Metric Based Comparison of Networks
Graph and Node Level Metrics
Key entity, Group, Twitter reports

Sumeet Kumar
sumeetku@cmu.edu

Terminology

- an edge or tie is defined as any entry where $r_{ij} \neq 0$
  - for geodesic measures an edge weight is assumed to be a distance between nodes $i$ and $j$. If your edges are based on similarity (i.e. number of phone calls between nodes $i$ and $j$) inverted edge weights should be used.
- diagonal: $\forall r_{ij}$ where $i = j$ in bold above
- in most networks the diagonal is assumed to be zeros
- a network is said to be symmetric if $r_{ij} = r_{ji} \forall i, j$
  - symmetric networks are referred to as undirected networks
  - directed networks are not symmetric
Conventions

- Most measures assume the diagonal is 0
- If multiple types of ties, they are typically binarized, summed, and re-binarized
  - Binarization typically uses a > mean or in top third rule
- Many measures assume data is symmetric
  - Let R’ be the new matrix and R the original
  - Strong agreement (local aggregation)
    - R’ij = R’ji = 1 if both Rij & Rji = 1
  - Weak agreement
    - R’ij = R’ji = 1 if either Rij or Rji = 1
- Only square matrices are analyzed
  - Common: Drop columns for nodes for which there is missing information
  - Alternative: Include columns but fill with 0’s
  - Alternative: Assume missing data is random and fill it to reflect the average

Network Level Metrics Commonly Used

Metric
- Size
- Link count
- Density
- Isolate count
- Component count
- Reciprocity
- Characteristic path length
- Clustering coefficient
**Size**

- Number of nodes (people) in the network
- Matters because as size increases
  - Density decreases
  - Clustering increases
- Reflects network boundary
- Should always be included as a covariate

**Density**

- Number of ties, expressed as percentage of the number of ordered/unordered pairs

\[
R = \begin{bmatrix}
  r_{1,1} & r_{1,2} & \cdots & r_{1,n} \\
  r_{2,1} & r_{2,2} & \cdots & \vdots \\
  \vdots & \vdots & \ddots & \vdots \\
  r_{n,1} & \cdots & \cdots & r_{n,n}
\end{bmatrix}
\]

- \( m \) is the number of ordered or unordered edges
- \( \text{density} = \frac{m}{n(n-1)} \) for directed networks
- \( \text{density} = \frac{2m}{n(n-1)} \) for undirected networks

High Density (39%)
Avg. Dist. = 1.76

Low Density (25%)
Avg. Dist. = 2.27
Reciprocity (Mutuality, Symmetry)

- Mutual ties: A → B then B → A
- Some relations are inherently symmetric or asymmetric
  - Who did you have lunch with?
  - Who did you go to for advice?
- Reciprocity is calculated as the percent of ties that are reciprocated:

\[ R = \frac{(A_{ij} = 1) \& (A_{ji} = 1)}{(A_{ij} = 1) \lor (A_{ji} = 1)} \]

Characteristic Path Length

- Also referred to as average path length
- The average distance from a specific node \(i\) to all other nodes in the network is defined naturally as

\[ \bar{d}(i) = \frac{1}{n-1} \sum_{j=1}^{n} d(i, j) \]

  - Where \(d(i, j)\) is the geodesic distance between nodes \(i\) and \(j\)
- The characteristic path length of the network is defined as the average of these over all nodes in the network, or

\[ \bar{d} = \frac{1}{n} \sum_{i=1}^{n} \bar{d}(i) \]
Clustering Coefficient

- a measure of degree to which nodes in a graph tend to cluster together

- Defined as:

\[ C = \frac{3 \times \text{number of triangles}}{\text{number of connected triples of vertices}} = \frac{\text{number of closed triplets}}{\text{number of connected triples of vertices}} \]

\[ c = 1 \quad c = 1/3 \quad c = 0 \]

Geodesic Distance Matrix

\[
\begin{array}{cccccccc}
\text{a} & \text{b} & \text{c} & \text{d} & \text{e} & \text{f} & \text{g} \\
\hline
\text{a} & 0 & 1 & 2 & 3 & 2 & 3 & 4 \\
\text{b} & 1 & 0 & 1 & 2 & 1 & 2 & 3 \\
\text{c} & 2 & 1 & 0 & 1 & 1 & 2 & 3 \\
\text{d} & 3 & 2 & 1 & 0 & 2 & 3 & 4 \\
\text{e} & 2 & 1 & 1 & 2 & 0 & 1 & 2 \\
\text{f} & 3 & 2 & 2 & 3 & 1 & 0 & 1 \\
\text{g} & 4 & 3 & 3 & 4 & 2 & 1 & 0 \\
\end{array}
\]

© Steve Borgatti 2004
**Average Distance**

- Average geodesic distance between all pairs of nodes

**Core/Periphery**
c/p fit = 0.97, avg. dist. = 1.9

**Clique structure**
c/p fit = 0.33, avg. dist. = 2.4

---

**Diameter**

- Maximum distance between any pair of nodes

Diameter = 3

Diameter = 4
FLOW: Walks, Trails, Paths

- **Path**: can’t repeat node
  - 1-2-3-4-5-6-7-8
  - Not 7-1-2-3-7-4
- **Trail**: can’t repeat line
  - 1-2-3-1-7-8
  - Not 7-1-2-7-1-4
- **Walk**: unrestricted
  - 1-2-3-1-2-7-1-7-1
- But – different things flow differently through networks
  - Gift process
  - Currency process
  - Transport process
  - Postal process
  - Gossip process
  - E-mail process
  - Infection process
  - Influence process

