UNIVERSITY OF MANNHEIM

Web Mining

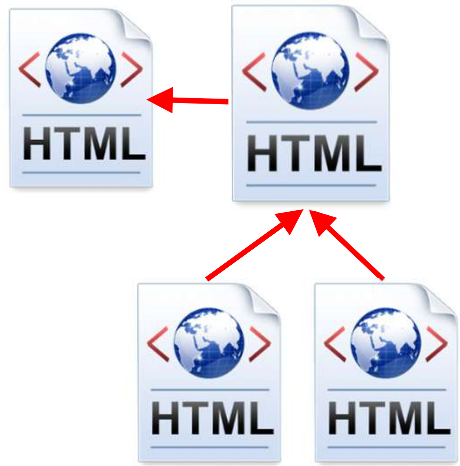
Web Structure Mining and Social Network Analysis

Prof. Dr. Christian Bizer

FSS 2025

Web Structure Mining

- **Definition**



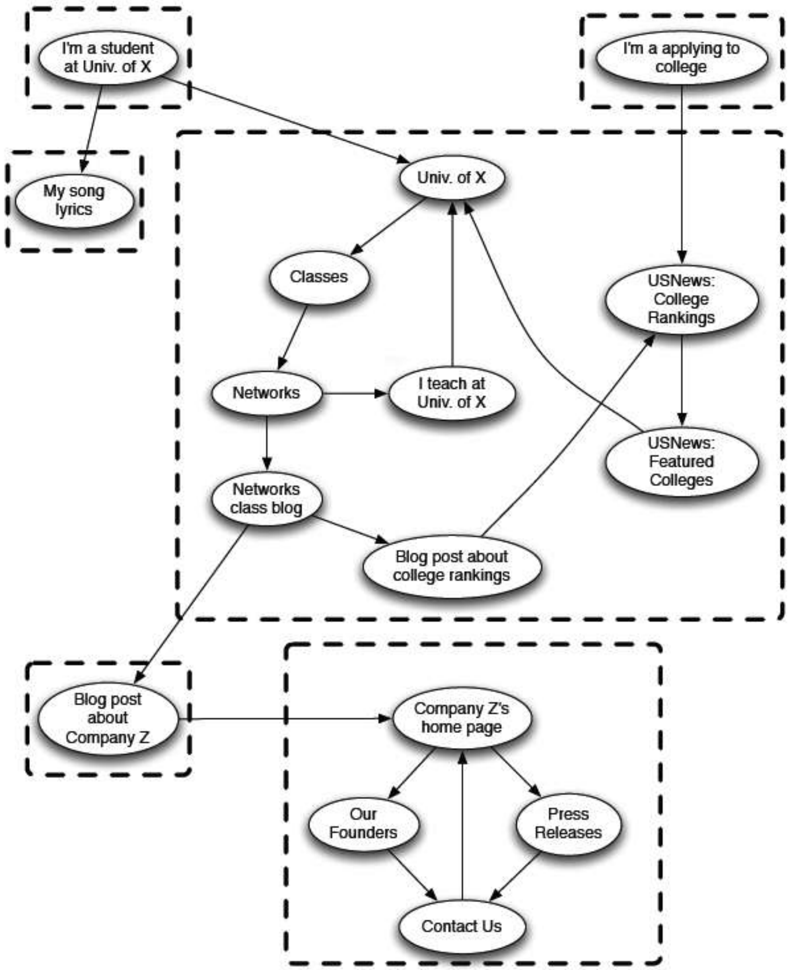
> **Discovery and interpretation of patterns in**
> 1. **the hyperlink structure of the Web**
> 2. **the social ties among actors that interact on the Web**

- **Typical sources of web graphs**
  1. web crawls including HTML pages and hyperlinks
  2. social networks representing relations between actors
  3. knowledge graphs that have been extracted from the Web
  4. other types of community data (discussion forums, email conversations, navigation paths …)

- **Web structure mining focuses on the structure, but is also often combined with content or usage mining techniques**

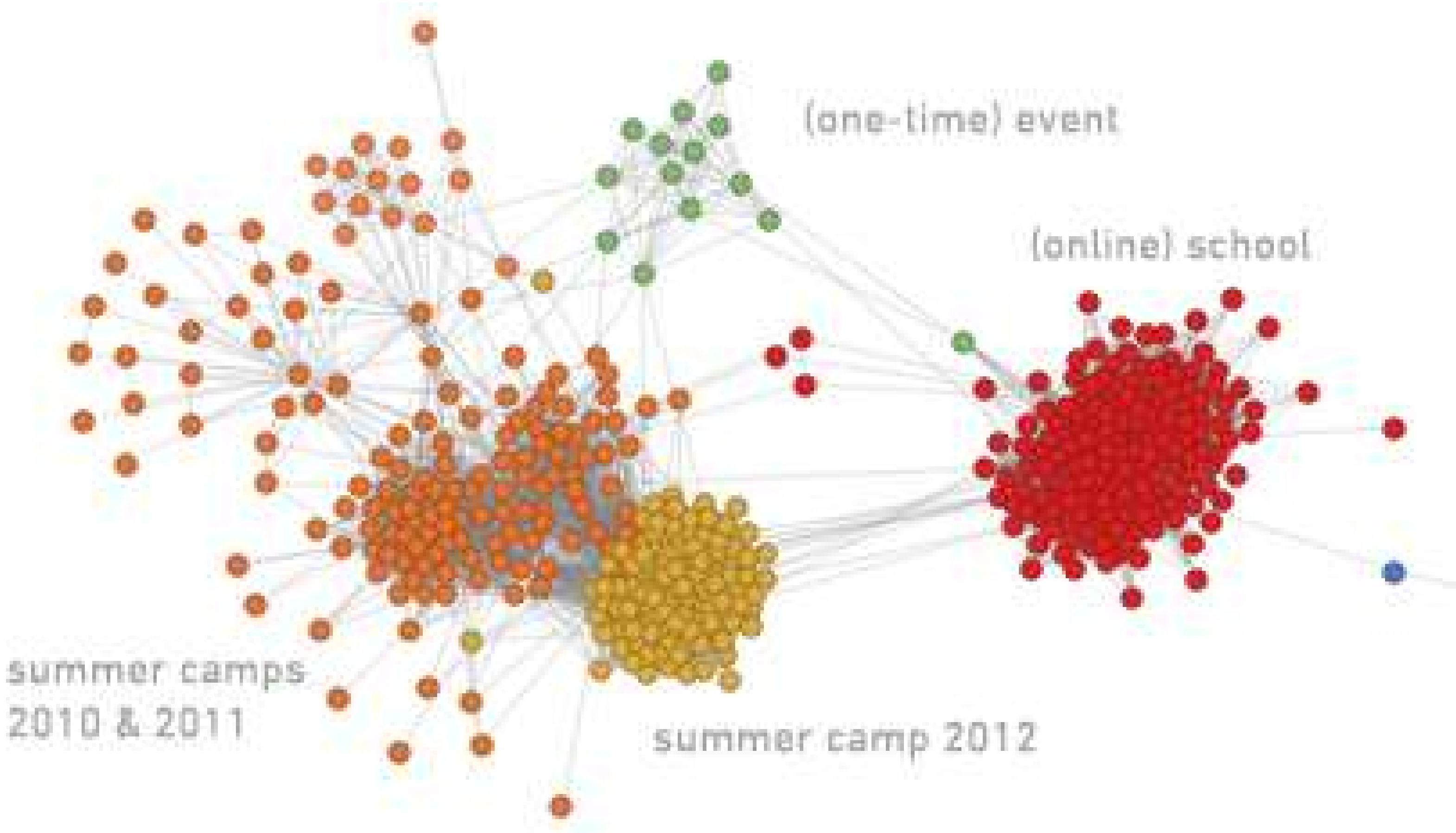
Hyperlink Graph

A hyperlink graph is a collection of hyperlinks between web pages which belong to web sites.



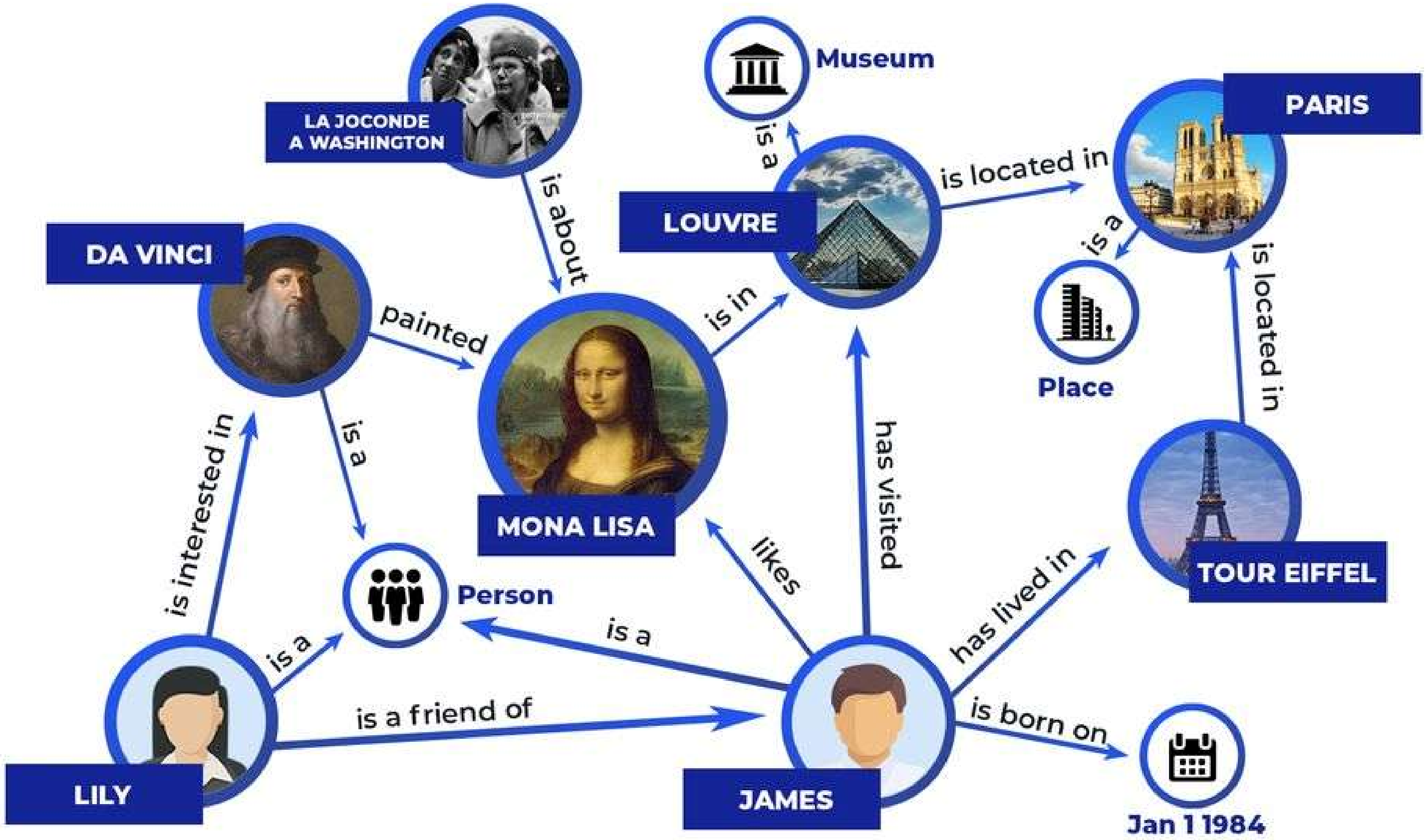
Social Network

A social network is a set of relations (e.g. friendship, interest, data exchange) between social entities, i.e. members of a social system (actors).



Knowledge Graph

A knowledge graph is a set of relations having different types (e.g. located in, painted, is interested in, is a) between entities (Mona Lisa, Louvre, Da Vinci) belonging to classes (e.g. persons, paintings, museums, places, dates).

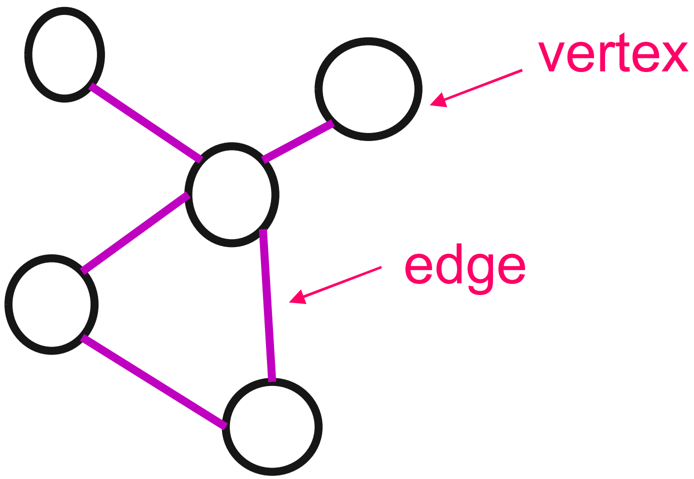


Chapter Outline

1. Describing Graphs
    1. Terminology and Metrics
2. Prominence
    1. Centrality
    2. Prestige
3. Community Detection
    1. Connected Components and K-Cores
    2. Clustering-based Techniques
4. Machine Learning with Graphs
    1. Link Prediction and Node Classification
    2. Node Embeddings
    3. Graph Neural Networks

1. Describing Graphs: Terminology and Metrics

A Graph is a collection of vertices that are connected by edges.



Network often refers to real systems

Graph: mathematical representation of a network

But often: “Network” ≡ “Graph”

Community	Points	Lines
Math	vertices	edges, arcs
Computer Science	nodes	links
Physics	sites	bonds
Sociology	actors	ties, relations

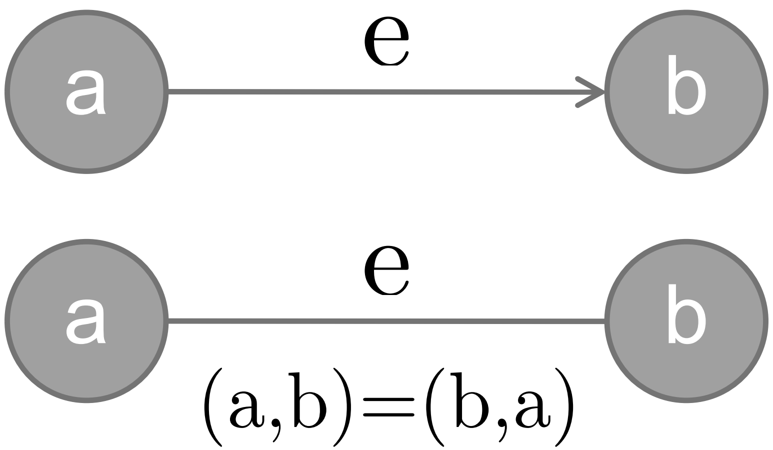
Graphs

A graph is an ordered pair $G=(V,E)$ where $V \neq \emptyset$ is a set of vertices and $E \subseteq V \times V$ is a set of edges.

