Chaos

Preserving the topological properties of complex networks in network sampling

Cite as: Chaos **32**, 033122 (2022); https://doi.org/10.1063/5.0076854
Submitted: 28 October 2021 • Accepted: 04 March 2022 • Published Online: 18 March 2022

Wen-tao Chen, An Zeng and Xiao-hua Cui

Preserving the topological properties of complex networks in network sampling

Cite as: Chaos **32**, 033122 (2022); doi: 10.1063/5.0076854
Submitted: 28 October 2021 · Accepted: 4 March 2022 ·
Published Online: 18 March 2022

Wen-tao Chen, An Zeng, and Xiao-hua Cui[a)]

AFFILIATIONS

School of Systems Science, Beijing Normal University, Beijing 100875, China

[a)]**Author to whom correspondence should be addressed:** xhcui@bnu.edu.cn

ABSTRACT

Extremely large-scale networks have received increasing attention in recent years. The development of big data and network science provides an unprecedented opportunity for research on these networks. However, it is difficult to perform analysis directly on numerous real networks due to their large size. A solution is to sample a subnetwork instead for detailed research. Unfortunately, the properties of the subnetworks could be substantially different from those of the original networks. In this context, a comprehensive understanding of the sampling methods would be crucial for network-based big data analysis. In our work, we find that the sampling deviation is the collective effect of both the network heterogeneity and the biases caused by the sampling methods themselves. Here, we study the widely used random node sampling (RNS), breadth-first search, and a hybrid method that falls between these two. We empirically and analytically investigate the differences in topological properties between the sampled network and the original network under these sampling methods. Empirically, the hybrid method has the advantage of preserving structural properties in most cases, which suggests that this method performs better with no additional information needed. However, not all the biases caused by sampling methods follow the same pattern. For instance, properties, such as link density, are better preserved by RNS. Finally, models are constructed to explain the biases concerning the size of giant connected components and link density analytically.

In this paper, we discuss the biases that arise during network sampling. Although there are many different sampling methods, we discuss three of the most common ones: random node sampling (RNS), breadth-first search (BFS), and the hybrid method. Overall, the sampling biases can be attributed to two causes: network heterogeneity and the biases caused by sampling methods. We mainly focus on the latter, as they are determined by the sampling methods themselves, and these biases can be regarded as systematic. Carrying out both empirical and analytical studies, we find that these biases vary with different network indicators. For the main contribution of this article, the systematic errors on the size of giant connected components and link density are discussed analytically using some models. These results can help us better understand the information gathered via sampling from large networks.

I. INTRODUCTION

A complex network is a model suitable for representing the relationship between agents of complex systems.[1,2] It has received sustained attention because of its simplicity and universality in terms of applications to systems, such as infrastructure systems,[3–8] socioeconomic systems,[9–11] and biological systems.[12–14] In recent years, the development of big data has established a solid foundation for the accumulation and analysis of various complex networks with the scales of networks becoming increasingly larger. On the one hand, it is an opportunity for studies of complex systems, as the accumulated data are sufficiently large to match the complexity of the system. The tools of statistical physics can be applied at this time as the scale of networks are sufficiently large to approach the thermodynamic limit, and some properties such as scale-free[15] become universal under this assumption. On the other hand, the large scale of networks makes it difficult to perform analysis directly on them. The solution is to sample a subnetwork from the original network and study its properties instead. However, this is only an approximate method since the differences concerning many topological properties between the original network and the sampled subnetwork have not yet been sufficiently understood.[16]

To be accurate, there are already some works that detect the differences in topological properties between subnetworks

obtained from the same real system[17,18] or by doing some sampling experiments.[19–23] These works suggest that the information gathered from networks will also be affected by the "noise" introduced by sampling methods. In this context, the effects of network sampling need to be studied to provide a comprehensive understanding of the information gathered via these methods.

Some seminal works have already been done to discuss the differences concerning some topological properties between the subnetworks and the original networks with statistical methods or empirical methods.

Among the work based on statistical methods, the earliest one can be traced back to the 1980s.[24] From the view of statistics, the indicators of subnetworks can be regarded as an estimate of those in the corresponding original networks. Therefore, statistical properties, such as bias, could be analyzed.[20] However, only a few properties can be studied in detail due to their analyzability. The most discussed indicators are degree[25,26] or indicators based on degree.[27] The sampling method is usually restricted to methods based on random samplings, such as random node sampling (RNS) or random link sampling (RLS).[20]

However, methods based on random sampling are not the only sampling methods applied, as the subnetworks sampled by these methods may exhibit poor connectivity due to the sparsity of real networks. Edge lists are usually used as the storage format of most complex networks. Methods based on neighbor exploration are usually more effective: start from a single node, explore its neighbors, and repeat. One should note that all the methods we discuss here sample nodes without replacement. These methods are also called traversal in some literature.[28] The most representative methods are breadth-first sampling (BFS) and depth-first sampling (DFS).

Combining these two kinds of sampling methods, there is the snowball sampling method,[29] which can be regarded as the BFS method with multiple starting points (root nodes). We find that there exists some confusion regarding the definition of this sampling method: in some literature, the snowball sampling method is referred to as BFS,[20,30] whereas it is defined as BFS starting from various root nodes in other literature.[25] In our work, the latter definition is used to distinguish snowball sampling from BFS.

It is hard to perform analytical analysis for the sampling methods concerning neighbor exploration due to the network heterogeneity. Hence, research is usually based on results obtained empirically. Some experiment-based works consider the difference of indicators between the subnetworks and the original networks under different sampling methods.[19,21–23] These works provide approximate answers to the question that which sampling method has better performance for some specific indicators.

The above works analyze the biases that arise during network sampling with some sampling methods. To eliminate these biases, some seminal works have been done to design more reasonable and effective sampling methods concerning different network structures, say nodes, edges, and connected induced subnetworks under two different data access assumptions.[28] Among these works, some sampling methods based on importance sampling are designed to help estimate degree or degree distribution.[28,31] Other works are aiming at sampling higher-order network structures (pattern) like three-node connected subgraphs[32–34] or connected subgraphs with more than three nodes.[35–38]

When analyzing the effectiveness of the abovementioned sampling methods, most researchers compare the similarities between the subnetworks and the original networks. In this context, the problem of network sampling is equivalent to the problem of network similarity. For this part of the work, different measures are created to directly represent topological differences between networks, or by using kernel methods.[39] The former works include those that focus on the maximum subgraph,[40] the minimum subgraph,[41] and the edit distance,[42] or using the breadth-first search algorithm with pruning to search the common part of two different networks.[43–45] However, the extremely high computational complexity of these algorithms makes them unsuitable for extremely large networks. Therefore, similarity measures based on kernel methods are studied[39] to ensure computational feasibility. These works include those that use the distribution of the shortest paths to define the dissimilarity between different networks,[46] or those that use the differences between the Laplacian matrix eigenvalues.[47]

The above works show that network sampling and network similarity issues have received widespread attention. In our work, we use three interrelated sampling methods to study the biases concerning some topological properties caused by network sampling. We conclude that the bias caused during the sampling process could be due to two main reasons: the heterogeneity of complex networks and the mechanisms of sampling methods. The former biases always exist because we estimate global information with partial information. These errors are similar to accidental errors, as they can sometimes be reduced by repeated experiments. However, the latter biases are determined by the sampling methods applied, and we regard these as systematic errors. In this paper, we focus on the latter biases to show how these systematic errors affect different network topological properties during network sampling. We use both empirical and analytical methods so that one can have a more comprehensive understanding of the biases caused by the sampling methods. For the main contribution of this article, we provide a theoretical analysis of systematic errors on the size of giant connected component (GCC) and link density.

The organization of this paper is as follows: In Sec. II, we briefly review some popular sampling methods and then we describe in detail the sampling methods we have considered. In Sec. III, we report the results of experiments on real networks. In Secs. IV and V, the properties of the size of giant connected component (GCC) and link density under different sampling methods are discussed to show their different underlying mechanisms. In Sec. VI, a conclusion is provided.

II. METHOD DESCRIPTION

As mentioned above, there are several different sampling methods on real networks. Considering the efficiency, none of them rely on the complex network properties. These different methods can be divided into two categories: those based on random sampling and those based on neighbor exploration.

For the first category, RNS or RLS samples nodes or links with uniform probability from the original network. Based on these methods, one can also sample nodes with the probability related to

their degree to obtain other methods. For the second category, there are classic BFS and DFS methods, which can also be modified to obtain other sampling methods. For example, ignoring nodes with a certain probability when using BFS, one can obtain what is called forest fire sampling.[48] As another example, the random walk sampling method is to randomly choose a neighbor as the next start node while performing BFS. Independent of the above sampling methods, whether the induced network resulting from sampling needs to be extra considered. An induced subnetwork includes all the links with two ends in the sampling set.

Although there are many different sampling methods, the most representative ones, as we mentioned, are RNS and BFS. Here, we provide a detailed description of these two sampling methods. For RNS,

- First, given the sampling rate S and the node number of the original network N, then we randomly sample SN nodes from the original network without replacement. The probability of each point being selected is the same.
- Then obtain the induced subnetworks as the one consisting of all the links with both ends in the sampled set.

For BFS, given the sampling rate S and the node number of the original network N:

- First, a root node is randomly selected into the sampling set.
- Second, all the nodes that are the neighbor of the nodes in the sampling set but not in the sampling set are detected (we sample without replacement).
- Third, all the nodes detected in the third step are sampled into the sampling set if the number of nodes in the sampling set is less than or equal to SN. Otherwise, only some of them are randomly selected into the sampling set to ensure that the number of nodes in the sampling set reaches SN.
- Then, repeat the second and third steps until the number of sampled nodes reaches SN. If there are no new neighbors and the number of sampling nodes is less than SN, then choose an alternative unsampled node as the new root node and repeat the second and the third steps.
- Finally, obtain the induced subnetworks as the one consisting of all the links with both ends in the sampled set.

