Social and Economic Networks

Representing and Measuring Networks

Xiang Sun

September 13–20, 2017

Outline

Section 1

Types of networks

Networks in the real world

- A network is a set of items (nodes or vertices) connected by edges or links.
- Systems taking the form of networks abound in the world.

Types of networks

- Social and economic networks: A set of people or groups of people with some pattern of contacts or interactions between them.
  - Facebook, friendship networks, business relations between companies, intermarriages between families, labor markets
  - Questions: Degree of connectedness, homophily, small-world effects
- Information networks: Connections of "information" objects.
  - Network of citations between academic papers, World Wide Web (network of Web pages containing information with links from one page to other), semantic (how words or concepts link to each other)
  - Questions: Ranking, navigation

Types of networks (Cont.)

- Technological networks: Designed typically for distribution of a commodity or service.
  - Infrastructure networks: e.g., Internet (connections of routers or administrative domains), power grid, transportation networks (road, rail, airline, mail)
  - Temporary networks: e.g., ad hoc communication networks, sensor networks, autonomous vehicles
  - Questions: Does network structure support performance? Fragility? Cascading failures?
- Biological networks: A number of biological systems can also be represented as networks.
  - Food web, protein interaction network, network of metabolic pathways

Network study

- Historical study of networks:
  - Mathematical graph theory: One of the pillars of discrete mathematics
    - Started with Euler's celebrated 1735 solution of the Konigsberg bridge problem.
  - Networks also studied extensively in sociology.
    - Typical studies involve circulation of questionnaires, leading to small networks of interactions.
- 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.

Network study: New analytical approach

- Find statistical properties that characterize the structure of these networks and ways to measure them.
- Create models of networks.
- Predict behavior of networks on the basis of measured structural properties and models.

Section 2

Graphs

Subsection 1

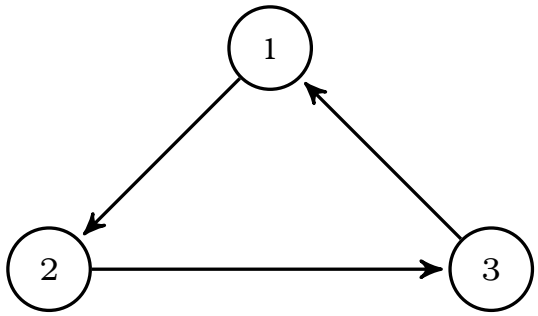
Graph representation

Graph

- We represent a network by a graph (N, g), which consists of a set of nodes $N = \{1, 2, \ldots, n\}$ and an $n \times n$ matrix $g = [g_{ij}]_{i,j \in N}$ (referred to as an adjacency matrix (邻接矩阵)), where $g_{ij} \in \{0, 1\}$ represents the availability of an edge from node i to node j.
- We refer to a graph as a directed graph (有向图) (or digraph) if $g_{ij} \neq g_{ji}$ and an undirected graph (无向图) if $g_{ij} = g_{ji}$ for all $i, j \in N$.
- When are directed/undirected graphs applicable?
  - Citation networks: directed
  - Friendship networks: undirected

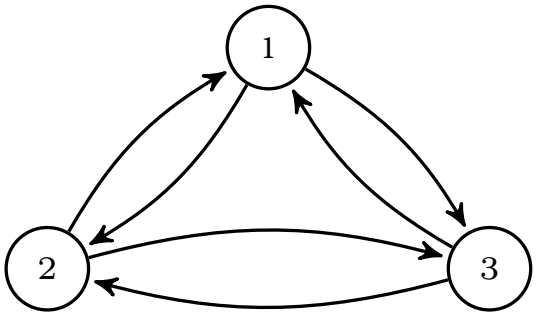
Graph: Example 1

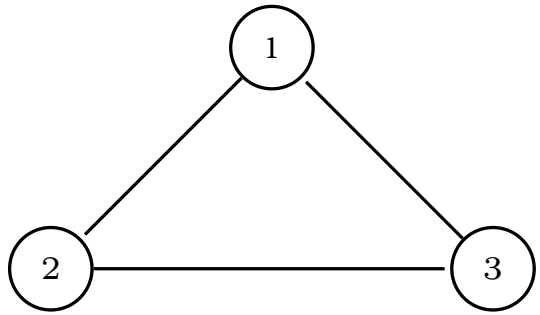
$$\begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix}$$



Graph: Example 2

$$\begin{bmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{bmatrix}$$





Graph (Cont.)

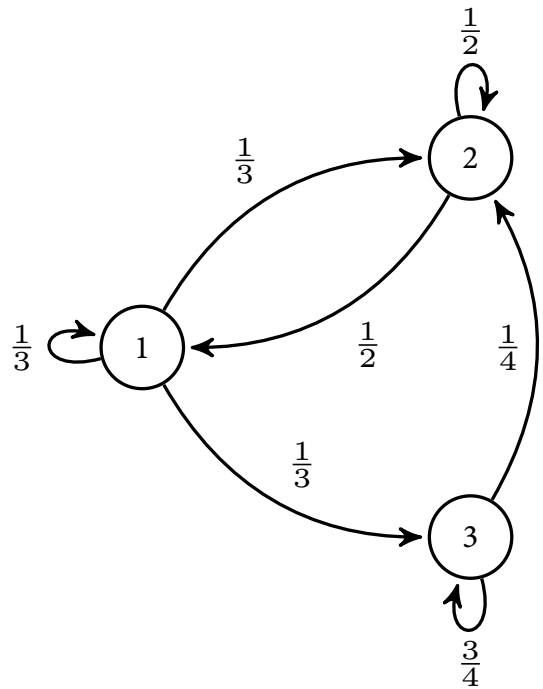
- Another representation of a graph is given by (N, E), where E is the set of edges in the network.
  - For directed graphs: E is the set of "directed" edges, i.e., $(i, j) \in E$.
  - For undirected graphs: E is the set of "undirected" edges, i.e., $\{i, j\} \in E$.
- In Example 1, $E_d = \{(1, 2), (2, 3), (3, 1)\}$.

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

Weighted graph

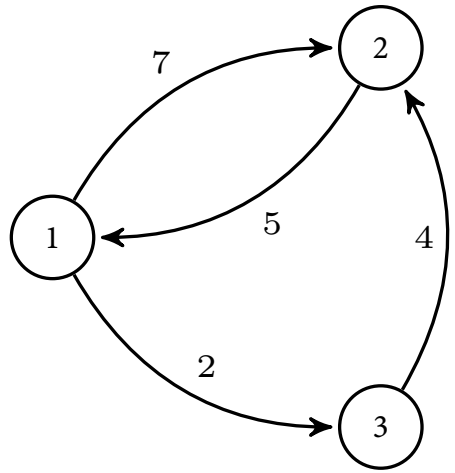
The edge weight $g_{ij} > 0$ can also take on non-binary values, representing the intensity of the interaction, in which case we refer to (N, g) as a weighted graph (权重图).

$$\begin{bmatrix} \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ \frac{1}{2} & \frac{1}{2} & 0 \\ 0 & \frac{1}{4} & \frac{3}{4} \end{bmatrix} \Longleftrightarrow$$



Weighted graph (Cont.)

$$\begin{bmatrix} 0 & 7 & 2 \\ 5 & 0 & 0 \\ 0 & 4 & 0 \end{bmatrix} \Longleftrightarrow$$



Conventions

- We will use the terms network and graph interchangeably.
- We will sometimes use the notation $(i, j) \in g$ (or $\{i, j\} \in g$) to denote $g_{ij} = 1$.
- Self-links or loops (圈) will often not have any real meaning or consequence, and so whether we set $g_{ii} = 1$ or $g_{ii} = 0$ as a default will most often (but not always!) be irrelevant. Unless otherwise indicated in what follows, assume that $g_{ii} = 0$ for all i.

Subsection 2

Walks, paths, and cycles

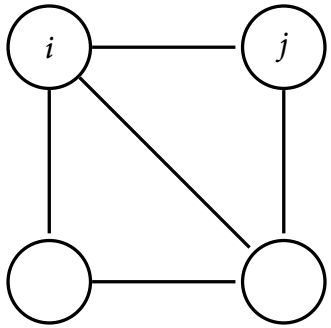
Walks, paths, and cycles

We consider "sequences of edges" to capture indirect interactions. For an undirected graph (N, g):

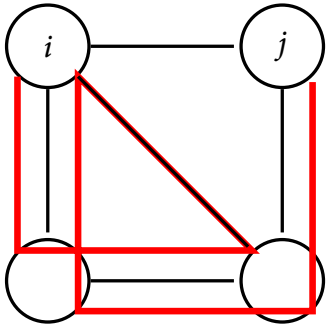
- A walk (链) is a sequence of edges $\{i_1, i_2\}, \{i_2, i_3\}, \ldots, \{i_{K-1}, i_K\}$.
- A path (路径) between nodes i and j is a sequence of edges $\{i_1, i_2\}, \{i_2, i_3\}, \ldots, \{i_{K-1}, i_K\}$ such that $i_1 = i$ and $i_K = j$, and each node in the sequence $i_1, i_2, \ldots, i_K$ is distinct.

  A path is a walk where there are no repeated nodes.
- A cycle (循环) is a path with a final edge to the initial node.
- A geodesic (测地线) between nodes i and j is a "shortest path" (i.e., with minimum number of edges) between these nodes.

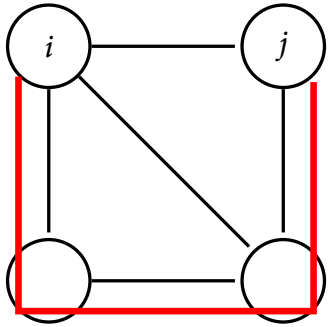
Walks, paths, and cycles: Example



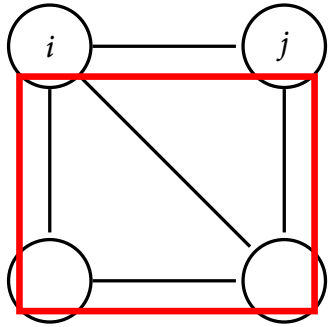
graph



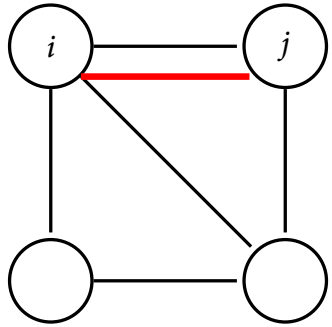
walk



path



cycle



geodestic

Walks, paths, and cycles (Cont.)

