6.207/14.15: Networks
Lectures 2 & 3: Graphs, Measures and Metrics

Outline

- Network representation
  - Graphs. Definitions. Notations.

- Graph properties, measurements and metrics
  - Diameter, Average Path Length, Degree Distributions
  - Clustering Coefficient, Centrality

Reading:

- Newman, Chapter 6 (skip Sections 6.8 and 6.12-6.14).
- Newman, Chapter 7, Sections 7.1, 7.6, and 7.7.

Network Study

- Historical study of networks:
  - Mathematical graph theory: One of the pillars of discrete mathematics
    - Started with Euler's 1735 solution of the Königsberg bridge problem.
  - Can you cross each bridge exactly once in a walk?

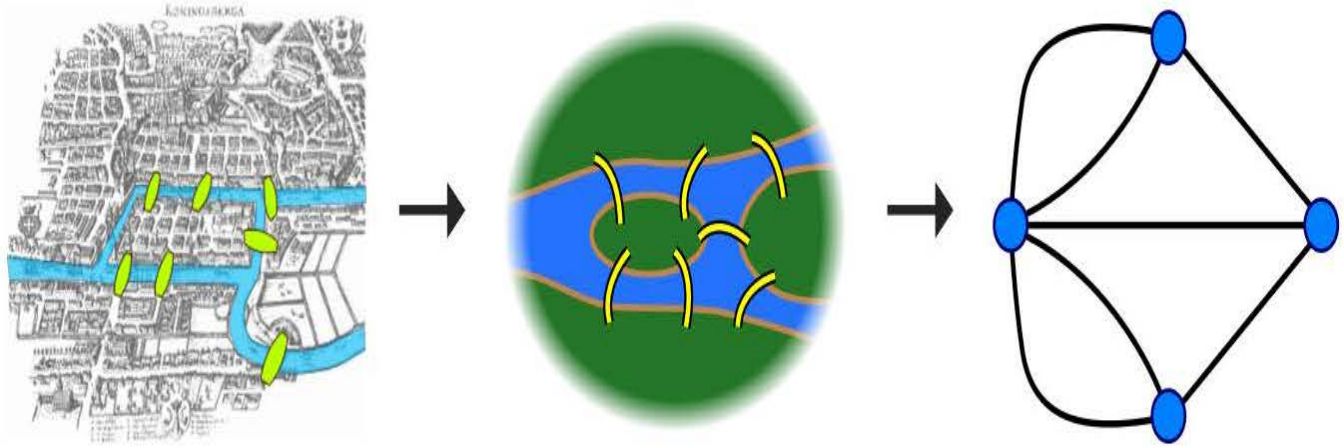


Figure: Islands in Königsberg, Prussia connected via Seven Bridges.

  - Networks also studied extensively in sociology.
    - Typical studies involve circulation of questionnaires/small networks.

Network Study

- Recent years witnessed a substantial change in network research.
  - From analysis of single small graphs (10-100 nodes) to statistical properties of large scale networks (million-billion nodes).
  - Motivated by availability of computers and computer networks that allow us to gather and analyze large scale data.

- New Analytical Approach:
  - Find statistical properties that characterize the structure of these networks and ways to measure them
  - Create models of networks (structure and dynamics)
  - Predict behavior of networks on the basis of measured structural properties and models

Graphs: Network Representation

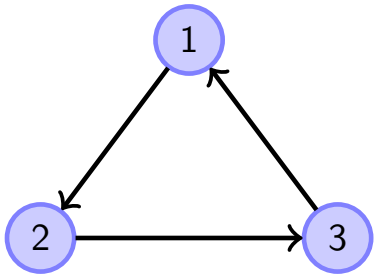
- A network G (also called a graph) is a set of nodes $N = \{1, \ldots, n\}$ joined by edges (or links).
- We will be mostly focusing on simple graphs: a graph with no self-edges or multi edges.
- A network is typically represented by its adjacency matrix which is an $n \times n$ matrix $A = [A_{ij}]_{i,j \in N}$, where

$$A_{ij} = \begin{cases} 1 & \text{if there is an edge from } j \text{ to } i, \\ 0 & \text{otherwise.} \end{cases}$$

