Analysis of biological networks: Real networks and random network models*

Lecturer: Roded Sharan Scribe: Elena Kyanovsky and David Hadas

Lecture 2, November 02, 2006

In this lecture we give an overview of real networks, discuss their characteristics and present random network models.

1 Real Networks

We start by presenting several commonly used examples for real networks.

1.1 Social networks

Social networks are amongst the first to be studied. *Six degrees of separation* is the theory that anyone on earth can be connected to any other person on the planet through a chain of acquaintances that has no more than five intermediaries. The theory was first proposed in 1929 by the Hungarian writer Frigyes Karinthy in a short story called Chains. In 1967, American sociologist Stanley Milgram devised a new way to test the theory, which he called "the small-world problem". He randomly selected people in the American Midwest to send packages to a target located in Boston, several thousand miles away. The senders knew the recipient's name, occupation, and general location. They were instructed to send the package to a person they knew on a first-name basis who they thought was most likely, out of all their friends, to know the target personally. That person would do the same, and so on, until the package was personally delivered to its target recipient. On average, it only took between five and seven intermediaries to get each package delivered. Milgram's findings were published in Psychology Today and inspired the phrase six degrees of separation. Social networks can be presented as a graph where:

- Nodes represent individuals.
- Edges represent social interactions between individuals, such as relatives, friends, etc.

1.2 Collaboration networks

Collaboration networks are another commonly used example for real networks. As an example, Figure 1 shows a co-authorship network whose center is the mathematician Paul Erdös. Erdös was an expert in the mathematics of networks. Erdös was known for traveling the world and collaborating with mathematicians on problems and proofs he found interesting. Many scientific papers resulted from these intense face-to-face collaborations. Erdös had 507 co-authors, some of whom collaborated with each other. These collaborations became so famous that soon mathematicians were keeping track of their Erdös Numbers. If you co-authored an article with Erdös you have a number of 1, if you co-authored a paper with one of his direct collaborators you have an Erdös Number of 2, and so on. Collaboration networks can be presented as a graph where:

*Based on a scribe by: Polsky Dima and Alperovich Sasha.

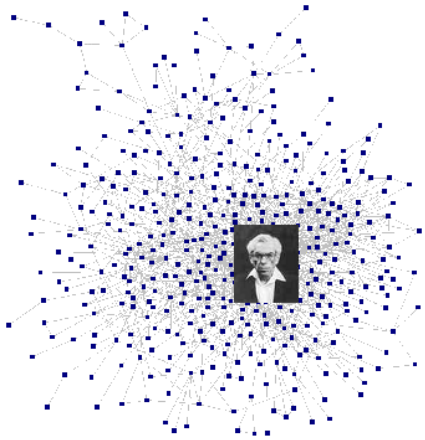


Figure 1: Source [1]. A collaboration network

- Nodes represent individuals.
- Edges represent acts of collaboration.

Another example for a collaboration network is that of film actors:

- Nodes represent film actors.
- Edges represent co-starring in the same film.

1.3 The Internet

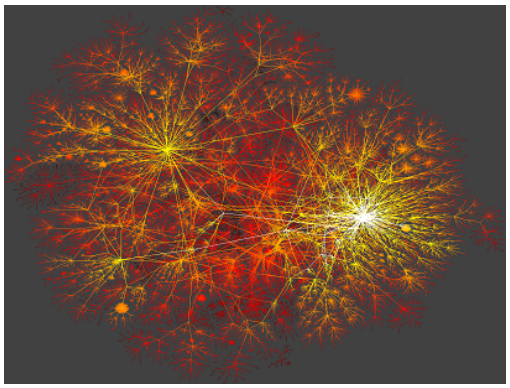
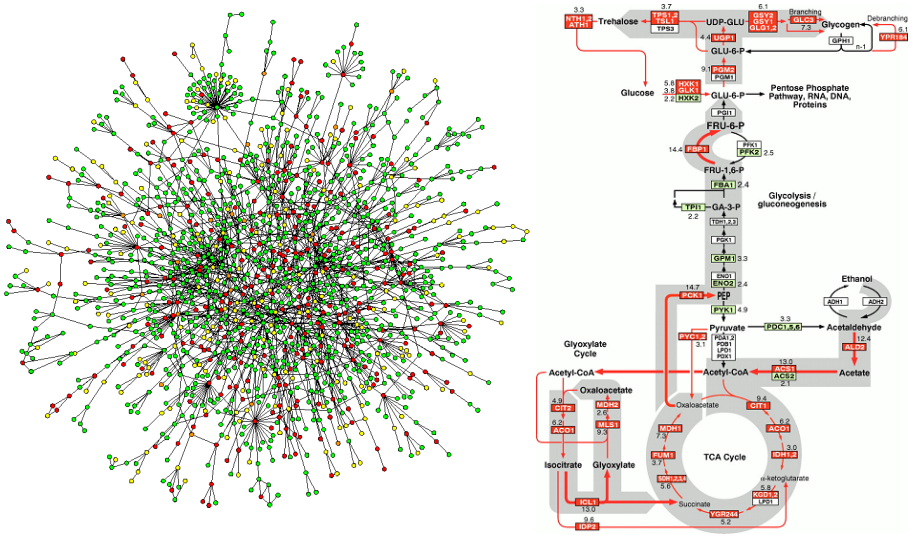


Figure 2: The Internet

Computer networks, for example the Internet network as illustrated in Figure 2, can be presented as a graph where:

- Nodes represent computers or routers.
- Edges represent physical links between the nodes.

1.4 Molecular networks



(a) (b) Metabolic network

Figure 3: Examples of molecular networks. (a) Protein-protein interaction network. (b) Metabolic network.

Molecular networks describe biological processes at different levels. In most cases the nodes represent genes or proteins and the edges represent molecular interactions. We review below some of the most studied networks.

Expression networks represent similarities at the transcriptional level. At such a graph:

- Nodes represent genes.
- Edges represent similarity of gene expression patterns across a set of conditions.

Protein-protein interaction networks (Figure 3(a)) represent interactions at the protein level:

- Nodes represent proteins.
- Edges represent physical interactions between proteins.

