Web Mining

Web Structure Mining

Prof. Dr. Christian Bizer
Prof. Dr. Simone Ponzetto

FSS 2017
**Web Structure Mining**

- **Definition**
  
  Discovery and interpretation of patterns in
  1. the hyperlink structure of the Web
  2. the social ties among actors that interact on the Web

- **Typical sources of structure data**
  1. Web crawls including HTML pages and hyperlinks
  2. crawls of the blogosphere
  3. social networks including explicit relations between actors (your Facebook friend network)
  4. other types of community data (discussion forums, email conversations, …)

- **Focuses on the structure, but can of course also be combined with content or usage mining techniques**
Social Media Analysis

Social Networking

Content Sharing

Blogs Micro-blogging

Wiki Forum

Social Media

facebook

go+

LinkedIn

delicious

social bookmarking

flickr

YouTube

twitter

WordPress

Epinions.com

Wikipedia

The Free Encyclopedia
Chapter Outline

1. Describing Networks
   1. Basic Terminology and Measures

2. Community Detection
   1. Connected Components
   2. K-Cores and Cliques
   3. Clustering-based Techniques

3. Prominence
   1. Centrality
   2. Prestige

4. Patterns in Large Real-World Networks
   1. Power-Law Distributions
   2. Small-World Phenomenon
Literature


Lots of thanks to Lada Adamic, Albert-László Barabási, Lei Tang and Huan Liu for the original versions of the slides.
Software that we will use in the Exercise

- Tool for the analysis and visualization of large networks
- Download: http://pajek.imfm.si/doku.php?id=pajek
A Graph is a collection of vertices that are connected by lines.

**Network** often refers to real systems

**Graph**: mathematical representation of a network

**But often**: “Network” ≡ “Graph”

<table>
<thead>
<tr>
<th>Community</th>
<th>Points</th>
<th>Lines</th>
</tr>
</thead>
<tbody>
<tr>
<td>Math</td>
<td>vertices</td>
<td>lines: edges, arcs</td>
</tr>
<tr>
<td>Computer Science</td>
<td>nodes</td>
<td>links</td>
</tr>
<tr>
<td>Physics</td>
<td>sites</td>
<td>bonds</td>
</tr>
<tr>
<td>Sociology</td>
<td>actors</td>
<td>ties, relations</td>
</tr>
</tbody>
</table>
Directed and Undirected Graphs

**Undirected**

Lines: undirected (*symmetrical*) ➔ *edge*

Graph:

```
Undirected edges:
• co-authorship links
• roads (mostly)
```

**Directed**

Lines: directed ➔ *arcs*

Digraph = directed graph:

```
Directed arcs:
• hyperlinks on the WWW
• following on Twitter
• phone calls
```
Vertex, Arc and Edge Attributes

Vertices, arcs and edges can have attributes.

Example of a network with vertex and arc attributes:
- girls’ school dormitory dining-table partners (Moreno, *The sociometry reader*, 1960)
- first and second choices shown
Degree: the number of edges connected to the vertex.

In directed graphs we can define an in-degree and out-degree. The (total) degree is the sum of in- and out-degree.

Source: a vertex with $k_{in} = 0$ and $k_{out} > 0$

Sink: a vertex with $k_{out} = 0$ and $k_{in} > 0$
Degree Distribution

Summarizes the degrees of all vertices.

Alternative representations:
1. A frequency count of the vertices of each degree
2. $P(k)$: probability that a randomly chosen vertex has degree $k$

$P(k) = \frac{N_k}{N}$
In-Degree Distribution of the WDC Hyperlink Graph

Covers 3.5 billion web pages and 128 billion hyperlinks, extracted from Common Crawl 2012

Displayed on log-log scale, meaning that left third covers over 99% of the mass.

Top In-Degree Websites

The Common Crawl WWW Ranking

Here you can browse a ranking of more than 100 million sites of the World Wide Web. Every single step leading to this ranking is open and accessible. Enjoy!

