Supplementary Information

Modelling the Yeast Interactome

Vuk Janjić[1], Roded Sharan[2] and Nataša Pržulj[1,*]

[1] Department of Computing, Imperial College London, London, United Kingdom
[2] Blavatnik School of Computer Science, Tel-Aviv University, Tel-Aviv 69978, Israel
[*] Corresponding author: N. Pržulj, Department of Computing, Imperial College London, London, United Kingdom; E-mail: natasha@imperial.ac.uk

Graphlet degree distribution agreement. Networks are commonly used to represent real-world relational data, thus the ability to compare networks is needed. However, comparing large networks is a computationally infeasible task, since in order to demonstrate similarity between two such networks, it is required to quantify the similarity between their exponentially many properties. Hence, various heuristics, such as the degree distribution, clustering coefficient, diameter, and relative graphlet frequency distribution have been sought.

Imposing similarity constraints on only a few of these properties can easily be achieved on two very large and different networks. For instance, it is easy to construct two networks with exactly the same degree distributions but whose structure and function differ substantially [1–3]. It is computationally infeasible to analyse *all* networks properties, but imposing a large number of such constraints will increase the likelihood of detecting genuine similarity between networks.

Graphlet degree distribution (GDD) agreement (GDDA) is a network comparison measure which captures 73 network similarity constraints. It generalises the notion of a degree distribution, which measures the number of nodes "touching" k edges, into 73 distributions measuring the number of nodes "touching" k graphlets; graphlets are small, connected, non-isomorphic subgraphs of a large network. In fact, the degree distribution is the first in the spectrum of 73 GDDs. Other properties captured by GDD include multi-edge paths, bi-partite sub-structures, triangles, squares, etc. GDDA then "combines and reduces" this large space of 73 graphlet degree distributions into an "agreement" measure — a number between 0 and 1, where 1 represents perfect agreement between networks. For more details on GDDA and its applications, see [3, 4].

Measuring model-to-data network similarity. When comparing real data to network models we used 6 measures of similarity (detailed in Methods): clustering coefficient, degree distribution, average shortest path, diameter, radius and GDDA. In the main manuscript text we present GDDA results and here we show that other network properties agree with GDDA.

To increase confidence when measuring model-to-data fit, we perform every comparison on 15 instances for each of the five random network models (see Methods for random network models) and take the mean and standard deviation over those 15 iterations; this gives us 18 (PPI data sets) $\times(1+14)$ (full network + functional sub-networks) $\times 5$ (network properties) $\times 5$ (random network models) $\times 15$ (instances of each model) $= 101,250$ similarity comparisons. This is *in addition* to the GDDA comparisons presented in the main text.

Since a network's clustering coefficient (CC) is a value in the $[0, 1]$ range, for each of the 5 random network models we take the CC average (and standard deviation) over all 15 instances and compare it to the CC of the real PPI network that the model is based on. The closer the model's value of average CC is to the CC of its corresponding real PPI network, the better that random model captures the CC of the real PPI network. We compare degree distribution between each of the 15 model network instances and their corresponding original PPI network by scaling and normalising the area under the probability distributions (so that they are easily comparable; see Methods) and then taking the average and standard deviation of the 15 resulting similarity values. Since diameter and radiality measures, d and r, can vary greatly from network to network, we summarise them into a single value as $e = \frac{d}{2r}$ which is in the $[0, 1]$ range. Again, the closer the average e value, $\bar{e}$, of the 15 model instances is to the e value of its corresponding PPI network, the better that model fits the original PPI network.

For instance, when looking at the full BioGRID network, GDDA gives STICKY as the best fitting model, followed by ER-DD, SF, GEO and ER models (see Figure 1 in the main text). The clustering coefficient also gives STICKY and ER-DD as best fitting models, followed by a SF, GEO and finally the ER model. The diameter and radiality measures also produce the same ordering of model-to-data fit as GDDA: STICKY, then ER-DD, followed by SF, GEO and ER. The degree distribution gives ER-DD as the best fit, but this is to be expected since the ER-DD is a degree-distribution preserving model and a perfect fit of 1.0 (from the $[0, 1]$ similarity range) is to be expected. Thus, apart from ER-DD, the degree distribution yields a result which is consistent with GDDA: STICKY as the best fitting model followed by SF, GEO and ER models.

