Multi-model inference of network properties from incomplete data

Michael P.H. Stumpf[1,2], Thomas Thorne[1]

[1]Centre for Bioinformatics, Division of Molecular Biosciences, Imperial College London, SW7 2AZ London, UK

[2]Instiute of Mathematical Sciences, Imperial College London, SW7 2AZ London, UK
http://www.imperial.ac.uk/theoreticalgenomics

Summary

It has previously been shown that subnets differ from global networks from which they are sampled for all but a very limited number of theoretical network models. These differences are of qualitative as well as quantitative nature, and the properties of subnets may be very different from the corresponding properties in the true, unobserved network. Here we propose a novel approach which allows us to infer aspects of the true network from incomplete network data in a multi-model inference framework. We develop the basic theoretical framework, including procedures for assessing confidence intervals of our estimates and evaluate the performance of this approach in simulation studies and against subnets drawn from the presently available PIN network data in *Saccaromyces cerevisiae*. We then illustrate the potential power of this new approach by estimating the number of interactions that will be detectable with present experimental approaches in sfour eukaryotic species, inlcuding humans. Encouragingly, where independent datasets are available we obtain consistent estimates from different partial protein interaction networks. We conclude with a discussion of the scope of this approaches and areas for further research.

1 Introduction

Molecular networks are widely seen to provide concise and coherent descriptions of the cellular machinery in living systems. While it is generally accepted, and indeed well established[16, 13], that in their current guise they offer only approximate representations of the complex processes in cells, tissues or organisms, their analysis has received great attention which has already produced some tangible results about the organization of molecular phenotypes and biological processes at the system-level. At the moment three types of molecular networks are currently being distinguished: metabolic networks (MN), gene regulatory networks (GRN) and protein interaction networks (PIN). This is, in many respects, a useful first-order distinction but ultimately it has to be kept in mind, that these networks do, in fact, interact with each other and are intricately linked.

The integrative analysis of biological networks poses considerable statistical challenges: for example, GRNs can only be inferred (at least in a high-throughput sense) indirectly from experimental data. Because the number of experiments is much less than the number of genes (typically by at least one to two orders of magnitude) statistical measures such as correlations are not well defined and sophisticated variance reduction approaches have to be employed in order to obtain results that can be treated as reliable (see Schäfer & Strimmer [25] for an excellent brief review).

Here we are concerned with developing a novel statistical approach that allows us to infer properties of the global network $\mathcal{N}$ from some subnetwork $\mathcal{S}$. We assume that $\mathcal{S}$ is generated by picking a subset of the nodes in the true network and considering only the interactions among them (see below for details). It has previously been shown [31, 30, 16, 35, 19] that the properties of subnets can differ quite considerably from those of the true network. Such differences will, in fact, be of a qualitative nature [31, 35] for most types of networks, even when sampling of nodes is essentially uniform and random.

We would, of course, not expect that the structure of real networks is adequately described by any of the theoretical network models that have been developed in the literature [14, 34]. Burda *et al.* [9] have shown, however, that the notion of network ensembles can nevertheless provide a useful tool for the statistcial analysis of real network data. But even the network ensembles may not reasonably be expected to provide adequate descriptions of real biological networks such as PINs. Because of this model-selection and multi-model inference (MMI) provide a natural framework for the analysis of complex networks: model selection determines the information that different (non-nested) models can capture about the available data, and MMI averages the statistical estimates of predictions obtained from different models, weighted by their relative explanatory powers. The strength of MMI lies in the fact that even if the true model is not among the set of models it is posible to obtain reliable statistical estimates or predictions for parameters describing a stochastic process/system.

Below we will give a brief introduction into model-selection and MMI. For the reasons outlined there we would expect that MMI offers a potential route towards predicting properties of the overall network, $\mathcal{N}$, from the properties of subnet, $\mathcal{S}$. This is particularly true for uncorrelated networks[18, 10] but can also be extended to correlated networks [4]. After introducing the new MMI approach we will evaluate its performance using simulated network and *S.cerevisiae* PIN data, before applying it to estimate the degree distributions of the global *S.cerevisiae* PIN as well as the sizes of PINs in a range of important model organisms. The manuscript concludes with a discussions of the scope and limitations of this approach in the analysis of complex biological networks.

2 Methods

We consider a complete network $\mathcal{N} = (\mathcal{V}, \mathcal{E})$ where $\mathcal{V}$ is the set of N nodes in the network (*i.e.* $N = |\mathcal{V}|$) and $\mathcal{E}$ is the set of M edges e_{ij} with $v_i, v_j \in \mathcal{V}$. In the simplest case we consider simple or Meier graphs[7], where self-interactions e_{ij} with $i = j$ and multiple edges between pairs of nodes cannot occur (this poses no serious constraint and can be lifted). We denote the number of nodes with degree k (*i.e.* nodes which have k interaction partners) in the network by $n(k)$.

We will also consider subnets of $\mathcal{N}$ denoted by $\mathcal{S} = (\mathcal{V}_\mathcal{S}, \mathcal{E}_\mathcal{S})$ where $\mathcal{V}_\mathcal{S} \subset \mathcal{V}$ and $\mathcal{E}_\mathcal{S} \subset \mathcal{E}$ with $e_{ij} \in \mathcal{E}_\mathcal{S}$ when $e_{ij} \in \mathcal{E}$ and $v_i, v_j \in \mathcal{V}_\mathcal{S}$. Furthermore we use $N_\mathcal{S} = |V_\mathcal{S}|$ and $M_\mathcal{S} = |E_\mathcal{S}|$ to denote the number of nodes and edges in the subnet, respectively and ρ to denote the ratio $\rho = N_\mathcal{S}/N$.

