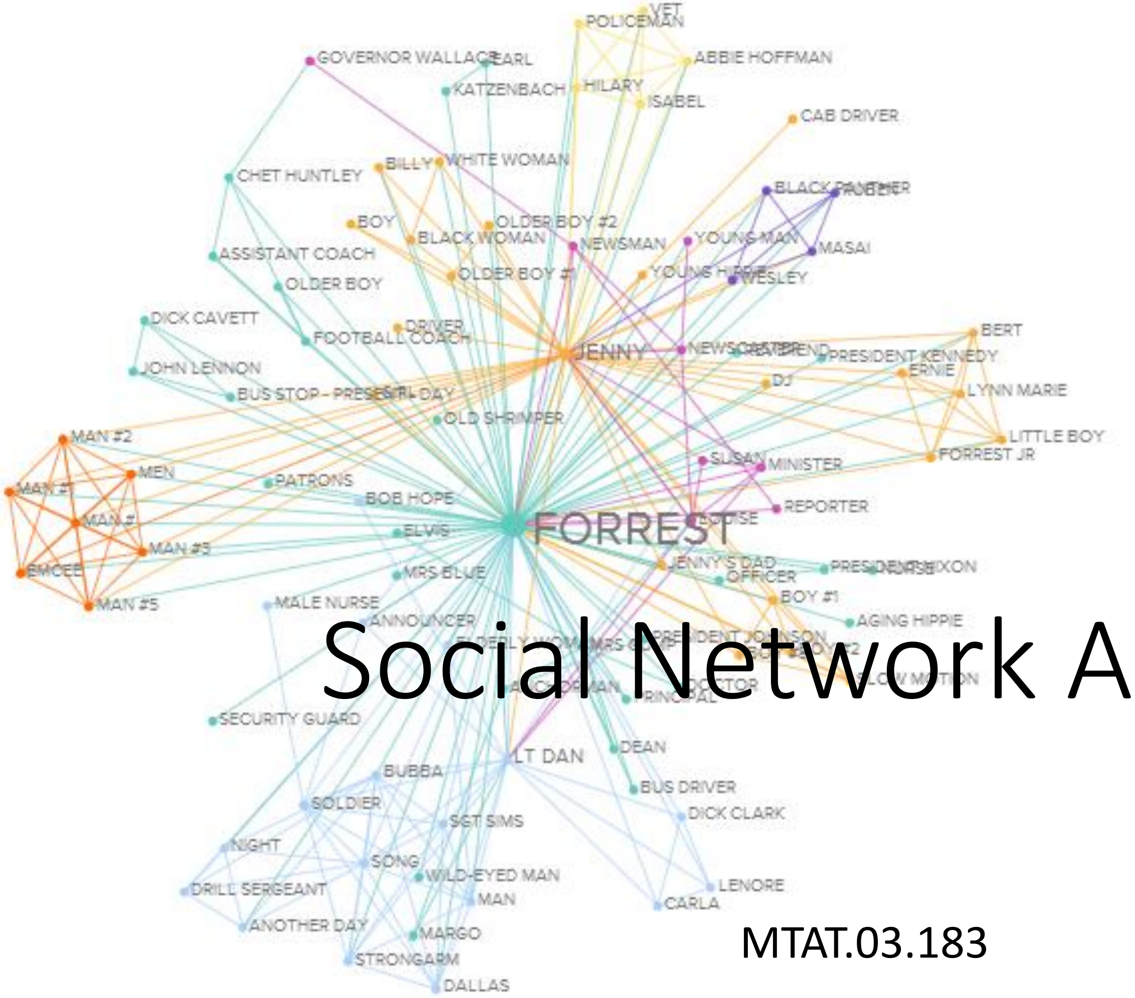
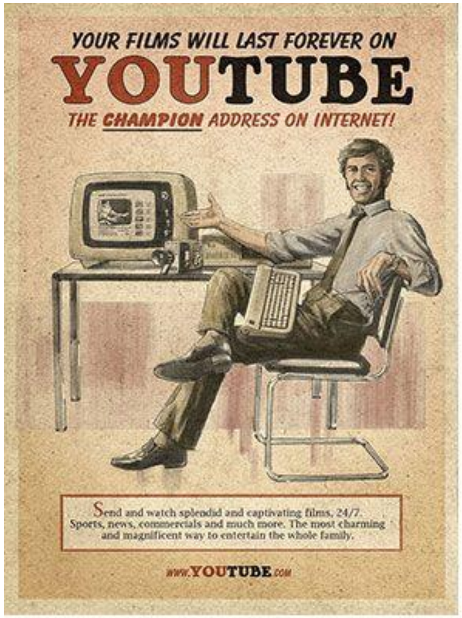


Social Network Analysis

MTAT.03.183

Anna Leontjeva



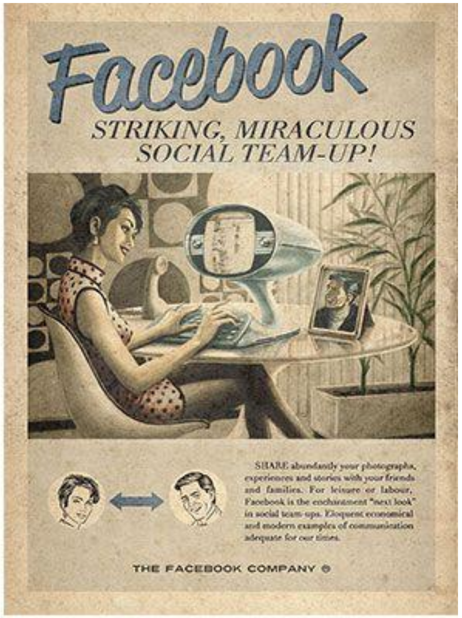
YOUR FILMS WILL LAST FOREVER ON

YOUTUBE

THE CHAMPION ADDRESS ON INTERNET!

Send and watch splendid and captivating films, 24/7.
Sports, news, commercials and much more. The most charming
and magnificent way to entertain the whole family.

www.YOUTUBE.com

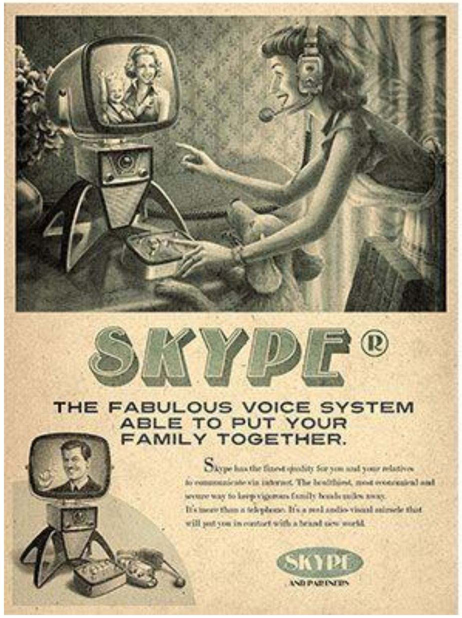


Facebook

STRIKING, MIRACULOUS
SOCIAL TEAM-UP!

SHARE abundantly your photographs, experiences and stories with your friends and families. For leisure or labour, Facebook is the enchantment "next look" in social team-ups. Eloquent economical and modern examples of communication adequate for our times.

THE FACEBOOK COMPANY ®

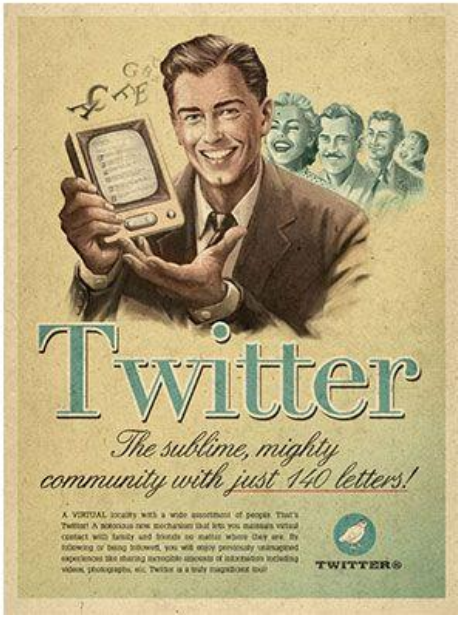


SKYPE®

THE FABULOUS VOICE SYSTEM
ABLE TO PUT YOUR
FAMILY TOGETHER.

Skype has the finest quality for you and your relatives to communicate via internet. The healthiest, most economical and secure way to keep vigorous family bonds miles away.
It's more than a telephone. It's a real audio-visual miracle that will put you in contact with a brand new world.

SKYPE
AND PARTNERS



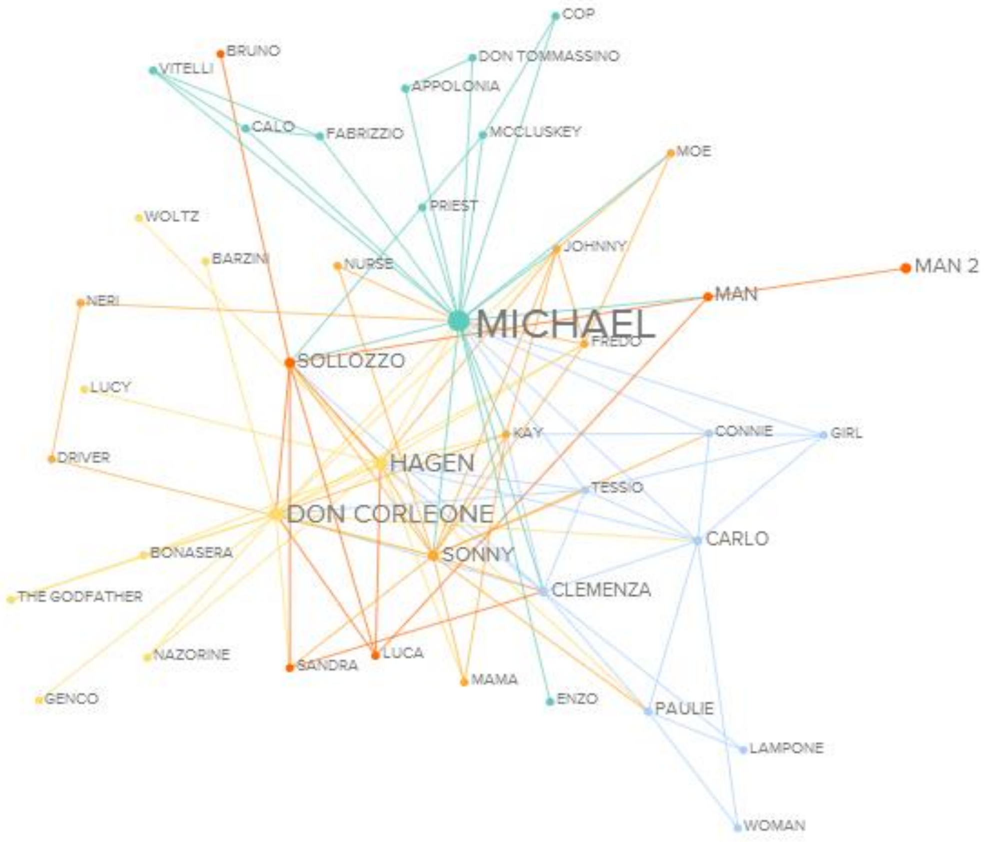
Twitter

The sublime, mighty
community with just 140 letters!

A VIRTUAL locality with a wide assortment of people. That's Twitter! A gorgeous new mechanism that lets you maintain virtual contact with family and friends no matter where they are. By following or being followed, you will enjoy previously unimagined experiences like sharing incredible amounts of information including videos, photographs, etc. Twitter is a truly magnificent tool!