Learn more »

<table>
<thead>
<tr>
<th>Harmonic centrality</th>
<th>Indegree centrality</th>
<th>Katz's index</th>
<th>PageRank</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>1. wordpress.org</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>1</td>
<td>2. youtube.com</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>23</td>
<td>3. gmpg.org</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>4. en.wikipedia.org</td>
<td>4</td>
<td>6</td>
</tr>
<tr>
<td>39</td>
<td>5. tumblr.com</td>
<td>5</td>
<td>7</td>
</tr>
<tr>
<td>3</td>
<td>6. twitter.com</td>
<td>6</td>
<td>5</td>
</tr>
<tr>
<td>4</td>
<td>7. google.com</td>
<td>7</td>
<td>9</td>
</tr>
<tr>
<td>6</td>
<td>8. flickr.com</td>
<td>8</td>
<td>14</td>
</tr>
<tr>
<td>172870</td>
<td>9. rralalabel.org</td>
<td>9</td>
<td>59</td>
</tr>
<tr>
<td>75</td>
<td>10. wordpress.com</td>
<td>10</td>
<td>30</td>
</tr>
<tr>
<td>2431063</td>
<td>11. mp3shake.com</td>
<td>11</td>
<td>44</td>
</tr>
</tbody>
</table>

http://wwwranking.webdatacommons.org/
Out-Degree Distribution of WDC Hyperlink Graph

Displayed on log-log scale.

Maximal out-degrees are much smaller than maximal in-degrees.

Strange shapes are SPAM networks.
Average Degree

\[
\langle k \rangle = \frac{1}{N} \sum_{i=1}^{N} k_i, \quad \langle k \rangle = \frac{2L}{N}
\]

N – the number of vertices in the graph

L – the number of lines in the graph

\[
\langle k^\text{in} \rangle = \frac{1}{N} \sum_{i=1}^{N} k^\text{in}_i, \quad \langle k^\text{out} \rangle = \frac{1}{N} \sum_{i=1}^{N} k^\text{out}_i, \quad \langle k^\text{in} \rangle = \langle k^\text{out} \rangle
\]

Warning: Might be meaningless because of power-law like degree distributions.
Graph Density

- Maximal number of the connections that may exist between vertices:
  - directed graph
    \[ L_{\text{max}} = N \times (N-1) \]
    since each of the \( N \) vertices can connect to \( (N-1) \) other vertices
  - undirected graph
    \[ L_{\text{max}} = \frac{N \times (N-1)}{2} \]
    since edges are undirected, count each one only once

- What fraction is present?

  \[ \text{Density} = \frac{L}{L_{\text{max}}} \]

  - For example, out of 12 possible connections, this graph has 7, giving it a density of
    \[ \frac{7}{12} = 0.583 \]
Clustering Coefficient

- Density of connections among one’s friends.
- What portion of your neighbors are connected?

\[ C_i = \begin{cases} \frac{k_i}{d_i \times (d_i-1)/2} & d_i > 1 \\ 0 & d_i = 0 \text{ or } 1 \end{cases} \]

- \( k_i \) number of edges (undirected) among \( V_i \)'s neighbors
- \( d_i \) degree of vertex \( V_i \)

- \( C_i \) in \([0,1]\)
Example: Clustering Coefficient

$C_i = \begin{cases} \frac{k_i}{d_i \times (d_i-1)/2} & d_i > 1 \\ 0 & d_i = 0 \text{ or } 1 \end{cases}$