---

Flow Terminology

- **Path**: the set of nodes and edges needed to get from one node to another
  - Paths for B
    - BA, BD, BDC, BDCE ...
- **Path length**
  - The number of steps/links
  - Popularly knows as degree – 6 degrees of separation
    - BA 1, BDC 2, BDCE 3
- **Cycle**: a path that crosses the same node
  - BDCEB
- **Distance between two nodes** is the length of the shortest path (aka geodesic)
- **Shortest path**
  - The path between two nodes that has the fewest links
Network Measures – Analysis Levels

- Network (complete graph) level
  - E.g., density
  - Is it easier to disrupt a cellular or hierarchical structure?
  - Use: Characterizing topology, comparing groups, high level change

- Dyadic level
  - E.g., frequency
  - Is there a pattern that money launderers follow?
  - Use: Locating trails

- Node level
  - E.g., centralities
  - Who has the power?
  - Use: Identifying key actors, events, resources ...

Individual vs. Network Level

- Individual behaviors are not independent of the network within which the behaviors occur.
- Individual network position is not independent of the network structure – being central in a centralized network is different than being central in a decentralized network.
Simple SNA Measures

<table>
<thead>
<tr>
<th>Measure</th>
<th>Definition</th>
<th>Meaning</th>
<th>Usage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Degree Centrality</td>
<td>Node with the most connections</td>
<td>In the know</td>
<td>Identifying sources for intel; Reducing information flow</td>
</tr>
<tr>
<td>Betweenness</td>
<td>Node in the most best paths Needs symmetric data</td>
<td>Connects groups</td>
<td>Typically has political influence, but may be too constrained to act</td>
</tr>
<tr>
<td>Eigenvector centrality</td>
<td>Node most connected to other highly connected nodes</td>
<td>Strong social capital</td>
<td>Identifying those who can mobilize others</td>
</tr>
<tr>
<td>Closeness</td>
<td>Node that is closest to all other nodes</td>
<td>Rapid access to all information</td>
<td>Identifying sources to acquire/transmit information</td>
</tr>
<tr>
<td>Betweenness - Centrality</td>
<td>High in betweenness but not degree centrality</td>
<td>Connects disconnected groups</td>
<td>Go-between; Reduction in activity by disconnecting groups</td>
</tr>
</tbody>
</table>

Key Points

- There is an analogous graph level metric for all node level metrics
- Many metrics have poor scale properties
- Local versus global influence
  - “atrophication of influence”
- Metrics are influenced by size and density
- Metrics may or may not take weighted links into account
Using ORA To Compare Networks

• There are many ways to compare networks. One way is to find changes in network metrics. In this session, We will learn three ways to compare network metrics:
  – Network comparison report
  – Run a report, but select both networks, the original and the updated network.
  – Change in Key Entities report

Using ORA To Compare Networks

• Three simple use cases
  – Delete a key-entity and observe the differences in the 'Network Comparison Report'
  – Delete an important Agent and observe the differences in the Twitter report
  – Delete a few central members of a group and observe the difference in the 'Change in Key Entities' report
Network Comparison Report

- Load ‘Twitter_Politifact_politics_search’ xml file
- Run key entity report to find the key entities
- Create a copy of the original network
- Delete one or two ‘key’ entities from the copied network.
- Run the Network comparison report and select both the networks.

Network Comparison Report

- Delete Nodes
Network Comparison Report

- Run Network comparison report
Network Comparison Report

• Run Network comparison report

NETWORK COMPARISON REPORT

Input date: Twitter_Politifact_politics_search, Twitter_Politifact_politics_copy
Start time: Wed Jan 13 07:14:41 2018

1. Differences
   This is a comparison of the two meta-networks: (A) Twitter_Politifact_politics_copy and (B) Twitter_Politifact_politics_search. The purpose of this report is to show what has changed, in terms of nodes and links, when going from A to B. The three cases are summarized in this table:

<table>
<thead>
<tr>
<th>Unchanged/Unmodified</th>
<th>A - B</th>
<th>B - A</th>
</tr>
</thead>
<tbody>
<tr>
<td>Common</td>
<td></td>
<td></td>
</tr>
<tr>
<td>New</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Symmetric difference

The symmetric difference of two meta-network A vs meta-network B is the number of nodes and links in A that are not in B. In set theoretic terms, this is the set difference (A - B). The percentages are computed as 100 * (A - B)/A, which reports the percentage of nodes in the different network that are in the first network.

<table>
<thead>
<tr>
<th>Network A</th>
<th>Network B</th>
<th>Nodes</th>
<th>Percent</th>
<th>Links</th>
<th>Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Twitter_Politifact_politics_search</td>
<td>Twitter_Politifact_politics_search_copy</td>
<td>0</td>
<td>0%</td>
<td>0</td>
<td>0%</td>
</tr>
<tr>
<td>Twitter_Politifact_politics_search</td>
<td>Twitter_Politifact_politics_search</td>
<td>2</td>
<td>0.03%</td>
<td>6335</td>
<td>10.79%</td>
</tr>
</tbody>
</table>

2. Nodeset size summary

This section shows the nodesets sizes of the meta-networks.

