Disentangling homophily, community structure and triadic closure in networks

Tiago P. Peixoto
Department of Network and Data Science, Central European University, 1100 Vienna, Austria and Department of Mathematical Sciences, University of Bath, Claverton Down, Bath BA2 7AY, United Kingdom

Network homophily, the tendency of similar nodes to be connected, and transitivity, the tendency of two nodes being connected if they share a common neighbor, are conflated properties in network analysis, since one mechanism can drive the other. Here we present a generative model and corresponding inference procedure that are capable of distinguishing between both mechanisms. Our approach is based on a variation of the stochastic block model (SBM) with the addition of triadic closure edges, and its inference can identify the most plausible mechanism responsible for the existence of every edge in the network, in addition to the underlying community structure itself. We show how the method can evade the detection of spurious communities caused solely by the formation of triangles in the network, and how it can improve the performance of edge prediction when compared to the pure version of the SBM without triadic closure.

I. INTRODUCTION

One of the most typical properties of social networks is the presence of homophily \cite{1,2,3,4}, i.e. the increased tendency of an edge to exist between two nodes if they share the same underlying characteristic, such as race, gender, class and a variety of other social parameters. More broadly, when the underlying similarity parameter is not specified a priori, the same homophily pattern is known as community structure \cite{5}. Another pervasive pattern encountered in the same kinds of network is transitivity \cite{6,7,8}, i.e. the increased probability of observing an edge between two nodes if they have a neighbor in common. Although these patterns are indicative of two distinct mechanisms of network formation, namely choice or constraint homophily \cite{9} and triadic closure \cite{10}, respectively, they are difficult to distinguish in non-longitudinal data. This is because both processes can result in the same kinds of observation: 1. the preferred connection between nodes of the same kind can induce the presence of triangles involving similar nodes, and 2. the tendency of triangles to be formed can induce the formation of groups of nodes with a higher density of connections between them, when compared to the rest of the network \cite{11,12}. This conflation means we cannot reliably interpret the underlying mechanisms of network formation merely from the abundance of triangles or observed community structure in network data.

In this work we present a solution to this problem, consisting in a principled method to disentangle homophily and community structure from triadic closure in network data, conditioned on mild modeling assumptions. This is achieved by formulating a generative model that includes community structure in a first instance, and an iterated process of triadic closure in a second. Based on this model, we develop a nonparametric Bayesian inference algorithm that is capable of identifying which edges are more likely to be due to community structure or triadic closure, in addition to the underlying community structure itself. What our approach demonstrates is that, while at first it seems that triadic closure and homophily generate similar patterns in network structure, the different mechanisms also leave behind particular traces in the network structure that can be used to disambiguate between the two.

Several authors have demonstrated that triadic closure can induce community structure and homophily in networks. Foster et al \cite{11,12} have shown that maximum entropy network ensembles conditioned on prescribed abundances of triangles tend to possess high modularity. A more recent analysis of this kind of ensemble by López et al \cite{13} showed that it is marked by a spontaneous size-dependent formation of “triangle clusters.” Bianconi et al \cite{14} have investigated a network growth model, where nodes are progressively added to the network, and connected in such a way as to increase the amount of triangles, and shown that it is capable of producing networks with emergent community structure. The effect of triangle formation on apparent community structure has been further studied by Wharrie et al \cite{15}, who showed that these patterns can even mislead methods specifically designed to avoid the detection of spurious communities in random networks. More recently, Asikainen et al \cite{16} have shown that iterated triadic closure can exacerbate homophily present in the original network, via a simple macroscopic model.

The approach presented in this work differs from the aforementioned ones primarily in that it runs in the reverse direction: instead of only defining a conceptual network model that demonstrates the interlink between triadic closure and homophily given prescribed parameters, the proposed method operates on empirical network data, and reconstructs the underlying generative process, decomposing it into distinct community structure and triadic closure components. As we show, this reconstruction yields a detailed interpretation of the underlying mechanisms of network formation, allowing us to identify macro-scale structures that emerge spontaneously from
micro-scale higher-order interactions \cite{17,18}, and in this way we can separate them from inherently macro-scale structures.

It is also worth mentioning some recent methods that have been proposed that use triangles as a means of finding communities in networks \cite{19,21}. Although these methods can be informative of the interplay between triangles and large-scale structure, they cannot explain the formation of the triangles themselves, or identify the contribution of pairwise homophily, as we do here. Likewise, there are also methods that reconstruct networks via compositions of higher-order building blocks \cite{22,23}, but which can make no statement about any existing large-scale homophily. Finally, a commonly used approach in the social sciences literature is to model the occurrence of triangles and homophily using exponential random graph models (ERGMs) \cite{24}. Generally, these models do not possess likelihoods that can be expressed in closed form, making their inference quite difficult without relying on approximations. Furthermore, when they are used to model the presence of triangles or other small subgraphs, they tend to possess extreme degeneracies \cite{11,25,26}, rendering them rather implausible models for clustered networks. Additionally, when they are combined with homophily, this only usually done with observed homophilic traits, not latent ones as we consider here.

Our method is based on the nonparametric Bayesian inference of a modified version of the stochastic block model (SBM) \cite{29,30} with the addition of triadic closure edges, and therefore leverages the statistical evidence available in the data, without overfitting. Importantly, our method is capable of determining when the observed structure can be attributed to an actual preference of connection between nodes, as described by the SBM, rather than an iterated triadic closure process occurring on top of a substrate network. As a result, we can distinguish between “true” and apparent community structure caused by increased transitivity. A key concept in the method that allows this distinction to be made is the principle of maximum parsimony: in situations where both transitivity and homophily serve as competing hypotheses, their relative plausibility is evaluated based not only on how well they can explain the data, but also on the amount of information needed to specify the particular model in the first place. As we also demonstrate, this decomposition yields an edge prediction method that tends to perform better in many instances than the SBM used in isolation.

We emphasize that our approach is capable of performing the decomposition between homophily and triadic closure from a single network observation without annotations. At first, this might seem at odds with formal results relating to similar, but distinct decomposition problems, that state that this kind of disentanglement is not possible from a single network observation. In particular, Chang et al \cite{31} considered a scenario of uncertain network measurement, and proved that, absent of any modeling assumption on how the edges of the network are initially placed, it is not possible to estimate the network structure from a single network observation. Similarly, Shalizi and Thomas \cite{32} famously proved that contagion (causal inheritance of traits due to peer influence) cannot be distinguished from homophily given a single network observation. Both of these statements rely on a lack of stipulation on how the networks are generated (which formally cannot be distinguished from making an explicit assumption that all networks are equally likely \textit{a priori}). However, whenever such stipulations are made, the situation changes. In particular, McFowland III and Shalizi \cite{33} have shown that as soon as the homophilic traits are latent (instead of being observed directly as considered in Ref. \cite{34}), and can be modeled as a SBM, the disentanglement becomes possible, even for a single network. Likewise, if we use the SBM as a structured prior distribution \cite{34}, it becomes possible to estimate the magnitude of the measurement error as well as to reconstruct noisy networks, even for a single network observation and when the error magnitude is unknown \textit{a priori}. Although the disentanglement problem that we consider here is different from the aforementioned ones, and the impossibility results do not carry over, nevertheless make use of the same kinds of modeling assumptions that make the other problems feasible.

Our manuscript is organized as follows. In Sec. \[II\] we describe our model, and its inference procedure. In Sec. \[III\] we demonstrate how it can be used disambiguate triadic closure from community structure in artificially generated networks. In Sec. \[IV\] we perform an analysis of empirical networks, in view of our method. In Sec. \[V\] we show how our model can improve edge prediction. We end in Sec. \[VI\] with a conclusion.

II. STOCHASTIC BLOCK MODEL WITH TRIADIC CLOSURE (SBM/TC)

Community structure and triadic closure are generally interpreted as different processes of network formation. With the objective of allowing their identification \textit{a posteriori} from network data, our approach consists in defining a generative network model that encodes both processes explicitly. More specifically, our generative model consists of two steps, with the first one being the generation of a substrate network containing “seminal” edges, placed according to an arbitrary mixing pattern between nodes, and an additional layer containing triadic closure edges, potentially connecting two nodes if they share a common neighbor in the substrate network (see Fig. \[1\]). The final network is obtained by “erasing” the identity of the edges, i.e. whether they are seminal or due to closure of a triangle. Conversely, the inference procedure consists in moving in the opposite direction, i.e. given a simple graph, with no annotations on the edges, we consider the posterior distribution of all possible divisions into seminal and triadic closure edges, weighted according to their plausibility.
The triadic closure edges are represented by an additional set of $N$ “ego” graphs $g$, attributed to each node $u$ of $A$, where $g(u)$ is the ego graph of node $u$. The ego graph $g(u)$ is allowed only to contain nodes that are neighbors of $u$ in $A$ (excluding $u$ itself), and edges that do not exist in $A$, so that any existing edge in $g(u)$ amounts to a triadic closure in $A$. The adjacency of $g(u)$ is given by

$$g_{ij}(u) = \begin{cases} 1, & \text{if } (i, j) \in g(u), \\ 0, & \text{otherwise.} \end{cases}$$

Let us denote the existence of an open triad $(i, u, j)$ in $A$ with

$$m_{ij}(u) = A_{ui}A_{uj}(1 - A_{ij}),$$

such that $m_{ij}(u) = 1$ if the open triad exists, or 0 otherwise, and we adopt the convention $A_{uu} = 0$ throughout. Therefore, an edge $(i, j)$ can exist in $g(u)$ only if $m_{ij}(u) = 1$. Based on this, the ego networks are generated independently with probability,

$$P(g(u)|A, p_u) = \prod_{i<j} \left[ p_u m_{ij}(u)^{g_{ij}(u)} [1 - p_u m_{ij}(u)]^{1-g_{ij}(u)} \right],$$

where $p_u \in [0, 1]$ is a probability associated with node $u$ that controls the degree to which its neighbors in $A$ end up connected in $g(u)$. This process may result in the same edge $(i, j)$ existing in different graphs $g(u)$, if $i$ and $j$ share more than one common neighbor in $A$. We therefore consider the resulting simple graph $G(A, g)$, constructed by ignoring any multiplicities introduced by the various ego graphs, i.e., with adjacency given by

$$G_{ij}(A, g) = \begin{cases} 1, & \text{if } A_{ij} + \sum_u g_{ij}(u) > 0, \\ 0, & \text{otherwise.} \end{cases}$$

