Representing And Measuring Networks

Rogayeh Dastranj Tabrizi
email: rda18@sfu.ca
Office: WMC 3607
Office Hours: Thursdays 12pm-2pm

Department of Economics
Simon Fraser University

Thanks to Matthew Jackson for access to his teaching resources.

10. Februar 2015

Simplifying The Complexity

- Global patterns of networks:
  - Degree distributions
  - Path lengths
- Segregation Patterns: node types and homophily
- Local Patterns
  - Clustering
  - Support
- Positions in networks
  - Neighbourhoods
  - Centrality, influence, ...

Representing Networks

- $N = \{1, 2, ..., n\}$ is the set of nodes, or **vertices, players, agents**.
- Connection between nodes is called links, or **edges, ties**:

1. They may have intensity (**weighted network**):
   - Hoe many hours do people spend together per week?
   - How much of one country's GDP is traded with another?
2. They may just be 0 or 1 (**unweighted network**):
   - Have two researchers written an article together?
   - Are two people "friends" on some social platform?
3. They may be Directed or Undirected:
   - coauthors, friends,..., relatives, spouses, ..., are mutual relationships
   - link from on web page to another, citations, following on social media..., one way

Graphs and Networks

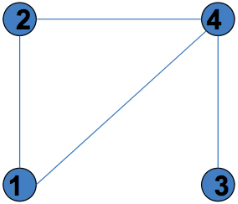
- $N = \{1, 2, ..., n\}$ is the set of nodes,
- $g_{n \times n}$ is a real-valued $n \times n$ matrix, where g_{ij} represent the relationship between i and j.
- In an unweighted network:

$$g_{ij} = \begin{cases} 1 & \text{if } ij \in g \\ 0 & \text{otherwise} \end{cases}$$

- Notation: $ij \in g$ indicates a link between i and j.
- Self-links or loops often do not have any real consequences or meaning. Unless otherwise is indicated, assume $g_{ii} = 0$.

Unweighted Undirected Networks

$$g = \begin{pmatrix} 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 1 & 1 & 1 & 0 \end{pmatrix}$$

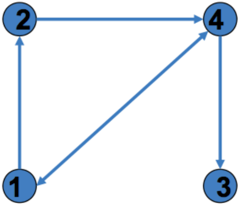


Or list the links:

- $g = \{\{1,2\},\{1,4\},\{2,4\},\{3,4\}\}$
- $g = \{12, 14, 24, 34\}$ or $g = \{21, 41, 42, 43\}$

Unweighted Directed Networks

$$g = \begin{pmatrix} 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 \end{pmatrix}$$

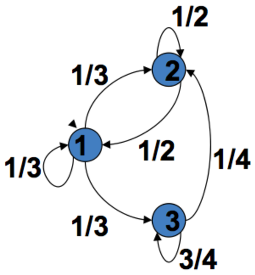


Notice that the order of nodes matter in a directed network

- $g = \{12, 14, 41, 24, 34\}$

Weighted Directed Networks

$$g = \begin{pmatrix} 1/3 & 1/3 & 1/3 \\ 1/2 & 1/2 & 0 \\ 0 & 1/4 & 3/4 \end{pmatrix}$$



Graphs and Networks

- A network is directed if it is possible that $g_{ij} \neq g_{ji}$.
- A network is undirected if $g_{ij} = g_{ji}$ for all nodes i, j.
- $g' \subset g$ indicates that:

$$g' \subset g \quad \Leftrightarrow \quad \{ij : ij \in g'\} \subset \{ij : ij \in g\}$$

- $g + ij$ indicates that a new network that is obtained by adding link ij from network g.
- $g - ij$ indicates that a new network that is obtained by deleting link ij from network g.

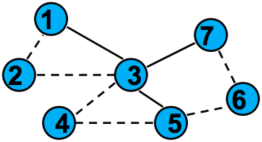
Basic Definitions

- Walk from i_1 to i_K : A sequence of links $\{i_1 i_2, i_2 i_3, ..., i_{K-1} i_K\}$ such that $i_{k-1} i_k \in g$ for all k in this walk.
- Often it is convenient simply to represent a **walk** as the corresponding sequence of nodes $(i_1, i_2, ..., i_K)$ such that $i_{k-1} i_k \in g$ for each k.
- Path from i_1 to i_K: is a **walk** where all the nodes are distinct.

