Sampling

Lena Trnovec[a,1] **and Loris Štrosar Grmek**[a,1]

[a] University of Ljubljana, Faculty of Computer and Information Science, Večna pot 113, SI-1000 Ljubljana, Slovenia

The manuscript was compiled on May 26, 2023

Sampling is a fundamental technique in network analysis used to reduce the size and complexity of graphs while preserving important properties. This article explores various sampling methods, including random selection, exploration techniques, and merging/aggregation methods, and their impact on network properties. The preservation of density and community structure are highlighted as key considerations in sampling. The self-similar scaling behavior of density allows for accurate insights into network characteristics using representative subsets. Different sampling techniques exhibit variations in preserving community structure, with some methods promoting community-like groups while others generating module-like groups. Overall, sampling enables efficient analysis, visualization, and inference in network analysis, providing valuable insights into complex networks.

Sampling refers to the technique of selecting a subset of vertices and/or edges from a larger, original graph. It has a broad range of applications in various fields, such as sociology, visualization, graph sparsification, and more. The motivation behind sampling is often to obtain a smaller graph that is more manageable for analysis and exploration. It can be applied when the entire graph is known, aiming to reduce its size, or when the graph is unknown, serving as a means to discover its structure.

Different sampling techniques, such as Vertex Sampling, Edge Sampling, and Traversal Based Sampling, are commonly used to achieve specific objectives. The article presents a taxonomy of graph sampling objectives and approaches, highlighting the relations between these approaches and providing a framework that connects theoretical analysis with practical implementation. One key interest lies in understanding which graph properties are preserved during the sampling process. If certain properties are preserved, efficient estimators can be constructed, and algorithms relying on those preserved properties can be expected to yield similar outputs on both the original and sampled graphs. This opens up avenues for accelerating graph algorithms systematically (1).

While sampled graphs are smaller in size, they may still exhibit similarities to the original graphs. The article explores the preservation of various classical and advanced graph properties and points out gaps in the existing research. While some theoretical studies and extensions are collected, a more systematic and neutral evaluation is necessary to shed light on further advancements in graph sampling studies.

Real-world networks pose challenges due to their large and evolving nature, and sampling offers a solution for their analysis and understanding. By reducing a network to a smaller sample, analysis and visualization become more feasible. Moreover, understanding the differences between complete original networks and their incomplete variants is crucial. Studies on network sampling have analyzed changes in network properties, such as degree distribution, clustering, connectivity, and more. Various sampling techniques have been compared based on their ability to preserve network properties, and efforts have been made to detect and correct biases in the sampling process. However, despite these endeavors, there is still much to be understood about the structural changes introduced by sampling and how network structure affects the performance of sampling techniques (2).

1. Overview

Snowball sampling, contact tracing, and random walks are network-based techniques used for sampling hidden populations in social network studies. These techniques are particularly valuable when studying populations such as drug users or illegal immigrants who are difficult to locate and interview. Snowball sampling involves identifying an initial member of the target population and then using their social network connections to find additional participants (Figure 1). This process creates a "snowball effect" and allows for a larger sample size (3).

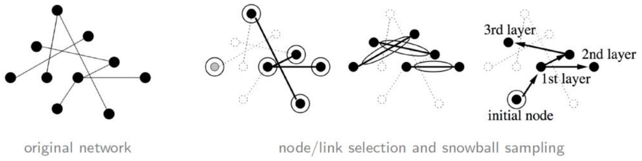


Fig. 1. Node/link selection and snowball sampling on networks.

Snowball sampling can be biased and may not provide accurate samples. Contact tracing, a similar technique used in disease incidence studies, traces the contacts of infected individuals to control outbreaks and collect data on disease spread, but it also has biases and incomplete samples. Random-walk sampling improves on bias issues by randomly selecting contacts for interviews, but it can be time-consuming. Respondent-driven sampling addresses the inability to directly name contacts by providing participants with tickets to distribute to acquaintances. However, it introduces challenges such as non-random distribution and participation refusal. Despite limitations, these techniques are widely used in social network studies to gain insights into hard-to-reach populations and disease transmission networks. (3).

Fractality and self-similarity are intriguing properties observed in real networks, revealing remarkable patterns within their structures.

All authors contributed equally to this work.

