

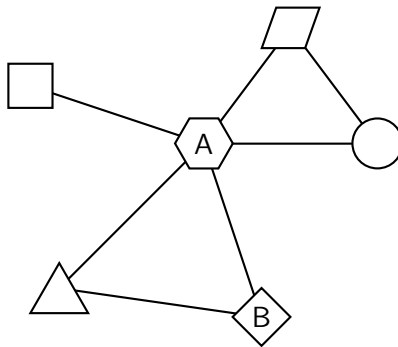
MS2a, Week 6, Model Solution

Rune Lyngsø

November 17, 2011

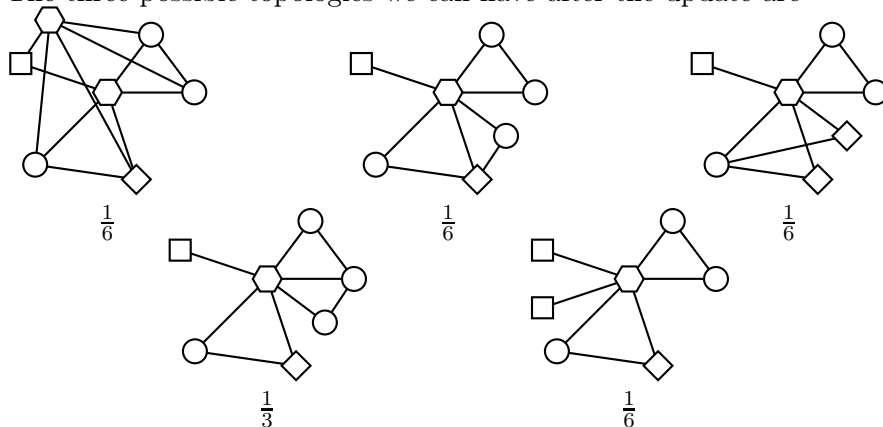
A Network Growth

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



- a. Assume the network grows as described in [1, Fig. 3b], 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.

The three possible topologies we can have after the update are



with probabilities as indicated. In the first case, descendants of A have 5 connections, in all other cases the descendant of A has 6 connections, yielding $5 \cdot 1/6 + 6 \cdot 5/6 = 35/6$ expected connections. Similarly, the descendant of B has 3 connections in the first two cases and 2 connection in all other cases, yielding $3 \cdot 1/3 + 2 \cdot 2/3 = 7/3$ connections.

How does the expected fraction of all network nodes A and B are connected to change with the update?

After the update there are 7 nodes, so descendants of A will be connected to $\frac{35/6}{7} = 5/6$ of all the network nodes in expectation, while descendants of B will be connected to $\frac{7/3}{7} = 1/3$ of all network nodes in expectation. This is exactly the same fraction that A and B are connected to in the original network.

- b. For node u let $P_u(i | n)$ 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 $P_u(i | n)$, including boundary conditions.

Boundary conditions are just the observed connectivities in the original network, so

$$P_u(i | 6) = \begin{cases} 1 & \text{if } u = A \text{ and } i = 5, u = \square \text{ and } i = 1, \text{ or } u \notin \{A, \square\} \text{ and } i = 2 \\ 0 & \text{otherwise} \end{cases}$$

Otherwise, the last duplication event either duplicated a node that u was connected to or a node that u was not connected to. In the first case, u must have been connected to $i - 1$ nodes before the duplication. In the second case, u must have been connected to i nodes before the duplication. These considerations give

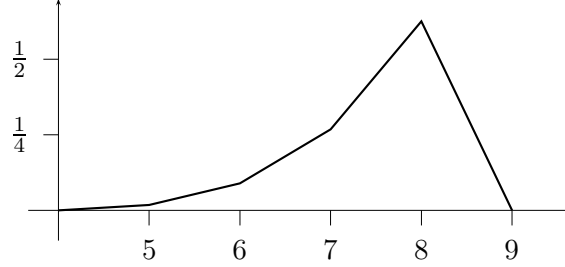
$$P_u(i | n) = \frac{i - 1}{n - 1} P_u(i - 1 | n - 1) + \frac{n - i - 1}{n - 1} P_u(i | n - 1)$$