The joint probability of the above process is then given by

$$P(G, g, A|p, b) = \prod_{u} P(g(u)|A, p_u),$$

where $1_{\{x\}}$ is the indicator function. Unfortunately, the marginal probability of the final graph

$$P(G) = \sum_{g, A, b} P(G, g, A|p, b) P(p) P(b) dp,$$

with $P(p)$ and $P(b)$ being prior probabilities, does not lend itself to a tractable computation. Luckily, however,
this will not be needed for our inference procedure. Instead, we are interested in the posterior distribution
\[ P(g, A, b | G) = \frac{P(G, g, A | b) P(b)}{P(G)}, \]  
(8)
which describes the probability of a decomposition of an observed simple graph \( G \) into its seminal graph \( A \), the underlying community structure \( b \), and the triadic closures represented by the ego graphs \( g \). (Although the marginal distribution \( P(G) \) appears in the denominator of the above equation, we will see later on that it is just a normalization constant that does not in fact need to be computed.) The marginal likelihood
\[ P(G, g, A | b) = P(G | A, g) P(g | A) P(A | b) \]  
(9)
can be computed easily via
\[
P(g | A) = \prod_u \int_0^1 P(g(u) | A, p) P(p) \, dp
\]
\[ = \prod_u \left[ \left( \frac{\sum_{i<j} m_{ij}(u)}{\sum_{i<j} g_{ij}(u)} \right)^{-1} \frac{1}{1 + \sum_{i<j} m_{ij}(u)} \right], \]
(10)
where we have used a uniform prior \( P(p) = 1 \), omitted for simplicity an indicator function setting \( P(g | A) = 0 \) if \( g_{ij} > 0 \) and \( m_{ij} = 0 \) for any \((i,j)\), and with the remaining likelihood term being only the indicator function, \( P(G | A, g) = 1_{\{G = G(A, g)\}} \). Although this choice of priors makes the calculation very simple, it implies that we expect the observed graphs to always have a large fraction of triadic closures. In Appendix\(^2\) we describe a slight modification of this model that makes it more versatile with respect to the abundance of triadic closures, at the expense of yielding somewhat longer expressions for the likelihood. We note that we made use of the modifications specified there in our ensuing analysis, as they can only improve the use of the model.

A. Iterated triadic closures

Triadic closures increase the number of edges in the network, and in this way can introduce opportunities for new triadic closures, involving both older and newer edges. This leads naturally to a dynamical model, where generations of triadic closures are progressively introduced in generation \( l \in [1, \ldots, L] \). For our formulation, it will be useful to define the cumulative network at generation \( l \), defined recursively by
\[ A^{(l)}_{ij} = \begin{cases} 1, & \text{if } A^{(l-1)}_{ij} + \sum_u g^{(l)}_{ij}(u) > 0, \\ 0, & \text{otherwise}, \end{cases} \]
(11)
with boundary conditions \( A^{(0)} = A \) (henceforth \( A \) refers solely to the seminal network, whereas e.g. \( A^{(0)} \) is the resulting network considering the first iteration of triadic closures, and \( \{A^{(l)}\} \) refers to the set of all generations, including the seminal network), and \( g^{(0)}(u) \) being empty graphs for all \( u \), and we will denote the final generation as \( A^{(L)} = G \). The formation of new triadic closure layers is done according to the probability
\[
P(g^{(l)}(u) | A^{(l-1)}, g^{(l-1)}, P_u^{(l)}) = \prod_{i<j} \left[ p^{(l)}_{ij} m^{(l)}_{ij}(u) \left( 1 - p^{(l)}_{ij} m^{(l)}_{ij}(u) \right)^{-1} g^{(l)}_{ij}(u) \right]. \]
(12)
where an open triad \((i, u, j)\) at generation \( l \) is denoted by
\[ m^{(l)}_{ij}(u) = w^{(l)}_{ij}(u) \left( 1 - A^{(l-1)}_{ij} \right), \]
(13)
so that \( m^{(l)}_{ij}(u) \in \{0,1\} \), where
\[ w^{(l)}_{ij}(u) = \begin{cases} 1, & \text{if } A^{(l-1)}_{ui} \sum_v g^{(l-1)}_{uj}(v) + A^{(l-1)}_{uj} \sum_v g^{(l-1)}_{ui}(v) > 0, \\ 0, & \text{otherwise}, \end{cases} \]
(14)
determines whether or not the open triad \((i, u, j)\) at generation \( l \) has at least one of the edges \((u,i)\) or \((u,j)\)

\(^2\) One may wonder if such a dynamical process would also make sense for the homophily part of the model, represented by the SBM. However, since homophily implies a conditional independence of the placement of the edges, it does not matter the order with which edges are added to the network, only their final placement.
formed exactly at the preceding generation \( l - 1 \). This restriction means that triadic closures at generation \( l \) can only close new triads that have been introduced at generation \( l - 1 \), not previously. The reason for this is a matter of identifiability: an edge at generation \( l \) that closes an open triad that has been formed at generation \( l' < l \) could also have been generated in any of the intermediate generations \([l', l - 1]\), thus introducing an inevitable ambiguity in the inference. The above restriction removes the ambiguity, and forces the new generations to form triadic closures which could not have existed in the preceding generations (see Fig. 2). Note that this restriction does not significantly alter the generality of the model, since the same final networks can still be formed with the similar probability despite it.

With the above, the joint likelihood of all generations is given by

\[
P(\{g^{(l)}\}, \{A^{(l)}\}|b, p) = P(A|b) \prod_{l=1}^{L} P(g^{(l)}|A^{(l-1)}, g^{(l-1)}, P_u^{(l)})
\]  \hspace{1cm} (15)

Following the same calculation as before, we obtain the individual marginal likelihood at each generation \( l \) as

\[
P(\{g^{(l)}\}, \{A^{(l)}\}|b) = P(A|b) \prod_{l=1}^{L} P(g^{(l)}|A^{(l-1)}, g^{(l-1)}).
\]  \hspace{1cm} (16)

with the individual terms in the product being entirely analogous to Eq. (16).

\[
P(g^{(l)}|A^{(l-1)}, g^{(l-1)}) = \prod_u \left[ \frac{\sum_{i<j} m_{ij}^{(l)}(u)}{\sum_{i<j} s_{ij}^{(l)}(u)} \right]^{-1} \frac{1}{1 + \sum_{i<j} m_{ij}^{(l)}(u)}
\]  \hspace{1cm} (17)

Finally, the posterior distribution for the reconstruction becomes

\[
P(\{g^{(l)}\}, \{A^{(l)}\}, b|G) = \frac{P(G, \{g^{(l)}\}, \{A^{(l)}\}|b) P(b)}{P(G)}.
\]  \hspace{1cm} (18)

Note that for \( L = 1 \) we recover the previous model. Having to specify \( L \) beforehand is not a strict necessity, since the inference will only occupy new generations if this yields a more parsimonious description of the network. [3]

---

[3] In more detail, we can recover the unconstrained model by substituting Eq. (13) with \( m_{ij}^{(l)}(u) = A_{ij}^{(l-1)} A_{ij}^{(l-1)} (1 - A_{ij}^{(l-1)}) \). Since the edges at each layer \( l \) are generated independently, this would only mean that more edges would be generated on top of each other across the layers. Since these multiple edges are removed in the end, this means that the unconstrained model would have a higher probability for forming edges that are possible in the earlier generations, since they could appear also in the later ones. But since we typically require only a very small number of generations, this is a very minor effect, and both models become very similar, while the constrained model is easier to infer.

[4] If we wanted to treat \( L \) as an unknown, we should introduce a prior for \( L, P(L) \), and include that in the posterior as well. However, with the parametrization in Appendix C generations which are unpopulated with edges have no contribution to the marginal likelihood. Therefore we can simply set \( L \) to be a sufficiently large value, for example \( L = C(N) \), since for later generations it is impossible to add new edges.

---

B. Inference algorithm

The posterior distribution of Eq. (18) can be written exactly, up to a normalization constant. However, this fact alone does not allow us to directly sample from this distribution, which can only be done in very special cases. Instead, we rely here on Markov chain Monte Carlo (MCMC), implemented as follows. We begin with an arbitrary choice of \( \{g^{(l)}\}, \{A^{(l)}\} \) and \( b \) that is compatible with our observed graph \( G \). We then consider modifications of these quantities, and accept or reject them according to the Metropolis-Hastings criterion [38, 39]. More specifically, we consider moves of the kind \( P(\{g^{(l)}\}, \{A^{(l)}\}|\{g^{(l')}\}, \{A^{(l')}\}) \), and accept them according to the probability

\[
\min \left( 1, \frac{P(\{g^{(l')}\}, \{A^{(l')}\}, b|G) P(\{g^{(l)}\}, \{A^{(l)}\}, b|G)}{P(\{g^{(l)}\}, \{A^{(l)}\}|\{g^{(l')}\}, \{A^{(l')}\})} \right)
\]  \hspace{1cm} (19)

which, as we mentioned before, does not require the computation of the intractable marginal probability \( P(G) \).

We also consider moves that change the community structure, according to a proposal \( P(b'|b) \) and accept with probability

\[
\min \left( 1, \frac{P(\{b\}|\{b'\}) P(b') P(b)}{P(\{b\})} \right)
\]  \hspace{1cm} (20)

For the latter we use the merge-split moves described in Ref. [40]. Iterating the moves described above eventually produces samples from the target posterior distribution. In Appendix C we specify the details of the particular move proposals we use.

