Systematic assessment of the quality of fit of the stochastic block model for empirical networks

Felipe Vaca-Ramírez [*] and Tiago P. Peixoto [†]
Department of Network and Data Science, Central European University, 1100 Vienna, Austria

(Received 6 January 2022; accepted 19 April 2022; published 18 May 2022)

We perform a systematic analysis of the quality of fit of the stochastic block model (SBM) for 275 empirical networks spanning a wide range of domains and orders of size magnitude. We employ posterior predictive model checking as a criterion to assess the quality of fit, which involves comparing networks generated by the inferred model with the empirical network, according to a set of network descriptors. We observe that the SBM is capable of providing an accurate description for the majority of networks considered, but falls short of saturating all modeling requirements. In particular, networks possessing a large diameter and slow-mixing random walks tend to be badly described by the SBM. However, contrary to what is often assumed, networks with a high abundance of triangles can be well described by the SBM in many cases. We demonstrate that simple network descriptors can be used to evaluate whether or not the SBM can provide a sufficiently accurate representation, potentially pointing to possible model extensions that can systematically improve the expressiveness of this class of models.

DOI: 10.1103/PhysRevE.105.054311

I. INTRODUCTION

The stochastic block model (SBM) [1,2] is an important family of generative network models used primarily for community detection [3] and link prediction [4]. In its simplest formulation, it describes a network formation mechanism where the nodes are divided into discrete groups, and the probability of an edge existing between two nodes is given as a function of their group memberships. Many variations of this idea exist, including mixed-membership SBMs [5], where nodes are allowed to belong to multiple groups, the degree-corrected SBM (DCSBM) [2], where nodes are allowed to possess arbitrary degrees, as well as several extensions to other domains, such as dynamical networks [6–8] and multilayer networks [7,9], to name a few.

SBMs also serve as generalizations of more fundamental random network models. The basic SBM has the Erdős-Rényi model [10] as a special case when there is a single group, and likewise the DCSBM recovers the configuration model [11] in the same situation. However, differently from these more fundamental models, the SBM possesses a set of parameters—the partition of the nodes and the affinities between groups—that is not trivially recoverable from observed networks. These parameters are *latent* information that need to be obtained via inference algorithms, which form the basis of the community detection methods that use this approach [3]. Furthermore, the SBM has a controllable level of complexity: by increasing the number of groups, we have the ability to express increasingly elaborate types of network structures, via arbitrary mixing patterns between the latent groups. In fact, despite its stylized nature, it can be shown that the SBM can approximate a broad class of generative models that are different from it [12], and its inference functions similarly to fitting a histogram to numeric data in order to estimate the underlying probability density—with the node groups playing a similar role to the histogram bins. However, the expressiveness of the SBM is not absolute, especially when the networks are *sparse*, i.e., when their average degree is much smaller than the total number of nodes. In such a situation, there is no guarantee that the SBM is capable of arbitrarily approximating the true underlying model, regardless of how we infer it: By increasing the model complexity we move from a situation where we are *underfitting*, i.e., extracting patterns that do not sufficiently capture all the features of the true model, to a situation where we are *overfitting*, i.e., incorporating randomness into the model description, which is also a deviation from the true model. When we find the most adequate inference that balances statistical evidence against model complexity to prevent overfitting, we might still be missing important features of the true model, simply because it cannot be sufficiently well captured under the SBM parametrization.

Here we are not interested in evaluating the SBM as a plausible generative process of networks across all domains, since it does not represent an ultimately credible mechanism for any of them. Instead, our objective is to assess how capable it is of providing a general *effective* description of empirical networks, and in which aspects and to what extent (and not *whether*) it tends to be misspecified. Understanding the limits of the SBM representation in empirical settings is therefore a nuanced undertaking that is likely to be affected by a variety of possible sources of deviations. Since the SBM tends to yield very good comparative performance in link prediction tasks [13,14], it is therefore known that it tends to outperform alternative models in capturing the structure of networks, but we still lack a more accurate assessment of its qualities and shortcomings in absolute terms.

In this work, we evaluate the quality of fit of the SBM in empirical contexts by performing *model checking* on Bayesian

[*] vaca_felipe@phd.ceu.edu
[†] peixotot@ceu.edu

inferences. Based on a diverse collection of 275 networks spanning various domains and several orders of size magnitude, we compare the values of many network descriptors computed on the observed network with what would be typically obtained with networks sampled from the inferred SBM. In this way, any significant discrepancy can be interpreted as a form of "residual" that points to a shortcoming of the SBM in capturing that particular network property.

Overall we find that the SBM is capable of encapsulating the network structure to a significant degree for a large fraction of the networks studied, but falls short of completely exhausting the modeling requirements in many cases. We find that for networks with very large diameter or a very slow mixing random walk [15] the SBM tends to provide a poor description. This includes, for example, many transportation networks—which are typically embedded in a low-dimensional space—as well as some economic networks.[1] However, for other kinds of networks the quality of fit tends to be good overall.

We proceed with describing in detail the model and inference procedure (Sec. II), our criteria to evaluate the quality of fit (Sec. III), the network corpus used (Sec. IV), and the results of our analysis for it (Sec. V). We finalize in Sec. VI with a conclusion.

II. MODEL AND INFERENCE

For our analysis we will use the microcanonical degree-corrected SBM (DCSBM) [2,17], which combines arbitrary mixing patterns between groups together with arbitrary degree sequences. It has as parameters the partition of the nodes into B groups, $\boldsymbol{b} = \{b_i\}$, with $b_i \in [1, B]$ being the group membership of node i, the degree sequence $\boldsymbol{k} = \{k_i\}$, where k_i is the degree of node i, and the edge counts between groups $\boldsymbol{e} = \{e_{rs}\}$ (or twice that number for $r = s$), given by $e_{rs} = \sum_{ij} A_{ij}\delta_{b_i,r}\delta_{b_i,s}$. Given these constraints, the network is generated with probability [17]

$$P(\boldsymbol{A}|\boldsymbol{k},\boldsymbol{e},\boldsymbol{b}) = \frac{\prod_{r<s} e_{rs}! \prod_r e_{rr}!! \prod_i k_i!}{\prod_{i<j} A_{ij}! \prod_i A_{ii}!! \prod_r e_r!}, \tag{1}$$

where $\boldsymbol{A} = \{A_{ij}\}$ is the adjacency matrix of an undirected multigraph with potential self-loops, and $e_r = \sum_s e_{rs}$.

All the networks we will be studying are undirected simple graphs, for which the above model can give only an approximation. As demonstrated in Ref. [18], the use of multigraph models based on the Poisson distribution (or equivalently, microcanonical models based on the pairing of half-edges, as above) cannot ascribe probabilities to simple edges (i.e., $A_{ij} = 1$) that are larger than $1/\mathrm{e} \approx 0.37$. This limits the applicability of such models on networks with heterogeneous density, due to either broad degree distributions or sufficiently dense communities, which are common properties of empirical networks. To address this limitation, we use the latent multigraph model of Ref. [18], where we assume that an underlying unobserved multigraph $\boldsymbol{A}$ is in fact responsible for the observed simple graph $\boldsymbol{G}$ simply via the removal of the edge multiplicities and self-loops:

$$P(\boldsymbol{G}|\boldsymbol{A}) = \prod_{i<j} \left(1 - \delta_{A_{ij},0}\right)^{G_{ij}} {\delta_{A_{ij},0}}^{1-G_{ij}}. \tag{2}$$

Note that $P(\boldsymbol{G}|\boldsymbol{A})$ can take only a value of 0 or 1, depending on whether $\boldsymbol{G}$ and $\boldsymbol{A}$ are compatible. Via this mathematical construction, the final model

$$P(\boldsymbol{G}|\boldsymbol{k},\boldsymbol{e},\boldsymbol{b}) = \sum_{\boldsymbol{A}} P(\boldsymbol{G}|\boldsymbol{A})P(\boldsymbol{A}|\boldsymbol{k},\boldsymbol{e},\boldsymbol{b}) \tag{3}$$

can express both arbitrary mixing patterns between groups as well as degree correction, without the limitations of the multigraph model for networks with large local densities [18]. The inference of this model is performed by sampling from the posterior distribution

$$P(\boldsymbol{A},\boldsymbol{k},\boldsymbol{e},\boldsymbol{b}|\boldsymbol{G}) = \frac{P(\boldsymbol{G}|\boldsymbol{A})P(\boldsymbol{A}|\boldsymbol{k},\boldsymbol{e},\boldsymbol{b})P(\boldsymbol{k},\boldsymbol{e},\boldsymbol{b})}{P(\boldsymbol{G})}, \tag{4}$$

which remains tractable. Here we use the merge-split Markov chain Monte Carlo (MCMC) algorithm described in Ref. [19] to efficiently sample from this distribution.

Note that for $P(\boldsymbol{k},\boldsymbol{e},\boldsymbol{b})$ we use the nonparametric microcanonical hierarchical priors and hyperpriors described in Refs. [17,20]. Importantly, this kind of approach determines the appropriate model complexity (via the number of groups) according to the statistical evidence available in the data. As has been shown in these previous works, this choice guarantees that only compressive inferences are made in a manner that prevents overfitting (finding a number of groups B that is too large), but also with a substantial protection against underfitting (finding a number that is too small), which tends to happen when noninformative priors are used instead.

In addition to the DCSBM we will also use the configuration model as a comparison, obtained by reshuffling the edges of the obtained network while preserving its degree sequence (here we use the edge-switching MCMC algorithm [11]). We note that the configuration model is an approximate special case of the DCSBM considered above when there is only a single group.[2] Therefore, whenever the Bayesian approach above identifies more than one group with a large probability, this automatically implies a selection of the DCSBM in lieu of the configuration model. This happens for every network that we consider in this work, meaning that the DCSBM is the favored model for all of them. Nevertheless, the configuration model serves as a good baseline to determine to what extent the quality of fit obtained with the DCSBM can be ascribed to the degree sequence alone or to the group-based mixing patterns uncovered.

III. ASSESSING QUALITY OF FIT

The approach we use to assess the quality of fit of the DCSBM is based on obtaining the *posterior predictive distribution* of certain network descriptors. More precisely, for a

[1] See Ref. [16] for a qualitative overview of the different network classifications we consider.

[2] This is only approximately true since the configuration model and the latent Poisson models are not identical, but sufficiently similar for the purposes of this work [18].

scalar network descriptor $f(\boldsymbol{G})$, its posterior predictive distribution is given by

$$P(y|\boldsymbol{G}) = \sum_{\substack{\boldsymbol{G}',\boldsymbol{A}',\boldsymbol{A}\\ \boldsymbol{k},\boldsymbol{e},\boldsymbol{b}}} \delta(y - f(\boldsymbol{G}'))P(\boldsymbol{G}'|\boldsymbol{A}') \times P(\boldsymbol{A}'|\boldsymbol{k},\boldsymbol{e},\boldsymbol{b})P(\boldsymbol{A},\boldsymbol{k},\boldsymbol{e},\boldsymbol{b}|\boldsymbol{G}), \tag{5}$$

where $\delta(x)$ is the Dirac delta function. In other words, for each inferred parameter set $(\boldsymbol{k}, \boldsymbol{e}, \boldsymbol{b})$, weighted according to its posterior probability, we sample a new network $\boldsymbol{G}'$ from the model defined above (which can be done in time $O(E + N)$ where E and N are the total number of edges and nodes, respectively, as we show in Appendix A), and obtain the descriptor value $y = f(\boldsymbol{G}')$.[3]

We can say that a model captures well the value of a descriptor if its predictive posterior distribution ascribes high probability to values that are close to what was observed in the original network. We can obtain a compact summary of the level of agreement in two different ways. The first measures the statistical significance of the deviation, e.g., via the z score [21]

$$z = \frac{f(\boldsymbol{G}) - \langle y\rangle}{\sigma_y}, \tag{6}$$

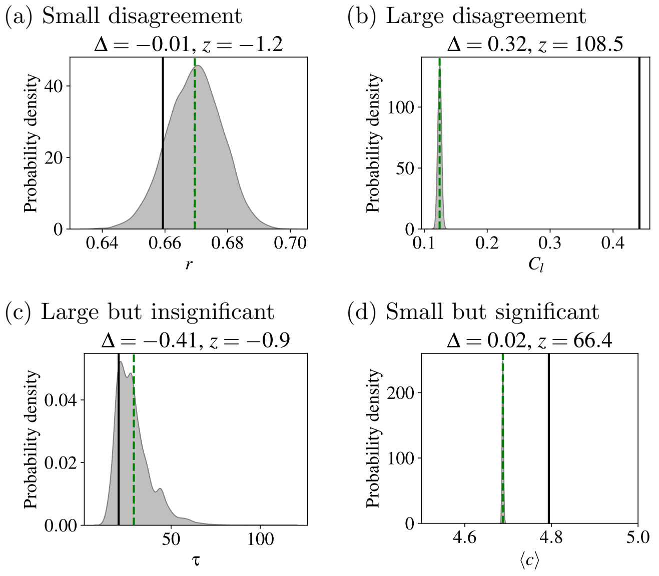
where $\langle y\rangle$ and σ_y are the mean and standard deviation of $P(y|\boldsymbol{G})$. The second criterion is the relative deviation, which here we compute in two different ways,

$$\Delta_1 = \frac{f(\boldsymbol{G}) - \langle y\rangle}{f(\boldsymbol{G})}, \quad \Delta_2 = \frac{f(\boldsymbol{G}) - \langle y\rangle}{f_{\max} - f_{\min}}, \tag{7}$$

depending on whether the descriptor values are bounded in a well-defined interval $[f_{\min}, f_{\max}]$ (Δ_2) or not (Δ_1).

The z score and relative deviation measure complementary aspects of the agreement between data and model, and represent different criteria which should be used together. While a high value of the z score can be used to reject the inferred model as a plausible explanation for the data, by itself it tells us nothing about how good an approximation it is. Conversely, the relative deviation tells us how well the descriptor is being reproduced by the model, but nothing about the statistical significance of the comparison.

In Fig. 1 we show examples that illustrate how the different criteria operate. In Figs. 1(a) and 1(b) we see examples that show good and bad agreements between model and data, respectively, according to both criteria simultaneously. In these cases, the conclusion is unambiguous: we either see no reason whatsoever to condemn the model, or we see a definitive reason to do so. However, in Figs. 1(c) and 1(d) we reach mixed conclusions. In Fig. 1(c) the model typically yields different values than observed in the data, but it still ascribes a large probability to it. We cannot condemn the model as an implausible explanation for the data, but it is conceivable that the true generative model would be more concentrated on the observed value. Conversely, in Fig. 1(d) we see a situation where the model ascribes close to zero probability to the actual descriptor value seen in the data, but, in absolute terms, the discrepancy is quite small. Although we find evidence to condemn the plausibility of the model, we could still claim that it is a good approximation.


