341 Introduction to Bioinformatics: Biological Networks

18th of February, 2010

Network Models (continued)

In last class we covered the following three network models:

1 Erdos-Renyi Random Graphs

2 Generalized Random Graphs

3 Small-world Networks

Today we cover:

4 Scale-free Networks

Scale-free networks ("SF") have power-law degree distributions: $P(k) = k^{-\gamma}$, where $\gamma isapositivenumber$. Such degree distributions are observed in many biological networks, where usually $2 < \gamma < 3$. There are many ways to generate SF networks. We will cover three that are the most commonly used to model biological networks.

4.1 Preferential Attachment Model (Barabasi-Albert, 1999)

In this network model (henceforth denoted by "SF-BA"), nodes are added to an existing network and newly added nodes preferentially attach to existing nodes with probability proportional to the degree of the target node. This process is repeated until the network size matches the empirical data set it is trying to model. As nodes with a high number of edges probabilistically receive more edges through this algorithm, it is also known as "Rich getting richer". In this model, the starting network strongly influences the properties of the resulting network. An illustration is given in Figure 1.

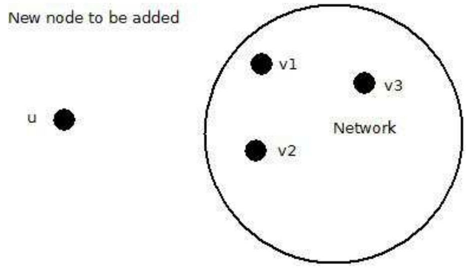

Figure 1: if $deg(v_1) > deg(v_2) > deg(v_3)$ then $p(uv_1 \in E) > p(uv_2 \in E) > p(uv_3 \in E)$

To generate this network ("growth process"):

1. Start with a small, connected network with N_0 nodes.
2. Add a node at each time step to the existing network and connect it's edges to an existing node with the probability proportional to the degree of the existing node. One standard method of doing this is to add edges to a node, v_i using this probability function: $\rho(v_i) = \frac{k_i}{\sum_{j=0}^{n} k_j}$. The newly introduced node is connected with $m \leq N_0$ edges to already present node, where N_0 is the starting number of nodes in the growth process.
3. Repeat 2 until the number of of nodes and the number of edges matches that of the empirical network to be modeled.

After t time steps, this network model has:

- $N_V(t) = N_0 + t$ nodes.
- $N_E(t) = mt$ edges.

SF-BA has a power-law degree distribution: $P(k) \propto k^{-\gamma_{\mathrm{BA}}}$ with $\gamma_{\mathrm{BA}} = 3$ that is invariant with time.

SF-BA is particularly effective at describing the Internet's link topologies - large sites tend to receive more links smaller ones resulting in this distribution.

4.2 Gene Duplication and Mutation (Vazquez et al., 2003)

Gene Duplication and Mutation scale-free network model ("SF-GD") is biologically motivated. It attempts to mimic gene duplication and mutation processes to influence the generation of the random network that still has scale-free degree distribution: the idea is that by mimicing gene duplication and mutation process that created biological networks, we might get a better network model for them. The network is generated in a similar manner to SF-BA: at each time step, a node is added to the network according to the following rules:

1. Start with a small, connected network.
2. Duplication:
   (a) Select node i at random.
   (b) A new node, i' is added and linked to all nodes that i is linked to; as such, the neighbourhoods of the two nodes are the same, i.e., $N(i) = N(i')$.
   (c) With probability p, add an edge between i and i'.
3. Divergence/mutation:
   (a) For each node j linked to nodes i and i' described above, choose one of the two links randomly and remove it with probability q.
4. Repeat steps 2 and 3 until the number of of nodes is similar to that of empirical network to be modeled.

The parameters of this gene duplication & mutation model are p and q.

