Proceedings of the Eighteenth International AAAI Conference on Web and Social Media (ICWSM 2024)

Tight Sampling in Unbounded Networks

Kshitijaa Jaglan[1*], **Meher Chaitanya Pindiprolu**[2*], **Triansh Sharma**[1], **Abhijeeth Reddy Singam**[1],
Nidhi Goyal[3], **Ponnurangam Kumaraguru**[1], **Ulrik Brandes**[2]
[1] IIIT Hyderabad
[2] Social Networks Lab, ETH Zürich
[3] IIIT Delhi
kshitijaa.jaglan@research.iiit.ac.in, mpindiprolu@ethz.ch, trianshsharma@gmail.com, singam.abhijeeth@gmail.com,
nidhig@iiitd.ac.in, pk.guru@iiit.ac.in, ubrandes@ethz.ch

Abstract

The default approach to deal with the enormous size and limited accessibility of many Web and social media networks is to sample one or more subnetworks from a conceptually unbounded unknown network. Clearly, the extracted subnetworks will crucially depend on the sampling scheme. Motivated by studies of homophily and opinion formation, we propose a variant of snowball sampling designed to prioritize the inclusion of entire cohesive communities rather than any kind of representativeness, breadth, or depth of coverage. The method is illustrated on a concrete example, and experiments on synthetic networks suggest that it behaves as desired.

Introduction

Online social networks such as Twitter are a valuable source of information for research on various questions in the social sciences, not least because they contain vast amounts of process-generated data (Antonakaki, Fragopoulou, and Ioannidis 2021). However, obtaining full-size networks from platforms is impossible for researchers due to access limitations and is prohibitive due to its volume. Sampling from massive online networks with millions or even billions of users thus presents a fundamental challenge for social network research (Ruths and Pfeffer 2014).

Typical sampling schemes rely on one or both of two main techniques (Ahmed, Neville, and Kompella 2013): retrieval based on attributes such as demographics or tweet content (node-based or edge-based sampling) and seed-set expansion by following incoming or outgoing links (topology-based sampling) (Kim et al. 2018; Leskovec and Faloutsos 2006).

Seed-set expansion is related to graph exploration and snowball sampling, where elements of a network of unknown size are discovered only through adjacency with already explored parts. If the underlying networks exhibit small-world characteristics, as many social networks do, the boundary of connectivity-based sampling methods quickly covers distant parts of the network. When the research goal is to study homophily and other social regularities, completeness of cohesive groups is a more important sampling criterion than coverage of the network. Our problem is, therefore, closely related to local clustering with seeds and especially relevant in conceptually unbounded networks such as Twitter.

To this end, we propose a novel snowball-type sampling scheme that is designed to prioritize sampling within the cohesive subgroups or local clusters around a given set of seed nodes in the (multiplex) network. The approach thus resembles seeded community detection, where the objective is to determine a locally dense subgraph containing a seed node or a set of seed nodes, except that in our case, the graph is only partially known. Common clustering objectives such as low conductance or high modularity (Chang et al. 2019; Zhang and Rohe 2018; Newman 2006) are difficult to optimize in such settings, because the use of methods such as approximate PageRank (Andersen, Chung, and Lang 2006) or random-walk techniques (Spielman and Teng 2004) require large parts of the graph in which a seeded community resides to be available. Without such information, and confronted with rapid expansion of the boundary around the sampled network, we prioritize the selection of nodes based on their likeliness to add to cohesive groups in the sample.

Our approach generalizes a technique known as maximum adjacency search (Cai and Matula 1993) that has been used prominently to find minimum graph cuts by repeatedly expanding from a seed node (Stoer and Wagner 1997). We replace the basic maximum-adjacency criterion with a generalized priority obtained from a combination of different forms of interaction in social media, such as likes, retweets, replies, and quotes, with empirically calibrated weights. Specifically for sampling subnetworks on Twitter, we prioritize profiles outside the current sample set that show maximum levels of engagement with profiles inside. The evolution of sampled networks is demonstrated on empirical and synthetic data, and we conclude that our method effectively prioritizes local clusters around seeds.

In summary, our main contributions are as follows.

- Application of a maximum-adjacency principle to snowball sampling to expand seed sets while staying within local communities.
- Generalization of the maximum-adjacency criterion to weighted multiplex networks. Specifically for Twitter, we propose an empirically calibrated weighting scheme

---

[*] These authors contributed equally.

to combine types of interaction.

- Provision of a Twitter dataset focusing on the interactions within communities that engage with a publicly available set of influential profiles.

Sampling Social Media Networks

The majority of methods used for social media networks begin by sampling a set of profiles, links, or interactions and expand the network by traversing (parts of) their neighborhoods. In this section, we provide a brief overview of various relevant sampling approaches.

Node-based sampling methods: These can be as simple as the subgraph induced by a set of nodes sampled uniformly at random (Ahmed, Neville, and Kompella 2013). This technique is simple to approximate the direct properties of nodes, such as degree distribution, but it does not preserve the connectivity. More sophisticated node-based sampling approaches, such as the PageRank and PageRank-with-restarts methods, choose sampled nodes based on their PageRank scores and construct the induced subgraph on them (Rozemberczki, Kiss, and Sarkar 2020).

Traversal-based sampling methods: Given one or more starting nodes ("seeds"), these are strategies for adding more nodes by traversing the underlying graph from already sampled nodes. Some strategies use traditional graph traversals such as breadth-first and depth-first search (Giudice and Ursino 2019). In each iteration, a new node is chosen depending on its earliest (breadth-first) or latest (depth-first) discovery time. Generally, a breadth-first search can result in a denser cover and has been shown to be biased towards high-degree nodes (Ye, Lang, and Wu 2010). The work by Kurant, Markopoulou, and Thiran (2011) addressed this bias by suggesting analytical solutions to correct it. Snowball sampling is another traversal-based sampling strategy that maintains the network connectivity using the breadth-first approach but suffers from boundary bias where the peripheral nodes sampled in the final iteration have many missing neighbors (Lee, Kim, and Jeong 2006). A large class of traversal-based sampling strategies are based on Random Walks (RW). RW sampling techniques commence a random walk (single or multi-dimensional) starting from seed nodes and construct a Markov chain by iteratively choosing a random neighbor (Gjoka et al. 2010; Ribeiro and Towsley 2010; Avrachenkov, Ribeiro, and Towsley 2010). These techniques are inherently biased towards high-degree nodes (Hu and Lau 2013). Metropolis-Hastings random walk sampling strategies overcome this bias by making the random walker visit low-degree nodes (Hübler et al. 2008; Stutzbach et al. 2006; Li et al. 2015). Liu et al. (2019) incorporates a novel hybrid jump mechanism in Metropolis-Hastings random walk to avoid repetitive sampling within a small connected component. Forest Fire sampling, a hybrid of random walk-based methods and snowball sampling, expands by burning a fraction of the outgoing links for each sampled node (Leskovec and Faloutsos 2006). This fraction is drawn randomly from a geometric distribution with mean $\frac{p}{1-p}$ (the recommended value of p is 0.7, implying that, on average, each selected node burns 2.33 neighbors). Maiya and Berger-Wolf (2010) proposed a community-preserving sampling approach by utilizing concepts from expander graphs to sample representative subgraphs that reflect the community structure of the original network by greedily constructing the sample with maximal expansion. Recently, Zhang et al. (2023) introduced expansion strategies for detecting clusters around seed nodes. These strategies involve including nodes in the sample through specific expansion techniques based on edge connectivity. However, all these approaches require large parts of the graph surrounding the seeded nodes to be available. In the next section, we provide a methodology to overcome the uncertainty of the unknown or unboundedness of the network to make the sampler stay within cohesive subgroups surrounding seeds.

Tight Sampling

Our goal is to sample subgraphs of social media networks in such a way that cohesive communities are covered in larger parts before expanding further into the underlying network. We refer to this as *tight* sampling. Since the network is assumed to be much larger than the targeted sample size, say, all of Twitter, we think of it as unbounded.

