Supplementary Material:
How threshold behaviour affects the use of subgraphs for network comparison

Tiago Rito\textsuperscript{1,2}, Zi Wang\textsuperscript{1}, Charlotte Deane\textsuperscript{1,2} and Gesine Reinert\textsuperscript{1,2}

\textsuperscript{1}Department of Statistics, University of Oxford, OX1 3TG, UK
\textsuperscript{2}OCISB/DTC Systems Biology, Department of Biochemistry, South Parks Road, OX1 3QU, UK

Email: tiago@stats.ox.ac.uk

1 Model \textit{versus} model comparisons using other subgraph-based scores

1.1 Relative Graphlet Frequency (RGF) distance

The relative graphlet frequency is given by $N_i(G)/T(G)$, where $N_i$ is the graphlet counts with $i \in \{1, \ldots, 29\}$ being the graphlet type/number (see Figure 1 in the main article) and $T(G) = \sum_{i=1}^{29} N_i(G)$ [2]. The RGF-distance between two graphs $G$ and $H$ is then defined as

$$\text{RGF-distance} = \sum_{i=1}^{29} |F_i(G) - F_i(H)|,$$

where $F_i(G) = -\log(N_i(G)/T(G))$.

Figure 1 shows the dependency of this score on the number of vertices and graph density of a network in same model comparisons. Although this score does not rely on automorphism orbits like the GDDA-based scores, we still see a very similar dependency in both cases. Note that this score is not an agreement, but a distance.

1.2 GDDA using the geometric mean

GraphCrunch [1] also calculates a Graphlet Degree Distribution Agreement (GDDA) by taking the geometric mean instead of the arithmetic mean.

*to whom correspondence should be addressed
Figure 1: RGF-distance dependence on the number of vertices and edges of a network in same model versus model comparisons. Average agreements of ER versus ER (A) and GEO-3D versus GEO-3D (B) graphs with 500, 1000 and 2000 vertices are plotted against graph density. Each value represents the average agreement of 50 networks. The graph density of the PPI networks considered (see Table 1 in the main article) is indicated in the top x axis. In (A), the thresholds for the appearance of graphlets for an ER graph with 500 and 1000 nodes are pointed out along the x axis. In (B), the thresholds for the appearance of 3-node graphlets are indicated for the GEO3D graphs with 500, 1000 and 2000 nodes; although error bars are not statistically informative, they were included to give a sense of the variability present in the GDDA values considered.
That is, with $D^j(G, H)$ defined in the main article as being the standardized Euclidean distance between the two scaled and normalized graphlet count vectors the networks $G$ and $H$ for a specific automorphism orbit $j$,

$$\text{GDDA (geometric mean)} = \left( \prod_{j=0}^{72} (1 - D^j(G, H)) \right)^{1/73}. $$

Again, the GDDA score using the geometric mean depends on the number of vertices and on the number of edges in the network. Figure 2 depicts same model versus model comparisons.

2 Model versus model comparisons for higher graph densities using GDDA

Model versus model comparisons for GEO3D and ER-DD were done for higher graph densities - up to 0.4 for GEO3D and 0.05 for ER (Figure 3). These plots suggest that for higher graph densities the score is more stable and less sensitive to the appearance of subgraphs.

3 Assessing Model Fit

To assess how well a random model network fits a given query network, we employ a protocol which consists in the comparison of two samples using two non-parametrics tests: a Monte Carlo test and a Wilcoxon Rank-Sum test. Both of these tests apply to two samples and test the null hypothesis that the samples come from the same distribution. The Monte Carlo test uses as alternative that the samples come from different distributions, whereas the Wilcoxon Rank-Sum test uses as alternative that the two different distributions are shifted versions of each other, with a non-zero shift.

For the Monte Carlo test, for our query network we generate $N$ random graphs from the given model, so that the number of edges and vertices are within 1% of the corresponding numbers for the query network. We calculate the GDDA scores between the query network and each of these $N$ random graphs and take the average over these GDDA agreements, resulting in one number; call it $S_0$. We then generate $M$ random graphs from the given model and use each of these as pseudo-query input graph; repeating the above procedure for each of these $M$ random graphs, we hence obtain $M$ averages over $N$ GDDA scores; call these $S_1, \ldots, S_{99}$. Now we order the vector $(S_0, \ldots, S_{99})$ in increasing order, $S_0 \leq S_1 \leq \ldots \leq S_{99}$. If $S_0$ is the $k^{th}$ of these numbers, then the $p$-value of the test is $k/M$. Thus, if $N = 99$ and if $S_0$ is smaller than all of $S_1, \ldots, S_{99}$, then the $p$-value is 0.01. With $N + 1 = 100$ observations, this is the lowest possible $p$-value for the test.
Figure 2: GDDA (geometric mean) dependence on the number of vertices and edges of a network in same model versus model comparisons. Average agreements of ER versus ER (A) and GEO-3D versus GEO-3D (B) graphs with 500, 1000 and 2000 vertices are plotted against graph density. Each value represents the average agreement of 50 networks. The graph density of the PPI networks considered (see Table 1 of the main article) is indicated in the top x axis. In (A), the thresholds for the appearance of the several 3-5 nodes graphlets for an ER graph with 500 and 1000 nodes are pointed out along the x axis. In (B), the thresholds for the appearance of 3-node graphlets are indicated for the GEO graphs with 500, 1000 and 2000 nodes; although error bars are not statistically informative, they were included to give a sense of the variability present in the GDDA values considered.
Figure 3: GDDA dependence on the number of vertices and edges of a network in same model versus model comparisons for higher graph densities. Average agreements of ER versus ER (A) and GEO3D versus GEO3D (B) graphs with 500 vertices are plotted against graph densities up to 0.05 for ER and 0.4 for GEO3D. Each value represents the average agreement of 50 networks. The graph density region of the PPI networks considered (see Table 1 of the main article) is indicated in the top $x$ axis. In (A), the thresholds for the appearance of the several 3- to 5-node graphlets for an ER graph with 500 nodes are pointed out along the $x$ axis. In (B), the thresholds for the appearance of 3, 4 and 5-node graphlets are indicated for the GEO3D graphs with 500 nodes.
Table 1: Assessing Model Fit: *p*-values obtained by employing Monte Carlo and Wilcoxon rank-sum tests for all PPI considered against GEO3D and ER-DD random graph models. The corresponding histograms can be found in Figures 4 and 5 in the reference letter for Monte Carlo and Wicoxon rank-sum tests respectively.

