Bioinformatics
doi.10.1093/bioinformatics/xxxxxx
Advance Access Publication Date: Day Month Year
Manuscript Category

OXFORD

SYSTEMS

Supplementary material for: Functional geometry of protein interactomes

Noël Malod-Dognin[1] and Nataša Pržulj[1,2,*]

[1] Department of Life Sciences, Barcelona Supercomputing Center, 08034 Barcelona, Spain
[2] ICREA, Pg. Lluís Companys 23, 08010 Barcelona, Spain

*To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

1 Generating random simplicial complexes

In the main document, we validate our approaches using randomly generated simplicial complexes, which we generate according to the ten random simplicical complex models described bellow. The first five models are based on randomly generated graphs, which are converted into so-called *clique complexes*, in which simplices connect nodes that belong to a clique in the graph:

- A *random clique complex* (RCC) is the clique complex of an Erdös-Rènyi random graph (Erdös and Rényi, 1959). We generate an Erdös-Rènyi graph by fixing the number of nodes in the graph, and then by adding edges between uniformly randomly chosen pairs of nodes until a given edge density is reached.
- A *Vietoris-Rips complex* (VRC) (Hausmann *et al.*, 1995) is the clique complex of a geometric random graph (Penrose, 2003). A geometric random graph represents the proximity relationship between uniformly randomly distributed points in a d-dimensional space. We generate geometric graphs by uniformly randomly distributing the desired number of nodes (points) in a 3-dimensional unit cube. Then, two nodes are connected by an edge if the Euclidean distance between the corresponding points is smaller than a distance threshold r. The distance threshold is chosen to obtain the desired edge density.
- A *scale-free complex* (SFC) is the clique complex of a Barabàsi-Albert scale-free graph (Barabási and Albert, 1999). The scale-free graph model constructed by preferential attachment generates graphs based on the "rich-get-richer" principle and are characterized by power-law degree distributions. We create a scale-free graph using an iterative process, in which the graph is grown by attaching new nodes each with m edges that are preferentially attached to the existing nodes with high degree (m is chosen to obtain the desired edge density).
- A *Watts-Strogatz complex* (WCS) is the clique complex of a small-world graph (Watts and Strogatz, 1998). Small-world graphs are characterized by short average path lengths and high clustering. We create a small word graph by constructing a regular ring lattice of n nodes and by connecting each node to its k neighbours, $k/2$ on each side (k is chosen to obtain the desired edge density). Then we uniformly randomly rewire 5% of the edges.
- An *nPSO complex* (nPSOC) is the clique complex of a non-uniform Popularity Similarity Optimization graph (Muscoloni and Cannistraci, 2018). Non-uniform Popularity Similarity Optimization graphs are geometric graphs in hyperbolic space that have realistic features, such as high clustering, small-worldness, scale-freeness and rich-clubness, with the additional possibility to control the community organization. We create a small word graph by constructing a regular ring lattice of n nodes and by connecting each node to its k neighbours, $k/2$ on each side (k is chosen to obtain the desired edge density), We use the graph generator from Muscoloni and Cannistraci (2018), in which the temperature parameter is set to 0.5, gamma is set to 0.3, and the number of communities is set to 50.

The five other models are extensions of the *Linial-Meshulam* model (Linial and Meshulam, 2006; Meshulam and Wallach, 2009), which originally consists in randomly connecting nodes with k-dimensional facets. We extended this model to randomly connect nodes with facets while following the facet distribution of an input simplicial complex. In this way, we can create Linial-Meshulam variant of the four clique complex-based models presented above:

- A *Linial-Meshulam random clique complex* (LM- RCC) is a Linial-Meshulam complex that follows the facet distribution of an input random clique complex.
- A *Linial-Meshulam Vietoris-Rips complex* (LM- VRC) is a Linial-Meshulam complex that follows the facet distribution of an input Vietoris-Rips complex.
- A *Linial-Meshulam scale-free complex* (LM-SFC) is a Linial-Meshulam complex that follows the facet distribution of an input scale-free complex.
- A *Linial-Meshulam Watts-Strogatz complex* (LM- WSC) is a Linial-Meshulam complex that follows the facet distribution of an input Watts-Strogatz complex.

