# Revisiting date and party hubs: Novel approaches to role assignment in protein interaction networks

Sumeet Agarwal[*]

Supervisor: Nick S. Jones[†]
Co-supervisors: Charlotte M. Deane[‡]and Mason A. Porter[§]

August 27, 2009

**Abstract**

A critical challenge in biology is to uncover the relation between the structure and biological function of an organism's protein-protein interaction network (the *interactome*). By studying the structure of this network, we hope to derive insights into the way proteins are organised. It has been proposed that proteins with a large number of interactions, called hubs, fall into two classes — 'date' and 'party' — that play important roles in the modular organisation of the interactome. This binary classification of hubs was based on the extent to which hubs are co-expressed with their interaction partners and subsequently used to attribute specific biological roles to them. However, despite being widely appreciated, the existence of a date–party hub dichotomy has proven to be quite controversial, thus we revisit this idea and examine possible alternative approaches for role assignment in the interactome.

Through an examination of previous results on date and party hubs, we demonstrate that the alleged bimodality of hub co-expression distributions is not robust to data heterogeneity. A noted property of date hubs is that they have a role in connecting disparate parts of the network; we use a betweenness centrality measure to show that this is not a generic property of date hubs but instead only a small subset of all date hubs are truly central. We also partition the networks into cohesive groups known as communities and use the (purely topological) node measures of within-community degree and participation coefficient to examine the extent to which hubs actually fall into a date/party grouping. We find no evidence for such a dichotomy based on these metrics. We then examine an alternative approach to studying topological roles by employing a betweenness centrality metric on links rather than nodes. We find that such link betweenness also does not correlate with co-expression, but it appears to have a significant correlation with protein colocalisation.

Our results suggest that there is currently little evidence for a clear date/party distinction. Instead, hubs in protein interaction networks seem to perform a variety of roles that fall along a continuum, and there is no strong correlation between these roles and co-expression. Our results also indicate that a link betweenness measure for interactions in such networks is related to similarity in both protein function and cellular location. We thus suggest that a distinction between local and global interactions might be relevant to understanding the organisation of the interactome.

Note: An extended version of this report is to be submitted for publication shortly to *PLoS Biology*.

## Introduction

Advances in molecular biology in recent years have allowed us to acquire a vast store of information about

[*] sumeet.agarwal@physics.ox.ac.uk

[†] Nick.Jones@physics.ox.ac.uk

[‡] deane@stats.ox.ac.uk

[§] porterm@maths.ox.ac.uk

proteins. We now know much about their makeup, their structural forms, the levels at which they are expressed in various situations, and their bindings and interactions. However, there remains a major disconnect between this new knowledge and the traditional study of biology, where living organisms are analysed by breaking them down into organs and organ systems and studying their respective functions. The challenge which has been receiving much attention in the last few years is that of going from the biochemistry of tens of thousands of proteins to the physiology of a relatively small number of high-level functions and processes [1, 2]. A key step in making this connection is to understand how groups of proteins combine to carry out various tasks. Thus, there has been a lot of interest in the study of the interactome, i.e., the set of all physical protein-protein interactions. The interactome can tell us how proteins convey signals to each other, and how coordination amongst them comes about. Given that even a relatively simple organism like baker's yeast (*Saccharomyces cerevisiae*) is thought to have nearly 18,000 protein-protein interactions [3], it is clear that a very complex system underlies the high-level biological functionality which we observe.

From a mathematical perspective, the interactome is a graph or network, where nodes represent proteins and links between them represent binary interactions. A study of the structure and organisation of this network may provide insightful abstractions. However, experimentally determined protein interaction networks do not capture the fact that the actual interactions that occur in vivo depend on prevailing physiological conditions. For instance, the actively expressed proteins vary among the tissues in an organism's body and also change over time. Thus, the specific parts of the interactome that are active, as well as their organisational form, may depend a great deal on where and when one examines the network [4, 5]. One way to incorporate such information is to use protein expression data from microarray experiments. This is challenging, but Han et al. [4] obtained an exciting result when they used such expression data to examine the extent to which hubs in the yeast interactome are co-expressed with their interaction partners, hubs being defined as proteins with degree 5 or more (the number of links emanating from a node in a network is referred to as the node's degree). Based on the averaged Pearson correlation coefficient (avPCC) of expression over all partners, they concluded that hubs fall into two distinct classes: those with a low avPCC (which they called *date* hubs) and those with a high avPCC (so-called *party* hubs). They inferred that these two types of hubs play different roles in the network: Party hubs coordinate single functions performed by a group of proteins that are all expressed at the same time, whereas date hubs are higher-level connectors between groups that perform varying functions and are active at different times or under different conditions.

The validity of the date/party hub distinction has since been debated in a sequence of papers [6–9], and there appears to be no consensus on the issue. However, the idea has been widely adopted in the literature (e.g., [3, 5, 10, 11]) and has been one of the more influential concepts in the study of protein interaction networks in recent years. Here we revisit the initial data and suggest possible problems with the statistical methodology that was employed. In particular, we show that the differing behaviour observed on the deletion of date and party hubs [4], which seemed to suggest that date hubs were more essential to global connectivity, was largely due to a just very small number of key date hubs. More generally, our results indicate that there is little correlation between co-expression and the structural roles of hubs in the network. A recent study by Taylor et al. [5] claimed to demonstrate the existence of 'intermodular' and 'intramodular' hubs — a categorisation along the same lines as date and party hubs — in the human interactome. However, we show that their observation of bimodality is susceptible to methodological changes.

Many real-world networks display some sort of modular organisation, as they can be partitioned into cohesive groups of nodes that have a relatively high ratio of internal to external connection densities. Such subnetworks, known as *communities*, often correspond to distinct functional units [12–14]. Several studies in recent years have considered the existence of community structure in protein-protein interaction networks [15–23]. Additionally, myriad algorithms have been developed for detecting communities in networks [13, 14]. We use the idea of community structure to take a new approach to the problem of hub classification by attempting to assign roles to hubs purely on the basis of network topology rather than on the basis of expression data. Our rationale is that the biological roles of date and party hubs are essentially topological in nature and should thus be identifiable from the network alone. Once we have partitioned the network into a set of meaningful communities, it is possible to compute statistics to measure the connectivity of each hub both within its own community and to other communities. One method for assigning relevant roles to nodes in a metabolic network was described by Guimerà and Amaral [24], and we follow an analogous procedure for the hubs in our protein interaction networks. We then examine the extent to which these roles match up with the date/party hypothesis, finding little evidence to support it.