Formally, we assume the existence of an infinite, initially unknown directed graph $G = (V, E)$ representing a vast social media network. Edges represent social relations between members of the network and will be described more concretely below, where we also introduce edge weights. We further assume that knowing a vertex $v \in V$, we can also obtain the set $N^-(v) = \{u \in V : (u, v) \in E\}$ of in-neighbors with edges directed to v; the set of out-neighbors $N^+(v)$ is defined symetrically.

Given a finite set $V_s \subset V$ of *seed* vertices, we want to extract a subgraph $G[S]$ induced by a finite set of sample vertices $S \subset V$ that includes the seeds, $V_s \subseteq S$. Starting from the seeds, vertices are sampled one at a time, and each newly sampled vertex must be an in-neighbor of a vertex sampled earlier. In other words, we aim for a sampling strategy that traverses edges backward. Thus, we successively add vertices that relate to those already included.

For notational simplicity, we omit timestamps and refer to the set of currently sampled vertices, or *insiders*, as S. Candidate vertices that may be sampled next are all in-neighbors $N^-(S) = \bigcup_{v \in S} N^-(v)$ not yet in S. We refer to the vertices in $N^-(S) \setminus S$ as *outsiders*.

The *boundary* $\partial(S)$ of a current sample S is the set of all edges directed from outsiders to insiders, i.e., the edges crossing a *directed cut*. Since our objective is to keep this boundary small, we sample outsiders that have the maximum number of edges directed to insiders. This is a directed version of the maximum-adjacency search and greedily removes edges from the boundary. Note that we do not know the in-neighborhood of a vertex prior to its sampling, so we can not guarantee whether the new boundary will be the smallest possible.

In summary, we sample a vertex-induced subgraph by expanding a set of seed vertices one vertex at a time, where the vertex selected is the outsider with the largest number of edges directed to insiders, i.e., by maximum-adjacency search. In the next section, we extend this principle to weighted graphs that integrate multiple types of relations and interactions in social media networks, and then validate the outcome.

Weighted Edges from Multiplex Relations

Social media typically combine multiple types of relations such as friending or following with interactions such as liking or forwarding. In order to sample subgraphs in which the most cohesively related groups are relatively intact, we propose an empirically calibrated aggregation into a single weighted relation. This will allow for straightforward generalization of the maximum-adjacency principle from counting edges to the sum of their weights.

As detailed in the following three subsections, weights are computed by deciding first on the patterns of interaction to distinguish, and then combining their re-scaled frequencies of occurrence.

Interaction Patterns

Because of our specific interest in social influence on Twitter, we consider four kinds of relations as indicators of engagement with information shared by other users via tweets: likes, retweets, replies, and quotes. First, the interaction pattern of a user i with a tweet t authored by i is represented by the characteristic vector $I_t(i,j) = x \in X$ of interaction types, $x \in X = \{0,1\}^4$. Here, binary values $\{0,1\}$ denote the presence or absence of a particular form of engagement from the set $\{$*like, retweet, reply, quote*$\}$. For example, if a user j retweets and quotes a tweet t of user i, there is a directed edge from j to i labeled with interaction pattern $I_t(i,j) = 0101$. We omit indices i and j if they are clear from the context. Note that for a single tweet and interacting user, we only consider the presence or absence of forms of engagement, not the number of their respective instances.

Frequency of Occurrence

When counting interaction patterns it is sometimes desirable to count occurrences of one pattern also toward the frequency of another, because it may or may not matter whether additional types of interaction are present. We distinguish three cases.

Distinct interaction patterns. A pair of a tweet and interacting user contributes to the frequency of an interaction pattern only if the user engages with the tweet in exactly this pattern. A user's engagement is counted as an occurrence of pattern $x = 1100$, for instance, if and only if the user *likes* and *retweets* and does not *reply* or *quote*.

Nested interaction patterns. A pair of a tweet and interacting user contributes to the frequency of an interaction pattern if the user engages with the tweet that includes this interaction pattern. A user's engagement is counted as an occurrence of pattern $x = 1100$, for instance, if and only if the user *likes* and *retweets* and does or does not *reply* or *quote*.

Audience-facing interactions (A-F). We posit that likes and replies are more personal forms of interaction and usually directed at the author of a tweet, whereas retweets and quotes tend to be aimed at visibility by signaling an interaction to followers. We therefore introduce a third method of counting by treating retweets and quotes as interchangeable types of interaction. We thus have $X = \{001, 010, 011, 100, 101, 110, 111\}$, reducing the effective number of patterns from 15 to seven. Merging of retweets and quotes has been applied in other studies, for instance, on the Higgs Boson Twitter dataset (De Domenico et al. 2013).

Importance Scaling

Interaction types occur at different rates and potentially signal different levels of engagement. Liking is the most frequent form of interaction but is therefore assumed to be less informative than, say, quoting. To determine the relative importance of interaction patterns, we therefore first assess their empirical prevalence and then assign a weight inversely proportional to it.

Assume we are given an empirical sample S of insiders as well as their tweets, and the interactions with them. Denote by $T \supseteq S$ the set of interacting users. Furthermore, let $n(i,x,j)$ denote the number of times that any user $j \in T$ engaged with the tweets of a user $i \in S$ using interaction pattern $x \in X$, and let $N = \sum_{i,x,j} n(i,x,j)$ be the overall number of pattern occurrences. Recall that, say, multiple replies of the same user to the same tweet are counted only once. To derive importance weights for the types of interaction, we first distinguish three approaches to normalizing frequencies.

Global normalization. Ignoring the users involved, the overall frequency of interaction pattern $x \in X$ is given by $n(x) = \sum_{i \in S, j \in T} n(i,x,j)$. The relative frequency of pattern x, normalized globally, is then defined as

$$\eta(x) = \frac{n(x)}{N} \ .$$

Source normalization. Users spreading information may see very different patterns of engagement with their tweets. An alternative approach is, therefore to normalize interaction patterns by the average engagement that sources of information receive,

$$\overleftarrow{\eta}(x) = \frac{1}{|S|} \sum_{i \in S} \frac{n(i,x)}{\overleftarrow{N}(i)}$$

where $n(i,x) = \sum\limits_{j \in T} n(i,x,t)$ and $\overleftarrow{N}(i) = \sum\limits_{x \in X} n(i,x)$.

Target normalization. Symmetrically, users interacting with information published by others may exhibit very different patterns of engagement. An alternative approach is to normalize interaction patterns by the average engagement that consumers of information display,

$$\overrightarrow{\eta}(x) = \frac{1}{|T|} \sum_{j \in T} \frac{n(x,j)}{\overrightarrow{N}(j)}$$

where $n(x, j) = \sum_{i \in S} n(i, x, t)$ and $\overrightarrow{N}(j) = \sum_{x \in X} n(x, j)$.

For the purpose of this paper, we are balancing all of the above three perspectives by determining a distribution that minimizes the sum of squared errors with respect to the alternatives, i.e., we find a non-negative vector $\eta^*(x)$ for the set of patterns X such that

$$\sum_{x \in X} (\eta^*(x) - \eta(x))^2 + (\eta^*(x) - \overleftarrow{\eta}(x))^2 + (\eta^*(x) - \overrightarrow{\eta}(x))^2$$

is minimum. In the spirit of Horvitz-Thompson importance sampling, we finally determine influence weights $\omega(x)$ for the interaction patterns as the inverse of their balanced normalized frequencies,

$$\omega(x) = \frac{1}{\eta^*(x)}$$

In practice, we use entries $\omega^*(x)$ rounded to two decimals for simplicity and robustness.

Table 1 shows the calculated weights for a Twitter dataset. More information about the users has been provided in later sections.

Experiments

We evaluate our sampling strategy by first creating synthetic networks in a controlled setting using stochastic blockmodels (SBM) (Holland, Laskey, and Leinhardt 1983). This allows us to generate networks with predefined communities for which we can monitor how they are sampled. Following this controlled assessment, we proceed to evaluate our sampling approach in an empirical context by expanding a seed set on Twitter into a directed weighted network. Our findings on both synthetic data and the empirical network indicate that our sampling strategy improves the coverage of cohesive communities when contrasted with random-based sampling approaches.