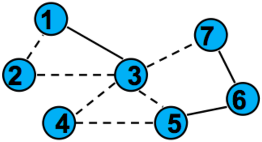
In other words, a **walk** may come back to a given node more than once, whereas a path is a walk that never hits the same node twice.

- A cycle is a **walk** where $i_i = i_k$.
- Geodesic: a **shortest path** between two nodes.

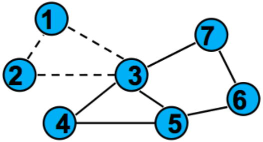
Walks, Paths, Cycles



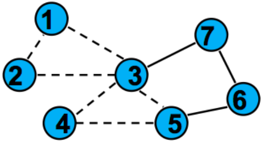
Path (and a walk) from 1 to 7:
1, 2, 3, 4, 5, 6, 7



Walk from 1 to 7 that is not a path:
1, 2, 3, 4, 5, 3, 7



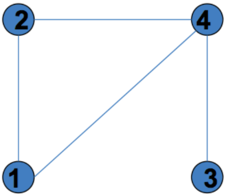
Simple Cycle (and a walk) from 1 to 1:
1, 2, 3, 1



Cycle (and a walk) from 1 to 1:
1, 2, 3, 4, 5, 3, 1

Counting Walks

$$g = \begin{pmatrix} 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 1 & 1 & 1 & 0 \end{pmatrix}$$

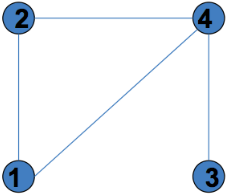


$$g^2 = \begin{pmatrix} 2 & 1 & 1 & 1 \\ 1 & 2 & 1 & 1 \\ 1 & 1 & 1 & 0 \\ 1 & 1 & 0 & 3 \end{pmatrix}$$

number of walks of length 2 from i to j

Counting Walks

$$g = \begin{pmatrix} 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 1 & 1 & 1 & 0 \end{pmatrix}$$



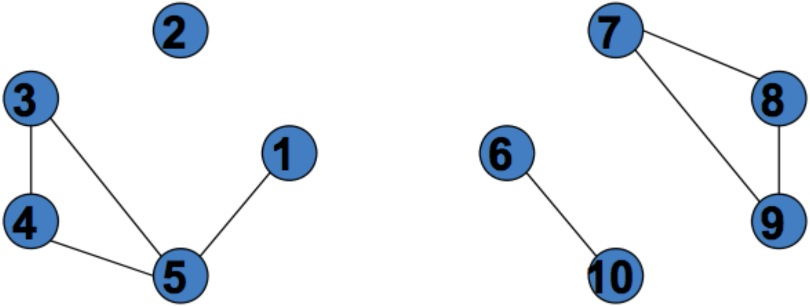
$$g^3 = \begin{pmatrix} 2 & 3 & 1 & 4 \\ 3 & 2 & 1 & 4 \\ 1 & 1 & 0 & 3 \\ 4 & 4 & 3 & 2 \end{pmatrix}$$

number of walks of length 3 from i to j

Components

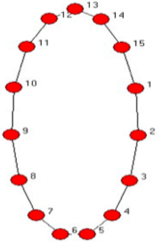
- In many applications it is important to track which nodes can reach which other nodes through a path.
  - Contagion, Learning, Diffusion of various behaviours, etc.
- Network (N, g) is Connected if every two nodes are connected by some path.
- Component: Maximal connected subgraph, i.e. (N', g') is a component of (N, g) such that:
  - $N' \neq \emptyset$, $N' \subset N$, $g' \subset g$,
  - (N', g') is connected,
  - if $i \in N'$, and $ij \in g$, then $j \in N'$ and $ij \in g'$.
- A link ij is a bridge in a network g if $g - ij$ has more components than g.

