(**S**ocial) **N**etwork **A**nalysis

Giulio Rossetti
Knowledge Discovery and Data Mining Laboratory (KDD) @ ISTI-CNR
giulio.rossetti@isti.cnr.it
@GiulioRossetti

ISTITUTO DI SCIENZA E TECNOLOGIE DELL'INFORMAZIONE "A. FAEDO"

KDD Lab

Ottobre 2021

About me...

Current Position

- Permanent Researcher @ CNR-ISTI, Italy
- External Prof. of “Social Network Analysis” @ University of Pisa
- Coordinator of SNA research @ KDD lab

Research interests

- Complex Networks, Epidemic Modelling, Polluted Information Environments

EU Projects Experience

- WP Leader: SoBigData++
- Unit-PI: HumMingBird, EPO
- Team member: DATASIM, SEEK, CIMPLEX, Track&Know, SAI

Software Libraries (Maintainer)

- NDlib: Network Diffusion
- CDlib: Community Discovery
- DyNetx: Dynamic Network modeling

Giulio Rossetti
CNR-ISTI, Italy
giulio.rossetti@isti.cnr.it

Agenda

Chapter 1: Why should we care about Complex Networks?

Chapter 2: Networks and Graphs

Chapter 3: Tie Strength & Resilience

Chapter 4: Centrality & Assortative Mixing

Chapter 5: Community Discovery

Chapter 6: Link Prediction

Chapter 7: Diffusion - Epidemics

Chapter 8: Diffusion - Opinions

Appendix: Case Studies @ KDD Lab

Chapter 1

Why should we care about Complex Networks?

Summary

- Complexity
- Real world networks
- Emergence of Network Science

Complex

[adj., v. kuh m-pleks, kom-pleks; n. kom-pleks]
adjective

1. Composed of many interconnected parts; compound; composite: a complex highway system.

2. Characterized by a very complicated or involved arrangement of parts, units, etc.: complex machinery.

3. So complicated or intricate as to be hard to understand or deal with: a complex problem.

Source: Dictionary.com

Complexity, a scientific theory which asserts that some systems display behavioral phenomena that are completely inexplicable by any conventional analysis of the systems' constituent parts. These phenomena, commonly referred to as emergent behaviour, seem to occur in many complex systems involving living organisms, such as a stock market or the human brain.

Source: John L. Casti, Encyclopædia Britannica

Complexity

Behind each complex system
there is a network,
that defines the interactions
between the components.

Suggested Reading

Complexity Explained
https://complexityexplained.github.io/

Examples of

Complex Systems

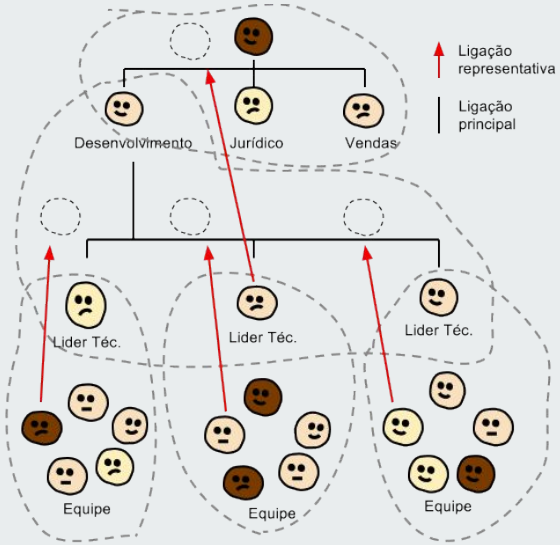
The Facebook “Social Graph”

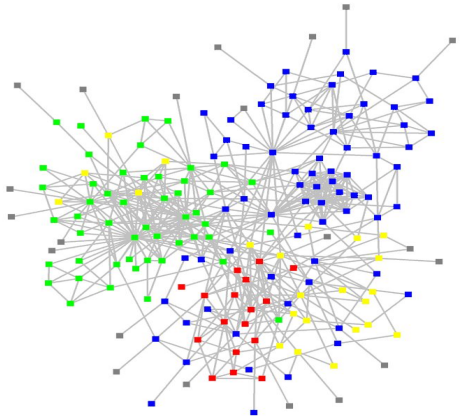
Keith Shepherd's "Sunday Best".
http://baseballart.com/2010/07/shades-of-greatness-a-story-that-needed-to-be-told/

Examples of

Complex Systems

The structure of an organization





departments

consultants

external experts

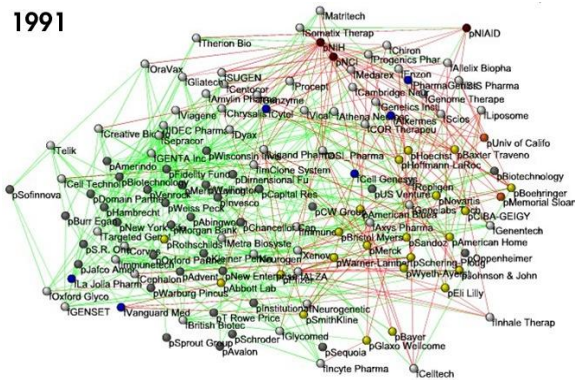
www.orgnet.com

Links:

- Collaborations
- Financial
- R&D

Nodes:

- Companies
- Investment
- Pharma
- Research Labs
- Public
- Biotechnology

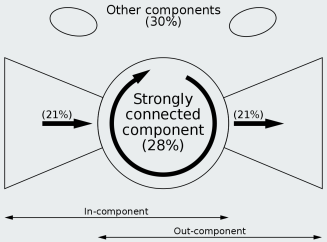


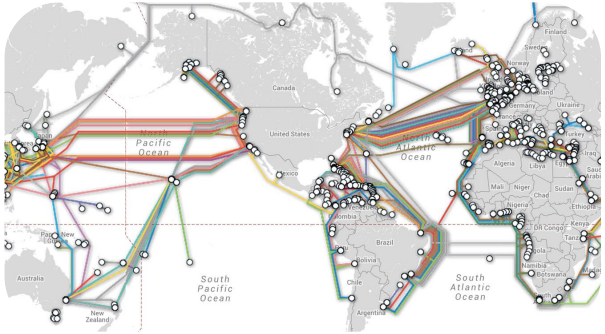
http://ecclectic.ss.uci.edu/~drwhite/Movie

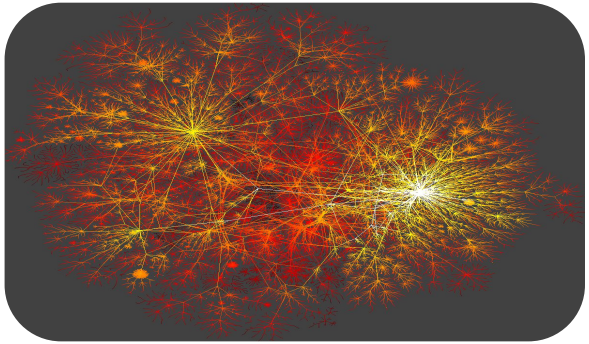
Examples of

Complex Systems

The Internet backbone,
The World Wide Web...

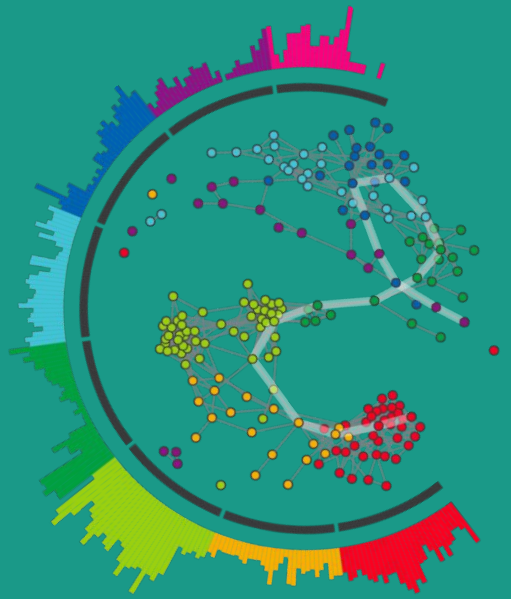






The role of networks

Behind each system studied in complexity there is an intricate wiring diagram, or a **network**, that defines the interactions between the component.



We will never understand **complex system** unless we map out and understand the networks behind them.

Examples of

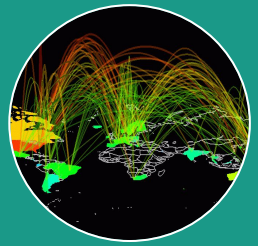
Real world Networks

Type: Social
Nodes: Individuals
Links: Social relationship

NetSci
network science society

Type: Scientific Collaborations
Nodes: Researchers
Links: Co-Authorships

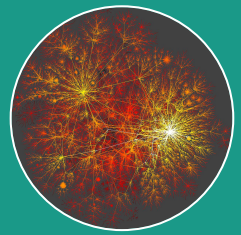
Type: Actor connectivity
Nodes: Actors
Links: Cast jointly



Type: Communication
Nodes: Phones, Airports..
Links: Phone calls, Flights..

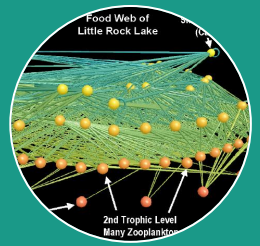
Examples of

Real world Networks (cont'd)



Type: Technological
Nodes: PC, Routers
Links: Physical lines

Type: Scientific Citation
Nodes: Papers
Links: Citations



Type: Biological
Nodes: Species
Links: Trophic interactions

Type: Mobility
Nodes: Individuals, Cars...
Links: Co-Location...

The Tools of

Modern Network Theory

1. Graph Theory
2. Social Network Theory
3. Statistical Physics
4. Computer Science
5. Biology
6. Statistics

Chapter 2

Networks & Graphs

Summary

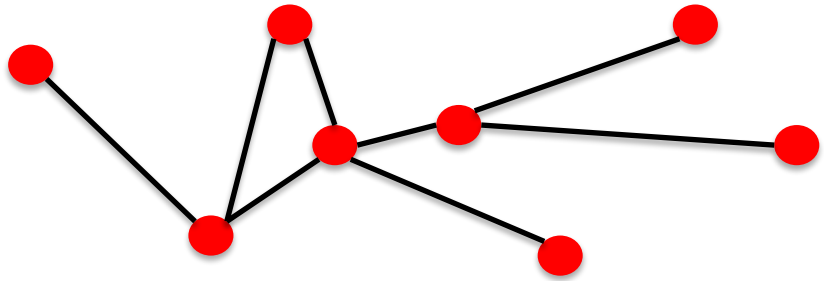
- Type of Networks
- Degree distribution
- Paths & Connectedness
- Clustering

Components of a Complex System

Networks or Graphs?

Network *<nodes, links>*
refers to real systems
(www, social network, metabolic network)

Graph *<vertices, edges>*
mathematical representation of a network
(web graph, social graph)



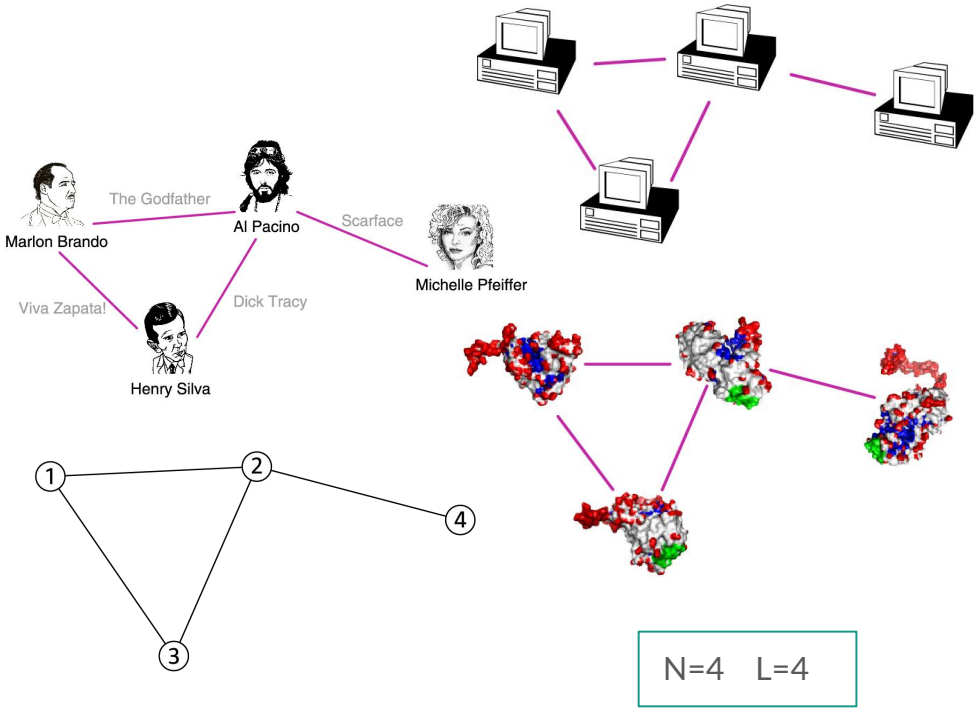
		Symbol
Components	nodes, vertices	N
Interactions	edges, links	L
System	network, graph	(N,L)

A Common Language

The choice of the proper network representation determines our ability to use network theory successfully.

In some cases there is a unique, unambiguous representation. In other cases, the representation is by no means unique.

The way we assign the links between a group of individuals will determine the nature of the question we can study.



If you connect individuals based on their first name

(e.g., *all Peters connected to each other*),

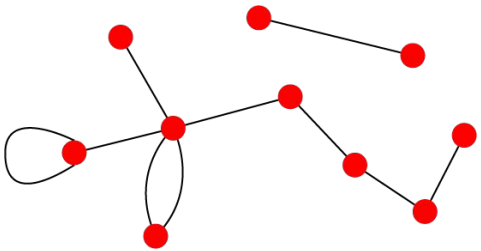
you will be exploring what?

It is a network, *nevertheless*.

Directedness

Undirected graphs

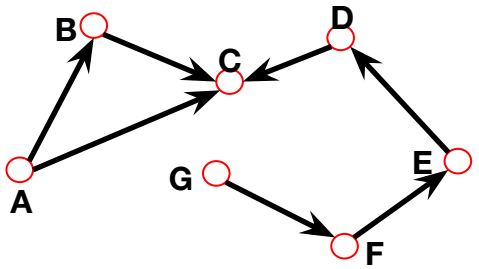
Links: undirected (*symmetrical*)



Examples of Undirected links
- Co-authorship links
- Actor network
- Protein interactions

Directed graphs (DiGraphs)

Links: directed (*arcs*).

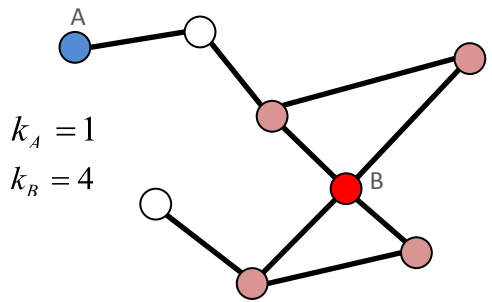


Example of Directed links
- URLs on the www
- Phone calls
- Metabolic reactions

Node Degree

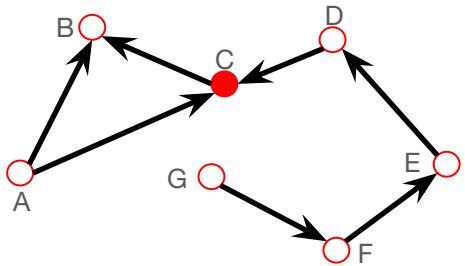
Undirected graphs

the number of links connected to the node



Directed graphs (DiGraphs)

we can define an in-degree and out-degree.
The (total) degree is the sum of in- and out-degree.



Source: a node with $k^{in} = 0$;

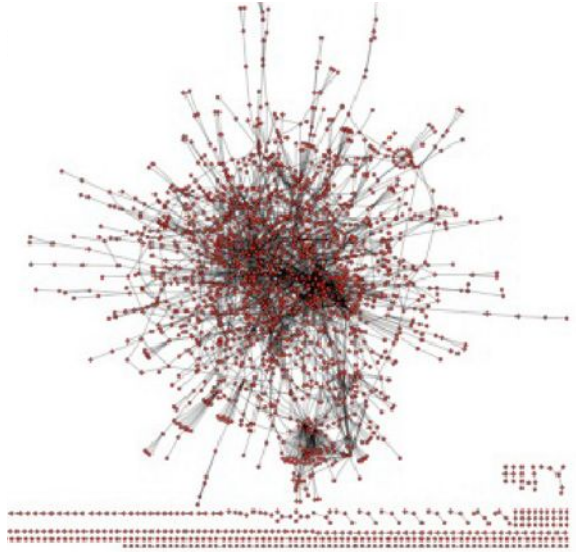
Sink: a node with $k^{out} = 0$.

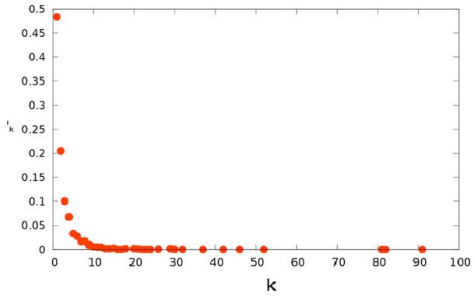
Degree Distribution

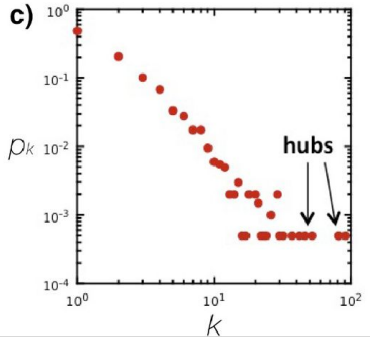
P(k): probability that a randomly chosen node has degree *k*

N_k = # nodes with degree k

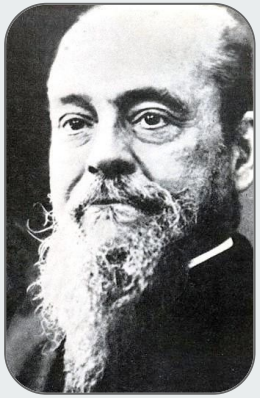
$P(k) = N_k / N$ ➔ plot





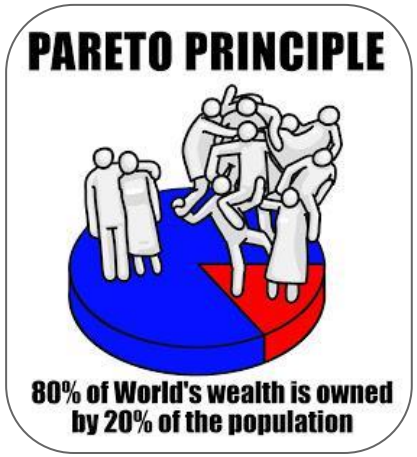


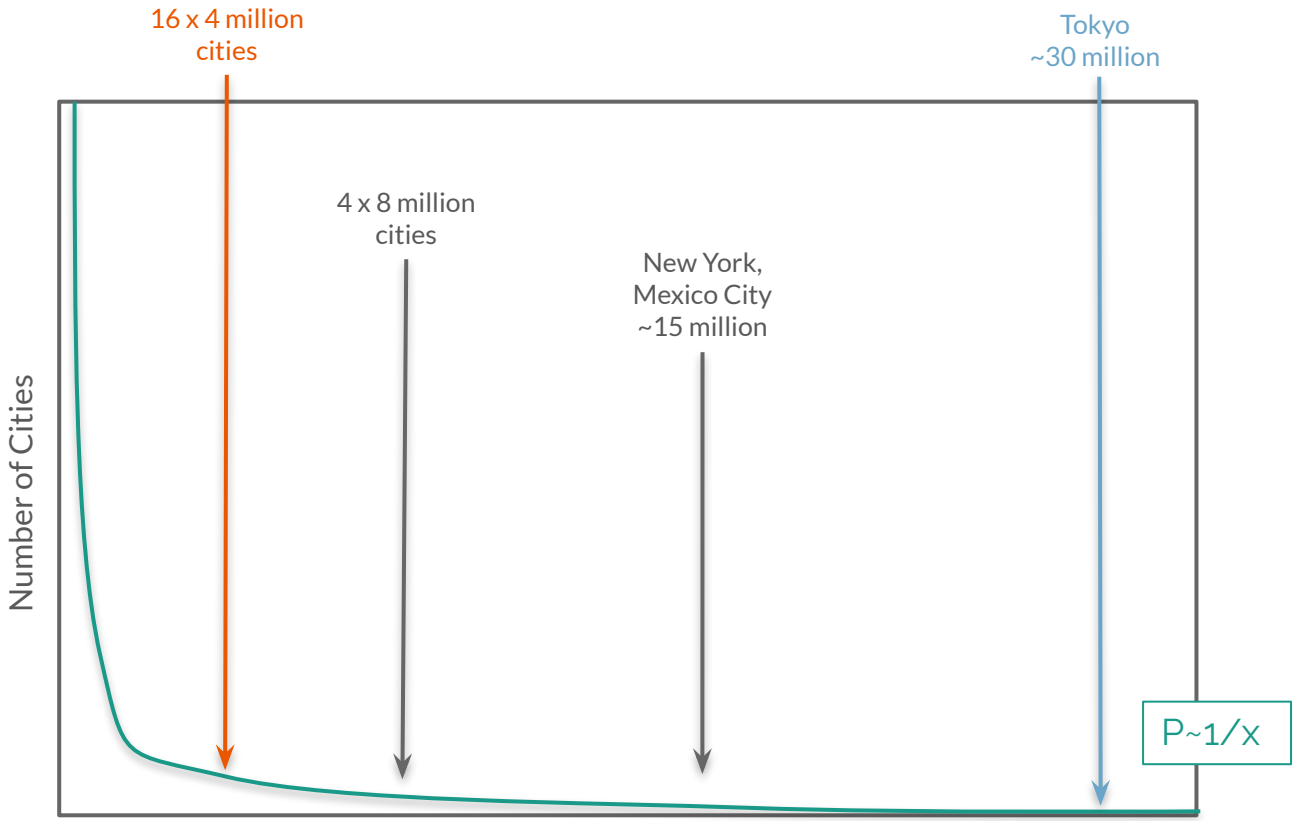
Hubs - 80/20 Rule



Vilfredo Federico Damaso Pareto (1848 – 1923)

Italian economist, political scientist and philosopher, who had important contributions to our understanding of income distribution and to the analysis of individuals choices.
A number of fundamental principles are named after him, like Pareto efficiency, Pareto distribution *(another name for a power-law distribution)*, the Pareto principle (or 80/20 law).





Hubs - Sizes of Cities:
there is an equivalent number of people living in cities of all sizes!

Paths and Connectedness

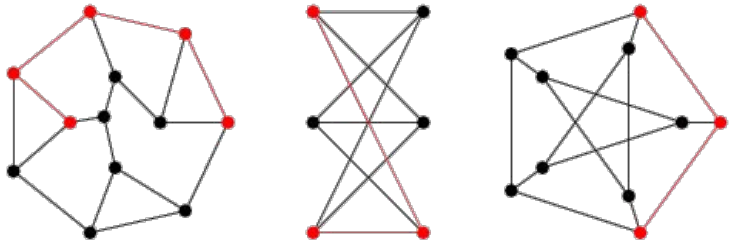
Paths

A *path is* a sequence of nodes in which each node is adjacent to the next one

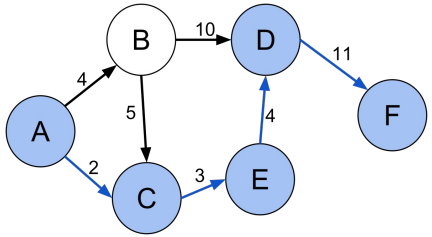
$P_{i0,in}$ of length *n* between nodes i_0 and i_n is an ordered collection of *n+1* nodes and *n* links

$$P_n = \{i_0, i_1, i_2, \ldots, i_n\}$$

$$P_n = \{(i_0, i_1), (i_1, i_2), (i_2, i_3), \ldots, (i_{n-1}, i_n)\}$$



Examples of phats in an undirected graph.

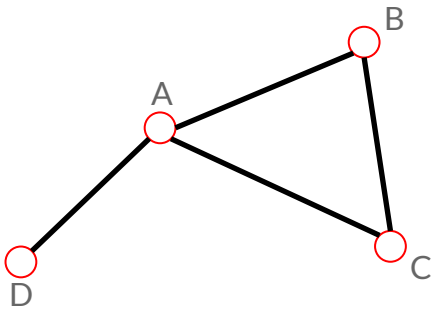


In a directed graph, the path can follow only the direction of an arrow.

Distance in a Graph

Undirected graphs

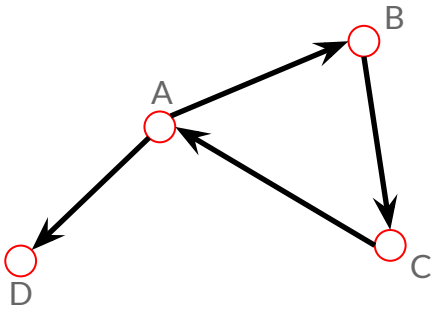
The *distance* (*shortest path*, *geodesic path*) between two nodes is defined as the number of edges along the shortest path connecting them.



*If the two nodes are disconnected, the distance is infinity.

Directed graphs (DiGraphs)

Each path needs to follow the direction of the arrows.



Thus in a digraph the distance from node A to B (on an AB path) is generally different from the distance from node B to A (on a BCA path).

History of

Six Degrees

Karinthy, Frigyes

1929:
Minden másképpen van (Everything is Different)
Láncszemek (Chains)

"Look, Selma Lagerlöf just won the Nobel Prize for Literature, thus she is bound to know King Gustav of Sweden, after all he is the one who handed her the Prize, as required by tradition. King Gustav, to be sure, is a passionate tennis player, who always participates in international tournaments. He is known to have played Mr. Kehrling, whom he must therefore know for sure, and as it happens I myself know Mr. Kehrling quite well."

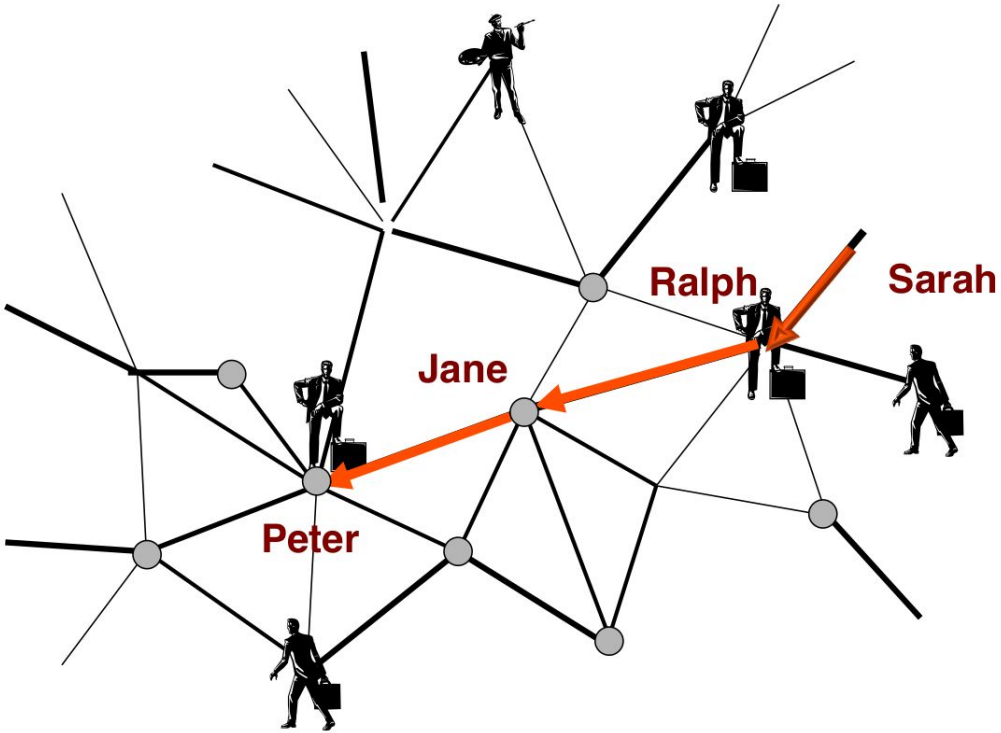
History of

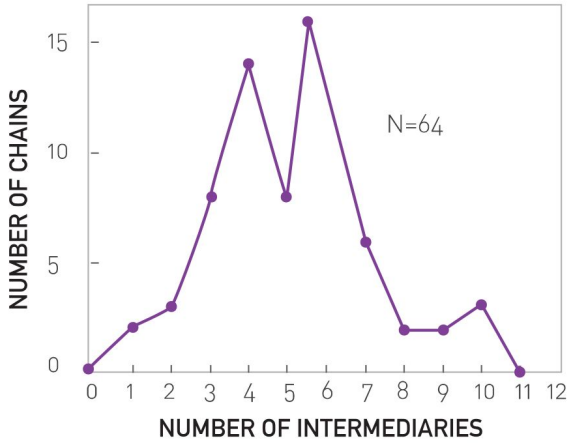
Six Degrees

1967: ***Stanley Milgram***

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 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 (POST CARDS AND ALL) TO A PERSONAL ACQUAINTANCE WHO IS MORE LIKELY THAN YOU TO KNOW THE TARGET PERSON. You may send the folder to a friend, relative or acquaintance, but it must be someone you know on a first name basis.*





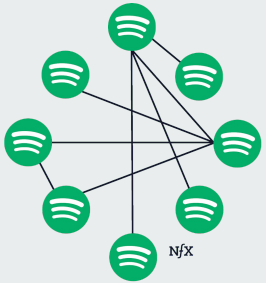
It’s a small, small world!