<table>
<thead>
<tr>
<th>Nodeset Name</th>
<th>Twitter_Politifact_politics_search</th>
<th>Twitter_Politifact_politics_search_copy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall meta-network</td>
<td>20045</td>
<td>20047</td>
</tr>
<tr>
<td>Agent</td>
<td>9113</td>
<td>9113</td>
</tr>
<tr>
<td>Location</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Tweet</td>
<td>10624</td>
<td>10624</td>
</tr>
</tbody>
</table>
Network Comparison Report

- Run Network comparison report

**Network Density Summary**

This section shows the network densities of the meta-networks.

<table>
<thead>
<tr>
<th>Network Name</th>
<th>Twitter_Politifact_politeness</th>
<th>Twitter_Politifact_politeness_copy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall meta-network</td>
<td>20045</td>
<td>20045</td>
</tr>
<tr>
<td>Agent x Agent - Mentioned-By</td>
<td>9113</td>
<td>9113</td>
</tr>
<tr>
<td>Agent x Agent - Mentioned-By</td>
<td>9113</td>
<td>9113</td>
</tr>
<tr>
<td>Agent x Agent - Reciprocated</td>
<td>9113</td>
<td>9113</td>
</tr>
<tr>
<td>Agent x Agent - Reciprocated</td>
<td>9113</td>
<td>9113</td>
</tr>
<tr>
<td>Agent x Hashing</td>
<td>9113</td>
<td>9113</td>
</tr>
<tr>
<td>Agent x Tweet - Senator</td>
<td>9113</td>
<td>9113</td>
</tr>
<tr>
<td>Hashing x Hashing - Co-Discusses</td>
<td>302</td>
<td>307</td>
</tr>
<tr>
<td>Tweet x Agent - Mentions</td>
<td>10624</td>
<td>10624</td>
</tr>
<tr>
<td>Tweet x Hashing</td>
<td>10624</td>
<td>10624</td>
</tr>
<tr>
<td>Tweet x Location</td>
<td>10624</td>
<td>10624</td>
</tr>
<tr>
<td>Tweet x Tweet - Retweeted-By</td>
<td>10624</td>
<td>10624</td>
</tr>
</tbody>
</table>

Comparison using Twitter Report

- Create a copy of the original network and delete two key nodes
Comparison using Twitter Report

- Run Twitter report, but select both meta networks

Comparison using Twitter Report

- Run Twitter report, observe the comparison.

Import Data Statistics

<table>
<thead>
<tr>
<th>Networks</th>
<th>Twitter_Politifact_politics_search</th>
<th>Twitter_Politifact_politics_search_copy</th>
</tr>
</thead>
<tbody>
<tr>
<td>First tweet date</td>
<td>19624</td>
<td>19624</td>
</tr>
<tr>
<td>Last tweet date</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Number of tweets</td>
<td>10391</td>
<td>10391</td>
</tr>
<tr>
<td>Number of tweets with a URL</td>
<td>10656</td>
<td>10656</td>
</tr>
<tr>
<td>Number of tweets</td>
<td>7864</td>
<td>7862</td>
</tr>
<tr>
<td>Number of tweets</td>
<td>23</td>
<td>23</td>
</tr>
<tr>
<td>Number of news agency tweets</td>
<td>104915</td>
<td>104915</td>
</tr>
<tr>
<td>Number of distinct hashtags</td>
<td>307</td>
<td>307</td>
</tr>
<tr>
<td>Number of distinct hashtags used more than once</td>
<td>162</td>
<td>162</td>
</tr>
<tr>
<td>Number of distinct words</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Number of distinct words used more than once</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Number of distinct locations</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>
Comparison using Twitter Report

- Run Twitter report, observe the comparison.

Tweet Statistics

This summarizes the characteristics of tweets.

<table>
<thead>
<tr>
<th>Network</th>
<th>Twitter_Politic Politics_search</th>
<th>Twitter_Politic Politics_search_copy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average mentions per tweet</td>
<td>0.645</td>
<td>0.645</td>
</tr>
<tr>
<td>Max mentions per tweet</td>
<td>913</td>
<td>913</td>
</tr>
<tr>
<td>Std dev mentions per tweet</td>
<td>1.130</td>
<td>1.130</td>
</tr>
<tr>
<td>Average distinct mentions per tweet</td>
<td>1.404</td>
<td>1.574</td>
</tr>
<tr>
<td>Max distinct mentions per tweet</td>
<td>12</td>
<td>12</td>
</tr>
<tr>
<td>Std dev distinct mentions per tweet</td>
<td>1.545</td>
<td>1.573</td>
</tr>
<tr>
<td>Average distinct hashtags per tweet</td>
<td>0.106</td>
<td>0.106</td>
</tr>
<tr>
<td>Max distinct hashtags per tweet</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Std dev distinct hashtags per tweet</td>
<td>0.049</td>
<td>0.049</td>
</tr>
<tr>
<td>Average distinct words per tweet</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>Max distinct words per tweet</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>Std dev distinct words per tweet</td>
<td>N/A</td>
<td>N/A</td>
</tr>
</tbody>
</table>

Recurring Top Ranked Agent - Twitter_Politic Politics_search

This chart shows the frequency of agents in the top-ranked mentions for each network. The values show the percentage of mentions for each agent. The bar chart displays the top three agents in the top network. If there are multiple network, then a separate bar chart is created for each network.

