NATIONAL RESEARCH
UNIVERSITY

Introduction to Network Science

Social Network Analysis. MAGoLEGO course

Leonid Zhukov

lzhukov@hse.ru
www.leonidzhukov.net/hse/2018/sna

National Research University Higher School of Economics
School of Data Analysis and Artificial Intelligence, Department of Computer Science

April 6, 2018

Class Technicalities

- Instructors: Leonid Zhukov, Ilya Makarov
- MAGoLEGO course: 1 module
- 10 lectures, 10 labs, 4 homeworks
- Final exam
- Schedule: Fridays, 18.10-21.00 (lecture + lab)
- Website: http://www.leonidzhukov.net/hse/2018/sna
- Emails: lzhukov@hse.ru, iamakarov@hse.ru

Helpful background

Theory:

- Discrete Mathematics
- Linear Algebra
- Probability Theory
- Differential Equations
- Algorithms and Data Structures

Programming experience:

- R, RStudio
- R libraries: igraph
- Visualization: Gephi

Textbooks

- "Network Science", Albert-Laszlo Barabasi, Cambridge University Press, 2016.
- "Networks, Crowds, and Markets: Reasoning About a Highly Connected World". David Easley and John Kleinberg, Cambridge University Press 2010.
- "Statistical Analysis of Network Data with R", Eric Kolaczyk, Gabor Csardi, Springer, 2014.
- "Social Network Analysis. Methods and Applications". Stanley Wasserman and Katherine Faust, Cambridge University Press, 1994

Lectures

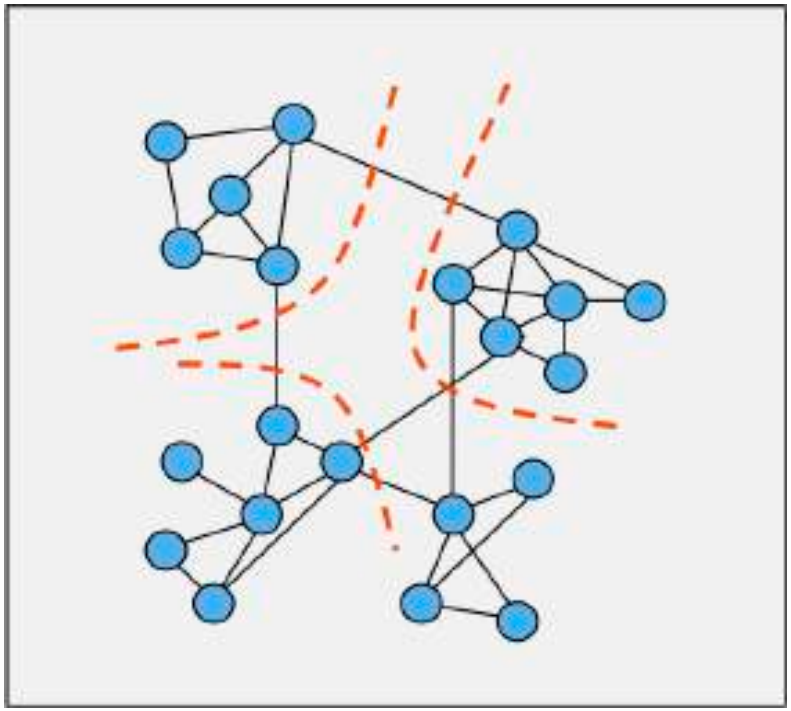
1. Introduction to network science
2. Descriptive network analysis
3. Mathematical models of networks
4. Node centrality and ranking on networks
5. Network communities
6. Network structure and visualization
7. Epidemics and information spreading in networks
8. Diffusion of innovation
9. Strategic network formation
10. Spatial models of segregation

Network science

- Sociology (SNA)
- Mathematics (Graphs)
- Computer Science (Graphs)
- Statistical Physics (Complex networks)
- Economics (Networks)
- Bioinformatics (Networks)

Terminology

- network = graph
- nodes = vertices, actors
- links = edges, relations
- clusters = communities



Examples: Social network

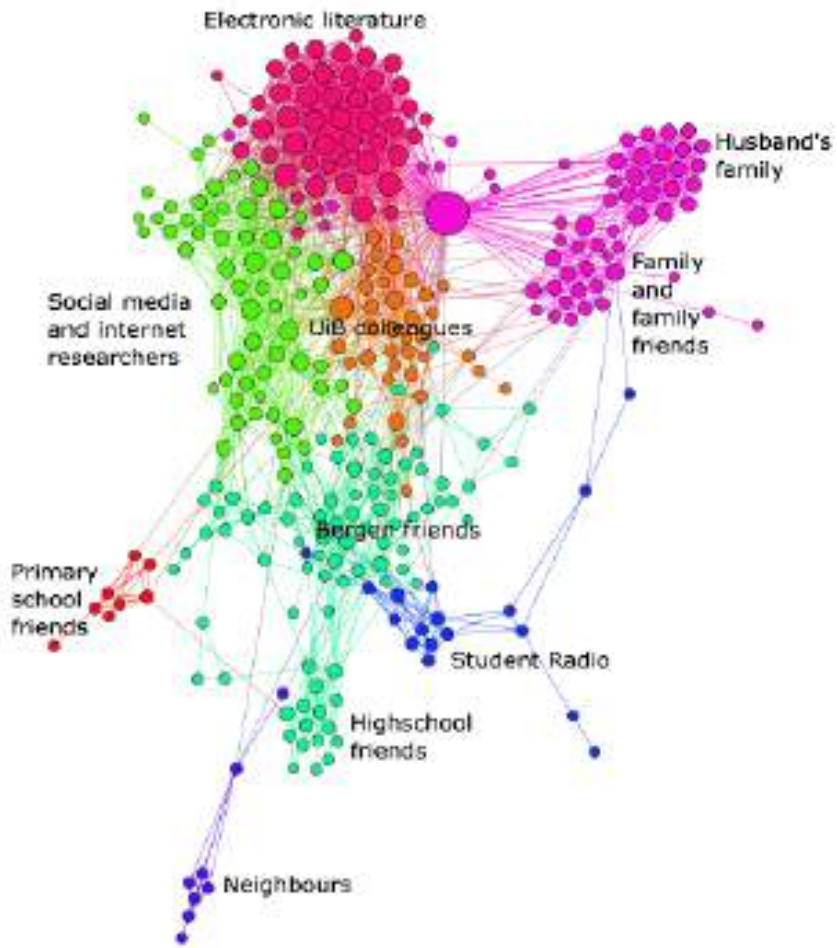
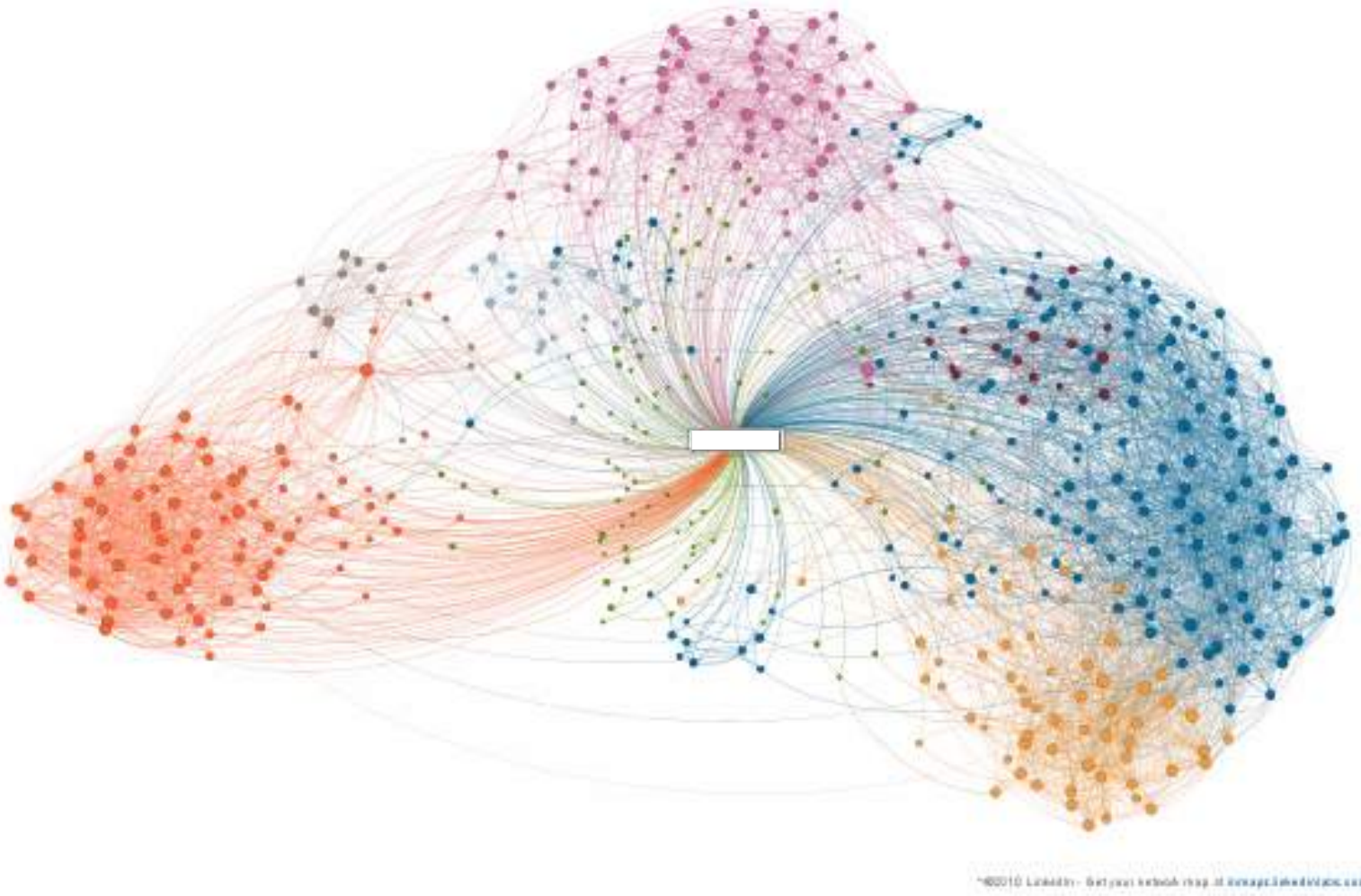


image from Jill Walker Rettberg, jilttxt.net

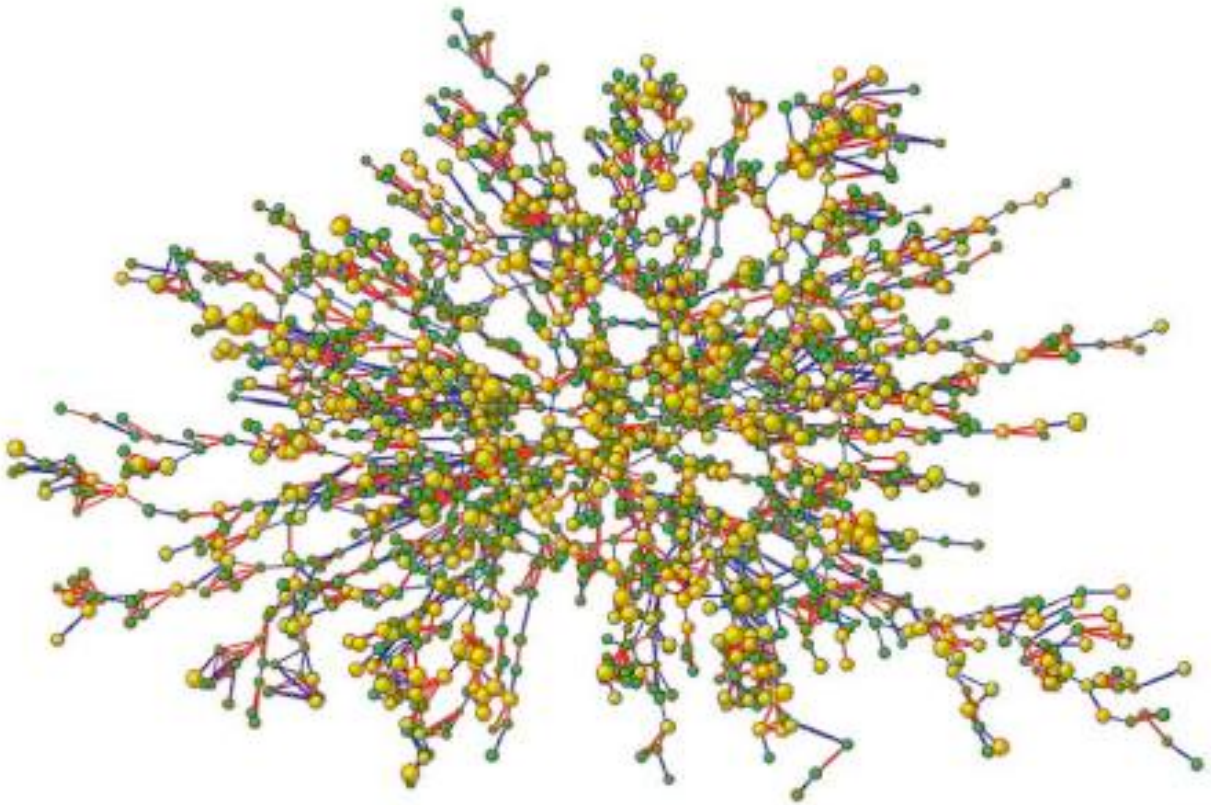
Examples: LinkedIn Map

LinkedIn contacts ego-centric network



Examples: Social network

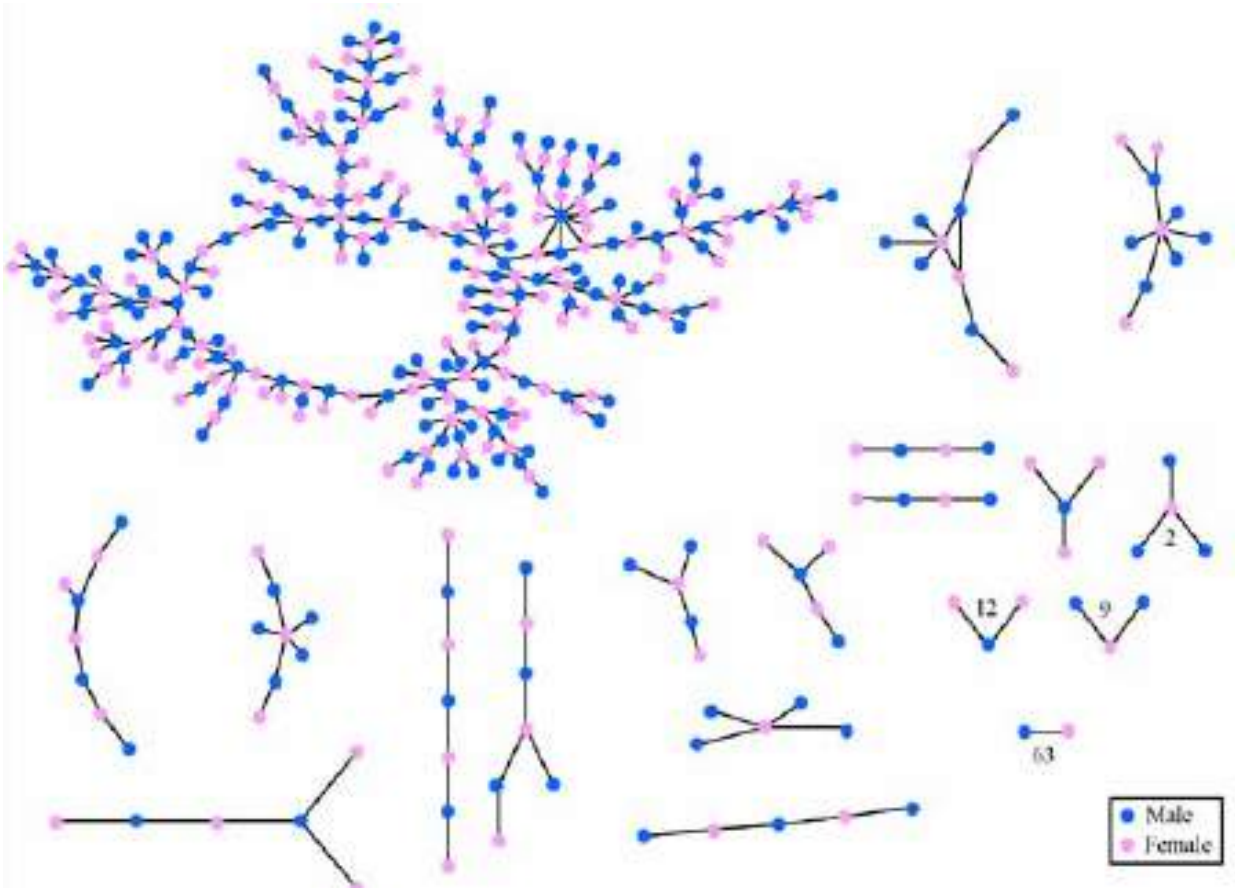
"The Spread of Obesity in a Large Social Network over 32 Years"



N. Christakis , J. Fowler 2007

Examples: high school dating network

"Chains of Affection: The Structure of Adolescent Romantic and Sexual Networks"



P. Bearman, J. Moody, K. Stovel, 2004

Examples: Political blogs

red-conservative blogs, blue -liberal, orange links from liberal to conservative, purple from conservative to liberal

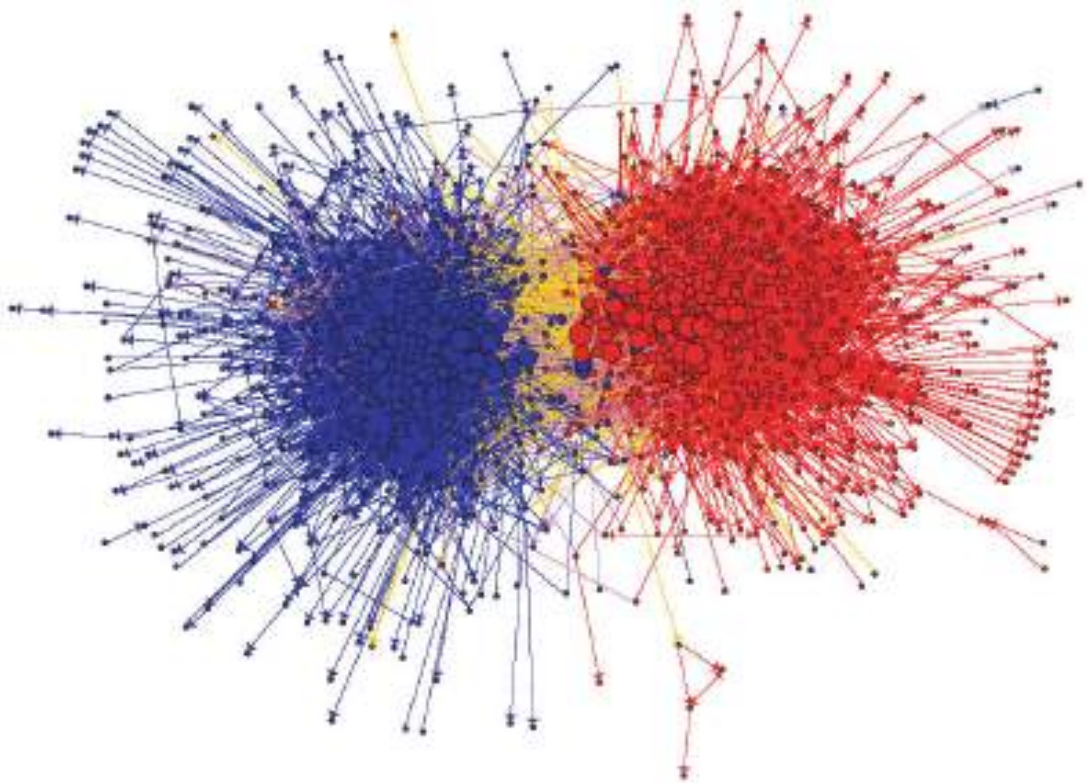


image from L. Adamic, N. Glance, 2005

Examples: Twitter

”#usa” hashtag diffusion, retweets - blue, mentions - orange

image from K. McKelvey et.al., 2012

Examples: Twitter

News about Bin Laden

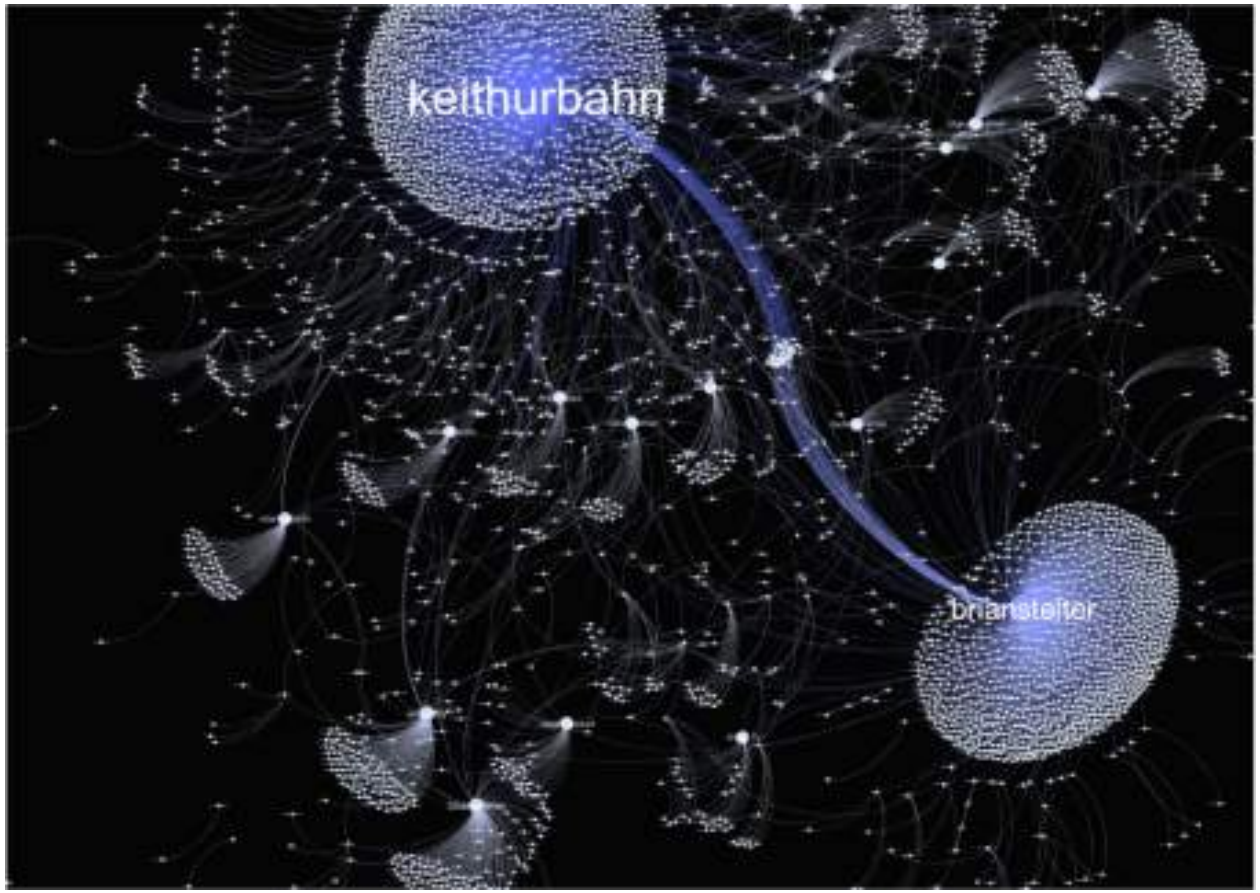
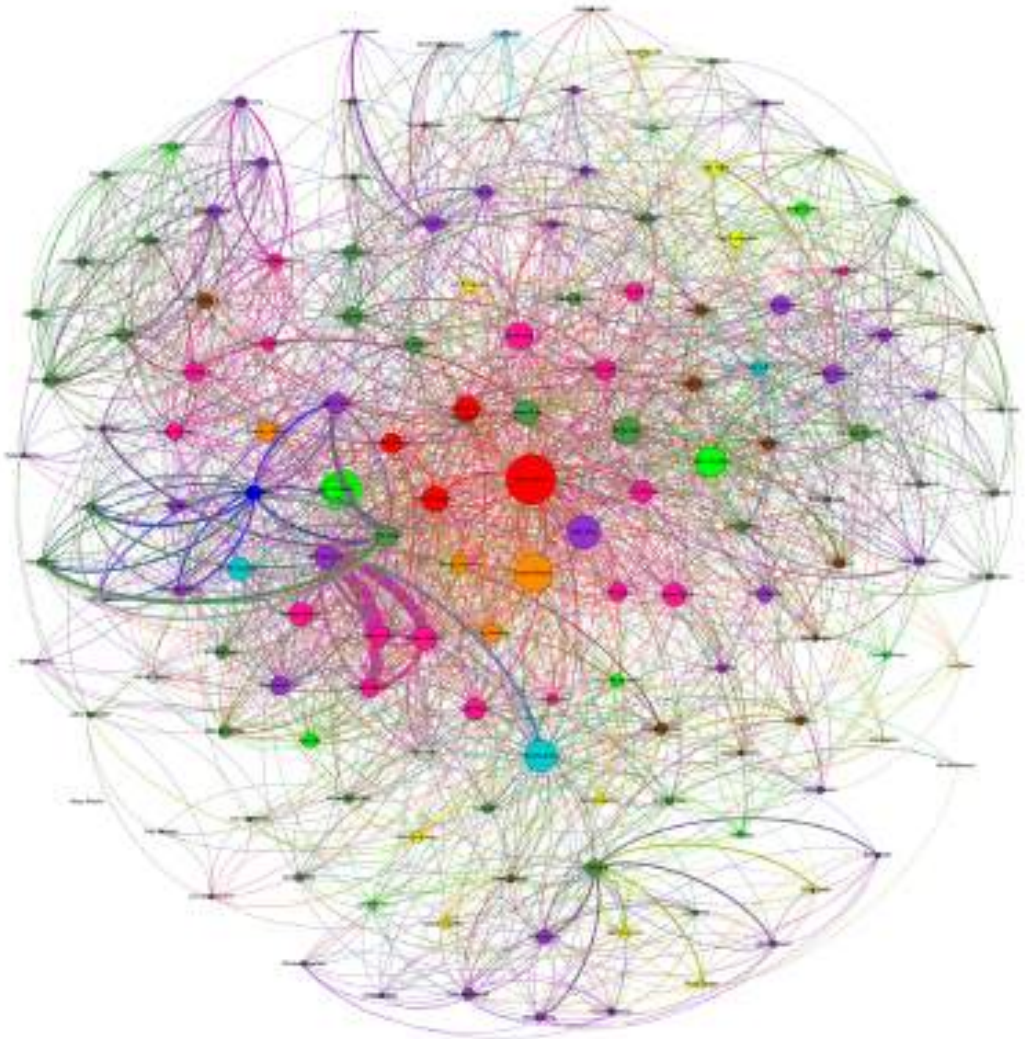


image from SocialFlow

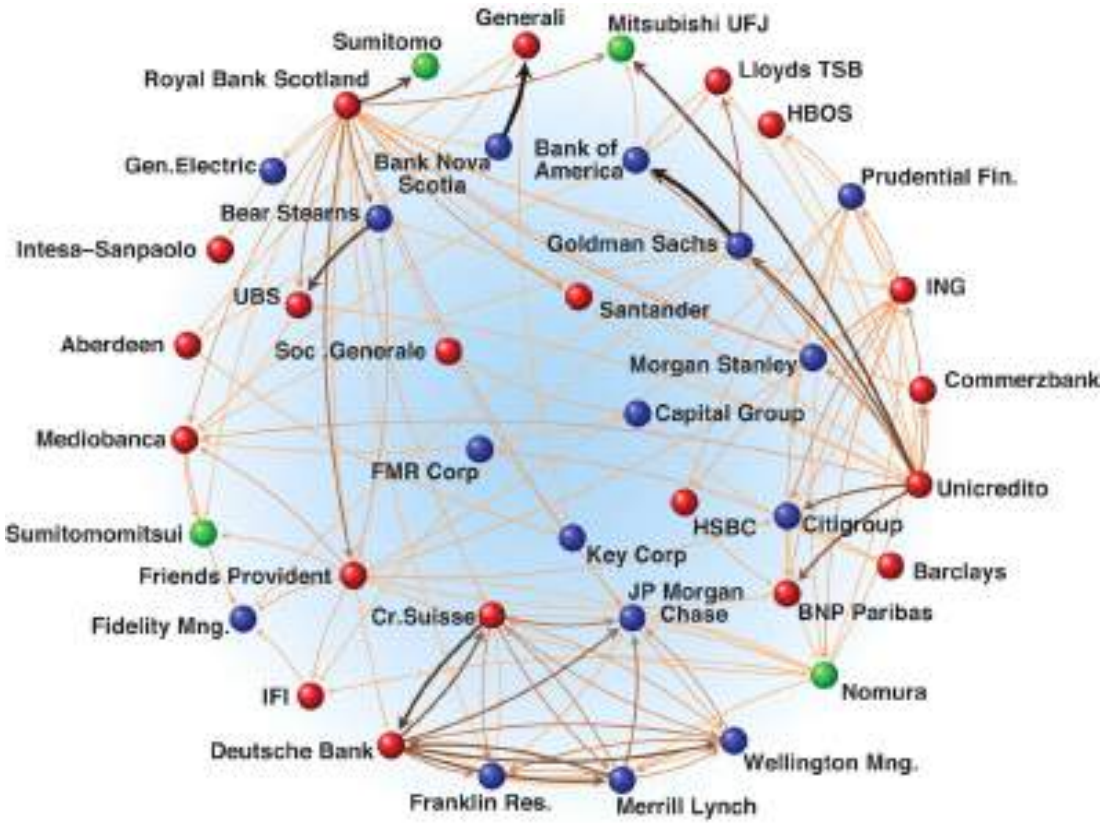
Examples: Emails

Enron emails



Examples: Finance

existing relations between financial institutions



F. Schweitzer, 2009

Examples: Transportation

Zurich public transportation map

image from http://www.visualcomplexity.com

Examples: Transportation

London bike share

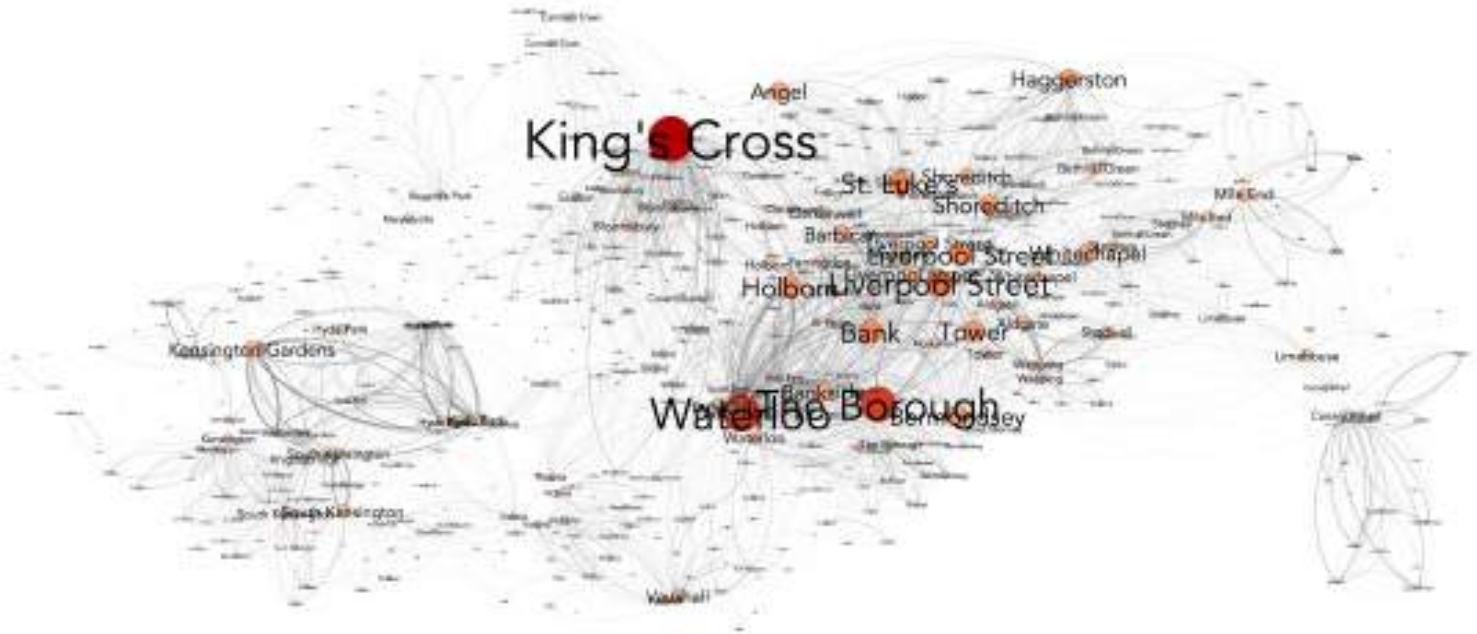
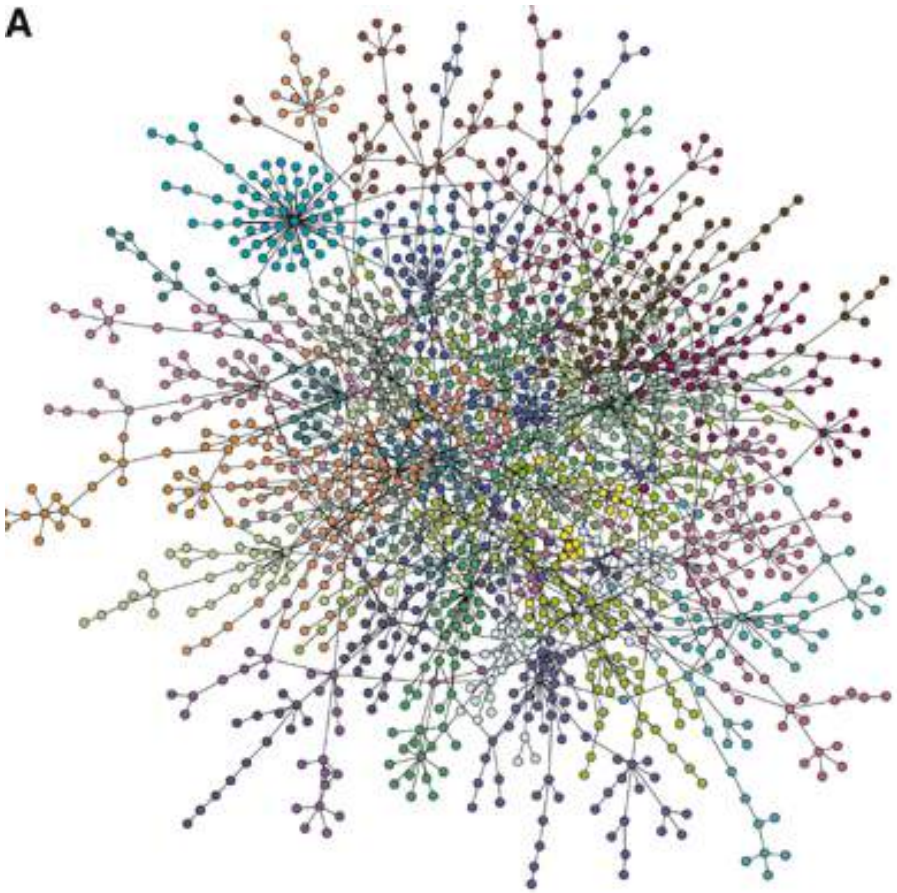


image from vartree.blogspot.com

Examples: Biology

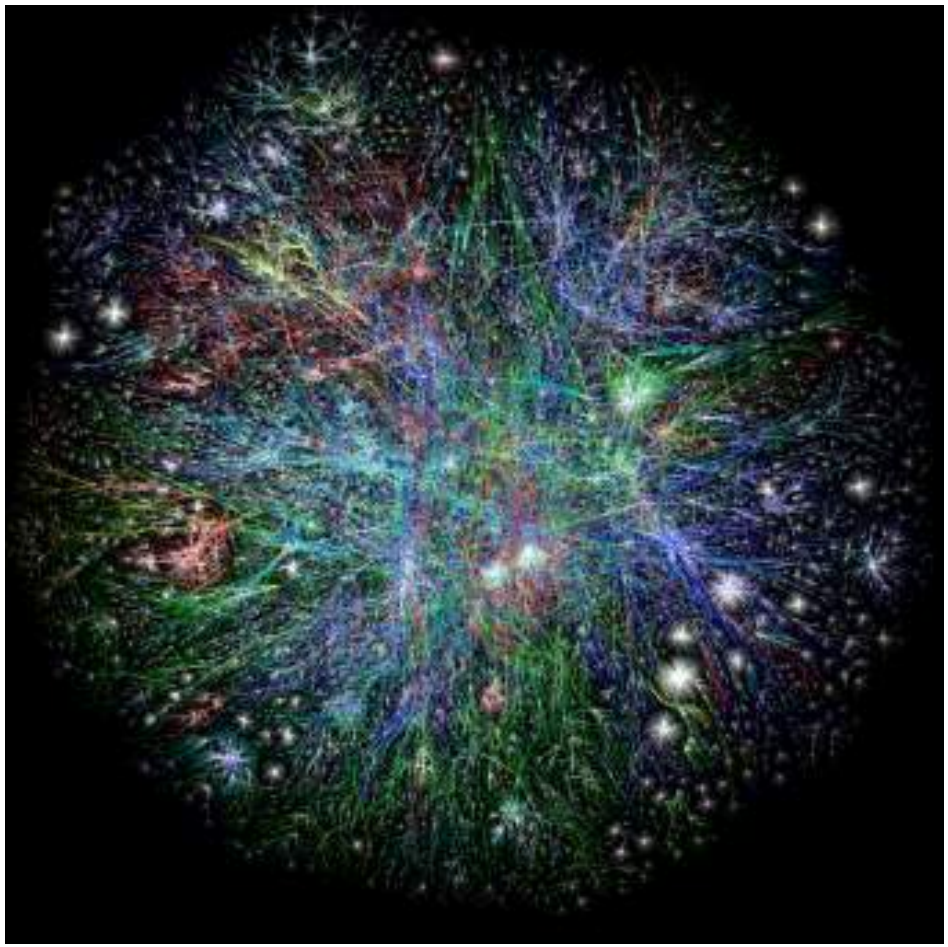
Yeast protein interaction network



H. Jeong et.al., 2001

Examples: Internet

Internet traffic routing (BGP)



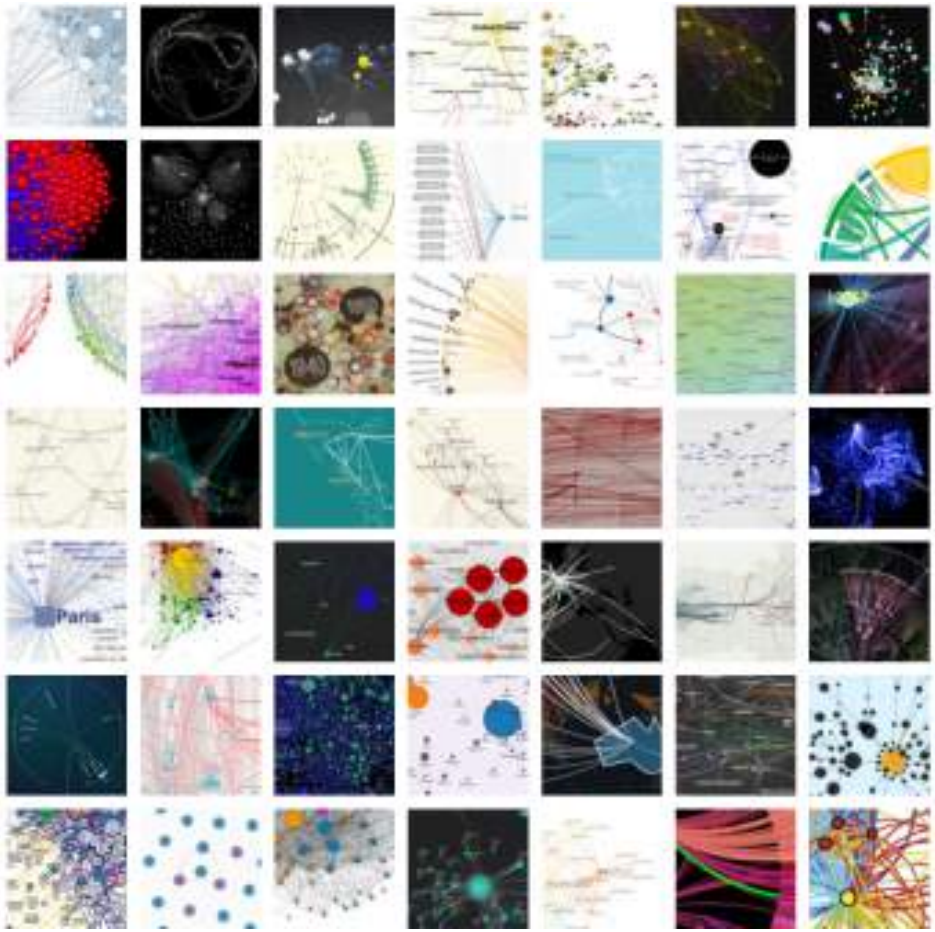
Barret Lyon, 2003

Examples: Facebook

Friendship graph 500 mln people

image by Paul Butler, 2010

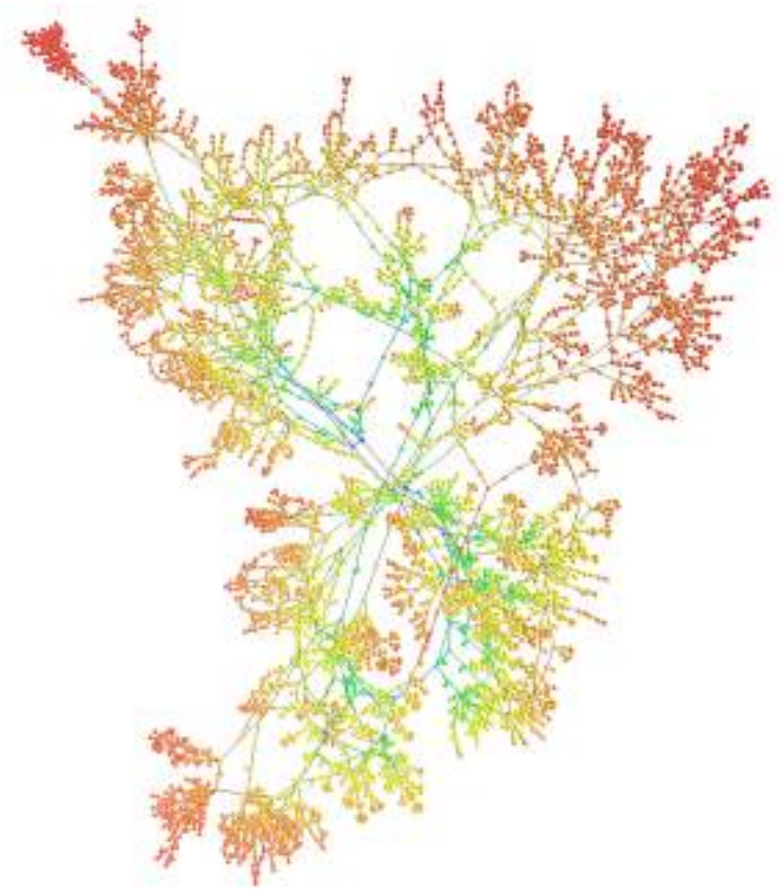
Visual complexity



www.visualcomplexity.com

Complex networks

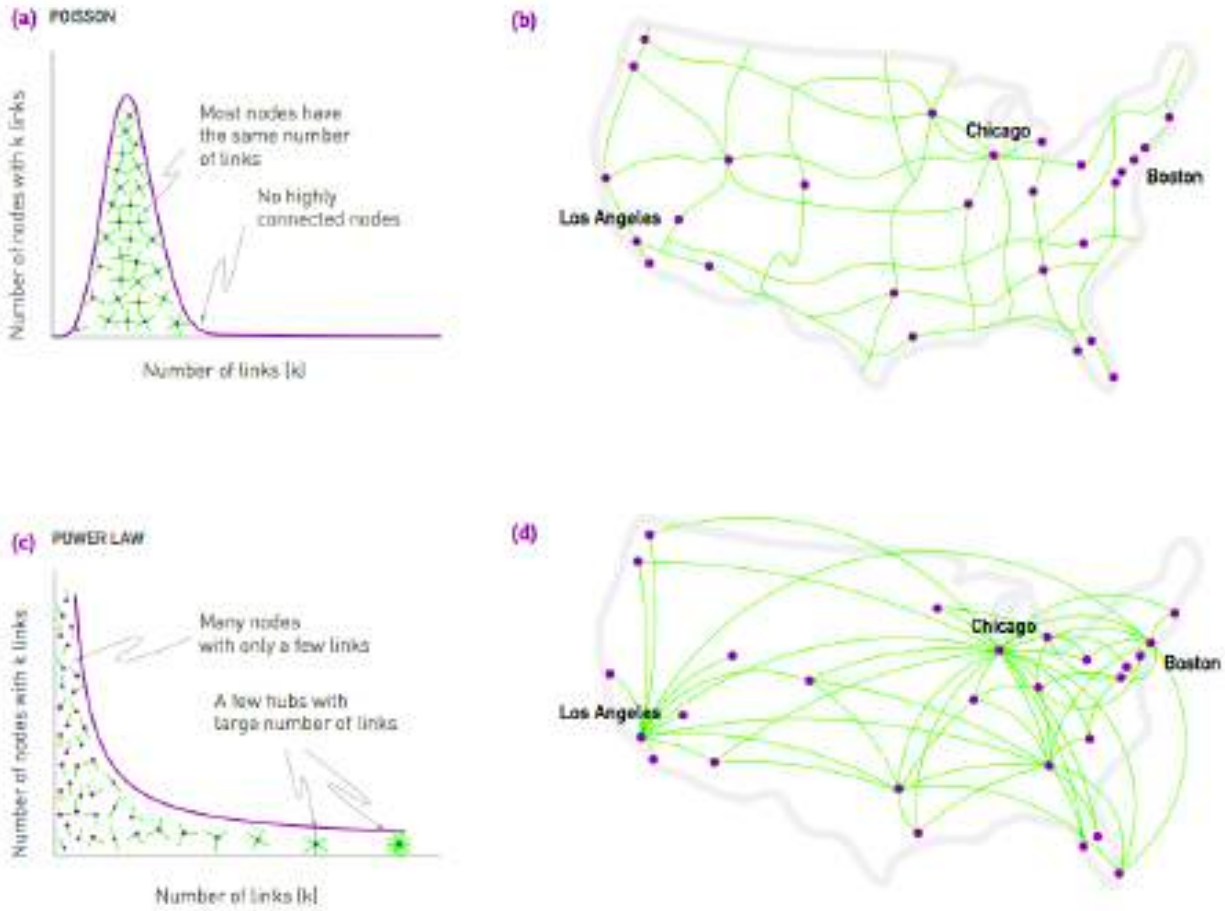
- not regular, but not random
- non-trivial topology
- universal properties
- everywhere
- complex systems



Complex networks

1. Power law node degree distribution: ”scale-free” networks
2. Small diameter and average path length: ”small world” networks
3. Hight clustering coefficient: transitivity

Power law



Frequency distribution of node degrees $f(k) \sim \frac{1}{k^{\gamma}}$

image from A.-L. Barabasi, 2016

Power law

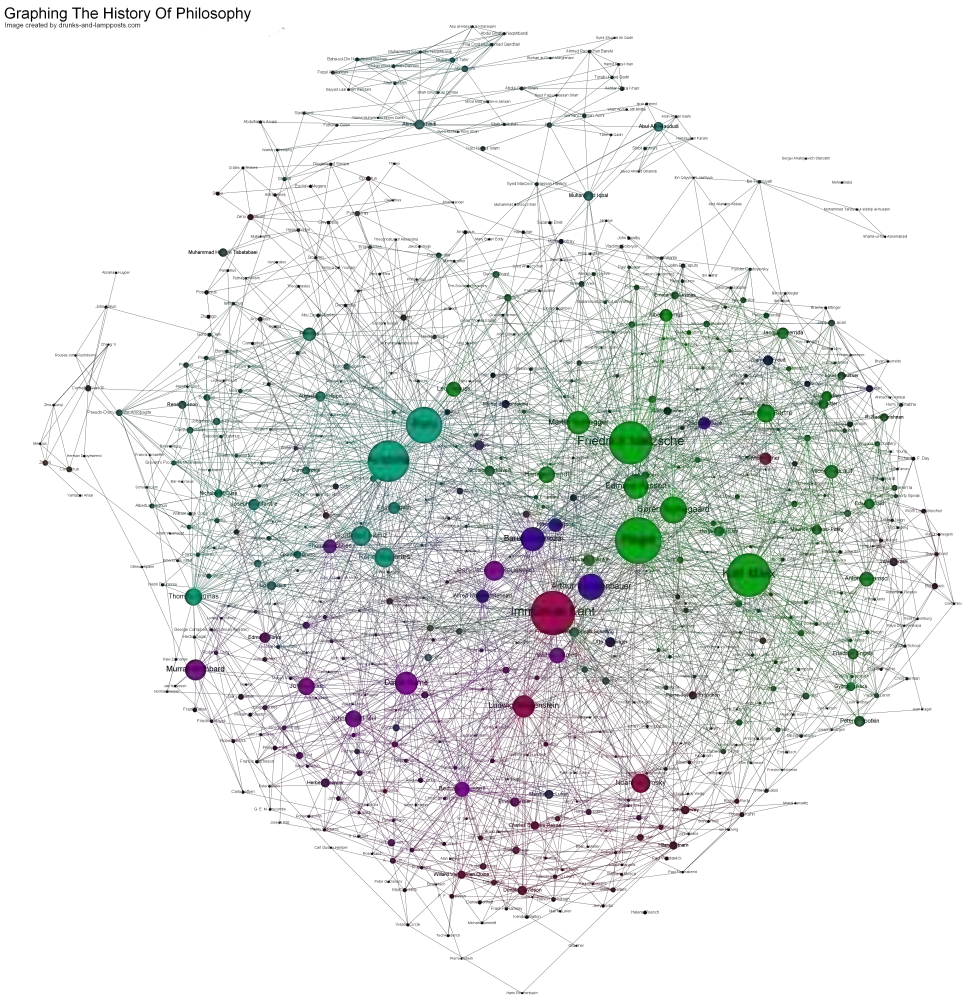


image from http://www.coppelia.io

Power law

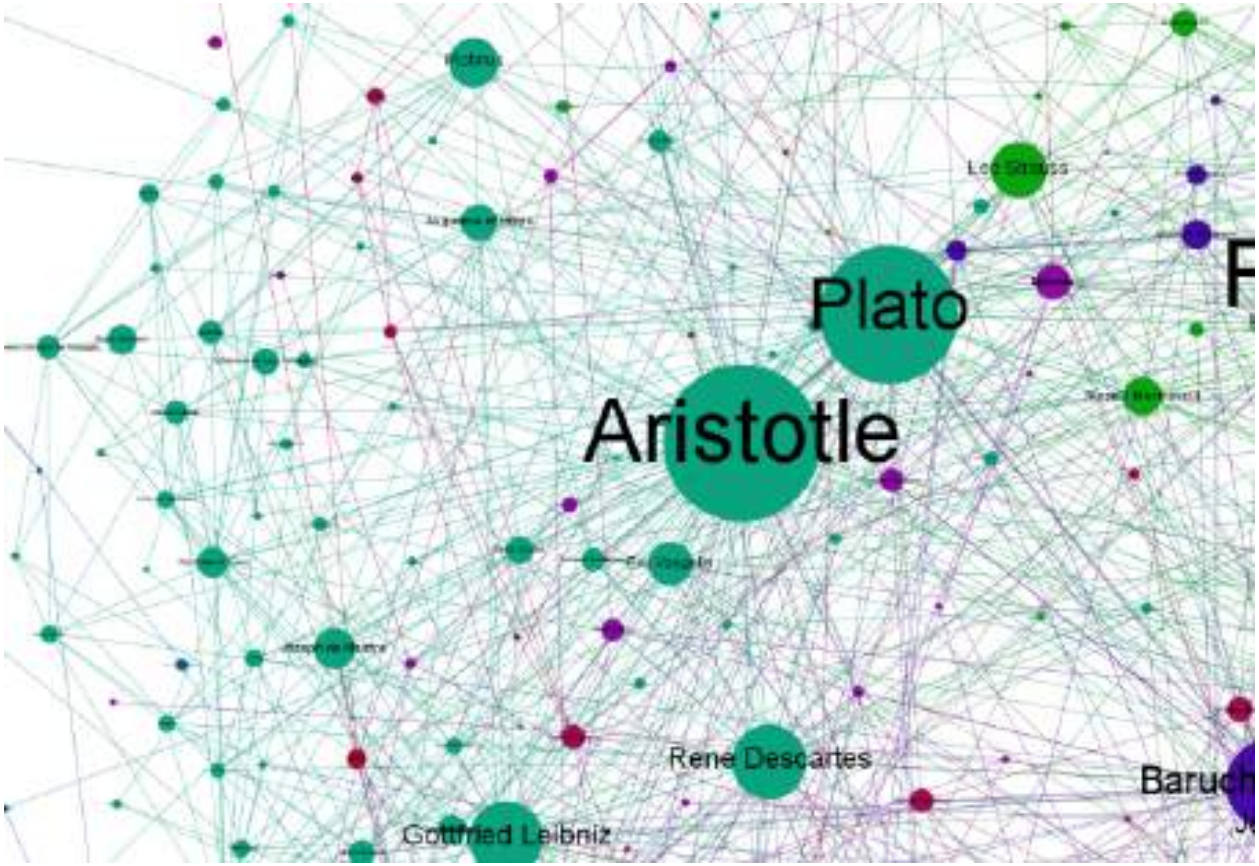


image from http://www.coppelia.io

Triads

The Strength of Weak Ties[1]

Mark S. Granovetter
Johns Hopkins University

Analysis of social networks is suggested as a tool for linking micro and macro levels of sociological theory. The procedure is illustrated by elaboration of the macro implications of one aspect of small-scale interaction: the strength of dyadic ties. It is argued that the degree of overlap of two individuals' friendship networks varies directly with the strength of their tie to one another. The impact of this principle on diffusion of influence and information, mobility opportunity, and community organization is explored. Stress is laid on the cohesive power of weak ties. Most network models deal, implicitly, with strong ties, thus confining their applicability to small, well-defined groups. Emphasis on weak ties lends itself to discussion of relations *between* groups and to analysis of segments of social structure not easily defined in terms of primary groups.

- "The Strength of Weak Ties", Mark Grannoveter, 1973
- "Spread of Information through a Population with Socio-Structural Bias. Assumption of Transitivity", Anatol Rapoport, 1953

Triadic closure

- strength of a tie
- high transitivity
- high clustering coefficient

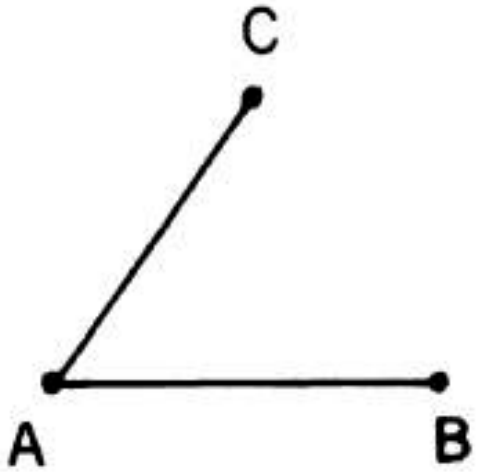


FIG. 1.—Forbidden triad

If A and B and B and C are strongly linked, the the tie between B and C is always present

Grannoveter, 1973

High clustering

Facebook friendship

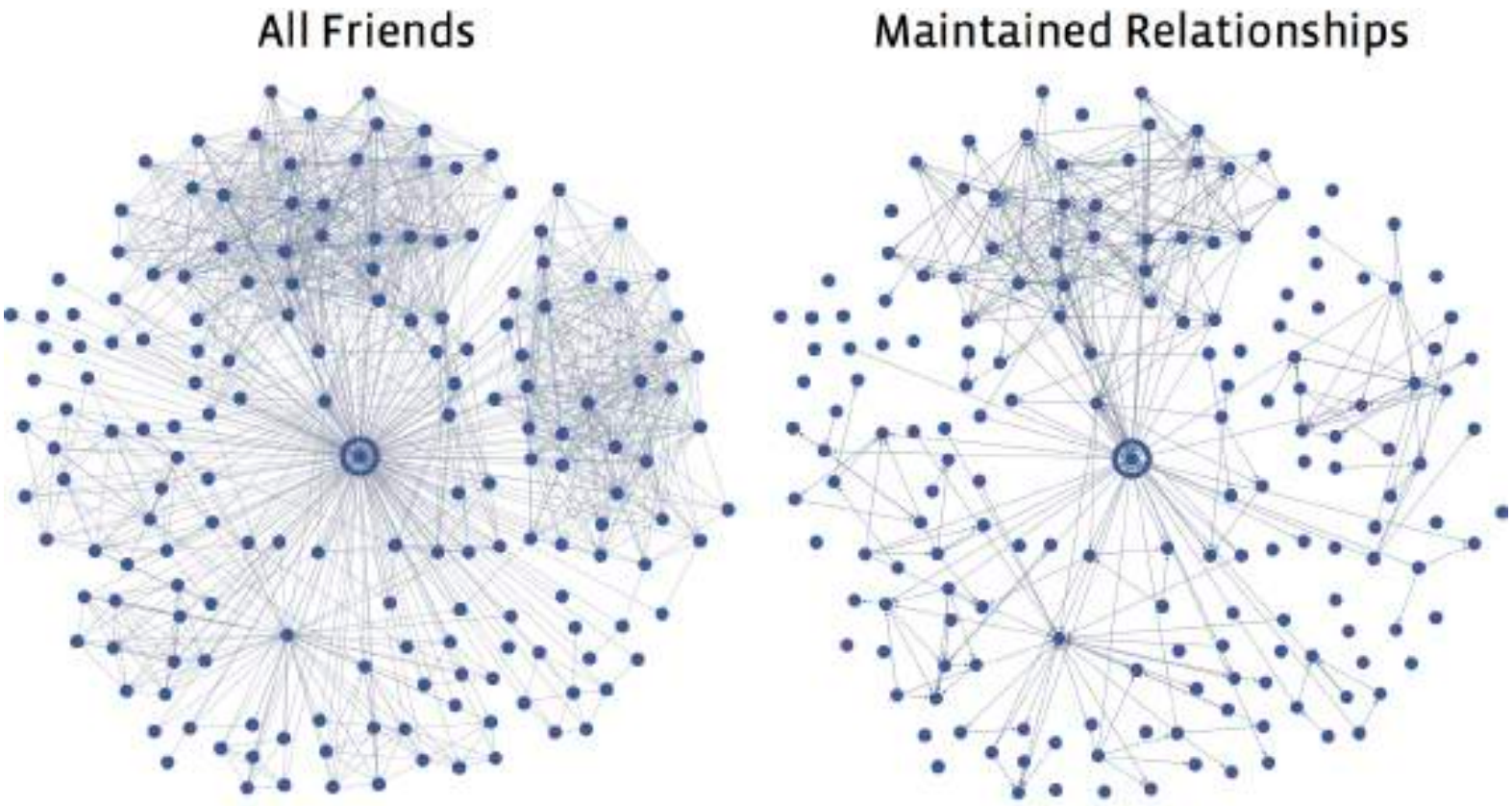
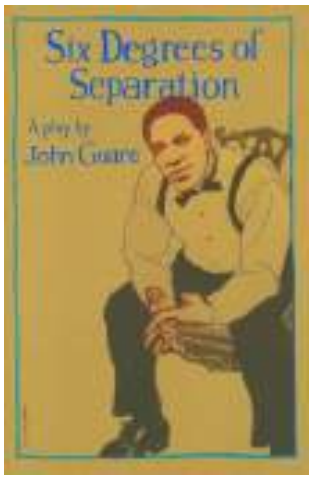


image from Cameron Marlow, Facebook

Six degrees of separation

"Any two people are on average separated no more that by six intermediate connections"

- Frigyes Karinthy, short story "Lancszemek" ("Chain-Links"), 1929.
- John Guare play (1991) and movie (1993), "Six Degrees of Separation"



Small world

An Experimental Study of the Small World Problem*

JEFFREY TRAVERS
Harvard University
AND
STANLEY MILGRAM
The City University of New York

Arbitrarily selected individuals (N=296) in Nebraska and Boston are asked to generate acquaintance chains to a target person in Massachusetts, employing "the small world method" (Milgram, 1967). Sixty-four chains reach the target person. Within this group the mean number of intermediaries between starters and targets is 5.2. Boston starting chains reach the target person with fewer intermediaries than those starting in Nebraska; subpopulations in the Nebraska group do not differ among themselves. The funneling of chains through sociometric "stars" is noted, with 48 per cent of the chains passing through three persons before reaching the target. Applications of the method to studies of large scale social structure are discussed.

- "The small-world problem". Stanley Milgram, 1967
- "An experimental study of the small world problem", Jeffrey Travers, Stanley Milgram, 1969

Stanley Milgram's 1967 experiment

HOW TO TAKE PART IN THIS STUDY

1. ADD YOUR NAME TO THE ROSTER AT THE BOTTOM OF THIS SHEET, so that the next person who receives this letter will know who it came from.

2. DETACH ONE POSTCARD. FILL IT OUT AND RETURN IT TO HARVARD UNIVERSITY. No stamp is needed. The postcard is very important. It allows us to keep track of the progress of the folder as it moves toward the target person.

3. IF YOU KNOW THE TARGET PERSON ON A PERSONAL BASIS, MAIL THIS FOLDER DIRECTLY TO HIM (HER). Do this only if you have previously met the target person and know each other on a first name basis.

4. IF YOU DO NOT KNOW THE TARGET PERSON ON A PERSONAL BASIS, DO NOT TRY TO CONTACT HIM DIRECTLY. INSTEAD, MAIL THIS FOLDER (POSTCARDS AND ALL) TO A PERSONAL ACQUAINTANCE WHO IS MORE LIKELY THAN YOU TO KNOW THE TARGET PERSON. You may send the folder

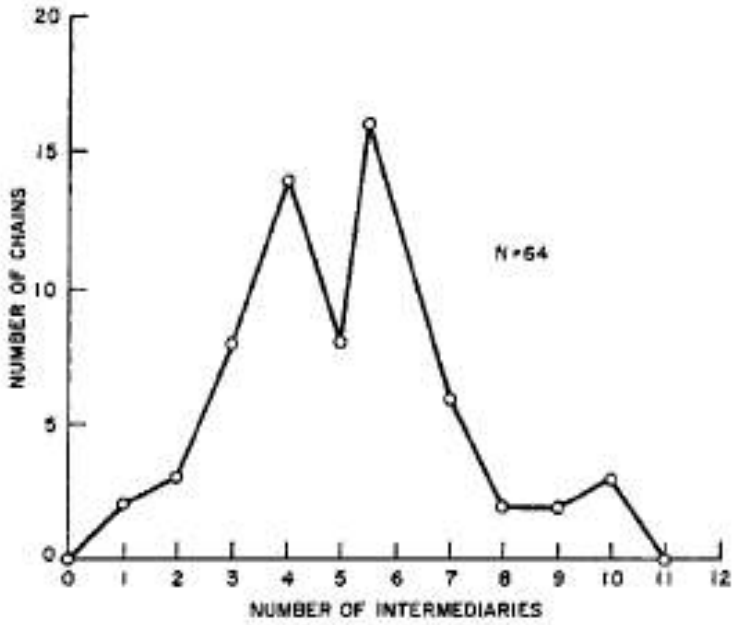
Stanley Milgram's 1967 experiment

- Starting persons:
  - 296 volunteers, 217 sent
  - 196 in Nebraska
  - 100 in Boston
- Target person - Boston stockbroker
- Information given: target name, address, occupation, place of employment, college, hometown

J. Travers, S. Milgram, 1969

Stanley Milgram's 1967 experiment

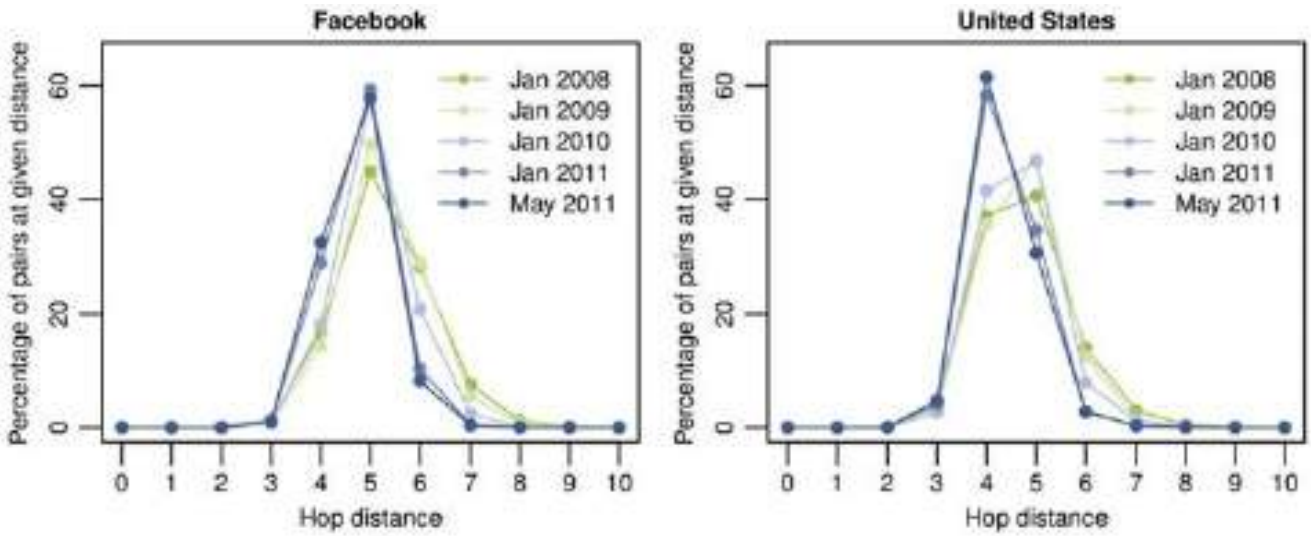
- Reached the target $N = 64(29\%)$
- Average chain length $\langle L \rangle = 5.2$
- Channels:
  - hometown $\langle L \rangle = 6.1$
  - business contacts $\langle L \rangle = 4.6$
  - from Boston $\langle L \rangle = 4.4$
  - from Nebraska $\langle L \rangle = 5.7$



J. Travers, S. Milgram, 1969

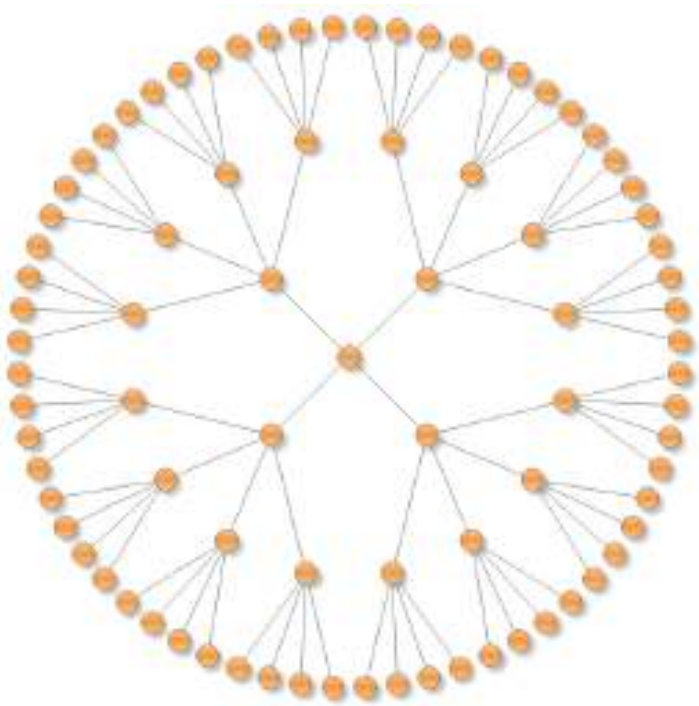
Small world

- Email graph:
  D. Watts (2001), 48,000 senders, $\langle L \rangle \approx 6$
- MSN Messenger graph:
  J. Lescovec et al (2007), 240mln users, $\langle L \rangle \approx 6.6$
- Facebook graph:
  L. Backstrom et al (2012), 721 mln users, $\langle L \rangle \approx 4.74$



figures from L.Backstrom, 2012

Simple model



An estimate: $z^d = N, d = \log N / \log z$

$N \approx 6.7$ bln, $z = 50$ friends, $d \approx 5.8$.