Synthetic Data

In this section, we will describe the process of generating synthetic network data with planted communities. These networks constitute the simplest meaningful situations in which the evolution of our sample can be observed most clearly.

Instances. We explore different networks generated using SBM by varying block sizes, inter/intra block densities, and seed node distributions. Specifically, we explore three distinct block size settings: (1) four blocks of sizes {400, 800, 1200, 1600}, (2) four blocks of sizes {800, 1200, 1600, 2000}, and (3) eight blocks of 1000 nodes each. For these three configurations, we derive the block probability matrix with consistent average degrees within each block ($\langle k' \rangle$) and a uniform ratio of intra-block to inter-block edges, denoted as r, across all blocks. In this context, we define several key parameters: n represents the total number of nodes in the SBM, ρ_{ij} signifies the inter-cluster probability between block i and block j, n_i denotes the size of the i^{th} block, m_{ii} represents the number of edges within block i, $m_{i,*}$ indicates the total count of edges between block i and all other blocks, and $\rho_{i,*}$ denotes the approximate density between block i and the other blocks. Given a specific configuration characterized by $\langle k' \rangle$, r, and b blocks with specified sizes, we derive the block probability matrix P as follows:

In the case of the diagonal elements of the matrix P, the value of ρ_{ii} is straightforwardly calculated as

$$\rho_{ii} = \frac{\langle k' \rangle}{n_i - 1}$$

However, for non-diagonal elements, we determine the value of ρ_{ij} using the ratio r, which represents the proportion of intra-block to inter-block edges as follows.

$$\begin{aligned} m_{ii} &= r \cdot m_{i,*} \\ \frac{n_i \cdot \langle k' \rangle}{2} &= r \cdot \rho_{i,*} \cdot n_i \cdot (n - n_i) \\ \rho_{i,*} &= \frac{\langle k' \rangle}{2 \cdot r \cdot (n - n_i)} \end{aligned}$$

In the case of a block (i, j) where $i \neq j$, since the previously calculated value is non-symmetric, we derive the final value for the respective cell by averaging ρ values with respect to both the row and column:

$$\rho_{ij} = \frac{\rho_{i,*} + \rho_{j,*}}{2} \tag{1}$$

where ρ_{ij} represents the value in block (i, j) for the block transition matrix P.

For our study, we have chosen the following values for r: $\frac{1}{b-1}$, 0.5, 1, 2, 4, 8. These values have been carefully selected to facilitate an evaluation of the sampler's performance across a spectrum of community structure definitions. The minimum value of r corresponds to a scenario in which, for each edge within block i, there are approximately $b - 1$ edges connecting block i to the other blocks. In this particular scenario, we observe a lack of distinct community structure, representing a case where our sampler struggles to identify clear community boundaries. By varying the values of r, we can effectively demonstrate the gradual changes in the sampler's performance.

It is important to highlight that while the intra-block average degree is fixed at $\langle k' \rangle$, the average degree of the entire network can vary due to the presence of inter-block edges (determined by r). Nevertheless, the process maintains uniform degree distributions across all blocks, ensuring that the sampler's preferences are not influenced solely by the presence of higher or lower-degree nodes in specific communities. In our study, we set the value of $\langle k' \rangle$ to 10 for all SBM configurations.

Selection of seeds. In the case of the first two network configurations, which consist of four blocks each, we conducted experiments involving varying numbers of seed nodes per block. Specifically, we selected 20 nodes per block from two blocks at a time, totaling 6 possible combinations. We repeated this process with 50 nodes per block, resulting

Interaction type	$\eta(x)$	$\overleftarrow{\eta}(x)$	$\overrightarrow{\eta}(x)$	$\eta^*(x)$	$\omega(x)$	$\omega^*(x)$
0001	2.2560	2.0575	1.4276	1.9137	0.5225	*0.52*
0010	7.9125	5.5468	5.9666	6.4753	0.1544	0.15
0011	0.3272	0.0645	0.0367	0.0047	22.3920	22.4
0100	6.0684	6.8172	6.2281	6.3712	0.1569	0.16
0101	0.0687	0.1092	0.0864	0.0881	11.3490	11.35
0110	0.0860	0.1641	0.1015	0.1172	8.5331	8.53
0111	0.0018	0.0034	0.0031	0.0028	360.1600	**360.1**
1000	72.3500	71.6710	72.4130	72.1440	0.0139	0.014
1001	0.3707	0.5174	0.3457	0.4113	2.4314	2.43
1010	1.0871	1.7172	1.3212	1.3752	0.7272	0.73
1011	0.0154	0.0172	0.0188	0.0171	58.4740	58.5
1100	9.3286	10.6870	11.3950	10.4700	0.0955	0.095
1101	0.1410	0.2118	0.1699	0.1743	5.7386	5.74
1110	0.2730	0.3984	0.4648	0.3787	2.6402	2.64
1111	0.0084	0.0169	0.0207	0.0153	65.1750	65.2

(a) Weights for Distinct interaction patterns

Interaction type	$\eta(x)$	$\overleftarrow{\eta}(x)$	$\overrightarrow{\eta}(x)$	$\eta^*(x)$	$\omega(x)$	$\omega^*(x)$
0001	2.8947	2.9979	2.1090	2.6672	0.3749	0.37
0010	9.4169	7.9285	7.9333	8.4263	0.1186	0.12
0011	0.0583	0.1020	0.0793	0.00799	12.5190	12.52
0100	15.9760	18.4080	18.4700	17.6180	0.0568	0.06
0101	0.2199	0.3414	0.2801	0.2805	3.5652	3.6
0110	0.3692	0.5828	0.5901	0.5141	1.9452	1.95
0111	0.0102	0.0203	0.0238	0.0181	55.1890	55.2
1000	83.5740	85.2370	86.1500	84.9870	0.0117	0.01
1001	0.5355	0.7633	0.5551	0.6180	1.6182	1.6
1010	1.3840	2.1497	1.8254	1.7864	0.5598	0.6
1011	0.0238	0.0341	0.0395	0.0324	30.8210	30.8
1100	9.7510	11.3150	12.0510	11.0390	0.0906	0.09
1101	0.1495	0.2287	0.1906	0.1896	5.2742	5.3
1110	0.2815	0.4153	0.4855	0.3941	2.5375	2.5
1111	0.0084	0.0169	0.0207	0.0153	65.1750	**65.2**

(b) Weights for Nested interaction patterns

Interaction type	$\eta(x)$	$\overleftarrow{\eta}(x)$	$\overrightarrow{\eta}(x)$	$\eta^*(x)$	$\omega(x)$	$\omega^*(x)$
001	19.0910	21.7480	20.8590	20.5660	0.0486	0.05
010	9.4170	7.9285	7.9333	8.4267	0.1187	0.12
011	0.4378	0.7052	0.6932	0.6121	1.6338	1.6
100	83.5740	85.2370	86.1500	84.9870	0.0117	0.01
101	10.4360	12.3060	12.7960	11.8460	0.0844	0.08
110	1.3840	2.1496	1.8254	1.7864	0.5599	0.6
111	0.3137	0.4663	0.5456	0.4419	2.2624	2.3

(c) Weights for Audience-facing interaction patterns

Table 1: Weights per interaction pattern involving $\{$*like, retweet, reply, quote*$\}$ for (a) distinct and (b) nested, and $\{$*like, reply, retweet+quote*$\}$ for (c) audience-facing interaction patterns respectively

in a total of 12 possible combinations. Furthermore, we explored the scenario in which each community was planted with 20 seed nodes, and similarly, we conducted experiments with 50 nodes per block.

