P.h.D. course
Universitat Autonoma de Barcelona, 2022

(**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"

Lecture 3

Micro, Meso & Macro: Different perspectives

Chapter 5

Micro: Centrality & Tie Strength

Summary

- Measuring Node importance
- The strength of weak ties

Reading

- Chapters 3 & 4 of Kleinberg's book

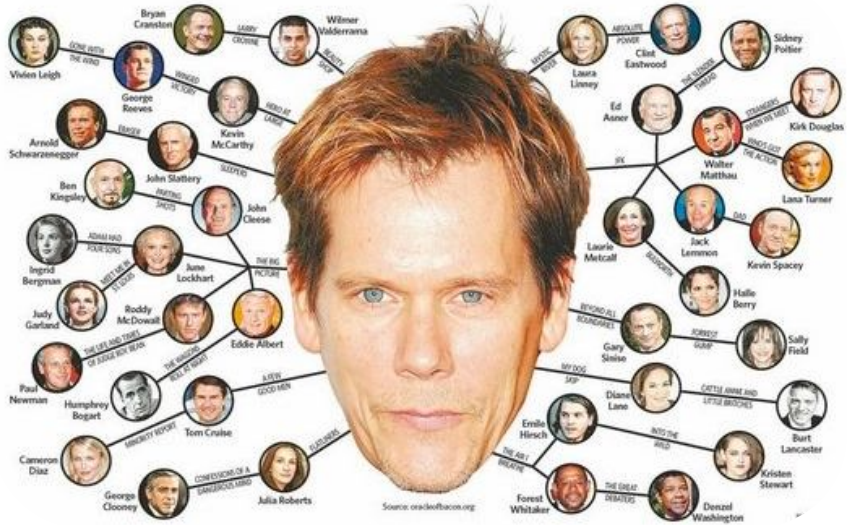
How important is a node in a network?

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.

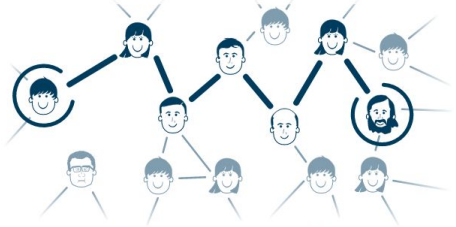
For instance:

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

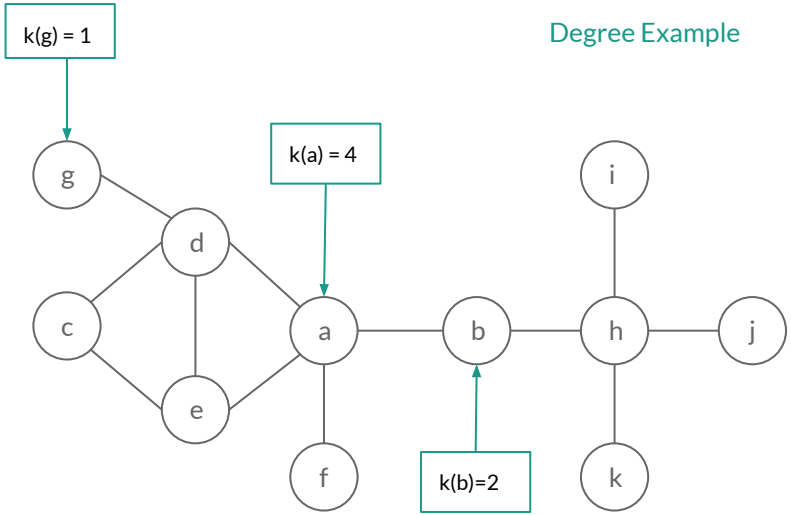
Name	Erdős number	Bacon number	Erdős–Bacon number
Daniel Kleitman	1	2	3[4]
Bruce Reznick	1	2	3[69]
Albert M. Chan	3[21][22][23]	1[24]	4
Nicholas Metropolis	2[10]	2[68]	4
Steven Strogatz	3[79][80][81]	1[a][c][82]	4[a][c]
Robert J. Marks II	3[44][45][46]	2[61][62][63]	5
Tom Porter	3 (in two ways)[6][7]	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 tv 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

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

Connectivity-based centralities

"influence based on the number of links a node has to other nodes in the network"

Recursive definitions

Recursive importance:

Important nodes are those connected to important nodes

Several centralities based on this idea:

- Eigenvector centrality
- PageRank
- Katz centrality
- ...

Idea:

1. Each node has a score (centrality)
2. If every node "sends" its score to its neighbors, the sum of all scores received by each node will be equal to its original score

$$x_i^{(t+1)} = \sum_{j=1}^{n} A_{ij} x_j^{(t)}$$

x_i is the centrality of node i

How to solve it *(power method)*:

1. Initialize scores to random values
2. Update the values according to the desired rule

Does it converge?

Yes, if the graph is *undirected* with a *single connected component* (Perron-Frobenius theorem)

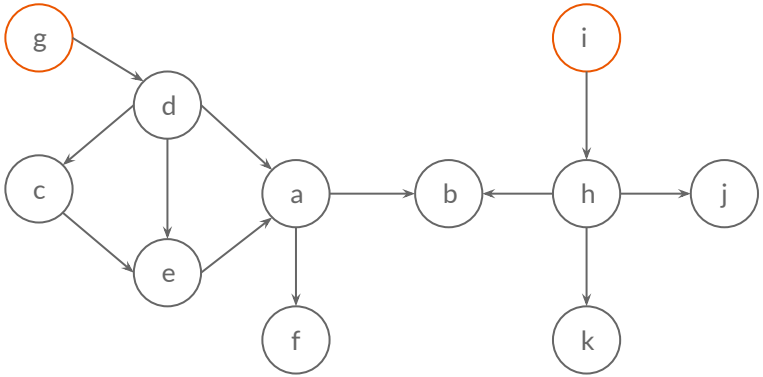
Eigenvector Centrality

A pair of eigenvector (***x***) and eigenvalue (***λ***) is defined by the relation:

$$Ax = \lambda x$$

- ***x*** is a vector of size N that can be interpreted as the nodes scores
- ***Ax*** yield a new vector of the same size which corresponds for each node to the sum of the received scores from its neighbors
- the equality implies that the new scores are proportional to the previous ones

Eigenvector Example



Problems:
In case of DiGraphs:
- Adjacency matrix is asymmetric
- 2 sets of eigenvectors
- 2 leading eigenvectors (use the incoming ones)

In presences of source nodes (0 in-degree):
- E(g) = 0
- E(d) = 0 as well since its incoming link comes from A
- ...

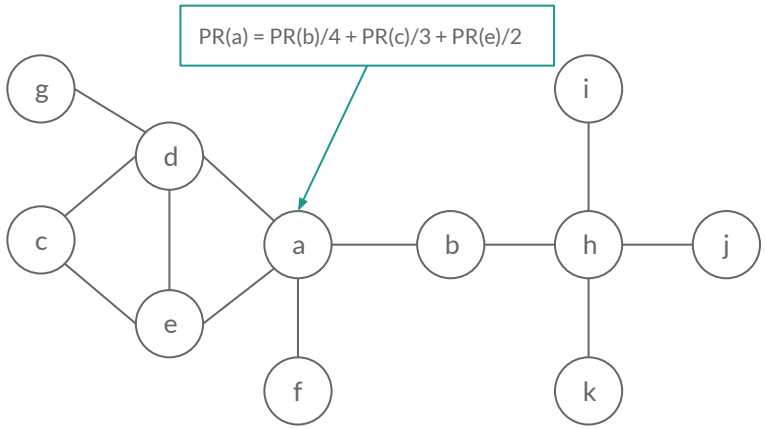
PageRank

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)

Geometric Centralities

"importance of a node depends on some function of its distances w.r.t. other nodes"

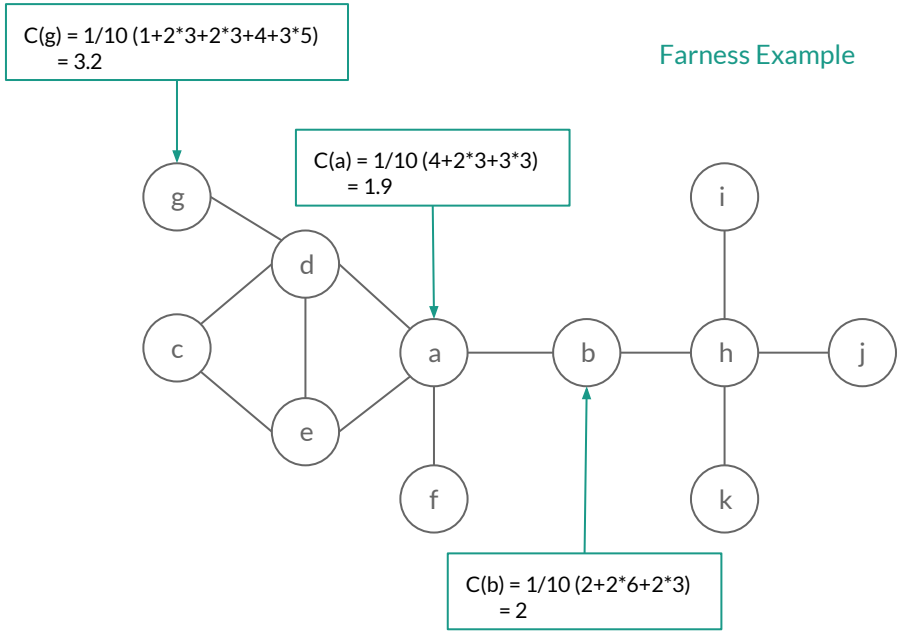
Closeness Centrality

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

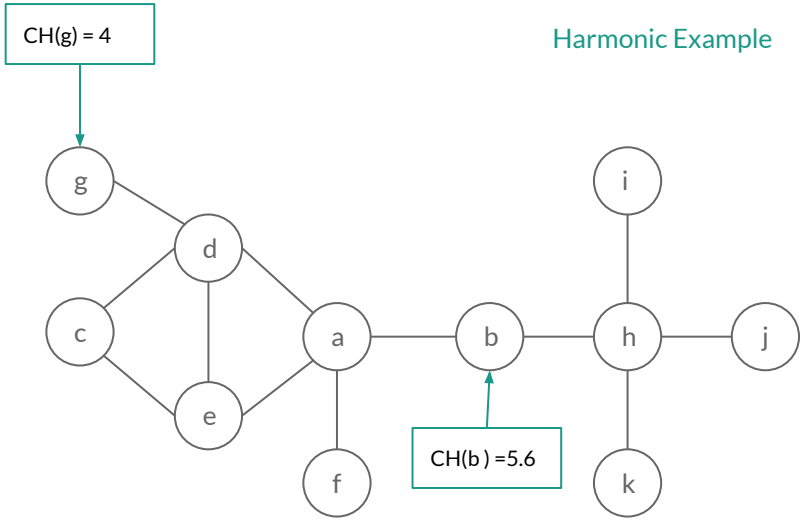


Closeness Formula

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

Harmonic Centrality

Harmonic mean of the geodesic (shorted paths) distances from a given node to all others.



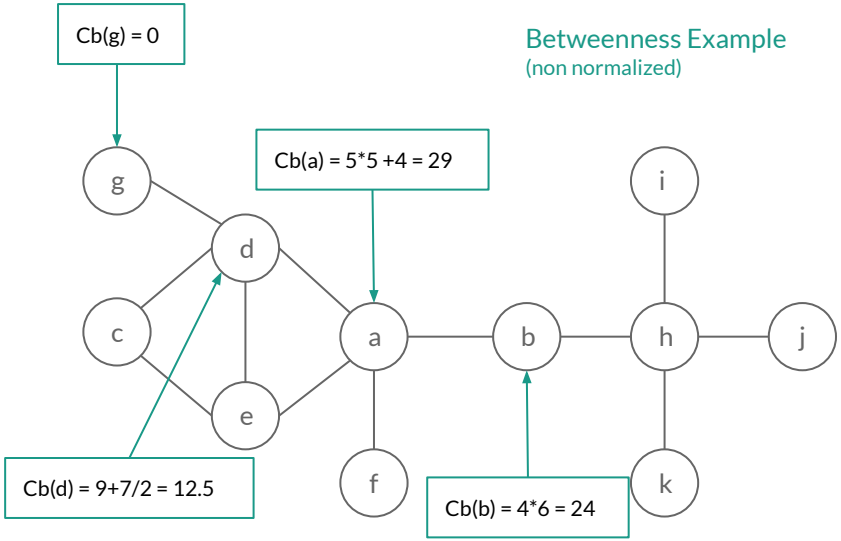
$$CH(i) = \frac{1}{n-1} \sum_{i \neq j} \frac{1}{d_{ij}}$$

- In case of no paths between two nodes i and j $d_{ij}=\infty$
- Well defined for disconnected graphs

Betweenness 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	Closeness	What's your average distance w.r.t. the rest of the network?
04	Harmonic	What's your harmonic average distance w.r.t. the rest of the network?
05	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

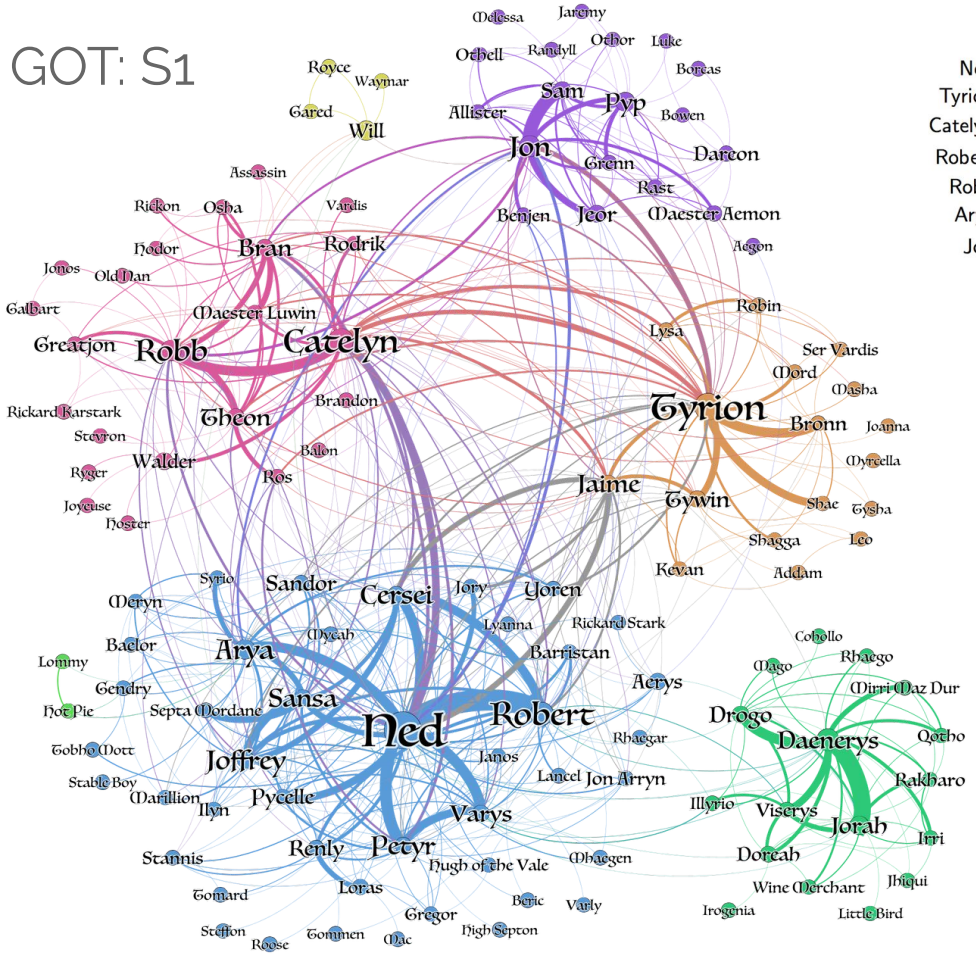
- E.g. If information does not spread through shortest paths, betweenness centrality is irrelevant

