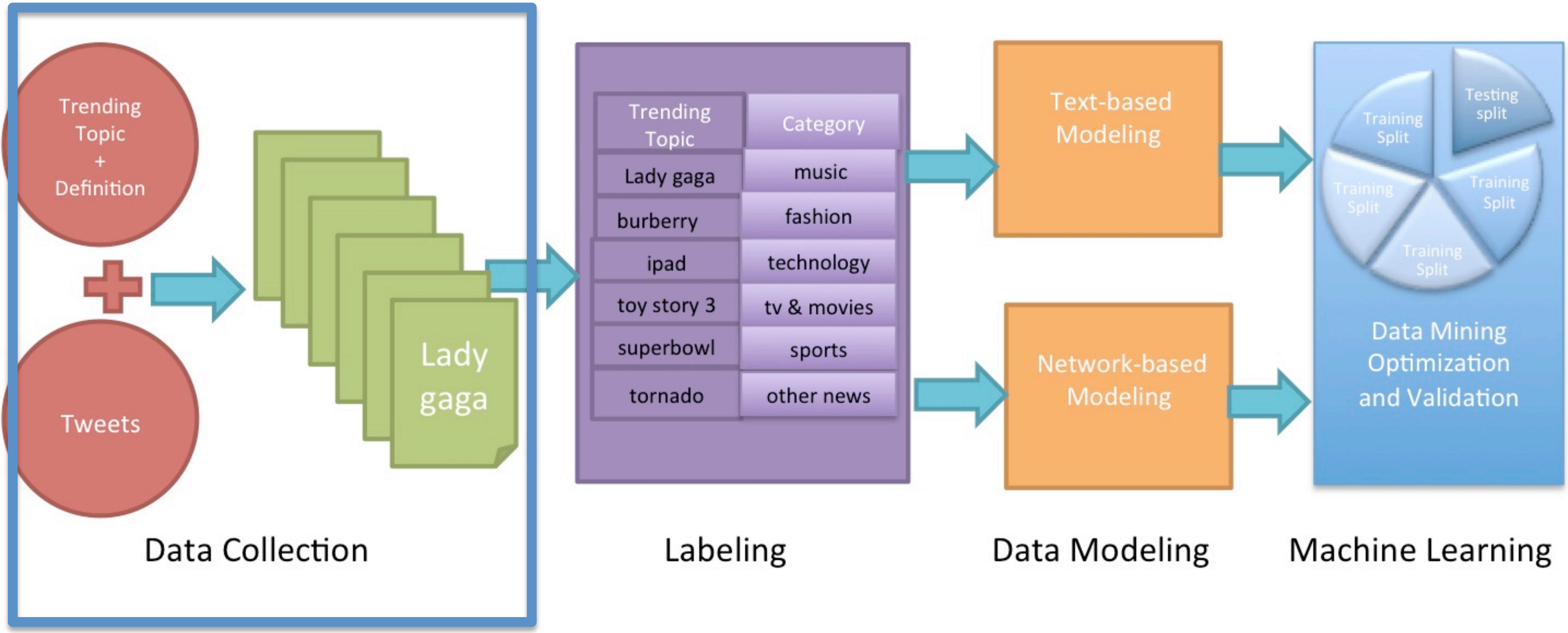
Social Analytics for BI

PART B

Extracting Social Network data for BI: workflow



Trending Topic	Category
Lady gaga	music
burberry	fashion
ipad	technology
toy story 3	tv & movies
superbowl	sports
tornado	other news

Data collection: sources and methods

- **Social network media:** access to comprehensive historic data sets and also real-time access to sources (Ex: Twitter, facebook, blogs..)
- **News data:** access to historic data and real-time news data sets, possibly through data licenses (cf. software license). Examples: Google news, Reuters press agency releases..
- **Public data**: access to scraped and archived important public data; available through RSS feeds, blogs or open government databases.
- **Programmable interfaces**: access to simple application programming interfaces (APIs) made available by resource owners to **scrape*** and store other available data sources that may not be automatically collected (e.g. the content of web sites) .

scraper: a program that periodically downloads the content (or selected parts) of a web page

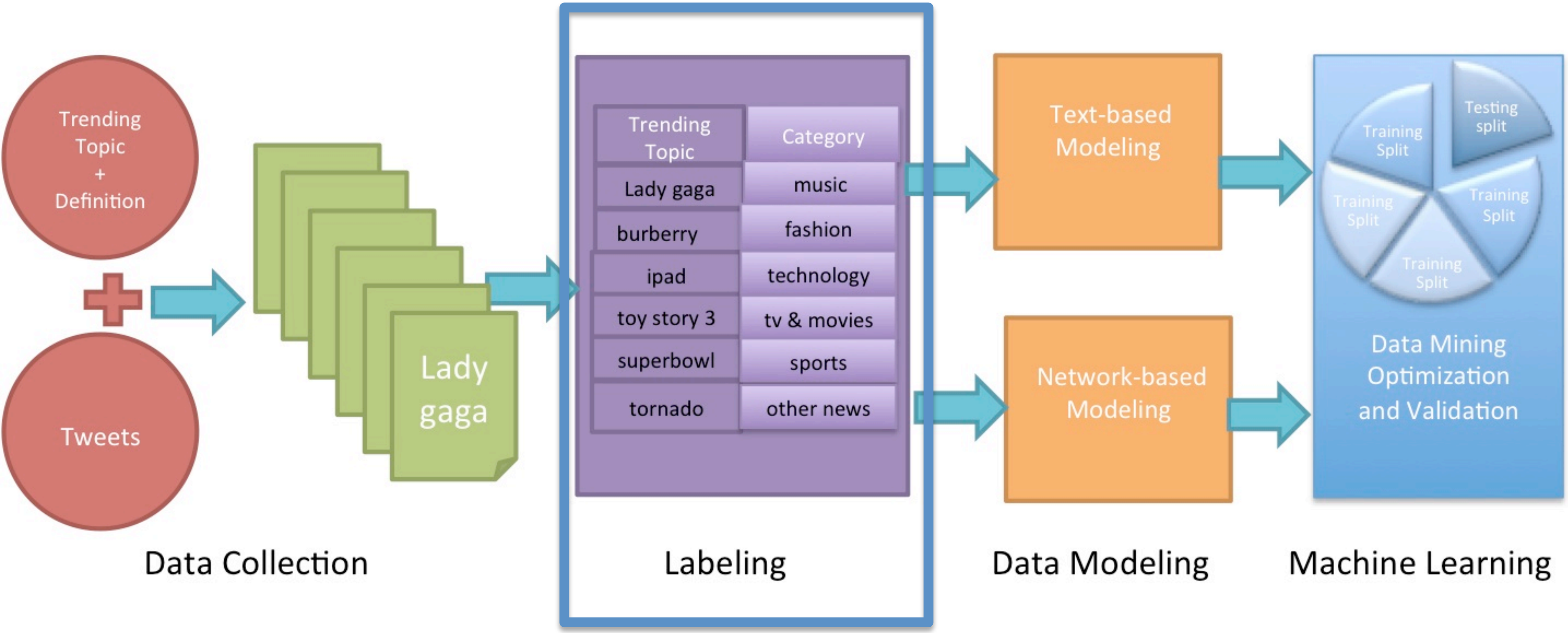
Social media providers

Freely available databases —can be freely downloaded, e.g., Wikipedia (http://dumps.wikimedia.org) and the Enron e-mail data set (http://www.cs.cmu.edu/*enron/) on emails exchanged within Enron

Data access via tools—sources that provide controlled access to their content/data. An example is Google Trends. Further subdivided into:

- Free sources—*repositories* that are freely accessible, but the tools protect or may limit access to the ‘raw’ data in the repository, such as the range of tools provided by Google (in G.Trends, you have trends, not numbers..).
- Commercial sources —data resellers that *charge for access* to their social media data. Gnip and DataSift provide commercial access to Twitter data through a partnership, and Thomson Reuters to news data.
- Data access via APIs —social media data repositories providing *programmable HTTP-based access to the data via APIs* (e.g., Twitter, Facebook and Wikipedia).

Extracting Social Network data for BI: workflow



Trending Topic	Category
Lady gaga	music
burberry	fashion
ipad	technology
toy story 3	tv & movies
superbowl	sports
tornado	other news

Labelling (tagging) social data

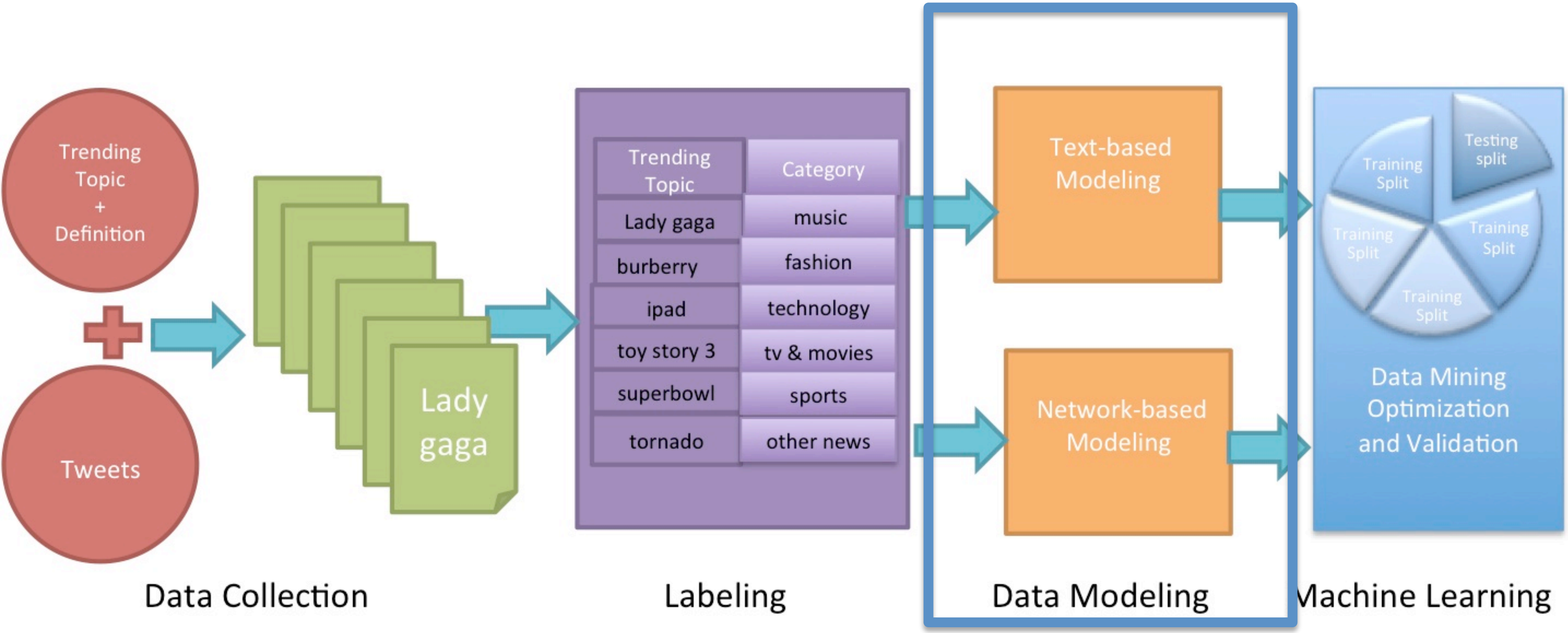
Preparing social data for storage

- Social media data is generated by humans and therefore is unstructured (i.e., it lacks a pre-defined structure or data model), algorithms are required to transform it **into structured data** before storage (data analytics need structured databases).
- Social data need then to be **preprocessed, tagged and then parsed** in order to analyze it.
- Adding extra information to the data (e.g, tagging the data with *sentiment* indicators, or with a *topic* indicator –e.g. music, politics..)- can be performed manually or via programs, which seek patterns or interpret the data using techniques such as data mining and text analytics.
- Algorithms exploit the linguistic, auditory and visual structure inherent in all of the forms of human communication.

Methods for social data processing

- **Computational statistics**—refers to computationally intensive statistical methods including resampling methods, Markov chain Monte Carlo methods, local regression, kernel density estimation and principal components analysis.
- **Machine learning**—algorithms for the autonomous acquisition and integration of knowledge learnt from experience, analytical observation, etc.
- **Natural Language Processing** — algorithms for part of speech tagging, syntactic analysis, semantic analysis

Extracting Social Network data for BI: workflow



Trending Topic	Category
Lady gaga	music
burberry	fashion
ipad	technology
toy story 3	tv & movies
superbowl	sports
tornado	other news

Methods for Data Modeling: **network-based** and text-based

Text-based modeling

- Will be analyzed later (sentiment analysis, topic extraction..)

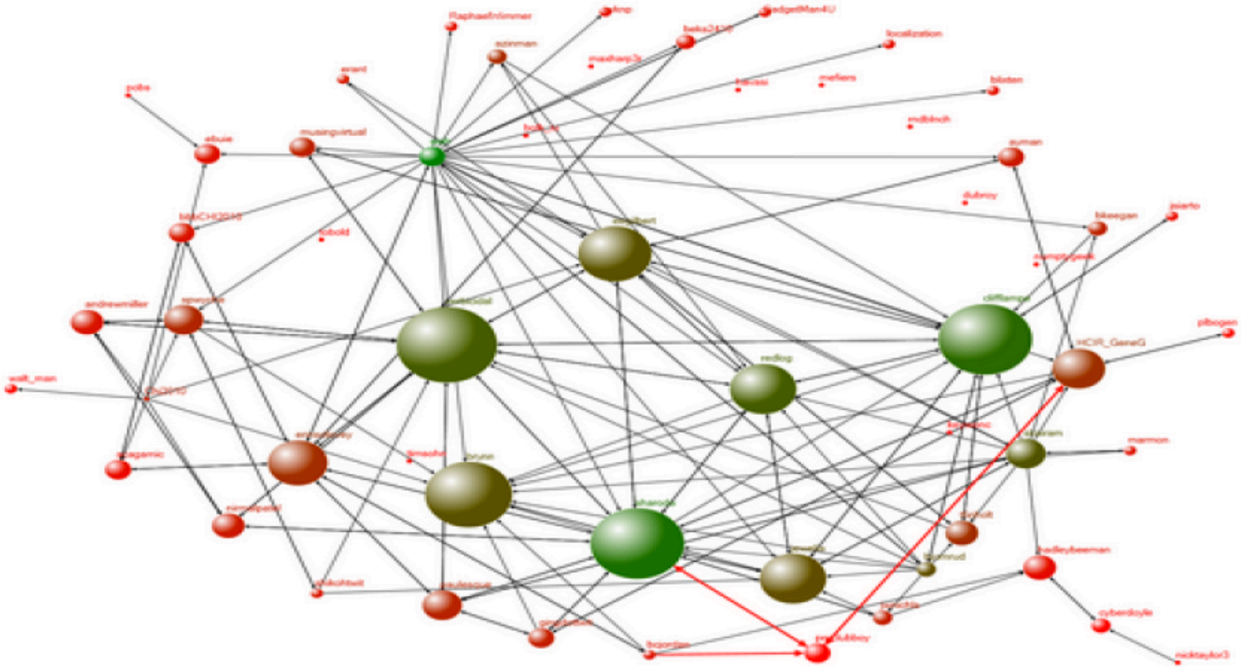
Network-based Modeling

- **Graph-based measures**: Based on the graph-structure of the network

Graph-based modeling

- Previously surveyed measures of influence, such as buzz, applause etc. are based on surface metrics (e.g. number of retweets, etc): graph-based measures go more in-depth.
- Objective here: **model the social network as a graph**
- 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 identify communities (community detection)
- Use graph-based methods to analyze (and predict) the “spread” of information

