Reconstructing Networks with Unknown and Heterogeneous Errors

Tiago P. Peixoto*

Department of Mathematical Sciences and Centre for Networks and Collective Behaviour, University of Bath, Claverton Down, Bath BA2 7AY, United Kingdom and ISI Foundation, Via Chisola 5, 10126 Torino, Italy

(Received 25 June 2018; revised manuscript received 23 August 2018; published 16 October 2018; corrected 22 January 2019)

The vast majority of network data sets contain errors and omissions, although this fact is rarely incorporated in traditional network analysis. Recently, an increasing effort has been made to fill this methodological gap by developing network-reconstruction approaches based on Bayesian inference. These approaches, however, rely on assumptions of uniform error rates and on direct estimations of the existence of each edge via repeated measurements, something that is currently unavailable for the majority of network data. Here, we develop a Bayesian reconstruction approach that lifts these limitations by allowing for not only heterogeneous errors, but also for single edge measurements without direct error estimates. Our approach works by coupling the inference approach with structured generative network models, which enable the correlations between edges to be used as reliable uncertainty estimates. Although our approach is general, we focus on the stochastic block model as the basic generative process, from which efficient nonparametric inference can be performed and yields a principled method to infer hierarchical community structure from noisy data. We demonstrate the efficacy of our approach with a variety of empirical and artificial networks.

DOI: 10.1103/PhysRevX.8.041011

Subject Areas: Complex Systems, Statistical Physics

I. INTRODUCTION

The study of network systems of various kinds constitutes a significant fraction of contemporary interdisciplinary research in physics, biology, computer science, and social sciences, among other disciplines [1]. This research is motivated in large part by the surging availability of network data during the past couple of decades, which describe the detailed interactions among constituents of large-scale complex systems, such as transportation networks, cell metabolism, social contacts, the Internet, and various others. Despite the widespread growth of this field, its relative infancy is still noticeable in some aspects. In particular, even though sophisticated and successful models of network structure and function have been proposed, as well as powerful data-analysis methods, most studies of empirical data are performed without taking into account measurement error. Most typically, real networks are represented as adjacency matrices, sometimes enriched with additional information such as edge weights and types, as well as various kinds of node properties, the validity of which is simply taken for granted. But, as is true for any empirical scenario, network data are subject to observational errors: Parts of the network might not have been recorded, and parts might be wrong. Although this problem has been recognized in the past in several studies [2–11], the practice of ignoring measurement error is still mainstream, and robust methods to take it into account are underdeveloped. This practice is in no small part due to the fact that most available network data contain no quantitative error-assessment information of any kind, thus preventing primary experimental uncertainties to be propagated up the chain of analysis.

In this work, we formulate a principled method to reconstruct networks that have been imperfectly measured. We do so by simultaneously formulating generative models of network structure—that incorporate degree heterogeneity, modules, and hierarchies—as well as models of the noisy measurement process. By performing Bayesian statistical inference of this joint model, we are able to reconstruct the underlying network given an imperfect measurement affected by observational noise. Importantly, our method works also when a single measurement of the underlying network has been made and the noise magnitudes are unknown, which means it can be directly applied to the majority of network data without available error estimates. In addition to this, our method is capable of extracting hierarchical modular structure from such noisy networks, thus

*t.peixoto@bath.ac.uk

generalizing the task of community detection to this uncertain setting.

Our method is equally applicable when information on measurement error is available, either as repeated measurements or as estimated edge probabilities. For this class of data, we construct a general model that allows for heterogeneous errors that vary in different parts of the network. We show strong empirical evidence for the existence of this kind of heterogeneity and demonstrate the efficacy of our method to include it in the reconstruction.

Our method shares some underlying similarities with well-known model-based approaches of edge prediction [5,6] but is different from them in fundamental aspects. Most importantly, model-based edge-prediction methods yield *relative* probabilities of edges existing or not, given a generative model fitted to the observed data. These relative probabilities can be used to reconstruct a network, provided one knows how many edges are missing or spurious. Our method obviates the need for this information (which is, in general, unknown) and yields not only a reconstructed network but also the uncertainty estimate that must come with it, via a posterior distribution over all possible reconstructions. Thus, our method realizes the underlying promise of reconstruction that motivates most edge-prediction methods, but in a principled and nonparametric way.

We form the basis of our reconstruction scenario on Ref. [10], which defined a statistical inference method based on multiple measurements of network data, but here we use a different approach based on nonparametric Bayesian inference, combined with community detection. This approach yields a more powerful method that, differently from Ref. [10], can be applied also when the network data do not contain any kind of primary error estimate, such as when the edges and nonedges have been measured only once.

This work is organized as follows. In Sec. II, we formulate our Bayesian reconstruction framework. In Sec. II A, we present our measurement model, and in Sec. II B, we illustrate the use of our reconstruction method with some examples. In Sec. II C, we perform a detailed analysis of the reconstruction performance of the method as well as its use to provide estimates of various network properties. In Sec. II D, we employ our approach to some empirical network data without primary error estimates and evaluate their reliability. In Sec. II E, we extend our method to heterogeneous errors and use it to analyze network data with multiple measurements. In Sec. III, we show how our method can be extended to situations where the arbitrary error estimates are extrinsically provided, and we finalize in Sec. IV with a conclusion.

II. BAYESIAN NETWORK RECONSTRUCTION

The scenario we consider is one where, instead of a direct observation of a network $\boldsymbol{A}$, we perform a noisy measurement $\mathcal{D}$ that contains only indirect information about $\boldsymbol{A}$. The task of network reconstruction is then to obtain $\boldsymbol{A}$ from $\mathcal{D}$. The approach we take is to perform statistical inference, where first we model the network generating process via a probability

$$P(\boldsymbol{A}|\theta), \tag{1}$$

where θ are arbitrary model parameters. The entire data-generating process is then completed by modeling also the noisy measurement

$$P(\mathcal{D}|\boldsymbol{A},\phi), \tag{2}$$

conditioned on the generated network $\boldsymbol{A}$ (the "true" network) and some further parameters ϕ. Given this general setup, the reconstruction procedure consists of determining $\boldsymbol{A}$ from the posterior distribution

$$P(\boldsymbol{A}|\mathcal{D}) = \frac{P(\mathcal{D}|\boldsymbol{A})P(\boldsymbol{A})}{P(\mathcal{D})}, \tag{3}$$

where

$$P(\mathcal{D}|\boldsymbol{A}) = \int P(\mathcal{D}|\boldsymbol{A},\phi)P(\phi)d\phi \tag{4}$$

is the marginal probability of the measurements $\mathcal{D}$ and

$$P(\boldsymbol{A}) = \int P(\boldsymbol{A}|\theta)P(\theta)d\theta \tag{5}$$

is the prior probability for $\boldsymbol{A}$, summed over all possible parameter choices, weighted according to their (hyper)prior probabilities. The remaining term $P(\mathcal{D}) = \sum_{\boldsymbol{A}} P(\mathcal{D}|\boldsymbol{A})P(\boldsymbol{A})$ is a normalization constant that corresponds to the total probability—or *evidence*—for the observed measurement. In the above, the probabilities $P(\theta)$ and $P(\phi)$ encode our prior knowledge (or lack thereof) about the network generation and measurement processes, respectively. With these at hand, Eq. (3) assigns the probability of a given network $\boldsymbol{A}$ being responsible for measurement $\mathcal{D}$. Importantly, this distribution defines an ensemble of possibilities for the underlying network $\boldsymbol{A}$ that incorporates the amount uncertainty resulting from the measurement. This procedure contrasts with reconstruction approaches that attempt to reproduce a single network, although within the above framework we could also attempt to find the single most likely reconstruction that maximizes Eq. (3), i.e., *a maximum posterior point estimate*. However, as we see below, this is not the most appropriate point estimate, as it tends to incorporate noise from the data, biasing the reconstruction. Instead, we should consider the consensus of the full posterior distribution, which can also give us an estimation of uncertainty.

The above framework is general and can be used for any kind of generative and measurement processes. Here, we are interested in those that can be used to describe

the large-scale modular structures of networks, characterized by the partition of the nodes into groups $\boldsymbol{b} = \{b_i\}$, where $b_i \in \{1, \ldots, B\}$ is group membership of node i. The simplest and most commonly used model in this context is the stochastic block model (SBM) [12]

$$P(\boldsymbol{A}|\boldsymbol{\omega}, \boldsymbol{b}) = \prod_{i<j} \omega_{b_i,b_j}^{A_{ij}} (1 - \omega_{b_i,b_j})^{1-A_{ij}}, \tag{6}$$

where ω_{rs} is the probability of an edge existing between nodes of groups r and s. Alternatively, we could also consider a more realistic version called the degree-corrected SBM (DCSBM) [13]:

$$P(\boldsymbol{A}|\boldsymbol{\lambda}, \boldsymbol{\kappa}, \boldsymbol{b}) = \prod_{i<j} \frac{e^{-\kappa_i \kappa_j \lambda_{b_i,b_j}} (\kappa_i \kappa_j \lambda_{b_i,b_j})^{A_{ij}}}{A_{ij}!}, \tag{7}$$

where λ_{rs} controls the number of edges between groups r and s and κ_i the expected degree of node i. This model variant decouples the degrees from the group memberships, allowing for arbitrary degree variability inside modules, a feature often found to be more compatible with real networks [14]. (Note that the DCSBM generates multigraphs with $A_{ij} \in \mathbb{N}$, whereas the SBM above generates simple graphs with $A_{ij} \in \{0, 1\}$, as our framework requires. In Appendix D, we amend this inconsistency.) Using the above, we compute the marginal network probability as

$$P(\boldsymbol{A}) = \sum_{\boldsymbol{b}} P(\boldsymbol{A}|\boldsymbol{b}) P(\boldsymbol{b}), \tag{8}$$

with

$$P(\boldsymbol{A}|\boldsymbol{b}) = \int P(\boldsymbol{A}|\boldsymbol{\lambda}, \boldsymbol{\kappa}, \boldsymbol{b}) P(\boldsymbol{\kappa}|\boldsymbol{b}) P(\boldsymbol{\lambda}|\boldsymbol{b}) d\kappa d\lambda, \tag{9}$$

integrated over the remaining model parameters, weighted by their respective prior probabilities. However, although Eq. (9) can be computed exactly [14], the complete marginal of Eq. (8) cannot, as it involves an intractable sum over all possible network partitions. Hence, instead of computing directly the posterior of Eq. (3), we obtain the joint posterior [15]

$$P(\boldsymbol{A}, \boldsymbol{b}|\mathcal{D}) = \frac{P(\mathcal{D}|\boldsymbol{A}) P(\boldsymbol{A}|\boldsymbol{b}) P(\boldsymbol{b})}{P(\mathcal{D})}, \tag{10}$$

which involves only quantities that can be computed exactly, except $P(\mathcal{D})$, which, as we shortly see, is unnecessary for the inference procedure. We do the above without any loss, as the original posterior of Eq. (3) can be obtained by marginalization, i.e.,

$$P(\boldsymbol{A}|\mathcal{D}) = \sum_{\boldsymbol{b}} P(\boldsymbol{A}, \boldsymbol{b}|\mathcal{D}). \tag{11}$$

This result means that, if we can sample from the joint posterior $P(\boldsymbol{A}, \boldsymbol{b}|\mathcal{D})$, we can compute any estimate $\hat{y}$ of a network property $y(\boldsymbol{A})$ (e.g., the clustering coefficient) over the full marginal $P(\boldsymbol{A}|\mathcal{D})$ by averaging it over the joint posterior, i.e.,

$$\hat{y} = \sum_{\boldsymbol{A}} y(\boldsymbol{A}) P(\boldsymbol{A}|\mathcal{D}) = \sum_{\boldsymbol{A},\boldsymbol{b}} y(\boldsymbol{A}) P(\boldsymbol{A}, \boldsymbol{b}|\mathcal{D}). \tag{12}$$

The procedure we use to sample from the posterior distribution is Markov chain Monte Carlo (MCMC). We consider move proposals of the kind $P(\boldsymbol{b}'|\boldsymbol{A}, \boldsymbol{b})$ and $P(\boldsymbol{A}'|\boldsymbol{A}, \boldsymbol{b})$ for the partition and network, respectively, and accept the proposal according to the Metropolis-Hastings [16,17] probability

$$\min\left(1, \frac{P(\boldsymbol{A}', \boldsymbol{b}'|\mathcal{D}) P(\boldsymbol{A}|\boldsymbol{A}', \boldsymbol{b}') P(\boldsymbol{b}|\boldsymbol{A}', \boldsymbol{b}')}{P(\boldsymbol{A}, \boldsymbol{b}|\mathcal{D}) P(\boldsymbol{A}'|\boldsymbol{A}, \boldsymbol{b}) P(\boldsymbol{b}'|\boldsymbol{A}, \boldsymbol{b})}\right), \tag{13}$$

which enforces detailed balance. If the move proposals are ergodic (i.e., they allow every network $\boldsymbol{A}$ and partition $\boldsymbol{b}$ to be proposed eventually), this algorithm will generate samples from the posterior distribution $P(\boldsymbol{A}, \boldsymbol{b}|\mathcal{D})$ after a sufficiently large number of iterations (usually determined by requiring that statistical properties of the chain, such as average log-probability, become stationary). The ratio in Eq. (13) can be determined exactly without computing the intractable constant $P(\mathcal{D})$ in Eq. (10), making this method asymptotically exact. We give more technical details of our MCMC procedure in Appendix B.

The above setup is still sufficiently general that it can be used with any variant of the SBM. In particular, here we make extensive use of the hierarchical DCSBM (HDCSBM) [14,18], which differs from the DCSBM in that a nested hierarchy of priors and hyperpriors is used in place of the single prior $P(\boldsymbol{\lambda}|\boldsymbol{b})$ for the connections between groups. In this model, groups are clustered hierarchically into metagroups, yielding a nested hierarchical partition $\{\boldsymbol{b}^l\}$, where $\boldsymbol{b}^l$ is the partition of the groups in level l. As discussed in Refs. [14,18], this choice of structured priors removes a tendency of noninformative priors to underfit [19] and enables the detection of structures at multiple scales while at the same time remaining unbiased with respect to different types of mixing patterns. Its posterior distribution is obtained in the same fashion, following the framework above, simply by replacing $\boldsymbol{b} \to \{\boldsymbol{b}^l\}$.

In the following, whenever we mention that we sample from the posterior $P(\boldsymbol{A}|\mathcal{D})$, it is meant that we sample from the joint posterior $P(\boldsymbol{A}, \boldsymbol{b}|\mathcal{D})$ and marginalize over $\boldsymbol{b}$, as described above. The same is true when using the hierarchical model; i.e., we sample from $P(\boldsymbol{A}, \{\boldsymbol{b}^l\}|\mathcal{D})$ and marginalize over the hierarchical partitions $\{\boldsymbol{b}^l\}$.

The main difference from typical community detection based on statistical inference is that here we are interested in not only detecting modules in networks but also inferring the network itself. Therefore, both the network and its partition into (hierarchical) groups are inferred from indirect data. As we see, the simultaneous detection of modules offers a substantial advantage to the reconstruction task, as it allows correlations among edges to inform it, which means that we are able to perform reconstruction in situations which would otherwise be impossible. But, before we proceed, we need to model the measurement process itself, as we do in the following.

A. Noisy network measurements

Here, we consider the scenario used in Ref. [10], where the edges of a network are measured directly and repeatedly, but the process is noisy and potentially distorts the network. In particular, we assume that for each node pair (i,j) we perform n_{ij} distinct measurements and record x_{ij} positive outcomes; i.e., an edge is observed. For each observation, we have a probability p of observing a missing edge (i.e., a false negative) and a probability q of observing a spurious edge (i.e., a false positive), depending in each case on whether the underlying network possesses or not an edge (i,j). Thus, for each edge, the observation probability is distributed according to a binomial distribution, with a success rate that depends on whether an edge exists in the underlying network, i.e.,

$$P(x_{ij}|n_{ij},A_{ij},p,q) = \binom{n_{ij}}{x_{ij}}[(1-p)^{x_{ij}}p^{n_{ij}-x_{ij}}]^{A_{ij}}[q^{x_{ij}}(1-q)^{n_{ij}-x_{ij}}]^{1-A_{ij}}. \tag{14}$$

Thus, the joint likelihood for the whole set of measurements $\boldsymbol{x}=\{x_{ij}\}$ is

$$\begin{aligned}P(\boldsymbol{x}|\boldsymbol{n},\boldsymbol{A},p,q) &= \prod_{i<j}P(x_{ij}|n_{ij},A_{ij},p,q)\\ &= \left[\prod_{i<j}\binom{n_{ij}}{x_{ij}}\right](1-p)^{\mathcal{T}}p^{\mathcal{E}-\mathcal{T}}q^{\mathcal{X}-\mathcal{T}}(1-q)^{\mathcal{M}-\mathcal{X}-\mathcal{E}+\mathcal{T}},\end{aligned} \tag{15}$$

written in terms of the following summary quantities:

$$\mathcal{M}=\sum_{i<j}n_{ij}, \qquad \mathcal{X}=\sum_{i<j}x_{ij}, \tag{16}$$

$$\mathcal{E}=\sum_{i<j}n_{ij}A_{ij}, \qquad \mathcal{T}=\sum_{i<j}x_{ij}A_{ij}, \tag{17}$$

where $\mathcal{M}$ is the total number of measurements (edge or nonedge), $\mathcal{X}$ is the total number of observed edges, $\mathcal{E}$ is the total number of measured edges, and $\mathcal{T}$ is the total number of correctly observed edges. [20] From these quantities, we also identify the total number of false positives (spurious edges) as $\mathcal{X}-\mathcal{T}$ and of false negatives (missing edges) as $\mathcal{E}-\mathcal{T}$.

To proceed with our calculation, we need to specify the degree of prior knowledge we have on the error rates p and q. We can express this belief most naturally with a Beta distribution:

$$P(p|\alpha,\beta)=\frac{p^{\alpha-1}(1-p)^{\beta-1}}{\mathcal{B}(\alpha,\beta)}, \tag{18}$$

where $\mathcal{B}(x,y)=\Gamma(x)\Gamma(y)/\Gamma(x+y)$ is the Euler beta function, $\Gamma(x)$ is the gamma function, and likewise for $P(q|\mu,\nu)$, with hyperparameters μ and ν. As illustrated in Fig. 17 in Appendix A, a value of $\alpha=\beta=1$ encodes a maximum amount of prior ignorance with respect to p, which is then uniformly distributed in the unit interval. Conversely, values $\alpha\to\infty$ and $\beta\to\infty$ converge to a Dirac delta function centered at $\alpha/(\alpha+\beta)$, amounting to a maximum certainty for a particular value of p, and, therefore, intermediary values of α and β interpolate between these two extremes (and analogously for q with μ and ν), with which we can compute the integrated likelihood

$$\begin{aligned}P(\boldsymbol{x}|\boldsymbol{n},\boldsymbol{A},\alpha,\beta,\mu,\nu) &= \int P(\boldsymbol{x}|\boldsymbol{n},\boldsymbol{A},p,q)P(p|\alpha,\beta)P(q|\mu,\nu)dpdq\\ &= \left[\prod_{i<j}\binom{n_{ij}}{x_{ij}}\right]\frac{\mathcal{B}(\mathcal{E}-\mathcal{T}+\alpha,\mathcal{T}+\beta)}{B(\alpha,\beta)}\\ &\quad\times\frac{\mathcal{B}(\mathcal{X}-\mathcal{T}+\mu,\mathcal{M}-\mathcal{X}-\mathcal{E}+\mathcal{T}+\nu)}{\mathcal{B}(\mu,\nu)}.\end{aligned} \tag{19}$$

The noninformative case $\alpha=\beta=\mu=\nu=1$ simplifies further to

$$P(\boldsymbol{x}|\boldsymbol{n},\boldsymbol{A}) = \left[\prod_{i<j}\binom{n_{ij}}{x_{ij}}\right]\binom{\mathcal{E}}{\mathcal{T}}^{-1}\frac{1}{\mathcal{E}+1}\binom{\mathcal{M}-\mathcal{E}}{\mathcal{X}-\mathcal{T}}^{-1}\times\frac{1}{\mathcal{M}-\mathcal{E}+1}. \tag{20}$$

The above noninformative generative process can also be equivalently interpreted as first choosing the number of false positives $\mathcal{X}-\mathcal{T}$ uniformly from the interval $[0,\mathcal{M}-\mathcal{E}]$ and then selecting them uniformly at random from the possible set with $\binom{\mathcal{M}-\mathcal{E}}{\mathcal{X}-\mathcal{T}}$ elements, and similarly choosing the number of false negatives $\mathcal{E}-\mathcal{T}$ uniformly in

the interval $[0,\mathcal{E}]$ and the false negatives from the set of size $\binom{\mathcal{E}}{\mathcal{E}-\mathcal{T}} = \binom{\mathcal{E}}{\mathcal{T}}$.