- 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").
- Under the convention $g_{ii} = 0$, the matrix g^2 tells us number of walks of length 2 between any two nodes:
  - $(g \times g)_{ij} = \sum_k g_{ik} g_{kj}$.
  - Similarly, g^k tells us number of walks of length k.

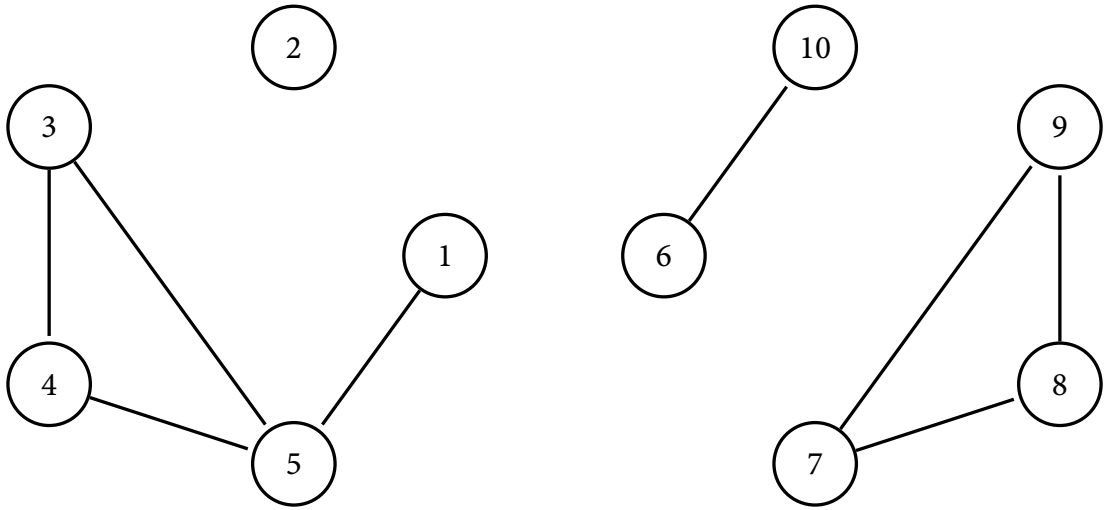
Subsection 3

Connectivity and components

Connectivity and components

- An undirected graph is connected (连通) if every two nodes in the network are connected by some path in the network.
- Components (分支) of a graph (or network) are the distinct maximally connected subgraphs.

Connectivity and components: Example



Four components:

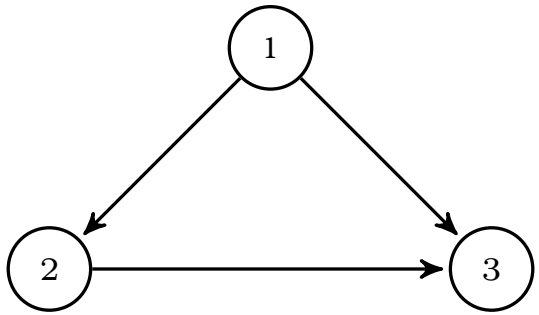
- the node 2 together with an empty set of links,
- the nodes $\{1, 3, 4, 5\}$ together with links $\{\{1, 5\}, \{3, 5\}, \{3, 4\}, \{4, 5\}\}$,
- the nodes 6 and 10 together with the link $\{\{6, 10\}\}$,
- and the nodes $\{7, 8, 9\}$ together with the links $\{\{7, 8\}, \{7, 9\}, \{8, 9\}\}$.

Connectivity and components: Example

A directed graph is

- connected (连通) if the underlying undirected graph is connected (i.e., ignoring the directions of edges).
- strongly connected (强连通) if each node can reach every other node by a "directed path".

Example: A directed graph that is connected but not strongly connected.

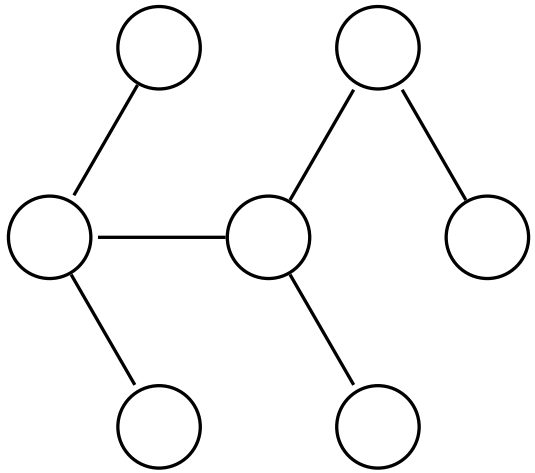


Subsection 4

Trees, stars, rings and complete graphs

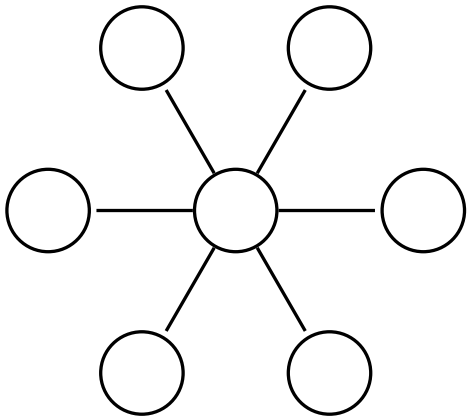
Trees

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

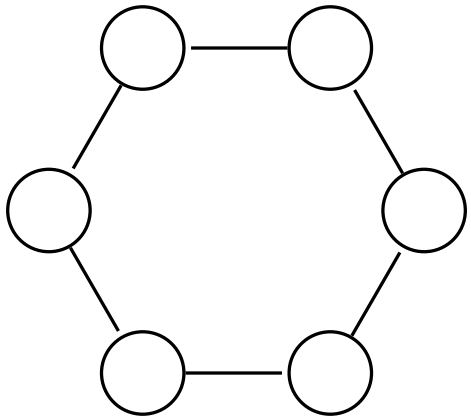


Stars, rings and complete graphs

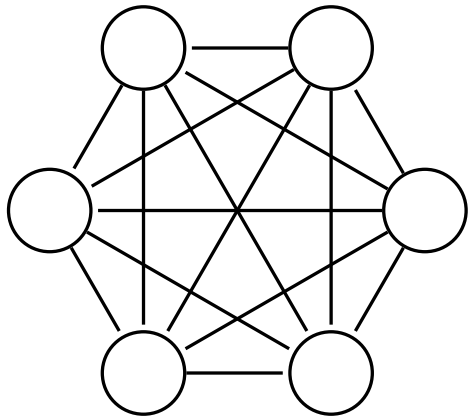
- A star (星) is a network such that there exists some node i such that every link in the network involves node i. In this case i is referred to as the center of the star.
- The complete graph (完全图) is one where all possible links are present, so one where where $g_{ij} = 1$ for all $i \neq j$.
- A ring (环) is a network that has a single cycle and such that each node in the network has exactly two neighbors.



Star



Ring



Complete graph

Subsection 5

Neighborhood and degree of a node

Neighborhood

- The neighborhood (邻域) of node i is the set of nodes that i is connected to:

$$N_i(g) = \{j \in N \mid g_{ij} = 1\}.$$

- The two-neighborhood of a node i is

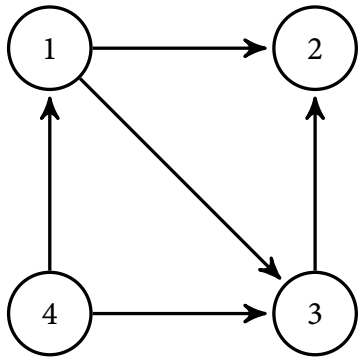
$$N_i^2(g) = N_i(g) \cup \left(\cup_{j \in N_i(g)} N_j(g)\right).$$

- The k-neighborhood of a node i is

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

Degree

- For undirected graphs:
  - The degree (度) of node i is the number of edges that involve i (i.e., cardinality of his neighborhood).
- For directed graphs:
  - Node i's in-degree (入度) is $\sum_j g_{ji}$.
  - Node i's out-degree (出度) is $\sum_j g_{ij}$.
- Node 1 has in-degree 1 and out-degree 2:



Section 3

Summary statistics and characteristics of networks

Summary statistics and characteristics of networks

- While a small network can be visualized directly by its graph (N, g), larger networks can be more difficult to envision and describe.
- Therefore, we define a set of summary statistics or quantitative performance measures to describe and compare networks (focus on undirected graphs):
  - Global patterns of networks
    degree distributions, diameter, average path length
  - Local patterns of networks
    clustering, transitivity
  - Positions in networks
    centrality

Benchmark: Random graphs

- Start with n nodes.
- Each link is formed independently with some probability p.
- Serves as a benchmark $G(n, p)$.
  - If we see some network out there in the real world and we know what the properties of a random network of n nodes with some probability p was, then we can compare the real world network to this benchmark network.
  - Does it look like something systematic is going on?
  - Does this network look systematically different than if nature had just picked links at random?
  - Does it look like something is systematically different?

Models

- Gilbert–Elliott model $G(n, p)$
- Erdos–Renyi model $G(n, M)$
- Scale-free network (Barabasi–Albert model) for fat tails
- Watts–Strogatz model for small worlds

Subsection 1

Degree distributions

Degree distribution

- For each node, the possible value of the degree could be one of $\{0, 1, \ldots, n-1\}$.

$\Rightarrow$ What is the proportion/fraction for each value?

