Geometric Local Structure in Biological Networks

Nataša Pržulj
Computer Science Department, University of California, Irvine, CA 92697-3435
E-mail: natasha@ics.uci.edu

***Abstract*— The recent explosion in biological and other real-world network data has created the need for improved tools for large network analyses. Several new mathematical techniques for analyzing local structural properties of large networks have recently been developed. Our work introduces small induced subgraphs of large networks, called graphlets. We use graphlets to develop "network signatures" that quantify local structural properties of a network. Based on these network signatures, we design two novel "network agreement" measures. These measures lead us to new, well-fitting geometric graph models of biological networks. Models are in turn used to design efficient heuristics.**

I. Introduction

Recent technological advances in experimental biology have yielded large amounts of biological network data. Examples span the range from biochemical networks, such as metabolic [48], transcriptional regulation [58], [26], [13], [35], [14], and protein-protein interaction [10] networks, to protein-structure [8] and neuronal wiring networks [69]. Many other real-world phenomena have also been described in terms of large networks, such as various types of social [43], [22] and technological [18] networks. Thus, understanding these complex phenomena has become an important scientific problem that has opened up a vibrant research area in the analysis and modeling of large, real networks.

Understanding the complex wiring of protein-protein interaction networks is a central problem of systems biology. In a *protein-protein interaction (PPI) network* (or a *graph*), proteins are modeled as *nodes* and two proteins are joined by an undirected *edge* (link) in the network if they can physically interact. Various biotechnologies have already produced large amounts of PPI network data for a number of organisms [62], [27], [19], [25], [21], [37], [20], [34], including human [61], [57], with hundreds of labs throughout the world continuously contributing to this pool of data. However, these networks are still largely incomplete, i.e., false negatives is the dominant type of noise in these data. They also contain false-positives produced by the noisy experimental techniques used to detect PPIs. In this paper we focus on studying PPI networks. However, the same types of analyses apply to other biological networks.

The amount of PPI data is expected to rapidly increase in the near future, paralleling the earlier explosion of genetic sequence data. For example, the PPI network of baker's yeast, *Saccharomyces cerevisiae*, has around $6,000$ nodes and currently over $78,000$ interactions between the proteins have been identified [65], with most of the interactions still being unknown. PPI networks of higher organisms, such as primates or plants, will be much larger. Therefore, understanding the complex wiring in PPI networks relies on advances in computational sciences. Algorithmic and modeling advances in this area will contribute directly to biological understanding and therapeutics.

II. Background: Graph Models and Comparisons

Analogous to genetic sequence comparison, comparing large cellular networks will bring insights into biological understanding, evolution, and disease. However, exhaustively comparing large networks is computationally infeasible, since it involves an NP-complete subgraph isomorphism problem [12]. Subsequently, such comparisons rely on heuristics commonly called *network properties*. Also, models are an important part of our understanding, since our ability to *model* real-world phenomena enables us to *reproduce* them and thus *understand* them. Various network models have been proposed and their fit to the real networks has been evaluated with respect to different network properties.

A. Network Properties

Easily computable macroscopic statistical *global properties* of large networks have extensively been examined. Based on these properties, network models have been proposed for cellular (and other real) networks if their global properties fit the global properties of real networks. The most widely used global network properties are the *degree distribution*, the *clustering coefficient*, the *clustering spectrum*, and the network *average diameter* [44]. The *degree* of a node is the number of edges touching the node and the *degree distribution* is the distribution of degrees of all nodes in the network, or equivalently, the probability that a randomly selected node of a network has degree k (commonly denoted by $P(k)$). Many large real-world networks have non-Poisson degree distributions with a power-law tail, $P(k) \sim k^{-\gamma}, \gamma > 0$; such networks have been termed *scale-free* [5]. However, networks with exactly the same power-law degree distributions can have vastly different structure affecting their function [55], [36]. The same holds for other global network properties [55]. Furthermore, global network properties of largely incomplete cellular networks do not tell us much about the true structure of the real networks; instead, they describe the network structure produced by the experimental sampling techniques used to obtain these networks (e.g. see [24]).

The *clustering coefficient,* C, is defined as the average probability that two neighbors of a given node are adjacent [67]. The *clustering spectrum*, $C(k)$, of a network is the distribution of average clustering coefficients of degree k

1-4244-1564-0/07/$25.00 ©2007 IEEE

nodes. Cellular networks have large clustering coefficients [67]. The *shortest path* between nodes x and y in a network is a path with the smallest number of links that have to be traversed to get from node x to node y [68]. The *average diameter* is defined as the average of shortest path lengths between all pairs of nodes in a network. Despite their large sizes, most cellular networks have small average diameters; this property is often referred to as the *small-world* property [67]. Clearly, the diameter of an incomplete network can be substantially different than that of the real network; this is another example of how global network properties can be misleading.

To overcome the above mentioned problems in modeling cellular networks based on their global properties, bottom-up *local* approaches to studying microscopic network structure have been proposed [40], [59], [55], [54]. Analogous to sequence motifs, *network motifs* have been defined as subgraphs that recur in a network at frequencies much higher than those found in randomized networks [40], [59], [39]. A *subgraph* of a network G with the set of nodes $V(G)$ and the set of edges $E(G)$ is a network whose nodes and edges belong to G. An *induced subgraph* H of G is a subgraph of G on $V(H) \subseteq V(G)$ set of nodes, such that edges $E(H)$ of H consists of *all* edges of G that connect nodes of $V(H)$. We introduced *graphlets* to denote small connected non-isomorphic induced subgraphs of large networks [55] and used their properties to compare large networks [55], [54] (see section III).

