JOURNAL OF THE ROYAL SOCIETY Interface

Incomplete and noisy network data as a percolation process

Michael P. H. Stumpf and Carsten Wiuf

J. R. Soc. Interface 2010 **7**, 1411-1419 first published online 8 April 2010
doi: 10.1098/rsif.2010.0044

JOURNAL OF THE ROYAL SOCIETY Interface

J. R. Soc. Interface (2010) **7**, 1411–1419
doi:10.1098/rsif.2010.0044
Published online 8 April 2010

Incomplete and noisy network data as a percolation process

Michael P. H. Stumpf[1,2,*] **and Carsten Wiuf**[1,3,*]

[1]*Centre for Bioinformatics, Division of Molecular Biosciences, and* [2]*Institute of Mathematical Sciences, Imperial College London, London SW72AZ, UK*
[3]*Bioinformatics Research Center, University of Aarhus, 8000 Aarhus C, Denmark*

We discuss the ramifications of noisy and incomplete observations of network data on the existence of a giant connected component (GCC). The existence of a GCC in a random graph can be described in terms of a percolation process, and building on general results for classes of random graphs with specified degree distributions we derive percolation thresholds above which GCCs exist. We show that sampling and noise can have a profound effect on the perceived existence of a GCC and find that both processes can destroy it. We also show that the absence of a GCC puts a theoretical upper bound on the false-positive rate and relate our percolation analysis to experimental protein−protein interaction data.

Keywords: complex networks; random graphs; protein interaction networks; sampling problems

1. INTRODUCTION

In many areas of the physical, engineering, life- and social sciences network data are now becoming increasingly abundant. In light of empirical network data, mathematical models of networks (Durrett 2006) and the statistical tools for their analysis have changed from the beginnings of random graph theory and improved considerably over the past 10 years. However, despite these advances, not much effort has been devoted to detailed investigation and modelling of the effects of erroneous network data.

While for technological networks sampling (Stumpf *et al.* 2005*b*) and noise are of relatively minor importance, for biological and social networks the situation is markedly different. Especially in biology, there appears to be a genuine trade-off between data quality and the quantity in which network data are being generated. In particular, in the context of protein-interaction network data, the false-positive and -negative rates have been well documented with reported error rates frequently exceeding 50 per cent (von Mering *et al.* 2002; Bork *et al.* 2004; Reguly *et al.* 2006). With such high error levels, it may not be automatically expected that inferences obtained from noisy networks are informative about the true network, and incorporating a detailed analysis of the effects of noise has now become widely accepted practice (Middendorf *et al.* 2004; Yook *et al.* 2004; deSilva *et al.* 2006). Yet, wherever analyses were repeated on artificially perturbed networks (noise was introduced by adding false-positive interactions and deleting reported interactions) the perturbed networks had similar properties to the original network (Middendorf *et al.* 2004; Yook *et al.* 2004; de Silva *et al.* 2006). The observation that some networks have similar properties when perturbed in this way may suggest that some balance between false positives and false negatives has already been obtained.

Of the reported genome-wide protein interaction surveys that have been published to date, only a small number (Ito *et al.* 2001; Gavin *et al.* 2002) do not show evidence for the existence of a giant connected component (GCC). This puzzling observation—*a priori* we would assume that the molecular machinery underlying living systems is highly interconnected and tightly linked—suggests that noise can induce a percolation transition (Stauffer & Aharony 1992) on real networks. Here we describe percolation transitions on general random networks due to network sampling (Han *et al.* 2005; Stumpf & Wiuf 2005; Stumpf *et al.* 2005*b*; de Silva *et al.* 2006) and noise in the network observation. This will serve to highlight some differences in the way noise and incompleteness of networks will affect our observations. We will also show that the very existence of a GCC sets an upper limit on the false-positive rate in interaction data.

2. NOISY AND INCOMPLETE NETWORKS AND PERCOLATION

Here we introduce the notion of noisy and incomplete network ensembles (NINEs), which extend the standard notion of network ensembles (Burda *et al.* 2001) in order to account for the differences between true networks, and the noisy and incomplete representations thereoff that we can observe.

*Authors for correspondence (m.stumpf@imperial.ac.uk, wiuf@birc.au.dk).

Received 1 February 2010
Accepted 18 March 2010

2.1. Notation

In order to define NINEs, we first assume that we have a true and complete network described by a graph,

$$\mathcal{G} = (\mathcal{V}, \mathcal{E}), \tag{2.1}$$

which consists of a set of nodes or vertices, $\mathcal{V}$, and a set of edges among the nodes, $\mathcal{E}$. The order and size of the network are defined as the number of nodes, N (or the size of $\mathcal{V}$), and edges, M (or the size of $\mathcal{E}$), respectively,

$$N = |\mathcal{V}| \quad \text{and} \quad M = |\mathcal{E}|.$$

Now let $\mathcal{V}_\mathrm{S}$ and $\mathcal{E}_\mathrm{S}$ denote subsets of $\mathcal{V}$ and $\mathcal{E}$, respectively, with the property

$$i \in \mathcal{V}_\mathrm{S} \text{ with probability } p \text{ if } i \in \mathcal{V} \tag{2.2}$$

and that $e(i,j) \in \mathcal{E}_\mathrm{S}$ iff $i,j \in \mathcal{V}_\mathrm{S}$ and $e(i,j) \in \mathcal{E}$. Thus, the graph

$$\mathcal{G}_\mathrm{S} = (\mathcal{V}_\mathrm{S}, \mathcal{E}_\mathrm{S}) \tag{2.3}$$

is the subgraph of $\mathcal{G}$ that is induced by $\mathcal{V}_\mathrm{S} \subseteq \mathcal{V}$. Trivially, we have $\mathcal{E}_\mathrm{S} \subseteq \mathcal{E}$ for noiseless data.

Equally, we define noisy networks via

$$\mathcal{G}_\mathrm{D} = (\mathcal{V}_\mathrm{D}, \mathcal{E}_\mathrm{D}), \tag{2.4}$$

where $\mathcal{V}_\mathrm{D} = \mathcal{V}$ and the set of edges $\mathcal{E}_\mathrm{D}$ is defined by

$$e(i,j) \in \mathcal{E}_\mathrm{D} \text{ with probability } \begin{cases} \rho \text{ if } e(i,j) \in \mathcal{E}, \\ \xi \text{ if } e(i,j) \notin \mathcal{E}. \end{cases} \tag{2.5}$$

Thus ρ and ξ are the true-positive and false-positive rates, respectively. Trivially, $E[M_\mathrm{D}] = \rho M + \xi\left(\binom{N}{2} - M\right)$.

