Network structure from rich but noisy data

M. E. J. Newman*

Department of Physics and Center for the Study of Complex Systems,

University of Michigan, Ann Arbor, MI 48109, USA

Driven by growing interest across the sciences, a large number of empirical studies have been conducted in recent years of the structure of networks ranging from the internet and the world wide web to biological networks and social networks. The data produced by these experiments are often rich and multimodal, yet at the same time they may contain substantial measurement error [1–7]. Accurate analysis and understanding of networked systems requires a way of estimating the true structure of networks from such rich but noisy data [8–15]. Here we describe a technique that allows us to make optimal estimates of network structure from complex data in arbitrary formats, including cases where there may be measurements of many different types, repeated observations, contradictory observations, annotations or metadata, or missing data. We give example applications to two different social networks, one derived from face-to-face interactions and one from self-reported friendships.

*Email: mejn@umich.edu

Most empirical studies of networks take a "naive" view of structural data, meaning that one assumes that the data *are* the network. For instance, in a study of a protein–protein interaction network [16–18] one might compile a list of known protein interactions and represent them as a network of protein nodes joined by interaction edges. But this network represents the pattern of *measured* interactions not the pattern of actual interactions. The two could, and probably do, differ substantially, because of both error in the measurements and missing data [5, 19]. As another example, in studies of friendship networks [20,21] one commonly assembles a network simply by asking people who their friends are. The resulting network thus represents who people *say* they are friends with, not who they are actually friends with. The two can differ if, for instance, participants and experimenters apply different standards for what constitutes a friendship, or if participants fail to report some friendships at all [1, 2, 8, 22].

At the same time, many studies return data much richer than just a simple measurement of connections. Protein–protein interaction networks, for example, are commonly assembled from the results of many complementary experiments involving a variety of techniques, further enriched by knowledge of protein function, genetics, or other features. Friendship networks can likewise be probed in different ways, using surveys, online data, observations of face-to-face interactions, and others, possibly enhanced with metadata on participant location, occupation, age, and many other characteristics. Taken together these many types of data may be able to give a more accurate and nuanced picture of network structure than any single one can alone.

The problem of determining network structure from experimental data, which often goes under the heading of *network reconstruction*, has been studied particularly in the biological sciences, for instance in the context of gene regulatory networks, metabolic networks, and protein networks [5, 12, 23, 24]. A range of methods have been developed for use with data from high-throughput laboratory techniques such as microarrays, RNA sequencing, and tandem affinity purification [19, 25–29]. The

issue of errors and unreliability in network data has also been recognized in the social sciences, where there has been extensive discussion of sources of error in social surveys, its effects on measurements, and ways of estimating and minimizing it [1,2,6–8]. There is also domain-specific literature on problems such as predicting missing nodes or edges in networks [9, 10, 30–32] and name disambiguation in bibliometrics [33–36], typically making use of assumptions about correlations in network structure. Combinations of these methods can be used to create hybrid algorithms for resampling and Monte Carlo estimation of network structure [9–11, 13, 15]. There is also a significant volume of work on the related problem of estimating network structure from non-network data—see Brugere *et al.* [37] for a review.

Here we present a general formalism for the optimal inference of network structure from rich but noisy data, and show how it can be applied to a range of data types. Generically, the question we want to answer is this: given the results of a set of measurements performed on a system of interest, what is our best estimate of the structure of the underlying network? The data could take many forms. They could be rich, hierarchical, multilevel, and multimodal, but they may also be unreliable and error prone. Some of the data may have no bearing at all on the network structure. Others maybe related only obliquely to it. And we may not know in advance which data are relevant and which are not, or how accurate any of the measurements are. Remarkably, under these seemingly daunting circumstances we can nonetheless make progress.

Suppose that we are interested in the structure of a certain n-node network and for the moment let us concentrate on the commonest case of an unweighted undirected network. (We describe some generalizations to weighted and directed data below and in the Supplementary Materials.) Let us denote the true structure of the network—which we do not know—by an $n \times n$ symmetric adjacency matrix $\mathbf{A}$, having elements $A_{ij} = 1$ if nodes i and j are connected by an edge and 0 otherwise. This structure, commonly called the *ground truth*, is the thing we are trying to estimate.

We now make a set of measurements of the system, measurements that can take many forms as discussed above, perhaps including direct measurements of network structure but also potentially including indirect measurements, metadata, or "red herrings" that have nothing to do with the network at all. The network structure and the data are related to one another by a *data model*, expressed in the form of a probability function $P(\text{data}|\mathbf{A},\theta)$ that specifies the probability of making the particular set of measurements we did, given the ground-truth network $\mathbf{A}$ plus, optionally, some additional model parameters, which we collectively denote by θ. In general, we do not know the form of this probability distribution—in most cases it will be a complicated function—but the option to include parameters θ allows us to specify a family of functions that encompass a broad spectrum of possibilities. Our goal will be, given such a family, first to determine the values of the parameters, which effectively chooses a particular member of the family and thereby fixes the relationship between the network structure and the data, and then, given those values, to estimate the network structure itself.

We write

$$P(\mathbf{A},\theta|\text{data}) = \frac{P(\text{data}|\mathbf{A},\theta)P(\mathbf{A})P(\theta)}{P(\text{data})}, \tag{1}$$

then, summing over all possible network structures $\mathbf{A}$, we get $P(\theta|\text{data}) = \sum_{\mathbf{A}} P(\mathbf{A},\theta|\text{data})$, which we maximize to find the most probable value of the parameters θ given the observed data, the so-called *maximum a posteriori* (or MAP) estimate. In fact, for convenience, we maximize not $P(\theta|\text{data})$ but its logarithm, whose maximum falls in the same place. Employing the well-known Jensen inequality $\log \sum_i x_i \geq \sum_i q_i \log(x_i/q_i)$, we can write

$$\log P(\theta|\text{data}) = \log \sum_{\mathbf{A}} P(\mathbf{A},\theta|\text{data}) \geq \sum_{\mathbf{A}} q(\mathbf{A}) \log \frac{P(\mathbf{A},\theta|\text{data})}{q(\mathbf{A})}, \tag{2}$$

where $q(\mathbf{A})$ is any probability distribution over networks $\mathbf{A}$ satisfying $\sum_{\mathbf{A}} q(\mathbf{A}) = 1$. It is trivially the case that exact equality between left- and right-hand sides of Eq. (2)

is achieved when

$$q(\mathbf{A}) = \frac{P(\mathbf{A}, \theta|\text{data})}{\sum_{\mathbf{A}} P(\mathbf{A}, \theta|\text{data})}, \tag{3}$$

and hence this choice maximizes the right-hand side with respect to q. A further maximization with respect to θ will then give us the optimal parameter values we seek. To put that another way, a double maximization of the right-hand side of (2) with respect to both q and θ will give us our answer for θ. This can be easily carried out by maximizing first with respect to $q(\mathbf{A})$ using Eq. (3) and then with respect to θ, repeating until the result converges. Differentiating (2) while holding $q(\mathbf{A})$ constant, we find the maximum with respect to θ to be the solution of

$$\sum_{\mathbf{A}} q(\mathbf{A}) \nabla_\theta \log P(\mathbf{A}, \theta|\text{data}) = 0. \tag{4}$$

Our calculation consists of iterating Eqs. (3) and (4) from random initial values to convergence. The final result is a value for the parameters θ, which we can then use to estimate the ground-truth network. In fact, however, it turns out that this last step is unnecessary: the calculations we have already performed give us the ground-truth network structure as a by-product, indeed they give us the entire posterior probability distribution over structures, since from Eq. (3) the quantity $q(\mathbf{A})$ is none other than $q(\mathbf{A}) = P(\mathbf{A}, \theta|\text{data})/P(\theta|\text{data}) = P(\mathbf{A}|\text{data}, \theta)$. In other words it is precisely the probability of the network having true structure $\mathbf{A}$ given the observed data and the parameters θ.