To discuss the effect on the indicators caused by these two methods, we construct a hybrid method (see Fig. 1) using a method similar to the snowball sampling method under different root node rates (root rate, i.e., proportion of root nodes in sampled nodes) as a natural way to connect RNS and BFS. The difference between the hybrid method and snowball sampling is that the latter usually focuses on the "wave," which means that the nodes located at the same distance from the root nodes are sampled at once. We do not require this but try to keep the groups related to the different root nodes the same size in the hybrid method.

Globally speaking, the hybrid method can be regarded as SRN repetitions of BFS, where S, R, and N denote the sampling rate, the root rate, and the number of sampled nodes in the whole process of sampling, respectively. Each repetition is BFS with the same sampling rate of $S/(SRN) = 1/RN$, with all the newly sampled nodes not being included in the other repetitions. The detailed description of this sampling method is a little complicated as we try to keep the size of each repetition the same, which can be found in S1 in the supplementary material with the form of pseudocode. For this hybrid method, one should notice that RNS and BFS are special cases of this sampling method: the hybrid method degenerates into RNS when $R = 1$, and it degenerates into BFS when $R = 0$. Therefore, the characteristics of this method become closer to that of RNS with larger R.

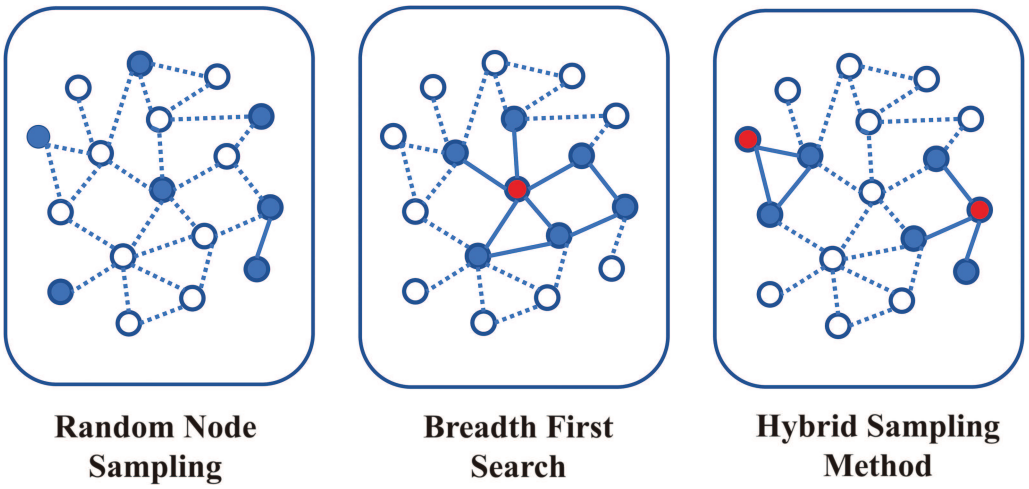

FIG. 1. RNS, BFS, and the hybrid sampling method. The solid nodes and lines represent those who are in the sample set. The hollow nodes and dotted lines denote those who are not in the sample set. We use the red nodes to represent the root nodes (the start nodes of sampling methods based on neighbor exploring). The hybrid methods with different root node rates (root rate, R, i.e., proportion of root nodes in sampled nodes) is a natural way to connect the other two methods.

When studying the differences concerning network topological properties, we need to choose some specific indicators and different measures for detailed research. The indicators used in this paper include link density, transitivity (proportion of triangles in all three-point groups), the size of giant connected component (the number of nodes in the giant connected accounts for all the nodes in the network, we use the symbol GCC to represent this measure), pairwise connectivity (the ratio of node pairs that are in the same connected component), degree distribution (degree dist), the average degree (avg degree), clustering coefficient distribution (CC dist), and the average clustering coefficient (avg CC). As for the measures to determine the difference of corresponding indicators, the absolute error is used. However, when it comes to the indicators concerning distributions, D-statistics in ks-test is applied as in some previous works.[19,23]

III. SAMPLING ON REAL NETWORKS

In this section, we report the results of empirical sampling experiments on real networks. There are already some previous works that show that the sampling methods will affect the topological properties in the subnetworks sampled from the same network,[19,20,22,23] which provide us with ideas for designing the experiments. The main results here are consistent with the previous evidence, i.e., the sampling methods do affect the information acquired from sampled subnetworks. However, as only the three sampling methods are focused on in this article, we find more interesting results in our experiments: both the regular pattern and the irregular pattern occur when we try to describe the change of the optimal sampling method. We attribute these results to the fact that the sampling deviation is the collective effect of both the network heterogeneity and the biases caused by the sampling methods. The regular pattern occurs when the latter plays a leading role. When the indicator is sensitive to network heterogeneity, an irregular pattern emerges.

Some real networks are used to perform empirical experiments first. Using these experiments, we want to provide some rough but intuitive results to show how the biases introduced by sampling methods affect the subnetworks. To highlight these systematic biases, we performed repeated experiments to reduce the impact caused by network heterogeneity. However, it is vital to use large networks because the discussion of this topic is valuable only in this situation. As a compromise between the above two issues, several networks with 1000 to 10 000 nodes are studied. These real networks include yeast protein–protein binding network generated by yeast two hybridization (Y2H), yeast protein–protein binding network generated via tandem affinity purification experiments (TAPs), high-energy theory collaborations (HEPs), Email (only the part corresponding to the original GCC), coauthorships in network science (NetSci), and hyperlinks between weblogs on US politics (USP). All of these networks are converted to undirected networks, the properties of which are shown in Table I.

We repeat the following two-stage experiments to obtain the empirical results:

- Given a certain sampling rate S, subnetworks are first sampled according to different root rates R.
- Then, we compare the difference of indicators under different root rates but the same sampling rate with the corresponding measures.

TABLE I. Description of real network used in sampling experiments. These networks are undirected without self-loops or transformed into the undirected network. We only use the networks with 1000–10 000 nodes because the experiments need to be repeated several times to reduce the impact of network heterogeneity.

Description	Node	Link	Link density	GCC
Y2H[49]	2111	2 203	9.89×10^{-4}	0.6907
TAP[50]	1373	6 833	7.26×10^{-3}	1
HEP[51]	8361	15 751	4.51×10^{-4}	0.6979
Email[52]	1133	5 451	8.50×10^{-3}	1
NetSci[53]	1589	2 742	2.17×10^{-3}	0.2385
USP[54]	1490	16 715	1.51×10^{-2}	0.8201

The first question we asked is whether there is an optimal root rate for all the different network properties. To help detect the optimal root rate (i.e., the optimal sampling method) under the given sampling rate S, we show the mean absolute errors (MAEs) or the mean absolute values of the corresponding metrics with heatmaps. Part of the results is presented in Fig. 2.

As a quick result, the times that each sampling method outperforms the others in preserving different indicators under distinct networks are counted. We display the frequencies that the sampling method performed the best in Table II, with triples to present these frequencies in their three positions. At most times, the hybrid method performed the best (among the indicators we considered), suggesting that this method seems to be the best one without any additional information. However, one should also realize that this method has a larger parameter space (root rate R). Thus, how to choose a proper parameter remains a question. In addition, not all the indicators share the same pattern, as link density and transitivity are usually better preserved by RNS. These results suggest that different indicators do not comply with the same pattern.

Then, the question arises: what are the differences between the underlying mechanisms of the biases concerning these distinct indicators. This problem leads us to conduct a more detailed theoretical analysis. All indicators used in this article can be divided into four groups. That is, the indicators that can be used to describe topological properties concerning the degree (avg degree, degree dist), clustering coefficient (avg CC, CC dist), connectivity (GCC, pairwise connectivity), and network density (link density, transitivity). We find that almost all the indicators in the same group show very similar patterns except for the pair of clustering coefficients (S2 in the supplementary material). One could find that the optimal root rates in the corresponding heatmaps show noisier patterns in S2.5 and S2.6 in the supplementary material. We attribute these phenomena to that the clustering coefficients being more sensitive to the subnetworks being sampled. As a result, the clustering coefficients in the subnetworks are more severely affected by network heterogeneity and these effects are more difficult to analyze.