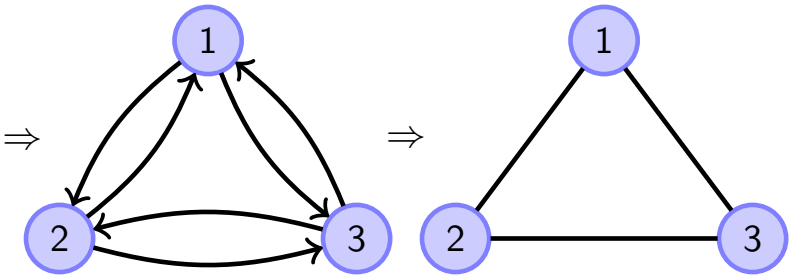
  - The edge weight A_{ij} can take on non-binary values (even negative values), representing the intensity of the interaction, in which case we refer to G as a weighted graph.
  - For simple graphs, diagonal elements are zero.
- We refer to a graph as a directed graph (or digraph) if $A_{ij} \neq A_{ji}$ and an undirected graph if $A_{ij} = A_{ji}$ for all $i, j \in N$.

Graphs: Network Representation

Example 1: $\begin{bmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \Rightarrow$



Example 2: $\begin{bmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{bmatrix} \Rightarrow$



Graphs: Network Representation

- Another representation of a graph is given by $G = (N, E)$, where $E = \{1, \ldots, m\}$ is the set of edges in the network.
  - *For directed graphs:* E is the set of "directed" edges, $(i, j) \in E$.
  - *For undirected graphs:* E is the set of "undirected" edges, $\{i, j\} \in E$.

- In Example 1, $E_d = \{(1, 2), (2, 3), (3, 1)\}$

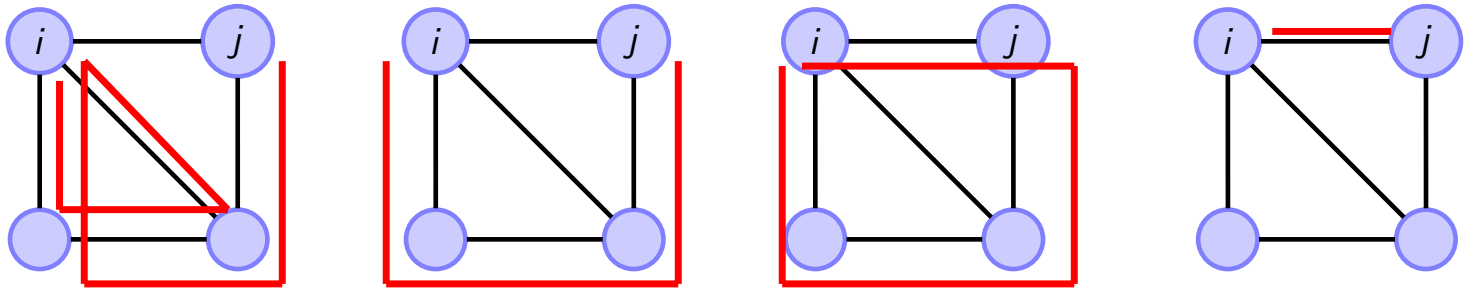
- In Example 2, $E_u = \{\{1, 2\}, \{1, 3\}, \{2, 3\}\}$

- When are directed/undirected graphs applicable?
  - Citation networks: directed
  - Friendship networks: undirected