Recurring Top Ranked Agent - Twitter_Politic Politics_search_copy

This chart shows the frequency of agents in the top-ranked mentions for each network. The values show the percentage of mentions for each agent. The bar chart displays the top three agents in the top network. If there are multiple network, then a separate bar chart is created for each network.
Again, we will use the same two networks, but select ‘Change in Key Entities’ report.
‘Change in Key Entity’ Report

- Observe the generated report

**Potentially Influential (betweenness centrality)**

The Betweenness Centrality of node $v$ in a network is defined as: $\gamma(v) = \sum_{s \neq t \neq v} \frac{\sigma_{st}(v)}{\sigma_{st}}$, where $\sigma_{st}$ is the total number of shortest paths from node $s$ to node $t$ and $\sigma_{st}(v)$ is the number of those paths that pass through $v$. The node is ranked by weight the more smaller the value is. Betweenness Centrality is a measure of the influence of an agent on others. The influence is measured by the number of shortest paths between other agents that are influenced by agent's actions.

**Number of Cliques (clique count)**

The number of distinct cliques to which each node belongs. A clique is defined as a group of items or more nodes that are all connected together and that cannot be made larger by adding another node.

**Input network:** Agents x Agents - Betweened by
Comparison using Group Report

- This is done the same way as we did the Comparison Using Twitter report.

Thank You!!
Density & Size are Negatively Correlated

- In STEP study we have data from 24 coalitions at baseline
- We correlated size and density and discovered a negative association as predicted:
- $R = -0.69$
Critical Personnel

- Individual whose absence will dramatically alter performance of organization; Individual who plays a special role
  - Only person who can do a task
  - Only person with certain organizationally critical knowledge
  - Person who keeps others in line, supported, feeling good about the organization
  - Person who is the only access point to certain organizationally critical knowledge
  - Only person who knows key people
  - Person who knows almost everything
- Examples
  - Lead scientist
  - Lifetime administrative assistant
  - One person lab/technician
  - Lone visionary

Critical Personnel

- Key players, network elite
- Those with power
- Those who, were they to leave, would reduce the organizations performance, adaptability, competence ...
- Direct identifiers
  - The centralities: e.g., degree
  - The exclusivities: e.g., task
  - The integrators: e.g., simmelian ties
  - The loads: e.g., workload and cognitive demand
- Indirect
  - Those who have access to, can influence, those who are critical
Brokerage

- Cutpoints
- Bridges
- Structural holes
- Embeddedness in triads
- Embeddedness in cliques

Node Level SNA Metrics Common Used: Centralities

- Degree Centrality
  - Node with the most connections
- Betweenness Centrality
  - Node in the most best paths
    - Requires symmetric data
- Eigenvector Centrality
  - Node connected best overall
    - Doesn’t work if there are components
- Closeness Centrality
  - Node that is closest to all other nodes

*Issue*: Measures are highly correlated
### Critical Actors
What Nodes Matter?

<table>
<thead>
<tr>
<th>Degree &amp; Eigenvector</th>
<th>Betweenness</th>
<th>Closeness</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.417 Mohamed Atta</td>
<td>0.334 Nawaf Alhazm</td>
<td>0.571 Mohamed Atta</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>In the know power</th>
<th>Connects groups</th>
<th>Rapid access</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mohamed Atta</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nawaf Alhazm</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

---

### Degree Centrality

- Number of edges incident upon a vertex
  - $d_9 = 6$, while $d_{10} = 1$
- Sum of degrees of all nodes is twice the number of edges in graph
- Average degree = density times (n-1)
- Index of exposure to what is flowing through the network
  - Gossip network: central actor more likely to hear a given bit of gossip
- Interpreted as opportunity to influence & be influenced directly
- Predicts variety of outcomes from virus resistance to power & leadership to job satisfaction to knowledge
Betweenness Centrality

How often a node lies along the shortest path between two other nodes
Computed as:

\[ b_k = \sum_{i,j} \frac{g_{ij}}{g_{kj}} \]

where \( g_{ij} \) is number of geodesic paths from \( i \) to \( j \) and \( g_{kj} \) is number of those paths that pass through \( k \)

Index of potential for gate-keeping, brokering, controlling the flow, and also of liaising otherwise separate parts of the network
Interpreted as indicating power and access to diversity of what flows; potential for synthesizing
Sometimes interpreted as “connecting” groups
Very “expensive” to compute

Adapted from Steve Borgatti 2004

Betweenness Centrality

NODE LEVEL –
Betweenness centrality describes the frequencies of nodes in the shortest paths between two indirectly connected nodes. The Betweenness Centrality of node \( v \) in a network is defined as: across all node pairs that have a shortest path containing \( v \), the percentage that pass through \( v \).