We conducted experiments with different seed node configurations in the network comprising eight communities, each containing 1000 nodes. These configurations included $[1] * 8$, $[10] * 8$, $[20] * 2$, and $[20] * 3$, where the notation $[i] * j$ indicates that there are i seed nodes in each of the j blocks, with the remaining blocks having zero seed nodes. We utilize random sampling to obtain the requisite number of nodes per block for all the aforementioned seed node configurations. Alternatively, we considered selecting nodes based on their degree centrality, both low and high, but throughout our experiments, we did not observe any significant disparities in the results.

Sampling: For a given synthetic defined by its P matrix and selection of seed users, we sample new nodes by employing the following expansion-based strategies:

1. **Maximum Adjacency Search (MAS)**. This strategy selects an outsider (non-seed node) with the highest number of edges incident to the insider set.
2. **Random Insider and MAS (RI_MAS).** This strategy randomly selects an insider, i, and selects an outsider incident to i based on maximum adjacency search.
3. **Random Outsider (RO).** This strategy randomly samples an outsider with uniform probability from the set of outsiders.
4. **Random Insider and Random Outsider (RI_RO).** We randomly select an insider followed by a random outsider incident to this insider.

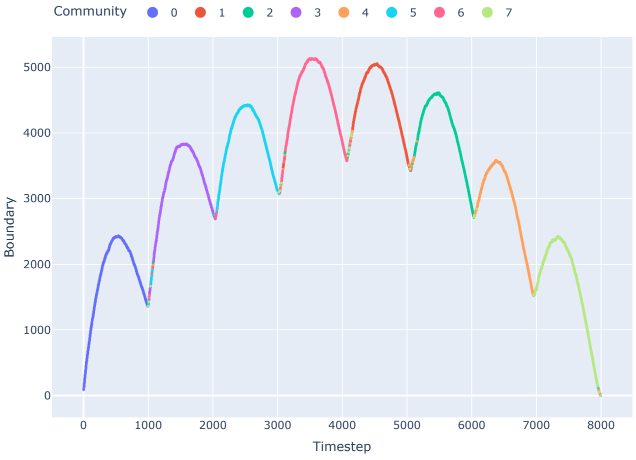
Evaluation: Our primary focus is directed towards the *boundary vs. timestep* plot to assess the synthetic networks we have constructed in light of our objective. Additionally, we make use of community size evolution plots to discern which community is being sampled at a given time. As an illustrative example of the outcomes we aim to achieve with our sampling scheme, consider Figure 1(a), which showcases the boundary's dynamic changes in one of the synthetic network configurations.

In Figure 1(a), notable inflection points are clearly visible around timesteps 1000, 2000, 3000, etc., along with the commencement of a corresponding steep increase in the size of one of the communities as depicted in Figure 1(b). Here, an inflection point refers to the timestep at which the sampler starts sampling a new community.

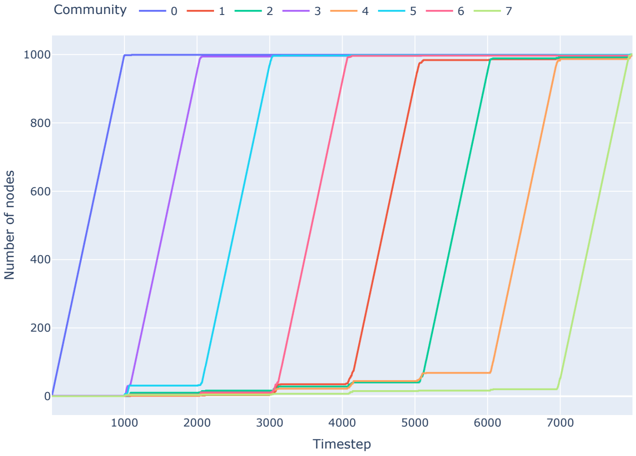
In stark contrast, when we investigate one of the random sampling methods, specifically Random Outsider (RO), applied to the identical SBM configuration, Figure 2(a) conspicuously lacks any discernible inflection points. Likewise, in Figure 2(b), as anticipated, we notice that the sizes of all communities are simultaneously increasing. This observation indicates that the sampled network does not exhibit a preference for obtaining one community at a time and does not account for community partitions. Similar behavior was observed with the other two random schemes, namely RI_RO and RI_MAS.

Having established our desired outcomes from the sampling scheme, we will now explore how its behavior varies across different configurations and assess the limits of detectability for inflection points by varying the values of r.

Selection and distribution of seeds. Throughout our experimentation, we observed that varying the choice of seed nodes had little to no discernible impact on sampling behavior. In other words, the sampling behavior remained largely consistent whether we opted for higher-degree, randomly selected, or lower-degree nodes. However, it is important to note that sampling does indeed depend on the distribution of seeds across blocks. This phenomenon can be attributed to the behavior of MAS, which tends to greedily favor the nodes with larger boundaries in an effort to minimize the boundary of the cluster.



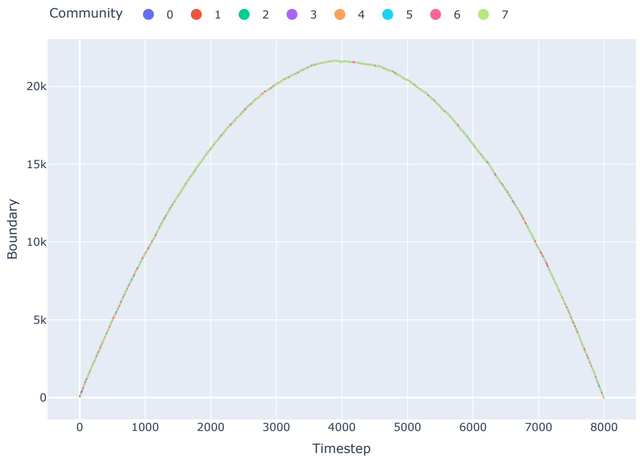
(a) Evolution of network boundary with time



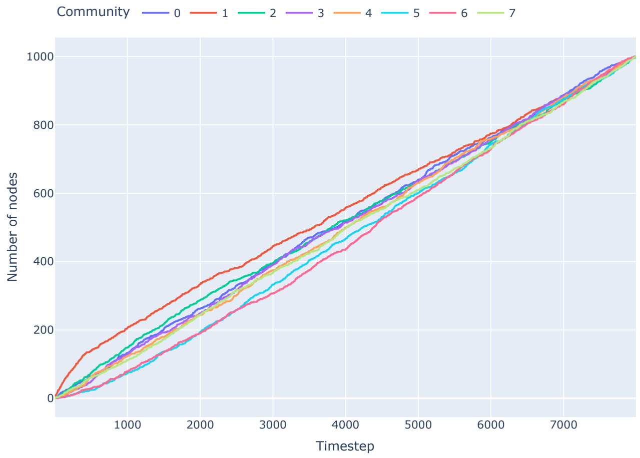
(b) Community size evolution

Figure 1: Sampling using MAS for a network of 8 blocks with 1000 nodes and one seed node each. The ratio of intra cluster to inter cluster edges (r) and average degree within the block $(\langle k' \rangle)$ are set as 4 and 10 respectively

Ratio of intra- to inter-cluster edges (r). In this study, we employ the ratio r as a metric to gauge the 'cohesiveness' of a community. A significant contrast in sampling behavior becomes apparent when $r \geq 2$, leading to the identification of inflection points signifying the transition from one community to another. Conversely, when $r = 1$, the inflection points on the boundary vs. timestep plot are not easily discernible, particularly in scenarios where seeds are distributed across multiple blocks of the SBM with differing block sizes. This phenomenon is exemplified in Figure 3(a) for the boundary vs. timestep plot of a network characterized by block sizes 800, 1200, 1600, 2000, $r = 1$, and 20 seed nodes per block.

