

Network Mining

Generalities, Complex Networks and Node-Centric Metrics

— Session 1 —

Outline

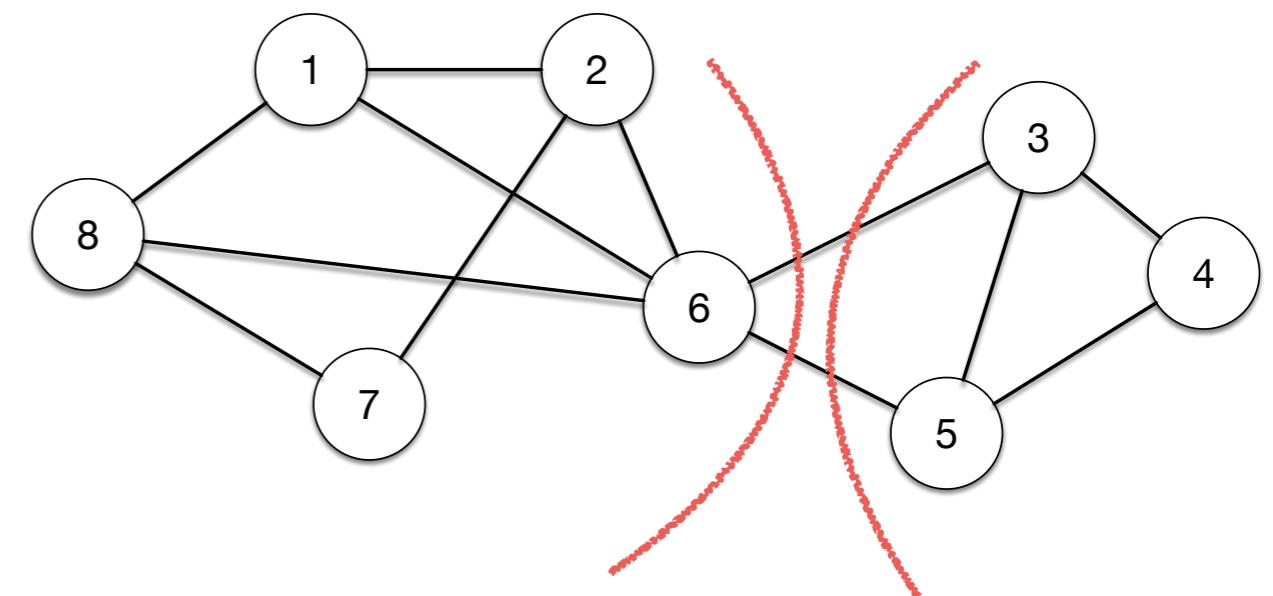
1. Introduction to network science
2. Complex network properties
 - A. Long tail distribution
 - B. Small world effect
 - C. Community structure
3. Node-centric metrics
4. Link Analysis
5. Influence Model

An interdisciplinary science

- ▶ Sociology (social networks)
- ▶ Mathematics (graphs)
- ▶ Computer science (graphs)
- ▶ Statistical physics (complex networks)
- ▶ Economics (networks)
- ▶ Bioinformatics (networks)

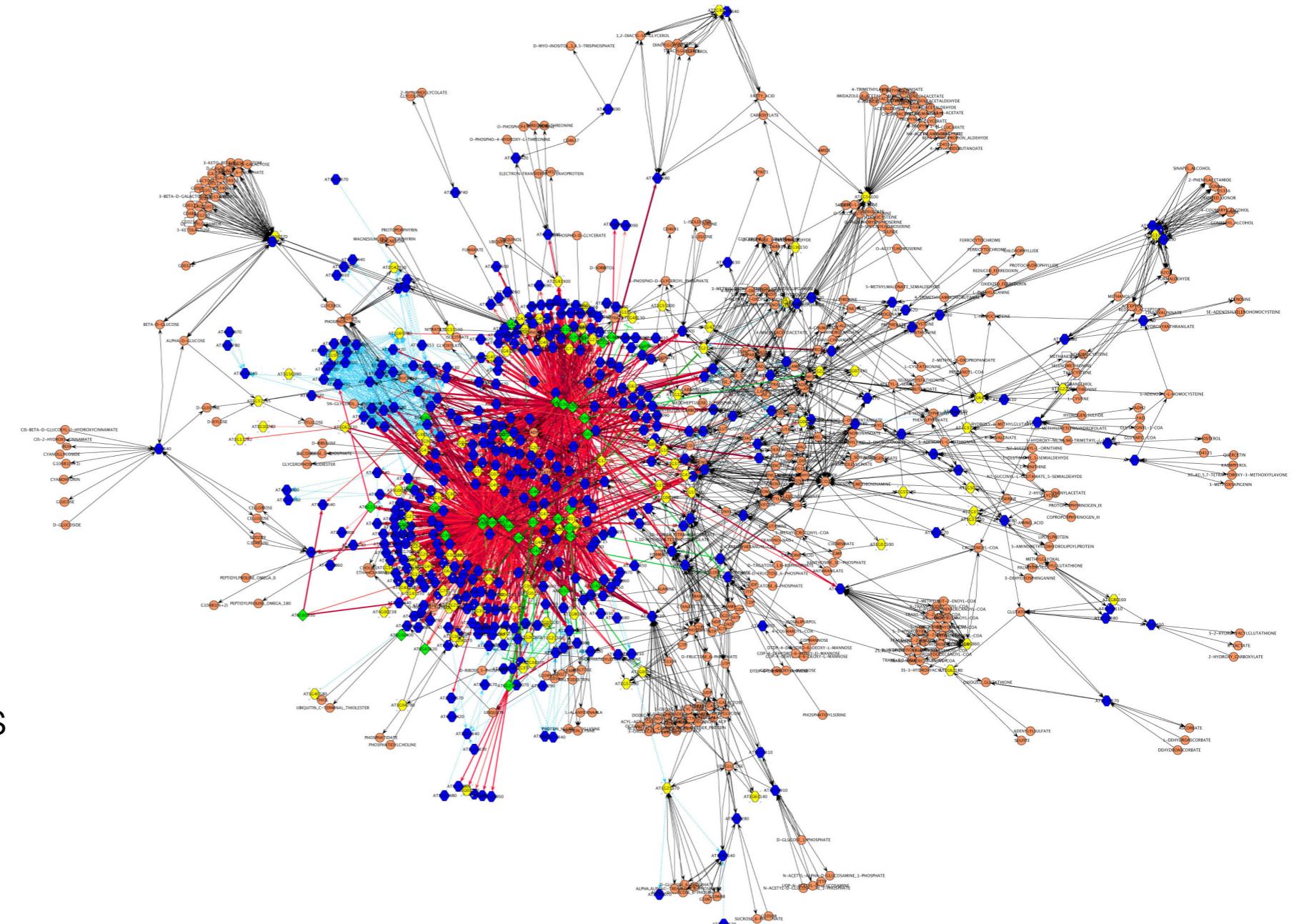
Let's speak the same language

- ▶ Network = graph
- ▶ Nodes = vertices, actors
- ▶ Links = ties, edges, relations
- ▶ Clusters = communities



Complex networks

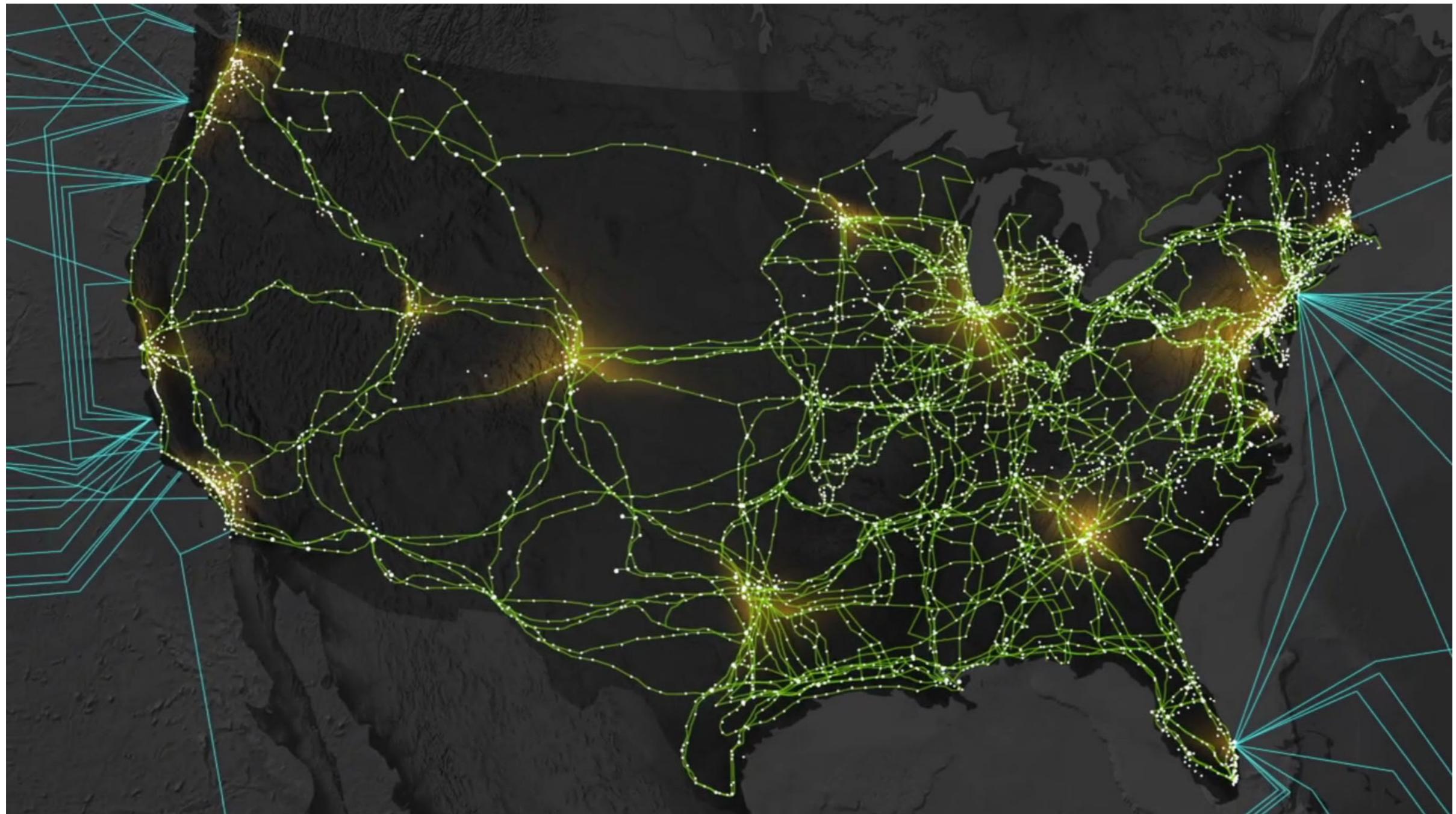
- ▶ Not regular
- ▶ Not random
- ▶ Scale-free networks
- ▶ Universal properties
- ▶ Very common
- ▶ Model complex systems



Source: <http://images.biomedsearch.com>

Complex networks

US internet infrastructure

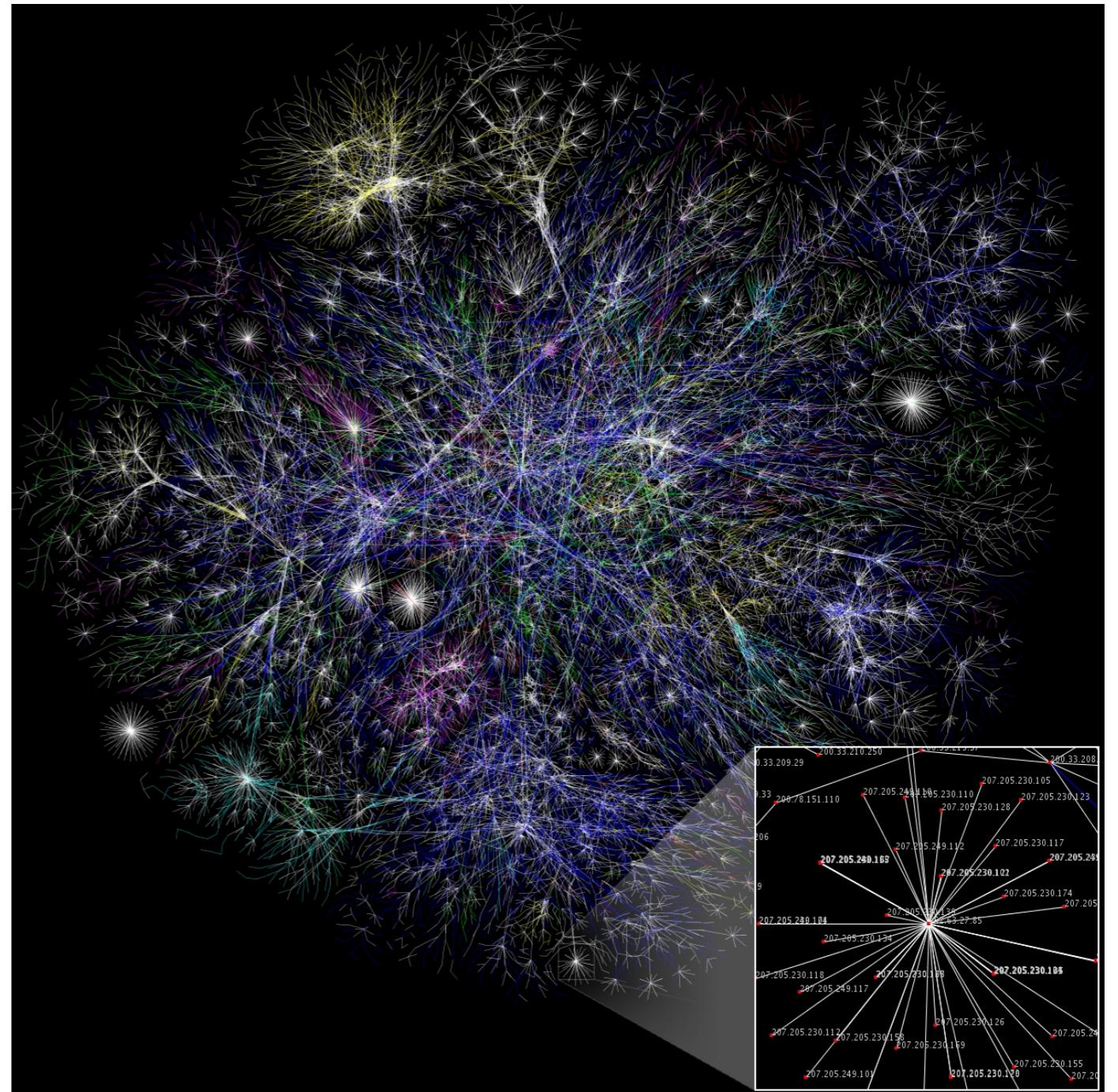


Source: <http://www.aec.at/nextidea/tracing-the-digitalphysical/>

Complex networks

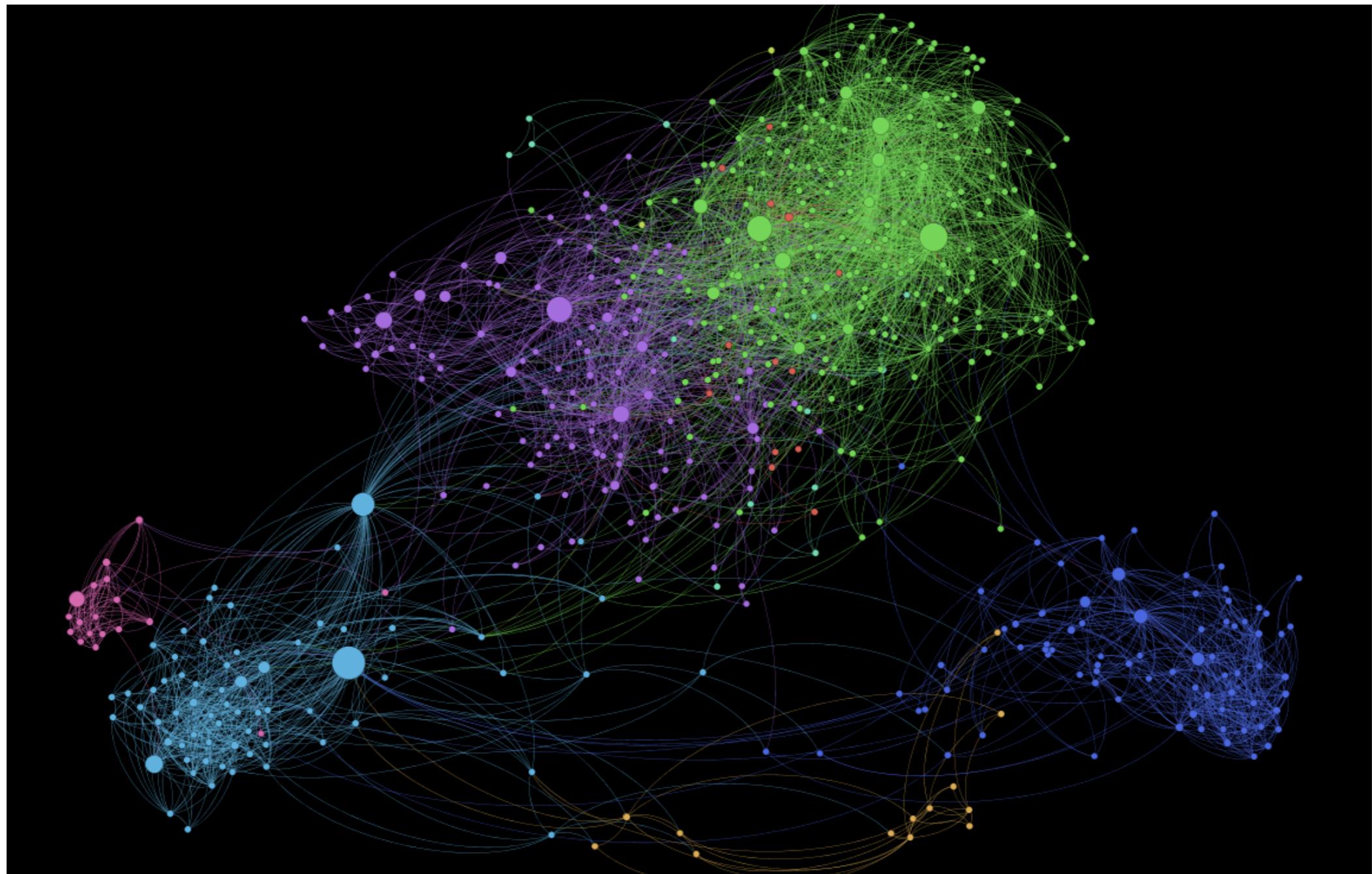
Class C networks

- ▶ Barrett Lyon, OPTE.org
- ▶ January 2015
- ▶ Node : IP address
- ▶ Link length : delay between two nodes



