

RESEARCH

Open Access



Sampling on networks: estimating spectral centrality measures and their impact in evaluating other relevant network measures

Nicolò Ruggeri* and Caterina De Bacco

*Correspondence:
nicolo.ruggeri@tuebingen.
mpg.de
Max Planck Institute
for Intelligent Systems,
Max-Planck Ring 4,
72076 Tübingen, Germany

Abstract

We perform an extensive analysis of how sampling impacts the estimate of several relevant network measures. In particular, we focus on how a sampling strategy optimized to recover a particular spectral centrality measure impacts other topological quantities. Our goal is on one hand to extend the analysis of the behavior of TCEC (Ruggeri and De Bacco, in: Cherifi, Gaito, Mendes, Moro, Rocha (eds) *Complex networks and their applications VIII*, Springer, Cham, pp 90–101, 2020), a theoretically-grounded sampling method for eigenvector centrality estimation. On the other hand, to demonstrate more broadly how sampling can impact the estimation of relevant network properties like centrality measures different than the one aimed at optimizing, community structure and node attribute distribution. In addition, we analyze sampling behaviors in various instances of network generative models. Finally, we adapt the theoretical framework behind TCEC for the case of PageRank centrality and propose a sampling algorithm aimed at optimizing its estimation. We show that, while the theoretical derivation can be suitably adapted to cover this case, the resulting algorithm suffers of a high computational complexity that requires further approximations compared to the eigenvector centrality case. Main contributions (a) Extensive empirical analysis of the impact of the TCEC sampling method (optimized for eigenvector centrality recovery) on different centrality measures, community structure, node attributes and statistics related to specific network generative models; (b) extending TCEC to optimize PageRank estimation.

Keywords: Sampling on network, Eigenvector centrality, PageRank, Centrality measures

Introduction

When investigating real-world network datasets we often do not have access to the entire network information. This is the case of large datasets, having limited storage capacity or limited resources during the data collection phase. Nevertheless, this should not prevent practitioners from analyzing an available network sample. In fact, evaluating network properties while accessing only a smaller sample is a relevant problem in various fields, ranging from modelling dynamical processes (De Choudhury et al. 2010; Sadikov et al. 2011), network statistics estimation (Leskovec and Faloutsos 2006), data compression (Adler and Mitzenmacher 2001) and survey design (Frank 2005). Imagining

that one could design the sampling scheme for data collection, then this should be done wisely, as this biases the estimates of the network properties aimed at investigating (Han et al. 2005; Lee et al. 2006; Kossinets 2006). The goal should be to design a sampling protocol that not only preserves the relevant network properties of the entire topology inside the sample, but that can be implemented efficiently. Most sampling strategies found in the literature (Leskovec and Faloutsos 2006) are empirically-driven and lack theoretical groundings. Recently, TCEC (Ruggeri and De Bacco 2020), a sampling algorithm to approximate in-sample eigenvector centrality (Bonacich 1972), whose main features are being theoretically grounded and computationally scalable, has been proposed. TCEC aims at preserving the relative eigenvector centrality ranking of nodes inside the sample. This is a centrality measure used in many disciplines to characterize the importance of nodes. However, this might not be the only property of interest when studying a network. The question is then how a sampling method, optimized to retrieve one particular property, performs in estimating other network-related measures. In this work we address this question by performing an extensive analysis of the behavior of TCEC in recovering several relevant network properties by means of empirical results on real and synthetic networks. In particular, we focus on estimating various centrality measures which have a very different characterization from eigenvector centrality and do not come from spectral methods. Then we investigate how community structure and covariate information are affected by the sampling. In addition, we analyze how sampling strategies behave in recovering relevant network statistics specific to various network generative models. We compare performance with other sampling strategies. Finally, we discuss what are the challenges preventing a trivial extension of TCEC on PageRank (Brin and Page 1998) score.

Related work

A large part of the scientific literature aiming at investigating sampling strategies on networks is based on empirical approaches (Blagus et al. 2017; Costenbader and Valente 2003) and focus on recovering standard topological properties like degree distribution, diameter or clustering coefficient (Leskovec and Faloutsos 2006; Morstatter et al. 2013; Stutzbach et al. 2009; Hübler et al. 2008; Stumpf and Wiuf 2005; Ganguly and Kolaczyk 2018; Antunes et al. 2018). To the best of our knowledge, TCEC sampling (Ruggeri and De Bacco 2020) is one of the first theoretical attempts in estimating eigenvalue centrality, which goes beyond heuristics or empirical reasoning. A closely related problem is that of estimating eigenvector centrality without observing any edge but only signals on nodes (Roddenberry and Segarra 2019; He and Wai 2020). A different but related research direction is to question the stability of centrality measures under perturbations (Segarra and Ribeiro 2015; Han and Lee 2016; Murai and Yoshida 2019). In the case of PageRank score, and more recently for Katz centrality as well (Lin et al. 2019), the focus of similar lines of research is based on the different objective of estimating single nodes' scores or approximating the external information missing for reliable within-sample estimation (Sakakura et al. 2014; Chen et al. 2004; Davis and Dhillon 2006), rather than estimating the relative ranking of nodes within a sample as we do here. Finally, focusing on temporal networks, Shao et al. (2017) propose a centrality measure suitable for this case and a method for its estimation using the network dynamics.

TCEC: sampling for eigenvector centrality estimation

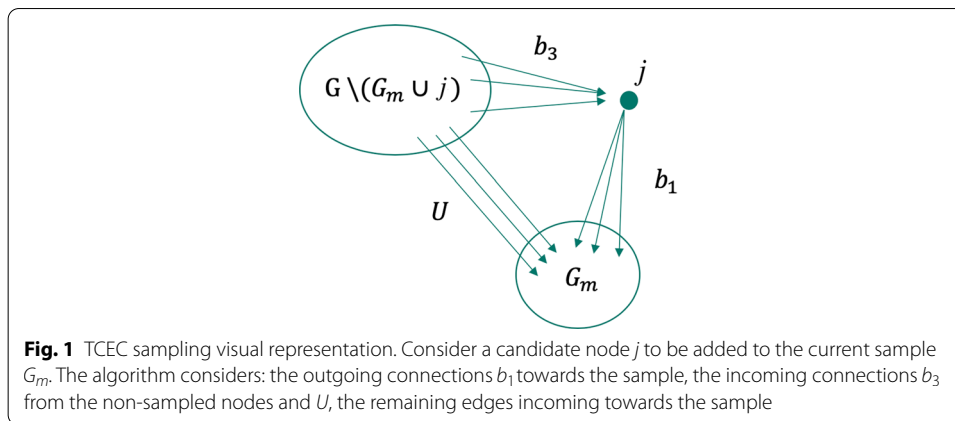
In this section we introduce the formalism and explain the main ideas behind the Theoretical Criterion for Eigenvector Centrality (TCEC) sampling algorithm (Ruggeri and De Bacco 2020). This method uses mathematical formalism from spectral approximation theory to approximate the eigenvector centralities of nodes in a subsample with their values in the whole graph. Consider a graph $G = (\mathcal{V}, \mathcal{E})$ where \mathcal{V} is the set of nodes and \mathcal{E} the set of edges; denote A its adjacency matrix with entries $A_{ij} \in \mathcal{R}_{\geq 0}$ the weight of an edge from i to j . Sampling a network can be defined as the problem of selecting a *principal submatrix* A_m' of size $m \leq |\mathcal{V}|$ induced by a subset of nodes $\mathcal{I} \subseteq \mathcal{V}$. The subsampled network is denoted as $G_m = (\mathcal{I}, \mathcal{E}_m)$, and $\mathcal{E}_m \subseteq \mathcal{E}$ is the set of edges in the subsample. In general, there can be several choices for selecting G_m . They should depend on the quantities aimed at preserving when sampling. TCEC selects G_m in order to minimize the sin distance $\sin(\mu_m, \tilde{\mu})$ between the eigenvector centrality $\tilde{\mu} \in \mathcal{R}^m$ in the subsample and the one on the same nodes, but calculated from the whole graph $\mu_m \in \mathcal{R}^m$; μ_m is a vector built from the whole-graph eigenvector centrality $\mu \in \mathcal{R}^V$, when selecting only the m entries corresponding to nodes in the subsample. Accessing $\sin(\mu_m, \tilde{\mu})$ without the knowledge of the whole graph is not possible. However, given that eigenvector centrality is a spectral method, i.e. is based on evaluating eigenvectors and eigenvalues, TCEC uses projection methods for spectral approximation to propose a bound on that distance and relate it to network-related quantities. This results in an algorithmic implementation of a sampling procedure that aims at minimizing that bound. Referring to Ruggeri and De Bacco (2020) for details, the algorithm briefly works as follows. Starting from an initial small random sample, it selects nodes in an *online* fashion: it adds to the current sample \mathcal{I} of size $k - 1$ one node at a time by selecting the best node from the set of non-sampled nodes $j \in \mathcal{V} \setminus \mathcal{I}$. The best candidate node j is the one that maximizes the following quantity made of network-related quantities:

$$(1 - \alpha) \left(\|b_1\|_2^2 + \|b_1^T U\|_2^2 - \|b_3\|_2^2 \right) + \alpha d_{in}^{G_k}(j), \quad (1)$$

where $b_1 \in \mathbb{R}^{k-1}$ are the edges pointing from j to the nodes already in the subsample, $b_2 \in \mathbb{R}$ is the entry corresponding to j , $b_3 \in \mathbb{R}^{n-k}$ are edges from nodes outside the sample towards j , $U \in \mathbb{R}^{k-1, n-k}$ are the edges from nodes outside the sample towards nodes in it, j excluded; $d_{in}^{G_k}(j)$ is the (weighted) in-degree of node j calculated considering only the incoming edges from nodes that are in the sample; $\alpha \in [0, 1]$ is an hyperparameter that can be tuned empirically. We present a diagram of the quantities involved in Fig. 1.

Empirical studies

We study the impact of sampling a network with TCEC on several relevant network properties different from eigenvector centrality. Namely, we investigate: (1) the distribution of the sampled nodes in terms of non-spectral centrality measures as *in-degree*, betweenness centrality and SpringRank (De Bacco et al. 2018); (2) the relationship between community structure and sampled nodes; (3) the preservation of the distribution of node attributes in the sampled network; (4) the impact of sampling on other model-specific statistics. For all these tasks, we compare with uniform random walk



sampling (RW), as this is the mainstream choice for many sampling scenarios, due to its favorable statistical and computational properties (Gjoka et al. 2010); it has also shown better performance in recovering eigenvector centrality than all other state-of-the-art algorithms analyzed against TCEC (Ruggeri and De Bacco 2020). In addition, in the absence of a best sampling protocol that works for all applications, we further show comparisons with various other algorithms; for sake of visualization, we move some of the results to the corresponding Appendices. In the following experiments we use the Kendall- τ correlation (Kendall 1990) to assess similarity between score vectors, as done in Ruggeri and De Bacco (2020).

Implementation details

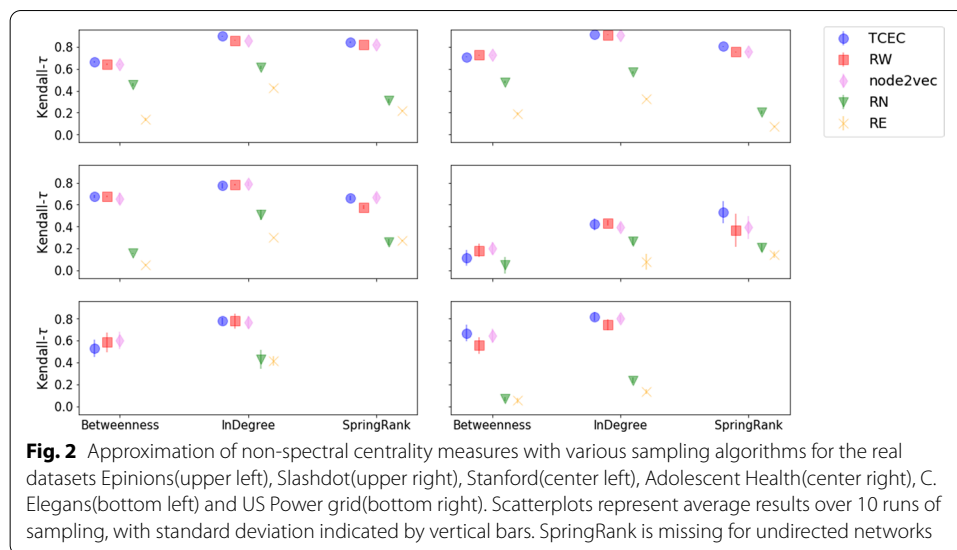
While we refer to Ruggeri and De Bacco (2020) for the detailed definitions of the parameters needed in the algorithmic implementation, we provide a summary of their values used in our experiments in the “Appendix 1”; we use the open-source implementation of TCEC available online.¹ In all the following experiments we sample to include 10% of the total nodes. We comment on the stability and effects of different samples sizes in “Appendix 2” by means of an empirical comparison.