2.1 The degree distribution and uncorrelated networks

The degree distribution $\Pr(k)$ defines the probability of observing a node in the network with degree k. For a given network we can define the *empirical degree distribution*

$$\Pr_e(k) = \frac{n(k)}{N}. \tag{1}$$

Below we will refer to $\Pr_e(k)$ as the *degree sequence* of a network.

An *uncorrelated random network* is one where the probability of hitting a node with degree k when traveling along a random edge depends only on the degree distribution and k. This probability is then given by

$$\frac{k\Pr(k)}{\langle k \rangle}. \tag{2}$$

In uncorrelated networks, as far as network structure is concerned, the degree sequence is a sufficient statistic for random graphs; *i.e.* the likelihood of any other network statistic (such as *clustering coefficient* or so-called *motif spectra* see [13] for definitions) depends only on the degree sequence and not on any further aspect of the data. This follows straightforwardly from the definition of uncorrelated networks[10]. Thus, to the extent that we are dealing with (approximately) uncorrelated, degree sequences or distributions are in fact very efficient summaries of the network data. Moreover, while it is possible to perform full likelihood analyses of models of network evolution[34], composite likelihood approaches which use the degree sequence require less run-time and are easily implemented[28, 29].

2.2 Model-Selection and Multi-Model Inference

We assume that we have a set of R probability models, $Q = \{Q_1, Q_2, \ldots, Q_R\}$, each of which is parameterized by a (potentially vector-valued) parameter θ_i. Formal model selection approaches allow us to compare the explanatory power (in an information theoretic sense) given the observed data, $\mathbf{D} = \{d_1, d_2, \ldots, d_n\}$, of these models (even if they are not nested and the standard likelihood ratio test is not applicable). Model-selection and multi-model inference are briefly reviewed here and discussed in detail in Burnham&Anderson [11]. An application to phylogenetic inferences is given in Strimmer and Rambaut [27].

2.2.1 Maximum Likelihood Inference

For each model we calculate the likelihood [12]

$$L(\theta_i) = \Pr(D|\theta_i) = \prod_{j=1}^{n} \Pr(d_j|\theta_i) \tag{3}$$

or, equivalently, the log-likelihood

$$\mathrm{lk}(\theta_i) = \log(\Pr(D|\theta_i) = \sum_{j=1}^{n} \log(\Pr(d_j|\theta_i)). \tag{4}$$

in terms of some dataset $D = \{d_1, d_2, \ldots, d_t\}$. The maximum likelihood estimates of the parameters θ_i which maximize the respective likelihood/log-likelihood are denoted by $\hat{\theta}_i$. The corresponding (maximal) value of the log-likelihood is denoted by lk_i.

2.2.2 Model Selection

Information criteria such as those due to Akaike or Schwartz assess the amount of information about the data that is apparently contained in the fitted model[1, 11]. The Akaike information criterion (AIC) chooses the model that minimizes the so-called Kullback-Leibler (K-L) distance [1]; it is defined by

$$\mathrm{AIC}_i = -2\mathrm{lk}_i + 2\nu_i, \tag{5}$$

where ν_i is the number of parameters describing probability model Q_i (*i.e.* $\nu_i = \dim(\theta_i)$). Thus the AIC explicitly biases against more complicated models (*i.e.* those with more parameters) unless they capture more information about the data, as measured by the log-likelihood.

In order to compare the different models we denote the AICof the model with the minimal AIC by $\mathrm{AIC}_{\min}$. For each model Q_i the difference between its AIC_i and the minimum

$$\Delta_i = \mathrm{AIC}_i - \mathrm{AIC}_{\min} \tag{6}$$

is then used to determine the *relative likelihood* of model i, Q_i, given all the other models. This is given by

$$\exp\left(\frac{-\Delta_i}{2}\right), \tag{7}$$

and accounts for the different numbers of parameters the models may contain. These relative likelihoods then, in turn, define the Akaike weights, ω_i, for each model, M_i,

$$\omega_i = \frac{\exp\left(\frac{-\Delta_i}{2}\right)}{\sum_{j=1}^{R} \exp\left(\frac{-\Delta_j}{2}\right)}. \tag{8}$$

The correct interpretation for the ω_i uses them as relative weights or probabilities, for a model to be true given all the other models in a set of models. Note that when the set $\mathbf{Q}$ is augmented we have to update the Akaike weights of all other models as well. Akaike weights refer only to a given configuration of probability models. In previous studies we have shown, that the Akaike model selection formalism always chooses a single model as the clearly best model (generally the stretched exponential or log-normal model; never, however, the pure scale-free model [28, 29]).

2.2.3 Multi-model inference

It is important to keep in mind, that complex biological data will have been generated by a process much more involved and intricate than any of the models we can typically study. Good theoretical models are — necessarily and by design — oversimplified representations of real processes. Thus the "real" model is not among our set of models, $\mathbf{Q}$. Model selection may help us to identify which model explains the data best in an information theoretic sense, and MMI will draw inferences from a weighted average over predictions made under the different candidate models. Interestingly, if the true model is among the candidate models then the MMI estimator will have smaller variance than the corresponding estimator for the true model.

