Missing and spurious interactions and the reconstruction of complex networks

Roger Guimerà[a,1] and Marta Sales-Pardo[a,b,c]

[a]Department of Chemical and Biological Engineering and [b]Northwestern Institute on Complex Systems, Northwestern University, Evanston, IL 60208; and [c]Northwestern University Clinical and Translational Science Institute, Northwestern University, Chicago, IL 60611

Network analysis is currently used in a myriad of contexts, from identifying potential drug targets to predicting the spread of epidemics and designing vaccination strategies and from finding friends to uncovering criminal activity. Despite the promise of the network approach, the reliability of network data is a source of great concern in all fields where complex networks are studied. Here, we present a general mathematical and computational framework to deal with the problem of data reliability in complex networks. In particular, we are able to reliably identify both missing and spurious interactions in noisy network observations. Remarkably, our approach also enables us to obtain, from those noisy observations, network reconstructions that yield estimates of the true network properties that are more accurate than those provided by the observations themselves. Our approach has the potential to guide experiments, to better characterize network data sets, and to drive new discoveries.

data reliability | block model | modularity | node roles | Bayesian inference

The structure of the network of interactions between the units of a system affects the system's dynamics and conveys information about the functional needs of the system, its evolution, and the role of individual units. For these reasons, network analysis has become a cornerstone of fields as diverse as systems biology and sociology (1). Unfortunately, the reliability of network data is often a source of concern. In systems biology, high-throughput technologies hold the promise to uncover the intricate processes within the cell but are also reportedly inaccurate. Protein interaction data provide, arguably, the most blatant example of data inaccuracy: In 2002, a systematic comparison of several high-throughput methods to a reference high-quality data set showed that these methods have accuracies below 20% (2). Additionally, different methods result in networks that have different topological properties (3), and the coverage of real interactomes is very limited: 80% of the interactome of yeast (3) and 99.7% of the human interactome (4, 5) are still unknown.

In the social sciences, missing data due to individual non-response and dropout (6), informant inaccuracy (7), and sampling biases (8) are also pervasive. Simulation studies have established that these inaccuracies can lead to fundamentally wrong estimates of network properties and to misleading conclusions (8), which is particularly worrisome at a time when social network analysis is being used for finding new friends and partners, singling out key individuals in organizations, and identifying criminal activity.

Despite these concerns, the issue of network reliability has only been addressed in a field-by-field basis [for example, to deal with protein–protein interactions (9, 10) or to take into account informant inaccuracy in social networks (7)], and in studies that only address parts of the problem [for example, to detect missing interactions (11)]. Therefore, a general framework to deal with the problem of data reliability in complex networks is lacking. Here, we develop such a framework. Specifically, we show that within our framework we can reliably (*i*) identify false negatives (missing interactions) and false positives (spurious interactions) as well as (*ii*) generate, from a single observed network, a reconstructed network whose properties (clustering coefficient, modularity,

assortativity, epidemic spreading threshold, and synchronizability, among others) are closer to the "true" underlying network than those of the observed network itself. We show that our approach outperforms previous attempts to predict missing and spurious interactions, and illustrate the potential of our method by applying it to a protein interaction network of yeast (12). We end by discussing how our approach will help to guide experiments and new discoveries, and to better characterize important data sets.

General Reliability Formalism

Consider an observed network with adjacency matrix A^O; $A^O_{ij} = 1$ if nodes i and j are connected and 0 otherwise. We assume that this observed network is a realization of an underlying probabilistic model, either because the network itself is the result of a stochastic process, because the measurement has uncertainty, or both (7).* Let us call $\mathcal{M}$ the set of generative models that could conceivably give rise to the observed network, and $p(M|A^O)$ the probability that $M \in \mathcal{M}$ is the model that gave rise to the observation A^O. If we could get a new observation of the network, the outcome would in general be different from A^O; our best estimate for the probability $p(X = x)$ for an arbitrary network property X is

$$p(X = x|A^O) = \int_{\mathcal{M}} dM\, p(X = x|M)\, p(M|A^O), \qquad [1]$$

where $p(X = x|M)$ is the probability that $X = x$ in a network generated with model M. Using Bayes theorem, we can rewrite Eq. **1** as

$$p(X = x|A^O) = \frac{\int_{\mathcal{M}} dM\, p(X = x|M)\, p(A^O|M)\, p(M)}{\int_{\mathcal{M}} dM'\, p(A^O|M') p(M')}, \qquad [2]$$

where $p(A^O|M)$ is the probability that model M gives rise to A^O among all possible adjacency matrices, and $p(M)$ is the a priori probability that model M is the correct one. We call $p(X = x|A^O)$ the reliability of the $X = x$ measurement.