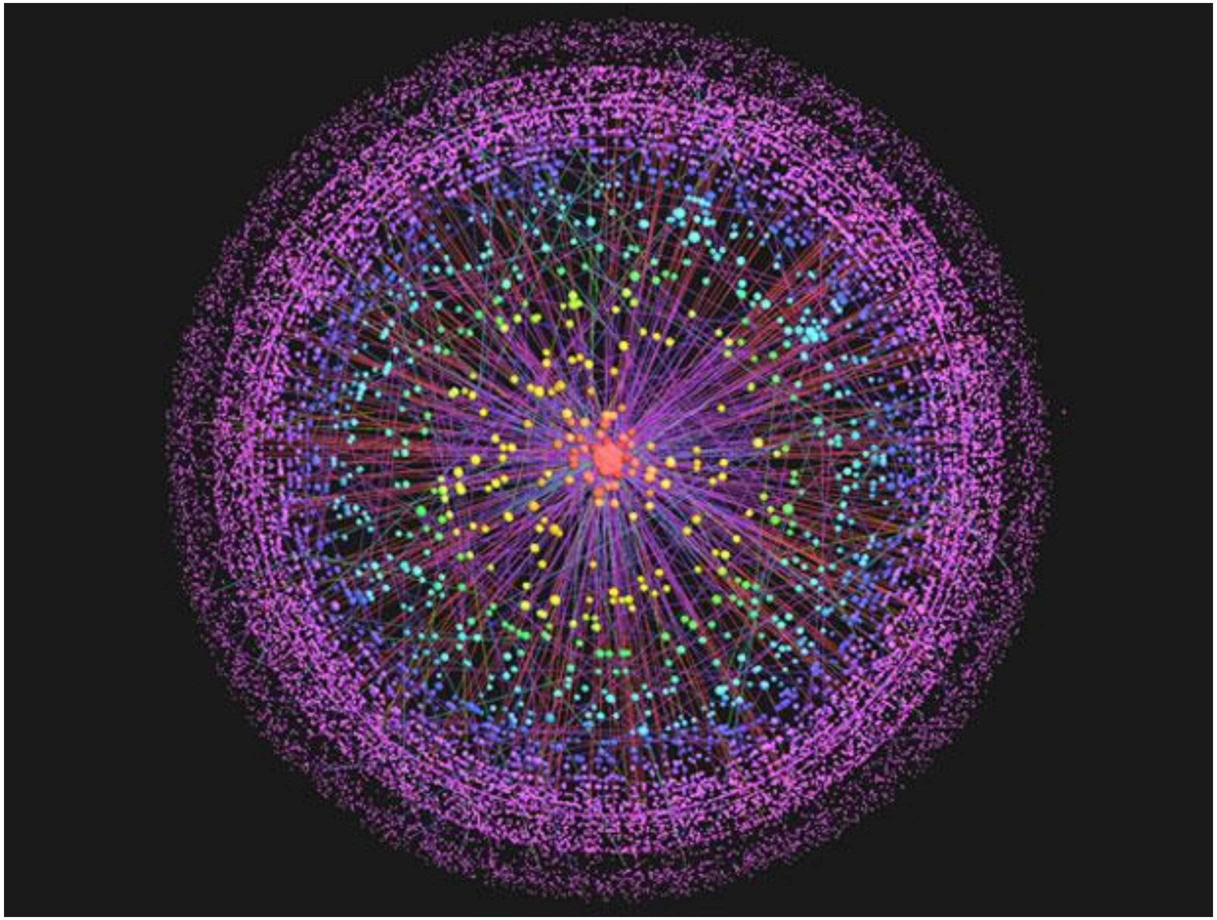
TWITTER®

By ad agency Moma Propaganda

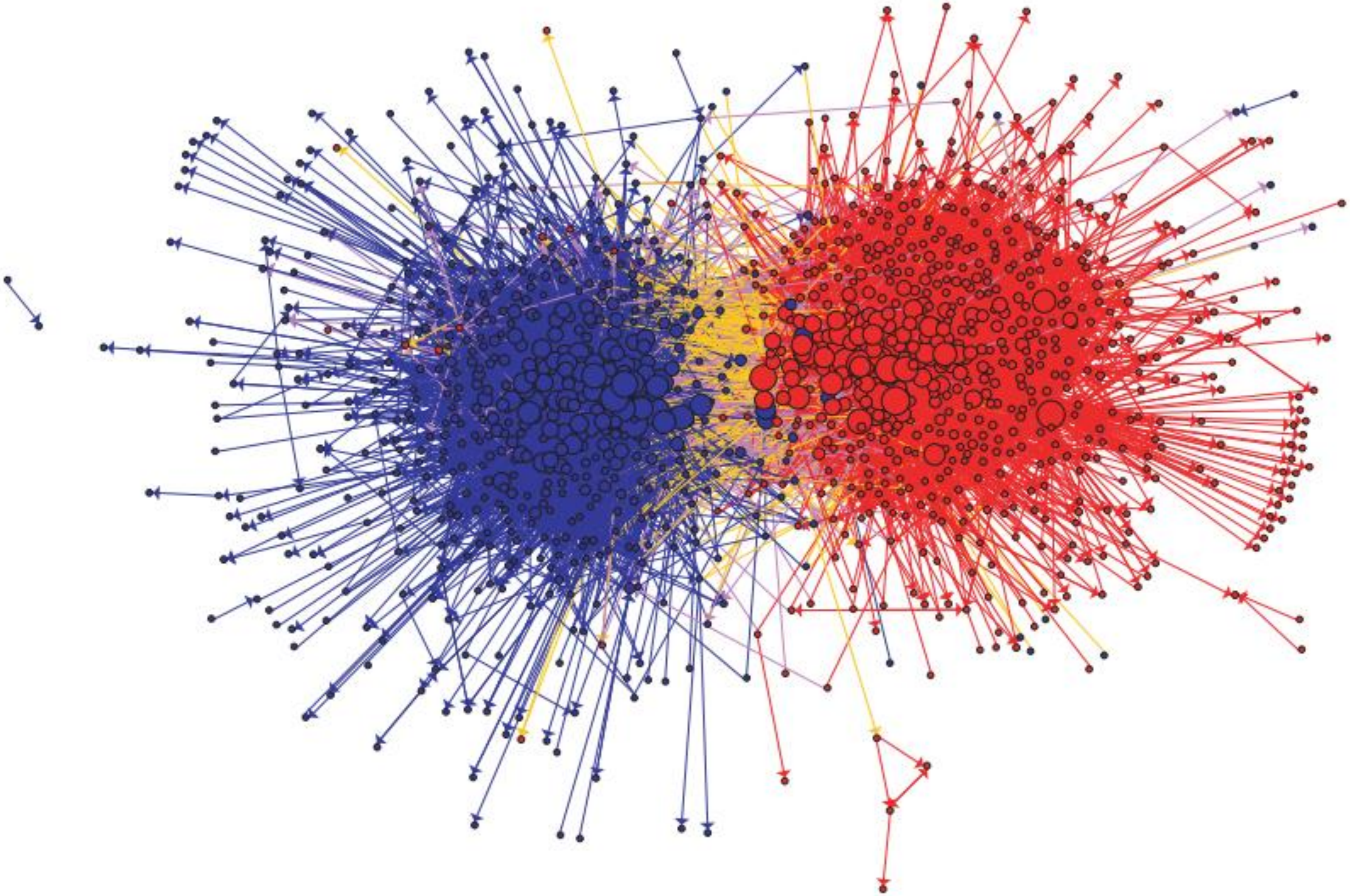


- http://moviegalaxies.com
- http://www.liveplasma.com/
- http://www.webdesignerdepot.com/2009/06/50-great-examples-of-data-visualization/

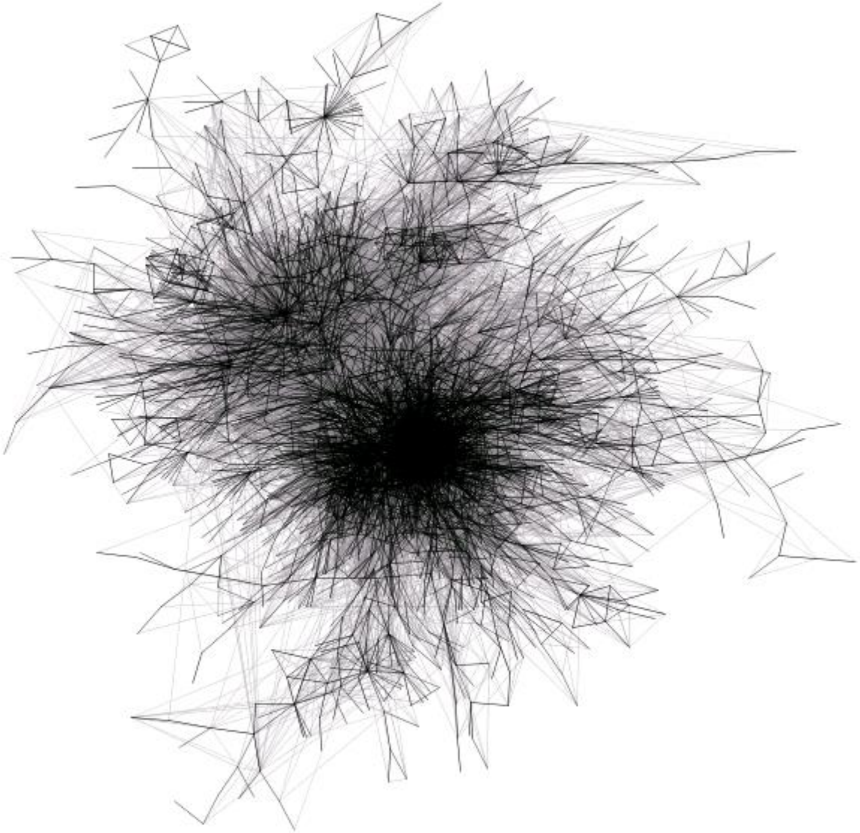
Internet

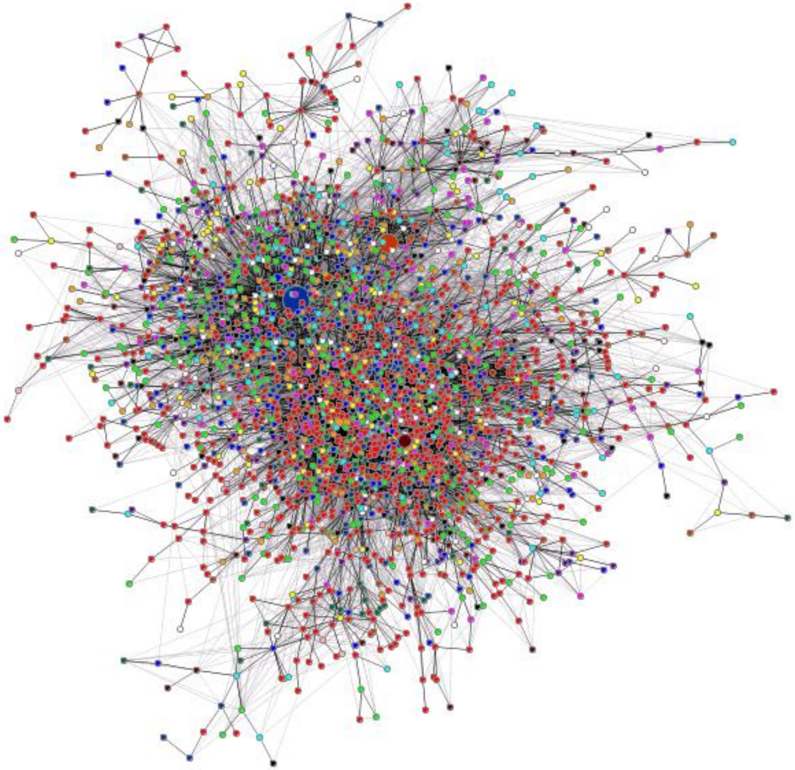


Political blogs



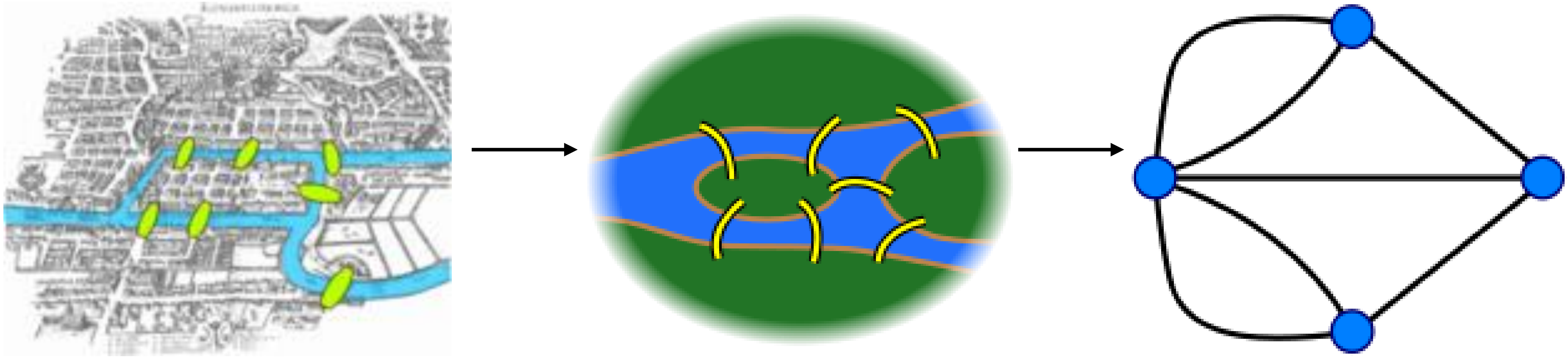
Why to study?





Background

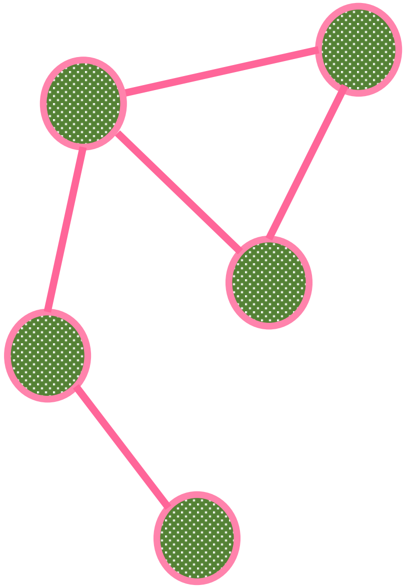
Königsberg bridge problem formulated in 1735 is considered to be the foundation of graph theory



“Can one walk across the seven bridges and never cross the same one twice?”