[1] To whom correspondence should be addressed. E-mail: lt89715@student.uni-lj.si

Definition (Fractality) *Fractality refers to the property of an object where it exhibits similarity to a part of itself, regardless of the scale of observation.*

This means that zooming in or out on different sections of the object reveals similar patterns or features. In the context of real networks, fractality implies that the network displays recurring motifs or structures that are present at various levels of organization, from local to global.

Definition (Self-similarity) *Self-similarity refers to the property of an object where it exhibits similar patterns or structures at different scales or levels of magnification*

This suggests consistent patterns and behaviors at different scales, indicating universal organizing principles that are independent of network size. This understanding provides valuable insights into the formation and evolution of complex networks. Fractality and self-similarity highlight recurring patterns and organizational principles, unraveling the fundamental mechanisms shaping real networks (4).

! Real networks, like social or information networks, are complex and vast. However, **any observed network is merely a sample of the true underlying network**. Sampling involves selecting a subset of nodes for analysis, and it can greatly impact the observed network's characteristics, especially its community structure. Sampling often promotes the formation and preservation of community structure in the observed network, leading to a stronger community organization compared to the true network. This recognition emphasizes the limitations and biases introduced by sampling in studying real networks (2).

A. Sampling motivation.

A.1. Whole data is not available. When you are limited by the API call limit or some similar restriction, you can make a sample of the graph by using one of the sample methods. You must know what is the point of interest to choose used method accordingly. In other words, you should know which graph properties are preserved by using the respective method.

A.2. Hidden population. Sampling nodes from a graph when not all nodes are directly observable or accessible. It arises when there is a subset of nodes in the graph that cannot be easily identified or included in the sampling process. Sampling from a graph with a hidden population can be challenging because the nodes of interest are not readily available for selection. Snowball sampling is often used in sociology to study a hidden population like drug abusers. They start the study with a small set of participants and expand it from there.

A.3. Visualization. Sampling is often used to facilitate easier visualization of large or complex graphs by reducing their size while preserving important structural characteristics.Large graphs can have thousands or even millions of nodes and edges, making it challenging to visualize them effectively. By sampling a subset of nodes and edges, the graph's size can be significantly reduced, allowing for better visualization on limited screen space.

A.4. Reduce test cost. Graph sampling can contribute to cost reduction in testing by reducing the scale and complexity of the graph being tested. Graphs can be computationally expensive to analyse, especially when dealing with large-scale networks. By sampling a subset of the graph, the computational and processing time required for testing can be significantly reduced.

B. Property preservation and property estimation. Property preservation and property estimation are closely related to each other. Preserving properties means ensuring that essential structural or statistical features of the original graph are retained after a particular operation or transformation is applied. If a graph is subjected to graph sparsification (reducing the number of edges while preserving key structural features), the goal is to preserve important properties such as connectivity patterns, clustering coefficients, or degree distributions. The aim is to create a sparser graph that still reflects the original graph's key properties and characteristics.

Property estimation, involves inferring or approximating specific properties of a graph based on a sample or subset of the graph's data. The goal is to estimate or predict various graph properties or characteristics when it is not feasible or practical to analyse the entire graph. If a graph is extremely large and computing certain properties directly is computationally expensive or time-consuming, property estimation techniques can be used to provide approximate values or statistical estimates.

Property preservation ensures that key properties are retained, while property estimation provides estimates or approximations when direct computation of properties on the entire graph is not feasible.

2. Methods

Properties that the method preserves are highlighted in green, properties that are not accurately captured by the method are highlighted in red.

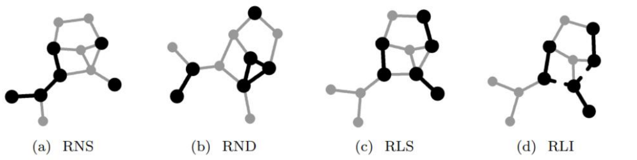
A. Random selection methods for global network sparsification. Sparsification is the process of reducing the number of nodes and/or links in the network while attempting to maintain its key structural and functional properties. For global network sparsification we use the following techniques (Fig. 2):

- *Random Node Selection (RNS):* Randomly selects nodes from the original network to form the sampled network. It does not consider any structural properties or connections between nodes during the sampling process.

- *Random Node Selection by Degree (RND):* Randomly selects nodes from the original network with a probability proportional to their degrees. Additionally, all the links connected to the selected nodes are included in the sampled network.