Given samples from the posterior distribution, we can use them to summarize it in a variety of ways. A useful quantity is the marginal probability \( \pi_{ij} \) of an edge \( (i, j) \) being seminal, which is given by

\[
\pi_{ij} = \sum_{\{g^{(l)}\}, \{A^{(l)}\}, b} \frac{A_{ij} P(\{g^{(l)}\}, \{A^{(l)}\}, b|G)}{P(G)}
\]  \hspace{1cm} (21)

Conversely, the reciprocal quantity,

\[
1 - \pi_{ij}
\]  \hspace{1cm} (22)

corresponds to the probability that edge \( (i, j) \) is due to triadic closure, occurring in any generation or ego graph.
Figure 3. (a) Example artificial network generated as a fully random graph with a geometric degree distribution, \( N = 100 \) nodes and \( E = 94 \) edges, and (b) a process of triadic closure based on network (a) with parameter \( p_u = 0.8 \) for every node, with closure edges shown in red. It is also shown the partition found by fitting the SBM to the resulting network, and the description length obtained. (c) The result of inferring the SBM/TC model, which uncovers a single partition — no community structure — and the closure edges shown in red (the thickness of the edges correspond to the marginal probabilities \( \pi_{ij} \) and \( 1 - \pi_{ij} \) for the seminal and closure edges, respectively). It is also shown the description length of the SBM/TC fit.

Therefore, the quantity \( \pi \) gives us a concise summary of posterior decomposition of a network, and we will use it throughout our analysis. (It is easy to devise and compute other summaries, such as the marginal probability of an edge belonging to a given triadic generation, or a particular ego graph, but we will not have use for those in our analysis.)

III. DISTINGUISHING COMMUNITY STRUCTURE FROM TRIADIC CLOSURE

Here we illustrate how triadic closure can be mistaken as community structure, and how our inference method is capable of uncovering it. We begin by considering an artificial example, where we first sample a fully random network with a geometric degree distribution, \( N = 100 \) nodes and \( E = 94 \) edges, as shown in Fig. 3a. This network does not possess any community structure, since the probability of observing an edge is just proportional to the product of the degrees of the endpoint nodes — indeed if we fit a DC-SBM to it, we uncover, correctly, only a single group. Conditioned on this network, Fig. 3b shows sampled triadic closure edges, according to the model described previously, where each node has the same probability \( p_u = 0.8 \) of having neighbors connected in their ego graphs. In the same figure we show the result of fitting the DC-SBM on the network obtained by ignoring the edge types. That approach finds five assortative communities, corresponding to regions of higher densities of edges induced by the random introduction of transitive edges. One should not, however, interpret the presence of these regions as a special affinity between the respective groups of nodes, since they are a result of a random process that has no relation to that particular division of the network — indeed, if we run the whole process again from the beginning, the nodes will most likely end up clustered in completely different “communities.” If we now perform the inference of our SBM with triadic closure (SBM/TC), we obtain the result shown in Fig. 3c. Not only are we capable of distinguishing the seminal from the triadic closure edges (AUC ROC = 0.92), but we also correctly identify the presence of a single group of nodes, which is in full accordance with the completely random nature in which the network has been generated. In other words, with the SBM/TC we are not misled by the density heterogeneity introduced by triadic closures into thinking that the network possesses real community structure, and we realize instead that they can be better explained by a different process.

In the artificial example considered above, the result obtained with the SBM/TC model is more appealing, since it more closely matches the known generative process that was used. However, in more realistic situations, we will need to decide if it provides a better description of the data without such privileged information. In view of this, we can make our model selection argument more formal in the following way. Suppose we are considering a partition \( b^{(1)} \) found with inferring the SBM on a given network, as well as another partition \( b^{(2)} \) and ego graphs
\{g^{(i)}\}$ found with the SBM/TC model. We can decide which one provides a better description of a network via the posterior odds ratio,
\[
\Lambda = \frac{P(b^{(2)}, \{g^{(i)}\}, H_{SBM/TC} | G)}{P(b^{(1)}, H_{SBM} | G)} = \frac{P(G, \{g^{(i)}\}, \{A^{(i)}\}, b^{(2)})}{P(G, b^{(1)})} \times \frac{P(H_{SBM/TC})}{P(H_{SBM})},
\]
where $P(H_{SBM/TC})$ and $P(H_{SBM})$ are the prior probabilities for either model. In case these are the same, we have
\[
\Lambda = e^{-\Sigma_{SBM/TC} - \Sigma_{SBM}},
\]
where $\Sigma_{SBM/TC}$ and $\Sigma_{SBM}$ are the description lengths of both hypotheses, given by
\[
\Sigma_{SBM/TC} = - \ln P(G, \{g^{(i)}\}, \{A^{(i)}\}, b^{(2)}),
\]
\[
\Sigma_{SBM} = - \ln P(G, b^{(1)}).
\]
The description length \[41\] measures the amount of information necessary to encode both the data and the model parameters, and hence accounts both for the quality of fit and the model complexity. The above means that the model that is most likely \textit{a posteriori} is the one that most compresses the data under its parametrization, and thus the criterion amounts to an implementation of Occam’s razor, since it points to the best balance between model complexity and fitness.

Before we employ the above criterion to select between both models considered, it is important to emphasize that the pure SBM is “nested” inside the SBM/TC, since the former amounts to the special case of the latter when the pure SBM is “nested” inside the SBM/TC, since the both models considered, it is important to emphasize that Occam’s razor, since it points to the best balance between model complexity and fitness.

Inference using SBM
\begin{align*}
\text{(a)} \quad & \text{Inference using SBM} \\
\text{(b)} \quad & \text{Inference using SBM/TC}
\end{align*}

Figure 4. Recovery of community structure for artificial networks generated from the PP model with added triadic closure, as described in the text, for networks with $N = 10^5$ nodes, average degree $\langle k \rangle = 5$, $B = 10$ planted groups, and uniform triadic closure probability $p_u = p$ shown in the legend. Figures (a) and (b) correspond to inferences done using the SBM, and (c) and (d) with the SBM/TC model. All results were averaged over 10 network realizations. The vertical dashed line marks the detectability transition value $c^*_u$, described in the text.

parsimonious and hence more probable account of the way the data was generated — which happens also to be the true one in this controlled setting.

We proceed with a more systematic analysis of how triadic closure can interfere in community detection with artificial networks generated by the SBM, more specifically the special case known as the planted partition model (PP), where the $B$ groups have equal size, and the number of edges between groups is given by
\[
c_{rs} = 2E \left[ \frac{c}{B} \delta_{rs} + \frac{1 - c}{B(B-1)} (1 - \delta_{rs}) \right],
\]
where $c \in [0,1]$ determines the affinity between the (dis)assortative groups. For this model, we know that there are critical values
\[
c^*_u = B \pm \frac{B - 1}{B(1/(B-1))}.
\]
such that if $c \in [c^*_u, c^*_u]$ then no algorithm can infer a partition that is correlated to the true one from a single network realization, as it becomes infinitely large $N \to \infty \[42\]$. Starting from a network generated with the PP model, we include triadic closure edges via the global probability $p_u = p$ for every node in the network. Based on the resulting network, we attempt to
recover the original communities, using the SBM and the SBM/TC model. A result of this analysis is shown in Fig. 4 where we compute the maximum overlap \( q \in [0, 1] \) between the inferred \( b \) and true partition \( b \), defined as

\[
q = \max_{\mu} \frac{1}{N} \sum_i \delta_{\mu(b_i), b_i},
\]

where \( \mu(r) \) is a bijection between the group labels in \( \hat{b} \) and \( b \), as well as the effective number of inferred groups \( B_e = e^S \), where \( S \) is the group label entropy

\[
S = -\sum_r \frac{n_r}{N} \ln \frac{n_r}{N}.
\]

As can be seen in Fig. 4a, the presence of triadic closure edges can have a severe negative effect on the recovery of the original partitions when using the SBM. In Fig. 4b we see that the number of groups uncovered can be orders of magnitude larger than the original partition, specially when the latter is not even detectable. This shows that the apparent communities that arise out of the formation of triangles substantially overshadow the underlying true community structure. The situation changes considerably when we use the SBM/TC instead, as shown Fig. 4c. In this case, the presence of triadic closure has no noticeable effect on the detectability of the true community structure, and we obtain a recovery performance indistinguishable from the SBM in the case with no additional edges. As seen in Fig. 4d the same is true for the number of groups inferred. These results seem to point to a robust capacity of the SBM/TC model to reliably distinguish between actual community structure, and the density fluctuations with result from triadic closures.

**IV. EMPIRICAL NETWORKS**

We investigate the use our method with a variety of empirical networks. We begin with a network of cooperation among students while doing their homework for a course at Ben-Gurion University. In Fig. 5a we show the network and a fit of the DC-SBM, which finds 9 assortative communities. Based on this result — and knowing that the partitions found by inferring the SBM as we do here point to statistically significant results that cannot be attributed to mere random fluctuations — we would be tempted to posit that these divisions are uncovering latent social parameters that could explain the observed cooperation between these groups of students. However, if we employ instead the SBM/TC, we obtain the result shown in Fig. 5b, which uncovers instead only a single group, and an abundance of triadic closure edges. This is not unlike the artificial example considered in Fig. 4 and points to a very different interpretation, namely there is no measurable a priori predisposition for students to work with each other in groups, and the resulting network stems instead from students choosing to work together if they already share a mutual partner. Indeed if we inspect the description lengths obtained with each model, we immediately recognize the SBM/TC as the most plausible explanation, and therefore we deem the community structure found by the SBM as an unlikely one by comparison.