Summary of network properties of this network model compared to network properties of real-world biological networks:

	Degree Distribution	Clustering Coefficient	Average Diameter
Real-World	Power Law	High	Small
SF-GD	Power Law	Low	Small
Comparison	✔	✗	✔

4.3 Hierarchical Scale-Free Networks

These networks attempt to preserve the degree distribution and network "modularity" via a fractal-like generation of the network.

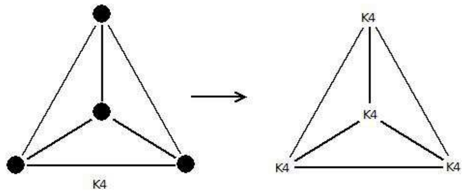

Figure 2: An example of how a simple hierarchical scale free network, based of the K_4 graph may be generated. Also see examples in ppt slides.

These graphs do not match any biological data and are highly unlikely to be found in data-sets unless engineered - one may liken them to crystal structures.

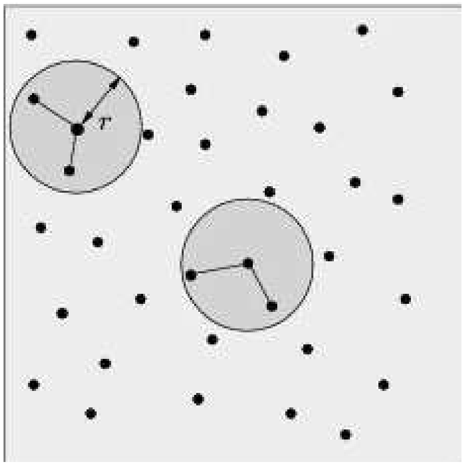

Figure 3: An illustration of constructing a geometric random graph. Source: http://www.nowpublishers.com/10.1561%5CNET13%5C1300000014%5Cnet01438x.gif

5 Geometric Random Graphs ("GEO")

GEO graphs are formed through the following method:

1. Take any metric space and, using a uniform random distribution, place nodes within it.

2. If any nodes are within radius r (calculated via any chosen distance norm for the space), they will be connected.

If we are modeling a real-world network $G(V, E)$ with $|V| = n$ and $|E| = m$, radius r in a GEO network must be selected so that the number of edges in the GEO graph is m, i.e., so that the size of the data and the GEO model network are the same.

There are many possible metric spaces (e.g., Euclidean space) and distance norms (e.g. the Euclidean distance (L_2), the Chessboard distance (L_∞), and the Manhattan/Taxi Driver distance (L_1)).

Summary of network properties of this network model compared to network properties of real-world biological networks:

	Degree Distribution	Clustering Coefficient	Average Diameter
Real-World	Power Law	High	Small
GEO	Poisson	High	Small
Comparison	-	✔	✔

In the geometric *random* graphs, the degree distribution follows a Poisson distribution. However, we can redefine the model, so that degree distribution becomes a power-law distribution. Also, gene duplications and mutations can be used to guide the growth process in geometric graphs. These concepts were explored in Przulj & Kuchaiev, Pacific Symposium on Biocomputing, 2009 & 2010.

5.1 Geometric Gene Duplication and Mutation ("GEO-GD")

This variant of geometric graph models, GEO-GD, attempts to model gene duplications and mutations as follows:

> Growth is governed by adding new nodes intended to model gene duplications and mutations, moderated by natural selection as follows. A duplicated gene starts at the same point in biochemical space as its parent, and then natural selection (which can be viewed as "evolutionary optimization") acts either to eliminate one, or cause them to slowly separate in the biochemical space.[1]

This variant also reproduces graphlet properties of the empirical dataset, which is another desirable feature, as well as these networks having power-law degree distributions.

6 Stickyness Index Based Model ("STICKY")

This is a biologically motived network model which is based on the notion of proteins interacting because they have complimentary binding domains. A stickyness index is a number based on the a protein's normalised degree in a PPI network and it is used to summarize the abundance and popularity of binding domains of the protein.