References

- Scale free networks. A.-L. Barabasi, E. Bonabeau, Scientific American 288, 50-59 (2003)
- Scale-Free Networks: A Decade and Beyond. A.-L. Barabasi, Science 325, 412-413 (2009)
- The Physics of Networks. Mark Newman, Physics Today, November 2008, pp. 33–38.

References

- The Small-World Problem. Stanley Milgram. Psychology Today, Vol 1, No 1, pp 61-67, 1967
- An Experimental Study of the Small World Problem. J. Travers and S. Milgram. . Sociometry, vol 32, No 4, pp 425-433, 1969
- Planetary-Scale Views on a Large Instant-Messaging Network. J. Leskovec and E. Horvitz. , Procs WWW 2008
- Four Degrees of Separation. L. Backstrom, P. Boldi, M. Rosa, J. Ugander, S. Vigna, WebSci '12 Procs. 4th ACM Web Science Conference, 2012 pp 33-42

NATIONAL RESEARCH UNIVERSITY

Descriptive Network Analysis

Social Network Analysis. MAGoLEGO course

Leonid Zhukov

lzhukov@hse.ru

www.leonidzhukov.net/hse/2018/sna

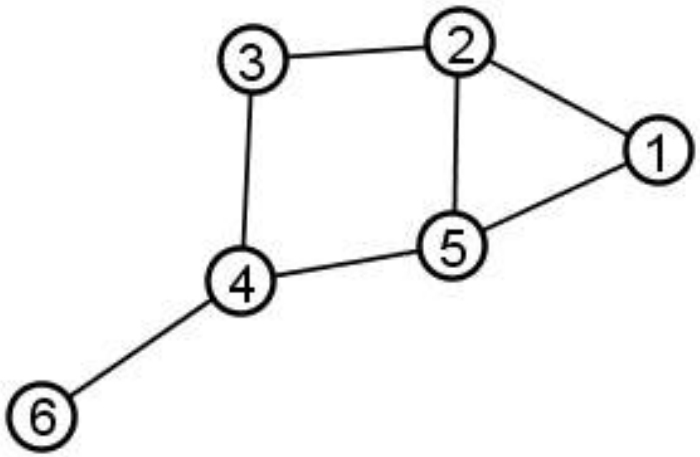
National Research University Higher School of Economics

School of Data Analysis and Artificial Intelligence, Department of Computer Science

April 13, 2018

Graphs

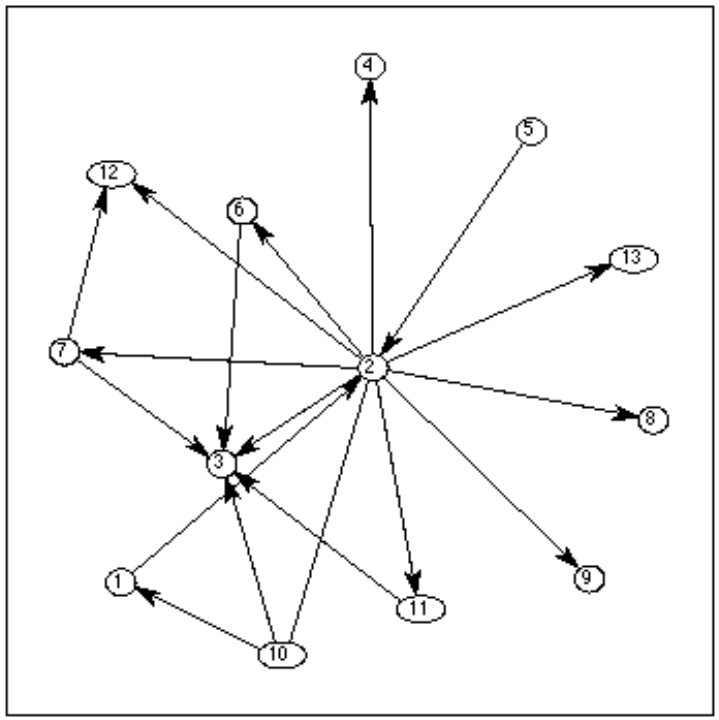
- A *graph* $G = (V, E)$ is an ordered pair of sets: a set of vertices V and a set edges E, where $n = |V|, m = |E|$
- An *edge* $e_{ij} = (v_i, v_j)$ is pair of vertices (ordered pair for directed graph)
- *Adjacency matrix* $A^{n \times n}$ is a matrix with nonzero element a_{ij} when there is an edge e_{ij}

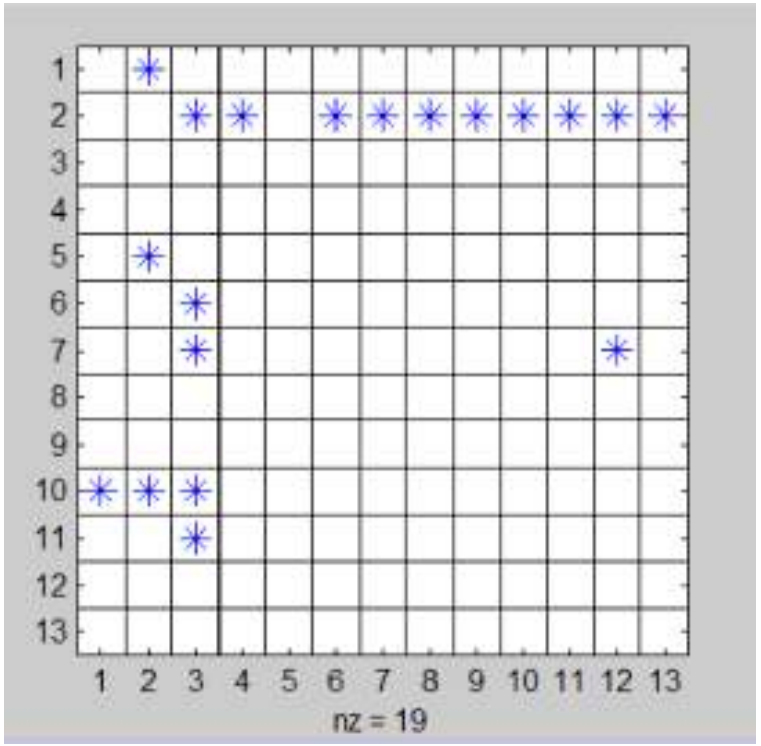


	[,1]	[,2]	[,3]	[,4]	[,5]	[,6]
[1,]	0	1	0	0	1	0
[2,]	1	0	1	0	1	0
[3,]	0	1	0	1	0	0
[4,]	0	0	1	0	1	1
[5,]	1	1	0	1	0	0
[6,]	0	0	0	1	0	0

Graphs and matrices

Graph $G(n, m)$, adjacency matrix $A_{ij}^{n \times n}$, edge $i \to j$, $m = nnz(A)$

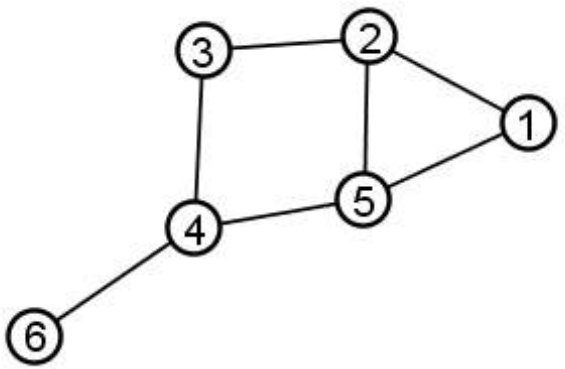




Node degree

- Two nodes/vertices are *adjacent* if they share a common edge
- An edge and a node on that edge are called *incident*.
- The *neighborhood* $\mathcal{N}(v)$ of a node v in a graph G is the set of nodes adjacent to v.
- The *degree* k_i of a nodes v_i is the total number of nodes adjacent to it, $k_i = |\mathcal{N}(v_i)|$
- Average node degree:

$$\langle k \rangle = \frac{1}{n} \sum_i k_i = \frac{2m}{n} = \frac{2|E|}{|V|}$$



Node degree

in directed networks:

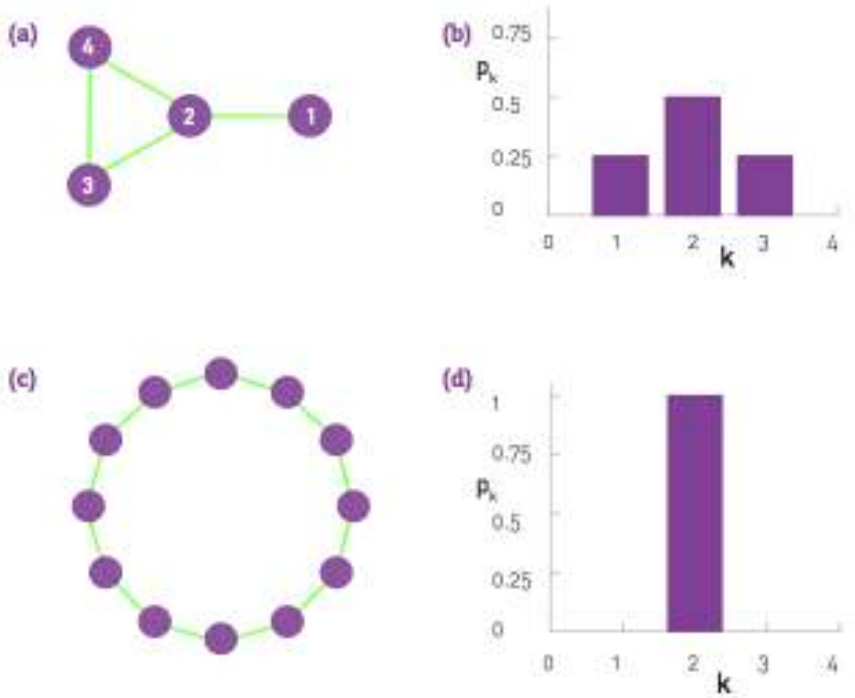
- k_i^{in} - incoming degree, number of edges/links pointing to node i
- k_i^{out} - outgoing degree, number of edges/links pointing from node i
- total node degree $k_i = k_i^{in} + k_i^{out}$
- Average in and out degrees are equal:

$$\langle k^{in} \rangle = \frac{1}{n} \sum_i k_i^{in} = \langle k^{out} \rangle = \frac{1}{n} \sum_i k_i^{out} = \frac{m}{n} = \frac{|E|}{|V|}$$

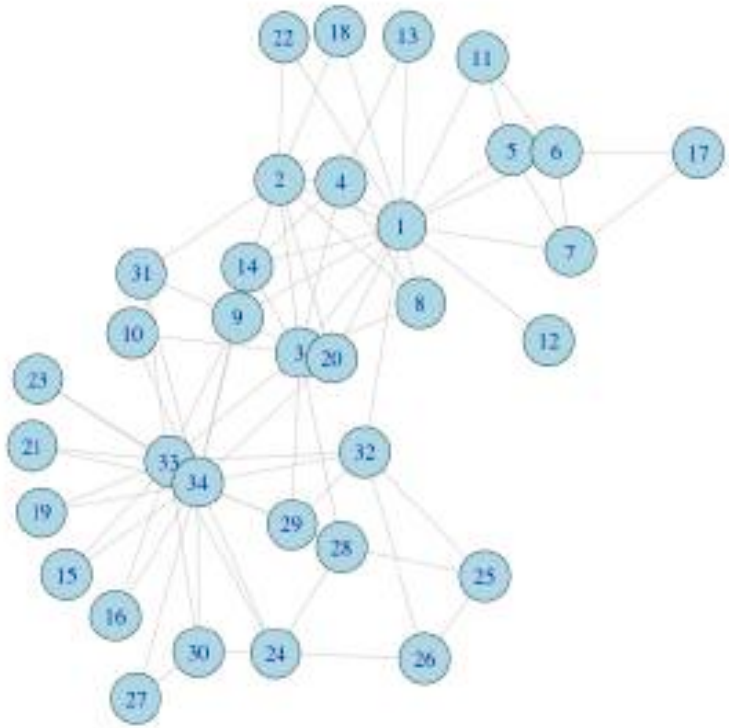
Degree distribution

- k_i - node degree, $k_i = 1, 2, \ldots k_{\max}$
- n_k - number of nodes with degree k, total nodes $n = \sum_k n_k$
- Degree distribution is a fraction of the nodes with degree k

$$P(k_i = k) = P(k) = P_k = \frac{n_k}{\sum_k n_k} = \frac{n_k}{n}$$

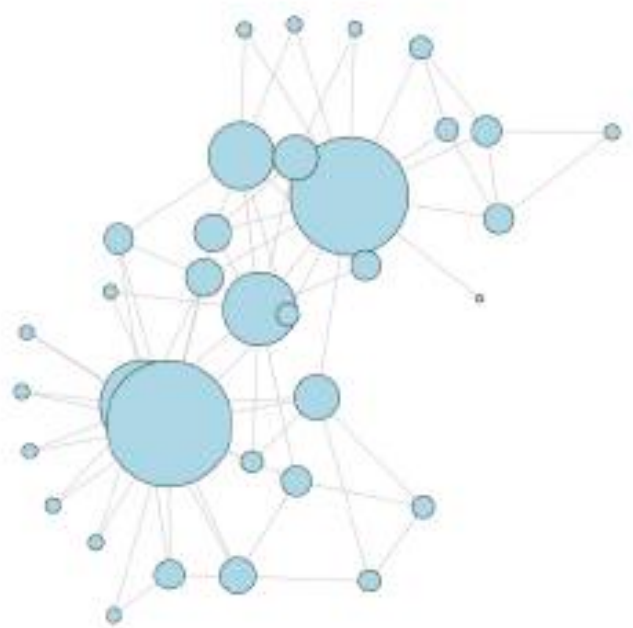


Node degrees



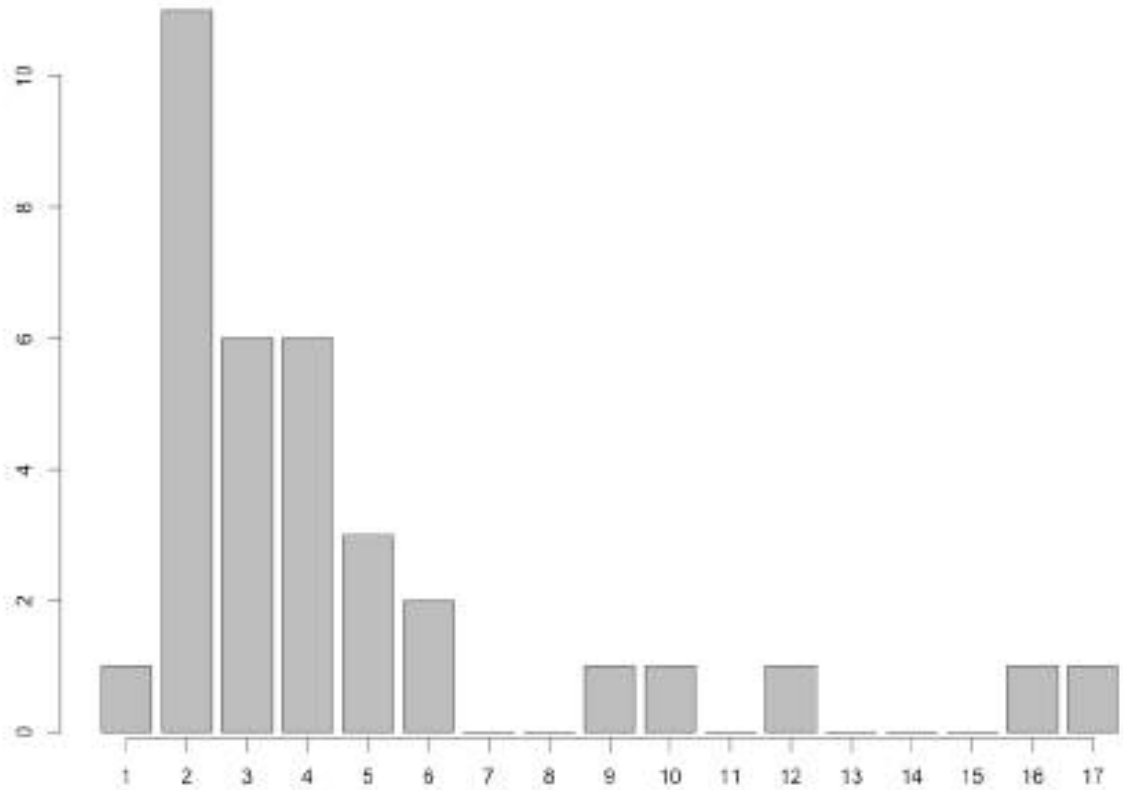
igraphdata: data(karate), igraph:plot()

Node degrees



igraphdata: data(karate), igprah:plot()

Node degree histogram

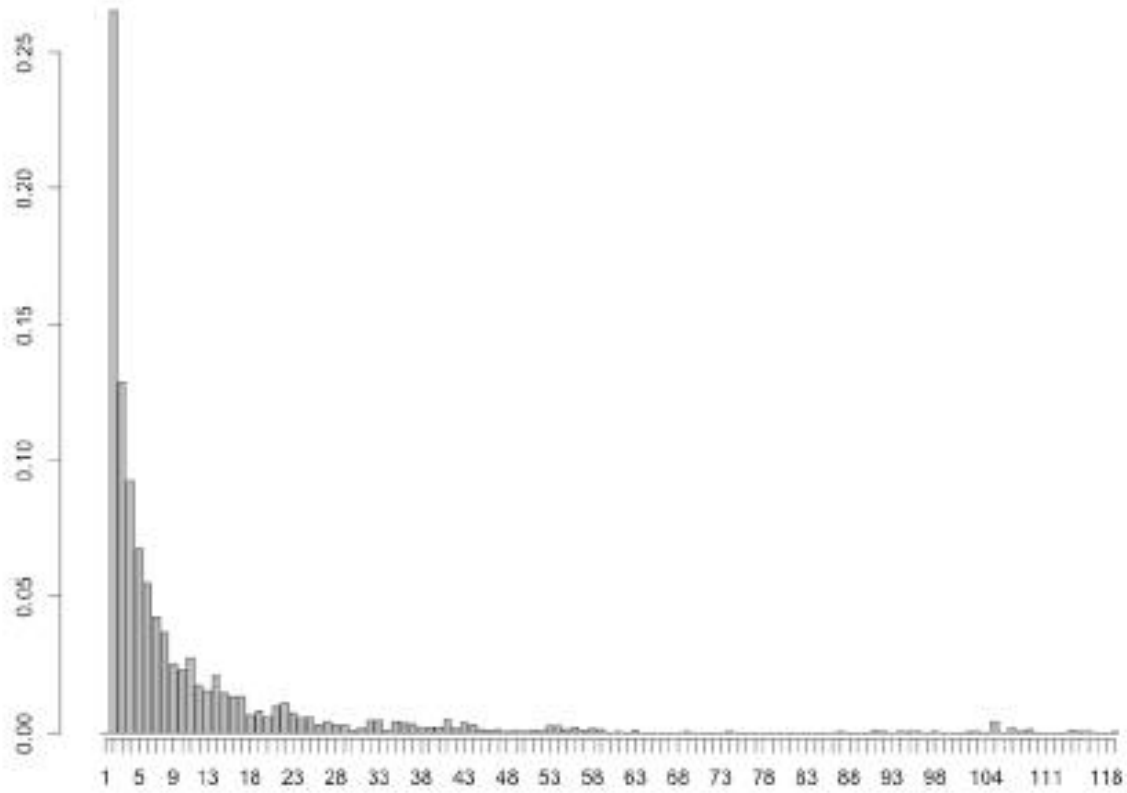


igraph: degree.distribution()

Degree distribution

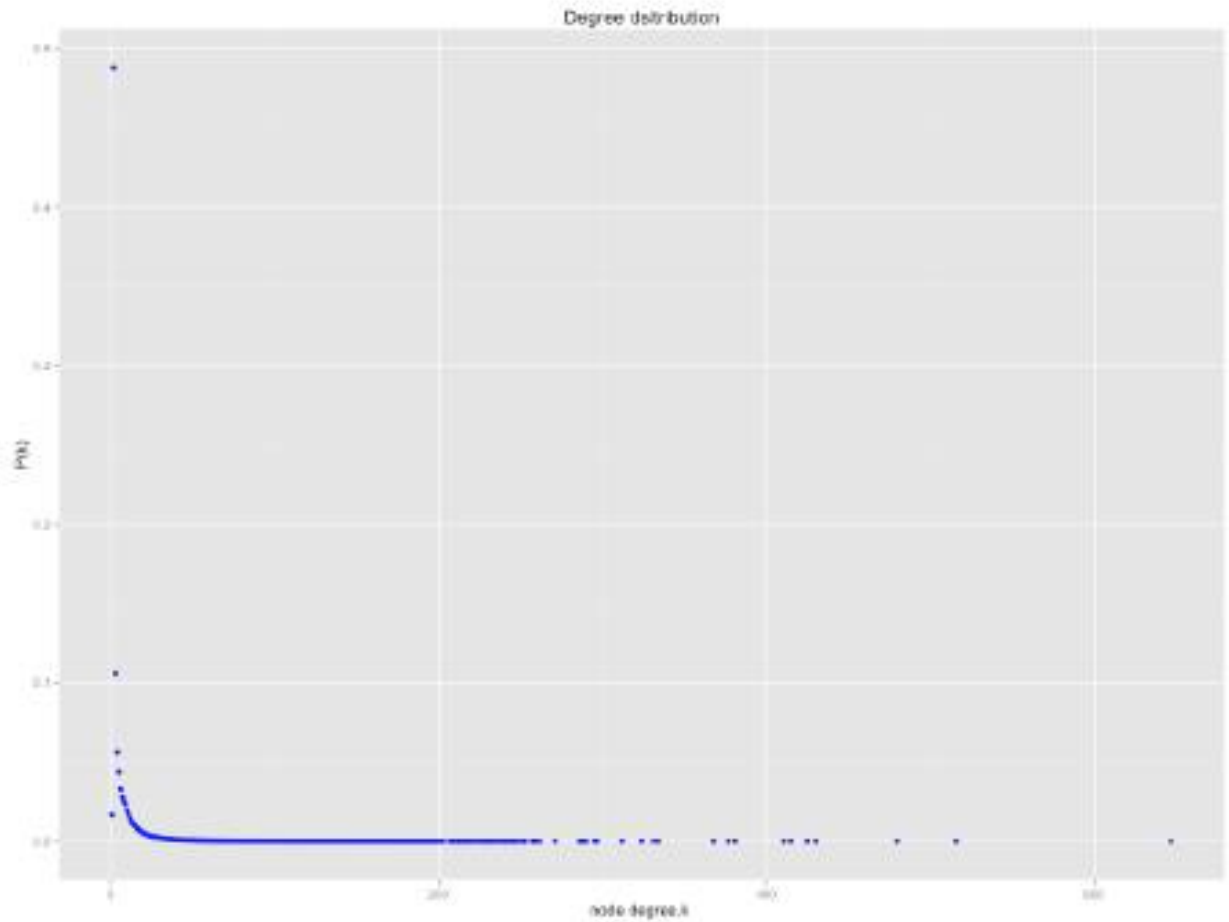
igraphdata: data(yeast)

Degree distribution



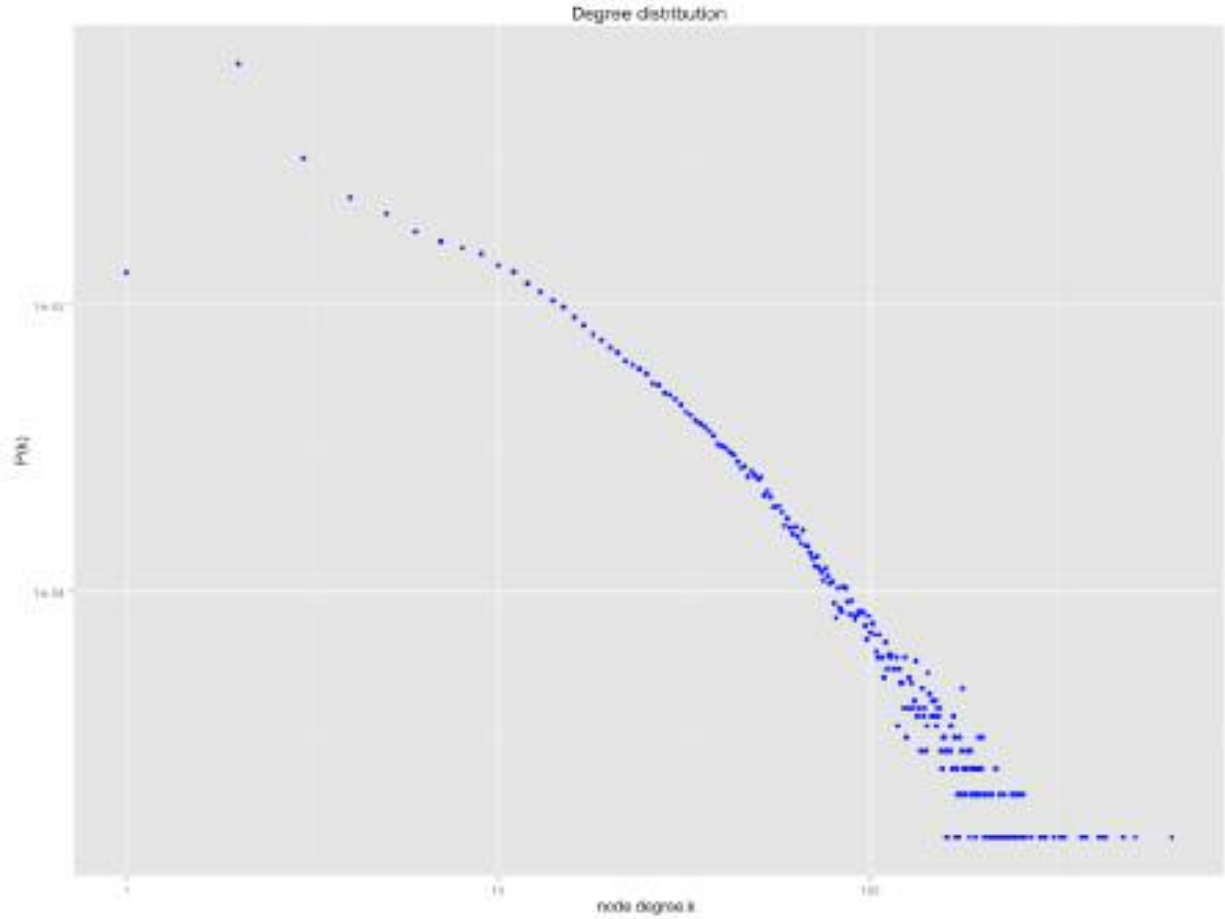
igraph: degree.distribution()

Power law degree distribution



Power law degree distribution

log-log scale



Discrete power law distribution

- Power law distribution

$$P(k) = Ck^{-\gamma} = \frac{1}{k^{\gamma}}C$$

- Log-log coordinates

$$\log P(k) = -\gamma \log k + \log C$$

$$y = -\gamma x + b$$

Distribution parameter estimation

- Maximum likelihood estimation of parameter γ:

$$\gamma = 1 + n \left[\sum_{i=1}^{n} \ln \frac{k_i}{k_{\min}} \right]^{-1}$$

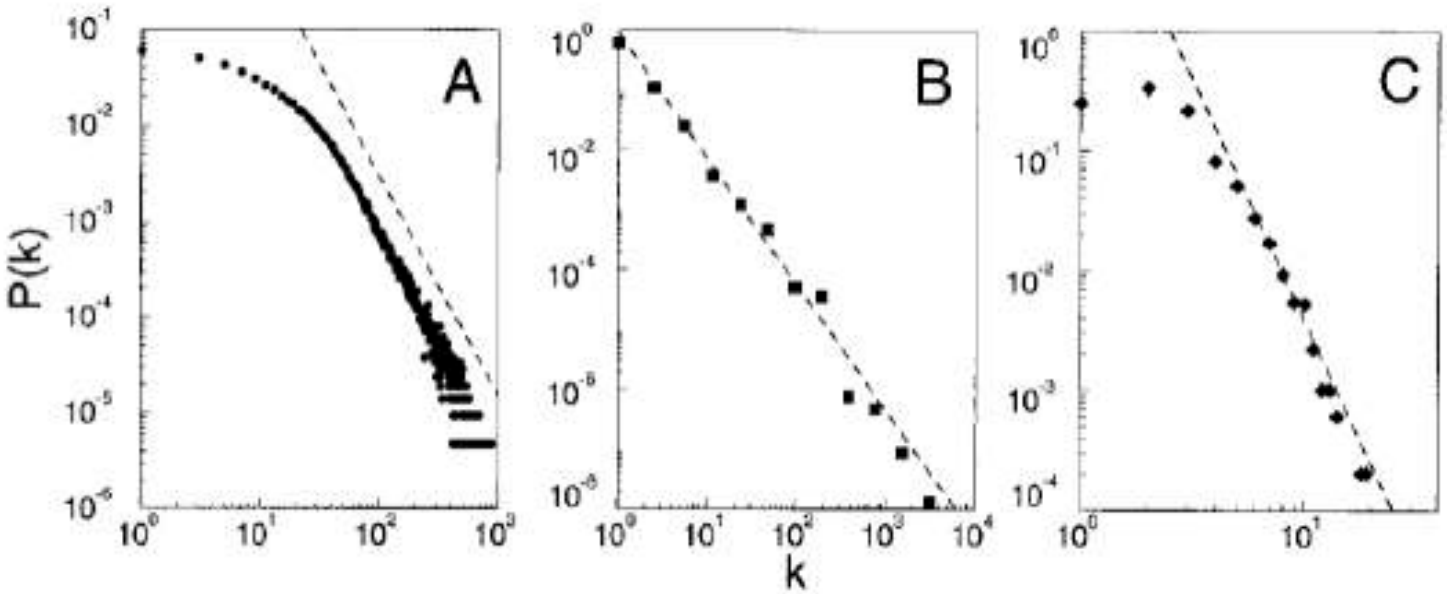
- error estimate

$$\sigma = \sqrt{n} \left[\sum_{i=1}^{n} \ln \frac{k_i}{k_{\min}} \right]^{-1} = \frac{\gamma - 1}{\sqrt{n}}$$

- Optimal value of k_{min} can be found using Kolmogorov-Smirnov test for optimal distribution fitting

igraph:power.law.fit()

Power law networks



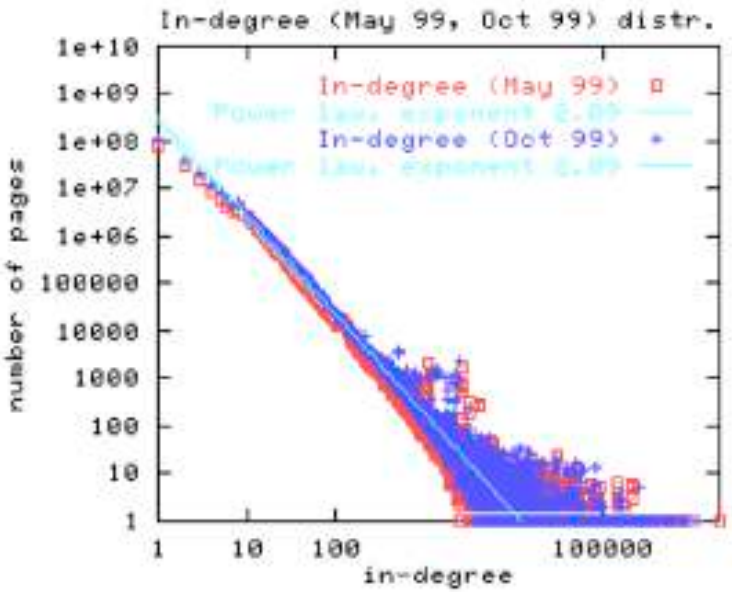
Actor collaboration graph, N=212,250 nodes, $\langle k \rangle = 28.8, \gamma = 2.3$

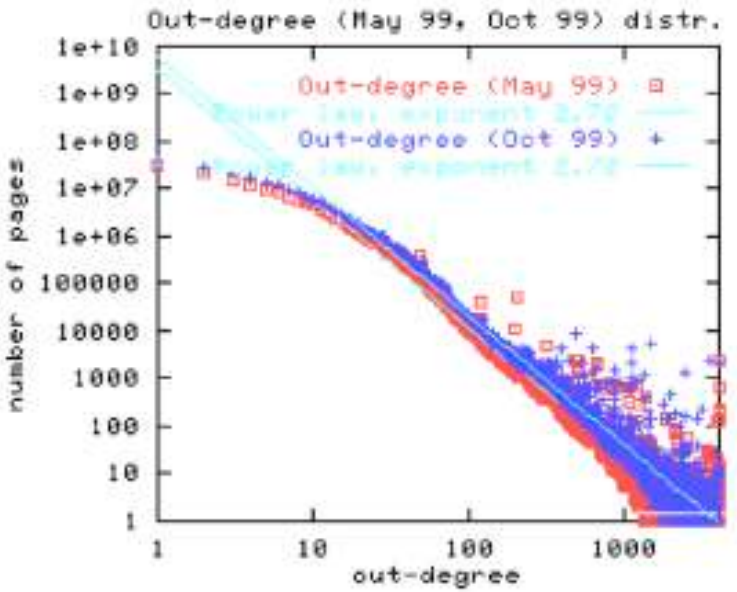
WWW, N = 325,729 nodes, $\langle k \rangle = 5.6, \gamma = 2.1$

Power grid data, N = 4941 nodes, $\langle k \rangle = 5.5, \gamma = 4$

Barabasi et.al, 1999

Power law networks



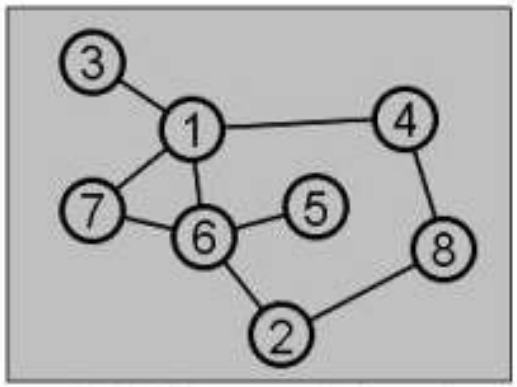


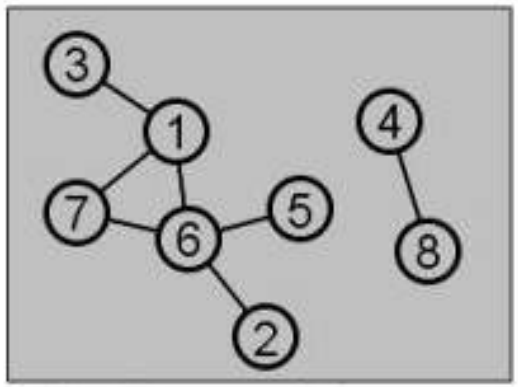
In- and out- degrees of WWW crawl 1999

Broder et.al, 1999

Graph connectivity

- A *path* from v_i to v_j is a sequence of edges that joins two vertices. (It also ordered list of vertices such that that there is an edge to the next vertex on the list)
- A graph is *connected* if there a paths between any two vertices.
- *Connected component* is a maximal connected subgraph of *G*





Graph connectivity

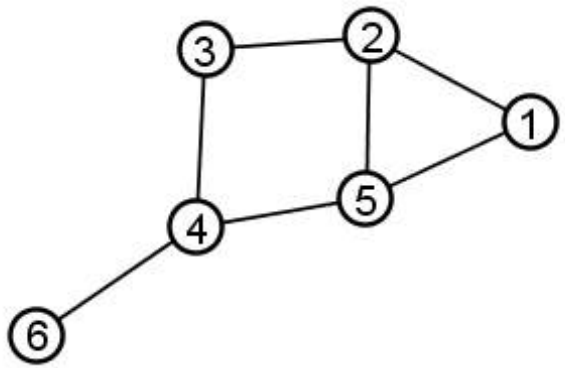
Connected components: 92

Component sizes: 2375 7 7 7 6 5 5 5 5 5 5 4 4 4 4

Graph connectivity

- The *distance* $d_G(v_i, v_j)$ between two vertices is the number of edges in the shortest path from v_i to v_j
- Graph *diameter* is the largest shortest path: $D = \max_{i,j} d_G(v_i, v_j)$
- Average path length:

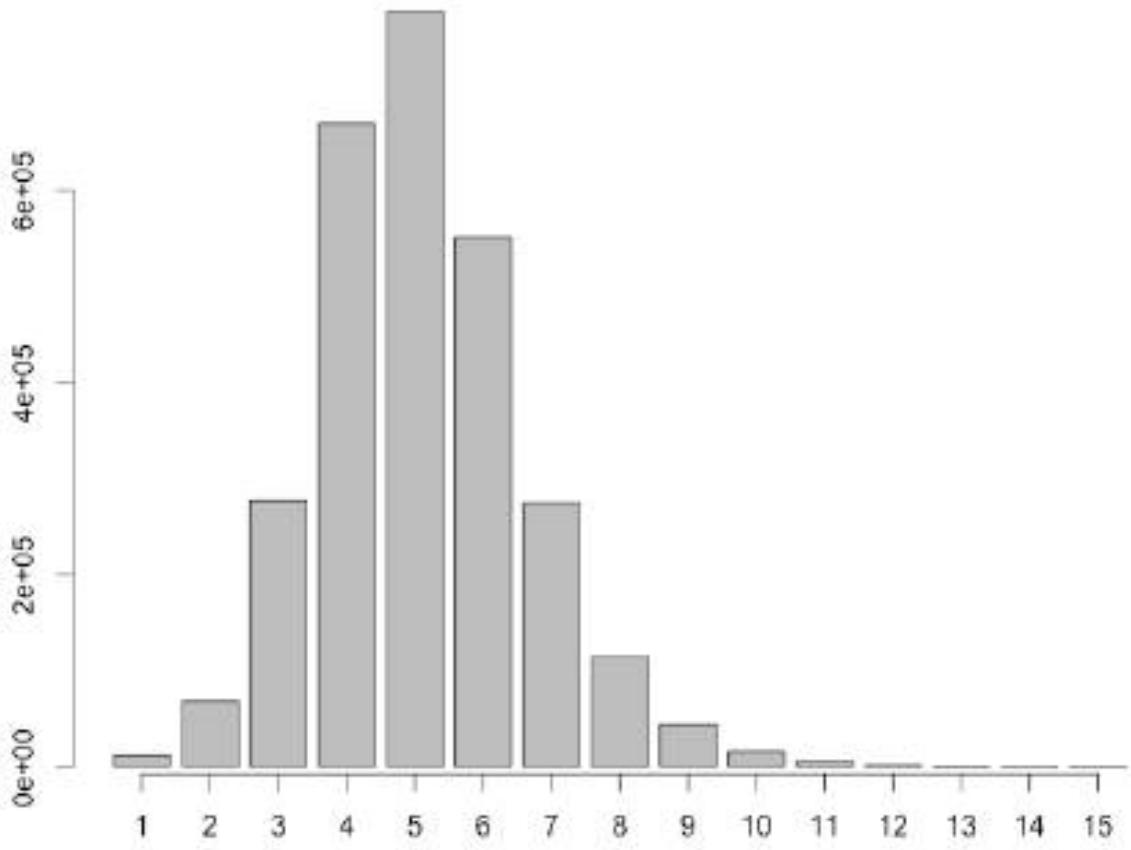
$$\langle L \rangle = \frac{1}{n(n-1)} \sum_{i \neq j} d_G(v_i, v_j)$$



igraph: shortest.paths(), diameter(), average.path.length(), path.length.hist()

Graph average path length

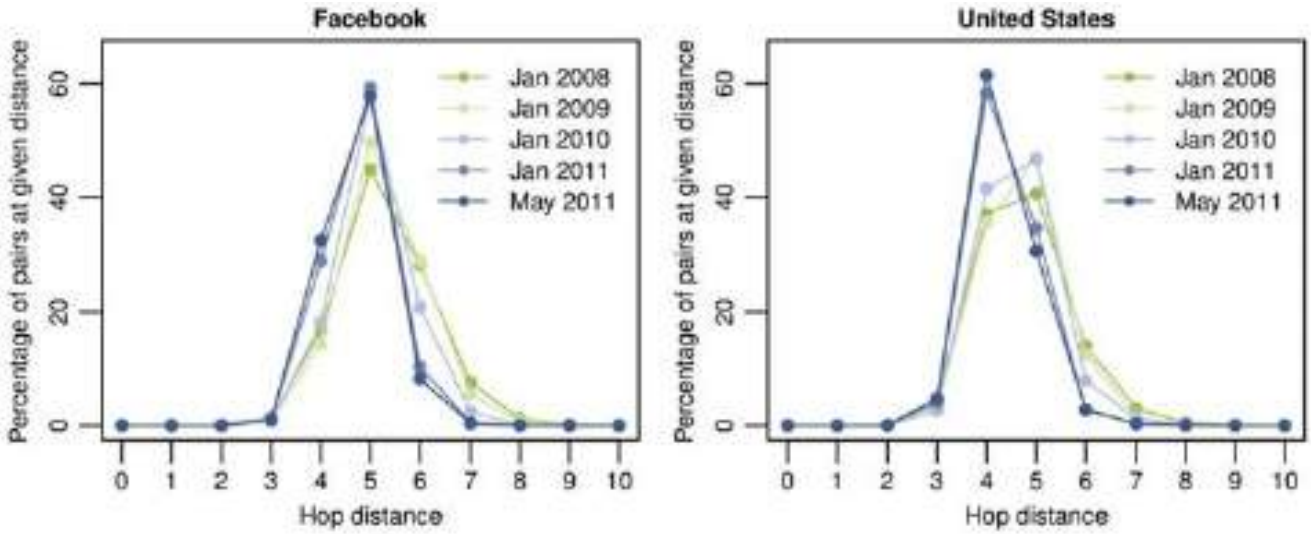
"Yeast" graph, $n = 2617$, $m = 11855$



Diameter $D = 15$, average path length $\langle L \rangle = 5.1$

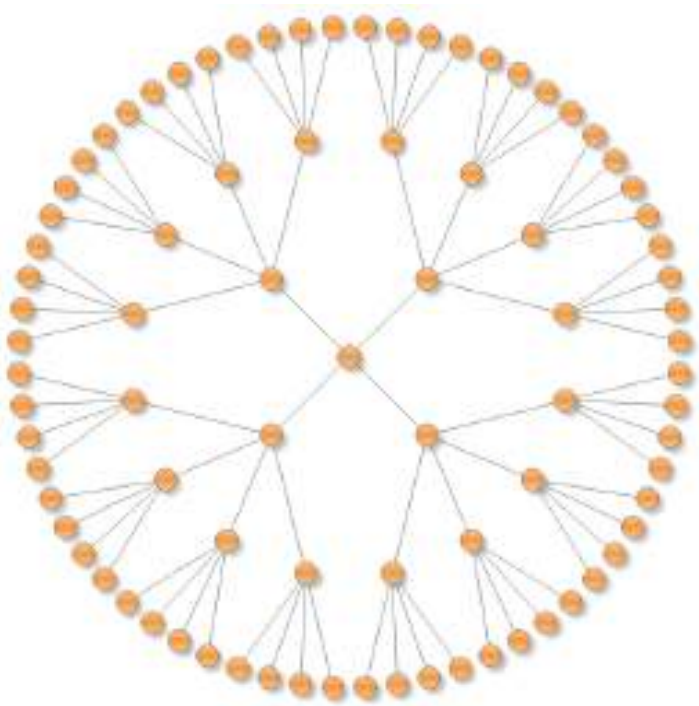
Small world

- Email graph:
  D. Watts (2001), 48,000 senders, $\langle L \rangle \approx 6$
- MSN Messenger graph:
  J. Lescovec et al (2007), 240mln users, $\langle L \rangle \approx 6.6$
- Facebook graph:
  L. Backstrom et al (2012), 721 mln users, $\langle L \rangle \approx 4.74$



figures from L.Backstrom, 2012

Simple model



An estimate: $z^d = N, d = \log N / \log z$

$N \approx 6.7$ bln, $z = 50$ friends, $d \approx 5.8$.

Triads, transitivity and clustering

Facebook friendship

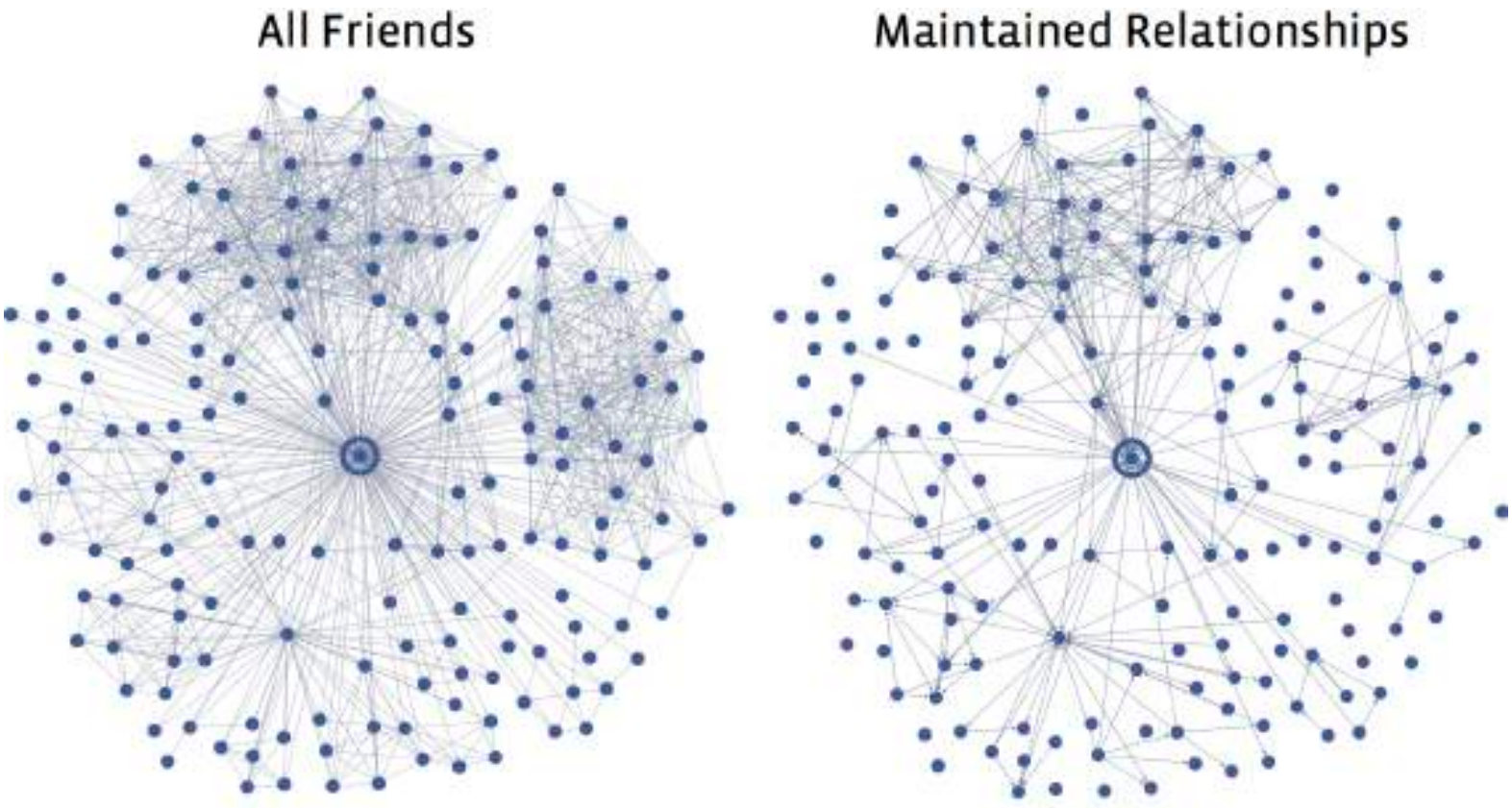


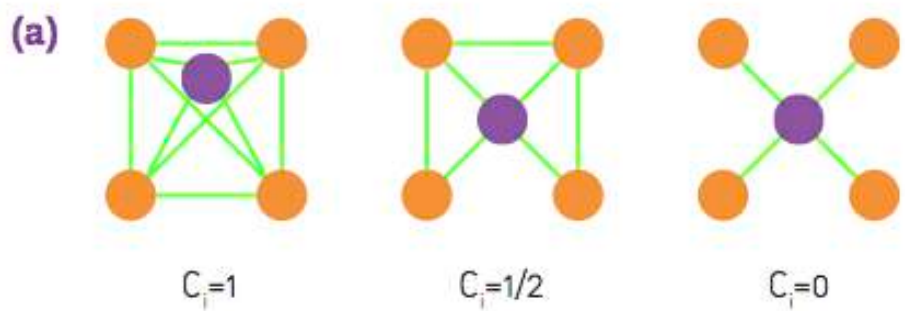
image from Cameron Marlow, Facebook

Clustering coefficient

How neighbors of a given node connected to each other

- *Local clustering coefficient* (per vertex):

$$C_i = \frac{\text{number of links in } \mathcal{N}_i}{k_i(k_i-1)/2}$$



- Average clustering coefficient:

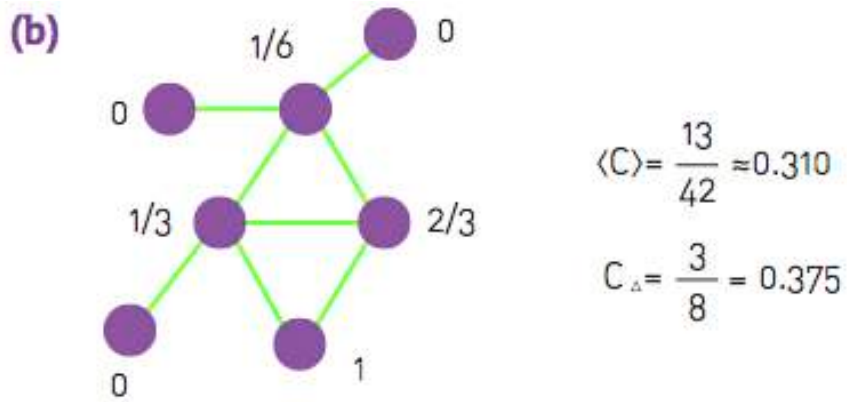
$$\bar{C} = \frac{1}{n}\sum_{i=1}^{n} C_i$$

igraph:transitivity(type="local")

Clustering coefficient

- *Global clustering coefficient*:

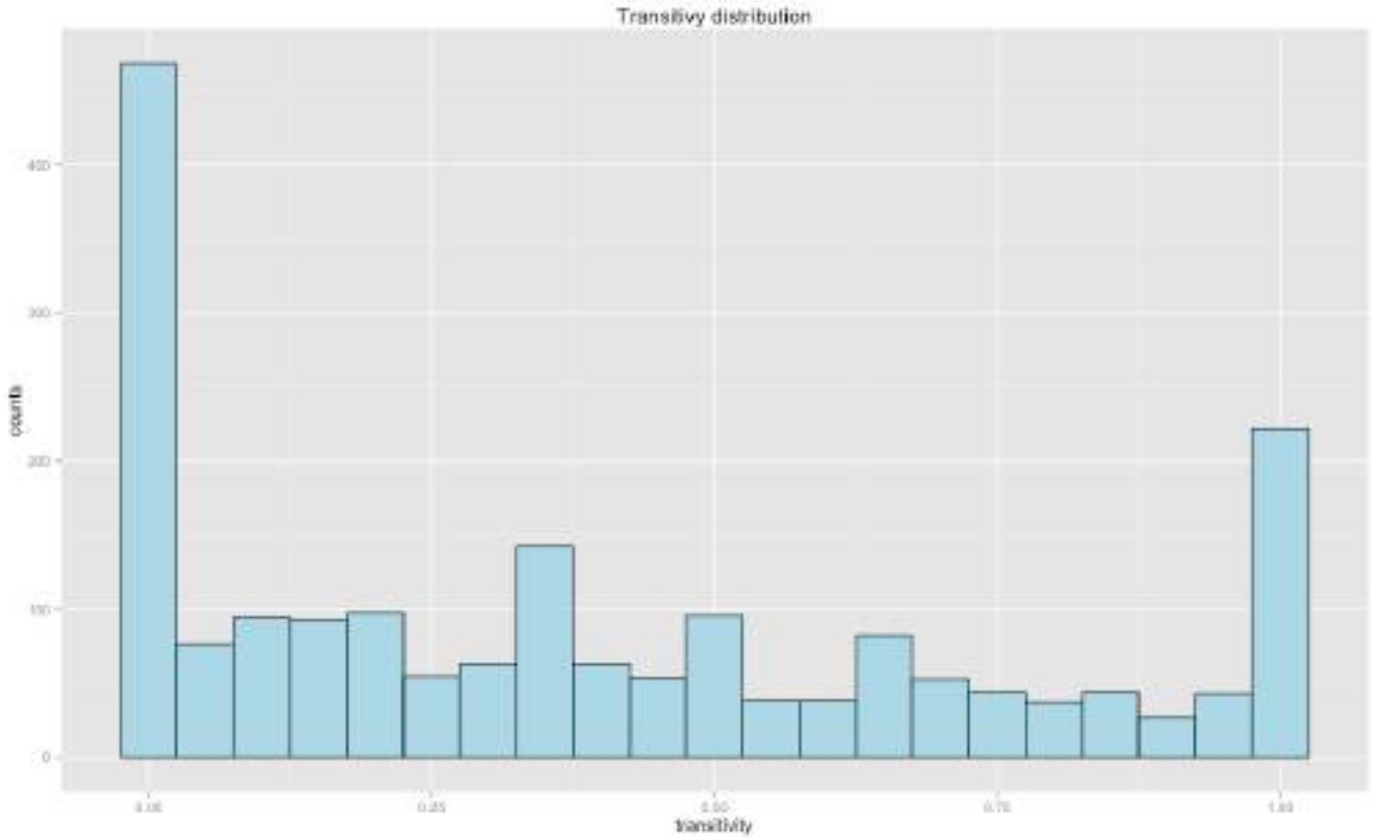
$$C = \frac{3 \times \text{number of triangles}}{\text{number of connected triplets of vertices}}$$



igraph:transitivity(type="global")

Clustering coefficient

Yeast graph



Global clustering coefficient: $C = 0.468$

Statistical properties

- Power-law degree distribution
- Small average path length
- High clustering coefficient (transitivity)
- Gigantic connected component

References

- Statistical Analysis of Network Data with R. Eric Kolaczyk, Gabor Csardi. Springer, 2014
- Social Network Analysis: Methods and Applications. S. Wasserman, K. Faust, Cambridge University Press, 1994
- Networks: An Introduction. Mark Newman. Oxford University Press, 2010.
- Power laws, Pareto distributions and Zipf's law, M. E. J. Newman, Contemporary Physics, pages 323–351, 2005.
- Power-Law Distribution in Empirical Data, A. Clauset, C.R. Shalizi, M.E.J. Newman, SIAM Review, Vol 51, No 4, pp. 661-703, 2009.

NATIONAL RESEARCH
UNIVERSITY

Mathematical models of networks

Social Network Analysis. MAGoLEGO course

Leonid Zhukov

lzhukov@hse.ru
www.leonidzhukov.net/hse/2018/sna

National Research University Higher School of Economics
School of Data Analysis and Artificial Intelligence, Department of Computer Science

April 20, 2018

Network models

network features:

- Power-law (heavy-tailed) degree distribution
- Small average distance (graph diameter)
- Large clustering coefficient (transitivity)
- Giant connected component

Generative models:

- Random graph model (Erdos & Renyi, 1959)
- Preferential attachement model (Barabasi & Albert, 1999)
- Small world model (Watts & Strogatz, 1998)

Random graph model

Graph $G\{E, V\}$, nodes $n = |V|$, edges $m = |E|$
Erdos and Renyi, 1959.

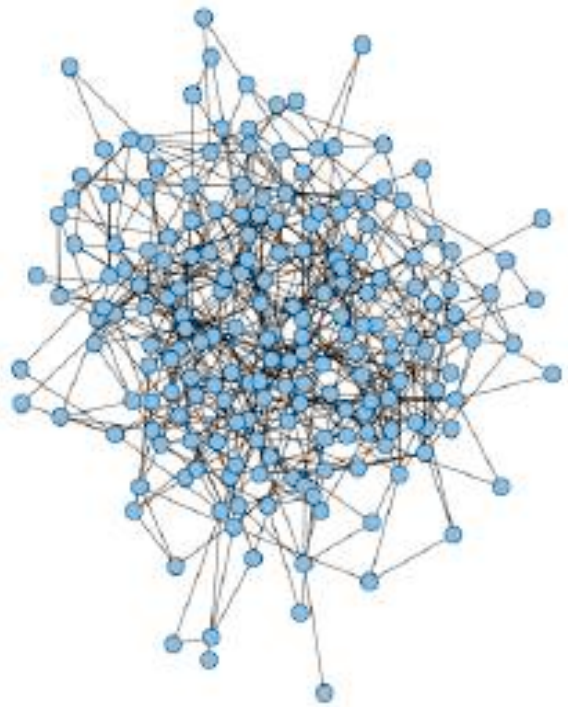
- $G_{n,p}$ - each pair out of $N = \frac{n(n-1)}{2}$ is connected with probability p,
  number of edges m - random number

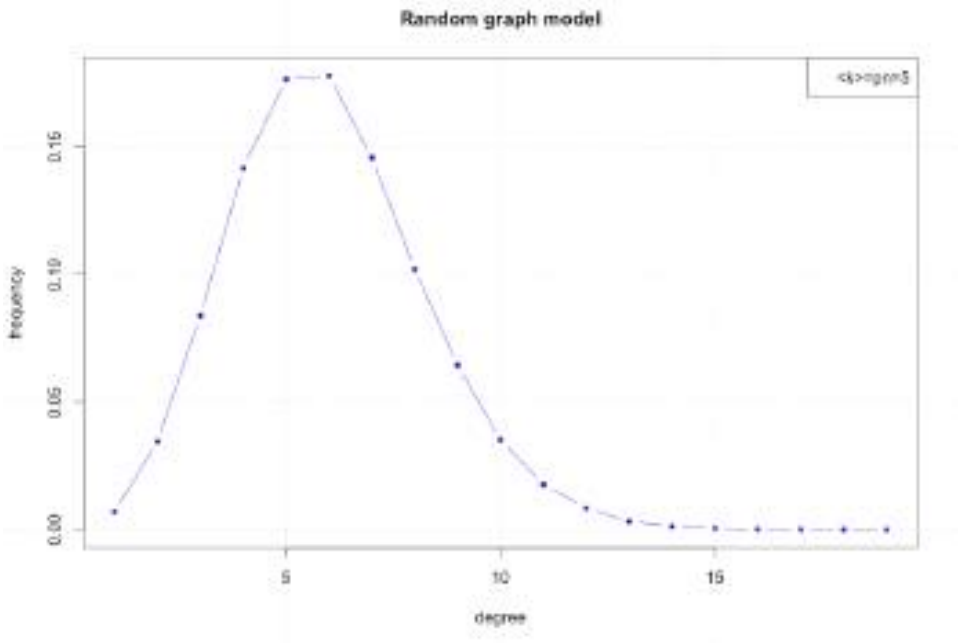
$$\langle m \rangle = p\frac{n(n-1)}{2}$$

$$\langle k \rangle = \frac{1}{n}\sum_i k_i = \frac{2\langle m \rangle}{n} = p\,(n-1) \approx pn$$

$$\rho = \frac{\langle m \rangle}{n(n-1)/2} = p$$

Random graph





Node degree distribution (Poisson distribution):

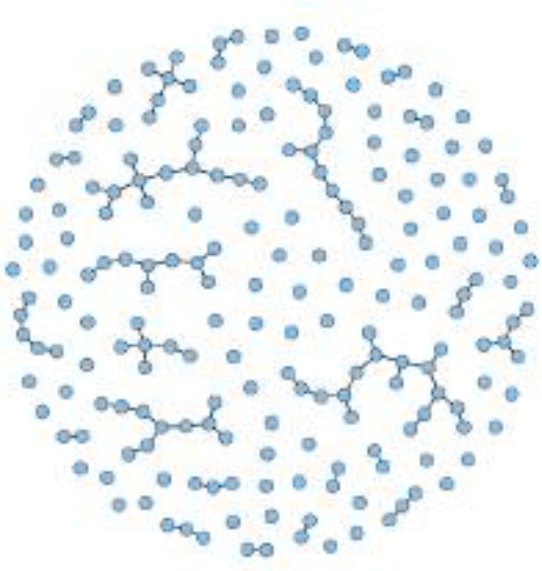
$$P(k_i = k) = \frac{\lambda^k e^{-\lambda}}{k!}, \quad \lambda = pn = \langle k \rangle$$

Phase transition

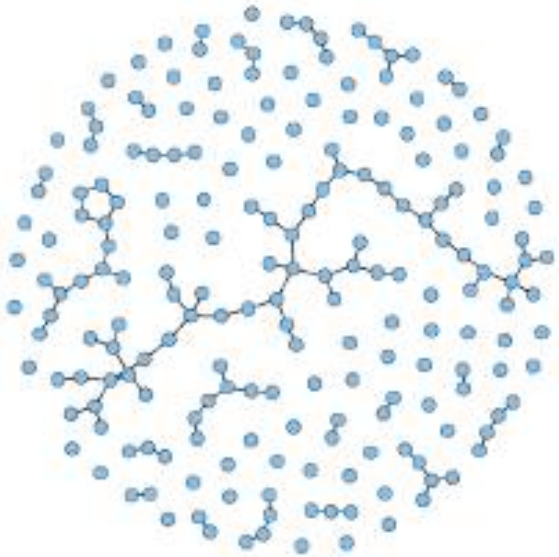
Consider $G_{n,p}$ as a function of p

- $p = 0$, empty graph
- $p = 1$, complete (full) graph
- There are exist critical p_c, structural changes from $p < p_c$ to $p > p_c$
- Gigantic connected component appears at $p > p_c$

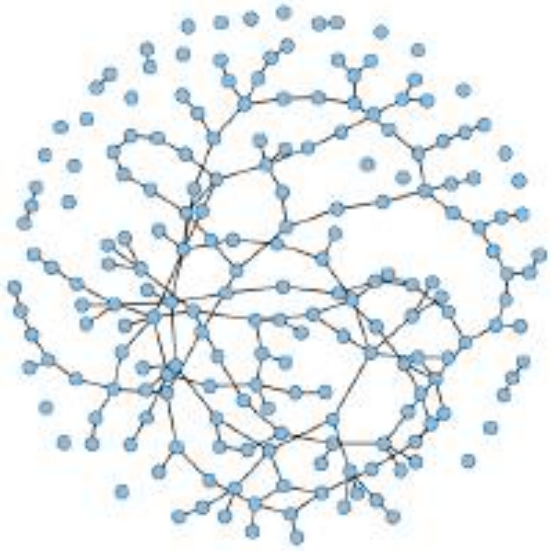
Random graph model



$p < p_c$



$p = p_c$



$p > p_c$

igraph:erdos.renyi.game()

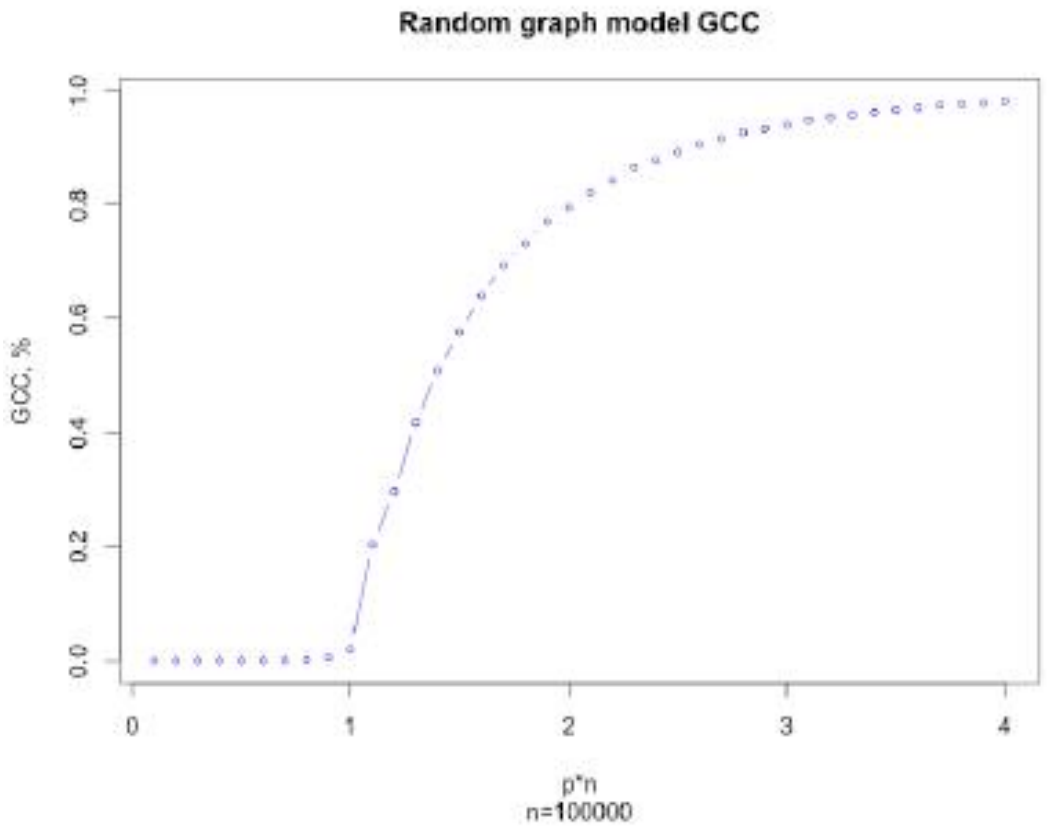
Phase transition

Graph $G(n, p)$, for $n \to \infty$, critical value $p_c = 1/n$

- when $p < p_c$, $(\langle k \rangle < 1)$ there is no components with more than $O(\ln n)$ nodes, largest component is a tree
- when $p = p_c$, $(\langle k \rangle = 1)$ the largest component has $O(n^{2/3})$ nodes
- when $p > p_c$, $(\langle k \rangle > 1)$ gigantic component has all $O(n)$ nodes

Critical value: $\langle k \rangle = p_c n = 1$- on average one neighbor for a node

Connected component



$$\langle k \rangle = pn$$

Clustering coefficient

- Clustering coefficient

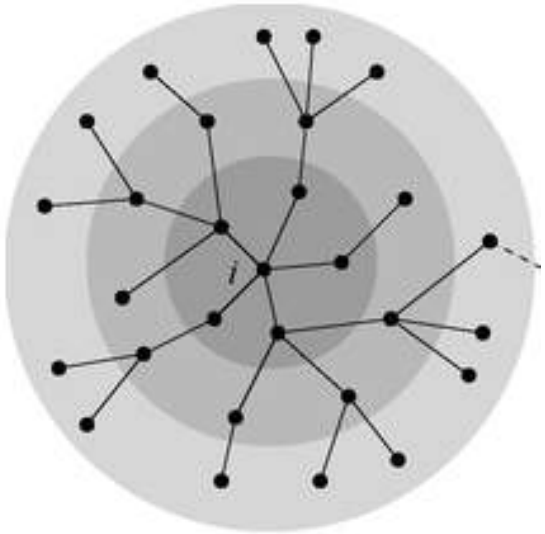
$$C(k) = \frac{\text{\#of links between NN}}{\text{\#max number of links NN}} = \frac{pk(k-1)/2}{k(k-1)/2} = p$$

$$C = p = \frac{\langle k \rangle}{n}$$

- when $n \to \infty$, $C \to 0$

Graph diameter

- $G(n, p)$ is locally tree-like (GCC) (no loops; low clustering coefficient)



- on average, the number of nodes d steps away from a node $\langle k \rangle^d$
- in GCC, around p_c , $\langle k \rangle^d \sim n$,

$$d \sim \frac{\ln n}{\ln \langle k \rangle}$$

Random graph model

- Node degree distribution function - Poisson:

$$P(k) = \frac{\lambda^k e^{-\lambda}}{k!}, \quad \lambda = pn = \langle k \rangle$$