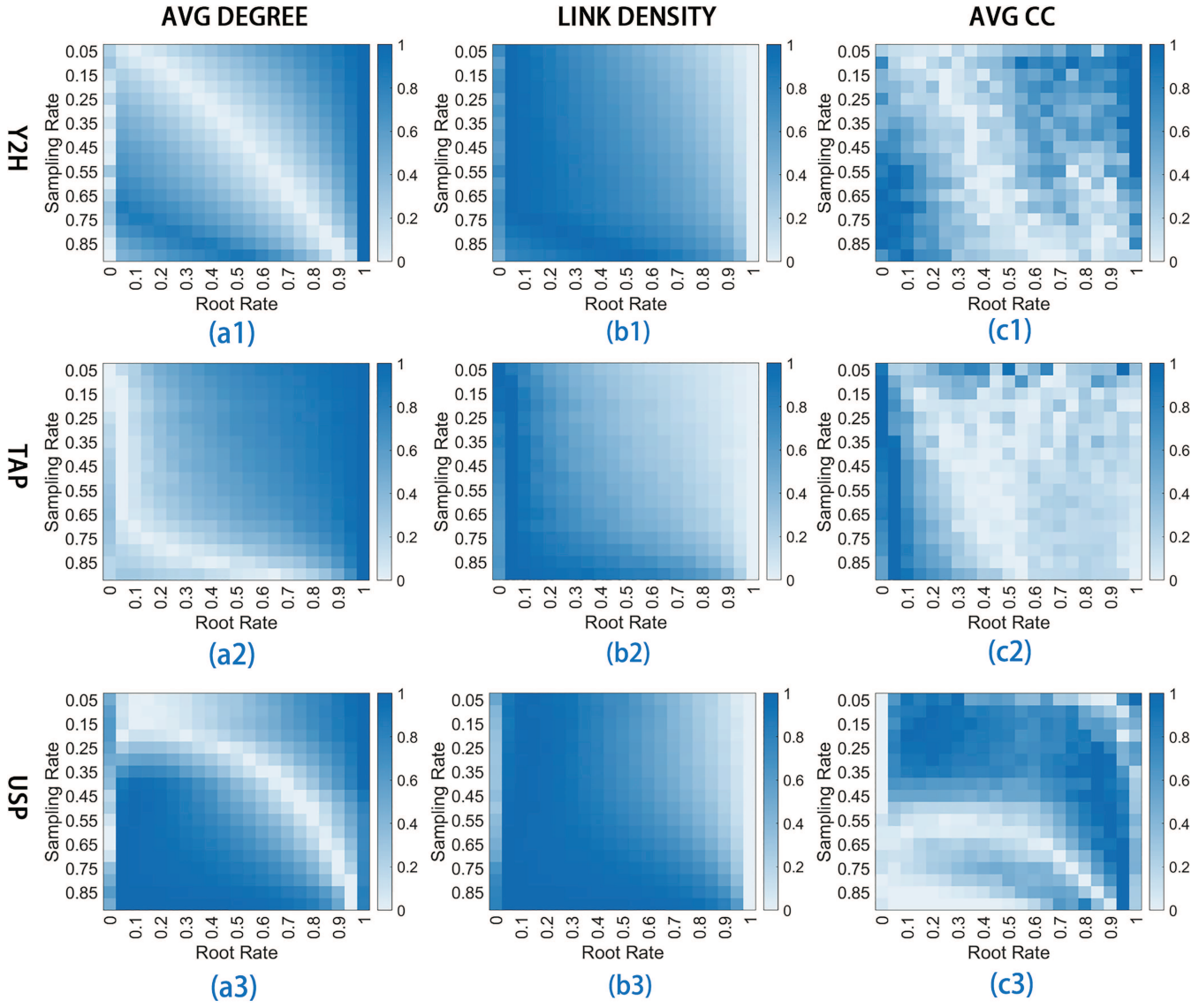

FIG. 2. Results of network sampling experiments. We use mean absolute errors (MAEs) to describe the ability that a specific sampling method can be used to preserve the corresponding network indicators in subnetworks. All the MAEs under the same sampling rate (i.e., in the same row of each graph) are rescaled [i.e., $(x - x_{min})/(x_{max} - x_{min})$] to help detect the optimal root rate. Note that the lighter the color, the smaller the MAE and the difference of the corresponding indicator. The results we show here are only concerning AVG DEGREE (labeled with a), LINK DENSITY (b), and AVG CC (c) under H2Y (labeled 1), TAP (2), and USP (3). More other results, including the MAE results before rescaling, can be found in S2 in the supplementary material. One can find that the changes of MAEs are regular (like graphs labeled with a or b) or irregular (c); network-independent (b) or network-dependent (a or c).

Finally, we choose GCC and link density for our further study. These two indicators, rather than the others, are chosen for the following two reasons: (1) There are already several works concerning degree,[25,26] and the analysis of the average degree is somewhat close to that of the link density, as they are similar by definition. (2) The analysis of pairwise connectivity is based on the distribution of the sizes of the connected components, and transitivity can be regarded as a higher-order property to describe the network density. The indicators of pairwise connectivity and transitivity are relatively more complicated, whereas GCC and link density share similar patterns with them. Therefore, we focus on the indicators of GCC and link density in Secs. IV and V.

TABLE II. The frequency of each sampling method outperforms the others at the preservation of the corresponding network indicator. We get the frequencies by these two steps: first, the sampling method (RNS, hybrid, or BFS) corresponding to the optimal root rate is found under the given sampling rate S. Then, the frequencies that each sampling method occurs are calculated under different sampling methods. The frequencies are presented in the three positions of triples. For example, (0.11, 0.89, 0) means that BFS accounted for 0.11, the hybrid method accounted for 0.89, and RNS accounted for 0 among all the optimal sampling methods appearing in the experiments. If two different sampling methods perform equally well, they win 0.5 experiments. Most times the hybrid method performs the best. However, RNS performs the best for link density and transitivity, which both describe the density of networks. When the original network is connected (TAP and Email), BFS performs better concerning GCC and pairwise connectivity. More detailed information can be found in S2 in the supplementary material.

Network	avg CC	CC dist	AVG degree	Degree dist	GCC	Pairwise connectivity	Link density	Transitivity
Y2H	(0, 1, 0)	(0, 1, 0)	(0.11, 0.89, 0)	(0, 1, 0)	(0, 1, 0)	(0, 1, 0)	(0, 0, 1)	(0, 0, 1)
HEP	(0, 1, 0)	(0, 1, 0)	(0.06, 0.94, 0)	(0, 1, 0)	(0, 1, 0)	(0, 1, 0)	(0, 0, 1)	(0, 0, 1)
TAP	(0, 0.94, 0.06)	(0, 1, 0)	(0.11, 0.89, 0)	(0.11, 0.89, 0)	(0.67, 0.33, 0)	(0.67, 0.33, 0)	(0, 0, 1)	(0, 0, 1)
Email	(0.06, 0.94, 0)	(0.17, 0.83, 0)	(0, 1, 0)	(0, 1, 0)	(0.53, 0.47, 0)	(0.53, 0.47, 0)	(0, 0, 1)	(0, 0, 1)
NetSci	(0, 1, 0)	(0, 1, 0)	(0, 1, 0)	(0, 1, 0)	(0, 1, 0)	(0, 1, 0)	(0, 0, 1)	(0.06, 0, 0.94)
USP	(0.56, 0.44, 0)	(0.61, 0.39, 0)	(0.11, 0.89, 0)	(0, 1, 0)	(0, 1, 0)	(0, 1, 0)	(0, 0, 1)	(0, 0, 1)

IV. THE IMPACT OF SAMPLING METHODS ON GCC

In this section, we present a theoretical analysis of the GCC in the subnetwork under different sampling methods. As far as we know, there are already some results focused on the giant connected component in subnetworks. For example, some rough empirical experiments show that BFS has a relatively better effect in preserving GCC.[23] Some works are focused on the number of connected components.[55,56] Here, we provide an analysis of the size of giant connected components (GCCs), a popular indicator in the field of a complex network.

We begin by presenting our empirical results using heatmaps so that one can easily detect the differences in optimal root rates concerning distinct indicators. To help distinguish the differences between the sampling methods, we rescaled all the MAEs under the same sampling rate, the same indicators, and the same network. Three representative results are selected and shown in Fig. 3 so that one can easily detect the pattern of their behaviors (the remaining results can be found in S2.3 in the supplementary material).

First, there are two different patterns among all these results. The optimal root rate varies with different sampling rates in the first pattern, whereas it remains the same in the second pattern. We notice that the appearance of these two different patterns is related to whether the original network is connected, as the second pattern only appears when the original network is connected.

Second, the optimal root rate increases with sampling rates in the disconnected networks. The corresponding results with respect to Y2H are shown in Fig. 4, where one can find more detailed information such as the optimal root rate and corresponding GCC sampled under different sampling rates (more results in S3 in the supplementary material). Considering that all the GCCs sampled approach the same value (the GCC in the original network) as the sampling rate approaches 1, we attribute this phenomenon to the fact that different sampling methods have distinct convergence speeds.

Third, one can see that the optimal root rates shift to the right in Fig. 3(a3) compared to (a1). Considering that USP has a much higher link density (1.51×10^{-2}) than TAP (7.26×10^{-3}), we hypothesize that the optimal root rate gets larger faster in the original networks with greater link density.

Then, we try to construct a model based on the random network to help reveal the underlying mechanisms analytically. We show that these phenomena exist even in simple random networks, ensuring that these biases are introduced by the sampling methods themselves rather than the heterogeneity of networks.

First, it is intuitive that the optimal sampling method is BFS if the original network is connected since the subnetwork sampled is also supposed to be connected. However, the optimal sampling method changes to the hybrid method when the original network is disconnected. This is because the subnetworks sampled by RNS usually share low connectivity, whereas those sampled by BFS tend to have higher connectivity. Therefore, the hybrid method has the best performance in disconnected networks because the subnetworks sampled by this method are neither overly nor poorly connected.