Milgram Experiment

Network Density

Network Density

Ratio of existing edges over possible ones.

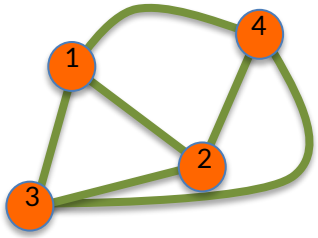


Low Density

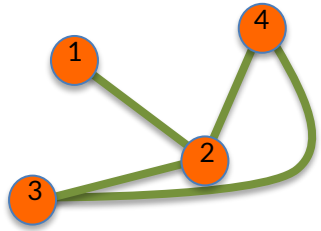
High Density

$$d(G) = \frac{L}{L_{max}}$$

Examples



Complete graph
d(G) = 1



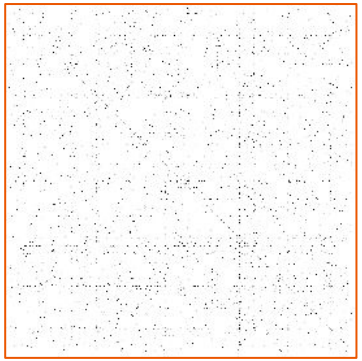
Sparse graph
d(G) = 2/3

Most networks observed in real systems are sparse

$L << L_{max}$

$\langle k \rangle << N-1$

$d(G) << 1$



Sparse
Adjacency matrix

WWW (ND Sample):	N=325,729;	$L=1.4\ 10^6$	$L_{max}=10^{12}$	$\langle k \rangle=4.51$
Protein (*S. Cerevisiae*):	N= 1,870;	L=4,470	$L_{max}=10^7$	$\langle k \rangle=2.39$
Coauthorship (Math):	N= 70,975;	$L=2\ 10^5$	$L_{max}=3\ 10^{10}$	$\langle k \rangle=3.9$
Movie Actors:	N=212,250;	$L=6\ 10^6$	$L_{max}=1.8\ 10^{13}$	$\langle k \rangle=28.78$

(Source: Albert, Barabasi, RMP2002)

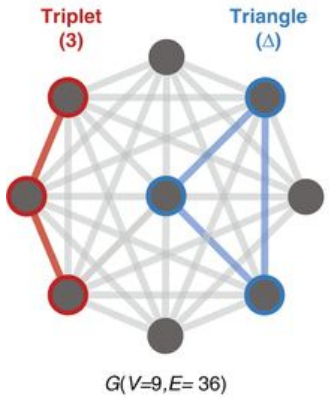
Clustering Coefficient

Clustering Coefficient

How “clustered” is my network?

Global Clustering coefficient

- Triangles and triplets
- C in [0,1]



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

Watts & Strogatz, Nature (1998)

Summarizing...

Central quantities in Network Science

Degree Distribution

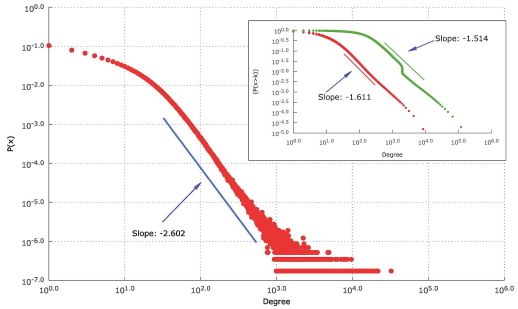
Presence of Hubs

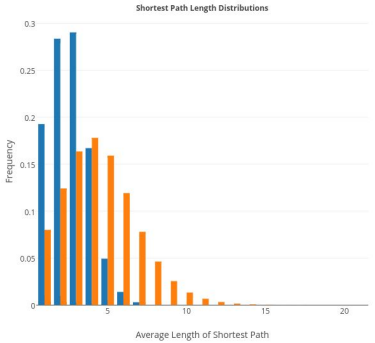
Path length

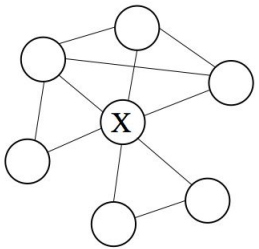
It's a small world!

Density & Clustering Coefficient

Social graphs are globally sparse:
However, locally, triangles are expected to close!







Network	Directed	Weighted	Multigraph	Self-loops
WWW	yes	no	yes	yes
Protein interactions	no	no	no	yes
Collaboration network	no	yes	yes	no
Mobile phone calls	yes	yes	no	no
Facebook Friendship	no	no	no	no

Real Networks can have multiple characteristics

Chapter 3

Tie Strength & Resilience

Summary

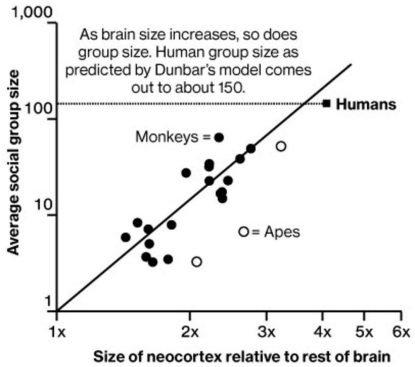
- Tie Strength
- Resilience/Robustness
- Failures and Attacks

How many friends does one person needs?

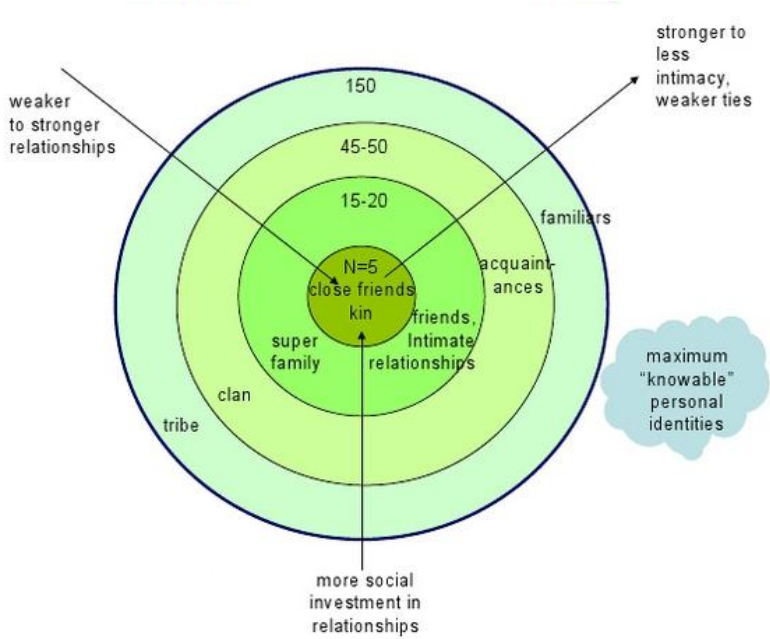
Not all ties in a social graph are the same

Dunbar's Number (Sociological Theory)

a suggested cognitive limit to the number of people with whom one can maintain stable social relationships



Considering the average human brain size and extrapolating from the results of primates, humans can comfortably maintain 150 stable relationships



In Dunbar's own words:

"the number of people you would not feel embarrassed about joining uninvited for a drink if you happened to bump into them in a bar"

Dunbar, Robin IM. Neocortex size as a constraint on group size in primates. Journal of human evolution 22.6 (1992): 469-493.

Dunbar, Robin. How many friends does one person need?: Dunbar's number and other evolutionary quirks. Faber & Faber, 2010.

The Strength of Weak Ties

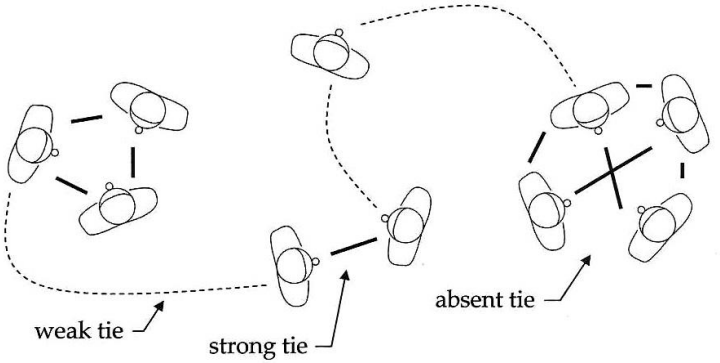
The strength of weak ties

Mark S. **Granovetter**, 1973

- (PhD Thesis)
  "How people get to know about new jobs?"
- Answer: Through personal contacts

Unexpected result:

Often acquaintances, **not** close friends… but *why*?



How to measure tie strength?

Granovetter's **dimensions of tie strength**:

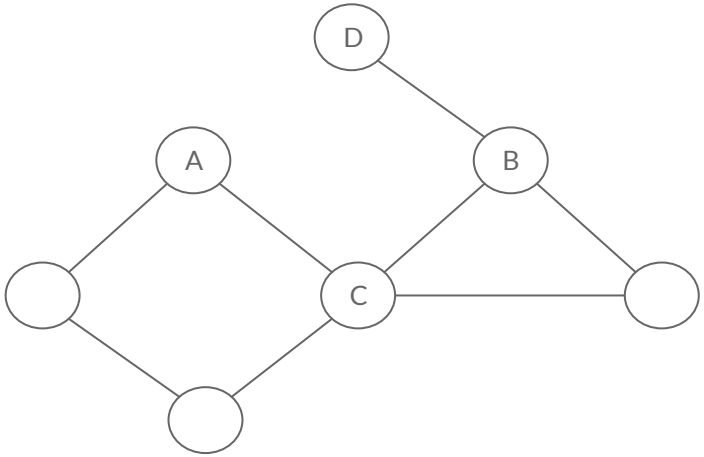
- the amount of time spent interacting with someone,
- the level of intimacy,
- the level of emotional intensity,
- and the level of reciprocity.

Granovetter, Mark S. "The strength of weak ties." Social networks. Academic Press, 1977. 347-367.

Triadic Closure

Social Intuition:

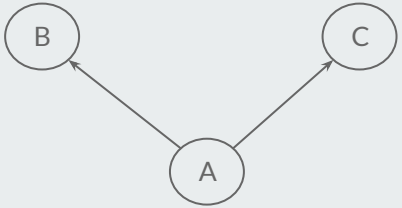
if two people in a network have a friend in common there is an increased likelihood that they will become friends themselves



Which is more likely to appear (A,B) or (A,D)?

Triadic Closure

Triadic Closure
implies
High Clustering Coefficient



(Social) Reasons for triadic closures

If B and C have a friend A in common then:

- B is more likely to meet C
  (since they spend time with A)
- B and C trust each other
  (since they have a friend in common)
- A has incentive to bring B and C together
  (as it is hard for A to maintain two disjoint relationships)

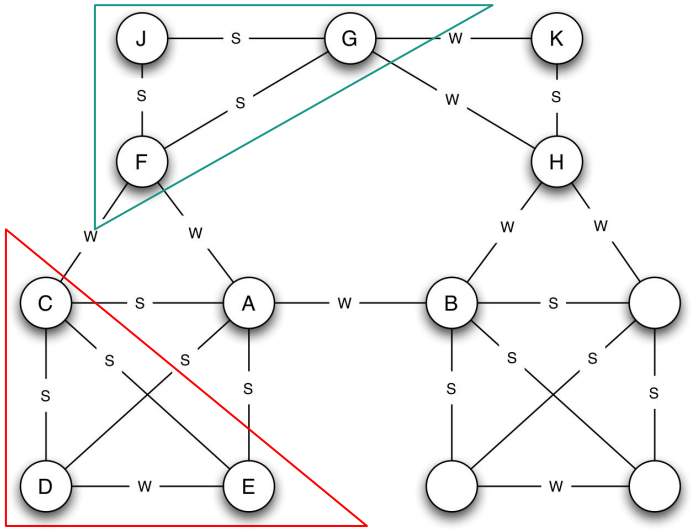
Strong Triadic Closure

Links in networks have strength;

- Friendship, Communication

We characterize links as either:

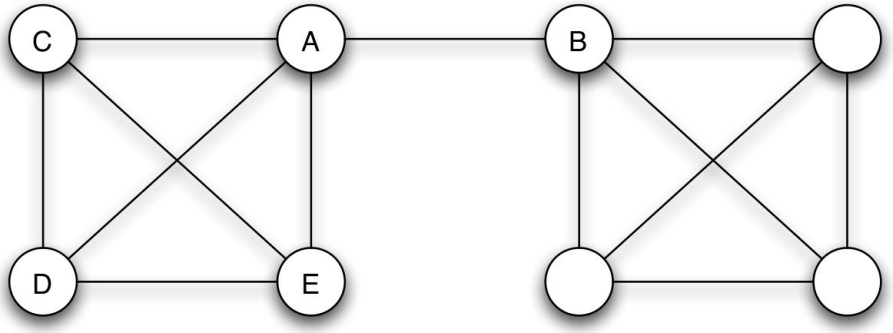
- Strong (friends), or
- Weak (acquaintances)



Strong Triadic Closure Property:

if A has strong links to B and C then there must be a link (B,C) (that can be strong or weak)

Bridges and Local Bridges



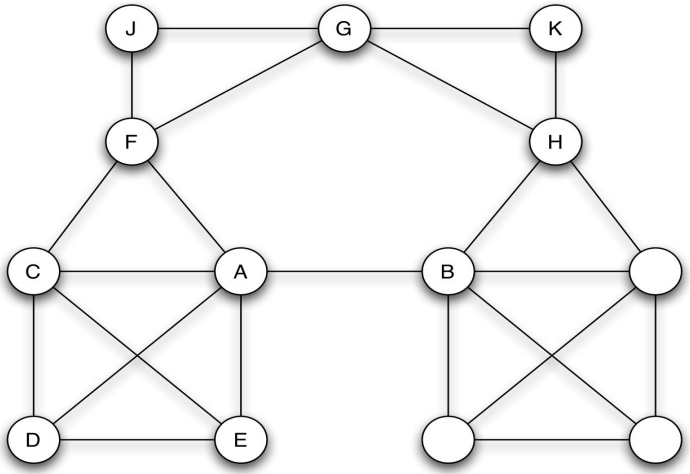
Strong Definition

Edge (A,B) is a bridge if deleting it would make A and B in two separate connected components

Bridges and Local Bridges

Edge (A,B) is a local bridge since A and B have no friends in common

Bridges are weak ties!



The span of a local bridge is the distance of the edge endpoints if the edge is deleted

Local bridges with long span are like real bridges

Measuring Tie Strength in Real Data

Social proximity and tie strength

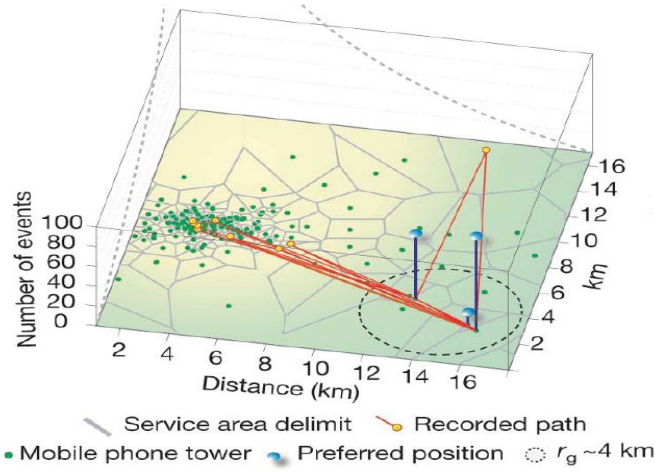
How connected are u and v in the social network.

- Various well-established measures of network proximity, based on the common neighbors (Jaccard, Adamic-Adar) or the structure of the paths (Katz) connecting u and v in the who-calls-whom network.

How intense is the interaction between u and v.

- Number of calls as **strength of tie**

Cell-phone network of 20% of country's population



when you call

where you call

who you call

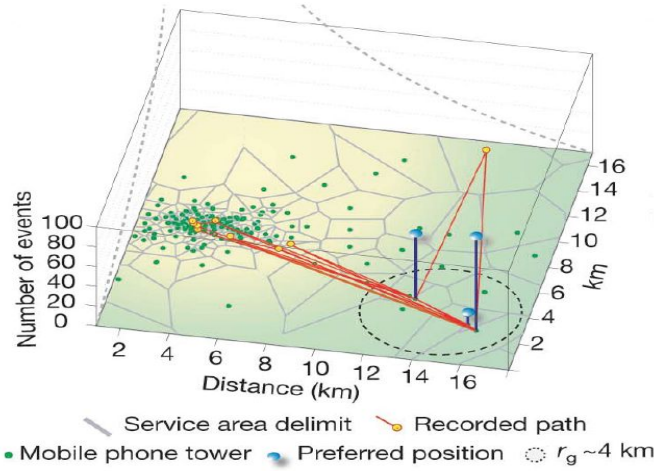
J.-P. Onnela, J. Saramaki, J. Hyvonen, G. Szabo, D. Lazer, K. Kaski, J. Kertesz, A.-L. Barabási. Structure and tie strengths in mobile communication networks. PNAS 104 (18), 7332-7336 (2007).

Strength of weak ties

First large scale empirical validation of Granovetter's theory

- Social proximity increases with tie strength
- Weak ties span across different communities

Cell-phone network of 20% of country's population



when you call

where you call

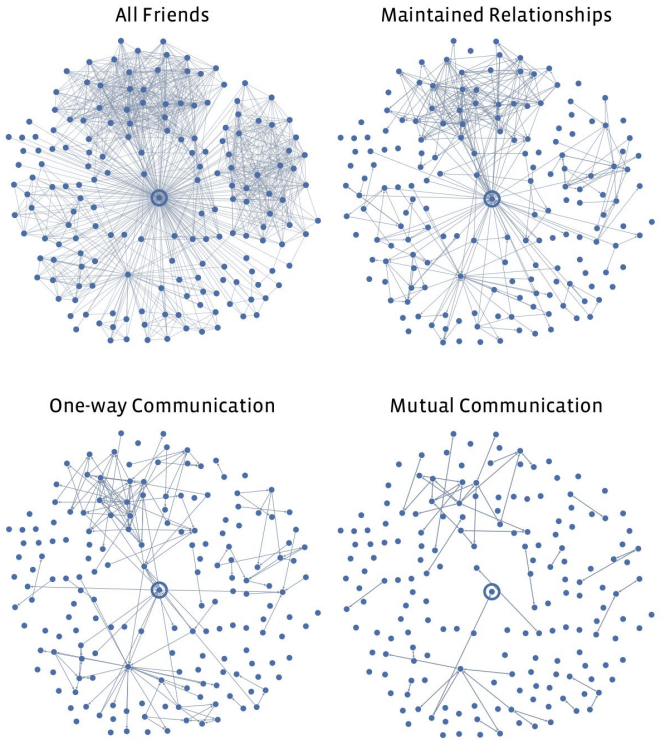
who you call

J.-P. Onnela, J. Saramaki, J. Hyvonen, G. Szabo, D. Lazer, K. Kaski, J. Kertesz, A.-L. Barabási. Structure and tie strengths in mobile communication networks. PNAS 104 (18), 7332-7336 (2007).

Ties Strength on Facebook

Different types of connections

- Mutual communication:
  both user sent messages eachother
- One-way communication:
  user messages where not reciprocated
- Maintained relationship:
  user clicked on content produced by his friend (no communication)

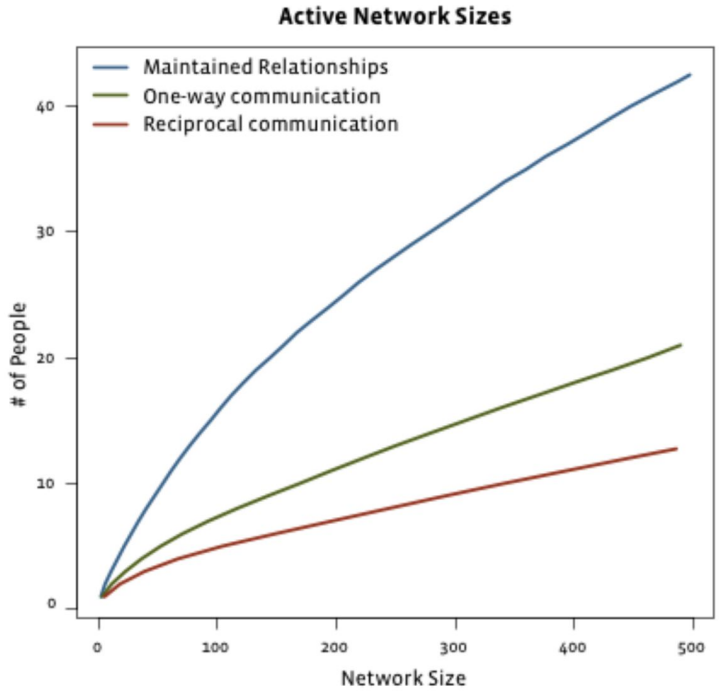


Cameron Marlow, Lee Byron, Tom Lento, and Itamar Rosenn. Maintained relationships on Facebook, 2009. http://overstated. net/2009/03/09/maintainedrelationships-on-facebook

Does tie strength affect network size?

Tie strength allows to:

- discriminate different type of contacts,
- categorize them by the involvement required to nurture them



Network Resilience

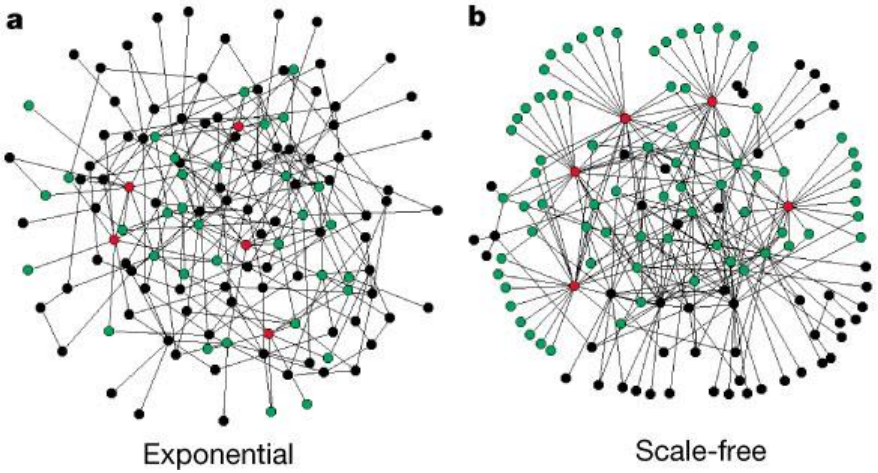
How robust is a complex network to node failures/attacks?

Network robustness and attack tolerance

How network topology is resistant against failure and targeted attacks

Numerical experiment:

1. Take a connected network
2. Remove nodes one at time
3. Observe the size of LCC
   (largest connected component)

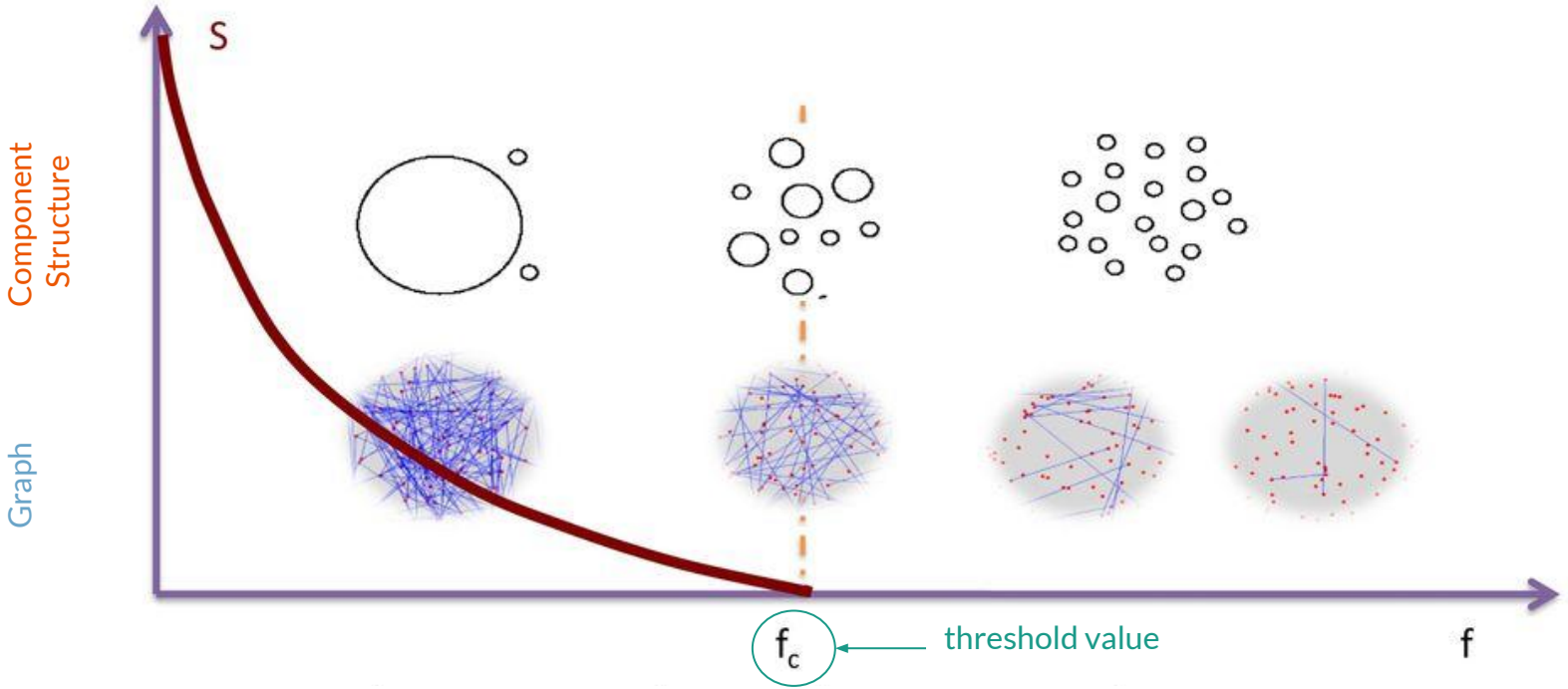


Different topologies, same parameters (N=130, <k>=3.3)

Node removal strategies:

1. Random removal (“failures”)
   e.g., random failure of internet routers

f = fraction of removed nodes



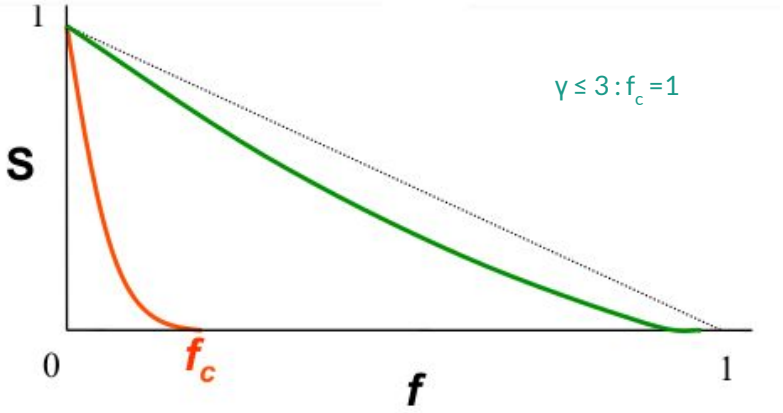
Inverse Percolation problem

Achille's Heel of Scale-free networks

The robustness of scale free networks is due to the hubs, which are difficult to hit by chance

Node removal strategies:

1. Random removal ("failures")
   e.g., random failure of internet routers
2. Remove nodes in descending order of their degrees ("attacks")
   i.e., hubs first



Examples:

- Terrorist attacks
- Efficient vaccination in epidemics

Connecting the dots...