Let \( G=(V, E) \) be the graph. Let \( n = |V| \), and fix an entity \( v \in V \).
For \( (u, w) \in V \times V \), let \( n_c(u, w) \) be the number of geodesics in \( G \) from \( u \) to \( w \).
If \( (u, w) \in E \), then set \( n_c(u, w) = 1 \).

In further defining betweenness centrality,
Let \( S = \{(u, w) \in V \times V \mid d_c(u, w) = d_c(u, v) + d_c(v, w) \} \)
Betweenness Centrality of the entity \( v \) = between/ ((n-1)(n-2)/2)

GRAPH LEVEL –
Let \( G=(V,E) \) represent the square network, and let \( n = |V| \)
let \( d_i \) = Betweenness Centrality of node \( i \)

Let \( \bar{d} = \max(d_i \mid 1 \leq i \leq n) \)

Then, Network Betweenness Centrality = \( (\sum_{1 \leq i \leq n} d_i - \bar{d})/(n - 1) \)
Inverted Betweenness

- Inverted betweenness and regular betweenness both measure the extent to which a node is on the critical path between many pairs of nodes.
- Regular betweenness – weighted data – when bigger is worse
  - The strength of the tie represents the delay or cost when going between nodes, so the higher the value of the edge, the less value the edge has in terms of flow.
  - High betweenness nodes have shortest paths to many parts of the system and without them flow would take longer, and greater volume can be handled.
  - The strength of the tie represents the value or frequency or degree of movement between nodes, so the higher the value of the edge, the more value the edge has in terms of flow.
  - High betweenness nodes are the most expensive, least frequented.
- Inverted betweenness - weighted data – when bigger is better
  - If the strength of tie is indicative of higher volume then it should contribute more to higher betweenness values.
  - The strength of the tie represents the delay or cost when going between nodes, so the higher the value of the edge, the less value the edge has in terms of flow and then invert this ...
  - High betweenness nodes are those with highest costs and could be removed with little impact.
  - The strength of the tie represents the value or frequency or degree of movement between nodes, so the higher the value of the edge, the more value the edge has in terms of flow and then invert this ...
  - High betweenness nodes are the most frequented.

Closeness Centrality

- Measured as:
  - Sum of distances to all other nodes
  - Computed as marginals of symmetric geodesic distance matrix.
- Closeness is an inverse measure of centrality.
- Index of expected time until arrival for given node of whatever is flowing through the network
  - Gossip network: central player hears things first.
Eigenvector Centrality

- Node has high score if connected to many nodes that are themselves well connected
- Computed as:

\[ \lambda v = Av \]

where \( A \) is adjacency network and \( V \) is eigenvector centrality. \( V \) is the principal eigenvector of \( A \)

- Indicator of popularity, “in the know”
- Like degree, is an index of exposure, risk
- Tends to identify centers of large cliques
- Often identified as leader of self-contained group
- Leader of Leaders

Moving Beyond Single Measures

*Issue: Centrality Measures are highly correlated*
Brokerage

Brokerage Roles

- Coordinator
- Representative
- Gatekeeper
- Consultant

Connectivity

- Line connectivity $\lambda(s,t)$ is the minimum number of lines that must be removed to disconnect $s$ from $t$
- Node connectivity $\kappa(s,t)$ is the minimum number of nodes that must be removed to disconnect $s$ from $t$
Cutpoints

- Nodes which, if deleted, would disconnect net

Bridge

- A tie that, if removed, would disconnect net
Structural Holes

Local Betweenness

Few structural holes

Many structural hole

Measured by:
- Burt's effective size
- Burt's constraint
- Everett & Borgatti's ego betweenness - This last is recommended

Structural Holes

Robert took over James' job. Entrepreneurial Robert expanded the social capital of the job by reallocating network time and energy to more diverse contacts.

It is the weak connections (structural holes) between Robert's contacts that provide his expanded social capital. Robert is more positioned at the crossroads of communication between social clusters within his firm and its market, and so is better positioned to craft projects and policy that add value across clusters.

Research shows that people like Robert, better positioned for entrepreneurial opportunity, are the key to integrating across functions and across the people of increasingly diverse backgrounds in today's flatter organizations. In research comparisons between managers like James and Robert, it is the people like Robert who get promoted faster, earn higher compensation, receive better performance evaluations, and perform more successfully on teams.

Slide from Ron Burt
Overall Tweet Network
Note there are a few sources that are picked up

Most Retweeted Actors
Sphere of Influence

- All ties to/from ego and the connections among the associated nodes in any meta-network
- Extension of ego net idea to meta-network
- Level = the selection path length allowed between an ego and alters in defining the size of the sphere
  - E.g., level 1 – all nodes one away from ego and the connections among them - typical ego net
  - E.g., level 2 – all nodes two away from ego and the connections among them

From Ego Networks to Spheres of Influence

- Who are they connected to
- What groups are they in
- What do they know
- What resources do they control
- What activities are they involved in
Hashtag network
Note – no direct linkage between Arabic topic-group and English
**Most Critical Hashtags**

These are ones that co-occur with other tags the most.

![Graph showing co-occurrence of hashtags]

**Multi-country Comparison**

Note Egyptian attack is against a background of violent events. Libya and Yemen are anomalous.