(a) Small disagreement

(b) Large disagreement

(c) Large but insignificant

(d) Small but significant

FIG. 1. Examples of posterior predictive distributions for some descriptors (see Table I for definitions) using the DCSBM, together with z score and relative deviation. The solid black line shows the empirical value of the descriptor $f(\boldsymbol{G})$, and the dashed green line the mean of the predictive posterior distribution. In (a) and (b) we see examples where employing both criteria reveal unambiguously good and bad agreements, respectively, between data and model. However, in (c) we see a situation where despite a substantial disagreement with respect to the relative deviation, the z score indicates that the model cannot be discarded as a plausible explanation for the data. In (d) we see a situation where the z score points to decisive rejection of the model, but the small relative deviation allows us to accept it as an accurate approximation.

Overall, since we know that a model like the DCSBM cannot possibly correspond to the true generative model of empirical networks, we should expect that in situations where the network is sufficiently large, and hence there is more abundant data, the values of the z score will tend to be high. Here we argue that since the objective of a model like the DCSBM is to obtain a good approximation of the underlying model, not an exact representation, the ultimate criterion is a combination of the two, where we may deem the model compatible with the data when *either* the z score *or* the relative deviation has a sufficiently low magnitude. For the purpose of clarity and simplicity of our analysis, we will consider the thresholds $|z| = 3$ and $|\Delta| = 0.05$ as reasonable choices to deem the model compatible with data, although our results will not depend on these particular choices, and we will always report the full range of values.

Before continuing, some important considerations regarding model checking should be made. While an excellent model should fulfill both of the above criteria simultaneously, we need to observe that a model that maximally overfits,

[3] The posterior predictive distribution for the configuration model is analogous, i.e., $P(y|\boldsymbol{G}) = \sum_{G'} \delta(y - f(\boldsymbol{G}'))P(\boldsymbol{G}'|\boldsymbol{k})$, where $\boldsymbol{k}$ are the observed degrees, and $P(\boldsymbol{G}|\boldsymbol{k})$ is the likelihood of the configuration model.

TABLE I. Network descriptors used in this work, with their respective symbol, range of values, and how the relative deviation was computed. More details on how the descriptors are computed are given in Appendix B.

Symbol	Descriptor	Range	Δ
r	Degree assortativity	$[-1, 1]$	Δ_2
$\langle c\rangle$	Mean k-core value	$[0, \infty]$	Δ_1
C_l	Mean local clustering coefficient	[0,1]	Δ_2
C_g	Global clustering coefficient	[0,1]	Δ_2
λ_1^A	Leading eigenvalue of the adjacency matrix	$[0, \infty]$	Δ_1
λ_1^H	Leading eigenvalue of the Hashimoto matrix	$[0, \infty]$	Δ_1
τ	Characteristic time of a random walk	$[0, \infty]$	Δ_1
$\varnothing$	Pseudodiameter	$[1, \infty]$	Δ_1
R_r	Node percolation profile (random removal)	$[0, 1/2]$	Δ_2
R_t	Node percolation profile (degree-targeted removal)	$[0, 1/2]$	Δ_2
S	Fraction of nodes in the largest component	$[0, 1]$	Δ_2

i.e., ascribes to the observed network a probability of one, and to any other a probability of zero, will achieve the best possible performance according to both relative deviation and statistical significance. This occurs because we are using the same data to perform both the model inference and evaluate its quality, which is an invalid approach for *model selection*. Therefore, it is important to recognize the crucial difference between model checking and model selection: the latter attempts to find the model alternative that is better justified according to statistical evidence, while the former simply finds systematic discrepancies between the inferred model and data. In our analysis, protection against overfitting is obtained via Bayesian inference, and we use model checking only to evaluate the discrepancies (indeed, the fact we find discrepancies to begin with shows that we cannot be massively overfitting). Another observation is that when performing multiple comparison over many networks and descriptors, some amount of "statistically significant" deviations are always expected, even if the models inferred correspond to the true ones, unless we incorporate the fact that we are doing multiple comparisons in our criterion of statistical significance, which would be the methodologically correct approach. We will not perform such a correction in our analysis, because we do not seek to demonstrate the absolute quality of DCSBM as an ultimately plausible hypothesis for network formation. As we will see from our results, such a correction would gain us very little.

Finally, in Table I we list the network descriptors that are used in this work. Our approach requires scalar values, so we constrained ourselves to this category, and furthermore we chose quantities that can be computed quickly, so that robust statistics from the predictive posterior distributions can be obtained. Given these restrictions, we then chose descriptors that measure different aspects of the network structure, both at a local and global levels. Further details on the network descriptors are given in Appendix B.

IV. NETWORK CORPUS

We base our analysis on a corpus containing 275 networks spanning various domains and several orders of size magnitude, as shown in Fig. 2. We have not collected every network at our disposal, but instead chosen networks that are as diverse as possible, in both size and domain, and avoided many networks that are closely related by belonging to the same subset. In Appendix C we give more details about the data sets used.

V. RESULTS

In Fig. 3 we show the summaries of the posterior predictive checks for each descriptor and network, for both models considered. We observe a wide variety of deviation magnitudes, for the same descriptors both across networks and across descriptors. As expected, the DCSBM results show systematically better agreement with the data when compared with the configuration model. Overall, the descriptors that show the

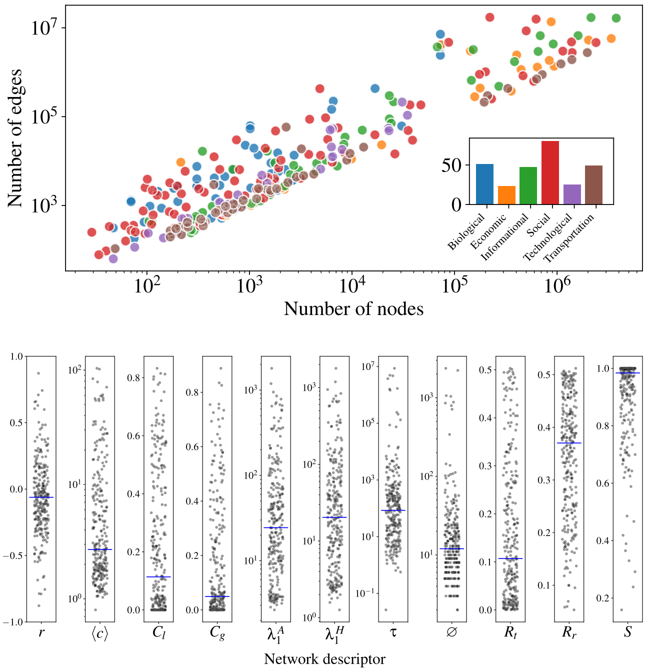


FIG. 2. (Top) Number of nodes and edges for the networks in the corpus used in this work and their domain composition (inset). (Bottom) Distribution of descriptor values for the networks in the corpus. The horizontal line marks the median values.

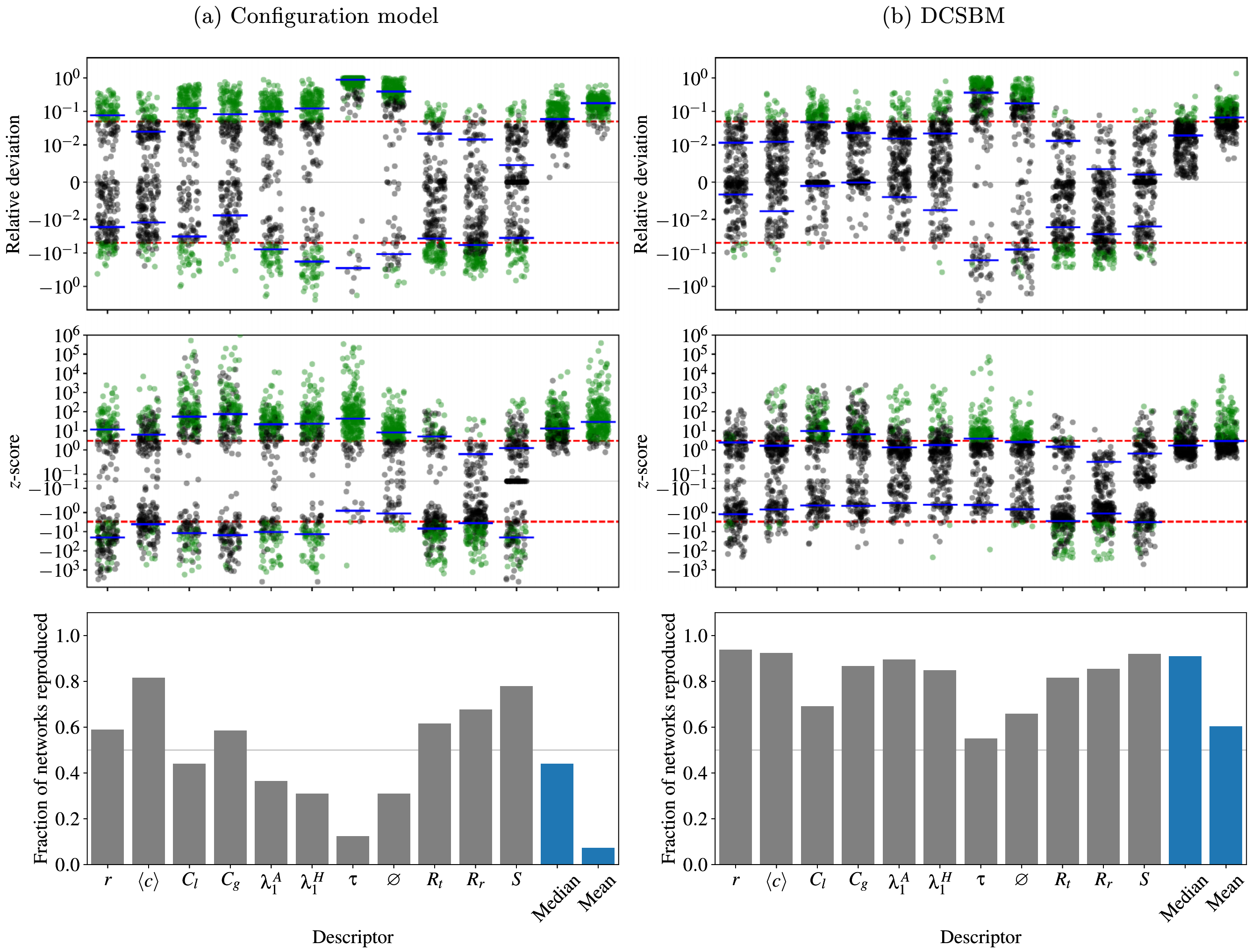


FIG. 3. Distribution of relative deviation (top), z score (middle), and fraction of networks reproduced (bottom) for (a) the configuration model and (b) the DCSBM, according to their respective predictive posterior distributions for each descriptor. We also show the median and mean of the absolute values for all descriptors for each network. The solid blue lines mark the negative and positive median values, and the dashed red line marks the values of $|\Delta| = 0.05$ and $|z| = 3$. The fraction of networks reproduced correspond to those that have the absolute value of either Δ or z below these thresholds. The points in green color correspond to the networks that are not reproduced according to this combined criterion.

worst agreement is the characteristic time of a random walk (τ) and the diameter ($\varnothing$), both of which are particularly high for networks that are embedded in two dimensions, and for which the DCSBM is an inaccurate approximation (more on this below). Nevertheless, there is no single descriptor that the DCSBM does not capture for fewer than 50% of the networks. For descriptors like S, R_r, R_t, and $\langle c \rangle$, the difference between the DCSBM and the configuration model are relatively minor, indicating that those can be captured to a substantial degree by the degree sequence alone.

When considering all descriptors simultaneously for each network, by either the median or mean of the absolute values of the z score and relative deviation, we observe that a substantial majority of the networks considered show good agreement with the DCSBM, as opposed to the small minority that agree with the configuration model. The difference between the median and the mean indicates that there is a sizeable fraction of the networks where the agreement is spoiled by a few outlier descriptors—typically τ and $\varnothing$.

The results obtained by the clustering coefficients are particularly interesting, since it is often the case that they are well reproduced by the DCSBM. This contrasts with what is commonly assumed, namely, that the DCSBM should not be able to capture the abundance of triangles often seen in empirical networks, because in the limit where the number of groups is much smaller than the total number of nodes, the DCSBM becomes locally tree-like [22], with a vanishing probability of forming triangles. Therefore, we may imagine that the situations where there is an agreement with the DCSBM are those where the clustering values are low. However, as we see in Figs. 4(a) to 4(d), this is not quite true, and we observe good agreements even when the clustering values are high. This illustrates a point made in Ref. [23], that it is possible to obtain an abundance of triangles with the SBM simply by increasing the number of groups, in which case it can be explained as a byproduct of homophily. Indeed this is a situation we see in Figs. 4(a) to 4(d), where both the relative deviation and z-score values can be quite small even for extremal values

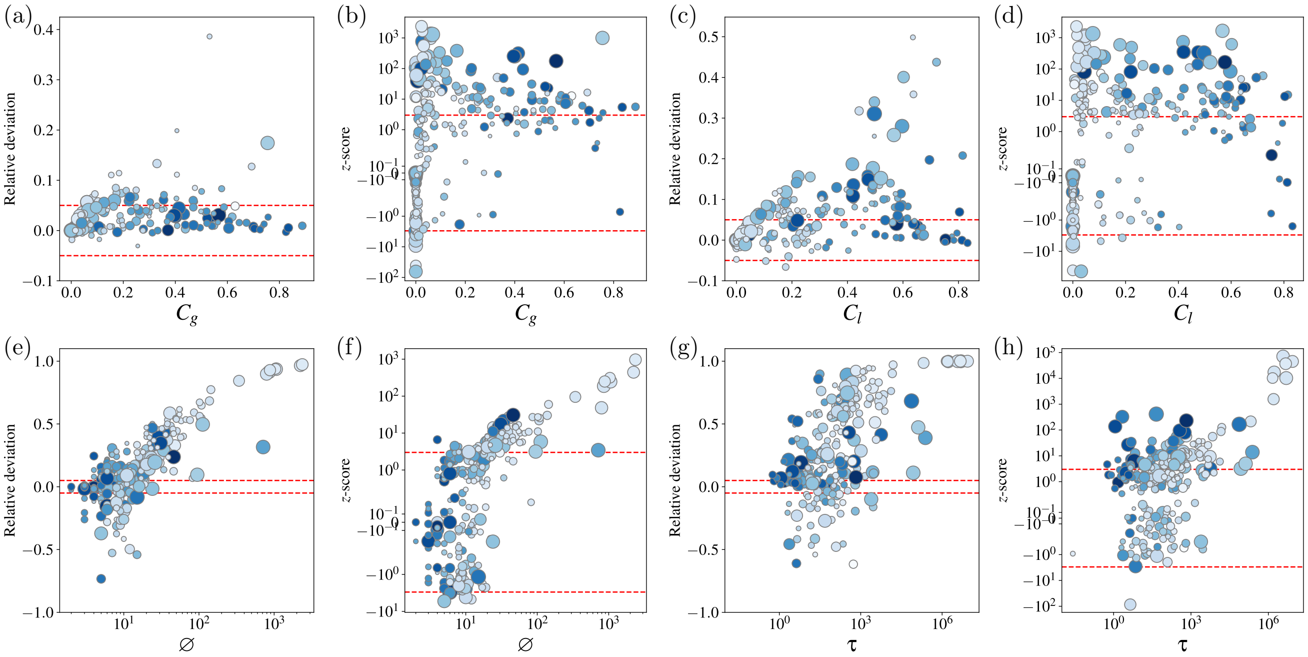


FIG. 4. Relative deviation and z-score values for the global and mean local clustering coefficients, C_g and C_l, as well as diameter and characteristic time of a random walk, $\varnothing$ and τ, as a function of their empirical values, for every network in the corpus, when using the DCSBM. The dashed red line marks the values of $|\Delta| = 0.05$ and $|z| = 3$. The size of the symbol corresponds to the logarithm of the number of edges in the network, and the darkness to the mean degree.

of clustering. However, we do notice a substantial variability between agreements, and a fair amount of instances where the DCSBM cannot capture the observed clustering values, even when they are moderate or even small. This seems to indicate that there are a variety of processes capable of resulting in high clustering values, with homophily being only one of them [23]. Overall, the mean local clustering values tend to be harder to reproduce than the global clustering values. In both cases, the z scores are systematically high, indicating that the clustering values are in general a good criterion to reject the DCSBM as a statistically plausible model, although the relative deviation values tend to be lower than what one would naively expect, meaning that the model can still serve as a reasonably accurate approximation for clustered networks in many cases.