Definitions

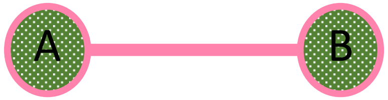
Network (graph)



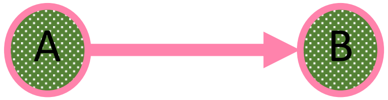
● nodes (vertices)

/ edges (arcs)

Edges

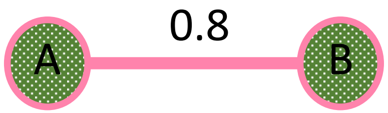


Undirected:
- A and B are friends
- A and B are family members
- A and B are students working on the same project



Directed:
- User A sent request of a friendship to user B
- A is parent of B
- Student A supervises student B

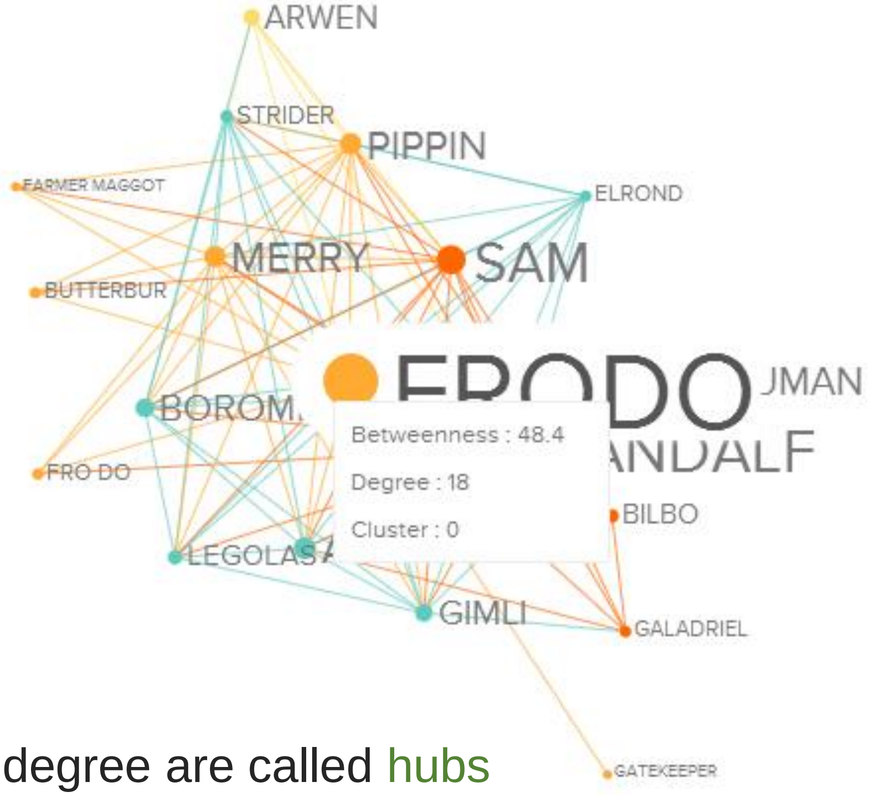
Edge attributes



- Weight (frequency of communication)
- Type (which member of family, position at company)
- Measures (degree, PageRank, centrality)

Degree

- the number of connections a node has to other nodes



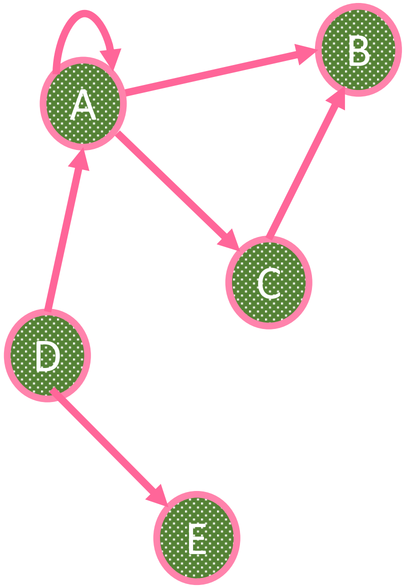
Betweenness : 48.4

Degree : 18

Cluster : 0

Nodes that have large degree are called hubs

Nodes



Indegree: number of directed edges incident on a node

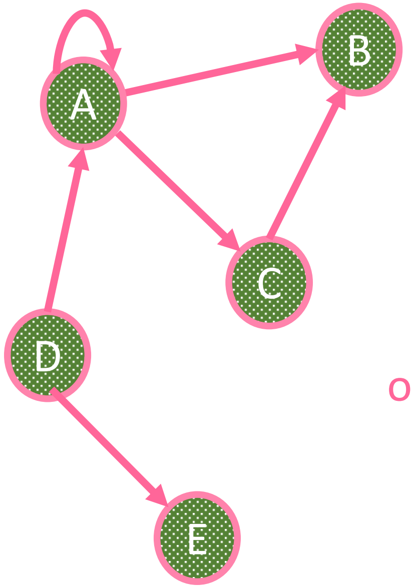
B indegree is 2

Outdegree: number of directed edges point out from a node

C outdegree is 1

Data representation

- Adjacency matrix

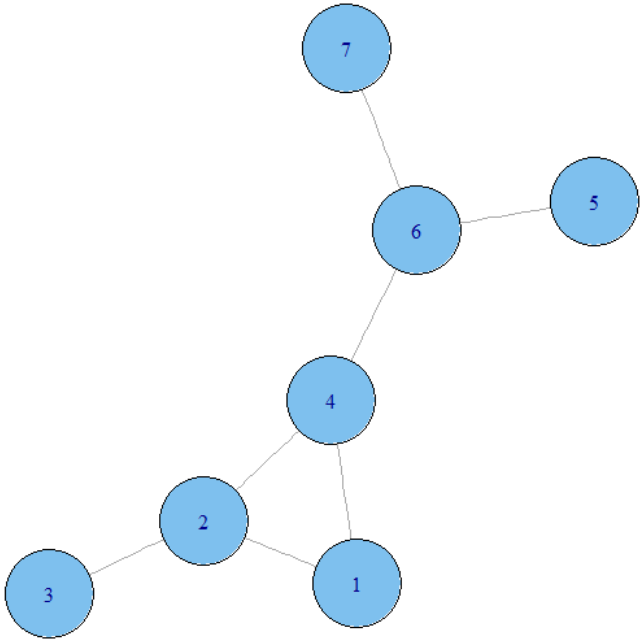


indegree

outdegree

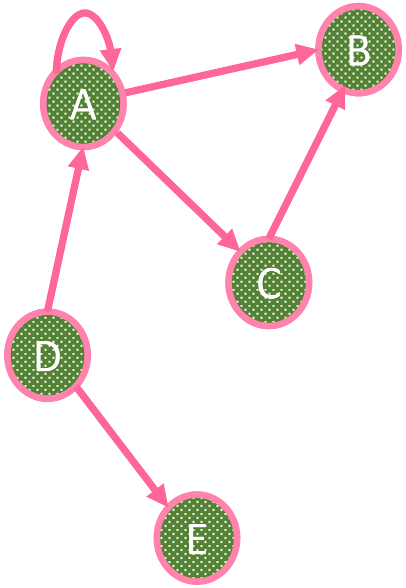
	A	B	C	D	E
A	1	1	1	0	0
B	0	0	0	0	0
C	0	1	0	0	0
D	1	0	0	0	1
E	0	0	0	0	0

Construct adjacency matrix for the following graph:



Data representation

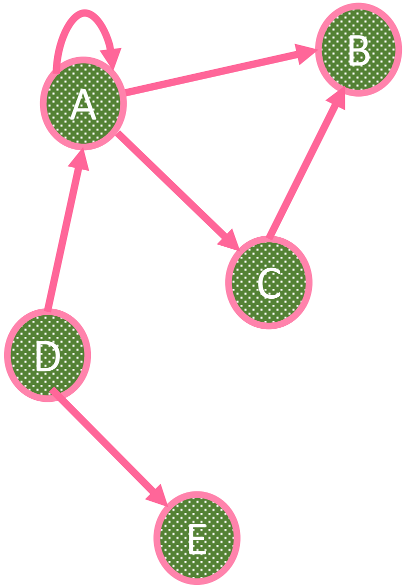
- Edge list



- A,A
- A,B
- A,C
- C,B
- D,A
- D,E

Data representation

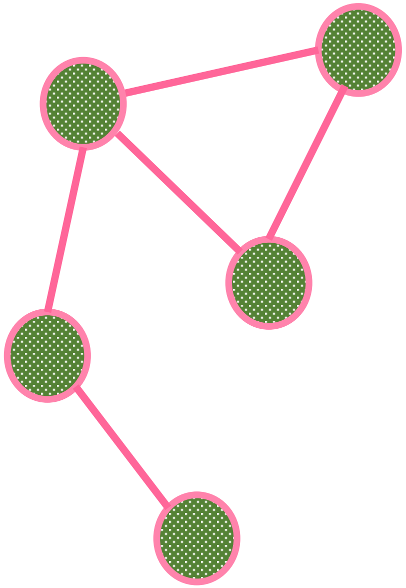
- Adjacency list



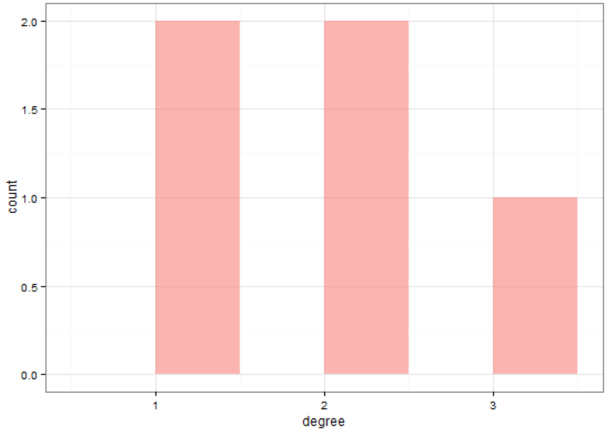
- A: A,B,C
- B:
- C: B
- D: A, E
- E:

Degree distribution

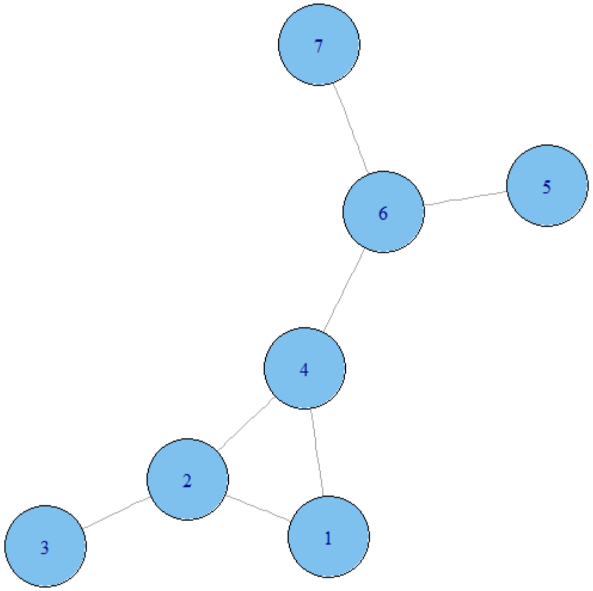
- a frequency count of the occurance of each degree



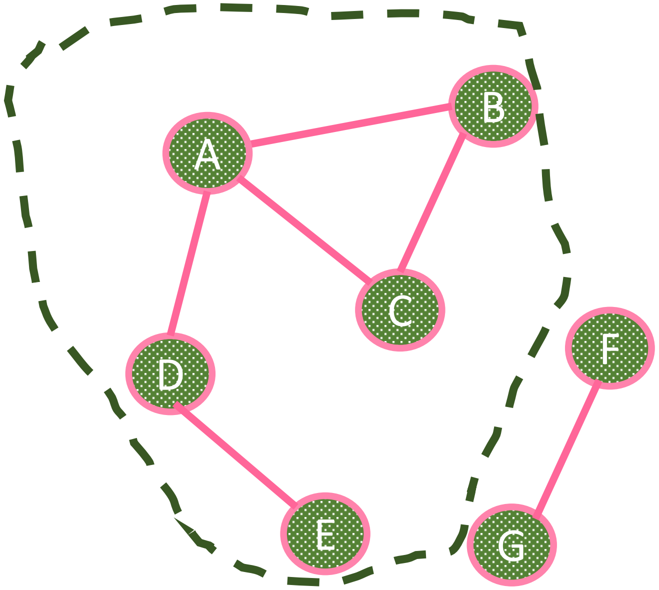
3, 2, 2, 1, 1



What is the degree distribution for the same graph?