- Average path length:

$$\langle L \rangle \sim \log(N) / \log\langle k \rangle$$

- Clustering coefficient:

$$C = \frac{\langle k \rangle}{n}$$

Preferential attachment model

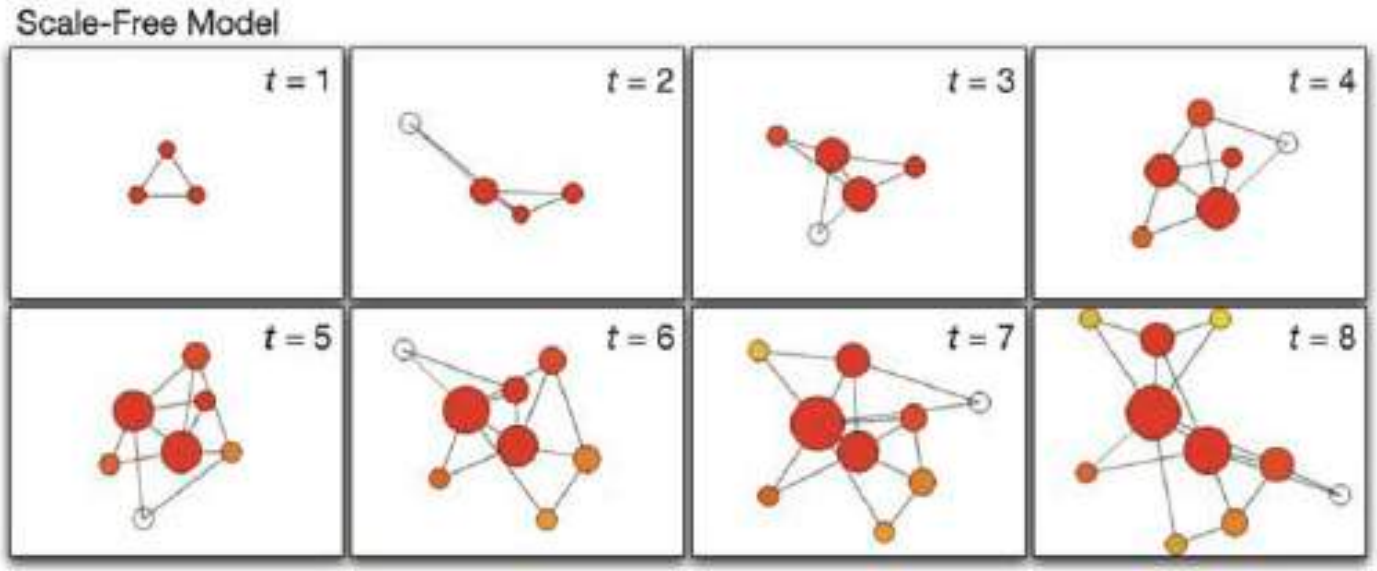
Barabasi and Albert, 1999
Dynamical growth model

- $t = 0$, n_0 nodes
- growth: on every step add a node with m_0 edges ($m_0 \leq n_0$), $k_i(i) = m_0$
- Preferential attachment: probability of linking to existing node is proportional to the node degree k_i

$$\Pi(k_i) = \frac{k_i}{\sum_i k_i} = \frac{k_i}{2m_0 t}$$

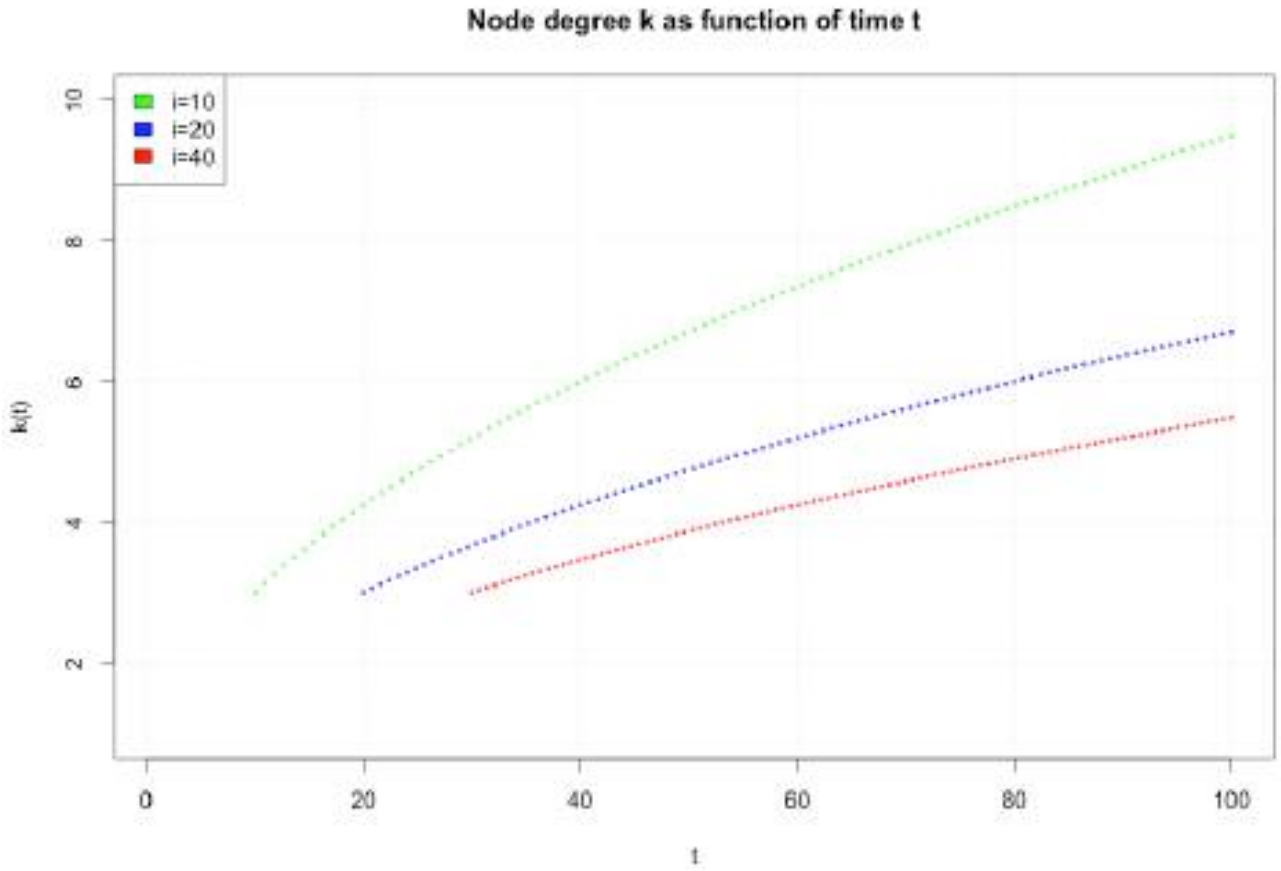
after t steps: $n_0 + t$ nodes, $m_0 t$ edges

Preferential attachment model



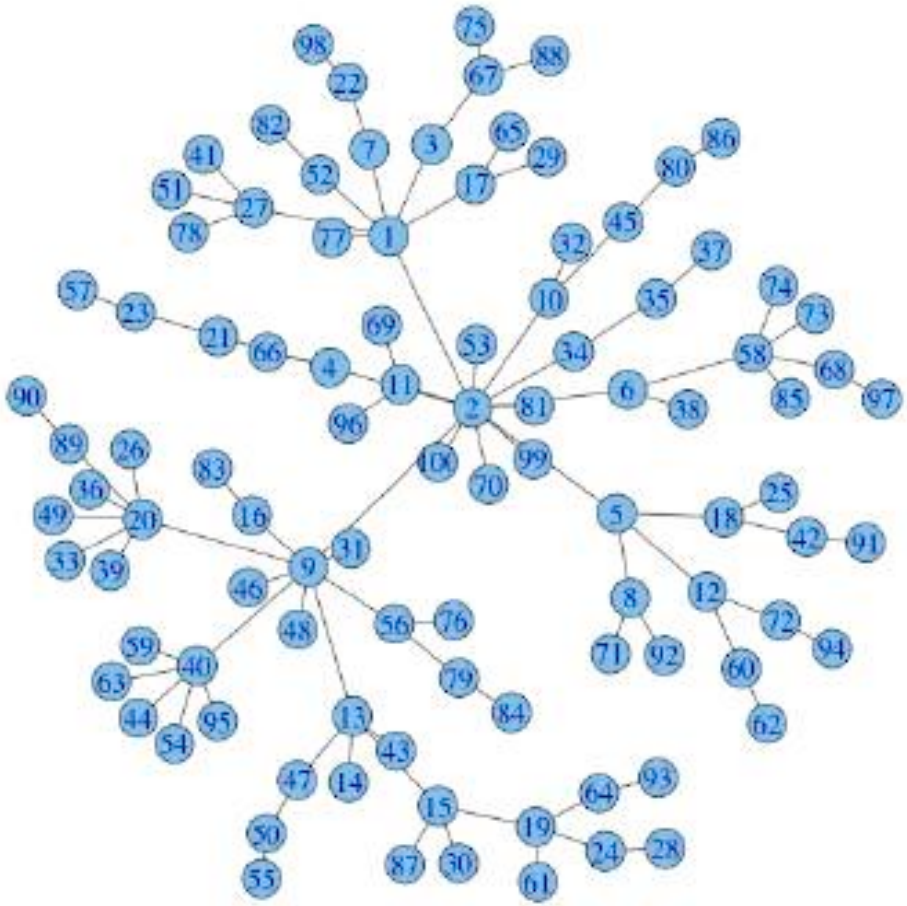
Barabasi, 1999

Preferential attachement

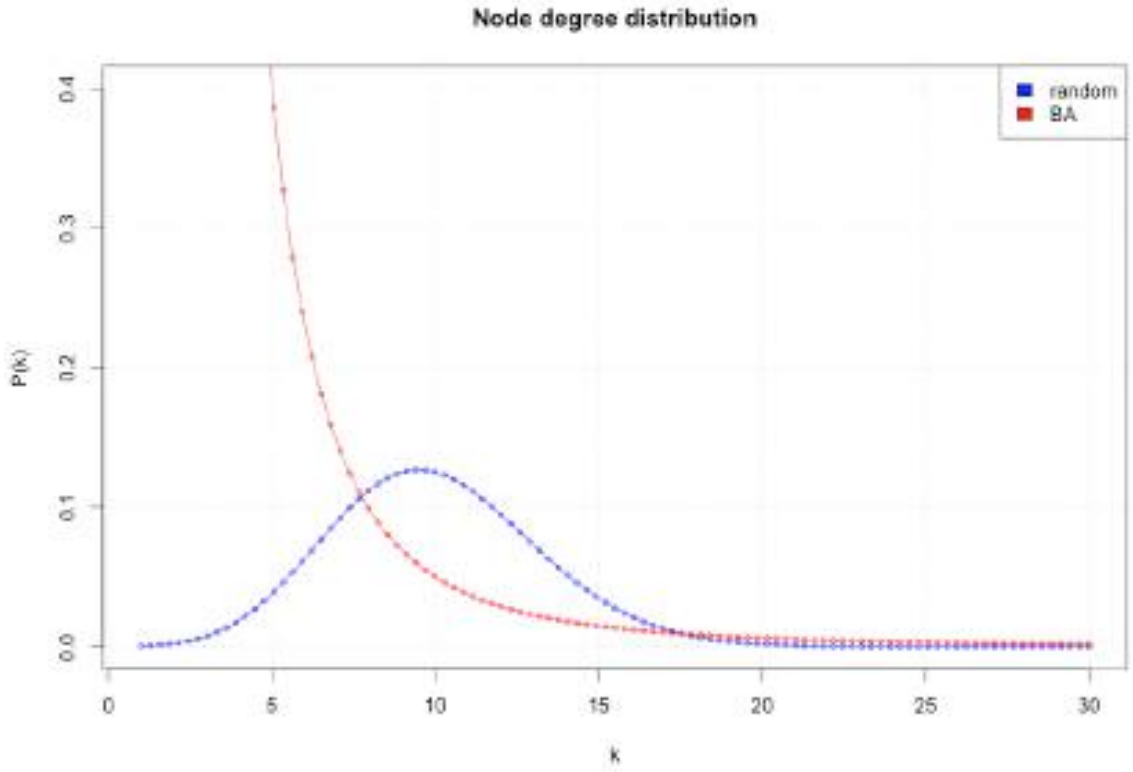


$$k_i(t) = m_0 \left(\frac{t}{i}\right)^{1/2}$$

Preferential attachement



Preferential attachement

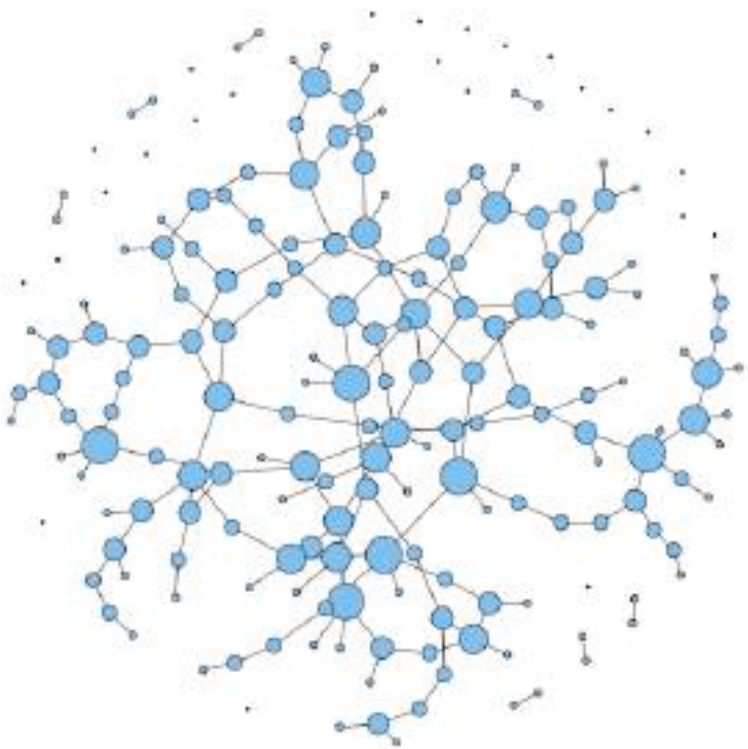


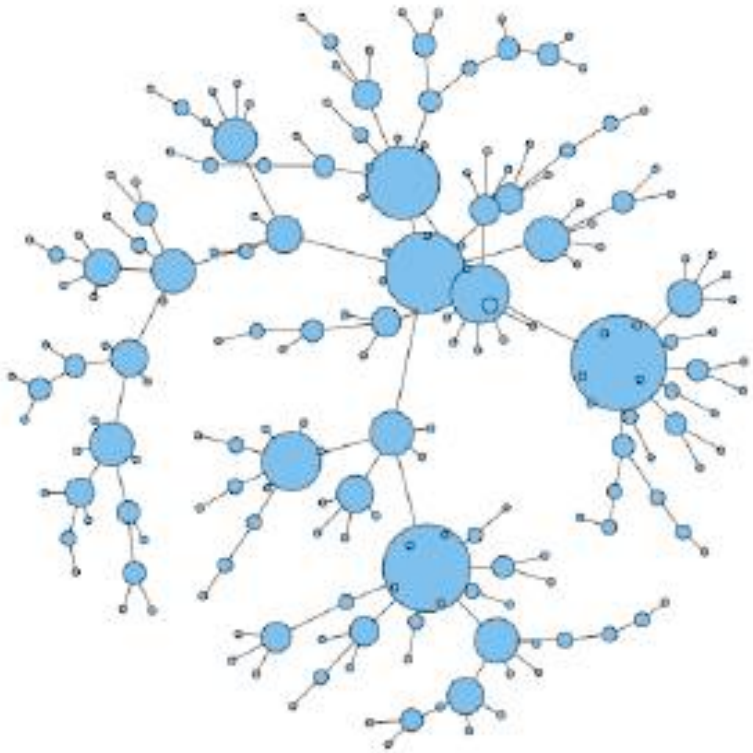
Node degree distribution:

$$P(k_i = k) = \frac{2m_0^2}{k^3}$$

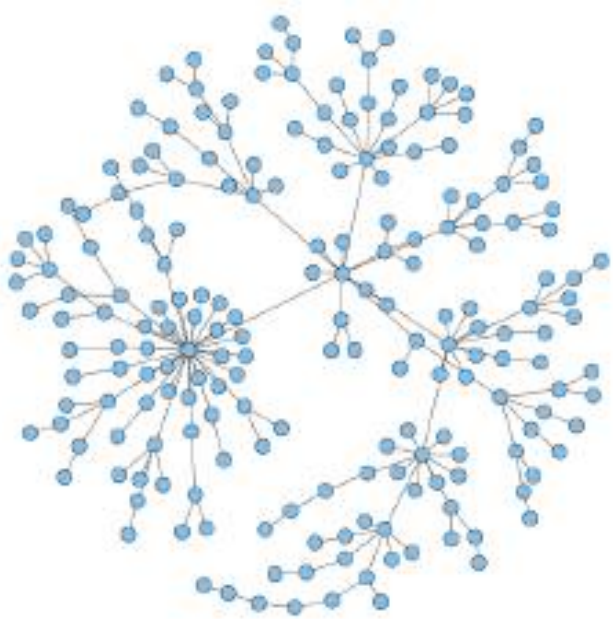
Preferential attachement

Preferential attachment vs random graph

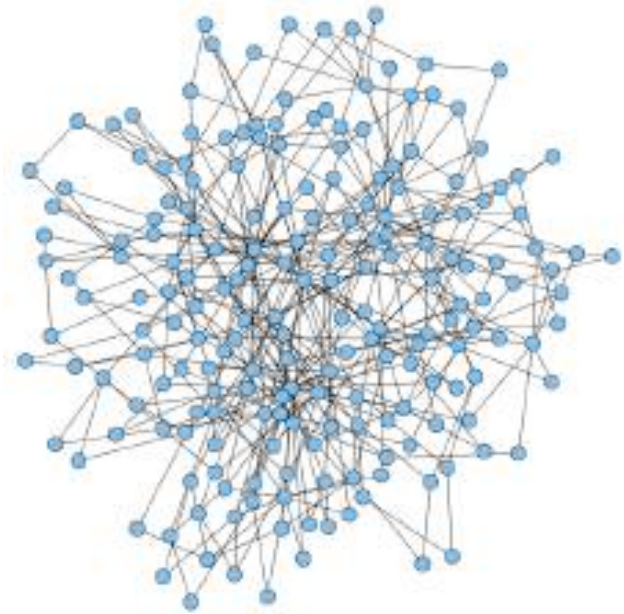




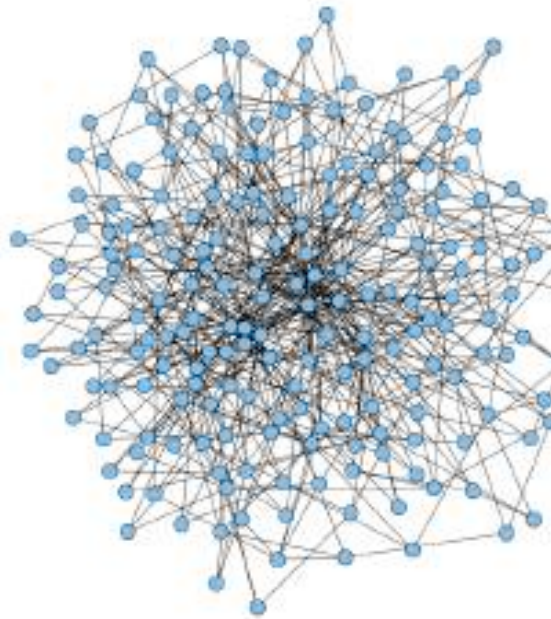
Preferential attachment model



$m_0 = 1$



$m_0 = 2$



$m_0 = 3$

igraph: barabasi.game()

Preferential attachment model

- Node degree distribution - power law):

$$P(k) = \frac{2m_0^2}{k^3}$$

- Average path length :

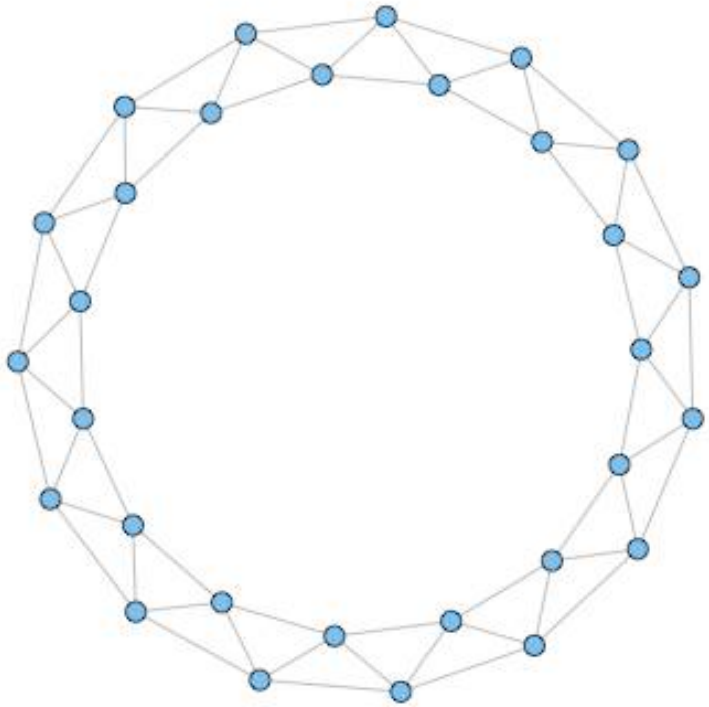
$$\langle L \rangle \sim \log(N) / \log(\log(N))$$

- Clustering coefficient (numerical result):

$$C \sim N^{-0.75}$$

Small world

Motivation: keep high clustering, get small diameter



Clustering coefficient $C = 1/2$

Graph diameter $d = 8$

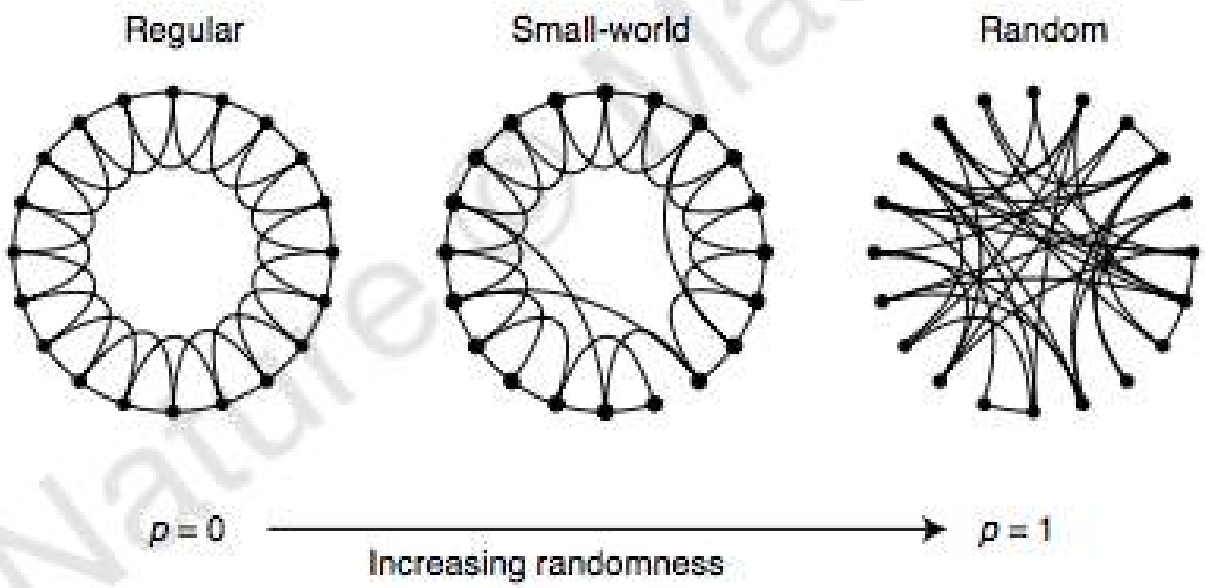
Small world

Watts and Strogatz, 1998

Single parameter model, interpolation between regular lattice and random graph

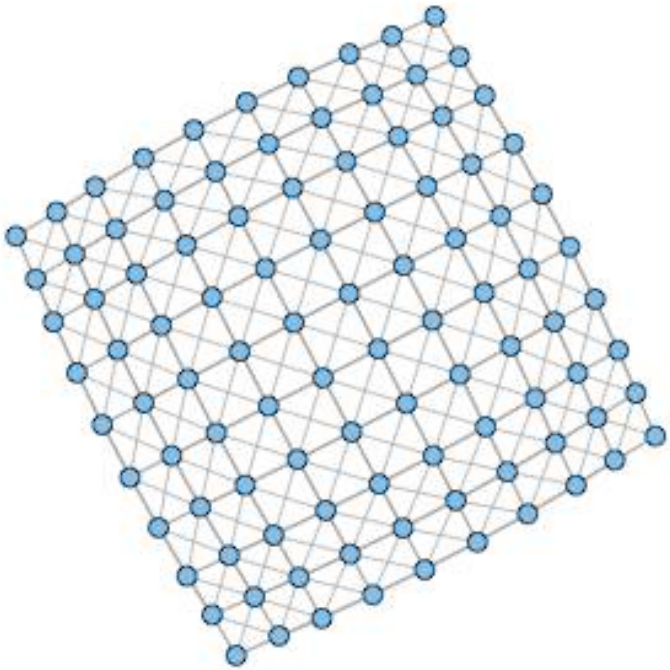
- start with regular lattice with n nodes, k edges per vertex (node degree), $k << n$
- randomly connect with other nodes with probability p, forms $pnk/2$ "long distance" connections from total of $nk/2$ edges
- $p = 0$ regular lattice, $p = 1$ random graph

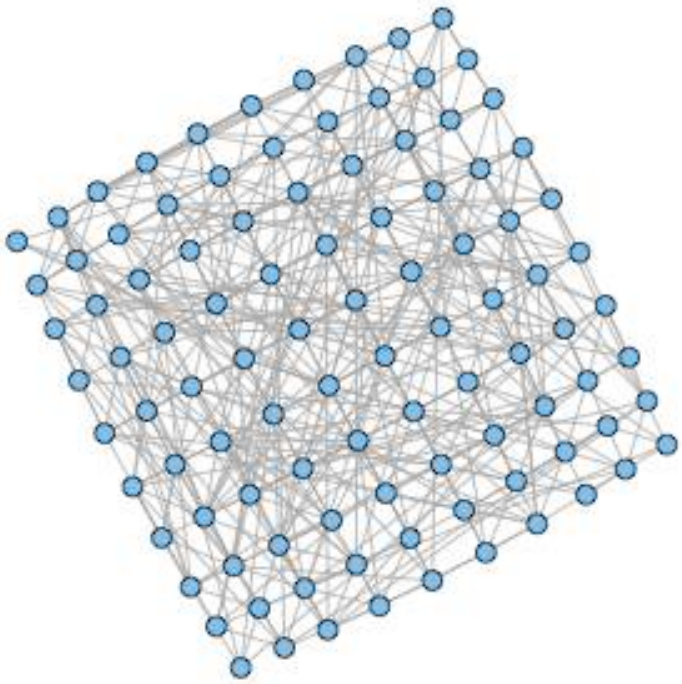
Small world



Watts, 1998

Small world model





20% rewiring:

ave. path length = 3.58 $\rightarrow$ ave. path length = 2.32

clust. coeff = 0.49 $\rightarrow$ clust. coeff = 0.19

igraph:watts.strogatz.game()

Small world model

- Node degree distribution function - Poisson like (numerical result)
- Average path length (analytical result) :

$$\langle L \rangle \sim \log(N)$$

- Clustering coefficient

$$C = const$$

Model comparison

	Random	BA model	WS model	Empirical networks
P(k)	$\frac{\lambda^k e^{-\lambda}}{k!}$	k^{-3}	poisson like	power law
C	$\langle k \rangle / N$	$N^{-0.75}$	const	large
$\langle L \rangle$	$\frac{\log(N)}{\log(\langle k \rangle)}$	$\frac{\log(N)}{\log\log(N)}$	$\log(N)$	small

References

- On random graphs I, P. Erdos and A. Renyi, Publicationes Mathematicae 6, 290–297 (1959).
- On the evolution of random graphs, P. Erdos and A. Renyi, Publicaton of the Mathematical Institute of the Hungarian Academy of Sciences, 17-61 (1960)
- Collective dynamics of small-world networks. Duncan J. Watts and Steven H. Strogatz. Nature 393 (6684): 440-442, 1998
- Emergence of Scaling in Random Networks, A.L. Barabasi and R. Albert, Science 286, 509-512, 1999

NATIONAL RESEARCH
UNIVERSITY

Node Centrality and Ranking on Networks

Social Network Analysis. MAGoLEGO course.
Lecture 4

Leonid Zhukov

lzhukov@hse.ru
www.leonidzhukov.net/hse/2018/sna

National Research University Higher School of Economics
School of Data Analysis and Artificial Intelligence, Department of Computer Science

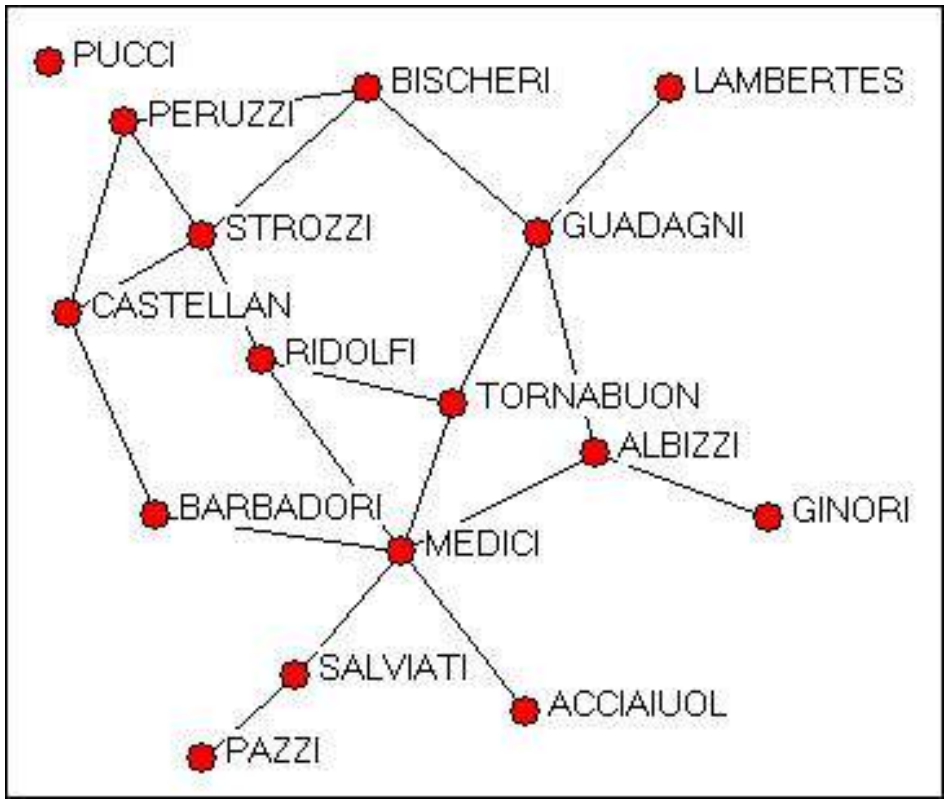
Centrality

Which vertices are important?

image from M.Grandjean, 2014

Centrality Measures

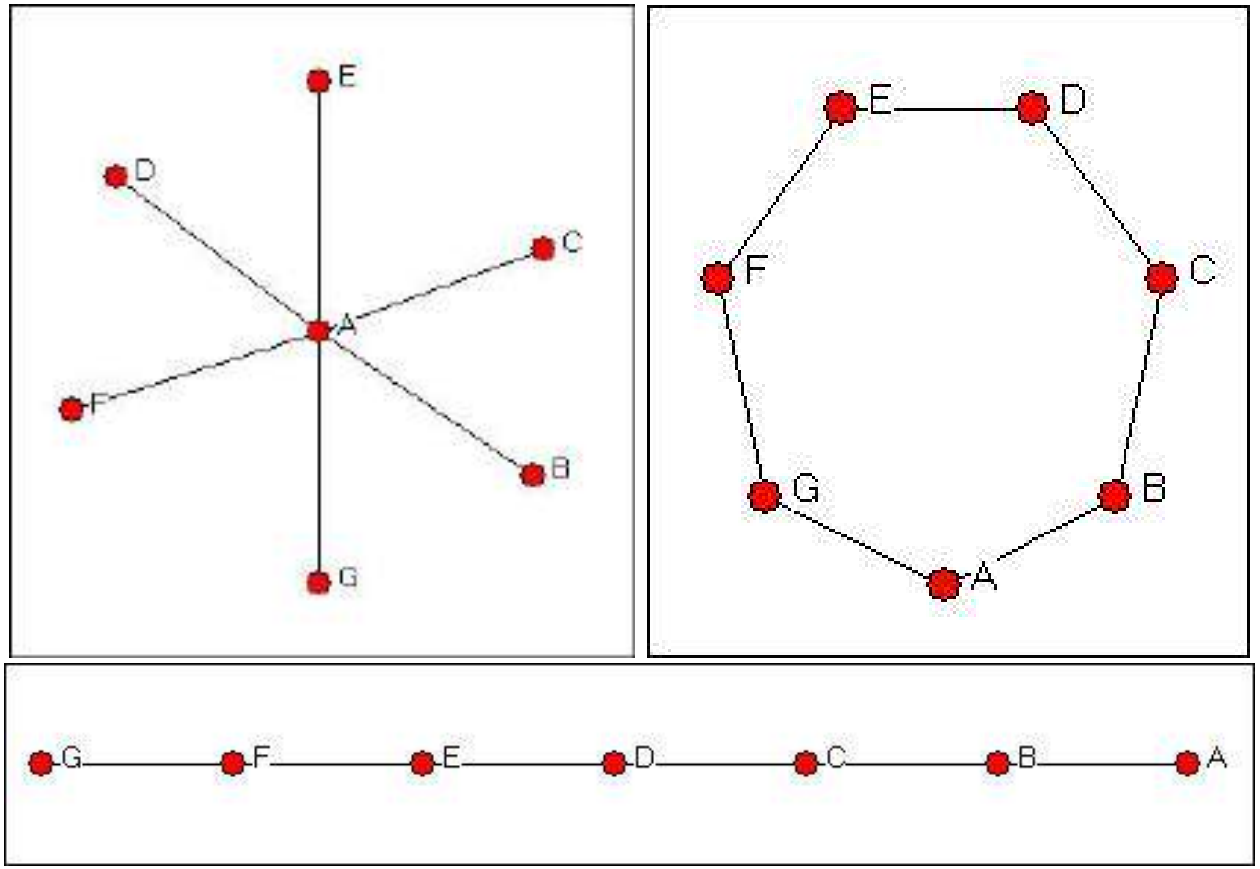
Determine the most ”important” or ”prominent” actors in the network based on actor location.



Marriage alliances among leading Florentine families 15th century.

Padgett, 1993

Three graphs



Star graph Circle graph Line Graph

Degree centrality

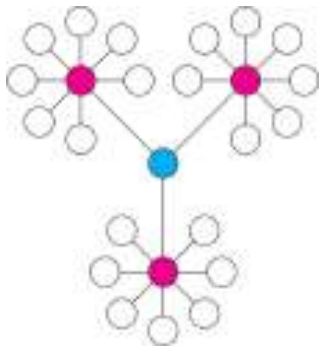
Degree centrality: number of nearest neighbors

$$C_D(i) = k(i) = \sum_j A_{ij} = \sum_j A_{ji}$$

Normalized degree centrality

$$C_D^*(i) = \frac{1}{n-1} C_D(i) = \frac{k(i)}{n-1}$$

High centrality degree -direct contact with many other actors



Closeness centrality

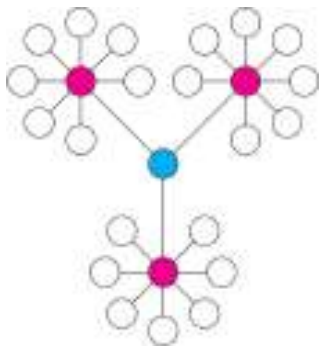
Closeness centrality: how close an actor to all the other actors in network

$$C_C(i) = \frac{1}{\sum_j d(i,j)}$$

Normalized closeness centrality

$$C_C^*(i) = (n-1)C_C(i) = \frac{n-1}{\sum_j d(i,j)}$$

High closeness centrality - short communication path to others, minimal number of steps to reach others



Betweenness centrality

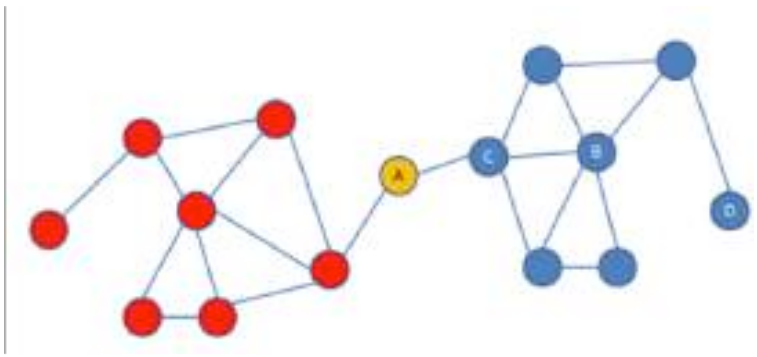
Betweenness centrality: number of shortest paths going through the actor $\sigma_{st}(i)$

$$C_B(i) = \sum_{s \neq t \neq i} \frac{\sigma_{st}(i)}{\sigma_{st}}$$

Normalized betweenness centrality

$$C_B^*(i) = \frac{2}{(n-1)(n-2)} C_B(i) = \frac{2}{(n-1)(n-2)} \sum_{s \neq t \neq i} \frac{\sigma_{st}(i)}{\sigma_{st}}$$

Hight betweenness centrality - vertex lies on many shortest paths
Probability that a communication from *s* to *t* will go through *i*



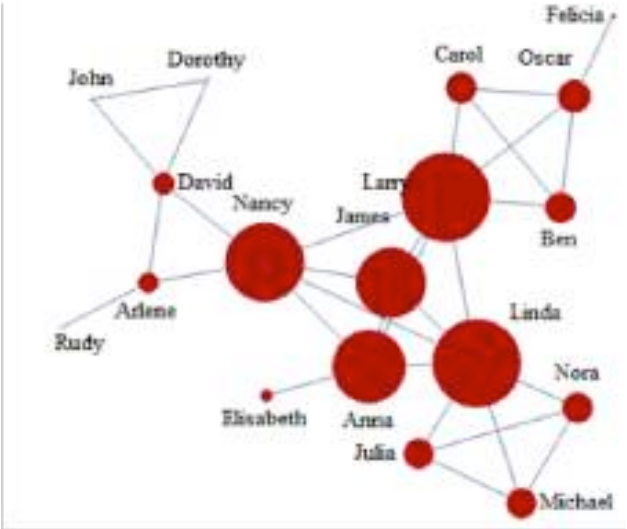
Eigenvector centrality

Importance of a node depends on the importance of its neighbors (recursive definition)

$$v_i \leftarrow \sum_j A_{ij} v_j$$

$$v_i = \frac{1}{\lambda} \sum_j A_{ij} v_j$$

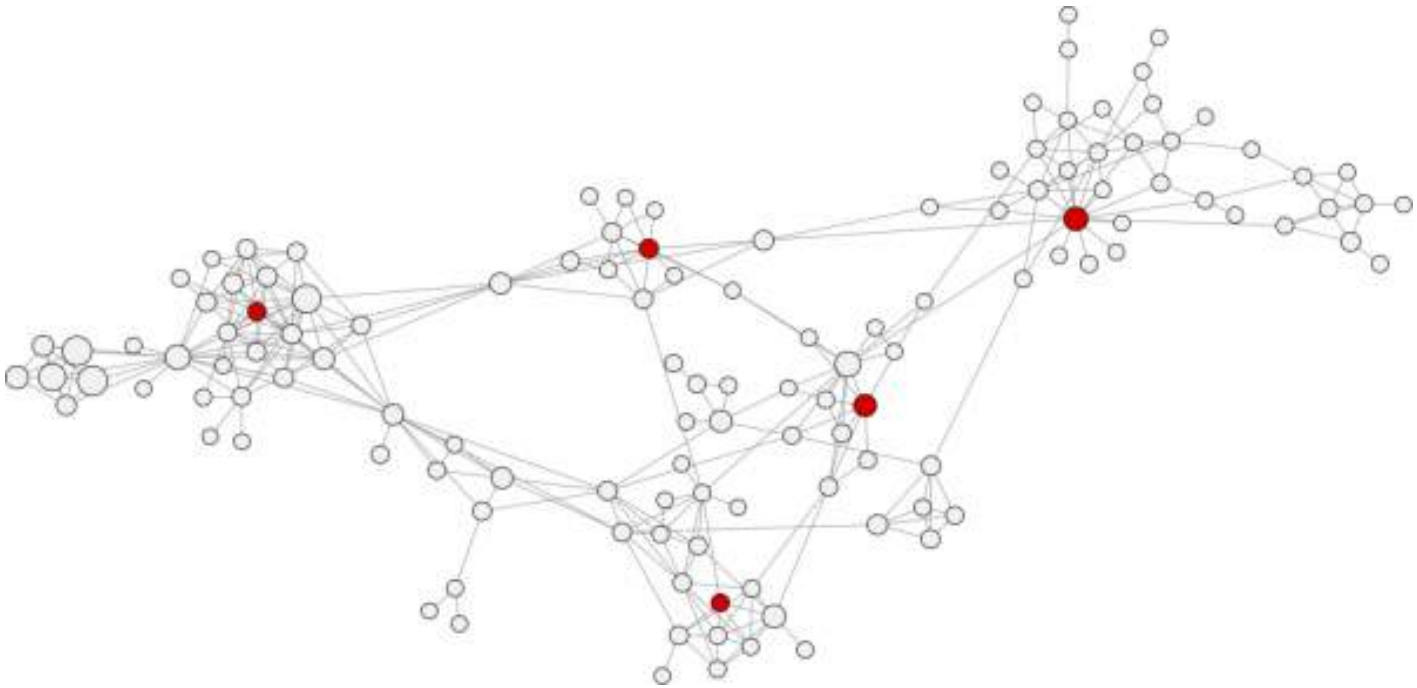
$$\mathbf{Av} = \lambda \mathbf{v}$$



Select an eigenvector associated with largest eigenvalue $\lambda = \lambda_1$, $\mathbf{v} = \mathbf{v}_1$

Centrality examples

Closeness centrality

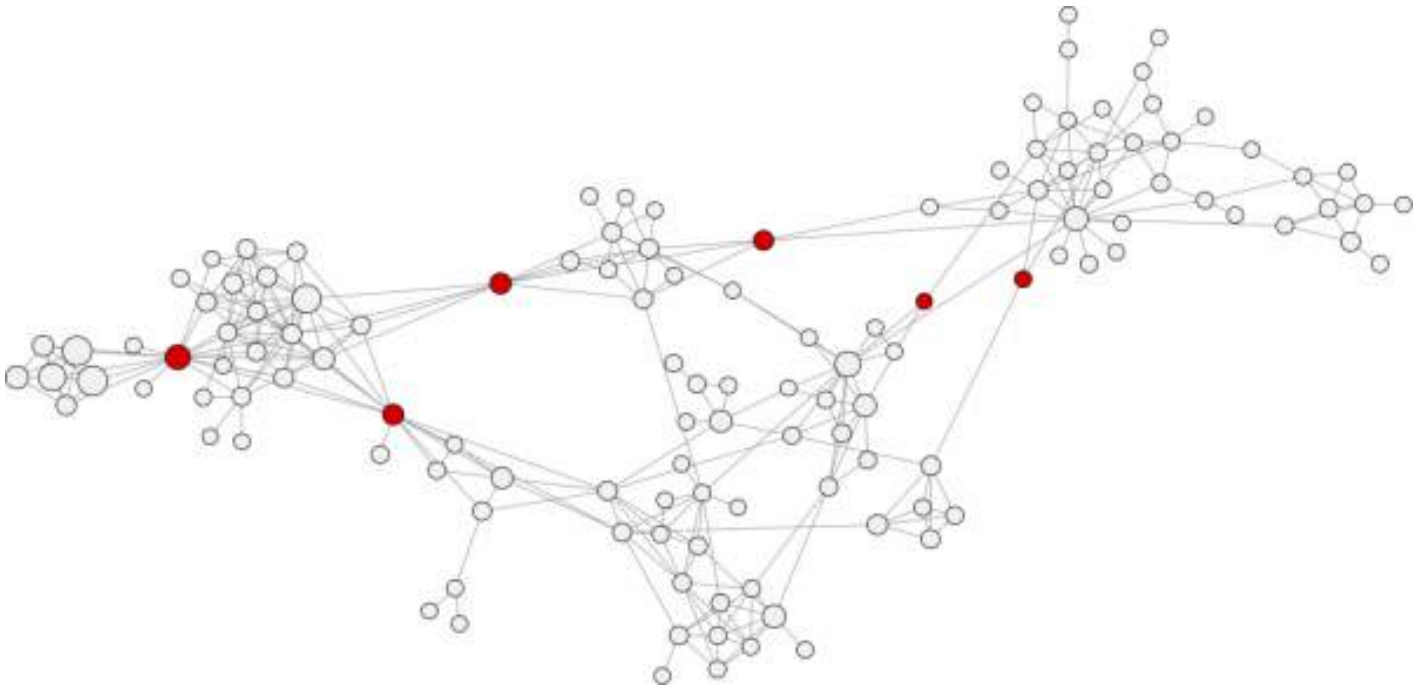


igraph:closeness()

from www.activenetworks.net

Centrality examples

Betweenness centrality

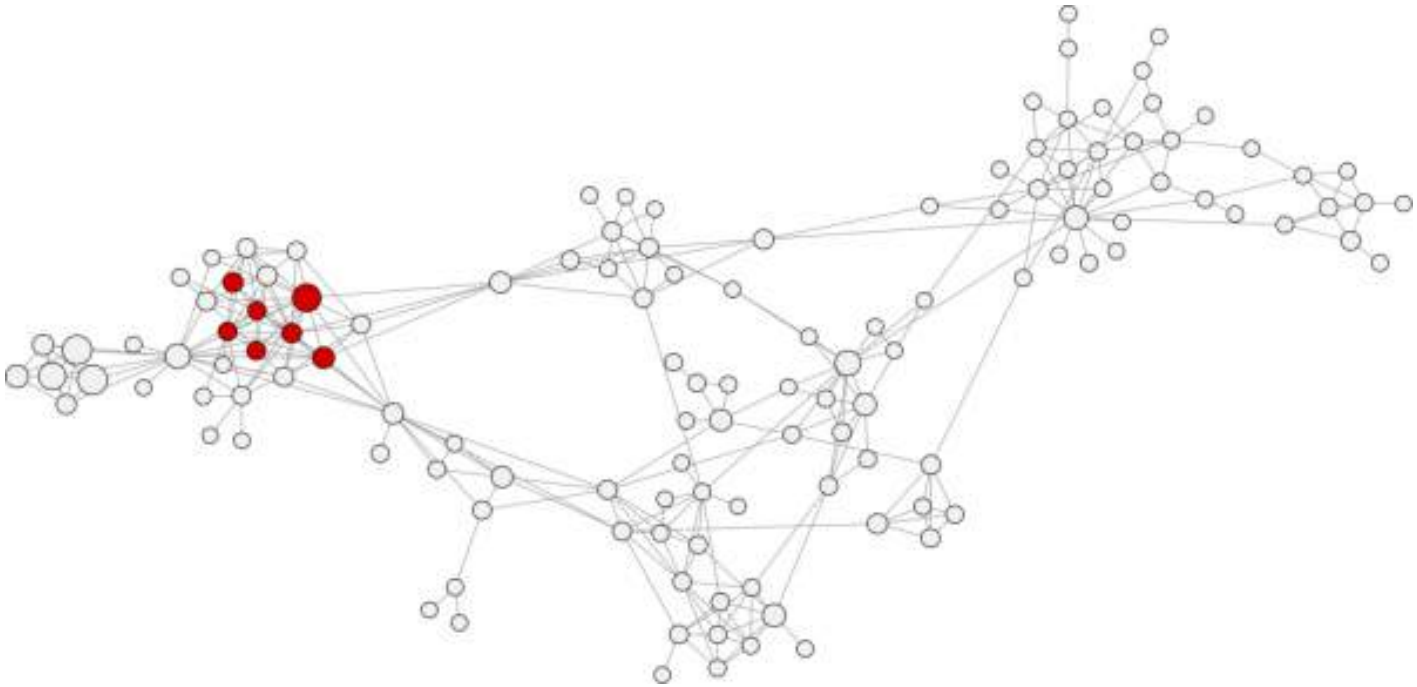


igraph:betweenness()

from www.activenetworks.net

Centrality examples

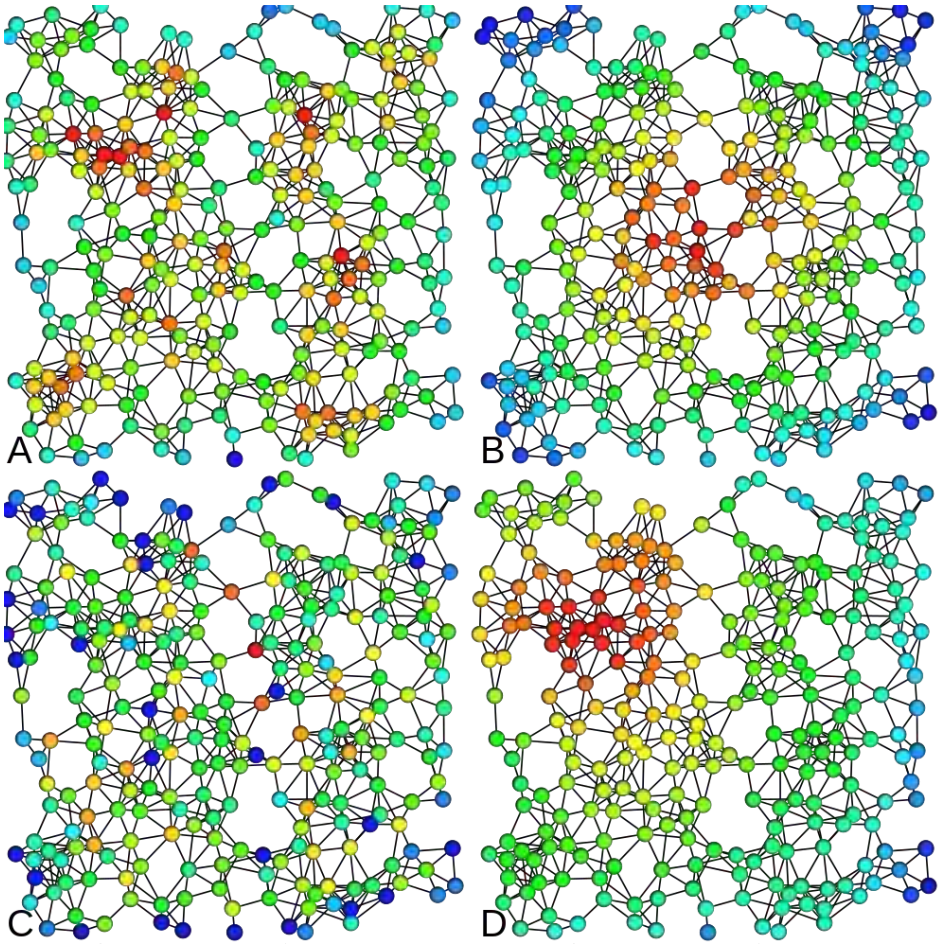
Eigenvector centrality



igraph:evcent()

from www.activenetworks.net

Centrality examples



from Claudio Rocchini

- A) Degree centrality
- B) Closeness centrality
- C) Betweenness centrality
- D) Eigenvector centrality

Centralization

Centralization (network measure) - how central the most central node in the network in relation to all other nodes.

$$C_x = \frac{\sum_i^N [C_x(p_*) - C_x(p_i)]}{\max \sum_i^N [C_x(p_*) - C_x(p_i)]}$$

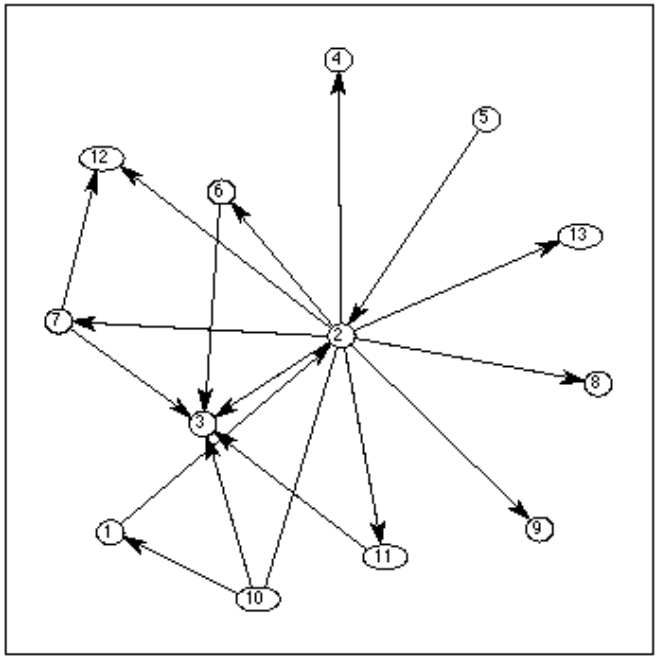
C_x - one of the centrality measures
p_* - node with the largest centrality value
max - is taken over all graphs with the same number of nodes (for degree, closeness and betweenness the most centralized structure is the star graph)

igraph: centralization.degree(), centralization.closeness(), centralization.betweenness(), centralization.evcent()

Linton Freeman, 1979

Directional relations

Directed graph: distinguish between choices made (outgoing edges) and choices received (incoming edges)



sending - receiving
export - import
cite papers - being cited

Centrality measures

All based on outgoing edges

- Degree centrality (normalized):

$$C_D^*(i) = \frac{k^{out}(i)}{n-1}$$

- Closeness centrality (normalized):

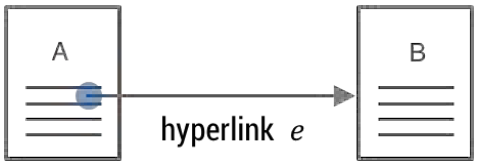
$$C_C^*(i) = \frac{n-1}{\sum_j d(i,j)}$$

- **Betweenness centrality (normalized):

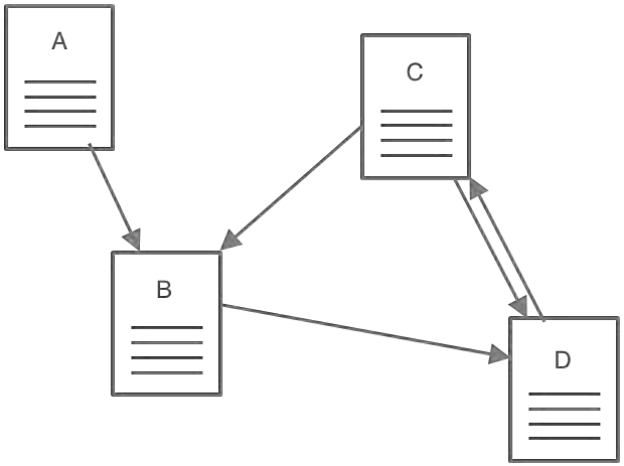
$$C_B^*(i) = \frac{1}{(n-1)(n-2)} \sum_{s \neq t \neq i} \frac{\sigma_{st}(i)}{\sigma_{st}}$$

Web as a graph

- Hyperlinks - implicit endorsements

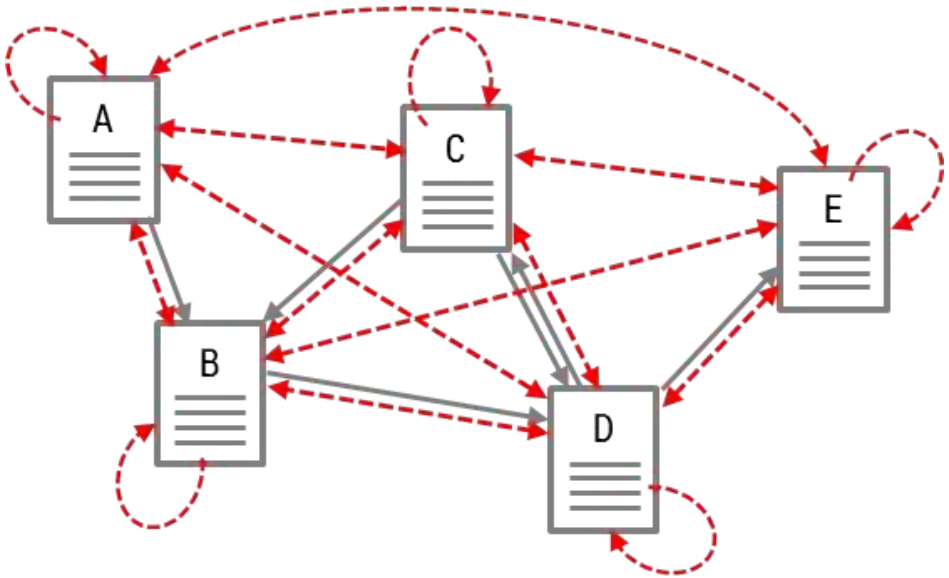


- Web graph - graph of endorsements (sometimes reciprocal)



PageRank

"PageRank can be thought of as a model of user behavior. We assume there is a "random surfer" who is given a web page at random and keeps clicking on links, never hitting "back" but eventually gets bored and starts on another random page. The **probability** that the random surfer visits a page is its **PageRank**."



Sergey Brin and Larry Page, 1998

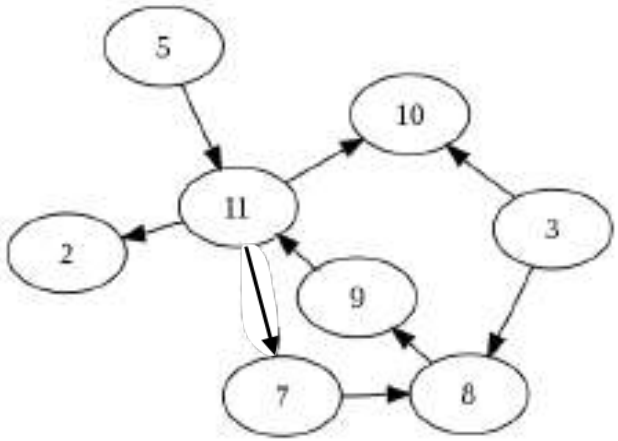
Random walk

- Random walk on graph

$$p_i^{t+1} = \sum_{j \in N(i)} \frac{p_j^t}{d_j^{out}} = \sum_j \frac{A_{ji}}{d_j^{out}} p_j$$

$$\mathbf{p}^{t+1} = \mathbf{P}^T \mathbf{p}^t$$

$$\mathbf{P} = \mathbf{D}^{-1}\mathbf{A}, \quad \mathbf{D}_{ii} = diag\{d_i^{out}\}$$



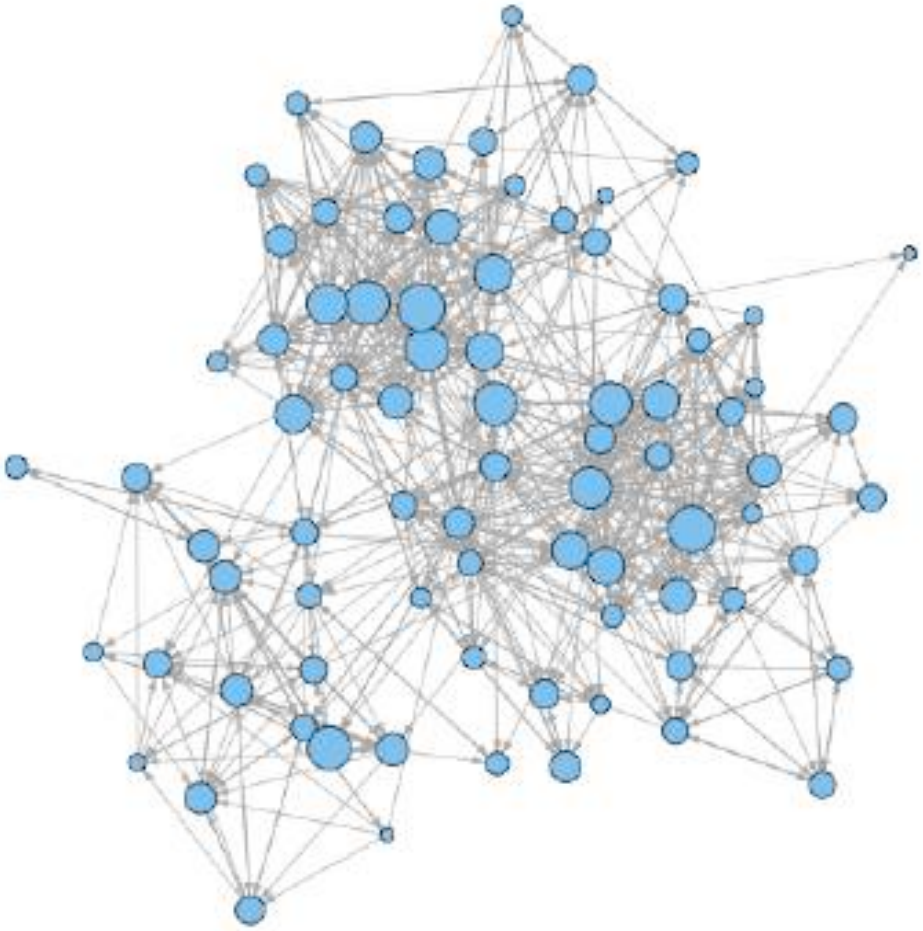
- with teleportation

$$\mathbf{p}^{t+1} = \alpha \mathbf{P}^T \mathbf{p}^t + (1-\alpha)\mathbf{v}$$

Perron-Frobenius Theorem guarantees existence and uniqueness of the solution $\lim_{t\to\infty} \mathbf{p} = \pi$

$$\pi = \alpha \mathbf{P}^T \pi + (1-\alpha)\mathbf{v}$$

PageRank



igraph: page.rank()

PageRank beyond the Web

1. GeneRank
2. ProteinRank
3. FoodRank
4. SportsRank
5. HostRank
6. TrustRank
7. BadRank
8. ObjectRank
9. ItemRank
10. ArticleRank
11. BookRank
12. FutureRank
13. TimedPageRank
14. SocialPageRank
15. DiffusionRank
16. ImpressionRank
17. TweetRank
18. TwitterRank
19. ReversePageRank
20. PageTrust
21. PopRank
22. CiteRank
23. FactRank
24. InvestorRank
25. ImageRank
26. VisualRank
27. QueryRank
28. BookmarkRank
29. StoryRank
30. PerturbationRank
31. ChemicalRank
32. RoadRank
33. PaperRank
34. Etc...