We move now to another social network, but this time of friendships between high school students. We show the results of our analysis in Fig. 6. Using the SBM we find \( B = 26 \) groups, shown in Fig. 6a, which at first seems like a reasonable explanation for this network. But instead, with the SBM/TC we find only \( B = 9 \) groups and a substantial amount of triadic closure edges, as shown in Fig. 6b. Differently from the previous example, the SBM/TC still finds enough evidence for a substantial amount of community structure, although with fewer groups than the pure SBM. The groups found with the SBM/TC have a strong correlation with the student grades, as shown in Fig. 6a, except for the 11th and 12th grades, which seem to intermingle more, and for which the model finds evidence of more detailed internal social structures. This indicates that most of the subdivisions of the grades found by the pure SBM are in fact better explained by triadic closure edges, and the a priori friendship preference within these grades are far more homogeneous than the SBM fit would lead us to conclude. One particularly striking feature of this analysis is that it imputes some seemingly clear communities entirely to triadic closure. A good example is the group highlighted with an arrow in Fig. 6a, formed by students in the 8th grade. According to the SBM/TC, this group has arisen due to the formation of triangles between an initially poorly connected subset of students, formed by all friends
Figure 6. (Top panel) Network of friendships between high school students — Adolescent health (comm26) [45]. (a) Fit of the SBM, yielding $B = 26$ communities. (b) Fit of the SBM/TC, uncovering $B = 9$ communities, with seminal (black) and triadic closure (red) edges shown separately in the left and right panels. (Bottom Panel) Network of collaborations between network scientists [46]. (c) Fit of the SBM, yielding $B = 27$ communities. (d) Fit of the SBM/TC, uncovering only $B = 3$ groups, and triadic closure edges shown in red. The thickness of the edges correspond to the marginal probabilities $\pi_{ij}$ and $1 - \pi_{ij}$ for the seminal and closure edges, respectively.

of a single student, rather than an initial affinity between them. Comparing the SBM and the SBM/TC models, we see that the latter has a substantially smaller description length value, and hence needs considerably less information to place all the edges in the network. We emphasize that this criterion takes into account not only the likelihood of the respective model but also on its complexity. In view of this, the SBM/TC hypothesis is objectively more parsimonious, and in the absence of further data should be considered more plausible than the pure SBM.

We move now to an additional example, this time of collaborations between researchers in network science [46], shown in Fig. 6. For this network, the SBM finds $B = 27$ communities. The interpretation here is the same as previous analyses of the same network, namely that these communities are groups that tend to work together, with the occasional collaboration across groups. On the other hand, when we employ the SBM/TC, the difference this time is quite striking. Most of the community structure found with the pure SBM vanishes and is replaced by a substrate network with a substantial “core-periphery” mixing pattern formed of two main groups, where the “core” (blue nodes) is composed of perceived initiators of the collaborations with the “periphery” (yel-
(a) Cooperation between students
(b) Adolescent health (comm26)
(c) Scientific collaborations in Network Science
(d) NCAA college football 2000

Figure 7. Posterior predictive distributions of the clustering coefficient, as described in the text, for the SBM and SBM/TC as indicated in the legend, for different datasets. The vertical line shows the empirical value $C(G)$.

The core-periphery pattern is not perfect, as we observe seminal edges between nodes of every type, but most commonly these exist between core and periphery nodes, and the core nodes themselves, who therefore seem to have a predisposition to wider collaborations. The difference between the description lengths of both models is substantial, indicating that the SBM/TC interpretation is indeed far more plausible.

Lastly, we consider the network of American football games between colleges during the fall of 2000 [47], shown in Fig. 8. For this network we observe an interesting result, namely the SBM and SBM/TC yield the exact same inference, corresponding very closely to the known division of the teams into “conferences” that tend to play with each other more frequently, which means that SBM/TC gives a negligible probability of triadic closure edges. Although we might expect this to occur for a network that has very few or no triangles, and therefore substantial evidence against triadic closure, this is not the case for the particular network in question, which has in fact an abundance of triangles, in addition to clear assortative communities. The reason for this is that, in this particular case, the SBM is fully capable of accounting for the triangles observed, which therefore can be characterized being a “side-effect” of the homophily between nodes of the same group, instead of an excess that needs additional explanation. We will revisit this particular case in the following, from a different angle.

One natural criticism of the SBM as a useful hypothesis for real networks, however stylized as it clearly is, is that it assumes that edges are placed independently with probability $O(B/N)$, for a network with $N$ nodes and $B$ groups, assuming the group affinities are uniform for all groups. One consequence of this is that the probability of observing a spontaneous triadic closure edge will also scale with $O(B/N)$. Therefore if $B \ll N$, we should not expect any abundance of triangles, which is at odds with what we observe in many empirical data. One problem with this logic is that we do not know a priori the precise relationship between $B$ and $N$ for finite empirical networks, and therefore we cannot rule out the SBM hypothesis based simply on an observed abundance of triangles. Auspiciously, with the SBM/TC at hand, we are the perfect position to evaluate the SBM in that regard, and understand how many of the observed triangles can be attributed to an incidental link placement due to community structure, or if they are instead better explained by explicit triadic closure edges. A common way of quantifying the amount of triangles in a network $G$ is via its clustering coefficient $C(G) \in [0, 1]$, which determines the fraction of triads in the network which are closed in a triangle, and is given by

$$C(G) = \frac{\sum_{ijk} G_{ij}G_{jk}G_{ki}}{\sum_{i} k_{i}(k_{i}-1)},$$  \hspace{1cm} (34)$$

where $k_{i} = \sum_{j} G_{ij}$ is the degree of node $i$. A meaningful way to evaluate whether a given model $P(G|\theta)$ with parameters $\theta$ can capture what is seen in the data is to
compute the posterior predictive distribution,

\[ P(C|G) = \sum_{G'} \delta(C - C(G')) \sum_{\theta} P(G'|\theta) P(\theta|G). \]  \hspace{1cm} (35)

This involves sampling parameters \( \theta \) from the posterior \( P(\theta|G) \), generating new networks \( G' \) from the model \( P(G'|\theta) \), and obtaining the resulting population of \( C(G') \) values, which can then be compared to the observed value \( C(G) \), and in this way we can determine if the model used is capable of capturing this aspect of the data. In Fig. 7 we show the results of this comparison for the SBM and SBM/TC (in Appendix D we give more details about how \( \theta \) should be chosen in each case) using three datasets. For three of the four networks we observe what one might expect: although the SBM is capable of accounting for a substantial amount of triangles (far more than one would expect by naively assuming \( B \ll N \)), it falls short of explaining what is actually seen in the data. The SBM/TC, on the other hand, accounts for a realm of possibilities that comfortably includes what is observed in the data, with a sufficiently high probability. For the remaining network in Fig. 7, NCAA college football 2000, as before, we observe a different picture. Namely, both models produce predictive posterior distributions that are essentially identical, and fully compatible with what is seen in the data. Therefore we can say with a fair amount of confidence that the fairly high clustering coefficient observed for this network can in principle be attributed to community structure alone, rather than triadic closure, contradicting the intuition obtained from the asymptotic case where \( B \ll N \), which is not applicable to this network.

We take the opportunity to emphasize that the results of Fig. 7 demonstrate how the SBM/TC model is significantly more well-behaved than ERGMs designed to reproduce triangle counts via a maximum-entropy formulation [25]. As demonstrated in [11, 26–28], these models define ensembles with strong degeneracies, with most sampled networks having either very low or very high triangle counts, but none with values similar to what is actually seen in the modelled networks. This is not a phenomenon we observe with the SBM/TC, where the clustering coefficient distributions are unimodal, and concentrated on the empirical values.

We extend the previous analysis to a larger set of empirical networks, as shown in Fig. 9 by summarizing the compatibility of the posterior predictive distribution via the z-score,

\[ z = \frac{C(G) - \langle C \rangle}{\sigma_C}, \]  \hspace{1cm} (36)

where \( \langle C \rangle \) and \( \sigma_C \) are the mean and standard deviation of the posterior predictive distribution. As we can see, there are a number of networks for which the z-score values lie in the plausible interval \([-2, 2]\) for both models, but there is a much larger fraction of the data for which the values for the SBM point to a decisive incompatibility, whereas the SBM/TC yields credible values more systematically.

Figure 9. (top) Values of the z-score for the posterior predictive distributions of the clustering coefficient, as described in the text, for the SBM and SBM/TC as indicated in the legend, for different datasets. The solid horizontal lines mark the values \(-2\) and \(2\). (middle) Values of the clustering coefficient (Eq. 34), computed for the original network, \( C(G) \), and for the inferred seminal network, \( C_\theta(G) \), averaged over the posterior distribution according to Eq. 37 as shown in the legend. (bottom) Values of effective number of inferred groups, as given by Eq. 33 for the SBM and SBM/TC as indicated in the legend.
We can further exploit the decomposition that the SBM/TC provides by quantifying precisely, for any given network, how much of the observed clustering can be attributed to triadic closure directly, or to community structure indirectly. We can do so by computing the mean clustering coefficient of the substrate seminal network from the posterior distribution,

\[ C_S(G) = \sum_{\{A^{(i)}\}, \{g^{(i)}\}, b} C(A) P(\{A^{(i)}\}, \{g^{(i)}\}, b | G). \]

We can then compare this value with the coefficient for the observed network \( C(G) \), as we show in Fig. 9. We identify a variety of scenarios, including situations where the seminal network (and hence the community structure) accounts for the majority of the observed clustering, but most commonly we observe that a substantial fraction can be attributed to more direct triadic closure. Nevertheless, in many cases the values of \( C_S(G) \) do not drop to negligible values, showing that the presence of triangles cannot be wholly attributed to either mechanism in these cases. Indeed, this variability seems to indicate that mere presence of a high or low density of triangles, as captured by the clustering coefficient, cannot be used by itself to evaluate whether triadic closure or community structure is the leading underlying mechanism of network formation.

