Social Analytics for BI

Two different aspects of social networks

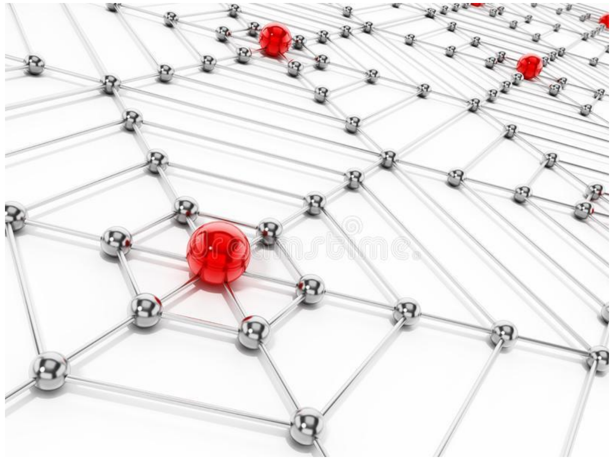
- **Social network analysis**: Analyze how the information propagates through the network. Based on network science.
  - Information diffusion
  - Influence maximization

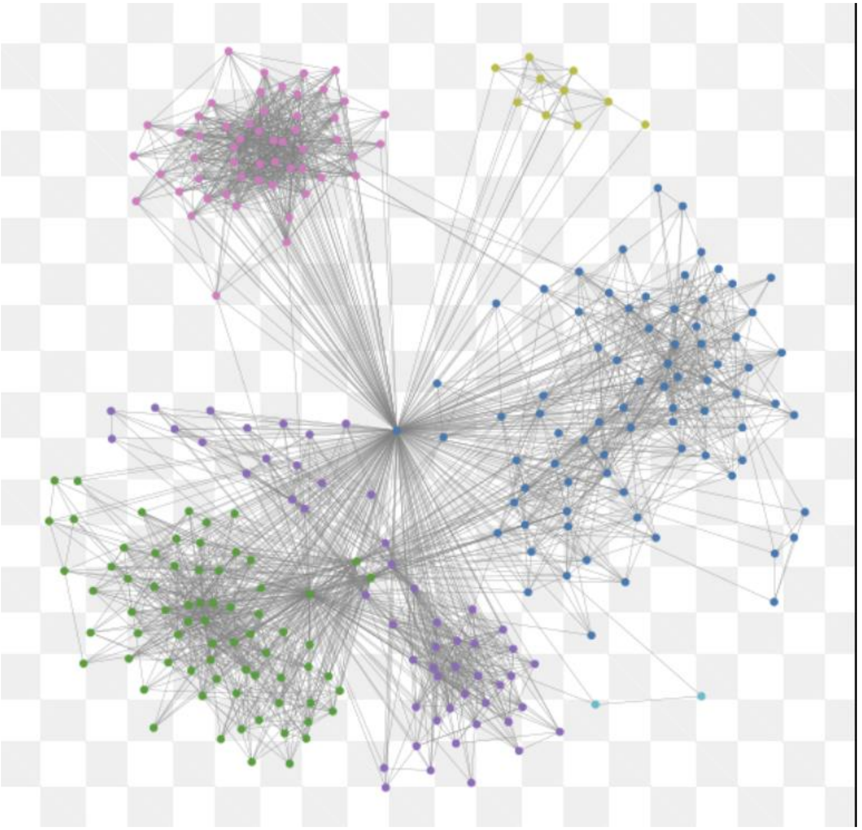
- **Social media mining**: Analyzing the information (messages) that users exchange. Based on ML and Natural Language Processing (NLP) algorithms.
  - Opinion mining
  - Topic mining

Social network analytics

- Social network analysis is a method by which one can analyze the *connections* across individuals or groups or institutions.
- The advantage of social network analysis is that it focuses on interaction (rather than on individual behavior).
- Network analysis allows us to examine how the configuration of networks influences how individuals, groups, organizations, or systems function.
- In particular: how the information and influence **propagates across the members of a network**

«Regular» networks and social networks





The information lies in this irregularity

- Center
- Bridges
- Communities (small worlds)

Centers, bridges and communities

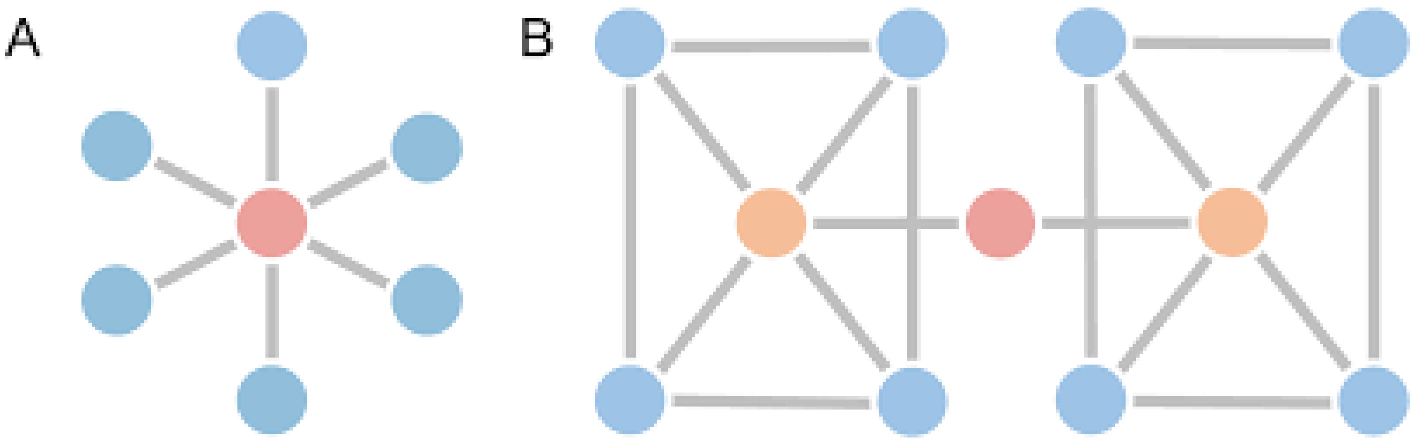
- Centers: those who have influence on others
- Bridges: those who connect different groups and facilitate the flow of information
- Communities: groups of similar ones

Centers, bridges and communities: what for?

- Centers: influencers that may promote a product
- Bridges: people that facilitate spread of influence (word of the mouth)
- Communities: "birds of a **feather** flock together." Exploit the *homophily principle* e.g., to create dedicated marketing campaign

Finding relevant players in SN: centers and bridges

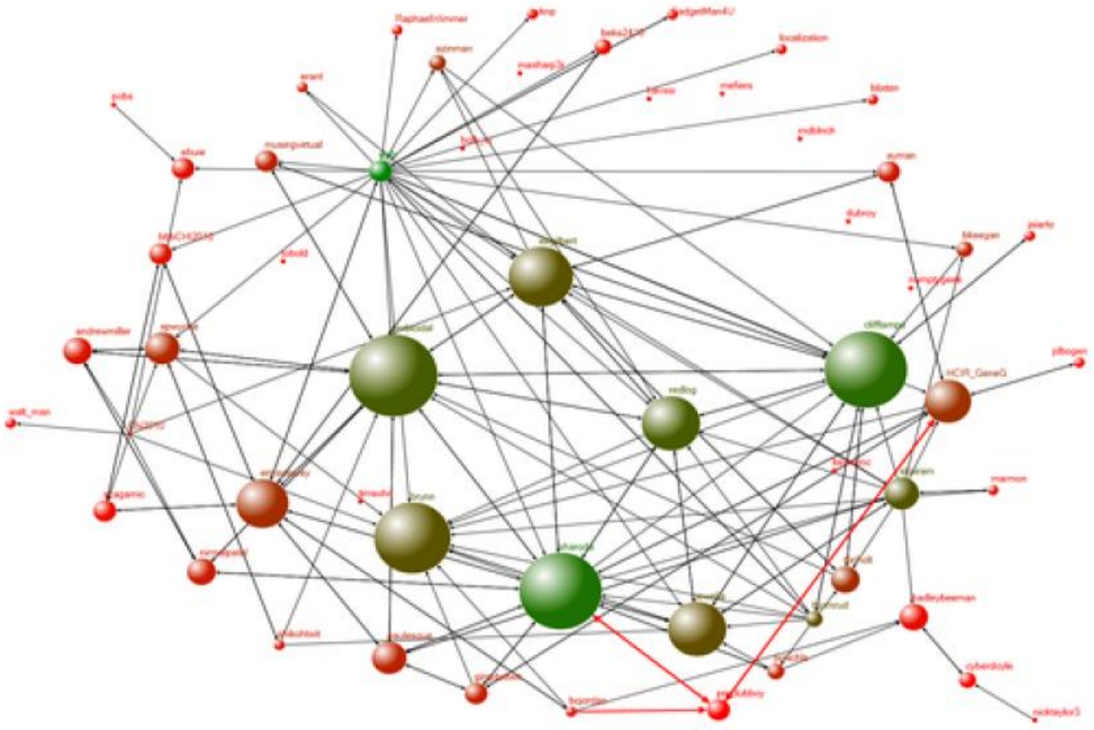
- To identify centers and bridges in a social network we need a
- A better formalization: social networks as graphs



Modeling a Social Network as a graph

NODE= “actor, vertices, points” i.e. the social entity who participates in a certain network

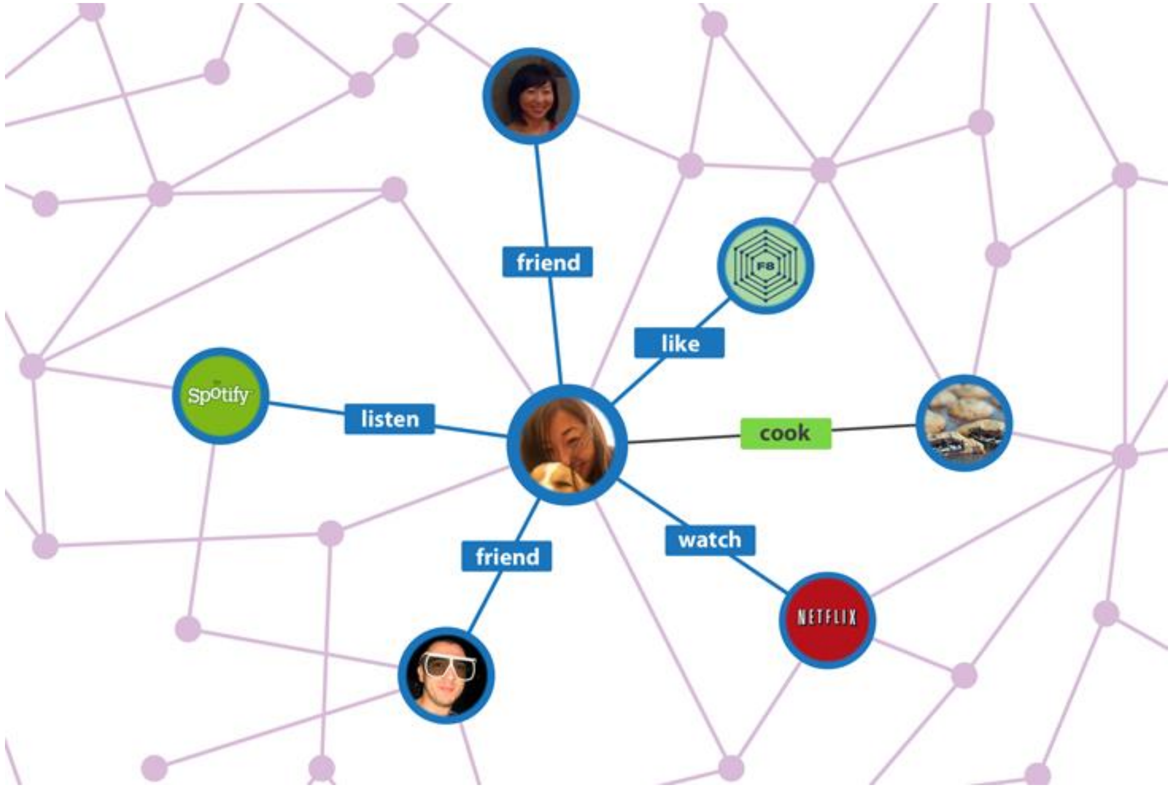
EDGE= “connection, edges, arcs, lines, ties” is defined by some type of relationship between these actors (e.g. friendship, reply/re-tweet, partnership between connected companies..)

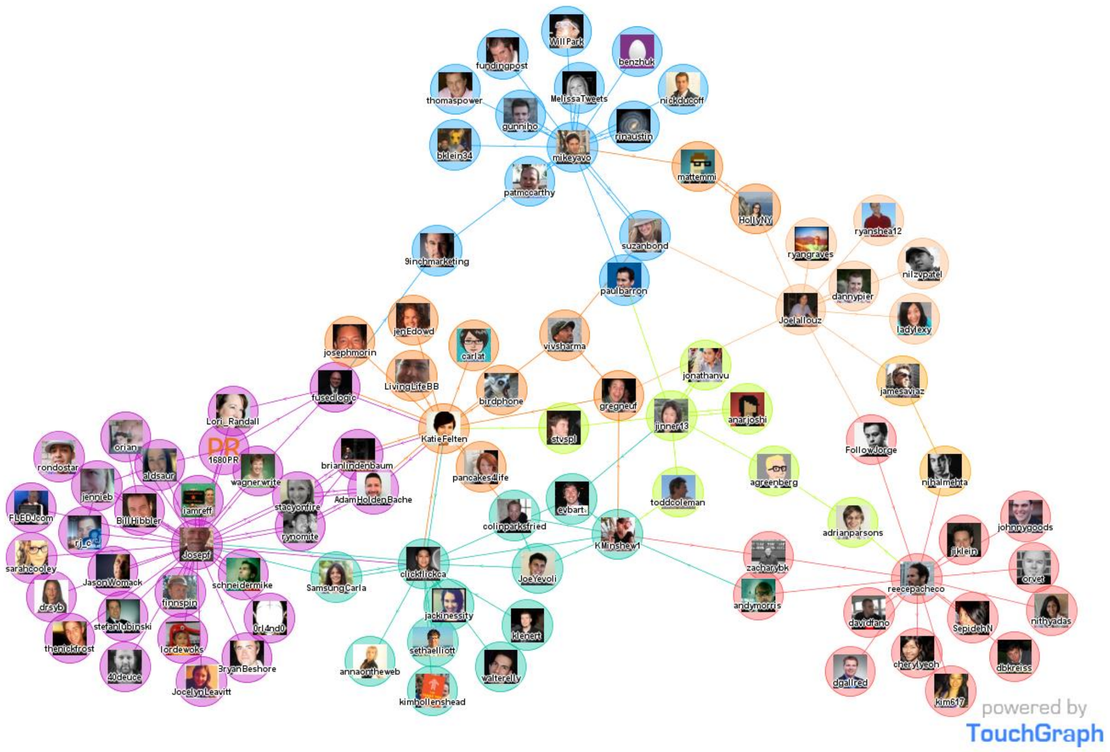


SN = graph

- A network can then be represented as a graph data structure
- We can apply a variety of measures and analyses to the graph representing a given SN
- Edges in a SN can be **directed or undirected** (e.g. friendship, co-authorship are usually undirected, emails are directed)