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

Using the above expression, we get the following table of probabilities

		i				
		5	6	7	8	
for $n \leq 9$:	n	6	1	0	0	0
		7	1/6	5/6	0	0
		8	1/21	5/21	5/7	0
		9	1/56	5/56	15/56	5/8

Plotting this we get



- d. Let $c_{u,n} = \sum_{i=0}^{\infty} iP_u(i | n)/n$ 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.

$$\begin{aligned}
 c_{u,n} &= \sum_{i=0}^{\infty} \frac{i}{n} P_u(i | n) = \sum_{i=1}^{\infty} \frac{i}{n} P_u(i | n) \\
 &= \sum_{i=1}^{\infty} \frac{i}{n} \left(\frac{i-1}{n-1} P_u(i-1 | n-1) + \frac{n-i-1}{n-1} P_u(i | n-1) \right) \\
 &= \sum_{i=0}^{\infty} \frac{i+1}{n} \frac{i}{n-1} P(i | n-1) + \sum_{i=1}^{\infty} \frac{i}{n} P(i | n-1) - \sum_{i=1}^{\infty} \frac{i^2}{n(n-1)} P(i | n-1) \\
 &= \sum_{i=1}^{\infty} \frac{i(i+1) + i(n-1) - i^2}{n(n-1)} P(i | n-1) \\
 &= \sum_{i=1}^{\infty} \frac{i}{n-1} P(i | n-1) \\
 &= c_{u,n-1} = c_{u,6}
 \end{aligned}$$

- 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).

For n nodes, each node has an average expected nc_n connections, so there are expected $n^2c_n/2$ edges. The expected number of new edges added when going from n nodes to $n+1$ nodes is nc_n . Hence,

$$c_n = \frac{2(n^2c_n/2 + nc_n)}{(n+1)^2} = \left(1 - \frac{1}{(n+1)^2}\right) c_n$$

According to Maple, $c_n \rightarrow \frac{6}{7}c_6$ for $n \rightarrow \infty$.

Are your results concordant with your results for $c_{u,n}$?

The average connectivity of a random node in the network is thus decreasing slightly over time, while the connectivity of descendants of a particular node in the original network remains unchanged, no matter which node in the original network we consider. This may at first appear paradoxical. The explanation is that the number of descendants of a particular node is inversely correlated with the expected connectivity of those descendants: when we copy the descendants of a node in the original network, no new connections are added to any of the descendants of that original node. So descendant types that are more prevalent than we would expect will also have lower expected connectivity than when we do not condition on prevalence.

- 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?

The ‘rich’ nodes, i.e. the nodes with high connectivity, have more incident edges that can be randomly chosen by deletion. The chance that a node loses a connection is directly proportional to how many other nodes it is connected to. So choosing an edge at random to delete will tend to even out the connectivity, for a network where every node will tend towards the same connectivity in the limit.

- 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?

In [1] towards the end of the *Topological Robustness* section on page 110, it is stated that highly interacting proteins have a smaller evolutionary distance to orthologues in other organisms than other proteins. This means that connections to highly connected nodes should have less chance of being deleted than edges that connect two nodes with low connectivity. The proposed connection loss model assumes all edges are deleted with equal probability, so the model is not fully 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 .

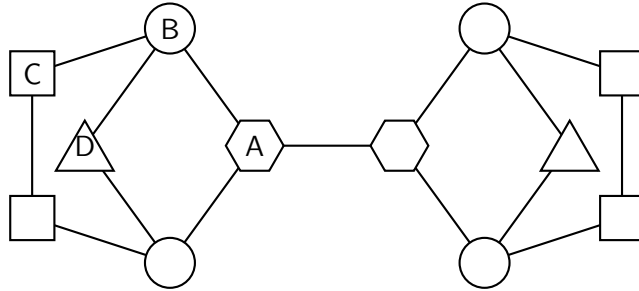
- h. Compute $C_b(v)$ for the nodes marked A and B in the network depicted in part A.