Then, we use the random network model[57] to create a disconnected system of two networks for further analysis. Let P_1, P_2 and N_1, N_2 denote the corresponding connection probabilities and numbers of nodes in different random networks, respectively. Without loss of generality, let $N_1 \geqslant N_2$; then, the original GCC is supposed to be $GR = \frac{N_1}{N_1+N_2}$. Note that as we restrict each of the random networks to be connected, the following equation holds:[58]

$$p_i > \frac{\log(N_i)}{N_i}, \quad i = 1, 2. \tag{1}$$

When BFS is used, the initial root node is located in the GCC with a proportion of GR. Let $N = N_1 + N_2$ and let S denote the sampling rate. In this situation, the expected GCC observed in the subnetwork is supposed to be

$$\hat{GCC}_1 = \begin{cases} 1, & S < GR, \\ \frac{N_1}{NS}, & S \geq GR. \end{cases} \tag{2}$$

However, when the initial root node is not located in the GCC, which occurs with probability $1 - GR$, the expected GCC observed in the subnetwork is supposed to be:

$$\hat{GCC}_2 = \begin{cases} 1, & 1 - GR > S, \\ \frac{N_2}{NS}, & 2 - 2GR > S \geq 1 - GR, \\ 1 - \frac{N_2}{NS}, & S \geq 2 - 2GR. \end{cases} \tag{3}$$

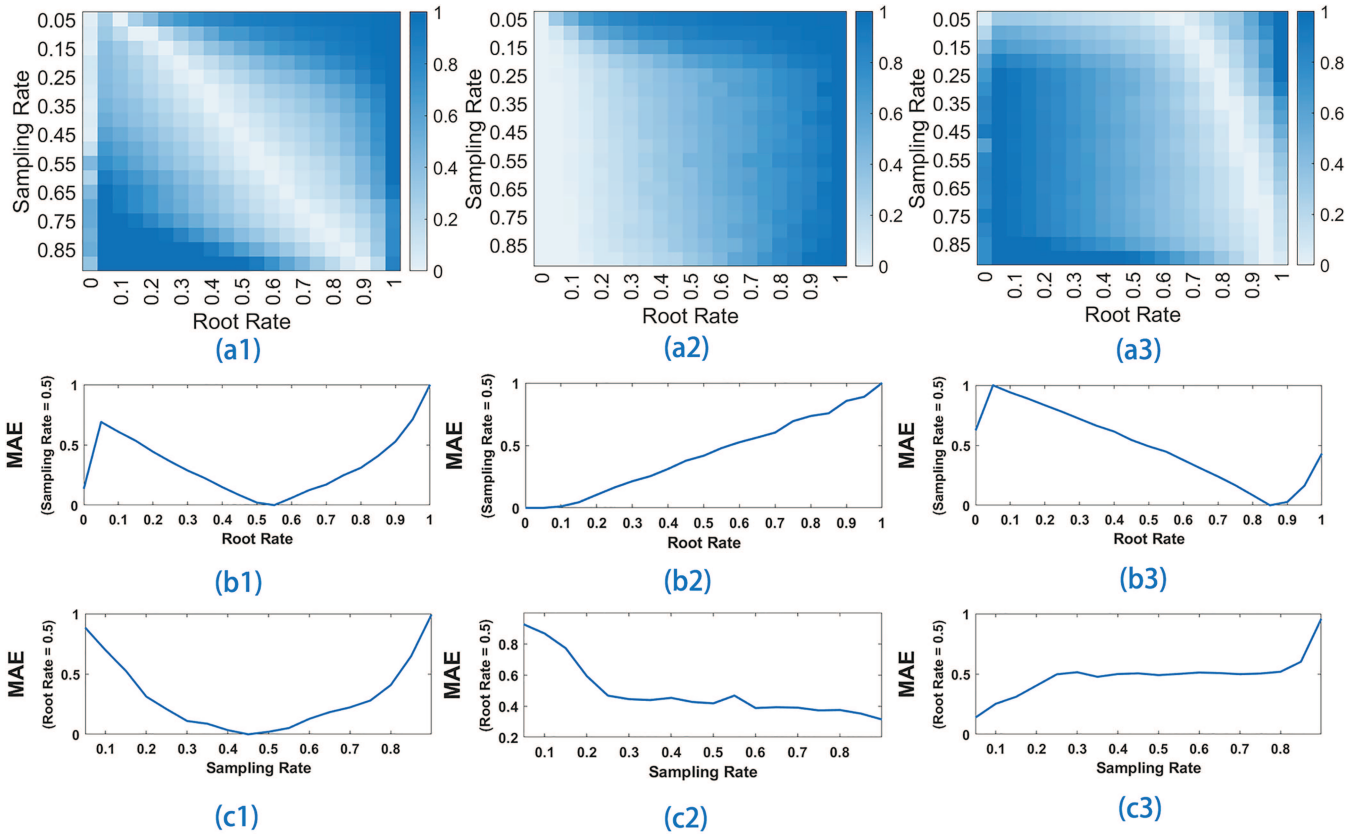

FIG. 3. The mean absolute errors (MAEs) of GCC under different sampling methods. We show the results corresponding to Y2H, TAP, and USP in the graphs that are in the first, second, and third columns, respectively. The heatmaps in the first row show the MAEs of GCC compared to the original networks under the given root rate and sampling rate. We rescale the MAEs under the same sampling rate to highlight the optimal root rates. The lighter the color, the smaller the MAE and the difference of GCC. The graphs in the second row show the curves of the MAEs under root rate $R = 0.5$ and different sampling rates, but the graphs in the third row show those curves under the sampling rate $S = 0.5$ and vary root rates. Note that only the network of TAP (labeled with 2) is connected. The result for Y2H (labeled 1) and USP (labeled 3) share the same pattern while the optimal root rate shifted to the right side due to the larger density of the original network. This result suggests that the optimal root rate will be affected by the GCC and link density of the original network. See more results in S2.3 in the supplementary material.

Now, we can calculate the expected $\hat{GCC}$ when doing BFS as there are only two situations that happened to the initial root node, that is, the root node located in the GCC of the original network or not. Given the sampling rate S, we get the expected GCC according to Eq. (2) if the first situation happens, and according to Eq. (3) if the second situation happens. Combining that the two situations happen with the probability of GR and $1 - GR$, respectively, we can get the expected $\hat{GCC}$ under BFS,

$$\begin{aligned} \hat{GCC} &= \hat{GCC}_1 \times GR + \hat{GCC}_2 \times (1 - GR) \\ &= \begin{cases} 1, & 1 - GR > S, \\ \frac{N_2^2}{N^2 S} + \frac{N_1}{N}, & GR > S \geq 1 - GR, \\ \frac{N_1^2 + N_2^2}{N^2 S}, & 2 - 2GR > S \geq 1 - GR, \\ \frac{N_2}{N} - \frac{N_2^2 - N_1^2}{N^2 S}, & S \geq 2 - 2GR. \end{cases} \end{aligned} \tag{4}$$

In relative terms, the analysis of GCC in the subnetworks under RNS is lightly more difficult, and it is often combined with the study of the percolation process in the networks. One can consider adding the process of sampling to the work where the relationship between the GCC of the random network and link probability is analyzed.[57,59]

Let S, k, and N denote the sampling rate, the average degree, and the number of nodes in the original network, respectively. Let N_G represent the number of nodes in the giant connected component of subnetworks ($N_G \leqslant NS$). The probability that node u is not located in the giant connected component of the subnetwork is

$$1 - \frac{N_G}{NS}. \tag{5}$$

Now, consider the above probability from another angle. The probability that the node u is not located in the GCC can be rewritten as the probability that the node u does not connect to the GCC via any other nodes. First, we focus on one of the nodes other than u in the subnetwork. We record this node as v (one should notice that

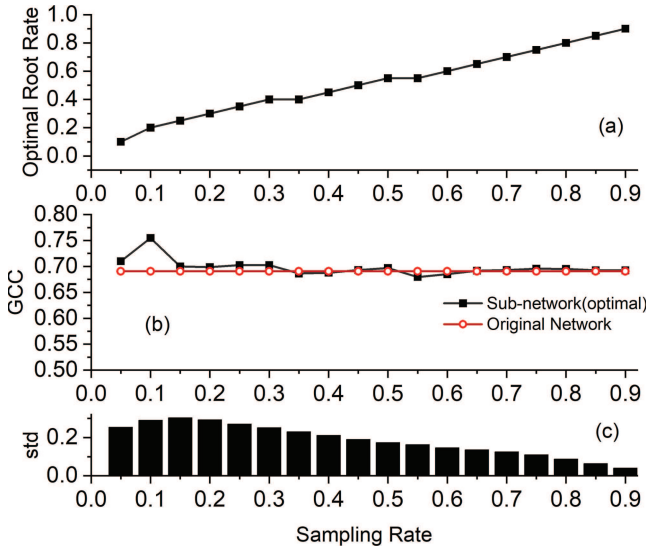

FIG. 4. The detailed results of GCC on Y2H. (a) shows the optimal root rate under different sampling rates, where the curve shows an upward trend as the sampling rate increases. (b) shows the GCC observed in the subnetwork under the corresponding optimal root rate. (c) shows the standard deviation of GCC under the same sampling rate and different root rates, where the overall downward trend suggests that the influence caused by the selection of root rates gradually decreases with the sampling rate. (Note that Y2H is disconnected, whereas in the connected situation, the curve shows another different pattern, which can be found in S3.1 supplementary material.)

there are $NS-1$ such nodes). Then, the situation that u does not connect to the GCC via v can be attributed to two situations:

- Situation 1: u does not connect to v.
- Situation 2: u does connect to v, but v does not connect to the giant connected component of the subnetwork.

For situation 1, the link between u and v does not exist with the probability of $1-\frac{k}{N}$. It is because the expected degree of u is Sk, and the nodes number of the subnetwork is SN, so the above probability can be calculated as $1-\frac{Sk}{SN}=1-\frac{k}{N}$. In addition to that, one can also think that the subnetworks sampled from random networks via RNS are also random networks that share the same link probability. For situation 2, the event that u does link to v happens with the probability of $\frac{k}{N}$ according to the derivation of situation 1. v does not connect to the GCC of the subnetwork with the probability of $1-\frac{N_G}{NS}$ according to Eq. (5). Therefore, situation 2 happens with the probability of $\frac{k}{N}(1-\frac{N_G}{NS})$.

Finally, we can calculate the probability that u does not connect to the GCC of the subnetwork via v as $1-\frac{k}{N}+\frac{k}{N}\left(1-\frac{N_G}{NS}\right)$ with some basic knowledge of combinatorial mathematics. Combining that there are $NS-1$ such nodes v, the probability that u does not connect to the giant connected component of the subnetwork is

$$\left[1-\frac{k}{N}+\frac{k}{N}\left(1-\frac{N_G}{NS}\right)\right]^{NS-1}. \tag{6}$$

As Eqs. (5) and (6) actually describe the same phenomenon, the following equation holds:

$$1-\frac{N_G}{NS}=\left[1-\frac{k}{N}+\frac{k}{N}\left(1-\frac{N_G}{NS}\right)\right]^{NS-1}. \tag{7}$$

With $N_G \geq 1$, the GCC in the subnetwork $\hat{GCC}$ is

$$\begin{aligned} \hat{GCC} &= \max\{N_G, 1\}/N, \\ s.t.\ \ln\left(1-\frac{N_G}{NS}\right) &= \left(\frac{k}{N^2S}-\frac{k}{N}\right)N_G. \end{aligned} \tag{8}$$

Here, equation $\ln(1-x)=-x+o(x^2)$ is used because $\frac{kN_G}{N^2S}$ is infinitesimal when $N\to\infty$.