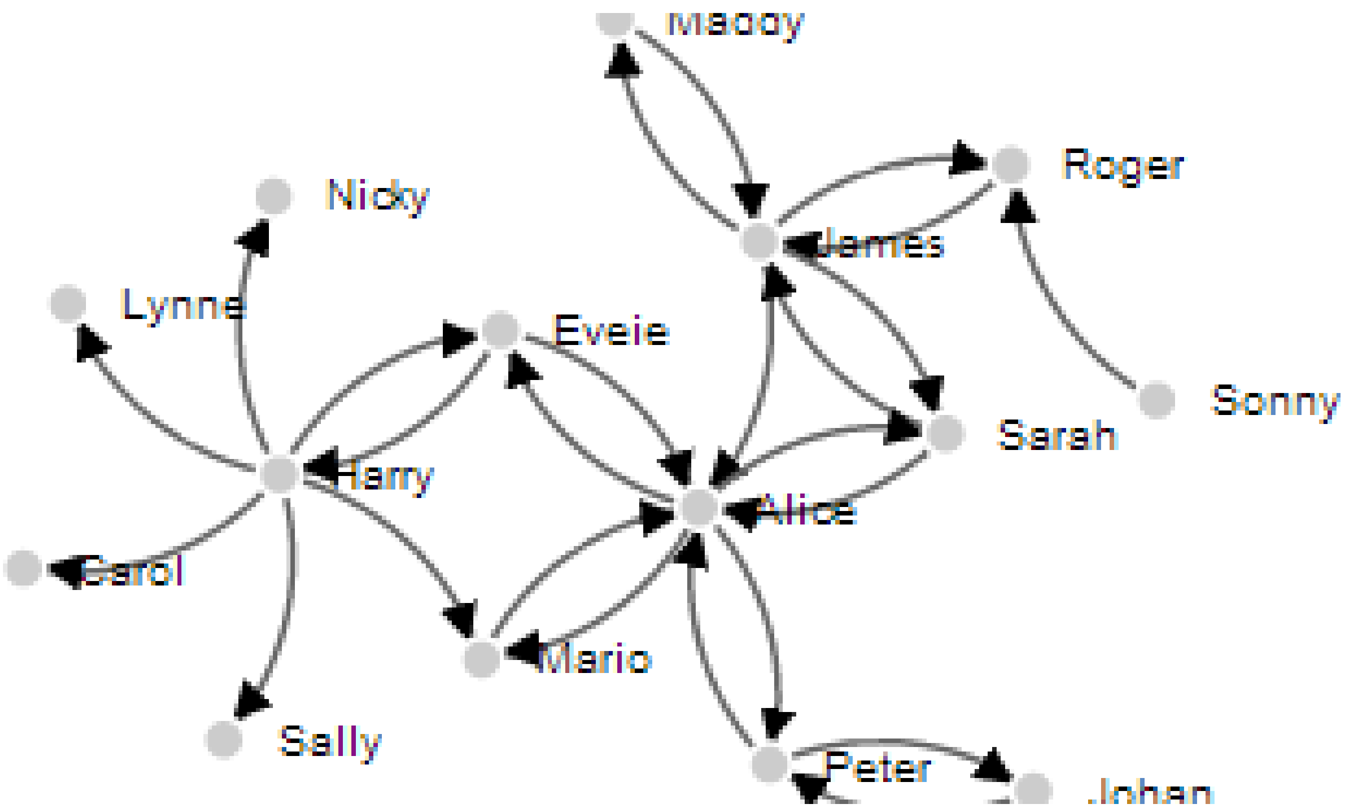
What is the meaning of edges?





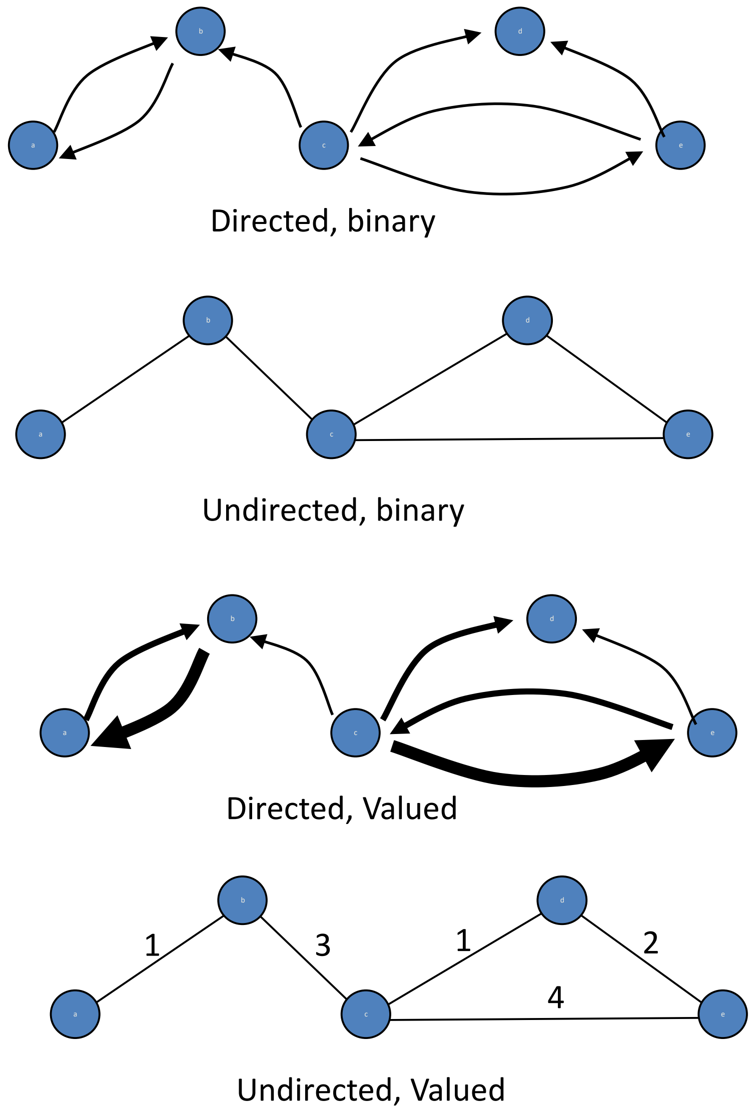
Facebook is undirected (friendship is mutual)

Twitter is a directed graph (friendship is not necessarily bidirectional)



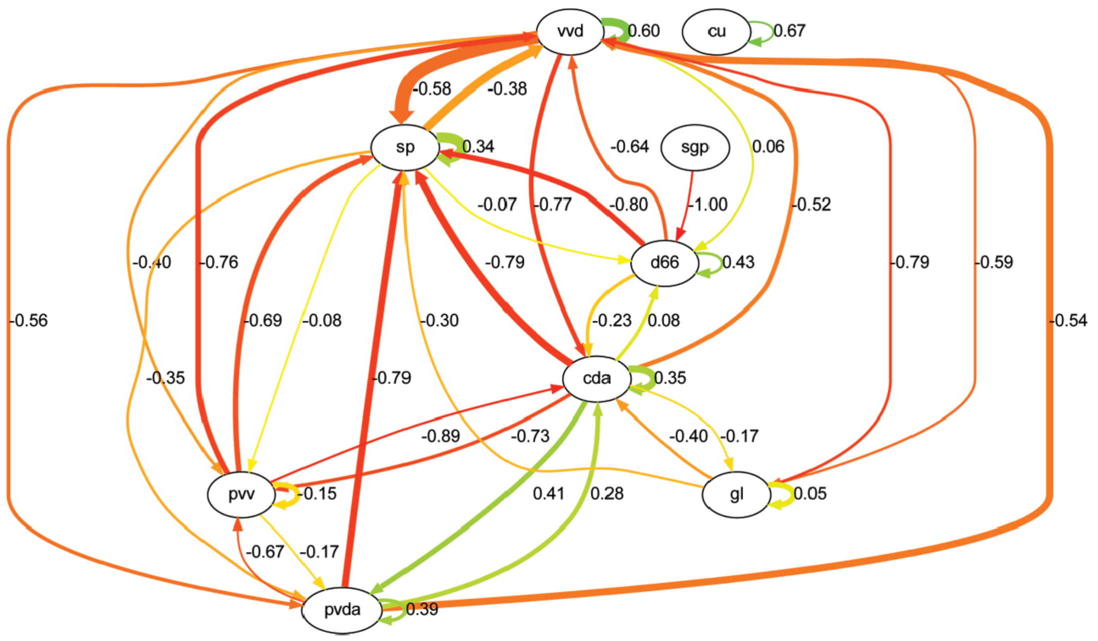
SN as a graph

- In general, a relation can be:
  Binary or Valued
  Directed or Undirected



Example of directed, valued:

- Sentiment relations among parties during a political campaign.
- Color: positive (green) negative (red).
- Intensity (thikness of edges): related to number of mutual references

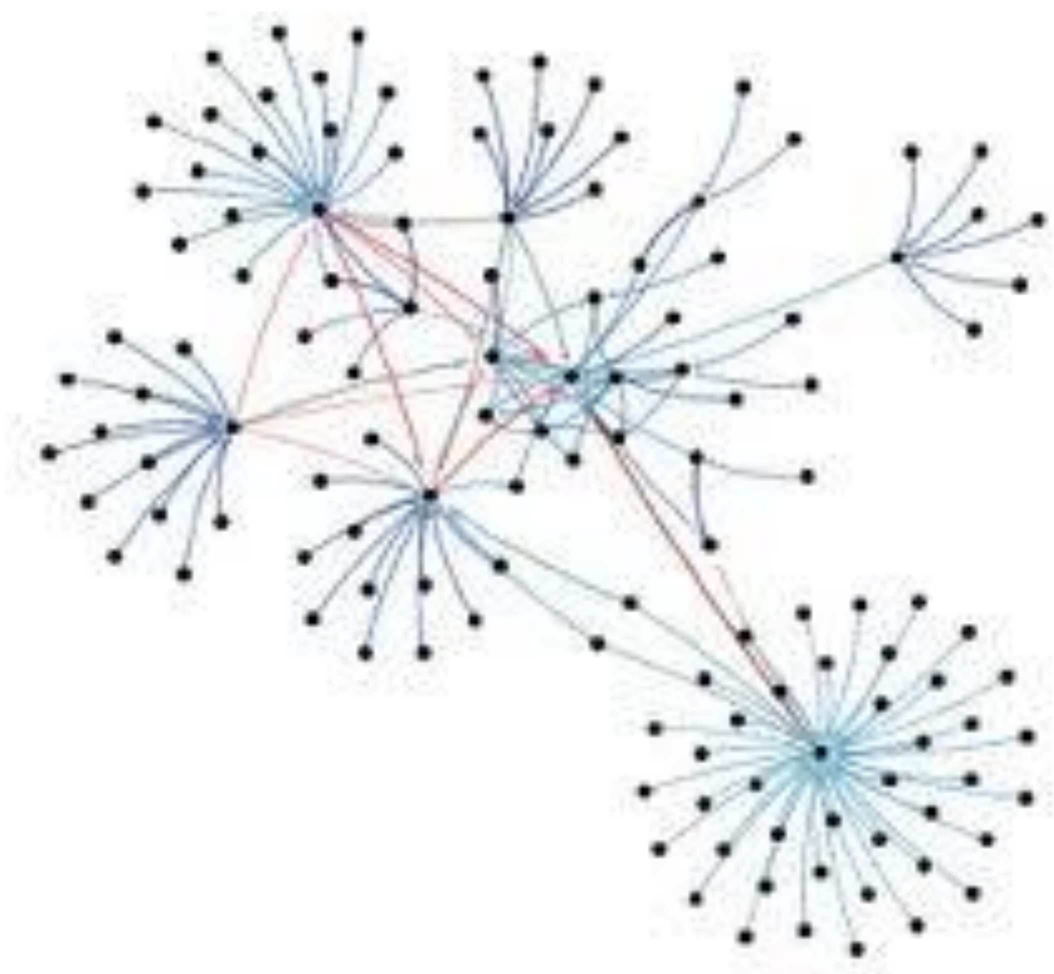


Graph-based measures of social influence

- **Use graph-based methods/algorithms to identify "relevant players" in the network**
  - **Relevant players = more influential, according to some criterion. (center, bridges)**
- Use graph-based methods to analyze the "spread" of information. (predictive)
- Use graph-based methods to identify global network properties and communities (community detection, a.k.o. clustering, descriprive/prescriptive)

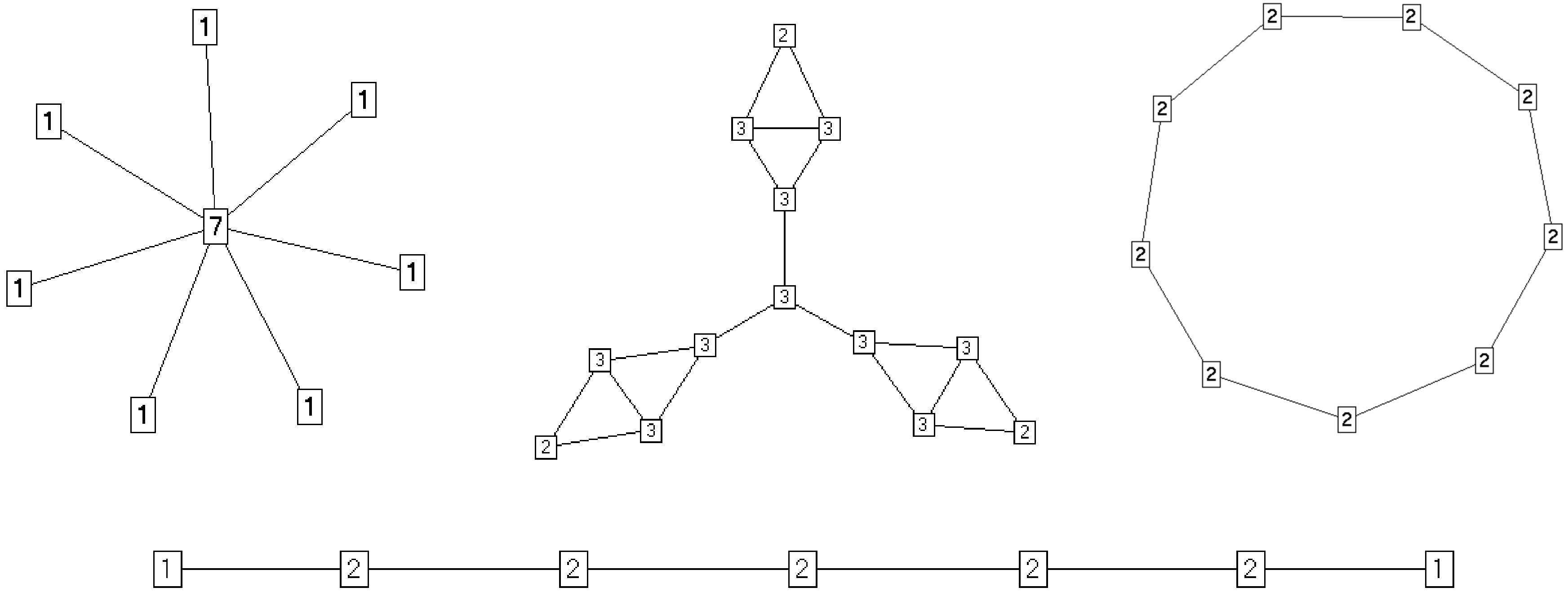
Graph-based measures of social influence: **key players**

- Using network theory, we can identify **key players** in a social network
- Key players are nodes (or actors, or vertexes) with some measurable **connectivity property**
- Two important concepts in a network are the ideas of **centrality** and **prestige** of an actor.
- Centrality more suited for undirected, prestige for directed
- Another important notion is that of **bridgeness**, or brokerage (people connecting other people)



Centrality degree calculation examples

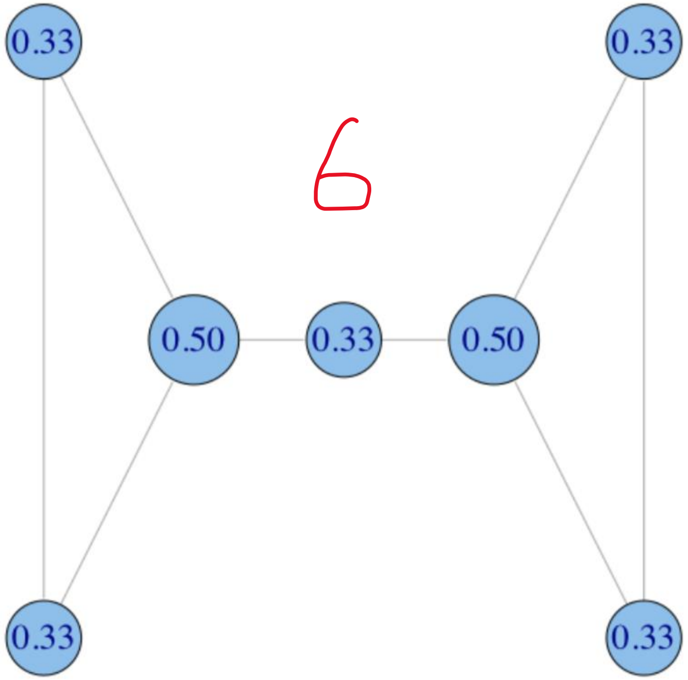
Degree is the number of ties, and the actor with the most ties is the most important:

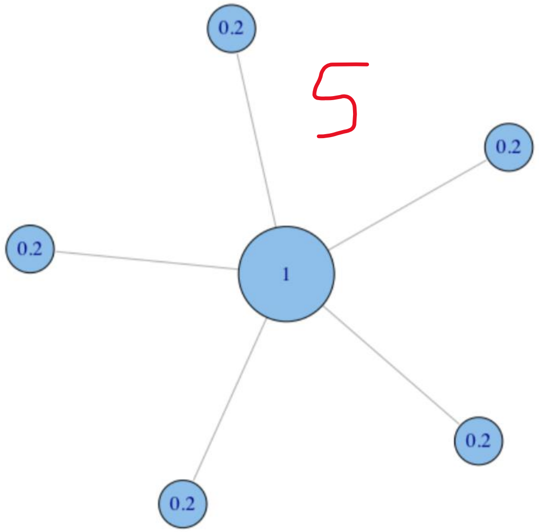


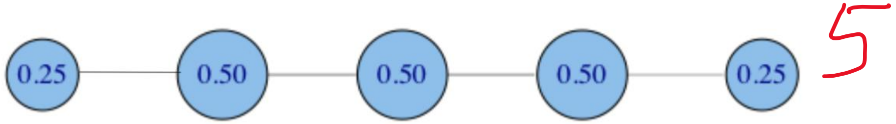
$$C_D = d(n_i) = X_{i+} = \sum_j X_{ij}$$

Centrality degree: normalization

Divide for the max possible value of C_D, i.e. N-1 for N nodes:



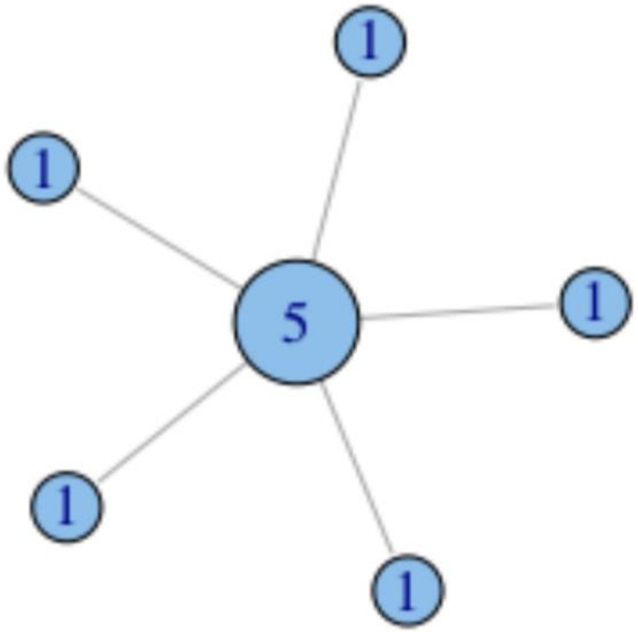




Another possible normalization is dividing by the max C_D observed value

Centrality Degree = number of connections

- Very simple measure, not always the best:
- You are “central” if you have many friends, but perhaps you are only very “social”
- Centrality degree does not account for the “importance” of your friends



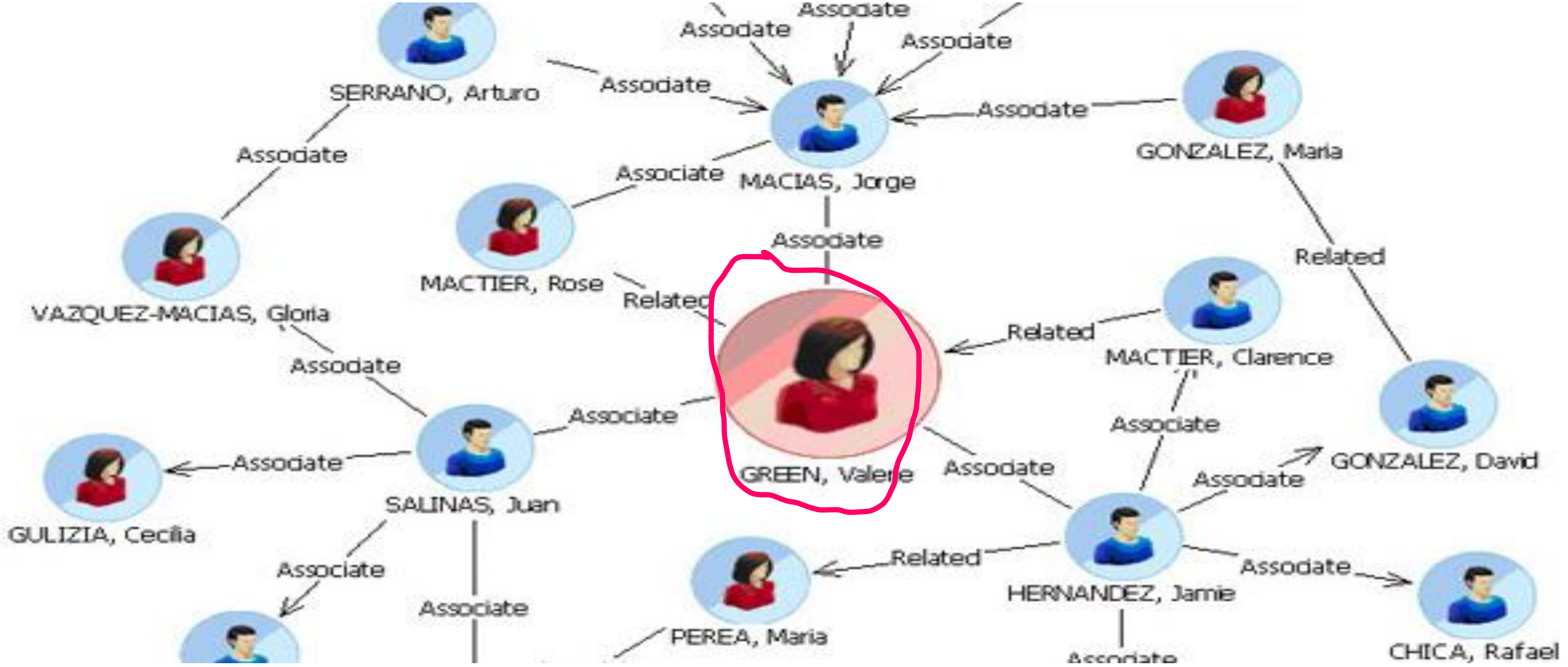
Centrality degree for directed networks (prestige)

- For **directed networks (e.g., Twitter)** direction is an important property of the relation.
- In this case we can define two different types of centrality (also called prestige for directed networks):
  - one for outgoing arcs (measures of influence),
  - one for incoming arcs (measures of support).
- Examples:
  - An actor has high influence, if he/she gives hints to several other actors (e.g. on Istagram).
  - An actor has high support, if a lot of people vote for him/her (many "likes")
  - An actor can be both an influencer and highly supported

Problem with degree centrality

- Degree Centrality depends on having many connections: but what if these connections are pretty isolated?
- A truly “central” node should be one connected to many other «powerful» nodes
- E.g. in a citation network: it is better to have fewer citations by very cited scientists than many citations by poorly cited scientists (being supported by other influencers)
- E.g. Mario Draghi being supported by Angela Merkel is better than Mario Draghi being supported by John Doe

Example

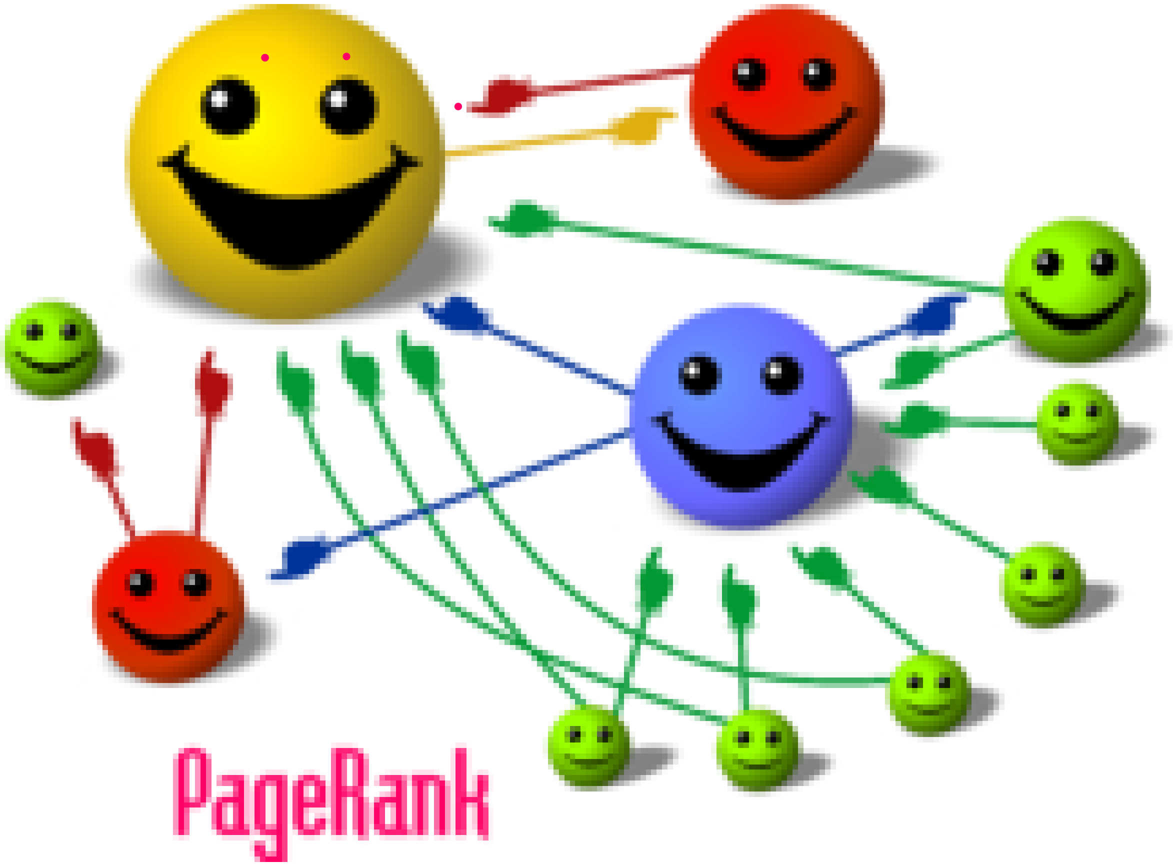


If Mrs. Green is the boss, employees referring directly to her are more important

Measuring prestige: Page Rank

- Page Rank is one of the main algorithms used by Google to rank web pages when you make a search (graph-based methods apply to any problem that can be modeled with a graph!)
- A complex method but basically the idea is that the rank (prestige) of a node depends on the rank of the other nodes pointing at that node

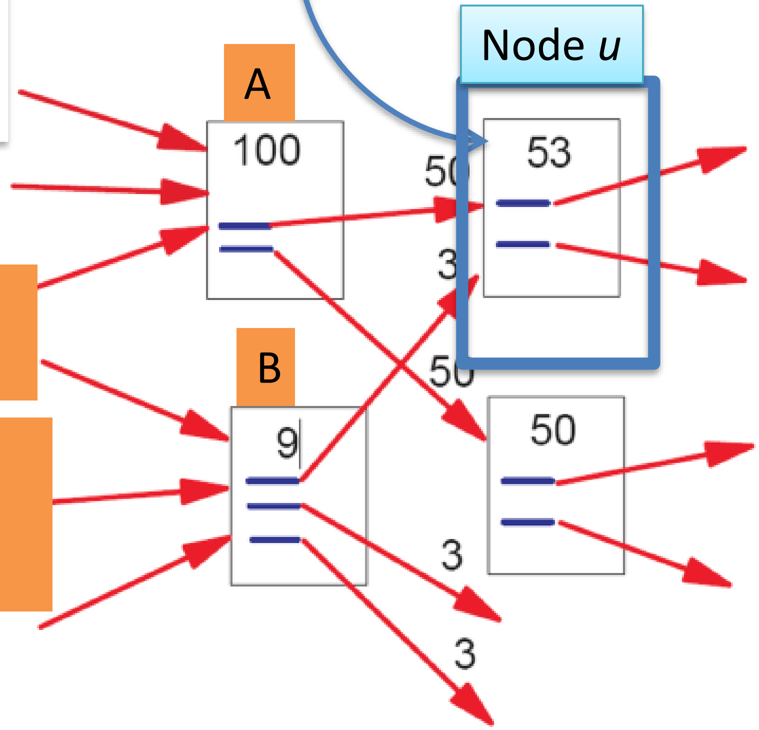
The basic principle of Page Rank



How does it works

$$PR(u) = \sum_{v \to u} \frac{PR(v)}{outlinks(v)}$$

Note this is a SIMPLIFIED formulation



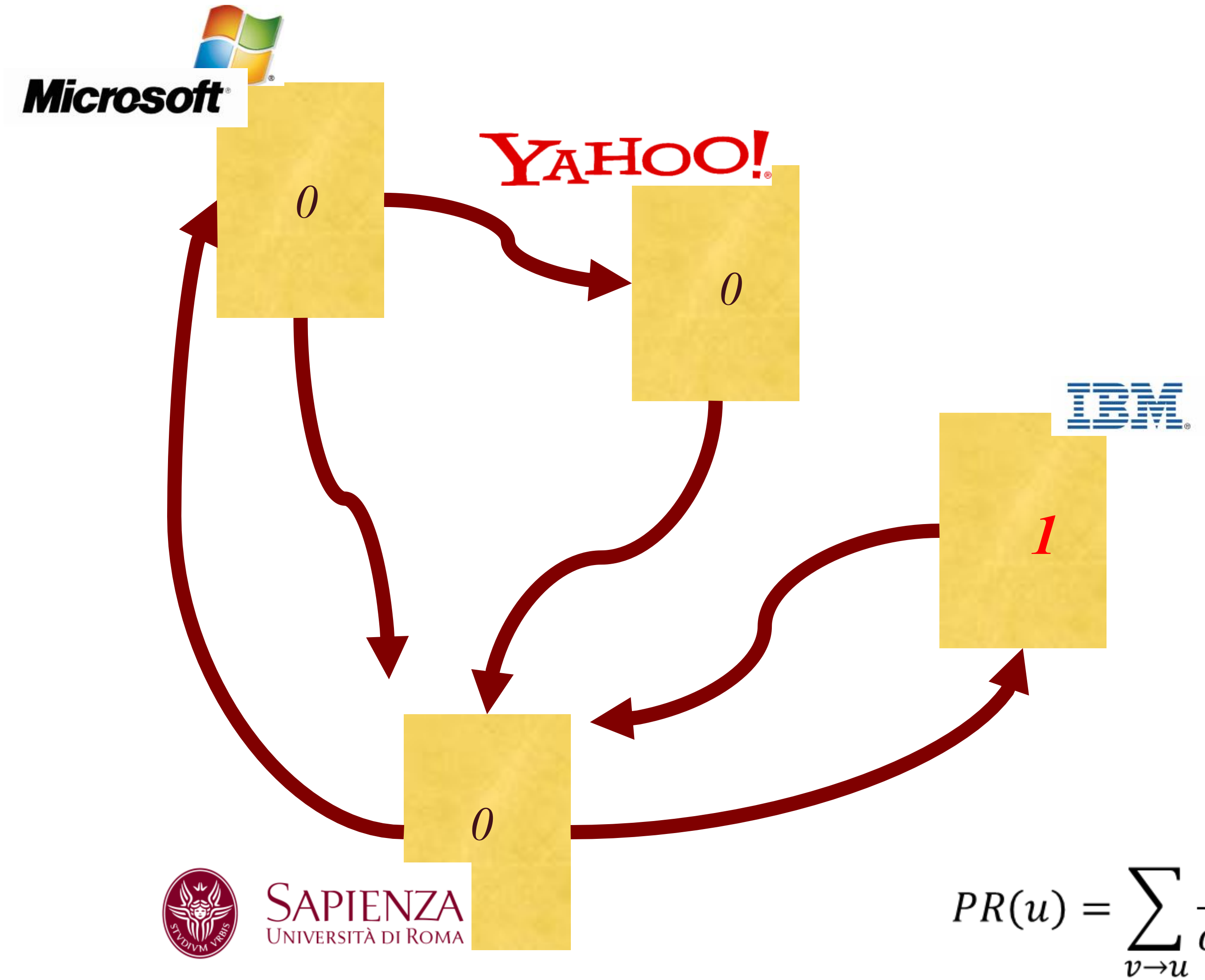
Example: Node *u* is pointed by 2 nodes, A and B with PR 100 and 9

A has 2 outlinks, B has 3. Some of these outlinks are connected to other nodes not shown here.