Modeling a Social Network as a graph



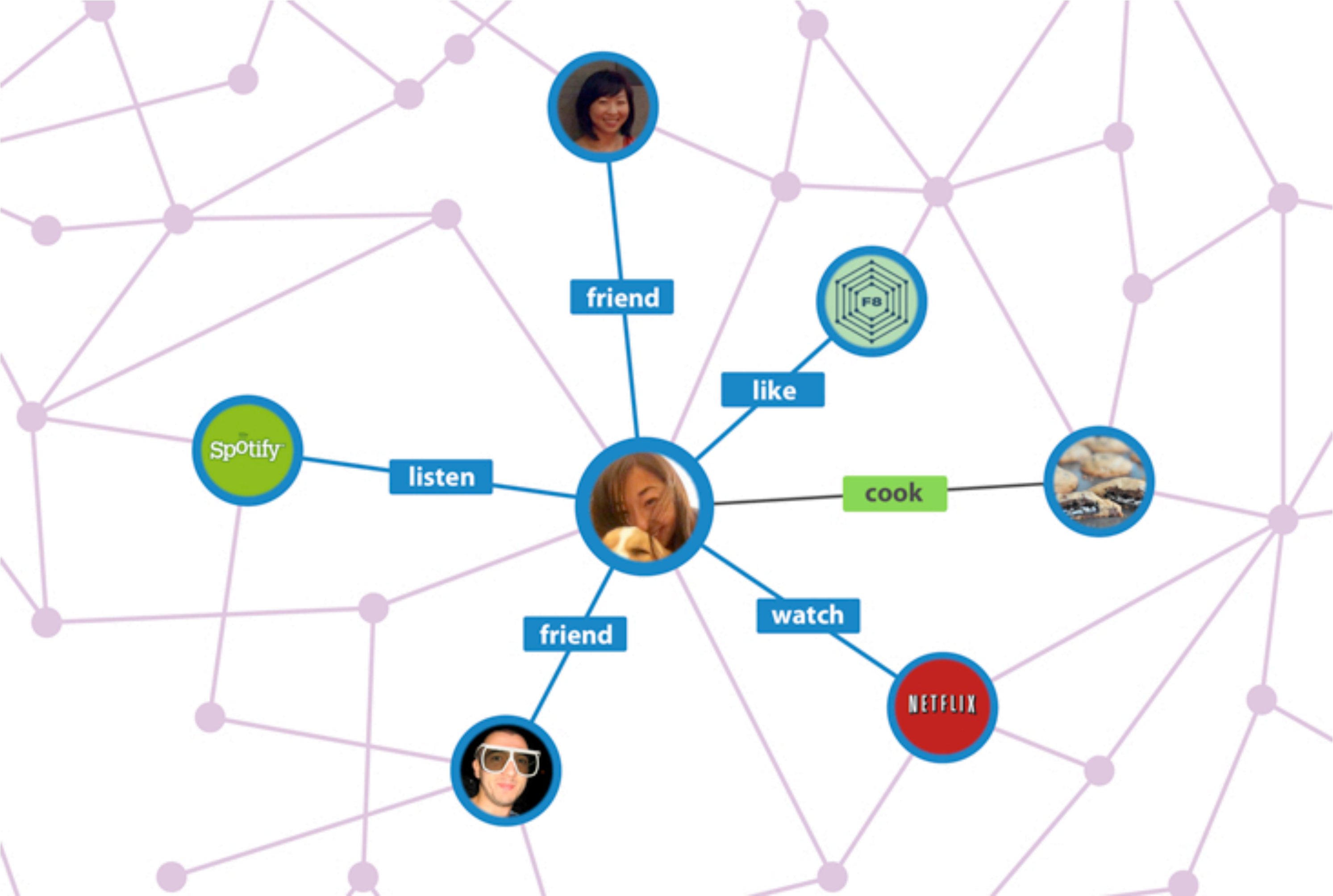
NODE= “actor, vertices, points” i.e. the social entity who participates in a certain network

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..)

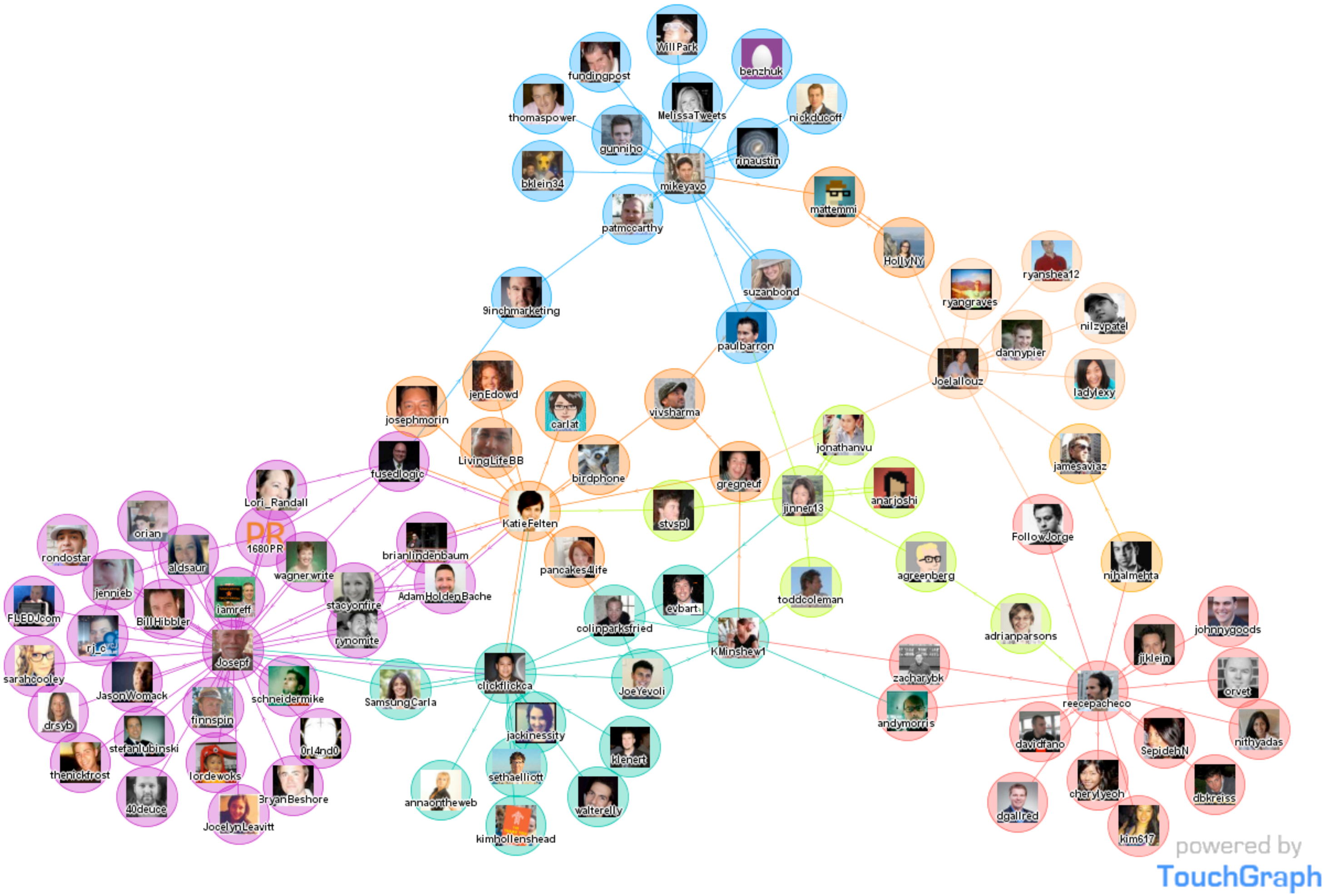
SN = graph

- A network can then be represented as a graph data structure
- We can apply a variety of measures and analysis 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)

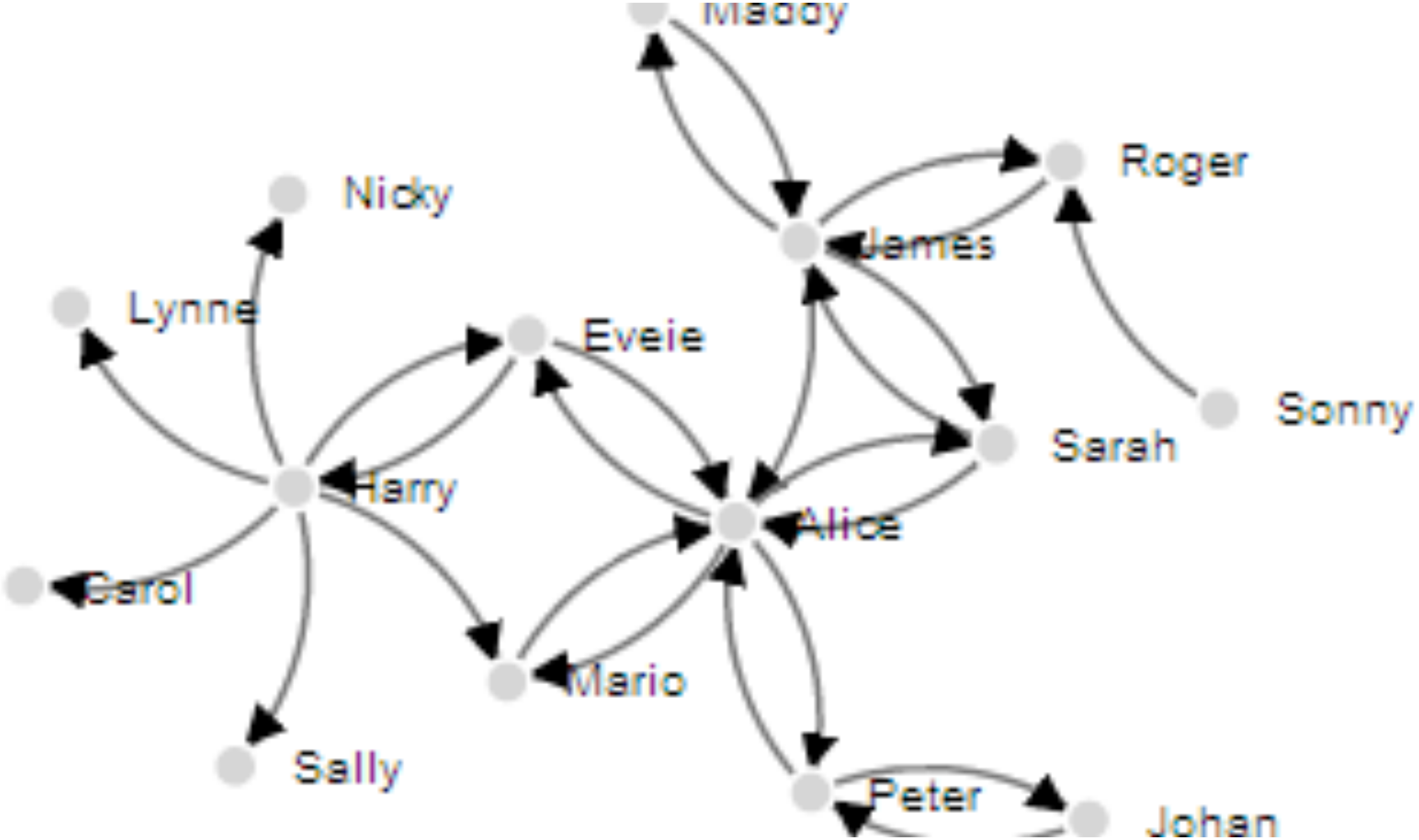
What is the meaning of edges?



Facebook in undirected (friendship is mutual)



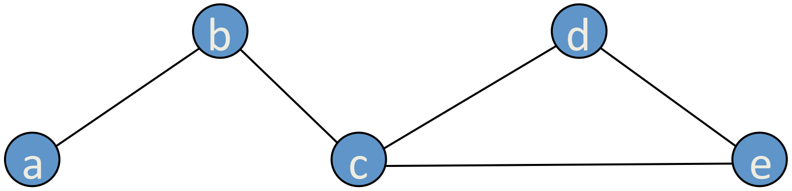
Twitter is a directed graph (friendship is not necessarily bidirectional)



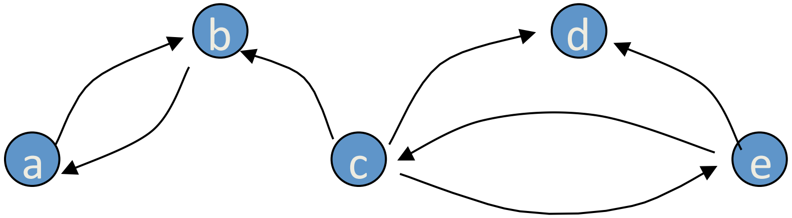
Social Network as a graph

In general, a relation can be:

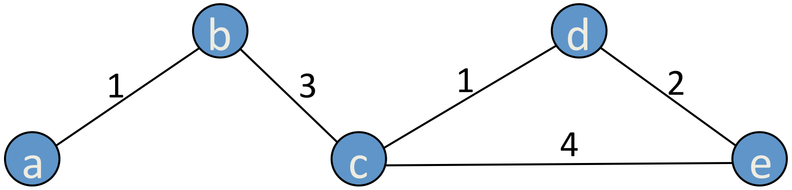
- Binary or Valued
- Directed or Undirected



