An Alternative to Crawlers: Scalable P2P Network Monitoring Using the Network Scale-Up Method

Juan Marcos Ramirez
Sergio Díaz-Aranda
juan.ramirez@networks.imdea.org
sergio.diaz@networks.imdea.org
IMDEA Networks Institute
Leganés, Madrid, Spain

Leonhard Balduf
TU Darmstadt
Darmstadt, Germany
leonhard.balduf@kom.tu-darmstadt.de

Antonio Fernández Anta
IMDEA Software Institute & IMDEA Networks Institute
Madrid, Spain
antonio.fernandez@software.imdea.org

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 the InterPlanetary File System (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.

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 [10, 27, 37, 41] and distributed data storage systems such as the InterPlanetary File System (IPFS) [6], 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 [12, 28] and Gnutella [30]. However, measuring P2P systems at scale typically relies on active probing and crawler-based techniques [18, 19, 35] 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 [34, 35].

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 [7, 8], and tracking diseases such as HIV/AIDS and COVID-19 [17, 39]. 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 [31] 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 [36]. It has gained attention as an external storage layer for blockchain-based applications, including non-fungible token (NFT) platforms, and decentralized applications [2, 3, 11, 13, 21]. IPFS is particularly well-studied in the literature for its structure, content accessibility, and node distribution [1, 4, 5, 29, 31, 36]. 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 the 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.

2 Background

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 [24]. 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 [24], 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)* [22], 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)* [23], 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 % [9].

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 [12, 28]), but the development of Bitcoin in 2008 [27] 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 [25] 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 [18, 19, 35].

IPFS. IPFS is a contemporary P2P network for data storage and retrieval [6] 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 [36]. 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 [4, 5, 31, 36] for its participants, overlay structure, data stored, and performance. A crawler for the IPFS DHT exists [20] 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.

3 Methodology and Datasets

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 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 $C_{v,H_t}/R_{v,t}$.

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 [18, 20]. 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 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 [15]:

$$\mathrm{NB} = \frac{1}{N}\sum_{i=1}^{N} \frac{|\hat{\rho}_i - \rho|}{\rho},$$

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 [16, 34, 35].

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 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 [14] and the Udger database [38] 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 multihoming, 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.* [5]. This approach selects the most frequently observed geolocation or cloud provider for that node based on resolved IP addresses.

4 Results

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 3.2. Fig. 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, 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 2(a) and 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.* [5]. 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).

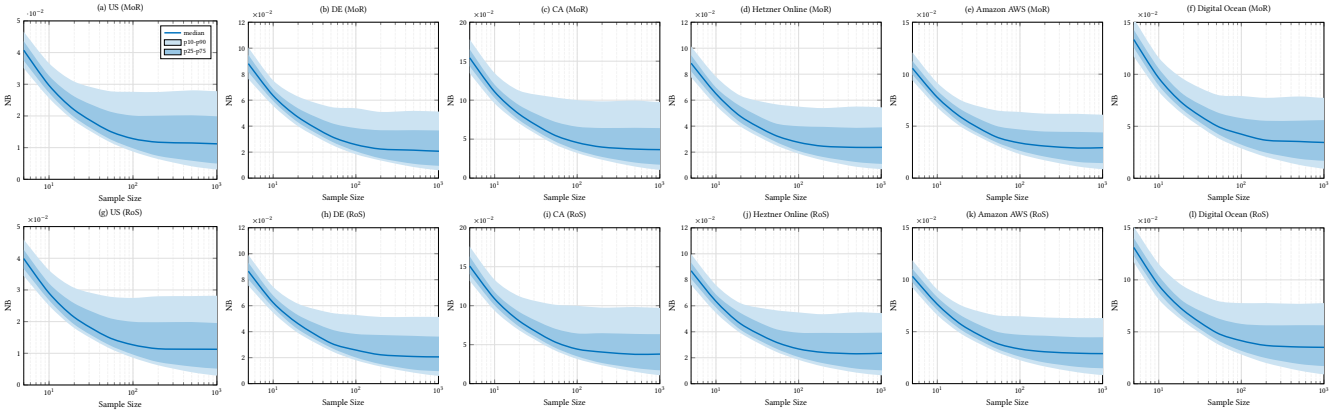

Figure 1: IPFS-Benchmark Dataset. Normalized bias (NB) of prevalence estimates as a function of sample size for six IPFS node attributes. The top row shows results are obtained with the MoR estimator, and the bottom row with the RoS estimator. Solid lines represent the median NB across the 360 snapshots, whereas the shaded regions indicate the $p25$-$p75$ and $p10$-$p90$ percentile intervals.