Complex networks

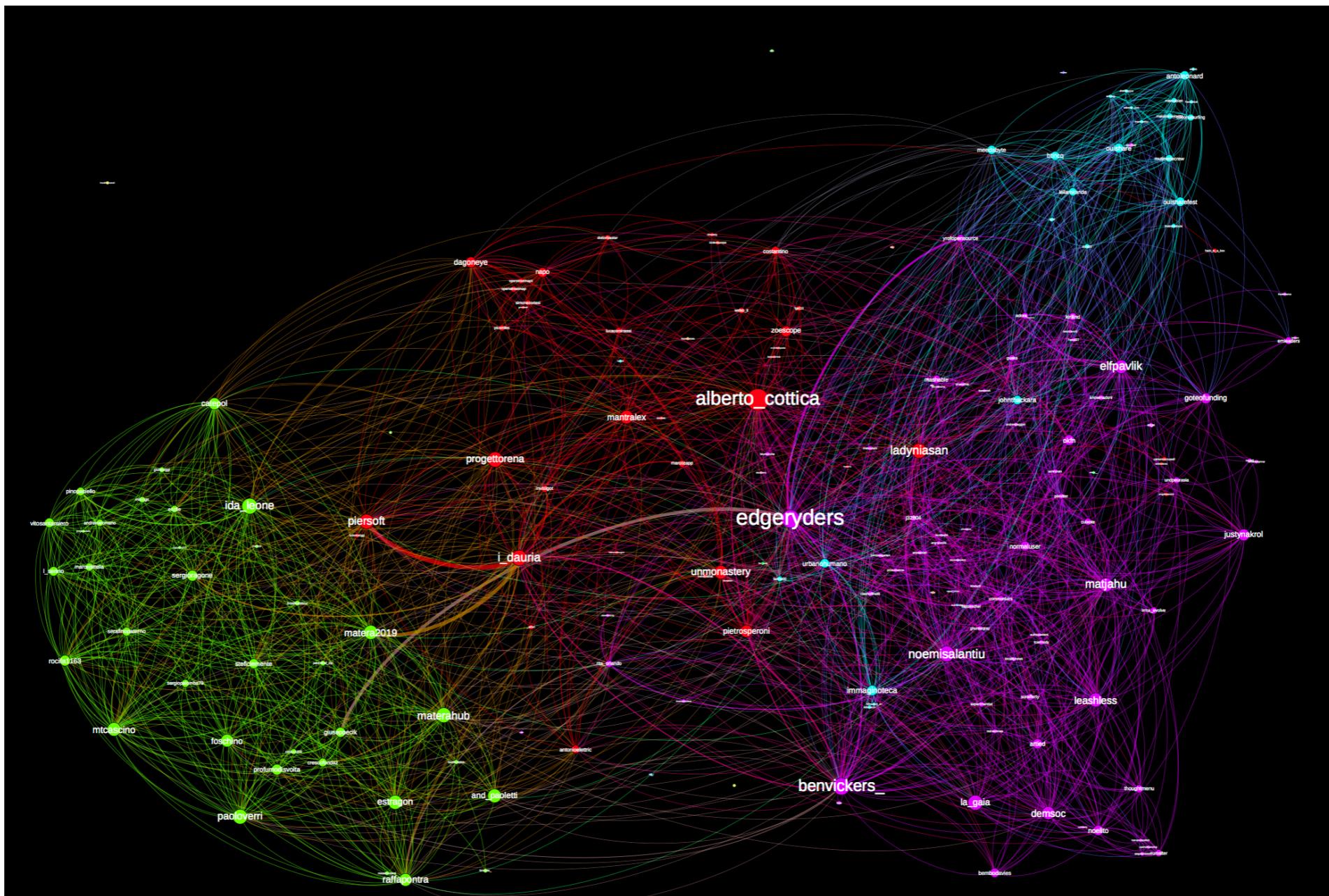
Study about Somalis from their Facebook networks



Source: <https://kimoquaintance.com/2011/08/22/what-can-we-learn-about-somalis-from-their-facebook-networks/>

Complex networks

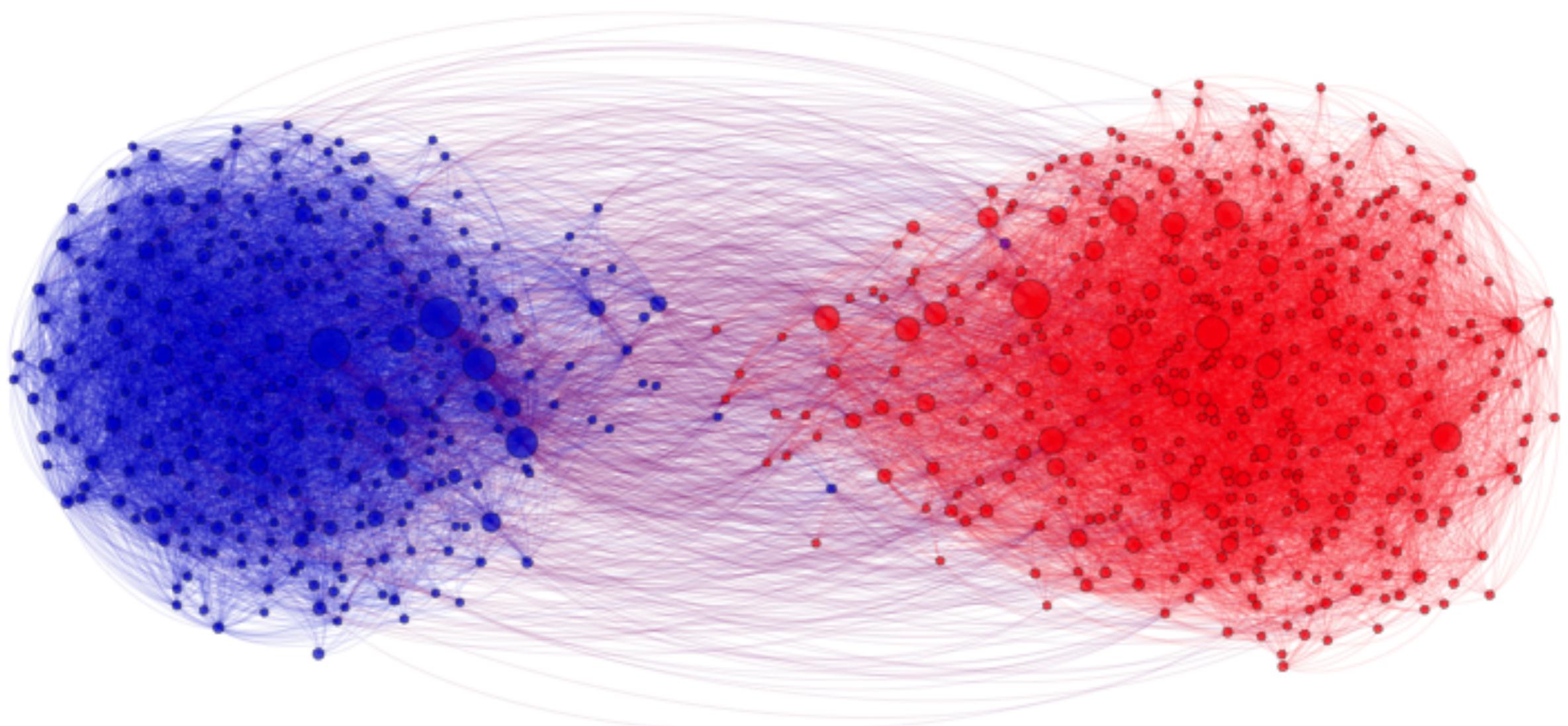
Twitter



Source: <http://www.cottica.net/2013/10/19/learning-from-the-twitterstorm-an-architecture-for-effortless-collaboration/>

Complex networks

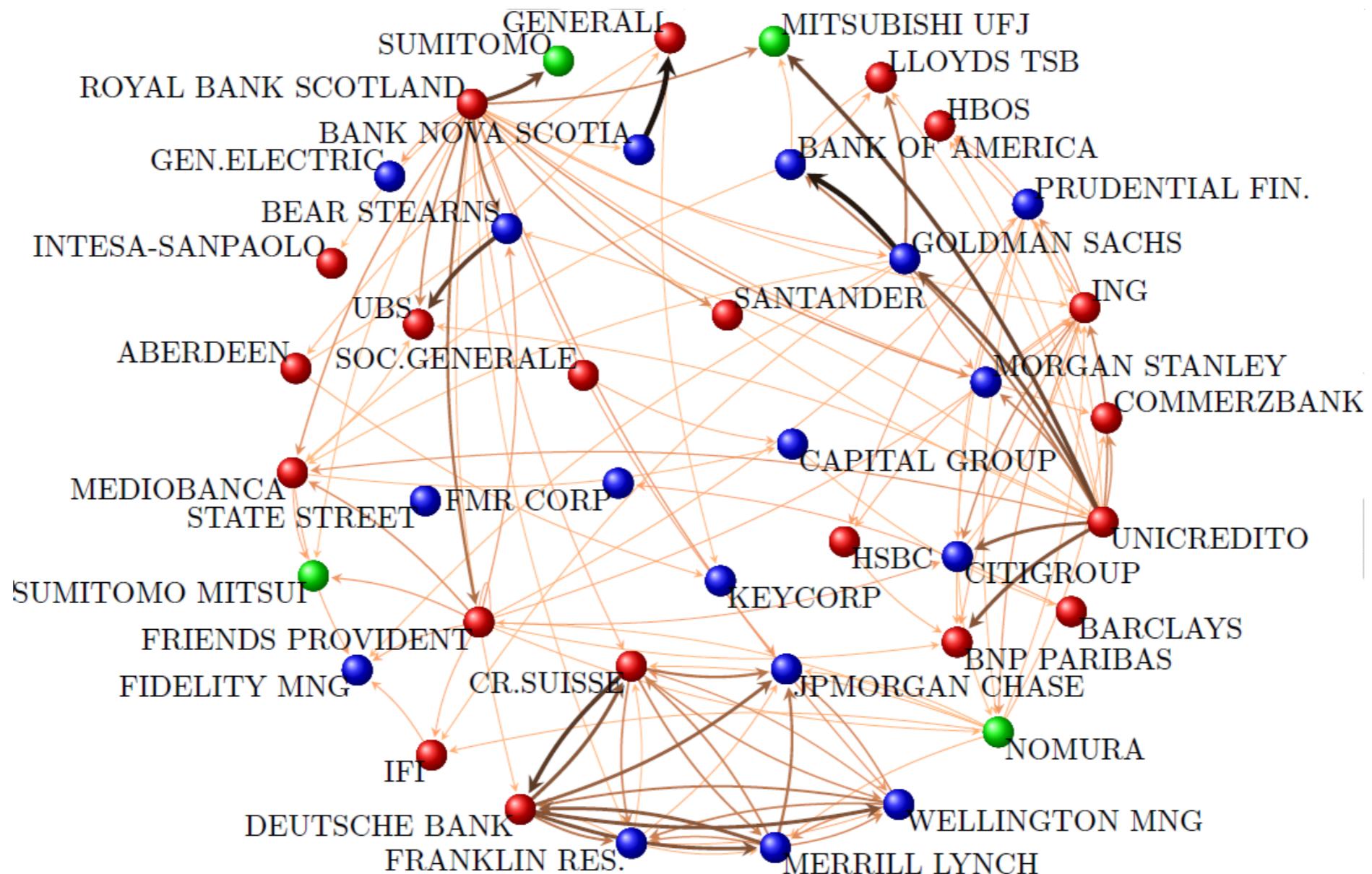
Political blogs



Source: <http://allthingsgraphed.com/2014/10/09/visualizing-political-polarization/>

Complex networks

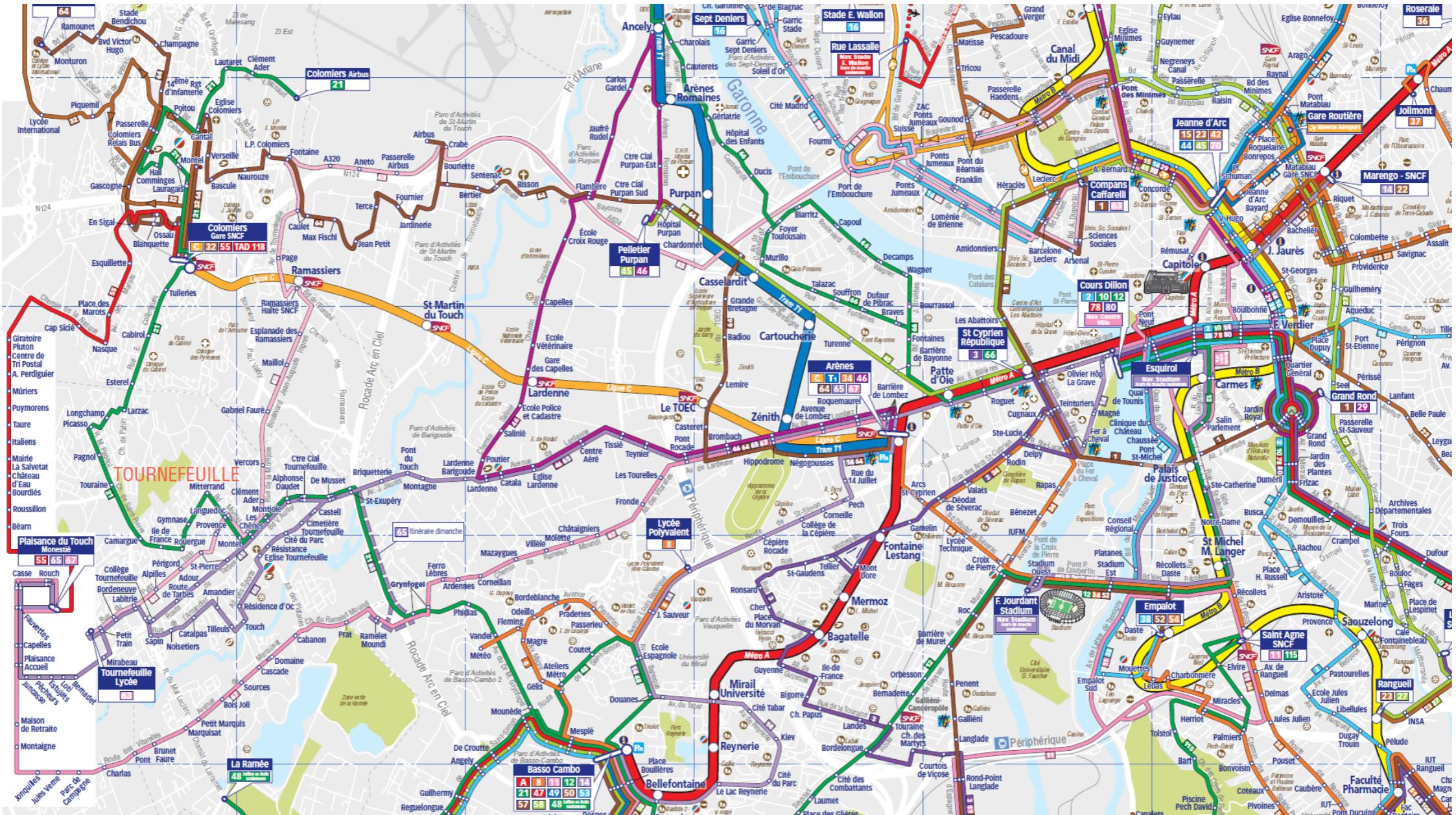
Finance



Source: <http://j-node.blogspot.fr/2011/10/network-of-global-corporate-control.html>