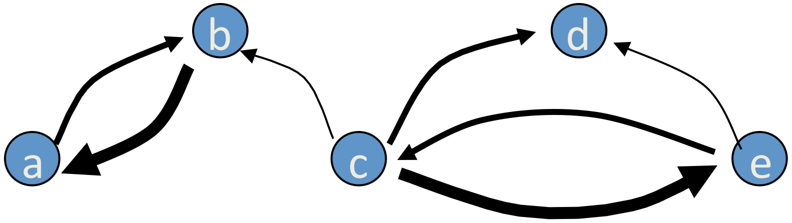
Undirected, binary



Directed, binary



Undirected, Valued

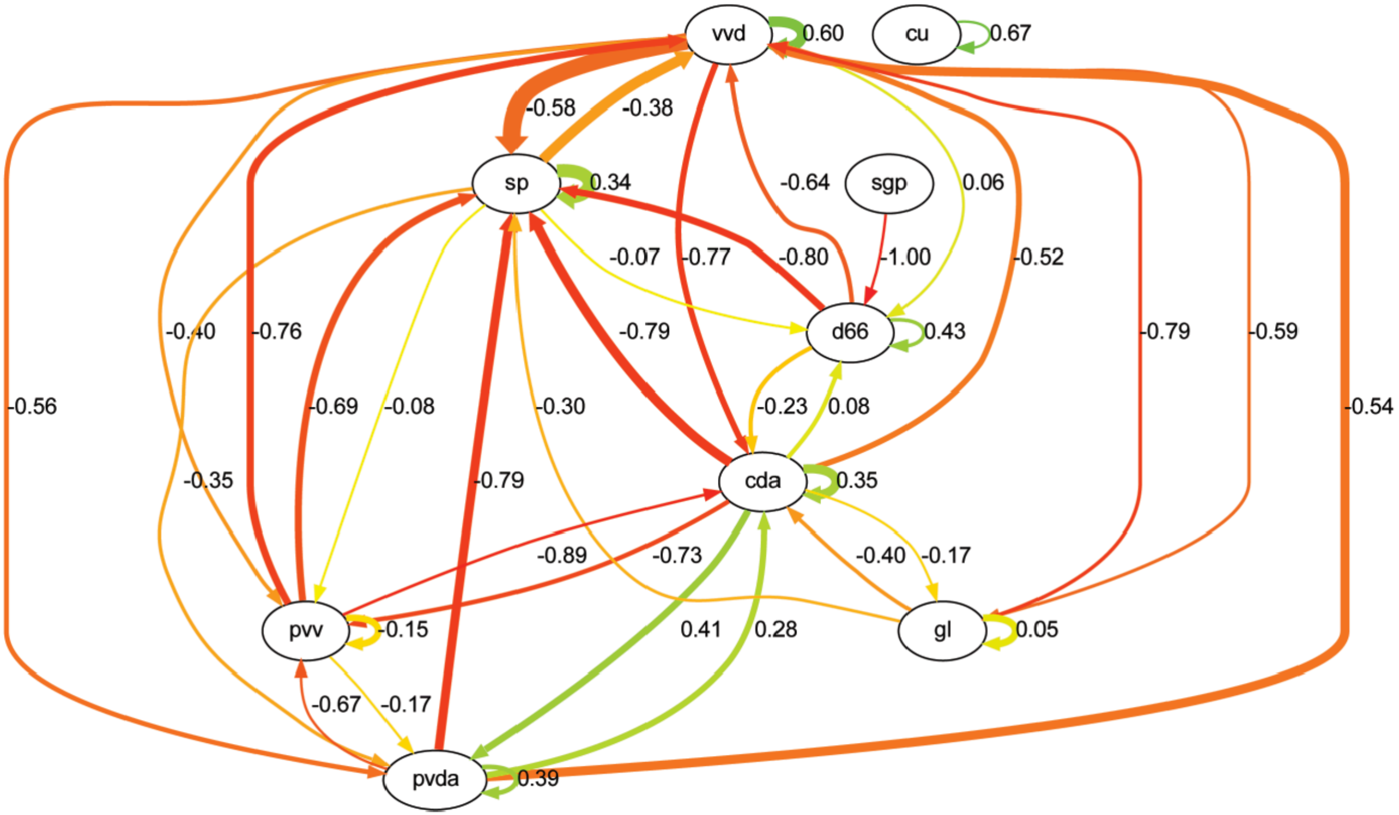


Directed, Valued

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**
- Use graph-based methods to analyze the “spread” of information
- Use graph-based methods to identify global network properties and communities (community detection)

Graph-based measures of social influence: key players

Key players

- Using graph 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)

Finding key players: Centrality

Finding out which is the most central node is important:

– It could help disseminating information in the network faster

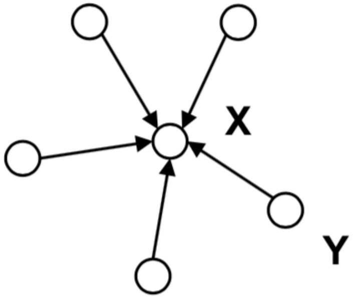
– It could help stopping epidemics

– It could help protecting the network from breaking

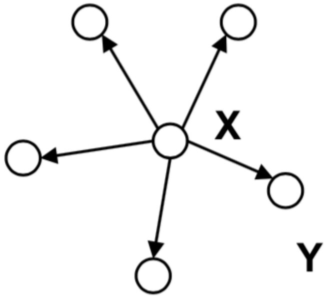
Centrality

- Conceptually, centrality is fairly straight forward: we want to identify which nodes are in the ‘center’ of the network.
- It has various meanings:

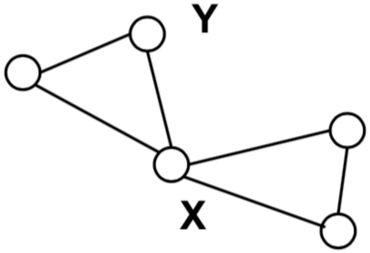
Centrality degree



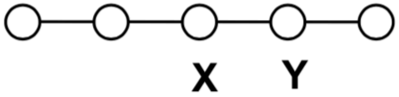
indegree



outdegree



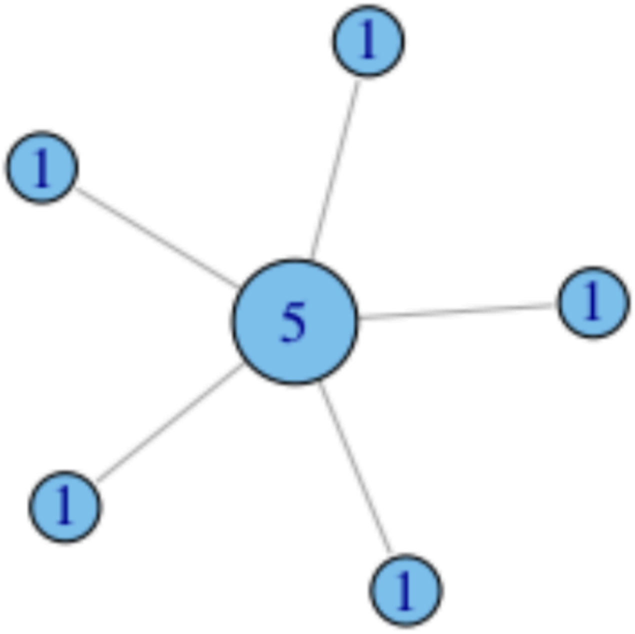
betweenness



closeness

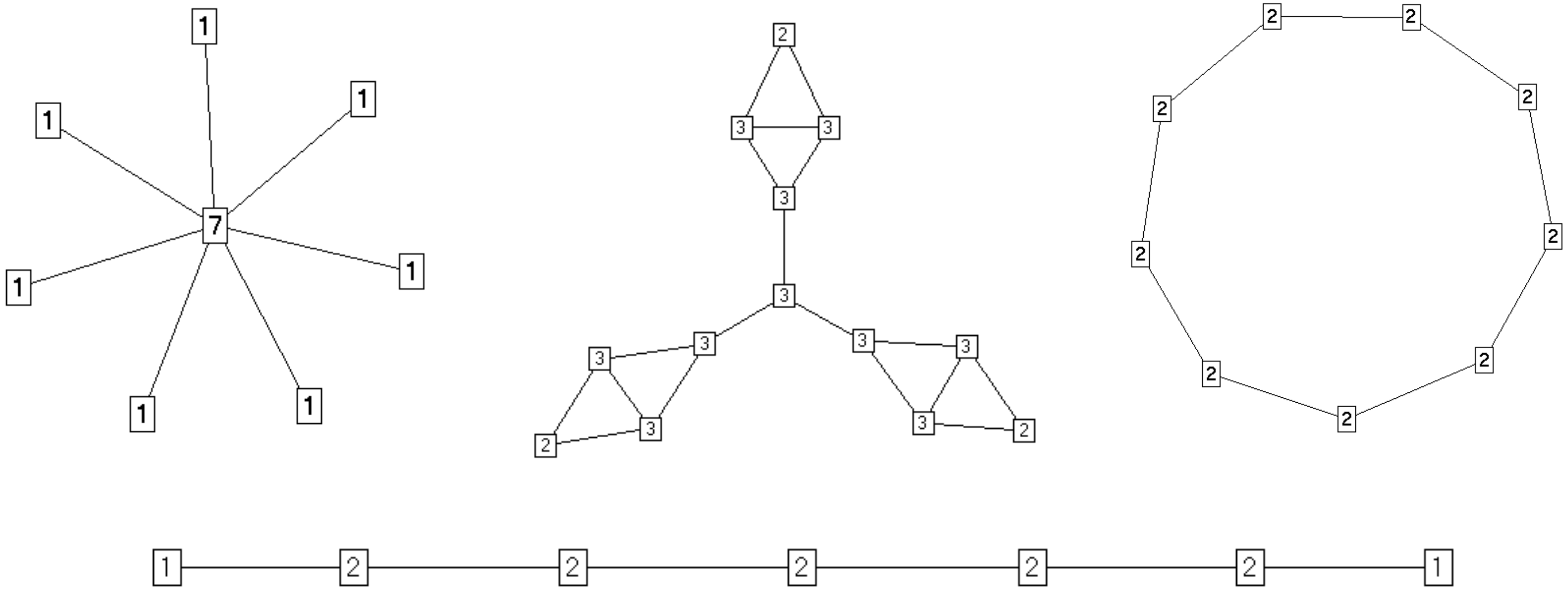
Centrality Degree = number of connections

- When is the number of connections the best centrality measure?
  - people who will do favors to you
  - people you can talk to / have a beer with



Centrality degree calculation examples

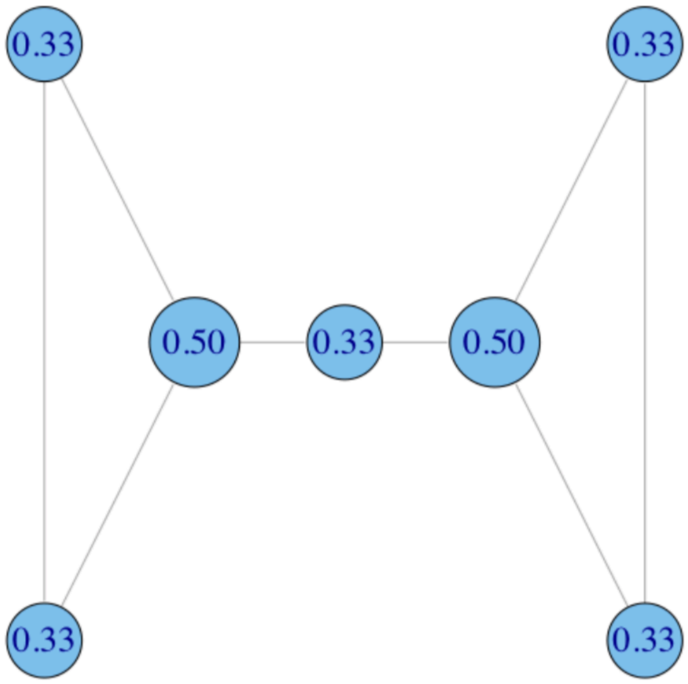
Degree is the number of ties, and the actor with the most ties is the most important:

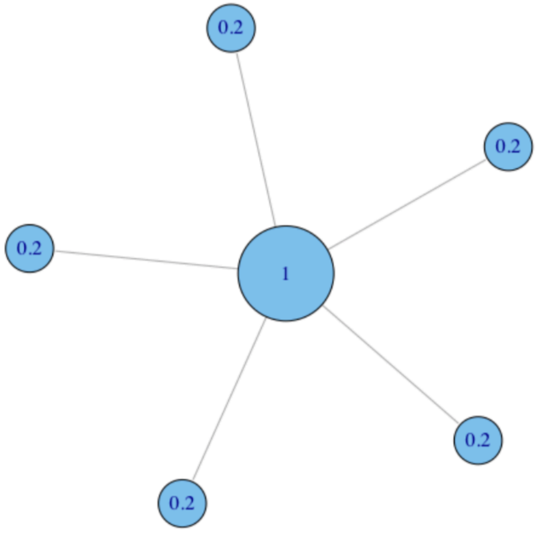


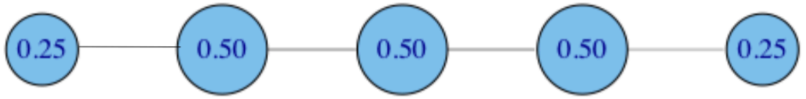
$$C_D = d(n_i) = X_{i+} = \sum_j X_{ij}$$

Centrality degree: normalization

Divide for the max number of nodes (N-1)





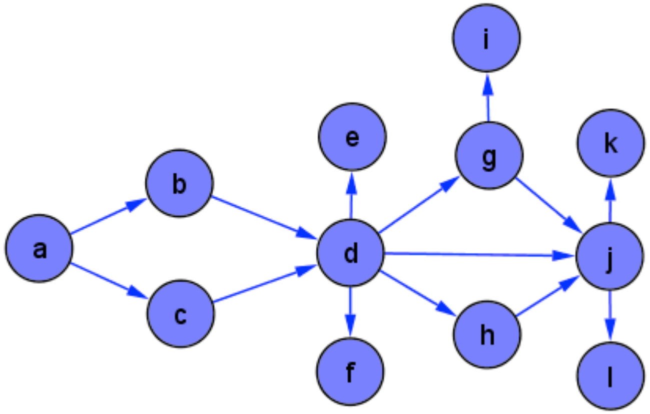


Centrality degree for directed networks (prestige)

- 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. in Yahoo! Answers).
  - An actor has high support, if a lot of people vote for him/her (many “likes”)

Measuring prestige: influence and support

- **Influence and support**: According to the direction/meaning of a relation, in and outdegree represent support or influence. (e.g., likes, votes for,. . .).



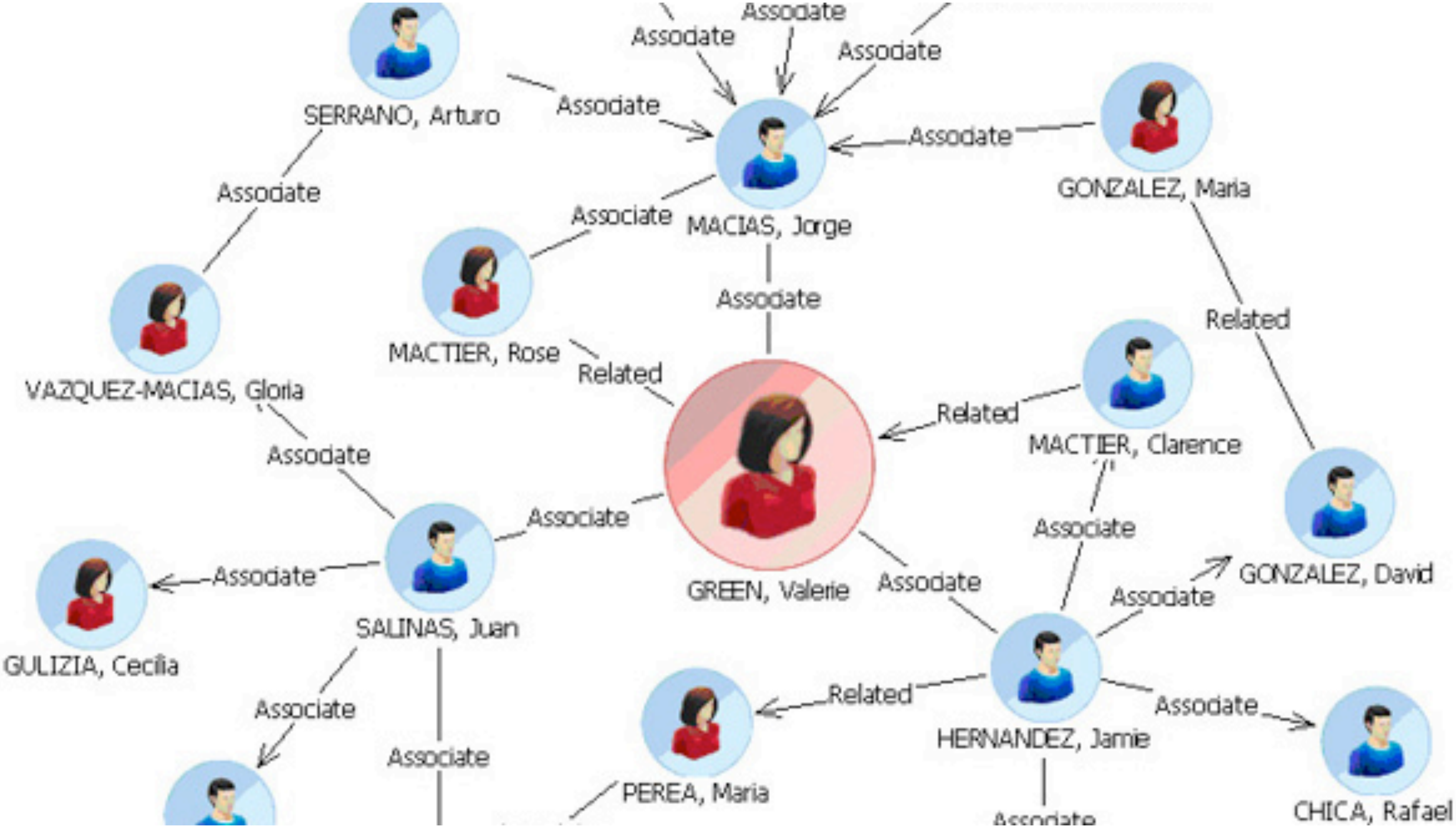
$$InDegree(x) = \#\ \ incoming\ \ edges(x)$$

$$InDegree^{N}(x) = \frac{\#\ \ incoming\ \ edges(x)}{\max_{y \in network}(InDegree^{N}(y))}$$

