Sampling cohesive communities in unbounded networks

Thesis submitted in partial fulfillment
of the requirements for the degree of

Master of Science
in
Computing and Human Sciences by Research

by

Kshitijaa Jaglan
2019115005
kshitijaa.jaglan@research.iiit.ac.in

International Institute of Information Technology, Hyderabad
(Deemed to be University)
Hyderabad - 500 032, INDIA
June 2024

International Institute of Information Technology
Hyderabad, India

CERTIFICATE

It is certified that the work contained in this thesis, titled 'Sampling cohesive communities in unbounded networks' by Kshitijaa Jaglan, has been carried out under my supervision and is not submitted elsewhere for a degree.

_____ _____

Date Adviser: Prof. Sushmita Banerji

_____ _____

Date Co-adviser: Prof. Ponnurangam Kumaraguru

In loving memory of a profound influence, forever in my heart.

To my parents, brother and friends,

Acknowledgments

I would like to begin by expressing my deepest gratitude to Prof. PK for his invaluable guidance, encouragement and support throughout my research journey. Since joining the lab in my fourth year, he has been a constant source of encouragement, positivity, and provided constructive feedback whenever I felt lost. I would like to thank him for being the pillar of strength whenever challenges emerged, helping me grow personally and professionally. His ability to connect with people is remarkable, allowing his students, including myself to gain exposure both locally and internationally.

I am also very thankful to Prof. Sushmita Banerji for her help and guidance throughout this wonderful journey. I would also like to show my gratitude to Prof. Don, who first introduced me to the world of research.

A special acknowledgement goes to Prof. Ulrik Brandes, whose unwavering support and passion for networks ignited the spark in me to pursue it further, and significantly shaped my research direction and academic aspirations. I am very thankful to him for always taking out time for discussions and motivating me to participate in different activities to make sure that my semester at ETH Zürich is a great and enjoyable learning experience. Even though I was at Social Networks Lab @ ETH Zürich for only a semester, my time there has helped me in gaining a lot of clarity about current and future directions in life. My visit wouldn't have been the same without Denise, who was always there to help in handling all the administrative tasks even before I arrived there. Her energy and enthusiasm is infectious and I am grateful for all her help.

I would also like to extend my gratitude to the Indian Council for Cultural Relations (ICCR) and the Swiss National Science Foundation (SNSF) for their generous financial support.

I am also grateful to my co-authors - Meher, Abhi, Triansh and Nidhi for their collaboration, contributions and going with me on this wonderful adventure. Meher has been like a big brother to me, especially during my semester at Zurich, helping me navigate my day-to-day life in a foreign country. Thank you Abhijeeth for always being a constant companion in exploring and debating research ideas, along with getting into petty discussions for fun and making delicious cookies for us all.

To Rishabh, thank you for being my go-to person throughout this college journey and listening to my rants, going to hospital for vaccination after cat bites, finding wonderful movies, listening to my elevator music and giving constant reminders to stay hydrated. You have been a beautiful big part of this wonderful yet challenging journey, and I am very grateful to have happened to meet you and get to know you so well.

To Harshita, thank you for being more than just a roommate and a close friend, but a sister with whom I could share countless bowls of maggi and life anecdotes. Despite the distance now, I know you are just a phone call away, and your unexpected conversations always light up my day.

To Simba, my cute yet silly cat, hearing your meows and mrrps as you came prancing through the corridor were enough to brighten up the day. I still miss you a lot and wish I was there with you during your last days. Thank you for showing me all the love and trust, I hope you're happy and getting loads of treats across the rainbow bridge.

Thank you Tushar, Ainesh, Shivansh, Kunwar, Aadu, Ashu and Pahul for always being ready to go on chaotic adventures, and being the bunch I can always rely on. A special thanks goes out to my lab mates at PreCog - Karuna, Arvindh and Vamshi, who motivated me to show up at lab regularly. Thank you to Keshav, Eshan, Anjali, Mihir, Kalp, Jai, the entire CHD family and others for enriching my college experience. Thank you Shreevignesh for letting me use your laptop to try to write this thesis. A big thanks to Ivana, Matt and Gordana for making Zurich more vibrant than it already is. Thank you Ishika for the beautiful bond we have since so many years, and I look forward to meeting you soon.

A heartfelt thank you to my family, whose regular check-ins were a reminder of their unwavering love, care and support. Their constant guidance towards independence, support to pursue my academic endeavours, and concern for my well-being have been a source of strength and comfort throughout my life. I am truly grateful to have been born to such lovely parents, and a truly warm-hearted brother. Since my mummy always wanted her name on a thesis, here we go - I am the daughter of Suman Jaglan and Subhash Jaglan, and sister of Antryksh Jaglan, and am truly thankful for all they have done in the last 22 years of my existence on this beautiful planet Earth.

Thank you dear reader, I hope you gain something worthwhile out of this thesis - a collective effort of many people - the way I did too!

Abstract

With today's social networks measuring up to millions and billions of nodes and edges, it becomes essential to devise methodologies to obtain a subgraph with the required properties. Network sampling is often one of the most important stages of obtaining and studying a network since the properties of a sampling scheme used directly influence the properties of the network under consideration. Driven by insights from studies on homophily and opinion formation, we introduce a variant of snowball sampling specifically tailored to prioritize the inclusion of entire cohesive communities. This approach deliberately avoids traditional aims like representativeness, breadth, or depth of coverage, which have been the focus of extensive research in the past. We propose the sampling scheme in the context of Twitter - a conceptually unbounded network, that can be transferred to other types of social networks with different types of interactions. The study is undertaken in two stages - we combine multiplex forms of interactions observed between users to construct a simple network, followed by using a variation of snowball sampling to iteratively sample nodes based on a priority determined by their connectivity with the currently sampled network. The efficacy and limitations of this approach are demonstrated through empirical analysis on synthetic networks, which are unweighted and undirected, generated via the Stochastic Block Model. Moreover, we utilize variants of our proposed sampling technique to gather dataset(s) from Twitter. The experiments in both real and synthetic networks suggest that the scheme behaves as desired.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

> Graphs stand or fall by their choice of nodes and edges.
>
> Watts & Strogatz

1.1 Motivation

The study of networks spans multiple domains - from trade networks [Furusawa and Konishi, 2007] and citation networks [Wallace et al., 2012] to even biological networks [Pavlopoulos et al.,] spanning millions of nodes and even more edges. The idea of social networks particularly has picked up pace in recent years, especially during COVID-19 [Prez-Escoda et al., 2020], when platforms like Instagram, Twitter, TikTok etc. became a popular window to the external world, with some studies even showing the correlation between structure of online social networks and geographic spread of COVID-19 [Kuchler et al.,].