Two vertices a and b are called adjacent if $(a,b) \in E$

directed edge/arc:

undirected edge:

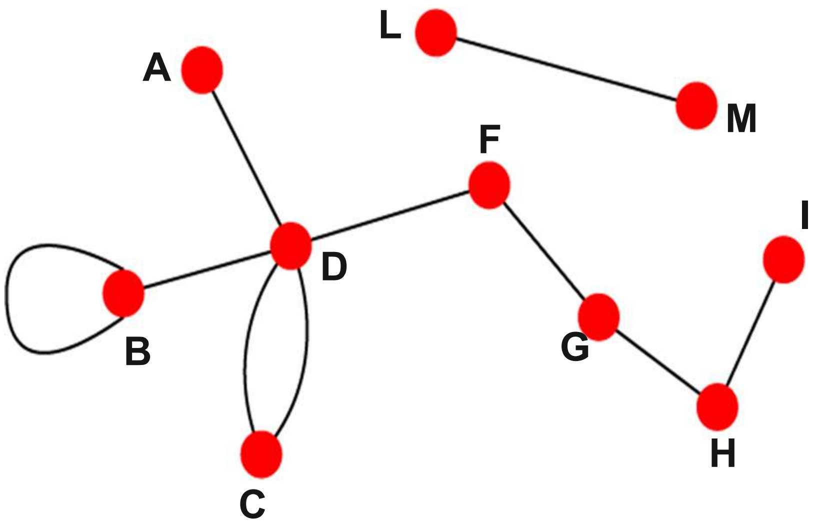


Examples: Directed and Undirected Graphs

Undirected Graph

undirected edges (*symmetrical*) ➔ edge

Graph:

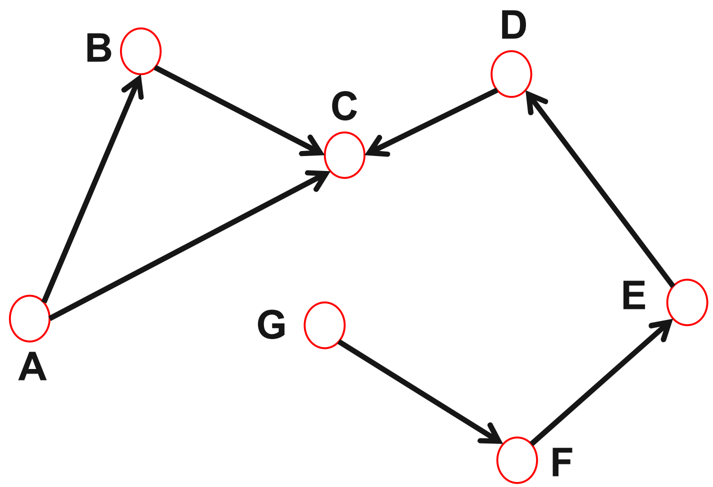


Undirected edges:
- co-authorship links
- roads (mostly)

Directed Graph

directed edges ➔ *arcs*

Digraph = directed graph:

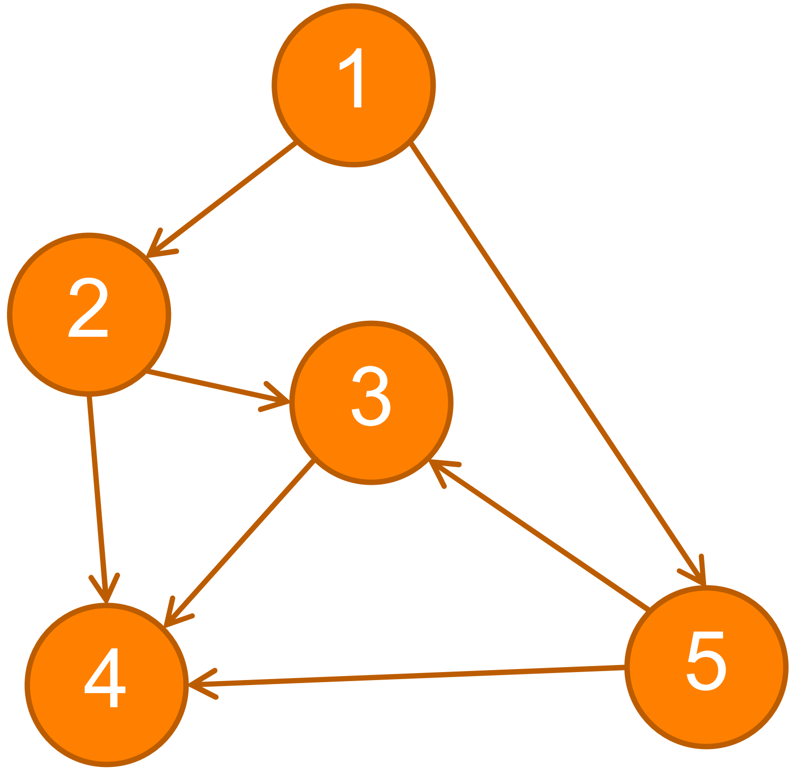


Directed arcs:
- hyperlinks on the WWW
- following on Twitter
- phone calls

Adjacency Matrix

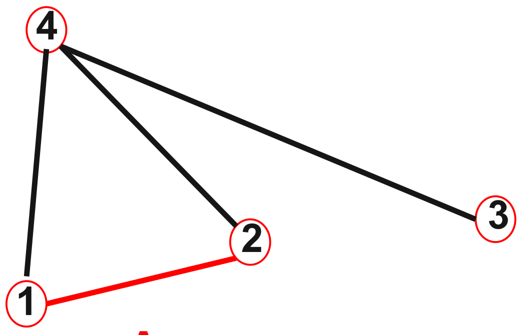
A graph can be represented as adjacency matrix.

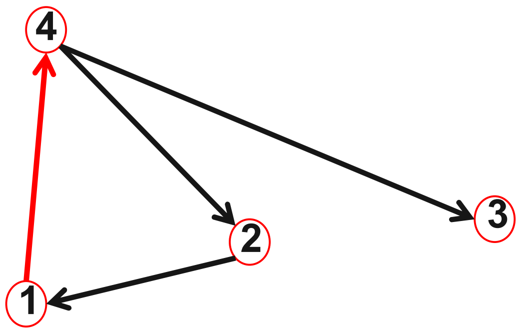
$$A = \begin{array}{cc} \begin{matrix} & & j & & \end{matrix} & \\ \begin{pmatrix} 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 \end{pmatrix} & i \end{array}$$



$$a_{ij} = 1 \Leftrightarrow (i, j) \in E$$

Adjacency Matrices for Directed and Undirected Graphs





A_{12}

$$A_{ij} = \begin{pmatrix} 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 1 & 1 & 1 & 0 \end{pmatrix}$$

A_{14}

$$A_{ij} = \begin{pmatrix} 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 \end{pmatrix}$$

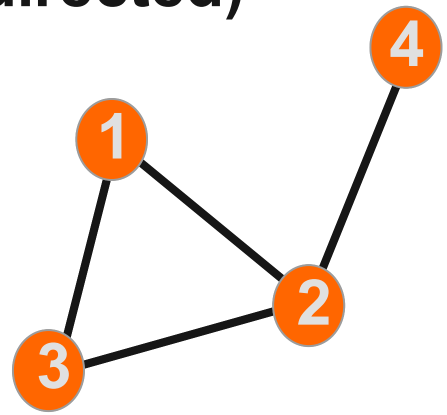
$\mathbf{A_{ij}=1}$ if there is a link between vertices *i* and *j*

$\mathbf{A_{ij}=0}$ if vertices *i* and *j* are not connected to each other.

Note that for an undirected graph (left) the matrix is symmetric.

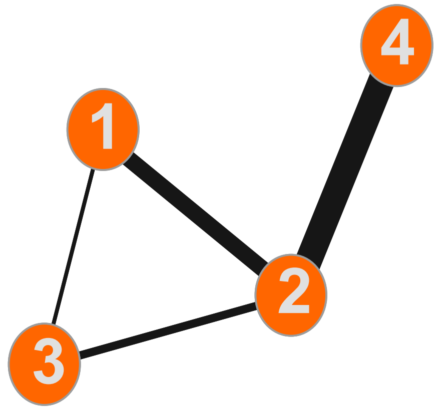
Weighted and Unweighted Graphs

Unweighted Graph (undirected)



$$A_{ij} = \begin{pmatrix} 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix}$$

Weighted Graph (undirected)

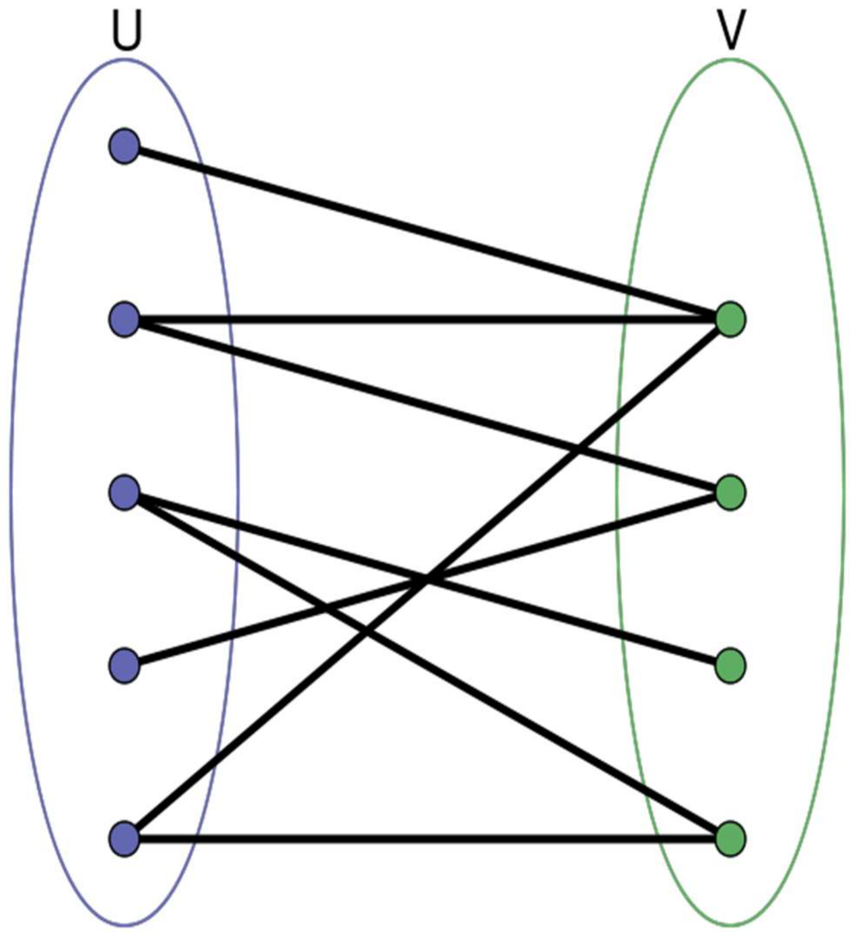


$$A_{ij} = \begin{pmatrix} 0 & 2 & 0.5 & 0 \\ 2 & 0 & 1 & 4 \\ 0.5 & 1 & 0 & 0 \\ 0 & 4 & 0 & 0 \end{pmatrix}$$

Example: Road networks (distance in miles)

Bipartite Graphs

Bipartite graph (or **bigraph**) is a graph whose vertices can be divided into two disjoint sets *U* and *V* such that every line connects a vertex in *U* to one in *V*; that is, *U* and *V* are independent sets.



Examples:

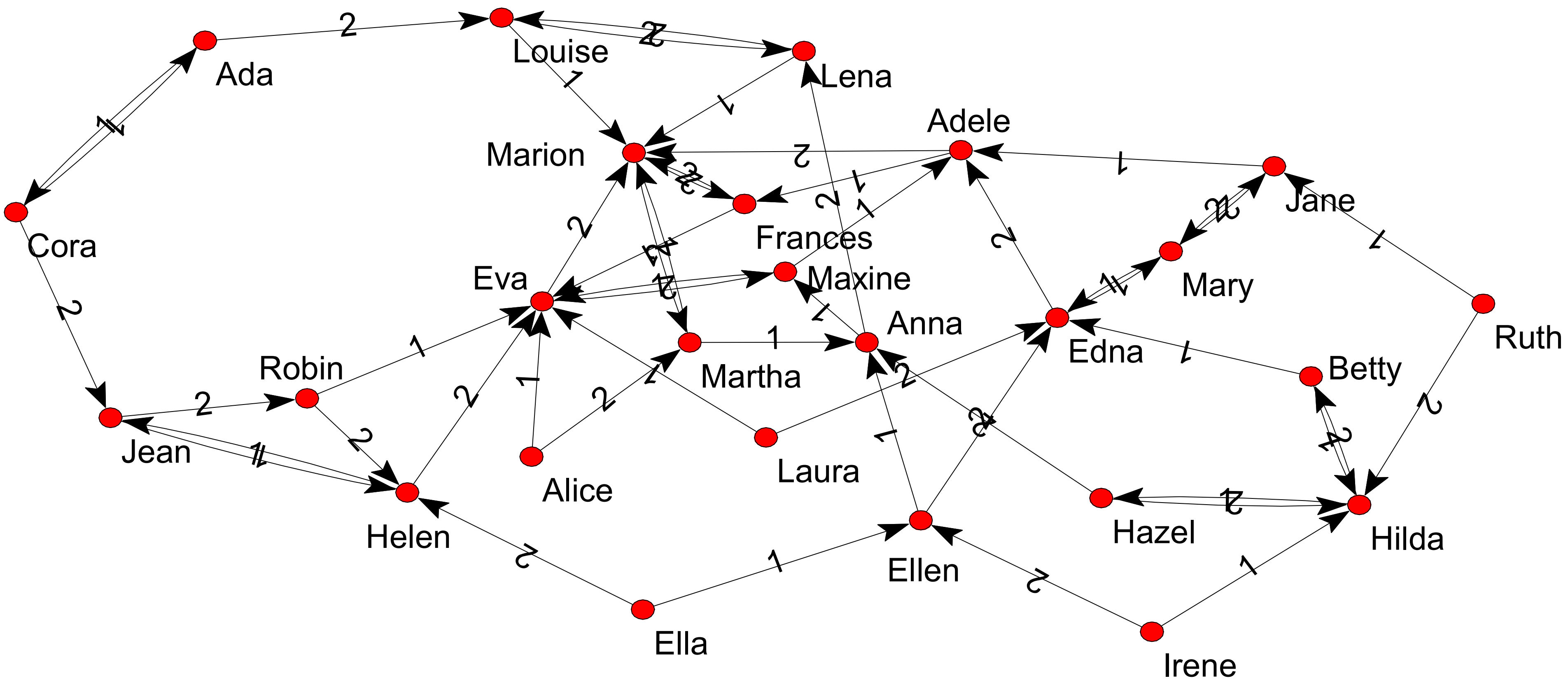
- movie/actor network
- disease/symptom network
- photo/tag network on Flickr
- customer/product recommendations

Vertex, Arc and Edge Attributes

Vertices, arcs and edges can have attributes.

Example of a network with vertex and arc attributes:

- girls' school dormitory dining-table partners (Moreno, *The sociometry reader*, 1960)
- first and second choices shown



Degree

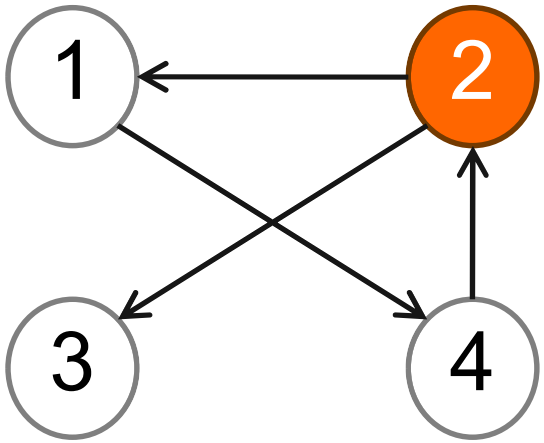
Degree: Number $C_D(v)$ of edges adjacent to v

In-degree: $$C_D^{in}(v) = \sum_{j=1, i \neq j}^{|V|} a_{ji}$$

Out-degree: $$C_D^{out}(v) = \sum_{j=1, i \neq j}^{|V|} a_{ij}$$

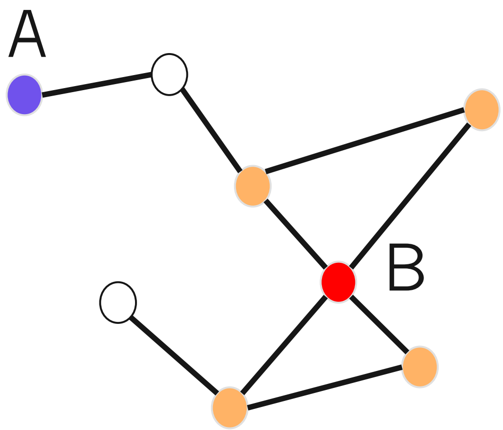
$$A = \begin{array}{cc} & j \\ \begin{pmatrix} 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix} & i \end{array}$$

$$C_D(v_2) = 3$$
$$C_D^{in}(v_2) = 1$$
$$C_D^{out}(v_2) = 2$$



Example: Degrees of Undirected and Directed Graphs

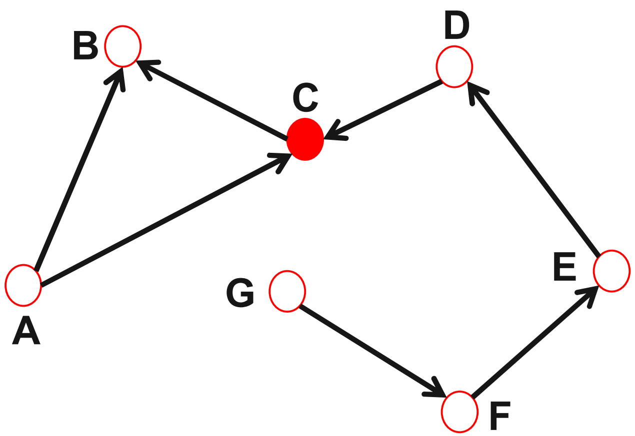
Undirected



Degree: the number of edges connected to the vertex.

$$k_A = 1 \qquad k_B = 4$$

Directed



In *directed graphs* we can define an in-degree and out-degree. The (total) degree is the sum of in- and out-degree.

$$k_C^{in} = 2 \quad k_C^{out} = 1 \quad k_C = 3$$

Source: a vertex with $k^{in} = 0$ and $k^{out} > 0$

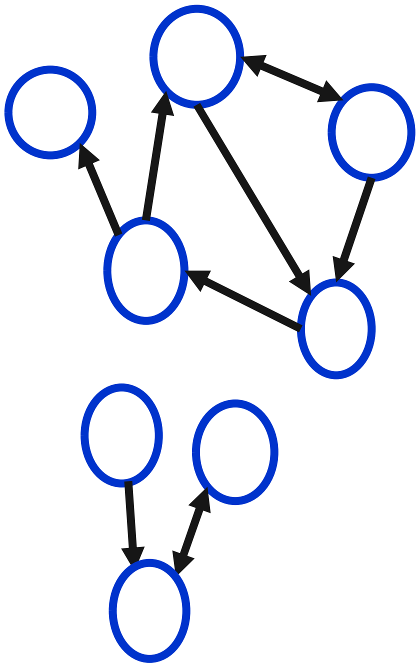
Sink: a vertex with $k^{out} = 0$ and $k^{in} > 0$

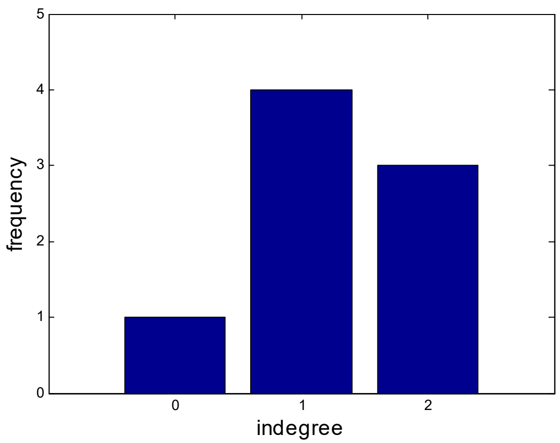
Degree Distribution

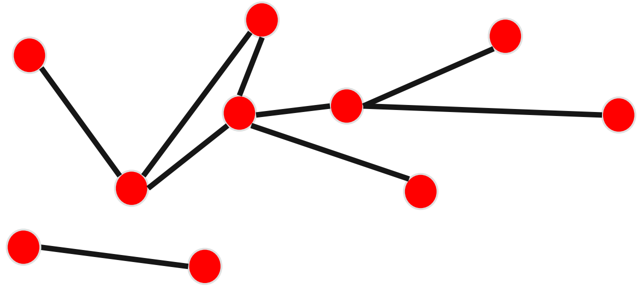
Summarizes the degrees of all vertices.

