LETTERS
https://doi.org/10.1038/s41567-018-0076-1
nature physics

Network structure from rich but noisy data

M. E. J. Newman

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.

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 ref. [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 may be related only obliquely to it. Furthermore, 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 Information.) 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

Department of Physics and Center for the Study of Complex Systems, University of Michigan, Ann Arbor, MI, USA. e-mail: mejn@umich.edu

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

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

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 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

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

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 equation (2) is achieved when

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

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 equation (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 equation (3) and then with respect to θ, repeating until the result converges. Differentiating equation (2) while holding $q(\mathbf{A})$ constant, we find the maximum with respect to θ to be the solution of

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

Our calculation consists of iterating equations (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 equation (3) the quantity $q(\mathbf{A}) = P(\mathbf{A},\theta|\text{data})/P(\theta|\text{data}) = P(\mathbf{A}|\text{data},\theta)$. 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 a 'reality mining' study[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 expectation-maximization equations give the following estimates for the three parameters:

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

(We use symbols with hats 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

$$Q_{ij} = \frac{\widehat{\rho}\,\widehat{\alpha}^{E_{ij}}(1-\widehat{\alpha})^{N-E_{ij}}}{\widehat{\rho}\,\widehat{\alpha}^{E_{ij}}(1-\widehat{\alpha})^{N-E_{ij}} + (1-\widehat{\rho})\widehat{\beta}^{E_{ij}}(1-\widehat{\beta})^{N-E_{ij}}} \quad (6)$$

The full calculation involves iterating equations (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 Supplementary Section 5) and can also be used as an input to further calculations (for instance, of community structure[14]).

Applying equations (5) and (6) to the reality mining data, the algorithm converges rapidly and reliably to parameter estimates $\widehat{\alpha} = 0.4242$, $\widehat{\beta} = 0.0043$ and $\widehat{\rho} = 0.0335$. 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 an 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 times or one time implies a low Q_{ij} (less than 0.1), so a single observation is probably a false alarm. However, 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 `Add Health' 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 ref. [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 Add Health 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 expectation-maximization algorithm converges quickly and reliably, giving a network-average estimated true-positive rate $\langle\widehat{\alpha}\rangle = 0.6083$, false-positive rate $\langle\widehat{\beta}\rangle = 0.0096$ and prior edge probability $\langle\widehat{\rho}\rangle = 0.0235$. 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 does not 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 it is not 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 Information for further discussion and a number of additional examples of possible models.

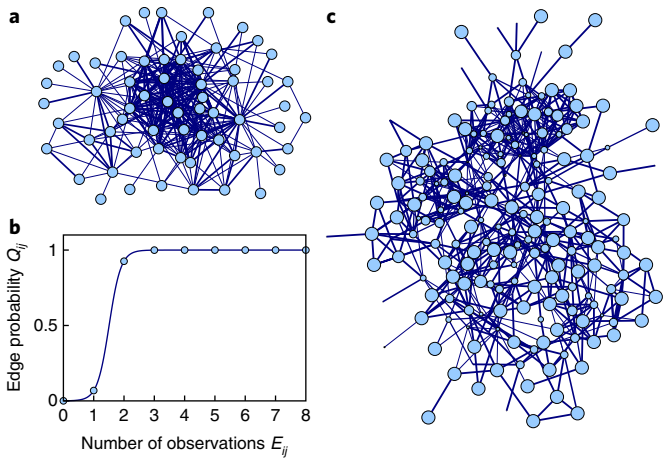

Fig. 1 | Application of the methods described here to two example networks. The expectation-maximization algorithm derived in this paper was applied to a data set of proximity measurements between a group of US university students (the `reality mining' study[39]) and to a friendship network derived from a survey of students in a US high school (the `Add Health' 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 20 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 Add Health 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—although these nodes may nonetheless have edges if another (reliable) node reports a connection. Unobserved edges and singleton nodes are again omitted.

Methods

Methods, including statements of data availability and any associated accession codes and references, are available at https://doi.org/10.1038/s41567-018-0076-1.

Received: 12 September 2017; Accepted: 2 February 2018;
Published online: 12 March 2018

References

1. Killworth, P. D. & Bernard, H. R. Informant accuracy in social network data. *Hum. Organ.* **35**, 269–286 (1976).
2. Marsden, P. V. Network data and measurement. *Annu. Rev. Sociol.* **16**, 435–463 (1990).
3. Lakhina, A., Byers, J., Crovella, M. & Xie, P. Sampling biases in IP topology measurements. In *Proc. 22nd Annual Joint Conf. of the IEEE Computer and Communications Societies* (Institute of Electrical and Electronics Engineers, New York, NY, 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. *Mol. Cell. Proteom.* **8**, 3–18 (2009).

