RUMORS, EPIDEMICS AND ECONOMIC CRISES: INFECTION PROCESSES IN NETWORKS

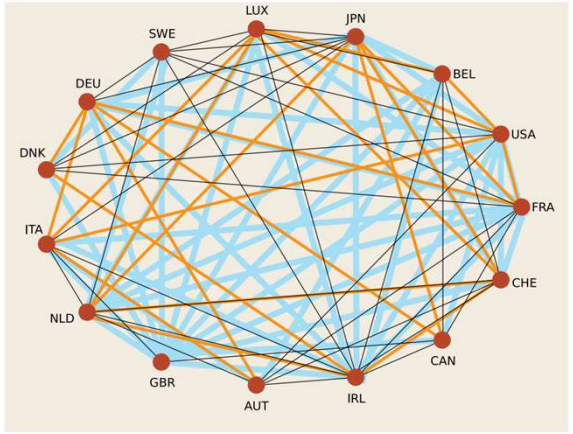
Famnitovo izleti v matematično vesolje

Miklós Krész

Univerza na Primorskem,
Fakulteta za Matematiko, Naravoslovje in Informacijske Tehnologije
February 19th, 2025

Prologue

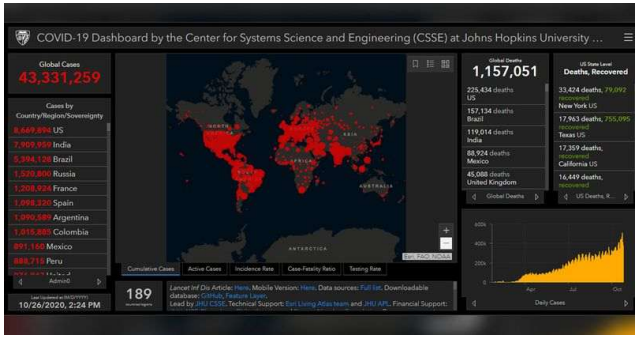
2008 financial crisis



Financial flows are extensive among the 15 advanced economies at the core of the global banking network.

Source: Minoiu, Camelia, and Javier A. Reyes, 2011, "A Network Analysis of Global Banking: 1978–2009," IMF Working Paper 11/74

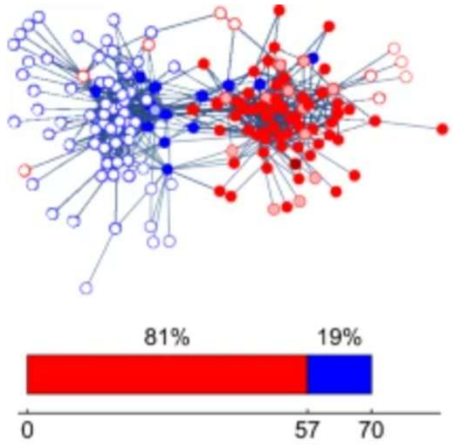
COVID-19



An interactive web-based dashboard

Source: John Hopkins University

Misinformation



„Echo chambers" in social media

Source: Avin, C., Daltrophe, H. & Lotker, Z. On the impossibility of breaking the echo chamber effect in social media using regulation. *Sci Rep* **14**, 1107 (2024).

„Data age”

- In 1960’s the sum of digital data was 8 GB …
- A revolution in the relationship with data (science, industry and society)

Source: Piackutatas.hu

Predicting the future?

Data is not enough, for information we also need models!

„Back to the future ...”

... 2013

Big Data/Data Science:

„Big Data is information about people's behavior instead of information about their beliefs. It's about the behavior of customers, employees, and prospects for your new business. This sort of Big Data comes from things like location data off of your cell phone or credit card, it's the little data breadcrumbs that you leave behind you as you move around in the world..”

„As a consequence, analysis of Big Data is increasingly about finding connections, connections with the people around you, and connections between people's behavior and outcomes.”

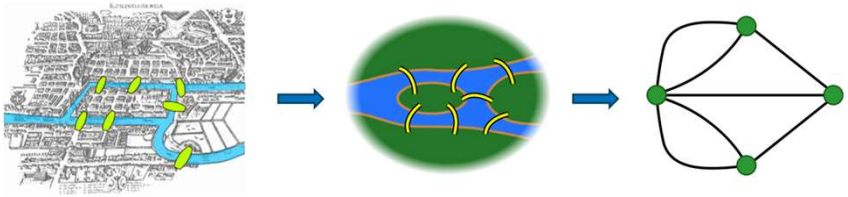
Alex „Sandy” Pentland

Source: Piackutatas.hu

Source: Wikipedia

... Graphs everywhere?

Graphs

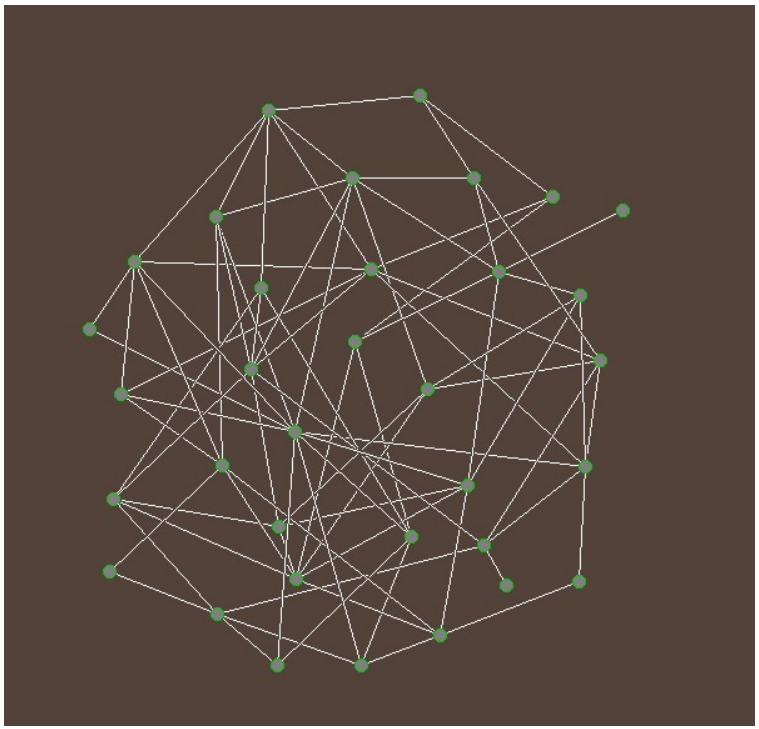


- Graph G = (V, E)
  - V: set of nodes or vertices
  - E: set of edges: (A, B), pairs of vertices A and B
- A graph can be directed or undirected
  - Edges are ordered or unordered pairs
- A graph can be weighted or unweighted
  - A weight function assigns weights to the edges
  - In real world graphs, weights can represent important information (distance or similarity)
- Degree of a vertex: the number of incident edges
- Sparse/dense graphs depending on the „edge density”