- *Random Link Selection (RLS):* RLS randomly selects links from the original network to construct the sampled network. Similar to RNS, it does not consider the network structure or the connectivity between nodes.

- *Random Link Selection with Induction (RLSI):* RLSI uses subgraph induction by selecting links randomly and including the nodes connected by those links in the sampled network. This technique preserves local structural properties by considering the connectivity between nodes.



(a) RNS (b) RND (c) RLS (d) RLI

Fig. 2. Random selection techniques applied to a small toy network. Highlighted nodes and links represent the samples obtained by different techniques (5)

B. Network exploration methods for local network sampling.

In this family of sampling techniques, the general approach involves randomly selecting a node and exploring its neighboring nodes (5, 6).

- *Snowball sampling (SBS):* It is a non-probability sampling technique, and it is often used in social science research. It is particularly useful when studying hidden populations. The goal is to sample nodes from a graph based on their connections. In snowball sampling on graphs, the process starts with selecting a set of initial nodes, often called "seeds," from which the sampling procedure begins. These seeds can be chosen randomly or based on specific criteria, such as high degree centrality or other relevant network measures. The initial set of nodes represents the starting point for the snowball sampling process.

- *Breadth-first exploration sampling (BFS):* A seed node is randomly selected, and its broad neighborhood obtained from a breadth-first search is included in the sample. BFS tends to select nodes with higher degree and performs well in matching degree distribution, average degree and clustering coefficient distribution but underestimates degree and betweenness centrality exponent. BFS is very similar to *Snowball sampling (SBS)*. BFS exhaustively expand the neighbourhood of current vertex while SBS only expands a fixed number of them.

- *Forest-fire sampling (FFS):* A broad neighborhood of a randomly selected seed node is retrieved using partial breadth-first search. The number of sampled links on each step follows a geometric distribution with mean $p/(1-p)$, where p is set to 0.7. FFS matches spectral properties well but fails to match path length and clustering coefficient.

- *Random walk sampling (RWS):* A random walk is simulated on the network, starting from a randomly selected seed node. The sample consists of links visited by the random walker, forming a connected subgraph. RWS performs well in matching transitivity, clustering coefficient distribution, and spectral properties but is biased towards nodes with high degree and fails to match degree distribution.

- *Metropolis-Hastings random walk (MHRW):* A random walk is simulated on the network, and on each step, the next-hop node is selected uniformly at random among the neighbors of the current node, or the random walker performs a self-loop. MHRW corrects the bias of RWS towards selecting nodes with higher degree but may get stuck in local communities.

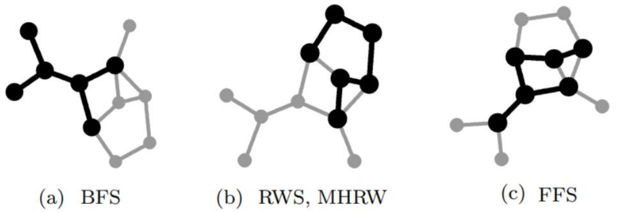


Fig. 3. Exploration techniques applied to a small toy network. Highlighted nodes and links represent the samples obtained by different techniques (5)

- *Random Node Neighbor (RNN) Sampling:* In RNN, a node is randomly selected along with all its outgoing neighbors. This technique imitates reading an edge file. It matches the out-degree distribution well but may not accurately capture in-degrees and community structure.

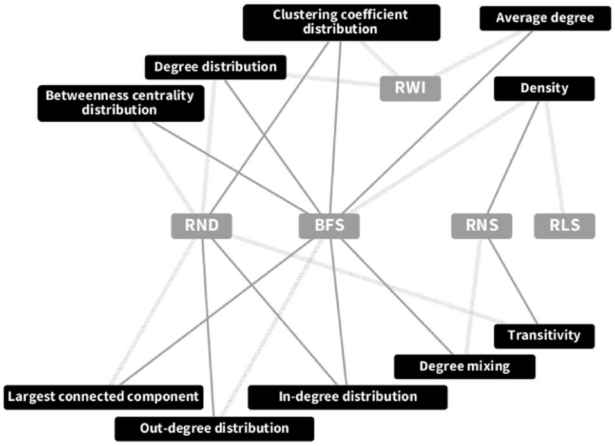


Fig. 4. Diagram of retained properties after application of different sampling methods.