Given k, N, and S, Eq. (8) can be solved by numerical methods. Here, we present both the results of the analytical solutions and the experimental results of sampling on the corresponding random networks (Fig. 5). The experimental results and the theoretical solutions fit very well. Thus, we can conclude that given the average degree, there will be a process of percolation phase transition in the network with increasing sampling rates. The critical sampling rate S^c satisfies

$$S^c k=1. \tag{9}$$

This conclusion can be found in some other works[57,59] that possibly use different models. Therefore, the critical sampling rate decreases with the original average degree (or link density).

Now, the above-mentioned system obtained by combining two separate random networks can be considered. Using the same symbols, the theoretical $\hat{GCC}$ can be determined by

$$\begin{aligned} \hat{GCC} &= \max\{N_{G1}, N_{G2}, 1\}/N, \\ s.t.\ \ln\left(1-\frac{N_{G1}}{N_1S}\right) &= \left(\frac{k_1}{N_1^2S}-\frac{k_1}{N_1}\right)N_{G1}, \\ \ln\left(1-\frac{N_{G2}}{N_2S}\right) &= \left(\frac{k_2}{N_2^2S}-\frac{k_2}{N_2}\right)N_{G2}. \end{aligned} \tag{10}$$

For convenience, let $N_1=600$, $N_2=400$, and $P_1=P_2=P$; here, we draw the curves of theoretical $\hat{GCC}$ under BFS and RNS (Fig. 6). It can be seen that the $\hat{GCC}$ under RNS approaches the original GCC from below, whereas the $\hat{GCC}$ under BFS approaches the original GCC from above.

Therefore, it is creditable that the $\hat{GCC}$ under the hybrid method will be located between the corresponding curves, although the GCC under the hybrid method seems difficult to derive precisely. Thus, the optimal sampling method that preserves GCC is the hybrid method.

In addition, the convergence behaviors under different methods can be compared using the partial derivative technique. For BFS, at the convergence point $S=1$, there is

$$\left.\frac{\partial \hat{GCC}}{\partial S}\right|_{S=1^-}=\left.\frac{N_2^2-N_1^2}{N^2S^2}\right|_{S=1^-}=\frac{N_2^2-N_1^2}{N^2}, \tag{11}$$

which is a nonzero constant [note that when $N_1=N_2$, the fourth process in Eq. (4) with respect to $S\geq 2-2GR$ does not occur, and $\left.\frac{\partial \hat{GCC}}{\partial S}\right|_{S=1^-}=-0.5$]. For RNS, the sampling rate is also S when only

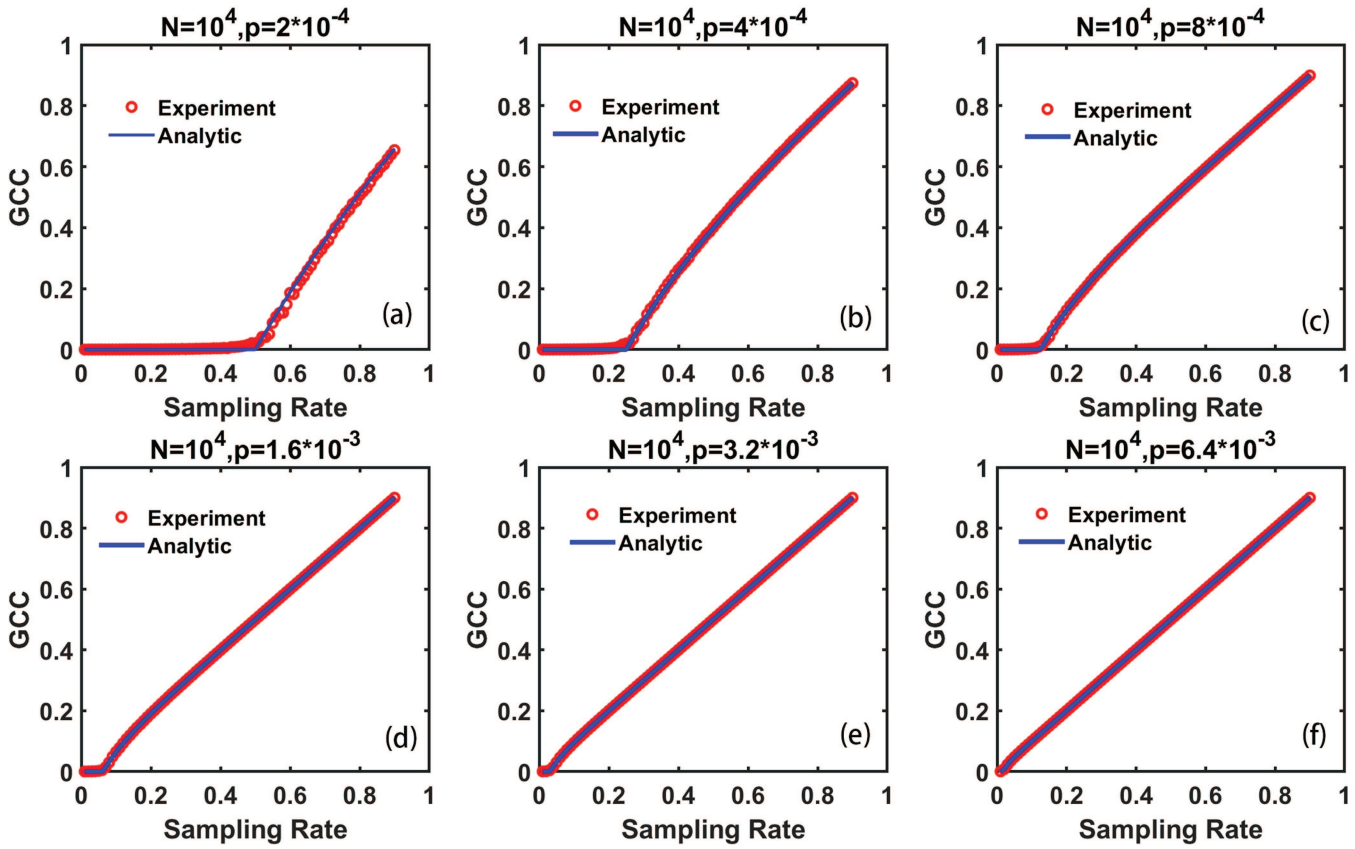

FIG. 5. Analytical solutions [according to Eq. (8), with $k = pN$] and the experimental sampling results of GCC in the subnetworks on random networks under different sampling rates and the given random networks. (a)–(f) show the experimental results under 10 000 nodes but different link probabilities. Ten experiments are performed at each sampling rate to take the average. The results show that the analytical solution fits well with experimental results. One can also find roughly that the percolation phase transition happens at the point S^c where $S^c k = 1$.

one of the random networks is considered in this situation. Using $\hat{G}_1$ to denote the GCC in the random network with N_1 nodes, we have the following equation according to Eq. (8):

$$\ln(1-\hat{G}_1) = \left(\frac{k_1\hat{G}_1}{N_1} - k_1\hat{G}_1 S\right). \tag{12}$$

One can take the partial derivative with respect to S on both sides,

$$\frac{\partial \hat{G}_1}{\partial S} = \frac{-k_1\hat{G}_1}{\frac{1}{\hat{G}_1-1} - \frac{k_1}{N_1} + k_1 S}. \tag{13}$$

Now, the whole system combined by two separate networks can be considered. Substitute $G\hat{C}C = \hat{G}_1 N_1/N$ in Eq. (13) near point $S = 1$, as the corresponding part of $\hat{G}_1$ is supposed to be the giant connected component of the whole system,

$$\left.\frac{\partial G\hat{C}C}{\partial S}\right|_{S=1^-} = \left.\frac{\partial G\hat{C}C}{\partial \hat{G}_1}\right|_{S=1^-}\left.\frac{\partial \hat{G}_1}{\partial S}\right|_{S=1^-} = \left.\frac{N_1}{N}\frac{\partial \hat{G}_1}{\partial S}\right|_{S=1^-} = \frac{-k_1\cdot G\hat{C}C}{\frac{N_1}{G\hat{C}C\cdot N - N_1} - \frac{k_1}{N_1} + k_1 S}. \tag{14}$$

Note that $G\hat{C}C\cdot N = N_1$ when $S \to 1^-$; thus, we have

$$\left.\frac{\partial G\hat{C}C}{\partial S}\right|_{S=1^-} = 0. \tag{15}$$

Comparing Eqs. (11) and (15), we can conclude that there exists a $\delta > 0$ such that the curve of GCC under RNS is flatter (which means higher convergence speed) than that under BFS when $S \in (1-\delta, 1)$. Therefore, the hybrid method shares closer properties with RNS and will have better performance at sufficiently big sampling rates, as both GCCs under RNS and BFS approach the GCC of the original network with an increased sampling rate. Thus, the mechanism

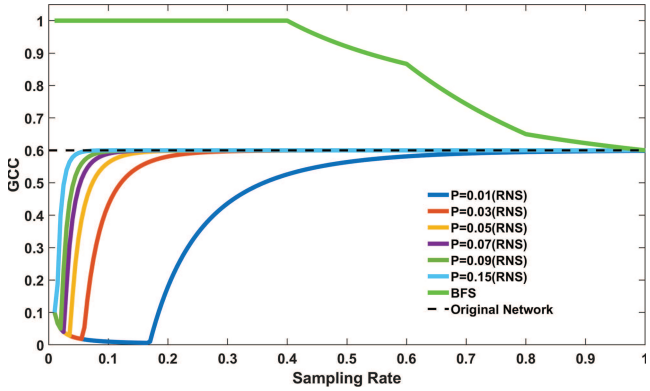

FIG. 6. Given $N_1 = 600$, $N_2 = 400$, and $P_1 = P_2 = P$, we show the analytical $G\hat{C}C$ under BFS and RNS corresponding to Eqs. (4) and (10) (note that the $G\hat{C}C$ under BFS does not change with P). The $G\hat{C}C$ sampled by RNS approaches the original GCC (GCC of original network) from the below, whereas the $G\hat{C}C$ sampled by BFS approaches the original GCC from the above. Therefore, the $G\hat{C}C$ sampled using the hybrid method is supposed to locate between the two curves, suggesting that the hybrid methods should perform the best. Notice that the $G\hat{C}C$ sampled using RNS has relatively good performance near *sampling rate* $= 1$. It is not accidental but determined by the different convergence speeds of these two methods.