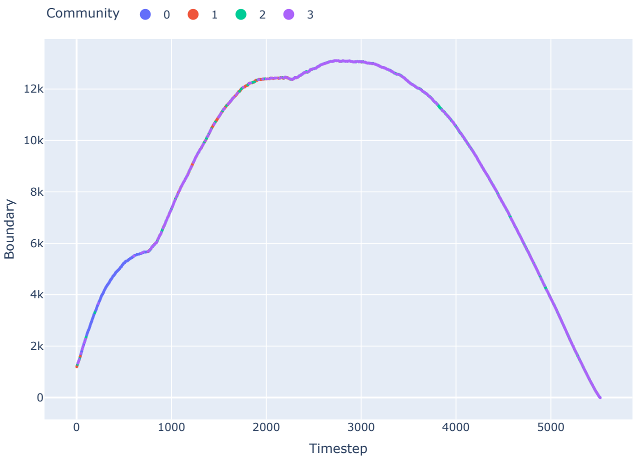


(a) Evolution of network boundary with time

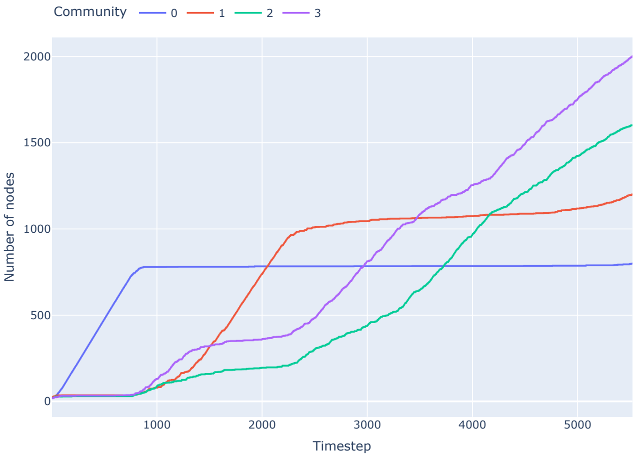


(b) Community size evolution

Figure 2: Sampling using RO (Random Outsider) for a network of 8 blocks with 1000 nodes and one seed node each. The ratio of intra cluster to inter cluster edges (r) and average degree within the block ($\langle k' \rangle$) are set as 4 and 10 respectively



(a) Evolution of network boundary with time



(b) Community size evolution

Figure 3: Sampling using MAS for a network of block sizes $\{800, 1200, 1600, 2000\}$ with 20 seed nodes per block. The ratio of intra cluster to inter cluster edges (r) and average degree within the block ($\langle k' \rangle$) are set as 1 and 10 respectively

As depicted in 3(b), following the sampling of community ‘0’, we observe that community ‘1’ attempts to be entirely sampled but becomes contaminated with nodes from communities ‘2’ and ‘3’. After approximately timestep 3200, no particular community exhibits a clear preference for complete sampling.

Nevertheless, in scenarios where communities are of equal size and seed nodes are uniformly distributed, we can still detect inflection points even when $r = 1$. Although less visible than those observed when $r \geq 2$, these inflection points remain detectable. For even lower values of r, the inflection points become less pronounced, and it becomes apparent that multiple communities are being sampled simultaneously.

Forecasting the exact sequencing of community sampling subsequently becomes complex, as it is influenced by a multitude of factors, including both intra-cluster edges and the inter-cluster edges between the community that has been sampled and those that remain unsampled, all of which impact the directed boundary of the sampled nodes.

Observations. Throughout our experimental investigations, we have discerned that the sampler's behavior is contingent upon the seeds' distribution and the intra- to inter-cluster ratio's value (r). However, when seeking to determine which community is being sampled at a given point in time, we have found that the plot depicting boundary vs. timestep tends to yield precise insights. We have observed that once a community is exhausted after sampling, a brief period of competition ensues among candidate communities, contending for the next sampling opportunity. The duration of this phase varies depending on the value of r. Specifically, for smaller values of r, this phase tends to be protracted, leading to the concurrent sampling of nodes from multiple communities. Conversely, this competition is relatively shorter for larger values of r. Eventually, the community with the highest boundary emerges as the winner, attracting a substantial influx of users who follow its initial lead. A higher value of r (such as $r \geq 2$) closely aligns with this ideal behavior since it results in a better-defined community structure, thereby reducing the likelihood of contention.

Empirical Data

As a case study, we expand a well-curated data set of topically relevant Twitter profiles by sampling additional profiles that form cohesive communities of engagement with them.

The selection of seed profiles is from the DISMISS dataset (Arya et al. 2022), comprising a cohort of $11,580$ highly networked individuals. Since we are looking for individuals engaging with information sources, we want our seed set to consist of influential profiles triggering engagement. DISMISS is seen as an ideal case in the context of the Indian political sphere. For our study we focus on a subset of these individuals as seed users, namely those with the 'category' label as 'civil society'. Here, 'category' indicates the 'primary industry' of the respective user and can have values like 'civil society', 'creative', etc.

To facilitate the study, we collected tweets posted by the seed users during July 2022 and use interactions received by these tweets to form a seed network as discussed in the previous section. During the process, we further filter out users to keep only those who posted at least one tweet in the said duration, resulting in $1{,}095$ users as the seed set, from the initial $1{,}184$ belonging to the 'civil society' category.

For the above $1{,}095$ seed users, we obtained $50{,}379$ tweets for the chosen duration. To ensure a balanced dataset and mitigate the potential influence of outlier tweets that may have garnered an exceptionally high number of interactions, we employed a ranking approach, focusing on the lower 90% of the tweet interactions. This curation process ultimately yielded a final set of $45{,}341$ tweets, which had received interactions from $379{,}514$ distinct Twitter users. Among the seed users, the number of authored tweets ranged from 1 to 534. This curated dataset forms the cornerstone for constructing an initial network, which subsequently serves as the foundational point for ongoing data collection efforts pertaining to Distinct, Nested, and A-F sampling schemes. The data collection was initiated in December 2022 under the presumption that interactions had reached a stable state by that time.

The interactions from these collected tweets were used to get weights per interaction type for all three schemes - distinct, nested, and audience-facing, as shown in Table 1.

Sampling: For the seed network generated above, along with the weights from Table 1, we sample the networks for distinct, nested, and audience-facing variants. Figure 4 shows the evolution of cohesive communities around the seed nodes using "distinct" sampling strategy. Along with the three variants, we also sample using four types of random sampling schemes for comparison.

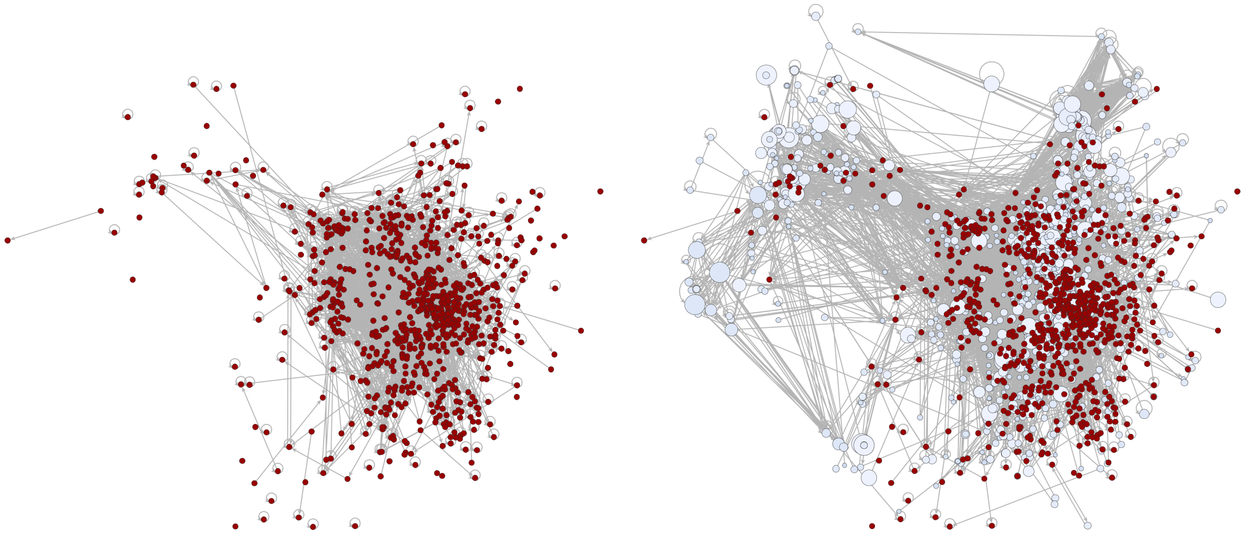
Our random node sampling strategy possesses two key attributes: selection probabilities and selection strategy. Selection probabilities can either be uniform (U), where all samples share an equal probability of being chosen, or weighted (W), where the probability is determined based on the priority/score computed using our sampling scheme. The selection strategy encompasses two options: 'direct' (D), in which one of the outsiders is chosen randomly with the specified selection probability, and 'staged' (S), which involves the selection of an insider at random, followed by the selection of one of the chosen insider's outsiders with the given selection probability. By combining these features, we derive four distinct random node sampling strategies, denoted as RS_DU, RS_DW, RS_SU, and RS_SW by considering all possible combinations. Table 2 provides an overview of the Twitter data sampled, using the three variants of our sampling scheme in conjunction with the random node-based sampling strategies. It is worth noting that due to the Twitter API shutdown, the sizes of the collected sampled networks differ, ranging from a minimum of 1,905 for RS_DW to a maximum of 5,515 for RS_SU.

