COMS 6998-006 Network Theory **Columbia University, Spring 2008**

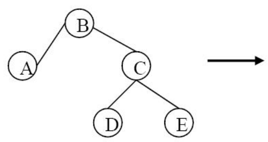
Lecture 5 – Properties of Networks: February 7, 2008

Lecturer: Dragomir Radev *Scribe: Anjali Koppal*

1 Properties of Networks

1.1 Mapping Graphs to Matrices

- Make a square matrix of dimension $|V| * |V|$ and fill in a 1 at position (i, j) if there is an edge between nodes i and j.
- Undirected graphs would map to symmetric matrices (edge A-B = edge B-A). If self-loops are disallowed, the diagonal will always have 0s. An example is shown in Figure 1



	A	B	C	D	E
A		1			
B	1		1		
C		1		1	1
D			1		
E			1		

Figure 1: A pictorial representation of mapping a graph to a matrix.

1.1.1 Mapping Properties of Matrices to Properties of Graphs

- Co-citation: Two papers B and C are co-cited if they are cited by the same paper A. We can count the number of co-citations for two papers B and C easily using matrices. Say the representative matrix is M and the ith row is B and the jth row is C. Then, the number of citations is simply the inner product between the ith and jth row. Figure 2 shows pictorially, an example of a graph depicting two papers being cited by one, and how the resulting co-citation value is computed.

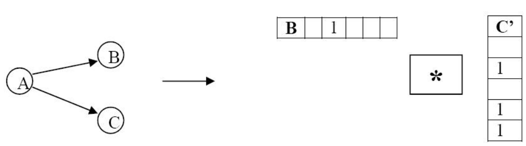


Figure 2: Pictorial Representation of Co-citation

- Bibliographic coupling: Two papers B and C are bibliographically coupled if they cite the same paper A (see Figure 3). The number of bibliographical couplings for papers i and j are the number of rows k such that M_{ki} and M_{kj} are equal to 1.
- Degree: It is easy to compute the degree of a graph by simply looking at its corresponding adjacency matrix (degree(i) = number of 1s in the ith row).
- Number of paths of length n: The number of paths between nodes i and j of length $n = A^n_{ij}$

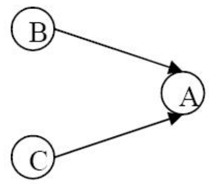


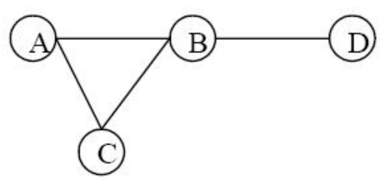
Figure 3: Pictorial Representation of Bibliographic Coupling

- Total number of paths between nodes i and j of all lengths: Sum up number of paths of all lengths : $\sum_{k=0}^{\infty} A_{jk}^{k}$
- Number of loops of length k: Sum up the values on the diagonal of the matrix (aka the trace) and then divide by k to avoid overcounting (otherwise we'll be counting a loop A->B->C->A as unique from B->C->A->B etc). In fact, it turns out the trace of a matrix is equal to the sum of the eigen-values (for square matrices). So the number of loops is dominated by the largest eigenvalue.
- Geodesic Path: The geodesic path between 2 nodes i and j is the the shortest path between i and j, I_{ij}.
- Diameter: The diameter of a graph is the largest shortest path between any two nodes on a graph, ie $max_{i,j} I_{ij}$

1.2 Assortativity: General Idea

- A measure of correlation. The questions we are trying to answer are:
  - do high degree nodes tend to connect with other high degree nodes? (positive correlation)
  - do high degree nodes tend to connect with low degree nodes?
- You can have either positive correlation ($\sigma^2 = 1$), negative correlation ($\sigma^2 > 0$), or random ($\sigma^2 < 0$)

1.2.1 Assortativity: Example



	Neighbours	Neighbours of neighbours
A	A-B	B-D, B-C, B-A (3)
	A-C	C-B, C-A (2)
B	B-A	A-B, A-C (2)
	B-C	C-A, C-B (2)
	B-D	D-B (1)
C	C-A	A-B, A-C (2)
	C-B	B-A, B-C, B-D (3)
D	D-B	B-A, B-C, B-D (3)

Figure 4: A worked out example of assortativity.

Let's look at one expression in the summation for Figure 4, specifically for A when

$$
\begin{aligned}
k_1 &= 2 \\
k_2 &= 2 \\
P(k_1, k_2) &= \frac{\#(\#neighbours = 2 \; and \; \#neighbours \; of \; neighbours = 2)}{\#(all \; values \; of \; neighbours \; and \; neighbours \; of \; neighbours)} \\
&= \frac{2}{16} \\
P(k_1) &= \frac{\#(\#neighbours = 2)}{\#(all \; values \; of \; neighbours)} \\
&= \frac{2}{4} \\
P(k_2) &= \frac{\#(\#neighbours \; of \; neighbours = 2)}{\#(all \; values \; of \; neighbours)} \\
&= \frac{4}{8}
\end{aligned} \tag{1}
$$

expression: 2 * 2 [(2/16) - (2/4)*(4/8)]

1.2.2 Newman Data [4]

Generalization: sociological networks seem to have a positive assortavity; technological networks have a negative assortavity.