The main advantage of the local properties is evident when we study networks with incomplete node and edge sets. While the local structures of these networks are more likely to be complete, the global properties are more likely to be biased. The most studied eucaryotic PPI network is that of the model organism baker's yeast, *S. cerevisiae*, for which we know all the nodes, since we know the genome, but only a small fraction of the edges is currently known and these edges are heavily clustered around proteins important for human disease. Furthermore, the biological experiments used to detect PPIs are of local nature. This is why we base our analysis of PPI networks on a variant of the local approach (see section III). It has been argued that global structural features of networks, such as the clustering coefficient, are intertwined with local structural properties [63]. In section III-B and in [54], we show that the degree distribution, commonly regarded as a global network property, is a part of a more complex new measure of local network structure.

B. Network Models

Various network models have been developed attempting to model real networks. The earliest such models are *Erdos-Renyi ("ER") random graphs* in which the probability p of an edge between any pair of nodes is distributed uniformly at random [15], [16], [17]. Random graphs have served as idealized models of gene networks [31], ecosystems [38], and the spread of infectious diseases [30] and computer viruses [33]. However, they fail to model power-law degree distributions and large clustering coefficients of real networks. Thus, *generalized random graphs ("ER-DD")* in which the edges are randomly chosen as in Erdos-Renyi random graphs, but the degree distribution is constrained to be the degree distribution of a real-world network at study, have been introduced and their properties studied [7], [41], [42], [45], [2]. Since these networks have the same low clustering coefficients as the ER graphs, other network models have been introduced. The most prominent such models are *small-world* networks [67], [46], [47], characterized by small diameters and large clustering coefficients, and *scale-free ("SF")* networks [60], [5], [6], [9], [36], that include an additional condition of scale-freeness of the degree distribution.

The degree distributions of metabolic reaction networks [28], the Internet backbone [18], the telephone call graph [1], the World Wide Web [11] and many other real networks decay as a power law. Thus, many variants of SF network growth models have been proposed. The most notable such models for PPI networks are those based on biologically motivated *gene duplication and mutation* network growth principles [23], [66], [49], [64]. In these models, networks grow by duplication of nodes (genes), and as a node gets duplicated, it inherits most of the neighbors (interactions) of the parent node, but gains some new neighbors as well.

III. Approach: Geometricity of PPI Networks

The focus of our attention is PPI networks. Thus, all networks that we study here have unweighted undirected edges. However, the same types of analyses can be generalized to other real-world networks with undirected or directed weighted edges. To provide meaningful comparisons between the data and model networks, all model networks used here have the same number of nodes and edges as the PPI network that they model.

A. Relative "Graphlet" Frequencies

We made the first step towards a *detailed* network comparison tool by introducing *graphlets*, small connected non-isomorphic induced subgraphs of a large network, and designing a new measure of local structural similarity between two networks based on graphlet frequency distributions [55]. Graphlets do not need to be over-represented in a network and this, along with being induced, distinguishes them from *network motifs* [40], [59]. The number of graphlets increases exponentially with the number of nodes, and thus, we restrict our attention to 2-, 3-, 4-, and 5-node graphlets. All 29 3-5-node graphlets are presented in Figure 1 Left (note that an *edge* is the only 2-node graphlet). Thus, our new measure of network similarity imposes 30 similarity constraints on networks being compared corresponding to the distribution of frequencies of 2-, 3-, 4- and 5-node graphlets. We define the new measure of *distance* between networks G and H as the sum over all 30 graphlets of the absolute values of the difference of logarithms of normalized graphlet frequencies; logarithms are used to even out the influences of frequent and infrequent graphlets onto the distance measure (see [55] for details). This new measure has been used to discover a new,