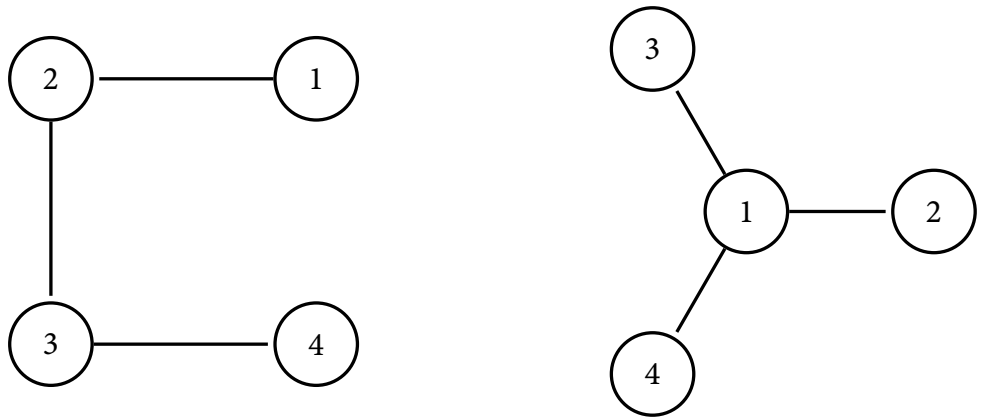
- The degree distribution (度分布) of a network is a description of relative frequencies of nodes that have different degrees.
- We use $P(d)$ to denote the fraction of nodes with degree d.

$\Rightarrow$ The degree distribution can be described as an n-dimensional vector:

$$\big(P(0), P(1), P(2), \ldots, P(n-1)\big).$$

Degree distribution (Cont.)

- Why do we care about the degree distribution?
- Example: Average degree tells only part of the story.



- These two networks have the same average degree $\frac{3}{2}$.
- However, they are quite different.

Degree distribution: Random graph

- The probability that a node has d links is binomial (二项分布):

$$\binom{n-1}{d} p^d (1-p)^{n-1-d}.$$

- For large n and constant $n \times p$, this is approximately a Poisson distribution (泊松分布):

$$\frac{1}{d!}(n-1)^d p^d e^{-(n-1)p}.$$

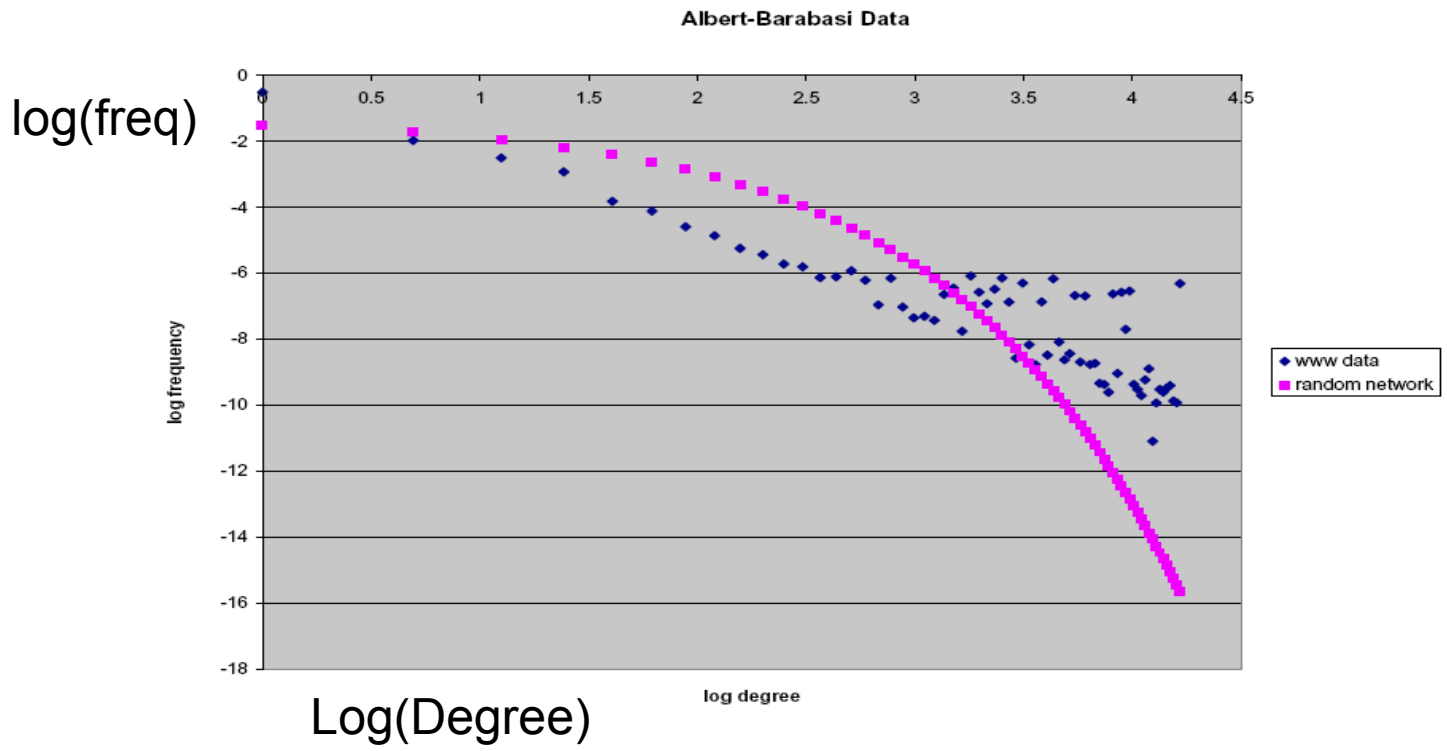
- Poisson/binomial distribution gives us an approximation.

Degree distribution: Fat tails

- Fat tails (肥尾): More high and low degree nodes than predicted at random.
- Price (1965): Citation network
  - Too many with 0 citations, too many with high numbers of citations to have citations drawn at random.
- Related to other settings (wealth, city size, word usage…): Pareto (1896), Yule (1925), Zipf (1949), Simon (1955).

Degree distribution: Fat tails (Cont.)

Albert, Jeong, Barabasi (1999)



Degree distribution: Fat tails (Cont.)

- Albert, Jeong, Barabasi (1999).
- $P_{out}(k)$ and $P_{in}(k)$ is the fractions of documents that have k outgoing and incoming links, respectively.
- They found that $P_{out}(k)$ and $P_{in}(k)$ follow a power law over several orders of magnitude, remarkably different not only from the Poisson distribution predicted by the classical theory of random graphs.

Degree distribution: Fat tails (Cont.)

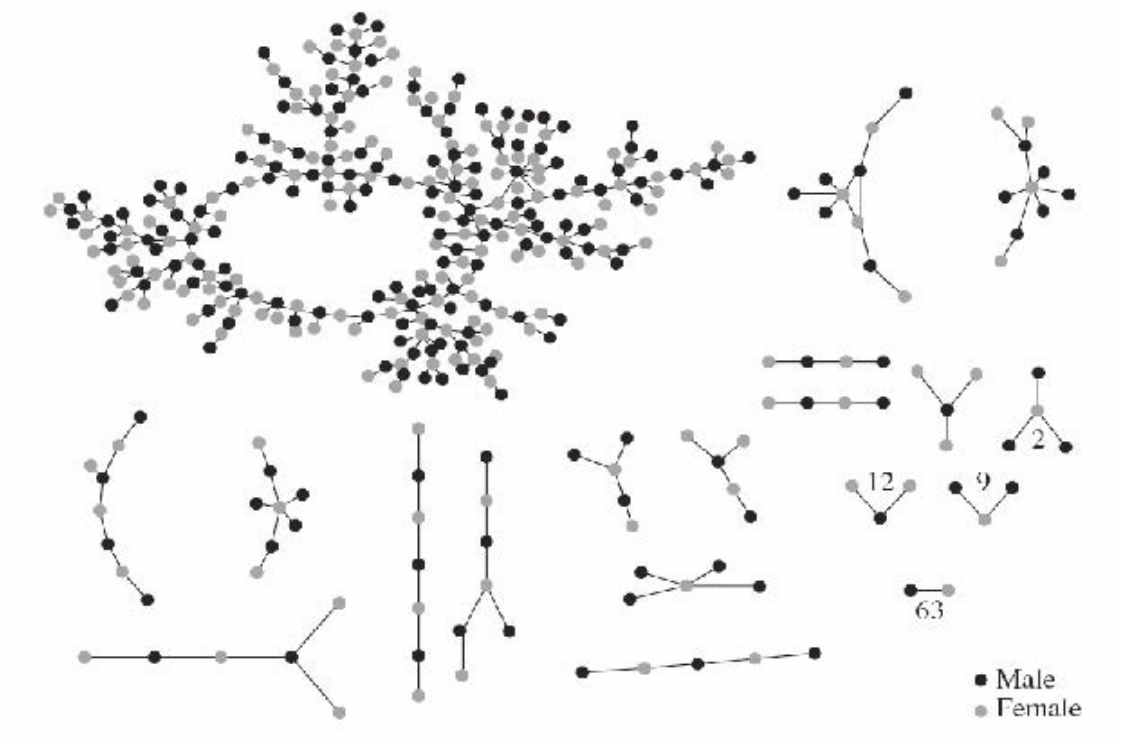
- To characterize the "fat tails", we consider the power-law distribution (幂分布):

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

  for some $\gamma > 0$ and $c > 0$.
- Appear linear on a log – log plot.
- Also known as a scale-free distribution (标度自由分布/无标度分布): a distribution that is unchanged (within a multiplicative factor) under a rescaling of the variable.

Poisson distribution vs power-law distribution

Bearman, Moody, and Stovel (2004): High school romance network



Poisson distribution vs power-law distribution (Cont.)

- Fit: Random graph 0.99 (better)
- Fit: Power-law distribution 0.84
- Degree distributions of different networks can have different properties.

Subsection 2

Diameter and average path length

Diameter and average path length

- How close are nodes to each other?
- How long does it take to get from one note to another node?
- How fast will information spread?
- How does it depend on network density?

