Estimating network structure from unreliable measurements

M. E. J. Newman

Phys. Rev. E **98**, 062321 — Published 26 December 2018

DOI: 10.1103/PhysRevE.98.062321

Estimating network structure from unreliable measurements

M. E. J. Newman
Department of Physics and Center for the Study of Complex Systems, University of Michigan, Ann Arbor, MI 48109, USA

Most empirical studies of networks assume that the network data we are given represent a complete and accurate picture of the nodes and edges in the system of interest, but in real-world situations this is rarely the case. More often the data only specify the network structure imperfectly—like data in essentially every other area of empirical science, network data are prone to measurement error and noise. At the same time, the data may be richer than simple network measurements, incorporating multiple measurements, weights, lengths or strengths of edges, node or edge labels, or annotations of various kinds. Here we develop a general method for making estimates of network structure and properties using any form of network data, simple or complex, when the data are unreliable, and give example applications to a selection of social and biological networks.

I. INTRODUCTION

Networks are widely used as a convenient quantitative representation of patterns of connection between the nodes, units, agents, or elements in complex systems, particularly technological, biological, and social systems. There has been an explosion of empirical work in the last two decades aimed at measuring and describing the structure of networks such as the Internet, the World Wide Web, road and airline networks, friendship networks, biochemical networks, ecological networks, and others [1].

A fundamental issue with empirical studies of networks, however, is that the data we have are often unreliable. Most measurement techniques for network structure suffer from measurement error of some kind. In biological networks such as metabolic or protein interaction networks, for example, traditional laboratory experimental error is a primary source of inaccuracy. There exist experimental methods for directly measuring interactions between proteins, such as affinity purification or yeast two-hybrid screens, but even under the best controlled conditions the exact same measurement repeated twice may yield different results [2–4]. Or consider the Internet, whose network structure is usually determined by examining either router tables or collections of traceroute paths. Both both router tables and traceroute paths, however, give only subsets of the edges in the network. Commonly one combines many tables or paths to provide better coverage, but even so it is well established that the resulting network structure contains significant errors [5, 6]. Social networks, such as friendship networks, provide another example. Such networks are typically measured using surveys, and the resulting data can contain errors of many kinds, including subjectivity on the part of respondents in surveys, missing data, and recording and coding errors [7–11].

Not all is bad news, however. Network data may be error-prone but they can also be very rich. Many studies produce not just simple measurements of network structure but multifaceted data sets that reflect structure from many different angles. A friendship network might be measured repeatedly, for instance, or measured in multiple ways using reported interactions, observed interactions, online data, or archival records. An ecological network such as a food web might combine field data of many different types as well as input from curated library studies. A data set for the World Wide Web will typically include not only the pattern of links between web pages but rich data on page content including word frequencies, headings, word positions, anchor text, and metadata.

In this paper we consider the problem of *network reconstruction*, deriving and demonstrating a broad class of methods that can be used to infer the true structure and properties of a network from potentially error-prone data in any format.

There has been a significant amount of previous work on network error and reconstruction, including discussion of the types of errors that can occur, for instance in social networks [7–11], biological networks [3, 12], and technological networks [5, 13], and studies of simulated errors that aim to determine what effect errors will have on estimates of network properties [14–19]. Methods for estimating true network structure from error-prone measurements have been developed in several fields, including sociology, statistics, physics, and computer science. Perhaps most closely related to the work reported here is that of Butts [9], who developed Bayesian methods for estimating how reliable social network survey respondents are in the answers they give. Our technical approach is different from that of Butts, but some of the example applications we consider address similar questions. The methods we develop generate a posterior distribution over possible network structures, and a number of other previous methods have been proposed for doing this, by ourselves and others, albeit using different approaches [20, 21]. Looking further afield, there is also a considerable body of work on estimating network structure from measurements of the evolution of networked dynamical systems such as coupled oscillators [22] or spreading processes [23, 24]. There has also been work on error correction strategies for networks, which can be viewed as a form of network reconstruction. Link prediction in particular—the task of identifying missing edges in networks—has received considerable

attention [20, 21, 25–27]. In the study of citation and collaboration networks a number of methods have been developed for name disambiguation, which can be thought of as a form of error correction for missing or extraneous nodes [28–32]. And a combination of methods of these kinds can be used to create hybrid "coupled classifier" algorithms for processing raw network data into more accurate and nuanced estimates of network quantities [33–36], often with a focus on a specific domain of study. Of particular interest is the large body of work on methods for analyzing and processing high-throughput laboratory data on biological and biochemical networks [37–42].

In a previous paper [43] we outlined a method for making optimal estimates of network structure using expectation–maximization (EM) algorithms, and in other work we and others have looked at methods for estimating networks in the presence of community structure [44, 45]. Here we build on this previous work and lay out a general formalism for inferring network structure from rich but noisy data sets. We focus specifically on the problem of inferring the positions of the edges in a network of given nodes. There are interesting questions to be asked about how one identifies the nodes in the first place, but these we will not tackle here.

II. APPROACH

The approach we present, which builds on our previous work in [43, 44], is based on model fitting and has two main components: a *network model* that represents how the network structure is generated (or, more properly, our belief about how it is generated), and a *data model* that represents how that structure maps onto the observed data. Given a set of observations, the method allows us to infer the parameters of both models as well as the entire posterior distribution over possible network structures. Features of interest in the network can also be estimated, in one of two different ways: one can inspect the parameters of the network model (as is done in community detection, for example) or one can calculate expected values of network metrics over the posterior distribution (as one might for things like degree distributions, path lengths, or correlations).

The requirement that we define network and data models is one aspect that distinguishes our method from other network reconstruction approaches. One might consider this requirement to be a disadvantage of the method, since it obliges us to make assumptions about our networks, but we would argue that this is a feature, not a bug. We argue that other methods are also making assumptions, though one may not notice them because they are often hidden from view. They may be implicit, for instance, in the decisions a programmer makes in developing code, or they may be the result of subconscious choices not even recognized by the researcher, but they nonetheless still affect the calculations [46]. One might implicitly assume, for instance, that edges are more likely between nodes that share a common neighbor [25], or that patterns of connections are similar between nodes that have other features in common [26]. In the literature on inferring network structure from dynamical processes taking place on network nodes, the dynamics is often effectively assumed to be driven by a particular dynamical model, although there have been efforts in recent years to develop model-free approaches [47, 48]. We believe it to be a desirable feature of the approach proposed here that it obliges us to acknowledge explicitly what assumptions we are making and formulate them in a precise manner.

Suppose then that we are interested in a particular network of n nodes, whose structure we will represent by an adjacency matrix $\mathbf{A}$. In the simplest case of an unweighted undirected network the adjacency matrix is an $n \times n$ symmetric matrix with elements $A_{ij} = 1$ if nodes i and j are connected by an edge and 0 otherwise. Our methods can also be applied to directed networks (represented by asymmetric matrices), weighted networks (represented by matrices containing values other than 0 and 1), and other more complicated forms if necessary, but for the moment we will concentrate on the undirected unweighted case.

We assume that the structure $\mathbf{A}$ of the network is initially unknown. This structure is sometimes called the *ground truth*. Our aim is to estimate the ground truth from the results of measurements of some kind. The measurements could take many forms: measurements of single edges, pathways, or subgraphs; repeated measurements or measurements made from the point of view of different participants or locations; metadata concerning edges or nodes; nonlocal or global properties of the network as a whole, such as densities, clustering coefficients, or spectral properties, or any of many other measurement types. Let us denote by D the complete set of data generated by the measurements performed on the system. We specifically do not assume that the data are reliable (they may contain errors of various kinds) or that they are complete (some parts of the network may not be measured). Our goal is to make the best estimate we can of the structure of the network given the available data.

This we do using probabilistic methods. We first define a network model, which represents our prior knowledge about the structure of the network. This model takes the form of a probability distribution $P(\mathbf{A}|\gamma)$, where γ denotes the parameters (if any) of the distribution. The parameters are normally unknown; we will show how to estimate their values shortly. The network model quantifies what we know about the structure of the network before we observe any data. The model could be a very simple one. For instance, in the (common) case where we know nothing about the structure of the network ahead of time, the model could be just a uniform (maximum-entropy) distribution, all structures being equally likely. In fact, we usually know at least a little more than this. For instance, almost all empirically observed networks are relatively sparse, meaning that only a small fraction of their possible edges are present. Armed with this addi-

tional knowledge, we might choose to employ a network model, such as a random graph, that favors (or at least can favor) networks with low edge density. We use models of this kind in several calculations in this paper. More complex choices are also possible and may be useful in some cases. If we are interested in performing community detection on our network, for example, then we might hypothesize that the network is drawn from a stochastic block model [49]. The parameters of the fitted block model can then tell us about the community structure, if any, in the observed network [44, 45, 50, 51].

Second, we hypothesize a measurement process or data model that describes how our empirical data D are generated from observations of the network, such that $P(D|\mathbf{A},\theta)$ is the probability of the data given the true structure of the network $\mathbf{A}$ and model parameters θ. Combining probabilities and applying Bayes rule, we then have

$$P(\mathbf{A},\gamma,\theta|D) = \frac{P(D|\mathbf{A},\theta)P(\mathbf{A}|\gamma)P(\gamma)P(\theta)}{P(D)}, \quad (1)$$

where $P(\gamma)$, $P(\theta)$, and $P(D)$ are the prior probabilities of the parameters and the data (which we assume to be independent). Summing over all possible values of the unknown adjacency matrix $\mathbf{A}$ (or integrating in the case of continuous-valued matrix elements), we get an expression for the posterior probability of the parameter values γ,θ given the data:

$$P(\gamma,\theta|D) = \sum_{\mathbf{A}} P(\mathbf{A},\gamma,\theta|D). \quad (2)$$

Our first goal will be to find the most likely values of the parameters by maximizing this posterior probability with respect to γ and θ, a so-called *maximum a posteriori* or MAP estimate.

In fact, as is often the case, it is convenient to maximize not the probability itself but its logarithm, which has its maximum in the same place. We make use of Jensen's inequality, which states that for any set of positive quantities x_i,

$$\log \sum_i x_i \geq \sum_i q_i \log \frac{x_i}{q_i}, \quad (3)$$

where q_i are an equal number of nonnegative quantities satisfying $\sum_i q_i = 1$. Applying this inequality to the log of Eq. (2) we have

$$\begin{aligned} \log P(\gamma,\theta|D) &= \log \sum_{\mathbf{A}} P(\mathbf{A},\gamma,\theta|D) \\ &\geq \sum_{\mathbf{A}} q(\mathbf{A}) \log \frac{P(\mathbf{A},\gamma,\theta|D)}{q(\mathbf{A})}, \end{aligned} \quad (4)$$

where $q(\mathbf{A})$ is any nonnegative function of $\mathbf{A}$ satisfying $\sum_{\mathbf{A}} q(\mathbf{A}) = 1$. It will be convenient to think of $q(\mathbf{A})$ as a probability distribution over networks $\mathbf{A}$.

It is straightforward to see that the exact equality in (4) is achieved, and hence the right-hand side of the inequality maximized, when

$$q(\mathbf{A}) = \frac{P(\mathbf{A},\gamma,\theta|D)}{\sum_{\mathbf{A}} P(\mathbf{A},\gamma,\theta|D)}. \quad (5)$$

Since this choice makes the right-hand side equal to $\log P(\gamma,\theta|D)$, a further maximization with respect to γ and θ will then give us the MAP estimate that we seek. To put that another way, maximization of the right-hand side of (4) with respect both to q and to γ and θ will give us the optimal values of the parameters.

This leads to a natural iterative algorithm for determining the values of the parameters: we perform the maximization by maximizing first over q with the parameters held constant, then over the parameters with q held constant, and repeat until we converge to the final answer.

The maximum over q is given by Eq. (5). The maximum over the parameters we find by differentiating. Taking derivatives of the right-hand side of Eq. (4) while holding $q(\mathbf{A})$ constant, we get

$$\sum_{\mathbf{A}} q(\mathbf{A}) \nabla_\gamma \log P(\mathbf{A},\gamma,\theta|D) = 0, \quad (6)$$

$$\sum_{\mathbf{A}} q(\mathbf{A}) \nabla_\theta \log P(\mathbf{A},\gamma,\theta|D) = 0, \quad (7)$$

where $\nabla_\gamma, \nabla_\theta$ denote derivatives with respect to the sets γ,θ of parameters of the two models. Alternatively, making use of Eq. (1), we have

$$\nabla_\gamma \log P(\gamma) + \sum_{\mathbf{A}} q(\mathbf{A}) \nabla_\gamma \log P(\mathbf{A}|\gamma) = 0, \quad (8)$$

$$\nabla_\theta \log P(\theta) + \sum_{\mathbf{A}} q(\mathbf{A}) \nabla_\theta \log P(D|\mathbf{A},\theta) = 0. \quad (9)$$

The solution of these equations gives us our values for γ,θ. Note that Eq. (8) depends only on the network model and its solution gives the parameter values for that model. Similarly, Eq. (9) depends only on the data model and gives the parameters for that model.