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 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

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

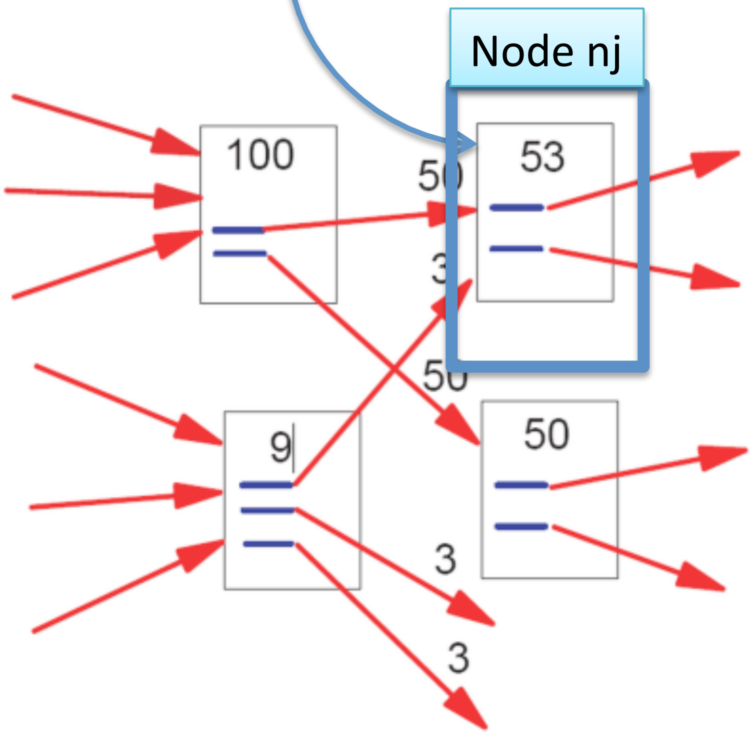
How does it work

$$PageRank\ of\ site = \sum \frac{PageRank\ of\ inbound\ link}{Number\ of\ links\ on\ that\ page}$$

OR

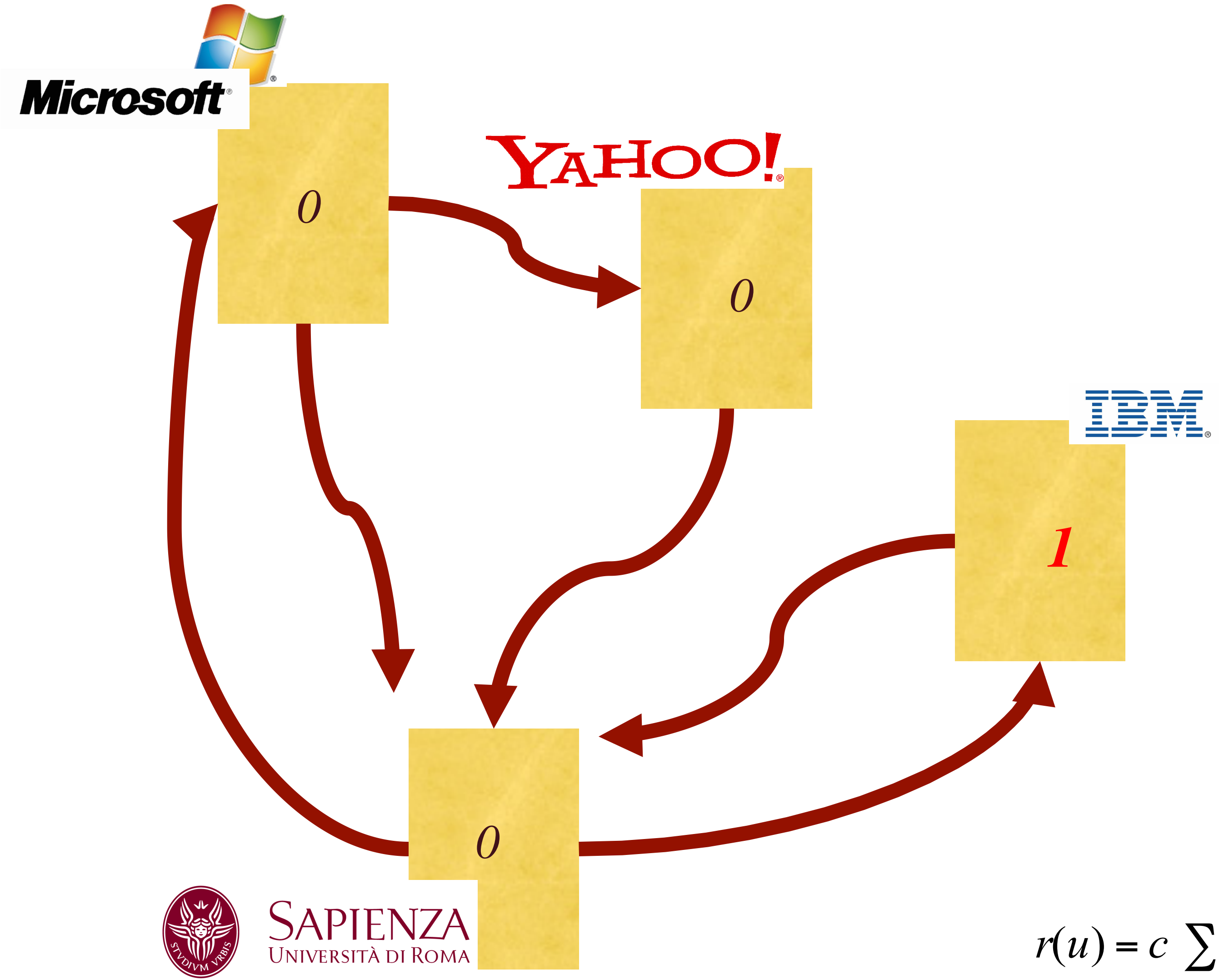
$$PR(u) = (1 - d) + d \times \sum \frac{PR(v)}{N(v)}$$

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

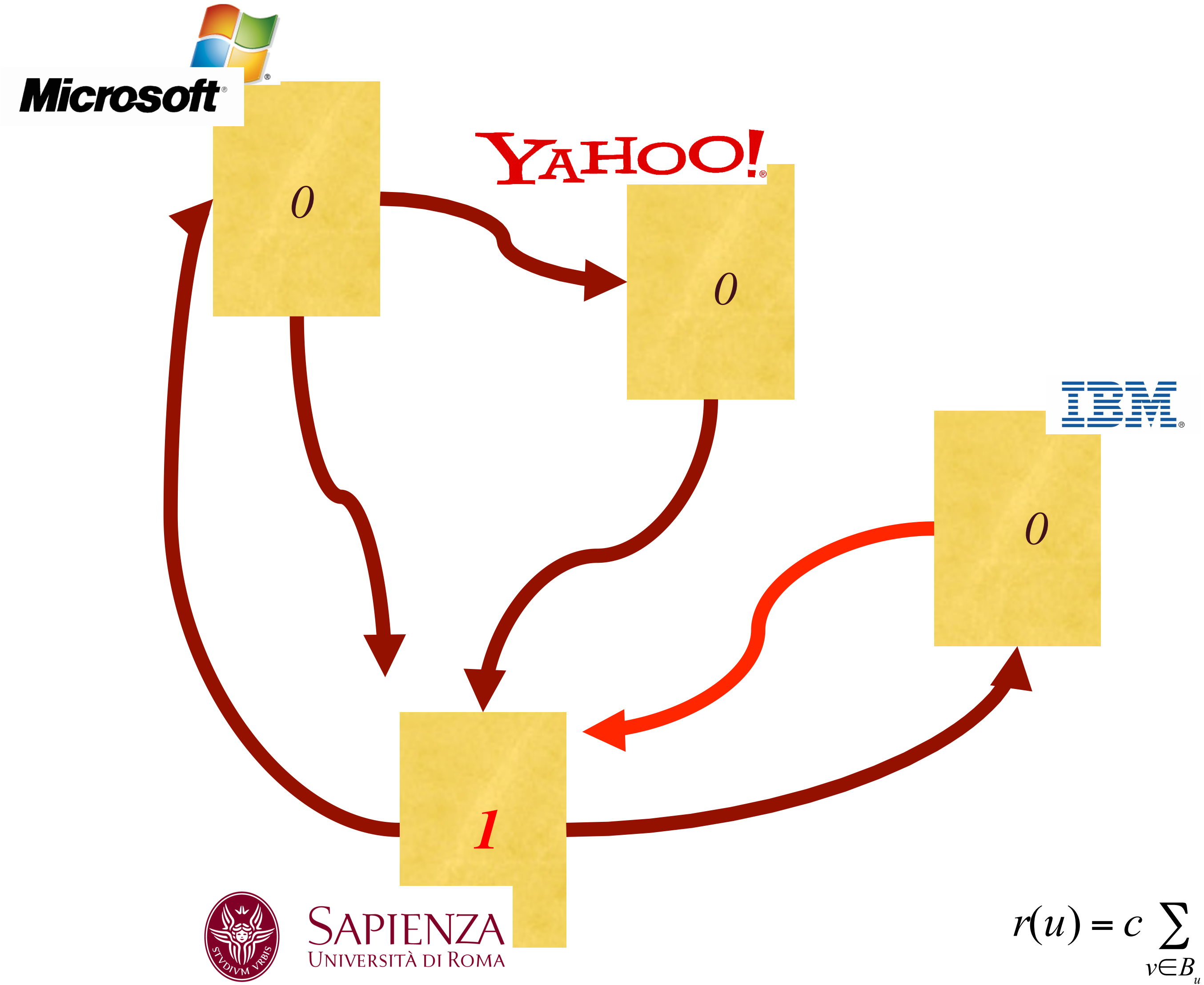


How is it calculated?

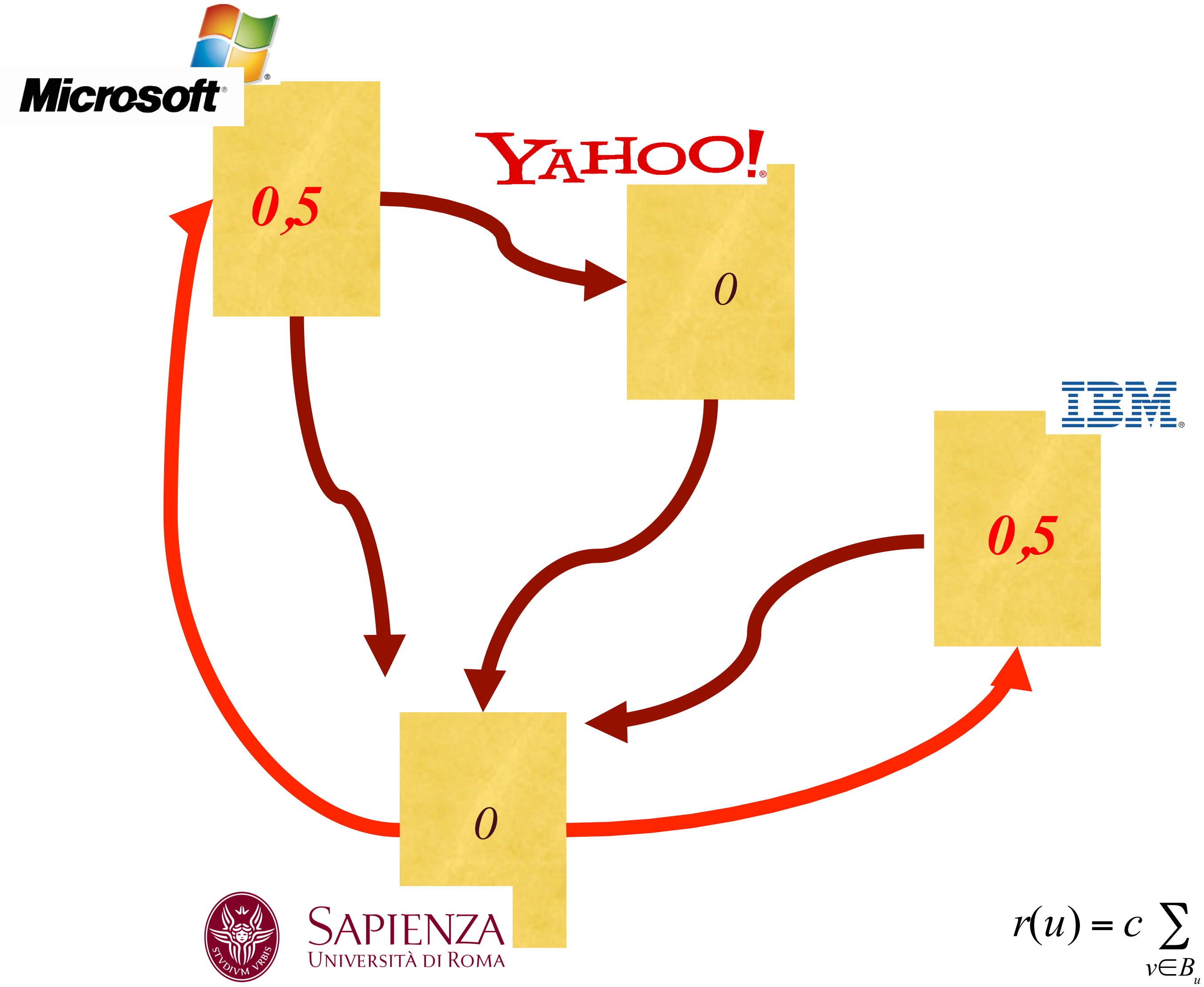
- The rank of a node depends on the rank of its pointing nodes..
- So it seems a circular problem: how can we compute it for all nodes?
- Start with a random guess of page rank values, and keep on adjusting values until values don’t change (steady state)
- In computer science, this is called RECURSION