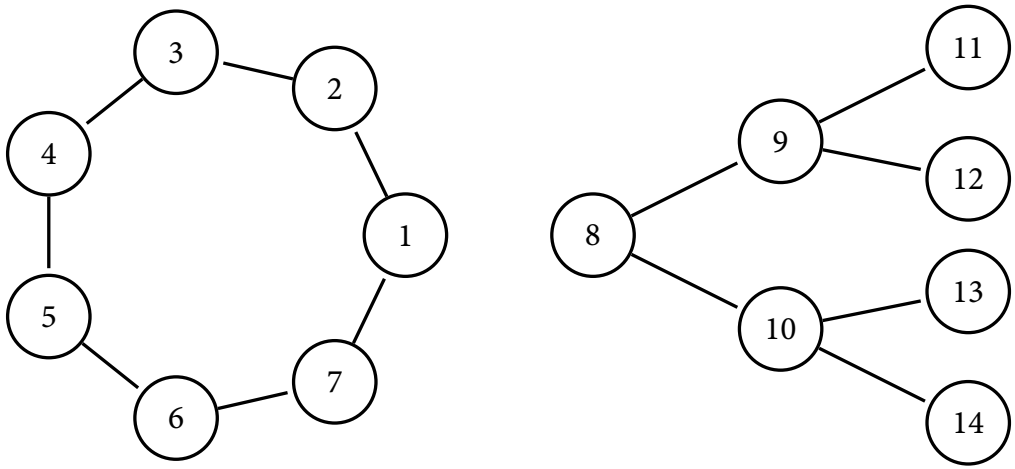
Diameter

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

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

- If the network is unconnected, then its diameter is the diameter for its largest component (giant component).

Diameter: Example



- Both networks have an average degree of 2, but they are very clearly different in structure.
- The diameter of a ring of n nodes is either $\frac{n}{2}$ or $\frac{n-1}{2}$.
- The diameter of a binary tree of n nodes is $2\log_2(n+1) - 2$:

$$1 + 2^1 + 2^2 + \cdots + 2^k = n \text{ and the diameter is } 2k.$$

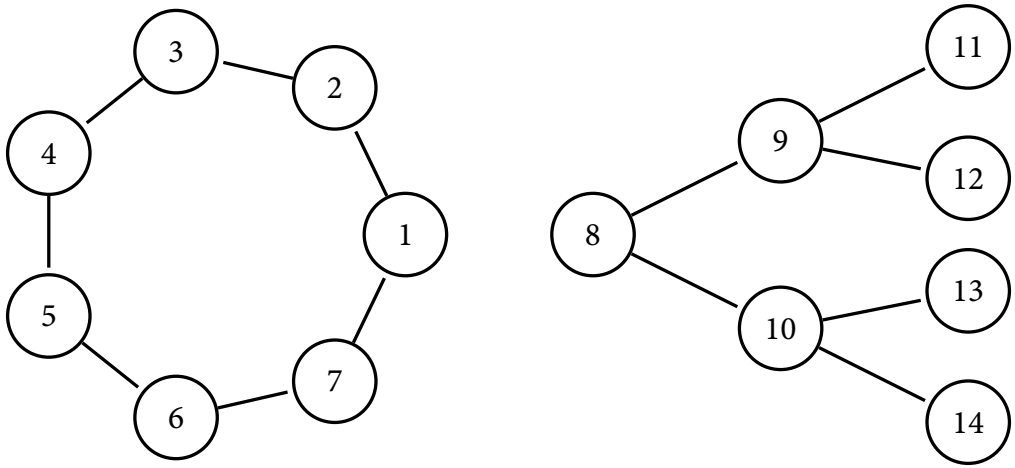
Average path length

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

$$\text{average path length} = \frac{\sum_{i \geq j} \ell(i,j)}{\binom{n}{2}}.$$

- If the network is unconnected, one often checks the average path length in the grant component.

Average path length: Example



- The average path length of a ring of n nodes is ?
- The average path length of a binary tree of n nodes is ?

Diameter vs average path length

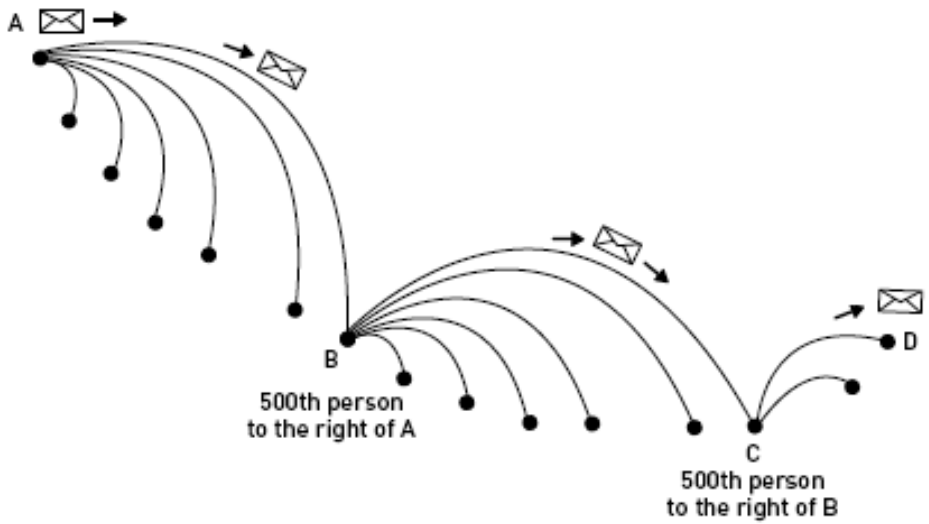
- Diameter: the maximum geodesic.

  Average path length: the average geodesic.
- Example: There is one pair of nodes which are very far from each other, but a lot of the other ones are relatively well connected to each other.
- Average path length is bounded from above by the diameter; in some cases, it can be much shorter than the diameter.

Small average path length and diameter

Milgram (1967) letter experiments:

- Median 5.5 for the 25% (fairly high in terms of response rate for this kind of participation in an experiment) that made it.
- People didn't get to see the network. (we are able to figure out what the most efficient path is if we know all the connections)



Small average path length and diameter: Real world

Co-authorship studies

	Biology	Economics	Math	Physics
Num of nodes	1520521	81217	253339	52909
Average degree	15.5	1.7	3.9	9.3
Average path length	4.9	9.5	7.6	6.2
Diameter	24	29	27	20

Small average path length and diameter: Real world (Cont.)

- WWW
  - Adamic, Pitkow (1999): mean 3.1 (85.4% possible of 50 million pages)
- Facebook
  - Backstrom et al. (2012): mean 4.74 (721 million users)

Why do we have short average path lengths

- Links are dense enough so that the network is connected almost surely:
$$d(n) \geq (1+\epsilon)\log(n) \text{ for some } \epsilon > 0.$$
- The network is not too complete: $\frac{d(n)}{n} \to 0$.
- Theorem: If $d(n) \geq (1+\epsilon)\log(n)$ for some $\epsilon > 0$ and $\frac{d(n)}{n} \to 0$. Then for large n, average path length and diameter are approximately proportional to $\frac{\log(n)}{\log(d(n))}$.
- This gives us an idea of why we end up with a very short average path lengths and short diameters if we had something in a random graph.

Subsection 3

Clustering

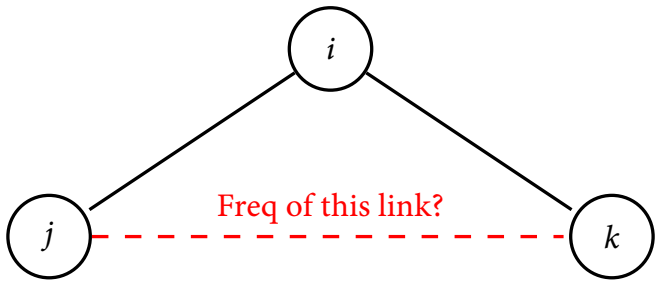
Clustering

- How dense is a network at a local level?

  What fraction of my friends are friends of each other?
- The individual clustering for a node i is

$$\mathrm{Cl}_i(g) = \frac{\text{number of triangles connected to } i}{\text{number of triples centered at } i},$$

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

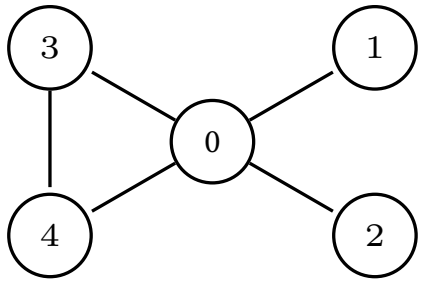


Clustering (Cont.)

- The average clustering coefficient (平均集群性) is

$$\mathrm{Cl}^{avg}(g) = \frac{1}{n}\sum_{i\in N}\mathrm{Cl}_i(g) = \frac{1}{n}\sum_{i\in N}\frac{\#\{kj \in g \mid k,j \in N_i(g)\}}{\#\{kj \mid k,j \in N_i(g)\}}.$$

- Example: The individual clustering for the nodes are $\frac{1}{6}$, 0, 0, 1 and 1. The average clustering coefficient for this network is $\frac{13}{30}$.

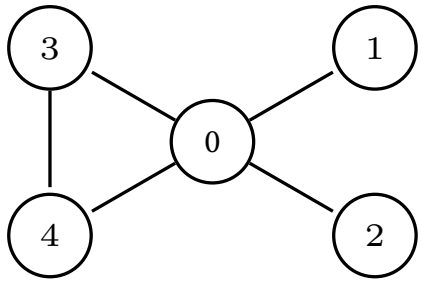


Clustering (Cont.)