The method derived here is an example of an expectation–maximization or EM algorithm [38]. As described the method is a general one that can be used with many different networks and data models. Let us see how it is applied in practice.

Our first example application is to a social network of US university students. The data come from the "reality mining" study of Eagle and Pentland [39], which aimed to establish the real-world social network of a set of individuals by measuring their physical proximity over time. The 96 students participating in the study were given

mobile phones that used special software to record when they were in proximity with one another. The resulting record of pairwise proximity measurements is both richer and poorer than a direct network measurement, in exactly the manner considered in this paper. It is richer in the sense that interactions between individuals may be measured repeatedly and not just once, but poorer in the sense that proximity is an error-prone indicator of actual interaction—two individuals may find themselves coincidentally in proximity, as they pass on the street say, without being acquainted or having any social interaction.

We take as our data set the measurements made during the reality mining study for eight consecutive Wednesdays in March and April of 2005. (We choose weekly observations to remove weekly periodic effects, and March and April because they fall during the university term.) This gives us eight sets of observations, one for each day, in which an observed edge means that two individuals were in physical proximity at some time during that day.

The data model we adopt for these data is a particularly simple one, in which the edge measurements—the observations of proximity—are assumed to be independent identically distributed random variables, conditioned on the ground truth A_{ij}. That is, the probability of observing an edge between nodes i and j depends only on the matrix element A_{ij} and in the same way for all i, j. This dependence can be parametrized by two quantities: the *true-positive rate* α, which is the probability of observing an edge where one truly exists, and the *false-positive rate* β, the probability of observing an edge where none exists. (Note that these are the empirical true- and false-positive rates—the frequency with which the measurements agree or disagree with the ground truth—rather than the true- and false-positive rates for our final inferred networks, which we cannot normally calculate.) In addition, we will assume a uniform prior probability ρ of the existence of an edge in any position, so that our model is parametrized by three parameters α, β, and ρ.

If for each node pair i, j we make N measurements and observe an edge to be

present in E_{ij} of them then, as shown in the Methods, our EM equations give the following estimates for the three parameters:

$$\hat{\alpha} = \frac{\sum_{i<j} E_{ij} Q_{ij}}{N \sum_{i<j} Q_{ij}}, \qquad \hat{\beta} = \frac{\sum_{i<j} E_{ij}(1 - Q_{ij})}{N \sum_{i<j} (1 - Q_{ij})}, \qquad \hat{\rho} = \frac{1}{\binom{n}{2}} \sum_{i<j} Q_{ij}. \tag{5}$$

(We use hatted symbols to denote estimated values of variables.) The quantity Q_{ij} appearing here is the posterior probability that there is an edge between nodes i and j for these parameter values, which is given by

$$Q_{ij} = \frac{\hat{\rho}\hat{\alpha}^{E_{ij}}(1-\hat{\alpha})^{N-E_{ij}}}{\hat{\rho}\hat{\alpha}^{E_{ij}}(1-\hat{\alpha})^{N-E_{ij}} + (1-\hat{\rho})\hat{\beta}^{E_{ij}}(1-\hat{\beta})^{N-E_{ij}}}. \tag{6}$$

The full calculation involves iterating Eqs. (5) and (6) until convergence is reached, and the results tell us the estimates of the three parameters α, β, and ρ, as well as the entire posterior probability distribution over possible ground-truth networks, which is given by $P(\mathbf{A}|\text{data}, \theta) = \prod_{i>j} Q_{ij}^{A_{ij}}(1 - Q_{ij})^{1-A_{ij}}$. The posterior distribution allows us to compute estimates of any other network quantities we might be interested in, such as degrees, correlations, or clustering coefficients (see Section S.5 in the Supplementary Materials) and can also be used as an input to further calculations, for instance of community structure [14].

Applying Eqs. (5) and (6) to the reality mining data, the algorithm converges rapidly and reliably to parameter estimates $\hat{\alpha} = 0.4242$, $\hat{\beta} = 0.0043$, and $\hat{\rho} = 0.0335$. The small value of β tells us that there are very few false positives: an edge is observed where none exists less than 1% of the time. On the other hand, even if the false-positive rate is low, the probability of being wrong when one does observe an edge can still be high. This probability, called the *false discovery rate*, is given by $(1 - \rho)\beta/[\rho\alpha + (1 - \rho)\beta]$, which has estimated value of 0.2270 in the present case, meaning that more than one in every five observed edges is in error. Moreover, the relatively small value of α implies that there are also a large number of false negatives: around 58% of pairs of individuals who are in fact connected in the underlying network

are not observed in proximity on any one day. This is understandable. Most people do not see all of their acquaintances every day.

Figure 1a shows the inferred ground-truth network, with edge thicknesses varying to indicate the probability Q_{ij} of individual edges. In Fig. 1b we show the relationship between the number of observations E_{ij} of a particular edge and the posterior probability Q_{ij}. As the figure shows, an edge observed only zero or one times implies a low Q_{ij} (less than 0.1), so a single observation is probably a false alarm. But two or more observations of the same edge result in a much larger Q_{ij} (greater than 0.9), indicating a strong inference that the edge exists in the ground truth. The sharp transition between low and high values of Q_{ij} means that it is possible to infer the presence or absence of edges with good reliability despite the high error rate in the data.

For our second example we study a more traditional friendship network, taken from the National Longitudinal Study of Adolescent Health (the "AddHealth" study) [21]. This study compiled networks of friendships between students at a number of US high schools by asking participants to name their friends. Again the data are both richer and poorer than a simple network measurement. They are richer in the sense that we have two measurements of each friendship, from the point of view of each of the two participants, but poorer in the sense that those measurements can (and often do) disagree, indicating that respondents are not reliable in the reports they give or that they are employing different standards for what constitutes a friendship. Following [8] we represent this situation by giving each participant i their own individual true- and false-positive rates α_i and β_i. Once again one can derive closed-form expressions for these parameters and for the posterior probabilities Q_{ij} of edges in the ground-truth network—see the Methods. The analysis can be applied to any of the schools in the AddHealth study; we use one of the smaller ones as our example, solely because it allows us to make a clear picture of the resulting network.

Again the EM algorithm converges quickly and reliably, giving a network-average

estimated true-positive rate $\langle\hat{\alpha}\rangle = 0.6083$, false-positive rate $\langle\hat{\beta}\rangle = 0.0096$, and prior edge probability $\langle\hat{\rho}\rangle = 0.0235$. These values indicate that non-existent friendships are rarely falsely reported as existing (low average β_i), although, once again, arguably the more interesting quantity is the false discovery rate, the probability of a friendship that *is* reported being false. This probability, which is equal to $(1-\rho)\beta_i/[\rho\alpha_i + (1-\rho)\beta_i]$, is significantly larger, having a network-average estimated value of 0.3309. In other words, about one in three reported friendships doesn't really exist. There is also a relatively high rate of failure to report friendships that do exist (many of the α_i are significantly less than 1). The latter is perhaps less surprising given the design of the study: students were limited to naming at most ten friends, so those with more than ten would be obliged to omit some.