Walks, Paths, and Cycles

- We consider "sequences of edges" to capture indirect interactions.
- For an undirected graph G:
  - A walk is a sequence of edges $\{i_1, i_2\}, \{i_2, i_3\}, \ldots, \{i_{K-1}, i_K\}$.
  - A path between nodes i and j is a sequence of edges $\{i_1, i_2\}, \{i_2, i_3\}, \ldots, \{i_{K-1}, i_K\}$ such that $i_1 = i$ and $i_K = j$, *and each node in the sequence* $i_1, \ldots, i_K$ *is distinct*.
  - A cycle is a path with a final edge to the initial node.
  - A geodesic between nodes i and j is a "shortest path" (i.e., with minimum number of edges) between these nodes.
- A path is a walk where there are no repeated nodes.
- The length of a walk (or a path) is the number of edges on that walk (or path).
- For directed graphs, the same definitions hold with directed edges (in which case we say "a path from node i to node j").

Walks, Paths, and Cycles



walk | path between i and j | cycle | shortest path

- *Note:* For simple graphs, we can calculate the number of walks of given length r between nodes i and j, $N_{ij}^{(r)}$, using the adjacency matrix:

$$N_{ij}^{(2)} = \sum_{k=1}^{n} A_{ik} A_{kj} = [A^2]_{ij}.$$

- Similarly, $N_{ij}^{(r)} = [A^r]_{ij}$.

Connectivity and Components

- An undirected graph is connected if every two nodes in the network are connected by some path in the network.
- Components of a graph (or network) are the distinct maximally connected subgraphs.
- A directed graph is
  - connected if the underlying undirected graph is connected (i.e., ignoring the directions of edges).
  - strongly connected if each node can reach every other node by a "directed path".

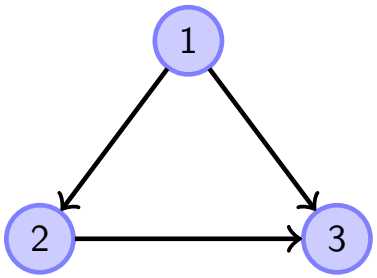


Figure: A directed graph that is connected but not strongly connected

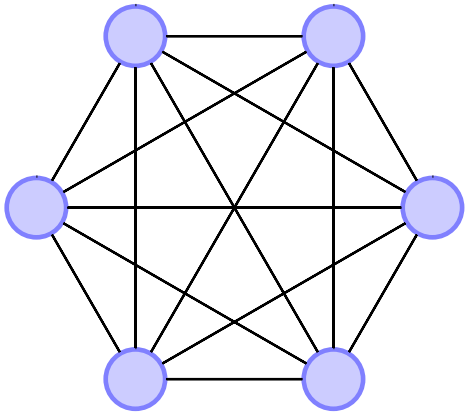
Special Graphs

- Hypergraphs: Graphs in which edges join more than two nodes (such edges are called hyper edges).
  - A social network representing families in a village.
  - Any network in which nodes connected by common membership of groups (also called affiliation networks).

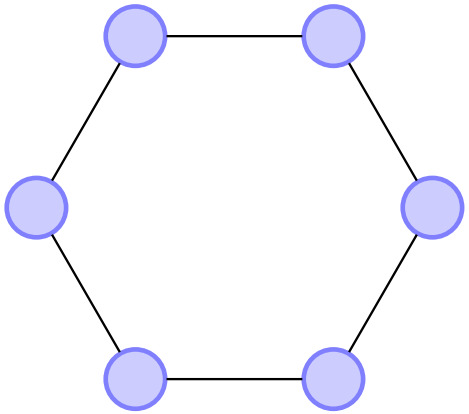
- Bipartite graphs: Graphs in which nodes decompose into two groups such that there are edges only between these groups.
  - Hypergraphs can be represented as a bipartite graph.

- A tree is a connected (undirected) graph with no cycles.
  - In a tree, there is a unique path between any two nodes.
  - A connected graph is a tree if and only if it has $n - 1$ edges.