Transcriptional networks describe the regulatory circuitry in the cell:

- Nodes represent genes.
- Edges are directed from a transcription factor to the genes it regulates.

Metabolic networks (Figure 3(b)) reflect the cells metabolic circuitry:

- Nodes represent metabolites.
- Edges are directed and represent reactions in which one metabolite turns into another.

Molecule networks reflect the atoms in a chemical component:

- Nodes represent atoms.

- Edges represent forces between the atoms in the molecule.

Chemical universe networks reflect the similarity between chemical components:

- Nodes represent molecules.
- Edges represent structural similarity between the molecules.

Chemical reaction networks reflect the chemical reactions between molecules:

- Nodes represent molecules.
- Edges represent chemical reactions between the molecules.

Neural networks reflect the buildup of a neural tissue:

- Nodes represent neurons.
- Edges are directed and represent synapse between an axon of a pre-synaptic neuron and dendrites of a post-synaptic neuron.

2 Graph Theoretic Concepts

In this section we will briefly cover general concepts from graph theory. We will focus on terms and measures that are commonly used to describe real networks.

2.1 Graphs and simple graphs

Graph, Network A collection of elements and a collection of binary relations between the elements. The graph is represented by $G = (V, E)$ where V is a collection of elements and E is a collection of binary relations between the elements.

Vertex, Node An element in a graph. We will use v, u, s and t to denote specific vertices in a graph. We use $N = |V|$ to represent the the number of vertices in a graph (also known as its size).

Edge, Link A relation between two elements of a graph. We will use (v, u) to denote an edge between vertices v and u. Each edge connects two vertices. The edges of a graph may be *directed*, representing causal relationships, or *undirected* representing a two way relation. Properties of elements or relations may be indicated by node and edge labels, or by *edge weights*. We use $m = |E|$ to represent the number of edges in a graph.

A Simple Graph A graph is termed a *simple graph* if it has no multiple edges or loops connecting a vertex to itself. A *tree* is a graph in which every two vertices are connected by exactly one path.

Distance The *Distance* between two vertices u and v in a graph G is the length of a shortest path between them. When u and v are identical, their distance is defined to be 0. When u and v are unreachable from each other, their distance is defined to be infinity.

Connected components Let $G = (V, E)$ be a graph. $H = (V', E')$ is a *subgraph* of G if $V' \subseteq V$ and $E' \subseteq E \cap (V' \times V')$. H is connected if there is a path between every pair of vertices in G. A *connected component* of G is defined as a maximal connected subgraph of G. Finally, a *giant component* is a component that contains a constant fraction of the graphs nodes.

2.2 Degree distributions

The *Degree sequence* for a graph is the vector $(d_1, d_2, , d_n)$ holding the degree information d_v of each vertex v in the graph. The *Degree distribution* for the graph denoted by $P(k)$ represents the probability of a vertex in the graph to have a degree of k. The degree distribution can be calculated using:

$$P(k) = \frac{|\{v|d_v = k\}|}{N} \tag{1}$$

where d_v is the degree of vertex v and N is the number of vertices in the graph. A network can be characterized in terms of its Degree Distribution, i.e. the distribution function of degrees in the network. Note that directed graphs would have two Degree Distributions, namely *In-degree distribution* and *Out-degree distribution*.

We denote the *maximum degree* by k_{max}, the largest degree over all vertices in the graph. The *average degree* in a graph is denoted $d \equiv \sum_k kP(k)$, i.e., the first moment (average) of the degree distribution function. Note that the number of edges in the graph is given by $m = \frac{Nd}{2}$.

We will next present several distribution functions that are commonly used when describing the probability function $P(k)$ of a vertex to have a degree k. Note that the m-th moment of $P(k)$ is defined by:

$$M_m \equiv \sum_{k=1}^{\infty} k^m P(k) \tag{2}$$

- Poisson: $P(k) = \frac{e^{-d}d^k}{k!}$ - This distribution describes networks whose degree distribution is highly concentrated around the mean and hardly any vertices have a degree further away from the mean.

- Exponential: $P(k) \propto e^{\frac{-k}{d}}$ - This distribution describes networks which have many vertices with low degrees and hardly any vertices with large degrees.

- Power-law: $P(k) \propto k^{-c}, k \neq 0, c > 1.$ - This distribution describes networks which contrasts with the Poisson and Exponential distributions. Although it is characterized by many vertices with low degrees, it also includes a small number of vertices with high vertex degrees. Such vertices are highly important to the network connectivity and serve as *hubs*. The power-law distribution have finite moments for $m < c - 1$.

  Power-law distributions are named *scale free* since they can be scaled without altering the distribution. If we denote the distribution by $p(x)$, scaling the distribution by a factor a results in $p(ax) = g(a)p(x)$. Thus, the power-law distribution has no natural scale and is scale invariant (in fact, this is the only distribution with such property).

2.3 Clustering coefficient

Another important measure seeking to evaluate how well is a network connected is the *Clustering coefficient*. Clustering coefficients take values in the range $[0, 1]$ and it measures the tendency of the network to form highly interconnected regions called *Clusters*.

The *Vertex clustering coefficient* C_v for vertex v is the proportion of links between the vertices within its neighborhood divided by the number of links that could possibly exist between them. Every link between two neighbors of v, for example s and t, creates a *Triangle*: a loop of three vertices, namely, v,s and t. The clustering coefficient of v is the percentage of triangles among the neighbors of v (i.e. the number of triangles divided by the number of all possible Triangles).

The vertex clustering coefficient can therefore be calculated by $C_v = \frac{t_v}{d_v(d_v-1)/2}$, where d is the degree of v and t is the number of v's triangles (number of links between the d neighbors of v). See Figure 4. We define for $d = 0$ and for $d = 1$ that $C_v = 0$. The *network clustering coefficient* is marked with C and is defined as the average of C_v over all vertices in a network. Therefore: $C = \frac{1}{N}\sum_v C_v$. For example, the clustering coefficient of a *Clique* is 1. A graph without triangles has a clustering coefficient of 0.