[1]The K-L distance is not a distance in the mathematical sense. In particular the triangular inequality is not fulfilled for the K-L distance. The term "distance" has been retained in conventional use.

Equally, estimates and predictions have been shown to have very good statistical properties when the true model is not among the candidate models [27].

When one of the Akaike weigths $\omega_i \gtrsim 0.9$ one often restricts inferences to this model (we need to remember, however, that it is not the correct model for the type of data we will generally be looking at). But when none of the models has a very high Akaike weight, we can average those parameters that are shared by different models, or predictors derived from the models in a straightforward manner. Let ϵ be the predicted value of some property of a system described by the models (below we will use the size of a network that is to be "predicted" from incomplete network data). Let $\hat{\epsilon}_i$ be the predicted value from model i. Then the model-averaged predictor, $\hat{\bar{\epsilon}}$, is given by

$$\hat{\bar{\epsilon}} = \sum_{i=1}^{R} \omega_i \hat{\epsilon}_i \tag{9}$$

Thus the estimate/prediction for ϵ resulting from each model is weighted by the model's Akaike weight. The variability of the estimator/predictor given by Eqn. (9) can be estimated by bootstrapping the data to obtain bootstrap values for $\omega_i$and $\hat{\epsilon}_i$.

2.3 Incomplete network data

We have previously studied the effects of unbiased and biased sampling from networks [28, 29] and generalize some of these results here. If the degree distribution of the network is given by $\mathrm{Pr}_{\mathcal{N}}$, and the probability for sampling a node is p, then the degree distribution of the resulting random subnet resulting for the subgraph induced by the sampled nodes is given by

$$\mathrm{Pr}_{\mathcal{S}}(k) = \sum_{l \geq k} \binom{l}{k} p^k (1-p)^{l-k} \mathrm{Pr}_{\mathcal{N}}(l). \tag{10}$$

With p the probability of sampling a node we can straightforwardly determine the probability of sampling an edge as

$$\hat{\pi} = p^2. \tag{11}$$

This can be straightforwardly generalized to non-random sampling. Assume that protein i is chosen with probability $p(i)$ with $0 \leq p \leq 1$ for all nodes $i \in \mathcal{V}$. In well defined limits we can make a mean-field approximation where

$$p(i) = \langle p \rangle. \tag{12}$$

and

$$\langle p \rangle = \frac{1}{N} \sum_{i=1}^{N} p(i) \tag{13}$$

is the average probability of picking a node (note that we can naturally also calculate higher moments to obtain corrections around the mean-field limit; this way we can, for example, recover behaviour such as the percolation transition as $\langle p \rangle$ decreases). Obviously we also have

$$\langle p \rangle = \rho \tag{14}$$

Given $p(i)$ and $p(j)$ the probability of including e_{ij} is given by

$$\pi_{ij} = p(i)p(j) \tag{15}$$

in agreement with Eqn. (11).Strictly speaking edges are not inherited independently but it is possible to show that given $M_{\mathcal{S}}$ and ρ an unbiased estimator for the number of edges in the full network, $\hat{M}$ is given by

$$\hat{M} = \frac{M_{\mathcal{S}}}{\rho^2}. \tag{16}$$

The properties of the estimator given by Eqn. (16) have been studied extensively [32] and have been shown to be remarkably accurate for simulated and real-datasets. Moreover, independent datasets for the same species result in similar estimates for M. Below we will derive estimators of properties of $\mathcal{N}$ from the subnet $\mathcal{S}$ using multi-model inference; in particular we will compare these with predictions based on Eqn. (16).

This approach also allows us to study other properties of incomplete network data. For example, we may calculate the expected overlap between different datasets: the fraction of edges shared among two subnets S_1 and S_2 is simply given by $\rho_1^2\rho_2^2$ where ρ_i is given by Eqn. (14) for subnet S_i. We can also superimpose noise (*i.e.* false positive and false negative interactions) onto the sampling process.

2.4 Inference of network properties from partial network data

By combining MMI with the results obtained above relating complete to partial network data, we can make predictions about structural properties of the complete but unobserved network. We can combine Eqns. (10) and (4) to obtain the log-likelihood for a model that describes the degree distribution of the true network and obtain

$$\mathrm{lk}(\theta_{i;\mathcal{N}}) \propto \sum_{k=0}^{k_{\max}} \log\left(\sum_{l\geq k}\binom{l}{k}p^k(1-p)^{l-k}\mathrm{Pr}_{\mathcal{N}}(l|\theta_i).\right) \times n(k) \tag{17}$$

where n_k is the number of nodes in the subnet with degree k and $k_{\max}$ is the maximum degree in the subnet. Note that only for very special choices of $\mathrm{Pr}_{\mathcal{N}}(l|\theta_i)$ will it be possible to evalue the inner sum in Eqn. (17) in closed form; generally we have to sum it up numerically.