- The overall clustering coefficient (整体集群性) $\mathrm{Cl}(g)$ is given by

$$\begin{aligned}\mathrm{Cl}(g) &= \frac{\sum_{i\in N} \#\{kj \in g \mid k,j \in N_i(g)\}}{\sum_{i\in N} \#\{kj \mid k,j \in N_i(g)\}} \\ &= \frac{3 \times \text{number of triangles in the network}}{\text{number of connected triples of nodes}}.\end{aligned}$$

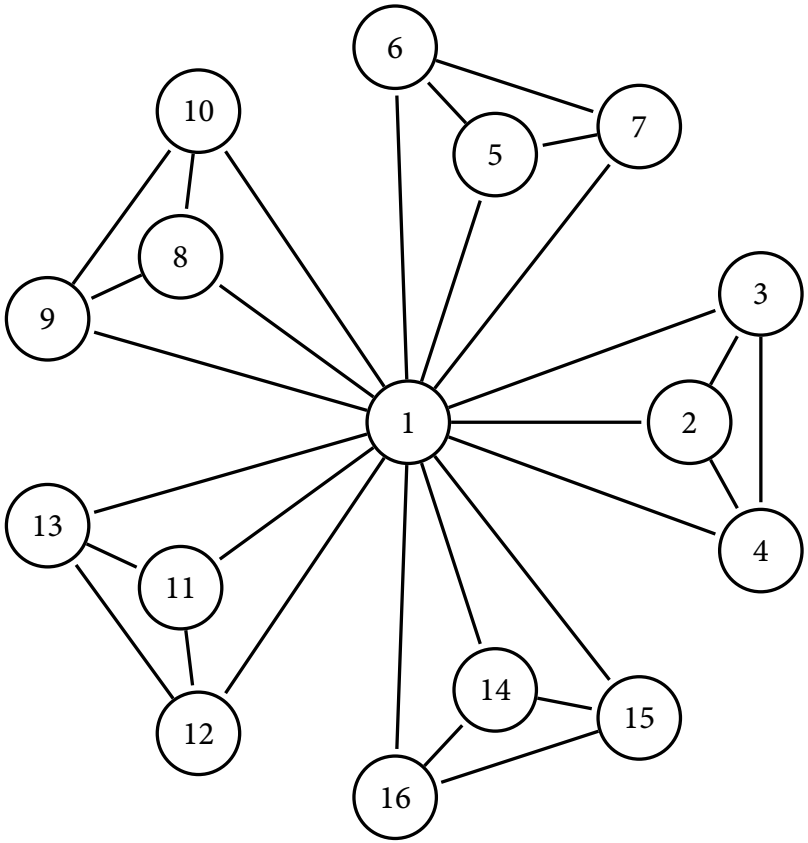
- Example: The overall clustering coefficient for this network is $\frac{3}{8}$.



Clustering (Cont.)

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

Average clustering vs overall clustering



If we have $3n + 1$ nodes, then

$$\mathrm{Cl}^{avg}(g) = \frac{1}{3n+1}(3n \times 1 + \frac{3n}{\binom{3n}{2}})$$

$$\to 1$$

$$\mathrm{Cl}(g) = \frac{3 \times n \times 4}{3n \times 3 + \binom{3n}{2}} \to 0$$

Clustering: Random graph

- The individual clustering for a node in random graph $G(n, p)$ is simply p.

$\Rightarrow$ The average clustering coefficient is p.

- The overall clustering coefficient is also p.
- Average and overall clustering tend to 0, if max degree is bounded and network becomes large.

  If np is constant (or the degree is bounded), then $p \to 0$ when $n \to \infty$.
- Random networks are going to tend to have very low clustering.

Clustering: Real world

Looking in data across a variety of different kinds of data sets, clustering are much higher than those would have occurred in random networks.

	Biology	Economics	Math	Physics
Num of nodes	1520521	81217	253339	52909
Average degree	15.5	1.7	3.9	9.3
Overall clustering	0.09	0.16	0.15	0.45
Random	0.00001019	0.00002093	0.00001539	0.00017578

Clustering: Real world (Cont.)

- Prison friendships: 0.31 (MacRae 1960) vs 0.134 (random network)
- WWW: 0.1078 for web links (Adamic 1999) vs 0.00023 (random network)
- Florentine marriages: 0.46 vs 0.29 (random network)

Subsection 4

Centrality

Centrality

- Is a node important, influential, central, or powerful?

$\Rightarrow$ To understand how difference nodes are positioned in a network.

- How to describe individual characterstics?
  - Degree
  - Clustering
  - Distance
  - Centrality

Four different measures

- Degree centrality: connectedness
- Closeness/decay centrality: ease of reaching other nodes
- Betweenness centrality: role as an intermediary, connector
- Eigenvector/Bonacich centrality/prestige: you are important if your friends are important.
- These things are capturing different ideas and different aspects.

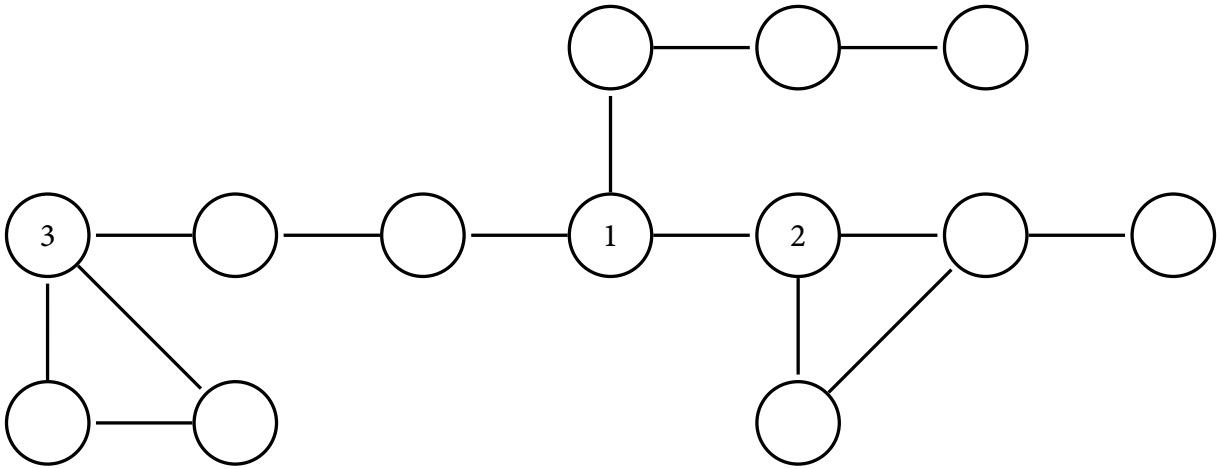
Degree centrality

- Degree centrality: for node i,

$$\frac{d_i(g)}{n-1}, \text{ where } d_i(g) \text{ is the degree of node } i.$$

- Degree captures the connectedness.
- In order to make it a measure between zero and one, we can normalize the degree (dividing $n-1$).

Degree centrality: Example



The degree centralities for nodes 1–3 are $\frac{3}{13-1} = \frac{1}{4}$.

Degree centrality: Issues

- Degree centrality misses a lot.
- Example: Node 3 has the same degree as node 1 or node 2.

  However, node 3 is less central than the other nodes.
- Degree centrality is not really gathering all of position, and it is just saying how big is your local neighborhood.
- It is not saying where you are positioned in the network, or how central you are in a deeper sense.

Closeness centrality

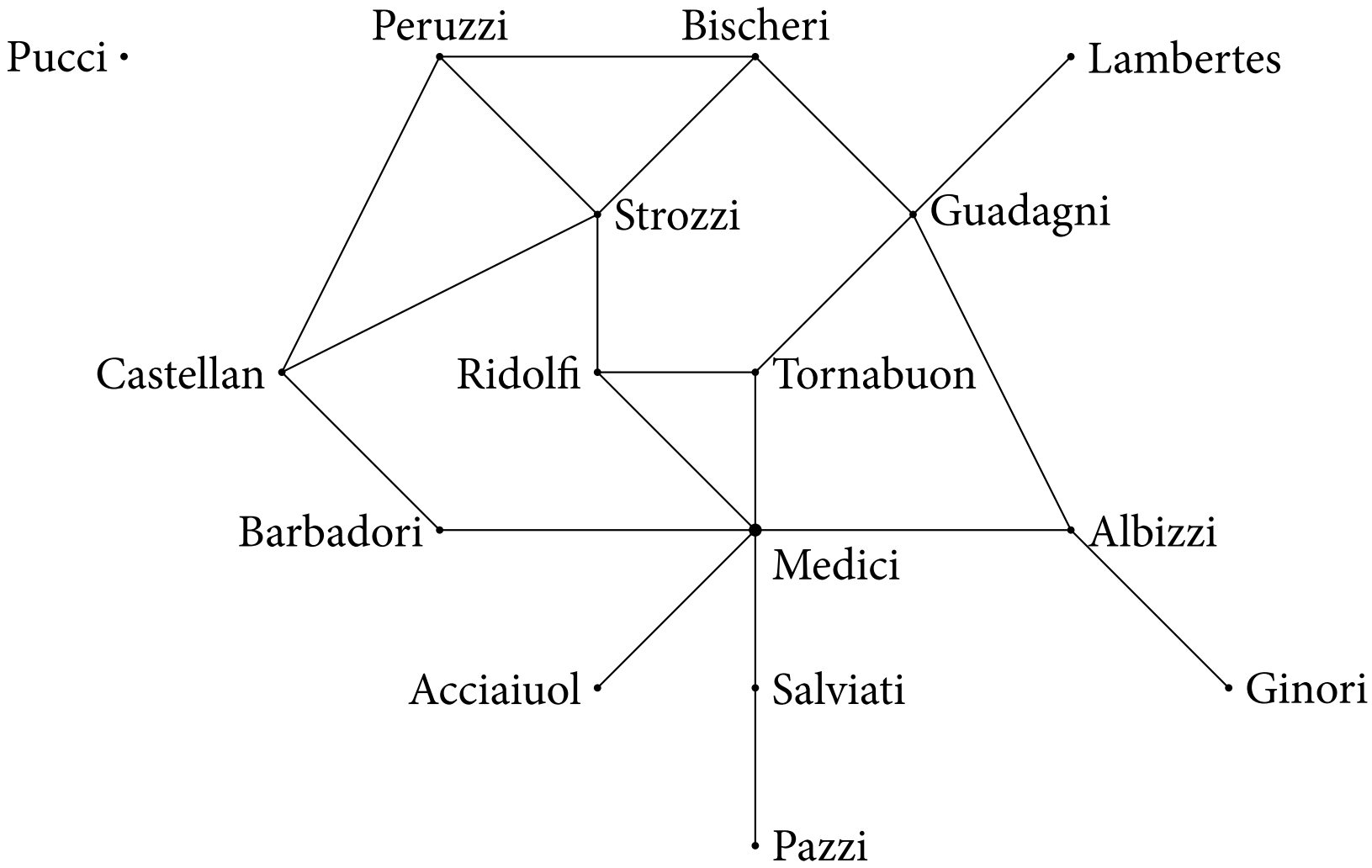
- Closeness centrality (接近中心性): for node i

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

  where $\ell(i,j)$ is the distance between i and j.
- It describes the relative distances to other nodes.