Stochastic Block Models

Given the generality of these arguments, the key to good estimates of reliability is to identify sets of models that are general, empirically grounded, and analytically or computationally tractable. Here, we focus on the family $\mathcal{M}_{BM}$ of stochastic block models (13, 14). In a stochastic block model, nodes are partitioned into
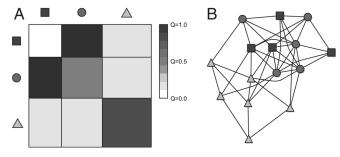
Fig. 1. Stochastic block models. A stochastic block model is fully specified by a partition of nodes into groups and a matrix Q in which each element $Q_{\alpha\beta}$ represents the probability that a node in group α connects to a node in group β. (*A*), A simple matrix of probabilities Q. Nodes are divided in three groups (which contain 4, 5, and 6 nodes, respectively) and are represented as squares, circles, and triangles depending on their group. The value of each element $Q_{\alpha\beta}$ is indicated by the shade of gray; for example, squares do not connect to other squares, and connect to triangles with small probability, but squares connect to circles with high probability. (*B*) A realization of the model in *A*. In this realization, the number of links between the square and the triangle group is $l_{\square\triangle} = 4$, whereas the maximum possible number of links between these groups is $r_{\square\triangle} = 24$.

groups and the probability that two nodes are connected depends only on the groups to which they belong (Fig. 1).

Stochastic block models are empirically grounded in that they capture two ubiquitous and fundamental properties of real complex networks. First, nodes in real networks are often organized into modules or communities (15, 16), which may overlap (17) or be hierarchically nested (11, 18, 19), so that connections are relatively more abundant within modules than between modules. In most real-world networks, this modularity is significantly larger than expected from chance (16, 20). Second, nodes in real networks fulfill distinct roles and connect to each other depending on these roles (13, 16). Role-to-role connections are not necessarily assortative (16, 21); that is, nodes with a certain role may or may not tend to connect with other nodes with the same role. In general, stochastic block models are particularly appropriate when nodes belong to groups and interact with each other depending on their group membership (regardless of whether interactions occur mostly within groups or between groups).

Stochastic block models are also appropriate in that they can capture other more general connectivity correlations in the network. For example, if people establish social connections with others according to age, then a block model that partitions individuals into age groups will capture some of the correlations in the network.

In general, complex networks result from a combination of mechanisms, including modularity, role structure, and maybe other factors. Although partitions into modules, roles, and age groups, for example, can be very different from each other, some block model in the $\mathcal{M}_{BM}$ family is likely to capture each of them separately; by sampling over all models $M \in \mathcal{M}_{BM}$ we capture a variety of correlations, ideally to the exact degree that they are relevant.

Additionally, stochastic block models are analytically tractable because in a stochastic block model the probability that nodes i and j are connected depends only on the groups to which they belong (14). Therefore, we can calculate the reliability of individual links and the reliability of entire networks.

Link Reliability: Missing and Spurious Interactions

The reliability of an individual link is $R_{ij}^L \equiv p_{BM}(A_{ij} = 1|A^O)$, that is, the probability that the link "truly" exists given our observation of the whole network (and our choice of the family of stochastic block models). Assuming no prior knowledge about the suitability of the models, we obtain (*Materials and Methods*)

$$R_{ij}^L = \frac{1}{Z} \sum_{P \in \mathcal{P}} \left(\frac{l_{\sigma_i\sigma_j}^O + 1}{r_{\sigma_i\sigma_j} + 2} \right) \exp[-\mathcal{H}(P)], \qquad [3]$$

where the sum is over partitions P in the space $\mathcal{P}$ of all possible partitions of the network into groups, σ_i is node i's group (in partition P), $l_{\alpha\beta}^O$ is the number of links in the observed network between groups α and β, and $r_{\alpha\beta}$ is the maximum possible number of links between α and β (Fig. 1). The function $\mathcal{H}(P)$ is a function of the partition

$$\mathcal{H}(P) = \sum_{\alpha \leq \beta} \left[\ln(r_{\alpha\beta} + 1) + \ln \binom{r_{\alpha\beta}}{l_{\alpha\beta}^O} \right], \qquad [4]$$

and $Z = \sum_{P \in \mathcal{P}} \exp[-\mathcal{H}(P)]$.

In practice, it is not possible to sum over all partitions even for small networks.[†] However, since Eq. 3 has the same mathematical form as an ensemble average in statistical mechanics (22), one can use the Metropolis algorithm to correctly sample relevant partitions (that is, partitions that significantly contribute to the sum) and obtain estimates for the link reliability (*Materials and Methods*).

