Nowcasting Temporal Trends Using Indirect Surveys

Ajitesh Srivastava[1], Juan Marcos Ramírez[2], Sergio Díaz-Aranda[2,3], Jose Aguilar[2], Antonio Ortega[1], Antonio Fernández Anta[2], Rosa Elvira Lillo[3]

[1] University of Southern California, USA
[2] IMDEA Networks Institute, Spain
[3] Universidad Carlos III, Spain
ajiteshs@usc.edu, juan.ramirez@imdea.org, sergio.diaz@imdea.org, jose.aguilar@imdea.org, aortega@usc.edu, antonio.fernandez@imdea.org, lillo@est-econ.uc3m.es

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.

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, Sims, and Costica 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, Bao, and Niu 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.

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, Bao, and Niu 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, Bao, and Niu 2021) for more details). Multiple NSUM extensions have been proposed (Laga, Bao, and Niu 2021) but all of them require to request or estimate PN sizes (Killworth et al. 1998b; Laga, Bao, and Niu 2021; Garcia-Agundez et al. 2021), or ask 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.

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 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 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 5, Theorem 3, and Theorem 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 3.1).
- We evaluate our approach in the estimation of COVID-19 cases in the US for a period of 18 months (Section 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.

2 Methodology

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 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, Bao, and Niu 2021). Under Assumption 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 1, we have the following theorem.

Theorem 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 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 1 and Theorem 1, the time series $\mathbb{E}(X_t)$ is proportional to 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 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 $0 \le \phi_t \le 1$.*

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 1. *If σ^2_D is the variance of the degree distribution,*

$$\sigma^2_{X_t} = f_t(\mu^2_D(\phi_t - f_t) + \mu_D(1 - \phi_t) + \sigma^2_D \phi_t)\,. \qquad (1)$$

Further, $\mu_D f_t(1 - \mu_D f_t) \le \sigma^2_{X_t} \le f_t(\sigma^2_D + \mu^2_D(1 - f_t))$.

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 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 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 2. *For any $\lambda > 0$, $P(|\bar{X}_t/\mu_D - f_t| > \lambda) \le P(|\bar{Y}_t - f_t| > \lambda)$, if the variance of degree distribution $\sigma^2_D \le \mu_D(\mu_D - 1)(1 - \phi_t)/\phi_t$*

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 $\sigma^2_D \le (\delta_{max} - \mu_D)(\mu_D - 1)$. To satisfy this bound, and still violate the assumption on degree variance in Theorem 2, would require $f_t \ge 0.2$, 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 2 also motivates framing the indirect survey questions in such a way that the variance of the graph is small.

2.2 Leveraging Smoothness

The hidden population in many real-life processes is **driven by smooth dynamic processes**, leading to the following.

Observation 2. *$|\Delta f_t| \le \epsilon_{f,1} f_t$ and $|\Delta^2 f_t| \le \epsilon_{f,2} f_t$, for some small $\epsilon_{f,1}, \epsilon_{f,2} \ge 0$.*

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 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 3.1 for simulation details). We repeat the analysis for $\sigma^2_{X_t}$. We calculated σ_{X_t} by setting $\mu_D = 15, \sigma^2_D = 100, \phi_t = f_t$. 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 3 and 4) that help ***bound the result of aggregating over smooth sequences***.

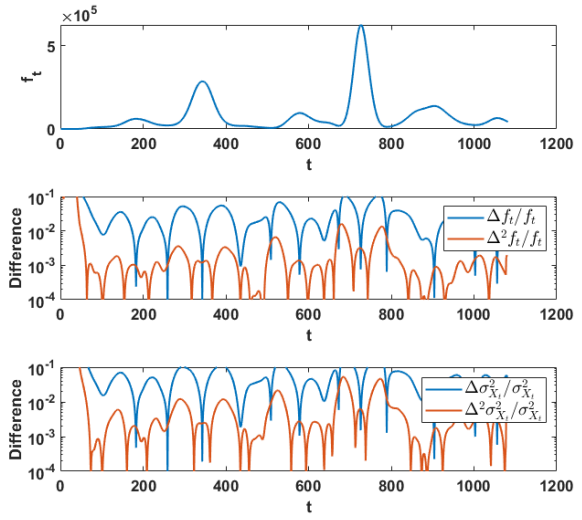

Figure 1: The first and second differences for f_t and σ_t are small – derived from denoised COVID-19 cases in California. t is the number of days since January 23, 2020.

Lemma 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. *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.

Demonstrating the Advantage of Smoothing We will use the fact that typical real-world signals to be estimated, f_t, are smooth (Observation 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 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 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 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 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 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 3 and Assumption 2, we are ready to prove the following theorem.