  It tracks how close a given node is to any other node.
- If I am a distance 1 from everyone, then the closeness centrality is 1.

Closeness centrality: Example



Closeness centrality: Example

- Ignoring the Pucci now because if we add them to everybody and we think of everybody has being infinitely distant from them, then everybody would have closeness centrality of zero.
- Medici: $\frac{14}{25}$.
- Strozzi: $\frac{14}{32}$.
- Guadagni: $\frac{14}{26}$.
- Tornabuon: $\frac{14}{29}$.
- Ridolfi: $\frac{14}{28}$.

Decay centrality

- Decay centrality (衰弱中心性): for node i,

$$\sum_{j \neq i} \delta^{\ell(i,j)}$$

  where $\delta \in (0, 1)$ is the decay factor.
- It captures the proximity between a given node and every other node weighted by the decay.

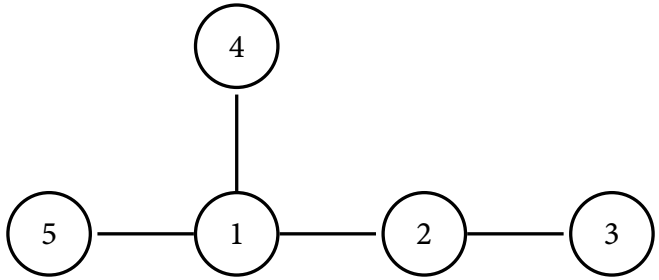
Decay centrality (Cont.)

- As δ approaches 1, the decay centrality measures how large a component a node lies in.
- As δ approaches 0, the decay centrality gives infinitely more weight to closer nodes than farther nodes.

$\Rightarrow$ It becomes degree.

- Different decay factors may give different order for nodes.

Closeness/decay centrality: Example



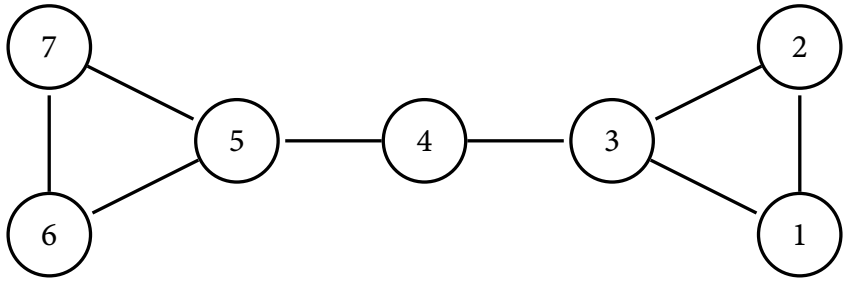
- The closeness centrality of node 1:

$$\frac{n-1}{\sum_{j\neq i}\ell(i,j)} = \frac{n-1}{1+2+1+1} = \frac{4}{5}.$$

- The decay centrality of node 1, with $\delta = 0.5$:

$$\sum_{j\neq i}\delta^{\ell}(i,j) = \delta + \delta^2 + \delta + \delta = 1.75.$$

Degree/closeness/decay centrality: Example



	Node 1	Node 3	Node 4
Degree	$\frac{2}{6}$	$\frac{3}{6}$	$\frac{2}{6}$
Closeness	$\frac{6}{15}$	$\frac{6}{11}$	$\frac{6}{10}$
Decay $\delta = 0.5$	1.5	2	2
Decay $\delta = 0.75$	3.1	3.7	3.8
Decay $\delta = 0.25$	0.59	0.84	0.75

Decay centrality: Normalization

- The normalized decay centrality:

$$\frac{\sum_{j \neq i} \delta^{\ell(i,j)}}{(n-1)\delta}.$$

- $(n-1)\delta$ is the lowest decay possible.

Betweenness centrality

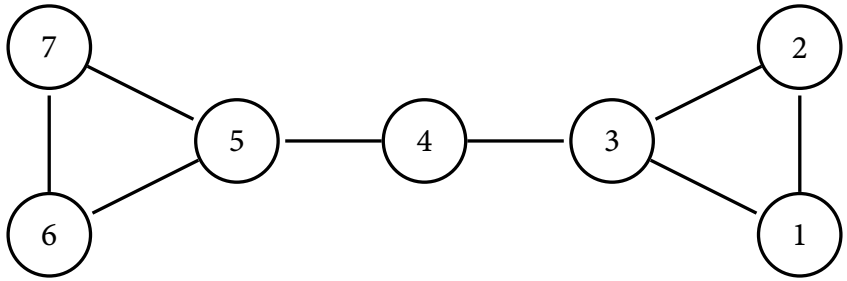
- Let $P(ij)$ denote the number of geodesics between i to j.
- Let $P_k(ij)$ denote the number of geodesics between i and j that k lies on.
- Betweenness centrality (中介中心性): for each k,

$$\sum_{i,j:\ i\neq j, i\neq k, j\neq k} \frac{P_k(ij)/P(ij)}{\binom{n-1}{2}}.$$

  If there are no paths connecting i and j, we set $P_k(ij)/P(ij) = 0$.
- It captures how well situated a node is in terms of paths that it lies on.

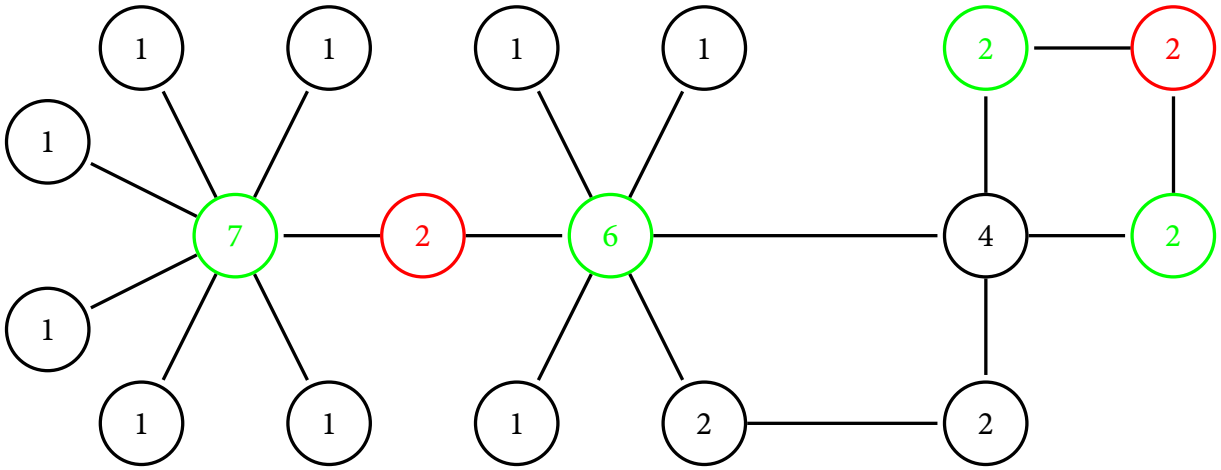
  If two nodes (they are not directly connected) want to deal with each other, then they might have to go through somebody that they are connected with.

Betweenness centrality: Example



	Node 1	Node 3	Node 4
Degree	$\frac{2}{6}$	$\frac{3}{6}$	$\frac{2}{6}$
Closeness	$\frac{6}{15}$	$\frac{6}{11}$	$\frac{6}{10}$
Decay $\delta = 0.5$	1.5	2	2
Decay $\delta = 0.75$	3.1	3.7	3.8
Decay $\delta = 0.25$	0.59	0.84	0.75
Betweenness	0	$\frac{8}{15}$	$\frac{9}{15}$

Eigenvector centrality: Motivation



- Each red node has degree 2, whereas left red node's friends have degree 6 and 7.
- In some sense the they are better connected than the right red node's friends.
- The idea of eigenvector centrality is that your importance comes from being connected to other important.

Eigenvector centrality

- Assumption: Each node's centrality is proportional to the sum of its neighbors' centralities.

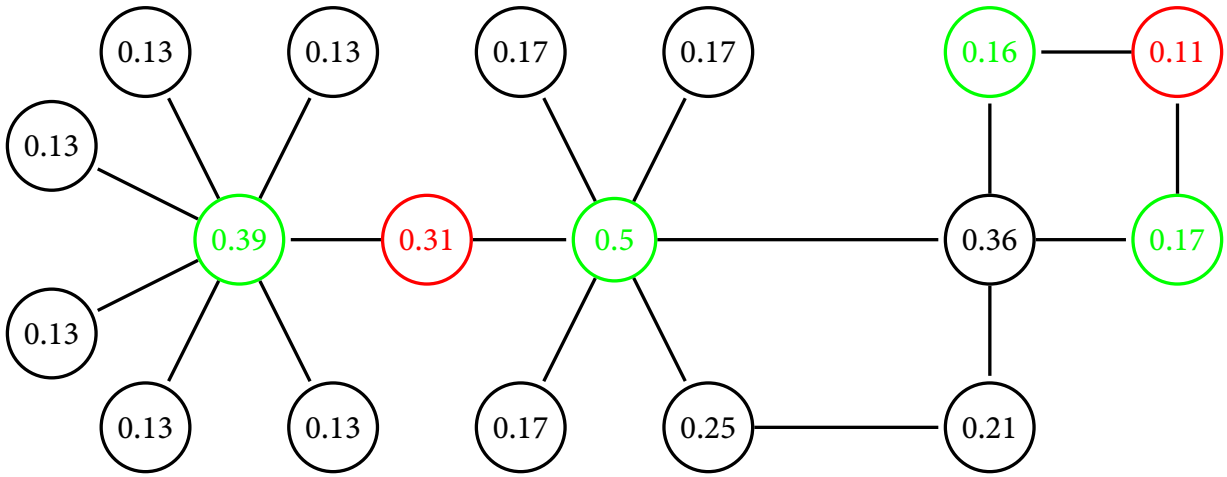
  That is, C_i is proportional to $\sum_{j \in N_i(g)} C_j$.