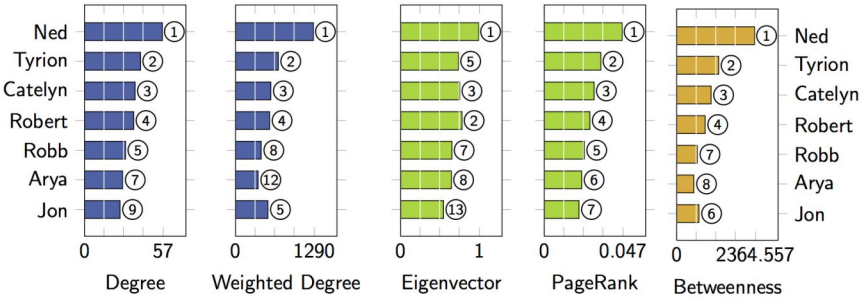
Centrality measures should be used with caution (a) for exploratory purposes and (b) for characterisation

Understanding Centralities

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

GOT: S1





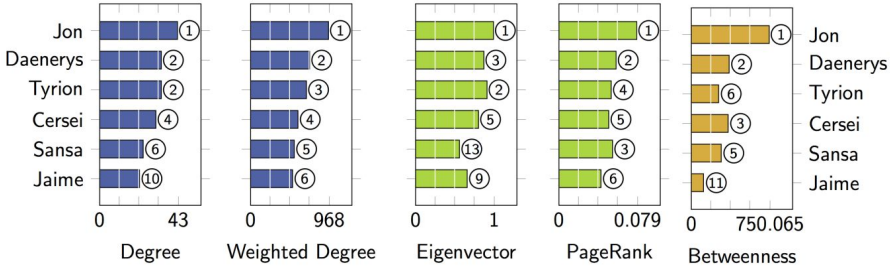
Node Label: PageRank
Node Size: Betweenness Centrality

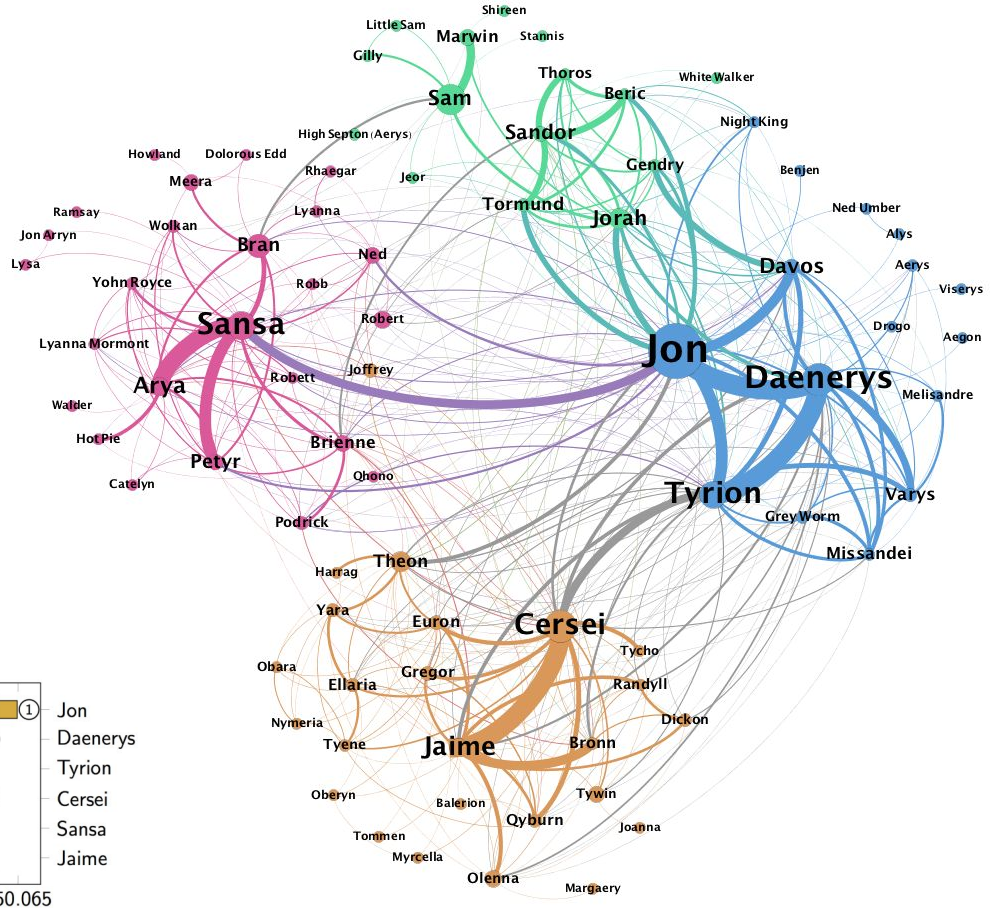
Edge Size: #interactions
Colors: Community (with Louvain)

GOT: S7

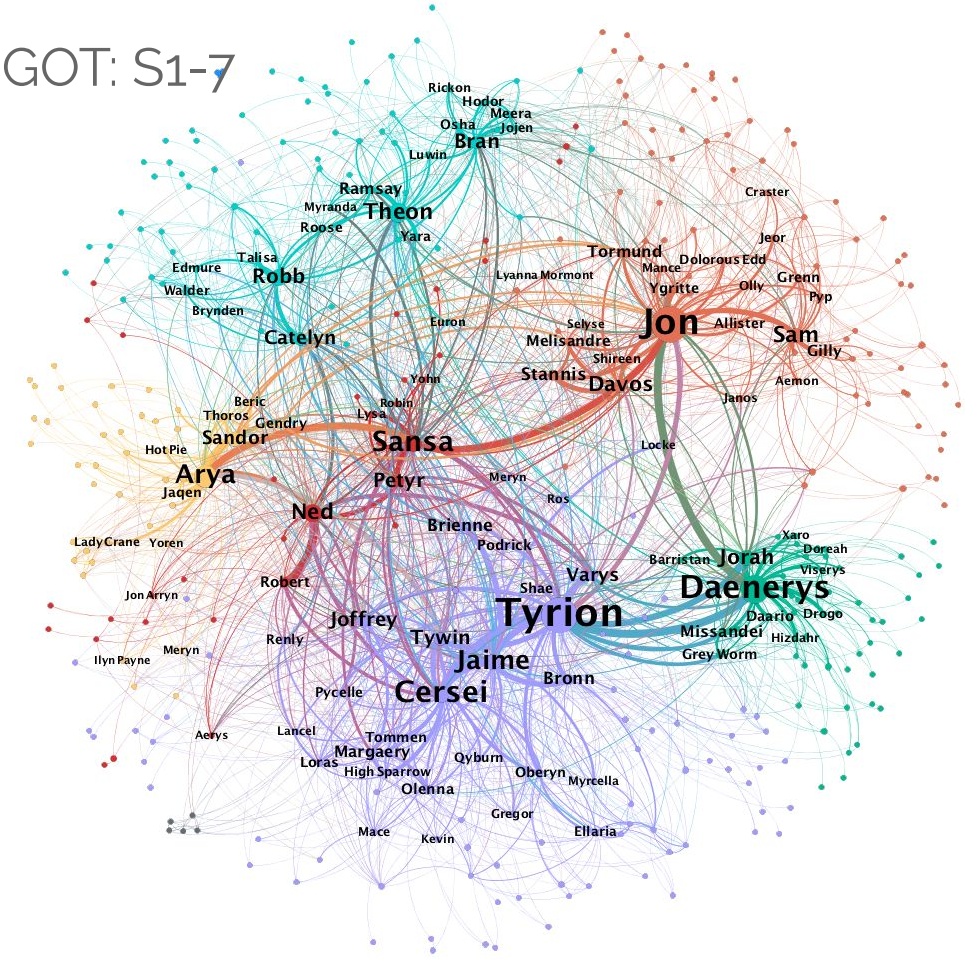
Node Label: PageRank
Node Size: Betweenness Centrality

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





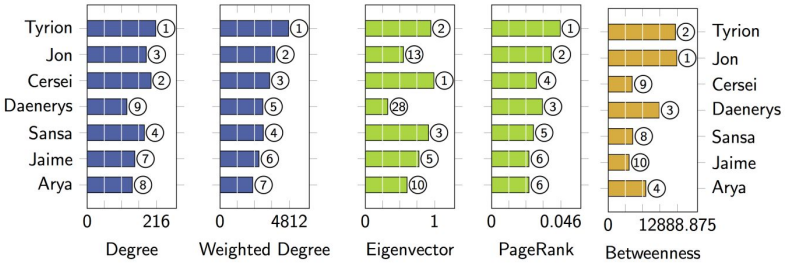
GOT: S1-7



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

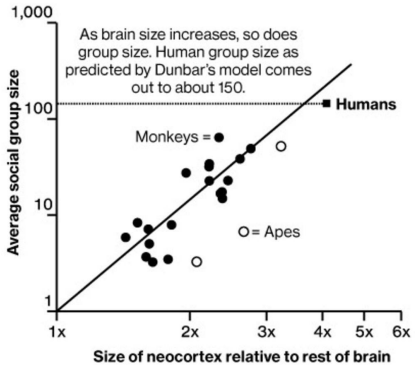
The Strength of Weak Ties

How many friends does one person needs?

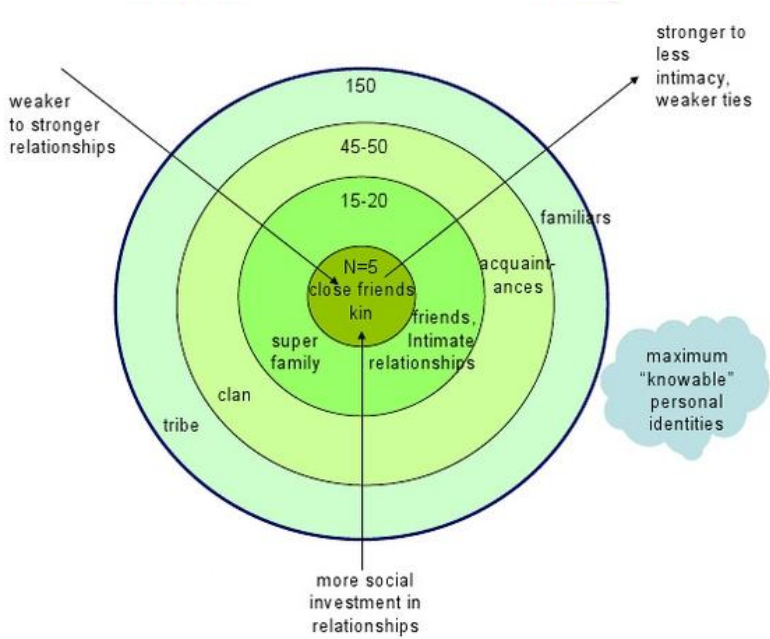
Not all ties in a social graph are the same

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.

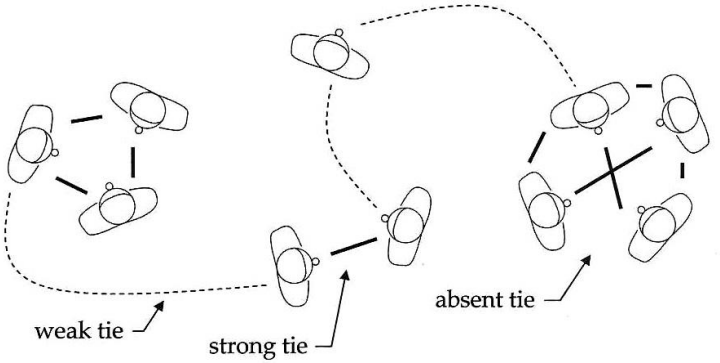
The strength of weak ties

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**:

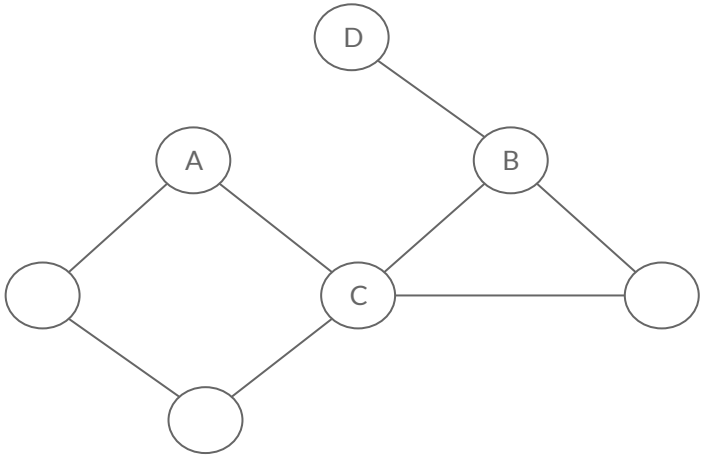
- the amount of time spent interacting with someone,
- the level of intimacy,
- the level of emotional intensity,
- and the level of reciprocity.