Figure 1c shows the inferred network of friendships, with edge widths again indicating the probability Q_{ij} that an edge exists, and node sizes now varying to indicate how reliable the nodes are, in terms of the fraction of reported friendships that actually exist (which is equal to one minus the false discovery rate, also called the precision). Reports made by nodes depicted with large diameter are reliable, those made by smaller nodes are not. Armed with these results, one can now calculate a multitude of further quantities, including any function of network structure.

These are just two examples of possible applications. The particular data models applied here are quite flexible and could be applied to other networks, but there are also many other models one could use. Note for instance that the two models above both make the assumption that edges are conditionally independent. This works well for these particular examples but is by no means a requirement. The methods described can be applied to models with dependent edges too, which might be appropriate for instance for data sets derived from longitudinal (time-dependent) network studies. See the Supplementary Materials for further discussion and a number of additional examples of possible models.

Methods

In the reality mining example, edge observations are assumed to be independent (Bernoulli) random variables, conditioned on the ground truth A_{ij} for the appropriate node pair i, j, with true-positive rate α and false-positive rate β. Suppose that for each node pair i, j we make N_{ij} measurements and observe an edge to be present in E_{ij} of those measurements. Then, under this independent edge model,

$$P(\text{data}|\mathbf{A}, \theta) = \prod_{i<j} \left[\alpha^{E_{ij}}(1-\alpha)^{N_{ij}-E_{ij}}\right]^{A_{ij}} \left[\beta^{E_{ij}}(1-\beta)^{N_{ij}-E_{ij}}\right]^{1-A_{ij}}. \tag{7}$$

If the prior probability of an edge in any position is ρ then the prior probability of the entire network is $P(\mathbf{A}|\rho) = \prod_{i<j} \rho^{A_{ij}}(1-\rho)^{1-A_{ij}}$. We also assume that the prior probability distributions on α, β, and ρ themselves are all uniform in the interval $[0, 1]$. Combining Eqs. (1) and (7) we then have

$$P(\mathbf{A}, \theta|\text{data}) = \frac{1}{P(\text{data})} \prod_{i<j} \left[\rho\alpha^{E_{ij}}(1-\alpha)^{N_{ij}-E_{ij}}\right]^{A_{ij}} \left[(1-\rho)\beta^{E_{ij}}(1-\beta)^{N_{ij}-E_{ij}}\right]^{1-A_{ij}}. \tag{8}$$

Taking the log, substituting into Eq. (4), and differentiating with respect to α, we find that the maximum a posteriori estimate $\hat{\alpha}$ of the true-positive rate satisfies

$$\sum_{\mathbf{A}} q(\mathbf{A}) \sum_{i<j} A_{ij} \left(\frac{E_{ij}}{\hat{\alpha}} - \frac{N_{ij} - E_{ij}}{1-\hat{\alpha}} \right) = 0. \tag{9}$$

Defining the posterior probability of an edge between i and j by $Q_{ij} = P(A_{ij} = 1|\text{data}, \theta) = \sum_{\mathbf{A}} q(\mathbf{A}) A_{ij}$ and rearranging Eq. (9), we then get

$$\hat{\alpha} = \frac{\sum_{i<j} E_{ij} Q_{ij}}{\sum_{i<j} N_{ij} Q_{ij}}. \tag{10}$$

Similarly, differentiating with respect to β and ρ we arrive at

$$\hat{\beta} = \frac{\sum_{i<j} E_{ij}(1-Q_{ij})}{\sum_{i<j} N_{ij}(1-Q_{ij})}, \qquad \hat{\rho} = \frac{1}{\binom{n}{2}} \sum_{i<j} Q_{ij}. \tag{11}$$

For the data set considered here the N_{ij} all take the same value N, in which case Eqs. (10) and (11) reduce to Eq. (5).

To calculate $q(\mathbf{A})$ we evaluate (8) at the estimated parameter values and substitute the result into Eq. (3) to get

$$\begin{aligned} q(\mathbf{A}) &= \frac{\prod_{i<j} \bigl[\hat{\rho}\hat{\alpha}^{E_{ij}}(1-\hat{\alpha})^{N_{ij}-E_{ij}}\bigr]^{A_{ij}} \bigl[(1-\hat{\rho})\hat{\beta}^{E_{ij}}(1-\hat{\beta})^{N_{ij}-E_{ij}}\bigr]^{1-A_{ij}}}{\sum_{\mathbf{A}} \prod_{i<j} \bigl[\hat{\rho}\hat{\alpha}^{E_{ij}}(1-\hat{\alpha})^{N_{ij}-E_{ij}}\bigr]^{A_{ij}} \bigl[(1-\hat{\rho})\hat{\beta}^{E_{ij}}(1-\hat{\beta})^{N_{ij}-E_{ij}}\bigr]^{1-A_{ij}}} \\ &= \prod_{i<j} \frac{\bigl[\hat{\rho}\hat{\alpha}^{E_{ij}}(1-\hat{\alpha})^{N_{ij}-E_{ij}}\bigr]^{A_{ij}} \bigl[(1-\hat{\rho})\hat{\beta}^{E_{ij}}(1-\hat{\beta})^{N_{ij}-E_{ij}}\bigr]^{1-A_{ij}}}{\sum_{A_{ij}=0,1} \bigl[\hat{\rho}\hat{\alpha}^{E_{ij}}(1-\hat{\alpha})^{N_{ij}-E_{ij}}\bigr]^{A_{ij}} \bigl[(1-\hat{\rho})\hat{\beta}^{E_{ij}}(1-\hat{\beta})^{N_{ij}-E_{ij}}\bigr]^{1-A_{ij}}} \\ &= \prod_{i<j} Q_{ij}^{A_{ij}}(1-Q_{ij})^{1-A_{ij}}, \end{aligned} \tag{12}$$

where

$$Q_{ij} = \frac{\hat{\rho}\hat{\alpha}^{E_{ij}}(1-\hat{\alpha})^{N_{ij}-E_{ij}}}{\hat{\rho}\hat{\alpha}^{E_{ij}}(1-\hat{\alpha})^{N_{ij}-E_{ij}} + (1-\hat{\rho})\hat{\beta}^{E_{ij}}(1-\hat{\beta})^{N_{ij}-E_{ij}}}. \tag{13}$$

Note that if we make no measurements for a pair of nodes i, j, so that $N_{ij} = E_{ij} = 0$ (the case of "missing data"), this expression correctly gives Q_{ij} equal to the estimated prior edge probability $\hat{\rho}$.

Turning to the AddHealth friendship network example, measurements of edges in this data set come from unilateral statements made by participants. Let E_{ij} in this case represent the number of times node i identifies node j as a friend. (Normally this number will be zero or one, but we allow arbitrary values for the sake of generality.) In effect, E_{ij} constitutes a directed network, and self-reported friendship networks are sometimes depicted as being directed. We consider the underlying ground-truth network, however, to be undirected. Only our observations of it are directed.

Study participants may vary in the reliability with which they identify their

friends. A participant whose identifications agree, by and large, with those of their friends, is probably a reliable observer; one whose identifications disagree is probably not. We do not have to impose these assumptions on our calculation, however. They will be automatically reflected in the solution found by the EM algorithm.

