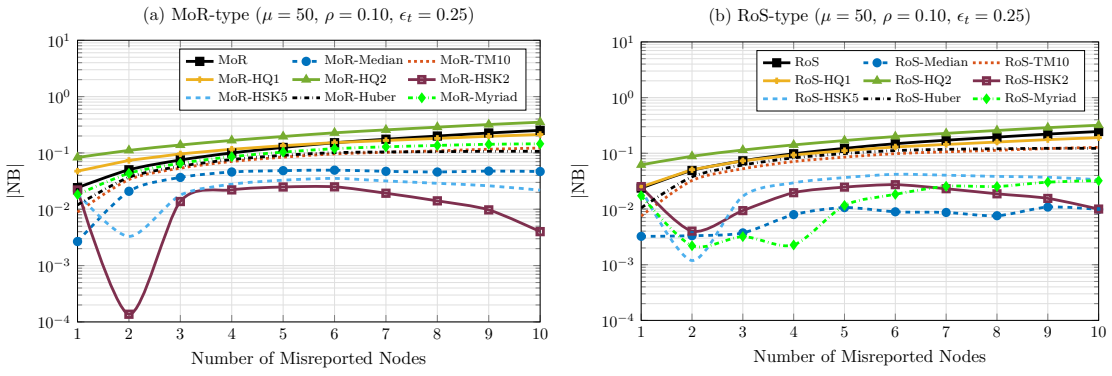


Figure 5.S.21: Erdős–Rényi ($\mu = 50$ and $\rho = 10\%$). |NB| versus the number of misreported nodes (recall error) for the groups of robust estimators: (a) MoR-type and (b) RoS-type.

5.S.D. Scale-Free Model

5.S.D.1. Without Contamination

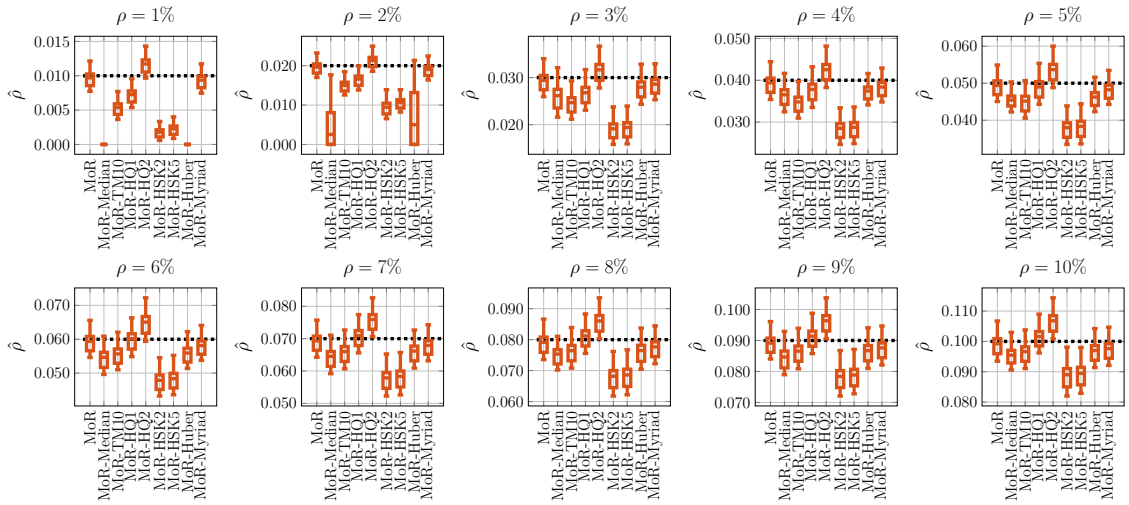


Figure 5.S.22: Scale-Free ($\mu = 50$). Boxplots yielded by robust MoR-type methods in scenarios without contamination for different prevalence levels.

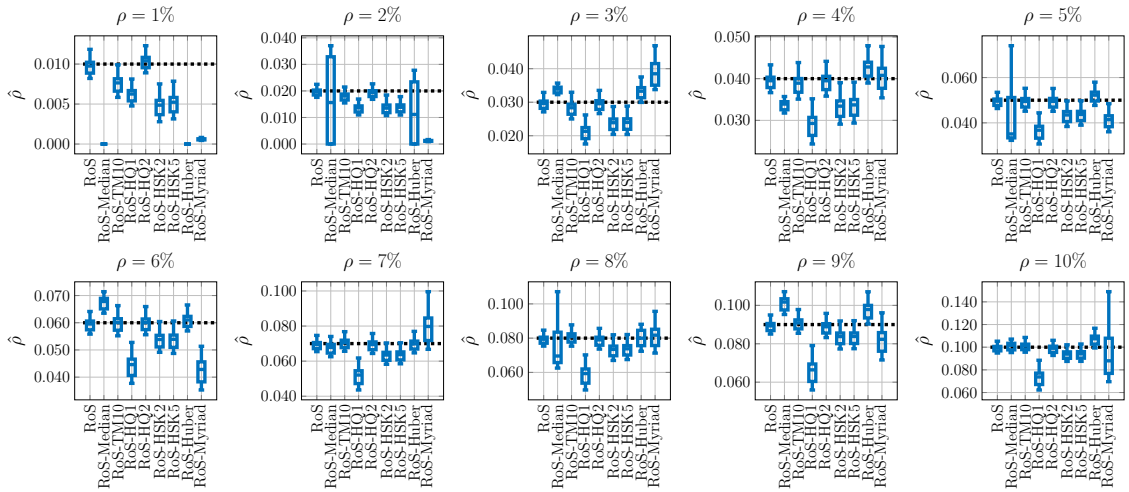


Figure 5.S.23: Scale-Free ($\mu = 50$). Boxplots yielded by robust RoS-type methods in scenarios without contamination for different prevalence levels.

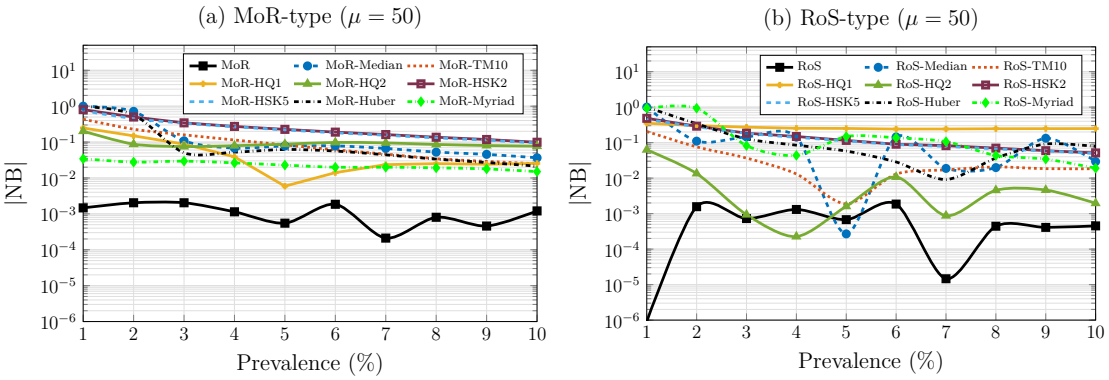


Figure 5.S.24: Scale-Free ($\mu = 50$). |NB| versus the prevalence without contamination for the groups of robust estimators: (a) MoR-type and (b) RoS-type.

5.S.D.2. With Contamination

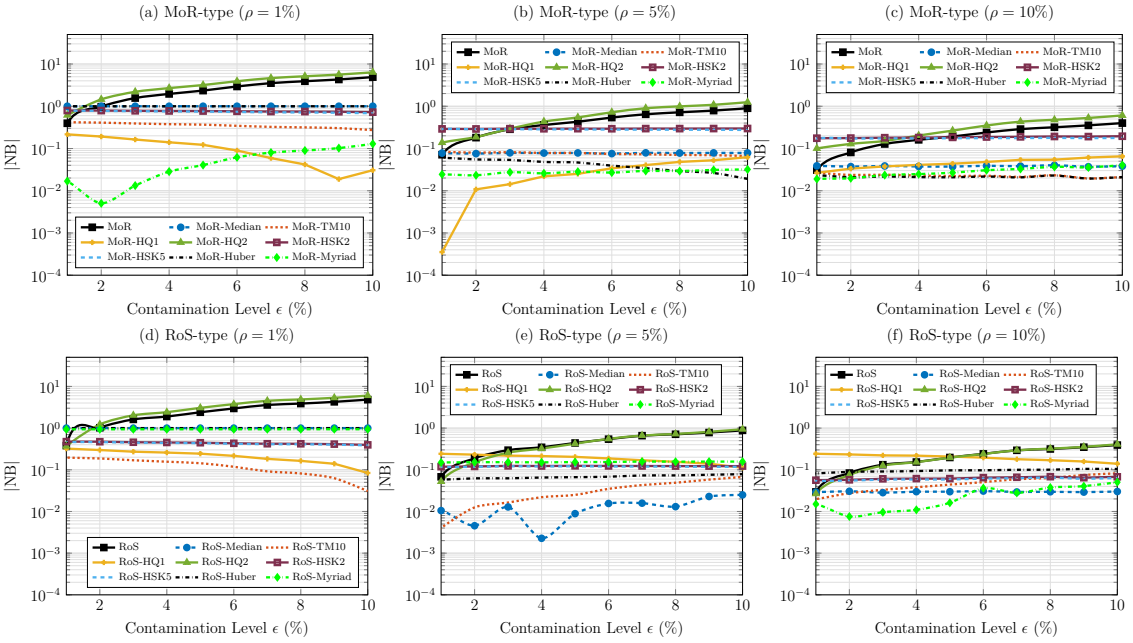


Figure 5.S.25: Scale-Free ($\gamma = 2.50$). |NB| versus the contamination level for MoR-type methods and prevalences: (a) 1%, (b) 5%, (c) 10%; RoS-type methods and prevalences: (d) 1%, (e) 5%, (f) 10%.

5.S.D.3. Transmission Error

See Section 5.S.C.3.

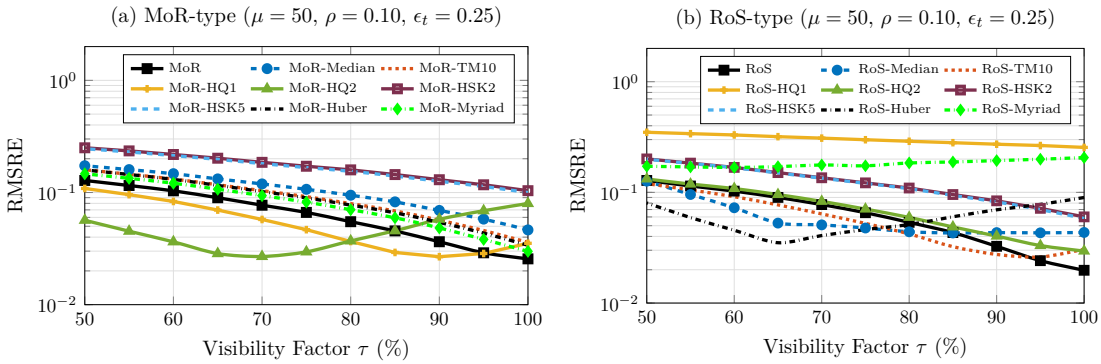


Figure 5.S.26: Scale-Free ($\mu = 50$ and $\rho = 10\%$). RMSRE versus the transmission error level for the groups of robust estimators: (a) MoR-type and (b) RoS-type.

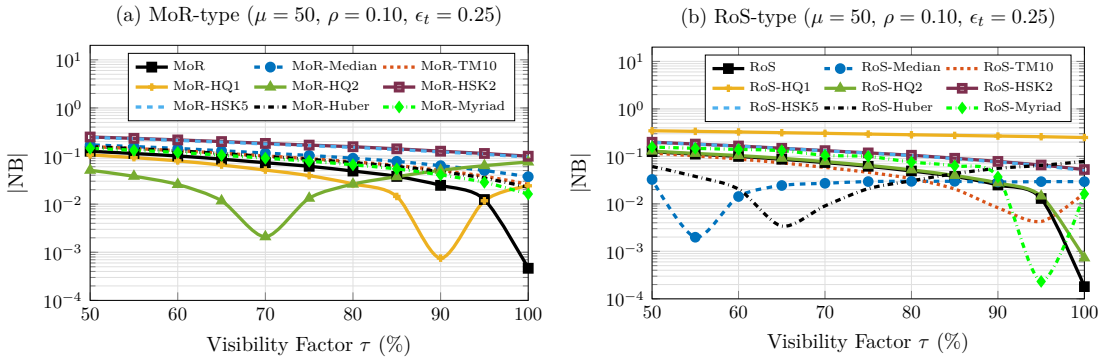


Figure 5.S.27: Scale-Free ($\mu = 50$ and $\rho = 10\%$). |NB| versus the transmission error level for the groups of robust estimators: (a) MoR-type and (b) RoS-type.

5.S.D.4. Recall Error

See Section 5.S.C.4.

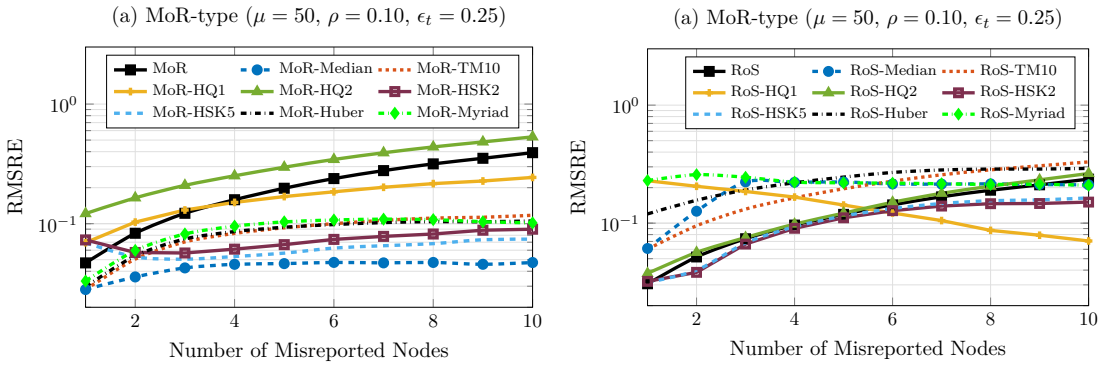


Figure 5.S.28: Scale-Free ($\mu = 50$ and $\rho = 10\%$). RMSRE versus the number of misreported nodes (recall error) for the groups of robust estimators: (a) MoR-type and (b) RoS-type.

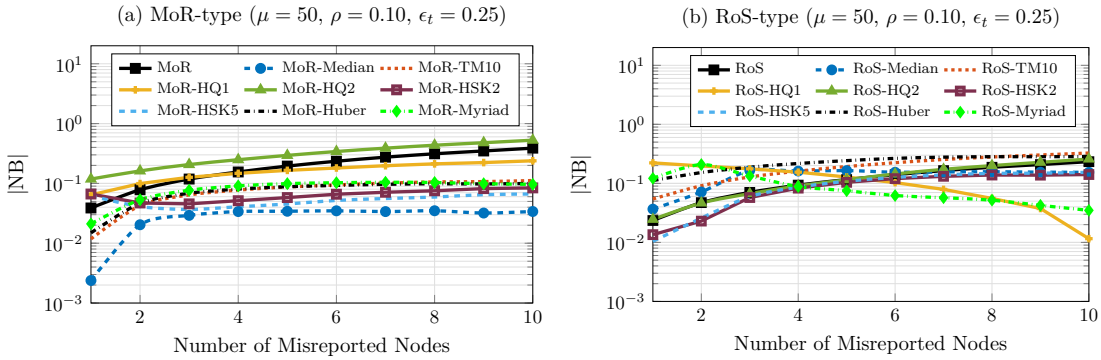


Figure 5.S.29: Scale-Free ($\mu = 50$ and $\rho = 10\%$). |NB| versus the number of misreported nodes (recall error) for the groups of robust estimators: (a) MoR-type and (b) RoS-type.

5.S.E. Stochastic Block Model

5.S.E.1. Without Contamination

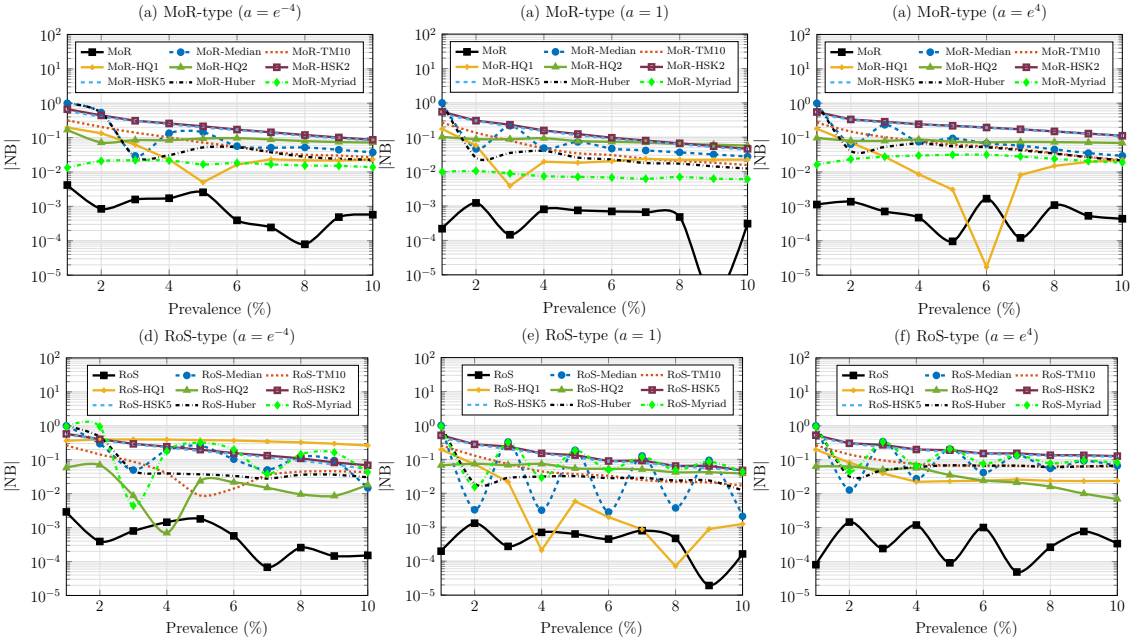


Figure 5.S.30: Stochastic Block Model. |NB| versus the prevalence level for different values of the connectivity factor a. MoR-type methods: (a) $a = e^{-4}$, (b) $a = 1$, (c) $a = e^{4}$; RoS-type methods: (d) $a = e^{-4}$, (e) $a = 1$, (f) $a = e^{4}$.

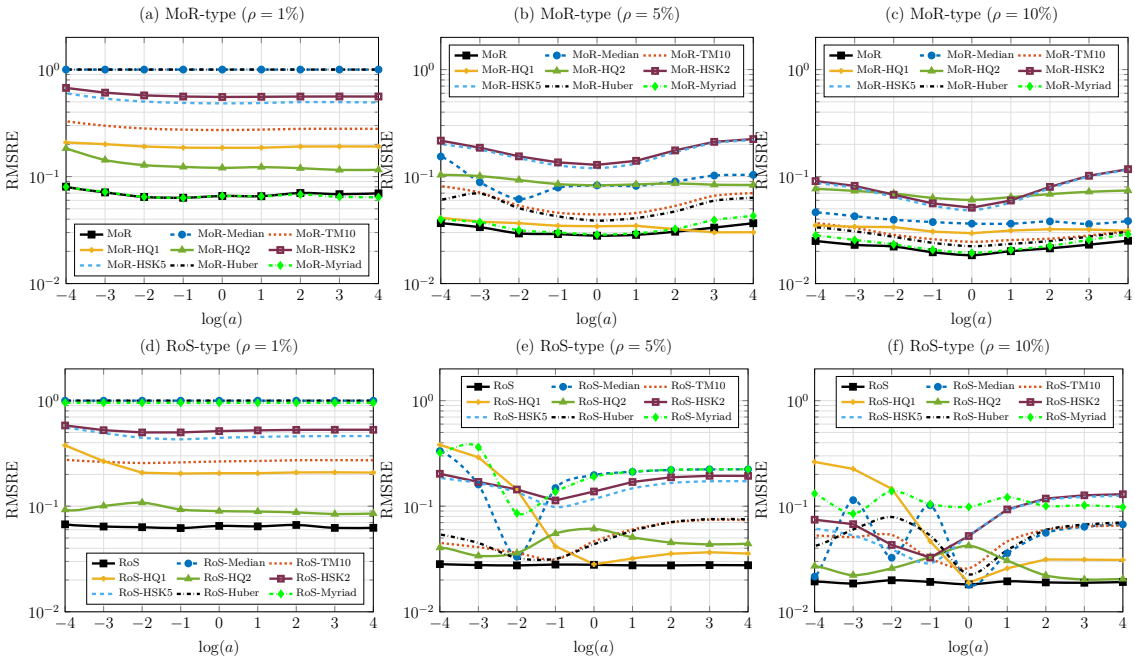


Figure 5.S.31: SBM. RMSRE versus the factor a for three different prevalence levels. MoR-type methods: (a) $\rho = 1\%$, (b) $\rho = 5\%$, (c) $\rho = 10\%$; RoS-type methods: (d) $\rho = 1\%$, (e) $\rho = 5\%$, (f) $\rho = 10\%$.

5.S.E.2. Performance versus the Connectivity Factor Without Contamination

Figures 5.S.31(a)-(c) present the RMSRE produced by the MoR-type methods concerning the logarithm of the connectivity factor across three different prevalence values: 1%, 5%, and 10%. The non-robust MoR and MoR-Myriad estimators perform the best at all three prevalence levels across the entire range of connectivity factors. Notably, for a prevalence of 10%, the errors of the various methods tend to increase as the connectivity factors move away from zero, indicating significant deviations from the Erdős–Rényi network model. Finally, Figs 5.S.31(d)-(f) depict the RMSRE generated by the RoS-type methods versus the logarithm of the connectivity factor for the three prevalence levels. In this case, the non-robust RoS estimator obtains the best results across the connectivity factor interval for the three prevalence values.

5.S.E.3. With Contamination

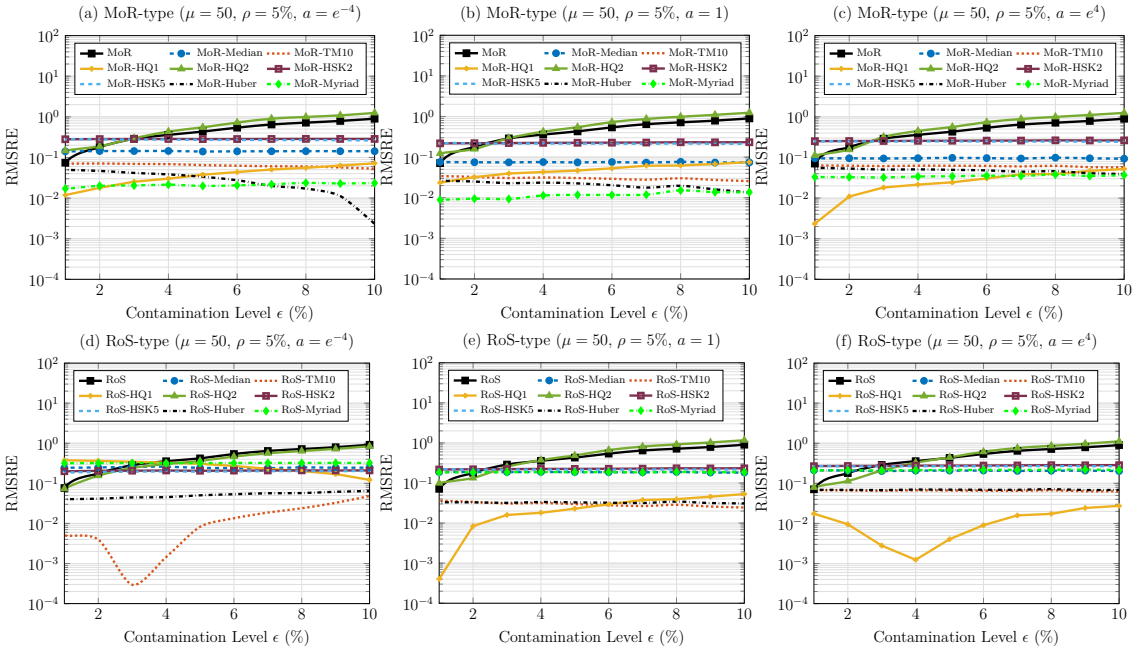


Figure 5.S.32: SBM ($\gamma = 2.50$). $|\mathrm{NB}|$ versus the contamination level for MoR-type methods, prevalence $\rho = 5\%$ and connectivity factors: (a) $a = e^{-4}$, (b) $a = 1$, (c) $a = e^{4}$; RoS-type methods, prevalence $\rho = 5\%$ and connectivity factors: (d) $a = e^{-4}$, (e) $a = 1$, (f) $a = e^{4}$.

Chapter 6

Network Scale-up Methods on Aggregated Relational Data to Estimate the Outcome of Elections

In *Social Networks*, 86, 312-324, 2026.
https://doi.org/10.1016/j.socnet.2026.04.006

Jorge M. Arevalillo[1,2] Juan Marcos Ramírez[3] Sergio Díaz-Aranda[3,1]
Jose Aguilar[1,4] Antonio Fernández Anta[6,1] Rosa E. Lillo[1,5]

[1] uc3m-Santander Big Data Institute, Universidad Carlos III de Madrid
[2] Department of Statistics and Operational Research, UNED, Spain.
[3] IMDEA Networks Institute, Spain
[4] Universidad de Los Andes, Venezuela
[5] Department of Statistics, Universidad Carlos III de Madrid, Spain
[6] IMDEA Software, Spain

Abstract

The Network Scale-Up Method (NSUM) is an estimation framework that aims to determine the size of hidden or hard-to-reach populations from questions such as "How many people do you know who belong to the target population?" The information collected from these questions is commonly referred to as aggregated relational data (ARD), and the estimation of hidden population sizes using NSUM in ARD has been widely applied to key problems in sociology and public health. Note that this approach has been widely used to estimate the size of populations subject to legal or social restrictions, such as sex workers and drug users, who are typically excluded from the formal census. Although voting intention is not a social stigma, this information has become a privacy-sensitive issue, particularly in polarized political contexts, and thus poses a challenge to determining the share of party support. In this work, we introduce a methodology for estimating vote-intention shares using NSUM techniques on ARD. The methodology involves the design of the indirect survey to collect ARD regarding the voting intention of the survey participants' contacts, jointly with other auxiliary questions, the processing of the data with appropriate filters to eliminate outliers, the study of sample stratification strategies, and finally, the support share estimation for each political group by using different

NSUM techniques. The methodology is applied to estimate voting outcomes in the 2023 Spanish general elections, using the Madrid, Andalusia, and Valencia regions as experimental scenarios. Overall, the resulting estimates are competitive with those published by leading private and public survey institutes, despite using a significantly smaller number of participants.

Keywords: Aggregated relational data, Indirect surveys, Network scale-up methods, Voting intention shares

6.1. Introduction

Direct surveys have been commonly used to estimate election results. In recent years, the results of surveys conducted by companies and institutions have not been free of public controversy due to their poor estimates. Many factors could explain this finding, such as coverage errors due to underestimating segments of the population, a poorly designed sampling strategy, or biases introduced by undecided voters in surveys, among others. Recent examples include surveys for the July 2023 general elections in Spain, which failed to estimate parliamentary seats (Fernández-Roldán and Barnfield, 2024), and surveys for the first round of the October 2023 presidential elections in Argentina, which failed to predict the leading candidate.

An alternative to direct surveys is indirect reporting, in which respondents are asked to report the number of their contacts in specific groups. This approach has at least three significant advantages over direct surveys. First, the surveys are designed to avoid the collection of personal information from participants and thus protect their privacy (i.e., they make the study anonymous), thereby reducing some privacy risk. Secondly, they have broader coverage with the same number of participants, since respondents are asked about their contacts. Thirdly, there are techniques that allow estimation of the subpopulations of interest in the survey, known in the literature as network scale-up methods (NSUM), which, in a broad sense, refer to the process of estimating network or subpopulation size using aggregated relational data (ARD) (Laga et al., 2021). ARDs are obtained by reporting on the different groups known to the respondent (e.g., their network of contacts).