C. Merging/aggregation methods for network simplification.

Merging or aggregation methods simplify networks by combining nodes or groups of nodes into larger entities. These techniques reduce network complexity while preserving structural characteristics (4).

- *The box covering method:* It is used to understand the scale-invariant properties (degree distribution, fractal dimension) of networks. It involves dividing the network into boxes based on node distances (Figure 5). If we cover the percolation cluster with N_B boxes of linear size ℓ_B. The fractal dimension or box dimension d_B is then given by

  $$N_b\ \ell_b^{-d_b}$$

  The renormalization process replaces each box with a single node, preserving properties like the degree distribution. Finding the minimum number of boxes to cover the network is challenging, but different coverings yield the same exponent. The method provides insights into network structures and their self-similar properties (4, 7).

- *Cluster growing:* In complex networks with broad degree distributions, the box counting method is not directly applicable. Instead, we can use the cluster growing method as an alternative. In this method, a seed node is randomly chosen, and a cluster of nodes connected to the seed is grown by considering a minimum distance (ℓ) between nodes. This process is repeated for multiple seed nodes, and the average "mass" of the resulting clusters ($\langle M_c \rangle$, defined as the number of nodes in the cluster) is calculated as a function of ℓ. It is a technique used to analyze the self-similar properties of complex networks with broad degree distributions, and it provides insights into the scale-invariant behavior of these networks (4, 7).
- *Community aggregation:* is an expansion method in network sampling that preserves the community structure. It involves grouping nodes into communities and selecting representative nodes from each community for the sample. This approach simplifies the network while retaining important community properties, making it useful for analyzing large-scale networks with complex community structures. It provides a manageable sample that allows researchers to study community characteristics and dynamics effectively.

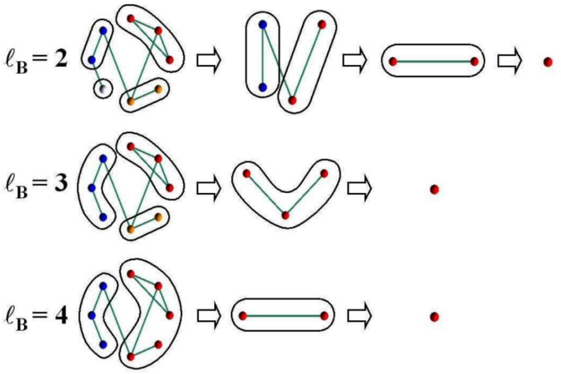


Fig. 5. Box covering method. The system is divided into boxes of size ℓ_B, with each box containing nodes connected within a minimum distance ℓ_B. The boxes are then replaced by renormalized nodes, and connections between renormalized nodes are established based on the connections between the original boxes. This process simplifies the network while preserving connectivity patterns. (4)

3. Comparison

A. Density. Network density refers to the number of connections within a network. It is determined by the proportion of actual connections in the graph with all the possible connections. Typically, it is calculated as the ratio of the actual number of connections divided by the number of all possible connections between nodes. Result ranges between 0 and 1, where the first means that the graph is not connected at all and the second tells us that we have a fully connected graph.

Certain self-organized networks display some properties relating to the density of connections and the size of the network. A notable example is found in utility networks across various European countries, where despite significant variations in size, the average degrees of these networks tend to be similar. In the article named Universal fractal scaling of self-organized networks (8), they examined the size and connection density of 47 self-organized networks of various origin. There was an obvious linear relationship between the network size N and the connection density d. They fit the data with d = $7.89N^{-0.986}$, and it revealed a power law relationship between the size and density of the networks. The scaling exponent tends towards negative one, signifying a fractal nature with 1/f properties. Despite the vast diversity of networks, there is a power law association between size and density.