6. Handcock, M. S. & Gile, K. J. Modeling social networks from sampled data. *Ann. Appl. Stat.* **4**, 5–25 (2010).
7. Lusher, D., Koskinen, J. & Robins, G. *Exponential Random Graph Models for Social Networks: Theory, Methods, and Applications* (Cambridge Univ. Press, Cambridge, 2012).
8. Butts, C. T. Network inference, error, and informant (in)accuracy: A Bayesian approach. *Soc. Netw.* **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 *Proc. 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 *Proc. International Conf. on Social Informatics (SocInfo 2017), no. 10540 in Lecture Notes in Computer Science* (eds Ciampaglia, G. et al.) 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. *Behav. Sci.* **6**, 279–291 (1961).
21. Resnick, M. D. et al. Protecting adolescents from harm: Findings from the National Longitudinal Study on Adolescent Health. *J. Am. Med. Assoc.* **278**, 823–832 (1997).
22. Bernard, H. R. & Killworth, P. D. Informant accuracy in social network data II. *Human. Commun. Res.* **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. Royal 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 Res.* **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. *J. Social Struct.* **10**, 1–29 (2009).
32. Kim, M. & Leskovec, J. The network completion problem: Inferring missing nodes and edges in networks. In *Proc. 2011 SIAM International Conf. on Data Mining* (eds Liu, B. et al.) 47–58 (Society for Industrial and Applied Mathematics: Philadelphia, PA, 2011).
33. Smalheiser, N. R. & Torvik, V. I. Author name disambiguation. *Annu. Rev. Inf. Sci. Technol.* **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 Rec.* **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 Trans. Knowl. Data Eng.* **24**, 975–987 (2012).
37. Brugere, I., Gallagher, B. & Berger-Wolf, T. Y. Network structure inference, a survey: Motivations, methods, and applications. *ACM Comput. Surv.* **1**, 1 (2016).
38. Dempster, A. P., Laird, N. M. & Rubin, D. B. Maximum likelihood from incomplete data via the EM algorithm. *J. Royal Stat. Soc. B* **39**, 185–197 (1977).
39. Eagle, N. & Pentland, A. Reality mining: Sensing complex social systems. *J. Personal Ubiquitous Comput.* **10**, 255–268 (2006).

Acknowledgements

The author thanks E. Bruch, G. Cantwell, T. Martin, G. Reinert and M. Riolofor 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 programme project designed by J. R. Udry, P. S. Bearman and K. 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. A special acknowledgment is due to R. R. Rindfuss and B. Entwisle for assistance in the original design. Anyone interested in obtaining data files from Add Health should contact Add Health, Carolina Population Center, 123 W. Franklin Street, Chapel Hill, NC 27516-2524 (addhealth@unc.edu). No direct support was received from grant P01-HD31921 for this analysis.

Author contributions

M.E.J.N. designed and conducted the research and wrote the paper.

Competing interests

The author declares no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41567-018-0076-1.

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence and requests for materials should be addressed to M.E.J.N.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

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,

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

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 equations (1) and (7), we then have

$$\begin{aligned} P(\mathbf{A},\theta|\text{data}) &= \frac{1}{P(\text{data})} \prod_{i<j} [\rho\alpha^{E_{ij}}(1-\alpha)^{N_{ij}-E_{ij}}]^{A_{ij}} \\ &\quad \times [(1-\rho)\beta^{E_{ij}}(1-\beta)^{N_{ij}-E_{ij}}]^{1-A_{ij}} \end{aligned} \tag{8}$$

Taking the log, substituting into equation (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 equation (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

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

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

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

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

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 Add Health 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. However, we consider the underlying ground-truth network 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, generally, 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 expectation-maximization 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

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

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 equation (1), and taking logs, we arrive at the log-likelihood:

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

Applying equation (4), performing the derivatives and rearranging, we then find the following estimates for the parameters:

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

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 equations (1) and (14) and using $A_{ij} = A_{ji}$, we write

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

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

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

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 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. Finally, 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. The reality mining data[39] are available at http://realitycommons.media.mit.edu/realitymining.html and the high-school friendship data[21] are available at http://www.cpc.unc.edu/projects/addhealth/documentation/publicdata.