In general, direct surveys have limitations when dealing with the study populations sharing some "sensitive" characteristic, like, for example, sex workers, LGBTQ+ individuals, people with HIV/AIDS, refugees and immigrants, formerly incarcerated individuals, religious or ethnic minorities, etc., due to social stigma. This drawback also applies to electoral surveys, as many individuals may conceal their political opinions for privacy reasons.

It is important to note that over the past 35 years, research on NSUM has focused on quantifying the size of such populations from ARD. Nevertheless, these techniques

have not been widely applied in public opinion surveys used to estimate the electoral support for political parties, making our proposal an innovative contribution to this field. Several previous works have used indirect surveys to estimate electoral results (Murr, 2011; Galesic et al., 2018; Graefe, 2014, 2016; Bruine de Bruin et al., 2022). However, none of them present how to use the NSUM methods to estimate electoral outcomes, so they have not analyzed the behavior of existing methods (stochastic, frequency-based, etc.). Additionally, none of these works described how to prepare the data collected from indirect electoral surveys to make the estimations, nor do they present a methodology for using indirect surveys in electoral processes.

This article proposes a methodology based on indirect surveys and NSUM to estimate electoral process voting outcomes. In that sense, our methodology proposes a framework for conducting indirect surveys and then applies NSUM-based approaches to estimate vote percentages by political group. Furthermore, the article examines the indirect reporting approach in the context of the July 2023 Spanish general elections. Thus, the main contribution is the design of a methodology based on indirect surveys and the NSUM to estimate electoral outcomes. Some other contributions of this work are:

- The design of a survey structure/format of indirect surveys for electoral contexts.
- The characterization of filters to clean surveys of atypical values.
- The comparison of different NSUM approaches in estimating election outcomes. In particular, the following methods were compared: the Maximum Likelihood Estimator (MLE), the Plug-in Maximum Likelihood Estimator (PIMLE), the Means of Sums (MoS), and the Naïve.
- A comparative analysis of the results of our approach with those of other direct surveys conducted in Madrid, Andalusia, and Valencia during the aforementioned electoral process.

The rest of the paper is organized as follows: the next section deals with an overall review of related work about NSUM. Sections 6.3 and 6.4 are the core parts of our work as they present a unified methodology covering the survey design and data collection, the data processing step, and the application of NSUM approaches for the estimation of election outcomes in real cases. Specifically, Section 6.3.3 addresses the approaches based on NSUM that will be applied to the estimation of election voting outcomes. The methodology is applied to the Spanish general elections of July 23rd, 2023, in Madrid, Valencia, and Andalusia regions. Section 6.5 describes the resulting estimations, which are compared with the results published by private and public survey houses and with the ground truth outcome. Finally, Sections 6.6 and 6.7 finish this work with a discussion of the application of the proposed methodology and some concluding remarks.

6.2. Related Work

Indirect surveys have emerged as a strategy to collect information quickly and anonymously, and then estimate the size of hidden subpopulations from such indirect survey data. To estimate unknown subpopulations, it is necessary to collect information about the contact network of surveyed individuals. The initial idea behind NSUM was introduced by Bernard et al. (1989), and consists of using ARD to obtain the network size (degree) of the participants, and then using this information to estimate the unknown subpopulation. Subsequently, works based on this approach continue to appear (Killworth et al., 1998b,d; Habecker et al., 2015b). Particularly relevant is the estimator proposed by Killworth et al. (1998d) since it has become the most employed NSUM estimation procedure (Ezoe et al., 2012; Snidero et al., 2012; Wang et al., 2015; Nikfarjam et al., 2016). The NSUM usually uses questions about groups whose size is known to estimate each surveyed individual's network size. Also, McCarty et al. (2001) use several questions that cover all the possible acquaintances, and then assign the degree to the sum of the responses. Additionally, Garcia-Agundez et al. (2021) directly ask about the size of the personal network. The NSUM approaches have been applied in epidemiological problems (UNAIDS, 2017; Maltiel et al., 2015; Teo et al., 2019; Feehan et al., 2016), health and social issues such as drug users (Kadushin et al., 2006; Salganik et al., 2010), cancer patients (Carletti et al., 2017), abortions (Sully et al., 2020), and sex workers (Jing et al., 2018).

On the other hand, several works have been developed to address the estimation procedure with NSUM. New Bayesian models have been proposed to handle biases (Zheng et al., 2006; McCormick and Zheng, 2012; Maltiel et al., 2015; Teo et al., 2019) and the correlation between groups (Laga et al., 2023). Other works add direct survey data in order to create new estimators or adjust previous methods (Feehan and Salganik, 2016; Verdery et al., 2019; Parsons et al., 2022). Finally, ARD has been used as well to produce a latent network representation (McCormick and Zheng, 2015), reconstruct network (Breza et al., 2020), and estimate network properties (Breza et al., 2019, 2023). Two review works about NSUM and ARD (Laga et al., 2021; McCormick, 2020) collect most of the existing approaches in the literature.

In the context of electoral surveys, Sio and Weber (2020) define a model that explains election outcomes on the basis of party strategy. The article traces a three-step process: (1) The configuration of electoral risks and opportunities (which is captured through public opinion surveys) guides party communication (measured with Twitter data) such that (2) parties design their campaigns strategically, which in turn, (3) improves their electoral performance (measured using official statistics). The article (Schmitt et al., 2020) analyses the second-order election (SOE) model originally developed by Reif and Schmitt in 1980, which describes voter behavior in the second round of two-round elections. This model suggests that, relative to the behavior described by the first-order model (describing voter behavior in the

first round), turnout is lower, large parties lose, and small and ideologically extreme parties win. This work analyses such hypotheses of the SOE model in the 2004 and 2014 European elections. Skoric et al. (2020) report a meta-analysis examining the predictive power of social media data by focusing on (1) sentiment analysis and (2) analysis of structural features. The study shows that machine learning-based estimates are generally better than those derived from pre-existing lexica and that a combination of structural features and sentiment analyses provides the most accurate predictions.

Some other works have exploited direct surveys to analyze specific aspects of electoral processes: polarization, the behavior of sectors of the population, opinions on specific issues, etc., which is an interesting preamble to the use of indirect surveys for these same studies. For example, Hernández et al. (2021) analyze the effects of the election on affective polarization. The article studies the origins of affective polarization by exploring two mechanisms: changes in ideological polarization and in the intensity of party identification. de Geus et al. (2022) investigate the prevalence and correlates of sexism in the British political context, using a measure of ambivalent sexism that distinguishes between hostile and benevolent sexist attitudes. They use two online surveys and find that more than half of the population holds some sexist attitudes, which can be predicted by gender, education, religiosity, and authoritarian values. The work of Schroedel et al. (2020) analyzes the impact of political trust on Native American electoral participation using direct surveys. They use in-person survey responses conducted in Native communities with the support of tribal leaders. Respondents expressed high levels of distrust in nontribal government and voting methods, and this distrust had an impact on the decision to participate in nontribal elections. As we can see, these latest works touch on sensitive issues such as polarization, sexism, and trust in government. Indirect surveys could improve the quality of the information retrieved by expanding the coverage and by protecting the privacy of participants (they may respond more truthfully and thus produce less bias).

There are other works that have discussed the use of indirect surveys and, especially, of collective questions that inquire about a population or the immediate social circles of respondents, in a variety of elections in different countries. For example, Murr (2011) explored the "wisdom of crowds" in forecasting elections. The approach relies on citizen prediction, which aggregates citizens' local expectations to directly forecast constituencies. Using the 2010 British Election Study data, the study highlighted that group size significantly influences the accuracy of predictions. This approach can directly forecast seat shares, although sometimes it yields a high 15% error. Another aspect, which is not analyzed by the approach, is the number of respondents needed to ensure a good forecasting capacity. Finally, the aggregation of local expectations to predict does not consider the effect of electoral campaigns, which can influence the opinion of the respondent. On the other hand, Galesic et al. (2018)

demonstrated that social-circle questions, which inquire about the voting intentions of participants' social contacts, offer improved predictions. These questions accurately forecasted the 2016 US and 2017 French presidential election outcomes on national, state, and individual levels. They also helped explain last-minute voting intention changes and provided insights into the dynamics of echo chambers among supporters of different candidates. The article describes how the survey was conducted (waves of study) that differed in the types of questions asked (three types of questions: questions about one's own intention, about one's social circle and about the winner of the election) and when they were asked (they were asked together, separately, etc.), but it does not give details of how the survey results were estimated (nor does it show these results). Graefe (2014) compared the accuracy of vote expectation surveys for forecasting US presidential elections from 1932 to 2012 to predictions from four established methods: expert judgment, traditional polls, prediction markets, and quantitative models. Also, Graefe (2016) evaluated vote expectation surveys for forecasting elections in the 2013 German election. It concluded that representative surveys are a valuable, low-cost method for forecasting, especially for small-scale elections where traditional methods like representative polls or prediction models may not be feasible or cost-effective. One concern with using this approach is that people's expectations may be influenced by their preferences. In the case of elections, this means that respondents tend to predict that their preferred candidate will win. While this approach does follow Condorcet's jury theorem, which implies that the probability of a group reaching a correct decision increases rapidly toward 100 percent as the number of individual estimates increases, this bias is not eliminated. In other words, even if you use Condorcet's jury theorem, which combines individual estimates into one group estimate, this is correct if you manage to distribute the sample well across all candidates. Finally, a recent work of Bruine de Bruin et al. (2022) found that social-circle questions outperform own-intention questions in predicting elections with numerous political parties, specifically the Netherlands' 2017 general election and Sweden's 2018 general election. This suggests that social-circle questions remain effective even in complex electoral environments. However, the work does not detail the characteristics of the samples, the environment, or how the calculations are carried out to ensure good results.

In the reviewed literature, we find methods based on indirect reports according to the concept of social-circle questions, but they do not provide a unified methodological framework, and do not show how to use the NSUM methods to estimate the result of an electoral process by assessing the voting intention for each political group. This is relevant since there are different NSUM approaches (frequentist, Bayesian, and network models), each one having properties that advise its suitability for the application in specific cases. Knowing them and how to use them, would allow us to know the method that better fits each context. Furthermore, there are no relevant works that indicate how to prepare the data collected from electoral indirect surveys, and in general, there is a research gap in the literature dedicated to analyzing the

quality of electoral indirect surveys.

6.3. Methodology

Our proposal relies on a methodology that unifies the main steps of the estimation problem as follows: First, data are collected via indirect surveys, for which a well-suited design is required in advance. In the second step, the data are processed to improve data quality by removing irrelevant information (e.g., outliers or aberrant responses). Finally, standard NSUM procedures (see Table 6.1) are used to estimate the election results for each political group. Hence, the result of the methodology will be the estimations of the vote proportion for each political party: $\hat{p}_u^{Naive}$, $\hat{p}_u^{MoS}$, $\hat{p}_u^{MLE}$ and $\hat{p}_u^{PIMLE}$. Each step of the proposed methodology is described in the next sections.

6.3.1. Survey Design and Data Collection

At this stage, the basic requirements for implementing the NSUM are established. They are concerned with the following issues involved in survey design and data collection:

- Selection of the digital survey platform used to carry out the social network survey. Pollfish or Prolific are examples of well-established platforms in the market.
- The sampling framework, which should include at least the geographic and socio-demographic limits of the survey.
- The size of the survey, which is usually limited by budget restrictions.
- The preparation of the questionnaire that includes the "how many..." questions about the political options and blank vote (it includes blank ballot, abstention, and "other parties"), as well as questions regarding the auxiliary known populations.
- The possible stratified sampling strategy by using the age, gender, or some other social attributes of surveyed individuals.

In particular, we used the Pollfish platform to conduct online surveys, which is age-biased (unbalanced toward younger respondents). The first time we used it for the regional elections in Madrid (see Supplementary Material 6.S.A.3), we observed this imbalance, so we proposed a post-stratification strategy (i.e., the definition of population weights). In the case of stratified sampling, it was implemented directly on the platform, and age was selected as the stratification variable because the sample's age distribution differed from the population distribution. The ideal is to conduct a study of all potential stratification variables to select the one to use. Thus, several strategies could be proposed: (i) choose the variable whose stratified sample is closest to the real one; (ii) carry out sampling using a different variable in each round; (iii) use a metric from information theory (for example, entropy) to select the

variable that provides more information. This is one possible extension of this work.

6.3.2. Data Processing

Once the survey has been conducted, the collected responses must be processed. This phase concerns the preparation of the survey data. A key task is data cleaning to eliminate inconsistent answers and outliers. To do this, a set of rules must be established in advance to ensure data quality in accordance with specified criteria. Such criteria should include at least the following points:

- Identification and elimination of cases with inconsistent (abnormal) network size.
- Identification and removal of responses declaring an abnormal number of blank votes.
- Identification and elimination of too many polarized responses.

Their specific application to real use cases is addressed in Section 6.4.2.

6.3.3. Estimation of Voting Intention

We propose to use NSUM for estimating the size of populations sharing some "sensitive" characteristic, which in our case corresponds to the number of potential voters for each political group. Our proposal is based on the simple idea of assessing these populations through indirect survey data, which collects the voting intention for the different political parties that participate in the electoral process, by means of questions about the network of respondents as follows: "How many people do you know who will vote for political party X?". Our approach mitigates many legal privacy concerns, as respondents provide information only about non-identifiable individuals in their networks.

In this context, we must consider different sources of uncertainty arising from indirect survey reporting. The main ones can be described as follows:(i) *barrier effects* are concerned with biases produced by the fact that the probability of knowing someone in a population depends on the characteristics of the participants; (ii) *transmission errors* are produced by the fact that the participant does not know the status of their acquaintances; (iii) *recall errors* occur when the participants inaccurately remember the number of their contacts that belong to some subpopulations. NSUM approaches are influenced by the previous sources of uncertainty.

The NSUM relies on the following simple idea: first, the degree (network size) of respondents is estimated using a set of auxiliary questions about populations of known sizes (usually data from official statistics) within their personal networks. Then, the survey data collected from participants' networks are combined with the degree estimates to construct an estimate of the unknown target population. In

order to formalize the previous ideas along with their implementation by different NSUM estimation strategies, we now introduce some notation:

- Let N be the size of the overall population.
- Let n be the size of the survey (number of respondents).
- Let Q be the number of available political options and let us denote by N_u the unknown size of the population voting for the u-th political group (blank vote option included): $u = 1, \ldots, Q$.
- Let y_{iu} be the number of people the respondent i declares to know that will vote for a specific political option u, and let us denote by $y_i = (y_{i1}, \ldots, y_{iQ})$ the vector collecting such responses for all the political groups.
- Let L be the number of auxiliary subpopulations, with known sizes $\mathcal{N}_k : k = 1, \ldots, L$, which are usually obtained from official statistics.
- Let us denote by x_{ik} the answer of respondent i to the k-th auxiliary question, and let us denote by $x_i = (x_{i1}, \ldots, x_{iL})$ the responses given by respondent i to the L auxiliary questions.

The first ad hoc strategy would ignore the survey data from the auxiliary populations and would use the network vote declaration for the different political groups to estimate the network size of respondent i by $\hat{d}_i = \sum_{v=1}^{Q} y_{iv}$. Hence, the Naïve estimator for the unknown N_u would be

$$\hat{N}_u^{Naive} = \frac{1}{n} \sum_{i=1}^{n} N \cdot \frac{y_{iu}}{\sum_{v=1}^{Q} y_{iv}} : u = 1, \ldots, Q. \tag{6.1}$$

A second Naïve strategy to estimate N_u would rely on the extrapolation to the general population of the declared voting intention to political group u by all the respondents to the survey. Such extrapolation would be given by

$$\hat{N}_u^{Naive2} = N \cdot \frac{\sum_{i=1}^{n} y_{iu}}{\sum_{v=1}^{Q} \sum_{i=1}^{n} y_{iv}} : u = 1, \ldots, Q. \tag{6.2}$$

Standard NSUM approaches use sums and averages as the building blocks to obtain the estimation of an unknown target population. The most popular methods include the Maximum Likelihood Estimator (MLE), the Plug-in Maximum Likelihood estimator (PIMLE), and the means of sums (MoS). The mathematical expressions giving the estimations of the degree and the size of the unknown population are shown in Table 6.1. The table uses the notation employed in the review work (Laga et al., 2021) and in the references to the seminal works that proposed each method. While MLE relies on random mixing assumptions – specifically, that the reported number of contacts in the hidden population follows a Binomial distribution with probability of success equal to the proportion of the hidden population – PIMLE and MoS only assume that the ratios y_{iu}/d_i are representative on average. All the

Method	Degree estimation	Estimation of hidden population	Reference
PIMLE	$\hat{d}_i = N \cdot \frac{\sum_{k=1}^{L} x_{ik}}{\sum_{k=1}^{L} \mathcal{N}_k}$	$\hat{N}_u^{PIMLE} = \frac{1}{n} \sum_{i=1}^{n} N \cdot \frac{y_{iu}}{\hat{d}_i}$	Killworth et al. (1998a)
MLE	$\hat{d}_i = N \cdot \frac{\sum_{k=1}^{L} x_{ik}}{\sum_{k=1}^{L} \mathcal{N}_k}$	$\hat{N}_u^{MLE} = N \cdot \frac{\sum_{i=1}^{n} y_{iu}}{\sum_{i=1}^{n} \hat{d}_i}$	Killworth et al. (1998c)
MoS	$\hat{d}_i = \frac{1}{L} \sum_{k=1}^{L} N \cdot \frac{x_{ik}}{\mathcal{N}_k}$	$\hat{N}_u^{MoS} = \frac{1}{n} \cdot \sum_{i=1}^{n} N \cdot \frac{y_{iu}}{\hat{d}_i}$	Habecker et al. (2015a)
MoS2	$\hat{d}_i = \frac{1}{L} \sum_{k=1}^{L} N \cdot \frac{x_{ik}}{\mathcal{N}_k}$	$\hat{N}_u^{MoS2} = N \cdot \frac{\sum_{i=1}^{n} y_{iu}}{\sum_{i=1}^{n} \hat{d}_i}$	—
Naïve	$\hat{d}_i = \sum_{v=1}^{Q} y_{iv}$	$\hat{N}_u^{Naive} = \frac{1}{n} \sum_{i=1}^{n} N \cdot \frac{y_{iu}}{\hat{d}_i}$	—
Naïve2	$\hat{d}_i = \sum_{v=1}^{Q} y_{iv}$	$\hat{N}_u^{Naive2} = N \cdot \frac{\sum_{i=1}^{n} y_{iu}}{\sum_{i=1}^{n} \hat{d}_i}$	—

Table 6.1: Standard NSUM approaches based on sums and averages.

referenced approaches are implemented in the `networkscaleup` R package (Laga et al., 2022).

Note that a second alternative for the MoS method would arise from the following combination (see the fourth row of Table 6.1).

$$\hat{N}_u^{MoS2} = N \cdot \frac{\sum_{i=1}^{n} y_{iu}}{\sum_{i=1}^{n} \hat{d}_i} \text{ with } \hat{d}_i = \frac{1}{L} \sum_{k=1}^{L} N \cdot \frac{x_{ik}}{\mathcal{N}_k}. \tag{6.3}$$

For the sake of completeness, we are also including in the table the expressions for the classical estimators, along with expressions for both Naïve estimation methods given by formulas (6.1) and (6.2), and the MoS2 method (6.3).

Let us denote by $\hat{N}_u$ the estimator of the number of votes obtained by a scale-up method for any of the available political choices: $u = 1, \ldots, Q$. Once the size, N_u, of the unknown population has been estimated, we propose to estimate the voting percentage of the u-th political group by means of the following expression:

$$\hat{p}_u = \frac{\hat{N}_u}{\sum_{v=1}^{Q} \hat{N}_v} : u = 1, 2, \ldots, Q. \tag{6.4}$$

Depending on the network scale-up method being applied, we will get the versions Naïve, Naïve2, MoS, MoS2, MLE, and PIMLE for the estimation of the proportion. The estimations obtained by these methods will be denoted as follows: $\hat{p}_u^{Naive}$, $\hat{p}_u^{Naive2}$, $\hat{p}_u^{MoS}$, $\hat{p}_u^{MoS2}$, $\hat{p}_u^{MLE}$ and $\hat{p}_u^{PIMLE}$, where the superscript indicates the method and the subscript denotes the political group.

It is worthwhile noting that:

$$\hat{N}_u^{Naive2} = N \cdot \hat{p}_u^{Naive2} \text{ and } \hat{N}_u^{MoS2} = N \cdot \hat{p}_u^{MoS2}$$

since $\sum_{v=1}^{Q} \hat{N}_v^{Naive2} = \sum_{v=1}^{Q} \hat{N}_v^{MoS2} = N$.

In this work, we only consider the versions provided by the methods MLE, PIMLE, MoS, and Naïve because it can be proved that $\hat{p}_u^{MLE} = \hat{p}_u^{MoS2} = \hat{p}_u^{Naive2}$, as the next proposition shows.

Proposition 6.1. *Let us consider the estimators $\hat{p}_u^{MoS2}$, $\hat{p}_u^{Naive2}$, and $\hat{p}_u^{MLE}$ of the voting proportion, being defined by the quantities:*

$$\hat{p}_u^{Naive2} = \frac{\hat{N}_u^{Naive2}}{\sum_{v=1}^{Q} \hat{N}_v^{Naive2}} \text{ , } \hat{p}_u^{MoS2} = \frac{\hat{N}_u^{MoS2}}{\sum_{v=1}^{Q} \hat{N}_v^{MoS2}} \text{ and } \hat{p}_u^{MLE} = \frac{\hat{N}_u^{MLE}}{\sum_{v=1}^{Q} \hat{N}_v^{MLE}},$$

respectively. Then, it holds that

$$\hat{p}_u^{Naive2} = \hat{p}_u^{MoS2} = \hat{p}_u^{MLE}.$$

Proof. From the expressions of the MLE estimator that appear in Table 6.1, it can be obtained that:

$$\hat{p}_u^{MLE} = \frac{\sum_{i=1}^{n} y_{iu}}{\sum_{v=1}^{Q} \sum_{i=1}^{n} y_{iv}}.$$

On the other hand, as the sum $\sum_{i=1}^{n} \hat{d}_i$, with $\hat{d}_i$ being the estimator of the degree in $\hat{N}_u^{MoS2}$, does not depend on the unknown voting populations, we get:

$$\hat{p}_u^{MoS2} = \frac{\sum_{i=1}^{n} y_{iu}}{\sum_{v=1}^{Q} \sum_{i=1}^{n} y_{iv}} = \hat{p}_u^{MLE}.$$

Using similar arguments, we also obtain that $\hat{p}_u^{Naive2} = \hat{p}_u^{MLE}$. □

It can also be noted that when the PIMLE method is applied to expression 6.4, it turns out a quantity independent of the totals $\mathcal{N}_k$ in the auxiliary information. This fact is proved by simple calculations as follows: from the first row of Table 6.1 we get

$$\hat{N}_u^{PIMLE} = \frac{1}{n} \sum_{i=1}^{n} \frac{y_{iu}}{\frac{\sum_{k=1}^{L} x_{ik}}{\sum_{k=1}^{L} \mathcal{N}_k}} = \left(\frac{1}{n} \sum_{k=1}^{L} \mathcal{N}_k \right) \sum_{i=1}^{n} \frac{y_{iu}}{\sum_{k=1}^{L} x_{ik}},$$

which in turn implies that

$$\hat{p}_u^{PIMLE} = \frac{\hat{N}_u^{PIMLE}}{\sum_{v=1}^{Q} \hat{N}_v^{PIMLE}} = \frac{\sum_{i=1}^{n} \frac{y_{iu}}{\sum_{k=1}^{L} x_{ik}}}{\sum_{v=1}^{Q} \sum_{i=1}^{n} \frac{y_{iv}}{\sum_{k=1}^{L} x_{ik}}}$$

as we have just asserted.

6.4. Application of the Methodology

The proposed methodology is applied to the Spanish general elections of July 23rd, 2023 (national elections) in specific communities (Madrid, Andalusia, and Valencia). The application of each step of the methodology is described in the next sections.

6.4.1. Survey Design and Data Collection

The data were collected by indirect survey reporting using the facilities provided by Pollfish Survey Tools (https://www.pollfish.com/). Below, we describe the survey design and how the information has been obtained.

- One round was conducted during the electoral campaign (fifteen days before the voting day).
- We do not limit the number of contacts of surveyed individuals, so it is left open for him/her to specify the size of his/her network.
- **Sampling framework:** The surveyed people are individuals who live in the Madrid, Andalusia, or Valencia regions, are at least 18 years old, and have a potential voting network in these regions. Moreover, the respondents are asked only about their social network contacts who belong to these regional electoral censuses.
- **Sample size:** $n = 200$ with stratification by age for the general elections (more details about stratification in Section 6.4.3).
- **Target unknown populations:** The main political groups presenting candidates for the general elections: PP, PSOE, Vox, and Sumar, the latter arising from the alliance between Más Madrid and Podemos. Adding the option of Blank vote yields in this case $Q = 5$.
- **Auxiliary known populations:** People with age in the intervals $(18, 30]$, $(31, 54]$ and > 54, self-employed people, unemployed, and physicians who belong to the respondent network. The real data for the auxiliary populations were obtained from official statistics provided by the Instituto Nacional de Estadística (INE, https://www.ine.es/), a public institution that manages all official statistics in Spain.
- **Questionnaire:** Both data sources, target and auxiliary, are collected from respondents by asking "How many people who ... do you know?" questions.

Additionally, the Pollfish platform provides demographic information on surveyed individuals, including age, income, and education. Note that the survey design assumes respondents are already aware of their contacts' voting intentions; that is, the questionnaire did not prompt respondents to inquire about these intentions directly during data collection. To mitigate potential biases arising from uncertainty about contacts' voting intentions, the survey includes auxiliary questions that assess respondents' familiarity with the demographic characteristics of their immediate environment, e.g., "How many of your contacts considered above are between 18 and 34 years old?" (see Section 6.S.A.1 in the Supplementary Material). Note also that MLE and PIMLE methods indirectly use these auxiliary questions to determine voting intention. Specifically, these methods integrate the auxiliary questions to determine the size of each respondent's network (Killworth et al., 1998a,c).

The estimand in our study is the proportion of voting intentions among contacts

for whom the respondent knows the political preference. That is, we estimate vote shares within the subnetwork of politically legible ties, rather than the full social network. In this study, respondents' political preferences are unavailable. We observe aggregated political preferences of their contacts, as reported by the respondents. This distinction is important for understanding what is being measured and how the results should be interpreted. The analysis relies on several arguments that explain why the observed preferences of contacts can still provide a useful approximation of broader political patterns. First, the analysis is conducted in a high-salience context. During general elections, political discussion is widespread, and voting intentions are often disclosed in social contexts, such as conversations with family, friends, or coworkers. This increases the likelihood that respondents are aware of the political preferences of a meaningful portion of their contacts. Second, the NSUM estimators use aggregate information from many respondents. Hence, individual-level selection effects, such as differences in political knowledge among respondents, partially average out when computing overall estimates. Taken together, these elements support the idea that, even without direct information on respondent preferences, the reported preferences of contacts can offer a reasonable and informative view of political support patterns in the population.

Finally, barrier effects in NSUM are not solely determined by age but also by the interaction between respondent characteristics and the social visibility of each target group. Different political groups may face distinct barriers depending on polarization, stigma, group size, and the contexts in which political discussions occur. In our analysis, age stratification and known population corrections mitigate some of these effects but do not eliminate all sources of heterogeneity. For this reason, we interpret the estimates as aggregates that average across respondent- and group-specific barriers. We complement these considerations with a sensitivity analysis in the Supplementary Material (Section 6.S.C), which examines the performance of the various NSUM approaches under deviations from random mixing.