In contrast, we observe a different behavior for the diameter and characteristic time of a random walk, which are the least well reproduced descriptors, as shown in Figs. 4(e) to 4(h). For both these descriptors—which are closely related, since a network with a large diameter will also tend to result in a slow mixing random walk—it is rare to find a network with very high empirical values which the DCSBM is able to accurately describe. Therefore it seems indeed that the DCSBM offers an inadequate ansatz to describe the structure of these networks, even by optimally adjusting its complexity.

In Fig. 5 we show how the model assessment depends on the size of the network. As one could expect, the z-score values tend to increase for larger networks, as more evidence becomes available against the plausibility of the DCSBM as the true generative model. However, the values of the relative deviation do not change appreciably for larger networks, indicating that it remains a good approximation regardless of the size of the system.[4]

In Fig. 6 we show a summary of the fraction of all networks for which we obtain good agreement with either model, according to the network domains. Overall, we see that most domains show similar levels of agreements, except transportation and economic networks. Transportation networks are often embedded in two-dimensional spaces, resulting in large diameters and slow-mixing random walks. The economic networks considered also tend to show large values of these quantities, so the explanation for their discrepancy is the same.

A. Predicting quality of fit

Now we address the question of whether it is possible to predict the quality of fit of both models considered based solely on the empirical values of the networks descriptors. If we can isolate the descriptors which are most predictive, this would give us a general direction in which more accurate models could be constructed.

In order to evaluate the predictability, we frame it as a binary classification problem, where to each network i is ascribed a binary value $y_i = 0$ if we have simultaneously $|z_i| > 3$ and $|\Delta_i| > 0.05$, or otherwise $y_i = 1$. The feature vector for each network is composed of the empirical values of the descriptors, $\boldsymbol{x}_i = (r, \langle c\rangle, C_l, C_g, \lambda_1^A, \lambda_1^H, \tau, \varnothing, R_r, R_t, S, E)$, with the addition of the number of edges E. For each network i, we train a random forest classifier on the entire corpus with

[4] Sampling issues with MCMC could also contribute to the elevated z scores for larger networks, as we discuss in Appendix A.

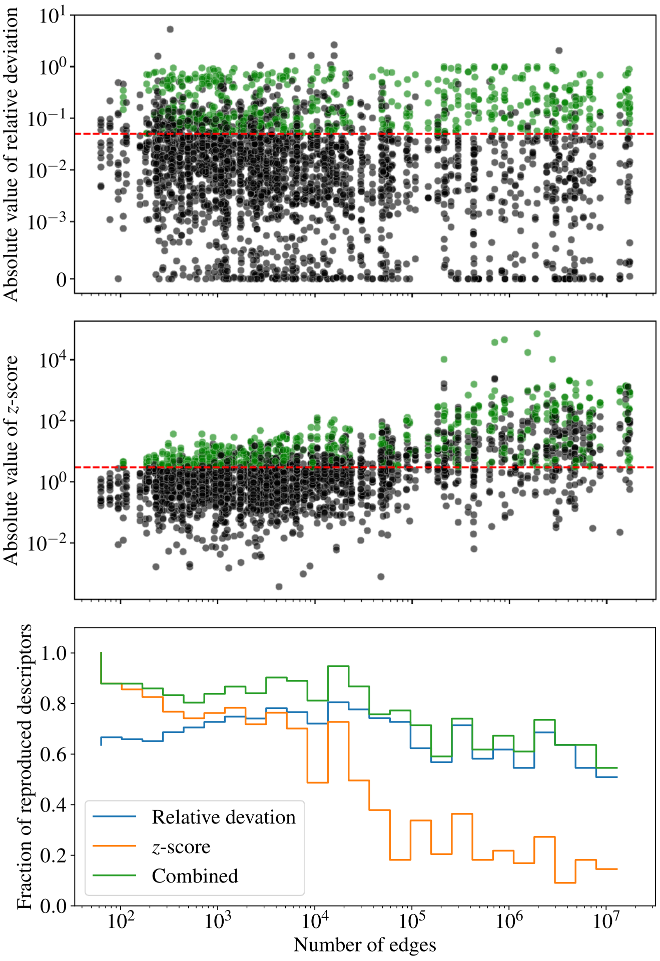


FIG. 5. Absolute value of the relative deviation (top), z score (middle) and fraction of reproduced descriptors (bottom), as a function of the number of edges, for every network in the corpus. The dashed red line marks the values of $|\Delta| = 0.05$ and $|z| = 3$. The fraction of descriptors reproduced correspond to those that have the value of either Δ or z below these thresholds. The points in green color correspond to the descriptors that are not reproduced according to this combined criterion.

that network removed, and evaluate the prediction score on the held-out network. We then repeat this procedure for all networks in the corpus, and evaluate how well the classifier is able to predict the binary label. We present the results of this experiment in Fig. 7 (top) which shows the receiver operating characteristic (ROC) curve, where the true positive rate and the false positive rate are plotted for all threshold values used to reach a classification. The area under the ROC curve (AUC), shown in the legend, can be equivalently interpreted as the probability that a randomly chosen true positive has a prediction score higher than a randomly chosen true negative. For the DCSBM and configuration model, we obtain an AUC value of 0.91 and 0.88, respectively. This indicates a fairly high predictability, from which we can conclude that it is indeed often possible to tell whether the models will provide a good or bad agreement, based only on the descriptor values.

Further insight can be obtained by inspecting the importance of each descriptor in the overall classification. We compute this via the so-called Gini importance [24], defined as the total decrease in node "impurity" (i.e., how often a node in decision tree contributes to a decision), weighted by the proportion of samples that reach that node, averaged over all trees in the classifier.[5] The results can be seen in Figs. 7(b) and 7(c). In both cases, we see that the number of edges is the most predictive descriptor, which is compatible with what we had already seen in Fig. 5, namely, that the larger the networks are, the easier it becomes to reject a model according to the z score. Otherwise, as one would expect, the importance of the remaining descriptors is largely compatible with their reproducibility shown in Fig. 3, where the descriptors that agree the least with the inferred models tend to be the most useful at predicting quality of fit beforehand.

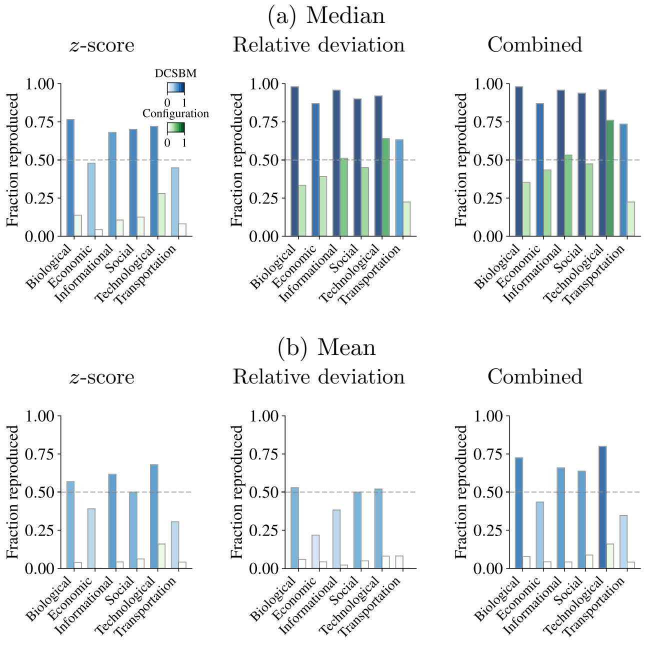


FIG. 6. Fraction of reproduced networks according to their domain, considering the (a) median and (b) mean values of either the z score, the relative deviations, or their combined values, for both models (as shown in the legend). When the combined values are used, this means that a model is deemed compatible with a network when we obtain either $|\Delta| < 0.05$ or $|z| < 3$.

This analysis allows us to emphasize two points: the characteristic time of a random walk τ and the diameter $\varnothing$, both extremal quantities of the network structure that are closely related, are the most difficult descriptors to be captured by the DCSBM. Therefore, an extension of the model that would cater for these properties would bring the most benefit across all networks. However, beyond these two descriptors, there is no substantial difference between the ones that remain, indicating that there is no obvious direction that would bring a systematic modeling improvement over all networks. On the other hand, as we show in Appendix B, the descriptor values and their predictive posterior deviations show nontrivial correlations, which means that if some of them are specifically

[5] We also computed different a measure, called permutation importance, which leads to very similar results (not shown).

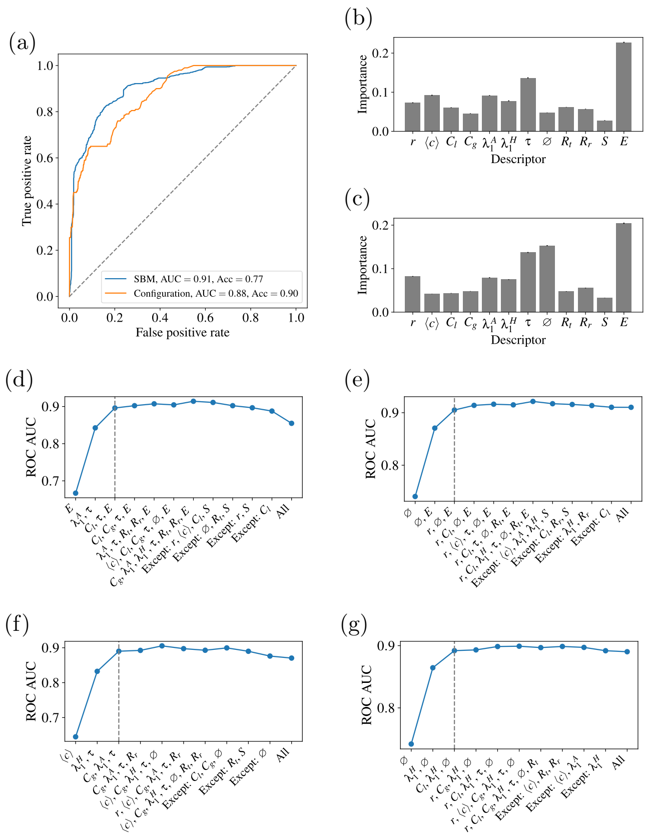


FIG. 7. Predictiveness of the quality of fit of the generative models considered, according to the empirical descriptor values, framed as a binary classification problem, as described in the text. (a) ROC curve for a leave-one-out random-forest classifier, (b) Gini feature importance for the configuration model, (c) same as (b) but for the DCSBM. Panels (d) and (e) show the best ROC AUC obtained for a set of descriptors of a given size, for the configuration model and DSCBM, respectively. Panels (f) and (g) show the same as (d) and (e), respectively, but with the number of edges excluded from the analysis.

targeted, it could potentially improve the quality of fit of other descriptors.

In order to understand what is the minimal amount of information required to predict the suitability of both models, and in this way remove the redundancy provided by the different descriptors, we computed the best ROC AUC obtained by a combination of descriptors of a given size, as shown in Figs. 7(d) and 7(e). In both cases we see that the predictability is saturated by only few descriptors.[6] In the case of the configuration model most of the predictability is already achieved by a combination of (C_l, τ, E). For the DCSBM we get instead $(r, \varnothing, E)$. If we remove the number of edges from the set of features (since it is not informative on the actual network structure), we obtain instead (C_g, λ_1^A, τ) and $(C_l, \lambda_1^H, \varnothing)$, for the configuration model and DCSBM, respectively. It should be emphasized that if a descriptor does not appear in the minimal set this does not mean it is not predictive of the quality of fit, only that it offers largely redundant information in that regard. Thus, for both models if we replace $\varnothing$ with τ or λ_1^H with λ_1^A, etc, we get similar results. This suggests that, besides spatial embeddedness (which influence $\varnothing$ and τ the most), the addition of explicit mechanisms for triangle formation (which affects $C_g, C_l, \lambda_1^H, \lambda_1^A$ directly) might improve the overall expressiveness of the DCSBM—which in fact has been observed in a more limited data set [23].

VI. CONCLUSION

We performed a systematic analysis of posterior predictive checks of the SBM on a diverse corpus of empirical networks, spanning a broad range of sizes and domains. Using a variety of network descriptors, we observed that the SBM is able to accurately capture the structure of the majority of networks in the corpus. The types of networks that show the worst agreement with DCSBM tend to possess a large diameter and a slow mixing of random walks—features that are commonly associated with a low-dimensional spatial embedding, and a violation of the "small-world" property. For the other kinds of networks the agreement tends to be fairly good, even for many networks with an abundance of triangles, in contradiction to what is commonly assumed to be possible with this class of models.

We have also identified the minimal set of network descriptors capable of predicting the quality of fit of the SBM, which is composed of the network diameter and characteristic time of a random walk as the most important, followed by clustering as a secondary feature. This points to the most productive directions in which this class of models could be improved.