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 et al., 2021]. However, obtaining full-size networks from platforms is practically impossible for researchers due to access limitations, and prohibitive due to 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].

Common sampling schemes rely on one or both of two main techniques [Ahmed et al., 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.

1.2 Proposed sampling scheme

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 using methods such as approximate PageRank [Andersen et al., 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.

1.3 Key contributions

Our main contributions are as follows:

1. Application of a maximum-adjacency principle to snowball-sampling to expand seed sets while staying within local communities.

2. Generalization of the maximum-adjacency criterion to weighted multiplex networks. Specifically for Twitter, we propose an empirically calibrated weighting scheme to combine types of interaction.

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

1.4 Thesis organisation

The thesis has been divided into seven chapters, where Chapter 2 discusses the background of the problem in details and outlines the related work. Chapter 3 formally introduces the problem and notation, followed by Chapter 4 where we discuss the proposed methodology for both construction of the graph, and sampling itself. Chapter 5 and 6 discuss experiments on synthetic and empirical Twitter data respectively, including data preparation and evaluation. We share the current state of work as a follow up to the research done for this thesis in Chapter 7. We then conclude the thesis and discuss limitations and future work in Chapter 8.

Chapter 2

Past Work

This chapter presents an overview of various sampling methods used for studying social media networks, primarily focusing on how these approaches begin by sampling a set of profiles, links, or interactions and then expand the network by exploring neighboring elements. This process is fundamental in understanding and analyzing the structure and dynamics of social media networks.

2.1 Node-based Sampling Methods

Node-based sampling methods range from straightforward techniques like uniform random sampling to more nuanced methods that consider node properties and network structure. For instance, *uniform random node sampling* [Ahmed et al., 2013] involves selecting nodes randomly without preference, providing a straightforward method to approximate direct node properties such as degree distribution where nodes with either high-degree or low-degree nodes can be given preference. However, this method may not preserve the connectivity of the underlying graph, possibly leading to isolated nodes or multiple connected components in the sampled data.

Stratified Sampling [Chaudhuri et al., 2005] takes a more structured approach by categorizing nodes into different groups or strata, often based on attributes like node degree, centrality, or other relevant characteristics, followed by sampling a specified number or fraction of nodes from each stratum. This method ensures a more representative sample across different node types, enhancing the quality and diversity of the sampled network. More sophisticated node-based sampling approaches, such as the *PageRank* and *PageRank-with-restarts* based sampling methods, choose sampled nodes based on their PageRank scores and construct the induced subgraph on them [Rozemberczki et al., 2020].

2.2 Edge-based Sampling Methods

Similar to node-based sampling techniques, edge-based sampling methods also exist where an edge is selected randomly or based on specific attributes that reflect the network's structure. In contrast to node selection, choosing an edge automatically includes the nodes at both of its endpoints in the sampled

graph. Typically, this results in a partially induced subgraph formed from these sampled edges without adding any additional edges from the original network. *Uniform random edge sampling*, analogous to its node-based counterpart, involves selecting edges at random from the graph. [Ribeiro and Towsley, 2010] demonstrate that random edge sampling is more effective than random node sampling for estimating characteristics like the tail of the out-degree distribution and other properties. A variant of uniform random edge sampling method is *Total Induction Edge Sampling (TIES)* [Ahmed et al.,], which creates a subgraph induced on nodes at the endpoints of at least one sampled edge. This approach helps mitigate the bias toward nodes with higher degrees often seen in uniform random edge sampling.

2.3 Traversal-Based Sampling Methods

Traversal-based sampling starts with one or more seed nodes and expands by exploring the neighbourhood of the already sampled nodes. Traditional traversal methods like *Breadth-First Search* (BFS) and *Depth-First Search* (DFS) [Giudice and Ursino, 2019] are commonly used, with BFS tending to cover the network more broadly by choosing the earliest (breadth-first) discovered node, and DFS exploring it more deeply by choosing the latest (depth-first) node. Generally, BFS can result in a denser cover and has been shown to be biased towards high-degree nodes [Ye et al., 2010]. The work by [Kurant et al., 2011] addressed this bias by suggesting analytical solutions to correct the said bias.

Snowball Sampling is another form of traversal-based sampling strategy that aims to maintain network connectivity using the breadth-first approach but can suffer from boundary bias, that is, nodes sampled in the later iterations have many missing neighbours [Lee et al., 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 et al., 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] incorporate 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.

Chapter 3

Tight sampling of cohesive communities[1]

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. We can visualise a sample bounded case for better understanding in the figure 3.1.