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

Triadic Closure

Social Intuition:

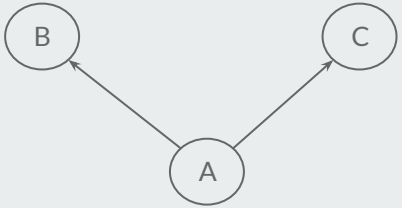
if two people in a network have a friend in common there is an increased likelihood that they will become friends themselves



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)

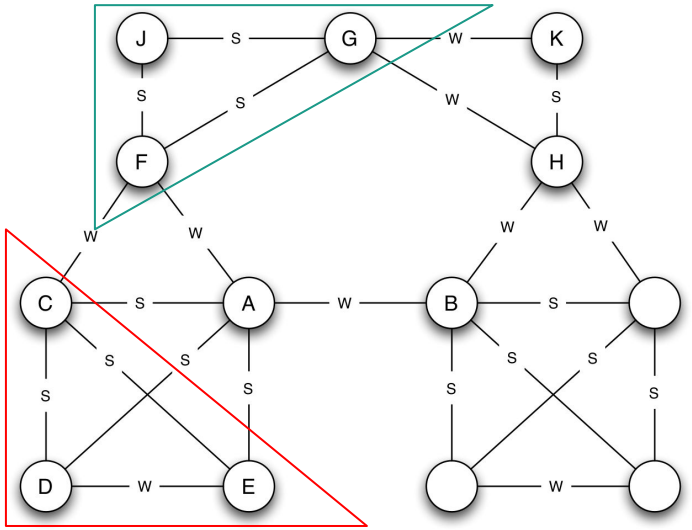
Strong Triadic Closure

Links in networks have strength;

- Friendship, Communication

We characterize links as either:

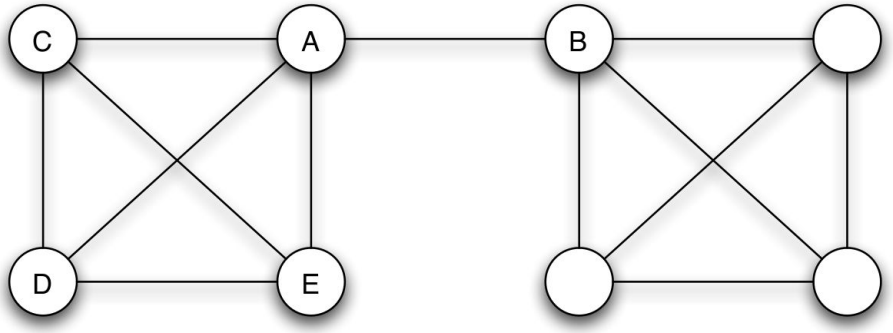
- Strong (friends), or
- Weak (acquaintances)



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

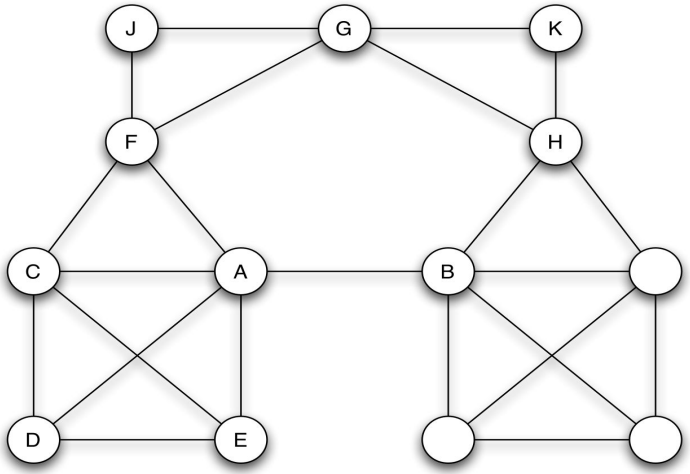


Strong Definition

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



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

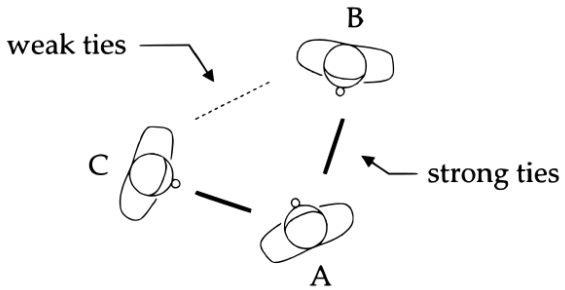
Local Bridges and Weak ties

Claim:

If a node A satisfies Strong Triadic Closure and it is involved in at least two strong ties, then any local bridge adjacent to A must be a weak tie

Proof (by contradiction):

- A satisfies Strong Triadic Closure
- Let (A,B) be a local bridge and a strong tie
- Then (B,C) must exist because of Strong Triadic Closure
- But then (A,B) is not a bridge



Measuring Tie Strength in Real Data

Social proximity and tie strength

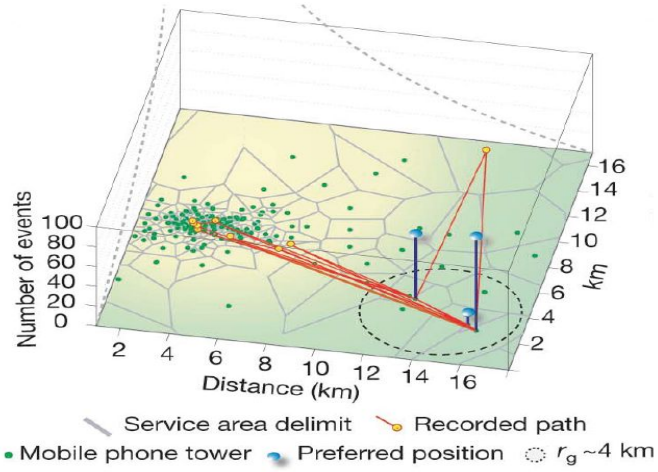
How connected are u and v in the social network.

- 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

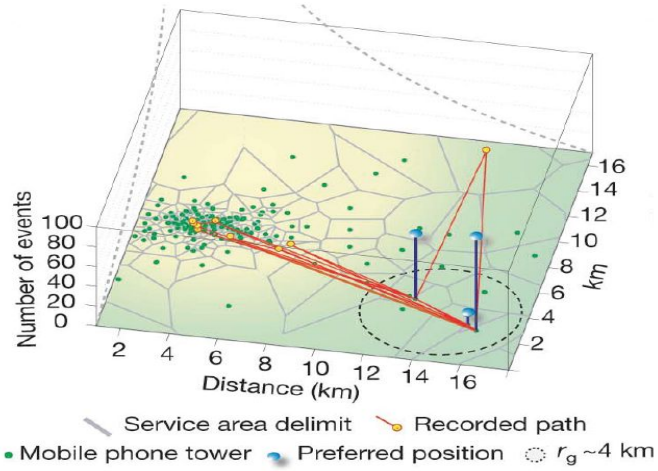
J.-P. Onnela, J. Saramaki, J. Hyvonen, G. Szabo, D. Lazer, K. Kaski, J. Kertesz, A.-L. Barabási. Structure and tie strengths in mobile communication networks. PNAS 104 (18), 7332-7336 (2007).

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).

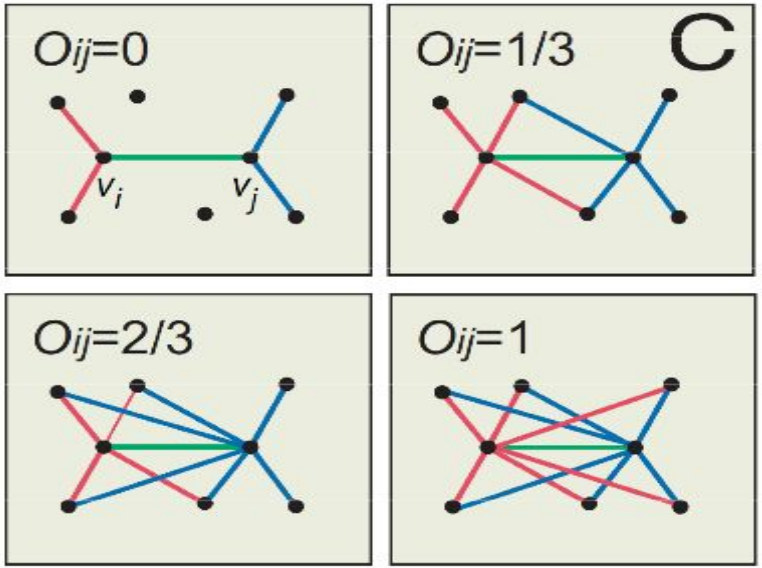
Measuring Tie Strength

Neighborhood Overlap

$$\mathbf{O}_{\mathrm{ij}} = \frac{n(\mathrm{i}) \cap \mathrm{n}(\mathrm{j})}{\mathrm{n}(\mathrm{i}) \cup \mathrm{n}(\mathrm{j})}$$

where n(i) is the neighbor set of i

If Overlap = 0 the edge is a local bridge

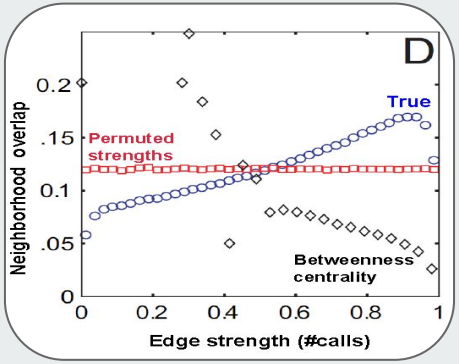


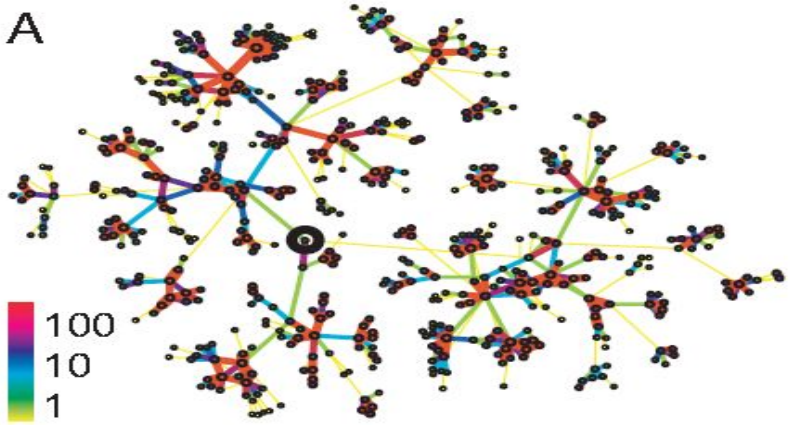
J.-P. Onnela, J. Saramaki, J. Hyvonen, G. Szabo, D. Lazer, K. Kaski, J. Kertesz, A.-L. Barabási. Structure and tie strengths in mobile communication networks. PNAS 104 (18), 7332-7336 (2007).

Strength, overlap and betweenness

Overlap and Betweenness are inversely correlated:

- Weak ties: the higher the number of shortest paths crossing an edge, the lower the overlap among the endpoints of the edge



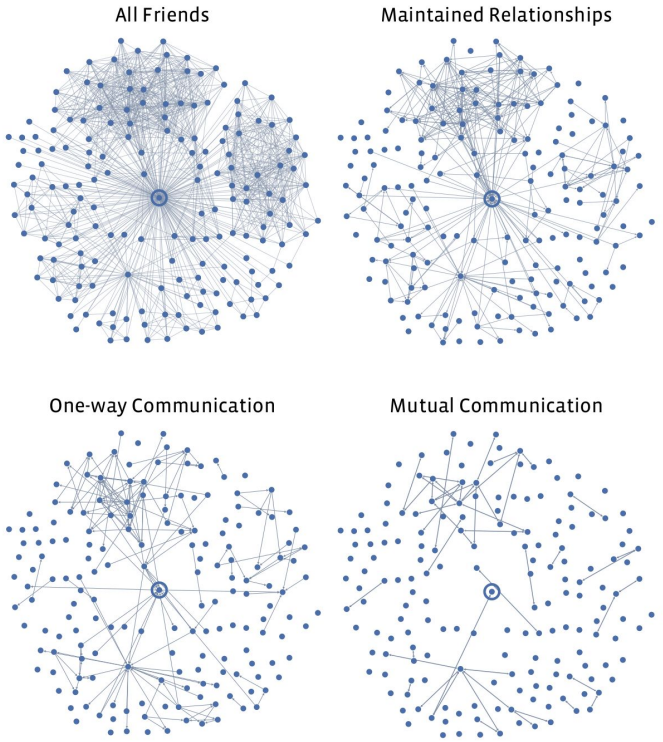


J.-P. Onnela, J. Saramaki, J. Hyvonen, G. Szabo, D. Lazer, K. Kaski, J. Kertesz, A.-L. Barabási. Structure and tie strengths in mobile communication networks. PNAS 104 (18), 7332-7336 (2007).

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)

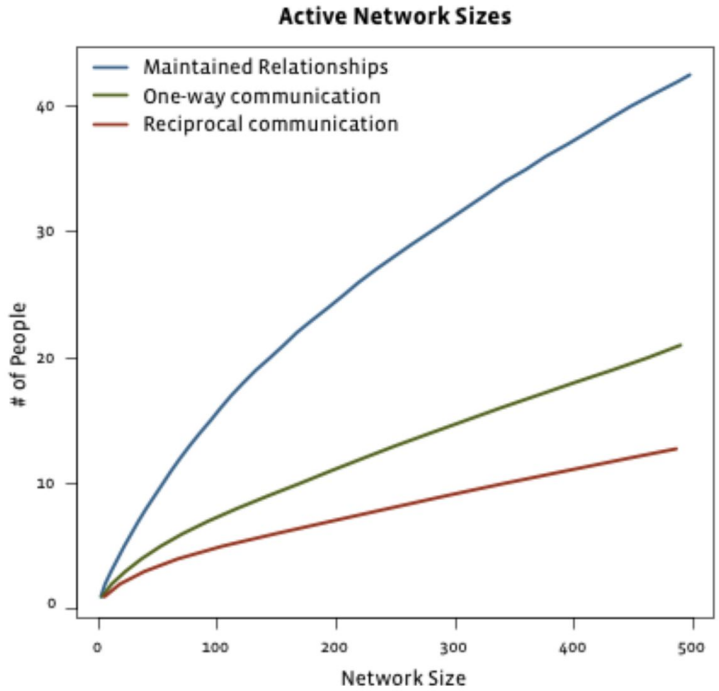


Cameron Marlow, Lee Byron, Tom Lento, and Itamar Rosenn. Maintained relationships on Facebook, 2009. http://overstated. net/2009/03/09/maintainedrelationships-on-facebook

Does tie strength affect network size?

Tie strength allows to:

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



Chapter 5

Conclusion

Take Away Messages

1. In social contexts individuals tend to cluster following homophilic patterns
2. Different topologies suffers from different vulnerabilities (node failures & attacks)

Suggested Readings

- Chapters 3 & 4 of Kleinberg's book

What's Next

Chapter 6:
Meso: Community Discovery

Chapter 6

Meso: Community Discovery

Summary

- What's a Community?
- Communities in static networks
- Evaluation & Benchmarking

Reading

- Chapter 9 of Barabasi's book
- Fortunato's survey

Community Discovery

A brief Introduction

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

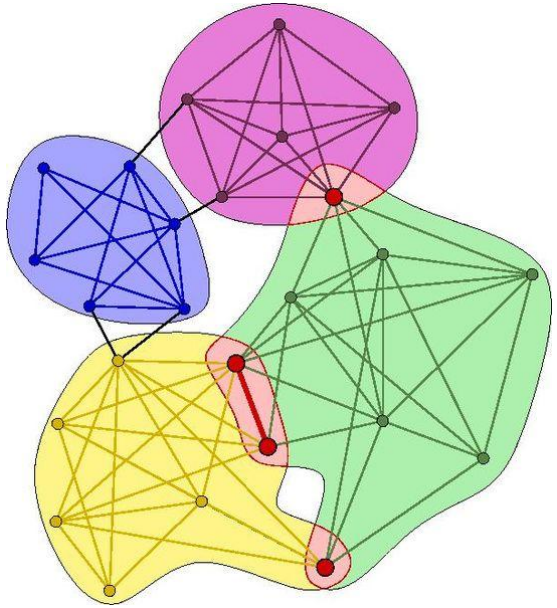
S. Fortunato. Community detection in graphs. Physics Reports 486 (2010).

