Models and algorithms for network immunization

Aris Gionis

Basic Research Unit, HIIT

University of Helsinki

a brief introduction...

- ...originally from Greece
- BS, University of Athens, Greece
- MS and PhD, Stanford University, USA
- PhD adviser: Rajeev Motwani
- Thesis title: *"Algorithms for similarity search and clustering in large data sets"*, July, 2003
- in Basic Research Unit, HIIT, Finland, since August 2003

Basic Research Unit, HIIT

- research
  - Heikki Mannila
  - Panayiotis Tsaparas
  - Niina Haiminen, Evimaria Terzi
  - external collaborators: Foto Afrati, Christos Faloutsos, Spiros Papadimitriou, Alex Hinnenburg, ...
- co-supervising students
  - Niina Haiminen, Evimaria Terzi
- teaching courses
  - data mining, approximation algorithms, computational complexity, spectral methods for data mining

Research paradigm in BRU

- develop novel data analysis techniques for use in other sciences
- combine basic research in computer science with applications
  - look at data analysis problems arising in practice
  - abstract new computational concepts from them
  - analyze, develop new computational methods
  - take the results into practice

$\Rightarrow$ theoretical work in algorithms and foundations of data analysis can have fast impact in the application areas

$\Leftarrow$ the applications feed interesting novel questions to theoretical research

Recent projects

- sequence analysis
  - biology, genetics, physics, telecommunications
- analysis of spatial data
  - biology, ecology
- ordering problems
  - paleontology
- clustering
- analysis of 0–1 matrices

...rest of the talk...

Models and algorithms for network immunization

joint work with George Giakkoupis, Evimaria Terzi, and Panayiotis Tsaparas

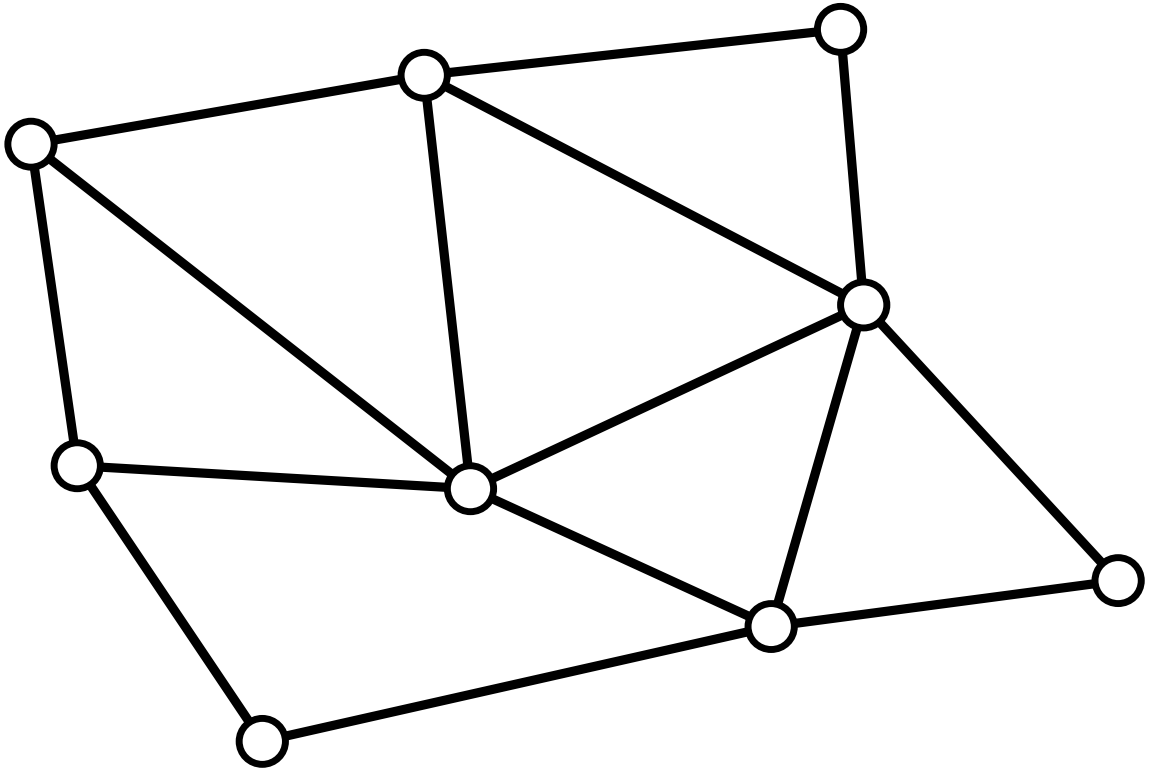
Genome segmentations

joint work with Niina Haiminen, Evimaria Terzi, Heikki Mannila

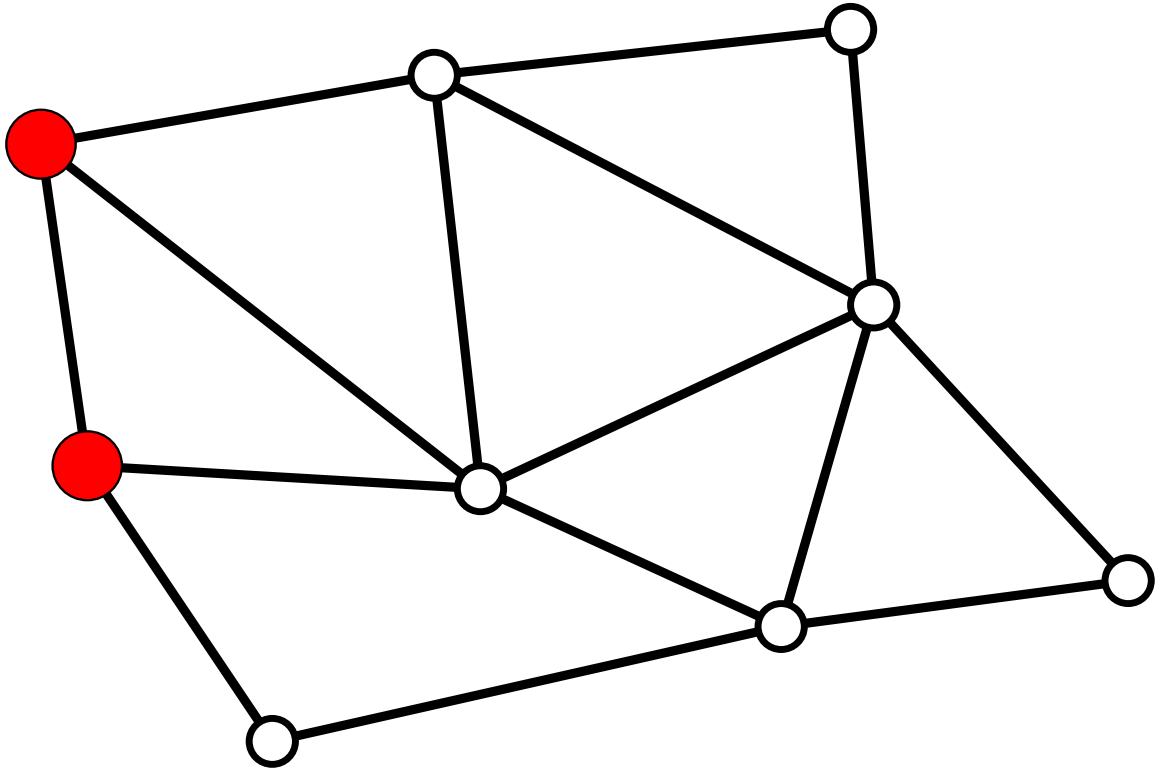
Motivation

- many natural or man-made systems are organized as networks
  - internet, web, social networks, protein networks, etc.
- operation is threaten by the propagation of a harmful entity through the network
  - diseases in social networks
  - gossip or panic in social networks
  - failures in power grids
  - computer viruses on the internet
- can we restrict the spread of the virus in the network?