We can also consider the case of a noisy–incomplete network by simply replacing $\mathcal{V}$ and $\mathcal{E}$ by $\mathcal{V}_\mathrm{S}$ and $\mathcal{E}_\mathrm{S}$ in equations (2.4) and (2.5). We will later introduce different schemes for specifying p, ρ and ξ, and study their interplay in shaping observed network data.

2.2. Percolation on networks

Below we consider general networks and seek to derive conditions for the GCC to emerge in the context of noise and incompleteness. The relative size of a component is defined as the number of nodes in the component divided by the number of non-zero degree nodes. The GCC has non-zero relative size as the size of the network becomes large.

In the theoretical section we assume that the network is uncorrelated and large and hence it is sufficient to know the degree sequence or the probability distribution generating the degree sequence. Networks given by their degree sequence were first studied by Hakimi (1962), and later in more detail by Bender & Canfield (1978) and Molloy & Reed (1995, 1998). The latter also studied percolation processes, i.e. in the context of random graphs the emergence of the GCC (Bollobás & Riordan 2006). We shall, however, refer to these graphs as Bender–Canfield (BC) random graphs. This has since been studied repeatedly in relationship to real networks (Callaway *et al.* 2000; Newman *et al.* 2001). The recent monographs by Durrett (2006) and Chung & Lu (2006) provide excellent surveys of this area from (predominantly) probabilistic and combinatorial perspectives, respectively.

The central result of Molloy & Reed (1995), frequently referred to as the Molloy–Reed (MR) criterion is given as follows.

Theorem 1. *A Graph* $\mathcal{G} = (\mathcal{V},\mathcal{E})$ *has a GCC with high probability as* $\mathrm{N} \to \infty$ *iff*

$$z_2 > z_1, \tag{2.6}$$

where z_1 *and* z_2 *refer to the average numbers of nearest and next-nearest neighbours, respectively.*

The proof of this statement can be found in Molloy & Reed (1995). Different perspectives are also offered by Durrett (2006). Here, we are content with analysing the behaviour of the GCC using equation (2.6) in the context of noisy and incomplete networks. In particular we are interested in uncorrelated large BC networks, where we need only to consider the degree distribution, $\mathrm{Pr}(k)$ (rather than the actual sequence of integers). Below, we denote the first and second moments of the degree distribution by $\langle k \rangle$ and $\langle k^2 \rangle$, respectively.

2.3. Calculating $\mathbf{z}_1$ *and* $\mathbf{z}_2$

Calculation of z_1 and z_2 is straightforward for the cases we consider and well covered in the literature, e.g. Dorogovtsev & Mendes (2003) or Burda & Krzywicki (2004); here it is only briefly repeated for the sake of completeness. The number of nearest neighbours, z_1, is given by the average degree,

$$z_1 = \langle k \rangle = \sum_{k'=0}^{\infty} k' \mathrm{Pr}(k'). \tag{2.7}$$

In uncorrelated graphs, the probability of a random edge ending in a node of degree l is

$$\frac{l\mathrm{Pr}(l)}{\langle k \rangle}. \tag{2.8}$$

We can obtain z_2 by summing over the probabilities of a node having degree k' multiplied by the average degrees of the k neighbours, i.e.

$$\begin{aligned} z_2 &= \sum_{k'} k' \mathrm{Pr}(k') \sum_{l} l \frac{(l+1)\mathrm{Pr}(l+1)}{\langle k \rangle} \\ &= \sum_{k'} k' \mathrm{Pr}(k') \frac{\langle k^2 \rangle - \langle k \rangle}{\langle k \rangle} = \langle k^2 \rangle - \langle k \rangle. \end{aligned} \tag{2.9}$$

For classical or Erdös–Rényi (Erdös & Rényi 1959, 1960) random graphs, which are characterized by an approximately Poisson degree distribution, the MR criterion yields the well-known result that the GCC appears as $\lambda > 1$, where λ is the average degree in the network.

Equations (2.7) and (2.9) together with equation (2.6) allow us to study the percolation behaviour of uncorrelated BC graphs by simply determining the functional form for the degree distribution in the incomplete, $\mathcal{G}_{\rm S}$, and noisy, $\mathcal{G}_{\rm D}$, network ensembles. Below, we will study the percolation transitions in incomplete and noisy BC networks, and we will briefly comment on the validity of these results for real correlated networks (Berg & Lässig 2002).

3. PERCOLATION PROCESSES DUE TO INCOMPLETE AND UNRELIABLE NETWORK DATA

Noise and incompleteness are known to affect present network data sets; in particular, most biological network data sets are subject to considerable uncertainty (Han *et al.* 2005; de Silva *et al.* 2006). For incomplete network data, several studies have recently made some theoretical progress (Stumpf & Wiuf 2005; Stumpf *et al.* 2005*b*; Lee *et al.* 2006; Wiuf & Stumpf 2006), while investigations of the effects of noise have thus far been confined to semi-rigorous assessments of experimental data sets, in particular regarding protein interaction data (von Mering *et al.* 2002; Hart *et al.* 2006).

3.1. *Percolation due to network sampling*

Let p be the probability of sampling a node to be included in the set of subnet vertices $\mathcal{V}_{\rm S}$. We can either specify $p \in (0,1)$ or set p to the fraction of sampled nodes $p = N_{\rm S}/N$. The degree distribution in the subnet ${\rm Pr}_{\rm S}(k)$ is then given by (Stumpf & Wiuf 2005; Lee *et al.* 2006),

$$\Pr_{\rm S}(k) = \sum_{l \geq k} \binom{l}{k} p^k (1-p)^{l-k} \Pr(l). \qquad (3.1)$$

Thus, the degree distributions in the true network and random subnets are generally of a different functional form.

The number of nearest and next-nearest neighbours in the subnet, $z_1^{\rm S}$ and $z_2^{\rm S}$, respectively, are straightforward to calculate and we can find the condition for which the MR criterion, equation (2.6), is fulfilled. We obtain for the number of nearest neighbours in the subnet,

$$z_1^{\rm S} = p z_1 \qquad (3.2)$$

and the next nearest neighbours,

$$z_2^{\rm S} = \langle k^2 \rangle_S - \langle k \rangle_S = p^2 \langle k^2 \rangle - p^2 \langle k \rangle = p^2 z_2. \qquad (3.3)$$

Taken together with the MR criterion we can thus derive the critical value for the sampling fraction which indicates the appearance of the GCC,

$$p_c = \frac{\langle k \rangle}{\langle k^2 \rangle - \langle k \rangle} = \frac{z_1}{z_2}. \qquad (3.4)$$

From equation (3.4), it is obvious that for degree distributions with a diverging second moment, the GCC only disappears as $p_c \to 0$; of course, this is only relevant in the limit, $N \to \infty$.