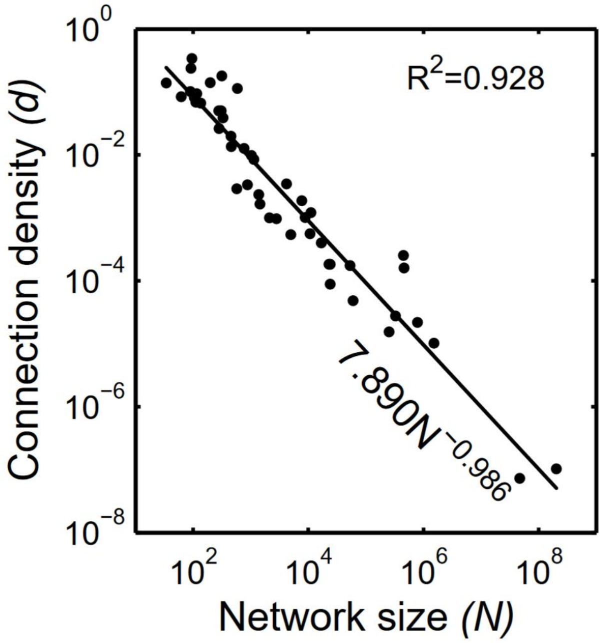


Fig. 6. Relationship between network size and density on logarithmic scale. Every point corresponds to a network based on existing literature. (8)

The observation of self-similar scaling in the density of real-world networks implies that the density exhibits a scale-free characteristic. In other words, as the network grows larger or undergoes renormalization, the density follows a specific pattern of increase or decrease, but its overall value remains consistent. This indicates that the density scaling is independent of the specific level of detail or resolution at which the network is examined. The connection between self-similar scaling of density and sampling lies in the preservation of density properties during the sampling process. The self-similar scaling behaviour suggests that sampling a subset from a larger network can provide valid insights into the density characteristics of the entire network. This enables researchers to analyse complex networks efficiently and draw meaningful conclusions based on representative subsets. (9)

B. Communities. A property which appears in many networks is community structure. Where nodes are more densely connected to each other within their own community than to nodes in other communities. Communities are groups of nodes that exhibit strong internal connectivity and relatively weaker connectivity with nodes outside their community. Understanding

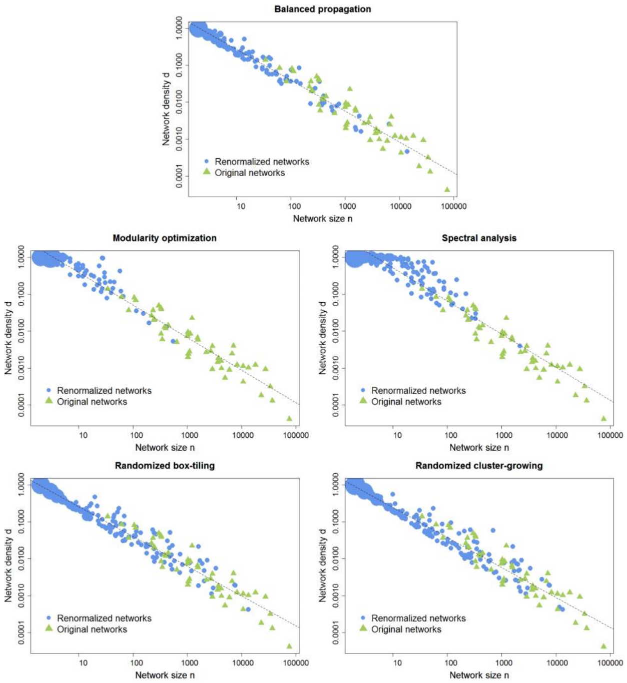


Fig. 7. Power-law scaling relationship between network density and size of 50 diverse real-world networks, using various renormalization techniques. Green triangles representing the original networks and blue circles representing their renormalized versions. The size of the symbols reflects the number of networks sharing the same size and density. (9)

community structure in graphs is crucial for various applications, as it reveals underlying patterns, functional modules, and organizational principles within complex networks.

In the article named, Sampling promotes community structure in social and information networks (2), they defined S as groups of nodes and T as the subset of nodes that represents a linking pattern. In other words, the pattern of nodes connections from S to the other nodes. Nodes in set of T are chosen to maximize the number of links between S and T and minimize the number of links between S and T complement. When S = T it characterizes community and S ∩ T = ∅ characterizes modules. Those are two extreme cases, and everything else are groups that are mixtures of the two.

The Jaccard index is defined as:

$$\tau(S, T) = \frac{|S \cap T|}{|S \cup T|}$$

$\tau \in [0, 1]$, and it can determine the type of group and corresponding linking pattern T. When $\tau = 1$ there are communities and τ represents modules. In between are mixtures.

RLS and FFS performs different from other methods. Their samples contain fewer groups, and almost all groups are modules, so the parameter τ approaches zero for all the networks. In comparison, of the original network to sampled one, those two methods contain fewer links from the original network. In numbers that means that sampled networks contain only 3% links of the original while other methods consists of around 16% of original links and that means that those networks are much sparser. For that percentage to mean something, we must state that created samples are 15% of original networks.

