Additional File 1 for 'The Function of Communities in Protein Interaction Networks at multiple scales'

Anna C F Lewis, Nick S Jones , Mason A Porter and Charlotte M Deane*

Email: Anna C F Lewis - lewis@stats.ox.ac.uk; Nick S Jones - nick.jones@physics.ox.ac.uk; Mason A Porter - porterm@maths.ox.ac.uk; Charlotte M Deane - deane@stats.ox.ac.uk;

*Corresponding author

Separation of Interactions into types A and P

The IntAct database [1] gives interaction types from the Molecular Interaction ontology [2] directly. It contains 23632 interactions of type A and 26611 of type P. The Mint database [3] uses the Molecular Interaction interaction detection type ontology, the broad categories of which are biophysical, biochemical, and protein complementation assay. The biochemical techniques give evidence of association (type A interactions), and the biophysical and protein complementation assays give evidence of physical interactions (type P). Using this division, there are 13347 A type interactions and 10407 P type interactions. The BioGrid database [4] uses its own evidence types. Those giving evidence of P type interactions are reconstituted complex, PCA, Co-crystal structure and yeast-two-hybrid. Those giving evidence of type A interactions are affinity capture, biochemical activity, co-fractionation, co-purification and Far Western. Details of these experimental types can be found on the BioGrid website, www.thebiogrid.org. There are 35716 A type interactions and 13142 P type interactions. All of the numbers that we report are based on interactions between proteins that each have a SGD reference number (Saccharomyces Genome Database, www.yeastgenome.org, [5]).

Convention for identifying communities at different partitions

To relate the partition at one value of the resolution parameter λ to that at another (which is useful for visualisation), we require a convention for labelling communities. Here we use a method based on the overlap of shared nodes [6]. A convention based on links rather than nodes gives nearly identical results. Let the communities in the first partition (which here is that at the highest resolution) be labelled

$K_1, .., K_s$, and those in the next partition be labelled $L_1, ..., L_t$. Then for each pair of communities, $\{K_i, L_j\}$, we have

$$W_{ij} = \frac{|K_i \cap L_j|}{|K_i \cup L_j|}, \tag{1}$$

where $|B|$ denotes the cardinality (number of elements) of the set B. Starting with the largest value of W_{ij}, we relabel community i as community j. Relabelling proceeds with the next largest W_{ij}, as long as community i is not yet relabelled, until all communities have been relabelled. If $s > t$, we introduce a new label.

Broad classification of protein types

The protein types considered are all the GO Biological Process terms in the GOslim ontology [7] that are annotated to at least 200 proteins in yeast. They are (numbers of proteins in brackets): 1. DNA metabolic process (357); 2. protein modification process (465); 3. transport (859); 4. response to stress (458); 5. membrane organization (208); 6. RNA metabolic process (715); 7. vesicle-mediated transport (280); 8. response to chemical stimulus (298); 9. cellular lipid metabolic process (204); 10. cellular carbohydrate metabolic process (220) and 11. chromosome organization (338).

References

1. Kerrien S, Alam-Faruque Y, Aranda B, Bancarz I, Bridge A, Derow C, Dimmer E, Feuermann M, Friedrichsen A, Huntley R, et al.: **IntAct–open source resource for molecular interaction data**. *Nucleic acids research* 2007, **35**(Database issue):D561.

2. Hermjakob H, Montecchi-Palazzi L, Bader G, Wojcik J, Salwinski L, Ceol A, Moore S, Orchard S, Sarkans U, von Mering C, et al.: **The HUPO PSI's molecular interaction format—a community standard for the representation of protein interaction data**. *Nature Biotechnology* 2004, **22**(2):177–183.

3. Zanzoni A, Montecchi-Palazzi L, Quondam G, Helmer-Citterich M, Cesareni G: **MINT: a Molecular INTeraction database**. *FEBS Letters* 2002, **513**:135–140.

4. Stark C, Breitkreutz BJ, Reguly T, Boucher L, Breitkreutz A, Tyers M: **BioGRID: a general repository for interaction datasets**. *Nucleic acids research* 2006, **34**(Database Issue):D535.

5. Cherry JM, Adler C, Ball C, Chervitz SA, Dwight SS, Hester ET, Jia Y, Juvik G, Roe T, Schroeder M, et al.: **SGD: Saccharomyces genome database**. *Nucleic Acids Research* 1998, **26**:73.