It is worth emphasizing that the consistency analysis that we have performed, which compares *a posteriori* the modeling assumptions with the actual properties seen in the data, is possible only if these assumptions are made explicitly via a generative model. Community detection methods that are only descriptive in nature (such as modularity maximization [25]) cannot be used for these purposes. Not only are these methods not guided by statistical evidence and prone to systematic overfitting, but they also provide no direct way to scrutinize the validity of their implicit assumptions [26].

One of the limitations of our analysis is that it is conditioned on the set of descriptors used, and thus shortcomings or successes of the model with respect to other properties not analyzed are not uncovered. A natural extension of our work would be to consider an even broader set of descriptors that could reveal more relevant dimensions for the comparison. This kind of analysis is open ended, as there is no short supply of possible network descriptors. We hope our work will motivate further study in this direction, and with a larger variety of generative models within or beyond the SBM family.

[6] Since we optimized exhaustively for all descriptor combinations of a given size, care should be taken to avoid overfitting, despite the leave-one-out cross-validation, because the optimization was performed on the same set of networks. Because of this, we consider always the smallest set of descriptors that reaches a ROC AUC close to the optimum, not the actual optimum, which is likely to be overfitting.

ACKNOWLEDGMENT

The computational results presented have been achieved using the Vienna Scientific Cluster (VSC).

APPENDIX A: POSTERIOR PREDICTIVE SAMPLING

As described in the main text, we obtain samples from the posterior predictive distribution of Eq. (5) by first sampling from the posterior distribution of Eq. (4) using MCMC and then generating new networks from the inferred models. More specifically, we sample $(\boldsymbol{A}, \boldsymbol{k}, \boldsymbol{e}, \boldsymbol{b})$ from

$$P(\boldsymbol{A}, \boldsymbol{k}, \boldsymbol{e}, \boldsymbol{b}|\boldsymbol{G}) = \frac{P(\boldsymbol{G}|\boldsymbol{A})P(\boldsymbol{A}|\boldsymbol{k}, \boldsymbol{e}, \boldsymbol{b})P(\boldsymbol{k}, \boldsymbol{e}, \boldsymbol{b})}{P(\boldsymbol{G})}, \qquad \text{(A1)}$$

using the merge-split MCMC of Ref. [19], together with the agglomerative initialization heuristic of Refs. [20,27] and the multigraph edge moves of Ref. [18]. For networks of size up to $E = 10^5$ edges we observe good equilibration of the MCMC runs, but for large networks it becomes too slow. For these large networks we settle for a point estimate of the partition $\boldsymbol{b}$ obtained by several runs of the initialization algorithm and keeping the best result, and then we equilibrate the chain according to $\boldsymbol{A}$ alone (which affects $\boldsymbol{k}$ and $\boldsymbol{e}$), which tends to happen quickly. We have verified that performing this calculation several times yields very similar results. The only noticeable outcome of this shortcut for larger networks is that it tends to reduce the variance of the posterior predictive distributions, which can potentially contribute to the elevated z scores we obtained in our analysis. However, since the relative deviation values we obtained did not seem to depend on the size of the network, this gives us confidence that this approach does not introduce significant biases.

Given a sample $(\boldsymbol{A}, \boldsymbol{k}, \boldsymbol{e}, \boldsymbol{b})$, we are interested only in $(\boldsymbol{k}, \boldsymbol{e}, \boldsymbol{b})$ (and hence samples from their marginal distribution), so we discard $\boldsymbol{A}$ and sample a new multigraph $\boldsymbol{A}'$ from the model of Eq. 1. This can be done exactly with an efficient algorithm that works similarly to what was proposed in Refs. [28,29], but is valid for the microcanonical model: Given the parameters $(\boldsymbol{k}, \boldsymbol{e}, \boldsymbol{b})$ we proceed by creating for each group r a multiset of candidate nodes $\boldsymbol{v}_r$, containing k_i copies of each node i with $b_i = r$. Then, for each group pair (r, s) with $r \leqslant s$ and $e_{rs} > 0$, we repeat the following three steps for an e_{rs} number of times (or $e_{rs}/2$ if $r = s$):

(1) We sample a node i from the multiset $\boldsymbol{v}_r$ uniformly at random, and we remove it from the multiset.

(2) We sample a node j from the multiset $\boldsymbol{v}_s$ uniformly at random, and we remove it from the multiset.

(3) We add an edge (i, j) to $\boldsymbol{A}$ (i.e., increment A_{ij} by one, or two if $i = j$).

The resulting multigraph $\boldsymbol{A}$ is sampled exactly with a probability given by Eq. (1). Since the number of nonzero entries of $\boldsymbol{e}$ cannot be larger than the total number of edges E, the whole algorithm finishes in time $O(N + E)$, where N is the number of nodes.

Given a sample $\boldsymbol{A}$, we obtain a simple graph $\boldsymbol{G}$ simply by removing all self-loops and truncating the edge multiplicities:

$$G_{ij} = \begin{cases} 1, & \text{if } A_{ij} > 0 \text{ and } i \neq j, \\ 0, & \text{otherwise.} \end{cases} \qquad \text{(A2)}$$

Finally, given $\boldsymbol{G}$ we compute the network descriptor $f(\boldsymbol{G})$ of interest.

A C + + implementation of every algorithm used in this analysis is freely available as part of the `graph-tool` library [30].

APPENDIX B: NETWORK DESCRIPTORS

Below are the definitions of the descriptors used in our analyses.

Degree assortativity, r: Defined as [31]

$$r = \frac{\sum_{kk'} kk'(m_{kk'} - m_k m_{k'})}{\sigma_k \sigma_{k'}},$$

where $m_{kk'}$ is the fraction of edges with endpoints of degree k and k', $m_k = \sum_{k'} m_{kk'}$, and σ_k is the standard deviation of m_k.

Mean k core, $\langle c\rangle$: The k core is a maximal set of vertices such that its induced subgraph only contains vertices with degree larger than or equal to k. The k-core value c_i of node i is the largest value of k for which i belongs to the k core. The mean value is then

$$\langle c\rangle = \frac{1}{N}\sum_i c_i.$$

This can be computed in time $O(N + E)$ according to the algorithm of Ref. [32].

Mean local clustering coefficient, C_l: The local clustering coefficient [33] of node i is given by

$$C_i = \frac{\sum_{jk} G_{ij}G_{ki}G_{jk}}{k_i(k_i - 1)}.$$

It measures the fraction of pairs of neighbors that are also connected. The mean value is then just

$$C_l = \frac{1}{N}\sum_i C_i.$$

Global clustering coefficient, C_g: The global clustering coefficient of is given by

$$C_g = \frac{\sum_{ijk} G_{ij}G_{ki}G_{jk}}{\sum_i k_i(k_i - 1)}.$$

It measures the fraction of connected triads that close to form a triangle.

Leading eigenvalue of adjacency matrix, λ_1^A: The leading eigenvalue of the adjacency matrix is the largest value of λ which solves

$$\boldsymbol{G}\boldsymbol{x} = \lambda\boldsymbol{x},$$

where $\boldsymbol{x}$ is the associated eigenvector.

Leading eigenvalue of Hashimoto matrix, λ_1^H: The leading eigenvalue of the Hashimoto (a.k.a. nonbacktracking) matrix [34] is the largest value of λ which solves

$$\boldsymbol{H}\boldsymbol{x} = \lambda\boldsymbol{x},$$

where $\boldsymbol{x}$ is the associated eigenvector, and $\boldsymbol{H}$ is an asymmetric $E \times E$ matrix with entries defined as

$$H_{k\to l, i\to j} = \begin{cases} 1 & \text{if } G_{kl} = G_{ij} = 1, l = i, k \neq j, \\ 0 & \text{otherwise.} \end{cases}$$

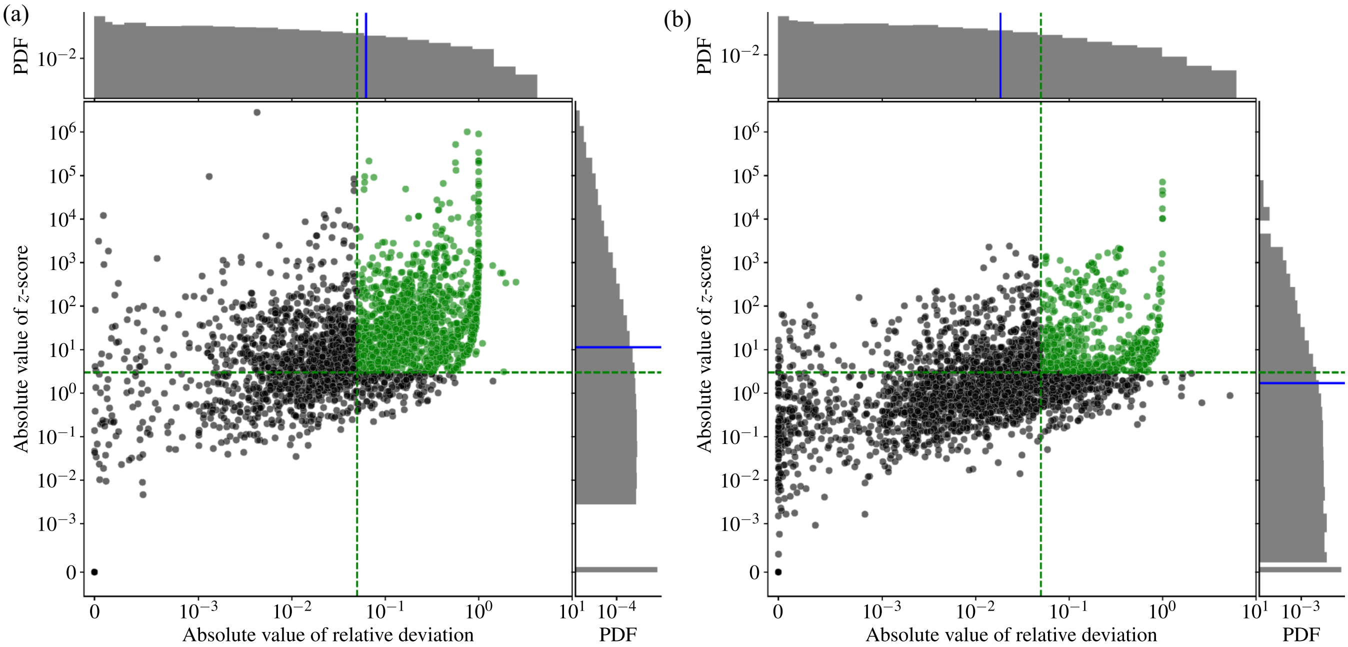


FIG. 8. Absolute value of the z score versus absolute value of relative deviation, for every descriptor value and network in the corpus, according to (a) the configuration model and (b) the DCSBM. The dashed lines mark the values $|z| = 3$ and $|\Delta| = 0.05$, and the histograms the marginal distributions. The solid blue lines mark the median values.

Characteristic time of a random walk, τ: The characteristic time of a random walk is obtained via the second largest eigenvalue $\lambda_2^T \in [0, 1]$ of the transition matrix $\boldsymbol{T}$, with entries

$$T_{ij} = \frac{G_{ij}}{k_j},$$

where $k_i = \sum_j G_{ji}$. It is defined as

$$\tau = -\ln \lambda_2^T.$$

If the network is disconnected, we compute τ only on the largest component.

Pseudodiameter, $\varnothing$: The pseudodiameter is an approximate graph diameter. It is obtained by starting from an arbitrary source node, and finding a target node that is farthest away from the source. This process is repeated by treating the target as the new starting node, and ends when the graph distance no longer increases. This graph distance is taken to be the pseudodiameter. The algorithm runs in time $O(N + E)$.

If the network is disconnected, $\varnothing$ is taken as the maximum of pseudodiameters of the connected components.

Node percolation profile (random removal), R_r: We chose a random node order and remove nodes sequentially from the graph according to it. If S_i is the fraction of nodes in the largest component after the ith removal, then the profile value is

$$R_r = \frac{1}{N}\sum_i S_i.$$

The value is averaged over several node orderings.

Node percolation profile (targeted removal), R_t: The computation is the same as R_r, but the nodes are always removed in decreasing order of the degree.

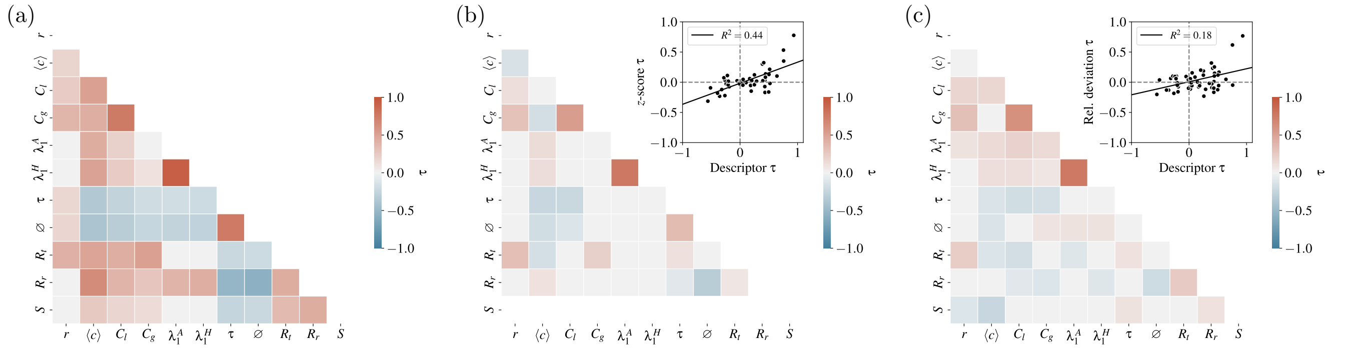


FIG. 9. (a) Kendall’s correlation coefficient τ between pairs of descriptor values across all networks in the corpus. Panels (b) and (c) show the same but for z score and relative deviation values, respectively, according to the DCSBM. The insets show the correlation between coefficients from each respective panel and panel (a).

TABLE II. Descriptions of network data sets.