The performance of RLS and FFS can be understood based

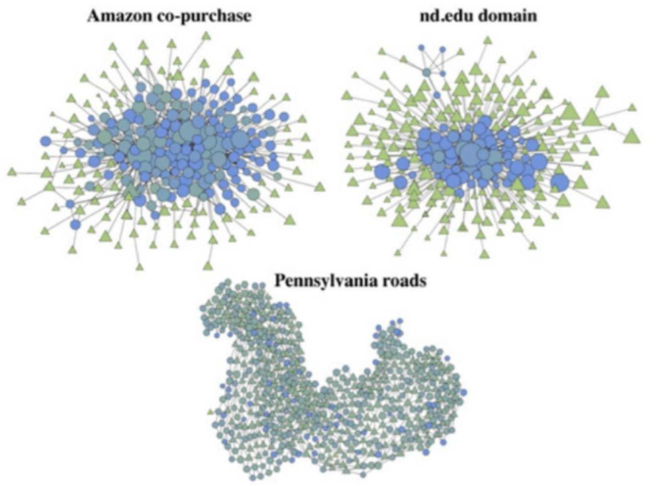


Fig. 8. Density of network structure in renormalized versions of three distinct real-world systems. Green triangles represent original networks and blue circles represent renormalized networks. (9)

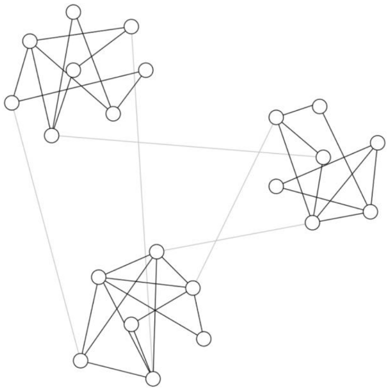
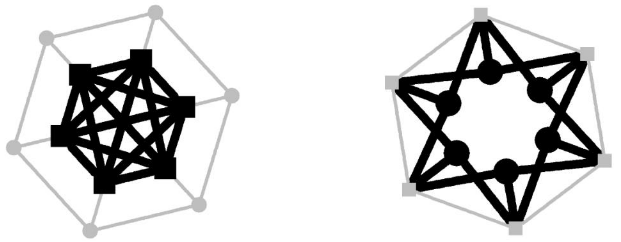


Fig. 9. An example of community structure. There are three communities that are densely connected, and there is much lower density of connection between each of them. (10)

on their respective definitions. In RLS, the sample includes randomly selected links, resulting in high variance. Consequently, the sample often contains numerous components with sparse connections, exhibiting a structure resembling modules. In contrast, FFS generates samples consisting of a single connected component. As a result, the sparsely connected nodes also form groups that exhibit module characteristics.

RND, RLI, BFS, and EXS perform similarly across all networks. In sampled information networks, the presence of mixtures decreases while communities become more prominent. As a result, the value of τ is higher in the sampled networks compared to the original networks. By comparing the number of groups and the parameter τ between the original networks and their samples, findings demonstrate differences. The original networks exhibit a considerably greater number of groups with a notably smaller τ compared to the sampled networks. The sampled social networks exhibit a larger parameter τ, indicating the presence of more community-like groups, while



(a) Community (b) Module

Fig. 10. Toy examples of groups of nodes in networks. Densely connected groups of nodes where $S = T$ and disconnected groups where $S \cap T = \emptyset$. (2)

the sampled information networks display fewer module-like groups. These results highlight alterations in the network structure caused by the sampling process, regardless of the network type or the specific sampling technique employed.

In RND, nodes with higher degrees are more likely to be included in the sample, and a similar bias is present in RLI. Consequently, densely connected groups of nodes have a higher probability of being sampled, while sparser regions of the network are not. On the other hand, BFS and EXS sample the broader neighbourhood of a randomly selected seed node, resulting in a sampled network that represents a connected component. BFS is biased towards sampling nodes with higher degrees and tends to overestimate the clustering coefficient.

To summarize, alterations in the structure of node groups introduced by sampling persist across different network types and various sampling methods.