Source: https://thebittimes.com/how-traditional-finance-world-could-cause-traditional-global-financial-crisis-tbt3286.html

Random graphs and social networks

Erdős-Rényi model (1959)

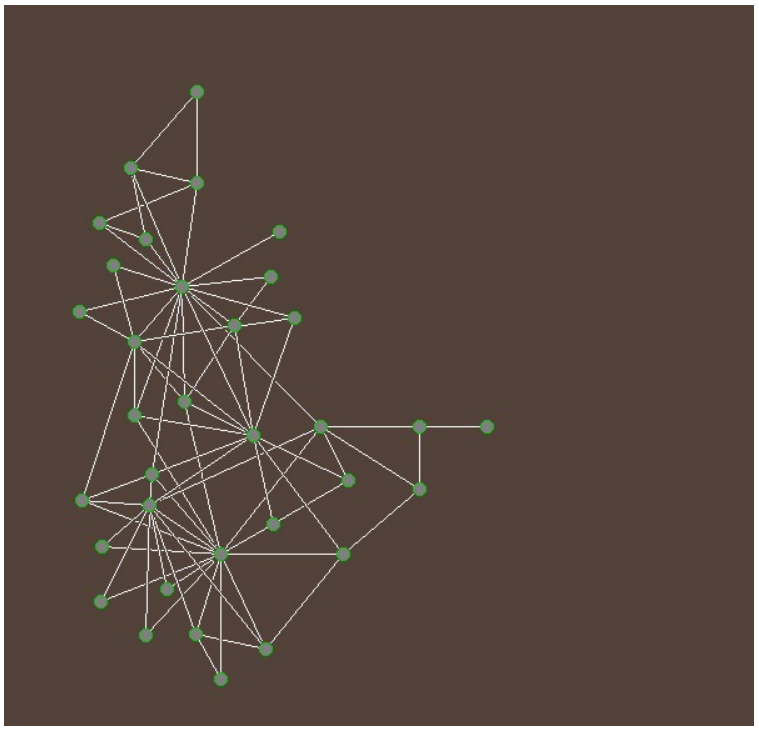


The „density" probability $0 \leq p \leq 1$ as a „coin flipping" for each pair of nodes.

Pál Erdős
1913-1996

Alfréd Rényi
1921-1970

Zachary's karate club (1977)



Zachary, W. W. (1977). "An Information Flow Model for Conflict and Fission in Small Groups". *Journal of Anthropological Research*. **33** (4): 452–473.

Erdős-Rényi modell degree distribution

Probability that a node has degree k for all $0 \leq k \leq (n-1)$ with n being the number of nodes?

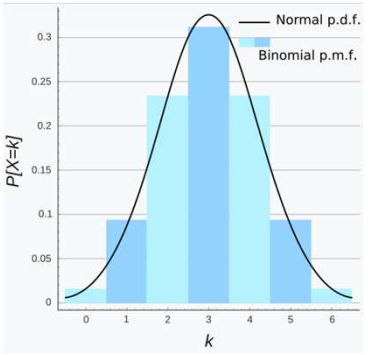
- A node has degree zero if all coin flips are "tails"
- A node has degree $(n-1)$ if all coin flips are "heads"
- For a node having degree k, the $(n-1)$ coin flips must have resulted in k "heads" and $(n-1-k)$ "tails"
- Probabaility with k „heads" and $(n-1-k)$ „tails":

$$p^k(1-p)^{n-1-k}$$

- There are exactly "(n−1) choose k" ways in which this outcome can occur:

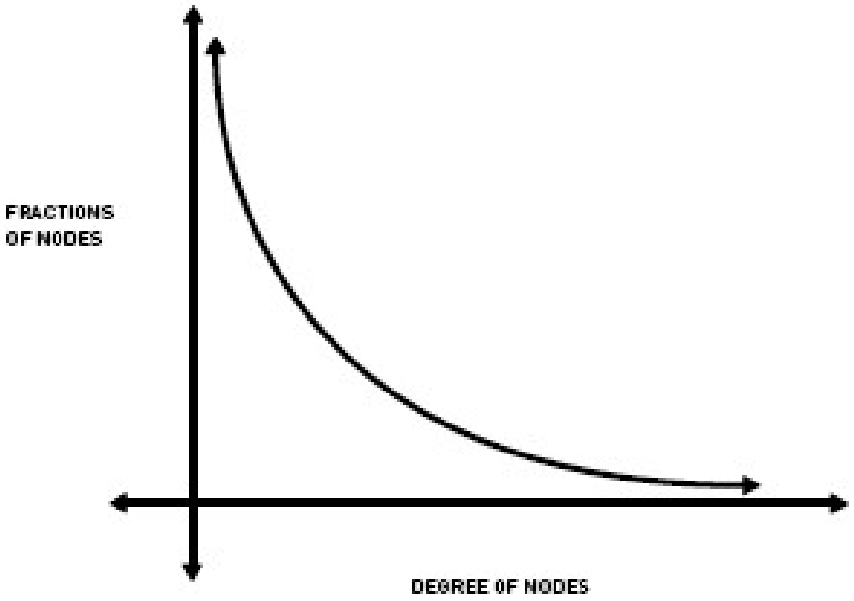
$$\binom{n-1}{k} p^k(1-p)^{n-1-k}$$

- Thus, the probability that a given node has degree k is given by the ***Binomial distribution*** with approximated for large n by the ***Normal distribution***

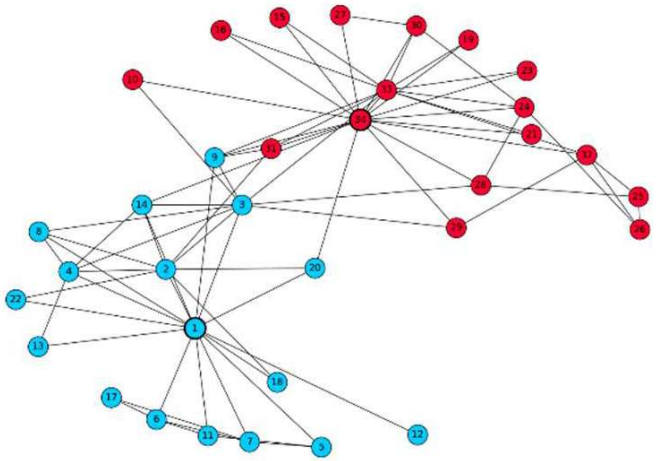


Small-world graphs

- „Small" number of edges (~ no. of vertices by constant)
- „Small" diameter (length of the shortest path between the most distanced nodes)
- „Dense" parts (highly clustered)
- Degree-distribution by power-law ($P(k) \sim k^{-\gamma}$, $\gamma>0$, typically $2<\gamma<3$)