With the integrated likelihood in place, we can finally complete the posterior distribution of Eq. (3) with $\mathcal{D} = (\boldsymbol{n}, \boldsymbol{x})$, which in this case becomes

$$P(\boldsymbol{A}|\boldsymbol{n},\boldsymbol{x},\alpha,\beta,\mu,\nu) = \frac{P(\boldsymbol{x}|\boldsymbol{n},\boldsymbol{A},\alpha,\beta,\mu,\nu)P(\boldsymbol{A})}{P(\boldsymbol{x}|\alpha,\beta,\mu,\nu)}. \quad (21)$$

For $P(\boldsymbol{A})$, we use the SBM and sample $\boldsymbol{A}$ using MCMC from the joint posterior $P(\boldsymbol{A},\boldsymbol{b}|\boldsymbol{n},\boldsymbol{x},\alpha,\beta,\mu,\nu)$, as discussed previously.

Even though we integrate over the error probabilities p and q in the above, we can nevertheless obtain their posterior estimates by averaging from the above posterior

$$P(p|\boldsymbol{n},\boldsymbol{x},\alpha,\beta,\mu,\nu) = \sum_{\boldsymbol{A}} P(p|\boldsymbol{n},\boldsymbol{x},\boldsymbol{A},\alpha,\beta)P(\boldsymbol{A}|\boldsymbol{n},\boldsymbol{x},\alpha,\beta,\mu,\nu), \quad (22)$$

using the posterior for p conditioned on the network $\boldsymbol{A}$,

$$P(p|\boldsymbol{n},\boldsymbol{x},\boldsymbol{A},\alpha,\beta) = \frac{p^{\mathcal{E}-\mathcal{T}+\alpha-1}(1-p)^{\mathcal{T}+\beta-1}}{\mathcal{B}(\mathcal{E}-\mathcal{T}+\alpha,\mathcal{T}+\beta)}, \quad (23)$$

and likewise for q with

$$P(q|\boldsymbol{n},\boldsymbol{x},\boldsymbol{A},\mu,\nu) = \frac{q^{\mathcal{X}-\mathcal{T}+\mu-1}(1-q)^{\mathcal{M}-\mathcal{X}-\mathcal{E}+\mathcal{T}+\mu-1}}{\mathcal{B}(\mathcal{X}-\mathcal{T}+\mu,\mathcal{M}-\mathcal{X}-\mathcal{E}+\mathcal{T}+\nu)}. \quad (24)$$

In the following, we most often assume the noninformative case $\alpha=\beta=\nu=\mu=1$, corresponding to the maximum lack of prior knowledge about the measurement noise. In order to unclutter our expressions, if this is the case, we simply omit those hyperparameters from the posterior distribution, i.e., $P(\boldsymbol{A}|\boldsymbol{n},\boldsymbol{x}) \equiv P(\boldsymbol{A}|\boldsymbol{n},\boldsymbol{x},\alpha=1,\beta=1,\mu=1,\nu=1)$.

1. Single edge measurements

As we increase the number of measurements n_{ij} of each pair of nodes, we should expect also to increase the accuracy of the reconstruction, resulting in a posterior distribution $P(\boldsymbol{A}|\boldsymbol{n},\boldsymbol{x})$ that is very sharply peaked around the true underlying network. Although this scenario is plausible, and indeed desirable under controlled experimental conditions, it is not representative of the majority of the network data that are currently available. In fact, inspecting comprehensive network catalogs such as the Koblenz Network Collection (KONECT) [21] and the Colorado Index of Complex Networks (ICON) [22] reveals a very pauper set of network data that can be cast under this setting of repeated measurements. On the contrary, the vast majority of them offer only a single adjacency matrix without quantitative error estimates of any kind. Needless to say, this omission is no reason to assume that they do not, in fact, contain errors, only that they have not been assessed or published.

Here, we propose an approach of assessing the uncertainty of this dominating kind of network data by interpreting it as a single measurement with unknown errors rates, using the framework outlined above. In more detail, we assume that $n_{ij}=1$ for every pair i, j and that the single measurements $x_{ij} \in \{0,1\}$ correspond to the reported adjacency matrix. The lack of knowledge about the underlying error rates p and q can be expressed by choosing $\alpha=\beta=\mu=\nu=1$, in which case it is assumed that they both lie *a priori* anywhere in the unit interval. [23] At first, we may wonder if this approach has any chance of succeeding, since the lack of knowledge about the error rates means that the network could have been modified in arbitrary ways such that the true underlying network is radically different from what has been observed. Indeed, if we define the distance between measured and generated networks,

$$d(\boldsymbol{A},\boldsymbol{x}) = \sum_{i<j}|A_{ij}-x_{ij}| = (\mathcal{E}-\mathcal{T})+(\mathcal{X}-\mathcal{T}), \quad (25)$$

which equals the sum of false negatives and false positives, we have that, according to Eq. (20), the expected distance over many measurements is

$$\bar{d}(\boldsymbol{A}) = \sum_{\boldsymbol{x}} d(\boldsymbol{A},\boldsymbol{x})P(\boldsymbol{x}|\boldsymbol{n},\boldsymbol{A}) = \binom{N}{2}/2, \quad (26)$$

which is half the maximum possible distance of $\binom{N}{2}$, which might lead us to conclude that our noise model will invariably destroy the network beyond the possibility of reconstruction, regardless of its original structure. What changes this picture is the fact that the posterior distribution $P(\boldsymbol{A}|\boldsymbol{x},\boldsymbol{n})$ of Eq. (21) is, in fact, more concentrated on the generated network than implied by the above and, ultimately, depends crucially on our generative process $P(\boldsymbol{A})$. The first point can be made by assuming a fully random generative model,

$$P(\boldsymbol{A}|\omega) = \prod_{i<j}\omega^{A_{ij}}(1-\omega)^{1-A_{ij}}, \quad (27)$$

which means that the true networks being measured are assumed to be completely random, given a particular density ω. The full prior can be obtained by a noninformative assumption $P(\omega)=1$, which yields

$$P(\boldsymbol{A}) = \int P(\boldsymbol{A}|\omega)P(\omega)d\omega \quad (28)$$

$$= \left(\binom{N}{2}\right)^{-1} \frac{1}{\binom{N}{2}+1}, \tag{29}$$

with $E = \sum_{i<j} A_{ij} = \mathcal{E}$ being the total number of edges, which is equivalent to sampling to the total number of edges from the interval $[0, \binom{N}{2}]$ and then a fully random graph with that number of edges. Combining this result with Eq. (20) yields the posterior distribution, which can be written as the product of two conditional probabilities:

$$P(\boldsymbol{A}|\boldsymbol{x},\boldsymbol{n}) = P(\boldsymbol{A}|\boldsymbol{x},\mathcal{T},\mathcal{E})P(\mathcal{T},\mathcal{E}|\boldsymbol{x}) \tag{30}$$

with

$$P(\boldsymbol{A}|\boldsymbol{x},\mathcal{T},\mathcal{E}) = \binom{\mathcal{X}}{\mathcal{X}-\mathcal{T}}^{-1} \binom{\binom{N}{2}-X}{\mathcal{E}-\mathcal{T}}^{-1} \tag{31}$$

corresponding to the uniform sampling of $\boldsymbol{A}$ with exactly $\mathcal{E}-\mathcal{T}$ false negatives and $\mathcal{X}-\mathcal{T}$ false positives, and

$$P(\mathcal{T},\mathcal{E}|\boldsymbol{x}) \propto \frac{[\mathcal{T}\leq\mathcal{E}][\mathcal{T}\leq\mathcal{X}]}{(\mathcal{E}+1)[\binom{N}{2}-\mathcal{E}+1]} \tag{32}$$

with $[\ldots]$ being the Inverson bracket that equals 1 if the condition inside it is true, or 0 otherwise, determines the posterior probability of the number of false negatives and false positives, up to a normalization constant. Although this distribution decays for values of $\mathcal{E}$ larger than 0, the decay is slow with approximately $1/\mathcal{E}$, and, hence, on average, the inferred networks $\boldsymbol{A}$ sampled from $P(\boldsymbol{A}|\boldsymbol{x},\boldsymbol{n})$ will be dense, yielding large distances $d(\boldsymbol{A},\boldsymbol{A}^*)$ if the true generated network $\boldsymbol{A}^*$ is sparse. Although the posterior distribution of false negatives and positives resulting from $P(\mathcal{T},\mathcal{E}|\boldsymbol{x})$ is not uniformly distributed in the allowed interval, it is also not sufficiently concentrated to enable any reasonable accuracy in the reconstruction, regardless of how large the network is. What changes this situation considerably is to replace the fully random model of Eq. (28) by a more structured model. The key observation here is that the modifications induced by the error rates p and q affect uniformly every edge and nonedge, and, thus, with structured models, we can exploit the observed correlations in the measurements $\boldsymbol{x}$ to infer the underlying network $\boldsymbol{A}$ and, in fact, even the error rates p and q, which are *a priori* unknown.

We illustrate this property by considering the nondegree-corrected SBM, where networks are generated with probability

$$P(\boldsymbol{A}|\boldsymbol{\omega},\boldsymbol{b}) = \prod_{i<j} \omega_{b_i b_j}^{A_{ij}} (1-\omega_{b_i b_j})^{1-A_{ij}}. \tag{33}$$

The final likelihood for the measurements $\boldsymbol{x}$ in this case are identical to an effective SBM, given by

$$P(\boldsymbol{x}|\boldsymbol{n},p,q,\boldsymbol{\omega},\boldsymbol{b}) = \sum_{\boldsymbol{A}} P(\boldsymbol{x}|\boldsymbol{n},\boldsymbol{A},p,q)P(\boldsymbol{A}|\boldsymbol{\omega},\boldsymbol{b}) \tag{34}$$

$$= \prod_{i<j} \omega'^{\,x_{ij}}_{b_i b_j} (1-\omega'_{b_i b_j})^{1-x_{ij}}, \tag{35}$$

where

$$\omega'_{rs} = (1-p-q)\omega_{rs} + q \tag{36}$$

are the new effective SBM probabilities that have been scaled and shifted by the measurement noise. Suppose, for simplicity, that we know the true network partition $\boldsymbol{b}$ and that the number of groups is very small compared to the number of nodes in each group. In this situation, the posterior distribution for $\boldsymbol{\omega}'$ should be tightly peaked around the maximum likelihood estimate $\hat{\boldsymbol{\omega}}'$:

$$\hat{\omega}'_{rs} = (1-p-q)\omega_{rs} + q = \frac{e_{rs}}{n_r n_s}, \tag{37}$$

where $e_{rs} = \sum_{ij} x_{ij}\delta_{b_i,r}\delta_{b_j,s}$ is the number of observed edges between groups r and s (or twice that for $r=s$) and n_r is the number of nodes in group r. The joint posterior distribution for p and q is then asymptotically given by

$$\begin{aligned} P(p,q|\boldsymbol{x},\boldsymbol{n},\boldsymbol{b}) &\propto \int P(\boldsymbol{x}|\boldsymbol{n},p,q,\boldsymbol{\omega},\boldsymbol{b})P(\boldsymbol{\omega}|\boldsymbol{b})d\boldsymbol{\omega} \\ &\propto \prod_{r\leq s}\int_0^1 \delta[(1-p-q)\omega_{rs}+q-e_{rs}/n_r n_s]P(\omega_{rs}|\boldsymbol{b})d\omega_{rs} \\ &\propto \prod_{r\leq s}\left[0\leq\frac{e_{rs}/n_r n_s-q}{1-p-q}\leq 1\right]\frac{P(\frac{e_{rs}/n_r n_s-q}{1-p-q}|\boldsymbol{b})}{1-p-q}, \end{aligned} \tag{38}$$

up to normalization, where $[\ldots]$ is again the Inverson bracket. The constraints above imply that the inferred error rates are bounded by the maximum and minimum inferred connection probabilities, i.e.,

$$\hat{q} \leq \min_{rs} \frac{e_{rs}}{n_r n_s}, \tag{39}$$

$$\hat{p} \leq 1 - \max_{rs} \frac{e_{rs}}{n_r n_s}. \tag{40}$$

These bounds mean that, if we have not observed many edges between groups r and s, this implies that q could not have been very large. If, instead, we do observe many edges between these groups, then it means that the value of p could not have been very large either [see Figs. 1(a) and 1(b)]. The inequalities of Eqs. (39) and (40) hold for every pair of groups r and s, but the values of p and q are global. Therefore, as long as the inferred SBM probabilities are

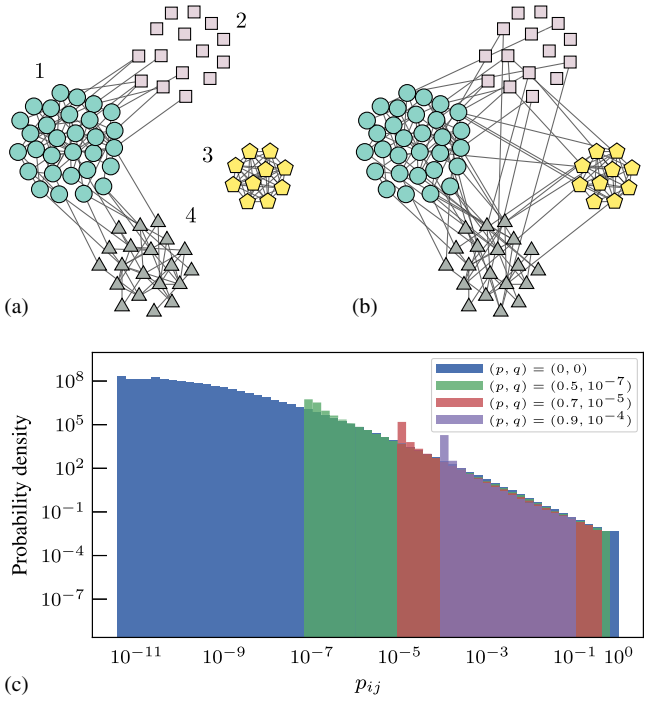

FIG. 1. (a) Illustration of a hypothetical measured network, with *a priori* unknown errors, but from which error estimates can be made: The lack of edges between groups 2 and 3, 3 and 4, 2 and 4, and 1 and 3 implies that the probability q of missing edges is likely to be low. Similarly, the large internal density of group 3 (which forms a clique of ten nodes) implies that the missing edge probability p must be low as well. (b) How the network in (a) would look for higher values of p and q. (c) The distribution of marginal edge probabilities p_{ij} between every node pair, for a fit of the HDCSBM on the openflights data (see Appendix E), measured with different values of the noise parameters (p, q). As the noise magnitudes increase, the probabilities become less heterogeneous and concentrate in narrower intervals. Hence, the inference of broad connection probabilities from the data rules out the existence of strong noise in the measurement.

sufficiently *heterogeneous*, they should constrain the inferred error rates to narrow intervals—which also constrains the inferred number of false negatives and false positives [see Fig. 1(c)]. [24] On the other hand, if the model probabilities are homogeneous, the posterior distribution for the errors is broad, and the quality of the reconstruction is poor. Therefore, the success of this approach depends ultimately on the observed networks being sufficiently structured and of our models being capable of describing them.

The above means that we have a better chance of accurate reconstruction if our models are capable of detecting heterogeneous connection probabilities among nodes. A fully uniform model like the Erdős-Renyi of Eq. (28) (equivalent to a SBM with only one group) exhibits the worst possible performance. The DCSBM, on the other hand, should, in general, perform better than the SBM, since it is capable of capturing degree heterogeneity inside groups, which is a common feature of many networks [13,14]. The HDCSBM [14,18] should perform even better, since its tendency not to underfit means it can detect statistically significant structures at smaller scales.

Finally, it must also be noted that, when performing only single measurements, there remains an unavoidable identification problem, where it becomes impossible to fully distinguish a network that has been sampled from a SBM with parameters $\boldsymbol{\omega}$ and error rates p and q from the same network sampled from a SBM with parameters $\boldsymbol{\omega}'$ given by Eq. (36) and error rates $p = q = 0$ (and, in fact, any interpolation between these two extremes). This uncertainty, however, is reflected in the variance of the posterior distribution and serves as a worst-case estimation of the error rates, which ultimately can be improved by either incorporating better prior knowledge (e.g., via the hyperparameters α, β, ν, and μ) or performing multiple measurements.