Resilience and Tie Strength

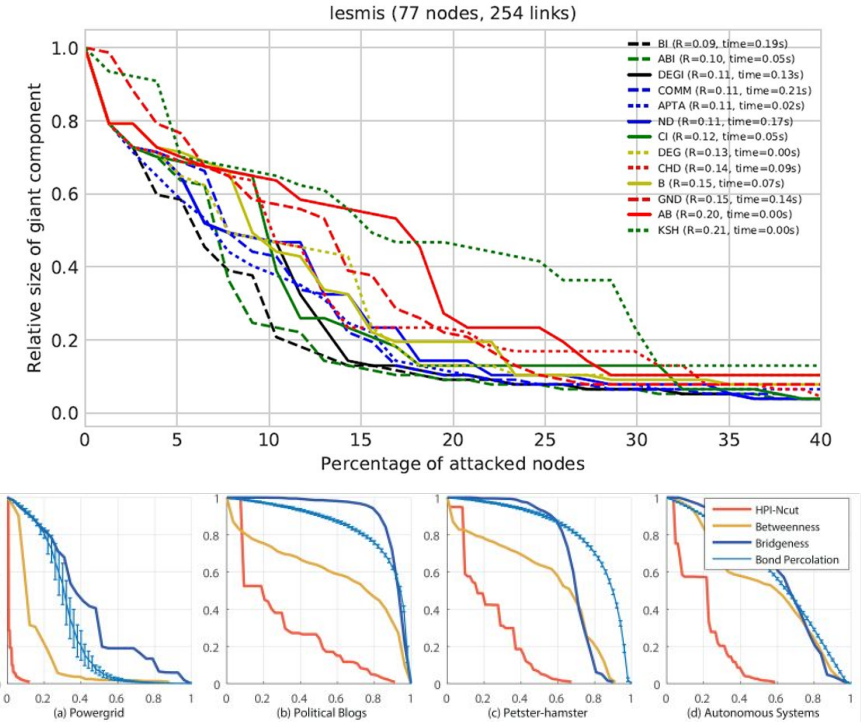
How to target a network?

Not only nodes can fail/being targeted

Identifying local/real bridges can lead to extremely efficient attacks

Node attacks:
- Centrality based
- Community based
- ...

Edge attacks:
- Edge Betweenness centrality removal
- Neighbour overlap removal
- ...



Ren, Xiao-Long, et al. "Underestimated cost of targeted attacks on complex networks." Complexity 2018 (2018).
Wandelt, Sebastian, et al. "A comparative analysis of approaches to network-dismantling." Scientific reports 8.1 (2018): 13513.

Chapter 4

Centrality & Assortative Mixing

Summary

- Measuring Node importance
- Do Birds of a feather flock together?

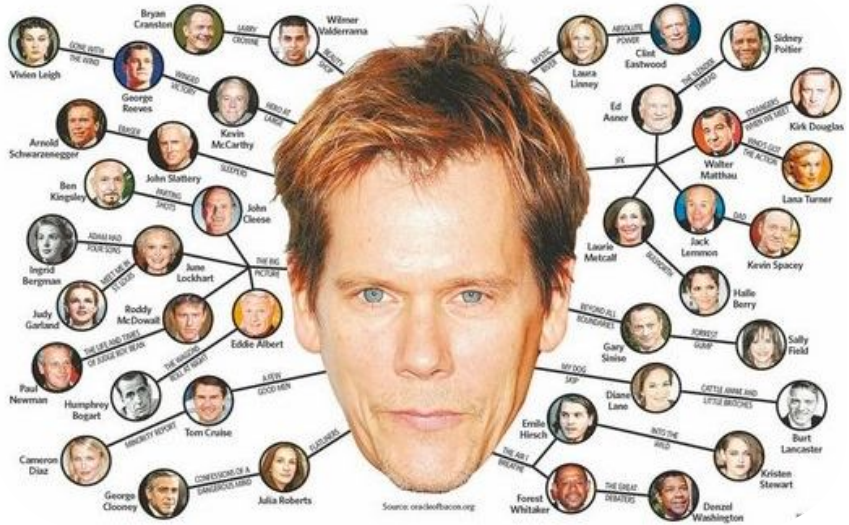
How important is a node in a network?

We can measure nodes importance using so-called centrality.

Bad term:
nothing to do with being central in general

Usage:

- Some centralities have straightforward interpretation
- Centralities can be used as node features for machine learning on graph



https://oracleofbacon.org/

Where are you?

It is always possible, once *fixed a context*, to measure our distance from a “focal” node.

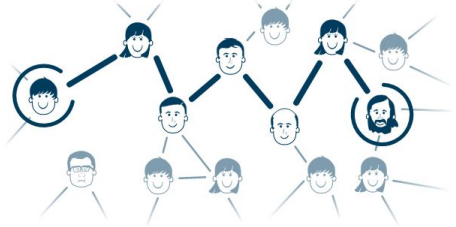
For instance:

Movie Stars:
- Bacon number

Researchers:
- Erdos number

Are such “focal” nodes really different from the others?

Co-authorship distance computation



Find the path between two authors:

Paul Erdős | Giulio Rossetti

Paul Erdős
co-authored 2 papers with Shlomo Moran
co-authored 7 papers with Ronny Lempel
co-authored 4 papers with Fabrizio Silvestri
co-authored 2 papers with **Giulio Rossetti**
distance = 4

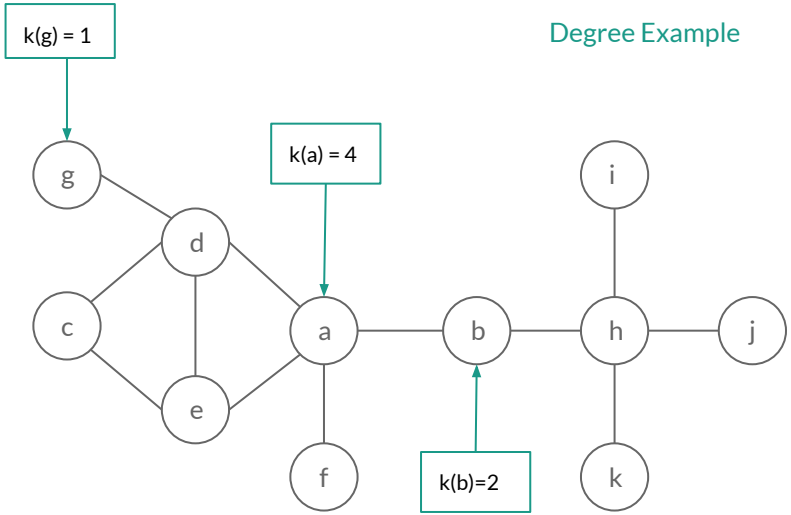
Name	Erdős number	Bacon number	Erdős–Bacon number
Daniel Kleitman	1	2	3[4]
Bruce Reznick	1	2	3[69]
Albert M. Chan	3[21][22][23]	1[24]	4
Nicholas Metropolis	2[10]	2[68]	4
Steven Strogatz	3[79][80][81]	1[a][c][82]	4[a][c]
Robert J. Marks II	3[44][45][46]	2[61][62][63]	5
Tom Porter	3 (in two ways)[6][7]	2[a][c][8][9]	5[a][c]
Richard Thaler	3[86][87][88]	2[89][a][90]	5
Doron Zeilberger	2[93][35]	3[a][94][95][96]	5[a][c]
Misha Collins	4[25][26][27][28]	2[29][30]	6
William A. Dembski	4[44][45][46][47]	2[a][48]	6[a]
Richard Feynman	3	3	6[10]
Ken Goldberg	3[49][50][51]	3[52][53][54]	6
Stephen Hawking	4	2[a]	6[20]

Degree Centrality

How many neighbors does a node have?

Often enough to find important nodes
(e.g., main characters of a series talk with more people)

But not always
- Twitter users with the most contacts are spam
- Webpages/wikipedia pages with most links are often lists of references



k = number of links

$A_{i,j}$ { 1 if i and j are connected, / 0 otherwise

$$k_i = \sum_{j=1}^{n} A_{ij}$$

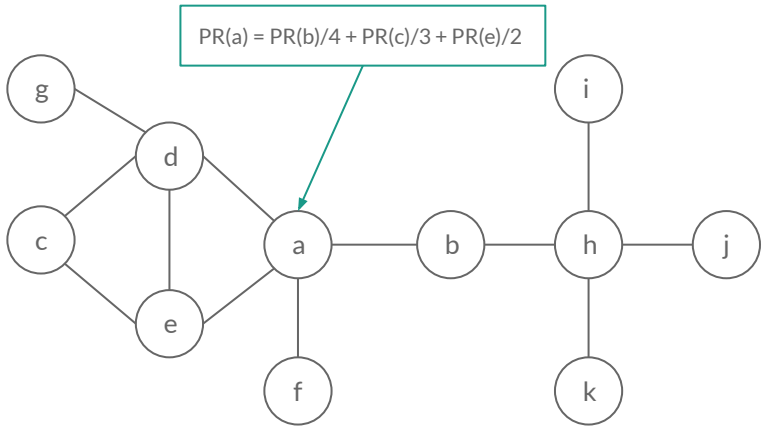
PageRank

Main idea: The PageRank computation can be interpreted as a Random Walk process with restart

Probability that the RW will be in node i next step depends only on the current node *j* and the transition probability
$j \rightarrow i$ determined by the stochastic matrix

- Consequently this is a first-order Markov process
- Stationary probabilities (i.e., when walk length tends towards ∞) of the RW to be in node *i* gives the PageRank of the node

PageRank Example



$$PR(x) = \frac{1-\alpha}{N} + \alpha\left(\sum_{k=1}^{n} \frac{PR(k)}{C(k)}\right)$$

Teleportation probability: the parameter α gives the probability that in the next step of the RW will follow a Markov process or with probability 1-α it will jump to a random node

- α<1, it assures that the RW will never be stuck at nodes with k_{out} = 0, but it can restart the RW from a randomly selected other node (usually α=0.85)

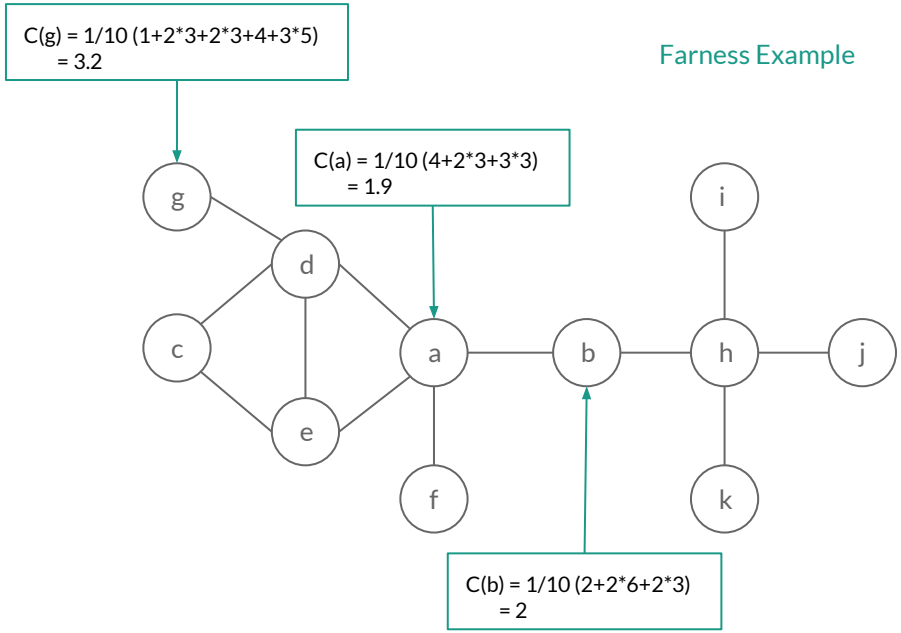
Closeness Centrality

Farness: average of length of shortest paths to all other nodes

Closeness: inverse of the Farness (normalized by number of nodes)

- Highest closeness = More central
- Closeness=1: directly connected to all other nodes
- Well defined only on connected networks

Farness Example



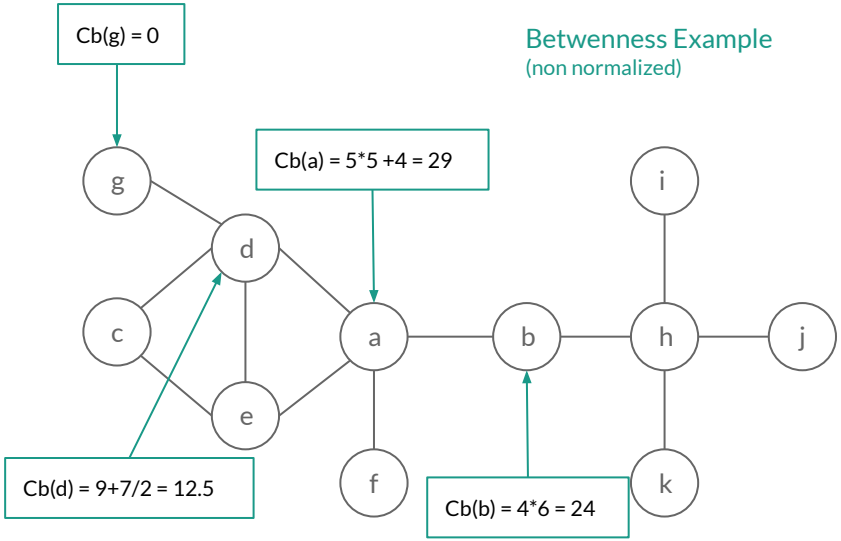
Closeness Formula

$$C_{cl}(i) = \frac{n-1}{\sum_{d_{ij}<\infty} d_{ij}}$$

Betwenness Centrality

Number of shortest paths that go through a node.

- **Assumption:** important vertices are bridges over which information flows
- **Practically:** if information spreads via shortest paths, important nodes are found on many shortest paths



$\sigma_{jk}(i) =$ number of geodesic path from j to k via i: $j \to \ldots \to i \to \ldots \neg k$
$\sigma_{jk} =$ number of geodesic path from j to k : $j \to \ldots \to k$

Definition $$C_b(i) = \sum_{j \neq k} \frac{\sigma_{jk}(i)}{\sigma_{jk}}$$

Normalized def. $$C_b(i) = \frac{1}{n^2} \sum_{j \neq k} \frac{\sigma_{jk}(i)}{\sigma_{jk}} \quad \text{where} \quad C_b \in [0, 1]$$

00	Degree	How many friends do you have?
01	Eigenvector	Are you connected to important nodes?
02	PageRank	How many important interactions do you have?
03	Katz	What's your degree of influence?
04	Closeness	What's your average distance w.r.t. the rest of the network?
05	Harmonic	What's your harmonic average distance w.r.t. the rest of the network?
06	Betwenness	How much do you help the network to stay connected?

Connectivity-based centralities

Geometric centralities

Each centrality measures is a proxy of an underlying network process.

If such a process is irrelevant for the network than the centrality measure makes no sense

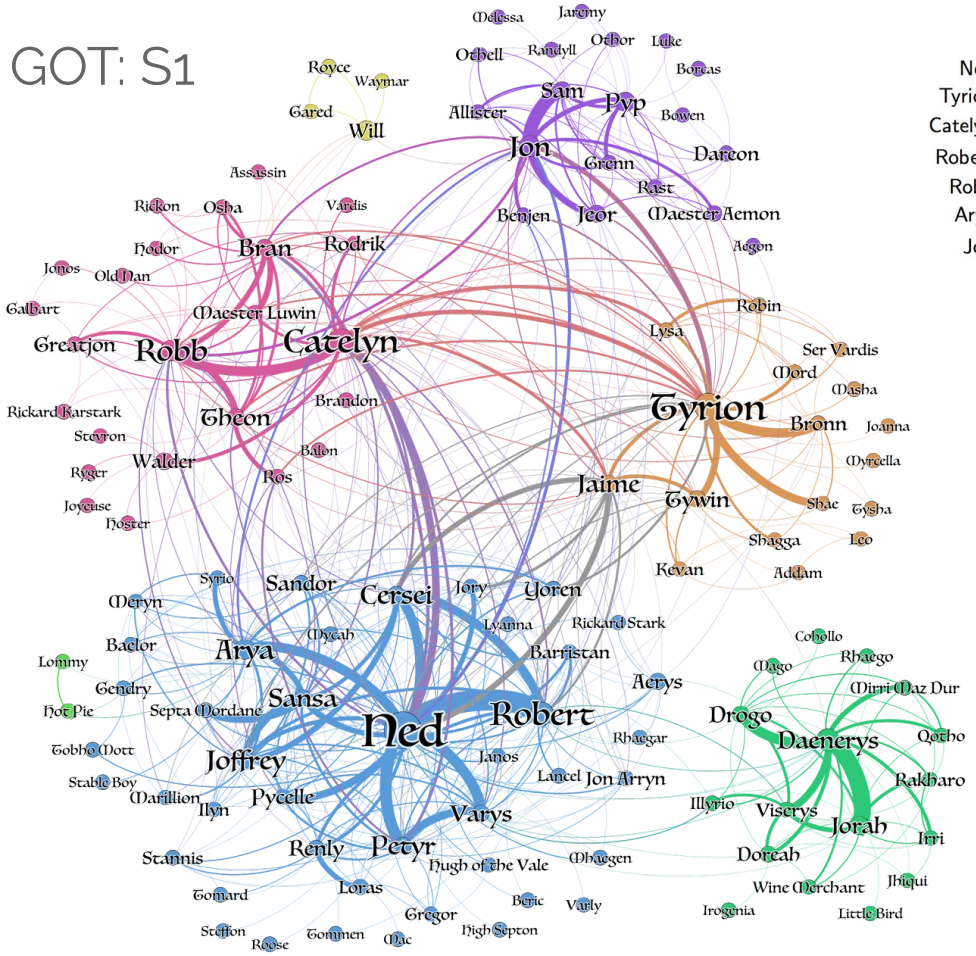
- E.g. If information does not spread through shortest paths, betweenness centrality is irrelevant

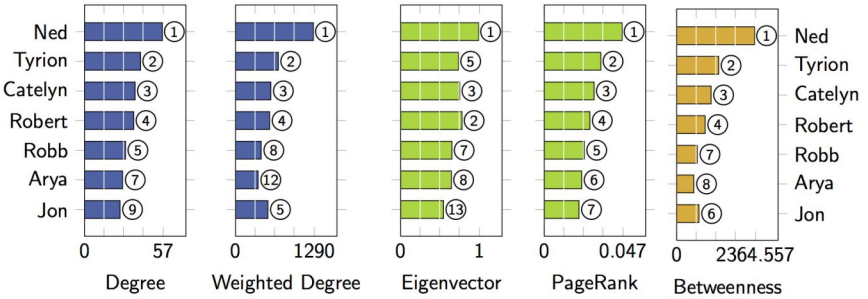
Centrality measures should be used with caution for (a) for exploratory purposes and (b) for characterisation

Understanding Centralities

Data and Viz @mathbeveridge
www.networkofthrones.wordpress.com

GOT: S1





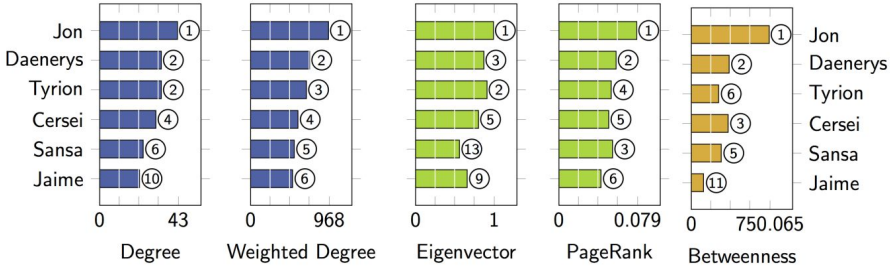
Node Label: PageRank
Node Size: Betweenness Centrality

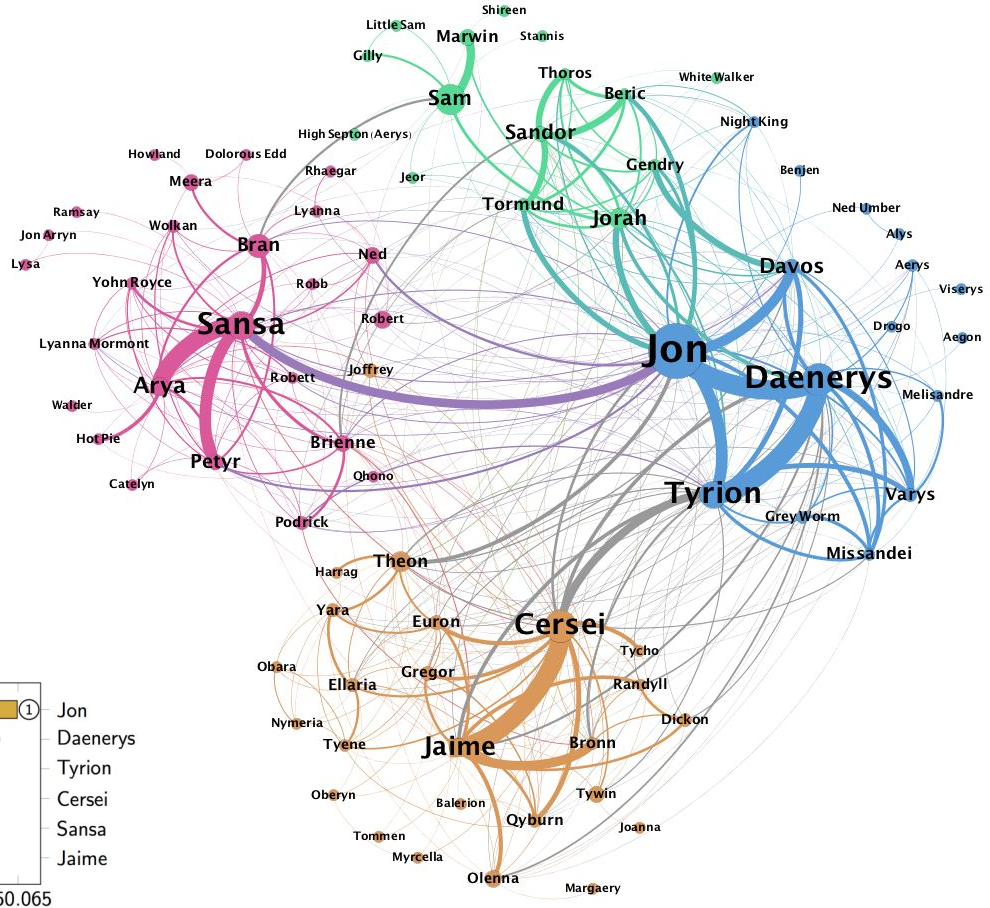
Edge Size: #interactions
Colors: Community (with Louvain)

GOT: S7

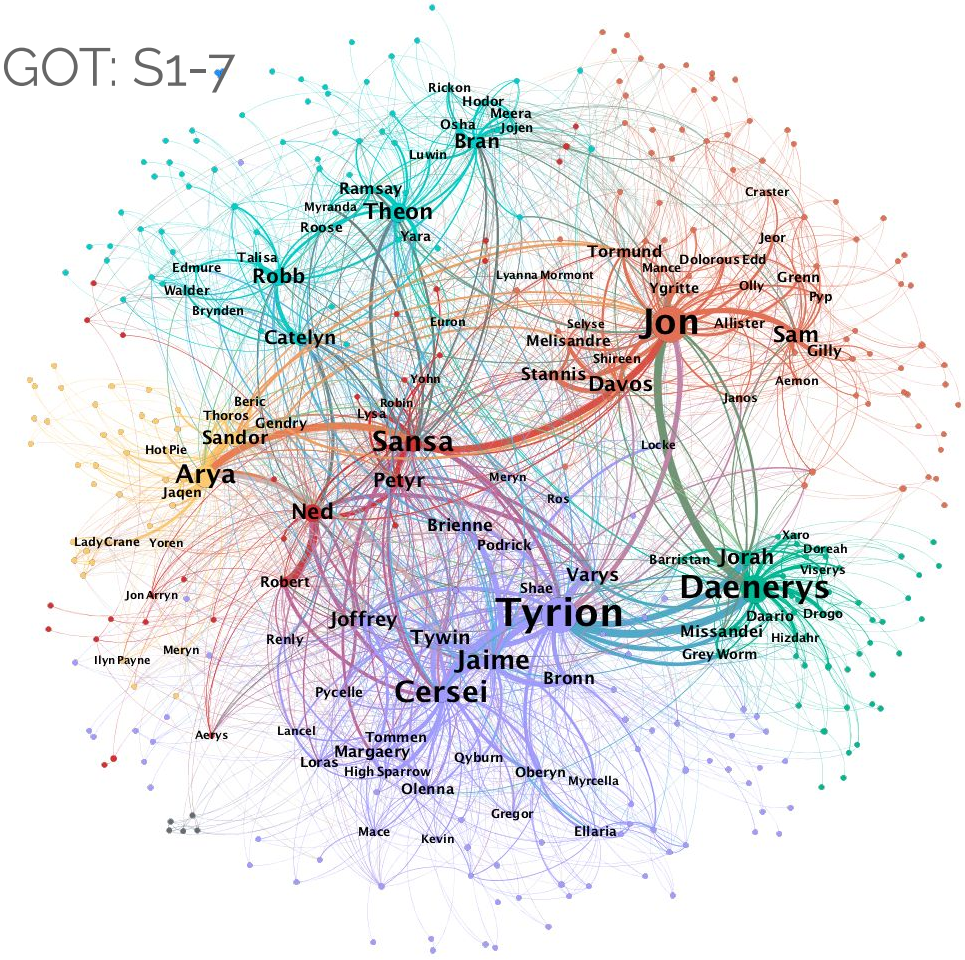
Node Label: PageRank
Node Size: Betweenness Centrality

Edge Size: #interactions
Colors: Community (with Louvain)





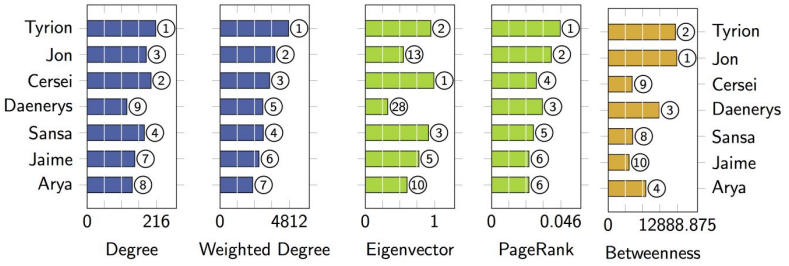
GOT: S1-7



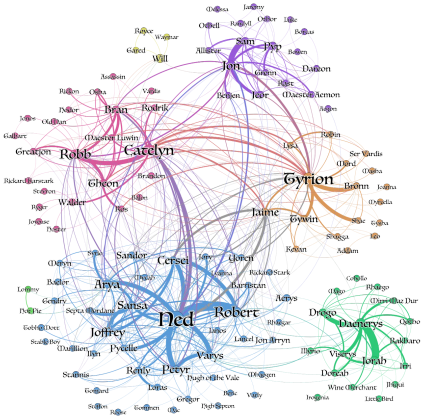
All characters interactions
(up to the last season... data are coming)

Node Label: PageRank
Node Size: Betweenness Centrality

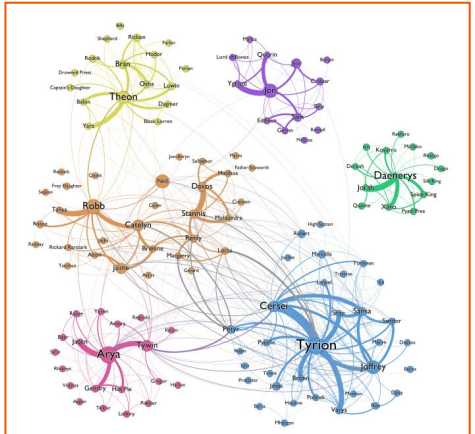
Edge Size: #interactions
Colors: Community (with Louvain)



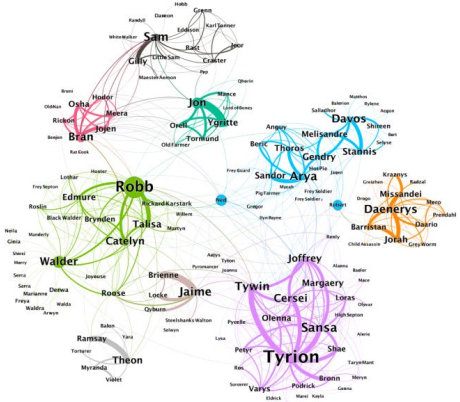
More on: www.networkofthrones.wordpress.com



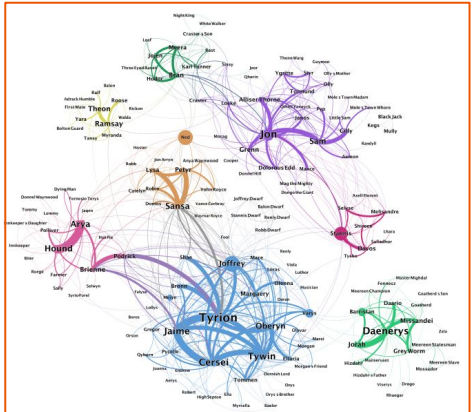
Season 1: Ned



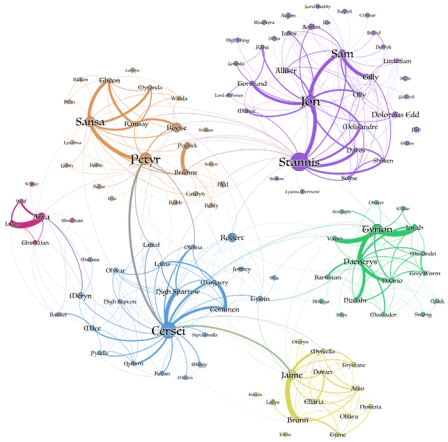
Season 2: Tyrion



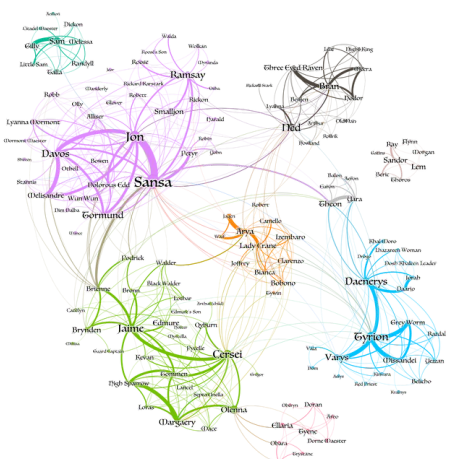
Season 3: Rob



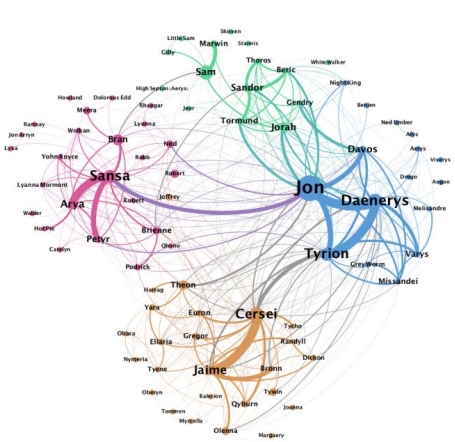
Season 4: Tyrion



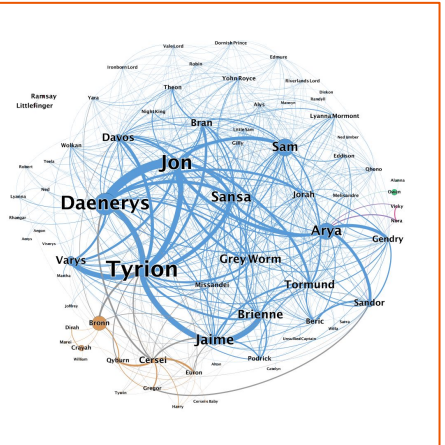
Season 5: Cersei



Season 6: Sansa



Season 7: Jon



Season 8: Tyrion

Do Birds of a Feather Flock Together?