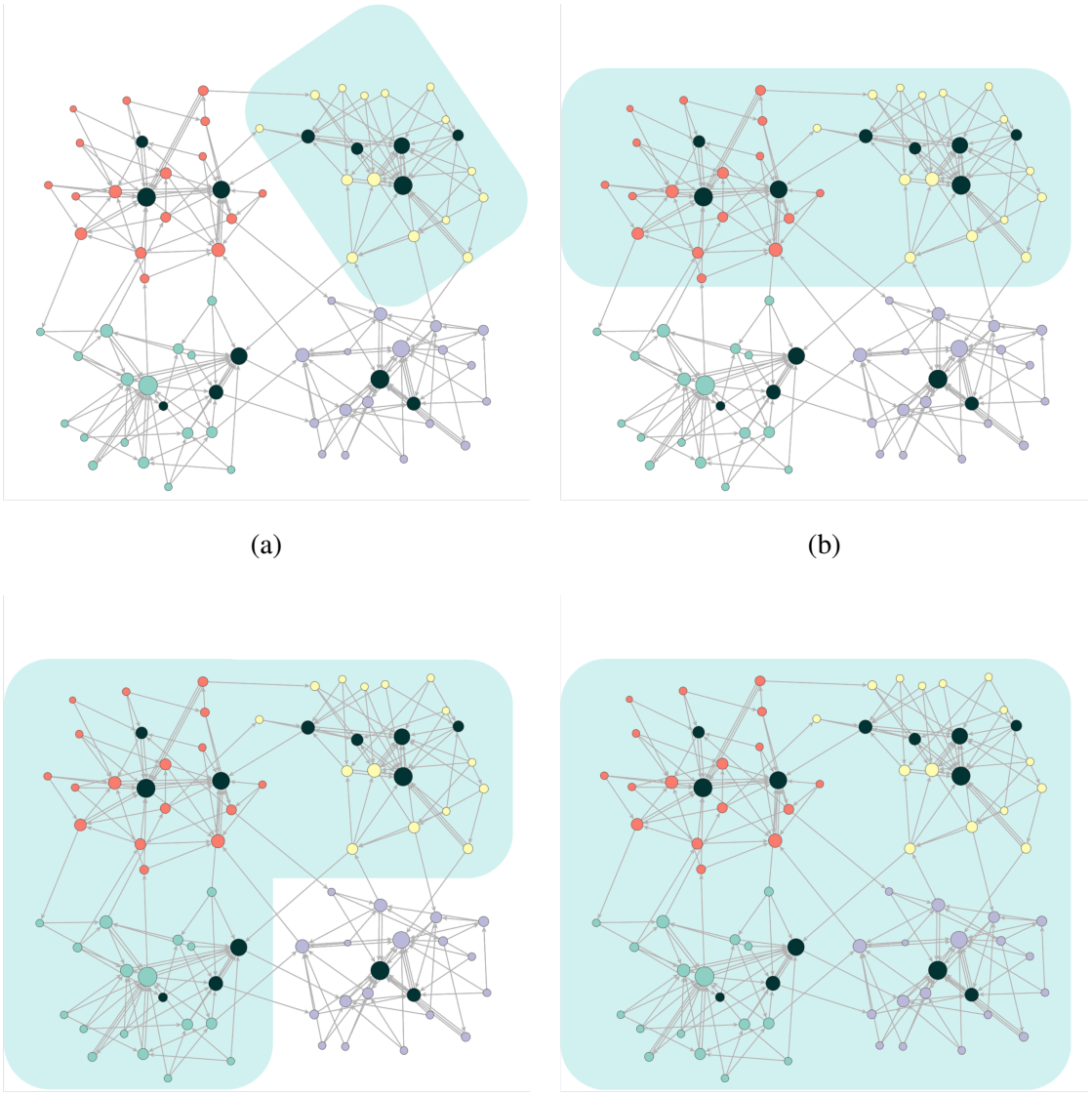
Formally, we assume the existence of an infinite, initially unknown directed graph $G = (V, E)$ representing a vast social media network. Edges represent social relations between members of the network and will be described more concretely below, where we also introduce edge weights. We further assume 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 symmetrically.

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 backwards. 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 maximum-adjacency search, and greedily removes edges from the boundary. Note that we do not know the in-neigborhood of a vertex prior to its sampling, so that we can not make any guarantees whether the new boundary is smallest possible.

[1]The next 4 chapters are a part of *Kshitijaa Jaglan*, Meher Chaitanya*, Triansh Sharma, Abhijeeth Singam, Nidhi Goyal, Ponnurangam Kumaraguru, Ulrik Brandes.* ***Tight Sampling in Unbounded Networks.*** *Forthcoming in ICWSM 2024 - The 18th International AAAI Conference on Web and Social Media, June 3 - June 6, 2024, Buffalo, New York, USA*



(a) (b)

(c) (d)

Figure 3.1: Demonstration of tight sampling on a dummy bounded network where black nodes represent seed nodes, and red, yellow, green and purple represent one community each. The shaded blue region represents the nodes sampled at a given time, where the timesteps are in the order a $<$ b $<$ c $<$ d.

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.

Chapter 4

Network construction and proposed sampling scheme

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.

4.1 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 is 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.

4.2 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.

4.2.1 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*.

4.2.2 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 including 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 or does not *reply* or *quote*.

4.2.3 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 interchangable 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].

4.3 Importance scaling

Interaction types occur at different rates and therefore 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.

4.3.1 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

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

4.3.2 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 $\overleftarrow{N}(i) = \sum\limits_{x \in X} n(i, x)$.

4.3.3 Target normalization.

Symmetrically, users interacting with information published by others may exhibit very different patterns of engagement. An alternative approach is therefore 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 $n(x, j) = \sum\limits_{i \in S} n(i, x, t)$ and $\overrightarrow{N}(j) = \sum\limits_{x \in X} n(x, j)$.

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 4.1 shows the calculated weights for a Twitter dataset collected using the proposed sampling scheme with distinct, nested and audience-facing interaction patterns. More information about the users has been provided in Chapter 6.