- A *Linial-Meshulam nPSO complex* (LM-nPSOC) is a Linial-Meshulam complex that follows the facet distribution of an input nPSO complex.

For each model we choose three node sizes, 1,000, 2,000, and 3,000 nodes, and three edge densities, 0.5%, 0.75% and 1%. We generated 25 random simplicial complexes for each model and each of these node sizes and edge densities. Hence, in total, we generated $10 \times 3 \times 3 \times 25 = 2{,}250$ random simplicial complexes. We chose these node sizes and edge densities to roughtly mimic the sizes and densities of real-world data used in the main paper.

2 Redundancies between the counts of simplet degrees

Analogous to graphlets, the statistics of different simplet orbits are not independent of each other. The reason behind this is the fact that smaller simplets are induced sub-simplicial complexes of larger simplets. For 2- to 4-node simplets, there are four non-redundant dependency equations between the simplet degrees of a given node u:

$$\binom{u_1}{2} = u_2 + u_4 + u_5, \tag{1}$$

$$\binom{u_2}{1}\binom{u_1 - 2}{1} = 3u_9 + 2u_{11} + 2u_{14} + u_{18} + u_{20} + u_{23}, \tag{2}$$

$$\binom{u_3}{1}\binom{u_1 - 1}{1} = \begin{array}{l} u_7 + u_{12} + u_{15} + 2u_{16} + 2u_{17} \\ +2u_{19} + 2u_{21} + 2u_{22} \end{array}, \tag{3}$$

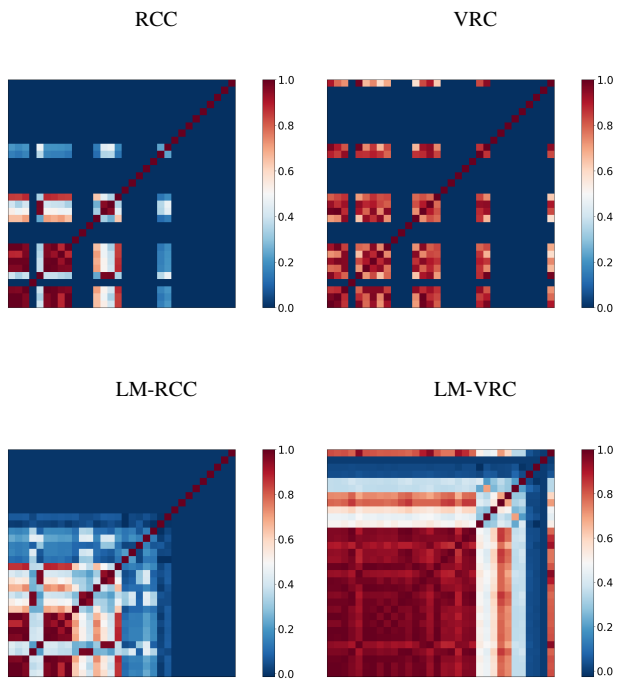
$$\binom{u_1}{3} = \begin{array}{l} u_9 + u_{11} + u_{14} + u_{18} + u_{20} + u_{23} + u_{24} + u_{25} \\ +u_{26} + u_{27} + u_{28} + u_{29} + u_{30} + u_{31} + u_{32} \end{array}. \tag{4}$$

We used these equations to assess the correctness of our exhaustive simplet counter.

