Introduction to Network Science

Yury Dvorkin

New York University

January 29, 2020

- Instructor: Yury Dvorkin
- Grader: Guanhong Liu
- Course length: 2 modules (Network Science and Network Optimization)

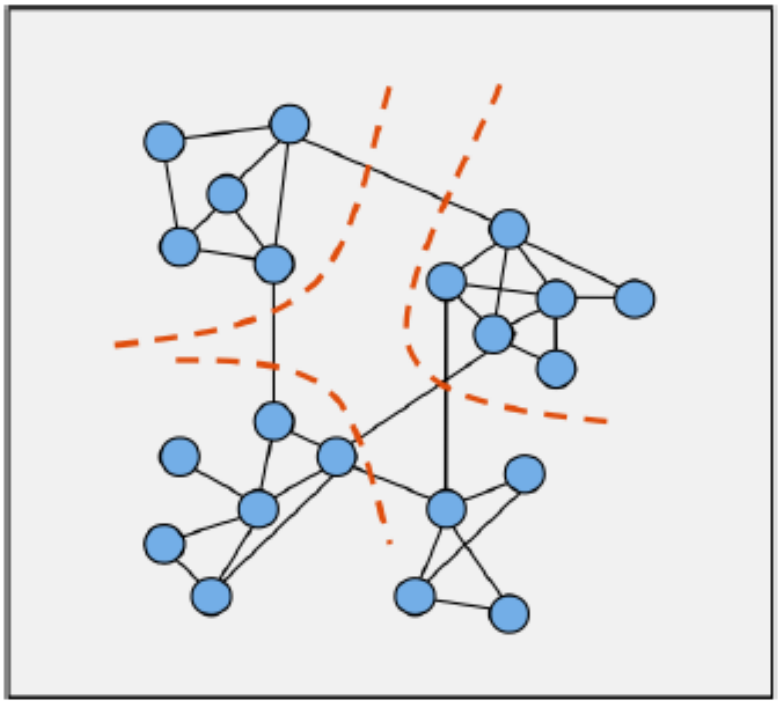
- Statistical mechanics of complex networks, R. Albert and A-L. Barabasi, Rev. Mod. Phys. 74, 47, 2002
- The Structure and Function of Complex Networks, M. E. J. Newman, SIAM Review, 45, pp 167-256, 2003
- Evolution of Networks, S. N. Dorogovtsev and J. F. F. Mendes, Adv.Phys., pp 1079-1187, 2002
- Complex networks: Structure and dynamics, S. Boccaletti et al., Physics Reports, Volume 424, Issue 4-5, p. 175-308, 2006

- 1. Statistical properties of networks
- 2. Network structure
- 3. Processes on networks
- 4. Predictions on networks (ML)

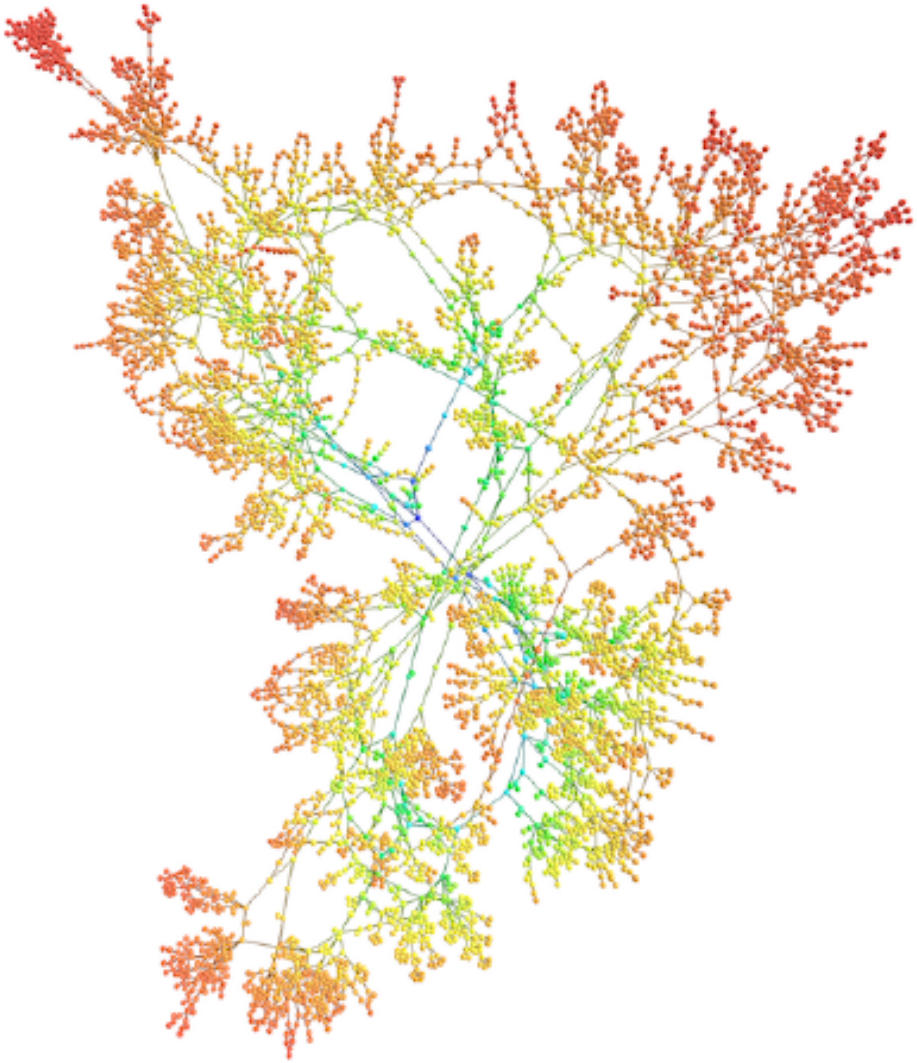
- Introduction to network science
- Power laws
- Random graphs
- Small world and dynamical growth models
- Centrality measures
- Link analysis
- Structural equivalence
- Network communities
- Graph partitioning algorithms
- Community detection

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

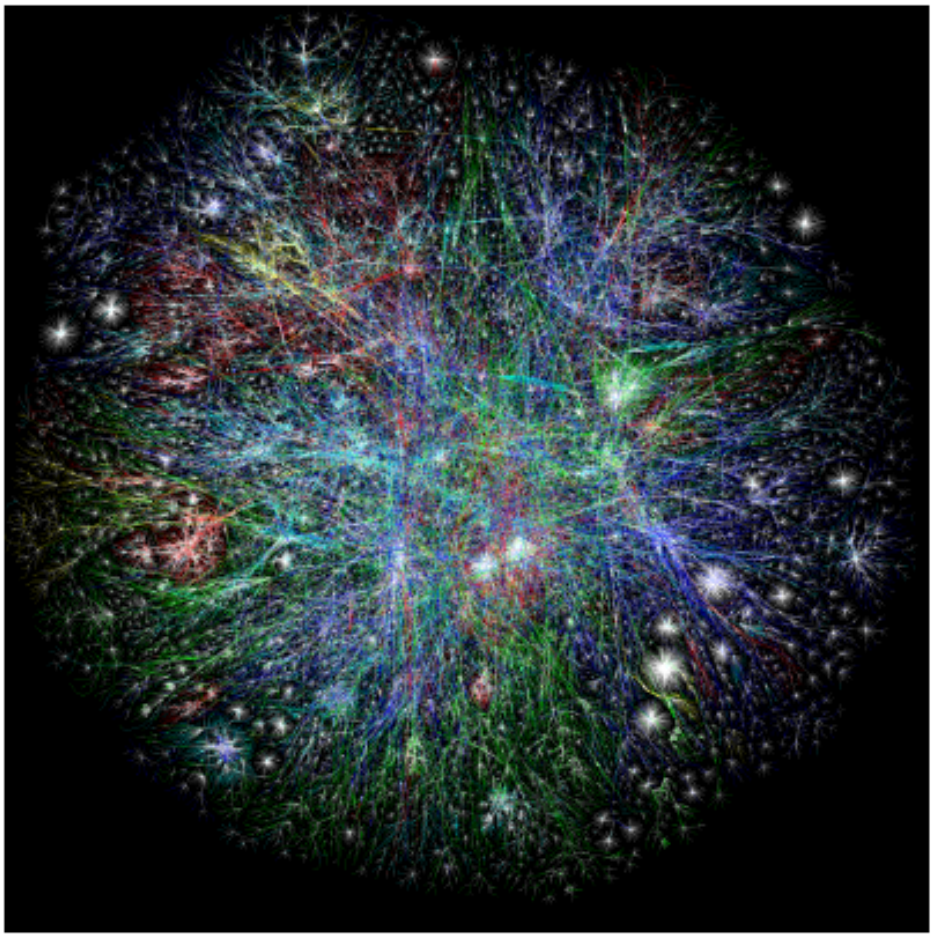
- network = graph
- nodes = vertices, actors
- links = edges, relations
- clusters = communities



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



- Internet traffic routing (BGP)



Barret Lyon, 2003

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

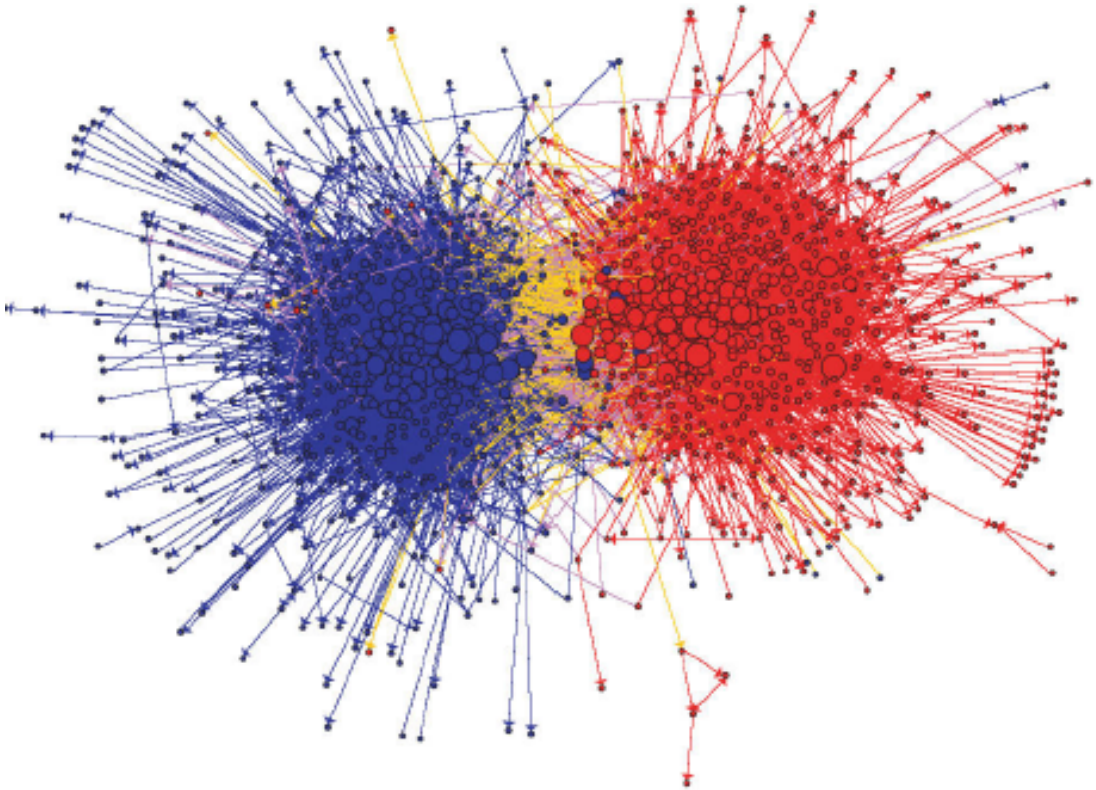
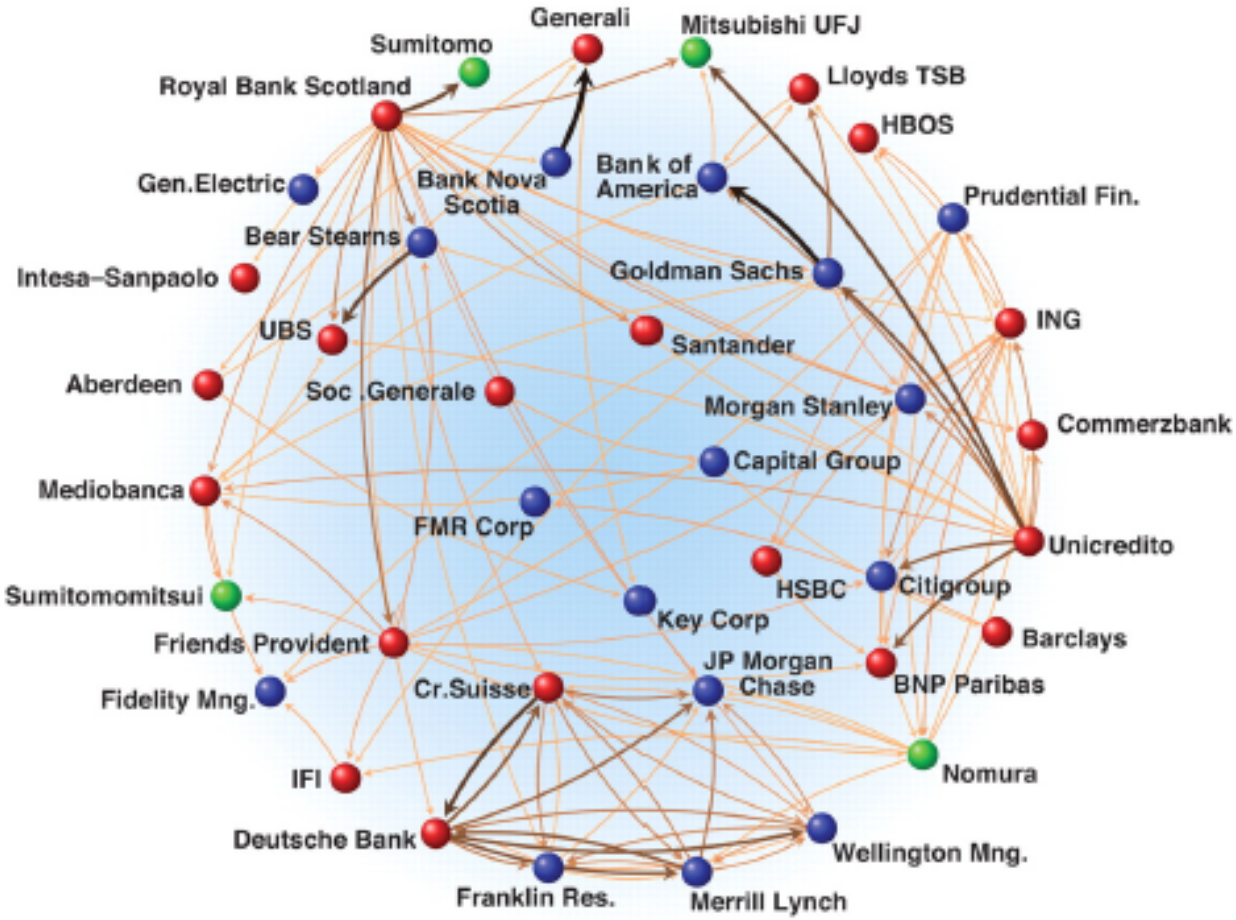


image from L. Adamic, N. Glance, 2005

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

image from K. McKelvey et.al., 2012

- existing relations between financial institutions



F. Schweitzer, 2009

- Zurich public transportation map

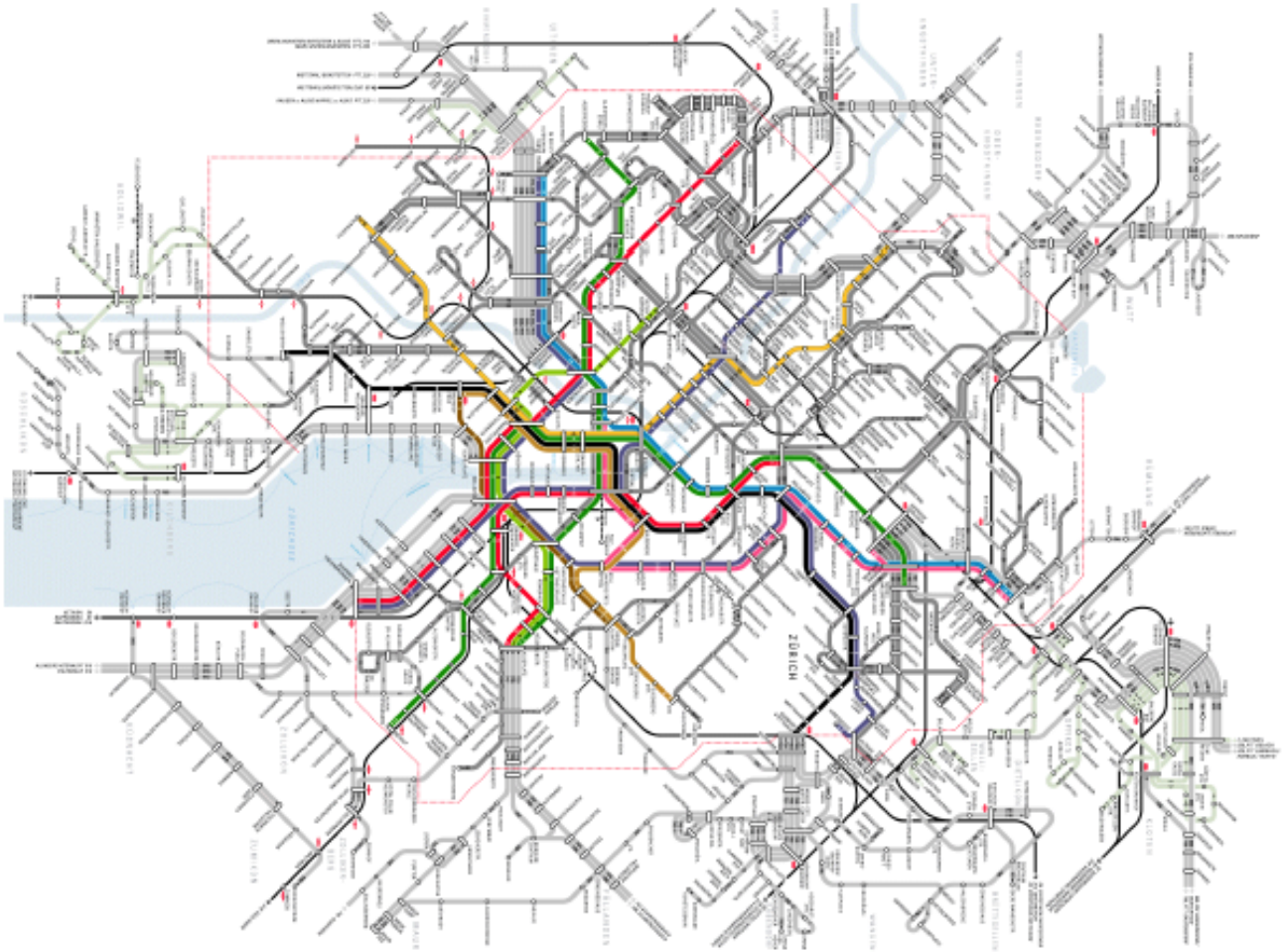
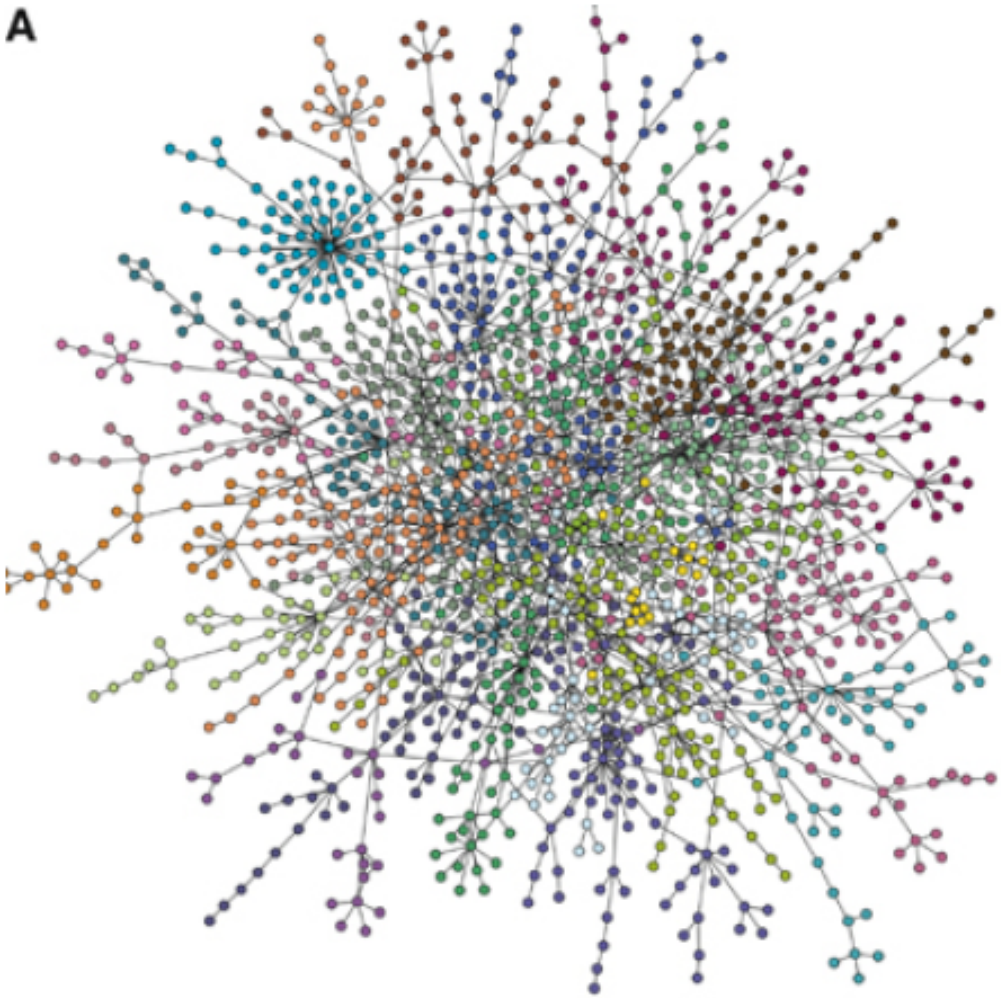
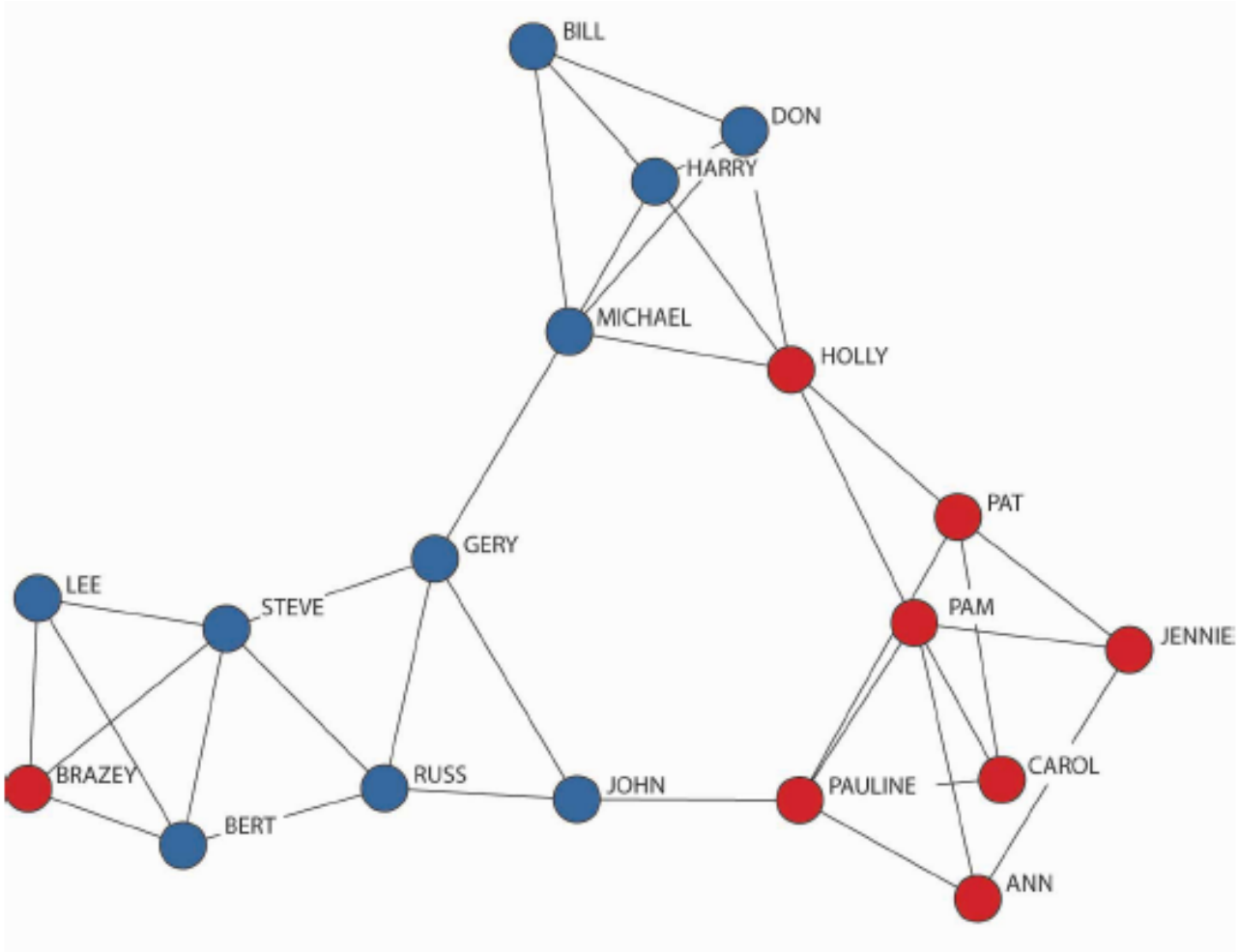