Components

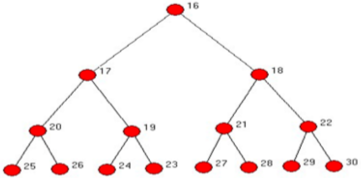


Diameter

- Diameter: largest geodesic in the network. If the network is unconnected, then the largest geodesic of the largest component.
- Another measure is average path length, which is less prone to outliers.



Diameter is either $n/2$ or $(n-1)/2$.



Diameter is on the order of $2 \log 2(n+1)$.

Trees and Stars

There are a few particular network structures that are commonly referred to:

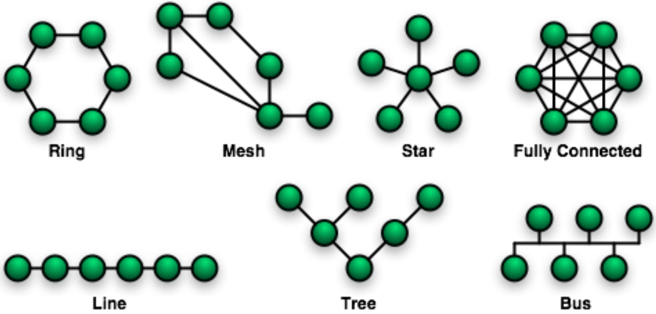
- Tree: A connected network that has **no cycles**.
- Forest: A network such that each **component** is a tree. Any network that has no cycle is a forest.
- Star: There exist a node such that every link in the network involves i. i is the center of the star.

Facts about Trees:

- A connected network is a tree if and only if there has $n-1$ links.
- In a tree, there is unique path between any two links.

Circles and Complete Networks

- Complete Network is one in which all possible links are present, so that $g_{ij} = 1$ for all $i \neq j$.
- Circle is a network that has a single cycle and is such that each node has exactly two neighbours.



Neighbourhood

- Neighbourhood of node i is a set of nodes that i is linked to.

$$N_i(g) = \{j : g_{ij} = 1\}$$

- Given some nodes S, the neighbourhood of S is the union of the neighbourhoods of its members:

$$N_S(g) = \bigcup_{i \in S} N_i(g) = \{j : \exists i \in S, g_{ij} = 1\}$$

- k-neighbourhood of i: All nodes that can be reached from i by walks of length no more than k:

$$N_i^k(g) = N_i(g) \cup \left(\bigcup_{j \in N_i(g)} N_j^{k-1}(g) \right)$$

Degree and Network Density

- Degree: of a node is the number of links that involves that node:

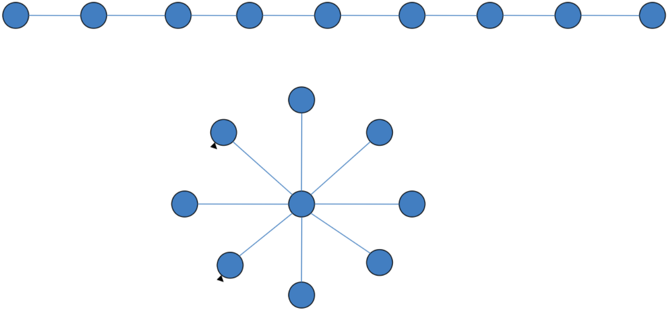
$$d_j(g) = \#N_i(g) = \#\{j : g_{ij} = 1\}$$

- In terms of **directed networks**:
  - In-degree: $d_j(g) = \#\{j : g_{ji} = 1\}$
  - Out-degree: $d_j(g) = \#\{j : g_{ij} = 1\}$

- Network Density: keeps track of the relative fraction of links present in the network and is equal to **average degree divided by** $n - 1$.