![Graph showing tweeting rates per hour across countries]
Characteristics of networks relation to Measures

<table>
<thead>
<tr>
<th>Measure</th>
<th>Allow/Ignore Self-Loops</th>
<th>Symmetric/Asymmetric</th>
<th>Binary/Weighted</th>
<th>Connected/Disconnected</th>
</tr>
</thead>
<tbody>
<tr>
<td>Degree</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>no</td>
</tr>
<tr>
<td>Betweenness</td>
<td>no</td>
<td>yes</td>
<td>yes</td>
<td>no</td>
</tr>
<tr>
<td>Closeness</td>
<td>no</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>Eigenvector</td>
<td>yes</td>
<td>no</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>Clustering Coefficient</td>
<td>yes</td>
<td>yes</td>
<td>no</td>
<td>no</td>
</tr>
</tbody>
</table>

Alternative Measures

- **Page rank**
  - Iteratively Weighted degree centrality
  - Page rank of node i is $PR(p_i)$
  - $D$ is the dampening factor
  - $L(p_j)$ is the outdegree of node j

  $$PR(p_i) = \frac{1 - d}{N} + d \sum_{p_j \in M(p_i)} \frac{PR(p_j)}{L(p_j)}$$

- **Authority Centrality**
  - A node is authority-central to the extent that its in-links are from nodes that have many out-links.
  - The authority of a node are calculated by: $a(i) = \sum_{(j,i) \in E} h(j)$,
    - Where $a(x)$ is the authority of node $x \in \{i, j\}$
Clustering Coefficient
(Carley et al., 2013: 845-847)

NODE LEVEL –
Measures the degree of clustering in a network by averaging the
clustering coefficient of each node, which is defined as the density of the
node’s ego network.

Let $G = (V, E)$ be the graph representation of a square network.
Define for each node $v \in V$ its Clustering Coefficient $CC_v$:
let $G_v = \text{ego network of entity } v$

Then Clustering Coefficient for entity: entity $v = CC_v = \text{density}$
($G_v$)

GRAPH LEVEL -

Then Clustering Coefficient for graph: graph $= \frac{\sum_{v \in V} CC_v}{|V|}$

Bonacich

- Bonacich’s power measure corresponds to the notion that the power
  of a vertex is recursively defined by the sum of the power of its
  alters.
  - The nature of the recursion involved is then controlled by the power exponent:
    positive values imply that vertices become more powerful as their alters become
    more powerful (as occurs in cooperative relations), while negative values imply
    that vertices become more powerful only as their alters become weaker (as
    occurs in competitive or antagonistic relations).
  - The magnitude of the exponent indicates the tendency of the effect to decay
    across long walks; higher magnitudes imply slower decay.
  - One interesting feature of this measure is its relative instability to changes in
    exponent magnitude (particularly in the negative case). If your theory motivates
    use of this measure, you should be very careful to choose a decay parameter on
    a non-ad hoc basis.

  American Journal of Sociology, 92, 1170-1182.
Bonacich's power centrality measure

\[ C_{BP}(\alpha, \beta) = \alpha (I - A)^{-1} A 1, \]

- \( \beta \) is an attenuation parameter (set here by exponent)
- \( A \) is the graph adjacency matrix.
- \( \alpha \) acts as a scaling parameter, and is set here (following Bonacich (1987)) such that the sum of squared scores is equal to the number of vertices.
  - This allows 1 to be used as a reference value for the \textquoteleft middle\textquoteright of the centrality range.
- When \( \beta \to 1/\lambda_{A1} \) (the reciprocal of the largest eigenvalue of \( A \)), this is to within a constant multiple of the familiar eigenvector centrality score; for other values of \( \beta \), the behavior of the measure is quite different.
- \( \beta \) gives positive and negative weight to even and odd walks, respectively, as can be seen from the series expansion \( C_{BP}(\alpha, \beta) = \alpha \sum (\beta^k A^{k+1} 1) \) which converges so long as \( |\beta| < 1/\lambda_{A1} \).
- The magnitude of \( \beta \) controls the influence of distant actors on ego's centrality score, with larger magnitudes indicating slower rates of decay.
- High rates imply a greater sensitivity to edge effects.

Homophily/Diffusion

- Because people associate with others like themselves, their networks are often closed to outside information
- My friends think and do the same things I think and do
- Even it they don't, I project on to them
- Makes diffusion a difficult and slow process
- Yet, once someone I know and am close to adopts something, I am likely to find it relevant for me
- [Strong ties important for adoption, weak ties for information and network level diffusion.]
Explanatory Mechanisms - Flow

- Flows mechanism
  - Interpersonal transmissions, flows, influence
  - Relational / Connectionist
  - Ties as pipes
  - Linkages

- Social Capital
  - Individual level
    - Nan Lin’s social resource theory
      - People who have ties to important, wealthy, knowledgeable people are better off
  - Group Level
    - Putnam’s bowling alone
      - Communities with dense helping & trust ties are better off

- Diffusion
  - Individual level
    - Granovetter’s strength of weak ties theory
      - People who with more weak ties have better chance of hearing novel information (e.g., about jobs)
    - Lave & Wenger’s community of practice theory
      - Developing shared culture through interaction
  - Group Level
    - Granovetter’s community strength argument
      - Communities with lots of weak ties can coordinate more effectively than those with strong ties