3.2. *Percolation due to noisy network data*

Here we consider how observational noise affects the GCC, i.e. if each edge is observed with (true-positive) probability $\rho \in [0,1]$, and each non-existing edge is observed with (false-positive) probability $\xi \in [0,1]$, do we observe the GCC or not?

We start again by determining the degree distribution in a network with given error rates, ρ and ξ. Assuming that a node has degree k in $\mathcal{G}$, the degree of a node in $\mathcal{G}_{\rm D}$ is the sum of two binomial variables, one controlling the number of false positives, the other the number of false negatives. Hence, the node has degree l in $\mathcal{G}_{\rm D}$ with probability,

$$\begin{aligned}\Pr_{\rm D}(l|k) &= \sum_{i=0}^{k} \binom{k}{i} \rho^i (1-\rho)^{k-i} \times \binom{N-1-k}{l-i} \\ &\quad \times \xi^{l-i} (1-\xi)^{N-k-1-l+i} \\ &= \frac{(N-1-k)!}{l!(N-1-k-l)!} (1-\rho)^k (1-\xi)^{N-1-k-l} \xi^l \\ &\quad \times {}_2F_1\left(-k, -l, N-k-l, \frac{\rho(1-\xi)}{\xi(1-\rho)}\right),\end{aligned} \qquad (3.5)$$

where ${}_2F_1$ $(a,\ b,\ c,\ x)$ is the hypergeometric function (Gradshteyn *et al.* 1994; Arfken & Weber 2005). We note that $l > k$ may occur as the sum of the numbers of observed true positive and false positive edges may be larger than the number of true edges. For the degree distribution of the noisy network, $\mathcal{G}_{\rm D}$, we thus have

$$\begin{aligned}\Pr_{\rm D}(l) &= \sum_{k=0}^{N-l} \frac{(N-1-k)!}{l!(N-1-k-l)!} (1-\rho)^k (1-\xi)^{N-1-k-l} \xi^l \\ &\quad \times {}_2F_1\left(-k, -l, N-k-l, \frac{\rho(1-\xi)}{\xi(1-\rho)}\right) \Pr(k).\end{aligned} \qquad (3.6)$$

While this expression is clearly intractable, the two moments $\langle k \rangle_{\rm D}$ and $\langle k^2 \rangle_{\rm D}$ have tractable analytic expressions in terms of $\langle k \rangle$ and $\langle k^2 \rangle$. Here we find

$$\langle k \rangle_{\rm D} = (\rho - \xi)\langle k \rangle + (N-1)\xi \qquad (3.7)$$

and

$$\begin{aligned}\langle k^2 \rangle_{\rm D} &= (\rho-\xi)^2 \langle k^2 \rangle - (\rho-\xi)^2 \langle k \rangle \\ &\quad + 2(N-2)\xi(\rho-\xi)\langle k \rangle + \xi^2 (N-1)(N-2) + \langle k \rangle_{\rm D}.\end{aligned} \qquad (3.8)$$

Next, we define $\delta = \rho - \xi$; then $z_1^{\rm D}$ and $z_2^{\rm D}$ are given by

$$z_1^{\rm D} = \delta\langle k \rangle + (N-1)\xi \qquad (3.9)$$

and

$$\begin{aligned}z_2^{\rm D} &= \delta^2 \langle k^2 \rangle - \delta^2 \langle k \rangle + 2(N-2)\xi\delta\langle k \rangle \\ &\quad + \xi^2 (N-1)(N-2).\end{aligned} \qquad (3.10)$$

Note, that many real networks show average degrees that are not comparable to the entire network size N. It suggests that it is reasonable to consider the situation

$\xi \approx 0$ with $\Xi = N\xi$, and $\delta \approx \rho$ (assuming $\rho \gg \xi$) in which case z_1^{D} and z_2^{D} become

$$z_1^{\mathrm{D}} = \rho\langle k\rangle + \Xi \tag{3.11}$$

and

$$z_2^{\mathrm{D}} = \rho^2\langle k^2\rangle - \rho^2\langle k\rangle + 2\Xi\rho\langle k\rangle + \Xi^2. \tag{3.12}$$

The parameter Ξ is essentially the expected number of false positives per node. Together with equation (2.6), equations (3.11) and (3.12) allow us to determine the critical boundary where the GCC appears. For fixed ρ (or Ξ) we can thus determine the critical value of Ξ (or ρ) where the percolation transistion occurs. In contrast to network sampling, we see that depending on the parameter values the GCC might appear even if it is absent in the true network.

One important observation is that if $\Xi \geq 1$ then $z_2^{\mathrm{D}} > z_1^{\mathrm{D}}$ always, i.e. if more than one false positive is expected per node a GCC is present. This is not surprising since the number of false positives approximately follows a Poisson distribution with parameter Ξ for large networks. It also transpires from equations (3.11) and (3.12) that ρ essentially affects the number of nearest and next-nearest neighbours in the same way as p does under network sampling; though in addition there is a term that also depends on Ξ. The quantities z_1^{S} and z_2^{S} might be obtained from z_1^{D} and z_2^{D} by letting $\rho = p$ and $\Xi = 0$.

3.3. Percolation due to the combined effects of noise and sampling

Combining sampling and noise is straightforward. There is, however, a subtle—so subtle to be undetectable in practice in fact—dependence on the order of sampling and noise. If we first choose the set of nodes, $\mathcal{V}_{\mathrm{S}}$ to assay for interactions, and then observe interactions subject to noise, we find

$$z_1^{\mathrm{SD}} = p\delta\langle k\rangle + (pN-1)\xi \tag{3.13}$$

and

$$z_2^{\mathrm{SD}} = p^2\delta^2\langle k^2\rangle - p^2\delta^2\langle k\rangle + 2p(Np-2)\xi\delta\langle k\rangle + (pN-1)(pN-2)\xi^2. \tag{3.14}$$

Should we, however, consider a subnet drawn from an already global but noisy network data set then we have

$$z_1^{\mathrm{DS}} = p\delta\langle k\rangle + p(N-1)\xi = pz_1^{\mathrm{D}} \tag{3.15}$$

and

$$z_2^{\mathrm{DS}} = p^2\delta^2\langle k^2\rangle - p^2\delta^2\langle k\rangle + 2p^2(N-2)\xi\delta\langle k\rangle + p^2(N-1)(N-2)\xi^2 = p^2 z_2^{\mathrm{D}} \tag{3.16}$$