$\Rightarrow$ $C_i = a \sum_j g_{ij} C_j = a(gC)_i$.

$\Rightarrow$ $C = agC$, where C is the column vector $(C_1, C_2, \ldots, C_n)$.

- Actually, C is an eigenvector of g (associated with the eigenvalue a^{-1}).
- It is a self-referential concept.

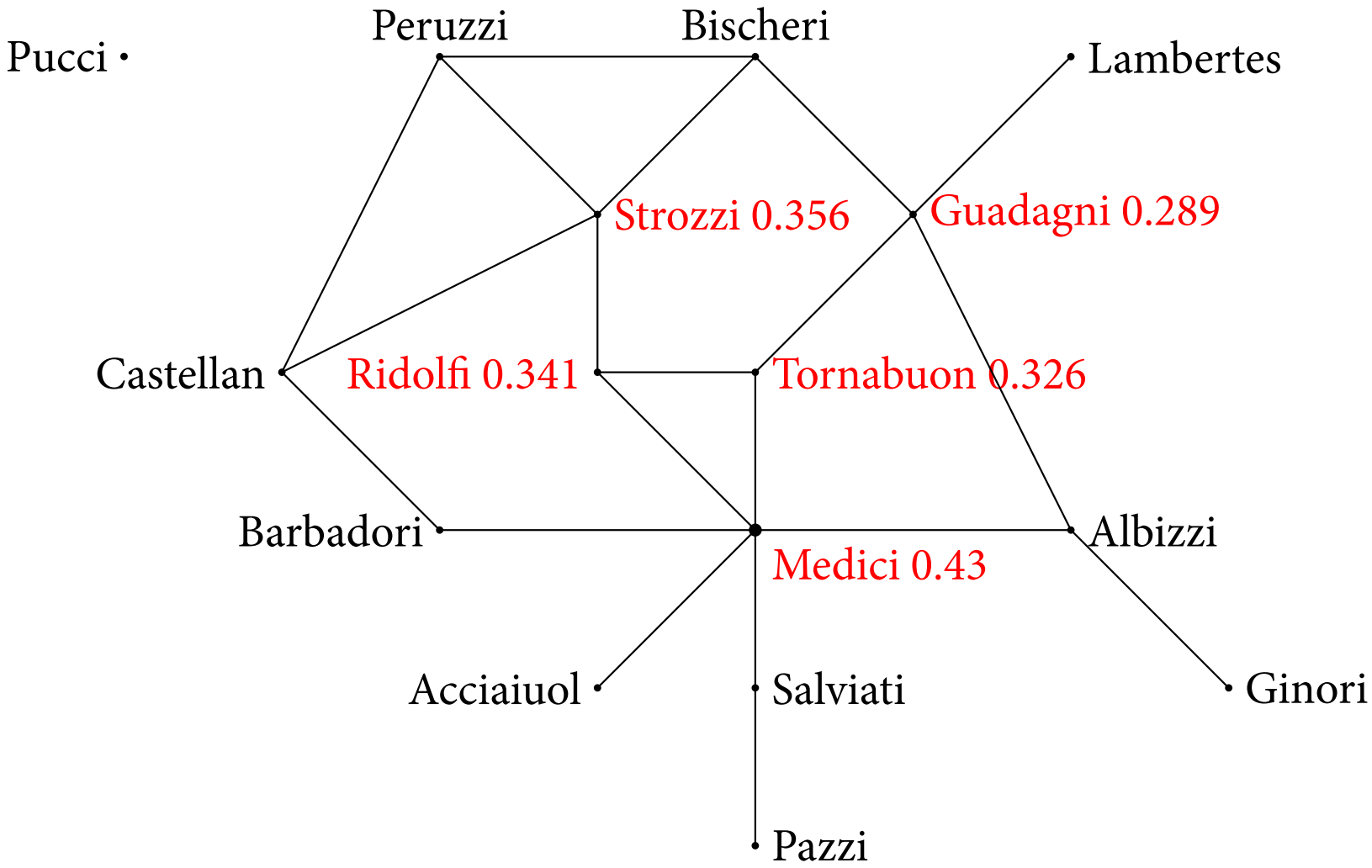
Eigenvector centrality: Example



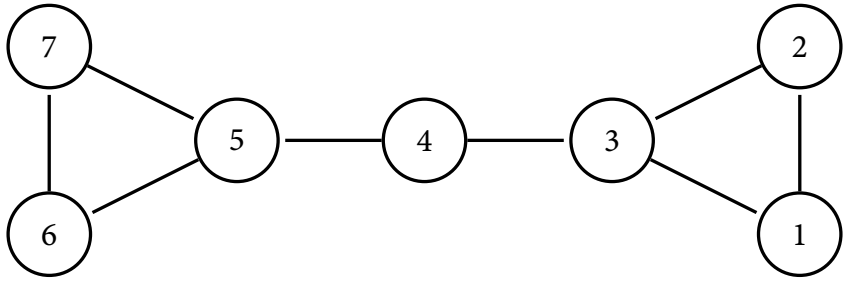
Eigenvector centrality: Computation

- $C = agC$, C is an eigenvector of g.
- There are many possible eigenvectors.
- Two steps to find the conventional one:
  - Look for one with the largest eigenvalue.
    The largest eigenvalue and an associated eigenvector are nonnegative by Perron-Frobenius Theorem (for connected graphs).
  - Normalize the eigenvector such that the sum of entries is 1.
- `EigenvectorCentrality` in Mathematica.

Eigenvector centrality: Example



Eigenvector centrality: Example



	Node 1	Node 3	Node 4
Degree	$\frac{2}{6}$	$\frac{3}{6}$	$\frac{2}{6}$
Closeness	$\frac{6}{15}$	$\frac{6}{11}$	$\frac{6}{10}$
Decay $\delta = 0.5$	1.5	2	2
Decay $\delta = 0.75$	3.1	3.7	3.8
Decay $\delta = 0.25$	0.59	0.84	0.75
Betweenness	0	$\frac{8}{15}$	$\frac{9}{15}$
Eigenvector	0.127678	0.171462	0.146365

Eigenvector centrality: Example (Cont.)

centrality.nb - Wolfram Mathematica 10.3

File Edit Insert Format Cell Graphics Evaluation Palettes Window Help

```
g = Graph[{1 ↔ 2, 2 ↔ 3, 3 ↔ 1, 3 ↔ 4, 4 ↔ 5, 5 ↔ 6, 6 ↔ 7, 7 ↔ 5}]
```

```
DegreeCentrality[g]
{2, 2, 3, 2, 3, 2, 2}

ClosenessCentrality[g]
{0.4, 0.4, 0.545455, 0.6, 0.545455, 0.4, 0.4}

BetweennessCentrality[g]
{0., 0., 8., 9., 8., 0., 0.}

EigenvectorCentrality[g]
{0.127678, 0.127678, 0.171462, 0.146365, 0.171462, 0.127678, 0.127678}

KatzCentrality[g, 1/3, {2/3, 2/3, 1, 2/3, 1, 2/3, 2/3}]
{3.125, 3.125, 4.25, 3.5, 4.25, 3.125, 3.125}

KatzCentrality[g, 1/3, {2, 2, 3, 2, 3, 2, 2}]
{9.375, 9.375, 12.75, 10.5, 12.75, 9.375, 9.375}

KatzCentrality[g, 1/4, {2, 2, 3, 2, 3, 2, 2}]
{4.94118, 4.94118, 6.82353, 5.41176, 6.82353, 4.94118, 4.94118}
```

100%

Eigenvector centrality: Applications

- Google Page rank: score of a page is proportional to the sum of the scores of pages linked to it.
- Random surfer model: start at some page on the web, randomly pick a link, follow it, repeat …

Katz prestige

- Katz prestige (Katz 声望) of node i is defined to be a sum of the prestige of i's neighbors divided by their respective degrees.

$$P_i^K(g) = \sum_{j \neq i} g_{ij} \frac{P_j^K(g)}{d_j(g)}.$$

- Katz prestige vs eigenvector centrality:
  - Katz prestige is corrected by how many neighbors j has.

Katz prestige: Computation

- Katz prestige of node i:

$$P_i^K(g) = \sum_{j \neq i} g_{ij} \frac{P_j^K(g)}{d_j(g)}.$$

- Let $\hat{g}_{ij} = \frac{g_{ij}}{d_j(g)}$ for each i and j.

$\Rightarrow$ Then $\hat{g}$ is the normalized adjacency matrix g so that each column sum is 1.

- Let $P^K(g)$ be the column vector $(P_1^K(g), P_2^K(g), \ldots, P_n^K(g))$, then

$$P^K(g) = \hat{g} P^K(g).$$

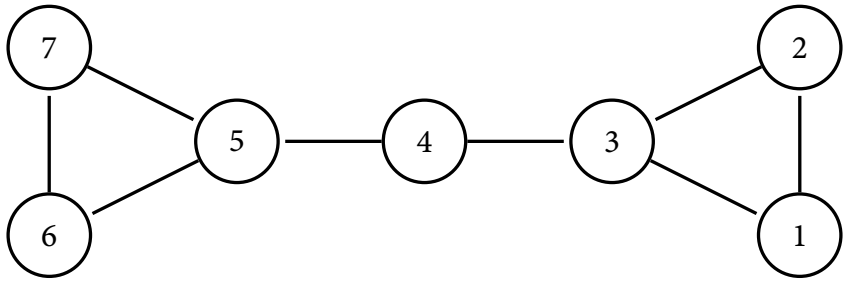
Katz prestige: Computation

- $P^K(g) = \hat{g}P^K(g)$.

$\Rightarrow$ $P^K(g)$ is an eigenvector of $\hat{g}$ associated with eigenvalue 1.

- Since $\hat{g}$ is a column stochastic matrix, Perron–Frobenius theorem implies the existence of a nonnegative eigenvector associated with eigenvalue 1.