	$\eta(x)$		$\overleftarrow{\eta}(x)$		$\overrightarrow{\eta}(x)$		$\eta^*(x)$		$\omega(x)$		$\omega^*(x)$	
Interaction type	Distinct	Nested	Distinct	Nested	Distinct	Nested	Distinct	Nested	Distinct	Nested	Distinct	Nested
0001	2.2560	2.8947	2.0575	2.9979	1.4276	2.1090	1.9137	2.6672	0.5225	0.3749	*0.52*	*0.37*
0010	7.9125	9.4169	5.5468	7.9285	5.9666	7.9333	6.4753	8.4263	0.1544	0.1186	0.15	0.12
0011	0.3272	0.0583	0.0645	0.1020	0.0367	0.0793	0.0047	0.0799	22.3920	12.5190	22.4	12.52
0100	6.0684	15.9760	6.8172	18.4080	6.2281	18.4700	6.3712	17.6180	0.1569	0.0568	0.16	0.06
0101	0.0687	0.2199	0.1092	0.3414	0.0864	0.2801	0.0881	0.2805	11.3490	3.5652	11.35	3.6
0110	0.0860	0.3692	0.1641	0.5828	0.1015	0.5901	0.1172	0.5141	8.5331	1.9452	8.53	1.95
0111	0.0018	0.0102	0.0034	0.0203	0.0031	0.0238	0.0028	0.0181	360.1600	55.1890	**360.1**	55.2
1000	72.3500	83.5740	71.6710	85.2370	72.4130	86.1500	72.1440	84.9870	0.0139	0.0117	0.014	0.01
1001	0.3707	0.5355	0.5174	0.7633	0.3457	0.5551	0.4113	0.6180	2.4314	1.6182	2.43	1.6
1010	1.0871	1.3840	1.7172	2.1497	1.3212	1.8254	1.3752	1.7864	0.7272	0.5598	0.73	0.6
1011	0.0154	0.0238	0.0172	0.0341	0.0188	0.0395	0.0171	0.0324	58.4740	30.8210	58.5	30.8
1100	9.3286	9.7510	10.6870	11.3150	11.3950	12.0510	10.4700	11.0390	0.0955	0.0906	0.095	0.09
1101	0.1410	0.1495	0.2118	0.2287	0.1699	0.1906	0.1743	0.1896	5.7386	5.2742	5.74	5.3
1110	0.2730	0.2815	0.3984	0.4153	0.4648	0.4855	0.3787	0.3941	2.6402	2.5375	2.64	2.5
1111	0.0084	0.0084	0.0169	0.0169	0.0207	0.0207	0.0153	0.0153	65.1750	65.1750	65.2	**65.2**

(a) Weights per interaction pattern involving *like, retweet, reply, quote* for distinct and nested interactions.

	$\eta(x)$	$\overleftarrow{\eta}(x)$	$\overrightarrow{\eta}(x)$	$\eta^*(x)$	$\omega(x)$	$\omega^*(x)$
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

(b) Weights per interaction pattern involving *like, retweet, reply, quote* for audience-facing interactions.

Table 4.1: Weights per interaction pattern involving distinct, nested, and audience-facing interactions.

Chapter 5

Experiments - Synthetic data

We evaluate our sampling strategy by first creating synthetic networks in a controlled setting using stochastic blockmodels (SBM) [Holland et al., 1983] - a type of probabilistic model used to generate networks where the nodes are partitioned into subgroups called blocks or clusters. This approach allows us to generate networks with predefined communities and monitor their sampling. 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 results in improved coverage of cohesive communities as compared to random-based sampling approaches.

5.1 Data preparation

In this section, we 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.

5.1.1 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 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:

$$\rho_{ij} = \frac{\rho_{i,*} + \rho_{j,*}}{2} \tag{5.1}$$

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.

5.1.2 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 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.

5.2 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.

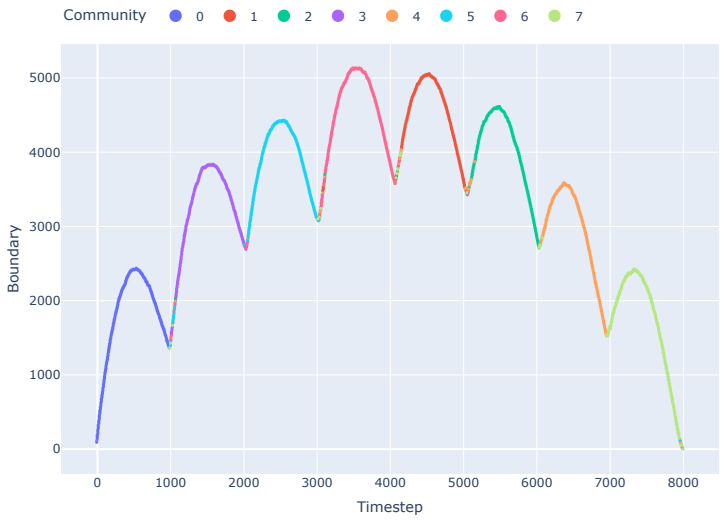
3. **Random Outsider (RO).** This strategy randomly samples an outsider with uniform probability from the set of outsiders.

4. **Random Insider and Random Outsider (RI_RO).** We randomly select an insider followed by a random outsider incident to this insider.

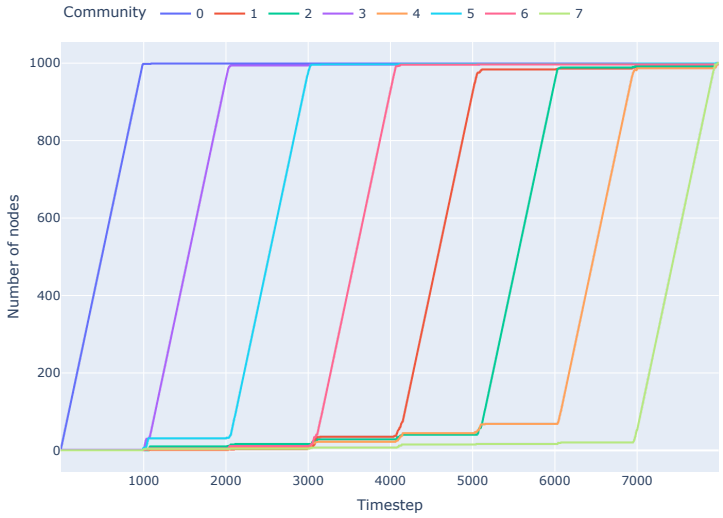
5.3 Evaluation

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 5.1(a), which showcases the boundary's dynamic changes in one of the synthetic network configurations.

In Figure 5.1(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 5.1(b). Here, an inflection point refers to the timestep at which the sampler starts sampling a new community.



(a) Evolution of network boundary with time



(b) Community size evolution

Figure 5.1: Sampling using MAS for a network of 8 blocks with 1000 nodes and one seed node each. The values of r and $\langle k' \rangle$ are set as 4 and 10, respectively. By comparing (a) and (b), we observe that the inflection points of (a) at 1000, 2000 etc, indicating significant changes in boundary evolution correspond to points in (b) where a new community starts getting sampled. The gradual increase in the boundary suggests that the sampler is entering more connected parts of the network, encountering more edges than it is absorbing in the network.