An alternative way to think about topological roles in networks is to define measures on links rather than on nodes. We use a measure of link significance known as *betweenness centrality* [12,25] and examine its relation to phenomena such as protein co-expression and functional overlap. Here as well we find little evidence of a significant correlation with co-expression. However, there seems to be a stronger relation between link betweenness and functional similarity of the interacting proteins, so that link-centric role definitions might have some utility.

In summary, we have examined the proposed division of hubs in the protein interaction network into the date and party categories from several different angles, demonstrating that prior results in favor of a date/party dichotomy appear to be susceptible to various kinds of changes in the data and methods used. Observed differences in network vulnerability to attacks on the two hub types seem to arise from only a small number of particularly important hubs. Furthermore, a more detailed analysis of network structure and the roles of hubs within it suggests that the picture is more complicated than a simple dichotomy, as proteins in the interactome show a variety of topological characteristics, and these are not correlated with the co-expression of the proteins' interaction partners. On the other hand, investigating link betweenness centralities reveals an interesting relation to the functional linkage of proteins, suggesting that a framework incorporating a more nuanced notion of roles for both nodes and links might provide a better framework for understanding the organisation of the interactome.

# Methodology and Results

## Revisiting Date and Party Hubs

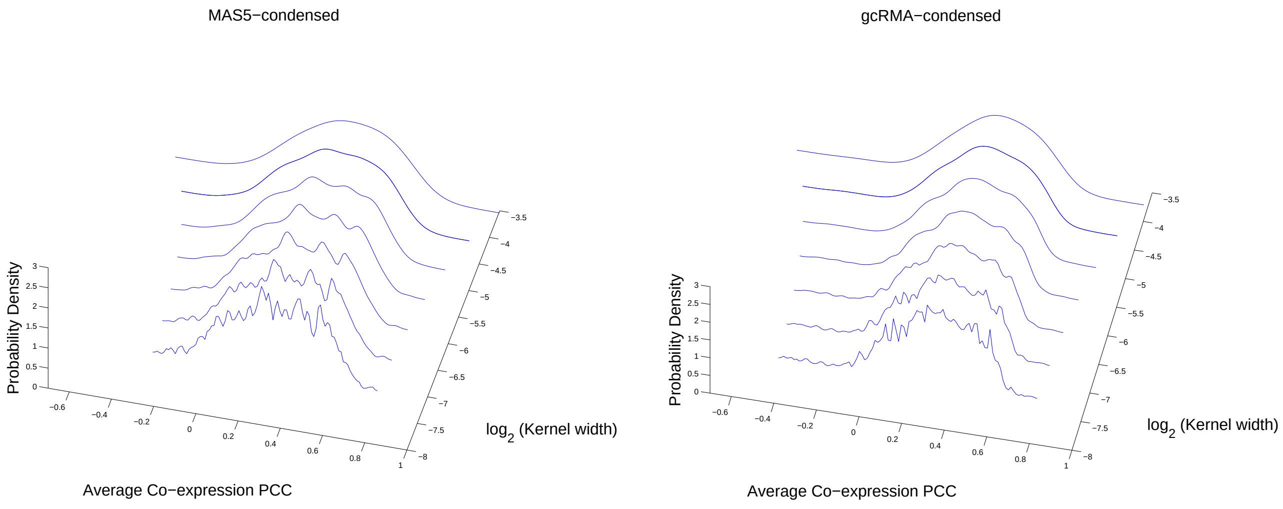
The definition of date and party hubs is based on the expression correlation of hubs in the protein interaction network with their interactors. Specifically, the avPCC was computed for each hub and its distribution was observed to be bimodal. A date/party threshold value of avPCC (for a given expression data set) was defined so as to best separate the two modes [4].

Recent support for the idea of date and party hubs appeared in a paper which considered data relating to the human interactome; the authors found bimodal distributions of avPCC values [5]. We used an interaction data set provided by the authors of Ref. [5] as an updated version of the one used in their paper (sourced from the Online Predicted Human Interaction Database [26]), and found that the bimodal distribution of hub co-expression observed by them is not robust to methodological changes. For instance, raw intensity data from microarray probes has to be processed and normalised in order to obtain comparable expression values for each gene [27]. The expression data used in Ref. [5] (taken from the human GeneAtlas [28]) was normalised using the MAS5 algorithm; when we repeated the analysis using the same data normalised with the gcRMA algorithm instead, we obtained significantly different results. This is shown in Figure 1, which depicts avPCC distributions for hubs (defined, as per Ref. [5], as the top 15% of nodes by degree) in the two cases. Density plots have been obtained for varying smoothing kernel widths. The gcRMA-processed data does not appear to lead to a substantially bimodal distribution at any kernel width, whereas the MAS5-processed data appears to give bimodality for only a relatively narrow range of widths, and could just as easily be regarded as trimodal.

We also find variability across different interaction data sets: For instance, we analysed the recent protein-fragment complementation assay (PCA) data set [29] and found no clear evidence of a bimodal distribution of hubs along date/party lines (data not shown). Even in cases where bimodality is seen, it might be arising as an artefact of combining different types of interaction data; there are beleived to be significant and systematic biases in which types of interactions each data-gathering method is able to obtain [3, 23, 29]. For instance, analysing avPCC values for hubs in networks obtained from Y2H or AP/MS alone [3], we find that 100% (259/259) are date hubs in the former but that only about 30% (56/186) are date hubs in the latter.

One of the key pieces of evidence used to argue that date and party hubs have distinct topological properties was the apparent observation of different effects upon their deletion from the network. Removing date hubs seemed to lead to very rapid disintegration into multiple components, whereas removal of party hubs had much less effect on global connectivity [4, 7]. However, it has been observed that removing just the top 2% of hubs by degree from the comparison of deletion effects obviates this difference, suggesting that the observation is actually due to just a few extreme date hubs [8]. In order to study this in greater detail, we used node betweenness centrality [25], which is a way of quantifying the importance of individual nodes or links to the connectivity of a network. The (geodesic) betweenness centrality of a node/link is defined as the number of pairwise shortest paths in the network that pass through that object [12, 25].

We found that in the original 'filtered yeast interactome' (FYI) data set [4], date hubs have on average somewhat higher betweenness centralities ($1.79 \times 10^4$ for 91



(a)

(b)

Figure 1: Probability density plots of the distribution of hub avPCC values for human interaction data from OPHID (provided by Taylor et al. [5]). Gene expression data from GeneAtlas [28], normalised using (a) MAS5 and (b) gcRMA [27]. Curves obtained using a normal smoothing kernel function at varying window widths.