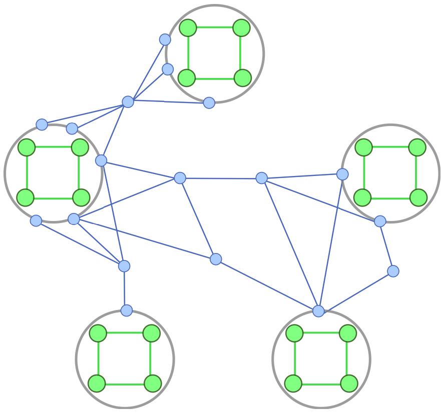


Figure SF1: **Illustration of "GEO-STICKY duality" in PPI networks.** Protein interactions inside functional modules of PPI networks are organized geometrically (🟢) while proteins and interactions that are shared among modules that link different functionalities behave in a STICKY manner (🔵).

Network	num. of nodes	num. of edges
BioGRID	5,891	74,642
Literature curated	1,533	2,839
Affinity capture luminescence	15	19
Affinity capture MS	4,804	43,972
Affinity capture RNA	3,932	6,369
Affinity capture western	2,923	8,237
Biochemical activity	2,054	5,469
Co-crystal structure	560	372
Co-fractionation	718	764
Co-localization	449	492
Co-purification	984	1,352
Far western	104	77
FRET	128	122
PCA	1,744	5,007
Protein-peptide	399	653
Protein-RNA	501	514
Reconstituted complex	2,176	4,112
Yeast two-hybrid	3,557	11,171

Table ST1: **Yeast protein-protein interaction (PPI) networks.** The first entry (BioGRID) is the full network of PPIs from BioGRID (i.e., includes physical interactions from all experimental evidence). The second entry (Literature curated) represents a PPI network constructed from a set of literature curated PPIs given in Reguly et al. (2006) [5]. All subsequent networks are derived from BioGRID based on the experimental evidence supporting each interaction: a sub-network is comprised of interactions detected by one screening technology.

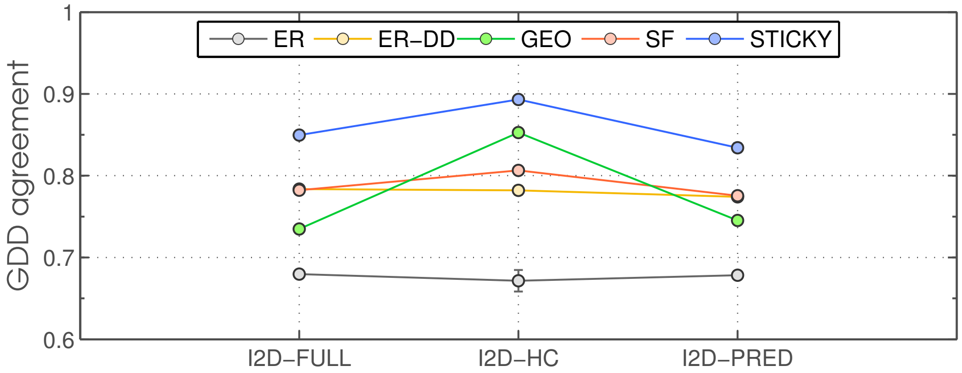


Figure SF2: **Random graph models of human PPI networks.** The fit of five random graph models to human PPI networks. The three data sets are: 1) I2D-FULL (all interactions from I2D), 2) I2D-HC (only the high confidence subset of I2D interactions), and 3) I2D-PRED (only predicted interactions from I2D). Data explained in the main text.