In stark contrast, when we investigate one of the random sampling methods, specifically Random Outsider (RO), applied to the identical SBM configuration, Figure 5.2(a) conspicuously lacks any dis-

cernible inflection points. Likewise, in Figure 5.2(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.

5.3.1 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.

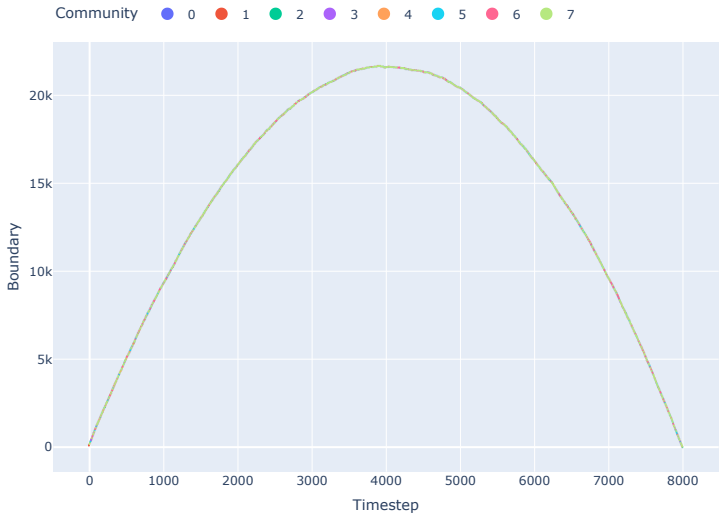
5.3.2 Ratio of intra- to inter-cluster edges (r)

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 5.3(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.

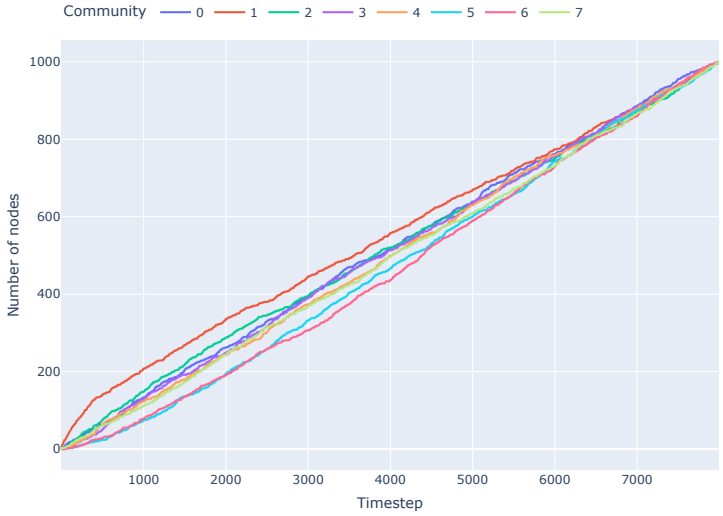
As depicted in 5.3(b), following the sampling of community '0', we observe that community '1' attempts to be entirely sampled but becomes contaminated with nodes from communities '2' and '3'. After approximately timestep 3200, no particular community exhibits a clear preference for complete sampling.

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 be-



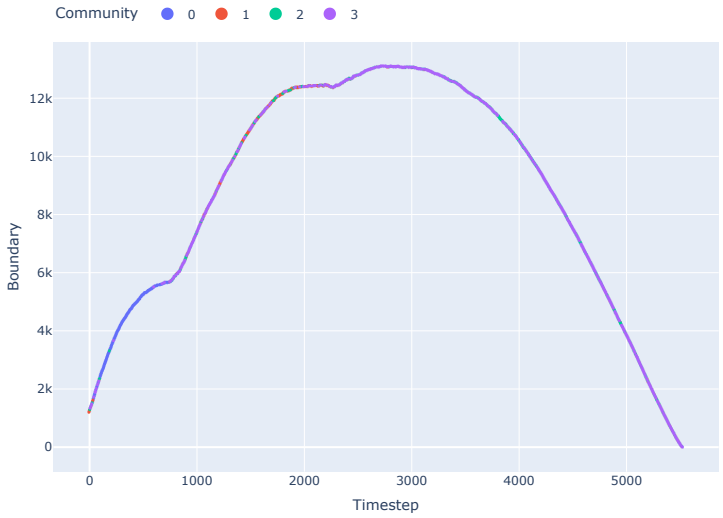
(a) Evolution of network boundary with time. We observe that all colours overlap, indicating simultaneous sampling of all the communities.



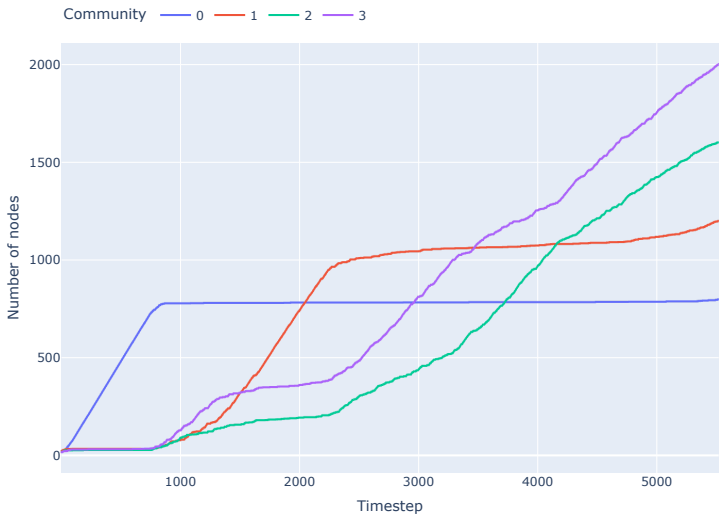
(b) Community size evolution

Figure 5.2: 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

tween the community that has been sampled and those that remain unsampled, all of which impact the directed boundary of the sampled nodes.