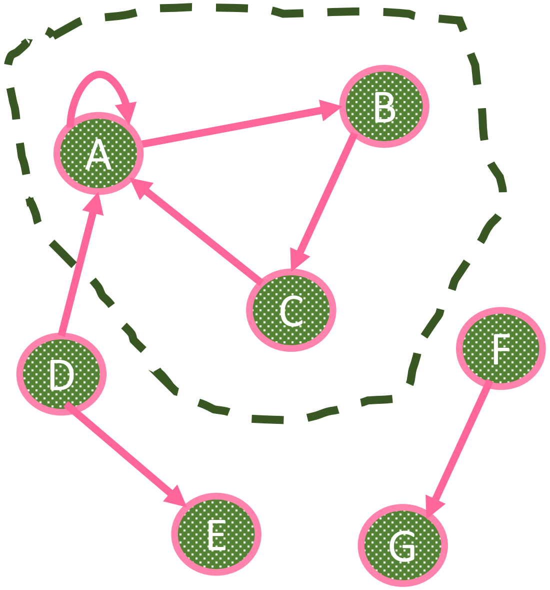


Connected components



- A connected component is a maximal connected subgraph of undirected graph *G.*

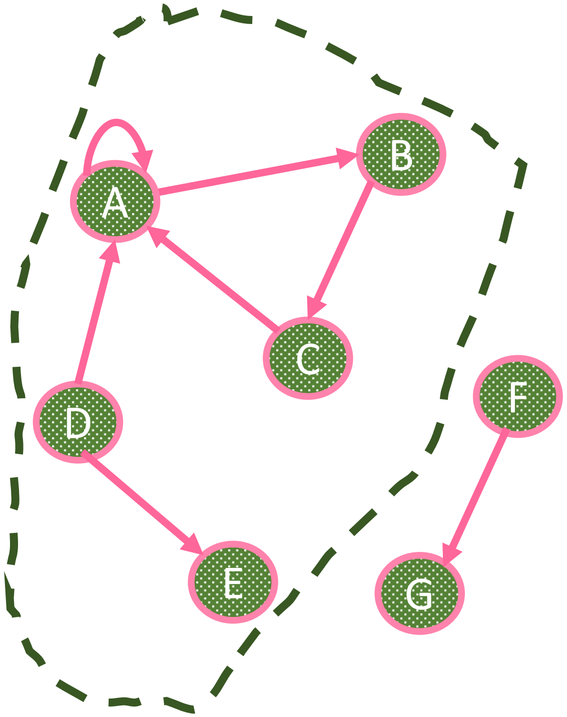
Connected components



- A connected component is a maximal connected subgraph of undirected graph *G.*

- A strongly connected component: each node is reachable from every other node through directed edges

Connected components

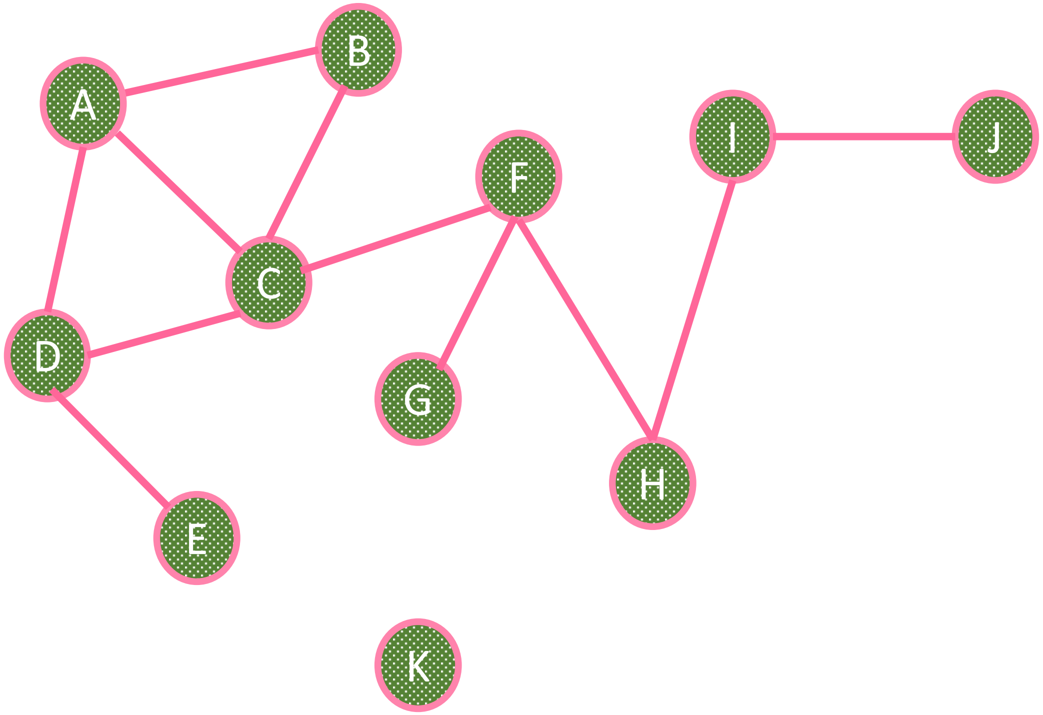


- A connected component is a maximal connected subgraph of undirected graph *G.*

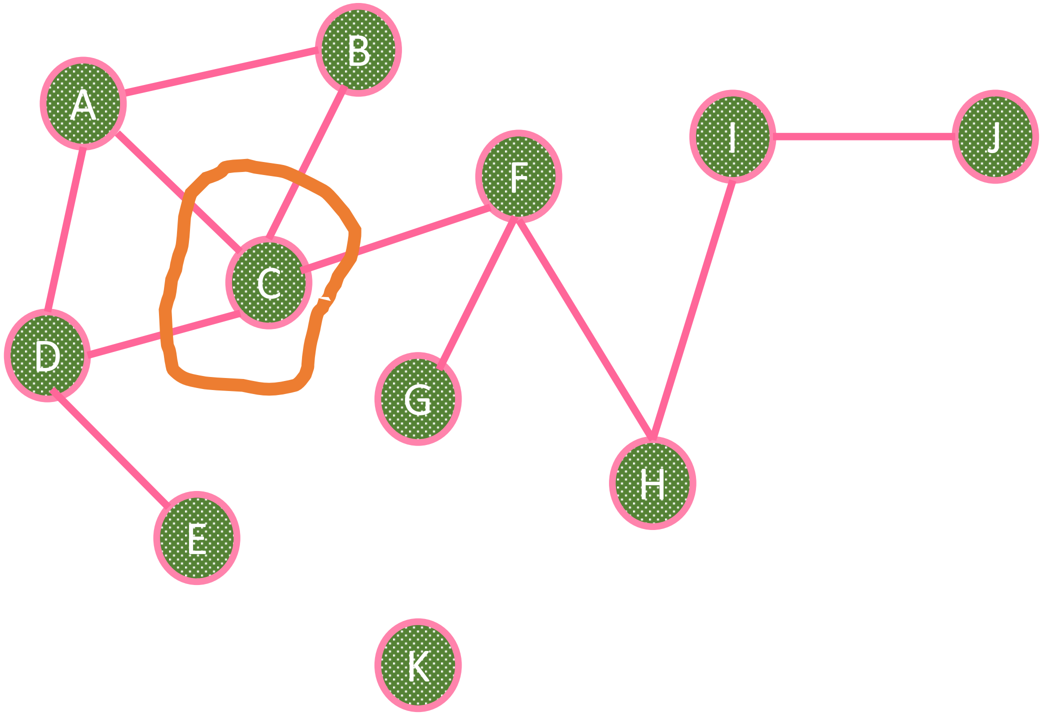
- A strongly connected component: each node is reachable from every other node through directed edges

- A weakly connected component each node is reachable from every other node through undirected edges

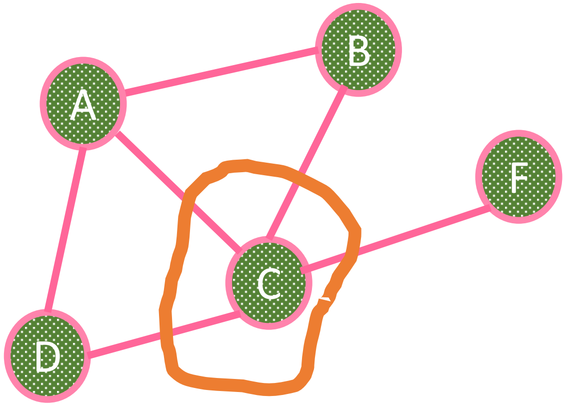
Ego network



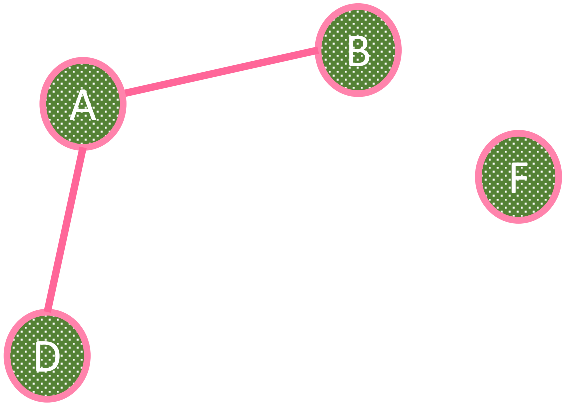
Ego network



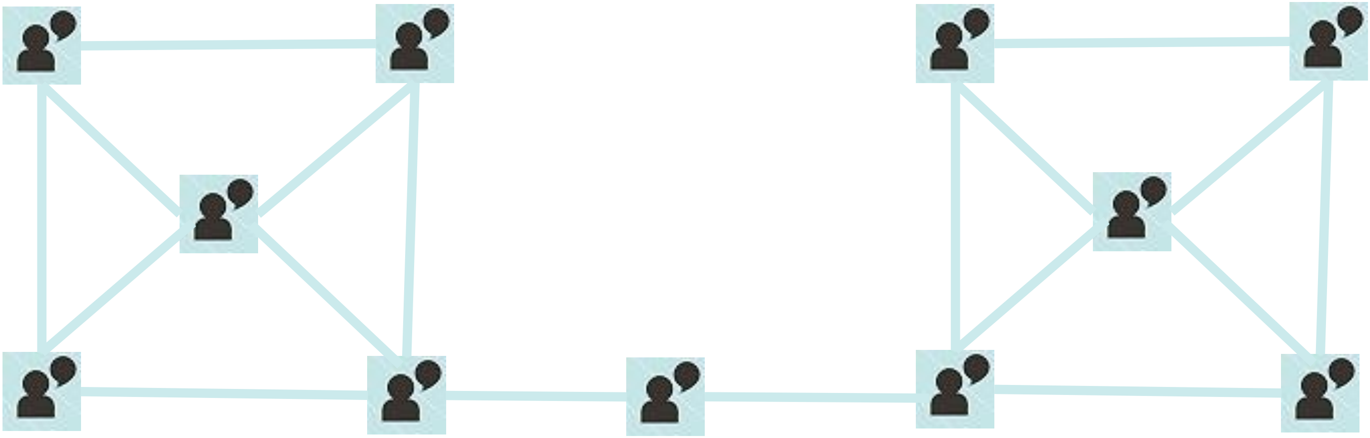
Ego network



Ego network minus ego

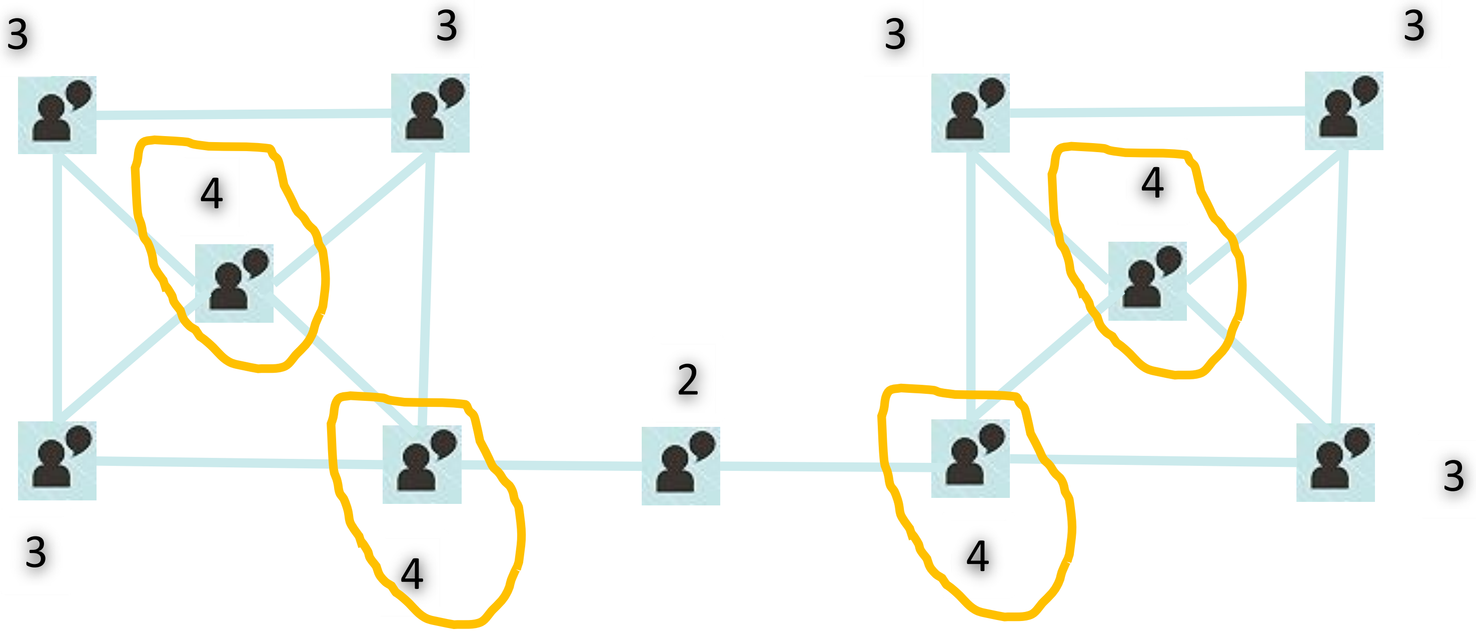


Centrality measures



Who is the most important?

Degree centrality



Nodes with higher degree centrality are "popular" nodes, can reach more nodes directly.