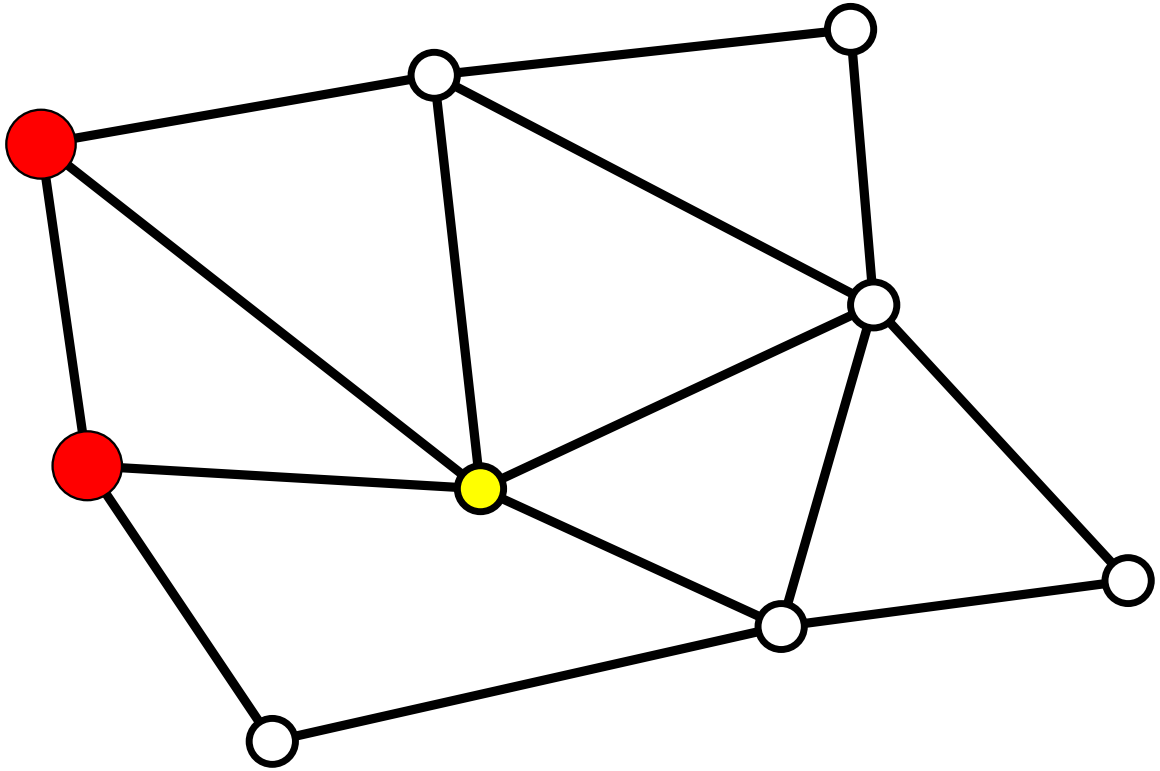
Virus spread



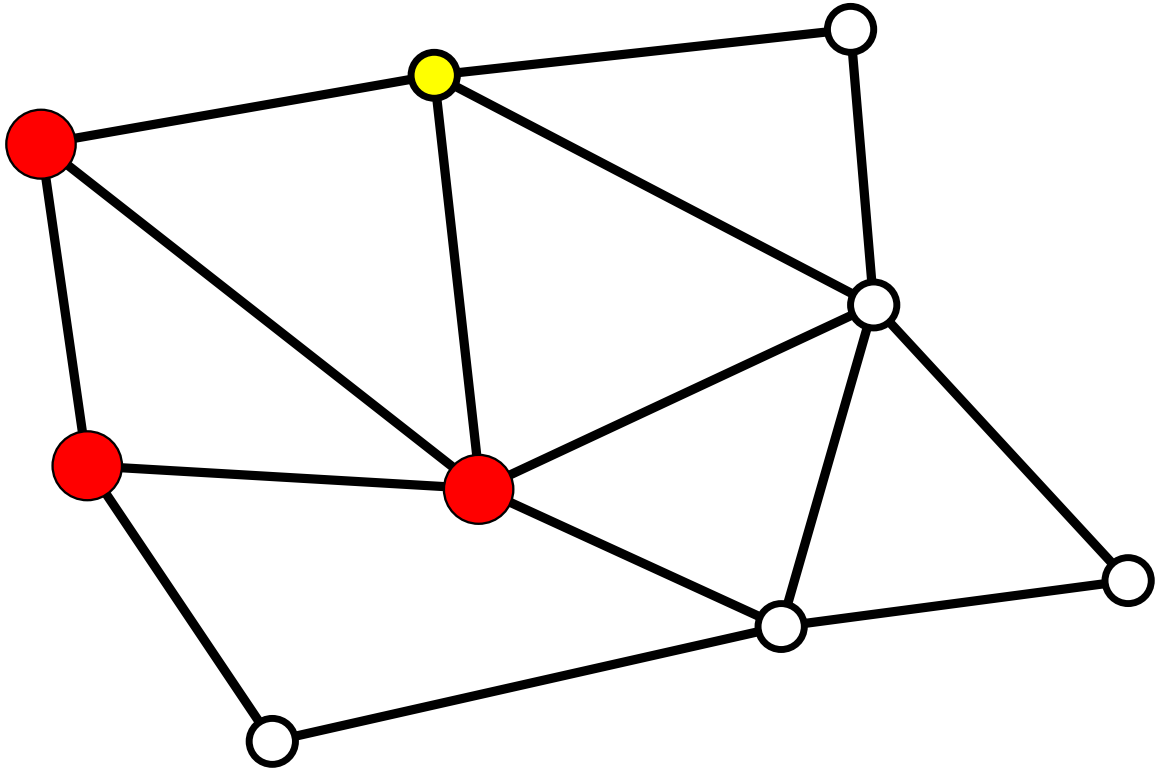
Virus spread



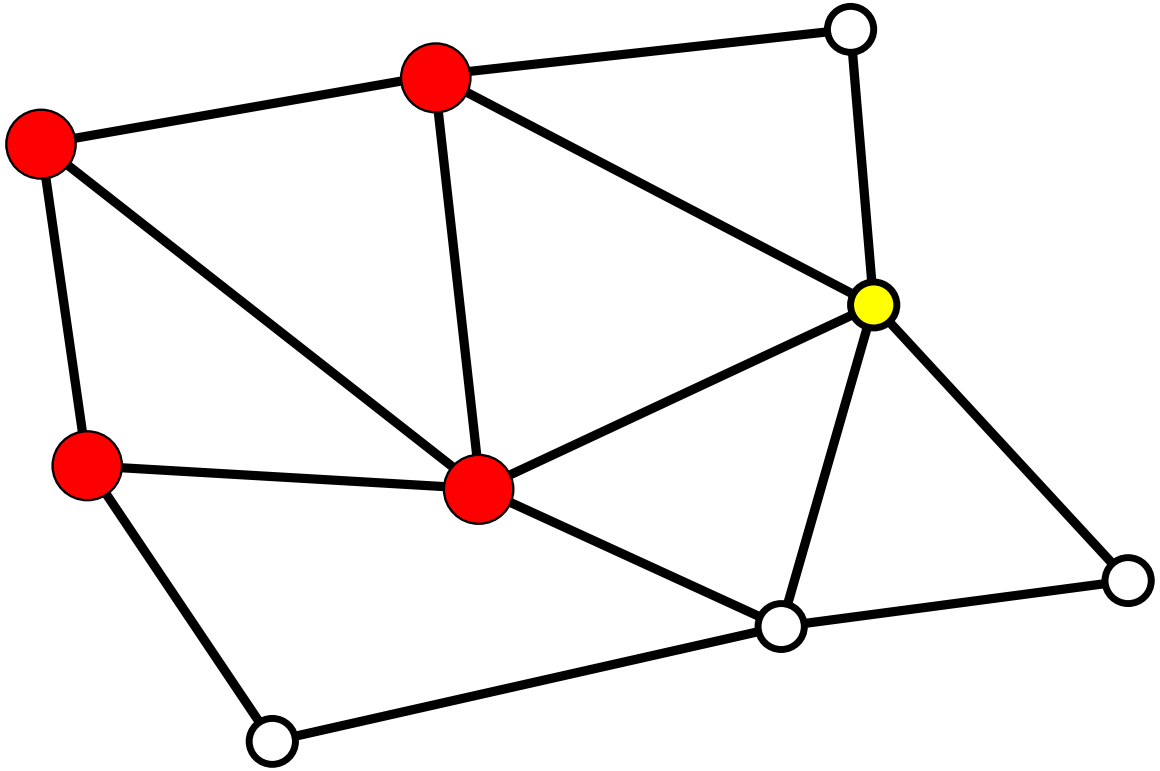
Virus spread



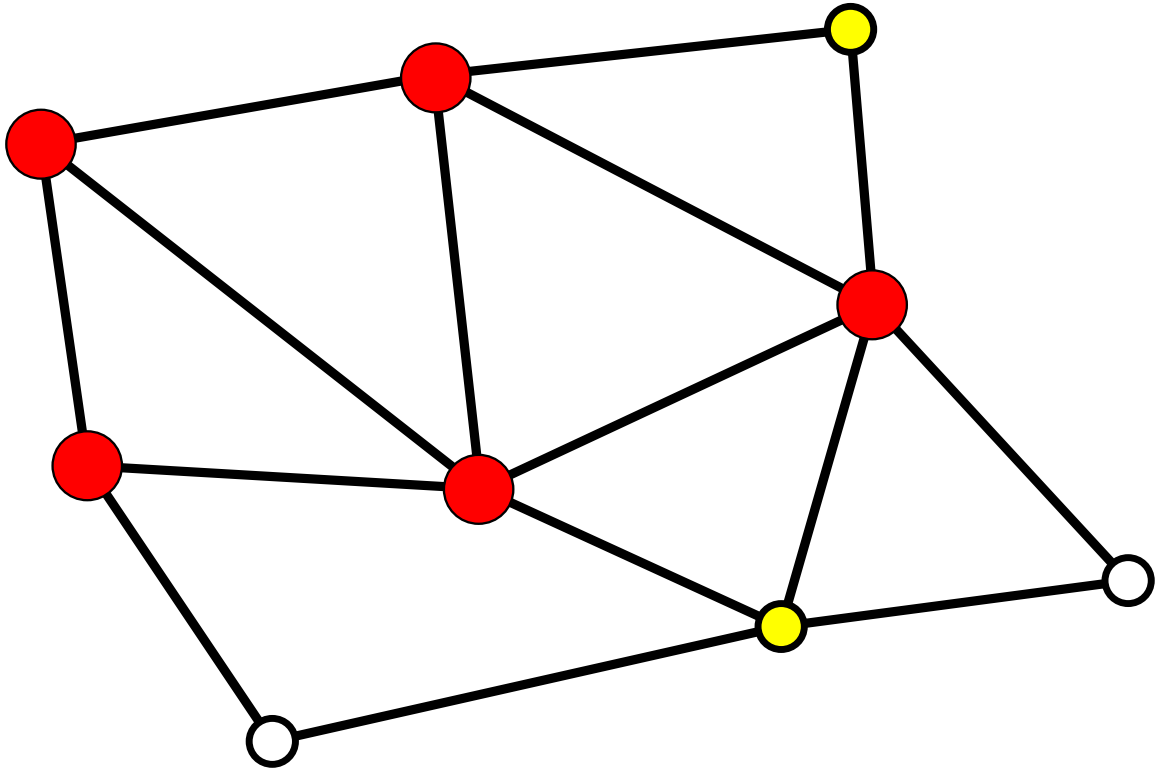
Virus spread



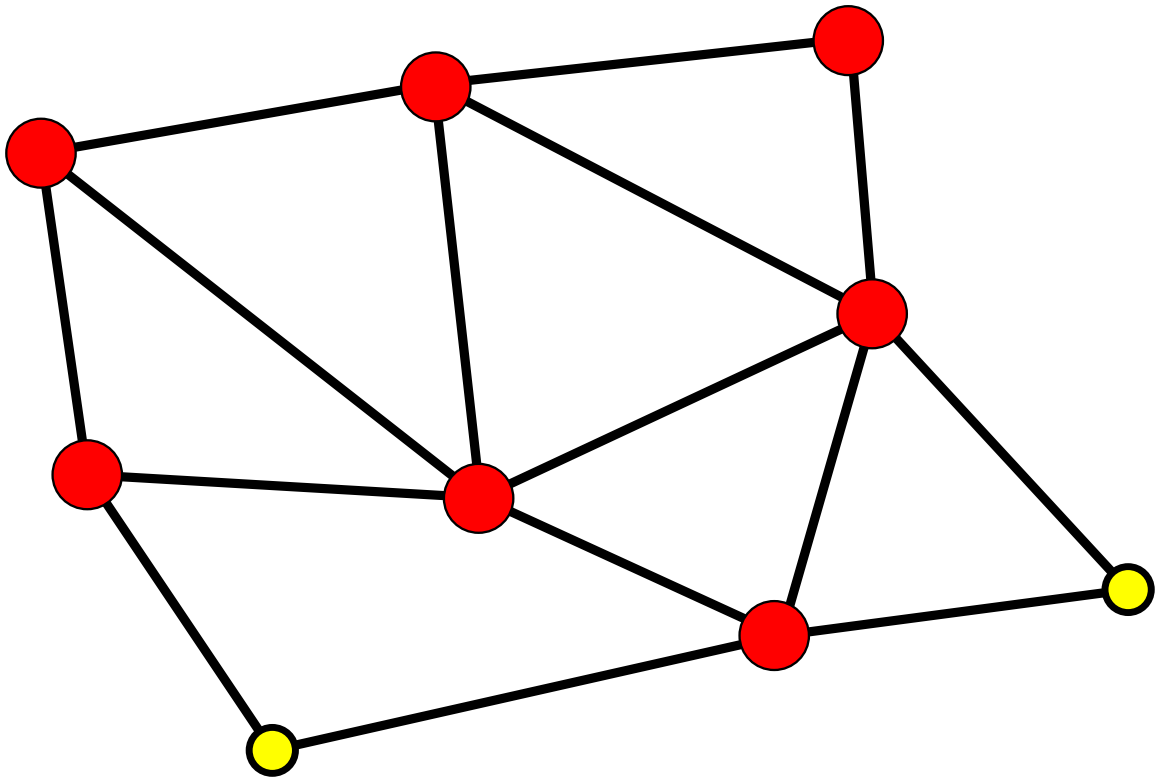
Virus spread



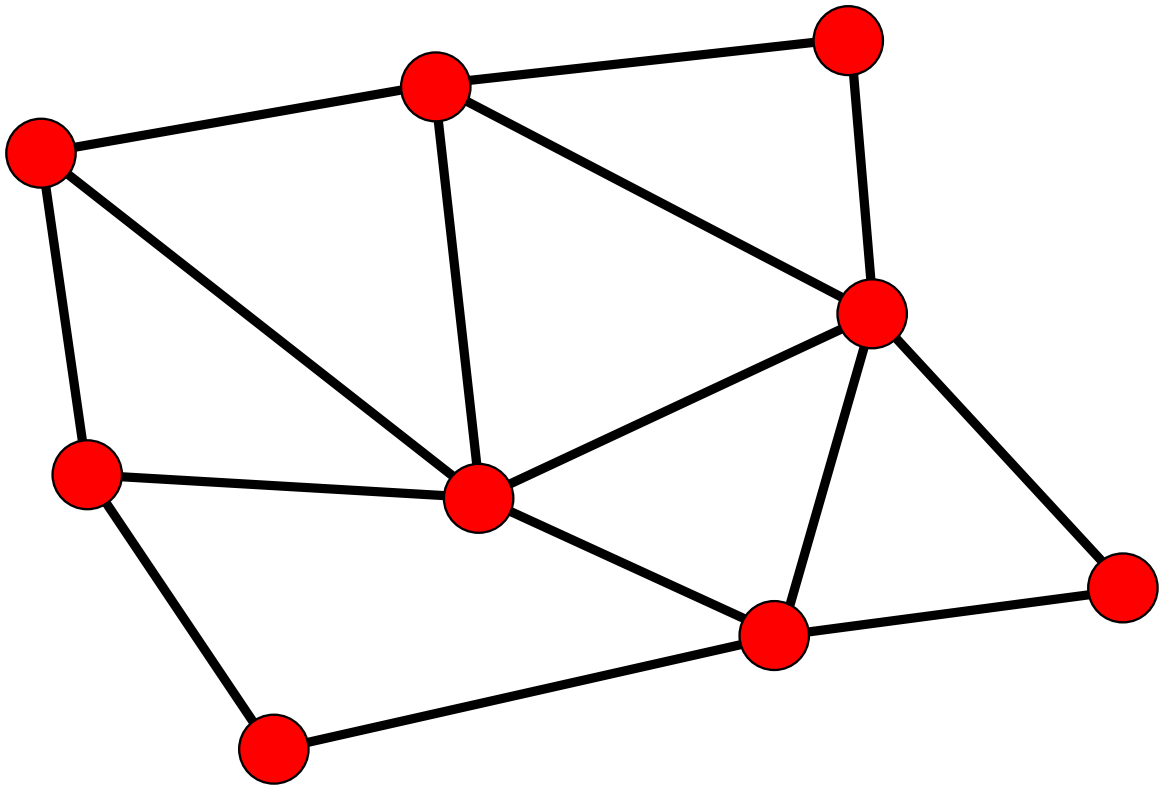
Virus spread



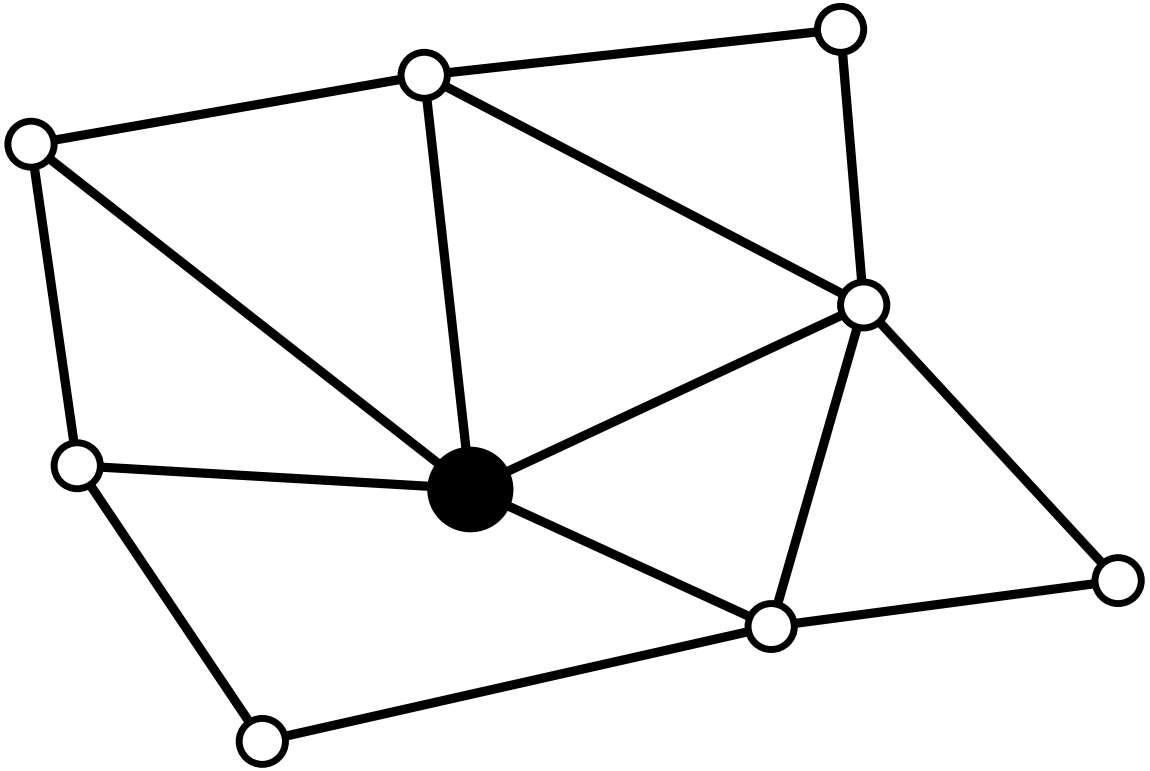
Virus spread



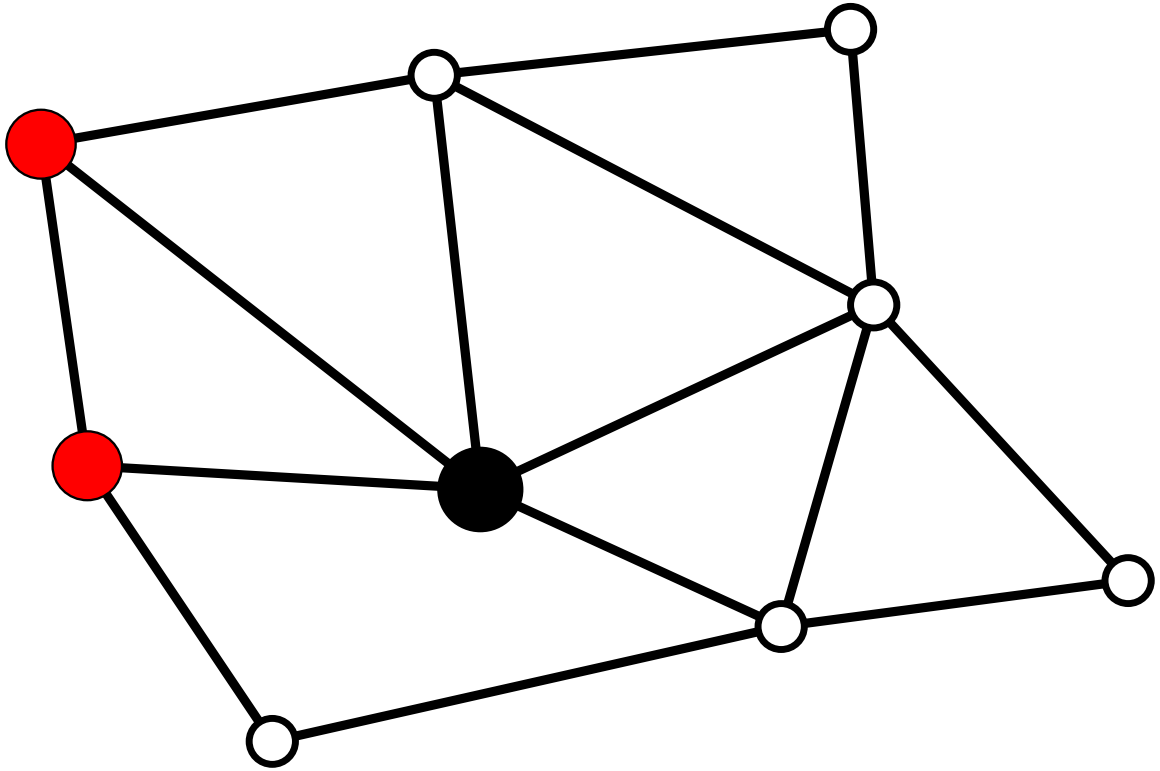
Virus spread



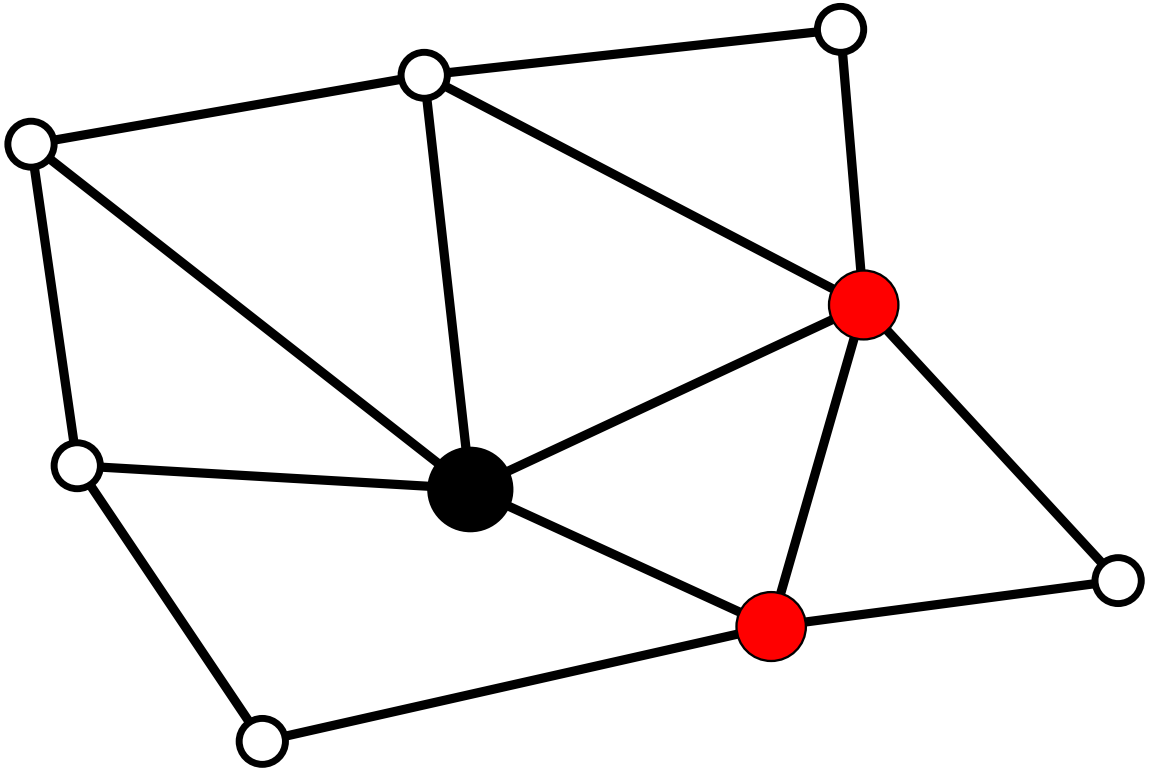
Restrain the spread



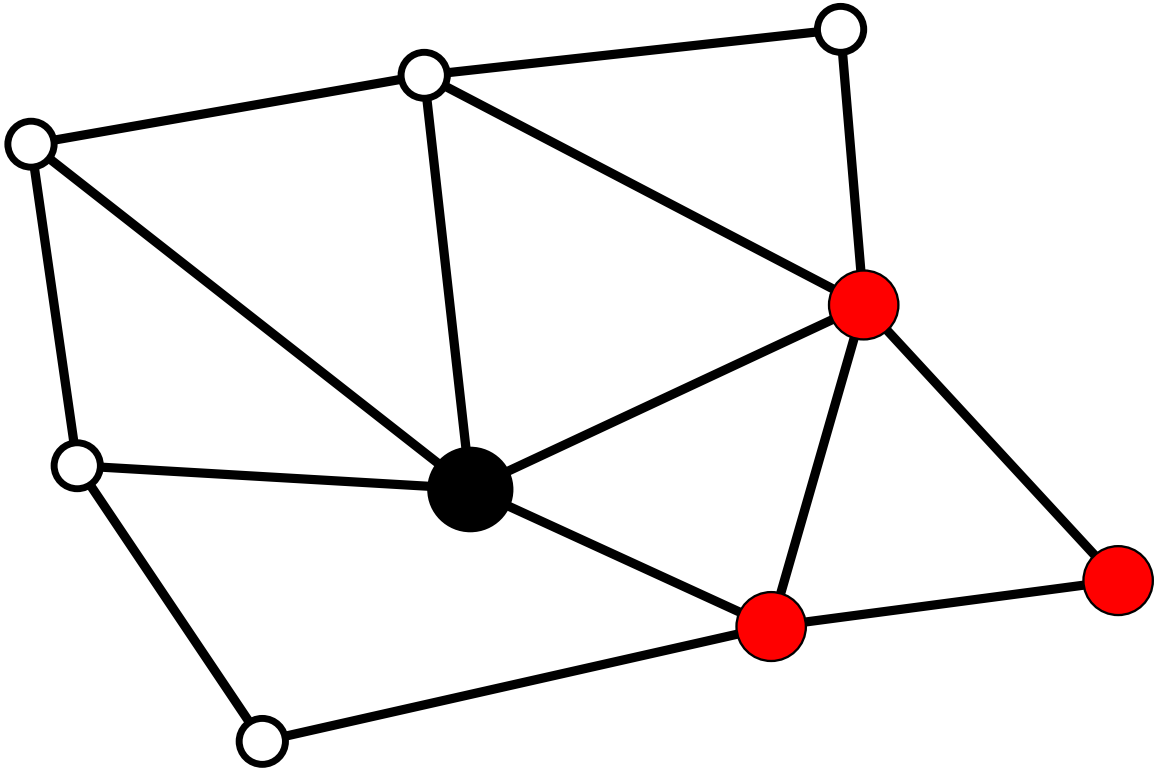
Restrain the spread



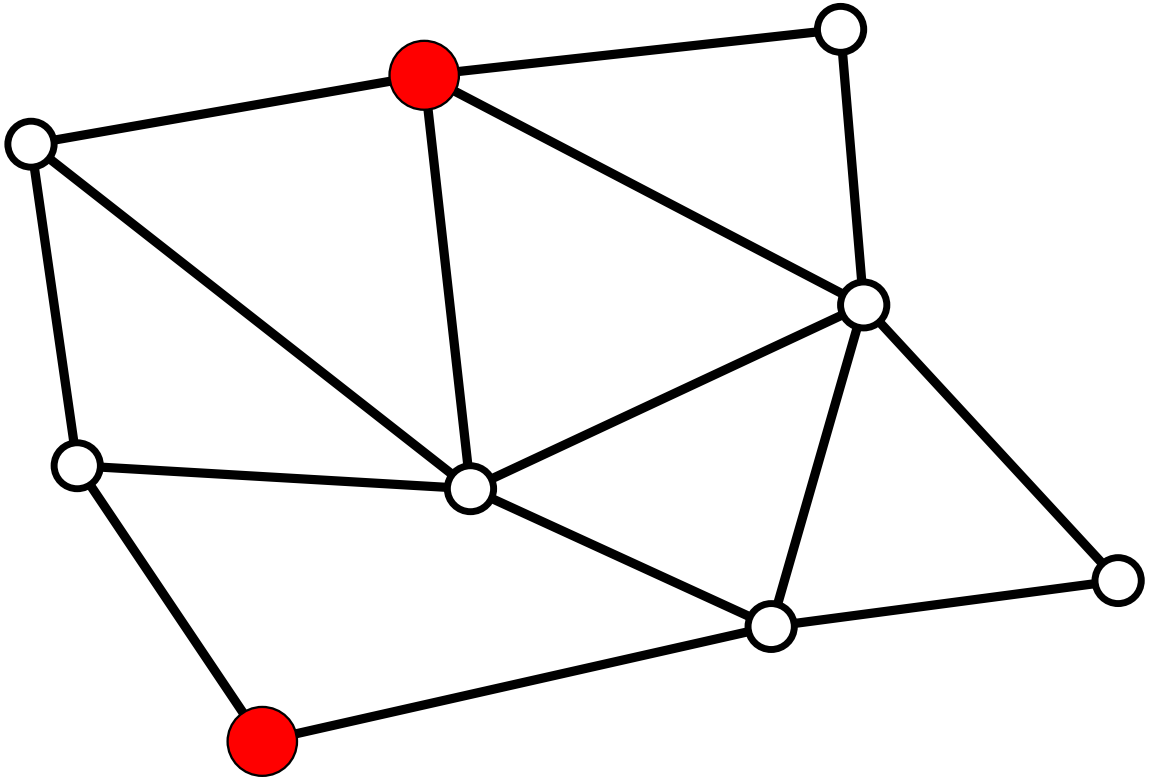
Restrain the spread



Restrain the spread



Naive virus injection



General framework

- network $G = (V, E)$ over which the virus propagates
- virus-propagation model (can be probabilistic)
- adversary who injects copies of the virus in the network
  - blind
  - adaptive

$\Rightarrow$ immunization algorithm:

given a network, budget k, and a virus-propagation model

find k nodes to immunize so that the spread is minimized

What is the spread?