$d_6 = 4$, $N_6 = \{4, 5, 7, 8\}$

$k_6 = 4$ as $e(4,5)$, $e(5,7)$, $e(5,8)$, $e(7,8)$

$C_6 = 4/(4*3/2) = 2/3$

Average Clustering Coefficient

$C = (C_1 + C_2 + \ldots + C_n) / N$

$C = 0.61$ for the left network

In a random graph, the expected average clustering coefficient is $0.19$

⇒ graph has some community structure
**Adjacency Matrix**

\[ A_{ij} = \begin{cases} 1 & \text{if there is a link between vertices } i \text{ and } j \\ 0 & \text{if vertices } i \text{ and } j \text{ are not connected to each other.} \end{cases} \]

\[ A_{ij} = \begin{pmatrix} 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 1 & 1 & 1 & 0 \end{pmatrix} \quad \text{and} \quad A_{ij} = \begin{pmatrix} 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 \end{pmatrix} \]

Note that for a directed graph (right) the matrix is not symmetric.
Adjacency Matrix and Node Degrees

**Undirected**

Adjacency Matrix:

\[ A_{ij} = \begin{pmatrix} 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 1 & 1 & 1 & 0 \end{pmatrix} \]

Degree:

\[ k_i = \sum_{j=1}^{N} A_{ij} \]

Number of lines:

\[ L = \frac{1}{2} \sum_{i=1}^{N} k_i = \frac{1}{2} \sum_{ij} A_{ij} \]

**Directed**

Adjacency Matrix:

\[ A_{ij} = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 \end{pmatrix} \]

Degree:

\[ k_{ij}^\text{in} = \sum_{i=1}^{N} A_{ij} \]

\[ k_{ij}^\text{out} = \sum_{j=1}^{N} A_{ij} \]

Number of lines:

\[ L = \sum_{i=1}^{N} k_{ij}^\text{in} = \sum_{j=1}^{N} k_{ij}^\text{out} = \sum_{i,j} A_{ij} \]
Graphology

Unweighted / Simple
(undirected)

\[ A_{ij} = \begin{pmatrix}
0 & 1 & 1 & 0 \\
1 & 0 & 1 & 1 \\
1 & 1 & 0 & 0 \\
0 & 1 & 0 & 0 \\
\end{pmatrix} \]

Weighted
(undirected)

\[ A_{ij} = \begin{pmatrix}
0 & 2 & 0.5 & 0 \\
2 & 0 & 1 & 4 \\
0.5 & 1 & 0 & 0 \\
0 & 4 & 0 & 0 \\
\end{pmatrix} \]

Example: Road networks (distance in miles)
Graphology

Self-Interactions

Multigraph
(undirected)

Example: Webpages that may link to themselves

Example: Social networks with different types of interactions

$$A_{ij} = \begin{pmatrix} 1 & 1 & 1 & 0 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 \end{pmatrix}$$

$$A_{ij} = \begin{pmatrix} 0 & 2 & 1 & 0 \\ 2 & 0 & 1 & 3 \\ 1 & 1 & 0 & 0 \\ 0 & 3 & 0 & 0 \end{pmatrix}$$
Bipartite graph (or bigraph) is a graph whose vertices can be divided into two disjoint sets $U$ and $V$ such that every line connects a vertex in $U$ to one in $V$; that is, $U$ and $V$ are independent sets.

Examples:

- movie/actor network
- disease/symptom network
- photo/tag network on Flickr
Vertex Attributes

- vertices are described by attributes in many real-world networks
  - e.g. social network with vertex attributes name, birthdate, address, interests, …
- combining these attributes with measures such as degree often reveals interesting insights
  - e.g. number of friends on Facebook in relation to age and location

A path is a sequence of vertices in which each vertex is adjacent to the next one.

\[ P_{i_0, i_n} \] of length \( n \) between vertices \( i_0 \) and \( i_n \) is an ordered collection of \( n+1 \) vertices and \( n \) lines.

\[ P_{i_0, i_n} = \{i_0, i_1, i_2, \ldots, i_n\} \]

\[ P_{i_0, i_n} = \{(i_0, i_1), (i_1, i_2), (i_2, i_3), \ldots, (i_{n-1}, i_n)\} \]

- A path can intersect itself and pass through the same line repeatedly. Each time a line is crossed, it is counted separately.
- A legitimate path on the graph on the right:
  \[ \text{A B C B C A D E E B A} \]
- In a directed graph, the path can follow only the direction of the arcs.
The distance (shortest path, geodesic path) between two vertices is defined as the number of edges along the shortest path connecting them.