In practice, however, the differences are so small that the order in which noise is added to the network and nodes are sampled from it does not matter, i.e.

$$z_1^{\mathrm{DS}} \approx z_1^{\mathrm{SD}} \tag{3.17}$$

and

$$z_2^{\mathrm{DS}} \approx z_2^{\mathrm{SD}}, \tag{3.18}$$

which is valid for $pN \gg 1$. The percolation transition again occurs when $z_1^{\mathrm{DS}} = z_2^{\mathrm{DS}}$. We reiterate that the MR criterion is derived in the limit of a large network and we note that as $N \to \infty$ or $N\xi \to \Xi$ the order in which sampling and noise enter the problem becomes irrelevant. But for very small networks this order may indeed matter.

3.4. Examples—theoretical network ensembles

We illustrate the percolation transitions using classical random graphs (Bollobás 2001) and scale-free random graphs (Barabasi & Albert 1999; Barabasi *et al.* 2001).

3.4.1. Classical random graphs. Classical random graphs are characterized by a binomial, or for sufficiently large graphs, Poisson degree distribution,

$$\Pr(k) = \frac{\lambda^k \exp(-\lambda)}{k!}. \tag{3.19}$$

The first and second moments are given by

$$\langle k\rangle = \lambda \quad \text{and} \quad \langle k^2\rangle = \lambda^2 + \lambda. \tag{3.20}$$

The percolation transition thus occurs at the critical sampling fraction

$$p_c = \frac{1}{\lambda}, \tag{3.21}$$

and a GCC is present if $p > 1/\lambda$. The noise-induced transition occurs at the critical values fulfilling

$$\rho_c\lambda = 1 - \Xi_c, \tag{3.22}$$

and a GCC is present if

$$\rho\lambda > 1 - \Xi. \tag{3.23}$$

Note that this condition reduces to equation (3.21) for $\Xi = 0$, as it should.

3.4.2. Scale-free random graphs. Scale-free random graphs exhibit a power-law degree distribution,

$$\Pr(k) = \frac{k^{-\gamma}}{\zeta_\gamma}, \tag{3.24}$$

where ζ_x is Riemann's zeta function and acts as a normalizing constant. For $\gamma \leq 3$, the second moment diverges and therefore neither sampling nor noise will induce a percolation transition on a scale-free random network unless $\gamma > 3$. Below we assume that we are dealing with a finite (though potentially very large, i.e. $N > 10^6$, scale-free random graph) with a degree sequence generated by drawing N random numbers from the distribution given by equation (3.24); in order to qualify as a proper degree distribution the sum of the degree sequence has, of course, to be even.

The sampling transition occurs at a critical value of the sampling fraction given by

$$p_c = \frac{\zeta_{\gamma-1}}{\zeta_{\gamma-2} - \zeta_{\gamma-1}}. \tag{3.25}$$

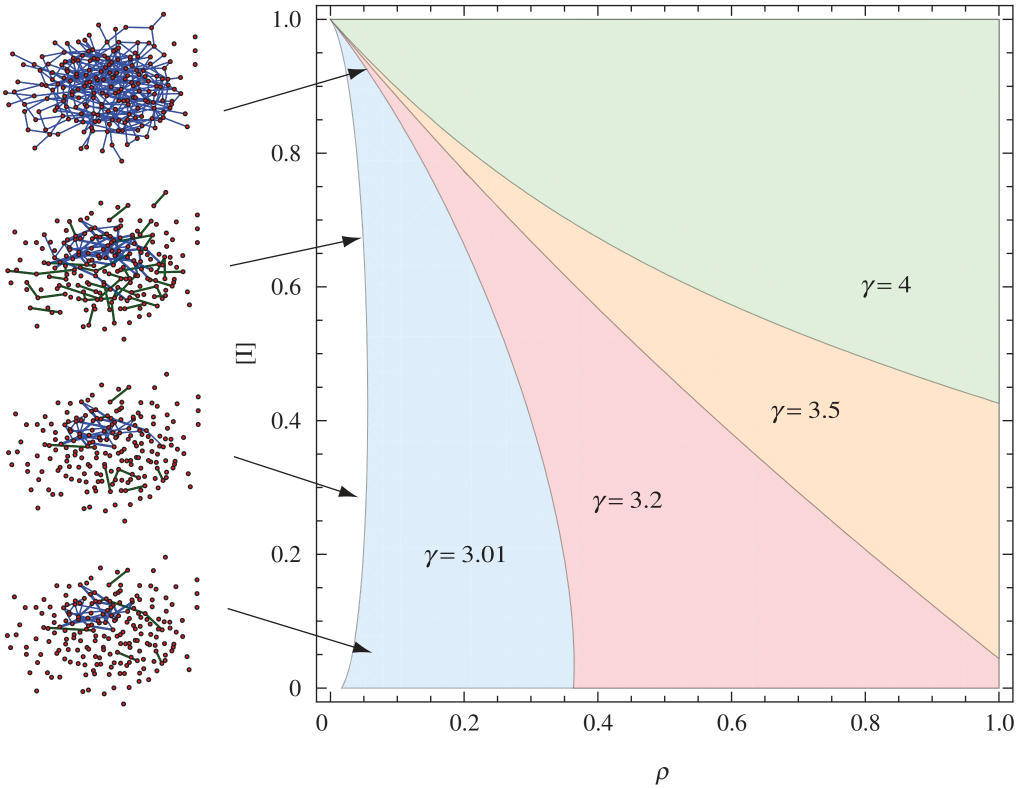

Figure 1. Contour-plot of the fraction of nodes contained in the GCC for different fractions of true-positive and false-positive edges in a network (area labels refer to the whole area to the right and above of the corresponding curves). Only when the fraction of true and false positives are both very small, will the GCC disappear. On the left-hand side of the figure we show illustrative examples of cases which correspond (loosely) to areas in the (ρ, Ξ) plane indicated by the arrow-tips. Real edges are indicated in blue, whereas false-positive edges are drawn in green.