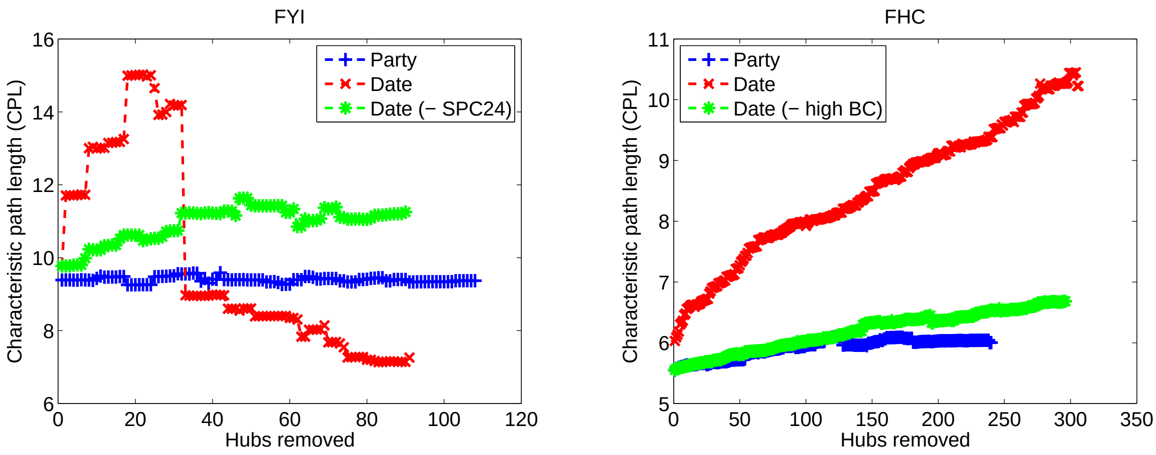
date hubs versus $1.07 \times 10^4$ for 108 party hubs). However, there happens to be one date hub (SPC24, a highly connected protein involved in chromosome segregation [30]) which has an exceptionally high betweenness in this network, and when the set of date hubs minus this one hub is attacked, we find the observed difference between date and party hubs is greatly reduced (Figure 2(a)).

It was subsequently shown that the FYI network was particularly sparse, and as more data became available the updated 'filtered high-confidence' (FHC) data set was used to perform the same analysis [7]; in this case the network did not break down on attacking date hubs, but nevertheless displayed a substantially greater increase in characteristic path length than seen for party hub deletion. For FHC too, date hubs have on average higher betweenness values ($3.7 \times 10^4$ for 306 date hubs versus $2.15 \times 10^4$ for 240 party hubs). However, the larger average is due almost entirely to a small number of hubs with unusually high betweenness, as removing the top 10 date hubs by betweenness (which all had values higher than any party hub) nearly equalised the averages. Furthermore, the removal of just these 10 hubs from the set of attacked date hubs is sufficient to virtually obviate the difference with party hubs, as shown in Figure 2(b). Notably, the set of 10 high-betweenness hubs includes prominent proteins such as Actin (ACT1), Calmodulin (CMD1), and the TATA binding protein (SPT15), which are known to be central to important cellular processes. Thus, we can account for the critical nodes for network connectivity using just a few major hubs, and most of the proteins that are classified as date hubs appear to be no more critical than the party hubs. It is also evident that the 10 key hubs in the FHC network show a wide range of avPCC values (Figure 2(c)), further weakening the claim that there is an inverse relation between a hub's avPCC and its central role in the network.

## Communities in the Interactome

Many real-world networks can be divided naturally into close-knit subnetworks called communities. The investigation of algorithms for detecting communities in networks has received considerable attention in recent years [13, 14].

From an intuitive standpoint, communities should consist of groups of nodes, such that there are many links between nodes in the same group but few links between nodes in different groups. To detect communities algorithmically, this notion must be formalised. One of the most popular ways of doing this is to optimise the quality function known as graph 'modularity' [32,33]. Supposing that an unweighted network with $n$ nodes and $m$ links



(a)

(b)

| Protein | Degree | AvPCC | Functions |
|---|---|---|---|
| SMT3 | 42 | 0.08 | Not known; suppressor of MIF2 mutations |
| PAB1 | 25 | 0.28 | Important mediator of the roles of the poly(A) tail in mRNA biogenesis, stability and translation |
| HSP82 | 37 | 0.19 | Maturation, maintenance and regulation of proteins involved in cell cycle control and signal transduction |
| GLC7 | 35 | -0.01 | Glycogen metabolism, meiosis, translation, chromosome segregation, cell polarity, cell cycle progression |
| ACT1 | 35 | 0.13 | Cell motility |
| CDC28 | 202 | 0.06 | Essential for the completion of the start, the controlling event, in the cell cycle |
| PSE1 | 24 | 0.28 | Nuclear import of ribosomal proteins; protein secretion |
| SPT15 | 50 | 0.12 | Regulation of gene expression by RNA polymerase II |
| CMD1 | 46 | 0.05 | Mediates the control of a large number of enzymes and other proteins |
| RPO21 | 58 | 0.05 | DNA-directed RNA polymerase |

(c)

Figure 2: (a) Effects of hub deletion on network connectivity. ‘Date (− SPC24)’ refers to the set of date hubs minus the protein SPC24. In each case, we used the complete FYI network [4] consisting of 1379 nodes as the starting point and then deleted all hubs in the given set from the network in order of decreasing degree. The characteristic path length is the mean of the lengths of all finite paths between two nodes in the network. (b) Effects of hub deletion on network connectivity for the FHC network [7]. ‘Date (− high BC)’ refers to the set of date hubs minus the 10 hubs with the highest betweenness centrality (BC) values. We used the upper bound on the BC for party hubs as a threshold to define these 10 ‘high BC’ date hubs. (c) List of the 10 high-betweenness FHC hubs, with degrees, avPCC values (as computed using the ’Compendium’ expression data set [4, 31]), and functional annotations from UniProt [30].

is divided into $N$ communities $(C_1, C_2, \cdots, C_N)$. Let $k_i$ denote the degree (number of links) of node $i$ and let $A$ be the $n \times n$ adjacency matrix, so that $A(i,j)$ is 1 if nodes $i$ and $j$ have a link between them and 0 if they do not. The modularity $Q$ is then given by [33]

$$Q = \frac{1}{2m} \sum_{l=1}^{N} \sum_{i,j \in C_l} \left( A_{ij} - \frac{k_i k_j}{2m} \right) , \qquad (1)$$

where $k_i k_j/(2m)$ is the expected number of links between nodes $i$ and $j$ in a network with the same expected degree distribution but with links placed at random. Graph modularity thus captures how many more links there are within the specified communities than one would expect to see by chance in a network with no modular structure. Note, however, that (1) assumes a particular null model that explicitly preserves the expected degree distribution in the random setting. It is possible to employ other null models [14], though this one is employed most commonly.