6.4.2. Data Processing

Data cleaning is needed to remove inconsistent responses from people who claim to know an aberrant number of voters for a specific political party, or who report an abnormal number of blanks in their network. There are also respondents who report knowing the voting intention of a highly unusual number of contacts in their social network; such cases are more likely to result in *recall errors*, so we should account for them by implementing specific filtering to remove them before computing the NSUM estimation. Based on these considerations, we establish several rules for assessing and removing outliers and "dirty" information. The following data cleaning criteria are considered in advance:

- Removal of the responses at the top 20% of voting declaration network size $\sum_{u=1}^{Q} y_{iu}$.

This criterion results from an exploratory data analysis using a boxplot, a statistical visualization method that allows for outlier identification. Successive applications of the boxplot are performed at different atypical thresholds until we obtain a boxplot with no outliers. In all the cases under study, the resulting boxplots suggest the elimination of the top 20% network sizes.

- Removal of the responses declaring a network with blank voters above a specific threshold. The median absolute deviation (MAD) criterion, with a deviation of 5, is used to set the blank voting threshold.
- Removal of the responses declaring a majority (above 95%) of the network voting intention to a specific party.

Based on the previous rules, several scenarios are possible: the first one would be not applying any filtering criteria and processing all the survey responses. The second one would involve either using these rules separately or combining them to filter out surveys that may contain inconsistent responses. For the sake of completeness, the appendix with the Supplementary Material also includes tables (see Table 6.S.1) displaying the estimations that result from the application of NSUM procedures when no filtering is considered. Just to illustrate the number of surveys removed after data cleaning, we observe that the simultaneous application of the previous rules would result in 128 surveys.

On the other hand, a simple exploratory data analysis warns of potential estimation errors arising from *barrier effects* (see (Laga et al., 2021)). An illustrative example is shown in Figure 6.1, which shows contingency tables for the distribution of the age and the voting declaration of the network against the age of the respondents for a survey without any stratification deployed in the Madrid regional election process of May 2023 to estimate the voting intention. Here, we provide only contingency tables for the indirect survey data collected in one of the rounds deployed (see the Supplementary Material for more information about this survey).

6.4.3. Sampling Stratification for NSUM-based Estimation Procedures

In order to handle the biases that may arise from survey data collection, we propose one alternative for adapting NSUM procedures to the vote estimation problem. This alternative relies on the simple idea of introducing stratification (weights) into the problem. The sampling strategy assigns weights to the sampling scheme in order to upweight or downweight cases; we call this approach "sampling stratification".

Pollfish survey platform provides a service for sample stratification by age at a higher cost; the service allows stratification for a prefixed age ranking given by the platform, as shown in the first row of Table 6.2. The sampling proportions in the second row of the table were set on the basis of the population proportions provided by INE official statistics. This sampling stratification strategy is also applied to the designs of Andalusia and Valencia region surveys; the sampling proportions are

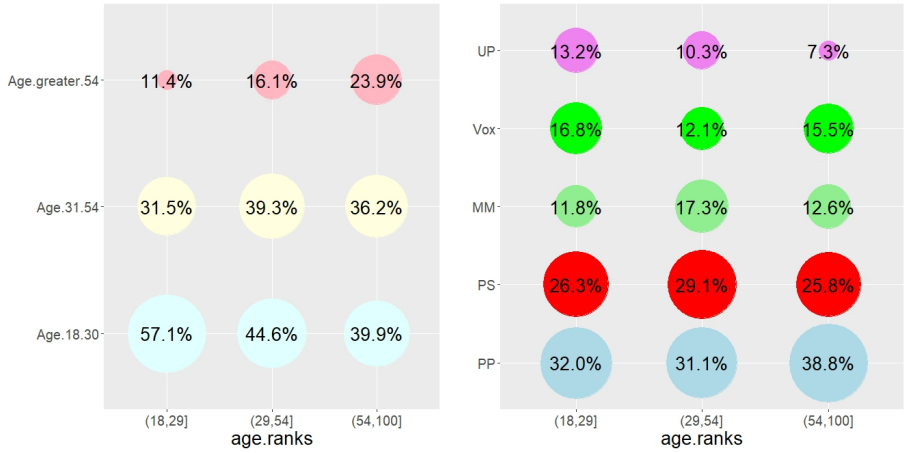


Figure 6.1: Age declaration from the network of respondents against the age of the respondent (left). Voting declaration from the network of respondents against the age of respondent (right) (Numbers in percentages). The parties' acronyms are UP-Unidas Podemos, Vox, MM-Más Madrid, PS-Partido Socialista de Madrid, and PP-Partido Popular.

shown in Table 6.2.

Madrid	Age	$[18, 24]$	$[25, 34]$	$[35, 44]$	$[45, 54]$	≥ 55
	Sampling prop (%)	8.78	12.32	15.66	20.23	43.01
Andalusia	Age	$[18, 24]$	$[25, 34]$	$[35, 44]$	$[45, 54]$	≥ 55
	Sampling prop (%)	9.32	12.98	16.43	19.64	41.64
Valencia	Age	$[18, 24]$	$[25, 34]$	$[35, 44]$	$[45, 54]$	≥ 55
	Sampling prop (%)	8.79	11.65	15.33	19.96	44.26

Table 6.2: Weights by age for the stratified sampling scheme by region

6.5. Results

The proposed methodology has been applied to datasets obtained from indirect surveys conducted in the communities of Madrid, Andalusia, and Valencia during the electoral campaigns of 2023 Spanish general elections. Indirect surveys were collected using a stratified sampling strategy, with stratification proportions based on age, as described in Table 6.2. Furthermore, the results of the various NSUM estimators have been extrapolated to the representative seats of each political party within the Spanish parliament for each region (Madrid, Andalusia, and Valencia). The allocation of representative seats for the political groups was determined using the D'Hondt method, which is explained in the Supplementary Material 6.S.B. Then, we compare the resulting seat estimations with those published by a famous survey company in Spain and the official estimations launched by the public reference center for sociological studies in Spain. The results are discussed separately for each region.

Method	Political parties				Ideological blocs		
	Vox	PP	PSOE	Sumar	PP+Vox	PSOE+Sumar	Blank
Naïve	16.1	35.1	28.1	16.2	51.2	44.3	4.6
MoS	14.9	37.8	27.5	16.1	52.7	43.6	3.7
MLE	15.4	36.3	29.1	16.4	51.7	45.5	2.8
PIMLE	14.6	39.2	27.6	14.7	53.9	43.4	3.8
Real outcome	**13.9**	**40.2**	**27.7**	**15.3**	**54.1**	**43.0**	**2.9**

Table 6.3: Spanish general elections, July 2023. Indirect survey collected in the Community of Madrid. Estimation of the percentage of votes by political group and blank votes after filtering out samples that contained a significant number of blank votes. Estimated percentage for the ideological blocs (conservative: PP+Vox and progressive: PSOE+Sumar).

6.5.1. Results for the General Elections (Madrid)

An online survey was conducted using the Pollfish platform to collect data for the Spanish general elections held on July 23, 2023. The survey was implemented in the Community of Madrid from July 13 to 21, concluding two days before election day. In particular, our study gathered information on voting intentions and auxiliary questions related to the closest contacts of 200 surveyed individuals. The English version of the questionnaire is available in the Supplementary Material, Section 6.S.A.1.

In the initial endeavor, and for illustrative purposes, we calculated voting intention rates by applying the different NSUM estimators detailed in Table 6.1 to the raw data. Table 6.S.1 in the Supplementary Material shows the estimated proportions for each NSUM estimator by political party. This table also displays percentages for the conservative (PP+Vox) and progressive (PSOE+Sumar) ideological blocs. This scenario serves as an example that justifies the need to carry out data processing steps. The table includes the percentages of blank votes alongside the official results in Madrid, suggesting that the blank vote is considerably overrepresented. More precisely, the official outcome for the blank vote is 2.9%, while the estimated rates obtained by the Naïve, MoS, MLE, and PIMLE methods are 10.4%, 14.8%, 21.8%, and 14.9%, respectively. Overestimation can be attributed to respondents who misinterpret what a Blank vote means. To improve our estimates, only the filter that removes responses declaring many blank voters above a specific threshold is implemented. The preprocessing stage yielded 151 valid responses, covering 4375 individuals, with an average of approximately 29.0 indirect votes per respondent.

Table 6.3 shows the estimated vote proportions for each political party, derived from applying NSUM methodologies following the aforementioned blank vote filtering process. Additionally, the table presents the estimated voting shares for the conser-

vative (PP+Vox) and progressive (PSOE+Sumar) blocs. Table 6.3 also displays the official results of the Spanish general elections for the Community of Madrid. Note that the estimated percentages are closer to the real outcomes than the estimations derived from raw survey data. For example, the blank voting proportion is 2.9%, and the percentages obtained by Naïve, MoS, MLE, and PIMLE estimators are 4.6%, 3.7%, 2.8%, and 3.8%, respectively. Furthermore, we can observe that PIMLE provides the closest estimates to the actual outcome. Notice that political preferences cannot be considered a "hidden population" in the strict sense. In contrast, these groups can exhibit structured patterns by homophily, geography, demographics, or ideology. To assess the effects of homophilious mixing on voting intention share estimation, the Supplementary Material in Section 6.S.C presents a sensitivity analysis of the NSUM estimators across different levels of homophily using synthetic graphs.

Finally, Table 6.4 shows the extrapolation of the NSUM results using the D'Hondt method (Bormann and Golder, 2013) to estimate the number of Madrid representative seats by political group in the Spanish national parliament. For the sake of completeness and comparison, the table also provides the results published by a famous company of surveys in Spain (GAD3, https://www.gad3.com) and the official forecast of the national public reference center for sociological studies in Spain (CIS, *Centro de Investigaciones Sociológicas*, https://www.cis.es/). The last two rows of Table 6.4 display the official results excluding (CERA excluded) and including (CERA included) the votes of citizens, registered in the Spanish electoral system, who reside in foreign countries (Censo Electoral de Residentes Ausentes, CERA). We can observe that the PIMLE estimator gives the most accurate seat estimates when compared to other polling results, such as GAD3 and CIS, as well as to other NSUM approaches. For instance, if we consider the PIMLE estimator and the true political party representation outcomes, the conservative bloc obtains 21 seats (PIMLE: PP + Vox = 16 +5, True: PP + Vox = 16 +5), and the progressive bloc obtains 16 seats (PIMLE: PSOE + Sumar = 11 + 5, True: PSOE + Sumar = 10 + 6).

Table 6.S.4 in the Supplementary Material presents the vote percentages and the 95% confidence intervals derived from NSUM estimators for the Community of Madrid. The confidence intervals were calculated using an approximate variance method (Killworth et al., 1998c), accounting for the stratified nature of our survey design. Specifically, we first computed the variance for each age group separately. Then, the overall population variance was estimated as a weighted sum of the variances, with the stratification rates in Table 6.2 serving as the weights. This table also includes the official results. This table indicates that most official results fall within the confidence intervals. The official results only fall outside the confidence intervals for the Naïve approach in the PP and blank options. Furthermore, Table 6.S.5 displays the vote percentages and the 95% confidence intervals obtained from 10,000 respondent-level stratified bootstrap replicates. In each replicate, respondents

Method	Voting percentage				Seats (D'Hondt method)			
	Vox	PP	PSOE	Sumar	Vox	PP	PSOE	Sumar
GAD3 (July 23)	NA	NA	NA	NA	5	17	10	5
CIS (June 23)	NA	NA	NA	NA	4-5	14-15	10-11	7-8
Naïve (July 23)	16.1	35.1	28.1	16.2	6	14	11	6
MoS (July 23)	14.9	37.8	27.5	16.1	6	15	11	5
MLE (July 23)	15.4	36.3	29.1	16.4	6	14	11	6
PIMLE (July 23)	14.6	39.2	27.6	14.7	5	16	11	5
Real (CERA excluded)	**13.9**	**40.2**	**27.7**	**15.3**	**5**	**15**	**11**	**6**
Real (CERA included)	**13.9**	**40.2**	**27.6**	**15.3**	**5**	**16**	**10**	**6**

Table 6.4: Spanish general elections, July 2023. Indirect survey collected in the Community of Madrid. Estimated percentages of votes by political group using the NSUM approaches after filtering out samples that contained a significant number of blank votes. Extrapolation of proportion estimates to representative seats using the D'Hondt method and comparison with other studies. Official results obtained from the Spanish general elections in Madrid, including (CERA included) and excluding (CERA excluded) the votes of Spanish nationals who reside in foreign countries.

are resampled within strata, and the full estimation pipeline is rerun, including the preprocessing and vote-share estimation. The resulting vote shares from the bootstrap analysis were then mapped to seats using the D'Hondt rule. Figure 6.S.3 shows the bootstrap distributions of seat allocations by party and NSUM estimator. Concretely, each panel depicts the empirical distribution of the number of seats obtained by a given party under a particular NSUM approach (rows: Naïve, MoS, MLE, and PIMLE; columns: PP, PSOE, Vox, and Sumar). Vertical dashed lines indicate the official outcome.

6.5.2. Results for the General Elections (Andalusia)

This study conducted an online survey in Andalusia using the Pollfish platform to gather information on voting intention for the Spanish general elections. Specifically, the data collection was performed from July 13 to 17, 2023. In this case, the preprocessing stage produced 137 valid responses, encompassing 2892 individuals, with an average of 21.11 contacts per survey participant. To evaluate the impact of the preprocessing steps on the vote share and seat allocation estimates, we present results across a grid of configurations, varying three dimensions: (i) the trimming level applied to the voting declaration network size (80%, 85%, 90%, and 95%), (ii) the threshold used in the median absolute deviation (MAD) rule (3, 4, and 5), and (iii) the inclusion or exclusion of the rule that removes respondents for whom more than 95% of reported alters support the same party. The results are presented in Tables 6.S.10 and 6.S.11 in the Supplementary Material for four estimation methods (Naïve, MoS, MLE, and PIMLE) and are reported in terms of both vote shares

Method	Political parties				Ideological blocs		
	Vox	PP	PSOE	Sumar	PP+Vox	PSOE+Sumar	Blank
Naïve	15.2	35.1	28.3	12.6	50.3	40.8	8.9
MoS	13.1	37.1	27.6	13.4	50.2	41.00	8.8
MLE	13.9	35.6	27.6	14.7	49.5	42.3	8.2
PIMLE	12.7	36.6	30.6	12.3	49.3	42.9	7.8
Real outcome	**15.3**	**36.4**	**33.5**	**12.0**	**51.7**	**45.5**	**2.8**

Table 6.5: Spanish general elections, July 2023. Indirect survey collected in Andalusia. Estimation of the percentage of votes by political group and blank votes after filtering out responses by using all data cleaning criteria. Estimated percentage for the ideological blocs (conservative: PP+Vox and progressive: PSOE+Sumar).

and implied seat allocations. Table 6.5 displays the percentage voting estimations obtained by the NSUM estimators for each political party after applying the data processing criteria described in Section 6.4.2. We can observe that the Naïve method outperforms the others in estimating the conservative political options. In contrast, the PIMLE approach seems better for estimating the progressive political options and the PP while underestimating Vox results.

The extrapolation of the estimated vote proportions to the national parliament seats from Andalusia using the D'Hondt method is given in Table 6.6. The official results, both with and without the CERA vote, are provided for completeness; although, in this case, the CERA vote did not change the final representation. The NSUM-based estimates of parliamentary seats are compared with the results published by GAD3 and CIS during the campaign. Overall, we can observe that NSUM outcomes are competitive with GAD3 and CIS estimations, with the MLE method given the most accurate results.

Tables 6.S.6 and 6.S.7 in the Supplementary Material present the percentage voting estimates along with 95% confidence intervals using the approximate variance and bootstrap methods, respectively. Moreover, Figure 6.S.4 displays the bootstrap distributions of seat allocations by party and NSUM technique. In this case, the official results of the political parties and ideological blocs lie within the confidence intervals. However, the NSUM operators significantly overrepresent the blank voting option with confidence intervals larger than official outcomes.

6.5.3. Results for the General Elections (Valencia)

For our third use case, we deployed an online survey in the Valencian Community using the Pollfish platform to gather information on voting intention for the general elections of July 23, 2023. This survey was deployed from July 13 to 21, 2023; it

Method	Voting percentage				Seats (D'Hondt method)			
	Vox	PP	PSOE	Sumar	Vox	PP	PSOE	Sumar
GAD3 (July 23)	NA	NA	NA	NA	6	29-31	21-23	3
CIS (June 23)	NA	NA	NA	NA	5-7	27-28	18-20	8
Naïve (July 23)	15.2	35.1	28.5	12.3	10	24	19	8
MoS (July 23)	13.1	37.1	27.6	13.4	9	25	18	9
MLE (July 23)	13.9	35.6	27.6	14.7	10	24	18	9
PIMLE (July 23)	12.7	36.6	30.6	12.3	8	25	20	8
Real (CERA excluded)	**15.3**	**36.4**	**33.5**	**12.0**	**9**	**25**	**21**	**6**
Real (CERA included)	**15.3**	**36.4**	**33.5**	**12.0**	**9**	**25**	**21**	**6**

Table 6.6: Spanish general elections, July 2023. Indirect survey collected in Andalusia. Estimated percentages of votes by political group using the NSUM approaches after filtering out responses using all data cleaning criteria. Extrapolation of proportion estimates to representative seats using the D'Hondt method and comparison with other studies. Official results obtained from the Spanish general elections in Andalusia, including (CERA included) and excluding (CERA excluded) the votes of Spanish nationals who reside in foreign countries.

was launched for individuals who live in this Community as a sampling framework. The Pollfish platform provided 134 valid responses from the data preprocessing stage, with 1872 potential individuals and 13.97 votes per respondent. In the case of Valencia, the results are shown in the Tables 6.7 and 6.8. The NSUM methods generally understate the progressive option with respect to the conservative one (see Table 6.7). The method providing the closest results to the real outcome is the PIMLE, followed by the MoS. Regarding the extrapolation of the estimates to determine the representative seats using the D'Hondt method, we can see in Table 6.8 that CERA inclusion or exclusion does not affect the distribution of the seats in Valencia. Additionally, the NSUM methods correctly estimate the seats of the conservative (Vox+PP) and progressive (PSOE+SUMAR) blocs, with the GAD3 and CIS surveys failing in such estimations, the latter one giving a different majority than the one provided by the real outcome.

Finally, in the Supplementary Material, Tables 6.S.8 and 6.S.9 show the estimated voting intention alongside the 95% confidence intervals with the approximate variance and bootstrap approaches. The bootstrap distributions of seat allocations by party and NSUM technique are depicted in Figure 6.S.5. The official results fall within the confidence interval for nearly all methods and political parties, except for the estimated percentages obtained by the Naïve method for PSOE. Additionally, similar to the results in Andalusia, the NSUM approaches overrepresent the blank voting option, with confidence intervals larger than the official results.

Method	Political parties				Ideological blocs		
	Vox	PP	PSOE	Sumar	PP+Vox	PSOE+Sumar	Blank
Naïve	16.5	31.8	28.1	14.3	48.3	42.4	9.4
MoS	16.6	32.2	28.3	14.3	48.8	42.6	8.6
MLE	15.6	32.4	27.7	14.1	48.0	41.8	10.3
PIMLE	16.5	32.3	27.7	15.0	48.8	42.7	8.6
Real outcome	**15.7**	**34.9**	**32.1**	**15.2**	**50.6**	**47.3**	**2.2**

Table 6.7: Spanish general elections, July 2023. Indirect survey collected in the Valencian Community. Estimation of the percentage of votes by political group and blank votes after filtering out respondents at the top 20% network size who also declare a network voting intention for a specific party above 95%. Estimated percentage for the ideological blocs (conservative: PP+Vox and progressive: PSOE+Sumar).

Method	Voting percentage				Seats (D'Hondt method)			
	Vox	PP	PSOE	Sumar	Vox	PP	PSOE	Sumar
GAD3 (July 23)	NA	NA	NA	NA	5	14	11	3
CIS (June 23)	NA	NA	NA	NA	4	11	11	7
Naïve (July 23)	16.5	31.8	28.1	14.3	6	12	10	5
MoS (July 23)	16.0	32.1	28.8	15.0	6	12	10	5
MLE (July 23)	15.6	32.4	27.7	14.1	6	12	10	5
PIMLE (July 23)	16.5	32.3	27.7	15.0	6	12	10	5
Real (CERA excluded)	**15.7**	**34.9**	**32.1**	**15.2**	**5**	**13**	**11**	**4**
Real (CERA included)	**15.7**	**34.9**	**32.1**	**15.2**	**5**	**13**	**11**	**4**

Table 6.8: Spanish general elections, July 2023. Indirect survey collected in the Valencian Community. Estimated percentages of votes by political group using the NSUM approaches after filtering out respondents at the top 20% network size who also declare a network voting intention for a specific party above 95%. Extrapolation of proportion estimates to representative seats using the D'Hondt method and comparison with other studies. Official results obtained from the Spanish general elections in the Valencian Community, including (CERA included) and excluding (CERA excluded) the votes of Spanish nationals who reside in foreign countries.

6.6. Discussion on the Use of the Methodology in the Case Study

In this section, we analyze the use of our methodology in the Case Study to summarize the key findings from its application. Figure 6.2 contains the steps of the methodology followed in the Case Study, along with the specific tasks/decisions performed in each step. These steps are detailed below.

6.6.1. Survey Design and Data Collection

In this first step, the following tasks were performed. First, the platform for conducting the surveys was chosen. In our case, the Pollfish platform was chosen because it had broad coverage of the regions under study. Furthermore, this platform allows for the collection of demographic data such as the age of participants.

For the design of the survey, the following method was used. Each participant was asked: "How many of your contacts will vote for this party for the Spanish general election?" We then added up the responses for each party to obtain the number of contacts. Thus, we defined the network based on the contacts whose voting intentions are known to the participants. This approach minimizes transmission error, since in this case, all contacts have known voting intentions, such that the numerator and denominator refer to the same network. This approach does not compromise the estimates, as long as the hidden subpopulations are evenly distributed in the network (Laga et al., 2021; Díaz-Aranda et al., 2025). Furthermore, there is another source of transmission error that can affect estimation outcomes, which is that participants misjudge their contacts' voting intentions, a bias that our framework does not correct for. However, given the overall proximity of our estimates to the actual results, this suggests that the influence of this bias is not significant.

In turn, in the first survey, we observed a barrier effect because younger people were more likely to know people their own age, who in turn were more likely to vote for specific parties. Age stratification is used to mitigate this effect. Although we know that age stratification on its own cannot make contact composition representative, we use it to reduce the impact of undesirable biases.

Specifically, to obtain a stratified representative sample, age was used in the fourth task because it was provided by the tool, and we could easily link it to INE official statistics. If we had another demographic variable with official statistical values, and we could obtain that information from the respondents through the survey platform, such variables could also be used for stratification. Our methodology is not tied to age as a demographic variable for stratification.

Based on everything indicated above, the surveys were collected through the Pollfish platform.

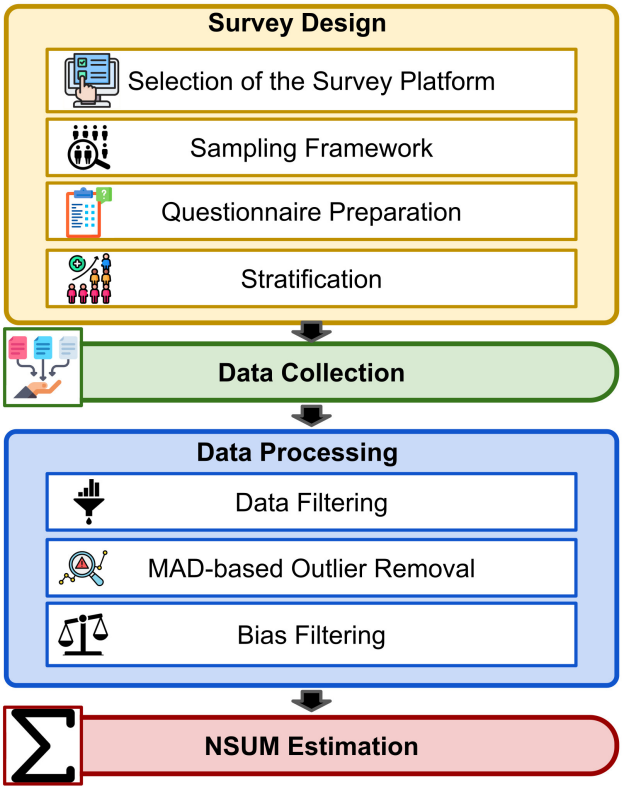


Figure 6.2: Flowchart of the methodology

6.6.2. Data Processing

Three tasks were used for data preparation. The first task was to eliminate 20% of the cases that report the highest number of contacts' voting intention in their social network, a threshold based on an empirical analysis which uses box plots as mechanisms to detect outliers (Figure 6.3 shows an example of the box plot for Madrid, with outliers (left) and without outliers (right)). While the box plot method eliminated 16.6% of the responses in the Madrid survey, 20% in Andalusia, and 18.7% in Valencia, a conservative unified criterion was used, and 20% was applied to all. This filter seeks to mitigate the effect of responses with disproportionately large network sizes, which could be due to declaration errors, ambiguous interpretation of the question, or unrepresentative extreme cases that distort the overall analysis. This allows the analysis to focus on the 80% of the responses that, due to their size, are presumably more representative of the phenomenon under study. The declarations' extremely large voter biases are often outliers that can exert a disproportionate influence on the aggregate results, such as by skewing the mean, reducing representativeness, or introducing "noise" or atypical cases that do not reflect the general behavior being studied.

The second task was to eliminate responses that reported a network with blank voters above a threshold defined by MAD equal to 5. This method allows for the detection of observations that deviate significantly from the general behavior of the

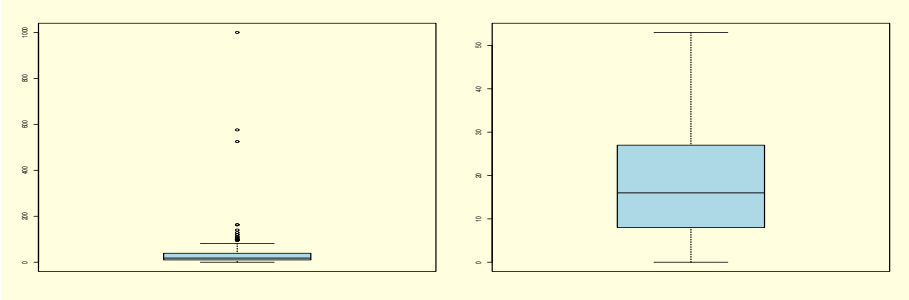


Figure 6.3: Boxplots of declared network size with (left) and without (right) outliers for Madrid surveys.

sample. By eliminating these cases, the analysis focuses on networks where there is discernible partisan voting intention, improving the clarity and relevance of the results for the specific objective of studying the distribution of votes between parties. The application of this filter ensures the quality and consistency of the results, minimizing the impact of responses that are potentially unrepresentative, erroneous, or influenced by distorted perceptions of the respondent's social environment. A high percentage of blank voters within a network can indicate several situations that can distort the analysis of partisan voting intention, such as: (i) Protest or widespread discontent: This could reflect a protest against the current political system, rather than a preference for a particular partisan option, which does not help in the analysis of the distribution of voting intentions among parties and introduces "noise". (ii) Lack of decision or extreme indecision: The network could be composed of individuals who have not yet decided on their vote or who do not feel represented by any partisan option, which is meaningless in an analysis of voting intentions. (iii) Problems in understanding the question: This could indicate that participants did not understand the question. (iv) Bias in the composition of the network: It is possible that the network is a subgroup of individuals who tend to vote blank in a greater proportion than the general population, which could bias the results.

In the third task, responses stating that more than 95% of their network would vote for the same party were eliminated, as these types of statements may indicate biased, exaggerated, or unrepresentative responses of the actual social environment. Such responses could distort the analysis, as they do not reflect the typical diversity of opinions on real social networks, thus compromising the validity of the results. This threshold mitigates potential social desirability biases, extreme information bubbles (echo chambers where information within the network may be highly biased toward one political perspective), or errors in self-perception of the political environment (leading to an unrepresentative sample).