Complex networks

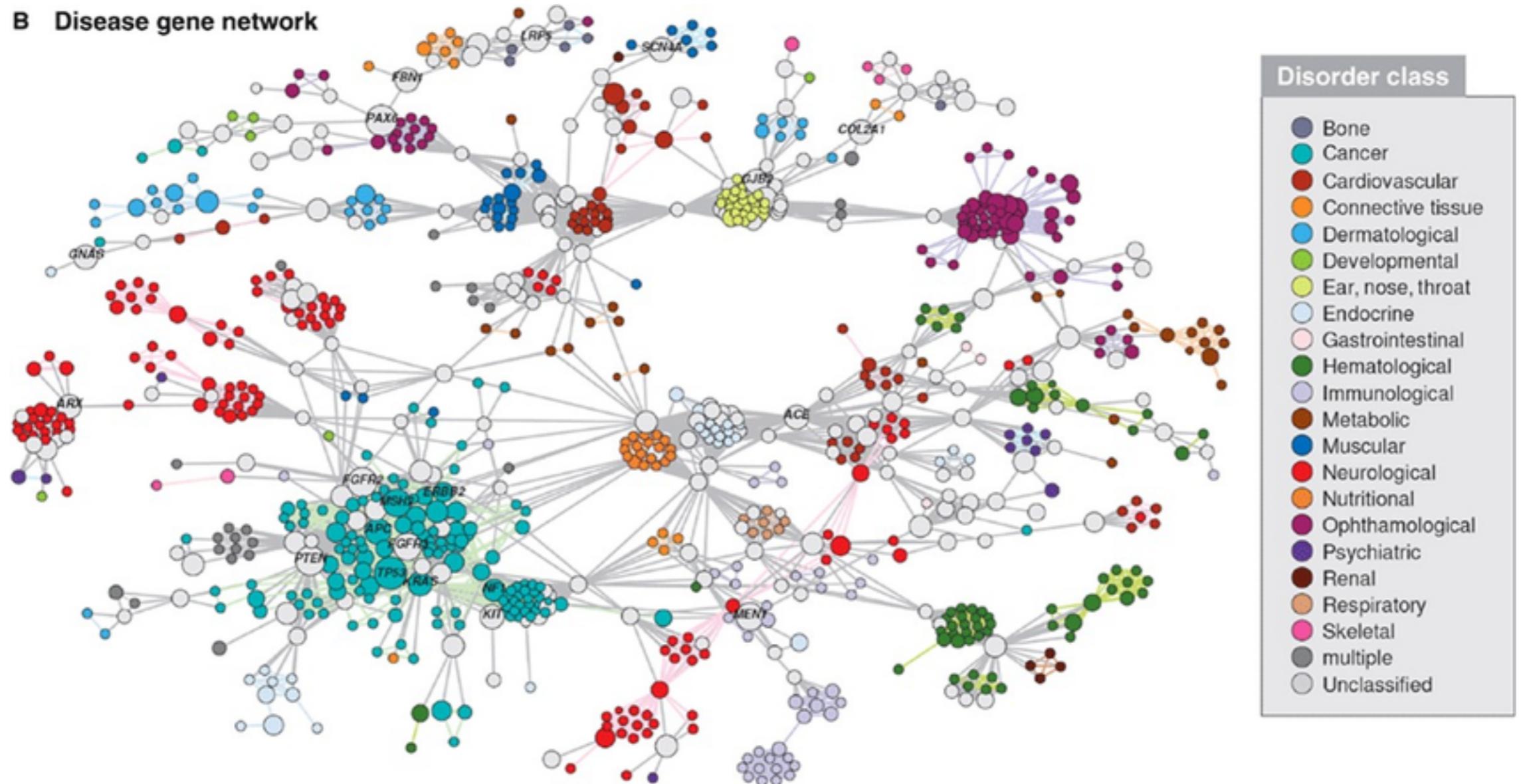
Transport



Source: Tisséo

Complex networks

Biology



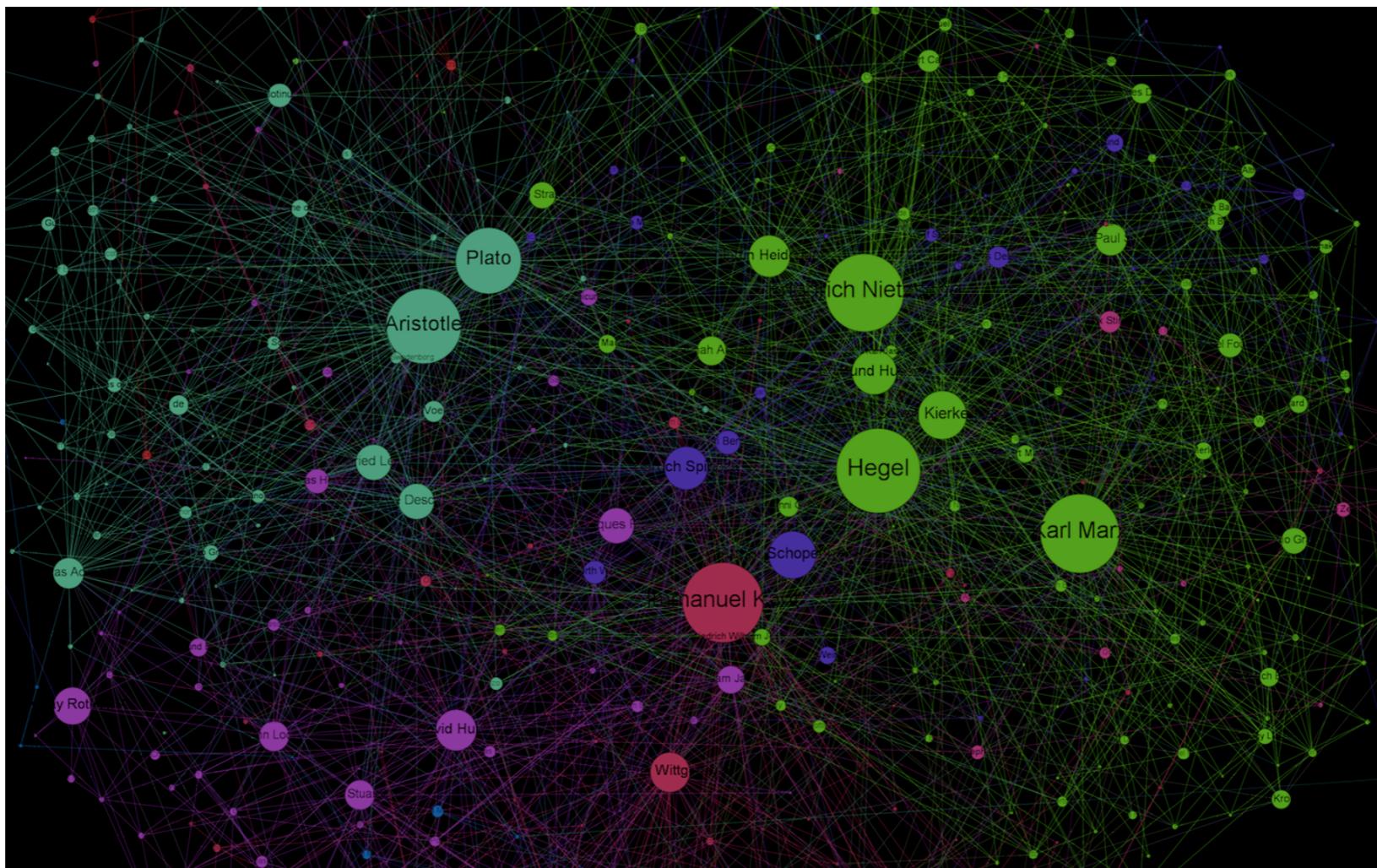
Source: Human disease classification in the postgenomic era: A complex systems approach to human pathobiology, 2007

Complex networks

Properties

- ▶ Complex network share common properties
 - ▶ Long-tail distribution (a.k.a. power law degree distribution)
 - ▶ Small world effect
 - ▶ Strong community structure

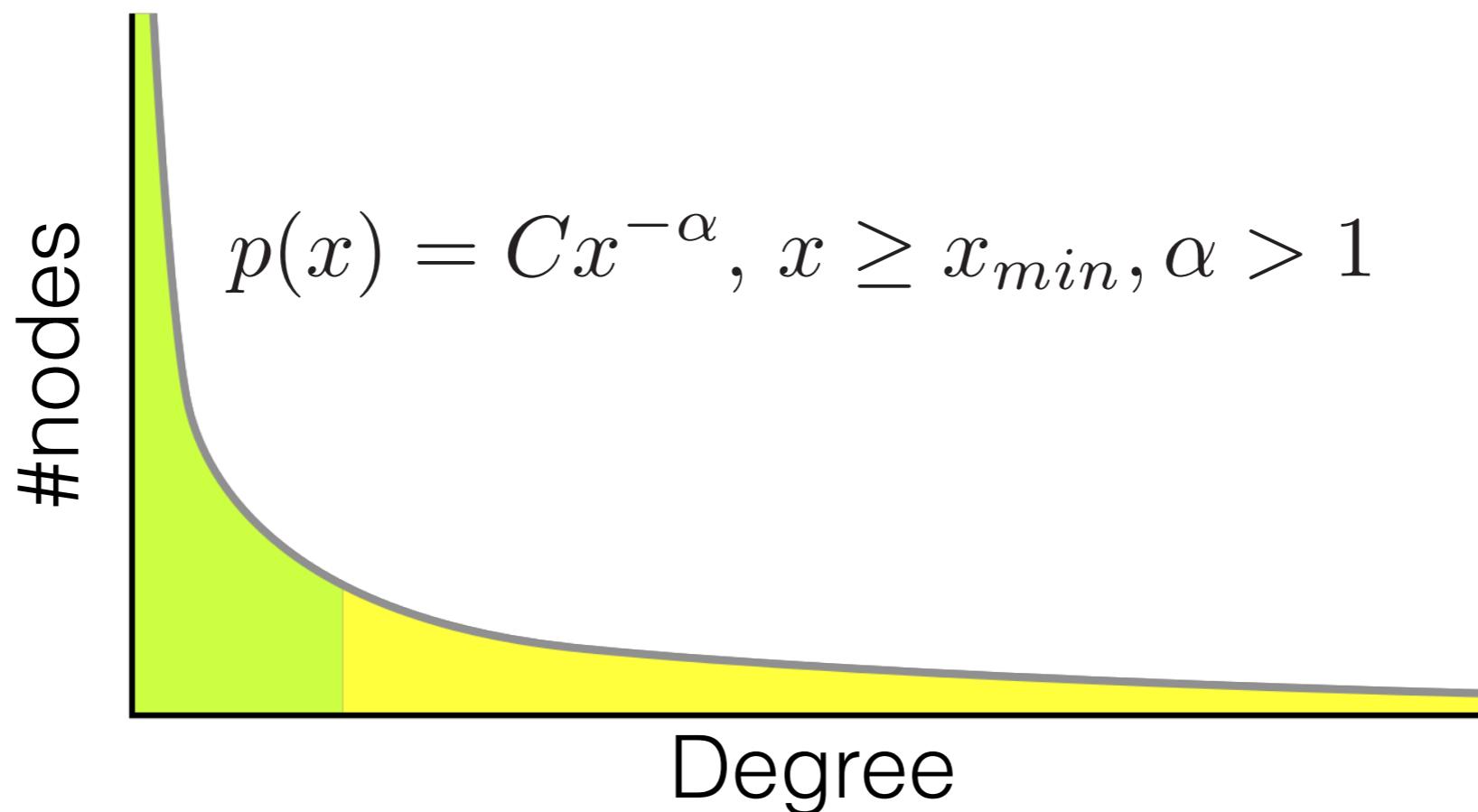
Long-tail distribution



- ▶ Few philosophers have many relations
- ▶ Many philosophers have few relations

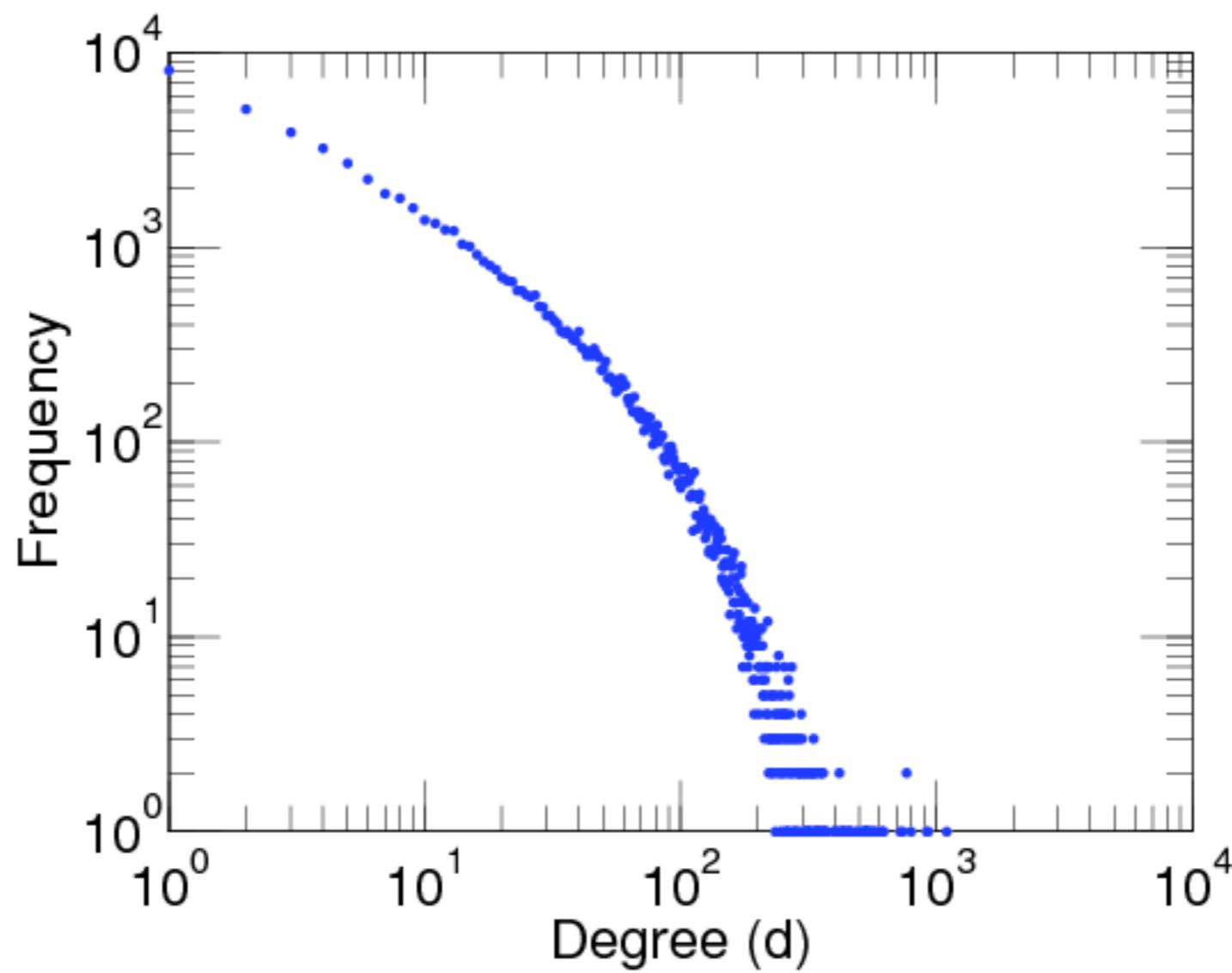
Long-tail distribution

- Degree distribution in large-scale networks often follows a power law (long-tail distribution or scale-free distribution)

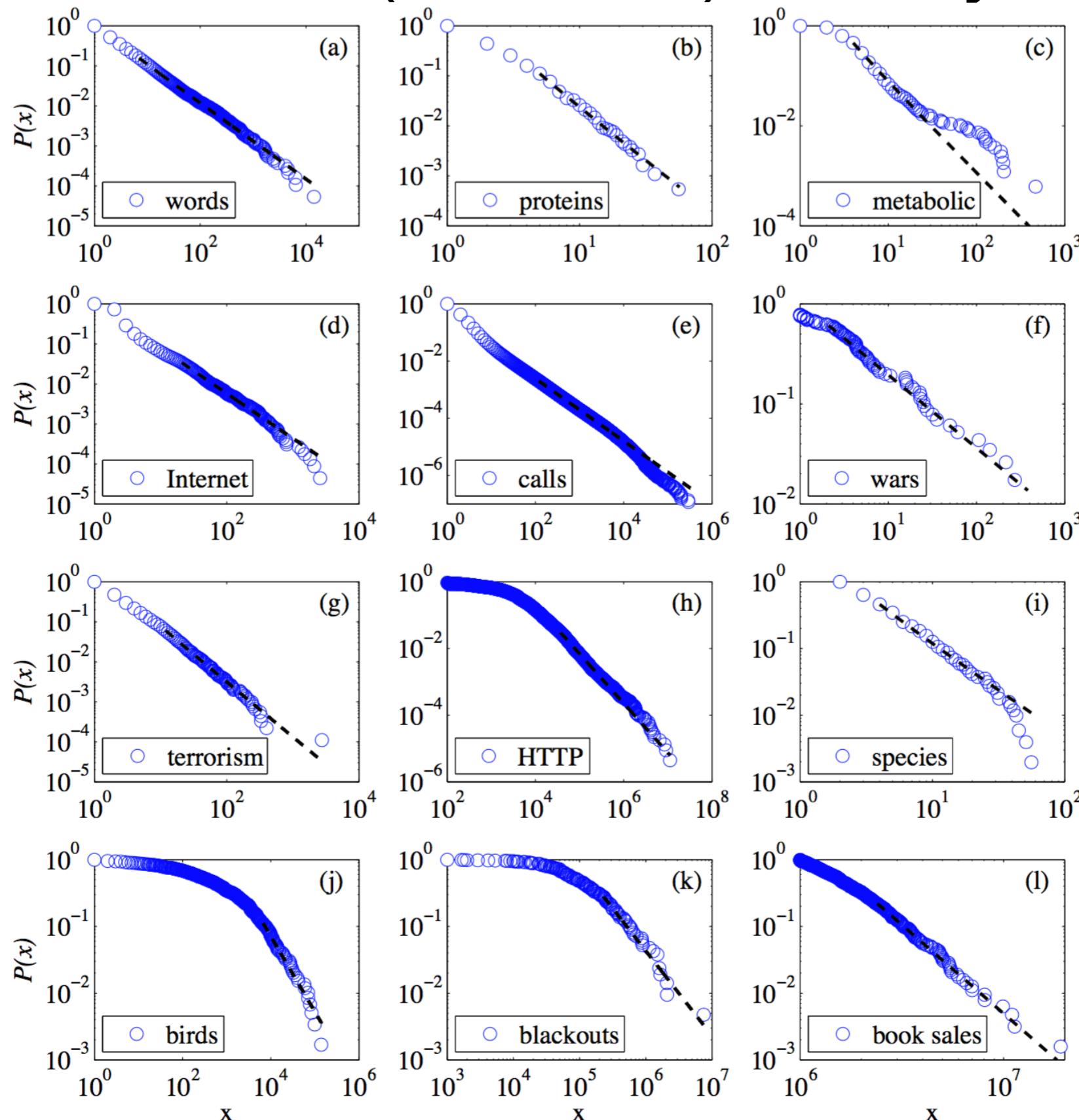


Long-tail distribution

- ▶ Long-tail distribution becomes a straight line if plot in log-log scale



Power-law is (almost) everywhere



Six degree of separation

- ▶ « *Any two people are on average separated no more than by six intermediate connections* »
- ▶ **A quite old topic**
 - ▶ *Chain-links* by Frigyes Karinthy in 1929 (short story)¹
 - ▶ John Guare play (1991)
 - ▶ Fred Schepisi movie (1993)

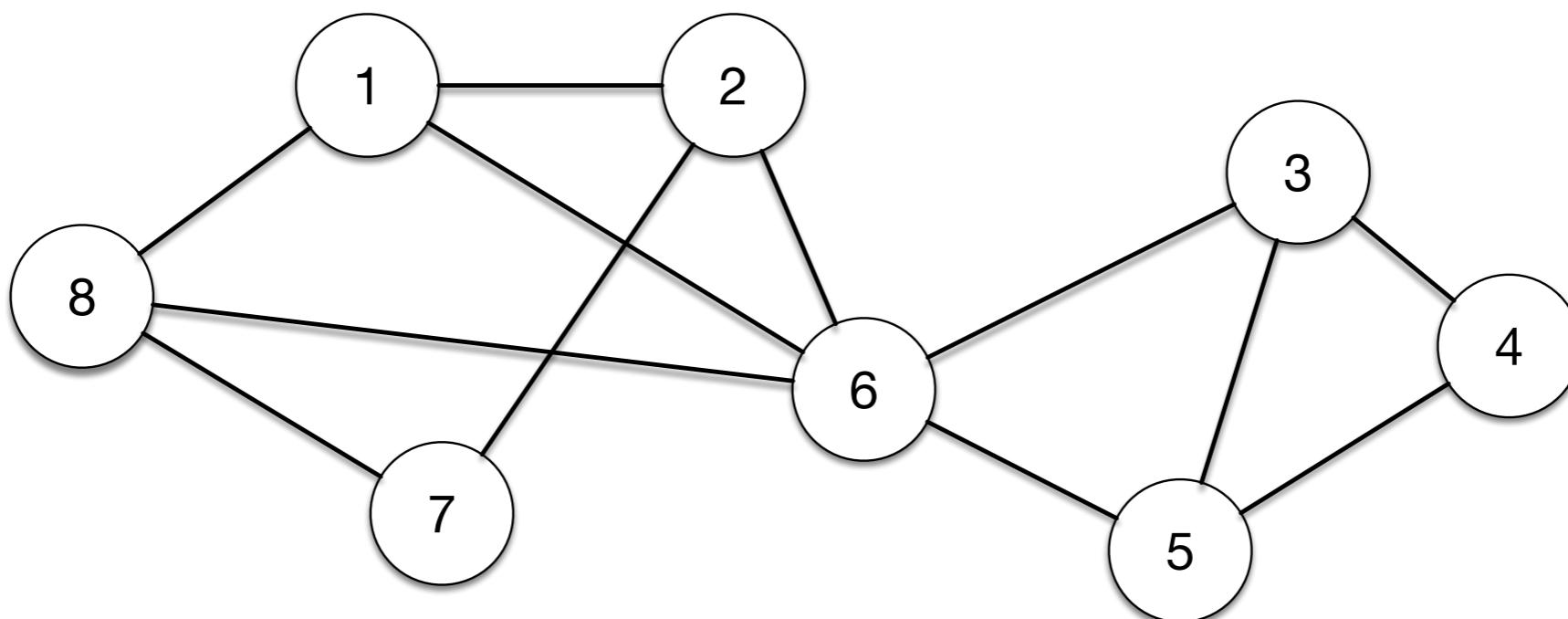
¹ An english translation is available at https://djjr-courses.wdfiles.com/local--files/soc180%3Akarinthy-chain-links/Karinthy-Chain-Links_1929.pdf

Small world effect

- ▶ Six Degrees of Separation
- ▶ A famous experiment conducted by Travers and Milgram in 1969
 - ▶ Subjects (296) were asked to send a chain letter to his/her acquaintance in order to reach a target person
 - ▶ The average path length is around **5.5**
- ▶ Verified on a planetary-scale instant messenger network of 180M users (Leskovec and Horvitz 2008)
 - ▶ The average path length is 6.6

Diameter

- ▶ Common measures to calibrate the small world effect:
 - ▶ Diameter
 - ▶ Average shortest path length



Community structure

- ▶ Community:
 - ▶ People (nodes) who interact with each other more frequently than with those outside the community
 - ▶ Can be seen as a partition of dense subgraph
- ▶ Many different definitions of what a community is
 - ▶ Dozens of scientific papers to detect communities

One possible intuition behind community

- ▶ Friends of a friend are likely to be friends as well
- ▶ Clustering coefficient: a very popular and widely used network metric
- ▶ Clustering coefficient measures the above intuition
 - ▶ Can be seen as the density of connections among one's friends

Local clustering coefficient

Directed

$$C_i = \frac{|\{e_{jk} : v_j, v_k \in N_i, e_{jk} \in E\}|}{k_i(k_i - 1)}$$

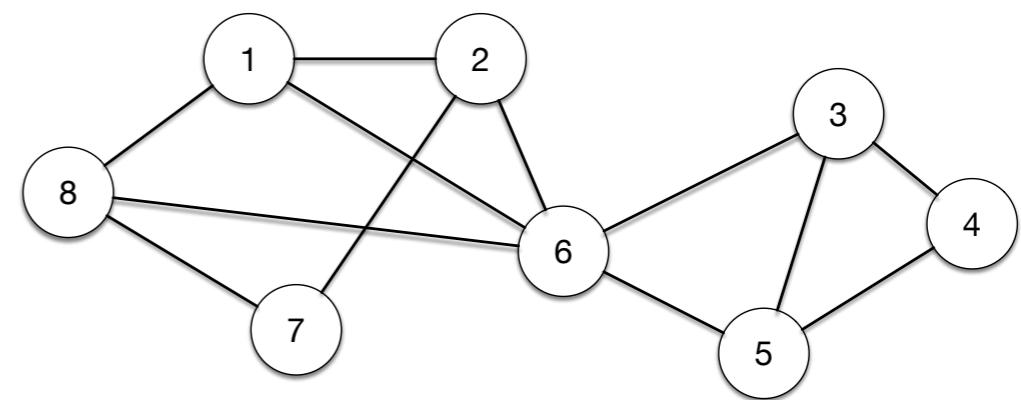
Undirected

$$C_i = \frac{2|\{e_{jk} : v_j, v_k \in N_i, e_{jk} \in E\}|}{k_i(k_i - 1)}$$

With e_{jk} , the edge (if exists) between vertices v_k and v_k ,
 N_i the neighborhood of vertex v_i and $k_i = |N_i|$

**Average Clustering Coefficient
(a.k.a. transitivity)**

$$\bar{C} = \frac{\sum_i C_i}{|V|}$$



C_1 ?