Label	Function
A	Cell cycle progression/meiosis
B	Nuclear-cytoplasmic transport
C	ER-Golgi traffic
D	RNA processing
E	Signaling/stress response
F	Chrom. seg./kinetoch./spindle/microtub.
G	Protein degredation/proteosome
H	DNA replication & repair/HR/cohesion
I	Chromatin/transcription
J	Golgi/endosome/vacuole sorting
K	Protein folding & glycosylation/cell wall
L	Metabolism/mitochondria
M	Ribosome/translation
N	Cell polarity/morphogenesis

Table ST2: **Functional annotation categories for the yeast interactome.** The 14 functional categories used for annotating *S. cerevisiae* proteins.

Label	Function
A	death
B	protein metabolism
C	signal transduction
D	other metabolic processes
G	developmental processes
I	DNA metabolism
J	cell-cell signaling
H	cell adhesion
K	other biological processes
M	RNA metabolism
F	stress response
N	transport
L	cell organization and biogenesis
E	cell cycle and proliferation

Table ST3: **Functional annotation categories for the human interactome.** The 14 functional categories used for annotating *H. sapiens* proteins.

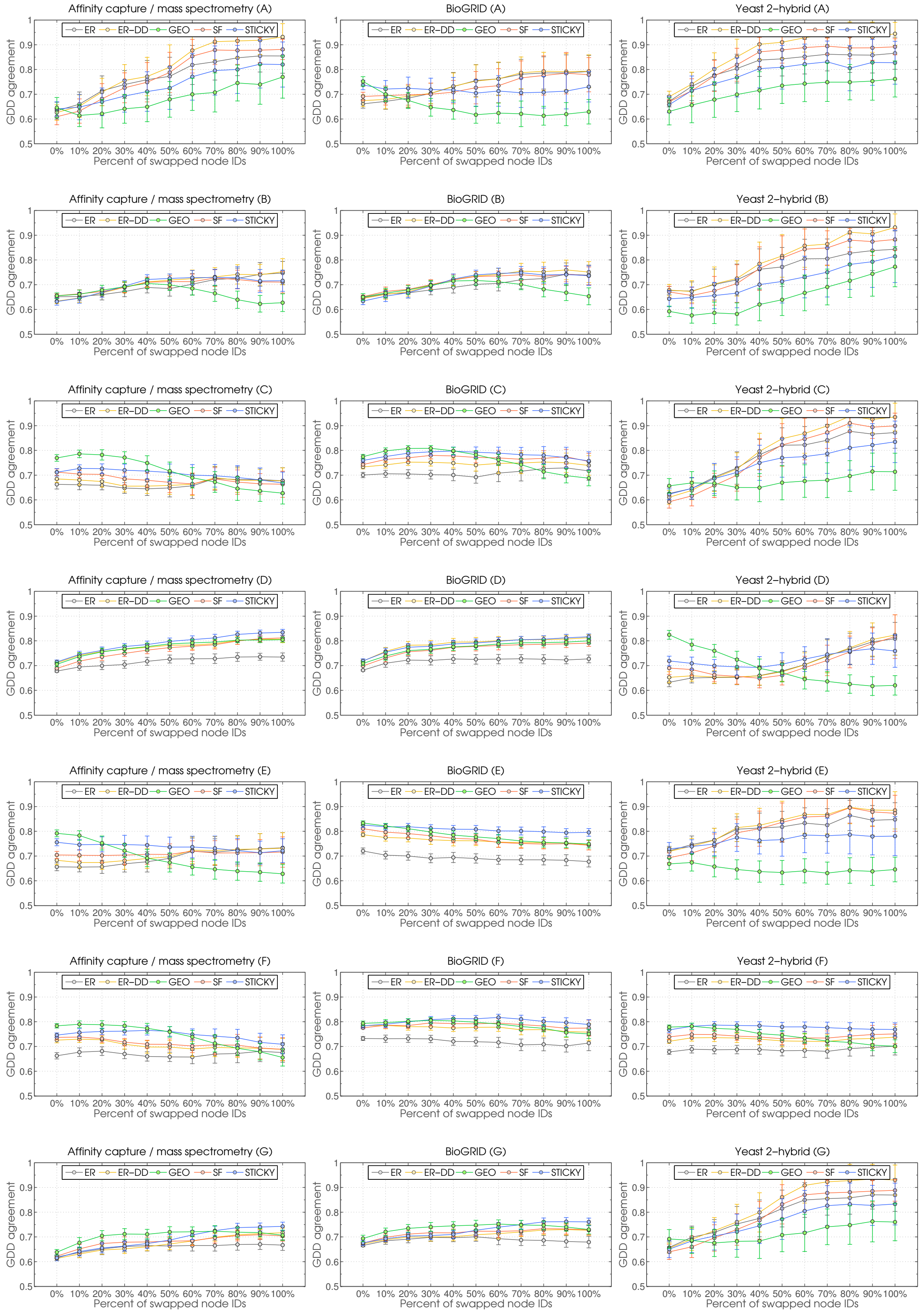
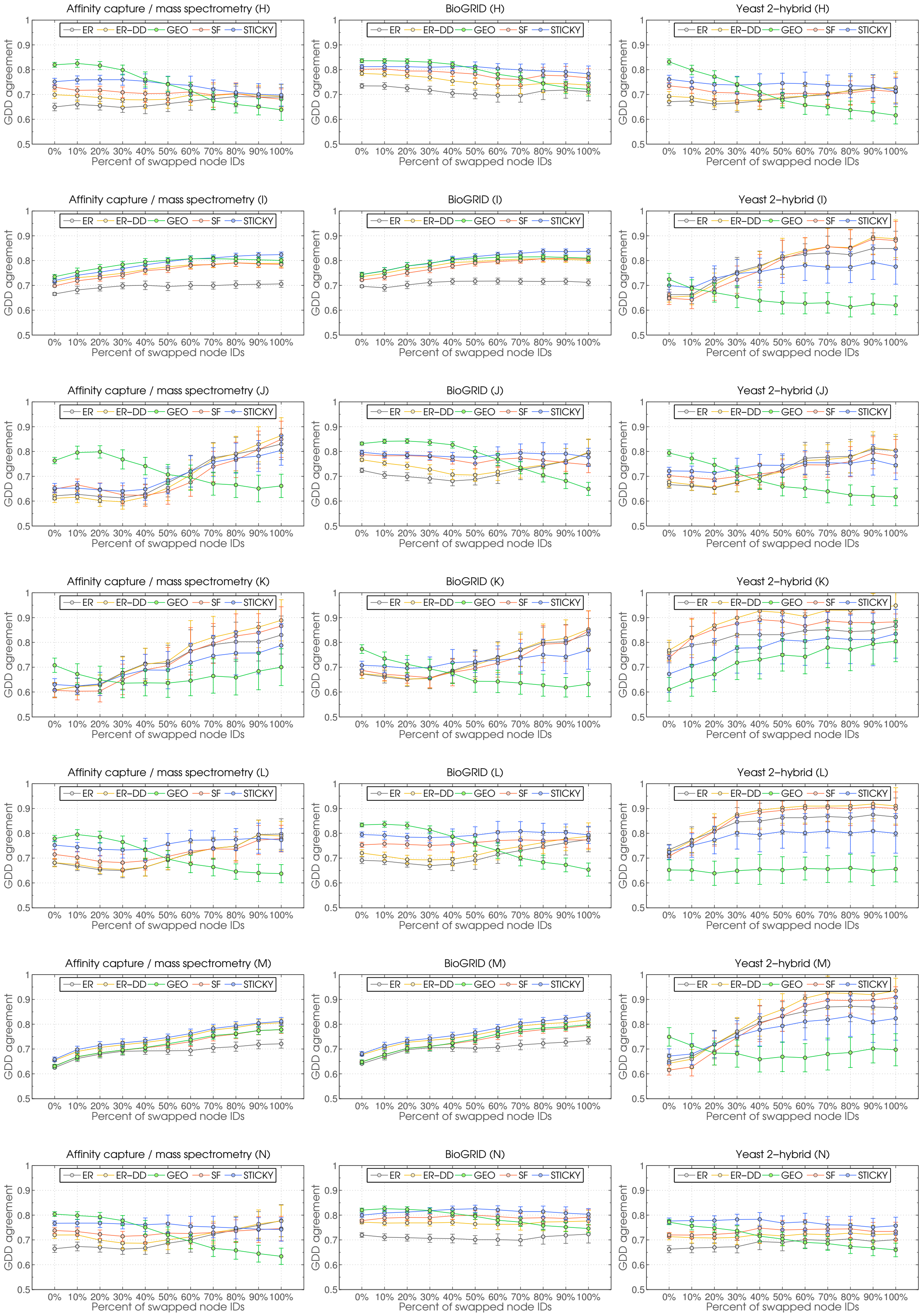