The basic **assumption** of this model is that a high degree protein has many binding domains. This assumption has certain issues - notably that some proteins would need to be huge to have such large binding domains as suggested by this assumption. In reality, the proteins will not be interacting with their bound proteins at all times, and the set of bound proteins at any one time may change. This is illustrated in Figure 4

A pair of proteins is more likely to interact under this model if both proteins have high stickyness indicies: the probability of an edge between two nodes is the product of those nodes' stickyness indicies.

	Degree Distribution	Clustering Coefficient	Average Diameter
Real World	Power Law	High	Small
"STICKY"	Matches data	Matches data	Matches data
Comparison	✔	✔	✔

[1] N. Przulj, O. Kuchaiev, A. Stevanovic, and W. Hayes, Geometric Evolutionary Dynamics of Protein Interaction Networks, Proceedings of the 2010 Pacific Symposium on Biocomputing (PSB), Big Island, Hawaii, January 4-8, 2010.

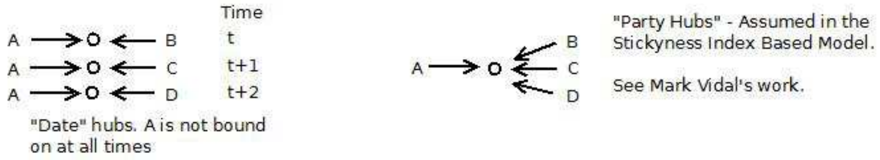
Time
A
B
t
A
C
t+1
A
D
t+2
"Date" hubs. A is not bound on at all times
"Party Hubs" - Assumed in the Stickyness Index Based Model.
See Mark Vidal's work.
A
B
C
D

Figure 4: A, B, C and D are different proteins

For further reading, see N. Przulj and Des Higham, Modelling Protein-Protein Interaction Networks via a Stickiness Index, Journal of the Royal Society Interface, volume 3, number 10, pages 711 - 716, 2006.

7 Interplay Between Network Topology and Biological Functions

7.1 Lethality and Centrality in PPI Networks (Jeong et al., Nature, 2001)

This study found the phenotypic consequence of singe gene deletion in yeast is affected by the topological position of it's protein product in the PPI network. This study was over a network of 1820 proteins and 2240 interactions - with a power law distribution (and clearly sparse).

The power law implies that the network is tolerant to random errors, but is intolerant to the removal of ”hubs” - the top degree-ranked nodes. When these hubs were removed, the network diameter increased rapidly leaving a less well connected network.

This study found that topology influences error tolerence: less connected nodes should be less essential than highly connected nodes. It also found that highly connected proteins have a central role in the network architecture and are 3 times more likely to essential than proteins of lower degrees.

7.2 Specificity and Stability in Topology of PPI networks (Maslov et al., Science, 2002)

This study made use of a yeast PPI network of 3278 proteins and 4549 interactions. This study explored the correlations in the connectivities of nodes by calculating the likelyhood, $p(k_0, k_1)$ that two proteins with degrees k_0 and k_1 are connected to each other. This study found that:

- There is a tendancy of highly connected nodes to interact with low degree nodes.

- There is a reduced likelyhood that a pair of hub nodes will interact with each other.
- There is a tendancy of proteins with degree between 4 and 9 to interact with each other (this seems to demonstrate that they belong to protein complexes).

During this study, the average connectivity, k_1 of neighbours of a node was calculated as a function of the degree of that node, k_0. This was used to find that k_1 shows a gradual decline with respect to k_0, i.e. degree correlation is negative.

The observed spectrum of degree of hub neighbours is consistent with the existence of functional modules which are organised around individual hubs - both the hub and it's neighbour hubs tend not to be connected directly.

This may imply network robustness through the suppression of the propigation of attacks over a network: if one hub is damaged, it is unlikely to affect all other hubs in the network. The reduced branching ratio around hubs provides a certain degree of protection against attacks of these nodes.