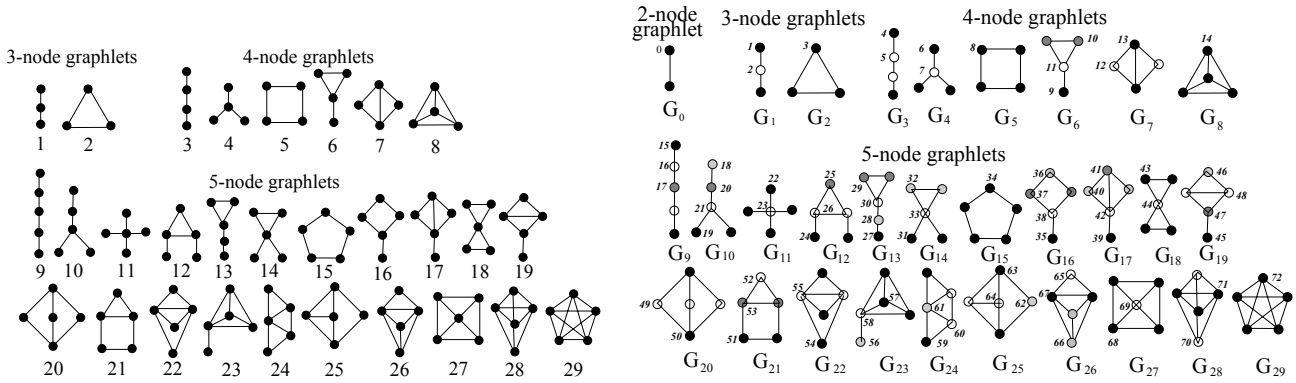

Fig. 1. **Left.** All 3-, 4-, and 5-node graphlets, ordered within groups from the least to the most dense with respect to the number of edges when compared to the maximum possible number of edges in the graphlet and numbered from 1 to 29 [55]. **Right.** Automorphism orbits $0, 1, 2, \ldots, 72$ for the thirty 2-, 3-, 4-, and 5-node graphlets $G_0, G_1, \ldots, G_{29}$. In a graphlet G_i, $i \in \{0, 1, \ldots 29\}$, nodes belonging to the same orbit are of the same shade [54].

better fitting, geometric random graph model for PPI networks (see section III-C below).

B. "Graphlet Degree Distribution (GDD)"–Based Network Comparison

In the next step, we introduce a new, more demanding, systematic measure of a network's local structure that imposes 73 similarity constraints on networks being compared. In particular, we generalize the degree distribution, which measures the number of nodes "touching" k edges, into distributions measuring the number of nodes "touching" k graphlets. Note that *an edge* is the only *graphlet with two nodes* (graphlet denoted by G_0 in Figure 1 Right). Thus, the degree distribution measures the following: how many nodes "touch" one G_0, how many nodes "touch" two G_0s, ..., how many nodes "touch" k G_0s. Note that there is nothing special about graphlet G_0 and that there is no reason not to apply the same measurement to other graphlets. Thus, in addition to applying this measurement to an edge, i.e., graphlet G_0, as in the degree distribution, we also apply it to the 29 graphlets $G_1, G_2, \ldots G_{29}$ presented in Figure 1 Right.

When we apply this measurement to graphlets $G_0, \ldots, G_{29}$, certain topological issues arise. For example, for graphlet G_1, we ask how many nodes touch a G_1; note that it is topologically relevant to distinguish between nodes touching a G_1 at an end or at the middle node. This is due to a G_1 admitting an automorphism (defined below) that maps its end nodes to each other and the middle node to itself. Formally, an *isomorphism* g from a graph X to a graph Y is a bijection of nodes of X to nodes of Y such that xy is an edge of X if and only if $g(x)g(y)$ is an edge of Y; an *automorphism* is an isomorphism from a graph to itself. The automorphisms of a graph X form a *group*, called the *automorphism group of* X, and commonly denoted by $Aut(X)$. If x is a node of graph X, then the *automorphism orbit* of x is $Orb(x) = \{y \in V(X) | y = g(x)$ for some $g \in Aut(X)\}$, where $V(X)$ is the set of nodes of graph X. Thus, end nodes of a G_1 belong to one automorphism orbit, while the mid-node of a G_1 belongs to another. Note that graphlet G_0 (i.e., an edge) has only one automorphism orbit, as does graphlet G_2; graphlet G_3 has two automorphism orbits, as does graphlet G_4, graphlet G_5 has one automorphism orbit, graphlet G_6 has three automorphism orbits etc. (see Figure 1 Right). In Figure 1 Right, we illustrate the partition of nodes of graphlets $G_0, G_1, \ldots, G_{29}$ into automorphism orbits (or just *orbits* for brevity); henceforth, we number the 73 different orbits of graphlets $G_0, G_1, \ldots, G_{29}$ from 0 to 72, as illustrated in Figure 1 Right. Analogous to the *degree distribution*, for each of these 73 orbits, we count the number of nodes touching a particular graphlet at a node belonging to a particular orbit. For example, we count how many nodes touch one triangle (i.e., graphlet G_2), how many nodes touch two triangles, how many nodes touch three triangles etc. We need to separate nodes touching a G_1 at an end-node from those touching it at a mid-node; thus we count how many nodes touch one G_1 at an end-node (i.e., at orbit 1), how many nodes touch two G_1s at an end-node, how many nodes touch three G_1s at an end-node etc. and also how many nodes touch one G_1 at a mid-node (i.e., at orbit 2), how many nodes touch two G_1s at a mid-node, how many nodes touch three G_1s at a mid-node etc. In this way, we obtain 73 distributions analogous to the degree distribution (actually, the degree distribution is the distribution for the 0^{th} orbit, i.e., for graphlet G_0). Thus, the degree distribution, which has been considered to be a global network property, is one in the *spectrum* of 73 *"graphlet degree distributions (GDDs)"* measuring local structural properties of a network. Note that GDDs are measuring *local* structure, since they are based on small local network neighborhoods. The distributions are unlikely to be statistically independent of each other, although we have not yet worked out the details of their interdependence.