This is an example of an expectation–maximization or EM algorithm [52, 53], a standard tool for statistical inference in situations where some data are unknown or hidden from us—in this case the network structure $\mathbf{A}$. Implementation of the algorithm involves choosing random initial values for the parameters γ,θ and then iterating Eq. (5) and Eqs. (8) and (9) until convergence is reached. The EM algorithm can be proved to converge to a local maximum of the posterior probability, but not necessarily to the global maximum we would like to find. In practice, therefore, one often performs repeated runs, starting from different initial values, to test for consistent convergence.

The output of the EM algorithm is a set of values for the parameters γ,θ. These are "point estimates," representing the single most likely values for the quantities in

question. There exist other (Bayesian) methods that can compute entire posterior distributions over parameters but in most of the applications we consider such an approach is unnecessary. The quantity of data embodied in the networks we study is typically large enough that the parameter values are quite precisely determined, meaning that the posterior distributions are sharply peaked, and hence the EM algorithm tells us everything we want to know. Just as in traditional statistical mechanics, the fact that we are studying a large system makes the point estimates highly accurate. (One exception occurs when we use a model that has an extensive number of parameters. In this case a Bayesian approach may give additional information that cannot be derived from the EM algorithm, but we will not pursue such approaches in this paper.)

After calculating the parameter values the next step would normally be to use them in Eq. (1) to find the probability distribution over networks $\mathbf{A}$. It turns out, however, that this is unnecessary, since the network structure can be deduced from results we have already calculated. Note that Eq. (5) can be written as

$$q(\mathbf{A}) = \frac{P(\mathbf{A},\gamma,\theta|D)}{P(\gamma,\theta|D)} = P(\mathbf{A}|D,\gamma,\theta). \tag{10}$$

In other words, $q(\mathbf{A})$ is the probability that the network has structure $\mathbf{A}$ given the observed data and our values for the parameters γ,θ. Thus the EM algorithm already gives us the entire posterior distribution over possible ground-truth network structures. In many cases this posterior probability distribution is the primary object of interest in the calculation, capturing both the network structure itself and the uncertainty in that structure.

Once we have this distribution, any other network quantity we are interested in, including degrees, correlations, clustering coefficients, and so forth, can be estimated from it. Specifically, for any quantity $X(\mathbf{A})$ that is a function of the network structure $\mathbf{A}$, the expected value, given the observed data and the parameter estimates, is

$$\mu_X = \sum_{\mathbf{A}} X(\mathbf{A}) P(\mathbf{A}|D,\gamma,\theta), \tag{11}$$

and the variance about that value is

$$\sigma_X^2 = \sum_{\mathbf{A}} [X(\mathbf{A}) - \mu_X]^2 P(\mathbf{A}|D,\gamma,\theta). \tag{12}$$

It is not always possible to perform the sums over $\mathbf{A}$ in these expressions analytically. In cases where they cannot be done, numerical approximations using Monte Carlo sampling can give good answers in reasonable time.

The values of the model parameters may also be of interest, both for the network model and for the data model. In cases where the parameters of the network model correspond to meaningful network quantities, they can give us useful information, as in the case of community detection using the stochastic block model [44, 45]. More commonly, however, it is the parameters of the data model that are of interest because they quantify the measurement process and hence can give us insight into the reliability of the data and the types of error they may contain.

III. NETWORK MODELS

Applying the methods of the previous section requires us to choose the models we will use: the network model, which describes the prior probability distribution over networks, and the data model, which describes how the data are related to the network structure. In this and the following section we give some examples of possible choices, starting with network models.

The network models most commonly used for structural inference in networks are random graph models in which the edges are (conditionally) independent random variables. The best known examples are the (Bernoulli) random graph, the configuration model, the stochastic block model, and their many variants.

A. The random graph

The simplest of network models is the standard random graph, in which every pair of distinct nodes i,j is connected by an edge with equal probability ω. For this model the probability $P(\mathbf{A}|\gamma)$ becomes

$$P(\mathbf{A}|\omega) = \prod_{i<j} \omega^{A_{ij}} (1-\omega)^{1-A_{ij}}. \tag{13}$$

Despite its simplicity, this model works well for many of the calculations we will look at. In the absence of evidence to the contrary, simply assuming that all edges are equally likely is a sensible approach. We do need to choose a prior probability $P(\omega)$ for the parameter ω. In the calculations we perform we will assume that all values of this parameter are equally likely, so that $P(\omega) = 1$.

B. Edge types

Various extensions of the simple random graph are possible. For instance, one could have a model in which instead of just two edge states (present/not present) we have three or more. In a social network, for instance, one might divide pairs of individuals into those who are not acquainted, somewhat acquainted, or well acquainted. Such states could be represented by adjacency matrix elements with values 0, 1, and 2, with corresponding probabilities ω_0, ω_1, and ω_2. More generally any number k of states could be represented by $A_{ij} = 0 \ldots k-1$ and probabilities $\omega_0 \ldots \omega_{k-1}$, subject to the constraint that $\sum_{m=0}^{k-1} \omega_m = 1$. Then the probability of a particular net-

work is

$$P(\mathbf{A}|\omega) = \prod_{i<j} \omega_{A_{ij}} = \prod_{i<j} \prod_{m=0}^{k-1} \omega_m^{\delta_{m,A_{ij}}}. \tag{14}$$

A variant of this type of model is one in which the edges are signed, meaning that they can have both positive and negative values. Such signed networks are sometimes used, for instance, to represent social networks in which interactions can be both positive and negative—friendship and animosity [54]. In the simplest case the elements of the adjacency matrix take three values 0, +1, and −1, and the probability $P(\mathbf{A}|\omega)$ is an obvious variation on Eq. (14).

C. Poisson edge model

In many calculations with network models one assumes not Bernoulli (i.e., zero/one) random variables for the edges but Poisson ones. That is, rather than placing edges with probability ω or not with probability $1-\omega$, one places a Poisson distributed number of edges with mean ω. This results in a network that can contain more than one edge between a given pair of nodes—a so-called multiedge—which is in a sense unrealistic since most observed networks do not have multiedges. However, the probability of having a multiedge is of order ω^2, which is typically negligible in the common case of a sparse network where ω is small, and hence the Poisson and Bernoulli models generate essentially the same ensemble in the sparse case. At the same time the Poisson model is often significantly easier to work with and has become favored for many applications. Commonly, in addition to multiedges, one also allows self-edges, placing a Poisson-distributed number of such edges at each node with mean $\frac{1}{2}\omega$. By convention a self-edge is represented by an adjacency matrix element $A_{ii} = 2$ (not 1). The factor of $\frac{1}{2}$ in the density of self-edges compensates for the 2 in the definition of A_{ii}, so that the expected value of all adjacency matrix elements is simply ω.

For this model the equivalent of Eq. (13) is

$$P(\mathbf{A}|\omega) = \prod_{i<j} \frac{\omega^{A_{ij}}}{A_{ij}!} \mathrm{e}^{-\omega} \prod_i \frac{\bigl(\frac{1}{2}\omega\bigr)^{A_{ii}/2}}{\bigl(\frac{1}{2}A_{ii}\bigr)!} \mathrm{e}^{-\omega/2}, \tag{15}$$

and the log of this probability is

$$\begin{aligned}\log P(\mathbf{A}|\omega) &= \tfrac{1}{2}\sum_{ij}\bigl(A_{ij}\log\omega - \omega\bigr) \\ &\quad - \sum_{i<j}\log A_{ij}! - \sum_i \bigl[\tfrac{1}{2}A_{ii}\log 2 + \log(\tfrac{1}{2}A_{ii})!\bigr].\end{aligned} \tag{16}$$

Here we have separated out terms that do not depend on ω. When we perform a derivative as in Eq. (8), these terms will vanish, leaving an especially simple result for ω.

Note also that this model and the previous one both use values of A_{ij} other than zero and one, but the values have different meanings. In the model of Section III B they represent different types of edges; in the model of this section they represent multiedges.

D. Stochastic block model

A more complex model, well studied in the networks literature, is the stochastic block model. First proposed in the 1980s by Holland *et al.* [49], the stochastic block model is a model of community structure in networks, although with large numbers of blocks it can also function as a model of very general kinds of network structure [55, 56]. In essence, the model consists of a set of random graphs stitched together into a larger network. We take n nodes and divide them into some number k of groups labeled by integers $1 \ldots k$, with μ_r being the probability that a node is assigned to group r. Then we place undirected edges between distinct nodes independently such that the probability of an edge between a given pair of nodes depends only on the groups that the nodes belong to.

The model is simplest when written using the Poisson formulation of Section III C: between nodes i, j belonging to groups r, s we place a number of edges which is Poisson distributed with mean ω_{rs}, except for self-edges $i = j$ for which the mean is $\frac{1}{2}\omega_{rr}$. The edge frequencies ω_{rs} thus dictate the relative probabilities of within- and between-group connections. In the most widely studied case, the diagonal elements ω_{rr} are chosen to be larger than the off-diagonal ones, so that edges are more likely within groups than between them, a type of structure known as assortative mixing or homophily. Other types of structure are also possible, however, and are observed in some networks, such as disassortative structure in which between-group edges are more likely than in-group ones [57].

Let us denote by g_i the label of the group to which node i is assigned. Then, given the parameters μ_r and ω_{rs}, the probability of generating a complete set of group assignments $\mathbf{g} = \{g_i\}$ and a network $\mathbf{A}$ in this model is

$$\begin{aligned}P(\mathbf{g},\mathbf{A}|\mu,\omega) = \prod_i \mu_{g_i} \prod_{i<j} \frac{\omega_{g_i g_j}^{A_{ij}}}{A_{ij}!} \exp\bigl(-\omega_{g_i g_j}\bigr) \\ \times \prod_i \frac{\bigl(\frac{1}{2}\omega_{g_i g_i}\bigr)^{A_{ii}/2}}{\bigl(\frac{1}{2}A_{ii}\bigr)!} \exp\bigl(-\tfrac{1}{2}\omega_{g_i g_i}\bigr).\end{aligned} \tag{17}$$

which has logarithm

$$\begin{aligned}\log P(\mathbf{g},\mathbf{A}|\mu,\omega) &= \sum_i \log\mu_{g_i} + \tfrac{1}{2}\sum_{ij}\bigl(A_{ij}\log\omega_{g_i g_j} - \omega_{g_i g_j}\bigr) \\ &\quad - \sum_{i<j}\log A_{ij}! - \sum_i \bigl[\tfrac{1}{2}A_{ii}\log 2 + \log\bigl(\tfrac{1}{2}A_{ii}\bigr)!\bigr].\end{aligned} \tag{18}$$

Again this separates terms that involve the parameters μ and ω from those that do not, making derivatives like those in Eq. (8) simpler.

In the formalism considered in Section II, the structure of the network is the only kind of unobserved data, but in the stochastic block model there are two kinds: the network $\mathbf{A}$ and the group assignments $\mathbf{g}$. Our EM algorithm carries over straightforwardly to this case, but with the joint distribution of $\mathbf{g}$ and $\mathbf{A}$ taking the place of the distribution of $\mathbf{A}$ alone. When combined with a suitable data model, this approach allows us to infer both the network structure and the community structure from a single calculation. An alternative approach, considered by Le and Levina [45], is to assume that the community structure is known via other means and only the network structure is to be inferred using the EM algorithm. In that case, Eq. (17) is replaced with

$$P(\mathbf{A}|\mathbf{g},\mu,\omega) = \prod_{i<j} \frac{\omega_{g_i g_j}^{A_{ij}}}{A_{ij}!} \exp\bigl(-\omega_{g_i g_j}\bigr) \times \prod_i \frac{\bigl(\frac{1}{2}\omega_{g_i g_i}\bigr)^{A_{ii}/2}}{\bigl(\frac{1}{2}A_{ii}\bigr)!} \exp\bigl(-\tfrac{1}{2}\omega_{g_i g_i}\bigr), \tag{19}$$

and $\mathbf{g}$ is treated as "data" with known values. Then the EM algorithm once again gives a posterior distribution on the network structure alone. In their work Le and Levina found the community structure using a traditional spectral algorithm.