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."*

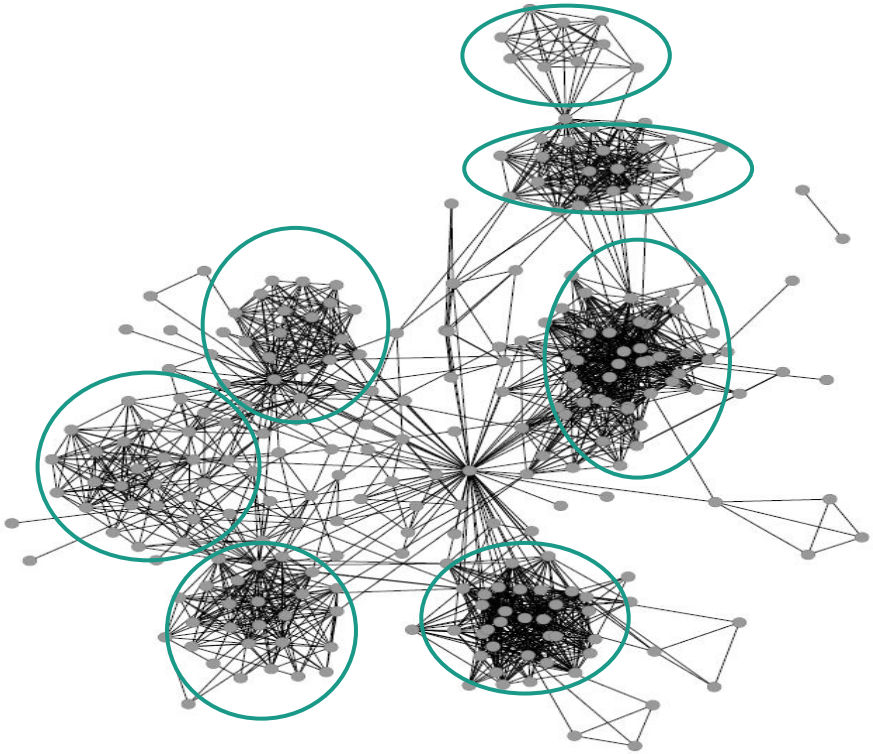
or, equivalently

> *"A set of nodes more 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…

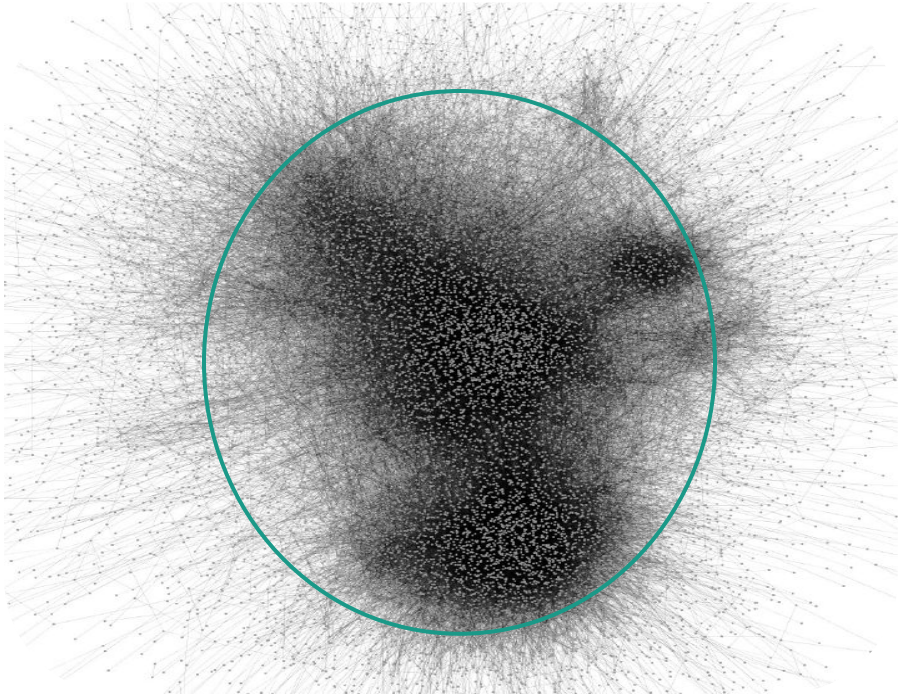
- i.e., using a Force directed layout



Real world networks? Too complex for visual analysis

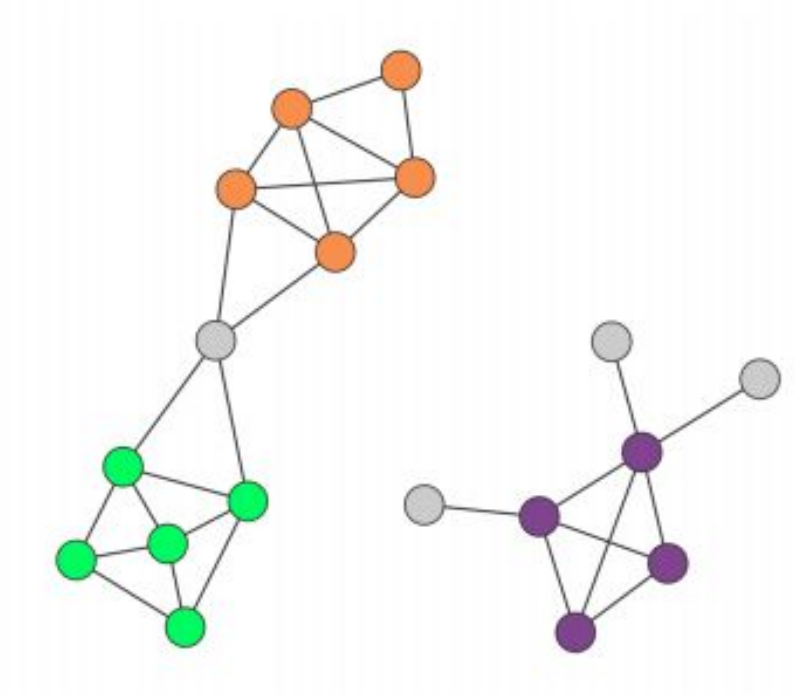
We can't easily identify (e.g., visually) different communities

We need automated procedures!

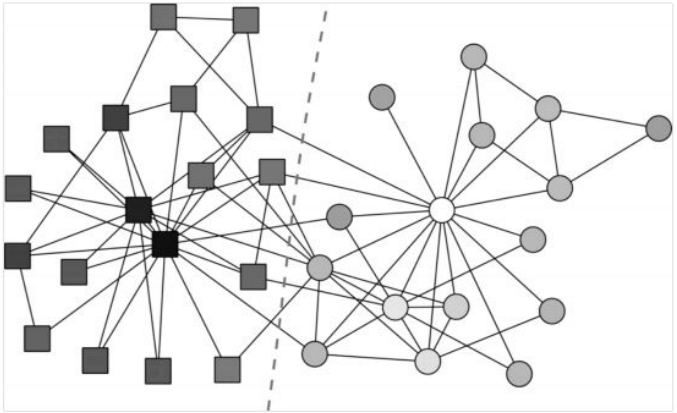


Communities: a few Hypothesis

- **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



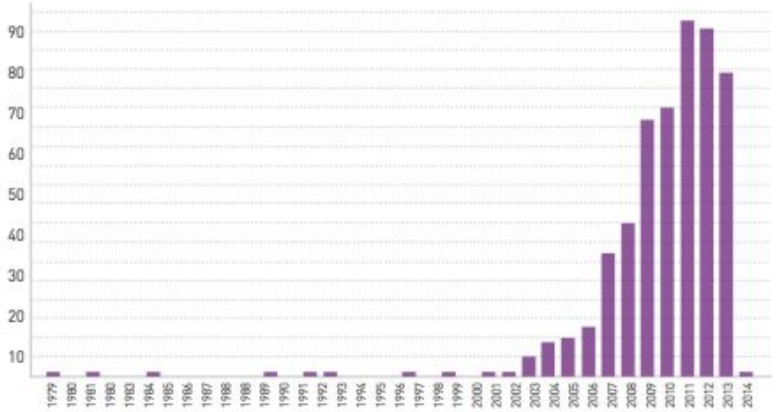
A first example…



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/

Community Discovery

The nightmare of an ill-posed problem

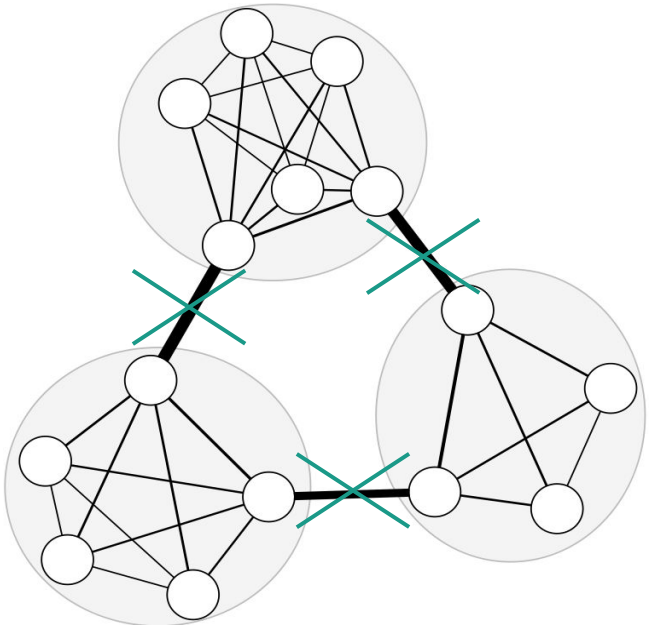
(a few algorithms and their rationale…)

Idea

Bridge Detection

"Communities as components of the network obtained by removing bridges"

Partitioning, usually top-down, approaches



Algorithms in this family:

- Girvan Newman (edge betweenness), ...

Algorithm

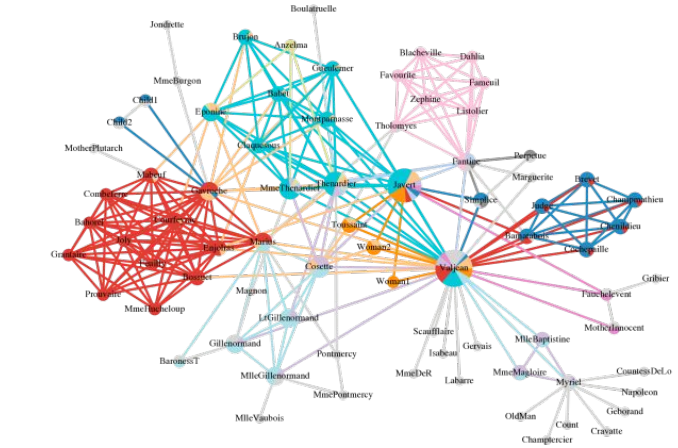
Girvan-Newman

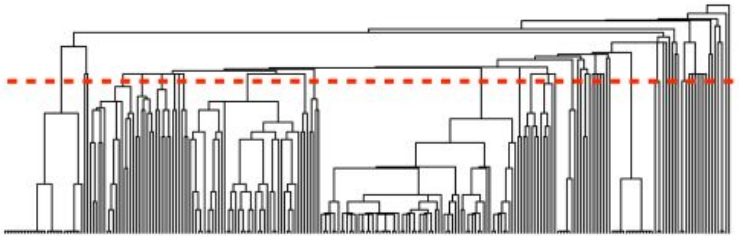
Steps

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.



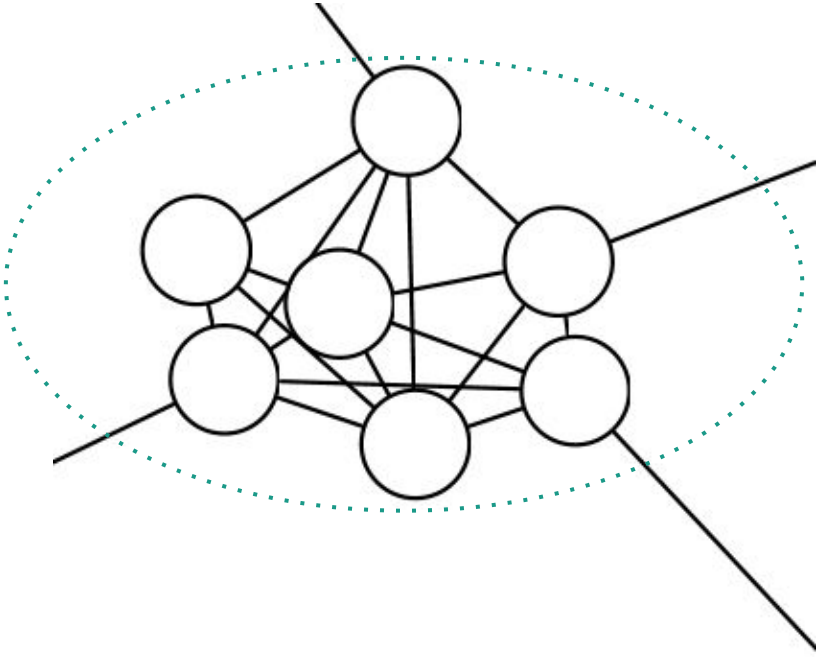


Idea

Internal Density

"Communities as a sets of densely connected entities"

Each community must have a number of edges significantly higher than what expected in a random graph



Algorithms in this family:

- Greedy Modularity, Louvain, ...

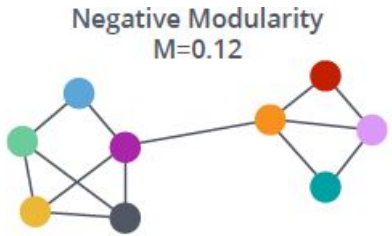
Idea

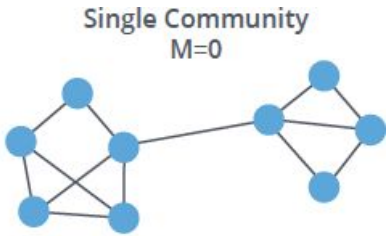
Internal Density

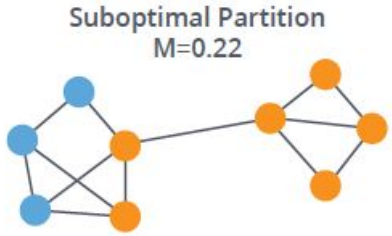
How to assure high density?

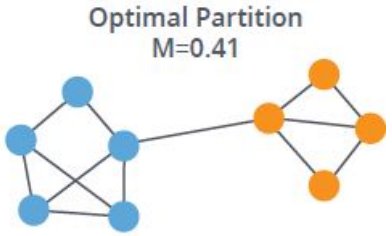
General Idea:

- define a quality function that measures the density of a community and then try to maximize it









Modularity [-1, 1]

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

Null Model
expected density

1 if i,j in same community,
0 otherwise

Algorithm

Louvain

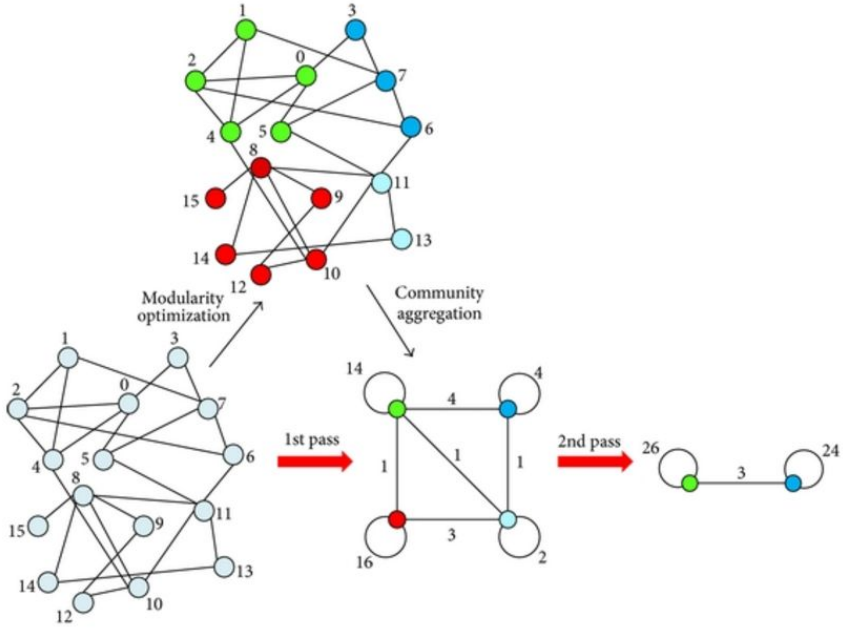
In order to maximize this value efficiently, the Louvain Method has two phases that are repeated iteratively.