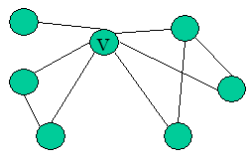


Figure 4: Calculating clustering coefficient of a vertex. vertex v above has $d = 6$, $e = 3$ and therefore $C_v = \frac{3}{(6*5)/2} = 0.2$

2.4 Centrality measures

Some of the most important measures of a network seek to estimate the *Centrality* of a vertex within a graph. The centrality of a vertex in a social network for example can indicates how important a person is, within the social network. As another example, the centrality of a vertex in a biological network may indicate how important a protein or a gene is, within a cell. Various centrality measures can be offered, we will discuss here two of them: degree and betweeness.

A node u is called a *Neighbor* of a node v, if it is connected to it by an edge. The *Neighborhood* of a node v, denoted as $N(v)$, is the set of all neighbors of v. The *Degree* of a vertex v is the number of edges connected to the vertex and equivalently the size of its neighborhood (the number of neighbors or relationship which the node has). The Degree is the most basic centrality measure and is marked $d(v)$ or d_v. The Degree therefore shows the centrality of the vertex in the network. In many real networks, a well connected vertex represent a more important one. Note that vertex of directed graphs are denoted by two Degrees, namely *In-degree* and *Out-degree*.

Another important centrality measure is the *Betweenness* of a vertex v, also known as *Betweenness centrality*. Betweenness of a vertex is a measure indicating how central this vertex is to the graph. Betweeness is most often calculated as the fraction of shortest paths between node pairs that pass through the node of interest. The betweeness measure $C_b(v)$ of a node v can be calculated in the following way:

Let $\sigma_{st} = \sigma_{ts}$ to be the number of shortest paths between two nodes s and t ($\sigma_{ss} = 1$).
Let $\sigma_{st}(v)$ be the number of shortest paths between two nodes s and t that goes through node v.
Then, the betweenness centrality $C_b(v)$ of any vertex v can be computed as:

$$C_b(v) = \sum_{s \neq v \neq t} \frac{\sigma_{st}(v)}{\sigma_{st}} \tag{3}$$

2.5 Small world

One of the phenomenons of real networks is that the average distance through the network from one vertex to another is small compared to the network size. In social networks this effect is known as the small-world effect. We will analyze this effect on a random graph example by estimating *Average distance* in a random graph. Since a random network is characterized by low clustering coefficient, it is locally tree-like. As a result, assuming that the graph degree is d, the number of vertices at distance i from a vertex is approximately $\sum_{k=0}^{i} d^k$ (See Figure 5).

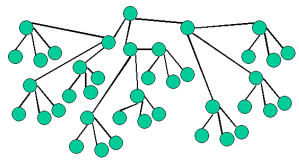


Figure 5: Number of vertices at distance 3 with degree of 3

We define l to be an average distance in the random network. Network size is $N \sim d^l$. Hence, average distance is $l \sim lnN/lnd$. The average distance in a random graph is logarithmically increasing with graph size. In other worlds, the average distance is related to $\ln N$ and remains small, yet dependent on N. Figure 6 shows that the average distance l in many real network cases is found to be fairly small as compared to the size of the network.

	network	type	n	m	z	ℓ
social	film actors	undirected	449 913	25 516 482	113.43	3.48
	company directors	undirected	7 673	55 392	14.44	4.60
	math coauthorship	undirected	253 339	496 489	3.92	7.57
	physics coauthorship	undirected	52 909	245 300	9.27	6.19
	biology coauthorship	undirected	1 520 251	11 803 064	15.53	4.92
	telephone call graph	undirected	47 000 000	80 000 000	3.16	
	email messages	directed	59 912	86 300	1.44	4.95
	email address books	directed	16 881	57 029	3.38	5.22
	student relationships	undirected	573	477	1.66	16.01
	sexual contacts	undirected	2 810			
information	WWW `nd.edu`	directed	269 504	1 497 135	5.55	11.27
	WWW Altavista	directed	203 549 046	2 130 000 000	10.46	16.18
	citation network	directed	783 339	6 716 198	8.57	
	Roget's Thesaurus	directed	1 022	5 103	4.99	4.87
	word co-occurrence	undirected	460 902	17 000 000	70.13	

Figure 6: Source [10]. Average distances l for various real networks with size indicated as n.

3 Network models - part I

The study of real-world networks, such as biological networks, has evoked the need for random models. Such models would allow researchers to better understand real-world networks, and to predict their behavior. We present here some of the common models developed for this purpose.

3.1 ER (Random) graphs

Random graphs are one of the most studies type of networks. Random graphs were presented by Erdös/Rényi [5] in the 1950s and 1960s. Erdös/Rényi characterized random networks and showed that many of the properties of such networks can be calculated analytically. Random graphs are also named *ER graphs* after the initials of Erdös/Rényi.

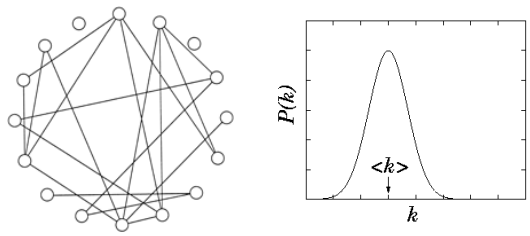


Figure 7: Source [5]. Degree distribution of a random graph, and an example of such a graph.

3.1.1 Constructing a random graphs

A random network network can be constructed in several ways. A simple way to construct a random network with N vertices is using the following algorithm:

For each pair of vertices (u, v) in the graph: Connect the two vertices with an edge by chance p and do not connect the two vertices by chance $1 - p$.

Note that the algorithm produces a simple graph.

3.1.2 Characterizing random graphs

The degree distribution ER graphs have binomial degree distribution. For large networks, the distribution is asymptotically Poisson. The ER graph degree distribution is concentrated around the mean value and exponentially decays (see Figure 7). The average degree is $d = (N - 1)p \approx Np$. The variance of the binomial distribution is $d_2 \approx Np(1 - p)$.

Since in a random graph, each two vertices would have an edge between them in probability of p, it follows that the network clustering coefficient is $C = p$. Most of the networks we shall discuss are characterized by a finite Graph Degree d and a very large graph size N. For such networks, since: $d \approx Np$, it follows that $C = p \sim \frac{1}{N}$. Hence the clustering coefficient of such networks follows $\frac{1}{N}$ meaning that for as N increases the clustering coefficient goes to zero. In other words, the random graph behaves in a way similar to a Tree.