Likewise, the noise-induced transition occurs when

$$\rho_c^2 \frac{\zeta_{\gamma-3} - \zeta_{\gamma-2}}{\zeta_{\gamma-1}} + (2\Xi_c - 1)\rho_c \frac{\zeta_{\gamma-2}}{\zeta_{\gamma-1}} + \Xi_c(\Xi_c - 1) = 0, \tag{3.26}$$

and if the left side is larger than zero for a given value of (ρ,Ξ) then a GCC appears. As stated before, this is automatically fulfilled for $\Xi \geq 1$. The areas where a GCC exists are shown in figure 1 for different values of the exponent γ.

Note that for some fixed values of ρ and $\Xi = 0$ the GCC is present, disappears for small non-zero values of Ξ and then reappears for larger values of Ξ. The explanation for this phenomomen is that the size of the GCC is relative to the number of non-zero degree nodes: for $\Xi = 0$ there are many zero degree nodes; when Ξ increases they become connected in many small disconnected components and for large Ξ, the small components are connected to form a GCC.

4. PERCOLATION ANALYSIS FOR FINITE SYSTEMS

Most real networks are finite in size and we would not expect to see a sharp percolation transition (i.e. the disappearance of a GCC) for either noise or sampling processes on networks. In figure 2, we show the effect of sampling on three different networks of order $N = 5000$; each network has 15 000 edges but was generated in a different way. From each network we sampled, at random, a fraction p of the nodes in the network and determined the components of the resulting induced subgraph. The networks considered are an Erdös–Rényi random graph (with $\lambda = 3$), a network generated by the Barabasi–Albert (BA) construction (Barabasi & Albert 1999) (at each timestep a new node was added and preferentially attached to the existing graph with three edges), and a corresponding BC random graph (i.e. we took the degree sequence of a BA graph and generated a randomly rewired graph with the same degree sequence).

The fraction of nodes which are contained in the GCC is virtually identical for the BA network and its BC counterpart. This suggests that, as far as the percolation behaviour is concerned, the type of network which is grown to order N under the BA preferential attachment model is not different from a corresponding BC network ensemble. That is the degree sequence determines the percolation behaviour as suggested by the MR criterion (Molloy & Reed 1995). For the Erdös–Rényi random graph we observe, of course, differences, and the transition appears to be somewhat sharper. In fact, as predicted by equation (3.19), the transition occurs in the vicinity of $1/\lambda = \frac{1}{3}$.

For noisy networks, the lack of simple analytic expressions makes the analysis of simulations somewhat more complicated. What we find in extensive

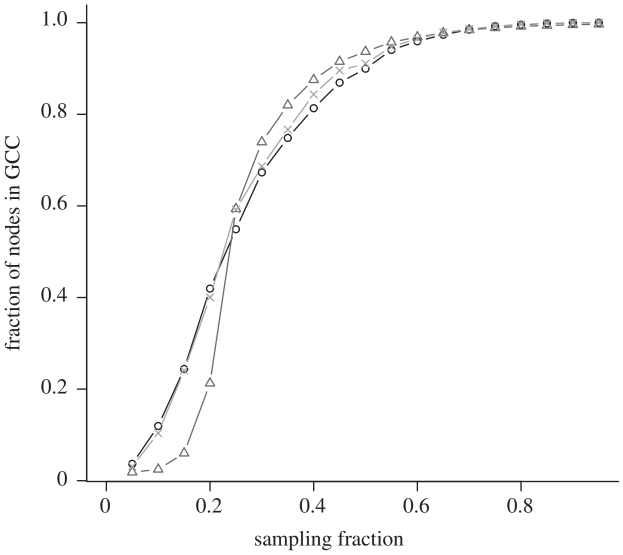

Figure 2. Average fraction of nodes in subnets of a network with 5000 nodes and 15 000 edges against sampling fraction/subnet size. Shown are results obtained for a scale-free network grown under the BA preferential attachment model (black), the corresponding BC graph (see text for details) (grey), and an Erdös–Rényi random graph. Averages were taken over 1000 independent simulated networks (black circles, scale-free RGG; crosses, scale-free BC; triangles, ER, random graph).

simulations studies, however, is that the GCC is remarkably stable against noise (see figure 1). Only when both the true-positive rate, ρ, and the false-positive rate, ξ, are sufficiently small, will the GCC vanish. It follows straightforwardly from equation (2.5) that the expected number of false-positive edges in a network—we always assume independence among edges—is given by

$$E_{\xi}[M^*] = \xi\left(\binom{N}{2} - M\right). \tag{4.1}$$

As most real networks are sparse, $\binom{N}{2}$ dominates M, and therefore for moderate values of $E_{\xi}[M^*]$ will generally be sufficiently large to ensure that a GCC exists.

5. ERROR BOUNDS FOR BIOLOGICAL DATA

Protein interaction network (PIN) data have been rightly chastised for the high error rates in the experimental assays. But the very fact that a GCC is absent can be used in order to put an upper bound on the false-positive rate, ξ, or, equivalently, the average number of false-positive edges per node, Ξ. Only two high-throughput surveys (we considered all yeast and human PIN data sets deposited in the IntAct database, Kerrien *et al.* 2007) show no evidence for a GCC; some of the basic properties of these networks are summarized in table 1. Using the expressions for z_1^{DS} and z_2^{DS}, equations (3.15) and (3.16), respectively, we obtain an inequality for the existence of a GCC which depends on Ξ, ρ and p.

Table 1. Details of two protein–protein interaction networks. N_{GCC} is the number of nodes which belong to the largest component in the data set. We have used recent estimates (Hart *et al.* 2006; Stumpf *et al.* 2008) of approximately 40 000 interactions among the 6000 or so protein-coding yeast genes in order to estimate $\hat{p}$ and $\langle\hat{k}\rangle \approx 6.7$; for humans, the average degree is expected to be approximately $\langle\hat{k}\rangle \approx 25$. Only the data sets of Ho *et al.* (2002) and Gavin *et al.* (2002) show no evidence for a GCC.

experiment	N_{S}	M_{S}	N_{GCC}	$\hat{p}$
Uetz *et al.* (2000)	1328	1438	921	0.221
Ho *et al.* (2002)	871	694	95	0.145
Gavin *et al.* (2002)	726	367	20	0.121
Ito *et al.* (2002)	3245	4449	2808	0.541
Gavin *et al.* (2006)	1432	6532	1359	0.238
Krogan *et al.* (2006)	2676	7076	2527	0.446
Rual *et al.* (2005)	1527	2671	1286	0.069
Stelzl *et al.* (2005)	1665	3119	1568	0.076

For real networks, however, we do not know the moments of the degree distribution, $\langle k\rangle$ and $\langle k^2\rangle$. For yeast, a range of studies has estimated the size of the true PIN and with the estimated number of edges, $\hat{M}$, we may in turn estimate $\langle\hat{k}\rangle = \hat{M}/N$, which still leaves us with $\langle k^2\rangle$ unknown. If we further set $\hat{p} = N_S/N$ we can write as:

$$\hat{p}\left(\rho\langle\hat{k}\rangle + \Xi\right) - \hat{p}^2\left(\rho^2\langle k^2\rangle - \rho^2\langle\hat{k}\rangle + 2\Xi\rho\langle\hat{k}\rangle + \Xi^2\right) \le 0 \tag{5.1}$$

to identify the region for which a GCC should exist according to the MR criterion.