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

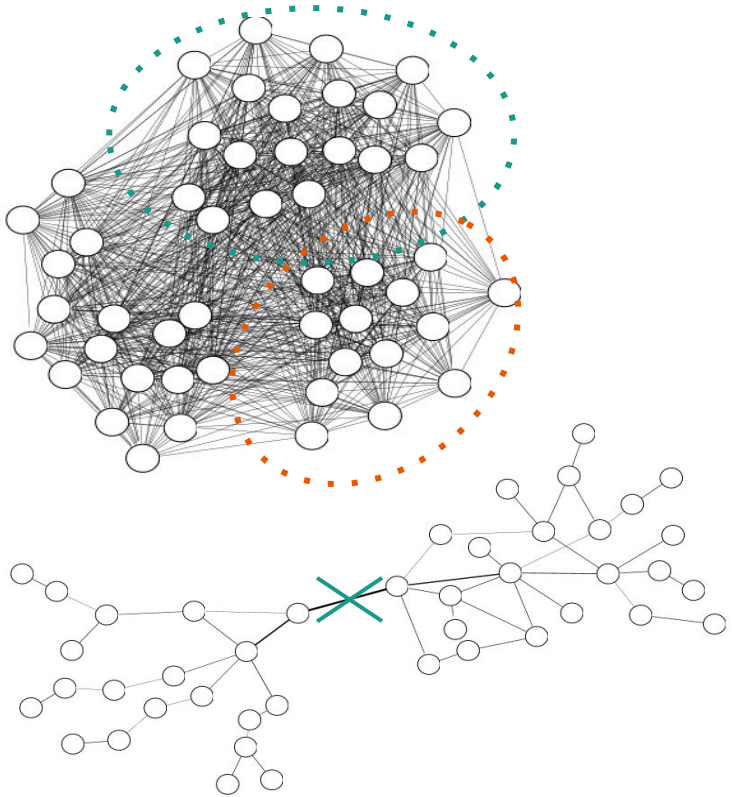


VD Blondel, et al. Fast unfolding of communities in large networks. Journal of statistical mechanics: theory and experiment (2008)

Density Vs. Bridges

These two definitions seems very similar...
Are they equivalent?

- In some networks yes;
- In dense network there are no clear bridges.
- For very sparse networks a density definition will fail, even if we can detect some bridges

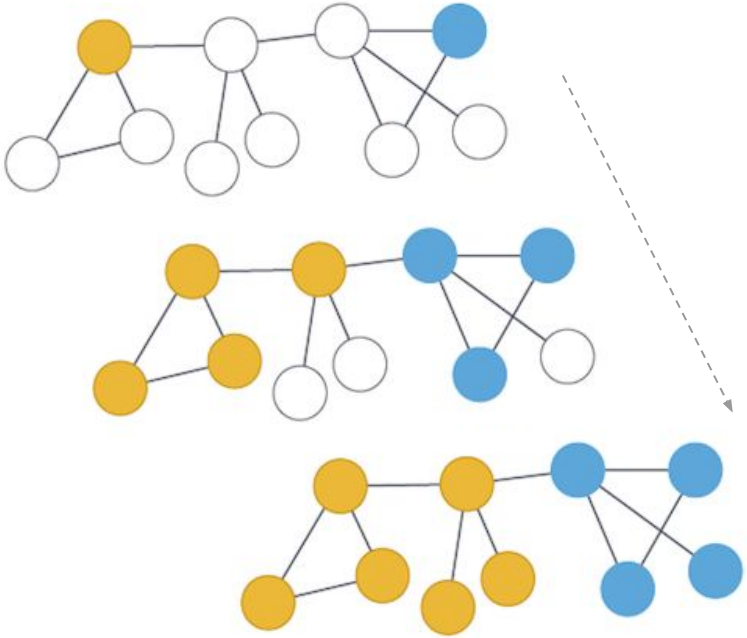


Idea

Percolation

"Communities as sets of nodes grouped together by the propagation of a same property, action or information"

Usually percolation approaches do not optimize an explicit quality function.



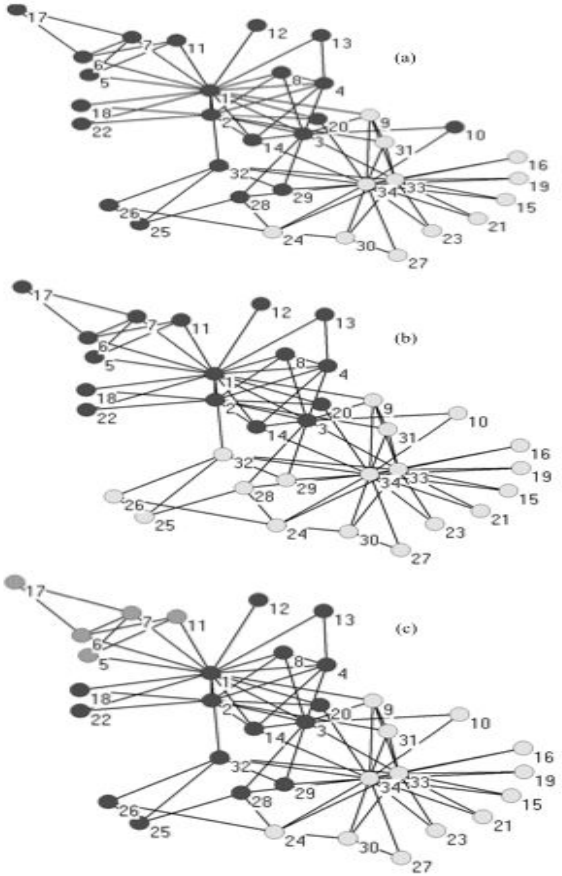
Algorithms in this family:

- Label Propagation
- Demon, Angel
- ...

Algorithm

Label Propagation

1. Each node has an unique label (i.e. its id)
2. In the first (setup) iteration each node, with probability α, change its label to one of the labels of its neighbors;
3. At each subsequent iteration each node adopt as label the one shared *(at the end of the previous iteration)* by the majority of its neighbors;
4. We iterate until consensus is reached.



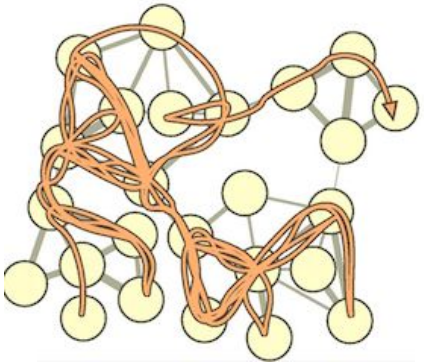
Idea

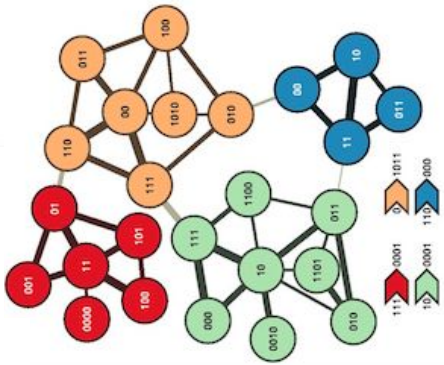
Entity Closeness

"Communities as sets of nodes that can reach any member of their group crossing a very low number of edges, significantly lower than the average shortest path in the network"

Idea:

Minimize the distances among nodes, implicitly avoiding the presence of bridges within communities





Algorithms in this family:

- Infomap (Conductance Optimization)
- ...

Algorithm

Infomap

The core of the algorithm follows closely the Louvain method:

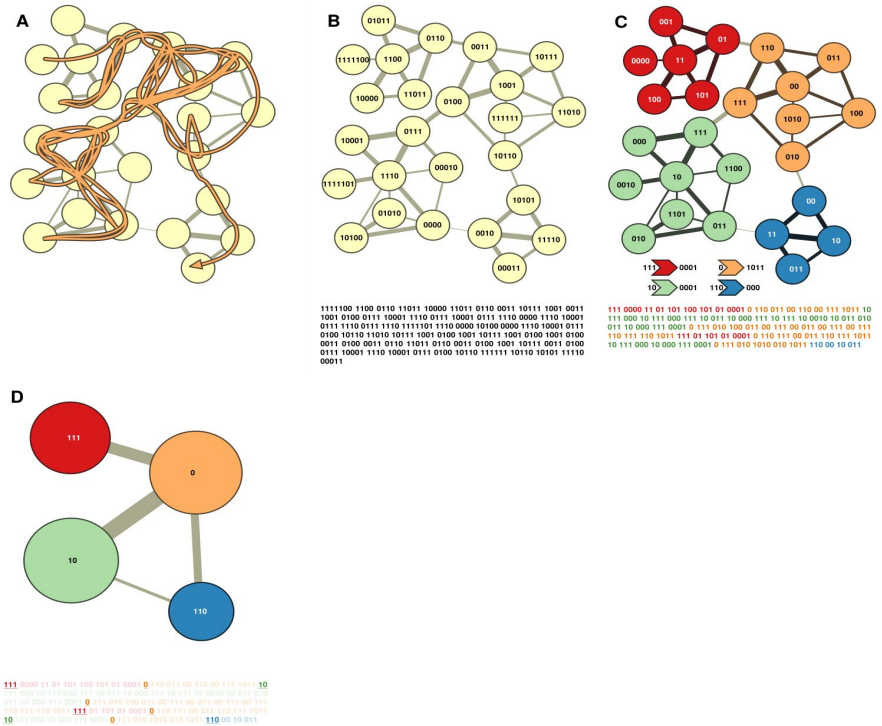
- Phase 1:
  Each node is moved to the neighboring module that results in the largest decrease of the map equation.

- 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
- $a(S) = \sum_{i \in S} \sum_{j \in V} a_{ij}$



Idea

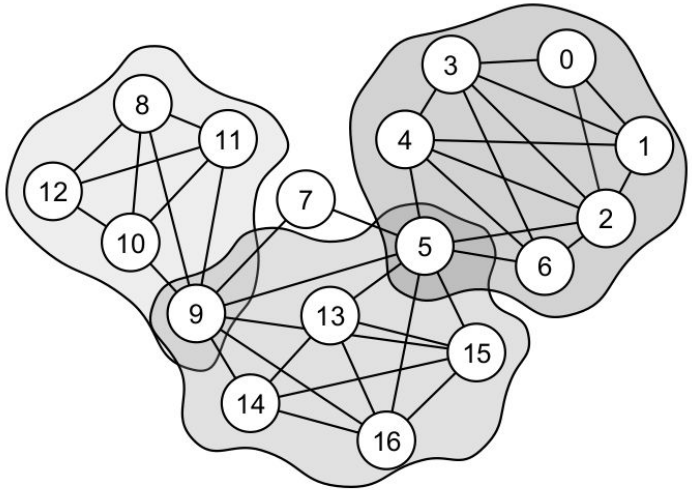
Structure Definition

“Communities as sets of nodes having a precise number of edges among them, distributed in a precise topology defined by a number of rules”

Idea:

Identify precise patterns within a network

(e.g., cliques, quasi-cliques, ...)

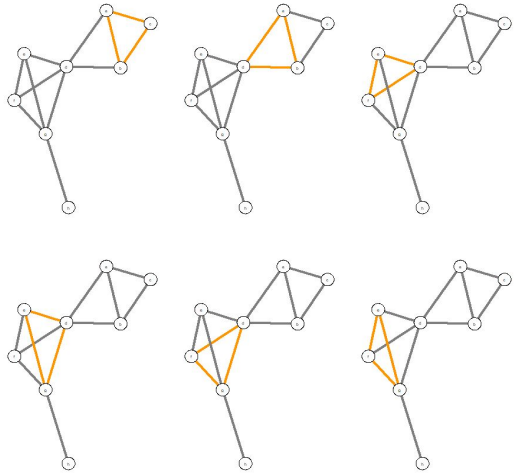


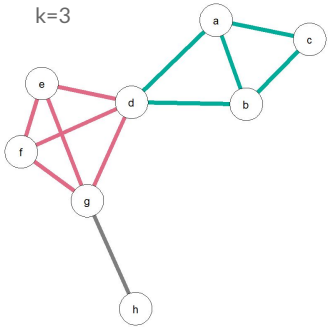
Algorithms in this family:

- k-cliques, ...

Algorithm

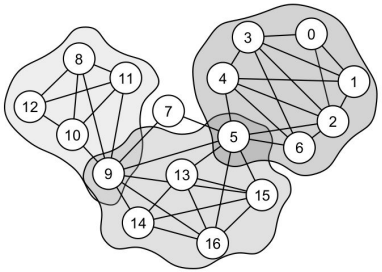
k-Cliques





A very popular algorithm: **k-cliques**

- Also this case is different from the density definition: node 7 is in some sense “dense” (is in a triangle), but outside of any community



Algorithm steps:

1. Identify *k-cliques*, which are fully connected networks with k nodes. (The smallest possible k would be k = 3. Otherwise, the cliques would be only edges.)

2. A community is defined as a set of adjacent k-cliques, that is, k-cliques that share exactly *k-1* nodes.
   With k = 3, two 3-cliques are adjacent if they share exactly two nodes (equivalent to an edge).

Community Discovery

Evaluation strategies

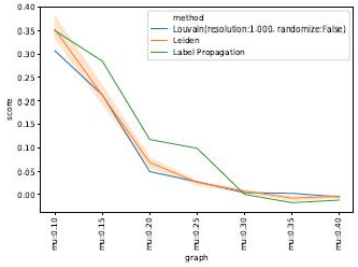
Strategies

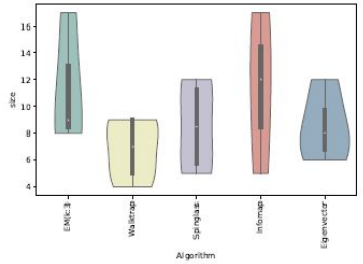
Internal Evaluation

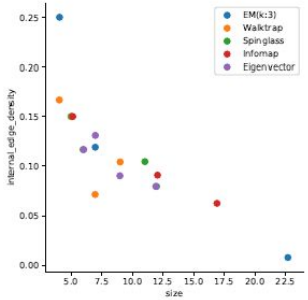
- Partition quality function
  (i.e., modularity, conductance, density…)
- Community characterization
  (i.e., size distribution, overlap distribution…)
- Execution time and Complexity

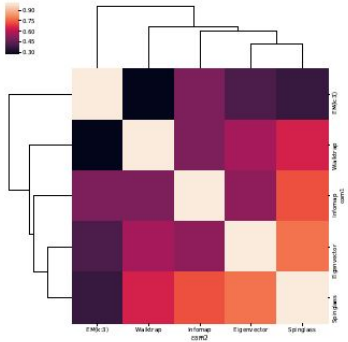
External Evaluation

- Ground truth testing
  (or partitions comparison)







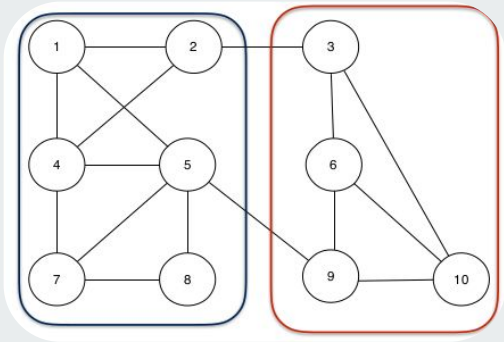


Quality Functions

Internal Evaluation

Several fitness functions can be defined to assess the quality of a partition.