Path and diameter

- A path between nodes is a sequence of edges that connect a sequence of non-repeating nodes.
- A shortest path is the path that connects pair of nodes via smallest number of edges.
- An average shortest path for graph G is an average number of steps along the shortest paths for all possible pairs of network nodes
- A diameter is the longest shortest path

Betweenness centrality

- quantifies the number of times a node acts as a bridge along the shortest path between two other nodes.

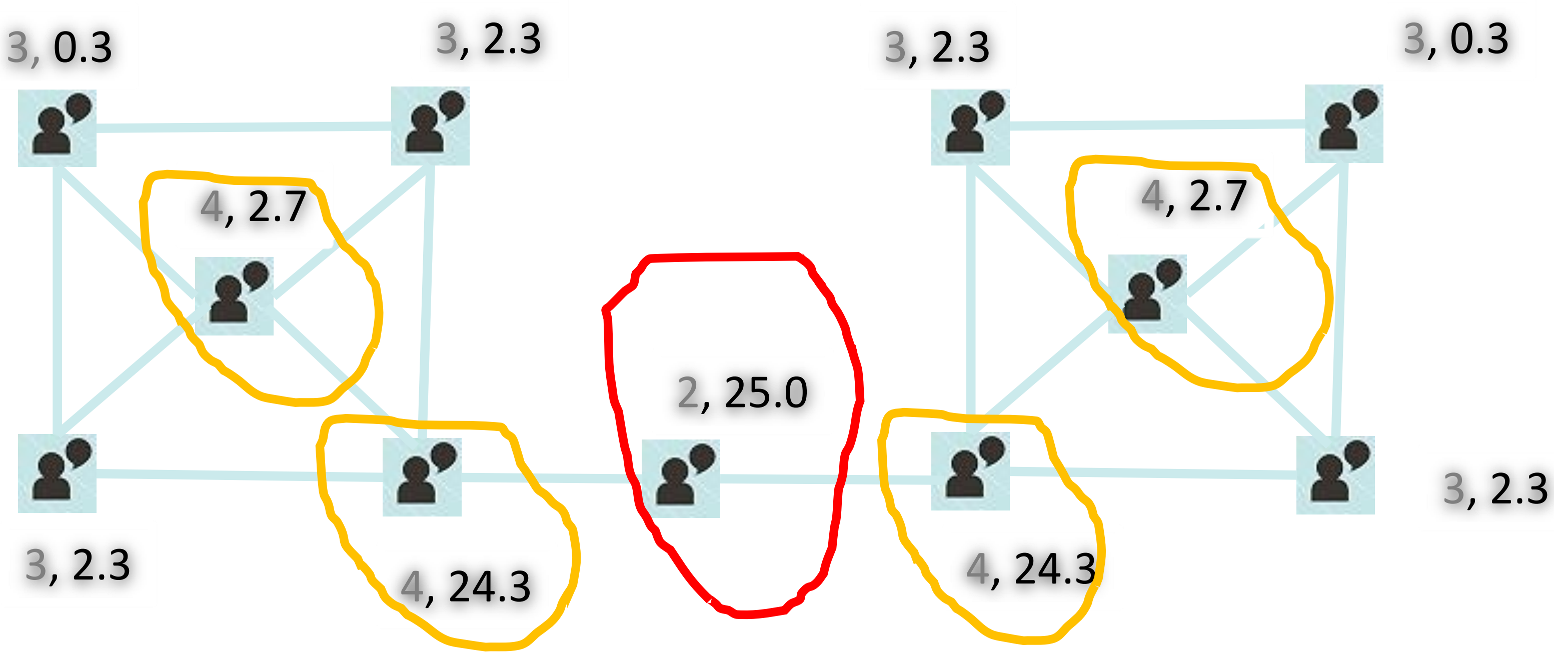
The betweenness for a node v computed as follows:

1. For each pair of nodes (k, l) compute the shortest paths between them.

2. For each pair of nodes (k, l) calculate the fraction of shortest paths that pass through the node v

3. Sum the fraction over all pair of vertices (k, l)

Betweenness centrality



Nodes with higher betweenness centrality are „bottlenecks“.

Closeness centrality

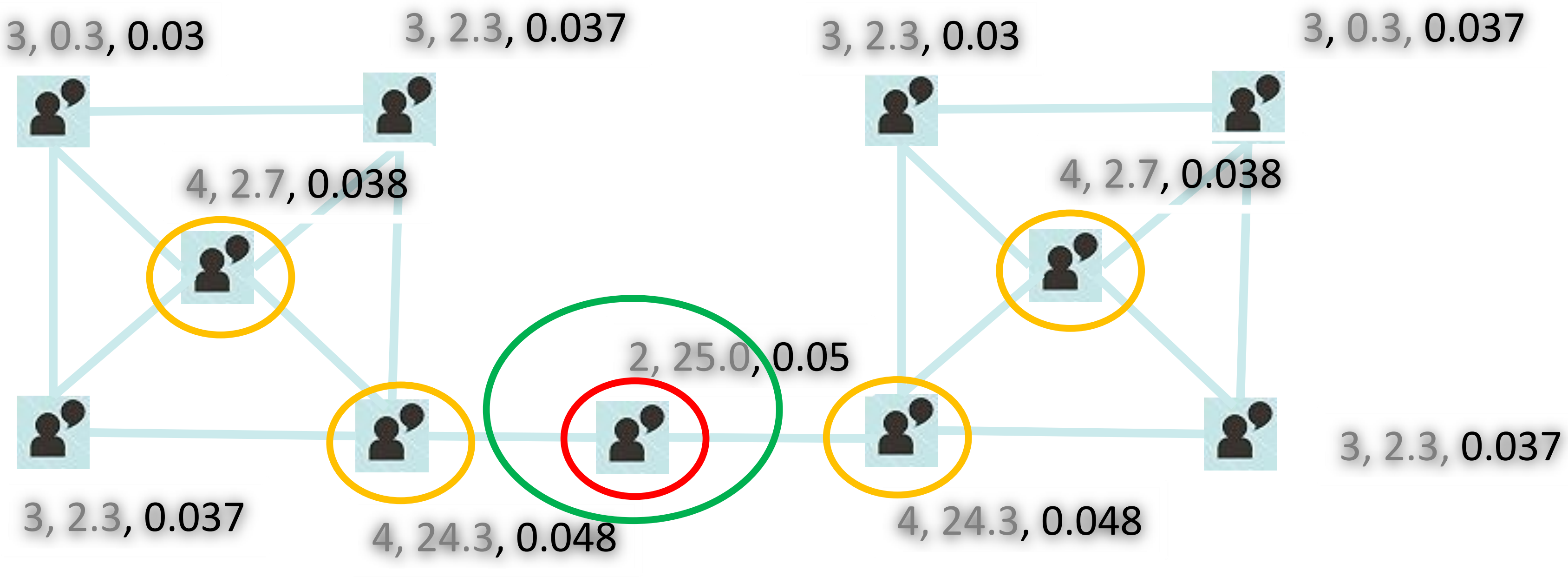
- measures „the speed“ with which information can reach other nodes from a starting node

The closeness for a node v computed as follows:

- Calculate the mean length of all shortest path from the node v to all other nodes in the network

- Take the reciprocal of the calculated value.

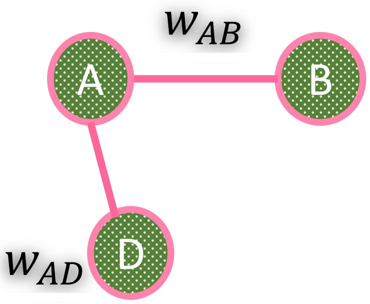
Closeness centrality



Eigenvector centrality

- measures the influence of a node in a graph. Google's PageRank is a variant of this measure.

$$C(A) = w_{AD} \times C(B) + w_{AB} \times C(D)$$

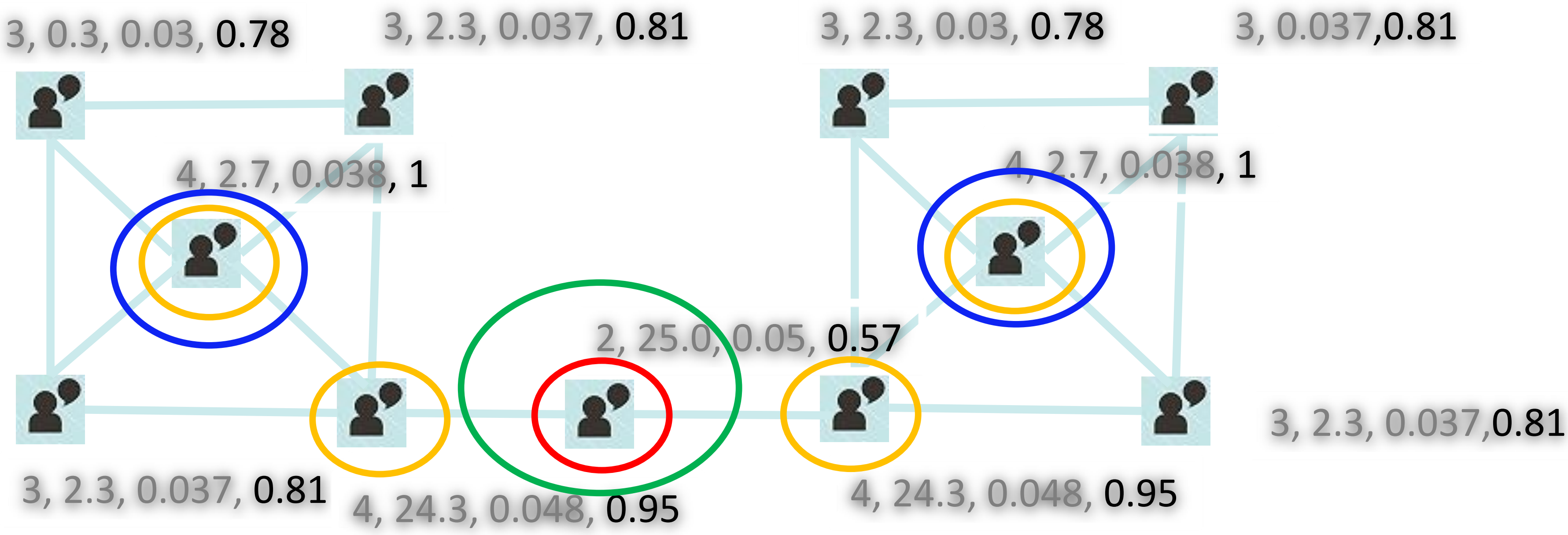


$$C(\alpha, \beta) = \alpha(I - \beta R)^{-1} R1$$

- α is a scaling vector
- β weight of how important the centrality of neighbors
- R is the adjacency matrix
- I is identity matrix
- 1 is a matrix of ones

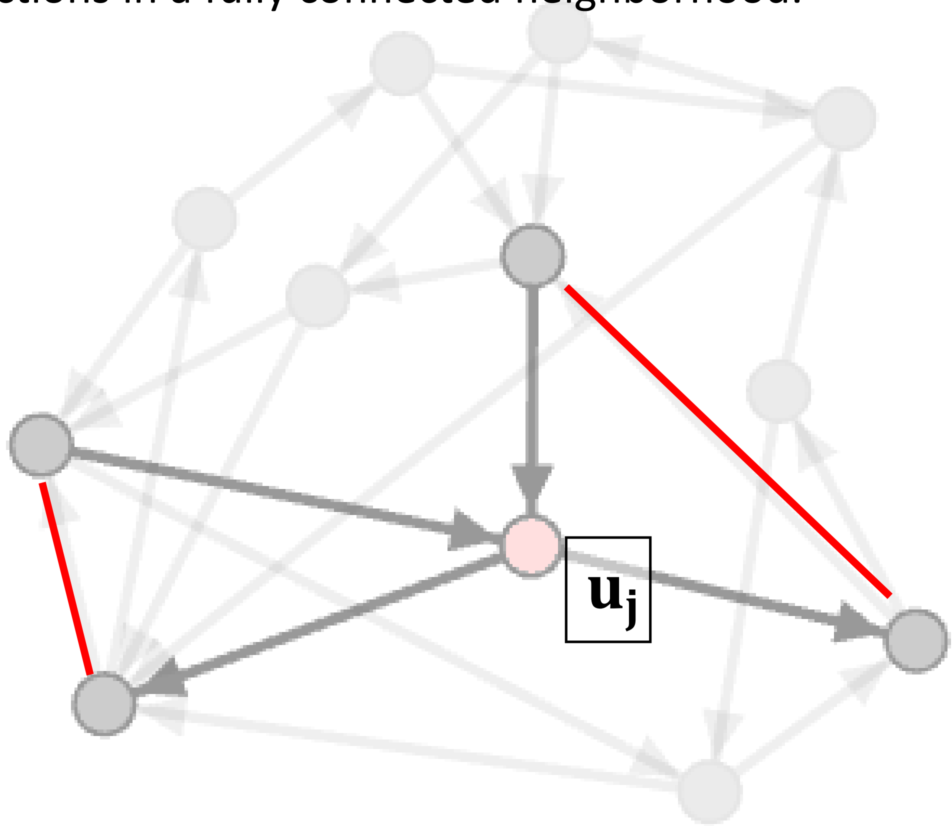
How central you are depends on how central your neighbors are.