These regions are shown in figure 3 for the only two published data sets that do not exhibit a GCC (according to the usual definitions). Depending on the second moment of the degree distribution, $\langle k^2\rangle$, the absence of a GCC indicates fairly low true- and false-positive rates. From the results in figure 4, we conclude that for those PIN data sets for which we do not observe a GCC, the fraction of false positive edges must be below 50 per cent, a number that has sometimes been suggested (von Mering *et al.* 2002; Bader *et al.* 2004).

6. CONCLUSIONS

Networks offer a convenient and powerful framework for the analysis of structured and complex processes and data. But network data, especially in biology, are also subject to considerable uncertainty and highly incomplete. Here we have studied the effects of incompleteness and noise on the existence of a GCC in theoretical network ensembles and experimental network data. We have introduced the notion of NINEs as a suitable conceptual framework to deal with observational vagaries of experimental network data sets. For uncorrelated networks we have derived the criteria which determine the existence of a GCC (with high probability). We have furthermore performed some simulations for finite-size network ensembles which show that the theoretical results describe the average behaviour of realistic networks.

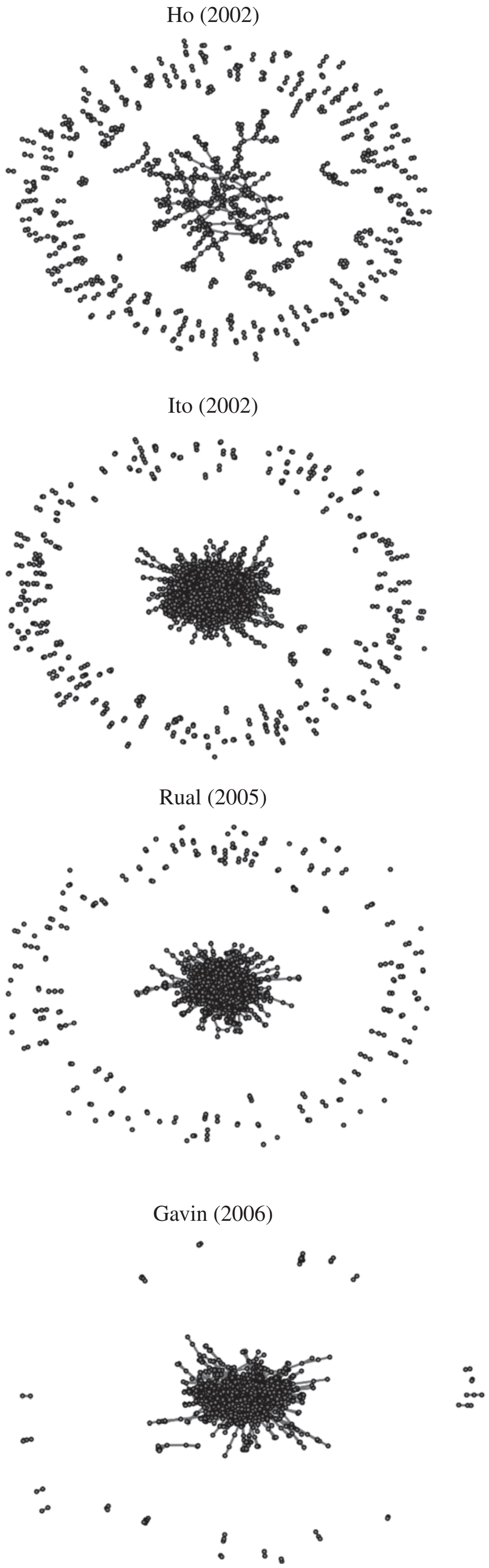

Figure 3. Representations of eight networks in the IntAct database. The two early data sets generated by mass spectrometry (Ho *et al.* 2002; Gavin *et al.* 2002) do not show evidence for a GCC, as may well be expected given that they focus on protein complexes. The details of the data sets are provided in table 1.

We have seen that the GCC's persistence under sampling but especially under the effects of noise depends strongly on the second moment of the degree distribution (by virtue of the MR criterion). If networks are scale-free with exponent $\gamma \leq 3$, then the divergence of $\langle k^2 \rangle$ means that a GCC will persist for all levels of noise. For real data sets that are finite—and almost certainly not scale-free in the classical sense (Przulj