„Real clusters" of the Zachary graph

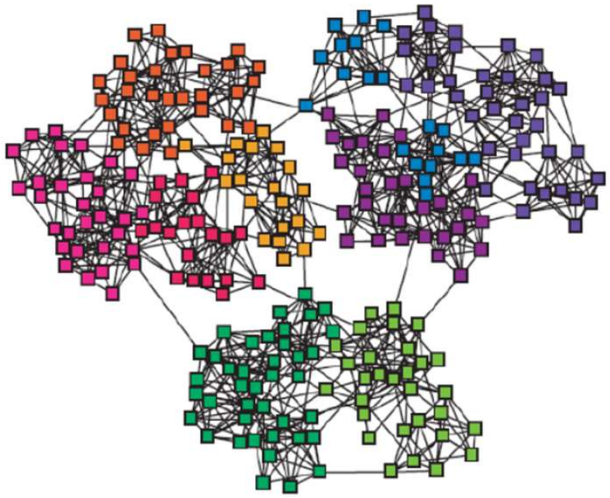


Predictive models for the real processes?

Graph Clustering: community detection in graphs

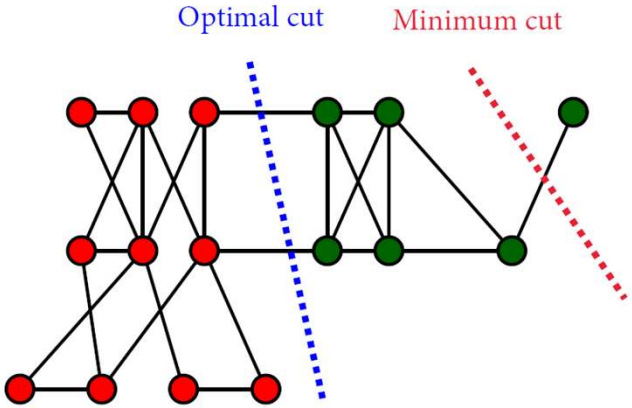
Clustering: Unsupervised learning, partitioning a data set

Communities: partition of nodes forming dense subgraphs



Evaluation: what is a „good" clustering?
- Which is a „good" objective function?

Example:
minimum cut – minimize weight of connections between groups



Normalized cut (Shi-Malik 97):

$$ncut(A,B) = \frac{cut(A,B)}{vol(A)} + \frac{cut(A,B)}{vol(B)}$$

Vol (A) is the sum of weights of edges having at least one endpoint in *A*

Minimum cut does not consider internal cluster connectivity

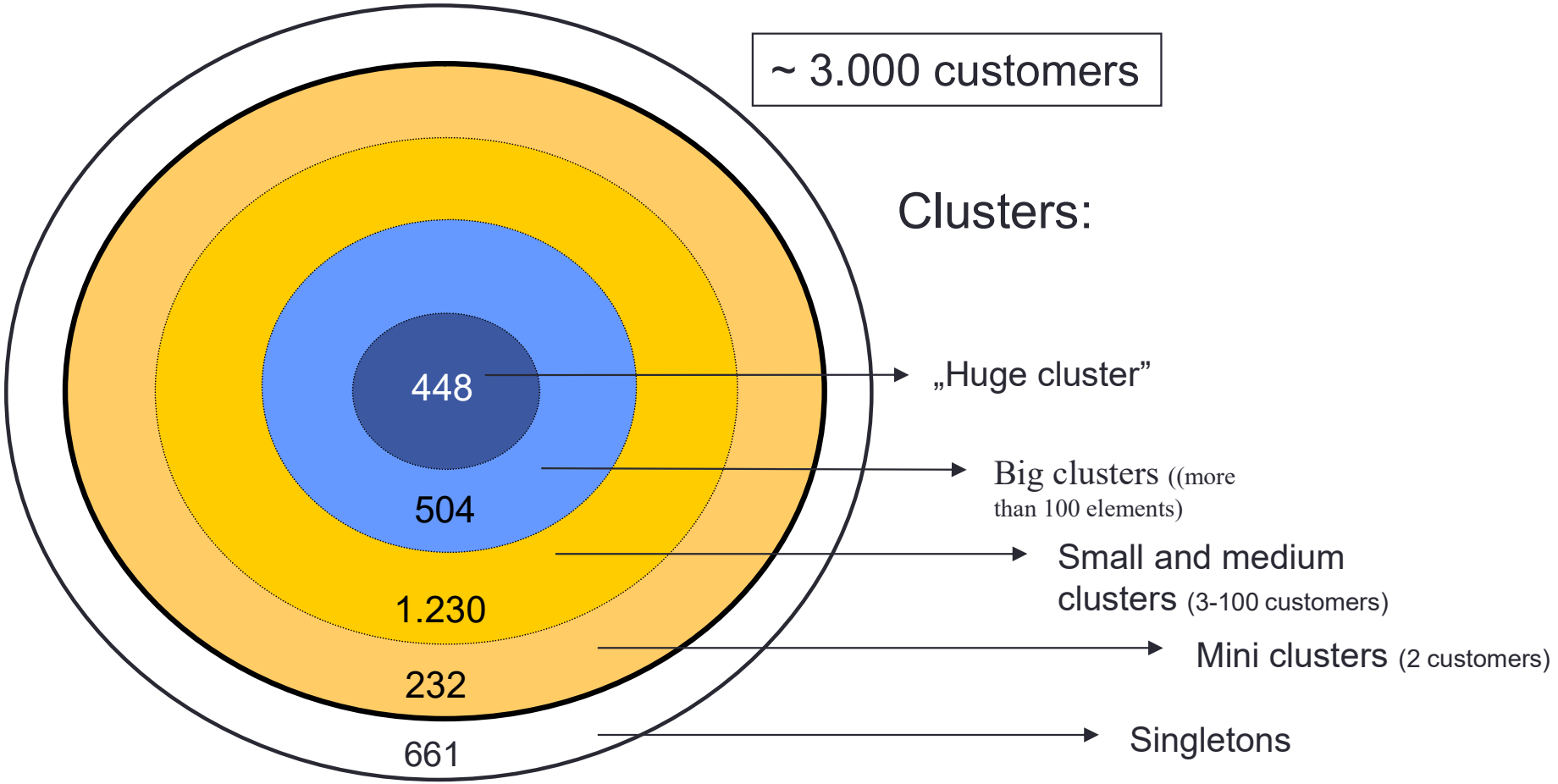
Business example: churn (customer attrition) prediction

Churn probability:
The probabilty for each customer to „move out” in a given time period.