Another aspect of the suitability of triadic closure as a more plausible network model is that it tends to come together with a less pronounced inferred community structure, since part of the density heterogeneity found is attributed to the former mechanism, rather than the latter. In Fig. 9 we characterize this difference by the effective number of groups found with both models. We see that the discrepancy between them is once again quite varied, where in some cases it can be quite small, indicating that triadic closure plays a minor role, while in other cases it can be quite extreme, indicating the dominant role that triadic role has in the respective networks.

Overall, what we seem to extract from these empirical networks is that, in the majority of cases (though not all), the observed structure seems to be better explained by a heterogeneous combination of underlying mixing patterns with a further distortion by an additional tendency of forming triangles. The precise balance between these two components vary considerably in general, and needs to be assessed for each individual network.

V. EDGE PREDICTION AND NETWORK RECONSTRUCTION

As with every kind of empirical assessment, network data is subject to measurement errors or omissions. A common use of network models is to predict such erroneous and missing information from what is more precisely known \([48, 49]\). The SBM has been successfully used as such a model \([34, 49]\), since the latent group assignments and the affinities between them can be learned from partial network information, which in turn can be used to infer what has been distorted or left unobserved. Another common approach to edge prediction consists of attributing a higher probability to a potential edge if it happens to form a triangle \([50]\). As we have been discussing in this work, these two properties — group affinity and triadic closure — point to related but distinct

![Figure 10. Distributions of Precision and Recall values, according to the SBM and SBM/TC model, for four empirical networks, and a fraction \( f = 0.05 \) of omitted edges and corresponding number of omitted non-edges. The results were obtained for 200 different realizations of missing edges and non-edges.](image-url)
processes of edge formation, and approaches of edge prediction that rely exclusively on either one will be maximally efficient only if it happens to be the dominant underlying mechanism, which, as we have seen in the last section, is typically not the case. However, with the SBM/TC model we have introduced, it should be possible to accommodate both mechanisms at the same time, and in this way improve edge prediction is more realistic settings. In the following, we show how this can be done, and demonstrate it with a few examples.

The scenario we consider is the general one presented in Ref. [34], where we make \( n_{ij} \) measurements of node pair \((i, j)\) and record the number of times \( x_{ij} \) that an edge has been observed. Based on this data, we infer the underlying network \( G \) according to the posterior distribution

\[
P(G|n, x) = \frac{P(x|G, n)P(G)}{P(x|n)},
\]

with \( n = \{n_{ij}\} \) and \( x = \{x_{ij}\} \). The measurement model corresponds to a situation where the probabilities of observing missing and spurious edges, \( p \) and \( q \) respectively, are uniform, leading to

\[
P(x|G, n, p, q) = \prod_{i<j} \left( \frac{n_{ij}}{x_{ij}} \right) \left[ (1-p)^{x_{ij}} + q^{n_{ij} - x_{ij}} \right]^{G_{ij} \times} \left[ g^{+ij} (1-q)^{n_{ij} - x_{ij}} \right]^{1-G_{ij}}. \tag{39}
\]

Assuming that both \( p \) and \( q \) are unknown a priori, i.e. \( P(p) = P(q) = 1 \), amounts to the marginal probability [31]

\[
P(x|G, n) = \int P(x|G, n, p, q)P(p)P(q) \, dp \, dq \tag{40}
\]

\[
= \left[ \prod_{i<j} \left( \frac{n_{ij}}{x_{ij}} \right) \right] \left( \frac{\mathcal{E}}{\mathcal{E} + 1} \right)^{-1} \frac{1}{\mathcal{E} + 1} \times \left( \frac{\mathcal{M} - \mathcal{E}}{\mathcal{T} - \mathcal{J}} \right)^{-1} \frac{1}{\mathcal{M} - \mathcal{E} + 1}, \tag{41}
\]

where we have

\[
\mathcal{M} = \sum_{i<j} n_{ij}, \quad \mathcal{X} = \sum_{i<j} x_{ij}, \tag{42}
\]

\[
\mathcal{E} = \sum_{i<j} n_{ij} G_{ij}, \quad \mathcal{T} = \sum_{i<j} x_{ij} G_{ij}. \tag{43}
\]

The network model comes into play via the prior \( P(G) \).

For the SBM/TC model this is

\[
P(G) = \sum_{\{g^{(l)}\}, \{A^{(l)}\}, b} P(G, \{g^{(l)}\}, \{A^{(l)}\}, b). \tag{44}
\]

Once more, we avoid an intractable computation, by sampling instead from a joint posterior with the model parameters, i.e.

\[
P(G, \{g^{(l)}\}, \{A^{(l)}\}, b | n, x) = \frac{P(x|G, n)P(G, \{g^{(l)}\}, \{A^{(l)}\}, b)}{P(x|n)} \tag{45}
\]

so that the desired posterior distribution can be obtained by marginalization

\[
P(G|n, x) = \sum_{\{g^{(l)}\}, \{A^{(l)}\}, b} P(G, \{g^{(l)}\}, \{A^{(l)}\}, b | n, x). \tag{46}
\]

In order to perform our comparison, we consider the following particular setup for the data \( (n, x) \). Given a true network \( G \) we select a random subset \( P_t \) of the edges ("true positives"), and an equal-sized random subset \( N_t \) of "non-edges" ("true negatives"), i.e., node pairs \((i, j)\) for which \( G_{ij} = 0 \), such that \( |P_t| = |N_t| = f E \), where \( f \in [0, 1] \) is a free parameter and \( E \) is the total number of edges. We then set \( n_{ij} \rightarrow \infty \) for all node pairs neither in \( P_t \) nor in \( N_t \), with \( x_{ij} = n_{ij} \) if \( G_{ij} = 1 \) and \( x_{ij} = 0 \) otherwise — these are parts of the network about which we are perfectly certain. For the node pairs in \( P_t \) and \( N_t \) we set \( n_{ij} = x_{ij} = 0 \), meaning we lack any data about them. We then compute the posterior marginal probability

\[
p^{ij} = \sum_{G} G_{ij} P(G|n, x), \tag{47}
\]

and use it to evaluate the quality of the reconstruction. We do so by computing the Precision and Recall, defined as

\[
\text{Precision} = \frac{\sum_{(i,j) \in P_t} p^{ij}}{\sum_{(i,j) \in P_t \cup N_t} p^{ij}}, \tag{48}
\]

\[
\text{Recall} = \frac{\sum_{(i,j) \in P_t} p^{ij}}{|P_t|}, \tag{49}
\]

which measures the fraction of correctly predicted edges, relative to the total number of edges predicted, or the total number of true edges, respectively.

In Fig. 10 we show the results of the above analysis for some of the networks studied previously, using both the SBM/TC model and the pure SBM. For most of them, the SBM/TC model yields a superior predictive performance, sometimes substantially. This shows that while community detection via the SBM can to some extent detect the patterns induced by triadic closure, the more explicit SBM/TC model does a better job at this, corroborating the model selection arguments we have used previously. For networks of games between American football college teams, the situation is once again different, and we observe indistinguishable results between the SBM and SBM/TC. For this network, as the previous analysis has established, triadic closure seems to play a insignificant role, despite the relative abundance of triangles. As a consequence, in this case the SBM/TC model offers no advantage in edge prediction, but importantly, it does not degrade it either.
In a recent work, Ghasemian et al. [51] have performed a large-scale analysis of over two hundred edge prediction methods on over five hundred networks belonging to various domains. Although the overall conclusion of that work was that no single method dominates on every data, the predictive performance of the different methods were far from uniform, with the method above based on the SBM providing the single best performance overall. Interestingly, the situations where the SBM approach yielded inferior performance were precisely for social networks, for which some predictors based on triadic closure performed better. Although our results above fall short of a thorough and systematic analysis of the wide domains of network data, since we consider only a handful of networks, they nevertheless seem to give good indication that combining group affinity with triadic closure could potentially eliminate this shortcoming for this particular class of network data.

VI. DISCUSSION

We have presented a generative model and corresponding inference scheme that is capable of differentiating community structure from triadic closure in empirical networks. We have shown that although these features are typically conflated in traditional network analysis, our method can pick them apart, allowing us to tell us whether an observed abundance of triangles is a byproduct of an underlying homophily between nodes, or whether they arise out of a local property of transitivity. Likewise, we have also shown how our method can evade the detection of spurious communities, which are not due to homophily, but arise instead simply out of a random formation of triangles.

Our approach shows how local and global (or mesoscale) generative processes can be combined into a single model. Since it contains a mixture of both mechanisms, our method is able to decompose them for a given observed network according to their inferred contributions. By employing our method on several empirical networks, we were able to demonstrate a wide variety of scenarios, containing everything from a large number of triangles caused predominantly by triadic closures, by a mixture of community structure and triadic closures, and by community structure alone. These findings seem to indicate that local and global network properties tend to mix in nontrivial ways, and we should refrain from automatically concluding that an observed local property (e.g. large number of triangles) cannot have a global cause (e.g. group homophily), and likewise an observed global property (e.g. community structure) cannot have a purely local cause (e.g. triadic closure). Our explicit mixture approach could in principle be extended also to other types of local structures such as reciprocity in directed networks [54], or higher-order motifs, bringing further insights into how these local properties are entangled with global ones.

Several authors had shown before that triadic closure can induce the formation of community structure in networks [11,16]. This introduces a problem of interpretation for community detection methods that do not account for this, which, to the best of our knowledge, happens to be the vast majority of them. This is true also for inference methods based on the SBM, which, although they are not susceptible to finding spurious communities formed by a fully random placement of edges [55] (unlike non-inferential methods, which tend to overfit [59]), they cannot evade those arising from triadic closure [15]. Our approach provides a solution to this interpretation problem, allowing us to reliably rule out triadic closure when identifying communities in networks.

We have also shown how incorporating triadic closure together with community structure can improve edge prediction, without degrading the performance in situations where it is not present. This further demonstrates the usefulness of approaches that model networks in multiple scales, combining multiple edge generation mechanisms, and points to a general way of systematically improving our understanding of network data.