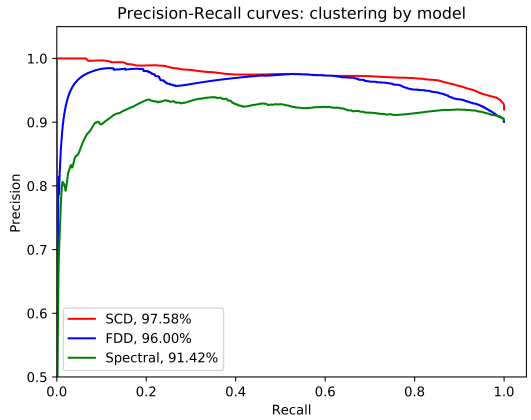
3 Supplementary figures and tables

Orbit, i	Weight, o_i
1	1
2, 3, 4, 5	3
6, 8, 9, 10, 13, 24, 26, 30, 31, 32	3
7, 11, 12, 14, 15, 16, 17, 18, 19, 21, 22, 23, 25, 27, 28, 29	4
20	5

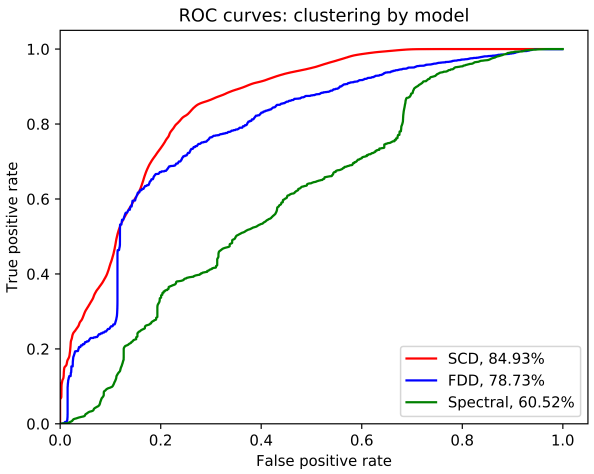
Table 1. The orbit weights. Weight o_i is the number of orbits that orbit i depends on, including itself. For instance, the count of orbit 2 (the middle of a three node path) of a node depends on its count of orbit 0 (i.e. its node degree) and on itself, so $o_2 = 2$. For orbit 9, o_9 = 3, since it is affected by orbits 0, 2, and itself.



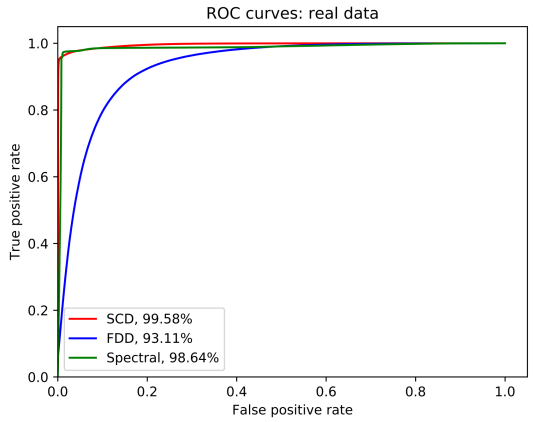
Supplementary Fig. 1: **SCMs of sample simplicial complexes from four random simplicial complex models.** The four simplicial complexes from RCC, VRC, LM-RCC, and LM-VRC models have been generated with 2,000 nodes and the edge density of 0.75%. Note that SCMs of all networks of this size and density coming from a particular model look similar. Hence, these four SCMs are representative of these models at these sizes and densities.



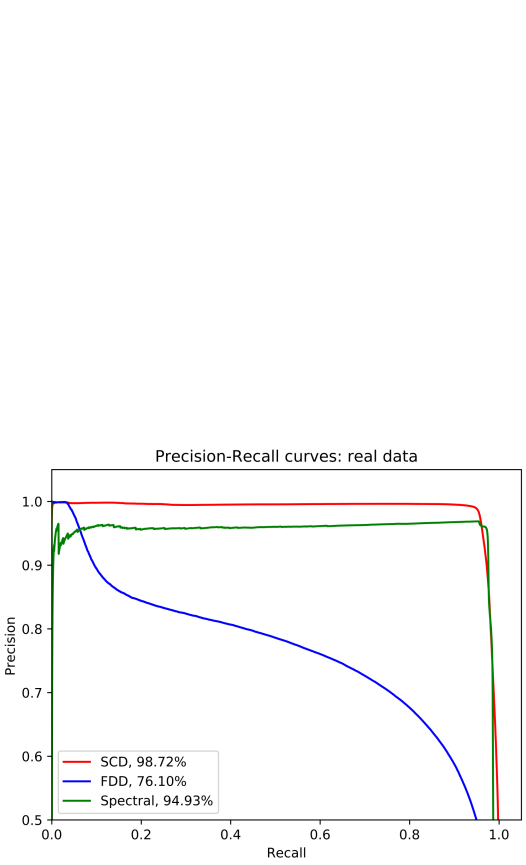
Supplementary Fig. 2: **Clustering randomly generated simplicial complexes.** The Precision-Recall curves that are achieved when using the three distance measures (color coded, simplet correlation distance in red, facet distribution distance in blue, spectral distance in green) to cluster together the 2,250 randomly generated simplicial complexes into the models that generated them.