3.2 The Scale-free model by Barabási and Albert

The m-th moment of the power-law distribution is finite for $m < c - 1$. Models with such degree distribution are characterized by a few central highly connected nodes known as the *hubs*.

Scale free networks and networks whose vertex distribution follows the power-law rule. In such networks, some nodes act as "highly connected hubs" (high degree), although most nodes are of low degree.

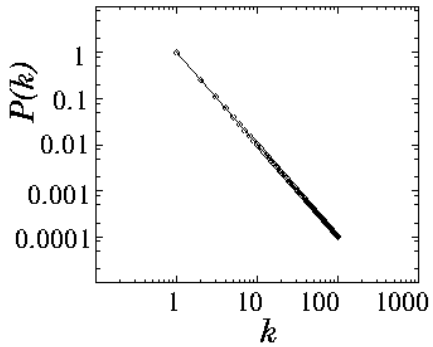


Figure 8: Source [2]. Power-law degree distribution in log-log representation.

Figure 8 shows a plotting of the power-law distribution in log-log representation. Under a log-log representation, a Power-Law distribution is characterized by a linear line. Albert and Barabasi [2] showed in 1999 that many real networks behave like scale-free networks.

Power-law degree distribution $P(k) \propto K^{-c}$ is characterized by a small number of highly connected hubs. Due to the existence of the central hubs, Scale Free networks are highly connected as compared to a random network.

4 Properties of real network

A In the following we introduce several basic properties of real networks [4].

4.1 Vertex degrees

The Degree is an important centrality measure is real networks. As an example of the importance of degree in the biological context, lets discuss an experiment performed by [8]. In this experiment a Yeast protein-protein interaction network is investigated. The network has 1870 proteins as nodes, connected by 2240 identified direct physical interactions. The vertex degree of was found to be following a Power-Law distribution. Also, it was found that on average less connected proteins prove to be less essential than highly connected ones.

For example, while proteins with five or less links constitute 93% of the total number of proteins it was found that only 21% of them are essential. In contrast, only 0.7% of the yeast proteins with known phenotypic profile have more than 15 links but single deletion of 62% of these proves lethal. This implies that highly-connected proteins, those with high Degree have not only a central role in the networks architecture but are three times more likely to prove essential to the yeast than proteins with low number of links to other proteins. The results are summarized in Figure 9 showing the % of essential proteins vs. a node Degree.

4.2 Vertex degree distribution

Many examples of real networks seem to have a Vertex Degree Distribution similar to that described by scale-free networks.

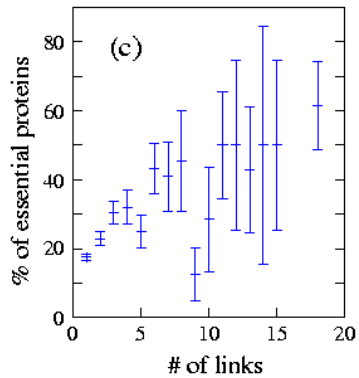


Figure 9: Source [8]. The essentiality of a gene as a function of its degree in the yeast protein-protein interaction network.

4.2.1 The Internet

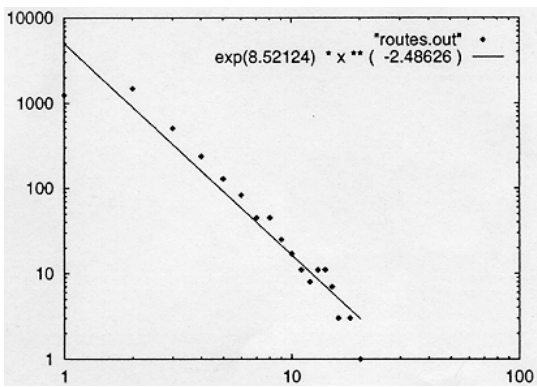


Figure 10: Source [7]. Degree distribution of the Internet.

When analyzing the Internet as a graph with Routers as Nodes and Physical links between routers as edges, we may consider the degree distribution function of the nodes. M. Faloutsos, P. Faloutsos, and C. Faloutsos [7] showed on 1999 that the degree distribution of the Internet can be approximated using: $P(k) \sim k^{-2.5}$ Using a log-log scale, we can see that the Internet network behaves like a scale-free network. Figure 10 shows the Power-law distribution graph of the Internet on a log-log scale.

4.2.2 Film actors

Barabási and Albert [2] analyzed in 1999 the collaboration graph of movie actors. This graph represents a well documented example of a social network. Each actor is represented by a vertex, two actors being connected if they were cast together in the same movie. The probability that an actor has k links (characterizing his or her popularity) has a power-law tail for large k, following $P(k) \sim k^{-2.3}$

Figure 11 shows the Power-law distribution graph of the actors social network.

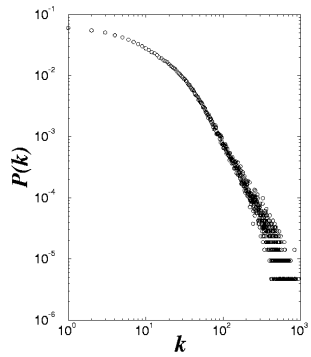


Figure 11: Source [2]. Degree distribution of the film actors collaboration network.

4.2.3 Protein interaction networks

Yook [11] analyzed in 2004 a variety of Protein Interaction networks and found the probability that a protein interacts with k other proteins (characterizing his or her popularity) has a power-law, following $P(k) \sim k^{-2.5}$

Figure 12 shows the Power-law distribution graph of the protein interaction network. It is interesting to note that each protein network by itself as well as all taken together seem to all follow the power-law distribution with $c = 2.3$.

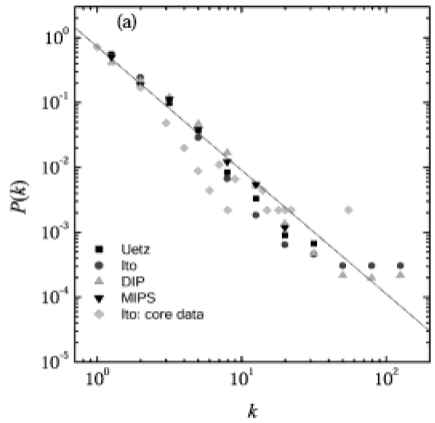


Figure 12: Source [11]. Degree distribution of the protein-protein interaction network.