Using this framework, we can detect communities by maximising graph modularity over all possible network partitions. Because this problem is known to be NP-complete [34], reliably finding the global maximum is computational intractable even for small networks. Thankfully, there exist a number of good computational heuristics that can be used to obtain good local maxima [13,14,35]. The approach we employed uses the physical interpretation of this problem as finding the ground state of a Potts spin glass [36]. The nodes can be treated spins, with links representing ferromagnetic interactions and lack of links corresponding to antiferromagnetic interactions. Under a natural choice of parameters, finding the ground state is then equal to finding the maximum modularity partitioning (with each spin state corresponding to a community). Thus, we can recast modularity maximisation as an energy minimisation problem, and then apply an appropriate optimisation algorithm. Here we have used a spectral bisection algorithm [37].

In principle, one should be able to view a categorisation of hubs according to the date/party dichotomy directly in the network structure, as the two kinds of hubs are posited to have different neighbourhood topologies. In order to identify the community structure of the various interaction networks that we examined, we employ the method described above. In Figure 3, we show the network partition (with nodes coloured according to community) that results from applying such an optimisation to the largest connected component of the FYI data set [4].

In order to assess how well the obtained topological communities reflect functional organisation, we used annotations from the GO database [39] to define their Information Content (IC). GO provides a controlled vocabulary for describing genes and gene products such as proteins using a limited set of annotation terms. It consists of three separate ontologies — one each for biological process, cellular component, and molecular function. For each community, we computed the $p$-value of the most-enriched GO annotation term; the frequency of this term within its community is highest relative to its background frequency in the entire network. To do this, we used the hypergeometric distribution, which corresponds to random sampling without replacement. The extent of enrichment can then be gauged using IC [40], which is defined as

$$IC = -\log_{10}(p) , \qquad (2)$$

where $p$ denotes the $p$-value.

In Table 1, we summarise the results of calculating the IC measure for communities detected on two of the yeast interaction data sets. For comparison, we also examine a random partition of FYI into communities with the same size distribution as the actual ones. It is clear that there is on average very significant functional enrichment within the detected communities. In particular, it is far greater than could be expected by chance. This is in accordance with previous studies on communities in protein interaction networks [18,20–23]. Thus, the topology of the interaction network provides a great deal of information about functional organisation.

## Topological Properties and Node Roles

Given that one can find functionally meaningful communities based on interaction data alone, it is natural to ask whether something like the date/party distinction can also be observed based only on interaction data. We thus leave gene expression data to one side for the moment and focus on what can be inferred about node roles purely from network topology. Guimerà and Amaral [24] have proposed a scheme for classifying nodes into topological roles in a modular network according to their pattern of intramodule and intermodule connections. Their classification uses two statistics for each node — within-community degree and participation coefficient — and divides the plane that they define into seven role boxes.

The *within-community degree* refers to the number of connections a node has within its own community. It is normalized here to a $z$-score, which for the $i^{th}$ node is given by the formula

$$z_i = \frac{\kappa_i - \bar{\kappa}_{s_i}}{\sigma_{\kappa_{s_i}}} , \qquad (3)$$

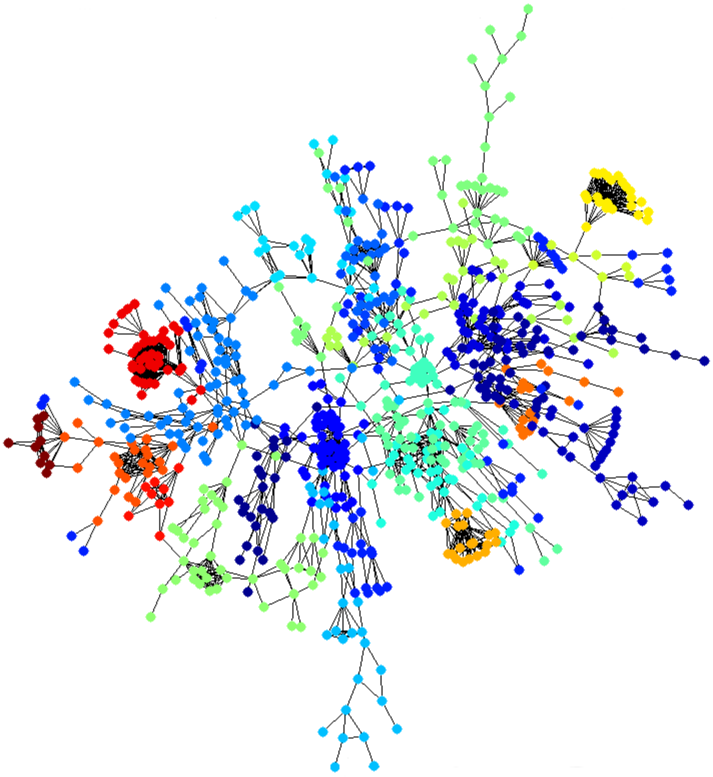


Figure 3: Community structure in the largest connected component of the FYI network; the different colours correspond to different communities (25 in all). The graph modularity value for this partition is -0.8784. We generated this visualisation using the Kamada-Kawai algorithm [38].

| **Data set** | **Communities** | **MF IC** | | | **CC IC** | | | **BP IC** | | | **Best IC** | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Min | Max | Avg | Min | Max | Avg | Min | Max | Avg | Min | Max | Avg |
| FYI | 25 | 2.05 | 43.09 | 14.36 | 4.28 | 51.60 | 17.18 | 2.99 | 35.74 | 15.72 | 4.81 | 51.60 | 20.15 |
| FYI | 25 (random) | 1.28 | 2.78 | 1.88 | 1.25 | 3.00 | 2.07 | 1.46 | 3.04 | 2.13 | 1.46 | 3.04 | 2.36 |
| FHC | 63 | 1.47 | 51.37 | 11.22 | 0.11 | 68.18 | 16.40 | 1.73 | 98.51 | 17.08 | 1.97 | 98.51 | 20.08 |