Figure SF3: (Continues on next page.)



Affinity capture / mass spectrometry (H)

BioGRID (H)

Yeast 2–hybrid (H)

Affinity capture / mass spectrometry (I)

BioGRID (I)

Yeast 2–hybrid (I)

Affinity capture / mass spectrometry (J)

BioGRID (J)

Yeast 2–hybrid (J)

Affinity capture / mass spectrometry (K)

BioGRID (K)

Yeast 2–hybrid (K)

Affinity capture / mass spectrometry (L)

BioGRID (L)

Yeast 2–hybrid (L)

Affinity capture / mass spectrometry (M)

BioGRID (M)

Yeast 2–hybrid (M)

Affinity capture / mass spectrometry (N)

BioGRID (N)

Yeast 2–hybrid (N)

Figure SF3: (Caption on next page.)

Figure SF3: (Previous page.) **Robustness of sub-modules' random modelling approach.** Functional sub-modules "A" through "N". The x-axis contains the percentage of re-labelled nodes (0%–100% in increments of 10%, where 0% corresponds to the original network) and y-axis contains the GDDA of the resulting network. The letter in brackets in the title of each plot is the functional sub-module being modelled. Node re-labelling enables the networks to preserve *all* topological properties, thus effectively testing the robustness of the approach. The results are consistent across the three data sets (Affinity capture / mass spec. in the left hand column, BioGRID in the middle, and yeast two-hybrid in the right hand column): the geometricity of the functional sub-modules decreases (GEO model) as the randomness increases (ER and ER-DD); this is more apparent on sub-modules that have sufficient nodes and edges to be outside of the "region of instability" and be modelled with confidence (e.g., BioGRID sub-modules C, G, H, J, K, L, N; Affinity capture / mass spec sub-modules C, E, F, G, H, J, K, L, N; Yeast two-hybrid sub-modules C, D, F, H, I, J, M, N).

References

[1] Doyle, J. C. *et al.* The “robust yet fragile” nature of the Internet. *Proc. Natl. Acad. Sci. U.S.A* **102**, 14497–14502 (2005).

[2] Li, L., Alderson, D., Doyle, J. C. & Willinger, W. Towards a theory of scale-free graphs: Definition, properties, and implications. *Internet Mathematics* **2**, 431–523 (2005).

[3] Pržulj, N., Corneil, D. G. & Jurisica, I. Modeling interactome: Scale-free or geometric? *Bioinformatics* **20**, 3508–3515 (2004).

[4] Pržulj, N. Biological network comparison using graphlet degree distribution. *Bioinformatics* **23**, e177–e183 (2007).

[5] Reguly, T. *et al.* Comprehensive curation and analysis of global interaction networks in Saccharomyces cerevisiae. *Journal of Biology* **5**, 11 (2006).