1.3 Centrality Measures

- Intuitively, it is a measure of how important a node is. In the first example (Figure 5, left), it is natural to believe the middle node is central because it has the maximum degree. But in the second example, even though the middle node has a smaller degree than the other nodes, it is somehow more important to the graph.

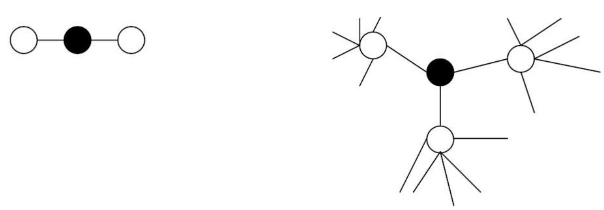


Figure 5: An intuitive notion of centrality. The shaded node on the left is central. The shaded node on the right is central even though its degree is smaller than other nodes

1.3.1 Degree Centrality

- Central node = node with maximum indegree
- Central node = node with maximum outdegree

1.3.2 Betweenness Centrality

- It measures how important a node is by counting the number of shortest paths that it is a part of.
- In Figure 6 # paths between B and D: 2
  remove C -> there will still be a path between B and D. So C is not that important that path.
  remove B -> all paths to A are broken => B is quite central.
  The matrix of shortest paths is shown in 7

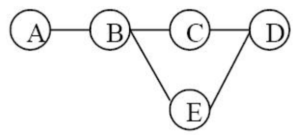


Figure 6: Betweenness Centrality example

	A	**B**	**C**	**D**	**E**
A	-	A-B	A-B-C	A-B-C-D A-B-E-D	A-B-E
B	B-A	-	B-C	B-C-D B-E-D	B-E
C	C-B-A	C-B	-	C-D	C-D-E C-B-E
D	D-C-B-A D-E-B-A	D-C-B D-E-B	D-C	-	D-E
E	E-B-A	E-B	E-B-C E-D-C	E-D	-

Figure 7: Matrix of shortest paths from the example.

$$
\begin{aligned}
Betweenness\,Centrality(C) &= (fraction-paths-broken(A,B)\\
&+ fraction-paths-broken(A,C)\\
&+ fraction-paths-broken(A,D)\\
&+ fraction-paths-broken(A,E)\\
&+ fraction-paths-broken(B,C)\\
&+ fraction-paths-broken(B,D)\\
&+ fraction-paths-broken(B,E)\\
&+ fraction-paths-broken(C,D)\\
&+ fraction-paths-broken(C,E)\\
&+ fraction-paths-broken(D,E))/10\\
&= (0+1+0.5+0+1+0.5+0+1+1+0)/10 = 0.5
\end{aligned} \tag{2}
$$

Other observations:
Betweenness centrality(E) = Betweenness centrality(C) [symmetry]
Betweenness centrality(B) > Betweenness centrality(C) [B turns out to be the center]

1.3.3 Graph-theoretic Centrality

$c = argmin_i(max_j d(i,j))$

1.3.4 Closeness Centrality

Mean geodesic distance between a node and all its reachable nodes. Here, a smaller value implies the node is more central. So in practice, the inverse is used as a numerical measure of centrality.

1.3.5 Eigenvector Centrality

Measure the importance of a node by the importance of its neighbours.
Eg: pagerank: measures the importance of a web page (see Figure 8).

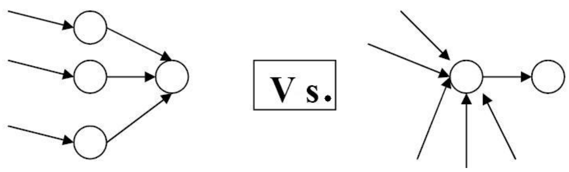


Figure 8: Pagerank: is a page important if it is linked many pages who are each linked by few, or if it is linked by few pages with a lot of links.

1.3.6 h-index

- A popular method to measure the centrality of academic papers.
- h-index = number of your papers h that have been cited at least h times.
- Example
  P: 10, 4, 1,
  Person P has 4 papers that have been cited 10, 4, 1 and no papers.
  h-index = 2 (2 papers have been cited at least twice).
- h-index has been implemented:
  Software: [1]
  ACL Anthology Network: [3]

1.4 Readings

- [4], pages 1-11 (required) and pages 12-19 (optional)
- [2], Optional

References

[1] *h-number (or h-index)*. URL: `http://www.brics.dk/~mis/hnumber.html`.

[2] K.Kilkki. "A Practical Model for Analyzing Long Tails". In: *First Monday* 12 (2007).

[3] University of Michigan. *The ACL Anthology Network*. URL: `http://tangra.si.umich.edu/clair/anthology`.

[4] M.E.J Newman. "Random Graphs as Models of Networks". In: *ArXiv Condensed Matter e-prints* (2002).