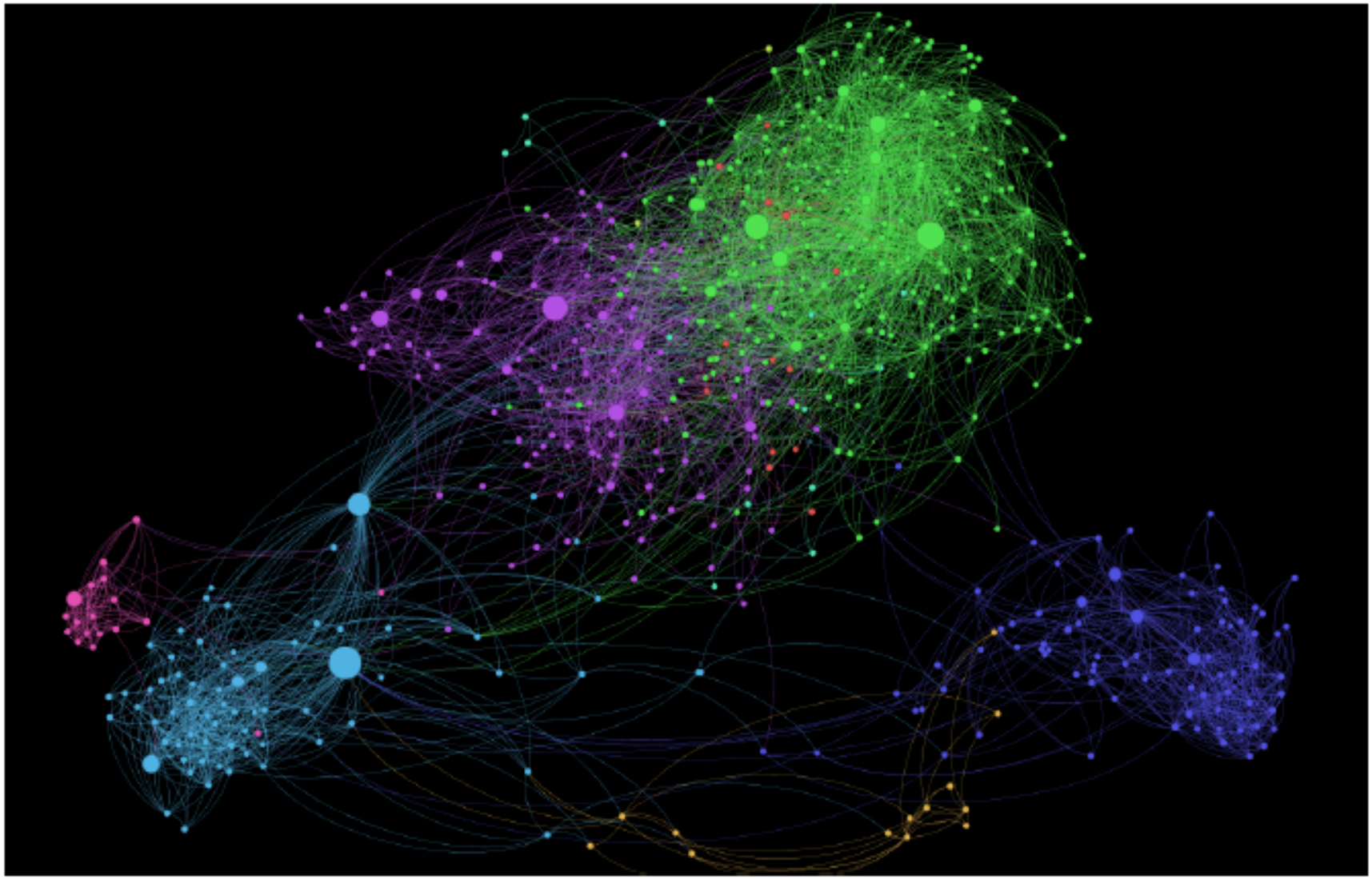
image from http://www.visualcomplexity.com

- Yeast protein interaction network



H.Jeong et.al., 2001


BILL
DON
HARRY
MICHAEL
HOLLY
PAT
GERY
LEE
STEVE
PAM
JENNIE
BRAZEY
RUSS
JOHN
PAULINE
CAROL
BERT
ANN



- Friendship graph 500 mln people

image by Paul Butler, 2010

- Power law node degree distribution: "scale-free" networks
- Small diameter and average path length: "small world" networks
- Hight clustering coefficient: transitivity

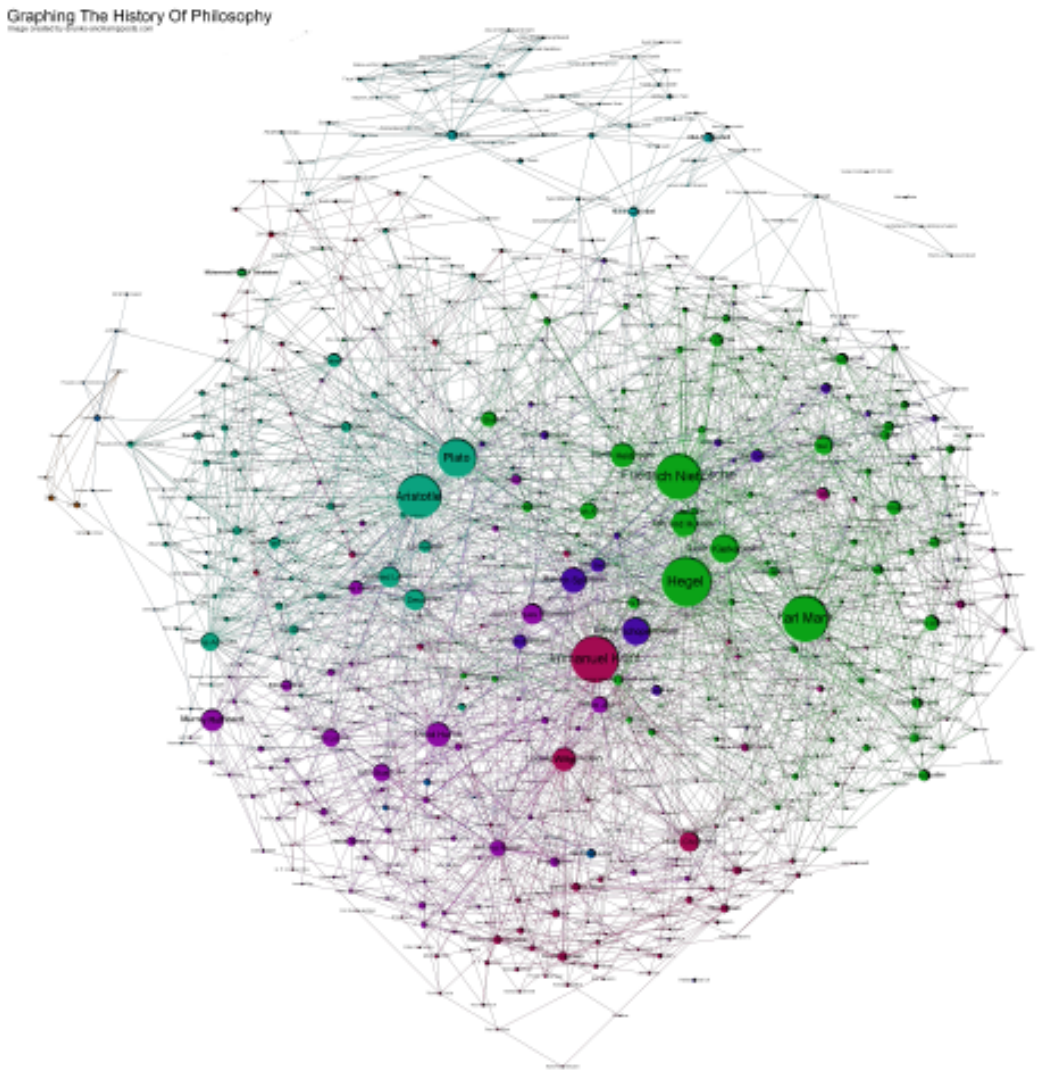


image from http://www.coppelia.io

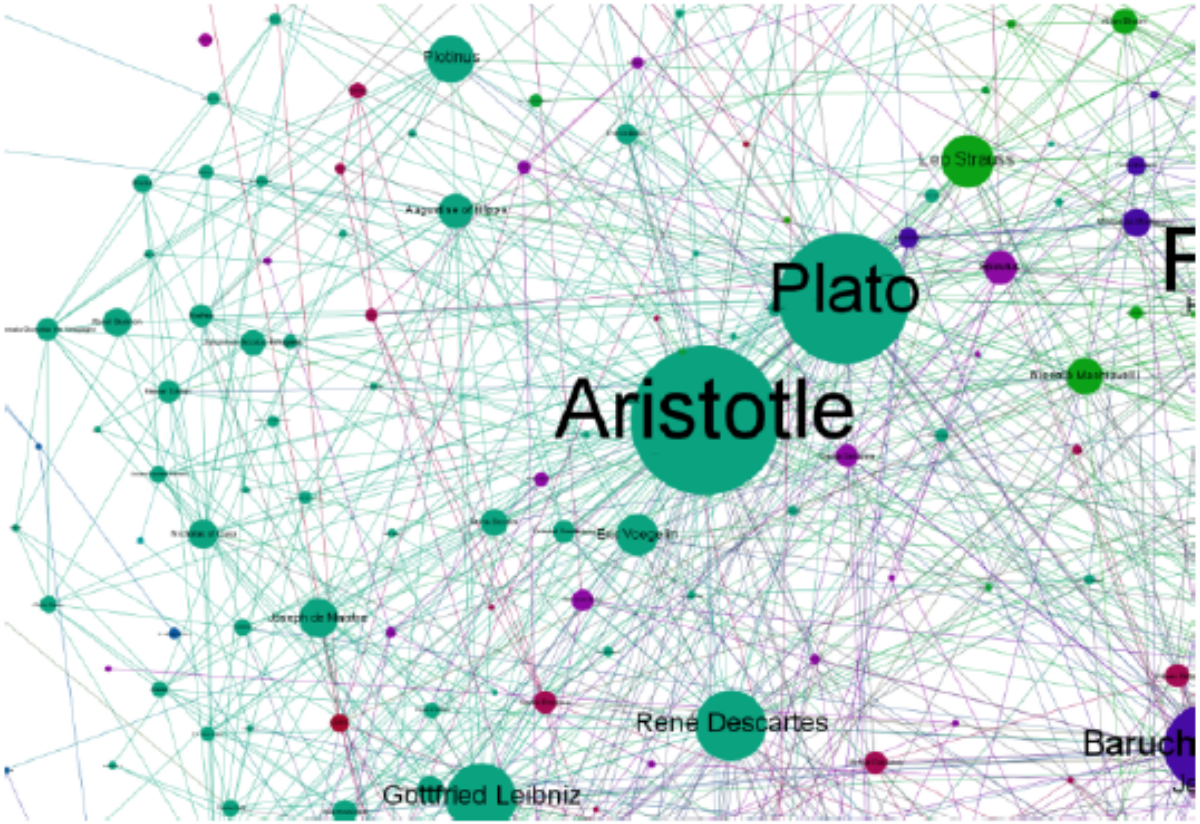


image from http://www.coppelia.io

- Node degree distribution in scale-free network:

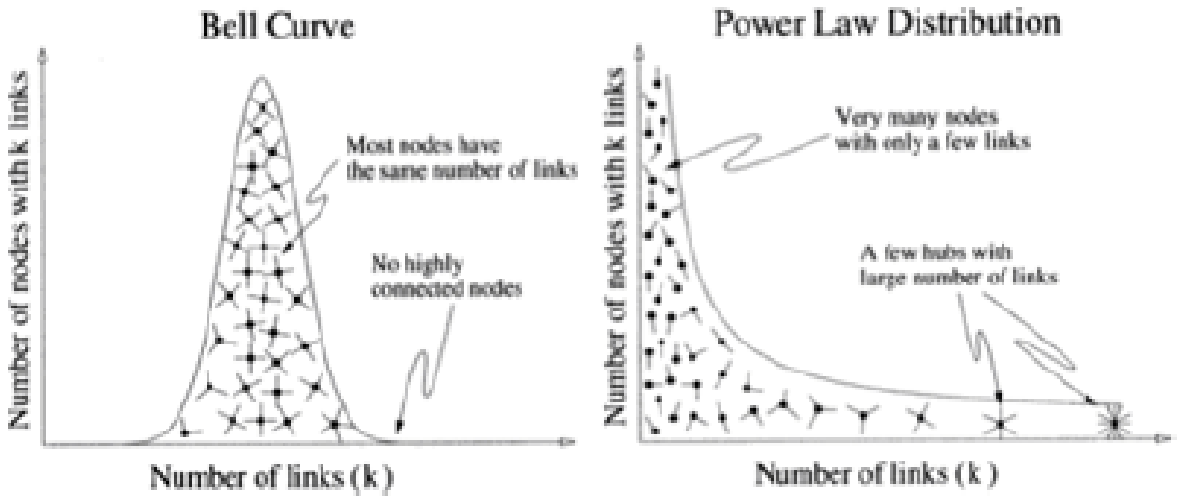
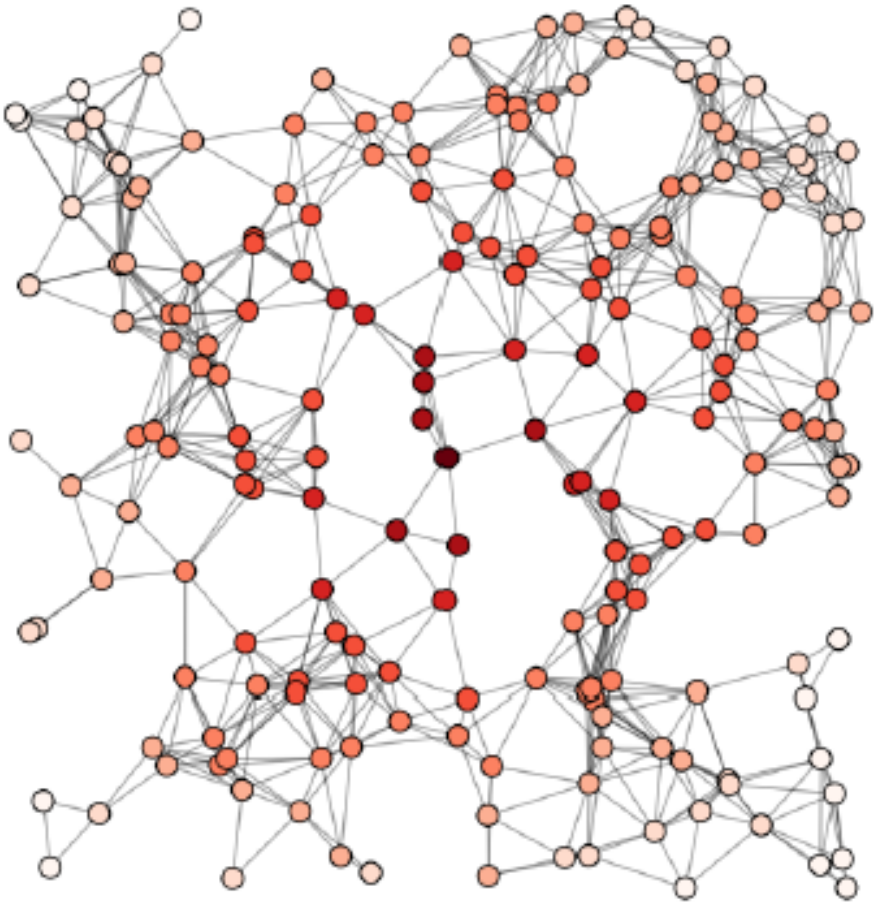
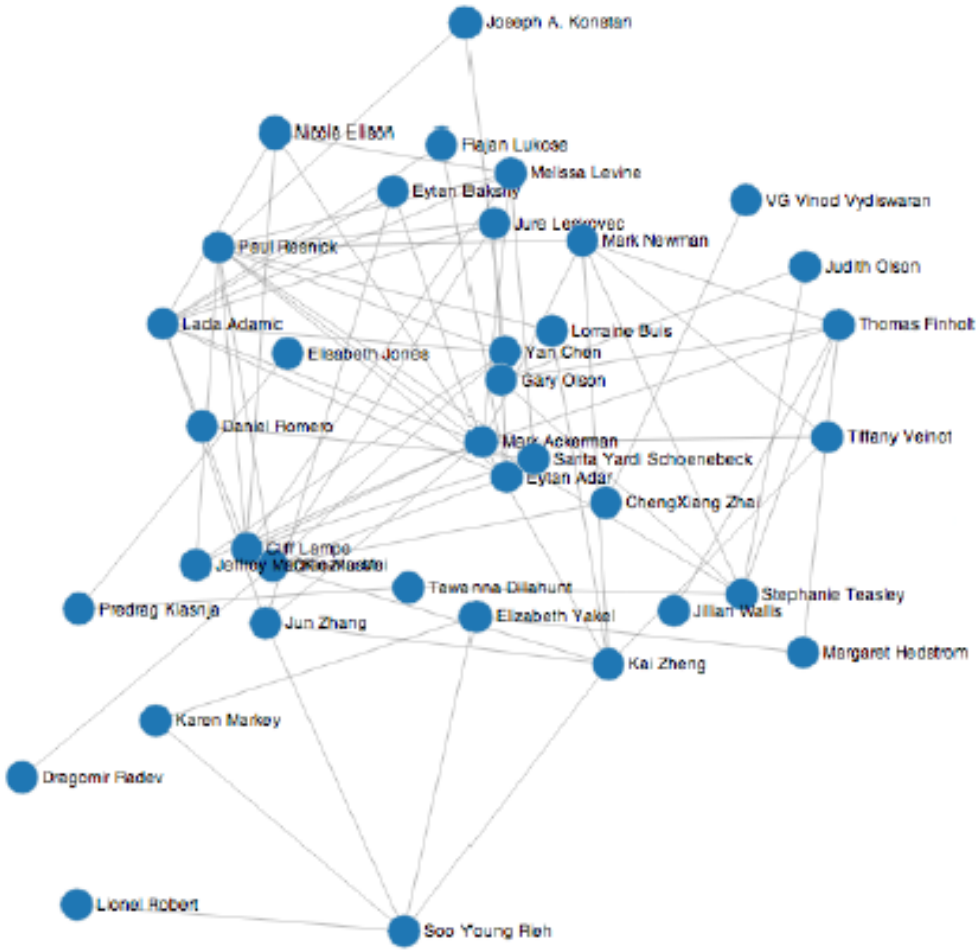


image from A.-L. Barabasi, 2002

- Random geometric graph

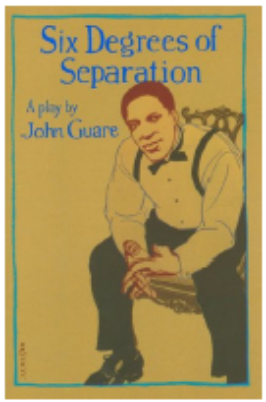


- Co-author network



"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"




© Al Satterwhite

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

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.

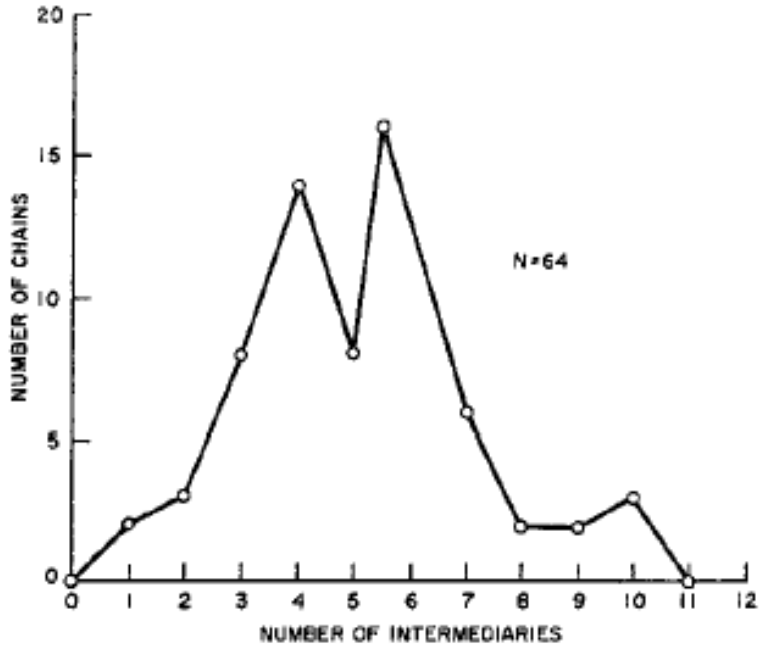
3. IF YOU KNOW THE TARGET PERSON ON A PERSONAL BASIS, MAIL THIS FOLDER DIRECTLY TO HIM (HER). Do this only if you have previously met the target person and know each other on a first name basis.