5 Ghasemian et al. [51] considered only a simplified version of the method described, where only the best-scoring partition was used, instead of an average over the posterior distribution. Furthermore, they have used only the version of the SBM with non-informative priors, which is known to underfit, as opposed to the nested SBM [56,52] which removes this problem. Accounting for both of these issues have been shown before to improve edge prediction systematically [53], and could potentially have pushed the result of the analysis in Ref. [51] even more in favor of the SBM approach.

6 Overfitting here means that the number of communities found is too large, and that the method can even find communities in completely random networks.

[1] J. Miller McPherson and Lynn Smith-Lovin, “Homophily in Voluntary Organizations: Status Distance and the Composition of Face-to-Face Groups,” American Sociological Review 52, 370–379 (1987), publisher: [American Sociological Association, Sage Publications, Inc.].

[2] Wesley Shrum, Neil H. Cheek, and Saundra MacD. Hunter, “Friendship in School: Gender and Racial Homophily,” Sociology of Education 61, 227–239 (1988), publisher: [Sage Publications, Inc., American Sociological Association].

[3] James Moody, “Race, School Integration, and Friendship Segregation in America,” American Journal of Sociology 107, 679–716 (2001), publisher: The University of Chicago Press.
missing and Aaron Clauset, Cristopher Moore, and M. E. J. Newman, “Finding community structure in networks,” Proceedings of the National Academy of Sciences 103, 4761–4766 (2006).

[39] Jérôme Kunegis, “KONECT,” in Proceedings of the 22nd International Conference on World Wide Web - WWW ’13 Companion, (ACM Press, 2013).

[40] Michael Fire, Ramu Puzis, and Yuval Elovici, “Organization mining using online social networks,” (2013), arXiv:1303.3741v2 [cs.SI].

[41] Aaron Clauset, Ellen Tucker, and Matthias Sainz, “The Colorado Index of Complex Networks,” (2016), accessible at https://icon.colorado.edu.

[42] James Moody, “Peer influence groups: identifying dense clusters in large networks,” Social Networks 23, 261–283 (2001).

[43] M. E. J. Newman, “Finding community structure in networks using the eigenvectors of matrices,” Physical Review E 74 (2006), 10.1103/physreve.74.036104.

[44] M. Girvan and M. E. J. Newman, “Community structure in social and biological networks,” Proceedings of the National Academy of Sciences 99, 7821–7826 (2002).

[45] James Moody, “Finding community structure in networks using the eigenvectors of matrices,” Physical Review E 74 (2006), 10.1103/physreve.74.036104.

[46] M. Girvan and M. E. J. Newman, “Community structure in social and biological networks,” Proceedings of the National Academy of Sciences 99, 7821–7826 (2002).

[47] Maron Clauset, Cristopher Moore, and M. E. J. Newman, “Hierarchical structure and the prediction of missing links in networks,” Nature 453, 98–101 (2008).

[48] Roger Guimerà and Marta Sales-Pardo, “Missing and spurious interactions and the reconstruction of complex networks,” Proceedings of the National Academy of Sciences 106, 22073–22078 (2009).

[49] Roger Guimerà and Marta Sales-Pardo, “Missing and spurious interactions and the reconstruction of complex networks,” Proceedings of the National Academy of Sciences 106, 22073–22078 (2009).

[50] Lada A Adamic and Eytan Adar, “Friends and neighbors on the Web,” Social Networks 25, 211–230 (2003).

[51] Amir Ghasemian, Homa Hosseinmardi, and Aaron Clauset, “Evaluating Overfit and Underfit in Models of Network Community Structure,” IEEE Transactions on Knowledge and Data Engineering , 1–1 (2019).

[52] Jérôme Kunegis, “KONECT,” in Proceedings of the 22nd International Conference on World Wide Web - WWW ’13 Companion, (ACM Press, 2013).

[53] Michael Fire, Ramu Puzis, and Yuval Elovici, “Organization mining using online social networks,” (2013), arXiv:1303.3741v2 [cs.SI].

[54] Neo D. Martinez, “Artifacts or attributes? effects of resolution in ecological analyses,” Ecology Letters 10, 99–108 (2007).

[55] Michael Fire, Lena Tenenboim-Chekina, Rami Puzis, Ofir Lesser, Lior Rokach, and Yuval Elovici, “Computing Reviews,” 2021.
tationally efficient link prediction in a variety of social networks,” ACM Transactions on Intelligent Systems and Technology 5, 1–25 (2013).

[34] Pablo M. Gleiser and Leon Danon, “Community Structure In Jazz,” Advances in Complex Systems 06, 565–573 (2003).

[37] Wayne W. Zachary, “An information flow model for conflict and fission in small groups,” Journal of Anthropological Research 33, 447–473 (1977).

[38] Donald Ervin Knuth, *The Stanford GraphBase: a platform for combinatorial computing* (AcM Press New York, 1993).

[40] Daniel B. Larremore, Aaron Clauset, and Caroline O. Buckee, “A network approach to analyzing highly recombinant malaria parasite genes,” PLoS Computational Biology 9, e1003268 (2013).

[41] Brian Hayes, “Connecting the dots,” American Scientist 94, 400 (2006).

[42] Boris Pasternak and Ivor Ivask, “Four unpublished letters,” Books Abroad 22, 399 (1970)

[43] Rossana Mastrandrea, Julie Fournet, and Alain 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).

[44] Manlio De Domenico, Andrea Lancichinetti, Alex Arenas, and Martin Rosvall, “Identifying modular flows on multilayer networks reveals highly overlapping organization in social systems,” Phys. Rev. X 5, 011027 (2015), arXiv:1408.2925v1 [physics.soc-ph].

[45] Boris Pasternak and Tvo Ivask, “Four unpublished letters,” Books Abroad 44, 196 (1970)

[46] Rossana Mastrandrea, Julie Fournet, and Alain 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).

[47] R. Guimerà, L. Danon, A. Diaz-Guilera, F. Giralt, and A. Arenas, “Self-similar community structure in a network of human interactions,” Physical Review E 68, 036123 (2003), 10.1103/physreve.68.066103.

**Appendix A: Latent multigraph SBM**

The marginal likelihood of Eq. 1 is in fact obtained for a multigraph model [46], where the adjacency entries can take any natural value, \( A_{ij} \in \mathbb{N} \). Although we could in principle ignore this discrepancy, since this kind of model generates simple graphs as a special case, this comes at the expense of a reduced expressiveness of the model [37], since this kind of multigraph model cannot describe the placement of single edges with high probability, or account for the emergent degree-degree correlations that must be present in simple graphs. Instead, here we take the approach proposed in [34, 37], and consider a **latent multigraph** \( \mathbf{A}' \), with \( A'_{ij} \in \mathbb{N} \), which is then converted into a simple graph \( \mathbf{A}(\mathbf{A}') \) simply by ignoring the edge multiplicities, i.e.

\[
A_{ij} = \begin{cases} 
1 & \text{if } A'_{ij} > 0, \\
0 & \text{otherwise}.
\end{cases}
\]  

The latent multigraph \( \mathbf{A}' \) is generated according to Eq. 1, which means the simple graph \( \mathbf{A} \) is generated according to

\[
P(\mathbf{A} | \mathbf{b}) = \sum_{\mathbf{A}'} \mathbf{1}_{\{A(A') = \mathbf{A}\}} P(\mathbf{A}' | \mathbf{b}).
\]  

Instead of working with this marginal probability directly (which is intractable), we infer the latent edge multiplicities as well, from a joint posterior distribution

\[
P(\mathbf{g}, \mathbf{A}', \mathbf{b} | \mathbf{G}) = \frac{P(\mathbf{G}, \mathbf{g} | \mathbf{A}')} {P(\mathbf{G})} P(\mathbf{A}' | \mathbf{b}) P(\mathbf{b}),
\]  

where the simple graph \( \mathbf{A}(\mathbf{A}') \) is used for the triadic closure likelihood \( P(\mathbf{G}, \mathbf{g} | \mathbf{A}) \). In this way, the inference procedure is the same as the one described in the main text, with the only modification that we need to infer the integer values of \( \mathbf{A}' \) rather than its binary values.

**Appendix B: Expected density of transitivity**

As mentioned in the main text, the choice of priors used for Eq. 10 makes the calculation very simple, but it implies that we expect the observed graphs to always have a large fraction of triadic closures. An outcome of this is that the probability of observing a final graph without any triadic closure, i.e. \( \sum_{i<j} g_{ij}(u) = 0 \), is given by

\[
P(\mathbf{g}' | \mathbf{A}) = \prod_u \left[ 1 + \sum_{i<j} m_{ij}(u) \right]^{-1}
\]  

which is exponentially suppressed for a large number of nodes \( N \). Even though we are interested in modeling networks which do posses some amount of triadic closure, we should be **a priori** more agnostic about the actual amount, as to also accommodate situations where this property is not abundant. We can address this by noting that the likelihood of Eq. 10 can be alternatively interpreted as the one of a fully equivalent model given by

\[
P(\mathbf{g} | \mathbf{A}) = \prod_u \sum_{\mathbf{E}_u} P(\mathbf{g}(u) | \mathbf{A}, \mathbf{E}_u) P(\mathbf{E}_u | \mathbf{A}),
\]  

where

\[
P(\mathbf{g}(u) | \mathbf{A}, \mathbf{E}_u) = \frac{1}{\sum_{\mathbf{E}_u} \mathbf{g}(u)}
\]  

is the probability of uniformly sampling an ego graph \( \mathbf{g}(u) \) with exactly \( \mathbf{E}_u = \sum_{i<j} g_{ij}(u) \) edges, and

