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

(Received 11 May 2021; revised 13 September 2021; accepted 26 October 2021; published 6 January 2022)

Network homophily, the tendency of similar nodes to be connected, and transitivity, the tendency of two nodes to be 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.

DOI: 10.1103/PhysRevX.12.011004 Subject Areas: Complex Systems, Statistical Physics

I. INTRODUCTION

One of the most typical properties of social networks is the presence of homophily [1–4], 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 [5]. Another pervasive pattern encountered in the same kind of network is transitivity [6–8], 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 [9] and triadic closure [10], respectively, they are difficult to distinguish in nonlongitudinal data. This is because both processes can result in the same kind 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 form can induce the formation of groups of nodes with a higher density of connections between them, when compared to the rest of the network [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.* [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.* [13] showed that it is marked by a spontaneous size-dependent formation of "triangle clusters." Bianconi *et al.* [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; they have shown that it is capable of producing networks with emergent community structure. The effect of triangle

*peixotot@ceu.edu

formation on apparent community structure has been further studied by Wharrie *et al.* [15], who showed that those patterns can even be misleading for methods specifically designed to avoid the detection of spurious communities in random networks. More recently, Asikainen *et al.* [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 the 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 macroscale structures that emerge spontaneously from microscale higher-order interactions [17,18], and in this way, we can separate them from inherently macroscale structures.

It is also worth mentioning some recent methods that have been proposed that use triangles as a means of finding communities in networks [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 [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) [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 [11,25–28], rendering them rather implausible models for clustered networks. Additionally, when they are combined with homophily, this is 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) [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, which state that this kind of disentanglement is not possible from a single network observation. In particular, Chang *et al.* [31] considered a scenario of uncertain network measurement and proved that, absent 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 [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 *a priori*). However, whenever such stipulations are made, the situation changes. In particular, McFowland III and Shalizi [33] have shown that as soon as the homophilic traits are latent (instead of being observed directly as considered in Ref. [32]) 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 [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 *a priori*. Although the disentanglement problem that we consider here is different from the aforementioned ones, and the impossibility results do not carry over, we nevertheless make use of the same kinds of modeling assumptions that make the other problems feasible.

Our paper 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 to 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 *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.

We denote the seminal edges with an adjacency matrix $\boldsymbol{A}$, and for its generation, we use the degree-corrected stochastic block model (DC-SBM) [35], conditioned on a partition $\boldsymbol{b}$ of the nodes into B groups, where $b_i \in [1, B]$ is the group membership of node i, which has a marginal distribution given by [36]

$$P(\boldsymbol{A}|\boldsymbol{b}) = \frac{\prod_{r<s} e_{rs}! \prod_r e_{rr}!! \prod_i k_i!}{\prod_{i<j} A_{ij}! \prod_i A_{ii}!! \prod_r e_r!} \times \prod_r \frac{\prod_k \eta_k^r!}{n_r! q(e_r, n_r)} \times \left(\begin{matrix} \frac{B(B+1)}{2} + E - 1 \\ E \end{matrix}\right)^{-1}, \tag{1}$$

where $e_{rs} = \sum_{ij} A_{ij} \delta_{b_i,r} \delta_{b_j,s}$ is the number of edges between groups r and s (or twice that for $r = s$), $e_r = \sum_s e_{rs}$, $k_i = \sum_j A_{ij}$ is the degree of node i, $n_r = \sum_i \delta_{b_i,r}$ is the number of nodes in group r, $\eta_k^r = \sum_i \delta_{b_i,r} \delta_{k_i,k}$ is the number of nodes in group r with degree k, $E = \sum_{ij} A_{ij}/2$ is the total number of edges, and $q(m, n)$ is the number of restricted partitions of integer m into at most n parts. We refer to Ref. [36] for a detailed derivation of this marginal likelihood, including also the extension for hierarchical partitions that is straightforward to incorporate, as well as latent multigraphs [37] (see the Appendix A), both of which we used in our analysis. This model is capable of generating networks with arbitrary degree distributions and mixing patterns between groups of nodes, including homophily [30,38].

The triadic closure edges are represented by an additional set of N "ego" graphs $\boldsymbol{g}$, attributed to each node u of $\boldsymbol{A}$, where $\boldsymbol{g}(u)$ is the ego graph of node u. The ego graph $\boldsymbol{g}(u)$ is only allowed to contain nodes that are neighbors of

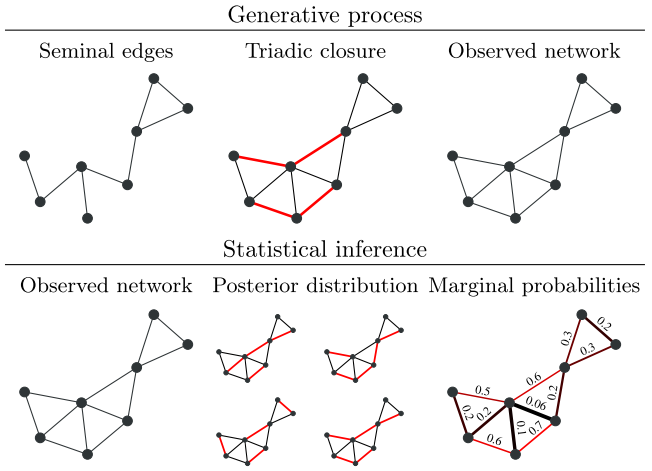


FIG. 1. Schematic representation of the generative process considered (top) and the associated inference procedure (bottom). The generative process consists in the placement of seminal edges according to a SBM and the addition of triadic closure edges conditioned on the seminal edges (shown in red). The inference procedure runs in the reverse direction, and given an observed graph, it produces a posterior distribution of possible divisions of seminal and triadic closure edges, with which edge marginal probabilities on the edge identities can be obtained.

u in $\boldsymbol{A}$ (excluding u itself) and edges that do not exist in $\boldsymbol{A}$, so any existing edge in $\boldsymbol{g}(u)$ amounts to a triadic closure in $\boldsymbol{A}$. The adjacency of $\boldsymbol{g}(u)$ is given by

$$g_{ij}(u) = \begin{cases} 1 & \text{if } (i,j) \in \boldsymbol{g}(u) \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

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

$$m_{ij}(u) = A_{ui} A_{uj} (1 - A_{ij}), \tag{3}$$

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 $\boldsymbol{g}(u)$ only if $m_{ij}(u) = 1$. Based on this, the ego networks are generated independently with probability

$$P(\boldsymbol{g}(u)|\boldsymbol{A}, p_u) = \prod_{i<j} [p_u m_{ij}(u)]^{g_{ij}(u)} [1 - p_u m_{ij}(u)]^{1 - g_{ij}(u)}, \tag{4}$$

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

$$G_{ij}(\boldsymbol{A},\boldsymbol{g}) = \begin{cases} 1 & \text{if } A_{ij} + \sum_u g_{ij}(u) > 0 \\ 0 & \text{otherwise.} \end{cases} \tag{5}$$

The joint probability of the above process is then given by

$$P(\boldsymbol{G},\boldsymbol{g},\boldsymbol{A}|\boldsymbol{p},\boldsymbol{b}) = \mathbf{1}_{\{\boldsymbol{G}=\boldsymbol{G}(\boldsymbol{A},\boldsymbol{g})\}} P(\boldsymbol{A}|\boldsymbol{b}) \prod_u P(\boldsymbol{g}(u)|\boldsymbol{A}, p_u), \tag{6}$$

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

$$P(\boldsymbol{G}) = \sum_{\boldsymbol{g},\boldsymbol{A},\boldsymbol{b}} \int P(\boldsymbol{G},\boldsymbol{g},\boldsymbol{A}|\boldsymbol{p},\boldsymbol{b}) P(\boldsymbol{p}) P(\boldsymbol{b}) \mathrm{d}\boldsymbol{p}, \tag{7}$$

with $P(\boldsymbol{p})$ and $P(\boldsymbol{b})$ being prior probabilities, does not lend itself to a tractable computation. Luckily, however, this is not needed for our inference procedure. Instead, we are interested in the posterior distribution

$$P(\boldsymbol{g},\boldsymbol{A},\boldsymbol{b}|\boldsymbol{G}) = \frac{P(\boldsymbol{G},\boldsymbol{g},\boldsymbol{A}|\boldsymbol{b})P(\boldsymbol{b})}{P(\boldsymbol{G})}, \tag{8}$$

which describes the probability of a decomposition of an observed simple graph $\boldsymbol{G}$ into its seminal graph $\boldsymbol{A}$, the underlying community structure $\boldsymbol{b}$, and the triadic closures represented by the ego graphs $\boldsymbol{g}$. [Although the marginal distribution $P(\boldsymbol{G})$ appears in the denominator of the above equation, we will see later that it is just a normalization constant that does not, in fact, need to be computed.] The marginal likelihood

$$P(\boldsymbol{G},\boldsymbol{g},\boldsymbol{A}|\boldsymbol{b}) = P(\boldsymbol{G}|\boldsymbol{A},\boldsymbol{g})P(\boldsymbol{g}|\boldsymbol{A})P(\boldsymbol{A}|\boldsymbol{b}) \tag{9}$$

can be computed easily via

$$\begin{aligned} P(\boldsymbol{g}|\boldsymbol{A}) &= \prod_u \int_0^1 P(\boldsymbol{g}(u)|\boldsymbol{A}, p) P(p) \mathrm{d}p \\ &= \prod_u \left[\binom{\sum_{i<j} m_{ij}(u)}{\sum_{i<j} g_{ij}(u)}^{-1} \frac{1}{1 + \sum_{i<j} m_{ij}(u)} \right], \end{aligned} \tag{10}$$

where we have used a uniform prior $P(p) = 1$ and omitted, for simplicity, an indicator function setting $P(\boldsymbol{g}|\boldsymbol{A}) = 0$ if $g_{ij} > 0$ and $m_{ij} = 0$ for any (i,j), and with the remaining likelihood term being only the indicator function, $P(\boldsymbol{G}|\boldsymbol{A},\boldsymbol{g}) = \mathbf{1}_{\{\boldsymbol{G}=\boldsymbol{G}(\boldsymbol{A},\boldsymbol{g})\}}$. 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 B, 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 naturally leads to a dynamical model, where generations of triadic closures are progressively introduced to the network. [39] We can incorporate this in our model via "layers" of ego graphs $\boldsymbol{g}^{(l)}$ representing edges 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

