Graph Mining: Social network analysis and Information Diffusion

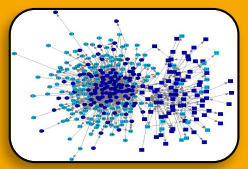
Davide Mottin, Konstantina Lazaridou

Hasso Plattner Institute

Graph Mining course Winter Semester 2016

Lecture road

Introduction to social networks



Real word networks' characteristics

Information diffusion

Paper presentations

- Link to form: https://goo.gl/ivRhKO
- Choose the date(s) you are available
- Choose three papers according to preference (there are three lists)
- Choose at least one paper for each course part
- Register to the mailing list: https://lists.hpi.uni-potsdam.de/listinfo/graphmining-ws1617

What is a social network?

- Oxford dictionary

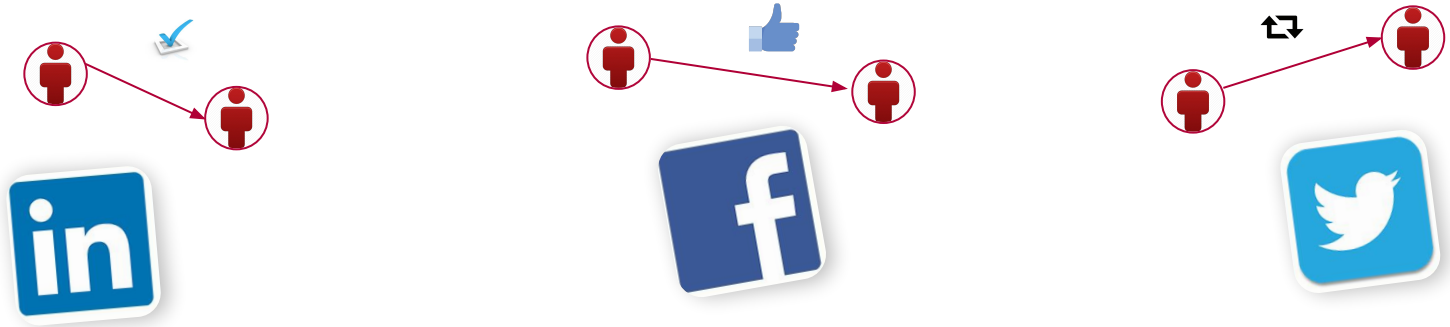
A social network is a network of social interactions and personal relationships

What is an online social network?

- Oxford dictionary

An online social network (**OSN**) is a dedicated website or other application, which enables users to communicate with each other by posting information, comments, messages, images, etc.

- Computer science : A network that consists of a finite number of users (nodes) that interact with each other (edges) by sharing information (labels, signs, edge directions etc.)



Definitions

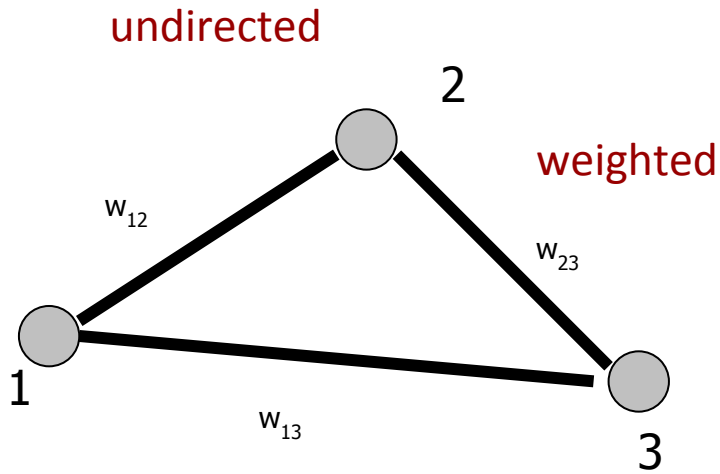
- Vertex/Node : a user of the social network (Twitter user, an author in a co-authorship network, an actor etc.)
- Edge/Link/Tie : the connection between two vertices referring to their relationship or interaction (friend-of relationship, re-tweeting activity, etc.)

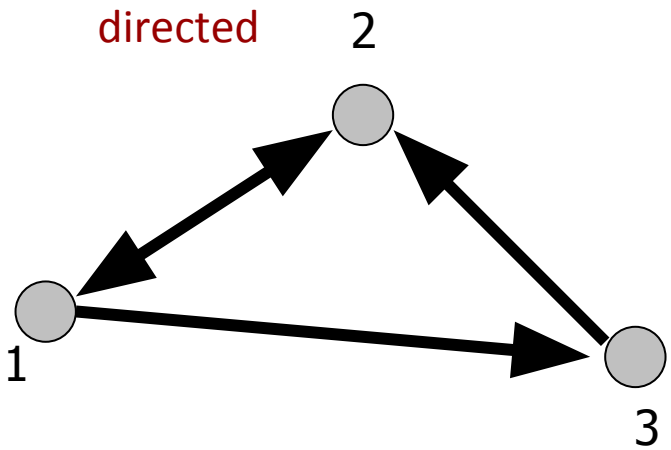
Graph G=(V,E)

symbols	meaning
V	users
E	connections
deg(u)	node degree: number of edges incident to the node u
$d_{in}(u)$	node indegree: number of incoming edges to u
$d_{out}(u)$	node outdegree: number of outgoing edges from u
SP(u,v)	shortest path between u, v
$\max_{u,v} SP(u,v)$	network diameter: longest shortest path

A path from u to v is a sequence of edges from u to v

Graph types





$V = \{1, 2, 3\}$

$E=\{(1,2),(1,3),(2,3)\}$

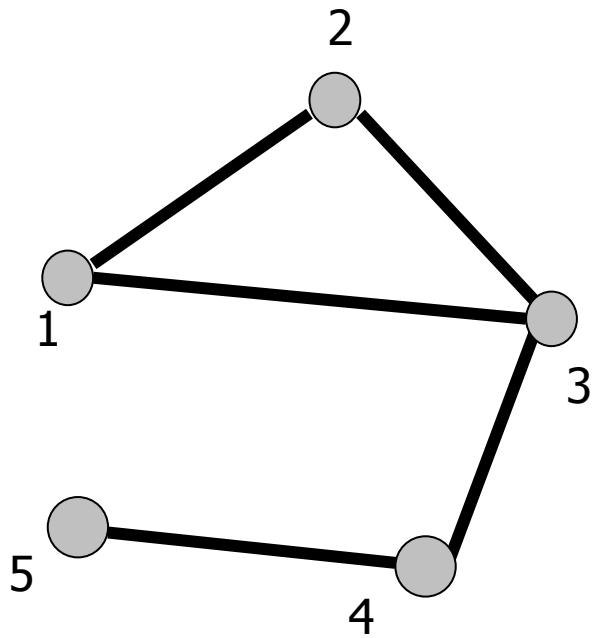
$E=\{<1,2>,<2,1><1,3>,<2,3>\}$

$E=\{(1,2,w_{12}),(1,3,w_{13}),(2,3,w_{23})\}$

Graph representations

Adjacency matrix

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



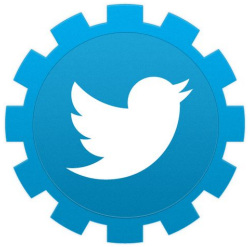
List of edges

(1,2)
(2,3)
(1,3)
(3,4)
(4,5)