\[
P(\mathbf{E}_u | \mathbf{A}) = \frac{1}{1 + \sum_{i<j} m_{ij}(u)}
\]
is the probability of uniformly sampling the number of edges in $g(u)$ in the allowed range $[0, \sum_{i<j} m_{ij}(u)]$. This interpretation allows us to do a small modification of our model that makes it more versatile, namely we separate the nodes into two sets, according to an auxiliary binary variable $t_u \in \{0, 1\}$, such that if $t_u = 0$ then the corresponding ego graph has no edges, $P(E_u | A, t_u = 0) = \delta_{0,E_u}$, otherwise it has a nonzero number of edges, sampled uniformly as

$$P(E_u | A, t_u = 1) = \frac{1 - \delta_{E_u,0}}{\sum_{i<j} m_{ij}(u)}.$$  

The modified marginal distribution becomes then

$$P(g | A) = \sum_t \left[ \prod_u \sum_{E_u} P(g(u) | A, E_u) P(E_u | A, t) \right] P(t | A)$$

$$= \left\{ \prod_u \left[ \left( \frac{\sum_{i<j} m_{ij}(u)}{\sum_{i<j} g_{ij}(u)} \right) \sum_{i<j} m_{ij}(u) \right] \right\} \delta_{1, n[\sum_{i<j} g_{ij}(u)]}$$

$$\times \frac{1}{1 + \sum_u \Theta \left[ \sum_{i<j} m_{ij}(u) \right]}$$

$$\times \frac{1}{1 + \sum_u \Theta \left[ \sum_{i<j} m_{ij}(u) \right]}$$

$$\sum_u g_{ij}(u) = 0,$$ it is generated with probability

$$P(g' | A) = \frac{1}{1 + \sum_u \Theta \left[ \sum_{i<j} m_{ij}(u) \right]} = O \left( \frac{1}{N} \right),$$

which is relatively large and no longer exponentially suppressed for large $N$, meaning that our model can also accommodate the same kinds of networks that are sampled from the pure SBM, without triadic closures. This does not mean that the modified model generates typical networks with substantially smaller number of transitive edges, only that the variance with respect to this property is larger, and the model is thus more indifferent about the potential networks that are possible to be observed.

As mentioned in the main text, this modification makes the SBM fully nested inside the SBM/TC, as we have

$$P(G, g', A = G | b) = \frac{P(G | b)}{1 + \sum_u \Theta \left[ \sum_{i<j} m_{ij}(u) \right]}$$

$$\geq \frac{P(G | b)}{N + 1},$$

with $g'$ being empty ego graphs, and the last equality achieved if $\sum_{i<j} m_{ij}(u) > 0$ for every node $u$.

1. **Iterated triadic closure**

For the generalized model with iterated triadic closures, the marginal likelihood is also analogous to Eq. [B8]
We then sample another generation which is independent on the value of the edge is given by random. The number of ego graphs that are relevant for the current generation implemented with the following moves. The first kind is to select again an edge such that \( l \) are always rejected in this way. To select first an edge \((i,j)\) in \( G \) as well as a generation \( l \), both uniformly at random, and an ego node \( u \) that is relevant to edge \((i,j)\) at generation \( l \), also uniformly at random. The number of ego graphs that are relevant for this edge is given by

\[
u_{ij}^{(l)} = \sum_u A_{ui}^{(l-1)} (1 - A_{uj}^{(l-1)}),
\]

which is independent on the value of \( g_{ij}^{(l)}(u) \) for any \( l' \). We then sample another generation \( l' \neq l \) and proceed

\[
\min \left( 1, \frac{P(\{g^{(l')}\}, \{A^{(l')}\}, b|G)P(i, j, u, v, l, l' | \{g^{(l')}\}, \{A^{(l')}\})}{P([g^{(l')}], \{A^{(l')}\}, b|G)P(i, j, v, u, l', l | \{g^{(l')}\}, \{A^{(l')}\})} \right) = \min \left( 1, \frac{P(\{g^{(l')}\}, \{A^{(l')}\}, b|G)}{P([g^{(l')}], \{A^{(l')}\}, b|G)} \right)
\]

which is independent on the actual move probabilities, since they remain the same after and before the move. Note that invalid moves that result in \( g_{ij}^{(l)} < 0 \) or \( A_{ij}^{(l)} < 0 \) are always rejected in this way.

In addition, we also make a second kind of move by selecting again an edge \((i,j)\) in \( G \) as well as a generation \( l \), both uniformly at random, and an ego node \( u \) that is relevant to edge \((i,j)\) at generation \( l \), with the same probability as before. We then move the move \( g_{ij}^{(l)} = g_{ij}^{(l)} + 1 \) with probability \( 1/2 \), and accept again according to

\[
\min \left( 1, \frac{P(\{g^{(l')}\}, \{A^{(l')}\}, b|G)}{P([g^{(l')}], \{A^{(l')}\}, b|G)} \right).
\]

In case \( l = 0 \) is selected, the move is different, due to the multigraph nature of \( A \). We make instead the proposal \( A_{ij} \rightarrow A_{ij}' \) according to a geometric distribution with mean \( A_{ij} + 1 \).

\[
P(A_{ij}'|A_{ij}) = \frac{A_{ij} + 1}{A_{ij} + 2} \frac{1}{A_{ij} + 2}.
\]
1. Algorithmic complexity

We can break down the time complexity of the above algorithm as follows. At any given time, we keep track of all relevant ego graphs for each edge \((i, j)\) in \(G\), those that have edge \((i, j)\) in them, as well as the number of edges \(E_n^{(l)} = \sum_{i<j} \delta_{ij}^{(l)}(u)\) of every ego graph. Based on this bookkeeping, whenever an entry \(g_{i,j}^{(l)}(u)\) (or \(A_{ij}\) if \(l = 0\)) is modified, to compute the log-likelihood difference we need only to evaluate the common neighbors of \(i\) and \(j\) or the new or removed open and closed triads \((i, j, v)\) or \((v, i, j)\) that affect generation \(l + 1\), both of which can be computed in \(O(k_i + k_j)\). As a result, a whole “sweep” of the MCMC algorithm, where each edge in \(G\) had a chance to be moved by one of the proposals considered, has an overall complexity of \(O(N(k^2))\), since each node \(i\) has \(k_i\) edges that need to be moved at every sweep, each of which requiring time \(O(k_i + k_j)\), with \(j\) being the other endpoint.

For the partition part of the algorithm, the overall complexity of a sweep, where every node had a chance to be moved to a different group, is \(O(E + N)\), independent of the number of groups being occupied [10].

Combining the two kinds of moves gives us an overall complexity of \(O(N(k^2))\) per sweep, which for sparse graphs with \((k^2) = O(1)\) amounts to \(O(N)\). This means that it is possible, at least in principle, to apply this algorithm for very large networks.

On top of the time it takes to perform sweeps of the MCMC chain, there is also the mixing time of the Markov chain, which determines how long one needs to wait before usable samples from the posterior distribution are made. It is difficult to estimate the mixing time, as it depends heavily on the actual network structure being considered, but we found that the algorithm gives usable results in reasonable time even for networks with over a hundred thousand to a million edges, although we did not attempt a detailed investigation of networks which are much larger than this.

We have evaluated the quality of the algorithm with the analysis presented in Fig. 3, where networks from the SBM/TC model were generated, and the inference was performed in them. By comparing the obtained results with the true values of the latent parameters, we observed that the triadic closure component was identified with excellent accuracy, and the SBM component was identified with an accuracy indistinguishable from when considering only the pure SBM case, all the way down to the detectability transition. This gives us a very good amount of confidence that the method converges, at least in controlled scenarios.

When applied to empirical networks, the diagnostics performed were to run the algorithm many times and evaluate if similar results are produced, which happened to be the case with the data analyzed.

A reference implementation of this algorithm is freely made available as part of the graph-tool library [57].

Appendix D: Predictive posterior distribution

The predictive posterior distribution considered in the main text is

\[
P(C|G) = \sum_{\theta} \delta(C - C(G')) \sum_{\theta} P(G'|\theta) P(\theta|G),
\]

(D1)

where \(\theta\) are the parameters of model \(P(G|\theta)\). Here we specify more precisely how these parameters are chosen and sampled for the SBM/TC model. The marginal likelihood for the SBM given by Eq. (1) can be written equivalently as [36]

\[
P(A|b) = P(A|k, e, b) P(k|e, b) P(e|b),
\]

(D2)

where the likelihood of the microcanonical DC-SBM is given by

\[
P(A|k, e, b) = \prod_{i<j} A_{ij}! \prod_i A_i! \prod_k k_i! \exp\left(-\sum_{l=1}^{k_i} \frac{k_i!}{k_i!}\right),
\]

(D3)

and prior for the degrees is

\[
P(k|e, b) = \frac{\prod_k k_i!}{n_i! i_q(e,r,n)},
\]

(D4)

and the prior for the edge counts between groups is

\[
P(e|b) = \left( \frac{2(B+1)}{E} + E - 1 \right)^{-1}.
\]

(D5)

For the triadic closure edges we have the likelihood \(P(g(u)|A, p_u)\) of Eq. (4), which, given a uniform prior \(p_u = 1\), gives us a Beta posterior distribution

\[
P(p_u|g(u), A) = \frac{\sum_{i<j} g_{ij}(u)m_{ij}(u) (1 - p_u) \sum_{i<j} (1 - g_{ij}(u)) m_{ij}(u)}{B \left( \sum_{i<j} g_{ij}(u)m_{ij}(u), \sum_{i<j} (1 - g_{ij}(u)) m_{ij}(u) \right)},
\]

(D6)

where \(B(x, y)\) is the Beta function. Based on this parametrization, our predictive posterior distribution is obtained by setting \(\theta = (\{p^{(l)}\}, k, e, b)\), amounting to
\[ P(C|G) = \sum_{\{g^{(l)}\}} \int d(p^{(l)}) \delta \left\{ C - C \left[ G(A, \{g^{(l)}\}) \right] \right\} \left[ \prod_{l,u} P(g^{(l)}(u)|p^{(l)}_u, A) \right] P(A|k, e, b) \times \left[ \prod_{l,u} P(p^{(l)}_u|g^{(l)}(u), A') \right] P(\{g^{(l)}\}, A', k, e, b|G). \]

Operationally, this just means running our inference algorithm to obtain our latent variables \( \{g^{(l)}\}, \{A^{(l)}\}, k, e \) and \( b \), and the triadic closure propensities \( p^{(l)} \) from its posterior, using that to obtain a new seminal network \( A \) from the same SBM, together with its new ego graphs \( \{g^{(l)}\} \), and then finally computing the resulting clustering coefficient.