If the two nodes are disconnected, the distance is infinity.

In directed graphs each path needs to follow the direction of the arcs. Thus in a digraph the distance from vertex A to B (on an AB path) is generally different from the distance from vertex B to A (on a BCA path).
Network Diameter and Average Distance

**Diameter**

\( d_{\text{max}} \) the maximum distance between any pair of nodes in the graph.

**Caution:** Some people use the term ‘diameter’ to be the average shortest path length.

**Average Distance / Average Shortest Path Length**

\(<d>\) for a connected graph:

\[
\langle d \rangle = \frac{1}{N(N-1)} \sum_{i,j} d_{ij}
\]

where \( d_{ij} \) is the distance from vertex \( i \) to vertex \( j \) and \( N \) is the number of vertices.

- The average shortest path length distinguishes an easily navigable network from one which is complicated and inefficient (i.e. for information or mass transport).
Exercise: Characterizing a Graph

Please calculate the following measures for the graph below:

1. Degree distribution
2. Clustering coefficients $C_2$ and $C_3$
3. Diameter $d_{\text{max}}$
4. Optional: Average shortest path length $<d>$
2. Community Detection

A **community** is a set of actors between which interactions are (relatively) frequent.

- Communities are also called groups, cohesive subgroups, clusters, or modules in different contexts.
- Finding a community in a social network is to identify a set of nodes such that they interact with each other more frequently than with those nodes outside the group.

- Applications: Recommendation based on communities, network compression, visualization of huge networks
Subjectivity of Community Definition

A densely-knit community?
Each component is a community?

Perception of community is subjective/task dependent (unsupervised learning)
2.1 Graph Connectivity and Components

**Connected graph:** Any two vertices of an undirected graph are joined by a path.

A disconnected graph is made up by two or more **connected components**.

**Bridge:** if we remove it, the graph becomes disconnected (red line).

**Cut vertex:** if we remove it, the graph becomes disconnected (F and A).
Giant Component and Isolates

- If the largest component encompasses a significant fraction of the graph, it is called the **giant component**.
- The other smaller components are called **isolates**.
Connectivity of Directed Graphs

Strongly connected directed graph: has a path from each vertex to every other vertex and vice versa (e.g. AB path and BA path).

Weakly connected directed graph: is connected if we disregard the arc directions.
Connected Components of Directed Graphs

- **Strongly Connected Component (SCC)**
  Maximal subgraph in which every vertex can be reached from every other vertex by following directed arcs. Interesting for web crawling.

  - Strongly connected components
    - B C D E
    - A
    - G H
    - F

- **Weakly Connected Component (WCC)**
  Maximal subgraph in which every vertex can be reached from every other vertex by following lines in either direction.

  - Weakly connected components
    - A B C D E
    - G H F
Exercise: Components

How many strongly connected components has the graph below?
Example: Strong Components and Vertex Attributes

- Contours: Strong Components
- Numbers: Family-Friendship-Grouping
Four major parts (Broder et al., WWW2000)

- **Central Strongly Connected Component (SCC)**
  - pages that can reach one another along hyperlinks
  - about 30% of the Web (normal pages)

- **IN Group**
  - can reach SCC via directed path but cannot be reached from it
  - about 20% (maybe new pages or boring ones)

- **OUT Group**
  - can be reached from SCC but cannot reach it
  - about 20% (maybe company pages that don’t link)

- **Tendrils**
  - cannot reach SCC and cannot be reached by it
  - about 20%

- **Unconnected**
  - about 10%
Bow-Tie and SCCs in WDC Hyperlink Graph

- IN much larger than OUT: 31% vs. 6%
- LSCC much larger compared to Broder et al.: 51% vs. 30%
- The Web has become more connected in the last 14 years

2.2 K-Cores

A K-Core is a maximal undirected subgraph in which each vertex has at least degree k within the subgraph.