Average clustering coefficient

A graph has a strong community structure if its average clustering coefficient is much higher than the cluster coefficient of a random graph having the same number of vertices and a similar number of edges

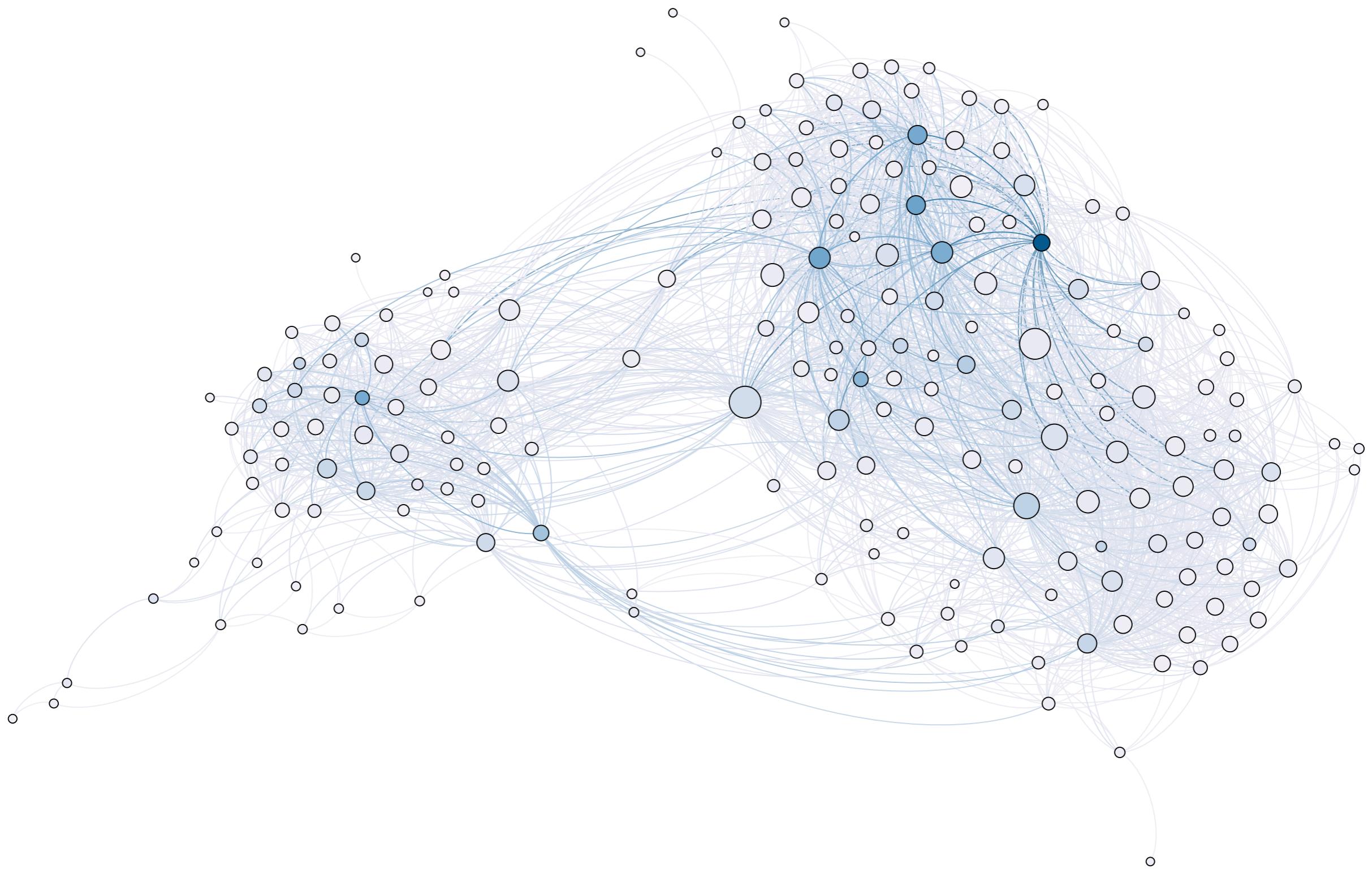
A graph having an average clustering coefficient greater than 0.15 is usually considered as a graph having a strong community structure

Outline

3. Node-centric metrics

- A. Centrality
- B. Prestige
- C. Page Rank
- D. Hubs and Authority
- E. Metrics comparison

Importance of Nodes



Importance of Nodes

Centrality and Prestige

- Not all nodes are equally important
- An actor is prominent if the ties of the actor make the actor particularly visible to the other actors in the network
- *Visibility is not only measured by direct ties, but also by indirect ties through intermediaries*
- Knock and Burt distinguished two types of visibility: centrality and prestige
- **A** is the adjacency matrix of the network

Centrality and Prestige

- ▶ Sociologic point of view
- ▶ Both measures rely on the importance of actors in the social network
- ▶ Actor centrality
 - involvement with other actors, many ties, source or recipient
 - Mostly over undirected networks
- ▶ Actor prestige
 - Object of many ties, ties directed to an actor
 - Only over directed networks

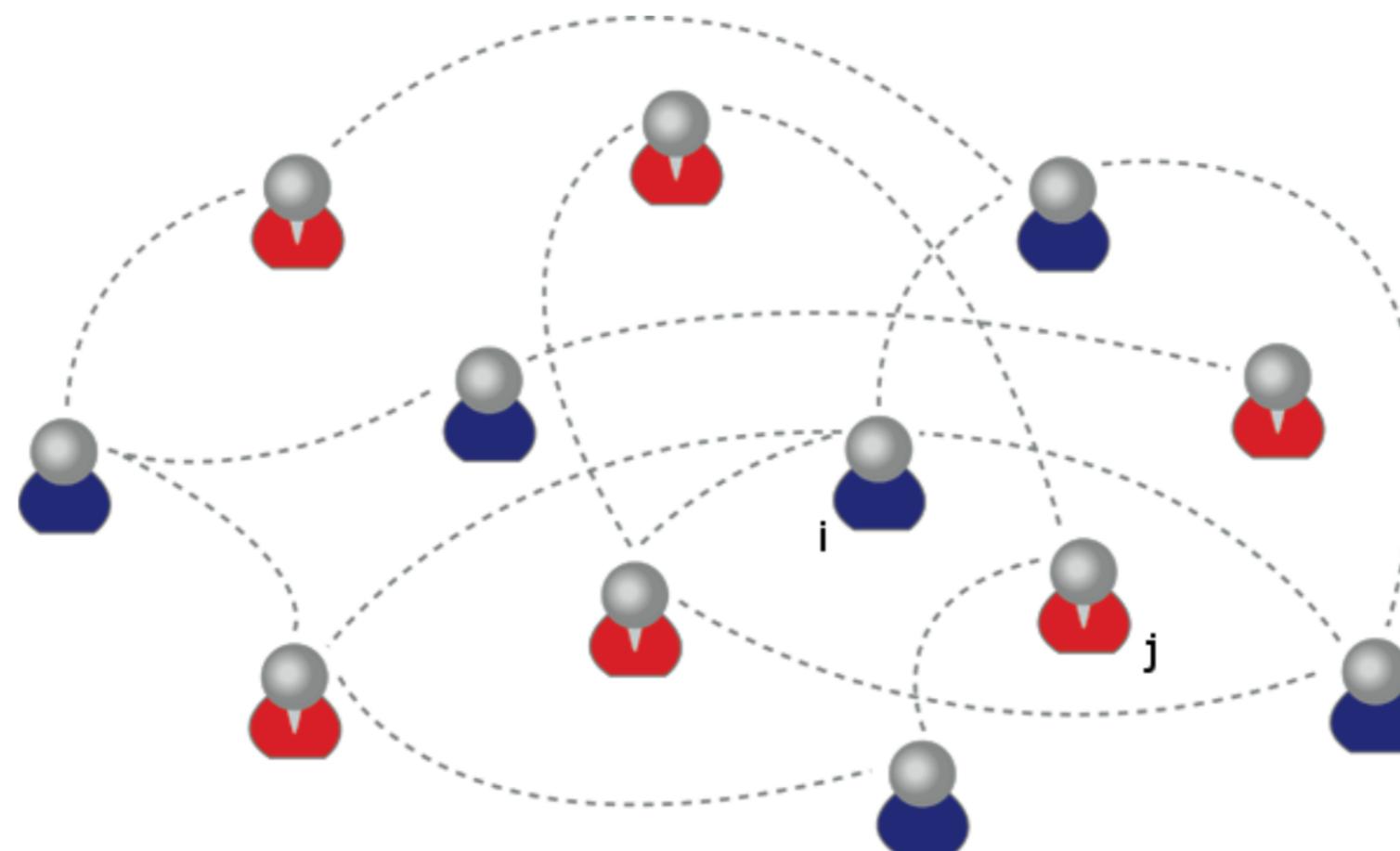
Centrality

- Not concerned with whether prominence is due to the receiving or the transmission of many ties
- What is important is that the actor is involved
- In undirected graphs, a central actor is involved in many ties.
- ***Commonly used prestige metrics***
 - Degree
 - Closeness
 - Betweenness
 - Eigenvector / Katz

Centrality

Intuition

The more connected a user, the more important



Centrality

Degree - Undirected

Degree centrality: number of nearest neighbours

$$C_D(i) = \sum_j A_{ij} = \sum_j A_{ji}$$

Normalized degree centrality

$$C_D^*(i) = \frac{1}{n-1} C_D(i)$$

n is the number of nodes in G

Interpretation

- High value: direct contact with many other actors
- Low value: not active, peripheral actor

Centrality

Degree - Directed

Degree centrality: number of nearest neighbours

$$C_D(i) = \sum_j A_{ij}$$

Normalized degree centrality

$$C_D^*(i) = \frac{1}{n-1} C_D(i)$$

n is the number of nodes in G

Interpretation

- High value: direct contact with many other actors
- Low value: not active, peripheral actor

Centrality

Closeness

Closeness centrality: how close an actor to all other actors in network

$$C_C(i) = \frac{1}{\sum_j d(i, j)}$$

Where $d(i, j)$ is the length of the shortest path between vertices v_i and v_j

Normalized closeness centrality

$$C_C^*(i) = (n - 1)C_C(i)$$

Can be adapted to the directed scenario taking direction of edges into account

Interpretation

- ▶ High value: actor in the center can quickly interact with all others, short communication path to others, minimal number of steps to reach others

Centrality

Betweenness

Measures how much a node has control over all possible pairs of nodes. If it is part of the shortest paths of many pairs, it will be central

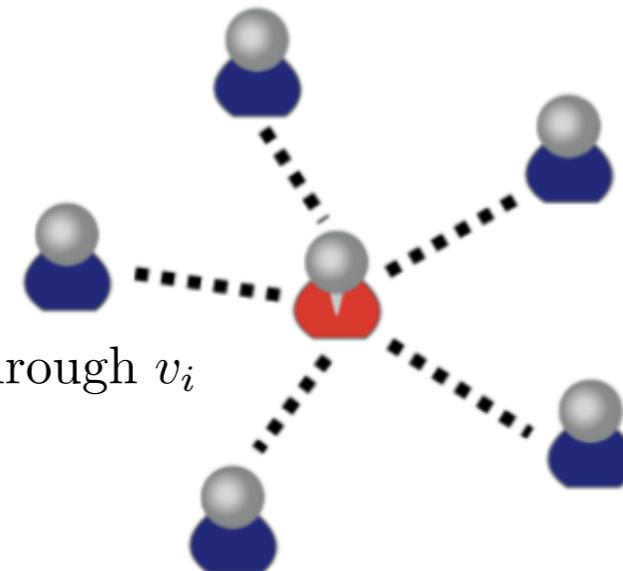
Undirected

$$C_B(i) = \sum_{i \neq j \neq k} \frac{p_{jk}(i)}{p_{jk}}$$

p_{jk} is the number of shortest paths between v_j and v_k

$p_{jk}(i)$ is the number of shortest paths between v_j and v_k passing through v_i

Can be normalized by $(n-1)(n-2)/2$



Directed

Same as for undirected

Normalized by $(n-1)(n-2)$

Centrality

Eigenvector

Eigenvector centrality: the importance of a node depends on the importance of its neighbours (recursive definition)

$$C_E(i) = \frac{1}{\lambda} \sum_{j \neq i} A_{i,j} C_E(j)$$

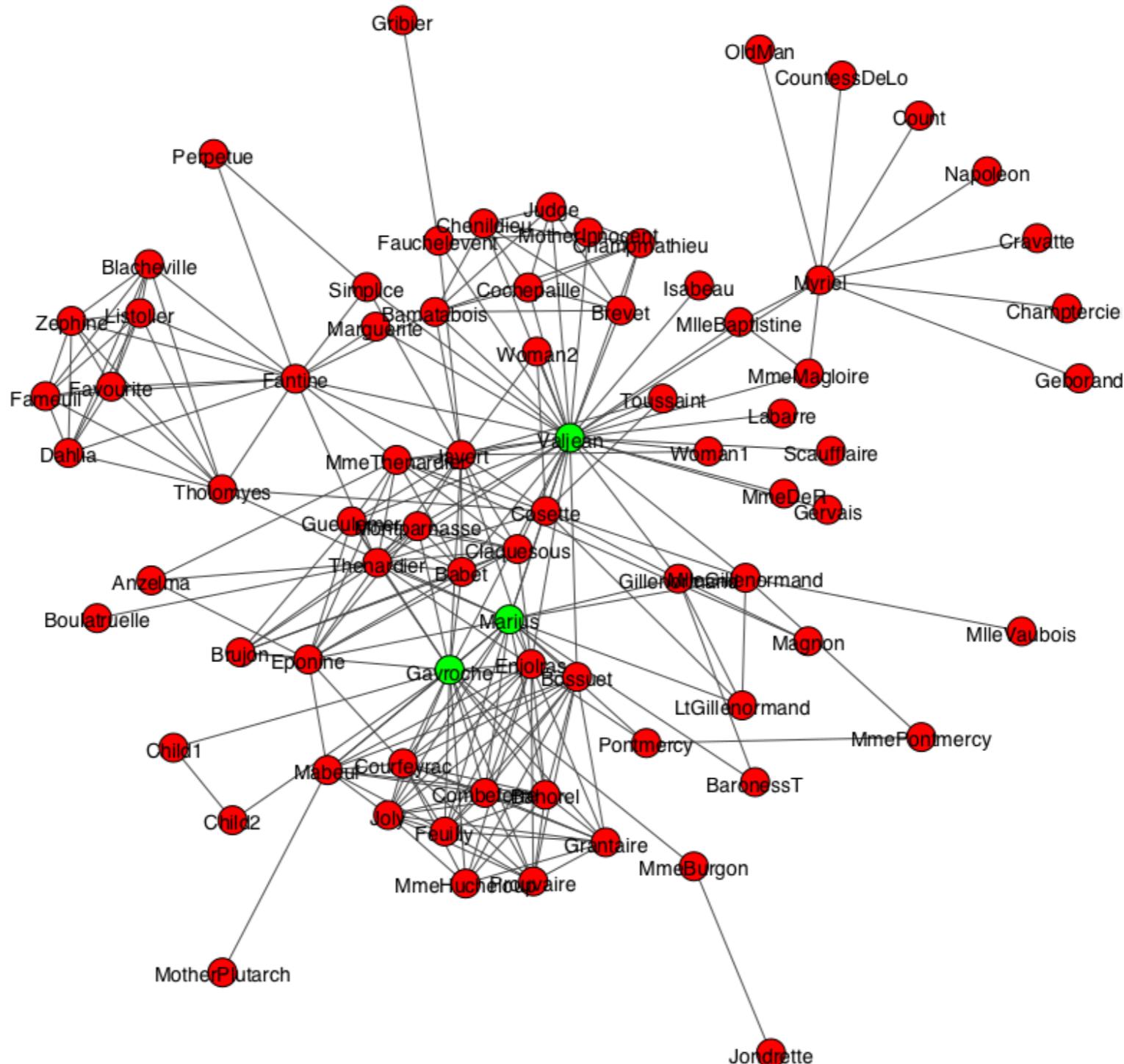
With the vector notation: $\mathbf{Ax} = \lambda \mathbf{x}$