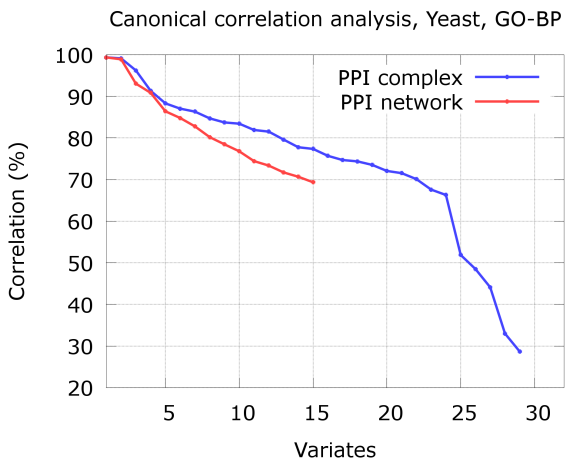
Supplementary Fig. 3: **Clustering randomly generated simplicial complexes.** The ROC curves that are achieved when using the different distance measures (color coded, simplet correlation distance in red, facet distribution distance in blue, spectral distance in green) to cluster together the 2,250 randomly generated simplicial complexes into the models that generated them.



Supplementary Fig. 5: **Clustering real-world simplicial complexes.** The ROC curves that are achieved when using the three distance measures (color coded, simplet correlation distance in red, facet distribution distance in blue, spectral distance in green) to cluster together the 1,775 real-world simplicial complexes according to their data types.



Supplementary Fig. 4: **Clustering real-world simplicial complexes.** The Precision-Recall curves that are achieved when using the three distance measures (color coded, simplet correlation distance in red, facet distribution distance in blue, spectral distance in green) to cluster together the 1,775 real-world simplicial complexes according to their data types.



Supplementary Fig. 6: **Canonical correlation analysis for yeast.** For both models of yeast interactomes (PPI network and PPI complex), we plotted for each variate the corresponding correlation value (only statistically significantly correlated variates are presented, with canonical correlation p-value $\leq 5\%$).

References

Barabási, A.-L. and Albert, R. (1999). Emergence of scaling in random networks. *Science*, **286**(5439), 509–512.

Erdös, P. and Rényi, A. (1959). On random graphs. *Publicationes Mathematicae*, **6**, 290–297.

Hausmann, J.-C. *et al.* (1995). On the vietoris-rips complexes and a cohomology theory for metric spaces. *Annals of Mathematics Studies*, **138**, 175–188.

Linial, N. and Meshulam, R. (2006). Homological connectivity of random 2-complexes. *Combinatorica*, **26**(4), 475–487.

Meshulam, R. and Wallach, N. (2009). Homological connectivity of random k-dimensional complexes. *Random Structures & Algorithms*, **34**(3), 408–417.

Muscoloni, A. and Cannistraci, C. V. (2018). A nonuniform popularity-similarity optimization (npso) model to efficiently generate realistic complex networks with communities. *New Journal of Physics*, **20**(5), 052002.

Penrose, M. (2003). Random geometric graphs. *Oxford Studies in Probability*, **5**.
Watts, D. J. and Strogatz, S. H. (1998). Collective dynamics of "small-world" networks. *nature*, **393**(6684), 440.