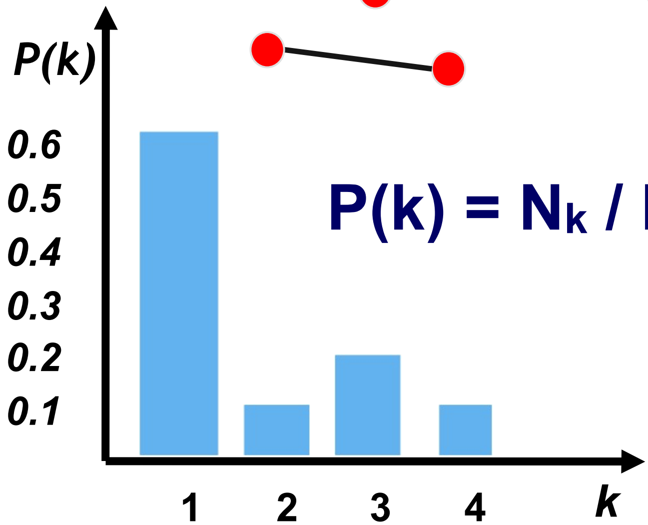
Alternative representations:

1. A frequency count of the vertices of each degree
2. P(k): probability that a randomly chosen vertex has degree k



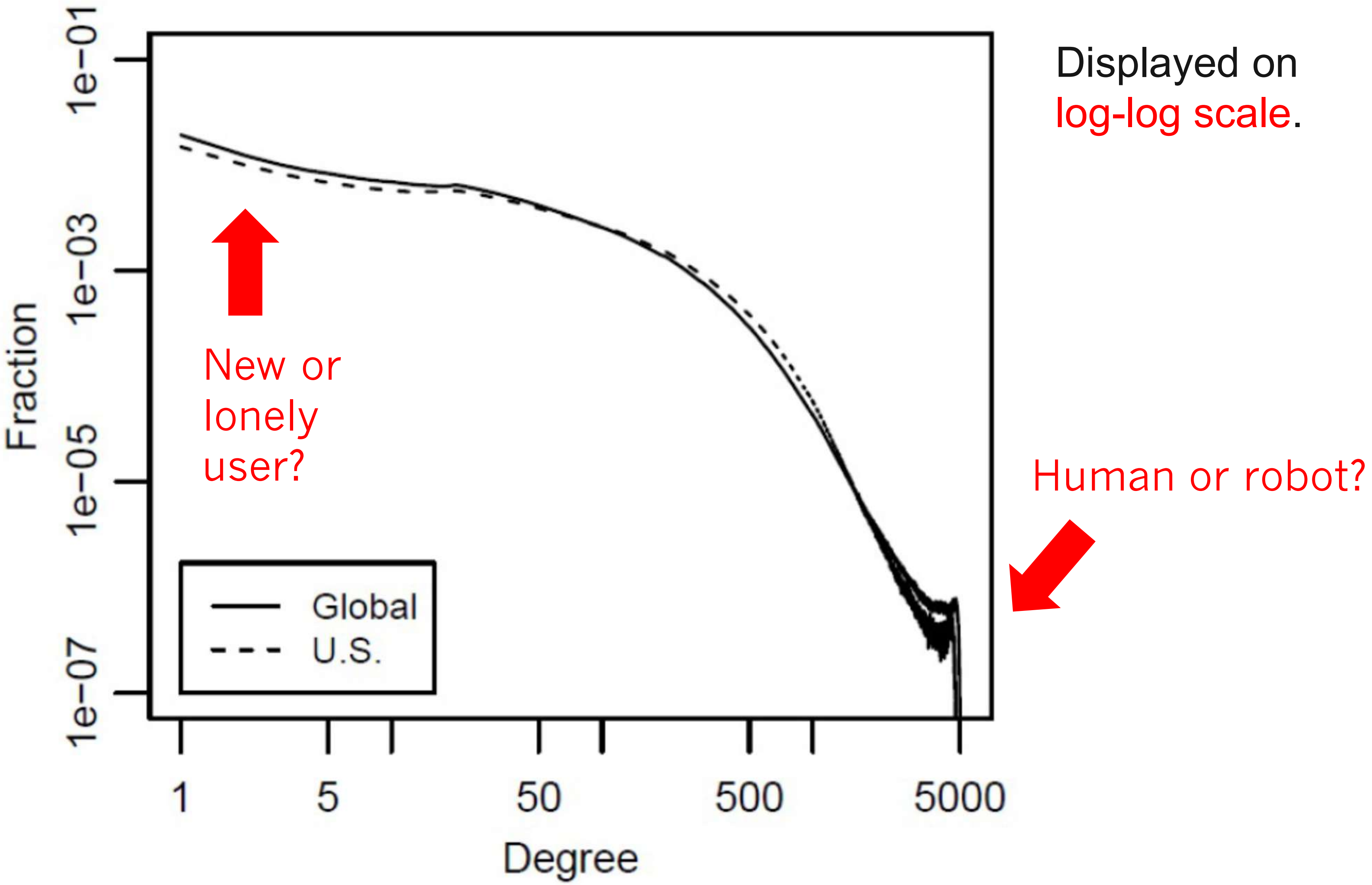






$$P(k) = N_k / N$$

Degree Distribution: Friendship on Facebook



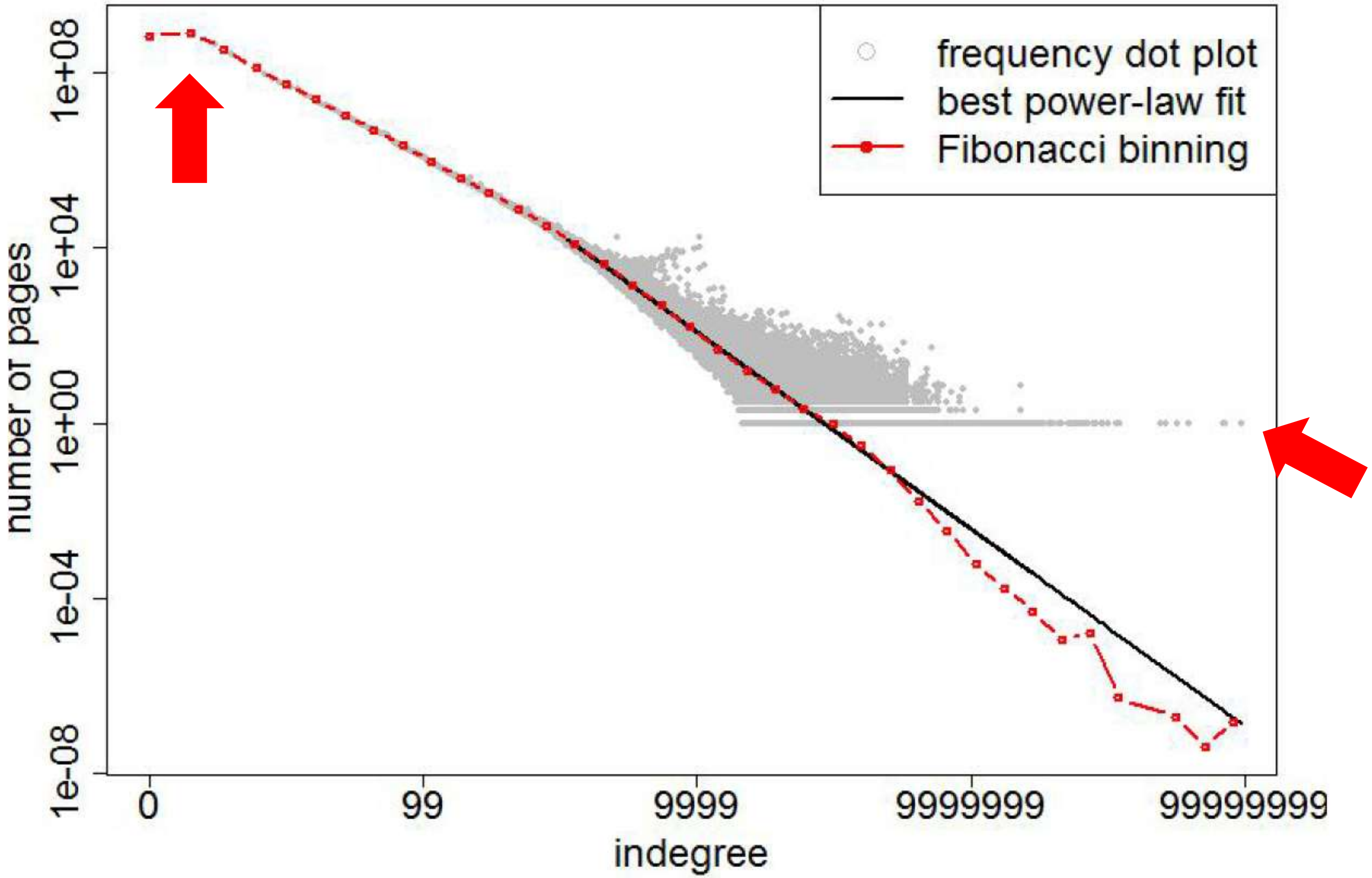
Displayed on log-log scale.

Source: Zafarani, et al: Social Media Mining. Cambridge University Press, 2014.

In-Degree Distribution of the WDC Hyperlink Graph

Covers 3.5 billion web pages and 128 billion hyperlinks, extracted from Common Crawl 2012

Displayed on log-log scale, meaning that left third covers over 99% of the mass.



Meusel, Vigna, Lehmberg, Bizer: Graph Structure in the Web - Revisited. 23rd Conference on World Wide Web (WWW2014).
Website: http://webdatacommons.org/hyperlinkgraph/

Top In-Degree Websites

The Common Crawl WWW Ranking

Here you can browse a ranking of more than 100 million sites of the World Wide Web. Every single step leading to this ranking is open and accessible. Enjoy!

Learn more »

Jump to... (prefix) Search

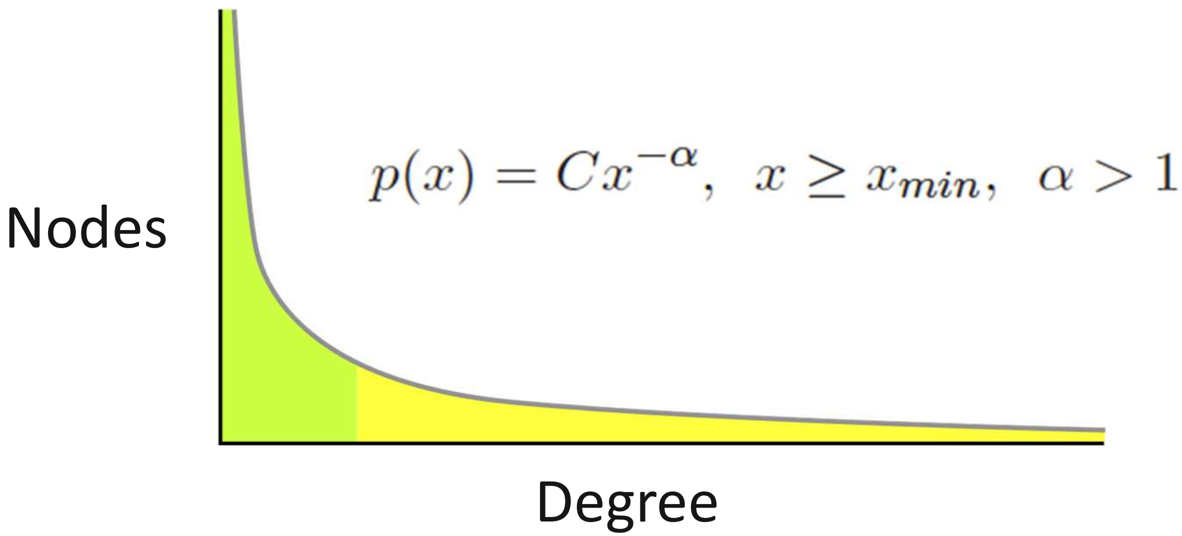
Compare ranks

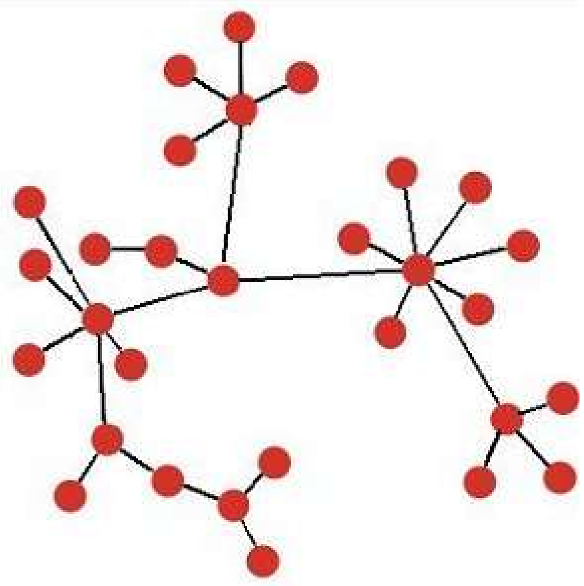
Harmonic centrality	Indegree centrality	Katz's index	PageRank
5	1. wordpress.org	1	2
1	2. youtube.com	2	3
23	3. gmpg.org	3	1
2	4. en.wikipedia.org	4	6
39	5. tumblr.com	5	7
3	6. twitter.com	6	5
4	7. google.com	7	9
6	8. flickr.com	8	14
172870	9. rtalabel.org	9	59
75	10. wordpress.com	10	30
2431063	11. mp3shake.com	11	44

http://wwwranking.webdatacommons.org/

Power-Law Distributions

- Degree distribution in large-scale networks often follow (approximately) a power law.