Cost of retaining an existing customer is far less than acquiring a new one.

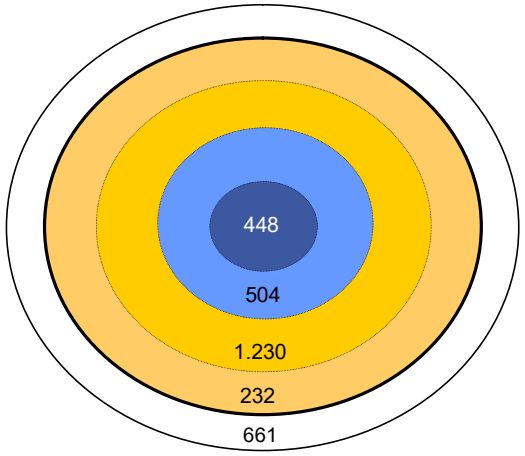
Churn prediction by clustering

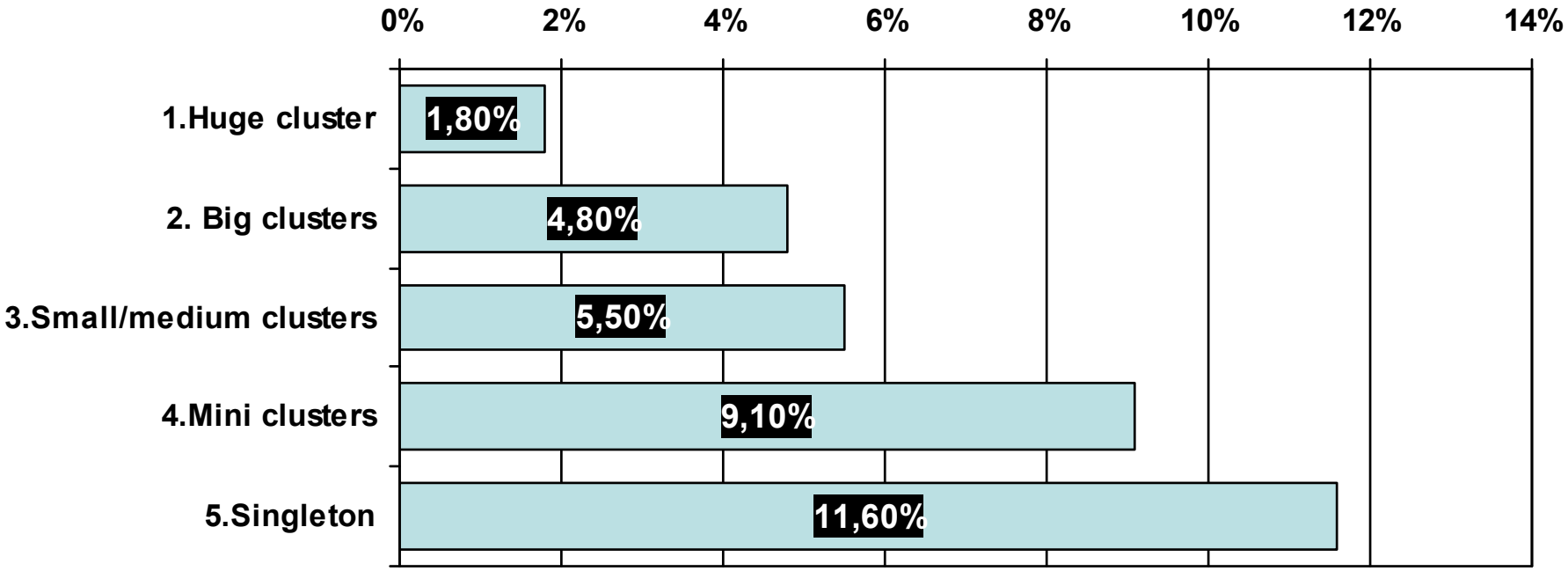
A pilot project at a mobile company for a special group („fleet”) of approx. 3000 customers



Churn results

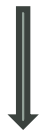
What is the churn ratio in half a year in the given segments?





Limitations of community detection

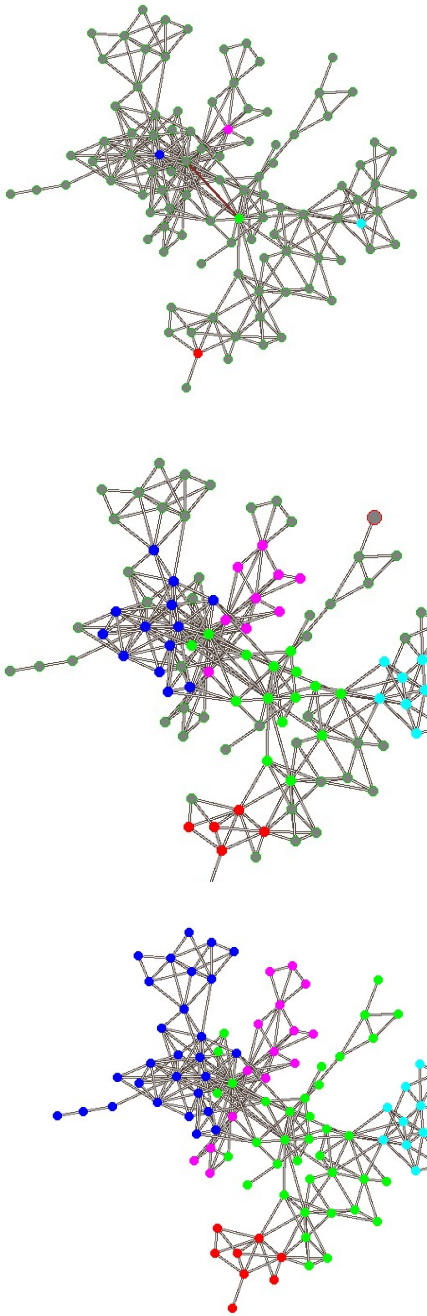
- Objective function can highly depend on the application – no general methodology
- Methods are based on the graph structure, not specialized on the real process
- In addition to predicting the expected outcome, intervention strategy should be identified. E.g. who are the critical persons in the club to be „targeted” in order to prevent the future event?