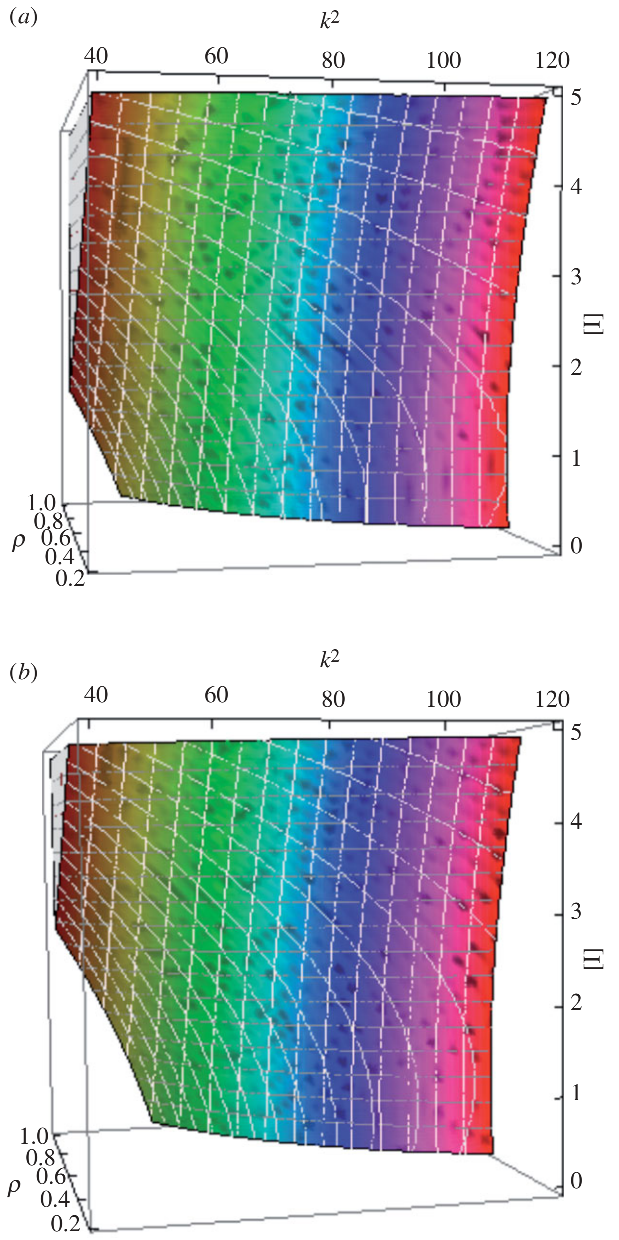

Figure 4. The coloured regions indicate combinations in ($\langle k^2 \rangle$, ρ, Ξ)-space for which we will observe a GCC. (*a*) The corresponds to the case of $\hat{p}$ in Ito *et al.* (2001), (*b*) to the case of $\hat{p}$ in Gavin *et al.* (2002). Perhaps somewhat unexpectedly we also observe a pronounced effect of $\hat{p}$ ($\xi \approx 14.5\%$ in (*a*) and $\approx 12.1\%$ in (*b*)) on the boundary of the parameter space allowing for a GCC with high probability.

et al. 2004; Stumpf *et al.* 2005*a*; Tanaka *et al.* 2005; Khanin & Wit 2006)—we can, however, use the emergence of a GCC as (weakly) indicative of the relative effects of false-positive to true-positive error rates.

We have illustrated this in the context of published high-throughput protein interaction data in yeast and humans. Only two data sets fail to exhibit a GCC. The data of Gavin *et al.* (2002) and Ho *et al.* (2002) focused on interaction in protein complexes and therefore we would not necessarily expect a GCC to be visible in the data. Previous analyses of error rates in protein interaction data have sometimes focused on overlap between different sets of data (von Mering *et al.* 2002; Hart *et al.* 2006), which we expect to systematically overestimate error rates in such data. However, we are aware that our analysis relies on errors being distributed evenly in the network. This assumption is naturally debatable and further theoretical developments might be called for.

Fruitful applications of the present analysis could arise when reversing our approach: how much can we learn from incomplete and noisy data about the true (but unobserved) network? The notion of NINEs introduced above allows us to address this question rigorously. For any realistic complex system, we may start from the assumption that observed data were sampled from a connected graph. There have been only preliminary attempts (e.g. Stumpf *et al.* (2008)) at reverse-engineering the global topology of networks from partial (often poor) data. Especially, in the social sciences Robins & Pattison (2001) where network sampling is fraught with all manner of methodological, social and ethical problems, and in systems biology, where technological problems frequently prevent us from seeing e.g. protein interactions that are conditional on post-translational modifications, we see the need and much scope for suitable inferential procedures.

The authors thank Sophie Lèbre and Thomas Thorne for helpful discussions. MPHS is a Royal Society Wolfson Research Merit Award holder and acknowledges support from the BBSRC; CW acknowledges research support from the Danish Research Councils.

REFERENCES

Arfken, G. & Weber, H. 2005 *Mathematical methods for physicists*, 6th edn. New York, NY: Academic Press.

Bader, J. S., Chaudhuri, A., Rothberg, J. M. & Chant, J. 2004 Gaining confidence in high-throughput protein interaction networks. *Nat. Biotechnol.* **22**, 78–85. (doi:10.1038/nbt924)

Barabasi, A. & Albert, R. 1999 Emergence of scaling in random networks. *Science* **286**, 509–512. (doi:10.1126/science.286.5439.509)

Barabasi, A., Ravasz, E. & Vicsek, T. 2001 Deterministic scale-free networks. *Physica A* **299**, 559–564. (doi:10.1016/S0378-4371(01)00369-7)

Bender, E. & Canfield, E. 1978 The asymptotic number of labeled graphs with given degree sequence. *J. Combin. Theory A* **24**, 296–307. (doi:10.1016/0097-3165(78)90059-6)

Berg, J. & Lässig, M. 2002 Correlated random networks. *Phys. Rev. Lett.* **89**, 228701. (doi:10.1103/PhysRevLett.89.228701)

Bollobás, B. 2001 *Random graphs.* New York, NY: Cambridge University Press.

Bollobás, B. & Riordan, O. 2006 *Percolation.* Cambridge, UK: Cambridge University Press.

Bork, P., Jensen, L. J., von Mering, C., Ramani, A. K., Lee, I. & Marcotte, E. M. 2004 Protein interaction networks from yeast to human. *Curr. Opin. Struct. Biol.* **14**, 292–299. (doi:10.1016/j.sbi.2004.05.003)

Burda, Z. & Krzywicki, A. 2004 Uncorrelated random networks. *Phys. Rev. E* **67**, 046118. (doi:10.1103/PhysRevE.67.046118)

Burda, Z., Correia, J. D. & Krzywicki, A. 2001 Statistical ensemble of scale-free random graphs. *Phys. Rev. E* **64**, 046118. (doi:10.1103/PhysRevE.64.046118)

Callaway, D., Newman, M., Strogatz, S. & Watts, D. 2000 Network robustness and fragility: Percolation on random graphs. *Phys. Rev. Lett.* **85**, 5468–5471. (doi:10.1103/PhysRevLett.85.5468)

Chung, F. & Lu, L. 2006 *Complex graphs and networks.* Providence, RI: American Mathematical Society.

de Silva, E., Thorne, T., Ingram, P., Agrafioti, I., Swire, J., Winf, C., Stumpf, M. PH. & Agrafioti, I. 2006 The effects of incomplete protein interaction data on structural and evolutionary inferences. *BMC Biol.* **4**, 39. (doi:10.1186/1741-7007-4-39)

Dorogovtsev, S. & Mendes, J. 2003 *Evolution of networks.* Oxford, UK: Oxford University Press.

Durrett, R. 2006 *Random graph dynamics.* Cambridge, UK: Cambridge University Press.

Erdös, P. & Rényi, A. 1959 On random graphs I. *Publ. Mat. Debr.* **5**, 290–297.

Erdös, P. & Rényi, A. 1960 On the evolution of random graphs. *Magyar Tud. Akad. Mat. Kutató Int. Közl.* **5**, 17–61.