In addition to RW, for performance comparison we compare with several commonly employed sampling algorithms. Namely, uniform sampling on nodes (RN) (Leskovec and Faloutsos 2006; Ahmed et al. 2012; Wagner et al. 2017), uniform sampling on edges (RE), node2vec (Grover and Leskovec 2016) (with exploration parameters $p = 2$, $q = 0.5$, i.e. depth-first oriented search) and snowball expansion sampling (EXP) (Maiya and Berger-Wolf 2010).

Non-spectral centrality measures behavior

We analyzed the performance of TCEC in estimating non-spectral centrality measures in various real world datasets (see “Appendix 1” for more details): C. Elegans, the neural network of the nematode worm *C. elegans* (Watts and Strogatz a); US Power grid, an electrical power grid of the western US (Watts and Strogatz a); Epinions, a

¹ https://github.com/cdebacco/tcec_sampling.



who-trusts-whom network based on the review site *Epinions.com* (Takac and Zabovsky 2012); Slashdot, a social network based on the reviews website *Slashdot.org* community (Leskovec et al. 2009); Stanford, a network of hyperlinks of the *stanford.edu* domain (Leskovec et al. 2009); Adolescent Health, a network of friendship between schoolmates (Moody 2001). Together these networks cover different domains (transportation, social, biological, communication), directed and undirected topologies, sizes (from order of 10^2 to 10^5 nodes) and sparsity levels (from average degree of 2.67 to 20.29). We use all algorithms listed above with the exception of EXP sampling, since it has already proven to perform poorly in eigenvector centrality approximation (Ruggeri and De Bacco 2020) and is computational too slow to be deployed on networks of the sizes considered here; in “Appendix 3” we show the computational efficiency of the various algorithms.

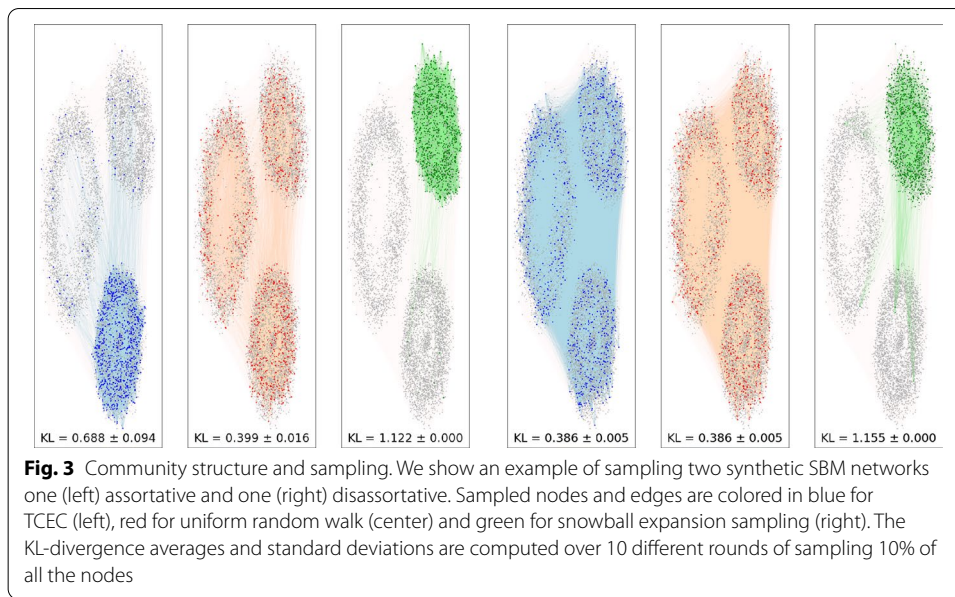
We consider three different centrality measures: (1) in-degree centrality, which corresponds to the in-degree of a node; (2) betweenness centrality, a measure that captures the importance of a node in terms of the number of shortest paths that need to pass through it in order to traverse the network; (3) SpringRank (De Bacco et al. 2018), a physics-inspired probabilistic method to rank nodes from directed interactions which yields rank distributions relatively different than that of spectral measures, like eigenvector centrality. Together, these three provide a diverse set of methods to characterize a node’s importance. Notably, none of these are based on spectral methods, as opposed to the theoretical grounding behind TCEC.

As we show in Fig. 2, all the considered non-spectral measures are well approximated by RW-like algorithms and TCEC on all datasets, with TCEC performing slightly better on average. A big gap can be observed instead with respect to uniform random sampling (RN and RE). We argue that this might be caused by the loss of discriminative edges when not taking into account the topology at sampling time, thus resulting in poor performance in recovering any edge-based centrality measure. For the sake of completeness, we include similar plots for the estimation of two spectral centrality measures, PageRank (Brin and Page 1998) and Katz centrality (Katz 1953) in “Appendix 4”. Finally, while it is difficult to model analytically how these results depend on the sample size, empirically

we see that performance increases as the sample grows, similarly to what observed for recovering eigenvector centrality in Ruggeri and De Bacco (2020). The magnitude of this increment depends on what measure is being tracked and the specific dataset, however the relative performance of the various sampling strategies seems to be coherent with the results reported here in Fig. 2, see “Appendix 2”. As a final consideration, one may wonder how much of the approximation capabilities of TCEC with respect to non-spectral centralities, such as in-degree, could be explained by correlations with the eigenvector centrality itself. In fact, one would expect RN sampling to perform best at least in the approximation of nodes’ degrees. We remark the following facts. First, in the datasets considered above, the correlation between eigenvector centrality and in-degree varies from high values, 0.84 and 0.80 for the Epinions and Slashdot datasets respectively, to values as low as 0.03 for the Stanford dataset, see “Appendix 5”, while Kendall’s correlation is quite high in both these extreme cases. It is therefore not possible to directly attribute the good approximation of degree centrality to the correlation between the latter and eigenvector centrality. We conjecture that this is rather due to the character of sampling methods based on moving between adjacent nodes, like RW, TCEC and node-2vec, which tend to preserve connections between nodes and therefore degrees. In fact, uniform sampling strategies (which do not move between adjacent nodes) as RN and RE, perform significantly worse and consistently across datasets and measures. In addition, in the presence of tight-knit communities, the former algorithms are more likely to sample the majority of a node’s neighborhood, without getting lost in other regions of the graph, as we show in the Section 3.3.

Community structure preservation

We investigate how the sampling algorithms impact a network’s underlying community structure. To this end, we study the distribution of the community memberships of sampled nodes in synthetic networks generated with Stochastic Block Model (SBM) (Holland et al. 1983) of size $N = 10^4$ nodes divided in 3 communities. Sampling protocols can be sensitive to the topological structure of the network (assortative or homophilic, disassortative or heterophilic) and to the balance of group sizes (Wagner et al. 2017). These can all impact how the different groups are represented in the sample and other factors such as individuals’ perception biases (Lee et al. 2019). We thus run tests on both types of structures and using various levels of balance for the communities. Specifically, we consider (1) balanced assortative networks: two groups of 3000 nodes and one of 4000, within-block probability of connection $p_{in} = 0.05$ and between-blocks $p_{out} = 0.005$; (2) unbalanced assortative networks: groups of sizes 1000, 3000 and 6000 respectively, same p_{in} and p_{out} as in (1); (3) balanced disassortative networks: same group division as in (3) but within-block probability of connection $p_{in} = 0.005$ and between-blocks $p_{out} = 0.05$. We compare TCEC with RW, which was shown to be robust in representing groups in the sample (Wagner et al. 2017) and EXP sampling, since it has been explicitly built to sample community structure. All algorithms start sampling from a node belonging to the group of smallest size. We observe two qualitatively different trends in the way nodes are chosen. RW yields samples of nodes more homogeneously distributed across communities, in all network structures. TCEC, instead, tends to select nodes within the block where it has been initialized. A possible

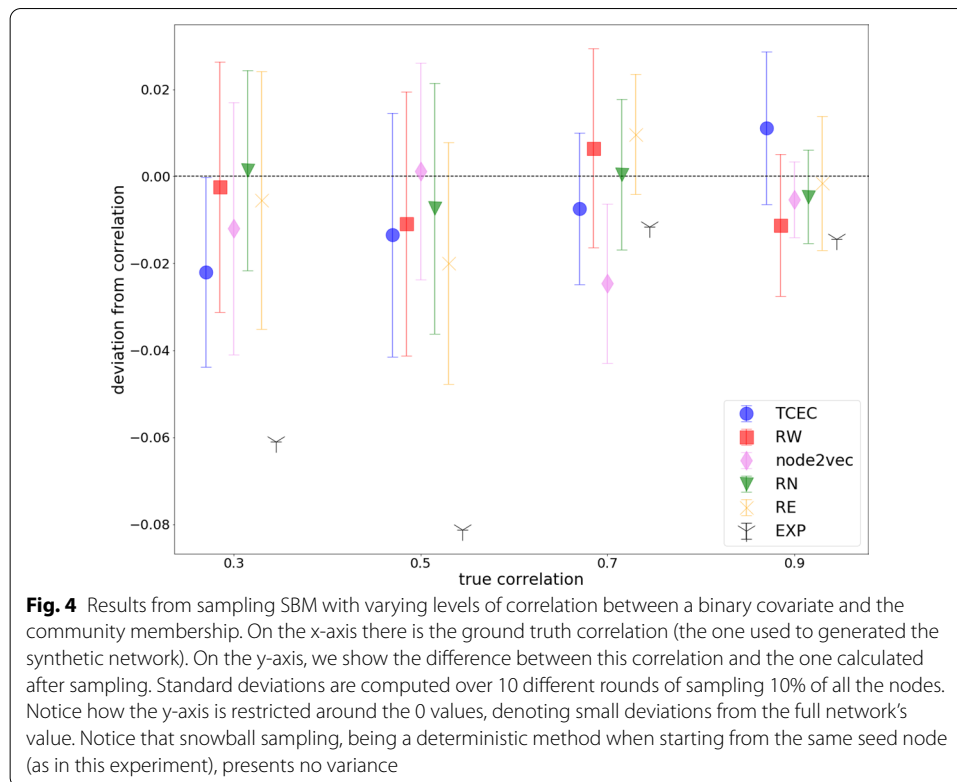


explanation for this behavior is given by the peculiar form of the TCEC score of Eq. (1). In general, the algorithm tends to select nodes with a large $\|b_1\|_2$ and small $\|b_3\|_2$, i.e. many connections towards the sample and few connections from outside the sample. A likely choice is to then select nodes within the same community, where this combination holds. However, for the standard SBM case one can show (see “Appendix 9”) that the contribution of these two terms cancels out, thus the term that matters is $b_1^T U$. This represents the number of common neighbors that are in the sample between j and any other $\ell \in G' \setminus \{j\}$. Assuming, for simplicity, homogenous communities, in the assortative case it is reasonable to assume that the initial random walk biases the sample to have more nodes of color as the initial seed, say r . With this initial bias, one can show that, on average, nodes j of color r have higher score, because they have higher values of $\|b_1^T U\|_2$ than other $\ell \in G'$ not of color r . Instead, for the disassortative case, we have the opposite result of nodes of color different than r having higher score. In this case, the sampling dynamics keep jumping to nodes of different colors, hence the sample is more balanced. One can generalize similar arguments to more complex topologies, for instance degree-corrected SBM, where nodes’ degree also contribute to their scores (not only in the case $\alpha > 0$). However, the theoretical generalization becomes more cumbersome as we add more details to the model. With the same reasoning we recover the Erdős-Rényi case by assuming a connection probability equal for all edges, i.e. independent of the nodes’ colors. In this case, one can show that the only term that matters is the (weighted) in-degree $d_{in}^{G_k}(j)$, if we assume $\alpha > 0$. This matches with the intuition that, given the uniform edge probability, in an Erdős-Rényi network all nodes are statistically identical, hence there should be no best candidate to distinguish, unless we look at a specific realization where random fluctuations dominate. These graphs may not be appropriate to evaluate the performance of a centrality measure, as all nodes are concentrated around the same score of centrality. Finally, expansion sampling remains confined in a single block, as it is a deterministic algorithm. Results are presented in

Fig. 3, where we also report the KL-divergence (Kullback and Leibler 1951) between the communities distribution in the sample and the whole network for all sampling algorithms (the communities are the known ground-truth used for the SBM synthetic generation). The KL-divergence is a measure of discrepancy between probability distributions, which is 0 if they perfectly overlap, and gets larger as the difference between them grows. Thus, higher values signal higher discrepancy between the in-sample block distribution and the one calculated on the entire network. To account for different number of nodes in the whole network and the sample, here we consider the frequency of each group as community distribution. This can be observed graphically in Fig. 3 (left) for the assortative homogeneous structure i). Here the higher KL divergence is due to a more pronounced clustering of sampled nodes in one single block. The nodes selected by RW are more scattered around different blocks, while TCEC tends to select nodes within a single block and expansion sampling is completely confined to the initial one. Similar results hold for case ii), as defined above, and are presented in “Appendix 6”. For the disassortative structure iii), however, results differ. In this case, TCEC and RW tend to explore the network in a similar manner. A lower KL-divergence from the ground truth signals the fact that blocks are sampled more uniformly. While for RW this phenomenon is explained by the stochasticity of the neighborhood exploration, for TCEC it is caused by the way the algorithm works in selecting candidate nodes with high out-degrees towards the sample but small in-degrees from outside of it, as shown in Fig. 1. In disassortative networks these likely candidates belong to different communities, thus the more homogeneous exploration. Expansion sampling is still confined inside the starting block as in the previous case.