We use the link reliability to identify missing and spurious interactions in network observations. We evaluate the performance of our approach using five high-quality networks: the social network of interactions between people in a karate club (23), the social network of frequent associations between 62 dolphins (24), the air transportation network of Eastern Europe (25), the neural network of the nematode *Caenorhabditis elegans* (26), and the metabolic network of *Escherichia coli* (27, 28). All of these networks have been manually curated and are widely used in the literature as model systems. Therefore, in what follows we assume that each of these networks is the "true" network A^T and is error-free. We then generate hypothetical observations A^O by adding or removing random connections from A^T, and evaluate the ability of our approach to recover the features of the true network.[‡]

To quantitatively study missing interactions, we generate observed networks A^O by removing random links from the true network A^T. We then estimate the link reliability R_{ij}^L for each of these false negatives ($A_{ij}^O = 0$ and $A_{ij}^T = 1$), as well as for the true negatives ($A_{ij}^O = 0$ and $A_{ij}^T = 0$). We measure the algorithm's ability to identify missing interactions by ranking the reliabilities (in decreasing order) and calculating the probability that a false negative has a higher ranking than a true negative (11). Similarly, we quantify the ability to identify spurious interactions by adding random links to the true network, obtaining and ranking the link reliabilities (again, in decreasing order), and calculating the probability that a false positive ($A_{ij}^O = 1$ and $A_{ij}^T = 0$) is ranked lower than a true positive ($A_{ij}^O = 1$ and $A_{ij}^T = 1$).

In Fig. 2, we compare our approach to the hierarchical random graph (HRG) approach of Clauset et al. (11) and to a local algorithm based on the number of common neighbors between each pair of nodes [*Materials and Methods*; see also Fig. S1 in supporting information (SI) *Appendix* for a comparison to other local algorithms] (11, 29). We find that, except for one network, our approach consistently outperforms all others at identifying both missing interactions and spurious interactions. Our approach is also the only one that performs consistently well for all networks (unlike local algorithms, which work well for some networks but very poorly for others) and for both missing and spurious interactions (unlike the HRG algorithm, which performs comparatively

---

[†]The number of distinct partitions of N elements into groups is $\sum_{k=1}^{N} \frac{1}{k!} \sum_{l=1}^{k} \binom{k}{l} (-1)^{k-l} l^N$, which grows faster than any finite power of N.

[‡]By adding and removing connections in this way, we are implicitly focusing on random errors; we discuss at the end how our approach can also deal with systematic (or, in general, correlated) errors.
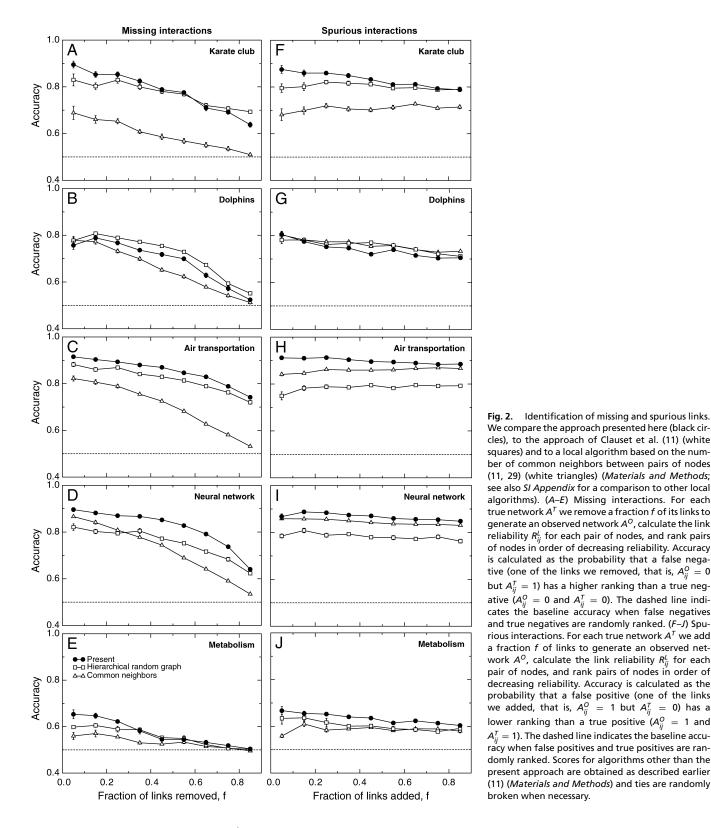
Fig. 2. Identification of missing and spurious links. We compare the approach presented here (black circles), to the approach of Clauset et al. (11) (white squares) and to a local algorithm based on the number of common neighbors between pairs of nodes (11, 29) (white triangles) (*Materials and Methods*; see also *SI Appendix* for a comparison to other local algorithms). (*A–E*) Missing interactions. For each true network A^T we remove a fraction f of its links to generate an observed network A^O, calculate the link reliability R_{ij}^L for each pair of nodes, and rank pairs of nodes in order of decreasing reliability. Accuracy is calculated as the probability that a false negative (one of the links we removed, that is, $A_{ij}^O = 0$ but $A_{ij}^T = 1$) has a higher ranking than a true negative ($A_{ij}^O = 0$ and $A_{ij}^T = 0$). The dashed line indicates the baseline accuracy when false negatives and true negatives are randomly ranked. (*F–J*) Spurious interactions. For each true network A^T we add a fraction f of links to generate an observed network A^O, calculate the link reliability R_{ij}^L for each pair of nodes, and rank pairs of nodes in order of decreasing reliability. Accuracy is calculated as the probability that a false positive (one of the links we added, that is, $A_{ij}^O = 1$ but $A_{ij}^T = 0$) has a lower ranking than a true positive ($A_{ij}^O = 1$ and $A_{ij}^T = 1$). The dashed line indicates the baseline accuracy when false positives and true positives are randomly ranked. Scores for algorithms other than the present approach are obtained as described earlier (11) (*Materials and Methods*) and ties are randomly broken when necessary.