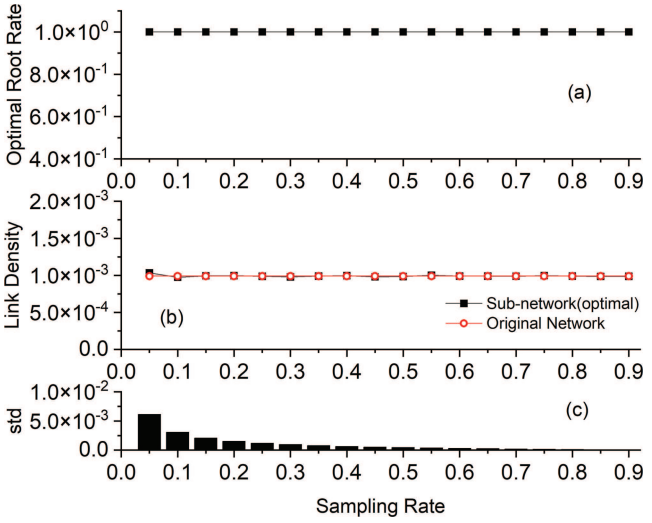

FIG. 7. The detailed results of link density on Y2H. (a) shows the optimal root rate with respect to different sampling rates. Note that the link density in the subnetwork does not change with different sampling rates and performed the best under RNS. (b) shows the link density observed in the subnetwork under the corresponding optimal root rate. (c) shows the standard deviation of link density under the same sampling rate and different root rates. One can compare this figure with Fig. 4 to distinguish the different patterns between GCC and link density. (More experimental results can be found in S3.2 in the supplementary material.)

by which the optimal root rate increases with the sampling rate has been revealed.

Finally, one should note that the distribution of nodes sampled in different connected components does not change with different sampling rates. According to Eq. (9), it is easier for the giant connected components to emerge in the networks with higher link density. However, the $G\hat{C}C$ under BFS has nothing to do with link density in the original network according to Eq. (4), as we have restricted both the number and the size of connected components in the original network. Therefore, the optimal root rate increases faster in denser networks.

These results can be generalized to more common situations where there are several connected components and the link density varies across different components. We can obtain the following conclusions: the GCC of the subnetwork under RNS is mainly influenced by the link density of the original GCC, as the giant connected component more easily emerges with a higher link density. However, the decisive factor of GCC under BFS is the size of the GCC in the original network, as it determines the probability for the initial node to be located in the right place. Therefore, the hybrid method will be affected by both factors. In addition, it is intuitive that the situation in which the original network is connected can be regarded as a special case in which there is only a single connected component.

V. THE IMPACT OF SAMPLING METHODS ON LINK DENSITY

Section IV shows that the optimal sampling method for GCC is always located in the parameter space of the hybrid method or BFS. However, from the experiments, it can be seen that RNS always has the best performance in preserving the link density in subnetworks (Fig. 7). In this section, the properties of link density are discussed. There are lots of works focused on the analysis of the sampled average degree or degree distribution,[20,25,26,28,31] which are very similar to the link density we focused on in this section. However, all of these works focus on random sampling methods, and we provide an analysis under the BFS and the hybrid sampling methods here. Besides that, some previous empirical results are consistent with our conclusion concerning link density.[23]

Specifically, random networks whose nodes share the same degree are used to construct the sampling dynamics so that we can focus on the biases introduced by sampling methods rather than those caused by heterogeneity. We show that the underlying mechanism of bias concerning the link density is different from that of GCC.

Consider the increase in links when a new node is taken, which is related to the size of the sample set: in cases where the size of the sample set is very small, almost no new link will be sampled. However, when the sample set is nearly the same size as the original network, almost all the links connected to that new node will be sampled. Thus, the following equation holds:

$$E^R_{n+1} = E^R_n + \frac{n}{N}k, \tag{16}$$

where E^R_n denotes the link number of the subnetwork under RNS, and n and N denote the node numbers of the subnetwork and the original network, respectively.

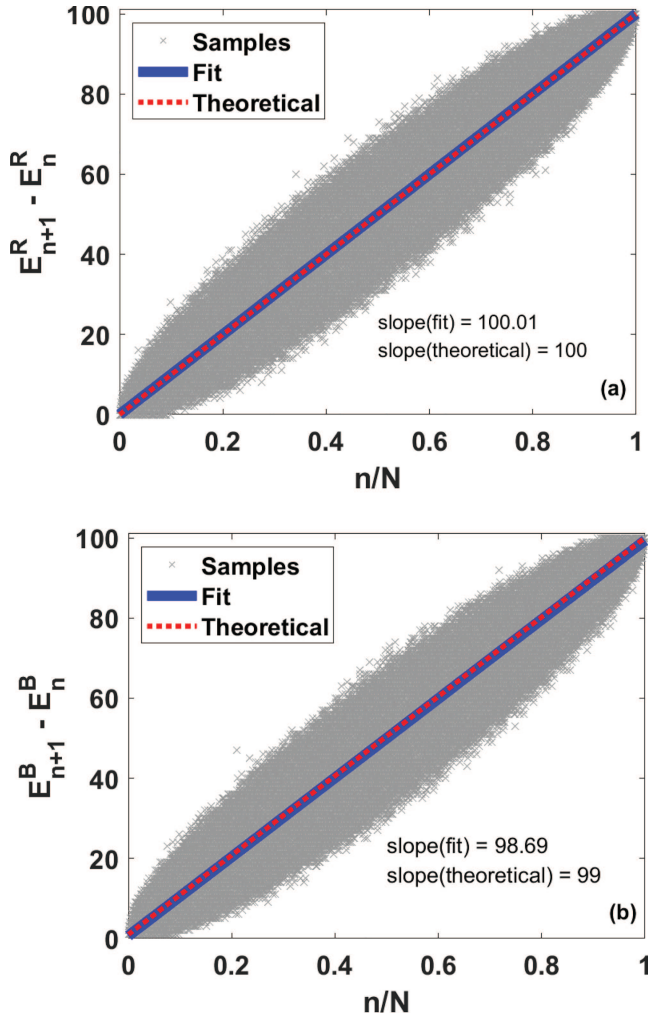

FIG. 8. The numerical results for Eqs. (16) and (17). $N = 5000$, $k = 100$, and every sampling experiment are performed 1000 times. (a) shows the results corresponding to RNS and (b) shows the results corresponding to BFS. The relative deviations between the fitted slope and the theoretical slope are 1.1×10^{-6} under RNS and 3.1×10^{-3} under BFS.

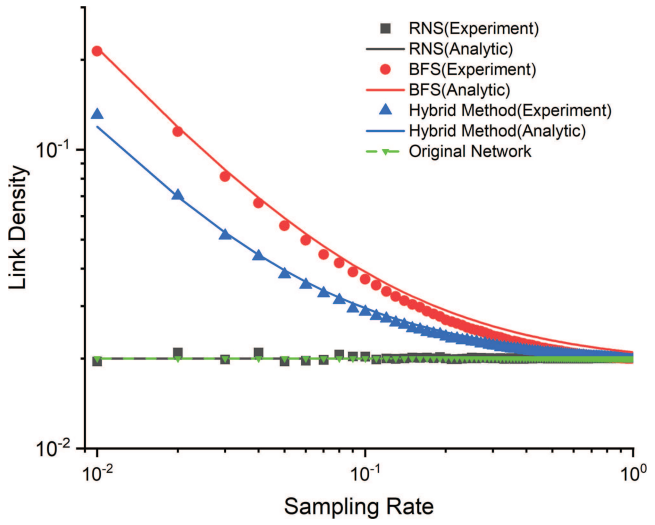

FIG. 9. The experimental results and analytical solution of RNS, BFS, and the hybrid method [according to Eqs. (16)–(18), $R = 0.5$] on the random network ($k = 20$, $N = 1000$). All the corresponding analytical results and experimental results fit well. The results under RNS are always near the link density of original networks. However, the results under BFS and the hybrid method decrease to the original link density with sampling rates. These results explain why RNS always has the best ability to keep the link density of the original network. Fifty repeated experiments are performed to take the average at each sampling rate in the experimental part.

However, in the case of BFS, usually at least one link will be sampled when a new node is acquired because of its neighbor exploring property. We say "usually" because it does not hold when the sampled nodes jump from one connected component to another. Here, we assume that the number of connected components is much smaller than the number of nodes, as is often the case in real networks; then, the above situation can be ignored. We have

$$E^B_{n+1} = E^B_n + 1 + \frac{n}{N}(k-1), \tag{17}$$

where E^B_n denotes the link number of the subnetwork under BFS. The difference between Eqs. (16) and (17) can be attributed to the fact that a part of the original degree is taken outside to guarantee the property of neighbor exploration. In random networks, nodes are homogeneous, which means that Eq. (17) applies not only to BFS but also to other sampling methods based on the property of neighbor exploration.