$$A_{ij}^{(l)} = \begin{cases} 1 & \text{if } A_{ij}^{(l-1)} + \sum_u g_{ij}^{(l)}(u) > 0 \\ 0 & \text{otherwise,} \end{cases} \tag{11}$$

with boundary conditions $\boldsymbol{A}^{(0)} = \boldsymbol{A}$ (henceforth, $\boldsymbol{A}$ refers solely to the seminal network, whereas, e.g., $\boldsymbol{A}^{(1)}$ is the resulting network considering the first iteration of triadic closures, and $\{\boldsymbol{A}^{(l)}\}$ refers to the set of all generations, including the seminal network), and $\boldsymbol{g}^{(0)}(u)$ being empty graphs for all u, and we denote the final generation as $\boldsymbol{A}^{(L)} = \boldsymbol{G}$. The formation of new triadic closure layers is done according to the probability

$$\begin{aligned} &P(\boldsymbol{g}^{(l)}(u)|\boldsymbol{A}^{(l-1)}, \boldsymbol{g}^{(l-1)}, p_u^{(l)}) \\ &\quad = \prod_{i<j} \left[p_u^{(l)} m_{ij}^{(l)}(u) \right]^{g_{ij}^{(l)}(u)} \left[1 - p_u^{(l)} m_{ij}^{(l)}(u) \right]^{1 - g_{ij}^{(l)}(u)}, \end{aligned} \tag{12}$$

where an open triad (i,u,j) at generation l is denoted by

$$m_{ij}^{(l)}(u) = w_{ij}^{(l)}(u)(1 - A_{ij}^{(l-1)}), \tag{13}$$

so that $m_{ij}(u) \in \{0,1\}$, where

$$w_{ij}^{(l)}(u) = \begin{cases} 1 & \text{if } A_{ui}^{(l-1)} \sum_v g_{uj}^{(l-1)}(v) + A_{uj}^{(l-1)} \sum_v g_{ui}^{(l-1)}(v) > 0 \\ 0 & \text{otherwise} \end{cases} \tag{14}$$

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) 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

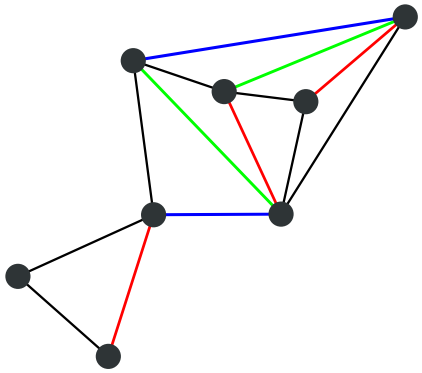


FIG. 2. Example network illustrating how iterated triadic closures are implemented in the model. The initial network (black edges) receives the first generation of triadic closures (red edges). The second generation (green edges) can only close triads involving at least one edge of the first generation (red). The third generation (blue edges), in turn, can only close triads involving at least one edge belonging to the second generation (green).

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 that 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 similar probability [40].

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

$$P(\{\boldsymbol{g}^{(l)}\},\{\boldsymbol{A}^{(l)}\}|\boldsymbol{b},\boldsymbol{p}) = P(\boldsymbol{A}|\boldsymbol{b})\prod_{l=1}^{L}\prod_{u}P(\boldsymbol{g}^{(l)}(u)|\boldsymbol{A}^{(l-1)},\boldsymbol{g}^{(l-1)},p_u^{(l)}). \quad (15)$$

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

$$P(\{\boldsymbol{g}^{(l)}\},\{\boldsymbol{A}^{(l)}\}|\boldsymbol{b}) = P(\boldsymbol{A}|\boldsymbol{b})\prod_{l=1}^{L}P(\boldsymbol{g}^{(l)}|\boldsymbol{A}^{(l-1)},\boldsymbol{g}^{(l-1)}), \quad (16)$$

with the individual terms in the product being entirely analogous to Eq. (10),

$$P(\boldsymbol{g}^{(l)}|\boldsymbol{A}^{(l-1)},\boldsymbol{g}^{(l-1)}) = \prod_{u}\left[\binom{\sum_{i<j}m_{ij}^{(l)}(u)}{\sum_{i<j}g_{ij}^{(l)}(u)}^{-1}\frac{1}{1+\sum_{i<j}m_{ij}^{(l)}(u)}\right]. \quad (17)$$

Finally, the posterior distribution for the reconstruction becomes

$$P(\{\boldsymbol{g}^{(l)}\},\{\boldsymbol{A}^{(l)}\},\boldsymbol{b}|\boldsymbol{G}) = \frac{P(\boldsymbol{G},\{\boldsymbol{g}^{(l)}\},\{\boldsymbol{A}^{(l)}\}|\boldsymbol{b})P(\boldsymbol{b})}{P(\boldsymbol{G})}. \quad (18)$$

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 [41].

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 the Markov chain Monte Carlo (MCMC) method, implemented as follows. We begin with an arbitrary choice of $\{\boldsymbol{g}^{(l)}\}$, $\{\boldsymbol{A}^{(l)}\}$, and $\boldsymbol{b}$ that is compatible with our observed graph $\boldsymbol{G}$. We then consider modifications of these quantities and accept or reject them according to the Metropolis-Hastings criterion [42,43]. More specifically, we consider moves of the kind $P(\{\boldsymbol{g}'^{(l)}\},\{\boldsymbol{A}^{(l)\prime}\}|\{\boldsymbol{g}^{(l)}\},\{\boldsymbol{A}^{(l)}\})$ and accept them according to the probability

$$\min\left(1,\frac{P(\{\boldsymbol{g}'^{(l)}\},\{\boldsymbol{A}^{(l)\prime}\},\boldsymbol{b}|\boldsymbol{G})}{P(\{\boldsymbol{g}^{(l)}\},\{\boldsymbol{A}^{(l)}\},\boldsymbol{b}|\boldsymbol{G})}\times\frac{P(\{\boldsymbol{g}^{(l)}\},\{\boldsymbol{A}^{(l)}\}|\{\boldsymbol{g}'^{(l)}\},\{\boldsymbol{A}^{(l)\prime}\})}{P(\{\boldsymbol{g}'^{(l)}\},\{\boldsymbol{A}^{(l)\prime}\}|\{\boldsymbol{g}^{(l)}\},\{\boldsymbol{A}^{(l)}\})}\right) \quad (19)$$

which, as we mentioned before, does not require the computation of the intractable marginal probability $P(\boldsymbol{G})$. We also consider moves that change the community structure, according to a proposal $P(\boldsymbol{b}'|\boldsymbol{b})$ and accept with probability

$$\min\left(1,\frac{P(\boldsymbol{A}|\boldsymbol{b}')P(\boldsymbol{b}')P(\boldsymbol{b}|\boldsymbol{b}')}{P(\boldsymbol{A}|\boldsymbol{b})P(\boldsymbol{b})P(\boldsymbol{b}'|\boldsymbol{b})}\right). \quad (20)$$

For the latter, we use the merge-split moves described in Ref. [44]. 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 π_{ij} of an edge (i,j) being seminal, which is given by

$$\pi_{ij} = \sum_{\{\boldsymbol{g}^{(l)}\},\{\boldsymbol{A}^{(l)}\},\boldsymbol{b}} A_{ij}P(\{\boldsymbol{g}^{(l)}\},\{\boldsymbol{A}^{(l)}\},\boldsymbol{b}|\boldsymbol{G}). \quad (21)$$

Conversely, the reciprocal quantity,

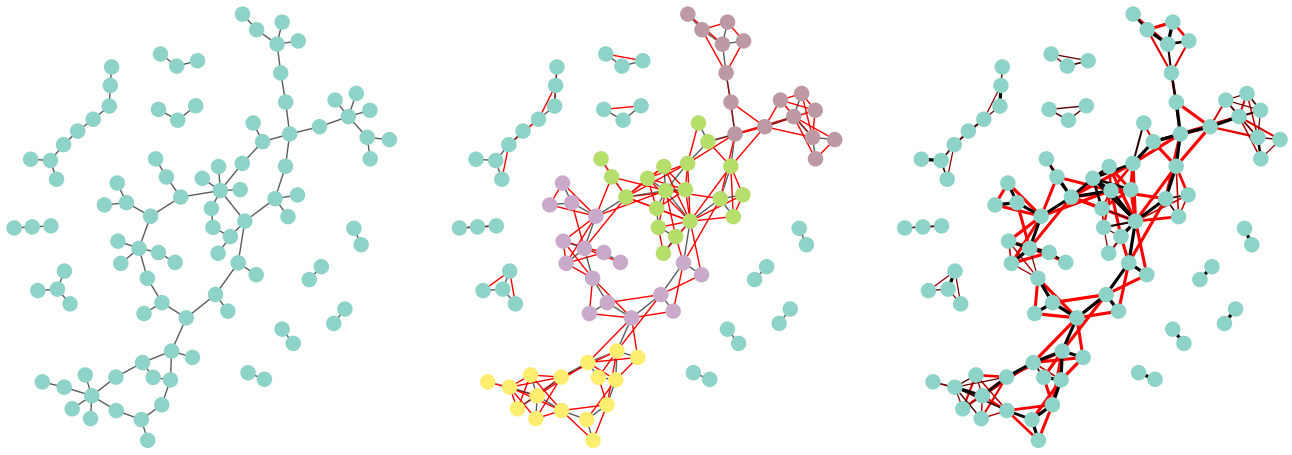
$$1-\pi_{ij}, \quad (22)$$

corresponds to the probability that edge (i, j) is due to triadic closure, occurring in any generation or ego graph. Therefore, the quantity $\boldsymbol{\pi}$ gives us a concise summary of posterior decomposition of a network, and we 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 do 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. 3(a). 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. 3(b) 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. However, one should not 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. 3(c). 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 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 $\boldsymbol{b}^{(1)}$ found by inferring the SBM on a given network, as well as another partition $\boldsymbol{b}^{(2)}$ and ego graphs $\{\boldsymbol{g}^{(l)}\}$ found with the SBM/TC



(a) Random seminal edges

(b) Triadic closure edges and spurious communities found with SBM ($\Sigma_{\mathrm{SBM}} = 801.7$ nats)

(c) Inference of the SBM/TC model ($\Sigma_{\mathrm{SBM/TC}} = 590.6$ nats)

FIG. 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. We also show the partition found by fitting the SBM to the resulting network and the description length obtained. (c) 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 corresponds to the marginal probabilities π_{ij} and $1 - \pi_{ij}$ for the seminal and closure edges, respectively). We also show the description length of the SBM/TC fit.