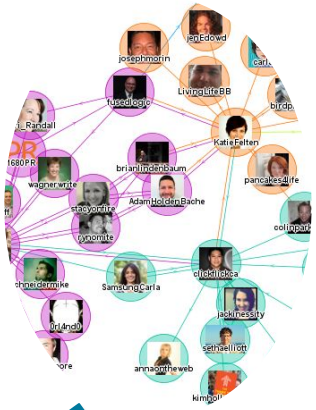
Adjacency list

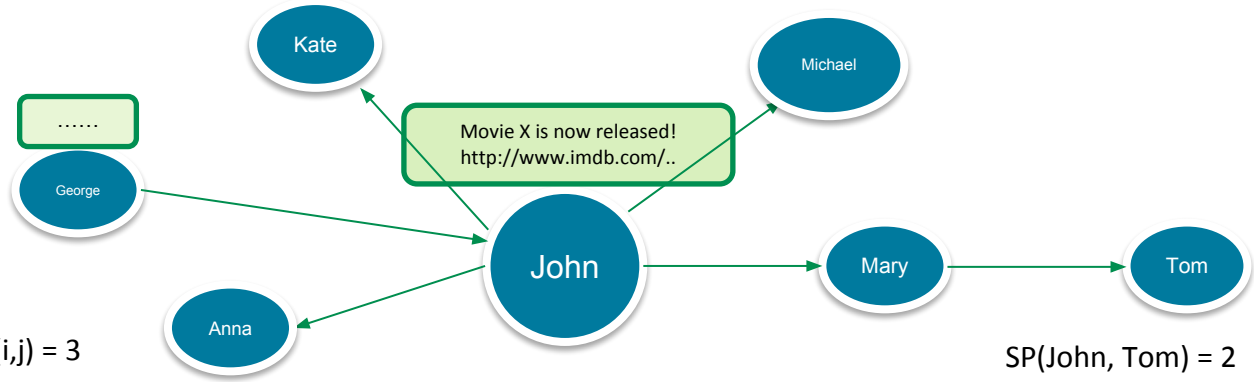
1: [2, 3]
2: [1, 3]
3: [1, 2, 4]
4: [3, 5]
5: [4]

Small directed graph example



{movie X, 2016, …}





$\max_{i,j} SP(i,j) = 3$

$d_{out}(John) = 4$

$SP(John, Tom) = 2$

$d_{in}(Mary) = 1$

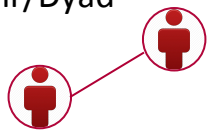
Interaction and concepts in social networks

- Social interaction types

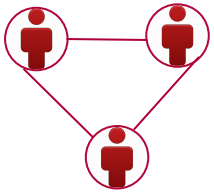
Roles	Friend of, Brother/Father of, Co-worker of
Affective	Likes/Dislikes, Loves/Hates
Proximity	Knows, Shares interests
Actions/Communication	Talks to, Comments on, Is with

- Interesting concepts and patterns in the network

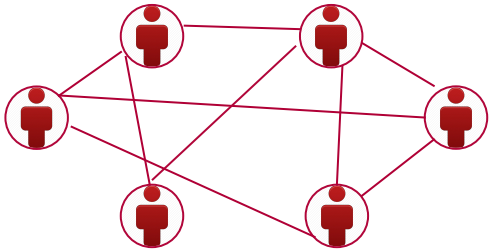
Pair/Dyad



Triangle



Group/Community



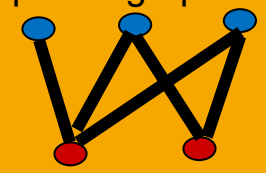
Wasserman, S. and K. Faust, Social Network Analysis, 1994, Cambridge

Possible network analysis focuses

- Individuals
  - How reachable a user *x* is? Is *x* trusted or a spammer?
- Overall network
  - Are there dense communities?
  - How large is the connected component of the network?
- Visualization
  - What are the hidden patterns?
- Structure
  - How many triangles does the network have? Is it a bipartite graph?

connected graph: every pair of nodes is connected

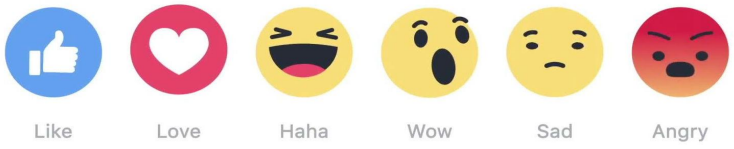
bipartite graph:



Famous graph structures and families, Wikipedia

Additional (more semantic) research questions..

- What do users think about each other?



- Is there any authority in the network?

President Obama @POTUS · 5.09
Happy Labor Day - longest streak of job growth ever, unemployment cut in half. That's what hardworking Americans can do. Let's keep going!

16 хил. 57 хил.

- What is discussed in the communities?

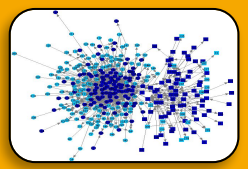
- How fast does the news spread in the network?

How does society benefit from Social Network analysis?

- Micro/Macro analysis
  - Better user experience with personalized recommendations
  - Influential users raise awareness on important issues
  - Understanding general public's opinion
- Anticipating natural hazards
- Epidemics detection
- Crime detection
  - Preventing terrorist attacks
  - Spam/Fraud detection

Lecture road

Introduction to social networks



Real world networks' characteristics

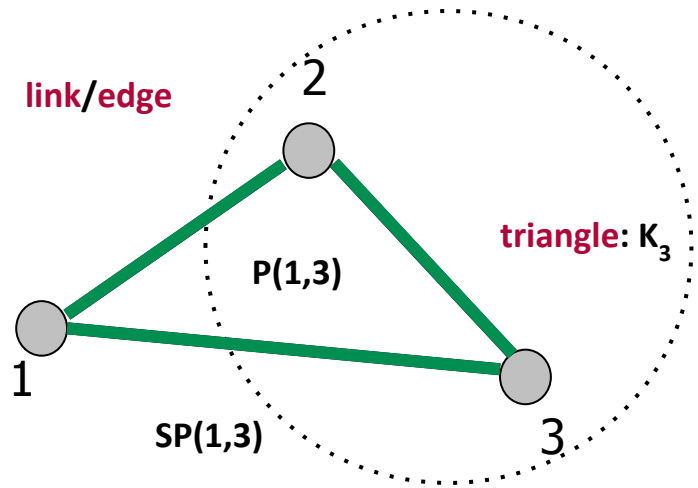
Information diffusion

Complex networks

- Graphs modelling *real systems* with non-trivial topological features
  - WWW
  - Internet router networks
  - Email networks
  - Social networks (co-authorship networks, friendship networks, etc.)
  - Brain networks
  - Protein-protein interaction networks
  - etc.