When calculating the Akaike weigths it is important to keep in mind that the sampling fraction is an additional parameter and ν_i is incremeneted by one for all models. To apply Eqn. (17) we replace p by the estimate

$$\hat{p} = \frac{N_{\mathcal{S}}}{N} = \rho. \tag{18}$$

For each model we thus obtain an estimate for the parameters θ_i that describe the degree distribution of the true network (rather than the subnet on which all previous efforts have concentrated). The model averaged estimate for the degree distribution $\hat{\bar{f}}(k)$ is thus given by

$$\hat{\bar{f}}(k) = \sum_{i=1}^{R} \omega_i \mathrm{Pr}(k|Q_i(\hat{\theta}_i)). \tag{19}$$

Probability model	**Degree distribution Pr**$(k;\theta)$	
Poisson	$\exp(-\lambda)\frac{\lambda^k}{k!}$	for all $k \geq 0$
Exponential	$C\exp(-\lambda k)$	for all $k \geq 0$
Scale-free	0	for $k = 0$
	$k^{-\gamma}/\zeta(\gamma)$	for $k > 0$
Lognormal	$C\frac{e^{-\ln((k-\beta)/m)^2/(2\sigma^2)}}{(k-\beta)\sigma\sqrt{2\pi}}$	for all $k \geq 0$
Stretched exponential	0	for $k < 0$
	$C\frac{\alpha}{m}\left(\frac{k}{m}\right)^{\alpha-1}\exp(-(k/m)^{\alpha})k^{-\gamma}$	for $k > 0$

Table 1: The five network models and their degree distributions, Pr$(k;\theta)$ used in the analysis, below. Wherever it appears, C denotes the normalizing constant such that $\sum_k$ Pr$(k;\theta) = 1$. In [28, 29] we analyzed the effects of other, such as trunacted scale-free models, but these were still found to be inferior to the lognormal and stretched exponential distributions.

$\hat{\bar{f}}(k)$ is, of course, a degree distribution and we have, *e.g.*

$$\sum_{k=0}^{\infty} \hat{\bar{f}}(k) = 1. \tag{20}$$

An approximate assessment of the variation in the estimator (19) can be obtained by bootstrapping the nodes directly, to obtain a bootstrap replicate for the empirical degree distribution in the subnet $\text{Pr}^*_{\mathcal{S}}(k)$. We thus get a standard error for $\hat{\bar{f}}(k)$ for each value of k. In the present treatment these are approximately independent. For sufficiently large networks (those with several hundred or more nodes) this does not appear to matter very much.

One quantity we are particularly interested in is the size of the interactome. The number of edges in a network with N nodes and degree sequence Pr$(k)$is given by

$$M = \frac{N}{2}\sum_{k=0}^{k_{\max}} k \times \text{Pr}(k). \tag{21}$$

Here we will estimate the total number of edges in the whole network, $\hat{M}$, directly from Eqn. (19),

$$\hat{M}_{\mathcal{N}} = \frac{N_{\mathcal{N}}}{2}\sum_{k=0}^{k_{\max}} k \times \hat{\bar{f}}(k). \tag{22}$$

This can be compared with the non-parametric estimate

$$\hat{M}'_{\mathcal{N}} = \frac{M_{\mathcal{S}}}{\hat{p}^2} \tag{23}$$

The estimator (23) does not depend on the degree distribution in the true network and is a good estimator for data-sets that where generated by a systematic, un-biased experimental process.[2]

[2]Bias of a sample refers to whether or not the subnet induced by the sampled nodes has systematically different properties from subnets that are induced by randomly selected nodes. In this respect "bias" for biological networks is only likely to result for extremely severe ascertainment bias. This may for example occur if baits and preys are chosen for proteins which are *a priori* known to be involved in the same process or the same cellular compartment. For large data-sets or high-throughput experiments this is extremely unlikely.