In our calculations we employ a data model in which each node i has its own true-positive rate α_i and false-positive rate β_i. Then the likelihood of a set of observations given a ground-truth network $\mathbf{A}$ is

$$\begin{aligned} P(\text{data}|\mathbf{A},\theta) = \prod_{i<j} &\bigl[\alpha_i^{E_{ij}}(1-\alpha_i)^{N_{ij}-E_{ij}}\bigr]^{A_{ij}} \bigl[\alpha_j^{E_{ji}}(1-\alpha_j)^{N_{ji}-E_{ji}}\bigr]^{A_{ji}} \\ &\times \bigl[\beta_i^{E_{ij}}(1-\beta_i)^{N_{ij}-E_{ij}}\bigr]^{1-A_{ij}} \bigl[\beta_j^{E_{ji}}(1-\beta_j)^{N_{ji}-E_{ji}}\bigr]^{1-A_{ji}}, \end{aligned} \tag{14}$$

where N_{ij} is the total number of observations of node j made by node i. Note that we explicitly include terms in E_{ij} and E_{ji} separately, since these numbers are distinct. (On the other hand, $A_{ij} = A_{ji}$ since the ground-truth network is assumed undirected. We write A_{ij} and A_{ji} separately in the above expression purely to preserve symmetry.)

Again assuming a prior probability of ρ on each ground-truth edge and uniform priors on the parameters, applying Eq. (1), and taking logs, we arrive at the log-likelihood:

$$\begin{aligned} \log P(\mathbf{A},\theta|\text{data}) = \sum_{i<j} \bigl[&A_{ij}E_{ij}\log\alpha_i + A_{ij}(N_{ij}-E_{ij})\log(1-\alpha_i) \\ &+ A_{ji}E_{ji}\log\alpha_j + A_{ji}(N_{ji}-E_{ji})\log(1-\alpha_j) \\ &+ (1-A_{ij})E_{ij}\log\beta_i + (1-A_{ij})(N_{ij}-E_{ij})\log(1-\beta_i) \\ &+ (1-A_{ji})E_{ji}\log\beta_j + (1-A_{ji})(N_{ji}-E_{ji})\log(1-\beta_j) \\ &+ A_{ij}\log\rho + (1-A_{ij})\log(1-\rho)\bigr] - \log P(\text{data}). \end{aligned} \tag{15}$$

Applying Eq. (4), performing the derivatives, and rearranging, we then find the fol-

lowing estimates for the parameters:

$$\hat{\alpha}_i = \frac{\sum_j E_{ij} Q_{ij}}{\sum_j N_{ij} Q_{ij}}, \qquad \hat{\beta}_i = \frac{\sum_j E_{ij}(1-Q_{ij})}{\sum_j N_{ij}(1-Q_{ij})}, \qquad \hat{\rho} = \frac{1}{\binom{n}{2}} \sum_{i<j} Q_{ij}. \tag{16}$$

As before, Q_{ij} is the posterior probability of an edge between i and j, which can be calculated by a method analogous to the one we used for our first model above. Combining Eqs. (1) and (14) and using $A_{ij} = A_{ji}$, we write

$$P(\mathbf{A}, \theta|\text{data}) = \frac{1}{P(\text{data})} \prod_{i<j} \bigl[\rho \alpha_i^{E_{ij}} (1-\alpha_i)^{N_{ij}-E_{ij}} \alpha_j^{E_{ji}} (1-\alpha_j)^{N_{ji}-E_{ji}}\bigr]^{A_{ij}} \times \bigl[(1-\rho) \beta_i^{E_{ij}} (1-\beta_i)^{N_{ij}-E_{ij}} \beta_j^{E_{ji}} (1-\beta_j)^{N_{ji}-E_{ji}}\bigr]^{1-A_{ij}}. \tag{17}$$

We evaluate this probability at the estimated values of the parameters and the complete posterior distribution over ground-truth networks $\mathbf{A}$ is then given by

$$q(\mathbf{A}) = P(\mathbf{A}|\text{data}, \theta) = \frac{P(\mathbf{A}, \theta|\text{data})}{\sum_{\mathbf{A}} P(\mathbf{A}, \theta|\text{data})} = \prod_{i<j} Q_{ij}^{A_{ij}} (1-Q_{ij})^{1-A_{ij}}, \tag{18}$$

where

$$Q_{ij} = \frac{\hat{\rho} \hat{\alpha}_i^{E_{ij}} (1-\hat{\alpha}_i)^{N_{ij}-E_{ij}} \hat{\alpha}_j^{E_{ji}} (1-\hat{\alpha}_j)^{N_{ji}-E_{ji}}}{\hat{\rho} \hat{\alpha}_i^{E_{ij}} (1-\hat{\alpha}_i)^{N_{ij}-E_{ij}} \hat{\alpha}_j^{E_{ji}} (1-\hat{\alpha}_j)^{N_{ji}-E_{ji}} + (1-\hat{\rho}) \hat{\beta}_i^{E_{ij}} (1-\hat{\beta}_i)^{N_{ij}-E_{ij}} \hat{\beta}_j^{E_{ji}} (1-\hat{\beta}_j)^{N_{ji}-E_{ji}}}. \tag{19}$$

Note that this expression is explicitly symmetric with respect to the indices i and j, as it should be, since $Q_{ij} = Q_{ji}$ by definition.

This calculation returns not only an estimate of the ground-truth network but also an estimate of the reliability of each of the nodes, parametrized by their true-positive and false-positive rates, which tell us both how often a node truthfully reports an edge that does exist and how often it falsely reports an edge that does not. Note that even in the (common) case where each edge is observed at most once, so that E_{ij} can take only the values zero and one, the parameter estimates $\hat{\alpha}_i$ and $\hat{\beta}_i$ and the

posterior probabilities Q_{ij} can take a wide range of values, by contrast with the case of the reality mining network, where there are only as many possible values of Q_{ij} as there are values of E_{ij} (see Fig. 1b). For instance, even if both of nodes i and j report the existence of an edge between them ($E_{ij} = E_{ji} = 1$), if neither node is considered reliable then the algorithm may say that the probability Q_{ij} of the edge actually existing is low. If either of them is considered reliable, on the other hand, then Q_{ij} will be larger. And if one is unreliable and claims an edge, while the other is reliable but does not, then Q_{ij} will be particularly small.

Data availability statement: The two data sets used in this paper are freely available on the web from the original authors of the respective studies. The "reality mining" data were collected by Eagle and Pentland [39] and can be downloaded from `http://realitycommons.media.mit.edu/realitymining.html`. The high-school friendship data are from the National Longitudinal Study of Adolescent Health [21] (the "AddHealth" study) and links to the data files can be found at `http://www.cpc.unc.edu/projects/addhealth/documentation/publicdata`.

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

This work uses data from Add Health, a program project designed by J. Richard Udry, Peter S. Bearman, and Kathleen Mullan Harris, and funded by a 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. Persons interested in obtaining data files from Add Health should contact Add Health, Carolina Population Center, 123 W. Franklin Street, Chapel Hill, NC 27516–2524 (addhealthunc.edu). No direct support

was received from grant P01–HD31921 for this analysis.

Author contributions: MEJN designed and conducted the research and wrote the paper.

References

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

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

[3] Lakhina, A., Byers, J., Crovella, M. & Xie, P. 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).

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

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

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

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

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

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

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

[11] Namata, G. M., Kok, S. & Getoor, L. 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).