There are many ways to "reduce" the large quantity of numbers representing 73 sample distributions. In [54], we describe a way that is based on the observed GDDs in the data and model networks, termed network *GDD agreement*; there may be better ways and finding them is an obvious future direction. To compare two networks G and H, we

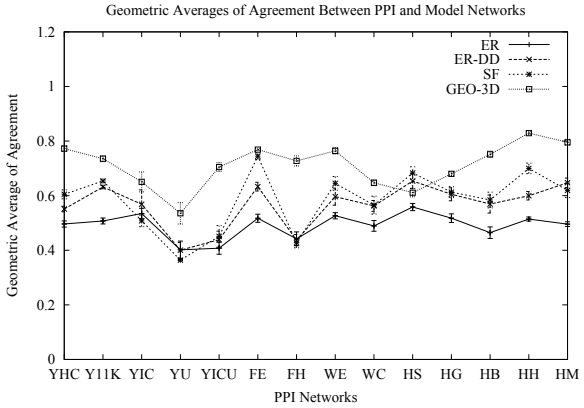

Fig. 2. Agreements between the fourteen PPI networks and their corresponding model networks. Labels on the horizontal axes represent the 14 PPI networks. Averages of agreements between 25 model networks and the corresponding PPI network are presented for each random graph model and each PPI network, i.e., at each point in the Figure; the error bar around a point is one standard deviation below and above the point (in some cases, error bars are barely visible, since they are of the size of the point) [54].

first compute all 73 GDDs of each. We normalize the 2×73 distributions each to have a total area of unity, to meaningfully compare distributions with different areas. We compute a_j, the *agreement in distribution* j, for $j = 0, \ldots, 72$, where $a_j = 1$ means that the networks have identical normalized distributions, and $a_j = 0$ means the networks are very different. Finally, the *average agreement* across all a_j provides one number that compares two networks' GDD agreement; it is a number between 0 and 1, where 1 means that that G and H have identical GDDs, and 0 means that their GDDs are far away (see [54] for details).

C. A Geometric Random Graph Model of PPI Networks

In a *geometric random graph*, nodes correspond to uniformly randomly distributed points in a metric space and there is an edge between two nodes, if the corresponding points in the metric space are close enough according to some distance norm [50]. Note that biological entities, such as genes, and proteins as gene products, exist in a multi-dimensional (likely metric) biochemical space. This space does not only include the 3-dimensional Euclidean space of protein folds with time being the additional 4^{th} dimension; it is likely to be very high-dimensional including protein binding domains, post-translational modifications, small molecules etc. as dimensions. It has been widely accepted and recently verified that genomes evolve through a series of gene duplication and mutation events [32]. Gene duplications and mutations are naturally modeled in the above mentioned currently unknown biochemical space. As the first approximation and a proof of principle, we use 2-, 3-, and 4-dimensional Euclidean boxes with Euclidean distance norm to construct geometric *random* graphs corresponding to the PPI networks (denoted by "GEO-2D", "GEO-3D", and "GEO-4D", respectively) [55].

Using the above described new *distance* measure of network similarity (see section III-A), we find that the high-confidence PPI networks of yeast *S. cerevisiae* [65] and fruitfly *D. melanogaster* [21] are more accurately modeled by geometric random graphs than by scale-free models [55], [53], including a random scale-free model [7], a preferential attachment scale-free model [5], and a gene duplication-mutation scale-free model [64]. The extent of the improvement of the fit of geometric random networks is such that even perturbing the yeast PPI network [65] by random additions, deletions and rewiring of 30% of the edges introduces a much smaller error when compared to the error from modeling the network by scale-free, or other random graph models [55]. In addition, we show that three out of four global network parameters (degree distribution, average diameter, clustering coefficient, and clustering spectrum) also show an improved fit between the experimentally-determined PPI networks and the geometric random graphs than between the PPI networks and the scale-free networks [55], [53].

Next, we undertake a large-scale scientific computing task by implementing the above described new method of computing network *agreement* (see section III-B) and using it to compare agreements across the four random graph models of fourteen real PPI networks. We analyze a total of $1,414$ networks: fourteen eucaryotic PPI networks of varying confidence levels and 25 model networks per random graph model corresponding to each of the fourteen PPI networks, where random graph models are ER, ER-DD, SF, and GEO-3D and the PPI networks analyzed are those of the eucaryotic organisms yeast *S. cerevisiae* [62], [27], [65], fruitfly *D. melanogaster* [21], nematode worm *C. elegans* [37], and human [61], [57], [4], [51], [70]. These PPI networks are obtained by various experimental techniques and are of varying confidence levels, including human curation. The largest of these networks have around $7,000$ nodes and over $20,000$ edges. For each of the fourteen PPI networks and each of the four random graph models, we compute averages and standard deviations of GDD agreements between the PPI and the 25 networks belonging to the same random graph model. The results show that almost all of the fourteen eucaryotic PPI networks are better modeled by geometric random graphs than by Erdos-Renyi [15], random scale-free, or Barabasi-Albert scale-free [5] networks (denoted by ER, ER-DD, and SF in Figure 2, respectively). This further confirms that a biological description of the (possibly *metric*) *space* of PPIs may help us model and understand them.

D. Efficient Estimation of Graphlet Frequency Distributions in PPI Networks

Inspired by the new geometric network model, we designed two efficient heuristic algorithms for estimating the graphlet frequency distribution in PPI networks [56]. These algorithms, along with the heuristic method of Kashtan *et al.* [29], are the first steps towards scalable tools for reliable estimation of local network structure. Such heuristic approaches will be necessary in the future, since exhaustive searches are already becoming computationally infeasible even for the currently available, largely incomplete PPI networks. In addition, humans have

less than $30,000$ genes, each of which can have 4-6 splice variants; therefore, including more than 200 possible post-translational protein modifications, humans are expected to have at least hundreds of thousands of proteins and millions of interactions between them. Plant genome sizes are much larger than the genome size of humans [52], [3] and thus, plants will have even larger proteomes and interactomes. Therefore, any cellular network comparison tool will need to be based on reliable and scalable heuristic algorithms that exploit the network structure of the data. Computing GDDs is scalable because it is "embarrassingly parallel" (i.e., each step can be computed independently from every other step, and thus each step can run on a separate processor to achieve quicker results), but heuristics are still needed to reduce the total CPU time.