Sampling Scheme	Insiders	Nodes	Edges in insider network	Total edges	Tweets
Distinct	8,721	609,609	208,628	1,545,420	161,471
Nested	4,698	525,531	98,889	1,182,774	91,966
A-F	3,919	513,466	93,476	1,149,281	84,267
RS_DU	1,976	417,439	5,438	745,871	50,856
RS_DW	1,905	410,061	8,383	744,067	51,191
RS_SU	5,515	600,858	28,536	1,070,803	74,463
RS_SW	3,355	527,265	34,127	1,023,682	62,872

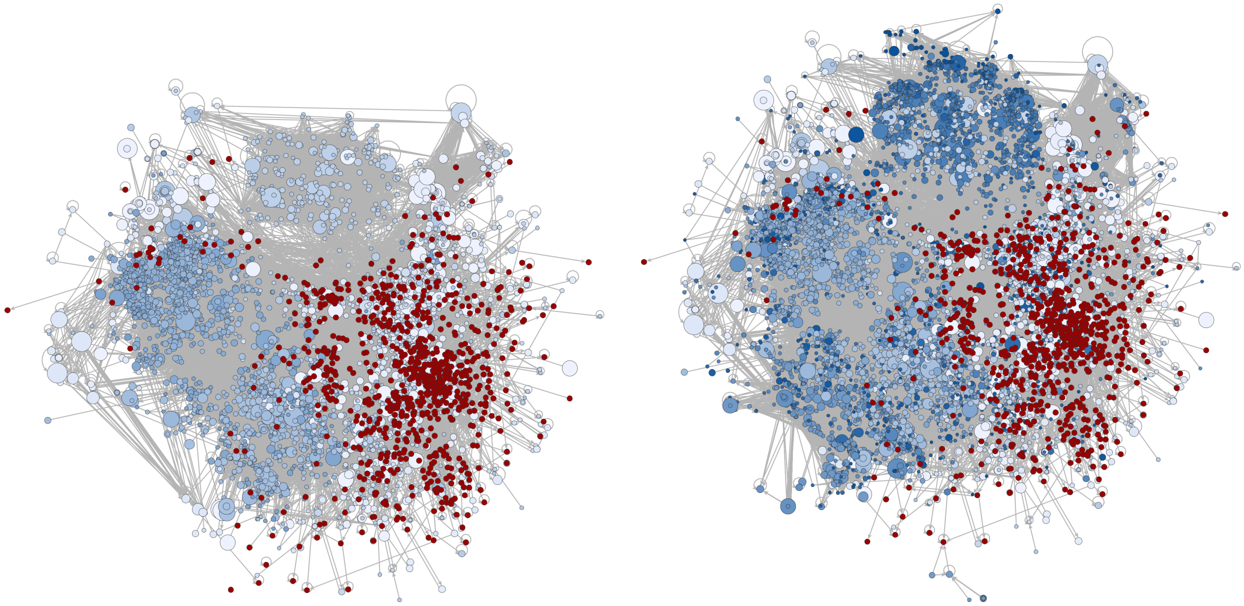
Table 2: Dataset statistics for the sampled Twitter network using the three variants of our sampling scheme and four variants of random sampling.

Evaluation: The fundamental distinction between the Twitter network we collected and the synthetic networks we generated pertains to the definition of a community. In the case of synthetic networks, a community was explicitly defined as a block utilized in configuring the Stochastic Block Model (SBM). In contrast, with the Twitter network, we lack a definitive "community label" and must rely on obtaining it without a guarantee of accuracy. In an effort to potentially assign community labels to each node, we apply the Louvain community detection algorithm to the collected Twitter network. Analogous to our approach with synthetic networks, we employ these community labels to explore potential correlations between the initiation of community sampling and the boundary vs. timestep plot of the entire network.

As observed in synthetic networks, we anticipate the presence of inflection points in the "boundary vs timestep" plots, indicating the transition from sampling one community to the next. In the context of the Twitter network, we notice a similar pattern, although occasional instances occur where two communities are sampled concurrently. For instance, in the case of sampling using the Audience-Facing (A-F) approach, exemplified in Figure 5(b), we discern distinct segments where only one community is sampled at any given time. However, there are intervals during which, alongside the primary community, certain nodes from a background



(a) Seed nodes (DISMISS data) (b) after 1000 timesteps

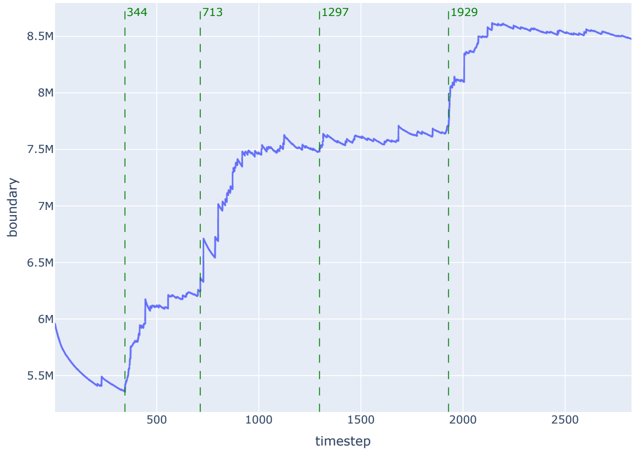


(c) After 4000 timesteps (d) After termination (8000 timesteps)

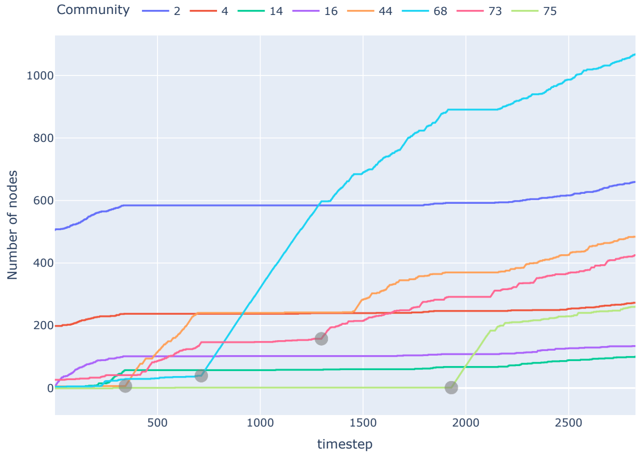
Figure 4: Growth of the Twitter network during the expansion of the DISMISS data by sampling nodes with maximum priority according to distinct interaction patterns. Each time step corresponds to the move of an outsider into the insider set. The layout is determined by the final network and constant across subfigures. Seed nodes are in red, and other nodes vary from light to dark blue based on the time of sampling. Non-seed node sizes reflect their priority at the time of sampling. While nodes added early are strongly interacting with the seed set, spillover into other cohesive communities is readily observed.

community are also included in the insider set. This occurrence is linked to scenarios where the priority of nodes is identical, signifying they possess similar weighted directed boundary values. A notable example of this behavior can be found in the time range between timesteps $t = 800$ and $t = 1200$, where we observe substantial growth in the community labeled as “68” (blue), while a few nodes are added to the community labeled as “73” (pink).