$C_b(A) = 8$ as the removal will break the network into three unconnected components of sizes 2, 2, and 1 where the shortest paths within each component are unaffected. So we only need to consider paths connecting nodes in different components, of which there are $\frac{2 \cdot 2(2+1) + 1 \cdot 4}{2} = 8$.

There are no shortest paths *passing through* B (evidently any shortest path connecting B with another node terminates at B), so removing B will not affect any of the shortest paths connecting the remaining $5 \cdot 4/2 = 10$ pairs of nodes. Hence, $C_b(B) = 0$.

- 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.

For the example to work, we need to have two nodes whose removal breaks the same large set of shortest paths. Consider the graph:

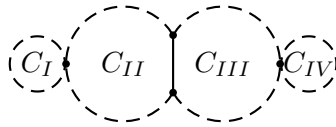


It has four types of nodes, as illustrated. It essentially has two separate components, with the only connection between the components being the edge between the two nodes of type A. So removing a node of type A breaks all paths connecting nodes in different components. However, removing a node of type A will not affect shortest paths within its component, or within the other component. So $C_b(A) = 30$.

Removing a node of type B breaks the shortest path connecting its type C neighbour with any node in the other component and the type A and D nodes in its own component, so $C_b(B) = 8$. Removing a type C node will only break the shortest path connecting its two neighbours, so $C_b(C) = 1$. Removing a type D node does not increase the shortest distance between any other pair of nodes, as the type A node in the same component can be used instead, so $C_b(D) = 0$. In conclusion, the two nodes with highest C_b score are the two type A nodes.

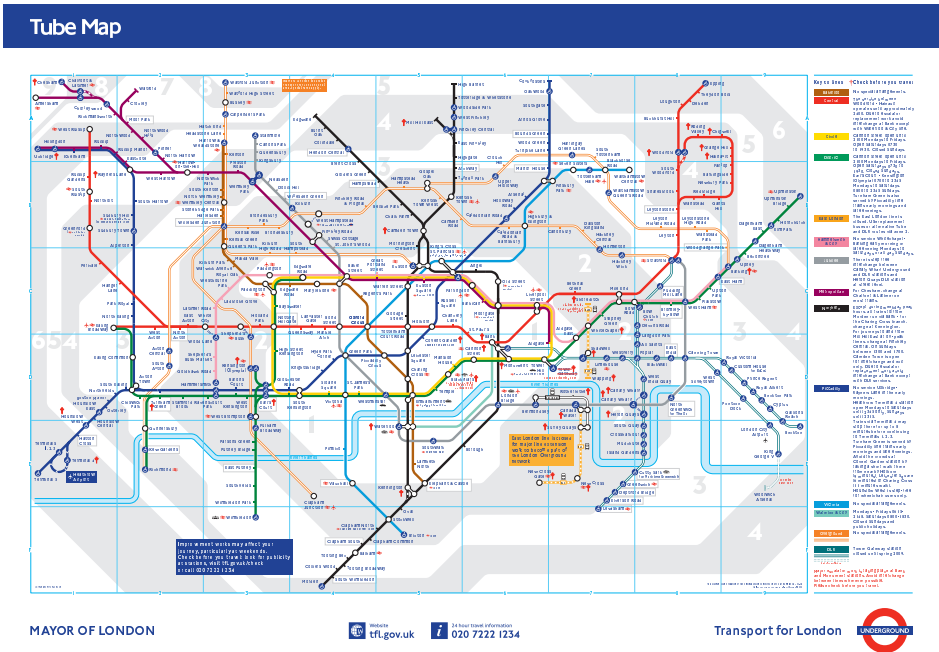
However, removing both type A nodes only increases the 30 inter-component distances. Within each component we can use the type D node to connect the two type B nodes with just two steps. So $C_b(\{A, A\}) = 25$. Removing just one type A node will increase the 25 inter-component distances and removing a type B node in the other component will further increase the distance between its neighbouring type C node and the type A and the type D node in this component. So $C_b(\{A, B'\}) = 27$.