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

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

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

[15] Casiraghi, G., Nanumyan, V., Scholtes, I. & Schweitzer, F. From relational data to graphs: Inferring significant links using generalized hypergeometric ensembles. In Ciampaglia, G., Mashhadi, A. & Yasseri, T. (eds.) *Proceedings of the International Conference on Social Informatics (SocInfo 2017)*, no. 10540 in Lecture Notes in Computer Science, 111–120 (Springer, Berlin, 2017).

[16] Uetz, P. *et al.* A comprehensive analysis of protein–protein interactions in saccharomyces cerevisiae. *Nature* **403**, 623–627 (2000).

[17] Ito, T. *et al.* A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc. Natl. Acad. Sci. USA* **98**, 4569–4574 (2001).

[18] Giot, L., Bader, J. S., Brouwer, C. *et al.* A protein interaction map of Drosophila melanogaster. *Science* **302**, 1727–1736 (2003).

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

[20] Rapoport, A. & Horvath, W. J. A study of a large sociogram. *Behavioral Science* **6**, 279–291 (1961).

[21] Resnick, M. D. *et al.* Protecting adolescents from harm: Findings from the National Longitudinal Study on Adolescent Health. *Journal of the American Medical Association* **278**, 823–832 (1997).

[22] Bernard, H. R. & Killworth, P. D. Informant accuracy in social network data II. *Human Communications Research* **4**, 3–18 (1977).

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

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

[25] Overbeek, R. *et al.* Wit: Integrated system for high-throughput genome sequence analysis and metabolic reconstruction. *Nucleic Acids Res.* **28**, 123–125 (2000).

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

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

[28] Margolin, A. A. *et al.* ARACNE: An algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinformatics* **7**, S7 (2006).

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

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

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

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

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

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

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

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

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

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

[39] Eagle, N. & Pentland, A. Reality mining: Sensing complex social systems. *Journal of Personal and Ubiquitous Computing* **10**, 255–268 (2006).

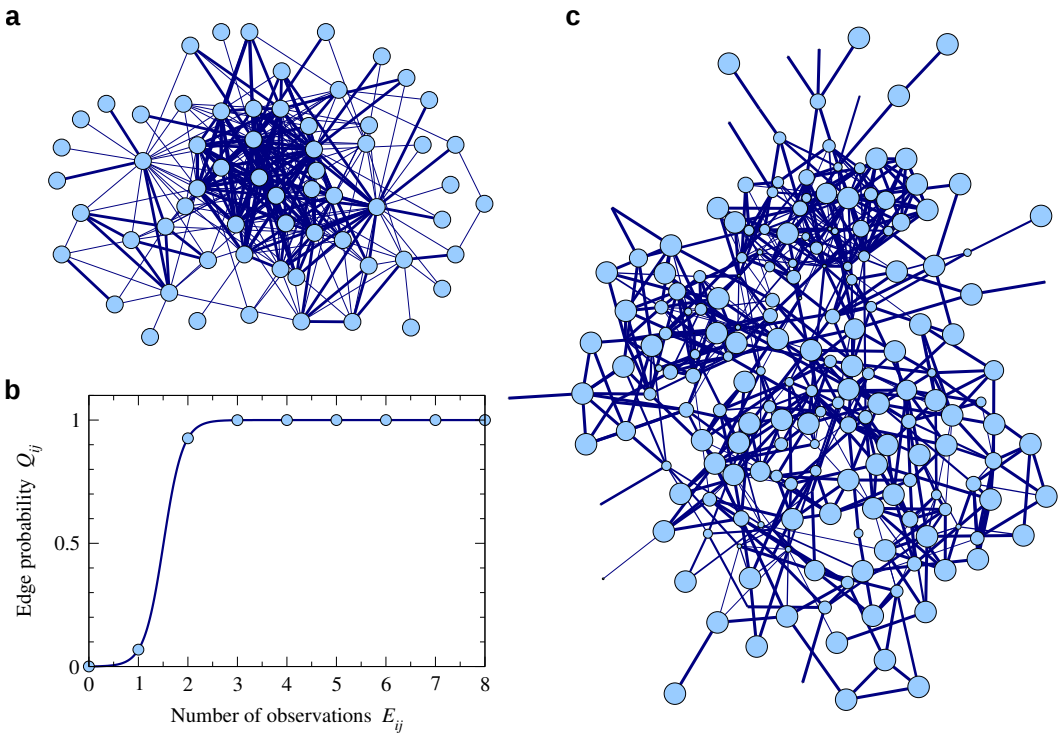


Figure 1: **Application of the methods described here to two example networks.** The EM algorithm derived in this paper was applied to a data set of proximity measurements between a group of US university students (the "reality mining" study [39]) and to a friendship network derived from a survey of students in a US high school (the "AddHealth" study [21]). (a) Inferred ground-truth network for the reality mining data set. Edge widths indicate the inferred probabilities Q_{ij}. Edges that are never observed are omitted, as are singleton nodes with no observed edges. The figure reveals a dense core of about twenty nodes that are with high probability connected to one another and a sparser periphery of nodes for whom the surety of connection is much lower. The thickest edges shown have $Q_{ij} > 0.999$, while the thinnest have $Q_{ij} < 0.1$. (b) Inferred edge probability as a function of the number of observations E_{ij} for the reality mining data set, showing a relatively sharp transition between $E_{ij} = 1$ and $E_{ij} = 2$. (c) Inferred network for the AddHealth friendship data. Edge widths again indicate inferred probabilities, while node diameters are proportional to the so-called precision $\rho\alpha_i/[\rho\alpha_i + (1-\rho)\beta_i]$, which is the estimated fraction of reported friendships that actually exist. Some nodes are invisible because they are unreliable—their precision is very small—though these nodes may nonetheless have edges if another (reliable) node reports a connection. Unobserved edges and singleton nodes are again omitted.

Network structure from rich but noisy data: Supplementary materials

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

S.1 Additional results for the reality mining network

Our EM algorithm works by finding the values of the model parameters that give the best fit of the data model to the observed data. The method does not, however, guarantee that we will get a *good* fit. Even the best fit may still be a bad one if the model itself is not capable of capturing the form of the data. As an analogy, imagine a set of data points on a graph that follow an intrinsically curved path across the page. We can fit a straight line through such points, but even the best fit will not be a good one. There simply is no good fit of a straight line to curved data.

For the case of the reality mining data set of mobile phone proximities, a "good fit" to the data means one that captures accurately the numbers of proximity observations for pairs of individuals in the network. Since the observations are assumed independent, only their number matters and not other features such as the specific days on which proximity is observed. Figure S1 shows a histogram of pairs of individuals in the network as a function of the number of days on which they are observed in proximity. Because the network is sparse and a large majority of pairs never meet, most of the weight of the histogram is in the "zero observations" bin, although significant numbers of pairs fall in the other bins as well. The circles show the values of the same quantities for the best-fit model—the one given by the parameter values in the paper. As the figure shows, the fit is a reasonably good one, although there is some deviation between data and fit if one looks closely.