Node attribute preservation

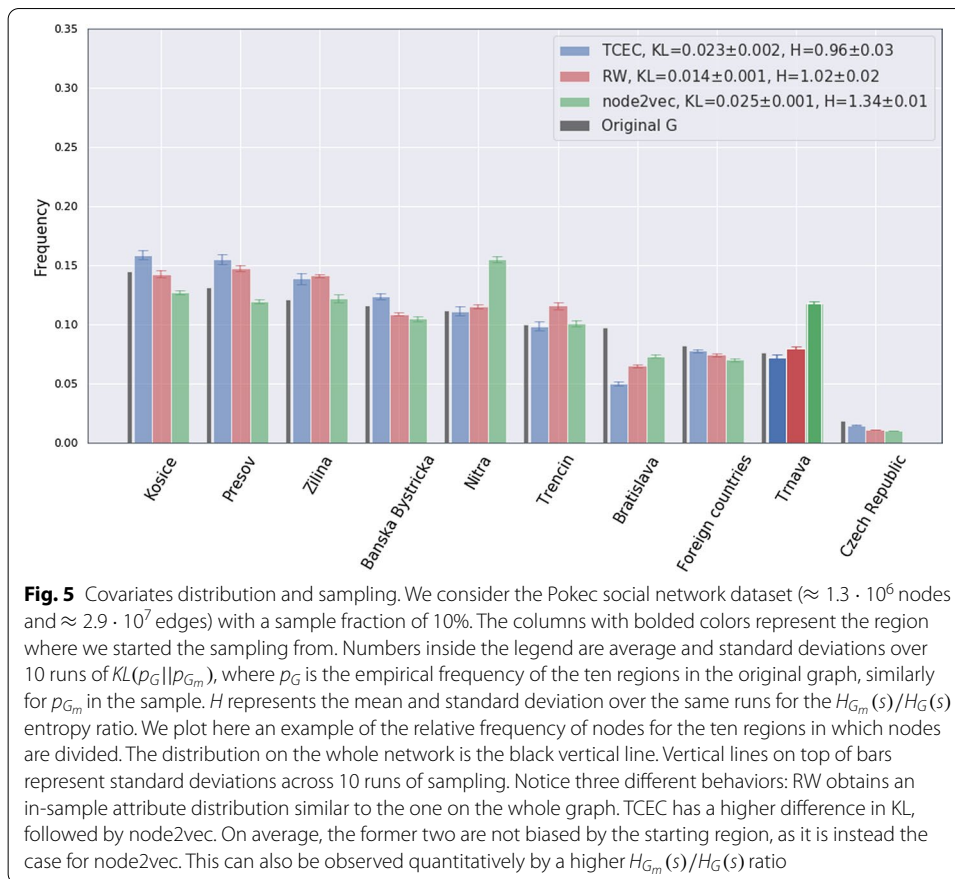
Another relevant question is whether node attributes are affected by the way the network is sampled. This is particularly important in cases where extra information is known, along with the network’s topological structure. For instance, in relational classification, network information is exploited to label individuals (e.g. recovering nodes’ attributes); classification performance can significantly change based on the sampling protocol adopted (Ahmed et al. 2012; Espín-Noboa et al. 2018). Here we assume that attribute information is used a posteriori to analyze the results, but not taken as input to bias sampling. In general, when performing statistical tests on sampled networks’ covariates, we work under the assumption that their distribution is similar to that of the original network. However, this assumption is not necessarily fulfilled when performing arbitrary sampling. Notice that this is a related but different problem than the one above of community structure preservation. In that case, we were explicitly imposing that communities were correlated with network structure. If attributes well correlate with communities, then we should see a similar pattern with the results of the previous section. We test this on synthetic networks with both communities and node attributes, where a parameter tunes their correlations. Namely, we use a 2-block SBM generative model with varying levels of correlation between the community and a binary covariate (Contisciani et al. 2020). We then measure the correlation between community and covariate after sampling. Results are presented in Fig. 4. As can be observed, the original correlation is preserved in the samples by most algorithms, as the relative deviation from the



real value (y-axis) is low with respect to the true correlation (x-axis), suggesting a regular behavior of covariates in the samples, despite not being taken into account as input for the sampling algorithm. For TCEC, the performance seems to increase with correlation values, while for the other algorithms we cannot distinguish a clear monotonic pattern.

But if the community-attribute correlation is not there or if this is not trivial, then we might see something different if we only look at attributes (e.g. if a practitioner is not interested in community detection). In this case, we can only assume correlation with network structure, but this may not be valid depending on the real dataset at hand. We test this behavior by studying the Pokecdataset (Takac and Zabovsky 2012). This is a social network representing connections between people in form of online friendships. In addition, the dataset contains extra covariate information on nodes, i.e. attributes about the individuals. In our case we focus on one of them, the geo-localization of users in one of the ten regions (the eight Slovakian regions, Czech Republic and one label for all other foreign countries) where the social network is based. We compare the distribution of this covariate in the full network with that on the nodes sampled by RW, TCEC and node2vec, with exploration parameters $p = 2$, $q = 0.5$, i.e. depth-first oriented search. We omit results from other sampling algorithms, to focus on the interpretation of these relevant cases. The choice of node2vec is motivated by its frequent implementation for node embedding tasks.² As node embeddings are often used for

² Note that node2vec is not explicitly using any covariate in input. Rather, it infers embeddings on nodes based on the observed network; these are then often used as inferred “feature” vectors to be subsequently given in input for machine learning tasks, e.g. classification.



regression or classification tasks, along with network covariates, it is thus relevant for our task here. We run the algorithms starting from seed nodes within different regions, as the choice of the initial sample of labeled seed nodes can impact the final in-sample attribute distribution (Wagner et al. 2017). As before, we measure KL-divergence between the empirical attribute distribution on the entire network against that found within the sample. A graphical representation of one example of the results is given in Fig. 5. We notice different behaviors for the various sampling methods. While all algorithms recover a covariate distribution close the ground truth, slightly better performances are achieved, in order, from RW, TCEC and node2vec, with average KL values ranging from 0.01 to 0.04 respectively. However, a peculiar trend can be observed in relation to the starting region. In fact, the final sample is biased towards over representing the seed region for node2vec, as opposed to a comparable homogeneity obtained by TCEC and RW. This is a subtle result, as this over representation is not shown by the KL values. Instead, it can be measured by the entropy ratio $H_{G_m}(s)/H_G(s)$ between the entropy $H_{G_m}(s) = -p_{G_m}(s) \log p_{G_m}(s) - (1 - p_{G_m}(s)) \log(1 - p_{G_m}(s))$ of a binary random variable representing whether a node in the sample belongs to the seed region s or not, over $H_G(s)$, the same quantity but calculated over all nodes in the graph. In words, this measures the discrepancy of the frequency of the particular attribute corresponding to the seed region between in-sample nodes and the whole network. Assuming that all the frequencies, in-sample and whole network, are less than 0.5 (which is the case in our

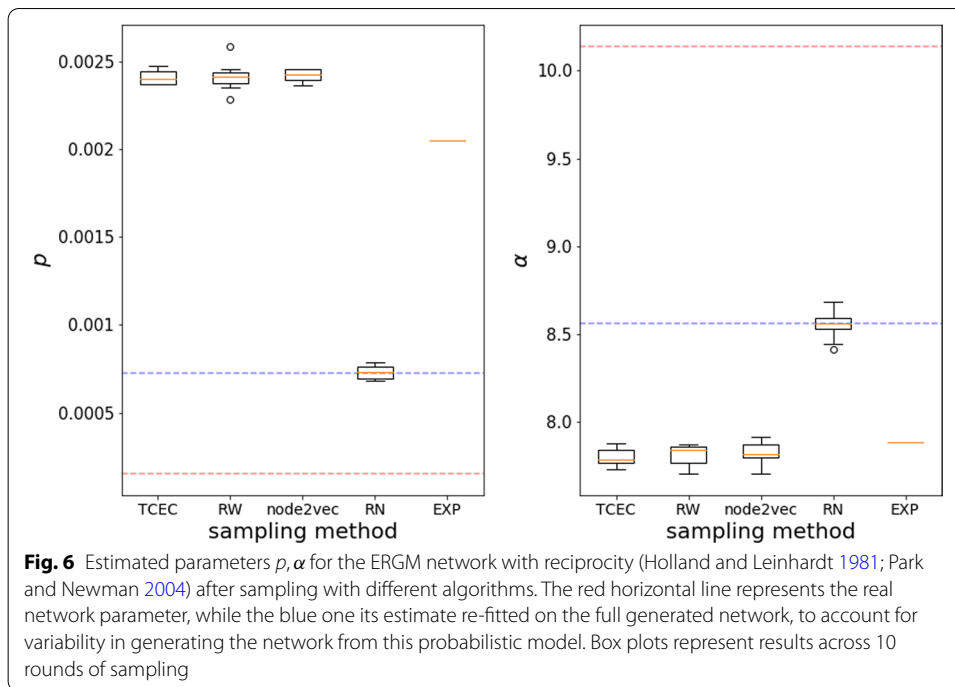
experiments), than values close to 1 denote high similarity, greater than 1 means over representation and less than one under representation of a particular attribute. In all but two starting regions, node2vec has a significantly high entropy ratio: for various seed regions this is higher than 1.19 whereas the maximum values obtained by TCEC and RW are both less than 1.12. Quantitatively, this shows the magnitude of the over representation in the sample induced by node2vec; instead, TCEC and RW do not yield any significant bias towards the starting region. An example of this behavior is plotted in Fig. 5, all the other starting regions are given in “Appendix 7”.

Further statistics

Another interesting question is whether other relevant network statistics are preserved in a controlled setting where we generate synthetic data with a particular underlying structure. We explore this on two different types of synthetic networks where we have control over the generative process. First we generate an Exponential Random Graph Model (ERGM), a probabilistic model popular in social sciences where a set of sufficient statistics has a predefined expected value. We use one of the few ERGMs with an analytical formulation,³ the reciprocity model of Holland and Leinhardt (Holland and Leinhardt 1981); here the network has a fixed number n of nodes and two parameters, p , regulating the probability of directed connections, and α , regulating the expected reciprocity (fraction of nodes with edges connecting them in both directions) (Park and Newman 2004). We set $n = 10^4$, $p = 1.51 \cdot 10^{-4}$, $\alpha = 10.1$. The values of p and α have been chosen as estimates based on the social network Slashdot, since for this type of social network reciprocity and density are particularly relevant and descriptive measures. After generating the synthetic network, we sample 10% of its nodes and recompute on it the parameters for the two sufficient statistics: p being the expected number of directed edges, is estimated as $E/N(N-1)$, where E and N are the number of edges and nodes in the sample; α is estimated reversing the formula (53) in (Park and Newman 2004), using the observed value of network reciprocity on the sample and the estimate of p . Results are presented in Fig. 6. We can observe that the only sampling strategy successfully recovering the same parameters as the original ones is RN, while all the others achieve similarly biased results. A possible interpretation is given by the way neighbors are chosen at sampling time by all but random node sampling. All algorithms but the latter incorporate a bias towards incoming (e.g. TCEC and node2vec) or outgoing connections (e.g. RW). This does not allow a balanced selection of edges in both directions, thus biasing the estimate of α . For similar reasons, by choosing neighboring nodes, this algorithms retrieve denser samples, which naturally yield a higher estimate of p .

