Centrality
Heat (1995)
Is counting the edges enough?
Stanford Social Web (ca. 1999)

network of personal homepages at Stanford
different notions of centrality

In each of the following networks, X has higher centrality than Y according to a particular measure.

- indegree
- outdegree
- betweenness
- closeness
review: indegree
trade in petroleum and petroleum products, 1998, source: NBER-United Nations Trade Data
Q: high indegree

- Which countries have high indegree (import petroleum and petroleum products from many others)
  - Saudi Arabia
  - Japan
  - Iraq
  - USA
  - Venezuela
review: outdegree
trade in petroleum and petroleum products, 1998
source: NBER-United Nations Trade Data
Q: low outdegree

- Which country has low outdegree but exports a significant quantity (thickness of the edges represents $$\text{value of export}$$) of petroleum products
  - Saudi Arabia
  - Japan
  - Iraq
  - USA
  - Venezuela
trade in **crude** petroleum and petroleum products, 1998, source: NBER-United Nations Trade Data
Undirected degree, e.g. nodes with more friends are more central.

Assumption: the connections that your friend has don't matter, it is what they can do directly that does (e.g. go have a beer with you, help you build a deck...)

putting numbers to it
divide degree by the max. possible, i.e. (N-1)
How much variation is there in the centrality scores among the nodes?

Freeman’s general formula for centralization (can use other metrics, e.g. gini coefficient or standard deviation):

$$C_D = \frac{\sum_{i=1}^{g} [C_D(n^*) - C_D(i)]}{[(N - 1)(N - 2)]}$$

centralization: skew in distribution
degree centralization examples

\[ C_D = 1.0 \]

\[ C_D = 0.167 \]

\[ C_D = 0.167 \]
example financial trading networks

real-world examples

high in-centralization: one node buying from many others

low in-centralization: buying is more evenly distributed
In what ways does degree fail to capture centrality in the following graphs?
Stanford Social Web (ca. 1999)

network of personal homepages at Stanford
Brokerage not captured by degree
Constraint

REALLY?

HEY, I DO BUSINESS ONLY WITH YOU.
constraint

YOU GUYS ARE MY FAVORITES!
intuition: how many pairs of individuals would have to go through you in order to reach one another in the minimum number of hops?
Betweenness: definition

\[ C_B(i) = \sum_{j<k} g_{jk}(i) / g_{jk} \]

Where \( g_{jk} = \) the number of shortest paths connecting \( jk \)
\( g_{jk}(i) = \) the number that actor \( i \) is on.

Usually normalized by:

\[ C_B'(i) = C_B(i) / [(n - 1)(n - 2)/2] \]

number of pairs of vertices excluding the vertex itself
Betweenness on toy networks

- non-normalized version:
Betweenness on toy networks

- non-normalized version:

A lies between no two other vertices
B lies between A and 3 other vertices: C, D, and E
C lies between 4 pairs of vertices (A,D),(A,E),(B,D),(B,E)

note that there are no alternate paths for these pairs to take, so C gets full credit
Betweenness on toy networks

- non-normalized version:
Betweenness on toy networks

- non-normalized version:
  - why do C and D each have betweenness 1?
  - They are both on shortest paths for pairs (A,E), and (B,E), and so must share credit:
    - \( \frac{1}{2} + \frac{1}{2} = 1 \)
Q: betweenness

- What is the betweenness of node E?
Lada’s old Facebook network: nodes are sized by degree, and colored by betweenness.
Q: high betweenness, low degree

- Find a node that has high betweenness but low degree
Find a node that has low betweenness but high degree.
What if it’s not so important to have many direct friends?
Or be “between” others
But one still wants to be in the “middle” of things, not too far from the center
need not be in a brokerage position
Closeness is based on the length of the average shortest path between a node and all other nodes in the network.