- network $G = (V, E)$
- adversary plants r viruses (blindly or adaptively)
- $N_r \subseteq V$: set of nodes selected by adversary
- expected number of infected nodes: $S(N_r, G)$
- spread: $S_r(G) = \max_{N_r} S(N_r, G)$
- expected spread: $\widehat{S}_r(G) = E_{N_r}[S(N_r, G)]$

Example of immunization algorithms

- immunize a random node
- immunize the node with the largest degree

Virus-propagation models

- problem as stated above is too general
  - e.g., no formal specification language for all possible virus-propagation models
- concentrate on two specific virus-propagation models:
  - independent cascade, and
  - dynamic propagation,

  ...but similar ideas can be applied to other models, too

Some background models on epidemics

- Susceptible-Infected-Removed (SIR)
  - susceptible (healthy) nodes do not have the virus but they can catch it if exposed to somebody who does
  - infected nodes have the virus and they can pass it
  - removed (or recovered) have immunity, cannot catch the virus again and cannot pass it on
- Susceptible-Infected-Susceptible (SIS)
  - susceptible nodes
  - infected nodes can be healed and become susceptible again

Epidemics background

- traditional studies do not take into account the network structure
  - nodes become infected or recovered with uniform probabilities
- modern studies do take into account network topology
- epidemic threshold
  - β: infection rate, δ: healing rate, $\lambda = \beta/\delta$: effective spreading rate
  - $\exists \lambda_c$ s.t. if
  - $\lambda \geq \lambda_c$ a non-zero fraction of nodes becomes infected (SIR)
  - $\lambda \geq \lambda_c$ virus spreads and becomes persistent (SIS)
  - $\lambda < \lambda_c$ virus dies out exponentially fast (SIS)

Epidemics background

- many studies of special cases
- power-law networks do not have (non-zero) epidemic thresholds
- studies of immunizing the highest degree nodes
- immunization in the case of unknown network topology
  - immunizing the adjacent node of a random node works well for skewed-degree networks
- ...

Our approach

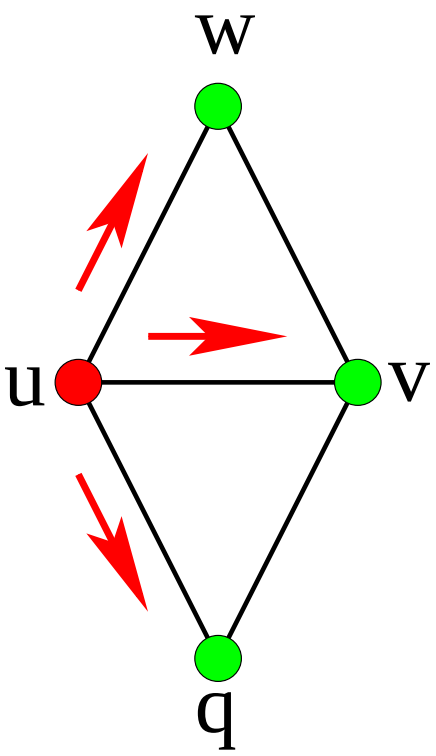
- algorithmic approach to the immunization problem
- extensive experimentation

- virus-propagation models considered:
  - independent cascade, and
  - dynamic propagation

Independent cascade

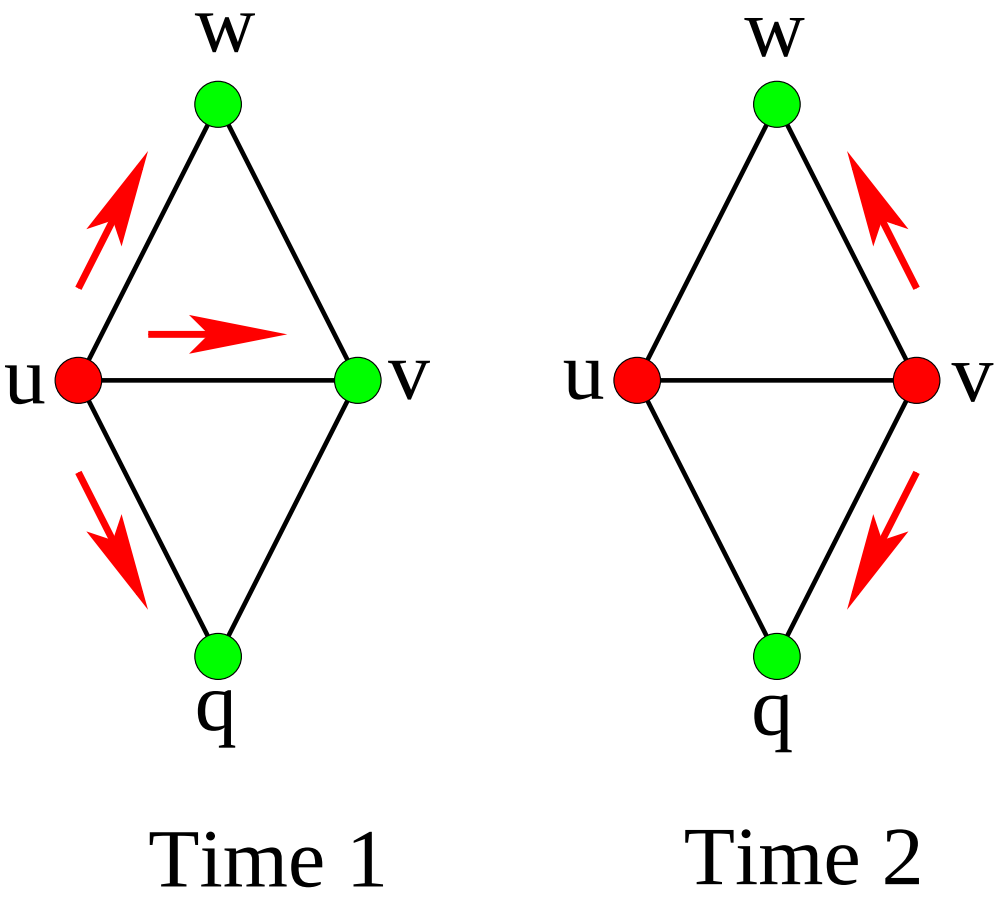
- initially the adversary plants r viruses in the network
- assume node u becomes infected for first time at time t:
  - u attempts to infect all currently uninfected neighbors v
  - it succeeds with probability p
  - if u succeeds then v becomes infected
  - otherwise u never attempts to infect v again