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

- 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

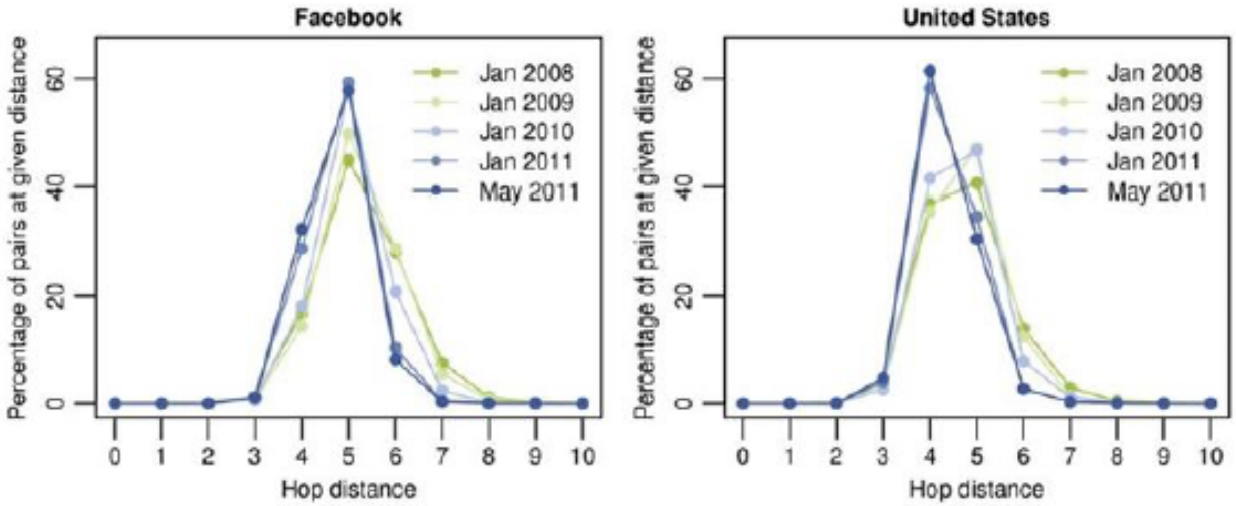
J. Travers, S. Milgram, 1969

- 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 $\langle L \rangle = 5.7$

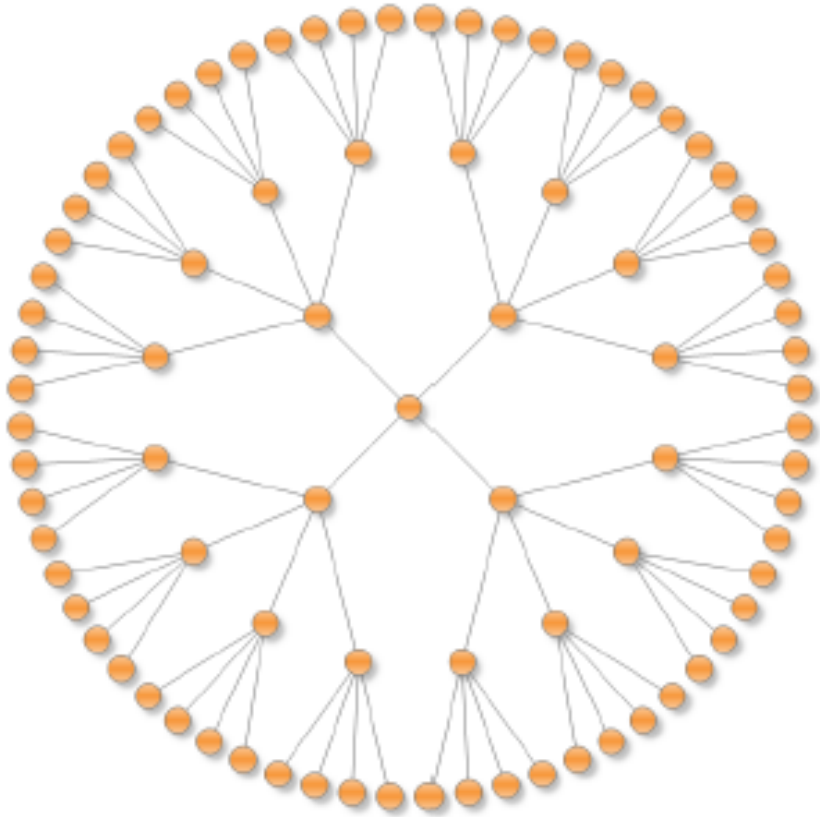


J. Travers, S. Milgram, 1969

- 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, $\langle L \rangle \approx 4.74$

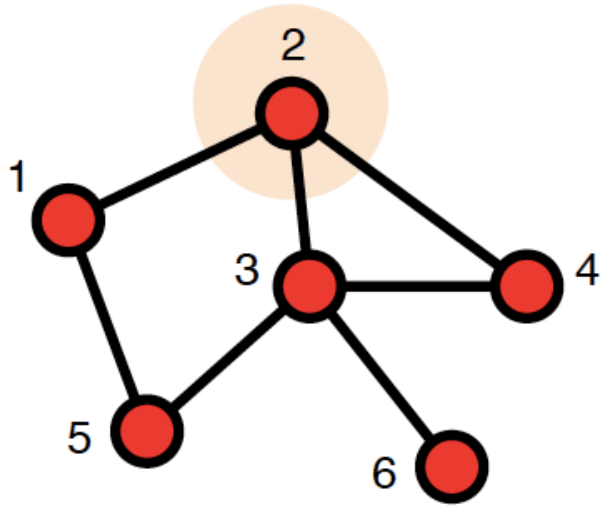


figures from L.Backstrom, 2012



An estimate: $z^d = N, d = \log N / \log z$
$N \approx 6.7\ bln,\ z = 50\ friends,\ d \approx 5.8.$

- 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. 3338.



undirected

unweighted

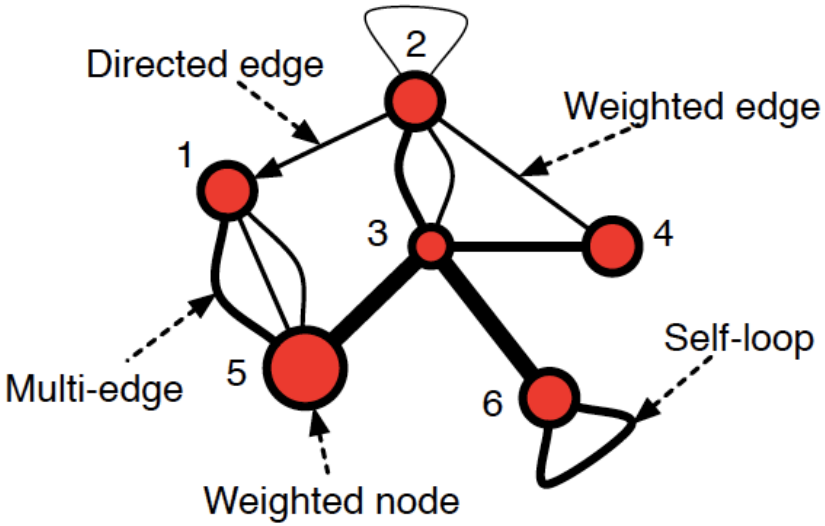
no self-loops

adjacency matrix

A	1	2	3	4	5	6
1	0	1	0	0	1	0
2	1	0	1	1	0	0
3	0	1	0	1	1	1
4	0	1	1	0	0	0
5	1	0	1	0	0	0
6	0	0	1	0	0	0

adjacency list

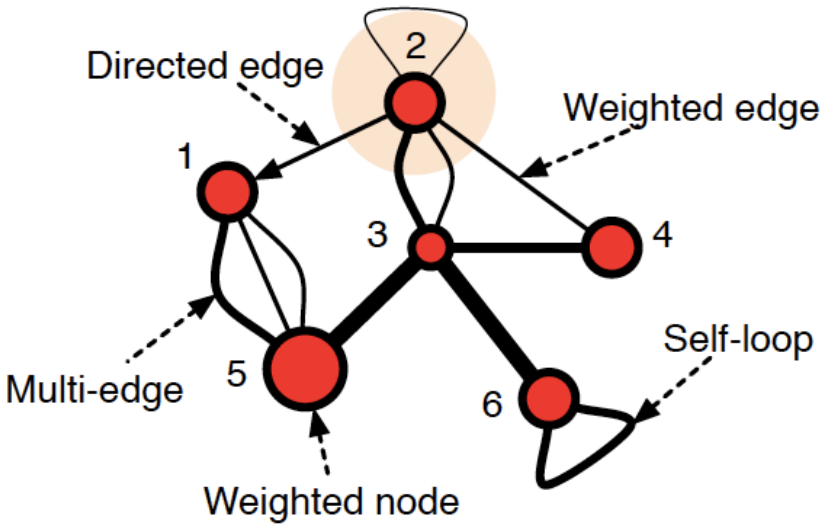
A
$1 \rightarrow \{2, 5\}$
$2 \rightarrow \{1, 3, 4\}$
$3 \rightarrow \{2, 4, 5, 6\}$
$4 \rightarrow \{2, 3\}$
$5 \rightarrow \{1, 3\}$
$6 \rightarrow \{3\}$



~~undirected~~

~~unweighted~~

~~no self-loops~~



adjacency matrix

A	1	2	3	4	5	6
1	0	0	0	0	$\{1,1,2\}$	0
2	1	$\frac{1}{2}$	$\{2,1\}$	1	0	0
3	0	$\{2,1\}$	0	2	4	4
4	0	1	2	0	0	0
5	$\{1,1,2\}$	0	4	0	0	0
6	0	0	4	0	0	2

adjacency list

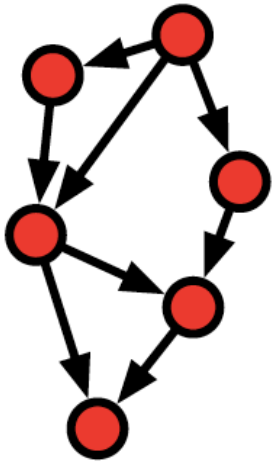
A	
1	$\rightarrow \{(5,1),(5,1),(5,2)\}$
2	$\rightarrow \{(1,1),(2,\frac{1}{2}),(3,2),(3,1),(4,1)\}$
3	$\rightarrow \{(2,2),(2,1),(4,2),(5,4),(6,4)\}$
4	$\rightarrow \{(2,1),(3,2)\}$
5	$\rightarrow \{(1,1),(1,1),(1,2),(3,4)\}$
6	$\rightarrow \{(3,4),(6,2)\}$

citation networks

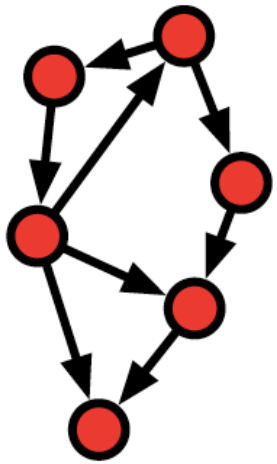
foodwebs*

epidemiological

others?



directed acyclic graph



directed graph

WWW

friendship?

flows of goods, information

economic exchange

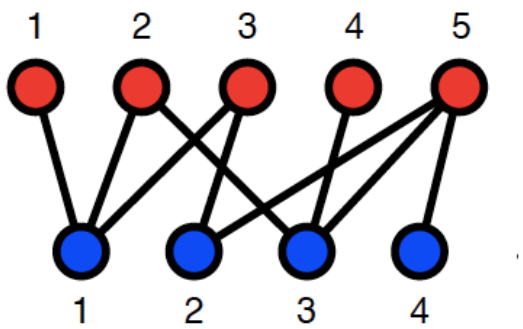
dominance

neuronal

transcription

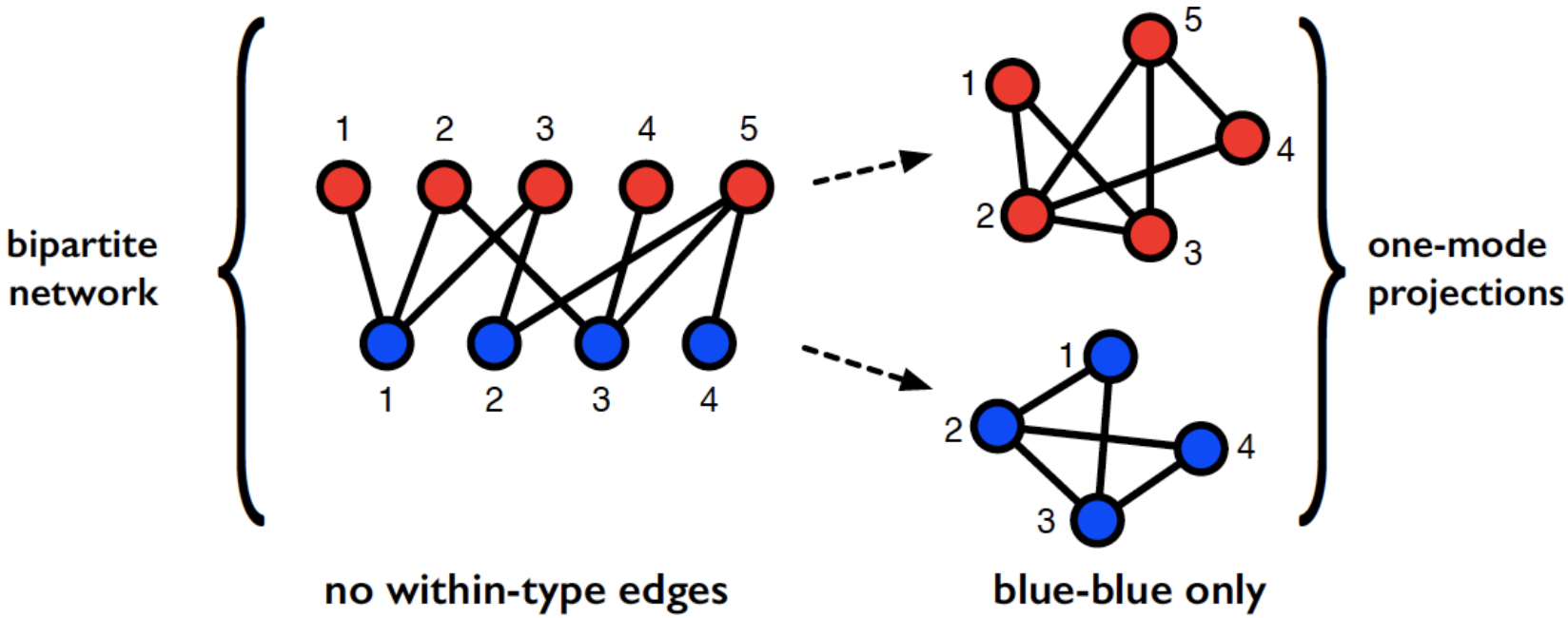
time travelers

bipartite network

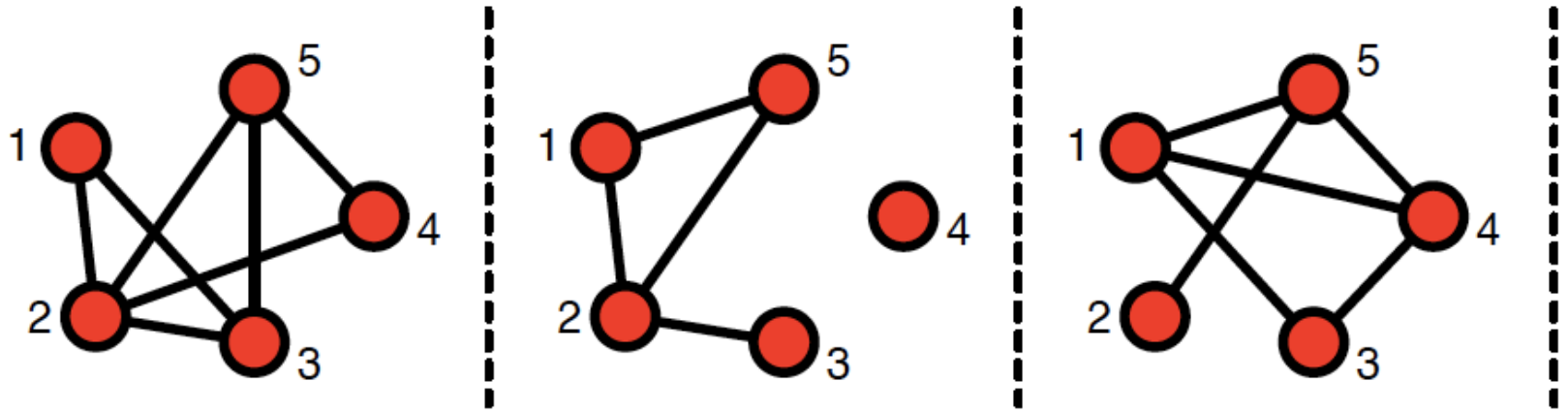


no within-type edges

authors & papers	people & locations (checkins)
actors & movies/scenes	metabolites & reactions
musicians & albums	genes & substrings
people & online groups	words & documents
people & corporate boards	plants & pollinators



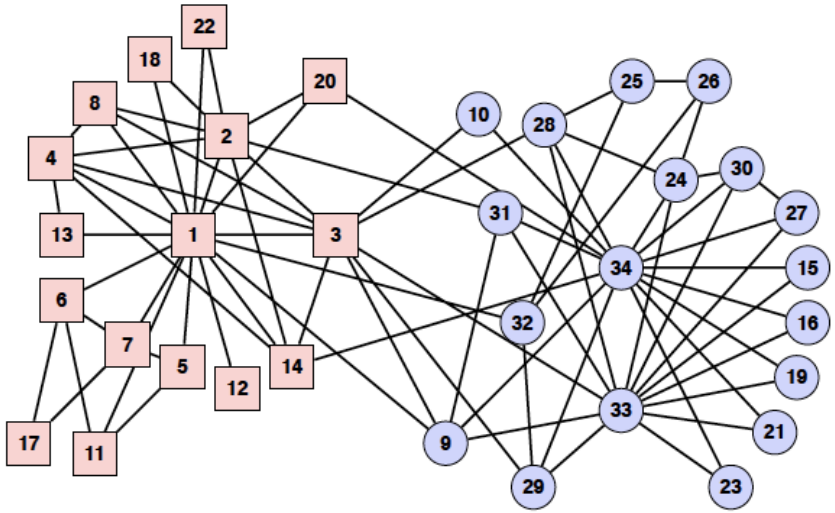
authors & papers	people & locations (checkins)
actors & movies/scenes	metabolites & reactions
musicians & albums	genes & substrings
people & online groups	words & documents
people & corporate boards	plants & pollinators



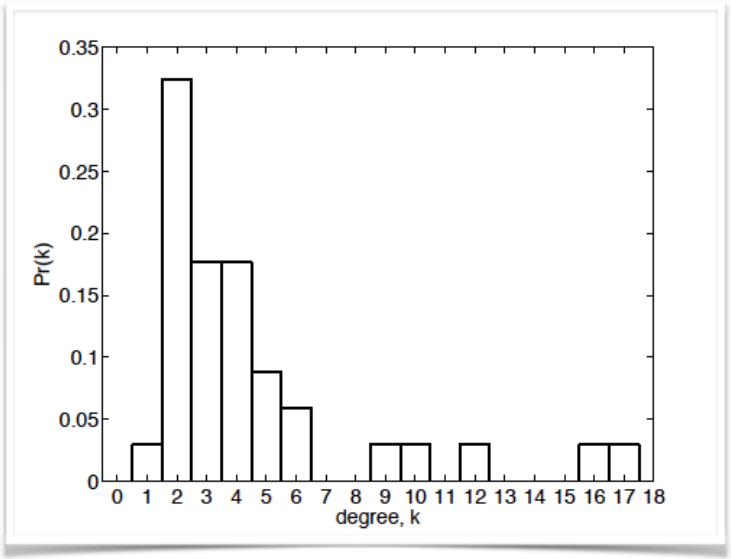
any network over time

discrete time (snapshots), edges (i, j, t)

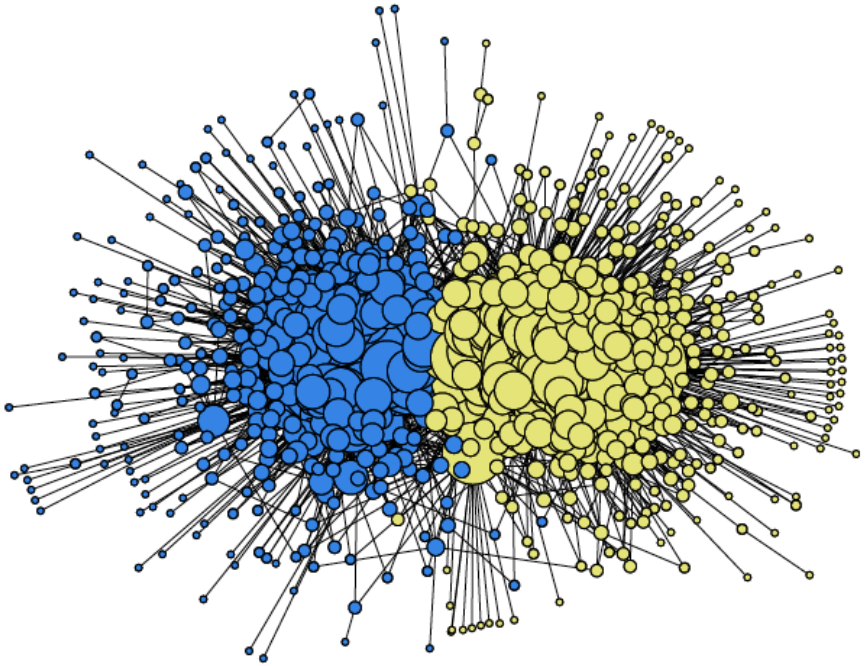
continuous time, edges $(i, j, t_s, \Delta t)$



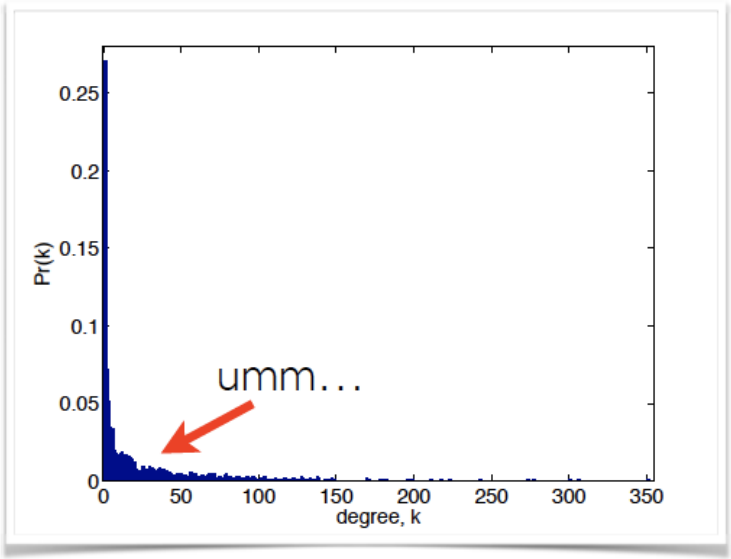
Zachary karate club



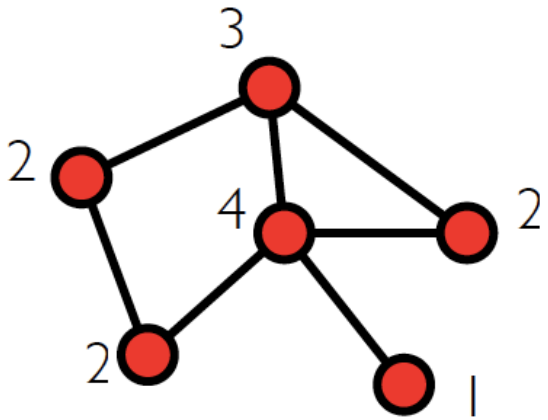
Zachary, *J. Anthropological Research* **33**, 452–473 (1977).



political blogs*



Adamic and Glance, *WWW Workshop on the Weblogging Ecosystem* (2005)
Karrer and Newman, *Physical Review E* **83**, 016107 (2010)



degree:
number of connections k

$$k_i = \sum_j A_{ij}$$

when does node degree matter?