The second synthetic model that we consider is the SpringRank generative model. In short, this model defines a network with a built-in hierarchical structure. In words, nodes have a real-valued score parameter s_i denoting their strength, or prestige, and this determines the likelihood of observing a directed and weighted tie between two nodes

³ The reciprocity model, being analytical, is much more efficient to sample from. Other non-analytical models, e.g. ERGM with fixed number of triads, which require Monte Carlo sampling, are computationally too expensive to run on large system sizes like the ones explored here.



(it is only valid for directed networks). Formally, the adjacency matrix A of a graph with n nodes is drawn with the following probabilistic model:

$$s_i \sim N\left(1, \frac{1}{\alpha\beta}\right)$$

$$A_{ij}|s_i, s_j \sim \text{Poisson}\left\{c \exp\left[-\frac{\beta}{2}(s_i - s_j - 1)^2\right]\right\}, \quad (2)$$

for $i, j = 1, \dots, n$ and α, β, c parameters tuning the scores' variance, the hierarchy strength and the network sparsity respectively. In particular, β can be seen as an inverse temperature: the larger its value the stronger the hierarchical relationship of nodes described by the scores is, thus its impact in determining the observed adjacency matrix. We investigate how various sampling algorithms successfully retrieve samples of the graph for which the inferred scores respect the ground truth ones. We generate a graph and its scores according to (2), and sample it with the methods listed above. We then infer the scores as described in De Bacco et al. (2018). The parameters are fixed as $n = 10^5, \alpha = 0.1, c = 0.01$. We vary β in $\{0.1, 1, 10\}$ to check if a stronger signal in the edges, obtained by increasing β , reflects in a better recovery of nodes' scores, regardless the sampling algorithm. Results are presented in Fig. 7. As in the ERGM case, we can compare against two reference values. In this synthetic case we know the real underlying scores that the network has been generated with, but these determine the adjacency matrix of observations in (2) only stochastically. For this reason we compare the scores computed in the samples against the ones computed on the whole graph, rather than with the ground truth ones. As it can be observed, TCEC, RW and node2vec have similar recovery capacity in terms of scores computed on the yielded samples, with TCEC showing the best performance for stronger hierarchical structures ($\beta = 1$); instead, RN and

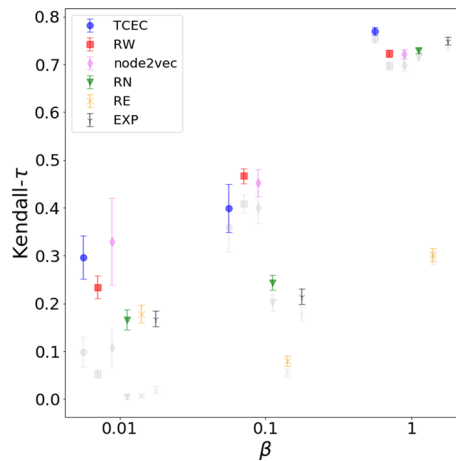
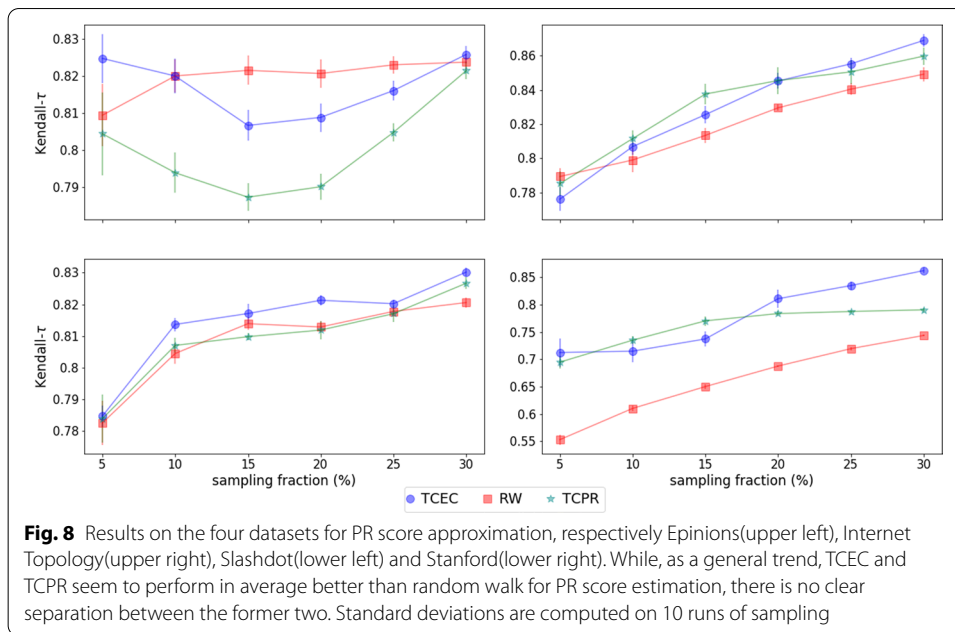


Fig. 7 Kendall- τ correlation between SpringRank scores inferred on samples and full graph, generated with the SpringRank generative model for varying level of inverse temperature β . Scores with respect to ground truth SpringRank scores are represented in shaded gray. Standard deviations are computed across 10 rounds of sampling. Notice that here EXP has variance, as in these experiments we varied the initial seed node

EXP sampling achieve comparable results only for high signal strength, i.e. high β . RE never achieves good performances compared to the other methods. For completeness, Kendall- τ measures with respect to ground truth scores are represented in shaded gray in the plots.

Sampling for PageRank estimation

In this section we discuss the challenges preventing an effective extension of the theoretical framework behind TCEC to PageRank score (PR) (Brin and Page 1998), i.e. a method for sampling networks theoretically grounded on the same ideas, but aiming at better approximating PageRank, rather than eigenvector centrality. In fact, arguably counterintuitively, there is no trivial generalization of TCEC for PageRank. Instead, it is necessary to make further assumptions that result in an algorithmic scheme that is equivalent to TCEC in practice, from our empirical observations. Here we explain the main challenges and refer to the “Appendix 8” for detailed derivations of how to address them. PageRank considers a different adjacency matrix A_{PR} , which is strongly connected (as the network is complete) and stochastic (the rows are normalized to 1). This is built from the original A . Both these features, not present for the eigenvector centrality case, are the cause of the additional complexity of sampling for PageRank. The PR score is defined as the eigenvector centrality computed on A_{PR} . At a first glance, this may lead to a straightforward generalization of TCEC sampling by simply applying the algorithm to A_{PR} . However, this simple scheme hinders in fact one main challenge, which makes this generalization theoretically non trivial. TCEC yields the matrix A_{G_m} (the adjacency of the sampled network G_m), which is a submatrix of the original A ; having a submatrix is a requirement for the validity of the sine distance bound at the core of TCEC. Instead, in the case of PageRank, the matrix of the sampled network A_{PR,G_m} is *not* a submatrix of A_{PR} ; this is because A_{PR} is a stochastic matrix, which requires knowing the degree of each node in advance to normalize each row. This information is in general not known



a priori. We fixed this problem introducing an approximation (see “Appendix 8”) which allows to use the theoretical criterion of Eq. (1) in this case as well. However, we still face a computational challenge. Due to the nature of PageRank, which allows jumps to non-neighboring nodes, albeit with low probability, the networks behind A_{PR} and A_{PR,G_m} are both complete. This results in a much higher computational cost of the sampling algorithm. Even though we proposed ways to fix this issue as well (see “Appendix 8”) and thus combined these two considerations into an efficient algorithmic implementation (which we refer to as TCPR) analogous to TCEC, empirical results for this are poor. In practice, TCEC performs better in recovering the PR scores of nodes in the sample. As noted above for non spectral centrality measures, also in this case this approximation cannot simply be attributed to the correlation between PR and EC; in fact, this is high for Epinions, Internet Topology and Slashdot (respectively 0.580, 0.524, 0.715) but -0.001 for Stanford, see “Appendix 5”. Rather, it has to be attributed to the structural properties of the networks recovered from the various sampling procedures.

TCEC versus TCPR for PageRank approximation

We compare the approximation of the PageRank score as obtained on samples from RW, TCEC and TCPR, via Kendall- τ correlation with the true score, which were assumed to be available in these experiments. A higher correlation signals a better recovery of the relative ranks between nodes. We do so on the Epinions, Internet Topology, Slashdot and Stanford network. The Internet Topology (Zhang et al. 2005) represents the (undirected) Internet Autonomous Systems level topology.

For these experiments we set the TCEC randomization probability to 0.5, to achieve better approximation scores. Figure 8 shows a noticeable improvement of TCEC in most of the networks, both as a function of the sampling ratio and compared to RW for in-sample PR ranking recovery. However, we do not observe such a pattern for TCPR,

which performs better than TCEC only for few datasets and sample ratio combinations. As the theoretical groundings behind the two are similar, we argue that using the L_1 -norm in TCPR (see “[Appendix 8](#)”), which is inherently less discriminative of the L_2 -norm behind TCEC, seems to affect this difference in performance. Another possible cause is the extra assumption of in-sample nodes’ degrees linearly scaling with sample size. Large deviations from this assumption could sensibly impact the quality of the goodness criterion at hand.

Conclusions

Designing a sampling protocol when the whole-network information is not accessible is a task that has to be performed wisely. In fact, the choice of the sampling algorithm biases the analysis of relevant network quantities performed on the sample. We investigated here the impact on various centrality measures, community structure, node attribute distribution and further statistics relevant to specific instances of network generative models that sampling techniques have. We studied in particular the performance of TCEC, a theoretically grounded sampling method aimed at recovering eigenvector centrality on such network properties within the sample and compared with other sampling approaches. The goal was to understand whether a sampling algorithm optimized to preserve a specific global and spectral network measure, is indirectly preserving also other network quantities. We empirically found that on various networks, the performance of TCEC, as well as that of other algorithms, varies with the task, further suggesting to an end-user that the choice of the sampling strategy should be made thoroughly and according to the goal. In particular, in some tasks there is high performance similarity with other routines that sample by moving between adjacent nodes, like RW and node2vec, their performances all differ significantly from strategies based on random uniform sampling. Instead, for tasks like recovering PageRank, a spectral measure, or SpringRank values for strong hierarchical structures it performs better than uniform random walk. In addition, while RW yields community structure homogeneously distributed across blocks, TCEC tends to select nodes inside the starting community, however partially reaching out to other blocks. Finally, studying a large online social network, it recovers in-sample attribute distributions close to the ones of the whole graph. It does not show any significant bias towards the seed region, as it is instead the case for node2vec, which is over representing the starting regions. We discussed possibilities of extending TCEC to the case of PageRank and showed the challenges associated to this task and the remedies to them. However, the resulting algorithm performs comparably well to TCEC on recovering PageRank values. We focused here in showcasing the impact of sampling on three different relevant tasks that have broad relevance in network datasets and presented example of further statistics covering more specific scenarios (i.e. networks with reciprocity or hierarchical structure). We have not considered the case where networks change in time, it would be interesting to measure the robustness of sampling strategies against the dynamics of network structure.

Abbreviations

TCEC: Theoretical Criterion Eigenvector Centrality sampling; TCPR: Theoretical Criterion PageRank centrality sampling; RW: Random Walk sampling; RN: Random Nodes sampling; RE: Random Edges sampling; EXP: snowball expansion sampling; SBM: Stochastic Block Model; ERGM: Exponential Random Graph Model.

Acknowledgements

We are grateful to Mirta Galesic for helpful discussions.

Authors' contributions

All authors contributed to developing the ideas, writing of the manuscript, read and approved the final version.