7.3 Gene Essentiality and the Topology of PPI networks (Coulomb et al., Proceedings of the Royal Society B, 2005)

In this paper, the following mutations were studied in the context of PPI network topology:

- **Lethal:** Single gene mutations which cause cell death.
- **Synthetically Lethal:** Combination of mutations in 2 genes causes cell death.
- **Viable:** Cell survives gene mutation.

They found that the strong correlation of gene essentiality and cell robustness in PPI networks was due to the use of PPI networks which had inherent biases in them. These included:

- Essential genes tend to be more studied than the viable genes - as such they may have inflated degrees (or conversely, the viable ones have suppressed degrees from being under-studied).
- The use of literature curation.
- The biotechnological biases previously discussed.

The study showed that the dispensibility of a gene is only weakly related to it's degree, suggesting that network topology has little influence on essentiality & robustness. More specifically, the average degree of essential and non-essential genes were 2.2 and 1.8 respectively - a difference factor of only 1.2. Similar results were found when analysing sythetically lethal and non-essential genes.

The main conclusions of this study were:

A: Physiological consequences of gene deletions are only weakly related to gene degrees in PPI networks.

B: k_1, the average degree of a node's neighbours, does not vary significantly between essential and non-essential genes, irrespective of their degree. This suggests that the essentiality of a gene does not seem to be rleated to the average degree of it's neighbours.

C: Clustering coefficients cannot be reliably associated with gene essentiality.

D: The average distance separating query genes from their synthetically lethal partners is similar to the average distance separating query genes from the set of non-essential genes: the distribution of these distances was found to be almost identical.

These conclusions are compatible with the hypothesis that the network topology is not under evolutionary constraints, but is instead a consequence of the construction process of the network. They are, however, at odds with the previous two studies.

7.4 Functional Topology in PPI Networks

This was a study of a yeast PPI network of 2401 proteins and 11000 interactions. The hetwork has a power law degree distribution.

Results of this study:

- Viable proteins were found to have degrees half that of lethal ones... although the interactions of the lethal genes tend to be studied more: they may be proportionally over-represented compraed to the viables in the network.

- Lethal proteins were found to be more frequent in the top 3% of nodes (ranked by degree) compared to viable nodes

- Lethals had a higher frequency in the group of proteins which were articulation points (AP's) [2] and hubs than did synthetically lethal and viable proteins.

- It was found that viable proteins tend to be on alternate pathways; this redundancy may explain why mutations of them were not lethal. This idea is demonstrated in figure 5, which shows that even though the grey node has been deleted, interactions still can take place through the other two paths from the top node to the bottom node.

[2] Articulation Points, or AP's are nodes, which if they are removed result in the disruption of a network's structure, i.e. part of the network becomes disconnected

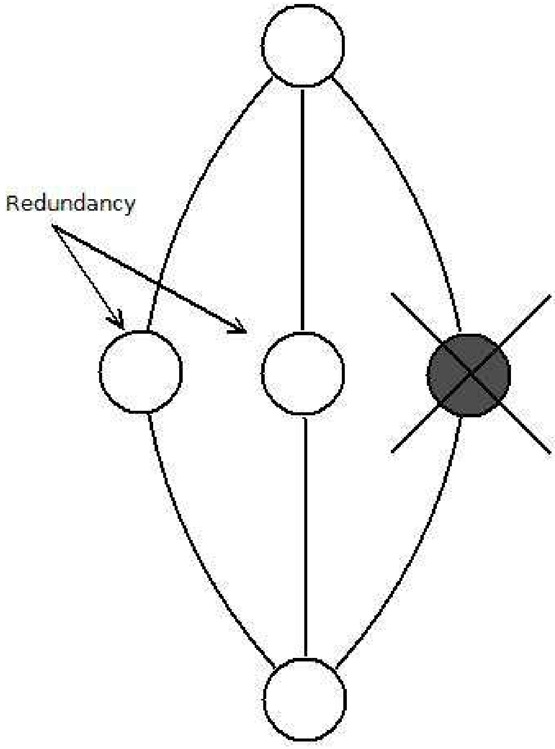

Figure 5: Redundancy in PPI networks