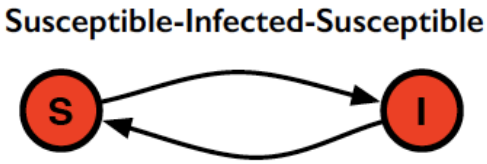
spreading processes on networks

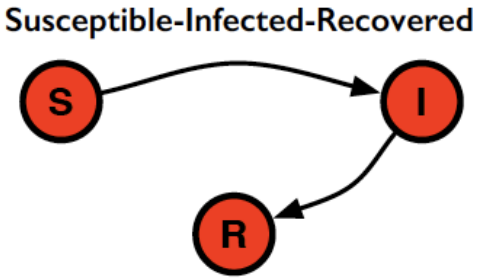
biological (diseases)

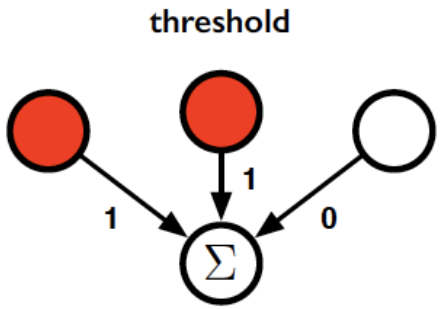
- SIS and SIR models

social (information)

- SIS, SIR models
- threshold models



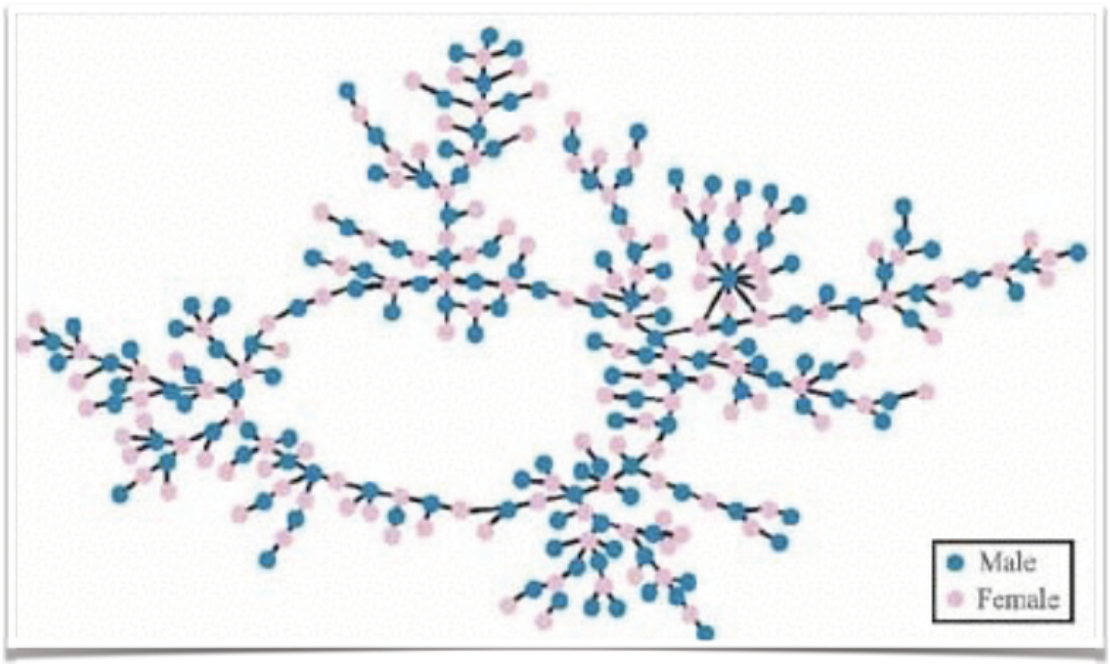




Chains of Affection: The Structure of Adolescent Romantic and Sexual Networks 2004

Peter S. Bearman, *Columbia University*
James Moody, *Ohio State University*
Katherine Stovel, *University of Washington*

- relationship network in "Jefferson High"
- this subgraph is 52% of school
- who are most important disease spreaders?



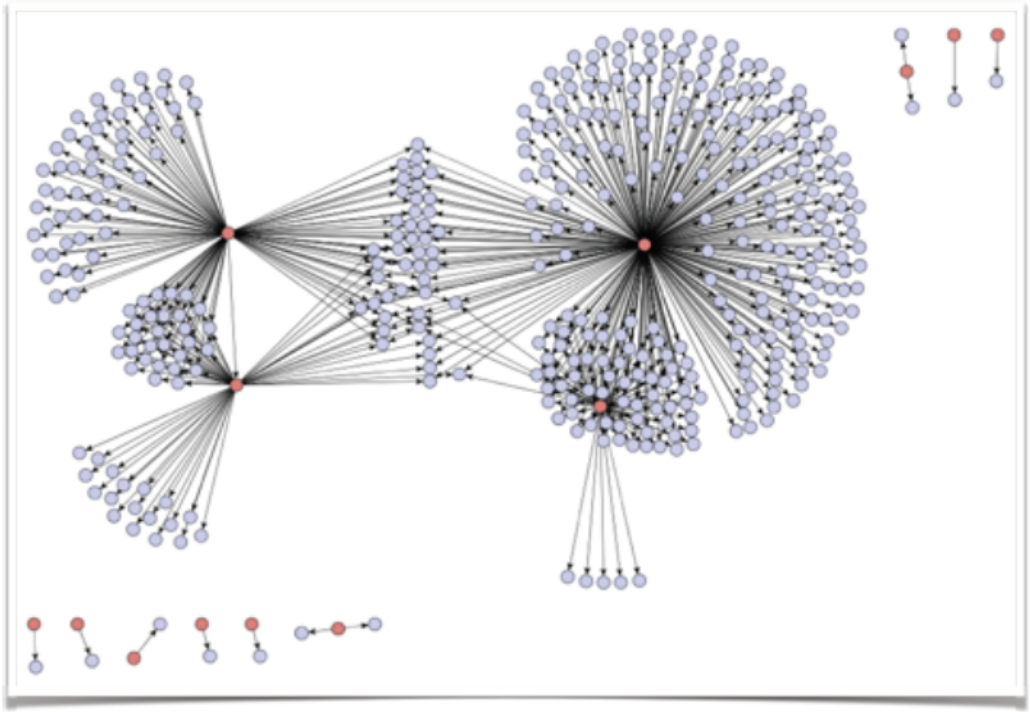
Bearman, et al., *Amer. J. Sociology* **110**, 44–91 (2004)

The Dynamics of Viral Marketing

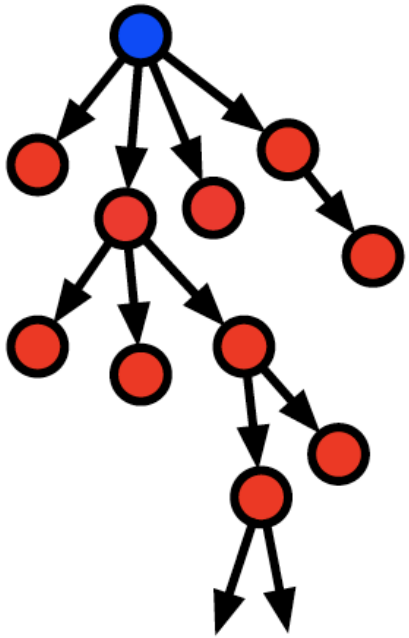
2007

JURE LESKOVEC LADA A. ADAMIC BERNARDO A. HUBERMAN

- amazon.com viral marketing
- viral trace for "*Oh my Goddess!*" community
- very high degrees!
- most attempts to "influence" fail



Leskovec et al., *ACM Trans. on the Web* 1, article 5 (2007)

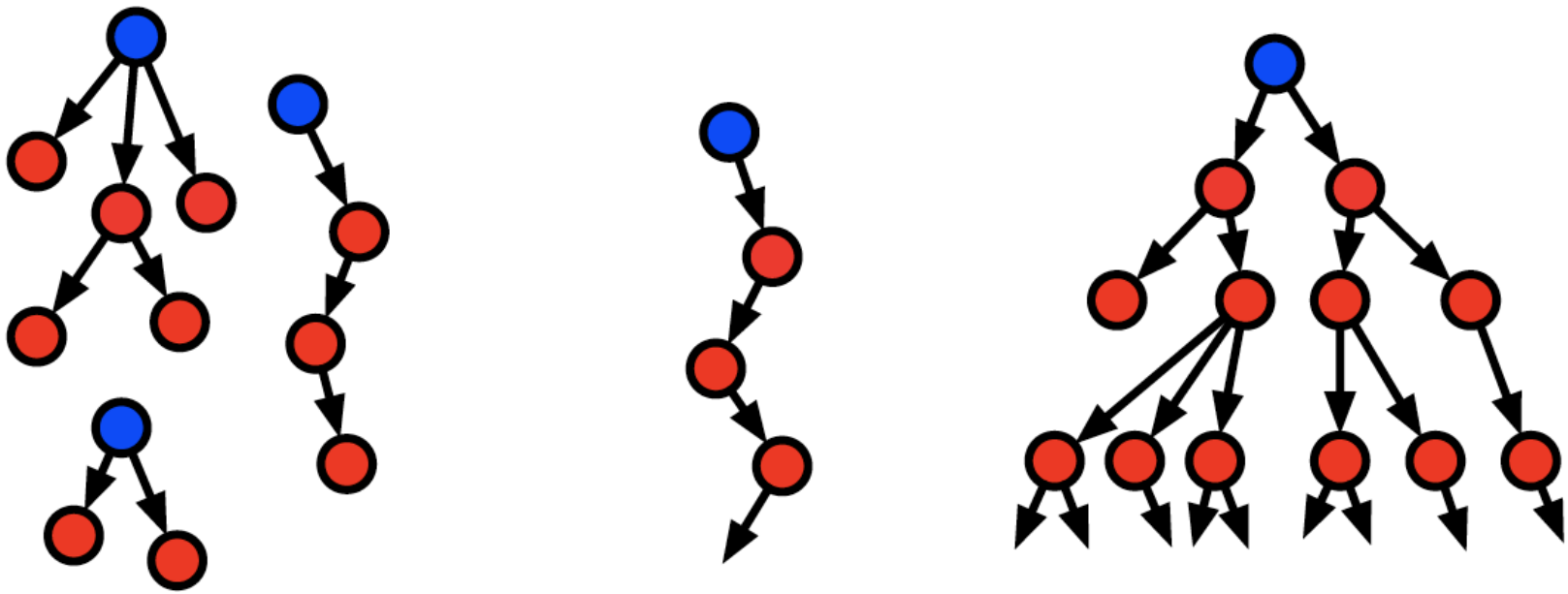


$R_0 = 0.923\ldots$

cascade
epidemic
branching process
spreading process

R_0 = net reproductive rate
= average degree $\langle k \rangle$

caveat:
ignores network structure, dynamics, etc.



$R_0 < 1$

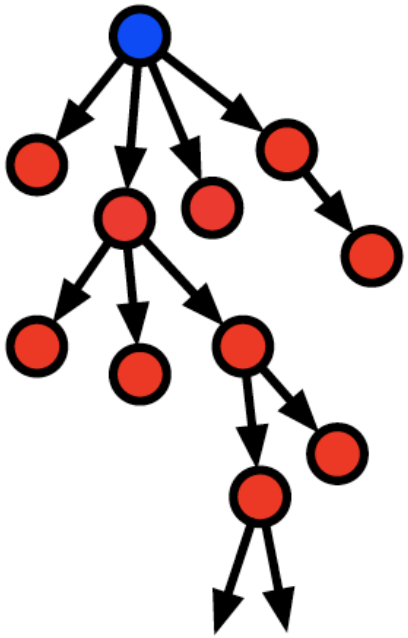
"sub-critical"
small outbreaks

$R_0 = 1$

"critical"
outbreaks of all sizes

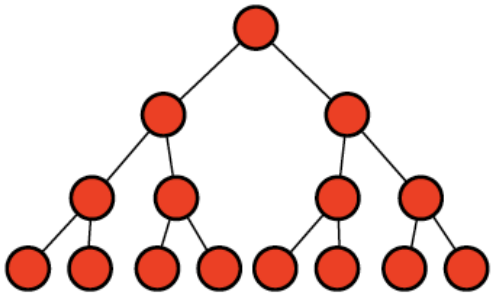
$R_0 > 1$

"super-critical"
global epidemics



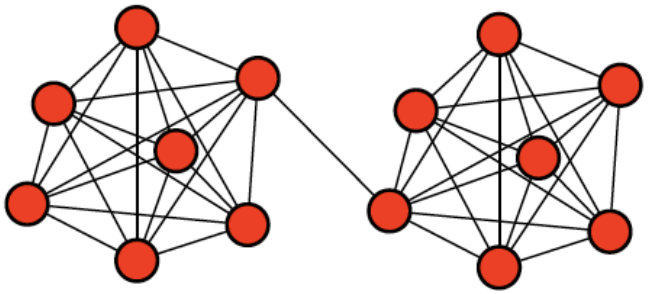
disease	R0	vaccination minimum
Measles	5-18	90-95%
Chicken pox	7-12	85-90%
Polio	5-7	82-87%
Smallpox	1.5-20+	70-80%
H1N1 influenza	1.0-3.0	

all super-critical



bigger cascades

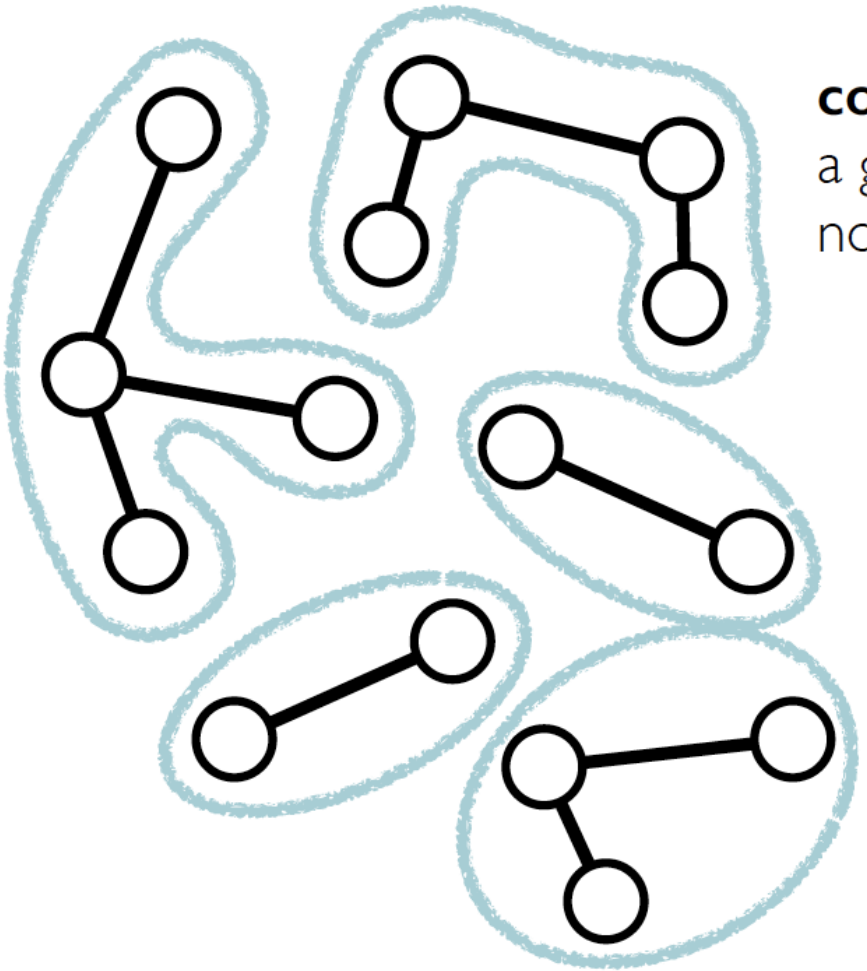
- low overlap among neighbors
- more *expander*-like
- higher transmission probability
- lower activation threshold



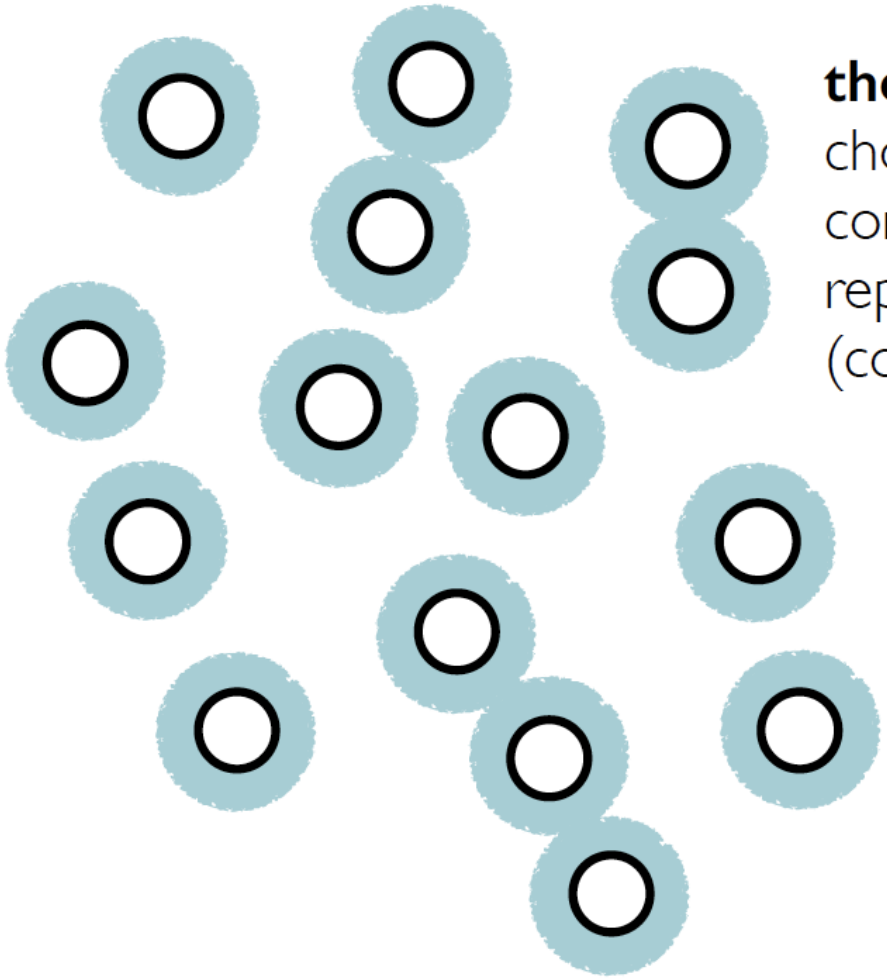
smaller cascades

- larger overlap among neighbors
- more triangles
- smaller "communities"
- more spatial-like organization
- lower transmission probability
- higher activation threshold

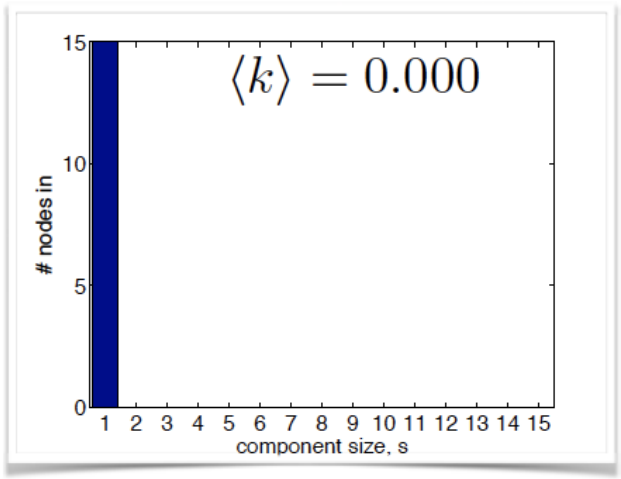
Volz, *J. Math. Bio.* **56**, 293–310 (2008)
Bansal et al., *J. Royal Soc. Interface* **4**, 879–891 (2007)
Karrer and Newman, *Phys. Rev. E* **82**, 016101 (2010)
Salathe and Jones, *PLoS Comp. Bio.* **6**, e1000736 (2010)

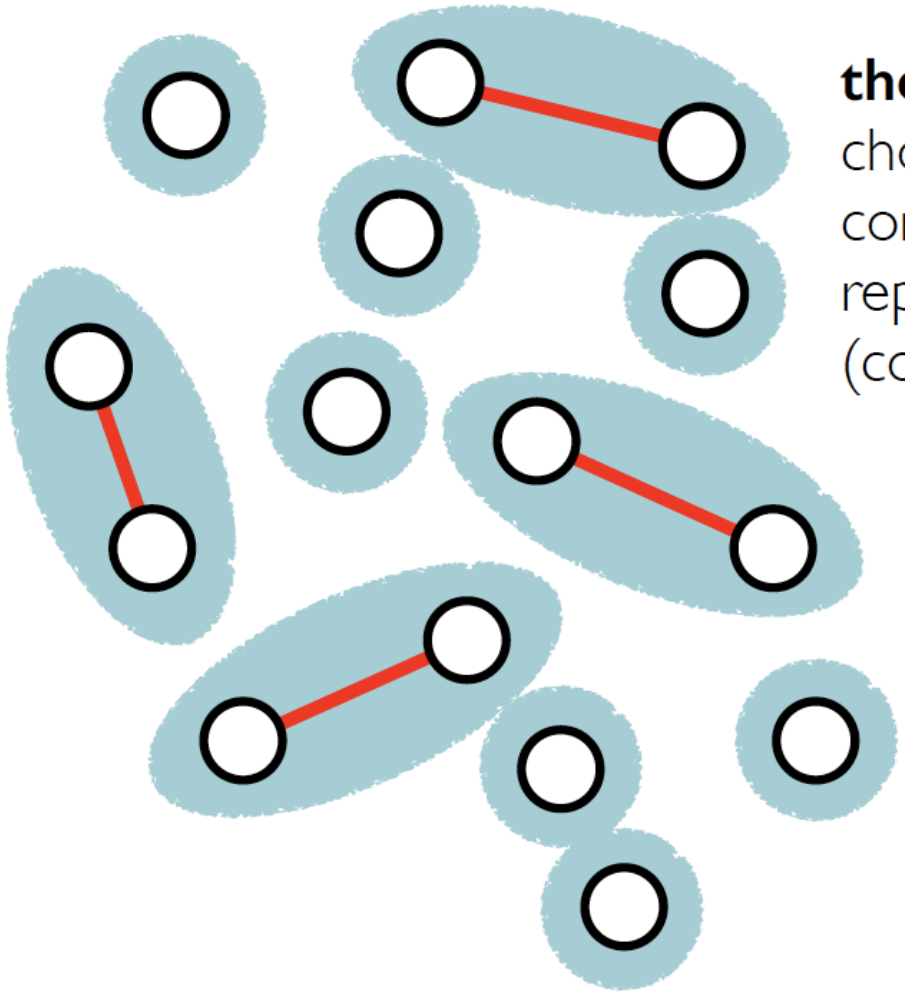


component:
a group of connected nodes

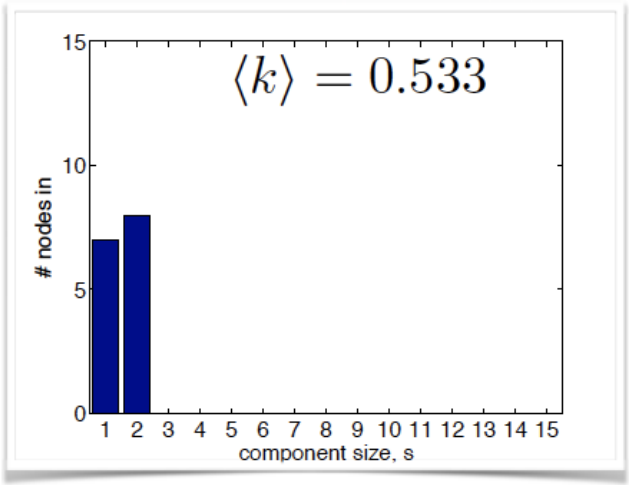


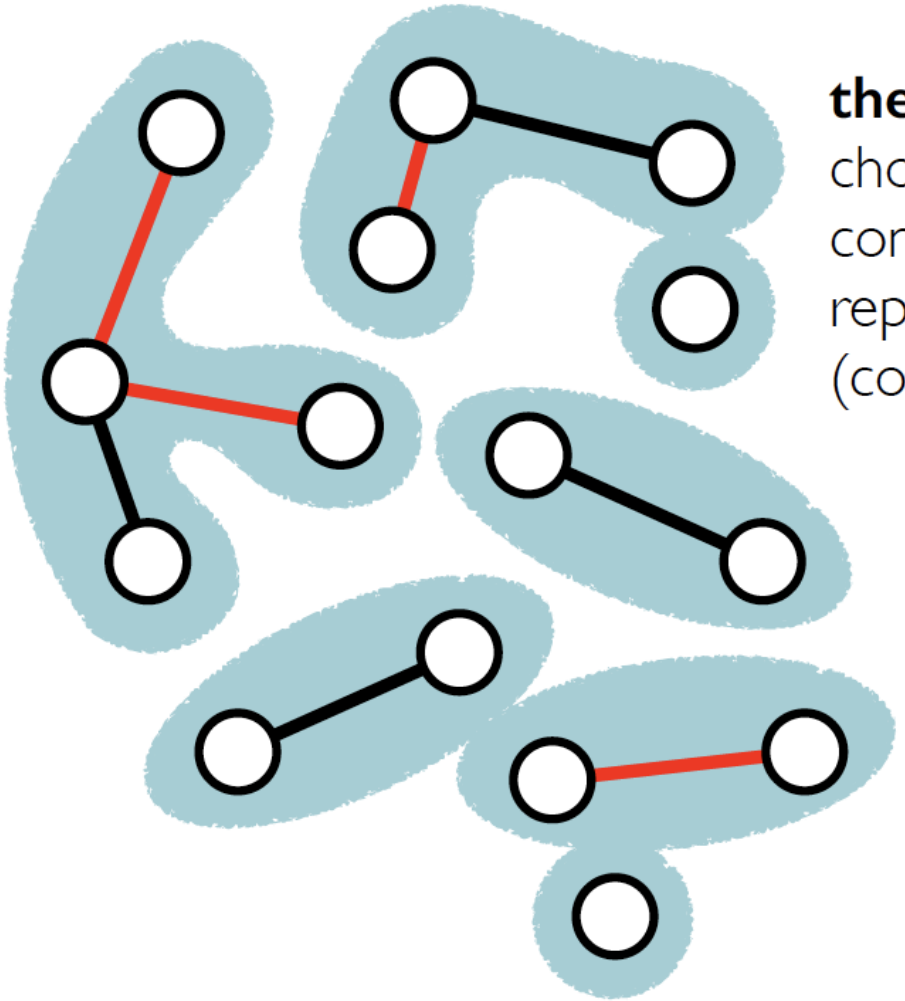
the percolation game:
choose X random pairs
connect them
repeat
(count components)