$$r(u) = c \sum_{v \in B_u} \frac{r(v)}{N_v}$$

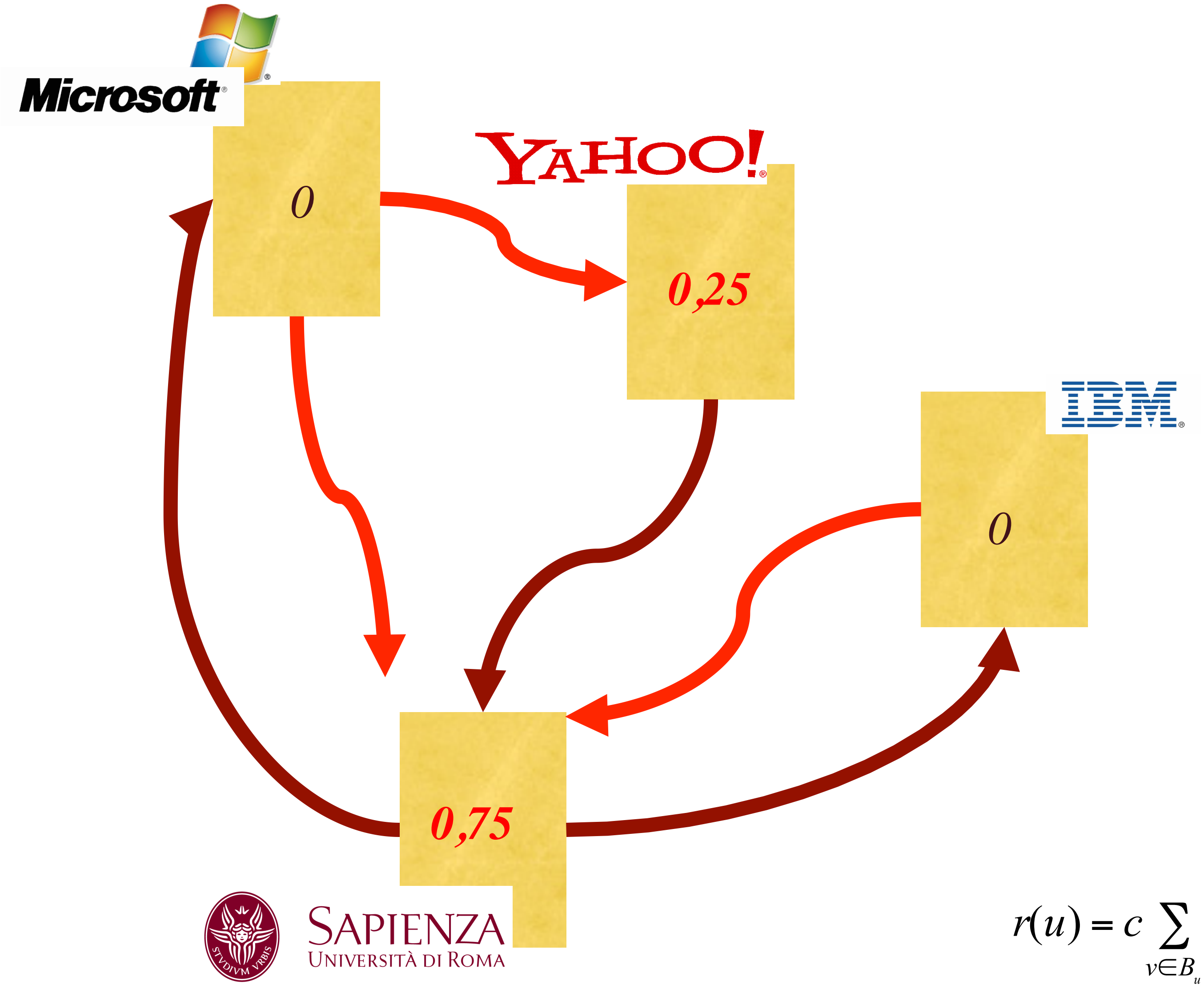


$$r(u) = c \sum_{v \in B_u} \frac{r(v)}{N_v}$$



SAPIENZA
UNIVERSITÀ DI ROMA

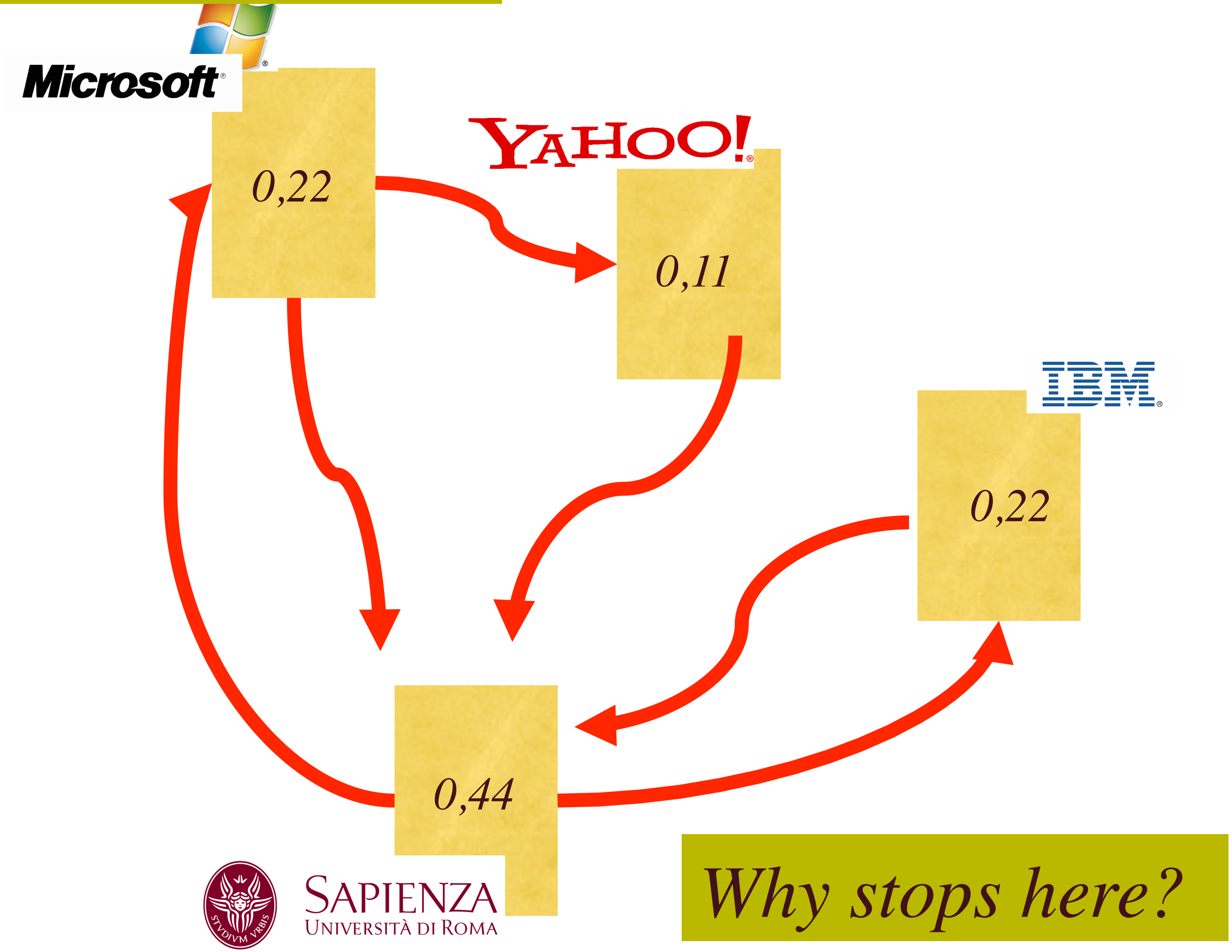
$$r(u) = c \sum_{v \in B_u} \frac{r(v)}{N_v}$$



SAPIENZA
UNIVERSITÀ DI ROMA

$$r(u) = c \sum_{v \in B_u} \frac{r(v)}{N_v}$$

After several iterations..

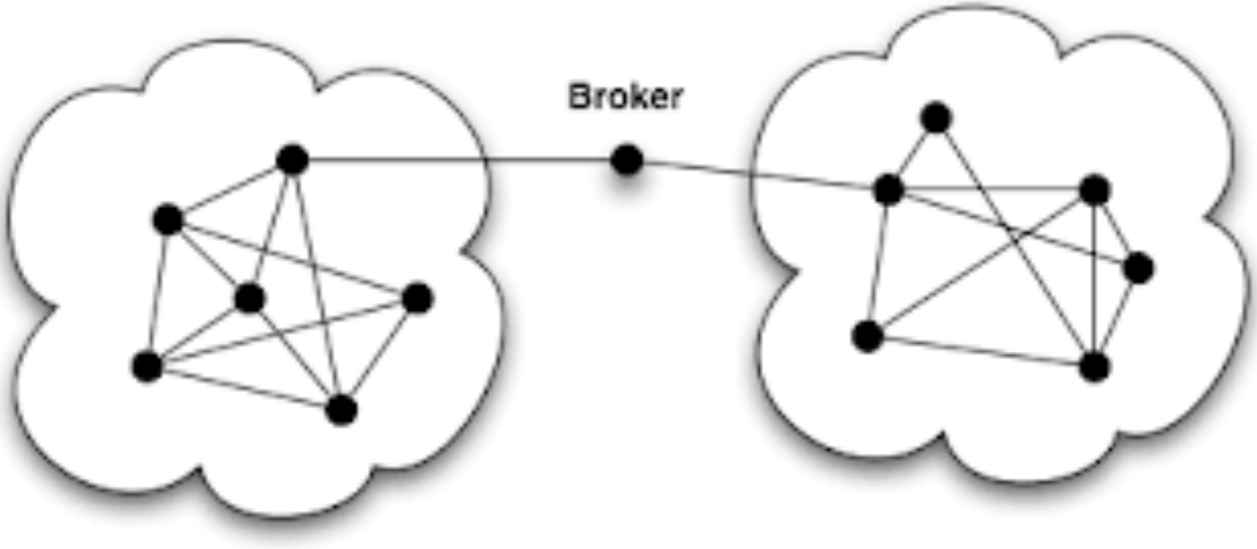


Why stops here?

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

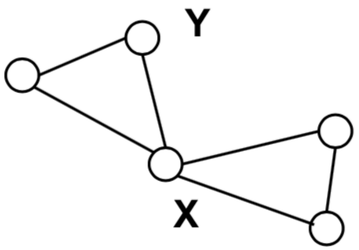
Example of bridge

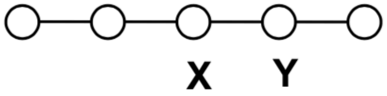


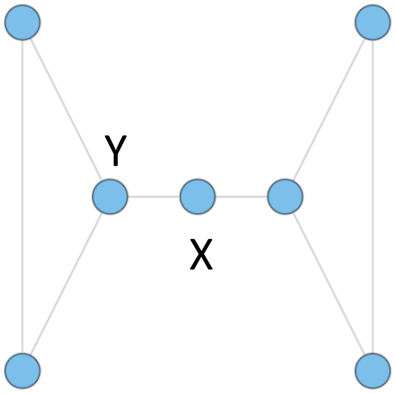
Algorithms to identify bridges (also called brockers) are all based on some measure of the **graph connectivity**.

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?

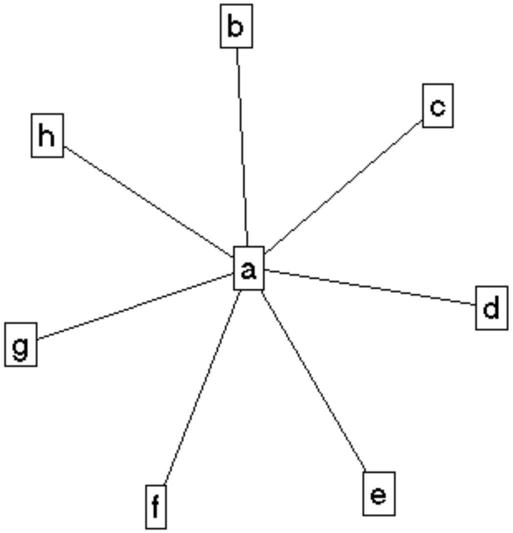


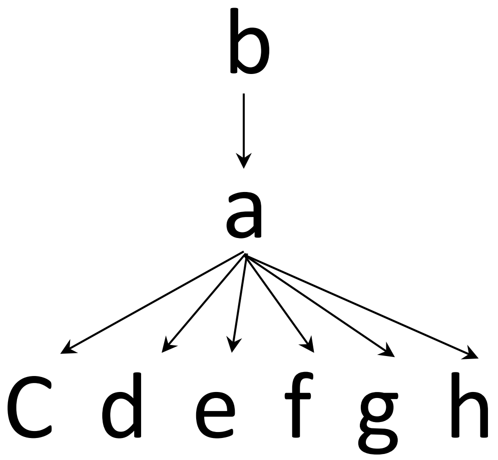




Formally: Betweenness Centrality

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

a

C d e f g h

Betweenness Centrality

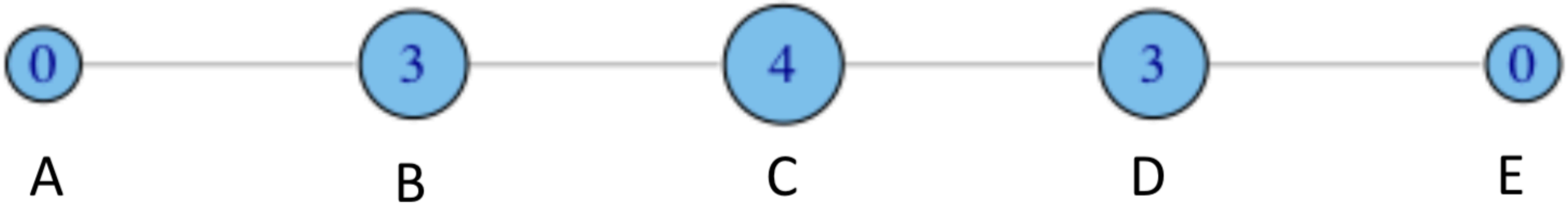
$$C_B(n_i) = \sum_{j<k} g_{jk}(n_i) / g_{jk}$$

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).