B. Empirical examples

Before we proceed further with a systematic analysis of our reconstruction method, we illustrate its behavior with some empirical data that are likely to contain errors and omissions. We begin with the network of social associations between 62 terrorists responsible for the 9/11 attacks [25,26]. The existence of an edge between two terrorists is established if there is evidence that they interacted directly in some way, e.g., if they attended the same college or shared an address. Clearly, this approach is inherently unreliable, as either investigators may fail to record evidence or the evidence recorded may be simply erroneous. Nevertheless, although this potential unreliability is acknowledged in Refs. [25,26], it is not assessed quantitatively, and the data presented there are a single adjacency matrix with no error estimates. Therefore, it serves as a suitable candidate for the application of our reconstruction method. When applied to this data set, our approach yields the results seen in Fig. 2, which shows the marginal posterior probability of each possible edge in the network, in addition to the hierarchical modular structured captured by the HDCSBM. Our method identifies the organization into a few largely disconnected cells, typical of terrorist groups. When ranking the potential edges according to their marginal posterior probability, as shown in Fig. 2(c), we have that all observed edges are more likely to be true edges than any of the nonedges, indicating a fair degree of inferred reliability. The observed nonedges have a probability substantially smaller than the observed edges of being edges, with the sole exception of a connection between Mohamed Atta (one of the main leaders) and Waleed al-Shehri, which is not considered in Refs. [25,26] but to which our method ascribes a reasonably high probability of 0.48. Atta is connected to all members of al-Shehri's group, and, according to the HDCSBM, the sole missing link

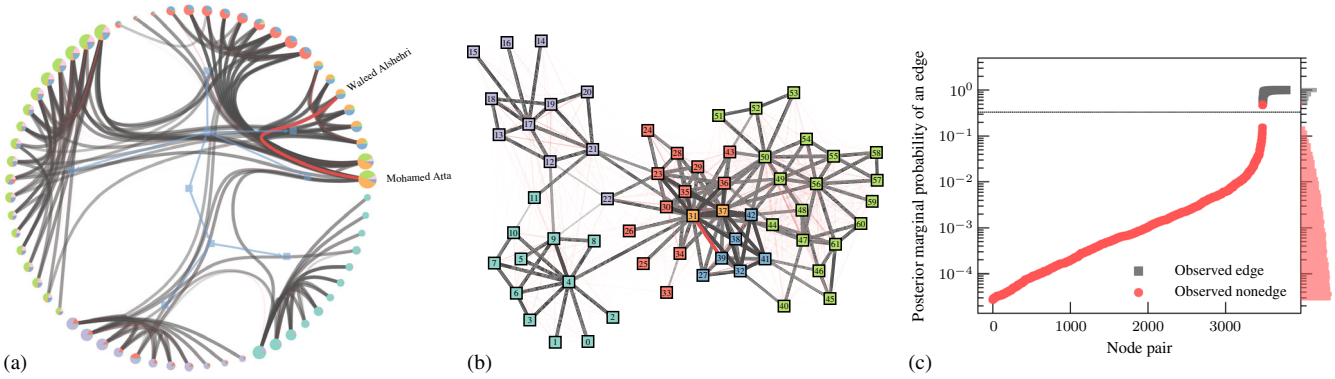

FIG. 2. Network of social associations between 9/11 terrorists [25,26]. This network is measured by potentially unreliable means, but no quantitative error estimates are known, and no repeated measurements were made. In (a) and (b) is shown the inferred network according to our method—which does not require direct error estimates or repeated measurements—where the edge thickness indicates the posterior marginal probability of an edge existing. In (a), the inferred hierarchical structure is shown, with pie charts on the nodes indicating the marginal probabilities of group memberships, and, in (b), a spatial layout of the same network shows the lowest level of the hierarchy as the node colors. The edge shown in red is inferred as existing with a large probability, despite not being measured. Other potentially missing edges are also shown in red, with a probability given by their thickness and opacity. In (c) is shown the marginal probability of edge existence for all node pairs, indicating a fair amount of inferred reliability—with the exception of the single missing edge highlighted in (a) and (b)—despite the lack of direct error estimates in the data. The horizontal line marks a $1/3$ probability as a visual aid. The missing edge corresponds to a connection between Mohamed Atta and Waleed Alshehri, which is not considered in Refs. [25,26] but is corroborated by reports that they shared an apartment in Berlin and met previously in Spain.

between them is therefore suspicious. Indeed, journalistic reports place both individuals occasionally sharing an apartment in Berlin [27] and meeting at least once in Spain [28], prior to the attacks, which seems to corroborate our reconstruction. The remaining observed nonedges have a probability of 0.15 or smaller, which should not be outright discarded, and could serve as candidates for further investigation.

We now move to another social network, namely, the interactions between 34 members of a karate club, originally studied by Zachary [29]. This network is widely used to evaluate community detection methods, after its use for this purpose in Ref. [30]. It is recorded just before the split of the club into two disjoint groups after a conflict, and many community detection methods are capable of accurately predicting the split by detecting communities from this snapshot. However, not only does the original publication of Ref. [29] omit any assessment of measurement uncertainties, but also it clearly contains one obvious error: The adjacency matrix $\boldsymbol{A}$ published in the original study, although it is supposed to be symmetric, contains two inconsistent entries with $A_{ij} \neq A_{ji}$, for $(i,j) = (23,34)$, creating an ambiguity about the existence of this particular edge [31]. The authors of Ref. [30] make the decision of assuming $A_{23,34} = 1$, even though there seems to be no obvious reason to decide either way *a priori*. The vast majority of other works in the area follow suit (possibly inadvertently), thus incorporating this potential, though arguably small, error in their analysis. Here, we tackle this reconstruction problem by mapping the uncertain data set of Ref. [29] to our framework. Since each node pair (i,j) is also presented reversed (j,i), we consider these as independent measurements such that $n_{ij} = 2$ for every pair (i,j). Since the measurements are consistent for all but one pair, we have $x_{ij} = 2$ or 0, except for the offending entry with $x_{(23,34)} = 1$. Based on this setup, we employ our reconstruction approach to obtain $P(\boldsymbol{A}|\boldsymbol{n},\boldsymbol{x})$, using as generative processes the Erdős-Rényi (ER) model (equivalent to a SBM with only one group, $B = 1$), the configuration model (CM) (equivalent to a DCSBM with $B = 1$), and the HDCSBM. As we see in Fig. 3, the ER model is incapable of disambiguating the data, as it cannot be used to detect any structure in it, and ascribes a posterior probability of 0.5 to the uncertain edge. Both the CM and the HDCSBM, however, ascribe high probabilities for the edge, of 0.87 and 0.93, respectively. The CM approach is able to recognize that, since node 34 is a hub in the network, an edge connecting to it is more likely to occur than not, and the HDCSBM can further use the fact that both nodes belong to the same group. Therefore, it seems like the choice made by the authors of Ref. [30] of assuming $A_{23,34} = 1$ is fortuitous, and the *de facto* instance of this network used by the majority of researchers is the one mostly likely to correspond to the original study.

In the following, we move to a systematic analysis of the reconstruction method, based on empirical and simulated data.

C. Reconstruction performance

Before we evaluate the performance of the reconstruction approach, we must first decide how to quantify it. As a

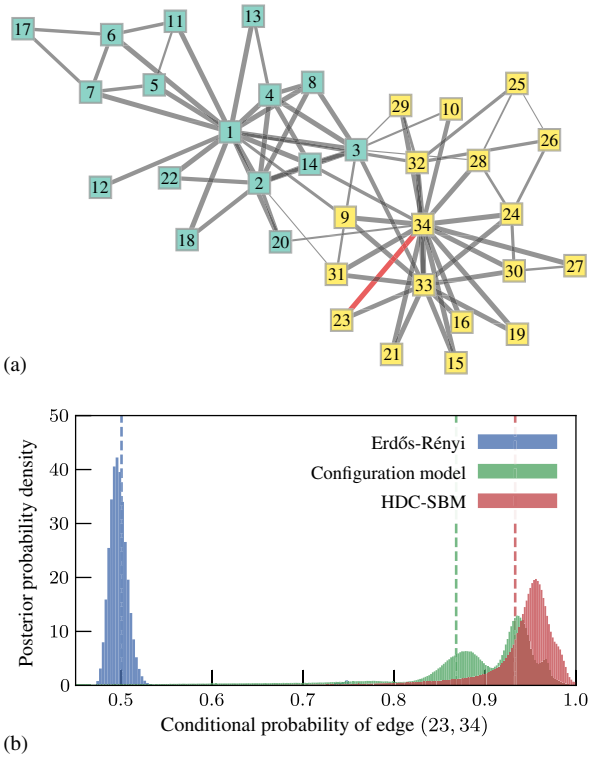

FIG. 3. Inferred Zachary's karate club network using the uncertain data from the original publication [29], which contain an ambiguous edge (23,34), as explained in the text. (a) Layout of the reconstructed network showing the posterior edge probabilities as edge thickness, according to the HDCSBM, and the ambiguous edge in red. The node colors correspond to a sample from the posterior distribution of the node partitions. (b) Posterior probability density of the probability of edge (23,34), conditioned on the remaining edges and model parameters, for the different model variants indicated in the legend and explained in the text. The vertical dashed lines indicate the distribution averages, corresponding to the marginal posterior probability of the edge.

criterion of how close an inferred network $\hat{\boldsymbol{A}}$ is to the true network $\boldsymbol{A}^*$ underlying the data, we use the distance of Eq. (25):

$$d(\hat{\boldsymbol{A}},\boldsymbol{A}^*) = \sum_{i<j}|\hat{A}_{ij} - A^*_{ij}|.$$

A successful reconstruction method should seek to find an estimate $\hat{\boldsymbol{A}}$ that minimizes this distance. However, since we do not have direct access to the true network $\boldsymbol{A}^*$, the best we can do is to consider the average distance over the posterior distribution given the noisy data:

$$\bar{d}(\hat{\boldsymbol{A}}) = \sum_{\boldsymbol{A}} d(\hat{\boldsymbol{A}},\boldsymbol{A})P(\boldsymbol{A}|\boldsymbol{x},\boldsymbol{n}) \tag{41}$$

$$= \sum_{i<j}|\hat{A}_{ij} - \pi_{ij}|, \tag{42}$$

where

$$\pi_{ij} = \sum_{\boldsymbol{A}} A_{ij}P(\boldsymbol{A}|\boldsymbol{x},\boldsymbol{n}) \tag{43}$$

is the marginal posterior probability of edge (i,j). If we minimize $\bar{d}(\hat{\boldsymbol{A}})$ with respect to $\hat{\boldsymbol{A}}$, we obtain

$$\hat{A}_{ij} = \begin{cases} 1 & \text{if } \pi_{ij} > 1/2, \\ 0 & \text{if } \pi_{ij} < 1/2, \end{cases} \tag{44}$$

for $\pi_{ij} \neq 1/2$. Equation (44) defines what is called a maximum marginal posterior (MMP) estimator, and it leverages the consensus of the entire posterior distribution of all possible networks for the estimation of every edge. Operationally, it can be obtained very easily by sampling networks from the posterior distribution and computing how often each edge is observed, yielding an estimate for $\boldsymbol{\pi}$ and, hence, $\hat{\boldsymbol{A}}$.

Given the above criterion, we evaluate the reconstruction performance by simulating the noisy measurement process. We do this evaluation by taking a real network $\boldsymbol{A}^*$ (which for this purpose we are free to declare to be error-free), obtaining a measurement $\boldsymbol{x}$ given error rates p and q, and measuring each edge and nonedge the same number of times $n_{ij} = n$. We choose p arbitrarily, and $q = pE/[\binom{N}{2} - E]$, where E is the number of edges in $\boldsymbol{A}^*$, so that the measured networks have the same average density as $\boldsymbol{A}^*$. Given a final measurement $\boldsymbol{x}$, we sample inferred networks $\boldsymbol{A}$ from the posterior distribution $P(\boldsymbol{A}|\boldsymbol{x},\boldsymbol{n})$ and compute the MMP estimate $\hat{\boldsymbol{A}}$ from the marginal distribution $\boldsymbol{\pi}$. The quality of the reconstruction is then assessed according to the similarity to the true network $\boldsymbol{A}^*$, $S(\hat{\boldsymbol{A}},\boldsymbol{A}^*) \in [0,1]$, defined as

$$S(\hat{\boldsymbol{A}},\boldsymbol{A}^*) = 1 - \frac{d(\hat{\boldsymbol{A}},\boldsymbol{A}^*)}{\sum_{i<j}\hat{A}_{ij} + A^*_{ij}}, \tag{45}$$

where $d(\hat{\boldsymbol{A}},\boldsymbol{A}^*)$ is the distance defined in Eq. (25). A value of $S(\hat{\boldsymbol{A}},\boldsymbol{A}^*) = 1$ indicates perfect reconstruction and $S(\hat{\boldsymbol{A}},\boldsymbol{A}^*) = 0$ the situation where $\hat{\boldsymbol{A}}$ and $\boldsymbol{A}^*$ do not share a single edge [34].

In Figs. 4(a) and 4(e) are shown the results of this procedure with the political blogs and openflights networks (see Appendix E). As a baseline, in both figures we show the direct similarity $S(\boldsymbol{x},\boldsymbol{A}^*)$ of the data obtained with $n = 1$ to the true network $\boldsymbol{A}^*$, as dashed curves. In both cases, the similarity of the inferred network $S(\hat{\boldsymbol{A}},\boldsymbol{A}^*)$ to the true network is larger than the one obtained with the direct observation $S(\boldsymbol{x},\boldsymbol{A}^*)$ for the vast majority of the parameter range, indicating systematic positive reconstruction even with single measurements. Expectedly, the quality of reconstruction increases progressively with a larger number

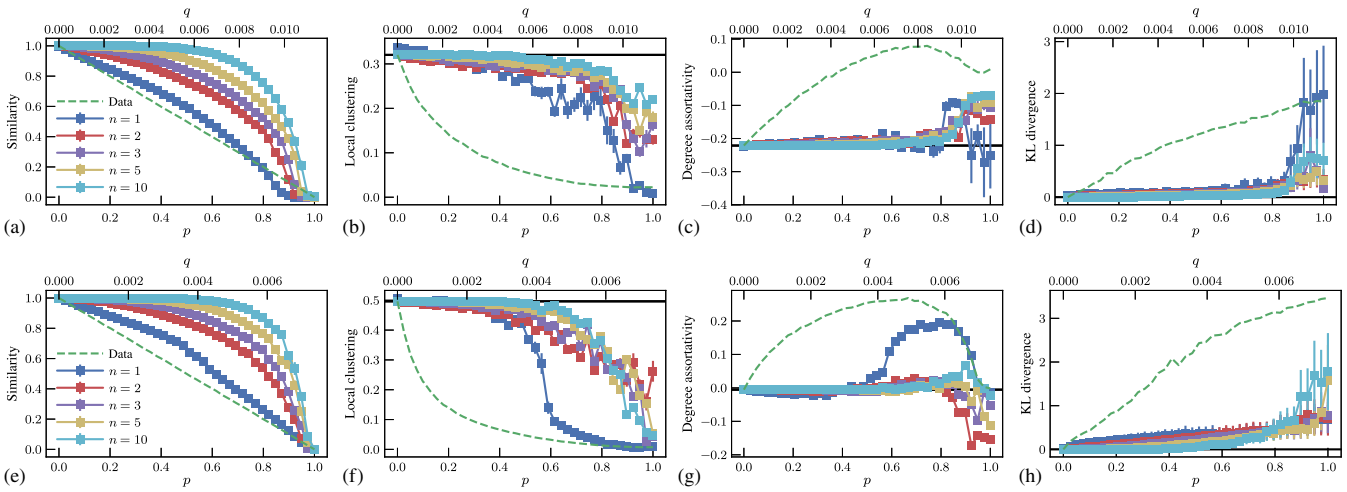

FIG. 4. Reconstruction performance for political blogs (top row) and openflights (bottom row) networks. In each case, the empirical network is considered as the true network, and simulated measurements are made for several values of missing edge probability p, with a spurious edge probability $q = pE/[\binom{N}{2} - E]$. (a),(e) Similarity of the MMP estimator to the true network, $S(\hat{\boldsymbol{A}}, \boldsymbol{A}^*)$, as a function of p, and for several values of the number of repeated measurements, n. (b),(c),(f),(g) Posterior average local clustering and degree assortativity coefficients, according to the same legend as (a) and (e). (d),(g) KL divergence between true and inferred degree distributions, as discussed in the text. In all cases [(a)–(h)], the dashed curve shows the corresponding value obtained directly with the measured data with $n = 1$, and the solid horizontal line marks the true value corresponding to perfect reconstruction.

of measurements n, with the similarity eventually approaching one. Although perfect reconstruction is not possible with single measurements when the noise is large, it is a noteworthy and nontrivial fact that the distance to the true network always decreases when performing it. This performance is possible only due to the use of a structured model such as the HDCSBM that can recognize the structure in the data and extrapolate from it. If one would use a fully random model in its place, the similarity would be zero in the entire range, if $n = 1$ (although it would improve for $n > 1$).

A particularly interesting outcome of the successful reconstruction is that the noise magnitudes p and q can be determined as well, even though they are not *a priori* known. As shown in Fig. 5, the posterior probabilities for p and q are very close to the true values used, even for single measurements. (The precision of the inferred values of q is generally higher than of p, as we are dealing with sparse networks, with vastly more nonedges than edges.) For the openflights data, the accurate noise recovery occurs only for moderate magnitudes, and a strong discrepancy is observed for values around $p \gtrsim 0.5$. In such situations, prior knowledge of the noise values could have aided the reconstruction for $n = 1$, but, otherwise, any benefit from this information would have been marginal. Again, the noise recovery becomes asymptotically exact as we increase the number of measurements and is already very accurate for $n = 2$.

We note that the results in Fig. 4 remain largely unchanged if the underlying network considered is sampled from the DCSBM with parameters inferred from the original data (not shown).

1. Estimating summary quantities

In addition to or instead of the network itself, we may want to estimate a given scalar observable $y(\boldsymbol{A})$ that acts as a summary of some aspect of the network's structure. In this case, we should seek to minimize the squared error with respect to the true network $\boldsymbol{A}^*$:

$$[\hat{y} - y(\boldsymbol{A}^*)]^2, \tag{46}$$

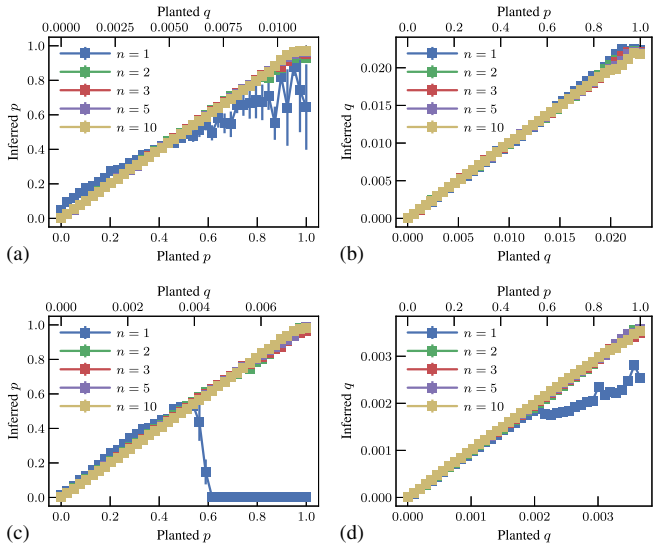

FIG. 5. Inferred values of noise magnitude p and q as a function of the planted values, for the same simulated measurements described in Fig. 4, for political blogs [(a),(b)] and openflights [(c),(d)] networks.

where $\hat{y}$ is our estimated value. Like before, without knowing $\boldsymbol{A}^*$, the best we can do is minimize the squared error averaged over the posterior distribution:

$$\sigma_{\hat{y}}^2 = \sum_{\boldsymbol{A}} [\hat{y} - y(\boldsymbol{A})]^2 P(\boldsymbol{A}|\boldsymbol{n}, \boldsymbol{x}). \tag{47}$$

Minimizing $\sigma_{\hat{y}}^2$ with respect to $\hat{y}$ yields the posterior mean estimator

$$\hat{y} = \sum_{\boldsymbol{A}} y(\boldsymbol{A}) P(\boldsymbol{A}|\boldsymbol{n}, \boldsymbol{x}). \tag{48}$$

We can also obtain the uncertainty of this estimator by computing its variance of Eq. (47) so that the uncertainty of $\hat{y}$ is summarized by its standard deviation $\sigma_{\hat{y}}$.

It is important to emphasize that, in general, $\hat{y} \neq y(\hat{\boldsymbol{A}})$, with $\hat{\boldsymbol{A}}$ being the MMP estimator of Eq. (44). In other words, the best estimate for $y(\boldsymbol{A}^*)$ (i.e., with minimal squared error) is not the same as the value obtained for the best estimate of $\boldsymbol{A}^*$ (i.e., with minimal distance).