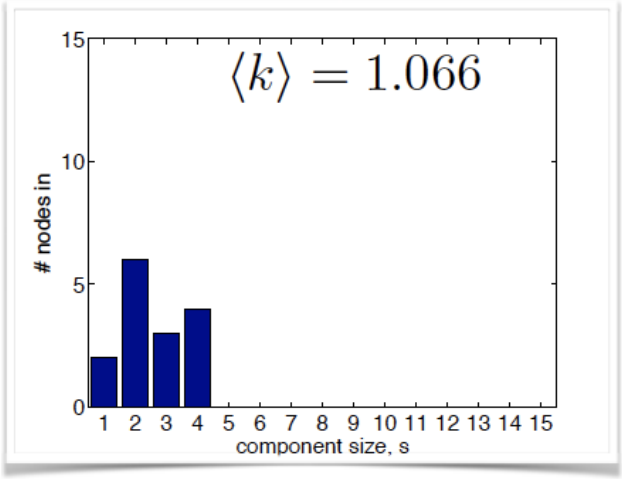


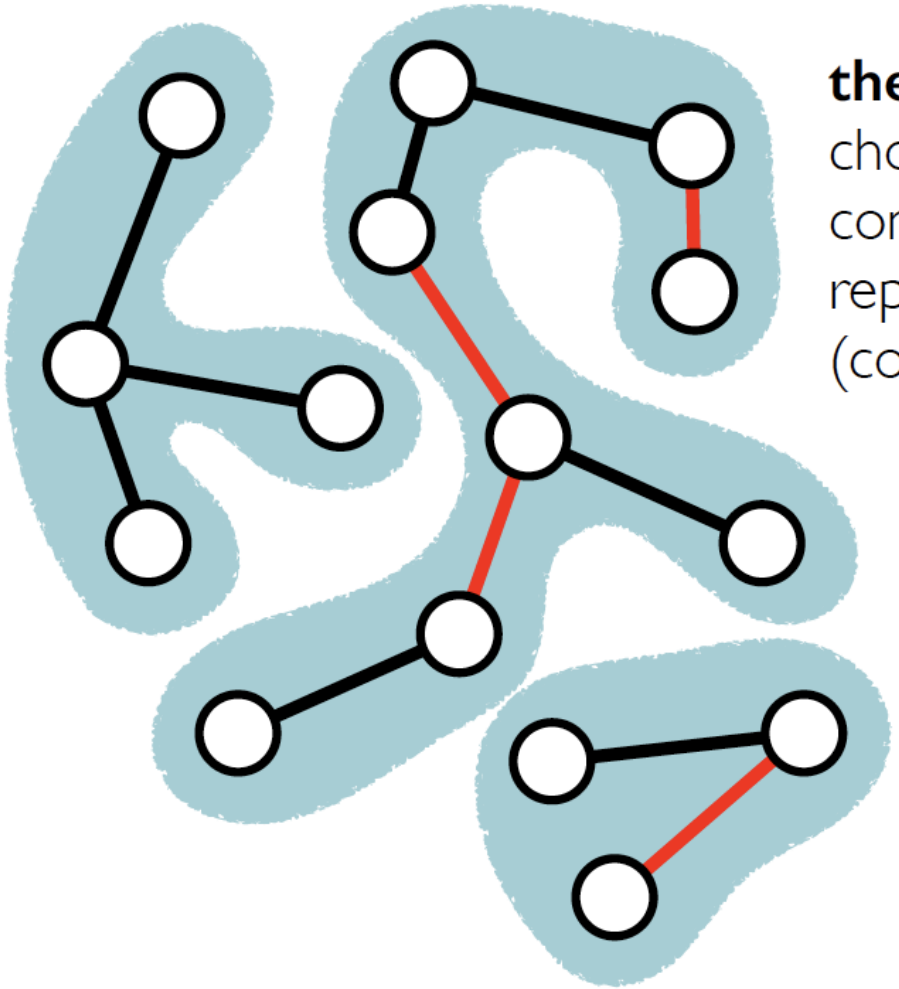
the percolation game:
choose X random pairs
connect them
repeat
(count components)



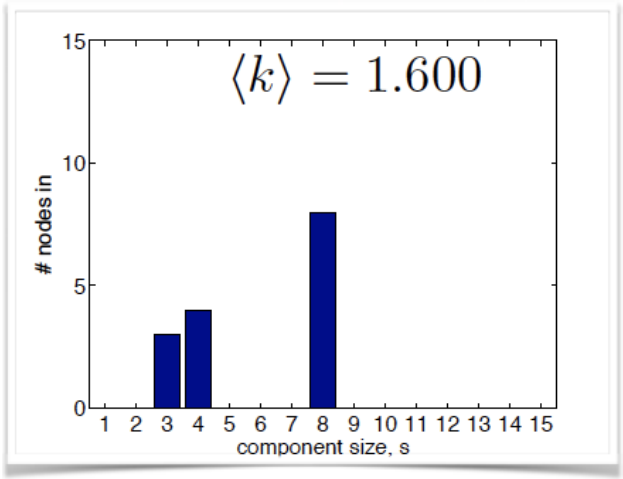


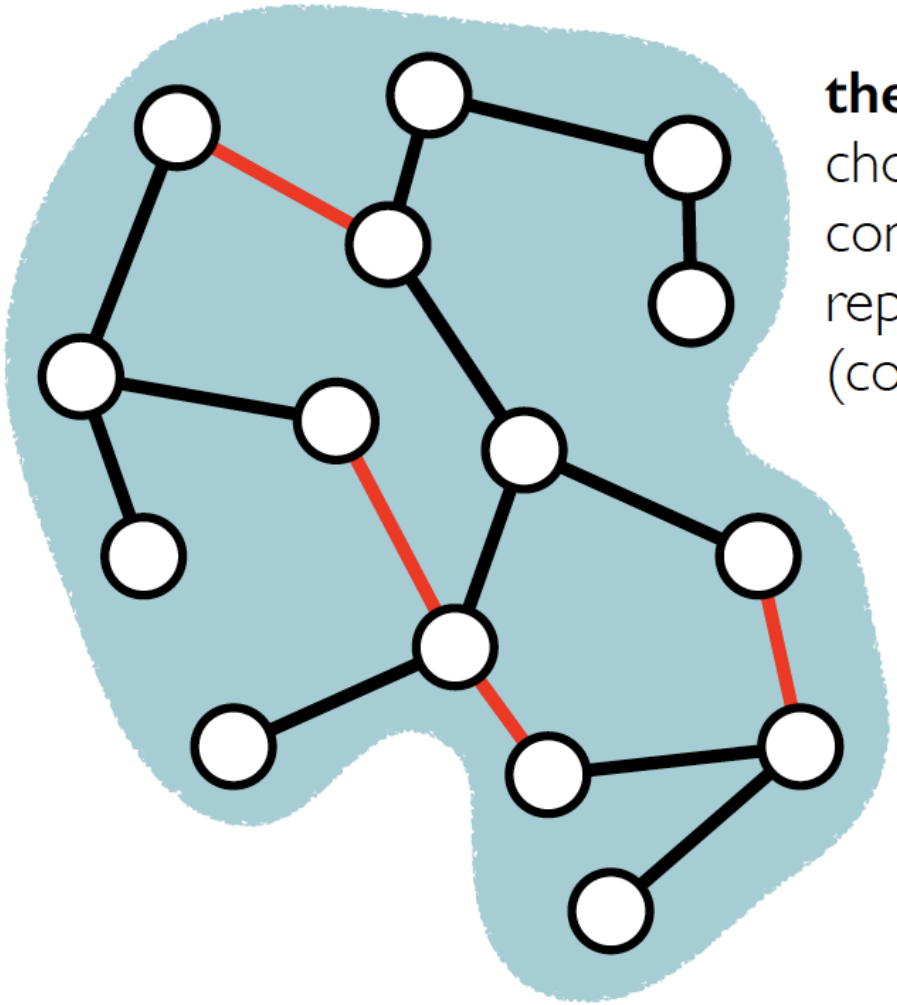
the percolation game:
choose X random pairs
connect them
repeat
(count components)



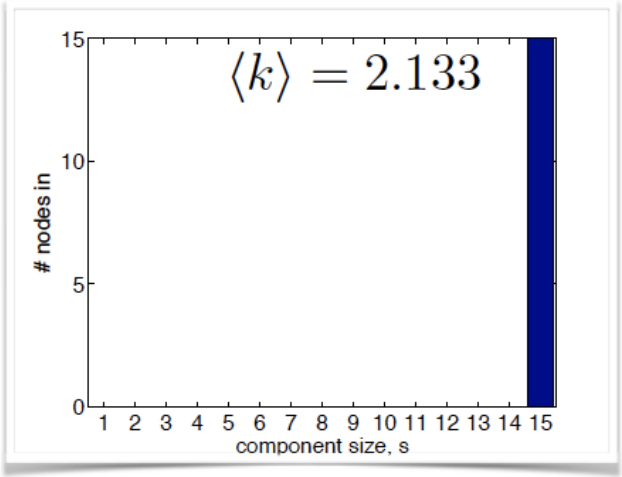


the percolation game:
choose X random pairs
connect them
repeat
(count components)

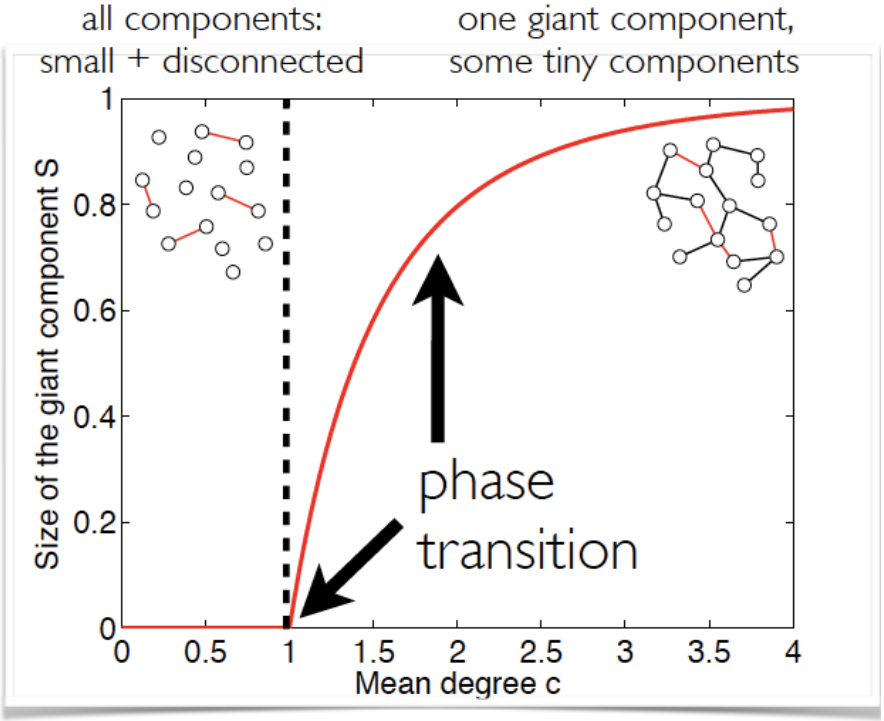


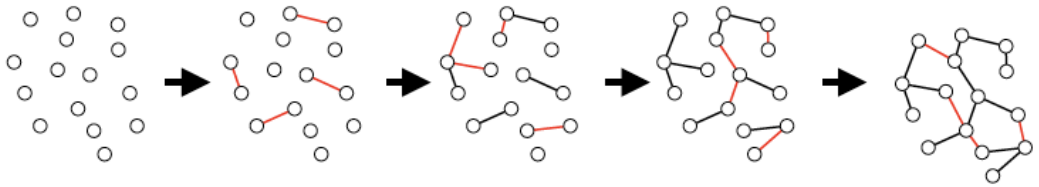


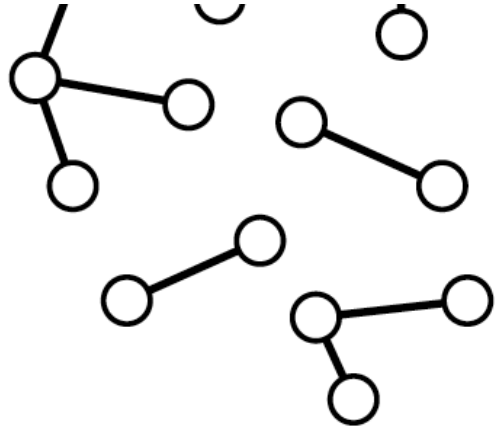
the percolation game:
choose X random pairs
connect them
repeat
(count components)



- add edges randomly
- at first, components are small and disconnected
- at critical value, these components begin linking
- beyond, all nodes in single "giant" component







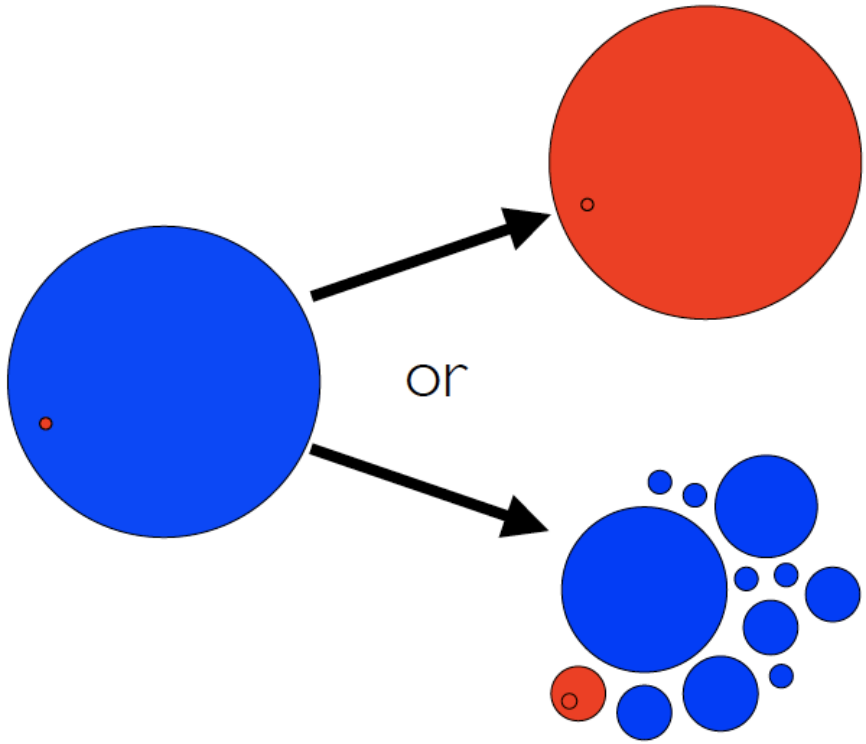
- component = connected group
- component dynamics are independent (no information flow)
- *phase transition in global connectivity*: from smoothly increasing **mean degree**

open questions:

- other network properties + phase transitions
- adaptive wiring
- local vs. global connectivity rules

how could we halt the spread?

- break network into disconnected pieces

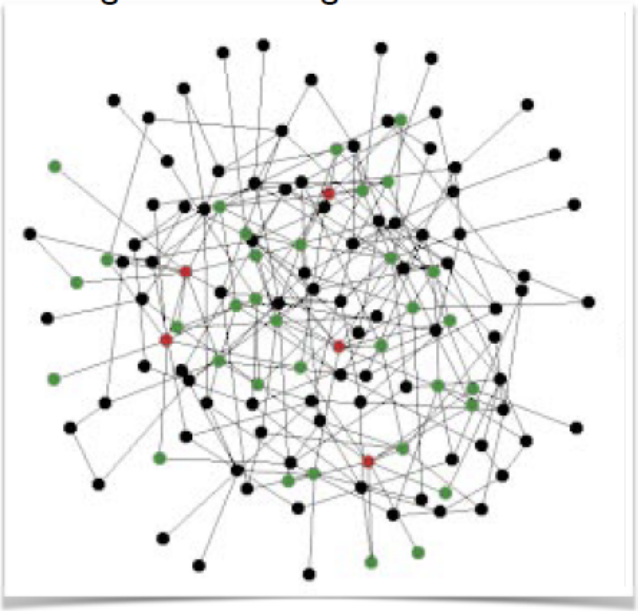


two networks

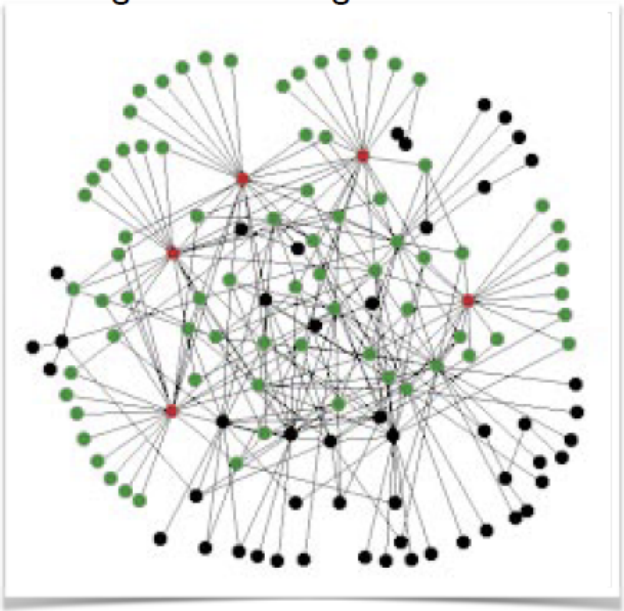
Error and attack tolerance of complex networks 2000

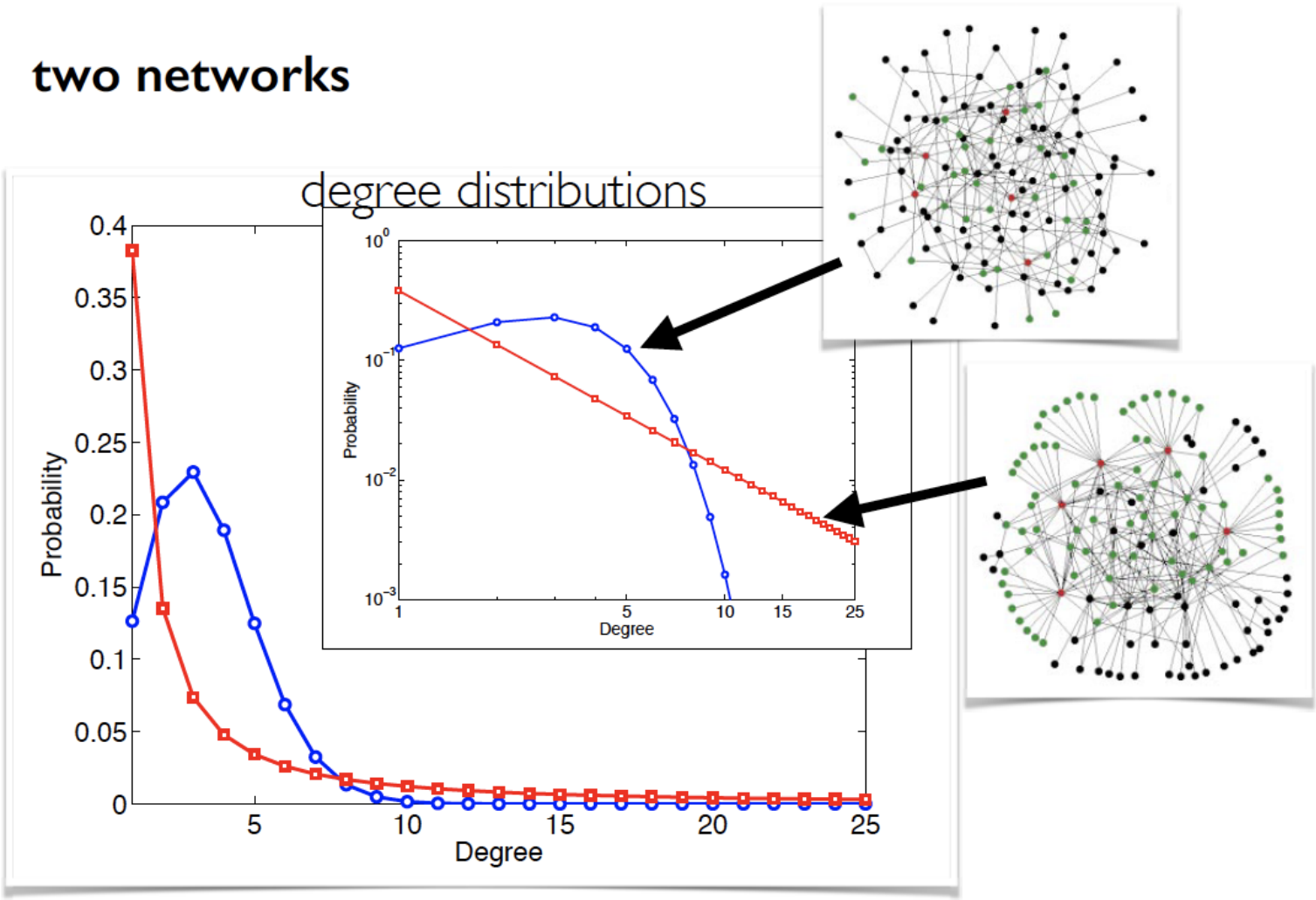
Réka Albert, Hawoong Jeong & Albert-László Barabási

homogeneous in degree



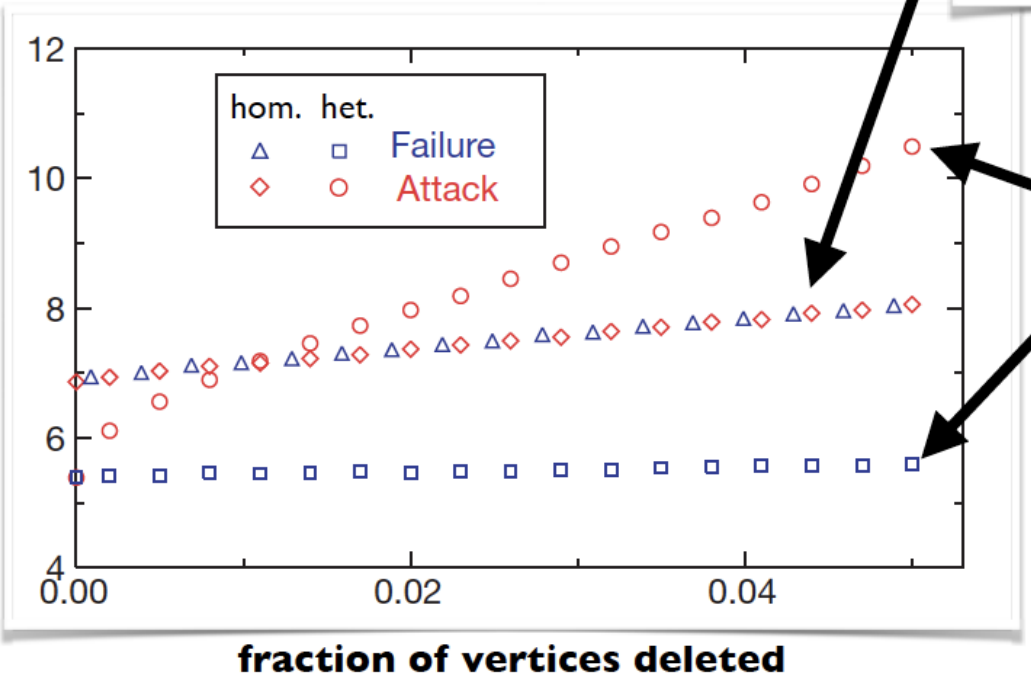
heterogeneous in degree

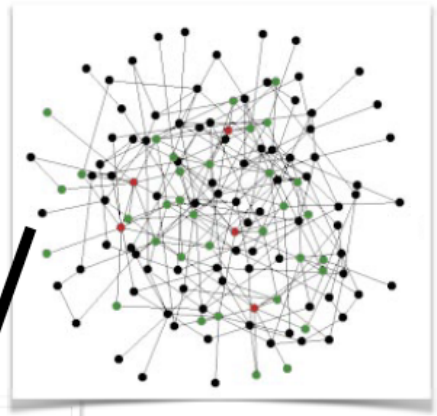


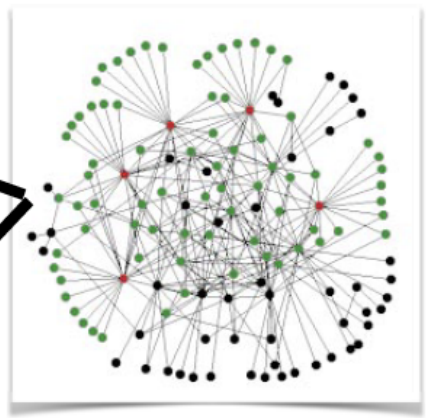

two networks
degree distributions
Probability
Degree
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
5
10
15
20
25
10^0
10^{-1}
10^{-2}
10^{-3}
1

strategy: delete vertices

1. uniformly at random ("failure")
2. in order of degree ("attack")