4. Practical example - Estimation by random-walk sampling

1. Simplifying the representation of networks, we utilize simple undirected graphs and focus solely on their largest connected component. In our case, we have 5 different networks: Java class dependency network (java.net), nec overlay map of the Internet (nec.net), sample of Facebook social network (facebook.net), Enron e-mail communication network (enron.net), and a small part of Google web graph (www_google.net). The data is represented in table 1.

Table 1. Network Statistics

Graph	Nodes	Edges	Degree
'java'	2,378 (0)	14,619 (0)	12.30 (2,166)
'nec'	75,885 (0)	357,317 (0)	9.42 (13,346)
'facebook'	63,392 (0)	816,831 (0)	25.77 (1,098)
'enron'	84,384 (0)	297,314 (1,425)	7.05 (1,728)
'www_google'	855,802 (0)	4,291,352 (0)	10.03 (6,332)

2. Applying a random-walk sampling technique, we repeatedly sample nodes from the networks until we have sampled 15% of the total nodes, allowing for duplicates. Let's denote the number of sampled nodes and their corresponding degree sequence as s and $k_1, ..., k_s$ respectively. To estimate the average degree of the network, we employ a biased average

$$\frac{\sum_i k_i}{s}$$

and also determine a corrected estimate

$$\frac{s}{\sum_i k_i^{-1}}$$

.

Code:

```
def estimate_k(G, sample = 0.15):
    g = [list(G[i]) for i in G.nodes()]

    i = random.randint(0, len(g) - 1)
    sumk, sumk_1 = len(g[i]), 1 / len(g[i])
    s = 1

    while s < sample * len(g):
        i = random.choice(g[i])
        sumk += len(g[i])
        sumk_1 += 1 / len(g[i])
        s += 1

    return sumk / s, s / sumk_1
```

3. Comparing both estimates to the true average degree of the network, we can see that the corrected estimate is much closer to ground truth (table 2).

Table 2. Biased average vs. corrected estimate

Graph	True av. degree	Estimated	Corrected
'java'	12.30	598.61	11.87
'nec'	9.42	1303.37	9.70
'facebook'	25.77	90.12	26.51
'enron'	7.05	168.39	6.86
'www_google'	10.03	172.52	9.80

Conclusions

In conclusion, sampling is a vital technique in network analysis that helps reduce the size and complexity of graphs while preserving important properties. It allows for more manageable analysis and visualization of large networks. Sampling methods, such as random selection and exploration techniques, as well as merging/aggregation methods, are employed to achieve specific objectives. Density and community structure are two essential properties affected by sampling. The self-similar scaling behavior of density allows for valid insights into network characteristics using representative subsets. Sampling techniques also impact community structure, with some methods promoting community-like groups while others generate module-like groups. Overall, sampling enables efficient analysis, visualization, and inference, leading to valuable insights in network analysis.

1. Pili Hu and Wing Cheong Lau. A survey and taxonomy of graph sampling. *CoRR*, abs/1308.5865, 2013. URL http://arxiv.org/abs/1308.5865.
2. Neli Blagus, Lovro Šubelj, Gregor Weiss, and Marko Bajec. Sampling promotes community structure in social and information networks. *Physica A*, 432:206–215, 2015.
3. Mark E. J. Newman. *Networks.* Oxford University Press, Oxford, 2nd edition, 2018.

4. Chaoming Song, Shlomo Havlin, and Hernan A. Makse. Self-similarity of complex networks. *Nature*, 433(7024):392–395, 2005.
5. Neli Blagus, Lovro Šubelj, and Marko Bajec. Empirical comparison of network sampling: How to choose the most appropriate method? *Physica A*, 477:136–148, 2017.
6. Jure Leskovec and Christos Faloutsos. Sampling from large graphs. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 631–636, Philadelphia, PA, USA, 2006.
7. T. Vicsek. *Fractal Growth Phenomena*. World Scientific, Singapore, 2nd edition, 1992.
8. Paul J. Laurienti, Karen E. Joyce, Qawi K. Telesford, Jonathan H. Burdette, and Satoru Hayasaka. Universal fractal scaling of self-organized networks. *Physica A*, 390(20):3608–3613, 2011.
9. Neli Blagus, Lovro Šubelj, and Marko Bajec. Self-similar scaling of density in complex real-world networks. *Physica A*, 391(8):2794–2802, 2012.
10. Michelle Girvan and Mark EJ Newman. Community structure in social and biological networks. *Proceedings of the national academy of sciences*, 99(12):7821–7826, 2002.