Another way to assess the quality of the results is to rerun the algorithm with an independent data set from the same source to see if we get a similar outcome. A nice feature of the reality mining study is that we have exactly such an independent data set available. Recall that the results given in the paper are based on observations made on eight consecutive Wednesdays. It is straightforward to perform the same analysis using data from a different day of the week. Figure S2 shows the network structure inferred from the Wednesday data (the same structure depicted in Fig. 1a in the main paper) alongside the equivalent structure inferred from data for eight Thursdays over the same time period. As the figure shows, the two networks are qualitatively similar, with a dense core and sparse periphery. Some notable features, such as the tightly-connected satellite group of nodes to the left of center in the figure, are common to both networks. But some individual details also differ from one network to the other—edges present in one are absent in the other and so forth. This is natural: the whole point about error-prone data is that if we measure the same thing twice we do not expect to get exactly the same result. Some

*Email: mejn@umich.edu

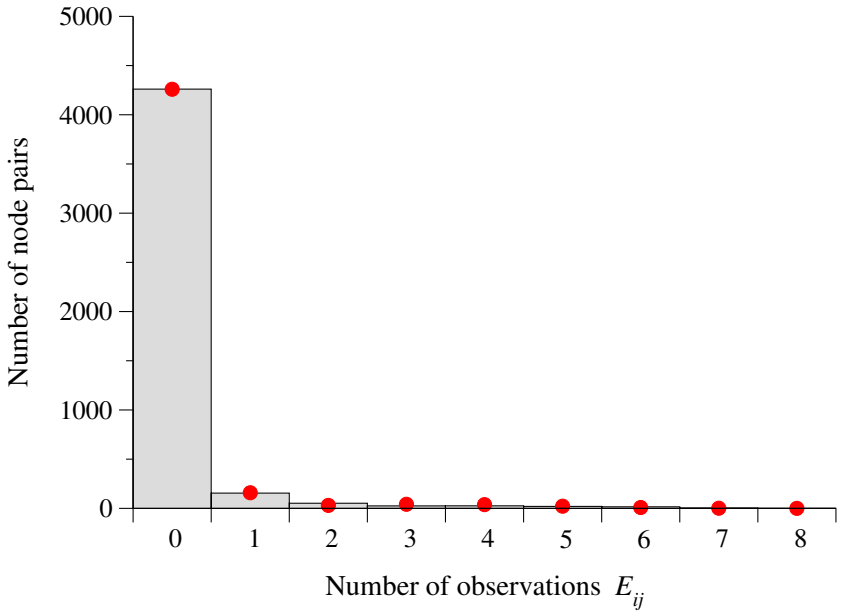


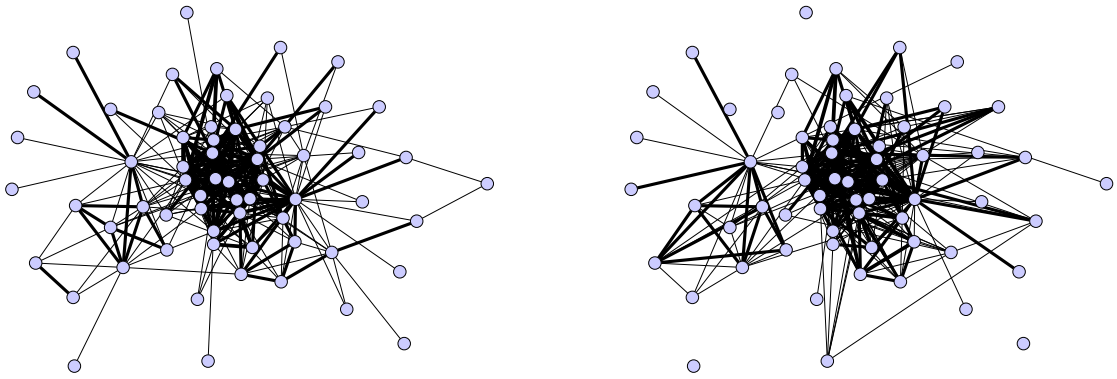
Figure S1: **Comparison of observations and fitted model for the reality mining data.** The histogram shows the number of node pairs i, j with each possible value of E_{ij}, the number of days on which the corresponding pair of individuals were observed in proximity. The circles represent the predictions made by the model for the parameter values that give the best fit to the data.

variation between different measurements is expected, and indeed the extent of this variation could in principle be used to estimate the size of the experimental error. In this case, however, we don't necessarily need to do this, since the model parameters—the true- and false-positive rates—already give us an estimate of the size of our errors.

S.2 Tests against synthetic data

How do we know if our method gives good results? It gives us a best estimate (in the maximum-likelihood sense) of a network and its parameters, but is that estimate actually good? Under normal circumstances we cannot compare our estimated network to the true structure to evaluate performance because, by definition, we do not know the true structure. A common alternative therefore is to evaluate performance against computer-generated or "synthetic" data. One can create a network with known structure (or use one that already exists) and artificially introduce errors into some fraction of its edges, then see whether our algorithms are able to accurately recover the true structure of the network, errors notwithstanding.

Figure S3 shows results from a set of such tests using the independent edge data model of Eqs. (5) and (6) in the main paper. In these tests we generated random networks of 200 nodes with average node degree 10 and then randomly introduced both false positive and false negative errors. For simplicity we fixed the true-positive rate and precision for the introduced errors to take the same values. For each network we generated either $N = 4$, 8, or 16 sets of measurements then fed the resulting data into our EM algorithm. The plot shows the performance of the algorithm in terms of the *recall*, i.e., the fraction of edges that were correctly recovered by the algorithm as a function of the rate of introduced errors.



(a) Wednesdays (b) Thursdays

Figure S2: **Networks inferred from observations made on different days of the week.** The "reality mining" network was reconstructed in two different ways, using proximity observations made on Wednesdays and on Thursdays. (a) The network reconstructed from observations made on Wednesdays, as described in the main text. (This is the same network as in Fig. 1a, but redrawn for the purposes of this figure.) (b) The equivalent network reconstructed from observations made on Thursdays over the same eight-week period in March and April 2005. We limit ourselves to the same set of nodes as in panel (a), laid out in the same positions, to allow easy comparison between networks.

As the figure shows, the success of the algorithm depends, as one would expect, on both the number N of measurements available to it and the quality of the data. As the error rate goes to 1 the algorithm fails to recover the network at all, but it has significant success for lower error rates. When both the true-positive rate and the precision of the measurements are 0.5, for instance, the algorithm succeeds with about 70% recall even with only 4 measurements of each network to work with, and with 16 measurements it has better than 98% recall.

S.3 Other data models

We have given two examples of possible data models. There are however many others that could be used within the inference framework described, depending on the specific data available and the questions one wants to answer.

Edge strengths or weights: A simple variation on the model we used for the reality mining data set is one in which the underlying network can have edges with different strengths. Many social network studies only consider pairs of individuals to be "acquainted" or "not acquainted." But a more nuanced representation might divide them into "not acquainted," "casual acquaintances," or "well acquainted," and the frequency with which people meet might well differ between these classes: casual acquaintances might be more likely to meet than people who don't know each other at all, but less likely than people who are close friends.