(a) Evolution of network boundary with time. We observe that after timestep $\sim$ 900, the remaining colours overlap, indicating that the communities represented by red, green and purple are being sampled simultaneously.



(b) Community size evolution

Figure 5.3: 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

5.3.3 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.

Chapter 6

Experiments - Empirical data

Moving from undirected and unweighted synthetic networks in the previous chapter, in this chapter, we analyze and understand the behaviour of our sampling scheme on a weighted and directed Twitter network.

6.1 Configuration

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 Chapter 4. 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 interaction frequencies - distinct, nested and audience-facing, and have been shown in Table 4.1.

6.2 Sampling

For the seed network generated above, along with the weights from Table 4.1, we sample the networks for distinct, nested and audience-facing variants. 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 6.1 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 6.1: Dataset statistics for the sampled Twitter network using the three variants of our sampling scheme and four variants of random sampling.

6.3 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.

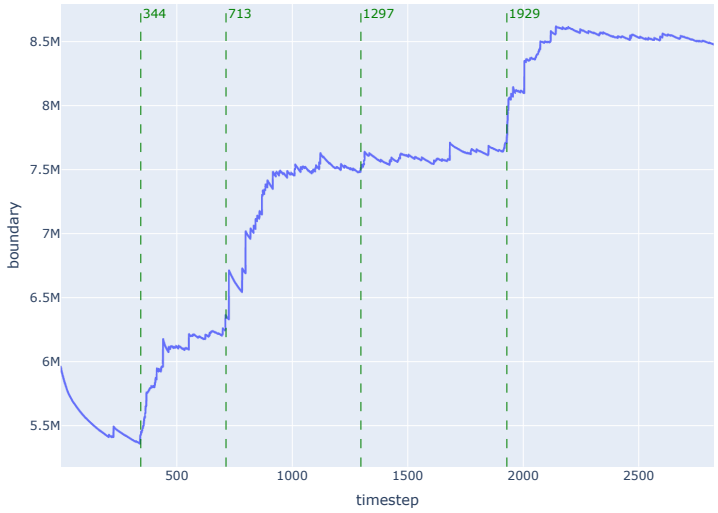
As observed in synthetic networks, we anticipate the presence of inflection points in the "boundary vs timestep" plots, indicating the transition from sampling one community to the next. In the context of the Twitter network, we notice a similar pattern, although occasional instances occur where two communities are sampled concurrently. For instance, in the case of sampling using the Audience-Facing (A-F) approach, exemplified in Figure 6.1(a), 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 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 labelled 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 6.1(b), 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 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 $\{344, 713, 1929\}$.

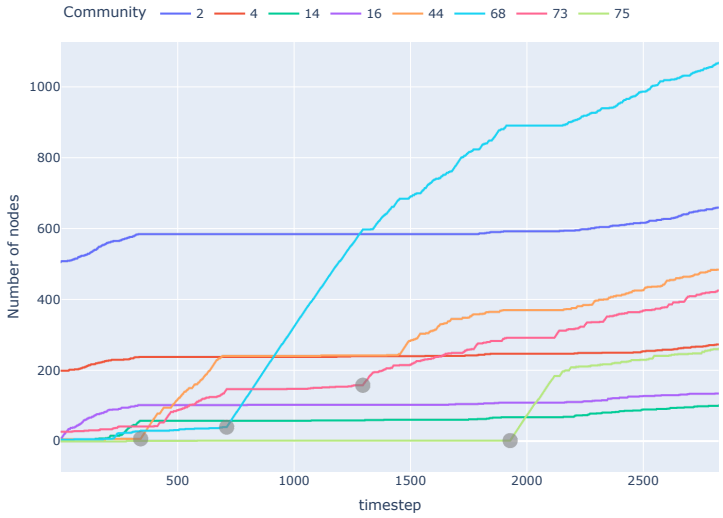
6.3.1 Structure based metrics

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.



(a) Evolution of network boundary with time



(b) Community size evolution

Figure 6.1: The four shaded circles in (b) indicate points where significant sampling of a new community starts. For corresponding timesteps in (a), we observe that 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 (a). However, every small rise in (b) might not correspond to a different community being sampled in (b). Hence, we tend to focus on bigger jumps in value of boundary timestep = $\{344, 713, 1929\}$ where a different community begins sampling as can be seen through the steep rise in community size.

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.

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
	Audience-Facing (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 6.2: Structural evaluation of networks obtained using priority-based and random schemes on Twitter. The bold values signify the highest or lowest values as per the chosen metric. To ensure comparison across sampling schemes despite different sampled network sizes, we consider the subgraphs sampled till the minimum common network size (one for priority based schemes, and one for random sampling) and calculate metrics for that snapshot.

As presented in Table 6.2, 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 6.2 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.

Chapter 7

Using long-range network regularities for node attribute prediction

In this chapter, we explore the application of datasets derived through the sampling approach detailed in earlier chapters. Our analysis primarily utilizes datasets gathered from "Audience-Facing" interaction frequencies. As indicated in Table 6.2, these datasets exhibit superior cohesion compared to those using Distinct or Nested interaction frequencies.

The sampling study was motivated by the objective of obtaining tightly clustered datasets conducive to downstream analytical tasks that necessitate such cohesion. In the current phase of our research, we leverage the concept of homophily to discern long-range patterns within the network, which are then applied in two key areas:

1. **Prediction of Node Attributes:** By exploiting the long-range regularities identified, we aim to predict node attributes. This prediction is intended to align with observed network patterns, facilitating the accurate characterization of nodes based on their network interactions.

2. **Identification of Fraudulent Actors:** Following the prediction of node attributes, our subsequent analysis focuses on detecting discrepancies between observed and predicted attributes. Significant disparities suggest deviations from established network patterns, potentially identifying nodes associated with fraudulent activity.

7.1 Selection of Attribute

Prior to initiating the studies described above, it is essential to select appropriate node attributes for prediction. The selection process is guided by evaluating potential attributes against the following criteria:

1. **Coverage**: The selected attribute should be prevalent across a significant portion of the nodes within the dataset. For instance, an attribute with 60% coverage means that 60% of the nodes have a recorded value for that attribute. This ensures a sufficient data volume for robust analysis.