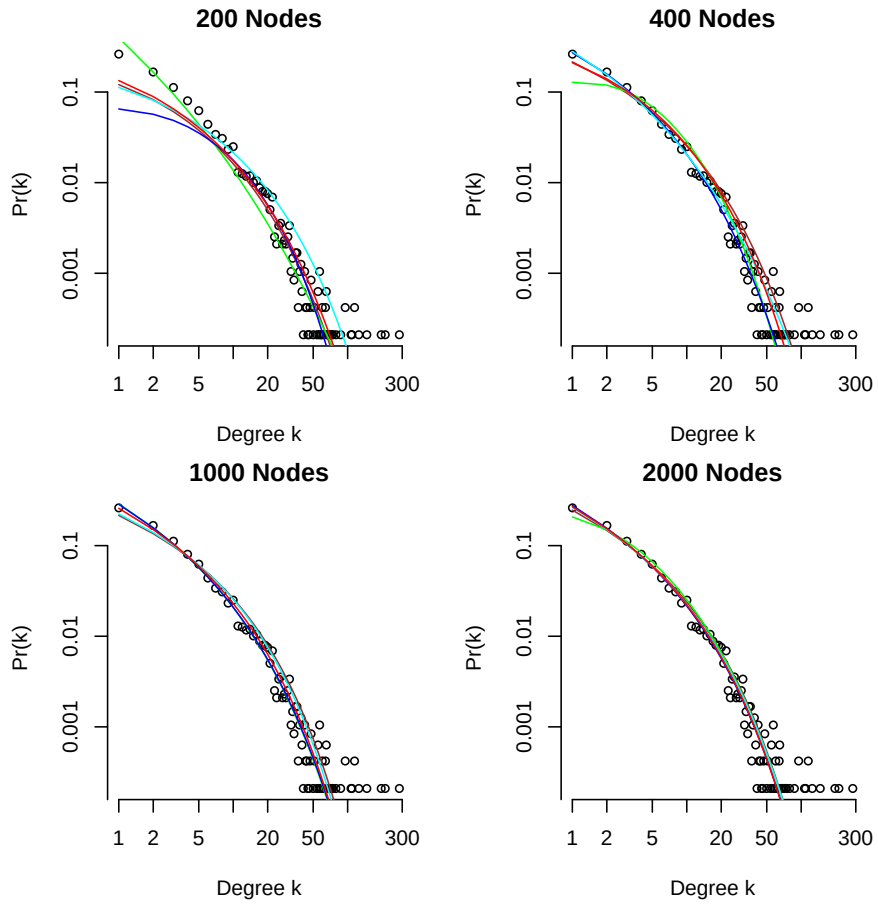

Figure 1: Empirical degree distribution of present yeast PIN dataset (black circles) and MMI estimates for the degree distribution resulting from five independent (shown in black, blue, red, green, cyan) subnets generated by sampling sets of 200, 400, 1000 and 2000 nodes, respectively (continuous curves).

The probability models $Q_i(\theta_i)$ can be chosen relatively freely: we are not *per se* interested in which model describes the data best but in predicting properties of $\mathcal{N}$ and each model may give some insight. In the particular instance of biological networks, however, it makes sense to limit the models for the degree sequence to those which have slowly decaying ("fat") tails and in addition to the power-law distribution we will use stretched exponentials and log-normal distribution. It is also possible to consider mechanistic models such as the popular *duplication-attachement* models[34]. In general, model selection and MMI work best, however, for a moderate number of well suited models tends to perform best.

Here we consider networks which are uncorrelated; given present data-sets this is justified. This condition can be relaxed and it is possible to incorporate correlated network ensembles should that be of importance.

3 Assessing the power of MMI from incomplete networks

In order to assess the power of MMI in the context of protein interaction networks we have assessed the power of this approach in three ways:

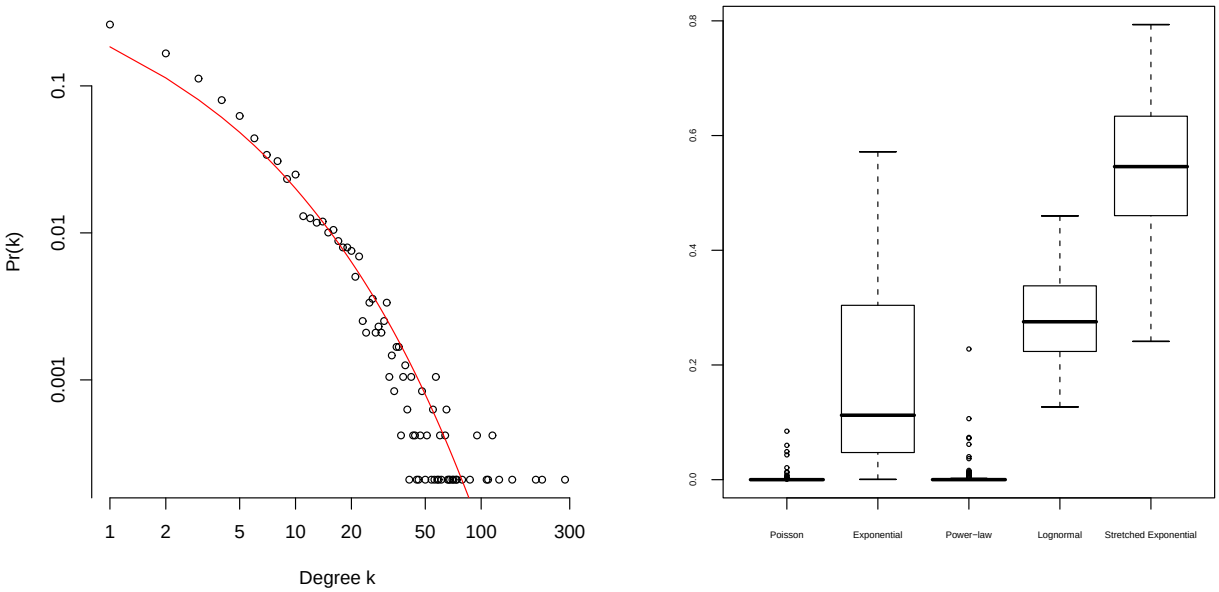

Figure 2: Empirical degree distribution of present yeast PIN dataset (black circles) and bagged MMI estimate for the degree distribution resulting from 200 bootstrap replicated of the degree distribution of a subnet containing 200 nodes. The boxplot shows the distribution of Akaike weights for the different models in table 1. Note that classical powerlaws, against expectation, appear to capture relatively little information contained in the degree distribution.

1. simulated data under a model not among the candidate models.
2. simulated data under a model which is among the candidate models.
3. random subnets generated from the presently available *S.cerevisiae* PIN data.

In each case we seek to infer the degree distribution and the total number of interactions in the complete dataset. Good performance was found in each case but here we focus on results obtained from the third approach.

In figure 1 we show the estimates for the degree distribution resulting from taking subnets of the present yeast PIN in the database of interacting proteins (`http://dip.doe-mbi.ucla.edu`) with 200, 400, 1000 and 2000 nodes, respectively; the true network contains interactions among 4773 proteins. These were generated by picking N_S nodes and looking at the induced subgraph; this was then used to fit each of the five probability models to the data and then averaging the degree distribution across models weighted by their respective Akaike weights.