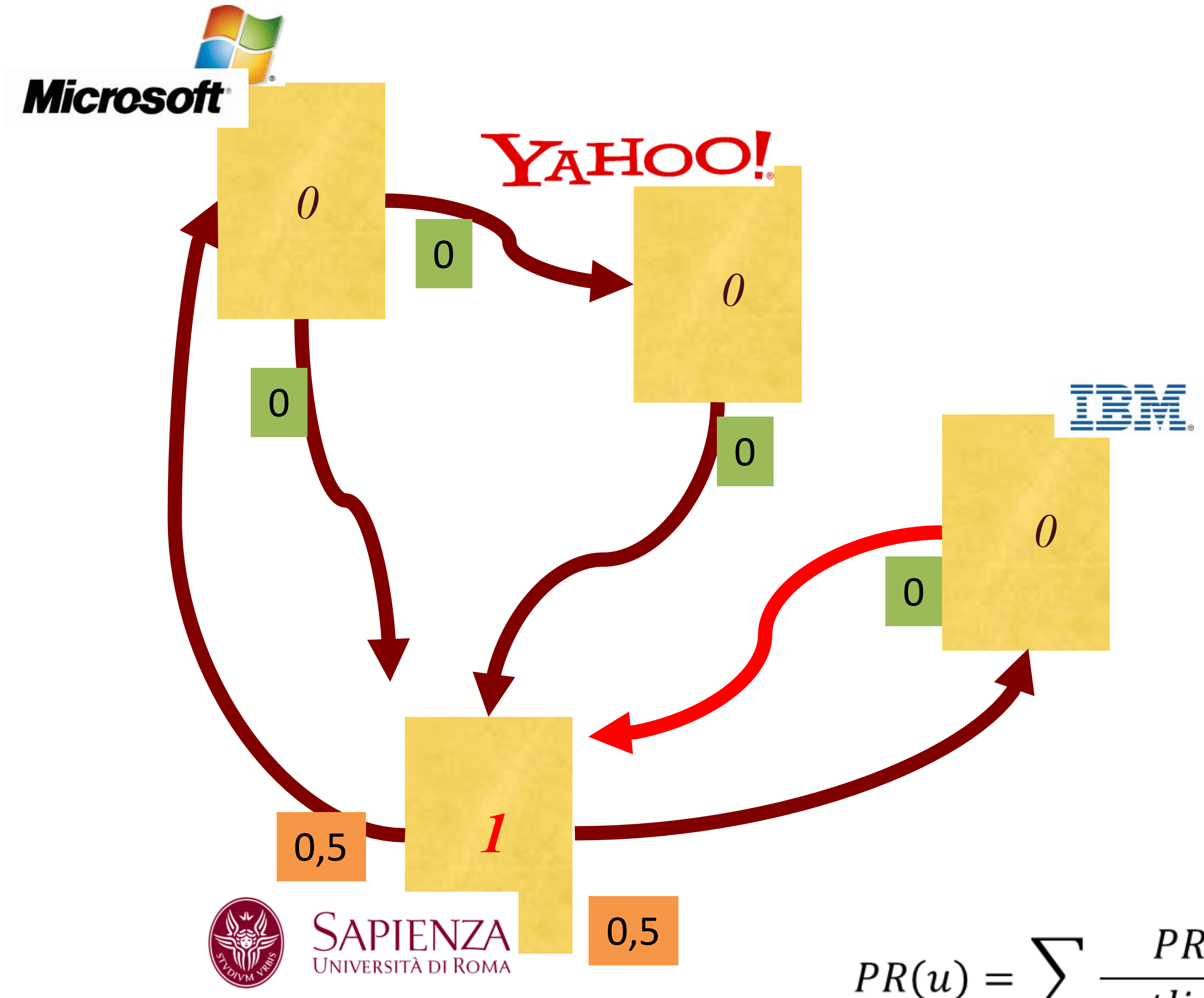
PR(u)=100/2 + 9/3=53

How is it calculated?

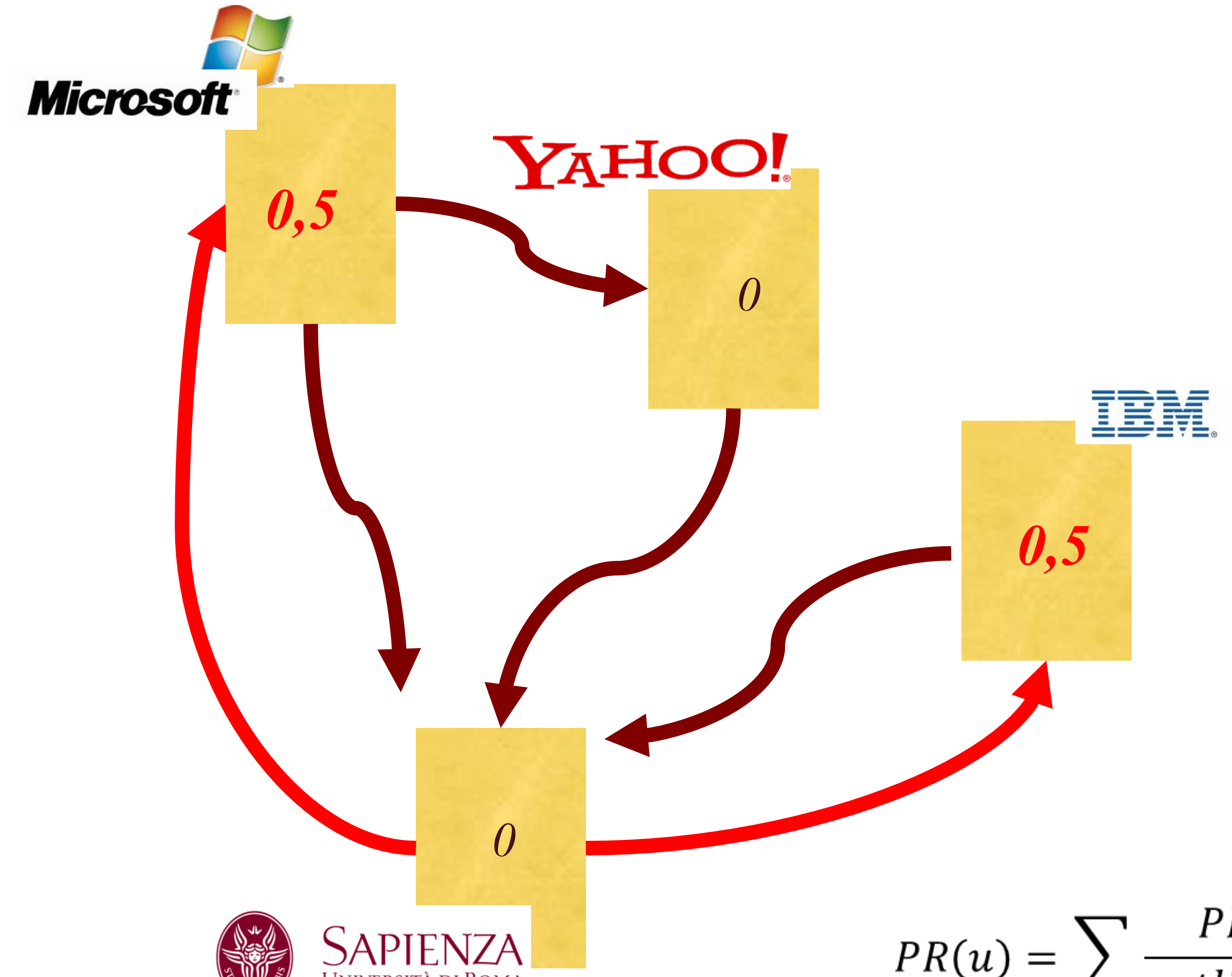
- **The rank of a node *u* depends on the rank of its pointing nodes *v*.**
- So it seems a "circular" problem: how can we compute it for all nodes?
- Start with a *random guess* of page rank values, and keep on adjusting values until values don't change (steady state)
- In computer science, this is called RECURSION