Name	Description	N	E	Domain
blumenau_drug	A network of drug-drug interactions, extracted from 18 months of electronic health records (EHRs) from the city of Blumenau in southern Brazil [37].	75	181	Biological
budapest_connectome (1)	Brain graphs derived from connectomes of 477 people, computed from the Human Connectome Project [38].	1015	53 586	Biological
budapest_connectome (2)		1015	62 552	Biological
celegans_2019 (1)	Networks among neurons of both the adult male and adult hermaphrodite worms *C. elegans*, constructed from electron microscopy series, to include directed edges (chemical) and undirected (gap junction), and spanning including nodes for muscle and nonmuscle end organs [39].	514	2832	Biological
celegans_2019 (2)		575	4500	Biological
celegans_2019 (3)		454	4172	Biological
celegans_2019 (4)		469	1433	Biological
celegans_interactomes (1)	Ten networks of protein-protein interactions in *C. elegans* (nematode), from yeast two-hybrid experiments, biological process maps, literature curation, orthologous interactions, and genetic interactions [40].	2724	13 564	Biological
celegans_interactomes (2)		912	22 738	Biological
celegans_interactomes (3)		537	517	Biological
collins_yeast	Network of protein-protein interactions in *S. cerevisiae* (budding yeast), measured by co-complex associations identified by high-throughput affinity purification and mass spectrometry (AP/MS) [41].	1622	9070	Biological
ecoli_transcription (1)	Network of operons and their pairwise interactions for *E. coli* [42].	423	519	Biological
foodweb_baywet	Networks of carbon exchanges among species in the cypress wetlands of south Florida, USA . One network covers the wet and the other the dry season [43].	128	2075	Biological
foodweb_little_rock	A food web among the species found in Little Rock Lake in Wisconsin, USA [44].	183	2434	Biological
fresh_webs (1)	Trophic-level species interactions in streams in New Zealand and Maine and North Carolina, USA [45].	94	424	Biological
fresh_webs (2)		107	965	Biological
genetic_multiplex (1)	Multiplex networks representing different types of genetic interactions, for different organisms. Layers represent (i) physical, (ii) association, (iii) colocalization, (iv) direct, and (v) suppressive, (vi) additive, or synthetic genetic interaction [46].	2640	3677	Biological
genetic_multiplex (2)		1005	1155	Biological
genetic_multiplex (3)		313	325	Biological
genetic_multiplex (4)		6570	223 542	Biological
human_brains (1)	Networks of neural interactions extracted from human patients using the Magnetic Resonance One-Click Pipeline (MROCP), where nodes are voxels of neural tissue and edges represent connections by single fibers [47].	1215	13 768	Biological
human_brains (2)		200	1231	Biological
human_brains (3)		139	873	Biological
human_brains (4)		1771	3645	Biological
human_brains (5)		1105	19 543	Biological
human_brains (6)		1527	3939	Biological
human_brains (7)		70	1219	Biological
human_brains (8)		200	2808	Biological
human_brains (9)		70	1301	Biological
human_brains (10)		1632	5218	Biological
human_brains (11)		16 783	430 493	Biological
human_brains (12)		72 783	2 411 659	Biological
human_brains (13)		72 783	3 720 694	Biological
human_brains (14)		72 783	4 205 222	Biological
human_brains (15)		72 783	7 175 769	Biological

(*Continued.*)

Name	Description	N	E	Domain
`interactome_figeys`	A network of human proteins and their binding interactions [48].	2239	6432	Biological
`interactome_yeast`	A network of protein-protein binding interactions among yeast proteins [49].	1870	2203	Biological
`kegg_metabolic` (1)	Metabolic networks of various species, as extracted from the Kyoto Encyclopedia of Genes and Genomes (KEGG) database in March 2006 [50].	1031	2485	Biological
`kegg_metabolic` (2)		1917	5803	Biological
`kegg_metabolic` (3)		505	1144	Biological
`macaque_neural`	A network of cortical regions in the Macaque cortex [51].	47	313	Biological
`malaria_genes` (1)	Networks of recombinant antigen genes from the human malaria parasite *P. falciparum* [52].	307	2684	Biological
`malaria_genes` (2)		307	3961	Biological
`malaria_genes` (3)		307	7579	Biological
`malaria_genes` (4)		307	2812	Biological
`messal_shale`	A network of feeding links among taxa based on the 48-million-yr-old uppermost early Eocene Messel Shale [53].	700	6395	Biological
`nematode_mammal`	A global interaction web of interactions between nematodes and their host mammal species, extracted from the helminthR package and data set [54].	30 516	61 597	Biological
`plant_pol_kato`	A bipartite network of plants and pollinators from Kyoto University Forest of Ashu, Japan, from 1984 to 1987 [55].	772	1206	Biological
`plant_pol_robertson`	A bipartite network of plants and pollinators, from southwestern Illinois, USA [56].	1884	15 255	Biological
`reactome`	A network of human proteins and their binding interactions, extracted from Reactome project [57].	6327	146 160	Biological
`yeast_transcription`	Network of operons and their pairwise interactions, via transcription factor-based regulation, within the yeast *S. cerevisiae* [58].	916	1081	Biological
`amazon_copurchases`	Network of items for sale on amazon.com and the items they “recommend” [155].	403 394	2 443 408	Economic
`amazon_ratings`	A bipartite network of users and products on Amazon.com [156].	3 376 972	5 743 258	Economic
`bookcrossing`	Bipartite network representing people and the books they have interacted with, from the BookCrossing website [157].	445 801	1 149 739	Economic
`corporate_directors`	Bipartite network of directors and the companies on whose boards they sit, spanning 54 countries worldwide, constructed from data collected by the *Financial Times* [158].	356 638	376 918	Economic
`dbpedia_starring`	A bipartite network of movies and the actors that played in them, as extracted from Wikipedia by the DBpedia project [112].	157 184	281 396	Economic
`dbpedia_team`	Bipartite network of the affiliations (employment relations) between professional athletes and their teams, as extracted from Wikipedia by the DBpedia project [112].	935 627	1 366 466	Economic
`discogs_affiliation`	A large bipartite network of the affiliations (contractual relations) among musical artists and record labels, as given in the discogs.com database [63].	2 025 594	5 302 276	Economic
`epinions`	A bipartite network of users and the products they rated on the website Epinions.com [159].	876 252	13 668 320	Economic
`eu_procurements`	A bipartite network of public EU procurement contracts, from 2008 to 2016, between issuing buyers (public institutions such as a ministry or city hall) and supplying winners (a private firm) [160].	839 824	1 841 009	Economic
`eu_procurements_alt` (1)	Networks representing the annual national public procurement markets of 26 European countries from 2008 to 2016, inclusive [160].	552	588	Economic
`eu_procurements_alt` (2)		585	588	Economic
`eu_procurements_alt` (3)		1038	1009	Economic
`eu_procurements_alt` (4)		1098	1118	Economic
`eu_procurements_alt` (5)		2189	2320	Economic
`eu_procurements_alt` (6)		1656	3132	Economic
`eu_procurements_alt` (7)		2097	2518	Economic

(*Continued.*)

Name	Description	*N*	*E*	Domain
eu_procurements_alt (8)		9877	11 185	Economic
eu_procurements_alt (9)		19 438	23 191	Economic
fao_trade	Multiplex network representing trade relationships between countries from the Food and Agricultural Organization of the United Nations [161].	214	9420	Economic
github	The bipartite project-user membership network of the software development hosting site GitHub [162].	177 386	440 237	Economic
jester	Two bipartite networks of users and jokes, extracted from the online joke recommender system Jester [163].	73 521	4 136 360	Economic
stackoverflow	A bipartite network of users and the posts they have favorited, from the online Q&A site Stack Overflow [63].	641 876	1 301 942	Economic
digg_votes	A bipartite network between users and stories on digg.com from 2009 [164].	142 962	3 010 898	Economic
adjnoun	A network of word adjacencies of common adjectives and nouns in the novel *David Copperfield* by Charles Dickens [68].	112	425	Informational
bag_of_words	Five text collections in the form of bags-of-words [69,70].	67 963	3 710 420	Informational
baidu	Four networks from Chinese online encyclopedias Baidu [71].	2 141 300	17 014 946	Informational
berkstan_web	The web graph of the University of California at Berkeley and Stanford universities [72].	685 231	6 649 470	Informational
bible_nouns	A network of noun phrases (places and names) in the King James Version of the Bible [73].	1773	9131	Informational
citeseer	Citations among papers indexed by the CiteSeer digital library [74].	384 413	1 736 145	Informational
cora	Citations among papers indexed by CORA, from 1998, an early computer science research paper search engine [75].	23 166	89 157	Informational
dblp_cite	Citations among papers contained in the DBLP computer science bibliography [76].	12 590	49 636	Informational
dbtropes_feature	A bipartite network of artistic works (movies, novels, etc.) and their tropes (stylistic conventions or devices), as extracted from tvtropes [63].	152 093	3 232 134	Informational
discogs_label	Two bipartite networks of the affiliations between musical labels and either musical genres or musical "styles," as given in the discogs.com database [63].	270 786	481 661	Informational
edit_wikibooks (1)	Two bipartite user-page networks extracted from Wikipedia, about books [77].	1162	1213	Informational
edit_wikibooks (2)		1584	1748	Informational
edit_wikibooks (3)		7177	7732	Informational
edit_wikinews (1)	Two bipartite user-page networks extracted from Wikipedia, about news events [77].	2511	4986	Informational
edit_wikinews (2)		4523	8891	Informational
edit_wikinews (3)		5541	10 545	Informational
edit_wikinews (4)		2208	2753	Informational
edit_wikinews (5)		4457	5942	Informational
edit_wikiquote (1)	A bipartite user-page network extracted from Wikiquotes [77].	270	243	Informational
edit_wikiquote (2)		1041	1109	Informational
edit_wikiquote (3)		704	800	Informational
edit_wikiquote (4)		1333	2731	Informational
edit_wikiquote (5)		625	823	Informational
edit_wiktionary (1)	Three bipartite user-page networks extracted from Wiktionary, for French, German, and English [77].	271	285	Informational
edit_wiktionary (2)		289	276	Informational
edit_wiktionary (3)		1271	1270	Informational
edit_wiktionary (4)		8552	34 589	Informational
edit_wiktionary (5)		3016	6263	Informational
google_web	A web graph representing a crawl of a portion of the general WWW, from a 2002 Google Programming contest [72].	916 428	4 322 051	Informational
movielens_100k	Three bipartite networks that make up the MovieLens 100K data set, a stable benchmark data set of 100 000 ratings from 1000 users on 1700 movies [78].	24 129	71 154	Informational

(*Continued.*)

Name	Description	*N*	*E*	Domain
polblogs	A directed network of hyperlinks among a large set of ones in the USA [79].	1490	16 715	Informational
polbooks	A network of books about the USA [80].	105	441	Informational
scotus_majority (1)	Network of legal citations by the U.S. Supreme Court (SCOTUS) [81,82].	25 417	216 456	Informational
trec_web	A web graph network originally constructed in 2003 as a test bed for information-retrieval techniques, including web search engines [83].	1 601 787	6 679 248	Informational
unicodelang	A bipartite network of languages and the countries in which they are spoken, as estimated by Unicode [63].	868	1255	Informational
us_patents	Citations among patents in the USA, as found in the National Bureau of Economic Research (NBER) database, from 1975 to 1999 [84].	3 774 768	16 518 947	Informational
webkb (1)	Web graphs crawled from four computer science departments in 1998, with each page manually classified into one of seven categories: course, department, faculty, project, staff, student, or other [85].	286	493	Informational
webkb (2)		433	954	Informational
webkb (3)		300	565	Informational
webkb (4)		349	696	Informational
webkb (5)		348	16 625	Informational
wiki_science	A network of scientific fields, extracted from the English Wikipedia in early 2020 [86].	687	6523	Informational
word_adjacency (1)	Networks of word adjacency in texts of several languages including English, French, Spanish, and Japanese [87].	8325	23 841	Informational
word_adjacency (2)		2704	7998	Informational
word_assoc	A network of word associations showing the count of such associations as collected from subjects, from the Edinburgh Associative Thesaurus (EAT) [88].	23 132	297 094	Informational
wordnet	A network of English words from the WordNet, denoting relationships between words (synonymy, hyperonymy, meronymy, etc.) [89]	146 005	656 999	Informational
yahoo_ads	A network of words extracted from phrases on which advertisers bid, in Yahoo! advertisements [63].	653 260	2 931 698	Informational
7th_graders	A network of friendships among 29 seventh-grade students in Victoria, Australia [101].	29	250	Social
academia_edu	Snapshot of the follower relationships among users of academia.edu, a platform for academics to share research papers, scrapped in 2011 [102].	200 169	1 022 441	Social
add_health (1)	A directed network of friendships obtained through a social survey of high school students in 1994. The ADD HEALTH data are constructed from the in-school questionnaire; 90 118 students representing 84 communities took this survey in 1994–1995 [103].	900	1648	Social
add_health (2)		1929	7035	Social
add_health (3)		1282	3487	Social
add_health (4)		111	378	Social
add_health (5)		74	358	Social
add_health (6)		624	1745	Social
add_health (7)		1755	4017	Social
arxiv_authors (1)	Scientific collaborations between authors of papers submitted to arxiv.org [104].	26 197	14 484	Social
arxiv_collab	Collaboration graphs for scientists, extracted from the arXiv (physics) [105].	8361	15 751	Social
bitcoin_alpha	A network of who-trusts-whom relationships among users of the Bitcoin Alpha platform [106].	3783	14 124	Social

(*Continued.*)

Name	Description	N	E	Domain
`ceo_club`	A bipartite network of the memberships of chief executive officers and the social organizations (clubs) to which they belong, from the Minneapolis–St. Paul, Minnesota, USA area [107].	40	95	Social
`chess`	A network among chess players (nodes) giving the chess match outcomes (edges), for game-by-game results among the world's top chess players [108].	7301	55 899	Social
`copenhagen` (1)	A network of social interactions among university students within the Copenhagen Networks Study, over a period of 4 weeks, sampled every 5 minutes [109].	536	621	Social
`copenhagen` (2)		800	6418	Social
`copenhagen` (3)		568	697	Social
`crime`	A network of associations among suspects, victims, and/or witnesses involved in crimes in St. Louis, Missouri, USA in the 1990s [110].	1380	1476	Social
`cs_department`	Multiplex network consisting of five edge types corresponding to online and offline relationships (Facebook, leisure, work, coauthorship, lunch) between employees of the computer science department at Aarhus University, Denmark [111].	61	353	Social
`dbpedia_country`	A bipartite network of the affiliations between notable people and countries of the world, as extracted from Wikipedia via the DBpedia project [112].	592 414	624 402	Social
`dbpedia_occupation`	A bipartite network of the affiliations between notable people and occupations, as extracted from Wikipedia by the DBpedia project [112].	229 307	250 945	Social
`dnc`	A network representing the exchange of emails among members of the Democratic National Committee, USA, in the email data leak released by WikiLeaks in 2016 [63].	2029	10 429	Social
`dolphins`	An undirected social network of frequent associations observed among 62 dolphins (*Tursiops*) in a community living off Doubtful Sound, New Zealand, from 1994 to 2001 [113].	62	159	Social
`ego_social` (1)	Ego networks associated with a set of accounts of three social media platforms (Facebook, Google+, and Twitter) [114].	150	1693	Social
`ego_social` (2)		747	30 025	Social
`ego_social` (3)		452	12 513	Social
`email_company`	A network of emails among employee email addresses at a midsized manufacturing company [115].	167	3250	Social
`email_enron`	The Enron email corpus, containing all the email communication from the Enron corporation, which was made public as a result of legal action [116].	36 692	183 831	Social
`escorts`	A bipartite network of escort and individuals who buy sex from them in Brazil, extracted from a Brazilian online community for such ratings [117].	16 730	39 044	Social
`facebook_friends`	A small anonymized Facebook ego network, from April 2014. Nodes are Facebook profiles, and an edge exists if the two profiles are "friends" on Facebook [118].	362	1988	Social
`facebook_organizations` (1)	Six networks of friendships among users on Facebook who indicated employment at one of the target corporations [119].	320	2369	Social
`facebook_organizations` (2)		165	726	Social
`facebook_organizations` (3)		1429	19 357	Social
`facebook_organizations` (4)		3862	87 324	Social
`facebook_organizations` (5)		5793	30 753	Social
`facebook_organizations` (6)		5524	94 218	Social
`facebook_wall`	Friendship relationships and interactions (wall posts) for a subset of the Facebook social network in 2009, recorded over a 2-yr period [120].	46 952	183 412	Social
`fediverse`	An early snapshot of the federation network among web publishers using the ActivityPub protocol [121].	4860	426 351	Social