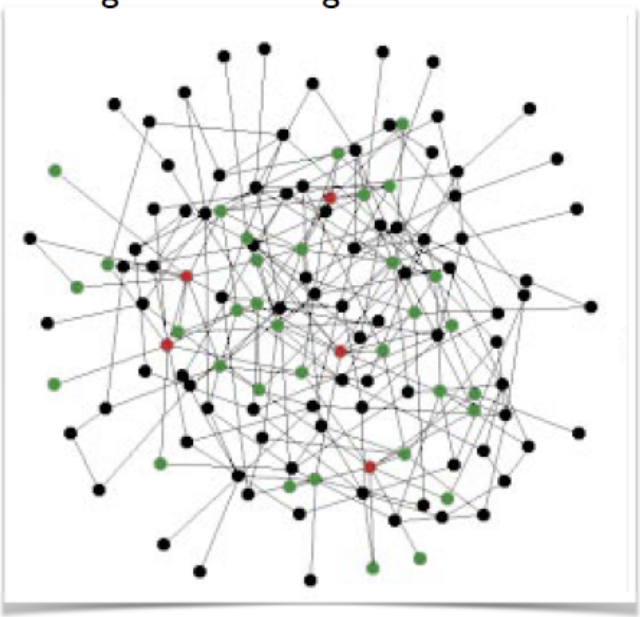




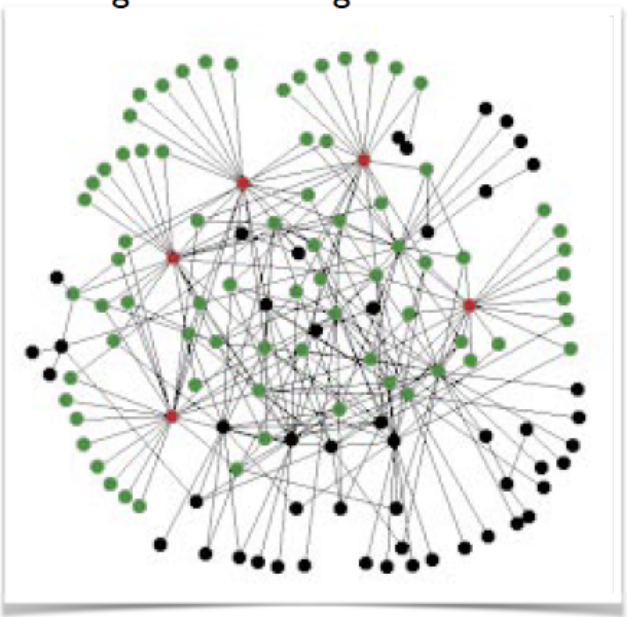
what promotes spreading?

- high-degree vertices*
- centrally-located vertices

homogeneous in degree



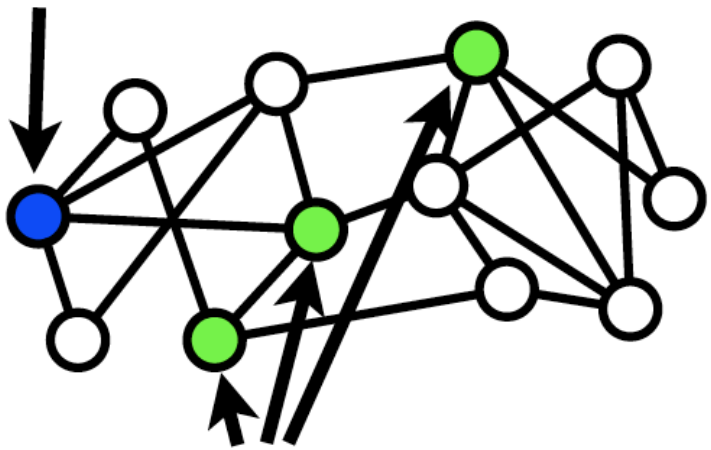
heterogeneous in degree



strategy: delete vertices

3. build "fire breaks"

patient 0



vaccinated = deleted

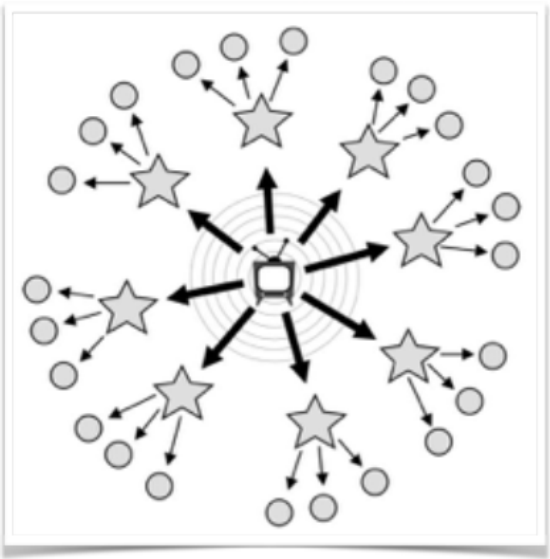
("fire break")

but, in social networks...

Influentials, Networks, and Public Opinion Formation

DUNCAN J. WATTS
PETER SHERIDAN DODDS*

2007



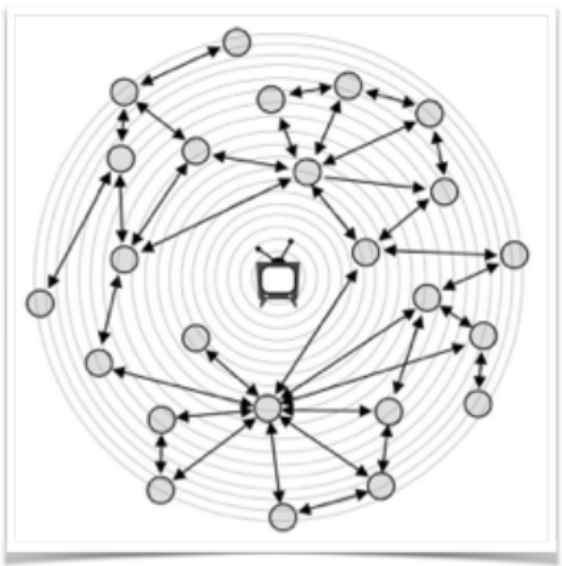
broadcast influence

- classic information marketing
- message saturation
- **degree** is most important

Influentials, Networks, and Public Opinion Formation

DUNCAN J. WATTS
PETER SHERIDAN DODDS*

2007

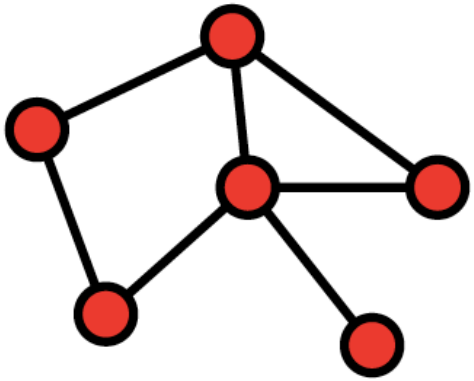


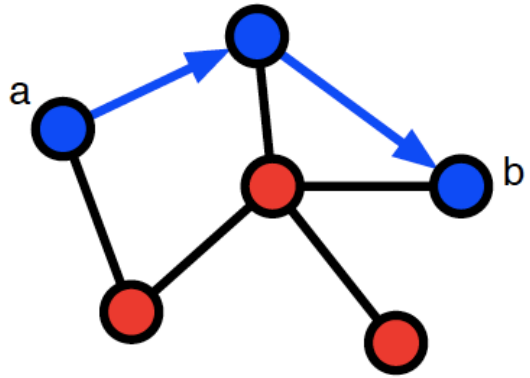
network influence

- "network" (decentralized) marketing
- high-degree = "opinion leader"
- high-degree alone = **irrelevant**
- a cascade requires a legion of *susceptibles* (a system-level property)

how to start a **social movement**?

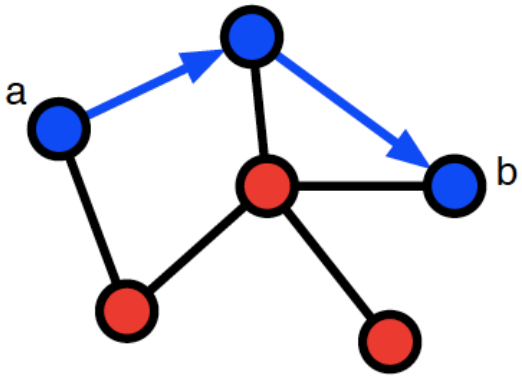
path





path:
number of "hops"
between two nodes

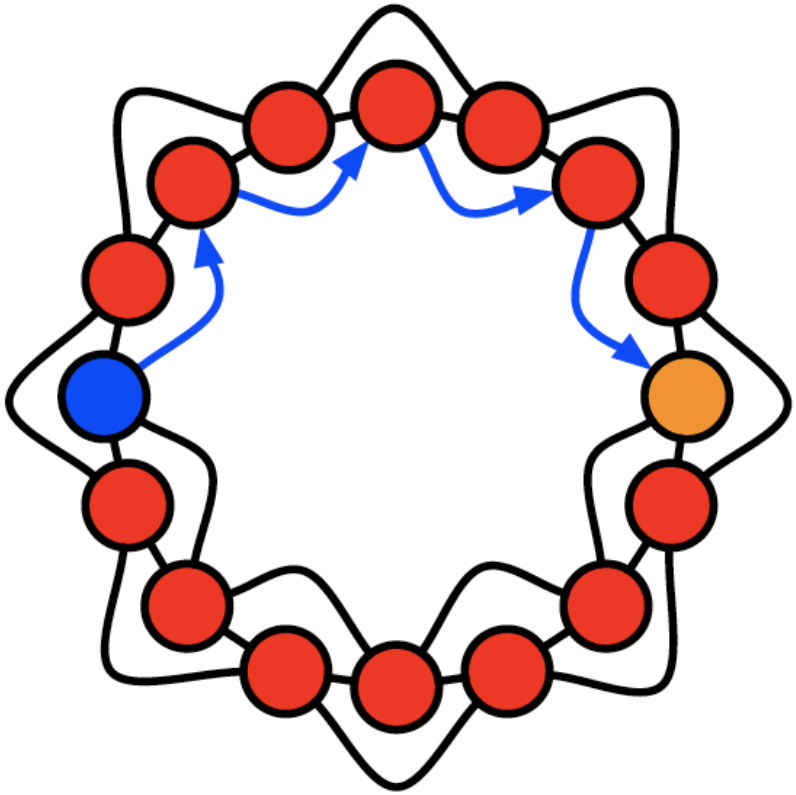
$$\ell_{a \to b} = 2$$



path:
number of "hops"
between two nodes

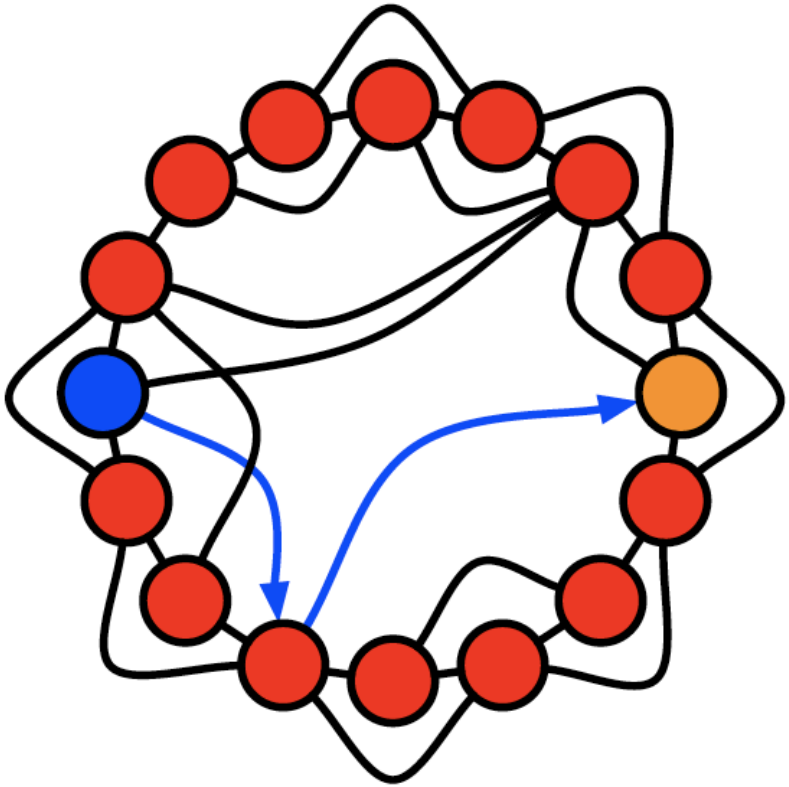
$$\ell_{a \to b} = 2$$

the longer the path,
the weaker the coupling



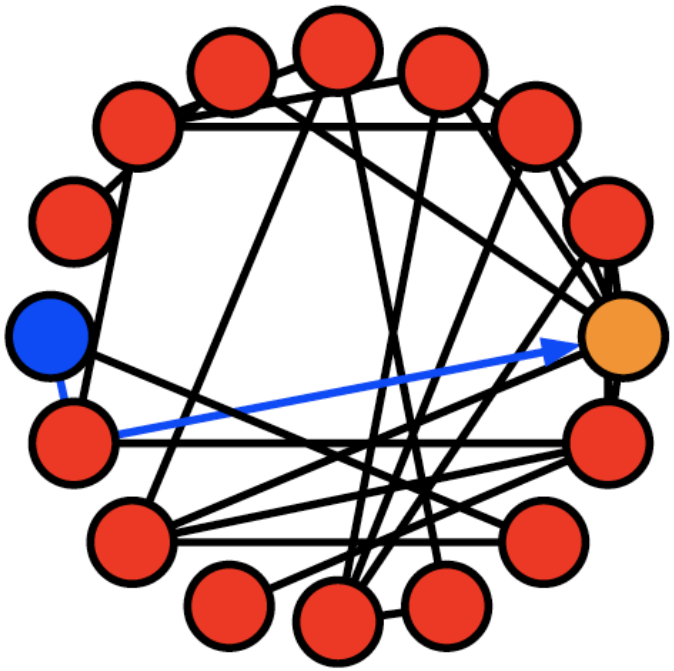
all links "local"

- most nodes far away
- high "clustering"



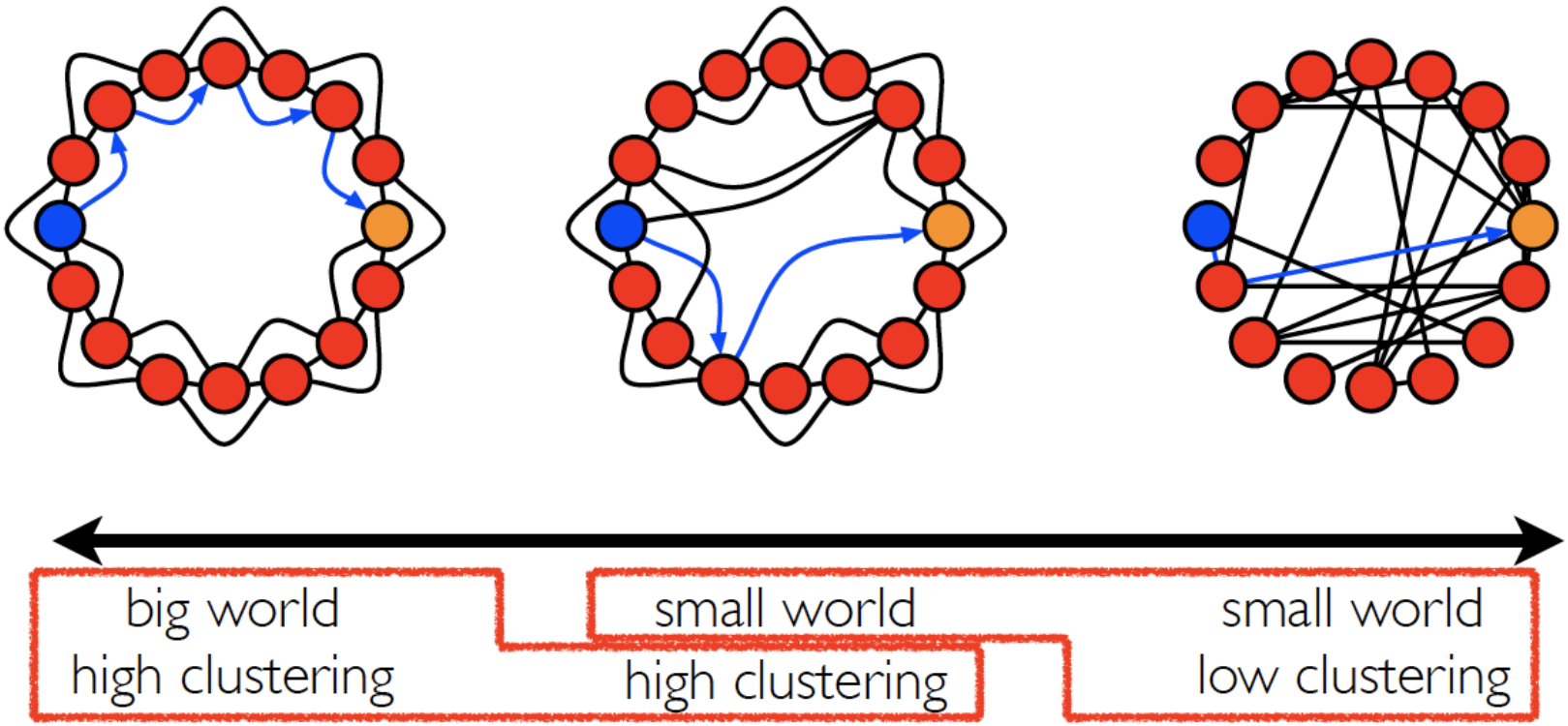
most links "local"
some links random

- most nodes near
- high "clustering"
- short paths can be found



all links random

- Erdos-Renyi graph
- most nodes near
- short paths hard to find
- no "clustering"


big world
high clustering
small world
high clustering
small world
low clustering

Geographic routing in social networks

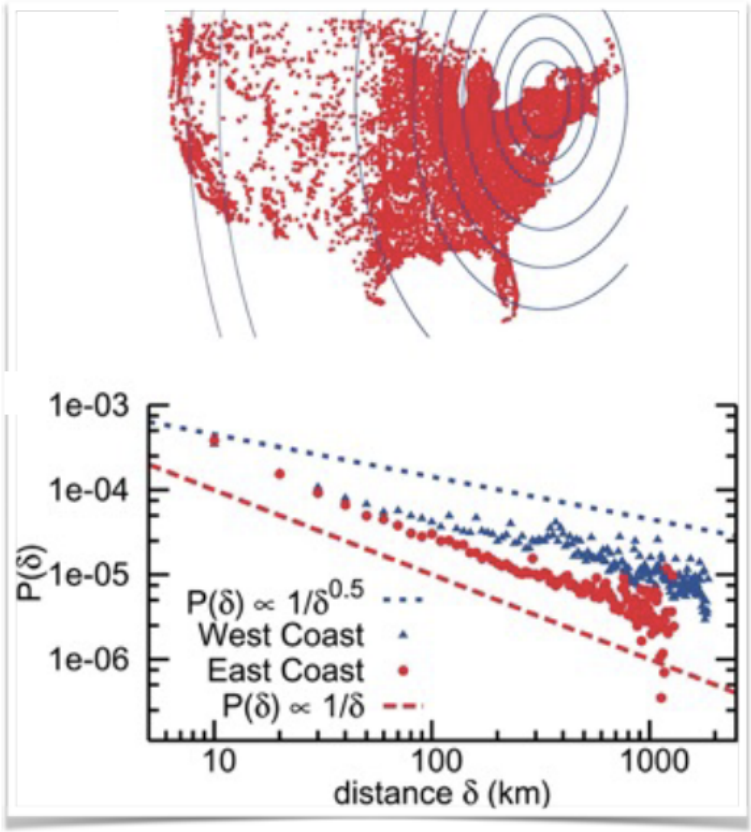
2005

David Liben-Nowell*†‡§, Jasmine Novak†, Ravi Kumar†¶, Prabhakar Raghavan¶‖, and Andrew Tomkins†¶

LiVEJOURNAL™

495,836 geo-located users

- most links "local"
- remaining links span all scales
- high clustering
- small "diameter"

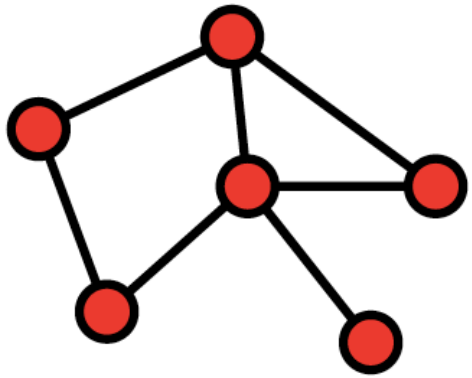


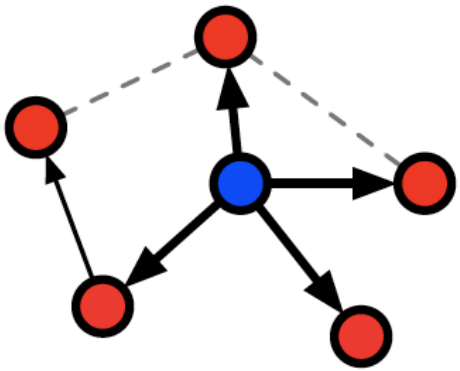
- path = sequence of edges $a \to \cdots \to b$
- path length related to coupling strength
- many short paths = global coupling
- social networks have short paths and many loops
- *nearly all* networks are "small world"

open questions:

- what processes shrink big / grow small worlds?
- social information filtering

position

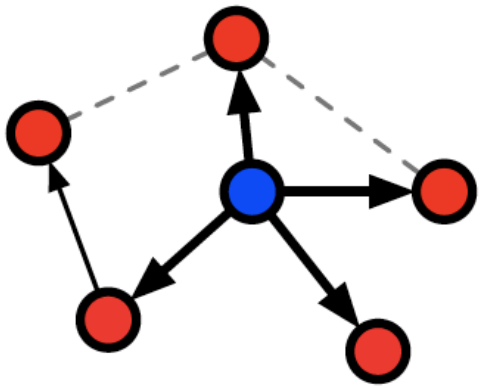




position = centrality:
measure of positional "importance"

geometric	harmonic centrality
	closeness centrality
	betweenness centrality
connectivity	degree centrality
	eigenvector centrality
	PageRank
	Katz centrality
	many many more…

Boldi & Vigna, arxiv:1308.2140 (2013)
Borgatti, *Social Networks* **27**, 55–71 (2005)



position = centrality:
harmonic, closeness centrality

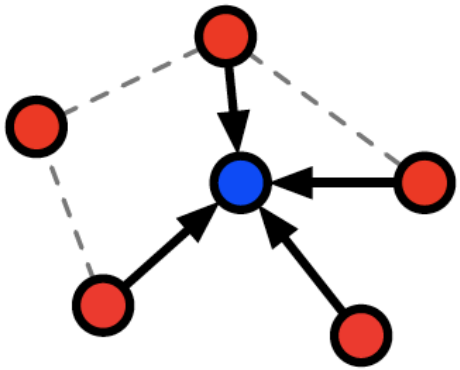
importance = being in "center" of the network

harmonic $$c_i = \frac{1}{n-1}\sum_{j \neq i}\frac{1}{d_{ij}}$$

length of shortest path

distance: $$d_{ij} = \begin{cases} \ell_{ij} & \text{if } j \text{ reachable from } i \\ \infty & \text{otherwise} \end{cases}$$

Boldi & Vigna, arxiv:1308.2140 (2013)
Borgatti, *Social Networks* **27**, 55–71 (2005)



position = centrality:
PageRank, Katz, eigenvector centrality

importance = sum of importances* of nodes that point at you

$$I_i = \sum_{j \to i} \frac{I_j}{k_j}$$

or, the left eigenvector of

$$\mathbf{A}\mathbf{x} = \lambda \mathbf{x}$$

Boldi & Vigna, arxiv:1308.2140 (2013)
Borgatti, *Social Networks* **27**, 55–71 (2005)

*modulo several technical details

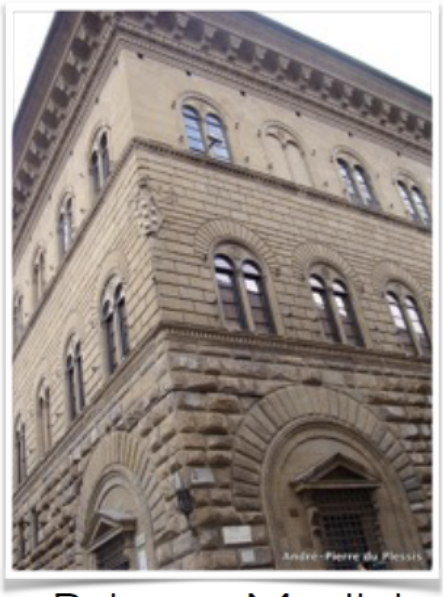
an example

Giovanni de Medici

Robust Action and the Rise of the Medici, 1400–1434[1]

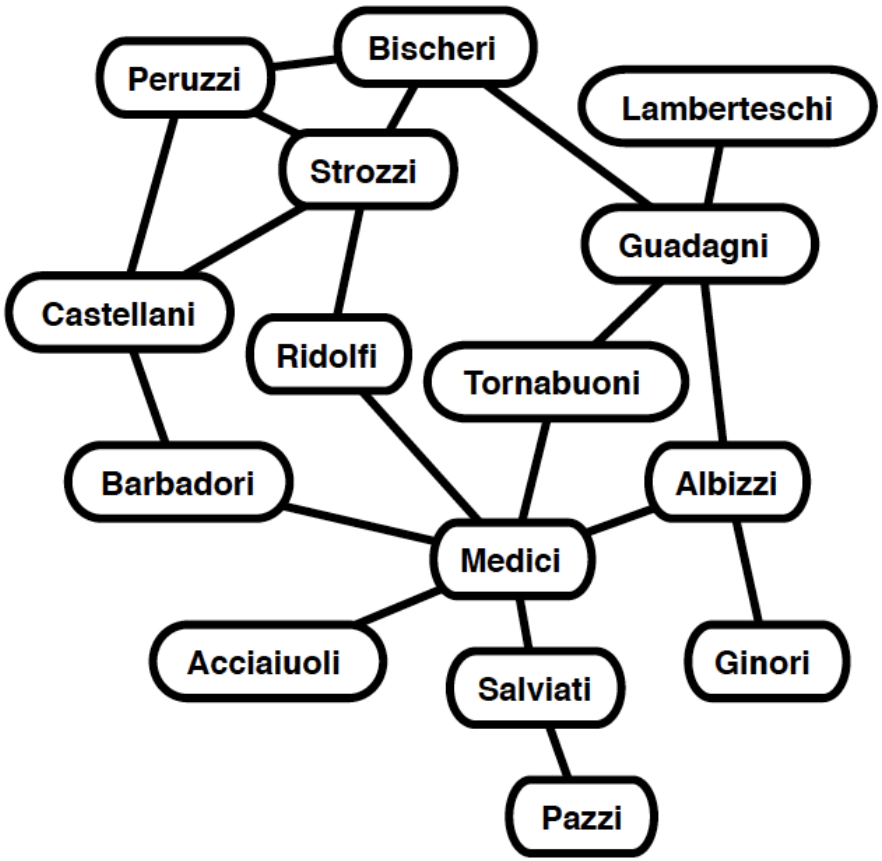
John F. Padgett and Christopher K. Ansell 1993

Duomo



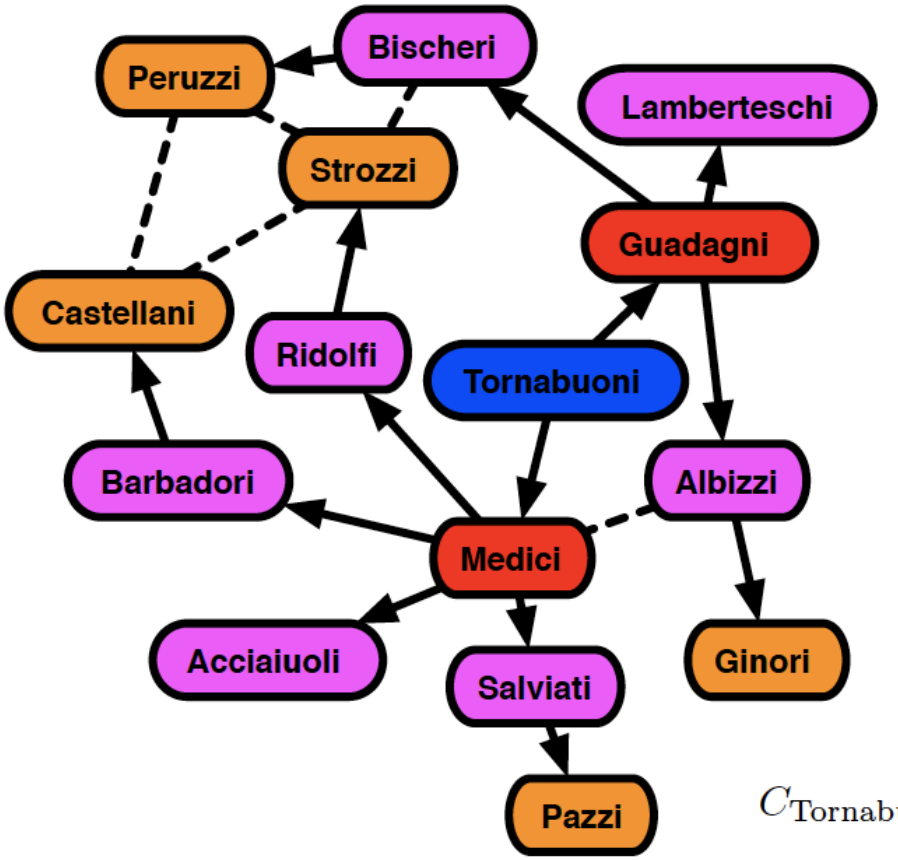
Palazzo Medici

Giovanni de Medici



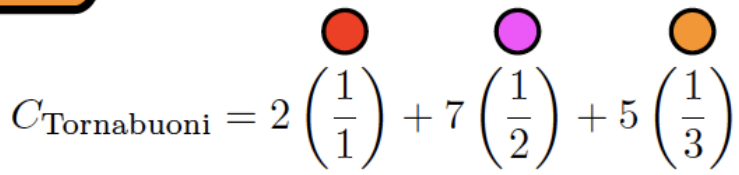
nodes: Florence families
edges: inter-family marriages

which family is most central?

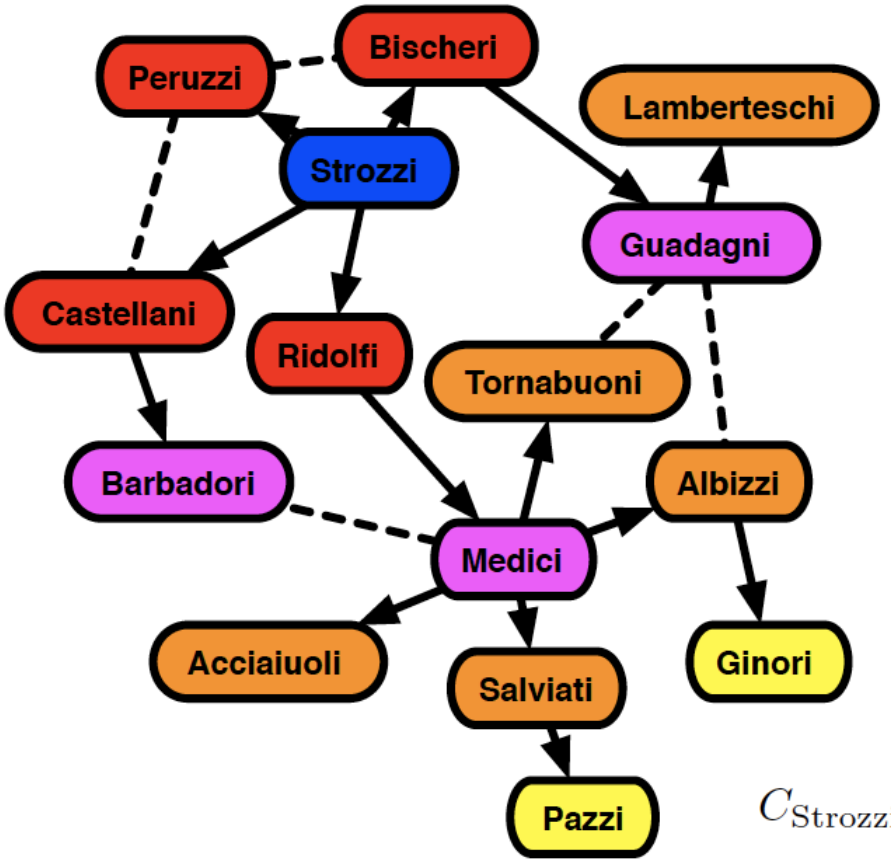


nodes: Florence families
edges: inter-family marriages

Tornabuoni?



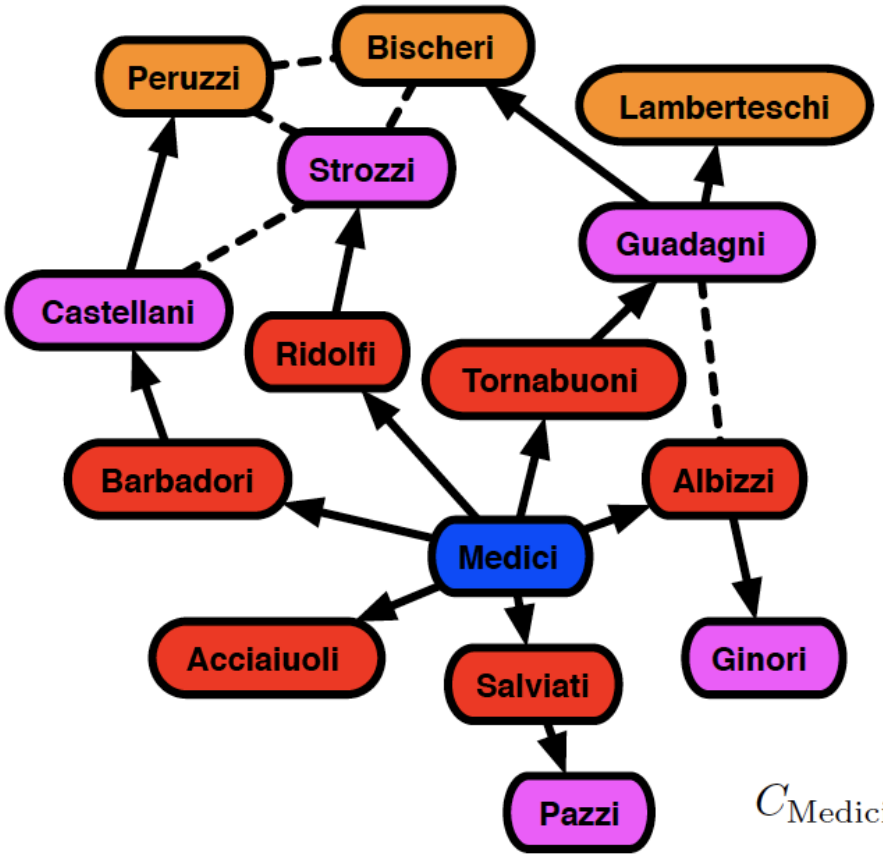
$$C_{\text{Tornabuoni}} = 2\left(\frac{1}{1}\right) + 7\left(\frac{1}{2}\right) + 5\left(\frac{1}{3}\right)$$
$$= 7.16\ldots$$



nodes: Florence families
edges: inter-family marriages

Strozzi?

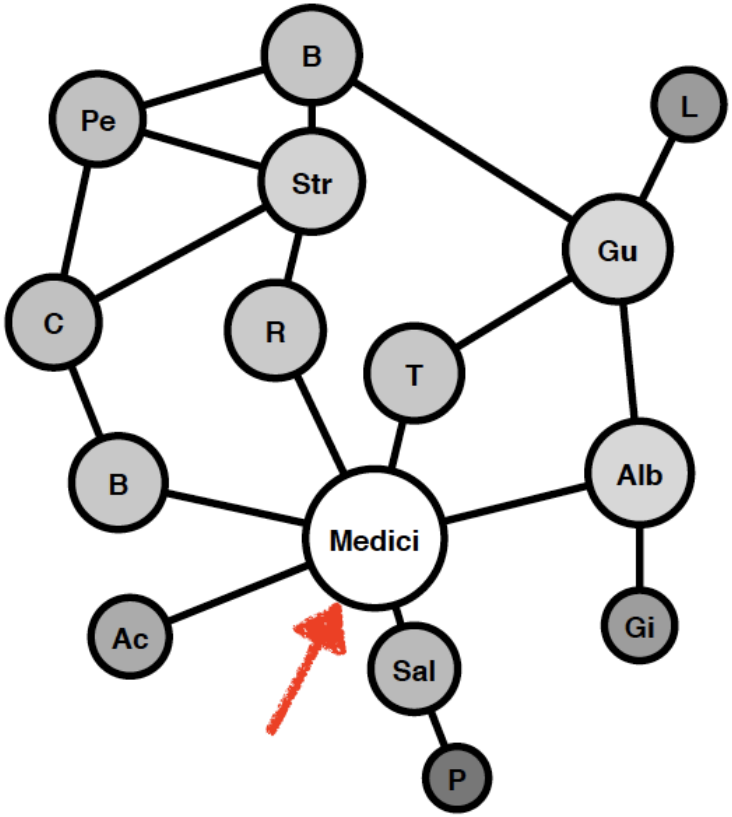
$$C_{\text{Strozzi}} = 4\left(\frac{1}{1}\right) + 3\left(\frac{1}{2}\right) + 5\left(\frac{1}{3}\right) + 2\left(\frac{1}{4}\right)$$
$$= 7.6\ldots$$



nodes: Florence families
edges: inter-family marriages

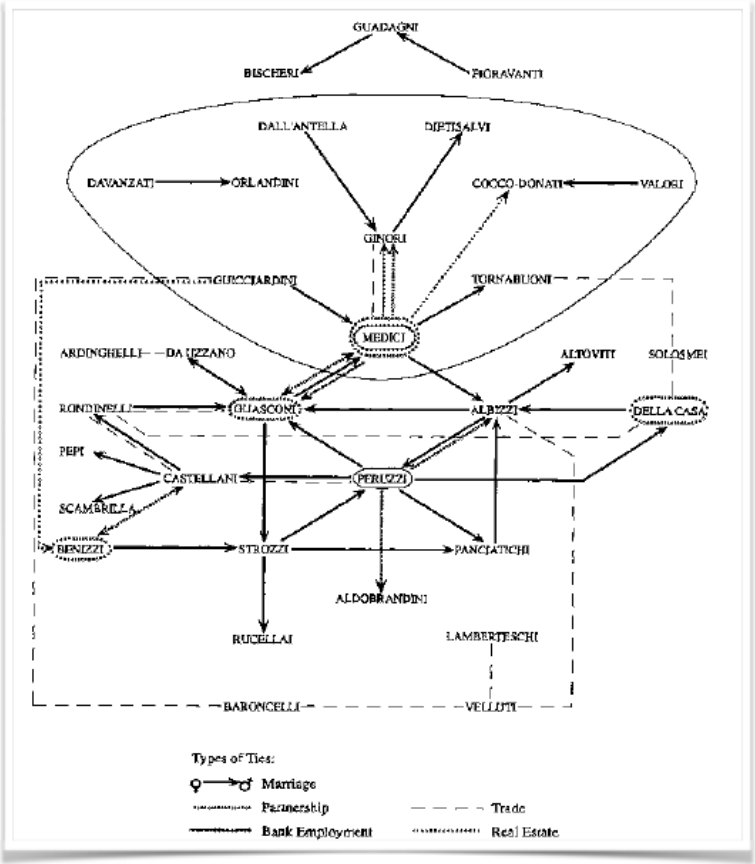
Medici.

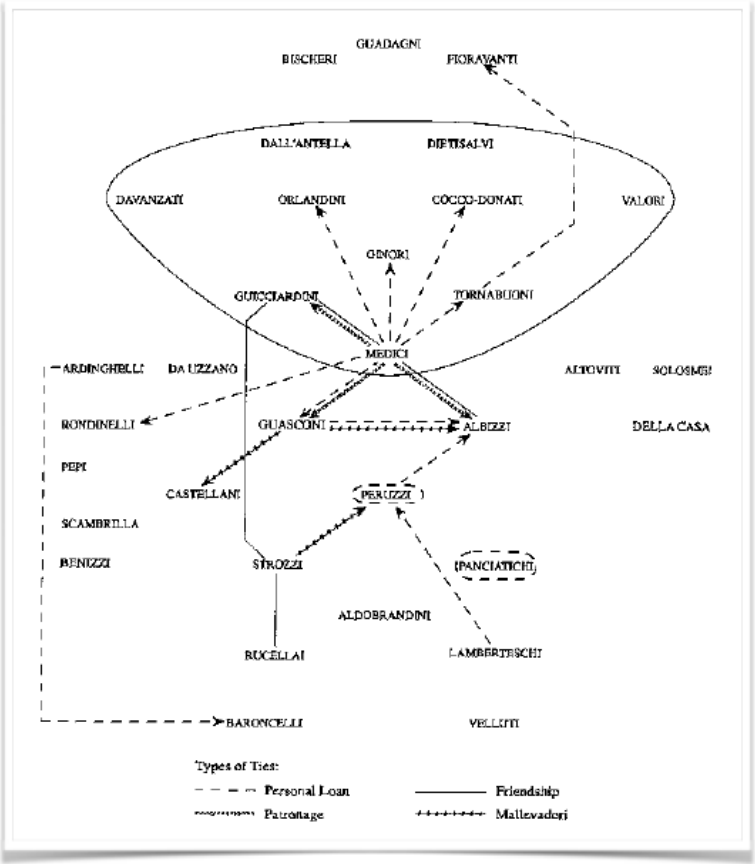
$$C_{\text{Medici}} = 6\left(\frac{1}{1}\right) + 5\left(\frac{1}{2}\right) + 3\left(\frac{1}{3}\right) = 9.5$$

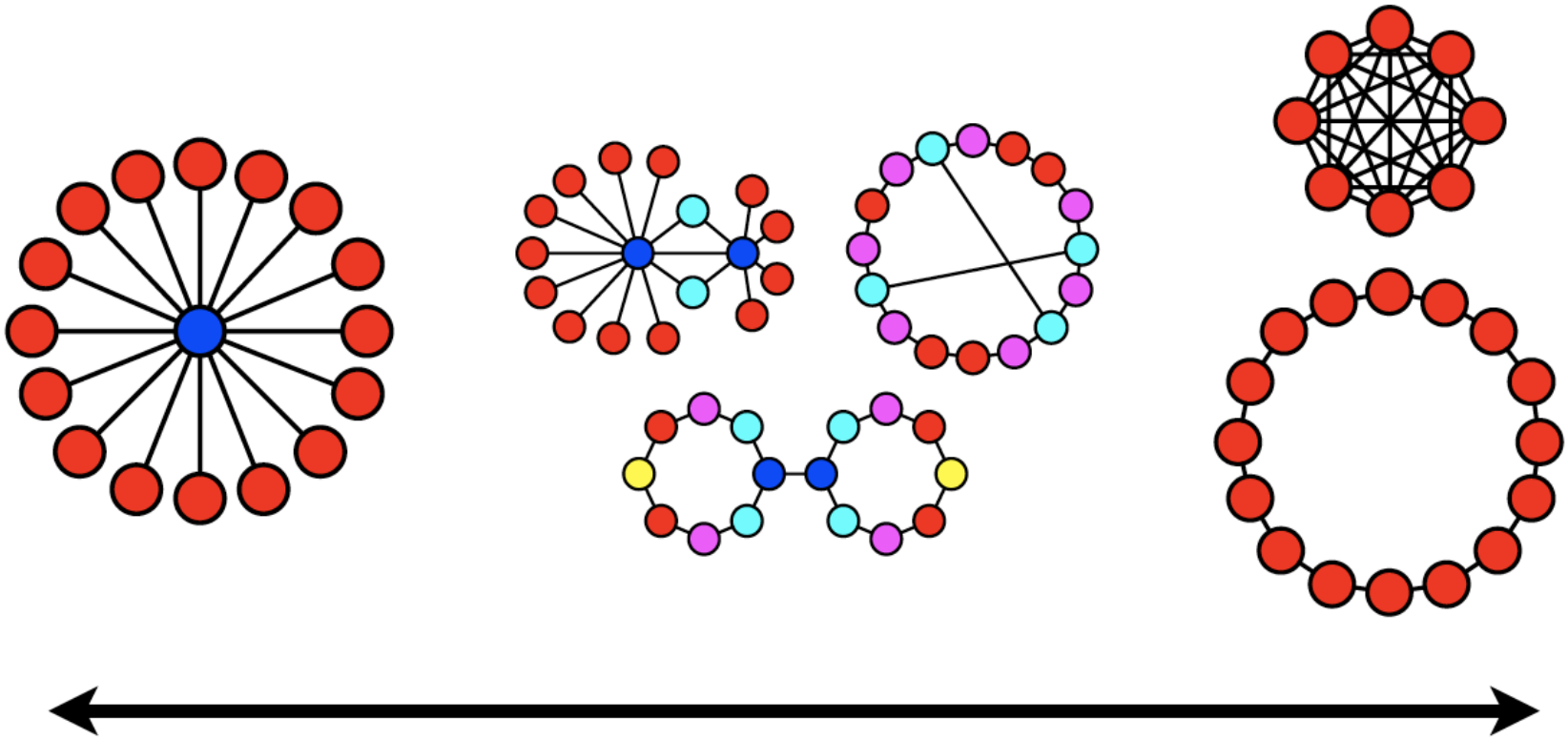


Medici	9.5
Guadagni	7.92
Albizzi	7.83
Strozzi	7.67
Ridolfi	7.25
Bischeri	7.2
Tornabuoni	7.17
Barbadori	7.08
Peruzzi	6.87
Castellani	6.87
Salviati	6.58
Acciaiuoli	5.92
Ginori	5.33
Lamberteschi	5.28
Pazzi	4.77

actually, it's complicated...