Table 1: Information Content (IC) of the most enriched term for each of the three GO ontologies (MF – Molecular Function; CC – Cellular Component; BP – Biological Process) and over all three ontologies combined ('Best IC'). We give the minimum, maximum, and average IC over all of the communities that we detected in a given data set. We generated the random communities for FYI using the same size distribution as the actual ones, i.e., the actual community labels of all proteins were removed and then randomly re-assigned, one label per protein.

where $s_i$ denotes the community label of node $i$, $\kappa_i$ is the number of links of node $i$ to other nodes in the same community $s_i$, the quantity $\bar{\kappa}_{s_i}$ is the average of $\kappa$ for all nodes in community $s_i$, and $\sigma_{\kappa_{s_i}}$ is the standard deviation of $\kappa$ in community $s_i$. The *participation coefficient* of node $i$ measures how its links are distributed amongst different communities. It is defined as [24]

$$P_i = 1 - \sum_{s=1}^{N} \left( \frac{\kappa_{is}}{k_i} \right)^2 , \tag{4}$$

where $N$ is the number of communities, $\kappa_{is}$ is the number of links of node $i$ to nodes in community $s$, and $k_i$ is the total degree of node $i$. The participation coefficient approaches 1 if the links of node $i$ are uniformly distributed amongst all communities (including its own) and is 0 if they are all within its own community.

We plot all nodes in the network in a two-dimensional space using coordinates determined by within-community degree and participation coefficient, and we divide the space into regions that correspond to 7 different node roles, as per Ref. [24]. We depict these 7 roles as demarcated regions in the plots in Figure 4, which shows the node roles for yeast and human data sets, computed based on the communities we detected by optimising modularity.

Some of the topological roles defined by this method would appear to correspond to those ascribed to date/party hubs. For instance, party hubs ought to be ‘provincial hubs’, which have many links within their community but few or none outside. Date hubs might be construed as ‘non-hub connectors’ or ‘connector hubs’, both of which have links to several different modules; they could also fall into the ‘kinless’ roles. We thus sought to examine the relationship between the date/party classification and this topological role classification. In Figure 4, we colour proteins according to their avPCC. In Figure 5, we present the same data in a more compact form, as we only show the hubs (defined in Ref. [7] as the top 20% of nodes ranked by degree) in the two interaction networks, plotting them according to node role and avPCC. The horizontal lines correspond to an avPCC of 0.5, which was the threshold used to distinguish date and party hubs in the yeast interactome [7].

One immediate observation from these results is that the avPCC threshold clearly does not carry over to the human data. In fact, all of the hubs in the latter have an avPCC of well below 0.5. Even if we utilize a different threshold in the human network, we find that there is little difference in the avPCC distribution across the topological roles, suggesting that no meaningful date/party categorisation can be made (at least for this data set). This might be the case because the human data set likely represents only a small fraction of the actual interactome. Additionally, it is derived from only one technique (Y2H) and is thus not multiply-verified like the yeast data set.

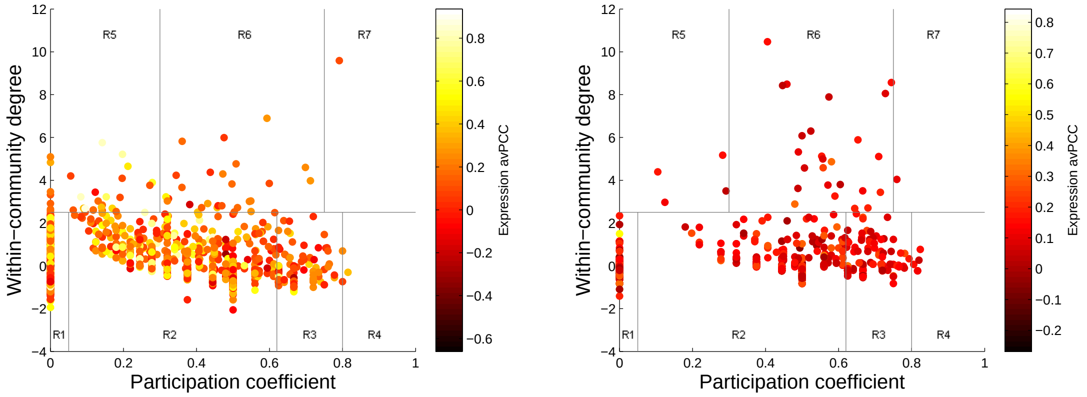
For yeast, we see that hubs below the threshold line (the supposed date hubs) include not only most of those that fall into the ‘connector’ roles but also many of the ‘provincial hubs’. On the other hand, those that lie above the line (the supposed party hubs) include mainly the provincial hub and peripheral categories. Although one can discern a difference in role distributions above and below the threshold, it is not very clear-cut and the so-called date hubs fall into all 7 roles. It would thus appear that even for yeast, the distribution of hubs is not bimodal, and the properties attributed to date and party hubs [4] do not seem to correspond very well with the actual topological roles that we estimate here. Indeed, these roles are more diverse than what can be explained using a simple dichotomy.

## The Roles of Interactions

Most research on interactome properties has focused on node-centric metrics, which draws on the perspective of individual proteins. Here we try an alternative approach that instead uses link-centric metrics in order to examine how the topological properties of interactions in the network relate to their function. In order to quantify the importance of a given link to global network connectivity, we use link betweenness centrality [12, 25]. We investigate the relationship between link betweenness and the expression correlation for a given interaction. If date and party hubs genuinely exist, one might expect a similar sort of dichotomy for interactions, with more central interactions having lower expression correlations and vice versa. In Figure 6, we depict all of the interactions in two yeast data sets, which we position on a plane based on the values of their link betweenness and interactor expression PCC. Additionally, we colour each point according to the level of functional similarity between the interacting proteins, as determined by overlap in GO (Cellular Component) annotations. In order to compute this functional similarity, we first define the set information content (SIC) [40] of each term in our ontology for a given data set. Suppose the complete set of proteins is denoted by $S$, and the subset annotated by term $i$ is denoted by $S_i$. The SIC of the term $i$ is then defined as

$$SIC(i) = -\log_{10} \left( \frac{|S_i|}{|S|} \right) . \tag{5}$$

Now suppose that we have two interacting proteins called $A$ and $B$. Let $S_A$ and $S_B$, respectively, denote their com-



(a)

(b)

Figure 4: Topological node role assignments for (a) yeast (FHC; 2,233 nodes, 63 communities) and (b) human (CCSB-HI1; 1,307 nodes, 38 communities) interaction data sets. Following Guimerà and Amaral [24], we designate the roles as follows: R1 – Ultra-peripheral; R2 – Peripheral; R3 – Non-hub connector; R4 – Non-hub kinless; R5 – Provincial hub; R6 – Connector hub; and R7 – Kinless hub. We colour proteins according to the avPCC of expression with their interaction partners. We computed expression avPCC using the stress response data set [41] (this being by far the largest of the expression data sets used in the original study [4]) for FHC and COXPRESdb [42] for CCSB-HI1. We assigned an avPCC of 0 to proteins for which no partner co-expression data was available.