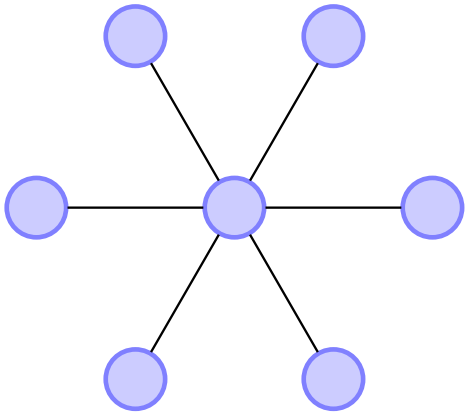
Special Graphs



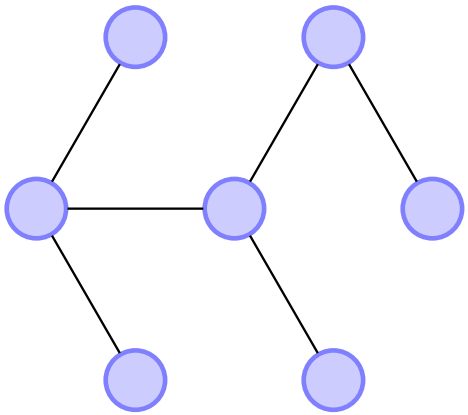
Complete graph



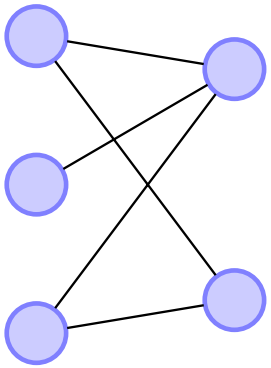
Ring



Star



Tree



Bipartite graph

actors movies

Neighborhood and Degree of a Node

- The neighborhood of node i is the set of nodes that i is connected to.
- For undirected graphs:
  - The degree of node i is the number of edges connected to i (i.e., cardinality of his neighborhood).
  - The degree of node i, k_i, can be written in terms of the adjacency matrix as
$$k_i = \sum_{j=1}^{n} A_{ij}.$$
  - An important relation that is used extensively relates number of edges to sum of degrees in the graph:
$$2m = \sum_{i=1}^{n} k_i.$$
  - Average node degree is given by $c = \frac{1}{n}\sum_{i=1}^{n} k_i$, or equivalently $c = \frac{2m}{n}$.

Neighborhood and Degree of a Node

- The maximum number of edges in a simple graph is $\binom{n}{2} = \frac{n(n-1)}{2}$.
- We will sometimes consider networks in which all nodes have the same degree. Such networks are called regular networks.
- For directed graphs:
  - Node i's in-degree is $\sum_{j=1}^{n} A_{ij}$ (number of incoming edges).
  - Node i's out-degree is $\sum_{j=1}^{n} A_{ji}$ (number of outgoing edges).

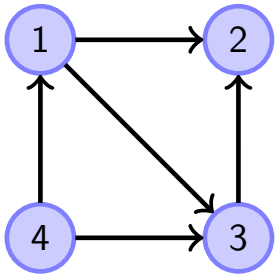


Figure: Node 1 has in-degree 1 and out-degree 2.

Königsberg bridge problem

- Can you cross each bridge exactly once in a walk?