Hubs and Authorities (HITS)

Citation networks. Reviews vs original research (authoritative) papers

- authorities, contain useful information, a_i
- hubs, contains links to authorities, h_i

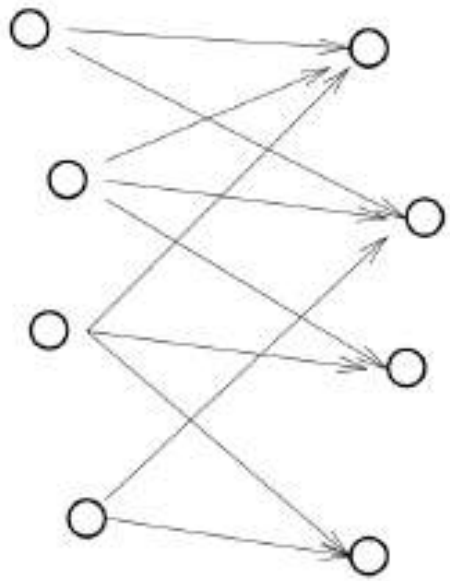
Mutual recursion

- Good authorities referred by good hubs

$$a_i \leftarrow \sum_j A_{ji} h_j$$

- Good hubs point to good authorities

$$h_i \leftarrow \sum A_{ij} a_j$$



HITS

System of linear equations

$$\begin{aligned} \mathbf{a} &= \alpha \mathbf{A}^T \mathbf{h} \\ \mathbf{h} &= \beta \mathbf{A} \mathbf{a} \end{aligned}$$

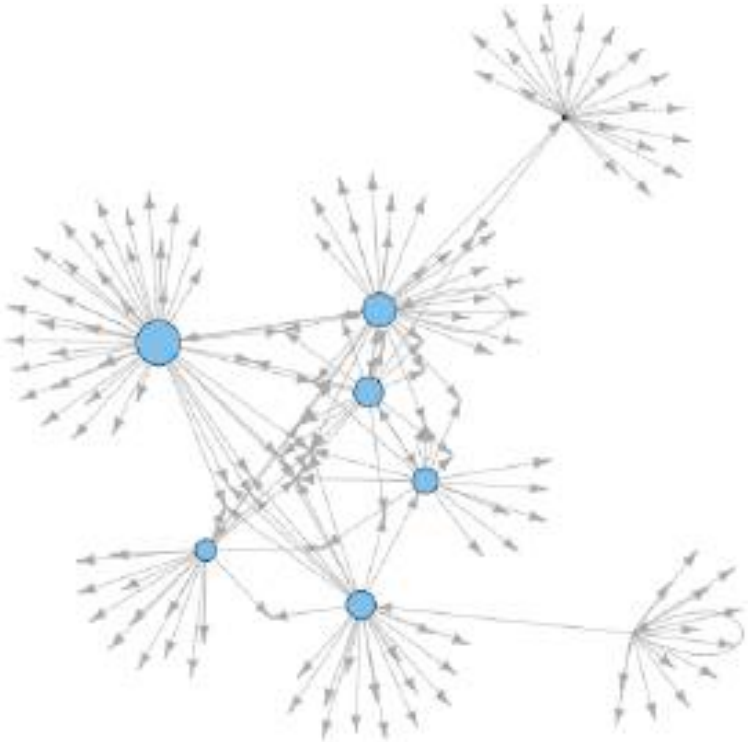
Symmetric eigenvalue problem

$$\begin{aligned} (\mathbf{A}^T \mathbf{A})\mathbf{a} &= \lambda \mathbf{a} \\ (\mathbf{A}\mathbf{A}^T)\mathbf{h} &= \lambda \mathbf{h} \end{aligned}$$

where eigenvalue $\lambda = (\alpha\beta)^{-1}$

Hubs and Authorities

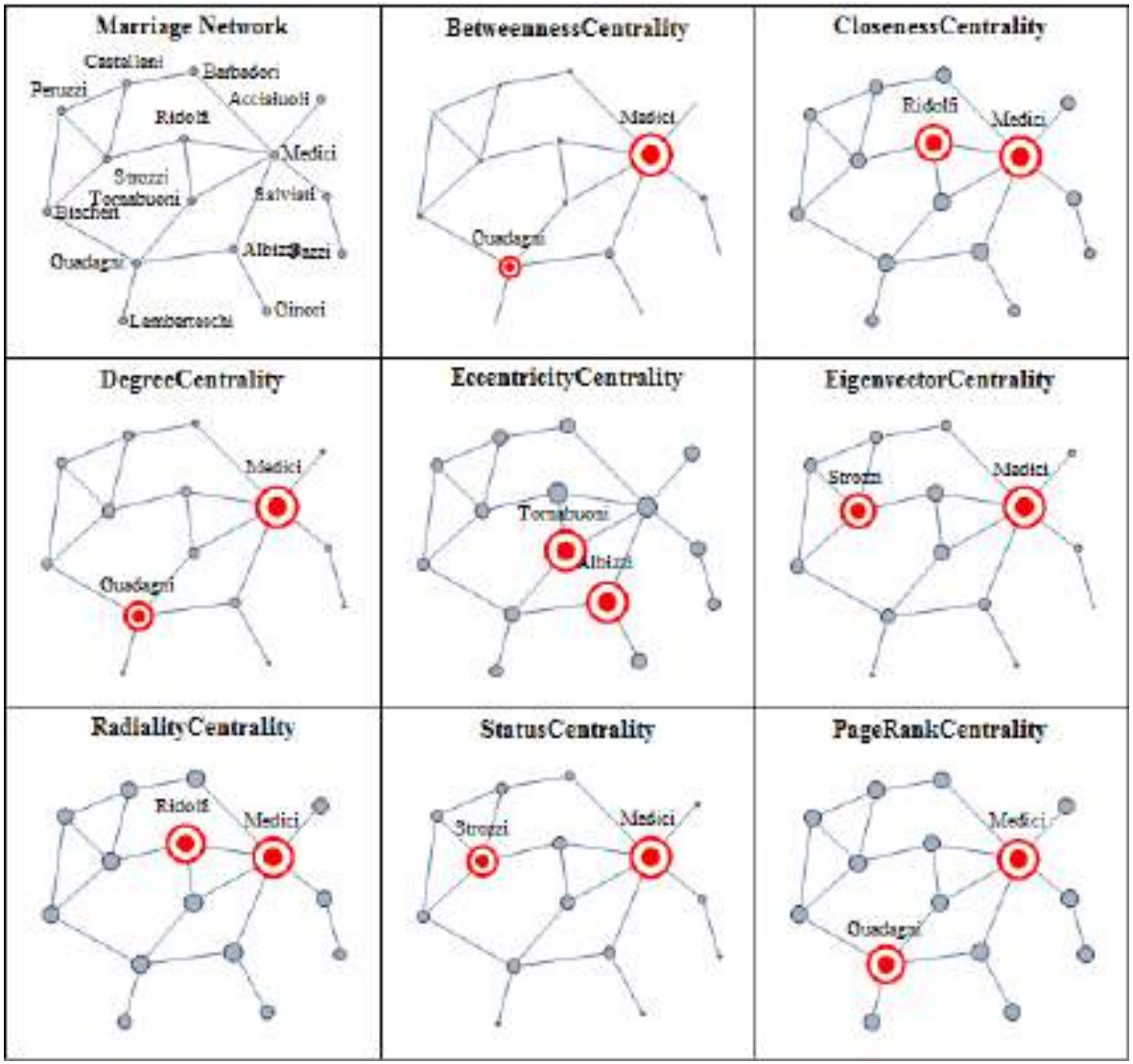
Hubs



Authorities

igraph: hub.score(), authority.score()

Florentines families



References

- Centrality in Social Networks. Conceptual Clarification, Linton C. Freeman, Social Networks, 1, 215-239, 1979
- Power and Centrality: A Family of Measures, Phillip Bonacich, The American Journal of Sociology, Vol. 92, No. 5, 1170-1182, 1987
- A new status index derived from sociometric analysis, L. Katz, Psychometrika, 19, 39-43, 1953.
- Eigenvector-like measures of centrality for asymmetric relations, Phillip Bonacich, Paulette Lloyd, Social Networks 23, 191-201, 2001
- The PageRank Citation Ranknig: Bringing Order to the Web. S. Brin, L. Page, R. Motwany, T. Winograd, Stanford Digital Library Technologies Project, 1998
- Authoritative Sources in a Hyperlinked Environment. Jon M. Kleinberg, Proc. 9th ACM-SIAM Symposium on Discrete Algorithms
- A Survey of Eigenvector Methods of Web Information Retrieval. Amy N. Langville and Carl D. Meyer, 2004
- PageRank beyond the Web. David F. Gleich, arXiv:1407.5107, 2014

NATIONAL RESEARCH UNIVERSITY

Network communities

Social Network Analysis. MAGoLEGO course. Lecture 5

Leonid Zhukov

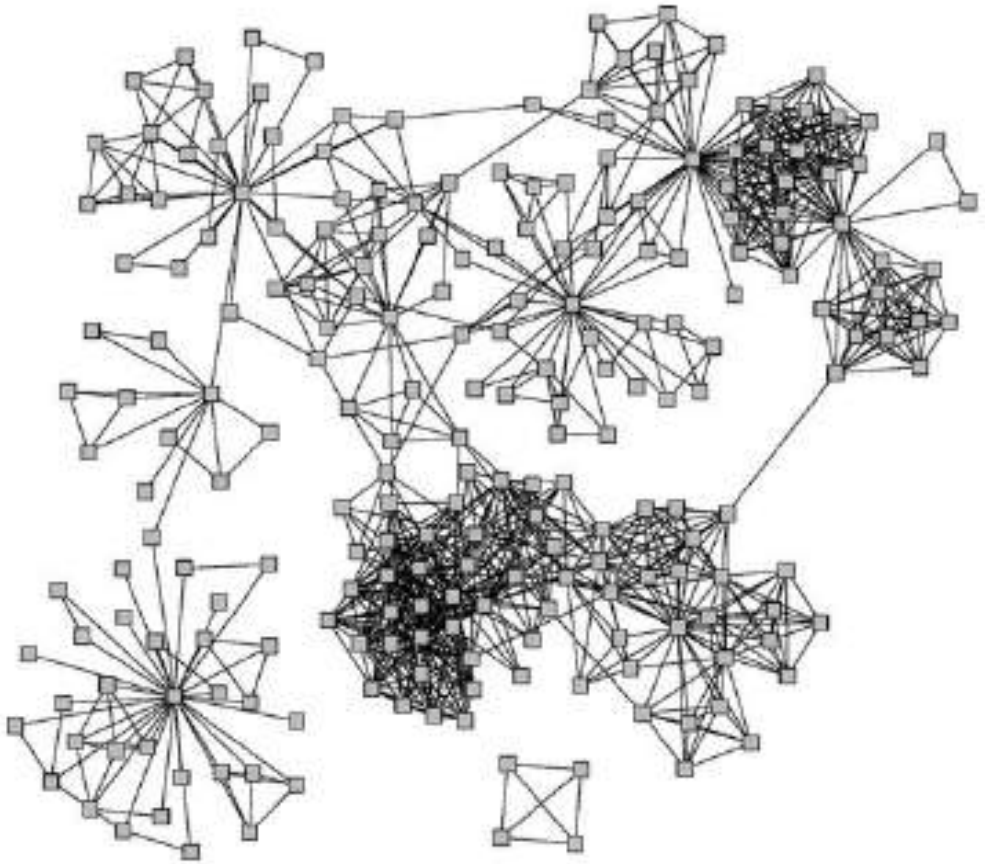
lzhukov@hse.ru

www.leonidzhukov.net/hse/2018/sna

National Research University Higher School of Economics

School of Data Analysis and Artificial Intelligence, Department of Computer Science

Network communities



Connected and undirected graphs

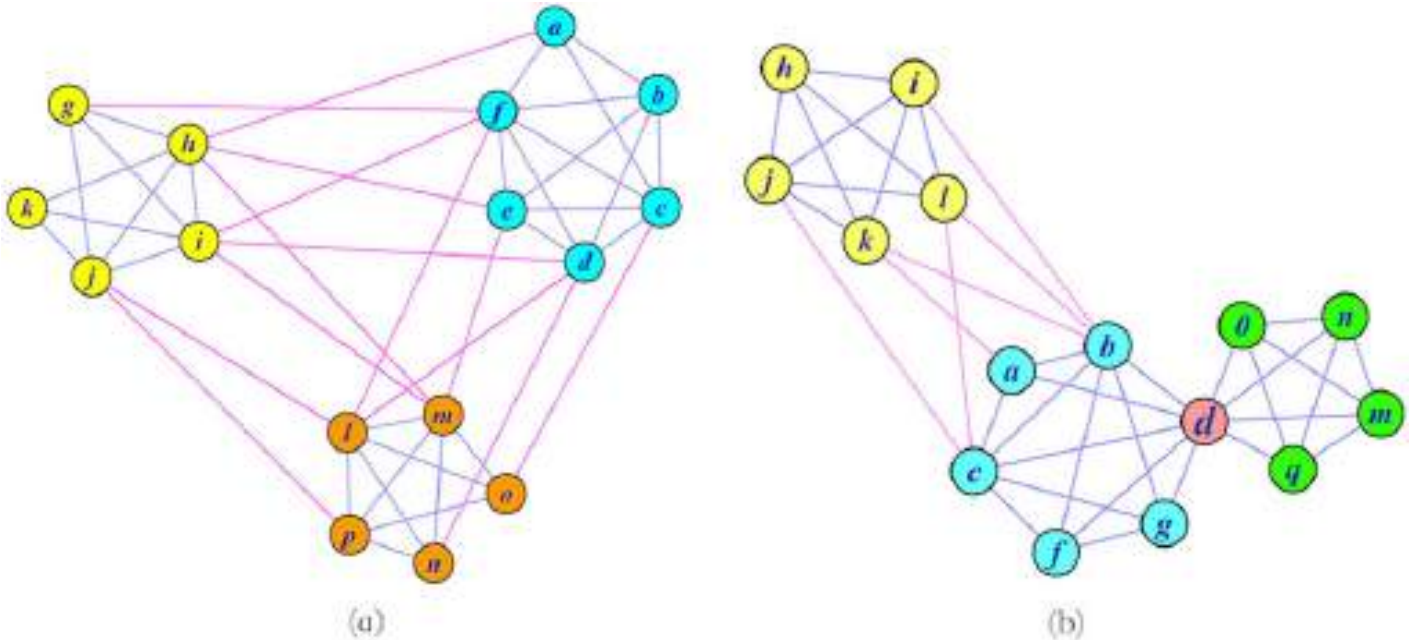
Network communities

What makes a community (cohesive subgroup):

- Mutuality of ties. Everyone in the group has ties (edges) to one another
- Compactness. Closeness or reachability of group members in small number of steps, not necessarily adjacency
- Density of edges. High frequency of ties within the group
- Separation. Higher frequency of ties among group members compared to non-members

Wasserman and Faust

Community types



Community types:

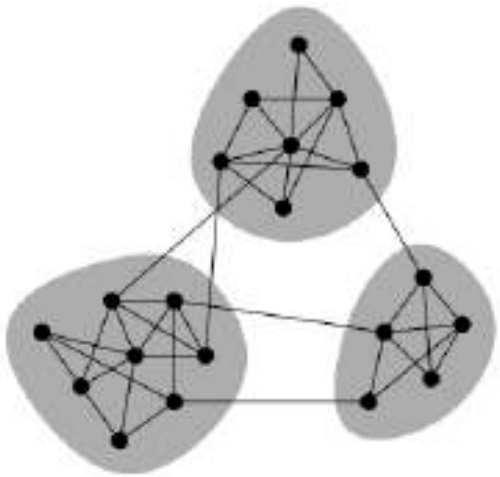
- Non-overlapping
- Overlapping

image from W. Liu , 2014

Network communities

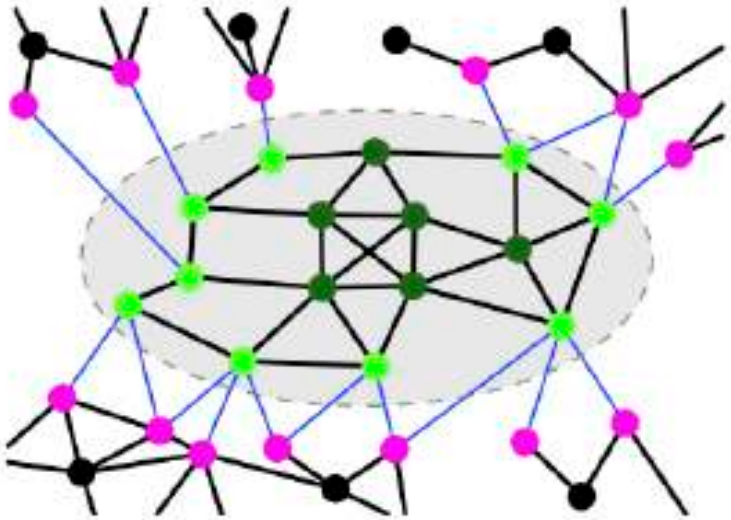
Definition

Network communities are groups of vertices such that vertices inside the group connected with many more edges than between groups.



- Community detection is an assignment of vertices to communities.
- Will consider non-overlapping communities, graph cuts

Graph cuts



Graph cut is a partition of the vertices of a graph $G(E, V)$ into two disjoint subsets: $V = V_1 + V_2$

$$Q = cut(V_1, V_2) = \sum_{i \in V_1, j \in V_2} e_{ij}$$

image from Fortunato 2016

Graph cuts

Graph $G(E, V)$ partition: $V = V_1 + V_2$

- Graph cut

$$Q = cut(V_1, V_2) = \sum_{i \in V_1, j \in V_2} e_{ij}$$

- Ratio cut:

$$Q = \frac{cut(V_1, V_2)}{||V_1||} + \frac{cut(V_1, V_2)}{||V_2||}$$

- Normalized cut:

$$Q = \frac{cut(V_1, V_2)}{Vol(V_1)} + \frac{cut(V_1, V_2)}{Vol(V_2)}$$

- Quotient cut (conductance):

$$Q = \frac{cut(V_1, V_2)}{\min\left(Vol(V_1), Vol(V_2)\right)}$$

where: $Vol(V_1) = \sum_{i \in V_1, j \in V} e_{ij} = \sum_{i \in V_1} k_i$

Modularity

- Compare fraction of edges within the cluster to expected fraction in random graph with identical degree sequence

$$Q = \frac{1}{4}(m_s - E(m_s))$$

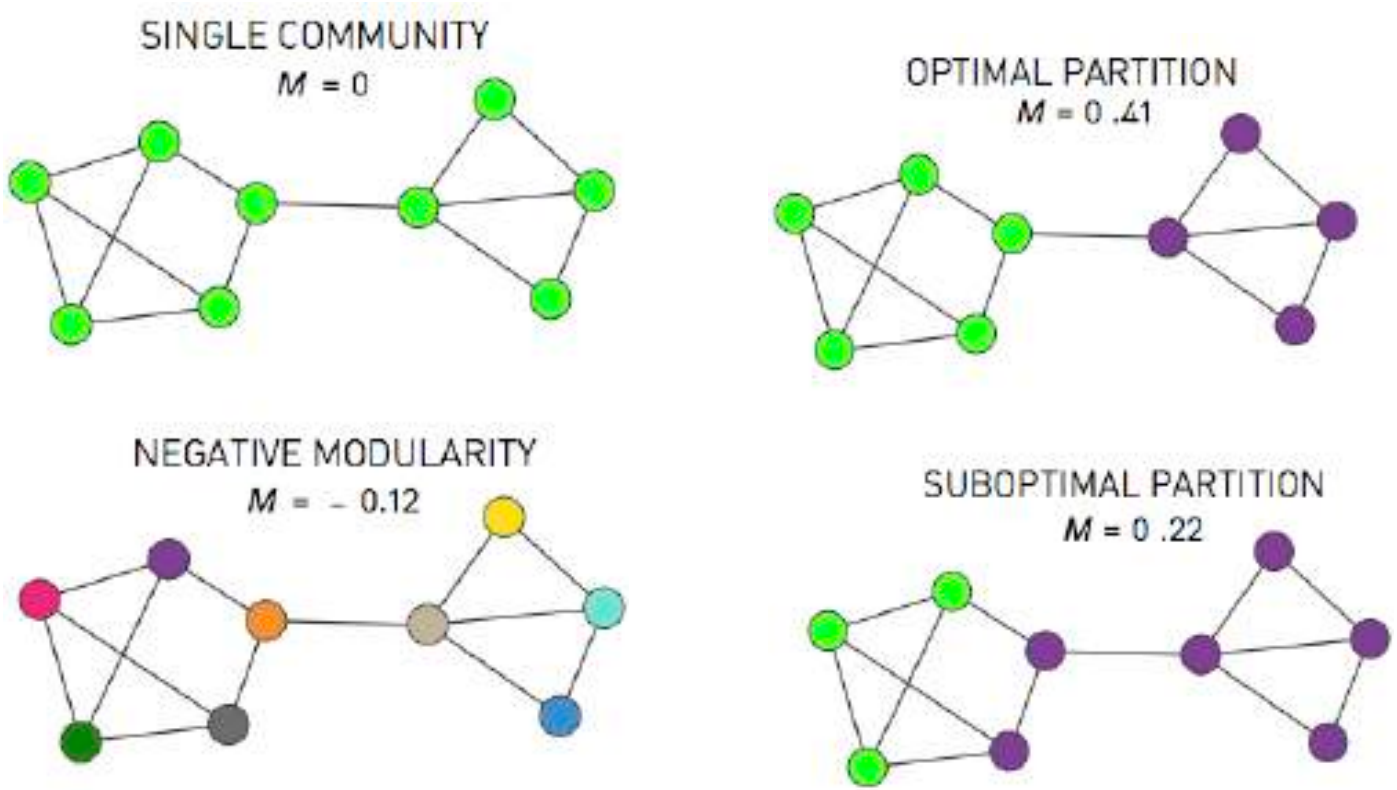
- Modularity score

$$Q = \frac{1}{2m}\sum_{ij}\left(A_{ij} - \frac{k_i k_j}{2m}\right)\delta(c_i, c_j), = \sum_{u}\left(\frac{m_u}{m} - \left(\frac{k_u}{2m}\right)^2\right)$$

m_u - number of internal edges in a community u,
k_u - sum of node degrees within a community

- Modularity score range $Q \in [-1/2, 1)$, single community $Q = 0$

Modularity



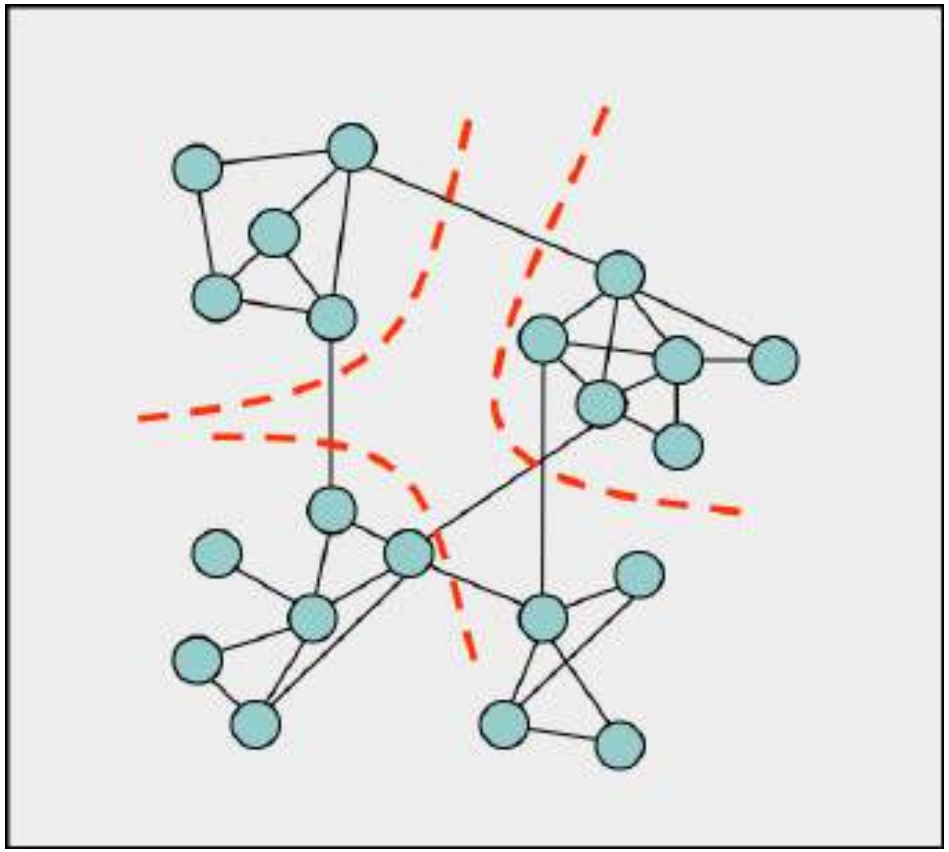
- The higher the modularity score - the better are communities

image from A.L. Barabasi 2016

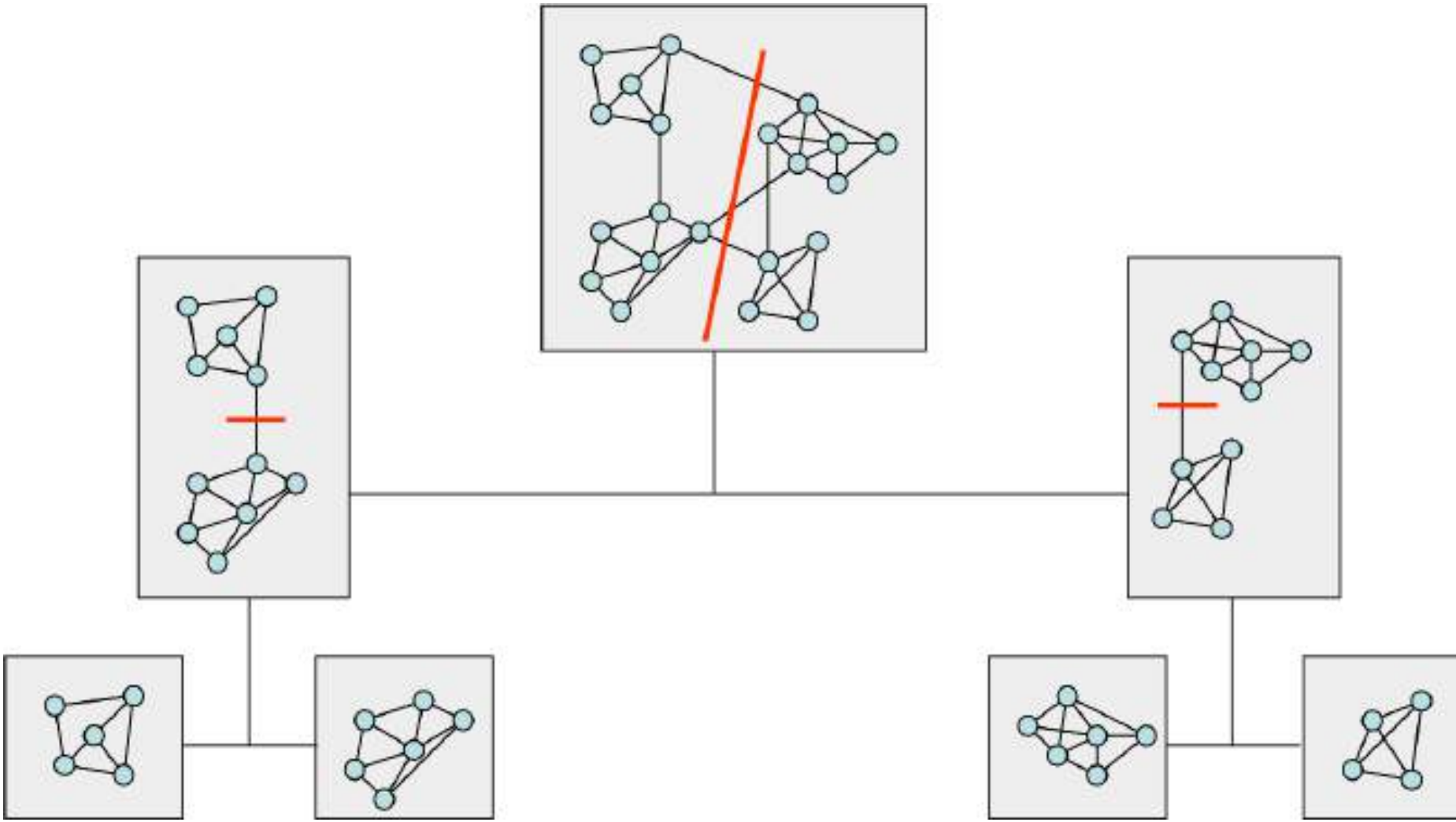
Community detection

- Consider only sparse graphs $m \ll n^2$
- Each community should be connected
- Combinatorial optimization problem:
  - optimization criterion (cut, conductance, modularity)
  - optimization method
- Exact solution NP-hard
  (bi-partition: $n = n_1 + n_2$, $n!/(n_1!n_2!)$ combinations)
- Solved by greedy, approximate algorithms or heuristics
- Recursive top-down 2-way partition, multiway partition
- Balanced class partition vs communities

Multiway partitioning



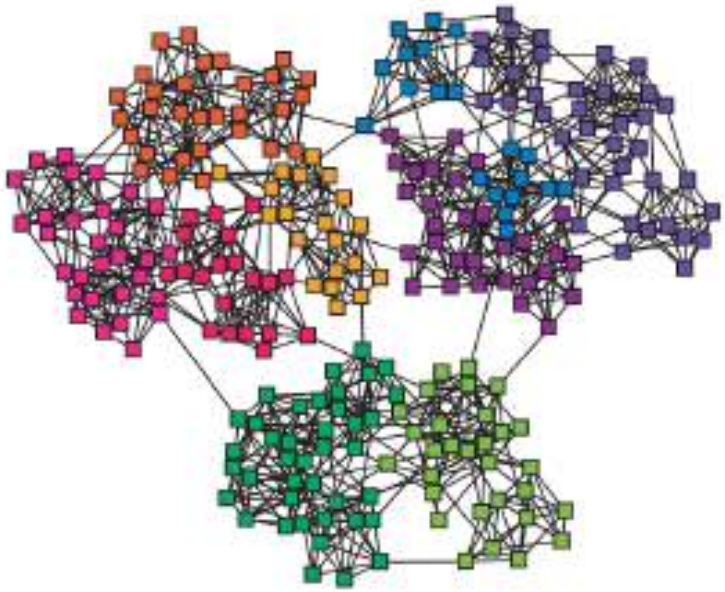
Recursive partitioning



Edge betweenness

Focus on edges that connect communities.
Edge betweenness -number of shortest paths $\sigma_{st}(e)$ going through edge e

$$C_B(e) = \sum_{s \neq t} \frac{\sigma_{st}(e)}{\sigma_{st}}$$



Edge betweenness algorithm

Newman-Girvan, 2004

Algorithm: Edge Betweenness

Input: graph G(V,E)

Output: Dendrogram/communities

repeat

For all $e \in E$ compute edge betweenness $C_B(e)$;

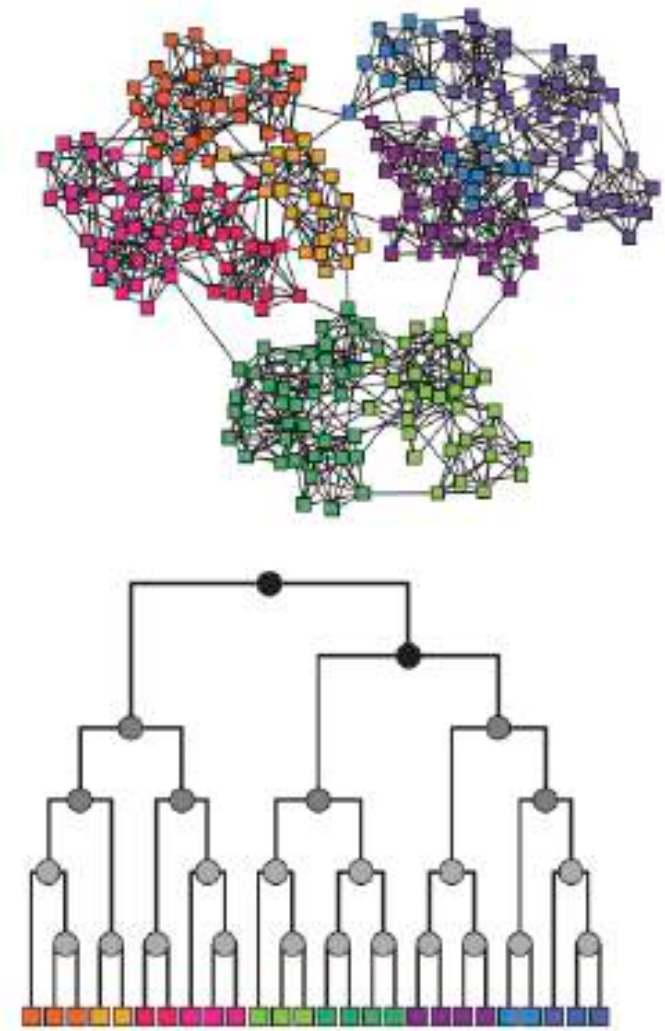
remove edge e_i with largest $C_B(e_i)$;

until *edges left*;

If bi-partition, then stop when graph splits in two components (check for connectedness)

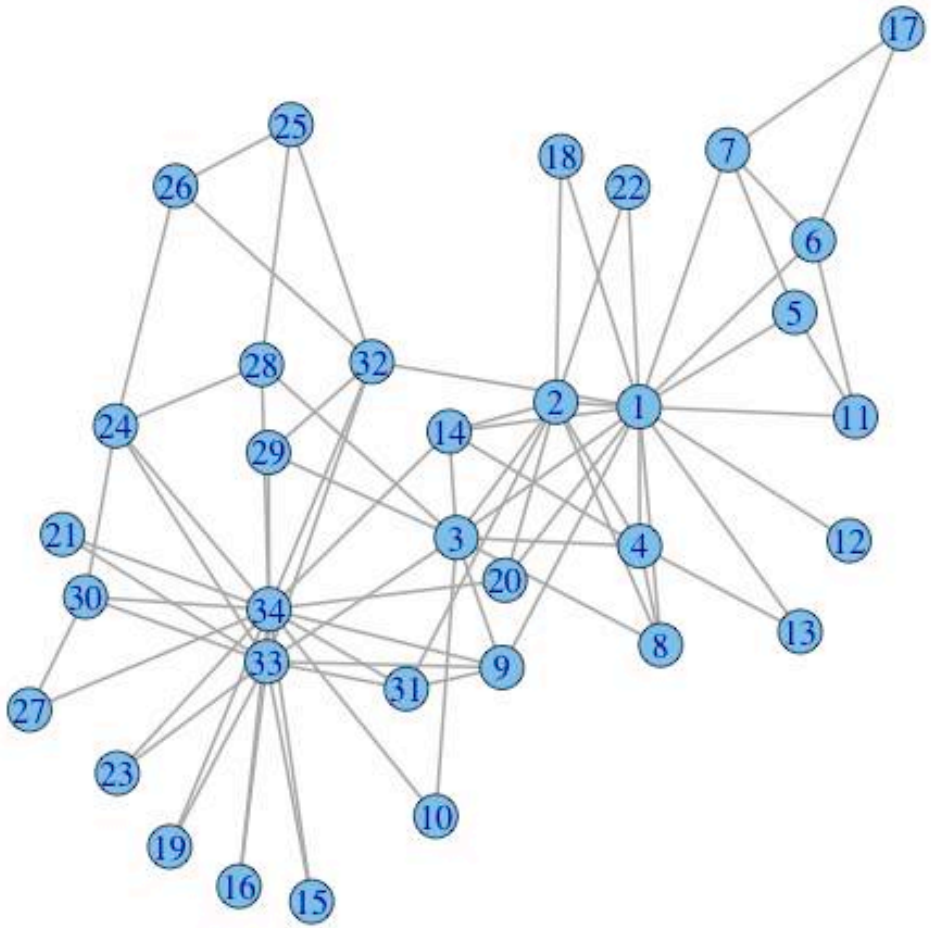
Edge betweenness

Hierarchical algorithm, dendrogram



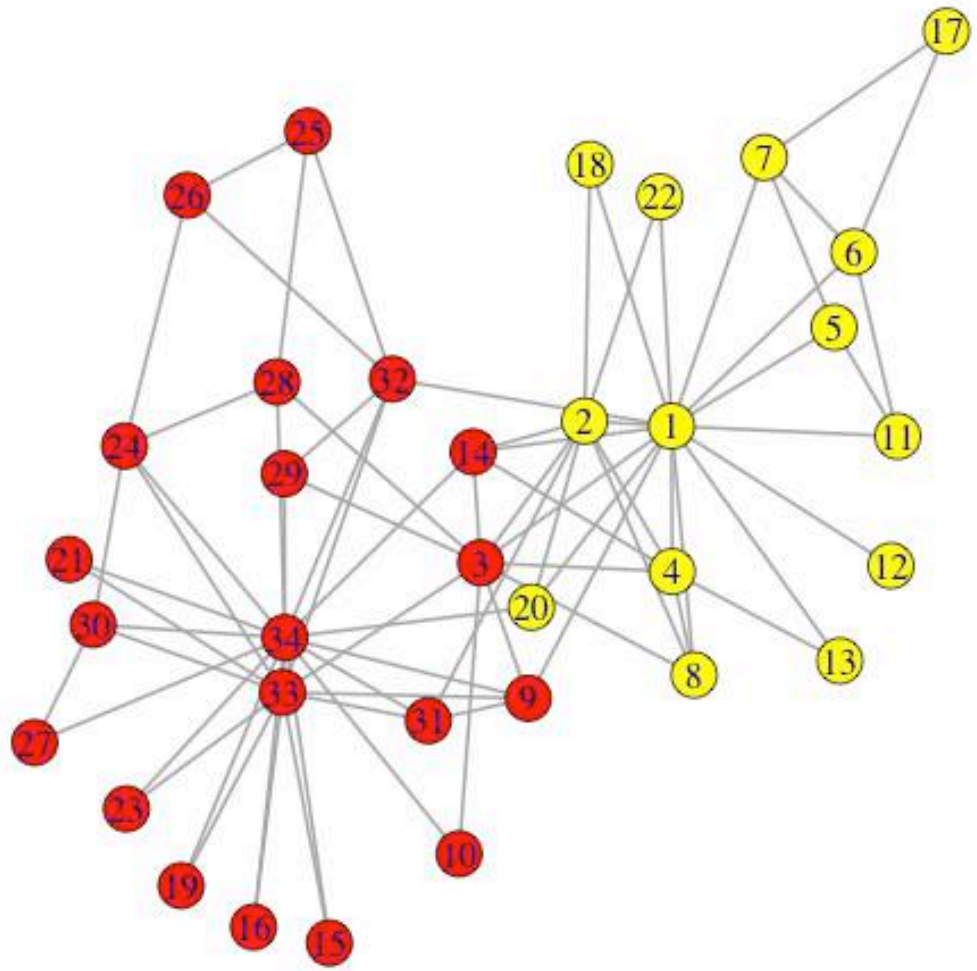
Edge betweenness

Zachary karate club



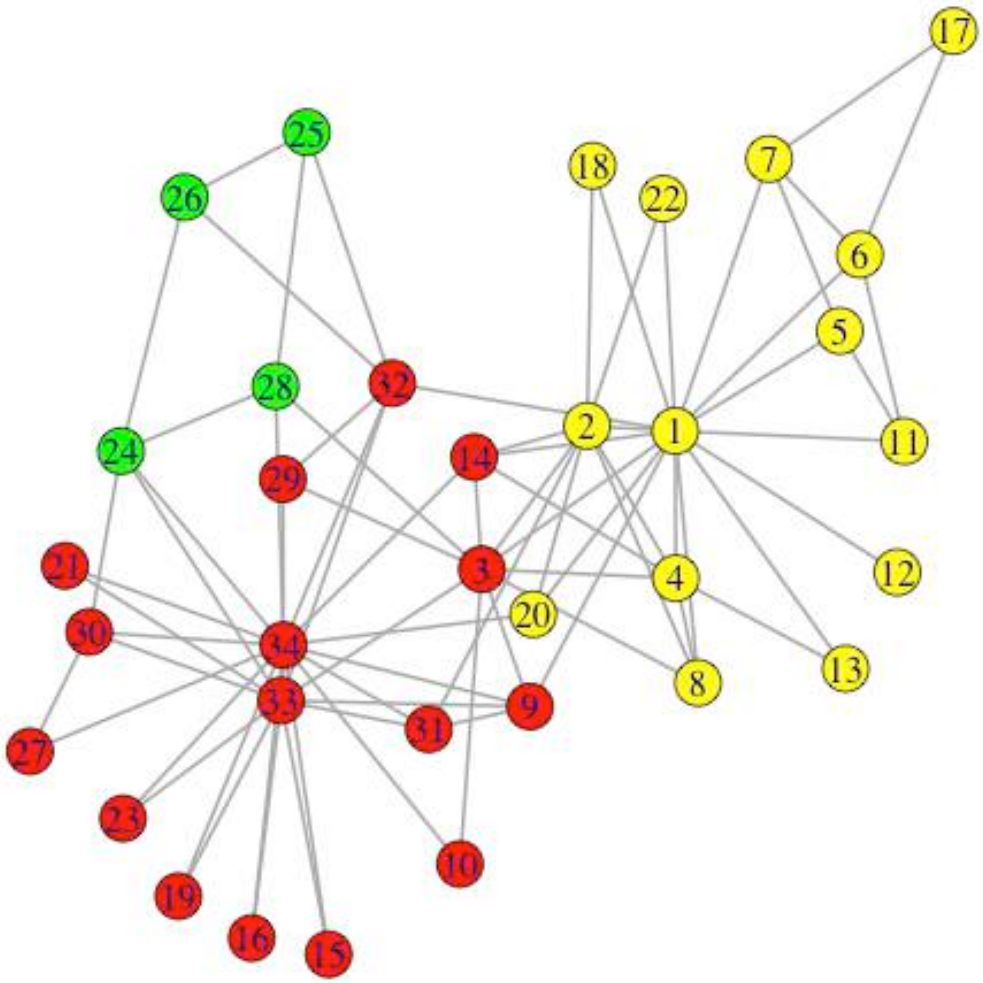
Edge betweenness

Zachary karate club

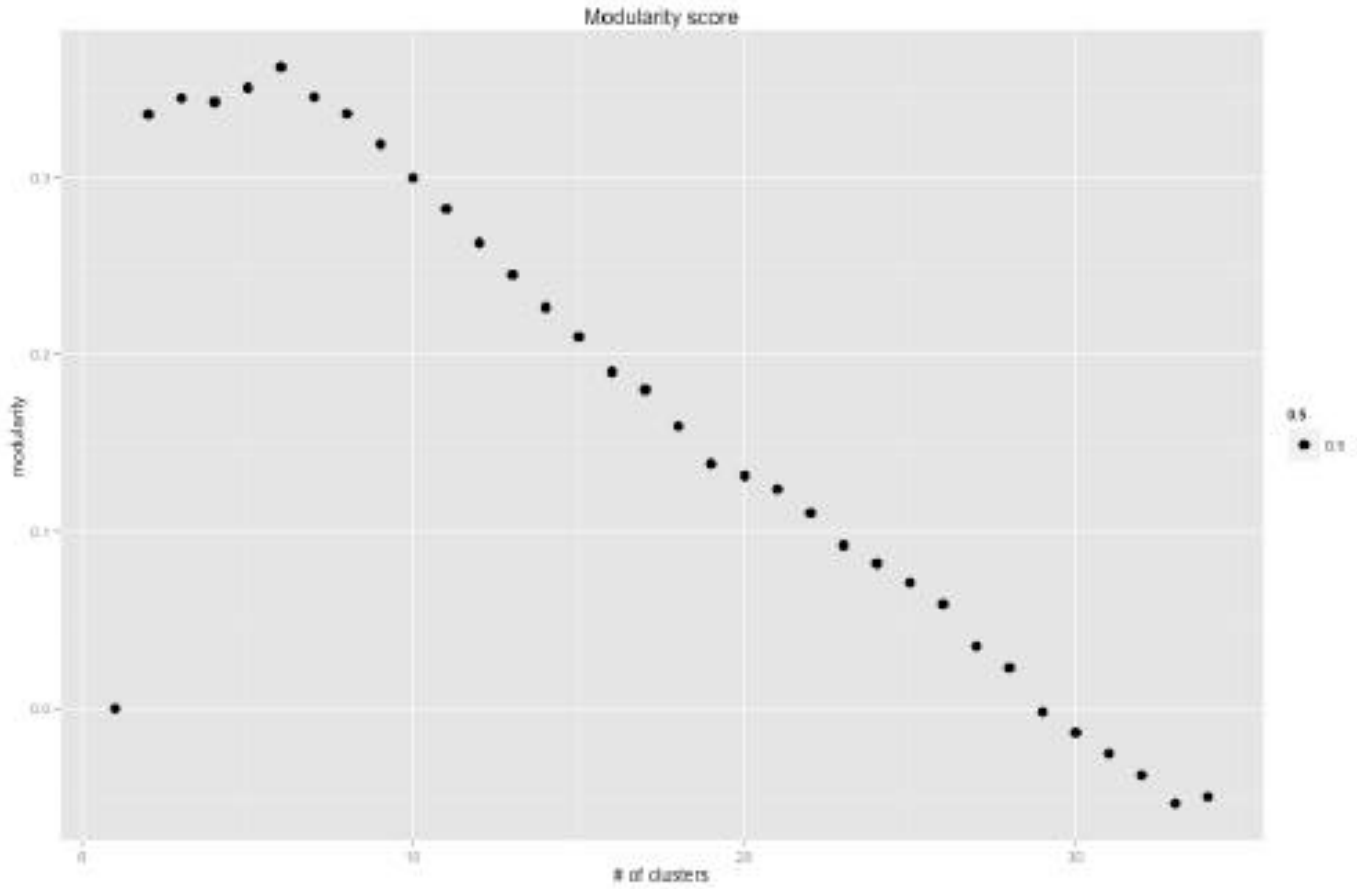


Edge betweenness

Zachary karate club



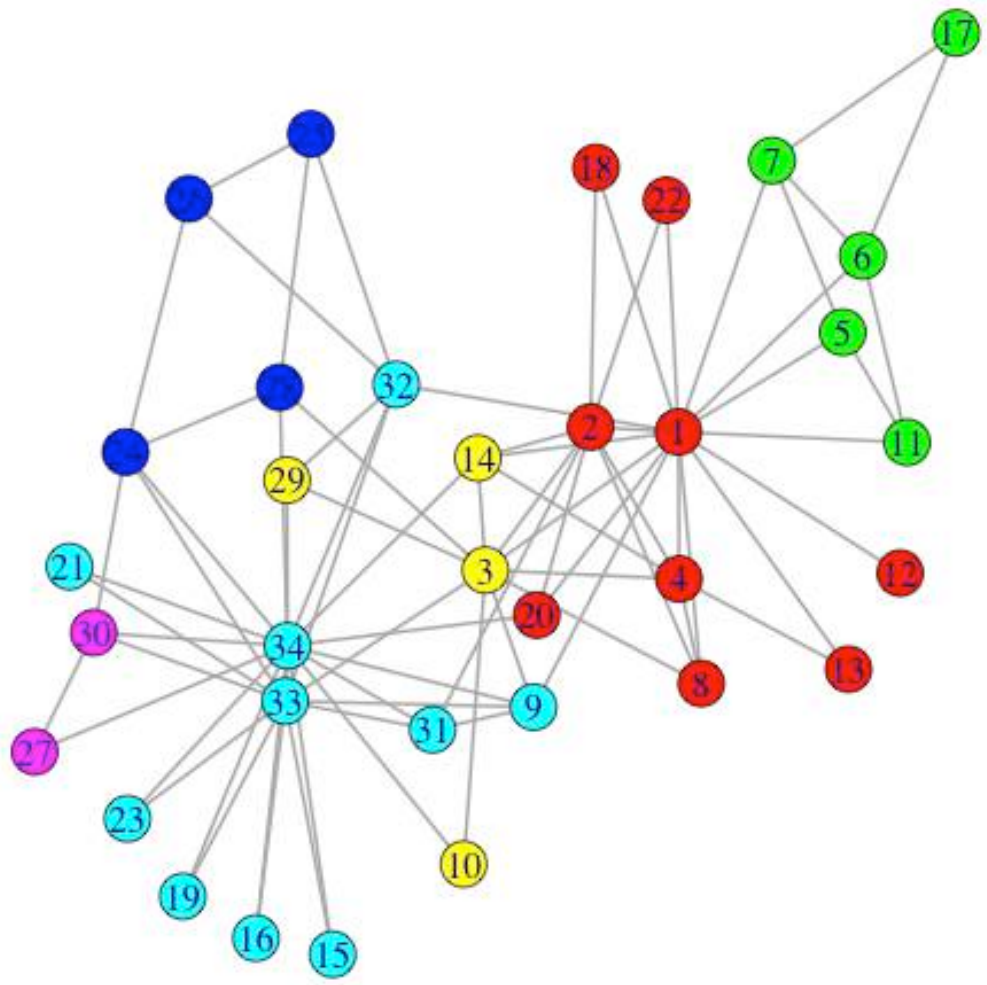
Edge betweenness



igraph:modularity()

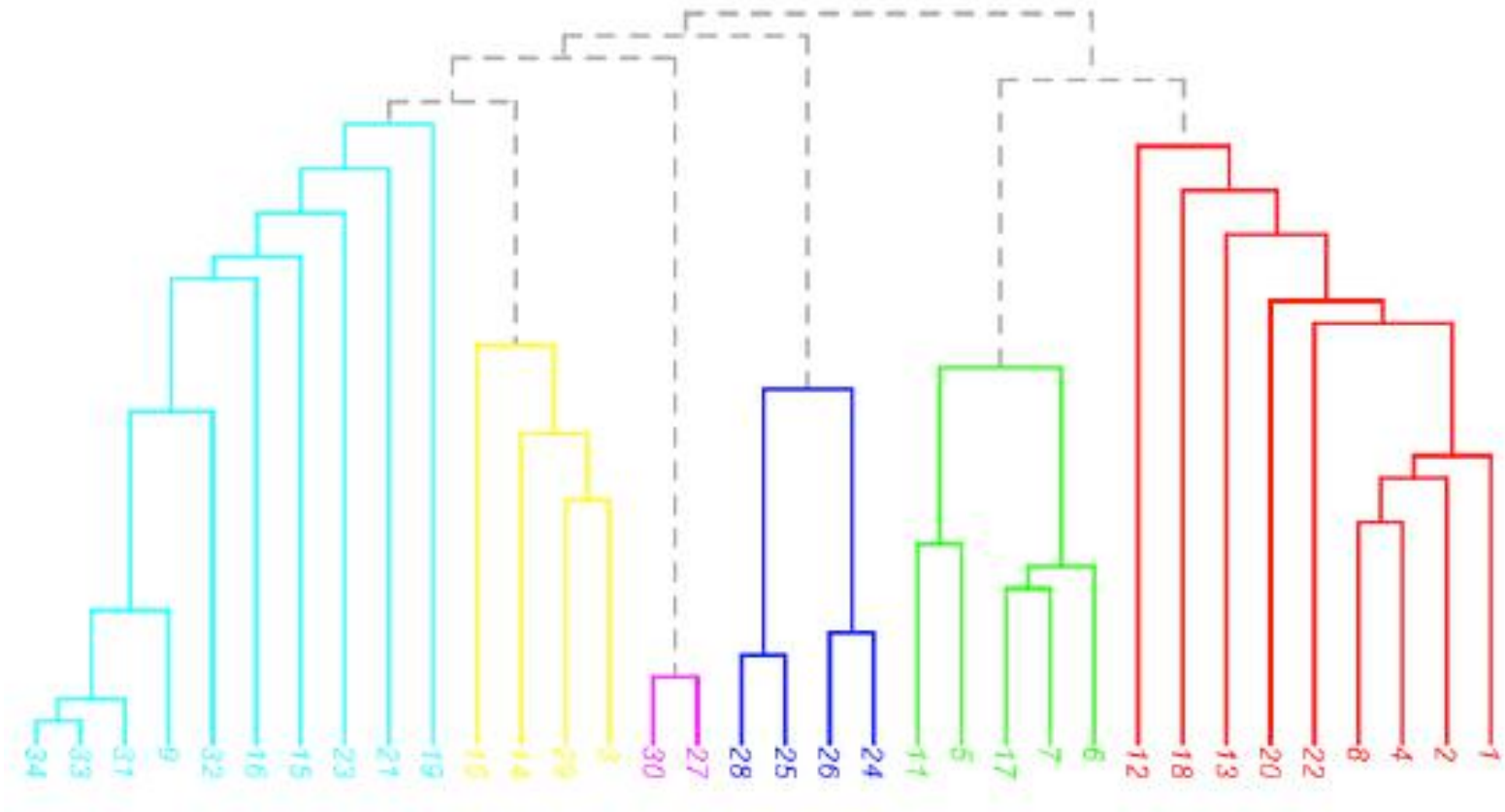
Edge betweenness

best: clusters = 6, modularity = 0.345



Edge betweenness

Zachary karate club



igraph:dendPlot()

Fast community unfolding

V.D. Blondel, J.-L. Guillaume, R. Lambiotte, E. Lefebvre, 2008 "The Louvain method"

- Heuristic method for greedy modularity optimization
- Find partitions with high modularity
- Multi-level (multi-resolution) hierarchical scheme
- Scalable

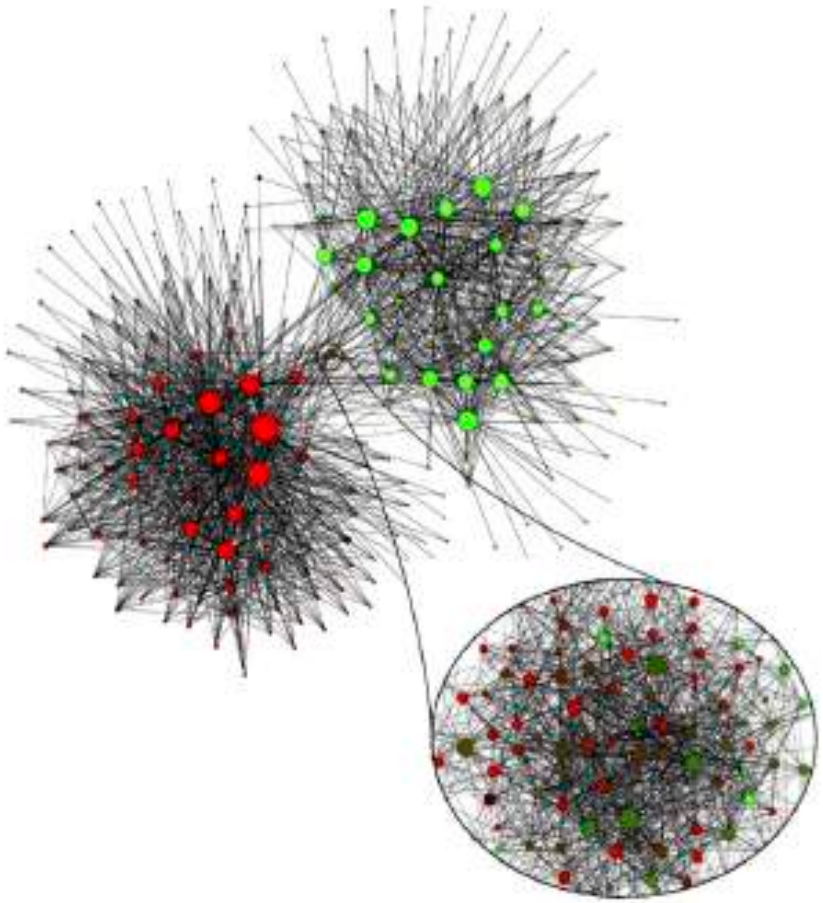
Modularity:

$$Q = \frac{1}{2m} \sum_{i,j} \left(A_{ij} - \frac{k_i k_j}{2m} \right) \delta(c_i, c_j)$$

V. Blondel et.al., 2008

Fast community unfolding

Multi-resolution scalable method



2 mln mobile phone network

V. Blondel et.al., 2008

Fast community unfolding algorithm

Input: Graph G(V,E)

Output: Communities

Assign every node to its own community;

repeat

  repeat

    For every node evaluate modularity gain from removing node from its community and placing it in the community of its neighbor;

    Place node in the community maximizing modularity gain;

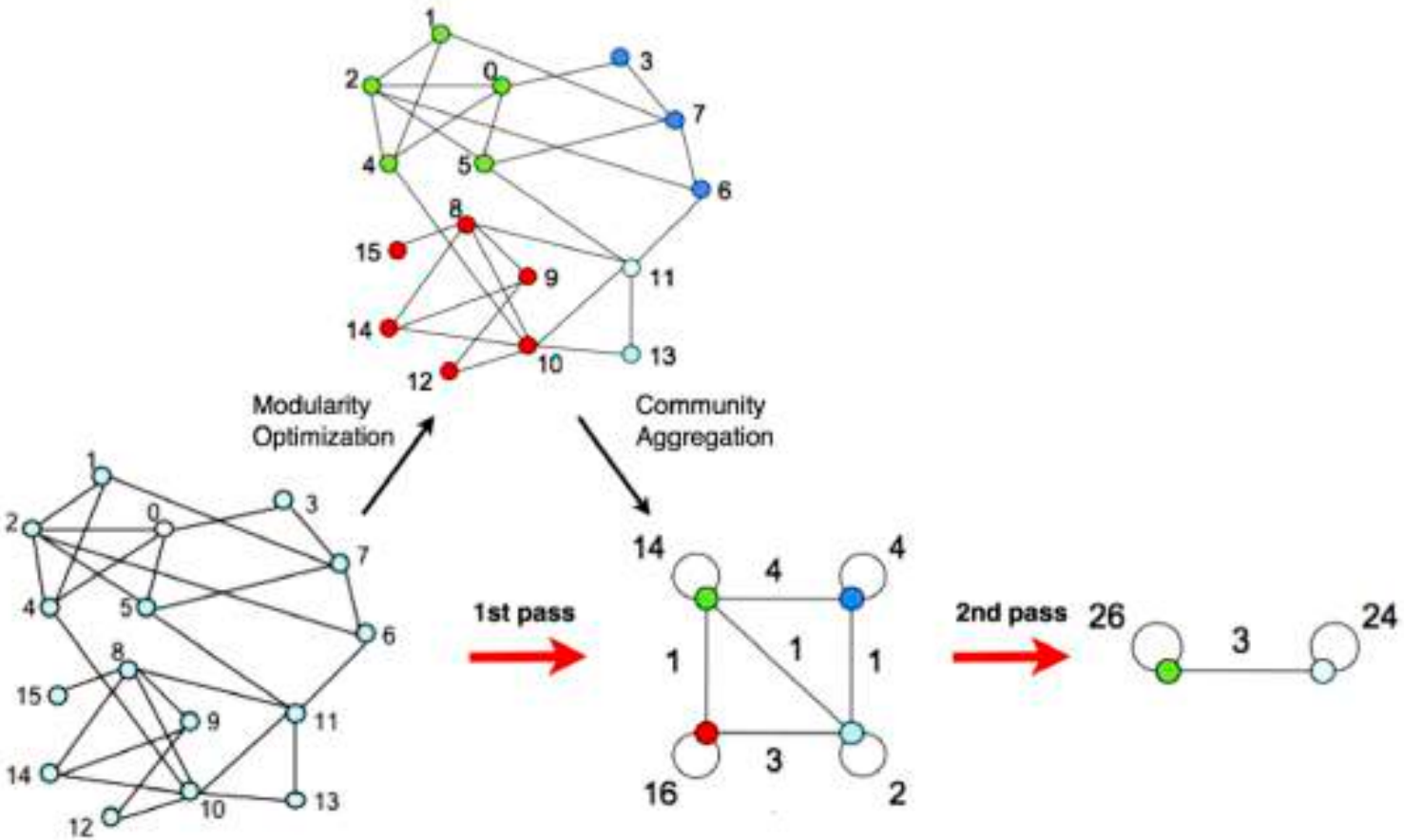
  until *no more improvement (local max of modularity)*;

  Nodes from communities merged into "super nodes" ;

  Weight on the links added up

until *no more changes (max modularity)*;

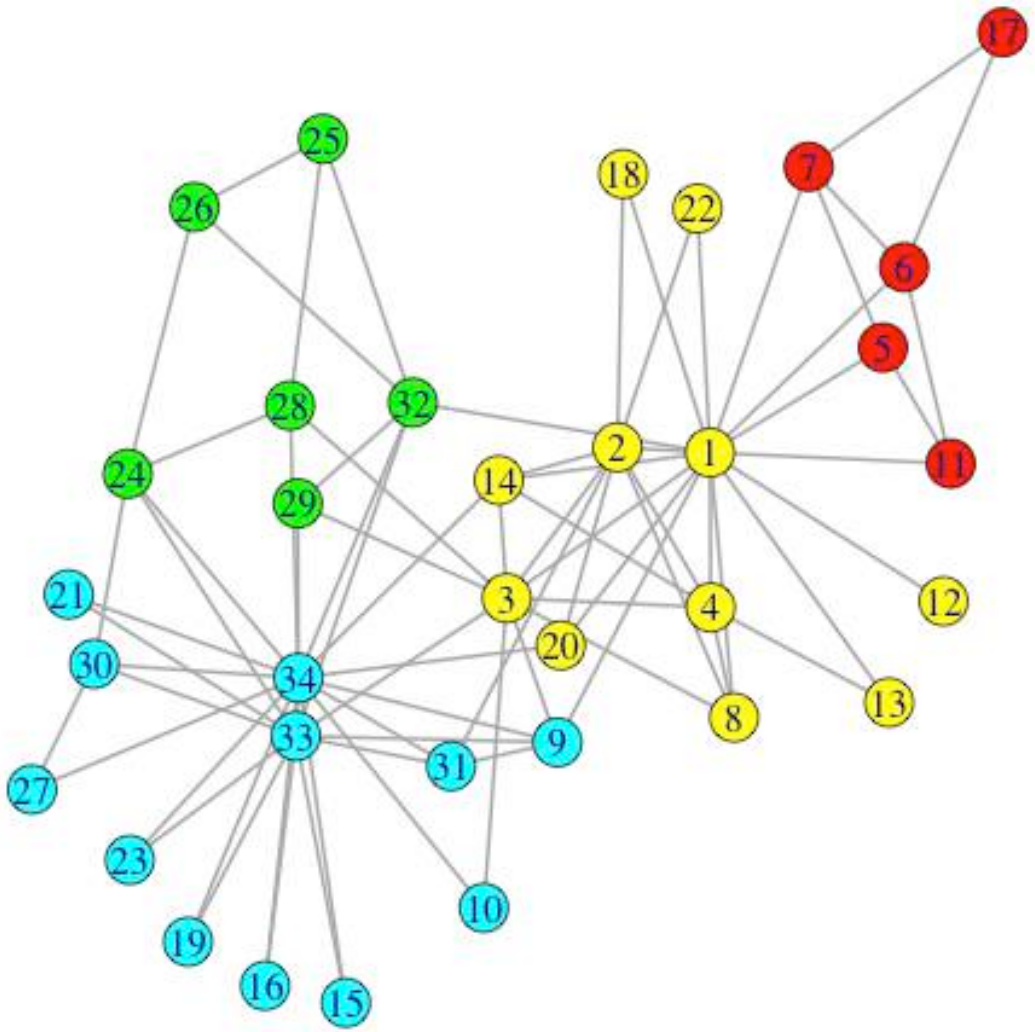
Fast community unfolding



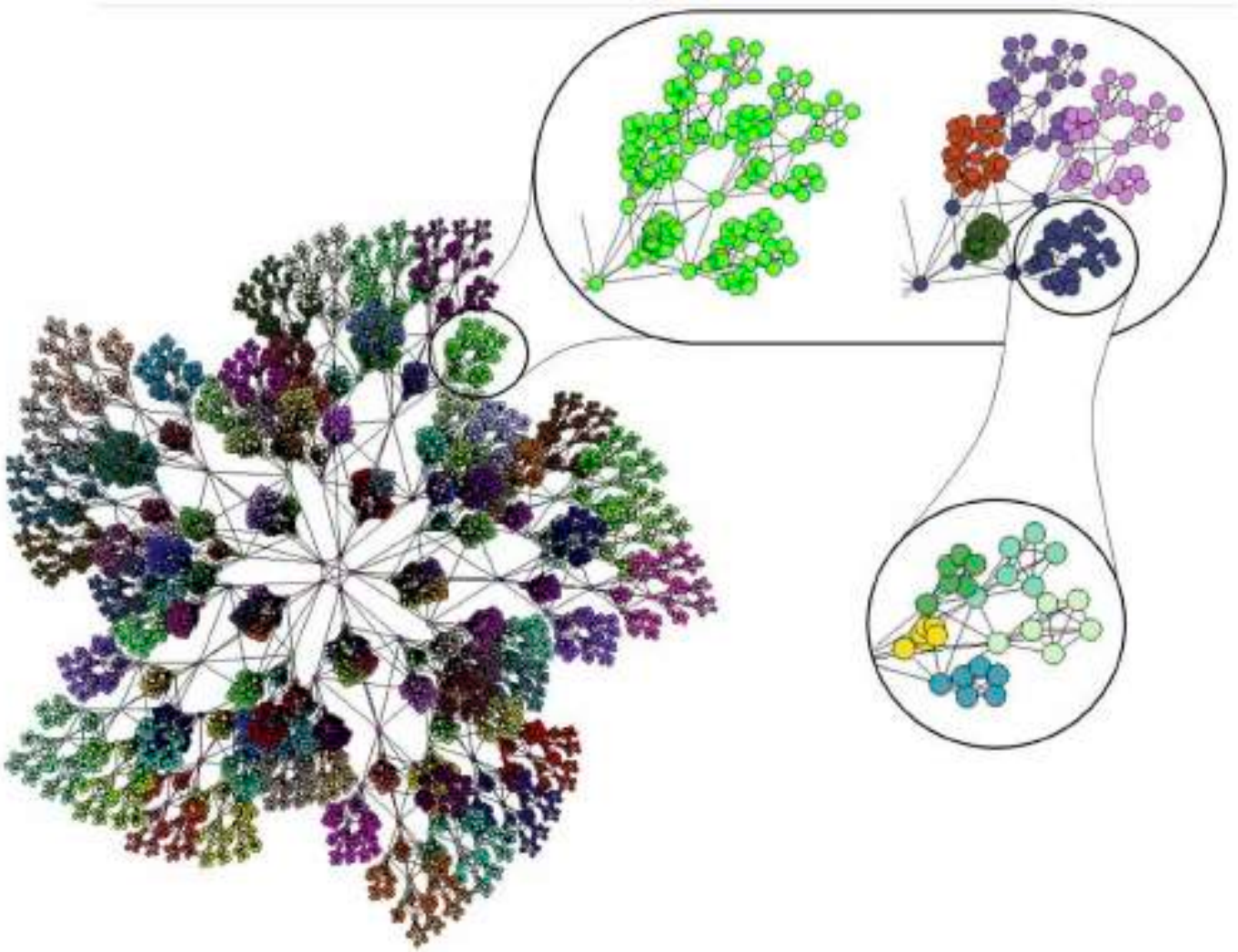
V. Blondel et.al., 2008

Fast community unfolding

clusters = 4, modularity = 0.445



Fast community unfolding



V. Blondel et.al., 2008

References

- S. Fortunato. Community detection in graphs, Physics Reports, Vol. 486, Iss. 3–5, pp 75-174, 2010
- S. E. Schaeffer. Graph clustering. Computer Science Review, 1(1):27–64, 2007.
- Modularity and community structure in networks, M.E.J. Newman, PNAS, vol 103, no 26, pp 8577-8582, 2006
- Finding and evaluating community structure in networks, M.E.J. Newman, M. Girvan, Phys. Rev E, 69, 2004
- U.N. Raghavan, R. Albert, S. Kumara, Near linear time algorithm to detect community structures in large-scale networks, Phys. Rev. E 76 (3) (2007) 036106.
- G. Palla, I. Derenyi, I. Farkas, T. Vicsek, Uncovering the overlapping community structure of complex networks in nature and society, Nature 435 (2005) 814?818.
- P. Pons and M. Latapy, Computing communities in large networks using random walks, Journal of Graph Algorithms and Applications, 10 (2006), 191-218.
- V.D. Blondel, J.-L. Guillaume, R. Lambiotte, E. Lefebvre, Fast unfolding of communities in large networks, J. Stat. Mech. P10008 (2008).

NATIONAL RESEARCH
UNIVERSITY

Network structure and visualization

Social Network Analysis. MAGoLEGO course. Lecture 6

Leonid Zhukov

lzhukov@hse.ru
www.leonidzhukov.net/hse/2018/sna

National Research University Higher School of Economics
School of Data Analysis and Artificial Intelligence, Department of Computer Science

Typical network structure

Core-periphery structure of a network

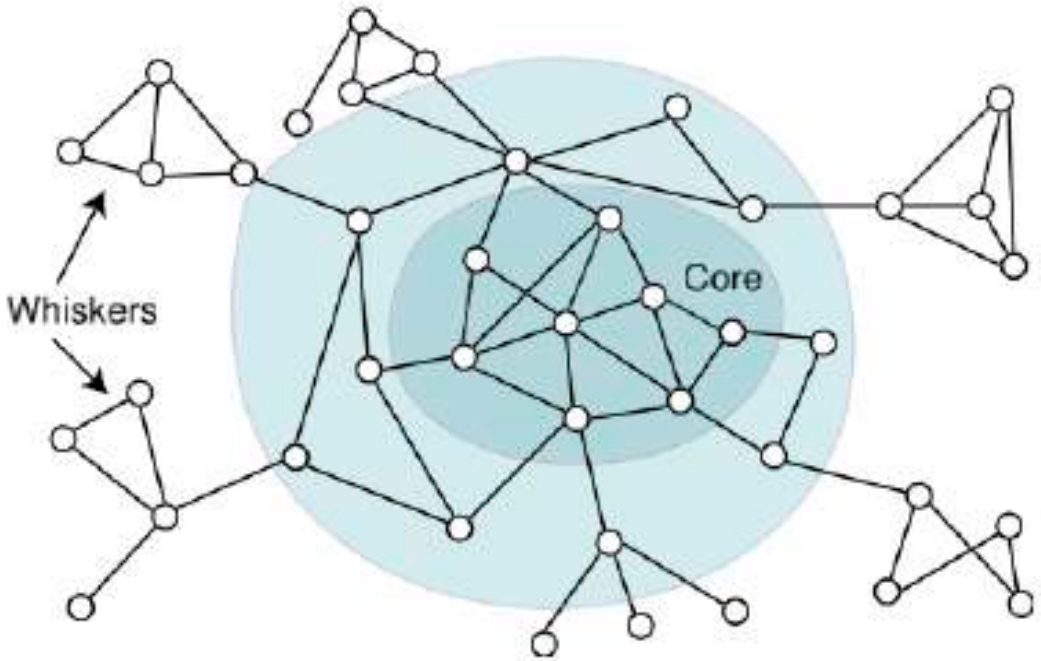
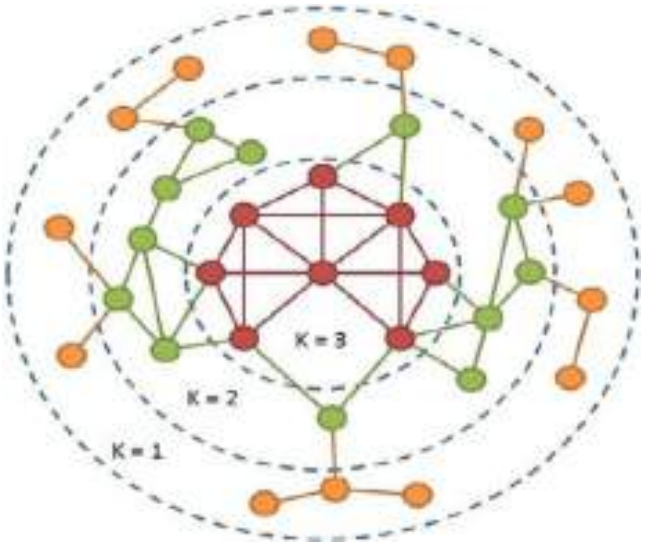


image from J. Leskovec, K. Lang, 2010

k-core decomposition

Definition

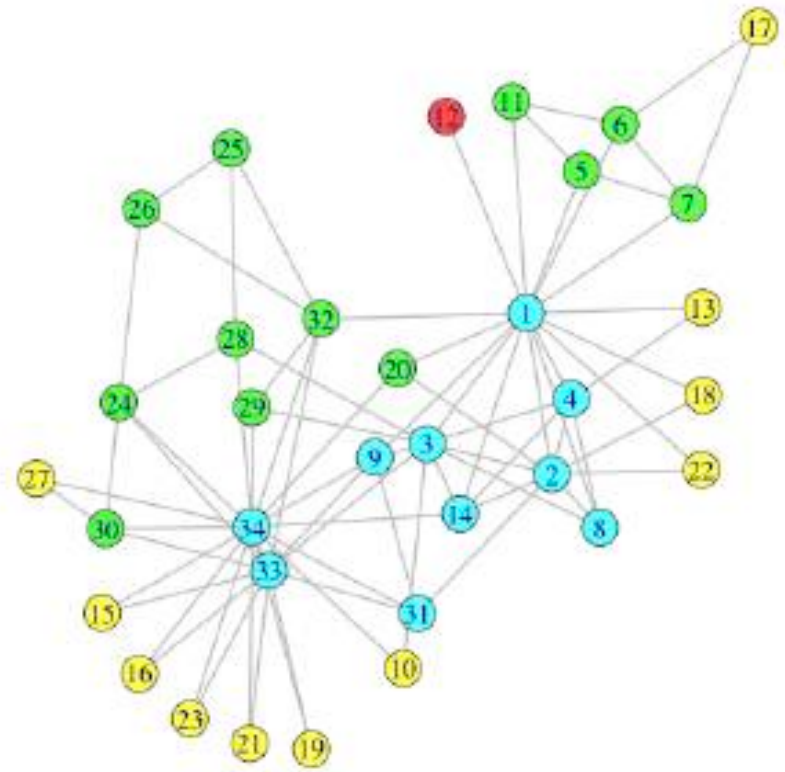
If from a given graph G = (V, E) recursively delete all vertices, and lines incident with them, of degree less than k, the remaining graph is the k-core.