Graph Theory Recap

- for the complex networks we will need ..



neighborhood:
N(1) = {2,3}

diameter: $\max_{x,y} SP(x,y) = SP(1,3)$

- deg(1) = 2
- deg(2) = 2
- deg(3) = 2

degree distribution:
(2/6, 2/6, 2/6)

Properties of complex networks

- Main classes
  - **Small-world networks**
    - degree of separation
    - high clustering coefficient
  - **Scale-free networks**
    - power law degree distribution
    - power law clustering coefficient distribution
- General properties
  - Homophily
  - Community structure (will be discussed in January)

Barabasi, E. Bonabeau , Scale-Free Networks, 2003, Scientific American
H.E. Stanley, L.A.N. Amaral, A. Scala, M. Barthelemy, Classes of small-world networks, 2000, PNAS

Small World Phenomenon

"***It is the strength of the weak ties that holds together a social network***"

Mark Granovetter,
American Journal of Sociology,
1973

Stanley Milgram experiment

"*Six degrees of separation*"

- Nebraska citizens were assigned to send letters to people in Boston
- Participants sent the letters only to people they know
- Results: the letters that reached Boston followed ~6-length paths (average path length)

Small world network characteristics (1): small distances

- Social networks have small short path distances
- They also have small diameter
  - Actors network in IMDB: $\max_{u,v} SP(u,v) = 7$

- **Definition$_1$** (recap)

The diameter of a network is defined as the maximum shortest path between two nodes, i and j

- **Definition$_2$** (for not connected graphs)

Effective diameter (the *90th percentile* distance) is sometimes used as a smoothed version instead

Leskovec, Kleinberg, Faloutsos, Graphs over Time: Densification Laws, Shrinking Diameters and Possible Explanations, 2005, KDD

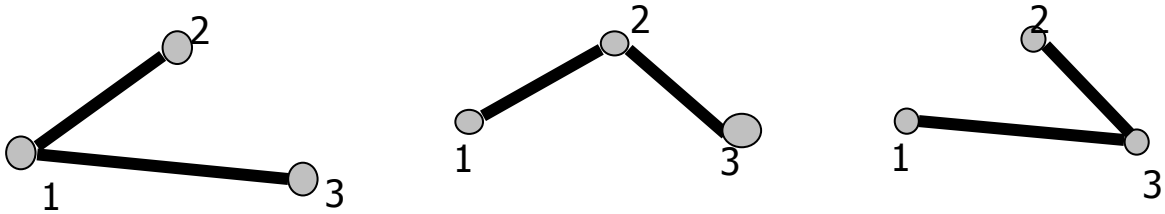
Small world network characteristics (2): high clustering coefficient

- Apart from small diameter, in small-world networks the clustering coefficient is also high
- **Definition$_3$**
  - *Clustering coefficient* is a measure of the degree to which the nodes in a graph tend to cluster together
- There are two types of clustering coefficient
  - Global (entire graph)
  - Local (individual vertices)
- The local clustering coefficient is related to the small world phenomenon

For the global clustering coefficient formula we need ..

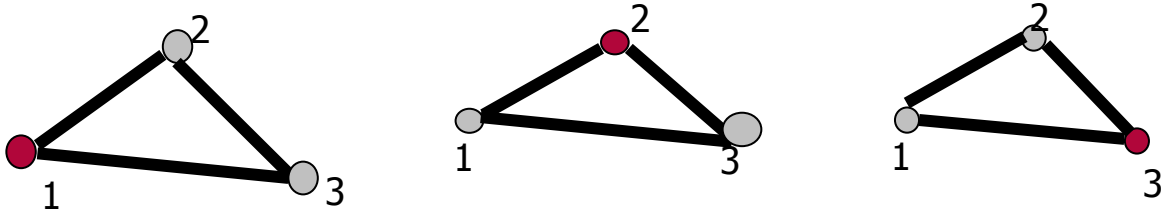
- Triplet

Open triplet



A triangle includes three closed triplets

Closed triplet



Clustering coefficient formulas

- Global coefficient gives an indication of the clustering in the whole network
- Local coefficient of a vertex i quantifies how close its neighbours are to being a clique
  - if the average local clustering coefficient is high (compared to its a random setting), then a graph is considered a small-world

Global clustering coefficient formula

$$C = \frac{3(\#\ \text{closed triplets})}{(\text{total triplets})}$$

Local clustering coefficient formula

$$Ci = \frac{\{e(u,v):\ u,v\ \epsilon\ N(i)\}}{|N(i)|(|N(i)|-1)}$$

Scale-free networks: Power laws

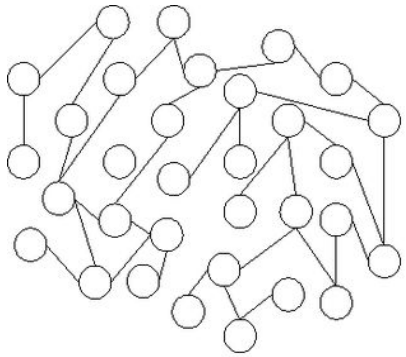
- A scale-free network is a network whose degree distribution follows a power law, at least asymptotically
- **A power-law implies that small occurrences are extremely common, whereas large instances are extremely rare**

The law holds also for..

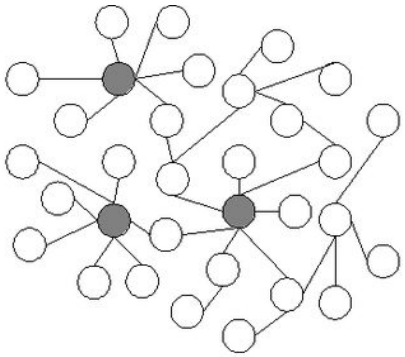
- city sizes
- incomes
- word frequencies
- earthquake magnitudes
- number of visits to a site
- number of pages within a site
- number of links to a page
- ……

Scale-free networks: Hubs and communities

- A hub refers to a vertex that connects to a lot of other vertices and communities
- Low-degree users form dense groups, which are connected to each other through hubs
- Hubs might have a key role in the network depending on the domain of study (politicians, celebrities, etc.)