APPLIED PHYSICAL SCIENCES

worse at detecting spurious interactions[§]). Our algorithm is also consistently the most accurate when applied to a number of model networks, including networks with hierarchically nested modules, networks with a strongly disassortative role structure, and non-modular scale-free networks (Fig. S2 in *SI Appendix*). We find that only when the network is strictly a hierarchical random graph, is the HRG approach slightly more accurate at predicting missing interactions than the BM approach (Sections 2 and 3 in *SI Appendix*). Remarkably, even for strict hierarchical random graphs, the BM approach is more accurate at identifying spurious interactions.

[§]A plausible explanation for this behavior is that, because in the HRG model most parameters are used to "fit" low-level features of the network (pairs of nodes, triplets of nodes, and so on), the HRG approach may overfit spurious links.

Network Reliability and Network Reconstruction

The success at detecting both missing and spurious interactions confirms that our approach is able to uncover the structural features of the true network A^T. The natural question is thus whether it is possible to "reconstruct" the observation A^O to gain greater insight into the global structure of A^T. This is difficult because, in general, adding a few candidate missing interactions and removing a few candidate spurious interactions does not give satisfactory network reconstructions (one of the main problems being that one does not know, a priori, how many missing and spurious interactions there are).

Therefore, the first step toward network reconstruction is to obtain the network reliability $R_A^N \equiv p_{BM}(A|A^O)$, that is, the probability that A is the true network given our observation A^O (and our choice of the family of stochastic block models). We obtain (*Materials and Methods*)

$$R_A^N = \frac{1}{Z} \sum_{P \in \mathcal{P}} h(A; A^O, P) \exp[-\mathcal{H}(P)], \qquad \textbf{[5]}$$

where

$$h(A; A^O, P) = \exp \left\{ \sum_{\alpha \le \beta} \left[\ln \left(\frac{r_{\alpha\beta} + 1}{2r_{\alpha\beta} + 1} \right) + \ln \left(\frac{\binom{r_{\alpha\beta}}{l_{\alpha\beta}^O}}{\binom{2r_{\alpha\beta}}{l_{\alpha\beta} + l_{\alpha\beta}^O}} \right) \right] \right\}, \qquad \textbf{[6]}$$

$l_{\alpha\beta}$ is the number of links in A between groups α and β, and $\mathcal{H}(P)$ and Z are the same as in Eq. **3**. Once more, we use the Metropolis algorithm to estimate R_A^N.

Given the network reliability $R_A^N = p_{BM}(A|A^O)$, the expected value of a property X

$$\langle X \rangle = \sum_A X(A) R_A^N \qquad \textbf{[7]}$$

over all possible networks A is a better estimate of $X(A^T)$ than $X(A^O)$. We find that in many situations $R_{A^T}^N \gg R_{A^O}^N$ (Fig. S13 in *SI Appendix*), which means that, presented only with an inaccurate observation A^O (and with the knowledge about complex networks embodied in the stochastic block model family), our approach is remarkably able to identify that A^T is a more likely network than the observation A^O itself. This confirms that, even without knowing A^T, it is possible to estimate a property $X(A^T)$ better than just by measuring that property on A^O (that is, better than assuming $X(A^T) = X(A^O)$).

Since summing over all possible networks in Eq. **7** is prohibitive, we use the approximation $\langle X \rangle \approx X(A^R)$, where A^R is the network that maximizes R_A^N (in other words, A^R is the maximum a posteriori estimate of A). The network A^R is what we call a network reconstruction, and we claim that $X(A^R)$ is, in general, a better estimate of $X(A^T)$ than $X(A^O)$. In practice, we build reconstructions by heuristically maximizing R_A^N, starting from A^O (*Materials and Methods*).