In Figs. 4(b), 4(c), 4(f), and 4(g), we see the results of the same experiment described above, where we attempt to recover the average local clustering coefficient and the degree assortativity of the original network. As with the similarity, the inferred values are closer to the true network's. However, in this case, the values for $n = 1$ are substantially closer to the true value for a large range of noise magnitudes and are often indistinguishable from it, which means that, even in situations where the posterior distribution of inferred networks yields a relatively poor similarity to the true network, as it cannot precisely correct the altered edges and nonedges, it still shares a high degree of statistical similarity with it and can accurately reproduce these summary quantities.

2. Estimating degree distributions

We can also estimate degree distributions $\hat{p}_k$, defined as the probability that a node has degree k, by treating them like a collection of scalar measurements and minimizing the squared error $\sum_k [\hat{p}_k - p_k(\boldsymbol{A})]^2$ averaged over the posterior distribution, which yields the same posterior mean estimator used so far:

$$\hat{p}_k = \sum_{\boldsymbol{A}} p_k(\boldsymbol{A}) P(\boldsymbol{A}|\boldsymbol{x}, \boldsymbol{n}). \tag{49}$$

The same estimator is also obtained when minimizing the Kullback-Leibler (KL) divergence,

$$KL(p(\boldsymbol{A})||\hat{p}) = \sum_k p_k(\boldsymbol{A}) \ln \frac{p_k(\boldsymbol{A})}{\hat{p}_k}, \tag{50}$$

over the posterior, which offers a more convenient way to compare distributions, as it can be interpreted as the amount of information "lost" when $\hat{p}_k$ is used to approximate $p_k(\boldsymbol{A})$.

For the estimation of the degree probabilities $p_k(\boldsymbol{A})$ for each individual network sampled from the posterior, we model the degrees $\boldsymbol{k} = \{k_i\}$ as a multinomial distribution [35]

$$P(\boldsymbol{k}|\{p_k\}) = \frac{N! \prod_k p_k^{n_k}}{\prod_k n_k!}, \tag{51}$$

where n_k is the number of nodes of degree k. The probabilities themselves are modeled by a uniform Dirichlet mixture, i.e., sampled uniformly from a simplex constrained by the normalization $\sum_{k=0}^{K} p_k = 1$:

$$P(\{p_k\}) = K! \delta\left(\sum_k p_k - 1\right), \tag{52}$$

where K is the largest possible degree. With this, the posterior mean becomes

$$p_k(\boldsymbol{A}) = \frac{n_k + 1}{N + K + 1}. \tag{53}$$

This estimation is superior to the more naive $p_k = n_k/N$, as it is less susceptible to statistical fluctuations due to a lack of data, such as when $n_k = 0$, although it approaches it for $N \gg K$ and $n_k \gg 1$.

In Figs. 4(d) and 4(h) are shown the KL divergence between the inferred and true distributions, for the same experiments as before. Like with the local clustering and assortativity coefficients, the reconstructed degree distributions remain very close to the true one, despite the continuously decreasing similarity for larger noise magnitudes. In Fig. 6 can be seen the true, measured, and reconstructed distributions for the political blog network, for a value of $(p, q) = (0.41, 0.0094)$. Despite the relatively high noise magnitudes, a single measurement of the network does fairly well in reconstructing the original

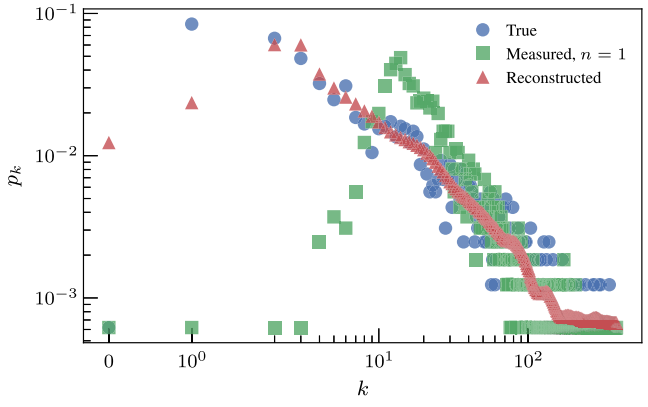

FIG. 6. True, measured (with $n = 1$), and reconstructed degree distributions of the political blog network, with noise magnitudes $(p, q) = (0.41, 0.0094)$.

distribution, failing mostly only for degrees zero and one, despite the significant distortion caused by the noisy measurement process.

3. Edge prediction: Network denoising and completion

The reconstruction task we have been considering shares many similarities with the task of model-based edge prediction [5,6] but is also different from it in some fundamental aspects. Most typically, edge prediction is formulated as a binary classification task [7], in which each missing (or spurious) edge is attributed a "score" (which may or may not be a probability), so that those that reach a prespecified discrimination threshold are classified as true edges (or true nonedges). This threshold is an input of the procedure, and usually the quality of the classification is assessed by integrating the true positive rate versus the false positive rate [also known as the receiver operating characteristic (ROC) curve] for all discrimination threshold values. This integration yields the area under the curve (AUC), which lies in the unit interval, and can be equivalently interpreted as the probability that a randomly selected true positive will be ranked above a randomly chosen true negative. Thus, a value of $1/2$ indicates a performance equivalent to a random guess, and a value of 1 indicates "perfect" classification (note that a classifier with an AUC value of 1 still requires the correct discrimination threshold as an input to fully recover the network).

In contrast, the reconstruction task considered here yields a full posterior distribution $P(\boldsymbol{A}|\boldsymbol{n},\boldsymbol{x})$ for the inferred network $\boldsymbol{A}$. Although this distribution can be used to perform the same binary classification task, by using the posterior marginal probabilities π_{ij} as the aforementioned "scores," it contains substantially more information. For example, the number of missing and spurious edges (and, hence, the inferred probabilities p and q) are contained in this distribution and thus do not need to be prespecified. Indeed, our method lacks any kind of *ad hoc* input, such as a discrimination threshold [note that the threshold $1/2$ in the MMP estimator of Eq. (44) is a derived optimum, not an input]. This property means that absolute assessments such as the similarity of Eq. (45) can be computed instead of relative ones such as the AUC.

Furthermore, the reconstruction approach can be used to recover summary quantities and perform error estimates, which is usually not directly possible in the binary classifier framing. In addition, reconstructed networks can contain spurious and missing edges simultaneously, whereas with traditional edge-prediction methods, they each require their own binary classification (with their own discrimination thresholds).

When doing edge prediction, one often distinguishes recovering from the effects of noise (i.e., an edge has been transformed into a nonedge, or vice versa)—to which we refer as *denoising*—and from a lack of observation (i.e., a given entry in the adjacency matrix is unknown)—to which we refer as *completion*. In each scenario, the scores are computed differently, yielding different classifiers. When performing reconstruction with our method, we inherently allow for any arbitrary combination of denoising and completion: If an entry is not observed, it has a value of $n_{ij} = 0$, which is different from it being observed with $n_{ij} > 0$ as a nonedge $x_{ij} = 0$. If the noise parameters p and q are zero, recovery via the posterior distribution amounts to a pure completion task for the entries with $n_{ij} = 0$, and likewise we have a pure denoising task if $n_{ij} > 0$ for every pair (i, j); otherwise, we have a mixture of these two tasks.

In Fig. 7, we illustrate some of these tasks, performed using our framework for the openflights data set, which we find to be representative of the majority investigated. In Figs. 7(a) and 7(b) are shown the results for edge ($q = 0$) and nonedge ($p = 0$) denoising, respectively. Given that this network is sparse, the probability of an edge is on average much smaller than that of a nonedge, which means that the edge denoising task is significantly harder than nonedge denoising, for which very high accuracy can be obtained even for $n = 1$ measurement per edge. Nevertheless, positive reconstruction is possible in each case, approaching a similarity of 1 as the number of measurements is increased.

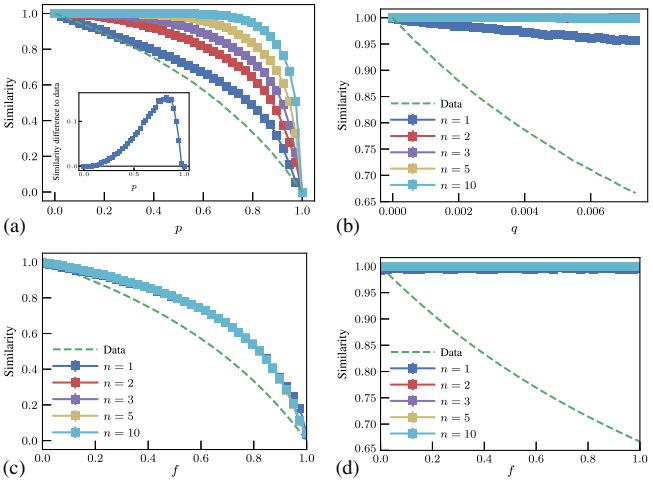

FIG. 7. (a) Edge denoising reconstruction performance for the openflights data, as a function of the missing edge probability p, for various n, and $q = 0$. The dashed curve shows the corresponding value obtained directly with the measured data with $n = 1$, and the inset shows the difference between the curve for $n = 1$ and the dashed curve. (b) The same as (a) but for nonedge denoising, with $p = 0$. The values of q are chosen to yield the same number of affected nonedges as edges in (a). (c) Edge completion reconstruction performance as a function of the fraction f of unobserved edges. The dashed line shows the value of similarity obtained by considering the unobserved edges as nonedges. (d) The same as (c) but for nonedge completion, as a function of the fraction f of unobserved nonedges. The dashed line shows the value of similarity obtained by considering the unobserved nonedges as edges.

We also perform network completion by choosing a fraction f of edges or nonedges, for which zero measurements are performed, $n_{ij} = 0$, while the remaining entries are observed n times, $n_{ij} = n$. In Figs. 7(c) and 7(d) are shown the reconstruction results for edge and nonedge completion, respectively. Like for denoising, nonedge completion is easier, approaching near perfection for the entire range of parameters, and for the same reason as before. For the completion tasks, however, the number of observations n for the nonaffected entries has a negligible effect in the reconstruction, and we observe near-optimal performance already for $n = 1$.

Although the number of edges and nonedges affected is the same for both our denoising and completion examples, the latter yields a larger rate of successful reconstruction for both edges and nonedges. This result is understood by noting that these tasks have a different number of unknowns. In the case of edge completion, on the one hand, for a given finite fraction f of nonobserved edges, we have $O(E)$ unknowns, which for sparse networks is $O(N)$. For edge denoising, on the other hand, for any fraction p of missing edges, for sparse networks we have, in principle, $O(N^2)$ possibilities for their placements, corresponding to all observed nonedges. For nonedge denoising and completion, the difficulty is comparable: For any fraction $f = O(1/N)$ left unobserved or $q = O(1/N)$ transformed into spurious edges, there are $O(N)$ unknowns, if the network is sparse. However, the actual number of unknowns for nonedge completion is strictly smaller, as it must involve only the fraction not observed, whereas for denoising it involves every observed edge.

This difference in performance shows how the correct interpretation of the data can be crucial—as absence of evidence is not evidence of absence. Unfortunately, most available data sets fail to make this distinction, including those few which actually provide some amount of error assessments, as they do not indicate which pairs of nodes have not been measured at all.

4. Detectability of modular structures

Our approach generalizes the task of community detection for networks with measurement errors. However, even in the case of error-free networks with planted community structure, this task is not always realizable. This situation is most often illustrated with a simple SBM parametrization known as the planted partition (PP) model:

$$\omega_{rs} = \omega_{\text{in}}\delta_{rs} + \omega_{\text{out}}(1 - \delta_{rs}), \tag{54}$$

with equal-sized groups, $n_r = N/B$. As shown in Ref. [38], the detection of communities from networks sampled from this model undergoes as a phase transition and becomes impossible for parameter values satisfying

$$N|\omega_{\text{in}} - \omega_{\text{out}}| < B\sqrt{\langle k\rangle}, \tag{55}$$

where $\langle k\rangle = N[\omega_{\text{in}} + (B-1)\omega_{\text{out}}]/B$ is the average degree of the network. This transition means that, even though a PP model may contain assortative community structure with $\omega_{\text{in}} > \omega_{\text{out}}$, the individual samples from the generative model are indistinguishable from a fully random graph if the inequality of Eq. (55) is fulfilled and, hence, contain no information useful for the recovery of the planted communities.

When considering measured networks, it is expected that the introduced errors make the detection task more difficult, as the noise removes information from the data. As we see in Sec. II A 1, when a single measurement of a SBM network is made with noise parameters p and q, it becomes indistinguishable from a SBM sample with effective probabilities $\boldsymbol{\omega}'$, given by Eq. (36). Applying this fact to the PP model yields a transition according to

$$N|\omega_{\text{in}} - \omega_{\text{out}}| < \frac{B\sqrt{(1-p-q)\langle k\rangle + qN}}{(1-p-q)}. \tag{56}$$

For positive error magnitudes $p > 0$ or $q > 0$, the above threshold is larger than Eq. (55). This result highlights how measurement noise can hinder the detection of large-scale structures if they are sufficiently weak and induce a phase transition in their detection. It also means that the reconstructions of the networks themselves are affected by the same transition, as our approach hinges on the detectability of these large-scale structures.

In Fig. 8 are shown the reconstruction results for PP network samples with $B = 2$ groups, for simulated measurements always using $q = 0$, but with either $p = 0$ or $p = 1/2$. Without measurement noise, $p = 0$, the detectability of the planted partition is possible all the way down to the detectability threshold of Eq. (55). Despite the lack of noise, the similarity with the true network is only slightly above 0.6 in the detectable region, which is because the

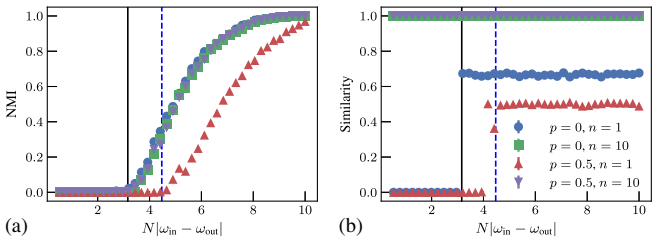

FIG. 8. (a) Normalized mutual information (NMI) between planted and inferred partitions for a PP model with $N = 10^4$, $B = 2$, $\langle k\rangle = 10$, and measurement errors $q = 0$ and p given in the legend, together with the number of measurements n. The black solid line marks the threshold of Eq. (55), and the blue dashed line the threshold of Eq. (56) with $(p, q) = (1/2, 0)$. (b) The same as in (a), but for the similarity $S(\hat{\boldsymbol{A}}, \boldsymbol{A}^*)$ between the inferred and true networks.

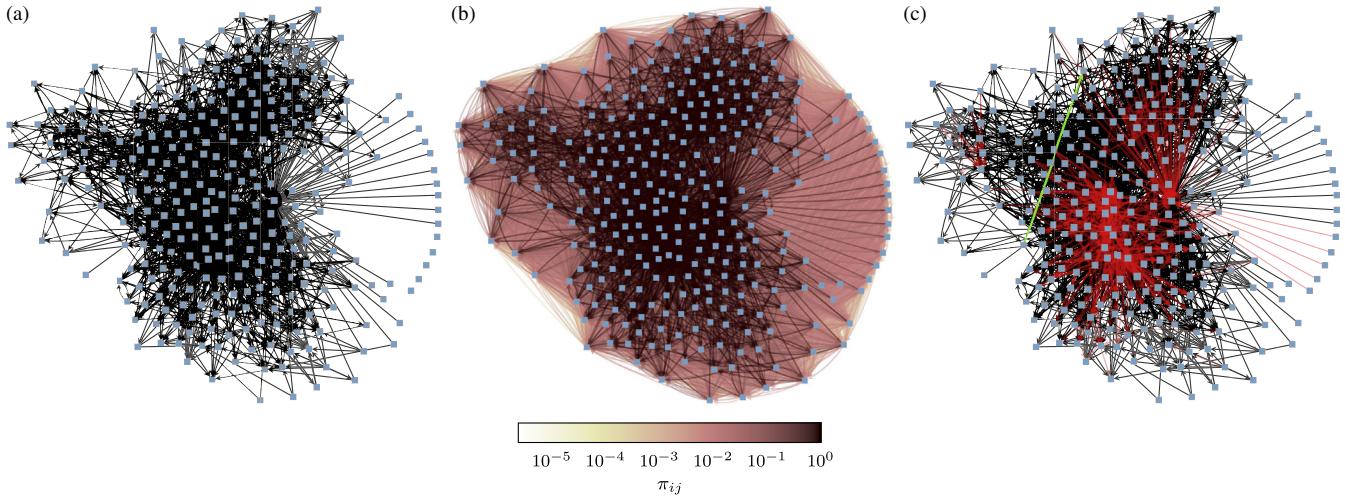

FIG. 9. (a) Measured neural network of the *C. elegans* worm [39]. (b) Marginal posterior distribution π_{ij} of the edges according to our reconstruction method, shown as edge colors. (c) MMP estimate of the network, with inferred missing edges shown in red and spurious edges shown in green.

probabilities in this ensemble are not sufficiently heterogeneous to rule out high noise values, as some of the empirical networks we consider. Below the transition, the similarity falls to zero, as the network becomes indistinguishable from a fully random one. Interestingly, this partial uncertainty about the network does not affect the inference of the node partition. If we increase the noise to $p = 1/2$, the partition recovery is possible up to the threshold of Eq. (56) when only $n = 1$ measurements are made. However, after sufficiently increasing n, the effects of noise are diminished, and the original threshold can be achieved. In this case, the similarity also becomes high even below the detectability threshold, where the community structure itself cannot be recovered, which is because the repeated measurements themselves yield sufficient information about the network structure, and the reconstruction no longer needs to rely on the network structure itself.

D. Reconstruction of empirical data and uncertainty assessment

A central advantage of our method is that it can be used to reconstruct noisy networks when only a single measurement has been made for each entry in the adjacency matrix and no error assessment is known. As the majority of network data can be cast into this framework, our method can be used to reconstruct them and give uncertainty assessments for quantities of interest. In this section, we discuss a few empirical examples.

We focus first on the neural network of the *Caenorhabditis elegans* worm. It is used extensively as a model organism, and it had its full neural network mapped in 1986 by White *et al.* [39]. The network measurement is done by electron microscopy of transverse serial sections of the animal's body of about 50 nm thickness, amounting to around 8000 images. Based on these images, the network is reconstructed by painstaking manual tracing of the neuron paths across the different

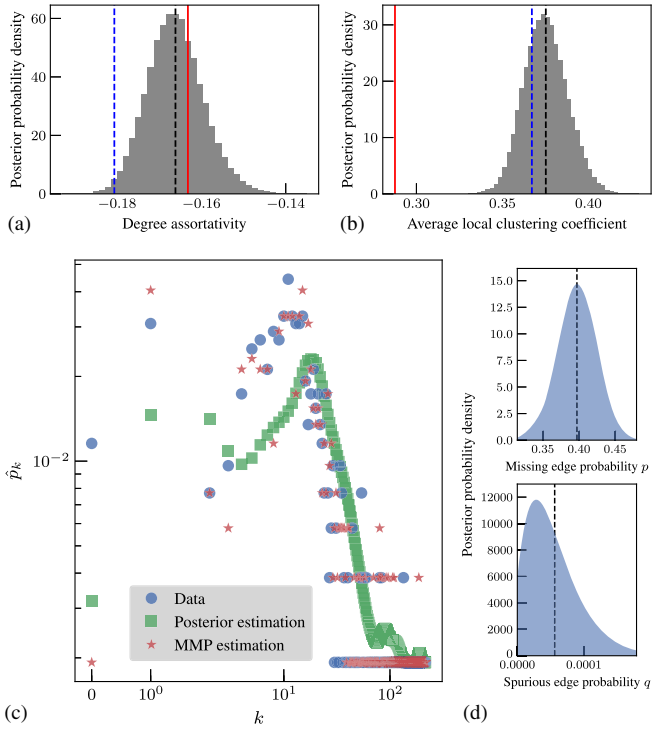

FIG. 10. Reconstruction statistics for the neural network of *C. elegans*. (a) Posterior distribution of the degree assortativity coefficient. The black dashed line marks the mean of the distribution, and the blue dashed line the value obtained for the MMP estimate $\hat{\boldsymbol{A}}$. The red solid line marks the value computed directly from the data. (b) The same as (a) but for the average local clustering coefficient. (c) Measured and estimated degree distributions. (d) Posterior distributions for the error probabilities p and q.