6.6.3. NSUM Estimation

The final stage consists of applying the different NSUM methods, then selecting the one that gave the best results for comparison purposes. Furthermore, in this case, the D'Hondt method was applied to determine the number of representatives per party, as it is the method of electoral representation used by the Spanish electoral system.

6.7. Conclusions

This paper introduces a comprehensive methodology for estimating the voting outcomes of electoral processes using aggregated relational data (ARD) from indirect surveys. Our proposal encompasses a unified approach that extends from the design of the indirect survey to the computation of vote estimations by utilizing several NSUM-based procedures. In essence, the proposed methodology contains three main stages: the first one focuses on designing the indirect survey to collect information on voting intention. Notice that this first step considers demographic information to enhance data reliability. Secondly, an ARD data processing step is addressed, with different filtering criteria being proposed and implemented for data cleaning. Finally, our approach estimates voting intention percentages by means of NSUM procedures. The proposed methodology was assessed to estimate the proportions of voting intentions and parliamentary seats in three Spanish regions – namely, the Community of Madrid, Andalusia, and the Valencian Community – for the general elections held on July 23, 2023. The results obtained are promising because they improved on the results reported by various direct surveys in all cases studied. Moreover, our proposal of using indirect surveys and NSUM to estimate the outcomes of elections has the advantage of reducing the costs of standard direct surveys, so it is expected to draw the attention of private and public poll organizations, providing an alternative to their survey planning.

The use cases developed for the general 2023 elections in Madrid, Andalusia, and Valencia demonstrate the potential of applying the proposed NSUM-based approach. In these cases, the survey was designed in accordance with a sample stratification scheme by age, with strata proportional to the ages of the voting adult population. When the proposed NSUM-based procedures are applied to estimate party voting proportions, their outputs provide promising results compared with those provided by public and private polling institutions, based on direct reporting surveys. In some instances, the proposed methodology highly surpasses traditional direct reporting estimations. Finally, it is worth noting the sample sizes for all the Spanish territory of GAD3 ($n = 3502$) and CIS ($n = 29201$) surveys. Thus, for example, in the case of GAD3, the number of samples from the smallest region considered (Community of Valencia, with a voting population of slightly more than 3.6 million, out of a Spanish

voting population of 35 million) can be estimated to be at least $3502 \times 3.6/35 = 360$, almost twice our sample size. In the case of CIS, these sample sizes are even larger.

One thing learned from this study is that sampling stratification may have an impact on NSUM resulting outcomes. To illustrate this, we conducted an initial analysis that did not consider stratification when collecting samples and then used weighted estimation (post-stratification) as an alternative method. This initial methodology was applied to the autonomic elections that took place in Madrid on May 28th, 2023 (regional elections). The application of each step of the post-stratification methodology is described in the Supplementary Material 6.S.A.3. The resulting estimations are worse than those obtained by applying the proposed sampling stratification strategy. Hence, we advocate the use of sampling stratification from the very beginning as it provides more accurate results. Summing up, the combination of sampling stratification and NSUM-based approaches provides better results than those obtained by GAD3 and CIS survey institutions from direct reporting and the ones obtained with the NSUM without sampling stratification. The previous findings open the door to future research, which would address a deep analysis of different sampling strategies to choose the strategy that guarantees a representative sample for any electoral process (see Section 3.1, where we make a first comment on possible strategies). In particular, future work will evaluate alternative stratification sampling strategies using other demographic variables, which would be possible provided that these variables are available both in the survey platform and in the official statistics. Additionally, NSUM estimators are affected by the homophily among supporters of different political parties. Another direction for future work is the estimation of homophily using ARD and the development of corrections to the NSUM estimators.

Although indirect surveys have advantages in terms of implementation costs and the protection of respondents and their contacts, our proposal has some limitations. The first drawback concerns the use of technological platforms to design and collect indirect surveys, which usually under-represent populations unfamiliar with digital environments or over-represent people accustomed to digital survey tools. More research is needed on sampling design to address this technological bias. A second limitation is that, although our methodology has been tested in different regions of Spain, such as Madrid, Valencia, and Andalusia, and the results are encouraging, we believe more field results are necessary in other countries and electoral contexts (such as the European elections) to evaluate the validity and reproducibility of the proposal. Hence, future work should include applying the methodology to other electoral processes, exploring in detail the role played by different sampling strategies, survey design, and weighting estimation mechanisms in the implementation of the NSUM and the resulting outcomes when deployed within different electoral systems. The third issue for future research would involve the study of alternative data processing filters that could be incorporated into the methodology to improve the accuracy of the estimations. Another limitation is the assumption we make when we

consider that respondents know the voting intentions of their circle of acquaintances and that they do not incur any political bias due to their voting intentions or due to uncertain knowledge about the voting intentions of their contacts (transmission error). This may be one of the causes of the overestimation of blank votes. Future work should analyze these potential biases in depth.

Ethical Approval

Participants were recruited through Pollfish and provided informed consent prior to participation. The consent explained the purpose of the study and the voluntary nature of participation. To minimize potential risks for respondents, we have designed the survey such that personal information is not required, asking questions solely about the contacts of the participants. It is worth noting that this information cannot identify users. However, captured data is processed by Pollfish as personal data according to the EU regulation for the protection of personal data (GPDR), as can be seen on its webpage, and according to the laws applicable to the users of this platform. Pollfish also claims to comply with applicable data protection law when receiving personal data from participants, i.e., respondents have been informed about Pollfish's privacy terms, and the compliance with the requirements of applicable law since they signed up for the Pollfish platform and agreed to the terms of use (see https://www.pollfish.com/terms/respondent). The participants have also been informed about the data, how they use it, where it goes, and how long it is kept. The raw data available in Pollfish and provided to the researchers includes, in addition to the responses to the survey questions, a number of other attributes of each participant. We did not use the raw data, but instead we extracted the following attributes provided by Pollfish: gender, year of birth, education (5 options), income (8 intervals), and employment status (10 options). The raw data were immediately deleted after this process. Data were stored on institutional secure storage, accessible only to the research team. Ethics approval was obtained from the Institutional Review Board (IRB) of the IMDEA Networks Institute, which approved this research in May and July 2023 for the regional and general elections, respectively.

Acknowledgement

This work was supported by grants TED2021-131264B-I00 (SocialProbing) and PID2019-104901RB-I00, funded by the Spanish Ministry of Science and Innovation - State Research Agency, Spain MCIN/AEI/10.13039/ 501100011033 and the European Union "NextGenerationEU"/PRTR. The research is part of the I+D+i project PID2022-137243OB-I00 founded by MCIN. JMA is supported by NextGenerationEU funds provided through the Spanish Ministry of Universities (Ministerio de Universidades, Gobierno de España). The work of SDA has been funded by the Comunidad

de Madrid predoctoral grant PIPF-2022/COM-24467.

The authors would like to thank the reviewers for their useful comments that helped to improve the draft version of this work.

Data availability

Data presented in this paper in aggregated form, as well as the codes used to process the data, can be provided by the authors upon a reasonable request.

References

H. R. Bernard, E. C. Johnsen, P. D. Killworth, and S. Robinson. Estimating the size of an average personal network and of an event subpopulation. In M. Kochen, editor, The Small World, pages 159–175. Ablex Publishing Corporation, Norwood, NJ, 1989.

N.-C. Bormann and M. Golder. Democratic electoral systems around the world, 1946–2011. Electoral Studies, 32(2):360–369, 2013.

E. Breza, A. G. Chandrasekhar, T. H. McCormick, and M. Pan. Consistently estimating graph statistics using aggregated relational data. arXiv preprint arXiv:1908.09881, 1, 2019.

E. Breza, A. G. Chandrasekhar, T. H. McCormick, and M. Pan. Using aggregated relational data to feasibly identify network structure without network data. American Economic Review, 110(8):2454–84, 2020.

E. Breza, A. G. Chandrasekhar, S. Lubold, T. H. McCormick, and M. Pan. Consistently estimating network statistics using aggregated relational data. Proceedings of the National Academy of Sciences, 120(21):e2207185120, 2023.

W. Bruine de Bruin, M. Galesic, R. Bååth, J. de Bresser, L. Hall, P. Johansson, T. Strandberg, and A. van Soest. Asking about Social Circles Improves Election Predictions Even with Many Political Parties. International Journal of Public Opinion Research, 34(1):edac006, 04 2022. ISSN 1471-6909. doi: 10.1093/ijpor/edac006. URL https://doi.org/10.1093/ijpor/edac006.

G. Carletti, N. Soriani, M. Mattiazzi, and D. Gregori. Suppl-1, m6: A social network approach to the estimation of perceived quality of health care. The Open Nursing Journal, 11:219, 2017.

R. de Geus, E. Ralph-Morrow, and R. Shorrocks. Understanding ambivalent sexism and its relationship with electoral choice in britain. British Journal of Political Science, 52(4):1564–1583, 2022. doi: 10.1017/S0007123421000612.

S. Díaz-Aranda, J. Aguilar, J. M. Ramírez, D. Rabanedo, A. Fernández Anta, and R. E. Lillo. Performance analysis of nsum estimators in social-network topologies. The American Statistician, 79(2):247–264, 2025.

S. Ezoe, T. Morooka, T. Noda, M. L. Sabin, and S. Koike. Population size estimation of men who have sex with men through the network scale-up method in japan. PloS one, 7(1):e31184, 2012.

D. M. Feehan and M. J. Salganik. Generalizing the network scale-up method: a new estimator for the size of hidden populations. Sociological methodology, 46(1): 153–186, 2016.

D. M. Feehan, A. Umubyeyi, M. Mahy, W. Hladik, and M. J. Salganik. Quantity versus quality: A survey experiment to improve the network scale-up method. American journal of epidemiology, 183(8):747–757, 2016.

A. Fernández-Roldán and M. Barnfield. Voters share polls that say what they want to hear: Experimental evidence from spain and the usa. International Journal of Public Opinion Research, 36(4):edae047, 2024.

M. Galesic, W. Bruine de Bruin, M. Dumas, A. Kapteyn, J. Darling, and E. Meijer. Asking about social circles improves election predictions. Nature Human Behaviour, 2(3):187–193, 2018.

A. Garcia-Agundez, O. Ojo, H. A. Hernández-Roig, C. Baquero, D. Frey, C. Georgiou, M. Goessens, R. E. Lillo, R. Menezes, N. Nicolaou, et al. Estimating the covid-19 prevalence in spain with indirect reporting via open surveys. Frontiers in Public Health, 9:658544, 2021.

A. Graefe. Accuracy of Vote Expectation Surveys in Forecasting Elections. Public Opinion Quarterly, 78(S1):204–232, 06 2014. ISSN 0033-362X. doi: 10.1093/poq/nfu008. URL https://doi.org/10.1093/poq/nfu008.

A. Graefe. Forecasting proportional representation elections from non-representative expectation surveys. Electoral Studies, 42:222–228, 2016. ISSN 0261-3794. doi: https://doi.org/10.1016/j.electstud.2016.03.001. URL https://www.sciencedirect.com/science/article/pii/S0261379416300798.

P. Habecker, K. Dombrowski, and B. Khan. Improving the network Scale-Up estimator: Incorporating means of sums, recursive back estimation, and sampling weights. PLoS One, 10(12):e0143406, 2015a.

P. Habecker, K. Dombrowski, and B. Khan. Improving the network scale-up estimator: Incorporating means of sums, recursive back estimation, and sampling weights. PloS one, 10(12):e0143406, 2015b.

E. Hernández, E. Anduiza, and G. Rico. Affective polarization and the salience of elections. Electoral Studies, 69:102203, 2021. ISSN 0261-3794. doi: https://doi.org/10.1016/j.electstud.2020.102203. URL https://www.sciencedirect.com/science/article/pii/S026137942030086X.

L. Jing, Q. Lu, Y. Cui, H. Yu, and T. Wang. Combining the randomized response technique and the network scale-up method to estimate the female sex worker population size: an exploratory study. Public Health, 160:81–86, 2018.

C. Kadushin, P. D. Killworth, H. R. Bernard, and A. A. Beveridge. Scale-up methods as applied to estimates of heroin use. Journal of Drug Issues, 36(2):417–440, 2006.

P. Killworth, E. Johnsen, C. McCarty, G. Shelley, and H. Bernard. A social network approach to estimating seroprevalence in the united states. Social Networks, 20 (1):23–50, 1998a.

P. D. Killworth, E. C. Johnsen, C. McCarty, G. A. Shelley, and H. R. Bernard. A social network approach to estimating seroprevalence in the united states. Social networks, 20(1):23–50, 1998b.

P. D. Killworth, C. McCarty, H. R. Bernard, G. A. Shelley, and E. C. Johnsen. Estimation of seroprevalence, rape, and homelessness in the united states using a social network approach. Evaluation Review, 22(2):289–308, 1998c.

P. D. Killworth, C. McCarty, H. R. Bernard, G. A. Shelley, and E. C. Johnsen. Estimation of seroprevalence, rape, and homelessness in the united states using a social network approach. Evaluation review, 22(2):289–308, 1998d.

I. Laga, L. Bao, and X. Niu. Thirty years of the network scale-up method. Journal of the American Statistical Association, 116(535):1548–1559, 2021.

I. Laga, L. Bao, and X. Niu. networkscaleup: Network Scale-Up Models for Aggregated Relational Data, 2022. URL https://CRAN.R-project.org/package=networkscaleup. R package version 0.1-1.

I. Laga, L. Bao, and X. Niu. A correlated network scale-up model: Finding the connection between subpopulations. Journal of the American Statistical Association, pages 1–10, 2023.

R. Maltiel, A. E. Raftery, T. H. McCormick, and A. J. Baraff. Estimating population size using the network scale up method. The annals of applied statistics, 9(3): 1247, 2015.

C. McCarty, P. D. Killworth, H. R. Bernard, E. C. Johnsen, and G. A. Shelley. Comparing two methods for estimating network size. Human organization, 60(1): 28–39, 2001.

T. H. McCormick. The network scale-up method. The Oxford Handbook of Social Networks, page 153, 2020.

T. H. McCormick and T. Zheng. Latent demographic profile estimation in hard-to-reach groups. The annals of applied statistics, 6(4):1795, 2012.

T. H. McCormick and T. Zheng. Latent surface models for networks using aggregated relational data. Journal of the American Statistical Association, 110(512):1684–1695, 2015.

A. E. Murr. "Wisdom of crowds"? A decentralised election forecasting model that uses citizens' local expectations. Electoral Studies, 30(4):771–783, 2011. ISSN 0261-3794. doi: https://doi.org/10.1016/j.electstud.2011.07.005. URL https://www.sciencedirect.com/science/article/pii/S0261379411000977.

A. Nikfarjam, M. Shokoohi, A. Shahesmaeili, A. A. Haghdoost, M. R. Baneshi, S. Haji-Maghsoudi, A. Rastegari, A. A. Nasehi, N. Memaryan, and T. Tarjoman. National population size estimation of illicit drug users through the network scale-up method in 2013 in iran. International Journal of Drug Policy, 31:147–152, 2016.

J. Parsons, X. Niu, and L. Bao. A bayesian hierarchical model for combining multiple data sources in population size estimation. The Annals of Applied Statistics, 16 (3):1550–1562, 2022.

M. Salganik, D. Fazito, M. Mello, et al. Estimating the number of heavy drug users in curitiba, brazil using multiple methods. Technical report, Technical report. UNAIDS, 2010.

H. Schmitt, A. Sanz, D. Braun, and E. Teperoglou. It all happens at once: Understanding electoral behaviour in second-order elections. Politics and Governance, 8 (1):6–18, 2020. doi: 10.17645/pag.v8i1.2513.

J. Schroedel, A. Berg, J. Dietrich, and J. M. Rodriguez. Political trust and native american electoral participation: An analysis of survey data from nevada and south dakota. Social Science Quarterly, 101(5):1885–1904, 2020. doi: https://doi.org/10.1111/ssqu.12840.

L. D. Sio and T. Weber. Issue yield, campaign communication, and electoral performance: a six-country comparative analysis. West European Politics, 43(3): 720–745, 2020. doi: 10.1080/01402382.2019.1655968.

M. M. Skoric, J. Liu, and K. Jaidka. Electoral and public opinion forecasts with social media data: A meta-analysis. Information, 11(4), 2020. doi: 10.3390/info11040187. URL https://www.mdpi.com/2078-2489/11/4/187.

S. Snidero, N. Soriani, I. Baldi, F. Zobec, P. Berchialla, and D. Gregori. Scale-up approach in cati surveys for estimating the number of foreign body injuries in the aero-digestive tract in children. International journal of environmental research and public health, 9(11):4056–4067, 2012.

E. Sully, M. Giorgio, and S. Anjur-Dietrich. Estimating abortion incidence using the network scale-up method. Demographic Research, 43:1651–1684, 2020.

A. K. J. Teo, K. Prem, M. I. C. Chen, A. Roellin, M. L. Wong, H. H. La, and A. R. Cook. Estimating the size of key populations for HIV in singapore using the network scale-up method. Sex Transm Infect, 95(8):602–607, 2019.

UNAIDS. Key populations atlas - data sources. 2017.

A. M. Verdery, S. Weir, Z. Reynolds, G. Mulholland, and J. K. Edwards. Estimating hidden population sizes with venue-based sampling: extensions of the generalized network scale-up estimator. Epidemiology (Cambridge, Mass.), 30(6):901, 2019.

J. Wang, Y. Yang, W. Zhao, H. Su, Y. Zhao, Y. Chen, T. Zhang, and T. Zhang. Application of network scale up method in the estimation of population size for men who have sex with men in shanghai, china. PloS one, 10(11):e0143118, 2015.

T. Zheng, M. J. Salganik, and A. Gelman. How many people do you know in prison? Journal of the American Statistical Association, 101(474):409–423, 2006.

Supplementary Data

6.S.A. Supplementary material

6.S.A.1. Questionnaire for the General Election in Madrid, Andalusia and Valencia

- **Q1. Scientific Study Information**
  We are a team of computer and statistics researchers. We collect data to estimate the voting intention in different Spain regions, with the goal of improving statistical estimation techniques. The following survey contains questions regarding the voting intention of the people that you know for the Spanish general election on July 23, 2023. These contacts can include family, friends, and colleagues whose voting intention you know.
  - Responses are anonymous. The researchers cannot identify the participants.
  - Data cannot be modified or deleted once the survey is completed.
  - The data collected may be made public in aggregated form under the *Open Data Commons Attribution License*.
  - The results of this study will not be published before the end of the Spanish general elections on July 23, 2023.

  You can participate in the survey even if you have answered it before.
  Are you willing to take part in the following survey?
  □ Yes, I'm above the age of 18, and I'm giving you my consent.
  □ No, I do not wish to take part.
  Consider only your contacts of voting age in Spain's general election on 23 July 2023.
- **Q2.** How many of your contacts will vote for PP (Partido Popular) for the Spanish general election?

Q3. How many of your contacts will vote for PSOE (Partido Socialista Obrero Español) for the Spanish general election?
Q4. How many of your contacts will vote for Vox for the Spanish general election?
Q5. How many of your contacts will vote for Sumar for the Spanish general election?
Q6. How many of your contacts will vote for other parties or blank for the Spanish general election?
Q7. How many of your contacts will not vote or you do not know their vote for the Spanish general election?
Q8. How many of your contacts considered above are between 18 and 34 years old?
Q9. How many of your contacts considered above are between 35 and 54 years old?
Q10. How many of your contacts considered above are over 54 years old?
Q11. How many of your contacts considered above are medical doctors?
Q12. How many of your contacts considered above are autonomous workers?
Q13. How many of your contacts considered above are unemployed?

Method	Political parties				Ideological blocs		
	Vox	PP	PSOE	Sumar	PP+Vox	PSOE+Sumar	Blank
Naïve	15.4	32.1	27.0	15.2	47.5	42.2	10.4
MoS	14.0	31.2	25.5	14.5	45.2	40.0	14.8
MLE	14.4	31.7	20.6	11.5	46.1	32.1	21.8
PIMLE	13.6	32.8	25.0	13.7	46.4	38.7	14.9
Real outcome	**13.9**	**40.2**	**27.7**	**15.3**	**54.1**	**43.0**	**2.9**

Table 6.S.1: Spanish general elections, July 2023. Indirect survey collected in the Community of Madrid. Estimation of the percentage of votes by political group and blank votes from the raw dataset. Estimated percentage for the ideological blocs (left-wing: PSOE+Sumar and right-wing: PP+Vox).

6.S.A.2. Questionnaire for the Regional Election in Madrid

- **Q1. Scientific Study Information**
  We are a team of computer and statistics researchers. We collect data to estimate the **voting intention** in different Spain regions, with the goal of improving statistical estimation techniques. The following survey contains questions regarding the voting intention of the people that you know. These contacts can include **family, friends, and colleagues** whose voting intention you know.
  - Responses are anonymous. The researchers cannot identify the participants.

– Data cannot be modified nor deleted once the survey is completed.
– The data collected may be made public in aggregated form under the *Open Data Commons Attribution License*.
– The results of this study will not be published before the end of the regional elections in Spain on May 28, 2023.

Are you willing to take part in the following survey?
□ Yes, I'm above the age of 18, and I'm giving you my consent.
□ No, I do not wish to take part

Note: 50% of the participants will read Text A and the other 50% will read Text B

Text A

Consider only your contacts who are of voting age in the Community of Madrid.

Text B

Consider only your 20 closest contacts who are of voting age in the Community of Madrid.

- **Q2.** How many of your contacts **have not decided their vote** for the election to the Assembly of the Community of Madrid?
- **Q3.** How many of your contacts will vote for **PP (Partido Popular)** for the election to the Assembly of the Community of Madrid?
- **Q4.** How many of your contacts will vote for **Más Madrid** for the election to the Assembly of the Community of Madrid?
- **Q5.** How many of your contacts will vote for **PSOE-M (Partido Socialista Obrero Español de la Comunidad de Madrid)** for the election to the Assembly of the Community of Madrid?
- **Q6.** How many of your contacts will vote for **Vox** for the election to the Assembly of the Community of Madrid?
- **Q7.** How many of your contacts will vote for **Unidas Podemos** for the election to the Assembly of the Community of Madrid?
- **Q8.** How many of your contacts **will not vote** for the election to the Assembly of the Community of Madrid?
- **Q9.** How many of your contacts **will vote blank** for the election to the Assembly of the Community of Madrid?
- **Q10.** How many of your contacts considered above are **between 18 and 30 years old**?
- **Q11.** How many of your contacts considered above are **between 31 and 54 years old**?
- **Q12.** How many of your contacts considered above are **over 54 years old**?
- **Q13.** How many of your contacts considered above are **medical doctors**?
- **Q14.** How many of your contacts considered above are **autonomous workers**?
- **Q15.** How many of your contacts considered above are **unemployed**?

Method	Political parties					Ideological blocs		
	Vox	PP	PSM	MM	UP	PP+Vox	PSM+MM+UP	BLANK
Naïve	16.3	33.3	21.3	11.8	9.7	49.6	42.8	7.6
MoS	16.3	36.5	19.8	10.6	10.0	52.8	40.4	6.8
MLE	12.4	29.8	17.7	19.1	5.6	42.2	42.4	15.4
PIMLE	15.8	35.6	19.8	11.1	10.6	51.4	41.5	7.1
Real outcome	**7.3**	**47.3**	**18.2**	**18.4**	**4.7**	**54.6**	**41.3**	**4.1**

Table 6.S.2: Regional elections: Indirect survey collected in the Community of Madrid. Estimation of the percentage of votes by political group and blank votes from the raw dataset. Estimated percentage for the ideological blocs (left-wing: PSM+MM+UP and right-wing: PP+Vox).

6.S.A.3. Results of the Regional Election in Madrid

Survey design and data collection

Below, we describe how the information has been obtained and how the survey has been designed for this electoral process.

- Three rounds of surveys were conducted on the third, second, and last week before the voting day for the regional elections.
- Initially, two situations are considered when asking the respondents: the first one is known as the fix-reach scenario, as the surveyed individual is asked to think about a given number of known contacts. In the second situation, we do not impose the limitation of the number of contacts of surveyed individuals, so it is left open for him/her to specify the size of his/her network (without reach scenario). Particularly, we consider a fixed-reach scenario with the 20 closest contacts and another scenario without reach. The reported data showed many inconsistencies in the fix-reach scenario, so we decided to work only with the without-reach scenario.
- **Sampling framework:** The surveyed people are individuals who live in the Madrid region, are at least 18 years old, and have a potential voting network also in the Madrid region. Moreover, the respondents are asked only about contacts in their social network, belonging to the Madrid regional electoral census.
- **Sample size:** $n = 100$ and no stratification.
- **Target unknown populations:** The main political groups presenting candidates for the regional elections: Partido Popular (PP), Partido Socialista de Madrid (PSM), Vox, Más Madrid (MM), and Unidas Podemos (UP); hence, $Q = 6$ as we are considering the blank vote as an alternative option.
- **Auxiliary known populations:** We have made the same considerations as in the national election (see previous section).
- **Questionnaire:** Similar to the questionnaire of the general election (see Appendix

6.S.A.2), where target and auxiliary data sources are collected from respondents by asking "How many people who ... do you know?" questions.

Data processing

Similar to the general election, we establish the same rules for the assessment and elimination of outliers and "dirty" information. Based on these rules, the same scenarios were considered. For the sake of completeness, the appendix with the Supplementary Material also includes tables (see Table 6.S.2) displaying the estimations that result from the application of NSUM procedures when no filtering is considered. Just to illustrate the number of surveys removed after data cleaning, we can observe that the simultaneous application of the previous rules would lead to 71 surveys for the regional elections.

Due to the existence of *barrier effects*, the networks surrounding the participants are not representative. We must introduce some ad hoc adjustments based on post-stratification (weighted estimation) in order to address the lack of representativeness and limit the impact of the barrier effect. The next section deals with the application of NSUM within the framework of stratification strategies to solve this problem.

NSUM estimation procedures

The biases that may arise from survey data collection can also be handled by an alternative strategy that relies on the idea of introducing weights to compute the estimations. The proposal is a weighted estimator, which is calculated by weighting NSUM estimations by age. This approach can be understood as a type of post-stratification that up-weights or down-weights the NSUM estimated voting proportions by age. As in our case, we have three age ranges, we compute NSUM estimators $\hat{p}_{u^j} : j = 1, 2, 3$ by each age range, and weight them as follows:

$$\hat{p}_{uw} = w_1 \cdot \hat{p}_{u^1} + w_2 \cdot \hat{p}_{u^2} + w_3 \cdot \hat{p}_{u^3} \tag{6.S.1}$$

The weights of the weighted estimator above must be set in advance so that the ages under-represented in the sample can be up-weighted and the over-represented ones can be down-weighted.

Let us denote by $w^p = (w_1^p, w_2^p, w_3^p)$ the vector containing proportions by age in the general population, and let us put $w^s = (w_1^s, w_2^s, w_3^s)$ to denote the same vector in the sample. The components of w^p can be obtained from official statistics. Then, we can set the weights in (6.S.1) as follows:

$$w_j = \frac{w_j^p / w_j^s}{\sum_{j=1}^{3} w_j^p / w_j^s} : j = 1, 2, 3. \tag{6.S.2}$$

The insertion of the weights above into the expression of the weighted estimator in (6.S.1), together with the application of the NSUM approaches from Table 6.1 by age range, will lead to different versions of the NSUM weighted estimation: $\hat{p}_{uw}^{NAIVE}$, $\hat{p}_{uw}^{MoS}$, $\hat{p}_{uw}^{MLE}$ and $\hat{p}_{uw}^{PIMLE}$.