We find that even for a network containing approximately 4% of the nodes (and hence $\approx 1.6\%$ of the interactions, the estimated shape of the degree distributions offers a reasonable description of the degree distribution of the true network. We then compared the performance of the model-averaged estimate for the degree distribution with the maximum-likelihood powerlaw fit to the degree distribution of the "complete network". Even for N_S the relative information captured by the model-averaged estimate obtained from the incomplete data is at least two orders of magnitude more than that of a powerlaw model fitted to the complete model. This is a very encouraging result.

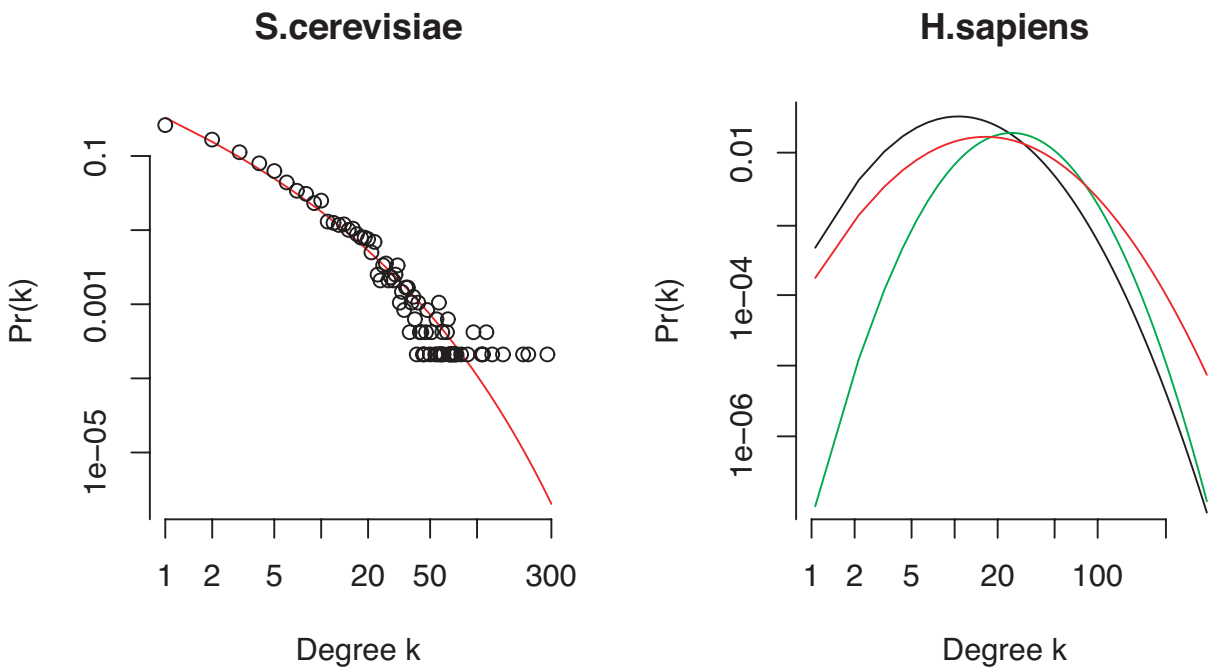

Figure 3: Empirical degree distribution of present yeast PIN dataset (black circles) and estimated degree distribution for the full yeast PIN. The three human estimates correspond to the datasets of Stelzl [26] (black), Rual [24] (red) and DIP (green); Individual human datasets were downloaded from the IntAct database (`www.ebi.ac.uk/intact`).

A further observation already apparent from figure 1 is that the variance of the estimate decreases as the fraction of sampled nodes increases. For smaller subnets it is therefore advisable to use variance-reduction techniques such as bagging (bootstrap aggregation) [8]. This is further shown in figure 2 where we show the bagged estimate in the left panel and box-plots of the Akaike weights in the right panel. For small sampling fractions we observe that several of the candidate models have appreciable weight, $\omega_i \gtrsim 1\%$. As the sampling fraction increases (for example from $N_S = 1000$ and 2000 in figure 1) we observe that only one large Akaike weight is typically observed $\omega_i \gtrsim 99\%$. For yeast this is typically either the log-normal or stretched exponential probability model. This may be a shortcoming and signal the inability of these probability models to describe degree distributions of real networks adequately.

4 Application to yeast and human PIN data

There are sufficient PIN data for four eukaryotic species: *H.sapiens*, *D.melanogaster*, *C.elegans* and *S.cerevisiae*. Here we focus on human and yeast PIN data and these cover approximately 6% (*H.sapiens*) to 85% (*S.cerevisiae*) of the protein coding genes. In figure 3 we show the degree distributions resulting from the DIP dataset in yeast and three largely independent human datasets for humans. In yeast we observe that the estimated degree distribution of the complete interactome describes the present dataset very well (also compared to previous studies [28, 29]) reflecting earlier findings that for sufficiently large subnets there are only negligible differences between the degree distribution of the true network and the subnet.