TABLE I. Reconstruction results for empirical networks with single measurements per edge and no available primary error assessments. Similarity refers to the average of $S(\boldsymbol{A},\boldsymbol{A}^*)$ over the posterior distribution. For each quantity (number of edges, degree assortativity, average local clustering) is shown the value directly obtained from the data (direct) and the average over the posterior distribution (estimated). The value of B_e is the posterior average of the effective number of inferred communities $e^{H(\boldsymbol{n})}$, with $H(\boldsymbol{n}) = -\sum_r (n_r/N)\ln(n_r/N)$, where n_r is the number of nodes in group r, being the entropy of the group size distribution. The values $\hat{p}$ and $\hat{q}$ are the posterior averages of the error rates. In all cases, the parentheses indicate the standard deviation over the posterior distribution. Data set descriptions are given in Appendix E.

			Edges		Degree assortativity		Local clustering				
Data set	Similarity	Nodes	Direct	Estimated	Direct	Estimated	Direct	Estimated	B_e	$\hat{p}$	$\hat{q}$
Karate club	0.94(4)	34	78	77(7)	−0.475 61	−0.49(5)	0.570 64	0.58(5)	2.7(6)	0.06(5)	0.012(10)
9/11 terrorists	0.96(2)	62	152	154(8)	−0.080 48	−0.096(20)	0.486 37	0.50(2)	5.4(5)	0.05(4)	0.003(2)
American football	0.857(16)	115	613	500(18)	0.162 44	0.18(7)	0.403 22	0.68(4)	12.7(3)	0.05(3)	0.0226(19)
Network scientists	0.9981(17)	379	914	915(3)	−0.081 68	−0.0823(18)	0.741 23	0.741(13)	29.6(14)	0.004(3)	$3.1(19)\times10^{-5}$
C. elegans neural	0.744(19)	302	2345	3950(160)	−0.163 20	−0.167(7)	0.287 52	0.378(12)	17.0(3)	0.41(2)	$6(3)\times10^{-5}$
Malaria genes	0.9981(15)	1103	2965	2973(9)	−0.300 13	−0.2997(20)	0	0(0)	30.8(3)	0.004(3)	$4(3)\times10^{-6}$
Power grid	0.80(7)	4941	6594	9900(1300)	0.003 46	0.043(17)	0.080 10	0.058(7)	15.6(7)	0.33(10)	$2.5(19)\times10^{-7}$
Political blogs	0.965(5)	1222	16 714	17 860(190)	−0.221 33	−0.2226(16)	0.320 25	0.343(5)	16.6(3)	0.066(10)	$4.4(17)\times10^{-5}$
DBLP citations	0.64(1)	12 590	49 744	106 000(2000)	−0.045 72	−0.0559(19)	0.117 18	0.164(7)	86.4(20)	0.529(11)	$9(5)\times10^{-9}$
Openflights	0.9916(9)	3286	39 430	40 100(70)	−0.005 31	−0.0071(11)	0.496 47	0.507(2)	117.1(5)	0.0167(18)	$1.0(3)\times10^{-7}$
Reactome	0.999 977(10)	6327	146 160	146 164(3)	0.244 87	0.244 87(4)	0.588 38	0.5887(3)	318.7(10)	$4.1(18)\times10^{-5}$	$1.3(8)\times10^{-7}$
cond-mat	0.999 986(13)	40 421	175 693	175 695(4)	0.186 33	0.186 33(2)	0.636 16	0.636 15(3)	1014(6)	$3(2)\times10^{-5}$	$3(2)\times10^{-9}$
Enron email	0.999 86(5)	36 692	183 831	183 885(18)	−0.110 76	−0.110 75(2)	0.496 98	0.496 92(8)	188.9(11)	0.000 28(10)	$2.9(19)\times10^{-9}$
Linux source	0.9973(3)	30 837	213 424	214 600(120)	−0.174 68	−0.174 67(7)	0.128 49	0.1322(10)	351.2(7)	0.0055(5)	$1.7(10)\times10^{-9}$
Brightkite	0.9985(3)	58 228	214 078	214 740(80)	0.010 82	0.011 00(11)	0.172 33	0.172 34(10)	151(3)	0.0029(5)	$1.7(12)\times10^{-8}$
PGP	0.997 99(9)	39 796	301 498	301 660(60)	0.000 76	0.000 49(8)	0.461 09	0.4617(2)	929(2)	0.002 27(16)	$3.35(18)\times10^{-7}$
Internet AS	0.999 67(13)	53 387	496 731	497 070(130)	−0.186 97	−0.186 959(17)	0.680 97	0.681 26(14)	218(16)	0.0007(3)	$1.0(8)\times10^{-9}$
Web Stanford	0.999 998 7(8)	281 903	2 312 497	2 312 494(4)	−0.112 44	−0.112 444 7(2)	0.597 63	0.597 634(3)	4168(2)	$1.0(2)\times10^{-6}$	$7(5)\times10^{-11}$
Flickr	0.999 976(13)	105 938	2 316 948	2 316 830(60)	0.246 85	0.246 823(16)	0.089 13	0.089 138(7)	617(2)	$6(3)\times10^{-7}$	$2.0(11)\times10^{-8}$

images. The reliability of the reconstruction procedure is discussed in Ref. [39], where human error in tracing the neuron bundles, the orientation of the neurons with respect to the transverse section, and poor image quality are identified as the main sources of potential errors. White *et al.* employ a series of error-mitigating procedures, such as detecting basic connection inconsistencies, exploiting the partial bilateral symmetry for suspect connections, and comparing with independent reconstructions of parts of the network. Although the authors of that work profess to be "reasonably confident" that the structure they present is "substantially correct," they do not exclude the possibility of remaining errors, nor do they quantify in any way the uncertainty of their measurements. Furthermore, the data commonly used for network analysis, which we also use here, are manually compiled by Watts and Strogatz [40], based on the original data of Ref. [39], and may contain further errors. The resulting data we use amount to $N = 302$ nodes and $E = 2345$ directed edges (note that five nodes are excluded in Ref. [40] for not having any connections; we include these nodes in our analysis, as it is suspicious that isolated neurons can exist and thus is probably a symptom of missing data).

When we employ our reconstruction procedure on the *C. elegans* data, we find the results shown in Figs. 9 and 10 and summarized in Table I. The MMP estimate of this network contains $\hat{E} = 2773$ edges, but the posterior distribution is significantly broad and contains on average $\langle E \rangle = 3950$ edges, meaning that there are many potential edges with low but non-negligible probabilities. We note that our reconstruction connects the isolated nodes in the data to the main hub in the network, which is an important neuron situated in the head of the worm. As seen in Fig. 10(a), the inferred degree assortativity coefficient is compatible with the value measured directly from data, and our method is capable of providing a confidence interval for this estimation. The same is not true for the average local clustering coefficient, as seen in Fig. 10(b), which is not compatible with the value measured directly from the data with any reasonable confidence.

For the *C. elegans* data, the inferred error rates are $(\hat{p}, \hat{q}) = (0.4, 6 \times 10^{-5})$. Although this result corresponds to a very high accuracy with respect to spurious edges, it indicates a low accuracy with respect to missing edges, and it implies that almost half of the original edges are misrepresented as nonedges. Although the consensus of the posterior distribution (represented by the MMP estimate) is reasonably close to the original data, with a similarity of 0.93, the similarity averaged over the posterior distribution is only 0.74, indicating a fair amount of uncertainty. This result seems to contradict the qualitative assessment of Ref. [39], which argues in favor of the reliability of their data. This discrepancy can be interpreted in two ways: (i) The assessment in Ref. [39] is too optimistic, and the data contain indeed more errors than anticipated. (ii) The data actually contain fewer errors than our method predicts, but the true network itself is not sufficiently structured to *rule out* errors in a manner that can be exploited by our method. However, even if case (ii) happens to be true, our method correctly projects an agnostic prior assumption about the error rates onto the posterior distribution, after being informed by the data. It then follows that more confidence in the data and in the existence of fewer errors must be accompanied by either more data (e.g., repeated measurements) or a more refined prior information on the error rates, obtained either by calibration or a quantitative study of the methods employed in Ref. [39]. As an illustration, in Fig. 11 is shown the posterior similarity with the data obtained with different choices of the hyperparameter β, using $\alpha = 1$, which control the prior knowledge on the value of p, with an average given by $\langle p \rangle = \alpha/(\alpha + \beta)$. A high accuracy of the data, with inferred similarities approaching one, is achieved only by a prior belief on p being on the order of 0.01 or smaller. This result means that one should trust the claimed high accuracy in Ref. [39] only if one is confident that the probability of an edge not being recognized as such was below one percent, which might very well be true but would need to be substantiated with further evidence. Although in situations such as these our method cannot fully resolve the discrepancy without further data, it serves as the appropriate framework in which to place the issue and shows that any analysis that takes the original measured data for granted, ignoring potential errors, inherently assumes more reliability than can be inferred from the data alone.

For other kinds of data, it is possible to obtain very accurate reconstructions with single measurements. As an example, we consider the network of collaborations in papers published in the cond-mat section of the arXiv.org preprint Web site in the period spanning from January 1,

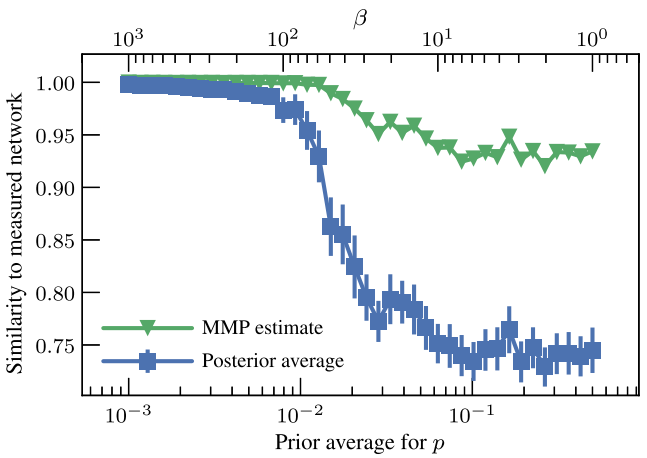

FIG. 11. Average similarity between the posterior samples and the measured *C. elegans* data as a function of the hyperparameter β (with $\alpha = 1$), which controls the prior belief on the probability p of missing edges (the average of which is shown in the x axis). For reference, the similarity for the MMP estimate is also shown.

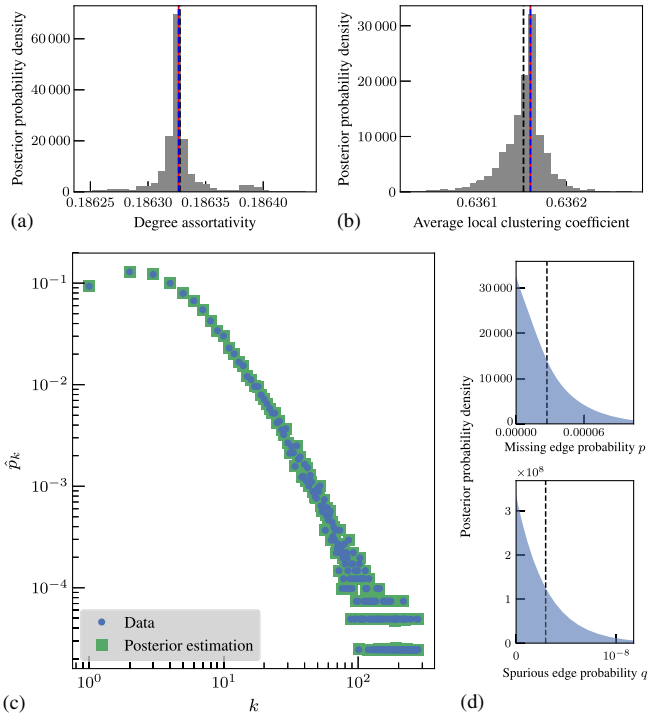

FIG. 12. Reconstruction statistics for the coauthorship network of arXiv.org. (a) Posterior distribution of the degree assortativity coefficient. The black dashed line marks the mean of the distribution, and the blue dashed line the value obtained for the MMP estimate $\hat{\boldsymbol{A}}$. The red solid line marks the value computed directly from the data. (b) The same as (a) but for the average local clustering coefficient. (c) Measured and estimated degree distributions. (d) Posterior distributions for the error probabilities p and q.

1995, to March 31, 2005, where authors are nodes and an edge exists if two authors published a paper together [41]. This network is compiled by crawling through the Web-site interface and could contain errors due to incorrect parsing [42]. When reconstructed using our method, however, we find that it is remarkably accurate, with very low error rates inferred as $(p, q) = (3 \times 10^{-5}, 3 \times 10^{-9})$. As can be seen in Fig. 12, all inferred properties match very closely the direct measurement—although our reconstruction is still useful in providing error estimates for them.

In Table I, we provide a summary of reconstruction results with our method to several empirical networks. We observe a tendency of larger networks to be more accurate than smaller ones. This is not a trivial result of there being more data but rather of these larger networks containing stronger structures which are informative of low measurement noise. If these networks were fully random, their reconstruction accuracy would have been very poor, regardless of their size.

E. Heterogeneous errors

So far, we consider only the situation where the error probabilities p and q are the same for every pair of nodes in the network. Although it is easy to imagine a simplified scenario where the same measurement instrument is used in every case, it is also easy to imagine situations where this is not an adequate representation of how a measurement is made. For example, in the case of the *C. elegans* neural network, the spatial proximity of the neurons may make it harder or easier to measure the edges and nonedges, thus impacting their error probabilities.

With this in mind, it is easy to generalize our framework to allow for individual error probabilities p_{ij} and q_{ij}, for missing and spurious edges between nodes i and j, respectively. Given a true underlying entry A_{ij} between these two nodes, its measurement probability is given by

$$P(x_{ij}|n_{ij}, A_{ij}, p_{ij}, q_{ij}) = \binom{n_{ij}}{x_{ij}} [(1-p_{ij})^{x_{ij}} p_{ij}^{n_{ij}-x_{ij}}]^{A_{ij}} [q_{ij}^{x_{ij}} (1-q_{ij})^{n_{ij}-x_{ij}}]^{1-A_{ij}}. \tag{57}$$

Using the same Beta priors as before, we can integrate over p_{ij} and q_{ij}, obtaining

$$\begin{aligned} P(x_{ij}|n_{ij},A_{ij},\alpha,\beta,\mu,\nu) &= \int P(x_{ij}|n_{ij},A_{ij},p_{ij},q_{ij}) P(p_{ij}|\alpha,\beta) P(q_{ij}|\mu,\nu) dp_{ij} dq_{ij} \\ &= \binom{n_{ij}}{x_{ij}} \left[\frac{\mathcal{B}(n_{ij}-x_{ij}+\alpha, x_{ij}+\beta)}{\mathcal{B}(\alpha,\beta)}\right]^{A_{ij}} \left[\frac{\mathcal{B}(x_{ij}+\mu, n_{ij}-x_{ij}+\nu)}{\mathcal{B}(\mu,\nu)}\right]^{1-A_{ij}}. \end{aligned} \tag{58}$$

With this result, we have the full conditional distribution for the measured network,

$$P(\boldsymbol{x}|\boldsymbol{n}, \boldsymbol{A}, \alpha, \beta, \mu, \nu) = \prod_{i<j} P(x_{ij}|n_{ij}, A_{ij}, \alpha, \beta, \mu, \nu), \tag{59}$$

with which we can obtain the posterior distribution of Eq. (3). However, unlike the case with uniform errors, the choice of hyperparameters is now vital. The noninformative assumption $\alpha = \beta = \mu = \nu = 1$ applied above makes the likelihood *independent* of the planted network $\boldsymbol{A}$, rendering the data completely uninformative as well, which means we must have some global information that specifies how the values of p_{ij} and q_{ij} are distributed. Although we could simply set (or fit) the values of the hyperparameters to values

different from one, we favor a nonparametric approach, and we include the hyperparameters in the posterior distribution,

$$P(\boldsymbol{A},\boldsymbol{b},\alpha,\beta,\mu,\nu|\boldsymbol{n},\boldsymbol{x}) = \frac{P(\boldsymbol{x}|\boldsymbol{n},\boldsymbol{A},\alpha,\beta,\mu,\nu)P(\boldsymbol{A}|\boldsymbol{b})P(\boldsymbol{b})P(\alpha,\beta,\mu,\nu)}{P(\boldsymbol{x}|\boldsymbol{n})}, \quad (60)$$

which requires their own hyperprior distribution $P(\alpha,\beta,\mu,\nu)$. Here, we are agnostic and use a constant prior $P(\alpha,\beta,\mu,\nu) \propto 1$, with an unspecified and unnecessary normalization constant, as it cancels out in the posterior distribution. [43] The inference algorithm is the same as before, but, in addition to move proposals for the network $\boldsymbol{A}$ and node partition $\boldsymbol{b}$, we make also move proposals for the hyperparameters.

Like in the uniform case, we can obtain the posterior distribution for the error probabilities via their conditional posteriors, i.e.,

$$P(p_{ij}|n_{ij},x_{ij},A_{ij},\alpha,\beta) = \frac{p^{A_{ij}(n_{ij}-x_{ij})+\alpha-1}(1-p)^{x_{ij}A_{ij}+\beta-1}}{\mathcal{B}(A_{ij}(n_{ij}-x_{ij})+\alpha, x_{ij}A_{ij}+\beta)} \quad (61)$$

and likewise for q_{ij} with

$$P(q_{ij}|n_{ij},x_{ij},A_{ij},\mu,\nu) = \frac{q^{(1-A_{ij})x_{ij}+\mu-1}(1-q)^{(1-A_{ij})(n_{ij}-x_{ij})+\nu-1}}{\mathcal{B}((1-A_{ij})x_{ij}+\mu,(1-A_{ij})(n_{ij}-x_{ij})+\nu)}, \quad (62)$$

averaged over the posterior distribution.

We note that, for heterogeneous error rates, the case with single measurements $n_{ij} = 1$ becomes less interesting. If we replace $n_{ij} = 1$ and $x_{ij} \in \{0,1\}$ in the above equations, they become identical to Eq. (15) for the case with uniform errors, if we make the substitution

$$p = \frac{\mathcal{B}(\alpha+1,\beta)}{\mathcal{B}(\alpha,\beta)} = \frac{\alpha}{\alpha+\beta}, \quad (63)$$

$$q = \frac{\mathcal{B}(\mu+1,\nu)}{\mathcal{B}(\mu,\nu)} = \frac{\mu}{\mu+\nu}. \quad (64)$$

In this situation, only the prior averages of p_{ij} and q_{ij} matter, not their variance. A uniform prior for α, β, μ, and ν is equivalent to Beta priors with parameters (1,0) for p and q computed via the equation above, [44] and hence this approach becomes completely identical to the one with uniform errors considered before. Therefore, there are no sufficient data in the single measurement case to detect heterogeneous errors of this kind, and thus a meaningful use of this method is confined to data with $n_{ij} > 1$. Note also that this equivalence implies that any error heterogeneity present in the data will be conflated with underlying network structure when single measurements are made. Ultimately, this conflation can only be resolved by making multiple measurements.