model. We can decide which one provides a better description of a network via the posterior odds ratio,

$$\Lambda = \frac{P(\boldsymbol{b}^{(2)}, \{\boldsymbol{g}^{(l)}\}, \mathcal{H}_{\mathrm{SBM/TC}}|\boldsymbol{G})}{P(\boldsymbol{b}^{(1)}, \mathcal{H}_{\mathrm{SBM}}|\boldsymbol{G})} \tag{23}$$

$$= \frac{P(\boldsymbol{G}, \{\boldsymbol{g}^{(l)}\}, \{\boldsymbol{A}^{(l)}\}, \boldsymbol{b}^{(2)})}{P(\boldsymbol{G}, \boldsymbol{b}^{(1)})} \times \frac{P(\mathcal{H}_{\mathrm{SBM/TC}})}{P(\mathcal{H}_{\mathrm{SBM}})}, \tag{24}$$

where $P(\mathcal{H}_{\mathrm{SBM/TC}})$ and $P(\mathcal{H}_{\mathrm{SBM}})$ are the prior probabilities for either model. In case these are the same, we have

$$\Lambda = \mathrm{e}^{-(\Sigma_{\mathrm{SBM/TC}} - \Sigma_{\mathrm{SBM}})}, \tag{25}$$

where $\Sigma_{\mathrm{SBM/TC}}$ and Σ_{SBM} are the description lengths of both hypotheses, given by

$$\Sigma_{\mathrm{SBM/TC}} = -\ln P(\boldsymbol{G}, \{\boldsymbol{g}^{(l)}\}, \{\boldsymbol{A}^{(l)}\}, \boldsymbol{b}^{(2)}), \tag{26}$$

$$\Sigma_{\mathrm{SBM}} = -\ln P(\boldsymbol{G}, \boldsymbol{b}^{(1)}). \tag{27}$$

The description length [45] measures the amount of information necessary to encode both the data and the model parameters and hence accounts for both the quality of the fit and the model complexity. Thus, the model that is most likely *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 there are zero triadic closure edges. In particular, if we use the more general parametrization described in Appendix B, in the situation with zero triadic edges (i.e., all $\{\boldsymbol{g}^{(l)}\}$ are empty graphs $\boldsymbol{g}_{\mathrm{empty}}$ and $\boldsymbol{A} = \boldsymbol{G}$), we have

$$P(\boldsymbol{G}, \{\boldsymbol{g}^{(l)} = \boldsymbol{g}_{\mathrm{empty}}\}, \boldsymbol{A} = \boldsymbol{G}, \boldsymbol{b}) \geq \frac{P(\boldsymbol{G}, \boldsymbol{b})}{N+1}. \tag{28}$$

Therefore, in general, we must have

$$\max_{\{\boldsymbol{g}^{(l)}\}, \{\boldsymbol{A}^{(l)}\}, \boldsymbol{b}} \ln P(\boldsymbol{G}, \{\boldsymbol{g}^{(l)}\}, \{\boldsymbol{A}^{(l)}\}, \boldsymbol{b}) \geq \max_{\boldsymbol{b}} \ln P(\boldsymbol{G}, \boldsymbol{b}) - \ln(N+1). \tag{29}$$

Since the last logarithm term becomes negligible for large networks, typically the use of the SBM/TC can only reduce the description length of the data. Therefore, in situations where there is no evidence for triadic closure, both models should yield approximately the same description length value.

In Fig. 3, we show the description lengths for both models for the particular example discussed previously, where we can see that the SBM/TC provides a substantially better compression of the data, therefore yielding a more parsimonious and hence more probable account of how the data were generated—which happens to also 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

$$e_{rs} = 2E\left[\frac{c}{B}\delta_{rs} + \frac{1-c}{B(B-1)}(1-\delta_{rs})\right], \tag{30}$$

where $c \in [0,1]$ determines the affinity between the (dis) assortative groups. For this model, we know that there are critical values

$$c^{*}_{\pm} = \frac{1}{B} \pm \frac{B-1}{B\sqrt{\langle k \rangle}}, \tag{31}$$

such that if $c \in [c^{*}_{-}, c^{*}_{+}]$, then no algorithm can infer a partition that is correlated to the true one from a single network realization, as it becomes infinitely large, $N \to \infty$ [46]. 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 models. A result of this analysis is shown in Fig. 4, where we compute the maximum overlap [47] $q \in [0,1]$ between the inferred $\hat{\boldsymbol{b}}$ and true partition $\boldsymbol{b}$, defined as

$$q = \max_{\mu} \frac{1}{N} \sum_{i} \delta_{\mu(\hat{b}_i), b_i}, \tag{32}$$

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

$$S = -\sum_{r} \frac{n_r}{N} \ln \frac{n_r}{N}. \tag{33}$$

As can be seen in Fig. 4(a), 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. 4(b), we see that the number of groups uncovered can be orders of magnitude larger than the original partition, especially 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

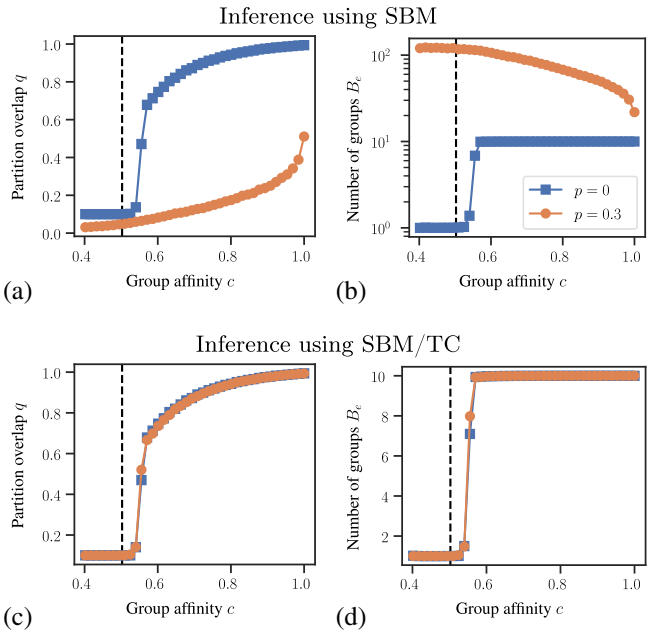
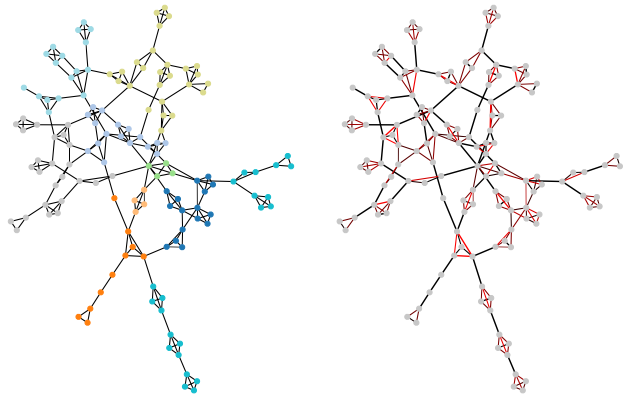


FIG. 4. Recovery of community structure for artificial networks generated from the PP model with added triadic closure, as described in the text, for networks with $N = 10^4$ nodes, average degree $\langle k \rangle = 5$, $B = 10$ planted groups, and uniform triadic closure probability $p_u = p$, shown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



(a) SBM, $\Sigma_{\mathrm{SBM}} = 1145.6$ nats (b) SBM/TC, $\Sigma_{\mathrm{SBM/TC}} = 935.1$ nats

FIG. 5. Network of cooperation between students [48]. (a) Fit of the SBM, yielding $B = 9$ communities. (b) Fit of the SBM/TC, uncovering a single community, and triadic closure edges shown in red. The thickness of the edges corresponds to the marginal probabilities π_{ij} and $1 - \pi_{ij}$ for the seminal and closure edges, respectively.

when we use the SBM/TC instead, as shown Fig. 4(c). 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. 4(c), 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 that 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 [48]. In Fig. 5(a), we show the network and a fit of the DC-SBM, which finds nine 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 [30]—we are tempted to posit that these divisions uncover 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. 5(b), which uncovers, instead, only a single group, and an abundance of triadic closure edges. This is not unlike the artificial example considered in Fig. 3, and it 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 now turn to another social network, but this time, one of friendships between high school students [49]. We show the results of our analysis in Fig. 6. Using the SBM, we find $B = 26$ groups, shown in Fig. 6(a), 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 seen in Fig. 6(b). Unlike 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. 6(b), 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 preferences 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

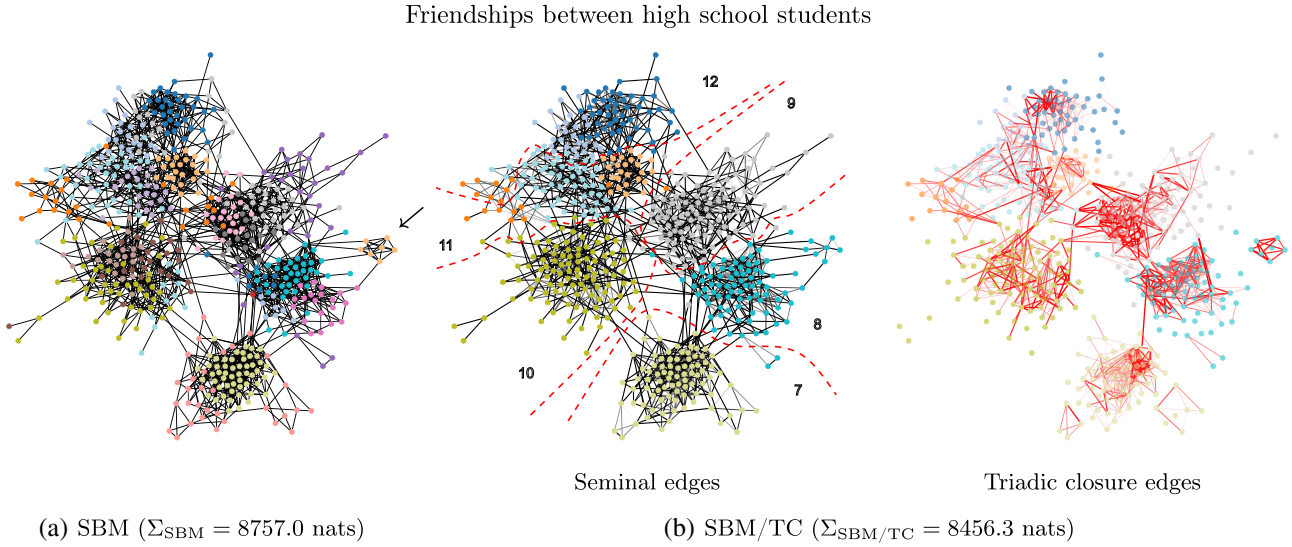


FIG. 6. Top panel: network of friendships between high school students—adolescent health (comm26) [49]. (a) Fit of the SBM, yielding $B = 26$ communities. (b) Fit of the SBM/TC, uncovering $B = 9$ communities, with seminal (black) and triadic closure (red) edges shown separately in the left and right panels. Bottom panel: network of collaborations between network scientists [50]. (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 corresponds to the marginal probabilities π_{ij} and $1 - \pi_{ij}$ for the seminal and closure edges, respectively.

communities entirely to triadic closure. A good example is the group highlighted with an arrow in Fig. 6(a), formed by students in the 8th grade. According to the SBM/TC, this group has arisen because of the formation of triangles between an initially poorly connected subset of students, formed by all friends 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 its complexity. In view of this, the SBM/TC hypothesis is objectively more parsimonious, and in the absence of further data, it should be considered more plausible than the pure SBM.