Results for the regional elections

This methodology is applied to the data coming from the indirect surveys carried out in the Madrid autonomic (regional) elections of 2023. Specifically, an online survey was conducted using the Pollfish platform to gather data on the regional election in the Community of Madrid that took place on May 28, 2023. The questionnaire was written in Spanish; its English version is provided in the Supplementary Material, Section 6.S.A.2. The Community of Madrid was chosen as a case study because it has a single province electoral district. In the Supplementary Material, Figs 6.S.1 and 6.S.2 display the demographic characteristics of the survey respondents extracted from the raw data and the samples retained after the filtering preprocessing stage described above. Observe that for the age groups of the filtered data, the proportions of respondents are: 18-24: 21.9%, 25-34: 21.9%, 35-44: 17.8%, 45-54: 19.2%, and >54: 19.2.2%. These proportions are different from those provided by INE official statistics. Therefore, we perform the weighting estimation procedure to adjust the target population estimates. The survey data reported that the average indirect votes from respondents' network is 19.9; as the data processing step has led to 71 valid responses, we definitely handled around 1420 records by means of an indirect declaration of voting intention.

The population proportions by age range in the region of Madrid are $w^p = (0.15, 0.42, 0.43)$ (data provided by INE official statistics); whereas in the sample they are $w^s = (0.23, 0.55, 0.22)$. As we can observe, the older segment is underrepresented in the survey, which may enhance the barrier effects highlighted in Figure 6.1. Hence, the weighted estimator proposed in (6.S.1) —with the weights being adjusted by means of (6.S.2) to handle both underrepresentation and overrepresentation — is employed to get the voting proportion estimations. In this case, the weights are $w = (0.20, 0.23, 0.57)$.

The computation of the proposed weighted estimator with the NSUM estimation procedures summarized in Table 6.1 gives the voting estimation outcomes of Table 6.S.3. The table displays the estimations obtained for each political group, from the conservative parties, Vox and PP, to the progressive ideological ones, PSM, MM, and UP, as well as the BLANK alternative used to capture the blank vote or the voting intention for other minority groups. Finally, the last row reports the true outcome of the elections.

The results show that the proposed approach tends to overestimate Vox and underestimate PP in the conservative bloc; meanwhile, PSM and UP are overestimated,

and MM is underestimated in the progressive bloc; note that the BLANK alternative is overestimated. These results may be explained by the recall errors effect (Laga et al., 2021) as people can usually figure out the ideological profile of the contacts in their network rather than the specific political group each contact will vote for. The issue gets clarified when the outcomes are organized and gathered as appearing on the right side of the table: PP+Vox (conservative bloc), PSM+MM+UP (progressive bloc with the latter two allied in Sumar for the national elections). Now, the results provided by the weighted NSUM procedure are closer to the true outcome of the regional elections, with the MLE method beating the other ones.

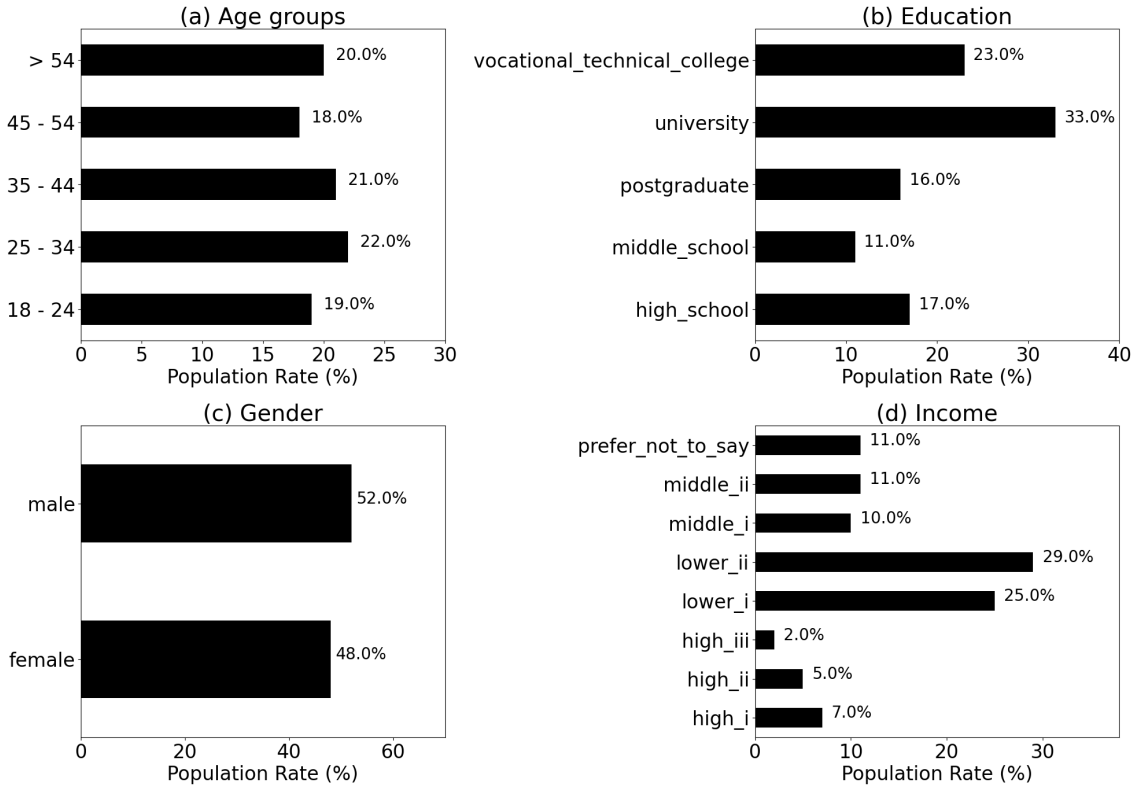


Figure 6.S.1: Madrid regional elections, May 2023. Indirect survey collected in the Community of Madrid. Demographics extracted from raw data. (a) Age groups, (b) Education, (c) Gender, and (d) Income.

6.S.A.4. Tables with NSUM-based Confidence Intervals

The following tables present the results we have obtained by applying NSUM methods, along with their respective confidence intervals (see Tables 6.S.4, 6.S.6 and 6.S.8).

Method	Political parties					Ideological blocs		
	Vox	PP	PSM	MM	UP	PP+Vox	PSM+MM+UP	BLANK
Naïve	14.0	34.2	21.1	12.7	11.0	48.2	44.8	7.0
MoS	15.2	34.6	19.8	12.9	11.0	49.8	43.7	6.5
MLE	12.3	40.2	20.5	11.6	9.4	52.5	41.5	6.0
PIMLE	14.3	35.3	20.4	12.2	10.8	49.6	43.4	7.1
Real outcome	**7.3**	**47.3**	**18.2**	**18.4**	**4.7**	**54.6**	**41.3**	**4.1**

Table 6.S.3: Estimation of the percentage of votes by political group for the regional elections.

Method	Political parties			
	Vox	PP	PSOE	Sumar
Naïve	16.06 (12.73-19.38)	35.07 (30.75-39.39)	28.07 (24.45-31.68)	16.23 (12.84-19.63)
MoS	14.90 (12.21-17.59)	37.81 (30.35-45.27)	27.47 (23.94-31.01)	16.08 (11.75-20.42)
MLE	15.42 (9.31-21.54)	36.29 (28.07-44.50)	29.09 (24.54-36.65)	16.37 (10.40-22.35)
PIMLE	14.63 (12.06-17.20)	39.24 (31.19-47.28)	27.65 (24.22-31.07)	14.71 (11.97-17.44)
Real outcome	**13.9**	**40.2**	**27.7**	**15.3**

Method	Ideological Blocs		
	PP+Vox	PSOE+Sumar	Blank
Naïve	51.13 (46.03-56.22)	44.29 (39.30-49.30)	4.57 (3.19-5.96)
MoS	52.71 (44.80-60.61)	43.56 (36.93-50.19)	3.74 (2.84-4.63)
MLE	51.71 (44.15-59.27)	45.47 (38.52-52.41)	2.82 (0.11-5.54)
PIMLE	53.87 (45.42-62.31)	42.35 (37.59-47.12)	3.78 (2.83-4.73)
Real outcome	**54.1**	**43.0**	**2.9**

Table 6.S.4: Spanish general elections, July 2023. Indirect survey collected in the Community of Madrid. Estimation of the percentage of votes by political group and blank votes and its 95% confidence interval after filtering out samples that contained a significant number of blank votes. Estimated percentage for the ideological blocs (conservative: PP+Vox and progressive: PSOE+Sumar).

Method	Voting Intention				
	Vox	PP	PSOE	Sumar	Blank
Naïve	15.70 (12.70-19.36)	33.75 (29.66-38.55)	27.59 (24.19-31.45)	15.75 (12.70-19.36)	5.53(3.78-11.22)
MoS	14.18 (11.16-18.33)	34.85 (27.61-43.25)	26.78 (21.59-31.79)	15.39 (11.42-20.56)	4.77 (3.14-21.77)
MLE	14.98 (9.37-21.11)	34.76 (23.74-43.19)	26.86 (15.80-34.36)	14.72 (8.52-20.67)	3.44 (2.41-34.95)
PIMLE	14.17 (11.05-17.94)	36.45 (28.76-45.32)	26.73 (21.72-31.83)	14.27 (10.99-18.16)	4.67 (3.07-21.02)
Real outcome	**13.9**	**40.2**	**27.7**	**15.3**	**2.9**

Table 6.S.5: Spanish general elections, July 2023. Indirect survey in the Community of Madrid. Bootstrap voting intention shares by political group and blank vote, and its 95% confidence interval.

Method	Political parties			
	Vox	PP	PSOE	Sumar
Naïve	15.25 (12.02-18.48)	35.07 (30.99-39.16)	28.52 (24.70-32.36)	12.29 (9.01-15.58)
MoS	13.09 (10.96-15.22)	37.10 (32.69-41.51)	27.55 (24.39-30.72)	13.44 (10.12-16.76)
MLE	13.91 (8.90-18.93)	35.58 (28.27-42.89)	27.59 (21.45-33.74)	14.69 (9.81-19.58)
PIMLE	12.68 (9.89-15.46)	36.59 (28.38-44.79)	30.63 (20.96-40.31)	12.31 (8.89-15.73)
Real outcome	**15.3**	**36.4**	**33.5**	**12.0**
Method	Ideological Blocs			
		PP+Vox	PSOE+Sumar	Blank
Naïve		50.32 (45.17-55.47)	40.82 (35.59-46.06)	8.85 (6.81-10.89)
MoS		50.19 (44.99-55.39)	40.99 (35.77-46.21)	8.82 (7.03-10.61)
MLE		49.49 (41.27-57.72)	42.29 (34.87-49.70)	8.22 (4.70-11.74)
PIMLE		49.26 (39.15-59.37)	42.94 (32.48-53.41)	7.80 (5.97-9.63)
Real outcome		**51.7**	**45.5**	**2.8**

Table 6.S.6: Spanish general elections, July 2023. Indirect survey collected in Andalusia. Estimation of the percentage of votes by political group and blank votes and its 95% confidence interval after filtering out samples that contained a significant number of blank votes. Estimated percentage for the ideological blocs (conservative: PP+Vox and progressive: PSOE+Sumar).

Method	Voting Intention				
	Vox	PP	PSOE	Sumar	Blank
Naïve	15.18 (12.04-18.47)	35.19 (31.15-39.42)	28.61 (24.90-32.57)	11.99 (8.83-15.41)	8.87 (6.74-11.34)
MoS	13.26 (10.33-16.38)	36.86 (32.07-41.76)	27.71 (23.89-31.74)	12.67 (8.86-17.01)	9.30 (6.80-12.31)
MLE	13.72 (10.69-17.04)	35.96 (31.18-41.05)	27.90 (23.73-32.36)	13.96 (9.45-18.82)	8.22 (6.18-11.08)
PIMLE	12.66 (10.13-15.61)	36.68 (32.25-40.93)	30.55 (25.40-35.82)	11.98 (7.87-16.83)	7.89 (5.35-10.92)
Real outcome	**15.3**	**36.4**	**33.5**	**12.0**	**2.8**

Table 6.S.7: Spanish general elections, July 2023. Indirect survey in Andalusia. Bootstrap voting intention shares by political group and blank vote, and its 95% confidence interval.

Method	Political parties			
	Vox	PP	PSOE	Sumar
Naïve	16.46 (13.12-19.80)	31.77 (27.78-35.76)	28.07 (24.22-31.92)	14.25 (10.88-17.63)
MoS	16.02 (13.23-18.81)	32.07 (28.27-35.88)	28.75 (25.38-32.13)	15.01 (12.20-17.81)
MLE	15.60 (7.75-23.45)	32.37 (21.16-43.58)	27.67 (17.98-37.36)	14.10 (6.59-21.61)
PIMLE	16.46 (10.65-22.27)	33.32 (24.89-39.76)	27.65 (21.65-33.65)	14.95 (9.60-20.30)
Real outcome	**15.7**	**34.9**	**32.1**	**15.2**
Method	Ideological Blocs			
		PP+Vox	PSOE+Sumar	Blank
Naïve		48.24 (42.91-53.57)	42.32 (37.37-47.27)	9.44 (6.88-12.00)
MoS		48.09 (42.71-53.48)	43.76 (39.08-48.44)	8.15 (6.43-9.87)
MLE		47.97 (35.22-60.72)	41.77 (30.18-53.37)	10.26 (3.49-17.02)
PIMLE		48.78 (37.69-59.88)	42.61 (34.10-51.12)	8.61 (5.00-12.22)
Real outcome		**50.6**	**47.3**	**2.2**

Table 6.S.8: Spanish general elections, July 2023. Indirect survey collected in the Valencian Community. Estimation of the percentage of votes by political group and blank votes and its 95% confidence interval after filtering out samples that contained a significant number of blank votes. Estimated percentage for the ideological blocs (conservative: PP+Vox and progressive: PSOE+Sumar).

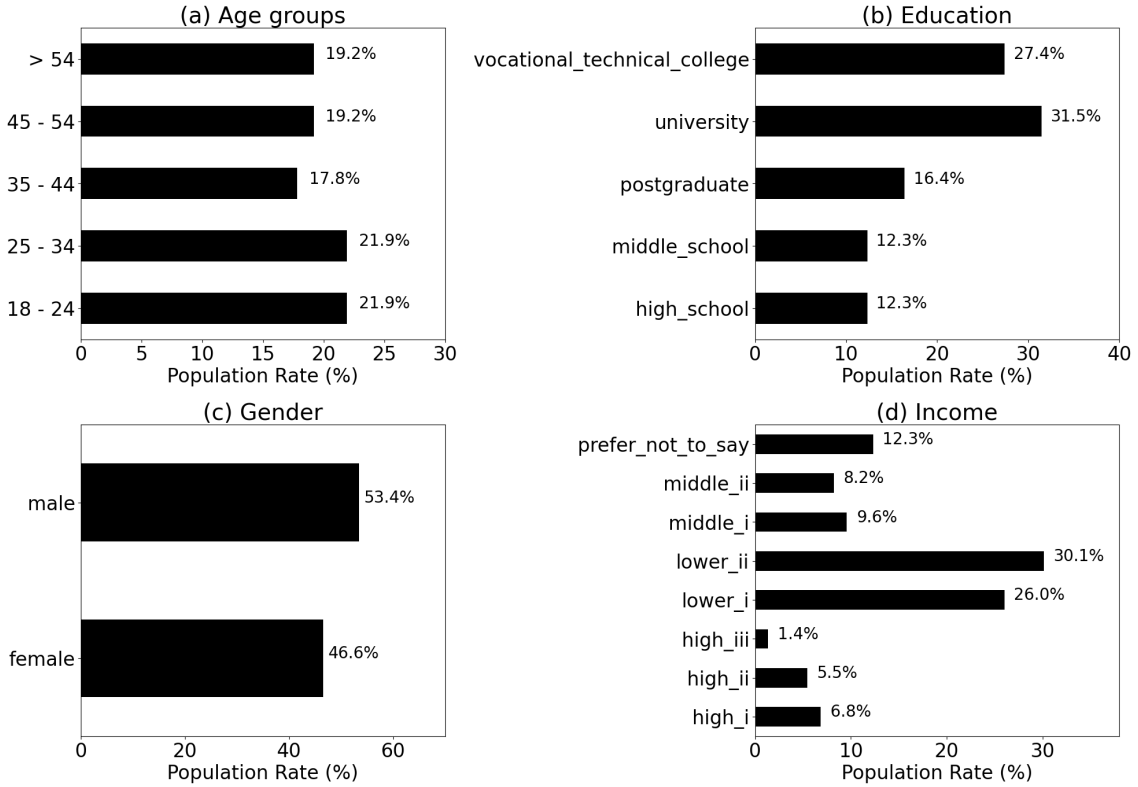


Figure 6.S.2: Madrid regional elections, May 2023. Indirect survey collected in the Community of Madrid. Demographics of the survey respondents after filtering out using the criteria of Section 6.3.2. (a) Age groups, (b) Education, (c) Gender, and (d) Income.

Method	Voting Intention				
	Vox	PP	PSOE	Sumar	Blank
Naïve	16.28 (13.07-19.81)	32.09 (28.18-36.16)	28.14 (24.38-32.03)	14.17 (10.98-17.69)	9.21 (6.67-11.55)
MoS	15.80 (12.36-19.50)	32.24 (27.88-36.75)	28.56 (24.73-32.63)	14.83 (11.31-18.77)	8.44 (6.03-10.86)
MLE	15.25 (12.06-18.93)	32.97 (28.28-38.03)	27.77 (23.82-32.11)	13.96 (10.68-17.70)	9.85 (6.44-12.82)
PIMLE	16.28 (12.71-20.03)	32.56 (28.09-37.09)	27.70 (24.04-31.64)	14.88 (11.31-18.83)	8.47 (6.03-10.92)
Real outcome	**15.7**	**34.9**	**32.1**	**15.2**	**2.2**

Table 6.S.9: Spanish general elections, July 2023. Indirect survey in the Valencian Community. Bootstrap voting intention shares by political group and blank vote, and its 95% confidence interval.

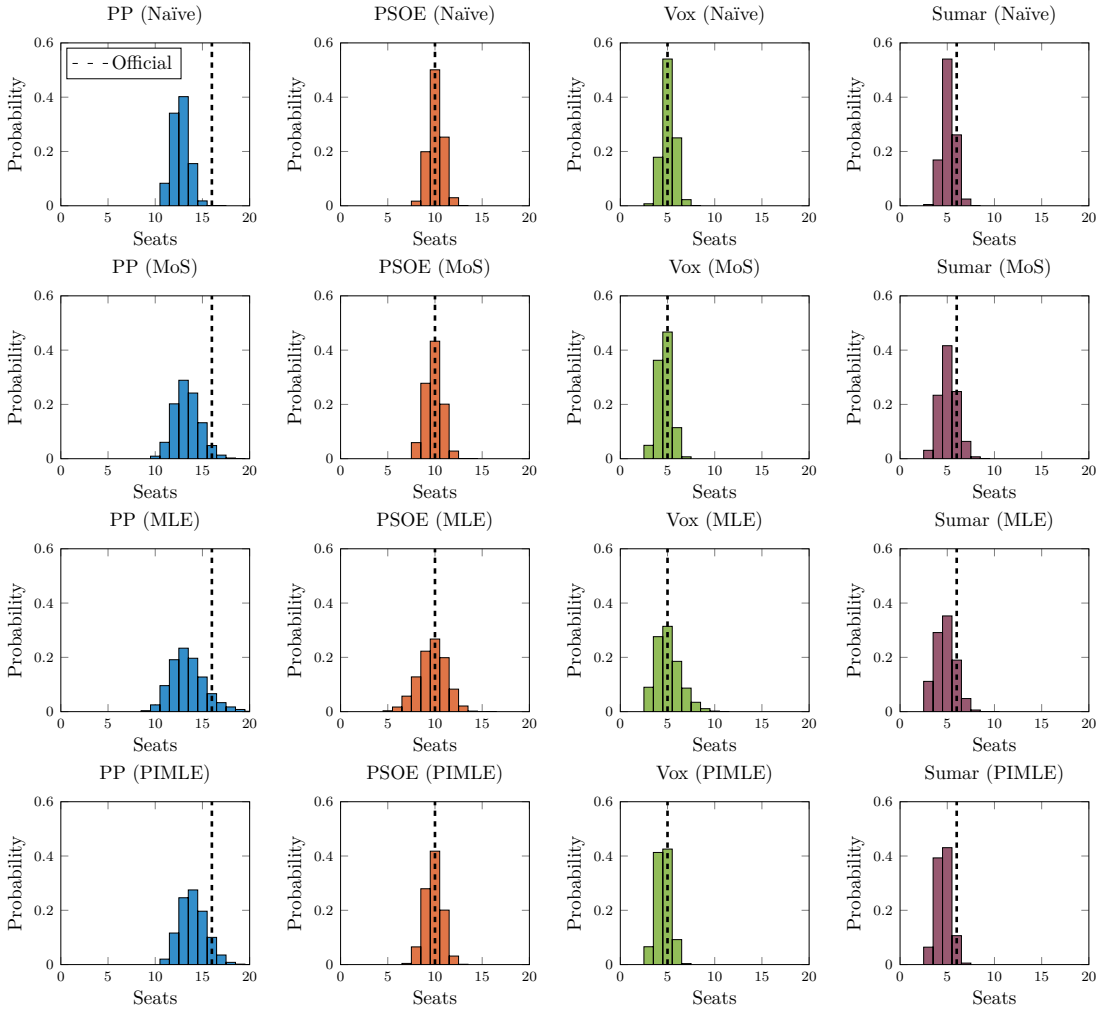


Figure 6.S.3: Bootstrap distributions of seat allocations by party and estimation method. Community of Madrid. Each panel shows the empirical distribution of the number of seats a given party obtains under a specific estimation method.

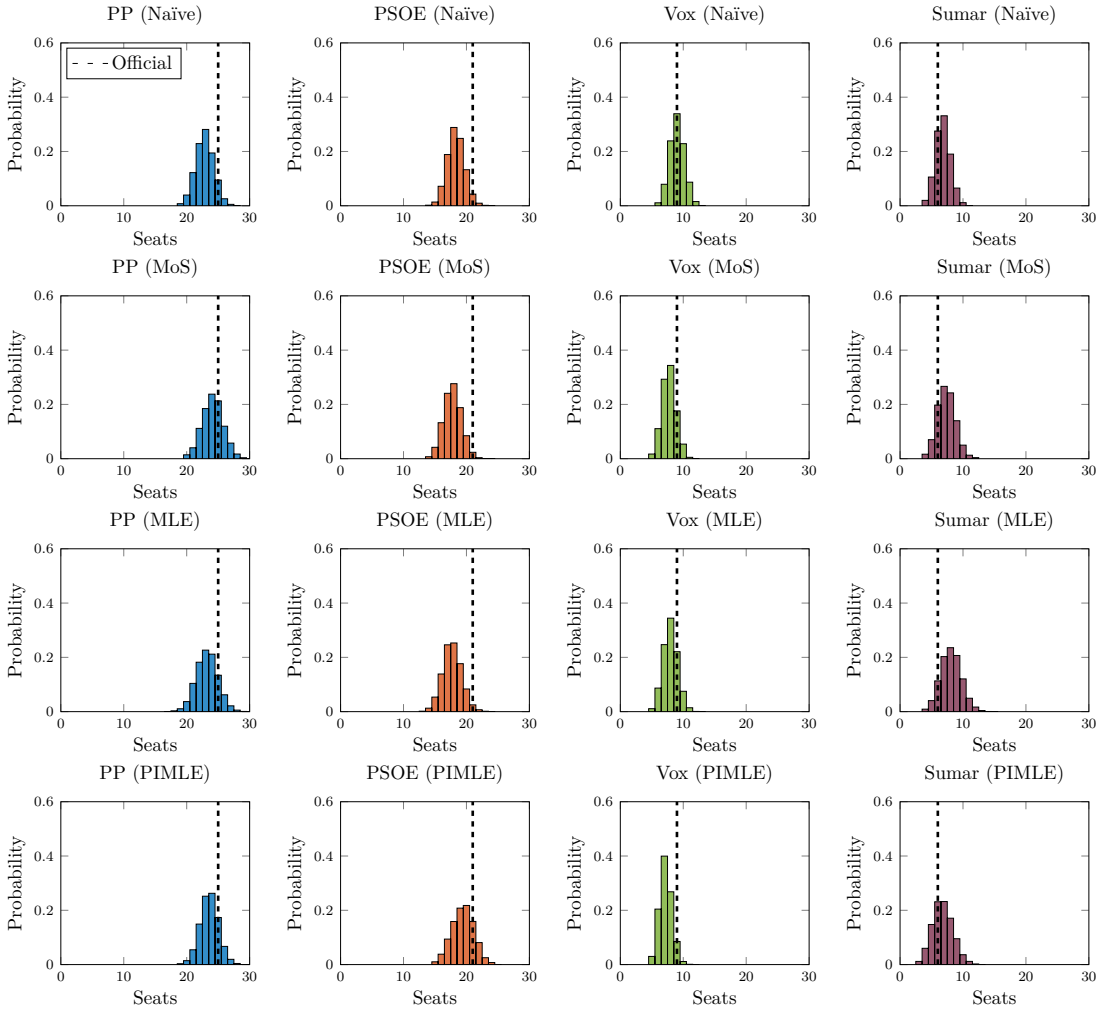


Figure 6.S.4: Bootstrap distributions of seat allocations by party and estimation method. Andalusia. Each panel shows the empirical distribution of the number of seats a given party obtains under a specific estimation method.

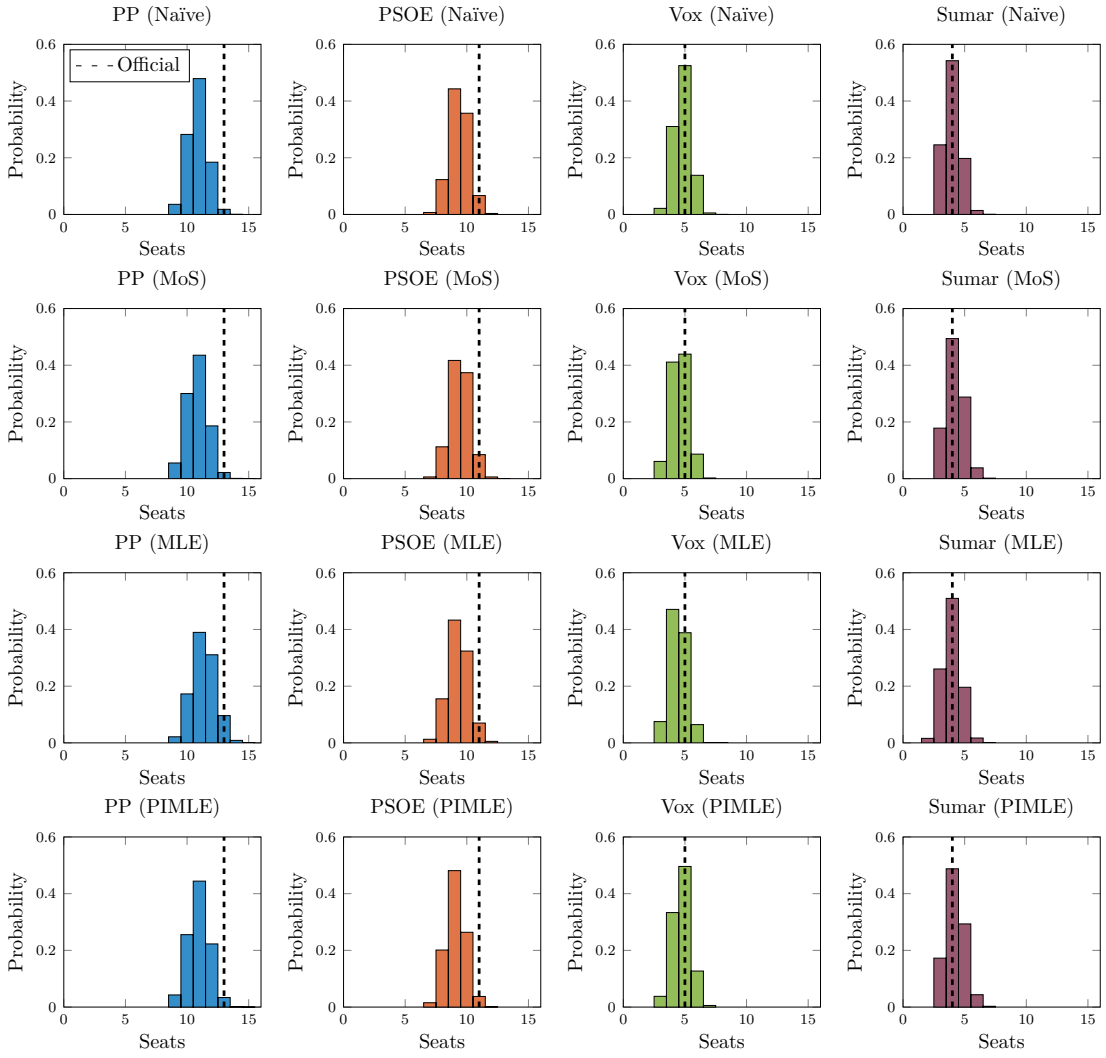


Figure 6.S.5: Bootstrap distributions of seat allocations by party and estimation method. Valencian Community. Each panel shows the empirical distribution of the number of seats a given party obtains under a specific estimation method.