(a) Random network

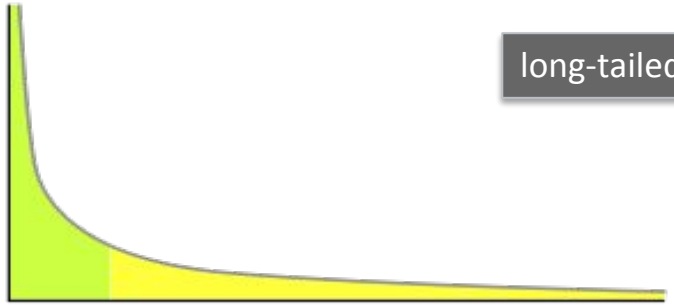


(b) Scale-free network

Power laws

- Scale-free networks also exhibit power laws in their clustering coefficient distribution (heavy tail)
- The clustering coefficient decreases as the node degree increases
- Other examples
  - There are few large earthquakes but many small ones
  - There are a few mega-cities, but many small towns
  - There are few words, such as 'and' and 'the' that occur very frequently, but many which occur rarely

80-20 rule



long-tailed distribution

Figure: Power law distribution example that demonstrates the distribution of popularity

Michalis Faloutsos, Petros Faloutsos and Christos Faloutsos.,On Power-Law Relationships of the Internet Topology, 1999, SIGCOMM

Zipf, Power-laws and Pareto

- Zipf law: "The rth largest city has y inhabitants"
  - the size of the rth largest occurrence of an event is inversely proportional to it's rank

$$y \sim r^{b}$$

- Pareto law: "r cities have y or more inhabitants"
  - the number of events larger than y is an inverse power of its rank

$$P[X > y] = r \sim y^{-1/b}$$

- Power law: "r cities have exactly y inhabitants"

$$P[X = y] = r \sim y^{-a} = y^{-a}$$

Power law distribution example

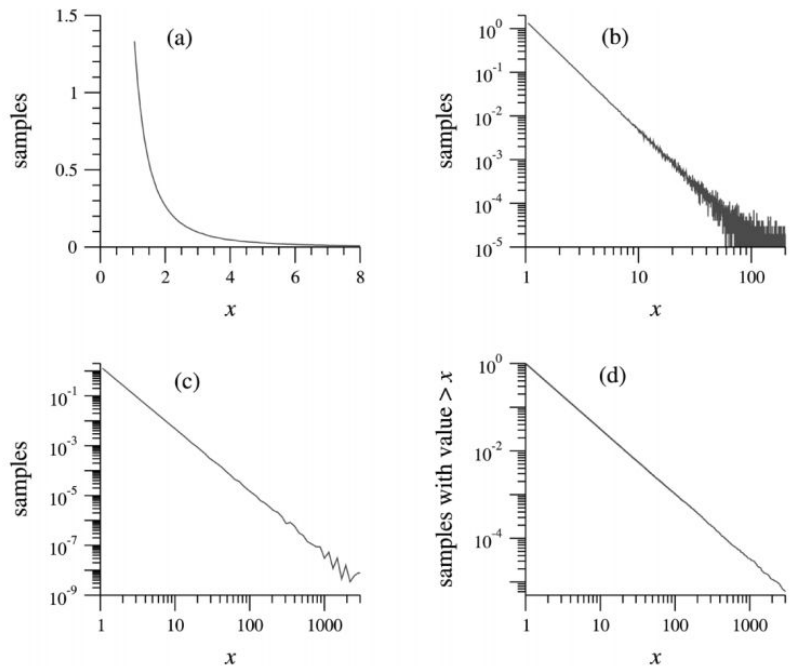


Figure:
(a) Histogram of the set of 1 million random numbers described in the text, which have a power-law distribution with exponent a= 2.5
(b) The same histogram on logarithmic scales. Notice how noisy the results get in the tail towards the right-hand side of the panel. This happens because the number of samples in the bins becomes small and statistical fluctuations are therefore large as a fraction of sample number.
(c) A histogram constructed using 'logarithmic binning'.
(d) A cumulative histogram or rank/frequency plot of the same data. The cumulative distribution also follows a power law, but with an exponent of a – 1 = 1.5.

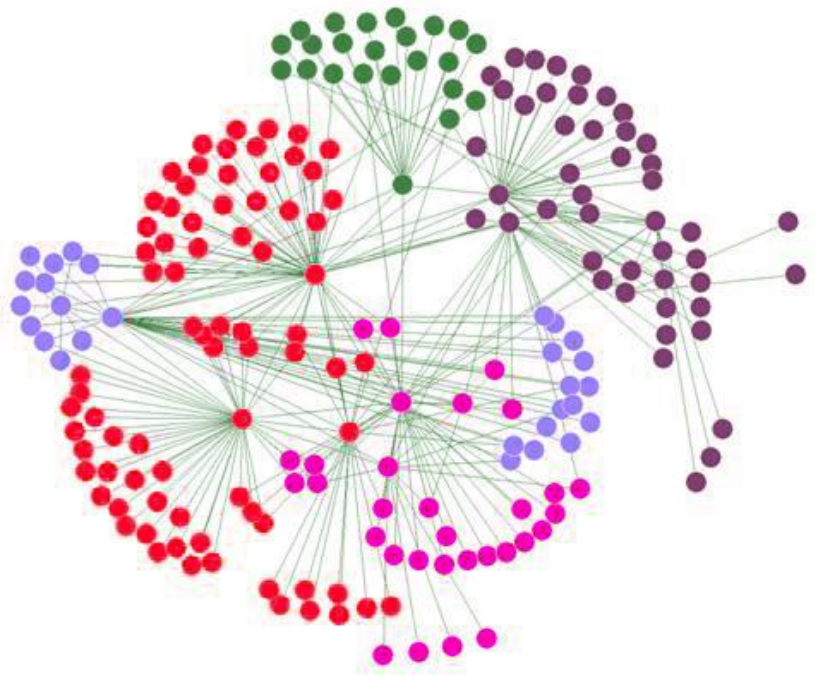
Mark E. J. Newman, Power Laws, Pareto Distributions and Zipf's Law, 2004, Contemporary Physics

Summary of scale-free network features

- Power laws
  - node degree distribution
  - clustering coefficient
- Very few users are popular (hubs)
- There are many small dense groups that everyone knows everyone (communities)
  - The communities connect to each other through hubs
  - Hubs don’t belong to dense communities
  - Hub removals could make the network fail! (isolated subgraphs)

A general property of real world graphs: Homophily: Love of the same!