Explanatory Mechanisms - Topology

- Topology mechanism
  - Convergent outcomes based on similar social environments
  - Structural
  - Ties as girders
  - Shapes

- Social Capital
  - Individual level
    - Burt’s structural holes argument
      - People whose personal networks have certain topological features are better off
  - Group Level
    - Bavelas’ centralization argument
      - For simple tasks, groups with centralized communication structures perform better

- Diffusion
  - Mimetic isomorphism (Dimaggio & Powell)
  - Similarity of attitude as function of structural equivalence (Erickson)
  - Adoption as a function of structural equivalence (Burt)
  - Adoption as a function of centrality (Coleman)
Network Elite

- Nodes that stand out as high/low on some measure
- Power
  - Bonacich power centrality = out-degree (row) centrality when \( \beta = 0 \)
  - Access to resources, information, people
  - Ability to mobilize others (reach)
  - Ability to control the flow of information
  - Ability to give orders
  - Ability to broker between groups

Identifying Network Elite

- Centrality Approach
  - How much matters
- Brokerage
  - Who you connect matters
- But ...
  - It matters what is flowing through the network
- It matters if network is multi-mode, multi-plex, multi-way
Centralities

- Degree Centrality
  - Node with the most connections

- Betweenness Centrality
  - Node in the most best paths
    - Requires symmetric data

- Eigenvector Centrality
  - Node connected best overall
    - Doesn't work if there are components

- Closeness Centrality
  - Node that is closest to all other nodes

Issue: Measures are highly correlated

Simple SNA Measures

<table>
<thead>
<tr>
<th>Measure</th>
<th>Definition</th>
<th>Meaning</th>
<th>Usage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Degree Centrality</td>
<td>Node with the most connections</td>
<td>In the know</td>
<td>Identifying sources for intel; Reducing information flow</td>
</tr>
<tr>
<td>Betweenness</td>
<td>Node in the most best paths</td>
<td>Connects groups</td>
<td>Typically has political influence, but may be too constrained to act</td>
</tr>
<tr>
<td></td>
<td>Needs symmetric data</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Eigenvector centrality</td>
<td>Node most connected to other highly connected nodes</td>
<td>Strong social capital</td>
<td>Identifying those who can mobilize others</td>
</tr>
<tr>
<td>Closeness</td>
<td>Node that is closest to all other nodes</td>
<td>Rapid access to all information</td>
<td>Identifying sources to acquire/transmit information</td>
</tr>
<tr>
<td>Betweenness - Centrality</td>
<td>High in betweenness but not degree centrality</td>
<td>Connects disconnected groups</td>
<td>Go-between; Reduction in activity by disconnecting groups</td>
</tr>
</tbody>
</table>
Centrality

- This is why central members are so important.
- They are seen as the same yet different than everyone else in the network.
- Little hope to get diffusion going unless central members embrace the idea.
- Yet, most ideas start on the periphery of the network because they are freed from social norms.
- Bridges

Centrality/Popularity

- One diffusion driver is the behavior of opinion leaders.
- OLs both reflect and drive the diffusion process.
- Identify OLs by those who receive many nominations (typically top 10-15%).
- Other measures of centrality—who is located at the center of the network.
Who Is “Key”?

Internet Alliances

Who Is “Key”?

Internet Alliances
Degree Centrality

- Degree – total number of edges/ nodes ego is connected to
  - Commonly thought of as a measure of influence or importance
- In Degree – total number of nodes that send edge to ego (column)
- Out Degree – total number of nodes that receive edge from ego (row)
- Sink – 0 in degree; Source – 0 out degree

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
</tr>
</thead>
<tbody>
<tr>
<td>N In</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>N Out</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Total</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
</tbody>
</table>

Total Degree Centrality

NODE LEVEL –
Total Degree Centrality : The normalized sum of an node's row and column degrees.
- Consider the matrix representation X of a square network with n entities. Each entity is assigned a value based on the sum of its rows and columns.
- The Total Degree Centrality for entity \( i \) = \( \frac{1}{2(n-1)} \sum_{i=1}^{n} \sum_{j=1}^{n} X(i,j) \)

GRAPH LEVEL –
A centralization of a square network based on the Total-Degree Centrality of each node.
- Let \( N \) be a unimodal network with n entities.
- Let \( d_i \) = Total Degree Centrality of entity \( i \)
- Let \( \bar{d} \) = \( \max \{d_i \mid 1 \leq i \leq n\} \)
- Then Total Degree Network Centralization = \( (\sum_{1 \leq i \leq n} \bar{d} - d_i)/(n - 2) \)
Typical Social Network Study

- Question?
- Please list all the Drs. to whom you have referred patients in the past six months.

1. Characteristics
   - One Way
   - One mode
   - One type of link
     - Binary
     - Directed
Steps in a Structural Analysis

- Collect network data.
  - Connections among people, knowledge, resources, events ...
- Enter data into ORA.
- Visualize.
- Generate Report.
- If multiple networks create combined measures.
- If needed look at some measures more indepth.
- Possibly drop isolates and pendants
- Check interpretations.

Egocentric

Please tell me, within the last six months, the first names or initials of up to 5 people you talk to most often about important matters?