Independent cascade — example



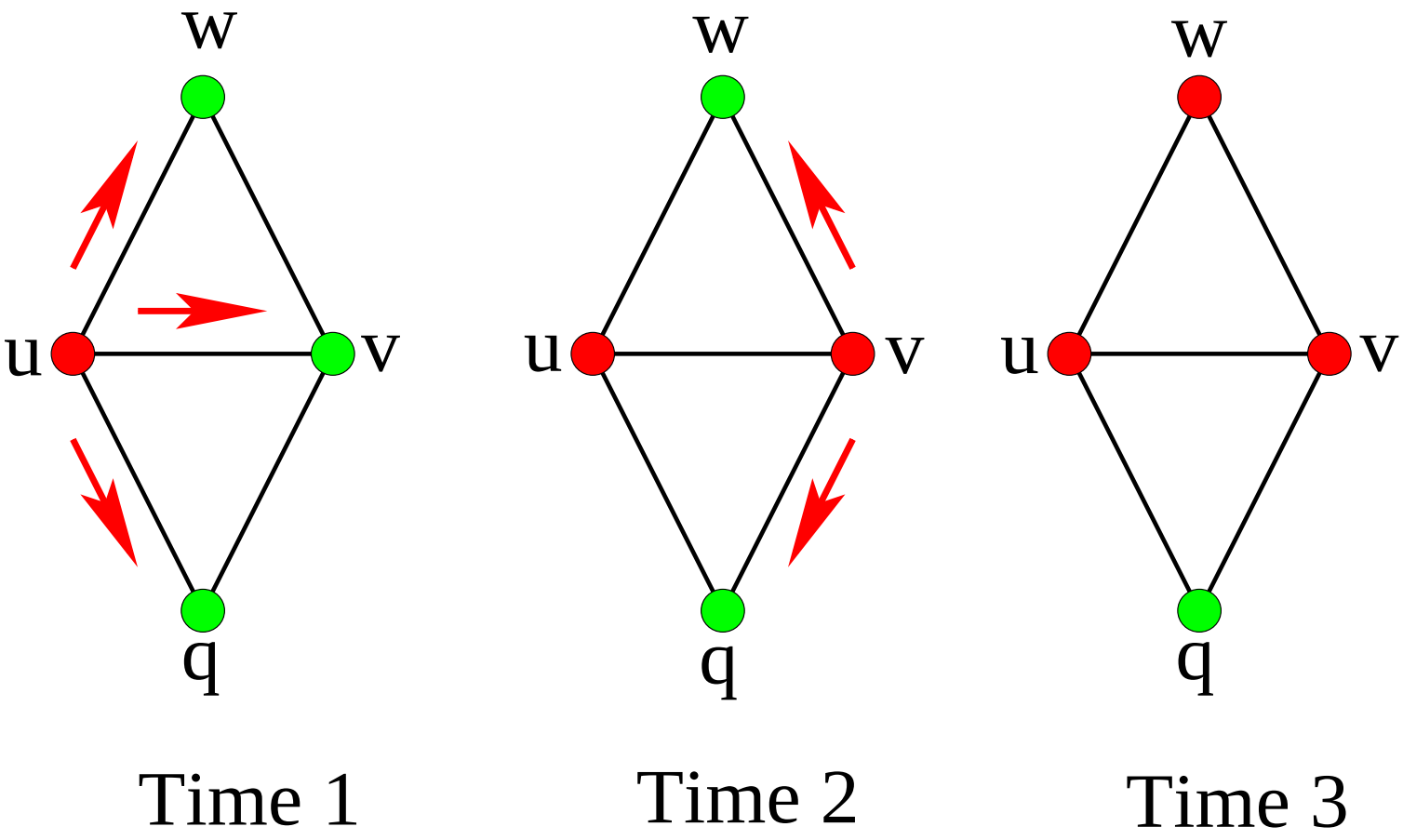
Time 1

Independent cascade — example

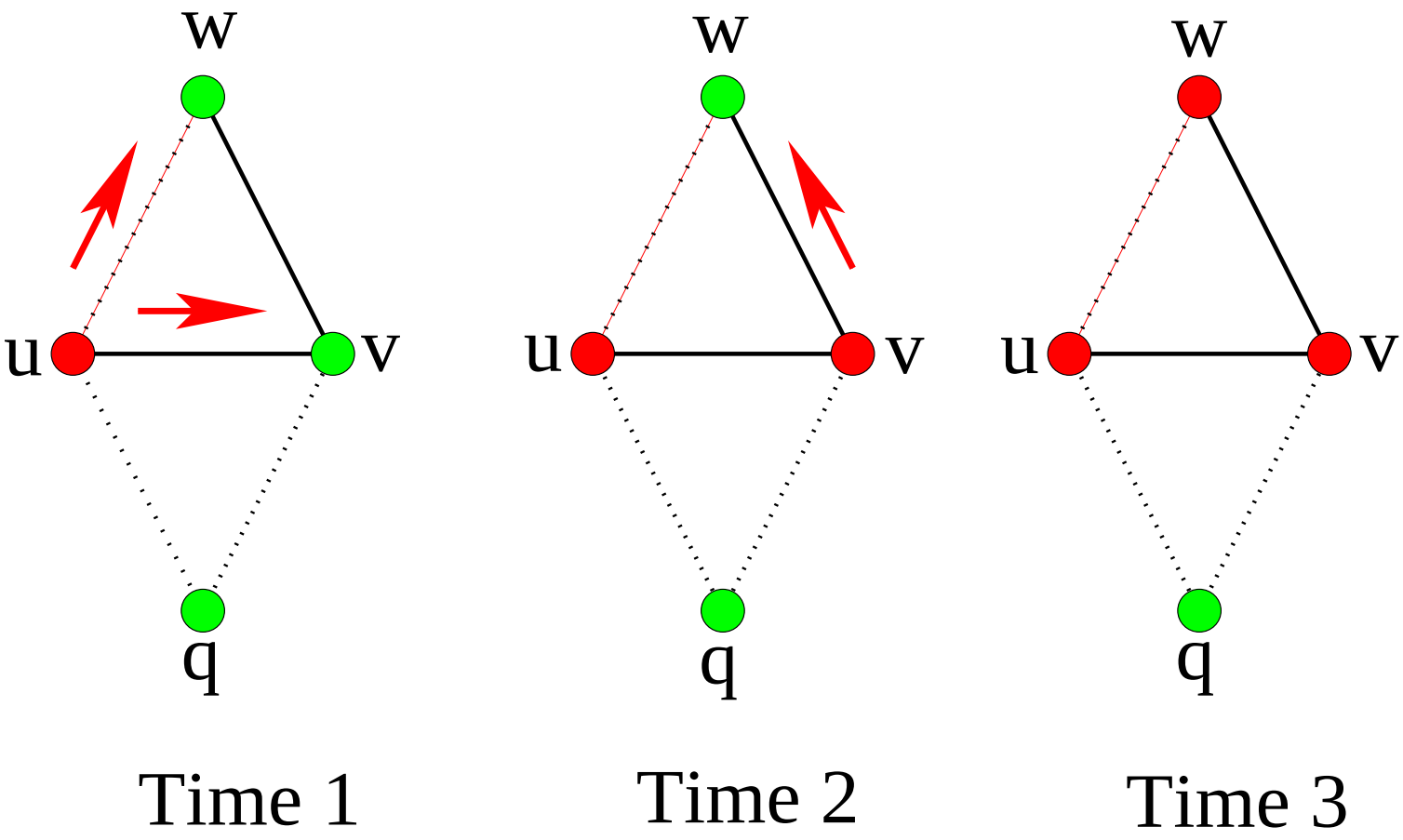


Time 1 Time 2

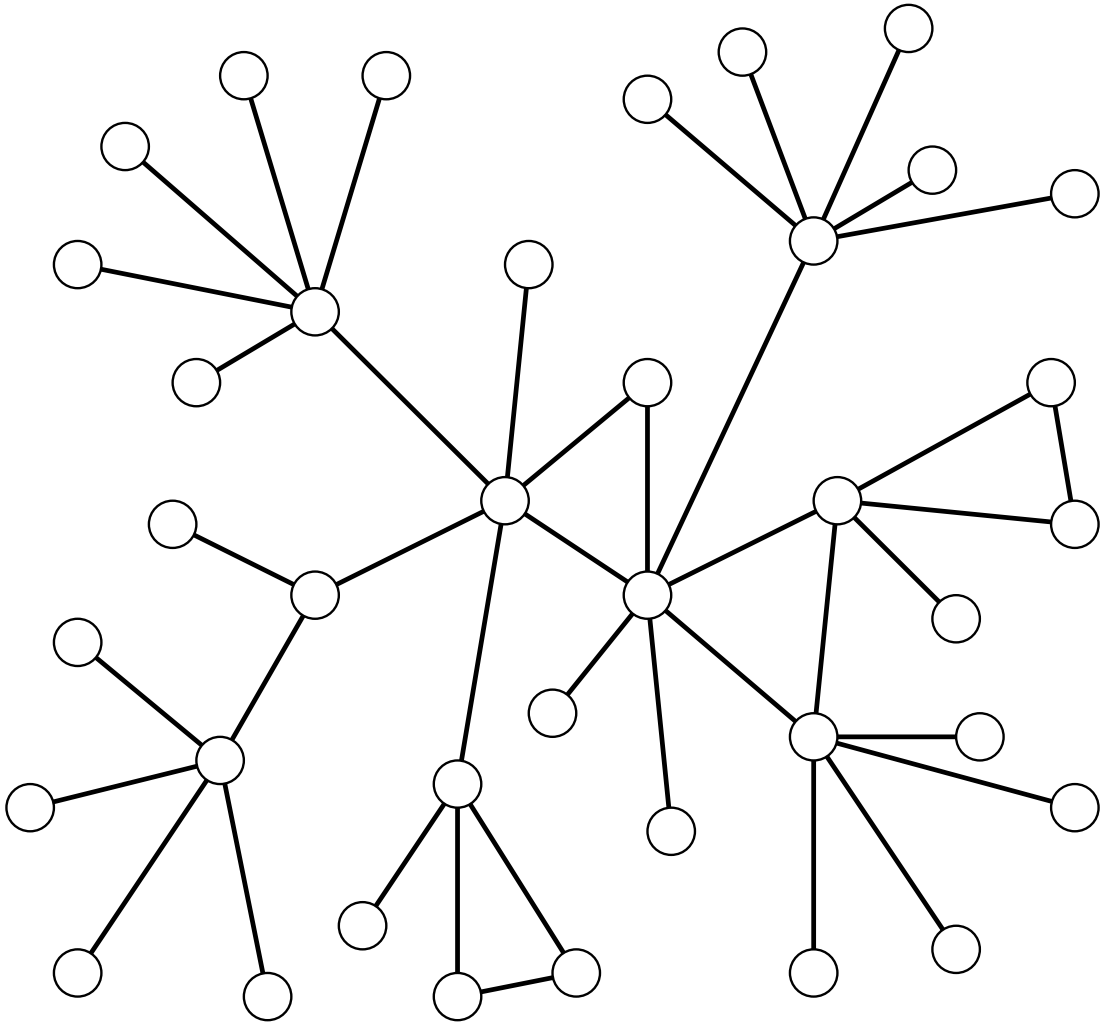
Independent cascade — example



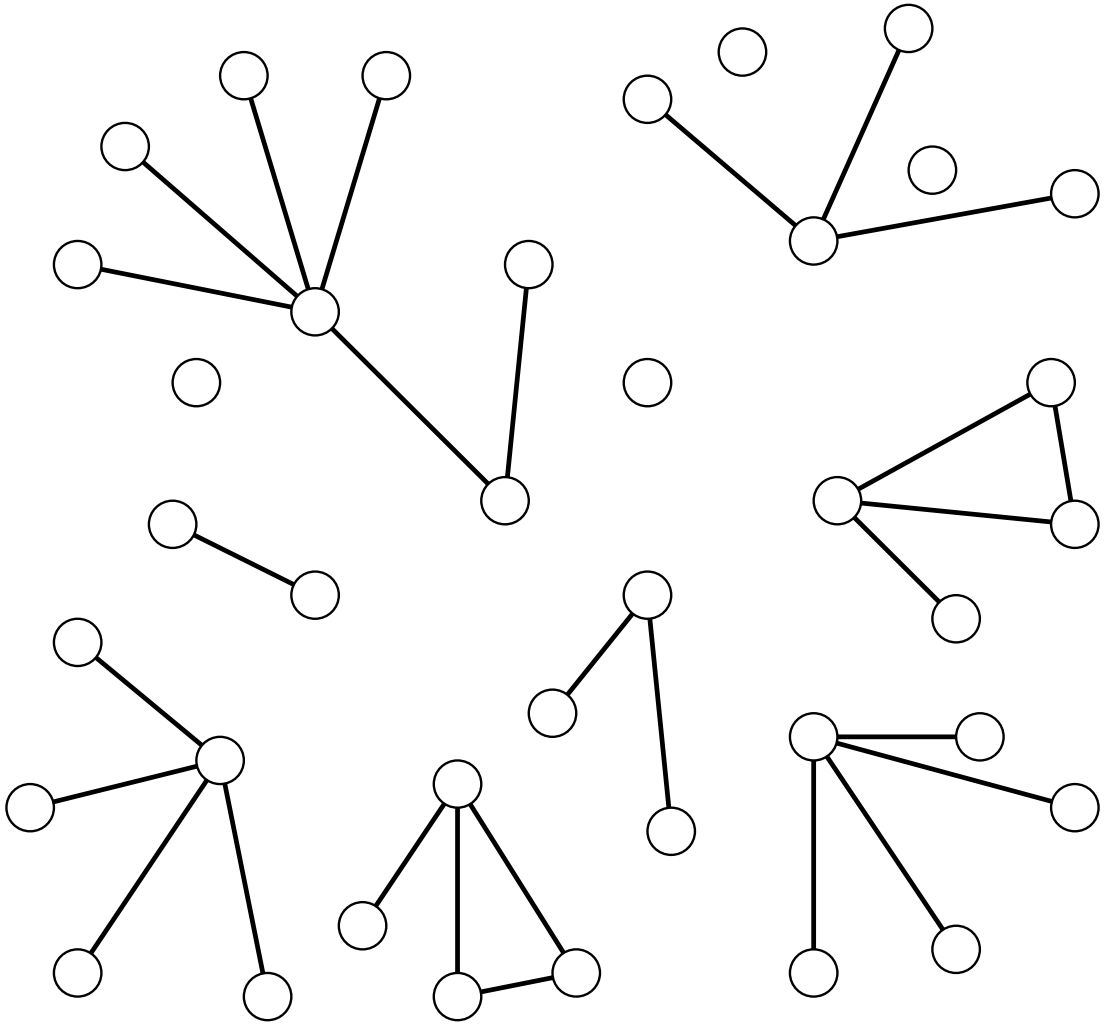
Independent cascade — example



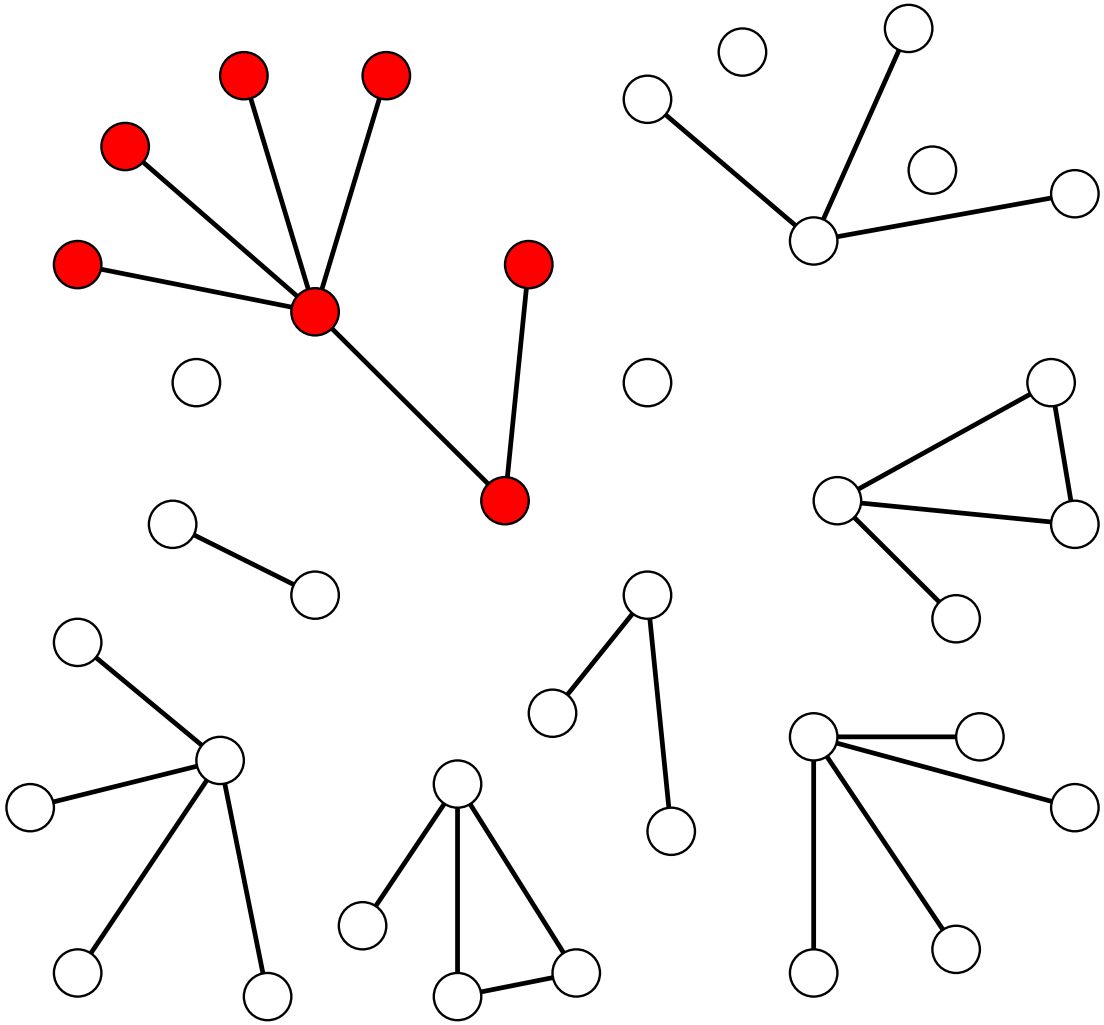
Independent cascade



Independent cascade



Independent cascade



Independent cascade

- given a sampling on network links with probability p
  - $S_1(G)$ is size of largest connected component (adaptive)
  - $\widehat{S}_1(G)$ is the average connected components size (blind)
- immunization problem:
  - remove k nodes from the network in order to minimize
  - size of r largest connected components, or
  - average size of connected component, respectively
- both $S_r(G)$ and $\widehat{S}_r(G)$ are NP-hard

Algorithm for the independent-cascade model

- greedy, i.e., immunize nodes one by one
- for the adaptive-adversary case:
  - at each iteration find the node that minimizes the expected size of the largest connected component in the resulting network
- for the blind-adversary case:
  - at each iteration find the node that minimizes the expected size of the average connected component in the resulting network

Computing the expectations

- sample many graphs over all the $2^{|E|}$ possible graphs
  - in each sample graph (u, v) exists with probability p

$\Rightarrow$ in each sampled graph

for each node u

find the size of the largest/average connected component in the graph resulting from removing (immunizing) u

select the node that minimizes the expectation (largest/average)

Dynamic-propagation

- a dynamic birth-death process that evolves over time
- virus propagates from node u to neighbor node v with probability β
- at each point in time, a node u that is infected heals with probability δ

Epidemic-threshold property

- Theorem. Consider network G with adjacency matrix M, propagation probability β, and healing probability δ.
  If $\beta/\delta < 1/\lambda_1(M)$ the expected time until the virus dies out is logarithmic in the number of nodes in the network, against an adaptive adversary

Epidemic threshold (cont.)

- what if β/δ large?
- notice that the virus eventually will die out
- dynamical model hard to analyze because of non linearities
- recent work by Ganesh et al. 2005 shows that if $\beta/\delta > 1/\eta(G)$ (isoparametric constant of the graph) then the expected time until the virus dies out is exponential with the size of the network

Multiple-copies model

- each node can have multiple copies of a virus
- infection probability refers to receiving one more copy
- healing probability refers to removing one copy
- more pessimistic than the single-copy model
- easier to analyze

Multiple-copies model

- at time t, node i has v_i^t copies
- $\mathbf{v}^t = [v_1^t, \ldots, v_n^t]$ vector of nodes' copies
- $\widehat{\mathbf{v}}^t$ expected value of $\mathbf{v}^t$
- then

$$\widehat{\mathbf{v}}^{t+1} = \Delta \widehat{\mathbf{v}}^t, \text{ where } \Delta = \beta M + \mathsf{diag}(1-\delta, \ldots, 1-\delta)$$

- Theorem. In the multiple-copies model the expected time until the virus dies out is logarithmic if $\beta/\delta < 1/\lambda_1(M)$ and it is unbounded if $\beta/\delta > 1/\lambda_1(M)$

Immunization problem for the dynamic model

- given network G and effective infection rate β/δ, immunize the minimum number of nodes in G, such that $\beta/\delta < 1/\lambda_1(M')$, where M' is the adjacency matrix of the immunized network
- we would like to use a greedy approach
- the problem becomes finding the node to immunize so that the eigenvalue of the adjacency matrix drops as much as possible

EIG algorithm for dynamic propagation

- $B \leftarrow M$
- while $\beta/\delta > 1/\lambda_1(B)$
  - compute w_1, the eigenvector of B that corresponds to $\lambda_1(B)$
  - find node u with the maximum value in w_1
  - Remove u from the graph and form new matrix B'
  - $B \leftarrow B'$

Intuition behind the EIG algorithm

- suppose that "susceptibility" of node i is captured by w_i
- probability of virus propagation between i and j: $p_{ij} = w_i w_j$
- healing probability of node i is $1 - w_i^2$
- system matrix $\Delta = \boldsymbol{w}\boldsymbol{w}^T$
- then $\lambda_1(\Delta) = ||\boldsymbol{w}||^2$ and corresponding eigenvector $\boldsymbol{w}$ (norm.)
- consider Δ' after immunizing node i
  (zero-ing the i-th row and column of Δ)
- now $\lambda_1(\Delta') = ||\boldsymbol{w}||^2 - w_i^2$

Intuition behind the EIG algorithm

- the principal eigenvalue gives an indication of the connectivity of the network
- large eigenvalue corresponds to a densely connected network
- the nodes with the maximum value in the first eigenvector are the ones that are most tightly interconnected
- removing these nodes reduces the graph connectivity
- in general EIG selects nodes with high degree, but not always (more global view)

Experimental setup – algorithms

- compare the performance of the algorithms against other strategies
  - MaxDegree
  - MaxDegreeIt
  - Random

Experimental setup – datasets

- synthetic datasets:
  - random graphs (Erdős-Rényi)
  - scale-free graphs (Barabási and Albert)
  - small-world graph (Watts, Watts and Strogatz)
- real datasets:
  - co-authorship graphs (representing social networks)
  - autonomous systems (internet topology)
  - power-grid (networks of electricity transfer)

Scale-free graphs (Barabási and Albert)

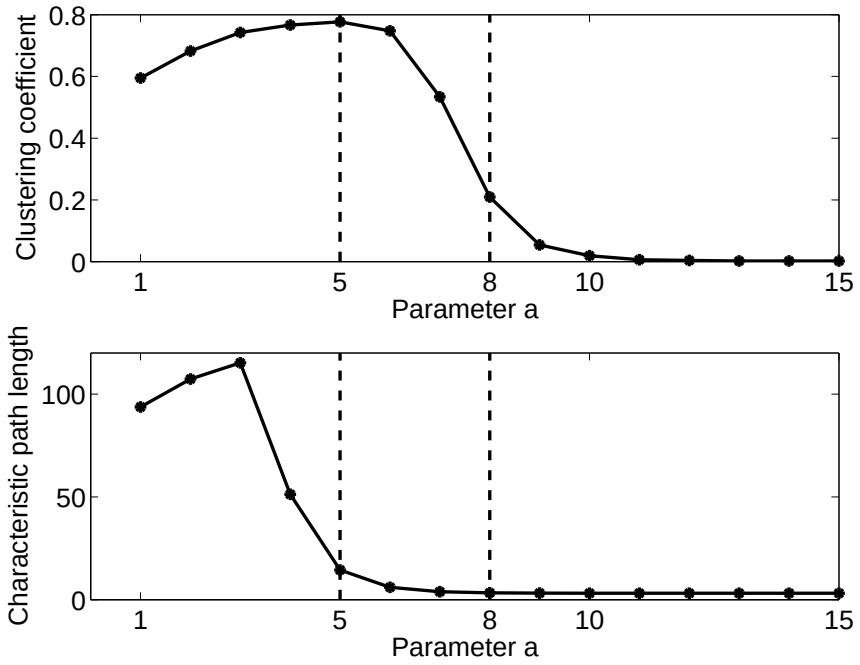
- preferential attachment
- nodes join the network sequentially
- each new node comes with m edges
- it connects its m edges to existing nodes, which are selected with probability proportional to their degrees
- simulates the rich gets richer effect
- results in power-law graphs with exponent 3

Small-world graphs

- Networks with
  high clustering coefficient and
  small average path length

Small-world graphs – Watts model

- generated using a parameter α
- intuitively α controls the probability that two nodes will be connected given the number of their common neighbors

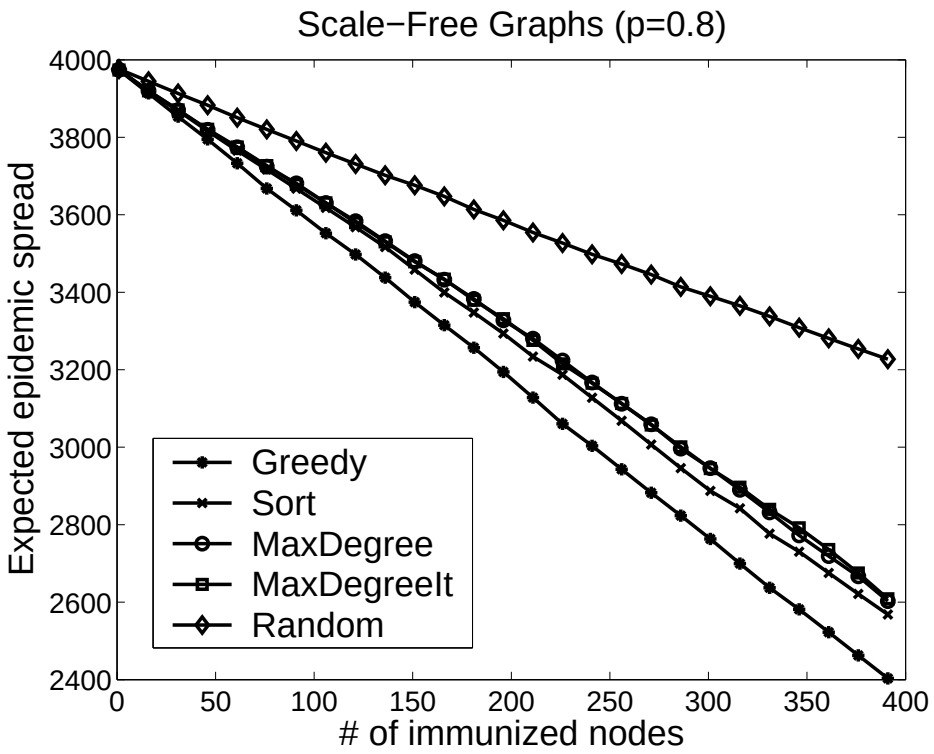


Small-world graphs – Watts-Strogatz model

- the generation process is governed by parameters q
- initially all nodes are on a ring lattice.
- each node has degree k
- each node is rewired to another random node with probability q