Now, we consider an additional example, this time of collaborations between researchers in network science [50], 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 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 from two main groups, where the "core" (blue nodes) is composed of perceived initiators of the collaborations with the "periphery" (yellow nodes), which end up being connected in the final network simply by virtue of the all-to-all nature of multiway collaborations, captured here by triadic closure edges. 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, which 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 [51], shown in Fig. 7. For this network, we observe an interesting result, namely, that 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 the 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 as being a "side effect" of the homophily between nodes of the same group, instead of an excess that needs additional explanation. We revisit this particular case in the following, from a different angle.

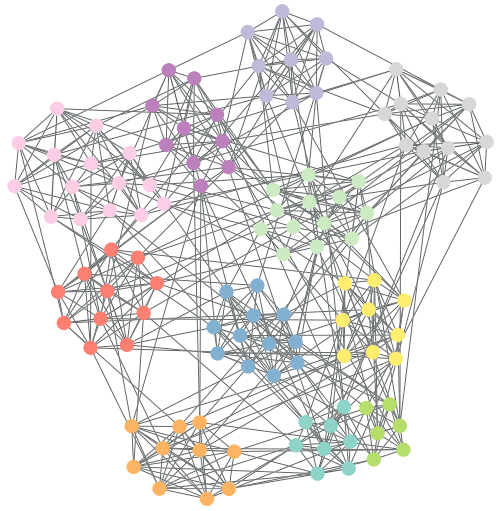


FIG. 7. Network of games between American college football teams (NCAA college football 2000) [51]. The node colors show the fit of the SBM and SBM/TC, both yielding the same $B = 11$ communities. The SBM yields a description length of $\Sigma_{\mathrm{SBM}} = 1761.1$ nats, and the SBM/TC, $\Sigma_{\mathrm{SBM/TC}} = 1767.6$ nats.

One natural criticism of the SBM as a useful hypothesis for real networks, however stylized 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 in 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 $\boldsymbol{G}$ is via its clustering coefficient $C(\boldsymbol{G}) \in [0, 1]$, which determines the fraction of triads in the network that are closed in a triangle, and is given by

$$C(\boldsymbol{G}) = \frac{\sum_{ijk} G_{ij} G_{jk} G_{ki}}{\sum_i k_i (k_i - 1)}, \tag{34}$$

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

$$P(C|\boldsymbol{G}) = \sum_{\boldsymbol{G}'} \delta(C - C(\boldsymbol{G}')) \sum_{\boldsymbol{\theta}} P(\boldsymbol{G}'|\boldsymbol{\theta}) P(\boldsymbol{\theta}|\boldsymbol{G}). \tag{35}$$

This involves sampling parameters $\boldsymbol{\theta}$ from the posterior $P(\boldsymbol{\theta}|\boldsymbol{G})$, generating new networks $\boldsymbol{G}'$ from the model $P(\boldsymbol{G}'|\boldsymbol{\theta})$, and obtaining the resulting population of $C(\boldsymbol{G}')$ values, which can then be compared to the observed value $C(\boldsymbol{G})$. In this way, we can determine if the model used is capable of capturing this aspect of the data. In Fig. 8, we show the results of this comparison for the SBM and SBM/TC (in Appendix D, we give more details about how $\boldsymbol{\theta}$ should be chosen in each case) using four data sets. 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

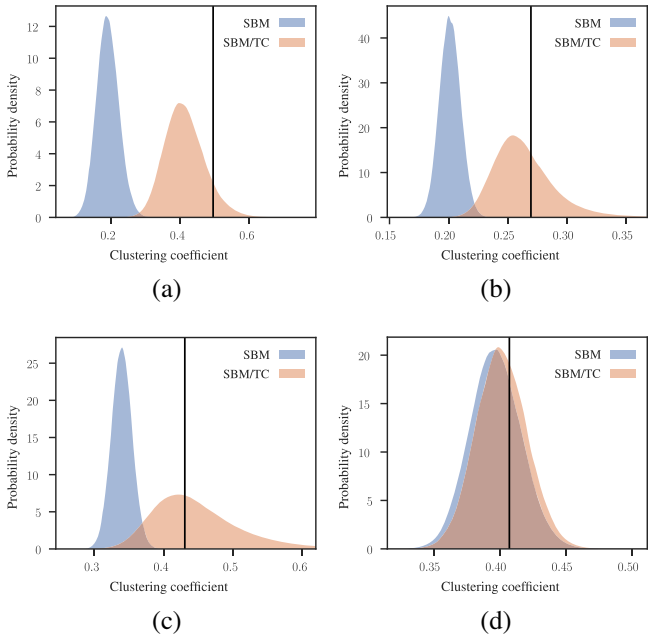


FIG. 8. Posterior predictive distributions of the clustering coefficient, as described in the text, for the SBM and SBM/TC as indicated in the legend, for different data sets. The vertical line shows the empirical value $C(\boldsymbol{G})$. (a) Cooperation between students (b) Adolescent health (comm26) (c) Scientific collaborations in Network Science (d) NCAA college football 2000.

in Fig. 8(c), 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. 8 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 Refs. [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 modeled 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,

$$z = \frac{C(\boldsymbol{G}) - \langle C \rangle}{\sigma_C}, \tag{36}$$

where $\langle C \rangle$ and σ_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.

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,

$$C_S(\boldsymbol{G}) = \sum_{\{\boldsymbol{A}^{(l)}\},\{\boldsymbol{g}^{(l)}\},\boldsymbol{b}} C(\boldsymbol{A}) P(\{\boldsymbol{A}^{(l)}\},\{\boldsymbol{g}^{(l)}\},\boldsymbol{b}|\boldsymbol{G}). \tag{37}$$

We can then compare this value with the coefficient for the observed network $C(\boldsymbol{G})$, 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 $C_S(\boldsymbol{G})$ 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 the 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 closure 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 varies considerably, in general, and needs to be assessed for each individual network.

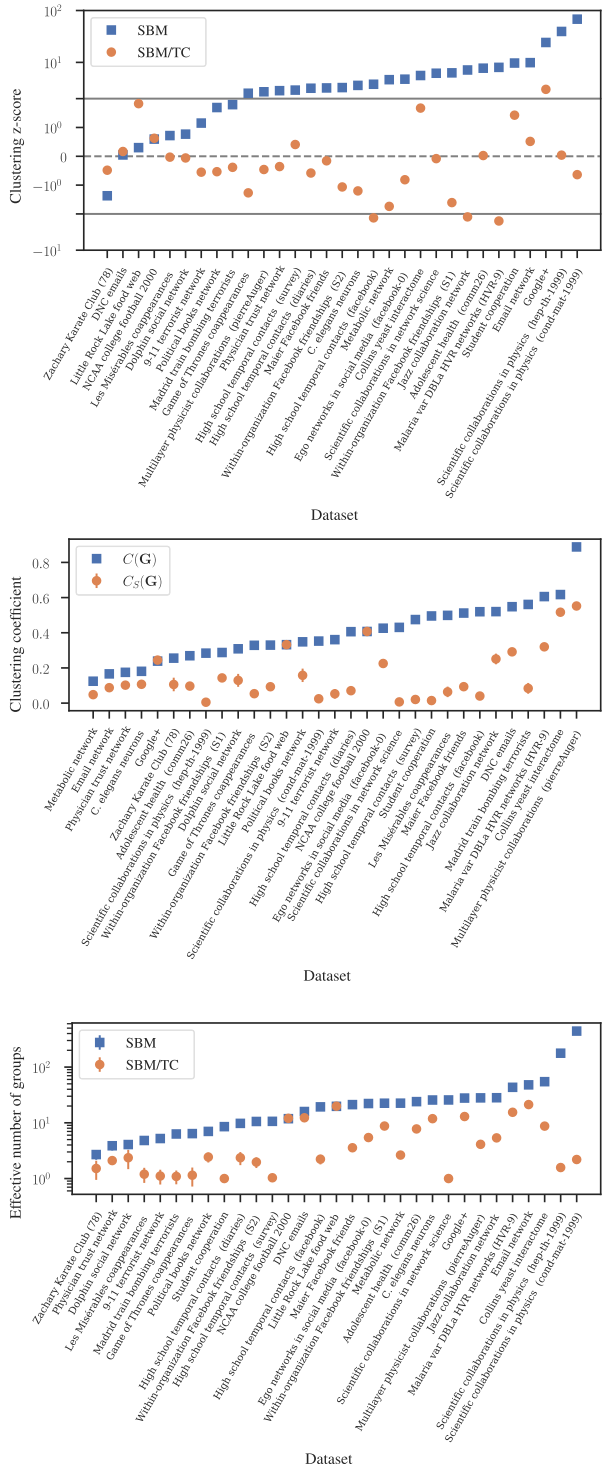


FIG. 9. Top panel: values of the z score for the posterior predictive distributions of the clustering coefficient, as described in the text, for the SBM and SBM/TC, as indicated in the legend, for different data sets. The solid horizontal lines mark the values -2 and 2. Middle panel: values of the clustering coefficient [Eq. (34)] computed for the original network, $C(\boldsymbol{G})$, and for the inferred seminal network, $C_S(\boldsymbol{G})$, averaged over the posterior distribution according to Eq. (37), as shown in the legend. Bottom panel: values of the effective number of inferred groups, as given by Eq. (33), for the SBM and SBM/TC, as indicated in the legend.