Despite the utilization of community labels generated from the data, we are still able to identify significant spikes in boundary values that correspond to the initiation of new communities. In Figure 5(a), we can observe these spikes corresponding to the commencement of communities “44,” “68,” and “75” at timesteps 344, 713, and 1929, respectively. Following this initiation, the boundary values stabilize briefly before witnessing another spike with the onset of a new community. It is worth highlighting that at timestep 1297, where although the expansion of community “73” is



(a) Community evolution of 8 largest communities using A-F



(b) Change in boundary using A-F

Figure 5: The four shaded circles in (b) indicate points where significant sampling of a new community starts. For corresponding timesteps in (b), we observe that the boundary shoots up before plateauing again. This is especially noticeable for timesteps = $\{344, 713, 1929\}$. For a smaller increase in community size, as seen for timestep = 1929, we still also observe a small rise in boundary value in (b). However, not every small rise in (a) corresponds to a different community being sampled in (b). Hence, we tend to focus on bigger jumps in the value of boundary timestep = $\{344, 713, 1929\}$ where a different community begins sampling, as can be seen through the steep rise in community size.

relatively modest, the boundary plot effectively captures it with a minor peak. In a real-world scenario, however, users would encounter a boundary-versus-timestep plot and need to discern points such as the one at timestep 1297, which might be easily overlooked amidst the noise. Nevertheless, the approach remains capable of capturing rising trends akin to those observed at timesteps $\{344, 713, 1929\}$.

Structure based metrics To gain insight into the characteristics of the cohesive communities obtained through sampling, we conduct a comprehensive evaluation using informative structural metrics, as outlined below:

1. *Average shortest path ($\langle L \rangle$):* The average directed path length along the shortest paths for all possible pairs of nodes.
2. *Clustering Coefficient:*
   (a) *Local clustering coefficient (CC_{local})*: The local clustering coefficient for a node i on a directed network is given by

$$CC_i = \frac{|e_{jk} : j, k \in N(i),\ e_{jk} \in E|}{deg(i)(deg(i) - 1)}$$

   where E denotes the set of edges in the graph and $N(i)$ denotes the open neighborhood of node i. The average local clustering coefficient, CC_{local}, is the mean of the local clustering coefficients of all nodes.

   (b) *Global clustering coefficient (CC_{global})*: This is given by the ratio of the number of closed triplets over all possible triplets in the network.
3. *Average degree ($\langle k \rangle$):* The average degree is the mean of all node degrees.

As the sizes of the sampled networks obtained through the three variants and four random schemes vary, we restrict our analysis to the subgraphs sampled up to the size of the smallest common network. This approach allows us to calculate metrics on networks of equal size, ensuring the comparability of results across different schemes while eliminating the influence of network size disparities. Using this approach, we have one common timestep for the priority-based sampling schemes and one for random sampling schemes.

Table 3 shows the above metrics for networks obtained using the seven sampling schemes.

Sampling scheme		CC_{local}	CC_{global}	$\langle L \rangle$	$\langle k \rangle$
Priority	Distinct	0.2566	**0.4239**	5.34	12.97
	Nested	0.3747	0.4145	4.62	21.65
	A-F	**0.4004**	0.4035	4.40	**26.49**
Random	RS_DU	0.0646	0.0698	5.25	3.40
	RS_DW	0.1360	0.0608	4.87	5.32
	RS_SU	0.1179	0.0559	4.95	4.81
	RS_SW	0.1237	0.0562	**4.33**	9.11

Table 3: Comparison of structure based metrics across priority based and random sampling schemes. The bold values signify the highest or lowest values as per the chosen metric.

As presented in Table 3, we observe that the clustering values (CC_{local} and CC_{global}) for the proposed priority-based sampling schemes are notably higher compared to any of the random sampling schemes. This disparity in values suggests that networks obtained through priority-based schemes exhibit stronger connectivity. Additionally, Table 3 reveals that the Audience Facing interactions variant outperforms all other variants in terms of CC_{local} metric, indicating a higher number of triads, with the exception of CC_{global} where the distinct variant maintains a slight advantage. It is also important to highlight that all the variants have a significant performance advantage over any random sampling schemes.

Conclusion and Future Work

We proposed a novel scheme for snowball-type sampling in unbounded networks designed to respect cohesive communities. Its intended purpose is the extraction of communities.

Our approach consists of two main parts, a sampling priority utilizing the maximum-adjacency principle, and a method to integrate modes of interaction into a single weighted directed graph. The latter is based on importance scaling and can be calibrated empirically as demonstrated in a prototypical case study. Computational experiments on synthetic and empirical data demonstrate that our method samples subgraphs with low inwards-directed conductance by keeping the boundary around the sampled region small. While the growth inside communities is almost perfect in the idealized setting of stochastic blockmodels, a similar evolution is observed in the case study on Twitter that motivated this research.

With an adapted construction of the weighted graph, our sampling strategy transfers to other social media networks, such as Reddit or Facebook, and it will be interesting to apply the maximum-adjacency principle in other settings such as the respondent-driven sampling of offline social networks (Heckathorn 1997).

Ethics Statement

In this work, we use Twitter API to gather a stream of tweets and their interactions. We leverage only publicly available information at the time of data collection. To comply with the terms of service, we will release our anonymized data (network structure) for research purposes only. To the best of our knowledge, no code of ethics has been violated throughout the data collection and experimental phase in this study.

Acknowledgments

This work was supported by the Swiss National Science Foundation (SNSF Grant number 209488) and Indian Council of Social Science Research (ICCR Grant number ICSSR-SNSF (SWISS)/JRP-7/2022-IC).

References

Ahmed, N. K.; Neville, J.; and Kompella, R. 2013. Network sampling: From static to streaming graphs. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 8(2): 1–56.

Andersen, R.; Chung, F.; and Lang, K. 2006. Local graph partitioning using pagerank vectors. In *2006 47th Annual IEEE Symposium on Foundations of Computer Science (FOCS'06)*, 475–486. IEEE.

Antonakaki, D.; Fragopoulou, P.; and Ioannidis, S. 2021. A survey of Twitter research: Data model, graph structure, sentiment analysis and attacks. *Expert Systems with Applications*, 164: 114006.

Arya, A.; De, S.; Mishra, D.; Shekhawat, G.; Sharma, A.; Panda, A.; Lalani, F.; Singh, P.; Mothilal, R. K.; Grover, R.; Nishal, S.; Dash, S.; Shora, S.; Akbar, S. Z.; and Pal, J. 2022. DISMISS: Database of Indian Social Media Influencers on Twitter. *Proceedings of the International AAAI Conference on Web and Social Media*, 16: 1201–1207.

Avrachenkov, K.; Ribeiro, B.; and Towsley, D. 2010. Improving random walk estimation accuracy with uniform restarts. In *Algorithms and Models for the Web-Graph: 7th International Workshop, WAW 2010, Stanford, CA, USA, December 13-14, 2010. Proceedings 7*, 98–109. Springer.

Cai, W.; and Matula, D. W. 1993. Partitioning by Maximum Adjacency Search of Graphs. *Partitioning Data Sets*, 19.

Chang, C.-H.; Chang, C.-S.; Chang, C.-T.; Lee, D.-S.; and Lu, P.-E. 2019. Exponentially Twisted Sampling for Centrality Analysis and Community Detection in Attributed Networks. *IEEE Transactions on Network Science and Engineering*, 6(4): 684–697.

De Domenico, M.; Lima, A.; Mougel, P.; and Musolesi, M. 2013. The anatomy of a scientific rumor. *Scientific reports*, 3(1): 1–9.