Usually, the best partition is the one that maximize (or minimize) a given fitness function in its worst case scenario (i.e., when computed on the worst community identified)



Approx. formulae
(for exercise only)

Internal Edge Density

$$\frac{2|E_C|}{|V_C|(|V_C|-1)}$$

E_C edges within C
V_C nodes within C

Worst case: **min**

Best-worst case: **max**

Average Node Degree

$$\frac{1}{|V_C|}\sum_{i \in C} d_i$$

d_i degree of node i

Worst case: **min**

Best-worst case: **max**

Modularity

$$\left(\frac{|V_c|}{|E|} - \frac{\deg_C}{2|E|}\right)^2$$

$\deg_C$ sum of degrees within C

$$\deg_C = \sum_{i \in C} d_i$$

Worst case: **min**

Best-worst case: **max**

Conductance

$$\frac{2|E_{OC}|}{2|E_C| + |E_{OC}|}$$

E_{OC} edges out of C

Worst case: **max**

Best-worst case: **min**

Yang, Jaewon, and Jure Leskovec. "Defining and evaluating network communities based on ground-truth." Knowledge and Information Systems 42.1 (2015): 181-213.

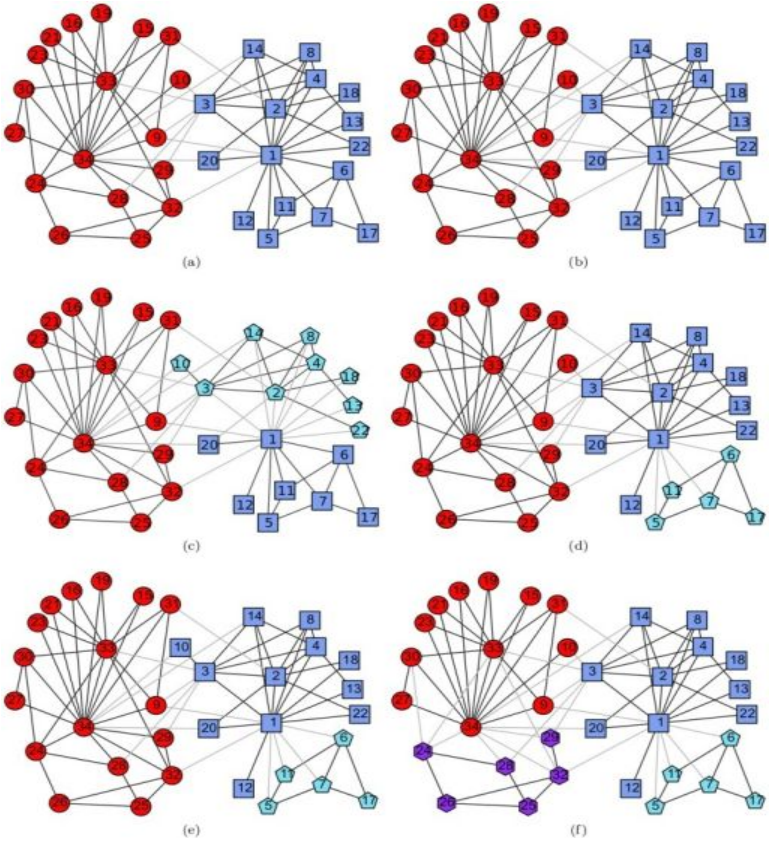
Ground truth testing

External Evaluation

Given a graph G, a ground truth partition P(G) and the set of identified communities C estimate the resemblance the latter has with P(G).

General Criticism(s)

- Different approaches generates communities following different criteria (“ill posed” problem”)
- It is not necessarily true that the ground truth represent the only valid semantic\topological partition for the analyzed graph.



Peel, et a. "The ground truth about metadata and community detection in networks." Science advances 3.5 (2017): e1602548.

External Evaluation

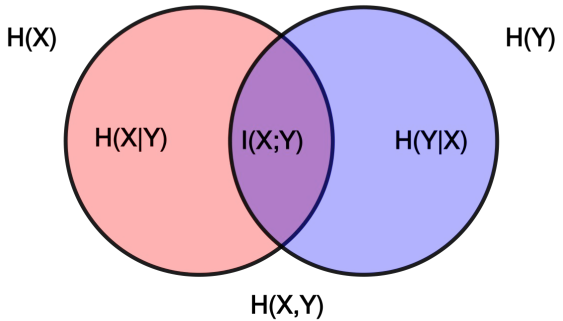
NMI

Normalized Mutual Information is a measure of *similarity* borrowed from information theory:

$$NMI(X,Y) = \frac{H(X)+H(Y)-H(X,Y)}{\frac{H(X)+H(Y)}{2}} \in [0,1]$$

- *H(X)* is the entropy of the random variable *X* associated to an identified community,
- *H(Y)* is the entropy of the random variable *Y* associated to a ground truth community,
- *H(X,Y)* is the joint entropy.

The higher the NMI the more similar the compared partitions are



Advantages

- Extensively used in literature

Drawbacks

- Computational complexity ~ $O(|C|^2)$ (where C is the community set)
- Needs to be approximated in case of overlapping partitions

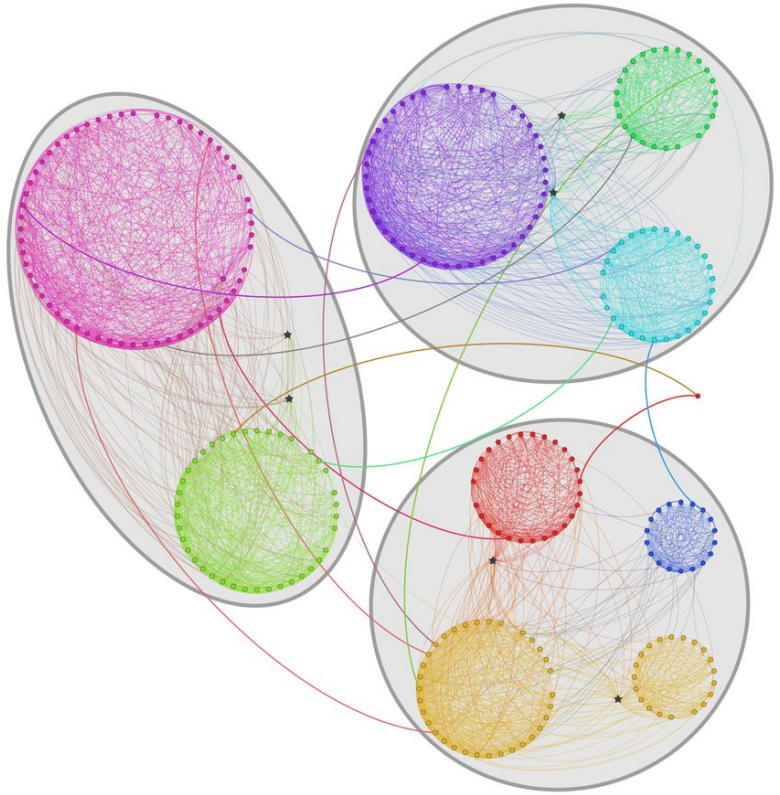
Synthetic Benchmarks

External Evaluation

Testing against topological ground truths

Synthetic graphs with embedded community structure (e.g., LFR)

- More stable than semantic ground truth partitions
- Community structure depends on the fitness function optimized by the chosen model
- Approximation of real world networks



Lancichinetti, Andrea, Santo Fortunato, and Filippo Radicchi. "Benchmark graphs for testing community detection algorithms." Physical review E 78.4 (2008): 046110.

Summarizing

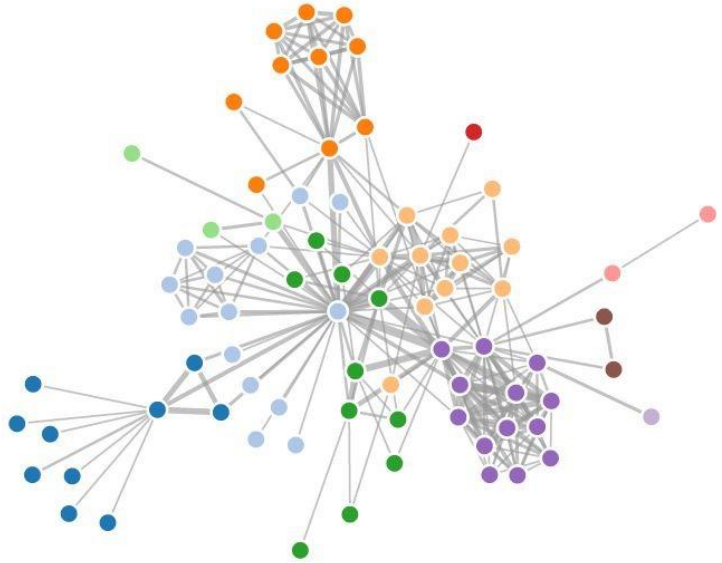
Community Discovery is, perhaps, the hottest topic in complex network analysis

Major issues:

- Problem definition
- Community evaluation

Problem specializations:

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



Chapter 6

Conclusion

Take Away Messages

1. Complex networks are composed by hidden meso-scale structures
2. Identify them is not a trivial task
3. Evaluate them is not a trivial task
4. Knowing them is fundamental to identify functional modules of a system

Suggested Readings

- Chapter 9 of Barabasi's book
- Fortunato's survey

What's Next

Chapter 7:
Macro: Assortativity & Resilience

Chapter 7

Macro: Assortativity & Resilience

Summary

- Do Birds of a feather flock together?
- Resilience/Robustness
- Failures and Attacks

Reading

- Chapters 3 & 4 of Kleinberg's book
- Chapter 8 of Barabasi's book

Do Birds of a Feather Flock Together?

Homophilic behaviors in complex networks

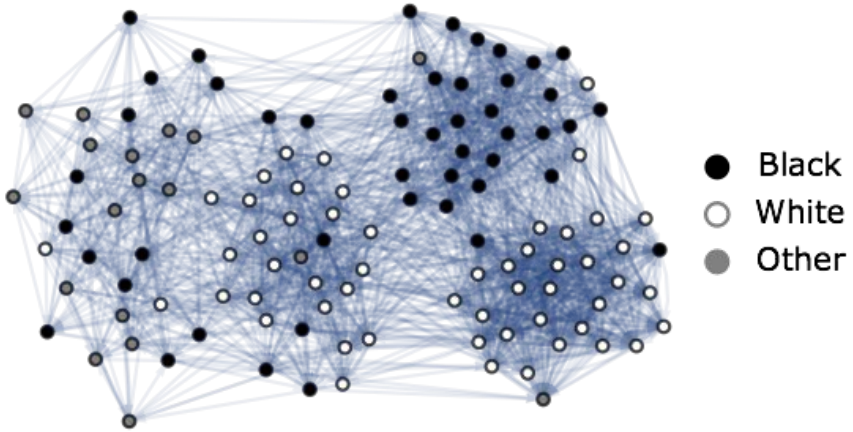
Homophily

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 (categoric) 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	

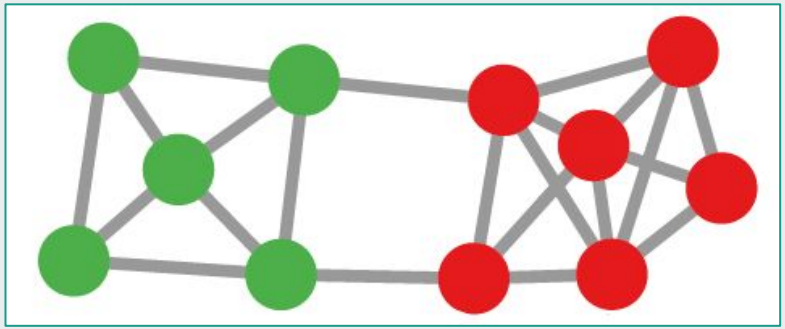
$$r = \frac{\sum_i e_{ii} \quad \sum_i a_i b_i}{\frac{(0.258 + 0.157 + 0.306 + 0.016) - ((0.323 \cdot 0.289) + (0.247 \cdot 0.204) + (0.377 \cdot 0.423) + (0.053 \cdot 0.083))}{1 - ((0.323 \cdot 0.289) + (0.247 \cdot 0.204) + (0.377 \cdot 0.423) + (0.053 \cdot 0.083))} = 0.621}$$
$$\sum_i a_i b_i$$

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

Assortative Mixing

Newman's assortativity - Example

Compute Newman's assortativity given the network below:



In undirected network $a_i = b_i$

Assortativity Matrix

	Green	Red	a_i
Green	0.363	0.090	0.453
Red	0.090	0.545	0.635
b_i	0.453	0.635	

Assortativity Index

$$r = \frac{((8/22)+(12/22))-((10/22)^2+(14/22)^2)}{1-((10/22)^2+(14/22)^2)} = \text{~0.766}$$

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

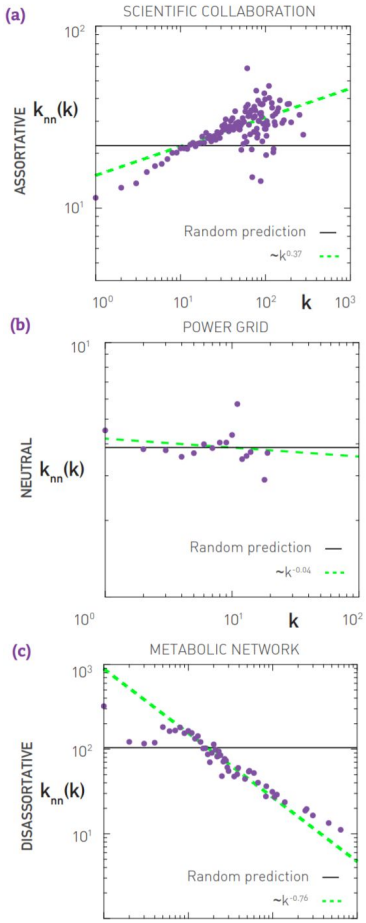
Assortativity by degree w.r.t. different network types

Degree correlation
(insights)

- Using a single number: Newman's assortativity
- Plotting $k_{nn}(k)$ in function of k, where

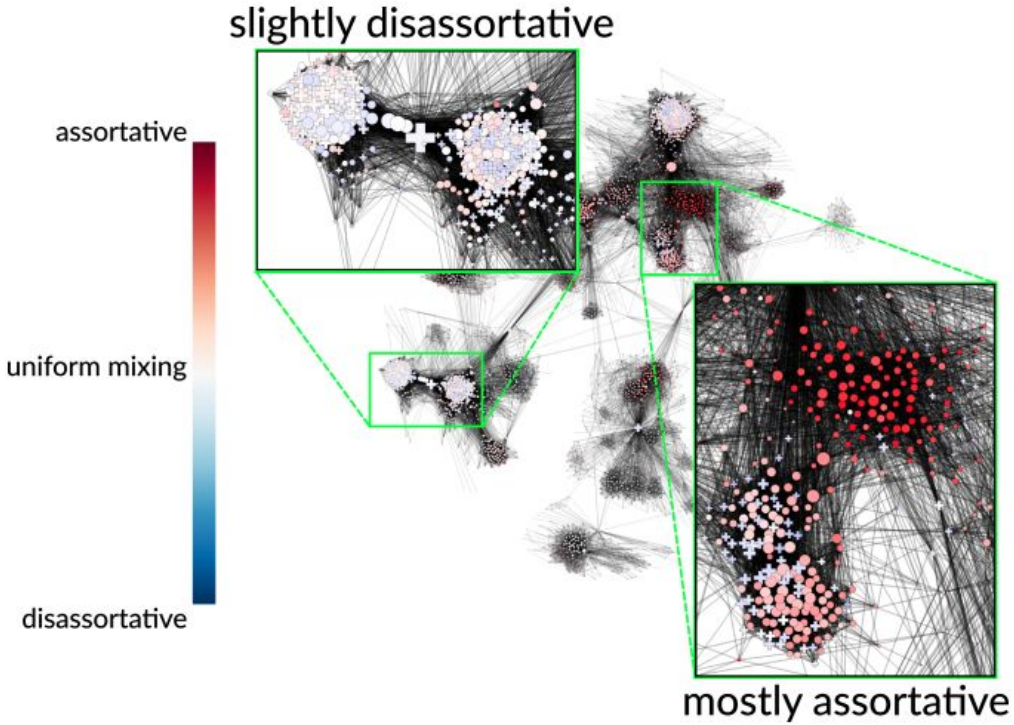
$$k_{nn}(k) = \sum_{k'} k' P(k' \mid k) \, ,$$

is a degree correlation function, e.g., the average degree of the neighbors of all degree-k nodes.