We test our network reconstruction approach by generating hypothetical observed networks A^O from the true test networks A^T described above. Each observation has a fraction of the true interactions removed (we call this fraction the observation error rate), and an identical number of random interactions added. In Fig. 3 we show the true air transportation network of Eastern Europe, as well as a hypothetical observation of this network (with an observation error rate of 20%) and the corresponding reconstruction. The reconstruction has 13% fewer missing and spurious interactions than the observation, and qualitatively it appears that individual node properties (specifically, degree and betweenness centrality) are also better captured by the reconstruction.
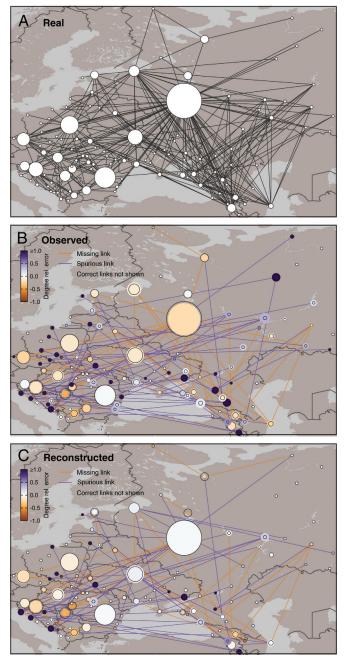

Fig. 3. Reconstruction of the air transportation network of Eastern Europe. (*A*) The true air transportation network. The area of each node is proportional to its betweenness centrality, with Moscow being the most central node in the network. (*B*) The observed air transportation network, which we build by randomly removing 20% of the real links and replacing them by random links. (*C*) The reconstructed air transportation network that we obtain, from the observed network, applying the heuristic reconstruction method described in the text and methods. For clarity, in *B* (respectively, *C*) we do not depict the correct links, but only (*i*) missing links in orange, which exist in the true network but not in the observation (reconstruction) and (*ii*) spurious links in blue, which do not exist in the true network but do exist in the observation (reconstruction). As in *A* the area of each node is proportional to its betweenness centrality, with the black circle representing the true betweenness centrality of each node. The color of each node represents the relative error in the degree of the node, with respect to the true degree. The observed network contains 60 missing and 60 spurious links, whereas the reconstruction only contains 52 of each (a 13% improvement). In general, node degree and betweenness centrality are also better captured in the reconstruction.
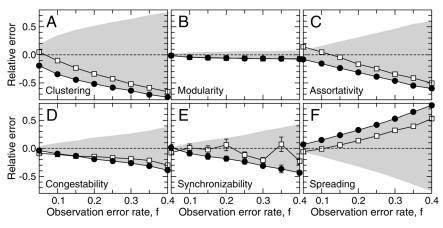
Fig. 4. Properties of observations and reconstructions of the air transportation network of Eastern Europe. In each case, the observation A^O is generated from the true network A^T by randomizing a fraction f of its links. The reconstruction A^R is generated from A^O as described in the text and methods. For each property X, we calculate the relative error of the observation $(X(A^O) - X(A^T))/X(A^T)$ (black circles) and of the reconstruction $(X(A^R) - X(A^T))/X(A^T)$ (white squares). Symbols represent the mean over 25 repetitions, and the error bars indicate the standard error of the mean. The shaded region corresponds to the region with smaller relative error (in absolute value) than the observation, so that squares within the shaded region correspond to reconstructions that provide better network-property estimates than the observation itself. (*A–C*) Static properties: clustering coefficient (30) (*A*), modularity (30) (*B*), assortativity (30) (*C*). (*D–F*) Dynamic properties: transportation congestability, that is, the maximum betweenness centrality in the network (31) (*D*), synchronizability, that is, the ratio between the largest eigenvalue and the smallest non-zero eigenvalue of the Laplacian matrix of the network (32) (*E*), spreading threshold, that is, the ratio between the first and the second moments of the degree distribution (33) (*F*).

However, from a systems perspective global network properties are more relevant than local node-level features. Therefore, the ultimate goal is to generate network reconstructions whose global properties are closer to those of the true network than those of the observations. To quantitatively investigate whether our approach accomplishes this aim, we calculate six network properties (static and dynamic) for observations and for the corresponding reconstructions of the air transportation network of Eastern Europe and compute the relative error with respect to the true value. As we show in Fig. 4, the reconstruction consistently improves the estimates of these properties. Only when the observed network contains <10% of errors it is better, for a few of the properties, to use the observed network rather than the reconstruction. We obtain similar results for the other test networks and other network properties (Figs. S8–S12 in *SI Appendix*).