- A k-core does not need to be connected. The one 3-core above consists of two parts.
- K-cores are nested, meaning every vertex in the 3-core also belongs to the 2-core.
K-Cores and Communities

- A vertex is displayed belonging to the highest k-core of which it is a part.

- Finding communities: Remove lower cores until the graph breaks up into multiple components that form cohesive subgroups.
Exercise: K-Cores

Count the number of 3-cores in the graph.

Vertex colors indicate the level of $k$:
- White: $k=1$
- Light gray: $k=2$
- Dark gray: $k=3$
- Black: $k=4$
2.3 Cliques

A clique is the maximal complete subgraph containing three vertices or more.

- Cliques are the strongest form of community as all vertices need to be directly adjacent to each other.

- Problems:
  - The definition is often too restrictive to detect communities in real-world networks (result: large number of very small cliques)
  - Finding cliques in larger graphs is computationally expensive.
Overlapping Complete Subgraphs

- Alternative less restrictive approach: Consider overlapping complete subgraphs (triads) as cohesive group.

- Triads overlap if they share one or more vertices.

- Finding overlapping complete subgraphs is supported by Pajek.
Procedure for the Analysis of Cohesive Subgroups

- Idea: Successively increase the minimal density required.
- After each step: Check if discovered groups make sense based on your knowledge about the application domain.
2.4 Divisive Hierarchical Clustering

- **Goal:** Build a *hierarchical structure of communities based on the network topology.*
  - Allows the analysis of a network at different resolutions.

- **Approach:** *Recursively remove the “weakest” edge*
  1. Find the edge with the least strength.
  2. Remove the edge and update the corresponding strength of each edge.
  3. Recursively apply the above two steps until a network is decomposed into desired number of components.
  4. Consider each component to form a community.
The strength of an edge can be measured by edge betweenness.

**Edge betweenness:** The number of shortest paths that pass along the edge.

The edge betweenness of $e(1, 2)$ is 4 ($=6/2 + 1$), as all the shortest paths from 2 to {4, 5, 6, 7, 8, 9} have to either pass $e(1, 2)$ or $e(2, 3)$, and $e(1, 2)$ is the shortest path between 1 and 2.

The edge with higher betweenness tends to be the bridge between two communities.
Divisive Clustering based on Edge Betweenness

Initial betweenness values

Table 3.3: Edge Betweenness

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>4</td>
<td>1</td>
<td>9</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
<td>0</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>4</td>
<td>0</td>
<td>9</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>9</td>
<td>0</td>
<td>9</td>
<td>9</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>10</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>6</td>
<td>3</td>
</tr>
<tr>
<td>6</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>10</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>6</td>
<td>3</td>
</tr>
<tr>
<td>7</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>6</td>
<td>6</td>
<td>0</td>
<td>2</td>
<td>8</td>
</tr>
<tr>
<td>8</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>9</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>8</td>
<td>0</td>
</tr>
</tbody>
</table>

After remove $e(4,5)$, the betweenness of $e(4,6)$ becomes 20, which is the highest;

After remove $e(4,6)$, the edge $e(7,9)$ has the highest betweenness value 4, and should be removed.
2.5 Islands

- Community detection method for weighted graphs, e.g.
  - number of interactions on a social network
  - number of times author A cited author B
  - overall duration of phone calls

An island is a maximal subnetwork of vertices connected directly or indirectly by lines with a value greater than the lines to vertices outside the subnetwork.
Backbone of the WDC Hyperlink Graph

Websites connected by more than 500,000 links. The colors refer to DMOZ categories.

http://webdatacommons.org/hyperlinkgraph/2012-08/topology.html
2.6 Graph Partition

A partition of a graph is a classification or clustering of the vertices in the graph such that each vertex is assigned to exactly one class of cluster.