6. Palla G, Barabási A-L, Vicsek T: **Quantifying social group evolution**. *Nature* 2007, **446**(7136):664–667.

7. Hong EL, Balakrishnan R, Dong Q, Christie KR, Park J, Binkley G, Costanzo MC, Dwight SS, Engel SR, Fisk DG, et al.: **Gene Ontology annotations at SGD: new data sources and annotation methods**. *Nucleic acids research* 2008, **36**(Database issue):D577.

8. Lancichinetti A, Fortunato S, Radicchi F: **Benchmark graphs for testing community detection algorithms**. *Physical Review E* 2008, **78**(4):46110.

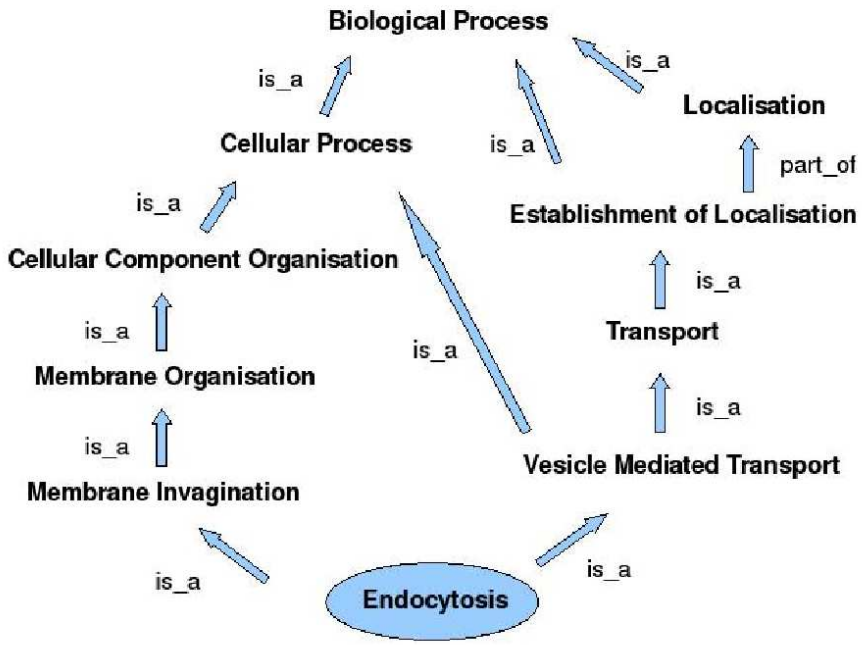

Figure 1: **Illustration of the structure of the GO.** Terms in each sub-ontology of the GO are related to each other via a directed acyclic graph.