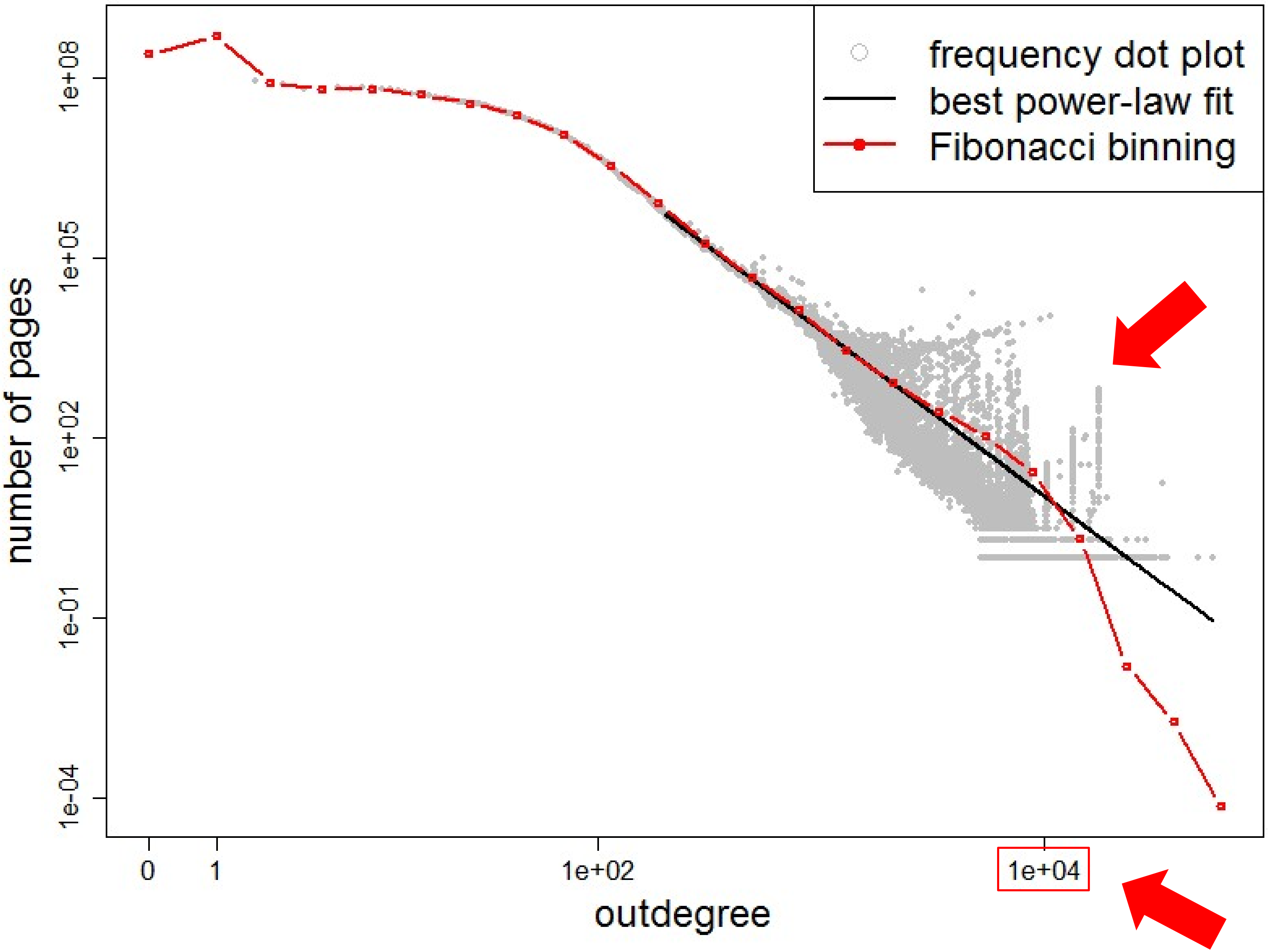
- The preferential attachment process (Barabási and Albert, 1999) explains power-law distributions: Vertices prefer to link to vertices having a high degree.
- Translates to “The rich get richer” or “The famous get more famous”.

Out-Degree Distribution of WDC Hyperlink Graph

Displayed on log-log scale.

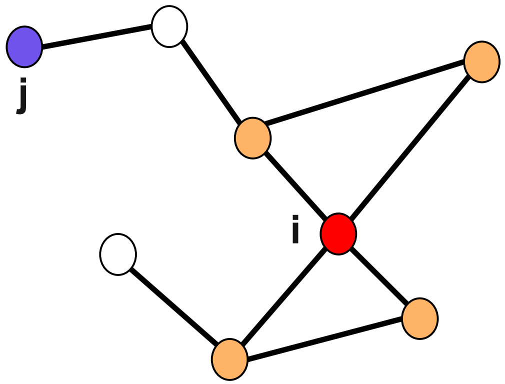
Maximal out-degrees are much smaller than maximal in-degrees.

Strange shapes are SPAM networks.



Average Degree

Undirected

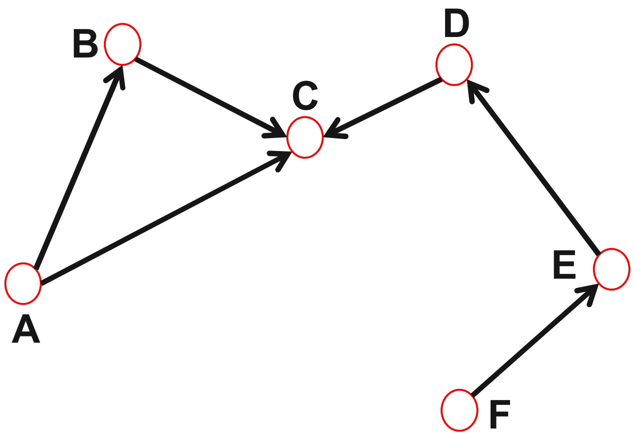


$$\langle k \rangle = \frac{1}{N}\sum_{i=1}^{N} k_i \qquad \langle k \rangle = \frac{2L}{N}$$

N – the number of vertices in the graph

L – the number of lines in the graph

Directed

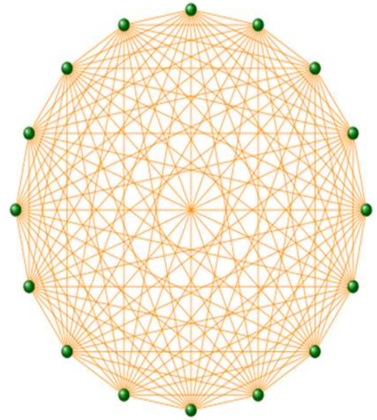


$$\langle k^{in} \rangle = \frac{1}{N}\sum_{i=1}^{N} k_i^{in}, \quad \langle k^{out} \rangle = \frac{1}{N}\sum_{i=1}^{N} k_i^{out}, \quad \langle k^{in} \rangle = \langle k^{out} \rangle$$

Warning: Average degree might be misleading because of power-law like degree distributions.

Graph Density

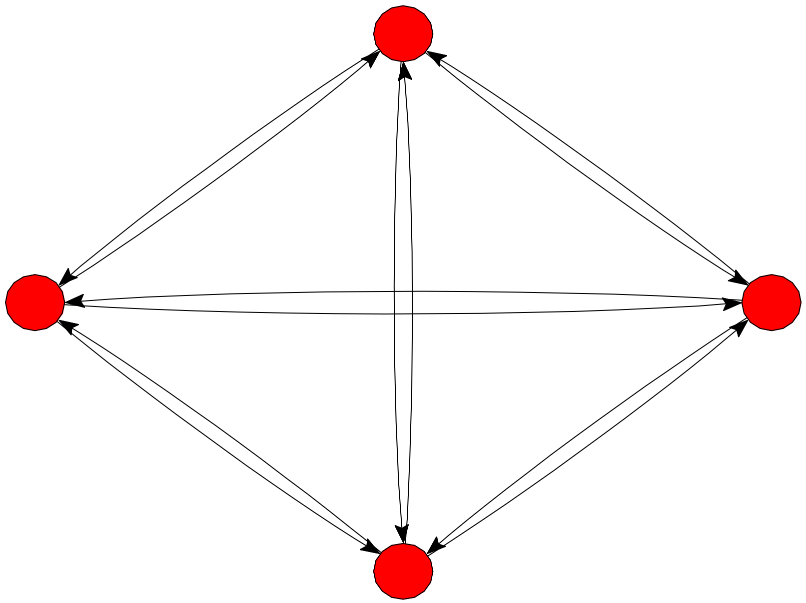
- **Maximal number of the connections that may exist between vertices:**
  - directed graph
    $L_{max} = N*(N-1)$
    since each of the N vertices can connect to (N-1) other vertices
  - undirected graph
    $L_{max} = N*(N-1)/2$
    since edges are undirected, count each one only once



- **What fraction is present?**

  Density = L / L_{max}

  - For example, out of 12 possible connections, a graph might have 7, giving it a density of 7/12 = 0.583



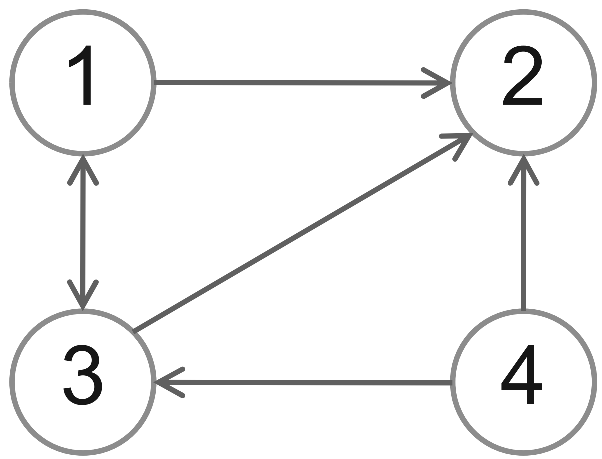
Example: Graph Density

$$density(G) = \frac{\sum_{i=1}^{|V|} \sum_{j=1, i \neq j}^{|V|} a_{ij}}{|V|(|V|-1)}$$

$$density(G) = \frac{6}{12} = 0.5$$

Density: degree of connectedness, i.e. number of existing edges in proportion to number of possible edges

G=(V,E)



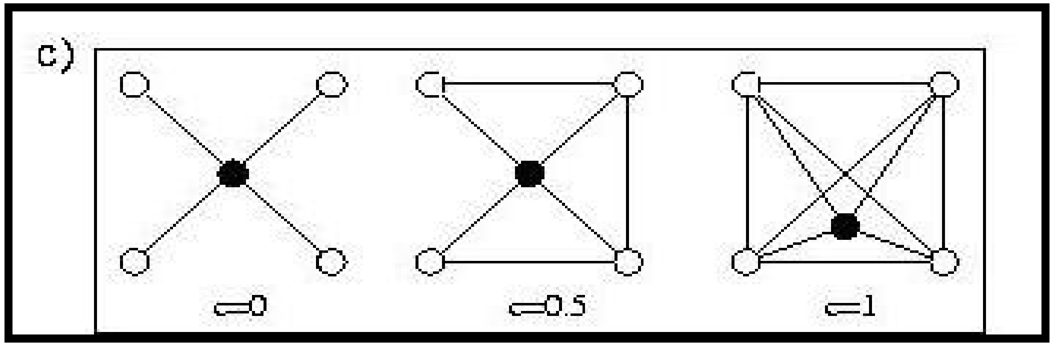
$$A = \begin{pmatrix} 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 \end{pmatrix}$$

Clustering Coefficient

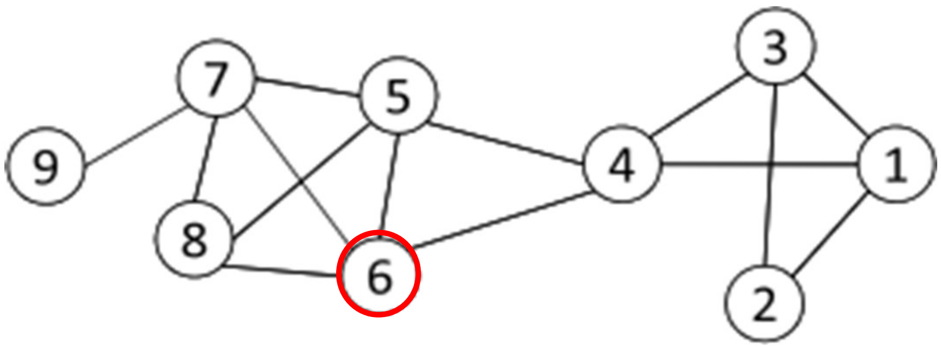
- Density of connections among one’s friends.
- What portion of your neighbors are connected?
- The clustering coefficient is kind of a “local” density measure

$$C_i = \begin{cases} \dfrac{k_i}{d_i \times (d_i - 1)/2} & d_i > 1 \\ 0 & d_i = 0 \text{ or } 1 \end{cases}$$

  - k_i number of edges (undirected) among V_i’s neighbors
  - d_i degree of vertex V_i
  - C_i in [0,1]



Example: Clustering Coefficient



$$C_i = \begin{cases} \frac{k_i}{d_i \times (d_i - 1)/2} & d_i > 1 \\ 0 & d_i = 0 \; or \; 1 \end{cases}$$

$d_6 = 4$, $N_6 = \{4, 5, 7, 8\}$

$k_6 = 4$ as e(4,5), e(5,7), e(5,8), e(7,8)

$C_6 = 4/(4*3/2) = 2/3$

Average Clustering Coefficient

$$C = (C_1 + C_2 + \ldots + C_n) / N$$

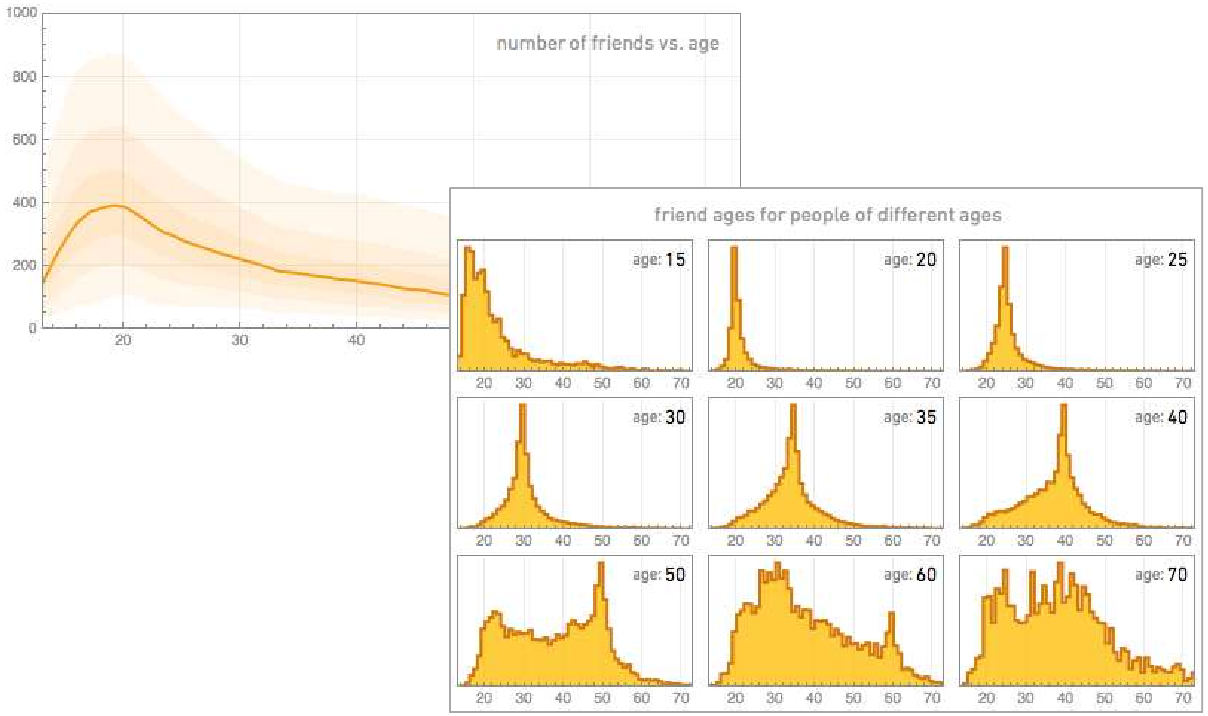
C = 0.61 for the left network

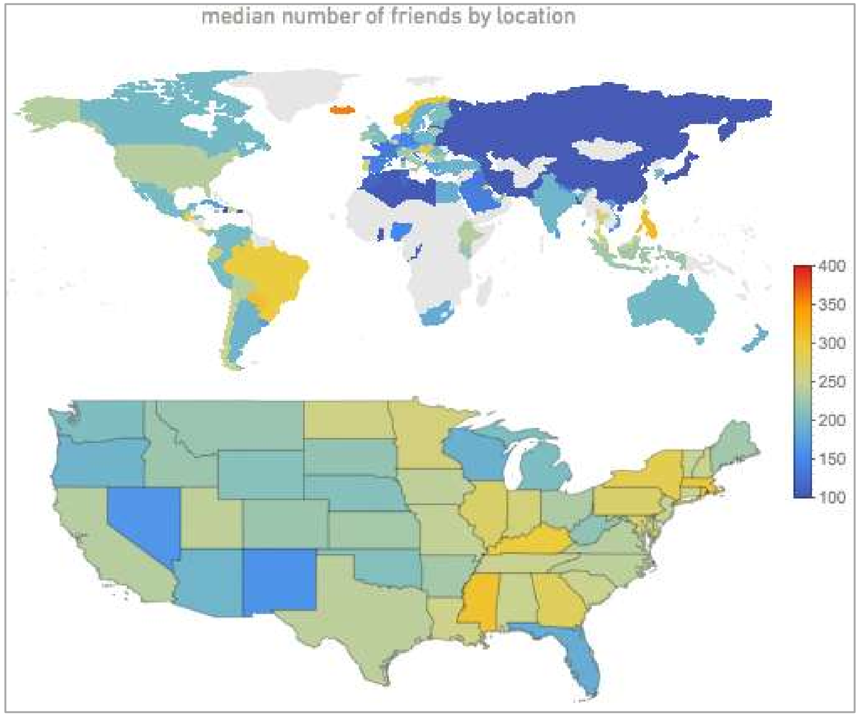
In a random graph, the expected average clustering coefficient is 0.19

➔ graph has some community structure

Vertex Attributes

- **vertices are described by attributes in many real-world networks**
  - e.g. social network with vertex attributes name, birthdate, address, interests, type …
- **combining these attributes with measures such as degree often reveals interesting insights**
  - e.g. number of friends on Facebook in relation to age and location





Source: http://blog.stephenwolfram.com/2013/04/data-science-of-the-facebook-world/

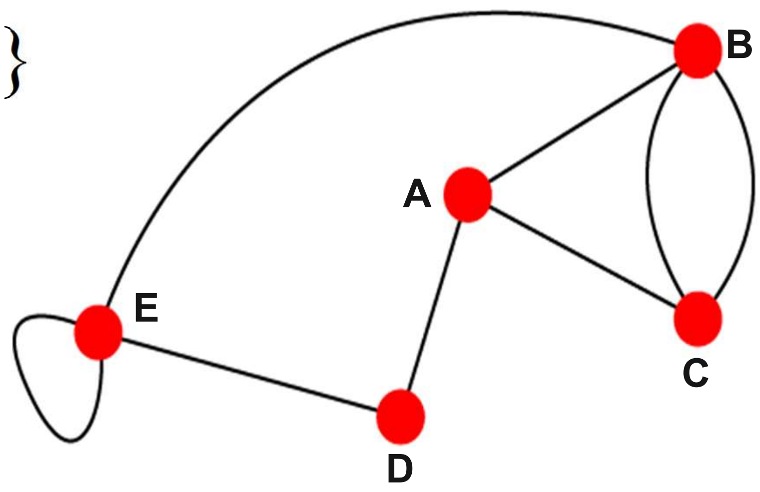
Paths

A path is a sequence of vertices in which each vertex is adjacent to the next one.