It is not enough to identify the potential „crisis situation”, we need to have strategic decision for „immunization” (state aid in financial crisis, vaccination strategy in pandemics, decreasing the „churn” etc)

Infection models

- Inputs
  - Graph G(V, E)
  - A weight w_e on all of the edges with $0 \leq w_e \leq 1$
  - Infection transmission model on neighboring nodes
  - The set of initial infectors A_0
- Output
  - The set of infected vertices A_t
- Iterative process
- States
  - Susceptible (S)
  - Infected (I)
  - Recovered (R)
- Origin
  - Epidemiology
  - Sociology
  - Economics



Independent Cascade Model

Independent Cascade Model

- Domingos-Richardson (2001)
- Kempe-Kleinberg-Tardos (2003)

Algorithm:

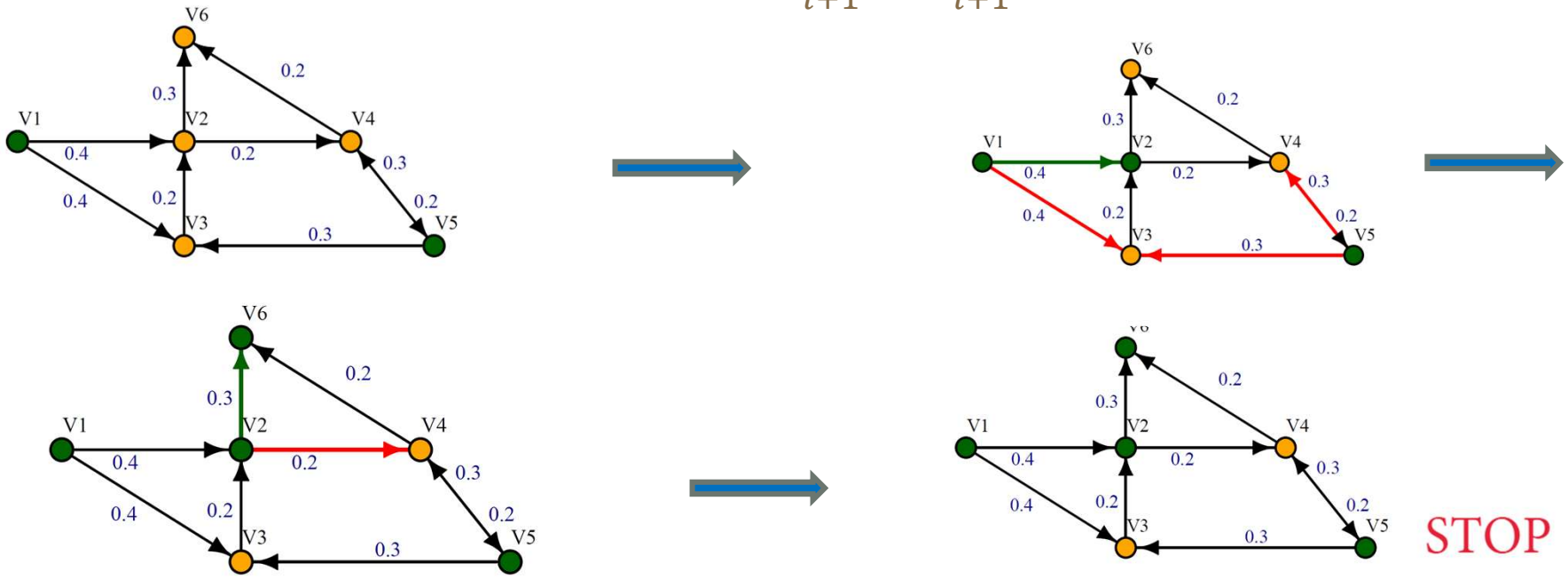
While $A_i \neq \emptyset$:

$A_i \leftarrow newly\ infected\ nodes$

$\forall v \in A_i\ tries\ to\ infect\ their\ neighbours\ with\ p(v,u)$

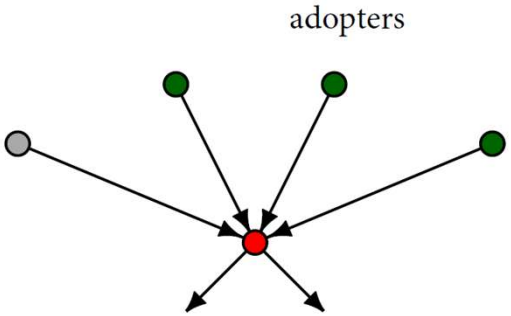
If $(the\ infection\ is\ succesful)$:

$A_{i+1} = A_{i+1} \cup u$



Linear Threshold Model

Linear Threshold Model
- Granovetter (1978)
- Kempe-Kleinberg-Tardos (2003)



Algorithm:

For each node *v* a random nonnegative threshold *T(v)* is generated

While $A_i \neq \emptyset$:

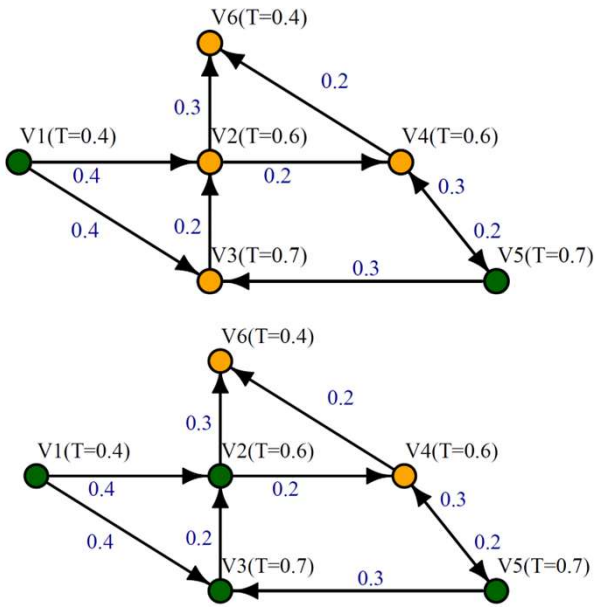
$A_i \leftarrow newly\ infected\ nodes$

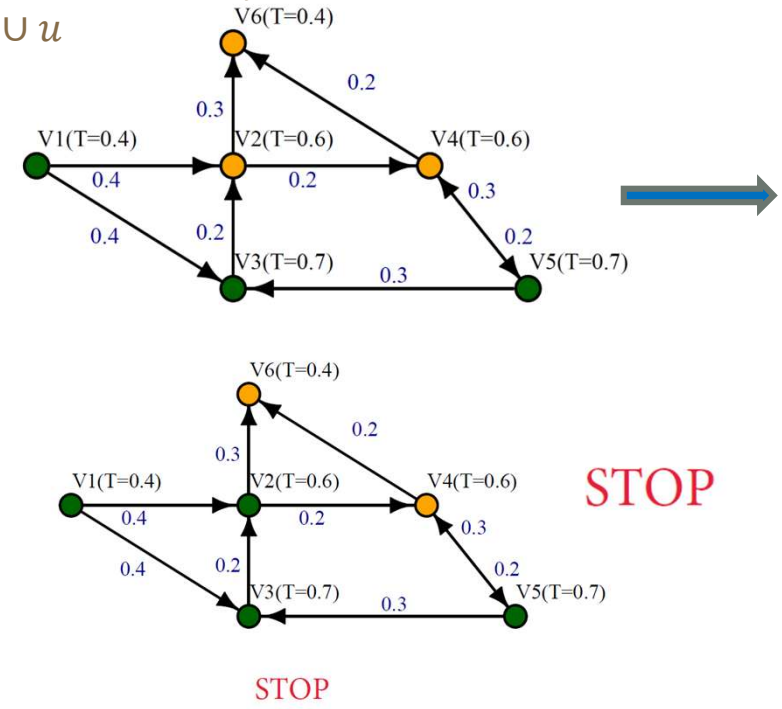
for each inactive node u, the infection is tried by the incoming edges *e*.

successful, if $\sum_e w_e \geq T(u)$

If $(the\ infection\ is\ succesful)$:

$A_{i+1} = A_{i+1} \cup u$





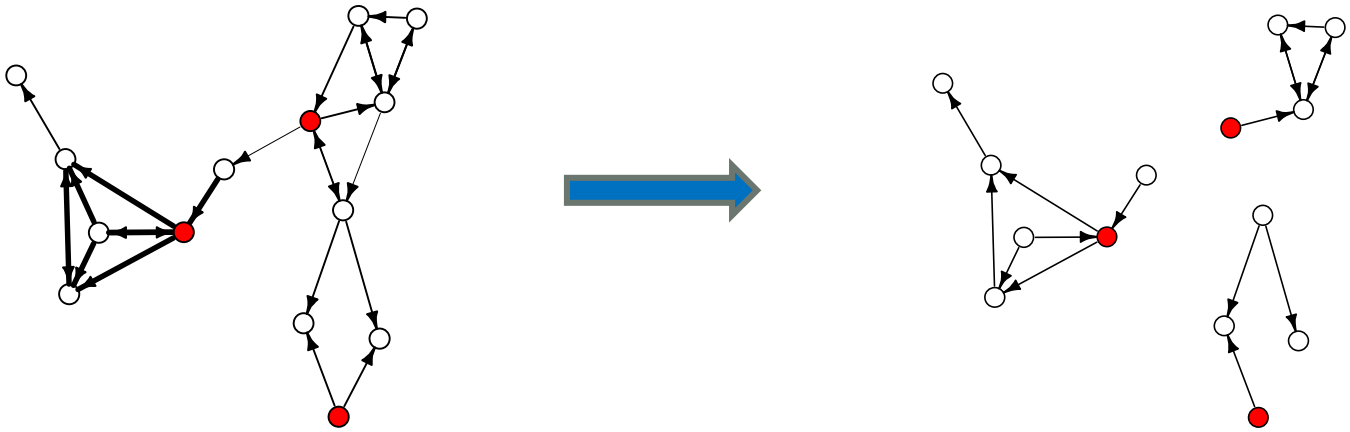
STOP

Determining the influence value

Stochastic process: expected value of infection probabilities (influence value)

Theorem. (Chen, 2010) Determining the influence value is Np-hard for both the Independent Cascade and the Linear Threshold Model.

Theorem. (Kempe-Kleinberg-Tardos, 2003) The influence value can be arbitrarily approximated by simulation for both the Independent Cascade and the Linear Threshold Model.



Complexity of algorithmic problems

Problems in ***P*** : solution in polynomial number of steps (with respect to the input size)

Problems in ***NP*** : certifying the solution in polynomial number of steps

King Arthur

Merlin

Source: Wikipedia

Example: Integer factorization. For every natural numbers n and $k<n$, does n have factor smaller than k besides 1?

NP-solution: As a positive answer a factorization with a factor smaller than k and as a negative answer a factorization with all factors at least k (with the primality test of Agrawal–Kayal–Saxena, 2002)

NP-hard problems : „polynomial subroutines" for any problems in NP

Complexity: The Limits of computation

Consider a problem for which there exist an algorithm with 2^n time complexity, i.e. it makes at most 2^n steps on any n-length input.

Which is the maximum size of an input for which a supercomputer can give a solution during the 14 thousand million years of the universe?

The theoretical bound for making an operation: $t = 2 \times 10^{32} s$

The life time of the universe: $T = 14 \times 10^9 \, years = 4.42 \times 10^{17} s$

The number of operations: $\frac{T}{t} = \frac{4.42 \times 10^{17} s}{2 \times 10^{32} s} = 2.21 \times 10^{49}$