Homophilic behaviors in complex networks

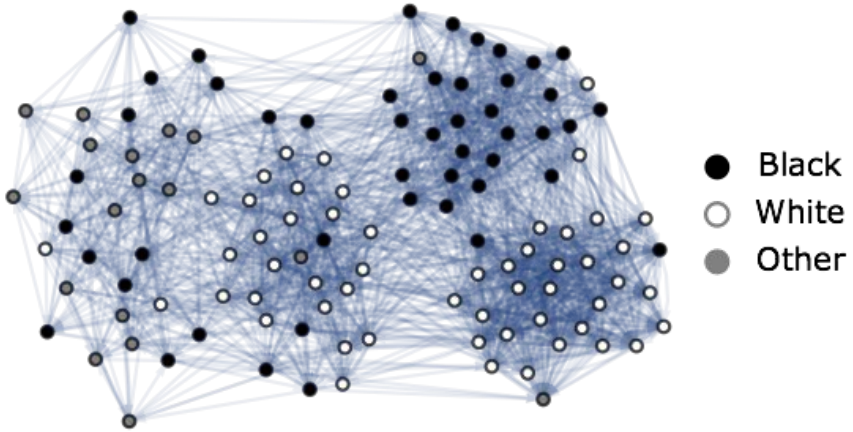
Homophily

Property of (social) networks that nodes of the same attitude tends to be connected with a higher probability than expected

- It appears as correlation between vertex properties of x(i) and x(j) if (i,j) ∈ E

Disassortative mixing:

Contrary of homophily: dissimilar nodes tend to be connected
(e.g., sexual networks, predator-prey)



Examples of Vertex properties

age, gender, nationality,
political beliefs, socioeconomic status,
obesity, ...

Homophily can be a link creation mechanism or consequence of social influence (and it is difficult to distinguish)

Assortative Mixing

(Newman's assortativity)

Quantify homophily while discrete node properties are involved

$$r = \frac{\sum_i e_{ii} - \sum_i a_i b_i}{1 - \sum_i a_i b_i}$$

where:

- e_{ij} fraction of links connecting nodes of type i and j
- a_i fraction of out-links from nodes of type a
- b_i fraction of in-links for type b nodes

Interpretation

- r=0: no assortative mixing
- r=1: perfectly assortative
- -1<r<0: disassortative mixing

		women				
		black	hispanic	white	other	a_i
men	black	0.258	0.016	0.035	0.013	0.323
	hispanic	0.012	0.157	0.058	0.019	0.247
	white	0.013	0.023	0.306	0.035	0.377
	other	0.005	0.007	0.024	0.016	0.053
	b_i	0.289	0.204	0.423	0.084	

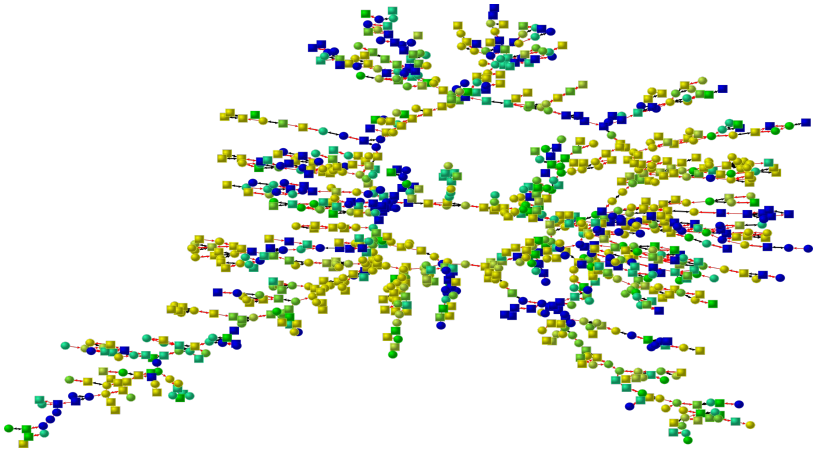
R = 0.621

Newman, Mark EJ. "Mixing patterns in networks." Physical Review E 67.2 (2003): 026126.

	network	type	size n	assortativity r
social	physics coauthorship	undirected	52 909	0.363
	biology coauthorship	undirected	1 520 251	0.127
	mathematics coauthorship	undirected	253 339	0.120
	film actor collaborations	undirected	449 913	0.208
	company directors	undirected	7 673	0.276
	student relationships	undirected	573	−0.029
	email address books	directed	16 881	0.092
technological	power grid	undirected	4 941	−0.003
	Internet	undirected	10 697	−0.189
	World-Wide Web	directed	269 504	−0.067
	software dependencies	directed	3 162	−0.016
biological	protein interactions	undirected	2 115	−0.156
	metabolic network	undirected	765	−0.240
	neural network	directed	307	−0.226
	marine food web	directed	134	−0.263
	freshwater food web	directed	92	−0.326

Assortativity by degree w.r.t. different network types

Case study: Happiness



James H. Fowler, Nicholas A. Christakis.
Dynamic Spread of Happiness in a Large Social Network:
Longitudinal Analysis Over 20 Years in the Framingham Heart Study
British Medical Journal 337 (4 December 2008)

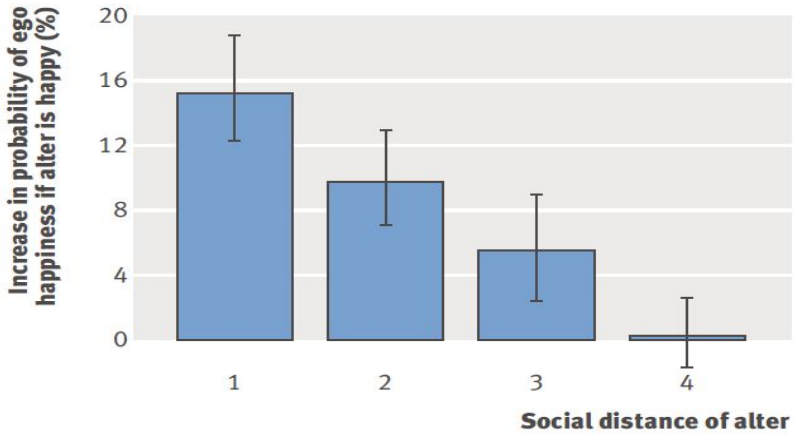
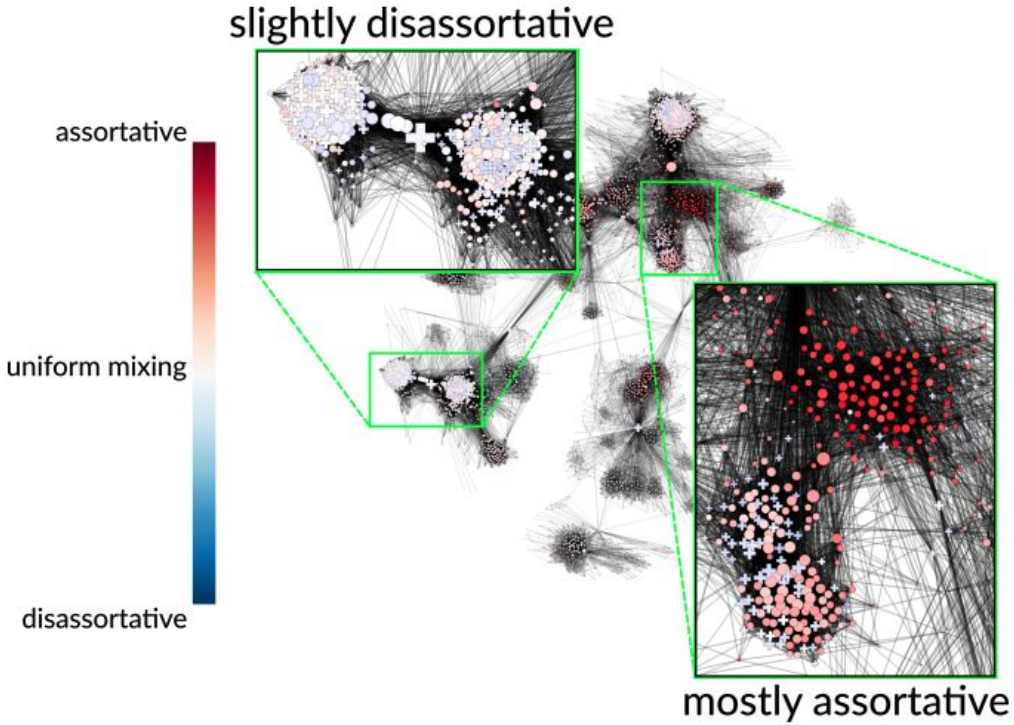


Fig 2 | Social distance and happiness in the Framingham social network. Percentage increase in likelihood an ego is happy if friend or family member at certain social distance is happy (instead of unhappy). The relationship is strongest between individuals who are directly connected but remains significantly >0 at social distances up to three degrees of separation, meaning that a person's happiness is associated with happiness of people up to three degrees removed from them in the network

Is a Global Measure enough?

“Sure I can work with the means, but I'd rather party with the outliers...”



Local assortativity of gender in a sample of Facebook friendships (McAuley and Leskovec 2012).

Different regions of the graph exhibit strikingly different patterns, suggesting that a single variable, e.g. global assortativity (Newman's), would provide a poor description of the system.

Limits of a global assortativity score

Multiscale Mixing Patterns

Idea:
A local measure that captures the mixing patterns within the local neighbourhood of a given node.

Trivial solution:
Consider only the node's neighbors
- issue with sample size (what about low degree nodes?)

Better approximation:
Considering the stable state of a RW (probability to reach a given node) to weight the edges

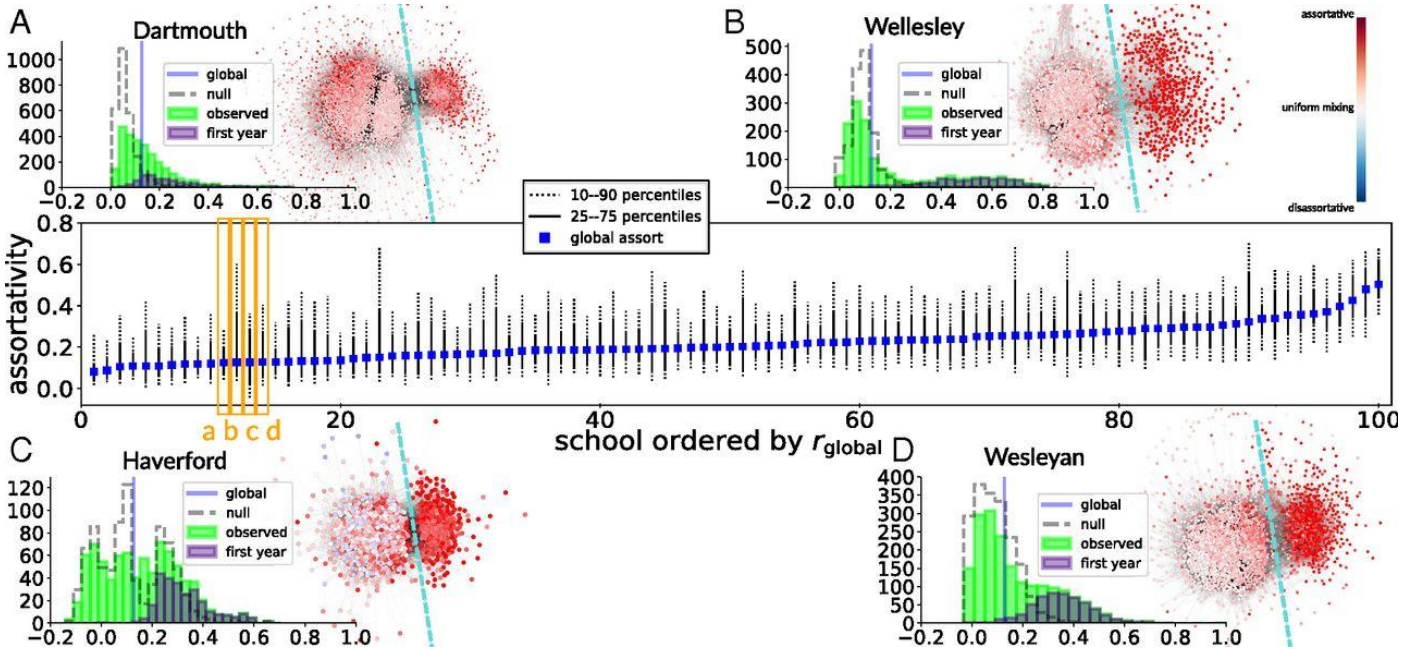
Issue:
Need to fix the value of α
- α=0 the RW stays put,
- α=1 the RW never restarts

Solution:
Integrate over all possible α (multiscale approach)

$$w_{\text{multi}}(i;\ell) = \int_0^1 w_\alpha(i;\ell)\,\mathrm{d}\alpha$$

Peel, Leto, Jean-Charles Delvenne, and Renaud Lambiotte. "*Multiscale mixing patterns in networks.*" *Proceedings of the National Academy of Sciences* 115.16 (2018): 4057-4062.

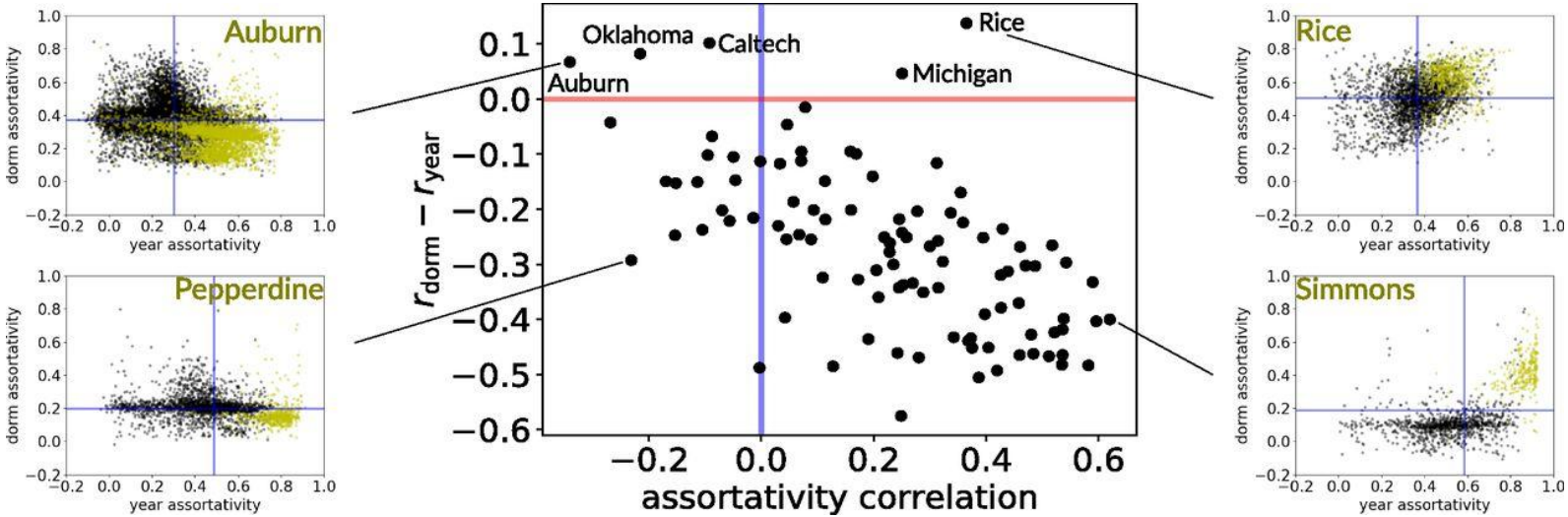
Evaluation on real data



Facebook100

Distribution of local assortativity for the "dorm" node feature

Evaluation on real data (cont'd)



Facebook100

Correlation of local assortativities by dorm and matriculation year (x axis)
and proportion of nodes which are more assortative by dorm than by year (y axis).

Chapter 5

Community Discovery

Summary

- What's a Community?
- Communities in static networks

Community Discovery

The aim of Community Discovery algorithms is to identify meso-scale topologies hidden within complex network structures

Why Community Discovery?

- "Cluster" homogeneous nodes relying on topological information

Major Problems:

- Community Discovery is an ill posed problem
  - Each algorithm models *different* properties of communities
- Different approaches comparison
- Context Dependency

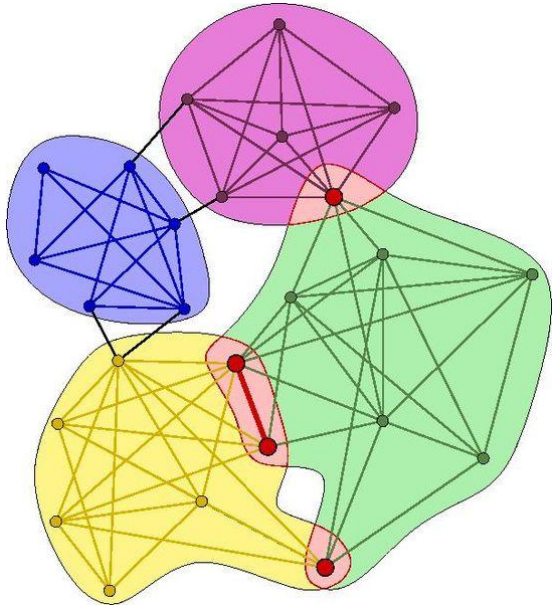
S. Fortunato. Community detection in graphs. Physics Reports 486 (2010).

Community Characteristics

Given the complexity of the problem a number of different typologies of approaches where proposed in order to:

Analyze:

- Directed\Undirected graphs
- Weighted\Unweighted graphs
- Multidimensional graphs
- …



Following:

- Top-Down\Bottom-Up partitioning
- …

Producing:

- Overlapping Communities
- Fuzzy Communities
- Hierarchical Communities
- Nested Communities
- …

But…what is it exactly a community?

Unfortunately does not exist a universally shared definition of what a community is…

A *general* idea is that a community should represent:

> *"A set of entities where each entity is closer, in the network sense, to the other entities within the community than to the entities outside it."*

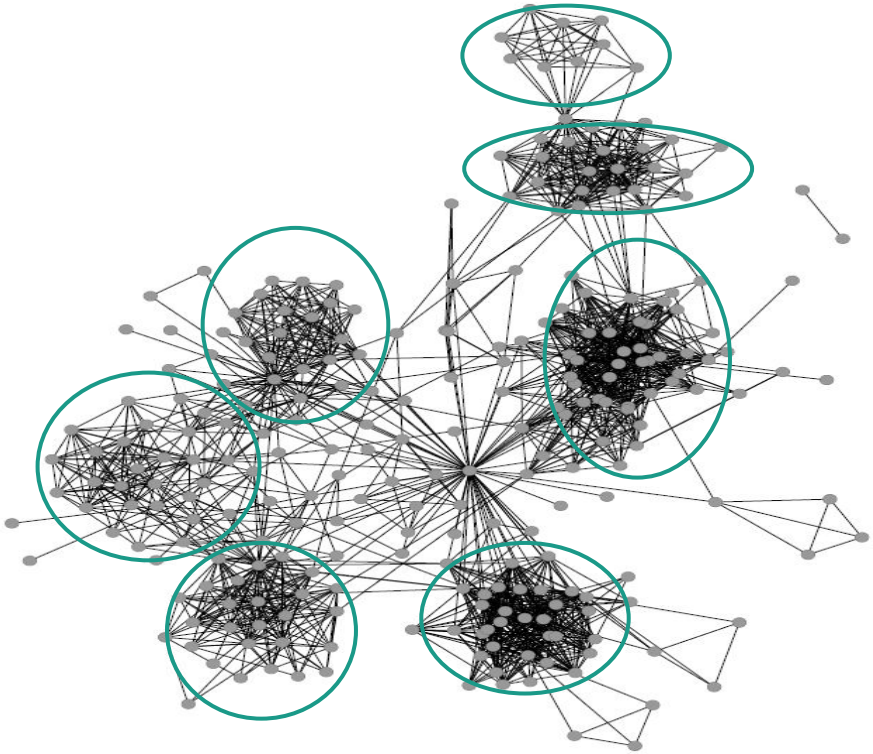
or, equivalently

> *"A set of nodes tightly connected within each other than with nodes belonging to other sets."*

Communities in Complex Networks

In simple, small, networks it is easy identify them by looking at the structure…

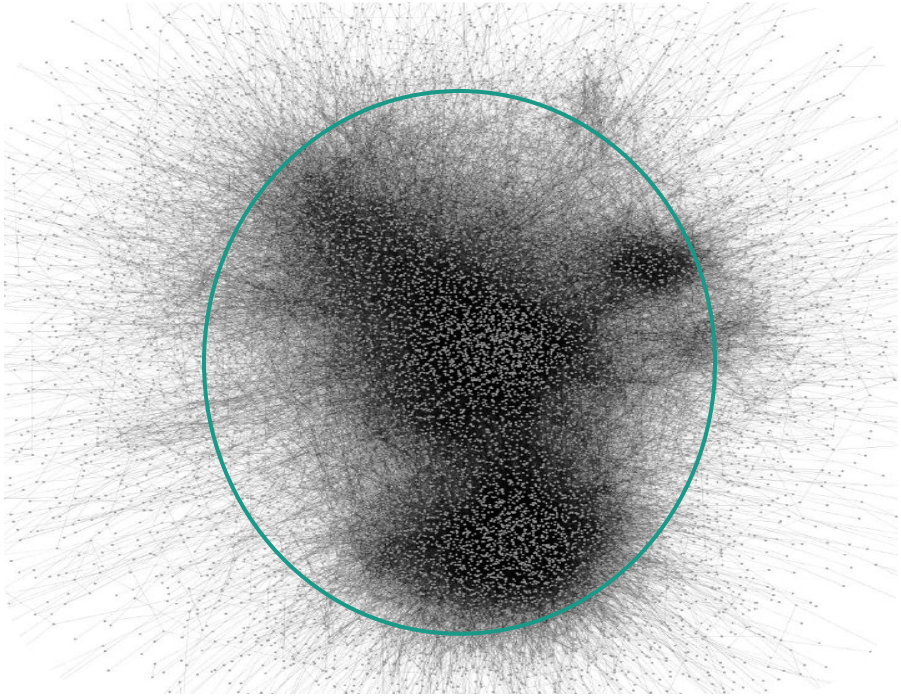
- i.e., using a Force directed layout



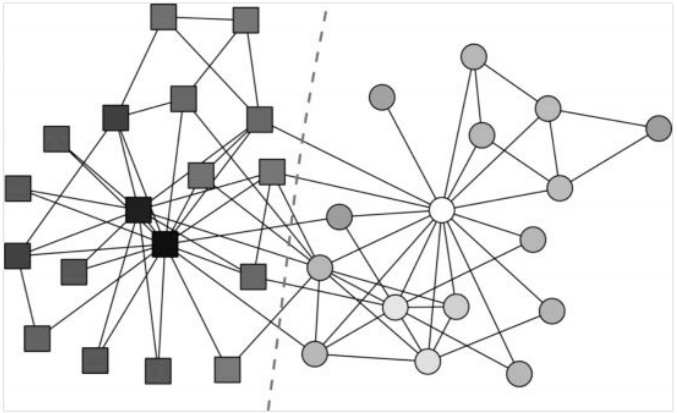
Real world networks? Too complex for visual analysis

We can't easily identify (e.g., visually) different communities

We need automated procedures!



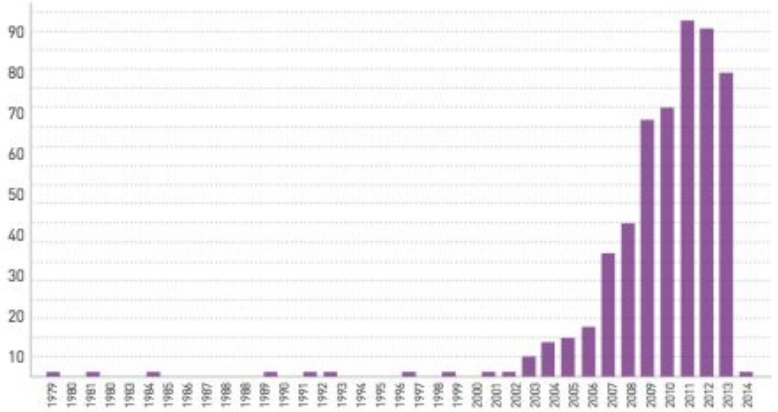
A first example…



Zachary's Karate Club

Communities emerge from the breakup of the Club

Citation history of the Zachary's Karate Club paper

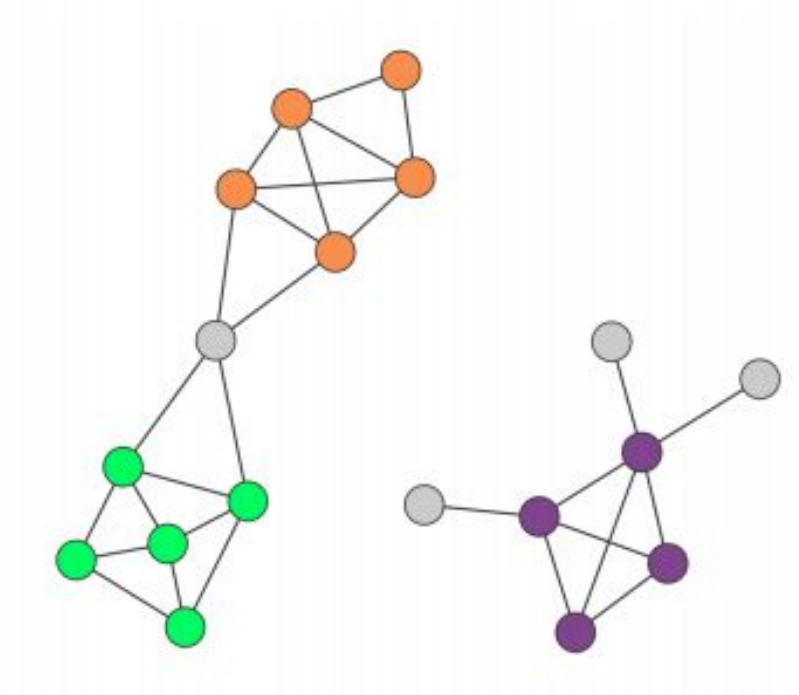


Karate Club Trophy

http://networkkarate.tumblr.com/

Communities: a few Hypothesis

- **H1**: The community structure is uniquely encoded in the wiring diagram of the overall network
- **H2**: A community corresponds to a connected subgraph
- **H3**: Communities are locally dense neighborhoods of a network



Taxonomy

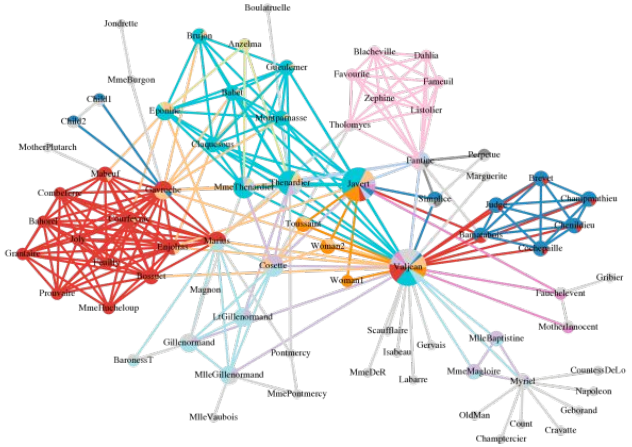
Girvan-Newman

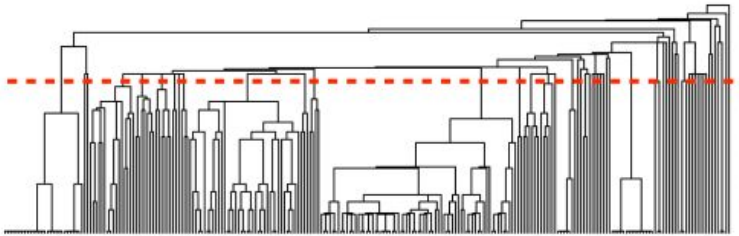
Steps

1. Compute the betweenness of all existing edges in the network;
2. Remove the edge(s) with the highest betweenness;
3. Recompute the betweenness for all edges;
4. Repeat steps 2 and 3 until no edges remain.