In the right part of figure 3 we find that different human PIN datasets yield different estimates for the degree distributions: while there is good agreement between the DIP dataset and the data

of Stelzl *et al.* [26], the data of Rual *et al.* [24] has a smaller mode. Such differences can be due to a number of factors which include details of the experimental design, choice of proteins to be tested for interactions, and the error rate inherent to high-throughput approaches to protein-protein interaction analysis [2]. An additional factor becomes apparent from our simulation studies which clearly indicate the importance of so-called orphan nodes [31], *i.e.* nodes which in the subnet have degree 0 but finite degree in the whole network. Their frequency in the subnet (which can be very high for fat-tailed degree distributions) also contains information about the structure of the network, which is lost when only connected nodes are reported. This will give rise to a decrease in $\Pr_{\mathcal{S}}(k)$ at small degrees $k \lesssim 10$ and may be partly responsible for the disagreement between the datasets observed here.

Estimates of the total number of interactions in the global *S.cerevisiae* and human PINs agree well with those obtained from the estimator (23) of $\approx 30,000$ and $\approx 700,000$, respectively. The present approach has, however, the advantage to provide realistic estimates of degree distribution and can be used to infer other properties of the network [20] as long as the true network is reasonably well described in terms of an ensemble of uncorrelated networks.

5 Conclusion

Above we have shown that MMI can be employed in order to make predictions about biological networks from present incomplete datasets. We have demonstrated the use of MMI using simulated datasets. In particular for uncorrelated networks such approaches offer a very promising path towards better understanding of networks, their global structure and organization, and their biological properties.

It is well known (see *e.g.* [11]) that MMI can offer a robust statistical approach to estimation and prediction even if the true model is not among the candidate models[27] (we have found this to be the case in simulations leaving out one of the models in table 1); the excellent fit of the MMI estimates to the degree distribution in figure 1 is an excellent illustration of the power of MMI.

All the models in table 1 are purely phenomenological; only for the power-law distribution is it possible to come up with a mechanistic model of growing networks which will ultimately give rise to a powerlaw degree distribution [3, 22, 15]. There are, however, a number of models which give rise to more realistic network ensembles such as duplication-attachment (DA) [34], duplication divergence [5] and multi-fractal models [14]. It is in principle possible to use such models in an MMI framework. Wiuf *et al.* [34] have shown that it is possible to fit DA models to network data, using the whole data and not just the degree distribution. The strength of MMI based on the degree distribution is largely in terms of computational speed and the ability to deal with noise and incomplete data (computational burdens on full likelihood treatments would be enormous). For uncorrelated networks it is furthermore possible to express many other properties of the networks in terms of the degree distribution. For the, intrinsically more interesting and realistic correlated networks [4, 21] it is possible to adapt the current approach and gain more realistic descriptions of network data in a statistical framework.

There are a number of obvious and important extensions of the present approach. Perhaps the two most pressing extensions are:

- More realistic (and perhaps mechanistic) probability models can be included in the MMI analysis of incomplete network data.
- In reality the sampling process is more complicated and depends also on the experimental set-up. This can be included and can be combined with an explicit probability model for false-positive and false-negative results. Prior information about error rates can be straightforwardly included in a more generalized approach.

From a practical perspective it may eventually also become more convenient to adopt a Bayesian model-selection perspective [17] which will allow a more flexible and consistent treatment of many models. In a hierarchical approach it will furthermore be possible to incorporate the effects of noise consistently. Finally, there has been recent progress in the analysis of correlated random networks [23, 21, 6], and in order to make these processes more relevant for the analysis of biological networks, we have to expand the present approach to deal with properties other than the degree distribution (*e.g.* the degree-degree distribution [4] will be a natural quantity).

Overall, our results suggest, that it is possible to predict aspects of complex biological networks from present incomplete datasets (R routines for MMI from network data are available from MPHS on request). MMI applied to *e.g.* protein interaction networks can generate testable hypothesies about biological systems, their organization and complexity. Using more realistic and mechanistic models of network evolution [34, 5] combined with MMI as described here may also provide us with insights into the evolutionary history and functional organization of biological networks. Quite generally, MMI can be used to predict properties of networks from partial data. For uncorrelated networks, where the degree sequence is statistically sufficient, this is particularly straightforward. In fitting data to networks we can however, also condition on other aspects of the data such as degree-degree correlations and/or the observed molular architecture of biological networks using suitable Importance sampling or Markov Chain formalisms [33].

References

[1] H. Akaike. Information measures and model selection. In *Proceedings of the 44th Session of the International Statistical Institute*, pages 277–291, 1983.

[2] Joel S Bader, Amitabha Chaudhuri, Jonathan M Rothberg, and John Chant. Gaining confidence in high-throughput protein interaction networks. *Nat Biotechnol*, 22(1):78–85, Jan 2004.

[3] AL Barabasi and R Albert. Emergence of scaling in random networks. *Science*, 286(5439):509–512, 1999.

[4] J. Berg and M. Lässig. Correlated random networks. *Phys.Rev.Lett.*, 89:228701, 2002.

[5] J. Berg, M. Lassig, and A. Wagner. Structure and evolution of protein interaction networks: A statistical model for link dynamics and gene duplications. *BMC Evolutionary Biology*, 5:51, 2004.

[6] M. Boguna and R. Pastor-Satorras. Class of correlated random networks with hidden variables. *Phys.Rev.E*, 68:036112, 2003.

[7] B. Bollobás. *Random Graphs*. Academic Press, 1998.