$$PR(u) = \sum_{v \to u} \frac{PR(v)}{outlinks(v)}$$

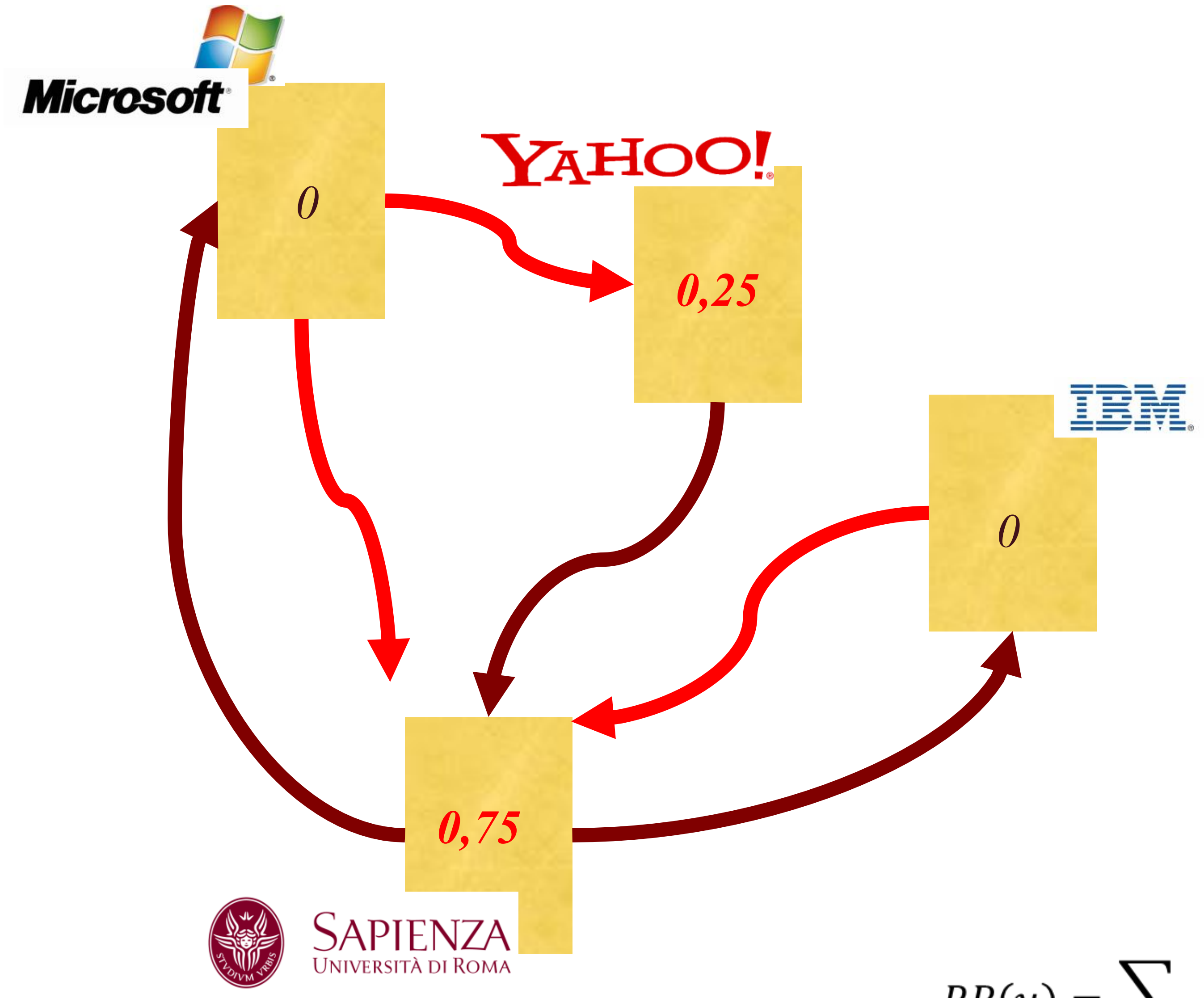


$$PR(u) = \sum_{v \to u} \frac{PR(v)}{outlinks(v)}$$



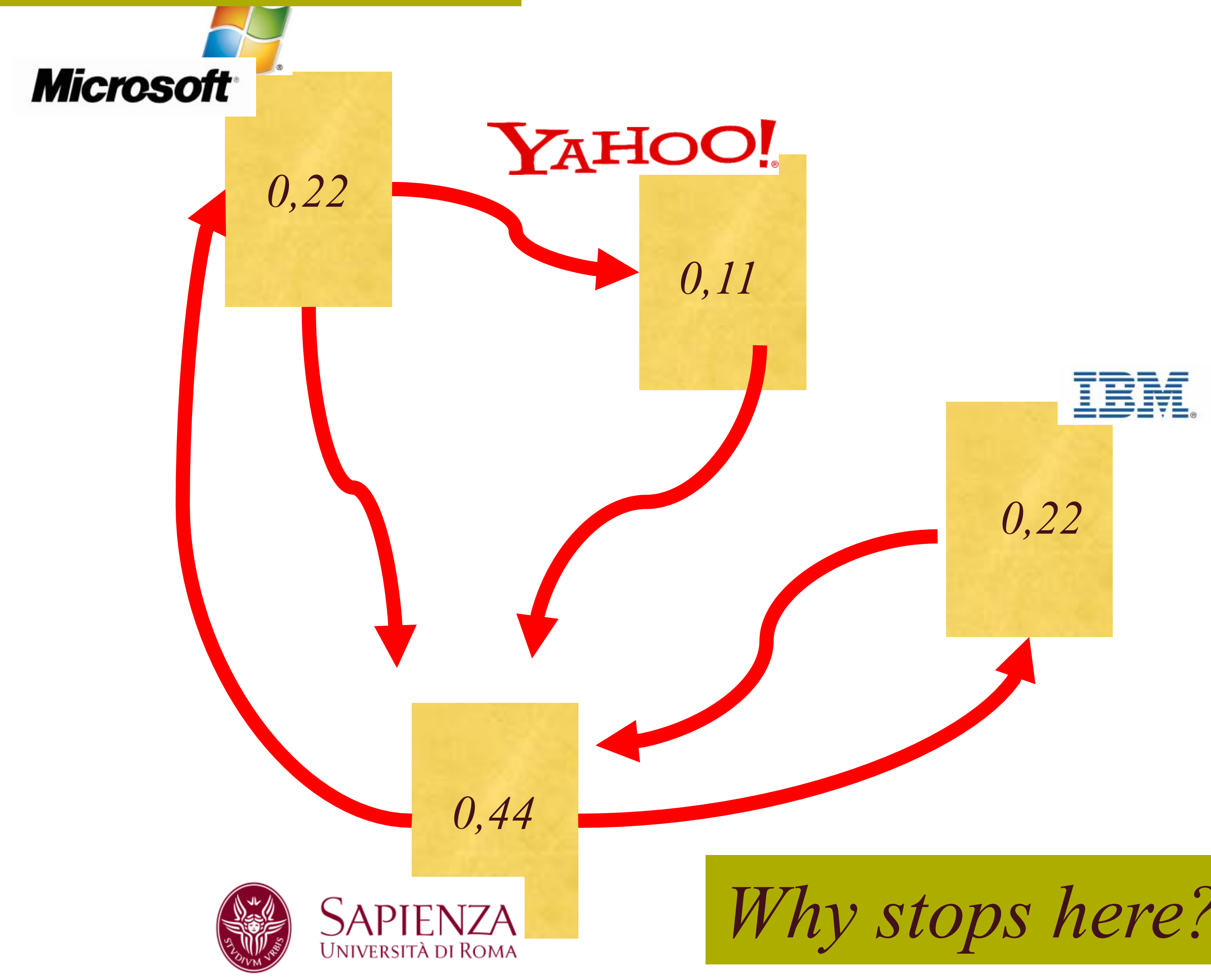
SAPIENZA
UNIVERSITÀ DI ROMA

$$PR(u) = \sum_{v \to u} \frac{PR(v)}{outlinks(v)}$$



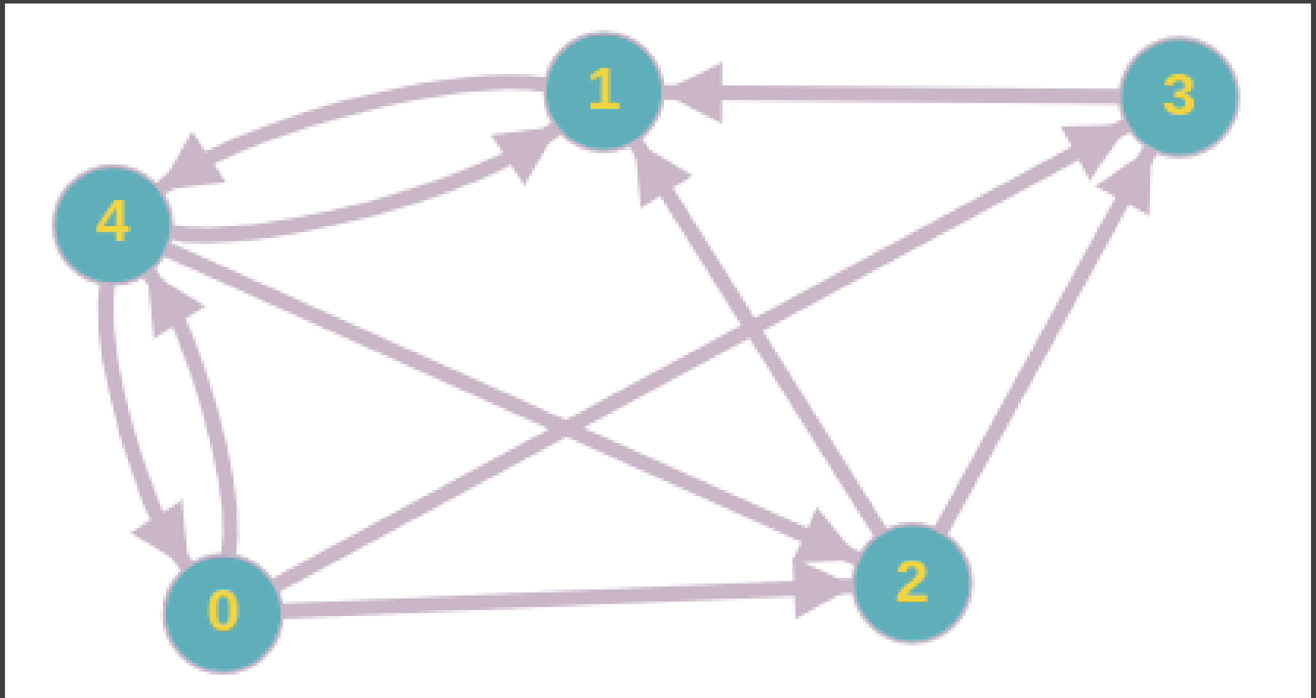
$$PR(u) = \sum_{v \to u} \frac{PR(v)}{outlinks(v)}$$

After several iterations..



Why stops here?

In class exercise step by step

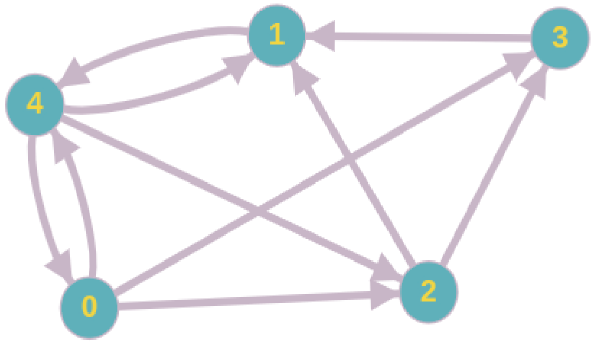


- Suppose that you start with the following “random guess” por the (page) ranks of the acors in a SN:
- r1=r2=r3=r4=0 r0=1
- Write the formulas to compute the new rank of each actor after the first iteration of the Page Rank algorithm.

Step 1

- Write the system of equations

$$PR(u) = \sum_{v \to u} \frac{PR(v)}{outlinks(v)}$$



$$\begin{cases} r_0 = \frac{r_4}{3} \\ r_1 = \frac{r_2}{2} + \frac{r_4}{3} + r_3 \\ r_2 = \frac{r_0}{3} + \frac{r_4}{3} \\ r_3 = \frac{r_2}{2} + \frac{r_0}{3} \\ r_4 = \frac{r_0}{3} + r_1 \end{cases}$$

Step 2, 3..

- Next, you put the initial vales into the formulas r1

  r1=r2=r3=r4=0 r0=1

  and you obtain:

  r_0=0; r_1=0; r_2=1/3; r_3=1/3; r_4=1/3

- Iterate: put these new values into the formulas, and continue until convergence (I am asking only the first two steps)

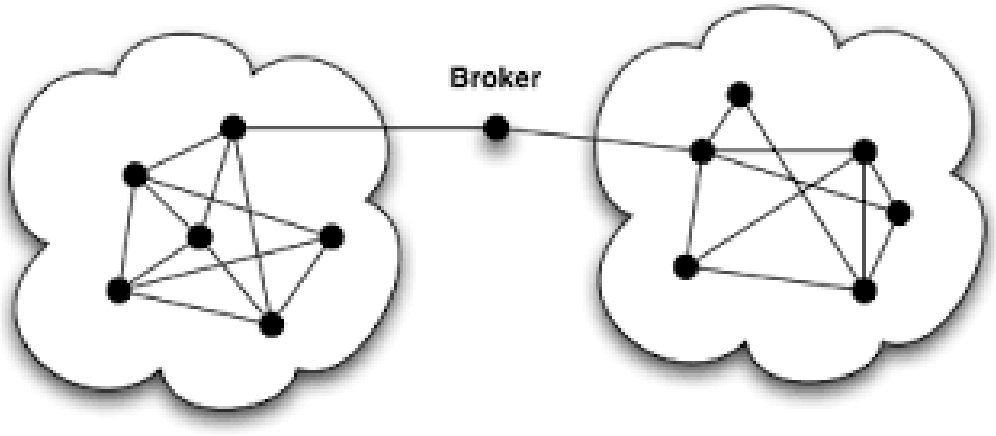
  r_0=1/9; r_1=(1/6+1/9+1/3); etc.

$$
\begin{cases}
r_0 = \frac{r_4}{3} \\
r_1 = \frac{r_2}{2} + \frac{r_4}{3} + r_3 \\
r_2 = \frac{r_0}{3} + \frac{r_4}{3} \\
r_3 = \frac{r_2}{2} + \frac{r_0}{3} \\
r_4 = \frac{r_0}{3} + r_1
\end{cases}
$$

More on finding “key players” : bridgeness

- Centrality degree, PageRank and other centrality measures tells us how a node (an individual in a social network) is “authoritative”
- There are other qualities we may want to compute, for example, the “bridgeness” (also called betweenness, brokerage, key separators..)
- People that link other people, acting as bridges
- Model based on **communication flow**: A person who lies on communication paths can control communication flow, and is thus important to ensure connections (flow of information) among groups

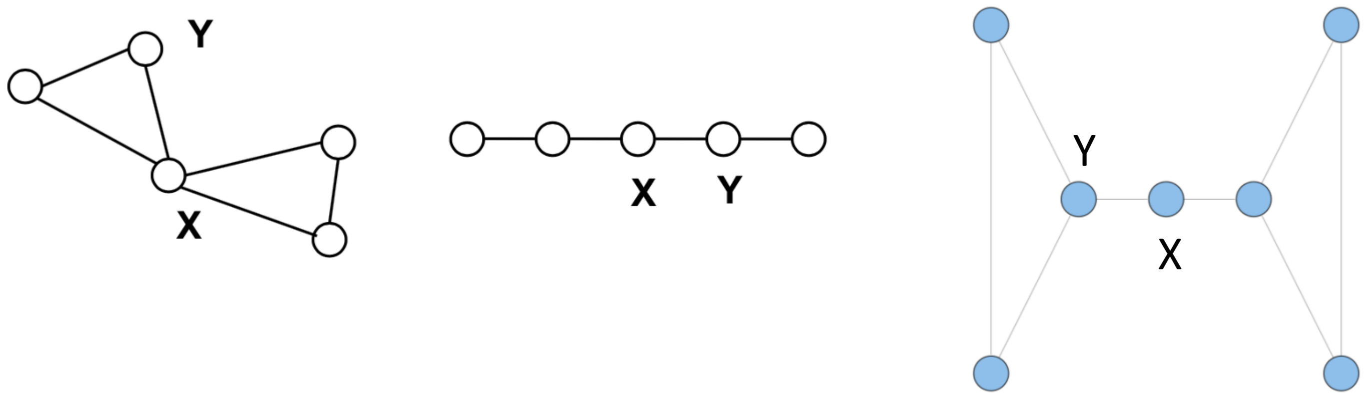
Example of bridge



- Algorithms to identify bridges (also called brockers) are all based on some measure of the **graph connectivity**.

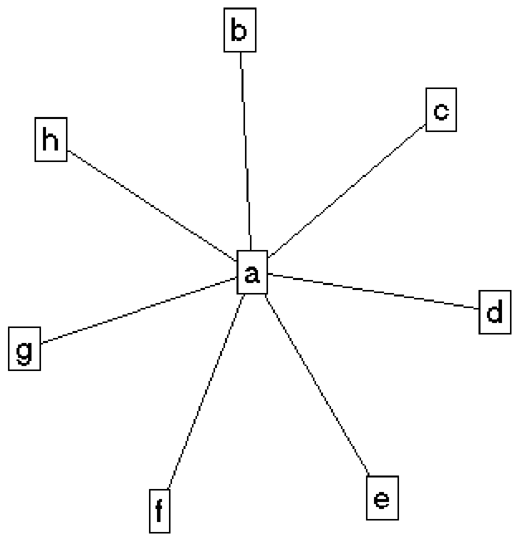
Betweenness: intuition

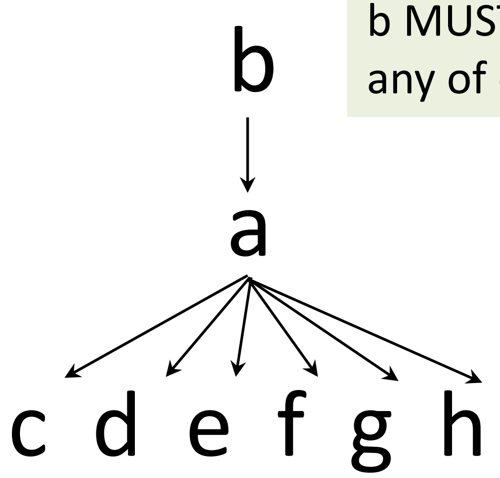
- Intuition: how many pairs of individuals would have to go through you in order to reach one another in the minimum number of steps?
- Who has higher betweenness, X or Y in these 3 examples?



Formally: Betweenness Centrality

Betweenness centrality counts the number of geodesic paths between *i* and *k* **that actor *j* resides on**. Geodesics are defined as the **shortest path** between points





b MUST go through a to reach any of c,d,e,f,g,h (and viceversa)

Any among b,c,d,e,f,g,h must go through a to reach any other

Betweenness Centrality

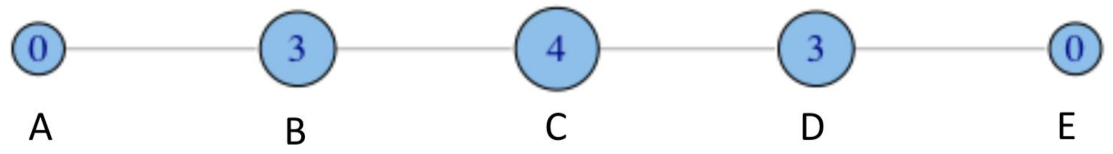
$$C_B(n_i) = \sum_{j<k} g_{jk}(n_i) / g_{jk}$$

Where g_{jk} = the number of geodesics (shortest) connecting *jk*, and g_{jk} (ni)= the number of such paths that node *i* is on (count also in the start-end nodes of the path).

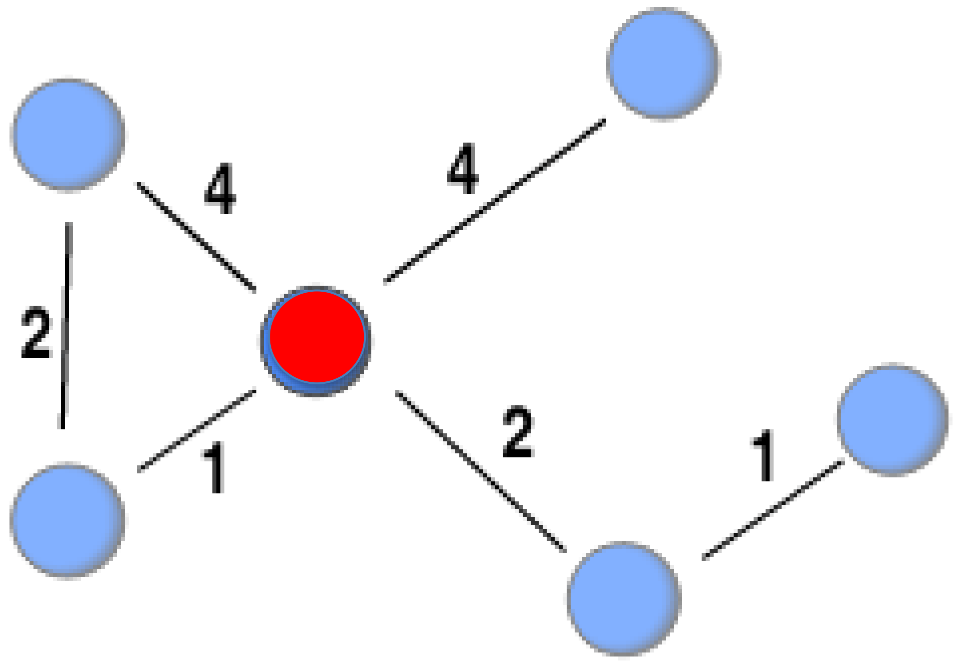
Can also compute **edge betweenness** in the very same way

Example of betweenness computation

- A lies between no two other vertices (betweenness is 0)
- B lies between A and 3 other vertices: C, D, and E (so any information from A to C,D,E or viceversa must flow trough B: betweenness is 3)
- C lies between 4 pairs of vertices (A,D),(A,E), (B,D),(B,E): betweenness is 4



Example of computation (bridgeness of the red node)