Such a situation could be represented using a weighted adjacency matrix $\mathbf{A}$ in which each element now has three possible values 0, 1, and 2, with corresponding prior probabilities ρ_0, ρ_1, and ρ_2 such that $\rho_0 + \rho_1 + \rho_2 = 1$. At the same time the two parameters α and β that we used in the previous model would now become three—say α_0, α_1, and α_2—representing the probability of observing an edge in each of the three states. With all other variables defined as before, the

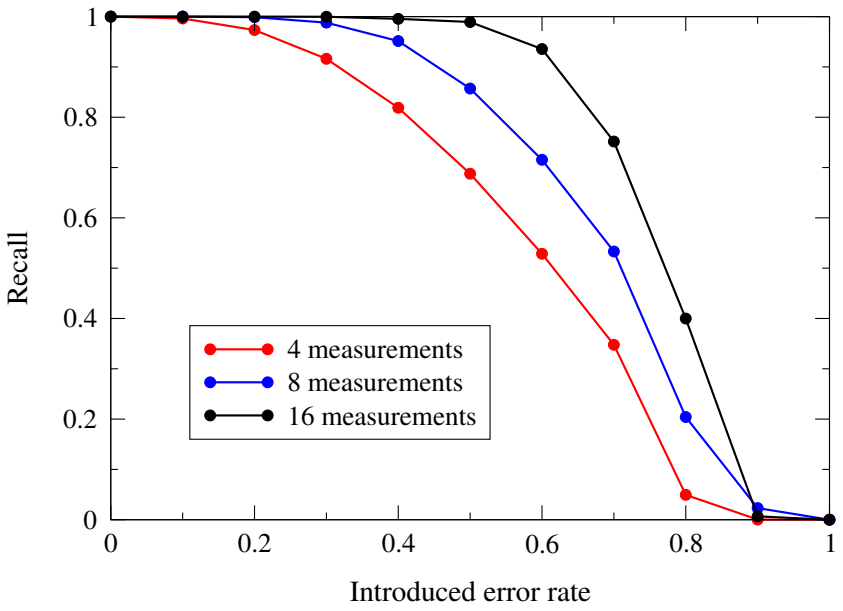


Figure S3: **Tests of the method on synthetically generated networks.** This plot shows the recall rate of our EM algorithm when used to reconstruct artificially generated networks with errors introduced into them at a controlled rate. For each data point we generated 100 networks of 200 nodes each using the standard (Bernoulli) random graph model and from each network we generated 4, 8, or 16 independent measurements of the network structure with random errors introduced with varying frequency. Both false positives and false negatives were introduced: the true-positive rate and precision were set to the same value, shown on the horizontal axis. The vertical axis measures the algorithm's ability to successfully recreate the original network from the error-prone data in terms of the recall, which is the fraction of edges correctly identified by the algorithm. An edge is considered to be identified if its existence is inferred to be more likely than not, i.e., if $Q_{ij} > \frac{1}{2}$. Error bars are comparable in size with the data points and are omitted.

log-likelihood would then take the form

$$\begin{aligned}\log P(\mathbf{A},\theta|\text{data}) = \sum_{i<j}\big\{ & \mathbb{1}_{A_{ij}=0}\big[E_{ij}\log\alpha_0 + (N_{ij}-E_{ij})\log(1-\alpha_0)\big] \\ & + \mathbb{1}_{A_{ij}=1}\big[E_{ij}\log\alpha_1 + (N_{ij}-E_{ij})\log(1-\alpha_1)\big] \\ & + \mathbb{1}_{A_{ij}=2}\big[E_{ij}\log\alpha_2 + (N_{ij}-E_{ij})\log(1-\alpha_2)\big] \\ & + \mathbb{1}_{A_{ij}=0}\log\rho_0 + \mathbb{1}_{A_{ij}=1}\log\rho_1 + \mathbb{1}_{A_{ij}=2}\log\rho_2\big\} \\ & - \log P(\text{data}),\end{aligned} \tag{S1}$$

where $\mathbb{1}$ is the indicator function. Then, applying Eq. (4) in the paper, we derive the estimates

$$\hat{\alpha}_w = \frac{\sum_{i<j} E_{ij}Q_{ij}^{(w)}}{\sum_{i<j} N_{ij}Q_{ij}^{(w)}}, \qquad \hat{\rho}_w = \frac{1}{\binom{n}{2}}\sum_{i<j} Q_{ij}^{(w)}, \tag{S2}$$

for $w = 0, 1, 2$, where $Q_{ij}^{(w)} = \sum_{\mathbf{A}} \mathbb{1}_{A_{ij}=w}\, q(\mathbf{A})$ is the posterior probability that $A_{ij} = w$, which is

given by

$$Q_{ij}^{(w)} = \frac{\hat{\rho}_w \hat{\alpha}_w^{E_{ij}} (1-\hat{\alpha}_w)^{N_{ij}-E_{ij}}}{\sum_{w'} \hat{\rho}_{w'} \hat{\alpha}_{w'}^{E_{ij}} (1-\hat{\alpha}_{w'})^{N_{ij}-E_{ij}}}. \tag{S3}$$

This approach can be extended to any number of levels or strengths of connection between node pairs—Eqs. (S2) and (S3) carry over unchanged. Interesting questions arise about how we decide the ideal number of levels to include in the calculation (if we don't know *a priori*), which can be addressed using generalizations of standard model selection methods. For instance, one could perform a χ^2 test on the distribution of values of E_{ij}, choosing the minimum number of levels for which the model is not rejected by the test to some predetermined degree of significance.

Note that within this framework the levels of the edges are not ordered: there is nothing in the mathematical formulation that stipulates that level 2 is "stronger" than level 1. In practice this means that the parameter values returned by the EM algorithm may be permuted from the canonical order—all permutations give equally good fits to the data. If we want the higher levels to correspond to stronger edges in the sense of greater values of α_w, then we may need to manually permute the levels after the algorithm completes its work.

Multimodal data: Another possibility is that of "multimodal" network data, by which we mean data that quantify the structure of a network in several different ways, such as a social network probed using traditional interviews or questionnaires, and then probed again using data from an online social networking site. Such data are sometimes referred to as "multilayer" networks [1, 2]. An example is the Copenhagen Networks Study [3], in which the interactions of a thousand individuals in Copenhagen were cataloged using measurements of face-to-face meetings, electronic communications, and online social networks.

Suppose that we have data that measure, directly or indirectly, a specific network $\mathbf{A}$ in several different ways or modes, which we label by integers $m = 1, 2, 3\ldots$ There is only one type of edge in the network itself—the matrix elements A_{ij} take the values 0 and 1 only—but they can be measured in multiple ways. The existence, or non-existence, of an edge between node pair i, j is measured $N_{ij}^{(m)}$ times in mode m. (The most likely values are $N_{ij}^{(m)} = 1$—the pair was observed once, the usual situation in most network studies—or $N_{ij}^{(m)} = 0$—the case of "missing data," where we have no information about a particular pair. For the sake of generality, however, we allow the possibility of higher values.) Generalizing our earlier models, we also define $E_{ij}^{(m)}$ to be the number of times an edge is actually observed between nodes i, j in mode m, and we assume the measurements to be independent, both for different modes and for different nodes, conditioned on the underlying ground truth A_{ij}. But we allow for the (likely) situation in which measurements in different modes have different levels of accuracy, meaning that there are different true- and false-positive rates for each mode m, which we denote α_m and β_m. In a social network, for instance, we might find that exchange of electronic communications such as emails or phone calls is a more reliable indicator of acquaintance than proximity measurements.