IV. Concluding Remarks

Research in biological networks is currently in its infancy and it has already become a vibrant, booming, and exciting new research area that is likely to flourish and make deep impacts on biological understanding, disease, and society in the decades to come. As such, it is rich in open important problems that we are currently only scratching the surface of and is promising to remain at the top of scientific endeavor.

Acknowledgment

This project was supported by the NSF CAREER IIS-0644424 grant.

References

[1] J. Abello, A. Buchsbaum, and J. Westbrook. A functional approach to external graph algorithms. *Lecture Notes in Computer Science*, 1461:332–343, 1998.

[2] W. Aiello, F. Chung, and L. Lu. A random graph model for power law graphs. *Experimental Mathematics*, 10:53–66, 2001.

[3] K. Arumuganathan and E. D. Earle. Nuclear DNA content of some important plant species. *Plant Mol. Biol. Rep.*, 9:208–218, 1991.

[4] G. D. Bader, D. Betel, and C. W. V. Hogue. BIND: the biomolecular interaction network database. *Nucleic Acids Research*, 31(1):248–250, 2003.

[5] A.-L. Barabási and R. Albert. Emergence of scaling in random networks. *Science*, 286(5439):509–12, 1999.

[6] A.-L. Barabási, R. Albert, and H. Jeong. Mean-fi eld theory for scale-free random networks. *Physica A*, 272:173–197, 1999.

[7] E. A. Bender and E. R. Canfi eld. The asymptotic number of labeled graphs with given degree sequences. *Journal of Combinatorial Theory A*, 24:296–307, 1978.

[8] H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, and P.E. Bourne. The protein data bank. *Nucleic Acids Research*, 28:235–242, 2000.

[9] S. Bornholdt and H. Ebel. World-wide web scaling exponent from Simon's 1955 model. *Physical Review E*, 64:046401, 2001.

[10] B.-J. Breitkreutz, C. Stark, and M. Tyers. The GRID: The general repository for interaction datasets. *Genome Biology*, 4:R23:R23.1–R23.3, 2003.

[11] A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, and J. Wiener. Graph structure of the web. *Computer Networks*, 33:309–320, 2000.

[12] S.A. Cook. The complexity of theorem-proving procedures. In *Proc. 3rd Ann. ACM Symp. on Theory of Computing*, pages 151–158. Assosiation for Computing Machinery, 1971.

[13] M. C. Costanzo, M. E. Crawford, J. E. Hirschman, J. E. Kranz, P. Olsen, L. S. Robertson, M. S. Skrzypek, B. R. Braun, K. L. Hopkins, P. Kondu, C. Lengieza, J. E. Lew-Smith, M. Tillberg, and J. I. Garrels. YPDTM, PombePDTM and WormPDTM: model organism volumes of the BioKnowledgeTM library, an integrated resource for protein information. *Nucleic Acids Research*, 29:75–79, 2001.

[14] E. H. Davidson, J. P. Rast, P. Oliveri, A. Ransick, C. Calestani, C.-H. Yuh, T. Minokawa, G. Amore, V. Hinman, C. Arenas-Mena, O. Otim, C. T. Brown, C. B. Livi, P. Y. Lee, R. Revilla, A. G. Rust, Z. J. Pan, M. J. Schilstra, P. J. C. Clarke, M. I. Arnone, L. Rowen, A. Cameron, D. R. McClay, L. Hood, and H. Bolouri. A genomic regulatory network for development. *Science*, 295(5560):1669–1678, 2002.

[15] P. Erdös and A. Rényi. On random graphs. *Publicationes Mathematicae*, 6:290–297, 1959.

[16] P. Erdös and A. Rényi. On the evolution of random graphs. *Publ. Math. Inst. Hung. Acad. Sci.*, 5:17–61, 1960.

[17] P. Erdös and A. Rényi. On the strength of connectedness of a random graph. *Acta Mathematica Scientia Hungary*, 12:261–267, 1961.

[18] M. Faloutsos, P. Faloutsos, and C. Faloutsos. On power-law relationships of the internet topology. *Computer Communications Review*, 29:251–262, 1999.

[19] A. C. Gavin, M. Bosche, R. Krause, P. Grandi, M. Marzioch, A. Bauer, J. Schultz, J. M. Rick, A. M. Michon, C. M. Cruciat, M. Remor, C. Hofert, M. Schelder, M. Brajenovic, H. Ruffner, A. Merino, K. Klein, M. Hudak, D. Dickson, T. Rudi, V. Gnau, A. Bauch, S. Bastuck, B. Huhse, C. Leutwein, M. A. Heurtier, R. R. Copley, A. Edelmann, E. Querfurth, V. Rybin, G. Drewes, M. Raida, T. Bouwmeester, P. Bork, B. Seraphin, B. Kuster, G. Neubauer, and G. Superti-Furga. Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature*, 415(6868):141–7, 2002.