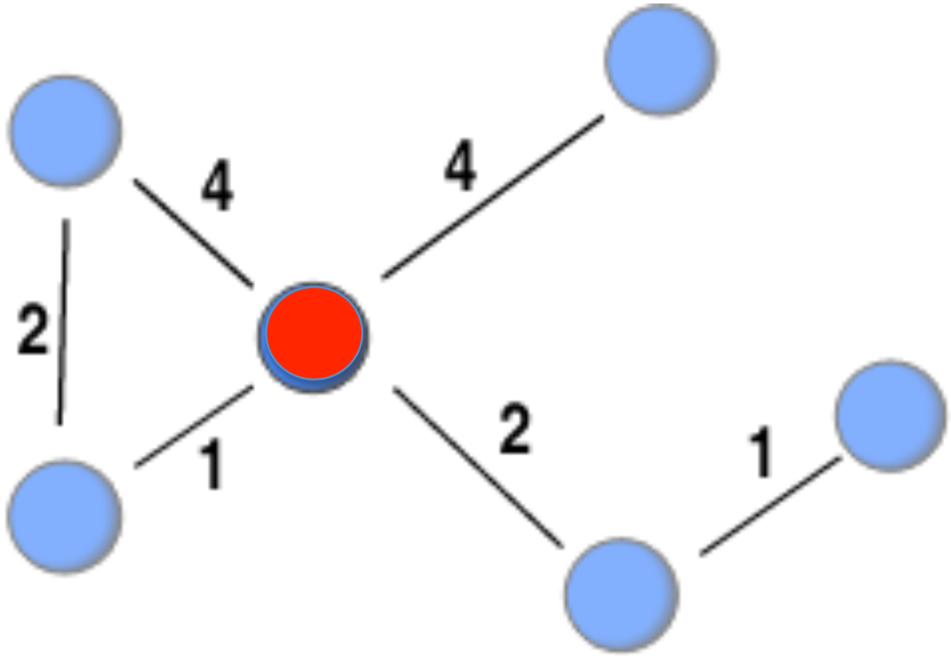
Can also compute **edge betweenness** in the very same way

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
- note that there are no alternative paths for these pairs to take, so C gets



Example of computation (bridgeness of the red node)



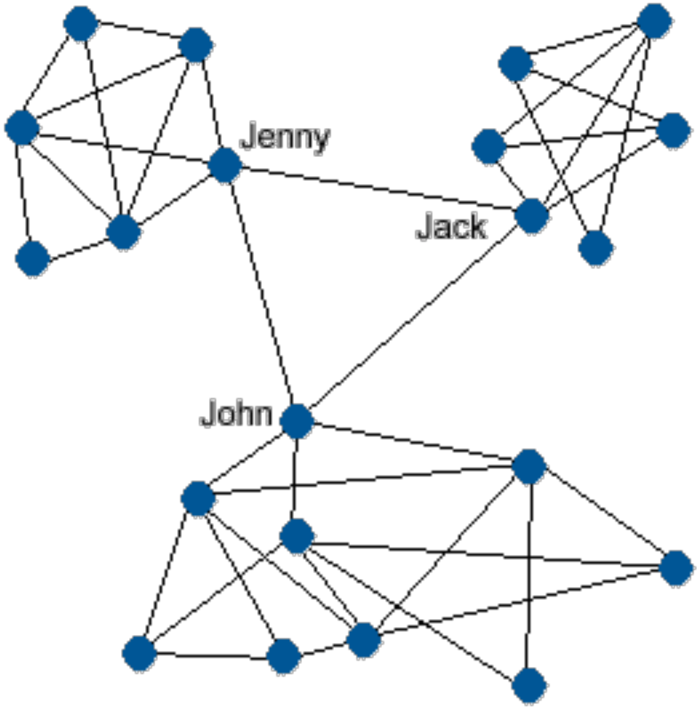
Many other measures of bridgeness

- 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 is 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 bridge sare actors that, if removed from the graph, **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

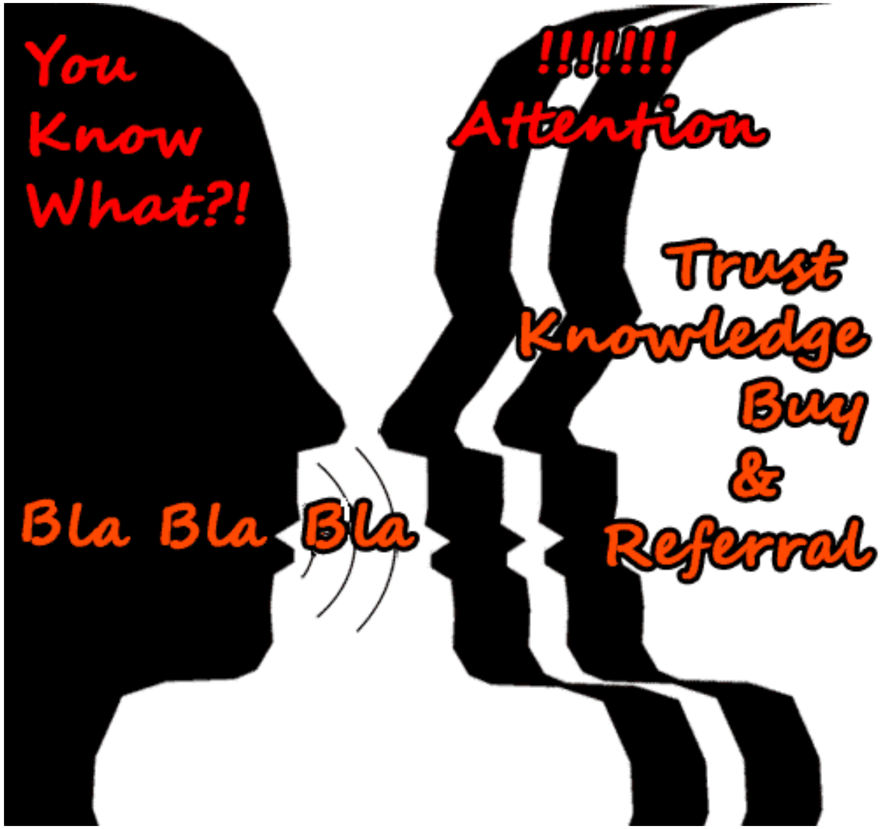
- Besides identifying key players (centrality, bridgeness) other types of information are relevant, and they require a global analysis of the network, not just single nodes
- E.g., If we wish to measure the likelyhood that an information originated anywhere in the network will reach you (spread of influence)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

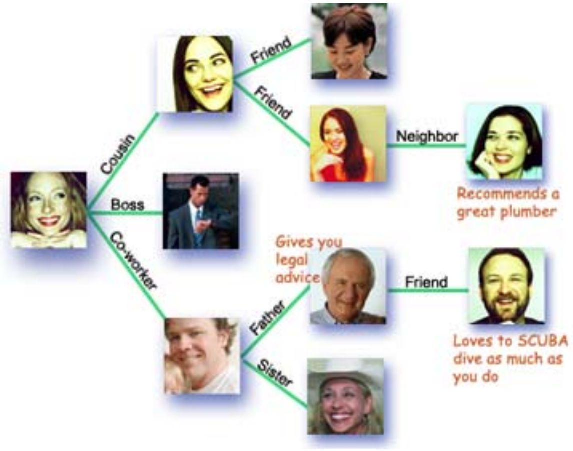
Influence Spread

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

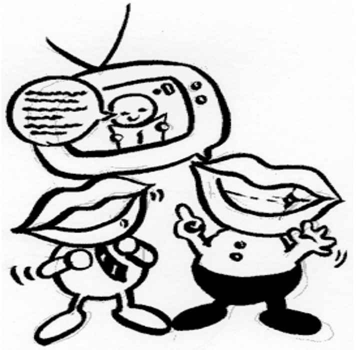


Social Network and Spread of Influence

- 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

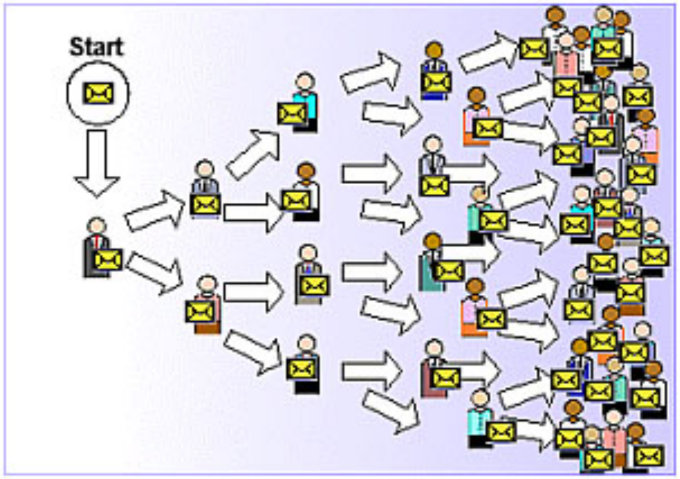
- Examples:
  - Hotmail grew from zero users to 12 million users in 18 months on a small advertising budget.
  - A company selects a small number of customers and ask them to try a new product. The company wants to choose a small group with largest influence.
  - Obesity grows as fat people stay with 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



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
- It also depends on the personal tendency of an actor to be influenced by others

Linear Threshold Model

- A node *v* has random threshold $\theta_v \sim U[0,1]$ *(this model the tendency to be influenced: the higher the threshold, the lower is the influence of others on an actor opinion)*
- A node *v* is influenced by each neighbor *w* according to a *weight* b_{vw} such that

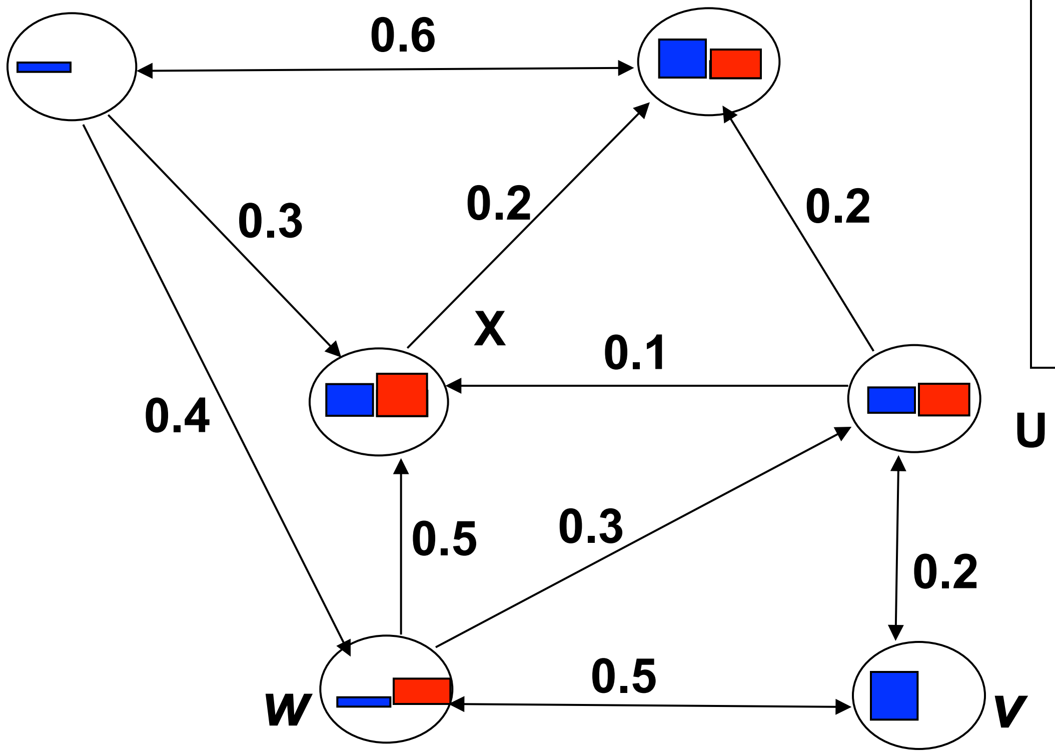
$$\sum_{w \text{ neighbor of } v} b_{v,w} \leq 1$$

- This model 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

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

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



○ Inactive Node

🟢 Active Node

🟦 Threshold

🟥 Active neighbors

Stop!

Outline

- Models of influence
  - Linear Threshold
  - Independent Cascade
- Influence maximization problem

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**
  - Interactions between nodes can help determine communities

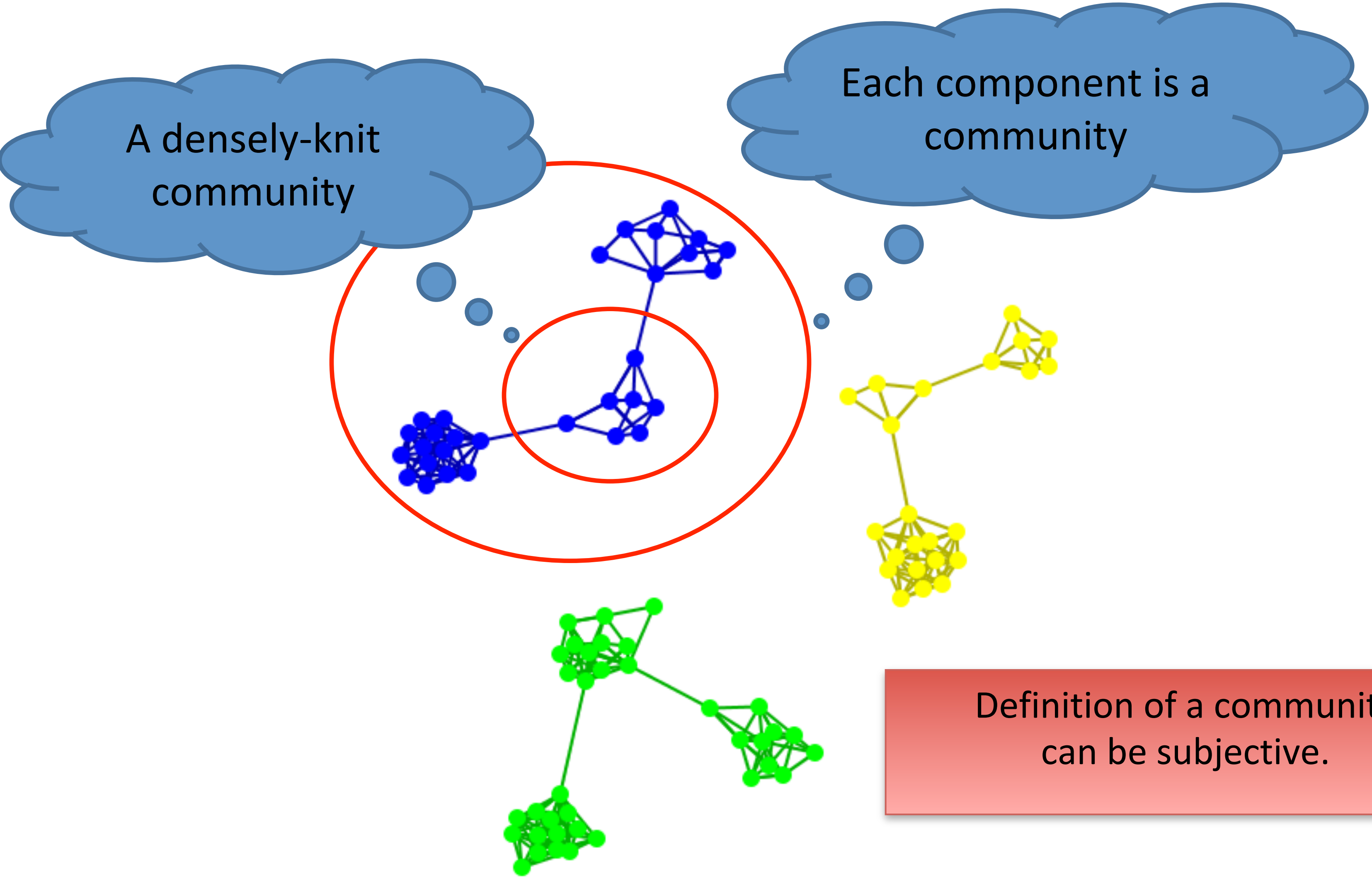
Communities in Social Media

- Two types of groups in social media
  - Explicit Groups: formed by user subscriptions (e.g. Google groups, Twitter lists)
  - Implicit Groups: implicitly formed by social interactions

- Some social media sites allow people to join groups, however it is still necessary to extract groups based on network topology
  - Not all sites provide community platform
  - Not all people want to make effort to join groups
  - Groups can change dynamically

- Network interaction provides rich information about the relationship between users
  - Can complement other kinds of information, e.g. user profile
  - Help network visualization and navigation
  - Provide basic information for other tasks, e.g. **recommendation**

Subjectivity of Community Definition



A densely-knit community

Each component is a community

Definition of a community can be subjective.