- Another characteristic of social network users is homophily (or assortativity), which refers to the users' tendency to connect to 'similar' with them users
- People tend to have friends with common interests



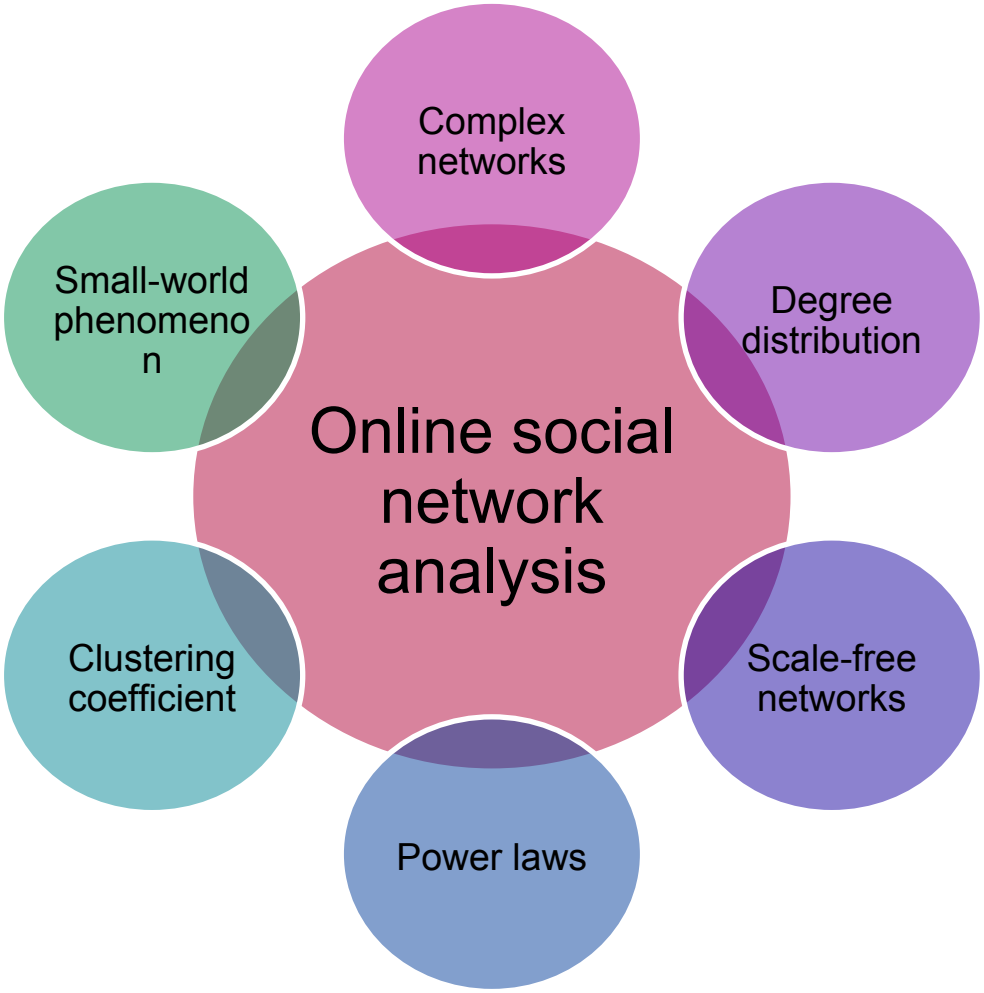
Miller McPherson, Lynn Smith-Lovin and James M. Cook, PBirds of a feather: Homophily in social networks, 2001, Annual Review of Sociology

Measurement and mechanisms

- Consider the different characteristics of the users (gender, race, age etc.)
- Given a characteristic ***c***, if the fraction of cross-***c*** edges is significantly less than expected, then there is evidence for homophily
- Homophily relies on two (inverse to each other) mechanisms
  - **Selection**: tendency of people to form friendships with others who are like them
  - **Socialization** or **Social Influence**: the existing social connections in a network are influencing the individual characteristics of the individuals

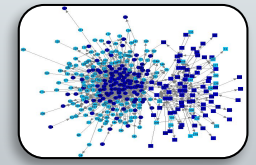
Halil Bisgin, Nitin Agarwal, Xiaowei Xu, Investigating Homophily in Online Social Networks, 2010, IEEE/WIC/ACM

Any questions so far?



Lecture road

Introduction to social networks



Real work networks' characteristics

Information diffusion

What is information diffusion?

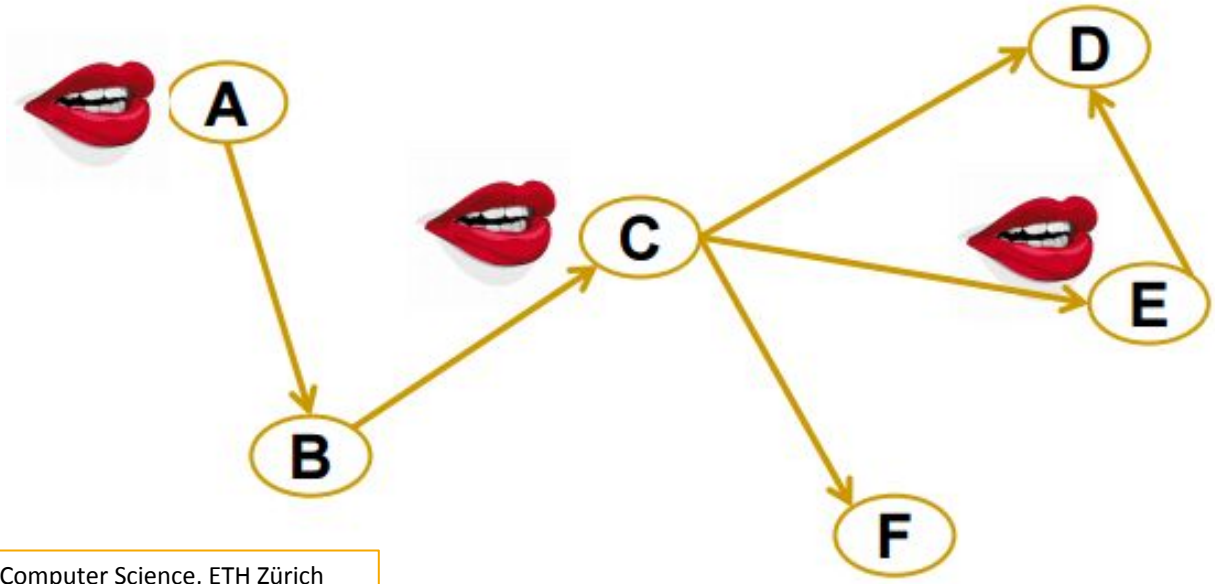
- Information is produced and consumed by millions of worldwide Internet users in online social networks
- Events, issues, interests, etc. happen and evolve very quickly in social networks
- Their capture, understanding, visualization and prediction are becoming critical expectations from both end-users and researchers

Motivation

- Analyzing new trends
- Preventing terrorist attacks
- Anticipating natural hazards
- Preventing spread of diseases
- Optimizing marketing campaigns

Interesting research questions

- Which topics diffuse faster?
- Which users play the most important role in spreading information?
- Which path is the information following?