V. EDGE PREDICTION AND NETWORK RECONSTRUCTION

As with every kind of empirical assessment, network data are 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 [52,53]. The SBM has been successfully used as such a model [34,53] 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 [54]. As we have been discussing in this work, these two properties—group affinity and triadic closure—point to related but distinct processes of edge formation and approaches to 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 in 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 $\boldsymbol{G}$ according to the posterior distribution

$$P(\boldsymbol{G}|\boldsymbol{n},\boldsymbol{x}) = \frac{P(\boldsymbol{x}|\boldsymbol{G},\boldsymbol{n})P(\boldsymbol{G})}{P(\boldsymbol{x}|\boldsymbol{n})}, \tag{38}$$

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

$$P(\boldsymbol{x}|\boldsymbol{G},\boldsymbol{n},p,q) = \prod_{i<j}\binom{n_{ij}}{x_{ij}}[(1-p)^{x_{ij}}p^{n_{ij}-x_{ij}}]^{G_{ij}} \times [q^{x_{ij}}(1-q)^{n_{ij}-x_{ij}}]^{1-G_{ij}}. \tag{39}$$

Assuming that both p and q are unknown *a priori*, i.e., $P(p) = P(q) = 1$, amounts to the marginal probability [34]

$$P(\boldsymbol{x}|\boldsymbol{G},\boldsymbol{n}) = \int P(\boldsymbol{x}|\boldsymbol{G},\boldsymbol{n},p,q)P(p)P(q)\mathrm{d}p\mathrm{d}q \tag{40}$$

$$= \left[\prod_{i<j}\binom{n_{ij}}{x_{ij}}\right]\binom{\mathcal{E}}{\mathcal{T}}^{-1}\frac{1}{\mathcal{E}+1} \times \binom{\mathcal{M}-\mathcal{E}}{\mathcal{X}-\mathcal{T}}^{-1}\frac{1}{\mathcal{M}-\mathcal{E}+1}. \tag{41}$$

where we have

$$\mathcal{M} = \sum_{i<j} n_{ij}, \qquad \mathcal{X} = \sum_{i<j} x_{ij}, \tag{42}$$

$$\mathcal{E} = \sum_{i<j} n_{ij} G_{ij}, \qquad \mathcal{T} = \sum_{i<j} x_{ij} G_{ij}. \tag{43}$$

The network model comes into play via the prior $P(\boldsymbol{G})$. For the SBM/TC model, this is

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

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

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

so that the desired posterior distribution can be obtained by marginalization,

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

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

$$p_{ij} = \sum_{\boldsymbol{G}} G_{ij} P(\boldsymbol{G}|\boldsymbol{n},\boldsymbol{x}), \tag{47}$$

and we use it to evaluate the quality of the reconstruction. We do so by computing the precision and recall, defined as

$$\text{Precision} = \frac{\sum_{(i,j)\in\boldsymbol{P}_t} p_{ij}}{\sum_{(i,j)\in\boldsymbol{P}_t\cup\boldsymbol{N}_t} p_{ij}}, \tag{48}$$

$$\text{Recall} = \frac{\sum_{(i,j)\in\boldsymbol{P}_t} p_{ij}}{|\boldsymbol{P}_t|}, \tag{49}$$

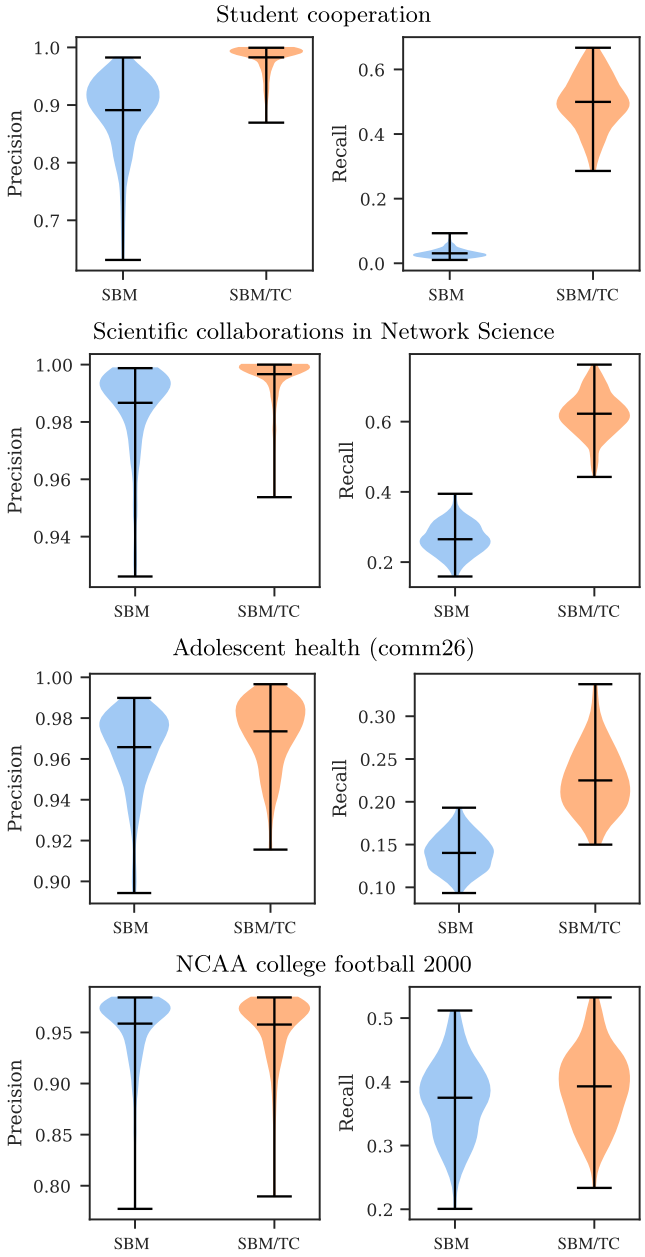


FIG. 10. Distributions of precision and recall values, according to the SBM and SBM/TC models, for four empirical networks, and a fraction $f = 0.05$ of omitted edges and the corresponding number of omitted nonedges. The results were obtained for 200 different realizations of missing edges and nonedges.

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, 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 an 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.* [55] have performed a large-scale analysis of over 200 edge prediction methods on over 500 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 was far from uniform, with the method above based on the SBM providing the single best performance overall [56]. 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 a 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 a 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 determine whether an observed abundance of triangles is a by-product 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, also be extended to other types of local structures such as reciprocity in directed networks [59] 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 [60] (unlike noninferential methods, which tend to overfit [61]), [62] 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.

APPENDIX A: LATENT MULTIGRAPH SBM

The marginal likelihood of Eq. (1) is in fact obtained for a multigraph model [36], where the adjacency entries can take any natural value, $A_{ij} \in \mathbf{N}$. 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] because 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 Refs. [34,37] and consider a *latent* multigraph $\boldsymbol{A}'$, with $\boldsymbol{A}'_{ij} \in \mathbf{N}$, which is then converted into a simple graph $\boldsymbol{A}(\boldsymbol{A}')$ simply by ignoring the edge multiplicities, i.e.,

$$A_{ij} = \begin{cases} 1 & \text{if } A_{ij}' > 0 \\ 0 & \text{otherwise.} \end{cases} \tag{A1}$$

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

$$P(\boldsymbol{A}|\boldsymbol{b}) = \sum_{\boldsymbol{A}'} \mathbf{1}_{\{\boldsymbol{A}(\boldsymbol{A}')=\boldsymbol{A}\}} P(\boldsymbol{A}'|\boldsymbol{b}). \tag{A2}$$

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

$$P(\boldsymbol{g}, \boldsymbol{A}', \boldsymbol{b}|\boldsymbol{G}) = \frac{P(\boldsymbol{G}, \boldsymbol{g}|\boldsymbol{A}(\boldsymbol{A}'))P(\boldsymbol{A}'|\boldsymbol{b})P(\boldsymbol{b})}{P(\boldsymbol{G})}, \tag{A3}$$

where the simple graph $\boldsymbol{A}(\boldsymbol{A}')$ is used for the triadic closure likelihood $P(\boldsymbol{G}, \boldsymbol{g}|\boldsymbol{A})$. In this way, the inference procedure is the same as the one described in the main text, with the only modification being that we need to infer the integer values of $\boldsymbol{A}'$ 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., $\sum_{uij} g'_{ij}(u) = 0$, is given by

$$P(\boldsymbol{g}'|\boldsymbol{A}) = \prod_u \left[1 + \sum\nolimits_{i<j} m_{ij}(u)\right]^{-1} \tag{B1}$$

$$= O\left(\frac{1}{[\langle k^2\rangle - \langle k\rangle]^N}\right), \tag{B2}$$

which is exponentially suppressed for a large number of nodes N. Even though we are interested in modeling networks that possess some amount of triadic closure, we should be *a priori* more agnostic about the actual amount, so 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

$$P(\boldsymbol{g}|\boldsymbol{A}) = \prod_u \sum_{E_u} P(\boldsymbol{g}(u)|\boldsymbol{A}, E_u) P(E_u|\boldsymbol{A}), \tag{B3}$$

where

$$P(\boldsymbol{g}(u)|\boldsymbol{A}, E_u) = \frac{\mathbf{1}_{\{E_u = \sum_{i<j} g_{ij}(u)\}}}{\binom{\sum_{i<j} m_{ij}(u)}{E_u}} \tag{B4}$$

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

$$P(E_u|\boldsymbol{A}) = \frac{1}{1 + \sum_{i<j} m_{ij}(u)} \tag{B5}$$

is the probability of uniformly sampling the number of edges in $\boldsymbol{g}(u)$ 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, $P(E_u|\boldsymbol{A}, t_u = 0) = \delta_{0,E_u}$; otherwise, it has a nonzero number of edges, sampled uniformly as

$$P(E_u|\boldsymbol{A}, t_u = 1) = \frac{1 - \delta_{E_u,0}}{\sum_{i<j} m_{ij}(u)}. \tag{B6}$$