(*Continued.*)

Name	Description	*N*	*E*	Domain
`flickr_groups`	Bipartite networks of the affiliations between users and groups on several online social network sites, including Flickr, YouTube, LiveJournal, and Orkut, extracted in 2007 [122].	499 610	8 545 307	Social
`football_tsevans`	A network of American football games between Division IA colleges during the regular season Fall 2000 [123,124].	115	613	Social
`foursquare` (1)	Two bipartite networks of users and restaurant locations in New York City, New York, USA on Foursquare, from 24 October 2011 to 20 February 2012 [125].	6410	9472	Social
`foursquare` (2)		4936	13 472	Social
`highschool`	A network of friendships among male students in a small high school in Illinois, USA, from 1958 [126].	70	274	Social
`hiv_transmission`	A set of networks of HIV transmissions between people through sexual, needle-sharing, or social connections, based on combining eight data sets collected from 1988 to 2001 [127].	35 229	48 889	Social
`hyves`	A network of friendships among users of Hyves, an online social networking site in the Netherlands (comparable to Facebook at the time) [128].	1 402 673	2 777 419	Social
`jazz_collab`	The network of collaborations among jazz musicians and among jazz bands, extracted from the Red Hot Jazz Archive digital database, covering bands that performed between 1912 and 1940 [129].	198	2742	Social
`karate` (1)	Network of friendships among members of a university karate club [130].	34	78	Social
`kidnappings`	Bipartite network of members of the Abu Sayyaf Group in the Philippines, and the kidnapping events they were involved in [131].	351	402	Social
`lastfm` (1)	User-band networks from the music website `last.fm` [132].	175 069	898 062	Social
`lesmis`	The network of scene coappearances of characters in Victor Hugo's novel *Les Miserables*. Edge weights denote the number of such occurrences [91].	77	254	Social
`libimseti`	A network of ratings given between users at Libimseti.cz, a Czech online dating website [133].	220 970	17 233 144	Social
`mislove_osn` (1)	Network structure for four large online social networks [122].	1 138 499	2 990 443	Social
`netscience`	A coauthorship network among scientists working on network science, from 2006. This network is a one-mode projection from the bipartite graph of authors and their scientific publications [68].	1589	2742	Social
`new_zealand_collab`	A network of scientific collaborations among institutions in New Zealand [134].	1511	4273	Social
`petster`	A network of friendships among users on `catster.com` and `dogster.com` [63].	623 766	15 695 166	Social
`physician_trust`	A network of trust relationships among physicians in four Midwestern (USA) cities in 1966 [135].	241	923	Social
`physics_collab`	Coauthorships among the Pierre Auger Collaboration of physicists [136].	514	6482	Social
`reality_mining`	A network of human proximities among students at Massachusetts Institute of Technology (MIT), Cambridge, Massachusetts, USA as measured by personal mobile phones [137].	96	2539	Social
`residence_hall`	A network of friendships among students living in a residence hall at Australian National University, Canberra [138].	217	1839	Social
`sp_high_school` (1)	Contacts and friendship relations between students in a high school in Marseilles, France, in December 2013 [139].	329	348	Social
`sp_high_school` (2)		329	406	Social
`sp_high_school_new`	Network of contacts between students in a high school in Marseilles, France [46].	126	1709	Social
`sp_hypertext`	Network of contacts among attendees of the ACM Hypertext 2009 conference [140].	113	2196	Social
`sp_office`	A temporal network of contacts between individuals, measured in an office building in France, from 24 June to 3 July 2013 [141].	92	755	Social

(*Continued.*)

Name	Description	N	E	Domain
sp_primary_school	Network of contacts among students and teachers at a primary school in Lyon, France, on consecutive days of in October 2009 [142].	242	8317	Social
student_cooperation	Network of cooperation among students in the "Computer and Network Security" course at Ben-Gurion University, Beersheba, Israel, in 2012 [143].	185	311	Social
swingers	A bipartite sexual affiliation network representing "swing unit" couples (one node per couple) and the parties they attended [144].	96	232	Social
twitter	A network of following relationships from Twitter, from a snowball sample crawl across "quality" users in 2009 [145].	465 017	833 540	Social
twitter_15m	A network representing follower-following relations among Twitter users associated with the 15-M Movement or anti-austerity movement in Spain, in the period April–May 2011 [146].	87 569	4 708 274	Social
twitter_higgs	Tweet reply network related to the discovery of the Higgs boson [147].	38 918	29 552	Social
ugandan_village	Complete friendship and health advice social networks among households in 17 rural villages bordering Lake Victoria in Mayuge District, Uganda in 2013 [148].	185	638	Social
us_agencies (1)	Web-based links between U.S. government agencies websites [149].	1796	47 686	Social
us_agencies (2)		234	515	Social
us_congress	Networks of bill co-sponsorship tendencies among U.S. congressmen and -women, from 1973 (93rd Congress) to 2016 (114th Congress) [150,151].	101	3914	Social
wiki_talk (1)	Interactions among users of 10 language-specific Wikipedias [63].	1181	2330	Social
wiki_talk (2)		3144	4098	Social
wikipedia-en-talk	Nodes in the network represent (English) Wikipedia users and a directed edge from node i to node j represents that user i at least once edited a talk page of user j [152].	2 394 385	4 659 565	Social
wikitree	A multigraph network representing child-parent connections among family members, collected in 2012 from WikiTree, an online genealogical website with 13+ million profiles [153].	1 382 751	4 810 045	Social
windsurfers	A network of interpersonal contacts among windsurfers in southern California during the fall of 1986. The edge weights indicate the perception of social affiliations majored by the tasks in which each individual was asked to sort cards with other surfer's name in the order of closeness [154].	43	336	Social
caida_as	Autonomous system (AS) relationships on the Internet, from 2004 to 2007 [59].	8020	18 203	Technological
gnutella (1)	Gnutella peer-to-peer file sharing network from 5–31 August 2002 [60].	6301	20 777	Technological
gnutella (2)		22 687	54 705	Technological
internet_as	A symmetrized snapshot of the structure of the Internet at the level of Autonomous systems (ASs), reconstructed from BGP tables posted by the University of Oregon Route Views Project [61].	22 963	48 436	Technological
internet_top_pop (1)	Assorted snapshots of internet graph at the point of presence (PoP) level (which lies between the IP and AS levels), collected from around the world and at various times. The earliest snapshots are for ARPANET (1969–1972), with a few more from before 2000 [62].	76	115	Technological
internet_top_pop (2)		145	186	Technological
internet_top_pop (3)		47	63	Technological
internet_top_pop (4)		197	243	Technological
internet_top_pop (5)		754	895	Technological
jdk	A network of class dependencies within the JDK (Java SE Development Kit) 1.6 [63].	6434	53 658	Technological
jung	A network of software class dependency within the JUNG 2.0 [64].	6120	50 290	Technological

(*Continued.*)

Name	Description	*N*	*E*	Domain
linux	A network of Linux (v3.16) source code file inclusion [63].	30 837	213 217	Technological
power	A network representing the Western States Power Grid of the USA [33].	4941	6594	Technological
route_views (1)	733 daily network snapshots denoting BGP traffic among autonomous systems (ASs) on the Internet, from the Oregon Route Views Project, spanning 8 November 1997 to 2 January 2000. Data collected by NLANR/MOAT [65].	103	239	Technological
route_views (2)		512	1181	Technological
route_views (3)		767	1734	Technological
route_views (4)		1486	3172	Technological
route_views (5)		6474	12 572	Technological
route_views (6)		6301	12 226	Technological
software_dependencies (1)	Several networks of software dependencies. Nodes represent libraries, and a directed edge denotes a library dependency on another [64,66].	388	514	Technological
software_dependencies (2)		838	1063	Technological
software_dependencies (3)		799	3579	Technological
software_dependencies (4)		550	1153	Technological
software_dependencies (5)		282	505	Technological
software_dependencies (6)		2124	4809	Technological
topology	An integrated snapshot of the structure of the Internet at the level of autonomous systems (ASs), reconstructed from multiple sources, including the RouteViews and RIPE BGP trace collectors, route servers, looking glasses, and Internet Routing Registry databases [67].	34 761	107 720	Technological
chicago_road	A transportation network of Chicago, Illinois, USA, from an unknown date (probably late 20th century) [90].	12 982	20 627	Transportation
contiguous_usa	A network of contiguous states in the USA, in which each state is a node and two nodes are connected if they share a land-based geographic border [91].	49	107	Transportation
eu_airlines	A multiplex network of airline routes among European airports, where each of the 37 edge types represents routes by a different airline [92].	450	2953	Transportation
euroroad	A network of international “E-roads,” mostly in Europe [93].	1174	1417	Transportation
faa_routes	A network of air traffic routes, from the U.S. FAA (Federal Aviation Administration) National Flight Data Center (NFDC) preferred routes database [94].	1226	2408	Transportation
london_transport	Multiplex network with three edge types representing links within the three layers of London train stations: Underground, Overground, and DLR [46].	369	430	Transportation
openflights	A network of regularly occurring flights among airports worldwide, extracted from the openflights.org data set [95].	3214	18 858	Transportation
openstreetmap (1)	The road network for the entire USA, as extracted from the OpenStreetMap project [96].	351	434	Transportation
openstreetmap (2)		354	350	Transportation
openstreetmap (3)		831	923	Transportation
openstreetmap (4)		1603	2188	Transportation
openstreetmap (5)		4240	5102	Transportation
openstreetmap (6)		8904	10 549	Transportation
openstreetmap (7)		724	1048	Transportation
openstreetmap (8)		2371	3295	Transportation
openstreetmap (9)		684	823	Transportation
openstreetmap (10)		500	780	Transportation
openstreetmap (11)		3377	4698	Transportation
openstreetmap (12)		1609	1972	Transportation
openstreetmap (13)		612	688	Transportation
openstreetmap (14)		209 734	297 196	Transportation

(*Continued.*)

Name	Description	N	E	Domain
roadnet (1)	Road networks from three U.S. states (California, Pennsylvania, and Texas), in which edges are stretches of road and vertices are intersections of roads [72].	1 971 281	2 766 607	Transportation
roadnet (2)		1 090 920	1 541 898	Transportation
roadnet (3)		1 393 383	1 921 660	Transportation
urban_streets (1)	Urban street networks, corresponding to 1-square-mile maps of 20 cities around the world [97,98].	179	230	Transportation
urban_streets (2)		240	339	Transportation
urban_streets (3)		467	691	Transportation
urban_streets (4)		248	418	Transportation
urban_streets (5)		697	1084	Transportation
urban_streets (6)		169	271	Transportation
urban_streets (7)		1840	2397	Transportation
urban_streets (8)		1496	2252	Transportation
urban_streets (9)		584	958	Transportation
urban_streets (10)		217	222	Transportation
urban_streets (11)		541	771	Transportation
urban_streets (12)		252	328	Transportation
urban_streets (13)		869	1307	Transportation
urban_streets (14)		192	302	Transportation
urban_streets (15)		210	323	Transportation
urban_streets (16)		2870	4375	Transportation
urban_streets (17)		335	494	Transportation
urban_streets (18)		488	729	Transportation
urban_streets (19)		169	196	Transportation
us_air_traffic	Yearly snapshots of flights among all commercial airports in the USA from 1990 to today [99].	2278	58 228	Transportation
us_roads (1)	The road networks of the 50 U.S. states and the District of Columbia based on U.S. Census 2000 TIGER/Line Files [100].	9559	14 841	Transportation
us_roads (2)		194 505	212 345	Transportation
us_roads (3)		330 386	431 398	Transportation
us_roads (4)		630 639	705 083	Transportation
us_roads (5)		716 215	886 897	Transportation

Fraction of nodes in the largest component, S: A component is a maximal set of nodes that are connected by a path. The largest component is the component with the largest number of nodes, and S is the fraction of all nodes that belong to it.

In Fig. 8 we show how the z scores and relative deviation values are related for every network descriptor, according to both models used. In Fig. 9 we show Kendall's τ correlation coefficient among the descriptor values themselves, as well as their z scores and relative deviations, according to the DCSBM. The insets show how the correlations among the deviations are themselves also correlated with the descriptor correlations.

APPENDIX C: DATA SET DESCRIPTIONS

Table II gives descriptions of the network data sets used in this work. The code names in the first row correspond to the respective entries in the Netzschleuder repository [35] where the networks can be downloaded. Some of the descriptions were obtained from the Colorado Index of Complex Networks [36].

For all networks, the versions considered in this work were transformed into simple graphs, i.e., symmetrized versions of directed networks and/or with parallel edges and self-loops removed.

[1] P. W. Holland, K. B. Laskey, and S. Leinhardt, Stochastic blockmodels: First steps, Social Netw. **5**, 109 (1983).

[2] B. Karrer and M. E. J. Newman, Stochastic blockmodels and community structure in networks, Phys. Rev. E **83**, 016107 (2011).

[3] T. P. Peixoto, Bayesian stochastic blockmodeling, in *Advances in Network Clustering and Blockmodeling*, edited by P. Doreian, V. Batagelj, A. Ferligoj (Wiley, 2019).

[4] R. Guimerà and M. Sales-Pardo, Missing and spurious interactions and the reconstruction of complex networks, Proc. Natl. Acad. Sci. USA **106**, 22073 (2009).

[5] E. M. Airoldi, D. M. Blei, S. E. Fienberg, and E. P. Xing, Mixed membership stochastic blockmodels, J. Mach. Learn. Res. **9**, 1981 (2008).

[6] K. S. Xu and A. O. Hero, Dynamic stochastic blockmodels for time-evolving social networks, IEEE J. Select. Topics Signal Proc. **8**, 552 (2014).