Application to a Protein Interaction Network

As we have discussed before, protein interaction networks are among the networks that may benefit the most from our approach. Ultimately, only experiments can prove our results useful; such experiments are, however, beyond the scope of this work. Nevertheless, here we show how our approach can help in directing the effort to refine protein interaction data.

We consider the protein interaction network of yeast that Gavin et al. (12) obtained using affinity purification and mass spectrometry (AP/MS). In AP/MS essays, a "bait" protein is used to detect "prey" proteins that interact with the bait directly or indirectly. Since bait and prey play different roles, we limit ourselves to the set of 991 proteins that are both viable baits and viable prey (10) (for example, we discard proteins that only appear as prey because prey-prey interactions cannot possibly be observed). We build a protein interaction network by connecting all pairs of proteins that are reported as a bait-prey pair at least once.[¶] From this network, we obtain the reliability for all pairs of proteins.

We evaluate how successful our algorithm is by considering those proteins among the 991 in the network that have been used once, and only once, as bait (some proteins are used as bait in

several independent essays). For a pair of these proteins A and B, a link in the network can represent two distinct situations: (*i*) The interaction was observed once (with A as bait and B as prey but not the other way around, or vice versa); (*ii*) the interaction was observed twice (both with A as bait and with B as bait). Since these experimentally "non-reproducible" and "reproducible" interactions (3,113 and 867 interactions in our network, respectively) are encoded identically in the network, it is interesting to see if our algorithm assigns lower reliability to the first and higher to the latter.

Remarkably, among the 100 interactions with the lowest link reliability according to our algorithm, only five are experimentally reproducible. Conversely, among the 100 interactions with the highest link reliability, as many as 65 are experimentally reproducible. The probabilities of observing by chance such a small number in the first case and such a large number in the latter case are $p_\leq = 3 \times 10^{-6}$ and $p_\geq = 2 \times 10^{-20}$, respectively. Our approach is therefore successfully separating interactions that are likely to be spurious from those that are likely to be correct, without using any biophysical or biochemical information.

Discussion

We have shown that our network reconstruction method allows for a better characterization of network data sets, which will be particularly useful in data sets that we know contain many inaccuracies, such as protein interactomes. We have also shown that our approach reliably identifies missing and spurious interactions in complex networks, so that we can identify suspect interactions for further experimental probing.

Interestingly, our method can also guide new discoveries. If a given interaction between i and j truly exists but our approach predicts a very low reliability for the interaction (or vice versa), it means that the function of the interaction is very specific (since the interaction is rare among nodes that are otherwise similar to i and j) and, therefore, functionally or evolutionarily important.

Finally, our approach is flexible enough to allow generalizations in several directions. Arguably the most important of these is the extension to arbitrarily sophisticated families of models. In particular, one could use models that are the "product" of a network model M_n (probably a block model) and an error model M_e that incorporates the relevant error structure (maybe another block model with non-uniform priors). The flexibility of our approach, along with its generality and its performance, will

---

[¶] Note that we do not advocate that this is the most appropriate procedure to analyze the structure of a protein interaction network (see ref. 10 for a detailed discussion). Rather, we use this procedure because it enables us to test whether our algorithm can separate the least reliable from most reliable interactions.

make it applicable to many areas where network data reliability is a source of concern.

Materials and Methods

Outline of the Reliability Calculations. Formally, a block model $M = (P, Q)$ is completely determined by the partition P of nodes into groups and the matrix Q of probabilities of linkage between groups, so that Eq. **2** in the main text can be rewritten as

$$p_{BM}(X = x|A^O) = \frac{1}{Z} \sum_{P \in \mathcal{P}} \int_{[0,1]^G} dQ\, p(X = x|P, Q) \times$$
$$\times p_{BM}(A^O|P, Q)\, p(P, Q),\qquad\qquad \textbf{[8]}$$

where $\mathcal{P}$ is the space of all possible partitions of the network into groups, G is the number of distinct group pairs, and Z is a normalizing constant.

Within the family of stochastic block models, one can evaluate the likelihood of each model M because the probability of any two nodes i and j being connected depends only on the groups to which they belong. We have that (14)

$$p_{BM}(A^O|P, Q) = \prod_{\alpha \leq \beta} Q_{\alpha\beta}^{l_{\alpha\beta}^O} \left(1 - Q_{\alpha\beta}\right)^{r_{\alpha\beta} - l_{\alpha\beta}^O},\qquad \textbf{[9]}$$

where $l_{\alpha\beta}^O$ is the number of links in A^O between nodes in groups α and β of P, and $r_{\alpha\beta}$ is the maximum number of such links (that is, the number of pairs of nodes such that one node is in α and the other is in β).