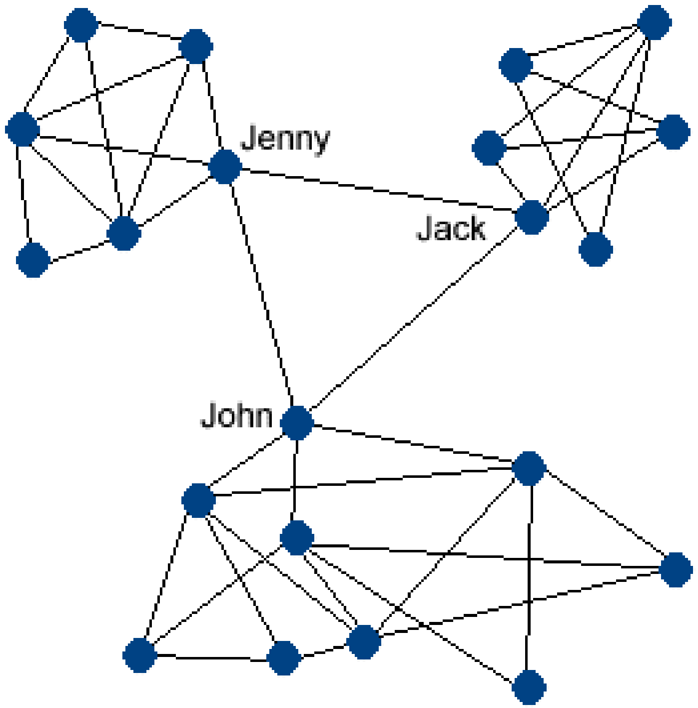
Many other measures of bridgeness

- Betweeness centrality, like centrality degree, is a local measure (based only on path counts)
- More sophisticated algorithms are available, based on the notion of graph connectivity
- The intuition is: what if we remove a node from the network? The highest the "damage" in term of connectivity, the highest the bridgeness value of the node

Finding Bridgeness /brokerage

Good bridges = actors that are indispensable for the flow of communication within the network

- As for graph representation, good bridges are actors that, <u>if removed from the graph</u>, **reduces graph connectivity.** For example, it causes the creation of disconnected components (*Jenny*, *Jack* and *John* in the graph)
- This is why bridges are also called brokers or key separators



Other graph-based social measures

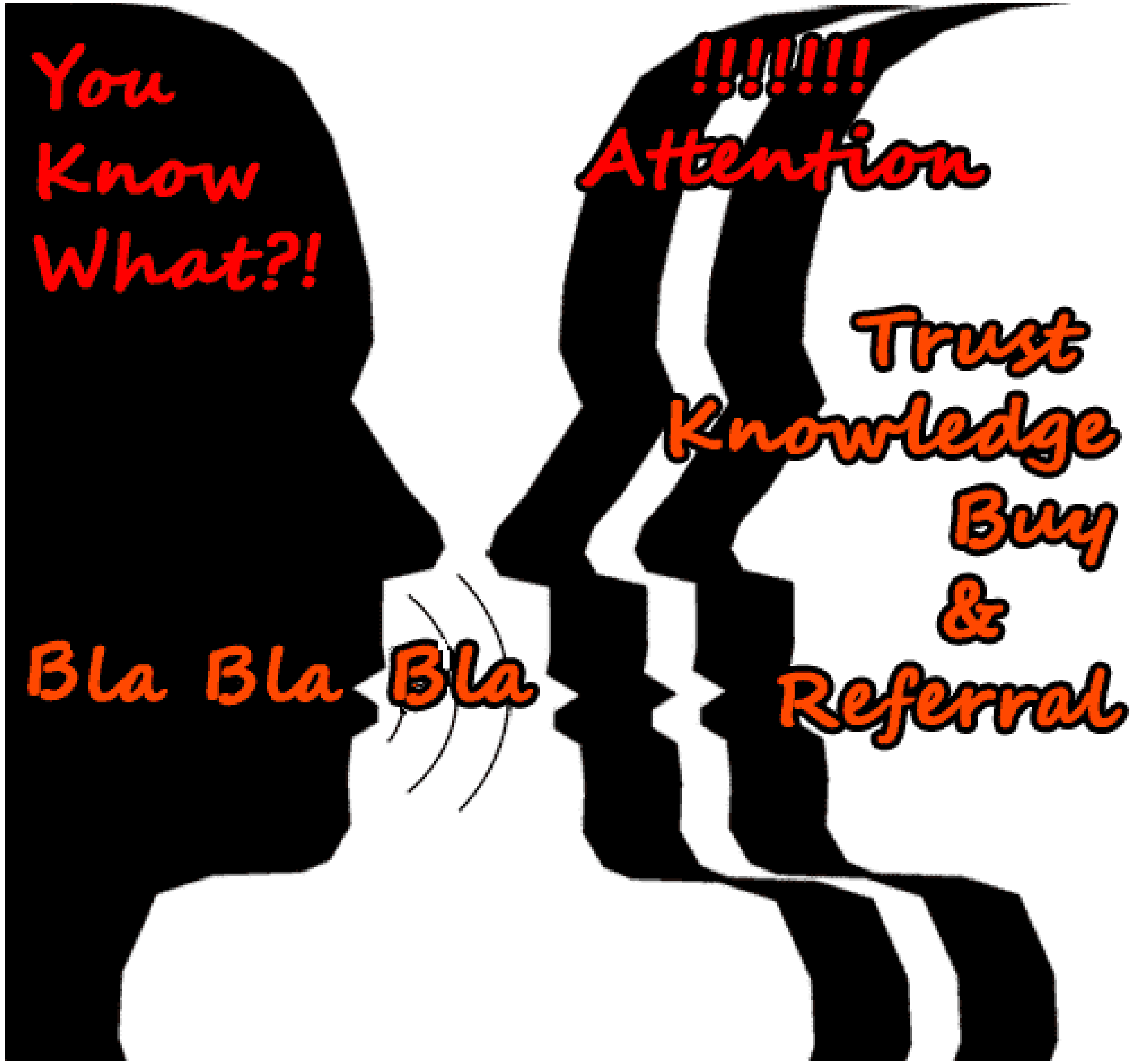
- Besides identifying key players (centrality, bridgeness) other types of information are relevant, and they require a global analysis of the network, not just single nodes
- E.g., If we wish to measure the likelyhood that an information originated anywhere in the network will reach you (**spread of influence**)
- Or, if we want to identify sub-groups (**communities**) within a network

Graph-based measures of social influence

- Use graph-based methods/algorithms to identify “relevant players” in the network
  - Relevant players = more influential, according to some criterion
- **Use graph-based methods to analyze the “spread” of information**
- Use graph-based methods to identify global network properties and communities (community detection)

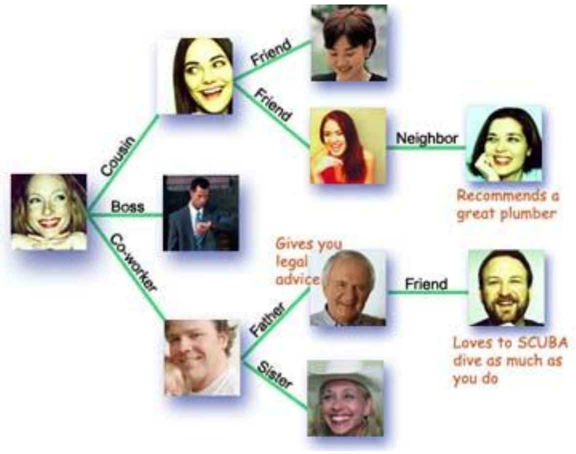
Influence Spread

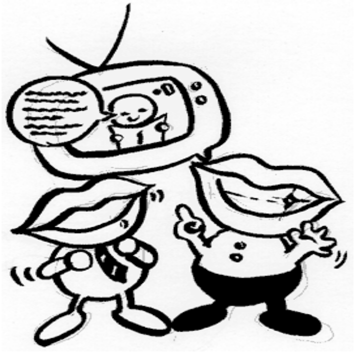
- We live in communities and interact with our friends, family and even strangers.
- In the process, we influence each other.



Social Network and Spread of Influence

- Social network plays a fundamental role as a medium for the spread of INFLUENCE among its members
  - Opinions, ideas, information, innovation…





- Direct Marketing takes the "word-of-mouth" effects to significantly increase profits (Gmail, Tupperware popularization, Microsoft Origami …)

Social Network and Spread of Influence

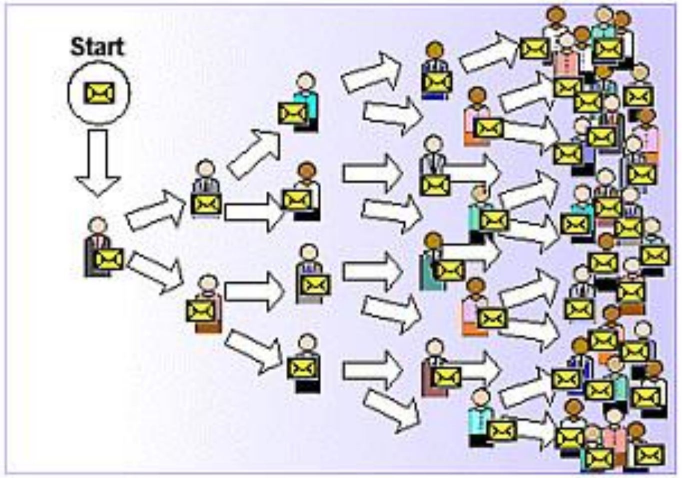
- Examples:
  - A company selects a small number of customers and ask them to try a new product. The company wants to choose a small group with largest influence.
  - Obesity grows as fat people stay with fat people (homofily relations)
  - Viral Marketing..

Viral Marketing

Identify influential customers

Convince them to adopt the product – Offer discount/free samples

These customers endorse the product among their friends



Spread of Influence analysis: Problem Setting

- Given
  - a limited budget B for initial advertising (e.g. give away free samples of product)
  - estimates for influence between individuals
- Goal
  - trigger a large cascade of influence (e.g. further adoptions of a product)
- Question
  - Which set of individuals should B target at?
- Application besides product marketing
  - spread an innovation
  - detect stories in blogs (gossips)
  - Epidemiological analysis

What we need

- Models of influence in social networks.
- Obtain data about particular network (to estimate inter-personal influence).
- Algorithms to maximize spread of influence.

A simple algorithm

- Linear Threshold Model (only the intuition..)
- The basic model implies that each actor is influenced by those he/she is linked to
- The influence depends on the **strength** of the relation between two actors (contagiousness)
- It also depends on the **personal tendency of an actor to be influenced** by others (resistance)

Linear Threshold Model

- A node *v* has some threshold θ_v *~ U[0,1] (this models the "tendency to be influenced": the higher the threshold (resistance), the lower is the influence of others on an person's opinion)*
- A node *v* is influenced by each neighbor *w* according to a *weight* b_{vw} such that

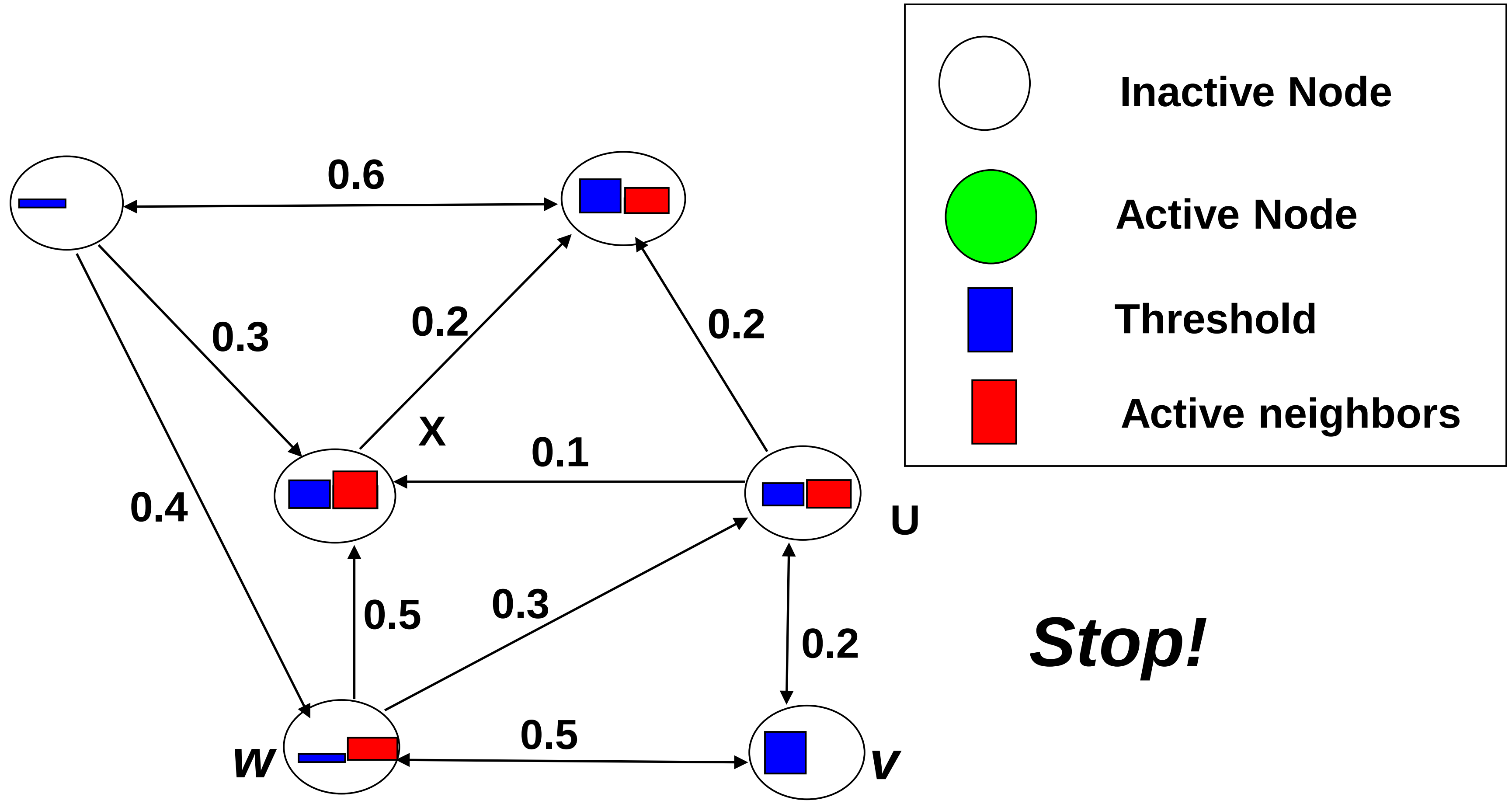
$$\sum_{w \text{ neighbor of } v} b_{v,w} \leq 1$$

- $b_{v,w}$ models the "strength of the relation" of actor v on actor w
- A node *v* becomes **active** when at least (weighted) θ_v fraction of its neighbors are active

$$\sum_{w \text{ active neighbor of } v} b_{v,w} \geq \theta_v$$

- We assume a **cumulative** effect of neighbours' influence on an actor!

Example (weights on edges are the $b_{u,v}$)



Stop!

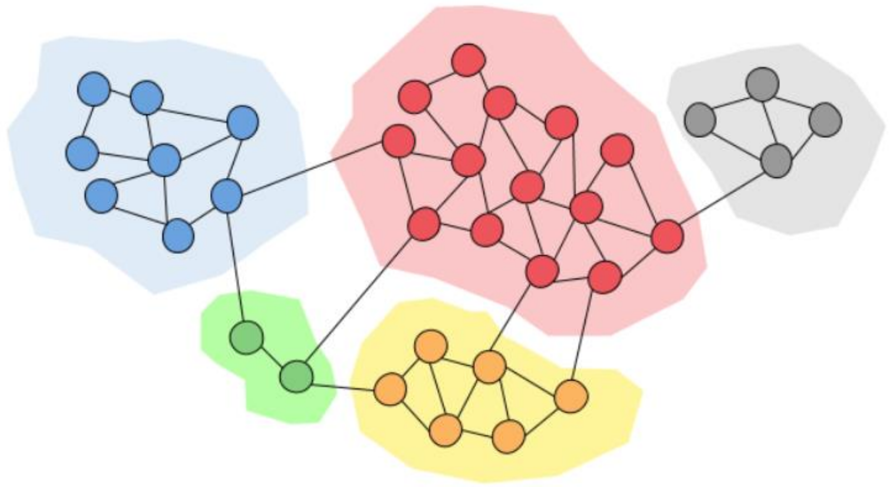
Influence Maximization Problem

- Problem:
  - Given a parameter *k* (**budget**), find a ***k*-node set *S*** to maximize f(S)
  - In simpler terms: find the minimum number of influencer to “reward”, given the budget, which maximizes the number of individuals that can be “influenced” (through a cascade process of influence propagation”
  - Several algorithms (you don’t need to learn..)