The log-likelihood for this model is given by

$$\begin{aligned}\log P(\mathbf{A}, \theta|\text{data}) = \sum_{i<j} \Big\{ \sum_m \big[& A_{ij} E_{ij}^{(m)} \log \alpha_m + A_{ij}(N_{ij}^{(m)} - E_{ij}^{(m)}) \log(1-\alpha_m) \\ & + (1-A_{ij}) E_{ij}^{(m)} \log \beta_m + (1-A_{ij})(N_{ij}^{(m)} - E_{ij}^{(m)}) \log(1-\beta_m) \big] \\ & + A_{ij} \log \rho + (1-A_{ij}) \log(1-\rho) \Big\} - \log P(\text{data}),\end{aligned} \tag{S4}$$

where ρ is once again the prior probability of an edge. Substituting this form into Eq. (4) of the main paper and performing the derivatives, we get

$$\hat{\alpha}_m = \frac{\sum_{i<j} E_{ij}^{(m)} Q_{ij}}{\sum_{i<j} N_{ij}^{(m)} Q_{ij}}, \qquad \hat{\beta}_m = \frac{\sum_{i<j} E_{ij}^{(m)} (1-Q_{ij})}{\sum_{i<j} N_{ij}^{(m)} (1-Q_{ij})}, \qquad \hat{\rho} = \frac{1}{\binom{n}{2}} \sum_{i<j} Q_{ij}, \tag{S5}$$

where $Q_{ij} = \sum_{\mathbf{A}} q(\mathbf{A}) A_{ij}$ is once again the posterior probability of an edge between nodes i and j. Following the same line of argument as in the Methods, we find that

$$Q_{ij} = \frac{\hat{\rho} \prod_m \hat{\alpha}_m^{E_{ij}^{(m)}} (1-\hat{\alpha}_m)^{N_{ij}^{(m)} - E_{ij}^{(m)}}}{\hat{\rho} \prod_m \hat{\alpha}_m^{E_{ij}^{(m)}} (1-\hat{\alpha}_m)^{N_{ij}^{(m)} - E_{ij}^{(m)}} + (1-\hat{\rho}) \prod_m \hat{\beta}_m^{E_{ij}^{(m)}} (1-\hat{\beta}_m)^{N_{ij}^{(m)} - E_{ij}^{(m)}}}. \tag{S6}$$

To understand how the different modes are weighted by the algorithm, it is helpful to consider the odds ratio for an edge between nodes i and j:

$$\frac{Q_{ij}}{1-Q_{ij}} = \frac{\hat{\rho}}{1-\hat{\rho}} \prod_m \left(\frac{\hat{\alpha}_m}{\hat{\beta}_m} \right)^{E_{ij}^{(m)}} \left(\frac{1-\hat{\alpha}_m}{1-\hat{\beta}_m} \right)^{N_{ij}^{(m)} - E_{ij}^{(m)}}. \tag{S7}$$

Note how, in modes m for which $\hat{\alpha}_m$ is large and $\hat{\beta}_m$ is small, the $E_{ij}^{(m)}$ observed edges contribute a large increase to the odds ratio (first term in parentheses) and the $N_{ij}^{(m)} - E_{ij}^{(m)}$ non-edges contribute a large decrease (second term). These modes are precisely the reliable ones—those with high true-positive rates and low false-positive rates—and hence it is appropriate that they contribute strongly to our inference of the network structure.

S.4 Computation of network properties

The primary output of our EM algorithms is the posterior probability distribution $q(\mathbf{A}) = P(\mathbf{A}|\text{data}, \theta)$ over possible ground-truth networks. Given this distribution, one can in principle calculate the expected value or distribution of any other quantity that depends on network structure, such as degree distributions, clustering coefficients, correlation measures, spectral properties, and so forth. If we have some quantity $X(\mathbf{A})$ whose value is a function of the network structure $\mathbf{A}$, then its expected value, given the observed data, is

$$\mu_X = \sum_{\mathbf{A}} q(\mathbf{A}) X(\mathbf{A}), \tag{S8}$$

and the variance about that expectation is

$$\sigma_X^2 = \sum_{\mathbf{A}} q(\mathbf{A}) [X(\mathbf{A}) - \mu]^2. \tag{S9}$$

These expressions are primarily of use for quantities whose distribution is approximately normal. In other cases one can compute the complete probability distribution of X thus:

$$P(X = x|\text{data}, \theta) = \sum_{\mathbf{A}} q(\mathbf{A}) \, \mathbb{1}_{X(\mathbf{A})=x}, \tag{S10}$$

where $\mathbb{1}$ is the indicator function again.

These expressions can be used in place of more traditional "naive" estimates of network properties, i.e., estimates made directly from the observed network data. They allow us to give best estimates of network quantities of interest as well as estimates of the uncertainty on those quantities. This approach puts network measurements on a similar standing to measurements in other fields of science, where providing error estimates on observed quantities is standard practice.

In some cases it is possible to employ the expressions above directly. Take for example the calculation of the degree of a node i. For any of the data models described in this paper, or any other model for which one can derive an explicit expression for the marginal probability Q_{ij} of an edge between two nodes, we can write the expected degree of node i as

$$d_i = \sum_{\mathbf{A}} q(\mathbf{A}) \sum_j A_{ij} = \sum_j \sum_{\mathbf{A}} q(\mathbf{A}) A_{ij} = \sum_j Q_{ij}. \tag{S11}$$

In other cases, particularly those in which the quantity of interest is a nonlocal function of network structure, such as a correlation function or an eigenvalue, it may not be possible to perform the sum over networks $\mathbf{A}$ in closed form, in which case one can estimate expectations, variances, or complete distributions using Monte Carlo sampling, whereby one draws a number of networks from the posterior distribution $q(\mathbf{A})$, computes the quantity of interest on each of them, and then calculates the desired statistics.

In the particular case in which the posterior distribution factors into independent probabilities on each edge—as in all of the models considered in this paper—Monte Carlo sampling of networks is trivial. One simply generates each edge independently with the appropriate probability Q_{ij}, and there exist straightforward algorithms for doing this efficiently [4]. In cases where the edges are not independent, one can generate networks using Markov chain importance sampling, in which one repeatedly makes small changes $\mathbf{A} \to \mathbf{A}'$ to the network, such as the addition or removal of a single edge, then accepts those changes with the standard Metropolis–Hastings acceptance probability

$$P_a = \begin{cases} q(\mathbf{A}')/q(\mathbf{A}) & \text{if } q(\mathbf{A}') < q(\mathbf{A}), \\ 1 & \text{otherwise.} \end{cases} \tag{S12}$$

As an example, consider the calculation of the clustering coefficient C, which is a measure of the density of triangles or "closed triads" of edges in a network. In social network terms, the clustering coefficient measures the average probability that two of your friends will also be friends with each other. It is defined by

$$C = \frac{3 \times (\text{number of triangles in network})}{(\text{number of connected triples})}, \tag{S13}$$

where a connected triple means an unordered pair of nodes that are both neighbors of the same third node [5].

Consider the example of the network of high school friendships in Fig. 1c in the main paper. Taking the output of our EM algorithm, we generate 1000 networks with edges sampled randomly with the probabilities Q_{ij} given by the algorithm, then calculate the clustering coefficient for each network and compute the mean and standard deviation of the 1000 resulting values. This gives us a best estimate of $C = 0.185 \pm 0.009$ for the clustering coefficient. For comparison, the "naive" value of the clustering coefficient for the same network, calculated directly from the raw friendship data,

is 0.269, in significant disagreement with the best estimate. If we were to calculate the clustering coefficient from the raw data in this case we would be introducing a substantial error.

References

[1] Boccaletti, S. *et al.* The structure and dynamics of multilayer networks. *Physics Reports* **544**, 1–122 (2014).

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

[3] Stopczynski, A. *et al.* Measuring large-scale social networks with high resolution. *PLOS One* **9**, e95978 (2014).

[4] Ramani, A. S., Eikmeier, N. & Gleich, D. F. Coin-flipping, ball-dropping, and grass-hopping for generating random graphs from matrices of edge probabilities. Preprint arxiv:1709.03438 (2017).

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