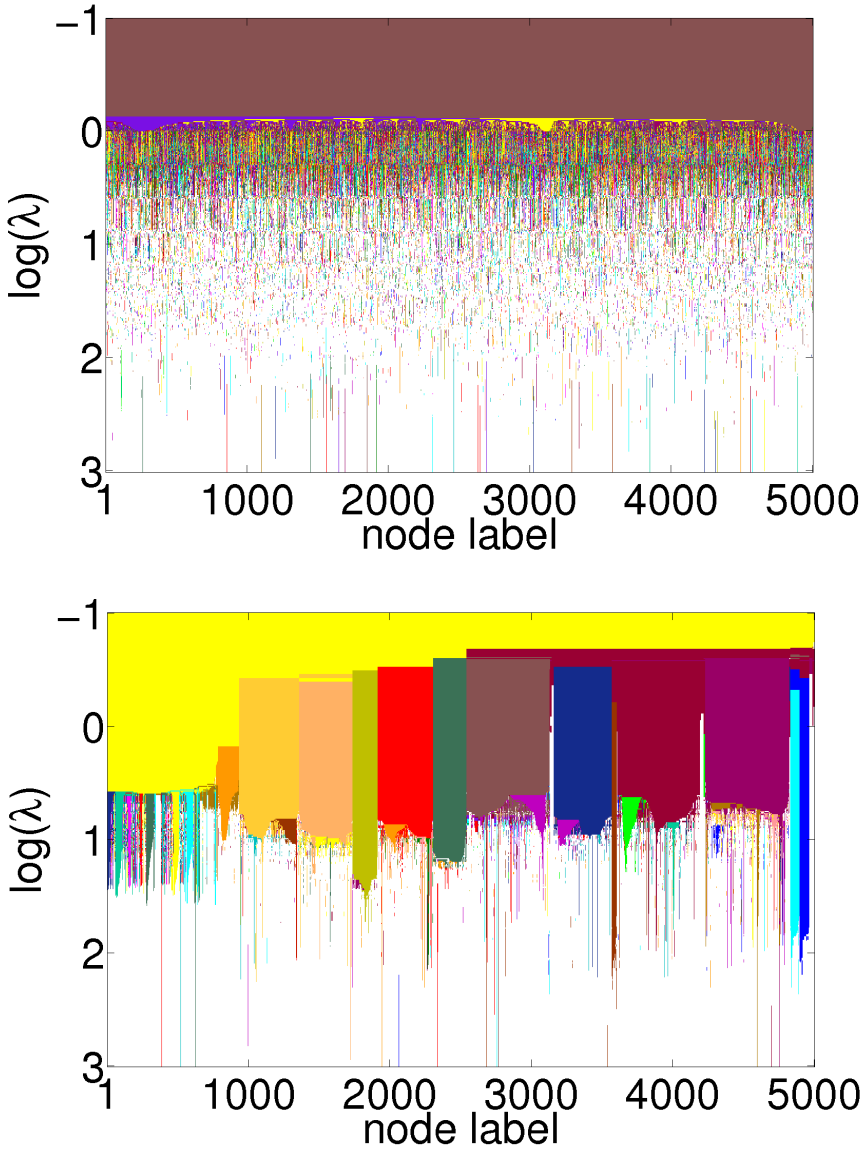

Figure 2: **As for Figure 1, but for a) An Erdös-Rényi random network and b) a network with strong community structure.** Both networks were designed to be of approximately the same size as the A and P networks (5000 nodes). The probability that two nodes are connected in the random network is the same as for the A network. We generated the network with community structure from code available at http://sites.google.com/site/santofortunato/inthepress2, which is reported in [8]. The parameters that we chose matched the statistics of the A network (average degree of 19, maximum degree of 1182), with additional parameters chosen as suggested default values (the exponent for the degree distribution is 2, the exponent for the community size distribution is 1, and the mixing parameter is 0.2).