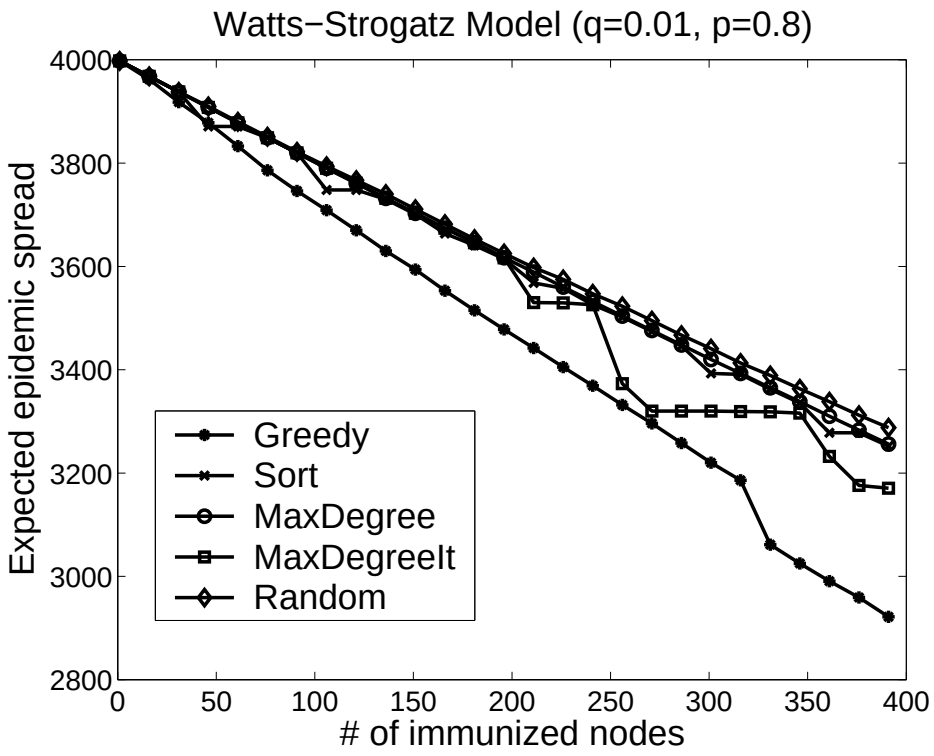
Independent cascade

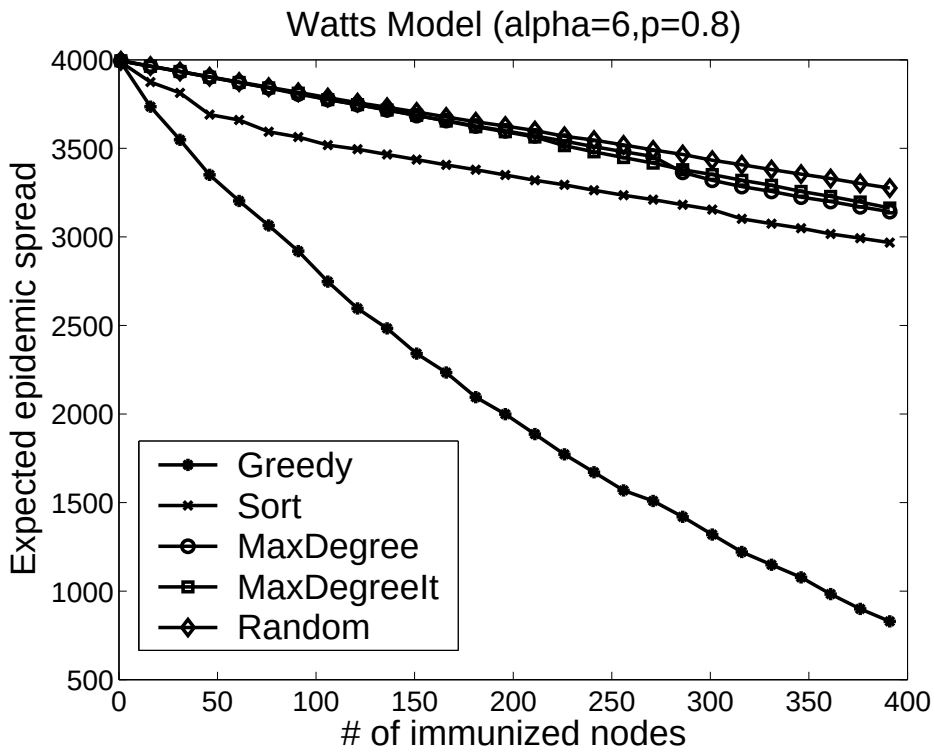
synthetic dataset – scale-free graphs



Independent cascade

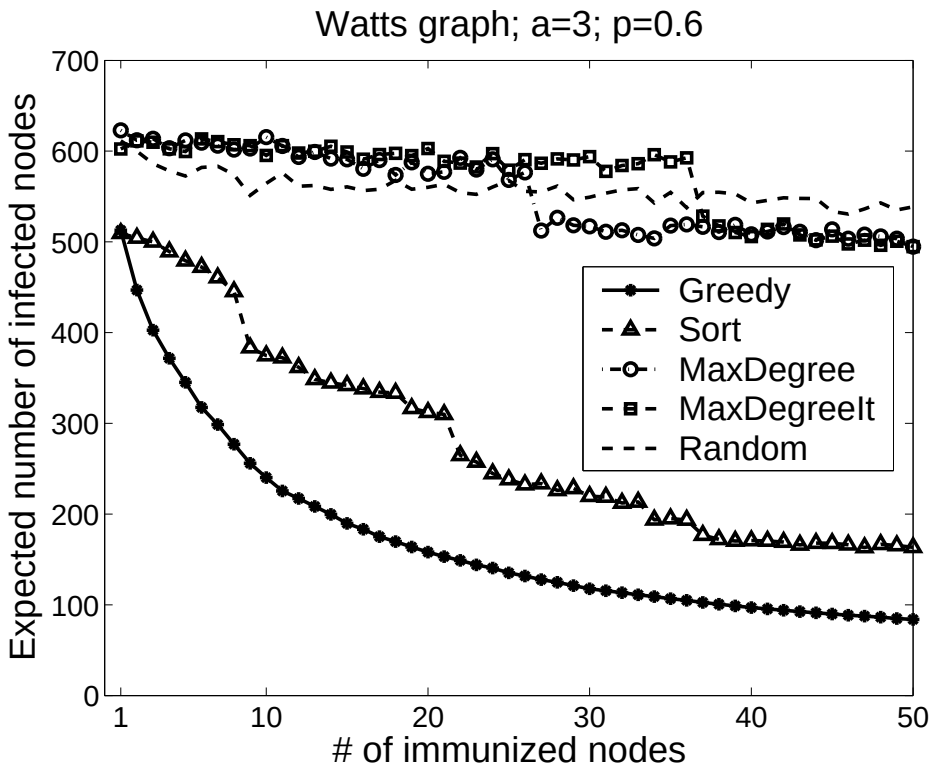
synthetic dataset – small-world graphs

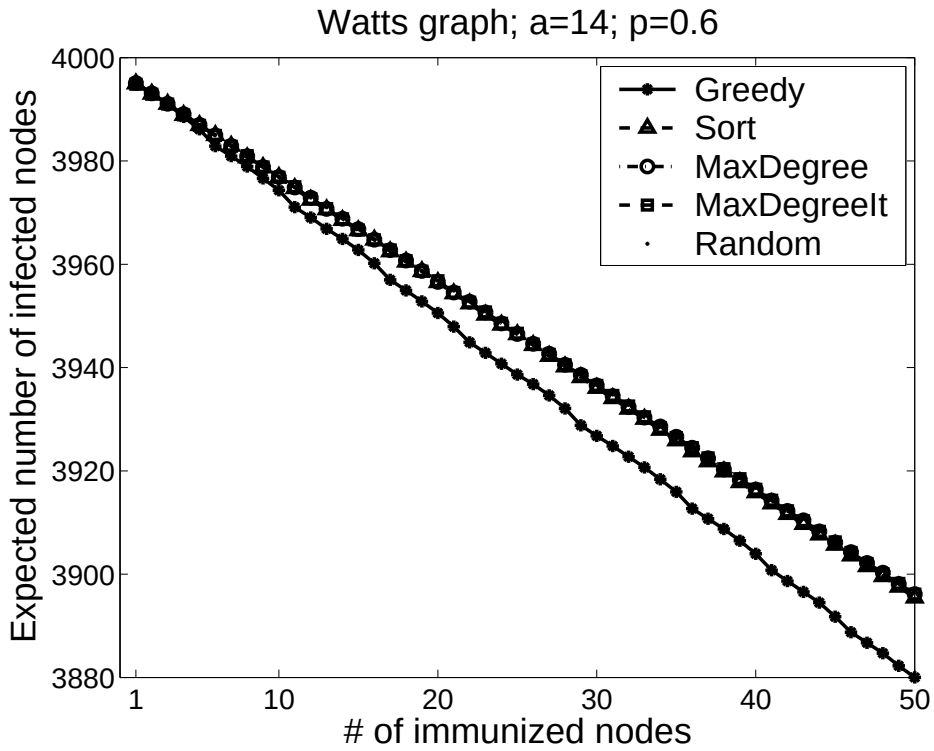




Independent cascade

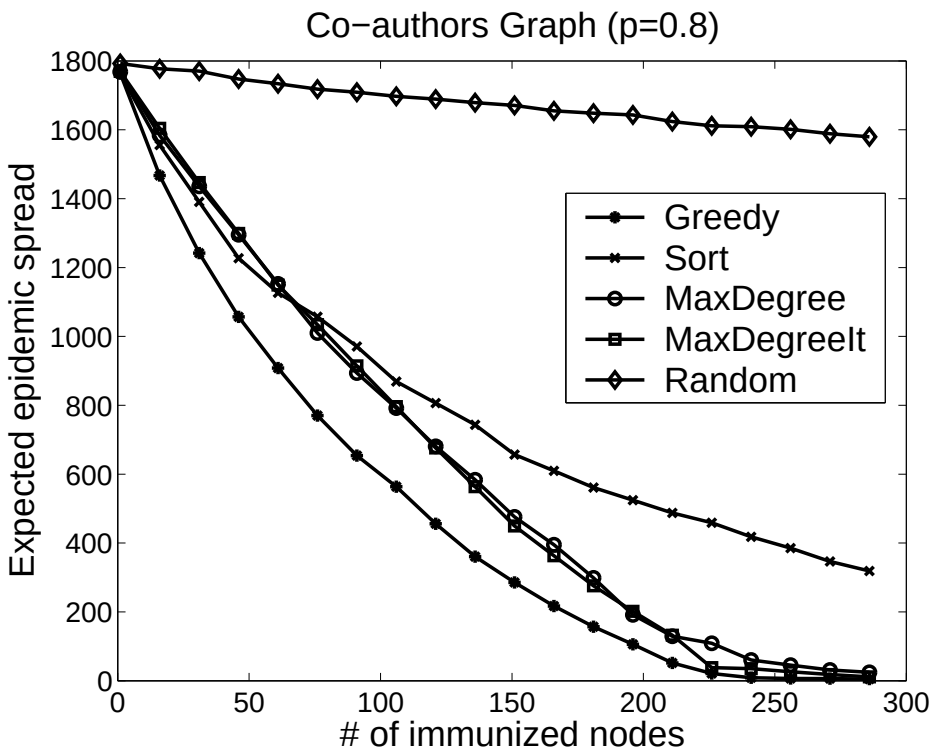
synthetic datasets – small-world graphs

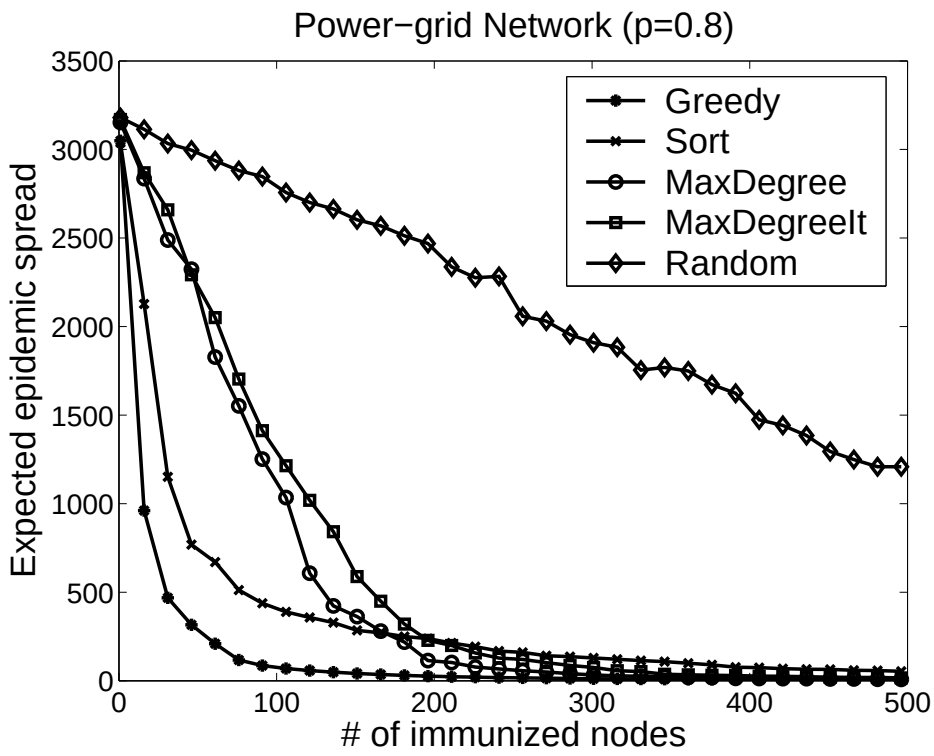




Independent cascade

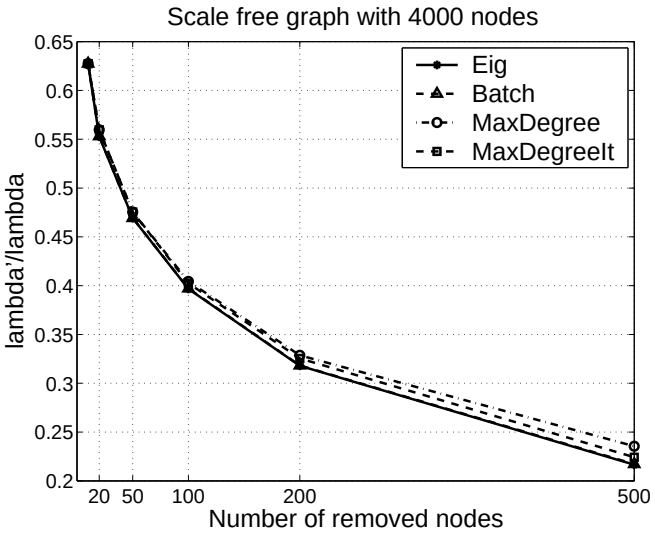
real datasets

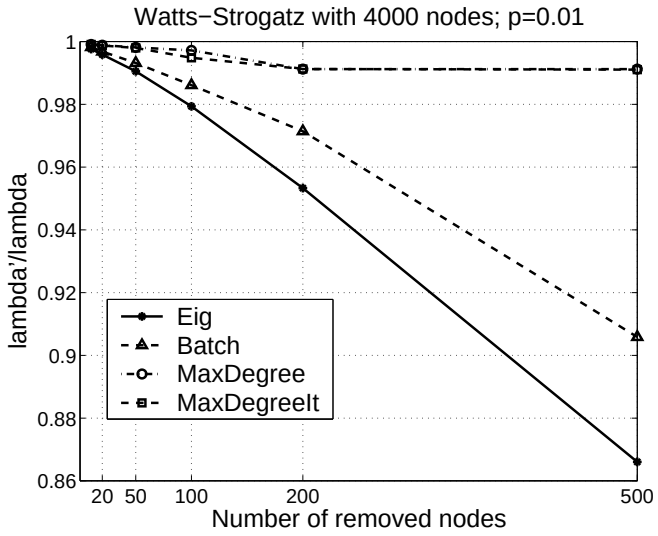


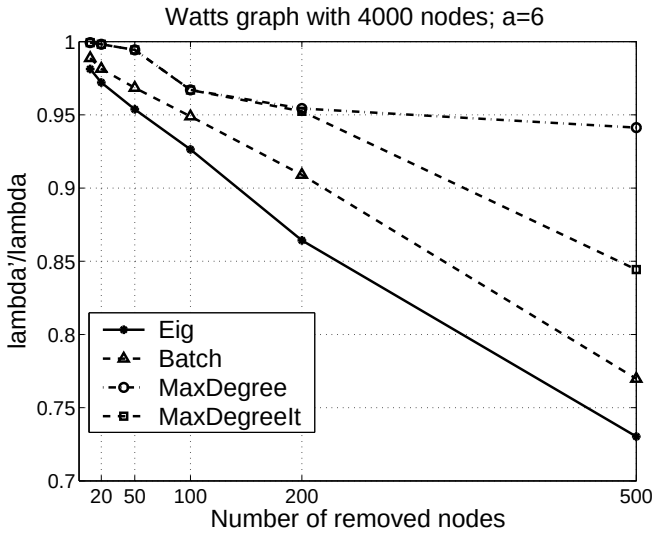


Dynamic propagation

synthetic datasets

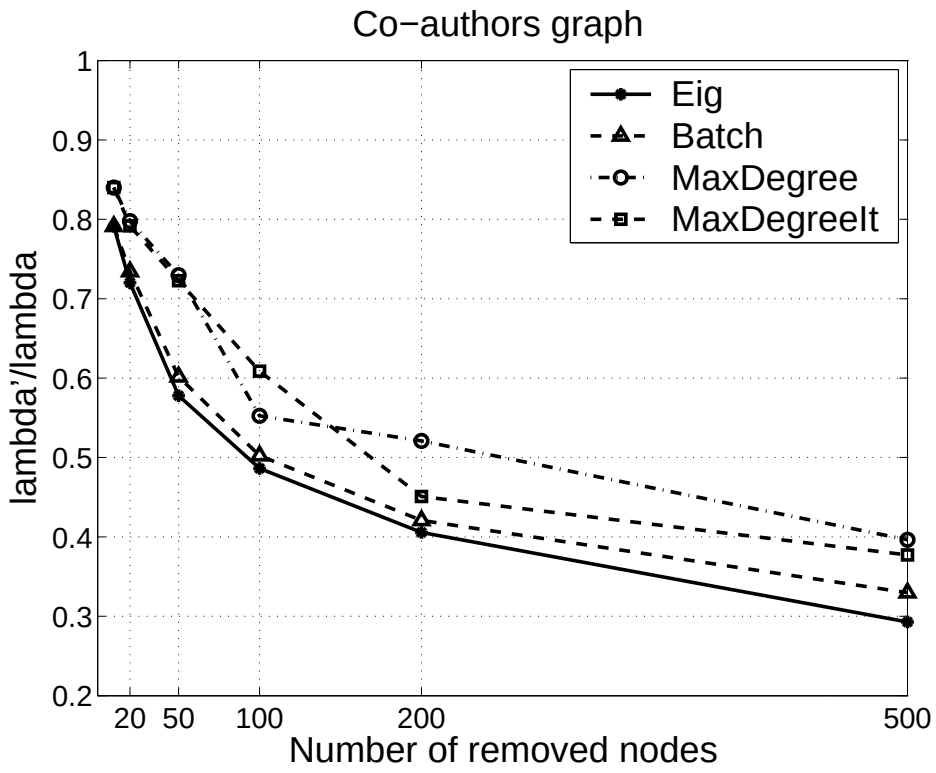


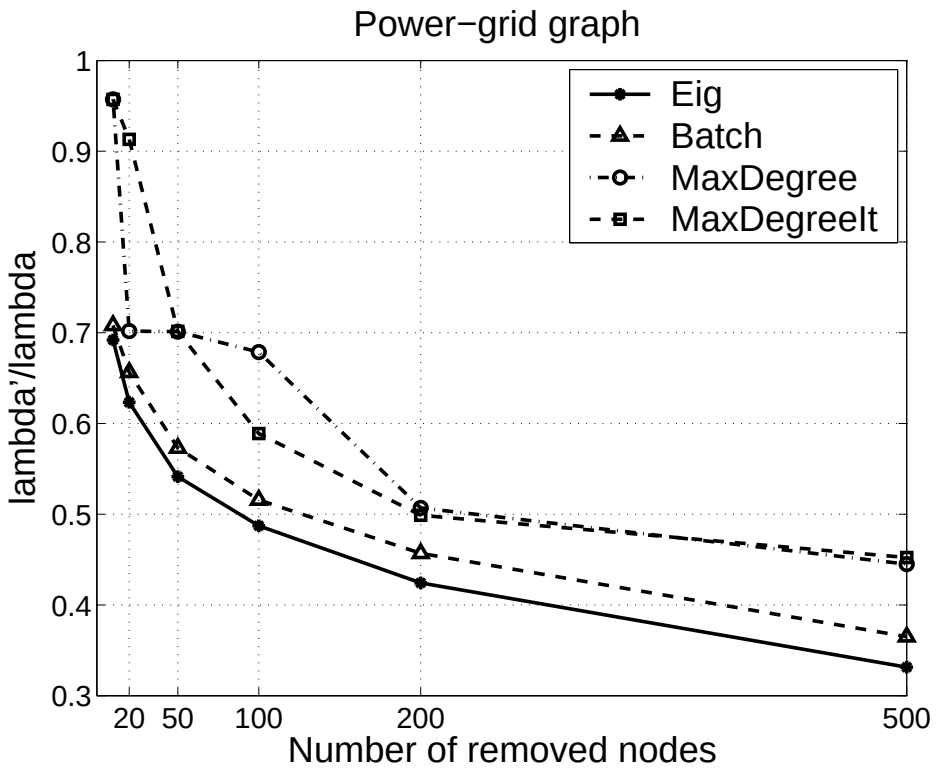




Dynamic propagation

real datasets





Conclusions

- network immunization problem under different virus propagation models
- greedy algorithms work well in practice
- applications in epidemiology and security of computer networks
- many open problems
  - can we do better than the greedy?
  - which node to remove in order to obtain the largest drop in the eigenvalue?

...complete change of topic...