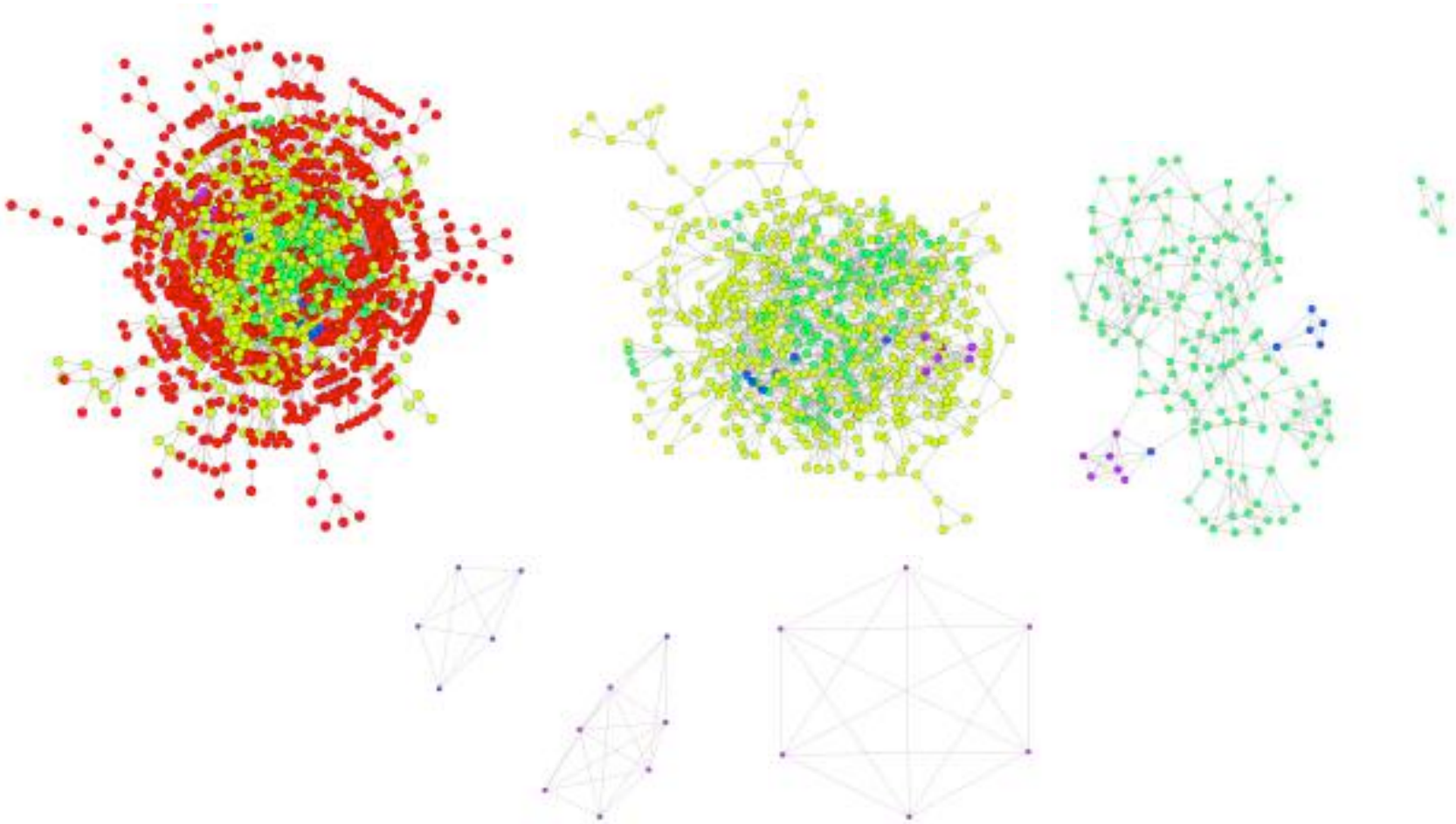
Every vertex in *k*-core has a degree $k_i \geq k$,
$(k+1)$-core is always subgraph of *k*-core
The core number of a vertex is the highest order of a core that contains this vertex

K-cores

Zachary karate club: 1,2,3,4 - cores



k-cores



k-cores: 1:1458, 2:594, 3:142, 4:12, 5:6

k-shells: 1:864-red, 2:452-pale green, 3:130-green, 5:6-blue,

Mixing patterns

Network mixing patterns

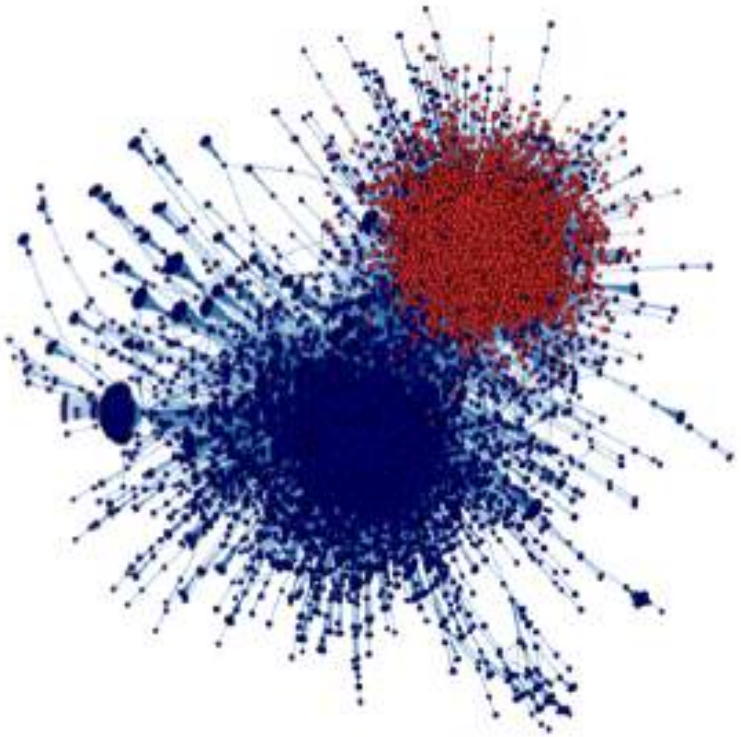
- **Assortative mixing**, ”like links with like”, attributed of connected nodes tend to be more similar than if there were no such edge
- **Disassortative mixing**, ”like links with dislike”, attributed of connected nodes tend to be less similar than if there were no such edge

Vertices can mix on any vertex attributes (age, sex, geography in social networks), unobserved attributes, vertex degrees

Examples:
assortative mixing - in social networks political beliefs, obesity, race
disassortative mixing - dating network, food web (predator/prey), economic networks (producers/consumers)

Assortative mixing

- Political polarization on Twitter: political retweet network ,red color - ”right-learning” users, blue color - ”left learning” users

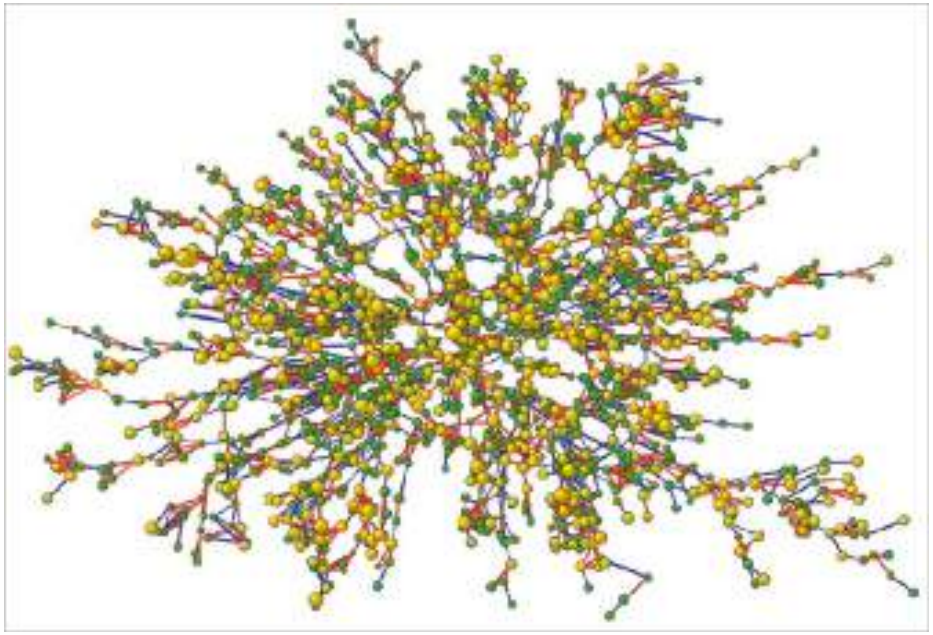


- Assortative mixing = homophily

Conover et al., 2011

Assortative mixing

- The Spread of Obesity in a Large Social Network over 32 Years



Node colors - person's obesity status: yellow denotes an obese person (body-mass index > 30) and green denotes a nonobese person.
Edge colors - relationship between them: purple denotes a friendship or marital tie and orange denotes a familial tie.

Christakis and Fowler, 2007

Assortativity measures

- **Discrete mixing** by categorical attribute (c_i -label: color, gender, ethnicity). How much more often do attributes match across edges than expected at random? Assortativity coefficient:

$$C = \frac{Q}{Q_{max}} = \frac{\sum_{ij}\left(A_{ij} - \frac{k_i k_j}{2m}\right)\delta(c_i, c_j)}{2m - \sum_{ij}\frac{k_i k_j}{2m}\delta(c_i, c_j)}$$

- **Mixing by scalar properties**, scalar value attribute (age, income, number of friends). Correlation of values across edges. Assortativity coefficient:

$$r = \frac{cov}{var} = \frac{\sum_{ij}\left(A_{ij} - \frac{k_i k_j}{2m}\right)x_i x_j}{\sum_{ij}\left(k_i\delta_{ij} - \frac{k_i k_j}{2m}\right)x_i x_j}$$

R igraph: assortativity.nominal(), assortativity()

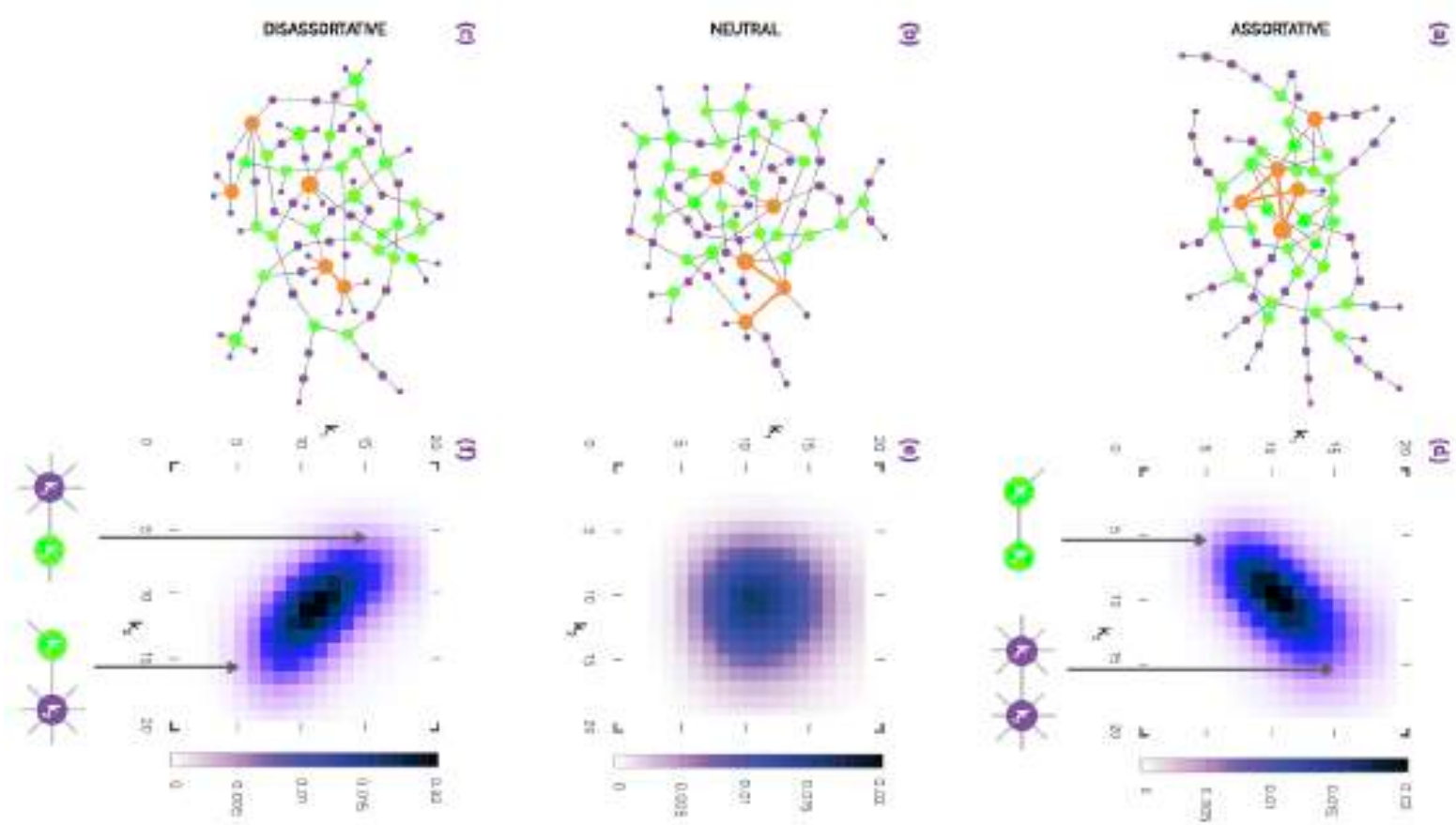
Mixing by node degree

- Assortative mixing by node degree, $x_i \leftarrow k_i - 1$

$$r = \frac{\sum_{ij} \left(A_{ij} - \frac{k_i k_j}{2m} \right) k_i k_j}{\sum_{ij} \left(k_i \delta_{ij} - \frac{k_i k_j}{2m} \right) k_i k_j}$$

R igraph: assortativity.degree()

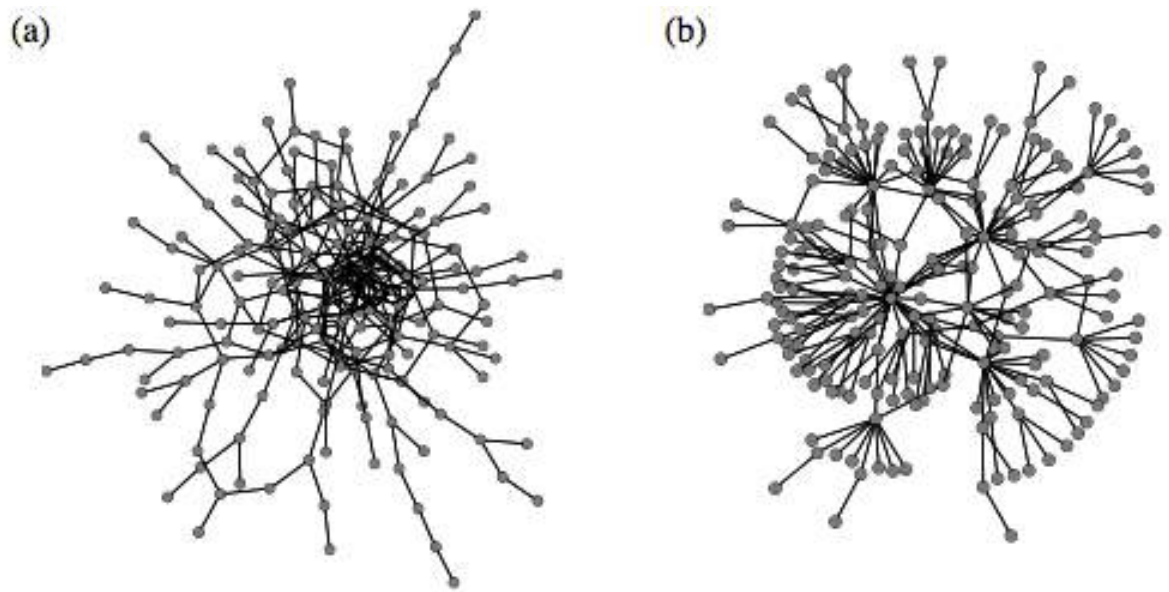
Mixing by node degree



from A-L. Barabasi

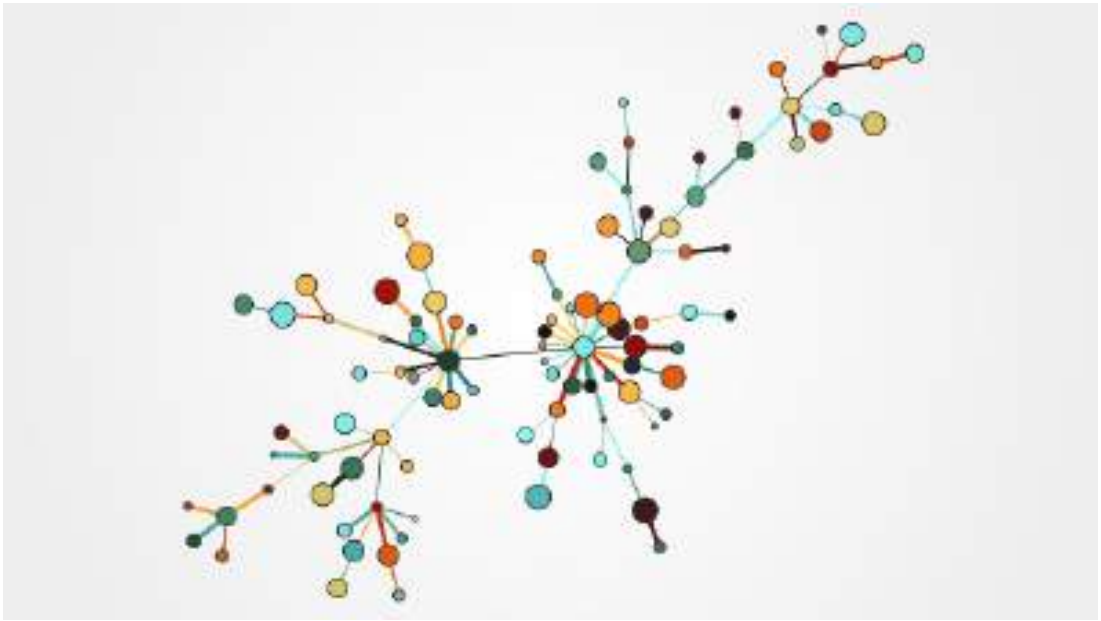
Mixing by node degree

- Assortative network: interconnected high degree nodes - core, low degree nodes - periphery
- Disassortative network: high degree nodes connected to low degree nodes, star-like structure



Assortative network Disassortative network

Network visualization



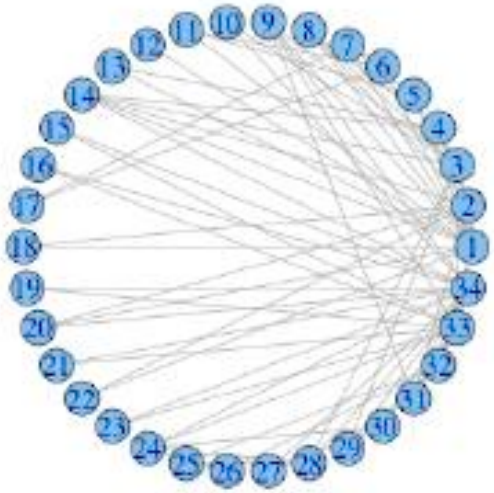
Network:

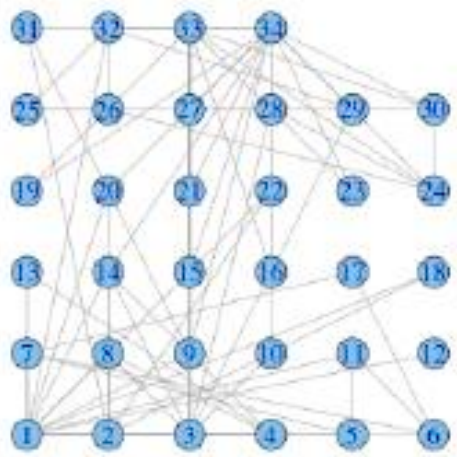
- nodes (+node attributes)
- edges (+edge attributes)

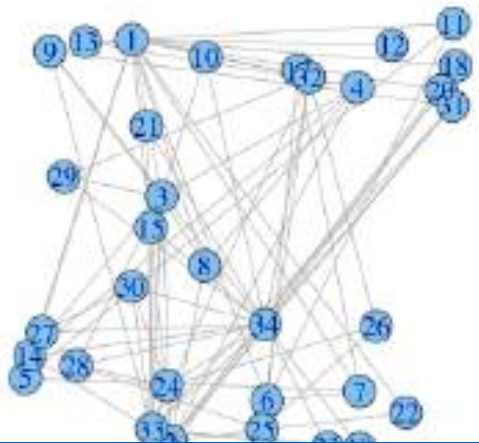
Graph (network) layout:

- nodes coordinates (x,y)

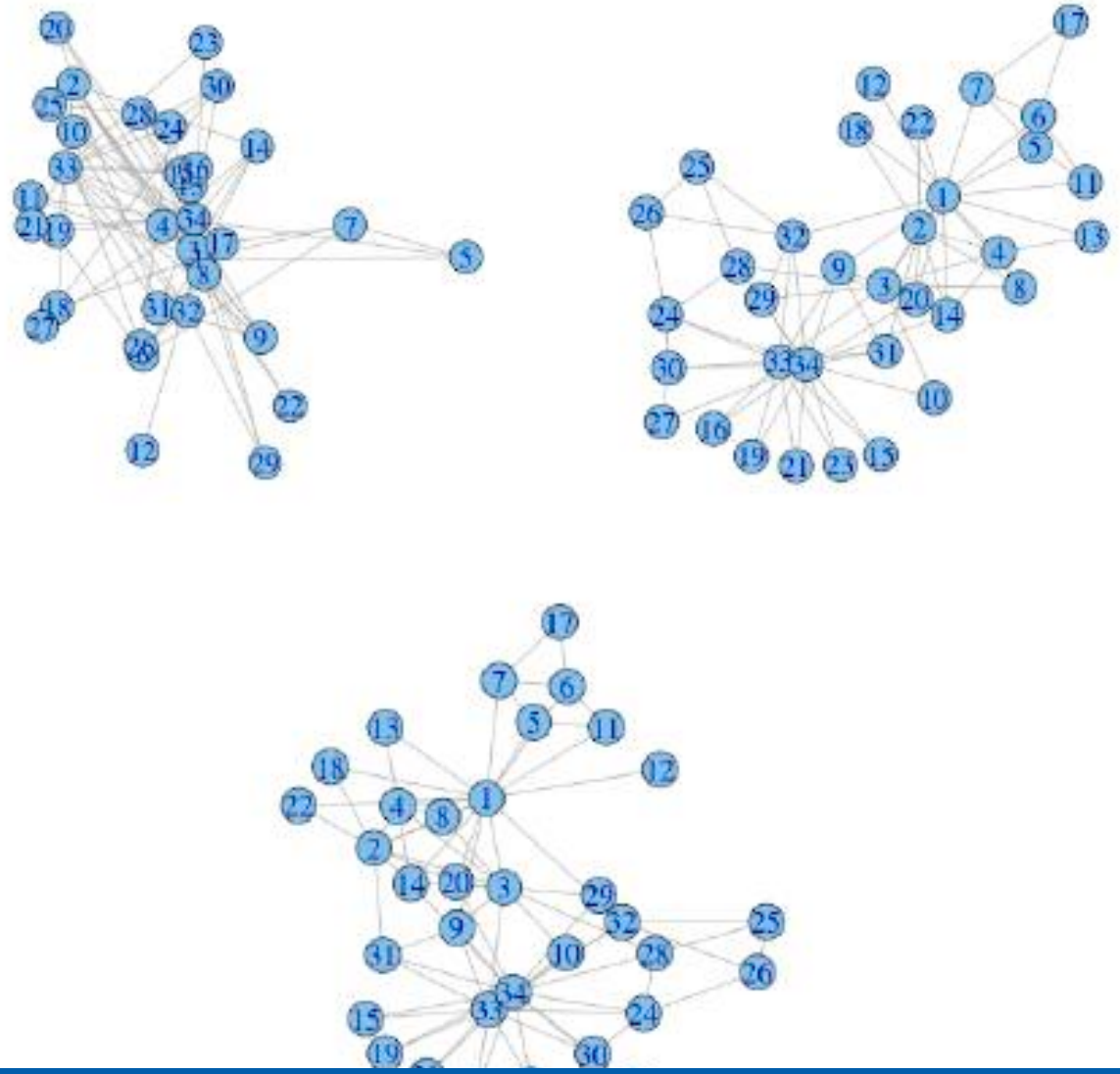
Simple graph layouts



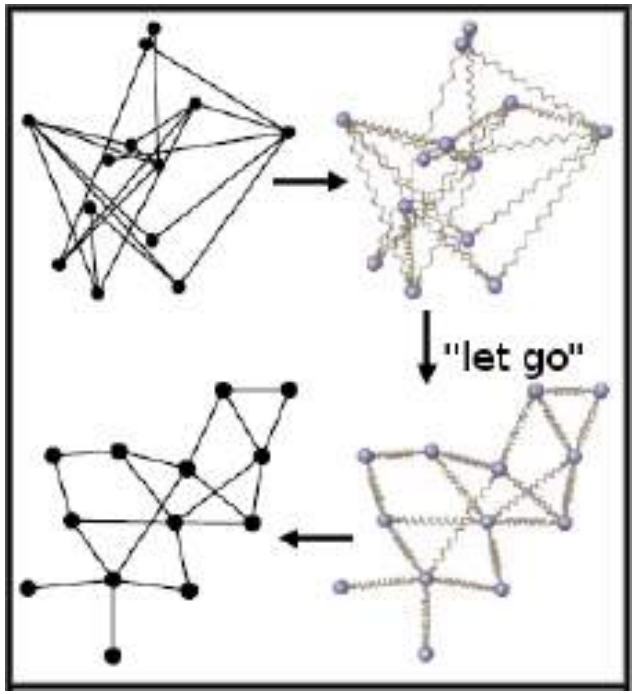




Force-directed layouts



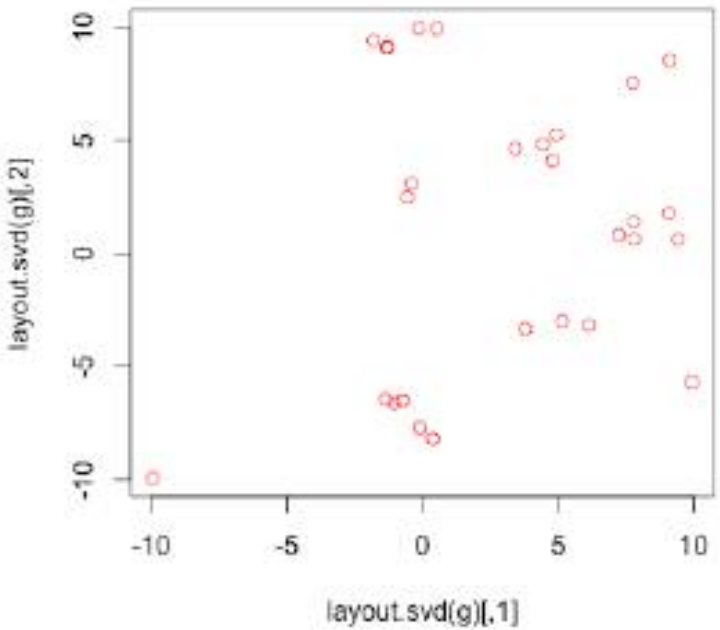
Force-directed layouts

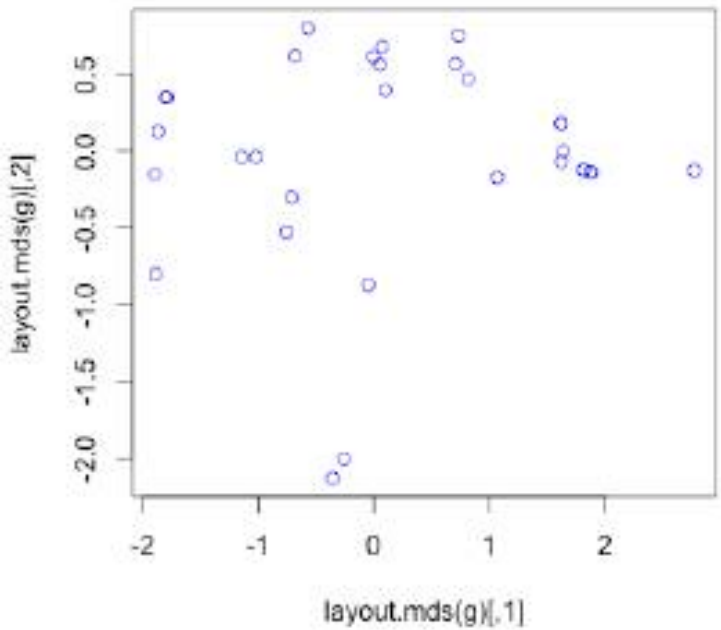


Kamada-Kawai model (stress minimization)

$$stress(X) = \sum_{i<j} w_{ij}(||X_i - X_j|| - d_{ij})^2$$

Low dimensional embeddings

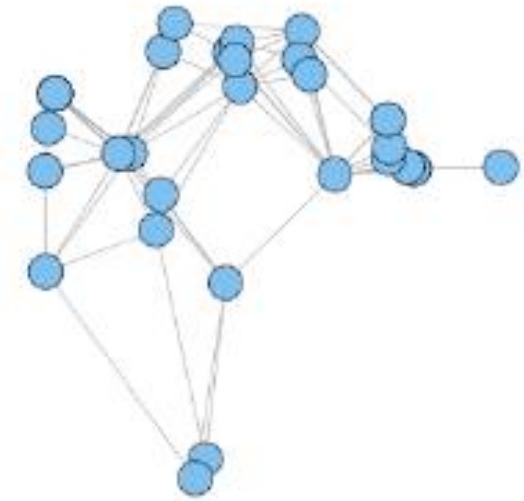


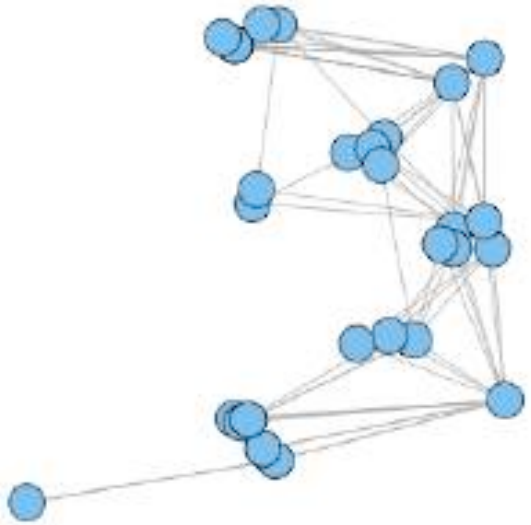


SVD (PCA)

MDS multidimensional scaling

Low dimensional embeddings



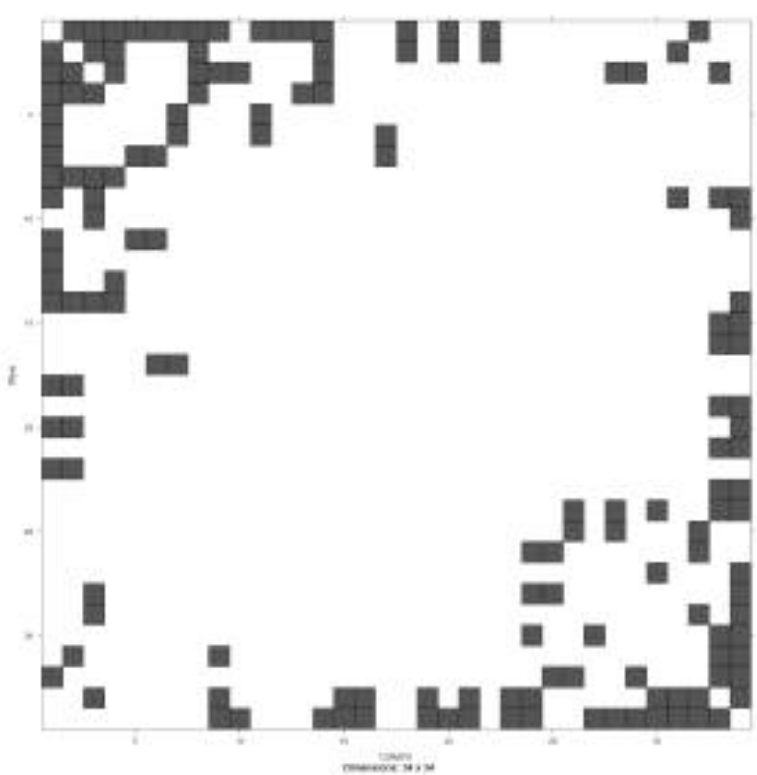


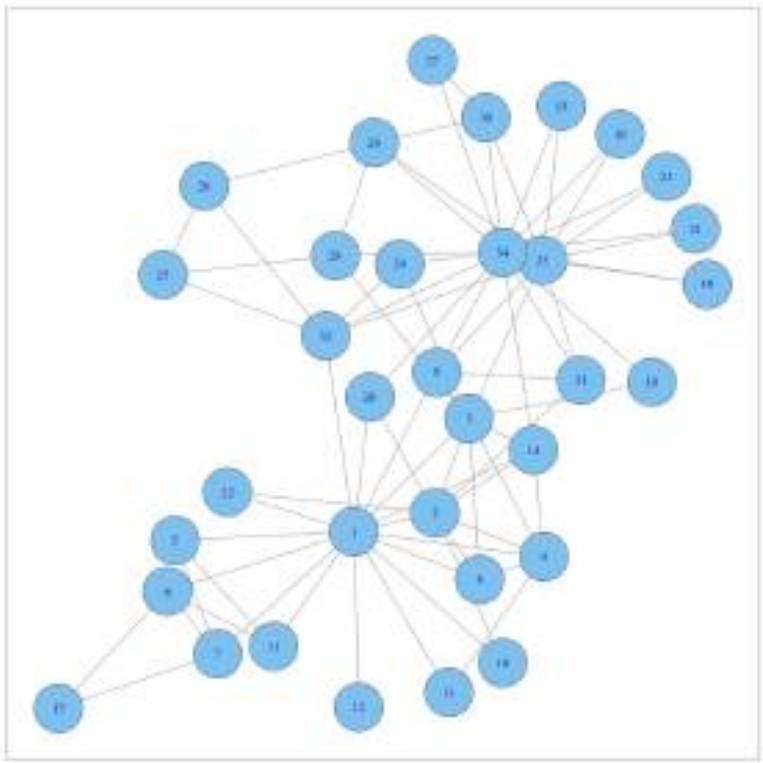
SVD (PCA)

MDS multidimensional scaling

R {igraph}: layout.svd(), layout.mds()

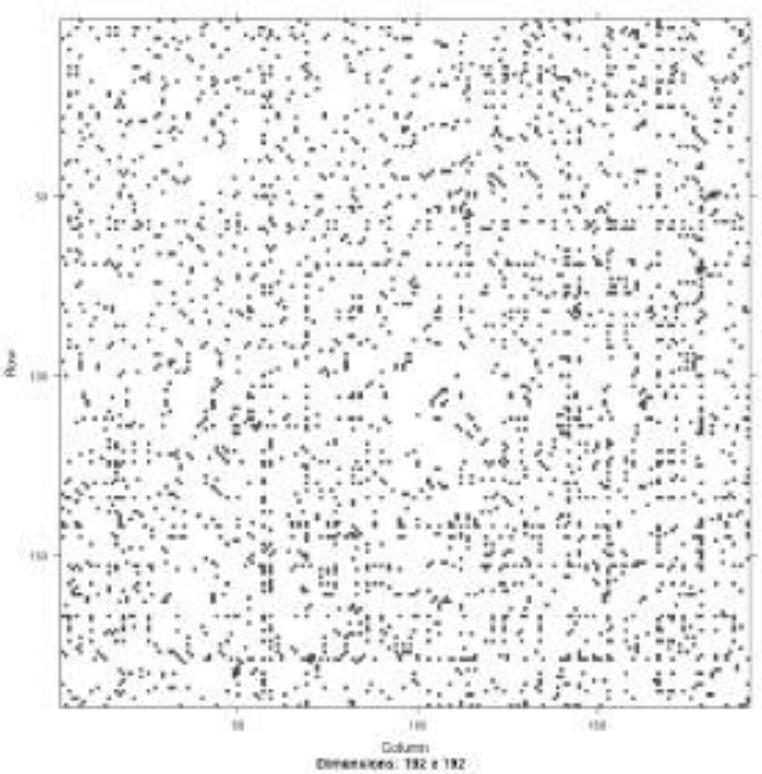
Sparse matrix visualization

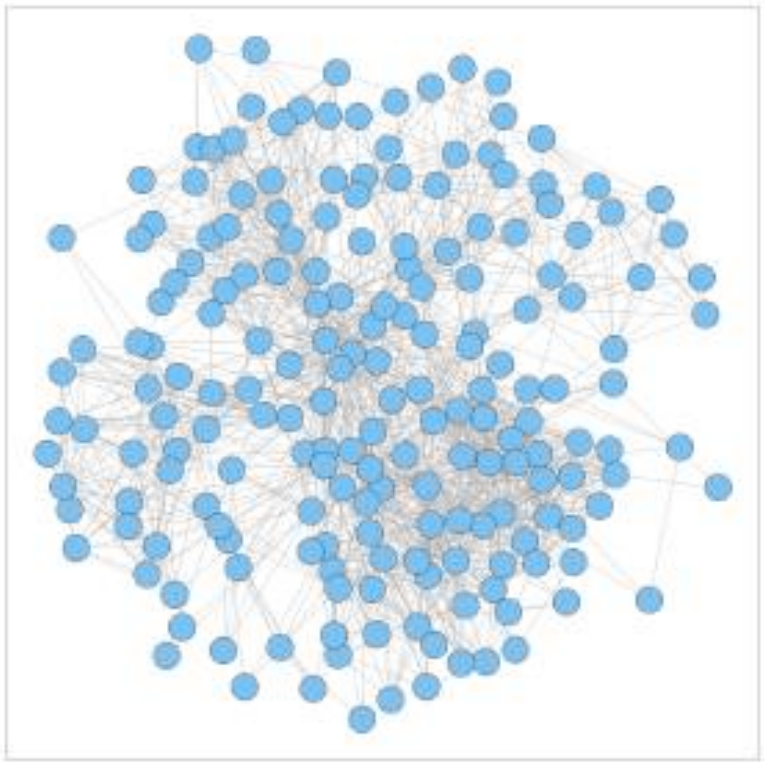




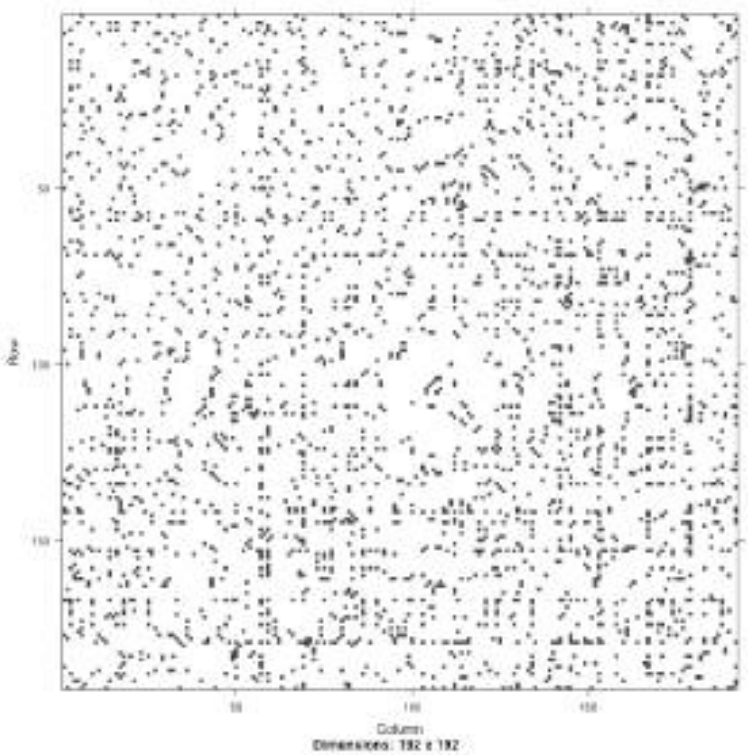
R {SparseM}:image()

Sparse matrix visualization





Sparse matrix visualization

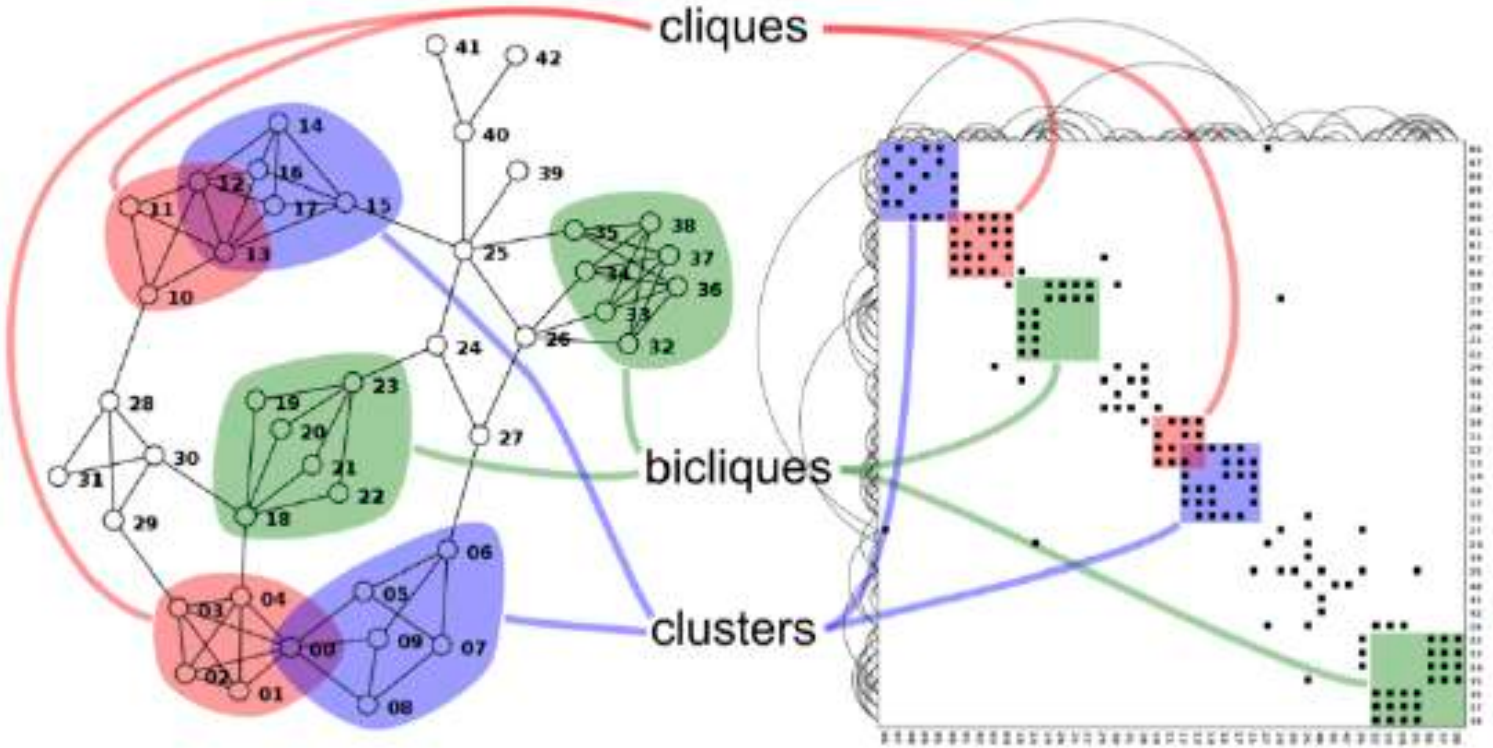


Matrix permutations (bandwidth reduction)

- Minimum degree ordering
- Reverse Cuthill-McKee ordering

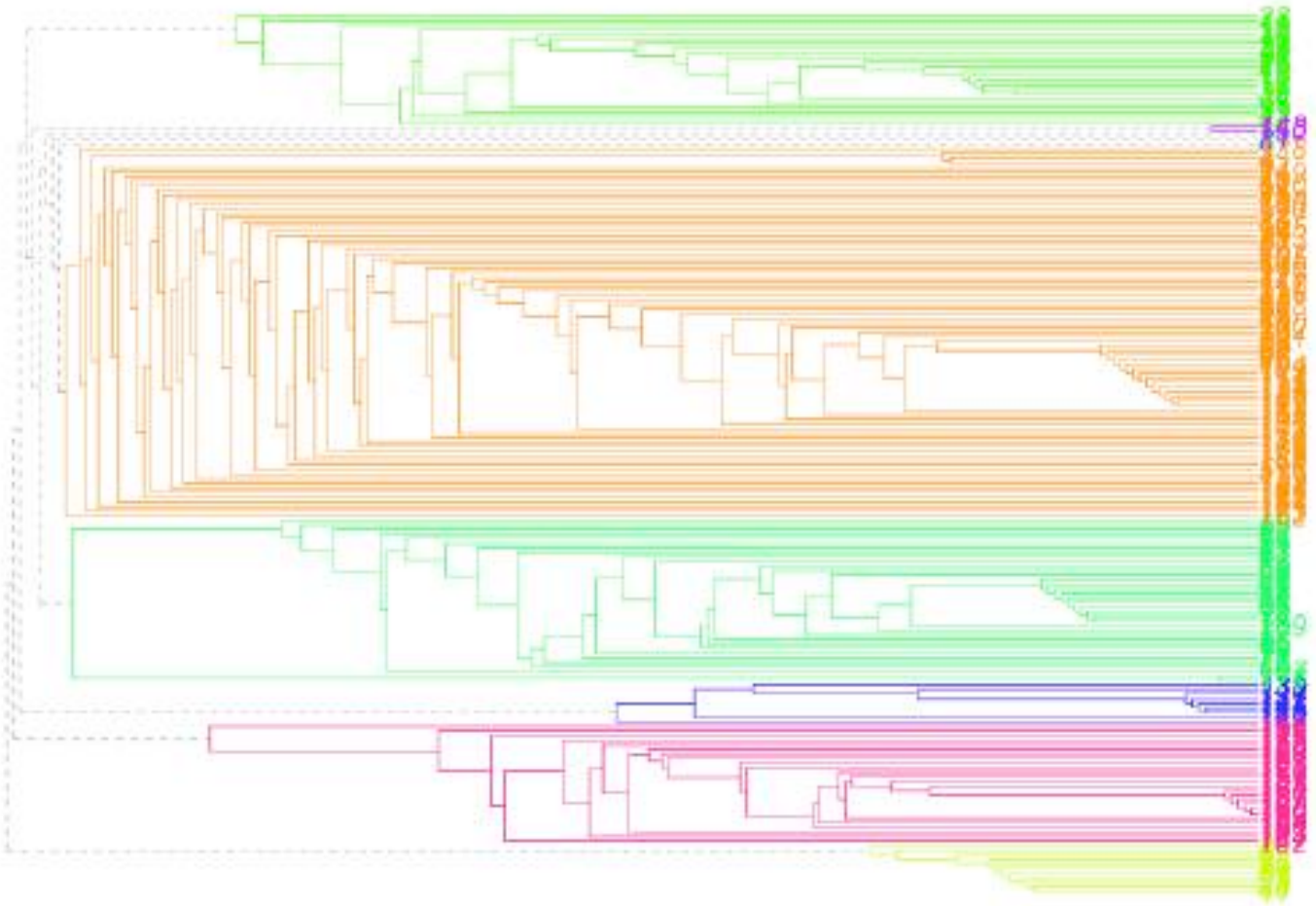
R {RBGL}:minDegreeOrdering(), cuthill.mckee.ordering()

Sparse matrix visualization



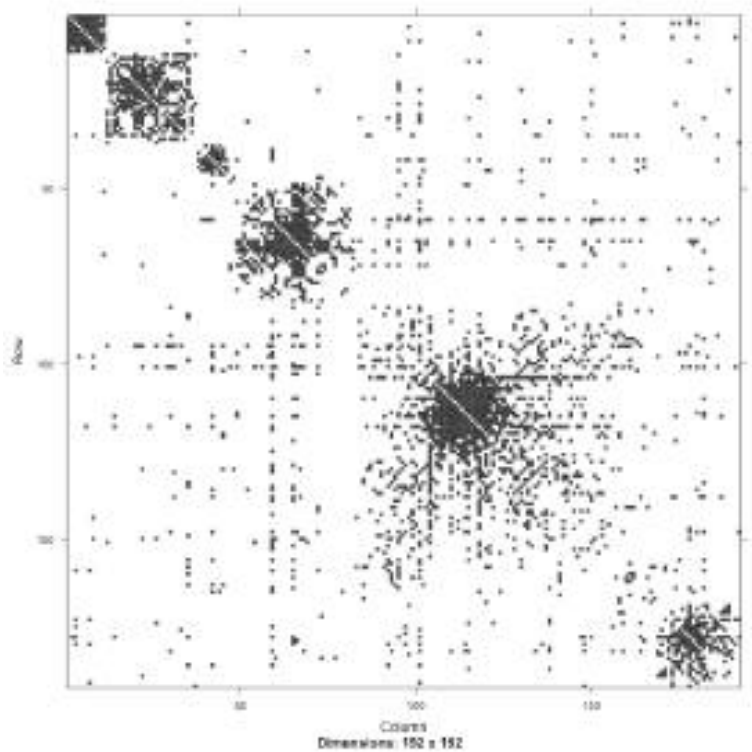
from M.J. McGuffin

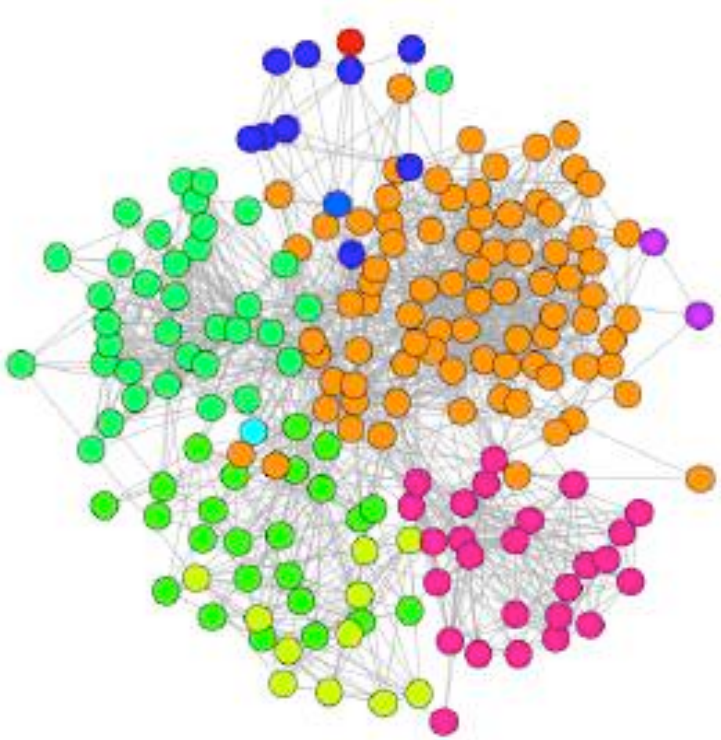
Hierarchical clustering



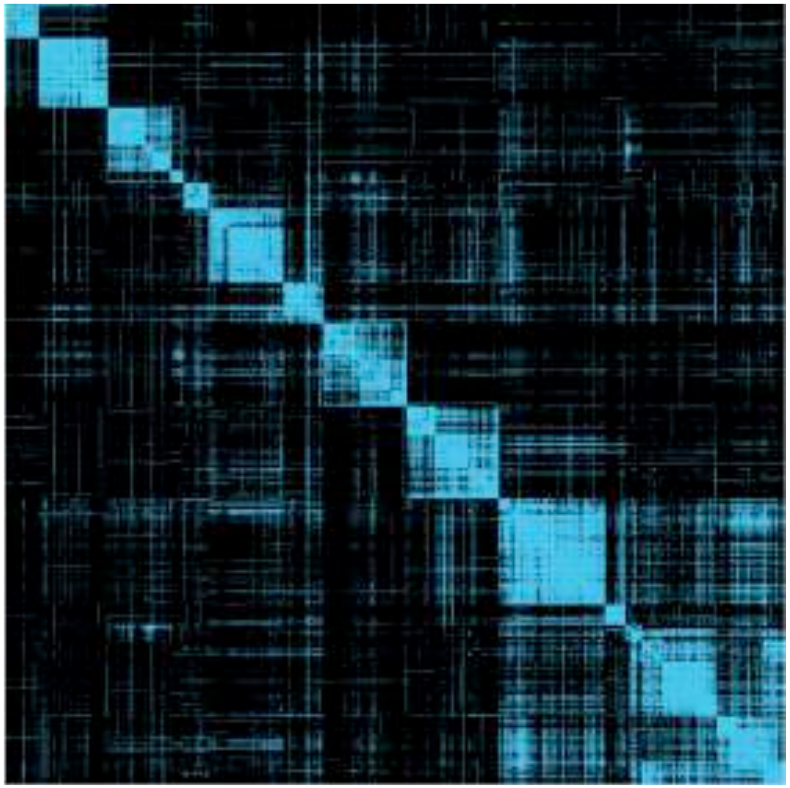
R {igraph}: dendPlot

Sparse matrix visualization





Sparse matrix visualization



Visualization tools

- Graphviz (http://www.graphviz.org)
- Gephi (http://www.gephi.org)
- yEd (http://www.yworks.com)
- Visone (http://www.visone.net)
- Pajek (http://pajek.imfm.si)

yEd

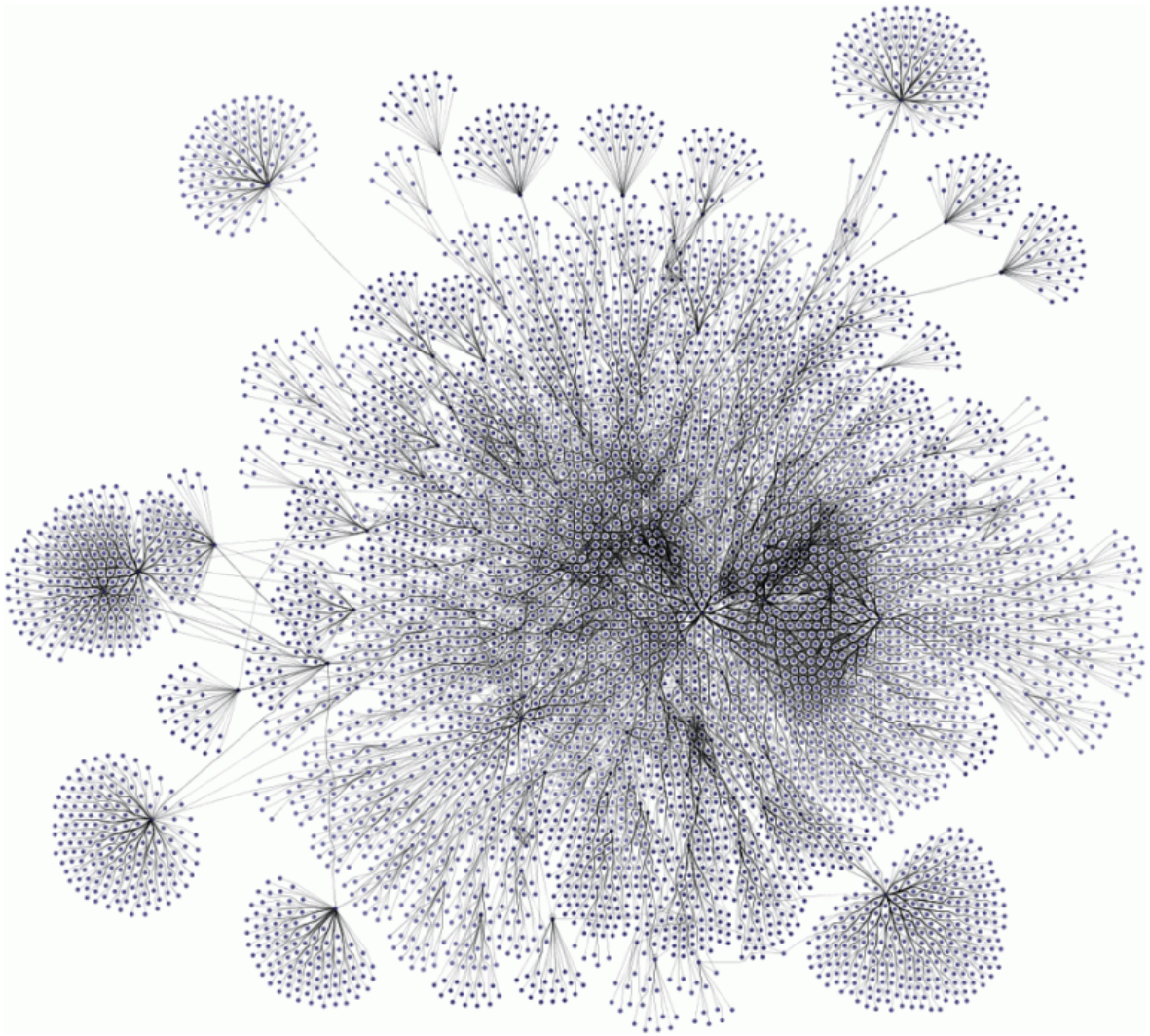
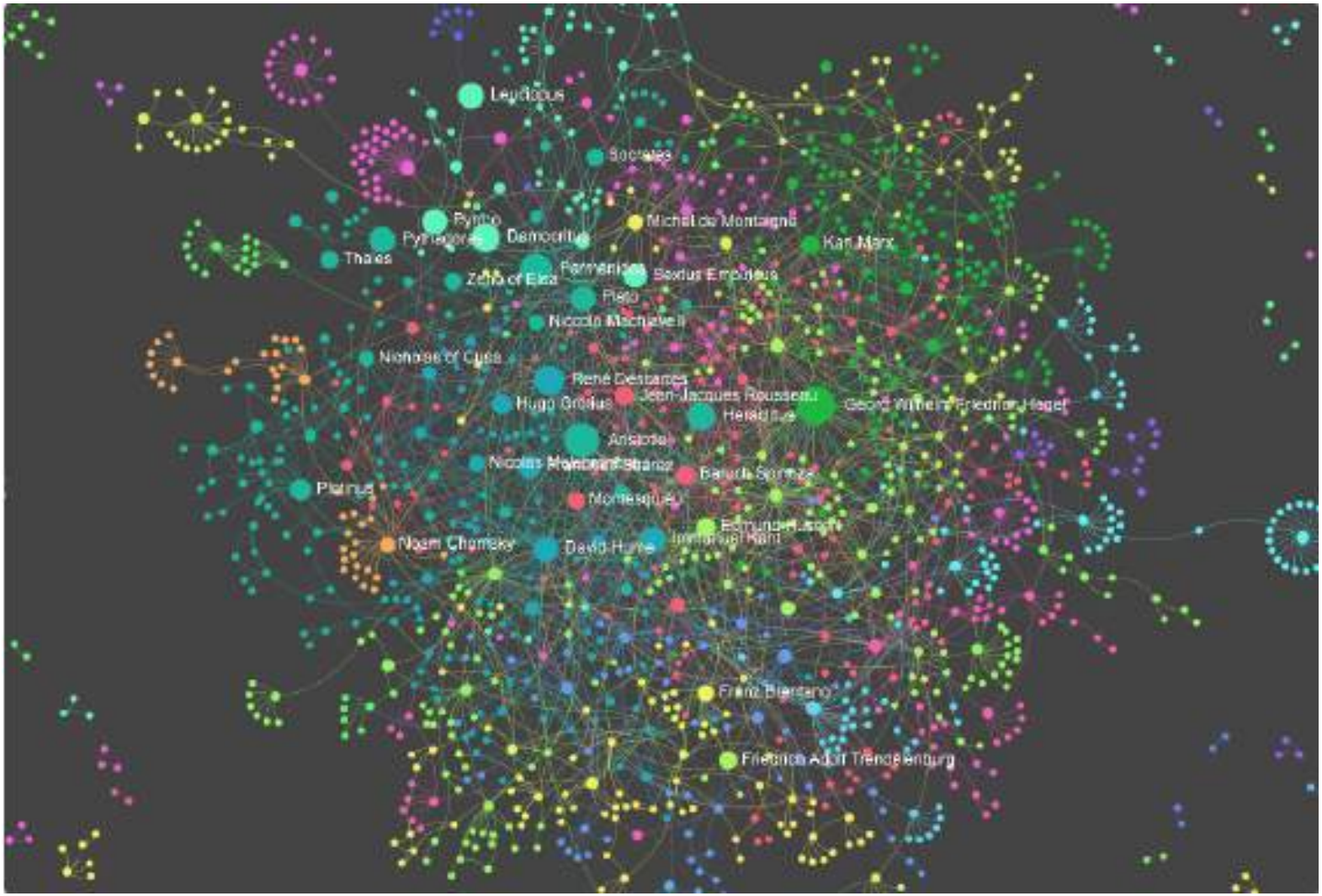
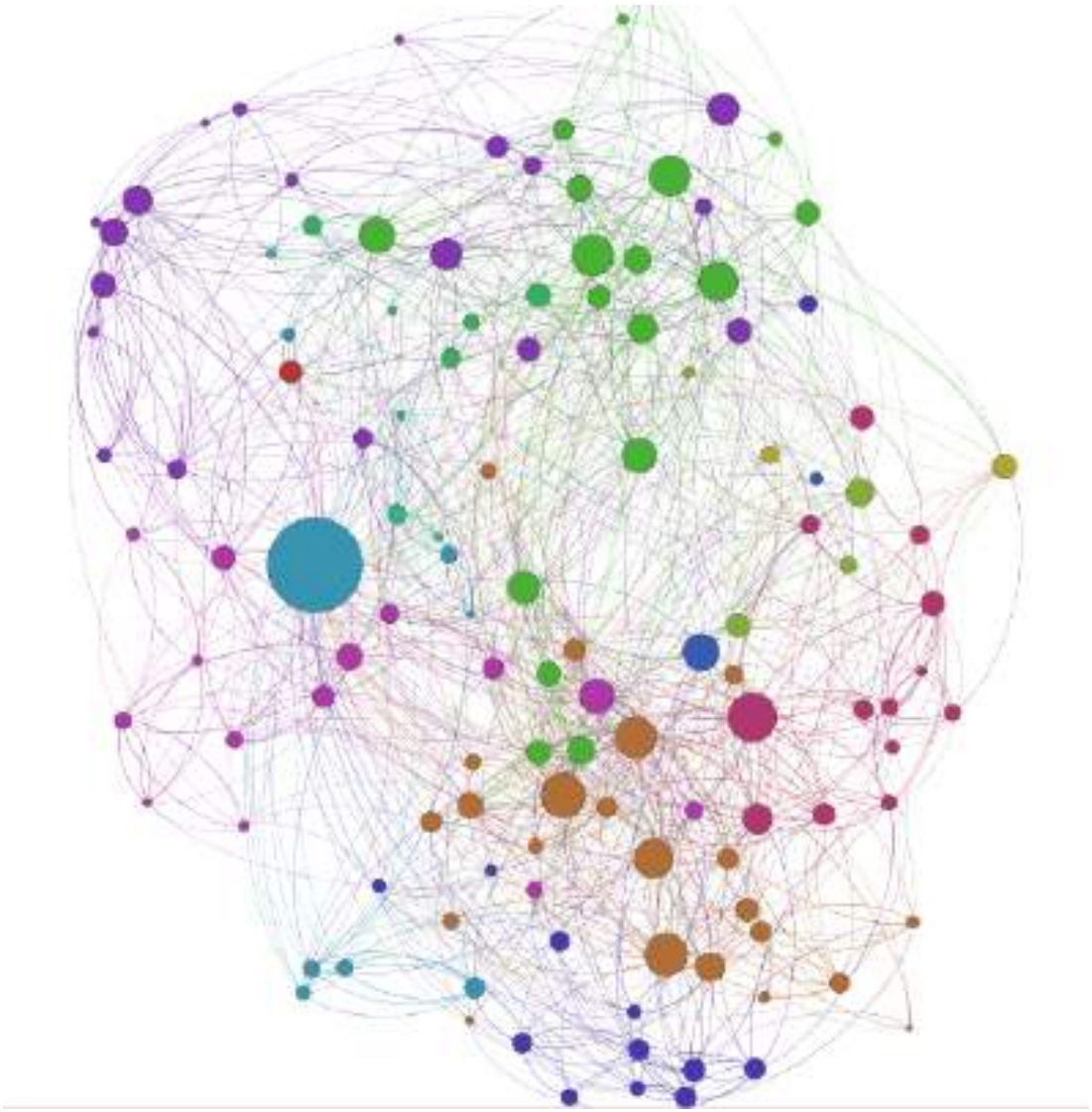


image from www.yworks.com

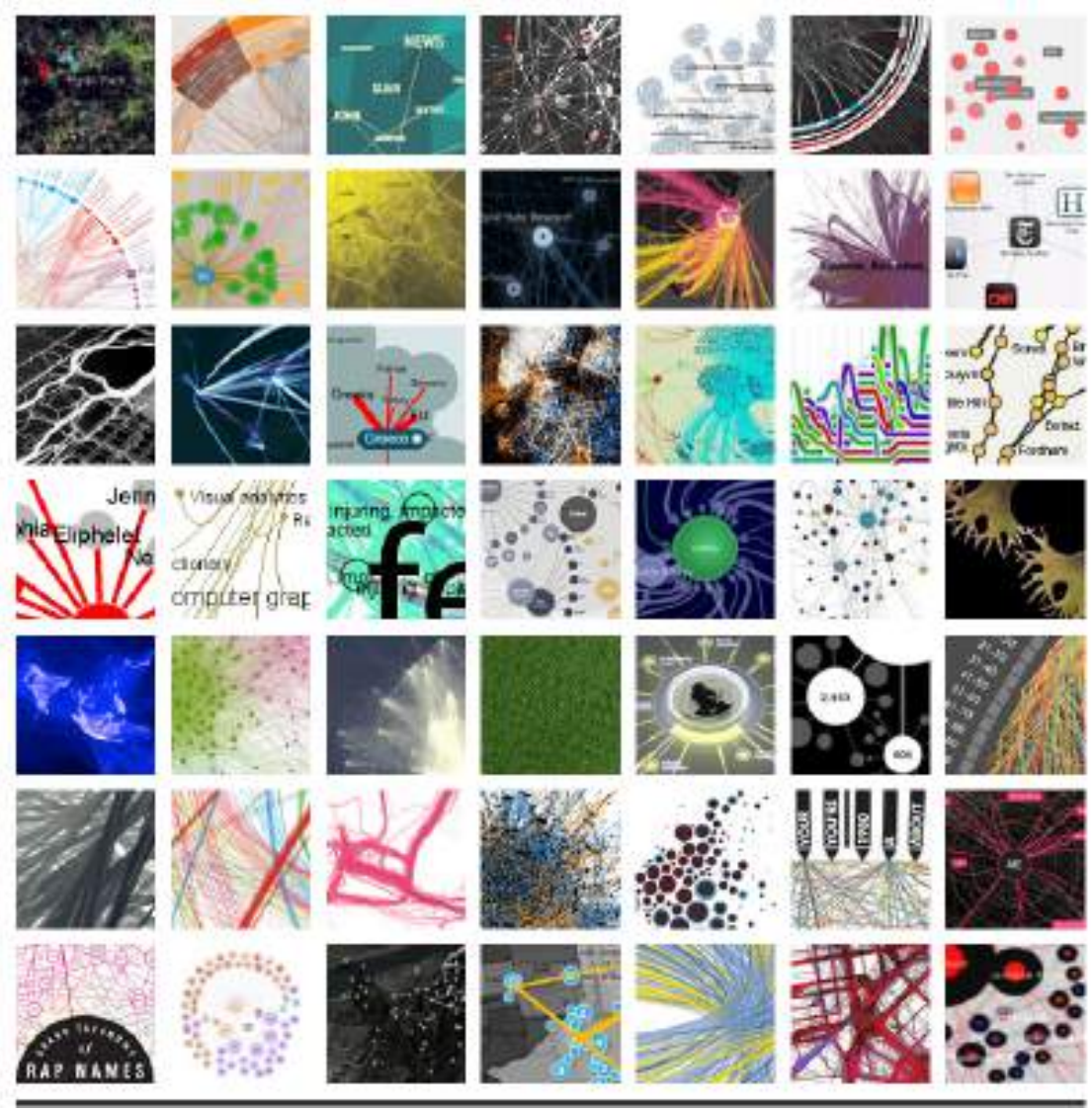
Gephi


Leucippus
Socrates
Pyrrho
Democritus
Michel de Montaigne
Karl Marx
Thales
Pythagoras
Zeno of Elea
Parmenides
Sextus Empiricus
Plato
Nicolo Machiavelli
Nicholas of Cusa
René Descartes
Hugo Grotius
Jean-Jacques Rousseau
Georg Wilhelm Friedrich Hegel
Heraclitus
Aristotle
Nicolas Malebranche
Baruch Spinoza
Pliny
Montesquieu
Edmund Husserl
Noam Chomsky
David Hume
Immanuel Kant
Franz Brentano
Friedrich Adolf Trendelenburg