Funding

Open Access funding enabled and organized by Projekt DEAL. Nicol  Ruggeri is supported by the Max Planck ETH Center for Learning Systems and Cyber Valley. Caterina De Bacco is supported by Cyber Valley.

Availability of data and materials

We provide a table with link to all datasets utilised in this work in "Appendix 1".

Competing interests

The authors declare that they have no competing interests.

Appendix 1: Details of the empirical implementation

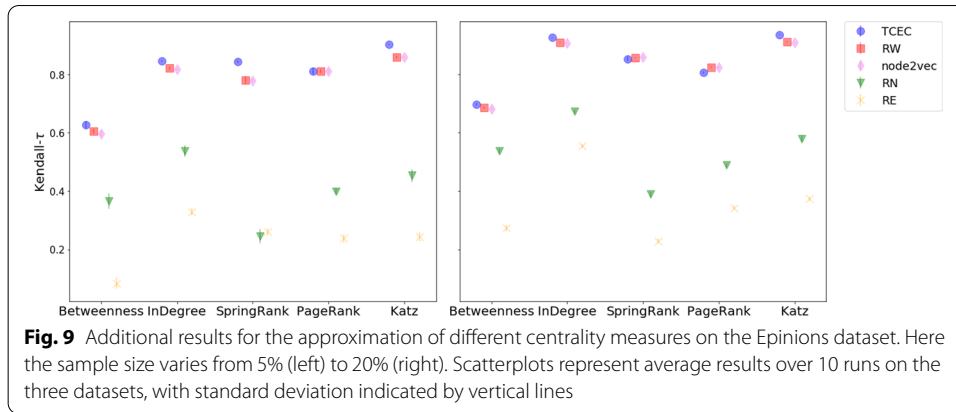
We set the leaderboard size to 100 for both TCEC and TCPR, and the α parameter for TCEC to 0 for undirected networks, 0.5 for directed ones. The explorations are initialized with a random walk sampling of 1/5 of the desired final sample size. The randomization level for neighborhood exploration is set to $p = 0.1$, meaning that 1/10 of the possible nodes are explored, unless specified otherwise. Real datasets were taken from the repositories *KONECT* (Kunegis 2013) and *SNAP* (Leskovec and Krevl 2014), as reported in Table 1.

Table 1 Datasets sources

Name	Source
C. Elegans	http://konect.uni-koblenz.de/networks/arenas-meta
Epinions	http://konect.uni-koblenz.de/networks/soc-Epinions1
Slashdot	https://snap.stanford.edu/data/soc-Slashdot0811.html
Internet Topology	http://konect.uni-koblenz.de/networks/topology ,
US Power grid	http://konect.uni-koblenz.de/networks/opsahl-powergrid
Adolescent Health	http://konect.uni-koblenz.de/networks/moreno_health
Stanford	https://snap.stanford.edu/data/web-Stanford.html
Pokec	https://snap.stanford.edu/data/so-Pokec.html

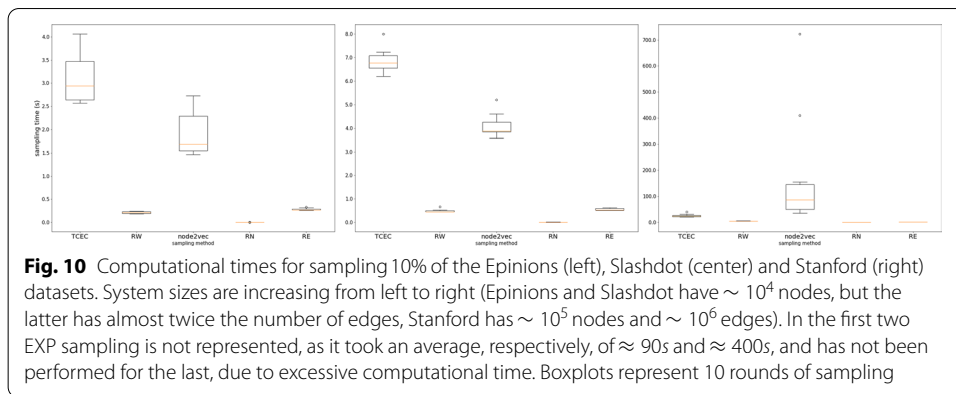
Appendix 2: Sample size stability

To test for the stability of the results presented with respect to the sample size, we present here plots similar to the ones in Figs. 2 and 11 and "Appendix 4". In Fig. 9 we plot the Kendall- τ correlation for the approximation of the same centrality measures on the Epinions dataset. To test different sample sizes, we sample 5% and 20% of the total nodes, as opposed to the 10% for the results presented in the main text (Fig. 2).



Appendix 3: Clock time for different sampling algorithms

We include in Fig. 10 a plot of the sampling time for the various algorithms considered on the Epinions, Slashdot and Stanford datasets. Notice that EXP sampling has only been performed on the first, smaller sized, two.



Appendix 4: Additional results for spectral centrality measures

We include results similar to the ones presented above for the approximation of non-spectral centrality measures for spectral measures PageRank and Katz centrality.

Appendix 5: Correlation between EC, PR and in-degree

We include in Table 2 the correlations between EC, PR and in-degree for all the real world datasets presented in the experimental section.

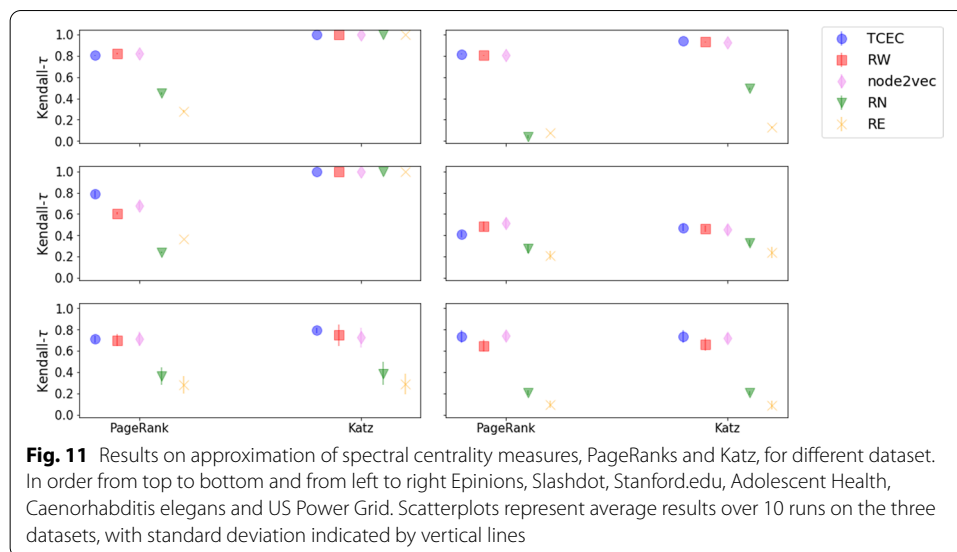


Table 2 Correlations between EC, PR and in-degree for the various datasets used in the experimental section

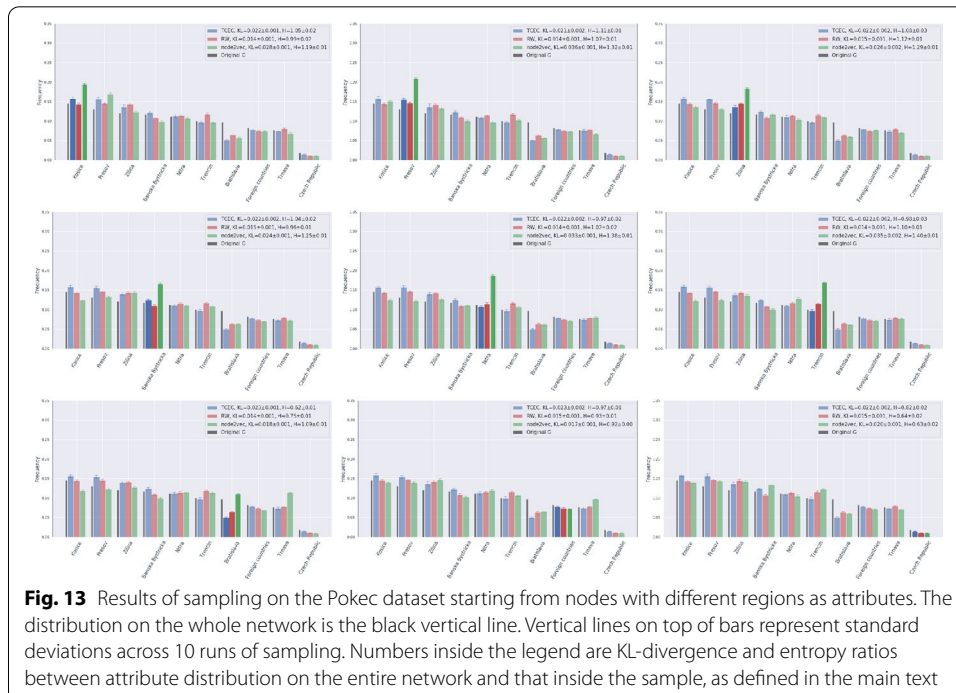
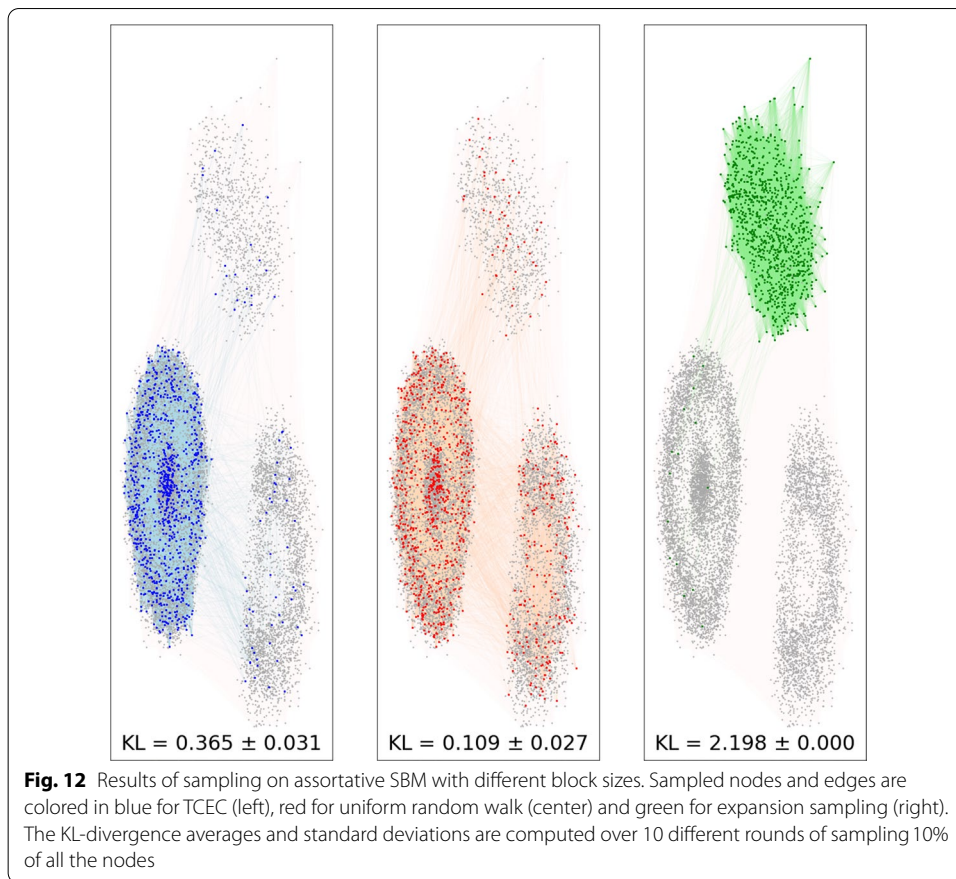
	InDeg versus EC	InDeg versus PR	EC versus PR
C. Elegans	0.925	0.997	0.900
Epinions	0.836	0.863	0.580
Slashdot	0.807	0.971	0.715
Internet Topology	0.707	0.949	0.524
US Power grid	0.261	0.946	0.086
Adolescent Health	0.239	0.715	0.317
Stanford	0.034	0.742	-0.001

Appendix 6: Additional results on SBM for community sampling

We include in Fig. 12 the results for sampling on an assortative SBM structure with three unbalanced groups of sizes 1000, 3000 and 6000 nodes, within-block connection probability $p_{in} = 0.05$ and between-blocks connection probability $p_{out} = 0.005$.