Graph-based measures of social influence

- Use graph-based methods/algorithms to identify “relevant players” in the network
  - Relevant players = more influential, according to some criterion
- Use graph-based methods to analyze the “spread” of information
- **Use graph-based methods to identify global network properties and communities (community detection)**

Community detection

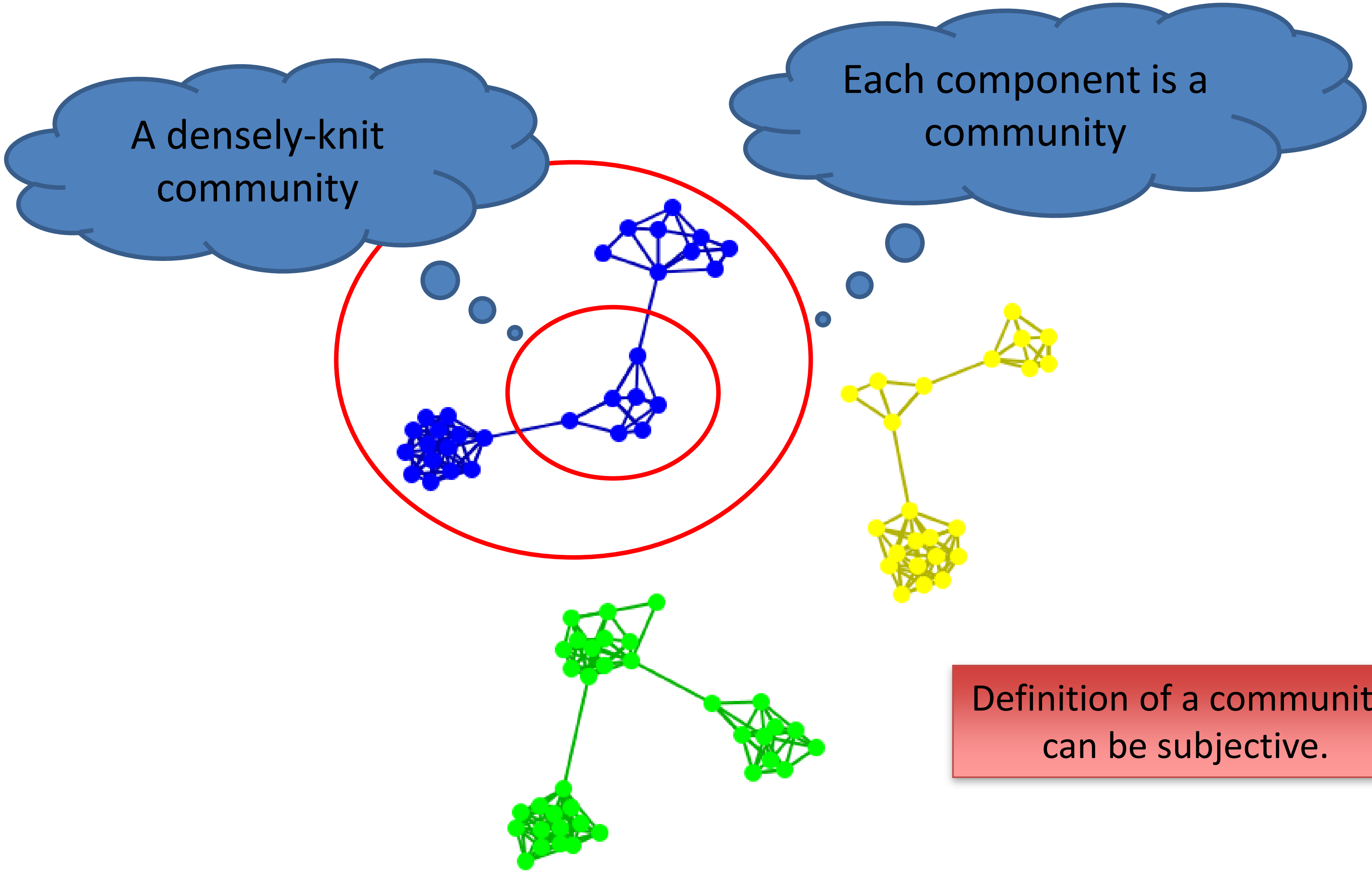


- Community: It is formed by individuals such that those within a group interact with each other **more frequently than with those outside the group** a.k.a. group, cluster, cohesive subgroup, module in different contexts
- Community detection: discovering groups in a network where individuals' group memberships are not explicitly given

Community detection

- Why communities in social media?
  - Human beings are social
  - Easy-to-use social media allows people to extend their social life in unprecedented ways
  - Difficult to meet friends in the physical world, but much easier to find friend online **with similar interests**
  - <u>Interactions</u> between nodes can help determine communities

Subjectivity of Community Definition



Definition of a community can be subjective.

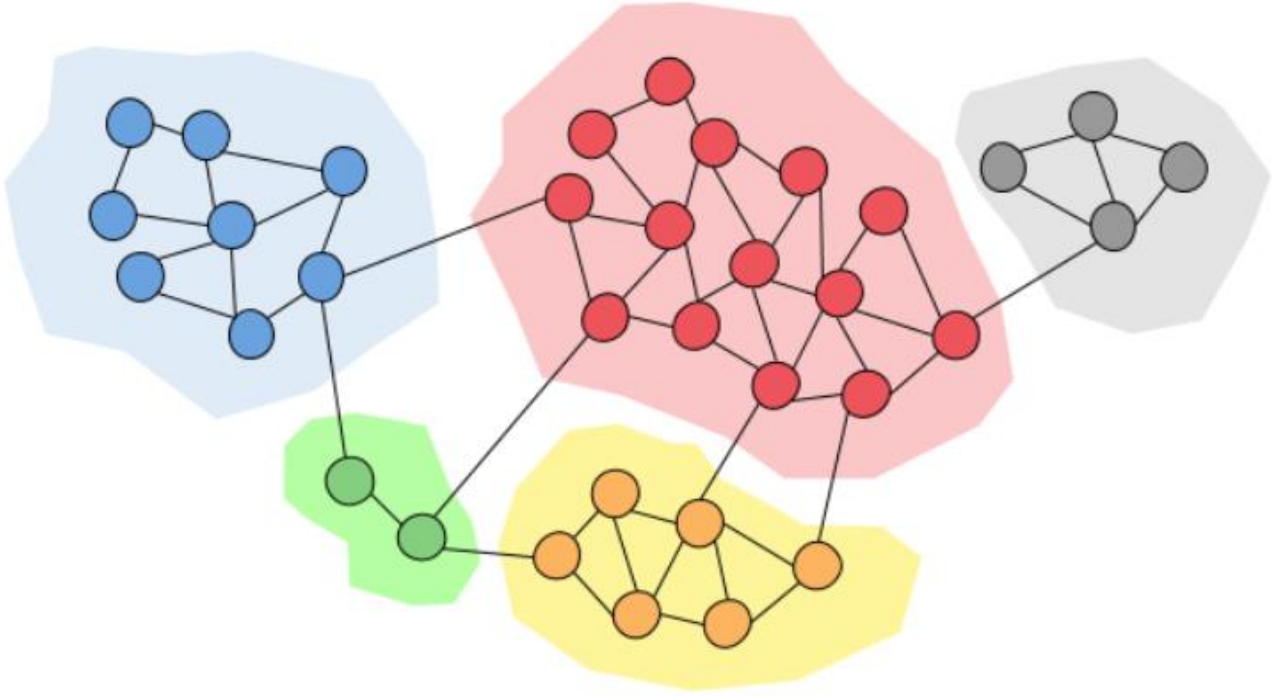
Community detection = clustering

The two problems are very similar

And share the same complexity

Community detection in graphs

- Many methods based on ML
- Example:
  - **Density-based**: find densest (with more connections) areas in the graph
  - **Spectral clustering**: find the edges that, if removed, maximally reduce the connectivity of the graph (where the connectivity is some precise quantitative measure)



Two different aspects of social networks

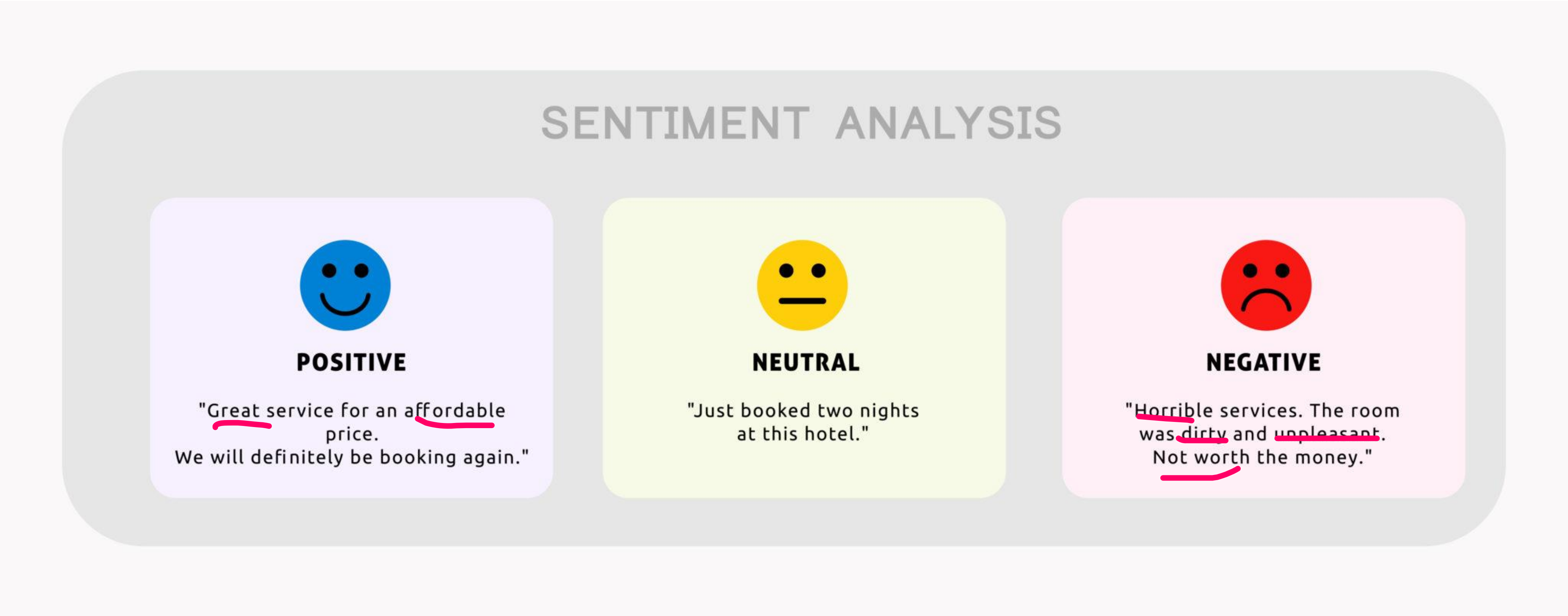
- **Social network analysis**: Analyze how the information propagates through the network. Based on network science.
  - Information diffusion
  - Influence maximization

- **Social media mining**: Analyzing the information (messages) that users exchange. Based on ML and Natural Language Processing (NLP) algorithms.
  - Opinion mining
  - Topic mining

Sentiment Analysis and Opinion Mining

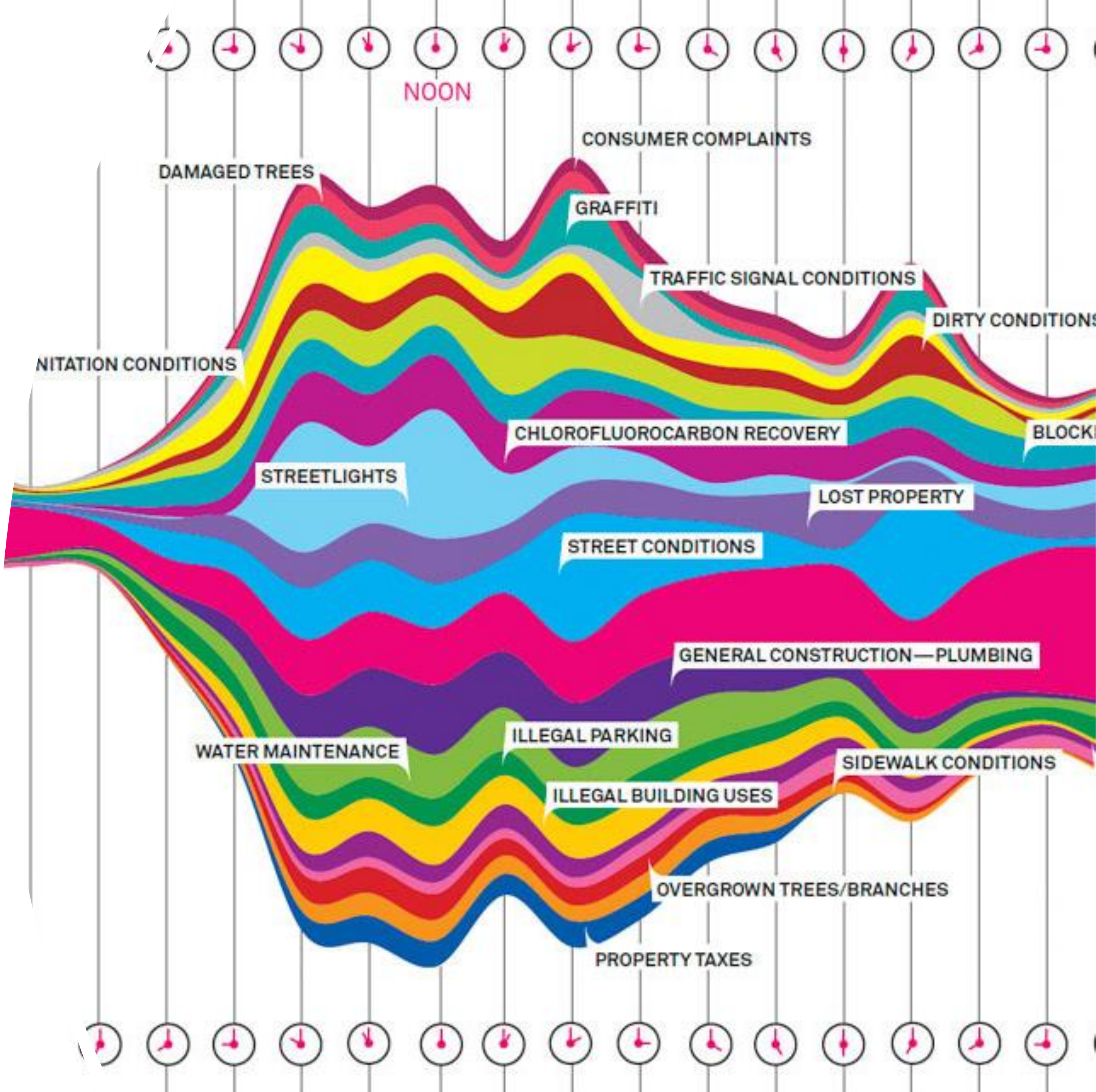
- It is a kind of classification task: given a text, determine its polarity (positive, negative, neutral)
- Based on machine learning methods for NLP