Closeness Centrality:

\[ C_c(i) = \left( \sum_{j=1}^{N} d(i,j) \right)^{-1} \]

Normalized Closeness Centrality

\[ C'_c(i) = (C_c(i))/(N - 1) \]
Closeness: toy example

\[ C'_c(A) = \left[ \sum_{j=1}^{N} \frac{d(A, j)}{N - 1} \right]^{-1} = \left[ \frac{1 + 2 + 3 + 4}{4} \right]^{-1} = \left[ \frac{10}{4} \right]^{-1} = 0.4 \]
Closeness: more toy examples
Q: high degree, low closeness

Which node has relatively high degree but low closeness?
How central you are depends on how central your neighbors are

\[ C(i) = w_{ij} \cdot C(j) + w_{ki} \cdot C(k) + w_{li} \cdot C(l) \]
Bonacich eigenvector centrality

\[ c_i(\beta) = \sum_j (\alpha + \beta c_j) A_{ji} \]

\[ c(\beta) = \alpha(I - \beta A)^{-1} A1 \]

- \( \alpha \) is a normalization constant
- \( \beta \) determines how important the centrality of your neighbors is
- \( A \) is the adjacency matrix (can be weighted)
- \( I \) is the identity matrix (1s down the diagonal, 0 off-diagonal)
- \( 1 \) is a matrix of all ones.
small $\beta \rightarrow$ high attenuation
  only your immediate friends matter, and their importance is factored in only a bit

high $\beta \rightarrow$ low attenuation
  global network structure matters (your friends, your friends' of friends etc.)

$\beta = 0$ yields simple degree centrality

$$c_i(\beta) = \sum_j (\alpha )A_{ji}$$
If $\beta > 0$, nodes have higher centrality when they have edges to other central nodes.

If $\beta < 0$, nodes have higher centrality when they have edges to less central nodes.
Why does the middle node have lower centrality than its neighbors when $\beta$ is negative?
Centrality in directed networks

- WWW
- food webs
- population dynamics
- influence
- hereditary
- citation
- transcription regulation networks
- neural networks
We now consider the fraction of all directed paths between any two vertices that pass through a node

\[ C_B(i) = \sum_{j,k} g_{jk}(i) / g_{jk} \]

Only modification: when normalizing, we have \((N-1)*(N-2)\) instead of \((N-1)*(N-2)/2\), because we have twice as many ordered pairs as unordered pairs

\[ C'_B(i) = C_B(i) / [(N-1)(N-2)] \]
A node does not necessarily lie on a geodesic (shortest path) from $j$ to $k$ if it lies on a geodesic from $k$ to $j$. 

Directed geodesics
Directed closeness centrality

- choose a direction
  - in-closeness (e.g. prestige in citation networks)
  - out-closeness

- usually consider only vertices from which the node \( i \) in question can be reached
Eigenvector centrality in directed networks

- PageRank (centrality) brings order to the Web:
  - it's not just the pages that point to you, but how many pages point to those pages, etc.
  - more difficult to artificially inflate centrality with a recursive definition

An important page, e.g. slashdot

If a web page is slashdotted, it gains attention

Many webpages scattered across the web
A random walker following edges in a network for a very long time will spend a proportion of time at each node which can be used as a measure of importance.
Problem with pure random walk metric:
- Drunk can be “trapped” and end up going in circles
Ingenuity of the PageRank algorithm

- Allow drunk to teleport with some probability
  - e.g. random websurfer follows links for a while, but with some probability teleports to a “random” page (bookmarked page or uses a search engine to start anew)
Example: probable location of random walker after 1 step

20% teleportation probability

PageRank

slide adapted from: Dragomir Radev
Example: probable location of random walker after 10 steps

Slide from: Dragomir Radev
Q: PageRank’s damping factor

- What happens to the relative PageRank scores of the nodes as you increase the teleportation probability (decrease the damping factor)?
  - they equalize
  - they diverge
  - they are unchanged