The end result of the Girvan–Newman algorithm is a dendrogram.

The leaves of the dendrogram are individual nodes.





Taxonomy

Louvain

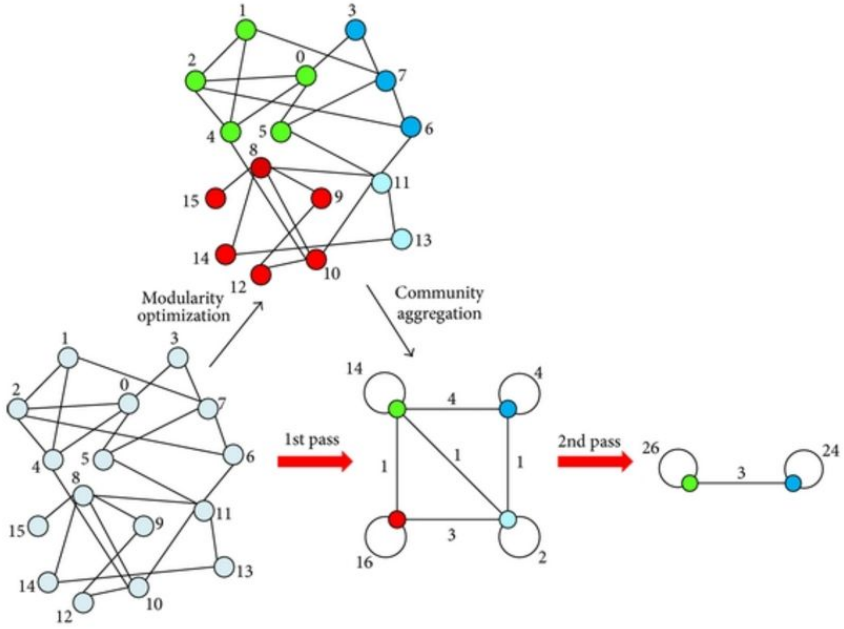
In order to maximize this value efficiently, the Louvain Method has two phases that are repeated iteratively.

Initialization:
Each node in the network is assigned to its own community.

- Phase 1:
  Each node is then moved into the adjacent community that guarantee the greatest modularity increase.

- Phase 2:
  A new graph is created: its nodes are the updated communities and weighted links connect them accounting for bridges in the original graph.

Phases 1 and 2 are repeated until modularity is maximized



VD Blondel, et al. Fast unfolding of communities in large networks. Journal of statistical mechanics: theory and experiment (2008)

Taxonomy

Infomap

The core of the algorithm follows closely the Louvain method:

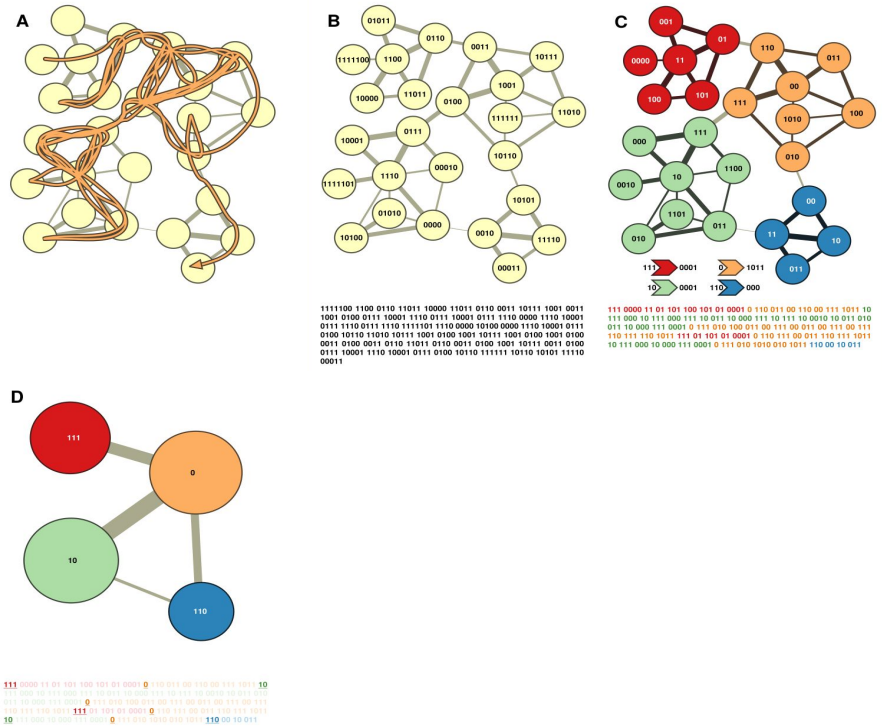
- Phase 1:
  Each node is moved to the neighboring module that results in the largest decrease of the map equation.

- Phase 2:
  The network is rebuilt, with the modules of the last level forming the nodes at this level.
  This hierarchical rebuilding of the network is repeated until the map equation cannot be reduced further.

Implicit optimization of the Conductance measure: $\phi(G) = \min_{S \subseteq V} \varphi(S)$

Where:

- $\varphi(S) = \frac{\sum_{i \in S, j \in \bar{S}} a_{ij}}{\min(a(S), a(\bar{S}))}$ is the conductance for a cut
- $(S, \bar{S})$ is a cut, and
- $a(S) = \sum_{i \in S} \sum_{j \in V} a_{ij}$



Taxonomy

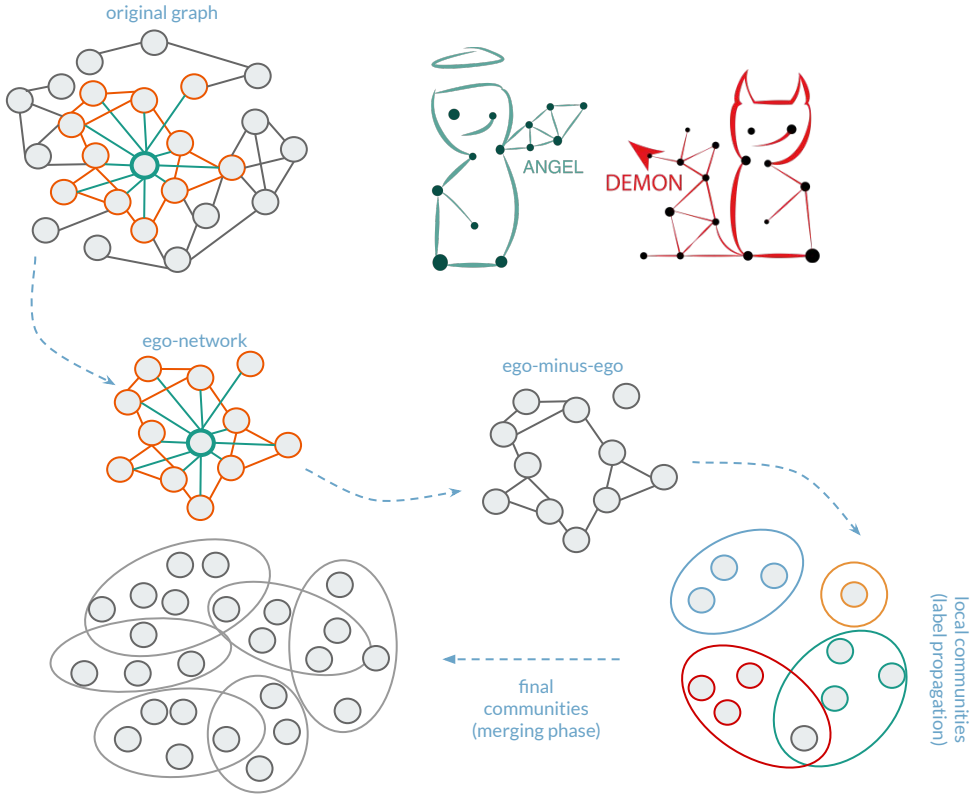
Demon/Angel

For each node n:

1. Extract the Ego Network of n
2. Remove n from the Ego Network
3. Perform a Label Propagation
4. Insert n in each community found
5. Update the raw community set C

For each local community c in C

6. Merge with “similar” ones in the set (given a threshold)
   (i.e. merge iff at most the ε% of the smaller one is not included in the bigger one)



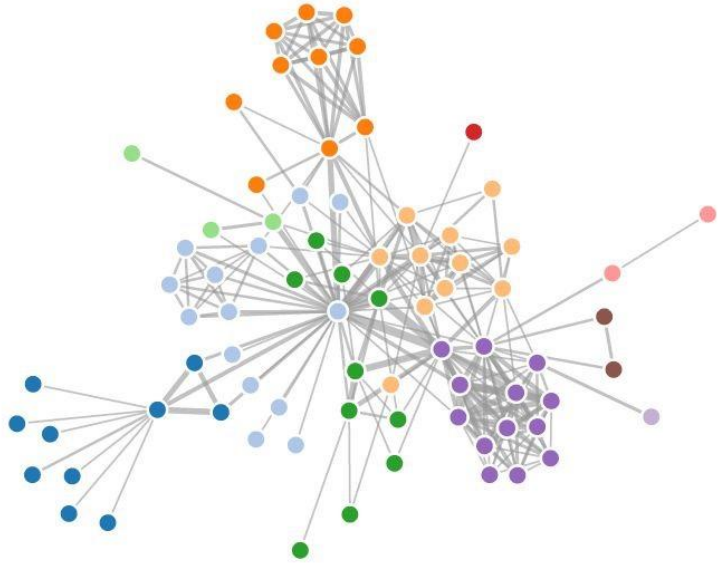
Community Discovery is, perhaps, the hottest topic in complex network analysis

Major issues:

- Problem definition
- Community evaluation

Problem specializations:

- Evolutionary Community Discovery
  (How do communities evolve in dynamic networks?)
- Multidimensional Community Discovery
- …



Chapter 6

Link Prediction

Summary

- Predicting Network Evolution
- Unsupervised approaches
- Supervised approaches
- Evaluation

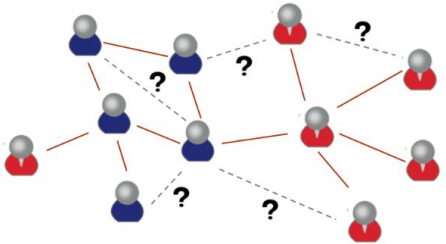
Link Prediction

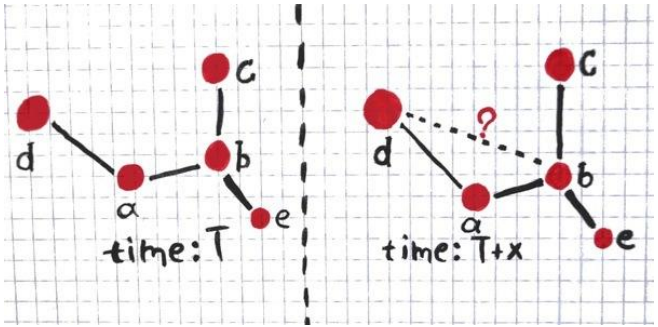
Goal

Understanding how networks evolve

Problem definition

Given a snapshot of a network at time **t**, (accurately) predict the edges that will appear in the network during the interval **(t, t+1)**





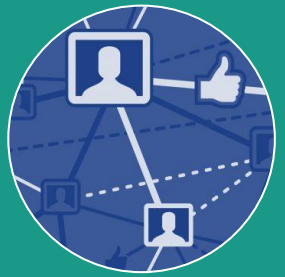
Liben-Nowell, David, and Jon Kleinberg. "The link-prediction problem for social networks." Journal of the American society for information science and technology 58.7 (2007): 1019-1031.

Examples of uses of

Link Prediction

Monitor terrorist networks – deducing possible interactions/missing links between terrorists (without direct evidence)

Suggest interactions or collaborations that haven't yet been exploited within an organization

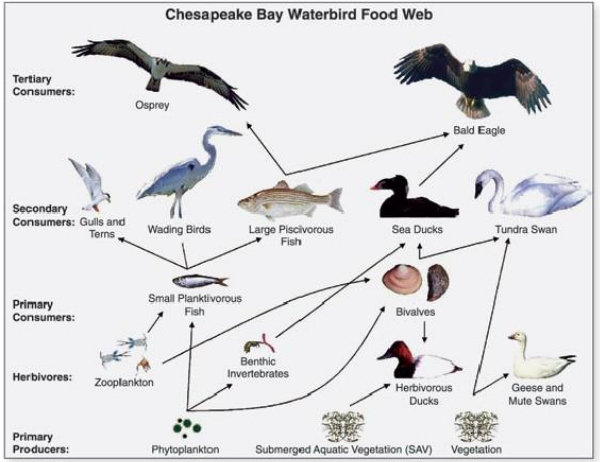


Friendship prediction (i.e., Facebook)

Link Prediction

Link prediction is used to predict future possible links in the network (e.g., Facebook).

Or, it can be used to predict missing links due to incomplete data (e.g., Food-webs)



PLOS ONE

RESEARCH ARTICLE

Link Prediction in Criminal Networks: A Tool for Criminal Intelligence Analysis

Giulia Berlusconi[1], Francesco Calderoni[1]*, Nicola Parolini[2], Marco Verani[2], Carlo Piccardi[3]*

1 Università Cattolica del Sacro Cuore and Transcrime, Milano, Italy, 2 MOX, Department of Mathematics, Politecnico di Milano, Milano, Italy, 3 Department of Electronics, Information and Bioengineering, Politecnico di Milano, Milano, Italy

* francesco.calderoni@unicatt.it (FC); carlo.piccardi@polimi.it (CP)

Concretizing an Intuition...

Scientists who are close in the network
(i.e., have common colleagues)

→ will likely collaborate in the future

Goal:

- make this intuitive notion precise and understand which measures of "proximity" leads to accurate predictions

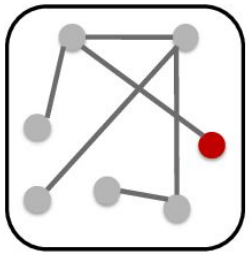
Link Prediction

Workflow

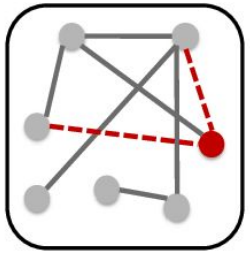
1. Consider as input a graph G at time t
2. Consider all the possible pairs of nodes (u,v)
3. Compute a link formation scores:

   score(u,v)
4. Build a list of all possible edges ordered by scores (from highest to lowest)
5. Verify, following that ordering, the predictions on the graph at time t+1

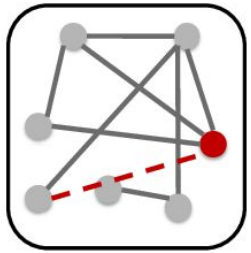
score is a measure of *proximity*



Time t



Time t+a



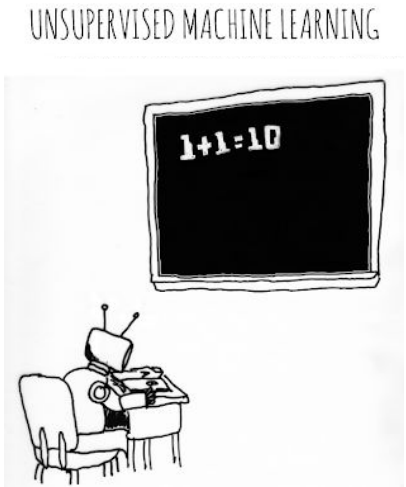
Time t+2a

Link Prediction

Approaches

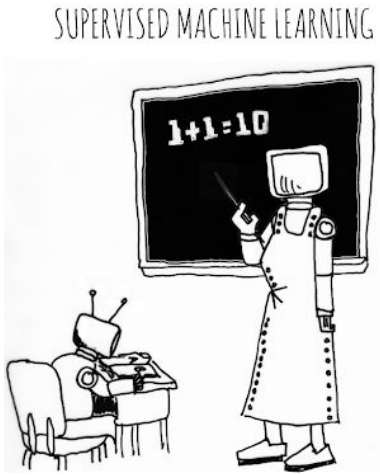
Unsupervised

Define a set of proximity measures unrelated to the particular network



Supervised

Extract knowledge from the network in order to improve prediction accuracy



Unsupervised Link Prediction

Unsupervised Link Prediction

Unsupervised measurements rely on different structural properties of networks

Neighborhood measures

- Common Neighbors, Adamic Adar, Jaccard, Preferential Attachment

Path-based measures

- Graph distance, Katz

Ranking

- Sim Rank, Hitting time, Page Rank

Liben-Nowell, David, and Jon Kleinberg. "The link-prediction problem for social networks." Journal of the American society for information science and technology 58.7 (2007): 1019-1031.

Unsupervised Link Prediction

Neighborhood measures

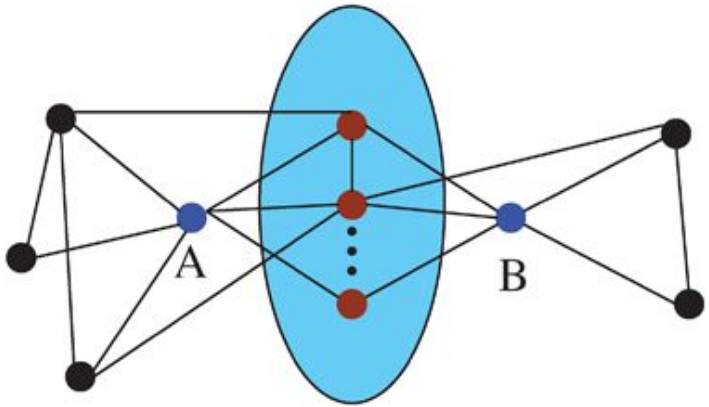
How many friends we have to share in order to become friends?

Common Neighbors:
the more friends we share,
the more likely we will become friends

$$\text{score}\,(u, v) = |\Gamma(u) \cap \Gamma(v)|$$

Jaccard:
the more similar our friends circles are,
the more likely we will become friends

$$\text{score}\,(u, v) = \frac{|\Gamma(u) n \Gamma(v)|}{\Gamma(u) \cup r(v)|}$$



Unsupervised Link Prediction

Neighborhood measures

How many friends we have to share in order to become friends?

Adamic Adar:

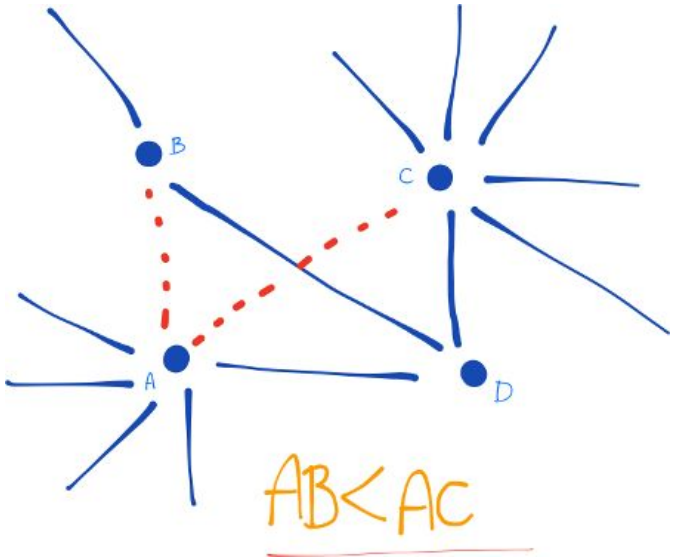
the more selective our mutual friends are,
the more likely we will become friends

$$\text{score}\ (u, v) = \sum_{z \in \Gamma(u) \cap \Gamma(v)} \frac{1}{\log(|\Gamma(z)|)}$$

Preferential Attachment:

the more friends we have,
the more likely we will become friends

$$\text{score}\ (u, v) = |\Gamma(u)| * |\Gamma(v)|$$



Unsupervised Link Prediction

Limits

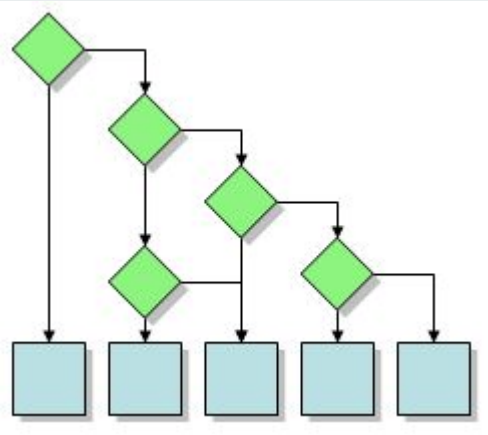
- Different kinds of networks are described by the same general closed formula
- An average overall performance between 6% and 12%

Measure comparison

- No single winner
- Almost all predictors outperform the random predictor
  ⇒ there is useful information in network topology

Supervised Link Prediction

Supervised Link Prediction



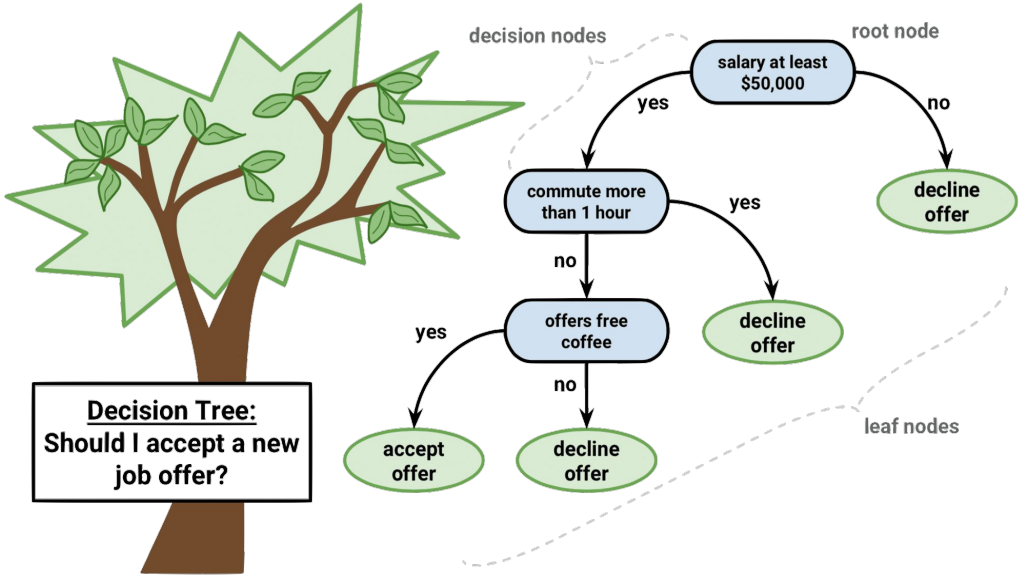
The process is now organized in 4 steps:

1. Split the data in train/test
2. **Learning a model on the train**
3. **Use the model for prediction**
4. Compare the prediction with the test

A natural way to do it:
build a "*classifier*" over a set of *network features*.

Staking Unsupervised Scores

Learn a Classifier (i.e., a Decision Tree) over unsupervised LP scores to generalize the assumption they made on the network growth model



Evaluation: ROC and PR curve

Precision Vs. Recall

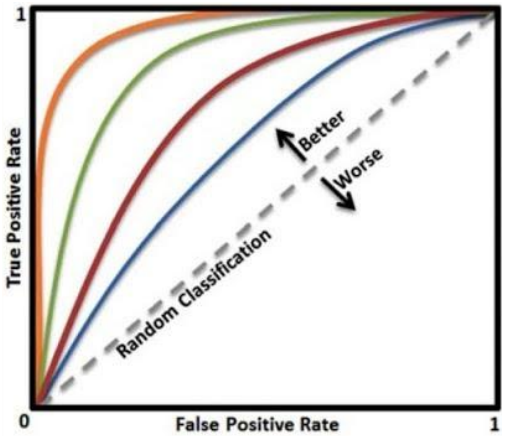
- Precision: PPV = TP/(TP+FP)
- Recall: TPR = TP/(TP+FN)

ROC (Receiver operating characteristic)

- 1-Specificity: FPR = FP/(FP+TN)
- Recall: TPR = TP/(TP+FN)

Note:

- ROC and PR spaces are isomorphic (the use of ROC is more widespread)
- Numerical comparison can be done using the AUROC (area under the ROC curve)



	p'	n'
p	TP	FN
n	FP	TN

Confusion Matrix

Link Prediction

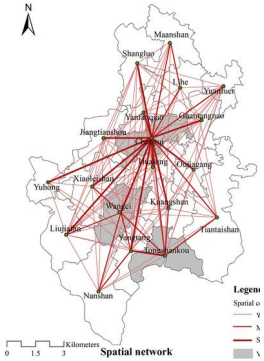
Something more...

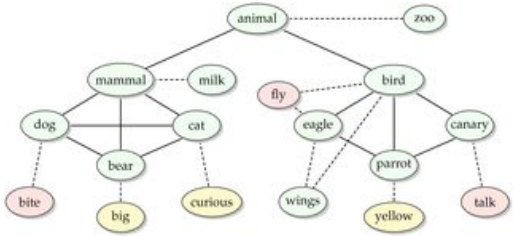
Accuracy could be improved extending simple models with more complex (even semantic) informations:

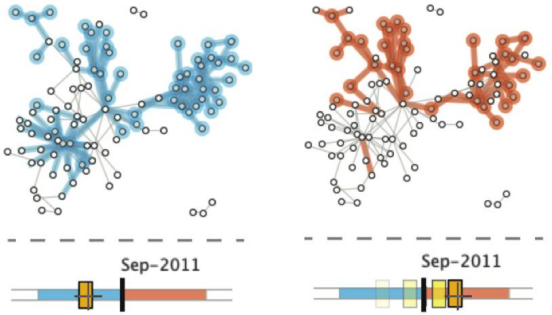
- Link strength
- Geographical information
- ...

Link Prediction needs to be revised while in some scenarios:

- **Dynamic Networks**
- Multiplex networks
- ...