Equations (16) and (17) can also be proven with some numerical methods under a certain accuracy: all one needs to do is to record the number of new edges $E_{n+1} - E_n$ and the corresponding n/N when a new node is taken under a given random network and certain sampling methods. Then, the method of linear regression can be used to fit $E_{n+1} - E_n$ with n/N. The results can be seen in Fig. 8.

However, as one can see, the deviation between the theoretical value and the real value under RNS is smaller than that under BFS. The deviation, as we think, can be divided into two parts. The first part comes from the scale of networks. The phenomenon of fractional rounding occurs in our model when the scale of the network is finite: the increase in links in the sample set between two steps can be a noninteger, which does not match the real situation. However, one can also easily prove that this effect gradually diminishes as N increases. The second source of deviation is that we assume that the average degree k is far smaller than the scale of networks N. When $k = N - 1$, Eq. (16) still holds; however, Eq. (17) has errors as one can easily figure out that Eq. (16) should be the correct form for BFS at this time.

The root rate R has a physical meaning in this process when the hybrid method is considered, as it denotes the probability of

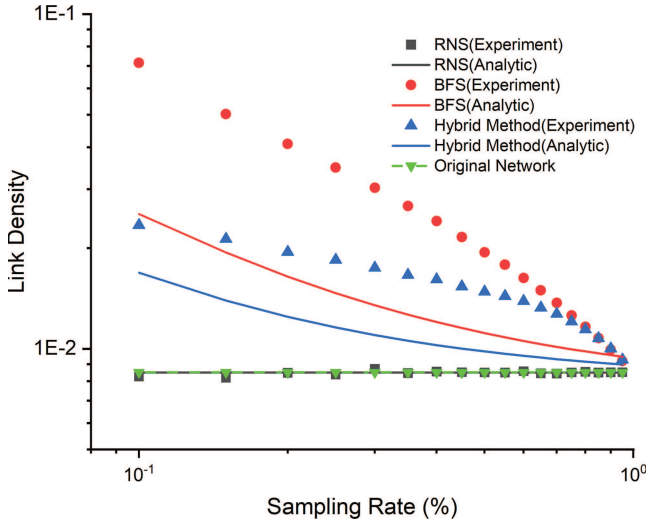

FIG. 10. The analytical solutions [according to Eqs. (16)–(18), where k is replaced with the average degree of the original network] and experimental results of link density with respect to TAP in the subnetworks are shown ($R = 0.5$). Fifty repeated experiments are performed to take the average at each sampling rate for the experimental part. The result shows that though the above analytical solutions do not fit well in real networks, they roughly describe the relations between the different methods. The results corresponding to the other networks can be found in S4 in the supplementary material.

sampling a node with RNS. Thus,

$$E^H_{n+1} = E^H_n + \frac{nk}{N} \times R + \left[1 + \frac{n}{N}(k-1)\right] \times (1-R), \tag{18}$$

where R and E^H_n denote the root rate and the link number under the hybrid method, respectively.

All of these equations can be solved analytically, and one can obtain the results of link density ρ_n with respect to different sample sizes n. Here are the results, and the proof can be found in S5 in the supplementary material,

$$\rho^R_n = \frac{k}{N}, \quad \rho^B_n = \frac{k-1}{N} + \frac{2}{n} \tag{19}$$

and

$$\rho^H_n = \frac{k-1+R}{N} + \frac{2-2\times R}{n}, \tag{20}$$

where ρ^R_n, ρ^B_n, and ρ^H_n denote the link density of the subnetwork under RNS, BFS, and the hybrid method, respectively. Interestingly, the link density under RNS does not depend on n and remains a constant, whereas the link density under BFS monotonically declines to $\frac{k+1}{N}$ as $n \to N$. When *root rate* $\in (0,1)$, the link density under the hybrid method is between those under BFS and RNS.

In addition, the link density ρ^T can be determined as $\rho^T = \frac{k}{N-1}$. One can find that this result is different from both ρ^R_n and ρ^B_N. However, we have $\rho^R_n \sim \rho^B_N \sim \rho^H_N \sim \rho^T$ as $N \to \infty$. These results show that there is a deviation in the theoretical results when N is finite.

The above results prove that the link density of the subnetwork under BFS decreases to the link density of the original network with sampling rates, whereas the link density under RNS is always near the link density of the original network. The link density under the hybrid method shares the same behavior as BFS but is closer to the link density of the original network. When the corresponding experiments are performed (see Fig. 9), we can see that the model fits well with the experimental results.

Due to the network heterogeneity, the above analytical results for BFS and the hybrid method do not hold on real networks if k is replaced with the average degree (Fig. 10; also find the extra similar results in S4 in the supplementary material). We find that the link density obtained by the experiment is usually larger than the analytical solution. Considering the fluctuation of degree and properties such as assortativity of real networks, these biases are caused by the heterogeneity of real networks rather than the sampling methods themselves.

VI. CONCLUSION

In this paper, the properties of subnetworks under distinct sampling methods are discussed. We conclude that the biases of network properties under sampling are mainly caused due to two reasons: the heterogeneity of complex networks and the biases introduced by the sampling methods. The former is mainly affected by the properties of networks, whereas the latter is determined by the sampling methods themselves and is amenable to analysis.

We mainly focus on the effect introduced by the sampling methods as it can be regarded as the systematic error while doing sampling. Specifically, we use both numerical and analytical methods to discuss the biases of network sampling under RNS, BFS, and the hybrid method. Empirically, the hybrid method has the best ability to preserve the properties of the original networks at most times, which may be regarded as the best method when there is no additional information. However, this is not true for all the indicators, as we find that the link density and transitivity, which describe the density of the networks, are better preserved by RNS. Therefore, one should be careful when trying to use subnetworks as substitutes for the original networks.

In addition, there was one phenomenon detected in all our experiments: With the increase in sampling rates, the standard deviations of all these properties concerning the same sampling rate and different root rates show a downward trend (see S6 in the supplementary material). This suggests that the effects caused by the choices of different sampling methods are less significant as the size of the subnetwork increases.

As the main contribution of this article, the properties of GCC and link density are studied to discuss the underlying mechanisms of biases caused by sampling methods themselves. Usually, BFS has the best ability to preserve the GCC of connected networks, whereas the hybrid method performs the best in disconnected networks. In the latter situation, we analytically prove that the biases concerning GCC under BFS are mainly influenced by the size of the GCC in the original networks, whereas GCC under RNS is determined by the link density of the original GCC. With these results, we can

explain the behaviors of the biases introduced by different sampling methods concerning GCC.

In contrast to GCC, the best sampling method preserving link density is always RNS. A process of sampling dynamics is constructed to compare the link density under different sampling methods to reveal the underlying mechanism. The results show that the link density under RNS is always near the link density of the original network regardless of the sampling rates, whereas that under BFS or the hybrid method approaches the link density of the original network with increasing sampling rates. Although this model is not a precise fit in real networks due to network heterogeneity, it accurately describes the relative relationship between the sampling methods. It also proves that the deviation appearing in the process of sampling on real networks is indeed the collective effect of both the network heterogeneity and the biases caused by the sampling methods.

Finally, on the topic of network similarity, our results show that the biases caused by sampling methods are different for distinct network properties. In this context, it is easy to sample two different subnetworks from the same original network where each subnetwork will have at least one topological property closer to the original network than the other. Therefore, it is not advisable to define the similarity of complex networks from a single perspective if the relative importance of the different network properties cannot be determined.

Our work helps us to clarify the sources of biases encountered when trying to gather information from large networks. Some models are constructed to provide baselines of biases caused by sampling methods in specific situations. Based on this work, one may have a deeper understanding of the information sampled from the networks. In this context, these results are of great significance for understanding the large and complex systems in the world surrounding us.

SUPPLEMENTARY MATERIAL

See the supplementary material for the experimental results concerning other indicators, formula derivation process, and other detailed information.

ACKNOWLEDGMENTS

This work was supported by the National Natural Science Foundation of China (NNSFC) (Grant No. 11775020), and we also want to express our gratitude for the helpful discussions among the members of the school.

AUTHOR DECLARATIONS

Conflict of Interest

The authors have no conflicts to disclose.

DATA AVAILABILITY

The data that support the findings of this study are available from the corresponding author upon reasonable request.