An alternative approach would be to aim at constructing a network where the shortest paths between many pairs have to use one of two nodes. The removal of one still leaves a shortest path of the same length through the other, but removal of both breaks all shortest paths. Consider for example a network consisting of four cliques (a set of nodes where all pairs are connected by an edge) as sketched in



Clique C_I shares one node with clique C_{II} , clique C_{II} shares two nodes with clique C_{III} , and clique C_{III} shares one node with clique C_{IV} . Let cliques C_I and C_{IV} consist of n nodes each, and cliques C_{II} and C_{III} consist of $3n$ nodes each. The only nodes with non-zero C_b values are the two nodes shared between cliques C_I and C_{IV} and their neighbouring clique. Removing both will disrupt approximately $13n^2$ shortest paths, while disrupting the two nodes shared between cliques C_{II} and C_{III} will disrupt approximately $16n^2$ shortest paths.

- 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?



The full list of C_b scores, when changing from one line to another is: Amersham: 0 Brixton: 0 Chesham: 0 Cockfosters: 0 Edgware: 0 Elephant & Castle: 0 Epping: 0 Harrow & Wealdstone: 0 Heathrow Terminal 5: 0 High Barnet: 0 Kensington (Olympia): 0 Mill Hill East: 0 Morden: 0 Mornington

Crescent: 0 Richmond: 0 Stanmore: 0 Upminster: 0 Uxbridge: 0 Walthamstow Central: 0 Watford: 0 West Ruislip: 0 Wimbledon: 0 Northwood Hills: 5 Swiss Cottage: 5 Preston Road: 8 Willesden Green: 10 Northwood: 11 Dollis Hill: 14 Stamford Brook: 16 Arsenal: 22 Chiswick Park: 23 Hainault: 29 Grange Hill: 33 Elephant and Castle: 35 Ravenscourt Park: 35 Northwick Park: 47 West Kensington: 50 Goodge Street: 58 Borough: 118 Covent Garden: 126 Bromley-by-Bow: 146 Stepney Green: 150 Marylebone: 169 St John's Wood: 172 Holloway Road: 197 Russell Square: 199 Neasden: 229 Whitechapel: 235 Charing Cross: 252 Shepherd's Bush Market: 257 Pinner: 258 Leicester Square: 259 Kilburn: 261 Goldhawk Road: 264 Blackhorse Road: 266 Burnt Oak: 266 Canons Park: 266 Croyley: 266 Heathrow Terminals 1, 2, 3: 266 Hillingdon: 266 Kenton: 266 Kew Gardens: 266 Lambeth North: 266 Oakwood: 266 Ruislip Gardens: 266 South Wimbledon: 266 Theydon Bois: 266 Totteridge and Whetstone: 266 Upminster Bridge: 266 Wimbledon Park: 266 Fairlop: 274 Cannon Street: 279 Monument: 284 Chigwell: 286 Aldgate: 326 Wood Lane: 336 Bow Road: 353 Tower Hill: 354 Regent's Park: 367 Mansion House: 368 Ealing Broadway: 424 Piccadilly Circus: 430 Caledonian Road: 438 Aldgate East: 442 Old Street: 471 Latimer Road: 482 Blackfriars: 490 West Acton: 506 North Harrow: 512 West Hampstead: 516 Barkingside: 518 Colindale: 530 Colliers Wood: 530 Debden: 530 Gunnersbury: 530 Heathrow Terminal 4: 530 Hornchurch: 530 Ickenham: 530 Queensbury: 530 South Kenton: 530 South Ruislip: 530 Southfields: 530 Southgate: 530 Tottenham Hale: 530 Woodside Park: 530 Chalfont & Latimer: 531 Angel: 533 Roding Valley: 538 Ladbroke Grove: 668 Temple: 681 Newbury Park: 766 Arnos Grove: 792 Chorleywood: 792 East Putney: 792 Elm Park: 792 Hatton Cross: 792 Hendon Central: 792 Kingsbury: 792 Loughton: 792 North Wembley: 792 Northolt: 792 Ruislip: 792 Seven Sisters: 792 Tooting Broadway: 792 West Finchley: 792 Westbourne Park: 861 Canning Town: 876 Oval: 887 St Paul's: 926 Park Royal: 944 St James's Park: 956 Chancery Lane: 969 North Ealing: 978 Alperton: 1007 Gants Hill: 1014 North Greenwich: 1018 Tottenham Court Road: 1041 Bounds Green: 1052 Brent Cross: 1052 Buckhurst Hill: 1052 Dagenham East: 1052 Greenford: 1052 Hounslow West: 1052 Putney Bridge: 1052 Rickmansworth: 1052 Ruislip Manor: 1052 Tooting Bec: 1052 Wembley Central: 1052 Royal Oak: 1067 Embankment: 1126 Sudbury Town: 1131 Canary Wharf: 1196 Kennington: 1196 Redbridge: 1267 Sudbury Hill: 1278 Holborn: 1283 Balham: 1310 Dagenham Heathway: 1310 Eastcote: 1310 Golders Green: 1310 Hounslow Central: 1310 Parsons Green: 1310 Perivale: 1310 Stonebridge park: 1310 Wood Green: 1310 Finchley Central: 1314 Vauxhall: 1345 Canada Water: 1380 South Harrow: 1431 Ealing Common: 1434 Queensway: 1501 Wanstead: 1519 Pimlico: 1521 Becontree: 1566 Bermondsey: 1566 Clapham South: 1566 East Finchley: 1566 Fulham Broadway: 1566 Hampstead: 1566 Hanger Lane: 1566 Harlesden: 1566 Hounslow East: 1566 Turnpike Lane: 1566 Lancaster Gate: 1604 Southwark: 1656 Wembley Park: 1736 Marble Arch: 1737 High Street Kensington: 1741 Belsize Park: 1820 Clapham Common: 1820 Highgate: 1820 Manor House: 1820 Osterley: 1820 Upney: 1820 West Brompton: 1820 Willesden Junction: 1820 Wood-