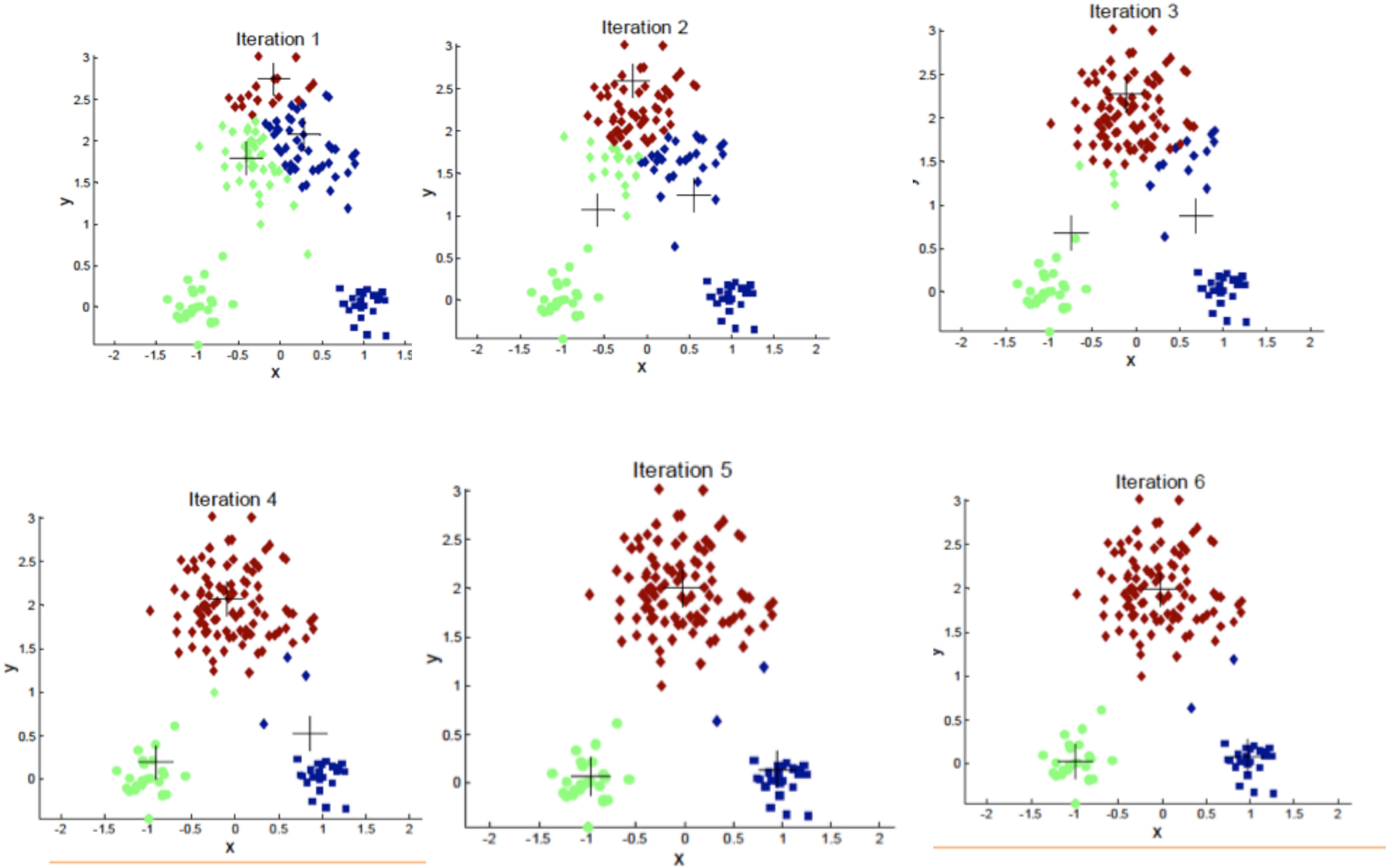
Community detection = clustering

- The two problems are very similar
- And share the same complexity

A very simple method

- K-means
- Input is the number of communities you wish to identify, K
- Algorithm:
  - Select K random nodes, each of these node represent the "cluster center"
  - Assign every other node to the cluster it is more close to
  - Compute the new cluster center
  - Iterate until clusters become stable (no more nodes are moved from one cluster to the other)

Example with K=3



Iteration 1

Iteration 2

Iteration 3

Iteration 4

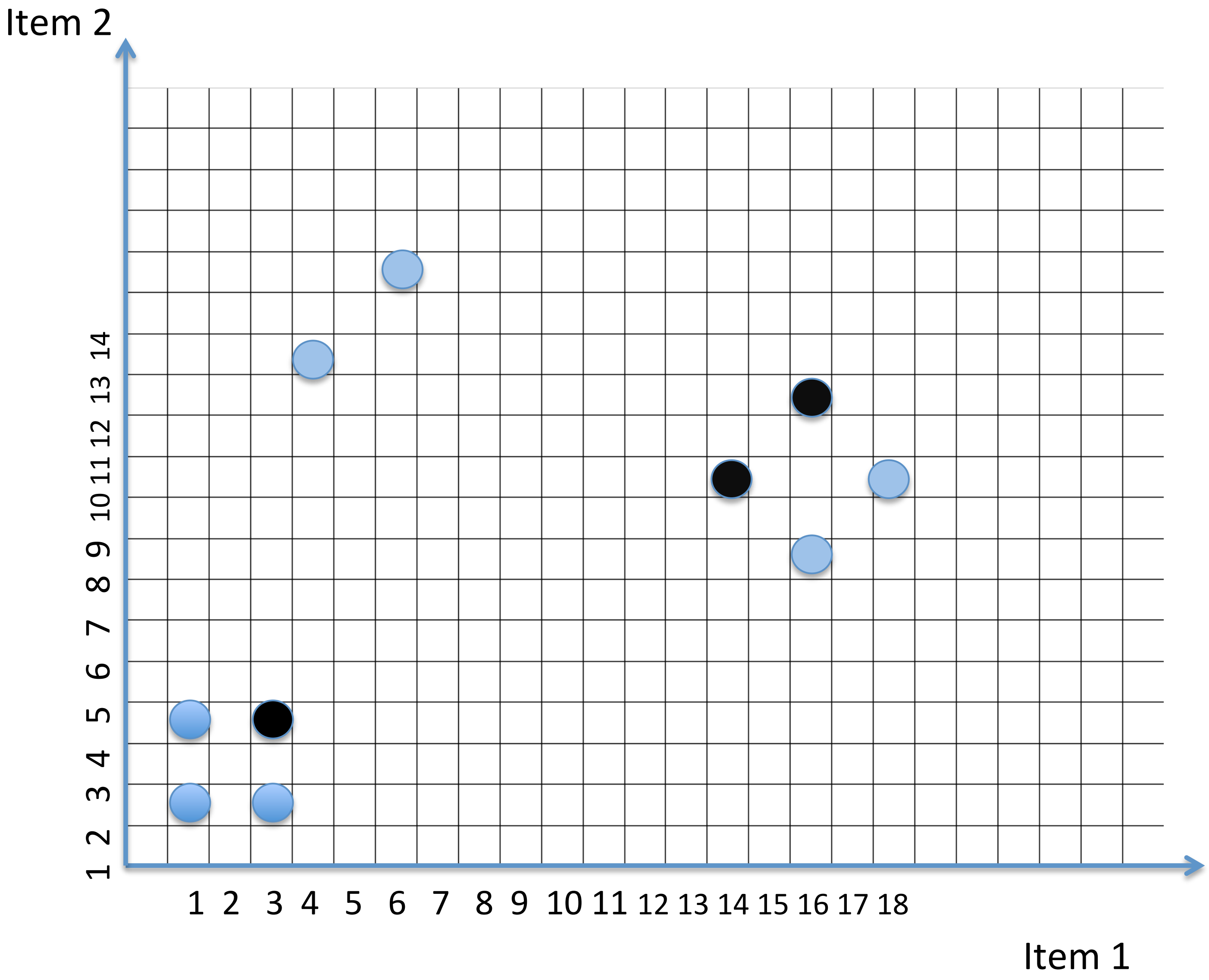
Iteration 5

Iteration 6

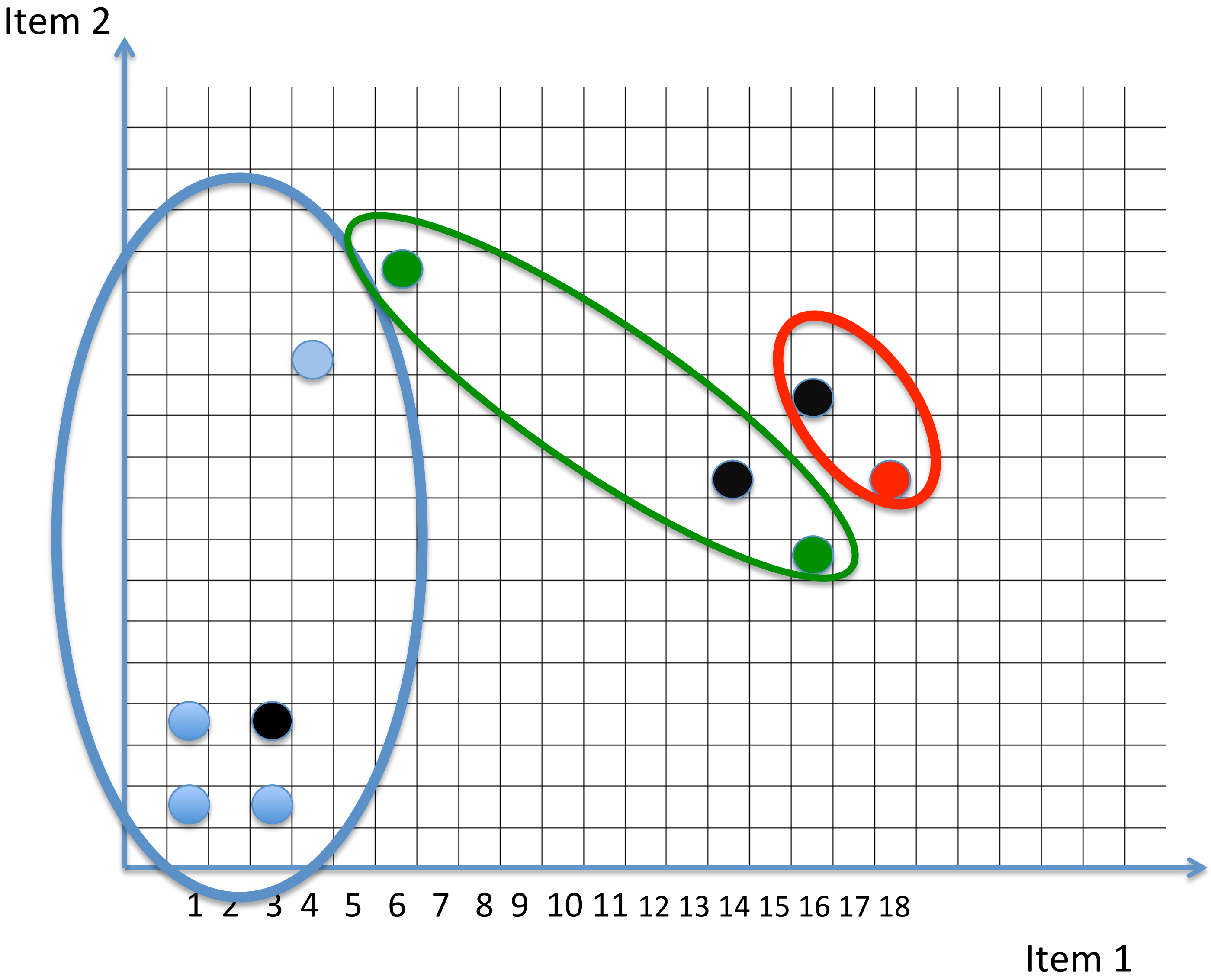
A step-by step example (customer clustering based on purchasing behaviours)

User ID	Item 1	Item 2
1	1	1
2	1	3
3	3	1
4	3	3
5	4	12
6	6	14
7	14	9
8	16	7
9	16	11
10	18	9

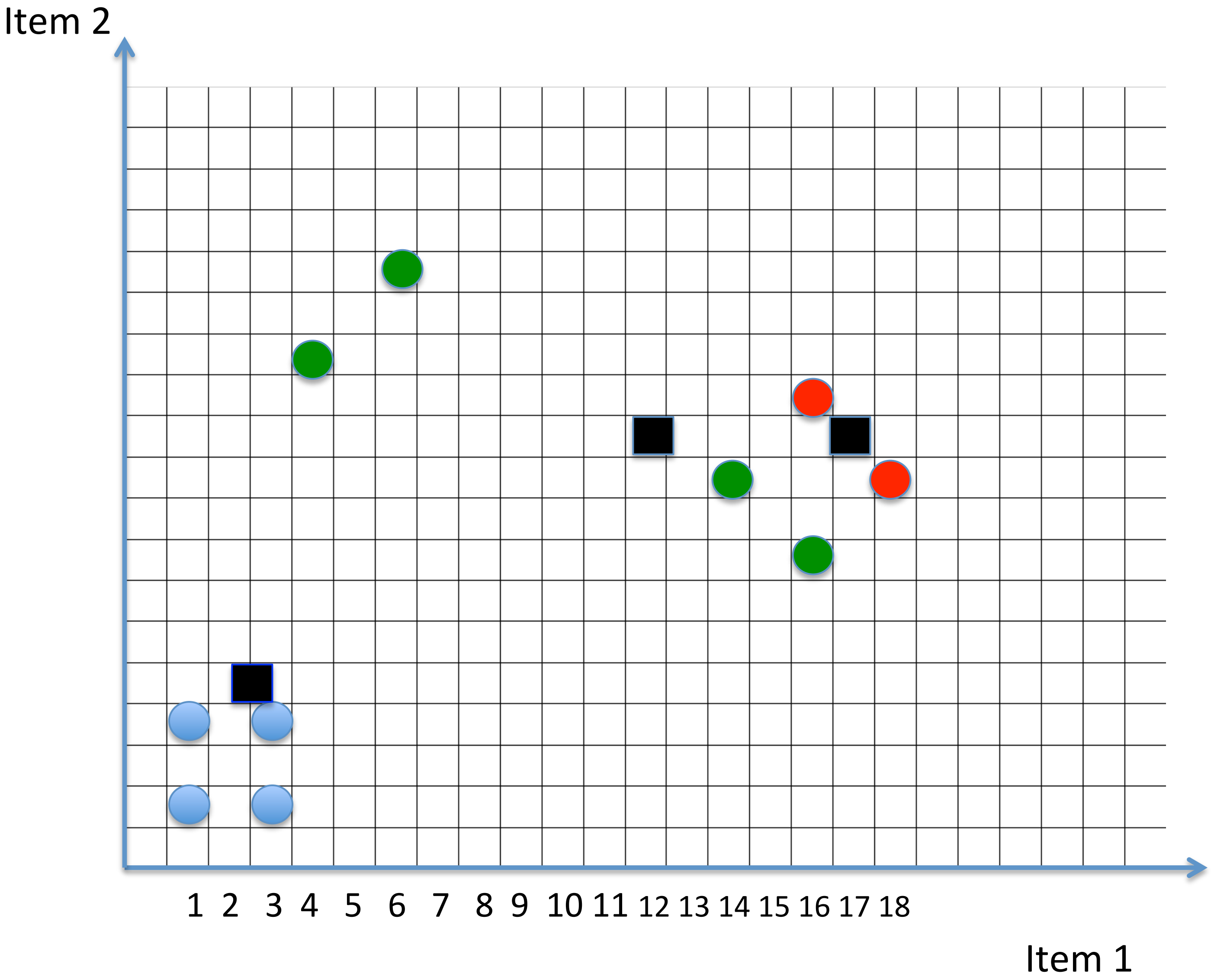
The table shows a list of users and their purchases (# of purchases for item)