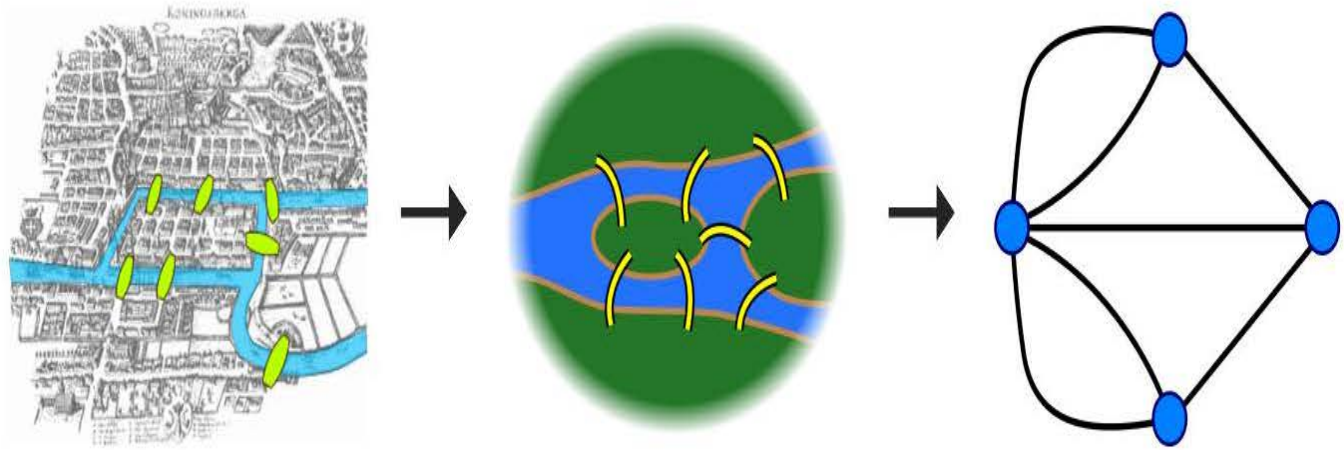


Figure: Islands in Königsberg, Prussia connected via Seven Bridges.

- Euler's insights:
  - Graph matters, not physical properties of the island, size, etc.
  - Suppose such a walk existed: starts at "node" u, ends at v. Then
    - Nodes u and v can have odd degrees
    - All other nodes must have even degrees
    - What about Königsberg bridge graph?

Degree Distributions

- The degree distribution, $P(d)$, of a network is a description of relative frequencies of nodes that have different degrees d.
  - For a given graph: $P(d)$ is a histogram, i.e., $P(d)$ is the fraction of nodes with degree d.
  - For a random graph model: $P(d)$ is a probability distribution.
- Two interesting degree distributions:
  - $P(d) \leq c\, e^{-\alpha d}$, for some $\alpha > 0$ and $c > 0$: The tail of the distribution **falls off faster than an exponential**, i.e., large degrees are unlikely.
  - $P(d) = c\, d^{-\gamma}$, for some $\gamma > 0$ and $c > 0$: Power-law distribution: The tail of the distribution is **fat**, i.e., there tend to be many more nodes with very large degrees.
    - Appear in a wide variety of settings including networks describing incomes, city populations, WWW, and the Internet
    - Also known as a scale-free distribution: a distribution that is unchanged (within a multiplicative factor) under a rescaling of the variable
    - Appear linear on a $\log - \log$ plot

Diameter and Average Path Length

- Let $l(i,j)$ denote the length of the shortest path (or geodesic) between node i and j (or the distance between i and j).
- The diameter of a network is the largest distance between any two nodes in the network:

$$\text{diameter} = \max_{i,j} l(i,j)$$

- The average path length is the average distance between any two nodes in the network:

$$\text{average path length} = \frac{\sum_{i \geq j} l(i,j)}{\frac{n(n-1)}{2}}$$

- Average path length is bounded from above by the diameter; in some cases, it can be much shorter than the diameter.
- If the network is not connected, one often checks the diameter and the average path length in the largest component.

Clustering Coefficient

- Measures the extent to which my friends are friends with one another.
- This clustering measure is represented by the overall clustering coefficient $Cl(G)$, given by

$$Cl(G) = \frac{3 \times \text{number of triangles in the network}}{\text{number of connected triples of nodes}},$$

where a "connected triple" refers to a node with edges to an unordered pair of nodes.

  - Note that $0 \leq Cl(G) \leq 1$.
  - $Cl(G)$ measures the fraction of triples that have their third edge filled in to complete the triangle.
  - Also referred to as network transitivity: measures the extent that a friend of my friend is also my friend.