ford: 2070 Archway: 2072 Barking: 2072 Boston Manor: 2072 Chalk Farm: 2072 Clapham North: 2072 Kensal Green: 2072 Moor Park: 2086 London Bridge: 2088 Bayswater: 2101 East Acton: 2116 Sloane Square: 2194 North Acton: 2224 South Woodford: 2279 White City: 2300 East Ham: 2322 Northfields: 2322 Queen's Park: 2322 Tufnell Park: 2322 Shepherd's Bush: 2499 Snaresbrook: 2523 Kentish Town: 2570 Kilburn Park: 2570 South Ealing: 2570 Upton Park: 2570 Stockwell: 2642 Knightsbridge: 2678 Holland Park: 2715 West Harrow: 2789 Hyde Park Corner: 2803 Maida Vale: 2816 Plaistow: 2816 Rayners Lane: 2921 Barbican: 3052 Warwick Avenue: 3060 Warren Street: 3092 Finsbury Park: 3093 Farringdon: 3158 Highbury & Islington: 3207 West Ham: 3213 Victoria: 3364 Acton Town: 3695 Bank: 3724 Turnham Green: 4014 Bond Street: 4216 Oxford Circus: 4326 Notting Hill Gate: 4337 Barons Court: 4403 Leytonstone: 4510 Hammersmith: 4518 Westminster: 4618 Moorgate: 4659 Leyton: 4712 Edgware Road: 4989 Waterloo: 5006 Stratford: 5021 Camden Town: 5039 Euston Square: 5275 Great Portland Street: 5313 Gloucester Road: 5461 South Kensington: 5521 Harrow-on-the-Hill: 5642 Bethnal Green: 5660 Paddington: 5875 Mile End: 6158 Euston: 6565 Earl's Court: 6791 Green Park: 7209 Liverpool Street: 7305 Finchley Road: 7656 King's Cross St Pancras: 9978 Baker Street: 11641

The most crucial stations are thus Baker Street, King's Cross St Pancras, and Finchley Road. An analysis of the possibility of a rational choice of targets in the 7/7 bombings is presented in [2].

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.