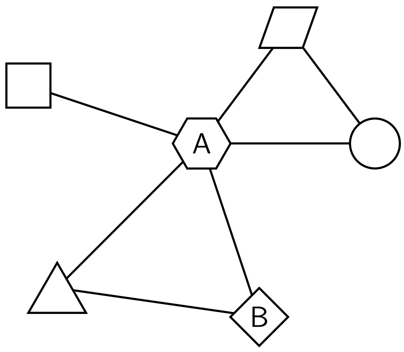
MS2a, Week 6

Rune Lyngsø

November 17, 2011

A Network Growth

Consider the network from [1, Fig. 3**b**]:



a. Assume the network grows as described in [1, Fig. 3**b**], by duplication of a gene (node) chosen uniformly at random. For the nodes marked A and B, respectively, compute the expected number of connections a random descendant has after a single duplication step, i.e. when the network has grown to seven nodes. How does the expected fraction of all network nodes A and B are connected to change with the update?

b. For node u let $P_u(i \mid n)$ denote the probability that a node u descendant is connected to exactly i other nodes when there are n nodes in the network. Write a recursion for $P_u(i \mid n)$, including boundary conditions.

c. Plot or tabulate $P_u(\cdot \mid n)$ for $n = 9$ for the node marked A in the above figure.

d. Let $c_{u,n} = \sum_{i=0}^{\infty} i P_u(i \mid n)/n$ denote the expected fraction of all nodes descendants of node u are connected to when there are n nodes in the network. Write an expression for $c_{u,n}$ in terms of $c_{u,6}$, *i.e.* the fraction of nodes u is connected to in the original network.

e. Let c_n denote the expected average fraction of nodes a node is connected to once the network has grown to n nodes, *e.g.* $c_6 = 7/18$. Write a recursion for c_n (boundary condition was just given). Are your results concordant with your results for $c_{u,n}$?

f. Connections between genes are not guaranteed eternal survival. Assume we extend the model such that in each round, after duplication of a gene, we choose a random edge uniformly at random and delete it. How will this affect the rich-get-richer nature of the model?

g. Loss of connections will usually result from evolutionary drift. If we assume that connection loss probability is positively correlated with evolutionary distance, is the connection loss model proposed above realistic?

B Network Stability

One measure of the importance of a node in a network is called the *Betweenness Centrality* (or BC from here on). It measures the number of pairs of other nodes that would have the shortest path connecting them disrupted if the node is eliminated from the network. More formally the BC of node v is

$$C_B(v) = \sum_{\substack{s \neq v \neq t \\ s \neq v}} \frac{\sigma_{s,t}(v)}{\sigma_{s,t}}$$

where $\sigma_{s,t}$ is the number of shortest paths connecting s to t and $\sigma_{s,t}(v)$ is the number of shortest paths connecting s to t that goes through v. There are several examples of biological networks where this can be viewed as a good measure of how crucial a node is for performance, e.g. in a regulatory network longer paths could introduce delays and in a metabolic network longer paths would usually result in increased overhead. We will use a slight modification of this measure,

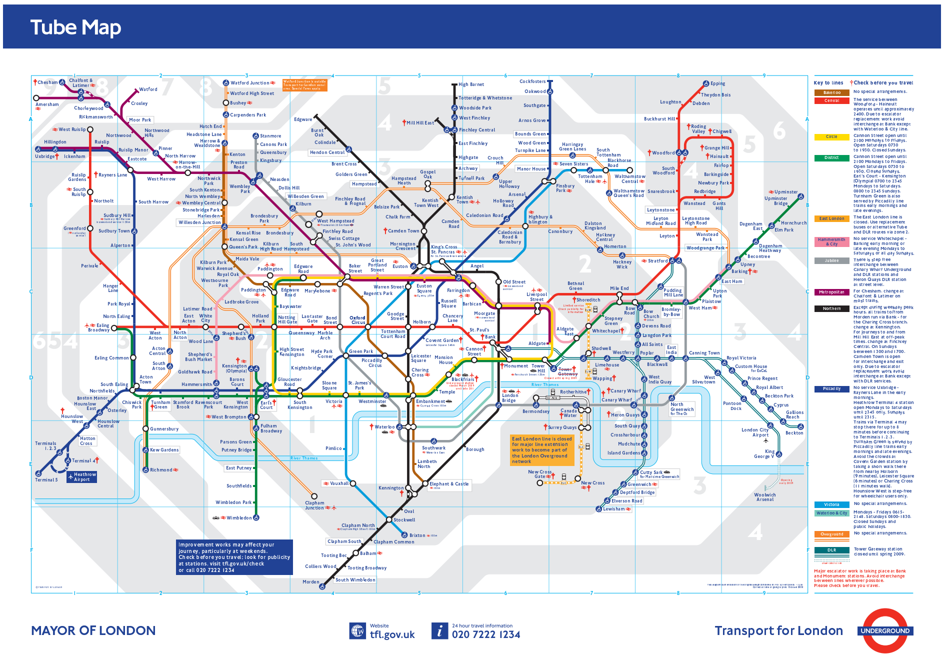
$$C_b(v) = \sum_{\substack{s \neq v \neq t \\ s \neq v}} \mathbb{1}_{\sigma_{s,t}(v) = \sigma_{s,t}} ,$$

i.e. we count the number of pairs that would see their shortest distance increased by the elimination of v.

h. Compute $C_b(v)$ for the nodes marked A and B in the network depicted in part A.

i. One can easily generalise the C_b score to sets, such that we count the number of pairs for which the shortest distance increases if all nodes in the set are eliminated. Design a network where the two element set with the highest C_b score is not the set of the two nodes with the highest C_b scores.

j. For transportation networks, the BC measure also captures how critical a node is. Which three stations on the map of the London Underground below would you guess have the highest C_b scores?



References

[1] A.-L. Barabási and Z. N. Oltvai. Network biology: Understanding the cell's functional organisation. *Nature Reviews Genetics*, 5:101–113, 2004.

[2] F. Jordán. Predicting target selection by terrorists: a network analysis of the 2005 london underground attacks. *International Journal of Critical Infrastructures*, 4(1/2):206–214, 2008.