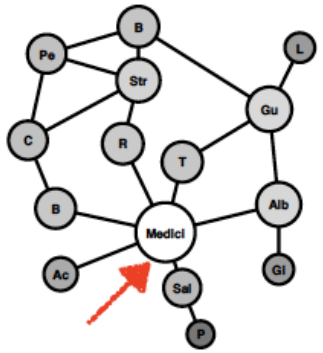





most
centralized
vast wilderness
of in-between
most
decentralized

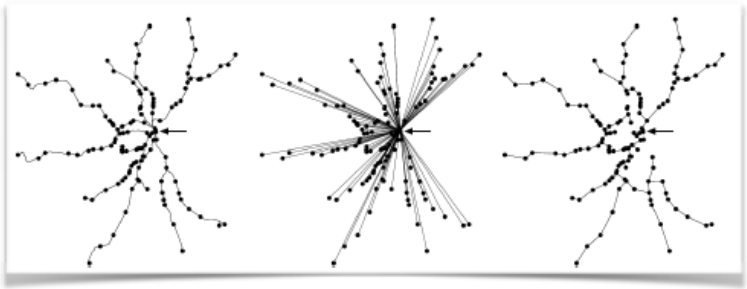
positions:

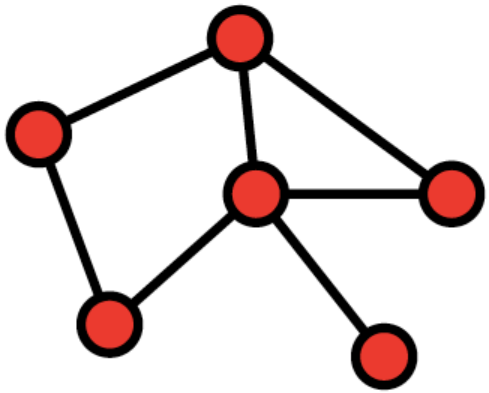
- *geometric* description of network structure
- core vs. periphery
- centrality = importance, influence



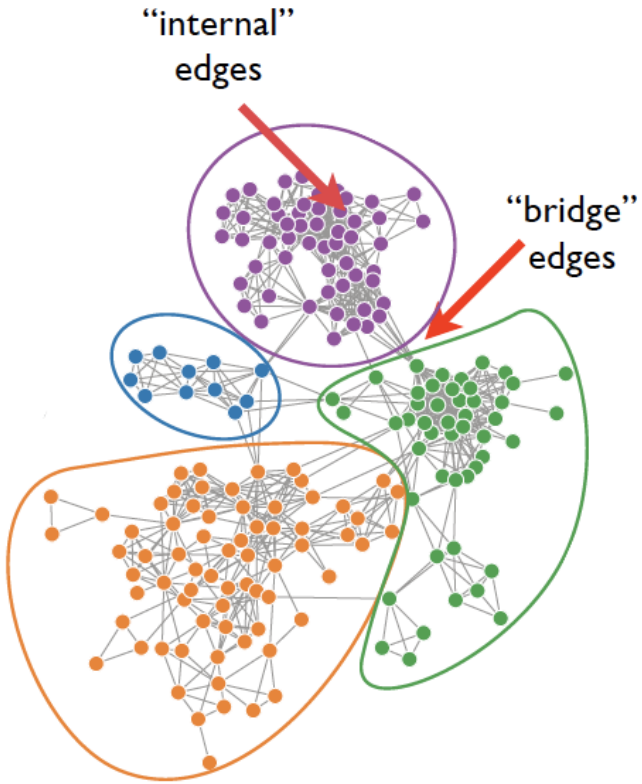
open questions:

- position and dynamics
- what does position predict?
- when does position *not* matter?





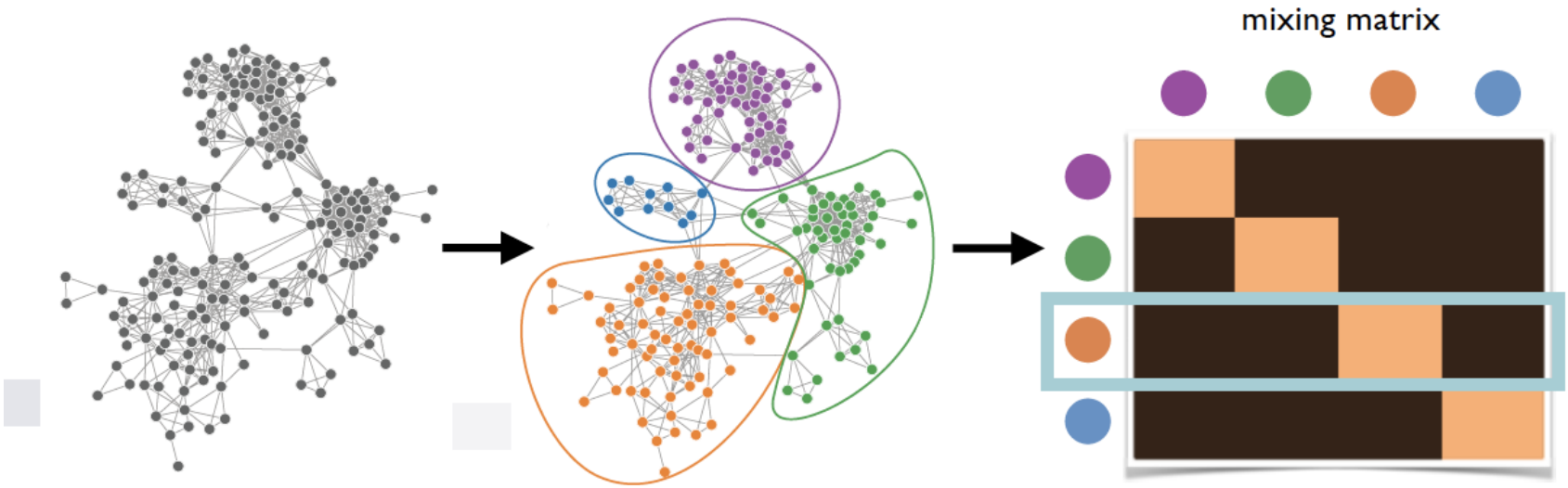
community structure

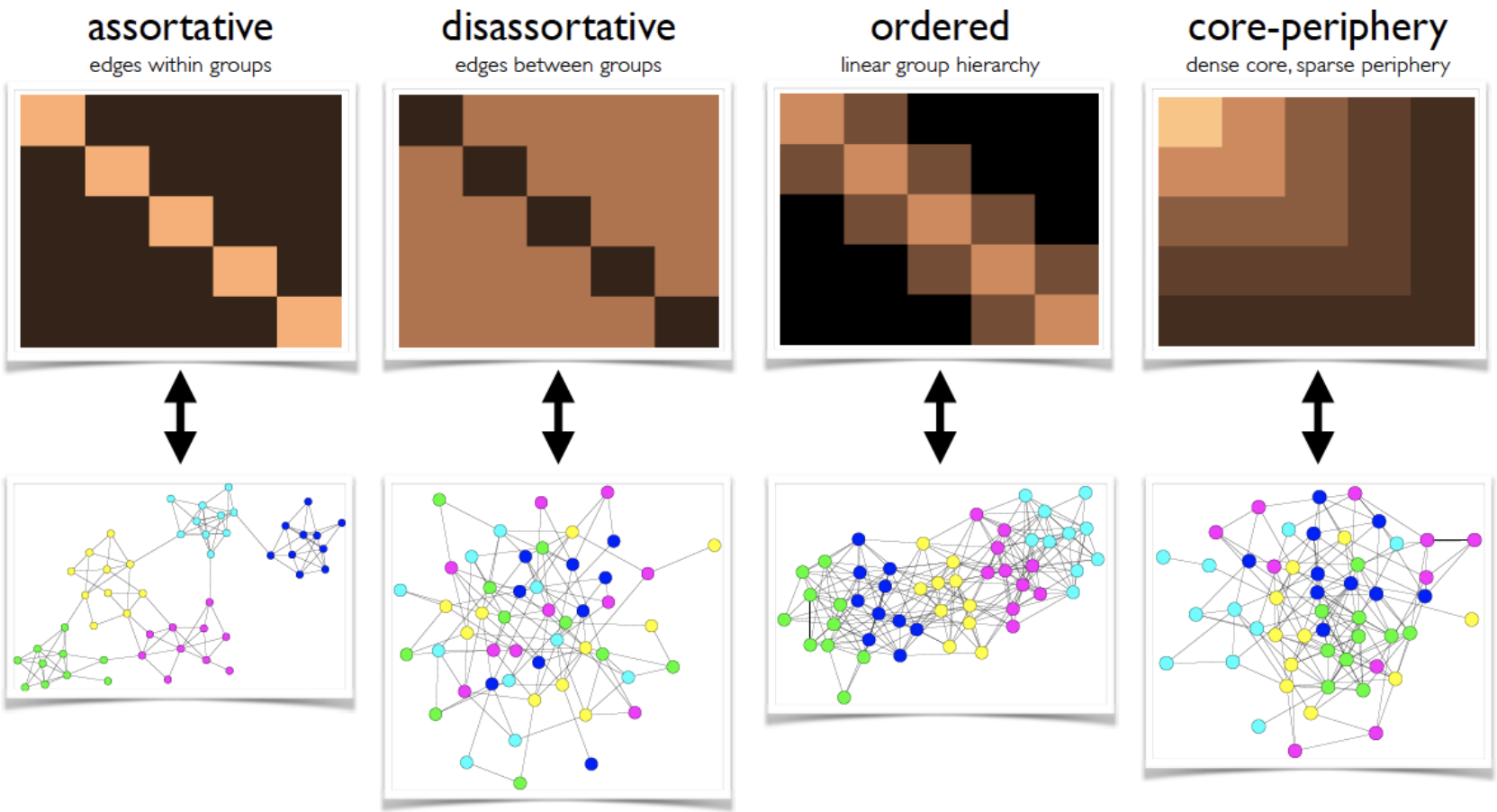


assortative community structure
(edges inside the groups)

community structure:
a *group of vertices* that
connect to *other groups*
in similar ways

community structure:
a *group of vertices* that
connect to *other groups*
in similar ways



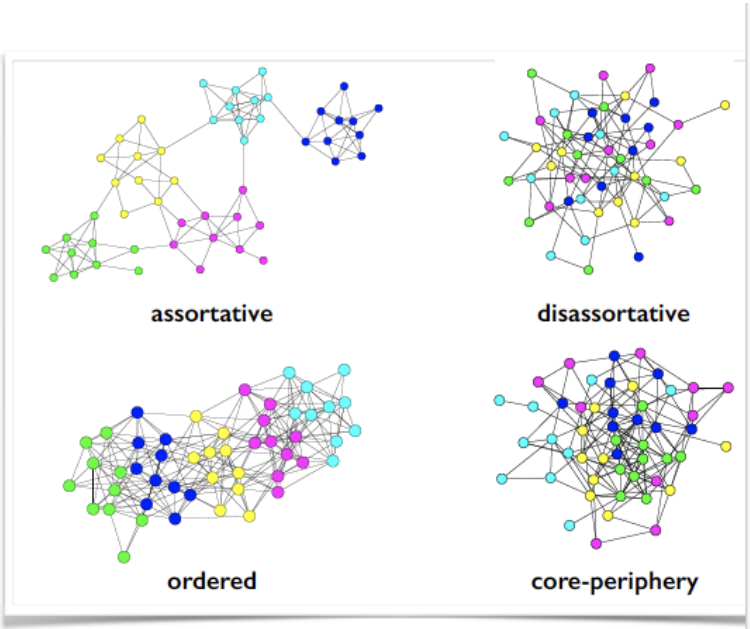

assortative
edges within groups
disassortative
edges between groups
ordered
linear group hierarchy
core-periphery
dense core, sparse periphery

- enormous interest, especially since 2000
- dozens of algorithms for extracting various large-scale patterns
- hundreds of papers published
- spanning Physics, Computer Science, Statistics, Biology, Sociology, and more
- this was one of the first:

Community structure in social and biological networks

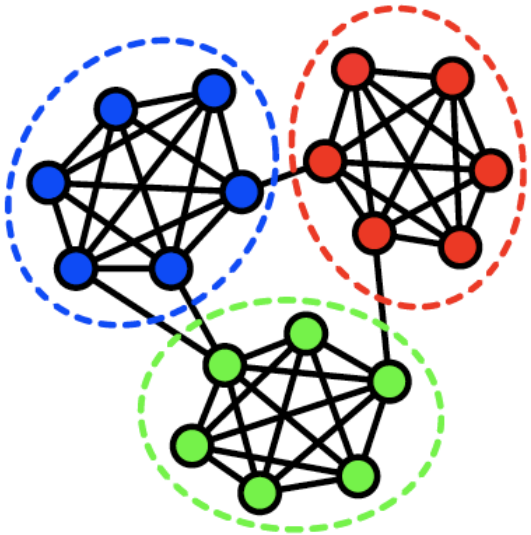
M. Girvan*†‡ and M. E. J. Newman*§ *PNAS* 2002

7200+ citations on Google Scholar



THE STRENGTH OF WEAK TIES: A NETWORK THEORY REVISITED

1983 *Mark Granovetter*



most new job opportunities from "weak ties"

- within-community links = strong
- bridge links = weak

why?

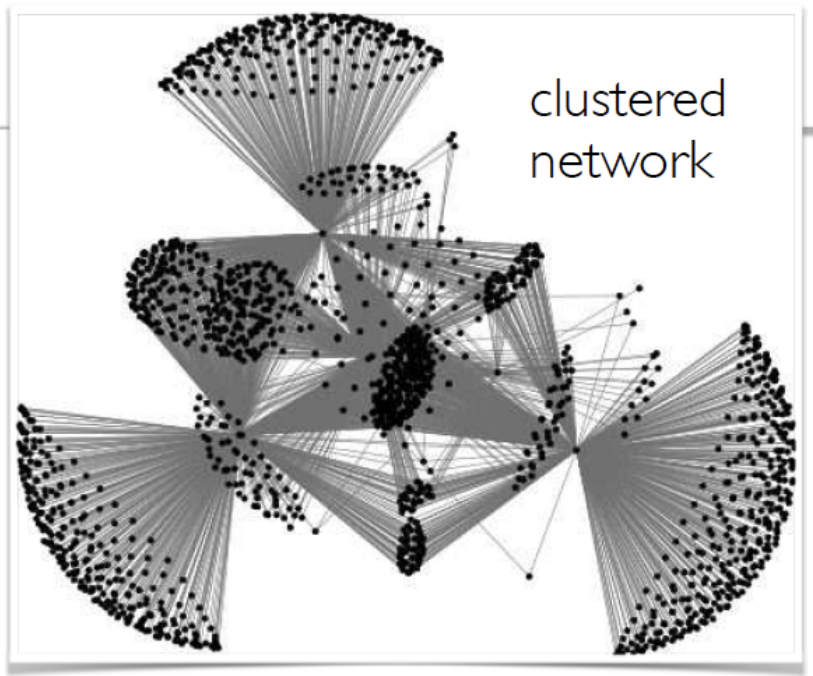
information propagates *quickly within* a community, but *slowly between* communities

Granovetter, *Amer. J. Sociology* **78**, 1360-1380 (1973)
Granovetter, *Sociological Theory* **1**, 201-233 (1983)

Rank	Size	Description
1	114538	General interest: politics; art/literature; general fiction; human nature; technical books; how things, people, computers, societies work, etc.
2	92276	The arts: videos, books, DVDs about the creative and performing arts
3	78661	Hobbies and interests I: self-help; self-education; popular science fiction, popular fantasy; leisure; etc.
4	54582	Hobbies and interests II: adventure books; video games/comics; some sports; some humor; some classic fiction; some western religious material; etc.
5	9872	classical music and related items
6	1904	children's videos, movies, music and books
7	1493	church/religious music; African-descent cultural books; homoerotic imagery
8	1101	pop horror; mystery/adventure fiction
9	1083	jazz; orchestral music; easy listening
10	947	engineering; practical fashion

purchases = interests

interests = clustered

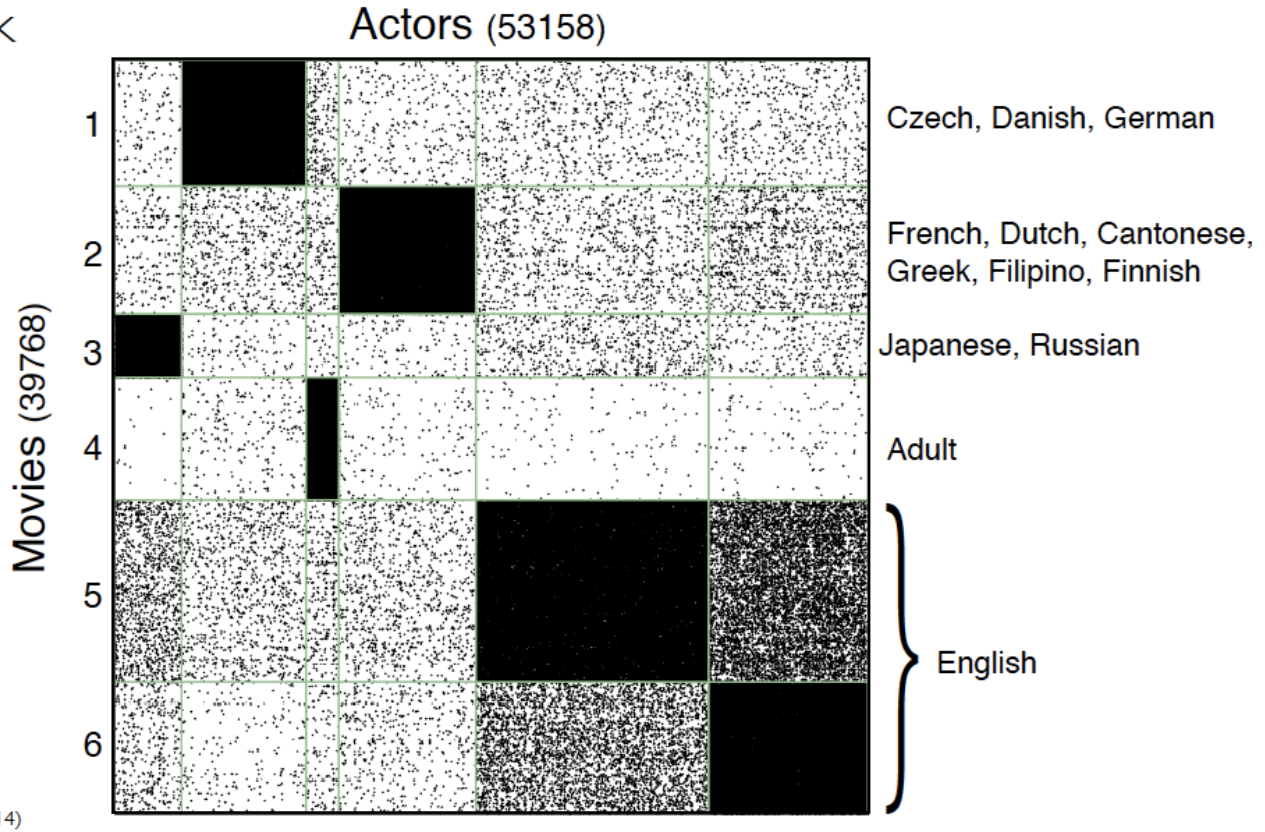


Clauset et al., *Physical Review E* **70**, 066111 (2004)

Efficiently inferring community structure in bipartite networks

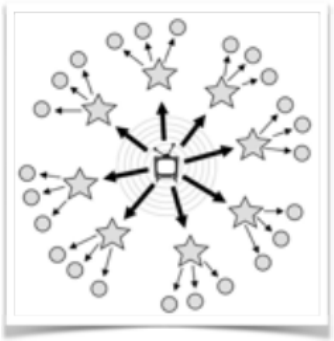
Daniel B. Larremore,[1,2] Aaron Clauset,[3,4,5] and Abigail Z. Jacobs[3] 2014

IMDB actor-movie bipartite network

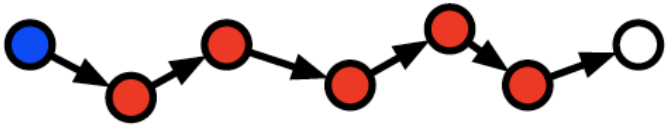


Larremore et al., *Physical Review E* **90**, 102805 (2014)

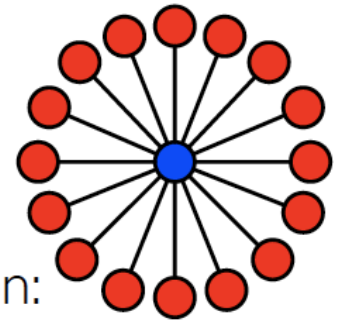
- *community = vertices with same pattern of inter-community connections*
- network macro-structure
- finding them like "network clustering"
- allow us to *coarse grain* system structure
  [decompose heterogeneous structure into homogeneous blocks]
- constrains network synchronization, information flows, diffusion, influence



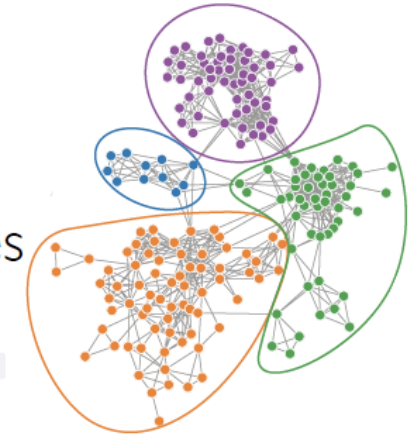
- paths:
  'distance' across the network



- node degree:
  individual connectivity

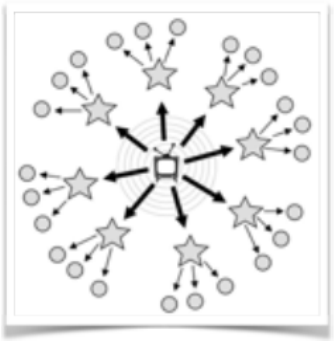


- position:
  centrality and status

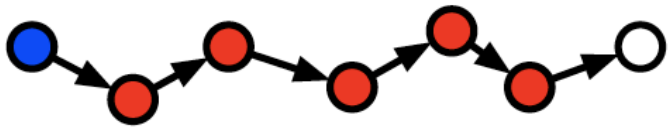


- communities
  and groups

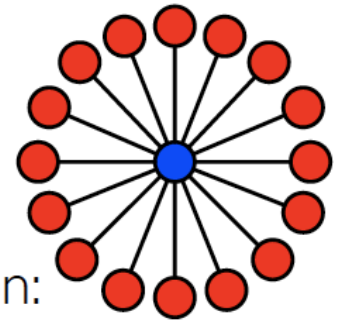
- many others...



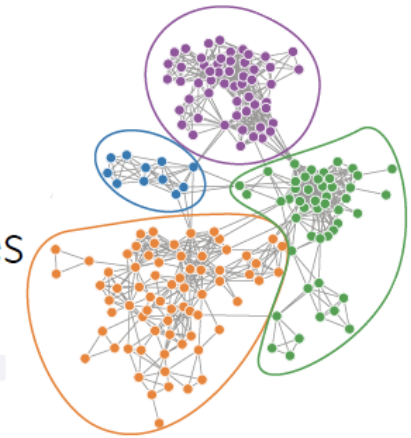
- paths:
  'distance' across the network



- node degree:
  individual connectivity



- position:
  centrality and status



- communities
  and groups

- many others...

Books:

- Newman: Sections 2-5

Other 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