E. The configuration model and degree correction

A potential issue with the models of the previous sections is that they all generate networks with Poisson degree distributions, which are quite unlike the strongly right-skewed degree distributions seen in many real-world networks [58, 59]. We can circumvent this issue and create more realistic models using *degree correction*.

The simplest example of a degree-corrected model is the configuration model, a random graph model that allows for arbitrary degree distributions [60, 61]. In the configuration model, one fixes the degree of each node separately and then places edges at random, but respecting the chosen node degrees. The standard way to do this is to place "stubs" of edges at each node, equal in number to the chosen degree, then join pairs of stubs together at random to create complete edges. It can be shown [1] that the number of edges falling between nodes i and j in such a network is Poisson distributed with mean $d_i d_j / \sum_k d_k$, where d_i is the degree of node i.

In our calculations we make use of a variant of the configuration model, similar to one proposed by Chung and Lu [62], in which, rather than employing edge stubs, one simply places a Poisson distributed number of edges between each pair of nodes with the appropriate mean. We define a set of real-valued parameters ϕ_i, one for each node i, then place a number of edges between each pair of nodes i, j which is Poisson distributed with mean $\omega\phi_i\phi_j$, or half that number if $i = j$. Note that, like the model of Section III C, this model can produce networks with self-edges or multiedges (or both), which is somewhat unrealistic. One commonly allows them nonetheless because it leads to technical simplifications and does not in practice make much difference in the common case of a sparse network.

As defined, the parameters ϕ_i and ω are not identifiable: one can increase all ϕ_i by any constant factor without changing the model if one also decreases ω by the square of the same factor. We can fix the values of the parameters by choosing a normalization for the ϕ_i. This can be done in several ways, all of which ultimately give equivalent results, but for present purposes a convenient choice is to set the average of the ϕ_i equal to 1:

$$\frac{1}{n}\sum_i \phi_i = 1. \tag{20}$$

This choice has the nice feature that the average of the elements of the adjacency matrix is then given by

$$\frac{1}{n^2}\Biggl[\sum_{i\neq j}\omega\phi_i\phi_j + 2\sum_i \tfrac{1}{2}\omega\phi_i^2\Biggr] = \frac{\omega}{n^2}\sum_{ij}\phi_i\phi_j = \omega. \tag{21}$$

Thus ω is the average value of an adjacency matrix element, just as in the earlier model of Section III C.

Given the parameters of the model, the probability $P(\mathbf{A}|\phi,\omega)$ of generating a particular network is

$$P(\mathbf{A}|\phi,\omega) = \prod_{i<j}\frac{(\omega\phi_i\phi_j)^{A_{ij}}}{A_{ij}!}\mathrm{e}^{-\omega\phi_i\phi_j} \times \prod_i \frac{(\frac{1}{2}\omega\phi_i^2)^{A_{ii}/2}}{(\frac{1}{2}A_{ii})!}\mathrm{e}^{-\omega\phi_i^2/2}, \tag{22}$$

and its log is

$$\log P(\mathbf{A}|\phi,\omega) = \tfrac{1}{2}\sum_{ij}A_{ij}\log\omega + \sum_{ij}A_{ij}\log\phi_i - \tfrac{1}{2}n^2\omega - \sum_{i<j}\log A_{ij}! - \sum_i\bigl[\tfrac{1}{2}A_{ii}\log 2 + \log(\tfrac{1}{2}A_{ii})!\bigr], \tag{23}$$

where we have made use of Eq. (20).

One can apply the same degree correction approach to the stochastic block model of Section III D, which leads to the so-called degree-corrected stochastic block model [51]. In this model we again divide nodes among k groups with probability μ_r of assignment to group r, but now between each pair of nodes i, j we place a number of edges that is Poisson distributed with mean $\omega_{rs}\phi_i\phi_j$, where r and s are respectively the groups to which nodes i and j belong. The additional factor of $\phi_i\phi_j$ allows us to control the degrees of the nodes and give the network essentially any degree distribution we desire. We can fix

the normalization of the parameters in various ways, for example by choosing the mean of ϕ_i to be 1 within each individual group thus:

$$\frac{1}{n_r}\sum_i \delta_{r,g_i}\phi_i = 1 \tag{24}$$

for all r, with $n_r = \sum_i \delta_{r,g_i}$ being the number of nodes in group r.

IV. DATA MODELS

We now turn to data models, meaning models of the measurement process. These models represent the way the data measured in our experiments depend on the underlying ground-truth network.

A. Independent edge measurements

Perhaps the simplest data model is one in which observations of edges are independent identically distributed Bernoulli random variables, conditioned only on the presence or absence of an edge in the same place in the ground-truth network. That is, we make a measurement on a node pair i, j and it returns a simple yes-or-no answer about whether the nodes are connected by an edge, which depends only on the adjacency matrix element A_{ij} for the same node pair and any parameters of the process, and is independent of other matrix elements or any other measurements we may make. That is not to say, however, that the answers we get need be accurate, and in general we will assume that they are not. In an error-prone world, our measurements will sometimes reflect the truth about whether an edge exists and sometimes they will not.

Consider the simplest case in which A_{ij} takes only the values zero and one. We can then parametrize the possible outcomes of a measurement by two probabilities: the true-positive rate α, which is the probability of observing an edge where one truly exists, and the false-positive rate β, which is the probability of observing an edge where none exists. (The two remaining possibilities, of true negatives and false negatives, occur with probabilities $1-\beta$ and $1-\alpha$ respectively, so no additional parameters are needed to represent the rates of these events.) The probability of observing an edge between nodes i and j can then be succinctly written as $\alpha^{A_{ij}}\beta^{1-A_{ij}}$ and the probability of not doing so is $(1-\alpha)^{A_{ij}}(1-\beta)^{1-A_{ij}}$.

We give an application of this model to an example data set in Section V C.

B. Multiple edge types

In Section III B we introduced a network model in which edges have several types, representing for instance different strengths of acquaintance in a social network. The k edge types were represented by integer values of adjacency matrix elements $A_{ij} = 0 \ldots k-1$. Observed data for such a network could take several forms. For instance, one can imagine situations in which it might be possible, via a measurement of some kind, to determine not only whether an edge exists between two nodes but also what type of edge it is. Such a situation could be represented by a set of variables that parametrize the probability of observing an edge of type j between a pair of nodes if there is an edge of type k in the ground truth. This, however, leads to a rather complicated data model. A simpler set-up is one in which measurements return only a yes-or-no answer about whether two nodes are connected by an edge and no information about edge type. This can be represented by a model with separate parameters $\alpha_0 \ldots \alpha_{k-1}$ equal to the probability of observing an edge given each of the different ground-truth edge states. Then the probability of observing an edge between nodes i and j is simply $\alpha_{A_{ij}}$ and the probability of not observing one is $1-\alpha_{A_{ij}}$.

C. Multimodal data

There are many cases where the data for a network consist not merely of one type of measurement but of two or more. For instance a social network might be measured by surveying participants using traditional questionnaires or interviews, but also by collecting social media data, email or text messages, or using observations of face-to-face interactions [63–65]. A protein–protein interaction network might be measured using a combination of co-immunoprecipitation, affinity purification, yeast two-hybrid screens, or other methods [66]. When represented as networks, such data are sometimes called multilayer or multiplex networks [67, 68].

Measurements of different types can be governed by different probabilities and errors. Assuming a ground-truth network represented by a simple binary adjacency matrix with $A_{ij} = 0$ or 1, one could define separate true- and false-positive probabilities α_m, β_m for each type of measurement. That is, α_m is the probability that a measurement of type m will reveal an edge between two nodes i, j where an edge truly exists ($A_{ij} = 1$), and β_m is the probability that such a measurement will reveal an edge where none exists ($A_{ij} = 0$). Then the total probability of observing an edge between i and j using a measurement of type m is $\alpha_m^{A_{ij}}\beta_m^{1-A_{ij}}$ and the probability of not observing one is $(1-\alpha_m)^{A_{ij}}(1-\beta_m)^{1-A_{ij}}$.

D. Directed edges and individual node errors

The models we have described so far assume undirected edges, but it is straightforward to generalize them to the case of directed networks. Directed versions of the basic network models exist already, such as directed versions

of the configuration model [61] or the stochastic block model [69]. Data models for directed networks are a natural generalization of the undirected versions. For instance, one could assume that the empirical observations of interactions between nodes in a directed network are independent directed Bernoulli random variables that depend on the underlying ground-truth edges, with appropriately defined true- and false-positive rates. In most cases the equations for the models are straightforward generalizations of those for the undirected case. We give an example in Section V D.

In some cases it is possible for the observations of edges to be directed even if the underlying ground-truth network is undirected, or *vice versa*. Perhaps the most prominent example of this phenomenon arises in the study of social networks such as friendship or acquaintance networks. In studies of these networks, by far the most common method for collecting data is simply to ask people who their friends or acquaintances are. This results in directed edge measurements in which the fundamental unit of data is a statement by person i that they are acquainted with person j. Often, however, we would consider the underlying network itself to be undirected—either two people are acquainted or they are not. This situation can again be represented with a relatively straightforward generalization of earlier data models in which directed observations depend on the underlying undirected ground truth, with appropriately defined true- and false-positive rates.

Directed measurements like these give rise to the possibility that two people may make contradictory statements about whether they are acquainted: person i may claim to know person j but person j may claim not to know i. Such unreciprocated claims are in fact common in social network studies [70]. In surveys of friendship among schoolchildren, for instance, only about a half of all claimed friendships are reciprocated [71]. Such a situation can arise naturally in the data model: if the true- and false-positive rates for observations are α and β as before, the probability of both of two individuals claiming acquaintance is $\alpha^{2A_{ij}}\beta^{2(1-A_{ij})}$, the probability of one but not the other doing so is $2[\alpha(1-\alpha)]^{A_{ij}}[\beta(1-\beta)]^{1-A_{ij}}$, and the probability of neither is $(1-\alpha)^{2A_{ij}}(1-\beta)^{2(1-A_{ij})}$.

An interesting alternative formulation, proposed by Butts [9], considers the case in which some individuals are more reliable in the reports they give than others. Variations in reliability could arise simply because some people take surveys more seriously than others, are more cooperative survey subjects, or take more care with their responses. But they could also arise because people have different perceptions of what it means to be acquainted: one person could have a relatively relaxed view in which they consider people to be acquaintances even if they barely know them, while another could adopt a stricter definition.

Such a situation can be represented by a data model in which there is a separate true-positive rate α_i and false-positive rate β_i for each node i. Then the probability for instance of i saying they are friends with j but j saying they are not is $[\alpha_i(1-\alpha_j)]^{A_{ij}}[\beta_i(1-\beta_j)]^{1-A_{ij}}$, and similar expressions apply for other patterns of observations. Using this kind of model allows us to infer not only the structure of the underlying network but also the individual true- and false-positive rates, which themselves may reveal interesting behaviors—see Section V F.

Surveys of social networks are not the only context in which directed measurements of undirected networks arise. For instance, there have been many studies of messaging behavior within groups of people: who calls whom on the telephone, who emails whom, who sends text messages to whom, and so forth [65, 72–74]. One can hypothesize that observations such as phone calls or emails are a noisy measurement of an underlying network of acquaintance, and hence use models such as those described above to infer the structure of the network from the observed pattern of messages.

E. Networks with multiedges

Some ground-truth networks may be *multigraphs*, meaning that they contain multiedges. True multigraphs are rare in real-world applications, but there are many networks which, though composed of single edges only, may nonetheless be conveniently represented as multigraphs, for instance using the Poisson formulation of Section III C. How should we define a data model for a multigraph? There are a number of types of data a measurement of such a network could return. It could, for instance, return an estimate of the multiplicity of an edge. In our work, however, we make a simpler assumption, similar to that of Section IV B, in which measurements return only a yes-or-no answer that there either is or is not an edge at a given position. This situation is most completely represented by a model with an infinite set of parameters α_k, representing the probability that upon making a measurement of a particular pair of nodes we observe an edge between them if there are exactly k edges in the corresponding position in the ground-truth network. In practice, however, since multiedges are rare in the examples we consider, only the first two of these parameters are of interest: α_0, which is the false-positive rate, and α_1, which is roughly, though not exactly, the true-positive rate.