Clustering Coefficient

- Another measure of clustering is defined on an individual node basis: The individual clustering for a node i is

$$Cl_i(g) = \frac{\text{number of triangles connected to vertex } i}{\text{number of triples centered at } i}.$$

- The average clustering coefficient is $Cl^{Avg}(g) = \frac{1}{n}\sum_i Cl_i(g)$.

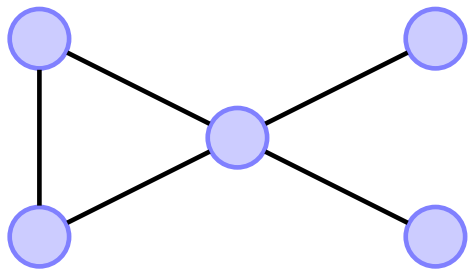


Figure: The overall clustering coefficient for this network is 3/8. The individual clustering for the nodes are 1, 1, 1/6, 0, and 0.

Centrality

- A micro measure that captures importance of a node's position in the network. Many different centrality measures:

- Degree centrality: For node i,

$$C_i = k_i, \quad \text{where } k_i \text{ is the degree of node } i.$$

  - For directed networks, both in-degree and out-degree can be used as centrality measures.
  - Simple, but intuitive: individuals with more connections have more influence and more access to information.
  - Does not capture "cascade effects": importance better captured by having connections to important nodes
    - for example, eigenvector centrality which we shall study soon