Genome segmentations

joint work with Niina Haiminen, Evimaria Terzi, Heikki Mannila

(k,h)-segmentation

- [Gionis and Mannila 03]
- given sequence $S = a_1, a_2, \ldots, a_n$
- we want to find k segments
- but only $h < k$ different *segment types* are allowed
- each of the k segments should be assigned to one of the h types
- find the best segmentation into k segments, the h types, and the assignment of each segment to one type

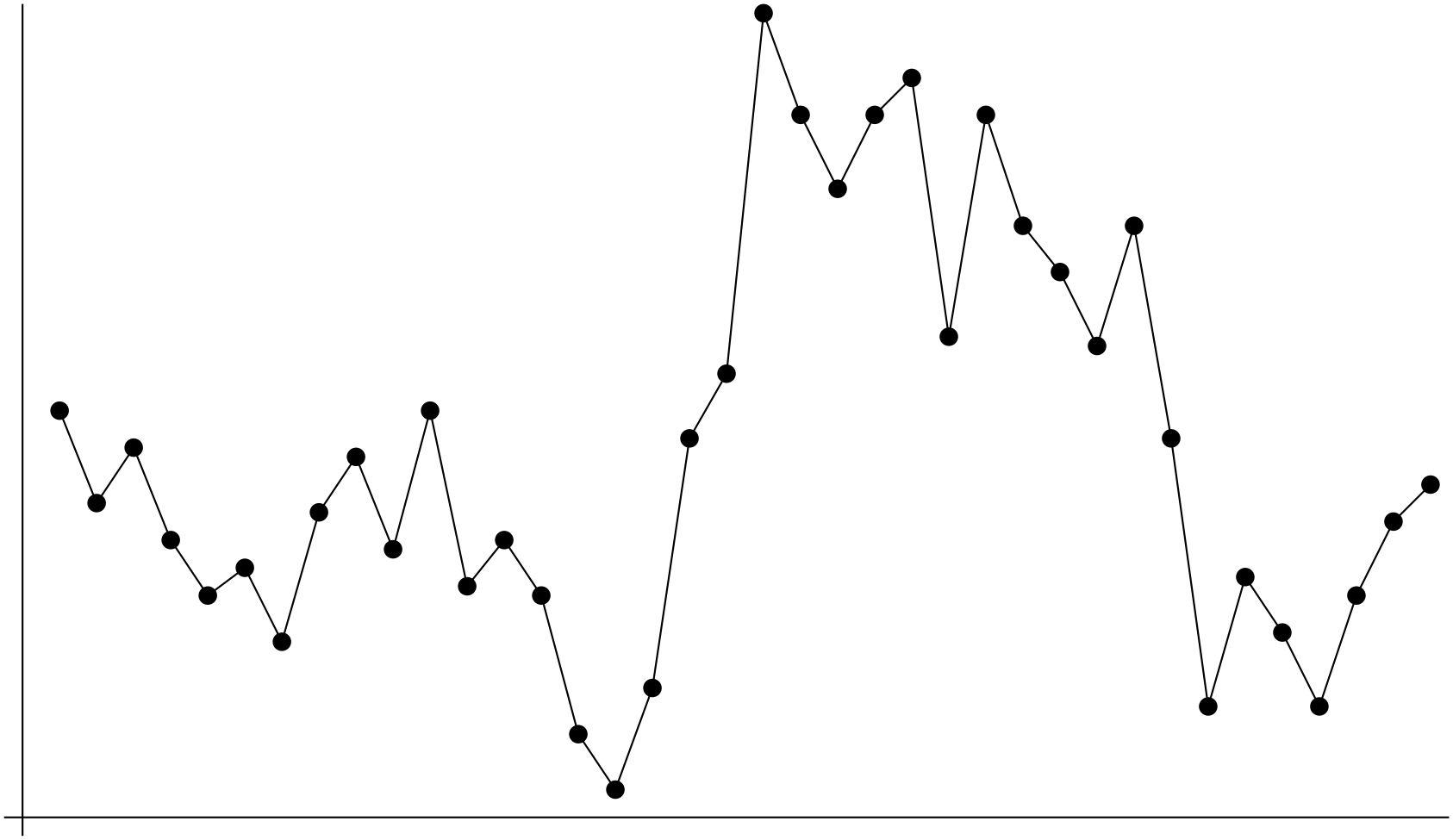
(k,h)-segmentation: problem definition

- assume piecewise constant representation, and L_2^2
- given sequence $S = a_1, a_2, \ldots, a_n$
- we want to find
  - partition of S into k segments $S_1, \ldots, S_k$,
  - h levels $l_1, \ldots, l_h$
  - assignment of segment j to level $l_j \in \{l_1, \ldots, l_h\}$

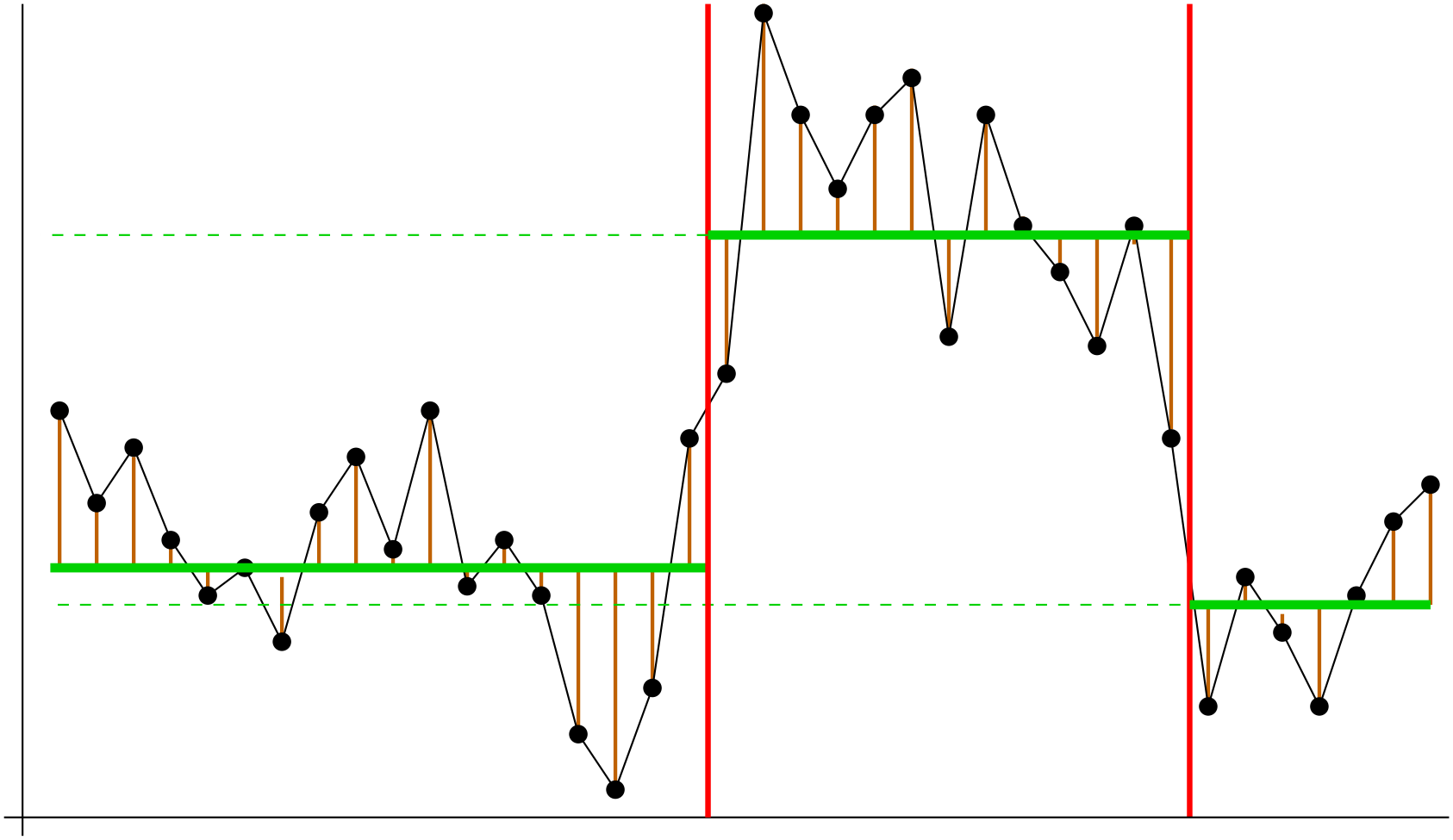
  in order to minimize the total error

$$R[n,k,h] = \sum_{j=1}^{k} \sum_{i=b_j}^{e_j} (a_i - l_j)^2$$

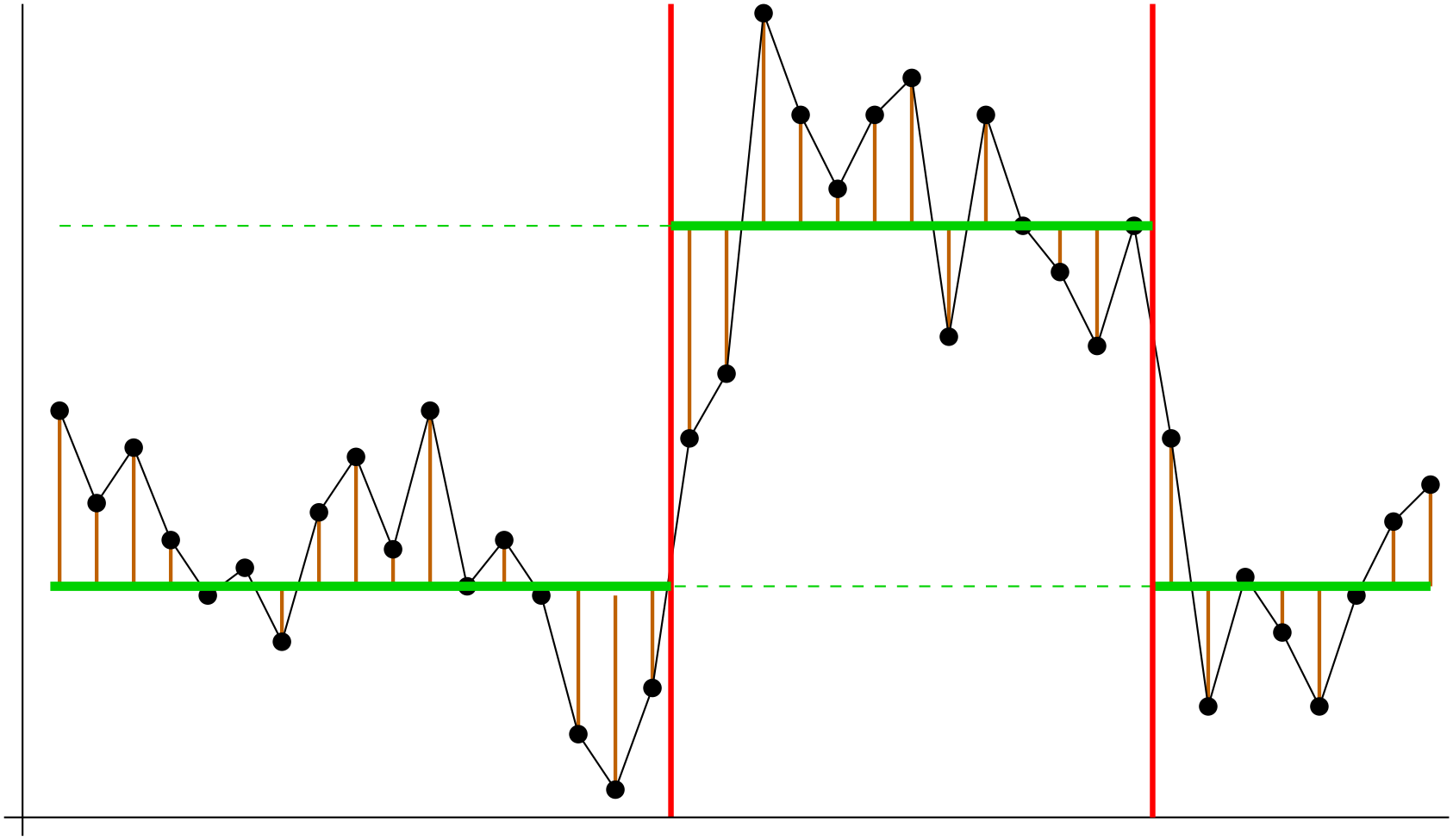
Example



Example: $k = 3$ and $h = 3$



Example: $k = 3$ and $h = 2$



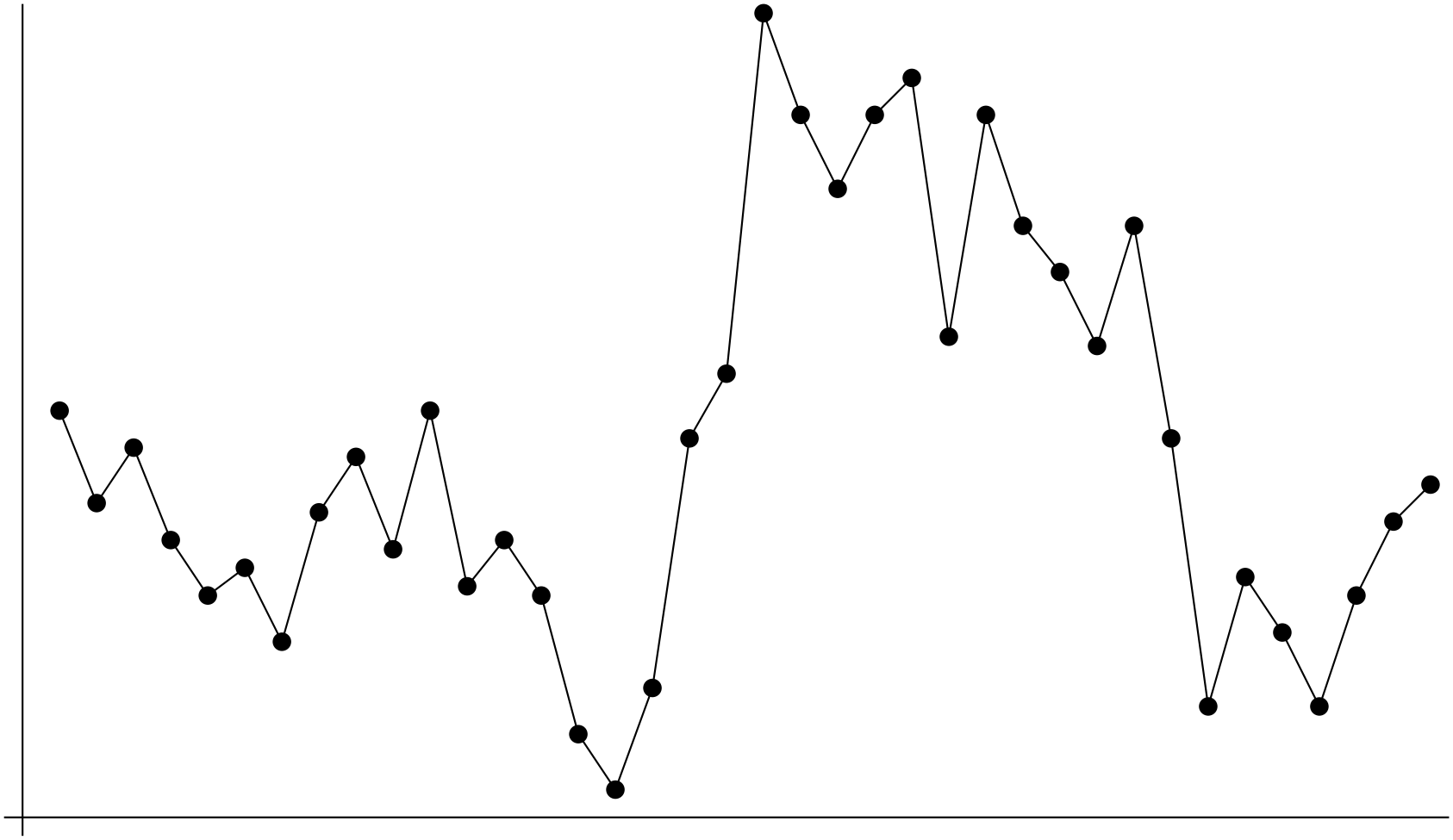
Some facts about the (k, h)-segmentation problem

- NP-Complete problem for multidimensional data ($d > 1$), w.r.t. L_1 and L_2 (contrast with k-segmentation, which is polynomial)
- generalizes k-segmentation and clustering
  - k-segmentation: $h = k$
  - clustering: $k = n$
- simple approximation algorithms that combine the above two subproblems
  - $d = 1$: 3-approximation for L_1, 5-approximation for L_2^2
  - $d > 1$: $(3 + \epsilon)$-approx. for L_1, $(1 + 4\alpha^2)$-approx. for L_2^2, where α is the best approximation factor for k-means

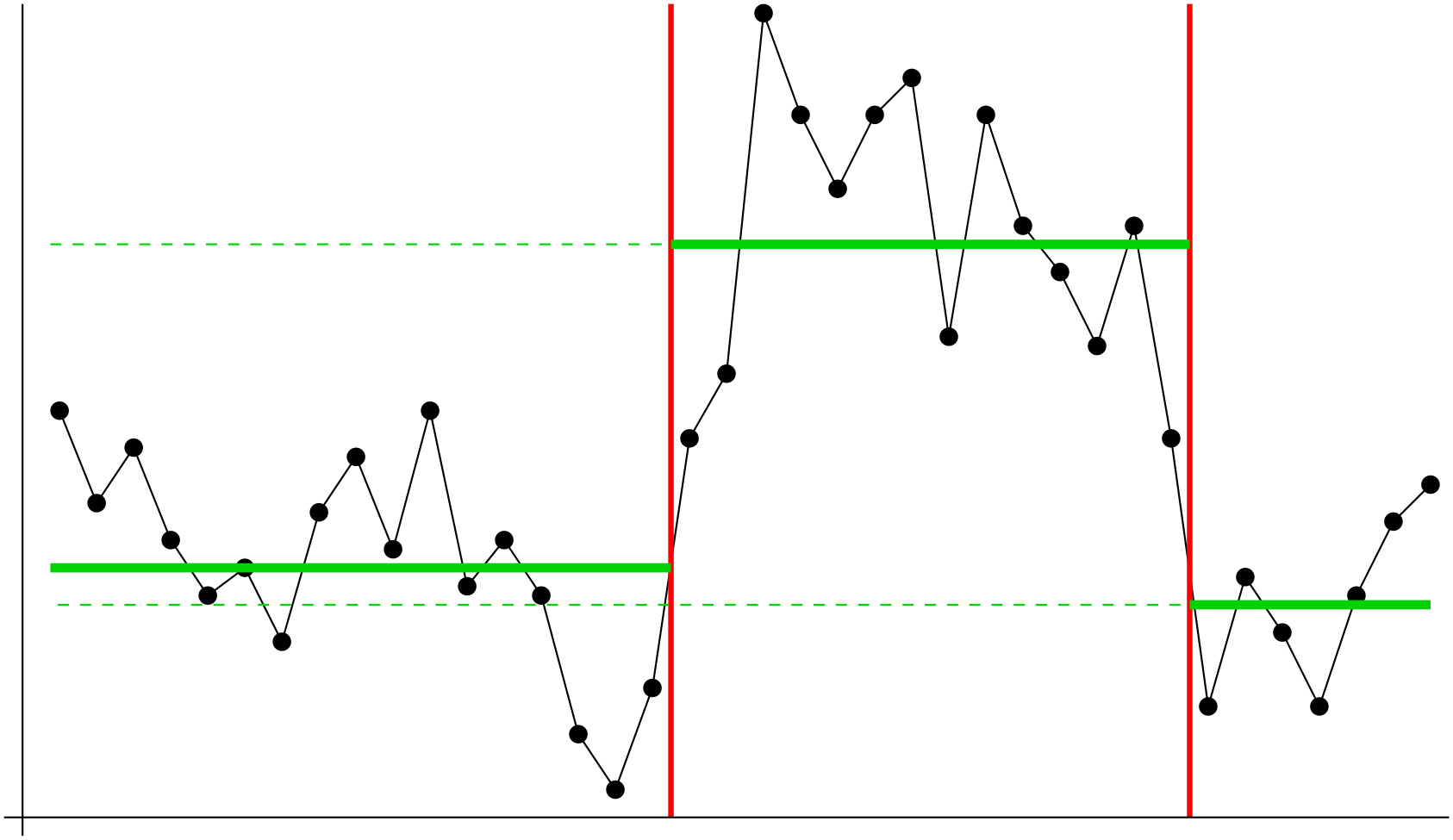
ClusterSegments algorithm

- solve k-segmentation problem and obtain segments $S_1, \ldots, S_k$
- consider the representative c_j for each segment S_j
  (mean, median, etc.)
- map segment S_j to a weighted point with value c_j and weight $w_j = |S_j|$
- cluster those k weighted points to h centers $L = \{l_1, \ldots, l_h\}$
- assign each segment to its closer center in L
- running time is $O(n^2 k)$ (from dynamic programming)