Seminar in Distributed Computing, Computer Science, ETH Zürich

Basic notions

Topic : A coherent set of semantically related terms that express a single argument

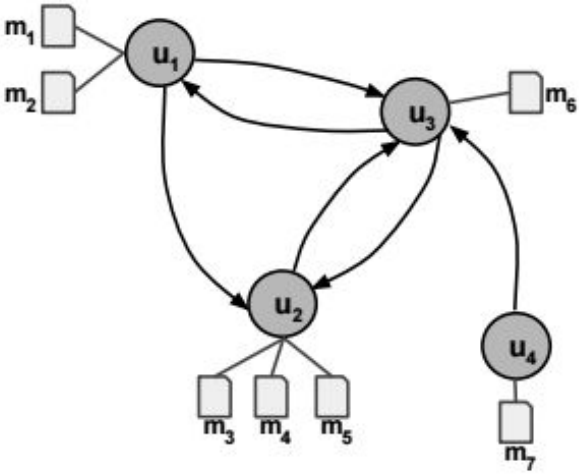
Bursty topic: A behavior associated to a topic within a time interval in which it has been extensively treated but rarely before and after.

Information cascade: People in the network adopt information and make decisions from inferences based on **earlier people's actions**.

Guille, Adrien and Hacid, Hakim and Favre, Cecile and Zighed, Djamel A., Information Diffusion in Online Social Networks: A Survey, 2013, ACM SIGMOD

OSN example

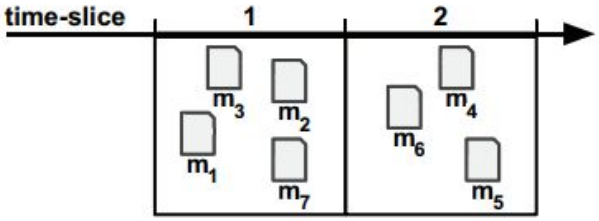
an OSN enriched by user messages.



u_1 is exposed to the messages published by u_3 and u_2

the stream of user messages

possible discretization



Information diffusion process

Activation sequence: An ordered set of nodes capturing the order in which the nodes of the network adopted a piece of information

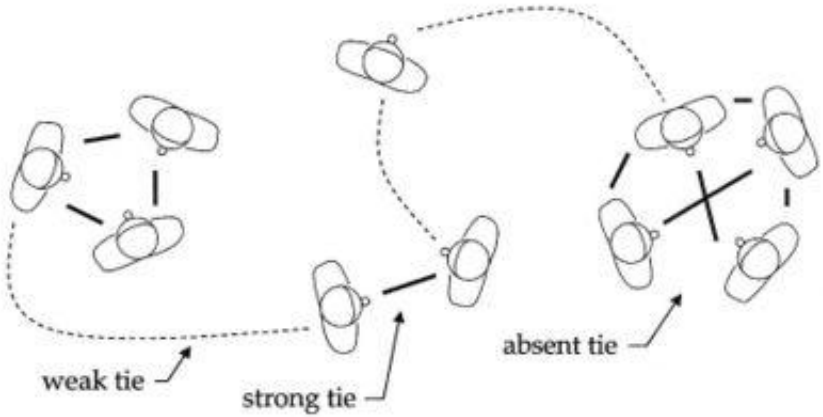
Spreading cascade: A directed tree having as a root the first node of the activation sequence. The tree captures the **influence between nodes** (branches represent who transmitted the information to whom) and unfolds in the same order as the activation sequence

How does news travel? Weak and strong ties

- Recap: Mark Granovetter's paper, “The Strength of Weak Ties, published in 1973”
- In social network theory weak and strong ties indicate the levels of friendship/intimacy/trust

regardless the intensity of the tie, if A is strongly connected to B and C, then B is connected to C

people know what their friends know



the novel news travel through the weak ties!

Research tasks

- Detection of popular topics
  - most used terms
  - time periods of trends
- Modelling information diffusion
  - capture/visualize the spreading process
  - predict the spreading process
- Identifying influential information spreaders
  - predict best influencers
  - maximize diffusion by utilizing them

using diffusion prediction models

we will focus on maximizing the spread of influence through a social network

Detection of popular (bursty) topics

- **Challenges**
  - topic detection in dynamic message streams
  - discretization of the stream
  - polysemous terms, combinations of terms, ...

A behavior associated to a topic within a time interval in which it has been extensively treated but rarely before and after

(Textual) Content
- tf-idf models
- LDA-based models

(Non Textual) Content
- utilization of images
- video

(Social) Structure
- authority values of users

overview of algorithms

Guille, Adrien and Hacid, Hakim and Favre, Cecile and Zighed, Djamel A., Information Diffusion in Online Social Networks: A Survey, 2013, ACM SIGMOD

Modelling information diffusion

- **Explanatory models**
  - infer the diffusion process
  - estimate the probability of a node to transmit information to its neighbor
- **Predictive models**
  - two important graph-based models

Independent Cascades (IC) ← sender-centric

Linear Threshold (LT) ← receiver-centric

J. Goldenberg, B. Libai, and E. Muller. **Talk of the network: A complex systems look at the underlying process of word-of-mouth.** Marketing Letters, 2001
M. Granovetter. **Threshold models of collective behavior**. American journal of sociology, 1978.
Seminar in Distributed Computing, Computer Science, ETH Zürich

Independent Cascades

- Directed graph
- Early adapters

A set of users who are first to adopt a piece of information and then trigger its diffusion

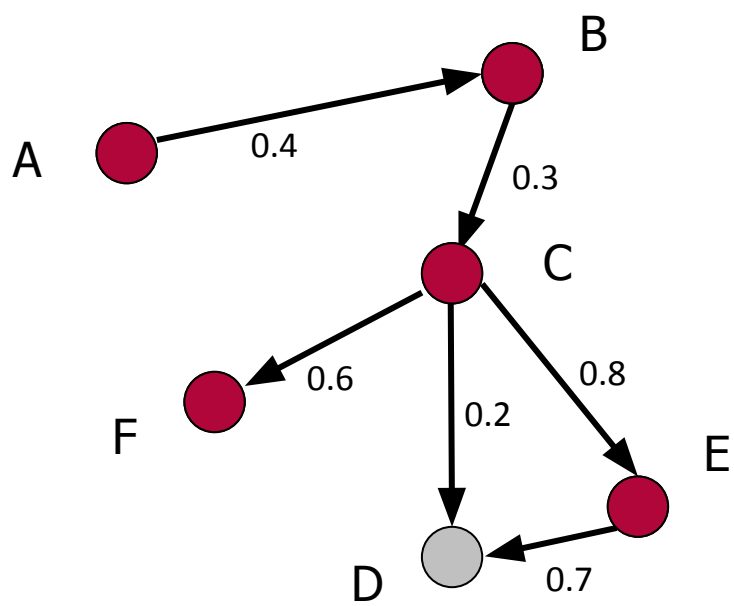
- Nodes can only be activated and activate their neighbors
- Diffusion probability associated to each edge
- Ending criterio: no transmission is possible

Independent Cascades (2)

- sender-centric
- single diffusion chance

- Activation probability: *p*
- Probability of not activating the neighbor: 1-*p*
- Early adopters S={A,C}