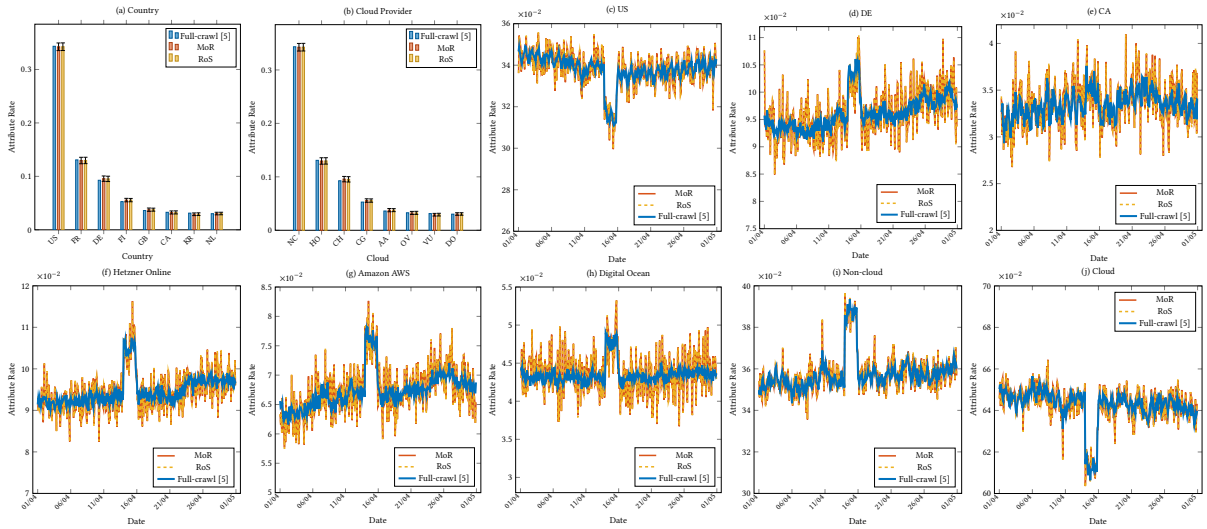

Figure 2: IPFS-Benchmark Dataset. Comparison of crawl-derived attribute prevalences with NSUM estimates (MoR and RoS) from 10 randomly sampled nodes at a snapshot captured on April 5, 2026. (a) country and (b) cloud-provider. (c)–(h) Temporal evolution of selected attributes. Evolution of (i) Non-cloud and (j) Cloud.

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, 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 1, where both estimators exhibited similar NB values.

Figures 2(c)–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 2(i) and 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.

4.2 Implementation Results

Figures 3(a) and 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 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 3(c)–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 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.

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 [24]. 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.

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. [36] 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 [19] to examine the topology and dynamics of IPFS. Henningsen et al. [20] analyzed connectivity and churn. More recently, Balduf et al. [5] 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 [26, 32, 33, 40]. 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.