V. COMPLETE ALGORITHMS AND EXAMPLE APPLICATIONS

Building a complete algorithm for inferring network structure from noisy data involves combining a suitable network model with a suitable data model. There are many such combinations we can construct from the models introduced in the previous sections. Here we give a selection of examples, along with illustrative applications.

A. Random graphs and independent measurements

Perhaps the simplest example of our methods is the combination of the standard (Bernoulli) random graph of Section III A with the independent edge data model of Section IV A. This turns the problem of network reconstruction into a standard binary classification problem. We will go through this case in detail.

To derive the EM equations for this combination of models, we first take the probability $P(\mathbf{A}|\omega)$ for the random graph from Eq. (13) and the uniform prior probability $P(\omega) = 1$ and substitute them into Eq. (8). Performing the derivative with respect to the single parameter ω, we get

$$\sum_{\mathbf{A}} q(\mathbf{A}) \sum_{i<j} \left[\frac{A_{ij}}{\omega} - \frac{1 - A_{ij}}{1 - \omega} \right] = 0. \tag{25}$$

Swapping the order of the summations and defining

$$Q_{ij} = \sum_{\mathbf{A}} q(\mathbf{A}) A_{ij}, \tag{26}$$

we find that

$$\omega = \frac{1}{\binom{n}{2}} \sum_{i<j} Q_{ij}, \tag{27}$$

where n is the number of nodes in the network, as previously.

The quantity Q_{ij} is equal to the posterior probability that there is an edge between nodes i and j—it is our estimate of the ground truth for this node pair given the observed data. Q_{ij} can be thought of as a generalization of the adjacency matrix. When it is exactly zero or one it has the same meaning as the adjacency matrix element A_{ij}: there definitely either is or is not a ground-truth edge between nodes i and j. For other values between zero and one it interpolates between these limits, quantifying our certainty about whether the edge exists. Equation (27) thus has the simple interpretation that the probability ω of an edge in our network is the average of the probabilities of the individual edges.

Turning to the data model, a crucial point to notice is that if measurements of different edges are truly independent, so that an observation (or not) of an edge between one node pair tells you nothing about any other node pair, then single measurements of node pairs are not enough to estimate the parameters of the model. It is well known that you cannot estimate the error on a random variable by making only a single measurement. You have to make at least two measurements. In the present context, this means that at least some edges in the network must be measured more than once to obtain an estimate of the true- and false-positive rates α and β.

Let us assume that we make some number N_{ij} of measurements of node pair i, j. Each measurement returns a yes-or-no answer about whether the nodes are connected by an edge, but repeated measurements may not agree, precisely because the measurements are noisy. So suppose that out of the N_{ij} measurements we make, we observe an edge on E_{ij} of them, and no edge on the remaining $N_{ij} - E_{ij}$. Plugging these definitions into the data model of Section IV A, we can write the probability of this particular set of observations as $\alpha^{E_{ij}}(1-\alpha)^{N_{ij}-E_{ij}}$ if there is truly an edge between i and j, and $\beta^{E_{ij}}(1-\beta)^{N_{ij}-E_{ij}}$ if there is not. Taking the product over all distinct node pairs, the probability for the entire data set can then be written

$$P(D|\mathbf{A},\alpha,\beta) = \prod_{i<j} \bigl[\alpha^{E_{ij}}(1-\alpha)^{N_{ij}-E_{ij}}\bigr]^{A_{ij}} \times \bigl[\beta^{E_{ij}}(1-\beta)^{N_{ij}-E_{ij}}\bigr]^{1-A_{ij}}. \tag{28}$$

Taking the log and substituting into Eq. (9), assuming that the priors on α and β are uniform, we get

$$\sum_{\mathbf{A}} q(\mathbf{A}) \sum_{i<j} A_{ij} \left[\frac{E_{ij}}{\alpha} - \frac{N_{ij} - E_{ij}}{1 - \alpha} \right] = 0, \tag{29}$$

$$\sum_{\mathbf{A}} q(\mathbf{A}) \sum_{i<j} (1 - A_{ij}) \left[\frac{E_{ij}}{\beta} - \frac{N_{ij} - E_{ij}}{1 - \beta} \right] = 0, \tag{30}$$

which can be rearranged to give

$$\alpha = \frac{\sum_{i<j} Q_{ij} E_{ij}}{\sum_{i<j} Q_{ij} N_{ij}}, \qquad \beta = \frac{\sum_{i<j} (1 - Q_{ij}) E_{ij}}{\sum_{i<j} (1 - Q_{ij}) N_{ij}}, \tag{31}$$

where Q_{ij} is as in Eq. (26) again.

It remains to calculate the value of Q_{ij}, which we do from Eq. (5). Combining Eqs. (1), (13), and (28) and substituting into (5), we find the following expression for $q(\mathbf{A})$:

$$\begin{aligned} q(\mathbf{A}) &= \frac{\prod_{i<j} \bigl[\omega\alpha^{E_{ij}}(1-\alpha)^{N_{ij}-E_{ij}}\bigr]^{A_{ij}} \bigl[(1-\omega)\beta^{E_{ij}}(1-\beta)^{N_{ij}-E_{ij}}\bigr]^{1-A_{ij}}}{\sum_{\mathbf{A}} \prod_{i<j} \bigl[\omega\alpha^{E_{ij}}(1-\alpha)^{N_{ij}-E_{ij}}\bigr]^{A_{ij}} \bigl[(1-\omega)\beta^{E_{ij}}(1-\beta)^{N_{ij}-E_{ij}}\bigr]^{1-A_{ij}}} \\ &= \prod_{i<j} \frac{\bigl[\omega\alpha^{E_{ij}}(1-\alpha)^{N_{ij}-E_{ij}}\bigr]^{A_{ij}} \bigl[(1-\omega)\beta^{E_{ij}}(1-\beta)^{N_{ij}-E_{ij}}\bigr]^{1-A_{ij}}}{\omega\alpha^{E_{ij}}(1-\alpha)^{N_{ij}-E_{ij}} + (1-\omega)\beta^{E_{ij}}(1-\beta)^{N_{ij}-E_{ij}}}. \end{aligned} \tag{32}$$

Then

$$Q_{ij} = \sum_{\mathbf{A}} q(\mathbf{A}) A_{ij} = \frac{\omega \alpha^{E_{ij}} (1-\alpha)^{N_{ij}-E_{ij}}}{\omega \alpha^{E_{ij}} (1-\alpha)^{N_{ij}-E_{ij}} + (1-\omega)\beta^{E_{ij}} (1-\beta)^{N_{ij}-E_{ij}}}. \tag{33}$$

The posterior distribution $q(\mathbf{A})$ can be conveniently rewritten in terms of Q_{ij} as

$$q(\mathbf{A}) = \prod_{i<j} Q_{ij}^{A_{ij}} (1-Q_{ij})^{1-A_{ij}}. \tag{34}$$

In other words, the probability distribution over networks is (in this special case) simply the product of independent Bernoulli distributions of the individual edges, with Bernoulli parameters Q_{ij}.

The complete EM algorithm now consists of the iteration of Eqs. (27), (31), and (33) from suitably chosen starting conditions until convergence. Typically one chooses random values of ω, α, and β for the initial conditions and proceeds from there.

Once the algorithm has converged we can estimate network quantities of interest using Eqs. (11) and (12). As a simple example, consider the average degree c of a node in the network. For a known network with adjacency matrix $\mathbf{A}$ the average degree is given by $c = (1/n)\sum_{ij} A_{ij}$. The mean (expected) value of the average degree given our posterior distribution $q(\mathbf{A})$ is thus

$$\begin{aligned} \mu_c &= \sum_{\mathbf{A}} q(\mathbf{A}) \frac{1}{n} \sum_{ij} A_{ij} = \frac{1}{n} \sum_{ij} \sum_{\mathbf{A}} q(\mathbf{A}) A_{ij} \\ &= \frac{1}{n} \sum_{ij} Q_{ij}. \end{aligned} \tag{35}$$

The estimated variance about this value is given by Eq. (12) to be

$$\begin{aligned} \sigma_c^2 &= \sum_{\mathbf{A}} q(\mathbf{A}) \biggl[\frac{1}{n} \sum_{ij} A_{ij} - \mu_c \biggr]^2 \\ &= \frac{1}{n^2} \sum_{\mathbf{A}} q(\mathbf{A}) \sum_{ijkl} A_{ij} A_{kl} - \mu_c^2 \\ &= \frac{1}{n^2} \sum_{ij} Q_{ij}(1-Q_{ij}). \end{aligned} \tag{36}$$

The approach of this section generalizes straightforwardly to the variant random graph of Section III C in which there is a Poisson distributed number of edges between each pair of nodes and the network can contain multiedges. Taking Eq. (16) and substituting into (8) we get

$$\sum_{\mathbf{A}} q(\mathbf{A}) \sum_{ij} \biggl[\frac{A_{ij}}{\omega} - 1 \biggr] = 0. \tag{37}$$

Here we have again assumed a uniform prior on ω, which is not strictly allowed in this case, since ω has an infinite range from 0 to ∞. One can, however, assume a uniform prior over a finite range and then make that range large enough to encompass the solution for ω.

Rearranging Eq. (37) for ω now gives

$$\begin{aligned} \omega &= \frac{1}{n^2} \sum_{\mathbf{A}} q(\mathbf{A}) \sum_{ij} A_{ij} = \frac{1}{n^2} \sum_{ij} \sum_{\mathbf{A}} q(\mathbf{A}) \sum_{k=0}^{\infty} k \delta_{k,A_{ij}} \\ &= \frac{1}{n^2} \sum_{ij} \sum_{k=0}^{\infty} k Q_{ij}(k), \end{aligned} \tag{38}$$

where

$$Q_{ij}(k) = \sum_{\mathbf{A}} q(\mathbf{A}) \delta_{k,A_{ij}} \tag{39}$$

is the posterior probability that there are exactly k edges between nodes i and j (or $\frac{1}{2}k$ edges when $i = j$). Alternatively, and perhaps more conveniently, we can write the estimated value of A_{ij} as

$$\hat{A}_{ij} = \sum_{k=0}^{\infty} k Q_{ij}(k), \tag{40}$$

in which case

$$\omega = \frac{1}{n^2} \sum_{ij} \hat{A}_{ij}. \tag{41}$$

As discussed in Section IV E, we will assume that, multiedges notwithstanding, measurements on node pairs i, j continue to return yes-or-no answers about the presence of an edge, with α_k being the probability of a yes if there are k ground-truth edges between i and j. Let E_{ij} represent the number of yeses out of a total of N_{ij} measurements, except for self-edges, for which the most natural definition is that E_{ii} represents twice the number of yeses and N_{ii} twice the number of measurements, by analogy with the definition of the adjacency matrix.

With these definitions, the equivalent of Eq. (28) for this model is

$$\begin{aligned} P(D|\mathbf{A},\alpha) = &\prod_{i<j} \prod_{k=0}^{\infty} \bigl[\alpha_k^{E_{ij}} (1-\alpha_k)^{N_{ij}-E_{ij}} \bigr]^{\delta_{k,A_{ij}}} \\ &\times \prod_{i} \prod_{k=0}^{\infty} \bigl[\alpha_k^{E_{ii}/2} (1-\alpha_k)^{(N_{ii}-E_{ii})/2} \bigr]^{\delta_{k,A_{ii}}}. \end{aligned} \tag{42}$$

Taking the log, substituting into Eq. (9), and assuming that the priors on the α_k are uniform, we then get

$$\sum_{\mathbf{A}} q(\mathbf{A}) \sum_{ij} \delta_{k,A_{ij}} \biggl[\frac{E_{ij}}{\alpha_k} - \frac{N_{ij}-E_{ij}}{1-\alpha_k} \biggr] = 0 \tag{43}$$

for all $k = 0 \ldots \infty$. Rearranging for α_k, we get

$$\alpha_k = \frac{\sum_{ij} Q_{ij}(k) E_{ij}}{\sum_{ij} Q_{ij}(k) N_{ij}}, \tag{44}$$

where $Q_{ij}(k)$ is defined in Eq. (39).