| Model   | Query   | Monte Carlo | Wilcoxon   | Reference letter |
|---------|---------|-------------|------------|-----------------|
| GEO3D   | YHC     | 0.01        | 6.68 × 10^{-66} | A               |
| ER-DD   | YHC     | 0.01        | 6.68 × 10^{-66} | B               |
| GEO3D   | YIC     | 0.01        | 6.81 × 10^{-66} | C               |
| ER-DD   | YIC     | 0.01        | 1.31 × 10^{-64} | D               |
| GEO3D   | HS      | 0.01        | 6.68 × 10^{-66} | E               |
| ER-DD   | HS      | 0.01        | 5.25 × 10^{-65} | F               |
| GEO3D   | HG      | 0.01        | 6.68 × 10^{-66} | G               |
| ER-DD   | HG      | 0.01        | 8.15 × 10^{-66} | H               |
| GEO3D   | BG-MS   | 0.01        | 6.68 × 10^{-66} | I               |
| ER-DD   | BG-MS   | 0.01        | 7.09 × 10^{-66} | J               |
| GEO3D   | BG-Y2H  | 0.01        | 6.68 × 10^{-66} | K               |
| ER-DD   | BG-Y2H  | 0.01        | 6.68 × 10^{-66} | L               |

For a more powerful test we employ the non-parametric Wilcoxon Rank-Sum test. Here we directly compare the GDDA scores, without averaging. For the query network, we generate *N* random graphs as before and store the resulting GDDA scores as an *N*-vector, we call this *Sample B*. *Sample A*, the result of *M* graphs from a model each compared with other *N* graphs from the same model, can comprise all the resulting *M* × *N* GDDAs. The two samples are then used to perform the test [5].

The *p*-values obtained upon comparison of the PPI considered with GEO3D and ER-DD random graph models are shown in Table 1. Histograms of the 99 averages of GDDA against the observed value can be seen in Figure 4 for several particular comparisons. The histograms of the samples used in the Wilcoxon Rank-Sum test are shown in Figure 5. The optimal bin size was calculated using the Matlab function sshist [4].

For the PPI networks YHC and BG-MS we also tested the STICKY model [3]. This model assumes that the probability of interaction increases with the expected degree of the vertices. Our results are the same as for ER, ER-DD and GEO comparisons: the *p*-values are 0.01 for the Monte Carlo test and 6.68 × 10^{-66} for the Wilcoxon test. The histograms of the samples used in the Wilcoxon Rank-Sum test are shown in Figure 6.

References

[1] T. Milenkovic, J. Lai, and N. Przulj. Graphcrunch: a tool for large network analyses. *BMC Bioinformatics*, 9:70–81, 2008.
Figure 4: Normalized histograms of average GDDA values. The *white* bar represents the observation in the Monte Carlo test, *i.e.* the average over N comparisons with the query network.
Figure 5: Normalized histograms of GDDA values. Histograms of GDDA values between a PPI network (see Table 1) and 99 graphs of a model network are represented by the bars in white. Histograms of GDDA values between graphs of the same model network, 99x99, are shown in gray.
Figure 6: Normalized histograms of GDDA values between the PPI networks BG-MS and YHC, and 99 graphs of the STICKY model are represented by the bars in white. Histograms of GDDA values between graphs of the same model network, 99x99, are shown in gray, i.e. STICKY versus STICKY comparisons.

[2] N. Przulj, D. G. Corneil, and I. Jurisica. Modeling interactome: scale-free or geometric? *Bioinformatics*, 20(18):3508–3515, Dec 12 2004.

[3] N. Pržulj and D.J. Higham. Modelling protein–protein interaction networks via a stickiness index. *Journal of the Royal Society Interface*, 3(10):711, 2006.

[4] H. Shimazaki and S. Shinomoto. A method for selecting the bin size of a time histogram. *Neural computation*, 19(6):1503–1527, 2007.

[5] G. Upton and I. Cook. *Understanding statistics*. Oxford University Press, 1996.