The modified marginal distribution then becomes

$$P(\boldsymbol{g}|\boldsymbol{A}) = \sum_{\boldsymbol{t}} \left[\prod_u \sum_{E_u} P(\boldsymbol{g}(u)|\boldsymbol{A}, E_u) P(E_u|\boldsymbol{A}, \boldsymbol{t})\right] P(\boldsymbol{t}|\boldsymbol{A}) \tag{B7}$$

$$= \left\{\prod_u \left[\binom{\sum_{i<j} m_{ij}(u)}{\sum_{i<j} g_{ij}(u)} \sum\nolimits_{i<j} m_{ij}(u)\right]^{\delta_{1,\Theta[\sum_{i<j} g_{ij}(u)]}}\right\} \binom{\sum_u \Theta[\sum_{i<j} m_{ij}(u)]}{\sum_u \Theta[\sum_{i<j} g_{ij}(u)]}^{-1} \times \frac{1}{1 + \sum_u \Theta[\sum_{i<j} m_{ij}(u)]}, \tag{B8}$$

where $\Theta[x] = \{1 \text{ if } x > 0, \text{else } 0\}$ is the Heaviside step function, and we have used the prior

$$P(\boldsymbol{t}|\boldsymbol{A}) = \sum_{N_t} P(\boldsymbol{t}|\boldsymbol{A}, N_t) P(N_t|\boldsymbol{A}), \tag{B9}$$

with

$$P(\boldsymbol{t}|\boldsymbol{A}, N_t) = \frac{\mathbf{1}_{\{\sum_u t_u = N_t\}} \prod_u \Theta[\sum_{i<j} m_{ij}(u)]^{t_u}}{\binom{\sum_u \Theta[\sum_{i<j} m_{ij}(u)]}{N_t}} \tag{B10}$$

and

$$P(N_t|\boldsymbol{A}) = \frac{1}{1 + \sum_u \Theta[\sum_{i<j} m_{ij}(u)]}. \tag{B11}$$

The above amounts to sampling in sequence the number of nodes N_t and the partition $\boldsymbol{t}$, both uniformly at random from the allowed range. Although these equations take longer to write, they are not much more difficult to use. As a result of this parametrization, if we consider again the particular graph with no triadic closures, i.e., $\sum_{uij} g'_{ij}(u) = 0$, it is generated with probability

$$P(\boldsymbol{g}'|\boldsymbol{A}) = \frac{1}{1+\sum_u \Theta[\sum_{i<j} m_{ij}(u)]} = O\left(\frac{1}{N}\right), \quad \text{(B12)}$$

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 a 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 observe.

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

$$P(\boldsymbol{G},\boldsymbol{g}',\boldsymbol{A}=\boldsymbol{G}|\boldsymbol{b}) = \frac{P(\boldsymbol{G}|\boldsymbol{b})}{1+\sum_u \Theta[\sum_{i<j} m_{ij}(u)]} \quad \text{(B13)}$$

$$\geq \frac{P(\boldsymbol{G}|\boldsymbol{b})}{N+1}, \quad \text{(B14)}$$

with $\boldsymbol{g}'$ being empty ego graphs, and the last equality is 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),

$$P(\boldsymbol{g}^{(l)}|\boldsymbol{A}^{(l-1)},\boldsymbol{g}^{(l-1)}) = \left\{\prod_u \left[\begin{pmatrix}\sum_{i<j} m_{ij}^{(l)}(u)\\ \sum_{i<j} g_{ij}^{(l)}(u)\end{pmatrix} \sum_{i<j} m_{ij}^{(l)}(u)\right]^{\delta_{0,\sum_{i<j} g_{ij}^{(l)}(u)}-1}\right\} \begin{pmatrix}\sum_u \mathbf{1}_{\{\sum_{i<j} m_{ij}^{(l)}(u)>0\}}\\ \sum_u \mathbf{1}_{\{\sum_{i<j} g_{ij}^{(l)}(u)>0\}}\end{pmatrix}^{-1} \frac{1}{1+\sum_u \mathbf{1}_{\{\sum_{i<j} m_{ij}^{(l)}(u)>0\}}}. \quad \text{(B15)}$$

APPENDIX C: MCMC MOVES

The MCMC algorithm described in the main text is implemented with the following moves. The first is to attempt to move an edge (i,j) in ego graph $\boldsymbol{g}^{(l)}(u)$ at its current generation $l \in [0,L]$ to another ego graph $\boldsymbol{g}^{(l')}(v)$ for $v \neq u$ at generation $l' \neq l$. We do so by first selecting an edge (i,j) in $\boldsymbol{G}$ 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

$$n_{ij}^{(l)} = \sum_u A_{ui}^{(l-1)}(1-A_{uj}^{(l-1)}), \quad \text{(C1)}$$

which is independent on the value of $g_{ij}^{(l')}(u)$ for any l'. We then sample another generation $l' \neq l$ and proceed in the same way to sample a relevant ego node v. In either case, if $l = 0$ is selected, then the choice of an ego graph is not made since we are simply selecting an entry (i,j) in $\boldsymbol{A}$ with probability 1. The final probability of selecting the move $(i,j,u,l) \to (i,j,v,l')$, assuming $l > 0$ and $l' > 0$, is given by

$$P(i,j,u,v,l,l'|\{\boldsymbol{g}^{(l)}\},\{\boldsymbol{A}^{(l)}\}) = \frac{1}{E_G n_{ij}^{(l)} n_{ij}^{(l')} L(L+1)}, \quad \text{(C2)}$$

where E_G is the number of edges in $\boldsymbol{G}$. Given this selection, we then make the change ${g'_{ij}}^{(l)}(u) = g_{ij}^{(l)} - 1$ and ${g'_{ij}}^{(l')}(u) = g_{ij}^{(l')} + 1$, and accept it with probability

$$\min\left(1, \frac{P(\{\boldsymbol{g}'^{(l)}\},\{\boldsymbol{A}^{(l)\prime}\},\boldsymbol{b}|\boldsymbol{G})P(i,j,u,v,l,l'|\{\boldsymbol{g}^{(l)}\},\{\boldsymbol{A}^{(l)\prime}\})}{P(\{\boldsymbol{g}^{(l)}\},\{\boldsymbol{A}^{(l)}\},\boldsymbol{b}|\boldsymbol{G})P(i,j,v,u,l',l|\{\boldsymbol{g}'^{(l)}\},\{\boldsymbol{A}^{(l)}\})}\right) = \min\left(1, \frac{P(\{\boldsymbol{g}'^{(l)}\},\{\boldsymbol{A}^{(l)\prime}\},\boldsymbol{b}|\boldsymbol{G})}{P(\{\boldsymbol{g}^{(l)}\},\{\boldsymbol{A}^{(l)}\},\boldsymbol{b}|\boldsymbol{G})}\right), \quad \text{(C3)}$$

which is independent of the actual move probabilities since they remain the same after and before the move. Note that invalid moves that result in $\boldsymbol{g}_{ij}^{(l)} < 0$ or $\boldsymbol{A}_{ij}^{(l)} < 0$ are always rejected in this way.

In addition, we also make a second kind of move by selecting again an edge (i,j) in $\boldsymbol{G}$ 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 make the move $g_{ij}'^{(l)} = g_{ij}^{(l)} \pm 1$ with probability $1/2$ and accept again according to

$$\min\left(1, \frac{P(\{\boldsymbol{g}'^{(l)}\}, \{\boldsymbol{A}^{(l)}\}, \boldsymbol{b}|\boldsymbol{G})}{P(\{\boldsymbol{g}^{(l)}\}, \{\boldsymbol{A}^{(l)}\}, \boldsymbol{b}|\boldsymbol{G})}\right). \tag{C4}$$

If $l = 0$ is selected, the move is different because of the multigraph nature of $\boldsymbol{A}$. Instead, we make the proposal $A_{ij} \to A_{ij}'$ according to a geometric distribution with mean $A_{ij} + 1$,

$$P(A_{ij}'|A_{ij}) = \left(\frac{A_{ij}+1}{A_{ij}+2}\right)^{A_{ij}'} \frac{1}{A_{ij}+2}. \tag{C5}$$

In this case, the acceptance probability changes to

$$\min\left(1, \frac{P(\{\boldsymbol{g}^{(l)}\}, \{\boldsymbol{A}^{(l)\prime}\}, \boldsymbol{b}|\boldsymbol{G}) P(A_{ij}|A_{ij}')}{P(\{\boldsymbol{g}^{(l)}\}, \{\boldsymbol{A}^{(l)}\}, \boldsymbol{b}|\boldsymbol{G}) P(A_{ij}'|A_{ij})}\right). \tag{C6}$$

Finally, the last kind of move involves a change in partition $\boldsymbol{b} \to \boldsymbol{b}'$ from the proposal $P(\boldsymbol{b}'|\boldsymbol{b})$, which is accepted with probability

$$\min\left(1, \frac{P(\boldsymbol{A}|\boldsymbol{b}')P(\boldsymbol{b}')P(\boldsymbol{b}|\boldsymbol{b}')}{P(\boldsymbol{A}|\boldsymbol{b})P(\boldsymbol{b})P(\boldsymbol{b}'|\boldsymbol{b})}\right). \tag{C7}$$

For the latter, we use the merge-split moves, combined with single-node moves, described in Ref. [44].

The moves above fulfill detailed balance, and when combined, they also preserve ergodicity since they allow every latent multigraph, decomposition into ego graphs, and node partition to be sampled. Thus, with sufficiently many iterations, the algorithm must eventually produce samples from the desired posterior distribution.

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 $\boldsymbol{G}$, those that have edge (i,j) in them, as well as the number of edges $E_u^{(l)} = \sum_{i<j} g_{ij}^{(l)}(u)$ of every ego graph. Based on this bookkeeping, whenever an entry $g_{ij}^{(l)}(u)$ (or A_{ij} if $l = 0$) is modified, to compute the log-likelihood difference, we only need 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 the 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 $\boldsymbol{G}$ had a chance to be moved by one of the proposals considered, has an overall complexity of $O(N\langle k^2\rangle)$ since each node i has k_i edges that need to be moved at every sweep, each of which requires 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 [44].

Combining the two kinds of moves gives us an overall complexity of $O(N\langle k^2\rangle)$ per sweep, which, for sparse graphs with $\langle k^2\rangle = O(1)$, 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, 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 a reasonable amount of time even for networks with over a hundred thousand to a million edges, although we did not attempt a detailed investigation of networks that are much larger than this.