Eigenvector centrality



Clustering coefficient

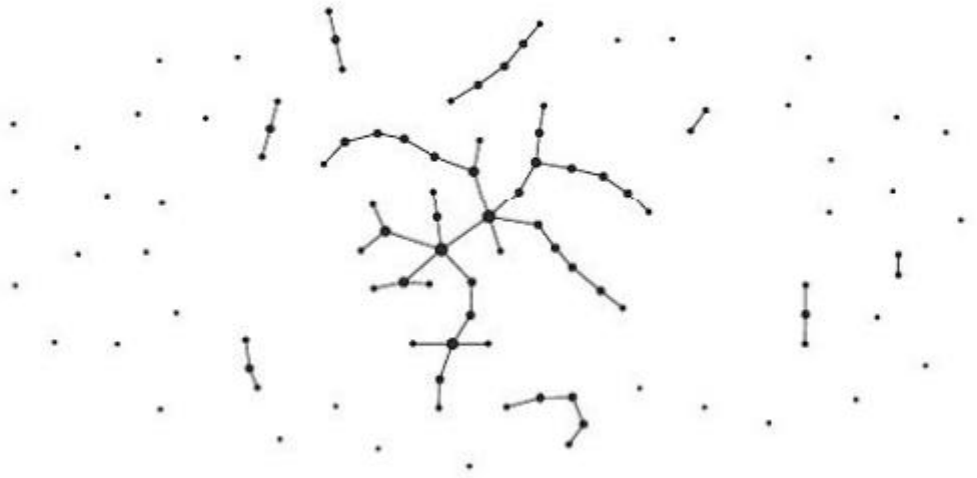
The ratio of the number of connections in the neighborhood of a user to the number of connections in a fully connected neighborhood.



Friend of my friend is my friend

Random graphs and its properties

- Random graphs are described by a random process that generates them.
- Provide the easiest model for any network (simple underlying assumptions)



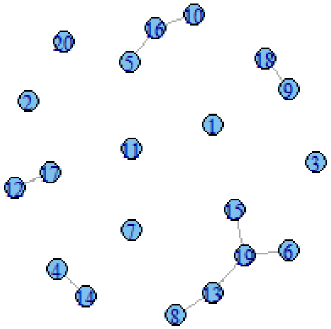
Erdős–Rényi model (1959)

Assumptions:

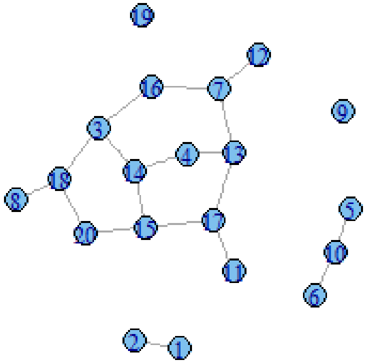
- nodes connect at random
- edges are undirected

Parameters for the model:

- probability of sharing an edge between two nodes (p) or
- total number of edges in the graph (M)

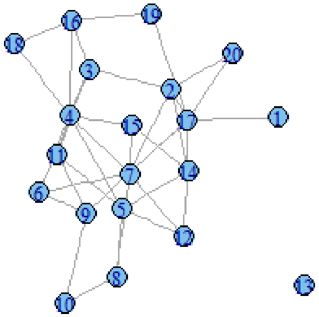


p = 1/20

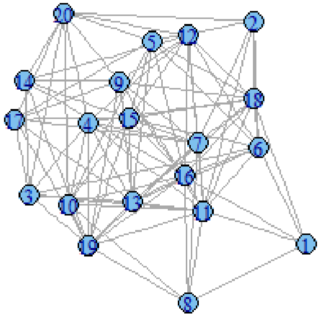


p = 1/10

p = 1/5



p = 1/2



How do we calculate the probability that a node has a degree d?

What is the probability that in the graph with 8 nodes and p = 0.5 a degree of a given node will be 2.

How do we calculate the probability that a node has a degree d?

The distribution of the degree of any particular node is binomial:

$$P(\deg(v) = k) = \binom{n-1}{k} p^k (1-p)^{n-1-k}$$

Average degree is then: $(n\text{-}1)p$

As the size of the network increases and probability p is the same, what happens to the average degree?

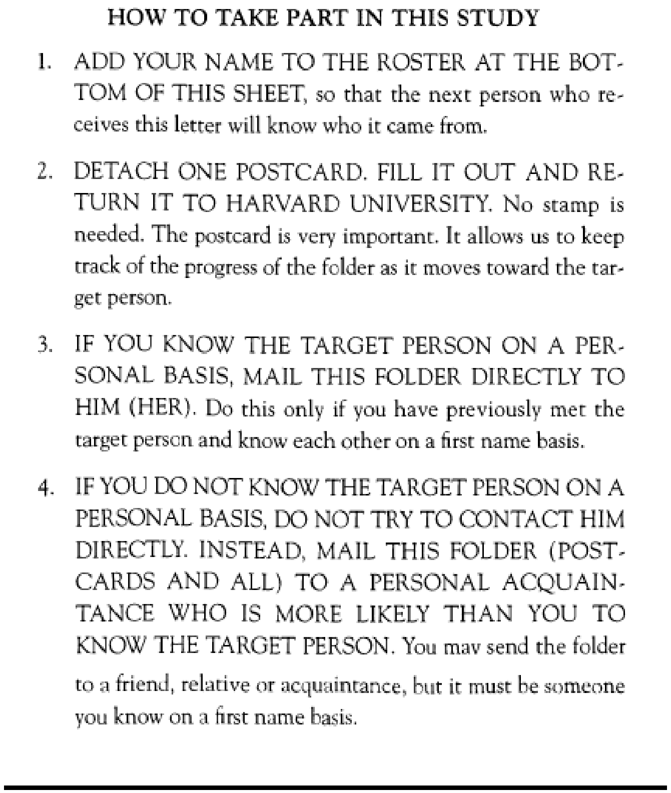
Moving towards more realistic networks

Social networks have properties that ER do not account for -> "similar" nodes are closer.

Milgram's experiment (1967)

20% of initiated chains reached the target
Average chain length = 6.5

6 degrees of separation



HOW TO TAKE PART IN THIS STUDY

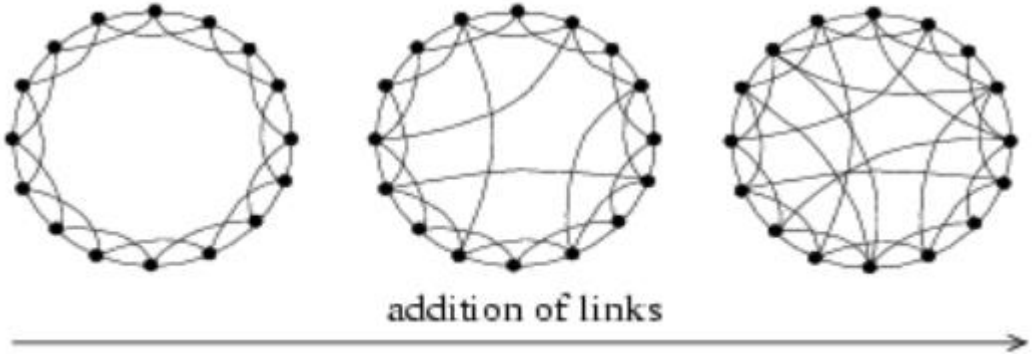
1. ADD YOUR NAME TO THE ROSTER AT THE BOTTOM OF THIS SHEET, so that the next person who receives this letter will know who it came from.
2. DETACH ONE POSTCARD. FILL IT OUT AND RETURN IT TO HARVARD UNIVERSITY. No stamp is needed. The postcard is very important. It allows us to keep track of the progress 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 mav send the folder to a friend, relative or acquaintance, but it must be someone you know on a first name basis.

Watts-Strogatz model (1998)

- Small-world effect: most real-world networks have a high clustering coefficient, but low average path length.

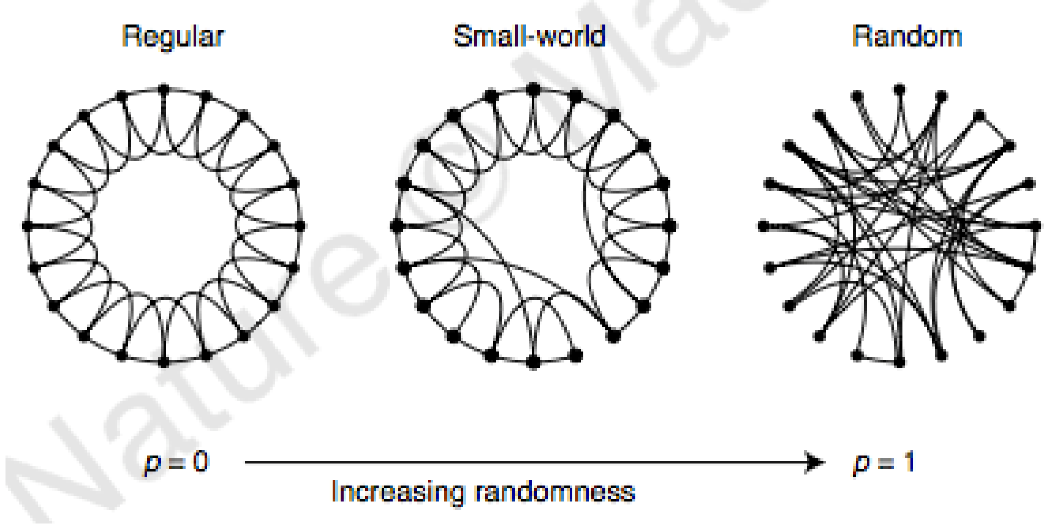
The generative Watt and Strogatz model:

- Build a ring lattice of n nodes and connect each node with its k clockwise neighbors on the ring
- Draw a random number between 0 and 1 for each edge
- Rewire each edge with probability p: if the edge‘s random number is smaller than p, keep the source node of the edge fixed, and choose a new target vertex uniformly at random from all other vertices



Add a fraction p of additional edges leaving underlying lattice intact

Watts-Strogatz small world graph

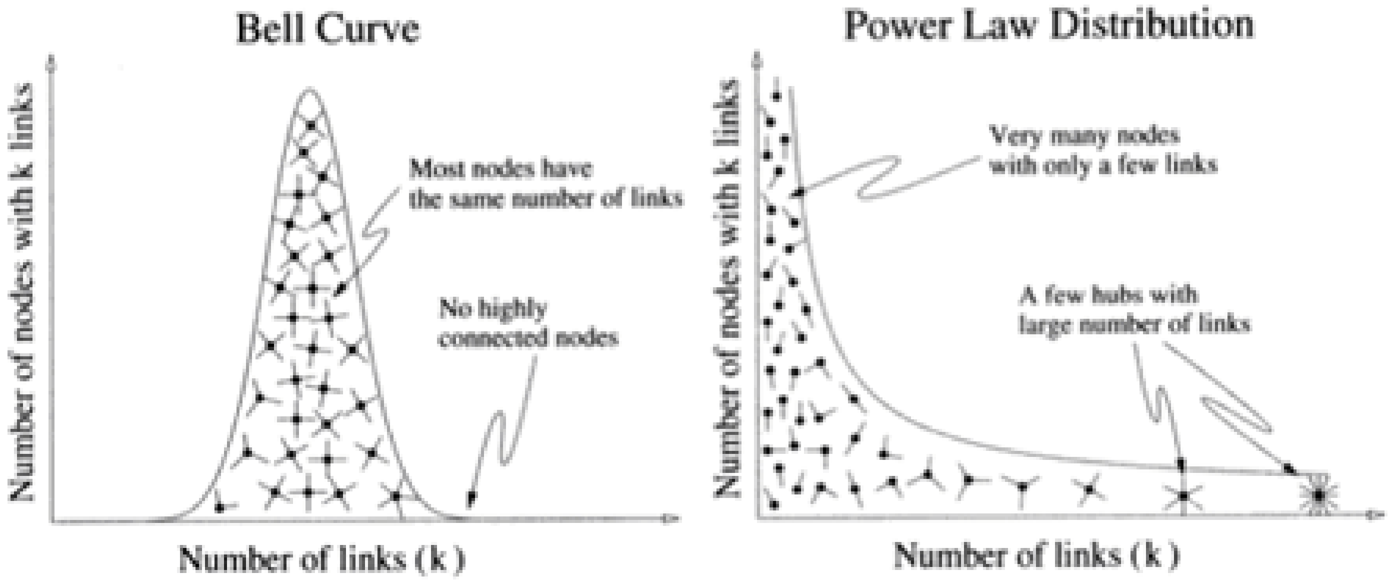


- Each node has k>= nearest neighbors
- You can tune the model by varying the probability p:

  small p: regular lattice

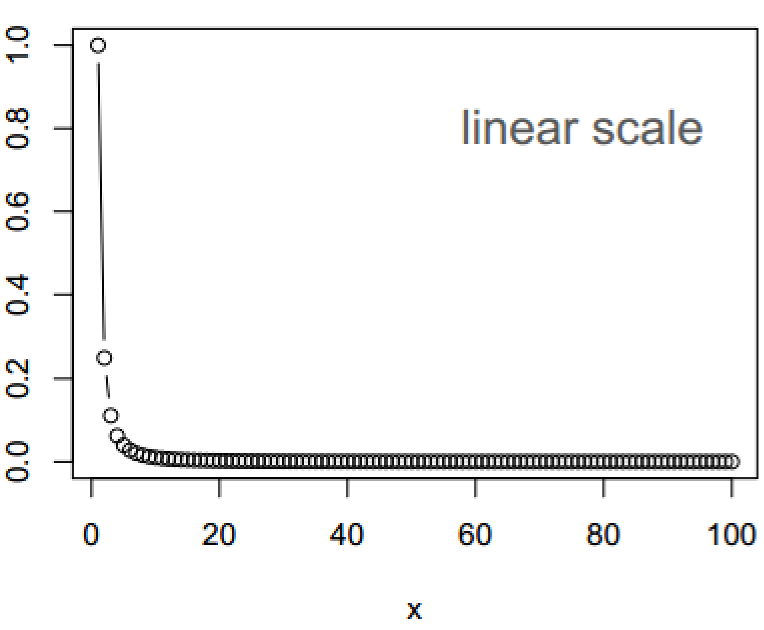
  large p: random graph

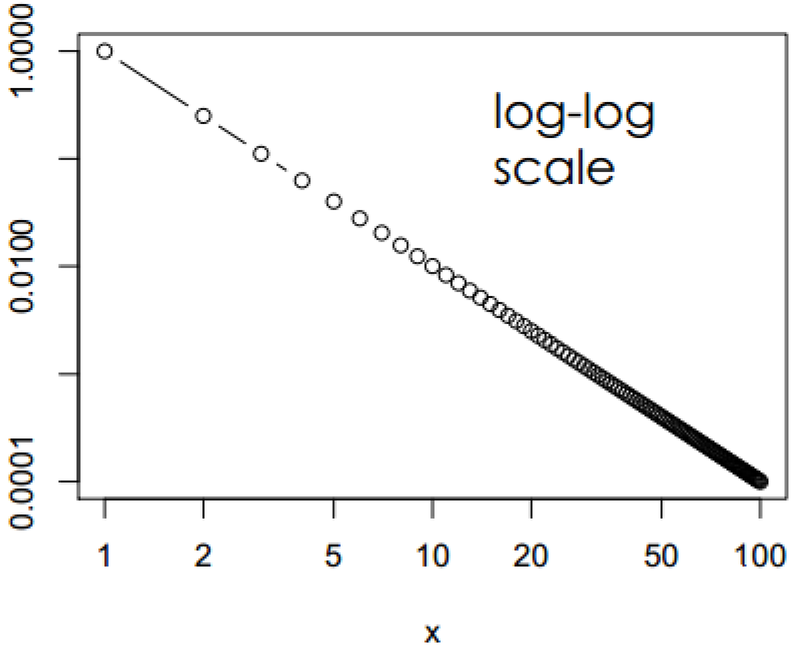
- The small-world network and random graph models assume that the degree in the graph will be normally distributed and do not deviate much.
- It seems that it is not the case for real-world networks.
- In the experiment in 1999 by A-L. Barabasi a small portion of the Web was crawled and it was discovered that the degree distribution follows Power Law.
- Turned out it is the case for many other real-world networks.



http://edgeperspectives.typepad.com/edge_perspectives/2007/05/the_power_of_po.html

Power Law





- Heavy-tailed distribution
- Straight line on a log-log plot
- Scale invariance

Probability observing an object of size ‘x’:
$p(x) = Cx^{-\alpha}$

Straight line on a log-log:

$$\ln\big(p(x)\big) = c - \alpha \ln(x)$$

As the exponent alpha increases, what happens to the downward slope of the line on a log-log plot?

Barabasi-Albert model

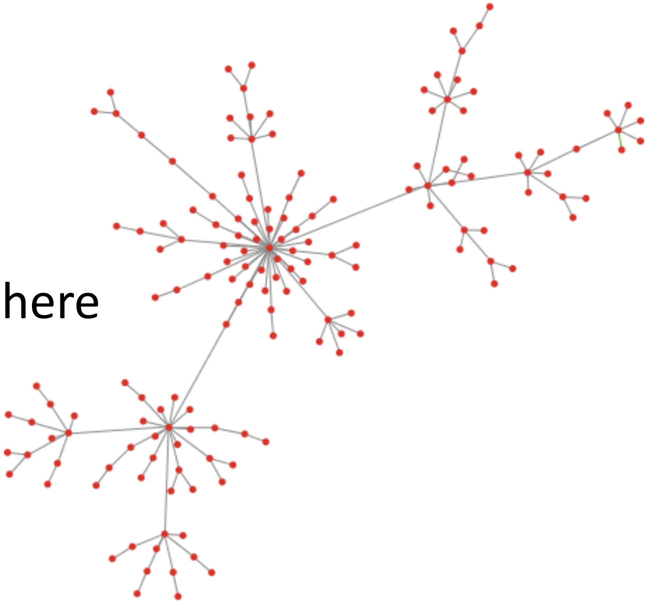
Produces scale-free network
Incorporates two important general concepts:

- Growth – number of nodes in the network increases over time
- Preferential attachment – rich gets richer

Generative algorithm:

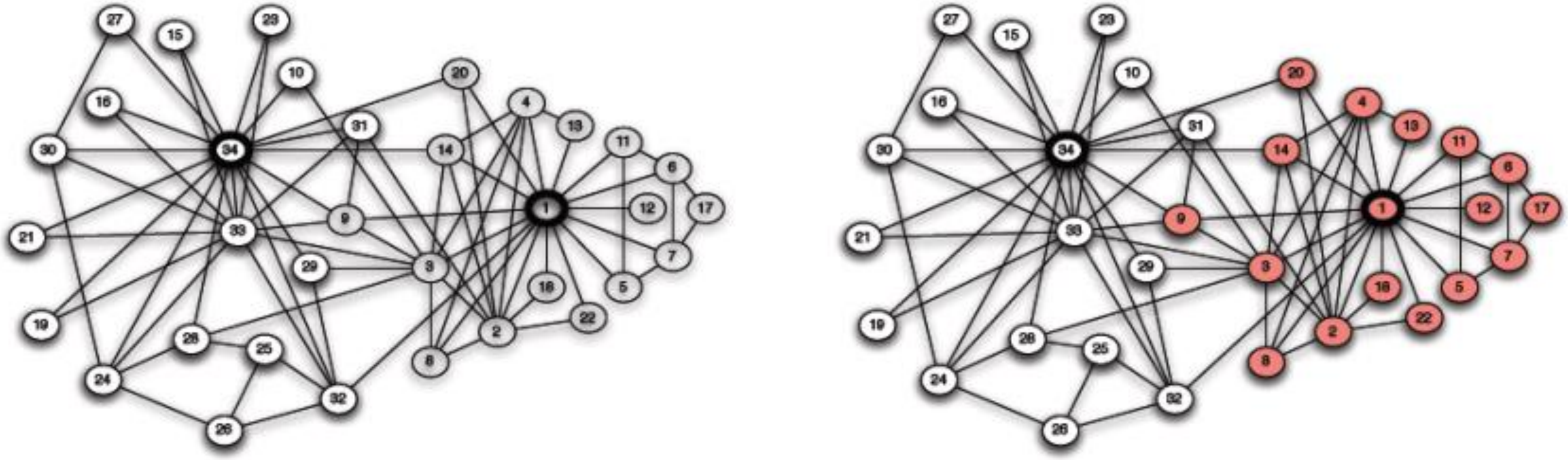
1. Start with a small random network
2. At each time step add a new node v.
3. Add m edges from v to the nodes that already there with the probability given by

$$p_i = \frac{\deg(v_i)}{\sum_j \deg(v_j)}$$



Are there ways to make the generative model even more realistic? How?

Communities



(a) *Karate club network*

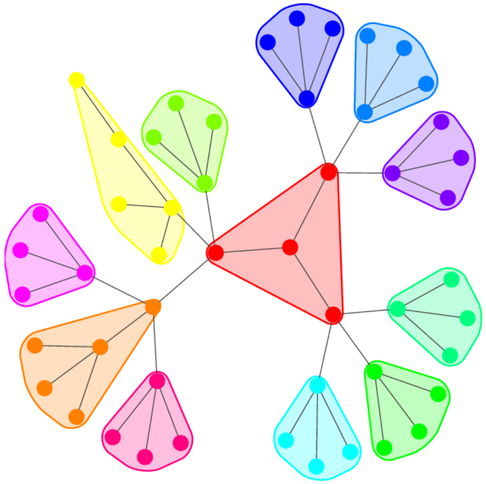
(b) *After a split into two clubs*

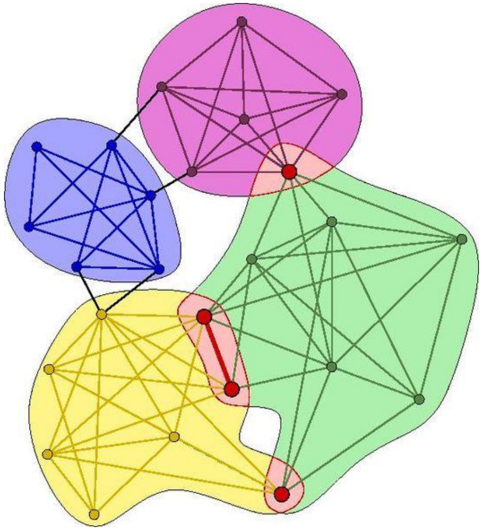
The first scientist at any conference on networks
who uses Zachary's karate club as an example is inducted into
the Zachary Karate Club, and awarded a prize.
http://networkkarate.tumblr.com/

What is the community

A graph is said to have community structure if the nodes can be grouped into sets of nodes such that

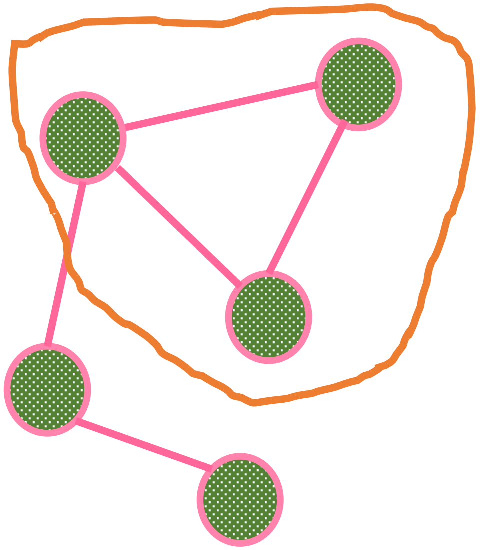
- density within a group is high and
- density between the groups is low



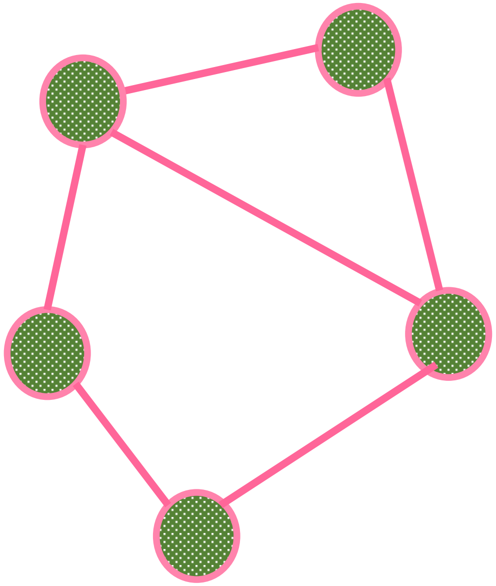


Community is a clique, if every member of the community has edges to every other member.

K-core is relaxed condition of a clique: each node within a community is connected to k other nodes in it.