$$2^n \leq 2.21 \times 10^{49}$$

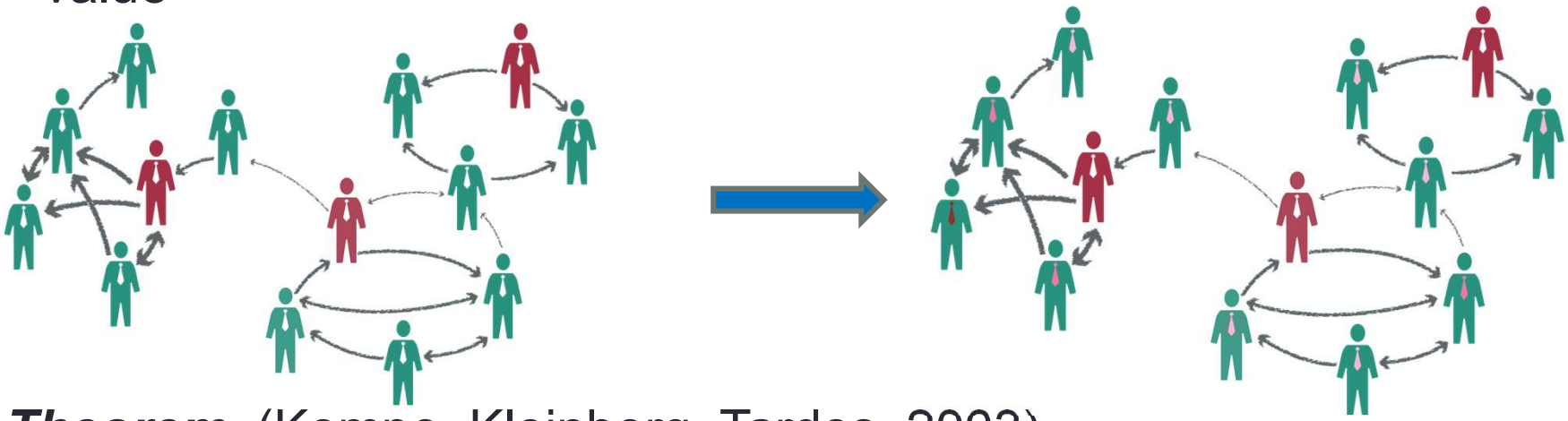
$$(n-1)\lg 2 = \lg\left(1.105 \times 10^{49}\right)$$

$$n - 1 = \frac{\lg\left(1.105 \times 10^{49}\right)}{\lg 2} = 162.93$$

$$n < 163$$

Influence maximization

- With fixed set size determine the influencers maximizing the influence value



Theorem. (Kempe, Kleinberg, Tardos, 2003)

The influence maximization problem is NP-hard, but can be approximated with *(1-1/e)* by the greedy method.

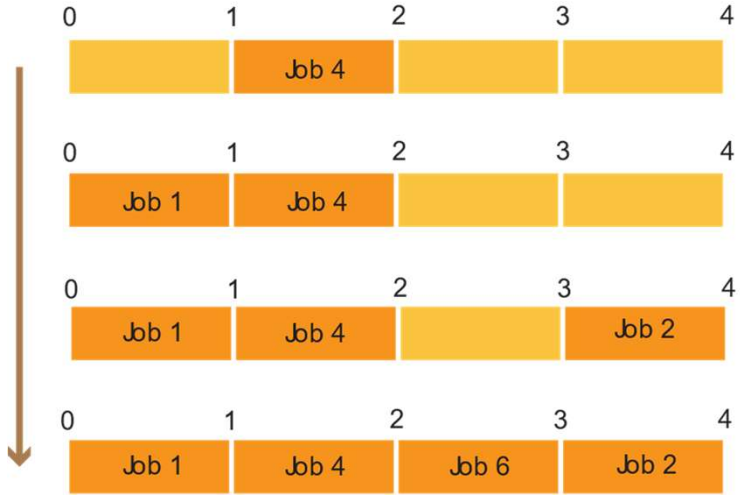
Note. It gives an approximation of at least 63%, but in practice it is much more efficient.

Greedy method

A **greedy algorithm** follows the problem-solving of making the locally optimal choice among the feasible solutions at each stage.

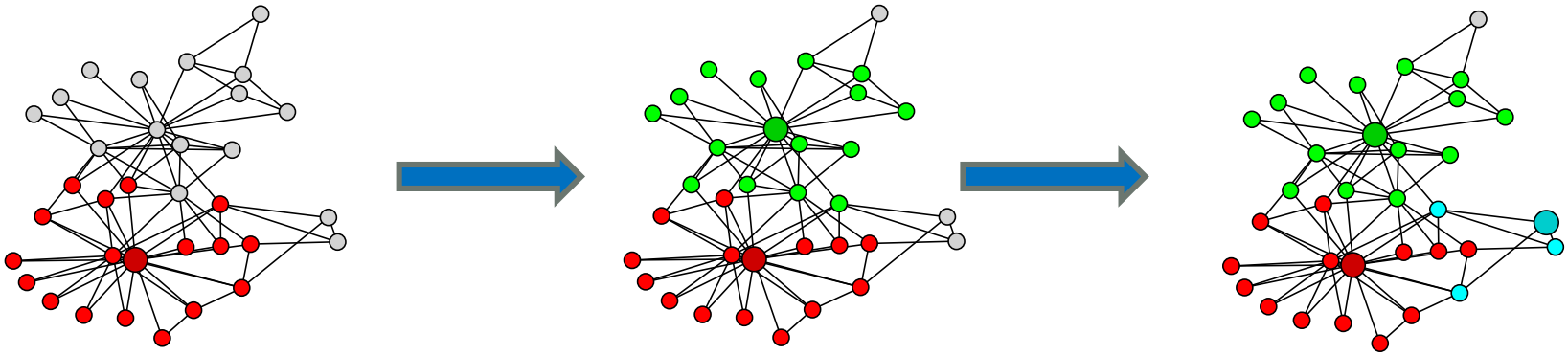
Example: Optimal unit job sequencing with deadlines.

Jobs	Deadlines	Profits
Job 1	1	200
Job 2	4	180
Job 3	2	190
Job 4	2	300
Job 5	1	120
Job 6	3	100



For most of the problems greedy does not provide optimal solution, but frequently used for approximation because of its efficiency.