Degree Distribution

- How is degree distributed in the network?
- Average degree only tells us part of the story:



Degree Distribution

- Degree Distribution of a network is description of the relative frequency of nodes that have different degrees.
- $P(d)$ is the fraction of nodes that have degree d under distribution p.
  - A **regular network** is one in which all nodes have the same degree.
  - A network is a **regular of degree** k if $P(k) = 1$ and $P(d) = 0 \; \forall d \neq k$.
- Another example is the scale-free or **power** degree distribution:

$$P(d) = cd^{-\gamma}$$

- The relative probabilities of degrees of a fixed relative ratio are the same independent of the scale of those degrees.

$$\frac{P(2)}{P(1)} = \frac{P(20)}{p(10)}$$

Overall Clustering

- What fraction of my friends are friends with each other?
- Overall Clustering:

$$
\begin{aligned}
Cl(g) &= \frac{\sum_i \#\{jk \in g | k \neq j, j \in N_i(g), k \in N_i(g)\}}{\sum_i \#\{jk | k \neq j, j \in N_i(g), k \in N_i(g)\}} \\
&= \frac{\sum_i g_{ij} g_{ik} g_{jk}}{\sum_i g_{ij} g_{ik}}
\end{aligned}
$$

- This is in fact the fraction of fully connected triples out of the potential triples in which at least two links are present.

Individual Clustering

- This measure is computed on node-by-node basis and then averaged across nodes. Individual Clustering:

$$Cl_i(g) = \frac{\#\{jk \in g | k \neq j, j \in N_i(g), k \in N_i(g)\}}{\#\{jk | k \neq j, j \in N_i(g), k \in N_i(g)\}}$$
$$= \frac{\sum_{j,k} g_{ij} g_{ik} g_{jk}}{\sum_{j,k} g_{ij} g_{ik}}$$

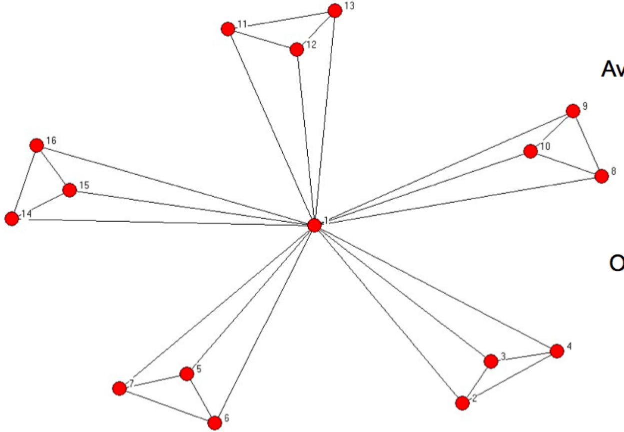
- This looks at all the pairs of nodes that are linked to i and then asks how many of them are connected to each other.
- Average Clustering:

$$Cl^{Avg}(g) = \sum_i Cl_i(g)/n$$

Difference In Clustering

- Under **average clustering**, one computes clustering for each node and then averages over all nodes.
- Whereas with **overall clustering**, the average is taken over all triplets.
- Average clustering gives more weight to low-degree nodes than does the clustering coefficient method.
- The average clustering for the Florentine marriage network is 3/20, where as the overall clustering coefficient is 9/47.

Difference In Clustering



Average tends to 1

Overall tends to 0

Position In The Network

How to describe individual characteristics?

- Degree
- Clustering
- Distance to other nodes
- Centrality and Influence

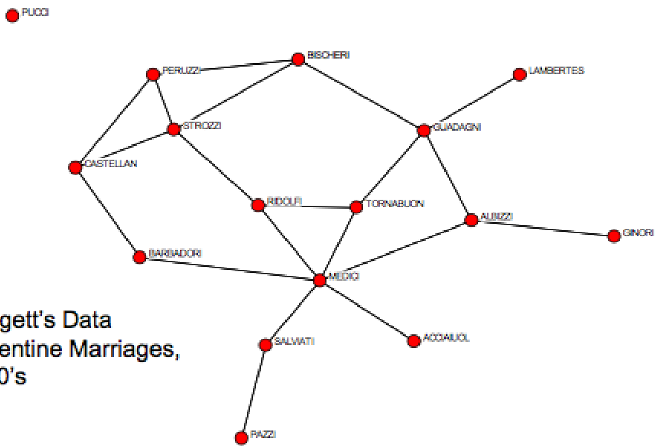
Centrality Meausres

- Degree: Measure of connectedness.
- Closeness, Decay: Ease of reaching other nodes.
- Betweenness: Importance as an intermediary, connector.
- Influence, Prestige, Eigenvectors: "not what you know, but who you know!"