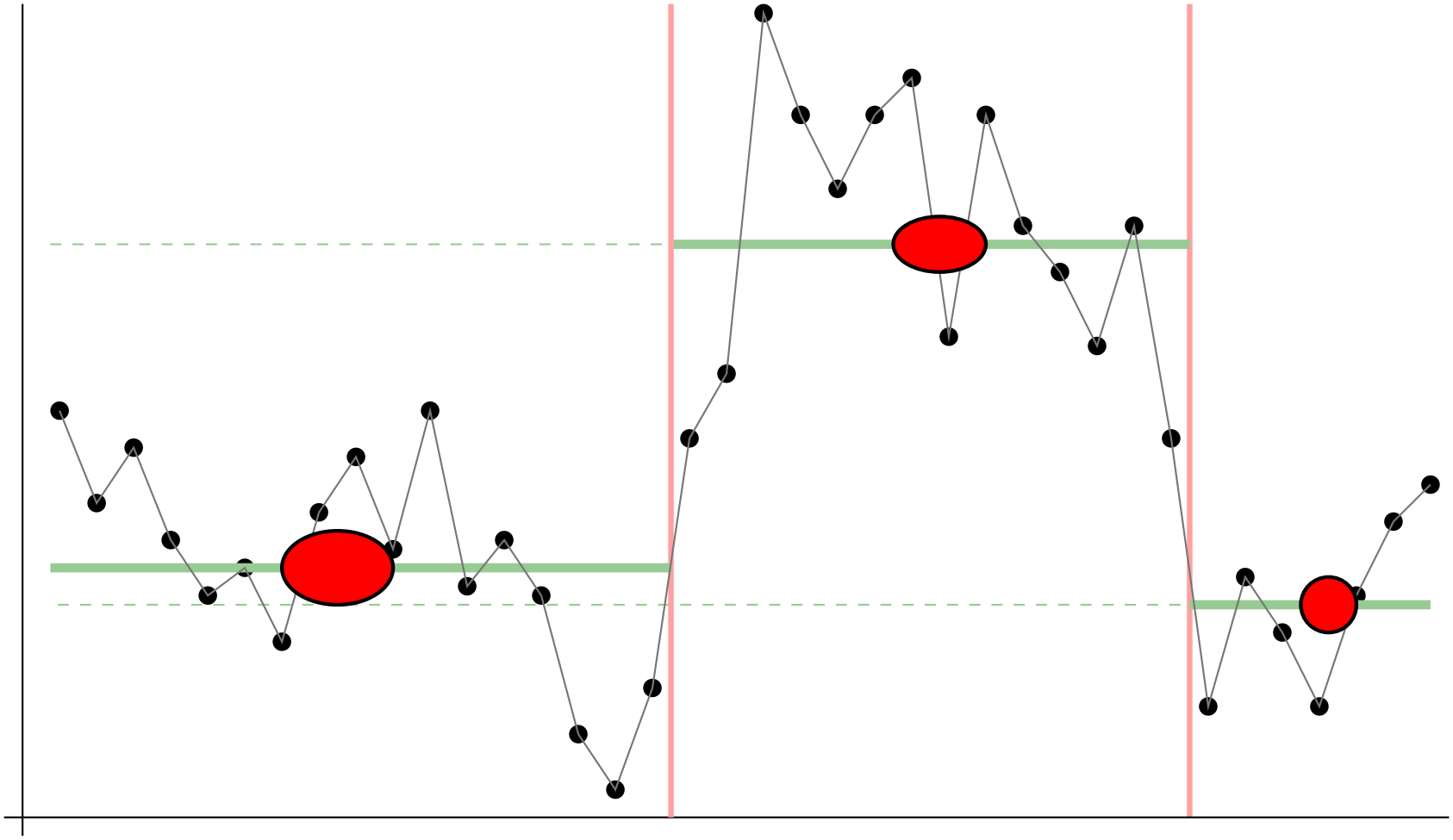
ClusterSegments example, $k = 3$, $h = 2$



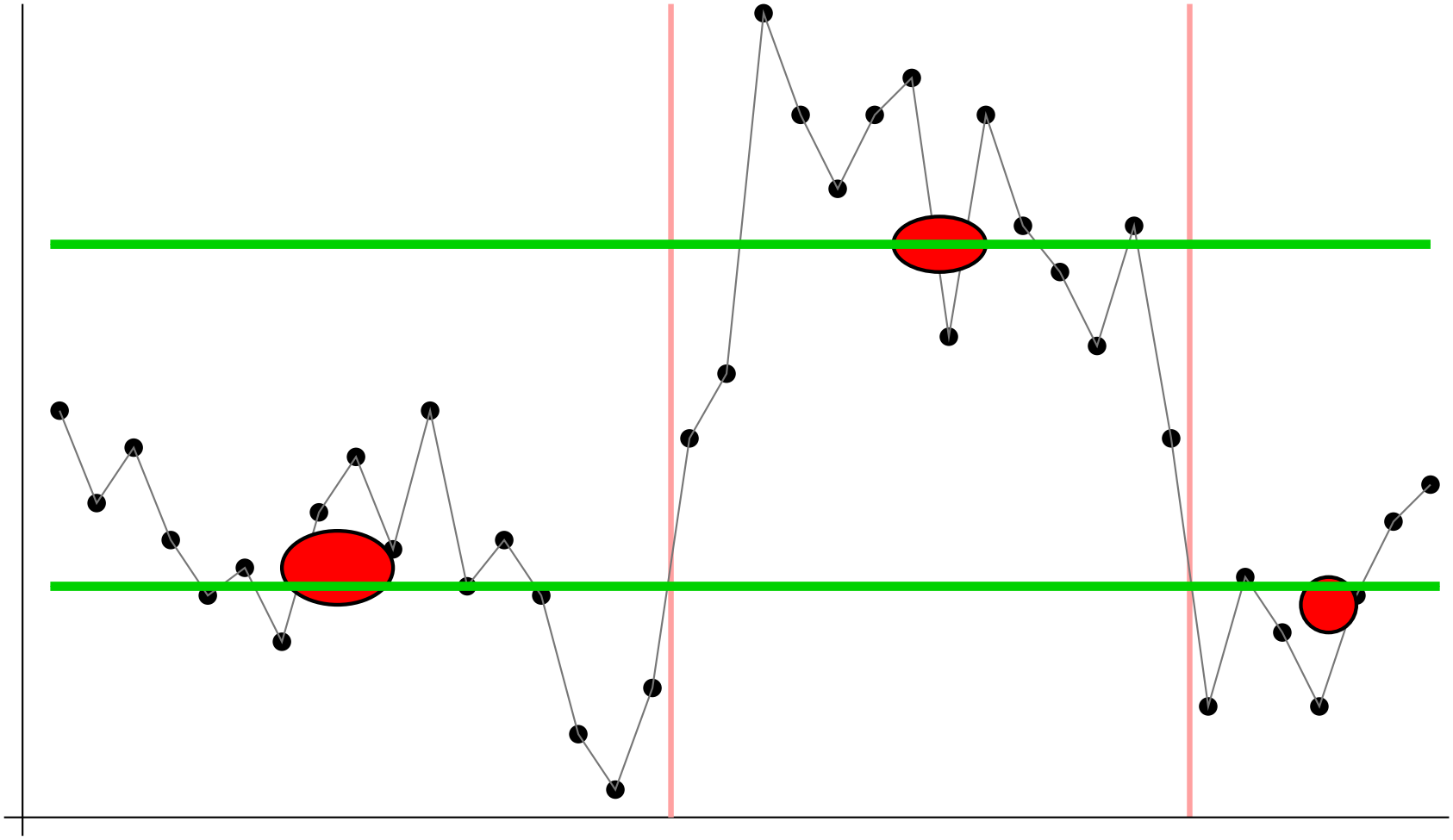
ClusterSegments example, $k = 3$, $h = 2$



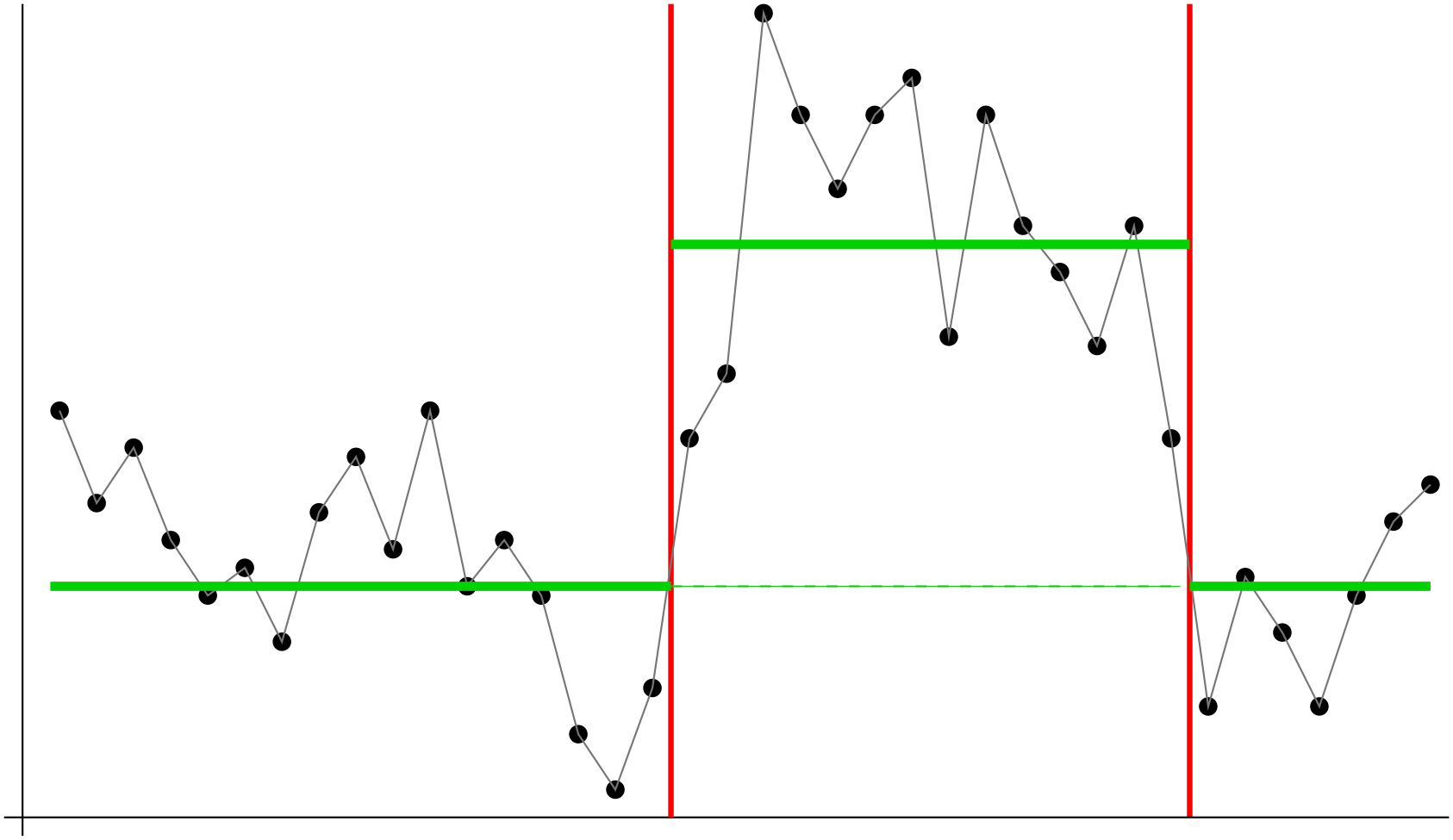
ClusterSegments example, $k = 3$, $h = 2$



ClusterSegments example, $k = 3$, $h = 2$



ClusterSegments example, $k = 3$, $h = 2$



ITERATIVE algorithm

- if we know the k best segments, we can find the h best levels
- if we know the h best levels, we can find the k best segments
- start from an initial solution,
  e.g., the one produced by the previous algorithm
- iterate:
  - keep segment boundaries fixed, find levels
  - keep levels fixed, find boundaries
- EM-style, fast convergence, good results

DNA segmentation

- segmentation: a powerful concept for examining the large-scale organization of DNA
- many examples of segments in DNA
  - (telomere, main-sequence, centromere)
  - (gene-rich, junk DNA)*
  - (regulatory region, gene, regulatory region, junk DNA)*
  - (microbial insert | viral insert | ancient mammalian)*
- goal is to understand the complexity of the genome organization based on segments and recurrent sources

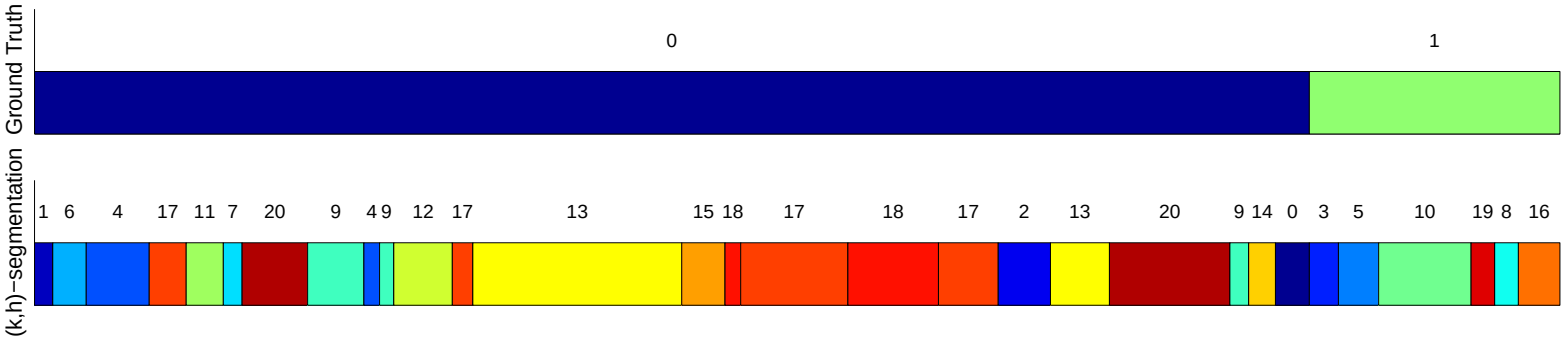
DNA segmentation

- existing approaches with top-down segmentation and greedy identification of similar segments [Bernaola-Galván et al. 96, Bernaola-Galván et al. 00, Li 01, Azad et al. 02]
- here we describe some of our own experiments with (k, h)-segmentation [Haiminen et al. 05]

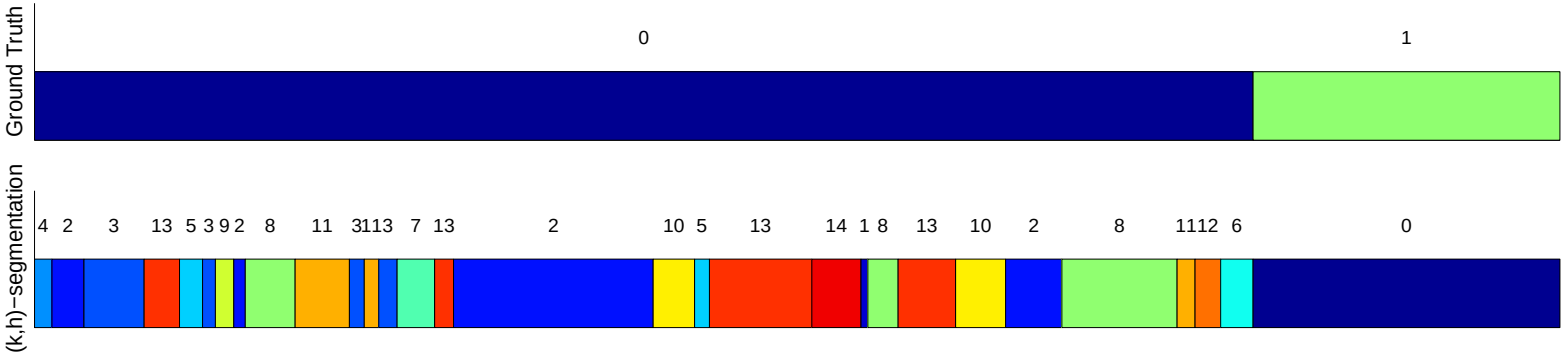
Distinguishing genomes of different species

- create many "semi-synthetic" datasets $HiSj$ by concatenating
  - Hi: the i-th chromosome of human with
  - Sj: the j-th chromosome of another species S
- apply (k, h)-segmentation and compare with the ground truth segmentation
- let $L = \{l_1, \ldots, l_h\}$ be the discovered sources in the concatenated sequence, and $\mathcal{L}_H$ and $\mathcal{L}_S$ be the distribution of lengths of sources of L in chromosomes H and S, resp.
- compare the variational distance between the two distributions
  - 0: identical distributions, 1: completely distinct distributions