Centrality

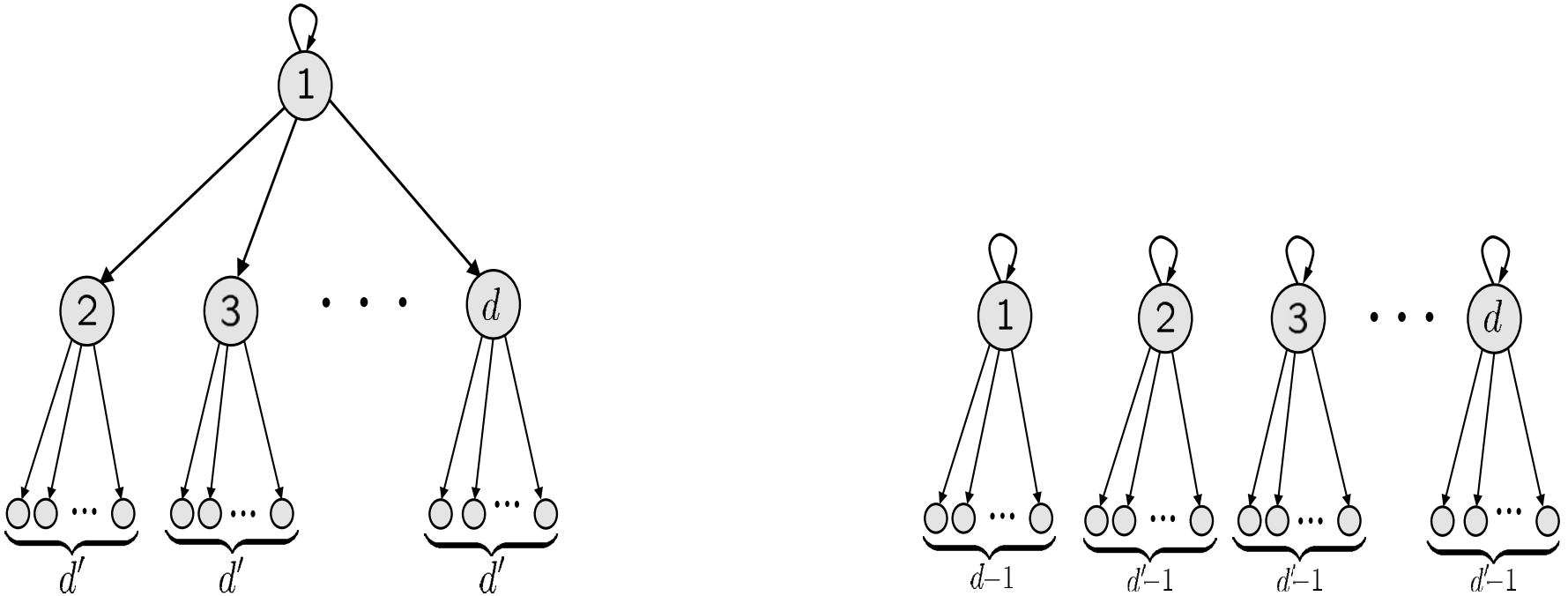


Figure: Degree centrality of Nodes 2,. . . , d are the same in both graphs.

Centrality

- Closeness centrality: Tracks how close a given node is to any other node: For node i, one such measure is

$$C_i = \left(\frac{1}{n-1} \sum_{j \neq i} d_{ij} \right)^{-1}$$

where d_{ij} is the distance between i and j.

- Nodes which are close to other nodes on average have high centrality: such nodes may have more direct influence on others.

Centrality

- Closeness centrality: An Example
- Network of movie actors: two actors are connected if they work together
- Highest centrality 0.4143 for Christopher Lee
  - Entered into the Guinness Book of World Records in 2007 for most screen credits
- Lowest centrality 0.1154 for Leia Zanganeh
  - An Iranian Theatre and Film Actress.

- Limitations
  - Spans a very small range.
  - For disconnected networks, leads to zero centrality for all nodes!

Centrality

- Betweenness centrality: Measures the extent to which a node lies on paths between other nodes.

- Let n_{st}^i be the number of shortest paths between nodes s to t that pass through a node i
- Let g_{st} be the total number of shortest paths from s to t.
- Then, betweenness centrality of node i is defined as

$$C_i = \sum_{s,t} \frac{n_{st}^i}{g_{st}}.$$

- Nodes with high betweenness centrality may have high influence since they control information passing between others.

Centrality

- Betweenness centrality differs from others: it is not a measure of how well-connected a node is.
- Moreover, it spans a wide range.
- Let us consider our example of Network of actors
- Highest centrality $7.47x10^8$ for Fernando Rey
  - A Spanish film, theatre, and television actor
  - Worked in both Europe and the United States
  - Starred in movies like French Connection (US), Triastana (Spanish), Cet obscur objet du dsir (French), Ese oscuro objeto del deseo (Spanish)
- Lowest centrality $8.91x10^5$.

Centrality

- And before we forget, in the network of actors
  - Highest degree centrality 98734 for Bess Flowers
  - She was an American actress best known for her work as an extra in hundreds of films
- So which of these centrality measure is really useful?
  - Well, it depends on what is the "objective"
- In a friendship network, degree centrality would correspond to who is the most popular kid. This might be important for certain questions.
- Closeness centrality would correspond to who is closest to the rest of the group, so this would be relevant if we wanted to understand who to inform or influence for information to spread to the rest of the network.
- Betweenness would be relevant if the thought experiment was which individuals would have to be taken out of the network in order to break the network into separate clusters.

Centrality and Power in a Network

- One application of this is to intermarriage network of Florentine families.
- The Medicis emerged as the most influential family in 15th century Florence. Cosimo de Medici ultimately formed the most politically powerful and economically prosperous family in Florence, dominating Mediterranean trade.
- The Medicis, to start with, were less powerful than many other important families, both in terms of political dominance of Florentine institutions and economic wealth.
- How did they achieve their prominence?
- It could just be luck (in social science, we have to be very careful to distinguish luck from a systematic pattern, and correlation from causation).
- An interesting explanation, eschewing luck, is offered by Padgett and Ansell (1993) "Robust Action and the Rise of the Medici"— they were the most powerful family because of their high betweenness centrality, which meant that they were part of many deals between families supported by marriage linkages.