Additional requirement : the eigenvector is non negative

Consequence (Perron-Frobenius theorem) : only the greatest eigenvalue works

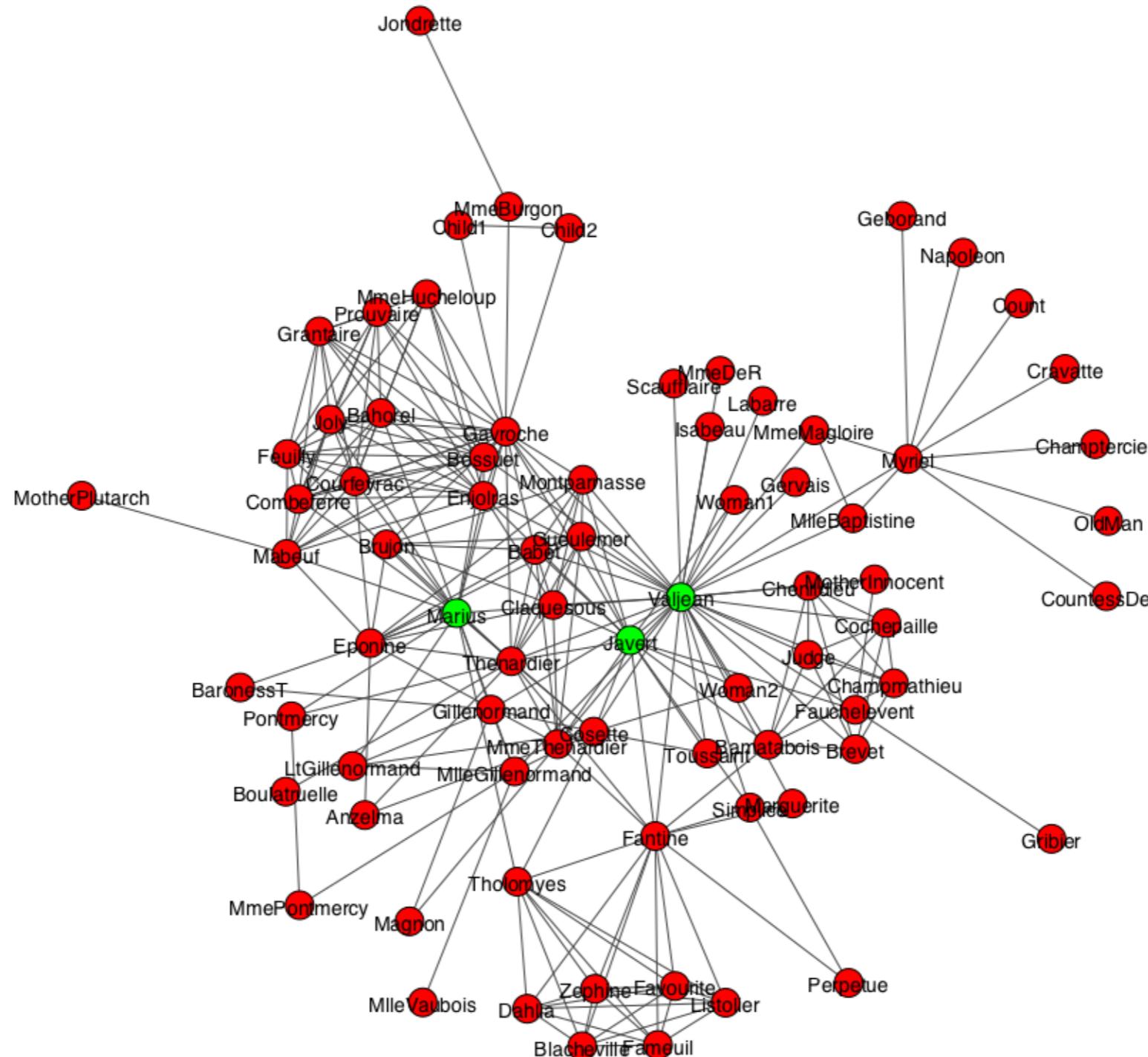
Centrality

Examples - Degree



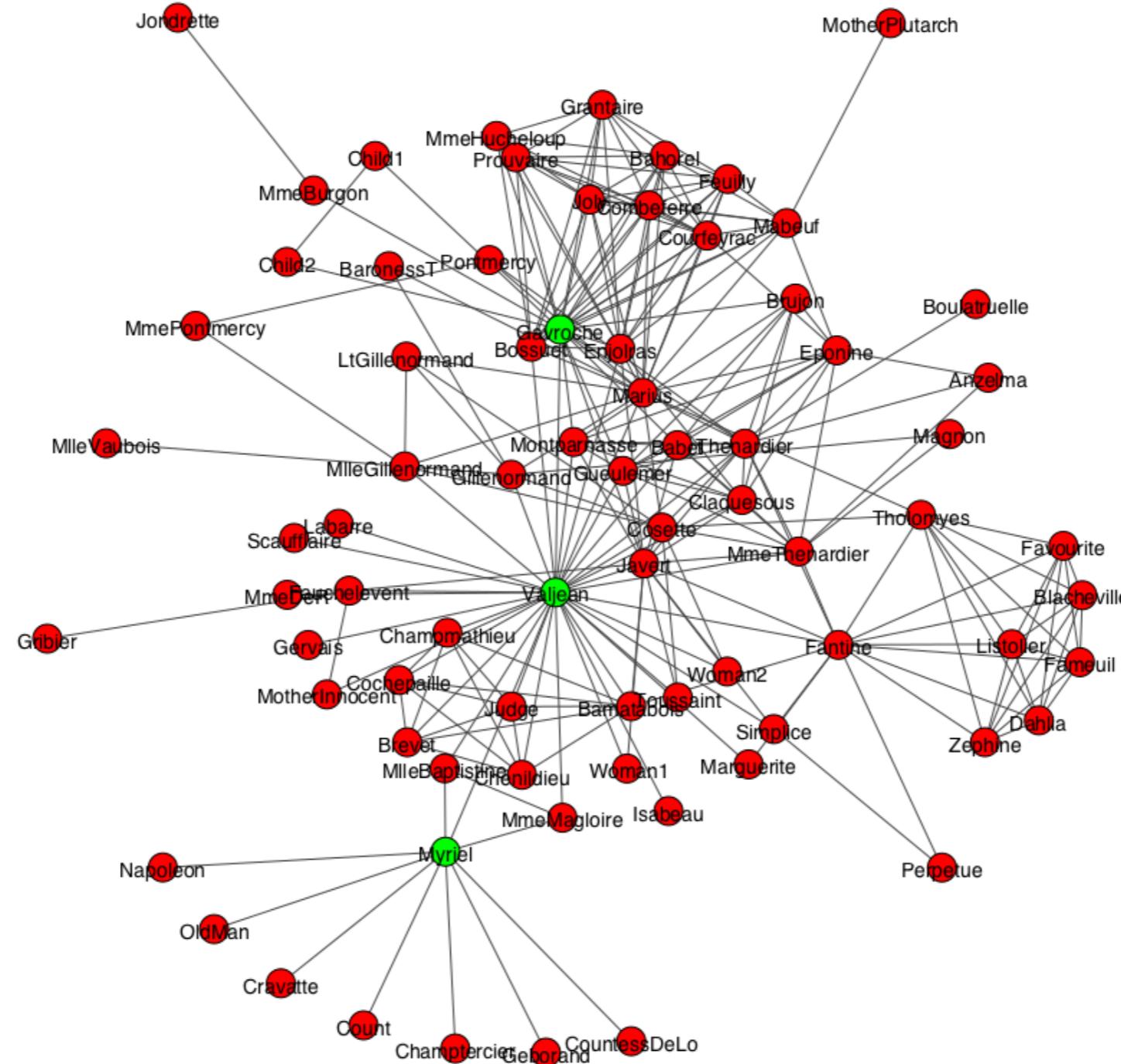
Centrality

Examples - Closeness



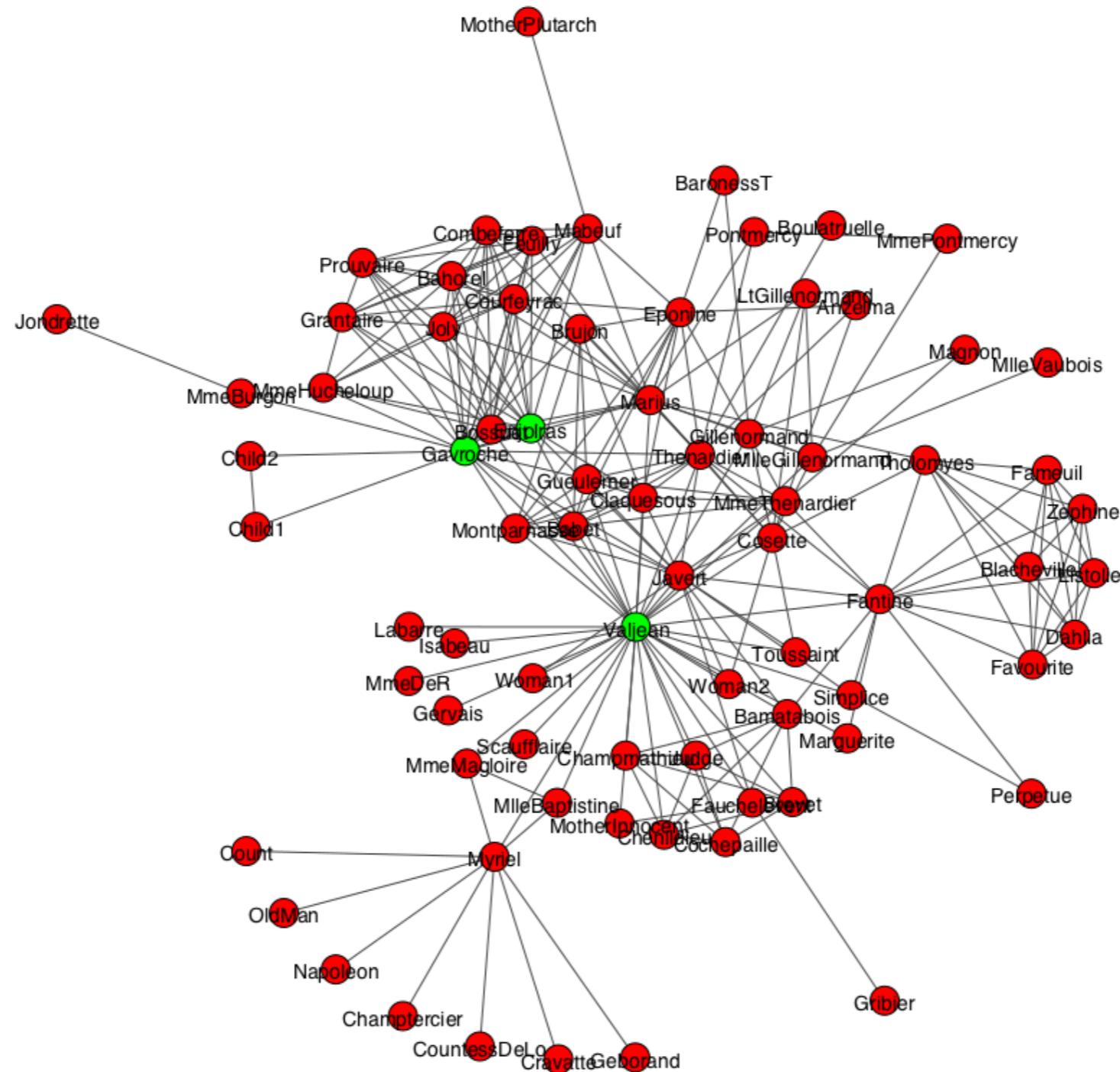
Centrality

Examples - Betweenness

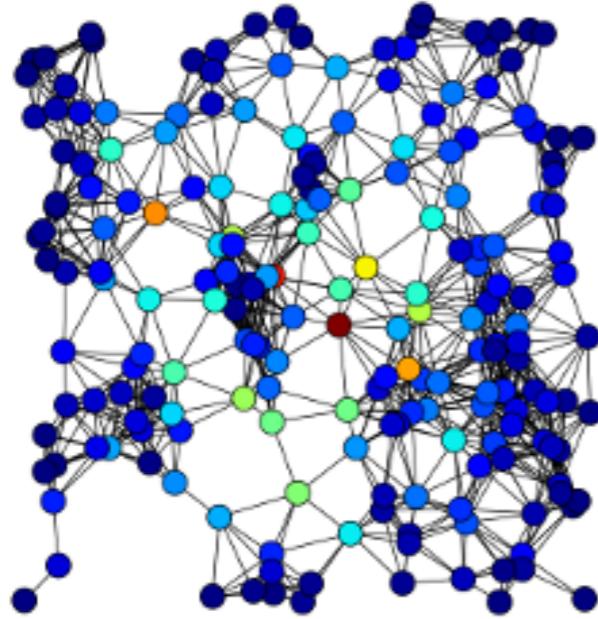


Centrality

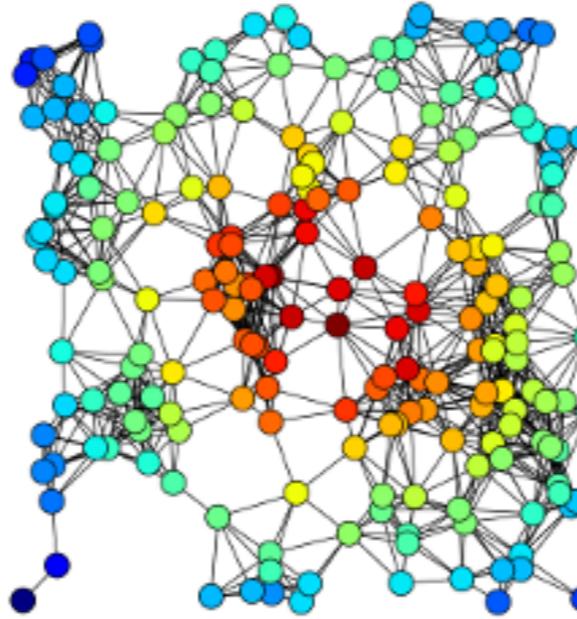
Examples - Eigenvector



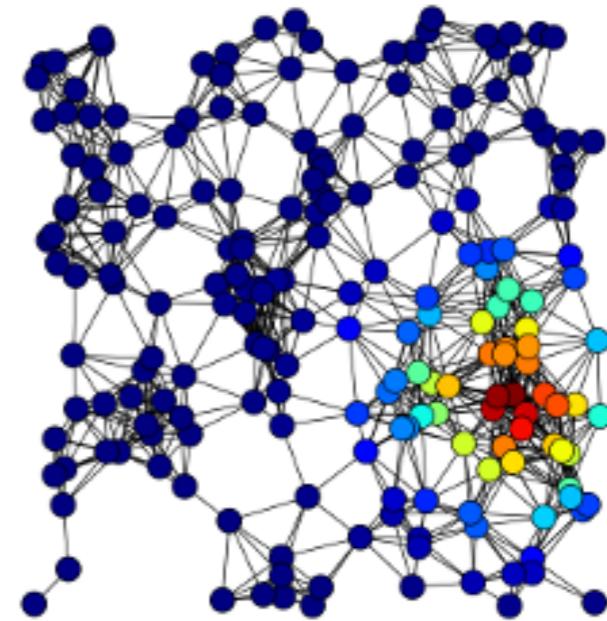
Centrality Examples



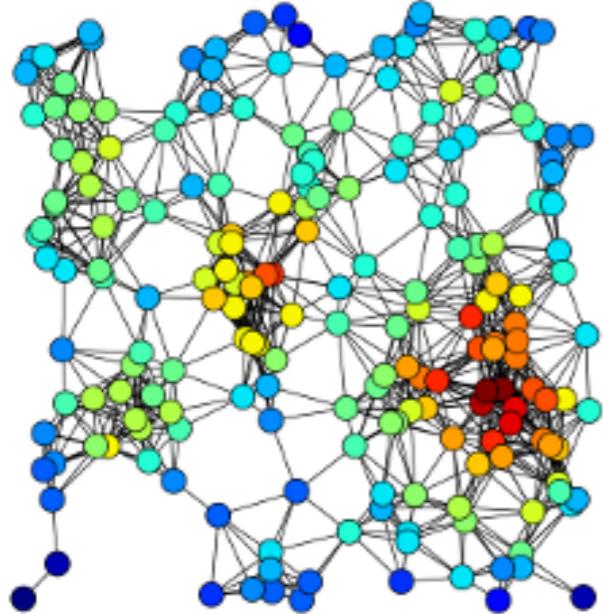
A



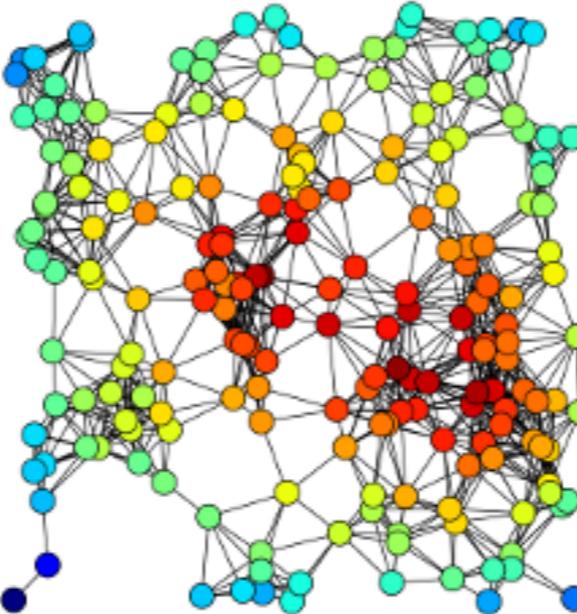
B



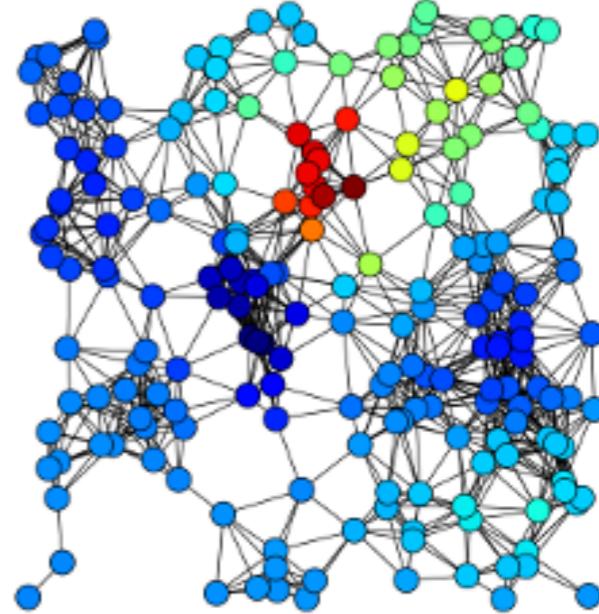
C



D



E



F

Source: Wikipédia

- A) Betweenness centrality, B) Closeness centrality, C) Eigenvector centrality, D) Degree centrality, E) Harmonic Centrality and F) Katz centrality

Take away message

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

	Low Degree	Low Closeness	Low Betweenness
High Degree		Embedded in cluster that is far from the rest of the network	Ego's connections are redundant - communication bypasses him/her
High Closeness	Key player tied to important/active players		Probably multiple paths in the network, ego is near many people, but so are many others
High Betweenness	Ego's few ties are crucial for network flow	Very rare cell. Would mean that ego monopolizes the ties from a small number of people to many others.	