Sentiment analysis example



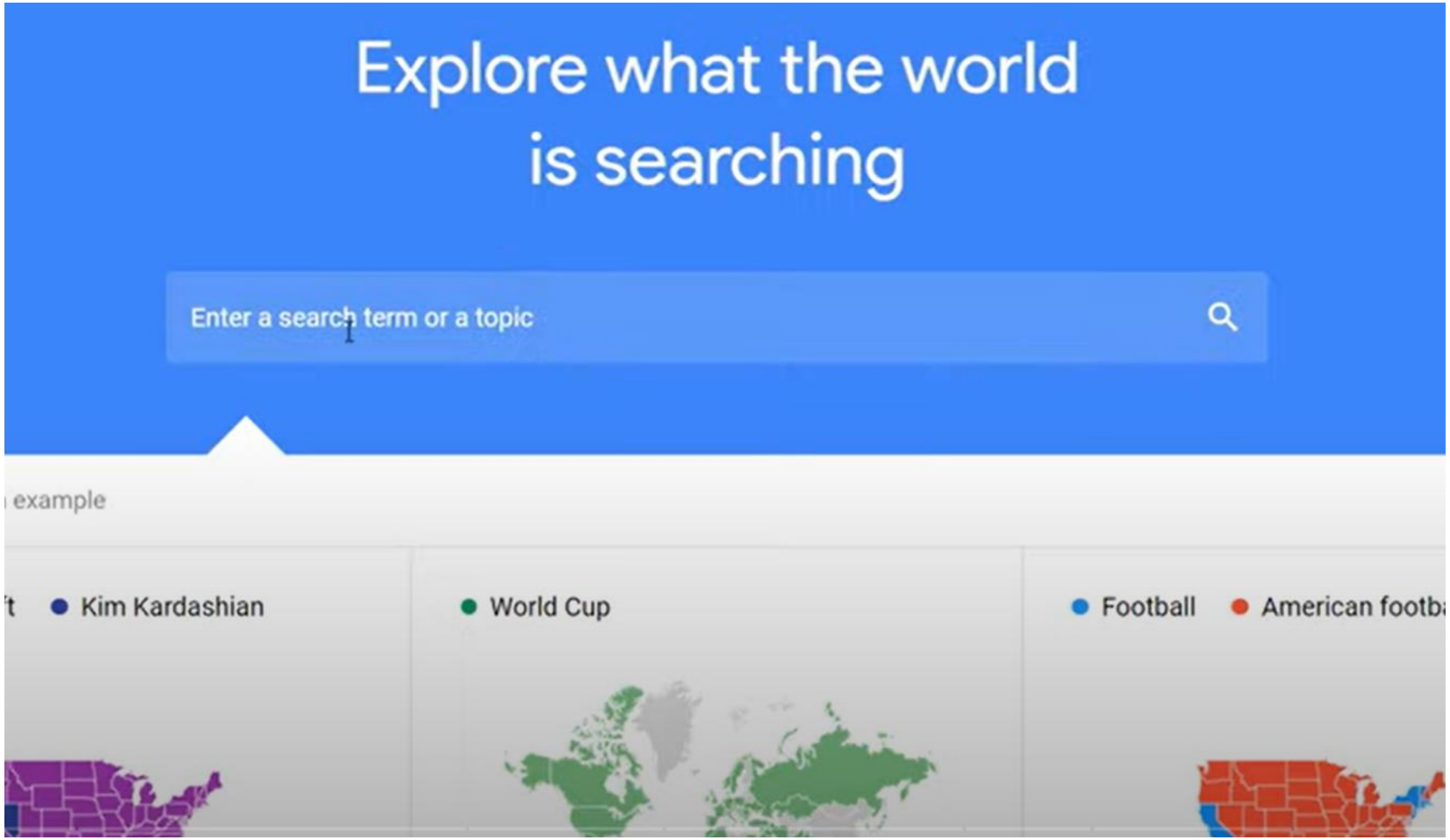
Topic mining

- Based on NLP techniques. Extract the relevant discussed topics in a social network and they popularity trends
- Example: citizen messages on an city council app to report problems

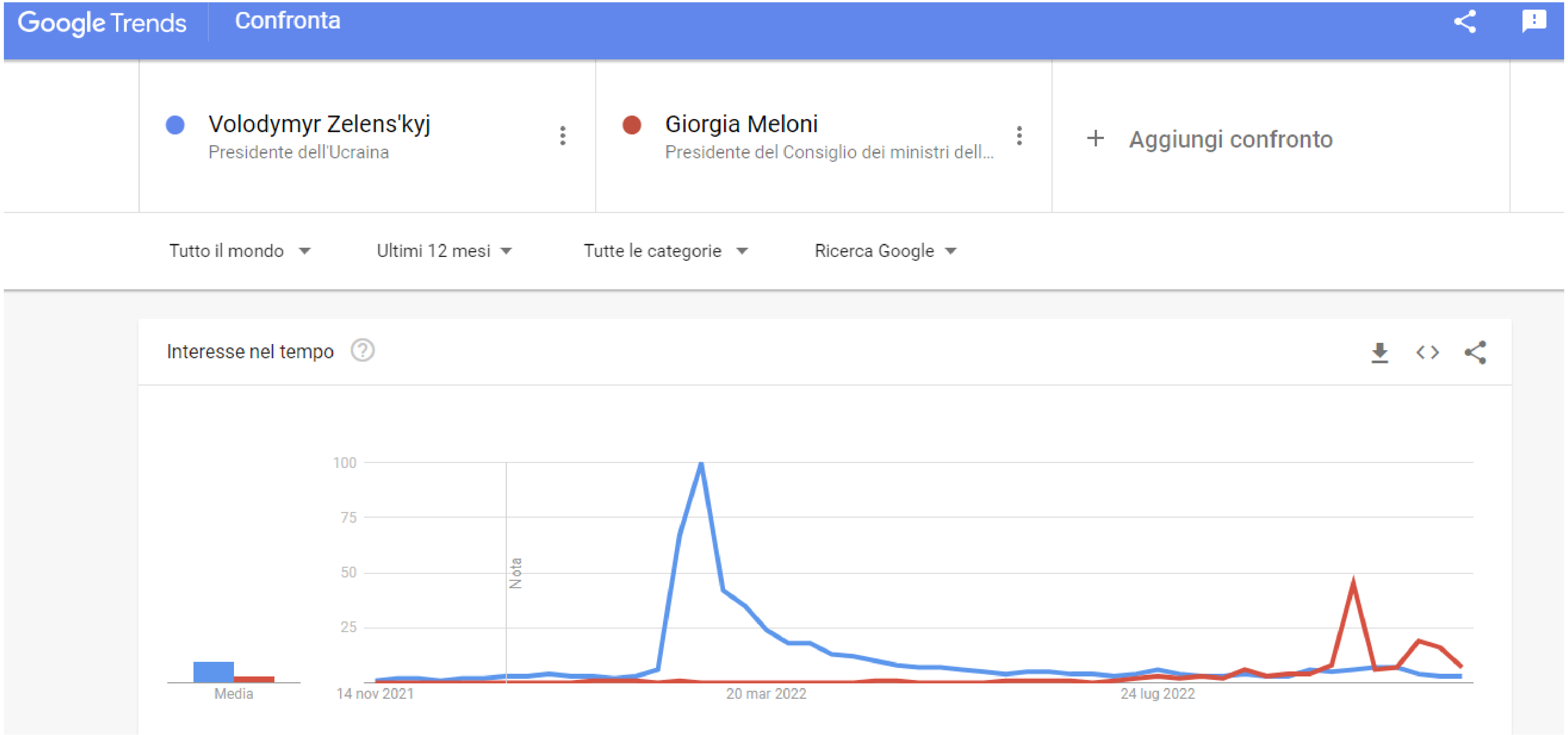


Google Trends: tool to analyse topic trends

Exploring topics

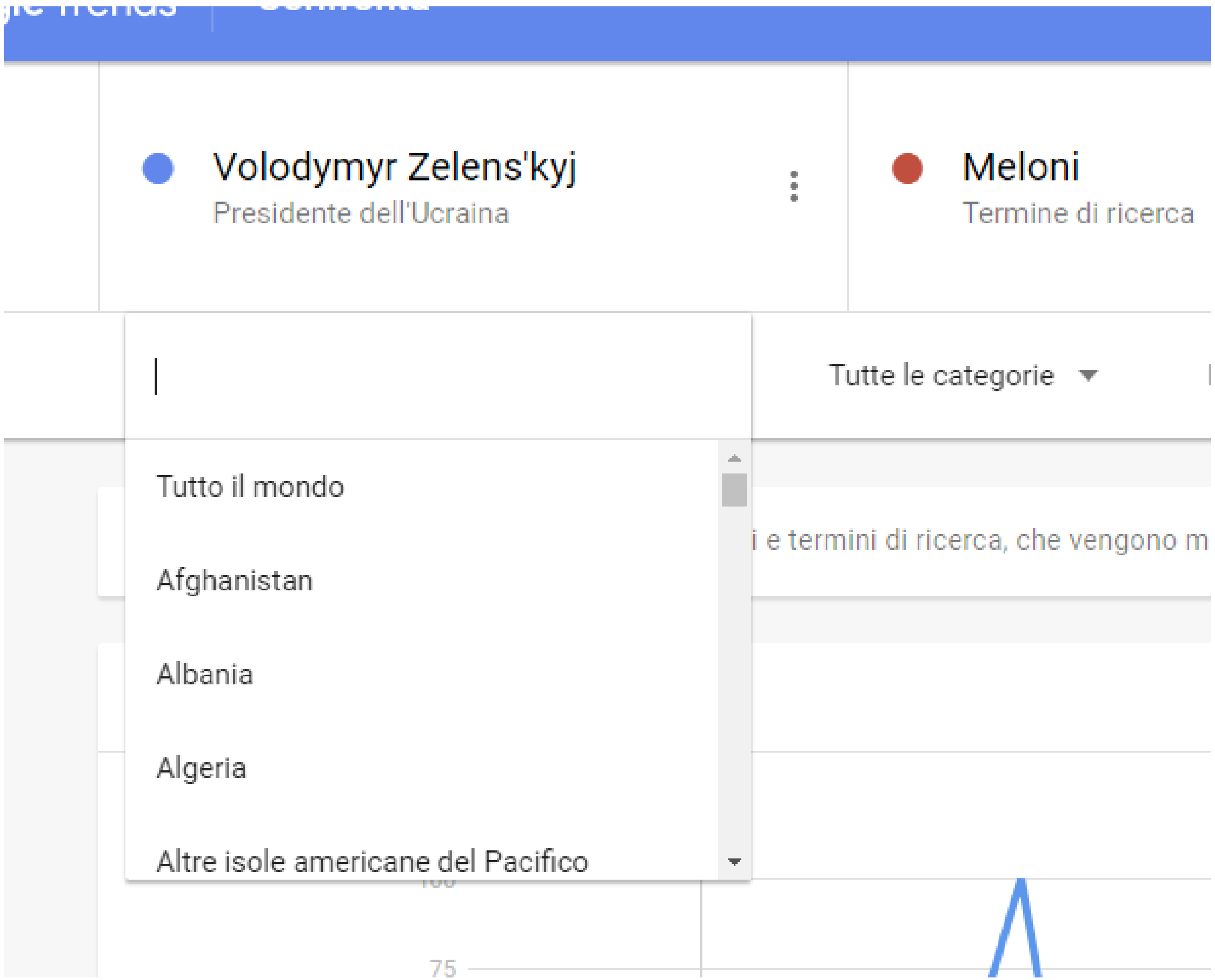


Comparing topics

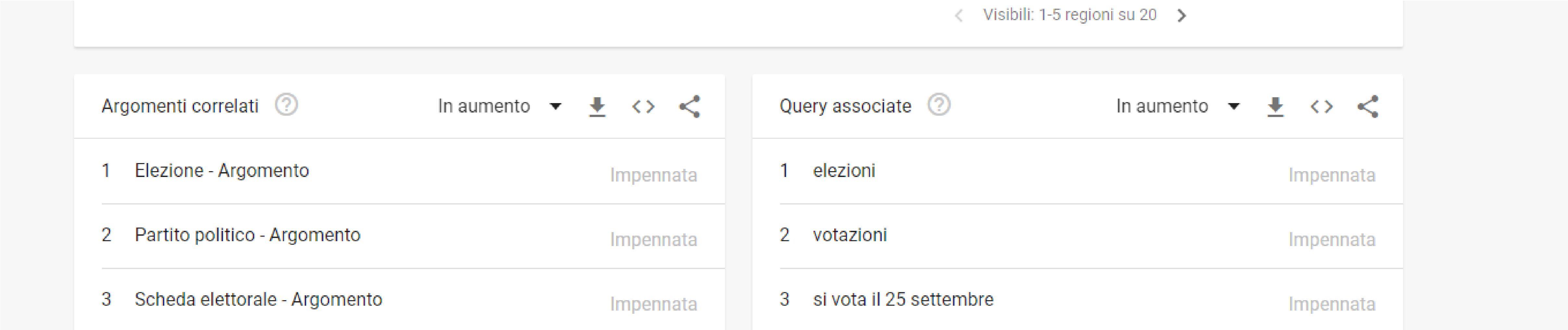


Find peaks

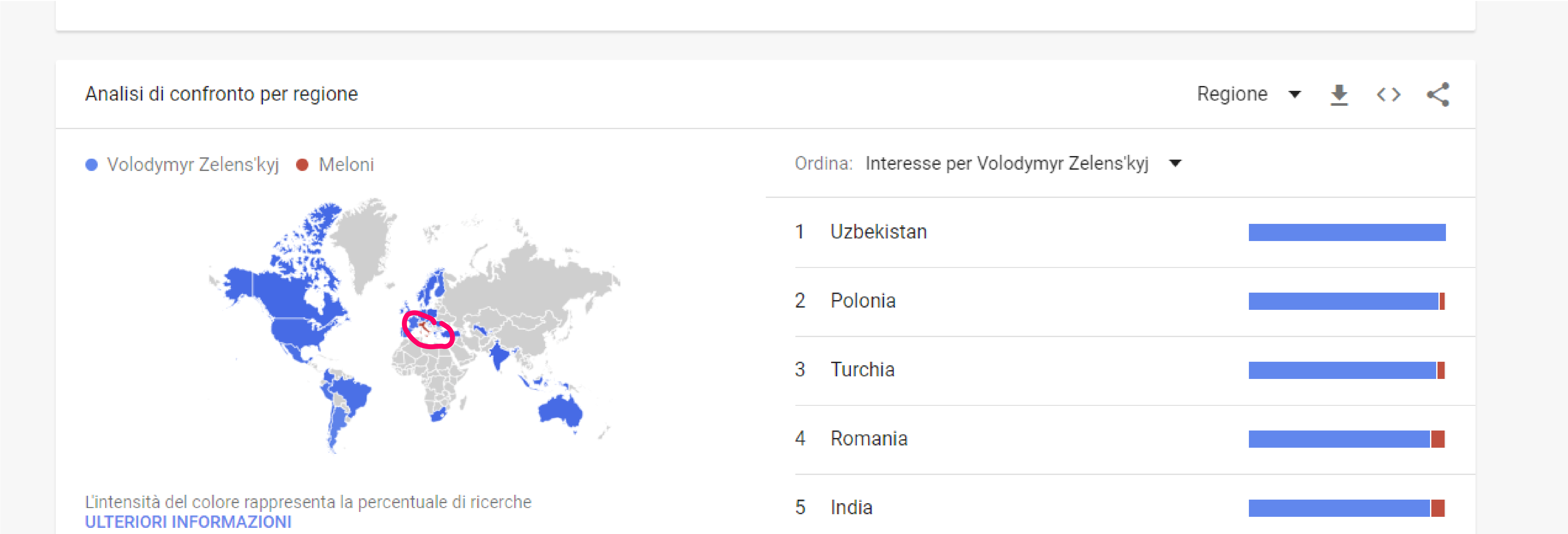
Adjust by region



Understand peaks



Regional distribution



In class exercise

- Use Google trends to explore popular topics and trends on the web and social media
- Consider a number of brands or products
- Find and compare relevant topics
- Find regions where topics are more popular
- Explore popularity over time
- If you find peaks, use web serch with dates to identify the relevant news