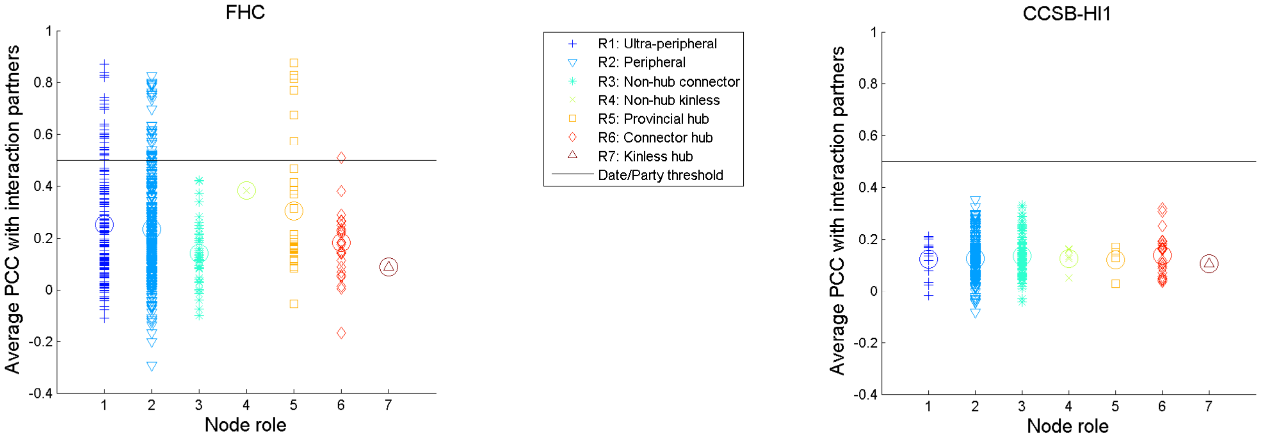


Figure 5: Node role versus average expression correlation with partners for hubs in yeast (FHC; 553 hubs with a minimum degree of 7) and human (CCSB-HI1; 326 hubs with a minimum degree of 4) networks. Larger circles represent means over all nodes in a given role. Note that 'hub' as used in the role names refers only to within-community hubs, but all of the depicted nodes are hubs in the sense that they have high degree. In each case, we determined the degree threshold so that approximately the top 20% highest-degree nodes are considered to be hubs. We also fixed the date/party avPCC threshold at 0.5, in accordance with Bertin et al. [7].

plete sets of annotations (consisting of not only their leaf terms but also all of their ancestors) from the ontology. Then the functional similarity of the proteins is given by

$$f(A,B) = \frac{\displaystyle\sum_{i \in (S_A \cap S_B)} SIC(i)}{\displaystyle\sum_{j \in (S_A \cup S_B)} SIC(j)} \,. \tag{6}$$

For the FHC data set, we find that there is no significant relation between expression PCC and link betweenness. For the FYI data set, we observe a dense cluster of interactions in the top left of the plot, but most of these are interactions within ribosomal complexes. If one removes such interactions from the data set, then here too one finds little relation between expression PCC and link betweenness. (Note that ribosomal proteins were already removed from FHC [7].) On the other hand, we find a fairly strong correlation between the logarithm of link betweenness and similarity in cellular component annotations, which can be used as a measure of colocalisation. In particular, there appears to be a natural threshold at the modal value of betweenness. This is somewhat reminiscent of the weak/strong tie distinction in social networks [43, 44], as the 'weak' (high betweenness) interactions serve to connect and transmit information between distinct spatial modules, which are composed predominantly of 'strong' (low betweenness) interactions. Figure 7 depicts the distribution of interactions involving a protein annotated with the GO term 'signal transduction', for both the FYI and FHC datasets. This further demonstrates the high-betweenness nature of such interactions.

# Discussion

In this report we have analysed modular organisation and the roles of hubs in protein interaction networks. We revisited the possibility of a date/party hub dichotomy and found significant points of concern. In particular, claims of bimodality in hub avPCC distributions do not appear to be robust across available interaction and expression data sets, and tests for the differences observed on deletion of the two hub types did not consider important outlier effects. Moreover, there is considerable evidence to suggest that the observed date/party distinction is at least partly an artefact of the different properties of the Y2H and AP/MS data sets.

In order to study the topological properties of hub nodes in greater detail, we partitioned protein interaction networks into communities and examined the statistics of the distributions of hub links. Our results show that hubs can exhibit an entire spectrum of structural roles and that there is little evidence to suggest a definitive date/party classification. We find, moreover, that co-expression of a hub with its partners is not a strong predictor of its topological role, and the overall extent of such co-expression varies considerably across the data sets that we examined.

As an alternative way of defining roles in the interactome, we have also investigated a link-centric approach, in which we study the topological properties of links as opposed to nodes. In particular, we examined betweenness centrality as an indicator of a link's importance to network connectivity. We found that this too does not correlate significantly with co-expression of the interacting proteins. For certain data sets, however, it does appear to correlate quite strongly with the functional similarity of the proteins. Additionally, there appears to be a threshold value of betweenness centrality beyond which one observes a sudden drop in functional similarity. We also found that the high-betweenness interactions are enriched for interactions involved in signalling. This suggests that a concept of intramodular versus intermodular interactions, somewhat analogous to the weak/strong tie dichotomy in social networks, might be useful. This sort of link-based role definition may also be applicable to other types of real-world networks, and an interesting general question in network science may be to examine how node- and link-based role definitions relate to each other for different kinds of networks.