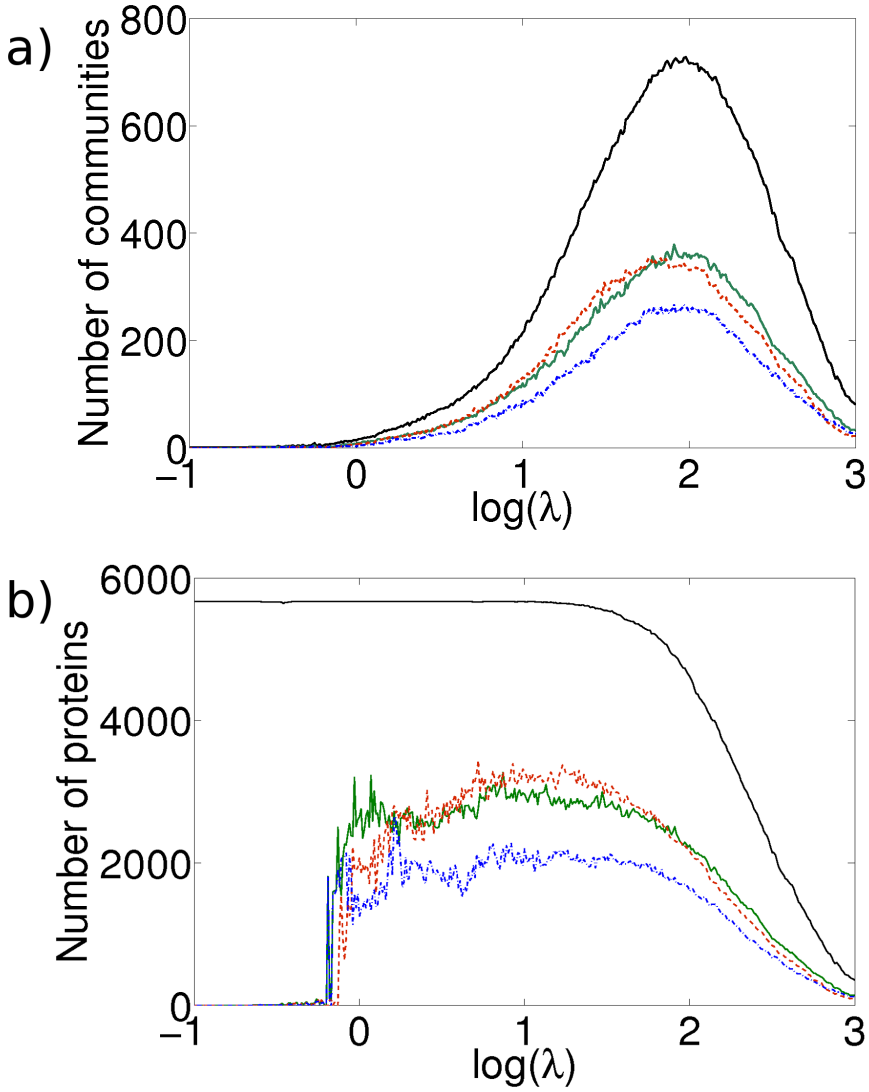

Figure 3: **As for Figure 2, but for the P network.** a) The number of communities with changing resolution parameter (solid black curve) b) The number of proteins p in communities of size four or more (solid black curve). We also show the numbers of communities and the number of proteins in such communities judged to be functionally homogeneous according to the GO similarity measure (green curves), the MIPS measure (dot-dashed blue curves) and the correlated growth similarity measure (dashed red curves). The curves are similar to the A network, and they show a similar proportion of proteins in functionally homogeneous communities. One difference is that there are more proteins in functionally homogeneous communities at a lower value of $\log(\lambda)$ than for the A network.

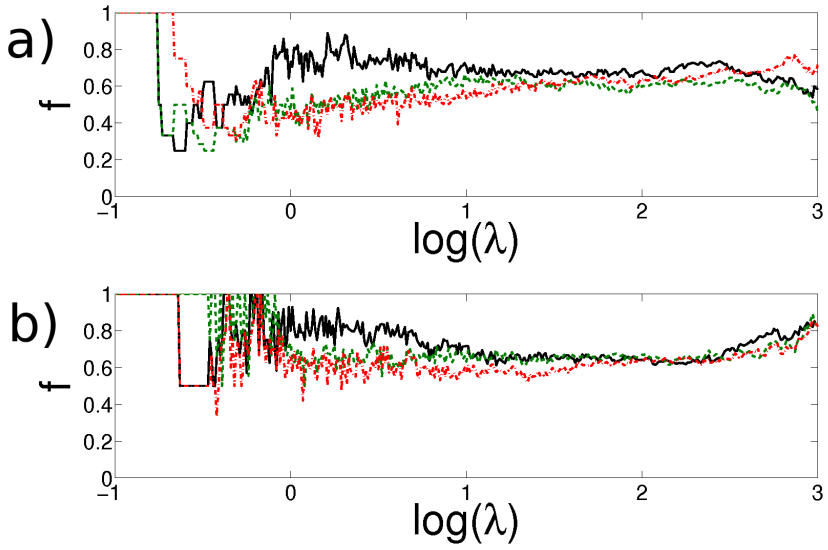

Figure 4: **The agreement in assessment of functional homogeneity between pairs of similarity measures.** For a) the A network and b) the P network, the fraction, f, of communities that are either both judged as functionally homogeneous or both not judged as functionally homogeneous under the G and C measures (black curve), the G and M measures (dark green dashed curve), and the M and C measures (red dot-dashed curve). The large degree of overlap between the measures derived from ontologies (G and M) with the measure derived from a single large scale experiment (C) gives us confidence in our ontology derived measures.