- Partitions are a central concept in Pajek.
- Can be generated based on
  1. existing vertex attributes.
  2. applying traditional classification or clustering algorithms to vertex attributes.
  3. analyzing the graph structure.
  4. combinations of all the above.
Partitioning of the WDC Hyperlink Graph by TLD

Visualization of the amount of links within and between partitions.

TLD = Top-Level-Domain

http://webdatacommons.org/hyperlinkgraph/topology.html
3. Prominence

Who are the “most important” actors in a social network?

Centrality
- A central actor is one involved in many edges.
- The direction of lines is not considered.

Prestige
- A prestigious actor is one who is the target of many arcs.
- The direction of arcs is considered.
3.1 Centrality

- Which nodes are most ‘central’?
  - Calculated for undirected graph

- Definition of ‘central’ varies by context / purpose:

- Local measure:
  - Degree centrality

- Relative to rest of network:
  - Closeness centrality
  - Betweenness centrality

- How evenly is centrality distributed among nodes?
  - Centralization
  - graph-level view
Degree Centrality

Idea: Measure centrality as the number of edges to other vertices in the graph.

Answers the question: How many people can a person directly reach or influence?

Degree Centrality

$$C_D(n_i) = d(n_i)$$

- Focuses only on direct choices (path length=1)

Normalized Degree Centrality

$$C'_{D}(n_i) = d(n_i) / N-1$$

- Degree divided by the maximal possible degree, i.e. number of vertices – 1
- Proportion of all nodes that are adjacent to $n_i$
Examples: Normalized Degree Centrality $C'_D$
Freeman’s general formula for centralization:

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

1. calculate the sum of differences in centrality between the most central vertex in a graph and all other vertices;
2. divide this quantity by the theoretically largest sum of differences in any graph of the same degree (star shape graph).

Value Range $[0,1]$
Examples: Degree Centralization

\[ C_D = 1.0 \]

\[ C_D = 0.167 \]

\[ C_D = 0.167 \]
Examples: Degree Centralization

Financial trading networks

high centralization: one node trading with many others

low centralization: trades are more evenly distributed
When degree isn’t everything

In what ways does degree fail to capture centrality in the following graphs?

In what contexts may degree be insufficient to describe centrality?

1. Ability to broker between groups
2. Likelihood that information originating from anywhere in the network reaches you
Betweenness Centrality

Intuition: How many pairs of individuals would have to go through you in order to reach one another in the minimum number of hops?

Assumptions:

- Interactions between two non-adjacent actors might depend on the other actors in the set of actors, especially the actors who lie on the paths between the two nodes.
- “Actor in the middle” between the others has some control over paths in the network – “interpersonal influence”.

Who has higher betweenness centrality, X or Y?
Betweenness Centrality: Definition

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

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

Usually normalized by dividing through maximal theoretical value for \( C'_b(i) \):

\[ C'_B(i) = C_B(i) / \left[ \frac{(n - 1)(n - 2)}{2} \right] \]

paths are symmetrical

number of vertices without the vertex itself

number of pairs of vertices excluding the vertex itself = shortest paths for each vertex
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.
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:

\[
\frac{1}{2} + \frac{1}{2} = 1
\]

Can you figure out why B has betweenness 3.5 while E has betweenness 0.5?
1. Can you spot nodes with high betweenness but relatively low degree?

2. Explain how this might arise.

- vertices are sized by degree centrality and
- colored by betweenness centrality
Closeness Centrality

- The measure focuses on how close an actor is to all the other actors in the network.
  - for instance to spread information or interact with others
  - or to be reached by information that spreads through the network

- Closeness centrality is based on the length of the average shortest path between a vertex and all vertices in the graph.

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

Closeness Centrality:

\[
C'_c(i) = C_c(i)(N - 1)
\]

Normalized Closeness Centrality:

star shape: each vertex has distance one to central vertex
Example: Closeness Centrality

\[ C'_c(A) = \left( \frac{\sum_{j=1}^{N} d(A, j)}{N - 1} \right)^{-1} \]

More toy examples:
Correlation of Centrality Metrics

- Generally different centrality metrics will be positively correlated.
- When they are not, there is likely something interesting about the vertex.