# Future Work

In the coming year, we hope to build on this work by attempting to integrate information from other kinds of biological networks, such as genetic interaction and regulatory networks and metabolic networks. Each of these represent different levels of biological organisation, but in principle they form a single network since there is constant interaction and flow of information between the different levels. By looking at the structural properties of the different levels and comparing them, we may be able to come up with more useful notions of things such as the role played by a given gene/protein/metabolite.

We would also like to use different kinds of network data to attempt to construct predictive models of biological dynamics, which allow us to track the network state through time. Some success has already been achieved in this direction [45], via learning of ordinary differential equation (ODE) models of gene expression dynamics. We will attempt to look at alternative ways of modeling these systems, in particular possible coarse-graining into qual-

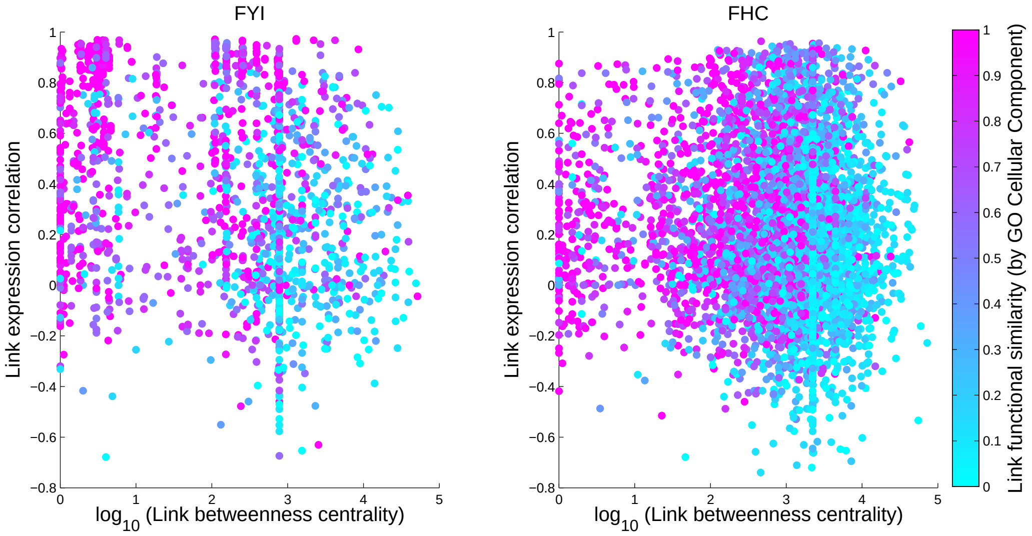


Figure 6: Link betweenness centralities and expression correlations, with points coloured according to average similarity of interactors' GO (Cellular Component) annotations, for two protein interaction data sets: FYI (778 nodes, 1,798 links) and FHC (2,233 nodes, 5,750 links).

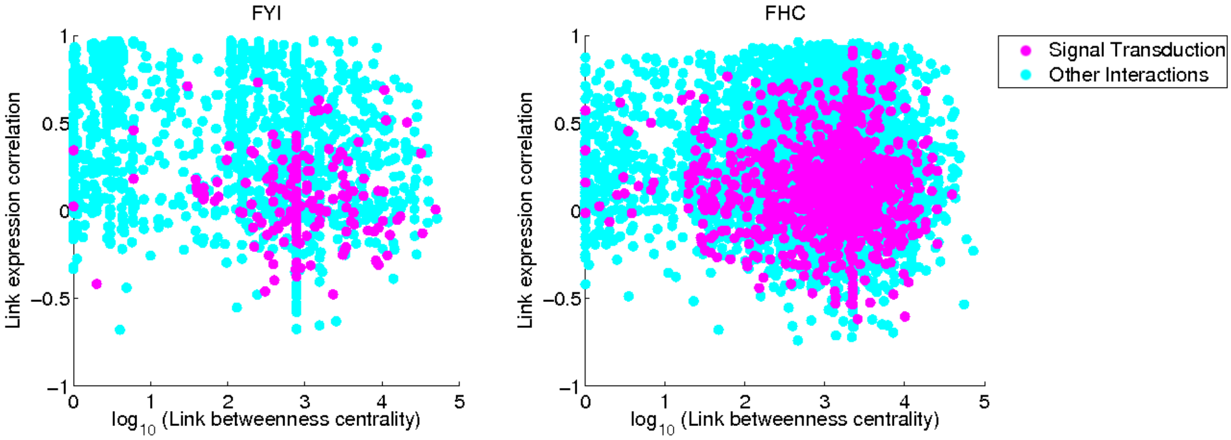


Figure 7: Link betweenness centralities and expression correlations, with points coloured according to whether the interaction involves a protein annotated with the GO term 'signal transduction', for two protein interaction data sets: FYI (778 nodes, 1,798 links) and FHC (2,233 nodes, 5,750 links).

itative dynamics [46] (such as just representing a gene as on/off, in the simplest case), as it is widely believed that biological systems tend to exist in a relatively small number of discrete states.

## Acknowledgments

We thank Patrick Kemmeren for providing the yeast expression data sets; Jeffrey Wrana for the curated OPHID human interaction and GeneAtlas expression data; Pao-Yang Chen and Waqar Ali for pointing us to relevant online data repositories; Peter Mucha and Stephen Reid for useful discussions and help with some code; and Anna Lewis for several useful discussions, as well as providing GO annotation data and Matlab code for computing functional similarities. We also thank Max Little for useful discussions.

## References

[1] Hartwell LH, Hopfield JJ, Leibler S, Murray AW (1999) From molecular to modular cell biology. Nature 402: C47–C52.

[2] Barabási AL, Oltvai ZN (2004) Network biology: Understanding the cell's functional organization. Nat Rev Genet 5: 101–113.

[3] Yu H, Braun P, Yildirim MA, Lemmens I, Venkatesan K, et al. (2008) High-quality binary protein interaction map of the yeast interactome network. Science 322: 104–110.

[4] Han JDJ, Bertin N, Hao T, Goldberg DS, Berriz GF, et al. (2004) Evidence for dynamically organized modularity in the yeast protein-protein interaction network. Nature 430: 88–93.

[5] Taylor IW, Linding R, Warde-Farley D, Liu Y, Pesquita C, et al. (2009) Dynamic modularity in protein interaction networks predicts breast cancer outcome. Nature Biotechnology 27: 199–204.

[6] Batada NN, Reguly T, Breitkreutz A, Boucher L, Breitkreutz BJJ, et al. (2006) Stratus not altocumulus: A new view of the yeast protein interaction network. PLoS Biology 4: e317.