[20] AC Gavin, P Aloy, P Grandi, R Krause, M Boesche, M Marzioch, C Rau, LJ Jensen, S Bastuck, B Dmpelfeld, A Edelmann, MA Heurtier, V Hoffman, C Hoefert, K Klein, M Hudak, AM Michon, M Schelder, M Schirle, M Remor, T Rudi, S Hooper, A Bauer, T Bouwmeester, G Casari, Gand Drewes, G Neubauer, JM Rick, B Kuster, P Bork, RB Russell, and G Superti-Furga. Proteome survey reveals modularity of the yeast cell machinery. *Nature*, 440(7084):631–636, 2006.

[21] L Giot, JS Bader, C Brouwer, A Chaudhuri, B Kuang, Y Li, YL Hao, CE Ooi, B Godwin, E Vitols, G Vijayadamodar, P Pochart, H Machineni, M Welsh, Y Kong, B Zerhusen, R Malcolm, Z Varrone, A Collis, M Minto, S. Burgess, L McDaniel, E Stimpson, F Spriggs, J Williams, K. Neurath, N Ioime, M Agee, E Voss, K Furtak, R Renzulli, N Aanensen, S Carrolla, E Bickelhaupt, Y Lazovatsky, A DaSilva, J Zhong, CA Stanyon, RL Jr Finley, KP White, M Braverman, T Jarvie, S Gold, M Leach, J Knight, RA Shimkets, MP McKenna, J Chant, and JM Rothberg. A protein interaction map of drosophila melanogaster. *Science*, 302(5651):1727–1736, 2003.

[22] M. Girvan and M. E. Newman. Community structure in social and biological networks. *Proc Natl Acad Sci U S A*, 99(12):7821–6, 2002.

[23] K.-I. Goh, B. Kahng, and D. Kim. Hybrid network model: the protein and the protein family interaction networks. *arXiv:q-bio.MN/0312009 v2, 28 March 2004*, 2004.

[24] J. D. H. Han, D. Dupuy, N. Bertin, M. E. Cusick, and Vidal. M. Effect of sampling on topology predictions of protein-protein interaction networks. *Nature Biotechnology*, 23:839–844, 2005.

[25] Y. Ho, A. Gruhler, A. Heilbut, G. D. Bader, L. Moore, S. L. Adams, A. Millar, P. Taylor, K. Bennett, K. Boutilier, L. Yang, C. Wolting, I. Donaldson, S. Schandorff, J. Shewnarane, M. Vo, J. Taggart, M. Goudreault, B. Muskat, C. Alfarano, D. Dewar, Z. Lin, K. Michalickova, A. R. Willems, H. Sassi, P. A. Nielsen, K. J. Rasmussen, J. R. Andersen, L. E. Johansen, L. H. Hansen, H. Jespersen, A. Podtelejnikov, E. Nielsen, J. Crawford, V. Poulsen, B. D. Sorensen, J. Matthiesen, R. C. Hendrickson, F. Gleeson, T. Pawson, M. F. Moran, D. Durocher, M. Mann, C. W. Hogue, D. Figeys, and M. Tyers. Systematic identifi cation of protein complexes in saccharomyces cerevisiae by mass spectrometry. *Nature*, 415(6868):180–3, 2002.

[26] T. Ishii, K.-I. Yoshida, G. Terai, Y. Fujita, and K. Nakai. DBTBS: a database of bacillus subtilis promoters and transcription factors. *Nucleic Acids Research*, 29:278–280, 2001.

[27] T. Ito, K. Tashiro, S. Muta, R. Ozawa, T. Chiba, M. Nishizawa, K. Yamamoto, S. Kuhara, and Y. Sakaki. Toward a protein-protein interaction map of the budding yeast: A comprehensive system to examine two-hybrid interactions in all possible combinations between the yeast proteins. *Proc Natl Acad Sci U S A*, 97(3):1143–7, 2000.

[28] H. Jeong, B. Tombor, R. Albert, Z. N. Oltvai, and A.-L. Barabási. The large-scale organization of metabolic networks. *Nature*, 407(6804):651–4, 2000.

[29] N. Kashtan, S. Itzkovitz, R. Milo, and U. Alon. Effi cient sampling algorithm for estimating subgraph concentrations and detecting network motifs. *Bioinformatics*, 20:1746–1758, 2004.

[30] S. Kauffman. *At Home in the Universe*. Oxford, New York, 1995.

[31] S. A. Kauffman. Metabolic stability and epigenesis in randomly constructed genetic nets. *Journal of Theoretical Biology*, 22:437–467, 1969.

[32] M. Kellis, B. W. Birren, and E. S. Lander. Proof and evolutionary analysis of the ancient gene duplication in yeast saccharomices cerevisiae. *Nature*, 428:617–624, 2004.

[33] J. O. Kephart and S. R. White. Directed-graph epidemiological models of computer viruses. *Proc. 1991 IEEE Comput. Soc. Symp. Res. Security Privacy*, pages 343–359, 1991.