REFERENCES

[1]R. Albert and A.-L. Barabási, "Statistical mechanics of complex networks," Rev. Mod. Phys. **74**, 47 (2002).
[2]M. Newman, *Networks* (Oxford University Press, 2018).
[3]D. Chowdhury, L. Santen, and A. Schadschneider, "Statistical physics of vehicular traffic and some related systems," Phys. Rep. **329**, 199–329 (2000).
[4]W. Li and X. Cai, "Statistical analysis of airport network of China," Phys. Rev. E **69**, 046106 (2004).
[5]R. Guimera and L. A. N. Amaral, "Modeling the world-wide airport network," Eur. Phys. J. B **38**, 381–385 (2004).
[6]R. V. Solé, M. Rosas-Casals, B. Corominas-Murtra, and S. Valverde, "Robustness of the European power grids under intentional attack," Phys. Rev. E **77**, 026102 (2008).
[7]G. Zeng, D. Li, S. Guo, L. Gao, Z. Gao, H. E. Stanley, and S. Havlin, "Switch between critical percolation modes in city traffic dynamics," Proc. Natl. Acad. Sci. U.S.A. **116**, 23–28 (2019).
[8]G. Zeng, J. Gao, L. Shekhtman, S. Guo, W. Lv, J. Wu, H. Liu, O. Levy, D. Li, Z. Gao *et al.*, "Multiple metastable network states in urban traffic," Proc. Natl. Acad. Sci. U.S.A. **117**, 17528–17534 (2020).
[9]W. W. Zachary, "An information flow model for conflict and fission in small groups," J. Anthropol. Res. **33**, 452–473 (1977).
[10]M. A. Serrano and M. Boguná, "Topology of the world trade web," Phys. Rev. E **68**, 015101 (2003).
[11]S. P. Borgatti, A. Mehra, D. J. Brass, and G. Labianca, "Network analysis in the social sciences," Science **323**, 892–895 (2009).
[12]S. L. Bressler, "Large-scale cortical networks and cognition," Brain Res. Rev. **20**, 288–304 (1995).
[13]E. Bullmore and O. Sporns, "Complex brain networks: Graph theoretical analysis of structural and functional systems," Nat. Rev. Neurosci. **10**, 186–198 (2009).
[14]J. Gao, B. Barzel, and A.-L. Barabási, "Universal resilience patterns in complex networks," Nature **530**, 307–312 (2016).
[15]M. Serafino, G. Cimini, A. Maritan, A. Rinaldo, S. Suweis, J. R. Banavar, and G. Caldarelli, "True scale-free networks hidden by finite size effects," Proc. Natl. Acad. Sci. U.S.A. **118**, e2013825118 (2021).
[16]E. D. Kolaczyk, *Topics at the Frontier of Statistics and Network Analysis:(re) Visiting the Foundations* (Cambridge University Press, 2017).
[17]A. Clauset and C. Moore, "Accuracy and scaling phenomena in internet mapping," Phys. Rev. Lett. **94**, 018701 (2005).
[18]S. González-Bailón, N. Wang, A. Rivero, J. Borge-Holthoefer, and Y. Moreno, "Assessing the bias in samples of large online networks," Soc. Netw. **38**, 16–27 (2014).
[19]J. Leskovec and C. Faloutsos, "Sampling from large graphs," in *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (ACM, 2006), pp. 631–636.
[20]S. H. Lee, P.-J. Kim, and H. Jeong, "Statistical properties of sampled networks," Phys. Rev. E **73**, 016102 (2006).
[21]M. Kurant, A. Markopoulou, and P. Thiran, "On the bias of BFS (breadth first search)," in *2010 22nd International Teletraffic Congress (lTC 22)* (IEEE, 2010), pp. 1–8.
[22]N. Blagus, L. Šubelj, and M. Bajec, "Assessing the effectiveness of real-world network simplification," Physica A **413**, 134–146 (2014).
[23]N. Blagus, L. Šubelj, and M. Bajec, "Empirical comparison of network sampling: How to choose the most appropriate method?," Physica A **477**, 136–148 (2017).
[24]O. Frank, "A survey of statistical methods for graph analysis," Sociol. Methodol. **12**, 110–155 (1981).
[25]Y. Zhang, E. D. Kolaczyk, B. D. Spencer *et al.*, "Estimating network degree distributions under sampling: An inverse problem, with applications to monitoring social media networks," Ann. Appl. Stat. **9**, 166–199 (2015).
[26]A. Ganguly and E. D. Kolaczyk, "Estimation of vertex degrees in a sampled network," in *2017 51st Asilomar Conference on Signals, Systems, and Computers* (IEEE, 2017), pp. 967–974.
[27]M. P. Stumpf, C. Wiuf, and R. M. May, "Subnets of scale-free networks are not scale-free: Sampling properties of networks," Proc. Natl. Acad. Sci. U.S.A. **102**, 4221–4224 (2005).

[28]M. Al Hasan, "Methods and applications of network sampling," in *Optimization Challenges in Complex, Networked and Risky Systems* (INFORMS, 2016), pp. 115–139.
[29]O. Frank, "Estimation of population totals by use of snowball samples," in *Perspectives on Social Network Research* (Elsevier, 1979), pp. 319–347.
[30]M. E. Newman, "Ego-centered networks and the ripple effect," Soc. Netw. **25**, 83–95 (2003).
[31]M. Gjoka, M. Kurant, C. T. Butts, and A. Markopoulou, "Walking in Facebook: A case study of unbiased sampling of OSNs," in *2010 Proceedings IEEE Infocom* (IEEE, 2010), pp. 1–9.
[32]C. E. Tsourakakis, U. Kang, G. L. Miller, and C. Faloutsos, "Doulion: Counting triangles in massive graphs with a coin," in *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (ACM, 2009), pp. 837–846.
[33]C. Seshadhri, A. Pinar, and T. G. Kolda, "Triadic measures on graphs: The power of wedge sampling," in *Proceedings of the 2013 SIAM International Conference on Data Mining* (SIAM, 2013), pp. 10–18.
[34]M. Rahman and M. A. Hasan, "Sampling triples from restricted networks using MCMC strategy," in *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management* (ACM, 2014), pp. 1519–1528.
[35]N. Kashtan, S. Itzkovitz, R. Milo, and U. Alon, "Efficient sampling algorithm for estimating subgraph concentrations and detecting network motifs," Bioinformatics **20**, 1746–1758 (2004).
[36]S. Wernicke, "Efficient detection of network motifs," IEEE/ACM Trans. Comput. Biol. Bioinf. **3**, 347–359 (2006).
[37]M. Rahman, M. A. Bhuiyan, M. Rahman, and M. Al Hasan, "Guise: A uniform sampler for constructing frequency histogram of graphlets," Knowl. Inf. Syst. **38**, 511–536 (2014).
[38]T. K. Saha and M. Al Hasan, "Finding network motifs using MCMC sampling," in *Complex Networks VI* (Springer, 2015), pp. 13–24.
[39]S. Ghosh, N. Das, T. Gonçalves, P. Quaresma, and M. Kundu, "The journey of graph kernels through two decades," Comput. Sci. Rev. **27**, 88–111 (2018).
[40]H. Bunke, "On a relation between graph edit distance and maximum common subgraph," Pattern Recognit. Lett. **18**, 689–694 (1997).
[41]M.-L. Fernández and G. Valiente, "A graph distance metric combining maximum common subgraph and minimum common supergraph," Pattern Recognit. Lett. **22**, 753–758 (2001).
[42]X. Gao, B. Xiao, D. Tao, and X. Li, "A survey of graph edit distance," Pattern Anal. Appl. **13**, 113–129 (2010).
[43]L. P. Cordella, P. Foggia, C. Sansone, and M. Vento, "A (sub)graph isomorphism algorithm for matching large graphs," IEEE Trans. Pattern Anal. Mach. Intell. **26**, 1367–1372 (2004).
[44]A. Jüttner and P. Madarasi, "Vf2++—An improved subgraph isomorphism algorithm," Discrete Appl. Math. **242**, 69–81 (2018).
[45]V. Carletti, P. Foggia, A. Saggese, and M. Vento, "Challenging the time complexity of exact subgraph isomorphism for huge and dense graphs with VF3," IEEE Trans. Pattern Anal. Mach. Intell. **40**, 804–818 (2018).
[46]T. A. Schieber, L. Carpi, A. Díaz-Guilera, P. M. Pardalos, C. Masoller, and M. G. Ravetti, "Quantification of network structural dissimilarities," Nat. Commun. **8**, 1–10 (2017).
[47]K. Deyasi, A. Chakraborty, and A. Banerjee, "Network similarity and statistical analysis of earthquake seismic data," Physica A **481**, 224–234 (2017).
[48]J. Leskovec, J. Kleinberg, and C. Faloutsos, "Graphs over time: Densification laws, shrinking diameters and possible explanations," in *Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining* (ACM, 2005), pp. 177–187.
[49]H. Jeong, S. P. Mason, A.-L. Barabási, and Z. N. Oltvai, "Lethality and centrality in protein networks," Nature **411**, 41–42 (2001).
[50]A.-C. Gavin, M. Bösche, R. Krause, P. Grandi, M. Marzioch, A. Bauer, J. Schultz, J. M. Rick, A.-M. Michon, C.-M. Cruciat *et al.*, "Functional organization of the yeast proteome by systematic analysis of protein complexes," Nature **415**, 141–147 (2002).
[51]M. E. Newman, "The structure of scientific collaboration networks," Proc. Natl. Acad. Sci. U.S.A. **98**, 404–409 (2001).
[52]R. Guimera, L. Danon, A. Diaz-Guilera, F. Giralt, and A. Arenas, "Self-similar community structure in a network of human interactions," Phys. Rev. E **68**, 065103 (2003).
[53]M. E. Newman, "Finding community structure in networks using the eigenvectors of matrices," Phys. Rev. E **74**, 036104 (2006).
[54]L. A. Adamic and N. Glance, "The political blogosphere and the 2004 us election: Divided they blog," in *Proceedings of the 3rd International Workshop on Link Discovery* (ACM, 2005), pp. 36–43.
[55]O. Frank, "Estimation of the number of connected components in a graph by using a sampled subgraph," Scand. J. Stat. **5**, 177–188 (1978); available at https://www.jstor.org/stable/4615713.
[56]J. M. Klusowski and Y. Wu, "Estimating the number of connected components in a graph via subgraph sampling," Bernoulli **26**, 1635–1664 (2020).
[57]P. Erdos and A. Renyi, "On random graphs I," Publ. Math. **4**, 3286–3291 (1959); available at http://snap.stanford.edu/class/cs224w-readings/erdos59random.pdf.
[58]A.-L. Barabási, "Network science," Philos. Trans. R. Soc. A: Math. Phys. Eng. Sci. **371**, 20120375 (2013).
[59]J. Gao, S. V. Buldyrev, S. Havlin, and H. E. Stanley, "Robustness of a network of networks," Phys. Rev. Lett. **107**, 195701 (2011).