PageRank.nlogo part of the built-in suite of network models for NetLogo
Centrality
- many measures: degree, betweenness, closeness, eigenvector
- may be unevenly distributed
  - measure via distributions and centralization
- in directed networks
  - indegree, outdegree, PageRank
- consequences:
  - benefits & risks (Baker & Faulkner)
  - information flow & productivity (Aral & Van Alstyne)
Some applications
(time permitting)
Hospital patient transfer network
Infection prevention strategies in a hospital patient transfer network

- random
- degree
- betweenness
- greedy

probability of becoming infected within a year from a random starting hospital

0 1/6 1/3
Identifying expertise

- The Response Time Gap

- The Expertise Gap
- Difficult to infer reliability of answers

Automatically ranking expertise may be helpful.

Zhang, Ackerman, Adamic, WWW’07
Java Forum

- 87 sub-forums
- 1,438,053 messages
- Community expertise network constructed:
  - 196,191 users
  - 796,270 edges
Thread 1: Large Data, binary search or hashtable? user A
Re: Large... user B
Re: Large... user C

Thread 2: Binary file with ASCII data user A
Re: File with... user C
Uneven participation

- ‘answer people’ may reply to thousands of others
- ‘question people’ are also uneven in the number of repliers to their posts, but to a lesser extent

\[ \alpha = 1.87 \text{ fit, } R^2 = 0.9730 \]

Cumulative probability

Number of people one replied to

Number of people one received replies from
Not Everyone Asks/Replies

The Web is a bow tie

The Java Forum network is an uneven bow tie

- Core: A strongly connected component, in which everyone asks and answers
- IN: Mostly askers.
- OUT: Mostly Helpers
fragment of the Java Forum
Human-rated expertise levels

- 2 raters
- 135 JavaForum users with >= 10 posts
- inter-rater agreement ($\tau = 0.74$, $\rho = 0.83$)
- for evaluation of algorithms, omit users where raters disagreed by more than 1 level ($\tau = 0.80$, $\rho = 0.83$)

<table>
<thead>
<tr>
<th>L</th>
<th>Category</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>Top Java expert</td>
<td>Knows the core Java theory and related advanced topics deeply.</td>
</tr>
<tr>
<td>4</td>
<td>Java professional</td>
<td>Can answer all or most of Java concept questions. Also knows one or some sub topics very well,</td>
</tr>
<tr>
<td>3</td>
<td>Java user</td>
<td>Knows advanced Java concepts. Can program relatively well.</td>
</tr>
<tr>
<td>2</td>
<td>Java learner</td>
<td>Knows basic concepts and can program, but is not good at advanced topics of Java.</td>
</tr>
<tr>
<td>1</td>
<td>Newbie</td>
<td>Just starting to learn java.</td>
</tr>
</tbody>
</table>
Algorithm Rankings vs. Human Ratings

simple local measures do as well (and better) than measures incorporating the wider network topology
automated vs. human ratings
Modeling expertise network formation

Control Parameters:
- Distribution of expertise
- Who asks questions most often?
- Who answers questions most often?
  - best expert most likely
  - someone a bit more expert

ExpertiseNet Simulator
suppose:

- expertise is uniformly distributed
- probability of posing a question is inversely proportional to expertise
  \[ p_{ij} = \text{probability a user with expertise } j \text{ replies to a user with expertise } i \]

2 models:

- **‘best’ preferred**
  \[ p_{ij} \sim e^{\beta(j-i)} / i \]

- **‘just better’ preferred**
  \[ p_{ij} \sim e^{\gamma(i-j)} / i \quad j>i \]
Visualization

Best “preferred”  just better
Degree correlation profiles

Java Forum Network

best preferred (simulation)

just better (simulation)
Algorithm selection

Preferred Helper: ‘best available’

Preferred Helper: ‘just better’
In the ‘just better’ model, a node is correctly ranked by PageRank but not by HITS.
Node centrality can reveal the relative importance of nodes within the network
Choose a measure appropriate to the question you are asking