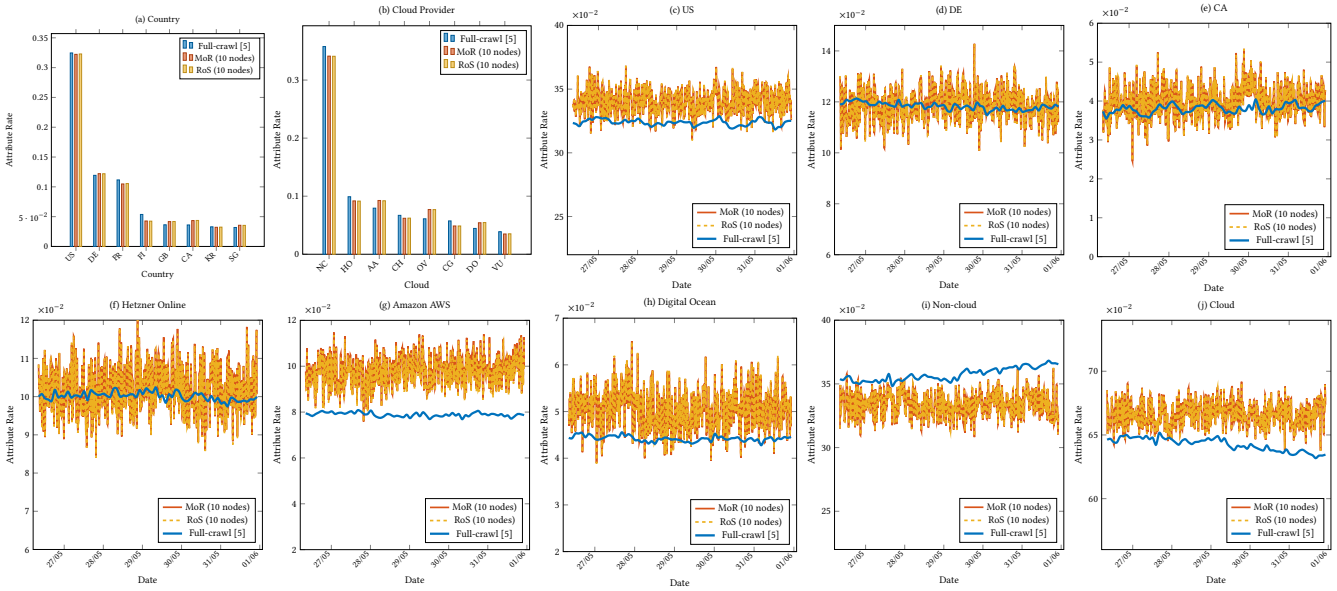

Figure 3: Bucket Dataset. Comparison of crawl-derived attribute prevalences with NSUM estimates (MoR and RoS) using the proposed measurement technique with 10 local nodes at a single snapshot. (a) country (b) cloud-provider. (c)–(h) Temporal evolution of selected attributes. Evolution of (i) Non-cloud and (j) Cloud.

7 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 slightly larger samples 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 [14, 38], 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.

Acknowledgments

The work of Sergio Díaz-Aranda has been funded by *Comunidad de Madrid* predoctoral grant PIPF-2022/COM-24467. 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 Grant Excelencia María de Maeztu CEX2024-001471-M funded by MICIU/AEI/10.13039/501100011033.

References

[1] Omar Abdullah Lajam and Tarek Ahmed Helmy. 2021. Performance Evaluation of IPFS in Private Networks. In *Proceedings of the 2021 4th International Conference on Data Storage and Data Engineering* (Barcelona, Spain) *(DSDE '21)*. Association for Computing Machinery, New York, NY, USA, 77–84. doi:10.1145/3456146.3456159

[2] Onur Ascigil, Sergi Reñé, Michał Król, George Pavlou, Lixia Zhang, Toru Hasegawa, Yuki Koizumi, and Kentaro Kita. 2019. Towards Peer-to-Peer Content Retrieval Markets: Enhancing IPFS with ICN. In *Proceedings of the 6th ACM Conference on Information-Centric Networking* (Macao, China) *(ICN '19)*. Association for Computing Machinery, New York, NY, USA, 78–88. doi:10.1145/3357150.3357403

[3] Leonhard Balduf, Martin Florian, and Björn Scheuermann. 2022. Dude, Where's my NFT: Distributed Infrastructures for Digital Art. In *Proceedings of the 3rd International Workshop on Distributed Infrastructure for the Common Good* (Quebec, Quebec City, Canada) *(DICG '22)*. Association for Computing Machinery, 1–6. doi:10.1145/3565383.3566106

[4] Leonhard Balduf, Sebastian Henningsen, Martin Florian, Sebastian Rust, and Björn Scheuermann. 2022. Monitoring Data Requests in Decentralized Data Storage Systems: A Case Study of IPFS. In *2022 IEEE 42nd International Conference on Distributed Computing Systems (ICDCS)*. IEEE, Boston, MA, USA, 658–668. doi:10.1109/ICDCS54860.2022.00069