P_{i_0,i_n} of length n between vertices i_0 and i_n is an ordered collection of $n+1$ vertices and n lines.

$$P_{i0,in} = \{i_0, i_1, i_2, \ldots, i_n\}$$

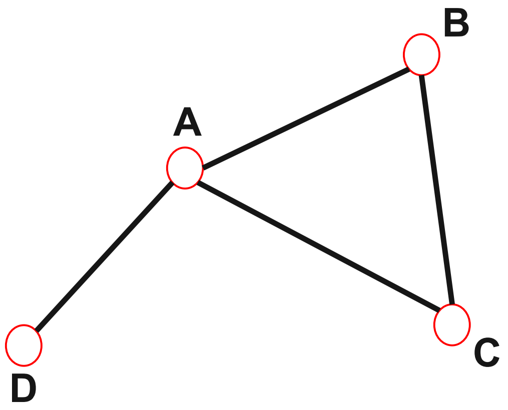
$$P_{i0,in} = \{(i_0, i_1), (i_1, i_2), (i_2, i_3), \ldots, (i_{n-1}, i_n)\}$$



- A path can intersect itself and pass through the same line repeatedly. Each time a line is crossed, it is counted separately
- A legitimate path on the graph on the right:

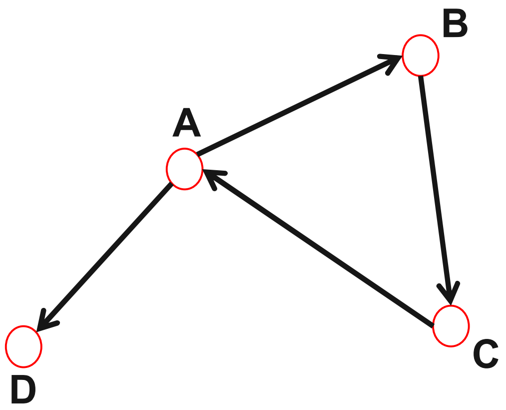
  A B C B C A D E E B A (length=11)
- In a directed graph, the path can follow only the direction of the arcs.

Shortest Path / Distance



The *distance (shortest path, geodesic path)* between two vertices is defined as the number of edges along the shortest path connecting them.

If the two nodes are disconnected, the distance is infinity.



In directed graphs each path needs to follow the direction of the arcs.

Thus, in a digraph the distance from vertex A to B (on an AB path) is generally different from the distance from vertex B to A (on a BCA path).

Network Diameter and Average Distance

Diameter

d_{max} the maximum distance between any pair of nodes in the graph.

Caution: Some people use the term ‘diameter’ to be the average shortest path length.

Average Distance / Average Shortest Path Length

<d> for a connected graph:

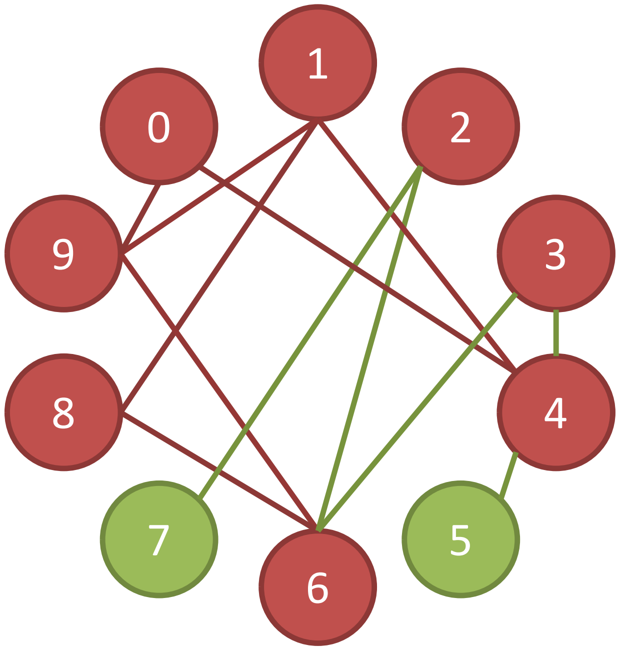
$$\langle d \rangle = \frac{1}{N(N-1)} \sum_{i,j} d_{ij}$$

where d_{ij} is the distance from vertex *i* to vertex j and N is the number of vertices.

– The average shortest path length distinguishes an easily navigable network from one which is complicated and inefficient (i.e. for information or mass transport).

Example: Diameter

The diameter of a graph is the maximum distance between any two nodes („longest shortest path“).

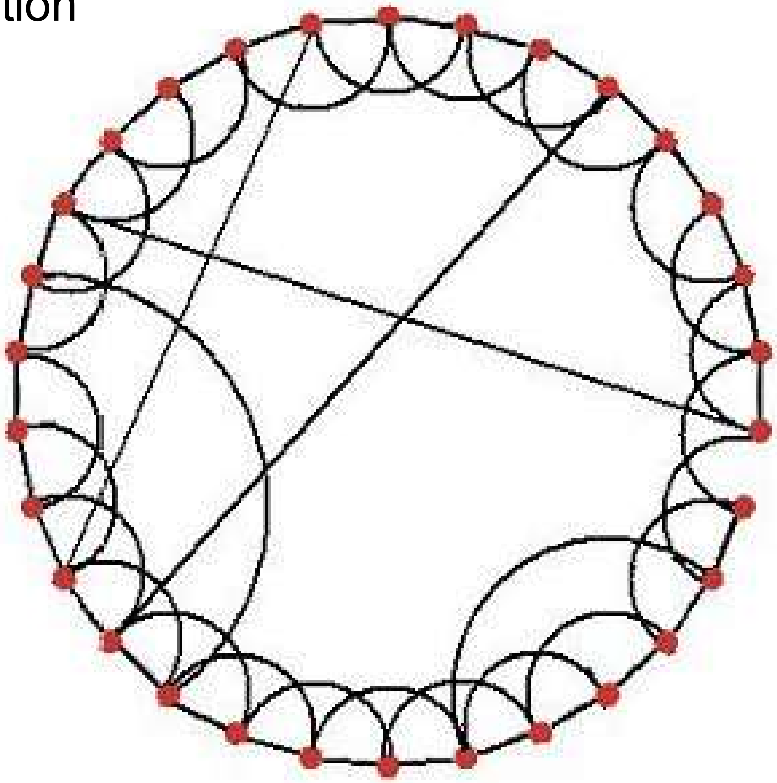


Diameter: 5 – because of path (7,2), (2,6), (6,3), (3,4), (4,5)

Small-World Phenomenon

Small-world networks are characterized by

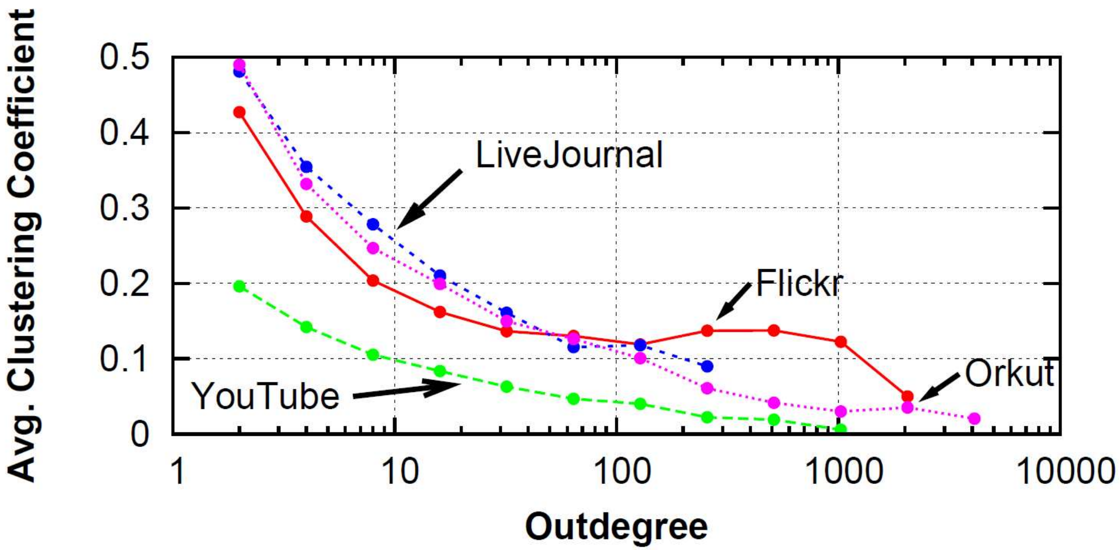
1. **high average clustering coefficient**
   - which indicates strong community structures
   - Explanation: Friends of a friend are likely to be friends as well
2. **small average shortest path length**
   - which is also known as “six degrees of separation” (Travers and Milgram,1969)
   - Explanation: Hub- or bridge-vertices connect communities and shorten the average path length
   - Hub- or bridge-vertices can be identified using betweenness centrality



Small-World Properties of Social Networks

Network	C	Ratio to Random Graphs Erdös-Rényi	Power-Law
Web [2]	0.081	7.71	-
Flickr	0.313	47,200	25.2
LiveJournal	0.330	119,000	17.8
Orkut	0.171	7,240	5.27
YouTube	0.136	36,900	69.4

The clustering coefficient is significantly higher compared to random networks

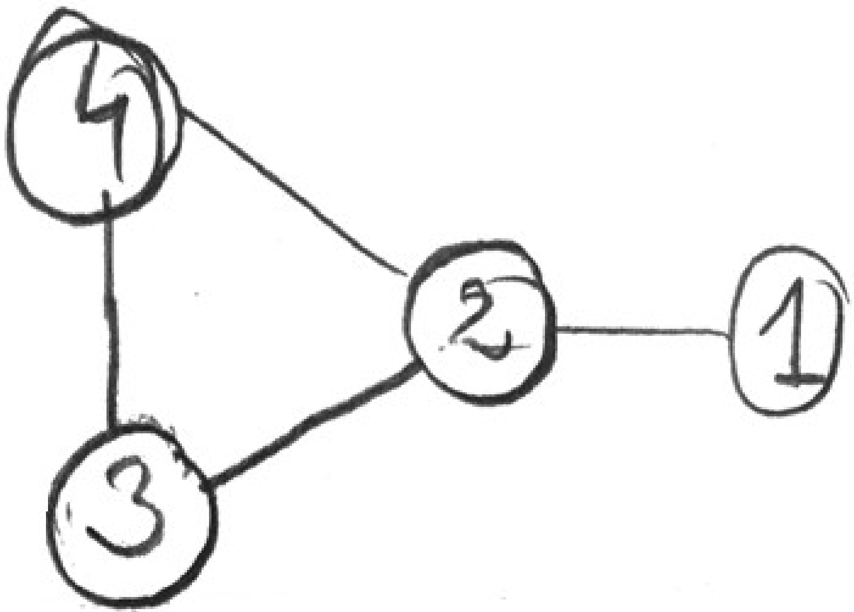


Users with few friends are more tightly clustered

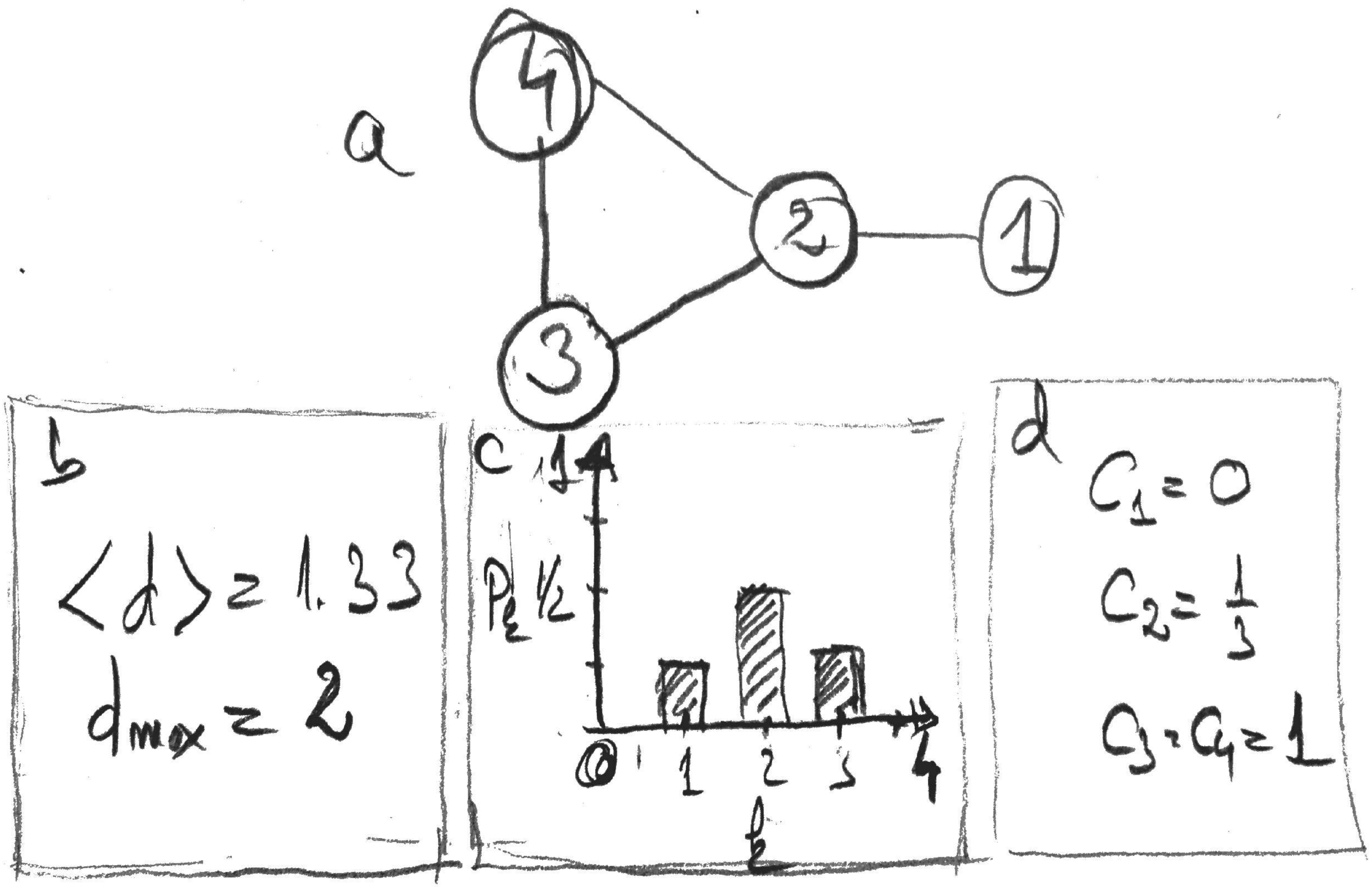
Source: Mislove, et al.: Measurement and Analysis of Online Social Networks. 2007.