Appendix 7: Extra plots for results on Pokec dataset

We include in Fig. 13 the plots of the results on the Pokec dataset, similar to Fig. 5, but where the seed point is chosen in each of the remaining regions. The bars relative to the initial region are bolded.



Appendix 8: TCPR: extending the model to PageRank score

Before introducing the theory behind TCPR, we begin with a review of the basic PageRank algorithm, introduce some notation and outline the main challenges and assumptions needed in the following derivations.

*Notation Consider a nonnegative adjacency matrix $A \in \mathbb{R}^{V,V}$. Then we build a new adjacency matrix A_{PR} , called *PageRank adjacency matrix*, defined as follows

$$A_{PR} := \gamma P + \frac{(1-\gamma)}{V} ee^T, \quad (3)$$

where e is the vector of ones of length V , $\gamma \in [0, 1)$ and P is defined as

$$P = Q + \frac{ed^T}{V} \quad (4)$$

$$d = e^T - e^T Q \quad (5)$$

$$Q = A \cdot \text{diag}\left(\frac{1}{d_1}, \dots, \frac{1}{d_V}\right), \quad (6)$$

with $d_j = \sum_i A_{ij}$ out-degree of node j and with the convention that for node with zero out-degree, named *dangling nodes*, we take $1/d_j = 0$.

For all the matrices P, Q, A_{PR} we can define two different quantities. Given a subset of nodes $\{1, \dots, m\}$ of V we have the principal submatrices $P_m, Q_m, A_{PR,m}$ relative to these nodes. But we can also compute the PageRank scores on the subgraph G_m . The matrices relative to G_m are instead noted as $P_{G_m}, Q_{G_m}, A_{PR,G_m}$. Notice that for the original case of eigenvector centrality we had the correspondence $A_m = A_{G_m}$, we will simply refer to this as A_m .

*Challenges In general, as we sample, we only know $d_{in}^{G_m}(i)$ but may not have access to $d_{in}^G(i)$ (in general $d_{in}^{G_m}(i) \leq d_{in}^G(i)$); this implies that the entries of A_{PR,G_m} are different than the submatrix $A_{PR,m}$ of A_{PR} induced by the nodes in G_m . We tackle this challenge by making an additional assumption: we assume that the degree $d_{in}^{G_m}(i) \approx \frac{m}{V} d_{in}(i)$, i.e. degrees of nodes in the sample scale linearly with the sample size m ; this is a necessary approximation for linking the two otherwise different matrices $A_{PR,m}$ and A_{PR,G_m} (which where instead equal for eigenvector centrality), its validity has been justified (Ganguly and Kolaczyk 2018) and thus we can use the theoretical criterion of Eq. (1) in this case as well. This fixes a theoretical challenge, however, we now face a computational one. Due to the nature of PageRank, which allows jumps to non-neighboring nodes, albeit with low probability, the networks behind A_{PR} and A_{PR,G_m} are both complete. This results in a much higher computational cost of the sampling algorithm. We reduce this by selecting candidate nodes to be added to the sample, in analogy with TCEC, among the *incoming* neighbors only, thus neglecting nodes that correspond to a non-zero entry of A_{PR} but do not correspond to an actual edge. This has also the advantage of excluding dangling nodes (i.e. nodes with out-degree zero) from the sample. Combining these two considerations, we obtain a sampling criterion similar to the one employed in TCEC; we name this TCPR (Theoretical Criterion PageRank).

*Adapting the theory to PageRank While for “vanilla” eigenvector centrality the matrix A was by hypothesis sparse, and therefore border exploration feasible, now the network represented by A_{PR} is complete. Border exploration, even if randomized by a level p , would be of cost $O(V)$. For TCEC the choice was to choose all *incoming* neighbours in the sample. Here we can do the same, but *only choosing incoming neighbours from the original network A* . This because an incoming connection in A has weight $\sigma(1/d_{out})$ in A_{PR} , while one due to the artificial edges in (4) and (3) have total weight $\sigma(\frac{k+1}{V})$, which is negligible for sample size $k \ll V$.

This is also in line with the observation that in many sampling scenarios we are not really able to pick nodes in the graph at random, but just explore neighbourhoods (Gjoka et al. 2010). Additionally, by only considering incoming neighbours, which have out-degree necessarily greater than 0, we exclude all the dangling nodes from the final sample.

Notice that the theorem in Ruggeri and De Bacco (2020) was comparing the principal eigenvectors of a matrix A and a principal submatrix A_m . In the case of PageRank, this is not applicable. In fact the matrix Q from eq (6) is normalized differently. In A , the rescaling is done on the full graph, while in A_m on the subgraph degrees. This means that $Q_m \neq Q_{G_m}$, and consequently $P_m \neq P_{G_m}$, $A_{PR} \neq A_{PR,m}$. This problem can be overcome by making a further assumption. In a pseudo-random choice of any subsample of size m , it is reasonable to assume that nodes' degrees scale linearly, i.e. $N_{j,G_m} \approx \frac{m}{V}N_j$. By holding this approximation as valid, and recalling that there are no dangling nodes in the subgraph, it is straightforward to check that $A_{PR,G_m} = \frac{V}{m}A_{PR,m}$. In particular, the eigenvector centrality for sampled nodes is the same in the complete graph G and the sampled one G_m , since $A_{PR,G_m}, A_{PR,m}$ have the same eigenvectors. This overcomes the first issue of linking the PR score on G_m and G , and we can simply sample nodes with the goodness criterion from Ruggeri and De Bacco (2020) on the page rank matrix A_{PR} .

We are left with the necessity of computing the goodness criterion efficiently.

*Efficient criterion computation Suppose, without loss of generality, that the sampled nodes are $\{1, \dots, k\}$ and the new node under evaluation $k+1$. Considering the PR adjacency matrix A_{PR} the quantities involved in the theoretical criterion (1) are:

$$b_1 = \alpha P_{1:k,k+1} + \frac{1-\alpha}{V}e \quad (7)$$

$$b_3 = \alpha P_{k+1,k+2:n} + \frac{1-\alpha}{V}e \quad (8)$$

$$U = \alpha P_{1:k,k+2:n} + \frac{1-\alpha}{V}\mathbb{1}, \quad (9)$$

where $e, \mathbb{1}$ are respectively a vector and matrix of all ones, of correct dimensions. Moreover, we need $b_1^T U$. By explicit calculations:

$$b_1^T U = \alpha^2 P_{1:k,k+1}^T P_{1:k,k+2:n} + \frac{\alpha(1-\alpha)}{V} P_{1:k,k+1} \mathbb{1}^T + \frac{\alpha(1-\alpha)}{V} e^T P_{1:k,k+2:n} + \left(\frac{1-\alpha}{V} \right)^2 e^T \mathbb{1}^T. \quad (10)$$

Implementing the computation of $b_1^T U$ in sparse arithmetic is not convenient, as it would anyway cost $O(k)$. Performing this increasingly costly operation for all (or some) of the nodes in the border at every new node sampled is not feasible. Here we optimise this computation explicitly. First, notice from equation (10) that many terms are independent on the sample. Therefore we compute the L_1 -norm (Kamvar et al. 2003) for all the vectors (7), (8), (10). In all the following computations we use the symbol \approx to indicate equality up to an additive constant independent on the sampled node k . a_{ij} stands for the element i, j of A .

- term b_1 :

$$\|b_1\|_1 \approx \alpha \|P_{1:k,k+1}\|_1 = \frac{\alpha}{N_k} \sum_{j \in G_k} a_{jk}$$

- term b_3 :

$$\begin{aligned} \|b_3\|_1 \approx \alpha \|P_{k+1,k+2:n}\|_1 &= \alpha \left(\sum_{\substack{i \notin G_k \cup \{k+1\} \\ i \text{ not dangling}}} \frac{a_{ki}}{N_i} + \sum_{\substack{i \notin G_k \cup \{k+1\} \\ i \text{ dangling}}} \frac{1}{n} \right) \\ &\approx \alpha \sum_{\substack{i \notin G_k \cup \{k+1\} \\ i \text{ not dangling}}} \frac{a_{ki}}{N_i} \end{aligned}$$

where the last equality is justified by the fact that, since $k+1$ cannot be dangling, $\{i \notin G_k \cup \{k+1\} : i \text{ dangling}\} = \{i \notin G_k : i \text{ dangling}\}$, which is independent on $k+1$.

- term $b_1^T U$. For this we need to split the computation in three, since from equation (10):

$$\|b_1^T U\|_1 \approx \alpha^2 \|P_{1:k,k+1}^T P_{1:k,k+2:n}\|_1 + \frac{\alpha(1-\alpha)}{n} \|P_{1:k,k+1}\|_1 + \frac{\alpha(1-\alpha)}{n} \|e^T P_{1:k,k+2:n}\|_1$$

For every node $j \in G_m$ define $\delta_j := \sum_{i \notin G_k \cup \{k+1\}} P_{ij}$ (which also depends on the sub-sample G_k and the new proposal node $k+1$, we omit the dependence in the notation). Then:

$$\begin{aligned}
\|P_{1:k,k+1}^T P_{1:k,k+2:n}\|_1 &= \frac{1}{N_k} \sum_{j \in G_k} a_{jk} \sum_{i \notin G_k \cup \{k+1\}} P_{ij} = \frac{1}{N_k} \sum_{j \in G_k} a_{jk} \delta_j \\
\|P_{1:k,k+1}\|_1 &= \sum_{i \notin G_k \cup \{k+1\}} \sum_{j \in G_k} P_{ji} \\
&= \sum_{\substack{i \notin G_k \cup \{k+1\} \\ i \text{ not dangling}}} \sum_{j \in G_k} \frac{a_{ji}}{N_i} + \sum_{\substack{i \notin G_k \cup \{k+1\} \\ i \text{ not dangling}}} \sum_{j \in G_k} \frac{1}{n} \\
&\doteq \sum_{\substack{i \notin G_k \cup \{k+1\} \\ i \text{ not dangling}}} \sum_{j \in G_k} \frac{a_{ji}}{N_i} \\
&\doteq \left(\sum_{\substack{i \notin G_k \\ i \text{ not dangling}}} \sum_{j \in G_k} \frac{a_{ji}}{N_i} \right) - \left(\sum_{j \in G_k} \frac{a_{jk}}{N_k} \right) \\
&\doteq -\frac{a_{jk}}{N_k} \sum_{j \in G_k} \\
\|e^T P_{1:k,k+2:n}\|_1 &= (n-k-1) \sum_{j \in G_k} P_{jk} = \frac{n-k-1}{N_k} \sum_{j \in G_k} a_{jk}
\end{aligned} \tag{11}$$

Now, why is expression (11) more efficient? Because we keep an updated calculation of the terms δ_j in memory. After the first random walk initialization we compute δ_j for every j in the sample. Then, whenever a node is added to the sample, they are updated. Namely, say that a node s is added to G_m . Then for all the outgoing neighbours j of s already in G_m , we perform the update $\delta_j \leftarrow \delta_j - P_{js} = \delta_j - \frac{a_{js}}{N_j}$. Summing up we get

$$\begin{aligned}
\|b_1^T U\|_1 &= \frac{1}{N_k} \left(\alpha^2 \sum_{j \in G_k} a_{jk} \delta_j + \frac{\alpha(1-\alpha)(n-k-2)}{n} \sum_{j \in G_k} a_{jk} \right) \\
&= \frac{\alpha}{N_k} \sum_{j \in G_k} a_{jk} \left(\frac{(1-\alpha)(n-k-2)}{n} + \alpha \delta_j \right)
\end{aligned}$$

Notice that all the quantities here are expressed as a sum over all the nodes in G_m . However, the summands depend on the edges of the new nodes to be added, and can therefore be performed in $O(d_{in})$ or $O(d_{out})$. As opposed to $O(k)$, this is constant with respect to the sample size. As a final remark, we would like to highlight the fact that it is much harder to find such a computational trick for the L_2 norm of the criterion vectors. This was instead possible for TCEC, where they had a simpler expression that allowed derivations.

Appendix 9: SBM and TCEC score

Here we show a theoretical analysis of how TCEC sampling works on a Stochastic Block Model as the one studied in Sec. 3.3. Specifically, we give approximations for the main quantities inside Eq. (1), $\|b_1\|_2^2$, $\|b_3\|_2^2$, $\|b_1^T U\|_2^2$.