[34] NJ Krogan, G Cagney, H Yu, G Zhong, X Guo, A Ignatchenko, J Li, S Pu, N Datta, AP Tikuisis, T Punna, JM Peregrn-Alvarez, M Shales, X Zhang, M Davey, MD Robinson, A Paccanaro, JE Bray, A Sheung, B Beattie, DP Richards, V Canadien, A Lalev, F Mena, P Wong, A Starostine, MM Canete, J Vlasblom, S Wu, C Orsi, SR Collins, S Chandran, R Haw, JJ Rilstone, K Gandi, NJ Thompson, G Musso, PS Onge, S Ghanny, Lam MHY, G Butland, AM Altaf-Ul, S Kanaya, A Shilatifard, E Oshea, JS Weissman, CJ Ingles, TR Hughes, J Parkinson, M Gerstein, SJ Wodak, A Emili, and JF Greenblatt. Global landscape of protein complexes in the yeast saccharomyces cerevisiae. *Nature*, 440(7084):637–643, 2006.

[35] T. I. Lee, N. J. Rinaldi, F. Robert, D. T. Odom, Z. Bar-Joseph, G. K. Gerber, N. M. Hannett, C. T. Harbison, C. M. Thompson, I. Simon, J. Zeitlinger, E. G. Jennings, H. L. Murray, B. Gordon, B. Ren, J. J. Wyrick, J.-B. Tagne, T. L. Volkert, E. Fraenkel, D. K. Gifford, and R. A. Young. Transcriptional regulatory networks in Saccharomyces cerevisiae. *Science*, 298(5594):799–804, 2002.

[36] L. Li, D. Alderson, R. Tanaka, J. C. Doyle, and W. Willinger. Towards a theory of scale-free graphs: definition, properties, and implications (extended version). *arXiv:cond-mat/0501169*, 2005.

[37] S Li, CM Armstrong, N Bertin, H Ge, S Milstein, M Boxem, P-O Vidalain, J-DJ Han, A Chesneau, T Hao, N Goldberg, DS Li, M Martinez, J-F Rual, P Lamesch, L Xu, M Tewari, SL Wong, LV Zhang, GF Berriz, L Jacotot, P Vaglio, J Reboul, T Hirozane-Kishikawa, Q Li, HW Gabel, A Elewa, B Baumgartner, DJ Rose, H Yu, S Bosak, R Sequerra, A Fraser, SE Mango, WM Saxton, S Strome, S van den Heuvel, F Piano, J Vandenhaute, C Sardet, M Gerstein, L Doucette-Stamm, KC Gunsalus, JW Harper, ME Cusick, FP Roth, DE Hill, and M Vidal. A map of the interactome network of the metazoan c. elegans. *Science*, 303:540–543, 2004.

[38] R. M. May. *Stability and Complexity in Model Ecosystems*. Princeton Univ. Press, Princeton, 1973.

[39] R. Milo, S. Itzkovitz, N. Kashtan, R. Levitt, S. Shen-Orr, I. Ayzenshtat, M. Sheffer, and U. Alon. Superfamilies of evolved and designed networks. *Science*, 303:1538–1542, 2004.

[40] R. Milo, S. S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon. Network motifs: simple building blocks of complex networks. *Science*, 298:824–827, 2002.

[41] M. Molloy and B. Reed. A critical point of random graphs with a given degree sequence. *Random Structures and Algorithms*, 6:161–180, 1995.

[42] M. Molloy and B. Reed. The size of the largest component of a random graph on a fixed degree sequence. *Combinatorics, Probability and Computing*, 7:295–306, 1998.

[43] M. E. Newman. Ego-centered networks and the ripple effect. *Social Networks*, 25:83–95, 2003.

[44] M. E. J. Newman. The structure and function of complex networks. *SIAM Review*, 45(2):167–256, 2003.

[45] M. E. J. Newman, S. H. Strogatz, and D. J. Watts. Random graphs with arbitrary degree distributions and their applications. *Physical Review E*, 64:026118–1, 2001.

[46] M. E. J. Newman and D. J. Watts. Renormalization group analysis in the small-world network model. *Physics Letters A*, 263:341–346, 1999.

[47] M. E. J. Newman and D. J. Watts. Scaling and percolation in the small-world network model. *Physical Review E*, 60:7332–7342, 1999.

[48] R. Overbeek, N. Larsen, G. D. Pusch, M. D'Souza, E. Selkov Jr, N. Kyrpides, M Fonstein, N. Maltsev, and E. Selkov. WIT: integrated system for high-throughput genome sequence analysis and metabolic reconstruction. *Nucleic Acids Research*, 28(1):123–125, 2000.

[49] R. Pastor-Satorras, E. Smith, and R. V. Sole. Evolving protein interaction networks through gene duplication. *Journal of Theoretical Biology*, 222:199–210, 2003.

[50] M. Penrose. *Geometric Random Graphs*. Oxford Univeristy Press, 2003.

[51] S. Peri, J. D. Navarro, T. Z. Kristiansen, R. Amanchy, V. Surendranath, B. Muthusamy, T. K. Gandhi, K. N. Chandrika, N. Deshpande, S. Suresh, B. P. Rashmi, K. Shanker, N. Padma, V. N iranjan, H. C. Harsha, N. Talreja, B. M. Vrushabendra, M. A. Ramya, A. J. Yatish, M. Joy, H. N. S hivashankar, M. P. Kavitha, M. Menezes, D. R. Choudhury, N. Ghosh, R. Saravana, S. Chandran, S. Mohan, C. K. Jonnalagadda, C. K. Prasad, C. Kumar-Sinha, K. S. Deshpande, and A. Pandey. Human protein reference database as a discovery resource for proteomics. *Nucleic Acids Res*, 32 Database issue:D497–501, 2004. 1362-4962 Journal Article.