[7] T. P. Peixoto, Inferring the mesoscale structure of layered, edge-valued, and time-varying networks, Phys. Rev. E **92**, 042807 (2015).

[8] A. Ghasemian, P. Zhang, A. Clauset, C. Moore, and L. Peel, Detectability Thresholds and Optimal Algorithms for Community Structure in Dynamic Networks, Phys. Rev. X **6**, 031005 (2016).

[9] N. Stanley, S. Shai, D. Taylor, and P. J. Mucha, Clustering network layers with the strata multilayer stochastic block model, IEEE Trans. Netw. Sci. Eng. **3**, 95 (2016).

[10] P. Erdős and A. Rényi, On random graphs, I, Publ. Math. (Debrecen) **6**, 290 (1959).

[11] B. Fosdick, D. Larremore, J. Nishimura, and J. Ugander, Configuring random graph models with fixed degree sequences, SIAM Rev. **60**, 315 (2018).

[12] S. C. Olhede and P. J. Wolfe, Network histograms and universality of blockmodel approximation, Proc. Natl. Acad. Sci. USA **111**, 14722 (2014).

[13] A. Ghasemian, H. Hosseinmardi, and A. Clauset, Evaluating overfit and underfit in models of network community structure, IEEE Trans. Knowl. Data Eng. **32**, 1722 (2019).

[14] A. Ghasemian, H. Hosseinmardi, A. Galstyan, E. M. Airoldi, and A. Clauset, Stacking models for nearly optimal link prediction in complex networks, Proc. Natl. Acad. Sci. USA **117**, 23393 (2020).

[15] L. Lovász, Random walks on graphs: A survey, Combinatorics, Paul Erdős Is Eighty **2**, 1 (1993).

[16] M. Newman, *Networks: An Introduction* (Oxford University Press, Oxford, 2010).

[17] T. P. Peixoto, Nonparametric Bayesian inference of the microcanonical stochastic block model, Phys. Rev. E **95**, 012317 (2017).

[18] T. P. Peixoto, Latent Poisson models for networks with heterogeneous density, Phys. Rev. E **102**, 012309 (2020).

[19] T. P. Peixoto, Merge-split Markov chain Monte Carlo for community detection, Phys. Rev. E **102**, 012305 (2020).

[20] T. P. Peixoto, Hierarchical Block Structures and High-Resolution Model Selection in Large Networks, Phys. Rev. X **4**, 011047 (2014).

[21] E. Kreyszig, H. Kreyszig, and E. J. Norminton, *Advanced Engineering Mathematics*, 10th ed. (John Wiley & Sons, Hoboken, NJ, 2010).

[22] A. Decelle, F. Krzakala, C. Moore, and L. Zdeborová, Asymptotic analysis of the stochastic block model for modular networks and its algorithmic applications, Phys. Rev. E **84**, 066106 (2011).

[23] T. P. Peixoto, Disentangling Homophily, Community Structure and Triadic Closure in Networks, Phys. Rev. X **12**, 011004 (2022).

[24] C. Gini, Measurement of inequality of incomes, Econ. J. **31**, 124 (1921).

[25] M. E. J. Newman, Modularity and community structure in networks, Proc. Natl. Acad. Sci. USA **103**, 8577 (2006).

[26] T. P. Peixoto, Descriptive vs. inferential community detection: Pitfalls, myths and half-truths, arXiv:2112.00183 [physics, stat] (2022).

[27] T. P. Peixoto, Efficient Monte Carlo and greedy heuristic for the inference of stochastic block models, Phys. Rev. E **89**, 012804 (2014).

[28] J. Parkkinen, J. Sinkkonen, A. Gyenge, and S. Kaski, A block model suitable for sparse graphs, in *Proceedings of the 7th International Workshop on Mining and Learning with Graphs (MLG 2009), Leuven* (2009).

[29] K. Rohe, J. Tao, X. Han, and N. Binkiewicz, A note on quickly sampling a sparse matrix with low rank expectation, J. Mach. Learn. Res. **19**, 3040 (2018).

[30] T. P. Peixoto, The `graph-tool` python library, 10.6084/m9.figshare.1164194, available at https://graph-tool.skewed.de (2014).

[31] M. E. J. Newman, Mixing patterns in networks, Phys. Rev. E **67**, 026126 (2003).

[32] V. Batagelj and M. Zaveršnik, Fast algorithms for determining (generalized) core groups in social networks, Adv. Data Anal. Class. **5**, 129 (2011).

[33] D. J. Watts and S. H. Strogatz, Collective dynamics of 'small-world' networks, Nature (London) **393**, 440 (1998).

[34] K.-i. Hashimoto, Zeta Functions of finite graphs and representations of p-Adic groups, in *Automorphic Forms and Geometry of Arithmetic Varieties*, Advanced Studies in Pure Mathematic, Vol. 15, edited by K. Hashimoto and Y. Namikawa (Academic Press, San Diego, 1989), pp. 211–280.

[35] T. P. Peixoto, The Netzschleuder network catalogue and repository, https://networks.skewed.de (2020).

[36] A. Clauset, E. Tucker, and M. Sainz, The Colorado Index of Complex Networks, https://icon.colorado.edu (2016).

[37] R. Brattig Correia, L. P. de Araújo Kohler, M. M. Mattos, and L. M. Rocha, City-wide electronic health records reveal gender and age biases in administration of known drug-drug interactions, npj Digit. Med. **2**, 74(2019).

[38] B. Szalkai, C. Kerepesi, B. Varga, and V. Grolmusz, The Budapest Reference Connectome Server v2.0, Neurosci. Lett. **595**, 60 (2015).

[39] S. J. Cook, T. A. Jarrell, C. A. Brittin, Y. Wang, A. E. Bloniarz, M. A. Yakovlev, K. C. Q. Nguyen, L. T.-H. Tang, E. A. Bayer, J. S. Duerr, H. E. Bülow, O. Hobert, D. H. Hall, and S. W. Emmons, Whole-animal connectomes of both *Caenorhabditis elegans* sexes, Nature (London) **571**, 63 (2019).

[40] N. Simonis, J.-F. Rual, A.-R. Carvunis, M. Tasan, I. Lemmens, T. Hirozane-Kishikawa, T. Hao, J. M. Sahalie, K. Venkatesan, F. Gebreab, S. Cevik, N. Klitgord, C. Fan, P. Braun, N. Li, N. Ayivi-Guedehoussou, E. Dann, N. Bertin, D. Szeto, A. Dricot, et al., Empirically controlled mapping of the *Caenorhabditis elegans* protein-protein interactome network, Nat Methods **6**, 47 (2009).

[41] S. R. Collins, P. Kemmeren, X.-C. Zhao, J. F. Greenblatt, F. Spencer, F. C. P. Holstege, J. S. Weissman, and N. J. Krogan, Toward a comprehensive atlas of the physical interactome of Saccharomyces cerevisiae, Mol. Cell. Proteomics **6**, 439 (2007).

[42] S. S. Shen-Orr, R. Milo, S. Mangan, and U. Alon, Network motifs in the transcriptional regulation network of Escherichia coli, Nat. Genet. **31**, 64 (2002).

[43] R. E. Ulanowicz and D. L. DeAngelis, Network analysis of trophic dynamics in south Florida ecosystems, U.S. Geological Survey Program on the South Florida Ecosystem, 114 (1999).

[44] N. D. Martinez, Artifacts or attributes? Effects of resolution on the Little Rock Lake food web, Ecol. Monogr. **61**, 367 (1991).

[45] R. M. Thompson and C. R. Townsend, Impacts on stream food webs of native and exotic forest: An intercontinental comparison, Ecology **84**, 145 (2003).

[46] M. De Domenico, A. Sole-Ribalta, S. Gomez, and A. Arenas, Navigability of interconnected networks under random failures, Proc. Natl. Acad. Sci. USA **111**, 8351 (2014).

[47] W. Gray Roncal, Z. H. Koterba, D. Mhembere, D. M. Kleissas, J. T. Vogelstein, R. Burns, A. R. Bowles, D. K. Donavos, S. Ryman, R. E. Jung, L. Wu, V. Calhoun, and R. J. Vogelstein, MIGRAINE: MRI graph reliability analysis and inference for connectomics, in *2013 IEEE Global Conference on Signal and Information Processing* (IEEE, New York, 2013).

[48] R. M. Ewing, P. Chu, F. Elisma, H. Li, P. Taylor, S. Climie, L. McBroom-Cerajewski, M. D. Robinson, L. O'Connor, M. Li, R. Taylor, M. Dharsee, Y. Ho, A. Heilbut, L. Moore, S. Zhang, O. Ornatsky, Y. V. Bukhman, M. Ethier, Y. Sheng *et al.*, Large-scale mapping of human protein-protein interactions by mass spectrometry, Mol. Syst. Biol. **3**, 89 (2007).

[49] S. Coulomb, M. Bauer, D. Bernard, and M.-C. Marsolier-Kergoat, Gene essentiality and the topology of protein interaction networks, Proc. R. Soc. B **272**, 1721 (2005).

[50] M. Huss and P. Holme, Currency and commodity metabolites: Their identification and relation to the modularity of metabolic networks, IET Syst. Biol. **1**, 280 (2007).

[51] M. P. Young, The organization of neural systems in the primate cerebral cortex, Proc. R. Soc. Lond. B **252**, 13 (1993).

[52] D. B. Larremore, A. Clauset, and C. O. Buckee, A network approach to analyzing highly recombinant malaria parasite genes, PLoS Comput. Biol. **9**, e1003268 (2013).

[53] J. A. Dunne, C. C. Labandeira, and R. J. Williams, Highly resolved early Eocene food webs show development of modern trophic structure after the end-cretaceous extinction, Proc. R. Soc. B **281**, 20133280 (2014).

[54] T. A. Dallas, A. A. Aguirre, S. Budischak, C. Carlson, V. Ezenwa, B. Han, S. Huang, and P. R. Stephens, Gauging support for macroecological patterns in helminth parasites, Global Ecol. Biogeogr. **27**, 1437 (2018).

[55] M. Kato, T. Kakutani, T. Inoue, and T. Itino, Insect-flower relationship in the primary beech forest of Ashu, Kyoto: An overview of the flowering phenology and the seasonal pattern of insect visits, Nihon Gomu Kyoukaishi **61**, 281 (1988).

[56] C. Robertson, *Flowers and Insects; Lists of Visitors of Four Hundred and Fifty-three Flowers, by Charles Robertson* (n.p., 1928).

[57] G. Joshi-Tope, Reactome: A knowledgebase of biological pathways, Nucleic Acids Res. **33**, D428 (2004).

[58] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon, Network motifs: Simple building blocks of complex networks, Science **298**, 824 (2002).

[59] The CAIDA AS relationships dataset (2009), https://www.caida.org/catalog/datasets/as-relationships/.

[60] M. Ripeanu and I. Foster, Mapping the gnutella network: Macroscopic properties of large-scale peer-to-peer systems, in *Peer-to-Peer Systems* (Springer, Berlin, 2002), pp. 85–93.

[61] B. Karrer, M. E. J. Newman, and L. Zdeborová, Percolation on Sparse Networks, Phys. Rev. Lett. **113**, 208702 (2014).

[62] S. Knight, H. X. Nguyen, N. Falkner, R. Bowden, and M. Roughan, The internet topology zoo, IEEE J. Sel. Areas Commun. **29**, 1765 (2011).

[63] J. Kunegis, Konect, in *Proceedings of the 22nd International Conference on World Wide Web—WWW '13 Companion* (ACM Press, 2013).

[64] L. Šubelj and M. Bajec, Software systems through complex networks science, in *Proceedings of the First International Workshop on Software Mining—SoftwareMining '12* (ACM Press, 2012).

[65] University of Oregon route views project (2001), http://www.routeviews.org/routeviews/.

[66] L. Šubelj and M. Bajec, Clustering assortativity, communities and functional modules in real-world networks, arXiv:1202.3188v1 [physics.soc-ph] (2012).

[67] B. Zhang, R. Liu, D. Massey, and L. Zhang, Collecting the internet AS-level topology, SIGCOMM Comput. Commun. Rev. **35**, 53 (2005).

[68] M. E. J. Newman, Finding community structure in networks using the eigenvectors of matrices, Phys. Rev. E **74**, 036104 (2006).

[69] D. Newman, Bag of words data set (2019), https://archive.ics.uci.edu/ml/datasets/bag+of+words.

[70] D. Dua and C. Graff, UCI machine learning repository, University of California at Irvine, School of Information and Computer Science (2019), https://archive.ics.uci.edu/ml/index.php.

[71] X. Niu, X. Sun, H. Wang, S. Rong, G. Qi, and Y. Yu, Zhishi.me—Weaving Chinese linking open data, in *The Semantic Web—ISWC 2011* (Springer, Berlin, 2011), pp. 205–220.

[72] J. Leskovec, K. J. Lang, A. Dasgupta, and M. W. Mahoney, Community structure in large networks: Natural cluster sizes and the absence of large well-defined clusters, Internet Math. **6**, 29 (2009).

[73] C. H. C. Römhild, Bible cross-references, https://www.chrisharrison.net/index.php/Visualizations/BibleViz.

[74] K. D. Bollacker, S. Lawrence, and C. L. Giles, CiteSeer, in *Proceedings of the Second International Conference on Autonomous Agents—AGENTS '98* (ACM Press, 1998).

[75] A. K. McCallum, K. Nigam, J. Rennie, and K. Seymore, Automating the construction of internet portals with machine learning, Inform. Retrieval **3**, 127 (2000).

[76] M. Ley, The DBLP computer science bibliography: Evolution, research issues, perspectives, in *String Processing and Information Retrieval* (Springer, Berlin, 2002), pp. 1–10.

[77] W. Foundation, Wikimedia downloads (2019), https://dumps.wikimedia.org/.

[78] GroupLens Research, MovieLens data sets (2019), https://grouplens.org/datasets/movielens/.

[79] L. A. Adamic and N. Glance, The political blogosphere and the 2004 U.S. election, in *Proceedings of the 3rd International Workshop on Link Discovery—LinkKDD '05* (ACM Press, 2005).

[80] B. Pasternak and I. Ivask, Four unpublished letters, Books Abroad **44**, 196 (1970).

[81] J. H. Fowler and S. Jeon, The authority of Supreme Court precedent, Soc. Netw. **30**, 16 (2008).

[82] J. H. Fowler, T. R. Johnson, J. F. Spriggs, S. Jeon, and P. J. Wahlbeck, Network analysis and the law: Measuring the legal importance of precedents at the U.S. Supreme Court, Polit. Anal. **15**, 324 (2007).

[83] P. Bailey, N. Craswell, and D. Hawking, Engineering a multipurpose test collection for web retrieval experiments, Inform. Process. Manag. **39**, 853 (2003).