SBM generative model

We start by reminding the main assumptions of the SBM. Nodes have colors $q_i \in \{1, \dots, K\}$. The $K \times K$ affinity matrix π contains the probability π_{kq} that there is an edge between nodes of color k and nodes of color q . We assume the simple case $\pi_{kq} = p_{in}$, when $k = q$ and $\pi_{kq} = p_{out}$, when $k \neq q$, and undirected networks, $\pi_{kq} = \pi_{qk}$. We then assume assortative structure so that $p_{out} = \epsilon p_{in}$ with $0 \leq \epsilon < 1$. Typically, $\epsilon \sim 10^{-1}, 10^{-2}$. The fraction of nodes of color k is $n_k = 1/K$, i.e. equal-size groups. The probability of an edge between i and j is $P(A_{ij}; q_i, q_j, \pi) = \pi_{q_i q_j}^{A_{ij}} (1 - \pi_{q_i q_j})^{1-A_{ij}}$.

SBM and sampling

We now assume that G_m has size $M = \alpha N$, where $0 < \alpha \leq 1$, this is the sample size ratio; values considered in our experiments are $\alpha \in [0.01, 0.4]$. We initialize the sample using a uniform-edge RW. We assume that if we initialize the walker on a node of color r , given we are far from equilibrium and the network is assortative, it is reasonable to expect that the initial sample contains more nodes of color r than what expected with equal-size colors, as in the original graph. Formally, we have $n_r^m = n_r + \delta = 1/K + \delta$, where $\delta > 0$ is the bias of color r inside the sample. This means that the sample G_m has $N_r^m = M n_r^m$ nodes of color r . Instead, the graph $G' = G \setminus G_m$, has $N'_r = N n_r - M n_r^m = (N - M) n_r - M \delta$, which implies $n'_r = n_r - M/(N - M) \delta \leq n_r$, i.e. there is a lower fraction of nodes of color r inside G' compared to the homogeneous one of the original network. It is convenient to denote the nodes that do not have color r with an index nr . Hence the nodes of colors other than r in the sample are $N_{nr}^m = M - N_r^m = M(1 - n_r - \delta)$, instead inside the remaining network there are $N'_{nr} = N - M - N'_r = (N - M)(1 - n_r + M/(N - M)\delta)$. We then assume that these nodes nr are homogeneously divided into the remaining $K - 1$ groups.

We now proceed selecting a $j \in G'$ and determining its TCEC score based on all the parameters defined above. As we will see, its score depends on the color q_j .

$\|b_1\|_2^2$

The quantity b_1 is a M -dimensional vector containing the edges from j towards nodes in G_m . The expected value of an edge $\mathbb{E}[A_{ji}] = \pi_{q_i, q_j}$, where $i \in G_m$, depends on their colors, hence also b_1 does. We then denote with $b_1(k)$ the value of b_1 when j has color k . Given there is a bias towards color r , we only need to distinguish the case of $k = r$ and $k \neq r$, again denoted as nr . The entries of $b_1(r)$ have only two different values, depending on what nodes of G_m are of color r or not. Similarly for $b_1(nr)$. If $q_i = r$, then $b_{1,i}(r) = p_{in}$, otherwise if $q_i \neq r$, $b_{1,i}(r) = p_{out}$. Considering that there are N_r^m and $N_{nr}^m/(K - 1)$ nodes of color r and of any other color in G_m , respectively, on average we have

$$\|b_1(r)\|_2^2 \approx \mathbb{E}[\|b_1(r)\|_2^2] = N_r^m p_{in} + N_{nr}^m p_{out} \quad (12)$$

$$\|b_1(nr)\|_2^2 \approx \mathbb{E}[\|b_1(nr)\|_2^2] = N_{nr}^m/(K-1)p_{in} + (M - N_{nr}^m/(K-1))p_{out} \quad . \quad (13)$$

The difference is:

$$\begin{aligned} \mathbb{E}[\|b_1(r)\|_2^2 - \|b_1(nr)\|_2^2] &= (p_{in} - p_{out}) \left(N_r^m - \frac{N_{nr}^m}{K-1} \right) \\ &= p_{in}(1 - \epsilon) \left(\frac{M}{K} + M\delta - \frac{M}{K} + \frac{M\delta}{K-1} \right) \\ &= M p_{in}(1 - \epsilon) \frac{K}{K-1} \delta. \end{aligned} \quad (14)$$

$$\|b_3\|_2^2$$

The quantity b_3 is a $(N - M)$ -dimensional vector containing the edges from G' towards j . Similar calculations as before can be done, accounting for size of G' and the fraction of nodes of color r is $n'_r \leq n_r \leq n_r^m$. Thus

$$\|b_3(r)\|_2^2 \approx \mathbb{E}[\|b_3(r)\|_2^2] = N'_r p_{in} + N'_r p_{out} \quad (15)$$

$$\|b_3(nr)\|_2^2 \approx \mathbb{E}[\|b_3(nr)\|_2^2] = N'_{nr}/(K-1)p_{in} + (N - M - N'_{nr}/(K-1))p_{out} \quad . \quad (16)$$

The difference is:

$$\begin{aligned} \mathbb{E}[\|b_3(r)\|_2^2 - \|b_3(nr)\|_2^2] &= (p_{in} - p_{out}) \left(N'_r - \frac{N'_{nr}}{K-1} \right) \\ &= p_{in}(1 - \epsilon) \left(\frac{N - M}{K} - M\delta - \frac{N - M}{K} - \frac{M\delta}{K-1} \right) \\ &= -M p_{in}(1 - \epsilon) \frac{K}{K-1} \delta. \end{aligned} \quad (17)$$

We thus obtain

$$\mathbb{E}[\|b_1(r)\|_2^2 - \|b_1(nr)\|_2^2] = \mathbb{E}[\|b_3(r)\|_2^2 - \|b_3(nr)\|_2^2], \quad (18)$$

hence, the only terms that matter for the selecting a node using the TCEC score of Eq. (1) in the SBM is the $\|b_1^T U\|_2^2$. Specifically, the node with highest score has color r if $\|b_1^T U(r)\|_2^2 - \|b_1^T U(nr)\|_2^2 > 0$, in the case $\alpha = 0$ inside Eq. (1). In case $\alpha > 0$, then one can simply assume that there are at least two nodes, one with color r and one with different color, with comparable and high degree, so that most of the TCEC score difference is due to the term $b_1^T U$.

$$\|b_1^T U\|_2^2$$

The quantity $b_1^T U$ is a $(N - M)$ -dimensional vector. It contains the number of common neighbors inside the sample that nodes in $G' \setminus \{j\}$ have with j , i.e. $(b_1^T U)_j = \sum_{i \in G_m} A_{ji} A_{\ell i}$, for $\ell \neq j \in G'$. Therefore we need to calculate quantities like $\mathbb{E}[\sum_{\ell \neq j \in G'} (\sum_{i \in G_m} A_{ji} A_{\ell i})^2]$.

Formally

$$\|b_1^T U\|_2^2 = \sum_{\ell \in G'} \left(\sum_{i \in G_m} A_{ji} A_{\ell i} \right)^2. \quad (19)$$

Fixing one $\ell \in G'$, the inner squared sum becomes

$$\left(\sum_{i \in G_m} A_{ji} A_{\ell i} \right)^2 = \sum_{i \in G_m} (A_{ji} A_{\ell i})^2 + 2 \sum_{i, h \in G_m, i \neq h} A_{ji} A_{\ell i} A_{jh} A_{\ell h} \quad (20)$$

$$= \sum_{i \in G_m} A_{ji} A_{\ell i} + 2 \sum_{i, h \in G_m, i \neq h} A_{ji} A_{\ell i} A_{jh} A_{\ell h}, \quad (21)$$

where we used the fact that we are assuming binary adjacency matrices, i.e. $A_{ij}^2 = A_{ij}$. Taking the expected value

$$\|b_1^T U\|_2^2 \approx \mathbb{E} \left[\sum_{\ell \neq j \in G'} \left(\sum_{i \in G_m} A_{ji} A_{\ell i} + 2 \sum_{i, h \in G_m, i \neq h} A_{ji} A_{\ell i} A_{jh} A_{\ell h} \right) \right] \quad (22)$$

$$= \sum_{\ell \neq j \in G'} \sum_{i \in G_m} \mathbb{E}[A_{ji} A_{\ell i}] + 2 \sum_{\ell \neq j \in G'} \sum_{i, h \in G_m, i \neq h} \mathbb{E}[A_{ji} A_{\ell i} A_{jh} A_{\ell h}]. \quad (23)$$

In a SBM, a common assumption is that of conditional independence: edges are conditionally independent, given the parameters. Thus $\sum_{\ell \neq j \in G'} \sum_{i \in G_m} \mathbb{E}[A_{ji} A_{\ell i}] = \sum_{i \in G_m} (\mathbb{E}[A_{ji}] \sum_{\ell \neq j \in G'} \mathbb{E}[A_{\ell i}])$.

$$\|b_1^T U\|_2^2 \approx \sum_{i \in G_m} \left(\mathbb{E}[A_{ji}] \sum_{\ell \in G'} \mathbb{E}[A_{\ell i}] \right) + 2 \sum_{i, h \in G_m, h \neq i} \mathbb{E}[A_{ji}] \mathbb{E}[A_{jh}] \sum_{\ell \neq j \in G'} \mathbb{E}[A_{\ell i}] \mathbb{E}[A_{\ell h}]. \quad (24)$$

The expression above depends on the colors of i, h, l, j being r or not. For instance, if $q_i = q_h = r$, then

$$\sum_{\ell \neq j \in G'} \mathbb{E}[A_{\ell i}] \mathbb{E}[A_{\ell h}] = \sum_{\ell \neq j \in G'} \mathbb{E}^2[A_{\ell i}] = N'_r p_{in}^2 + N'_{nr} p_{out}^2. \quad (25)$$

All of the other color combinations of terms scale with similar behavior. Completing the calculations accounting for the whole $\sum_{i, h \in G_m, h \neq i} \mathbb{E}[A_{ji}] \mathbb{E}[A_{jh}] \sum_{\ell \neq j \in G'} \mathbb{E}[A_{\ell i}] \mathbb{E}[A_{\ell h}]$ gives extra factors of the order of $(N_r^m)^2 p_{in}^2$. Because $p_{in} \sim \frac{1}{N}$ in the sparse regime, the total weight of this term is $(N_r^m)^2 p_{in}^2 \times N_r' p_{in}^2 \sim \frac{1}{N^2}$. This is negligible compared to the term $\sum_{i \in G_m} (\mathbb{E}[A_{ji}] \sum_{\ell \in G'} \mathbb{E}[A_{\ell i}])$. The quantity $\mathbb{E}[A_{ji}]$ can be estimated using same reasoning as done for b_1 above. The most important terms are thus:

$$\|b_1^T U\|_2^2(r) \approx N_r^m p_{in} (N'_r p_{in} + (N - M - N'_r) p_{out}) \quad (26)$$

$$+ N_{nr}^m p_{out} \left(\frac{N'_{nr}}{K-1} p_{in} + \left(N - M - \frac{N'_{nr}}{K-1} \right) p_{out} \right) \quad (27)$$

$$\|b_1^T U\|_2^2(nr) \approx \frac{N_{nr}^m}{K-1} p_{in} \left(\frac{N'_{nr}}{K-1} p_{in} + \left(N - M - \frac{N'_{nr}}{K-1} \right) p_{out} \right) \quad (28)$$

$$+ N_{nr}^m \left(\frac{K-2}{K-1} \right) p_{out} \left(\frac{N'_{nr}}{K-1} p_{in} + \left(N - M - \frac{N'_{nr}}{K-1} \right) p_{out} \right) \quad (29)$$

$$+ N_r^m p_{out} (N'_r p_{in} + (N - M - N'_r) p_{out}) \quad . \quad (30)$$

The difference between these two terms is

$$\begin{aligned} & \|b_1^T U\|_2^2(r) - \|b_1^T U\|_2^2(nr) \\ & \approx N_r^m N'_r (p_{in} - p_{out})^2 - \frac{N_{nr}^m}{K-1} \frac{N'_{nr}}{K-1} (p_{in} - p_{out})^2 \\ & \quad + (N - M) \frac{N_r^m}{K-1} p_{in}^2 (1 - \epsilon) \epsilon - (N - M) \frac{N_{nr}^m}{K-1} p_{in}^2 (1 - \epsilon) \epsilon \\ & = p_{in}^2 (1 - \epsilon) \left[\left(N_r^m N'_r - \frac{N_{nr}^m}{K-1} \frac{N'_{nr}}{K-1} \right) (1 - \epsilon) + (N - M) \epsilon \left(N_r^m - \frac{N_{nr}^m}{K-1} \right) \right]. \end{aligned} \quad (31)$$