We have evaluated the quality of the algorithm with the analysis presented in Fig. 4, 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 would 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 [63].

APPENDIX D: PREDICTIVE POSTERIOR DISTRIBUTION

The predictive posterior distribution considered in the main text is

$$P(C|\boldsymbol{G}) = \sum_{\boldsymbol{G}'} \delta(C - C(\boldsymbol{G}')) \sum_{\boldsymbol{\theta}} P(\boldsymbol{G}'|\boldsymbol{\theta}) P(\boldsymbol{\theta}|\boldsymbol{G}), \tag{D1}$$

where $\boldsymbol{\theta}$ are the parameters of model $P(\boldsymbol{G}|\boldsymbol{\theta})$. 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]

$$P(\boldsymbol{A}|\boldsymbol{b}) = P(\boldsymbol{A}|\boldsymbol{k},\boldsymbol{e},\boldsymbol{b}) P(\boldsymbol{k}|\boldsymbol{e},\boldsymbol{b}) P(\boldsymbol{e}|\boldsymbol{b}), \tag{D2}$$

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

$$P(\boldsymbol{A}|\boldsymbol{k},\boldsymbol{e},\boldsymbol{b}) = \frac{\prod_{r<s}e_{rs}!\prod_r e_{rr}!!\prod_i k_i!}{\prod_{i<j}A_{ij}!\prod_i A_{ii}!!\prod_r e_r!}, \tag{D3}$$

the prior for the degrees is

$$P(\boldsymbol{k}|\boldsymbol{e},\boldsymbol{b}) = \prod_r \frac{\prod_k \eta_k^r!}{n_r!q(e_r,n_r)}, \tag{D4}$$

and the prior for the edge counts between groups is

$$P(\boldsymbol{e}|\boldsymbol{b}) = \left(\!\!\binom{\frac{B(B+1)}{2}}{E}\!\!\right)^{-1}. \tag{D5}$$

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

$$P(p_u|\boldsymbol{g}(u),\boldsymbol{A}) = \frac{p_u^{\sum_{i<j}g_{ij}(u)m_{ij}(u)}(1-p_u)^{\sum_{i<j}(1-g_{ij}(u))m_{ij}(u)}}{\mathcal{B}(\sum_{i<j}g_{ij}(u)m_{ij}(u),\sum_{i<j}(1-g_{ij}(u))m_{ij}(u))}, \tag{D6}$$

where $\mathcal{B}(x,y)$ is the Beta function. Based on this parametrization, our predictive posterior distribution is obtained by setting $\theta=(\{\boldsymbol{p}^{(l)}\},\boldsymbol{k},\boldsymbol{e},\boldsymbol{b})$, amounting to

$$P(C|\boldsymbol{G}) = \sum_{\substack{\{\boldsymbol{g}^{(l)}\}\\ \{\boldsymbol{g}'^{(l)}\}\\ \boldsymbol{A},\boldsymbol{A}'\\ \boldsymbol{k},\boldsymbol{e},\boldsymbol{b}}} \int d\{\boldsymbol{p}^{(l)}\}\delta\{C - C[\boldsymbol{G}(\boldsymbol{A},\{\boldsymbol{g}^{(l)}\})]\}\left[\prod_{l,u}P(\boldsymbol{g}^{(l)}(u)|p_u^{(l)},\boldsymbol{A})\right]P(\boldsymbol{A}|\boldsymbol{k},\boldsymbol{e},\boldsymbol{b}) \times \left[\prod_{l,u}P(p_u^{(l)}|\boldsymbol{g}'^{(l)}(u),\boldsymbol{A}')\right]P(\{\boldsymbol{g}'^{(l)}\},\boldsymbol{A}',\boldsymbol{k},\boldsymbol{e},\boldsymbol{b}|\boldsymbol{G}). \tag{D7}$$

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

APPENDIX E: NETWORK DATA SETS

Below are descriptions of the network data sets used in this work. The codenames in parentheses correspond to the respective entries in the Netzschleuder repository [64] where the networks can be downloaded. Some of the descriptions were obtained from the Colorado Index of Complex Networks [65].

Adolescent health (`add_health`) [49]: 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. 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`) [66]: Cxollaboration 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 was considered. The particular networks named `cond-mat-1999`, `hep-th-1999` have been used, with the number of nodes and edges, (N,E), given by (13861, 44619), (5835, 13815), respectively.

Metabolic network (`celegans_metabolic`) [67]: 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`) [68,69]: 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`) [70]: 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`) [71]: 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`) [72]: An undirected social network of frequent associations observed among 62 dolphins (Tursiops) in a community living in Doubtful Sound, New Zealand, from 1994–2001. This network has $N = 62$ nodes and $E = 159$ edges.

Ego networks in social media (`ego_social`) [73]: Ego networks associated with a set of accounts of three social media platforms (Facebook, Google+, and Twitter). Data sets 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`) [74]: 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 give 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`) [75]: Six networks of friendships among users on Facebook who indicated employment at one of the target corporations. Companies range in size from small to large. Only edges between employees at the same company are included in a given snapshot. Node metadata give listed job type on the user's page. The particular networks named `S1`, `S2` have been used, with the number of nodes and edges, (N, E), given by (320, 2369), (165, 726), respectively.

Little Rock Lake food web (`foodweb_little_rock`) [76]: 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`) [51]: 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`) [77]: 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`) [78]: 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`) [79]: 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`) [80]: Network of friendships among members of a university karate club. Includes metadata for faction membership after a social partition. Note that there are two versions of this network, one with 77 edges and one with 78, because of 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`) [81]: The network of scene coappearances of characters in Victor Hugo's novel "Les Miserables." 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`) [82]: Networks of recombinant antigen genes from the human malaria parasite *P. falciparum*. Each of the nine networks shares the same set of vertices but has different edges, corresponding to the nine 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 include 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`) [50]: 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`) [83]: 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 represents 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 (`physics_collab`) [84]: 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 (`polbooks`) [85]: 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 (`sp_high_school`) [86]: 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 the number of nodes and edges, (N, E), given by (120, 502), (128, 658), (156, 1437), respectively.

Student cooperation (`student_cooperation`) [48]: 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, and 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 (`terrorists_911`) [87]: Network of individuals and their known social associations, centered around the hijackers that carried out the terrorist attacks on September 11, 2001. Associations were extracted after the fact from public data. The 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 (`train_terrorists`) [88]: 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, coparticipants in wars, and coparticipants in previous terrorist attacks. This network has $N = 64$ nodes and $E = 243$ edges.

Email network (`uni_email`) [89]: 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 $E = 10\,903$ edges.

[1] J. M. McPherson and L. Smith-Lovin, *Homophily in Voluntary Organizations: Status Distance and the Composition of Face-to-Face Groups*, Am. Soc. Rev. **52**, 370 (1987).

[2] W. Shrum, N. H. Cheek, and S. M. Hunter, *Friendship in School: Gender and Racial Homophily*, Soc. Educ. **61**, 227 (1988).

[3] J. Moody, *Race, School Integration, and Friendship Segregation in America*, Am. J. Soc. **107**, 679 (2001).

[4] M. McPherson, L. Smith-Lovin, and J. M. Cook, *Birds of a Feather: Homophily in Social Networks*, Annu. Rev. Sociol. **27**, 415 (2001).

[5] S. Fortunato, *Community Detection in Graphs*, Phys. Rep. **486**, 75 (2010).

[6] A. Rapoport, *Spread of Information through a Population with Socio-Structural Bias: I. Assumption of Transitivity*, Bull. Math. Biophys. **15**, 523 (1953).

[7] P. W. Holland and S. Leinhardt, *Transitivity in Structural Models of Small Groups:*, Comparative Group Studies **2**, 107 (1971).

[8] P. W. Holland and S. Leinhardt, *Local Structure in Social Networks*, in *Sociological Methodology*, edited by D. Heise (Jossey-Bass, San Francisco, 1975).

[9] G. Kossinets and D. J. Watts, *Origins of Homophily in an Evolving Social Network*, Am. J. Soc. **115**, 405 (2009).

[10] M. S. Granovetter, *The Strength of Weak Ties*, Am. J. Soc. **78**, 1360 (1973).

[11] D. V. Foster, J. G. Foster, P. Grassberger, and M. Paczuski, *Clustering Drives Assortativity and Community Structure in Ensembles of Networks*, Phys. Rev. E **84**, 066117 (2011).

[12] D. Foster, J. Foster, M. Paczuski, and P. Grassberger, *Communities, Clustering Phase Transitions, and Hysteresis: Pitfalls in Constructing Network Ensembles*, Phys. Rev. E **81**, 046115 (2010).

[13] F. A. Lopez and A. C. C. Coolen, *Transitions in Loopy Random Graphs with Fixed Degrees and Arbitrary Degree Distributions*, arXiv:2008.11002.

[14] G. Bianconi, R. K. Darst, J. Iacovacci, and S. Fortunato, *Triadic Closure as a Basic Generating Mechanism of Communities in Complex Networks*, Phys. Rev. E **90**, 042806 (2014).

[15] S. Wharrie, L. Azizi, and E. G. Altmann, *Micro-, Meso-, Macroscales: The Effect of Triangles on Communities in Networks*, Phys. Rev. E **100**, 022315 (2019).

[16] A. Asikainen, G. Iñiguez, J. Ureña-Carrión, K. Kaski, and M. Kivelä, *Cumulative Effects of Triadic Closure and Homophily in Social Networks*, Sci. Adv. **6**, eaax7310 (2020).

[17] F. Battiston, G. Cencetti, I. Iacopini, V. Latora, M. Lucas, A. Patania, J.-G. Young, and G. Petri, *Networks beyond Pairwise Interactions: Structure and Dynamics*, Phys. Rep. **874**, 1 (2020).

[18] A. R. Benson, R. Abebe, M. T. Schaub, A. Jadbabaie, and J. Kleinberg, *Simplicial Closure and Higher-Order Link Prediction*, Proc. Natl. Acad. Sci. U.S.A. **115**, E11221 (2018).

[19] G. Palla, I. Derényi, I. Farkas, and T. Vicsek, *Uncovering the Overlapping Community Structure of Complex Networks in Nature and Society*, Nature (London) **435**, 814 (2005).

[20] A. R. Benson, D. F. Gleich, and J. Leskovec, *Higher-Order Organization of Complex Networks*, Science **353**, 163 (2016).

[21] H. Yin, A. R. Benson, J. Leskovec, and D. F. Gleich, *Local Higher-Order Graph Clustering*, in *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Association for Computing Machinery, New York, NY, 2017), pp. 555–564.