Is a Global Measure enough?

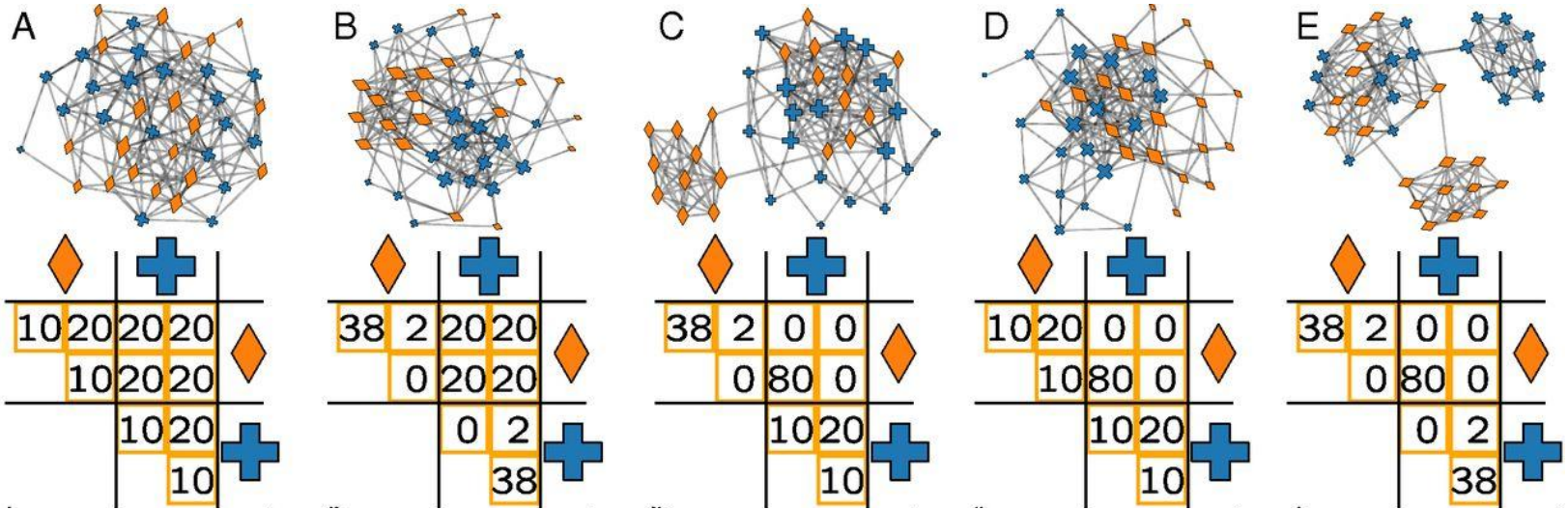
“Sure I can work with the means, but I'd rather party with the outliers...”



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



Five networks (top) of n=40 nodes and m=160 edges with the same global assortativity r=0

Moving toward a multiscale approach to measure assortativity

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?)

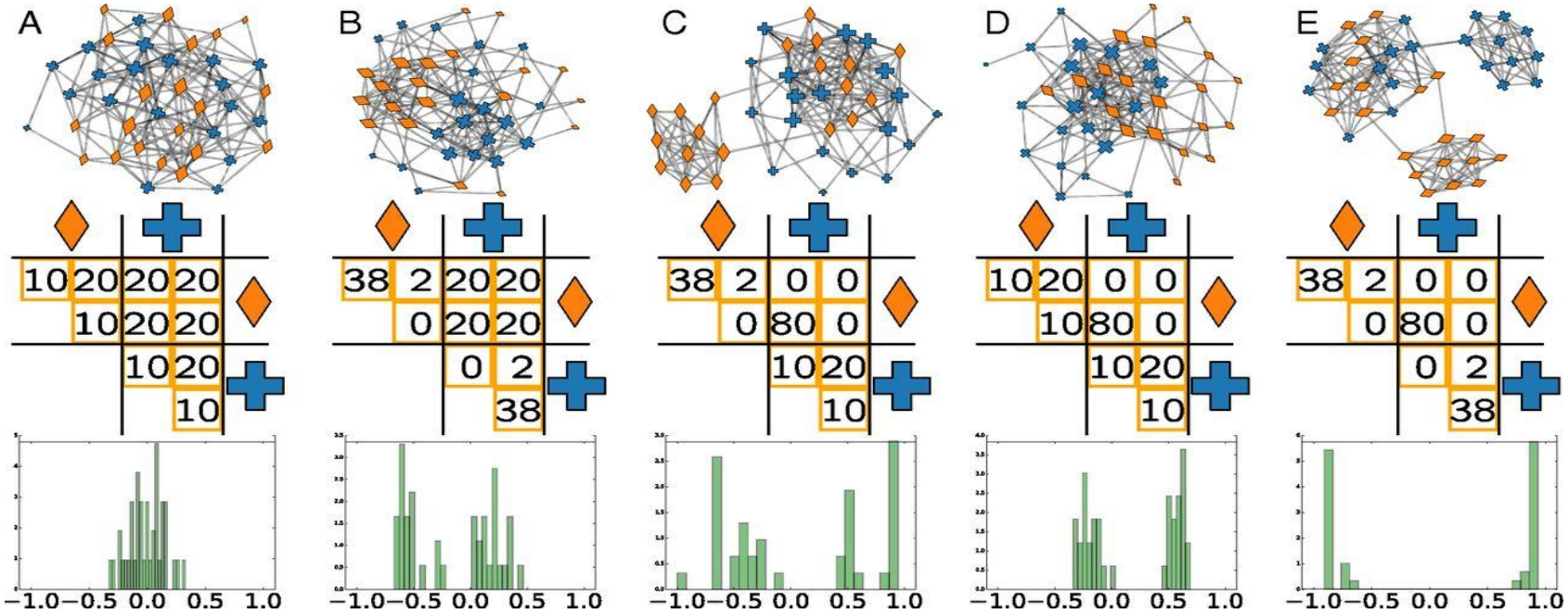
Better approximation:
Considering the stable state of a RW (probability to reach a given node) to weight the edges

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)

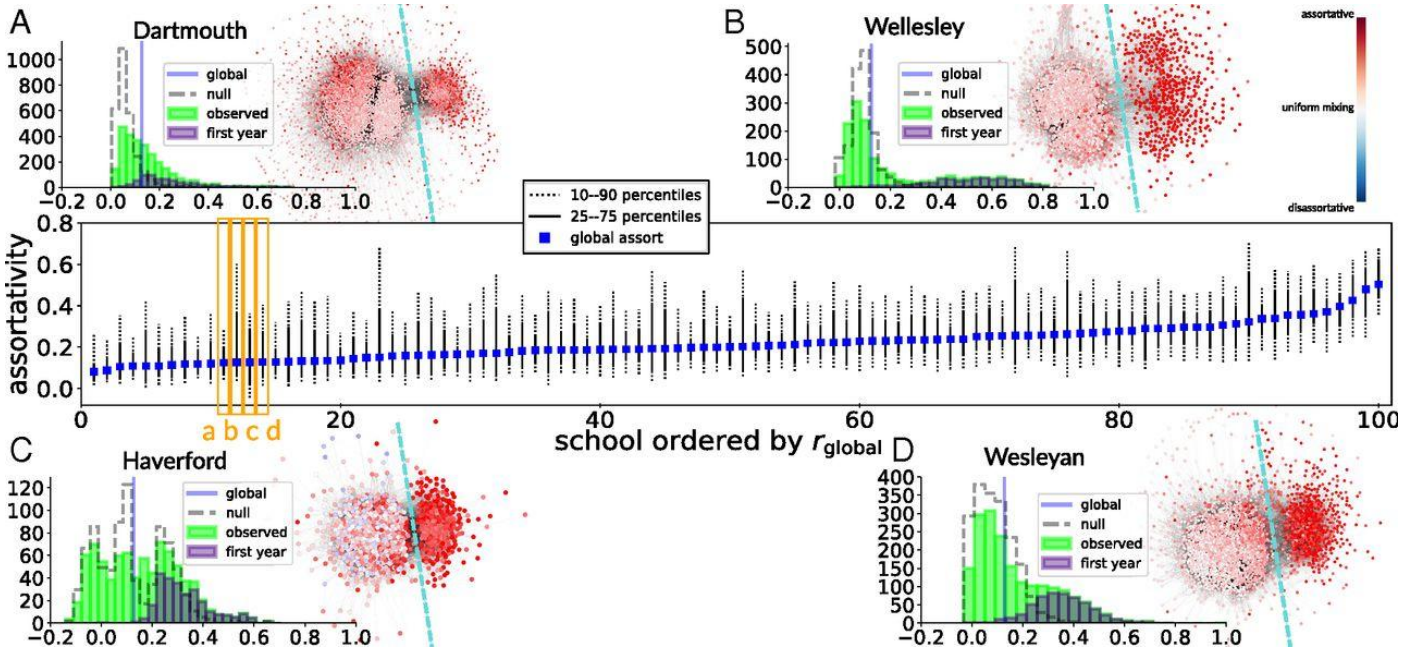
$$w_{\text{multi}}(i;\ell) = \int_0^1 w_\alpha(i;\ell)\,\mathrm{d}\alpha$$

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.



Peel's Quintet: Only A has a random-like assortment distribution

Evaluation on real data



Facebook100

Distribution of local assortativity for the "dorm" node feature

Network Resilience

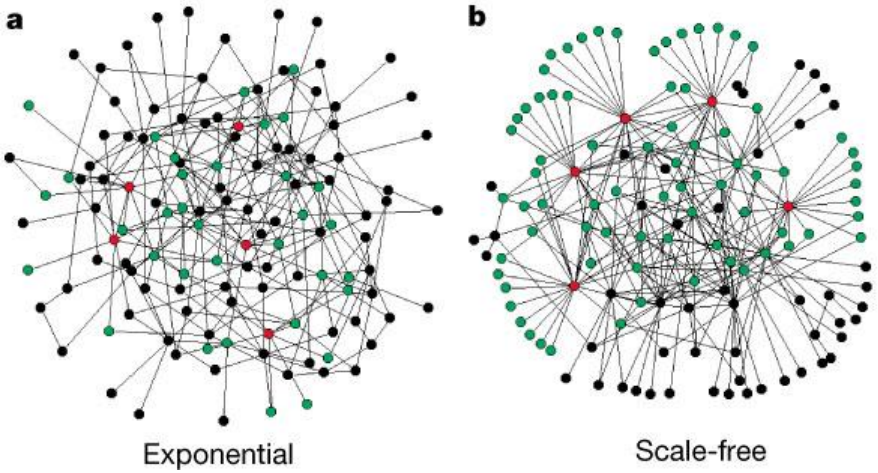
How robust is a complex network to node failures/attacks?

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)

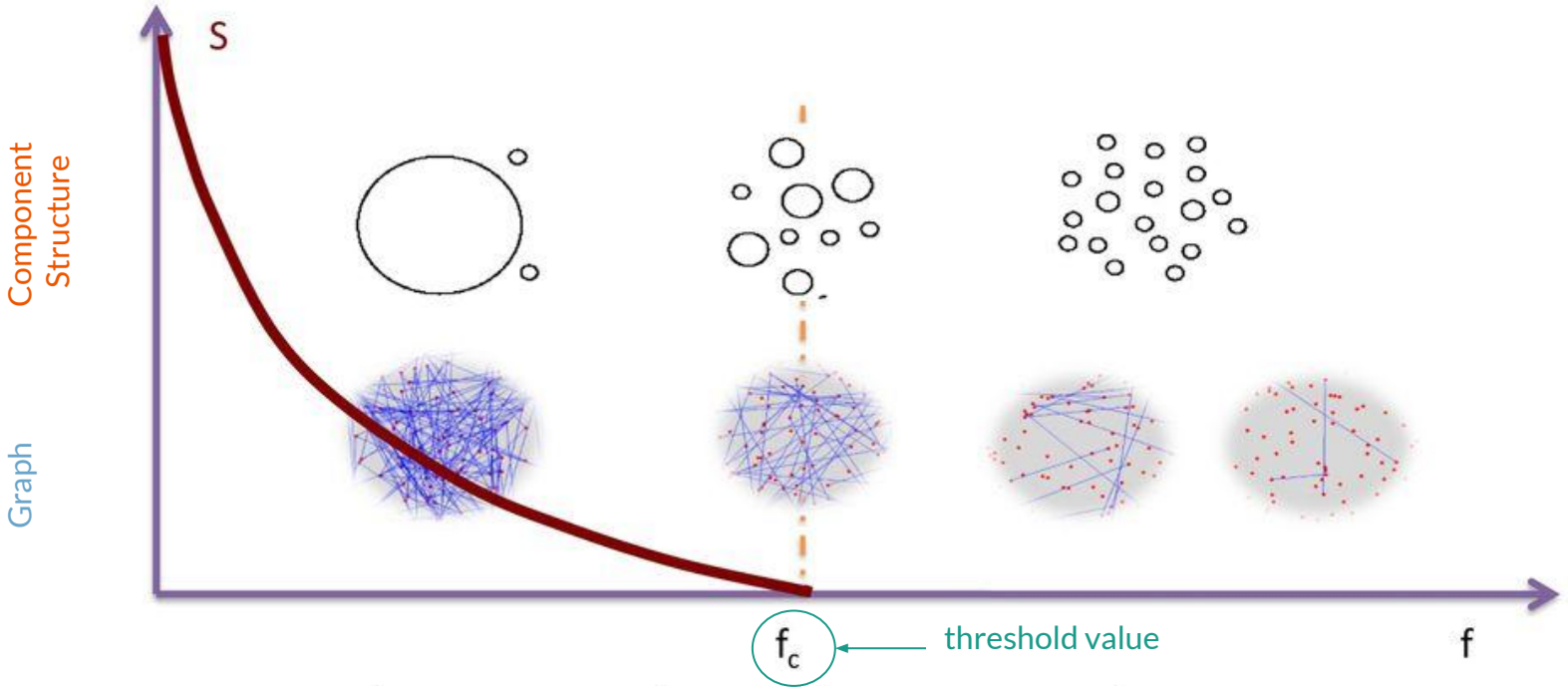


Different topologies, same parameters (N=130, <k>=3.3)

Node removal strategies:

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

f = fraction of removed nodes



Inverse Percolation problem

Molloy-Reed criterion for giant components

A giant cluster exists if each node is connected to at least two other nodes.