Following similar lines of argument to those for the Bernoulli model, Eq. (5) now tells us that the posterior distribution over networks $\mathbf{A}$ is

$$q(\mathbf{A}) = \prod_{i<j} \frac{\omega^{A_{ij}}/A_{ij}!\big[\alpha_{A_{ij}}^{E_{ij}}(1-\alpha_{A_{ij}})^{N_{ij}-E_{ij}}\big]}{\sum_{k=0}^{\infty}\omega^k/k!\big[\alpha_k^{E_{ij}}(1-\alpha_k)^{N_{ij}-E_{ij}}\big]} \prod_i \frac{(\frac{1}{2}\omega)^{A_{ii}/2}/(\frac{1}{2}A_{ii})!\big[\alpha_{A_{ii}}^{E_{ii}/2}(1-\alpha_{A_{ii}})^{(N_{ii}-E_{ii})/2}\big]}{\sum_{r=0}^{\infty}(\frac{1}{2}\omega)^r/r!\big[\alpha_{2r}^{E_{ii}/2}(1-\alpha_{2r})^{(N_{ii}-E_{ii})/2}\big]} = \prod_{i\le j} Q_{ij}(A_{ij}). \tag{45}$$

Then

$$Q_{ij}(k) = \frac{\omega^k/k!\big[\alpha_k^{E_{ij}}(1-\alpha_k)^{N_{ij}-E_{ij}}\big]}{\sum_{k=0}^{\infty}\omega^k/k!\big[\alpha_k^{E_{ij}}(1-\alpha_k)^{N_{ij}-E_{ij}}\big]} \tag{46}$$

for $i \neq j$ and

$$Q_{ii}(k) = \frac{(\frac{1}{2}\omega)^{k/2}/(\frac{1}{2}k)!\big[\alpha_k^{E_{ii}/2}(1-\alpha_k)^{(N_{ii}-E_{ii})/2}\big]}{\sum_{r=0}^{\infty}(\frac{1}{2}\omega)^r/r!\big[\alpha_{2r}^{E_{ii}/2}(1-\alpha_{2r})^{(N_{ii}-E_{ii})/2}\big]}. \tag{47}$$

In the common case of network that does not actually have any self-edges, however, one would not normally attempt to measure their presence, so $N_{ii} = E_{ii} = 0$ for all i and the latter expression simplifies to

$$Q_{ii}(k) = \frac{(\frac{1}{2}\omega)^{k/2}}{(\frac{1}{2}k)!}\mathrm{e}^{-\omega/2}, \tag{48}$$

which is simply the prior distribution on self-edges assuming the random graph model. In practice, for sparse networks where ω is small, it will often be an adequate approximation to simply set $Q_{ii}(0) = 1$ for all i and $Q_{ii}(k) = 0$ for $k > 0$, implying that there are no self-edges, which is true.

In theory, the evaluation of $Q_{ij}(k)$ from Eq. (46) requires us to first calculate all of the parameters α_k, of which there are an infinite number, in order to evaluate the denominator. In practice, however, most networks, as we have said, are sparse, having small values of ω, which means that all but the first two terms in the denominator can be neglected and only α_0 and α_1 need be calculated (which represent approximately the false-positive and true-positive rates for this data model). This in turn means that $Q_{ij}(k)$ is negligible for $k \geq 2$, so that $Q_{ij}(0) \simeq 1 - Q_{ij}(1)$. Thus we only really need to calculate one probability $Q_{ij}(1)$ for each node pair:

$$Q_{ij}(1) \simeq \frac{\omega\alpha_1^{E_{ij}}(1-\alpha_1)^{N_{ij}-E_{ij}}}{\alpha_0^{E_{ij}}(1-\alpha_0)^{N_{ij}-E_{ij}} + \omega\alpha_1^{E_{ij}}(1-\alpha_1)^{N_{ij}-E_{ij}}}, \tag{49}$$

which represents, roughly speaking, the probability that there is an edge between i and j, which is also (approximately) the expected value of the corresponding adjacency matrix element $\hat{A}_{ij} \simeq Q_{ij}(1)$.

B. Configuration model with independent measurements

The developments of the previous section can be extended in a straightforward manner to the more realistic configuration model introduced in Section III E. Substituting Eq. (23) into Eq. (8) and differentiating with respect to ω gives

$$\begin{aligned}\omega &= \frac{1}{n^2}\sum_{\mathbf{A}} q(\mathbf{A}) \sum_{ij} A_{ij} = \frac{1}{n^2}\sum_{ij}\sum_{k=0}^{\infty} k Q_{ij}(k) \\ &= \frac{1}{n^2}\sum_{ij}\hat{A}_{ij},\end{aligned} \tag{50}$$

just as in Eqs. (38) and (41), with $Q_{ij}(k) = \sum_{\mathbf{A}} q(\mathbf{A})\delta_{k,A_{ij}}$ as before and $\hat{A}_{ij} = \sum_k k Q_{ij}(k)$ being the estimated value of A_{ij}, Eq. (40). At the same time, differentiating with respect to ϕ_i, while enforcing the normalization condition (20) with a Lagrange multiplier, gives

$$\phi_i = n\frac{\sum_j \hat{A}_{ij}}{\sum_{ij}\hat{A}_{ij}}. \tag{51}$$

The equations for the data model parameters α_k are unchanged from the previous section, with α_k still being given by Eq. (44). And the posterior distribution over networks is once again given $q(\mathbf{A}) = \prod_{i\le j} Q_{ij}(A_{ij})$ but now with

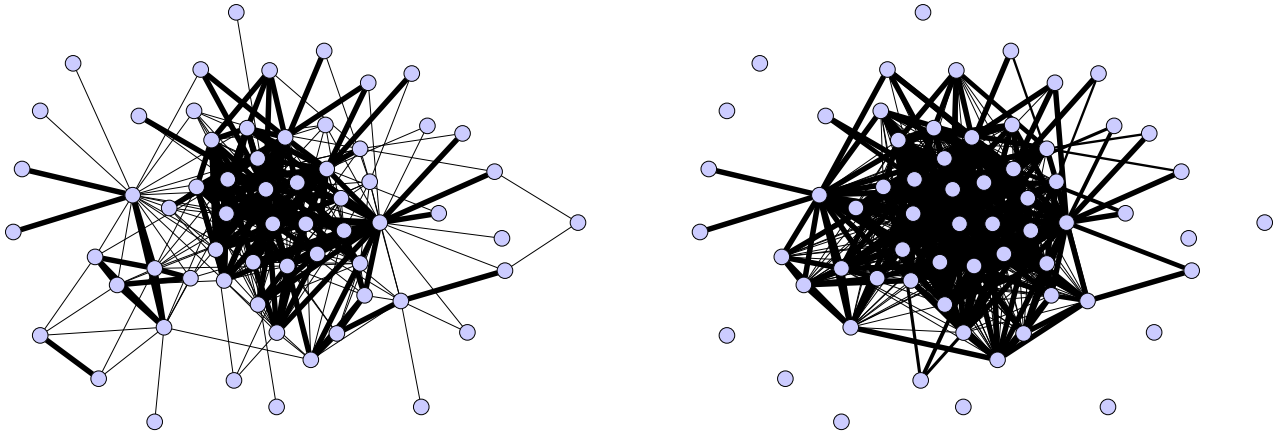
$$Q_{ij}(k) = \frac{(\omega\phi_i\phi_j)^k/k!\big[\alpha_k^{E_{ij}}(1-\alpha_k)^{N_{ij}-E_{ij}}\big]}{\sum_{k=0}^{\infty}(\omega\phi_i\phi_j)^k/k!\big[\alpha_k^{E_{ij}}(1-\alpha_k)^{N_{ij}-E_{ij}}\big]}, \tag{52}$$

and $Q_{ii}(k) = \mathrm{e}^{-\omega\phi_i^2/2}(\frac{1}{2}\omega\phi_i^2)^{k/2}/(\frac{1}{2}k)!$.

If we make the same assumption as we made at the end of the previous section, that ω is small and hence only the first two terms in the denominator of Eq. (52) need be included, then one need only calculate the quantities

$$Q_{ij}(1) \simeq \frac{\omega\phi_i\phi_j\alpha_1^{E_{ij}}(1-\alpha_1)^{N_{ij}-E_{ij}}}{\alpha_0^{E_{ij}}(1-\alpha_0)^{N_{ij}-E_{ij}} + \omega\phi_i\phi_j\alpha_1^{E_{ij}}(1-\alpha_1)^{N_{ij}-E_{ij}}}, \tag{53}$$

and $Q_{ij}(0) \simeq 1 - Q_{ij}(1)$, $\hat{A}_{ij} \simeq Q_{ij}(1)$ (and $Q_{ii}(0) \simeq 1$).



(a) (b)

FIG. 1: Two examples of inferred networks of connections between a subset of the participants in the "reality mining" study of Eagle and Pentland [75, 76]. (a) Network inferred using the Poisson random graph for the network model and the independent edges model of Section IV A for the data model. Edges with probability less than 0.01 are omitted, as are nodes that have no edges with probability 0.01 or greater. (b) Network inferred from the same data, but using the configuration model as the network model. For ease of comparison, the same set of nodes is shown in panel (b) as in panel (a), with each in the same spatial position.

C. Example application

As an example of the application of these methods, we turn to a data set we examined previously in [43]. The data come from the "reality mining" study of Eagle and Pentland [75, 76] and describe the interactions of a group of 96 university students. The goal of the study was to determine whether one could reconstruct networks of acquaintance—who actually knows whom—from data on physical proximity. Students in the study carried mobile phones equipped with special software that used Bluetooth radio technology to record when two of the phones where in close proximity with one another (a few meters). It is reasonable to suppose that people who are acquainted will sometimes be in close proximity, but it is also certainly the case that some acquaintances are rarely or never in proximity and that people may be in proximity and not be acquainted. You might sit next to someone on the bus, for example, or stand next to them in the line at the supermarket without ever knowing who they are. Thus proximity is a noisy measurement of acquaintance of exactly the kind considered here.

The study by Eagle and Pentland recorded detailed, time-resolved instances of pairwise proximity between participants over a period of several months in 2004 and 2005, but the data we study cover a smaller interval, being taken from eight consecutive Wednesdays in March and April of 2005. We limit ourselves to Wednesdays in order to factor out the (large) weekly variation in proximity patterns: lower rates of proximity are observed at weekends than on weekdays for instance. We also amalgamate all observations for each Wednesday into a single measurement: we consider two people to be observed together on a particular day if they are measured to be in proximity at any time during that day. The result is eight separate measurements of proximity for each pair of individuals. In the nomenclature of our models, $N_{ij} = 8$ for all i, j, and E_{ij} can take integer values from 0 up to 8. All possible values in this range are observed in the data.

Figure 1a shows what happens when we apply the algorithm of Eqs. (44) and (49) to these data. This algorithm assumes a simple random graph for the network model and (conditionally) independent edge measurements for the data model. The figure shows the resulting inferred network with edge thicknesses representing the posterior probabilities of the edges. As we can see there is a well connected core of about twenty nodes in the center of the picture, surrounded by a periphery with weaker connections. The thickest lines in the figure represent edges with probability of almost 1, while the thinnest represent edges with probability less than 0.1. Edges with probability less than 0.01 are omitted from the figure, as are nodes that have no connections above this threshold.

Figure 1b shows the same data analyzed using the algorithm of Eqs. (50), (51), and (53), which uses the configuration model as its network model. As we can see, this produces some changes in the inferred edge probabilities. Overall many of the same edges get high or low probability in both models but the configuration-model based algorithm gives a more "decisive" result than its random-graph counterpart, mostly assigning either very high or very low probabilities to edges, meaning that it is more certain whether edges do or do not exist.

One way to think about the configuration model in this

context is that it introduces correlations between edges that are not present when one uses the simple random graph. The presence of edges attached to a particular node i increases the inferred value of the node parameter ϕ_i via Eq. (51), and this in turn increases the probability of other edges being attached to the same the node via Eq. (53). Thus the presence of one edge makes the presence of another more likely.

One might ask which is the better of the two network models: the random graph or the configuration model? In fact, despite the visual differences between the networks in Fig. 1, the two do not differ very greatly. If we take the maximum-probability structure predicted by each algorithm (which is equivalent to assuming an edge to exist whenever $Q_{ij} > \frac{1}{2}$ and not otherwise), then we find that the two calculations agree on essentially all node pairs *except* those for which $E_{ij} = 1$, i.e., those for which proximity was observed on only one of the eight days of observation. For these pairs the random-graph-based calculation always concludes that the corresponding edge does not exist, whereas the configuration model sometimes says it does and sometimes says it doesn't. Thus the primary contribution of the configuration model in this instance is to give us more sensitivity in the case of node pairs with particularly sparse observations.