**Appendix E: Network datasets**

Below are descriptions of the network datasets used in this work. The codenames in parenthesis correspond to the respective entries in the Netzschleuder repository \[58\] where the networks can be downloaded. Some of the descriptions were obtained from the Colorado Index of Complex Networks \[59\].

**Adolescent health (add_health) [45]:** 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 1994-95. Some communities had only one school; others had two. Where there are two schools in a community students from one school were allowed to name friends in the other, the “sister school”. For this analysis, a symmetrized version of the original directed network has been used, considering only its largest connected component. The particular network named comm26 has been used. This network has \( N = 551 \) nodes and \( E = 2624 \) edges.

**Scientific collaborations in physics (arxiv_collab) [60]:** Collaboration graphs for scientists, extracted from the Los Alamos e-Print arXiv (physics), for 1995-1999 for three categories, and additionally for 1995-2003 and 1995-2005 for one category. For copyright reasons, the MEDLINE (biomedical research) and NCSTRL (computer science) collaboration graphs from this paper are not publicly available. For this analysis, only the largest connected component of the networks were considered. The particular networks named cond-mat-1999, hep-th-1999 have been used,

with number of nodes and edges, \( (N, E) \), given by \( (13861, 44619) \), \( (5835, 13815) \), respectively.

**Metabolic network (celegans_metabolic) [61]:** List of edges comprising the metabolic network of the nematode *C. elegans*. This network has \( N = 453 \) nodes and \( E = 4596 \) edges.

**C. elegans neurons (celegansneural) [62, 63]:** A network representing the neural connections of the Caenorhabditis elegans nematode. For this analysis, a symmetrized version of the original directed network has been used. This network has \( N = 297 \) nodes and \( E = 2359 \) edges.

**Collins yeast interactome (collins_yeast) [64]:** Network of protein-protein interactions in Saccharomyces cerevisiae (budding yeast), measured by co-complex associations identified by high-throughput affinity purification and mass spectrometry (AP/MS). For this analysis, only the largest connected component of the network was considered. This network has \( N = 1004 \) nodes and \( E = 8319 \) edges.

**DNC emails (dnc) [65]:** A network representing the exchange of emails among members of the Democratic National Committee, in the email data leak released by WikiLeaks in 2016. For this analysis, only the largest connected component of the network was considered. This network has \( N = 849 \) nodes and \( E = 12038 \) edges.

**Dolphin social network (dolphins) [66]:** An undirected social network of frequent associations observed among 62 dolphins (Tursiops) in a community living off Doubtful Sound, New Zealand, from 1994-2001. This network has \( N = 62 \) nodes and \( E = 159 \) edges.

**Ego networks in social media (ego_social) [67]:** Ego networks associated with a set of accounts of three social media platforms (Facebook, Google+, and Twitter). Datasets include node features (profile metadata), circles, and ego networks, and were crawled from public
sources in 2012. For this analysis, only the largest connected component of the network was considered. The particular network named facebook_0 has been used. This network has $N = 324$ nodes and $E = 2514$ edges.

Maier Facebook friends (facebook_friends) [68]: 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. Metadata gives the social context for the relationship between ego and alter. For this analysis, only the largest connected component of the network was considered. This network has $N = 329$ nodes and $E = 1954$ edges.

Within-organization Facebook friendships (facebook_organizations) [69]: Six networks of friendships among users on Facebook who indicated employment at one of the target corporation. Companies range in size from small to large. Only edges between employees at the same company are included in a given snapshot. Node metadata gives listed job-type on the user’s page. The particular networks named S1, S2 have been used, with number of nodes and edges, $(N, E)$, given by (320, 2369), (165, 726), respectively.

Little Rock Lake food web (foodweb_little_rock) [70]: A food web among the species found in Little Rock Lake in Wisconsin. Nodes are taxa (like species), either autotrophs, herbivores, carnivores or decomposers. Edges represent feeding (nutrient transfer) of one taxon on another. For this analysis, a symmetrized version of the original directed network has been used. This network has $N = 183$ nodes and $E = 2494$ edges.

NCAA college football 2000 (football) [47]: A network of American football games between Division IA colleges during regular season Fall 2000. This network has $N = 115$ nodes and $E = 613$ edges.

Game of Thrones coappearances (game_thrones) [71]: Network of coappearances of characters in the Game of Thrones series, by George R. R. Martin, and in particular coappearances in the book “A Storm of Swords.” Nodes are unique characters, and edges are weighted by the number of times the two characters’ names appeared within 15 words of each other in the text. This network has $N = 107$ nodes and $E = 352$ edges.

Google+ (google_plus) [72]: Snapshot of connections among users of Google+, collected in 2012. Nodes are users and a directed edge $(i, j)$ represents user $i$ added user $j$ to $i$’s circle. For this analysis, a symmetrized version of the original directed network has been used, considering only its largest connected component. This network has $N = 201949$ nodes and $E = 1496936$ edges.

Jazz collaboration network (jazz_collab) [73]: 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. This network has $N = 198$ nodes and $E = 2742$ edges.

Zachary Karate Club (karate) [74]: Network of friendships among members of a university karate club. Includes metadata for faction membership after a social partition. Note: there are two versions of this network, one with 77 edges and one with 78, due to an ambiguous typo in the original study. (The most commonly used is the one with 78 edges.). The particular network named 78 has been used. This network has $N = 34$ nodes and $E = 78$ edges.

Les Misérables coappearances (lesmis) [75]: The network of scene coappearances of characters in Victor Hugo’s novel “Les Misérables.” Edge weights denote the number of such occurrences. This network has $N = 77$ nodes and $E = 254$ edges.

Malaria var DBLa HVR networks (malaria_genes) [76]: Networks of recombinant antigen genes from the human malaria parasite P. falciparum. Each of the 9 networks shares the same set of vertices but has different edges, corresponding to the 9 highly variable regions (HVRs) in the DBLa domain of the var protein. Nodes are var genes, and two genes are connected if they share a substring whose length is statistically significant. Metadata includes two types of node labels, both based on sequence structure around HVR6. For this analysis, only the largest connected component of the network was considered. The particular network named HVR_9 has been used. This network has $N = 297$ nodes and $E = 7562$ edges.

Scientific collaborations in network science (netscience) [48]: 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. For this analysis, only the largest connected component of the network was considered. This network has $N = 379$ nodes and $E = 914$ edges.

Physician trust network (physician_trust) [77]: A network of trust relationships among physicians in four midwestern (USA) cities in 1966. Edge direction indicates that node $i$ trusts or asks for advice from node $j$. Each of the four components represent the network within a given city. For this analysis, a symmetrized version of the original directed network has been used, considering only its largest connected component. This network has $N = 117$ nodes and $E = 542$ edges.
Multilayer physicist collaborations \cite{physics_collab}: Two multiplex networks of coauthorships among the Pierre Auger Collaboration of physicists (2010-2012) and among researchers who have posted preprints on arXiv.org (all papers up to May 2014). Layers represent different categories of publication, and an edge’s weight indicates the number of reports written by the authors. These layers are one-mode projections from the underlying author-paper bipartite network. For this analysis, only the largest connected component of the network was considered. The particular network named pierreAuger has been used. This network has $N = 475$ nodes and $E = 7090$ edges.

Political books network \cite{polbooks}: A network of books about U.S. politics published close to the 2004 U.S. presidential election, and sold by Amazon.com. Edges between books represent frequent copurchasing of those books by the same buyers. The network was compiled by V. Krebs and is unpublished. This network has $N = 105$ nodes and $E = 441$ edges.

High school temporal contacts \cite{sp_high_school}: These data sets correspond to the contacts and friendship relations between students in a high school in Marseilles, France, in December 2013, as measured through several techniques. For this analysis, symmetrized versions of the original directed networks have been used, considering only their largest connected component. The particular networks named diaries, survey, facebook have been used, with number of nodes and edges, $(N, E)$, given by $(120, 502), (128, 658), (156, 1437)$, respectively.

Student cooperation \cite{student_cooperation}: Network of cooperation among students in the "Computer and Network Security" course at Ben-Gurion University, in 2012. Nodes are students, and edges denote cooperation between students while doing their homework. The graph contains three types of links: Time, Computer, Partners. For this analysis, only the largest connected component of the network was considered. This network has $N = 141$ nodes and $E = 297$ edges.

9-11 terrorist network \cite{terrorists_911}: Network of individuals and their known social associations, centered around the hijackers that carried out the September 11th, 2001 terrorist attacks. Associations extracted after-the-fact from public data. Metadata labels say which plane a person was on, if any, on 9/11. This network has $N = 62$ nodes and $E = 152$ edges.

Madrid train bombing terrorists \cite{train_terrorists}: A network of associations among the terrorists involved in the 2004 Madrid train bombing, as reconstructed from press stories after-the-fact. Edge weights encode four levels of connection strength: friendships, ties to Al Qaeda and Osama Bin Laden, co-participants in wars, and co-participants in previous terrorist attacks. This network has $N = 64$ nodes and $E = 243$ edges.

Email network \cite{uni_email}: A network representing the exchange of emails among members of the Rovira i Virgili University in Spain, in 2003. For this analysis, a symmetrized version of the original directed network has been used. This network has $N = 1133$ nodes and $E = 10903$ edges.