Or, equivalently:
The average degree of a node *i* linked to the GC, must be at least 2.

Can be shown to correspond to the following relation:

$$\kappa \equiv \frac{\langle k^2 \rangle}{\langle k \rangle} = 2$$

second moment of the degree distribution (i.e., variance)

- κ>2: Giant component exist
- κ<2: Many disconnected cluster

Malloy, Reed, Random Structures and Algorithms (1995); Cohen et al., Phys. Rev. Lett. 85, 4626 (2000).

Breakdown threshold

ER graphs

Random node removal changes

- The degree of individual nodes decrease by losing links via node removal (e.g., [k -> k' ≤ k])
- A node with degree k becomes a node with degree k' with probability

$$\binom{k}{k} f^{k-k'}(1-f)^{k'} \qquad \text{where } k' \le k$$

Remove k-k' links, each with probability f

Leave k' links untouched, each with probability 1-f

the degree distribution [P(k) -> P'(k')] after random removal of f fraction of nodes becomes

$$P'(k') = \sum_{k=k'}^{\infty} P(k) \binom{k}{k'} f^{k-k'}(1-f)^{k}$$

thus,

$$\langle k' \rangle_f = (1-f)\langle k \rangle$$

$$\langle k^2 \rangle_f = (1-f)^2 \langle k^2 \rangle + f(1-f)\langle k \rangle$$

Breakdown threshold

ER graphs

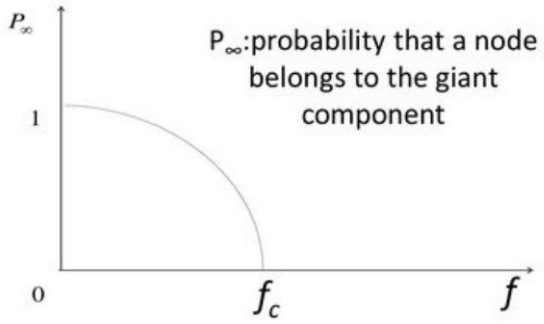
We know that,

$$\langle k' \rangle_f = (1-f)\langle k \rangle$$
$$\langle k^2 \rangle_f = (1-f)^2 \langle k^2 \rangle + f(1-f)\langle k \rangle$$
$$\kappa \equiv \frac{\langle k^2 \rangle}{\langle k \rangle} = 2$$

- κ>2: Giant component exist
- κ<2: Many disconnected cluster

Thus, the breakdown threshold becomes:

$$f_c = 1 - \frac{1}{\frac{\langle k^2 \rangle}{\langle k \rangle} - 1}$$



$$\kappa \equiv 1 - CN^{-\frac{3-\gamma}{\gamma-1}}$$

Breakdown threshold

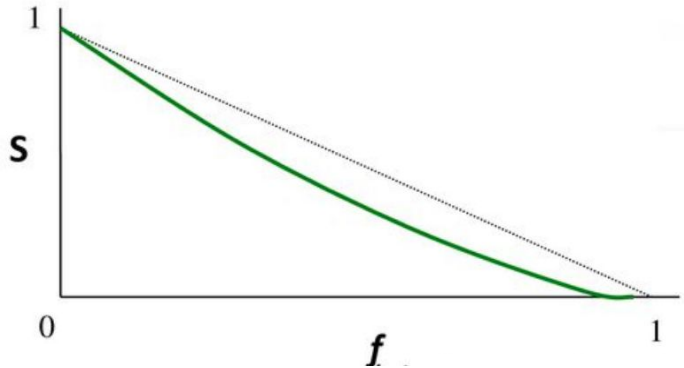
Resilience of Scale-free networks

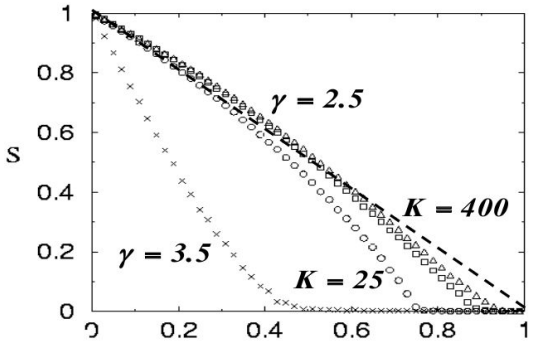
Scale-free graph with

$$P(k) = Ak^{-\gamma} \text{ with } k = m,\ldots,K$$

Scale-free networks do not appear to break apart under random failures.

Reason:
The likelihood of removing a hub is small.





Example Internet:

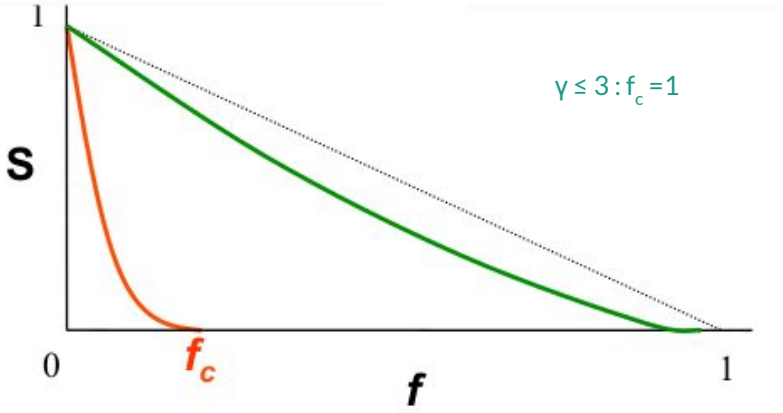
- Router level map, N=228,263, γ=2.1±0.1, k=28 -> f_c=**0.962**
- AS level map, N=11, γ=2.1±0.1, k=264 -> f_c=**0.996**

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:

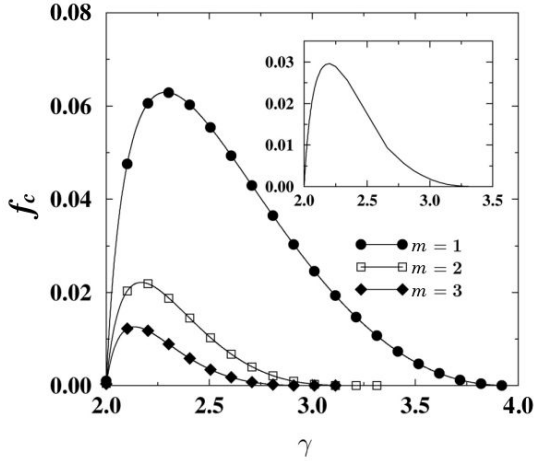
- Terrorist attacks
- Efficient vaccination in epidemics

Attack threshold for Scale-free networks

Attack problem:

- What if we remove a fraction *f* of the hubs?
- At what threshold f_c will the network fall apart (no giant component)?

$$f_c^{\frac{2-\gamma}{1-\gamma}} = 2 + \frac{2-\gamma}{3-\gamma} K_{\min} \left(f_c^{\frac{3-\gamma}{1-\gamma}} - 1 \right)$$



- f_c depends on γ;
  it reaches its max for γ<3
- f_c depends on K_{min}
  (m in the figure)
- f_c is tiny.
  Its maximum reaches only 6%, i.e. the removal of 6% of nodes can destroy the network in an attack mode.

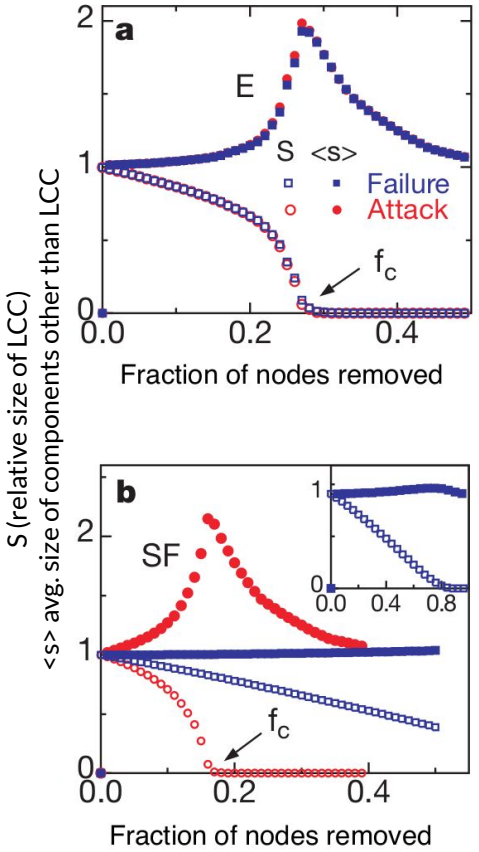
Example:
Internet γ=2.1, so 4.7% is the threshold

Network robustness & attack tolerance

Role of Hubs

Observations:

- Internet still works even when several servers are out of service
- Random vaccination is not effective in case of epidemic spreading



Poisson random graph

Both removal methods give the same result

The network falls apart after a finite fraction of nodes are removed

Scale-free network

Robust against random removal (blue)

Vulnerable against targeted attacks

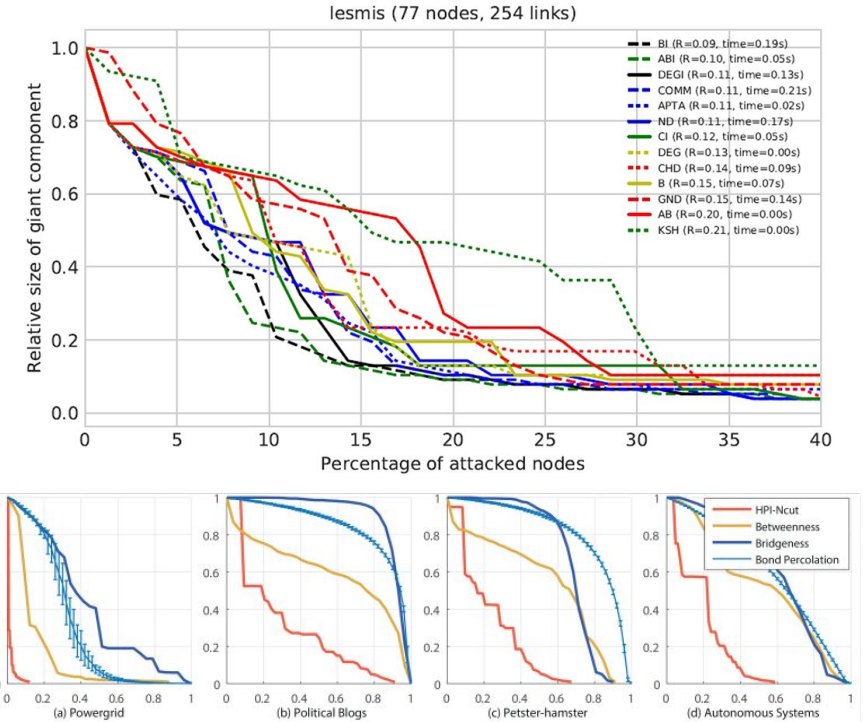
How to target a network?

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
- ...

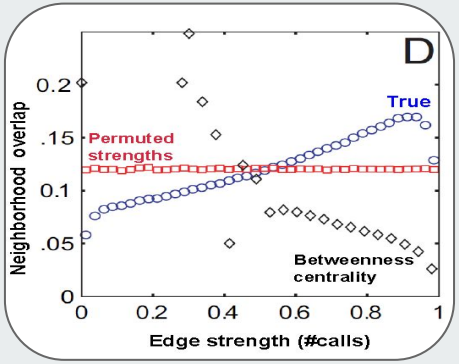


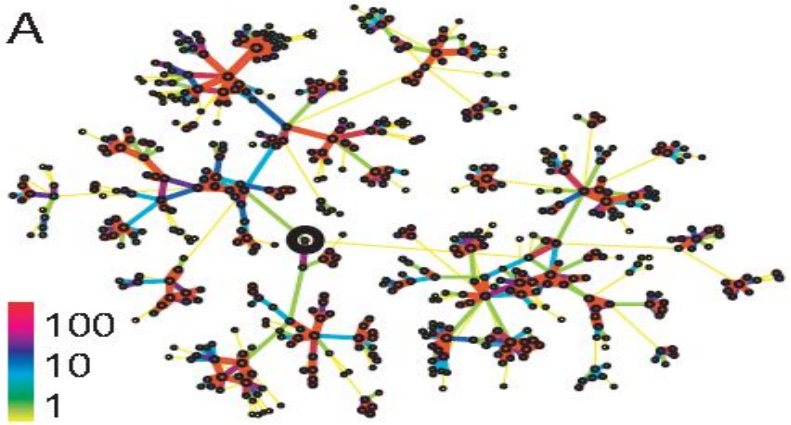
Ren, Xiao-Long, et al. "Underestimated cost of targeted attacks on complex networks." Complexity 2018 (2018).
Wandelt, Sebastian, et al. "A comparative analysis of approaches to network-dismantling." Scientific reports 8.1 (2018): 13513.

Strength, overlap and betweenness

Overlap and Betweenness are inversely correlated:

- Weak ties: the higher the number of shortest paths crossing an edge, the lower the overlap among the endpoints of the edge





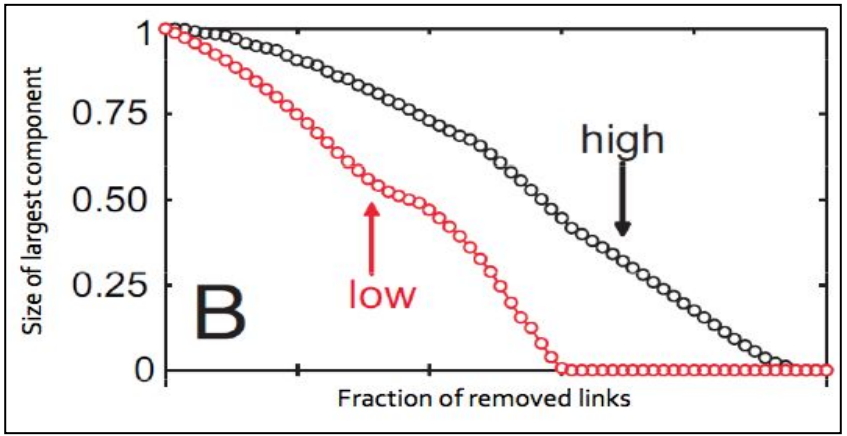
J.-P. Onnela, J. Saramaki, J. Hyvonen, G. Szabo, D. Lazer, K. Kaski, J. Kertesz, A.-L. Barabási. Structure and tie strengths in mobile communication networks. PNAS 104 (18), 7332-7336 (2007).

Removing links based on overlap

Two strategy of edge removal w.r.t. neighborhood overlap

- Low to high
- High to low

Removing bridges first (low overlap) destroy the network structure faster



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).

Chapter 7

Conclusion

Take Away Messages

1. In social contexts individuals tend to cluster following homophilic patterns
2. Different topologies suffers from different vulnerabilities (node failures & attacks)

Suggested Readings

- Chapters 3 & 4 of Kleinberg's book

What's Next

Lecture 4:
Dynamics of Networks