Topological Fractal Dimension of Networks of Protein–Protein Interaction Networks

Georgios Kalantzis[a] **and Andrei Stoica**[a]

[a]SABS, DTC

This manuscript was compiled on November 1, 2018

Please provide an abstract of no more than 250 words in a single paragraph. Abstracts should explain to the general reader the major contributions of the article. References in the abstract must be cited in full within the abstract itself and cited in the text.

PPIN | NetworkX | BioGrid | Network Science | Centrality Measures

1. Introduction & Prerequisites

Networks are representations of real systems where individual units are modelled as nodes and interactions between these units as links. Formally speaking, this corresponds to a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, with $\mathcal{V}$ and $\mathcal{E}$ standing for the set of nodes and edges respectively. However, nodes can be anything, ranging from regions of the human brain to electrical power plants. As a result, the study of networks pervades all of science, from neurobiology to statistical physics (1).

The set of nodes $\mathcal{V}$ has usually a finite number of elements. Links can be undirected or directed, unweighted or weighted. When dealing with undirected edges, a link can be defined as a pair of nodes (u, v); in directed graphs (u, v) and (v, u) correspond to different edges. In the case of weighted networks, links are also assigned with a real number characterising the importance of the association.

There are two main ways for representing networks, namely lists and adjacency matrices. Let $\mathbf{A}$ be the adjacency matrix of graph $\mathcal{G}$ containing n nodes. Then $\mathbf{A}$ is of size $n \times n$ and element A_{ij} is 1 if nodes i, j are connected, otherwise, $A_{ij} = 0$. In weighted networks non-zero elements of $\mathbf{A}$ are equal to the weights of edges. Moreover, real complex networks, although might contain thousands of nodes, are usually sparse in regards of edges and can be represented by sparse matrices which are special data structures.

After defining nodes and edges, the next important term is that of node-degree. The degree k_i of node i is defined as the number of edges linked to i. In directed graphs, the degree of a node might be discriminated in in-degree and out-degree, depending on whether the edges ending to or start from i are counted. For undirected networks, the degree can be computed by

$$k_i = \sum_{j=1}^{n} A_{ij} = \mathbf{e}_i^{\top}(\mathbf{A} \cdot \mathbf{e}),$$

where $\mathbf{e}$ stands for a column-vector full of ones and $\mathbf{e}_i$ has zeros everywhere except element i which is one. In other words, the product $\mathbf{A} \cdot \mathbf{e}$ gives the degree for every node. These are some simple indicators showing why adjacency matrices are important: they connect Network Science with Linear Algebra.

Graphs are very attractive tools to biological and medical research applications, since they can use for description of many mechanisms or interactions, for instance metabolic or cell signalling networks. Another important direction are the so-called Protein-Protein Interaction Networks (which will be denoted as PPIN), which represent physical contacts between proteins within a cell. In brief, proteins are macromolecules, consisting of one or more long chains of amino acid residues, which perform a vast array of functions within organisms, including catalysing metabolic reactions,

Significance Statement

Authors must submit a 120-word maximum statement about the significance of their research paper written at a level understandable to an undergraduate educated scientist outside their field of speciality. The primary goal of the Significance Statement is to explain the relevance of the work in broad context to a broad readership. The Significance Statement appears in the paper itself and is required for all research papers.

DNA replication, responding to stimuli, providing structure to cells and organisms, and transporting molecules from one location to another. Usually, the aforementioned procedures incorporate the cascade of many proteins, resulting in networks of interactions.

PPINs are characterised by three notable properties. First of all, PPINs show a small world effect meaning that there is great connectivity between proteins. More typically, the diameter (the maximum number of steps separating any two nodes) of such networks is small, regardless the number of nodes or edges. Such strong connectivity has important biological consequences, since it allows for an efficient and quick flow of signals within the network (2).

Moreover, PPINs are scale-free. This class of networks can describe a variety of complex systems where some nodes have a tremendous number edges (hubs), whereas most nodes have only a few (3). In this sense, the network appears to have no scale which provides important features. For instance, scale-free networks are very robust and stable since small perturbations have low effect. Furthermore, hubs in cancer-linked networks could be used for targeted attacks in drug discovery.

Finally, another crucial characteristic of PPINs is their modularity. The transitivity or clustering coefficient of a network is a measure of the tendency of the nodes to cluster together. High transitivity means that the network contains communities or groups of nodes that are densely connected internally. Generally speaking, structure always affects function (1). In biological networks particularly, finding these communities is very important, because they can reflect functional modules and protein complexes.