4.2.4 Metabolic networks

Jeong et al. writes in a publication from 2000 [9],

In his article, Jeong presents a systematic comparative mathematical analysis of the metabolic networks of 43 organisms representing all three domains of life. He shows that, despite significant variances in their individual constituents and pathways, these metabolic networks display the same topological scaling properties demonstrating striking similarities to the inherent organization of complex non-biological systems. In these networks, each Metabolite is represented by a vertex, two metabolites are connected if there is a

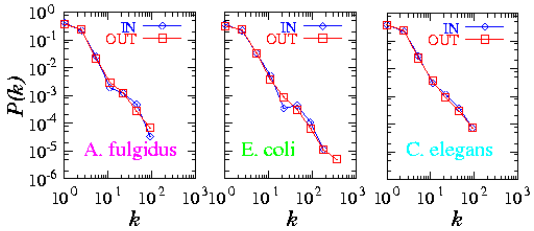


Figure 13: Source [9]. Degree distribution of metabolic networks.

biochemical reaction transforming one metabolite to the other. The edges are therefore directed resulting in a directed graph. The distribution of vertex degrees in the resulting metabolic networks follow a power-law distribution with $P(k) \sim k^{-2.2\pm0.2}$

Figure 13 shows some examples of the power-law distribution graph of the metabolic networks. It is interesting to note that both in-degree distribution and out-degree distribution follow the power-law distribution rule with the same value range of c. Generally, most of the real networks are scale free. Let us symmetrize the distributions of the above networks:

Network	Degree Distribution
Internet	$P(k) \propto k^{-2.5}$
Film Actors	$P(k) \propto k^{-2.3}$
Protein Interaction	$P(k) \propto k^{-2.5}$
Metabolic Networks	$P(k) \propto k^{-2.2\pm0.2}$

4.3 Clustering coefficients

Pruzlj [6] presents a comparison between the clustering coefficients of real networks. The analytical results from calculating the respective clustering coefficients from parallel random (ER) networks. See Figure 14. As can be seen, real world clustering coefficients are considerably higher than those anticipated using random networks. In Real World networks clustering is significant. For example, in the Internet network, the chance of two neighbors of a node to also be connected is about 24% compared to the analytical result based on a random network which yields 0.06%.

It is evident from the presented table that the measured C is significantly higher than the clustering coefficient of random networks.

5 Network models - part II

We will now continue our discussion about network models.

5.1 Scale free networks

5.1.1 Preferential attachment

In their article, Albert and Barabasi [2] indicated that a common property of many large networks is that the vertex connectivities follow a scale-free power-law distribution.

The *Preferential attachment* model offered by Albert and Barabasi suggested that networks are constructed by gradually adding new elements (see Figure 15) to an existing network and connecting it to some

Network	n	z	C measured	C for random graph
Internet [153]	6,374	3.8	0.24	0.00060
World Wide Web (sites) [2]	153,127	35.2	0.11	0.00023
power grid [192]	4,941	2.7	0.080	0.00054
biology collaborations [140]	1,520,251	15.5	0.081	0.000010
mathematics collaborations [141]	253,339	3.9	0.15	0.000015
film actor collaborations [149]	449,913	113.4	0.20	0.00025
company directors [149]	7,673	14.4	0.59	0.0019
word co-occurrence [90]	460,902	70.1	0.44	0.00015
neural network [192]	282	14.0	0.28	0.049
metabolic network [69]	315	28.3	0.59	0.090
food web [138]	134	8.7	0.22	0.065

Figure 14: Source [6]. Number of vertices n, mean degree z and clustering coefficient C for a number of different networks.

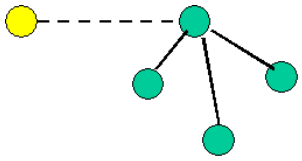


Figure 15: Preferential attachment: (i) networks expand continuously by the addition of new vertices, and (ii) new vertices attach preferentially to sites that are already well connected.

of the existing nodes. The model suggests that the probability that the newly added node a will be connected to an existing one depends on the existing node's degree

$$P(a - v) = \frac{d(v)}{\sum_{u \in V} d(u)} \tag{4}$$

To rationalize this model, Albert and Barabasi provide several examples;

- The Actor network grows by the addition of new actors to the system.
- The WWW grows exponentially over time by the addition of new Web pages
- The research literature constantly grows by the publication of new papers.

Consequently, a common feature of these systems is that the network continuously expands by the addition of new vertices that are connected to the vertices already present in the system. For example a new actor is most likely to be cast in a supporting role with more established and better-known actors. Therefore, the probability that a new actor will be cast with an established one is much higher than that the new actor will be cast with other less-known actors. Similarly, a newly created Web page will be more likely to include links to well-known popular documents with already-high connectivity, and a new manuscript is more likely to cite a well-known and thus much-cited paper than its less-cited and consequently less-known peer.

We also find such behavior in biological networks. For example, new genes are created by duplicating old genes. As a result, immediately after duplicating the duplicated gene is attached to the same nodes

to which original gene was connected. As a result if a gene is a hub, it would have a higher chance of connecting to the original gene and therefore higher chance of being connected to the duplicated gene. We will later show that $P(k) \propto k^{-3}$. An immediate result of the Preferential Attachment model is that a node with high degree will tend to continue and increase its degree, ("the rich get richer")

5.1.2 Maximum degree

We can estimate the maximum degree using a reasonable expectation that our network includes only one vertex with a degree equal to k_{max} Therefore it holds that for such graphs, $P(k_{max}) = \frac{1}{N}$. Since by definition we would have $P(k) = 0; k > k_{max}$, it stands that for finite graphs, we would expect

$$N \sum_{k \geq k_{max}} P(k) \sim 1 \tag{5}$$

If we now replace the sum with an integral, we will see that:

$$N \sum_{k \geq k_{max}} P(k) = N \int_{k_{max}}^{\infty} P(k) \sim 1. \tag{6}$$

Although this is only an estimation, it provides a powerful tool to evaluate the maximum degree given the number of vertices N and the distribution function $P(k)$. Since for scale free networks: $P(k) \equiv K^{-}c$ it holds that:

$$N \int_{k_{max}}^{\infty} K^{-c} = (-c+1)NK^{-c+1}|_{k_{max}}^{\infty} = 0 - (-c+1)NK_{max}^{-c+1} \sim 1 \tag{7}$$

or

$$K_{max} \sim N^{\frac{-1}{c-1}} \tag{8}$$

Although this is only an estimation, it provides us with a powerful tool to evaluate the maximum degree based on N and the distribution's power-law constant.