Prestige

Intuition

More subtle than centrality

- ▶ Distinguish incoming and outgoing links
- ▶ A prestigious node is often referenced
- ▶ Only incoming links are considered
- ▶ Only directed graphs

Commonly used prestige metrics

- ▶ Degree
- ▶ Closeness
- ▶ Rank

Prestige

Degree

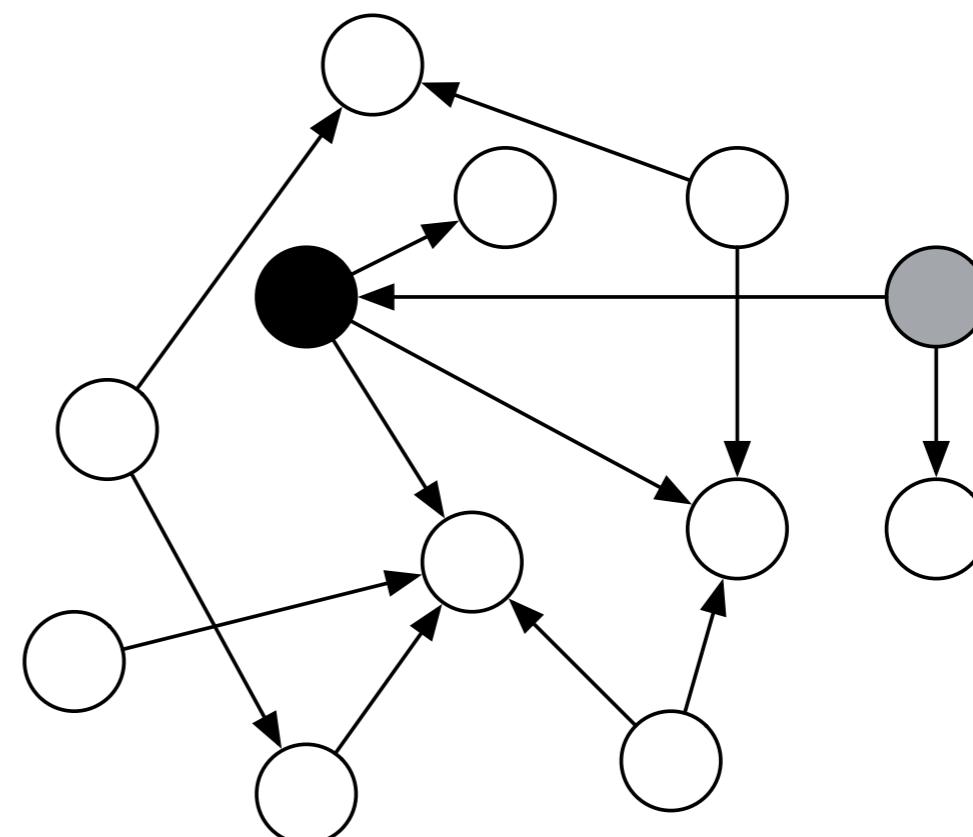
Definition

$$P_D(i) = \frac{d_I(i)}{n - 1}$$

$d_I(i)$ is the incoming degree of v_i

Prestige of ?

Prestige of ?



Prestige

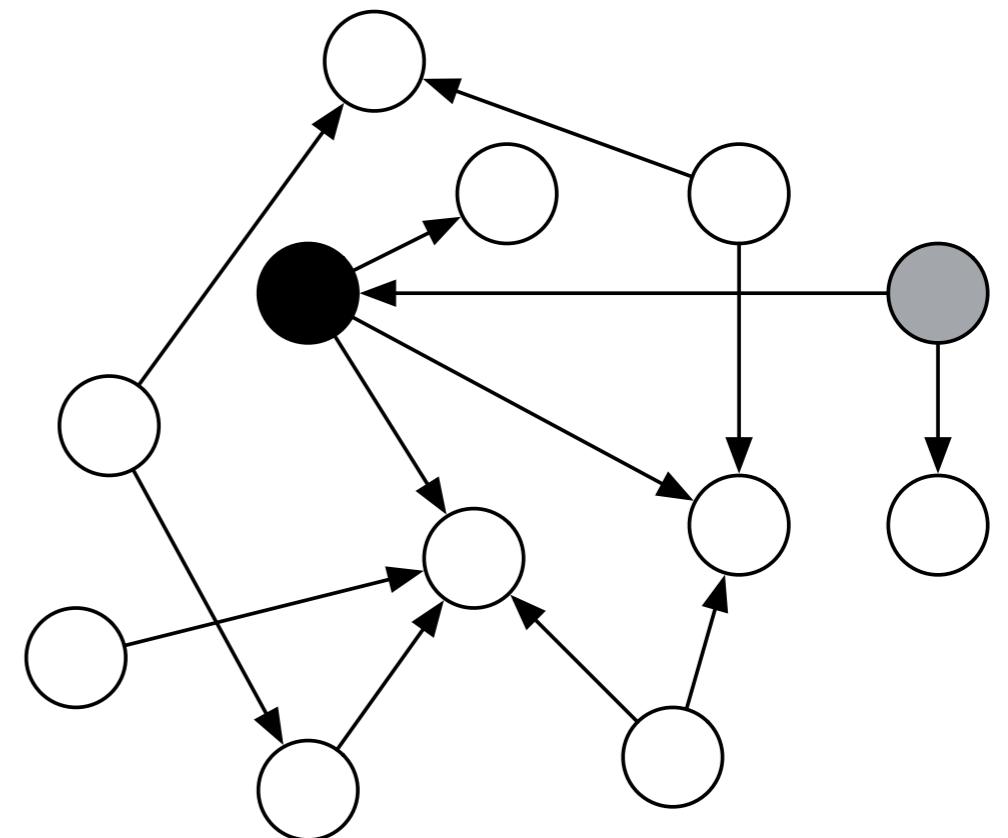
Closeness

Intuition

Paths going to the node are considered

Definition

$$P_B(i) = \frac{\sum_{j \in I_i} d(j, i)}{|I_i|}$$



I_i is the set of vertices v_x s.t. it exists a path between v_x and v_i

$d(j, i)$ is the length of the shortest path from v_j to v_i

Prestige

Rank

Intuition

The prestige of a node depends on the importance of its predecessors.

Definition

$$P_R(i) = \sum_{j \neq i} A_{ji} P_R(j)$$

Matrix representation

$$\mathbf{P} = \mathbf{A}^T \mathbf{P}$$

Page Rank (1)

Intuition

"PageRank can be thought of as a model of user behavior. We assume there is a "random surfer" who is given a web page at random and keeps clicking on links, never hitting "back" but eventually gets bored and starts on another random page. The probability that the random surfer visits a page is its PageRank."

Model

Create a stochastic matrix of the Web, denoted by **S**:

- Each page i corresponds to row i and column i of the matrix
- If page j has n successors (links) then the ij th cell of the matrix is equal to $1/n$ if page i is one of these n successors of page j , and 0 otherwise.

Page Rank (2)

- ▶ Initially each page has 1 unit of importance. At each round, each page shares importance it has among its successors, and receives new importance from its predecessors
- ▶ The importance of each page reaches a limit after some steps
- ▶ Random walk on a directed graph

$$\mathbf{p}^{t+1} = \mathbf{S}\mathbf{p}^t$$

\mathbf{p}^t is the Page Rank vector at iteration t

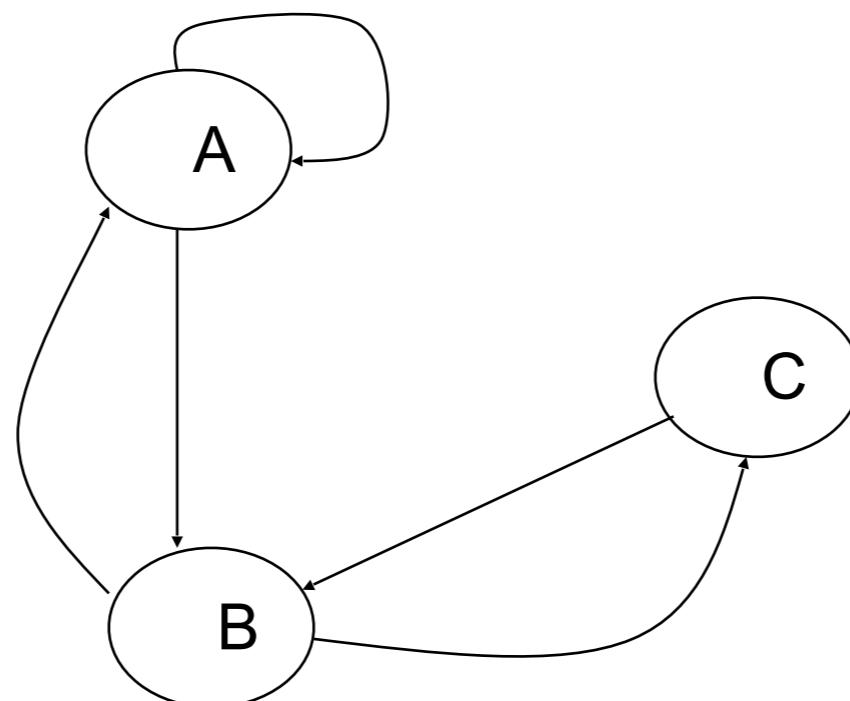
- ▶ Can also be seen as a Markov chain

Page Rank (3)

Example 1

Assume that the Web consists of only three pages (A, B, and C). The links among these pages are shown below.

Let $P=[a, b, c]$ be the vector of importances for these three pages



		S		
		A	B	C
Source	A	1/2	1/2	0
	B	1/2	0	1
C	0	1/2	0	

Page Rank (3)

Example 1

The equation describing the asymptotic values of these three variables is:

$$\begin{bmatrix} a \\ b \\ c \end{bmatrix} = \begin{bmatrix} 1/2 & 1/2 & 0 \\ 1/2 & 0 & 1 \\ 0 & 1/2 & 0 \end{bmatrix} \times \begin{bmatrix} a \\ b \\ c \end{bmatrix}$$

We can solve the equations like this one by starting with the assumption $a = b = c = 1$, and applying the matrix to the current estimate of these values repeatedly. The first four iterations give the following estimates:

$$\mathbf{a} = 1 \quad 1 \quad 5/4 \quad 9/8 \quad 5/4 \quad \dots \quad \mathbf{6/5}$$

$$\mathbf{b} = 1 \quad 3/2 \quad 1 \quad 11/8 \quad 17/16 \quad \dots \quad \mathbf{6/5}$$

$$\mathbf{c} = 1 \quad 1/2 \quad 3/4 \quad 1/2 \quad 11/16 \quad \dots \quad \mathbf{3/5}$$

Problems with Real Web Graphs

In the limit, the solution is $a=b=6/5$, $c=3/5$. That is, a and b each have the same importance, and twice of c .

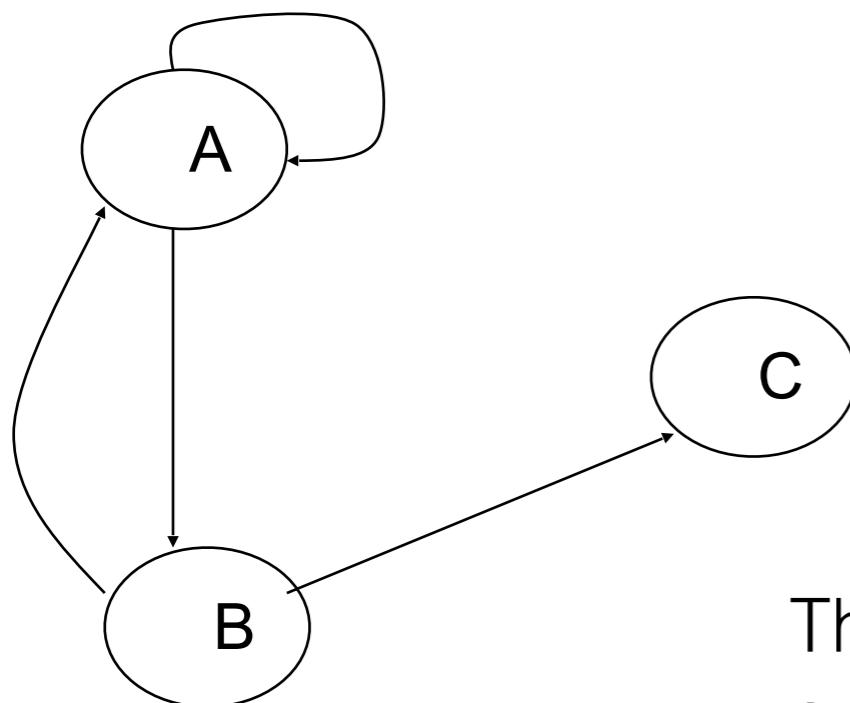
Problems with Real Web Graphs

- **Dead ends:** a page that has no successors has nowhere to send its importance.
- **Spider traps:** a group of one or more pages that have no links out.

Page Rank

Example 2

Assume now that the structure of the Web has changed. The new matrix describing transitions is:



$$\begin{bmatrix} a \\ b \\ c \end{bmatrix} = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} & 0 \\ \frac{1}{2} & 0 & 0 \\ 0 & \frac{1}{2} & 0 \end{bmatrix} \times \begin{bmatrix} a \\ b \\ c \end{bmatrix}$$

The first four steps of the iterative solution are:

$$a = 1 \quad 1 \quad 3/4 \quad 5/8 \quad 1/2$$

$$b = 1 \quad 1/2 \quad 1/2 \quad 3/8 \quad 5/16$$

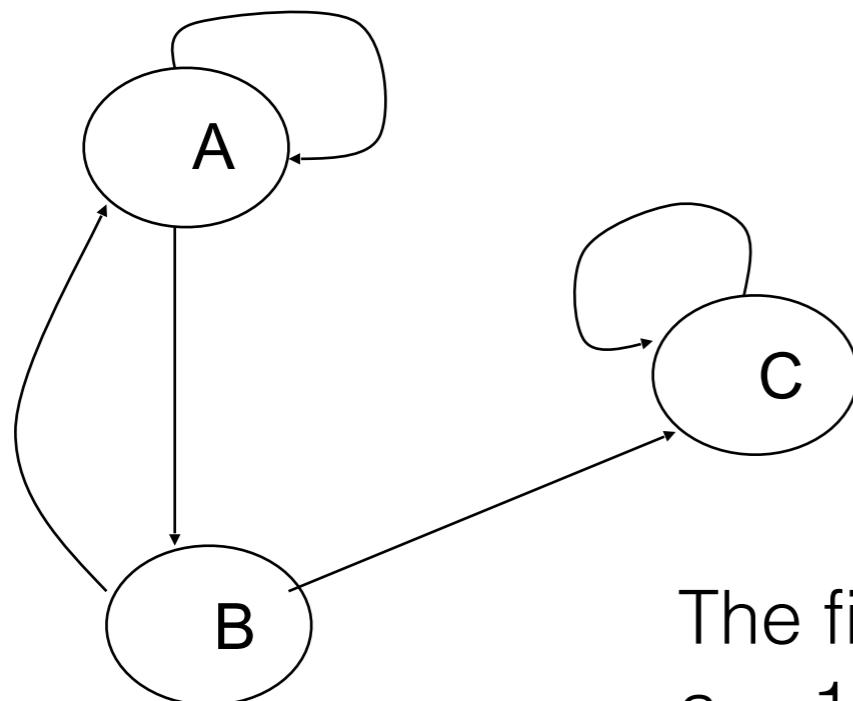
$$c = 1 \quad 1/2 \quad 1/4 \quad 1/4 \quad 3/16$$

Eventually, each of a, b, and c become 0.

Page Rank

Example 3

Assume now once more that the structure of the Web has changed. The new matrix describing transitions is:



$$\begin{bmatrix} a \\ b \\ c \end{bmatrix} = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} & 0 \\ \frac{1}{2} & 0 & 0 \\ 0 & \frac{1}{2} & 1 \end{bmatrix} \times \begin{bmatrix} a \\ b \\ c \end{bmatrix}$$

The first four steps of the iterative solution are:

$$a = 1 \quad 1 \quad 3/4 \quad 5/8 \quad 1/2$$

$$b = 1 \quad 1/2 \quad 1/2 \quad 3/8 \quad 5/16$$

$$c = 1 \quad 3/2 \quad 7/4 \quad 2 \quad 35/16$$

c converges to 3, and a=b=0.

Google Solution

Instead of applying the matrix directly, “**tax**” each page **some fraction of its current importance**, and distribute the taxed importance equally among all pages.

Example: if we use 20% tax, the equation of the previous example becomes:

$$a = 0.8 * (\frac{1}{2} * a + \frac{1}{2} * b + 0 * c)$$

$$b = 0.8 * (\frac{1}{2} * a + 0 * b + 0 * c)$$

$$c = 0.8 * (0 * a + \frac{1}{2} * b + 1 * c)$$

The solution to this equation is $a=7/11$, $b=5/11$, and $c=21/11$

HITS Algorithm

Hyperlink-Induced Topic Search

Key Definitions

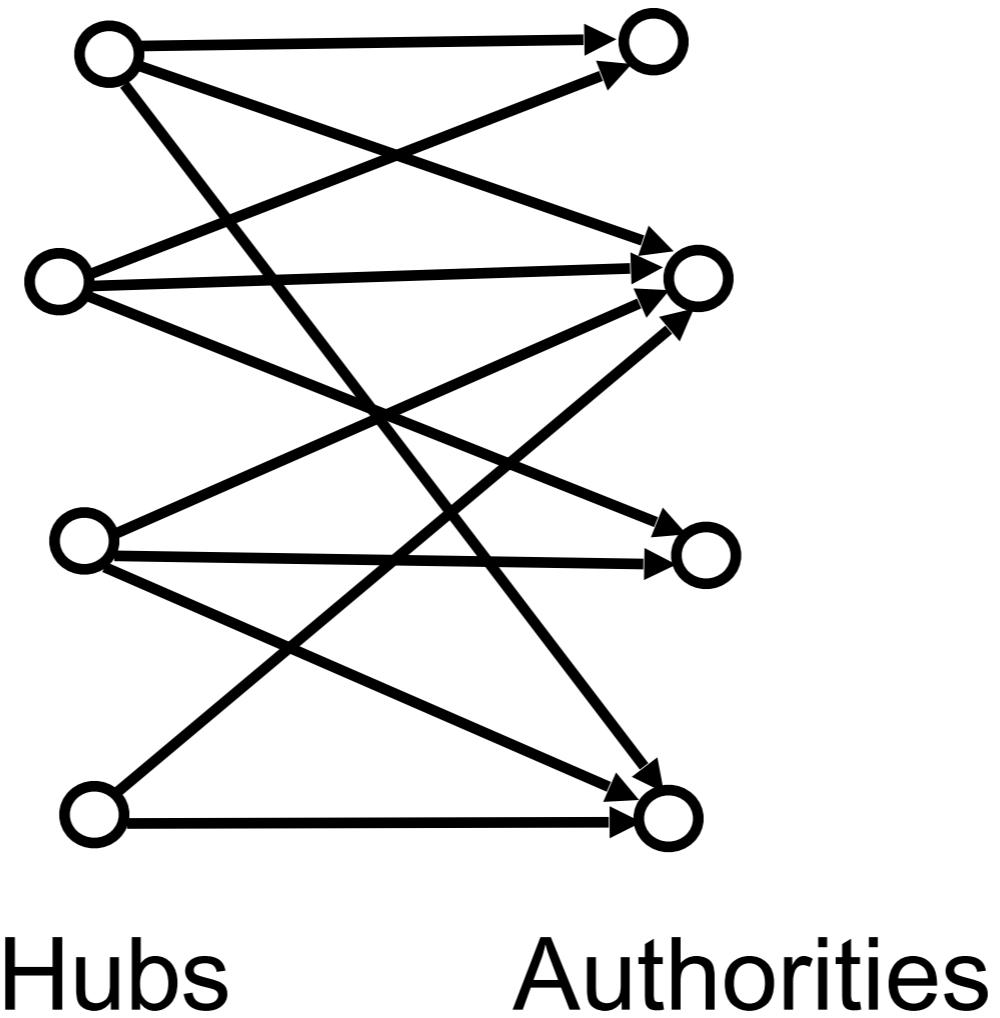
Authorities

- Relevant pages of the highest quality on a broad topic

Hubs

- Pages that link to a collection of authoritative pages on a broad topic

Hub-Authority Relations



Intuition

A **good hub** page is one that points to many **good authorities**

A **good authority** page is one that is pointed to by many **good hub** pages

Hyperlink-Induced Topic Search (HITS)