However, PPINs are not real but correspond to actual biological networks and occur after experimental procedures. As a result, data might contain noise and some observations can be less reliable since the record of molecular interactions is occasionally incomplete or patchy.

2. Construction of PPIN and Analysis

There are various software packages or programmatic methods available to build and analyse networks. Throughout this project we work mainly with Python and the NetworkX module. Furthermore, there are a lot of sources from which you can obtain and integrate PPI data, other than creating new experimental results. Current databases are distinguished in primary or predicting, depending on whether they just provide evidence or combine other resources for prediction. For the aims of this project we will be working with data from BioGrid, which is a primary curated biological database of protein-protein interactions, genetic interactions, chemical interactions, and post-translational modifications.

A. Hands-on BioGrid & NetworkX. PPINs can be created from edge-lists, which are text files containing rows of the form "$\mathrm{ID}_A - \mathrm{ID}_B$". The latter correspond to distinctive labels of nodes which can also be used as reference to other information sources. After loading an edge-file, NetworkX can initialise a graph with the provided edges.

B. Analysis & Illustration of PPIN.

C. Centrality & Important Proteins. By using PageRank, HITS and Degree Centrality we find that the five most "central" nodes are

Table 1. Top-5 Central Proteins in *Homo Sapiens*

Node Label	Protein	Gene	Function
114030	Cullin-3	CUL3	This protein plays a critical role in the polyubiquitination and subsequent degradation of specific protein substrates
113164	HMG20	UBC	It plays a key role in maintaining cellular ubiquitin levels under stress. Defects could lead to embryonic lethality.
108309	HuR	ELAVL1	RNA-binding protein that binds to the 3'-UTR region of mRNAs and increases their stability
113348	Exportin-1	XPO1	eukaryotic protein that mediates the nuclear export of proteins, rRNA, snRNA, and some mRNA.
113010	TP53	TP53	tumor suppressor protein containing transcriptional activation, DNA binding, and oligomerization domains.

D. Opt: Computational Complexity.

E. Opt: The Development of PPINs.

3. Computing the Topological Fractal Dimension

A. The Box Counting Method.

B. Implementation.

C. Computational Complexity.

4. TFD of PPIN & Conclusions

A. Dynamic changes of TFD & Interaction with other Databases.

B. Conclusions.

Manuscript Length. The maximum length of a Direct Submission research article is six pages and a Direct Submission Plus research article is ten pages including all text, spaces, and the number of characters displaced by figures, tables, and equations. When submitting tables, figures, and/or equations in addition to text, keep the text for your manuscript under 39,000 characters (including spaces) for Direct Submissions and 72,000 characters (including spaces) for Direct Submission Plus.

Data Archival. PNAS must be able to archive the data essential to a published article. Where such archiving is not possible, deposition of data in public databases, such as GenBank, ArrayExpress, Protein Data Bank, Unidata, and others outlined in the Information for Authors, is acceptable.

Language-Editing Services. Prior to submission, authors who believe their manuscripts would benefit from professional editing are encouraged to use a language-editing service (see list at www.pnas.org/site/authors/language-editing.xhtml). PNAS does not take responsibility for or endorse these services, and their use has no bearing on acceptance of a manuscript for publication.

Supporting Information (SI). Authors should submit SI as a single separate PDF file, combining all text, figures, tables, movie legends, and SI references. PNAS will publish SI uncomposed, as the authors have provided it. Additional details can be found here: policy on SI. For SI formatting instructions click here. The PNAS Overleaf SI template can be found here. Refer to the SI Appendix in the manuscript at an appropriate point in the text. Number supporting figures and tables starting with S1, S2, etc.

Authors who place detailed materials and methods in an SI Appendix must provide sufficient detail in the main text methods to enable a reader to follow the logic of the procedures and results and also must reference the SI methods. If a paper is fundamentally a study of a new method or technique, then the methods must be described completely in the main text.

ACKNOWLEDGMENTS. Please include your acknowledgments here, set in a single paragraph. Please do not include any acknowledgments in the Supporting Information, or anywhere else in the manuscript.