5.2 Random graphs

5.2.1 Why use random graphs?

The random graph is one of the oldest and most studied models of a network. Random graphs are a very useful model to compare with the real networks behavior. When we study a phenomenon at the real network, we can use a random model to realize if the phenomenon carries information or if it is random.

The ER model is can be analyzed analytically with ease with regard to many of its average properties. It is therefore a very handy tool for analyzing real networks. However, the ER model differs from real networks in two crucial ways: it lacks network clustering, and it has unrealistic Poissonian degree distribution.

5.2.2 Characterizing random graphs

A graph created by this model has N vertices, and for each pair of vertices (u, v), an edge exists between v and u with probability p. The Average number of edges in an ER graph can be calculated by: $\frac{1}{2}N(N-1)p$. The Average degree therefore $d = \frac{N(N-1)p}{N} = (N-1)p \approx Np$. As we can see ER Graphs have binomial degree distribution. For large networks, the distribution is asymptotically Poisson. Such distribution is concentrated around the mean value and exponentially decays (see Figure 7).

5.2.3 Giant component

We can estimate different properties of random graphs. As an example let us look on a *Giant component* creation, that is dependent on a p parameter. If p is growing the number of edges is growing and the graph becomes a connected component. At a graph with N vertices when $p = d$ is low, we can expect many small components. As p increases, a giant component of size sN (where s is a chosen constant) is formed. Let us now analyze when does a giant component is formed, and what is the expected size of a giant component.

Define u to be the probability that a random vertex doesn't exist at the giant component and s to be the supplement of u ($s = 1 - u$). If a vertex has k neighbors with probability $P(k) = e^{\frac{(-d)*d^k}{k!}}$, the probability that a vertex is not included in the giant component is equal to the probability that all his neighbors are not included into the giant component, which equals to u^k for a vertex of degree k.

Averaging this expression over all vertices results in:

$$u = \sum_k P(k)u^k \approx e^{-d} \sum_k \frac{d^k u^k}{k!} = e^{-d} e^{ud} = e^{d(u-1)}$$

$$S = 1 - e^{-dS}$$

This equation is not solvable, but it can be analyzed and we can find the cases when it will have the solutions.

- For $d < 1$ (each vertex has less that 1 edge) there is a single non-negative solution when $s = 0 \Rightarrow$ and Giant Component doesn't exist.
- For $d > 1$ there is non-zero solution, Giant Component exists that covers a constant fraction of the vertices.
- There is a phase transition at $d = 1$.

5.2.4 Drawbacks of random graphs

We have seen that a random graph is not a very good model to represent real networks. The model clearly has the following drawbacks:

1. It is characterized by a low clustering coefficient, in contrast to many known real networks.
2. It is characterized by a Poisson degree distribution, in contrast to power-law distribution as seen in real networks.

These significant differences between random networks and real networks limit the usefulness of random networks. Most significantly, in a random networks all vertices are alike, while real networks are characterized by a small number of vertices with very large degree while most vertices maintain a very low degree.

We will next describe the generalized random network model looking to overcome this difference.

5.3 Generalized random graphs

As discussed earlier, the degree of many real networks follows a power-law distribution. However, the random graph model we have discussed so far, do not reproduce such distribution.

In 1978, Rodney and Canfield [3] suggested to improve the approximation of real networks by controlling the degree distribution of the network. The suggested *Generalized random graph* model, creates a

graph based on a given degree sequence. That is, it uniformly picks a graph in which node i has degree k_i, for a given $\{k_i\}$ set.

This allows us to consider graphs with a specific degree distribution, and in particular graphs with power-law distributions. This model is helpful in analyzing the behavior of real networks, as will be seen below.

5.3.1 Construction using a matching algorithm

Let‘s consider how a generalized random network can be constructed.

Rodney and Canfield [3] suggested a matching algorithm that includes random assignments of vertices to edges whereas a vertex k is assigned k_d times.Assuming the degree of vertex k is k_d for every vertex:

1. Prepare k_d copies of vertex k.
2. Randomly assign copies to edges.

We consider the different copies of each vertex to be the same, we therefore may use different choices during the construction algorithm and still produce the same graph topology. In fact, each vertex that has k_d copies, contributes $d(v)!$ different construction choices. Therefore, the number of different choices that may produce each and every Graph Topology is: $\prod_v d(v)!$.

Hence, every graph topology has the same chance to be created using the presented algorithm and therefore the construction algorithm is uniform. The matching algorithm may introduce graphs with unity loops and duplicate vertices. Optionally, we may choose to reject such graphs and repeat the matching algorithm until we come up with a simple graph. Naturally, this works for graphs with low vertex degree. However, if we choose a graph with higher vertex degrees, we may fail to come up with a simple generalized random graph within a reasonable amount of trials.

1. Prepare k copies of each k-degree vertex.
2. Randomly assign copies to edges. This can be done by creating a two column table where each column contains all the copies in random order and creating a link for every row.
3. Repeat steps 1-2 if the resulting graph is not simple.

Using this algorithm, a graph is chosen uniformly at random from the collection of all graphs with a given degree sequence.

5.3.2 Construction using a switching algorithm

Another method to generate a generalized random graph is to pick any graph which conforms to the required degree distribution (i.e. node i has degree k_i, for a given $\{k_i\}$ set) and change it to using a long random series of edge crosses, until it becomes a generalized random graph. Using this algorithm, we randomly select two pairs of connected nodes and cross their edges : for example $u - v, s - t \implies u - s, v - t$ or $u - t, v - s$ (See Figure 16). The switching is legal only if the resulting graph remains simple. Therefore, if the resulting graph includes unit loops, we will skip this iteration.