**closeness centrality** denoted by color
**degree** denoted by size

**betweenness centrality** denoted by color
**degree** denoted by size
3.2 Prestige

Prestige refers to a class of prominence metrics which take the direction of arcs into account.

- Translates to: choices received
- Examples where direction matters:
  - Votes in an election
  - Hyperlinks on the WWW
  - Citations of scientific papers
- Examples when ‘prestige’ may not be the right word
  - dislikes
  - distrusts
The simplest vertex-level measure of prestige (in-degree)

The idea is that actors who are prestigious tend to receive many nominations or choices

- a paper that is cited by many others has high prestige
- a person nominated by many others for a reward has high prestige

Local measure as only the neighbors are taken into account.

Degree Prestige / Popularity

\[ P_D(n_i) = d_{in}(n_i) \]

Normalized Degree Prestige

\[ C_D'(n_i) = \frac{d_{in}(n_i)}{N-1} \]

Indegree divide by the maximal possible indegree

Proportion of all nodes that choose \( n_i \)
Input Domain

- Degree prestige only counts actors who are adjacent to actor $n_i$, but we might also want to take **indirect choices** into account.

The **Input Domain** of a vertex in a directed network is the number or percentage of all other vertices that are connected by a path to this vertex.

- Also called **Influence domain**, which makes for instance sense for the use case of following on Twitter.
Proximity Prestige

A Prestige measure based on distances in the input domain.

- Direct nominations (choices) should count more than indirect ones
- Nominations from second degree neighbors should count more than third degree ones

\[ P_p(n_i) = \frac{\text{fraction of all vertices that are in } i\text{'s input domain}}{\text{average distance from } i \text{ to vertex in input domain}} \]

\[ P_P(n_i) = \frac{(I_i / (g - 1))}{\sum d(n_i, n_j) / I_i} \]

Example:
Rank Prestige and Page Rank

**Rank Prestige**
- Prestige measure which considers the prestige of the actors who do the “choosing”.
- You are more prestigious if you have lots of other prestigious people in your input domain.

\[
P_R(i) = \sum_{(j,i) \in E} P_R(j)
\]

**Page Rank**
- Variation of rank prestige in which the prestige of a voting node is shared between all link targets.

\[
P_{PR}(i) = \sum_{(j,i) \in E} P_{PR}(j) / D_{out}(j)
\]

- Advantages of PageRank in the search context
  - hard to trick with SPAM links
  - The score is independent of actual search engine query

**Calculation of PageRank Score:** See Bing Lui: Web Data Mining. Chapter 7.3
Indegree and PageRank approximately correlate.

### Top Websites according to PageRank

<table>
<thead>
<tr>
<th>Harmonic centrality</th>
<th>Indegree centrality</th>
<th>Katz's index</th>
<th>PageRank</th>
</tr>
</thead>
<tbody>
<tr>
<td>23</td>
<td>3</td>
<td>3</td>
<td>1. gmpg.org</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>1</td>
<td>2. wordpress.org</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>2</td>
<td>3. youtube.com</td>
</tr>
<tr>
<td>98</td>
<td>21</td>
<td>33</td>
<td>4. livejournal.com</td>
</tr>
<tr>
<td>3</td>
<td>6</td>
<td>6</td>
<td>5. twitter.com</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
<td>4</td>
<td>6. en.wikipedia.org</td>
</tr>
<tr>
<td>99</td>
<td>5</td>
<td>5</td>
<td>7. tumblr.com</td>
</tr>
<tr>
<td>56711</td>
<td>50</td>
<td>52</td>
<td>8. promodj.com</td>
</tr>
<tr>
<td>4</td>
<td>7</td>
<td>7</td>
<td>9. google.com</td>
</tr>
<tr>
<td>394</td>
<td>54</td>
<td>56</td>
<td>10. networkadvertising.org</td>
</tr>
<tr>
<td>99</td>
<td>20</td>
<td>32</td>
<td>11. phpbb.com</td>
</tr>
<tr>
<td>19798</td>
<td>192</td>
<td>214</td>
<td>12. ytmnd.com</td>
</tr>
<tr>
<td>9715</td>
<td>18</td>
<td>19</td>
<td>13. milbelan.gov.cn</td>
</tr>
<tr>
<td>6</td>
<td>8</td>
<td>8</td>
<td>14. flickr.com</td>
</tr>
</tbody>
</table>