6.S.B. D'Hondt Method

The D'Hondt method, also known as the process of greatest divisors, allocates seats in parliaments in federative countries and aims for proportional representation among political parties. It is calculated as follows: once all votes have been counted, successive quotients are determined for each party. The party with the highest quotient wins a seat, and its quotient is recalculated. This process continues until the required number of seats is filled. Each quotient is calculated using the formula $quotient = V/(s + 1)$, where V represents the total number of votes received by that party, and s is the number of seats that party has won so far, initially 0 for all parties.

6.S.C. Sensitivity Analysis under Homophilous Mixing

To evaluate the precision of NSUM estimators when the assumptions of random mixing are compromised due to political homophily, we conducted a sensitivity analysis using synthetic networks. Specifically, we generated stochastic block model (SBM) networks containing $250,000$ nodes with a mean degree of 15. Each network was partitioned into five disjoint groups corresponding to political parties, with group prevalences reflecting the official results of the Spanish general elections in the Community of Madrid: 40.2% (PP), 27.7% (PSOE), 15.3% (Sumar), 13.9% (Vox), and 2.9% (Blank). We controlled for homophily by varying the ratio (α) between intra-group and inter-group connection probabilities. That means that a node has α times more probability of being connected with a member of its own group. Furthermore, for each network, we created 4 known-prevalence populations (10%, 5%, 2%, and 1%) to evaluate the accuracy of NSUM methods that require degree estimation, namely MoS, MLE, and PIMLE. We generated ten SBM networks for each value of α, ranging from 1 to 10, representing both weak and strong homophily effects. In this setting, the value of α equal to 10 implies that the nodes are ten times more likely to know someone from their own group, which is a significantly high level of homophily for networks with an average degree of 15.

Figure 6.S.6 reports the estimated prevalence $\hat{\rho}$ in % as a function of α obtained by the various NSUM estimators for each party. We obtained each point on the curves by averaging the results of 2000 realizations, in which, at each trial, we extracted a sample of size $|S| = 200$ from the corresponding SBM network with a particular α. As observed, the MLE method is highly sensitive to increasing homophily. For political parties with high prevalence (e.g., Party 1), the MLE increasingly overestimates prevalence as α increases, while for smaller parties (e.g., Parties 3–5), it significantly underestimates prevalence. The other estimators are affected, but the difference relative to the MLE is substantial. One possible cause is the strong random mixing assumption underlying the MLE. Concretely, MLE assumes that $y_{iu} \sim Bin(d_i, p_u)$, an assumption that is violated under homophily. In contrast, PIMLE, MoS, and Naïve only require conditions for the first moments, $E[y_{iu}] = d_i p_u$, which may still hold in some scenarios with homophily. Furthermore, the Naïve estimator is also affected, showing deviations from the true prevalence under strong homophily (large α). Although some bias is evident —particularly for medium-sized parties— the magnitude of distortion is significantly smaller than that of the MLE. Finally, the PIMLE approach closely tracks MoS across all values of α, and both outperform Naïve (which estimates the degree with the responses of the political parties), particularly with high homophily. These results suggest that correcting degrees using known populations, as done in PIMLE and MoS, partially mitigates homophily-induced bias without introducing instability. In summary, these results indicate that homophilous mixing induces systematic, directionally consistent

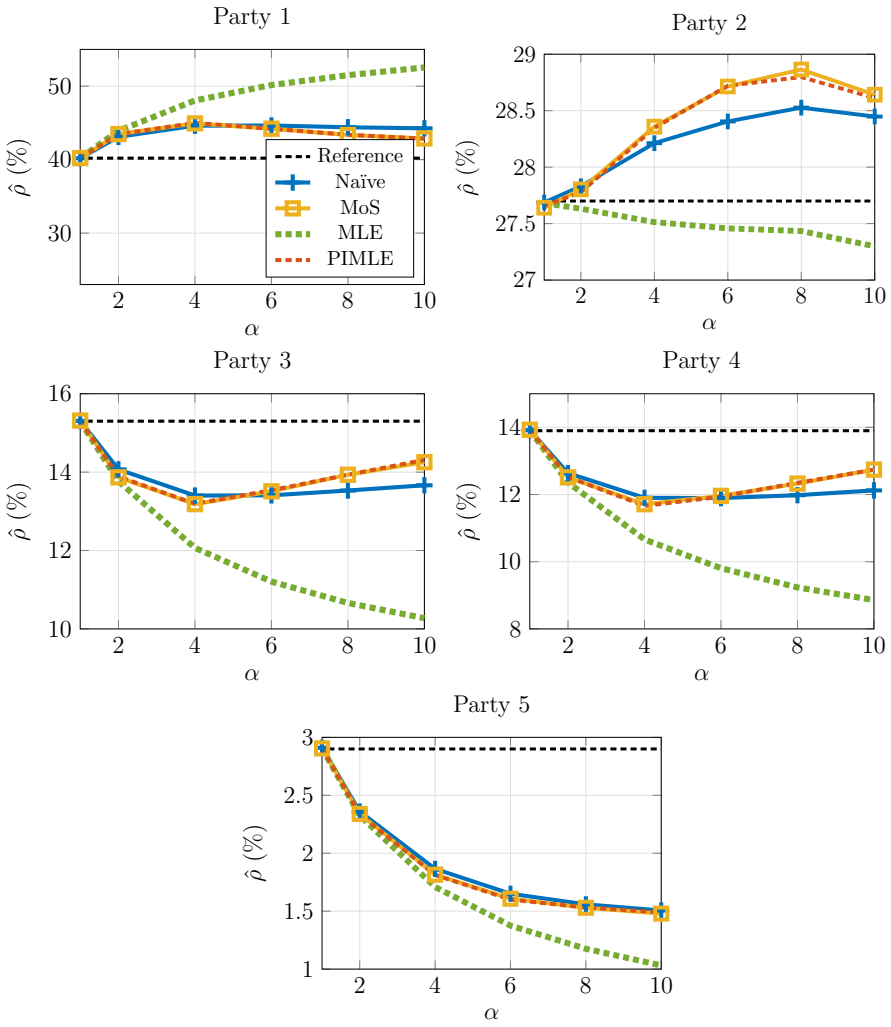


Figure 6.S.6: Estimated prevalence $\hat{\rho}$ in % against the homophily factor α obtained by the different NSUM estimators under evaluation for each party.

biases whose sign and magnitude depend on group size and position on the political spectrum.

Figure 6.S.7 displays the mean relative error as a function of the homophily factor α yielded by the various NSUM estimators for each party. In particular, the mean relative error is defined as

$$\frac{1}{|\mathcal{S}|}\sum_{i\in\mathcal{S}}\frac{\hat{\rho}_i-\rho}{\rho} \tag{6.S.3}$$

where $\mathcal{S}$ is the set of realizations. Notice that a positive error indicates overestimation, while a negative error indicates underestimation. Across all parties, the MLE method shows a rapid rise in absolute error as homophily increases, with errors ranging from $\pm 30\%$ to $\pm 60\%$ at realistic levels of homophily. This confirms that the MLE approach performs particularly poorly in politically structured networks. In contrast, Naïve, MoS, and PIMLE exhibit substantially lower relative error, even under strong homophily factors. Notably, the direction of bias is consistent: minority parties are underestimated, and majority parties are overestimated as homophily increases.

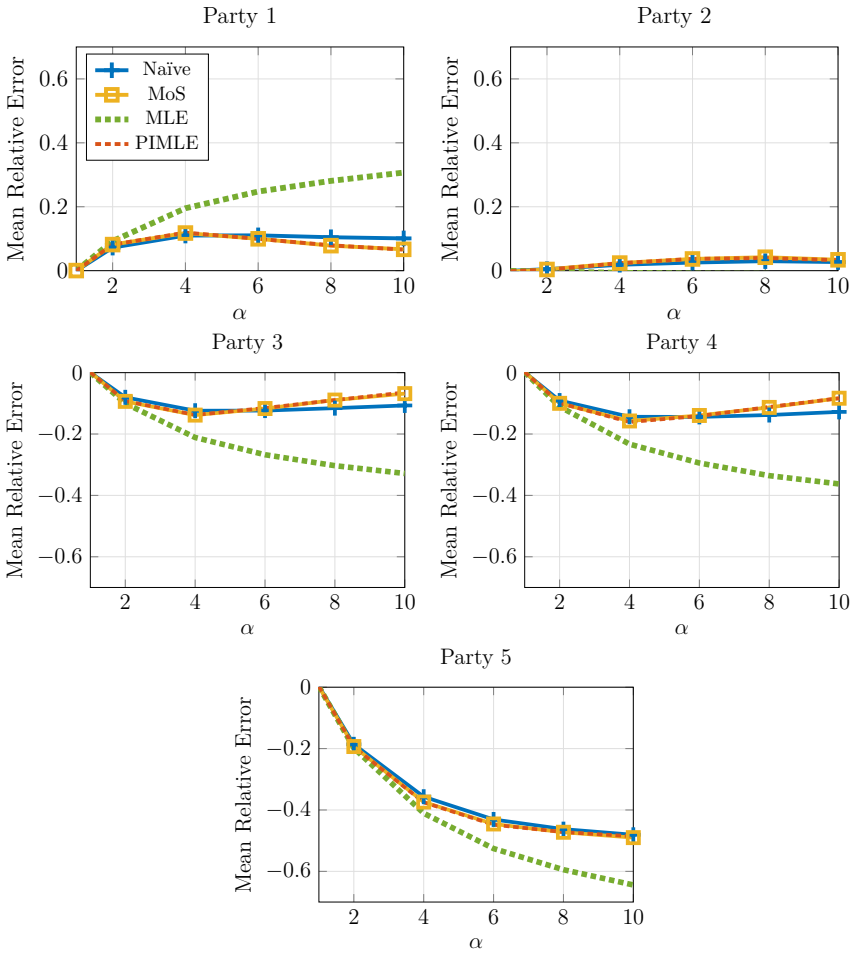


Figure 6.S.7: Mean Relative Error versus the homophily factor α obtained by the different NSUM estimators under evaluation for each party.

Overall, the sensitivity analysis shows that the main conclusions hold across realistic levels of homophily. At the same time, these results underscore the importance of accounting for bias direction when translating vote-share estimates into downstream quantities, such as seat allocations.

6.S.D. Robustness of Vote Share and Seat Allocation to Trimming and Filtering Choices

This section evaluates the sensitivity of the estimates of vote share and seat allocation to various filtering configurations. The goal is to assess whether the results reported in the main text depend on a specific, potentially ad hoc, choice of thresholds. To this end, we report results across a wide range of configurations, varying along three dimensions: (i) the trimming level applied to the voting-declaration network size (80%, 85%, 90%, and 95%), (ii) the cutoff used in the median absolute deviation (MAD) rule (3, 4, and 5), and (iii) whether the rule that removes respondents for whom more than 95% of reported alters support the same party is included or

excluded. Tables 6.S.10 and 6.S.11 present the results for Andalusia across four estimation methods: Naïve, MoS, MLE, and PIMLE, and report both vote shares and implied seat allocations.

Across all configurations, several consistent patterns emerge. First, the estimates of vote share are remarkably stable across trimming levels and MAD cutoffs. Adjusting the trimming from 80% to 95% results in variations of generally one to two percentage points across all major parties. However, these variations do not produce significant changes in party rankings and do not systematically favor any particular method. Second, seat allocations are even more stable than vote shares. For most parties, the number of seats remains unchanged across different trimming levels, MAD cutoffs, and the implementation of the homogeneity rule. When changes occur, they are usually limited to a single seat and do not affect the overall coalition structure. Third, increasing the MAD cutoff from 3 to 5 has only a marginal effect on both vote shares and seat shares, suggesting that the estimates are not highly sensitive to extreme observations. Overall, these results indicate that the proposed approach is not significantly affected by the choice of stopping rule or filtering method. Instead, the estimates are highly robust across a wide range of commonly used trimming and outlier-handling configurations, highlighting the stability of the proposed methodology in both polarized and more heterogeneous network settings.

95% rule	MAD cutoff	Trimming	Method	Voting shares (%)					Seats			
				Vox	PP	PSOE	Sumar	Blank	Vox	PP	PSOE	Sumar
with	3	80%	Naïve	15.0	35.7	29.2	12.7	7.3	10	24	19	8
			MoS	12.9	38.2	27.9	13.7	7.2	9	26	18	8
			MLE	13.3	36.1	28.7	15.5	6.5	10	24	19	8
			PIMLE	12.4	37.4	31.3	12.8	6.2	8	25	20	8
		85%	Naïve	14.9	35.9	29.5	12.7	7.0	10	24	19	8
			MoS	12.5	38.1	29.1	13.4	6.9	9	25	19	8
			MLE	13.5	36.7	29.6	14.6	5.7	9	24	19	9
			PIMLE	12.5	37.2	31.5	12.9	6.0	8	24	21	8
		90%	Naïve	14.6	35.8	29.9	12.8	6.8	9	24	20	8
			MoS	12.8	37.4	28.9	14.0	7.0	9	25	19	8
			MLE	12.5	36.4	31.1	14.8	5.3	9	24	20	8
			PIMLE	12.3	37.0	31.5	13.4	5.8	8	24	21	8
		95%	Naïve	14.5	35.5	29.7	13.6	6.7	9	23	20	9
			MoS	12.6	37.2	28.7	14.7	6.8	9	25	19	8
			MLE	11.9	34.6	30.1	18.5	4.9	12	22	20	7
			PIMLE	12.2	36.6	31.3	14.1	5.7	9	24	20	8
	4	80%	Naïve	15.2	35.4	28.9	12.7	7.7	10	24	19	8
			MoS	13.1	37.1	28.1	13.9	7.8	9	25	19	8
			MLE	13.5	35.9	28.2	15.3	7.1	10	24	18	9
			PIMLE	12.7	37.1	30.9	12.7	6.7	8	25	20	8
		85%	Naïve	15.1	35.7	29.2	12.6	7.4	10	24	19	8
			MoS	12.7	37.5	29.0	13.5	7.2	9	25	19	8
			MLE	13.6	36.4	29.2	14.5	6.2	9	24	19	9
			PIMLE	12.8	36.9	31.1	12.8	6.4	8	24	21	8
		90%	Naïve	14.8	35.7	29.6	12.7	7.3	10	24	19	8
			MoS	13.0	37.7	28.5	13.6	7.1	9	25	19	8
			MLE	12.7	36.4	30.6	14.5	5.8	9	24	20	8
			PIMLE	12.6	36.8	31.2	13.2	6.3	8	24	21	8
		95%	Naïve	14.6	35.3	29.4	13.5	7.1	9	24	19	9
			MoS	12.8	37.0	28.6	14.6	7.0	9	25	19	8
			MLE	12.1	34.7	29.8	18.1	5.4	11	23	19	8
			PIMLE	12.4	36.4	31.0	14.0	6.2	9	24	20	8
	5	80%	Naïve	15.2	35.1	28.5	12.3	8.9	10	24	19	8
			MoS	13.1	37.1	27.6	13.4	8.8	9	25	18	9
			MLE	13.9	35.6	27.6	14.7	8.2	10	24	18	9
			PIMLE	12.7	36.6	30.6	12.3	7.8	8	25	20	8
		85%	Naïve	15.1	35.3	28.9	12.2	8.5	10	24	19	8
			MoS	13.1	36.3	28.4	13.2	9.0	9	24	19	9
			MLE	13.8	36.3	28.7	13.9	7.4	9	24	19	9
			PIMLE	12.7	36.5	30.9	12.4	7.6	8	24	21	8
		90%	Naïve	14.8	35.4	29.2	12.3	8.3	10	24	19	8
			MoS	13.0	36.8	28.1	13.6	8.5	9	25	19	8
			MLE	12.9	36.2	30.1	14.0	6.9	9	24	20	8
			PIMLE	12.5	36.4	30.9	12.8	7.4	8	24	21	8
		95%	Naïve	14.6	35.0	29.1	13.1	8.2	10	24	19	8
			MoS	12.9	36.8	27.9	14.2	8.3	9	25	19	8
			MLE	12.3	34.6	29.3	17.4	6.4	11	23	19	8
			PIMLE	12.4	36.1	30.8	13.6	7.2	9	24	20	8

Table 6.S.10: Voting Shares and Seat Allocation for Political Groups in the July 2023 Spanish General Elections: Results from an Indirect Survey in Andalusia with Various NSUM Approaches and Filtering Configurations (with the 95% rule).

95% rule	MAD cutoff	Trimming	Method	Voting shares (%)					Seats			
				Vox	PP	PSOE	Sumar	Blank	Vox	PP	PSOE	Sumar
w/o	3	80%	Naïve	13.2	37.6	29.0	12.6	7.5	9	25	19	8
			MoS	11.3	38.6	28.9	14.0	7.2	9	26	19	7
			MLE	12.7	36.6	28.9	15.4	6.4	10	24	19	8
			PIMLE	11.1	39.6	30.6	12.6	6.1	8	26	20	7
		85%	Naïve	13.2	37.7	29.3	12.6	7.2	9	25	19	8
			MoS	11.6	39.5	28.4	13.8	6.7	9	26	19	7
			MLE	12.9	37.0	29.7	14.6	5.7	9	24	20	8
			PIMLE	11.2	39.3	30.8	12.7	5.9	8	26	20	7
		90%	Naïve	12.9	38.0	29.5	12.6	7.0	8	25	20	8
			MoS	11.0	40.3	28.2	13.9	6.6	9	27	18	7
			MLE	11.7	38.4	30.3	14.3	5.2	9	25	20	7
			PIMLE	11.0	39.6	30.6	13.0	5.7	8	26	20	7
		95%	Naïve	12.8	37.7	29.3	13.3	6.9	9	25	19	8
			MoS	11.0	39.6	28.0	14.8	6.5	10	26	18	7
			MLE	11.2	36.6	29.5	17.9	4.8	11	24	19	7
			PIMLE	10.9	39.3	30.5	13.7	5.6	9	25	20	7
	4	80%	Naïve	13.4	37.4	28.8	12.6	7.9	9	25	19	8
			MoS	11.7	39.5	27.8	13.4	7.6	9	27	18	7
			MLE	12.9	36.4	28.5	15.2	7.0	10	24	19	8
			PIMLE	11.4	39.3	30.3	12.5	6.5	8	26	20	7
		85%	Naïve	13.4	37.5	29.0	12.5	7.6	9	25	19	8
			MoS	11.9	40.3	27.6	13.1	7.1	8	27	18	8
			MLE	13.1	36.8	29.4	14.5	6.2	9	25	19	8
			PIMLE	11.5	39.0	30.5	12.6	6.3	8	26	20	7
		90%	Naïve	13.1	37.8	29.2	12.5	7.4	8	25	20	8
			MoS	11.1	39.9	27.8	14.2	6.9	9	27	18	7
			MLE	12.0	38.4	29.9	14.1	5.7	9	25	20	7
			PIMLE	11.3	39.4	30.3	12.9	6.1	8	26	20	7
		95%	Naïve	13.0	37.5	29.1	13.2	7.3	9	25	19	8
			MoS	10.9	39.8	27.5	14.9	6.8	10	26	18	7
			MLE	11.4	36.6	29.2	17.5	5.3	11	24	19	7
			PIMLE	11.2	39.0	30.2	13.6	6.0	9	25	20	7
	5	80%	Naïve	13.5	37.0	28.4	12.2	8.8	9	25	19	8
			MoS	11.6	40.1	27.0	13.0	8.4	9	27	18	7
			MLE	13.3	36.1	27.9	14.6	8.1	9	24	19	9
			PIMLE	11.4	38.8	30.1	12.2	7.6	8	26	20	7
		85%	Naïve	13.4	37.1	28.7	12.2	8.5	9	25	19	8
			MoS	12.1	39.2	27.3	12.7	8.7	8	27	18	8
			MLE	13.3	36.6	28.9	13.9	7.3	9	24	19	9
			PIMLE	11.5	38.6	30.3	12.3	7.4	8	26	20	7
		90%	Naïve	13.1	37.5	28.9	12.2	8.3	9	25	19	8
			MoS	11.5	39.3	27.8	13.2	8.2	9	26	19	7
			MLE	12.2	38.1	29.5	13.6	6.6	9	25	19	8
			PIMLE	11.3	38.9	30.1	12.6	7.1	8	26	20	7
		95%	Naïve	13.0	37.2	28.8	12.8	8.2	9	25	19	8
			MoS	11.5	38.7	27.8	14.0	8.0	9	26	19	7
			MLE	11.7	36.4	28.8	16.9	6.2	11	24	19	7
			PIMLE	11.2	38.6	30.0	13.2	7.0	8	26	20	7

Table 6.S.11: Voting Shares and Seat Allocation for Political Groups in the July 2023 Spanish General Elections: Results from an Indirect Survey in Andalusia with Various NSUM Approaches and Filtering Configurations (without the 95% rule).

Chapter 7

An Alternative to Crawlers: Scalable P2P Network Monitoring Using the Network Scale-Up Method

Juan Marcos Ramírez[1] Sergio Díaz-Aranda[1,2] Leonhard Balduf[3]
Antonio Fernández Anta[4,1]

[1] IMDEA Networks Institute, Spain
[2] uc3m-Santander Big Data Institute, Universidad Carlos III de Madrid
[3] TU Darmstadt, Germany
[4] IMDEA Software Institute, Spain

Abstract

Peer-to-peer (P2P) networks have remained an active area of research and development for decades. Originally used for data dissemination, they have later become instrumental to decentralized payment systems, blockchains, and Web3 applications following the introduction of Bitcoin by Satoshi Nakamoto in 2008. Since decentralization is a crucial property of these networks, they are continuously monitored to track attributes such as the geographical distribution of nodes, which serve as indicators of infrastructure concentration and resilience. However, monitoring typically relies on full-network crawls to derive complete, accurate snapshots, which requires substantial computational and network resources. In contrast, the Network Scale-up Method (NSUM) is a statistical tool in social network analysis that estimates the prevalence of subpopulations based on responses to questions such as "How many people do you know in a specific group?" This work leverages the information embedded in Kademlia routing tables to develop a lightweight NSUM-based measurement and estimation framework that infers global network attribute distributions from a small set of nodes, offering an alternative to full-network crawls. As a case study, we evaluate our approach on IPFS, a well-researched and widely used Web3 data storage and retrieval network. To the best of our knowledge, this is the first application of NSUM to decentralized network measurement. We evaluate our methodology on two datasets: the first comprises 360 full-network crawls collected over one month to assess our measurement-estimation approach across multiple simulated measurement scenarios, while the second dataset was collected through a deployment of the proposed lightweight measurement framework during May 2026. The results show that local Kademlia

routing views, with only 10 nodes, preserve sufficient information to accurately estimate global attribute distributions, achieving normalized biases below 10 %, and closely track temporal changes in network composition.

7.1. Introduction

The last two decades have seen a renewed interest in peer-to-peer (P2P) systems, driven in part by the development of cryptocurrencies such as Bitcoin and Ethereum (Nakamoto, 2008; Wood et al., 2014; Buterin, 2014; Tschorsch and Scheuermann, 2016) and distributed data storage systems such as the InterPlanetary File System (IPFS) (Benet, 2014), forming the foundation of the larger Web3 movement to decentralize the Web. These modern P2P systems are often measured for their properties, such as size, node populations, or geospatial node distribution, akin to measurements on earlier systems such as BitTorrent (Cohen, 2003; Pouwelse et al.) and Gnutella (Ripeanu et al., 2002). However, measuring P2P systems at scale typically relies on active probing and crawler-based techniques (Group, 2020, 2023; Stutzbach and Rejaie, 2006a) which are inherently intrusive because they try to extract information from all reachable nodes and pose challenges for capturing complete network snapshots due to churn (Stutzbach and Rejaie, 2006a,b).

The Network Scale-up Method (NSUM) is a statistical framework to estimate the proportion of subpopulations in social networks. NSUM has been applied to diverse domains, including disaster victims estimation (Bernard et al., 1988, 2001), and tracking diseases such as HIV/AIDS and COVID-19 (UNAIDS, 2010; Garcia-Agundez et al., 2021). Rather than relying on direct identification of nodes belonging to a specific sub-population, NSUM uses indirect reporting: a sample of nodes is queried about how many of their neighbors belong to the target group. NSUM offers two advantages: it enhances privacy by avoiding direct disclosure, and it increases coverage by enabling estimates even when full network visibility is unavailable.

IPFS implements a P2P network for decentralized content storage and retrieval (Shi et al., 2024) aimed at the next generation of the Web. It is built upon LibP2P, which forms the foundation of many modern P2P systems, such as Ethereum and Polkadot. IPFS is one of the largest P2P networks in operation today (Trautwein et al., 2022). It has gained attention as an external storage layer for blockchain-based applications, including non-fungible token (NFT) platforms, and decentralized applications (Ascigil et al., 2019; Chen et al., 2017; Das et al., 2022; Hoang et al., 2020; Balduf et al., 2022a). IPFS is particularly well-studied in the literature for its structure, content accessibility, and node distribution (Balduf et al., 2023; Shi et al., 2024; Abdullah Lajam and Ahmed Helmy, 2021; Balduf et al., 2022b; Prünster et al., 2022; Trautwein et al., 2022). The availability of tools to measure IPFS accurately, as well as the wide adoption of LibP2P as an

underlying technology for P2P networks, makes it a prime candidate to explore new measurement methodologies to characterize modern P2P systems.

Contributions. This paper introduces a measurement and estimation framework for P2P networks based on the NSUM. By utilizing information in routing tables, the proposed framework estimates the prevalence of node attributes from partial network views, offering an alternative to full-network crawls. Our contributions are fourfold: (1) We present, to the best of our knowledge, the first application of NSUM to decentralized network measurement and show how routing-table observations and node attributes can be mapped to the NSUM framework. (2) We design a lightweight measurement methodology that collects information from a small number of nodes and uses it with NSUM estimators to compute the prevalence of node attributes of the entire network. (3) We perform a large-scale evaluation using 360 IPFS distributed hash table (DHT) snapshots collected between April 1 and May 1, 2026, to assess the accuracy of the proposed framework under simulated measurement scenarios. (4) We implement the proposed lightweight measurement technique with 10 IPFS nodes to estimate the prevalence and track temporal changes in network attributes. To evaluate this framework, we consider geographic and cloud-provider distributions as representative attributes and quantify estimation accuracy using prevalence estimates and normalized bias.

Results. The routing table structure of the IPFS Kademlia DHT allows the implementation of a global estimator based on NSUM. This method achieves high accuracy in estimating key network features, such as the geospatial distribution of nodes and hosting patterns, while requiring fewer resources than methods based on full-network crawls. The estimation performance remains stable over time. Through extensive simulations, we evaluate two classical NSUM estimators observing a normalized bias (NB) of less than 10 % when sampling just 10 nodes. Additionally, our applied implementation of the method on 10 real IPFS nodes shows that network features can be estimated accurately using only local information. This approach is lightweight and faster than crawling, allowing for more frequent snapshots, which accurately track network features over time. Our method offers a less invasive, faster, and resource-efficient approach to characterizing P2P systems, paving the way for more ethical measurement practices. Compared to full-network crawls, our method interacts with a fixed number of nodes and finishes within seconds instead of minutes. Our findings likely generalize to other P2P networks with similar structures.