Initialization



Second iteration

Third iteration

- Final activated nodes σ(S)=5

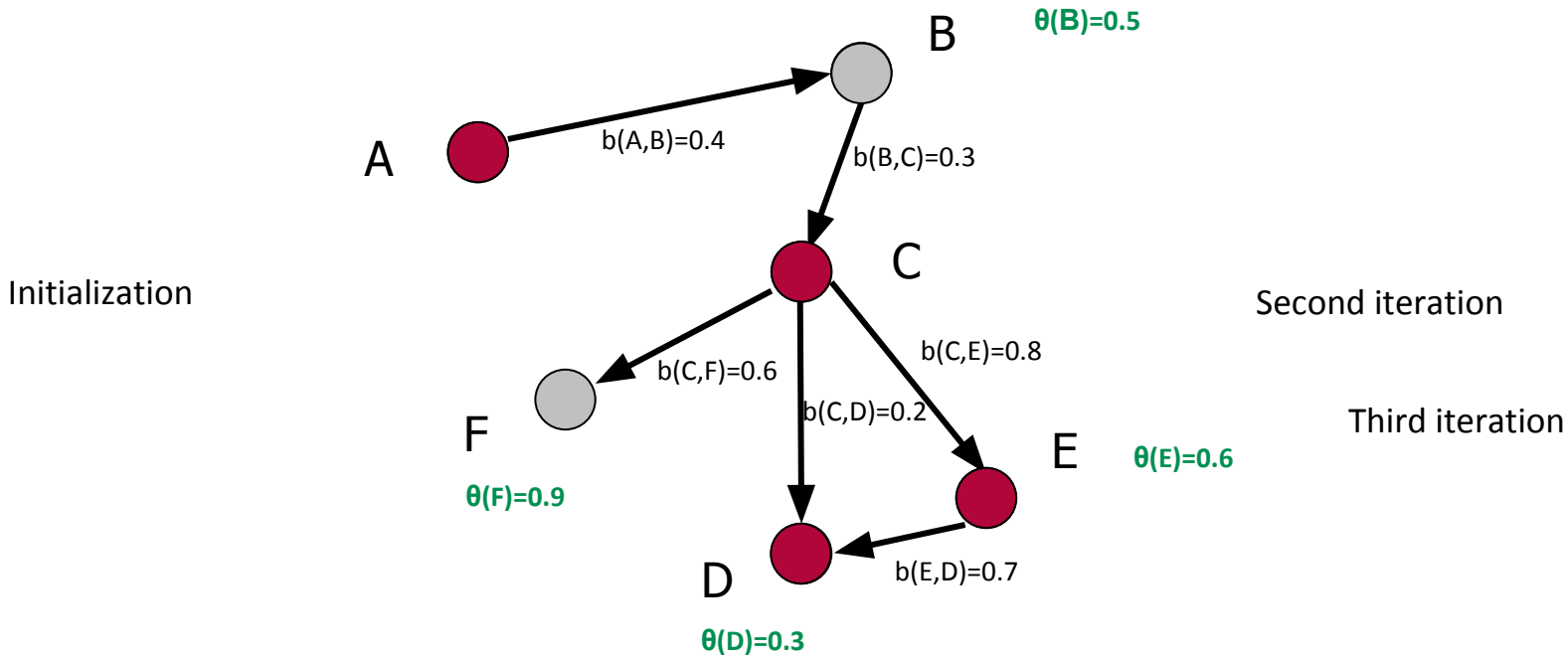
Linear Threshold

- Directed graph
- Early adapters
- Nodes can only be activated and activate their neighbors
- Ending criterion: no transmission is possible
- ~~Diffusion probability associated to each edge~~
    - Influence probability associated to each edge
    - Influence threshold for each node
- Activation criterion
    - sum of the influence probabilities exceeds a node's influence threshold

Linear Threshold (2)

- receiver centric
- more diffusion chances

- Random inactive node threshold: $\theta(v) \in [0,1]$
- Edge influence value: $b(u,v),\ u,v\ neighbors$
- Activation criterion: $\Sigma u:\ active\_neighbor\_of\_v\ b(u,v) >= \theta(v)$
- Early adopters S={A,C}



- Final activated nodes σ(S)=4

Identifying influential information spreaders

- Techniques
  - main focus on distinguishing passive from active users
  - some approaches are PageRank or HITS-based
  - few ones are topic sensitive
- Applications?
  - Trigger large cascades for social media campaigns!

Use best spreaders for influence maximization :

IC and Lt based algorithm from Krempe et. al

D. Kempe et al.,. Maximizing the spread of influence through a social network, KDD, 2003.

Problem definition

- OSN G (V,E)
- Model to capture the spread process (IC or LT)
- S: early adopters (seeds)
- $\sigma(S)$: final active nodes

- Task
  - Choose S^* such that:
    - $\sigma(S^*) = \max \sigma(S)$

- Initial problem: NP-hard!
- Reduced goal: maximum approximation
  - Choose S s.t: $\sigma(S) >= r*\sigma(S^*)$

NP-hardness

- Why is influence maximization an NP hard problem?

- IC: The problem is at least as difficult as the **Set Cover** problem!
- LT: The problem is at least as difficult as the **Vertex Cover** problem!