We consider two data sets which contain multiple measurements, in order to compare both approaches. We consider the reality mining data set, which records proximity interactions between voluntary students over time [45]. Following

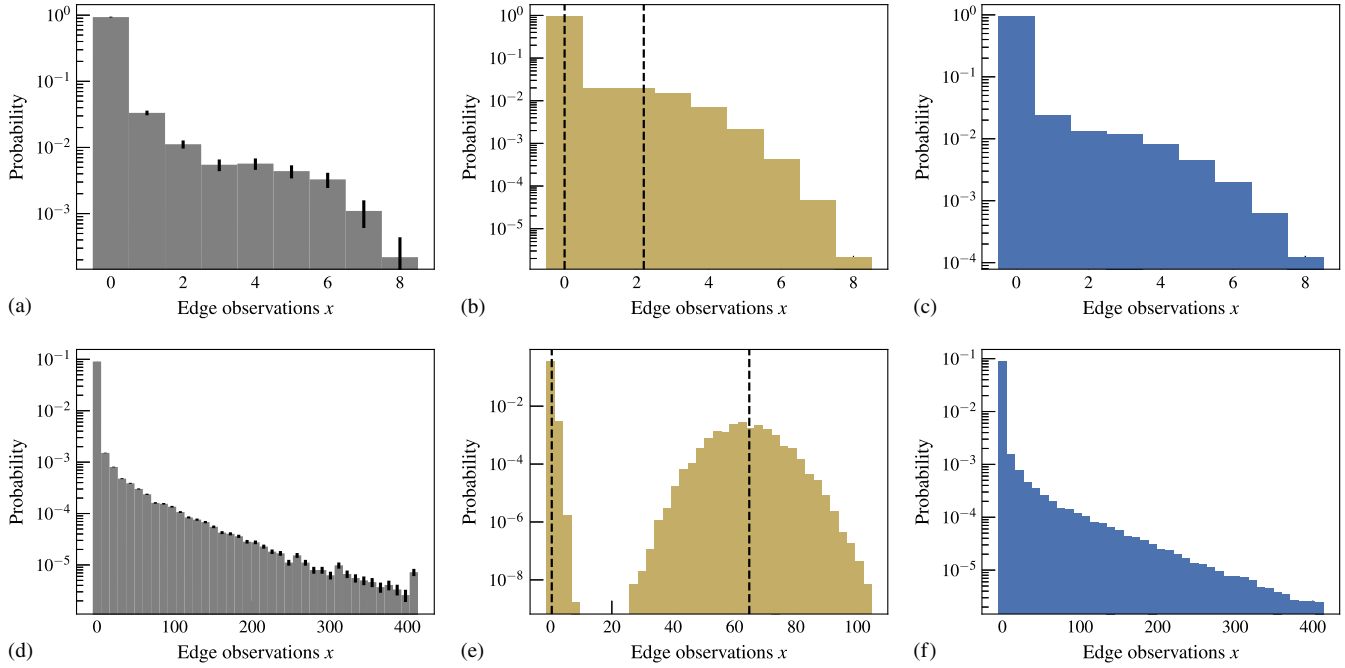

FIG. 13. Distribution of edge occurrences, x_{ij}, for the reality mining (top row) and human connectome (bottom row) data sets. (a),(d) Empirical data. (b),(e) Generated from inferred parameters, according to the uniform model. (c),(f) Generated from inferred parameters, according to the nonuniform model.

TABLE II. Reconstruction results for empirical networks with multiple measurements per edge. For each quantity is shown the value obtained using either the uniform or the nonuniform model, as indicated. The value of $B_e = e^{H(\boldsymbol{n})}$ is the effective number of inferred communities, computed as $H(\boldsymbol{n}) = -\sum_r (n_r/N)\ln(n_r/N)$, where n_r is the number of nodes in group r. The values $\hat{p}$ and $\hat{q}$ are the posterior averages of the error rates. In all cases, the parentheses indicate the standard deviation over the posterior distribution. Data set descriptions are given in Appendix E.

			Edges		Degree assortativity		Local clustering		B_e		$\hat{p}$		$\hat{q}$	
Data set	n	Nodes	Uniform	Nonuniform	Uniform	Nonuniform	Uniform	Nonuniform	Uniform	Nonuniform	Uniform	Nonuniform	Uniform	Nonuniform
Karate club	2	34	77.9(3)	95(6)	−0.475(3)	−0.43(5)	0.569(8)	0.63(5)	2.9(6)	2.9(6)	0.012(2)	0.49(3)	0.0011(3)	0.0004(13)
Reality mining	8	96	293(11)	280(20)	−0.23(3)	−0.23(3)	0.31(2)	0.29(2)	3.5(6)	3.4(6)	0.724(8)	0.71(3)	0.0007(2)	0.001(2)
School friends	6	2539	12 500(40)	8200(300)	0.258(4)	0.322(6)	0.1535(13)	0.188(3)	82.5(3)	80.2(3)	0.5064(11)	0.16(3)	$1.8(7)\times10^{-5}$	0.0002(9)
Human connectome	418	1015	23 020(16)	62 000(6000)	0.0008(5)	0.002(3)	0.6796(4)	0.68(7)	100.5(11)	51.26(19)	0.845 03(8)	0.93(9)	0.000 984 6(10)	$1(11)\times10^{-4}$

Ref. [10], as measurements we consider the state of the network during eight consecutive Wednesdays in March and April of 2005, so chosen to avoid weekly periodic events. In addition, we consider the human connectome, using data from the Budapest Reference Connectome [46] (which itself is based on primary data from the Human Connectome Project [47]). This data set contain records of the neuronal connections of 418 individuals, each of which we consider as a separate measurement.

For both data sets considered—as it is arguably always true whenever multiple network measurements are made—it is debatable whether there is really a true single network behind the measurements, as our method assumes. For example, in the reality mining data set, the underlying network could be changing over time, and the connectome can vary between individuals for physiological reasons rather than measurement error. In each case, however, we are free to keep the mathematical structure of our model in place and change its interpretation. We could, for instance, assume that the single network being inferred amounts simply to a consensus or a blueprint of the network, and the "error" rates p_{ij} and q_{ij} indicate the variability of each single edge or nonedge around this blueprint. Since both scenarios are generally conflated when making this kind of measurement, we can choose the interpretation that is most suitable according to the context.

In Figs. 13(a) and 13(d) are shown the distributions of the measured frequencies of edge occurrences, x_{ij}, for both data sets. For the human connectome, we observe a very broad distribution, with occurrences present in the entire possible range. In Figs. 13(b) and 13(e), we see the simulated results by sampling parameters from the posterior distribution and generating new data from them, using in this case the model with uniform errors. Whereas the results for reality mining are reasonably close to the data, the results for the human connectome show an obvious discrepancy, where the generated data are concentrated around two modes, corresponding to the frequencies of edges and nonedges. Indeed, for the uniform model, this separation is guaranteed to occur for any given $p \neq 1/2$ and $q \neq 1/2$ and a sufficiently large number of measurements. The fact that this separation is not observed in the data is a clear indication that the error rates are not uniform (or alternatively, but mathematically equivalently, that there is no single network behind the measurements). Indeed, when using the nonuniform model, it recovers the observed frequency almost perfectly, as seen in Figs. 13(c) and 13(f).

If we look more closely at the human connectome data, we see that both approaches give us different pictures of the underlying network structure. As is summarized in Table II, the uniform model yields a sparser network, which nevertheless seems more finely structured, with close to 100 effective groups detected. Conversely, the nonuniform model yields a denser network, with a more uniform structure and only half as many identified groups. In Fig. 14, we see more

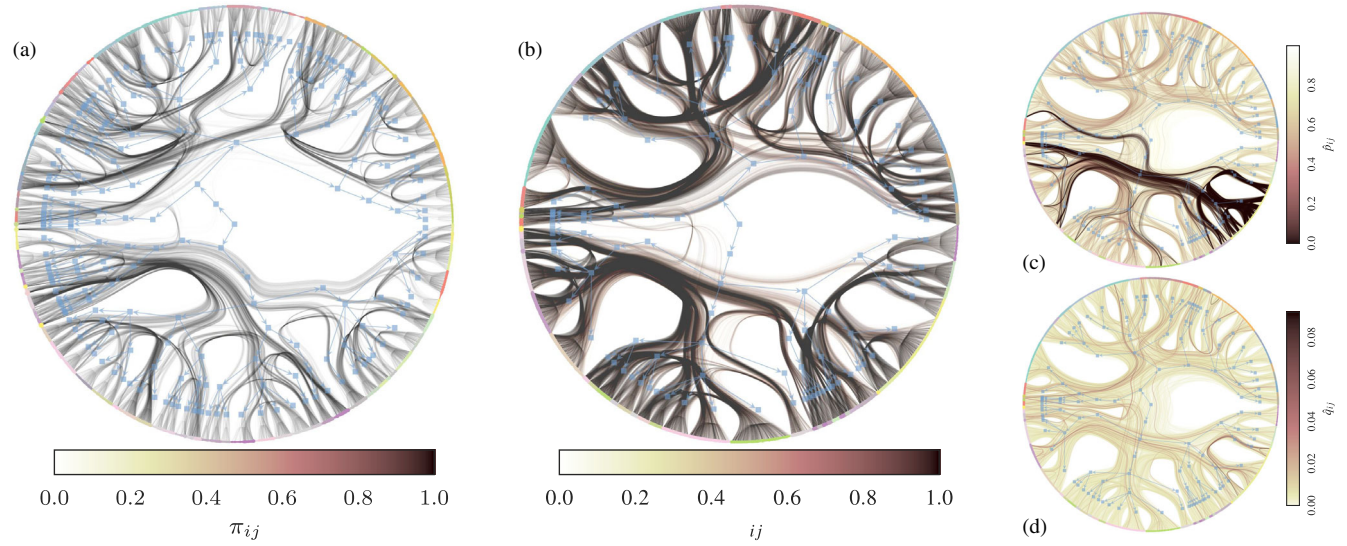

FIG. 14. Reconstruction results for the human connectome. (a) Marginal posterior distribution of edges π_{ij} and inferred hierarchical partition, according to the model with uniform errors. The upper hierarchy branch corresponds to the right hemisphere. (b) The same as (a) but with the nonuniform model. (c) Inferred missing edge probabilities p_{ij} for the nonuniform model. (d) The same as (c) but for the spurious edge probabilities q_{ij}.

clearly the differences between both results. Both are capable of uncovering the hemispherical divisions and the partial bilateral symmetry of the connectome. The nonuniform model can detect a larger number of edges, but it yields larger probabilities of missing edges p_{ij} which are heterogeneously distributed. In Fig. 14(c), it can be seen that the inferred p_{ij} are strongly correlated with the detected group structure and, in particular, seem to indicate a rather stable set of edges (low p_{ij}) that belong mostly to the left hemisphere. The uniform model, on the other hand, incorporates the variability of edge occurrences in the model itself, subdividing the groups further to accommodate it. Therefore, the nonuniform model gives a more faithful separation between the consensus and the variability around it.

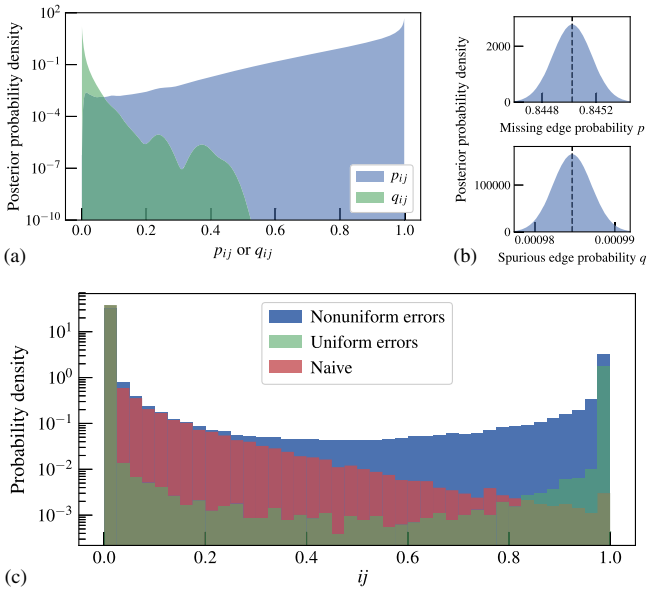

FIG. 15. Inferred uncertainties for the human connectome. (a) Posterior distribution of p_{ij} and q_{ij}, using the nonuniform model. (b) Posterior distribution of p and q, using the uniform model. (c) Distribution of posterior marginal edge probabilities π_{ij}, according to both model variants, as well as the naive estimate $\tilde{\pi}_{ij} = x_{ij}/n_{ij}$.

In Fig. 15, we can see the posterior distributions of p_{ij} and q_{ij} for the nonuniform model, as well p and q for the uniform model, showing how the former is indeed significantly more heterogeneous than the latter. In Fig. 15(c) is also shown the distribution of posterior probabilities π_{ij} for both models, in addition to the naive estimate $\tilde{\pi}_{ij} = x_{ij}/n_{ij}$. This naive estimate is crude, as it does not differentiate between the different sources of error (spurious or missing edge) and does not take into account the observed correlations between the different entries. Indeed, as Fig. 15(c) shows, it leads to very different results, which are not correctly justified, and should be avoided.

III. INCORPORATING EXTRINSIC UNCERTAINTY ESTIMATES

So far, we consider only situations where direct error estimates on the edges originate from repeated measurements. However, there are situations where primary error estimates are made under different formats. Here, we consider the scenario of Ref. [48], where an arbitrary measurement process is made which yields uncertainty assessments for each node pair, $Q_{ij} \in [0, 1]$, interpreted as conditionally independent probabilities, i.e.,

$$P_Q(\boldsymbol{A}|\boldsymbol{Q}) = \prod_{i<j} Q_{ij}^{A_{ij}}(1-Q_{ij})^{1-A_{ij}}. \tag{65}$$

In principle, we could use these probabilities as they are and generate networks and measure their properties from this distribution. But we could also extract from this information the measurement process which it represents and couple it with our reconstruction approach. This procedure gives us the advantage of being able to use the large-scale structure in the data to better inform our estimates of the underlying network.

The distribution $P_Q(\boldsymbol{A}|\boldsymbol{Q})$ implies the following noisy measurement process:

$$P(\boldsymbol{Q}|\boldsymbol{A}) = \frac{P_Q(\boldsymbol{A}|\boldsymbol{Q})P_Q(\boldsymbol{Q})}{P_Q(\boldsymbol{A})}, \tag{66}$$

with normalization constant

$$P_Q(\boldsymbol{A}) = \int P_Q(\boldsymbol{A}|\boldsymbol{Q})P_Q(\boldsymbol{Q})d\boldsymbol{Q}. \tag{67}$$

If we assume the priors on the edge uncertainties are identically distributed and conditionally independent, i.e.,

$$P_Q(\boldsymbol{Q}) = \prod_{i<j} P(Q_{ij}), \tag{68}$$

we have

$$P_Q(\boldsymbol{A}) = \prod_{i<j} \bar{Q}^{A_{ij}}(1-\bar{Q})^{1-A_{ij}}, \tag{69}$$

with $\bar{Q} = \int_0^1 QP(Q)dQ$. Combining these together, we have

$$P(\boldsymbol{Q}|\boldsymbol{A}) = P_Q(\boldsymbol{Q})\prod_{i<j}\left(\frac{Q_{ij}}{\bar{Q}}\right)^{A_{ij}}\left(\frac{1-Q_{ij}}{1-\bar{Q}}\right)^{1-A_{ij}}. \tag{70}$$

The above depends on an unknown prior $P_Q(\boldsymbol{Q})$. Determining it would require us to delve into the details of how this measurement is made, which is unavailable to us if all we know is $P_Q(\boldsymbol{A}|\boldsymbol{Q})$. However, since it is only a multiplicative constant that does not depend on the data or any latent variable, it does not affect the posterior distribution, and thus we do not need to determine it. The single aspect of this distribution that is relevant is its average $\bar{Q}$. By allowing only for a minor violation of the Bayesian ansatz, we can estimate this average directly from the data:

$$\bar{Q} = \frac{\sum_{i<j} Q_{ij}}{\binom{N}{2}}. \tag{71}$$

With this value, we can couple this arbitrary noise generating process with our overall framework by taking $\mathcal{D} = \boldsymbol{Q}$ and obtaining the posterior distribution

$$P(\boldsymbol{A}|\boldsymbol{Q}) = \frac{P(\boldsymbol{Q}|\boldsymbol{A})P(\boldsymbol{A})}{P(\boldsymbol{Q})}, \tag{72}$$

where $P(\boldsymbol{A})$ assumes that the network has been generated by a SBM. Note that $P(\boldsymbol{A}|\boldsymbol{Q}) \neq P_Q(\boldsymbol{A}|\boldsymbol{Q})$, as we are keeping the same noise-generating process but changing our prior assumption about the data. As desired, our prior is structured and is capable of detecting large-scale patterns—latent groups of nodes and their probabilities of connections, as well as node degrees and hierarchical structure—to inform our inference. This procedure also highlights the versatility of our framework, as we are free to replace the measurement model as appropriate.

Although our derivation is somewhat different, Eqs. (65)–(71) above are the same as in Ref. [48]. The resulting posterior of Eq. (72), however, is different, as our approach is nonparametric, and hence can be used to infer the number of groups and does not involve any approximations that rely on the network being sparse or locally treelike.

In Fig. 16, we show the results for the protein-protein interaction network of *Escherichia coli*, for which error estimates in the form of Q_{ij} probabilities are provided [49]. The probabilities are computed in an elaborate manner by combining seven sources of evidence for the existence of an interaction between two proteins. As seen in the figure, our method is able to detect prominent large-scale features that

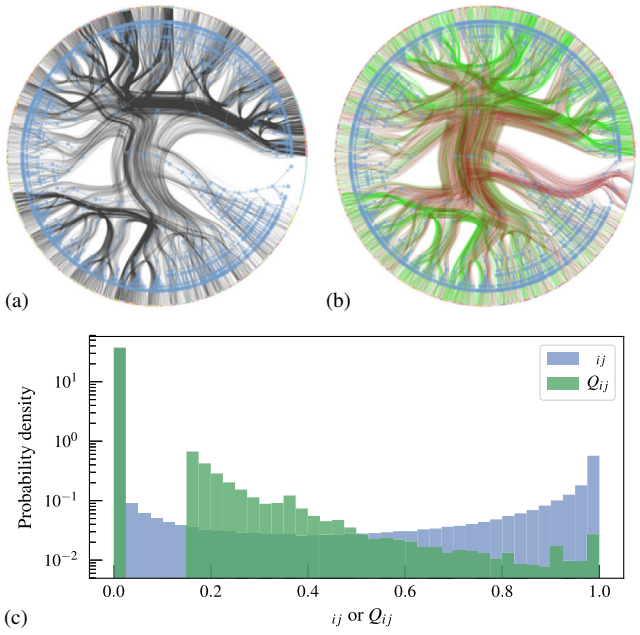

FIG. 16. (a) Inferred *E. coli* protein interaction network, according to uncertain data $\boldsymbol{Q}$, using the MMP estimator from the posterior $P(\boldsymbol{A}|\boldsymbol{Q})$. (b) Difference between (a) and the MMP estimator using the original uncertainties $\boldsymbol{Q}$ directly, via $P_Q(\boldsymbol{A}|\boldsymbol{Q})$ [Eq. (65)]. Green edges are those that are added in (a), and red ones are removed. [The hierarchical partition is the same as in (a) and is shown only as a visual aid.] (c) Distribution of marginal posterior probabilities π_{ij} and original uncertainties Q_{ij}.

help shape the posterior distribution. The resulting posterior probabilities are fairly different from the primary error estimates, showing that these observed correlations can be very informative for the reconstruction process.