[5] Leonhard Balduf, Maciej Korczyński, Onur Ascigil, Navin V. Keizer, George Pavlou, Björn Scheuermann, and Michał Król. 2023. The Cloud Strikes Back: Investigating the Decentralization of IPFS. In *Proceedings of the 2023 ACM Internet Measurement Conference* (Montreal QC, Canada) *(IMC '23)*. Association for Computing Machinery, New York, NY, USA, 391–405. doi:10.1145/3618257.3624797

[6] Juan Benet. 2014. IPFS - Content Addressed, Versioned, P2P File System. *CoRR* abs/1407.3561 (2014). arXiv:1407.3561 http://arxiv.org/abs/1407.3561

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

[8] H Russell Bernard, Peter D Killworth, Eugene C Johnsen, Gene A Shelley, and Christopher McCarty. 2001. Estimating the ripple effect of a disaster. *Connections* 24, 2 (2001), 18–22.

[9] Emily Breza, Arun G Chandrasekhar, Tyler H McCormick, and Mengjie Pan. 2020. Using aggregated relational data to feasibly identify network structure without network data. *American Economic Review* 110, 8 (2020), 2454–84.

[10] Vitalik Buterin. 2014. Ethereum: A Next-Generation Smart Contract and Decentralized Application Platform. (2014). https://ethereum.org/content/whitepaper/whitepaper-pdf/Ethereum_Whitepaper_-_Buterin_2014.pdf

[11] Yongle Chen, Hui Li, Kejiao Li, and Jiyang Zhang. 2017. An improved P2P file system scheme based on IPFS and Blockchain. In *2017 IEEE International Conference on Big Data (Big Data)*. IEEE, Boston, MA, USA, 2652–2657. doi:10.1109/BigData.2017.8258226

[12] Bram Cohen. 2003. Incentives Build Robustness in BitTorrent. In *Workshop on Economics of Peer-to-Peer systems*, Vol. 6. 68–72. http://jmvidal.cse.sc.edu/library/cohen03a.pdf

[13] Dipanjan Das, Priyanka Bose, Nicola Ruaro, Christopher Kruegel, and Giovanni Vigna. 2022. Understanding Security Issues in the NFT Ecosystem. In *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security* (Los Angeles, CA, USA) *(CCS '22)*. Association for Computing Machinery, New York, NY, USA, 667–681. doi:10.1145/3548606.3559342

[14] Maxmind Developers. 2023. GeoIP and GeoLite. https://dev.maxmind.com/geoip/. Accessed: 2025-10-13.

[15] Sergio Díaz-Aranda, Juan Marcos Ramírez, Jose Aguilar, Rosa E Lillo, and Antonio Fernández Anta. 2026. Robust network scale-up method estimators. *Social Networks* 84 (2026), 46–61.

[16] Jarret Falkner, Michael Piatek, John P. John, Arvind Krishnamurthy, and Thomas E. Anderson. 2007. Profiling a million user DHT. In *Proceedings of the 7th ACM SIGCOMM Internet Measurement Conference* (San Diego, CA, USA, 2007-10-24) *(IMC 2007)*, Constantine Dovrolis and Matthew Roughan (Eds.). ACM, 129–134. doi:10.1145/1298306.1298325

[17] Augusto Garcia-Agundez, Oluwasegun Ojo, Harold A Hernández-Roig, Carlos Baquero, Davide Frey, Chryssis Georgiou, Mathieu Goessens, Rosa E Lillo, Raquel Menezes, Nicolas Nicolaou, et al. 2021. Estimating the COVID-19 prevalence in Spain with indirect reporting via open surveys. *Frontiers in Public Health* 9 (2021), 658544.

[18] Trudi Group. 2020. Libp2p-Crawler. https://github.com/trudi-group/ipfs-crawler. Accessed: 2025-10-13.

[19] Trudi Group. 2023. IPFS Metrics Exporter Plugin. https://github.com/trudi-group/ipfs-metric-exporter. Accessed: 2025-10-13.