[22] A. E. Wegner and S. Olhede, *Atomic Subgraphs and the Statistical Mechanics of Networks*, Phys. Rev. E **103**, 042311 (2021).

[23] J.-G. Young, G. Petri, and T. P. Peixoto, *Hypergraph Reconstruction from Network Data*, Commun. Phys. **4**, 1 (2021).

[24] G. Robins, P. Pattison, Y. Kalish, and D. Lusher, *An Introduction to Exponential Random Graph (p^*) Models for Social Networks*, Soc. Networks **29**, 173 (2007).

[25] D. J. Strauss, *A Model for Clustering*, Biometrika **62**, 467 (1975).

[26] J. Park and M. E. J. Newman, *Statistical Mechanics of Networks*, Phys. Rev. E **70**, 066117 (2004).

[27] J. Park and M. E. J. Newman, *Solution of the Two-Star Model of a Network*, Phys. Rev. E **70**, 066146 (2004).

[28] R. Fischer, J. C. Leitão, T. P. Peixoto, and E. G. Altmann, *Sampling Motif-Constrained Ensembles of Networks*, Phys. Rev. Lett. **115**, 188701 (2015).

[29] P. W. Holland, K. Blackmond Laskey, and S. Leinhardt, *Stochastic Blockmodels: First Steps*, Soc. Networks **5**, 109 (1983).

[30] T. P. Peixoto, *Bayesian Stochastic Blockmodeling*, in *Advances in Network Clustering and Blockmodeling* (John Wiley & Sons, New York, 2019), pp. 289–332.

[31] J. Chang, E. D. Kolaczyk, and Q. Yao, *Estimation of Subgraph Density in Noisy Networks*, arXiv:1803.02488.

[32] C. R. Shalizi and A. C. Thomas, *Homophily and Contagion Are Generically Confounded in Observational Social Network Studies*, Socio. Methods Res. **40**, 211 (2011).

[33] E. McFowland III and C. R. Shalizi, *Estimating Causal Peer Influence in Homophilous Social Networks by Inferring Latent Locations*, J. Am. Stat. Assoc. **0**, 1 (2021).

[34] T. P. Peixoto, *Reconstructing Networks with Unknown and Heterogeneous Errors*, Phys. Rev. X **8**, 041011 (2018).

[35] B. Karrer and M. E. J. Newman, *Stochastic Blockmodels and Community Structure in Networks*, Phys. Rev. E **83**, 016107 (2011).

[36] T. P. Peixoto, *Nonparametric Bayesian Inference of the Microcanonical Stochastic Block Model*, Phys. Rev. E **95**, 012317 (2017).

[37] T. P. Peixoto, *Latent Poisson Models for Networks with Heterogeneous Density*, Phys. Rev. E **102**, 012309 (2020).

[38] The SBM is capable of modeling arbitrary kinds of mixing patterns between groups of nodes, with homophily (or assortative mixing) as a special case. Therefore, our approach is in fact able to disentangle arbitrary mixing patterns from triadic closure, not only homophily. However, homophily is the dominant pattern that causes an abundance of triangles and hence needs to be distinguished from triadic closure.

[39] 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 in what order the edges are added to the network, only their final placement.

[40] In more detail, we can recover the unconstrained model by substituting Eq. (13) with $m_{ij}^{(l)}(u) = A_{ui}^{(l-1)}A_{uj}^{(l-1)}(1 - A_{ij}^{(l-1)})$. 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. Because these multiple edges are removed in the end, this means that the unconstrained model would have a higher probability of forming edges that are possible in the earlier generations since they could also appear in the later ones. However, because 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.

[41] 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 B, generations that 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, $L = \binom{N}{2}$, since for later generations it is impossible to add new edges.

[42] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller, *Equation of State Calculations by Fast Computing Machines*, J. Chem. Phys. **21**, 1087 (1953).

[43] W. K. Hastings, *Monte Carlo Sampling Methods Using Markov Chains and Their Applications*, Biometrika **57**, 97 (1970).

[44] T. P. Peixoto, *Merge-Split Markov Chain Monte Carlo for Community Detection*, Phys. Rev. E **102**, 012305 (2020).

[45] P. D. Grünwald, *The Minimum Description Length Principle* (MIT Press, Cambridge, MA, 2007).

[46] 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).

[47] T. P. Peixoto, *Revealing Consensus and Dissensus between Network Partitions*, Phys. Rev. X **11**, 021003 (2021).

[48] 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, Heidelberg, 2012), pp. 584–595.

[49] J. Moody, *Peer Influence Groups: Identifying Dense Clusters in Large Networks*, Soc. Networks **23**, 261 (2001).
[50] M. E. J. Newman, *Finding Community Structure in Networks Using the Eigenvectors of Matrices*, Phys. Rev. E **74**, 036104 (2006).
[51] M. Girvan and M. E. J. Newman, *Community Structure in Social and Biological Networks*, Proc. Natl. Acad. Sci. U.S.A. **99**, 7821 (2002).
[52] A. Clauset, C. Moore, and M. E. J. Newman, *Hierarchical Structure and the Prediction of Missing Links in Networks*, Nature (London) **453**, 98 (2008).
[53] R. Guimerà and M. Sales-Pardo, *Missing and Spurious Interactions and the Reconstruction of Complex Networks*, Proc. Natl. Acad. Sci. U.S.A. **106**, 22073 (2009).
[54] L. A. Adamic and E. Adar, *Friends and Neighbors on the Web*, Soc. Networks **25**, 211 (2003).
[55] 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. U.S.A. **117**, 23393 (2020).
[56] Ghasemian *et al.* [55] 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 used only the version of the SBM with noninformative priors, which is known to underfit, as opposed to the nested SBM [36,57], which removes this problem. Accounting for both of these issues has been shown before to improve edge prediction systematically [58] and could potentially have pushed the result of the analysis in Ref. [55] even more in favor of the SBM approach.
[57] T. P. Peixoto, *Hierarchical Block Structures and High-Resolution Model Selection in Large Networks*, Phys. Rev. X **4**, 011047 (2014).
[58] T. Vallès-Català, T. P. Peixoto, M. Sales-Pardo, and R. Guimerà, *Consistencies and Inconsistencies between Model Selection and Link Prediction in Networks*, Phys. Rev. E **97**, 062316 (2018).
[59] H. Safdari, M. Contisciani, and C. De Bacco, *A Generative Model for Reciprocity and Community Detection in Networks*, Phys. Rev. Research **3**, 023209 (2021).
[60] R. Guimerà, M. Sales-Pardo, and L. A. N. Amaral, *Modularity from Fluctuations in Random Graphs and Complex Networks*, Phys. Rev. E **70**, 025101(R) (2004).
[61] A. Ghasemian, H. Hosseinmardi, and A. Clauset, *Evaluating Overfit and Underfit in Models of Network Community Structure*, IEEE transactions on knowledge and data engineering 1 (2019).
[62] Overfitting here means that the number of communities found is too large and that the method can even find communities in completely random networks.
[63] T. P. Peixoto, The `graph-tool` Python Library, available at https://graph-tool.skewed.de.
[64] T. P. Peixoto, *The Netzschleuder Network Catalogue and Repository*, accessible at https://networks.skewed.de.
[65] A. Clauset, E. Tucker, and M. Sainz, *The Colorado Index of Complex Networks*, accessible at https://icon.colorado.edu.
[66] M. E. J. Newman, *The Structure of Scientific Collaboration Networks*, Proc. Natl. Acad. Sci. U.S.A. **98**, 404 (2001).
[67] J. Duch and A. Arenas, *Community Detection in Complex Networks Using Extremal Optimization*, Phys. Rev. E **72**, 027104 (2005).
[68] J. G. White, E. Southgate, J. N. Thomson, and S. Brenner, *The Structure of the Nervous System of the Nematode Caenorhabditis elegans*, Phil. Trans. R. Soc. B **314**, 1 (1986).
[69] D. J. Watts and S. H. Strogatz, *Collective Dynamics of Small-World Networks*, Nature (London) **393**, 440 (1998).
[70] 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).
[71] J. Kunegis, *KONECT*, in *Proceedings of the 22nd International Conference on World Wide Web—WWW 13 Companion* (ACM Press, New York, 2013).
[72] 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).
[73] J. McAuley and J. Leskovec, *Discovering Social Circles in Ego Networks*, arXiv:1210.8182.
[74] B. F. Maier and D. Brockmann, *Cover Time for Random Walks on Arbitrary Complex Networks*, Phys. Rev. E **96**, 042307 (2017).
[75] M. Fire, R. Puzis, and Y. Elovici, *Organization Mining Using Online Social Networks*, arXiv:1303.3741.
[76] N. D. Martinez, *Artifacts or Attributes? Effects of Resolution on the Little Rock Lake Food Web*, Ecol. Monogr. **61**, 367 (1991).
[77] A. Beveridge and J. Shan, *Network of Thrones*, Math Horizons **23**, 18 (2016).
[78] 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).
[79] P. M. Gleiser and L. Danon, *Community Structure in Jazz*, Adv. Complex Syst. **06**, 565 (2003).
[80] W. W. Zachary, *An Information Flow Model for Conflict and Fission in Small Groups*, Journal of anthropological research **33**, 452 (1977).
[81] D. E. Knuth, *The Stanford GraphBase: A Platform for Combinatorial Computing* (ACM Press, New York, 1993).
[82] 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).
[83] J. Coleman, E. Katz, and H. Menzel, *The Diffusion of an Innovation among Physicians*, Sociometry **20**, 253 (1957).
[84] M. De Domenico, A. Lancichinetti, A. Arenas, and M. Rosvall, *Identifying Modular Flows on Multilayer Networks Reveals Highly Overlapping Organization in Social Systems*, Phys. Rev. X **5**, 011027 (2015).
[85] B. Pasternak and I. Ivask, *Four Unpublished Letters*, Books Abroad **44**, 196 (1970).
[86] 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).

[87] V. Krebs, *Uncloaking Terrorist Networks*, First Monday **7**, 941 (2002).

[88] B. Hayes, *Connecting the Dots*, Am. Scientist **94**, 400 (2006).

[89] R. Guimerà, L. Danon, A. Díaz-Guilera, F. Giralt, and A. Arenas, *Self-Similar Community Structure in a Network of Human Interactions*, Phys. Rev. E **68**, 065103 (2003).