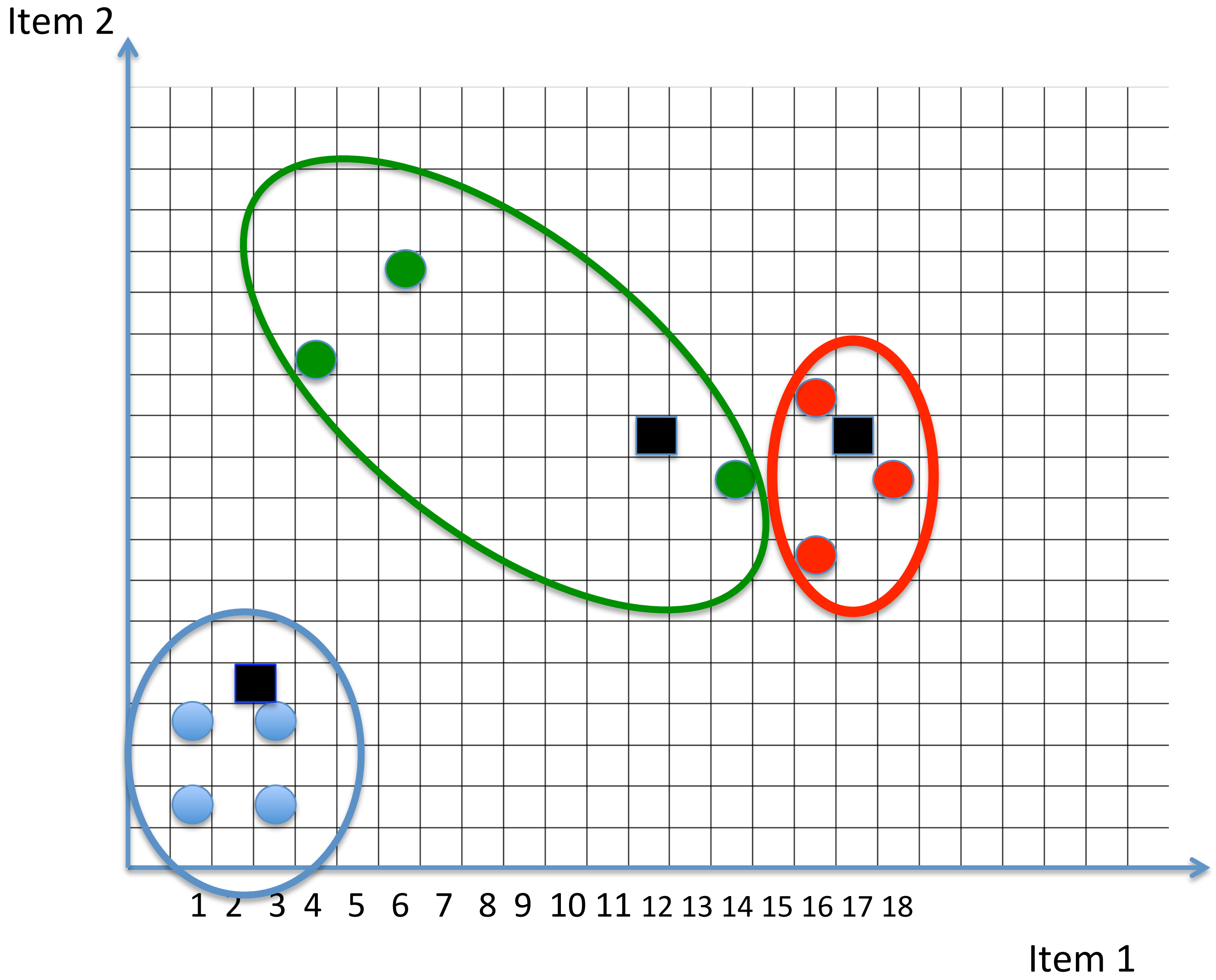
Each user is represented as a point in a bi-dimensional space. In step 1, we randomly pick up 3 users who are the initial cluster seeds.



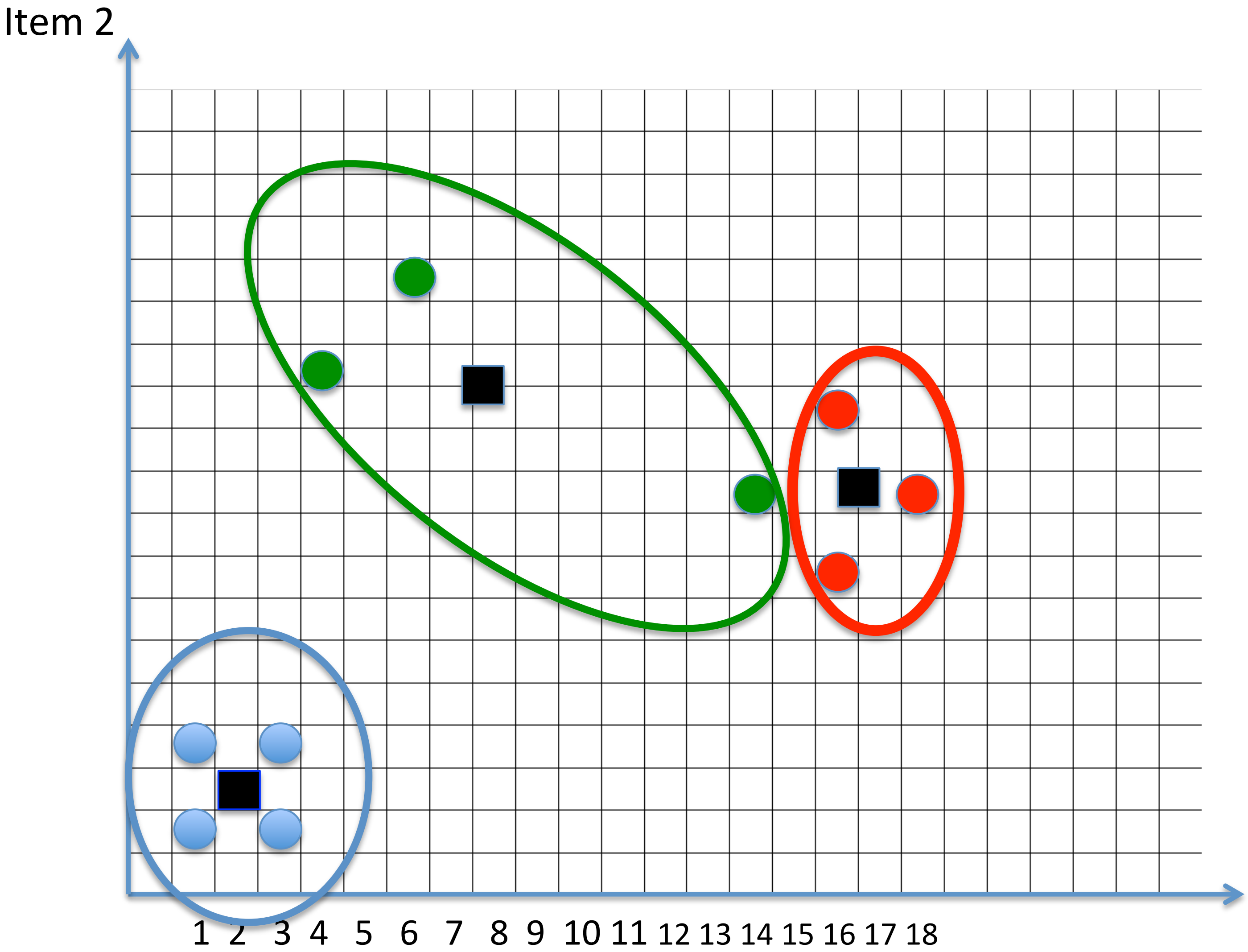
Each of the other users is assigned to the cluster whose center is the closest among the three initial seeds



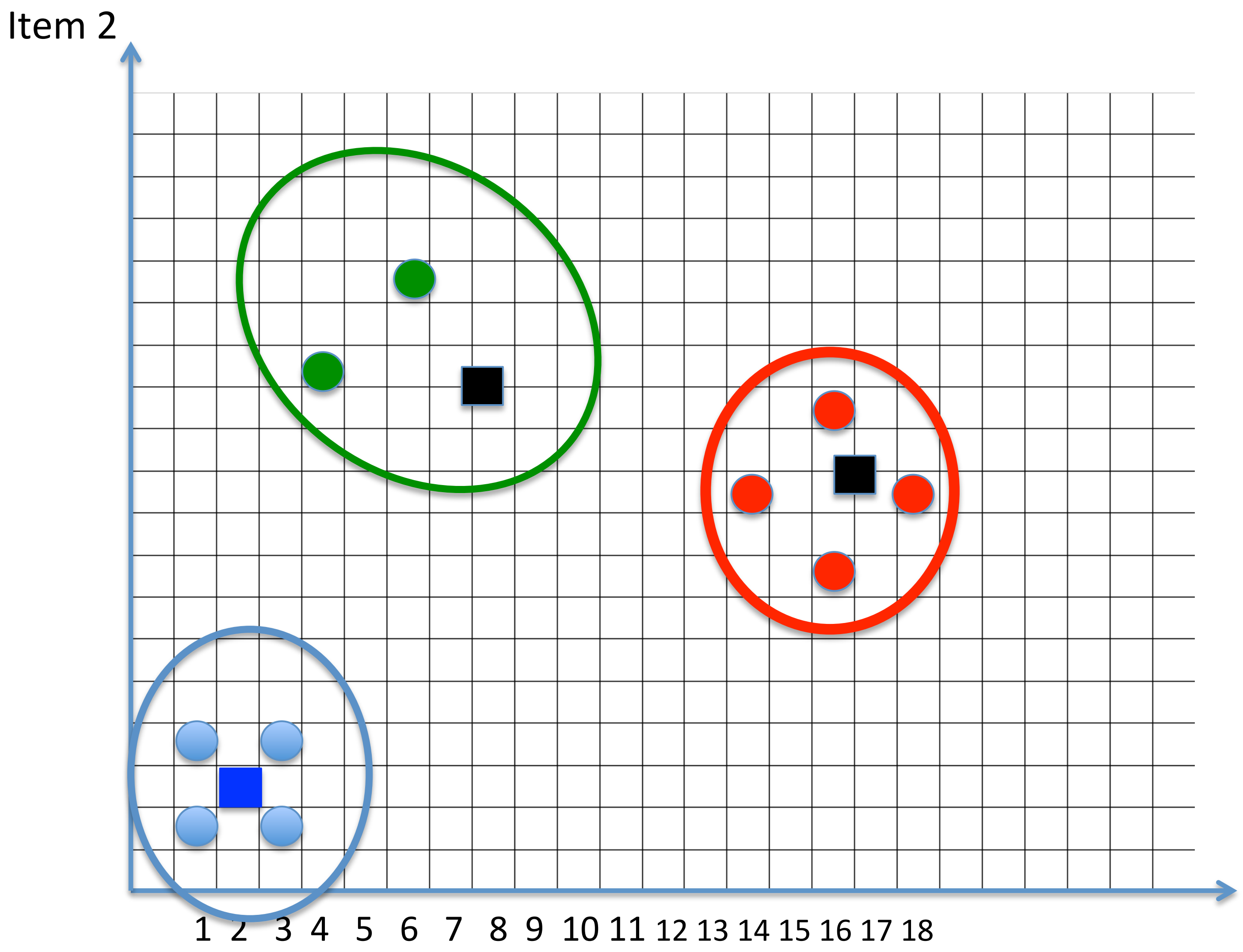
For each cluster, we compute a centroid (centroid are an “average” of the users of each cluster – indicated by squares)



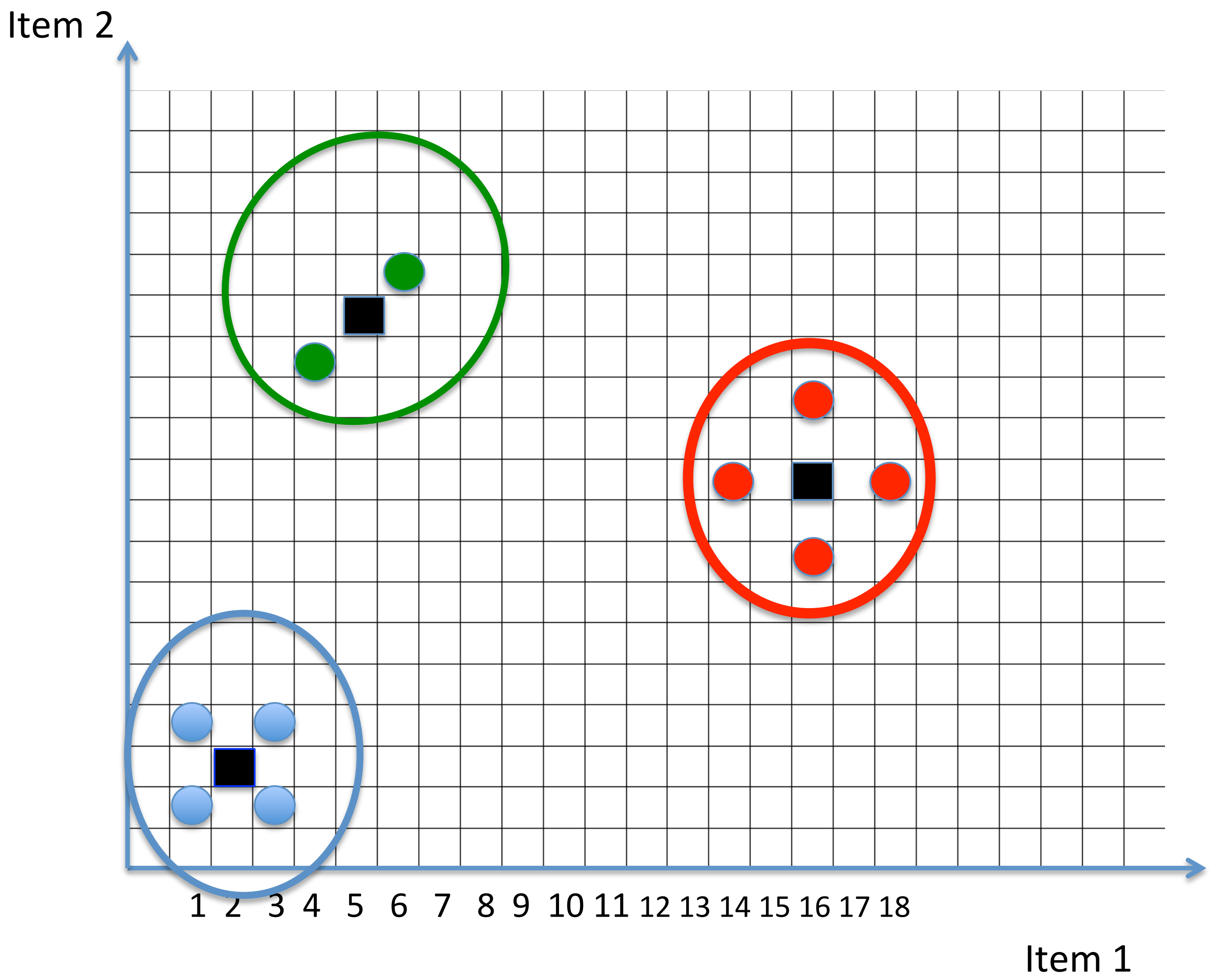
Using the new centroids, we re-assign all users to a centroid and re-compute clusters



Based on new clusters, we re-compute centroids



..and re-assign users to clusters, based on closest centroid



We re-adjust centroids, but now re-assignment of users to clusters produce NO changes. So we are done!