Exercise: Characterizing a Graph

- Please calculate the following measures for the graph below:
  1. Diameter d_{max}
  2. Degree distribution
  3. Clustering coefficients C_2 and C_3



Solution: Characterizing a Graph



C = 0.58

Solution: Characterizing a Graph

$$C_2 = \frac{1}{3 \times (3-1)/2} = \frac{1}{3}$$

$$C_3 = \frac{1}{2 \times (2-1)/2} = \frac{1}{1} = C_4$$

- d_{max} = 2 because of shortest path between 3 and 1 as well as between 4 and 1
- Degree distribution: $P(k=1) = \frac{1}{4}$ $P(k=3) = \frac{1}{4}$

$$P(k=2) = \frac{1}{2}$$

2. Prominence

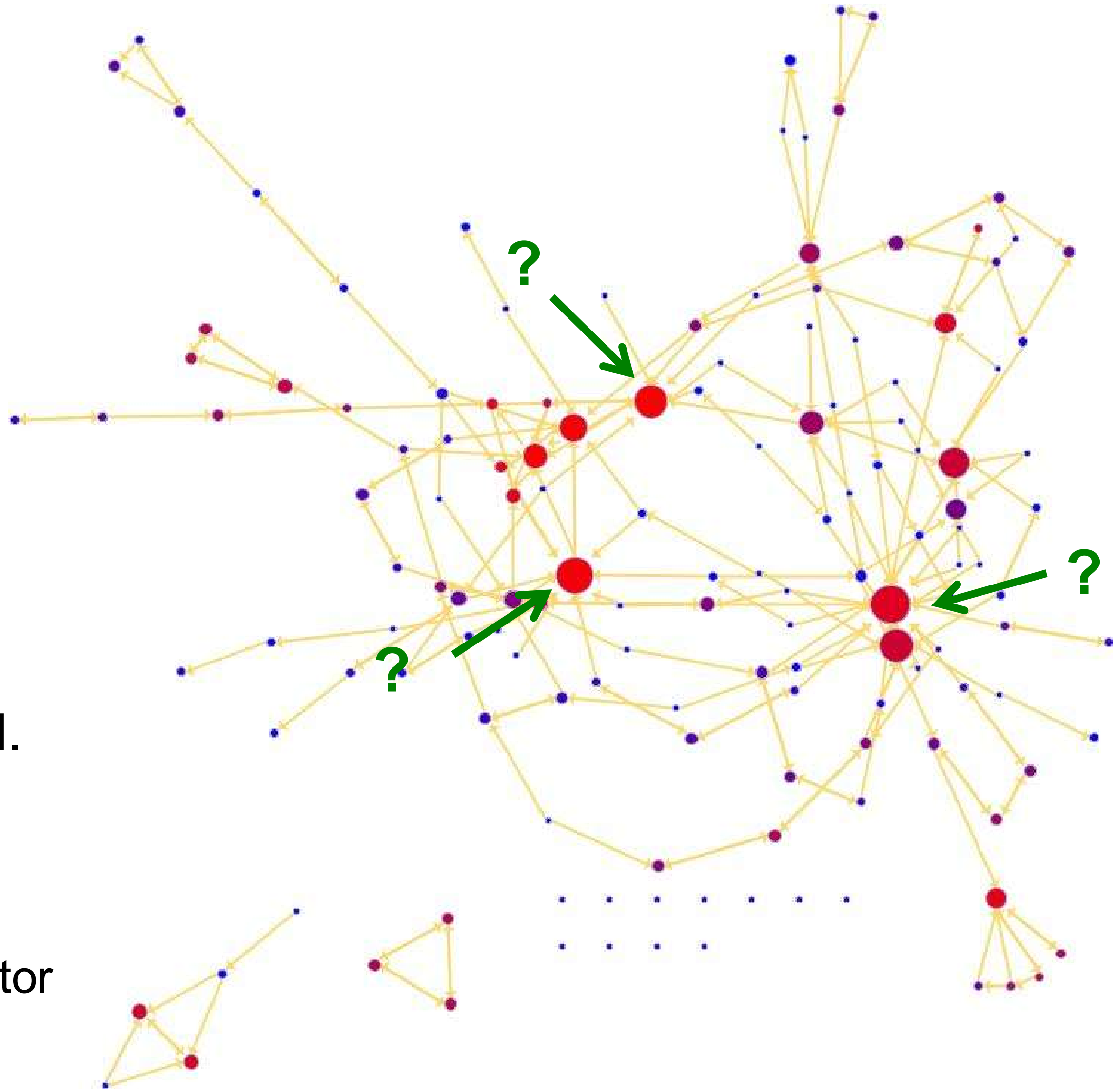
Who are the "most important" actors in a social network?

Centrality

- A central actor is one involved in many edges.
- The edge direction is not considered.

Prestige

- A prestigious actor is one who is the target of many arcs.
- The direction of arcs is considered.



- Possible interpretations
  - Centrality: „social power" of an actor
  - Prestige: „reputation" of an actor

2.1 Centrality

- Which nodes are most ‘central’?
  - Calculated for undirected graph
- Definition of ‘central’ varies by context / purpose:
- Local measure:
  - Degree centrality
- Relative to rest of network:
  - Closeness centrality
  - Betweenness centrality
- How evenly is centrality distributed among nodes?
  - Centralization
  - Graph-level view

Degree Centrality

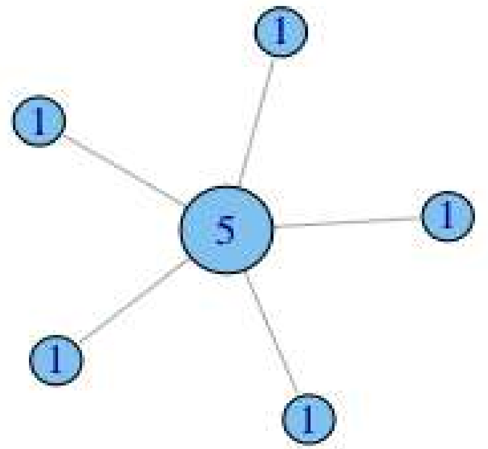
Idea: Measure centrality as the number of edges to other vertices in the graph.

Answers the question: How many people can a person directly reach or influence?

Degree Centrality

$$C_D(n_i) = d(n_i)$$

- Focuses only on direct choices (path length=1)



Normalized Degree Centrality

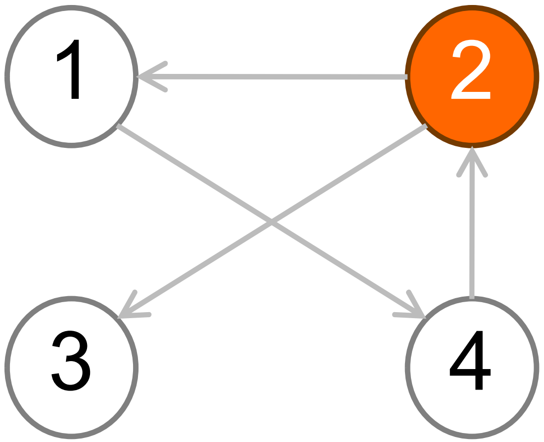
$$C'_D(n_i) = d(n_i) / N-1$$

- Degree divided by the maximal possible degree, i.e. number of vertices – 1
- Fraction of all nodes that are adjacent to n_i

Example: Degree Centrality

In undirected graphs, the centrality $C_D(v)$ of a node is its degree.

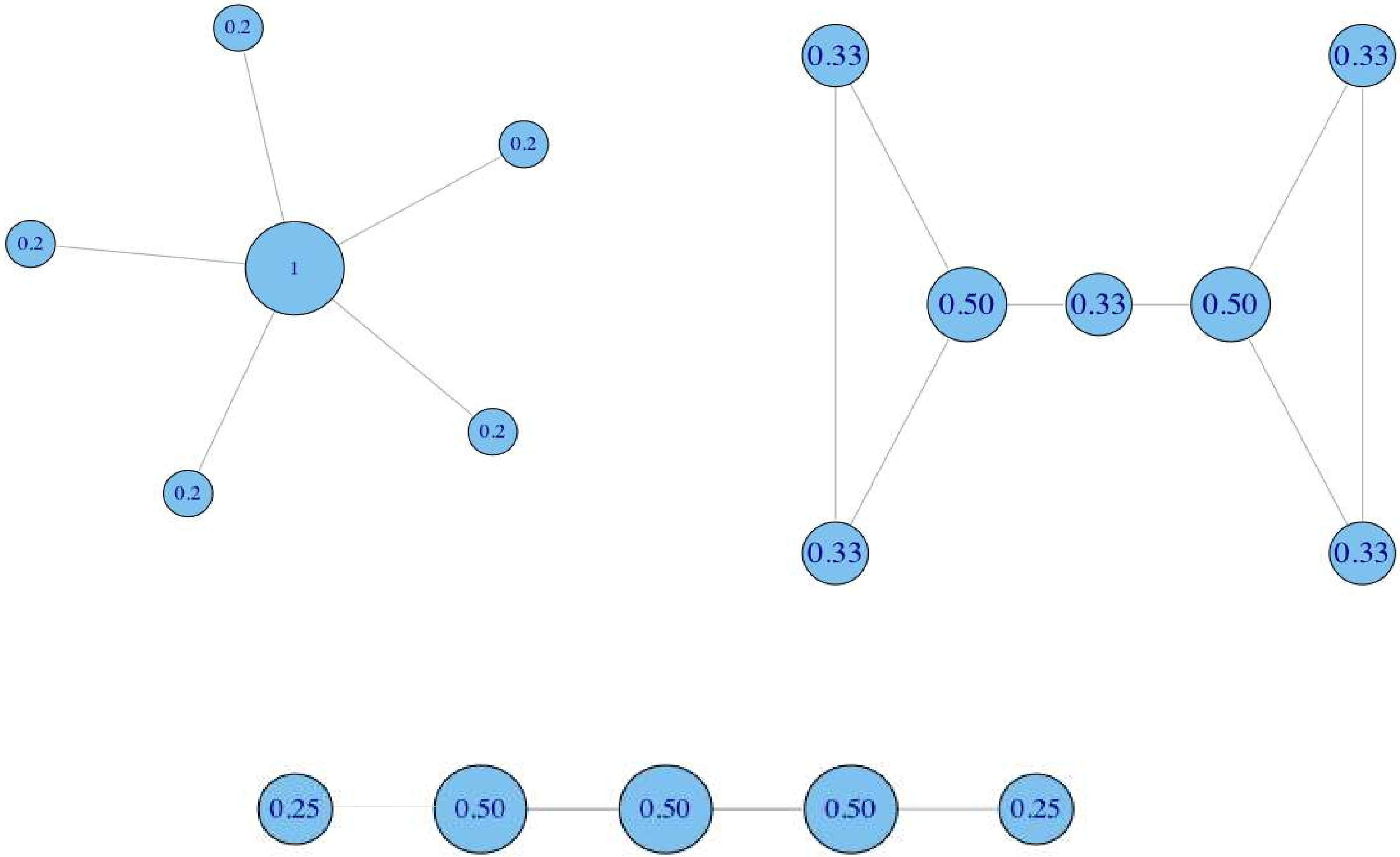
Normalization: $C'_D(v) = \dfrac{C_D(v)}{|V| - 1}$



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

$$C'_D(v_2) = \frac{2}{3}$$

Examples: Normalized Degree Centrality C'_D



Centralization

How much variation is there in the centrality scores among the vertices?

Freeman's general formula for centralization:

maximum value in the network

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

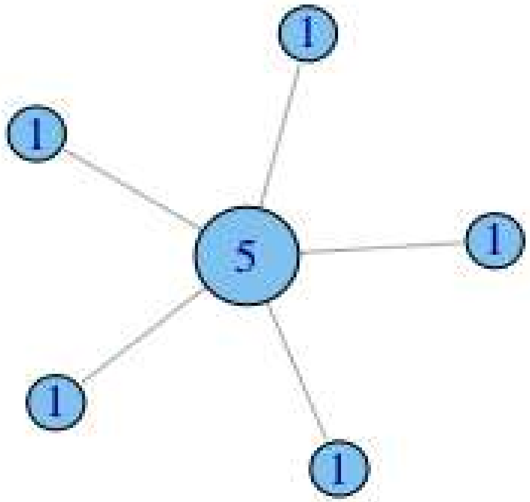
number of nodes without central node

maximal possible degree difference

1. calculate the sum of differences in centrality between the most central vertex in a graph and all other vertices;
2. divide this quantity by the theoretically largest sum of differences in any graph of the same degree (star shape graph).

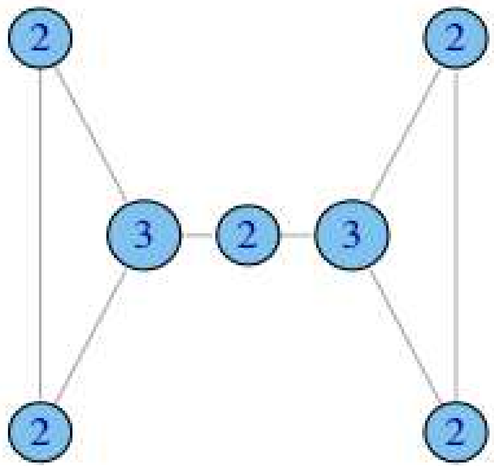
Value Range [0,1]

Examples: Degree Centralization



$C_D = 1.0$

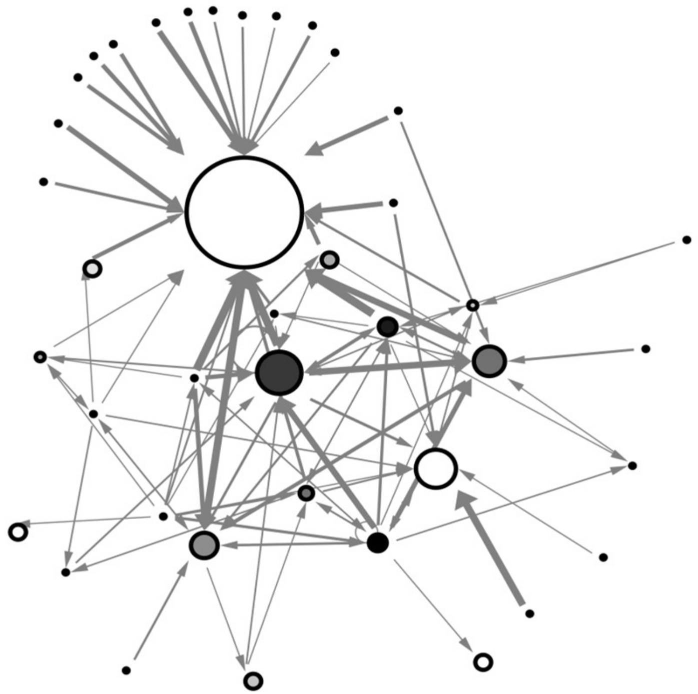
$C_D = 0.167$



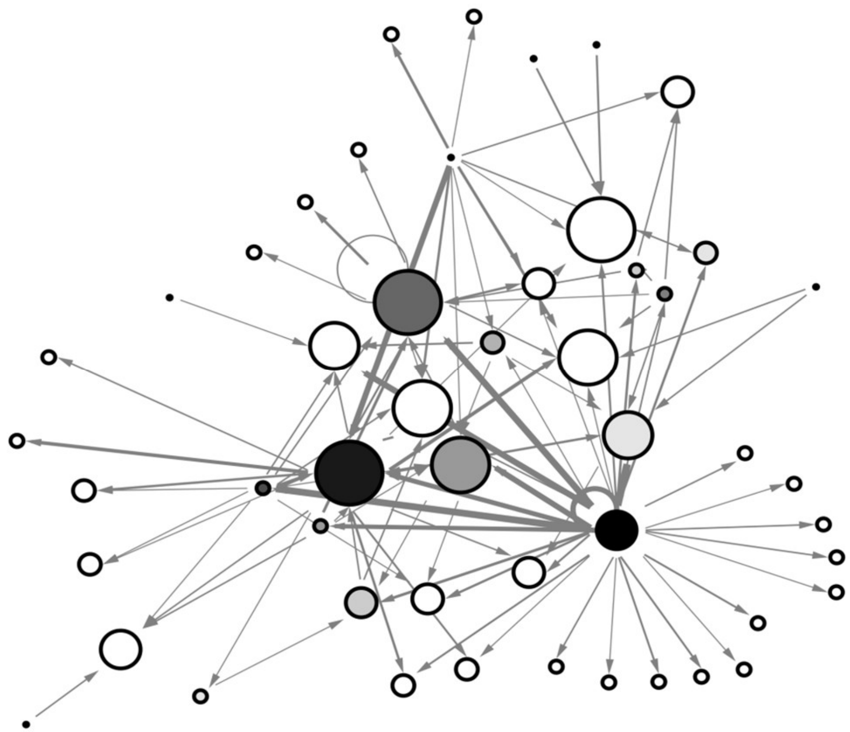
$C_D = 0.167$

Examples: Degree Centralization

Financial trading networks



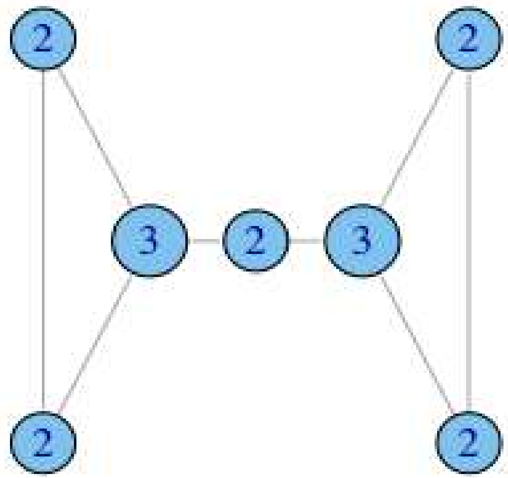
high centralization: one node trading with many others



low centralization: trades are more evenly distributed

When degree isn't everything

- **In what ways does degree fail to capture centrality in the following graphs?**