The first term inside the bracket is:

$$\begin{aligned} N_r^m N'_r - \frac{N_{nr}^m}{K-1} \frac{N'_{nr}}{K-1} &= \left(\frac{M}{K} + M\delta \right) \left(\frac{N-M}{K} - M\delta \right) \\ &\quad - \left(\frac{M}{K} - \frac{M\delta}{K-1} \right) \left(\frac{N-M}{K} + \frac{M\delta}{K-1} \right) \end{aligned} \quad (32)$$

$$= (N - M) \frac{M\delta}{K-1} - \frac{M^2\delta}{K-1} - (M\delta)^2 \frac{K-2}{(K-1)^2} K \quad (33)$$

$$= \frac{M\delta}{K-1} \left[N - 2M - M\delta \frac{K-2}{K-1} K \right] \quad . \quad (34)$$

The last term inside the brackets is:

$$(N - M) \epsilon \left(N_r^m - \frac{N_{nr}^m}{K-1} \right) = (N - M) \epsilon M\delta \frac{K}{K-1} > 0 \quad . \quad (35)$$

From this we can notice that, in the case of $M \ll N$, i.e. when the initial sample size is very small compared to the system size, we obtain that:

$$\|b_1^T U\|_2^2(r) - \|b_1^T U\|_2^2(nr) \approx p_{in}^2 (1 - \epsilon) \frac{M\delta}{K-1} [1 + \epsilon(K-1)] N > 0 \quad . \quad (36)$$

This means that the best scoring $j \in G'$, based on the TCEC criterion, has color r . The score difference with candidates of color different than r is also increasing with the bias δ , thus reinforcing this bias as more nodes of color r are added to the sample G_m . Instead,

if $M = \alpha N$ but with α finite and $0 < \alpha < 1$, and we also assume that $\epsilon \ll 1$ so that ϵ is small enough that the dominant term is that of Eq. (32):

$$\|b_1^T U\|_2^2(r) - \|b_1^T U\|_2^2(nr) \approx p_{in}^2 (1 - \epsilon) \frac{M \delta}{K - 1} \left[N - 2M - M \delta \frac{K - 2}{K - 1} K \right]. \quad (37)$$

The term inside the brackets can become negative as M increases. The sampling thus proceeds as following: at the beginning, when M is very small compared to N , δ is initially increasing. In turns, also M is increasing, as more nodes are added to the sample. However, at a certain point, as both M and δ increase, the best candidate will switch to be of color different than r , this happens when the term $\left[N - 2M - M \delta \frac{K - 2}{K - 1} K \right]$ becomes negative. This happens earlier when K is bigger. Finally, an Erdos–Renyj case is obtained when $\epsilon = 1$, i.e. $p_{in} = p_{out} = p$. In this case all the terms are zero, i.e. the TCEC score difference is zero and there is no preferred candidates based on the color. In the generalized TCEC score, i.e. $\alpha > 0$, then the node with higher degree $d_{in}^{G_m}(j)$ will be chosen as best candidate. Finally notice that in the disassortative case, i.e. $\epsilon > 1$, there is a change of sign that makes nodes of color different than r have higher score. In this case then the sampling dynamics keep jumping to nodes of different colors, hence the sample is more balanced.

Received: 12 March 2020 Accepted: 12 October 2020

Published online: 28 October 2020

References

- Adler M, Mitzenmacher M (2001) Towards compressing web graphs. In: Proceedings DCC 2001. Data compression conference. IEEE, pp 203–212
- Ahmed NK, Neville J, Kompella R (2012) Network sampling designs for relational classification. In: Sixth international AAAI conference on weblogs and social media
- Antunes N, Bhamidi S, Guo T, Pipiras V, Wang B (2018) Sampling-based estimation of in-degree distribution with applications to directed complex networks. arXiv preprint [arXiv:1810.01300](https://arxiv.org/abs/1810.01300)
- De Bacco C, Larremore DB, Moore C (2018) A physical model for efficient ranking in networks. *Sci Adv* 4(7):8260
- Blagus N, Šubelj L, Bajec M (2017) Empirical comparison of network sampling: how to choose the most appropriate method? *Physica A* 477:136–148
- Bonacich P (1972) Factoring and weighting approaches to status scores and clique identification. *J Math Sociol* 2(1):113–120
- Brin S, Page L (1998) The anatomy of a large-scale hypertextual web search engine. *Comput Netw ISDN Syst* 30(1–7):107–117
- Chen Y-Y, Gan Q, Suel T (2004) Local methods for estimating pagerank values. In: Proceedings of the thirteenth ACM international conference on information and knowledge management. ACM, pp 381–389
- Contisciani M, Power E, De Bacco C (2020) Community detection with node attributes in multilayer networks. arXiv preprint [arXiv:2004.09160](https://arxiv.org/abs/2004.09160)
- Costenbader E, Valente TW (2003) The stability of centrality measures when networks are sampled. *Soc Netw* 25(4):283–307
- Davis JV, Dhillon IS (2006) Estimating the global pagerank of web communities. In: Proceedings of the 12th ACM SIGKDD international conference on knowledge discovery and data mining. ACM, pp 116–125
- De Choudhury M, Lin Y-R, Sundaram H, Candan KS, Xie L, Kelliher A (2010) How does the data sampling strategy impact the discovery of information diffusion in social media? In: Fourth international AAAI conference on weblogs and social media
- Espín-Noboa L, Wagner C, Karimi F, Lerman K (2018) Towards quantifying sampling bias in network inference. *Companion Proc Web Conf* 2018:1277–1285
- Frank O (2005) Network sampling and model fitting. *Models and methods in social network analysis*, pp 31–56
- Ganguly, A., Kolaczyk, E.D (2018) Estimation of vertex degrees in a sampled network. In: 2017 51st asilomar conference on signals, systems, and computers. IEEE, pp 967–974
- Gjoka M, Kuran M, Butts CT, Markopoulou A (2010) Walking in Facebook: a case study of unbiased sampling of OSNs. In: 2010 Proceedings IEEE Infocom. IEEE, pp 1–9

- Grover A, Leskovec J (2016) node2vec: scalable feature learning for networks. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, pp 855–864
- Han J-DJ, Dupuy D, Bertin N, Cusick ME, Vidal M (2005) Effect of sampling on topology predictions of protein-protein interaction networks. *Nat Biotechnol* 23(7):839
- Han C-G, Lee S-H (2016) Analysis of effect of an additional edge on eigenvector centrality of graph. *J Korea Soc Comput Inf* 21(1):25–31
- He Y, Wai H-T (2020) Estimating centrality blindly from low-pass filtered graph signals. In: ICASSP 2020-2020 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, pp 5330–5334
- Holland PW, Laskey KB, Leinhardt S (1983) Stochastic blockmodels: first steps. *Soc Netw* 5(2):109–137
- Holland PW, Leinhardt S (1981) An exponential family of probability distributions for directed graphs. *J Am Stat Assoc* 76(373):33–50
- Hübler C, Krieger H-P, Borgwardt K, Ghahramani Z (2008) Metropolis algorithms for representative subgraph sampling. In: 2008 eighth IEEE international conference on data mining. IEEE, pp 283–292
- Kamvar SD, Haveliwala TH, Manning CD, Golub GH (2003) Extrapolation methods for accelerating pagerank computations. In: Proceedings of the 12th international conference on world wide web, pp 261–270
- Katz L (1953) A new status index derived from sociometric analysis. *Psychometrika* 18(1):39–43
- Kendall MG (1990) Rank correlation methods, 5th edn. A Charles Griffin Title. <https://www.bibsonomy.org/bibtex/2b5c89320f7c7f43cf6d7865d19a1a02c/asalber>
- Kossinets G (2006) Effects of missing data in social networks. *Soc Netw* 28(3):247–268
- Kullback S, Leibler RA (1951) On information and sufficiency. *Ann Math Stat* 22(1):79–86
- Kunegis J (2013) Konect: the Koblenz network collection. In: Proceedings of the 22nd international conference on world wide web, pp 1343–1350
- Lee E, Karimi F, Wagner C, Jo H-H, Strohmaier M, Galesic M (2019) Homophily and minority-group size explain perception biases in social networks. *Nat Hum Behav* 3(10):1078–1087
- Lee SH, Kim P-J, Jeong H (2006) Statistical properties of sampled networks. *Phys Rev E* 73(1):016102
- Leskovec J, Lang KJ, Dasgupta A, Mahoney MW (2009) Community structure in large networks: natural cluster sizes and the absence of large well-defined clusters. *Int Math* 6(1):29–123
- Leskovec J, Faloutsos C (2006) Sampling from large graphs. In: Proceedings of the 12th ACM SIGKDD international conference on knowledge discovery and data mining. ACM, pp 631–636
- Leskovec J, Krevl A (2014) SNAP datasets: Stanford large network dataset collection. <http://snap.stanford.edu/data>
- Lin M, Li W, Nguyen C-t, Wang X, Lu S (2019) Sampling based Katz centrality estimation for large-scale social networks. In: International conference on algorithms and architectures for parallel processing. Springer, pp 584–598
- Maiya AS, Berger-Wolf TY (2010) Sampling community structure. In: Proceedings of the 19th international conference on world wide web. ACM, pp 701–710
- Moody J (2001) Peer influence groups: identifying dense clusters in large networks. *Soc Netw* 23(4):261–283
- Morstatter F, Pfeffer J, Liu H, Carley KM (2013) Is the sample good enough? Comparing data from twitter's streaming API with twitter's firehose. In: Seventh international AAI conference on weblogs and social media
- Murai S, Yoshida Y (2019) Sensitivity analysis of centralities on unweighted networks. In: The world wide web conference. ACM, pp 1332–1342
- Park J, Newman ME (2004) Statistical mechanics of networks. *Phys Rev E* 70(6):066117
- Roddenberry TM, Segarra S (2019) Blind inference of centrality rankings from graph signals. arXiv preprint [arXiv:1910.10846](https://arxiv.org/abs/1910.10846)
- Ruggeri N, De Bacco C (2020) Sampling on networks: estimating eigenvector centrality on incomplete networks. In: Cherifi H, Gaito S, Mendes JF, Moro E, Rocha LM (eds) Complex networks and their applications VIII. Springer, Cham, pp 90–101
- Sadikov E, Medina M, Leskovec J, Garcia-Molina H (2011) Correcting for missing data in information cascades. In: Proceedings of the fourth ACM international conference on web search and data mining. ACM, pp 55–64
- Sakakura Y, Yamaguchi Y, Amagasa T, Kitagawa H (2014) An improved method for efficient pagerank estimation. In: International conference on database and expert systems applications. Springer, pp 208–222
- Segarra S, Ribeiro A (2015) Stability and continuity of centrality measures in weighted graphs. *IEEE Trans Signal Process* 64(3):543–555
- Shao H, Mesbahi M, Li D, Xi Y (2017) Inferring centrality from network snapshots. *Sci Rep* 7(1):1–13
- Stumpf MP, Wiuf C (2005) Sampling properties of random graphs: the degree distribution. *Phys Rev E* 72(3):036118
- Stutzbach D, Rejaie R, Duffield N, Sen S, Willinger W (2009) On unbiased sampling for unstructured peer-to-peer networks. *IEEE/ACM Trans Netw TON* 17(2):377–390
- Takac L, Zabovsky M (2012) Data analysis in public social networks. In: International scientific conference and international workshop present day trends of innovations, vol 1
- Wagner C, Singer P, Karimi F, Pfeffer J, Strohmaier M (2017) Sampling from social networks with attributes. In: Proceedings of the 26th international conference on world wide web, pp 1181–1190
- Watts DJ, Strogatz SH (1998) Collective dynamics of 'small-world' networks. *Nature* 393(6684):440
- Zhang B, Liu R, Massey D, Zhang L (2005) Collecting the internet as-level topology. *ACM SIGCOMM Comput Commun Rev* 35(1):53–61

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.