Theorem 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*

$$\lambda \geq w\epsilon_{f,1} / \left(1 - \left(1 + \frac{w\epsilon_{\sigma^2,1}}{1 - w\epsilon_{\sigma^2,1}}\right)\sqrt{n_t/n_w}\right). \quad (2)$$

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 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 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. *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*

$$\lambda \geq \gamma_f / \left(1 - \sqrt{\frac{n_t}{n_w}(1+\gamma_{\sigma^2})}\right), \tag{3}$$

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 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.

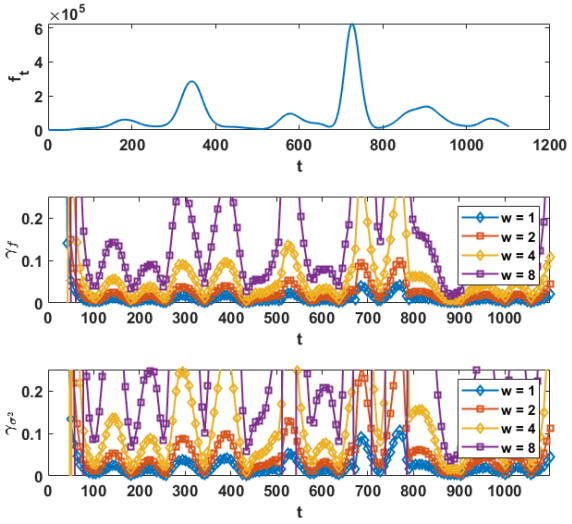

Figure 2: γ_f and γ_{σ^2} calculated over various window sizes for COVID-19 reported cases in California. The values are small as desired, particularly for small w.

3 Results and Analysis

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[1].

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 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 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

[1]Our code is available at https://github.com/GCGImdea/coronasurveys/tree/master/papers/2024-AAAI-Nowcasting-Temporal-Trends-Using-Indirect-Surveys.

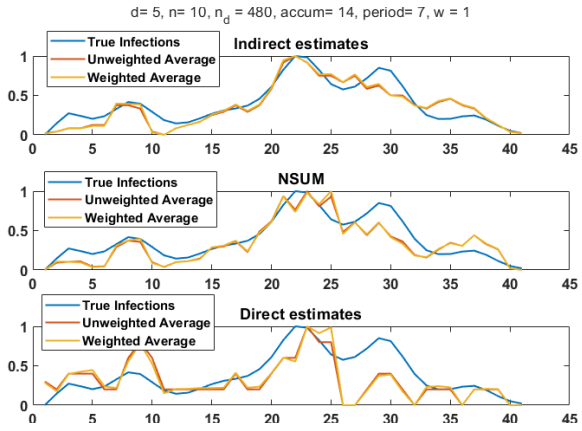

Figure 3: Result of one of the survey simulations.

have a similar effect as using $w > 1$ on $accum = 1$.

Figure 3 shows the result of the survey simulation with parameters $d = 5, n = 10, n_d = 480, accum = 14, period = 14$, and $w = 15$. Time-series obtained from smoothed indirect survey is 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 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 do 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 3 and 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.

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 in it 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