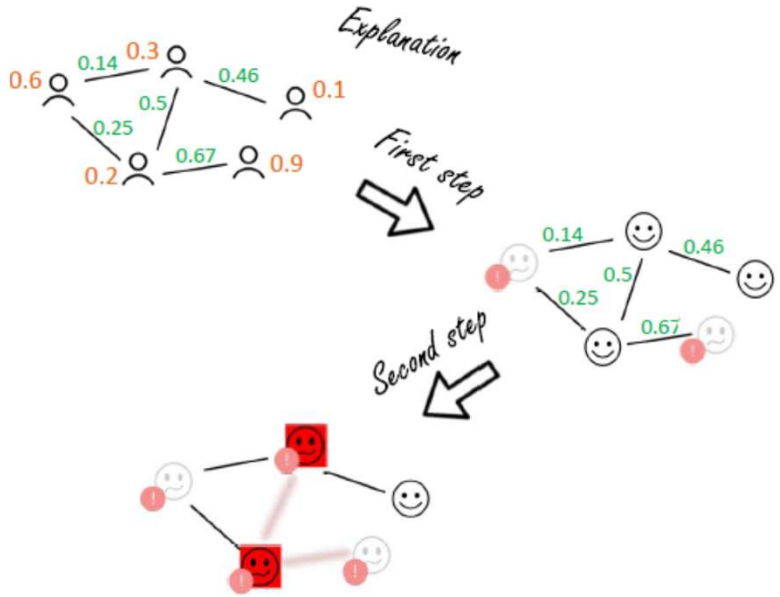
Greedy for influence maximization



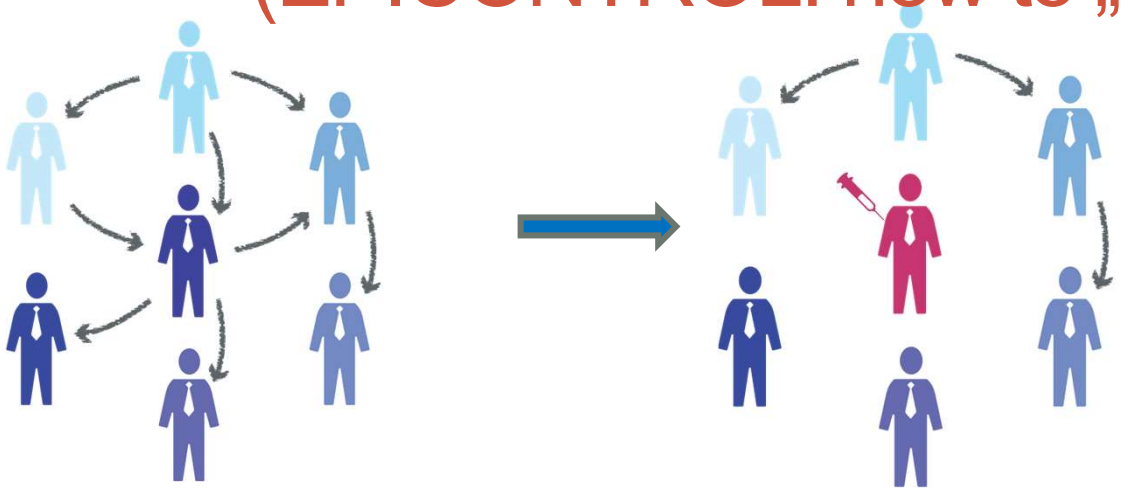
Influence value: submodular function

Generalized infection models

- Initial infections → A priori infection values (coin flipping for initially infected nodes)
- Expected (output) infection values → A posteriori infection values
- More realistic process
  - A priori infection values: external influence

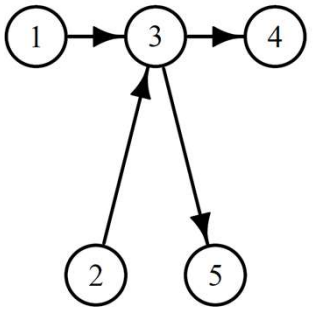


The Network Immunization Problem (EPICONTROL: how to „vaccinate”)



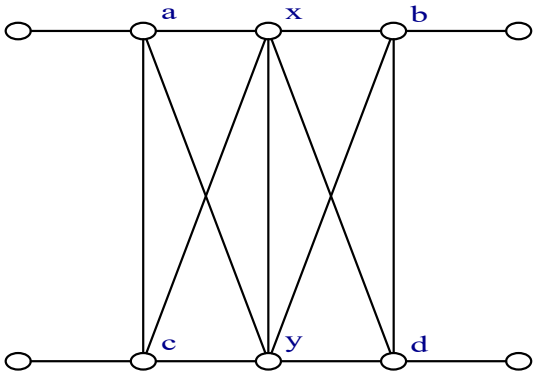
***EPICONTROL**:* finding a fixed size set of nodes such that by their removal the aposteriori infection will be minimum

Influence maximization and network immunization are different:

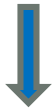


Properties of Network Immunization

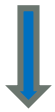
- Apriori infections: *0.1*
- Transmission infections*: 1.0*



- Best single choice: *a,b,c,d*
- Best pair choice: (*x,y)*



EPICONTROL is NOT submodular (no efficient greedy)

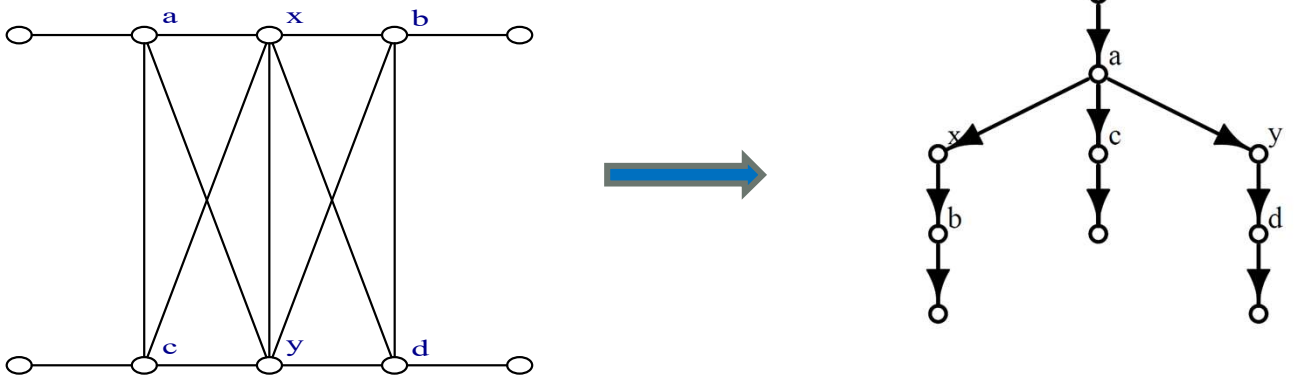


Optimal testing choice ?
(for vaccination)

Influence monitoring problem

Influence monitoring vs. EPICONTROL : Vaccinate those who were chosen for testing?

The influence monitoring maximizing the expected value of the number of nodes rooted from the given set of nodes in a random branching process.



For *k=1*, nodes *x* and *y* are optimal solutions, for *k=2 (x,y)*

The influence monitoring problem can be approximated with *(1-1/e)* by the greedy method. (by submodular property)

Epilogue

- Data and models are available, their integration are the main challenges (e.g. Which are the best fitting infection model, which are the realistic edge weights?) Modern Machine Learning can help a lot.

- Pros: In crisis situation we will have very sophisticated solutions.

- Cons: „Big Brother" knows „everything"

A still shot from Michael Radford's film '1984', based on George Orwell's novel

Thank you for your attention!