7.2. Background

7.2.1. The Network Scale-up Method

NSUM has been developed as a class of estimators designed to compute an accurate estimate of the size or prevalence of a subset of nodes—typically those sharing a specific attribute or property—using only the number of connections of a sample of nodes (Laga et al., 2021). Concretely, the NSUM estimators are the estimators that use only the number of outgoing connections that each sample node has into several groups. Given a network $G = (V, E)$, a subset of nodes H, and sample nodes $S \subset V$, the NSUM estimators use two types of variables:

- The out-degree, the total number of out-neighbor nodes of a node $v \in V$ in the graph G, denoted by R_v.
- The number of out-neighbors of a node $v \in V$ that belong to a given subset $H \subset V$, denoted by $C_{v,H}$.

The key intuition behind NSUM is that the proportion of connections a given node v has to nodes in H, expressed as $C_{v,H}/R_v$, carries meaningful information about the overall prevalence of that group in the network (Laga et al., 2021), where the prevalence of H, ρ_H, is expressed as $\rho_H = \frac{|H|}{|V|}$. The representativeness of the ratios considers several random-mixing assumptions for the classical NSUM estimators. The first estimator, known as *Mean of Ratios (MoR)* (Killworth et al., 1998a), assumes that $\mathbb{E}[C_{v,H}/R_v] = \rho_H$, and estimates prevalence as the average of the individual rates of the sample nodes S,

$$\hat{\rho}_H^{MoR} = \frac{1}{|S|} \sum_{v \in S} \frac{C_{v,H}}{R_v}.$$

The second estimator, *Ratio of Sums (RoS)* (Killworth et al., 1998b), assumes that $C_{v,H}$ are independent variables and follow the Binomial model $C_{v,H} \sim Bin(R_v, \rho_H)$. The prevalence is calculated as the maximum likelihood estimator, which corresponds to the ratio between the sum of the out-neighbor nodes belonging to H and the sum of the out-degree of all sampled nodes:

$$\hat{\rho}_H^{RoS} = \frac{\sum_{v \in S} C_{v,H}}{\sum_{v \in S} R_v}.$$

A major advantage of the NSUM estimators is that each sample node provides information about its neighbors. As a result, NSUM can leverage information from a substantially larger set of nodes without requiring direct access to each one. In particular, studies in economic networks have shown that this type of data reached the same conclusions while reducing costs by 70 % to 80 % (Breza et al., 2020).

7.2.2. Peer-to-Peer Networks and IPFS

P2P networks are a class of overlay networks in which equal peers communicate to share resources directly without the need for centralized servers. Early applications of this technology focused on file sharing (*e.g.*, BitTorrent (Pouwelse et al.; Cohen, 2003)), but the development of Bitcoin in 2008 (Nakamoto, 2008) extended the concept to cryptocurrencies. P2P networks can be divided into structured and unstructured networks depending on their overlay structure and search procedures. Unstructured networks such as Gnutella or Bitcoin use flooding mechanisms to disseminate messages through the network. Structured networks such as DHTs impose a logical structure upon the connections held by each node and introduce a distance metric, which ultimately allows efficient routing through the network towards a given target. Each node in a structured network maintains a routing table of other nodes.

Kademlia. Kademlia (Maymounkov and Mazières, 2002) is a DHT protocol used in many modern P2P systems (*e.g.*, BitTorrent, IPFS). Each participant in Kademlia is identified by a pseudorandom ID in a large key space, typically $[0, 2^{256})$. A node N maintains connections in a routing table consisting of k-buckets, each of which holds up to k entries. The buckets are constructed based on the XOR distance $d = a \oplus b$ between the node ID of N and the remotes. The first bucket contains peers with which N shares no common prefix, with each consecutive bucket containing peers that share one more bit of common prefix.

Crawling. It is possible to use the routing mechanism of structured overlays to discover and enumerate the participants of the network and their routing tables through a process known as crawling. A crawler starts with a set of nodes (*e.g.*, the bootstrap nodes of a network), and enumerates the contents of their routing table through specifically targeted routing queries. The newly learned nodes are added to a set of nodes to be crawled. This process repeats until no new peers are discovered. For Kademlia-based DHTs, a crawler typically issues requests targeted at specific k-buckets of a node, where each query returns up to k new peers. As typically only the first few buckets of a node's routing table are filled, not all buckets need to be probed to get a complete snapshot of the network. Numerous P2P systems have been crawled to describe their overlay structure and participants (Group, 2020, 2023; Stutzbach and Rejaie, 2006a).

IPFS. IPFS is a contemporary P2P network for data storage and retrieval (Benet, 2014) built on LibP2P. It uses a Kademlia DHT for routing. Nodes are identified via node IDs derived as the hash of their public key. LibP2P transparently multiplexes among multiple underlay connections, meaning that nodes can advertise and be reachable at multiple addresses. Underlay addresses in LibP2P are encoded as multiaddresses (Trautwein et al., 2022). These multiaddresses support various IP

versions (IPv4 or IPv6), transport protocols (TCP or QUIC), and network interfaces. The node population is segregated into DHT servers and clients. The former provide DHT functionality to other participants, storing key-value mappings and answering queries, whereas the latter only use the DHT without becoming a server themselves. The network is particularly well-studied in the literature (Trautwein et al., 2022; Balduf et al., 2022b, 2023; Shi et al., 2024) for its participants, overlay structure, data stored, and performance. A crawler for the IPFS DHT exists (Henningsen et al., 2020) that can enumerate all nodes and their routing tables. The crawler uses crafted FIND_NODE remote procedure calls (RPCs) to retrieve specific k-buckets of a remote node. The crawler sends between 4 and 20 RPCs per node, terminating once no new peers are learned.

7.3. Methodology and Datasets

7.3.1. NSUM IPFS Characterization

We present an implementation of the NSUM to estimate the prevalence of a subset of nodes sharing a specific characteristic in the IPFS DHT, such as being located in a particular country or hosted by a specific cloud provider, using fewer measurement nodes. In the following, we describe our method and then provide the theoretical justification behind the proposed NSUM implementation.

Our method considers a set of measurement nodes S in the IPFS network. We utilize information about their neighbors in the network, specifically information of the peers in their buckets. The nodes in S can either be operated for explicit measurement purposes, in which case their routing information can be obtained entirely locally, or existing nodes in the IPFS DHT can be crawled to obtain their buckets. We collect this routing information from S at a point in time t. From metadata about the neighbors (*e.g.*, their underlay addresses), we calculate the sought attributes of the peers in their routing table. We can then measure the number of outconnections $R_{v,t}$, and the number of out-neighbors C_{v,H_t} that belong to H_t, where H_t is a set of nodes sharing one attribute. We then apply the two estimators from Section 7.2.1 to arrive at the sought prevalence $\hat{\rho}_{H_t}$.

Node attributes are assumed to be located uniformly at random in the IPFS network from the perspective of the node ID space. This is reasonable, as node IDs are assigned pseudorandomly, not based on properties of the node. In addition, peers are connected based solely on proximity in the node ID space. For these reasons, the random variables C_{v,H_t} can be assumed to be independent and to follow the Binomial model $C_{v,H_t} \sim Bin(R_{v,t}, \rho_{H_t})$, which are the assumptions underlying the RoS estimator. In particular, this implies that by the law of total expectation

$$\mathbb{E}\left[\frac{\sum_{v\in S} C_{v,H_t}}{\sum_{v\in S} R_{v,t}}\right] = \mathbb{E}\left[\mathbb{E}\left[\frac{\sum_{v\in S} C_{v,H_t}}{\sum_{v\in S} R_{v,t}}\middle| R_v, v \in S\right]\right] = \mathbb{E}\left[\frac{\sum_{v\in S} \mathbb{E}\left[C_{v,H_t}\right]}{\sum_{v\in S} R_{v,t}}\right]$$

$$= \mathbb{E}\left[\frac{\sum_{v \in S} \rho_{H_t} R_{v,t}}{\sum_{v \in S} R_{v,t}}\right] = \rho_{H_t}.$$

Additionally, $\mathbb{E}[C_{v,H_t}/R_{v,t}] = \rho_{H_t}$, and therefore, MoR is unbiased because it is expressed as the sample mean of the variables $C_{v,H_t}/R_{v,t}$.

7.3.2. Datasets and Experiments

In order to validate our method and to showcase the feasible applicability in a real deployment, we collect two main datasets: (1) A dataset consisting of full-network crawls, and (2) data from a deployment of ten IPFS nodes, of which we collect the local bucket information. We then use these two datasets for two main experiments: (1) First, we utilize the crawls to simulate measurement studies with a number of randomly chosen measurement nodes S. This allows us to show the effectiveness of the NSUM estimators. (2) Then, using the deployment of ten nodes, we use only their local routing information, *i.e.*, the contents of their buckets, to drive the previously established estimators.

In both experiments, we compare the results obtained through these estimations to the ground truth of full-network crawls.

Crawl Dataset and Simulated Measurement Studies. To derive a ground truth for our estimators, we crawl the entire IPFS DHT using an existing crawler (Henningsen et al., 2020; Group, 2020). We crawl the network 12 times per day between April 1 and May 1, 2026, for a total of 360 snapshots. The crawler enumerates all participants of the DHT, the edges between them, and attributes of each node, such as their underlay IP addresses, client versions, and supported protocols. A crawl takes, on average, 6 min 48 s, with each single successfully crawled node taking on average 1.974 s. Each crawl generates two files: (1) a metadata file with peer IDs, multiaddresses, and supported protocols, and (2) a graph edge list showing directed P2P connections.

We remove nodes that are either unreachable or could not be fully enumerated. The network comprises 8166 nodes and 644 747 edges on average.

The objective of this dataset is to determine whether node attributes in the IPFS DHT can be accurately estimated from a reduced set of sampled nodes, thereby avoiding the need for exhaustive network crawls. For each snapshot, we generate a measurement scenario by randomly selecting a subset of nodes from the network. Each sampled node exposes the peer IDs contained in its DHT routing table. Within the NSUM framework, these routing-table entries are treated as observable neighbors, while node attributes such as geographic location or cloud-provider affiliation define the target populations whose prevalence we aim to estimate. Using only information collected from the sampled nodes, we estimate attribute prevalences through two classical NSUM estimators, MoR and RoS (cf. Section 7.2.1). To quantify estimation

accuracy, we compute the normalized bias NB, which measures the relative deviation between the estimated prevalence and the prevalence obtained from the corresponding crawl snapshot (Díaz-Aranda et al., 2026):

$$\mathrm{NB} = \frac{1}{N} \sum_{i=1}^{N} \frac{|\hat{\rho}_i - \rho|}{\rho},$$

where N is the number of realizations, ρ is the ground truth, and $\hat{\rho}_i$ is the NSUM prevalence estimate in the i-th realization.

In terms of complexity when compared to complete network crawls, this scenario considers crawls of $|S|$ randomly chosen nodes, and thus utilizes approximately a fraction $|S|/|V|$ of RPCs with respect to a complete crawl of all V nodes of the DHT. On the other hand, if we were to operate any node in S for measurement purposes specifically, its bucket contents would be available locally, thus eliminating the need to crawl it. Running the NSUM estimators themselves takes less than 1 ms. As such, the application of NSUM is much faster than crawling, and reduces the number of RPCs issued for measurement purposes. The former is important as previous works have shown that P2P networks must be measured quickly due to churn (Stutzbach and Rejaie, 2006b,a; Falkner et al., 2007).

Bucket Dataset and Real-World Deployment. In addition to the above complete crawls, we also operate ten IPFS nodes spread among three servers in Germany, and collect their local routing tables. This experiment shows the practical applicability of our proposed method and measures the accuracy of the derived results. Five nodes operate as DHT servers, and their routing tables are already part of the above crawl dataset. The other five nodes operate as DHT clients, which allows them to operate with fewer resources. DHT clients still maintain a routing table in order to use the DHT. Using the routing tables of these ten nodes, we can evaluate whether it is possible to utilize the NSUM estimators from a purely local perspective. For each node, we extract its routing table every 15 min between May 26 and June 1, 2026, resulting in 596 snapshots, while continuing to collect crawls as described above. Each node maintains an average of 169 peers in its buckets.

Kademlia routing tables are affected by temporal dynamics. In particular, as a security mechanism, bucket management policies prioritize long-lived, stable peers, preferentially retaining recently responsive nodes and evicting transient or unstable ones. Consequently, routing tables of long-lived nodes tend to over-represent highly available peers and under-represent churn-prone nodes. This effect may introduce survivorship bias into the observed bucket lists, increasing the probability of overlap among sampled nodes for stable peers. To mitigate this potential bias, we restart each node after every measurement round, which causes it to clear its routing table and re-join the network.

Feature Assignment.

Recall (cf. Section 7.2.2) how each node in the IPFS DHT network is uniquely identified by a peer ID. Metadata per node includes multiaddresses, which identify the network endpoints to reach the node. We map each IP address of a node to a geolocation using the MaxMind GeoLite2 database (Developers, 2023) and the Udger database Udger (2026) to link the IP addresses to known cloud providers. Because a single peer may advertise multiple IP addresses, the inferred attributes may not be consistent across all endpoints. For example, different addresses of the same node may resolve to distinct geographic locations or cloud infrastructures due to multi-homing, VPN usage, or distributed cloud deployments. To assign a single attribute to each peer ID, we adopt the majority-vote strategy proposed by Balduf et al. (2023). This approach selects the most frequently observed geolocation or cloud provider for that node based on resolved IP addresses.

7.4. Results

7.4.1. Estimator Benchmark

We begin by assessing the proposed NSUM-based measurement framework using the 360 snapshots of the IPFS-Benchmark Dataset as described in Section 7.3.2. Fig. 7.1 displays the median NB together with the $p10$-$p90$ and $p25$-$p75$ percentile intervals as a function of the sample size. Results are shown for six representative attributes: three countries (the United States (US), Germany (DE), and Canada (CA)) and three cloud providers (Hetzner Online, Amazon AWS, and Digital Ocean). The curves correspond to aggregates across the 360 snapshots.

The NB decreases monotonically with sample size across all attributes and both estimators. This behavior is expected because larger samples provide broader coverage of the DHT routing state and reduce sampling variability, yielding more accurate prevalence estimates. The largest gains occur when the sample size increases from about 5 to 50 nodes. Beyond roughly 100 nodes, the curves flatten, indicating diminishing returns. Secondly, even for sample sizes below 20 nodes, the NB remains relatively low. For instance, for attributes such as US, DE, and Hetzner Online, the median NB rapidly falls below 5 % with fewer than 50 sampled nodes. This result is particularly relevant because it suggests that accurate monitoring can be achieved using only a very small fraction of the network. Lastly, attributes with low prevalence are more difficult to estimate. In particular, CA and Digital Ocean consistently show errors larger than those in the US or Hetzner Online. This difference arises because NSUM estimators depend on observing a sufficient number of neighbors that share the target attribute. Consequently, for low-prevalence attributes, fewer observations are available, leading to greater variance and larger errors. Finally, RoS and MoR show very similar behavior, with both exhibiting nearly identical trends. This suggests that the main source of uncertainty is sampling variability,

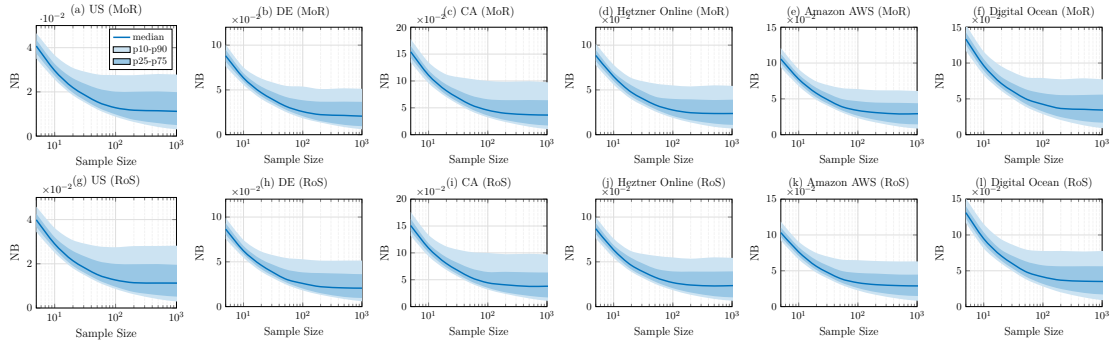


Figure 7.1: IPFS-Benchmark Dataset. Normalized bias (NB) of prevalence estimates as a function of sample size for six IPFS node attributes. The top row shows results are obtained with the MoR estimator, and the bottom row with the RoS estimator. Solid lines represent the median NB across the 360 snapshots, whereas the shaded regions indicate the $p25$-$p75$ and $p10$-$p90$ percentile intervals.

not estimator choice. RoS has a slight advantage at larger sample sizes, but the difference is small.

> **Takeaway 1:** NSUM-based estimation of IPFS attributes is feasible using a small number of sampled nodes. For all evaluated countries and cloud providers, both RoS and MoR achieve low NB with sample sizes well below those of full-network crawls. These findings indicate that the local views maintained by Kademlia routing tables preserve sufficient information about global attribute distributions to enable accurate estimation from partial observations and classical NSUM estimators.

Figures 7.2(a) and 7.2(b) compare the attribute prevalence estimates obtained using the MoR and RoS estimators against the prevalence values derived from a complete IPFS DHT crawl following the method of Balduf et al. (2023). The results correspond to a single snapshot collected on April 5, 2026, at 16:17 UTC. To simulate a real measurement scenario with limited server availability, the procedure randomly selects only 10 nodes from the network. For each estimator, the experiment is repeated 100 times, with a different set of sampled nodes in each realization. The reported MoR and RoS values correspond to the average prevalence across realizations, and the error bars represent the corresponding 95 % empirical confidence intervals. For visualization purposes, cloud-provider names are abbreviated as follows: NC (Non Cloud), HO (Hetzner Online), DO (DigitalOcean), VU (Vultr), AA (Amazon AWS), CH (Choopa), CG (Contabo GmbH), and OV (OVH).

For both countries and cloud providers, the estimates from MoR and RoS closely match the prevalence values from the crawl. Although the corresponding confidence intervals are slightly wider for attributes with lower prevalence, such as CH, CA,

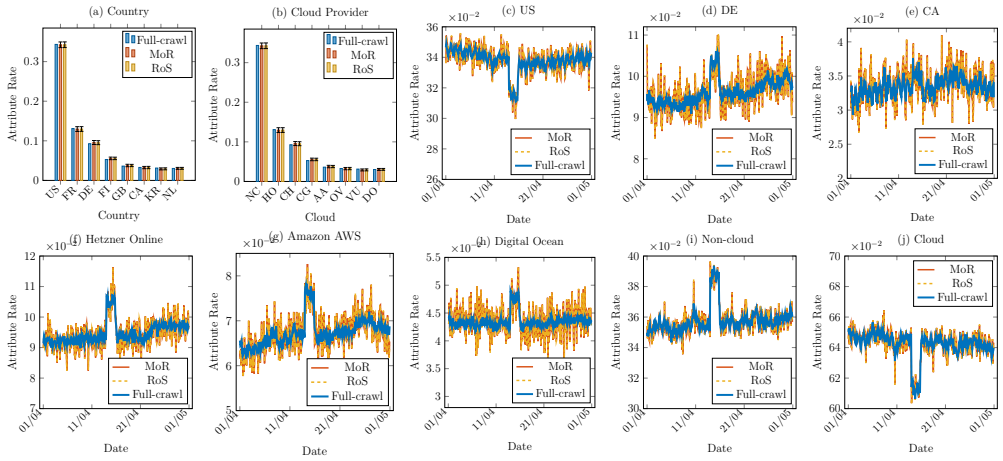


Figure 7.2: IPFS-Benchmark Dataset. Comparison of crawl-derived attribute prevalences with NSUM estimates (MoR and RoS) from 10 randomly sampled nodes at a snapshot captured on April 5, 2026. (a) country and (b) cloud-provider. (c)–(h) Temporal evolution of selected attributes. Evolution of (i) Non-cloud and (j) Cloud.

and DO, the estimated attribute rates remain close to the crawl-derived values. In addition, the MoR and RoS estimates are nearly indistinguishable across all attributes. This observation is consistent with the previous results shown in Figure 7.1, where both estimators exhibited similar NB values.

Figures 7.2(c)–7.2(h) extend this analysis to the temporal domain. The plots show the evolution of prevalence estimates across three countries (US, DE, and CA) and three cloud providers (Hetzner Online, Amazon AWS, and Digital Ocean) based on 10 sampled nodes. The ground-truth time series are derived from complete network crawls. Finally, Figures 7.2(i) and 7.2(j) present the temporal evolution of the aggregate Cloud and Non-cloud populations estimated using MoR and RoS, respectively. In all cases, the NSUM estimates follow the temporal dynamics observed in the crawl-derived ground truth. While the NSUM estimates exhibit greater short-term variability, they accurately capture major changes in network composition, including transient decreases in the prevalence of US nodes and corresponding adjustments in the distribution of cloud providers. These results indicate that a small number of sampled routing tables is sufficient not only to estimate attribute prevalence, but also to monitor temporal variations in the structure of the IPFS DHT.

Takeaway 2: The temporal dynamics of the IPFS DHT can be accurately monitored using a small number of randomly selected nodes. Even with data from just 10 nodes' routing tables, the MoR and RoS estimators accurately reflect both the current prevalence and the temporal evolution of attributes observed in complete network crawls.

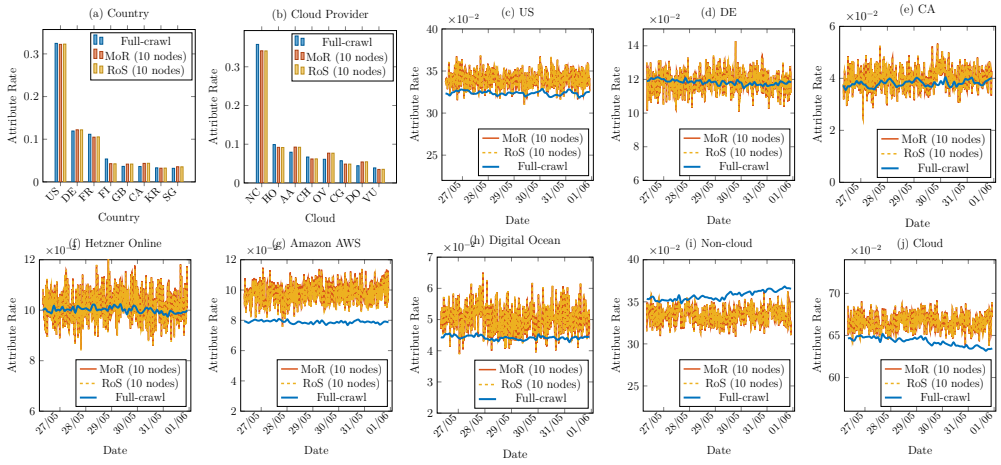


Figure 7.3: Bucket Dataset. Comparison of crawl-derived attribute prevalences with NSUM estimates (MoR and RoS) using the proposed measurement technique with 10 local nodes at a single snapshot. (a) country (b) cloud-provider. (c)–(h) Temporal evolution of selected attributes. Evolution of (i) Non-cloud and (j) Cloud.

7.4.2. Implementation Results

Figures 7.3(a) and 7.3(b) compare the attribute prevalences from a full-network crawl with estimates from the MoR and RoS estimators, using the proposed lightweight measurement framework (cf. Section 7.3.2). The results are based on a single snapshot and cover two classes of node attributes: countries and cloud providers. NSUM estimates are derived from information collected locally from 10 nodes. Both MoR and RoS closely match the attribute distributions from the full crawl across all evaluated countries and cloud providers. The agreement is particularly notable given the small number of sampled nodes. Although minor deviations can be observed in attributes with lower prevalence, the overall ranking and relative proportions of the categories are accurately preserved. From this static perspective, the results suggest that even a limited number of local Kademlia routing table observations contain sufficient information to recover global attribute distributions.

Figures 7.3(c)–7.3(h) extend the analysis to the temporal domain by comparing crawl-derived prevalences with estimates from the proposed measurement and estimation framework for three countries (US, DE, and CA) and three cloud providers (Hetzner Online, Amazon AWS, and Digital Ocean). The results are based on one week of measurements collected between May 26 and June 1, 2026. Owing to its significantly lower measurement overhead, the proposed framework enables more frequent observations: in this case, we collect routing-table snapshots from 10 local nodes every 15 min whereas full-network crawls are performed once every two hours. Although the NSUM estimates exhibit greater variability than the crawl-derived measurements, they consistently track the temporal evolution of the monitored attributes. MoR and RoS estimates capture trends in network composition, despite

relying on a very limited number of sampled nodes. This behavior indicates that lightweight routing-table measurements can be used not only to estimate attribute prevalence but also to monitor temporal changes in the IPFS DHT. Appendix 7.S.A presents the estimation results for 5 nodes with NAT and 5 nodes without NAT.

> **Takeaway 3:** Despite observing only 10 sampled nodes, the proposed NSUM-based measurement and estimation framework accurately captures both the static distributions and the temporal evolution of attributes in the IPFS DHT. These results indicate that local Kademlia routing-table views preserve sufficient information about the network composition, enabling monitoring and temporal characterization without repeated full-network crawls.

7.5. Discussion

We were able to show that NSUM is applicable to P2P measurements. This implementation differs from traditional applications of NSUM in statistics and sociology, and characteristics of IPFS actually mitigate limitations commonly encountered in those domains.

In the traditional application of NSUM in social networks, a representative sample of the population is required. Representative sampling is necessary in social networks because individuals often have different propensities for forming connections, and neighbors in the sample report the common contacts, particularly in highly clustered networks. These issues are largely absent in P2P networks, where attributes are uniformly distributed in the ID space. Another challenge found in social networks is the reporting behavior. Information is provided by sampled individuals and may be affected by recall errors, malicious responses, and a lack of transmission from the contacts (Laga et al., 2021). In contrast, peer information in IPFS is obtained directly from routing tables, eliminating these sources of reporting error and providing accurate measurements of node attributes of the peers. As for the complexity of our proposed method, we acknowledge that, while it greatly reduces the number of participants contacted and thus load on the network being generated, it cannot eliminate them entirely. Even in our proposed implementation using local bucket information the nodes themselves need to interact with the network prior to our measurements. Lastly, our simulations use randomly sampled nodes, and our implementation uses ten dedicated nodes. A middle ground and a potential solution to the question of how to randomly sample a number of nodes in a DHT, could be to operate just one measurement node and randomly sample among its peers.

7.6. Related Work

The popularity of IPFS has led to increased research focused on characterizing and understanding it. These studies have provided valuable information on the architecture of IPFS, including its P2P protocol design and node distribution. For instance, Trautwein et al. (2022) performed a large-scale measurement study that revealed the global reach of IPFS nodes across more than 2,700 Autonomous Systems and 152 countries, emphasizing its potential for decentralized content delivery. Other efforts have utilized crawler-based methods (Group, 2023) to examine the topology and dynamics of IPFS. Henningsen et al. (2020) analyzed connectivity and churn. More recently, Balduf et al. (2023) conducted a multimodal measurement study to examine the topology, traffic patterns, and hosting environments. These approaches rely on extensive crawling and probing, which can be resource-intensive and difficult to scale in decentralized environments. A number of works have proposed sampling techniques to accelerate estimates of the size or number of items in large-scale P2P systems such as KAD and the BitTorrent DHT (Wang and Kangasharju, 2013; Steiner et al., 2007b,a; Memon et al., 2009). These usually depend on crawling part of the entire keyspace exhaustively, which still potentially generates considerable load in the network.

In contrast, our approach introduces a lightweight, statistically grounded framework to characterize the IPFS network via NSUM using partial observations and indirect reporting.

7.7. Conclusions and Future Work

The IPFS DHT protocol enables the estimation of feature prevalence using a small sample set. Building on this property, classical NSUM estimators (RoS and MoR) provide accurate and consistent prevalence estimates of global properties, such as geographic distribution and hosting patterns, even when applied to random samples as small as 10 nodes. This finding underscores the potential of NSUM for low-overhead and less intrusive monitoring of large decentralized networks. While low-prevalence groups require a slightly larger number of nodes for precise estimation, the overall bias remains manageable, validating NSUM's robustness under realistic constraints.