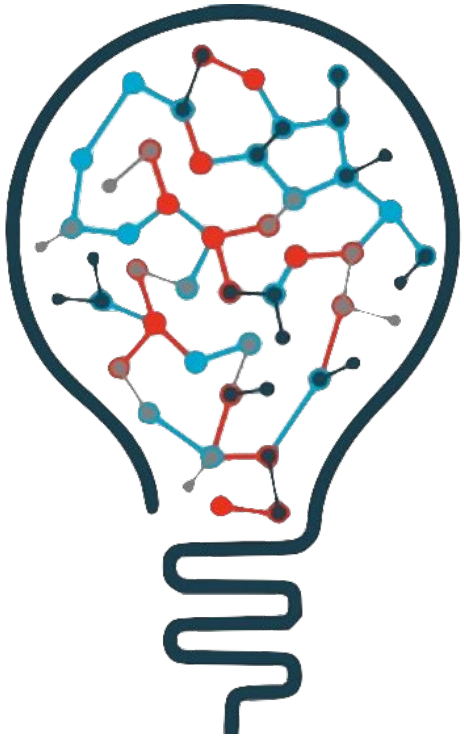
Key Messages

Predict new links that will arise in a network is **not easy**:

1. Networks are, usually, **sparse**
2. **Cold Start** Problem
    - What if I don't have enough information?
        - *Can I predict bridges?*

3. Huge **False Positive** prediction
    - *Bridges* !?!

4. Simple approaches are **"too simple"**
5. Complex approaches are **costly**



Chapter 7

Diffusion: Epidemics

Summary

- Probabilistic Epidemic Models
  - SI/SIS/SIR

Epidemic

Biological:
- Airborne diseases (flu, SARS, …)
- Venereal diseases (HIV, …)
- Other infectious diseases (HPV, …)
- Parasites (bedbugs, malaria, …)

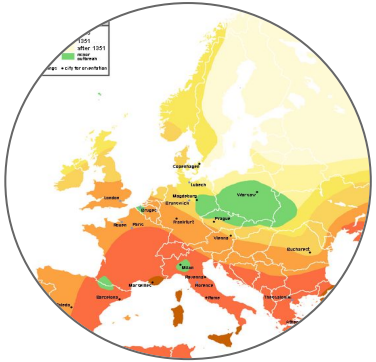
Digital:
- Computer viruses, worms
- Mobile phone viruses

Conceptual/Intellectual:
- Diffusion of innovations
- Rumors
- Memes
- Business practices

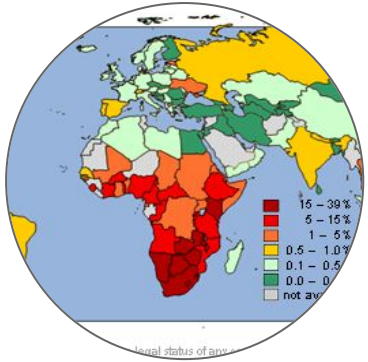
Epi + demos

upon *people*

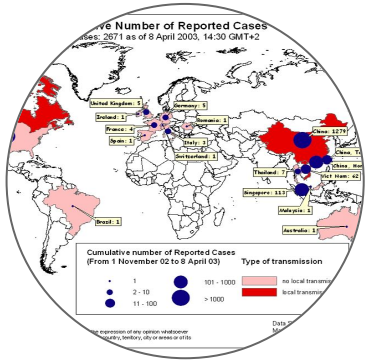
The Great Plague



HIV

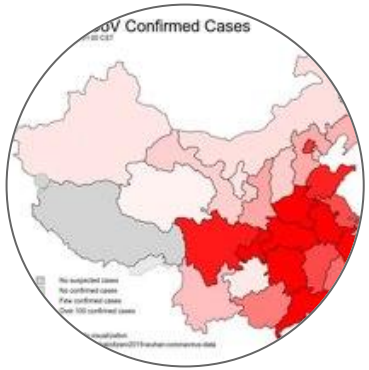


SARS



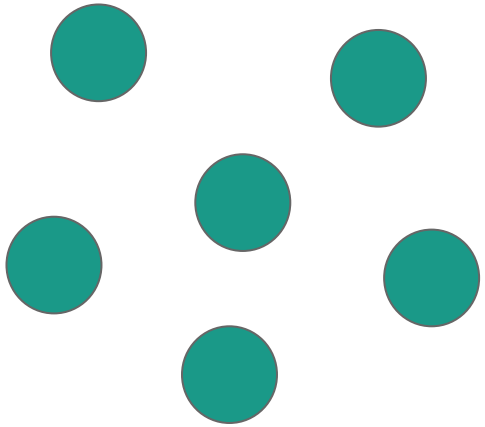
1918 Spanish flu

H1N1 flu

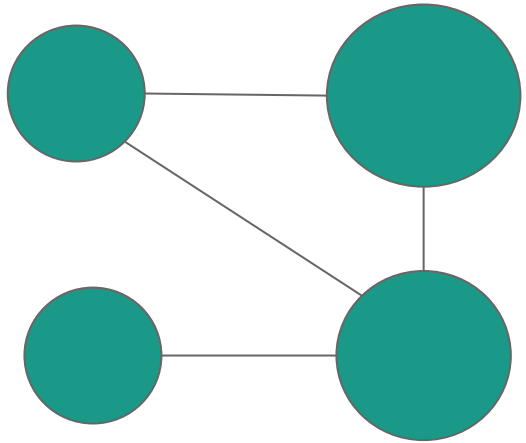


2019-nCoV

Biological: Notable Epidemic Outbreaks



Separate, small population
(hunter-gatherer society, wild animals)



Connected, highly populated areas
(cities)

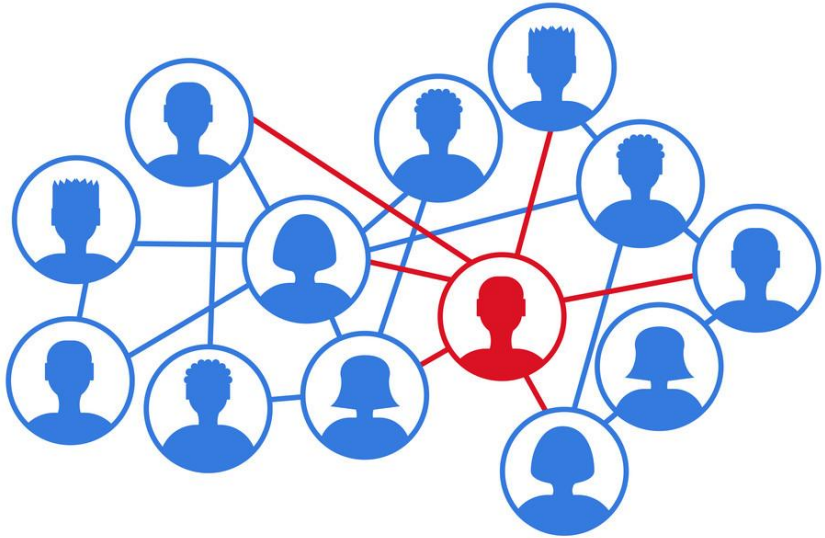
Human societies have "**crowd diseases**", which are the consequences of large, interconnected populations
(Measles, tuberculosis, smallpox, influenza, common cold, …)

Large population can provide the "fuel"

Topology matters

The described approaches assumed *homogenous mixing*, which means that each individual can infect *any* other individual.

In reality, epidemics spread along *links in a network:* we need to explicitly account for the role of the network in the epidemic process.



Probabilistic Epidemic Models

Compartmental Models

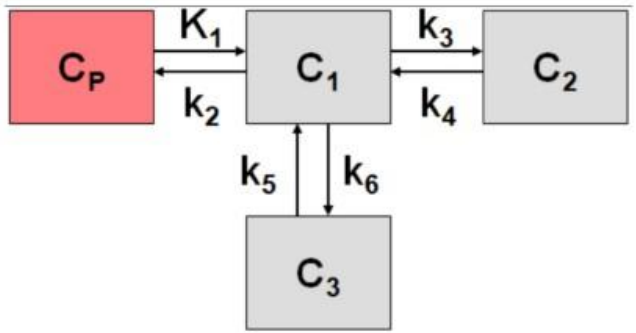
The framework is based on two hypotheses:

Compartmentalization:

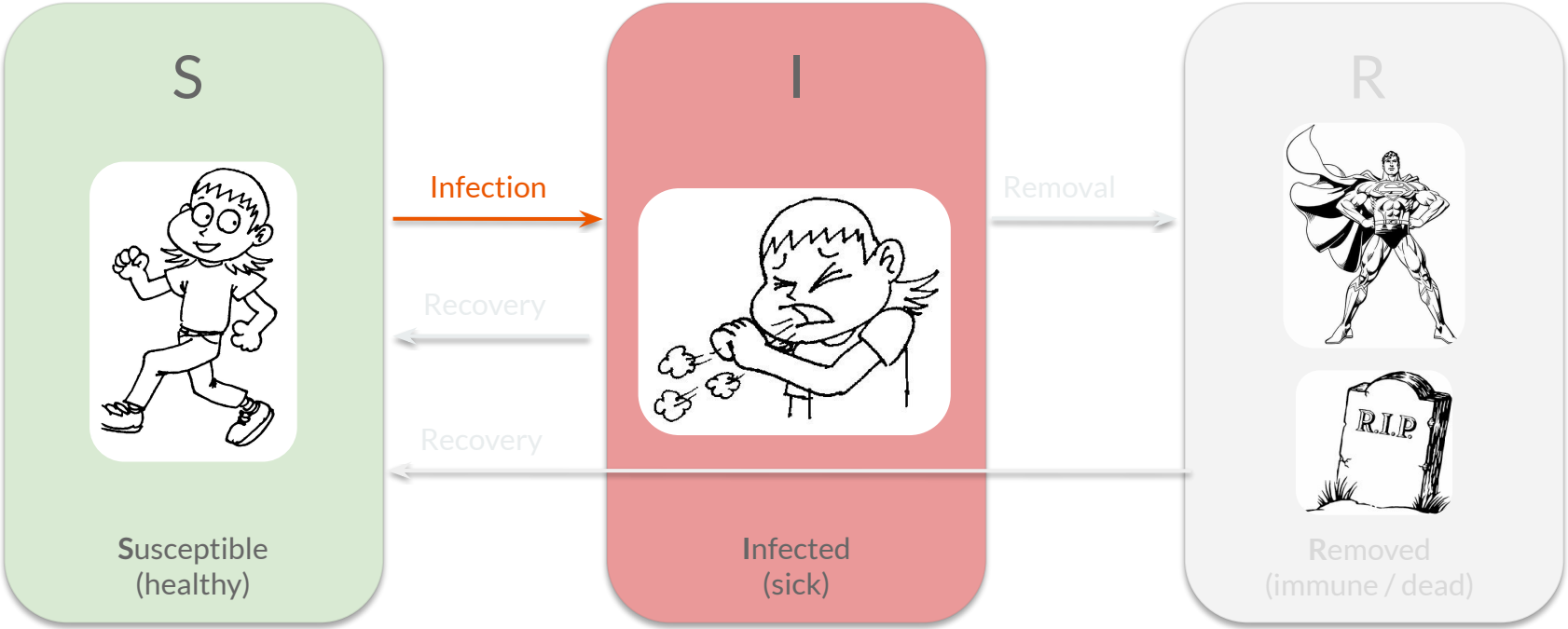
each individual is classified into distinct statuses. The simplest classification assumes that an individual can be in one of the states.

Homogeneous Mixing:

each individual has the same chance of coming into contact with an infected individual.



W. O. Kermack and Ag McKendrick. A Contribution to the Mathematical Theory of Epidemics. 1927



SI: The simplest model

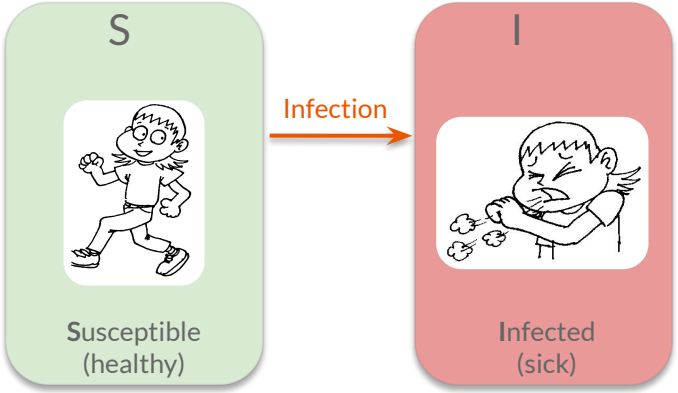
SI model

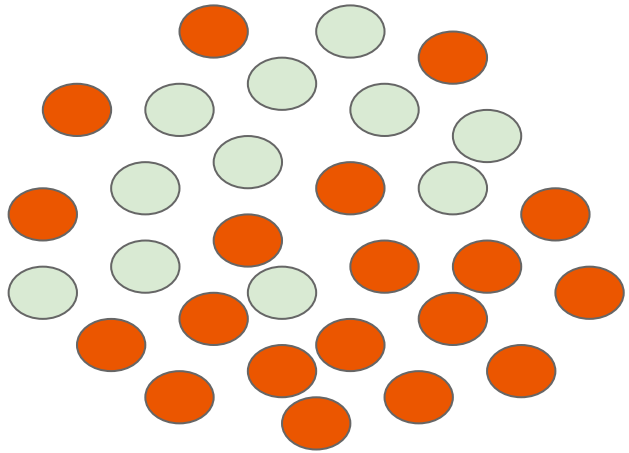
Each individual has β contacts with randomly chosen others individuals per unit time.

If there are *I* infected individual and *S* susceptible individuals, the average rate of new infection is βsi/N

with $s = S/N, \qquad i = I/N$

$$\frac{di}{dt} = \beta si = \beta i(1-i)$$

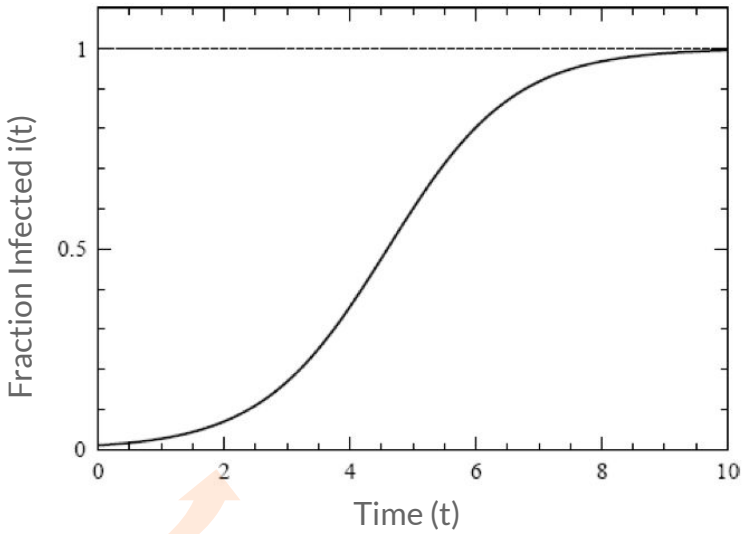




SI model

Behaviour

$$i(t) = \frac{i_0 \exp(\beta t)}{1 - i_0 + i_0 \exp(\beta t)}$$



If *i(t)* is small,

$$\frac{di}{dt} \approx \beta i$$

$$i \approx i_0 \exp(\beta t)$$

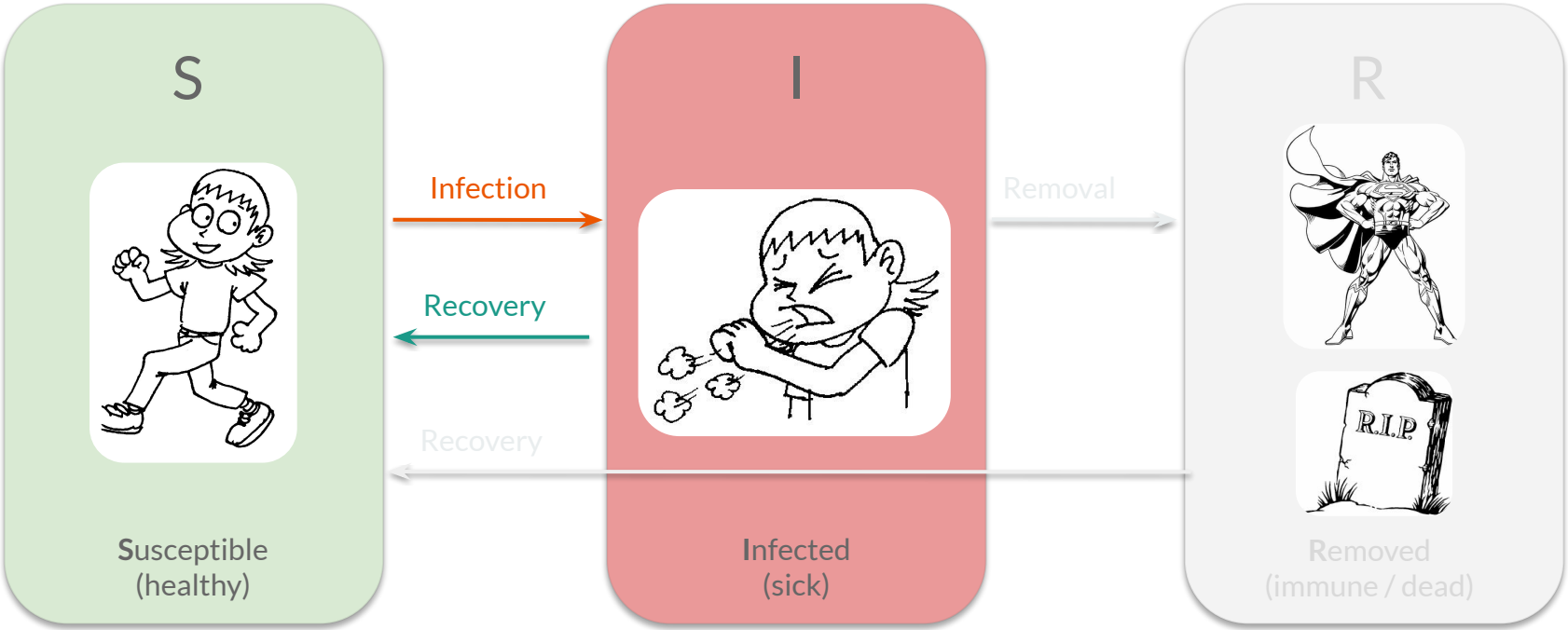
exponential outbreak

As i(t)~ 1.

$$\frac{di}{dt} \to 0$$

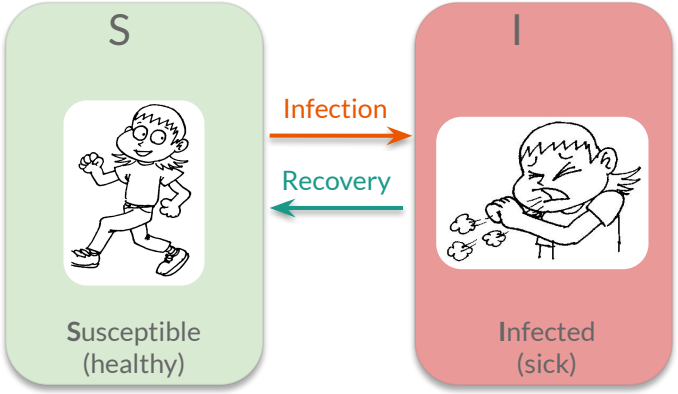
saturation

SI model:
the fraction infected increases until everyone is infected.



SIS: Common Cold

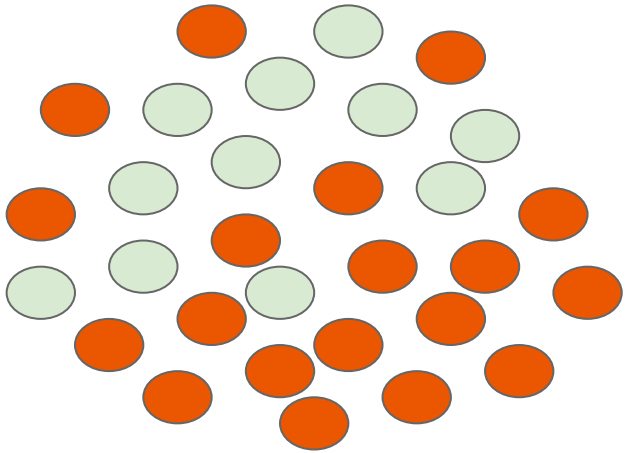
SIS model

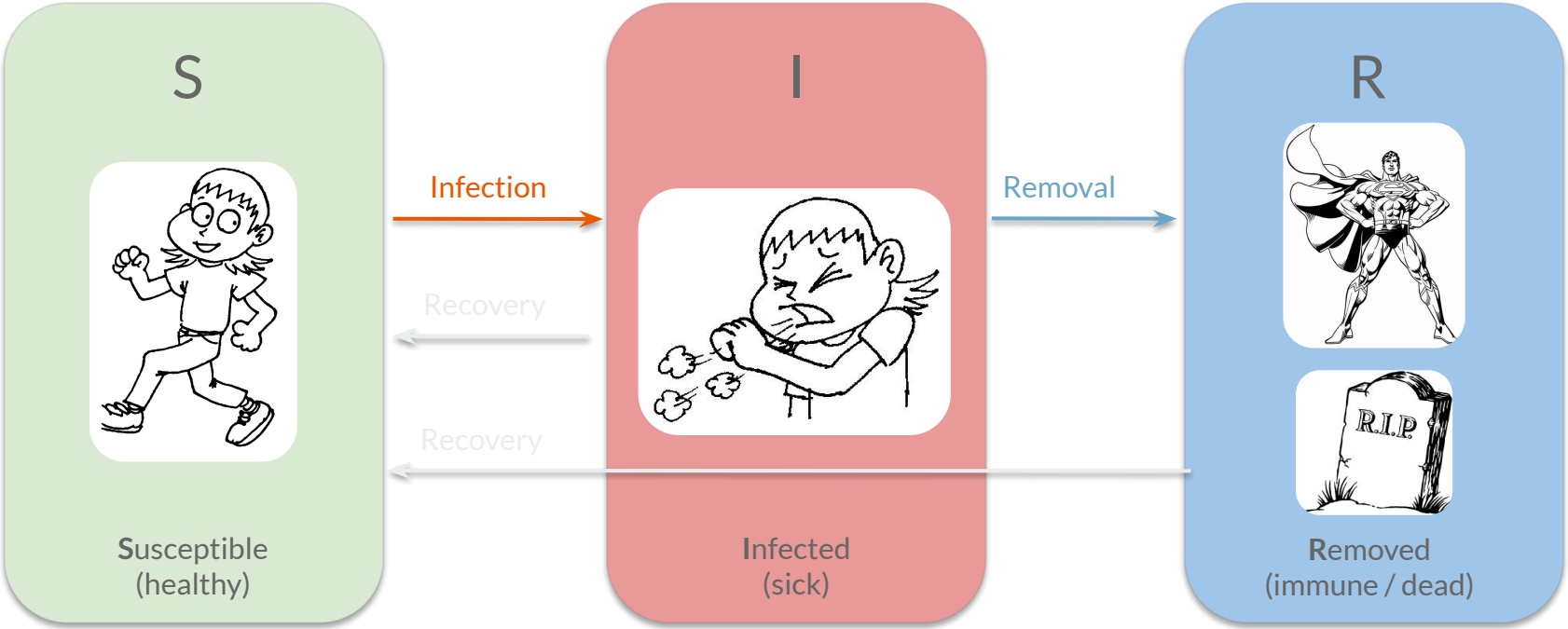


Modeling Common Cold

Each individual has β contacts with randomly chosen others individuals per unit time.

Each infected individual has μ probability of revert its status to susceptible





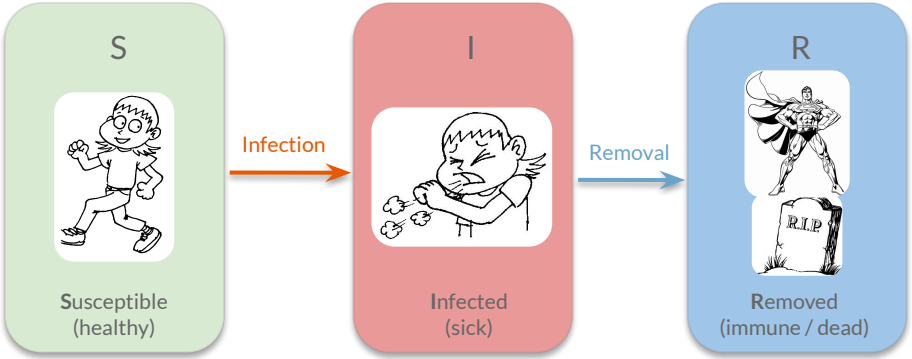
SIR: Flu, SARS, Plague

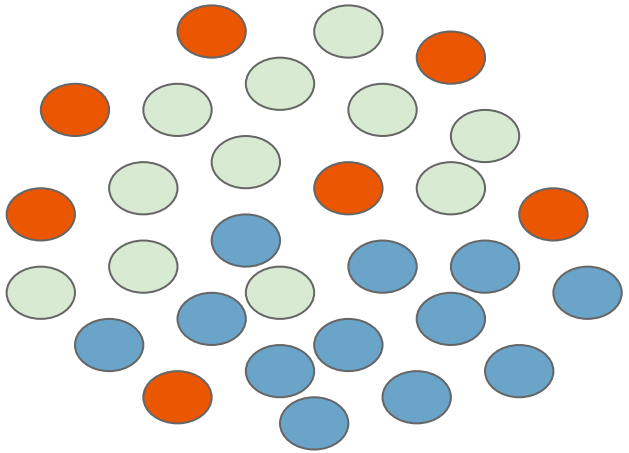
SIR model

Modeling Flu-like disease

Each individual has β contacts with randomly chosen others individuals per unit time.

Each infected individual has μ probability of becoming immune after being infected



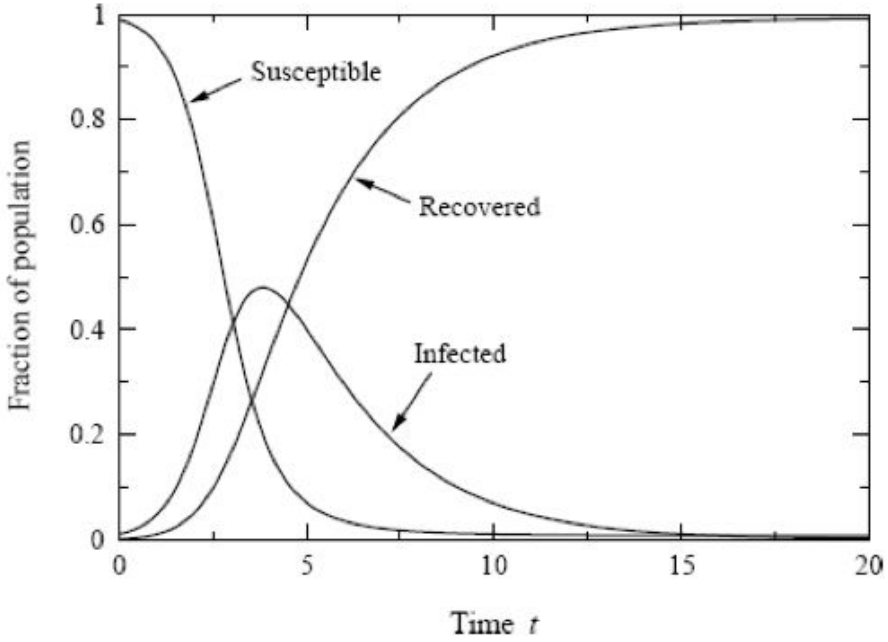


SIS model

Behaviour

$$\frac{\mathrm{d}s(t)}{\mathrm{d}t} = \beta\langle k\rangle i(t)\,[1 - r(t) - i(t)]$$
$$\frac{\mathrm{d}i(t)}{\mathrm{d}t} = -\mu i(t) + \beta\langle k\rangle i(t)\,[1 - r(t) - i(t)]$$
$$\frac{\mathrm{d}r(t)}{\mathrm{d}t} = \mu i(t).$$

SIR model:
the fraction infected peaks and the fraction recovered saturates.