Note that, among all possible block models, there is at least one whose likelihood is 1, namely, the block model in which each node is in a different block and each $Q_{\alpha\beta}$ is 1 or 0 depending on whether the corresponding nodes are connected or not. This model contributes to $p(X = x|A^O)$ much more than most other models. However, there is only one such model (or very few), whereas there are many models with, for example, four blocks. This "entropic" term prevents overfitting of the network by very detailed (and ultimately uninformative) block models.

Using that $p(A_{ij} = 1|P, Q) = Q_{\sigma_i \sigma_j}$ (where σ_i is the module of node i in partition P) and assuming no prior knowledge about the models (that is, $p(P, Q) = $ const.), one can use Eqs. **8** and **9** to obtain Eqs. **3**–**6** in the main text.

Metropolis Estimation of Link and Network Reliability. To estimate the link and network reliabilities given by Eqs. **3** and **5** we use the following procedure. We start by placing each of the N nodes in a group, which we choose with uniform probability from a set of N possible groups (that is, there are as many groups as nodes). In general, some of these N groups will be empty after the initial node assignment.

At each step we select a random node and attempt to move it to a randomly selected group. This update scheme is appropriate because (*i*) it results in an ergodic exploration of the space of possible partitions and (*ii*) it satisfies detailed balance (since the probability of choosing a move and its reverse are identical). To decide whether we accept the move, we calculate the change

$\Delta\mathcal{H}$ (Eq. **4**): If $\Delta\mathcal{H} \leq 0$, the change is automatically accepted; otherwise, the change is accepted with probability $\exp(-\Delta\mathcal{H})$.

The sampling procedure starts after an equilibration period, during which $\mathcal{H}$ decreases from an initial value to its equilibrium value. We sample the partition space by considering $S = 10^4$ partitions, each one separated from the previous one by a number of steps that is large enough for the two partitions to be reasonably uncorrelated (as measured by the mutual information between partitions).

Because the link and network reliabilities are ensemble averages over independent partitions, it is straightforward to parallelize the algorithm so that the partitions are obtained concurrently. Therefore, given enough computational resources, the reliabilities can be calculated even for large networks (probably up to millions of nodes) in relatively short times.

Benchmark Algorithms for the Identification of Missing and Spurious Interactions. The hierarchical random graph approach is described in detail in (11). We use the implementation provided by the authors (available at www.santafe.edu/~aaronc/hierarchy/hrg_20080819_predictHRG_v1.0.3.tgz), which we modified slightly to be able to study spurious as well as missing interactions.

We analyze three local algorithms: common neighbors, degree product, and Jaccard index (the last two in *SI Appendix* only). For each of these algorithms, the link "reliability" R_{ij}^L is defined as follows (note that, for these approaches, the "reliability" is not a probability, but just a score that enables us to rank node pairs):

- Common neighbors: $R_{ij}^L = \|\Gamma_i \cap \Gamma_j\|$, where Γ_i is the set of neighbors of node i, and $\|\ldots\|$ indicates the number of nodes in a set.
- Degree product: $R_{ij}^L = \|\Gamma_i\| \times \|\Gamma_j\|$.
- Jaccard index: $R_{ij}^L = \|\Gamma_i \cap \Gamma_j\| / \|\Gamma_i \cup \Gamma_j\|$.

Heuristic Network Reconstruction. The goal of the heuristic network reconstruction algorithm is to find $A^R = \arg\max_A R_A^N$, where $R_A^N = p(A|A^O)$ is the reliability of network A given observation A^O. Since exhaustive maximization of R_A^N is not possible, we use the following heuristic method. Start by evaluating the link reliabilities R_{ij}^L for all pairs of nodes in A^O; sort observed links ($A_{ij}^O = 1$) by increasing reliability, and observed non-links ($A_{ij}^O = 0$) by decreasing reliability. Then choose pairs of link/non-link in order: remove the link (which has a low reliability) and add the non-link (which has a high reliability), and accept the change if, and only if, R_A^N increases. Repeat this procedure, going down the lists, until we reject five consecutive attempts to swap a link/non-link pair. At this point, reevaluate R_{ij}^L and repeat the process. The algorithm stops when no link swaps are accepted.