How "connected" is a node?

- Degree Centrality captures connectedness.
- Normalize by $n-1$ Which the most possible number of connections in a network of size n.

Degree Centrality - Examples

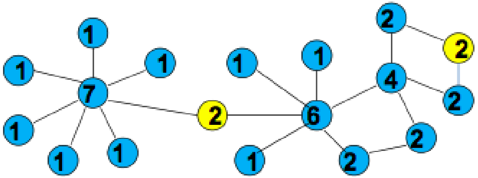


Medici = 6
Strozzi = 4
Guadagni = 4

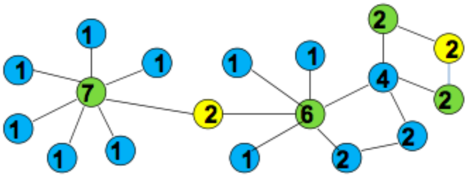
Padgett's Data
Florentine Marriages,
1430's

Degree Centrality - Examples

- Failure of degree centrality to capture reach of a node:



- More reach if connected to a 6 and 7 than a 2 and 2?



Eigenvalue Centrality

- Eigenvalue Centrality is proportional to the **sum of neighbours' centralities**:

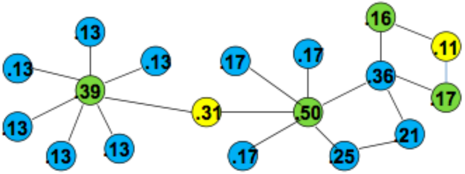
$$C_i \propto \sum_{j \;\; in N_i(g)} C_j$$

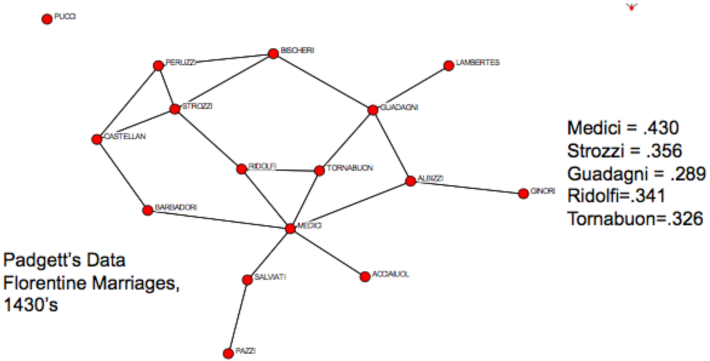
- More connections matter, but also accounts for how central they are!

$$\alpha C_i = \sum_{j \;\; in N_i(g)} C_j = \sum_j g_{ij} C_j$$
$$\Rightarrow \qquad \alpha \mathbf{C} = \mathbf{gC}$$

- **Google page rank**: score of a page is proportional to the sum of the scores of pages linked to it.

Eigenvalue Centrality - Example





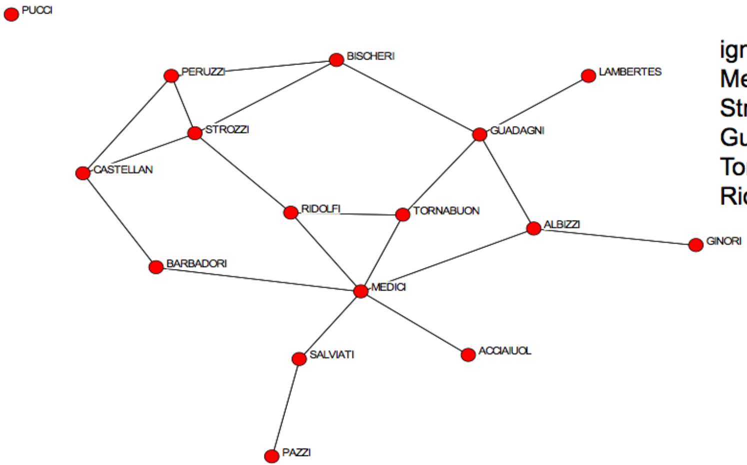
Closeness Centrality

- How close a given node is to any other node?
- Closeness Centrality: The inverse of the average distance between i and any other node j:

$$\frac{n-1}{\sum_{i \neq j} l(i,j)}$$

- $l(i,j)$ is the number of links in the shortest path between i and j.
- Scales directly with distance: twice as far is half as central.