More generally, the configuration model is considered to be the more accurate model in most real-world cases since it allows for realistic non-Poisson degree distributions similar to those seen in empirical networks [61]. Nonetheless, there maybe cases where the ordinary random graph is justified; which model one uses depends in the end on the assumptions one makes about the nature of the network. One might perhaps consider this a problem with the method. Other network reconstruction methods do not require one to make assumptions in this way. As discussed at the start of Section II, however, we would argue that these other methods are still making assumptions, though they may not be explicitly acknowledged. We feel that the approach we propose is preferable in that it requires us to make our assumptions clear and allows us to see directly what effect they have on the results.

Note also that our algorithms cannot make any statement about what network it is exactly that is represented in pictures like Fig. 1. Is it a network of who is friends with whom? Who knows whom? Who works with whom? The algorithm does not say. All we can say is that this is our best estimate of whatever network it is that is driving the observations. In the present case it is probably some amalgam of friendship, students who work together, students who go to class together, and so forth. "Acquaintance" might be a good umbrella term for this set of interactions, but in the end the network is most correctly defined as that network which causes people to be in proximity with one another.

Once we have the posterior distribution over networks, we can estimate any other network quantity of interest from it using Eqs. (11) and (12). For instance, we can calculate the mean degree c of the network, Eqs. (35) and (36). Using the configuration model version of our calculation and approximating Q_{ij} by $Q_{ij}(1)$ we find that $c = 5.55 \pm 0.05$. By contrast, a naive estimate of the mean degree, derived by simply aggregating all proximity observations for the eight days of data, gives $c = 6.23$, which is of the same order of magnitude, but nonetheless in significant disagreement (and lacking any error estimate). A more conservative estimate, in which we assume an edge only if there are proximity observations between a given pair of nodes on two or more days, gives a lower value of $c = 3.00$, again in substantial disagreement with our estimate from the posterior distribution.

Running time for the calculation is minimal, varying from a fraction of a second to a few seconds depending on model details and the programming language used. More generally, since the calculation requires the evaluation of $\mathrm{O}(n^2)$ probabilities Q_{ij}, we expect the running time to scale at least as n^2. Network algorithms running in $\mathrm{O}(n^2)$ time are typically feasible (with patience) for networks of up to hundreds of thousands or perhaps millions of nodes, putting our methods within reach for many large network data sets, though not, in their current form, for the very largest (there exist examples with billions of nodes or more).

D. Multimodal data

For our second example we consider the case discussed in Section IV C of a network whose edges are observed using a number of different methods or modes, labeled by $m = 1 \ldots M$. We will consider specifically a directed network, both to give an explicit example of an algorithm for directed edges and because it will be useful in Section V E, where we will apply the method to a directed data set. By convention, directed networks are represented by an adjacency matrix in which $A_{ij} = 1$ if there is an edge *from* node j *to* node i.

We will assume that the prior probability of any directed edge is ω as previously. Then, assuming a simple network in which there are no multiedges or self-edges, we have

$$P(\mathbf{A}|\omega) = \prod_{i \neq j} \omega^{A_{ij}} (1-\omega)^{1-A_{ij}}, \tag{54}$$

which is a trivial generalization of Eq. (13). Following the same line of argument as in Eq. (25) we can then show that

$$\omega = \frac{1}{n(n-1)} \sum_{i \neq j} Q_{ij}, \tag{55}$$

where $Q_{ij} = \sum_{\mathbf{A}} q(\mathbf{A}) A_{ij}$ is the posterior probability of a directed edge from node j to node i.

Now let the true- and false-positive rates for observations in mode m be α_m and β_m respectively, as described

in Section IV C. And let $N_{ij}^{(m)}$ be the number of measurements made of the presence or absence of an edge from j to i (usually zero or one, but other values are possible in principle) and $E_{ij}^{(m)}$ be the number of times an edge is in fact observed. Then the likelihood of the data D given the ground-truth network and parameters is

$$P(D|\mathbf{A},\alpha,\beta) = \prod_{i\neq j}\Biggl[\prod_{m=1}^{M} \alpha_m^{E_{ij}^{(m)}} (1-\alpha_m)^{N_{ij}^{(m)}-E_{ij}^{(m)}}\Biggr]^{A_{ij}} \times \Biggl[\prod_{m=1}^{M} \beta_m^{E_{ij}^{(m)}} (1-\beta_m)^{N_{ij}^{(m)}-E_{ij}^{(m)}}\Biggr]^{1-A_{ij}}. \tag{56}$$

Here we are assuming that measurements made in different modes are statistically independent, so that the probability of observing any given edge in any given combination of modes is a product over the probabilities of the individual modes. In the language of machine learning such an approach is called a naive Bayes classifier.

Following the same line of argument as in Eq. (28), we can then show that

$$\alpha_m = \frac{\sum_{i\neq j} Q_{ij} E_{ij}^{(m)}}{\sum_{i\neq j} Q_{ij} N_{ij}^{(m)}}, \qquad \beta_m = \frac{\sum_{i\neq j} (1-Q_{ij}) E_{ij}^{(m)}}{\sum_{i\neq j} (1-Q_{ij}) N_{ij}^{(m)}}, \tag{57}$$

while the equivalent of Eq. (33) for Q_{ij} is

$$Q_{ij} = \frac{\omega \prod_m \alpha_m^{E_{ij}^{(m)}} (1-\alpha_m)^{N_{ij}^{(m)}-E_{ij}^{(m)}}}{\omega \prod_m \alpha_m^{E_{ij}^{(m)}} (1-\alpha_m)^{N_{ij}^{(m)}-E_{ij}^{(m)}} + (1-\omega) \prod_m \beta_m^{E_{ij}^{(m)}} (1-\beta_m)^{N_{ij}^{(m)}-E_{ij}^{(m)}}}. \tag{58}$$

The EM algorithm now consists of the iteration of Eqs. (55), (57), and (58) from suitable starting values to convergence.

E. Example application

As an example of this algorithm, we consider an ecological network, a food web of predator-prey interactions between species. The specific example we look at is the early Eocene Messel Shale food web of Dunne *et al.* [77], a prehistoric food web of exactly $n = 700$ extinct taxa and their patterns of predation, reconstructed from paleontological evidence. Like many food webs, this one is pieced together from data derived from a variety of sources. In this case, the authors used ten different types of evidence to establish links between taxa, including such things as gut contents (the digested remains of one species were found in the fossilized gut of another), stratigraphic co-occurrence (evidence of interaction is present in one or more other nearby contemporaneous fossil deposits), or body size (larger animals eat smaller ones, so a difference in body sizes can suggest a predator-prey interaction).

What is particularly interesting about this data set for our purposes is that Dunne *et al.* made available not only the final form of the network but the details of which particular modes were observed for each edge in the network—gut contents, body size, etc. Thus the data set has exactly the "multimodal" form we considered in the previous section.

Of the ten modes of observation used by Dunne *et al.* one of them—"taxonomic uniformity"—is seen in virtually all edges (6126 edges out of a total of 6444, or 95%), which means in practice that it communicates almost no information. So we discard it, leaving $M = 9$ remaining modes of measurement in the data set. Each mode is listed as either observed or not for each edge in the network, meaning in effect that $N_{ij}^{(m)} = 1$ for all i, j and all m, and $E_{ij}^{(m)}$ is either zero or one. Applying Eqs. (55), (57), and (58) to these data and iterating to convergence, we then arrive at values for the true- and false-positive rates in each mode and probabilities Q_{ij} for the directed edges.

In addition to the data set itself, Dunne *et al.* published their own judgments about the structure of the network. For each edge that was observed in at least one of the modes they assigned a score, based on the data, indicating how confident they were that the edge in question actually exists in the network. They used a three-valued scale to say whether they judged there to be high, medium, or low certainty about each edge. If we similarly divide the edges of our inferred network into three categories, arbitrarily defining high certainty to be $Q_{ij} > 0.9$, low certainty to be $Q_{ij} < 0.1$, and medium certainty to be everything else, we find that our EM algorithm is able to reproduce the assessments of Dunne *et al.* for 5446 of the 6444 observed edges, or 84.5%. For comparison, a random guess would get only 33% correct. (Our choice

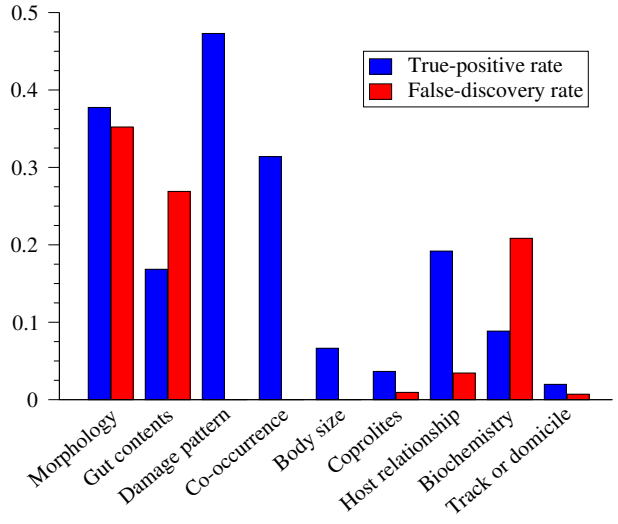


FIG. 2: Inferred values of the true-positive and false-discovery rates for each of the nine types of evidence considered in our analysis of the Messel Shale food web data set of Dunne *et al.* [77]. High true-positive rate indicates a type of evidence that is frequently observed when there truly is a link in the network; high false-discovery rate indicates that many observations of links using this type of evidence are in error.

of 0.1 and 0.9 for the cut-off lines is not based on any rigorous principle, and it is possible to get better agreement with the assessments of Dunne *et al.* by tuning the values carefully. Even the rough calculation presented here, however, shows that the EM algorithm is capable of extracting real insight from the data.)

In addition to the network itself, the inferred values of the parameters α_m and β_m are also of interest. Because this network (like most others studied in network science) is very sparse, all the β_m are small. To make them easier to interpret we reparametrize them in terms of the *false-discovery rate*, which is the probability that an observation of an edge is wrong. Applying Bayes' rule, the false-discovery rate for observations in mode m is given by

$$\begin{aligned} P(A_{ij}^{(m)} = 0 | E_{ij}^{(m)} = 1) & \\ &= P(E_{ij}^{(m)} = 1 | A_{ij}^{(m)} = 0) \frac{P(A_{ij}^{(m)} = 0)}{P(E_{ij}^{(m)} = 1)} \\ &= \frac{(1-\omega)\beta_m}{\omega\alpha_m + (1-\omega)\beta_m}. \end{aligned} \tag{59}$$

Figure 2 shows the estimated true-positive and false-discovery rates for each of the nine measurement modes. A high true-positive rate for a mode means that when an edge truly exists we will typically see evidence in this mode. A high false-discovery rate means that observations in this mode cannot be trusted because they are frequently false alarms.

The figure reveals that none of the modes of observation has a particularly high estimated true-positive rate—none is above 50%—but that "damage pattern" (evidence of damage to prey by predators) is, overall, the best of the bunch, having a true-positive rate of 47.3% and a zero false-discovery rate, meaning that if this mode is observed it is a reliable indication of predation. The latter initially appears to be an interesting and informative statement (it also applies to the "co-occurrence" and "body size" modes), but in fact it is not as useful as it might at first seem. The zero false-discovery rate occurs because this type of observation is usually seen in concert with other modes of evidence for the same edge, which together cause the algorithm to (correctly) conclude that the edge is present with high probability. Thus we can indeed reliably infer that an edge is present when this mode is observed, but even in the absence of this mode we would probably infer the same thing in most cases.

Among the other modes, "coprolites" and "track or domicile" have the lowest true-positive rates. And, perhaps surprisingly, "gut contents," which Dunne *et al.* consider the gold standard for establishing predation, is relatively poor on both measures, with a true-positive rate of only 16.8% (meaning observations in this mode are rare) and a false-discovery rate of 26.9% (meaning over a quarter of observations of this type turn out to be wrong). This makes gut contents the second-most unreliable mode of observation, after "morphology." The explanation for this result is essentially the opposite of that for "damage pattern" above: in a significant fraction of cases gut contents is the only type of observation in favor of an interaction. If an interaction truly exists then the probability that none of the other modes of evidence would be observed is low. When no other modes are observed, therefore, the algorithm concludes that there is a chance that the interaction does not in fact exist, and hence that the gut contents data are not wholly reliable.

F. Individual node errors

For our final example we consider the model of Section IV D in which the network is undirected but observations of it are directed and there are individual and potentially different error rates for each node. This model is particularly appropriate for acquaintance network data.