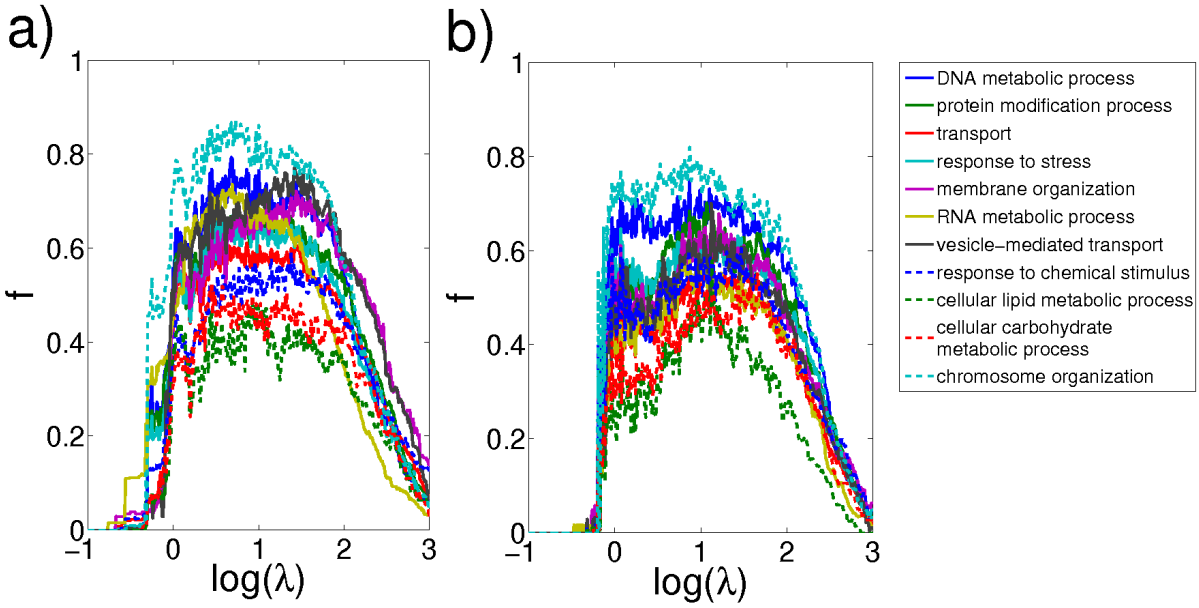

Figure 5: **The fraction of proteins of different types in functionally homogeneous communities as judged under the GO similarity measure.** The fraction, f, of proteins of particular types that are in functionally homogeneous communities in a) the A network and b) the P network, with changing resolution parameter. Some protein types are consistently more likely to be found in functionally homogeneous communities through changing resolution parameter. For example, proteins involved in chromosome organisation are much more likely to be in functionally homogeneous communities than proteins involved in metabolism. There are also some features that suggest 'good' resolutions for particular processes. The same patterns as for the A network hold for which types of protein tend to be classified in functionally homogeneous communities (see main text), but there do not appear to be any clear differences between protein types at varying resolutions in the P network, though some types have clearer peaks than others.

