Network Analysis

Maneesh Agrawala

CS 448B: Visualization
Fall 2017

Announcements
Final project

Design new visualization method (e.g. software)
- Pose problem, implement creative solution
- Design studies/evaluations less common but also possible (talk to us)

Deliverables
- Implementation of solution
- 6-8 page paper in format of conference paper submission
- Project progress presentations

Schedule
- Project proposal: Mon 11/6
- Project progress presentation: 11/13 and 11/15 in class (3-4 min)
- Final poster presentation: 12/6 Location: Lathrop 282
- Final paper: 12/10 11:59pm

Grading
- Groups of up to 3 people, graded individually
- Clearly report responsibilities of each member

Network Analysis

*Slides adapted from E. Adar’s / L. Adamic’s Network Theory and Applications course slides.*
Diseases

Transportation

http://diseasome.eu/

http://www.lx97.com/maps/
Lombardi, M. "George W. Bush, Harken Energy and Jackson Stephens, ca 1979–90"

Actors and movies (bipartite)
Characterizing networks

What does it look like?
Topics

Network Analysis
• Centrality / centralization
• Community structure
• Pattern identification
• Models

Tools for Network EDA
Centrality

How far apart are things?
Distance: shortest paths

Shortest path (geodesic path)
- The shortest sequence of links connecting two nodes
- Not always unique
- A and C are connected by 2 shortest paths
  - A → E → B → C
  - A → E → D → C

Distance: shortest paths

Shortest path from 2 to 3: 1

Diameter?
Distance: shortest paths

Shortest path from 2 to 3?

Most important node?
Centrality

Degree centrality (undirected)

\[ C_D = d(n_i) = A_{i+} = \sum_j A_{ij} \]
When is degree not sufficient?

Does not capture

Ability to broker between groups

Likelihood that information originating anywhere in the network reaches you
Betweenness

Assuming nodes communicate using the most direct (shortest) route, how many pairs of nodes have to pass information through target node?

Betweenness - examples

non-normalized:
Betweenness: definition

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

\( g_{jk} \) = the number of geodesics connecting \( j/k \)
\( g_{jk}(i) \) = the number that node \( i \) is on.

Normalization:

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

number of pairs of vertices excluding the vertex itself

When are \( C_d, C_b \) not sufficient?

Do not capture

Likelihood that information originating anywhere in the network reaches you
Closeness: definition

Being close to the center of the graph

Closeness Centrality:

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

Normalized Closeness Centrality

\[ C_c'(i) = \frac{(C_c(i))}{(N - 1)} = \frac{N - 1}{\sum_{j=1, j \neq i}^{N} d(i, j)} \]

Examples - closeness
Centrality in directed networks

Prestige ~ indegree centrality
Betweenness ~ consider directed shortest paths
Closeness ~ consider nodes from which target node can be reached
Influence range ~ nodes reachable from target node

Straight-forward modifications to equations for non-directed graphs

Characterizing nodes

<table>
<thead>
<tr>
<th>High Degree</th>
<th>Low Degree</th>
<th>Low Closeness</th>
<th>Low Betweenness</th>
</tr>
</thead>
<tbody>
<tr>
<td>Node's connections are redundant - communication bypasses him/her</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>High Closeness</th>
<th>Node links to a small number of important/active other nodes.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Many paths likely to be in network; node is near many people, but so are many others</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>High Betweenness</th>
<th>Node's few ties are crucial for network flow</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rare. Node monopolizes the ties from a small number of people to many others.</td>
<td></td>
</tr>
</tbody>
</table>

Low Degree: Node embedded in cluster that is far from the rest of the network.
Centralization – how equal

Variation in the centrality scores among the nodes

Freeman’s general formula for centralization:

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

Examples

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

\[ C_D = \frac{(5 - 5) + (5 - 1) \times 5}{(6 - 1)(6 - 2)} = 1 \]
Examples

Financial networks
Community Structure

How dense is it?

density = \( e / e_{\text{max}} \)

Max. possible edges:
- Directed: \( e_{\text{max}} = n^*(n-1) \)
- Undirected: \( e_{\text{max}} = n^*(n-1)/2 \)
Is everything connected?

Connected Components - Directed

Strongly connected components
- Each node in component can be reached from every other node in component by following directed links
  - B C D E
  - A
  - G H
  - F

Weakly connected components
- Each node can be reached from every other node by following links in either direction
  - A B C D E
  - G H F
Community finding (clustering)

Hierarchical clustering

Process:
- Calculate affinity weights $W$ for all pairs of vertices
- Start: $N$ disconnected vertices
- Adding edges (one by one) between pairs of clusters in order of decreasing weight (use closest distance to compare clusters)
- Result: nested components
Hierarchical clustering (path counts)

- Compute $C_b$ of all edges
- Remove edge $i$ where $C_b(i) = \max(C_b)$
- Recalculate betweenness

Betweenness clustering

Girvan and Newman 2002 iterative algorithm:
- Compute $C_b$ of all edges
- Remove edge $i$ where $C_b(i) = \max(C_b)$
- Recalculate betweenness
**Clustering coefficient**

Local clustering coefficient:

\[ C_i = \frac{\text{number of closed triplets centered on } i}{\text{number of connected triplets centered on } i} \]

Global clustering coefficient:

\[ C_G = \frac{3^* \text{ number of closed triplets}}{\text{number of connected triplets}} \]

\[ C_G = 3*1/5 = 0.6 \]

\[ C_i = 1/3 = 0.33 \]

**Pattern finding - motifs**

Define / search for a particular structure, e.g. complete triads

\[ \text{Define / search for a particular structure, e.g. complete triads} \]

\[ X \quad Y \quad Z \]

\[ a \quad b \quad c \quad d \]
Motifs can overlap in the network

http://mavisto.ipk-gatersleben.de/frequency_concepts.html

motif matches

http://mavisto.ipk-gatersleben.de/frequency_concepts.html

4 node subgraphs
Simulating network models

Small world network

Milgram (1967)
- Mean path length in US social networks
- ~ 6 hops separate any two people
Small world networks

Watts and Strogatz 1998

- a few random links in an otherwise structured graph make the network a small world

**regular lattice:** my friend’s friend is always my friend

**small world:** mostly structured with a few random connections

**random graph:** all connections random

---

**Defining small world phenomenon**

**Pattern:**

- high clustering
- low mean shortest path

**Examples**

- neural network of C. elegans,
- semantic networks of languages,
- actor collaboration graph
- food webs

\[
\ln(l_{network}) \approx \ln(N)
\]
Power law networks

Many real world networks contain hubs: highly connected nodes
Usually the distribution of edges is extremely skewed

![Diagram showing the distribution of edges and nodes]

Summary

Structural analysis
- Centrality
- Community structure
- Pattern finding

→ Widely applicable across domains