[84] B. Hall, A. Jaffe, and M. Trajtenberg, The NBER Patent Citation Data File: Lessons, insights and methodological tools, Tech. Rep. (National Bureau of Economic Research, Cambridge, MA, 2001).

[85] S. Slattery and M. Craven, Combining statistical and relational methods for learning in hypertext domains, in *Inductive Logic Programming* (Springer, Berlin, 1998), pp. 38–52.

[86] A. Calderone, A wikipedia based map of science, Figshare (2020), doi: 10.6084/m9.figshare.11638932.v5.

[87] R. Milo, S. Itzkovitz, N. Kashtan, R. Levitt, S. Shen-Orr, I. Ayzenshtat, M. Sheffer, and U. Alon, Superfamilies of evolved and designed networks, Science **303**, 1538 (2004).

[88] G. R. Kiss, C. Armstrong, R. Milroy, and J. Piper, An associative thesaurus of English and its computer analysis, Comput Liter. Studi., 153 (1973).

[89] C. Fellbaum, WordNet, in *Theory and Applications of Ontology: Computer Applications* (Springer Netherlands, Amsterdam, 2010), pp. 231–243.

[90] B. Stabler, Transportation network test problems (2021), https://github.com/bstabler/TransportationNetworks.

[91] D. E. Knuth, *The Stanford GraphBase: A Platform for Combinatorial Computing* (AcM Press, New York, 1993).

[92] A. Cardillo, J. Gómez-Gardeñes, M. Zanin, M. Romance, D. Papo, F. del Pozo, and S. Boccaletti, Emergence of network features from multiplexity, Sci. Rep. **3**, 1344 (2013).

[93] L. Šubelj and M. Bajec, Robust network community detection using balanced propagation, Eur. Phys. J. B **81**, 353 (2011).

[94] United States Federal Aviation Administration, Air traffic control system command center (2010), https://nasstatus.faa.gov/.

[95] openflights.org (2019).

[96] G. Boeing, Street network models and measures for every U.S. city, county, urbanized area, census tract, and Zillow-defined neighborhood, Urban Sci. **3**, 28 (2019).

[97] P. Crucitti, V. Latora, and S. Porta, Centrality measures in spatial networks of urban streets, Phys. Rev. E **73**, 036125 (2006).

[98] V. Latora, V. Nicosia, and G. Russo, *Complex Networks* (Cambridge University Press, Cambridge, 2017).

[99] Bureau of Transportation Statistics, T-100 domestic market (2017), https://www.transtats.bts.gov/Homepage.asp.

[100] 9th DIMACS implementation challenge—Shortest paths (2002), http://www.diag.uniroma1.it//challenge9/data/tiger/.

[101] R. Mathews, Secondary education in Victoria: The liberal dilemma, Melbourne Stud. Ed. **18**, 234 (1976).

[102] M. Fire, L. Tenenboim-Chekina, R. Puzis, O. Lesser, L. Rokach, and Y. Elovici, Computationally efficient link prediction in a variety of social networks, ACM Trans. Intell. Syst. Technol. **5**, 1 (2013).

[103] J. Moody, Peer influence groups: Identifying dense clusters in large networks, Soc. Netw. **23**, 261 (2001).

[104] J. Leskovec, J. Kleinberg, and C. Faloutsos, Graph evolution, ACM Trans. Knowl. Discov. Data **1**, 2 (2007).

[105] M. E. J. Newman, The structure of scientific collaboration networks, Proc. Natl. Acad. Sci. USA **98**, 404 (2001).

[106] S. Kumar, F. Spezzano, V. S. Subrahmanian, and C. Faloutsos, Edge weight prediction in weighted signed networks, in *2016 IEEE 16th International Conference on Data Mining (ICDM)* (IEEE, New York, 2016).

[107] K. Faust, Centrality in affiliation networks, Soc. Netw. **19**, 157 (1997).

[108] Kaggle, Chess ratings - Elo versus the Rest of the World, https://www.kaggle.com/c/chess/data.

[109] P. Sapiezynski, A. Stopczynski, D. D. Lassen, and S. Lehmann, Interaction data from the Copenhagen networks study, Sci Data **6**, 315(2019).

[110] S. Decker, C. W. Kohfeld, R. Rosenfeld, and J. Sprague, *St. Louis Homicide Project: Local Responses to a National Problem* (A report made to the community, St. Louis, 1991), pp. 22–23.

[111] M. Magnani, B. Micenkova, and L. Rossi, Combinatorial analysis of multiple networks, arXiv:1303.4986v1 [cs.SI] (2013).

[112] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. Ives, DBpedia: A nucleus for a web of open data, in *The Semantic Web* (Springer, Berlin, 2007), pp. 722–735.

[113] D. Lusseau, K. Schneider, O. J. Boisseau, P. Haase, E. Slooten, and S. M. Dawson, The bottlenose dolphin community of Doubtful Sound features a large proportion of long-lasting associations, Behav. Ecol. Sociobiol. **54**, 396 (2003).

[114] J. Mcauley and J. Leskovec, Discovering social circles in ego networks, ACM Trans. Knowl. Discov. Data **8**, 1 (2014).

[115] R. Michalski, S. Palus, and P. Kazienko, Matching organizational structure and social network extracted from email communication, in *Business Information Systems* (Springer, Berlin, 2011), pp. 197–206.

[116] B. Klimt and Y. Yang, The Enron corpus: A new dataset for email classification research, in *Machine Learning: ECML 2004* (Springer, Berlin, 2004), pp. 217–226.

[117] L. E. C. Rocha, F. Liljeros, and P. Holme, Simulated epidemics in an empirical spatiotemporal network of 50,185 sexual contacts, PLoS Comput Biol **7**, e1001109 (2011).

[118] B. F. Maier and D. Brockmann, Cover time for random walks on arbitrary complex networks, Phys. Rev. E **96**, 042307 (2017).

[119] M. Fire and R. Puzis, Organization mining using online social networks, Netw. Spat. Econ. **16**, 545 (2016).

[120] B. Viswanath, A. Mislove, M. Cha, and K. P. Gummadi, On the evolution of user interaction in Facebook, in *Proceedings of the 2nd ACM Workshop on Online Social Networks—WOSN '09* (ACM Press, 2009).

[121] E. Rochko, Map of the fediverse (2018), https://gist.github.com/Gargron/48e67b1b14723cd178c951fe7f373a38.

[122] A. Mislove, M. Marcon, K. P. Gummadi, P. Druschel, and B. Bhattacharjee, Measurement and analysis of online social networks, in *Proceedings of the 7th ACM SIGCOMM Conference on Internet Measurement—IMC '07* (ACM Press, 2007).

[123] M. Girvan and M. E. J. Newman, Community structure in social and biological networks, Proc. Natl. Acad. Sci. USA **99**, 7821 (2002).

[124] T. S. Evans, Clique graphs and overlapping communities, J. Stat. Mech. (2010) P12037.

[125] D. Yang, D. Zhang, Z. Yu, and Z. Yu, Fine-grained preference-aware location search leveraging crowdsourced digital footprints from LBSNs, in *Proceedings of the 2013 ACM International Joint Conference on Pervasive and Ubiquitous Computing* (ACM, 2013).

[126] D. Harper and J. S. Coleman, Introduction to mathematical sociology, Brit. J. Sociol. **16**, 260 (1965).

[127] M. Morris and R. Rothenberg, HIV transmission network metastudy project: An archive of data from eight network studies, 1988–2001, Inter-university Consortium for Political and Social Research (2011), doi: 10.3886/ICPSR22140.v1.

[128] R. Zafarani and H. Liu, Social computing data repository at Arizona State University (2009), http://socialcomputing.asu.edu/datasets/Hyves.

[129] P. M. Gleiser and L. Danon, Community structure in jazz, Adv. Complex Syst. **06**, 565 (2003).

[130] W. W. Zachary, An information flow model for conflict and fission in small groups, J. Anthropol. Res. **33**, 452 (1977).

[131] L. M. Gerdes, K. Ringler, and B. Autin, Assessing the Abu Sayyaf Group's strategic and learning capacities, Stud. Confl. Terror **37**, 267 (2014).

[132] Ò. Celma, The long tail in recommender systems, in *Music Recommendation and Discovery* (Springer, Berlin, 2010), pp. 87–107.

[133] J. Kunegis, G. Gröner, and T. Gottron, Online dating recommender systems, in *Proceedings of the 4th ACM RecSys Workshop on Recommender Systems and the Social Web—RSWeb '12* (ACM Press, 2012).

[134] S. Aref, D. Friggens, and S. Hendy, Analysing scientific collaborations of New Zealand institutions using scopusbibliometric data, in *Proceedings of the Australasian Computer Science Week Multiconference* (ACM, 2018).

[135] J. Coleman, E. Katz, and H. Menzel, The diffusion of an innovation among physicians, Sociometry **20**, 253 (1957).

[136] M. De Domenico, A. Lancichinetti, A. Arenas, and M. Rosvall, Identifying Modular Flows on Multilayer Networks Reveals Highly Overlapping Organization in Interconnected Systems, Phys. Rev. X **5**, 011027 (2015.

[137] N. Eagle and A. (Sandy) Pentland, Reality mining: Sensing complex social systems, Pers. Ubiquit. Comput. **10**, 255 (2006).

[138] L. C. Freeman, C. M. Webster, and D. M. Kirke, Exploring social structure using dynamic three-dimensional color images, Soc. Netw. **20**, 109 (1998).

[139] R. Mastrandrea, J. Fournet, and A. Barrat, Contact patterns in a high school: A comparison between data collected using wearable sensors, contact diaries and friendship surveys, PLoS ONE **10**, e0136497 (2015).

[140] L. Isella, J. Stehlé, A. Barrat, C. Cattuto, J.-F. Pinton, and W. Van den Broeck, What's in a crowd? Analysis of face-to-face behavioral networks, J. Theor. Biol. **271**, 166 (2011).

[141] M. Génois, C. L. Vestergaard, J. Fournet, A. Panisson, I. Bonmarin, and A. Barrat, Data on face-to-face contacts in an office building suggest a low-cost vaccination strategy based on community linkers, Netw. Sci. **3**, 326 (2015).

[142] J. Stehlé, N. Voirin, A. Barrat, C. Cattuto, L. Isella, J.-F. Pinton, M. Quaggiotto, W. Van den Broeck, C. Régis, B. Lina, and P. Vanhems, High-resolution measurements of face-to-face contact patterns in a primary school, PLoS ONE **6**, e23176 (2011).

[143] M. Fire, G. Katz, Y. Elovici, B. Shapira, and L. Rokach, Predicting student exam's scores by analyzing social network data, in *Active Media Technology* (Springer, Berlin, 2012), pp. 584–595.

[144] A.-M. Niekamp, L. A. G. Mercken, C. J. P. A. Hoebe, and N. H. T. M. Dukers-Muijrers, A sexual affiliation network of swingers, heterosexuals practicing risk behaviours that potentiate the spread of sexually transmitted infections: A two-mode approach, Soc. Netw. **35**, 223 (2013).

[145] M. De Choudhury, Discovery of information disseminators and receptors on online social media, in *Proceedings of the 21st ACM Conference on Hypertext and Hypermedia—HT '10* (ACM Press, 2010).

[146] S. González-Bailón, J. Borge-Holthoefer, A. Rivero, and Y. Moreno, The dynamics of protest recruitment through an online network, Sci. Rep. **1**, 197 (2011).

[147] M. De Domenico, A. Lima, P. Mougel, and M. Musolesi, The anatomy of a scientific rumor, Sci. Rep. **3**, 2980(2013).

[148] G. F. Chami, S. E. Ahnert, N. B. Kabatereine, and E. M. Tukahebwa, Social network fragmentation and community health, Proc. Natl. Acad. Sci. USA **114**, E7425 (2017).

[149] S. Kosack, M. Coscia, E. Smith, K. Albrecht, A.-L. Barabási, and R. Hausmann, Functional structures of US state governments, Proc. Natl. Acad. Sci. USA **115**, 11748 (2018).

[150] Z. P. Neal, A sign of the times? weak and strong polarization in the U.S. Congress, 1973–2016, Soc. Netw. **60**, 103 (2020).

[151] Z. Neal, The backbone of bipartite projections: Inferring relationships from co-authorship, co-sponsorship, co-attendance and other co-behaviors, Soc. Netw. **39**, 84 (2014).

[152] J. Leskovec, D. Huttenlocher, and J. Kleinberg, Signed networks in social media, in *Proceedings of the 28th International Conference on Human Factors in Computing Systems—CHI '10* (ACM Press, 2010).

[153] M. Fire and Y. Elovici, Data mining of online genealogy datasets for revealing lifespan patterns in human population, ACM Trans. Intell. Syst. Technol. **6**, 1 (2015).

[154] L. Freeman, S. Freeman, and A. Michaelson, On human social intelligence, J. Soc. Biol. Syst. **11**, 415 (1988).

[155] J. Leskovec, L. A. Adamic, and B. A. Huberman, The dynamics of viral marketing, ACM Trans. Web **1**, 5 (2007).

[156] E.-P. Lim, V.-A. Nguyen, N. Jindal, B. Liu, and H. W. Lauw, Detecting product review spammers using rating behaviors, in *Proceedings of the 19th ACM International Conference on Information and Knowledge Management—CIKM '10* (ACM Press, 2010).

[157] C.-N. Ziegler, S. M. McNee, J. A. Konstan, and G. Lausen, Improving recommendation lists through topic diversification, in *Proceedings of the 14th International Conference on World Wide Web—WWW '05* (ACM Press, 2005).

[158] A. Evtushenko and M. T. Gastner, Beyond fortune 500: Women in a global network of directors, in *Complex Networks and Their Applications VIII*, vol. 1, edited by H. Cherifi *et al.*, (Springer, Cham, 2020), pp. 586–598.

[159] P. Massa and P. Avesani, Controversial users demand local trust metrics: An experimental study on epinions.com community, in *AAAI*,vol. 5 (2005), pp. 121–126.

[160] J. Wachs, M. Fazekas, and J. Kertész, Corruption risk in contracting markets: A network science perspective, Int. J. Data Sci. Anal. **12**, 45 (2021).

[161] M. De Domenico, V. Nicosia, A. Arenas, and V. Latora, Structural reducibility of multilayer networks, Nat. Commun. **6**, 6864 (2015).

[162] S. Chacon and B. Straub, Github, in *Pro Git* (Apress, 2014), pp. 131–180.

[163] K. Goldberg, T. Roeder, D. Gupta, and C. Perkins, Eigentaste A Constant Time Collaborative Filtering Algorithm, Inform. Retrieval **4**, 133 (2001).

[164] T. Hogg and K. Lerman, Social dynamics of digg, EPJ Data Sci. **1**, 5(2012).