2. **Categorical vs. Continuous**: Our methodology is adaptable to both categorical and continuous attributes. For categorical attributes, we aim to calculate the probability that a node falls into a particular category. Conversely, with continuous attributes, we predict a specific numerical value and validate this against the actual measured value. Attributes that do not conform to these types (neither categorical nor continuous) are excluded from the current study.

7.2 Preliminary analysis of attributes

In the process of evaluating our sampled Twitter dataset, we have identified a set of available attributes, which we will refer to as candidate attributes. These attributes are detailed in Tables 7.1 and 7.2, representing user and tweet characteristics, respectively. It is important to note that the Twitter Academic API directly provides these attributes; thus, our analysis will rely exclusively on these predefined attributes without incorporating any synthetically created ones.

The rationale behind avoiding synthetic attributes stems from potential risks of data leakage. If synthetic attributes were derived from the structural properties of the nodesproperties that are also utilized in our predictive modelsit could introduce bias. Our objective is to assess the extent to which the graph's structure alone can reflect (non-structural) social interactions. Therefore, our focus remains on non-structural attributes that capture the intrinsic qualities of users and tweets, avoiding any overlap with the structural data used in our models.

7.2.1 User attributes

In our study, user-related attributes such as the number of followers, friends, lists, and status updates present challenges due to their volatility. These attributes can change frequently, especially for active users, over short periods like days or weeks. Given that our dataset spans an extended timeframe, the values of these attributes may have shifted, thus compromising their reliability for consistent analysis. This temporal variability contrasts with the stability of tweet-related attributes, where potential changes over time were mitigated during the data collection process, as detailed in Chapter 6.

Our preliminary analysis of the 'Entities' attribute, which aggregates hashtags, mentions, and URLs, shows considerable promise due to its broad coverage. However, the predominance of URLs among these data points suggests that while this attribute provides valuable insights, its diversity and non-uniformity across different entities make it challenging to define its scope for our use case.

Similarly, the 'Location' attribute, while frequently populated, often contains non-serious or 'mock' entries like "Universe" or "ur home," introducing significant noise and reducing its utility.

The 'Verified' attribute initially seemed promising due to its high coverage and categorical nature. However, verification status is influenced by factors beyond mere user characteristics or network homophily, such as external administrative criteria, which complicates its use in predicting based on network behaviour alone.

Attribute	Description	Structure	Coverage
Verified	If the user is verified by Twitter or not. On the Twitter UI, this is often represented by a blue tick mark	Boolean (True or False)	100% where 13% have the true value
Follower count	Number of followers of the respective user	A positive integer	100%
Friend count	Number of people followed by the respective user	A positive integer	100%
Listed count	Number of public lists this user is a member of	A positive integer	100%
Status count	Number of tweets (including retweets and quotes) the user has posted	A positive integer	100%
Entities	Hashtags, Mentions and URLs in the user's Twitter profile	A JSON object containing list of hashtags, mentions and URLs	100%
Location	Entry in the "location" field on the user profile. However, this can have values like "Universe"	A string	59%

Table 7.1: Attributes obtained as a part of user's profile metadata

Due to these limitations, we have decided not to utilize user attributes for predictive analysis in our study, focusing instead on more stable and reliable tweet-related data.

7.2.2 Tweet attributes

In evaluating tweet attributes for our analysis, coverage is defined as the percentage of tweets that possess a specific attribute. Unlike user attributes, which are linked directly to individual profiles, tweet attributes require aggregation for effective integration into an interaction network - constructed with users as nodes and the interactions between them as edges.

Attribute	Description	Structure	Coverage
Context Annotations	Describes what the respective tweet is talking about	List of (domain, entity) pairs. Example: (City, Mumbai), (Politician, Narendra Modi)	44%
Entities	Hashtags, Mentions, Cashtags, URLs etc. included in the tweet	A JSON object containing list of hashtags, mentions, cashtags and URLs	100%
Source	The device use to tweet	A string of the format ”Twitter for Web”, ”Twitter for Android” etc.	27%
Public Metrics	Counts of likes, retweets, quotes and replies	A dictionary where key-value pairs respresent the counts of likes, retweets, replies and quotes	100%
Geo	The geo-location linked with the tweet	A JSON object containing information like coordinates about the specific geo location	3%
Language	Language of the tweet	Language tags like ENG etc.	97% (3% are undefined)

Table 7.2: Attributes obtained as a part of Tweet metadata

Among the tweet attributes, ‘Entities’ poses similar challenges to its counterpart in user attributes, primarily because of its heterogeneity and the extensive range of data points it encompasses. However, other attributes like ‘Context annotations’, ‘Source’, ‘Public Metrics’, and ‘Language’ display promising characteristics for our study. These attributes are notable for their substantial coverage, making them valuable resources for analyzing and understanding the dynamics within our interaction network.

The structure of ‘Context annotations’ is inherently complex, typically presented as a list of tuples, with each tuple representing a domain-entity pair for a tweet. To standardize this attribute for consistent application across the network, we transform it into a multidimensional vector. Each dimension within this vector represents a specific entity, such as ‘Canoeing & Kayaking’, ‘FIFA’, ‘Apple - iPhone’, etc. In our study, each entity in the context annotation is analogous to a topic discussed by the user. Here, the user’s value for a particular topic represents the number of their tweets containing the said topic.

The number of dimensions d in this vector corresponds to the total number of unique entities identified in our dataset, 2970. This approach enables us to quantify and analyze the thematic engagements of users across the network, providing an understanding of the information discussed and interaction within the Twitter landscape.

7.2.3 Future goals and prediction baselines

Building on our initial analysis of tweet attributes, we aim to develop a methodology that effectively captures long-range interactions within the network. This involves exploring various baseline strategies to compare and validate our proposed methods. These baseline methods might include:

1. **Random Prediction based on probability of observed values**: This method generates predictions by randomly selecting attribute values, weighted by their observed frequency in the dataset. This strategy ensures that the probability of predicting a specific attribute value reflects its actual distribution within the data.