[20] Sebastian Henningsen, Martin Florian, Sebastian Rust, and Björn Scheuermann. 2020. Mapping the Interplanetary Filesystem. In *2020 IFIP Networking Conference (Networking)*. IEEE, Boston, MA, USA, 289–297.

[21] Van-Hoan Hoang, Elyes Lehtihet, and Yacine Ghamri-Doudane. 2020. Privacy-Preserving Blockchain-Based Data Sharing Platform for Decentralized Storage Systems. In *2020 IFIP Networking Conference (Networking)*. IEEE, Boston, MA, USA, 280–288.

[22] Peter D Killworth, Eugene C Johnsen, Christopher McCarty, Gene Ann Shelley, and H Russell Bernard. 1998. A social network approach to estimating seroprevalence in the United States. *Social networks* 20, 1 (1998), 23–50.

[23] Peter D Killworth, Christopher McCarty, H Russell Bernard, Gene Ann Shelley, and Eugene C Johnsen. 1998. Estimation of seroprevalence, rape, and homelessness in the United States using a social network approach. *Evaluation review* 22, 2 (1998), 289–308.

[24] Ian Laga, Le Bao, and Xiaoyue Niu. 2021. Thirty Years of The Network Scale-up Method. *J. Amer. Statist. Assoc.* 116, 535 (2021), 1548–1559. arXiv:https://doi.org/10.1080/01621459.2021.1935267 doi:10.1080/01621459.2021.1935267 PMID: 37994314.

[25] Petar Maymounkov and David Mazières. 2002. Kademlia: A Peer-to-Peer Information System Based on the XOR Metric. In *Peer-to-Peer Systems, First International Workshop, IPTPS 2002, Cambridge, MA, USA, March 7-8, 2002, Revised Papers (Lecture Notes in Computer Science, Vol. 2429)*, Peter Druschel, M. Frans Kaashoek, and Antony I. T. Rowstron (Eds.). Springer, 53–65.

[26] Ghulam Memon, Reza Rejaie, Yang Guo, and Daniel Stutzbach. 2009. Large-scale monitoring of DHT traffic. In *Proceedings of the 8th international conference on Peer-to-peer systems, IPTPS'09, Boston, MA, USA, April 21, 2009*, Rodrigo Rodrigues and Keith W. Ross (Eds.). USENIX, 11. http://www.iptps.org/papers-2009/memon.pdf

[27] Satoshi Nakamoto. 2008. Bitcoin: A Peer-to-Peer Electronic Cash System. (2008).

[28] Johan Pouwelse, Paweł Garbacki, Dick Epema, and Henk Sips. 2005. The Bittorrent P2P File-Sharing System: Measurements and Analysis. In *P2P Systems IV* (Berlin, Heidelberg, 2005) *(Lecture Notes in Computer Science)*. Springer, 205–216. doi:10.1007/11558989_19

[29] Bernd Prünster, Alexander Marsalek, and Thomas Zefferer. 2022. Total eclipse of the heart–disrupting the InterPlanetary file system. In *31st USENIX Security Symposium (USENIX Security 22)*. USENIX Association, Boston, MA, USA, 3735–3752.

[30] Matei Ripeanu, Ian T. Foster, and Adriana Iamnitchi. 2002. Mapping the Gnutella Network: Properties of Large-Scale Peer-to-Peer Systems and Implications for System Design. *CoRR* cs.DC/0209028 (2002). https://arxiv.org/abs/cs/0209028

[31] Ruizhe Shi, Ruizhi Cheng, Bo Han, Yue Cheng, and Songqing Chen. 2024. A Closer Look into IPFS: Accessibility, Content, and Performance. *Proc. ACM Meas. Anal. Comput. Syst.* 8, 2, Article 20 (May 2024), 31 pages. doi:10.1145/3656015

[32] Moritz Steiner, Ernst W. Biersack, and Taoufik En-Najjary. 2007. Actively Monitoring Peers in KAD. In *6th International workshop on Peer-To-Peer Systems, IPTPS 2007* (Bellevue, WA, USA, 2007-02-26), John R. Douceur and Roger Wattenhofer (Eds.).