The approach consists in two phases:

- ▶ It uses the query terms to collect a starting set of pages (200 pages) from an index-based search engine – root set of pages.
- ▶ The root set is expanded into a base set by including all the pages that the root set pages link to, and all the pages that link to a page in the root set, up to a designed size cutoff, such as 2000-5000.
- ▶ A weight-propagation phase is initiated. This is an iterative process that determines numerical estimates of hub and authority weights

Hub and Authorities

- ▶ Let **a** and **h** be vectors, whose i^{th} component corresponds to the degrees of authority and hubbiness of the i^{th} page. Then:
- ▶ **h = A × a**. That is, the hubbiness of each page is the sum of the authorities of all the pages it links to.
- ▶ **a = A^T × h**. That is, the authority of each page is the sum of the hubbiness of all the pages that link to it (**A^T** - transposed matrix).
- ▶ Then, **a = A^T × A × a h = A × A^T × h**

Hub and Authorities

Example

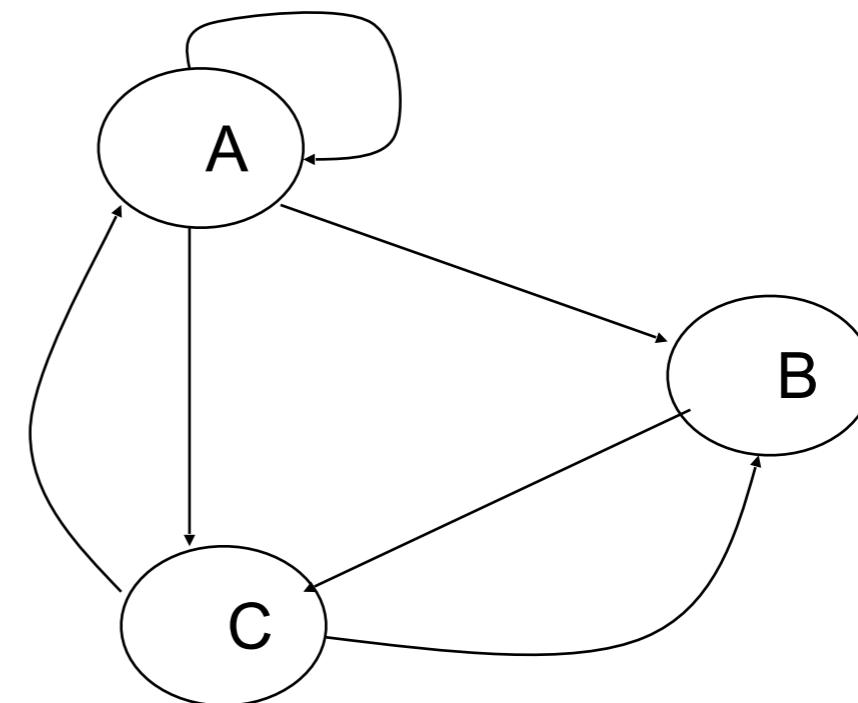
Consider the Web presented below.

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

$$A^T = \begin{bmatrix} 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{bmatrix}$$

$$AA^T = \begin{bmatrix} 3 & 1 & 2 \\ 1 & 1 & 0 \\ 2 & 0 & 2 \end{bmatrix}$$

$$A^TA = \begin{bmatrix} 2 & 2 & 1 \\ 2 & 2 & 1 \\ 1 & 1 & 2 \end{bmatrix}$$



Hub and Authorities

Example

If we assume that the vectors $h = [h_a, h_b, h_c]$ and $a = [a_a, a_b, a_c]$ are each initially $[1,1,1]$, the first three iterations of the equations for a and h are the following:

$a_a = 1$	5	24	114
$a_b = 1$	5	24	114
$a_c = 1$	4	18	84
$h_a = 1$	6	28	132
$h_b = 1$	2	8	36
$h_c = 1$	4	20	96

Metrics comparison

- **Pearson correlation coefficient** shows linear dependence between variables

$$r = \frac{\sum_i (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_i (X_i - \bar{X})^2} \sqrt{\sum_i (Y_i - \bar{Y})^2}}$$

- **Spearman rank correlation** coefficient (Sperman's rho) shows strength of monotonic association
 - Convert raw scores to ranks (X_i^* is the rank of X_i)

$$\rho = 1 - \frac{6 \sum_i (X_i^* - Y_i^*)^2}{n(n^2 - 1)}$$

Ranking comparison

- ▶ The **Kendall tau rank** distance is a metric that count the number of pairwise disagreement between two ranking lists

$$\tau = \frac{n_c - n_d}{n(n - 1)/2}$$

- n_c : the number of concordant pairs, n_d : the number of discordant pairs

- ▶ **Example**

Rank1	A	B	C	D	E
Rank 2	A	C	B	E	D

$$\tau = \frac{8 - 2}{5(5 - 1)/2} = 0.6$$

Strength of edges

Weak and Strong Edges

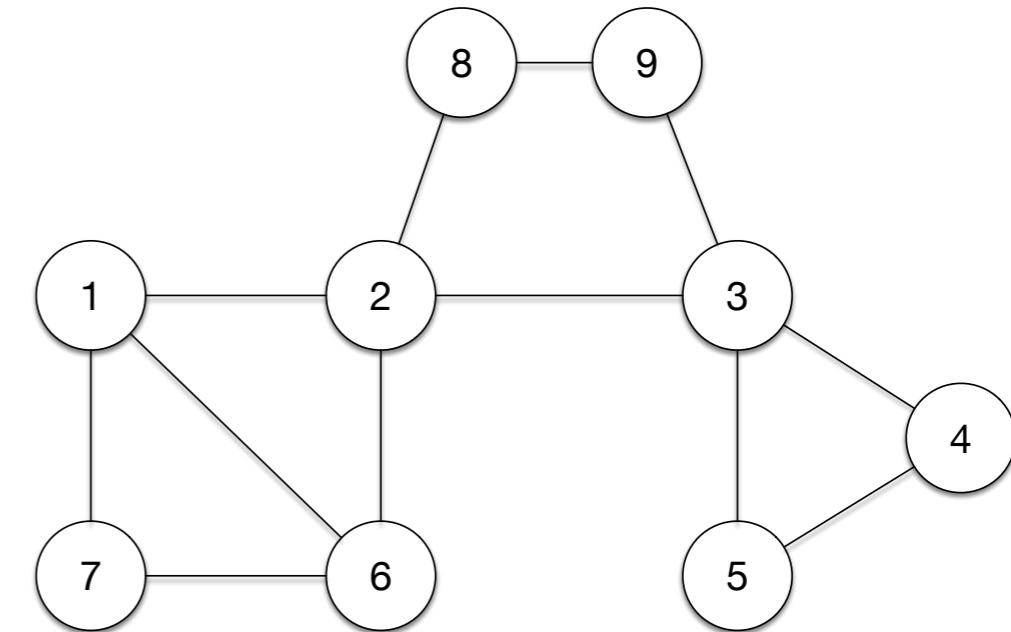
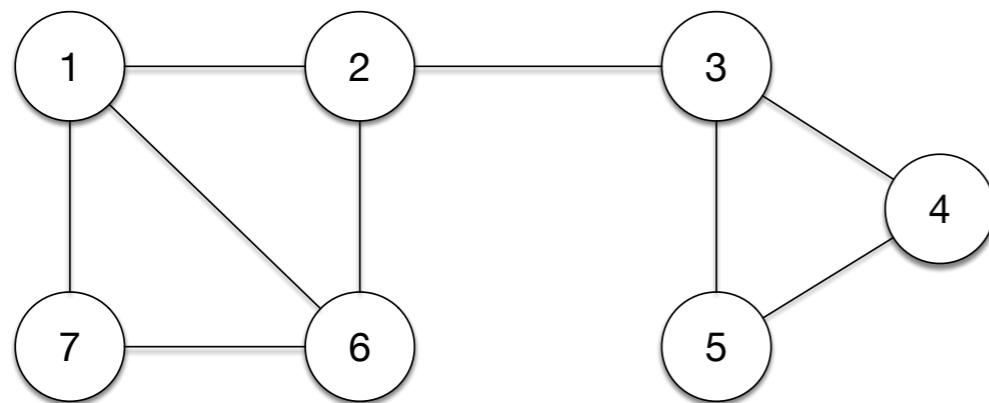
- ▶ Connections are not of the same strength
- ▶ Social networks:
 - Strong edges (close friends)
 - Weak edges (acquaintances)
- ▶ Weak and strong edges play different roles for community formation and information diffusion
- ▶ Strength of Weak Ties (*Granovetter, 1973*)
 - Occasional encounters with distant acquaintances can provide important information about new opportunities for job search

Connections in Social Media

- ▶ Social Media allow users to connect to each other more easily than ever
 - One user might have thousands of friends online
 - Who are the most important ones among your 300 FB friends?
- ▶ Need for a estimation of the importance of an edge in advanced analysis
 - Analyze network topology
 - Learn from User Profiles and Attributes
 - Learn from User Activities

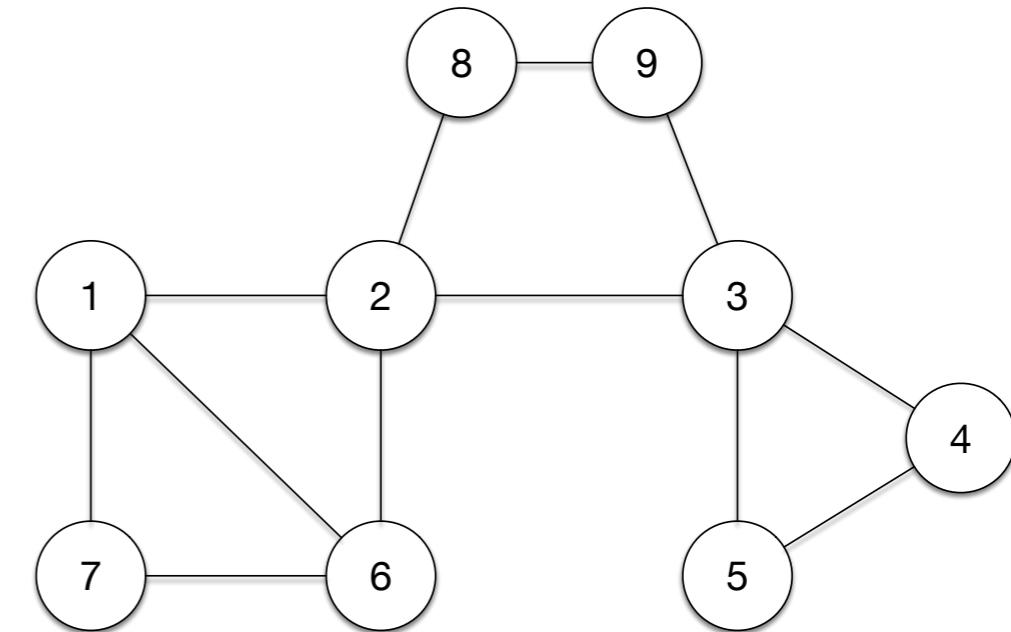
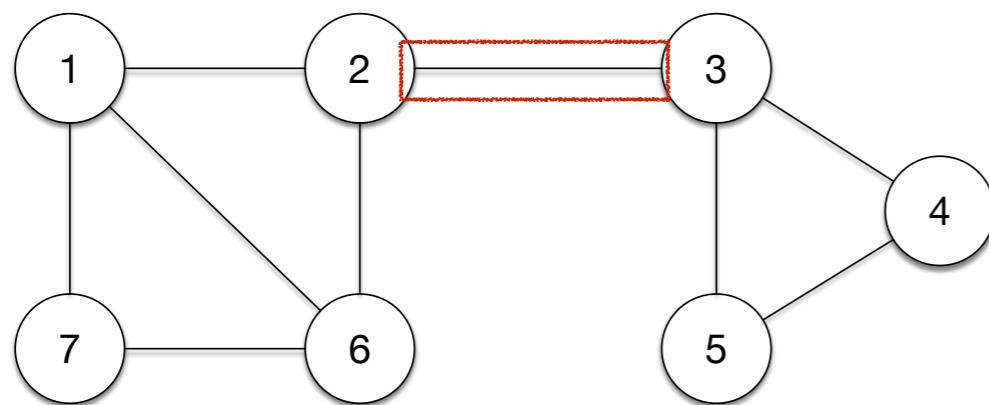
Analyze Network Topology

- ▶ Bridges that connect two different communities are weak edges
- ▶ An edge is a bridge if its removal results in disconnection of its terminal nodes



Analyze Network Topology

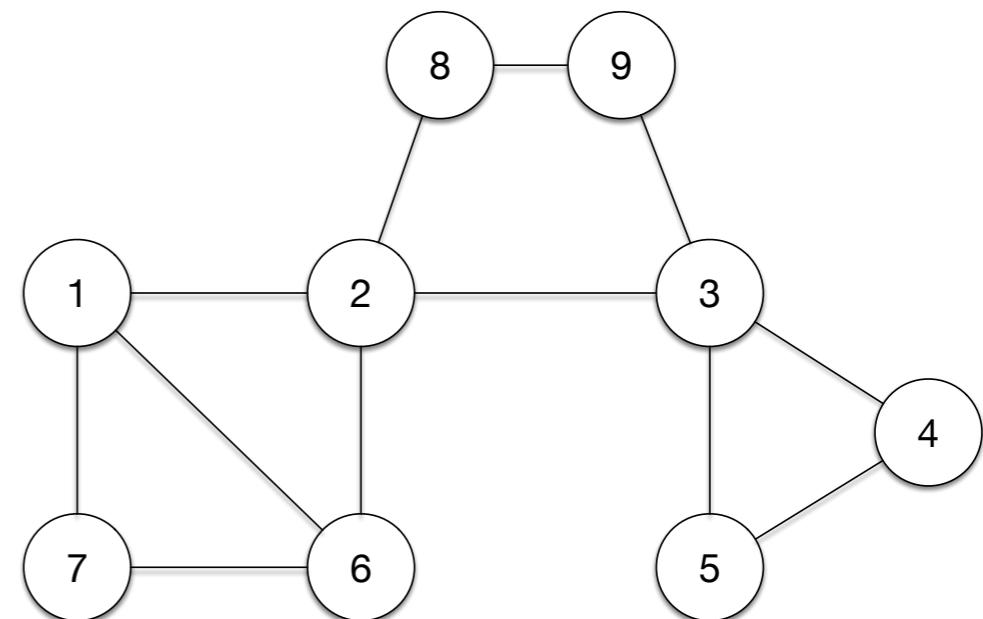
- ▶ Bridges that connect two different communities are weak edges
- ▶ An edge is a bridge if its removal results in disconnection of its terminal nodes



Shortcut Bridge

- ▶ Bridges are rare in real-life networks
- ▶ One can relax the definition by checking if the distance between two terminal nodes increases if the edge is removed
- ▶ The larger the distance, the weaker the edge

Compare (2,3)
and (2,6)

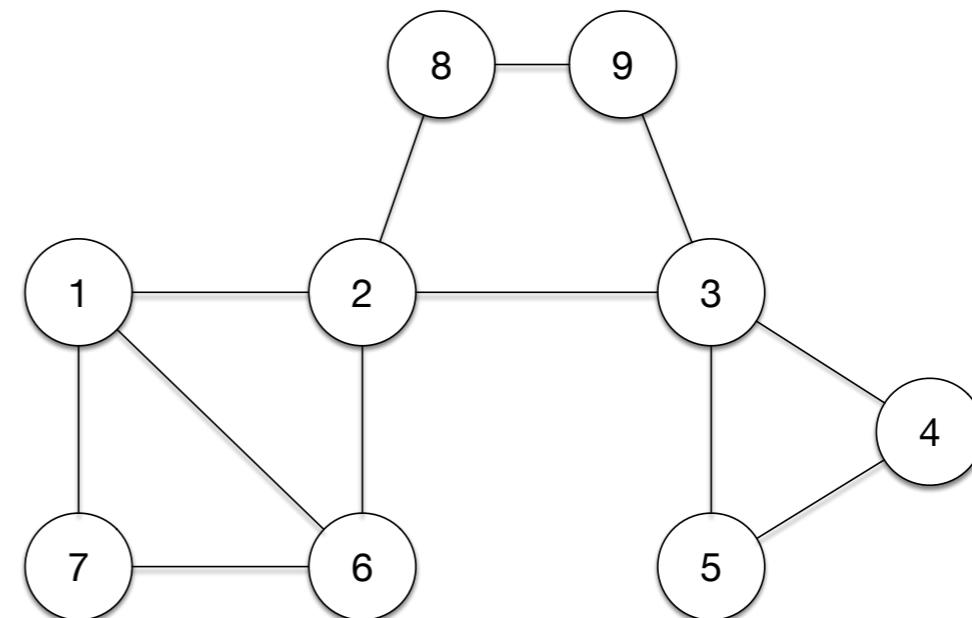


Neighbourhood Overlap

- Edge Strength can be measured based on neighbourhood overlap
- The larger the overlap, the stronger the edge

$$\text{overlap}(v_i, v_j) = \frac{|N_i \cap N_j|}{|N_i \cup N_j| - 2}$$

Compare (2,3)
and (2,6)



Learning from Profiles and Interactions

- ▶ On Twitter:
 - Friendship network is determined by the frequency two users talk each other, rather than the follower-followed network
 - The real friendship network is more influential in driving Twitter usage
- ▶ Facebook: some features help in accurately predicting strength of edges:
 - Friend-initiated posts, messages exchanged in wall post, number of mutual friends, ...
- ▶ Learning numeric link strength by maximum likelihood estimation
 - User profile similarity determines the strength
 - Link strength in turn determines user interaction
 - Maximize the likelihood based on observed profiles and interactions

Learning from User Activities

- ▶ One might learn how one influences his/her friends if the user activity log is accessible
- ▶ Depending on the adopted influence model
 - Independant cascading model
 - Linear threshold model
- ▶ Maximizing the likelihood of user activity given an influence model

Influence Modelling

Influence Modelling

- ▶ One of the fundamental question in order to understand the information diffusion, spread of new ideas, word-of-mouth (viral) marketing
- ▶ Well-known influence modelling methods
 - Independant cascading model
 - Linear threshold model

Common Properties of Influence Modelling Methods

- ▶ A social network is represented as a directed graph (each actor being a node)
- ▶ Each node is started as active or inactive
- ▶ A node, once activated, will potentially activate its neighbors
- ▶ Once activated, a node cannot be deactivated

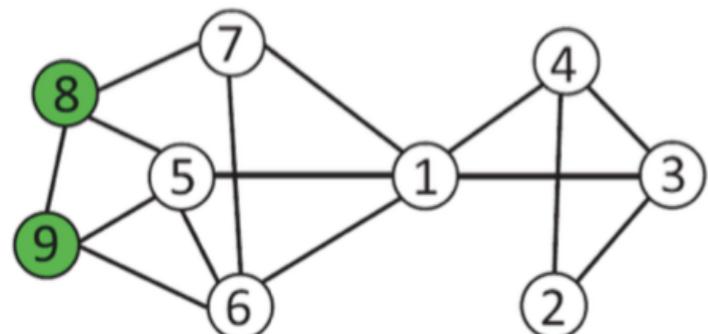
Linear Threshold Model

An actor would take an action if the number of his/her friends who have taken the action exceeds (reaches) a certain threshold