Gephi



Visual complexity



References

- V. Batagelj, M. Zaversnik. An O(m) Algorithms for Cores Decomposition of Networks. 2003
- L. da F. Costa, F. A. Rodrigues, et. al. Characterization of complex networks: A survey of measurements.Advances in Physics, Vol. 56, pp. 167-242, 2007
- R. Milo, S. Shen-Orr, S. Itzkovitz et al. Network motifs: simple building blocks of complex networks. Science 298 (5594): 824?827, 2002
- M. Newman. Mixing patterns in networks. Phys. Rev. E, Vol. 67, p 026126, 2003

NATIONAL RESEARCH
UNIVERSITY

Spreading phenomena in networks

Social Network Analysis. MAGoLEGO course. Lecture 7

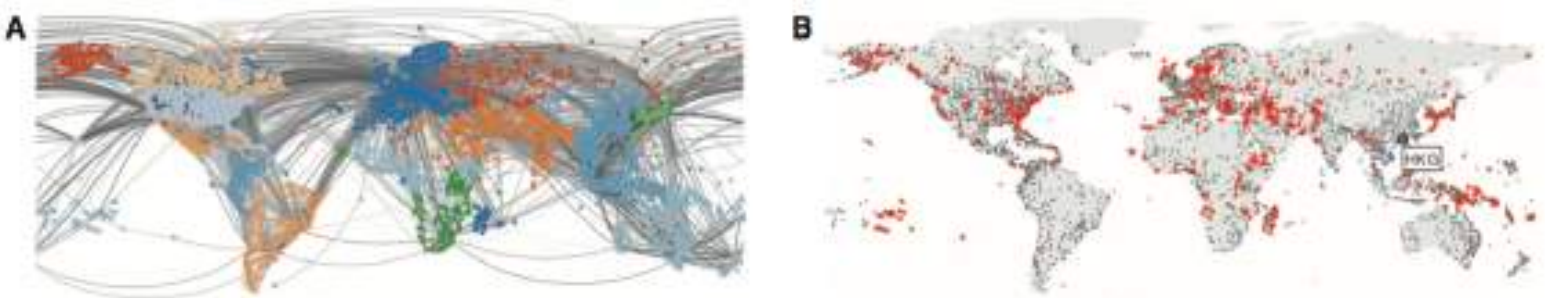
Leonid Zhukov

lzhukov@hse.ru
www.leonidzhukov.net/hse/2018/sna

National Research University Higher School of Economics
School of Data Analysis and Artificial Intelligence, Department of Computer Science

Global contagion

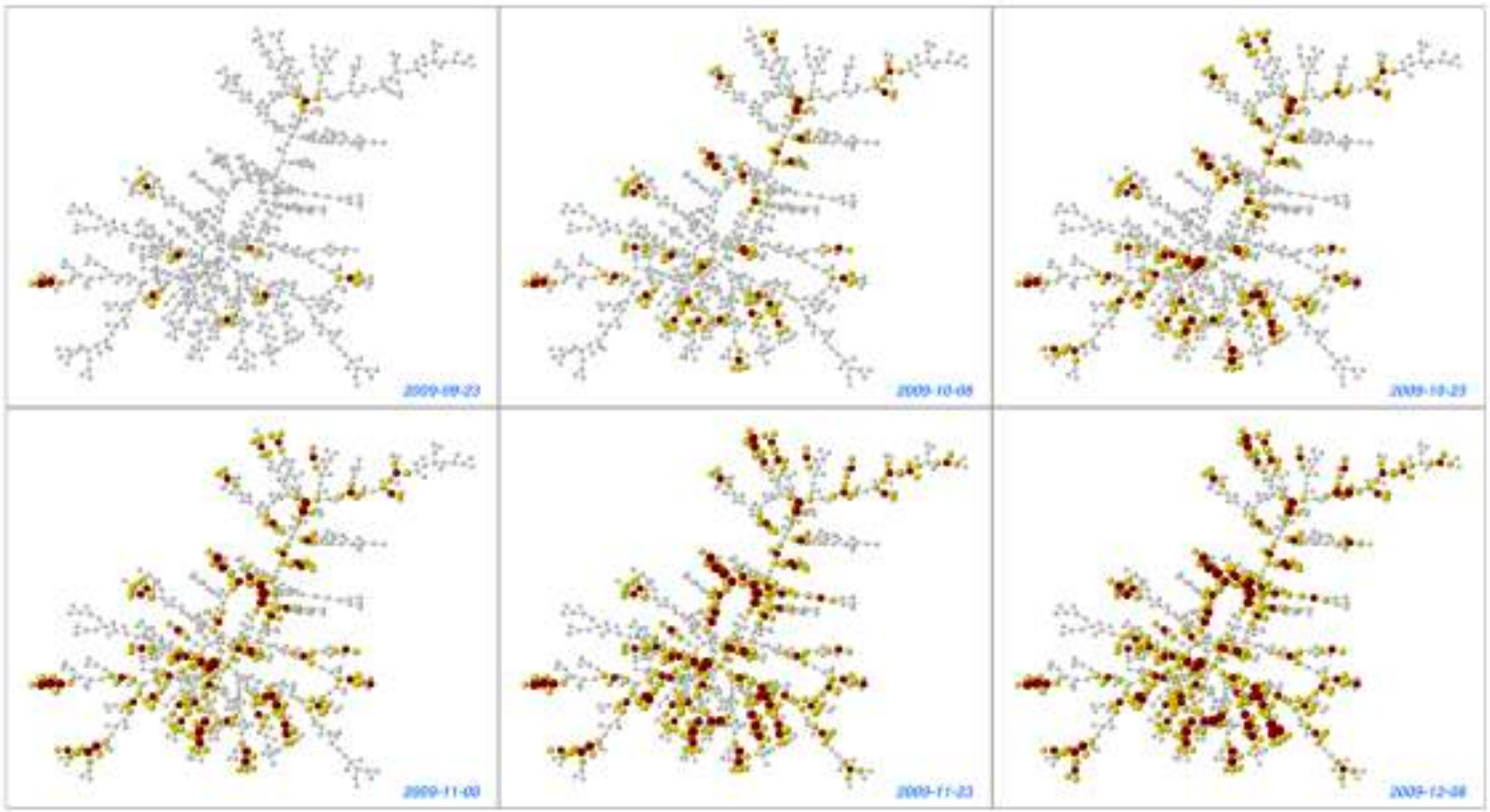
Outbreak of SARS in 2003, > 8000 cases, 10% fatality rate, 37 countries



Simulated model:
gray lines - passenger flow, red symbols epidemics location

D. Brockmann, D. Helbing, 2013

Flu contagion



Infected - red, friends of infected - yellow

N. Christakis, J. Fowler, 2010

Network epidemic model

- Given a network **G** of potential contacts
- Three states model: susceptible, infected, recovered states
- Probabilistic model (state of a node):
  $s_i(t)$ - probability that at *t* node *i* is susceptible
  $x_i(t)$ - probability that at *t* node *i* is infected
  $r_i(t)$ - probability that at *t* node *i* is recovered
- Model parameters:
  β - infection rate (probably to get infected on a contact in time δt)
  γ - recovery rate (probability to recover in a unit time δt)
- connected component - all nodes reachable
- network is undirected (matrix **A** is symmetric)
- if graph complete - fully mixing model
- Based upon models from mathematical epidemiology, W.O. Kermack and McKendrick, 1927

Probabilistic model

Two processes:

- Node infection:

$$P_{inf} \approx \beta s_i(t) \sum_{j \in \mathcal{N}(i)} x_j(t) \delta t$$

- Node recovery:

$$P_{rec} = \gamma x_i(t) \delta t$$

SI model

- SI Model

$$S \longrightarrow I$$

- Probabilities that node i: $s_i(t)$ - susceptible, $x_i(t)$ -infected at t

$$x_i(t) + s_i(t) = 1$$

- β - infection rate, probability to get infected in a unit time

$$x_i(t + \delta t) = x_i(t) + \beta s_i(t) \sum_j A_{ij} x_j(t) \delta t$$

- infection equations

$$\begin{aligned} \frac{dx_i(t)}{dt} &= \beta s_i(t) \sum_j A_{ij} x_j(t) \\ x_i(t) &+ s_i(t) = 1 \end{aligned}$$

SI model

SI Model

$$S \longrightarrow I$$

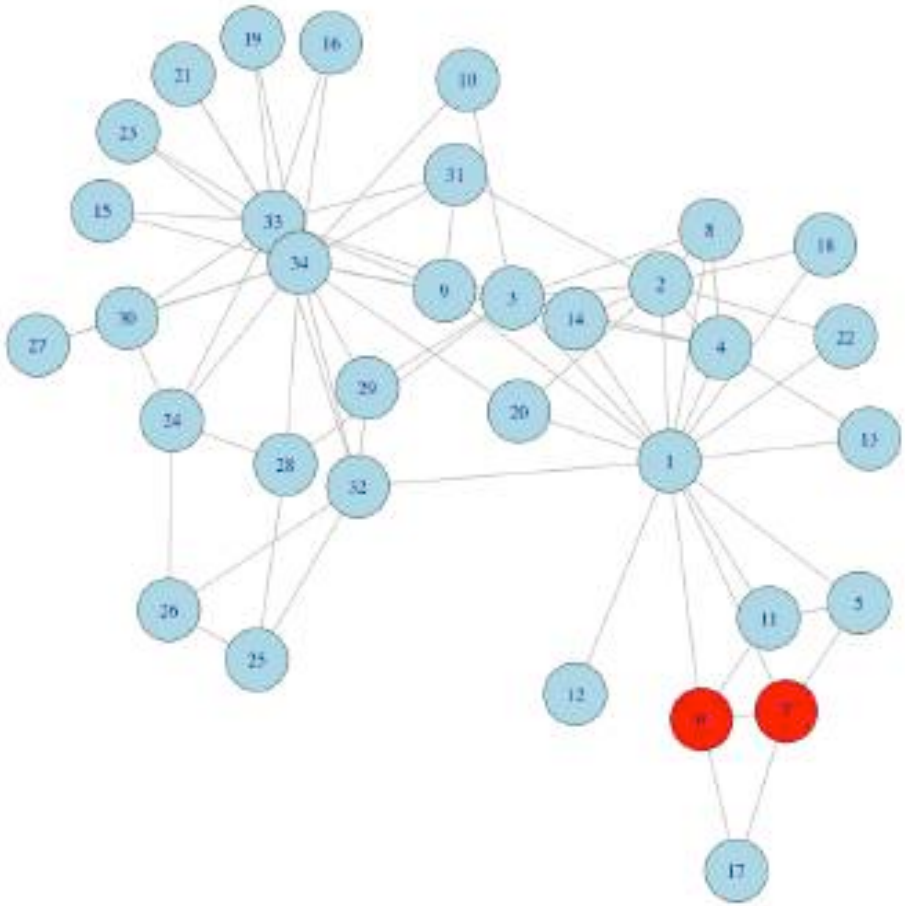
1. Every node at any time step is in one state $\{S, I\}$
2. Initialize c nodes in state I
3. On each time step each I node has a probability β to infect its nearest neighbors (NN), $S \rightarrow I$

Model dynamics:

$$I + S \quad \stackrel{\beta}{\longrightarrow} \quad 2I$$

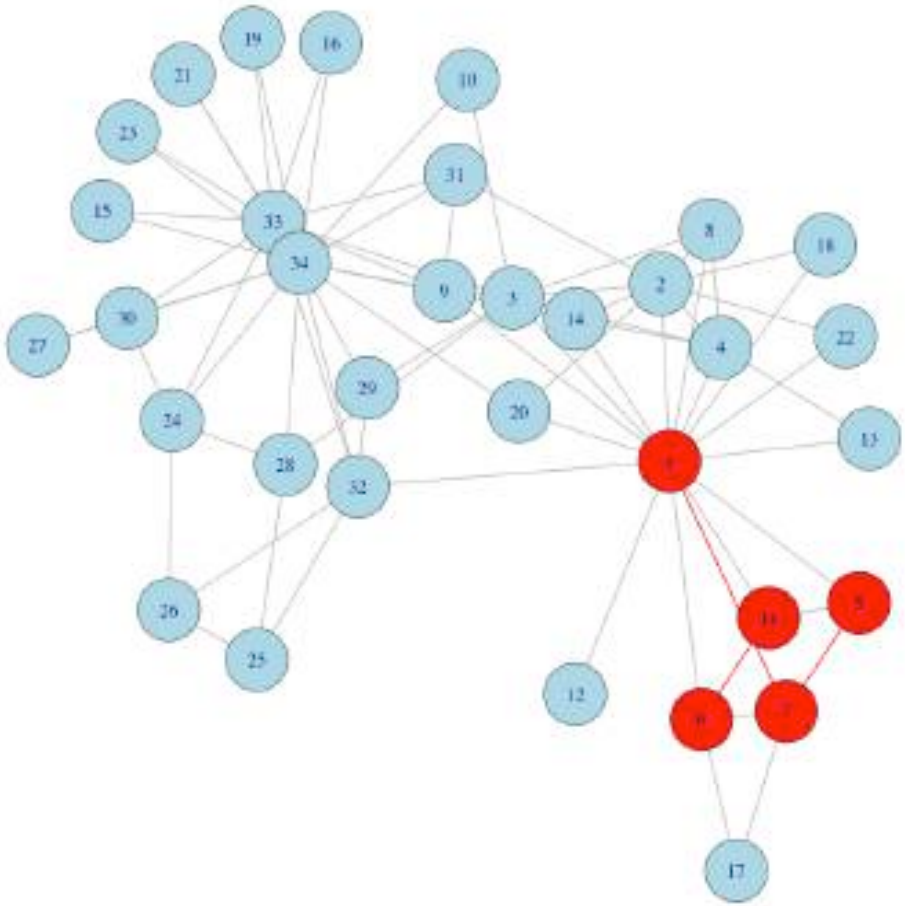
SI model

$\beta = 0.5$



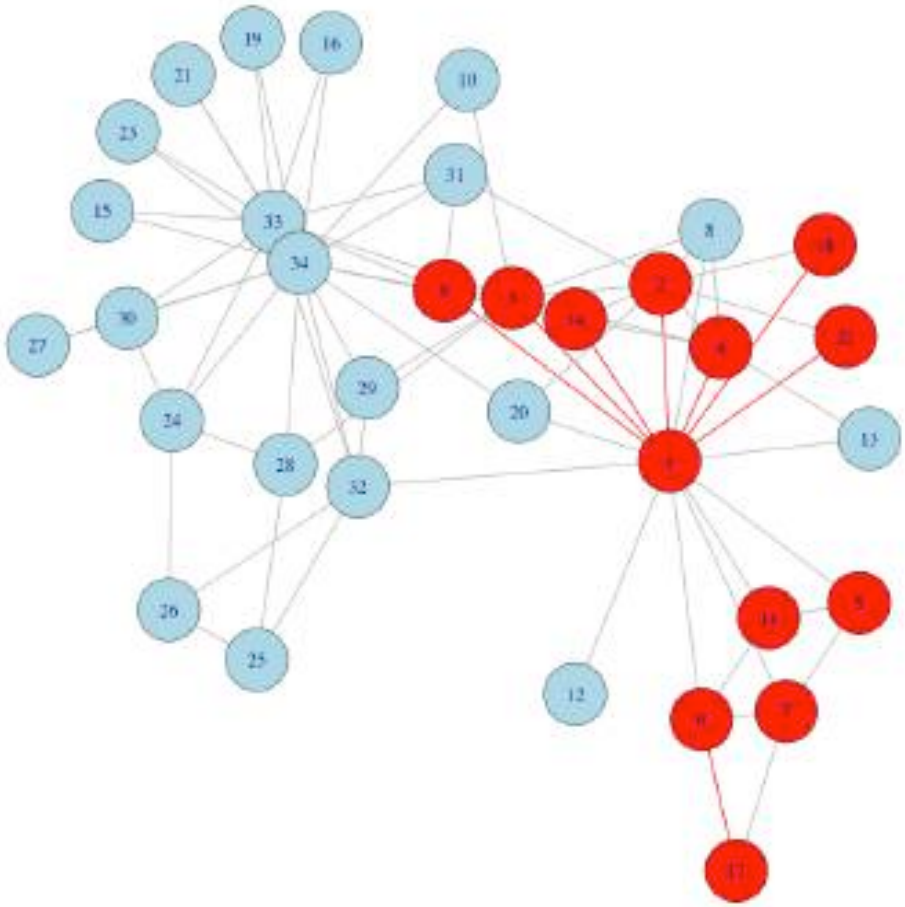
SI model

$\beta = 0.5$



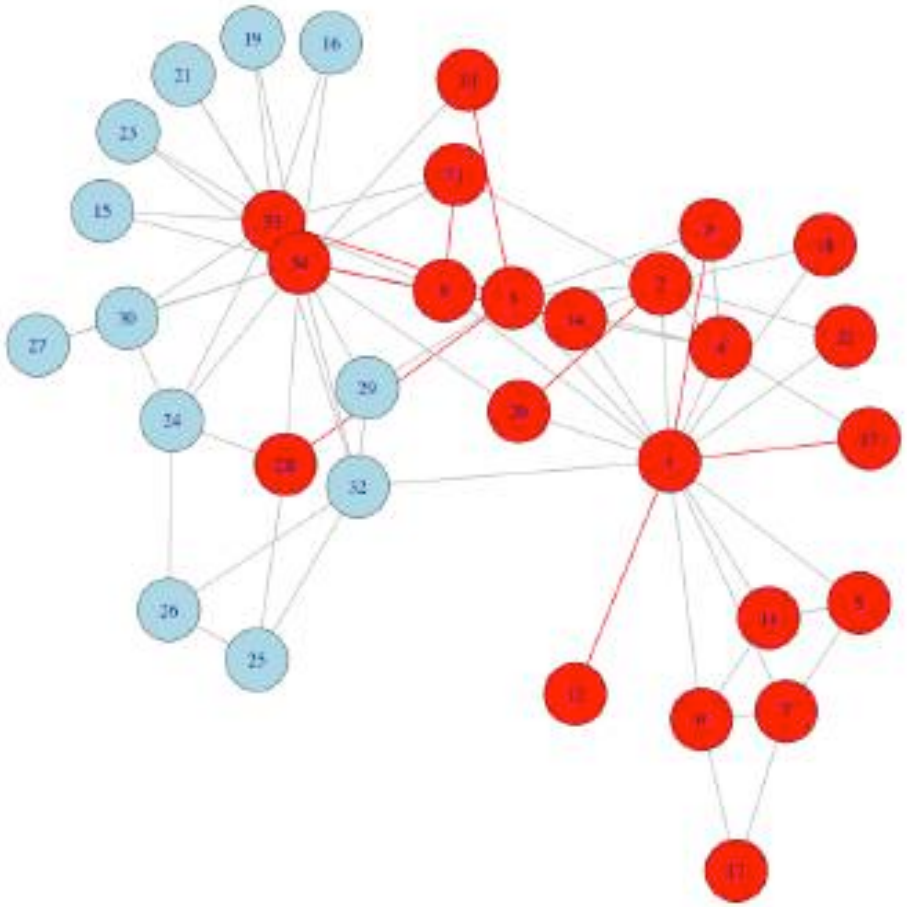
SI model

$\beta = 0.5$



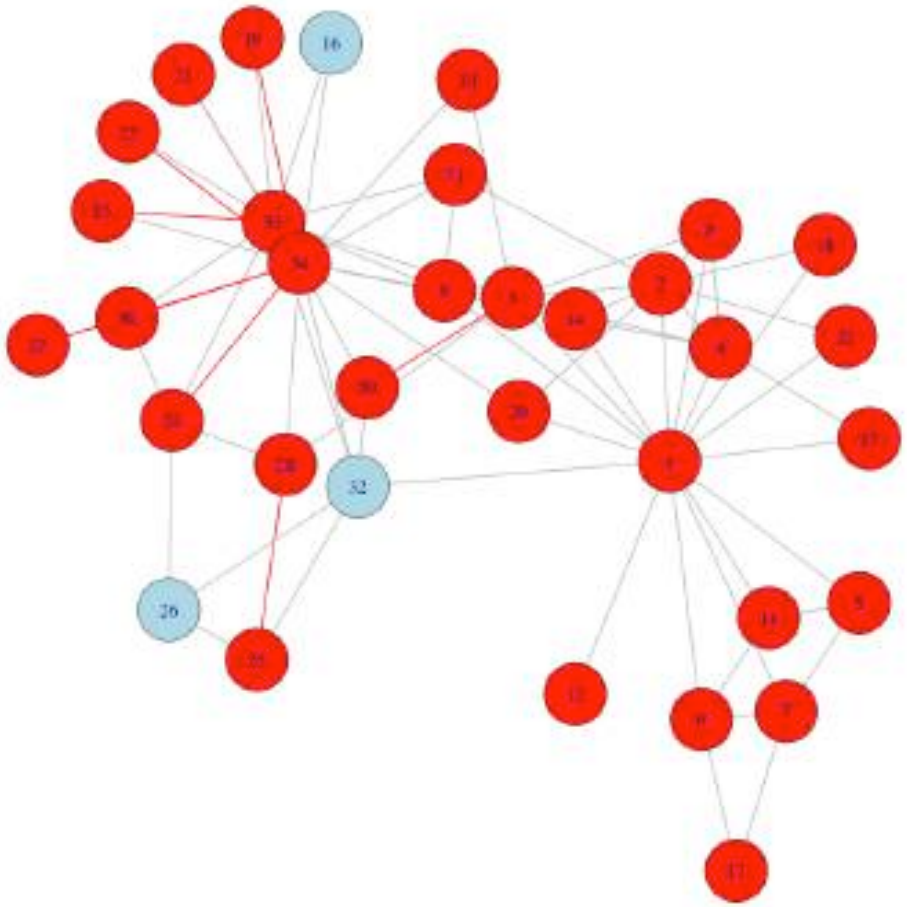
SI model

$\beta = 0.5$



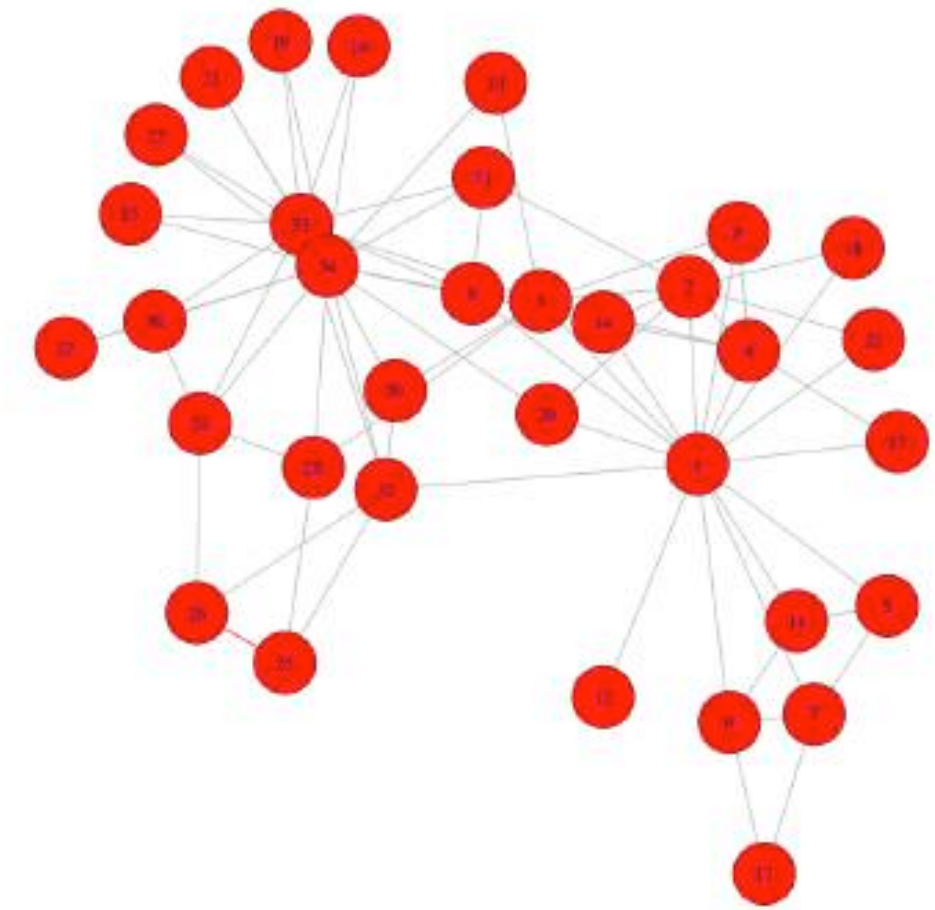
SI model

$\beta = 0.5$

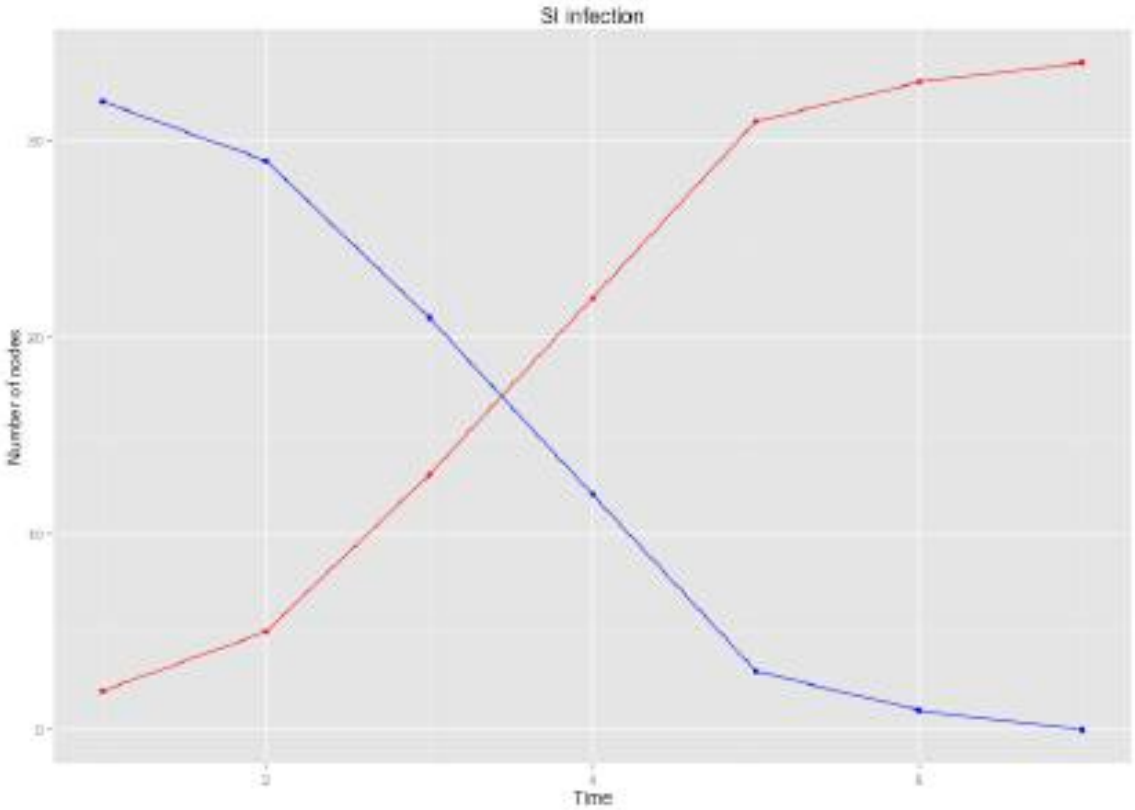


SI model

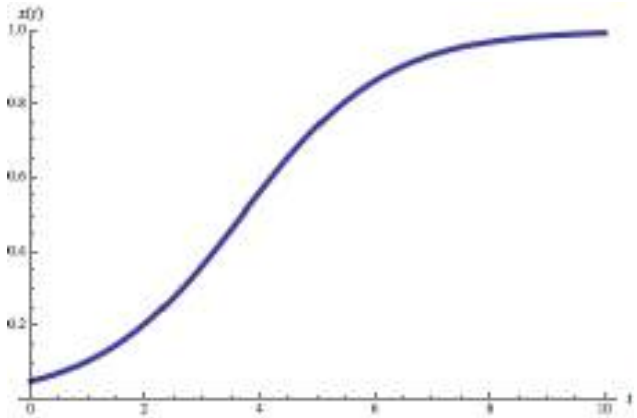
$\beta = 0.5$



SI model



SI model



1. growth rate of infections depends on λ_1
2. All nodes in connected component get infected $t \to \infty$
   $x_i(t) \to 1$

image from M. Newman, 2010

SIS model simulations

SIS Model

$$S \longrightarrow I \longrightarrow S$$

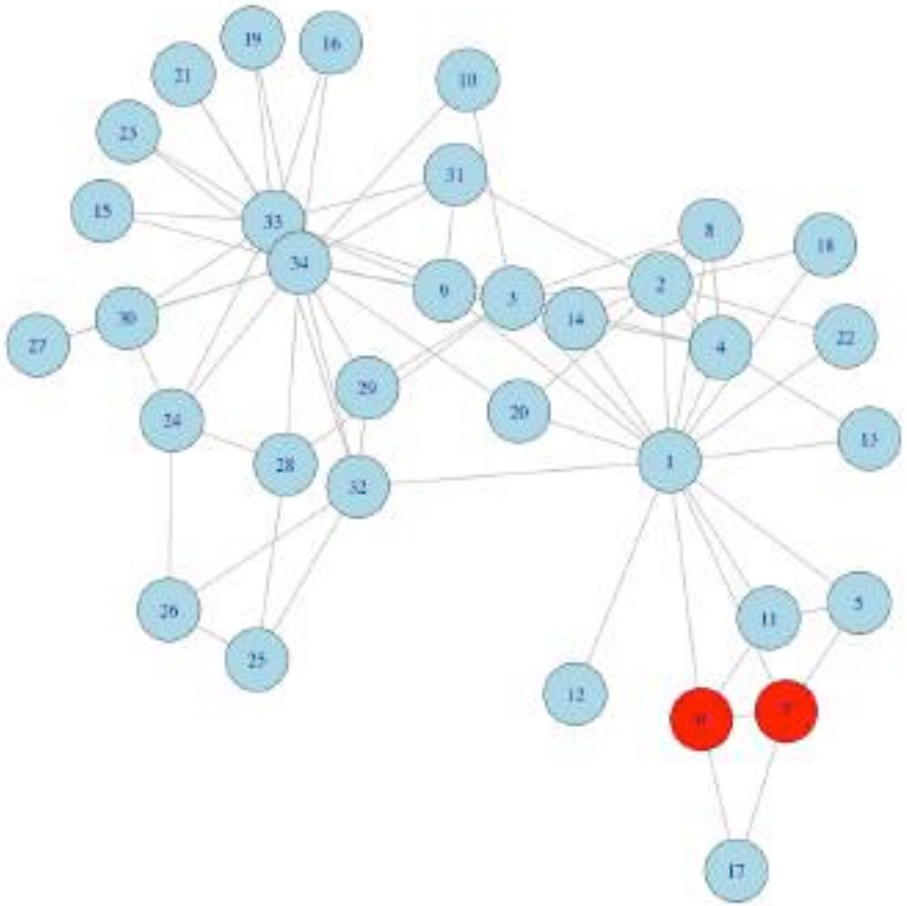
1. Every node at any time step is in one state $\{S, I\}$
2. Initialize c nodes in state I
3. Each node stays infected $\tau_\gamma = 1/\gamma$ time steps
4. On each time step each I node has a probability β to infect its nearest neighbors (NN), $S \to I$
5. After τ_γ time steps node recovers, $I \to S$

Model dynamics:

$$\begin{cases} I + S & \xrightarrow{\beta} & 2I \\ I & \xrightarrow{\gamma} & S \end{cases}$$

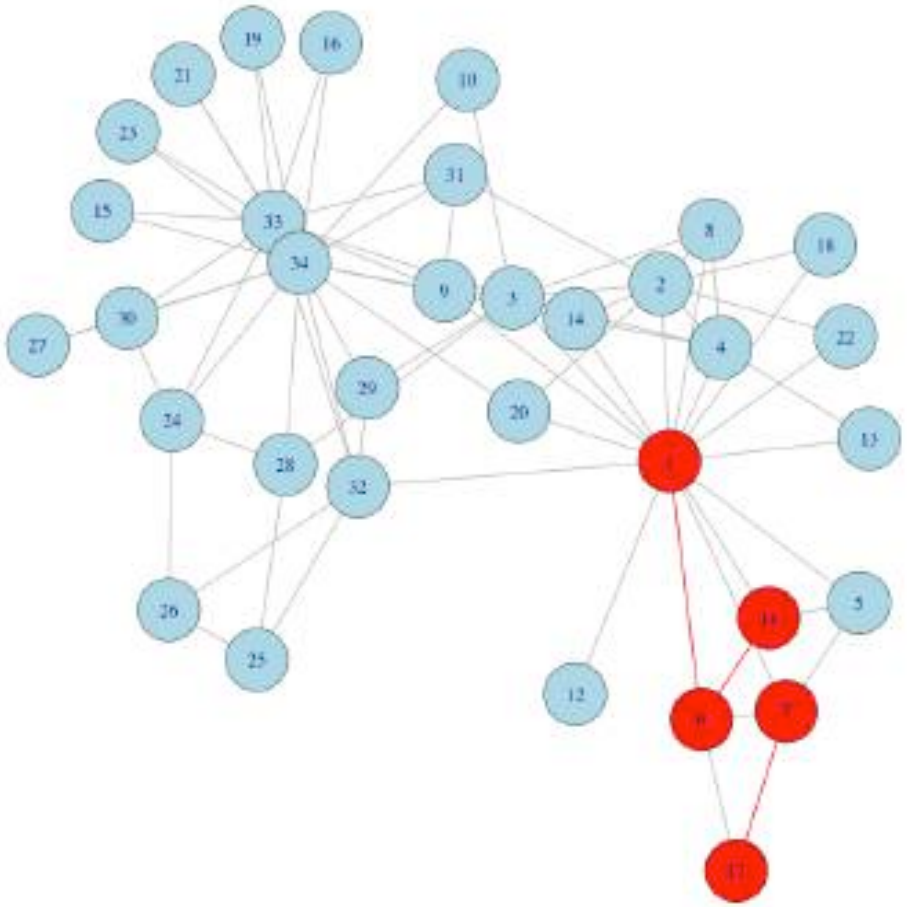
SIS model

$\beta = 0.5, \tau = 2$



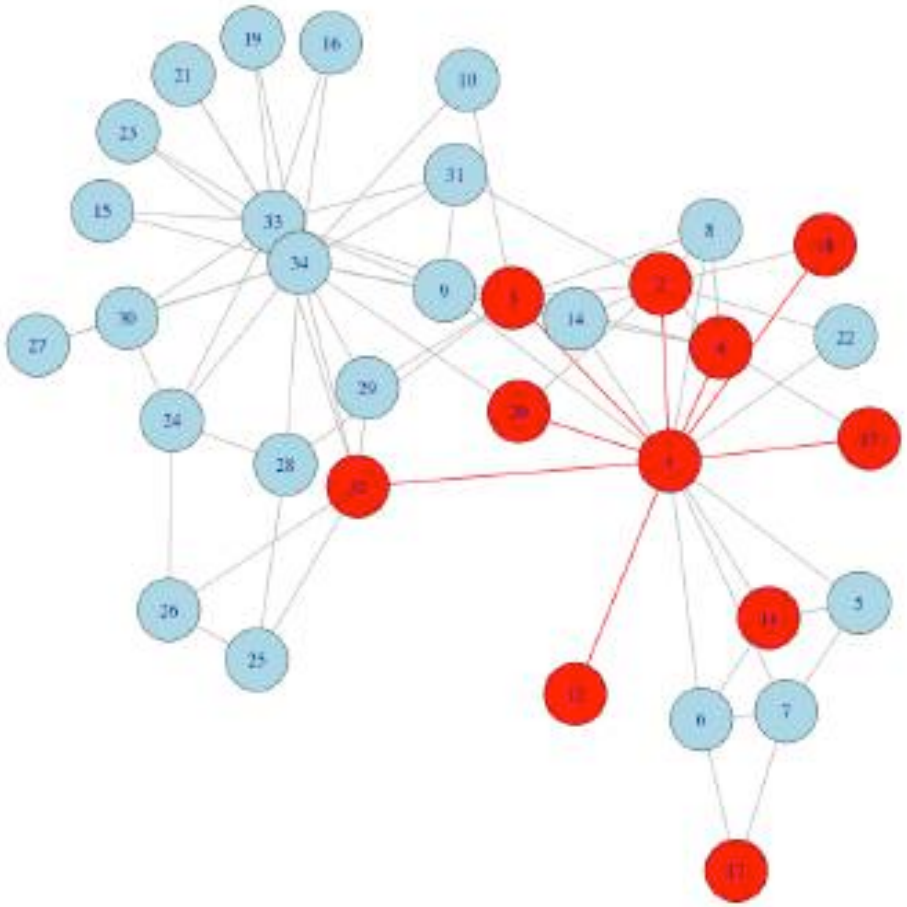
SIS model

$\beta = 0.5, \tau = 2$



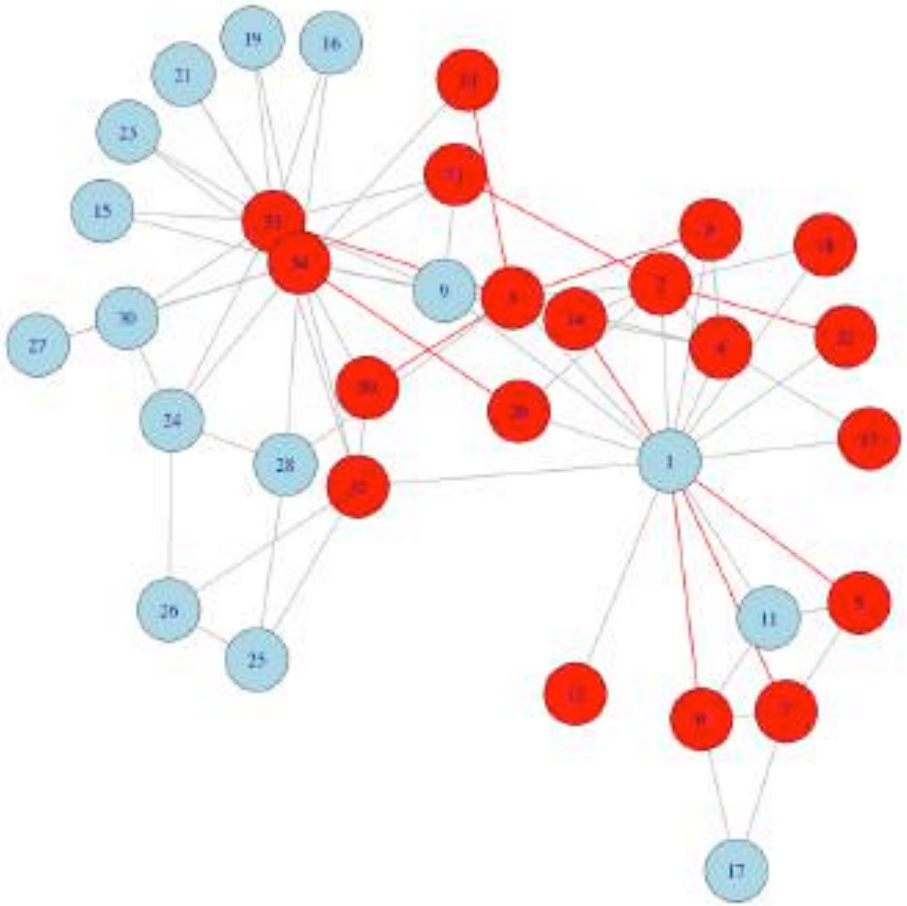
SIS model

$\beta = 0.5, \tau = 2$



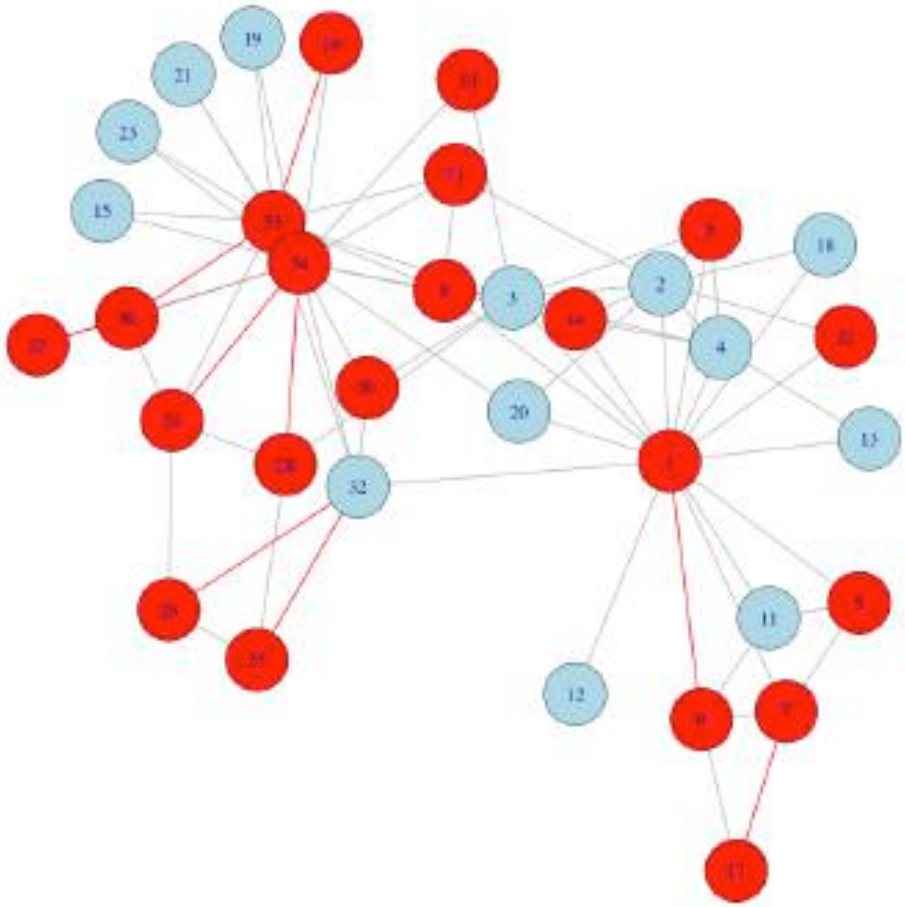
SIS model

$\beta = 0.5, \tau = 2$



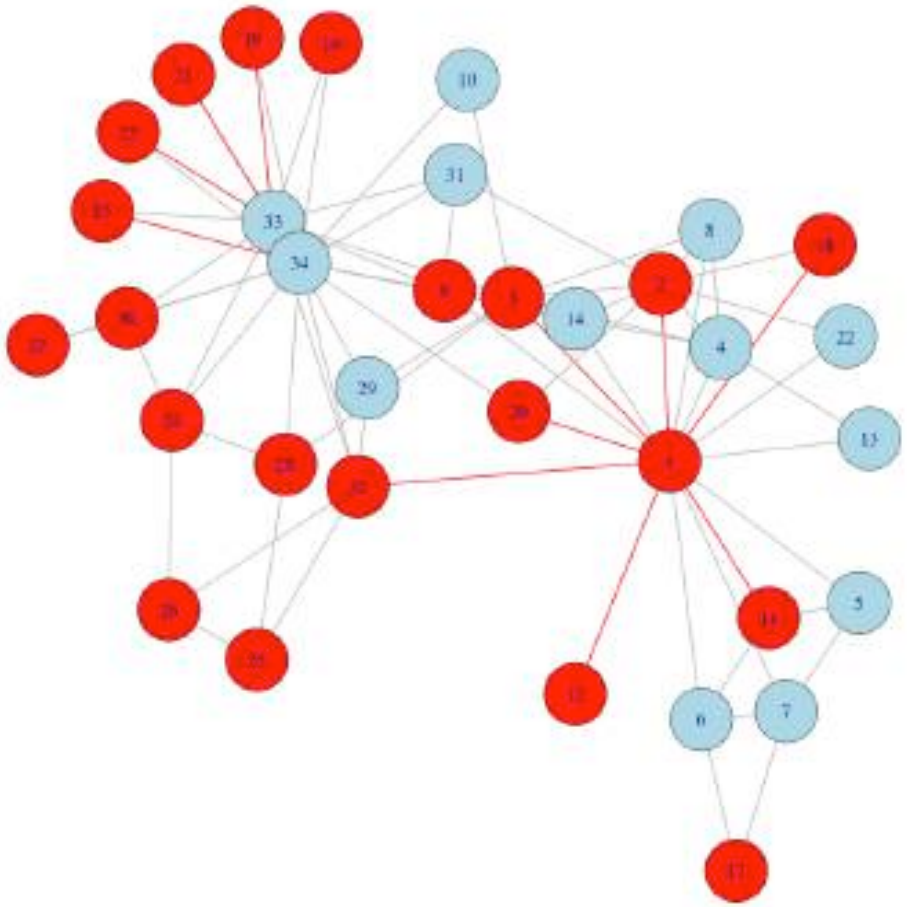
SIS model

$\beta = 0.5, \tau = 2$



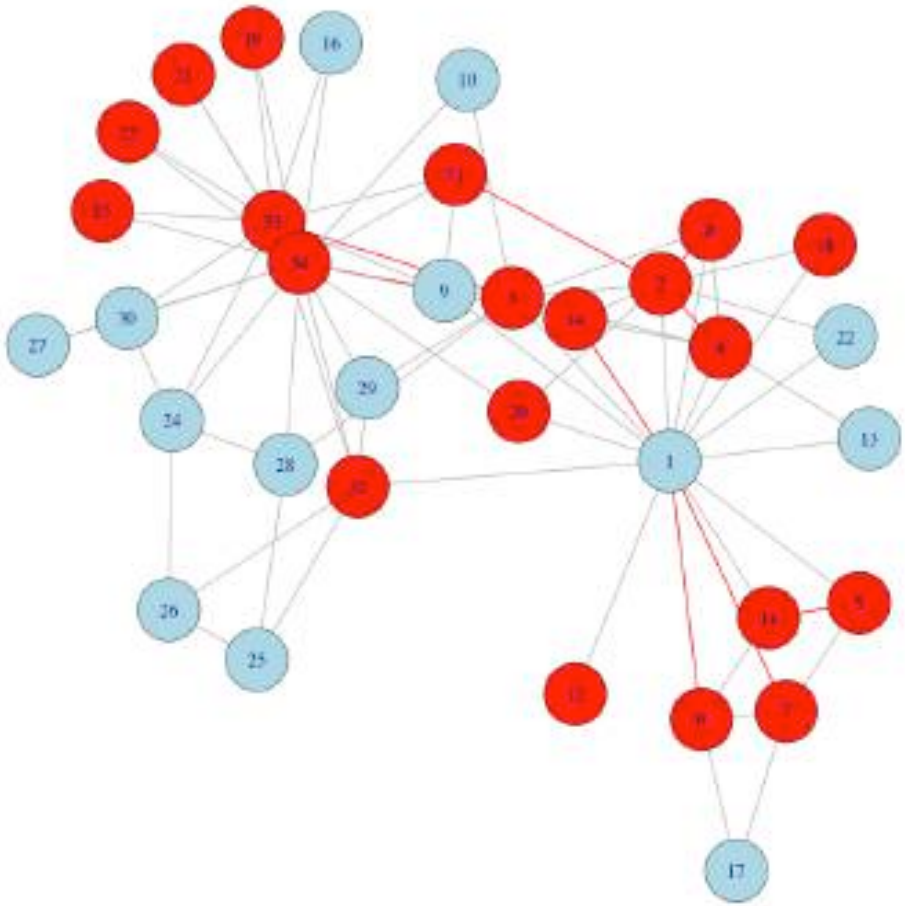
SIS model

$\beta = 0.5, \tau = 2$

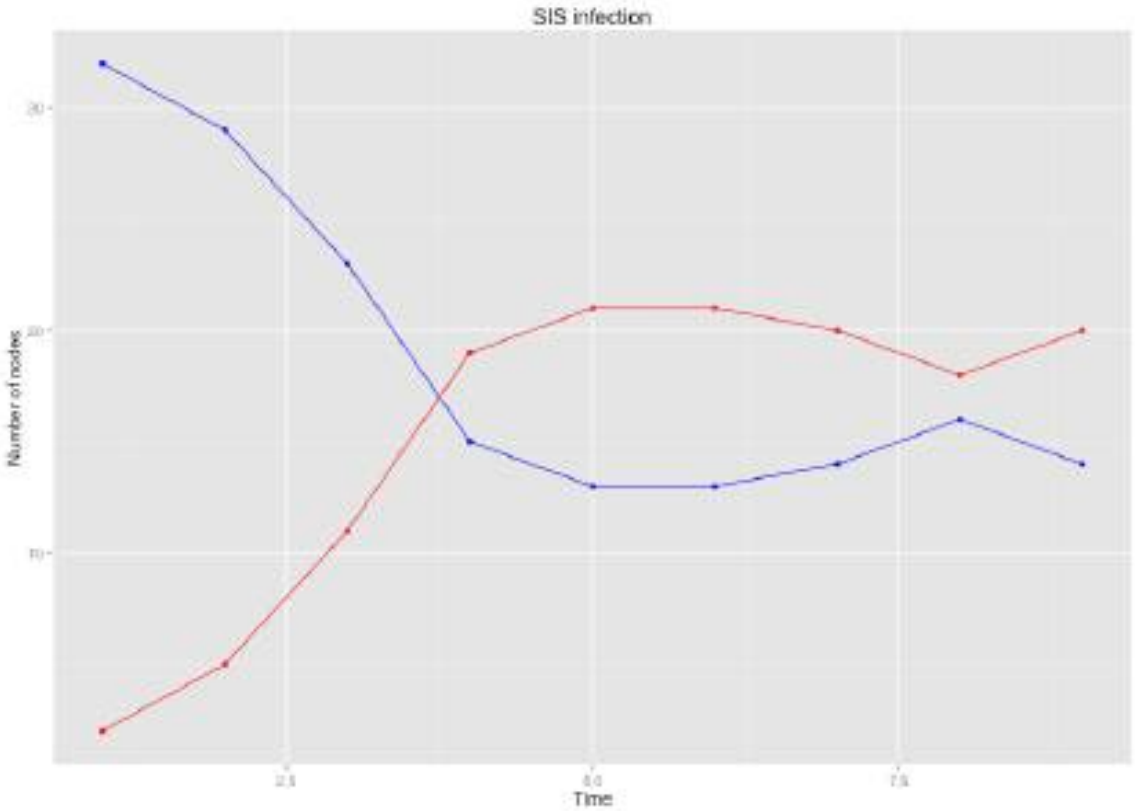


SIS model

$\beta = 0.5, \tau = 2$

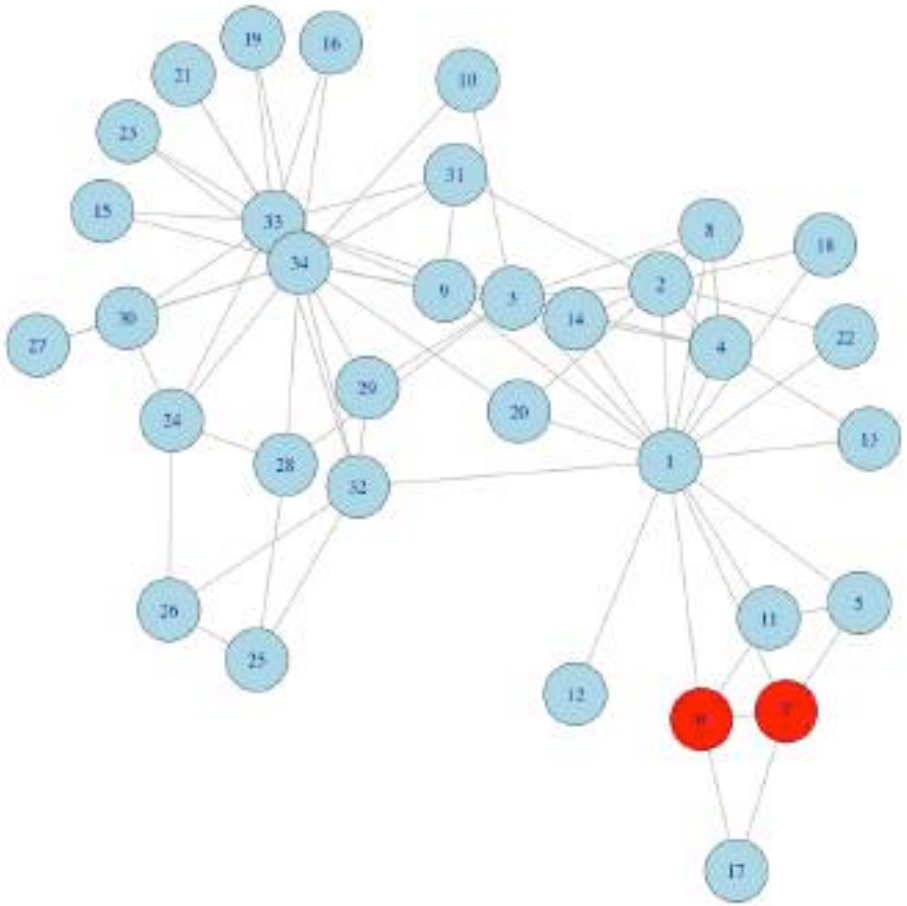


SIS model



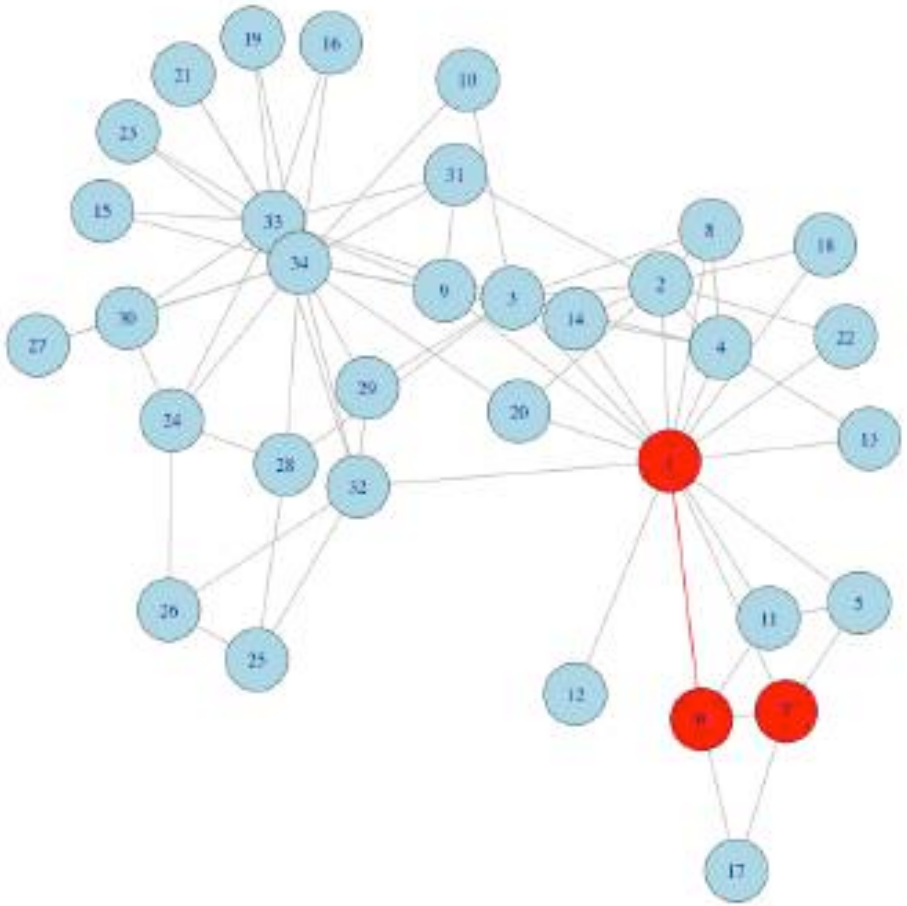
SIS model

$\beta = 0.2, \tau = 2$



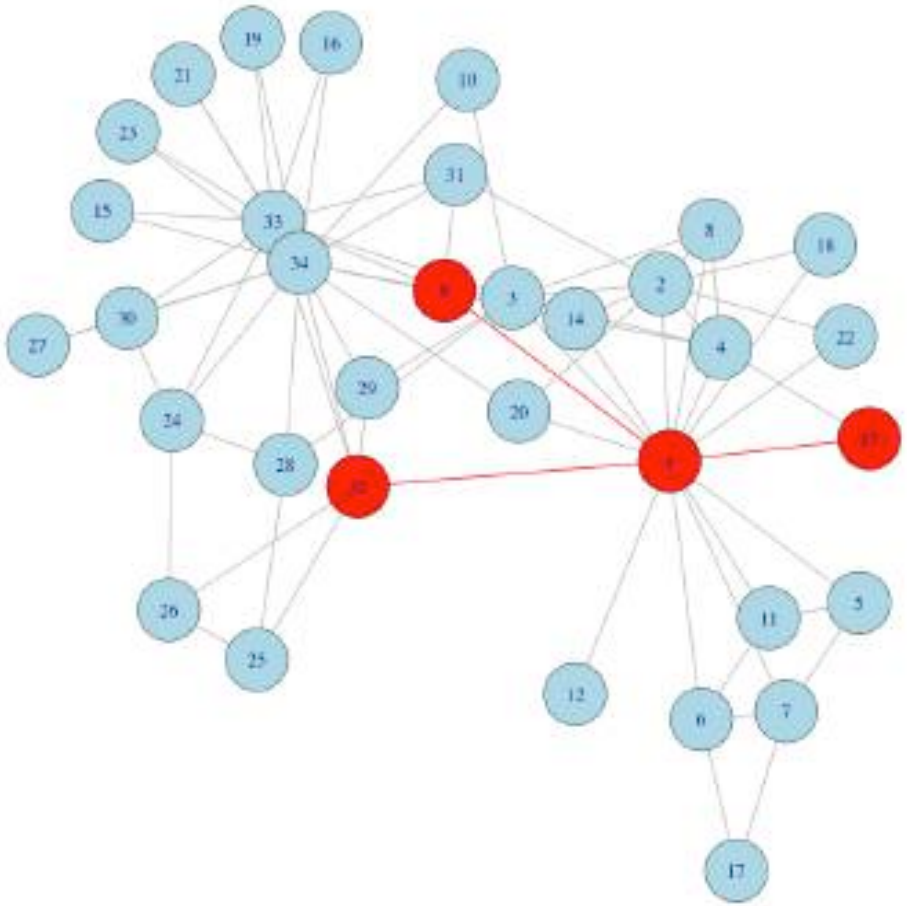
SIS model

$\beta = 0.2, \tau = 2$



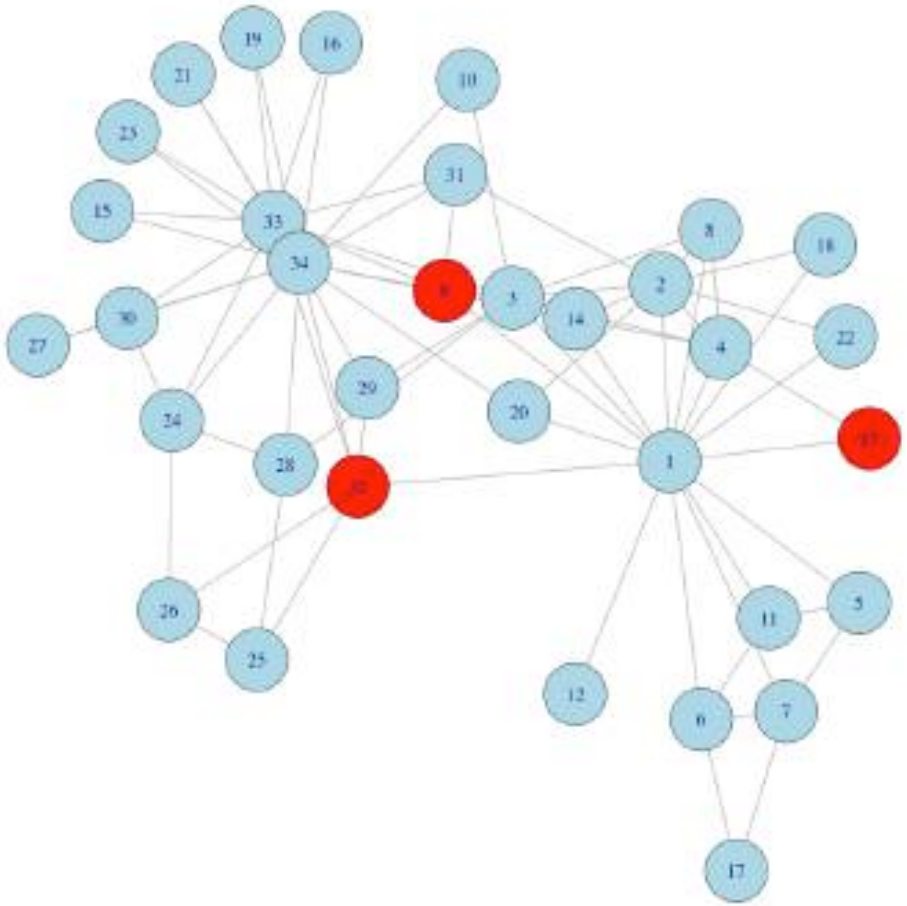
SIS model

$$\beta = 0.2, \tau = 2$$



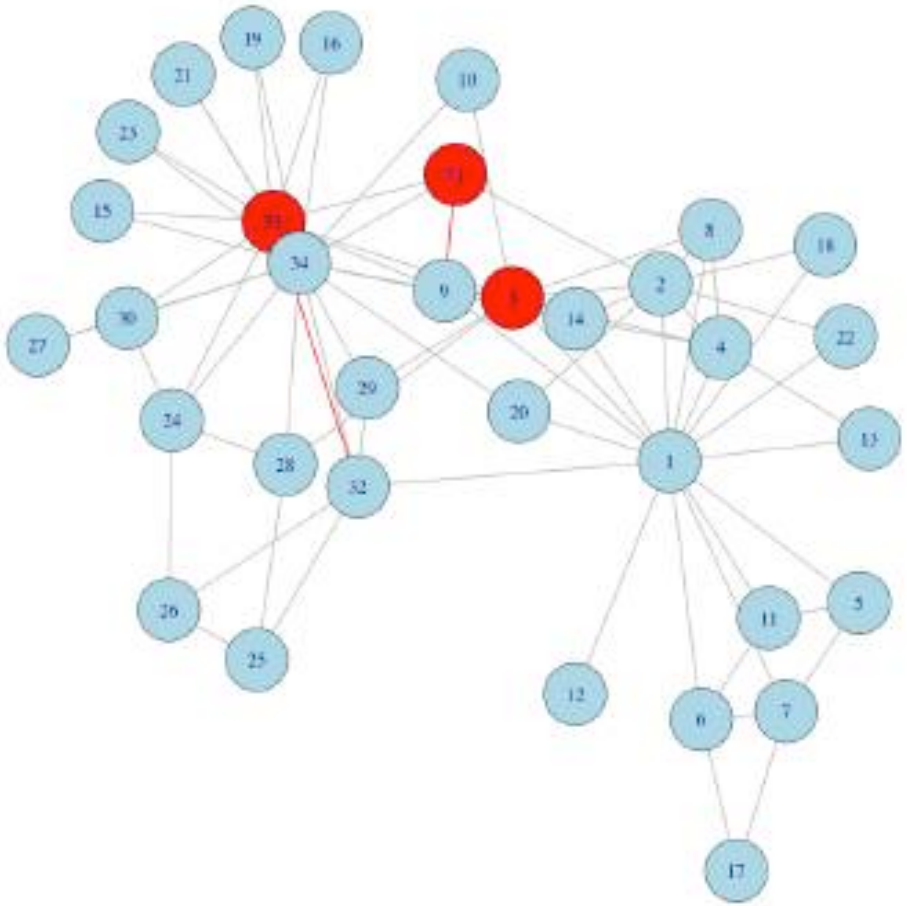
SIS model

$\beta = 0.2, \tau = 2$



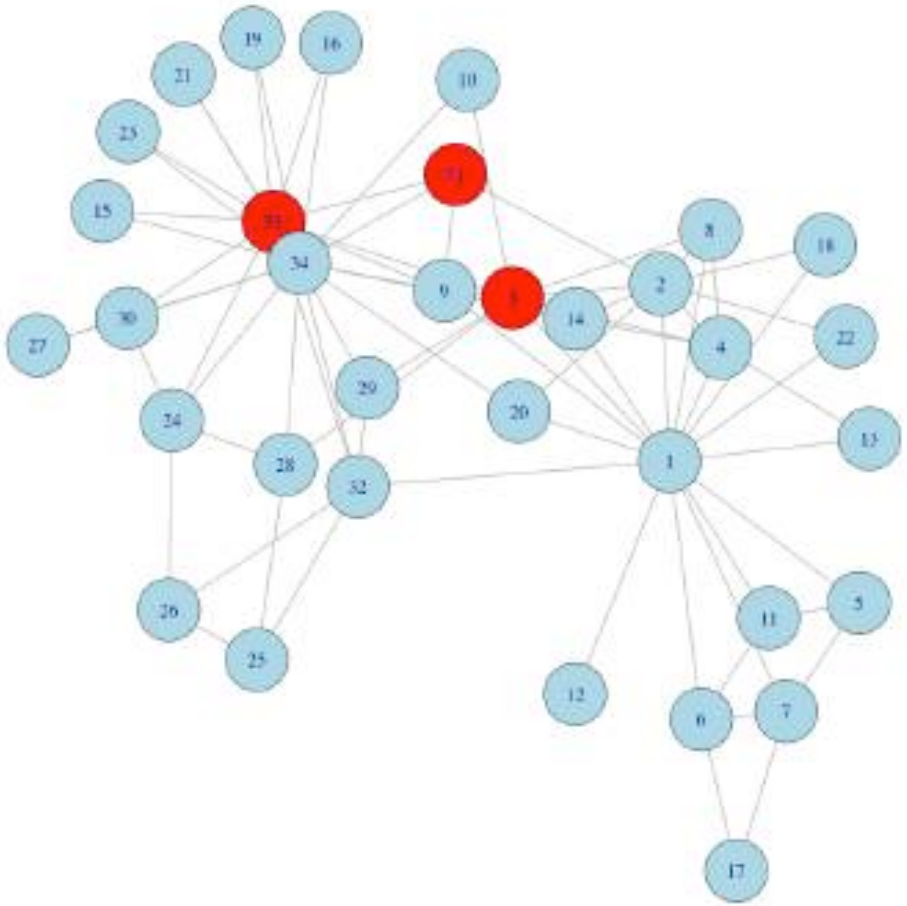
SIS model

$\beta = 0.2, \tau = 2$



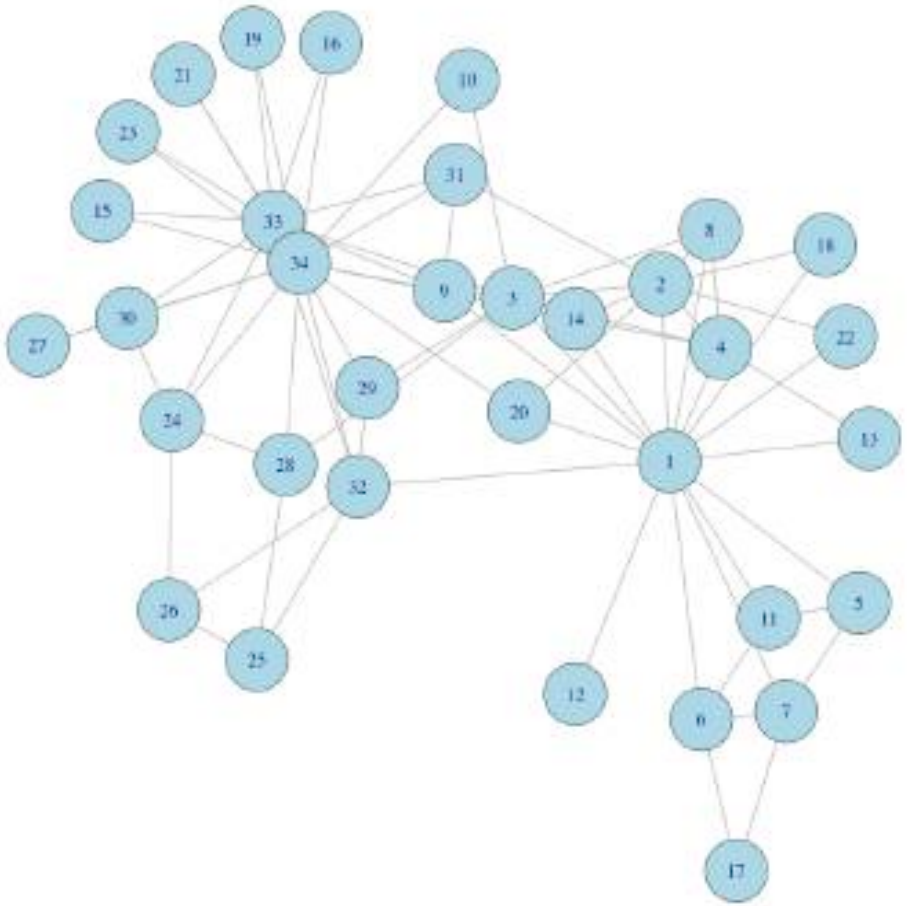
SIS model

$\beta = 0.2, \tau = 2$

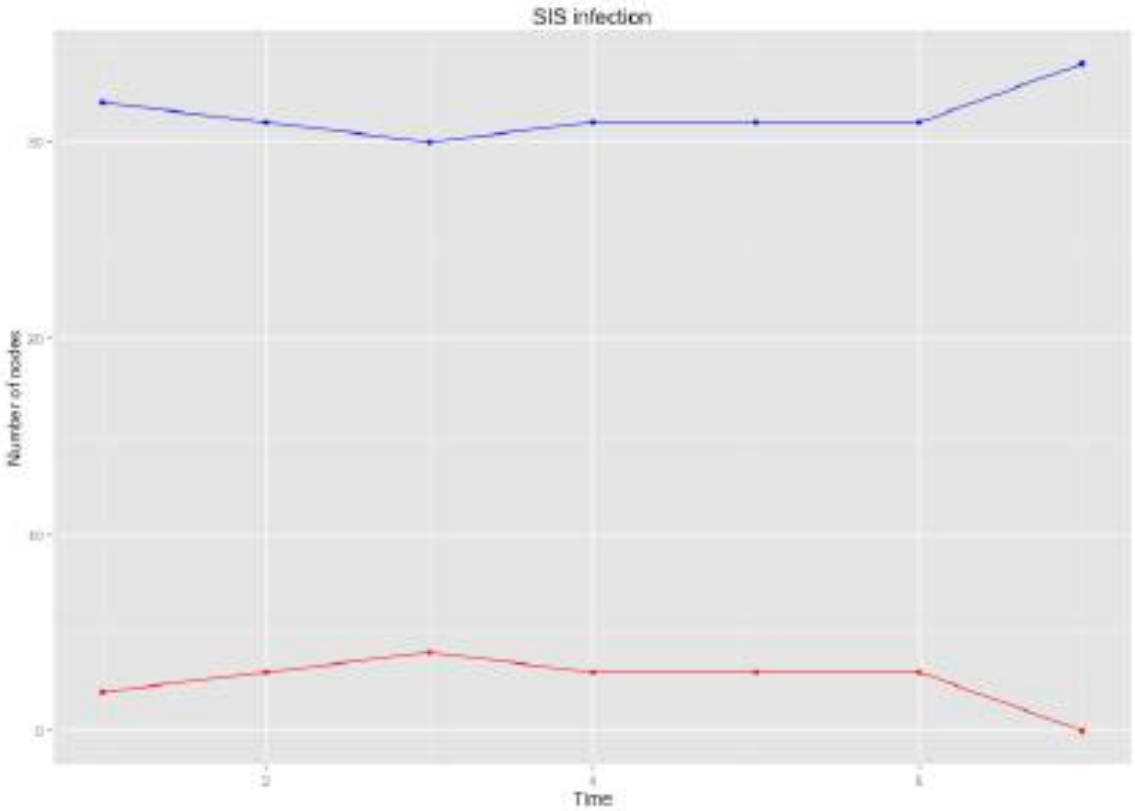


SIS model

$\beta = 0.2, \tau = 2$



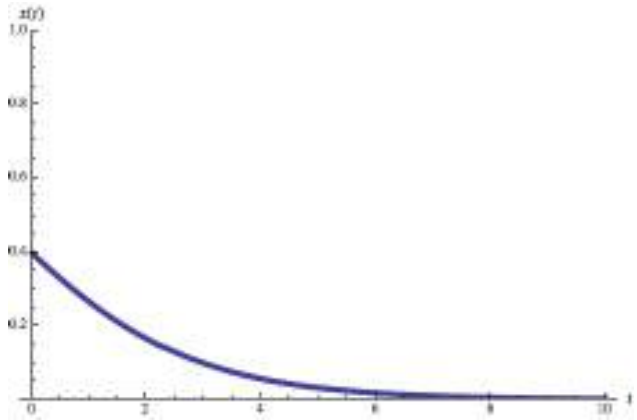
SIS model



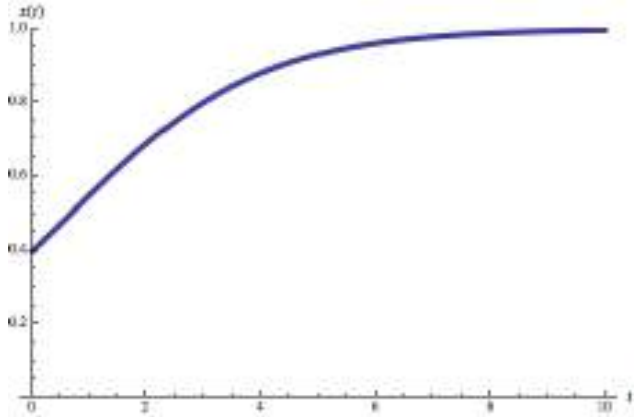
SIS model

Epidemic threshold R_0:

- if $\frac{\beta}{\gamma} < R_0$ - infection dies over time



- if $\frac{\beta}{\gamma} > R_0$ - infection survives and becomes epidemic



SIR model simulation

SIR Model

$$S \longrightarrow I \longrightarrow R$$

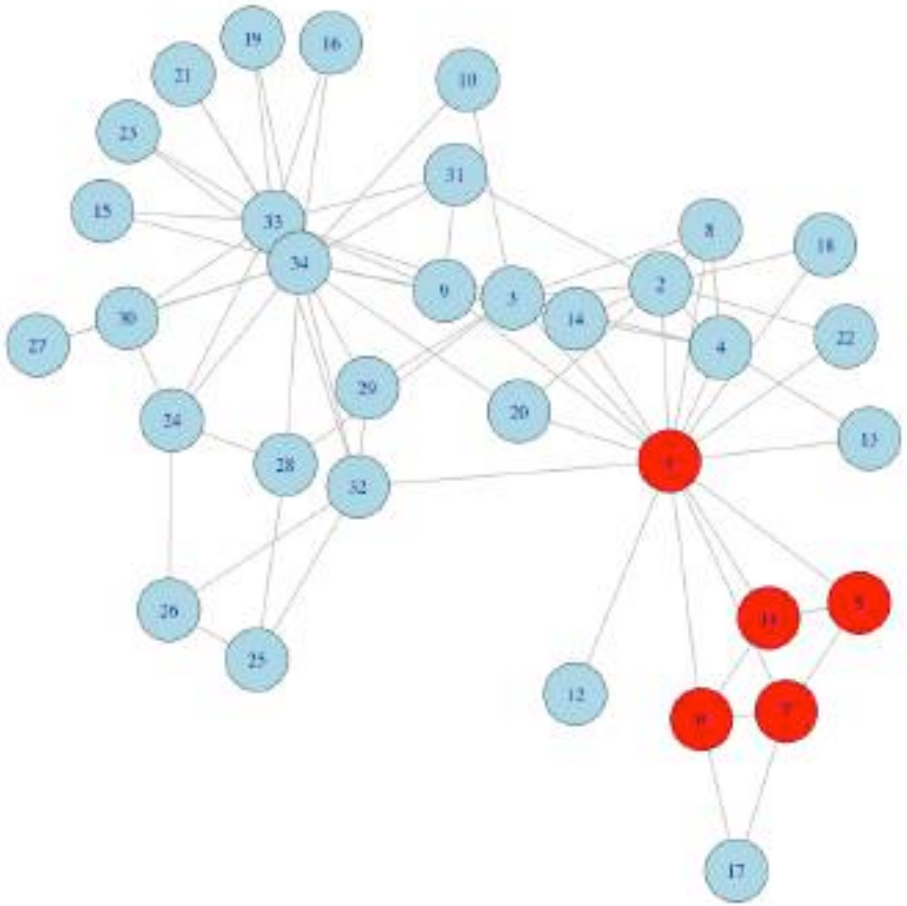
1. Every node at any time step is in one state $\{S, I, R\}$
2. Initialize c nodes in state I
3. Each node stays infected $\tau_\gamma = 1/\gamma$ time steps
4. On each time step each I node has a prabability β to infect its nearest neighbours (NN), $S \to I$
5. After τ_γ time steps node recovers, $I \to R$
6. Nodes R do not participate in further infection propagation

Model dynamics:

$$\begin{cases} I + S & \xrightarrow{\beta} & 2I \\ I & \xrightarrow{\gamma} & R \end{cases}$$

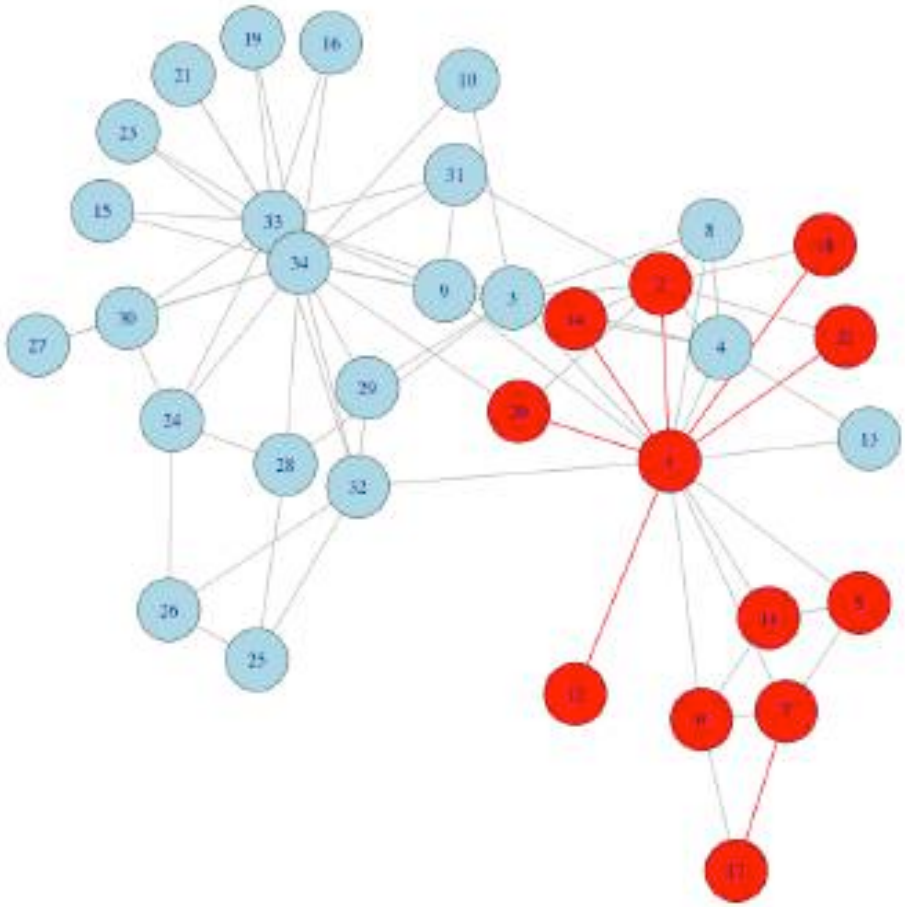
SIR model

$\beta = 0.5, \tau = 2$



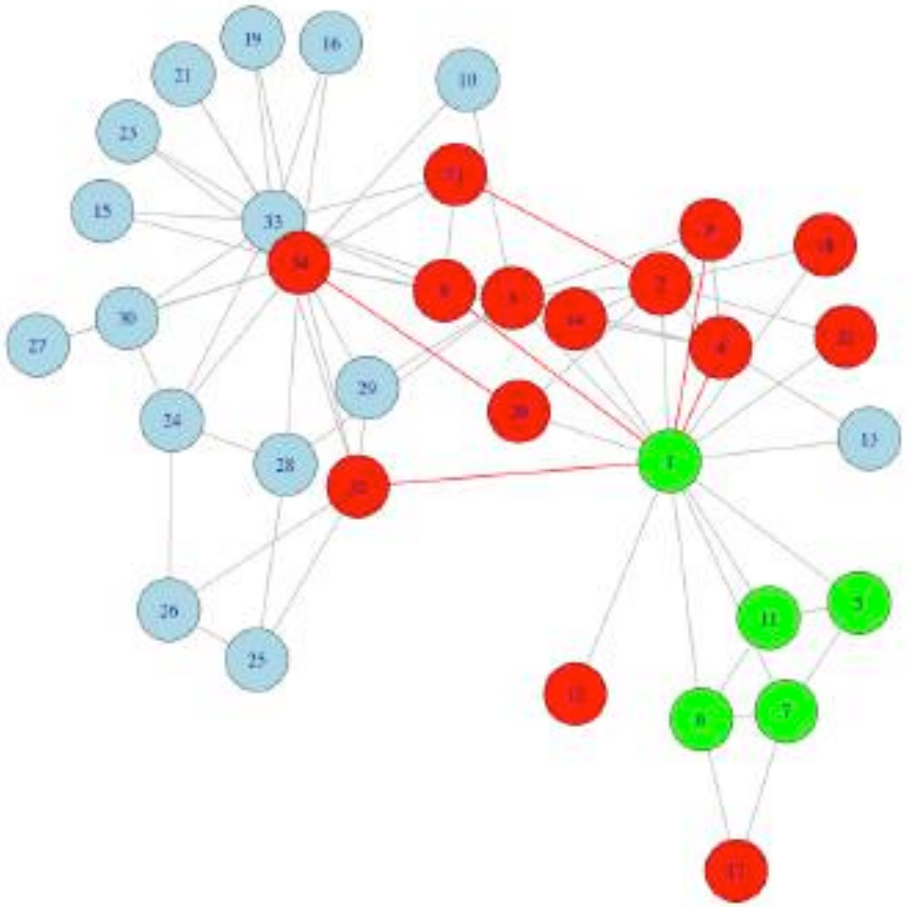
SIR model

$\beta = 0.5, \tau = 2$



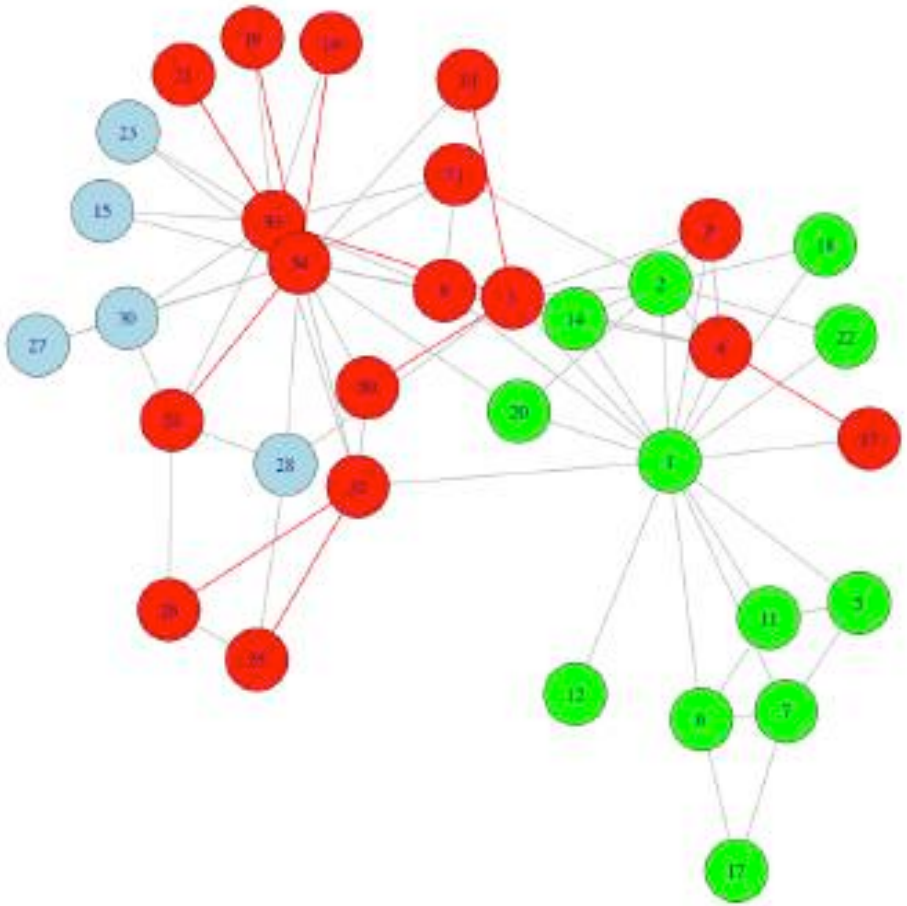
SIR model

$\beta = 0.5, \tau = 2$



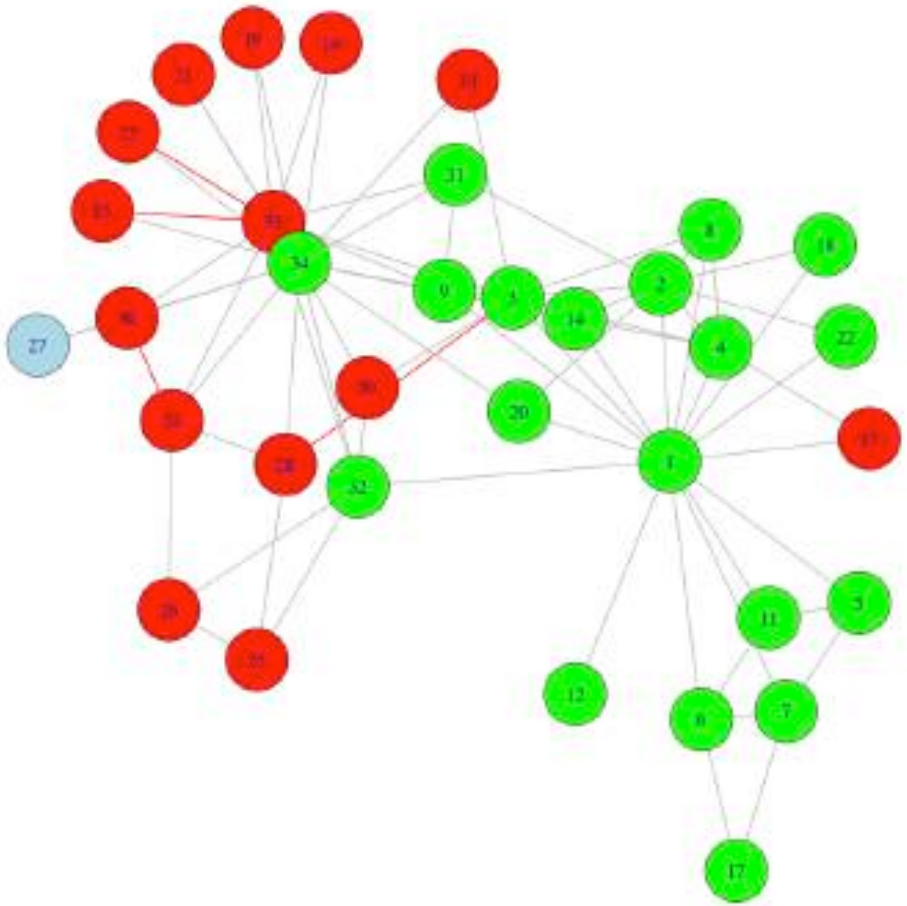
SIR model

$\beta = 0.5, \tau = 2$



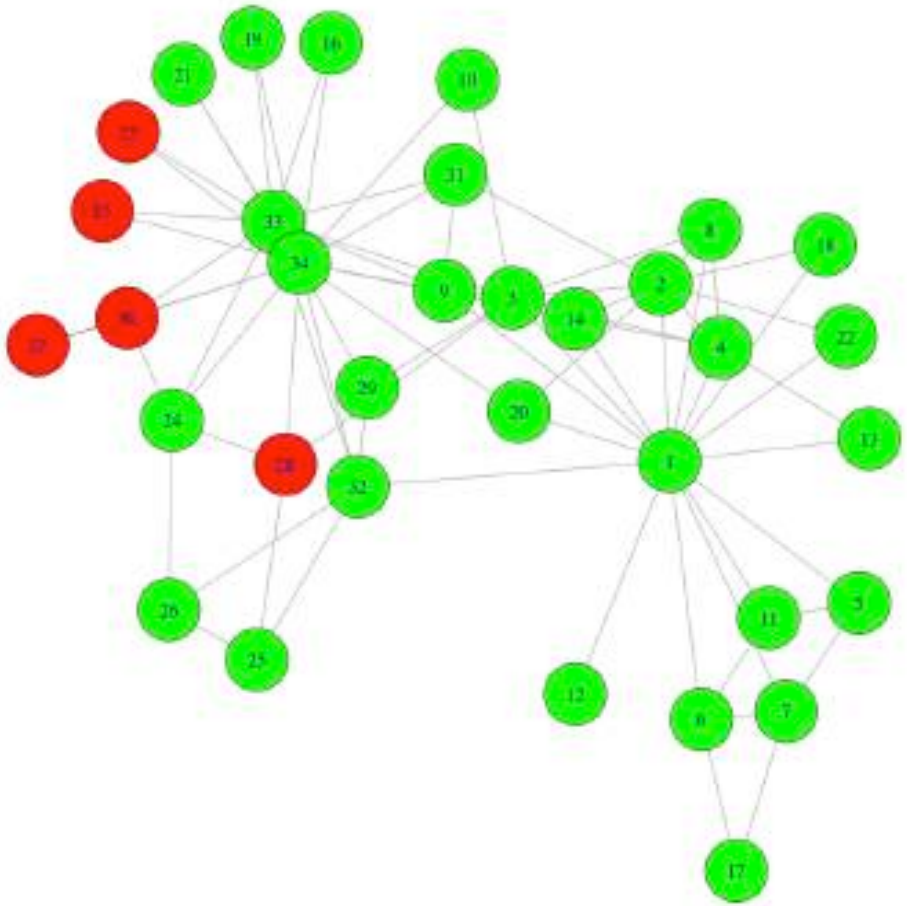
SIR model

$\beta = 0.5, \tau = 2$



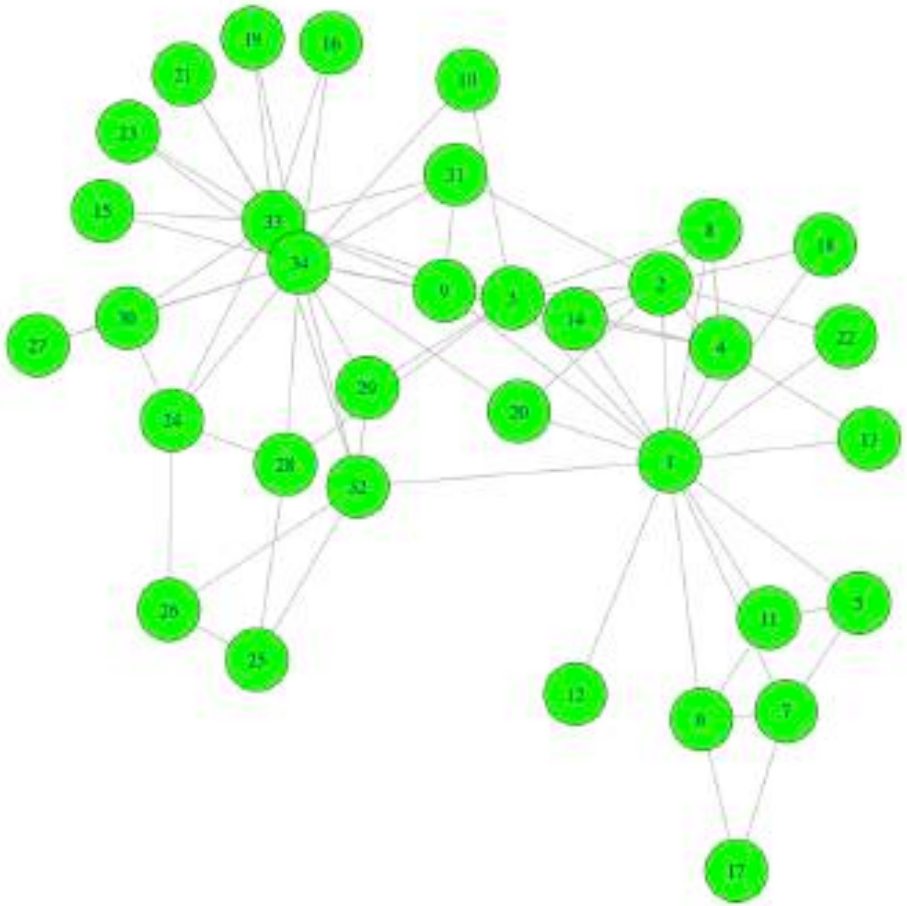
SIR model

$\beta = 0.5, \tau = 2$

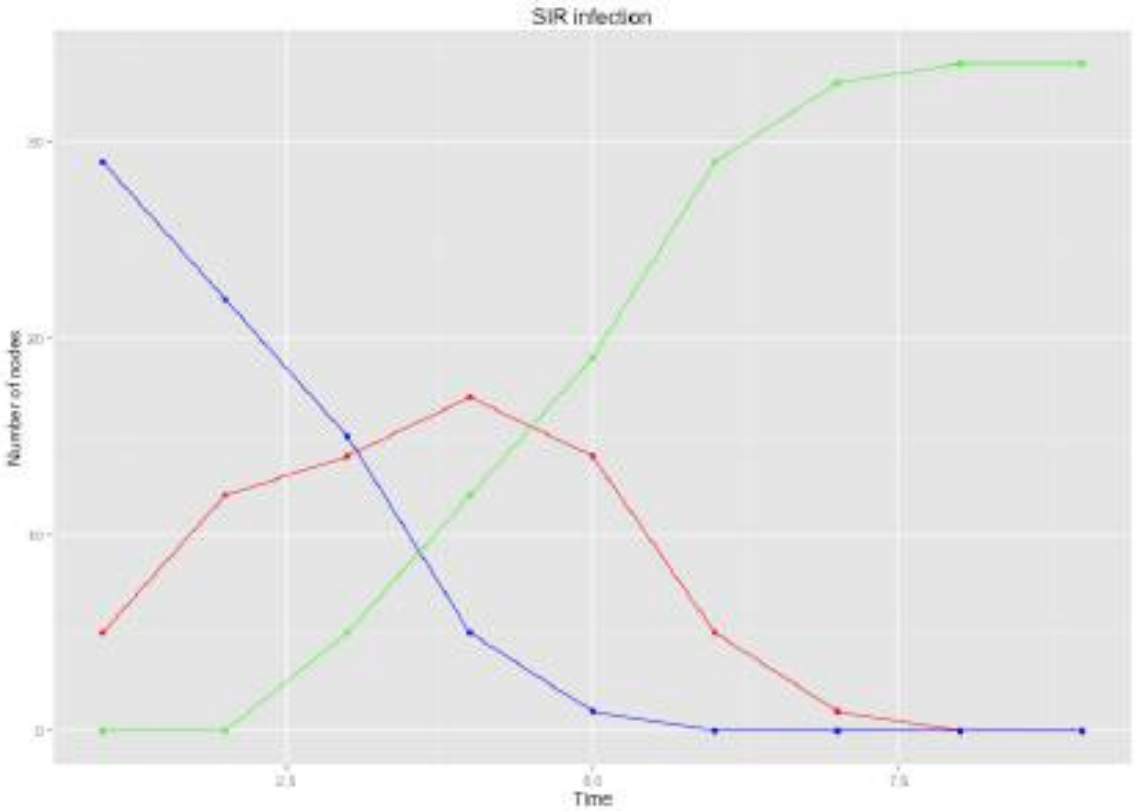


SIR model

$\beta = 0.5, \tau = 2$

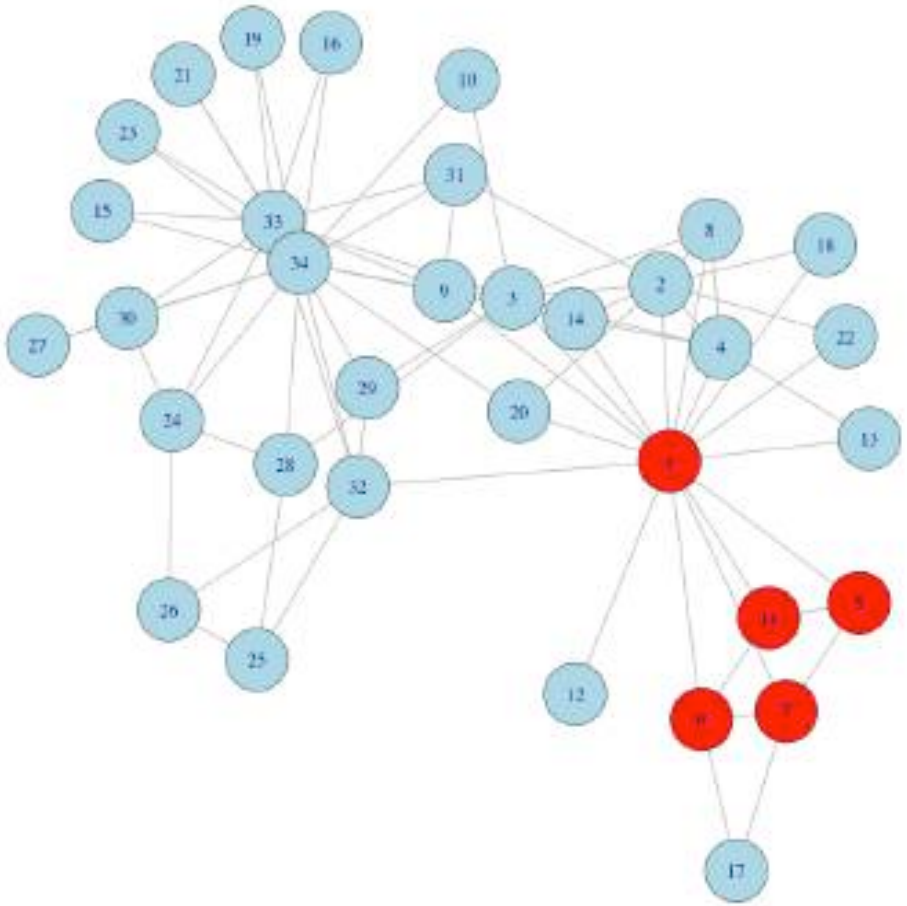


SIR model



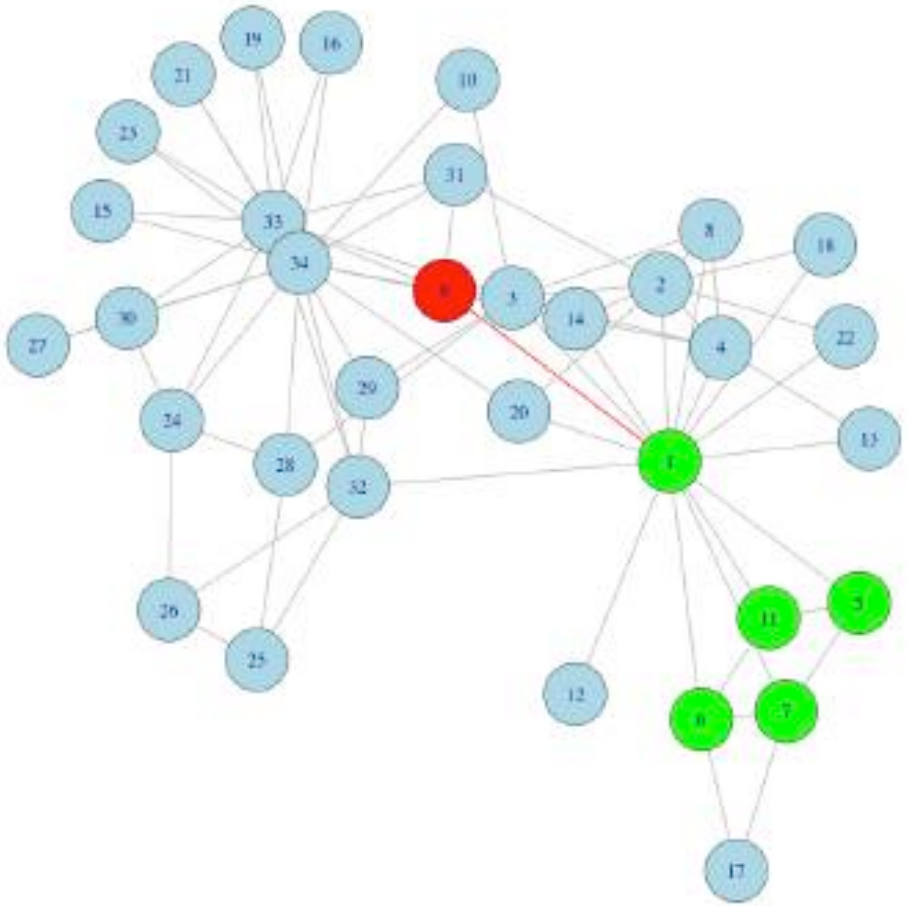
SIR model

$\beta = 0.2, \tau = 2$



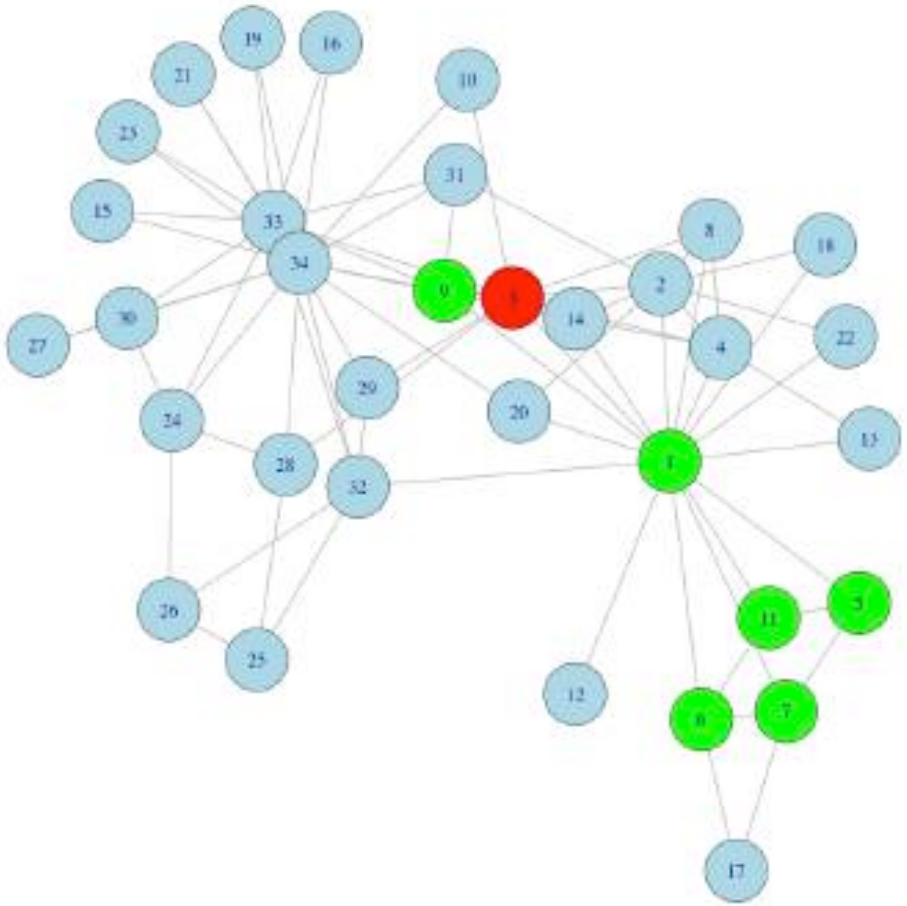
SIR model

$$\beta = 0.2, \tau = 2$$



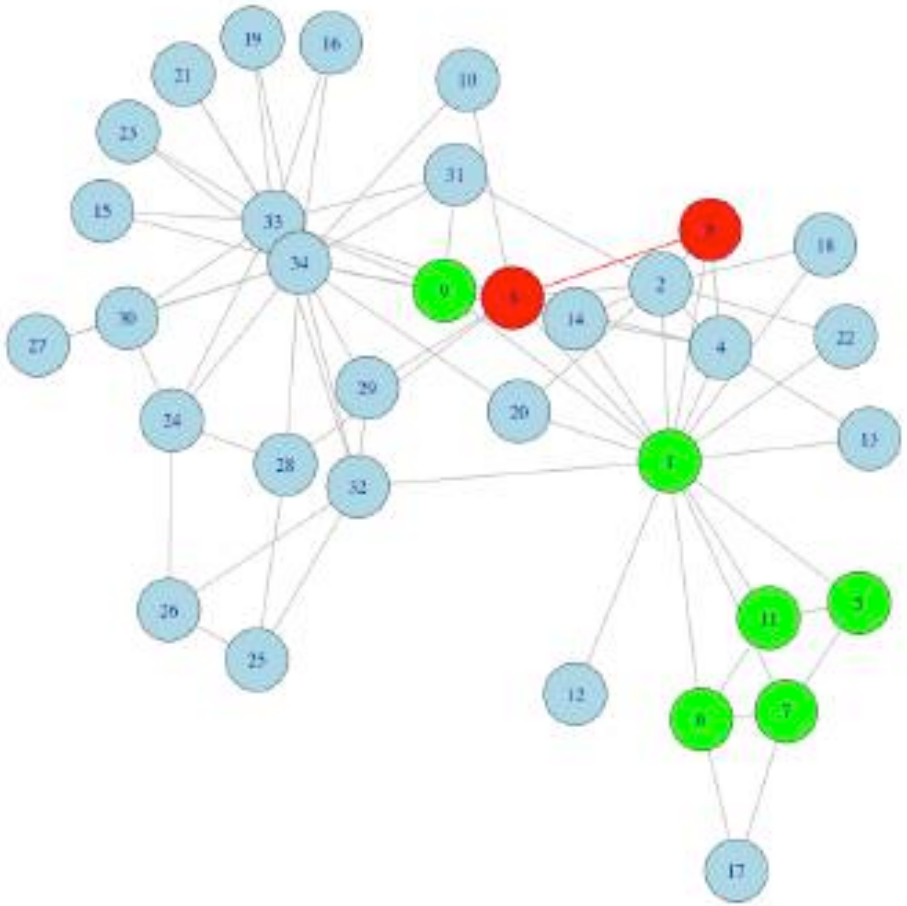
SIR model

$\beta = 0.2, \tau = 2$



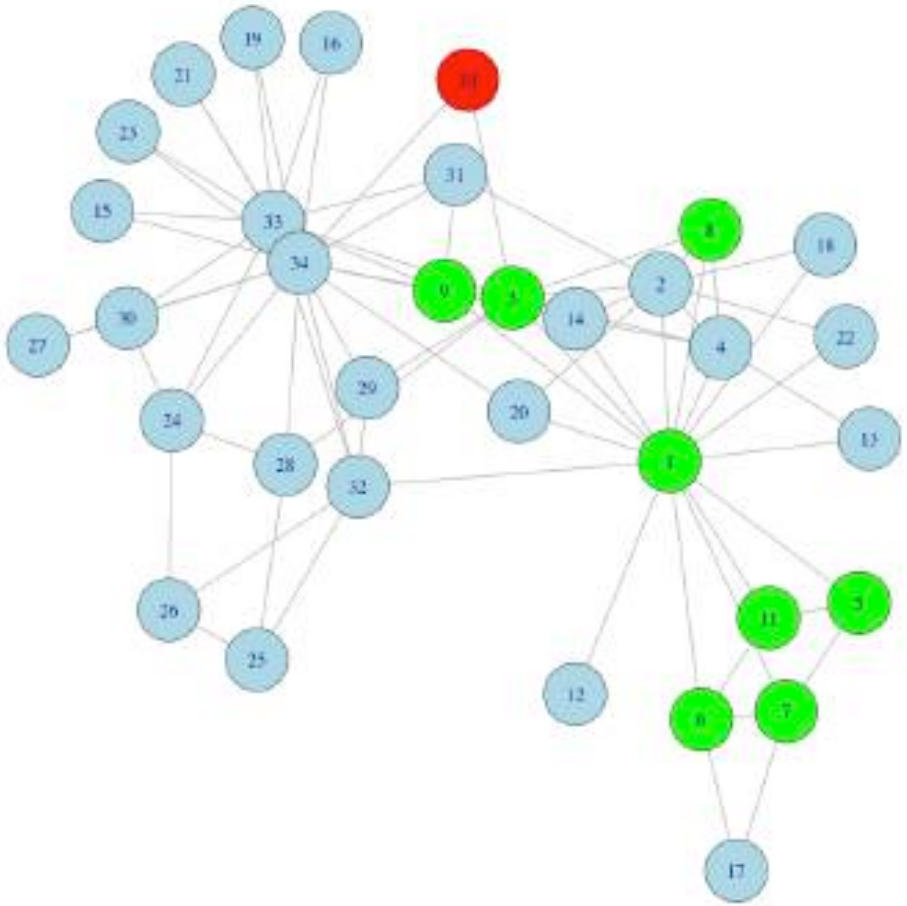
SIR model

$\beta = 0.2, \tau = 2$



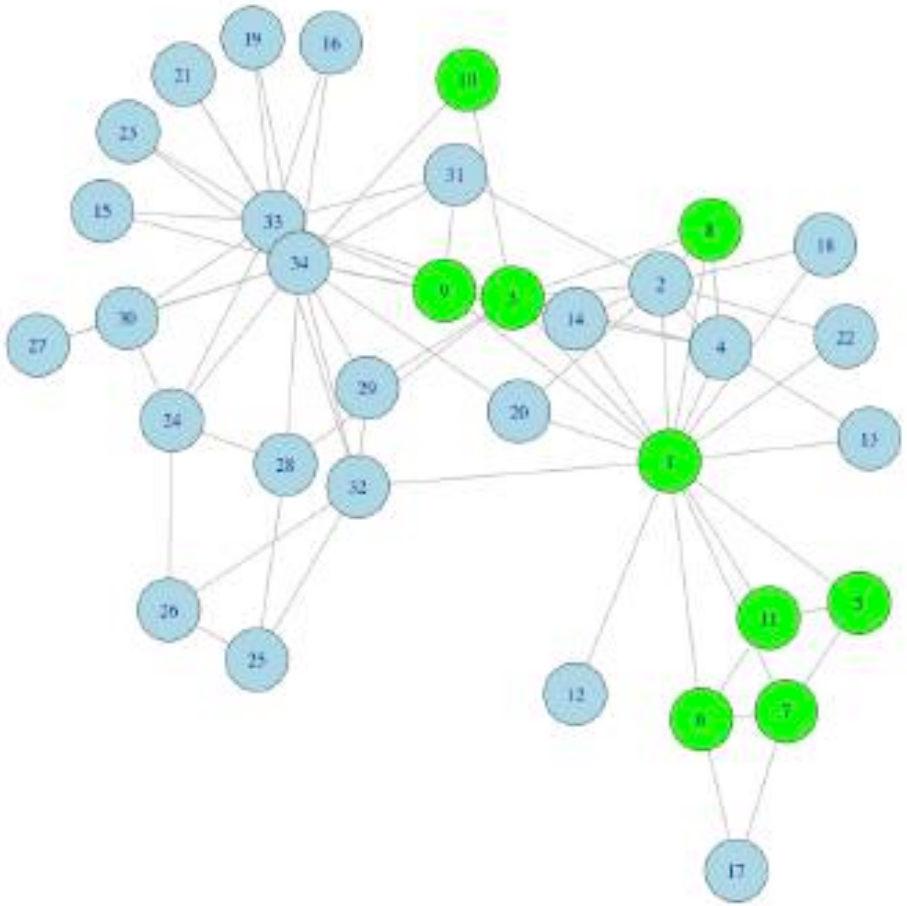
SIR model

$\beta = 0.2, \tau = 2$

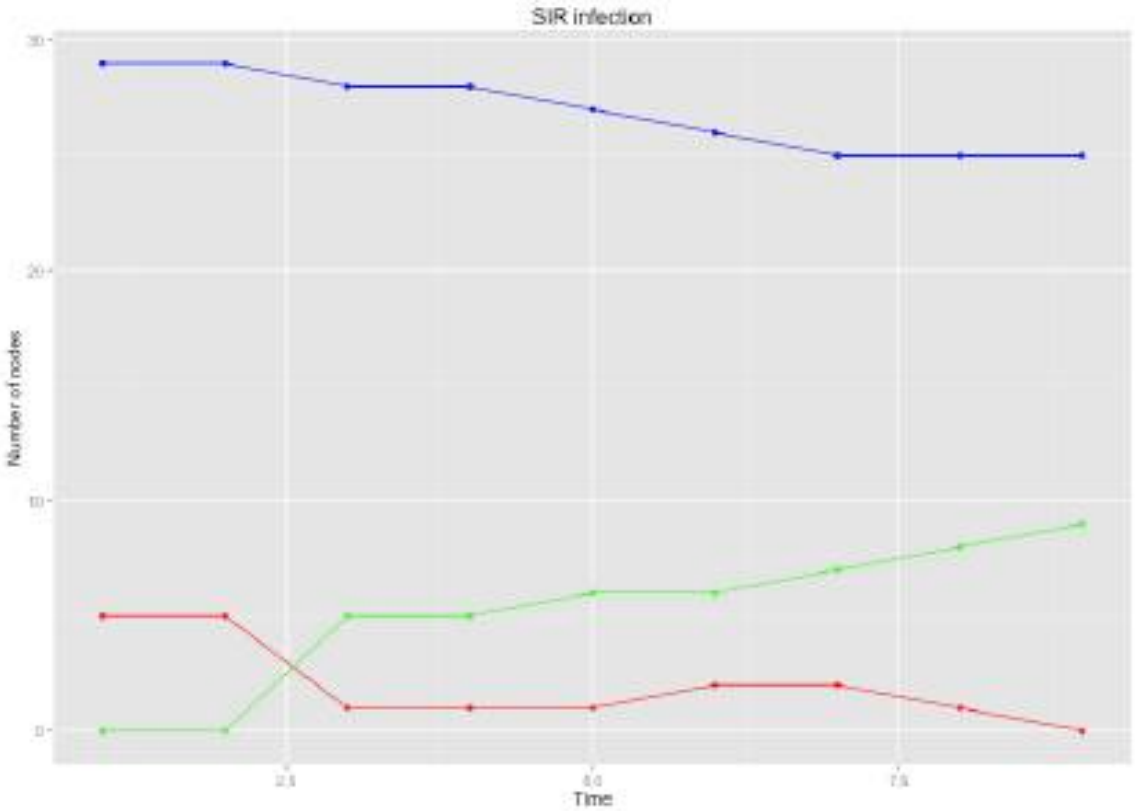


SIR model

$\beta = 0.2, \tau = 2$



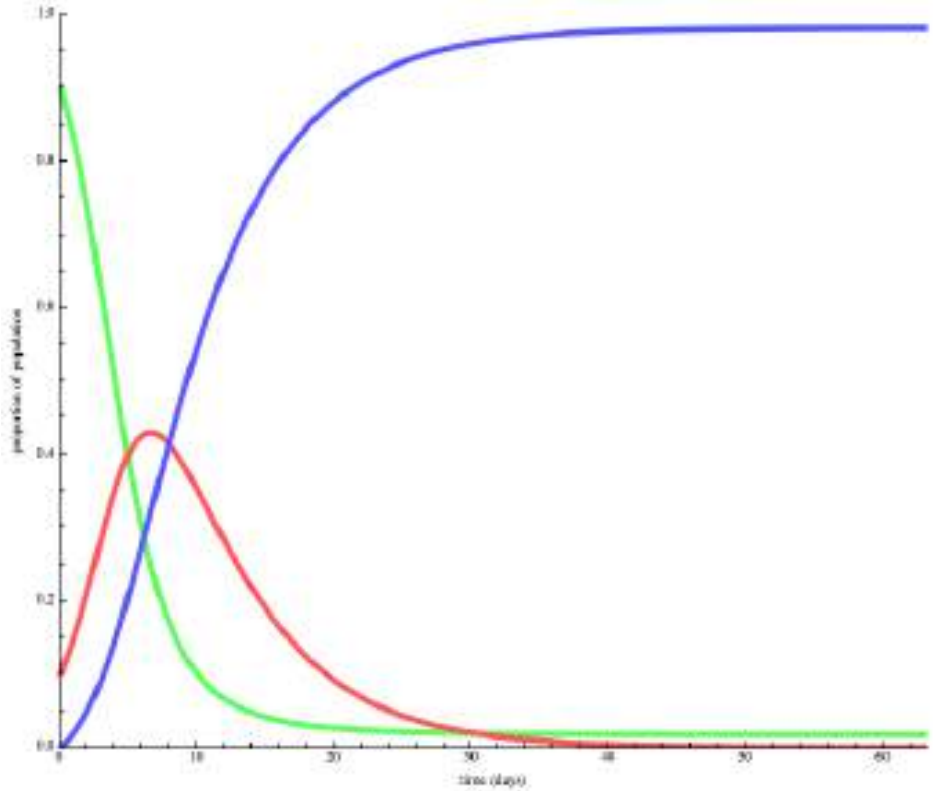
SIR model



SIR model

Epidemic threshold R_0:

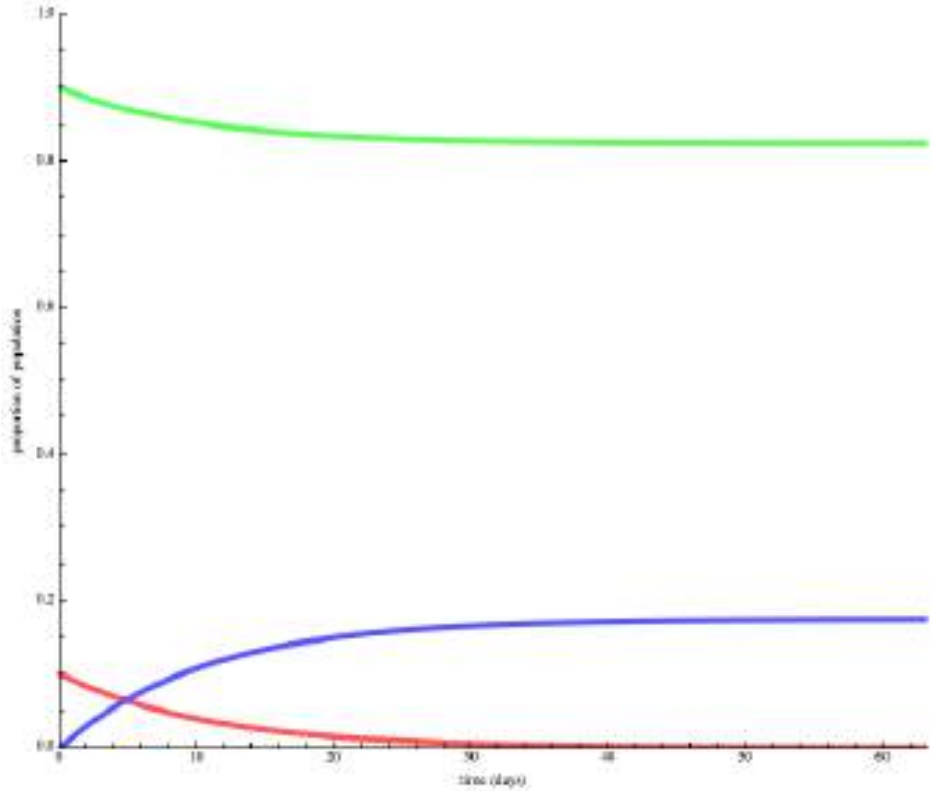
$\frac{\beta}{\gamma} > R_0$ - infection survives and becomes epidemic



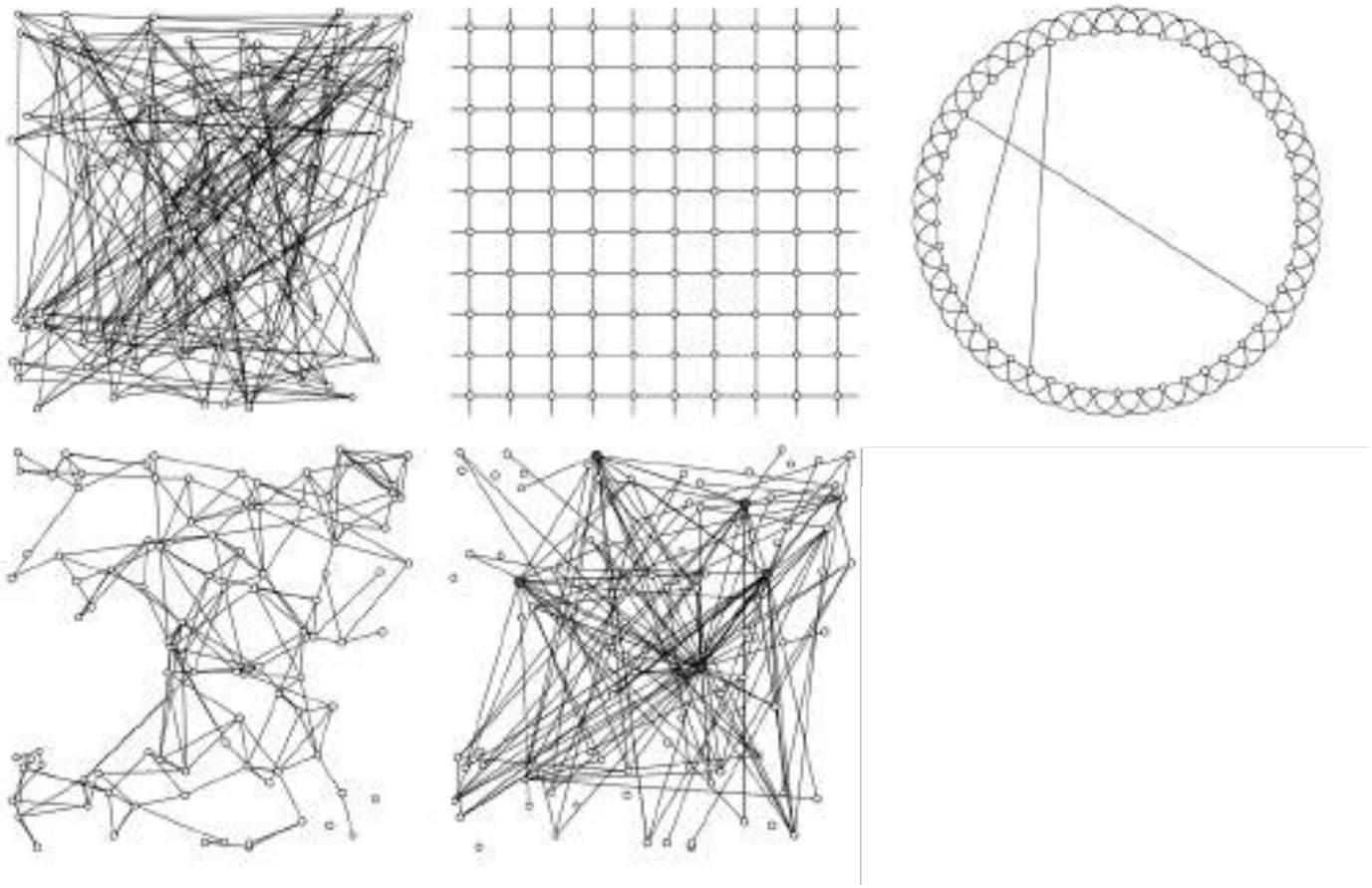
SIR model

Epidemic threshold R_0:

$\frac{\beta}{\gamma} < R_0$ - infection dies over time



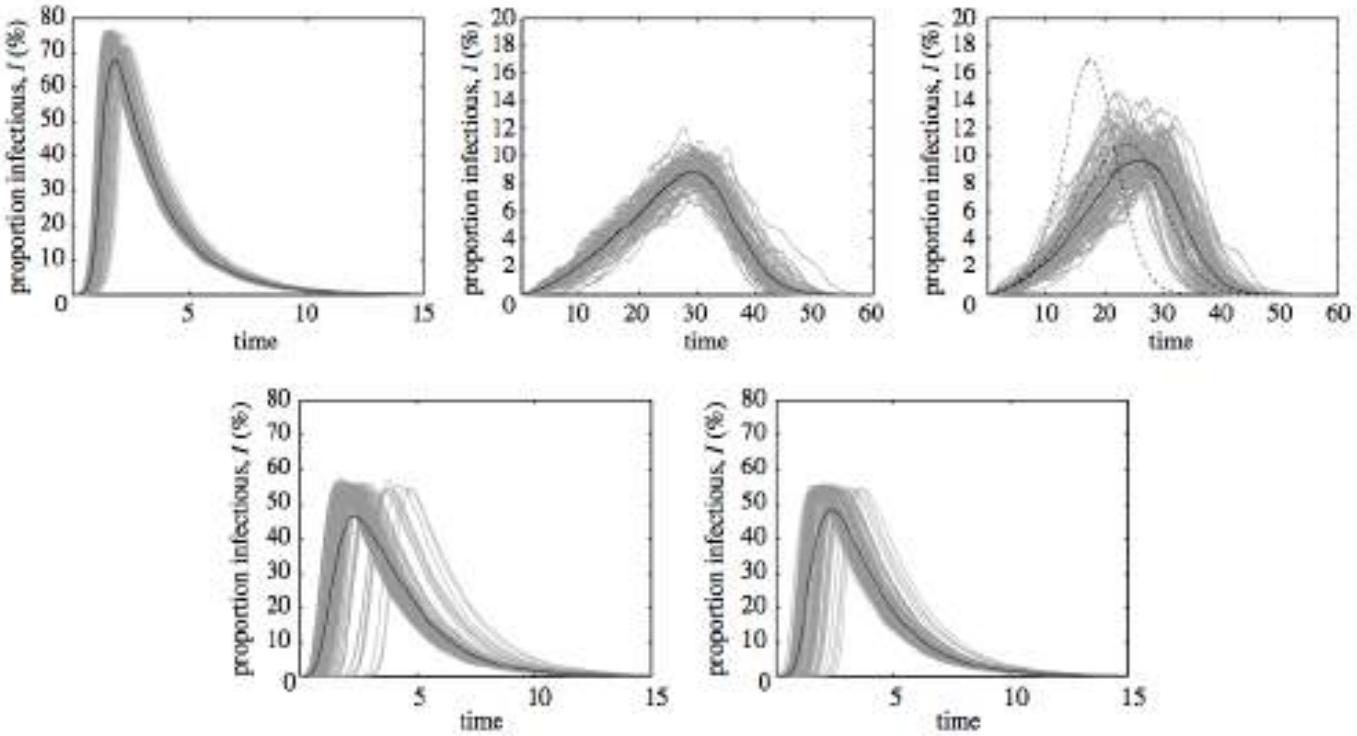
5 Networks, SIR



Networks: 1) random, 2) lattice, 3) small world, 4) spatial, 5) scale-free

image from Keeling et al, 2005

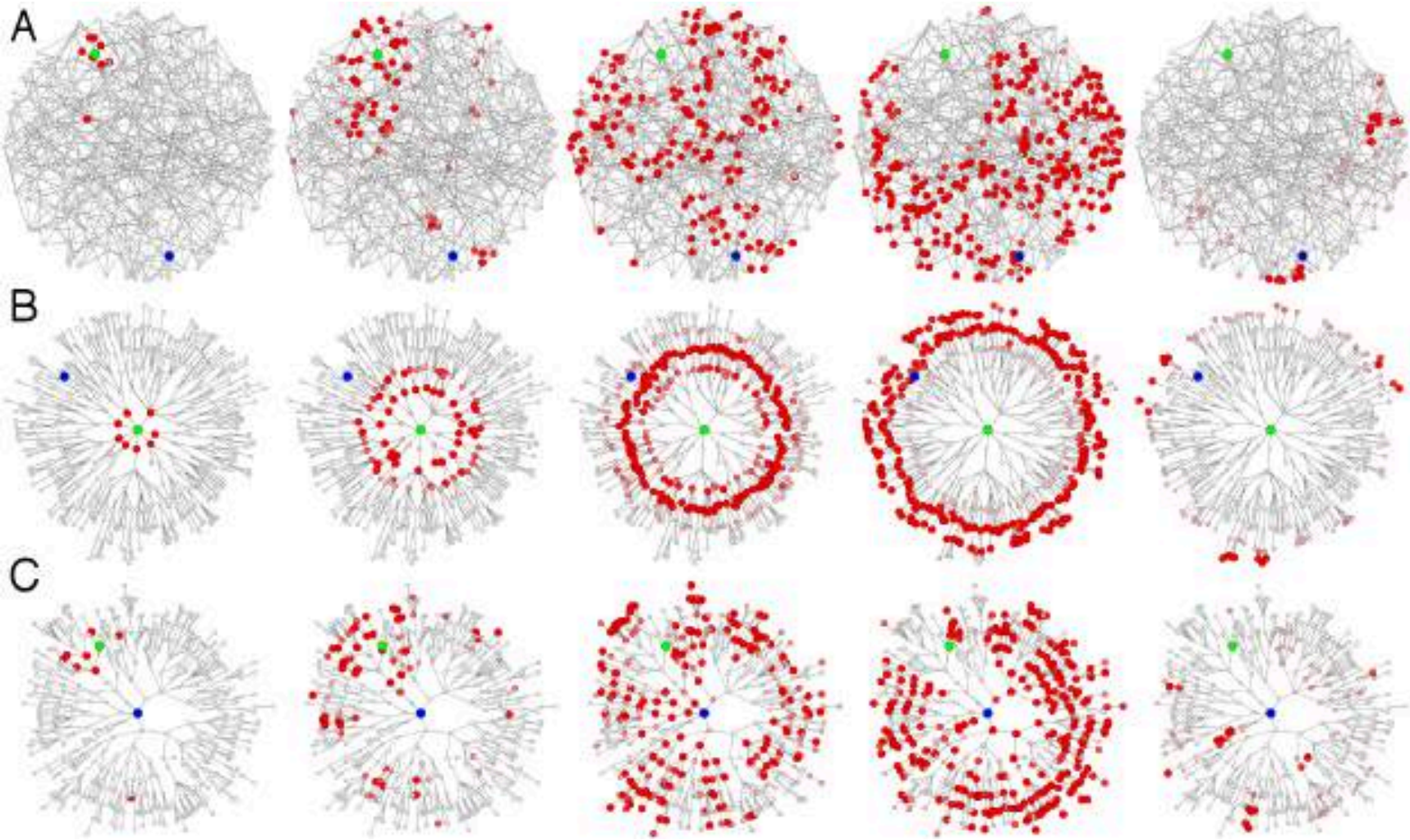
5 Networks, SIR



Networks: 1) random, 2) lattice, 3) small world, 4) spatial, 5) scale-free

Keeling et al, 2005

Effective distance



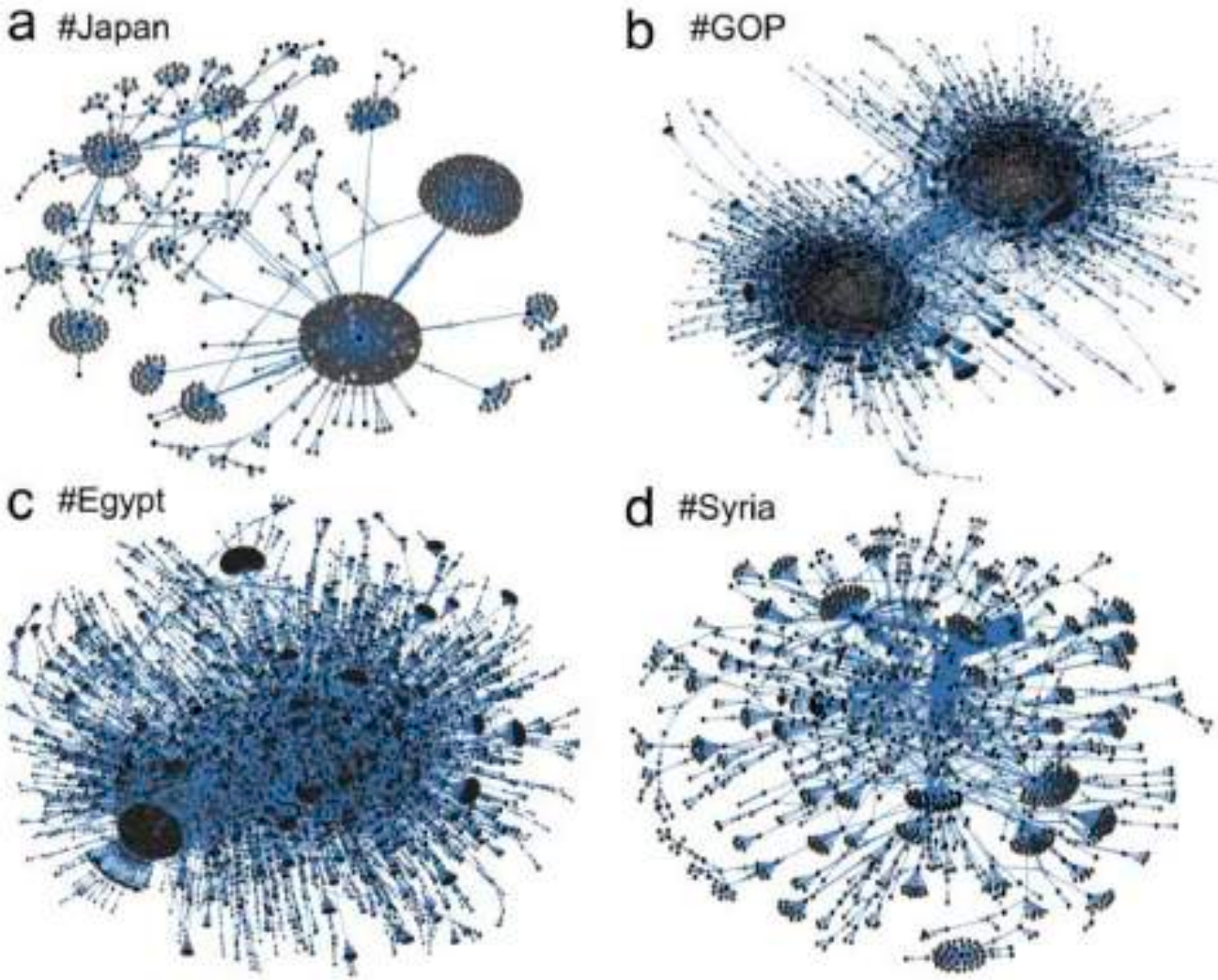
Social contagion

Social contagion phenomena refer to various processes that depend on the individual propensity to adopt and diffuse knowledge, ideas, information.