Table 2. Basic Information about PPINs from BioGrid 3.5.165 dataset

Organism	# Nodes	# Edges	Avg Degree	ConComps	Largest
Anopheles gambiae PEST	2	1	1.00	1	2
Apis mellifera	2	1	1.00	1	2
Arabidopsis thaliana Columbia	9626	36091	7.50	134	9389
Bacillus subtilis 168	3	2	1.33	2	2
Bos taurus	439	412	1.88	75	71
Caenorhabditis elegans	3954	8051	4.07	93	3753
*Candida albicans SC*5314	726	886	2.44	43	637
Canis familiaris	53	35	1.32	21	7
Cavia porcellus	9	5	1.11	4	3
Chlamydomonas reinhardtii	19	16	1.68	4	12
Chlorocebus sabaeus	11	7	1.27	4	3
Cricetulus griseus	32	24	1.50	8	16
Danio rerio	247	255	2.06	39	99
*Dictyostelium discoideum AX*4	24	20	1.67	6	5
Drosophila melanogaster	9197	55350	12.04	43	9113
*Emericella nidulans FGSC A*4	64	62	1.94	6	45
Equus caballus	4	2	1.00	2	2
*Escherichia coli K*12	2	1	1.00	1	2
*Escherichia coli K*12 *MC*4100 *BW*2952	10	9	1.80	2	8
*Escherichia coli K*12 *MG*1655	150	133	1.77	25	91
*Escherichia coli K*12 *W*3110	4063	181621	89.40	1	4063
Gallus gallus	391	421	2.15	42	230
Glycine max	45	40	1.78	8	14
Hepatitus C Virus	131	129	1.97	2	129
Homo sapiens	22840	321550	28.16	28	22798
Human Herpesvirus 1	174	195	2.24	1	174
Human Herpesvirus 2	7	4	1.14	3	3
Human Herpesvirus 3	4	2	1.00	2	2
Human Herpesvirus 4	240	235	1.96	7	185
Human Herpesvirus 5	91	80	1.76	12	35
Human Herpesvirus 6*A*	11	7	1.27	4	4
Human Herpesvirus 6*B*	7	4	1.14	3	3
Human Herpesvirus 7	2	1	1.00	1	2
Human Herpesvirus 8	714	689	1.93	45	378
Human Immunodeficiency Virus 1	1121	1306	2.33	1	1121
Human Immunodeficiency Virus 2	16	12	1.50	4	8
Human papillomavirus 16	14	12	1.71	2	12
Macaca mulatta	15	13	1.73	3	11
Meleagris gallopavo	2	2	2.00	1	2
Mus musculus	13021	38893	5.97	107	12793
*Mycobacterium tuberculosis H*37*Rv*	11	9	1.64	2	9
*Neurospora crassa OR*74*A*	12	10	1.67	2	8
Nicotiana tomentosiformis	2	2	2.00	1	2
Oryctolagus cuniculus	283	278	1.96	33	142
Oryza sativa Japonica	75	94	2.51	19	26
Ovis aries	2	1	1.00	1	2
Pan troglodytes	10	5	1.00	5	2
Pediculus humanus	2	1	1.00	1	2
Plasmodium falciparum 3*D*7	1227	2508	4.09	26	1179
Rattus norvegicus	3718	5282	2.84	123	3407
Ricinus communis	3	2	1.33	1	3
*Saccharomyces cerevisiae S*288*c*	7158	535782	149.70	1	7158
Schizosaccharomyces pombe 972*h*	4318	58739	27.21	33	4279
Selaginella moellendorffii	6	8	2.67	1	6
Simian Immunodeficiency Virus	19	16	1.68	4	8
Simian Virus 40	6	5	1.67	1	6
Solanum lycopersicum	45	109	4.84	7	21
Solanum tuberosum	5	3	1.20	3	2
Strongylocentrotus purpuratus	17	16	1.88	1	17
Sus scrofa	94	79	1.68	23	22
Tobacco Mosaic Virus	3	2	1.33	1	3
Ustilago maydis 521	4	4	2.00	1	4
Vaccinia Virus	8	6	1.50	3	3
Vitis vinifera	2	1	1.00	1	2
Xenopus laevis	1128	1223	2.17	61	959
Zea mays	21	13	1.24	10	3

1. Strogatz SH (2001) Exploring complex networks. *nature* 410(6825):268.
2. Institute EB (year?) Network analysis of protein interaction data: an introduction (https://www.ebi.ac.uk/training/online/course/network-analysis-protein-interaction-data-introduction). Accessed: 2018-10-31.
3. Barabási AL, Bonabeau E (2003) Scale-free networks. *Scientific american* 288(5):60–69.