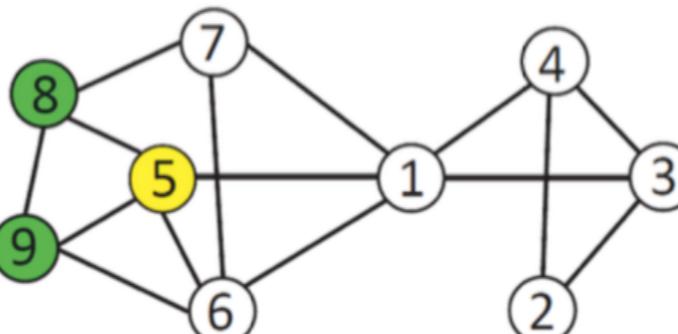
- ▶ Each node v chooses a threshold randomly from a uniform distribution in an interval between 0 and 1
- ▶ In each discrete step, all nodes that were active in the previous step remains active

Linear Threshold Model

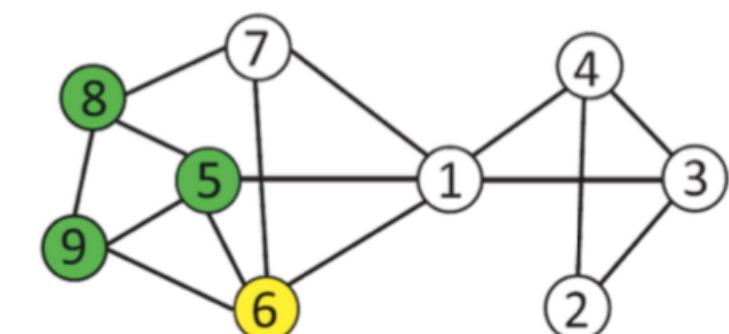
Diffusion Process (threshold = 50%)



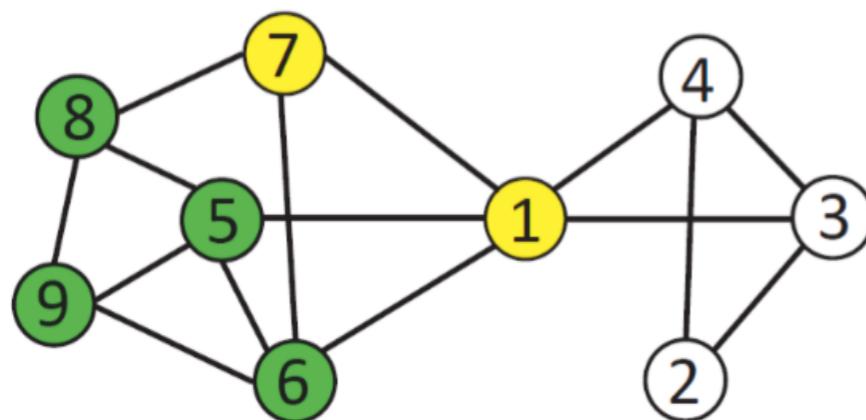
Step 0



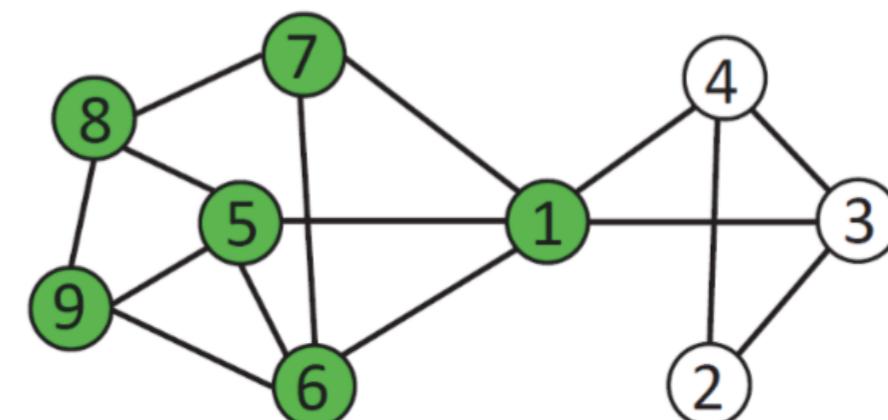
Step 1



Step 2



Step 3



Final Stage

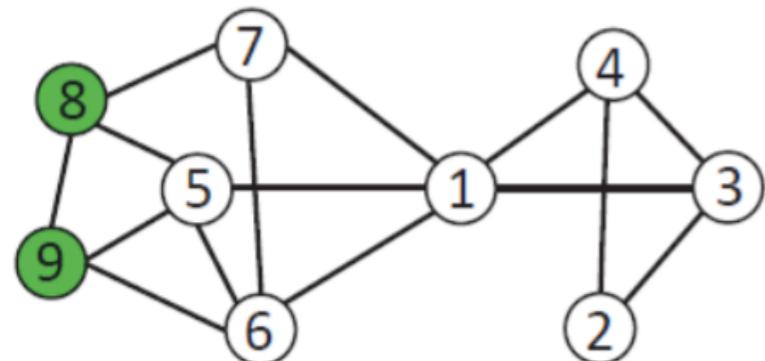
Source: Community Detection and Mining in Social Media

Independent Cascade Model

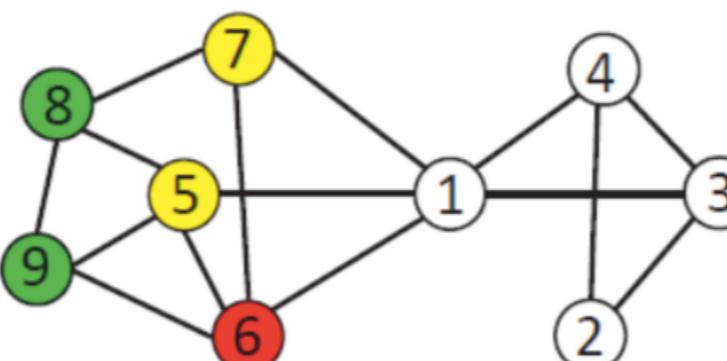
- ▶ Focuses on the sender's rather than the receiver's view
- ▶ A node w , once activated at step t , has one chance to activate each of its neighbors randomly
 - For a neighboring node v , the activation succeed with probability $p_{w,v}$ (e.g., $p = 0.5$)
- ▶ If the activation succeeds, v will become active at the next step
- ▶ In the subsequent rounds, w will not attempt to activate v anymore
- ▶ The diffusion process starts with an initial activated set of nodes and continues until no further activation is possible

Independent Cascade Model

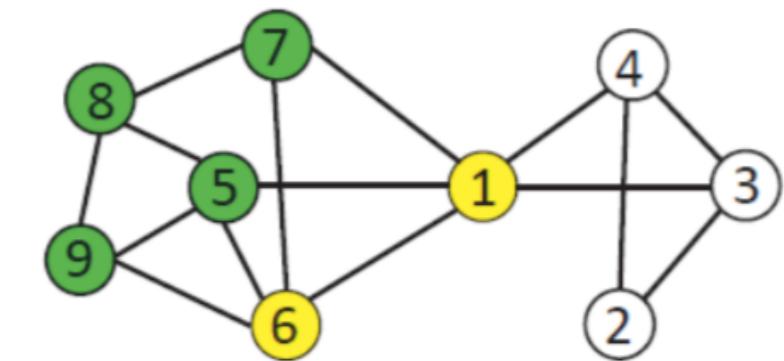
Diffusion Process



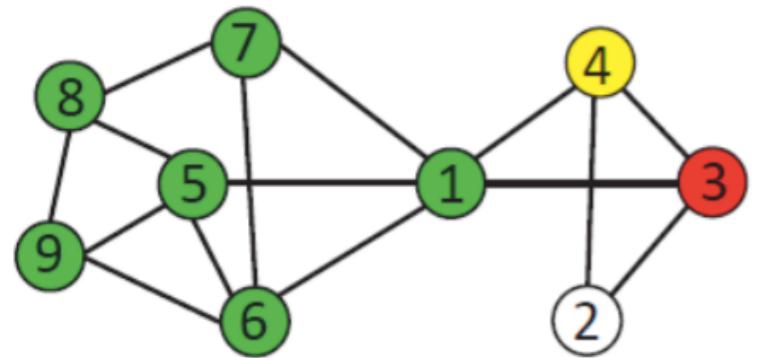
Step 0



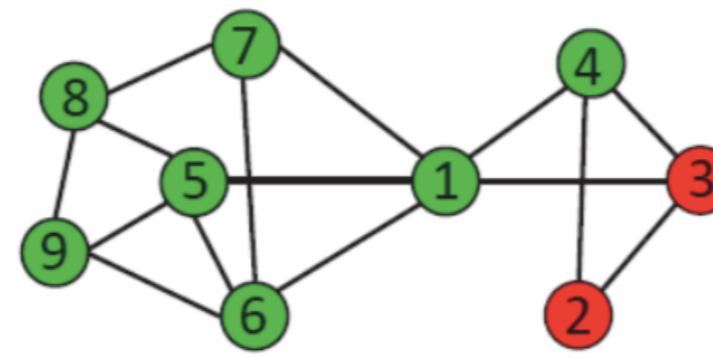
Step 1



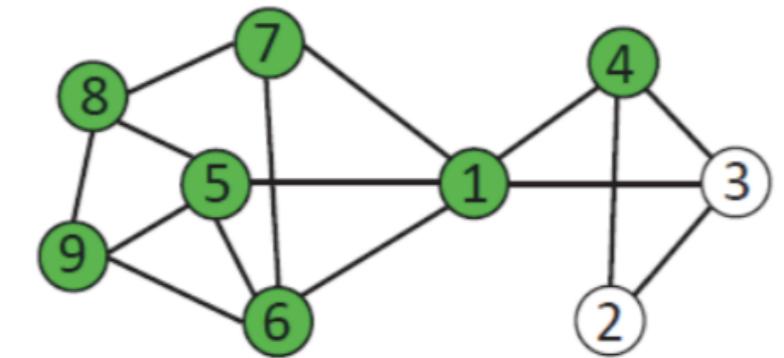
Step 2



Step 3



Step 4



Final Stage

Source: Community Detection and Mining in Social Media

Influence Maximization

- ▶ Given a network and a parameter k , which k nodes should be chosen to be in the activation set B in order to maximize the influence in terms of activated nodes at the end?
- ▶ Let $\sigma(B)$ be the expected number of nodes influenced by B . The problem can be formulated as:

$$\max_{B \subseteq V} \sigma(B) \text{ s.t. } |B| \leq k$$

Influence Maximization

A Greedy Approach

- ▶ Influence Maximization is a NP-hard problem
- ▶ Greedy approach are proved to give a solution that is 63% optimal
- ▶ **Greedy Algorithm**
 1. Start with $B = \emptyset$
 2. Evaluate $\sigma(v)$ for each node, and pick the node with maximum σ as the first node v_1 to form $B = \{v_1\}$
 3. Iterate on the rest of nodes to increase $\sigma(B)$ at each iteration
 4. Stop when no more node can be added to B

Correlation

It has been widely observed that user attributes and behaviors tend to correlate with their social networks

- ▶ Suppose we have a binary attribute with each node, e.g., whether or not being smoker
- ▶ If the attribute is correlated with the network, we expect actors sharing the same attribute value to be positively correlated with social connections
- ▶ That is, smokers are more likely to interact with other smokers, and non-smokers with non-smokers

Test for Correlation

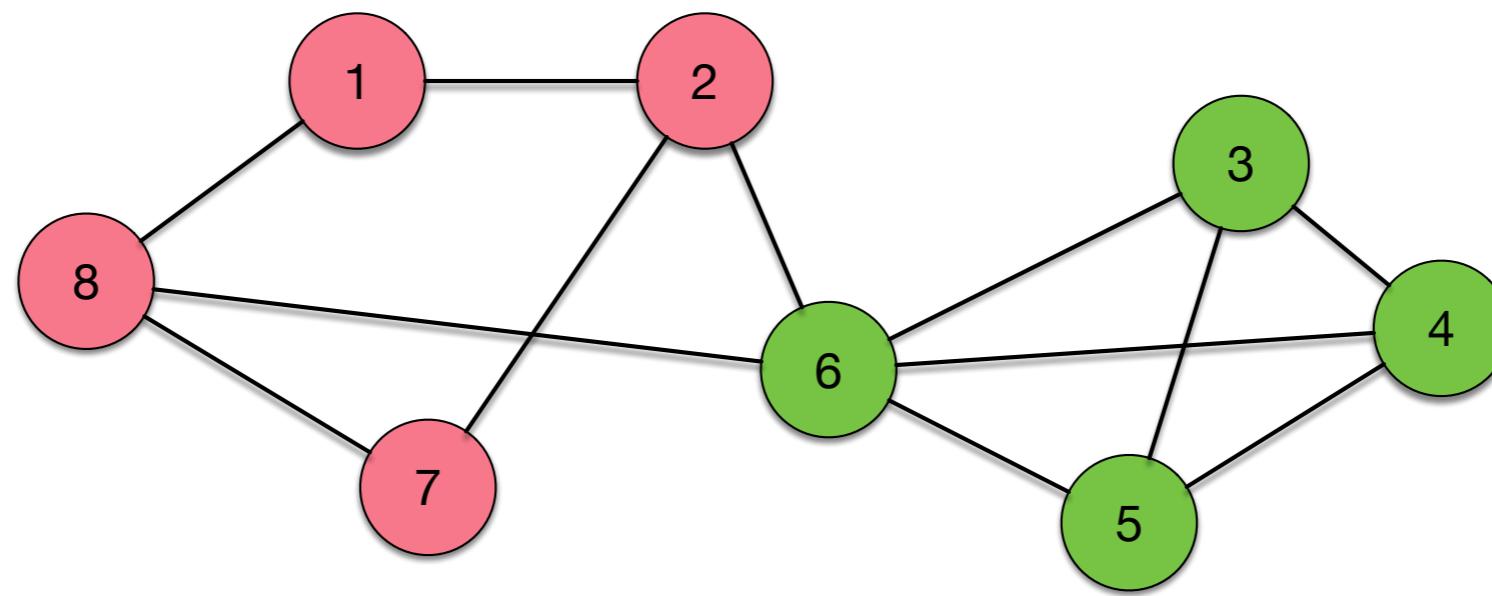
If the fraction of edges linking nodes with different attribute values are significantly less than the expected probability, then there is evidence of correlation

If connections are independent of the smoking behavior (p fraction are smokers and $1-p$ non-smoker)

One edge is expected to connect

- ▶ two smokers with probability $p \times p$,
- ▶ two non-smokers with probability: $(1 - p) \times (1 - p)$
- ▶ A smoker and a non-smoker: $2 \times p \times (1-p)$

Test for Correlation

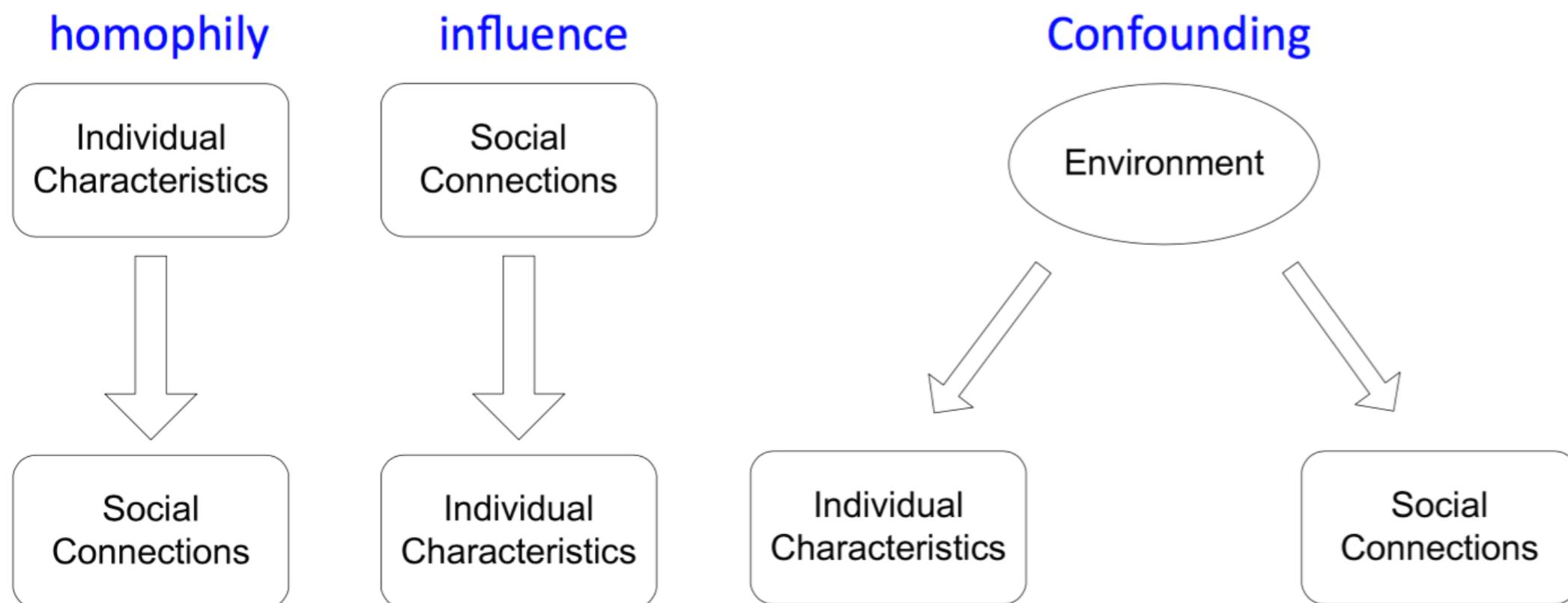


- In case of no correlation, the probability to connect a smokers and a non-smoker is $2 \times 1/2 \times 1/2 = 50\%$
- Here: $2/12 = 16.5\% < 50\%$
- A more formal way is to conduct a χ^2 test for independence of social connections and attributes (La Fond and Neville, 2010)

Correlation in Social Networks

It is well-known that there exists correlations between behaviors or attributes of adjacent actors in a social network

Three major social processes to explain correlations:



Source: Community Detection and Mining in Social Media

Correlation in Social Networks

- **Homophily**: tendency to link to others that share certain similarity with me
- **Confounding**: correlation between two actors can also be forged due to external influences, e.g., I will more likely be more connected to people living in my city than to randomly chosen people
- **Influence**: a process that causes behavioral correlations between connected nodes, e.g., switching to a mobile company

Influence or Correlation?

- ▶ In many studies about influence modelling, influence is determined by timestamps
- ▶ Shuffle test is an approach to identify whether influence is a factor associated with social system
- ▶ The probability of one node being active is a logistic function of the number of his/her active friends

$$p(a) = \frac{e^{\alpha ln(a+1)+\beta}}{1 + e^{\alpha ln(a+1)+\beta}}$$

with a the number of active friends and α measures social correlation (α and β are estimated using a maximum likelihood logistic regression)