IV. CONCLUSION

We have presented a general nonparametric Bayesian network reconstruction framework that couples a noisy measurement model with the SBM as a generative process. The posterior distribution of this joint model yields simultaneously an ensemble of possibilities for the underlying network as well as its large-scale hierarchical modular organization. As we have shown, this joint identification of the network structure enables the existence of correlations in the measured data to inform the network reconstruction. As a consequence, our method can be employed also when a single measurement of the network has been made—which is not possible with methods that do not exploit such correlations—and the error probabilities are unknown. This property makes our approach applicable to the dominating set of network data sets that do not provide primary error estimates of any kind and can extract from them not only the most likely underlying network but also error estimates for arbitrary network properties.

We have shown that our general methodology is versatile, allowing for different noise models. We have considered the situation where the error probabilities are heterogeneous, showing strong evidence for its existence in empirical data, and demonstrated the efficacy of our modified approach in capturing it. We have also shown how extraneous uncertainty estimations obtained with arbitrary methods can be incorporated into our approach, without requiring a detailed model for their generation.

The approach we have proposed is open ended and admits many extensions and generalizations. For example, although the SBM can be used to exploit edge correlations in favor of reconstruction, this can be further improved by considering more realistic models that include other kinds of correlations such as triadic closure [50] or latent spaces [51,52]. Furthermore, there is a wide range of possibilities for other kinds of noise models different from the ones considered here, including missing and duplicated nodes, and edge endpoint swaps (e.g., that can occur from crossings in imaging data). Additionally, network data often come with a wealth of node and edge annotations [53,54], with important special cases being weighted [55,56] and multilayer [57,58] networks. These extra data are potentially useful for reconstruction, although they also contain their own measurement errors. Determining the most appropriate and effective manner to model and exploit this extra information in reconstruction seems like fertile ground for future work.

ACKNOWLEDGMENTS

This research made use of the Balena High Performance Computing (HPC) Service at the University of Bath.

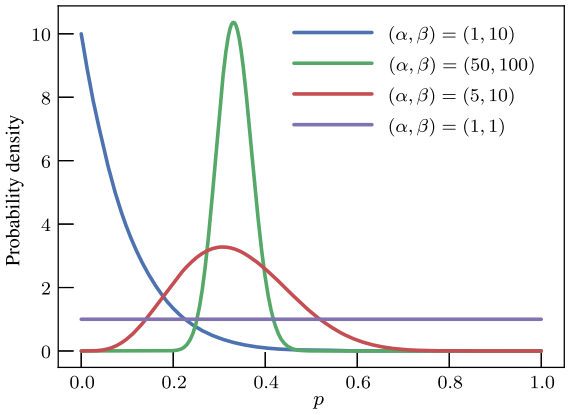

FIG. 17. Beta distributions for the noise magnitudes p and q allow us to control the degree of prior knowledge we have on their values. For example, the values $(\alpha,\beta) = (1, 10)$ represent an expectation that the value of p is relatively low, with a mode at 0 and average $\alpha/(\alpha+\beta) = 1/11 \approx 0.09$. The values $(\alpha,\beta) = (50, 100)$ express the relative certainty that the value of p is close to $1/3$, whereas the values $(\alpha,\beta) = (5, 10)$ represent the same average expectation but with less certainty. The values $(\alpha,\beta) = (1, 1)$ express the largest amount of uncertainty about the parameter p, in which case it is uniformly distributed in unit interval.

APPENDIX A: BETA PRIOR DISTRIBUTION

In Fig. 17 are shown examples of the Beta distribution of Eq. (18), for different choices of the hyperparameters α and β, illustrating their meaning with respect to the prior knowledge assumed for the missing edge probability p (and analogously for the spurious edge probability q and its hyperparameters μ and ν).

APPENDIX B: LATENT EDGE MCMC ALGORITHM

As described in the main text, we use a MCMC algorithm to sample from the posterior distribution

$$P(\boldsymbol{A}|\boldsymbol{\mathcal{D}}) = \frac{P(\boldsymbol{\mathcal{D}}|\boldsymbol{A})P(\boldsymbol{A})}{P(\boldsymbol{\mathcal{D}})}, \tag{B1}$$

where $\boldsymbol{A}$ is the network being inferred and $\boldsymbol{\mathcal{D}}$ is the measurement data. Since we are using structured distributions in place of $P(\boldsymbol{A})$, consisting of nonparametric formulations of the SBM, its computation in closed form is not tractable. Instead, we sample from the joint posterior

$$P(\boldsymbol{A},\boldsymbol{b}|\boldsymbol{\mathcal{D}}) = \frac{P(\boldsymbol{\mathcal{D}}|\boldsymbol{A})P(\boldsymbol{A}|\boldsymbol{b})P(\boldsymbol{b})}{P(\boldsymbol{\mathcal{D}})}, \tag{B2}$$

where $\boldsymbol{b}$ is the partition of nodes used for the SBM. If we sample from this distribution and ignore the values of $\boldsymbol{b}$, we obtain the desired marginal $P(\boldsymbol{A}|\boldsymbol{\mathcal{D}}) = \sum_{\boldsymbol{b}} P(\boldsymbol{A},\boldsymbol{b}|\boldsymbol{\mathcal{D}})$. However, we are often also interested in the partition itself,

as it gives information on the large-scale network structure, so we often use this in our analyses as well.

The MCMC algorithm consists of making proposals of the kind $P(\boldsymbol{b}'|\boldsymbol{A},\boldsymbol{b})$ and $P(\boldsymbol{A}'|\boldsymbol{A},\boldsymbol{b})$ for the partition and network, respectively, and accepting them according to the Metropolis-Hastings probability

$$\min\left(1,\frac{P(\boldsymbol{A}',\boldsymbol{b}'|\boldsymbol{D})P(\boldsymbol{A}|\boldsymbol{A}',\boldsymbol{b}')P(\boldsymbol{b}|\boldsymbol{A}',\boldsymbol{b}')}{P(\boldsymbol{A},\boldsymbol{b}|\boldsymbol{D})P(\boldsymbol{A}'|\boldsymbol{A},\boldsymbol{b})P(\boldsymbol{b}'|\boldsymbol{A},\boldsymbol{b})}\right), \tag{B3}$$

which does not require the computation of the intractable normalization constant $P(\mathcal{D})$. In practice, at each step in the chain, we make either a move proposal for $\boldsymbol{A}$ or $\boldsymbol{b}$, but not both at once. For the node partition, we use the move proposals similar to the ones used in Refs. [14,59], where for any given node i in group r we propose to move it to group s (which can be previously unoccupied, in which case it is labeled $s = B+1$) according to

$$P(b_i = r \to s|\boldsymbol{A},\boldsymbol{b}) = d\delta_{s,B+1} + (1-d)(1-\delta_{s,B+1})\sum_{t=1}^{B}P(t|i)\frac{e_{ts}+\epsilon}{e_t+\epsilon B}, \tag{B4}$$

where $P(t|i) = \sum_j A_{ij}\delta_{b_j,t}/k_i$ is the fraction of neighbors of i that belong to group t, $\epsilon > 0$ is a small parameter which guarantees ergodicity, and d is the probability of moving to a previously unoccupied group. [If $k_i = 0$, we assume $P(b_i = r \to s|\boldsymbol{A},\boldsymbol{b}) = d\delta_{s,B+1} + (1-d)(1-\delta_{s,B+1})/B$.] This move proposal attempts to use the currently known large-scale structure of the network to better inform the possible moves of the node, without biasing with respect to group assortativity. The parameters d and ϵ do not affect the correctness of the algorithm, only the mixing time, which is typically not very sensitive, provided they are chosen within a reasonable range (we used $d = 0.01$ and $\epsilon = 1$ throughout). When using the HDCSBM, we use the variation of the above for hierarchical partitions described in Ref. [14]. The move proposals above require only minimal bookkeeping of the number edges incident on each group and can be made in time $O(k_i)$, which is also the time required to compute the ratio in Eq. (B3), independent of how many groups are currently occupied.

For the network move proposals, we could have used simple edge-nonedge flips with

$$P(A'_{ij} = A_{ij} + \delta|\boldsymbol{A}) = \begin{cases} 1 & \text{if } A_{ij}+\delta = 1-A_{ij}, \\ 0 & \text{otherwise,} \end{cases} \tag{B5}$$

with $\delta \in \{-1,1\}$. But, in fact, since we operate with latent multigraphs, the moves are slightly different, as described in Appendix D. The correctness of the algorithm does not depend on the order or the frequency with which we attempt to update the entries (i,j), provided they are all eventually updated, so, in principle, we could choose them randomly each time. However, we find that this leads to poor mixing times, since most entries correspond to nonedges $A_{ij} = 0$ which tend to remain in that state. Instead, we choose the entries to update with a probability given by the current SBM,

$$P(i,j|\boldsymbol{A},\boldsymbol{b}) = \kappa_i\kappa_j m_{b_i,b_j}, \tag{B6}$$

with

$$\kappa_i = \frac{k_i+1}{\sum_j \delta_{b_j,b_i}k_j+1} \tag{B7}$$

being the probability of selecting node i from its group b_j, proportional to its current degree plus one, and

$$m_{rs} = \frac{e_{rs}+1}{\sum_{tu}e_{rs}+1} \tag{B8}$$

is the probability of selecting groups (r,s), where $e_{rs} = \sum_{ij}A_{ij}\delta_{b_i,r}\delta_{b_j,s}$. The above probabilities guarantee that every entry is eventually sampled, but it tends to probe denser regions more frequently, which we find to typically lead to faster mixing times. This sampling can be done in time $O(1)$, simply by keeping urns of vertices and edges according to the group memberships. The time required to compute the ratio in Eq. (B3) is also $O(1)$ for the DCSBM and $O(L)$ for the HDCSBM, where L is the hierarchy depth, again independent of the number of occupied groups.

When combining both move proposals above for the partition and network, the time required to perform V node proposals and M edge proposals is $O(\langle k\rangle V + M)$, where $\langle k\rangle$ is the average degree, which allows for the inference of very large networks, with up to millions of edges. A reference implementation of the above algorithm is freely available as part of the graph-tool library [60].

APPENDIX C: NONPARAMETRIC SBM FORMULATION

Here, we give a summary of the nonparametric SBMs used in this work, which are derived in detail in Ref. [14]. We begin with the Poisson DCSBM likelihood [13]:

$$P(\boldsymbol{A}|\boldsymbol{\lambda},\boldsymbol{\theta},\boldsymbol{b}) = \prod_{i<j}\frac{e^{-\theta_i\theta_j\lambda_{b_i,b_j}}(\theta_i\theta_j\lambda_{b_i,b_j})^{A_{ij}}}{A_{ij}!} \times \prod_i \frac{e^{-\theta_i^2\lambda_{b_i,b_i}/2}(\theta_i^2\lambda_{b_i,b_i}/2)^{A_{ii}/2}}{(A_{ii}/2)!}, \tag{C1}$$

which generates multigraphs with $A_{ij} \in \mathbb{N}$, and with self-loops allowed. By choosing the arbitrary parametrization $\sum_i \theta_i\delta_{b_i,r} = 1$ for every group r, λ_{rs} becomes the expected number of edges between groups r and s, and θ_i is

proportional to the expected degree of node i, $\theta_i = \langle k_i\rangle/\sum_s \lambda_{b_i,s}$. We use the noninformative prior for $\boldsymbol{\theta}$,

$$P(\boldsymbol{\theta}|\boldsymbol{b}) = \prod_r (n_r-1)!\delta\left(\sum_i \theta_i\delta_{b_i,r}-1\right), \tag{C2}$$

and $\boldsymbol{\lambda}$,

$$P(\boldsymbol{\lambda}|\boldsymbol{b}) = \prod_{r\le s} e^{-\lambda_{rs}/(1+\delta_{rs})\bar{\lambda}}/(1+\delta_{rs})\bar{\lambda} \tag{C3}$$

with $\bar{\lambda} = 2E/B(B+1)$, which results in the integrated marginal probability

$$\begin{aligned} P(\boldsymbol{A}|\boldsymbol{b}) &= \int P(\boldsymbol{A}|\boldsymbol{\lambda},\boldsymbol{\theta},\boldsymbol{b})P(\boldsymbol{\lambda}|\boldsymbol{b})P(\boldsymbol{\theta}|\boldsymbol{b})d\boldsymbol{\lambda}d\boldsymbol{\theta} \\ &= \frac{\bar{\lambda}^E}{(\bar{\lambda}+1)^{E+B(B+1)/2}} \times \frac{\prod_{r<s}e_{rs}!\prod_r e_{rr}!!}{\prod_{i<j}A_{ij}!\prod_i A_{ii}!!} \times \end{aligned} \tag{C4}$$

$$\prod_r \frac{(n_r-1)!}{(e_r+n_r-1)!} \times \prod_i k_i!, \tag{C5}$$

where $k_i = \sum_j A_{ij}$ is the degree of node i. As shown in Ref. [14], the above is equivalent to a microcanonical model given by

$$P(\boldsymbol{A}|\boldsymbol{b}) = P(\boldsymbol{A}|\boldsymbol{k},\boldsymbol{e},\boldsymbol{b})P(\boldsymbol{k}|\boldsymbol{e},\boldsymbol{b})P(\boldsymbol{e}|\boldsymbol{b}), \tag{C6}$$

with

$$P(\boldsymbol{A}|\boldsymbol{k},\boldsymbol{e},\boldsymbol{b}) = \frac{\prod_{r<s}e_{rs}!\prod_r e_{rr}!!\prod_i k_i!}{\prod_{i<j}A_{ij}!\prod_i A_{ii}!!\prod_r e_r!!}, \tag{C7}$$

$$P(\boldsymbol{k}|\boldsymbol{e},\boldsymbol{b}) = \prod_r \left(\!\!\binom{n_r}{e_r}\!\!\right)^{-1}, \tag{C8}$$

$$P(\boldsymbol{e}|\boldsymbol{b}) = \bar{\lambda}^E/(\bar{\lambda}+1)^{E+B(B+1)/2} \tag{C9}$$

being the corresponding noninformative priors. Following Ref. [14], we replace the microcanonical prior for the degrees with

$$P(\boldsymbol{k}|\boldsymbol{e},\boldsymbol{b}) = P(\boldsymbol{k}|\boldsymbol{\eta})P(\boldsymbol{\eta}|\boldsymbol{e},\boldsymbol{b}), \tag{C10}$$

where $\boldsymbol{\eta} = \{\eta_k^r\}$ are the degree frequencies of each group, with η_k^r being the number of nodes with degree k that belong to group r,

$$P(\boldsymbol{k}|\boldsymbol{\eta}) = \prod_r \frac{\prod_k \eta_k^r!}{n_r!} \tag{C11}$$

is a uniform distribution of degree sequences constrained by the overall degree counts, and finally

$$P(\boldsymbol{\eta}|\boldsymbol{e},\boldsymbol{b}) = \prod_r q(e_r,n_r)^{-1} \tag{C12}$$

is the distribution of the overall degree counts. The quantity $q(m,n)$ is the number of different degree counts with the sum of degrees being exactly m and that have at most n nonzero counts, given by

$$q(m,n) = q(m,n-1) + q(m-n,n). \tag{C13}$$

For the node partition, we use the prior

$$P(\boldsymbol{b}) = P(\boldsymbol{b}|\boldsymbol{n})P(\boldsymbol{n}|B)P(B) = \frac{\prod_r n_r!}{N!}\binom{N-1}{B-1}^{-1}N^{-1}, \tag{C14}$$

which is agnostic to group sizes.

The HDCSBM is obtained by replacing the uniform prior for $P(\boldsymbol{e}|\boldsymbol{b})$ by a nested sequence of SBMs, where the edge counts in level l are generated by a SBM at a level above:

$$P(\boldsymbol{e}_l|\boldsymbol{e}_{l+1},\boldsymbol{b}_l) = \prod_{r<s}\left(\!\!\binom{n_r^l n_s^l}{e_{rs}^{l+1}}\!\!\right)^{-1}\prod_r\left(\!\!\binom{n_r^l(n_r^l+1)/2}{e_{rr}^{l+1}/2}\!\!\right)^{-1}, \tag{C15}$$

where $\left(\!\binom{n}{m}\!\right) = \binom{n+m-1}{m}$ is the multiset coefficient. The prior for the hierarchical partition is obtained using Eq. (C14) at every level. We refer to Ref. [14] for further details.

Directed variations of the model above are straightforward [14], together with their noise models considered in the text, which simply require sums and products to go over all directed node pairs. We omit the expressions here for brevity, but we use the directed models whenever appropriate.

The hierarchical model above is constructed to be agnostic about several large-scale aspects of the network, including the degree distribution, the distribution of group sizes, and the mixing patterns. Because of its nonparametric nature, it can be used to infer the dimensions of the model, including the number of groups and hierarchy shape. The HDCSBM has the additional advantage that it can detect small but statistically significant groups in large networks, where the maximum number of detectable groups scales with $O(N/\ln N)$, as opposed to the $O(\sqrt{N})$ obtainable with nonhierarchical models [18,19].

APPENDIX D: ADAPTING MULTIGRAPH MODELS TO SIMPLE GRAPHS

The SBM variations considered in the previous section generate multigraphs with self-loops; however, the noise models considered in this work operate on simple graphs. The usual justification for the use of multigraph models on simple graph data is that in the sparse case they are approximately the same, since the probability of multiple

edges and self-loops being generated is very small. Although this is true for uniform SBMs, like the planted partition model considered in Sec. II C 4, it may not be true for the DCSBM when the degree distribution is sufficiently broad. In this situation, the simple and multigraph ensembles are no longer equivalent [61–63], and the use of the multigraph model in this case may lead to biases. Unfortunately, the simple graph formulations of the DCSBM cannot have their integrated likelihoods computed in closed form.