Gavin, M., Bosche, M., Krause, R., Grandi, P. *et al.* 2002 Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* **415**, 141–147. (doi:10.1038/415141a)

Gavin, A. *et al.* 2006 Proteome survey reveals modularity of the yeast cell machinery. *Nature* **440**, 631–636. (doi:10.1038/nature04532)

Gradshteyn, I., Ryzhik, I. & Jeffrey, A. 1994 *Table of integrals, series and products.* New York, NY: Academic Press.

Hakimi, S. 1962 On realizability of a set of integers as degrees of vertices of a linear graph I. *J. Soc. Ind. Appl. Math.* **10**, 496–506.

Han, J., Dupuy, D., Bertin, N., Cusick, M. & Vidal, M. 2005 Effect of sampling on topology predictions of protein–protein interaction networks. *Nat. Biotechnol.* **23**, 839–844. (doi:10.1038/nbt1116)

Hart, G., Ramani, A. & Marcotte, E. 2006 How complete are current yeast and human protein-interaction networks? *Genome Biol.* **7**, 120. (doi:10.1186/gb-2006-7-11-120)

Ho, Y. *et al.* 2002 Systematic identification of protein complexes in saccharomyces cerevisiae by mass spectrometry. *Nature* **415**, 180–183. (doi:10.1038/415180a)

Ito, T., Chiba, T., Ozawa, R., Yoshida, M., Hattori, M. & Sakaki, Y. 2001 A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc. Natl Acad. Sci. USA* **98**, 4569–4574. (doi:10.1073/pnas.061034498)

Ito, T., Ota, K., Kubota, H., Yamaguchi, Y., Chiba, T., Sakuraba, K. & Yoshida, M. 2002 Roles for the two-hybrid system in exploration of the yeast protein interactome. *Mol. Cell. Proteom.* **1**, 561–566. (doi:10.1074/mcp.R200005-MCP200)

Kerrien, S. *et al.* 2007 Intact–open source resource for molecular interaction data. *Nucleic Acids Res.* **35**, D561–D565. (doi:10.1093/nar/gkl958)

Khanin, R. & Wit, E. 2006 How scale-free are biological networks? *J. Comp. Biol.* **13**, 810–818. (doi:10.1089/cmb.2006.13.810)

Krogan, N. J. *et al.* 2006 Global landscape of protein complexes in the yeast *saccharomyces cerevisiae Nature* **440**, 637–643. (doi:10.1038/nature04670)

Lee, S., Kim, P. & Jeong, H. 2006 Statistical properties of sampled networks. *Phys. Rev. E* **73**, 016102. (doi:10.1103/PhysRevE.73.016102)

Middendorf, M., Etay, Z. & Wiggins, C. 2004 Inferring network mechanisms: The drosophila melanogaster protein interaction network. *Proc. Natl Acad. Sci. USA* **102**, 3192–3197. (doi:10.1073/pnas.0409515102)

Molloy, M. & Reed, B. 1995 A critical point for random graphs with a given degree distribution. *Random struct. Algorithms* **6**, 161–179.

Molloy, M. & Reed, B. 1998 The size of the giant component of a random graph with a given degree sequence. *Combin. Probab. Comput.* **7**, 295–305. (doi:10.1017/S0963548398003526)

Newman, M., Strogatz, S. & Watts, D. 2001 Random graphs with arbitrary degree distributions and their applications. *Phys. Rev. E* **64**, 026118. (doi:10.1103/PhysRevE.64.026118)

Przulj, N., Corneil, D. G. & Jurisica, I. 2004 Modeling interactome: scale-free or geometric? *Bioinformatics* (doi:10.1093/bioinformatics/bth436)

Reguly, T. *et al.* 2006 Comprehensive curation and analysis of global interaction networks in *saccharomyces cerevisiae. J. Biol.* **5**, 11. (doi:10.1186/jbiol36)

Robins, G. & Pattison, P. 2001 Random graph models for temporal processes in social networks. *J. Math. Soc.* **25**, 4–21.

Rual, J. *et al.* 2005 Towards a proteome-scale map of the human protein–protein interaction network. *Nature* **437**, 1173–1178. (doi:10.1038/nature04209)

Stauffer, D. & Aharony, A. 1992 *Introduction to percolation theory*, 2nd edn. London, UK: Taylor & Francis.

Stelzl, U. *et al.* 2005 A human protein–protein interaction network: a resource for annotating the proteome. *Cell* **122**, 957–968. (doi:10.1016/j.cell.2005.08.029)

Stumpf, M. & Wiuf, C. 2005 Sampling properties of random graphs: the degree distribution. *Phys. Rev. E* **72**, 036118. (doi:10.1103/PhysRevE.72.036118)

Stumpf, M., Ingram, P., Nouvel, I. & Wiuf, C. 2005*a* Statistical model selection methods applied to biological networks. *Trans. Comput. Syst. Biol.* **3**, 65–72 (doi:oai:arXiv.org:q-bio/0506013).

Stumpf, M., Wiuf, C. & May, R. 2005*b* Subnets of scale-free networks are not scale-free: the sampling properties of networks. *Proc. Natl Acad. Sci. USA* **102**, 4221–4224. (doi:10.1073/pnas.0501179102)

Stumpf, M., Thorne, T., de Silva, E., Stewart, R. *et al.* 2008 Estimating the size of the human interactome. *Proc. Natl Acad. Sci.* **105**, 6959. (doi:10.1073/pnas.0708078105)

Tanaka, R., Ti, T. & Doyle, J. 2005 Some protein interaction data do not exhibit power law statistics. *FEBS Lett.* **579**, 5140–5144. (doi:10.1016/j.febslet.2005.08.024)

Uetz, P., Giot, L., Cagney, G., Mansfield, T. *et al.* 2000 A comprehensive analysis of protein-protein interaction networks in *saccharomyces cerevisiae. Nature* **403**, 623–627. (doi:10.1038/35001009)

von Mering, C., Krause, R., Snel, B., Cornell, M. *et al.* 2002 Comparative assessment of large-scale data sets of protein–protein interactions. *Nature* **417**, 399–403. (doi:10.1038/nature750)

Wiuf, C. & Stumpf, M. 2006 Binomial sampling. *Proc. R. Soc. A* **462**, 1181–1195. (doi:10.1098/rspa.2005.1622)

Yook, S., Oltvai, Z. & Barabasi, A. 2004 Functional and topological characterization of protein interaction networks. *Proteomics* **4**, 928–942. (doi:10.1002/pmic.200300636)