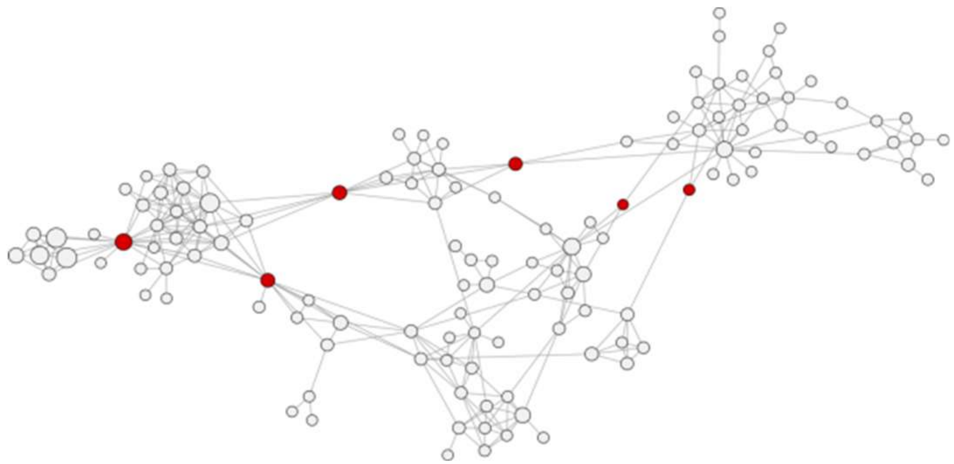


- **In what contexts may degree be insufficient to describe centrality?**
  1. Ability to broker between groups
  2. Likelihood that information originating from anywhere in the network reaches you

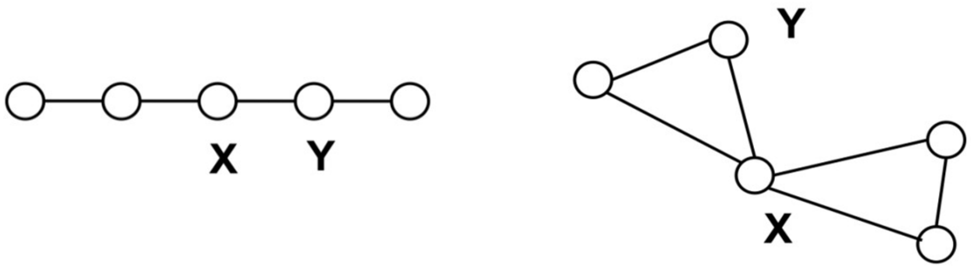
- **These use cases require measures that are relative to the rest of the network.**

Betweenness Centrality

Intuition: How many pairs of individuals would have to go through you in order to reach one another in the minimum number of hops?



- **Assumptions:**
  - Interactions between two non-adjacent actors might depend on the other actors in the set of actors, especially the actors who lie on the paths between the two nodes.
  - “Actor in the middle” between the others has some control over paths in the network – “interpersonal influence”.
- **Who has higher betweenness centrality, X or Y?**



Betweenness Centrality: Definition

$$C_B(i) = \sum_{j<k} g_{jk}(i) / g_{jk}$$

Where g_{jk} = the number of shortest paths connecting *jk*, and
$g_{jk}(i)$ = the number that actor *i* is on.

Usually normalized by dividing through maximal theoretical value for $C'_b(i)$:

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

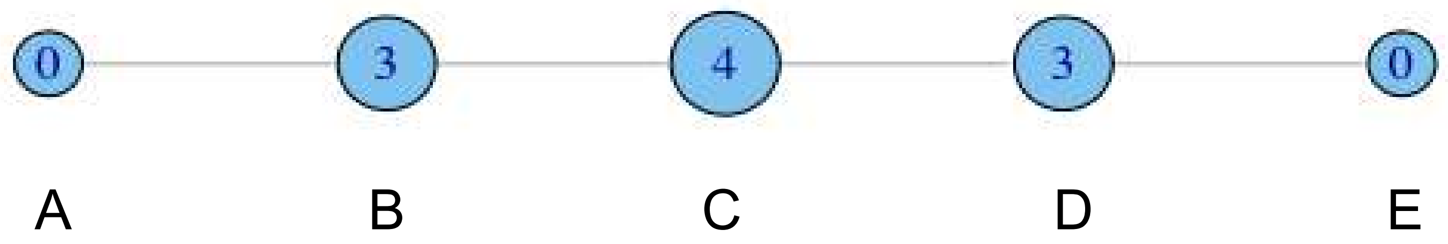
paths are symmetrical

number of vertices without the vertex itself

number of pairs of vertices excluding the vertex itself = shortest paths for each vertex

Example: Betweenness Centrality on Toy Networks

- **Non-normalized version:**

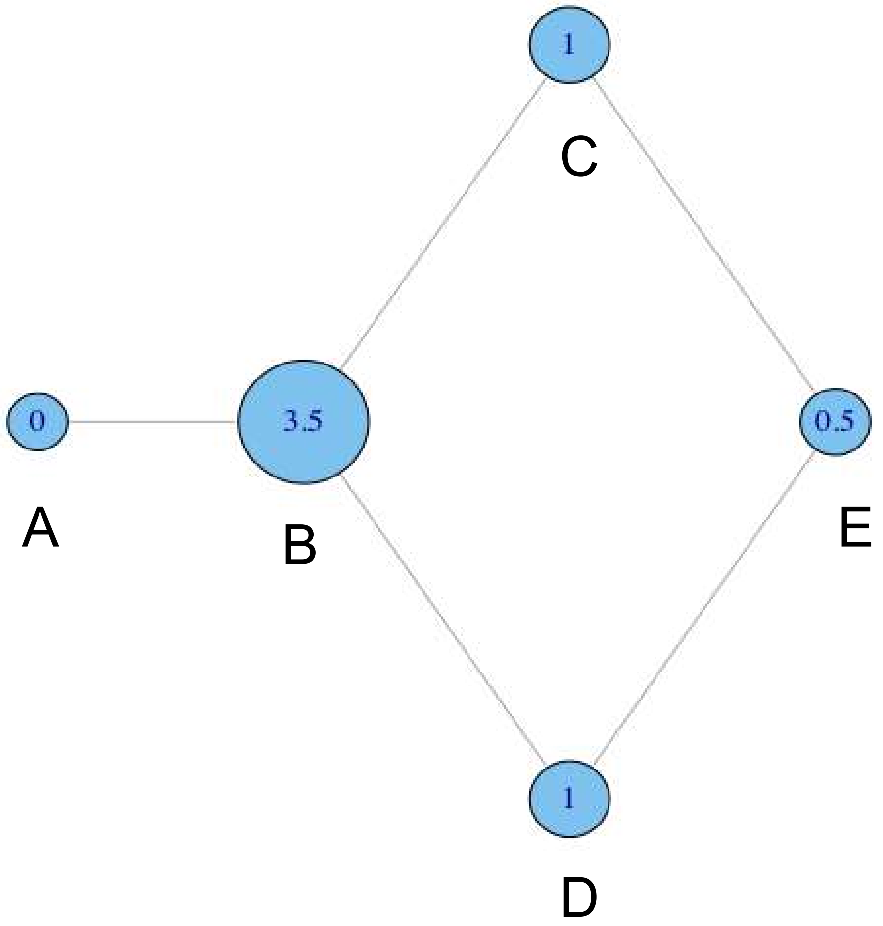


A lies between no two other vertices
B lies between A and 3 other vertices: C, D, and E
C lies between 4 pairs of vertices (A,D),(A,E),(B,D),(B,E)

Note that there are no alternate paths for these pairs to take, so C gets full credit.

Betweenness Centrality on Toy Networks

- **Non-normalized version:**



Why do C and D each have betweenness 1?

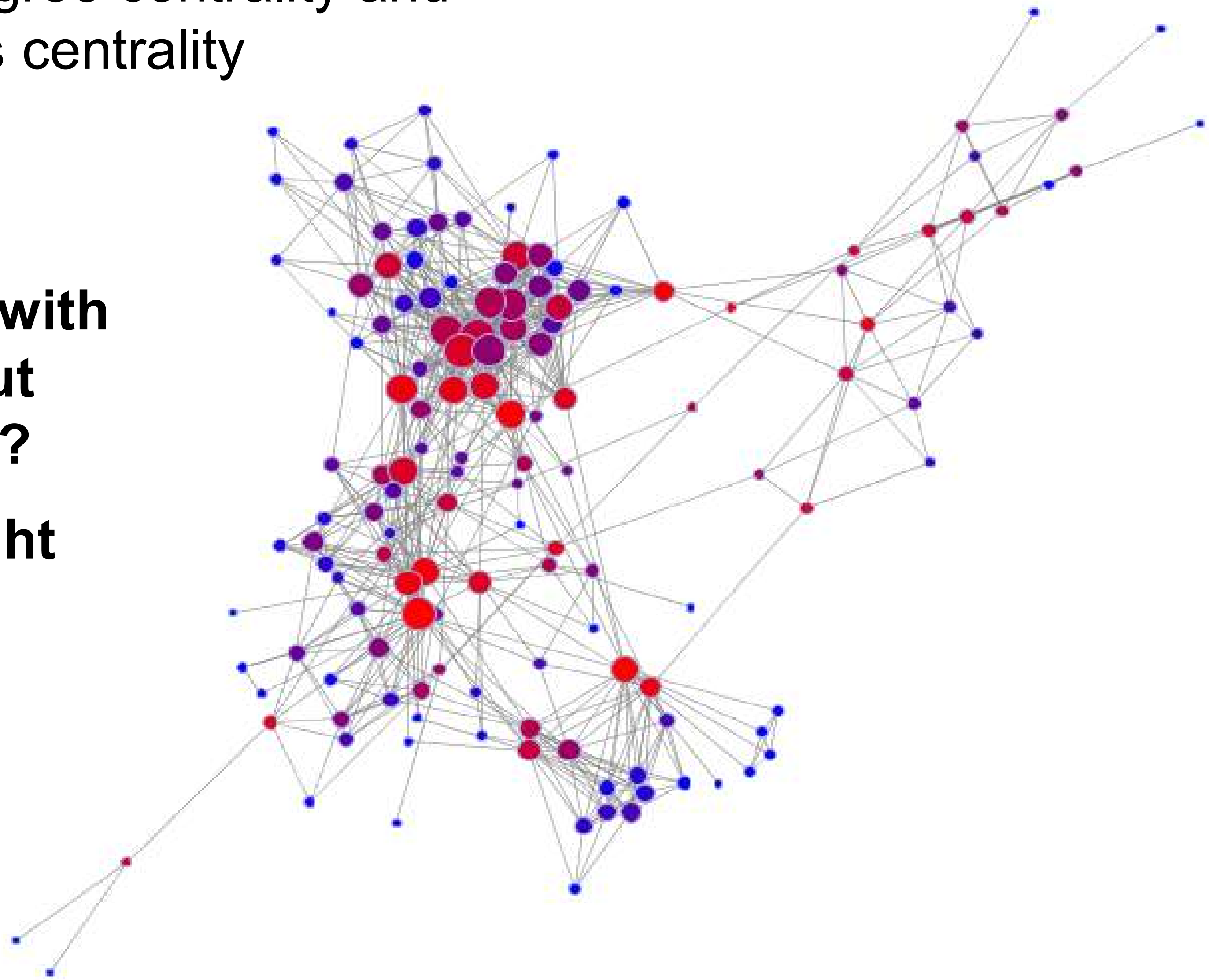
They are both on shortest paths for pairs (A,E), and (B,E), and so must share credit:

$\frac{1}{2}+\frac{1}{2} = 1$

Can you figure out why B has betweenness 3.5 while E has betweenness 0.5?

Example: Facebook Network

- vertices are sized by degree centrality and
- colored by betweenness centrality



1. **Can you spot nodes with high betweenness but relatively low degree?**
2. **Explain how this might arise**.

Closeness Centrality

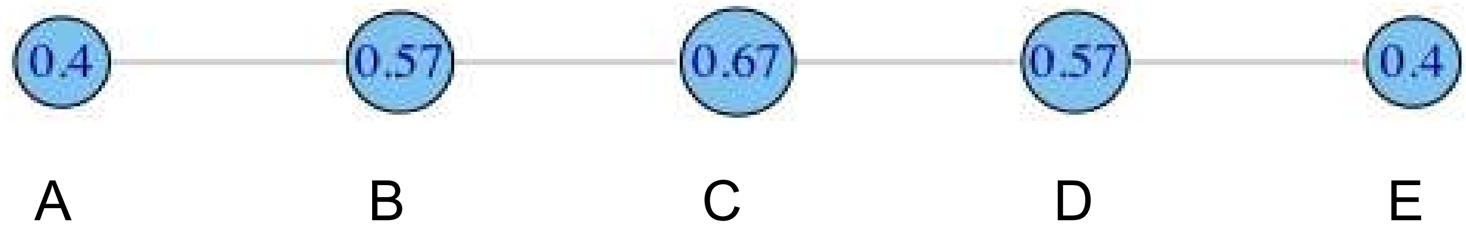
- **The measure focuses on how close an actor is to all the other actors in the network**
  - for instance to spread information or interact with others
  - or to be reached by information that spreads through the network
- **Closeness centrality is based on the length of the average shortest path between a vertex and all vertices in the graph.**

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

Normalized Closeness Centrality: $$C'_C(i) = C_C(i)(N-1)$$

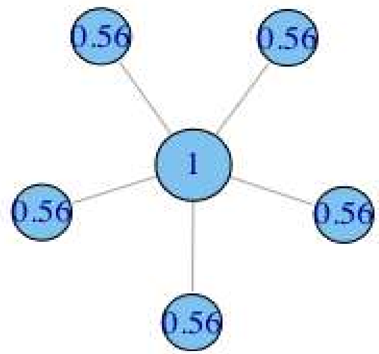
star shape: each vertex has distance one to central vertex

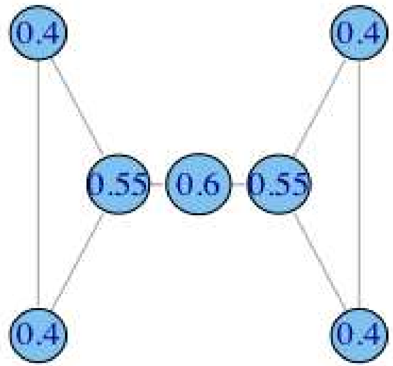
Example: Normalized Closeness Centrality



$$C'_c(A) = \left[\frac{\sum_{j=1}^{N} d(A,j)}{N-1} \right]^{-1} = \left[\frac{1+2+3+4}{4} \right]^{-1} = \left[\frac{10}{4} \right]^{-1} = 0.4$$

More toy examples:

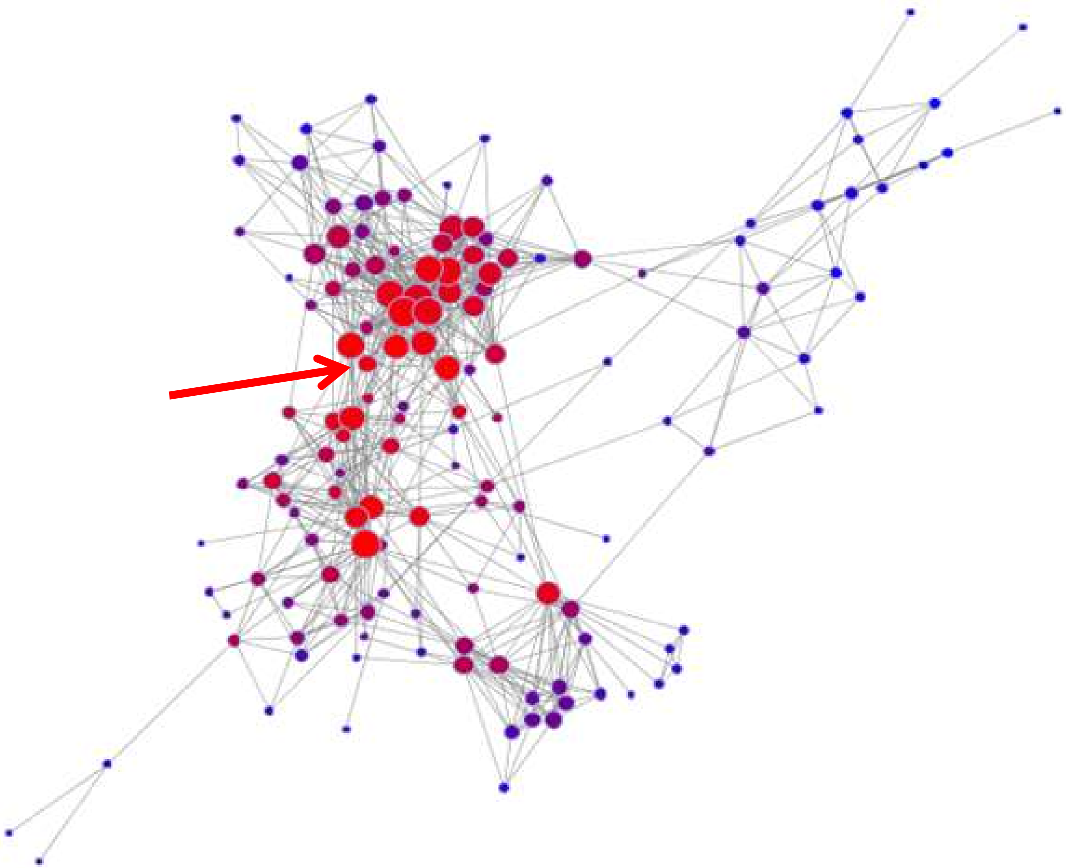




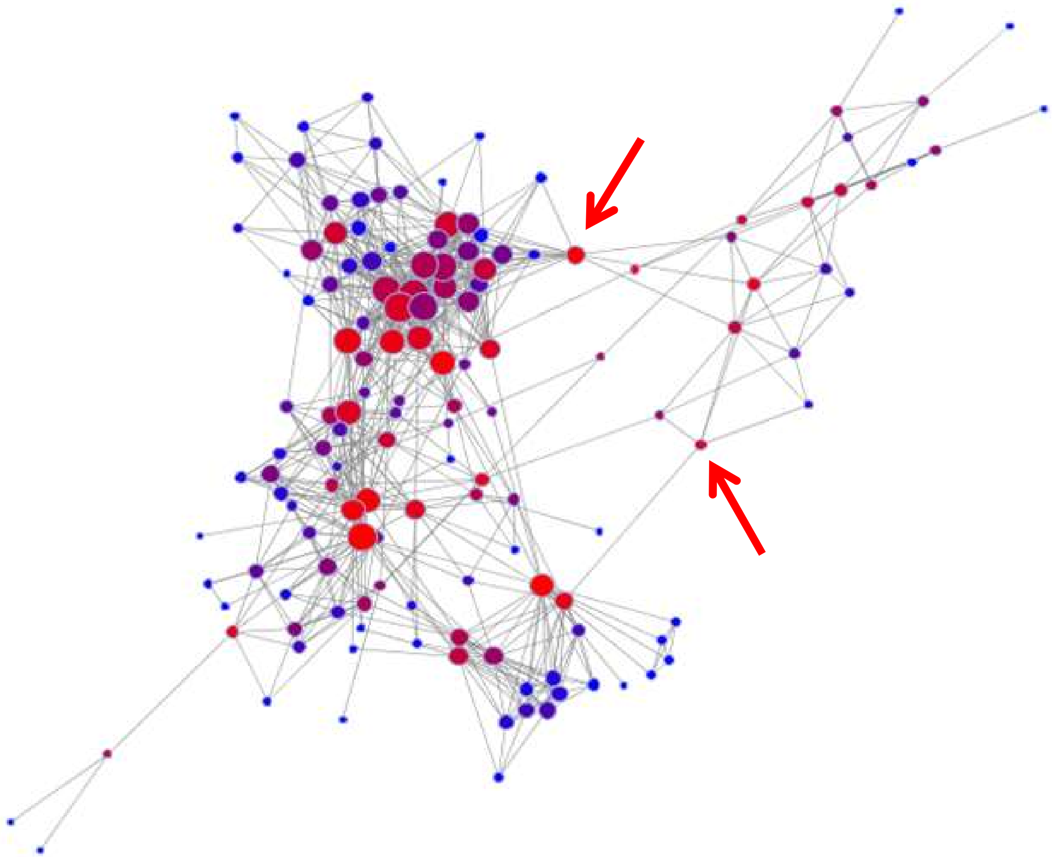
Correlation of Centrality Metrics

- Generally different centrality metrics will be positively correlated.
- When they are not, there is likely something interesting about the vertex.

closeness centrality denoted by color
degree denoted by size



betweeness centrality denoted by color
degree denoted by size



3.2 Prestige

Prestige refers to a class of prominence metrics which take the direction of arcs into account.

- Translates to: choices received
- Examples where direction matters:
  - votes in an election
  - hyperlinks on the WWW
  - likes on TikTok
  - citations of scientific papers
- Examples when ‘prestige’ may not be the right word
  - dislikes
  - distrusts

Degree Prestige / Popularity

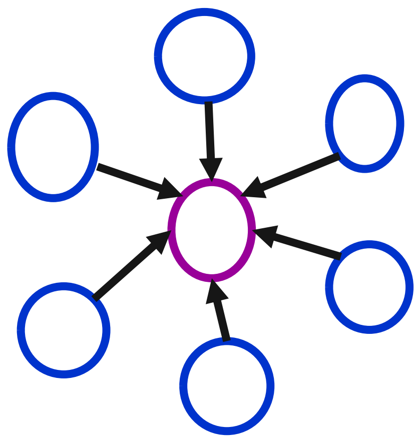
- The simplest vertex-level measure of prestige: in-degree
- The idea is that actors who are prestigious tend to receive many nominations or choices
  - a paper that is cited by many others has high prestige
  - a person nominated by many others for a reward has high prestige
- Local measure as only the neighbors are taken into account.

Degree Prestige / Popularity

$$P_D(n_i) = d^{in}(n_i)$$

Normalized Degree Prestige

$$C'_D(n_i) = d^{in}(n_i) / N-1$$



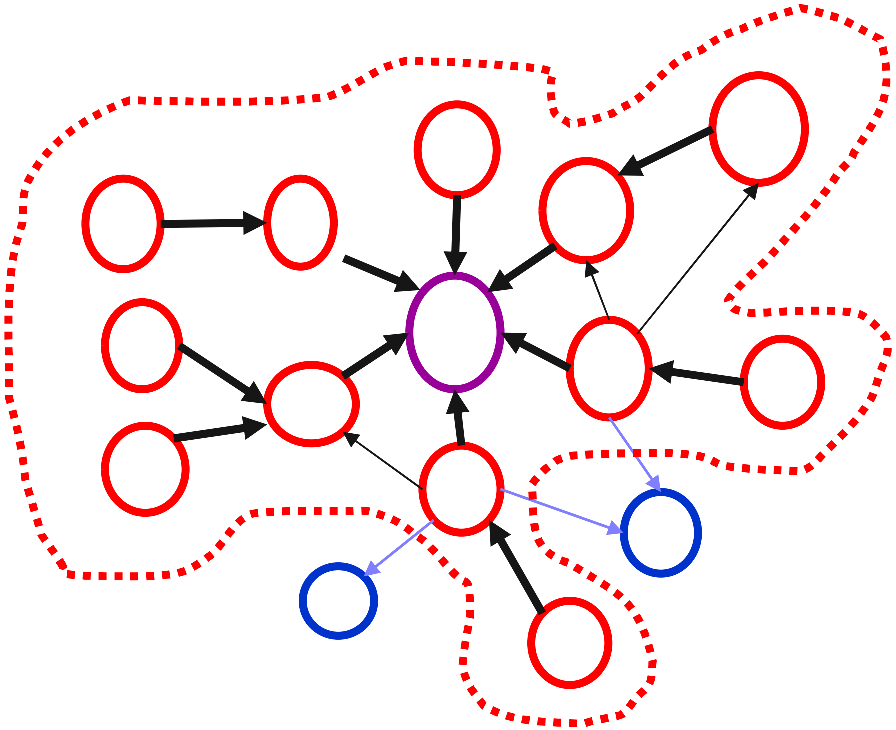
- Indegree divide by the maximal possible indegree
- Fraction of all nodes that choose n_i (e.g. fraction of a vote)

Input Domain

- Degree prestige only counts actors who are directly adjacent to actor n_i, but we might also want to take indirect choices into account.

The input domain of a vertex in a directed network is the number or percentage of all other vertices that are connected by a path to this vertex.

- Also called influence domain, which makes for instance sense for the use case of following on Twitter.



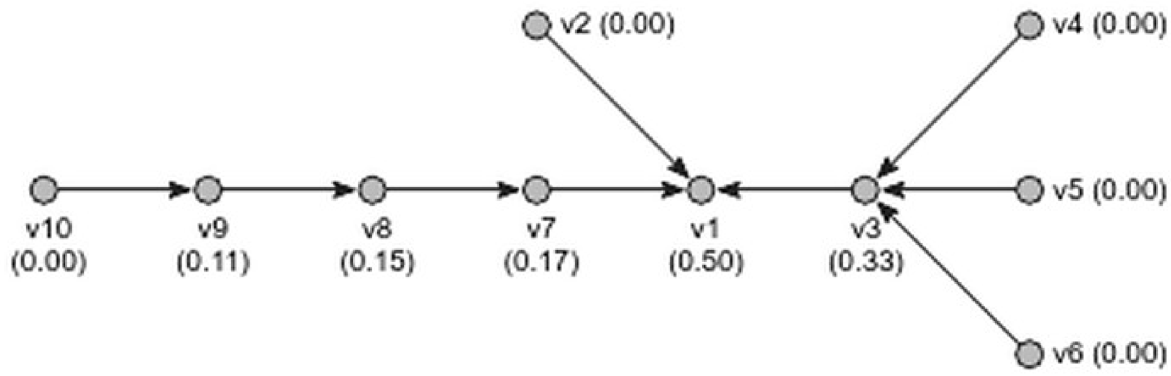
Proximity Prestige

- **Prestige measure based on distances in the input domain.**
  - Direct nominations (choices) should count more than indirect ones
  - Nominations from second degree neighbors should count more than third degree ones

$$P_p(v_i) = \frac{\text{fraction of all vertices that are in } i\text{'s input domain}}{\text{average distance from } i \text{ to vertex in input domain}}$$

$$P_P(v) = \frac{|I_v|/(|V|-1)}{\left(\sum_{i \in I_v} dist(i,v)\ \right)\ /\ |I_v|}$$

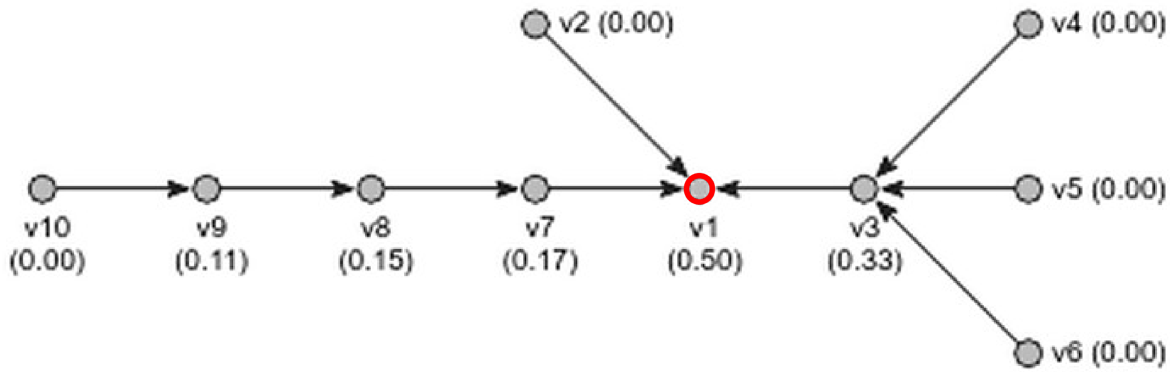
Example:



Example: Proximity Prestige of Vertex v1

$$P_P(v) = \frac{|I_v|/(|V|-1)}{(\sum_{i \in I_v} dist(i,v)\ )\ /\ |I_v|}$$

Example:



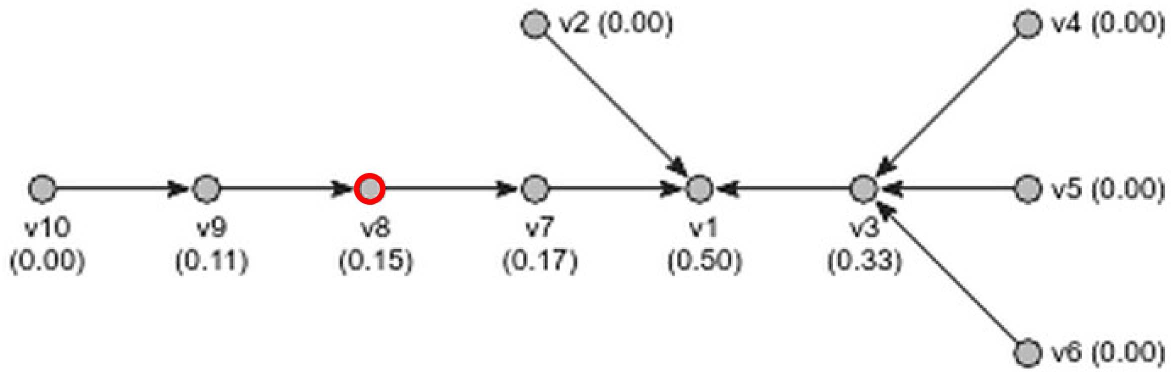
$$I_{v_1} = \{v_2, v_3, v_4, v_5, v_6, v_7, v_8, v_9, v_{10}\}$$

$$P_P(v_1) = \frac{9/9}{(1+1+2+2+2+1+2+3+4)/9} = 0,5$$

Example: Proximity Prestige of Vertex v8

$$P_P(v) = \frac{|I_v|/(|V|-1)}{(\sum_{i \in I_v} dist(i,v)\ )\ /\ |I_v|}$$

Example:



$$I_{v_8} = \{v_9, v_{10}\}$$

$$P_P(v_8) = \frac{2/9}{(2+1)/2} = 0,148148148$$

Rank Prestige and Page Rank

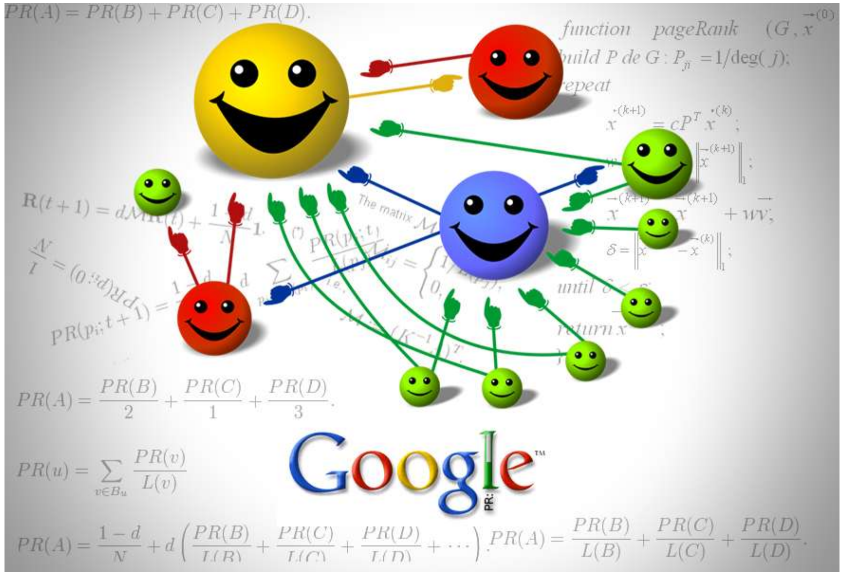
- **Rank Prestige**
  - Prestige measure which considers the prestige of the actors who do the “choosing”.
  - You are more prestigious if you have lots of other prestigious people in your input domain.

$$P_R(i) = \sum_{(j,i)\in E} P_R(j)$$

j: Vertex in the input domain of i

- **Page Rank**
  - Variation of rank prestige in which the prestige of a voting node is shared between all link targets.

$$P_{PR}(i) = \sum_{(j,i)\in E} P_{PR}(j) / D_{out}(j)$$



  - Advantages of PageRank in the search context
    - hard to trick with SPAM links
    - the score is independent of actual search engine query
  - Calculation of PageRank Score: See Bing Lui: Web Data Mining. Chapter 7.3

Literature

1. Zafarani, et al: Social Media Mining. Cambridge University Press, 2014. Free online version http://www.socialmediamining.info
2. Wasserman and Faust: Social Network Analysis. Cambridge University Press, 1994.
3. David Easley, Jon Kleinberg: Networks, Crowds, and Markets: Reasoning About a Highly Connected World. Cambridge University Press, 2010. Free online version http://www.cs.cornell.edu/home/kleinber/networks-book/
4. Bing Liu: Web Data Mining. 2nd Edition, Springer, 2011.