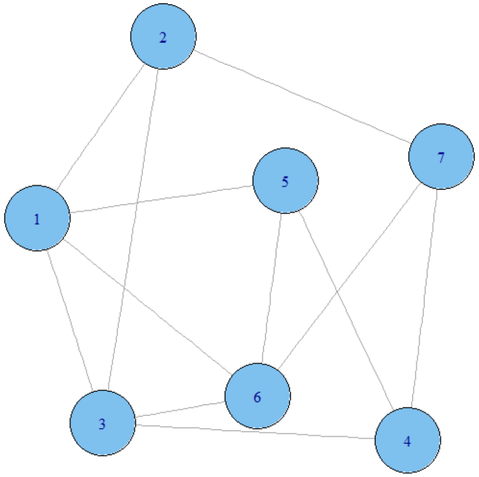


Clique of size 3



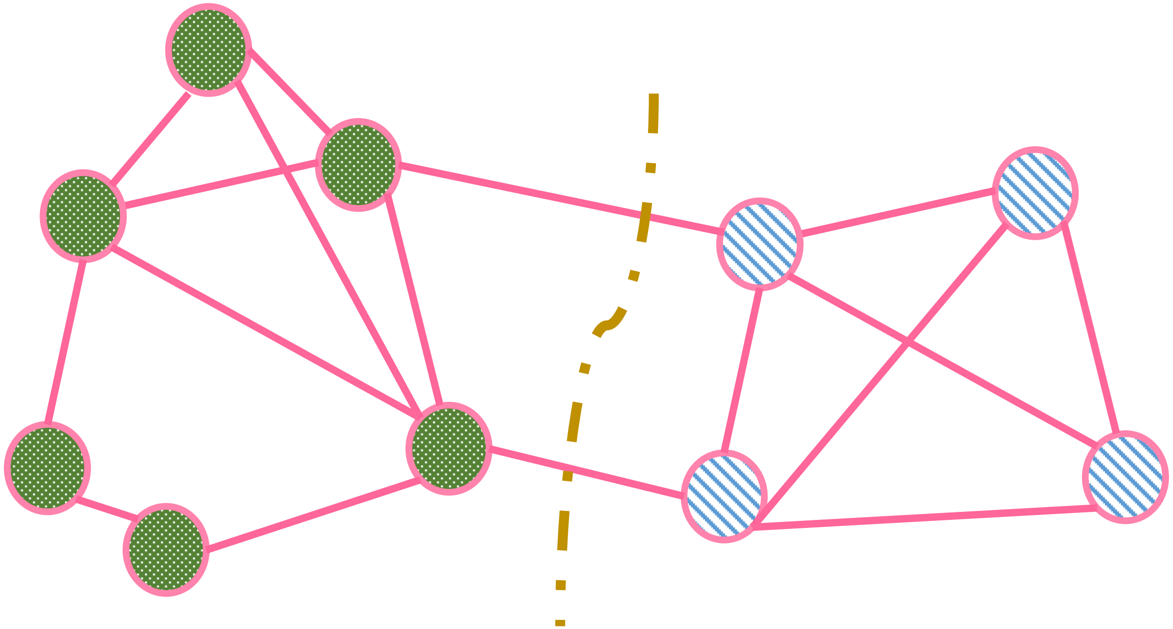
2-core

What is the k for the k-core community:



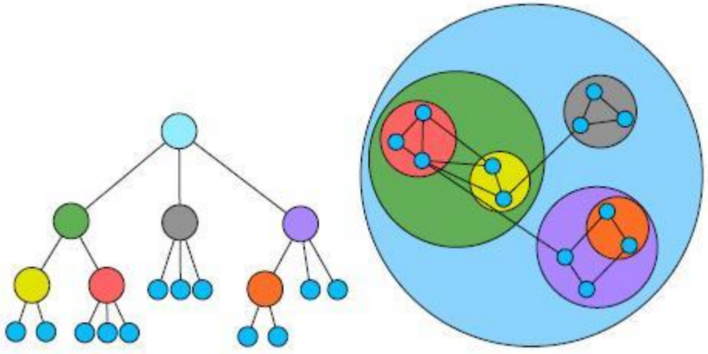
Community detection

Discovering the natural structure of the social network in an automated way.



Hierarchical clustering

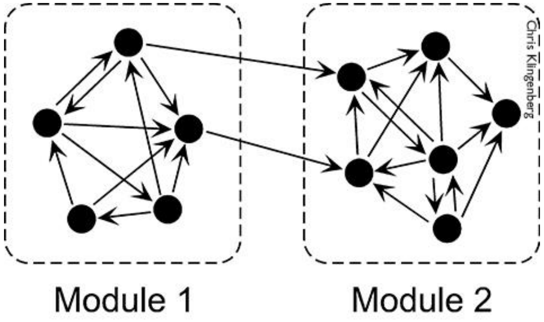
- Define and calculate similarity measure between all pairs of nodes
- Start with all nodes disconnected
- Add edges between pairs one by one according to the calculated measure
- Given structure allows to take different levels of aggregation for the communities



(a) Graph Hierarchy (b) Graph

Daniel, A., M. Tamara, and A. David, *GrouseFlocks: Steerable Exploration of Graph Hierarchy Space*.

Modularity



- Measure of goodness of the division into communities

$$Q = \frac{1}{2|E|}\sum_{vw}\left(A_{vw} - \frac{\deg(v)\deg(w)}{2|E|}\right)\delta(c_v, c_w)$$

Probability of an edge between two nodes is proportional to their degrees

1 if nodes are in the same community

Adjacency matrix

$$A_{vw} = \begin{cases} 1 \; if\ vertices\ v\ and\ w\ are\ connected \\ 0 \;\; otherwise \end{cases}$$

Q = 0 for a random network

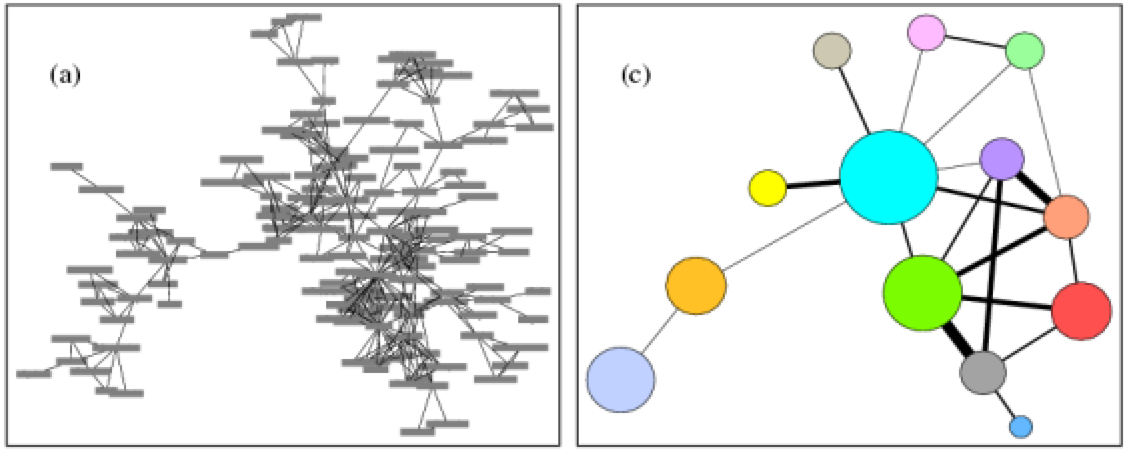
Louvain community detection

- a greedy optimization method that attempts to optimize the modularity of a partition of the network

Two steps performed iteratively:

- starts with small communities and optimizes modularity locally
- aggregates nodes belonging to the same community and builds a new network whose nodes are the communities

http://perso.uclouvain.be/vincent.blondel/research/louvain.html



By L.Adamic

Overlapping communities

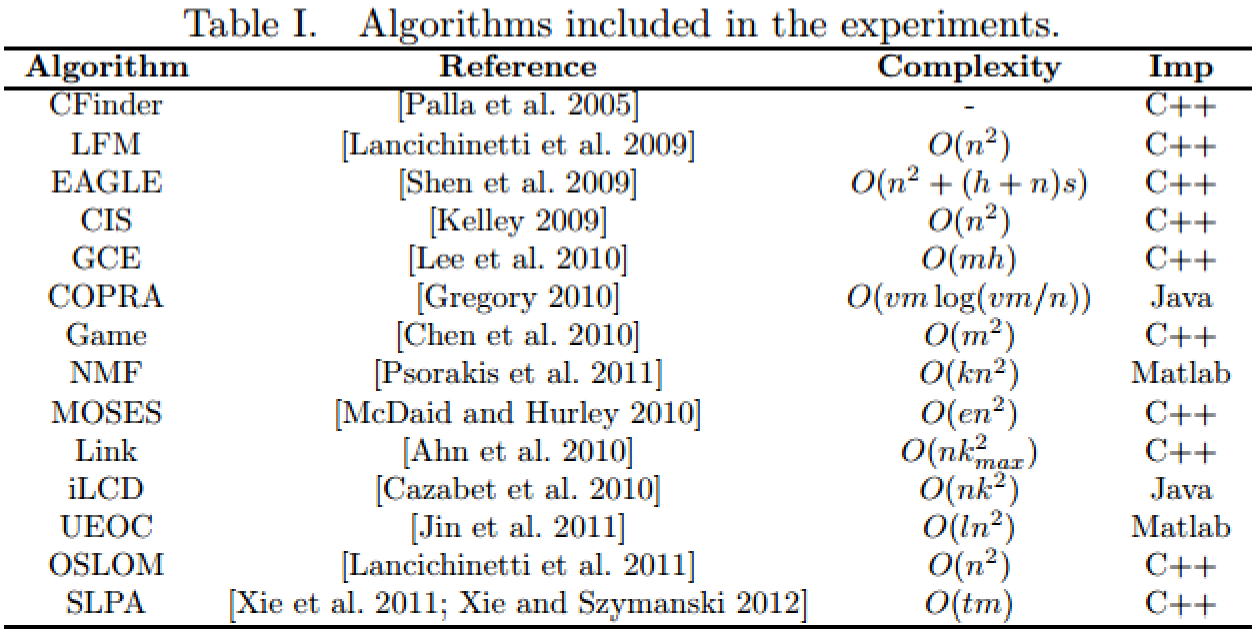


Table I. Algorithms included in the experiments.

Algorithm	Reference	Complexity	Imp
CFinder	[Palla et al. 2005]	-	C++
LFM	[Lancichinetti et al. 2009]	$O(n^2)$	C++
EAGLE	[Shen et al. 2009]	$O(n^2 + (h+n)s)$	C++
CIS	[Kelley 2009]	$O(n^2)$	C++
GCE	[Lee et al. 2010]	$O(mh)$	C++
COPRA	[Gregory 2010]	$O(vm \log(vm/n))$	Java
Game	[Chen et al. 2010]	$O(m^2)$	C++
NMF	[Psorakis et al. 2011]	$O(kn^2)$	Matlab
MOSES	[McDaid and Hurley 2010]	$O(en^2)$	C++
Link	[Ahn et al. 2010]	$O(nk_{max}^2)$	C++
iLCD	[Cazabet et al. 2010]	$O(nk^2)$	Java
UEOC	[Jin et al. 2011]	$O(ln^2)$	Matlab
OSLOM	[Lancichinetti et al. 2011]	$O(n^2)$	C++
SLPA	[Xie et al. 2011; Xie and Szymanski 2012]	$O(tm)$	C++

Jierui Xie et al.

Infomap

“ Infomap optimizes the map equation, which exploits the information-theoretic duality between the problem of compressing data, and the problem of detecting and extracting significant patterns or structures within those data.”

http://www.mapequation.org/code.html

For more intuitive explanation check

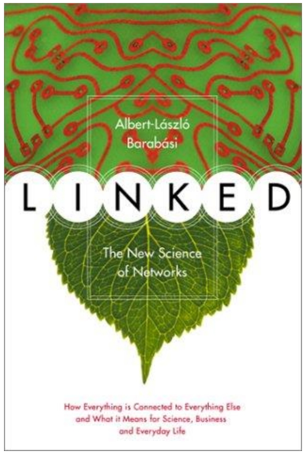
http://www.mapequation.org/apps/MapDemo.html

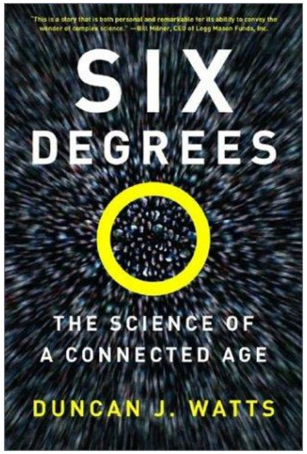
What we do in STACC

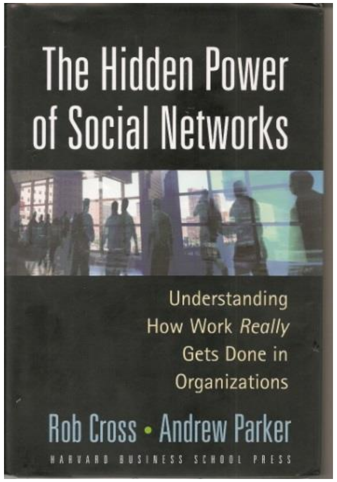
Software Technology and
Applications Competence Center

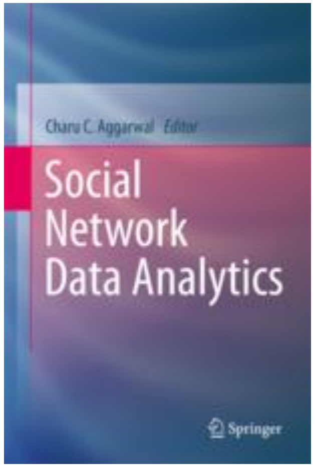
Social Network Analysis for

- Analysis of billions of users and their connections
- Methods for fast shortest path calculation
- Analysis of migration of Skype users
- Service adoption and its dynamics
- Network quality analysis
- Analysis of contact list burstiness
- Methods for fraud detection
- Community detection
- etc


Albert-László
Barabási
LINKED
The New Science
of Networks
How Everything is Connected to Everything Else
and What it Means for Science, Business
and Everyday Life


"This is a story that is both personal and remarkable for its ability to convey the
wonder of complex science." —Bill Miller, CEO of Legg Mason Funds, Inc.
SIX
DEGREES
THE SCIENCE OF
A CONNECTED AGE
DUNCAN J. WATTS


The Hidden Power
of Social Networks
Understanding
How Work Really
Gets Done in
Organizations
Rob Cross • Andrew Parker
HARVARD BUSINESS SCHOOL PRESS


Charu C. Aggarwal Editor
Social
Network
Data Analytics
Springer

. . .

Literature and links

- “Social Network Analysis” course by L. Adamic
- “Random Graphs, Small-Worlds, and Scale-Free Networks” by W-T. Balke and W. Siberski
- “Community detection in graphs”, S.Fortunato
- “Social Netowrk Analysis” by Giorgos Cheliotis
- “Linked” by A.-L. Barabasi
- “Finding community structure in very large networks” by A. Clauset, M.E.J. Newman, C. Moore
- http://datamining.typepad.com/gallery/blog-map-gallery.html
- Data: http://www-personal.umich.edu/~mejn/netdata/