Looking ahead, we plan to extend this approach beyond DHT nodes to incorporate all network participants. Furthermore, we aim to explore its applicability to other decentralized networks, such as Ethereum or other networks for which complete enumeration is difficult, and investigate adaptive sampling strategies that further reduce measurement overhead while maintaining accuracy. Finally, a current limitation of our work is that, while we can estimate the prevalence of certain subgroups, a measurement of the network size is needed to calculate the number of nodes per group.

Ethics and use of Generative AI

Our study raises a few ethical concerns. The crawls and routing tables we collect include IP addresses, but we do not attempt to associate these with the identities of specific individuals. For geolocation and cloud-hosting, we use the MaxMind GeoLite2 and Udger databases (Developers, 2023; Udger, 2026), which operate locally on our machines. Our work focuses solely on estimating network characteristics, not on behavior or personal data. We do not collect any additional datasets beyond those described in the article. In particular, our research does not involve handling or analyzing content data. As for generative AI, we use ChatGPT to revise previously written text. No text was authored solely by generative AI.

References

O. Abdullah Lajam and T. Ahmed Helmy. Performance evaluation of ipfs in private networks. In Proceedings of the 2021 4th International Conference on Data Storage and Data Engineering, DSDE '21, page 77–84, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450389303. doi: 10.1145/3456146.3456159. URL https://doi.org/10.1145/3456146.3456159.

O. Ascigil, S. Reñé, M. Król, G. Pavlou, L. Zhang, T. Hasegawa, Y. Koizumi, and K. Kita. Towards peer-to-peer content retrieval markets: Enhancing ipfs with icn. In Proceedings of the 6th ACM Conference on Information-Centric Networking, ICN '19, page 78–88, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450369701. doi: 10.1145/3357150.3357403. URL https://doi.org/10.1145/3357150.3357403.

L. Balduf, M. Florian, and B. Scheuermann. Dude, where's my nft: Distributed infrastructures for digital art. In Proceedings of the 3rd International Workshop on Distributed Infrastructure for the Common Good, DICG '22, page 1–6. Association for Computing Machinery, 11 2022a. ISBN 9781450399289. doi: 10.1145/3565383.3566106.

L. Balduf, S. Henningsen, M. Florian, S. Rust, and B. Scheuermann. Monitoring data requests in decentralized data storage systems: A case study of ipfs. In 2022 IEEE 42nd International Conference on Distributed Computing Systems (ICDCS), pages 658–668, Boston, MA, USA, 2022b. IEEE. doi: 10.1109/ICDCS54860.2022.00069.

L. Balduf, M. Korczyński, O. Ascigil, N. V. Keizer, G. Pavlou, B. Scheuermann, and M. Król. The cloud strikes back: Investigating the decentralization of ipfs. In Proceedings of the 2023 ACM Internet Measurement Conference, IMC '23, page 391–405, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400703829. doi: 10.1145/3618257.3624797. URL https://doi.org/10.1145/3618257.3624797.

J. Benet. IPFS - Content Addressed, Versioned, P2P File System. CoRR, abs/1407.3561, 2014. URL http://arxiv.org/abs/1407.3561.

H. Bernard, E. Johnsen, P. Killworth, and S. Robinson. How many people died in the Mexico City earthquake. Technical report, University of Florida, 1988.

H. R. Bernard, P. D. Killworth, E. C. Johnsen, G. A. Shelley, and C. McCarty. Estimating the ripple effect of a disaster. Connections, 24(2):18–22, 2001.

E. Breza, A. G. Chandrasekhar, T. H. McCormick, and M. Pan. Using aggregated relational data to feasibly identify network structure without network data. American Economic Review, 110(8):2454–84, 2020.

V. Buterin. Ethereum: A next-generation smart contract and decentralized application platform. 2014. URL https://ethereum.org/content/whitepaper/whitepaper-pdf/Ethereum_Whitepaper_-_Buterin_2014.pdf.

Y. Chen, H. Li, K. Li, and J. Zhang. An improved p2p file system scheme based on ipfs and blockchain. In 2017 IEEE International Conference on Big Data (Big Data), pages 2652–2657, Boston, MA, USA, 2017. IEEE. doi: 10.1109/BigData.2017.8258226.

B. Cohen. Incentives build robustness in bittorrent. In Workshop on Economics of Peer-to-Peer systems, volume 6, pages 68–72, 2003. URL http://jmvidal.cse.sc.edu/library/cohen03a.pdf.

D. Das, P. Bose, N. Ruaro, C. Kruegel, and G. Vigna. Understanding security issues in the nft ecosystem. In Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security, CCS '22, page 667–681, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450394505. doi: 10.1145/3548606.3559342. URL https://doi.org/10.1145/3548606.3559342.

M. Developers. Geoip and geolite. https://dev.maxmind.com/geoip/, 2023. Accessed: 2025-10-13.

S. Díaz-Aranda, J. M. Ramírez, J. Aguilar, R. E. Lillo, and A. F. Anta. Robust network scale-up method estimators. Social Networks, 84:46–61, 2026.

J. Falkner, M. Piatek, J. P. John, A. Krishnamurthy, and T. E. Anderson. Profiling a million user dht. In C. Dovrolis and M. Roughan, editors, Proceedings of the 7th ACM SIGCOMM Internet Measurement Conference, IMC 2007, pages 129–134. ACM, 2007. doi: 10.1145/1298306.1298325.

A. Garcia-Agundez, O. Ojo, H. A. Hernández-Roig, C. Baquero, D. Frey, C. Georgiou, M. Goessens, R. E. Lillo, R. Menezes, N. Nicolaou, et al. Estimating the covid-19 prevalence in spain with indirect reporting via open surveys. Frontiers in Public Health, 9:658544, 2021.

T. Group. Libp2p-crawler. https://github.com/trudi-group/ipfs-crawler, 2020. Accessed: 2025-10-13.

T. Group. Ipfs metrics exporter plugin. https://github.com/trudi-group/ipfs-metric-exporter, 2023. Accessed: 2025-10-13.

S. Henningsen, M. Florian, S. Rust, and B. Scheuermann. Mapping the interplanetary filesystem. In 2020 IFIP Networking Conference (Networking), pages 289–297, Boston, MA, USA, 2020. IEEE.

V.-H. Hoang, E. Lehtihet, and Y. Ghamri-Doudane. Privacy-preserving blockchain-based data sharing platform for decentralized storage systems. In 2020 IFIP Networking Conference (Networking), pages 280–288, Boston, MA, USA, 2020. IEEE.

P. D. Killworth, E. C. Johnsen, C. McCarty, G. A. Shelley, and H. R. Bernard. A social network approach to estimating seroprevalence in the united states. Social networks, 20(1):23–50, 1998a.

P. D. Killworth, C. McCarty, H. R. Bernard, G. A. Shelley, and E. C. Johnsen. Estimation of seroprevalence, rape, and homelessness in the united states using a social network approach. Evaluation review, 22(2):289–308, 1998b.

I. Laga, L. Bao, and X. Niu. Thirty years of the network scale-up method. Journal of the American Statistical Association, 116(535):1548–1559, 2021. doi: 10.1080/01621459.2021.1935267. URL https://doi.org/10.1080/01621459.2021.1935267. PMID: 37994314.

P. Maymounkov and D. Mazières. Kademlia: A peer-to-peer information system based on the XOR metric. In P. Druschel, M. F. Kaashoek, and A. I. T. Rowstron, editors, Peer-to-Peer Systems, First International Workshop, IPTPS 2002, Cambridge, MA, USA, March 7-8, 2002, Revised Papers, volume 2429 of Lecture Notes in Computer Science, pages 53–65. Springer, 2002.

G. Memon, R. Rejaie, Y. Guo, and D. Stutzbach. Large-scale monitoring of DHT traffic. In R. Rodrigues and K. W. Ross, editors, Proceedings of the 8th international conference on Peer-to-peer systems, IPTPS'09, Boston, MA, USA, April 21, 2009, page 11. USENIX, 2009. URL http://www.iptps.org/papers-2009/memon.pdf.

S. Nakamoto. Bitcoin: A peer-to-peer electronic cash system. 2008.

J. Pouwelse, P. Garbacki, D. Epema, and H. Sips. The bittorrent p2p file-sharing system: Measurements and analysis. In P2P Systems IV, Lecture Notes in Computer Science, pages 205–216. Springer. ISBN 978-3-540-31906-1. doi: 10.1007/11558989_19.

B. Prünster, A. Marsalek, and T. Zefferer. Total eclipse of the heart–disrupting the InterPlanetary file system. In 31st USENIX Security Symposium (USENIX Security 22), pages 3735–3752, Boston, MA, USA, 2022. USENIX Association.

M. Ripeanu, I. T. Foster, and A. Iamnitchi. Mapping the gnutella network: Properties of large-scale peer-to-peer systems and implications for system design. CoRR, cs.DC/0209028, 2002. URL https://arxiv.org/abs/cs/0209028.

R. Shi, R. Cheng, B. Han, Y. Cheng, and S. Chen. A closer look into ipfs: Accessibility, content, and performance. Proc. ACM Meas. Anal. Comput. Syst., 8(2), May 2024. doi: 10.1145/3656015. URL https://doi.org/10.1145/3656015.

M. Steiner, E. W. Biersack, and T. En-Najjary. Actively monitoring peers in KAD. In J. R. Douceur and R. Wattenhofer, editors, 6th International workshop on Peer-To-Peer Systems, IPTPS 2007, 2007a.

M. Steiner, T. En-Najjary, and E. W. Biersack. A global view of KAD. In Proceedings of the 7th ACM SIGCOMM Internet Measurement Conference, IMC 2007, pages 117–122. ACM, 2007b. doi: 10.1145/1298306.1298323.

D. Stutzbach and R. Rejaie. Understanding churn in peer-to-peer networks. In J. M. Almeida, V. A. F. Almeida, and P. Barford, editors, Proceedings of the 6th ACM SIGCOMM Internet Measurement Conference, IMC 2006, Rio de Janeriro, Brazil, October 25-27, 2006, pages 189–202. ACM, 2006a. doi: 10.1145/1177080.1177105. URL https://doi.org/10.1145/1177080.1177105.

D. Stutzbach and R. Rejaie. Capturing accurate snapshots of the gnutella network. In INFOCOM 2006. 25th IEEE International Conference on Computer Communications, Joint Conference of the IEEE Computer and Communications Societies. IEEE, 2006b. doi: 10.1109/INFOCOM.2006.347. URL https://doi.org/10.1109/INFOCOM.2006.347.

D. Trautwein, A. Raman, G. Tyson, I. Castro, W. Scott, M. Schubotz, B. Gipp, and Y. Psaras. Design and evaluation of ipfs: a storage layer for the decentralized web. In Proceedings of the ACM SIGCOMM 2022 Conference, SIGCOMM '22, page 739–752, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450394208. doi: 10.1145/3544216.3544232. URL https://doi.org/10.1145/3544216.3544232.

F. Tschorsch and B. Scheuermann. Bitcoin and beyond: A technical survey on decentralized digital currencies. IEEE Commun. Surveys Tuts., 18(3):2084–2123, 2016. doi: 10.1109/COMST.2016.2535718.

Udger. Udger Data v4. https://udger.com/, 2026. Accessed: 2026-05-15.

W. UNAIDS. Guidelines on estimating the size of populations most at risk to HIV. Geneva: World Health Organization, 2010.

L. Wang and J. Kangasharju. Measuring large-scale distributed systems: case of bittorrent mainline DHT. In Proc. of P2P, pages 1–10. IEEE, 2013. doi: 10.1109/P2P.2013.6688697. URL https://doi.org/10.1109/P2P.2013.6688697.

G. Wood et al. Ethereum: A secure decentralised generalised transaction ledger. Ethereum project yellow paper, 2014.

Appendices

7.S.A. Nodes with and without NAT

Figures 7.S.1(a) and 7.S.1(b) compare the attribute prevalences obtained from a full-network crawl with the corresponding MoR and RoS estimates derived from only five NATed nodes within a single snapshot of the Bucket Dataset. These nodes operate as DHT clients and are therefore not directly observable through conventional DHT crawling. Although moderate deviations are observed, the estimated prevalences preserve the relative proportions and ranking of the attributes. This result suggests that the proposed measurement and estimation framework based on NSUM remains effective regardless of whether the measurement nodes are publicly reachable or operate behind NAT.

Figures 7.S.1(c)-7.S.1(h) evaluate the temporal evolution of different network features considering only five NATed IPFS nodes. Although relying on a very limited number of NATed nodes, both estimators capture the overall temporal behavior of the monitored attributes. The NSUM estimates exhibit substantially larger fluctuations than time series derived from full-network crawls. Since the estimators operate on a much smaller set of observations, they are therefore more sensitive to sampling variability. A notable exception is Amazon AWS, for which both MoR and RoS consistently overestimate the prevalence relative to the curve obtained from crawls, suggesting the presence of attribute-specific visibility biases. These results show that even a very small set of NATed nodes provides sufficient information to monitor the temporal changes in attribute prevalences. Although estimation variance increases compared with the experiments using 10 nodes, the framework remains capable of tracking network dynamics without requiring publicly reachable measurement infrastructure or full-network crawls.

Figures 7.S.2(a) and 7.S.2(b) show the attribute prevalences from a full-network crawl, along with the corresponding MoR and RoS estimates derived from only five nodes without NAT, in a single snapshot of the Bucket Dataset. As shown, the estimated prevalences preserve the relative rates and ranking of the attributes for this snapshot. Figures 7.S.2(c)–7.S.2(h) depict the temporal behavior of the proposed NSUM-based framework when measurements are collected from only five publicly reachable nodes. Although these results are obtained from only five nodes, both estimators capture the overall temporal evolution of the monitored attributes. For all evaluated countries and cloud providers, the estimated prevalences remain close to the curve obtained from full-network crawls and follow the same temporal patterns. Furthermore, the NSUM estimates exhibit larger fluctuation than both the curves obtained from crawls and those obtained from 10 nodes. However, the fluctuations are mainly around the prevalences derived from full-network crawls, indicating that the primary source of error is increased estimation variability due to

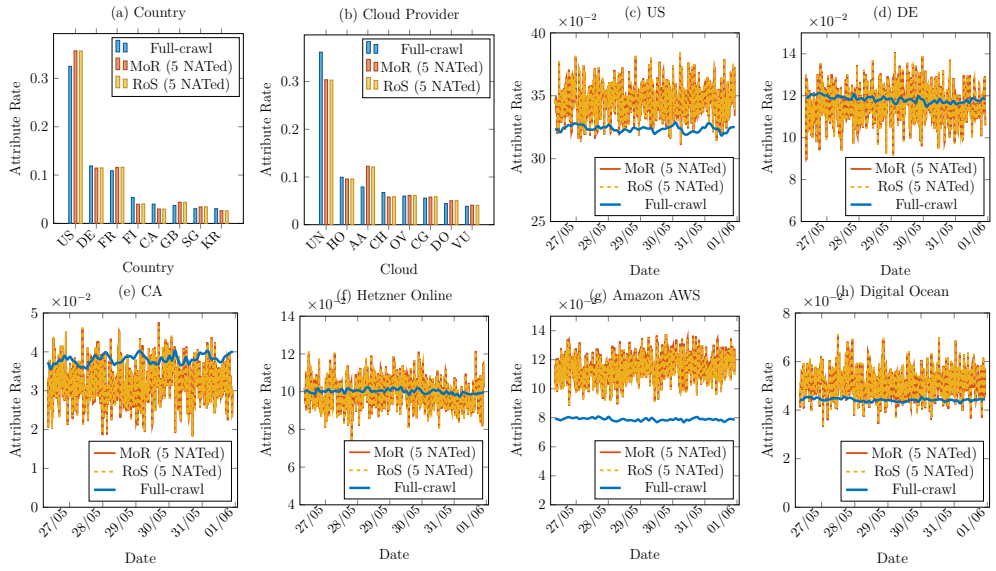


Figure 7.S.1: Bucket Dataset. Comparison of crawl-derived attribute prevalences with NSUM estimates (MoR and RoS) using the proposed measurement technique with 5 NATed nodes at a single snapshot. (a) country (b) cloud-provider. (c)–(h) Temporal evolution of selected attributes.

a small observation set, rather than significant estimation bias.

Unlike the NATed-node experiment shown in Fig. 7.S.1(g), no overestimation is observed for Amazon AWS. Instead, the MoR and RoS estimates remain approximately centered around the curve obtained from crawls throughout the measurement period. This suggests that publicly reachable nodes may provide a more representative view of the DHT, although the overall behavior remains consistent across both measurement scenarios. The results indicate that the proposed framework remains effective even when operating with only five publicly reachable nodes. Although the estimation variance is higher than in experiments using larger sample sizes, the framework accurately tracks the temporal dynamics of network attributes without requiring full-network crawls. Furthermore, these results support the feasibility of the proposed measurement and monitoring approach for large-scale Kademlia-based networks.

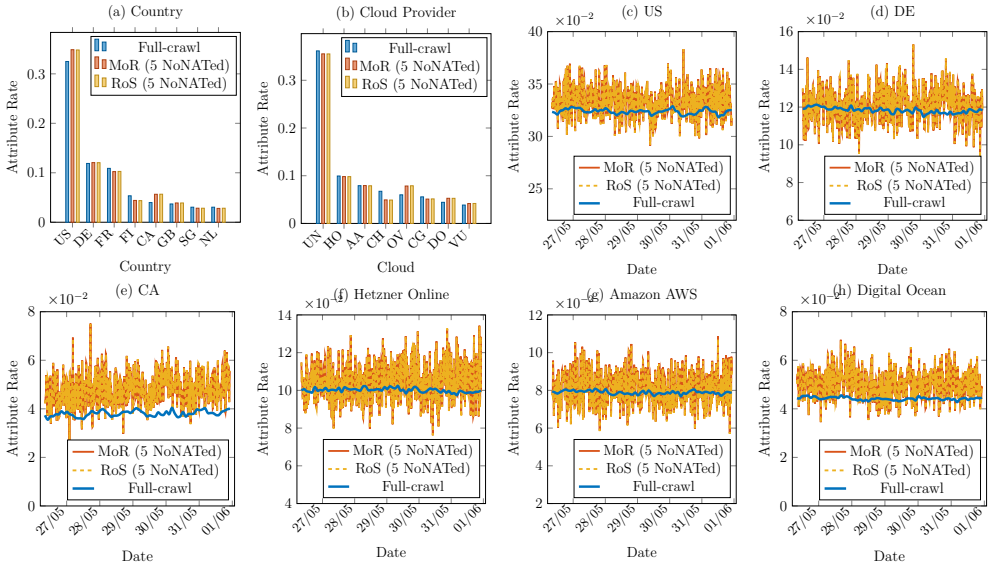


Figure 7.S.2: Bucket Dataset. Comparison of crawl-derived attribute prevalences with NSUM estimates (MoR and RoS) using the proposed measurement technique with 5 NoNATed nodes at a single snapshot. (a) country (b) cloud-provider. (c)–(h) Temporal evolution of selected attributes.

Chapter 8

Conclusion

This thesis contributes to the methodological development of the Network Scale-up Methods (NSUM) and Aggregated Relational Data (ARD) along several complementary directions. First, it provides an extensive empirical evaluation under a wide range of scenarios containing network and group structures and reporting conditions. Second, it establishes analytical guarantees for the error of the NSUM, their comparison with direct methods, and the estimation of trends using ARD. Third, it introduces robust estimation approaches designed to address limitations of traditional NSUM estimators. Finally, it extends the application of NSUM to new domains beyond hard-to-reach populations. More specifically, chapter by chapter, the main contributions of the thesis are as follows.

Chapter 2, published in *The American Statistician* https://doi.org/10.1080/00031305.2024.2421361, evaluated the performance of most of the NSUM estimators proposed in the literature under simulated scenarios. The study examined the impact of transmission error, recall error, network topologies, sample size, distribution of the unknown group, and the size, number, and intersection of known groups. The comparison of the large set of NSUM estimators revealed several important findings. First, some estimators exhibit substantial sensitivity to deviations from their underlying assumptions. Second, while transmission error affects all the estimators similarly–with the exception of GNSUM–recall error produces unreliable results for several estimators. Third, network topology influences the results: the best method depends on the network, and some methods display bias or excessive variability in certain settings. Fourth, reporting multiple nodes when the sample size increases does not impact the estimates in these networks excessively. Finally, the results suggested that the number of known groups used in the literature may be reduced significantly without compromising estimation accuracy.

Chapter 3, published in *ACM SIGKDD Conference on Knowledge Discovery and Data Mining* https://doi.org/10.1145/3711896.3736940, provided analytical error bounds for the NSUM, focusing on the multiplicative relative error of the NSUM estimators, assuming a simplified setting with only one unknown group. This article showed that identical ARD observations can arise from markedly different underlying networks and group structures, leading to potentially high estimation errors that can exceed $\sqrt{(n-1)/2}$, regardless of the NSUM technique and sample size, where n is the network size. Additionally, when the underlying network satisfies certain random structures, an upper bound was derived for RoS and MoR that decays exponentially

when the sample size increases. In particular, the probability of exceeding a small error can be achieved with high probability using only a logarithmic sample size, $O(\log n)$. Some refinements for Scale-Free networks were also provided. Finally, the theoretical findings were supported by an extensive set of numerical experiments in both synthetic and real data.

Chapter 4, published in *AAAI Conference on Artificial Intelligence* https://doi.org/10.1609/aaai.v38i20.30242, studied the estimation of temporal trends using ARD observations. A latent graph formulation was proposed to estimate the temporal evolution of the prevalence of unknown groups. The theoretical findings showed that, when the probability that a node belongs to the unknown group is independent of the in-degree, the expected response to the ARD observations is proportional to the size of the unknown group. Consequently, the temporal evolution can be estimated without requiring knowledge of the degrees of the participants. Applied for the estimation of time series of COVID-19 across several regions in the United States, this approach outperformed direct methods and traditional NSUM estimators. In addition, under latent networks with reasonably small variance and perfect knowledge of the average degree, estimation based on ARD achieves a lower probability of error than direct reporting for estimating the size of the unknown group. Finally, the temporal dynamics were incorporated through weighted moving averages, identifying conditions to improve the raw ARD-based estimates.

Chapter 5, published in *Social Networks* https://doi.org/10.1016/j.socnet.2025.08.002, introduced robust versions of RoS and MoR by replacing the sample means with robust estimators of location. The performance was evaluated on synthetic networks with several distributions, deviations of random mixing, and unusual data modeled through different types of contamination. The experiments revealed that RoS and MoR were particularly effective in scenarios without contamination but deteriorated significantly in contaminated scenarios. In contrast, the robust alternatives consistently outperform the traditional approaches in all the contaminated scenarios. Particularly, the robustification of the MoR Myriad outperformed other estimation methods (based on MoR) across various prevalence values and contamination levels. Methods based on M-estimators and trimmed means produced notable performance when the prevalence of the unknown group increased. Finally, methods based on medians tended to produce unstable results, likely due to the discrete nature of the ARD.

Chapter 6, published in *Social Networks* https://doi.org/10.1016/j.socnet.2026.04.006, implemented a methodology to estimate the outcome of elections using NSUM based on the ARD containing the perceived voting intention of the participants' contacts. The methodology proposed a framework for conducting indirect surveys and applying NSUM to estimate vote percentages by political group. Specifically, our methodology consisted of four stages: design of the survey questions, sampling procedures, outlier filtering, and selection of the NSUM estimators. The

methodology was evaluated in the July 2023 Spanish general elections in the regions of Madrid, Andalusia, and Valencia. The results showed that procedures based on NSUM can provide accurate estimates of electoral outcomes. Notably, the proposed methodology outperformed several well-known polling institutions in all considered regions despite relying on substantially smaller sample sizes.

Chapter 7 introduced the first methodology based on NSUM for estimating network attributes in structured distributed networks. Rather than reconstructing the entire network by exploiting routing mechanisms, as done by the traditional approaches based on crawlers, the proposed method used only information contained in the routing tables of a small set of nodes. This design considerably reduced the intrusiveness and computational cost of the measurement process, requiring only a few interactions with a fixed number of nodes, which was completed within seconds instead of minutes. Consequently, it enabled monitoring with more granularity while adhering to more ethical measurement practices. Extensive experiments on real network snapshots, together with a node deployment on IPFS, demonstrated that the proposed methodology can accurately estimate attribute prevalence using routing-table information from only 10 nodes.

The results presented throughout these chapters illustrate the methodological advances developed in this thesis for techniques based on ARD and NSUM. These contributions include both theoretical and empirical developments and have been implemented across a variety of social and technological scenarios, leading to publications in diverse research areas, including statistics, sociology, and machine learning. Although NSUM and ARD are relatively recent, with approximately 37 years of development, their popularity and scientific relevance have grown considerably, becoming an established research line with 153 publications between 2020 and 2025 (Lubbers et al., 2026). The breadth of applications explored in this thesis highlights the interdisciplinary nature of ARD and NSUM and their potential beyond traditional public health settings, such as statistics, computer science, machine learning, and network measurements. However, beyond the results presented in this thesis, we believe that these methodologies still offer substantial potential for new applications and methodological developments, opening multiple avenues for future interdisciplinary research.

Future work

The results obtained throughout this thesis have revealed several open directions for future research:

- *Evaluation under broader network structures.* Future work should evaluate NSUM estimators under additional network topologies, group configurations, and contamination. Chapter 2 and Chapter 5 show that these factors may significantly affect the performance of the estimators, but other combinations and cases remain

unexplored.

- *Bias correction of NSUM*. Although the proposed robust procedures mitigate the effects of unusual extreme values and distributional deviations, estimates may still be biased. Existing biased corrections are limited to a small number of estimators. Therefore, future work should aim to correct biases for a broader set of methods.
- *Unknown degrees*. Most methodological developments assume perfect knowledge of participants' personal network sizes. Future research should investigate estimation procedures and theoretical guarantees under unknown or contaminated degree information
- *Worst-case scenarios*. The substantial errors identified in some settings motivate the study of the minimal additional information required to overcome worst-case scenarios for NSUM.
- *Sensitivity of the analytical guarantees*. The presented analytical results in this thesis rely on assumptions of accurate reporting, random mixing, and perfect knowledge of the degrees. Extending these guarantees to more realistic settings would improve the practical applicability of NSUM estimators.
- *Error bounds in alternative network models*. The developed error bounds were derived under specific random network structures. Extending these analyses to other network models, such as Stochastic Block Models or Hyperbolic networks, which may also present deviations from the random-mixing assumptions, constitutes a relevant direction for future work.
- *Sample size selection*. Further research is needed on sample size criteria for other NSUM estimators under more general random networks, barrier effects, and contamination. Such results could be particularly valuable in low-resource contexts.
- *Multivariate methods*. Current robust approaches treat each ARD variable independently and therefore ignore the multivariate structure of ARD observations. Future work should explore multivariate robust methodologies capable of accounting for this structure. In this regard, we have already obtained promising results based on the use of statistical depth measures.
- *Sampling strategies*. The previous findings in the electoral application suggest that sampling design and stratification strategies are relevant for NSUM performance. Investigating optimal sampling methodologies across different domains represents another promising line of research.
- *Peer-to-peer networks*. Although the proposed NSUM implementation proved effective for characterizing network attributes, it relies on sampled observations rather than complete network enumeration. Future work should provide stronger analytical guarantees for the NSUM implementation, particularly when estimating low-prevalence attributes. In addition, other applications of NSUM to peer-to-peer network characterization, such as network size estimation, are still unexplored.
- *Privacy*. ARD-based surveys typically avoid asking respondents direct questions about membership in sensitive populations, which may encourage participation and reduce biases. However, it remains unclear to what extent sensitive attributes

can be inferred from ARD responses or in combination with other demographic data. A future line of work should investigate the privacy properties of ARD and quantify the risks of attribute inference and re-identification.

References

M. J. Lubbers, B. Völker, and M. Bojanowski. From hidden populations to social structure: Evolution of the network scale-up method, aggregated relational data, and their applications, 2026.