<table>
<thead>
<tr>
<th>Respondent</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>Name ________</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>No Yes Name ________</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>No Yes No Yes Name ________</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>No Yes No Yes No Yes Name ________</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>No Yes No Yes No Yes No Yes Name ________</td>
</tr>
</tbody>
</table>
Ego-centric Network Data are Node-based

Network Analysis

- Derives from graph theory
- Set of measures on graphs or networks
- Graph - binary matrix
- Network - weighted matrix
Terminology

- **Node**
  - entity, <person>, dot, point

- **Tie**
  - relation, link, edge, connection, <friendship>
  - vary in strength (weight), direction, type, confidence (another weight)

- **Caveat:** 80-90% of work/measures uses binary data

```
0 1 0 1 0
1 0 1 0 1
1 0 0 0 1
0 0 1 0 1
0 1 0 0 1
```

Key Graph Theoretic Concepts

- Directed —versus— undirected
  - Directed — commands
  - Undirected — works with

- **Strength**
  - Frequency of interaction
  - Distance

- **Adjacency**
  - Equivalent matrix

- **Walk; length**
  - Unrestricted; number of ties

- **Path**
  - Do not repeat a node

- **Trail**
  - Do not repeat a tie

- **Distance**
  - Shortest path (geodesic)

```
0 2 0 5 0
1 0 0 1 0
1 0 0 7 0
0 0 8 0 1
0 2 4 0 0
```
## Terminology

- **Degree** – total number of edges/ nodes ego is connected to
- **In Degree** – total number of nodes that send edge to ego
- **Out Degree** – total number of nodes that receive edge from ego
- **Sink** – 0 in degree; **Source** – 0 out degree

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>2</td>
<td>2</td>
<td>4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>2</td>
<td>4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>2</td>
<td>4</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

## Terminology

- **Mode** – types of nodes (number of entity classes)
  - Traditional SNA uses 1 mode data (sometimes 2)
  - Traditional Link analysis uses multi-mode
- **Multi-plex** – types of links
  - A multi-plex data set has multiple relations among nodes of the same mode/entity class
  - Most SNA data sets are single plex
  - Traditional Link analysis uses multi-plex data
- **“Way”** means dimensions: rows, columns, levels, etc.
  - Most SNA data sets are 2-way (row by column)
  - Most over time data sets are 3-way (1 matrix per time)
- **Meta-Network**
  - A set of networks defined over multiple entity classes, both multi-mode and multi-plex
  - Can be multi-way also
- **E.g., 3-way, 1-mode, single-plex**
  - Perceived social networks (CSS)
    - CSS – cognitive social structure
    - Each person gives their perception of who knows whom
  - Transactive memory of social relations
- **E.g., 3-way, 3-mode, multi-plex**
  - Transactive memory (over actors, knowledge, tasks) for existent and desired relations
**Ego Nets**

- **Ego Network**
  - All ties to/from ego and the connections among the associated nodes in a two-way, single mode, single plex data set
  - *Be careful – some authors only use the ties to/from B*
  - *Nora – finds these limited ego nets*
  - By convention – typically only used for actor or organization entity classes

```
0 1 0 1 0
1 0 0 1 0
1 0 0 0 1
0 0 1 0 1
0 1 1 0 0
```

B's ego Net

---

**Illustration of Control: Cognitive Demand**

- *The cognitive effort the individual has to do on average*
- **How many people do you interact with**
- **How many tasks do you do**
- **How much knowledge do you have**
- **How much knowledge is needed to do the tasks**
- **How many people do you need to interact with to do the tasks**
- **How many other tasks and so people depend on you**
- **How many other tasks and so people do you depend on**
1-mode ego network

Carter Administration meetings

Year 1
Year 4

FLOW: Walks, Trails, Paths

- Path: can’t repeat node
  - 1-2-3-4-5-6-7-8
  - Not 7-1-2-3-7-4
- Trail: can’t repeat line
  - 1-2-3-1-7-8
  - Not 7-1-2-7-1-4
- Walk: unrestricted
  - 1-2-3-1-2-7-1-7-1

- But – different things flow differently through networks
  - Gift process
  - Currency process
  - Transport process
  - Postal process
  - Gossip process
  - E-mail process
  - Infection process
  - Influence process

© Steve Borgatti 2004
Flow Terminology

- Path: the set of nodes and edges needed to get from one node to another
  - Paths for B
  - BA, BD, BDC, BDCE ...
- Path length
  - The number of steps/links
  - Popularly known as degree – 6 degrees of separation
  - BA 1, BDC 2, BDCE 3
- Cycle: a path that crosses the same node
  - BDCEB
- Distance between two nodes is the length of the shortest path (aka geodesic)
- Shortest path
  - The path between two nodes that has the fewest links

Network Measures – Analysis Levels

- Network (complete graph) level
  - E.g., density
  - Is it easier to disrupt a cellular or hierarchical structure?
  - Use: Characterizing topology, comparing groups, high level change

- Dyadic level
  - E.g., frequency
  - Is there a pattern that money launderers follow?
  - Use: Locating trails

- Node level
  - E.g., centralities
  - Who has the power?
  - Use: Identifying key actors, events, resources ...
Individual v. Network Level

- Individual behaviors are not independent of the network within which the behaviors occur.
- Individual network position is not independent of the network structure – being central in a centralized network is different than being central in a decentralized network.