[52] International Rice Genome Sequencing Project. The map-based sequence of the rice genome. *Nature*, 436:793–800, 2005.

[53] N. Pržulj. *Analyzing Large Biological Networks: Protein-Protein Interactions Example*. PhD thesis, University of Toronto, Canada, 2005.

[54] N. Pržulj. Biological network comparison using graphlet degree distribution. *Bioinformatics*, 23:e177–e183, 2006. Proceedings of the European Conference on Computational Biology (ECCB'06).

[55] N. Pržulj, D. G. Corneil, and I. Jurisica. Modeling interactome: Scale-free or geometric? *Bioinformatics*, 20(18):3508–3515, 2004.

[56] N. Pržulj, D. G. Corneil, and I. Jurisica. Efficient estimation of graphlet frequency distributions in protein-protein interaction networks. *Bioinformatics*, 22(8):974–980, 2006.

[57] J.-F. Rual, K. Venkatesan, T. Hao, T. Hirozane-Kishikawa, A. Dricot, N. Li, G. F. Berriz, F. D. Gibbons, M. Dreze, N. Ayivi-Guedehoussou, N. Klitgord, C. Simon, M. Boxem, S. Milstein, J. Rosenberg, D. S. Goldberg, L. V. Zhang, S. L. Wong, G. Franklin, S. Li, J. S. Albala, J. Lim, C. Fraughton, E. Llamosas, S. Cevik, C. Bex, P. Lamesch, R. S. Sikorski, J. Vandenhaute, H. Y. Zoghbi, A. Smolyar, S. Bosak, R. Sequerra, L. Doucette-Stamm, M. E. Cusick, D. E. Hill, F. P. Roth, and M. Vidal. Towards a proteome-scale map of the human protein-protein interaction network. *Nature*, 437:1173–78, 2005.

[58] H. Salgado, A. Santos-Zavaleta, S. Gama-Castro, D. Millan-Zarate, E. Diaz-Peredo, F. Sanchez-Solano, E. Perez-Rueda, C. Bonavides-Martinez, and J. Collado-Vides. Regulondb (version 3.2): transcriptional regulation and operon organization in escherichia coli k-12. *Nucleic Acids Research*, 29:72–74, 2001.

[59] S. S. Shen-Orr, R. Milo, S. Mangan, and U. Alon. Network motifs in the transcriptional regulation network of escherichia coli. *Nature Genetics*, 31:64–68, 2002.

[60] H. A. Simon. On a class of skew distribution functions. *Biometrika*, 42:425–440, 1955.

[61] U. Stelzl, U. Worm, M. Lalowski, C. Haenig, F.H. Brembeck, H. Goehler, M. Stroedicke, M. Zenkner, A. Schoenherr, S. Koeppen, J. Timm, S. Mintzlaff, C. Abraham, N. Bock, S. Kietzmann, A. Goedde, E. Toksoz, A. Droege, S. Krobitsch, B. Korn, W. Birchmeier, H. Lehrach, and E.E. Wanker. A human protein-protein interaction network: A resource for annotating the proteome. *Cell*, 122:957–968, 2005.

[62] P. Uetz, L. Giot, G. Cagney, T. A. Mansfield, R. S. Judson, J. R. Knight, E. Lockshon, V. Narayan, M. Srinivasan, P. Pochart, A. Qureshi-Emili, Y. Li, B. Godwin, D. Conover, T. Kalbfleish, G. Vijayadamodar, M. Yang, M. Johnston, S. Fields, and J. M. Rothberg. A comprehensive analysis of protein-protein interactions in saccharomyces cerevisiae. *Nature*, 403:623–627, 2000.

[63] A. Vazquez, R. Dobrin, D. Sergi, J.-P. Eckmann, Z. N. Oltvai, and A.-L. Barabasi. The topological relationship between the large-scale attributes and local interaction patterns of complex networks. *P.N.A.S.*, 101(52):17940–17945, 2004.

[64] A. Vazquez, A. Flammini, A. Maritan, and A. Vespignani. Modeling of protein interaction networks. *ComPlexUs*, 1:38–44, 2001.

[65] C. von Mering, R. Krause, B. Snel, M. Cornell, S. G. Oliver, S. Fields, and P. Bork. Comparative assessment of large-scale data sets of protein-protein interactions. *Nature*, 417(6887):399–403, 2002.

[66] A. Wagner. How the global structure of protein interaction networks evolves. *Proceedings of The Royal Society of London. Series B, Biological Sciences*, 270:457–466, 2003.

[67] D. J. Watts and S. H. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393:440–442, 1998.

[68] D. B. West. *Introduction to Graph Theory*. Prentice Hall, Upper Saddle River, NJ., 2nd edition, 2001.

[69] J.G. White, E. Southgate, J. N. Thomson, and S. Brenner. The structure of the nervous system of the nematode C. elegans. *Philosophical Transactions of the Royal Society of London - Series B: Biological Sciences*, 314:1–340, 1986.

[70] A. Zanzoni, L. Montecchi-Palazzi, M. Quondam, G. Ausiello, Helmer-Citterich M., and G. Cesareni. Mint: A molecular interaction database. *FEBS Letters*, 513(1):135–140, 2002.