Here, we adapt the multigraph models to simple graphs in a tractable and simple way by generating multigraphs and then collapsing the multiple edges. In other words, if $\boldsymbol{G}$ is a multigraph with entries $G_{ij} \in \mathbb{N}$, the collapsed simple graph $\boldsymbol{A}(\boldsymbol{G})$ has binary entries

$$A_{ij}(G_{ij}) = \begin{cases} 1 & \text{if } G_{ij} > 0 \text{ and } i \neq j, \\ 0 & \text{otherwise.} \end{cases} \tag{D1}$$

Therefore, if $\boldsymbol{G}$ is a multigraph generated by $P(\boldsymbol{G}|\theta)$, where θ are arbitrary parameters, then the corresponding collapsed simple graph $\boldsymbol{A}$ is generated by

$$P(\boldsymbol{A}|\theta) = \sum_{\boldsymbol{G}} P(\boldsymbol{A}, \boldsymbol{G}|\theta) \tag{D2}$$

$$= \sum_{\boldsymbol{G}} P(\boldsymbol{A}|\boldsymbol{G}) P(\boldsymbol{G}|\theta), \tag{D3}$$

with

$$P(\boldsymbol{A}|\boldsymbol{G}) = \begin{cases} 1 & \text{if } \boldsymbol{A} = \boldsymbol{A}(\boldsymbol{G}), \\ 0 & \text{otherwise.} \end{cases} \tag{D4}$$

Even if $P(\boldsymbol{A}|\theta)$ cannot be computed in closed form, the joint distribution $P(\boldsymbol{A}, \boldsymbol{G}|\theta) = P(\boldsymbol{A}|\boldsymbol{G})P(\boldsymbol{G}|\theta)$ is trivial, provided we have $P(\boldsymbol{G}|\theta)$ in closed form. Therefore, instead of directly sampling from the posterior distribution

$$P(\boldsymbol{A}, \boldsymbol{b}|\mathcal{D}) = \frac{P(\mathcal{D}|\boldsymbol{A})P(\boldsymbol{A}, \boldsymbol{b})}{P(\mathcal{D})}, \tag{D5}$$

we sample from the joint posterior

$$P(\boldsymbol{A}, \boldsymbol{G}, \boldsymbol{b}|\mathcal{D}) = \frac{P(\mathcal{D}|\boldsymbol{A})P(\boldsymbol{A}|\boldsymbol{G})P(\boldsymbol{G}, \boldsymbol{b})}{P(\mathcal{D})}, \tag{D6}$$

using MCMC, treating the values G_{ij} as latent variables, and then we marginalize

$$P(\boldsymbol{A}, \boldsymbol{b}|\mathcal{D}) = \sum_{\boldsymbol{G}} P(\boldsymbol{A}, \boldsymbol{G}, \boldsymbol{b}|\mathcal{D}), \tag{D7}$$

which is done simply by sampling from $P(\boldsymbol{A}, \boldsymbol{G}, \boldsymbol{b}|\mathcal{D})$ and ignoring the actual magnitudes of the G_{ij} values and the diagonal entries. This protocol yields an almost identical MCMC algorithm to the one described in Appendix B, with the only difference that we keep track of the values of G_{ij}, which are no longer binary but automatically give us A_{ij} [which are used for the computation of $P(\mathcal{D}|\boldsymbol{A})$]. The move proposals of the entries of G_{ij} are done by unity changes:

$$P(G'_{ij} = G_{ij} + \delta|\boldsymbol{G}) = \begin{cases} 1/2 & \text{if } G_{ij} > 0, \\ 1 & \text{if } G_{ij} = 0 \text{ and } \delta = 1, \\ 0 & \text{otherwise,} \end{cases} \tag{D8}$$

again for $\delta \in \{-1, 1\}$.

In the case of the DCSBM, the degree correction happens for the multigraph $\boldsymbol{G}$ and only indirectly for $\boldsymbol{A}$. But, since our model is nonparametric and the degrees of $\boldsymbol{G}$ are also generated from their own priors, it is also a perfectly valid and useful degree-corrected model for $\boldsymbol{A}$ as well.

APPENDIX E: DATA SETS

Here, we give brief descriptions of the data sets used in this work, with properties listed in Tables I and II.

1. Data without primary error estimates

Karate club.—Social network between 34 members of a karate club [29]. The version used in Table I is the same one used in Ref. [30], with $A_{23,34} = 1$ and hence 78 edges in total. In Table II, it is assumed that each repeated entry of the adjacency matrix reported in Ref. [29] amounts to a different measurement, so that $n_{ij} = 2$ and $x_{ij} = 2A_{ij}$ for all (i, j), except for $x_{23,34} = 1$.

9/11 terrorists.—Social associations between 62 terrorists responsible for the 9/11 attacks [25,26].

American football.—Network of American football games between division IA colleges during the regular season in fall of 2000 [30].

Network scientists.—Coauthorship network of scientists working on network science [64].

C. elegans neural.—Directed neural network of the *C. elegans* worm [39], manually compiled by Watts and Strogatz [40], based on the original data. The five nodes with zero degree omitted in Ref. [40] are included in our analysis, resulting in $N = 302$ nodes in total.

Malaria genes.—Bipartite gene-substring association network for malaria [65].

Power grid.—Western state power grid of the United States [40].

Political blogs.—Citations between political blogs during the 2004 presidential election in the United States [66].

DBLP citations.—Citation network of DBLP, a database of scientific publications [67].

Open flights.—Directed network of flights between worldwide airports, collected from the community-driven Web site [68].

Reactome.—Network of protein-protein interactions in humans [69].

cond-mat.—Network of collaborations in papers published in the cond-mat section of the arXiv.org preprint Web site in the period spanning from January 1, 1995 to March 31, 2005 [41].

Enron email.—Emails sent between employees of Enron between 1999 and 2003 [70].

Linux source.—Network of Linux source code files, with directed edges denoting that they include each other [21].

Brightkite.—Online social network from the defunct brightkite Web site.

PGP.—Global web of trust of the pretty-good-privacy (PGP) encryption protocol. Nodes are public keys, and directed edges indicate that one key digitally signed another [71].

Internet AS.—Directed network of internet autonomous systems, ca. 2009, as measured by the Center for Applied Internet Data Analysis (CAIDA) [72].

Web Stanford.—Directed network of hyperlinks between the web pages from the Web site of Stanford University [73].

Flickr.—Network of images in the image-sharing site Flickr [74], where two images are connected if they share metadata, such tags, groups, or location [75].

2. Data with primary error estimates

Reality mining.—Proximity interactions between voluntary students over time [45]. As measurements, we consider the state of the network during eight consecutive Wednesdays in March and April of 2005.

School friends.—Directed network of friendship between primary and high-school students [76]. Each student is asked repeatedly to list his or her best five female and five male friends.

Human connectome.—Neuronal connections in the human brain, measured for 418 individuals, each of which we consider as a separate measurement [46].

[1] A.-L. Barabási, *The Network Takeover*, Nat. Phys. **8**, 14 (2012).

[2] P. V. Marsden, *Network Data and Measurement*, Annu. Rev. Sociol. **16**, 435 (1990).

[3] C. T. Butts, *Network Inference, Error, and Informant (In) Accuracy: A Bayesian Approach*, Soc. Networks **25**, 103 (2003).

[4] D. Liben-Nowell and J. Kleinberg, *The Link-Prediction Problem for Social Networks*, J. Am. Soc. Inf. Sci. Technol. **58**, 1019 (2007).

[5] A. Clauset, C. Moore, and M. E. J. Newman, *Hierarchical Structure and the Prediction of Missing Links in Networks*, Nature (London) **453**, 98 (2008).

[6] 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).

[7] L. Lü and T. Zhou, *Link Prediction in Complex Networks: A Survey*, Physica (Amsterdam) **390A**, 1150 (2011).

[8] A. Žnidaršič, A. Ferligoj, and P. Doreian, *Non-Response in Social Networks: The Impact of Different Non-Response Treatments on the Stability of Blockmodels*, Soc. Networks **34**, 438 (2012).

[9] T. Squartini and D. Garlaschelli, *Maximum-Entropy Networks: Pattern Detection, Network Reconstruction and Graph Combinatorics* (Springer, New York, 2017).

[10] M. E. J. Newman, *Network Structure from Rich but Noisy Data*, Nat. Phys. **14**, 542 (2018).

[11] M. E. J. Newman, *Network Reconstruction and Error Estimation with Noisy Network Data*, arXiv:1803.02427.

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

[13] Brian Karrer and M. E. J. Newman, *Stochastic Block Models and Community Structure in Networks*, Phys. Rev. E **83**, 016107 (2011).

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

[15] It is important to distinguish between the network generation given by the prior of Eq. (9) and the reconstruction given by the posterior of Eq. (10). The former is a generative process that, even if it closely captures the large-scale structure present in the underlying network, it may deviate from it in important ways, e.g., lack an abundance of triangles or other properties not well described by the SBM, and thus generate the true network with only a very small probability. In contrast, the posterior of Eq. (10) corresponds to a distribution of networks that are "centered" around the observed data and will incorporate features that are present in it, even if they are not well described by the SBM prior (such as clustering and other "small-scale" properties).

[16] 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).

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

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

[19] T. P. Peixoto, *Parsimonious Module Inference in Large Networks*, Phys. Rev. Lett. **110**, 148701 (2013).

[20] Note that the binomial terms in Eq. (15), and those that follow it, depend only on the measurement data and not on $\boldsymbol{A}$, p, or q, so ultimately they do not contribute to the posterior distribution.

[21] J. Kunegis, in *Proceedings of the 22nd International Conference on World Wide Web*, WWW '13 Companion (Association for Computing Machinery, New York, 2013), pp. 1343–1350.

[22] A. Clauset, *Ellen Tucker, and Matthias Sainz, The Colorado index of complex networks*, https://icon.colorado.edu/, 2016.

[23] One could argue that being totally agnostic about the error rates p and q is too extreme, as in many cases they are likely to be small in some sense, even if we cannot precisely quantify how small at first. The answer to this objection is

that, to the extent that this vague belief can be quantified, it should be done so via the hyperparameters α, β, γ, and μ—as it can with our method—otherwise, we have little choice but to assume maximum ignorance.

[24] We stress that the bounds of Eq. (39) are strict only in the limit of a dense network with few groups and do not represent the posterior distribution found for arbitrary data. These bounds are presented just to convey the intuition of how structure heterogeneity can inform the error probabilities.

[25] V. E. Krebs, *Mapping Networks of Terrorist Cells*, Connections **24**, 43 (2002).

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

[27] The Washington Post, 2001, https://www.washingtonpost.com/wp-srv/nation/graphics/attack/hijackers.html.

[28] ABC Eyewitness News, 2001, https://web.archive.org/web/20030415011752/http://abclocal.go.com/wabc/news/WABC_092701_njconnection.html.

[29] W. W. Zachary, *An Information Flow Model for Conflict and Fission in Small Groups*, J. Anthropol. Res. **33**, 452 (1977).

[30] M. Girvan and M. E. J. Newman, *Community Structure in Social and Biological Networks*, Proc. Natl. Acad. Sci. U.S.A. **99**, 7821 (2002).

[31] To the best of our knowledge, this issue was first identified by Clauset [32], who assembled the alternative data set with $A_{23,34} = 0$ and hence 77 edges (as opposed to the more common variant with $A_{23,34} = 1$ and 78 edges) and made it available at his Web site ca. 2015 [33].

[32] Aaron Clauset (private communication).

[33] http://santafe.edu/~aaronc/data/zkcc-77.zip.

[34] Note that $S(\hat{\boldsymbol{A}},\boldsymbol{A}^*)$ differs from the measure of accuracy commonly used in binary classification tasks, defined as the fraction of entries in $\boldsymbol{A}^*$ (both zeros and ones) that are correctly estimated in $\hat{\boldsymbol{A}}$, which in this case amounts to $1 - d(\hat{\boldsymbol{A}},\boldsymbol{A}^*)/\binom{N}{2}$. This difference is because we are more typically interested in reconstructing sparse networks, where the number of zeros (nonedges) is far larger than ones (edges), such that $d(\hat{\boldsymbol{A}},\boldsymbol{A}^*) \ll \binom{N}{2}$, for all choices of sparse $\hat{\boldsymbol{A}}$ and $\boldsymbol{A}^*$, causing the accuracy to approach one simply because $\hat{\boldsymbol{A}}$ shares most of its nonedges with $\boldsymbol{A}^*$, even if they do not have a single edge in common. The similarity $S(\hat{\boldsymbol{A}},\boldsymbol{A}^*)$ fixes this problem by normalizing instead by the total number of edges observed in both networks. Note, however, that a value of $S(\hat{\boldsymbol{A}},\boldsymbol{A}^*) = 0$ does not imply that the distance $d(\hat{\boldsymbol{A}},\boldsymbol{A}^*)$ is maximal, only that it is large enough for both networks not to share any edge.

[35] This model is somewhat crude, as degrees of simple graphs need to be further constrained [36,37], but it serves our main purpose of evaluating reconstruction quality.

[36] V. Havel, *Poznámka o Existenci Konečných Grafú*, Časopis pro pěstování matematiky **080**, 477 (1955).

[37] S. L. Hakimi, *On Realizability of a Set of Integers as Degrees of the Vertices of a Linear Graph. I*, J. Soc. Ind. Appl. Math. **10**, 496 (1962).

[38] A. Decelle, F. Krzakala, C. Moore, and L. Zdeborová, *Inference and Phase Transitions in the Detection of Modules in Sparse Networks*, Phys. Rev. Lett. **107**, 065701 (2011).

[39] 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).

[40] D. J. Watts and S. H. Strogatz, *Collective Dynamics of "Small-World" Networks*, Nature (London) **393**, 440 (1998).

[41] M. E. J. Newman, *The Structure of Scientific Collaboration Networks*, Proc. Natl. Acad. Sci. U.S.A. **98**, 404 (2001).

[42] These kinds of data also tend to suffer from name ambiguity problems, where the same author appears under different names, due, for example, to alternative spellings. But since this problem causes node duplications to occur, it cannot be corrected with our method, which can address only spurious and missing edges.

[43] In fact, since α, β, μ, and ν are unbounded continuous variables, the constant prior cannot be normalized, making it improper. The way around this situation is to use instead a constant prior constrained to some domain of interest, outside of which it is zero. If this domain is large enough to contain the inferred values, the resulting posterior is very close to the one obtained with the improper prior, which is identical to the limit (if it exists) of the posterior distribution where the domain boundaries go to infinity.

[44] Note that Beta distributions with parameters (1,0) are also improper but yield meaningful results for the same reason given in Footnote.

[45] N. Eagle and A.(Sandy) Pentland, *Reality Mining: Sensing Complex Social Systems*, Personal Ubiquitous Comput. **10**, 255 (2006).

[46] B. Szalkai, C. Kerepesi, B. Varga, and V. Grolmusz, *Parameterizable Consensus Connectomes from the Human Connectome Project: The Budapest Reference Connectome Server v3.0*, Cognit. Neurodynamics **11**, 113 (2017).

[47] J. A. McNab, B. L. Edlow, T. Witzel, S. Y. Huang, H. Bhat, K. Heberlein, T. Feiweier, K. Liu, B. Keil, J. Cohen-Adad, M. D. Tisdall, R. D. Folkerth, H. C. Kinney, and L. L. Wald, *The Human Connectome Project and Beyond: Initial Applications of* 300 mt/m *Gradients*, NeuroImage **80**, 234 (2013).

[48] T. Martin, B. Ball, and M. E. J. Newman, *Structural Inference for Uncertain Networks*, Phys. Rev. E **93**, 012306 (2016).

[49] D. Szklarczyk, J. H. Morris, H. Cook, M. Kuhn, S. Wyder, M. Simonovic, A. Santos, N. T. Doncheva, A. Roth, P. Bork, L. J. Jensen, and C. von Mering, *The STRING Database in 2017: Quality-Controlled Protein-Protein Association Networks, Made Broadly Accessible*, Nucleic Acids Res. **45**, D362 (2017).

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

[51] P. D. Hoff, A. E. Raftery, and M. S. Handcock, *Latent Space Approaches to Social Network Analysis*, J. Am. Stat. Assoc. **97**, 1090 (2002).

[52] M. E. J. Newman and T. P. Peixoto, *Generalized Communities in Networks*, Phys. Rev. Lett. **115**, 088701 (2015).

[53] M. E. J. Newman and A. Clauset, *Structure and Inference in Annotated Networks*, Nat. Commun. **7**, 11863 (2016).

[54] D. Hric, T. P. Peixoto, and S. Fortunato, *Network Structure, Metadata, and the Prediction of Missing Nodes and Annotations*, Phys. Rev. X **6**, 031038 (2016).
[55] C. Aicher, A. Z. Jacobs, and A. Clauset, *Learning Latent Block Structure in Weighted Networks*, J. Complex Netw. **3**, 221 (2015).
[56] T. P. Peixoto, *Nonparametric Weighted Stochastic Block Models*, Phys. Rev. E **97**, 012306 (2018).
[57] M. Kivelä, A. Arenas, M. Barthelemy, J. P. Gleeson, Y. Moreno, and M. A. Porter, *Multilayer Networks*, J. Complex Netw. **2**, 203 (2014).
[58] T. P. Peixoto, *Inferring the Mesoscale Structure of Layered, Edge-Valued, and Time-Varying Networks*, Phys. Rev. E **92**, 042807 (2015).
[59] T. P. Peixoto, *Efficient Monte Carlo and Greedy Heuristic for the Inference of Stochastic Block Models*, Phys. Rev. E **89**, 012804 (2014).
[60] T. P. Peixoto, *The graph-tool python library*, figshare (2014), available at https://graph-tool.skewed.de.
[61] J. Park and M. E. J. Newman, *Origin of Degree Correlations in the Internet and Other Networks*, Phys. Rev. E **68**, 026112 (2003).
[62] S. Johnson, J. J. Torres, J. Marro, and M. A. Muñoz, *Entropic Origin of Disassortativity in Complex Networks*, Phys. Rev. Lett. **104**, 108702 (2010).
[63] D. Garlaschelli, F. den Hollander, and A. Roccaverde, *Ensemble Nonequivalence in Random Graphs with Modular Structure*, J. Phys. A **50**, 015001 (2017).
[64] M. E. J. Newman, *Finding Community Structure in Networks Using the Eigenvectors of Matrices*, Phys. Rev. E **74**, 036104 (2006).
[65] 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).
[66] L. A. Adamic and N. Glance, in *Proceedings of the 3rd International Workshop on Link Discovery, LinkKDD '05* (Association for Computing Machinery, New York, 2005), pp. 36–43.
[67] M. Ley, in *Proceedings of the 9th International Symposium on String Processing and Information Retrieval, SPIRE 2002* (Springer-Verlag, London, 2002), pp. 1–10.
[68] http://www.openflights.org.
[69] G. Joshi-Tope, M. Gillespie, I. Vastrik, P. D'Eustachio, E. Schmidt, B. de Bono, B. Jassal, G. R. Gopinath, G. R. Wu, L. Matthews, S. Lewis, E. Birney, and L. Stein, *Reactome: A Knowledgebase of Biological Pathways*, Nucleic Acids Res. **33**, D428 (2004).
[70] B. Klimt and Y. Yang, *Introducing the Enron Corpus*, in *CEAS 2004, Mountain View, 2004*, http://www.ceas.cc/papers-2004/168.pdf.
[71] O. Richters and T. P. Peixoto, *Trust Transitivity in Social Networks*, PLoS One **6**, e18384 (2011).
[72] Available at https://www.caida.org/data/.
[73] J. Leskovec, K. J. Lang, A. Dasgupta, and M. W. Mahoney, *Community Structure in Large Networks: Natural Cluster Sizes and the Absence of Large Well-Defined Clusters*, arXiv:0810.1355.
[74] http://flickr.com.
[75] J. Mcauley and J. Leskovec, *Discovering Social Circles in Ego Networks*, ACM Trans. Knowl. Discovery Data **8**, 1 (2014).
[76] J. Moody, *Peer Influence Groups: Identifying Dense Clusters in Large Networks*, Soc. Networks **23**, 261 (2001).

Correction: The caption to Fig. 3 contained typographical errors and has been fixed.