1. Amaral LAN, Ottino JM (2004) Complex networks: Augmenting the framework for the study of complex systems. *Eur Phys J B* 38:147–162.
2. von Mering C, et al. (2002) Comparative assessment of large-scale data sets of protein-protein interactions. *Nature* 417:399–403.
3. Yu H, et al. (2008) High-quality binary protein interaction map of the yeast interactome network. *Science* 322:104–110.
4. Stumpf MPH, et al. (2008) Estimating the size of the human interactome. *Proc Natl Acad Sci USA* 105:6959–6964.
5. Amaral LAN (2008) A truer measure of our ignorance. *Proc Natl Acad Sci USA* 105:6795–6796.
6. Schafer JL, Graham JW (2002) Missing data: Our view of the state of the art. *Psychol Methods* 7:147–177.
7. Butts CT (2003) Network inference, error, and informant (in)accuracy: A Bayesian approach. *Soc Networks* 25:103–140.
8. Kossinets G (2006) Effects of missing data in social networks. *Soc Networks* 28:247–268.
9. Jansen R, et al. (2003) A bayesian networks approach for predicting protein-protein interactions from genomic data. *Science* 302:449–453.
10. Chiang T, Scholtens D (2009) A general pipeline for quality and statistical assessment of protein interaction data with r and bioconductor. *Nat Protoc* 4:535–546.
11. Clauset A, Moore C, Newman MEJ (2008) Hierarchical structure and the prediction of missing links in networks. *Nature* 453:98–101.
12. Gavin AC, et al. (2006) Proteome survey reveals modularity of the yeast cell machinery. *Nature* 440:631–636.
13. White HC, Boorman SA, Breiger RL (1976) Social structure from multiple networks. I. Blockmodels of roles and positions. *Am J Sociol* 81:730–780.
14. Holland PW, Laskey KB, Leinhardt S (1983) Stochastic blockmodels: First steps. *Soc Networks* 5:109–137.
15. Girvan M, Newman MEJ (2002) Community structure in social and biological networks. *Proc Natl Acad Sci USA* 99:7821–7826.
16. Guimerà R, Sales-Pardo M, Amaral LAN (2007) Classes of complex networks defined by role-to-role connectivity profiles. *Nat Phys* 3:63–69.
17. Sawardecker EN, Sales-Pardo M, Amaral LAN (2009) Detection of node group membership in networks with group overlap. *Eur Phys J B* 67:277–284.
18. Sales-Pardo M, Guimerà R, Moreira AA, Amaral LAN (2007) Extracting the hierarchical organization of complex systems. *Proc Natl Acad Sci USA* 104:15224–15229.
19. Arenas A, Díaz-Guilera A, Pérez-Vicente CJ (2006) Synchronization reveals topological scales in complex networks. *Phys Rev Lett* 96:114102.
20. Guimerà R, Sales-Pardo M, Amaral LAN (2004) Modularity from fluctuations in random graphs and complex networks. *Phys Rev E* 70:025101.
21. Newman MEJ (2003) Mixing patterns in networks. *Phys Rev E* 67:026126.
22. Huang K (1987) *Statistical Mechanics* (Wiley, Hoboken, NJ), 2nd Ed.
23. Zachary W (1977) An information flow model for conflict and fission in small groups. *J Anthropol Res* 33:452–473.
24. Lusseau D, et al. (2003) The bottlenose dolphin community of Doubtful Sound features a large proportion of long-lasting associations. *Behav Ecol Sociobiol* 54:396–405.
25. Guimerà R, Mossa S, Turtschi A, Amaral LAN (2005) The worldwide air transportation network: Anomalous centrality, community structure, and cities' global roles. *Proc Natl Acad Sci USA* 102:7794–7799.
26. White JG, Southgate E, Thomson JN, Brenner S (1986) The structure of the nervous system of the nematode *C. elegans. Phil Trans R Soc B* 314:1–340.
27. Reed JL, Vo TD, Schilling CH, Palsson BØ (2003) An expanded genome-scale model of Escherichia coli K-12 (iJR904 GSM/GPR). *Genome Biol* 4:R54.
28. Guimerà R, Sales-Pardo M, Amaral LAN (2007) A network-based method for target selection in metabolic networks. *Bioinformatics* 23:1616–1622.
29. Liben-Nowell D, Kleinberg J (2007) The link-prediction problem for social networks. *J Am Soc Inf Sci Technol* 58:1019–1031.
30. Costa LDF, Rodrigues FA, Travieso G, Boas PRV (2007) Characterization of complex networks: A survey of measurements. *Adv Phys* 56:167–242.
31. Guimerà R, Díaz-Guilera A, Vega-Redondo F, Cabrales A, Arenas A (2002) Optimal network topologies for local search with congestion. *Phys Rev Lett* 89:248701.
32. Arenas A, Díaz-Guilera A, Kurths J, Moreno Y, Zhou C (2008) Synchronization in complex networks. *Phys Rep* 469:93–153.
33. Boguñá M, Pastor-Satorras R (2002) Epidemic spreading in correlated complex networks. *Phys Rev E* 66:047104.