[33] Moritz Steiner, Taoufik En-Najjary, and Ernst W. Biersack. 2007. A global view of KAD. In *Proceedings of the 7th ACM SIGCOMM Internet Measurement Conference, IMC 2007* (San Diego, CA, USA, 2007-10-24). ACM, 117–122. doi:10.1145/1298306.1298323

[34] Daniel Stutzbach and Reza Rejaie. 2006. 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* (Barcelona, Catalunya, Spain, 2006-04-23). IEEE. doi:10.1109/INFOCOM.2006.347

[35] Daniel Stutzbach and Reza Rejaie. 2006. Understanding churn in peer-to-peer networks. In *Proceedings of the 6th ACM SIGCOMM Internet Measurement Conference, IMC 2006, Rio de Janeriro, Brazil, October 25-27, 2006*, Jussara M. Almeida, Virgílio A. F. Almeida, and Paul Barford (Eds.). ACM, 189–202. doi:10.1145/1177080.1177105

[36] Dennis Trautwein, Aravindh Raman, Gareth Tyson, Ignacio Castro, Will Scott, Moritz Schubotz, Bela Gipp, and Yiannis Psaras. 2022. Design and evaluation of IPFS: a storage layer for the decentralized web. In *Proceedings of the ACM SIGCOMM 2022 Conference* (Amsterdam, Netherlands) *(SIGCOMM '22)*. Association for Computing Machinery, New York, NY, USA, 739–752. doi:10.1145/3544216.3544232

[37] Florian Tschorsch and Björn Scheuermann. 2016. Bitcoin and Beyond: A Technical Survey on Decentralized Digital Currencies. *IEEE Commun. Surveys Tuts.* 18, 3 (2016), 2084–2123. doi:10.1109/COMST.2016.2535718

[38] Udger. 2026. Udger Data v4. https://udger.com/. Accessed: 2026-05-15.

[39] WHO UNAIDS. 2010. Guidelines on Estimating the Size of Populations most at Risk to HIV. Geneva: World Health Organization.

[40] Liang Wang and Jussi Kangasharju. 2013. Measuring large-scale distributed systems: case of BitTorrent Mainline DHT. In *Proc. of P2P*. IEEE, 1–10. doi:10.1109/P2P.2013.6688697

[41] Gavin Wood et al. 2014. Ethereum: A secure decentralised generalised transaction ledger. *Ethereum project yellow paper* (2014).

A Nodes with and without NAT

Figures A1(a) and A1(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 A1(c)-A1(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 A2(a) and A2(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 A2(c)–A2(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 a small observation set, rather than significant estimation bias.

Unlike the NATed-node experiment shown in Fig. A1(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.