It was found empirically that each vertex should be switched on average 100 times to approximate generalized random graph.

The procedure :

1. Randomly select two links $(u, v), (j, k)$.

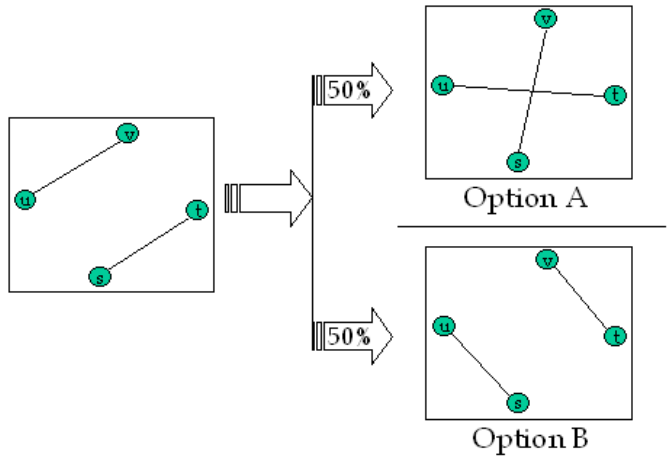


Figure 16: The Switching Algorithm

2. Choose the crossing direction by a 50% chance: either edges $(u, j), (v, k)$ or edges $(u, k), (j, v)$.

3. Check if the graph that will result if the crossing take place is a simple graph, if so, perform the crossing.

4. Repeat steps 1-3 for $100 * m$ successful crossings where m is the number of edges in the graph.

5.3.3 Probability of an edge between two vertices

When constructing a Generalized Random Chance, the change of for an edge to exist between two vertices, u and v depends on the respective degrees of the two vertices and the total number of edges m in the graph.

A good approximation of this chance can be evaluated to be: $p(u, v) = \frac{d(u)d(v)}{2m}$

This approximation can be rationalized as follows:

Assuming we first assign all copies of u. Overall there are now $d(u)$ edges where u is assigned. Next, we move on to assign v. Assuming m is considerably larger than $d(u)$ and $d(v)$, we have approximately $2m$ different choices to place each of the copies of v. So, overall we have $2m^{d(v)}$ different options.

Out of those options, any one of the $d(v)$ copies of v can meet any of the $d(u)$ copies of u, while the other copies of v are assigned in any other location. Overall $d(u)d(v)2m^{(d(v)-1)}$ different options.

To sum this approximation, we would note that:

$p(u, v) = \frac{d(u)d(v)2m^{(d(v)-1)}}{2m^{d(v)}} = \frac{d(u)d(v)}{2m}$

5.3.4 Model analysis

We will next evaluate different properties on the received graph and compare these values with the real graphs statistics. First, let estimate the average graph distance. We will assume that the clustering coefficient is negligible, $C \Rightarrow 0$. We will start by evaluating the number of vertices at distance 2 from vertex v with average degree d. We denote this number by: d_2.

The degree distribution of the first neighbor of a vertex is not the same as the degree distribution of vertices on the graph as a whole - $P(k)$. Because a high degree vertex has more edges connected to it, there is a higher chance that any given edge on the graph will be connected to this node, in precise proportion to the vertex's degree. Thus, the probability distribution of the degree of the vertex to reach an edge leads it proportional to $kP(k)$ for $P(k)$ vertices with k edges.

When reaching the neighboring node, we use one of its remaining edges to move on to the distance of 2 nodes. We will mark a vertex degree to be $(k+1)$ and therefore a neighbor would have k remaining (unvisited) edges. We define q_k as a probability of reaching a vertex that has k unvisited edges ($(k+1)$ edges altogether). q_k is proportional to $(k+1)P(k+1)$.

$$q_k = \frac{(k+1)P(k+1)}{\sum_k kP(k)} = \frac{(k+1)P(k+1)}{d}$$

We can look at q_k as at a probability that we reach a vertex with k free edges (this vertex initially has $(k+1)$ degree).

An unvisited edge of a vertex v is one that didn't participate in the computation when v is reached by the computation.

The average number of vertices at distance 2 from v is :

$$\sum_k kq_k = \frac{\sum k(k+1)P(k+1)}{\sum kP(k)} = \frac{\sum (k+1)^2P(k+1) - (k+1)P(k+1)}{d} = \frac{M_2 - d}{d}$$

where M_2 is the second moment of the degree distribution $P(k)$

¿From the above, we can calculate the mean number of neighbors at distance 2 for a given node. The result depends on the number of vertices at distance 1 and their degree distribution. We therefore obtain:

$$d_2 = [\sum_k kP(k)][\sum_k kq_k] = \frac{(M_2 - d)d}{d} = M_2 - d$$

We can extend this computation to all distances m. The number of the neighbors in distance m equals to taking a single step from the neighbors in distance $m-1$ and can be written as :

$$d_m = d_{m-1}\frac{d_2}{d} = \left(\frac{d_2}{d}\right)^{m-1} d$$

¿From this computation we can learn about the existence of a giant component in the graph. Depending on whether d_2 is greater than d or not, this expression will either diverge or converge exponentially as m becomes large.

- $d_2 > d$: the neighborhood grows exponentially and a giant component emerges.
- $d_2 < d$: the size of the $m-th$ neighborhood converges.
- The conclusion is that phase transition occurs at $d_2 = d$ that is equivalent to:

$$\sum_k P(k)k(k-2) > 0$$

  which has a critical value at $c = 3.48$

Most of the real networks with giant component have degree distribution in the $[2, 3]$ range.

It is interesting to note, that because of the $k(k-2)$ factor, vertices of degree zero or two don't contribute to the sum, and therefore the number of such vertices doesn't affect on the existence of giant component and doesn't alter the topology of the graph.

Let us compare the the properties of the generalized random graph with those of the ER random graph. This generalized random graph allows us to build a graph with a power-law degree distribution, while the

ER graph has Poisson degree distribution for large networks with variance . As a result, its variance is equal to average degree:

$$\sigma^2 = d = M_2 - d^2 \Rightarrow d_2 = d^2$$

Phase transition will occur at $d_2 = d$

$$d_2 = d^2 \Leftrightarrow d^2 = d \Leftrightarrow d = 1$$

This agrees with our previous calculations $d = 1$.