SIS model

Basic Reproductive Number

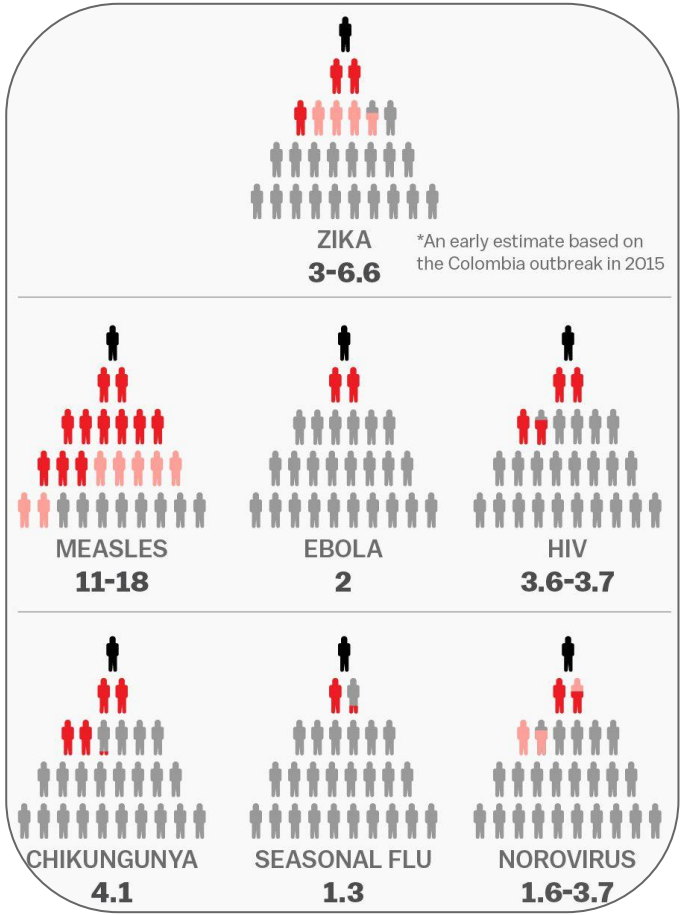
ƛ (also identified with R_0):
average # of infectious individuals generated by one infected in a fully susceptible population.

$$\lambda \equiv \frac{\beta}{\mu}$$

ƛ > 1: Outbreack

ƛ < 1: Die Out

Epidemic Threshold
if $\mu \approx$ □ then $i \longrightarrow 0$



Chapter 8

Diffusion: Opinion Dynamics

Summary

- Modeling Opinions
- Polarization, Fragmentation, Bias
- Modeling Algorithmic Bias

Opinion Dynamics

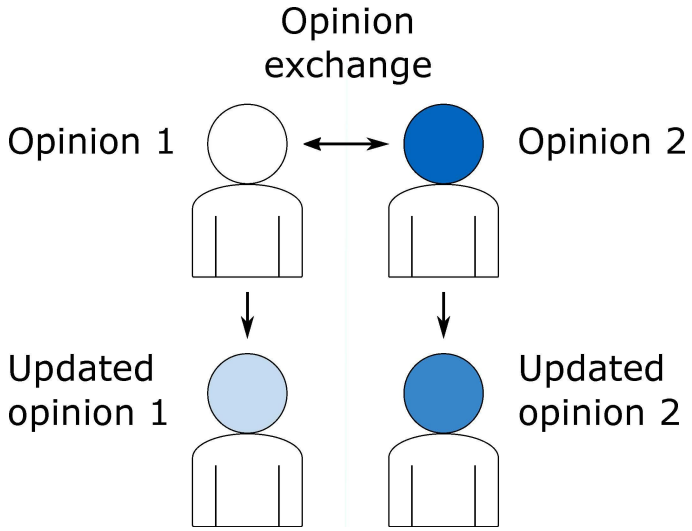
Model evolution of opinions in a population

Opinions are at the base of human behaviour

- understand behaviour - which mechanisms are important?
- trigger changes in behaviour ~ intervention methods in spreading, less explored

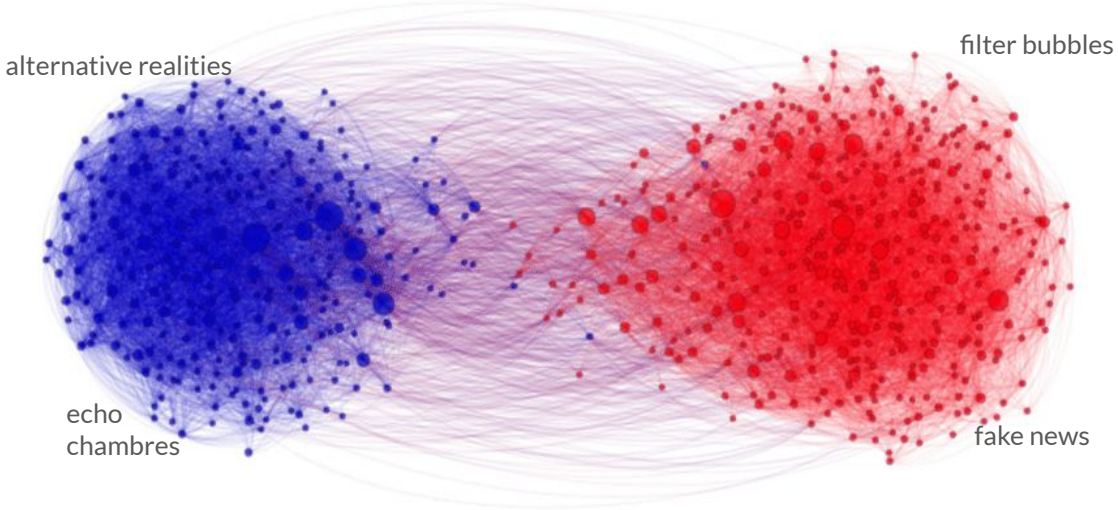
Broadly part of complex contagion modelling: peer effects through social network.

Simple representations of opinions - one variable.



Polarization and Fragmentation in Social Media

Polarization of the public debate

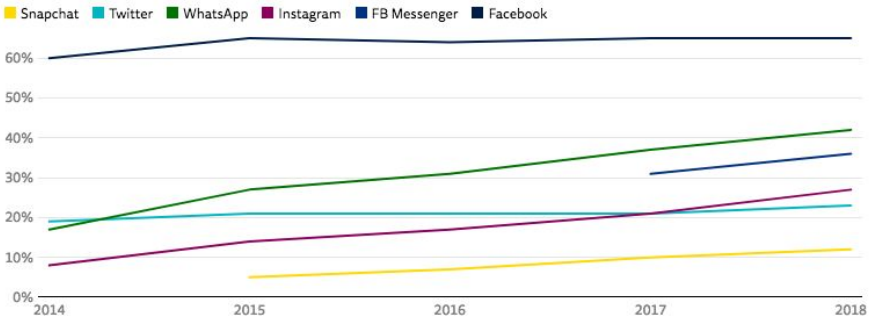


Adamic, Lada A., and Natalie Glance. "The political blogosphere and the 2004 US election: divided they blog." ACM (2005).

Online News Consumption

PROPORTION THAT USED EACH SOCIAL NETWORK FOR ANY PURPOSE IN THE LAST WEEK (2014–18)

Selected markets



Q12A. Which, if any, of the following have you used for any purpose in the last week?

Base: Total sample across selected markets: 2014 = 18859, 2015 = 23557, 2016 = 24814, 2017 = 24487, 2018 = 24735.

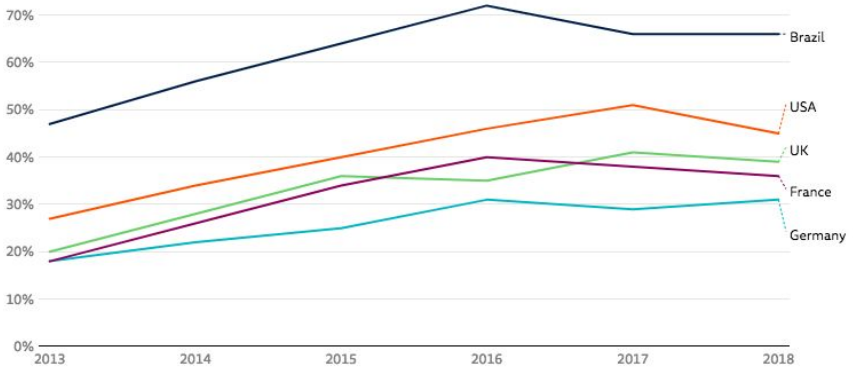
Note: From 2015–18, the 12 markets included are UK, US, Germany, France, Spain, Italy, Ireland, Denmark, Finland, Japan, Australia, Brazil. In 2014, we did not poll in Australia or Ireland.

Get the data • Embed

REUTERS INSTITUTE | UNIVERSITY OF OXFORD

PROPORTION THAT USED SOCIAL MEDIA AS A SOURCE OF NEWS IN THE LAST WEEK (2013–18)

Selected countries



Q3. Which, if any, of the following have you used in the last week as a source of news?

Base: Total 2013–2018 sample in each market.

Get the data • Embed

REUTERS INSTITUTE | UNIVERSITY OF OXFORD

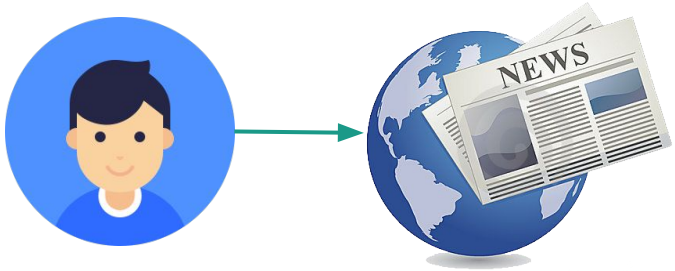
Online consumption of information

Interaction of

- *users*,
- with *media* content

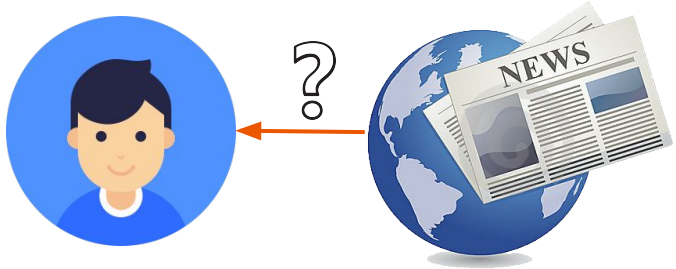
mediated by computer programs

1st scenario



Users actively search for news

2nd scenario



Users are passively fed of news

Online consumption of information

The aim of the computer programs is to maximise the usage of the platform

To fulfill such goal they carefully taylor the information shown to their users

Confirmation Bias

*"[is the] tendency to search for, interpret, favor, and recall information in a way that **confirms one's preexisting beliefs** or hypotheses."*

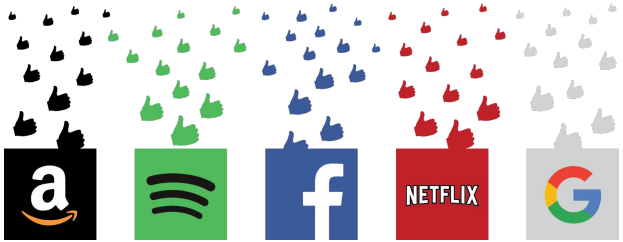
Recommender Systems

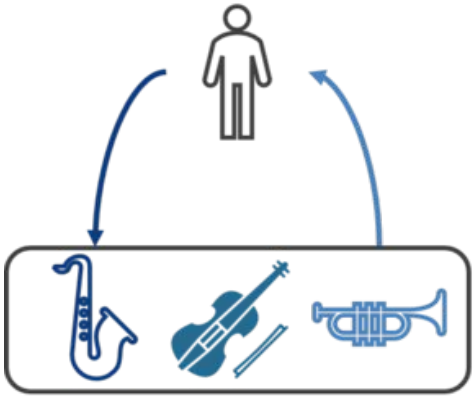
Leveraging user's history

Recommendations are built on top of user's past choices...

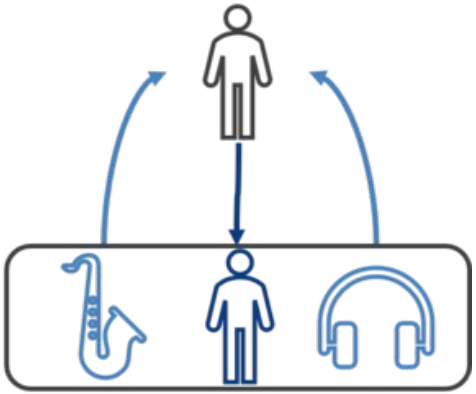
- type of news searched, product bought…

As well as on top of "similar" users' ones





A product is recommended that is **similar to products** the customer has already looked at.



The customer is shown products that customers with **similar data profiles** have found interesting.

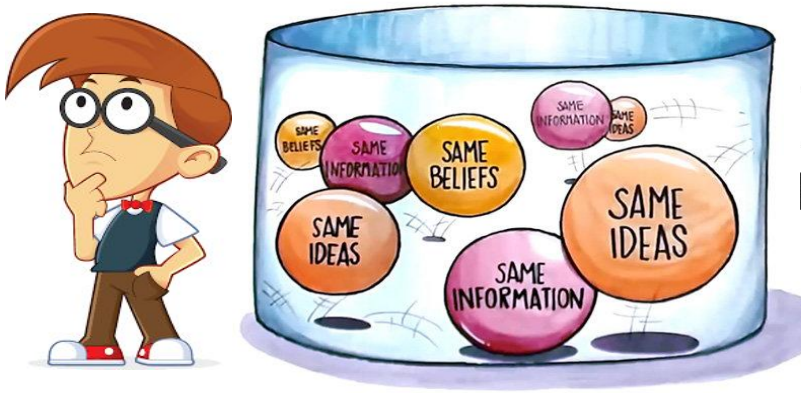
Online consumption of information

Users are mostly shown opinions that are close to their own (algorithmic bias)

- News about topics we like,
- Posts from close friends,
- ...

Users do not even get confronted with narratives different from their favorite ones

- or they get in contact with extreme opposite narratives



Modeling Algorithmic Bias

Models

Algorithmic Bias

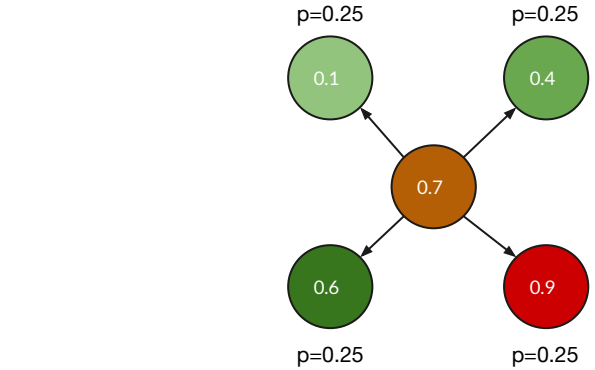
Modified Deffuant model

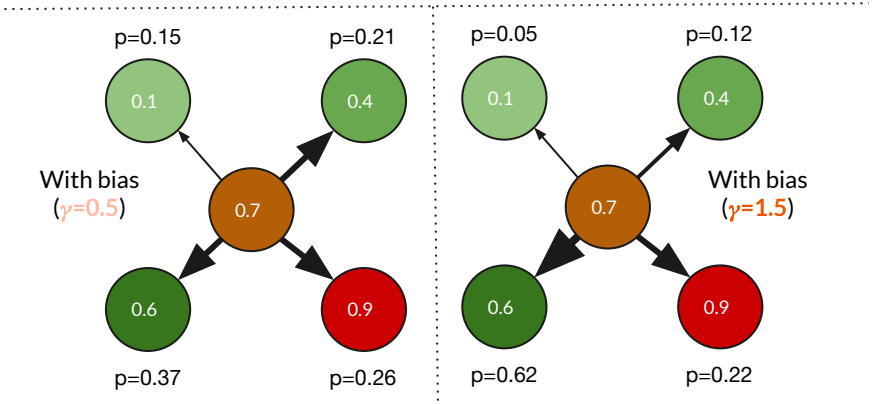
Probability to select interaction partner depends by

- the opinion distance, d_{ij}
- the bias strength, γ

$$p_i(j) = \frac{d_{ij}^{-\gamma}}{\sum_{k \neq i} d_{ik}^{-\gamma}}$$

The more similar the opinions, the more likely that the interaction will take place.





Sîrbu, Alina, et al. "Algorithmic bias amplifies opinion fragmentation and polarization: A bounded confidence model." PloS One (2019)

Deffuant

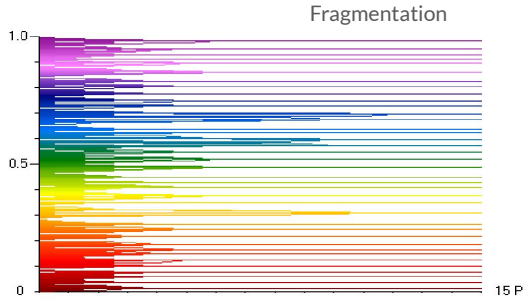
Simulations

Recap:

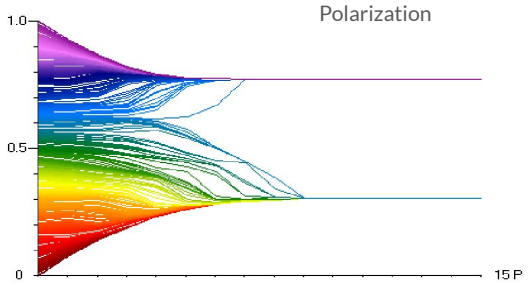
Reducing the bounded confidence threshold value opinion fragmentation (polarization) intensifies

Interpretation:

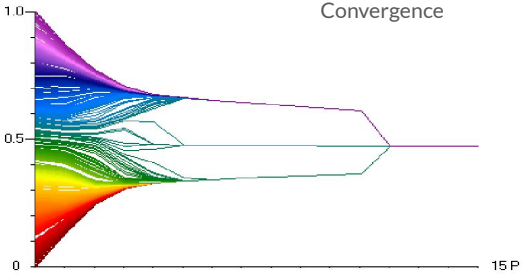
The larger the open-mindedness value, the more likely that consensus will be reached



$(a)\ \varepsilon_l = \varepsilon_r = 0.01$



$(b)\ \varepsilon_l = \varepsilon_r = 0.15$

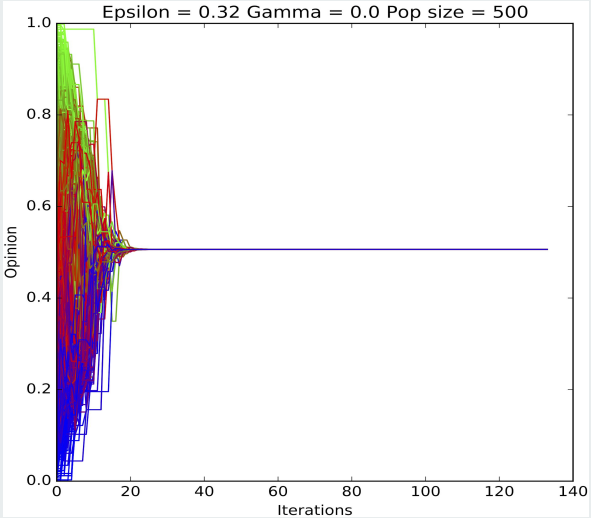


$(c)\ \varepsilon_l = \varepsilon_r = 0.25$

Deffuant

Without Bias

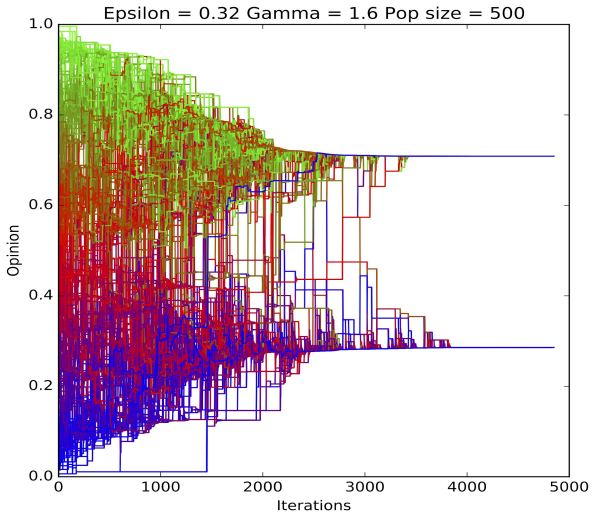
Convergence to common opinion



Deffuant

With Bias

Opinion Polarization, Fragmentation,
Convergence slow-down (instability)



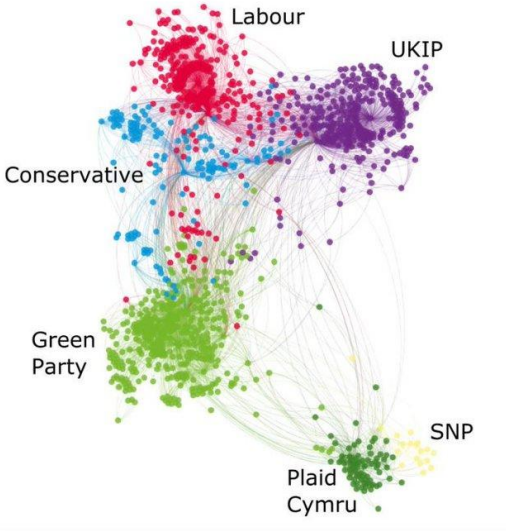
Algorithmic Bias

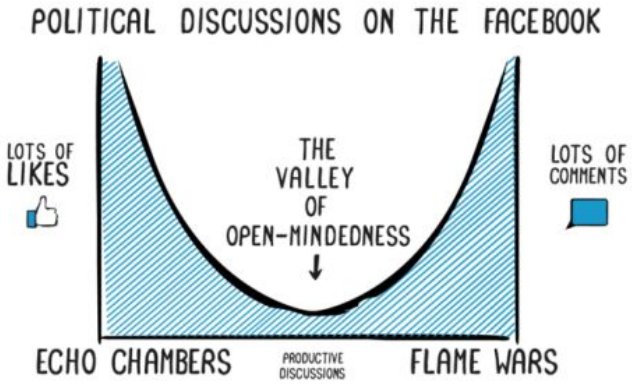
Is this the whole story?

Unfortunately, it is not.

The situation in reality is even worse

- Simulations performed in mean field
- The observed effects can be exacerbated by the topology of the social network





Appendix

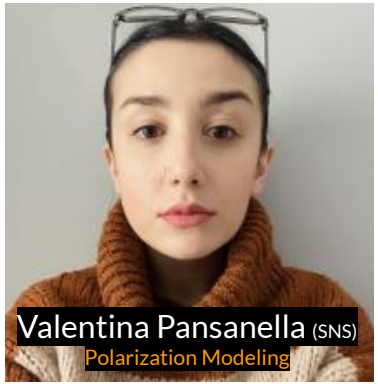
Complex Networks Analysis @KDD Lab

Knowledge Discovery & Data Mining Laboratory, ISTI-CNR & UNIPI & Scuola Normale

Team

Networks are ubiquitous in research.

Some of us use them as tools, others study their properties.



Applicative Contexts

Migration **S**tudies

Brain Drain, Turkey Crisis…

(H2020 project, HummingBird)

Where (and why) scholars move?

Bloomberg

Economics

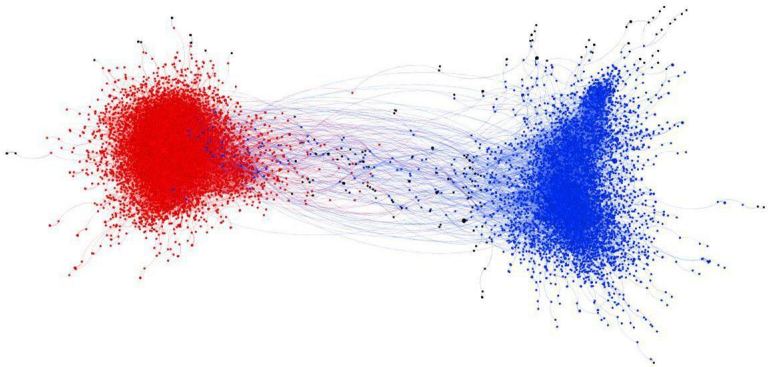
Europe Risks New Migration Meltdown as Erdogan Opens Floodgates

By Cagan Koc, Selcan Hacaoglu, and Nikos Chrysoloras

29 febbraio 2020, 12:37 CET *Updated on 1 marzo 2020, 09:21 CET*

Polluted **I**nformation **E**nvs

Echo Chambers, Social Pressure & Fake News

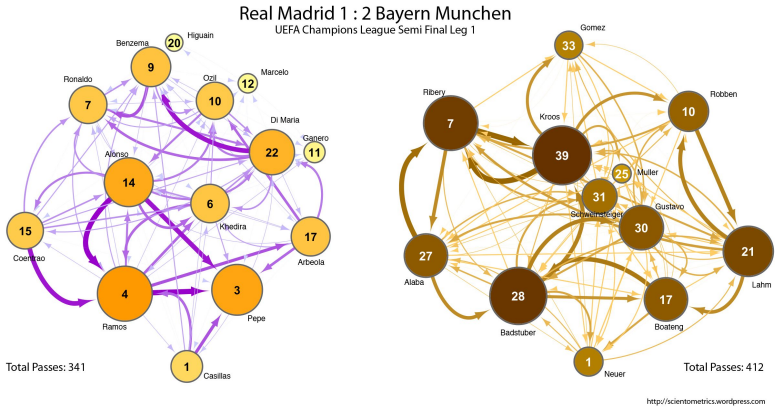


Identifying, modeling & dismantle echo chambers
Modeling Opinion Dynamics of Fake news perception
...

Can we model how fake news spread?
Is peer pressure playing a role?

Science of **S**uccess

Sport, Hit-Savyy, Start-ups,...



Modeling Sport with networks

Revising Rogers Innovators and their role...

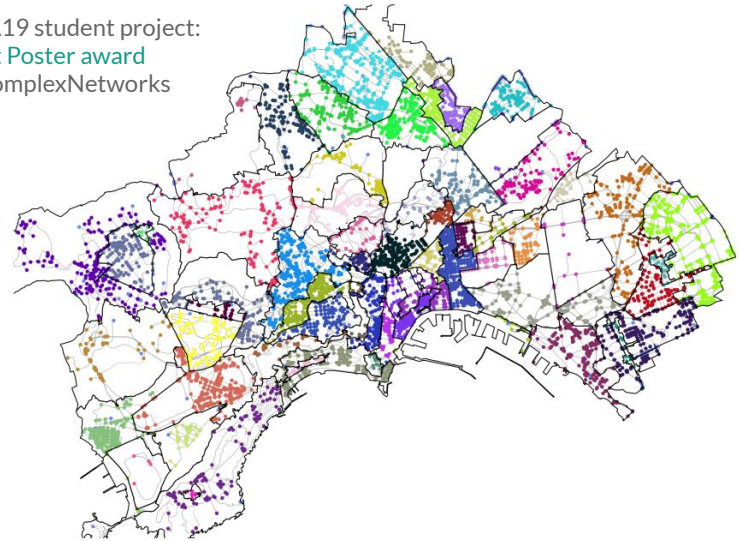
Can we predict start-ups success?

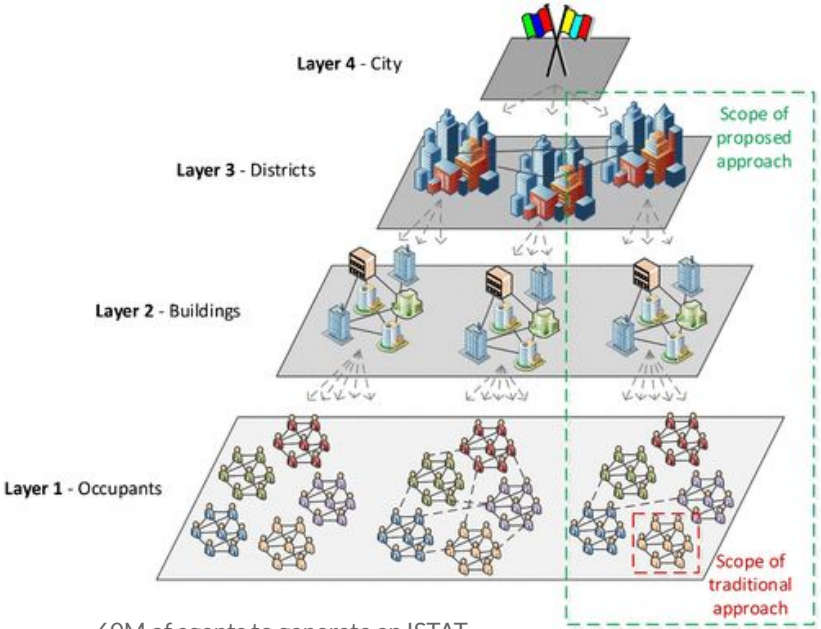
(one active project EPO, one H2020 Proposal, OrAIcle)

Others

Mobility, Urban planning, SimulaItalia...

SNA19 student project:
Best Poster award
@ComplexNetworks





60M of agents to generate an ISTAT data-driven social proxy for Italy

Recent Case Studies

The Three Dimensions of Social Prominence

Giulio Rossetti
KDD Lab, ISTI-CNR, Italy
giulio.rossetti@isti.cnr.it
@GiulioRossetti

et. all

Question:

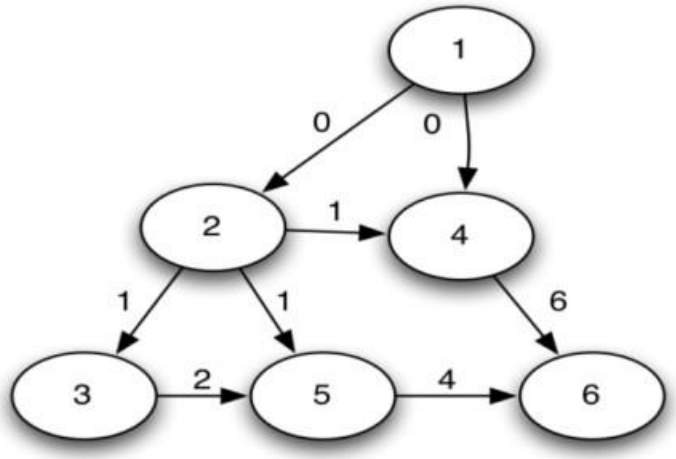
Who are the Social Influence “Leaders”?

What can we say by observing social influence patterns?