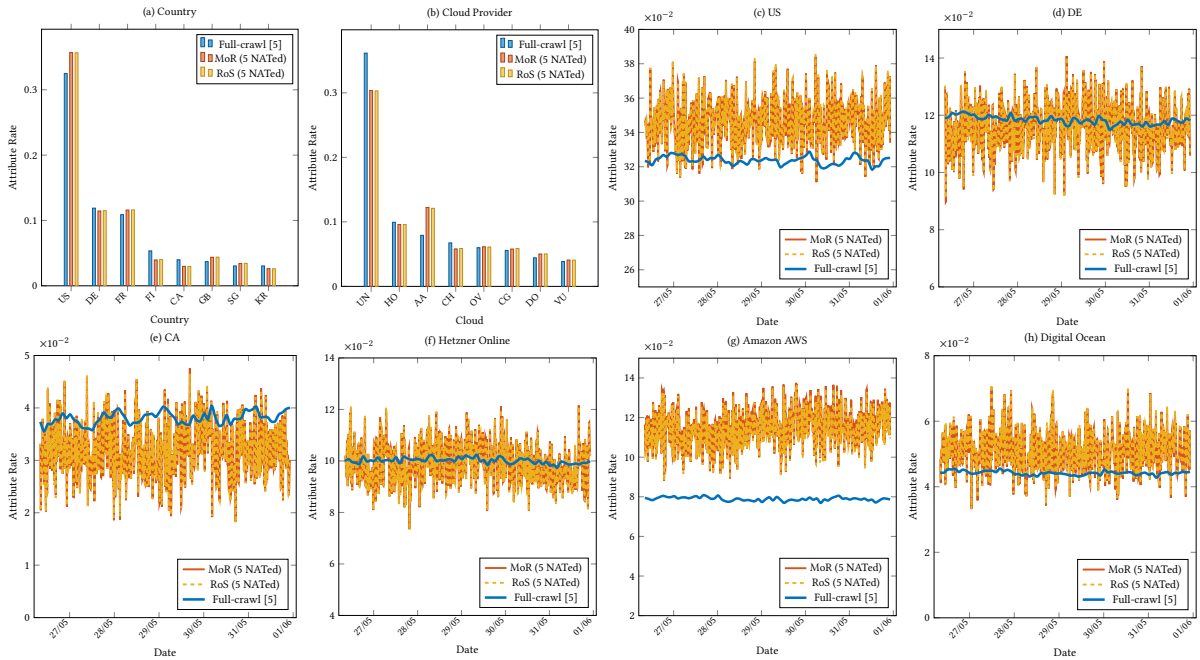

Figure A1: Bucket Dataset. Comparison of crawl-derived attribute prevalences with NSUM estimates (MoR and RoS) using the proposed measurement technique with 5 NATed nodes at a single snapshot. (a) country (b) cloud-provider. (c)–(h) Temporal evolution of selected attributes.

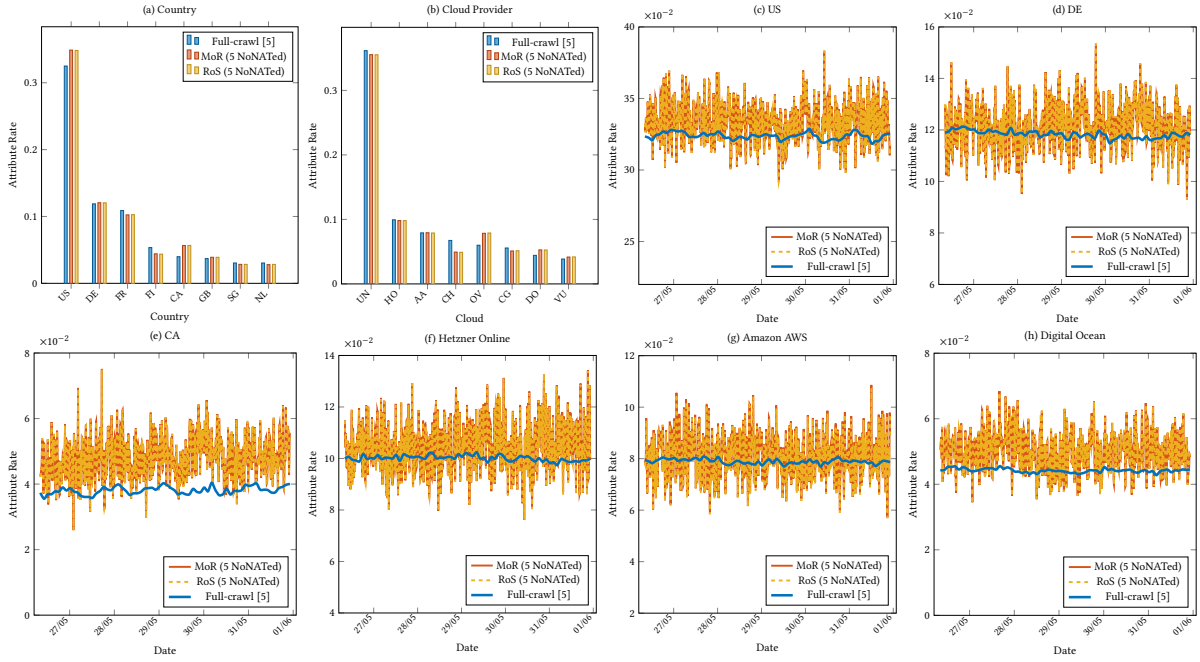

Figure A2: Bucket Dataset. Comparison of crawl-derived attribute prevalences with NSUM estimates (MoR and RoS) using the proposed measurement technique with 5 NoNATed nodes at a single snapshot. (a) country (b) cloud-provider. (c)–(h) Temporal evolution of selected attributes.