Suppose we have a social network of friendship or acquaintance and the structure of the network is measured by surveying people and asking them who their friends are. We use the configuration model as our network model, with the parameters ω and ϕ_i being given once again by Eqs. (50) and (51). For our data model we use a variant of the approach described in Section IV D and define α_{ik} to be the probability that individual i identifies another individual as a friend if there are k (undirected) edges between them in the ground-truth network. Then

the data likelihood given the ground truth is

$$P(D|\mathbf{A},\alpha) = \prod_{i\neq j}\prod_{k=0}^{\infty}\bigl[\alpha_{ik}^{E_{ij}}(1-\alpha_{ik})^{N_{ij}-E_{ij}}\bigr]^{\delta_{k,A_{ij}}} \times \prod_{i}\prod_{k=0}^{\infty}\bigl[\alpha_{ik}^{E_{ii}/2}(1-\alpha_{ik})^{(N_{ii}-E_{ii})/2}\bigr]^{\delta_{k,A_{ii}}}, \tag{60}$$

where E_{ij} is the number of times (out of N_{ij} total) that i identifies j as a friend (which under normal circumstances will be either zero or one) or twice that number when $i=j$.

Following the same lines of argument as previously, we then find that

$$\alpha_{ik} = \frac{\sum_j Q_{ij}(k)E_{ij}}{\sum_j Q_{ij}(k)N_{ij}}, \qquad Q_{ij}(k) = \frac{(\omega\phi_i\phi_j)^k/k!\bigl[\alpha_{ik}^{E_{ij}}(1-\alpha_{ik})^{N_{ij}-E_{ij}}\alpha_{jk}^{E_{ji}}(1-\alpha_{jk})^{N_{ji}-E_{ji}}\bigr]}{\sum_{k=0}^{\infty}(\omega\phi_i\phi_j)^k/k!\bigl[\alpha_{ik}^{E_{ij}}(1-\alpha_{ik})^{N_{ij}-E_{ij}}\alpha_{jk}^{E_{ji}}(1-\alpha_{jk})^{N_{ji}-E_{ji}}\bigr]}. \tag{61}$$

As with the model of Section V B, it will in the common case of a sparse network usually be adequate to compute only $Q_{ij}(1)$ and assume $Q_{ij}(0) = 1 - Q_{ij}(1)$ and $\hat{A}_{ij} = Q_{ij}(1)$, all other probabilities $Q_{ij}(k)$ with $k \geq 2$ being negligible.

G. Example application

As an example of this algorithm we consider data from the US National Longitudinal Study of Adolescent Health [78, 79], known colloquially as the "Add Health" study, a large-scale study of students in US middle and high schools conducted during the 1990s. Among other things the study asked students to identify their friends, but it was found that individuals often disagreed about friendships: as discussed in Section IV D, a substantial fraction of all claims of friendship are unreciprocated, implying a significant level of false positives, false negatives, or both in the data, and we can estimate these levels by applying our methods.

There were 84 populations surveyed in the Add Health study, where a population consisted of a high school and an associated feeder middle school. The populations ranged in size from dozens of students to thousands and the methods described here could be applied to any of them. In Fig. 3 we show results for a medium-sized population with 542 students. In this figure the widths of the edges once again vary to indicate the estimated probabilities Q_{ij}. In addition we vary the diameters of the nodes in proportion to the estimated degree parameter ϕ_i, which can be thought of as a measure of the sociability or popularity of individuals. We also vary the shades of the nodes to denote the reliability of their reports of friendships. As our measure of reliability we use the *precision*, which is the probability that a reported friendship is actually correct. As with the false-discovery rate of Section V E, an expression for the precision can be written using Bayes' rule:

$$P(A_{ij}=1|E_{ij}=1) = P(E_{ij}=1|A_{ij}=1)\frac{P(A_{ij}=1)}{P(E_{ij}=1)} = \frac{\omega\phi_i\phi_j\alpha_i}{\omega\phi_i\phi_j\alpha_i+\beta_i}. \tag{62}$$

The numbers we use to compute the shades of the nodes are the average value of this precision over all the friendships an individual reports.

The figure reveals a network with a dense core of strongly connected nodes (perhaps divided into two parts), plus a sparser periphery of more weakly connected nodes. Most nodes appear to have roughly the same value of ϕ_i (they appear about the same size), though a small subset seem to be "less sociable" (they appear smaller). Most nodes also have relatively low precision (lighter shades); only a handful, mostly in the interior of the figure, fall in the high-precision range (darker shades).

VI. CONCLUSIONS

In this paper we have developed in detail a class of expectation-maximization (EM) algorithms that allow one to infer the structure of an observed network from noisy, error-prone, or incomplete measurements. These algorithms take raw observational data concerning the structure of networks and return a posterior probability distribution over possible structures the network could take. This posterior distribution can then be used to estimate any other network quantity of interest along with the standard error on that estimate. In addition our algorithms also return values for a range of model parameters that quantify the mapping between the true structure of the network and the observed data, such as true- and false-positive rates for observation of individual edges. In many cases these parameters are of interest in their own right.

We have given three examples of practical applications

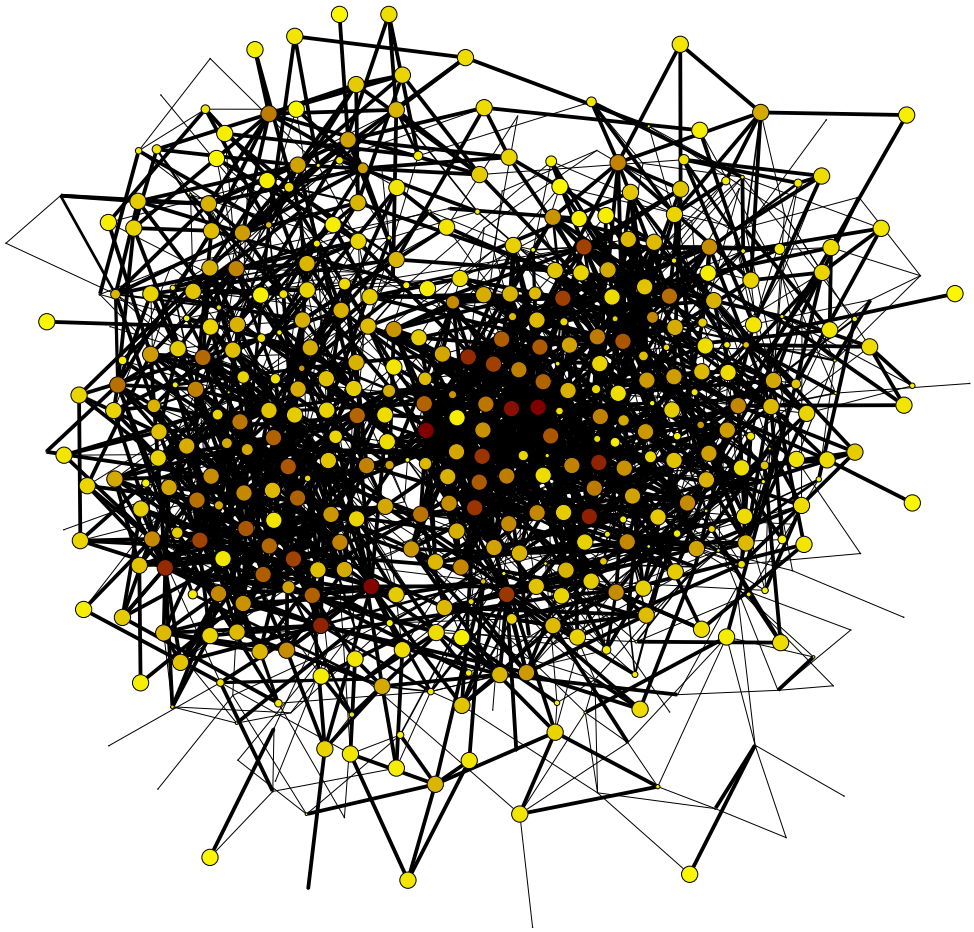


FIG. 3: The inferred network of friendships among the students in a medium-sized American high school and its feeder middle school. Edge thicknesses represent the inferred posterior probabilities of the edges. Node sizes represent the inferred values of the degree parameters ϕ_i for the configuration model. Shading of the nodes represents the estimated average precision of reports made by the corresponding individual, precision being the probability that a reported friendship actually exists. Lightest shades correspond to the lowest precision and darkest shades to the highest. Only edges that are reported to exist by at least one participant are shown, and nodes with no reported edges are omitted.

of our methods to previously published network data sets. In the first example we inferred the structure of a social network from repeated observations of physical proximity between pairs of people. In the second example we looked at a food web data set of predator-prey interactions among a group of species. Connections in the network are measured using a number of different techniques and, though none of techniques is very reliable, our methods allow us to combine them to make an estimate of the structure of the network. Our third example focused again on a social network, in this case of declared friendships between students in a US high school and middle school. In addition to allowing us to infer the structure of the friendship network, our algorithm in this case also gives us a measure of how accurately each student reports their own friendships.

One thing we have not done in this paper is look in detail at the case introduced in Section III D of networks that are generated from the stochastic block model (or its degree-corrected variant introduced in Section III E). Application of these models would allow us to simultaneously infer both the structure of the network and its division into communities. Recent work by Peixoto, appearing after the submission of this paper, describes one potential method for performing such a calculation. The interested reader is invited to look at Ref. [80].

Acknowledgments

The author thanks Elizabeth Bruch, George Cantwell, Jennifer Dunne, Travis Martin, Gesine Reinert, and Maria Riolo for useful comments. This work was funded in part by the US National Science Foundation under grants DMS–1407207 and DMS–1710848.

[1] M. Newman, *Networks*. Oxford University Press, Oxford, 2nd edition (2018).

[2] T. Ito, T. Chiba, R. Ozawa, M. Yoshida, M. Hattori, and Y. Sakaki, A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc. Natl. Acad. Sci. USA* **98**, 4569–4574 (2001).

[3] N. J. Krogan *et al.*, Global landscape of protein complexes in the yeast Saccharomyces cerevisiae. *Nature* **440**, 637–643 (2006).

[4] T. Rolland *et al.*, A proteome-scale map of the human interactome network. *Cell* **159**, 1212–1226 (2014).

[5] A. Lakhina, J. Byers, M. Crovella, and P. Xie, Sampling biases in IP topology measurements. In *Proceedings of the 22nd Annual Joint Conference of the IEEE Computer and Communications Societies*, Institute of Electrical and Electronics Engineers, New York (2003).

[6] D. Achlioptas, A. Clauset, D. Kempe, and C. Moore, On the bias of traceroute sampling. In *Proceedings of the 37th ACM Symposium on Theory of Computing*, Association of Computing Machinery, New York (2005).

[7] P. D. Killworth and H. R. Bernard, Informant accuracy in social network data. *Human Organization* **35**, 269–286 (1976).

[8] P. V. Marsden, Network data and measurement. *Annual Review of Sociology* **16**, 435–463 (1990).

[9] C. T. Butts, Network inference, error, and informant (in)accuracy: A Bayesian approach. *Social Networks* **25**, 103–140 (2003).

[10] M. S. Handcock and K. J. Gile, Modeling social networks from sampled data. *Annals of Applied Statistics* **4**, 5–25 (2010).

[11] D. Lusher, J. Koskinen, and G. Robins, *Exponential Random Graph Models for Social Networks: Theory, Methods, and Applications*. Cambridge University Press, Cambridge (2012).

[12] S. J. Wodak, S. Pu, J. Vlasblom, and B. Séraphin, Challenges and rewards of interaction proteomics. *Molecular & Cellular Proteomics* **8**, 3–18 (2009).

[13] A. Clauset and C. Moore, Accuracy and scaling phenomena in Internet mapping. *Phys. Rev. Lett.* **94**, 018701 (2005).

[14] S. L. Feld and W. C. Carter, Detecting measurement bias in respondent reports of personal networks. *Social Networks* **24**, 365–383 (2002).

[15] S. P. Borgatti, K. M. Carley, and D. Krackhardt, On the robustness of centrality measures under conditions of imperfect data. *Social Networks* **28**, 124–136 (2006).

[16] G. Kossinets, Effects of missing data in social networks. *Social Networks* **28**, 247–268 (2006).

[17] B. Karrer, E. Levina, and M. E. J. Newman, Robustness of community structure in networks. *Phys. Rev. E* **77**, 046119 (2008).