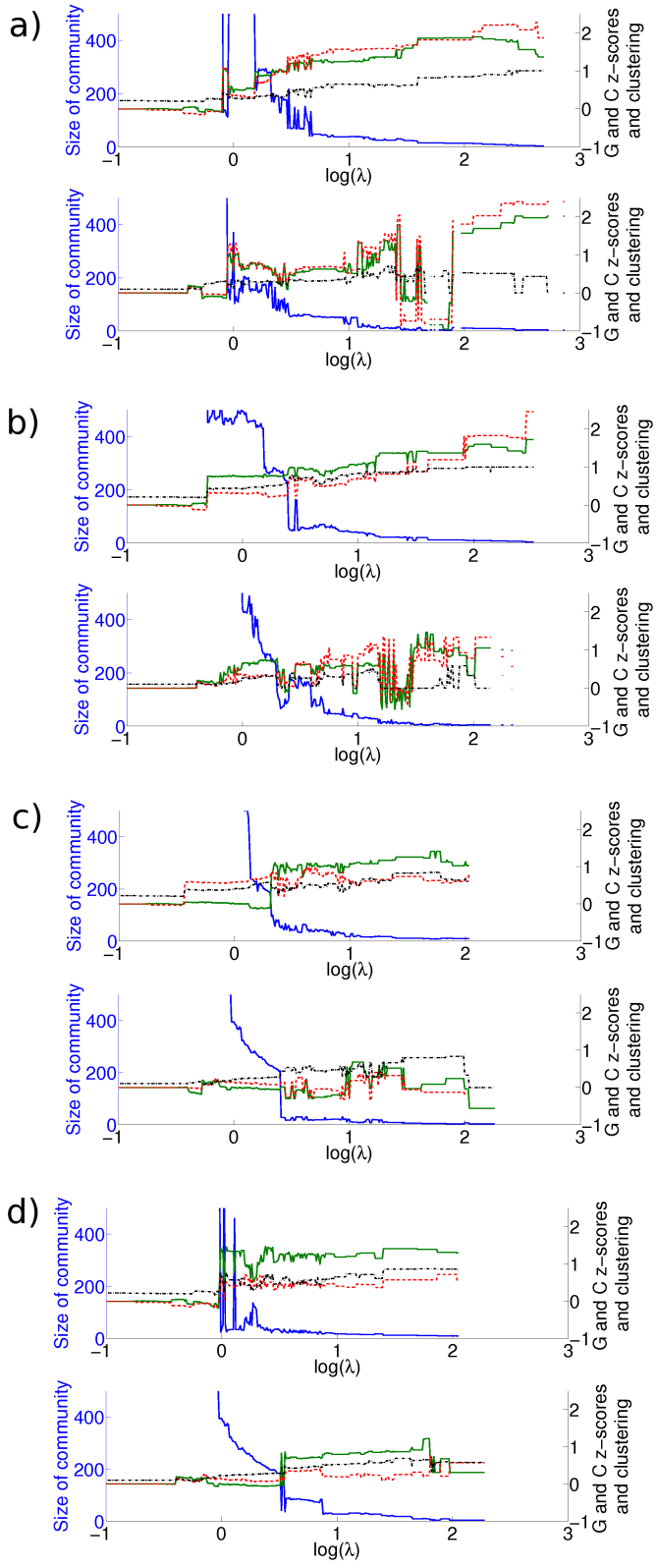

Figure 6: **Further examples as per Figure 5.** These figures display the same information as Figure 5, but for the proteins a) YAL002W, b) YAL011W, c) YAL016W, and d) YAL021C. We show the size (solid blue curve), mean clustering coefficient (dot-dashed black curve), mean z-score under the GO measure (solid green curve), and correlated growth measure (dashed red curve) with changing resolution for the A network (top) and P network (bottom). Gaps appear whenever the protein is assigned to a community of size three proteins or less. We give the names of proteins in several example communities, chosen as motivated by these figures, in Additional File 2.

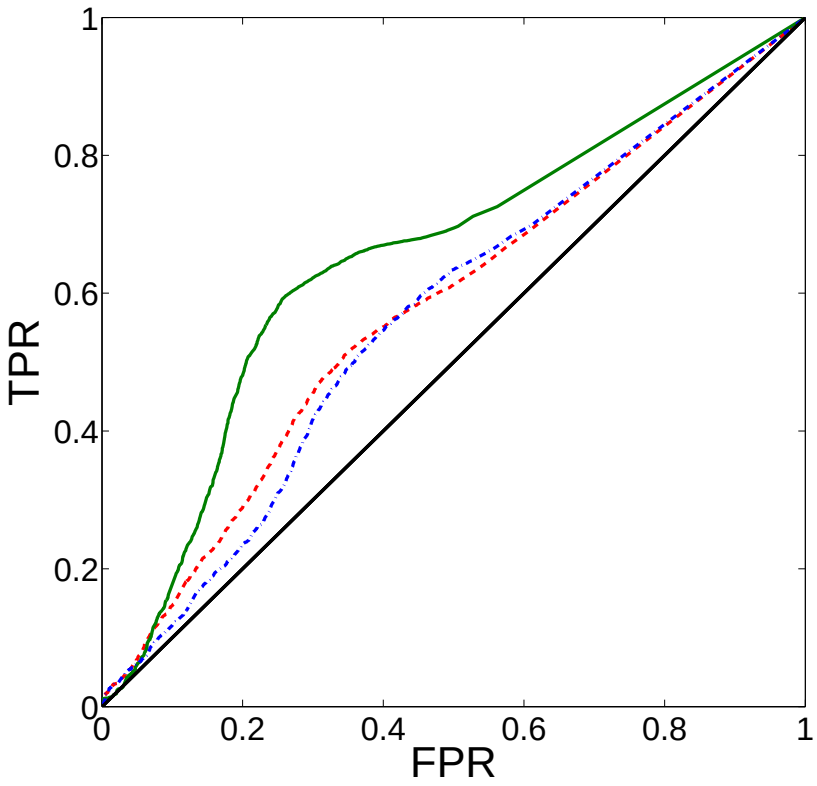

Figure 7: **As for Figure 6, but for the P network.** The Receiver Operating Characteristic (ROC) curve for using mean clustering coefficient as a predictor of functional homogeneity using the GO measure (solid green curve), MIPS measure (dot-dashed blue curve), and correlated growth measure (dashed red curve). We plot the false positive rate (FPR) versus the true positive rate (TPR). A random classifier would give the solid black line. As for the A network, we achieve the best predictive ability using the GO measure and the worst for the MIPS measure. The AUCs for the P network are in general lower than those for the A network (see Table 4).