Approximation with guarantee

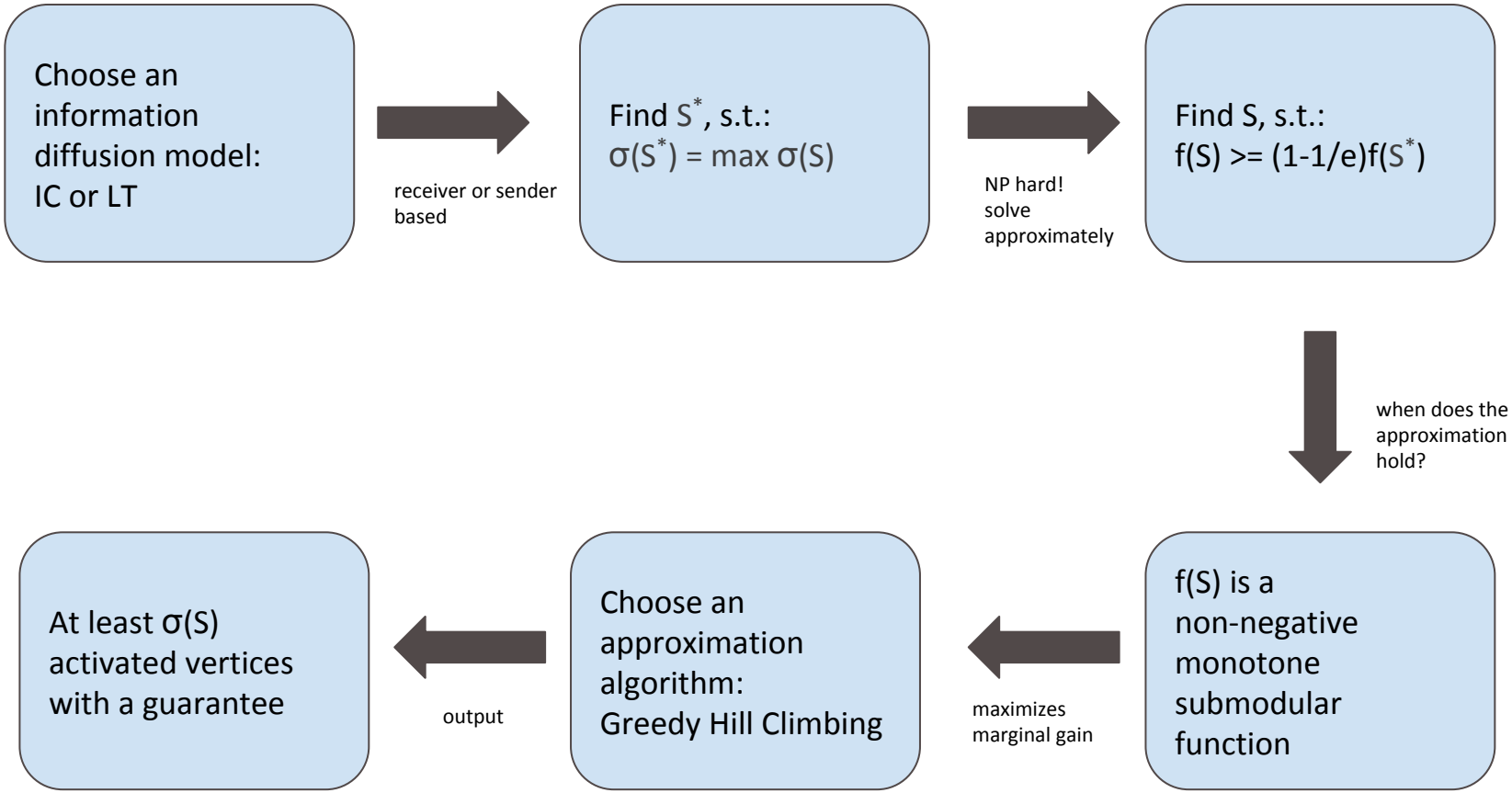
- $\sigma(S) >= r*\sigma(S^*)$
- … Find S, s.t:
  - $f(S) >= k*f(S^*)$, $k = (1 - 1/e)$
- The condition will be reach iff
  - σ is submodular!

the difference in the incremental value of the function *f* that a single element makes when added to an input set decreases as the size of the input set increases

$$f(S \bigcup \{v\}) - f(S) \geq f(T \bigcup \{v\}) - f(T)$$

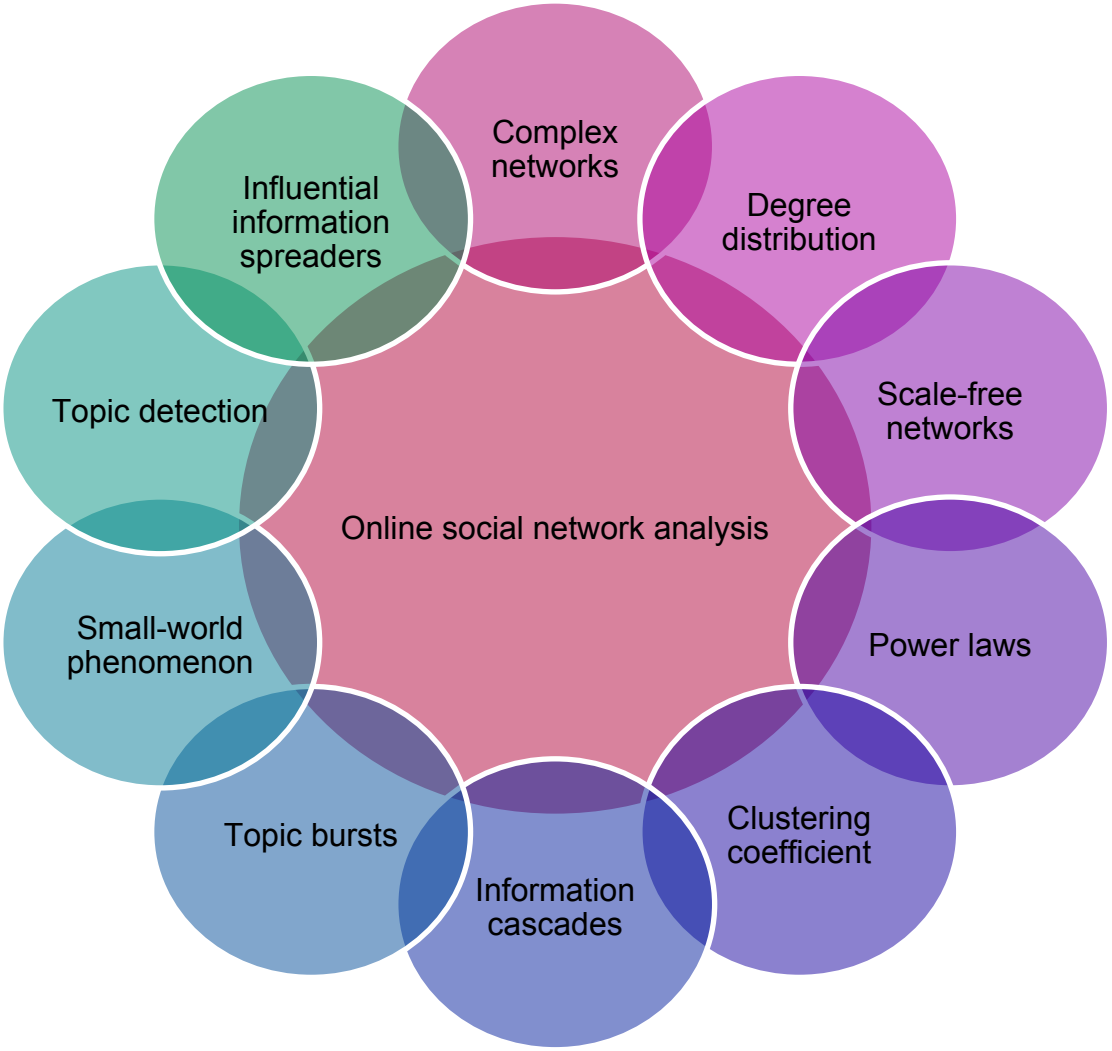
- For both IC and LT, $\sigma(S)$ can be proven to be submodular
- The proof is based on the fact that linear combinations of submodular functions are submodular as well

Krempe et. al, KKD, 2003



Thank you for your attention!

Are there any questions?



In the next episode …

Graph querying

Frequent subgraph mining

And much more …