[7] Bertin N, Simonis N, Dupuy D, Cusick ME, Han JDJ, et al. (2007) Confirmation of organized modularity in the yeast interactome. PLoS Biology 5: e153.

[8] Batada NN, Reguly T, Breitkreutz A, Boucher L, Breitkreutz BJ, et al. (2007) Still stratus not altocumulus: Further evidence against the date/party hub distinction. PLoS Biology 5: e154.

[9] Wilkins MR, Kummerfeld SK (2008) Sticking together? Falling apart? Exploring the dynamics of the interactome. Trends in Biochemical Sciences 33: 195–200.

[10] Kim PM, Lu LJ, Xia Y, Gerstein MB (2006) Relating three-dimensional structures to protein networks provides evolutionary insights. Science 314: 1938–1941.

[11] Komurov K, White M (2007) Revealing static and dynamic modular architecture of the eukaryotic protein interaction network. Mol Sys Bio 3: 110.

[12] Girvan M, Newman MEJ (2002) Community structure in social and biological networks. Proc Natl Acad Sci USA 99: 7821–7826.

[13] Fortunato S (2009) Community detection in graphs. E-print arXiv: 0906.0612.

[14] Porter MA, Onnela J-P, Mucha PJ (2009) Communities in networks. Notices of the American Mathematical Society (to appear). E-print arXiv: 0902.3788.

[15] Spirin V, Mirny LA (2003) Protein complexes and functional modules in molecular networks. Proc Natl Acad Sci USA 100: 12123–12128.

[16] Rives AW, Galitski T (2003) Modular organization of cellular networks. Proc Natl Acad Sci USA 100: 1128–1133.

[17] Yook SH, Oltvai ZN, Barabási AL (2004) Functional and topological characterization of protein interaction networks. Proteomics 4: 928–942.

[18] Dunn R, Dudbridge F, Sanderson CM (2005) The use of edge-betweenness clustering to investigate biological function in protein interaction networks. BMC Bioinformatics 6: 39.

[19] Gavin AC, Aloy P, Grandi P, Krause R, Boesche M, et al. (2006) Proteome survey reveals modularity of the yeast cell machinery. Nature 440: 631–636.

[20] Adamcsek B, Palla G, Farkas IJ, Derenyi I, Vicsek T (2006) Cfinder: locating cliques and overlapping modules in biological networks. Bioinformatics 22: 1021–1023.

[21] Chen J, Yuan B (2006) Detecting functional modules in the yeast protein-protein interaction network. Bioinformatics 22: 2283–2290.

[22] Maraziotis I, Dimitrakopoulou K, Bezerianos A (2008) An in silico method for detecting overlapping functional modules from composite biological networks. BMC Systems Biology 2: 93.

[23] Lewis ACF, Jones NS, Porter MA, Deane CM (2009) The function of communities in protein interaction networks. E-print arXiv: 0904.0989.

[24] Guimerà R, Amaral LAN (2005) Functional cartography of complex metabolic networks. Nature 433: 895–900.

[25] Freeman LC (1977) A set of measures of centrality based on betweenness. Sociometry 40: 35–41.

[26] Brown KR, Jurisica I (2005) Online predicted human interaction database. Bioinformatics 21: 2076–2082.

[27] Lim WK, Wang K, Lefebvre C, Califano A (2007) Comparative analysis of microarray normalization procedures: effects on reverse engineering gene networks. Bioinformatics 23: i282–i288.

[28] Su AI, Wiltshire T, Batalov S, Lapp H, Ching KA, et al. (2004) A gene atlas of the mouse and human protein-encoding transcriptomes. Proc Natl Acad Sci USA 101: 6062–6067.

[29] Tarassov K, Messier V, Landry CR, Radinovic S, Molina MM, et al. (2008) An in vivo map of the yeast protein interactome. Science 320: 1465–1470.

[30] The UniProt Consortium (2008) The universal protein resource (UniProt). Nucleic Acids Res 36: D190–D195.

[31] Kemmeren et al P (2002) Protein interaction verification and functional annotation by integrated analysis of genome-scale data. Molecular Cell 9: 1133–1143.

[32] Newman MEJ, Girvan M (2004) Finding and evaluating community structure in networks. Phys Rev E 69: 026113.

[33] Newman MEJ (2006) Modularity and community structure in networks. Proc Natl Acad Sci USA 103: 8577–8582.

[34] Brandes U, Delling D, Gaertler M, Gorke R, Hoefer M, et al. (2008) On modularity clustering. Knowledge and Data Engineering, IEEE Transactions on 20: 172–188.

[35] Danon L, Diaz-Guilera A, Duch J, Arenas A (2005) Comparing community structure identification. Journal of Statistical Mechanics: Theory and Experiment 2005: P09008.

[36] Reichardt J, Bornholdt S (2006) Statistical mechanics of community detection. Phys Rev E 74: 016110.

[37] Newman MEJ (2006) Finding community structure in networks using the eigenvectors of matrices. Phys Rev E 74: 036104.

[38] Kamada T, Kawai S (1989) An algorithm for drawing general undirected graphs. Information Processing Letters 31: 7–15.

[39] Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, et al. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nature Genet 25: 25–29.

[40] Resnik P (1995) Using information content to evaluate semantic similarity in a taxonomy. In: Proc. 14th Int'l Joint Conf. Artificial Intelligence. pp. 448–453.

[41] Gasch AP, Spellman PT, Kao CM, Carmel-Harel O, Eisen MB, et al. (2000) Genomic expression programs in the response of yeast cells to environmental changes. Mol Biol Cell 11: 4241–4257.

[42] Obayashi T, Hayashi S, Shibaoka M, Saeki M, Ohta H, et al. COXPRESdb: a database of coexpressed gene networks in mammals. Nucleic Acids Res 36: D77–D82.

[43] Rapoport A (1957) Contributions to the theory of random and biased nets. Bulletin of Mathematical Biophysics 19: 257–277.

[44] Granovetter MS (1973) The strength of weak ties. Amer J of Sociology 78: 1360–1380.

[45] Bonneau R, Facciotti MT, Reiss DJ, Schmid AK, Pan M, et al. (2007) A predictive model for transcriptional control of physiology in a free living cell. Cell 131: 1354–1365.

[46] Srinivasan A, King RD (2008) Incremental identification of qualitative models of biological systems using inductive logic programming. Journal of Machine Learning Research 9: 1475–1533.