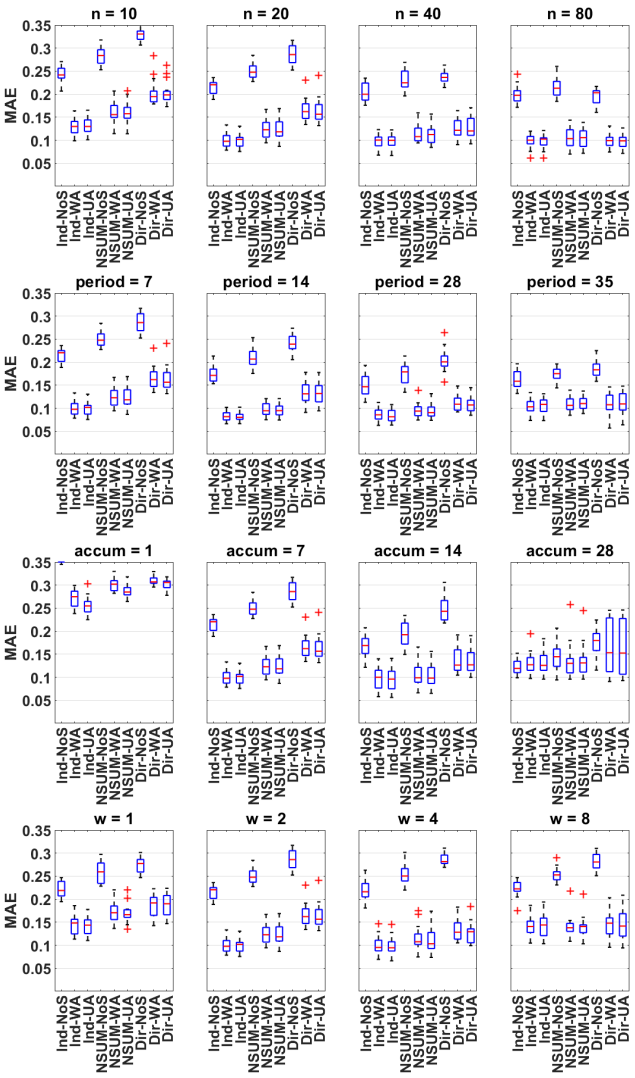

Figure 4: **1st row:** MAE vs n, the number of respondents. Indirect with moving average is the best for low n. All moving averages converge as n increases. **2nd row:** MAE vs *period* in the survey question. Smoothed results converge as *period* increases. **3rd row:** MAE vs *accum*, the accumulation 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.

dataset for each state.

Figure 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 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).

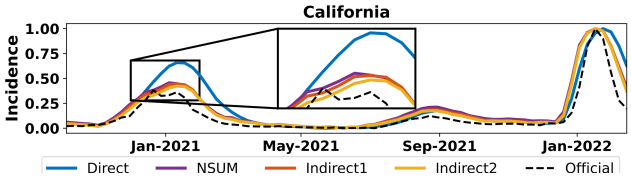

Figure 5: Normalized COVID-19 incidence estimated from the US CMU-CTIS data for California. The parameters of the proposed approach were set to $accum = 7$ and $w = 1$.

Table 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.

$accum$	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**

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, Du, and Gardner 2020). We observe that the curves obtained from indirect surveys are much similar to the official ones. To quantitatively evaluate the proposed approach, Table 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 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).

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 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.

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).

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.

Acknowledgements

This work was partially supported by grants TED2021-131264B-I00 (SocialProbing) and PID2019-104901RB-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.

Ethical Declaration

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 was obtained cannot be shared, as per the Data Use Agreements signed with Facebook and Carnegie Mellon University (CMU).

References

Alix-Garcia, J. M.; Sims, K. R.; and Costica, L. 2021. Better to be indirect? Testing the accuracy and cost-savings of indirect surveys. *World Development*, 142: 105419.

Astley, C. M.; et al. 2021. 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).

Bernard, H. R.; Johnsen, E. C.; Killworth, P. D.; and Robinson, S. 1989. Estimating the Size of an Average Personal Network and of an Event Subpopulation. *In The Small World*, 159–175.

Bhatia, R.; and Davis, C. 2000. A better bound on the variance. *The american mathematical monthly*, 107(4): 353–357.

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

Dietz, K. 1993. The estimation of the basic reproduction number for infectious diseases. *Statistical methods in medical research*, 2(1): 23–41.

Dong, E.; Du, H.; and Gardner, L. 2020. An interactive web-based dashboard to track COVID-19 in real time. *The Lancet infectious diseases*, 20(5): 533–534.

Dunbar, R. 2010. *How many friends does one person need? Dunbar's number and other evolutionary quirks*. Harvard University Press.

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

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

Garcia-Agundez, A.; et al. 2021. Estimating the COVID-19 prevalence in Spain with indirect reporting via open surveys. *Frontiers in Public Health*, 9: 658544.

Geldsetzer, P. 2020. 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.

Jing, L.; Lu, Q.; Cui, Y.; Yu, H.; and Wang, T. 2018. 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.

Kazemzadeh, Y.; Shokoohi, M.; Baneshi, M. R.; and Haghdoost, A. A. 2016. 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).

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

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

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

Oliver, N.; et al. 2020. 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.

Rossier, C. 2010. The anonymous third party reporting method. *Methodologies for estimating abortion incidence and abortion-related morbidity: a review*, 99–106.

Salganik, M.; et al. 2010. Estimating the number of heavy drug users in Curitiba, Brazil using multiple methods. Technical report, Technical report. UNAIDS.

Salganik, M. J.; et al. 2011. The game of contacts: estimating the social visibility of groups. *Social networks*, 33(1): 70–78.

Salomon, J. A.; et al. 2021. 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.

Scheers, N. 1992. A review of randomized response techniques. *Measurement and Evaluation in Counseling and Development*.

Srivastava, A.; Ramírez, J. M.; Díaz-Aranda, S.; Aguilar, J.; Ortega, A.; Fernández Anta, A.; and Lillo, R. E. 2023. Nowcasting Temporal Trends Using Indirect Surveys. arXiv:2307.06643.

Teo, A. K. J.; et al. 2019. Estimating the size of key populations for HIV in Singapore using the network scale-up method. *Sexually transmitted infections*, 95(8): 602–607.