[8] L. Breiman. Bagging predictors. *Machine Learning*, 24:123–140, 1996.

[9] Z. Burda, J. D. Correia, and A. Krzywicki. Statistical ensemble of scale-free random graphs. *Phys.Rev. E*, 64:046118, 2001.

[10] Z. Burda and A. Krzywicki. Uncorrelated random networks. *Phys.Rev.E*, 67:046118, 2004.

[11] K.P. Burnham and D.R. Anderson. *Model Selection and Multimodel Inference*. Springer, 1998.

[12] D.R. Cox and D.V. Hinkley. *Theoretical Statistics*. Chapman&Hall/CRC, 1974.

[13] E. de Silva and M.P.H. Stumpf. Complex networks and simple models in biology. *J.Roy.Soc. Interface*, 2005.

[14] S.N. Dorogovtsev, J.F.F. Mendes, and A.N. Samukhin. Multifractal properties of growing networks. *Europhys.Lett.*, 57:334–338, 2002.

[15] T.S. Evans. Complex networks. *Contemporary Physics*, 45(6):455–474, 2004.

[16] J.D.J. Han, D. Dupuy, N. Bertin, M.E. Cusick, and M. Vidal. Effect of sampling on topology predictions of protein-protein interaction networks. *Nature Biotech.*, 23:839–844, 2005.

[17] J. Hoeting, D. Maigan, A. Raftery, and C. Volinski. Bayesian model averaging. *Stat.Sci.*, 14:382–401, 1999.

[18] A. Krzywicki. Defining statistical ensembles of random graphs. *arXiv cond-mat*, page 0110574, 2001.

[19] S.H. Lee, P.J. Kim, and H Jeong. Statistical properties of sampled networks. *Phys.Rev.E*, 73:016102, 2006.

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

[21] M.E.J. Newman. Random graphs as models of networks. In S. Bornholdt and H.G. Schuster, editors, *Handbook of Graphs and Networks*. Wiley-VCH, 2003.

[22] MEJ Newman. The structure and function of complex networks. *SIAM Review*, 45(2):167–256, 2003.

[23] M.E.J. Newman, D.J. Watts, and S.J. Strogatz. Random graph models of social networks. *Proc.Natl.Acad.Sciences USA*, 99:2566–2572, 2002.

[24] JF Rual, K Venkatesan, T Hao, T Hirozane-Kishikawa, A Dricot, N Li, GF Berriz, FD Gibbons, M Dreze, N Ayivi-Guedehoussou, N Klitgord, C Simon, M Boxem, S Milstein, J Rosenberg, DS Goldberg, LV Zhang, SL Wong, G Franklin, S Li, JS Albala, J Lim, C Fraughton, E Llamosas, S Cevik, C Bex, P Lamesch, RS Sikorski, J Vandenhaute, HY Zoghbi, A Smolyar, S Bosak, R Sequerra, L Doucette-Stamm, ME Cusick,

DE Hill, FP Roth, and M Vidal. Towards a proteome-scale map of the human protein-protein interaction network. *Nature*, 437(7062):1173–8, 2005.

[25] J. Schäfer and K. Strimmer. Learning large-scale graphical gaussian models from genomic data. In *Proceedings of CNET 2004*, 2004.

[26] U Stelzl, U Worm, M Lalowski, C Haenig, FH Brembeck, H Goehler, M Stroedicke, M Zenkner, A Schoenherr, S Koeppen, J Timm, S Mintzlaff, C Abraham, N Bock, S Kietzmann, A Goedde, E Toks?z, A Droege, S Krobitsch, B Korn, W Birchmeier, H Lehrach, and EE Wanker. A human protein-protein interaction network: a resource for annotating the proteome. *Cell*, 122(6):957–68, 2005.

[27] K. Strimmer and A. Rambaut. Inferring confidence sets of possibly misspecified gene trees. *P.Roy.Soc.Lond. B*, 269:127–142, 2002.

[28] M.P.H. Stumpf and P.J. Ingram. Probability models for degree distributions of protein interaction networks. *Europhys.Lett.*, 71(1):152–158, 2005.

[29] M.P.H. Stumpf, P.J. Ingram, I. Nouvel, and C. Wiuf. Statistical model selection methods applied to biological networks. *Trans.Comput.Systems Biol.*, 3:65–72, 2005.

[30] M.P.H. Stumpf and C. Wiuf. Sampling properties of random graphs: the degree distribution. *Phys.Rev. E*, 72:036118, 2005.

[31] M.P.H. Stumpf, C. Wiuf, and R.M. May. Subnets of scale-free networks are not scale-free: the sampling properties of networks. *PNAS*, 102:4221–4224, 2005.

[32] M.P.H. Stumpf, C. Wiuf, Thorne T., E. de Silva, H. Jun an, and M. Lappe. Predicting the size of the human interactome. *submitted*, 2006.

[33] T. Thorne and M.P.H. Stumpf. Generating confidence intervals on biological networks. *submitted*, 2006.

[34] C. Wiuf, M. Brameier, O Hagberg, and M.P.H. Stumpf. A likelihood approach to the analysis of network data. *PNAS*, 103:7566–7570, 2006.

[35] C. Wiuf and M.P.H. Stumpf. Binomial sampling. *Proc.Royal.Soc.A*, 462:1181–1195, 2006.