- Similar to epidemiological models:
  - ”susceptible” - an individual who has not learned new information
  - ”infected” - the spreader of the information
  - ”recovered” - aware of information, but no longer spreading it
- Two main questions:
  - if the rumor reaches high number of individuals
  - rate of infection spread

Mem diffusion

Mem diffusion on Twitter



References

- Epidemic outbreaks in complex heterogeneous networks. Y. Moreno, R. Pastor-Satorras, and A. Vespignani. Eur. Phys. J. B 26, 521-529, 2002.
- Networks and Epidemics Models. Matt. J. Keeling and Ken.T.D. Eames, J. R. Soc. Interfac, 2, 295-307, 2005
- Simulations of infections diseases on networks. G. Witten and G. Poulter. Computers in Biology and Medicine, Vol 37, No. 2, pp 195-205, 2007
- Dynamical processes on complex networks. A. Barrat, M. Barthelemy, A. Vespignani Eds., Cambridge University Press 2008
- Dynamics of rumor spreading in complex networks. Y. Moreno, M. Nekovee, A. Pacheco, Phys. Rev. E 69, 066130, 2004
- Theory of rumor spreading in complex social networks. M. Nekovee, Y. Moreno, G. Biaconi, M. Marsili, Physica A 374, pp 457-470, 2007

NATIONAL RESEARCH
UNIVERSITY

Diffusion of Innovation

Social Network Analysis. MAGoLEGO course. Lecture 8

Leonid Zhukov

lzhukov@hse.ru
www.leonidzhukov.net/hse/2018/sna

National Research University Higher School of Economics
School of Data Analysis and Artificial Intelligence, Department of Computer Science

Spreading process

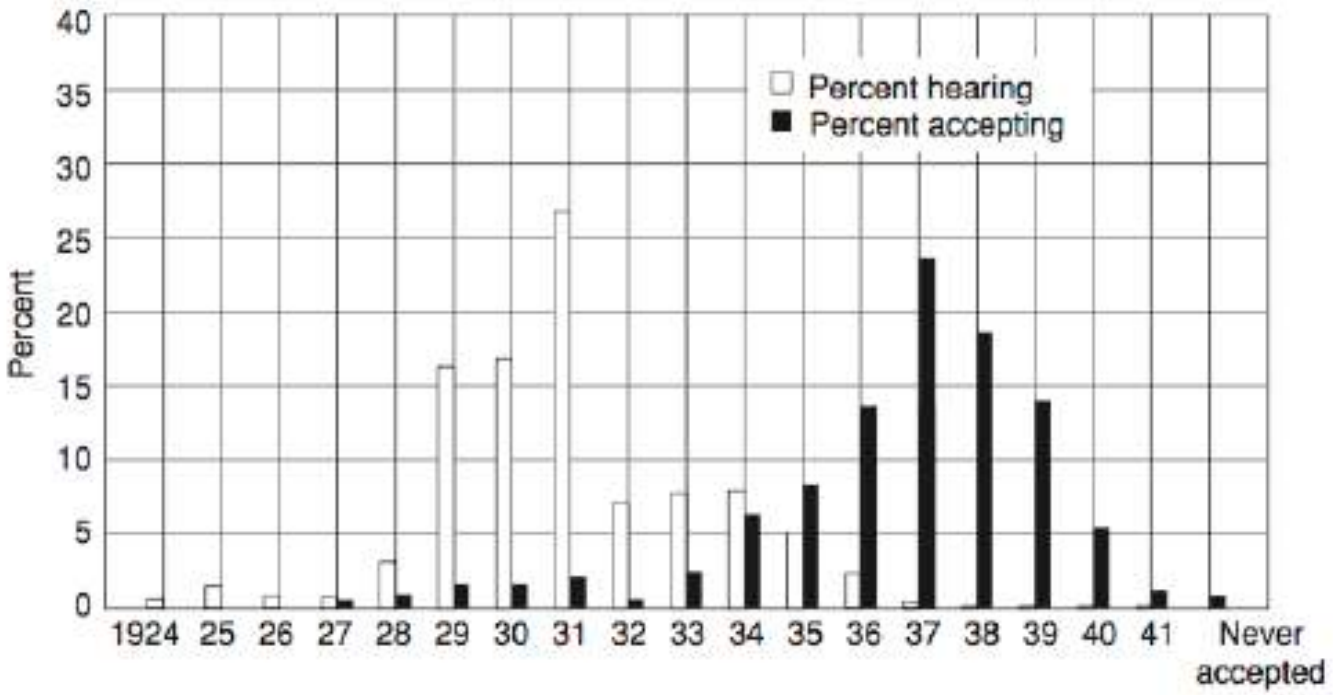
Propagation process:

- Information based models:
  - ideas, knowledge
  - virus and infection
  - rumors, news
- Decision based models:
  - adoption of innovation
  - joining political protest
  - purchase decision

Local individual decision rules will lead to very different global results.
"microscopic" changes $\rightarrow$ "macroscopic" results

Ryan-Gross study

Ryan-Gross study of hybrid seed corn delayed adoption (after first exposure)

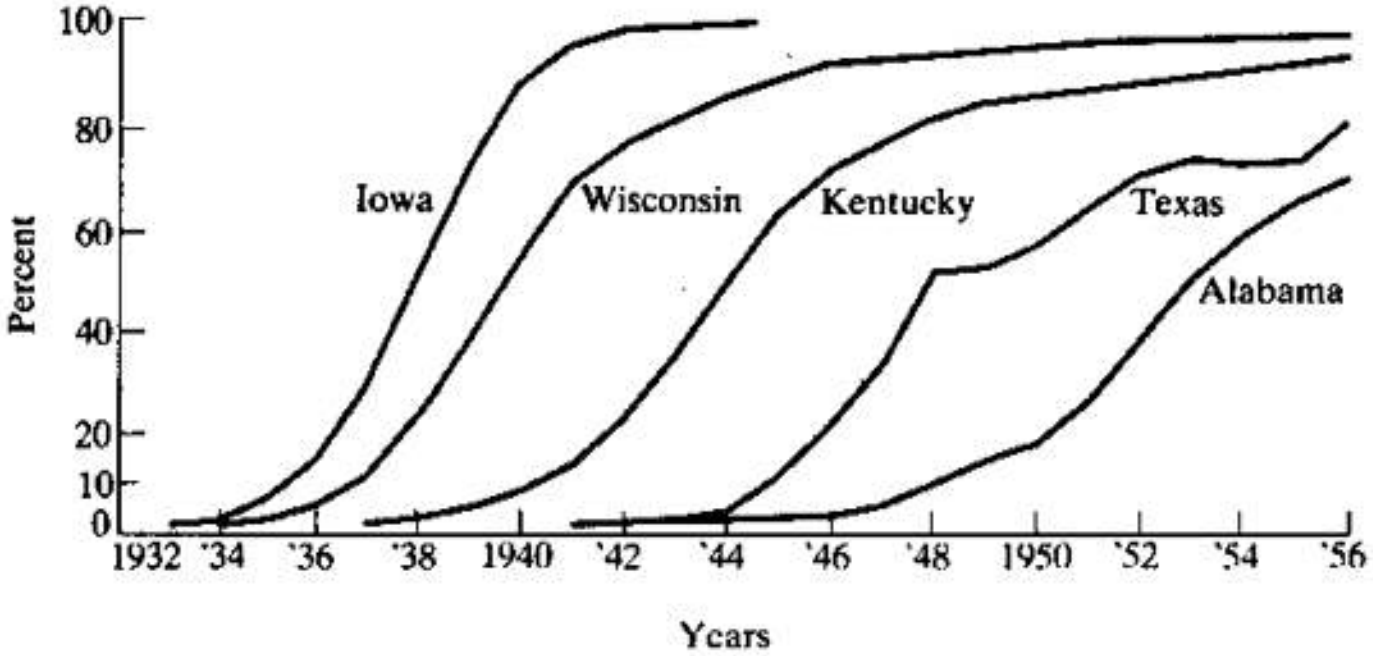


Information effect vs adopting of innovation

Ryan and Gross, 1943

Ryan-Gross study

Hybrid corn adoption

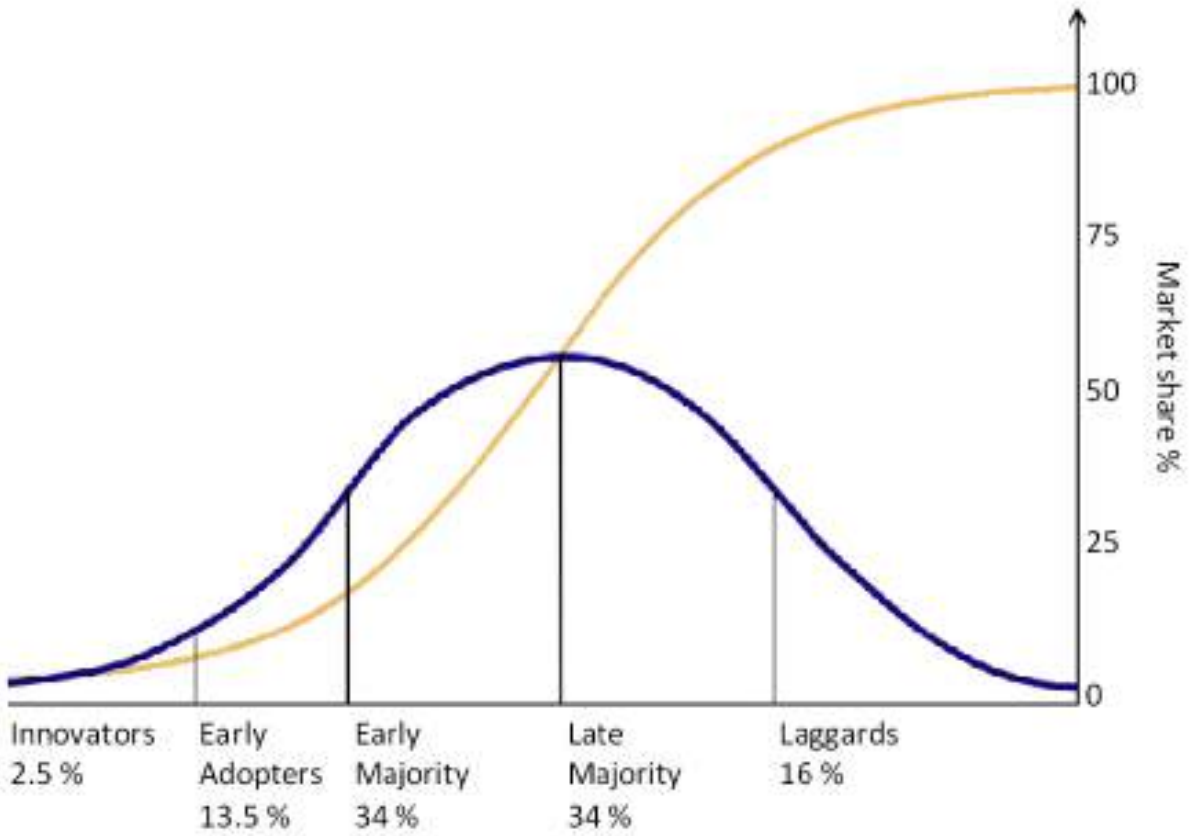


Percentage of total acreage planted

Griliches, 1957

Diffusion of innovation

Everett Rogers, "Diffusion of innovation" book, 1962



Frank Bass, 1969, "A new product growth model for consumer durables"

Diffusion of innovation

What influences potential adopters:

- relative advantage of the innovation
- compatibility with current ways of doing things
- complexity of the innovation
- triability - the ease of testing
- observability of results

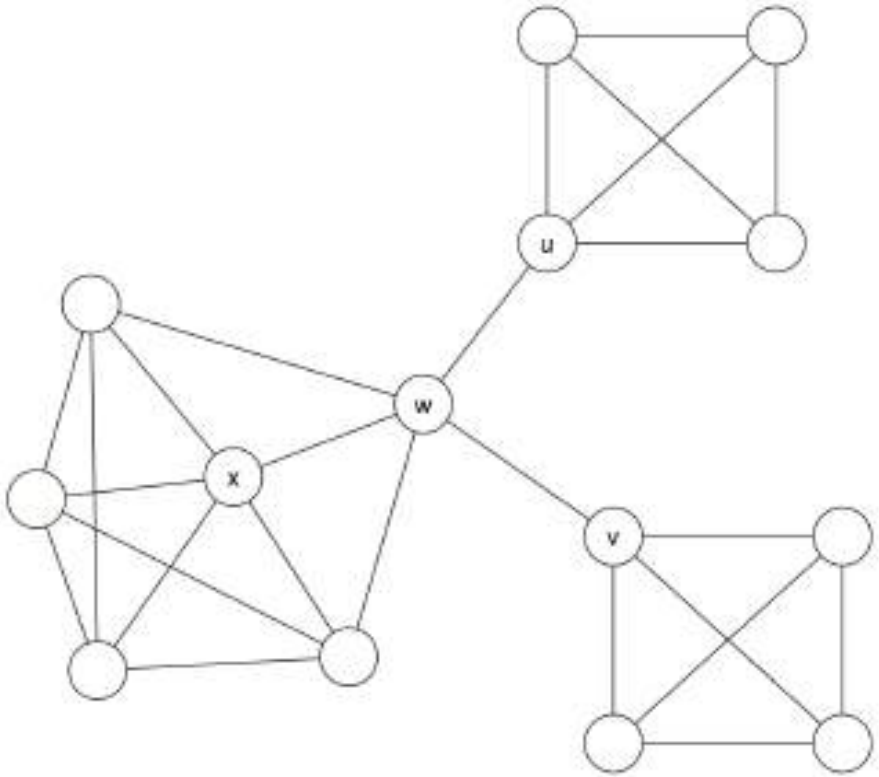
Some questions remain:

- how a new technology can take over?
- who different technologies coexist?
- what stops new technology propagation?

Everett Rogers,1962

Network structure

From the population level to local structure



Network coordination game

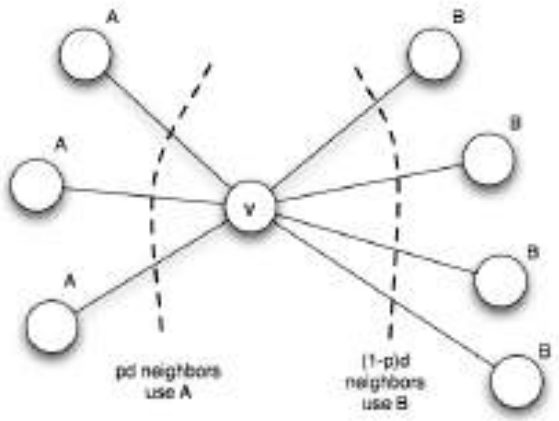
Local interaction game: Let *u* and *v* are players, and *A* and *b* are possible strategies
Payoffs

- if *u* and *v* both adopt behavior A, each get payoff $a > 0$
- if *u* and *v* both adopt behavior B, each get payoff $b > 0$
- if *u* and *v* adopt opposite behavior, each get payoff 0

		w: A	w: B
v	A	a, a	$0, 0$
	B	$0, 0$	b, b

Threshold model

Network coordination game, direct-benefit effect



v \ w	A	B
A	a, a	$0, 0$
B	$0, 0$	b, b

Node v to make decision A or B, p - portion of type A neighbors to accept A:

$$a \cdot p \cdot d > b \cdot (1 - p) \cdot d$$

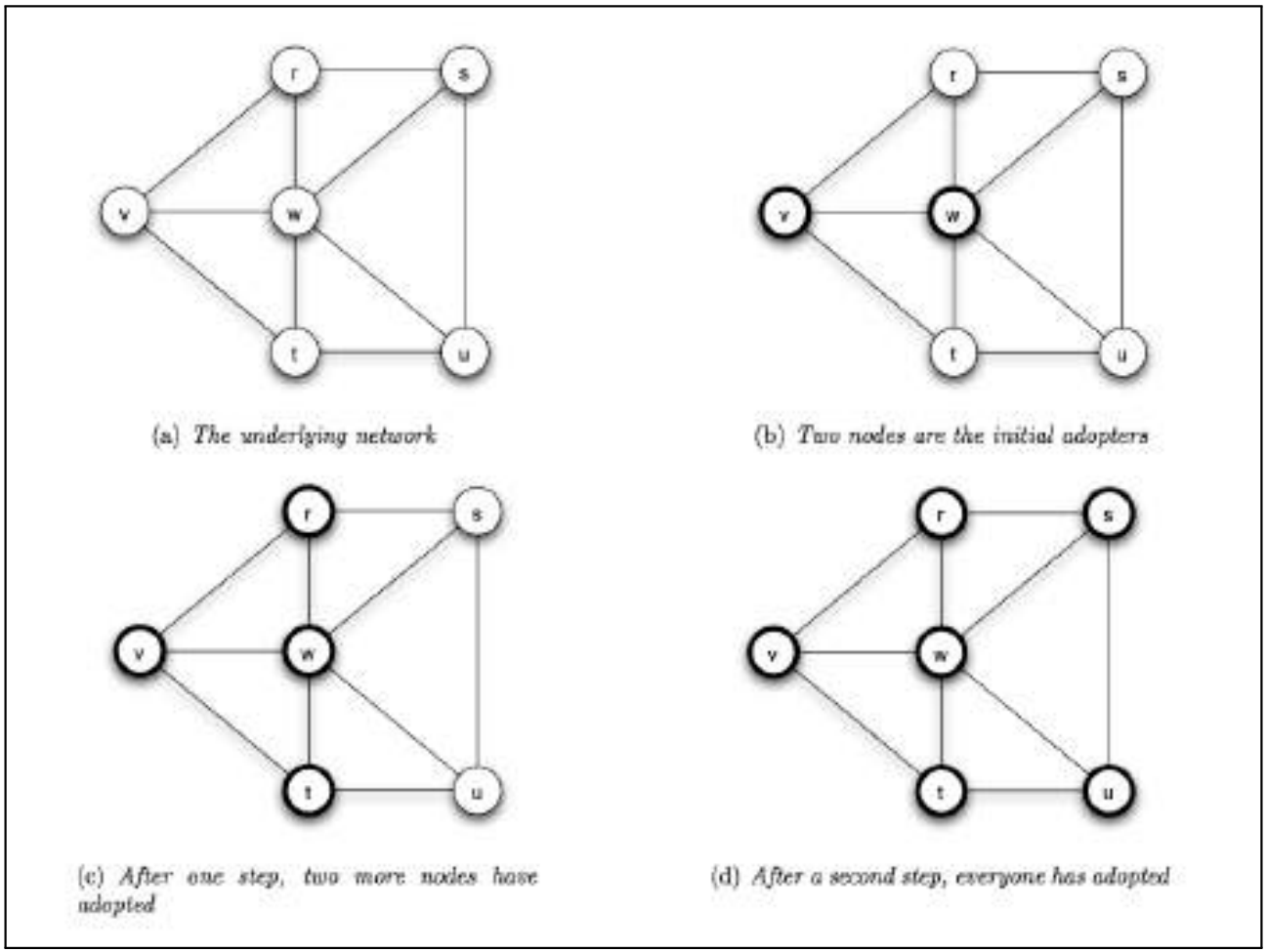
$$p \geq b/(a + b)$$

Threshold:

$$q = \frac{b}{a + b}$$

Cascades

Cascade - sequence of changes of behavior, "chain reaction"



(a) *The underlying network*

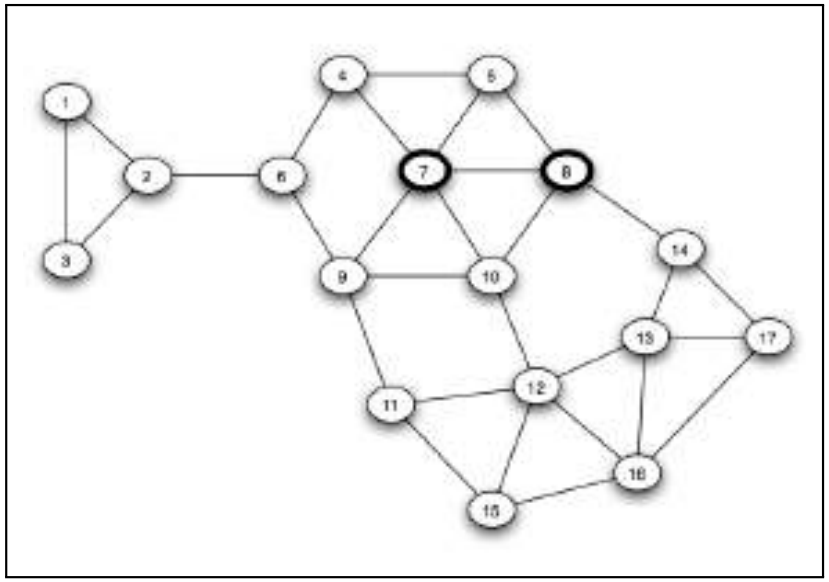
(b) *Two nodes are the initial adopters*

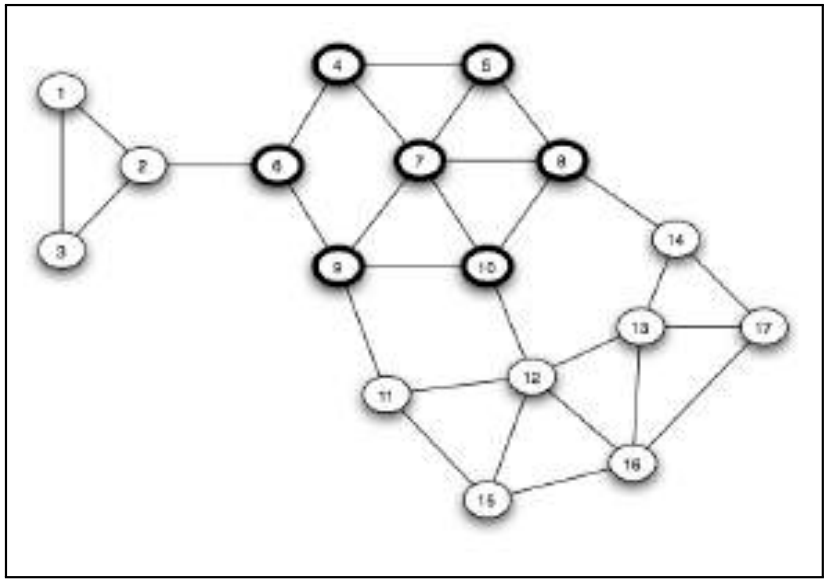
(c) *After one step, two more nodes have adopted*

(d) *After a second step, everyone has adopted*

Let $a = 3, b = 2$, threshold $q = 2/(2+3) = 2/5$

Cascade propagation

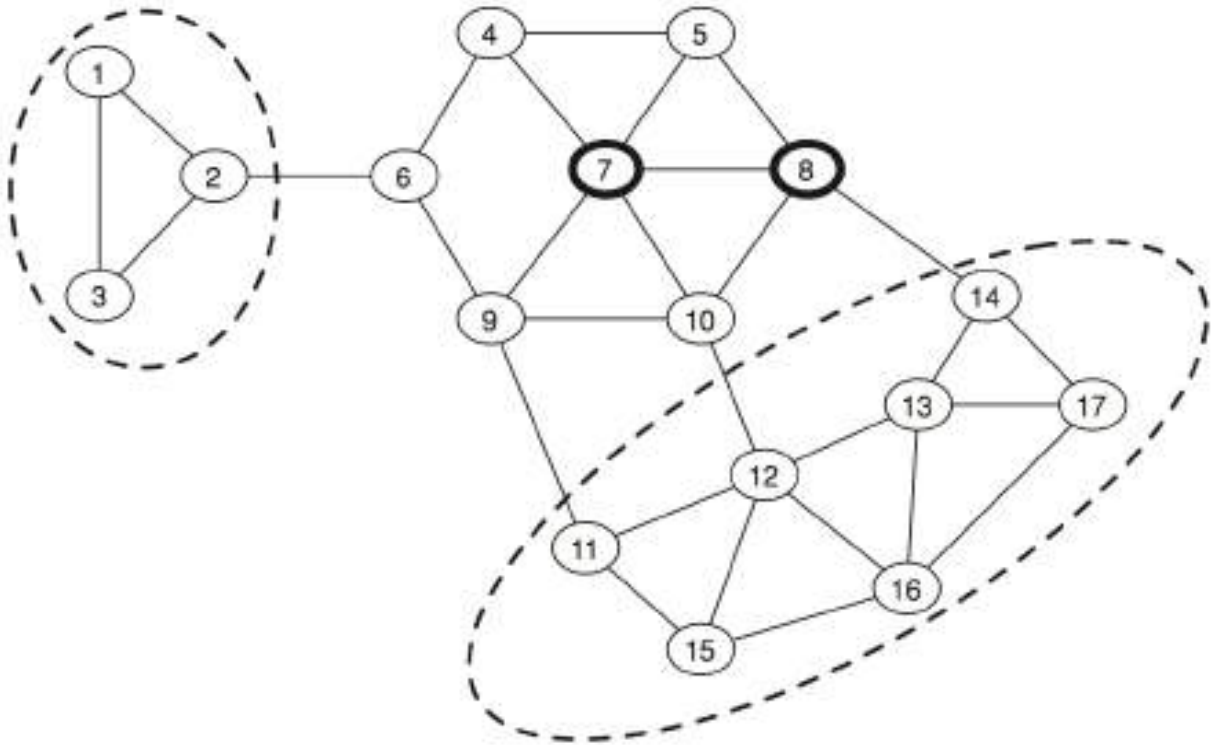




- Let $a = 3, b = 2$, threshold $q = 2/(2+3) = 2/5$
- Start from nodes 7,8: $1/3 < 2/5 < 1/2 < 2/3$
- Cascade size - number of nodes that changed the behavior
- Complete cascade when every node changes the behavior

Cascades and clusters

Group of nodes form a cluster of density ρ if every node in the set has at least fraction ρ of its neighbors in the set



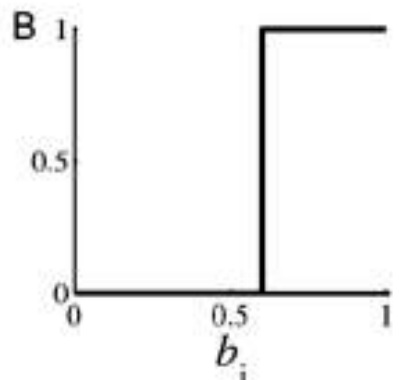
Both clusters of density $\rho = 2/3$. For cascade to get into cluster $q \leq 1 - \rho$.

Linear threshold model

- Influence comes only from NN $N(i)$ nodes, w_{ij} influence $i \to j$
- Require $\sum_{j \in N(i)} w_{ji} \leq 1$
- Each node has a random acceptance threshold from $\theta_i \in [0, 1]$
- Activation: fraction of active nodes exceeds threshold

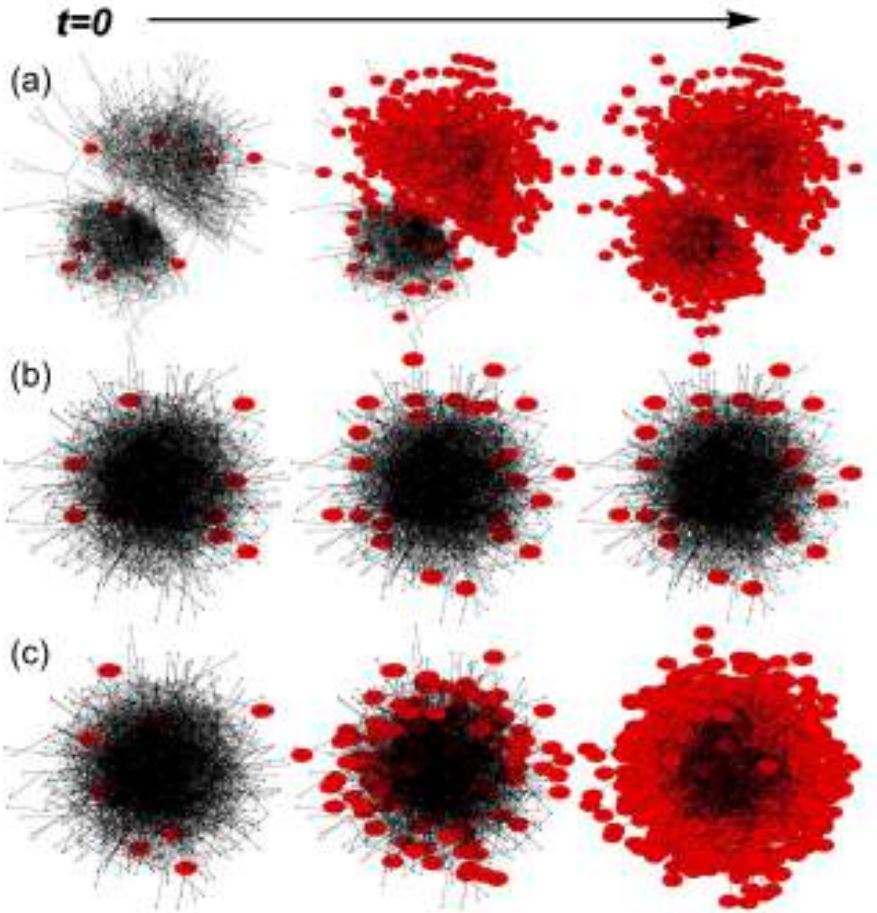
$$\sum_{\text{active } j \in N(i)} w_{ji} > \theta_i$$

- Initial set of active nodes A_o, iterative process with discrete time steps
- Progressive process, only nonactive $\to$ *active*



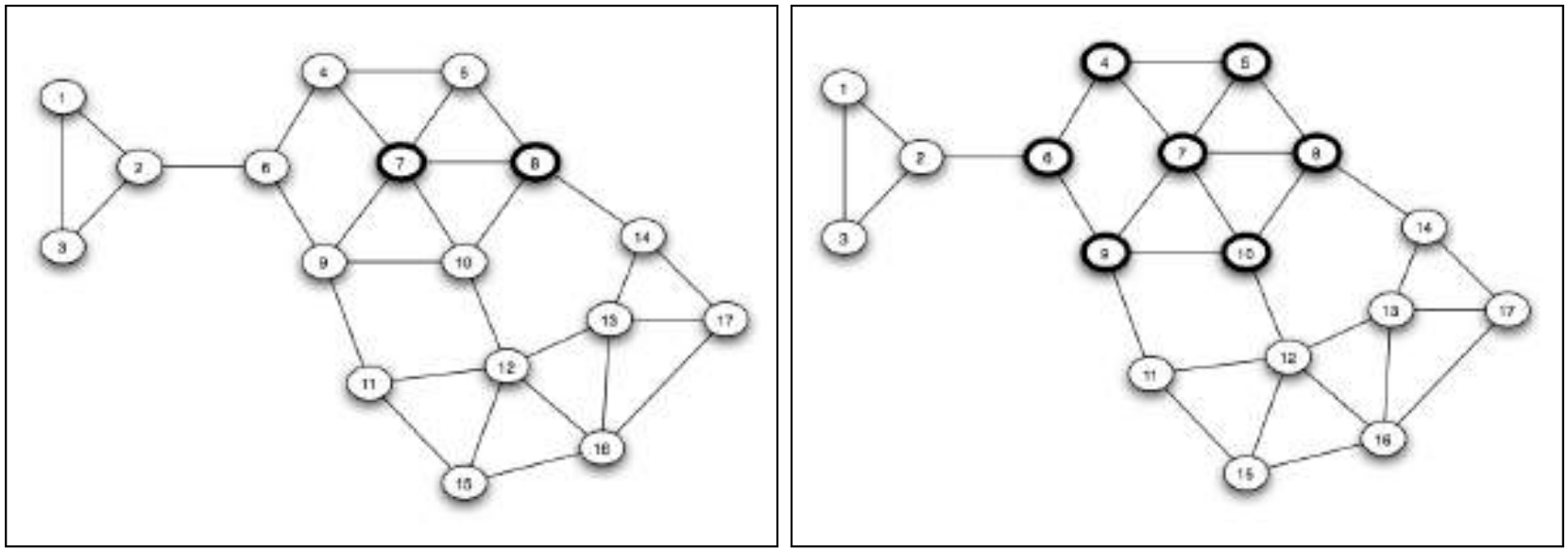
Cascades in random networks

multiple seed nodes



(a) Empirical network; (b), (c) - randomized network

Influence maximization problem



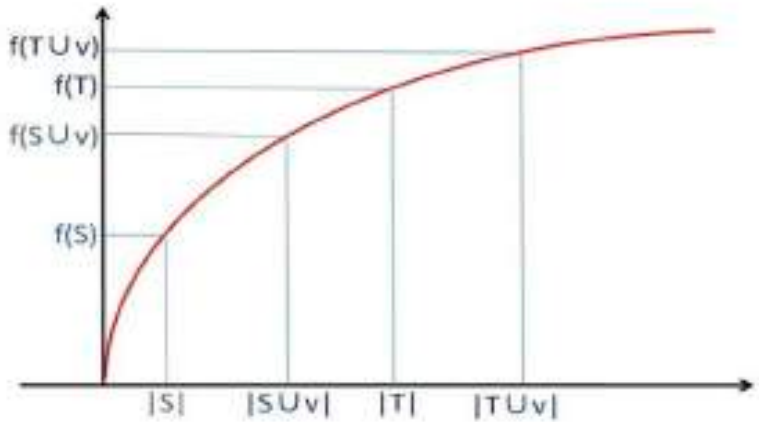
- Initial set of active nodes A_o
- Cascade size $\sigma(A_o)$ - expected number of active nodes when propagation stops
- Find *k*-set of nodes A_o that produces maximal cascade $\sigma(A_o)$
- *k*-set of "maximum influence" nodes
- NP-hard

D. Kempe, J. Kleinberg, E. Tardos, 2003, 2005

Influence maximization

Greedy maximization algorithm:
Given: Graph and set size k
Output: Maximum influence set A

1. Select a node v_1 that maximizes the influence $\sigma(v_1)$
2. Fix v_1 and find v_2 such that maximizes $\sigma(v_1, v_2)$
3. Repeat k times
4. Output maximum influence set: $A = \{v_1, v_2 \ldots v_k\}$



Influence maximization

Approximation algorithm

Algorithm: Greedy optimization

Input: Graph $G(V, E)$, k

Output: Maxumum influence set S

Set $S \leftarrow 0$

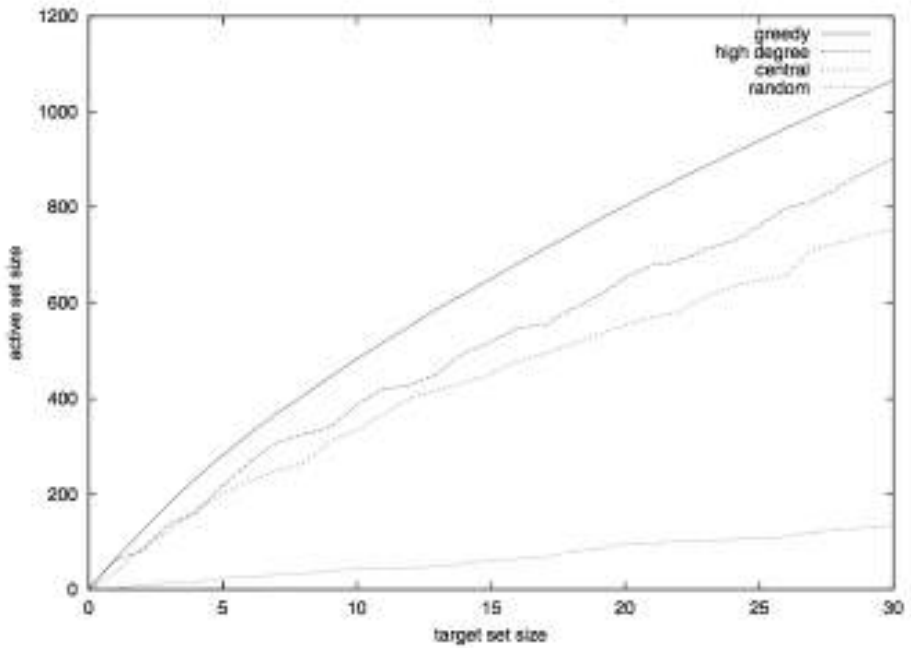
for $i = 1 : k$ **do**

  select $v = \arg\max_{u \in V \setminus S}(\sigma(S \cup \{u\}) - \sigma(S))$

  $S \leftarrow S \cup \{v\}$

Experimental results

Linear threshold model
network: collaboration graph 10,000 nodes, 53,000 edges



Greedy algorithm finds a set S such that its influence set $\sigma(S)$ is $\sigma(S) \geq (1 - \frac{1}{e})\sigma(S^*)$ from the true optimal (maximal) set $\sigma(S^*)$

D. Kempe, J. Kleinberg, E. Tardos, 2003

References

- Contagion, S. Morris, Review of Economic Studies, 67, p 57-78, 2000
- Maximizing the Spread of Influence through a Social Network, D. Kempe, J. Kleinberg, E. Tardos, 2003
- Influential Nodes in a Diffusion Model for Social Networks, D. Kempe, J. Kleinberg, E. Tardos, 2005
- A Simple Model of Global Cascades on Random Networks. D. Watts, 2002.

NATIONAL RESEARCH
UNIVERSITY

Spatial Models of Segregation

Social Network Analysis. MAGoLEGO course.
Lecture 9

Leonid Zhukov

lzhukov@hse.ru
www.leonidzhukov.net/hse/2018/sna

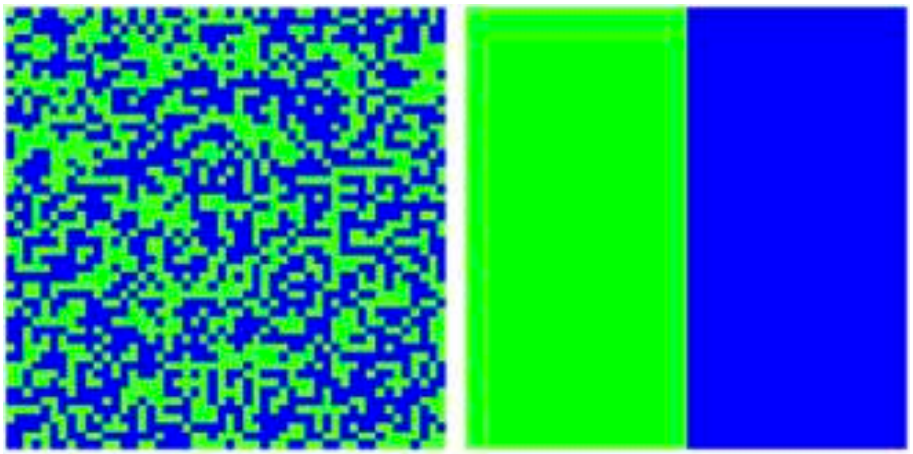
National Research University Higher School of Economics
School of Data Analysis and Artificial Intelligence, Department of Computer Science

Spatial model of segregation

"Dynamic Models of Segregation", Thomas Schelling, 1971

- Micro-motives and macro-behavior
- Personal preferences lead to collective actions
- Global patterns of spatial segregation from homophily at a local level
- Segregated race, ethnicity, native language, income
- Cities are strongly racially segregated. Are people that racists?
- Agent based modeling: agents, rules (dynamics), aggregation

Segregation



Integrated pattern Segregated pattern

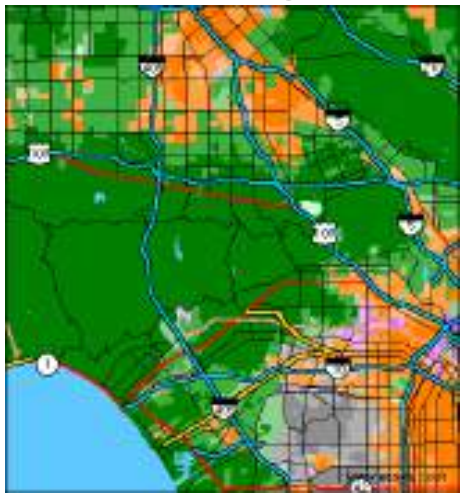
Racial segregation

New York

Washington

Chicago

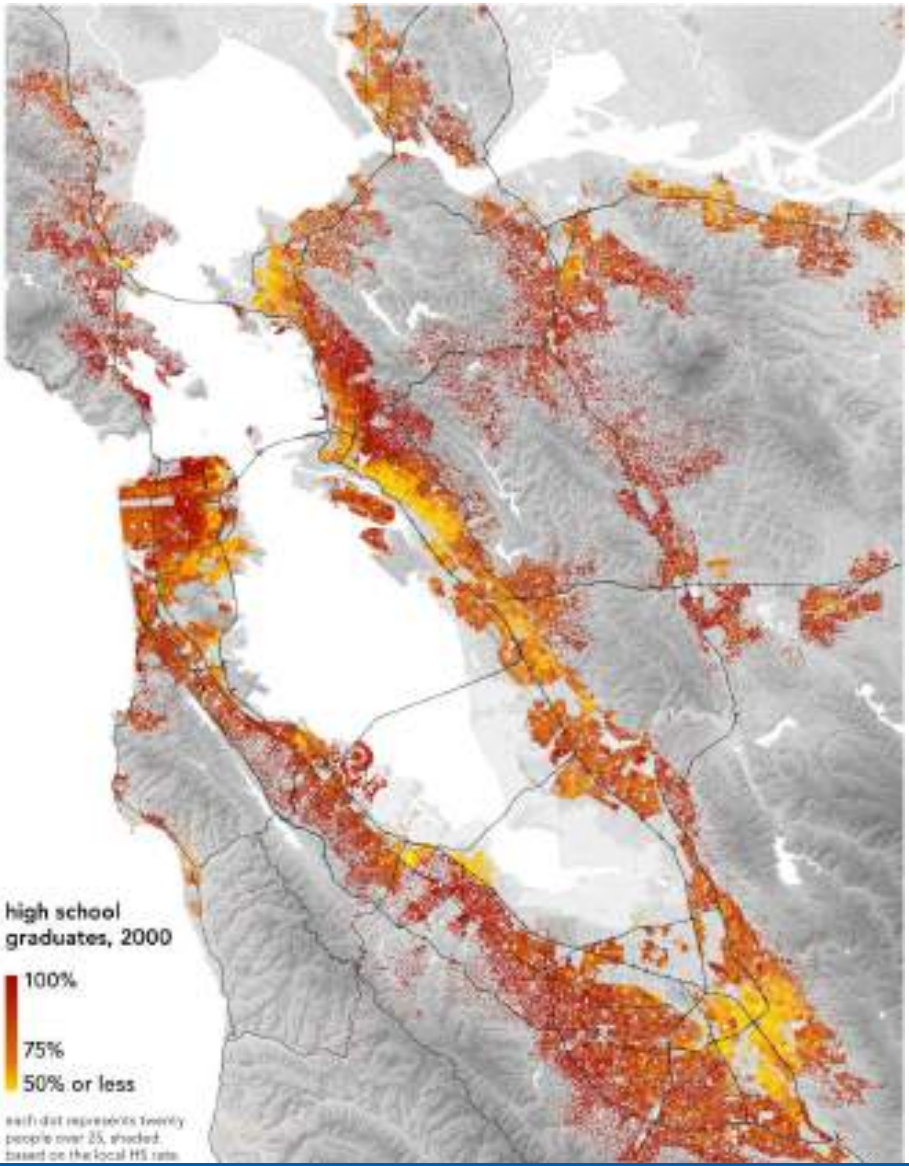
Seattle



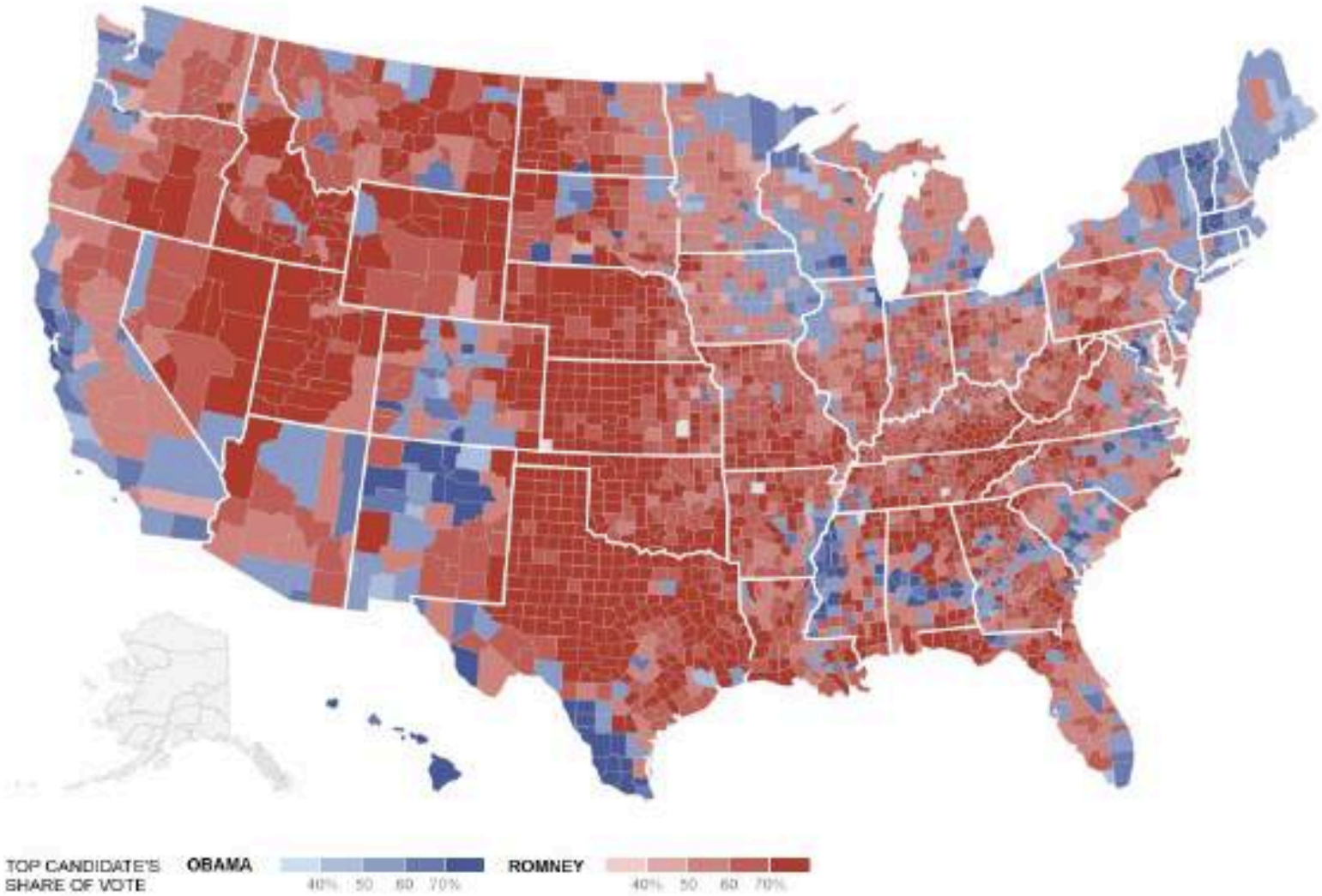
Los Angeles

Miami

Bay area high school graduates



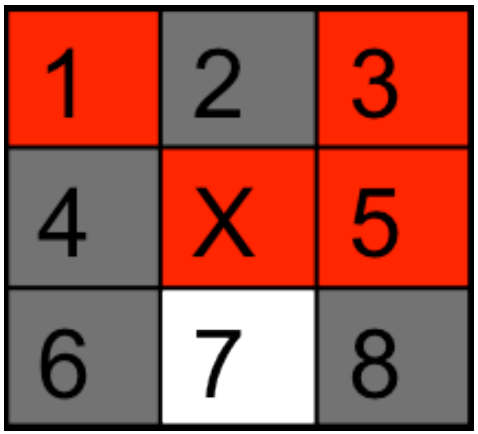
2012 US Presidential Elections Map



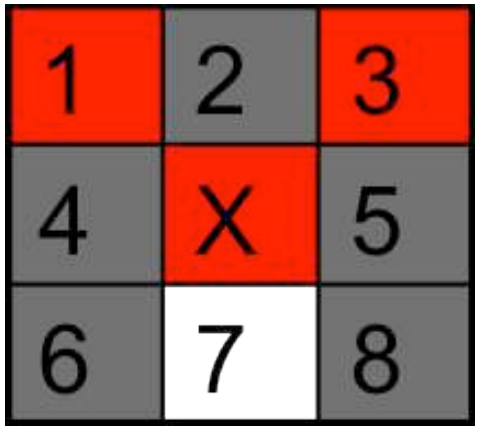
Schelling's model of segregation

- Population consists of 2 types of agents
- Agent reside in the cells of the grid (2-dimensional geography of a city), 8 neighbors
- Some cells contain agents, some unpopulated
- Every agent wants to have at least some fraction of agents (threshold) of his type as neighbor (satisfied agent)
- On every round every unsatisfied agent moves to a satisfactory empty cell.
- Continues until everyone is satisfied or can't move

Spatial segregation



satisfied agent



unsatisfied agent

- preference threshold $\lambda = 3/7$

Spatial segregation

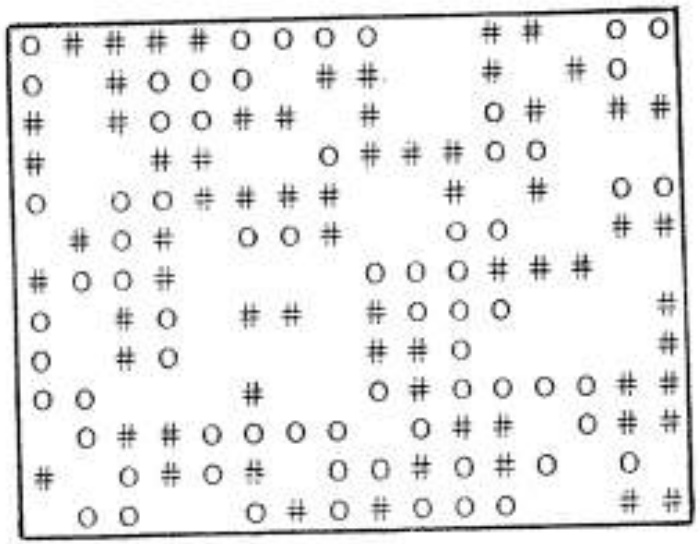


Fig. 7

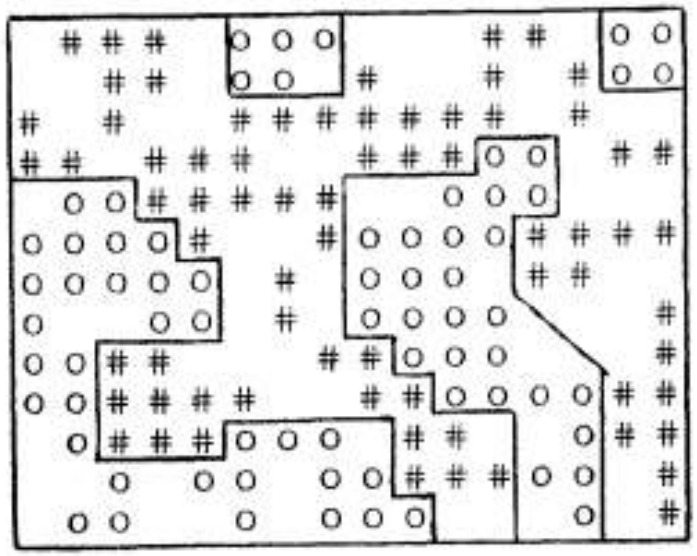
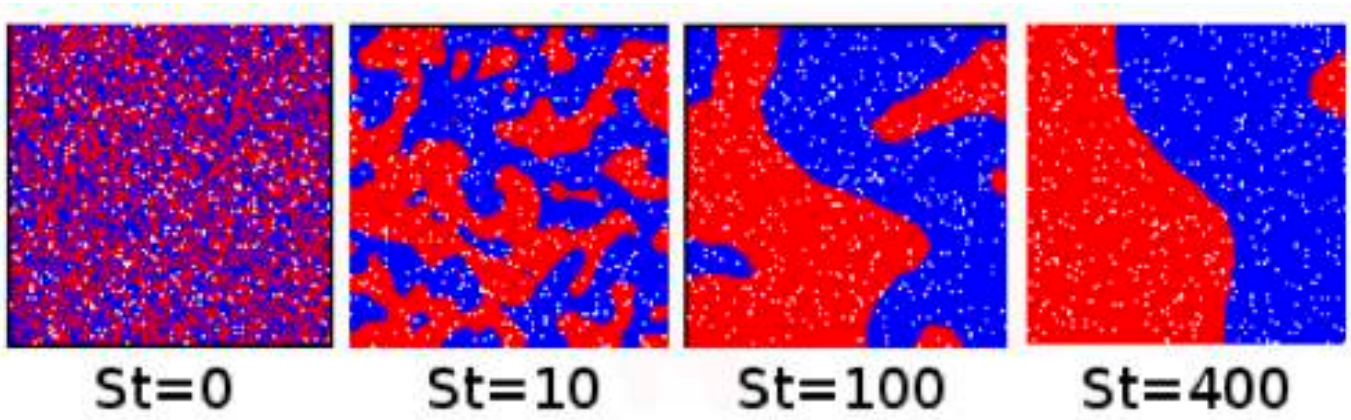


Fig. 10

T. Schelling, 1971

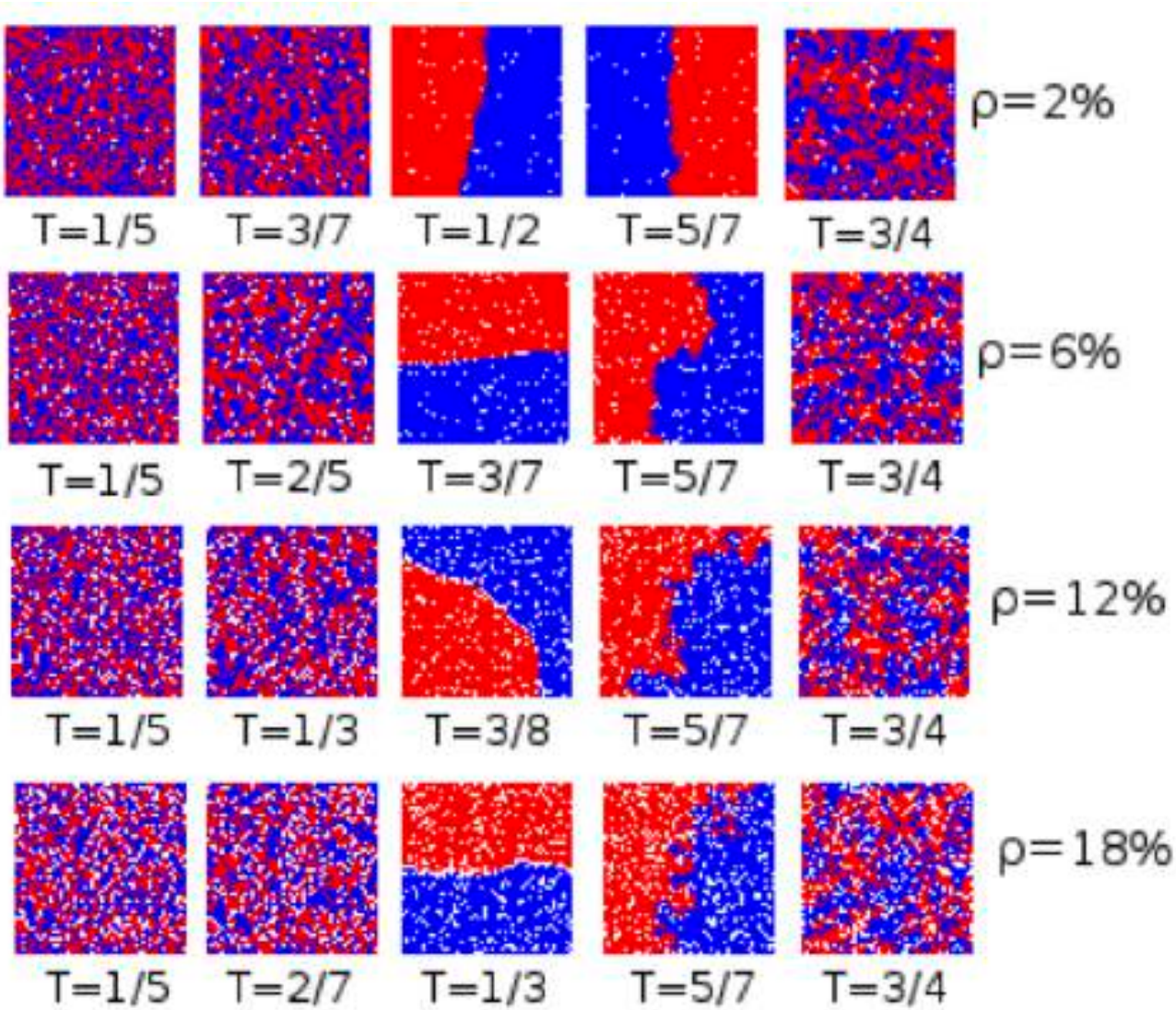
Spatial segregation

vacancy 5%, tolerance $\lambda = 0.5$



L. Gauvin et.al. 2009

Spatial segregation



Model

- N - nodes, θ - fraction of occupied by A and B

$$n_A + n_B = \theta \cdot N$$

- Proportion of "unlike" nearest neighbors, $k_i = \#NN$

$$P_i = \begin{cases} \#n_B/k_i, \text{ if } i \in A \\ \#n_A/k_i, \text{ if } i \in B \end{cases}$$

- Utility function, λ - sensitivity (tolerance threshold) level

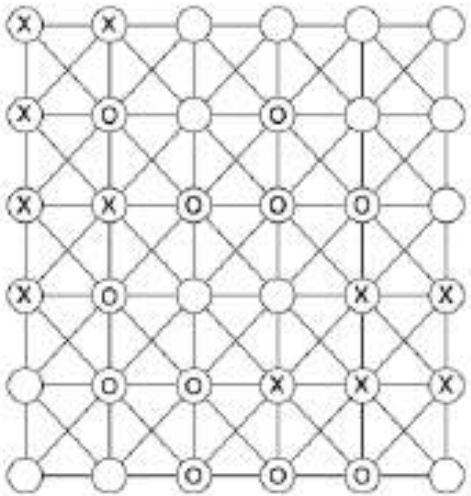
$$u_i = \begin{cases} 1, \text{ if } P_i \leq \lambda \\ 0, \text{ if } P_i > \lambda \end{cases}$$

- Every node moves to maximize its utility

Spatial segregation

X	X				
X	O		O		
X	X	O	O	O	
X	O			X	X
	O	O	X	X	X
		O	O	O	

(a)



(b)

Algorithm

- time steps $1..T$
- At every time step randomly select an agent, compute utility
- If utility is $u = 0$ move to an empty location to maximize utility
- Movements: 1) random location 2) nearest available location
- Repeat until either all utilities are maximized $\sum_i u_i = \theta N$ or reaches "frozen" state, no place to move, then $\sum_i u_i < \theta N$
- Total utility of society

$$U = \sum_i u_i$$

Measuring segregation

- Schilling's solid mixing index

$$M = \frac{1}{n_A + n_B} \sum_i P_i$$

- Freeman's segregation index

$$F = 1 - \frac{e^*}{E(e^*)}$$

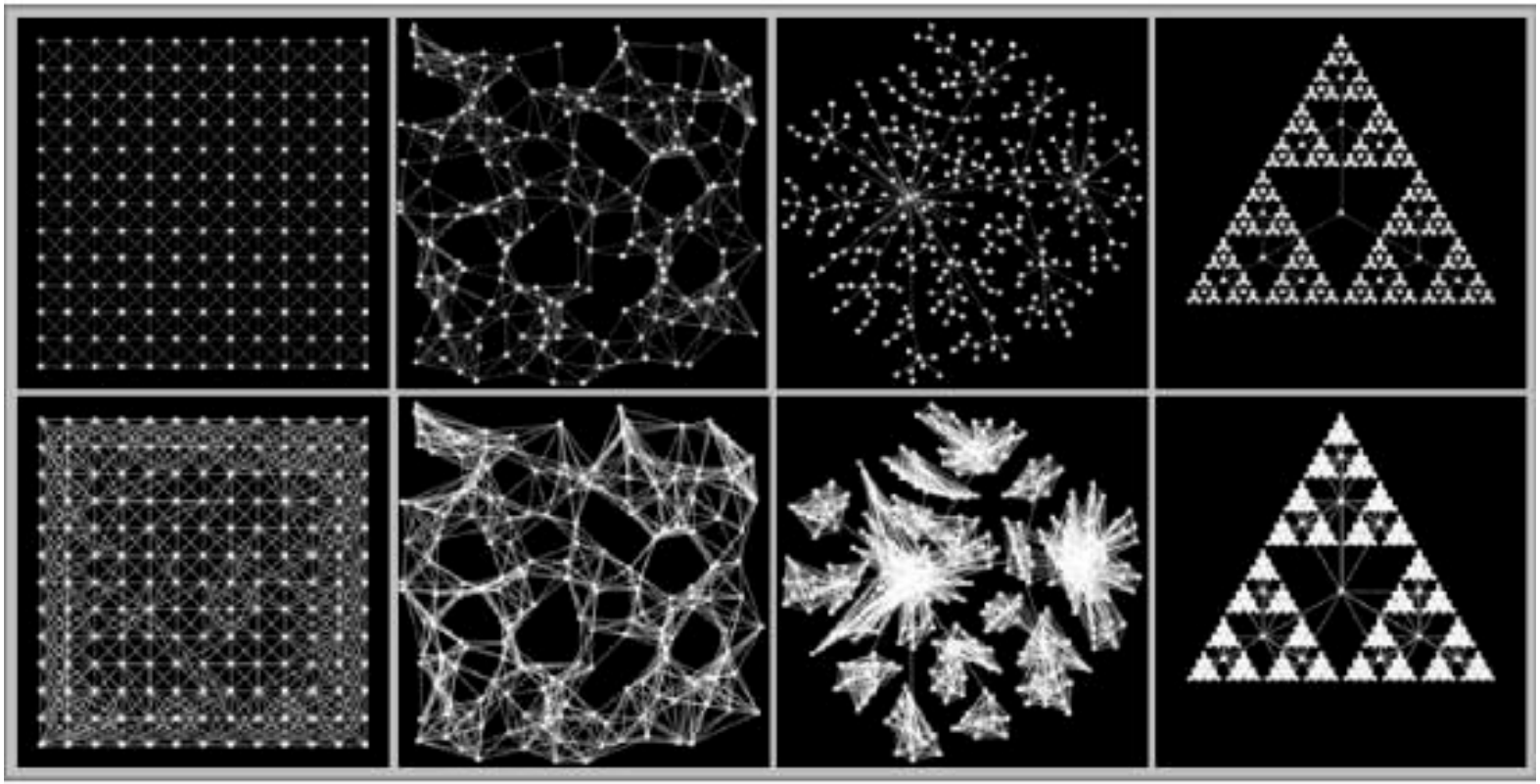
$e^* = \frac{e_{AB}}{(e_{AB}+e_{AA}+e_{BB})}$ - observed proportion of between group ties,
$E(e^*) = \frac{2n_A n_B}{(n_A+n_B)(n_A+n_B-1)}$ - expected proportion for random ties

- Assortative mixing

$$Q = \frac{1}{2m} \sum_{ij} (A_{ij} - \frac{k_i k_j}{2m}) \delta(c_i, c_j)$$

Spatial segregation on networks

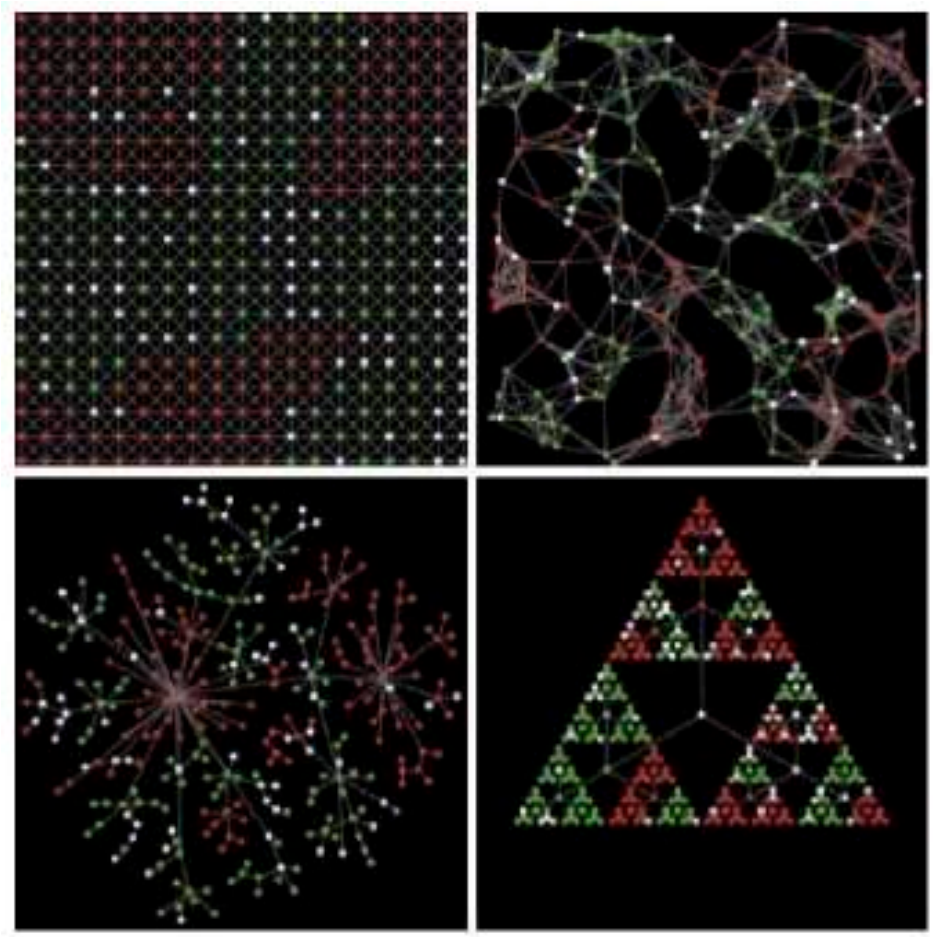
Fixed degree $k = 10$ neighboring graphs: regular, random, scale-free, fractal



Arnaud Banos, 2010

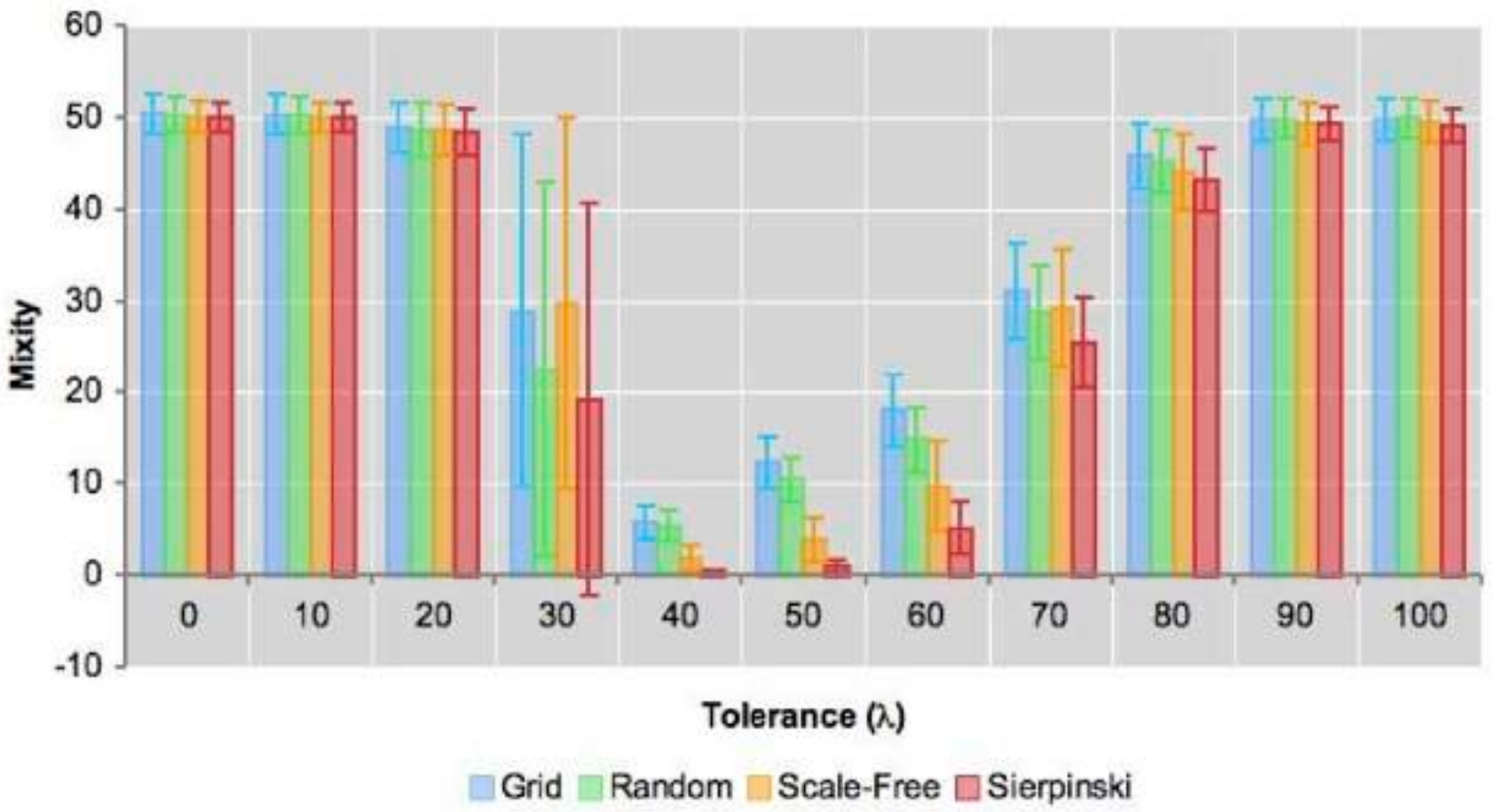
Spatial segregation on networks

$\lambda = 0.5, \theta = 0.8$



Banos, 2010

Spatial segregation on networks



Banos, 2010

Summary

- Spatial segregation is taking place even though no individual agent is actively seeking it (minor preferences, high tolerance)
- Network structure does affect segregation
- Fixed characteristics (race) can become correlated with mutable (location)

References

- Dynamic Models of Segregation, Thomas C. Schelling, 1971
- Segregation in Social Networks, Linton Freeman, 1978
- Gauvin L, Vannimenus J, Nadal JP. Phase diagram of a Schelling segregation model. The European Physical Journal B, 70:293-304, 2009
- Arnaud Banos. Network effects in Schelling's model of segregation: new evidences from agent-based simulations. 2010

Course summary

1. Introduction to network science
2. Descriptive network analysis
3. Mathematical models of networks
4. Node centrality and ranking on networks
5. Network communities
6. Network structure and visualization
7. Epidemics and information spread
8. Diffusion of innovation
9. Spatial model of segregation