Giudice, P. L.; and Ursino, D. 2019. Algorithms for Graph and Network Analysis: Traversing/Searching/Sampling Graphs. In *Encyclopedia of Bioinformatics and Computational Biology*.

Gjoka, M.; Kurant, M.; Butts, C. T.; and Markopoulou, A. 2010. Walking in facebook: A case study of unbiased sampling of osns. In *2010 Proceedings IEEE Infocom*, 1–9. Ieee.

Heckathorn, D. D. 1997. Respondent-Driven Sampling: A New Approach to the Study of Hidden Populations. *Social Problems*, 44(2): 174–199.

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

Hu, P.; and Lau, W. C. 2013. A survey and taxonomy of graph sampling. *arXiv preprint arXiv:1308.5865*.

Hübler, C.; Kriegel, H.-P.; Borgwardt, K.; and Ghahramani, Z. 2008. Metropolis algorithms for representative subgraph sampling. In *2008 eighth ieee international conference on data mining*, 283–292. IEEE.

Kim, H.; Jang, S. M.; Kim, S.-H.; and Wan, A. 2018. Evaluating Sampling Methods for Content Analysis of Twitter Data. *Social Media + Society*, 4(2): 2056305118772836.

Kurant, M.; Markopoulou, A.; and Thiran, P. 2011. Towards unbiased BFS sampling. *IEEE Journal on Selected Areas in Communications*, 29(9): 1799–1809.

Lee, S. H.; Kim, P.-J.; and Jeong, H. 2006. Statistical properties of sampled networks. *Phys. Rev. E*, 73: 016102.

Leskovec, J.; and Faloutsos, C. 2006. Sampling from large graphs. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, 631–636.

Li, R.-H.; Yu, J. X.; Qin, L.; Mao, R.; and Jin, T. 2015. On random walk based graph sampling. In *2015 IEEE 31st international conference on data engineering*, 927–938. IEEE.

Liu, L.; Wang, L.; Wu, W.; Jia, H.; and Zhang, Y. 2019. A novel hybrid-jump-based sampling method for complex social networks. *IEEE Transactions on Computational Social Systems*, 6(2): 241–249.

Maiya, A. S.; and Berger-Wolf, T. Y. 2010. Sampling community structure. In *Proceedings of the 19th international conference on World wide web*, 701–710.

Newman, M. E. 2006. Modularity and community structure in networks. *Proceedings of the national academy of sciences*, 103(23): 8577–8582.

Ribeiro, B.; and Towsley, D. 2010. Estimating and sampling graphs with multidimensional random walks. In *Proceedings of the 10th ACM SIGCOMM conference on Internet measurement*, 390–403.

Rozemberczki, B.; Kiss, O.; and Sarkar, R. 2020. Karate Club: an API oriented open-source python framework for unsupervised learning on graphs. In *Proceedings of the 29th ACM international conference on information & knowledge management*, 3125–3132.

Ruths, D.; and Pfeffer, J. 2014. Social media for large studies of behavior. *Science*, 346(6213): 1063–1064.

Spielman, D. A.; and Teng, S.-H. 2004. Nearly-linear time algorithms for graph partitioning, graph sparsification, and solving linear systems. In *Proceedings of the thirty-sixth annual ACM symposium on Theory of computing*, 81–90.

Stoer, M.; and Wagner, F. 1997. A simple min-cut algorithm. *Journal of the ACM (JACM)*, 44(4): 585–591.

Stutzbach, D.; Rejaie, R.; Duffield, N.; Sen, S.; and Willinger, W. 2006. On unbiased sampling for unstructured peer-to-peer networks. In *Proceedings of the 6th ACM SIGCOMM conference on Internet measurement*, 27–40.

Ye, S.; Lang, J.; and Wu, F. 2010. Crawling online social graphs. In *2010 12th International Asia-Pacific Web Conference*, 236–242. IEEE.

Zhang, J.; Chen, H.; Yu, D.; Pei, Y.; and Deng, Y. 2023. Cluster-preserving sampling algorithm for large-scale graphs. *Science China Information Sciences*, 66(1): 112103.

Zhang, Y.; and Rohe, K. 2018. Understanding regularized spectral clustering via graph conductance. *Advances in Neural Information Processing Systems*, 31.

Ethics Checklist

1. For most authors...
    - (a) Would answering this research question advance science without violating social contracts, such as violating privacy norms, perpetuating unfair profiling, exacerbating the socio-economic divide, or implying disrespect to societies or cultures? Yes
    - (b) Do your main claims in the abstract and introduction accurately reflect the paper's contributions and scope? Yes
    - (c) Do you clarify how the proposed methodological approach is appropriate for the claims made? Yes
    - (d) Do you clarify what are possible artifacts in the data used, given population-specific distributions? NA
    - (e) Did you describe the limitations of your work? Yes, they have been elaborated for different parameters in *Experimentation* section.
    - (f) Did you discuss any potential negative societal impacts of your work? No, we are proposing a new sampling scheme. We do not believe there is any negative societal impacts of our work.
    - (g) Did you discuss any potential misuse of your work? No, we are proposing a new sampling scheme. We do not believe there is a measurable chance of the scheme being misused.
    - (h) Did you describe steps taken to prevent or mitigate potential negative outcomes of the research, such as data and model documentation, data anonymization, responsible release, access control, and the reproducibility of findings? NA
    - (i) Have you read the ethics review guidelines and ensured that your paper conforms to them? Yes, we have read ethics review guidelines and our manuscript did not violate any of them.
2. Additionally, if your study involves hypotheses testing...
    - (a) Did you clearly state the assumptions underlying all theoretical results? Yes
    - (b) Have you provided justifications for all theoretical results? Yes
    - (c) Did you discuss competing hypotheses or theories that might challenge or complement your theoretical results? NA
    - (d) Have you considered alternative mechanisms or explanations that might account for the same outcomes observed in your study? Yes
    - (e) Did you address potential biases or limitations in your theoretical framework? Yes, we have elaborated it in *Experiments* section
    - (f) Have you related your theoretical results to the existing literature in social science? NA
    - (g) Did you discuss the implications of your theoretical results for policy, practice, or further research in the social science domain? No
3. Additionally, if you are including theoretical proofs...
    - (a) Did you state the full set of assumptions of all theoretical results? NA
    - (b) Did you include complete proofs of all theoretical results? NA
4. Additionally, if you ran machine learning experiments...
    - (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? NA
    - (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? NA
    - (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? NA

(d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? NA

(e) Do you justify how the proposed evaluation is sufficient and appropriate to the claims made? NA

(f) Do you discuss what is “the cost“ of misclassification and fault (in)tolerance? NA

5. Additionally, if you are using existing assets (e.g., code, data, models) or curating/releasing new assets...

(a) If your work uses existing assets, did you cite the creators? Yes

(b) Did you mention the license of the assets? NA

(c) Did you include any new assets in the supplemental material or as a URL? NA

(d) Did you discuss whether and how consent was obtained from people whose data you’re using/curating? Yes, we limit ourselves to using Twitter API/publicly available data.

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? Yes, our dataset does not contain any personally identifiable information. We leverage only publicly available information at the time of data collection.

(f) If you are curating or releasing new datasets, did you discuss how you intend to make your datasets FAIR? Yes, our dataset (upon release) will adhere to FAIR principles: Findable, Accessible, Interoperable and Reusable. The complete dataset would be made publicly available by ensuring anonymization. All data stored in format parsed and viewed easily. In addition to this, making it Interoperable and Reusable.

(g) If you are curating or releasing new datasets, did you create a Datasheet for the Dataset? No

6. Additionally, if you used crowdsourcing or conducted research with human subjects...

(a) Did you include the full text of instructions given to participants and screenshots? NA, since our study does not include human subjects.

(b) Did you describe any potential participant risks, with mentions of Institutional Review Board (IRB) approvals? NA

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? NA

(d) Did you discuss how data is stored, shared, and deidentified? NA