Closeness Centrality



ignoring Pucci:
Medici 14/25
Strozzi 14/32
Guadagni 14/26
Tornabuon 14/29
Ridolfi 14/28

Decay Centrality

- A richer way of measuring centrality based on closeness is considering a decay parameter $0 < \delta < 1$, and consider the proximity between a given node and every other node wighted by this parameter.
- Decay Centrality:

$$C_i^d(g) = \sum_{i \neq j} \delta^{l(i,j)}$$

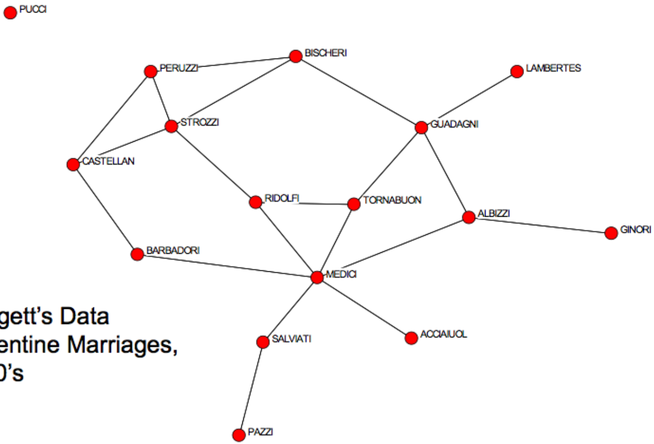
- When $\delta \to 1$, then $C_i^d(g)$ measures **component size**.
- When $\delta \to 0$, then the decay centrality gives infinitely more weights to the closers nodes than farther ones, i.e. becomes proportional to **Degree Centrality**.
- For intermediate values of δ, closer nodes have higher weights than less close nodes.

Betweenness Centrality

- Betweenness Centrality gives a measure of how well situated a node is in terms of the paths that it lies on.
- Let $P(i, j)$ be the number of geodesics between i and j,
- Let $P_k(i, j)$ be the number of geodesics between i and j that goes through k,
- Estimate how important k is in terms of connecting i and j, by looking at the ratio:

$$C_i^B(g) = \sum_{i \neq j, k \notin \{i,j\}} \frac{P_k(ij)/P(ij)}{(n-1)(n-2)/2}$$

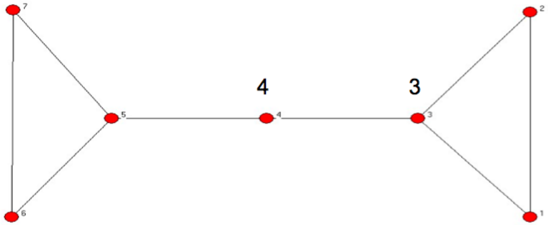
Betweenness Centrality



Padgett's Data
Florentine Marriages,
1430's

Medici = .522
Strozzi = .103
Guadagni = .255

Centrality Measures - Example



	Node 1	Node 3	Node 4
Degree	.33	.50	.33
Closeness	.40	.55	.60
N. Decay $\delta = .5$	.50	.67	.67
N. Decay $\delta = .75$	.69	.82	.84
N. Decay $\delta = .25$	.39	.56	.50
Betweenness	.00	.53	.60
Eigenvector	.47	.63	.54

Centrality Measures

- Degree Centrality: Measures connectedness.
- Closeness and Decay centrality: Measures ease of reaching others.
- Betweenness Centrality:Measures importance as an intermediary and a connector.
- Influence, Prestige, Eigenvectors Centrality: "not what you know, but who you know...".