5.3.5 Average distance

Finally we will calculate the average distance for a graph. Assume the graph has a giant component, that covers most of the graph. Since the neighborhood grows exponentially with the distance from the vertex, most of the vertices are concentrated on some distance l from a vertex. Let d_l be an average number of neighbors at distance l. For large graphs $d_l \sim N$, hence :

$$N = 1 + \sum_{m=1}^{l} d_m = 1 + \sum_{m=1}^{l} (\frac{d_2}{d})^{m-1}$$

Assuming $N >> d$ and $d_2 >> d$ we can deduce:

$$l = \frac{log(\frac{N}{d})}{log\frac{d_2}{d}} + 1$$

The average distance between two nodes grows logarithmically on the graphs' size.

For a special case of the ER graph:

$$d_2 = d \Rightarrow l = \frac{logN}{logd}$$

Figure 17 shows that average distance estimation for generalized graphs are pretty close to the corresponding measures in real networks.

5.3.6 Clustering coefficient

Last, we will evaluate the probability that the i neighbor of vertex v has an edge with its j neighbor. Each of vertex v neighbors has degree distribution q_k. Its i-th neighbor has $k(i)$ 'out-going' edges and its j-th neighbor has $k(j)$ 'out-going' edges. Thus, the probability that two of its neighbors (i and j) are connected is: $\frac{k(i)k(j)}{Nd}$. Taking an average for all i, j we obtain :

$$C = \frac{1}{Nd}\left(\sum_k q_k k\right)^2 = \frac{d}{N}\left(\frac{M_2 - d}{d^2}\right)^2$$

where $\frac{d}{N}$ is a cluster coefficient for ER graph.

Let us analyze $(\frac{M_2 - d}{d^2})$ factor. This coefficient is dependent on the second moment M_2.
M_2 can be very large, and it is finite, if and only if $c > 2$. Knowing that c is typically at range (2,3) for power-law networks, we can estimate M_2 for $c < 3$.

$$M_2 \approx k_{max}^{3-c} \approx N^{\frac{3-c}{c-1}}$$

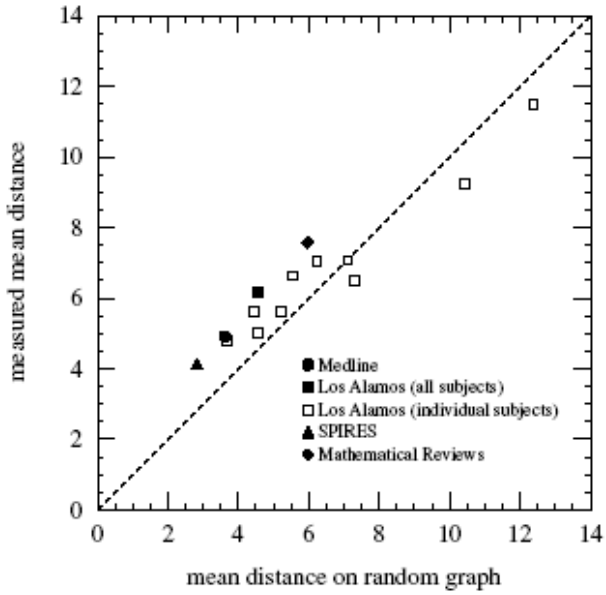


Figure 17: Source [10]. Estimated vs. real average distance for several real networks.

¿From cluster coefficient formula we obtain:

$$C \approx N^{-\beta}, \beta = \frac{3c-7}{c-1}$$

That leads to the following conclusions:

- When $c = 7/3$ there is a phase transition.
- When $c > 7/3$ - C tends to 0 with increasing of N
- When $c < 7/3$ - C increases with graph size

Thus, for a good random graph network model we expect cluster coefficient to be $2 < c < 3$. There is an improvement over the Erdös/Rényi random model: e.g for WWW estimate improves from 0.002 to 0.05 - only 2-fold away from the real coefficient of 0.11. Yet, real networks exhibit higher clustering coefficients than their random counterparts.

6 Summary

We have given an introduction to the use of random graphs as models of real-world networks and evaluated different properties for such a model. Most real networks have cluster coefficient $2 < c < 3$ and a giant component emerges when $c > 3.4788$. The average distance for such graphs is actually ultra-small: $O(\log \log N)$. Though there are certain drawbacks that differ random graphs from real networks (lack of network transitivity and unrealistic Poissonian degree distribution) this model is very popular and is the best studied model.

References

[1] Social networking in academia. http://www.orgnet.com/Erdos.html.

[2] R. Albert and A. L. Barabási. Emergence of scaling in random networks. *Science*, 286:509–512, 1999.

[3] E. A. Bender and E. R. Canfield. The asymptotic number of labeled graphs with given degree sequences. *J. Comb. Theory, Ser. A*, 24(3):296–307, 1978.

[4] S.N. Dorogovtsev and J.F.F. Mendes. *Evolution Of Networks - From Biological Nets to the Internet and WWW.* Oxford, 2002.

[5] P. Erdos and A. Renyi. On the evolution of random graphs. *MTA Mat. Kut. Int. K ozl.*, 5:17–61, 1960.

[6] N. Pruzlj et al. *Knowledge Discovery in Proteomics: Graph theory analysis of protein-protein interactions*. CRC Press, 2005.

[7] M. Faloutsos, P. Faloutsos, and C. Faloutsos. *On power-law relationships of the Internet topology*. ACM Press, 1999.

[8] H. Jeong, S. P. Mason, A. L. Barabási, and Z. N. Oltvai. Lethality and centrality in protein networks nature. *Nature*, 411:41–2, 2001.

[9] H. Jeong, B. Tombor, R. Albert, Z. N. Oltvai, and A. L. Barabási. The large-scale organization of metabolic networks. *Nature*, 407:651–654, 2000.

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

[11] S. H. Yook, Z. N. Oltvai, and A. L. Barabási. Functional and topological characterization of protein interaction networks. *Proteomics*, 4:928942, 2004.