Data: Last.fm

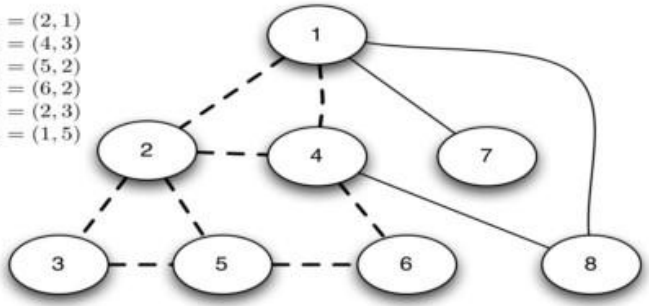
- 80.000 UK users
- 4.000.000 friendship ties
- 2 years of listening records

- Define what a diffusion "*leader*" is
- Identify three measures of social prominence (*width*, *depth*, *strength*)
- Analyze their relationships with the topological characteristics of prominent actors in a network
- Look for patterns distinguishing different objects spreading in a social network



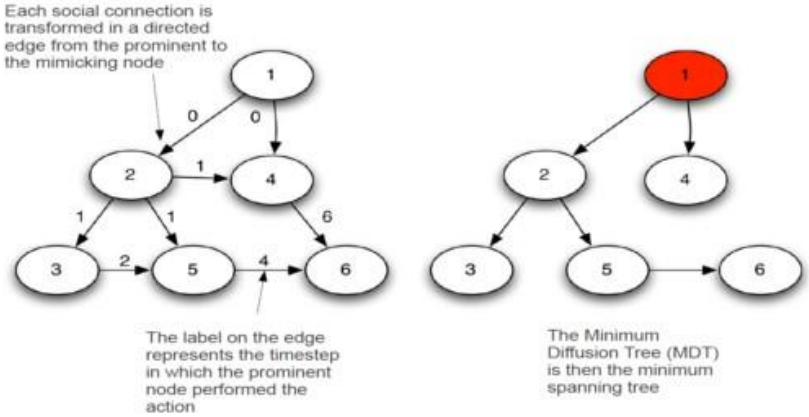
Who are the Leaders?

$a_{1,x} = (1, 0)$ $a_{4,y} = (2, 1)$
$a_{2,x} = (2, 1)$ $a_{7,y} = (4, 3)$
$a_{3,x} = (1, 2)$ $a_{8,y} = (5, 2)$
$a_{4,x} = (4, 6)$ $a_{6,y} = (6, 2)$
$a_{5,x} = (1, 4)$ $a_{1,y} = (2, 3)$
$a_{6,x} = (6, 7)$ $a_{2,y} = (1, 5)$



For each Artist we extract the **induced temporal subgraph** of its Listeners

We define Leader all those nodes that are the **first**, in their neighborhood to **adopt** the given artist

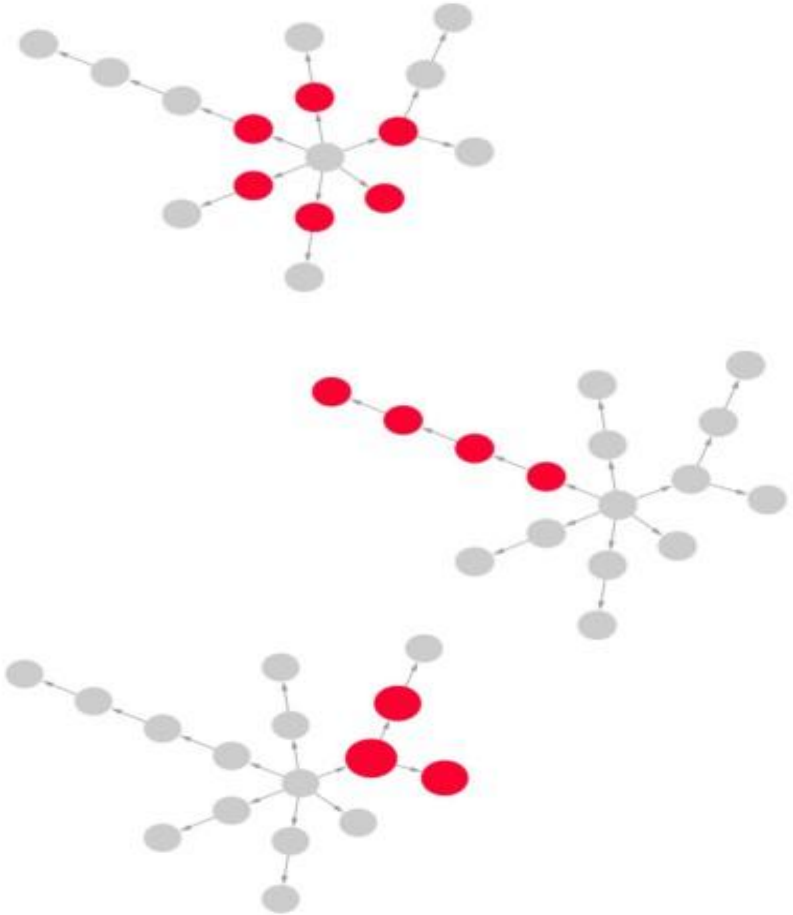


Social Prominence

A small set of users in a Social Network is able to anticipate (or influence) the behavior of the entire network.

We characterize them in terms of their diffusion tree properties:

- width
- depth
- strength

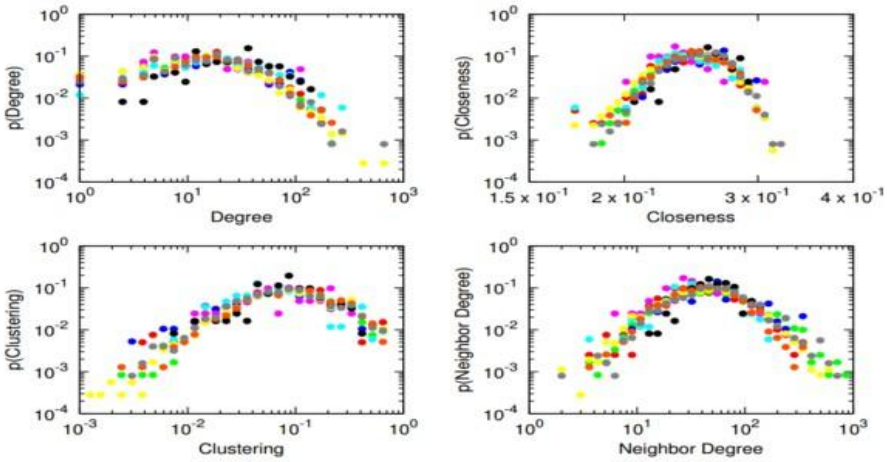


	Width	Strength	Degree	Clustering	Neigh Deg	Bet Centr	Clo Centr
AVG Depth	-0.03	**-0.23**	-0.08	0.05	-0.08	-0.02	**-0.13**
Width	-	0.01	**-0.31**	**0.13**	0.05	-0.07	**-0.59**
Strength	-	-	0.02	-0.02	0.03	0.00	0.04
Degree	-	-	-	**-0.16**	-0.02	**0.77**	**0.56**
Clustering	-	-	-	-	-0.05	-0.06	**-0.32**
Neigh Deg	-	-	-	-	-	-0.00	**0.39**
Bet Centr	-	-	-	-	-	-	**0.22**

Central nodes are characterized by low Depth & Width

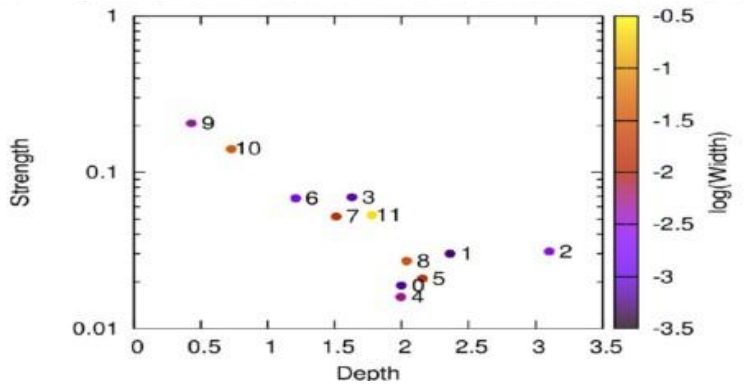
High Width are usually reached only by
nodes in tightly knit communities

There is a trade-off between Depth and Strength



Experimental Results: Hubs and central nodes are not “real” influencers

Cluster	size	dance	ele	folk	jazz	met	pop	punk	rap	rock
0	1822	1.25	1.13	**1.54**	1.37	1.50	0.76	1.31	1.13	1.10
1	136	1.28	1.55	1.28	**2.35**	0.78	0.73	0.64	1.35	0.70
2	664	0.59	0.87	0.98	0.48	0.95	0.97	**1.50**	1.20	1.19
3	482	1.26	1.16	1.09	1.12	0.91	0.80	**2.48**	1.24	0.89
4	973	1.14	1.20	1.15	**1.41**	0.80	0.91	0.66	0.97	0.97
5	512	**1.29**	0.96	0.95	1.09	1.10	0.97	0.33	1.06	1.01
6	682	0.89	0.79	0.61	0.64	**1.13**	1.08	1.07	1.08	1.01
7	124	0.75	**1.45**	0.35	0.64	0	1.09	0	1.02	0.62
8	524	0.93	1.01	1.12	0.91	**1.15**	1.07	0.43	0.95	0.87
9	937	0.40	0.46	0.19	0.23	0.45	**1.56**	0.13	0.37	1.06
10	232	0.72	0.57	0.27	0.99	0.38	**1.44**	0.38	0.46	1.00
11	612	0.74	0.94	0.71	0.40	0.70	**1.27**	0.07	0.68	0.83



Jazz:
[1] lowest width
[4] lowest strength
Not easy to be prominent

Pop:
[9, 10, 11]
Lowest depth, highest strength
Leaders for pop artists are embedded in groups of users very engaged with the new artist, but not prominent among their friends

Punk:
[2] high depth, low width and strength
Long cascades, exactly the opposite of the pop genre, similar to folk!

Dance:
]5] high depth, high width, low strength
Dance successes are studied to reach everyone, but in two days nobody remembers anything about them…

Results: Music Genres can be classified by diffusion trees

Homophilic Network Decomposition:
a community-centric analysis of OSNs

Giulio Rossetti
KDD Lab, ISTI-CNR, Italy
giulio.rossetti@isti.cnr.it
@GiulioRossetti

et. all

KDD Lab

CD·LIB

ISTITUTO DI SCIENZA E TECNOLOGIE DELL'INFORMAZIONE "A. FAEDO"

Skype Service Usage

Given an online platform we often we need to *estimate* how its services are used by the registered users (i.e., Skype Audio\Video call).

In particular we can be asked to answer the following questions:

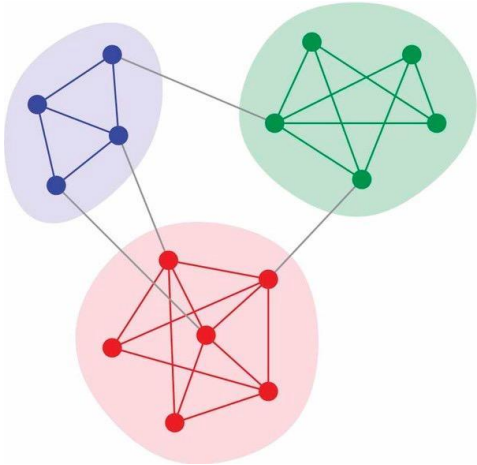
Q1: Can Service Usage be described as a function of the Network Data?

Q2: If so, at which scale should we analyze the network in order to perform a descriptive analysis?

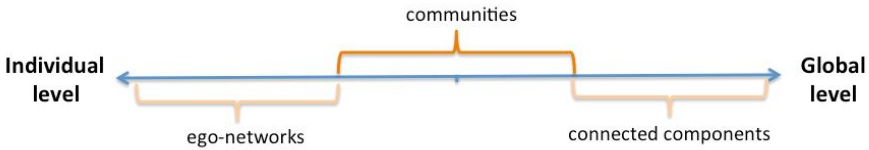
Observation Scale?

Problem:
Given the size of the dataset (several hundred millions of users) an individual level analysis can be redundant;

Identifying tight groups of “similar” users we can reduce the problem space



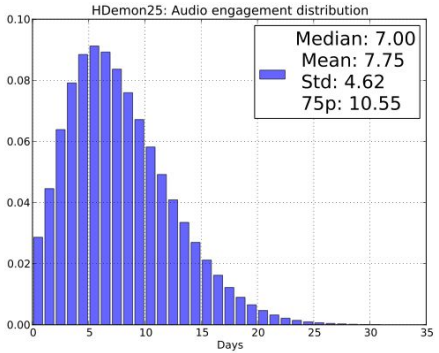
Community Discovery

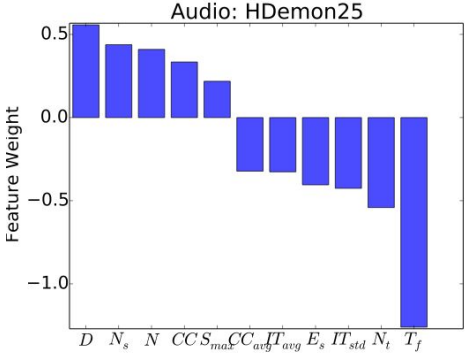


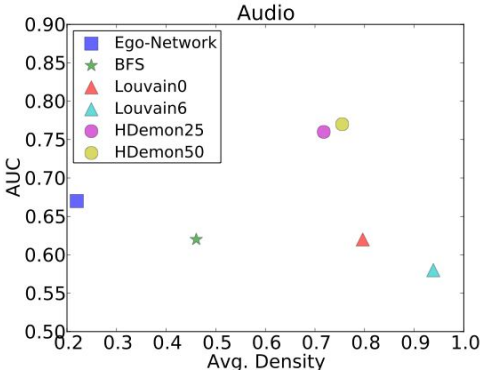
The smaller and denser communities are the better

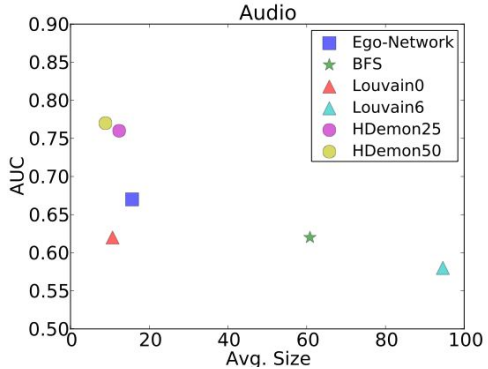
Demon outperforms Louvain, Ego-Nets and BFS

Topological, Temporal and Geographical features of communities are valuable activity level predictors









Experimental Results

A Complex Network approach to Semantic Spaces

Salvatore Citraro
University of Pisa, Italy
citraro@di.unipi.it
@SalvatoreCitraro1

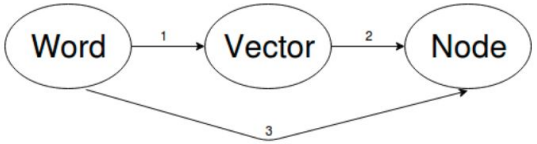
et. all

KDD Lab

CD·LIB

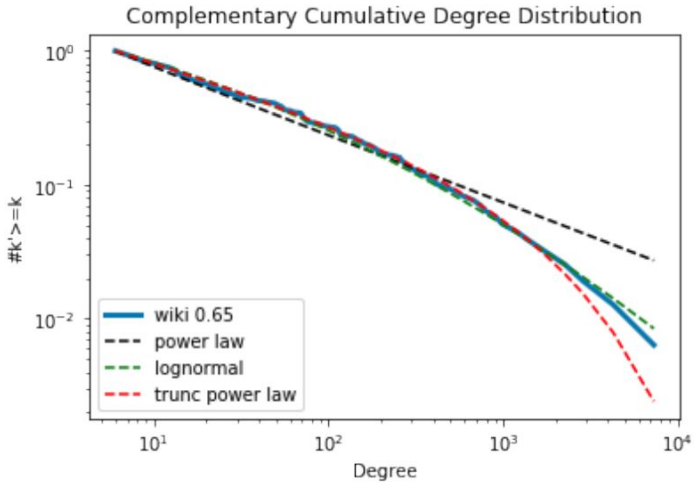
ISTITUTO DI SCIENZA E TECNOLOGIE DELL'INFORMAZIONE "A. FAEDO"

Semantic Spaces



Can we assume communities as proxies for semantic fields?

- A semantic field is a set of words grouped by one or more types of semantic properties



NB: We use the equivalence word-vector-node to model a mental lexicon

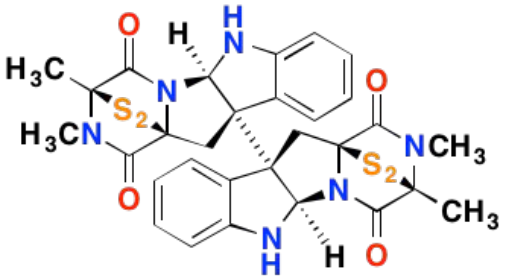
Infomap and **Louvain**:

Louvain: biggest community capture the domain of **chemistry**
Infomap: chemistry community breaks into smaller ones
subdisciplines: organic chemistry, biochemistry, pharmacology, etc

Demon (term "**Siria**"):

- Historical geographical entity:
  accadi, anatolia, assiro, babilonese, egizio, eufrate, ittiti, mesopotamia, persia, sumero

- Modern geographical entity:
  afghanistan, aleppo, arabia, armenia, cisgiordania, egitto, gaza, iran, iraq, isreale, libano, marocco, oman, pakistan, turchia, yemen



Experimental Results: Algorithms & Contexts

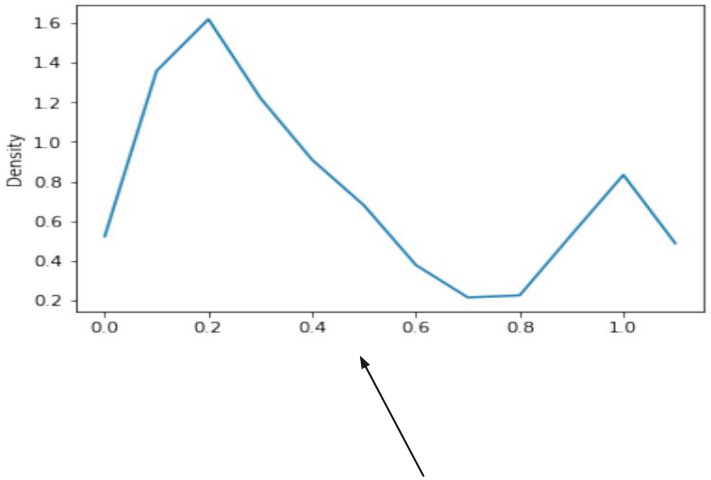
Ground Truth: Wikipedia disambiguation pages

✓ "**Stima**" (polysemy):

- Price of something:
  (lira, miliardo, dollaro, milione, sterlina, euro)
- Approximate measurement:
  (grossomodo, pressapoco, incirca)

✗ "**Pesca**" (homonymy):

- Communities related to the fishing activity and not to the fruit
  (anguilla, aragosta, salmone, polpo, tonno and barca, marinaio, pescatore, peschereccio and allevamento, allevare, intensivo)



It is not an easy task!

Experimental Results: Can we capture Polysemy or Homonymy?

Toward a Standard Approach for Echo Chambers Detection

Virginia Morini
ISTI-CNR, Italy
virginiamorini95@gmail.com
@Virgiiim

et. all

KDD Lab

CD·LIB

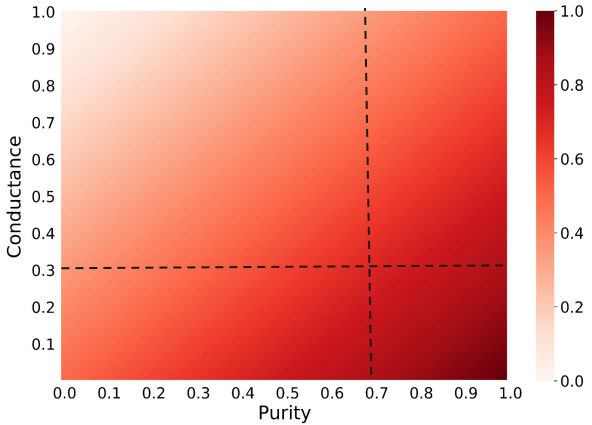
ISTITUTO DI SCIENZA E TECNOLOGIE DELL'INFORMAZIONE "A. FAEDO"

Echo Chambers

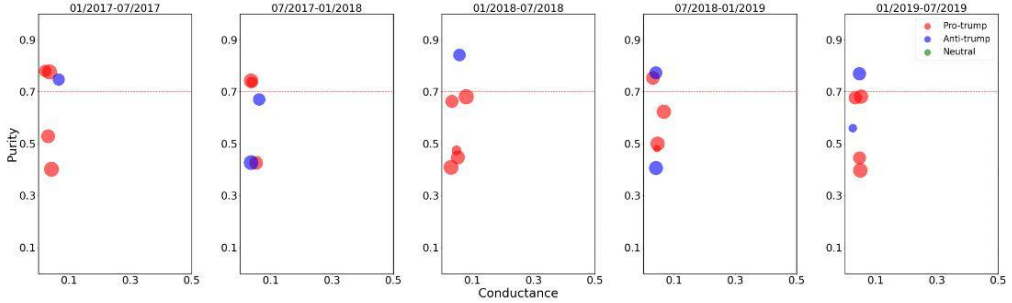
An Echo Chamber is a subset of the network nodes (users) who share the same ideology and tend to have dense connections primarily within the same group.

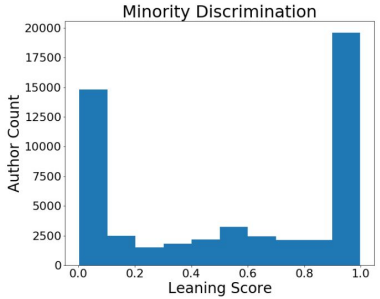
Idea

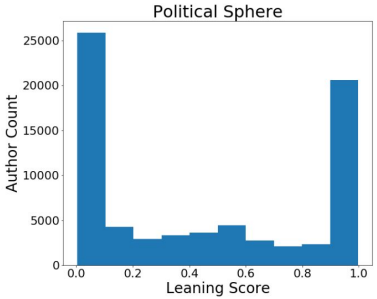
Apply Attributed-CD to find homogeneous clusters and identify the ones that are close-worlds

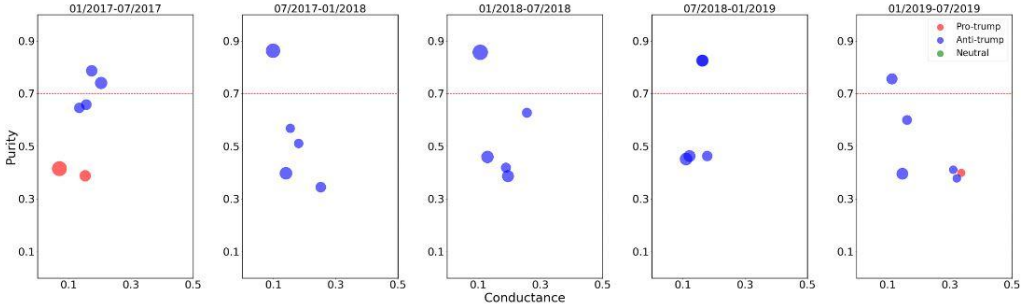


Minority Discrimination









Political Debate

Experiments: Echo Chambers in time (Reddit data)

(Re)Orgainizing Health
Services and Demands

Letizia Milli
University of Pisa, Italy
milli@di.unipi.it
@LetiziaMilli

Giulio Rossetti
KDD Lab, ISTI-CNR, Italy
giulio.rossetti@isti.cnr.it
@GiulioRossetti

et. all

KDD Lab

CD·LIB

NDlib
Network Diffusion Library

ISTITUTO DI SCIENZA E TECNOLOGIE DELL'INFORMAZIONE "A. FAEDO"

Resource Allocation

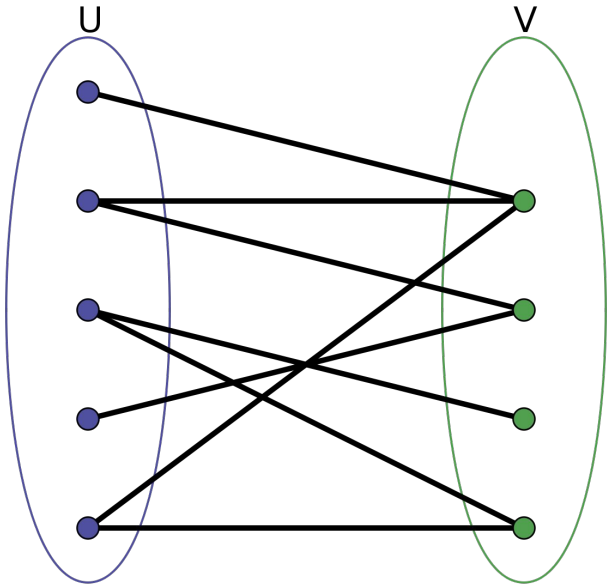
How to better allocate health resources to meet population demands?

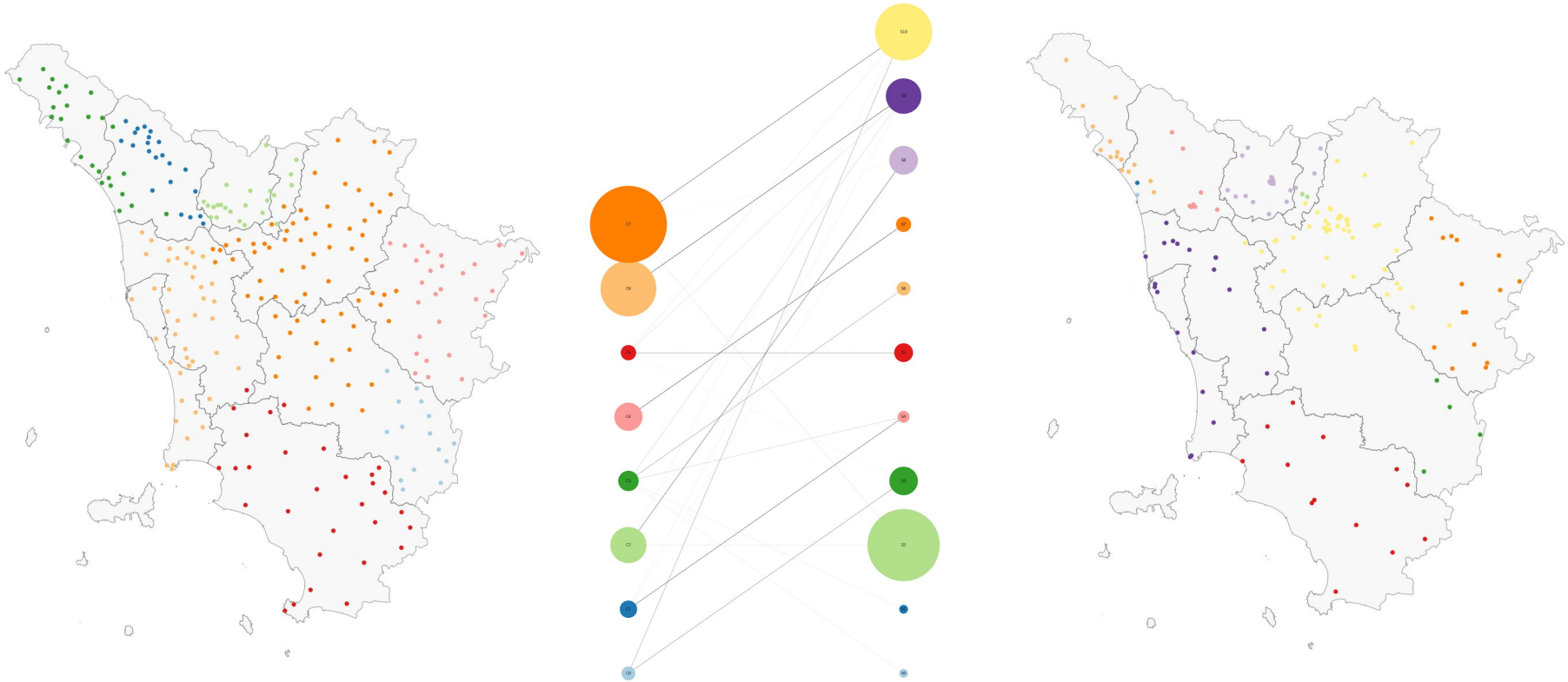
Data

- Specialistic visits
- Health centers (hospitals, clinics…)
- Patients municipalities

Expected Outcome:

- Municipality clusters
  (e.g., municipalities that uses the same services)
- Health centers clusters
  (e.g., centers that satisfy demands of a same set of municipalities)





Bi-Clustering: Municipalities (left) & Health Centers (right) -- Ophthalmology

(Social) Network Analysis

Conclusion

Take Away Message

Network Science opens countless opportunity of analysis

Email: giulio.rossetti@isti.cnr.it
Twitter: @GiulioRossetti