Genomes of different species — sample segmentations

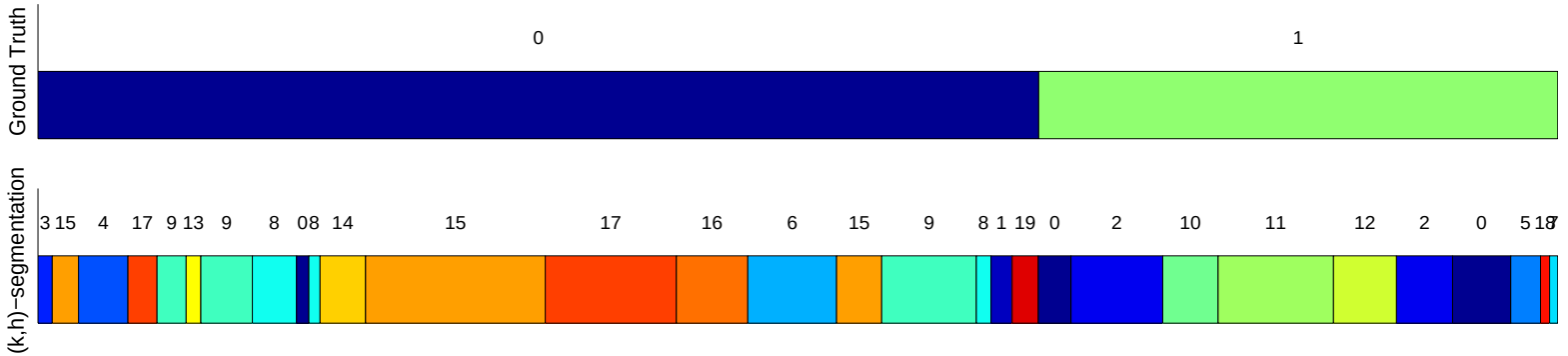


human 8 vs. chicken 8

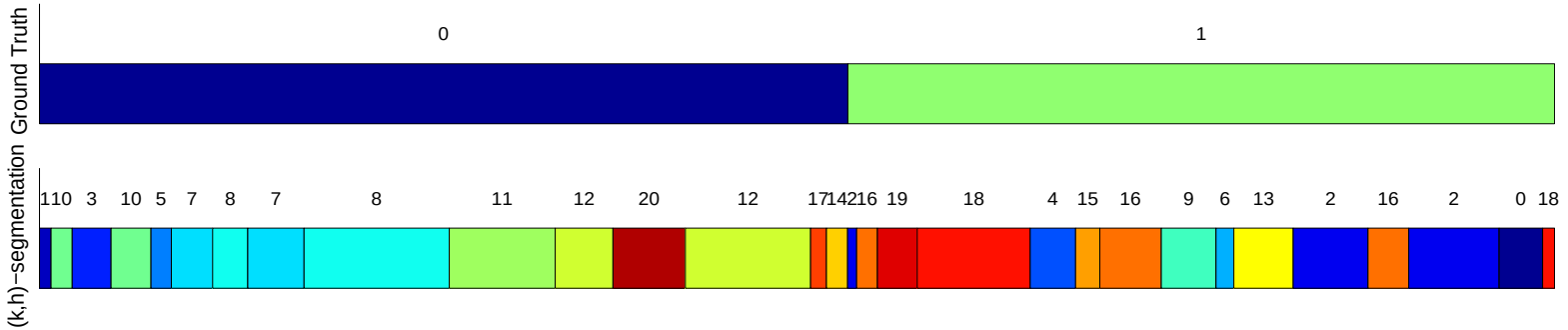


human 8 vs. zebra fish 8

Genomes of different species — sample segmentations

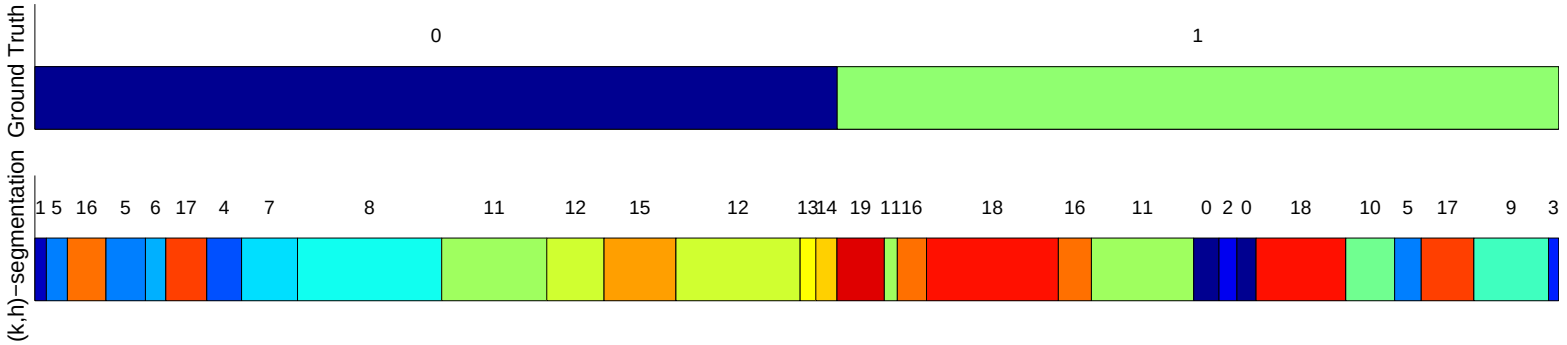


human 8 vs. dog 8

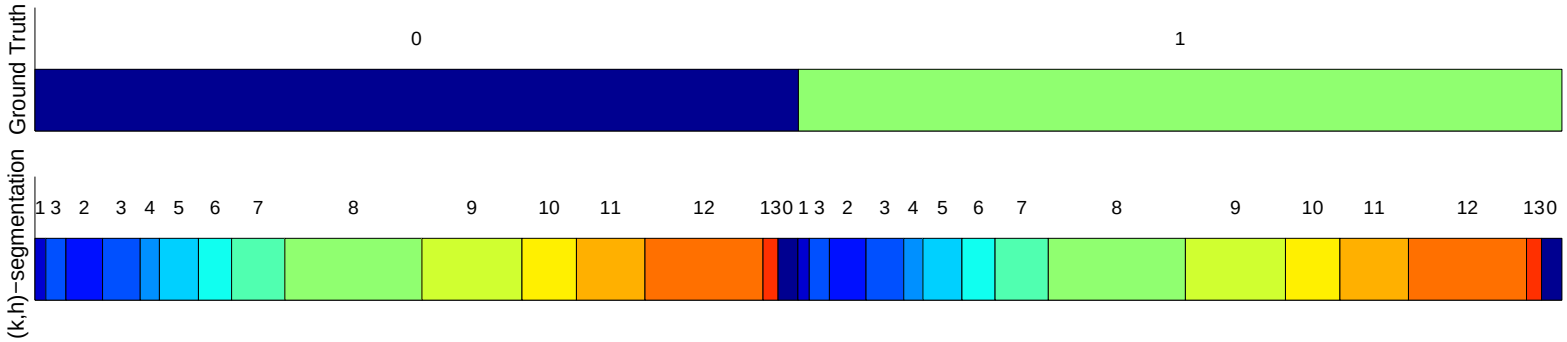


human 8 vs. mouse 8

Genomes of different species — sample segmentations

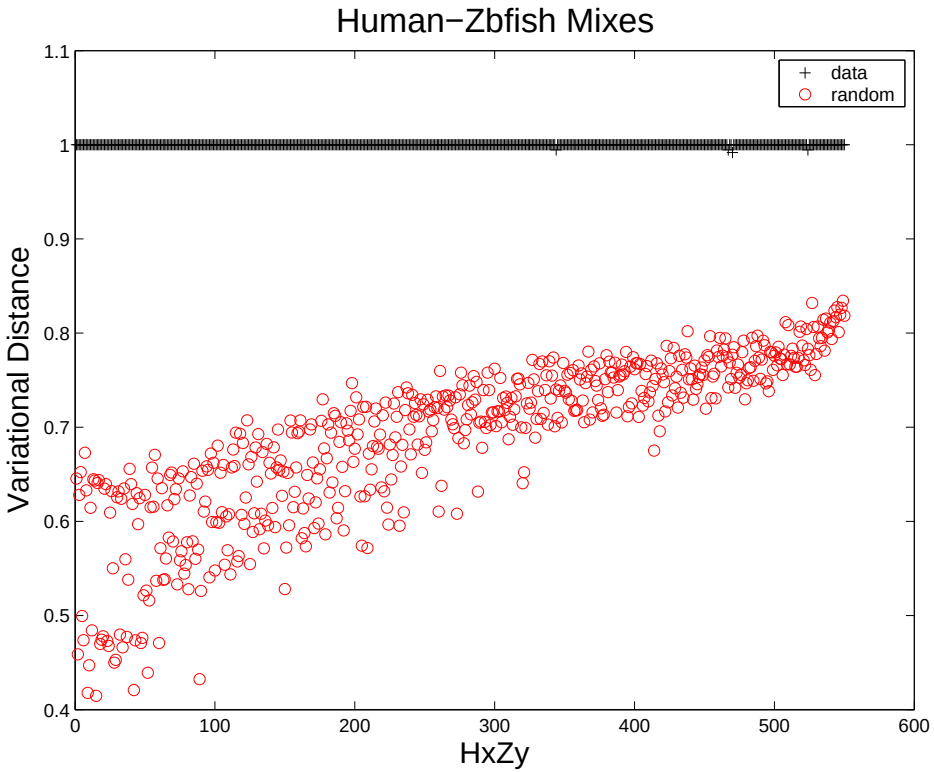


human 8 vs. chimp 8

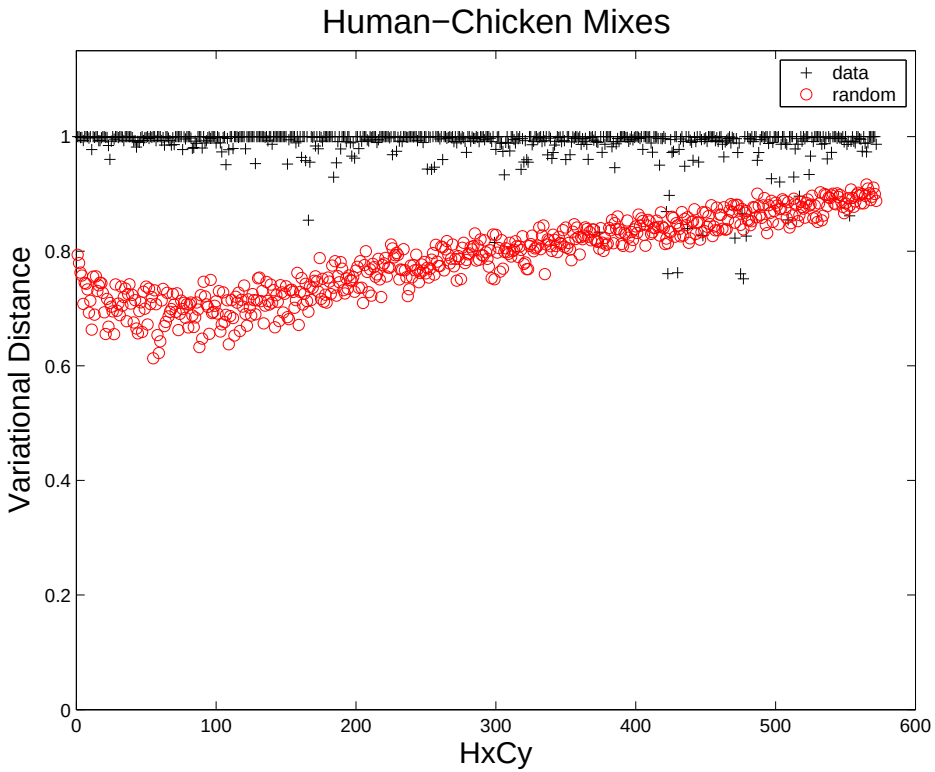


human 8 vs. human 8

Genomes of different species — variational distances

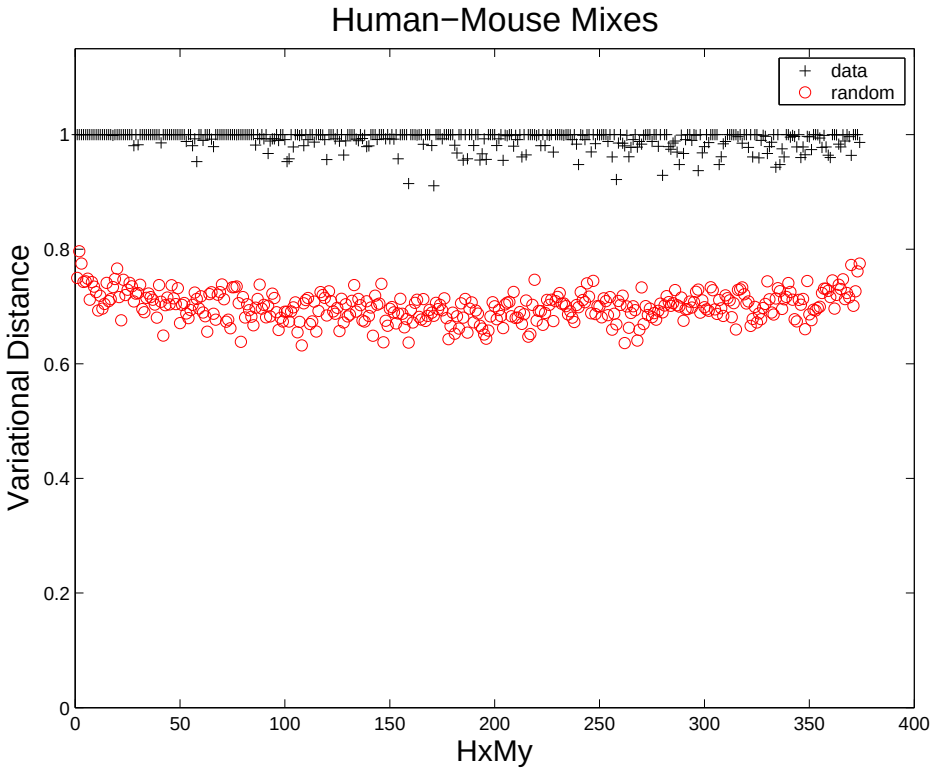


human vs. zebrafish

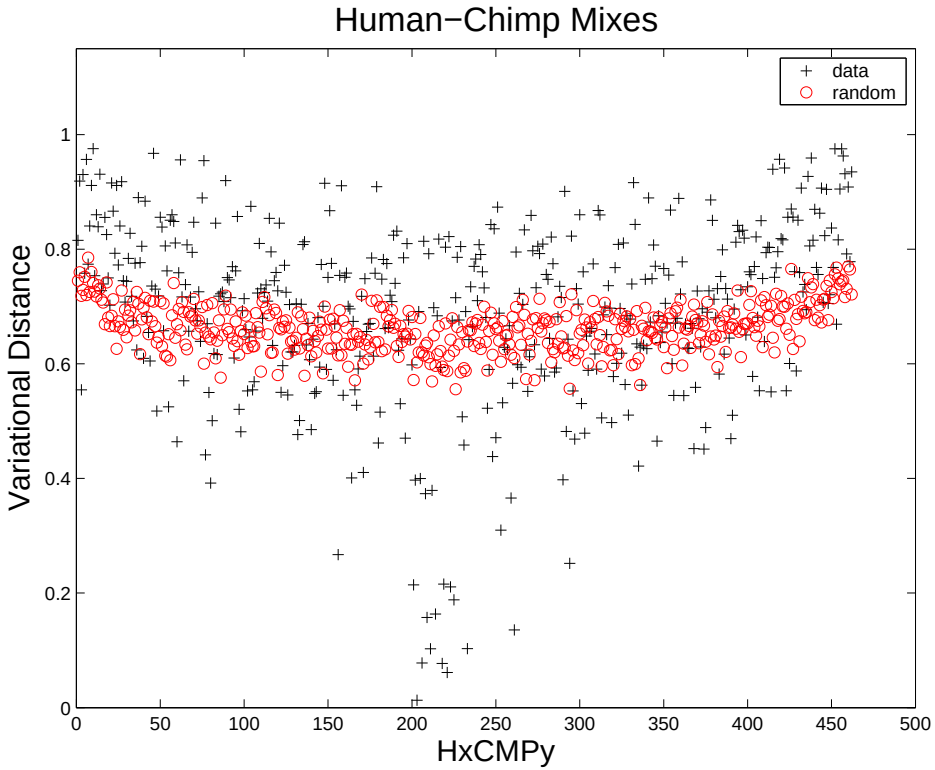


human vs. chicken

Genomes of different species — variational distances



human vs. mouse

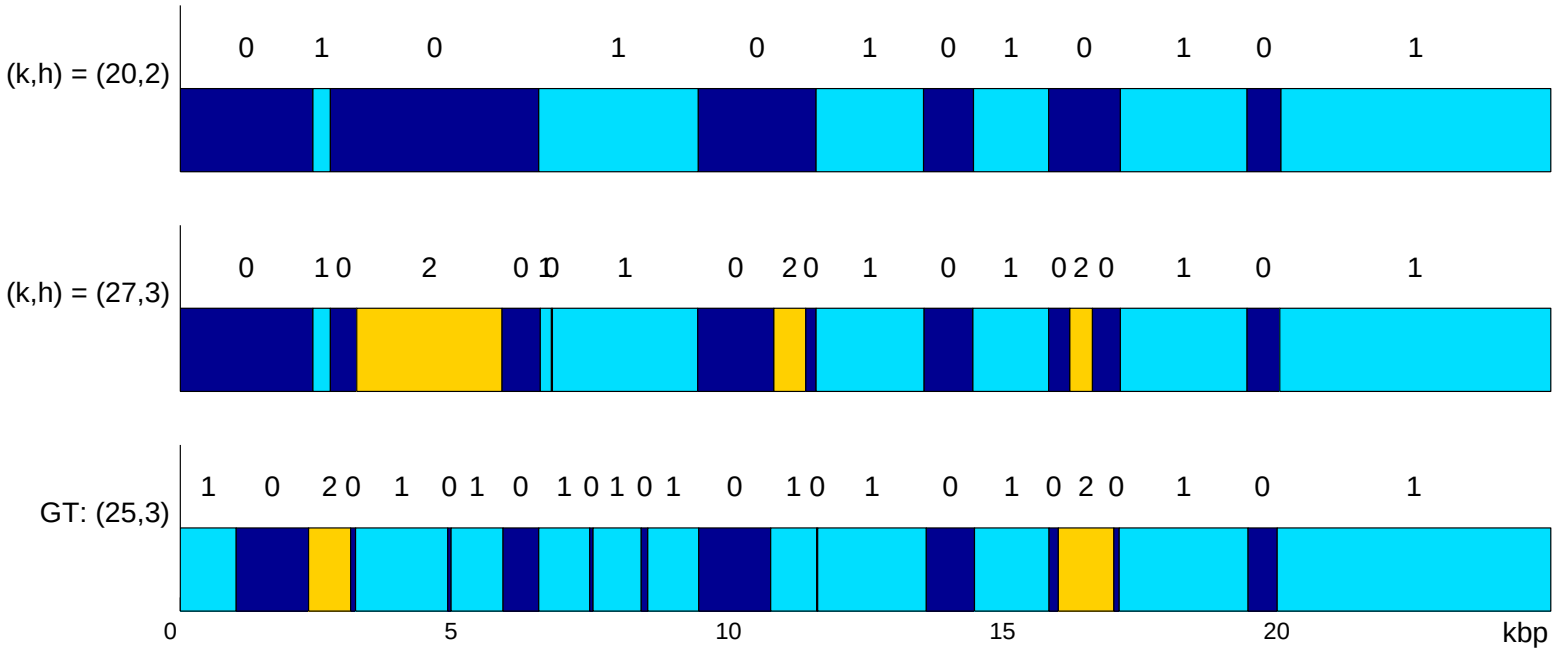


human vs. chimp

Distinguishing coding from non-coding regions

- *Rickettsia* bacterium region that includes 13 genes and non-coding in-between region
- 10 bp non-overlapping windows
- in each window features that capture the existence of codons

Distinguishing coding from non-coding regions



DNA segmentations — conclusions

- segmentation is promising tool for analyzing genomic sequences
- fascinating problem of understanding the structure of DNA

Thank you!

- for your attention
- Helger Lipmaa and Tarmo Uustalu for the invitation
- hope to learn more about CS research and theory in Estonia...
- ...hope to enjoy the weekend, too!