2. **Prediction Using Mean of $\leq k$-hop Neighbors**: This approach leverages the network structure by predicting an attribute based on the mean values of that attribute observed in the k-hop neighbourhood of a node. This method hypothesizes that nodes within close network proximities might exhibit similar attribute characteristics.

By integrating these baseline models, we can better assess the effectiveness of our proposed methodology in capturing and predicting the interactions based on long-range regularities. These comparisons not only help validate our approaches but might also provide insights into the influence of network structure on user behaviour and vice versa. This groundwork is crucial for advancing our understanding of network dynamics and enhancing predictive accuracy in social network analysis.

Chapter 8

Conclusion and limitations

> Networks are present everywhere. All we need is an eye for them.
>
> Albert-Laszlo Barabasi

We propose a novel scheme for snowball-type sampling in unbounded networks designed to respect cohesive communities. Its intended purpose is the extraction of communities, and can be used as a form of local community detection.

Our approach consists of two main parts, a sampling priority utilizing the maximum-adjacency principle, and a method to integrate modes of interactions such as likes, retweets, replies and quotes 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 on Twitter. 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.

While the proposed sampling scheme tackles the problem of getting cohesive subgroups instead of any form of representativity, it should be made sure that the expected properties of the sampled network align with the expected result of the sampling scheme. For studies where the aim is to get a representative sample, other classic schemes like random sampling, that have been extensively researched in the past, might be more suitable. As mentioned in subsection 5.3.3, how well the clusters are separated within the underlying network determines the extent to which a community can be identified. For example, in case of a synthetic network made using stochastic block model, we observe a limiting case when the ratio of density within clusters to that between clusters is one. Thus, in cases where the distinction between different clusters is not apparent, caution should be exercised, and due analysis should be conducted on the underlying network.

Our research paves the way for future investigations into network sampling aimed at identifying cohesive communities, potentially generating new datasets and research initiatives that focus on studying

the dynamics within these groups. We discuss our current work built on top of the networks collected through our sampling method in Chapter 7, and hope that the given examples demonstrate the need and scope of learning more about dynamics within cohesive social groups.

Related Publications

- *Kshitijaa Jaglan*, Meher Chaitanya*, Triansh Sharma, Abhijeeth Singam, Nidhi Goyal, Ponnurangam Kumaraguru, Ulrik Brandes.* **Tight Sampling in Unbounded Networks.** Forthcoming in ICWSM 2024 - The 18th International AAAI Conference on Web and Social Media, June 3 - June 6, 2024, Buffalo, New York, USA

Bibliography

[Ahmed et al.,] Ahmed, N., Neville, J., and Kompella, R. Network sampling via edge-based node selection with graph induction.

[Ahmed et al., 2013] 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 et al., 2006] 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)*, pages 475–486. IEEE.

[Antonakaki et al., 2021] 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 et al., 2022] 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 et al., 2010] 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*, pages 98–109. Springer.

[Cai and Matula, 1993] Cai, W. and Matula, D. W. (1993). Partitioning by maximum adjacency search of graphs. *Partitioning Data Sets*, 19.

[Chang et al., 2019] 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.

[Chaudhuri et al., 2005] Chaudhuri, A., Chaudhuri, A., Stenger, H., and Stenger, H. (2005). *Survey Sampling: Theory and Methods, Second Edition*. CRC Press, 2 edition.

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

[Furusawa and Konishi, 2007] Furusawa, T. and Konishi, H. (2007). Free trade networks. *Journal of International Economics*, 72(2):310–335.

[Giudice and Ursino, 2019] 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 et al., 2010] 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*, pages 1–9. Ieee.

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

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

[Hübler et al., 2008] 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*, pages 283–292. IEEE.

[Kim et al., 2018] 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.

[Kuchler et al.,] Kuchler, T., Russel, D., and Stroebel, J. JUE insight: The geographic spread of COVID-19 correlates with the structure of social networks as measured by facebook. 127:103314.

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

[Lee et al., 2006] Lee, S. H., Kim, P.-J., and Jeong, H. (2006). Statistical properties of sampled networks. *Phys. Rev. E*, 73:016102.

[Leskovec and Faloutsos, 2006] 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*, pages 631–636.

[Li et al., 2015] 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*, pages 927–938. IEEE.

[Liu et al., 2019] 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 and Berger-Wolf, 2010] Maiya, A. S. and Berger-Wolf, T. Y. (2010). Sampling community structure. In *Proceedings of the 19th international conference on World wide web*, pages 701–710.

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

[Pavlopoulos et al.,] Pavlopoulos, G. A., Secrier, M., Moschopoulos, C. N., Soldatos, T. G., Kossida, S., Aerts, J., Schneider, R., and Bagos, P. G. Using graph theory to analyze biological networks. 4(1):10.

[Prez-Escoda et al., 2020] Prez-Escoda, A., Jimnez-Narros, C., Perlado-Lamo-de Espinosa, M., and Pedrero-Esteban, L. M. (2020). Social networks engagement during the covid-19 pandemic in spain: Health media vs. healthcare professionals. *International Journal of Environmental Research and Public Health*, 17(14).

[Ribeiro and Towsley, 2010] 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*, pages 390–403.

[Rozemberczki et al., 2020] 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*, pages 3125–3132.

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

[Spielman and Teng, 2004] 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*, pages 81–90.

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

[Stutzbach et al., 2006] 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*, pages 27–40.

[Wallace et al., 2012] Wallace, M. L., Larivire, V., and Gingras, Y. (2012). A Small World of Citations? The Influence of Collaboration Networks on Citation Practices. *PLOS ONE*, 7(3):e33339. Publisher: Public Library of Science.

[Ye et al., 2010] Ye, S., Lang, J., and Wu, F. (2010). Crawling online social graphs. In *2010 12th International Asia-Pacific Web Conference*, pages 236–242. IEEE.

[Zhang et al., 2023] 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 and Rohe, 2018] Zhang, Y. and Rohe, K. (2018). Understanding regularized spectral clustering via graph conductance. *Advances in Neural Information Processing Systems*, 31.