Katz prestige: Example



$$P^K(g) = (2, 2, 3, 2, 3, 2, 2).$$

Katz prestige: Example (Cont.)

Katz.nb - Wolfram Mathematica 10.3

File Edit Insert Format Cell Graphics Evaluation Palettes Window Help

```
g = Transpose[{{0, 1/2, 1/2, 0, 0, 0, 0}, {1/2, 0, 1/2, 0, 0, 0, 0},
   {1/3, 1/3, 0, 1/3, 0, 0, 0}, {0, 0, 1/2, 0, 1/2, 0, 0},
   {0, 0, 0, 1/3, 0, 1/3, 1/3}, {0, 0, 0, 0, 1/2, 0, 1/2},
   {0, 0, 0, 0, 1/2, 1/2, 0}}]
```

$$\left\{\left\{0, \frac{1}{2}, \frac{1}{3}, 0, 0, 0, 0\right\}, \left\{\frac{1}{2}, 0, \frac{1}{3}, 0, 0, 0, 0\right\},\right.$$
$$\left\{\frac{1}{2}, \frac{1}{2}, 0, \frac{1}{2}, 0, 0, 0\right\}, \left\{0, 0, \frac{1}{3}, 0, \frac{1}{3}, 0, 0\right\},$$
$$\left.\left\{0, 0, 0, \frac{1}{2}, 0, \frac{1}{2}, \frac{1}{2}\right\}, \left\{0, 0, 0, 0, \frac{1}{3}, 0, \frac{1}{2}\right\}, \left\{0, 0, 0, 0, \frac{1}{3}, \frac{1}{2}, 0\right\}\right\}$$

```
Eigenvectors[g]
```

$$\left\{\left\{1, 1, \frac{3}{2}, 1, \frac{3}{2}, 1, 1\right\}, \left\{-1, -1, \frac{1}{4}\left(3-\sqrt{57}\right), 0, \frac{1}{4}\left(-3+\sqrt{57}\right), 1, 1\right\},\right.$$
$$\left\{1, 1, \frac{1}{4}\left(-9-\sqrt{33}\right), \frac{1}{2}\left(1+\sqrt{33}\right), \frac{1}{4}\left(-9-\sqrt{33}\right), 1, 1\right\}, \{0, 0, 0, 0, 0, -1, 1\},$$
$$\{-1, 1, 0, 0, 0, 0, 0\}, \left\{-1, -1, \frac{1}{4}\left(3+\sqrt{57}\right), 0, \frac{1}{4}\left(-3-\sqrt{57}\right), 1, 1\right\},$$
$$\left.\left\{1, 1, \frac{1}{4}\left(-9+\sqrt{33}\right), \frac{1}{2}\left(1-\sqrt{33}\right), \frac{1}{4}\left(-9+\sqrt{33}\right), 1, 1\right\}\right\}$$

100%

Katz prestige: Property

- Katz prestige:

$$P^K(g) = \hat{g}P^K(g).$$

- Katz prestige is only determined up to a scale factor: if $P^K(g)$ solves the above equation, then so does cP^K for any c.
- In undirected graph, it is easy to check that the solution to the above equation is the degree vector (or any rescaling of it).

$\Rightarrow$ $[P^K(g)]_i = d_i(g)$.

- This provides a justification for degree centrality.

Bonacich centrality

- The power/prestige of a node is a weighted (weighting by distance) sum of the walks that emanate from it.
- A walk of length 1 is worth a, a walk of length 2 is worth a^2, and so forth, for some parameter $a \in (0, 1)$.
- This scheme gives higher weights to walks of shorter distance.

Katz prestige-2

- Katz prestige-2:

$$P^{K2}(g,a) = ag\mathbf{1} + agag\mathbf{1} + a^2g^2ag\mathbf{1} + \cdots .$$

- Each node i has a base value $ad_i(g) = (ag\mathbf{1})_i$ for some $a > 0$.
- Add a times the base value of each node that it has a direct link to i: $a \times \sum_j g_{ij}(ag\mathbf{1})_j = (a^2g^2\mathbf{1})_i$.
- Add a^2 times the base value of each node that it has a walk of length 2 to i.
- And so forth.

Bonacich centrality

- Katz prestige-2:

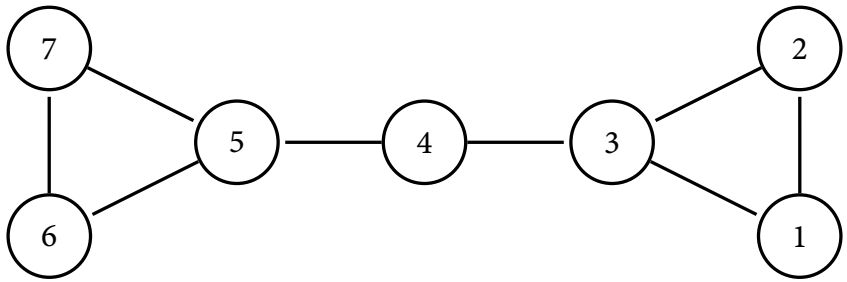
$$(I - ag)^{-1} ag\mathbf{1}.$$

- Bonacich centrality is a direct extension of Katz prestige-2:

$$(I - bg)^{-1} ag\mathbf{1},$$

where $a > 0$, $b > 0$, and b is sufficiently small such that the expression is well defined.

Bonacich centrality: Example



Bonacich centrality: Example (Cont.)

	Node 1	Node 3	Node 4
Degree	$\frac{2}{6}$	$\frac{3}{6}$	$\frac{2}{6}$
Closeness	$\frac{6}{15}$	$\frac{6}{11}$	$\frac{6}{10}$
Decay $\delta = 0.5$	1.5	2	2
Decay $\delta = 0.75$	3.1	3.7	3.8
Decay $\delta = 0.25$	0.59	0.84	0.75
Betweenness	0	$\frac{8}{15}$	$\frac{9}{15}$
Eigenvector	0.127678	0.171462	0.146365
Katz prestige-2 $a = 1/3$	3.125	4.25	3.5
Bonacich $a = 1, b = 1/3$	9.375	12.75	10.5
Bonacich $a = 1, b = 1/4$	4.94118	6.82353	5.41176

Bonacich centrality: Example (Cont.)

centrality.nb - Wolfram Mathematica 10.3

File Edit Insert Format Cell Graphics Evaluation Palettes Window Help

```
g = Graph[{1 ↔ 2, 2 ↔ 3, 3 ↔ 1, 3 ↔ 4, 4 ↔ 5, 5 ↔ 6, 6 ↔ 7, 7 ↔ 5}]
```

```
DegreeCentrality[g]
{2, 2, 3, 2, 3, 2, 2}

ClosenessCentrality[g]
{0.4, 0.4, 0.545455, 0.6, 0.545455, 0.4, 0.4}

BetweennessCentrality[g]
{0., 0., 8., 9., 8., 0., 0.}

EigenvectorCentrality[g]
{0.127678, 0.127678, 0.171462, 0.146365, 0.171462, 0.127678, 0.127678}

KatzCentrality[g, 1/3, {2/3, 2/3, 1, 2/3, 1, 2/3, 2/3}]
{3.125, 3.125, 4.25, 3.5, 4.25, 3.125, 3.125}

KatzCentrality[g, 1/3, {2, 2, 3, 2, 3, 2, 2}]
{9.375, 9.375, 12.75, 10.5, 12.75, 9.375, 9.375}

KatzCentrality[g, 1/4, {2, 2, 3, 2, 3, 2, 2}]
{4.94118, 4.94118, 6.82353, 5.41176, 6.82353, 4.94118, 4.94118}
```

100%

Subsection 5

Homophily

Homophily

- Many networks exhibit "homophily (趋同性)" by Lazarsfeld and Merton.
- This property refers to the fact that people are more prone to maintain relationships with people who are similar to themselves.

  race, gender, religion, …
- "Birds of a feather flock together" — Philemon Holland (1960).

  物以类聚，人以群分

Homophily: Real world

Friendship in high school in US:

Percent	52	38	5	5
	White	Black	Hispanic	Other
White	86	7	47	74
Black	4	85	46	13
Hispanic	4	6	2	4
Other	6	2	5	9

Homophily: Real world (Cont.)

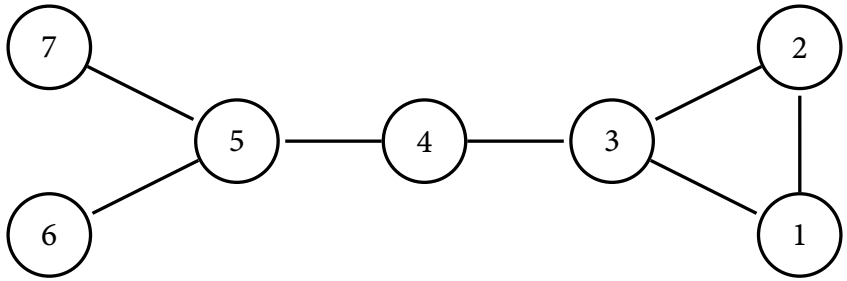
Friendship in high school in Dutch:

Percent	65	5	6	7	17
	Dutch	Moroccan	Turkish	Surinamese	Other
Dutch	79	24	11	21	47
Moroccan	2	27	8	4	5
Turkish	2	19	59	8	6
Surinamese	3	8	8	44	12
Other	13	22	14	23	30

Homophily: Reason

- Opportunity—contract theory
- Benefits/costs
- Social pressure
- Social competition

Homework: Question 1



Use software (Matlab, Mathematica, R, …) to compute the degree distribution, diameter, average path length, overall/average clustering coefficient, degree centrality, closeness centrality, decay centrality ($\delta = 0.5$), betweenness centrality, eigenvector centrality, Bonacich centrality ($a = 1, b = 0.5$), Katz prestige, and Katz prestige-2 ($a = 0.5$).

Homework: Question 2 and 3

Practical questions

Appendix: Poisson distribution

- $$\binom{n-1}{d} p^d (1-p)^{n-1-d}.$$
- $$\frac{1}{d!}(n-1)^d p^d e^{-(n-1)p} = \frac{\lambda^d e^{-d}}{d!},$$

  where $d = (n-1)p$.