[18] D. J. Wang, X. Shi, D. A. McFarland, and J. Leskovec, Measurement error in network data: A reclassification. *Social Networks* **34**, 396–409 (2012).

[19] N. Erman and L. Todorovski, The effects of measurement error in case of scientific network analysis. *Scientometrics* **104**, 453–473 (2015).

[20] A. Clauset, C. Moore, and M. E. J. Newman, Hierarchical structure and the prediction of missing links in networks. *Nature* **453**, 98–101 (2008).

[21] R. Guimerà and M. Sales-Pardo, Missing and spurious interactions and the reconstruction of complex networks. *Proc. Natl. Acad. Sci. USA* **106**, 22073–22078 (2009).

[22] W.-X. Wang, Y.-C. Lai, and C. Grebogi, Data based identification and prediction of nonlinear and complex dynamical systems. *Physics Reports* **644**, 1–76 (2016).

[23] A. Y. Lokhov, M. Mzard, H. Ohta, and L. Zdeborová, Inferring the origin of an epidemic with a dynamic message-passing algorithm. *Phys. Rev. E* **90**, 012801 (2014).

[24] X. Li and X. Li, Reconstruction of stochastic temporal networks through diffusive arrival times. *Nature Communications* **8**, 15729 (2017).

[25] D. Liben-Nowell and J. Kleinberg, The link-prediction problem for social networks. *J. Assoc. Inf. Sci. Technol.* **58**, 1019–1031 (2007).

[26] M. Huisman, Imputation of missing network data: Some simple procedures. *Journal of Social Structure* **10**, 1–29 (2009).

[27] M. Kim and J. Leskovec, The network completion problem: Inferring missing nodes and edges in networks. In B. Liu, H. Liu, C. Clifton, T. Washio, and C. Kamath (eds.), *Proceedings of the 2011 SIAM International Conference on Data Mining*, pp. 47–58, Society for Industrial and Applied Mathematics, Philadephia, PA (2011).

[28] N. R. Smalheiser and V. I. Torvik, Author name disambiguation. *Annual Review of Information Science and Technology* **43**, 287–313 (2009).

[29] C. A. D'Angelo, C. Giuffrida, and G. Abramo, A heuristic approach to author name disambiguation in bibliometrics databases for large-scale research assessments. *J. Assoc. Inf. Sci. Technol.* **62**, 257–269 (2011).

[30] A. A. Ferreira, M. A. Goncalves, and A. H. F. Laender, A brief survey of automatic methods for author name disambiguation. *SIGMOD Record* **41**, 15–26 (2012).

[31] J. Tang, A. C. M. Fong, B. Wang, and J. Zhang, A unified probabilistic framework for name disambiguation in digital library. *IEEE Transactions on Knowledge and Data Engineering* **24**, 975–987 (2012).

[32] T. Martin, B. Ball, B. Karrer, and M. E. J. Newman, Coauthorship and citation patterns in the Physical Review. *Phys. Rev. E* **88**, 012814 (2013).

[33] G. M. Namata, S. Kok, and L. Getoor, Collective graph identification. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Association of Computing Machinery, New York (2011).

[34] X. Han, Z. Shen, W.-X. Wang, and Z. Di, Robust reconstruction of complex networks from sparse data. *Phys. Rev. Lett.* **114**, 028701 (2015).

[35] G. Casiraghi, V. Nanumyan, I. Scholtes, and F. Schweitzer, From relational data to graphs: Inferring significant links using generalized hypergeometric ensembles. In G. Ciampaglia, A. Mashhadi, and T. Yasseri (eds.), *Proceedings of the International Conference on Social Informatics (SocInfo 2017)*, number 10540 in Lecture Notes in Computer Science, pp. 111–120, Springer, Berlin (2017).

[36] M. T. Angulo, J. A. Moreno, G. Lippner, A.-L. Barabási, and Y.-Y. Liu, Fundamental limitations of network reconstruction from temporal data. *J. Roy. Soc. Interface* **14**, 20160966 (2017).

[37] J. Forster, I. Famili, P. Fu, B. O. Palsson, and J. Nielsen, Genome-scale reconstruction of the Saccharomyces cerevisiae metabolic network. *Genome Research* **13**, 244–253 (2003).

[38] Y. Liu, N. J. Liu, and H. Y. Zhao, Inferring protein-protein interactions through high-throughput interaction data from diverse organisms. *Bioinformatics* **21**, 3279–3285 (2005).

[39] J. Schafer and K. Strimmer, An empirical Bayes approach to inferring large-scale gene association networks. *Bioinformatics* **21**, 754–764 (2005).

[40] A. A. Margolin, I. Nemenman, K. Basso, C. Wiggins, G. Stolovitzky, R. Dalla Favera, and A. Califano, ARACNE: An algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinformatics* **7**, S7 (2006).

[41] P. Langfelder and S. Horvath, WGCNA: An R package for weighted correlation network analysis. *BMC Bioinformatics* **9**, 559 (2008).

[42] J. D. Allen, Y. Xie, M. Chen, L. Girard, and G. Xiao, Comparing statistical methods for constructing large scale gene networks. *PLOS One* **7**, e29348 (2012).

[43] M. E. J. Newman, Network structure from rich but noisy data. *Nature Physics* **14**, 542–545 (2018).

[44] T. Martin, B. Ball, and M. E. J. Newman, Structural inference for uncertain networks. *Phys. Rev. E* **93**, 012306 (2016).

[45] C. M. Le and E. Levina, Estimating a network from multiple noisy realizations. Preprint arxiv:1710.04765 (2017).

[46] I. Brugere, B. Gallagher, and T. Y. Berger-Wolf, Network structure inference, a survey: Motivations, methods, and applications. *ACM Computing Surveys* **1**, 1 (2016).

[47] S. L. Brunton, J. L. Proctor, and J. N. Kutz, Discovering governing equations from data by sparse identification of nonlinear dynamical systems. *Proc. Natl. Acad. Sci. USA* **113**, 3932–3937 (2016).

[48] J. Casadiego, M. Nitzan, S. Hallerberg, and M. Timme, Model-free inference of direct network interactions from nonlinear collective dynamics. *Nature Communications* **8**, 2192 (2017).

[49] P. W. Holland, K. B. Laskey, and S. Leinhardt, Stochastic blockmodels: Some first steps. *Social Networks* **5**, 109–137 (1983).

[50] P. J. Bickel and A. Chen, A nonparametric view of network models and Newman–Girvan and other modularities. *Proc. Natl. Acad. Sci. USA* **106**, 21068–21073 (2009).

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

[52] A. P. Dempster, N. M. Laird, and D. B. Rubin, Maximum likelihood from incomplete data via the EM algorithm. *J. R. Statist. Soc. B* **39**, 185–197 (1977).

[53] G. J. McLachlan and T. Krishnan, *The EM Algorithm and Extensions*. Wiley-Interscience, New York, 2nd edition (2008).

[54] S. Wasserman and K. Faust, *Social Network Analysis*. Cambridge University Press, Cambridge (1994).

[55] C. Borgs, J. T. Chayes, L. Lovász, V. T. Sós, and K. Vesztergombi, Convergent sequences of dense graphs I: Subgraph frequencies, metric properties and testing. *Adv. Math.* **219**, 1801–1851 (2008).

[56] L. Lovász, *Large Networks and Graph Limits*, volume 60 of *American Mathematical Society Colloquium Publications*. American Mathematical Society, Providence, RI (2012).

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

[58] A.-L. Barabási and R. Albert, Emergence of scaling in random networks. *Science* **286**, 509–512 (1999).

[59] L. A. N. Amaral, A. Scala, M. Barthélemy, and H. E. Stanley, Classes of small-world networks. *Proc. Natl. Acad. Sci. USA* **97**, 11149–11152 (2000).

[60] M. Molloy and B. Reed, A critical point for random graphs with a given degree sequence. *Random Structures and Algorithms* **6**, 161–179 (1995).

[61] M. E. J. Newman, S. H. Strogatz, and D. J. Watts, Random graphs with arbitrary degree distributions and their applications. *Phys. Rev. E* **64**, 026118 (2001).

[62] F. Chung and L. Lu, The average distances in random graphs with given expected degrees. *Proc. Natl. Acad. Sci. USA* **99**, 15879–15882 (2002).

[63] N. Aharony, W. Pan, C. Ip, I. Khayal, and A. Pentland, Social fMRI: Investigating and shaping social mechanisms in the real world. *Pervasive and Mobile Computing* **7**, 643–659 (2011).

[64] J. Karikoski and M. Nelimarkka, Measuring social relations with multiple datasets. *Int. J. Social Computing and Cyber-Physical Systems* **1**, 98–113 (2011).

[65] A. Stopczynski, V. Sekara, P. Sapiezynski, A. Cuttone, M. M. Madsen, J. E. Larsen, and S. Lehmann, Measuring large-scale social networks with high resolution. *PLOS One* **9**, e95978 (2014).

[66] C. von Mering, L. J. Jensen, B. Snel, S. D. Hooper, M. Krupp, M. Foglierini, N. Jouffre, M. A. Huynen, and P. Bork, STRING: Known and predicted protein-protein associations, integrated and transferred across organisms. *Nucleic Acids Research* **33**, D433–D437 (2005).

[67] S. Boccaletti, G. Bianconi, R. Criado, C. I. del Genio, J. Gomez-Gardenes, M. Romance, I. Sendina-Nadal, Z. Wang, and M. Zanin, The structure and dynamics of multilayer networks. *Physics Reports* **544**, 1–122 (2014).

[68] M. De Domenico, C. Granell, M. A. Porter, and A. Arenas, The physics of multilayer networks. *Nature Physics* **12**, 901–906 (2016).

[69] Y. J. Wang and G. Y. Wong, Stochastic blockmodels for directed graphs. *Journal of the American Statistical Association* **82**, 8–19 (1987).

[70] E. Vaquera and G. Kao, Do you like me as much as I like you? Friendship reciprocity and its effects on school outcomes among adolescents. *Soc. Sci. Res.* **37**, 55–72 (2008).

[71] B. Ball and M. E. J. Newman, Friendship networks and social status. *Network Science* **1**, 16–30 (2013).

[72] H. Ebel, L.-I. Mielsch, and S. Bornholdt, Scale-free topology of e-mail networks. *Phys. Rev. E* **66**, 035103 (2002).

[73] J.-P. Onnela, J. Saramäki, J. Hyvönen, G. Szabó, D. Lazer, K. Kaski, J. Kertész, and A.-L. Barabási, Structure and tie strengths in mobile communication networks. *Proc. Natl. Acad. Sci. USA* **104**, 7332–7336 (2007).

[74] J. Leskovec and E. Horvitz, Planetary-scale views on a large instant-messaging network. In *Proceedings of the 17th International Conference on the World Wide Web*, pp. 915–924, Association of Computing Machinery, New York (2008).

[75] N. Eagle and A. Pentland, Reality mining: Sensing complex social systems. *Journal of Personal and Ubiquitous*

Computing **10**, 255–268 (2006).

[76] N. Eagle, A. Pentland, and D. Lazer, Inferring friendship network structure by using mobile phone data. *Proc. Natl. Acad. Sci. USA* **106**, 15274–15278 (2009).

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

[78] M. D. Resnick, P. S. Bearman, R. W. Blum, K. E. Bauman, K. M. Harris, J. Jones, J. Tabor, T. Beuhring, R. E. Sieving, M. Shew, M. Ireland, L. H. Bearinger, and J. R. Udry, Protecting adolescents from harm: Findings from the National Longitudinal Study on Adolescent Health. *Journal of the American Medical Association* **278**, 823–832 (1997).

[79] J. R. Udry, P. S. Bearman, and K. M. Harris, National Longitudinal Study of Adolescent Health (1997).

This research uses data from Add Health, a program project directed by Kathleen Mullan Harris and designed by J. Richard Udry, Peter S. Bearman, and Kathleen Mullan Harris at the University of North Carolina at Chapel Hill, and funded by grant P01–HD31921 from the Eunice Kennedy Shriver National Institute of Child Health and Human Development, with cooperative funding from 23 other federal agencies and foundations. Special acknowledgment is due Ronald R. Rindfuss and Barbara Entwisle for assistance in the original design. Information on how to obtain the Add Health data files is available on the Add Health website (http://www.cpc.unc.edu/addhealth). No direct support was received from grant P01–HD31921 for this analysis.

[80] T. P. Peixoto, Reconstructing networks with unknown and heterogeneous errors. *Phys. Rev. X* **8**, 041011 (2018).