Centrality and Power in a Network

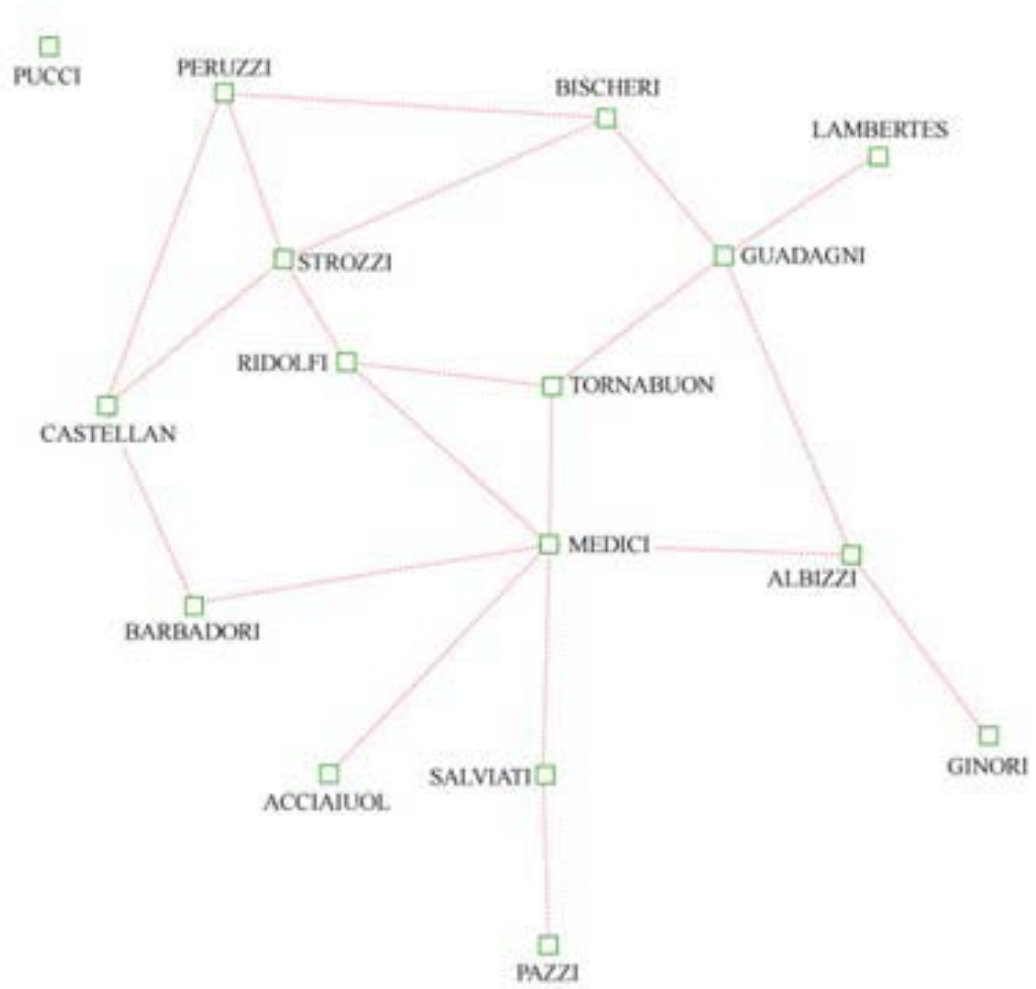


Figure: Political and friendship blockmodel structure (Padgett and Ansell 1993)

Centrality and Power in a Network

- It turns out that Medicis has a very high betweenness centrality 0.522.
- No other family has betweenness centrality greater than 0.255.
- So the Medicis may have played a central role in holding the network of influential families in Florence together and thus gained "power" via this channel.

14.15J/6.207J Networks
Spring 2018