Large-scale networks in the real world often demonstrate similar patterns:

1. Power-law distributions
2. Small-world phenomenon
Power-Law Distributions

- Degree distribution in large-scale networks often follows a power law.

\[ p(x) = Cx^{-\alpha}, \quad x \geq x_{\text{min}}, \quad \alpha > 1 \]

- The preferential attachment process (Barabási and Albert, 1999) explains power-law distributions: Vertices prefer to link to vertices with a high degree.

- Translates to “The rich get richer” or “The famous get more famous”.
Examples of Power Law Distributions

- Power law distribution becomes a straight line if plot in a log-log scale.
- Power laws are often hypothesized by simply looking at the plot.
Better Methods for detecting Power Laws

1. Fibonacci binning (shows differences to actual counts)
2. Goodness-of-Fit Tests (using for instance the plfit tool)
Small-world networks are characterized by

1. **high average clustering coefficient**
   - which indicates strong community structures.
   - Explanation: Friends of a friend are likely to be friends as well.

2. **small average shortest path length**
   - which is also known as “six degrees of separation” (Travers and Milgram, 1969)
   - Explanation: *Hub*- or bridge-vertices interconnect communities and help shortening the average path length.
   - *Hub*- or bridge-vertices can be identified using betweenness centrality
Small-World Properties of Social Networks

<table>
<thead>
<tr>
<th>Network</th>
<th>C</th>
<th>Ratio to Random Graphs Erdös-Rényi</th>
<th>Power-Law</th>
</tr>
</thead>
<tbody>
<tr>
<td>Web [2]</td>
<td>0.081</td>
<td>7.71</td>
<td>-</td>
</tr>
<tr>
<td>Flickr</td>
<td>0.313</td>
<td>47,200</td>
<td>25.2</td>
</tr>
<tr>
<td>LiveJournal</td>
<td>0.330</td>
<td>119,000</td>
<td>17.8</td>
</tr>
<tr>
<td>Orkut</td>
<td>0.171</td>
<td>7,240</td>
<td>5.27</td>
</tr>
<tr>
<td>YouTube</td>
<td>0.136</td>
<td>36,900</td>
<td>69.4</td>
</tr>
</tbody>
</table>

The clustering coefficient is significantly higher compared to random networks.

Users with few friends are more tightly clustered.

Download the WDC Hyperlink Graph

- [http://webdatacommons.org/hyperlinkgraph/](http://webdatacommons.org/hyperlinkgraph/)

- 4 aggregation levels:

<table>
<thead>
<tr>
<th>Graph</th>
<th>#Nodes</th>
<th>#Arcs</th>
<th>Size (zipped)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Page graph</td>
<td>3.56 billion</td>
<td>128.73 billion</td>
<td>376 GB</td>
</tr>
<tr>
<td>Subdomain graph</td>
<td>101 million</td>
<td>2,043 million</td>
<td>10 GB</td>
</tr>
<tr>
<td>1st level subdomain graph</td>
<td>95 million</td>
<td>1,937 million</td>
<td>9.5 GB</td>
</tr>
<tr>
<td>PLD graph</td>
<td>43 million</td>
<td>623 million</td>
<td>3.1 GB</td>
</tr>
</tbody>
</table>