CAI: Cerca i Anàlisi d'Informació
Grau en Ciència i Enginyeria de Dades, UPC

8. Network Analysis

December 8, 2019

Slides by Marta Arias, José Luis Balcázar, Ramon Ferrer-i-Cancho, Ricard Gavaldà, Department of Computer Science, UPC

Contents

Examples of complex networks

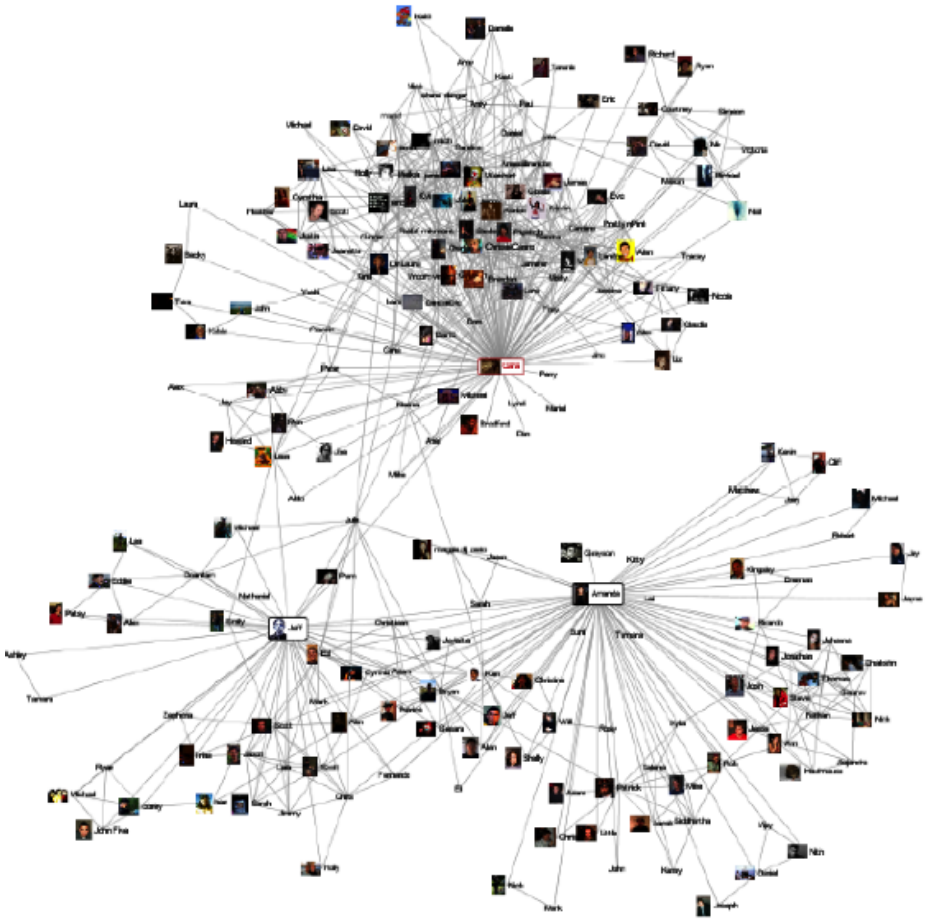
- Social networks
- Information networks
- Technological networks
- Biological networks

- The Web

Social networks

Links denote social “interactions”

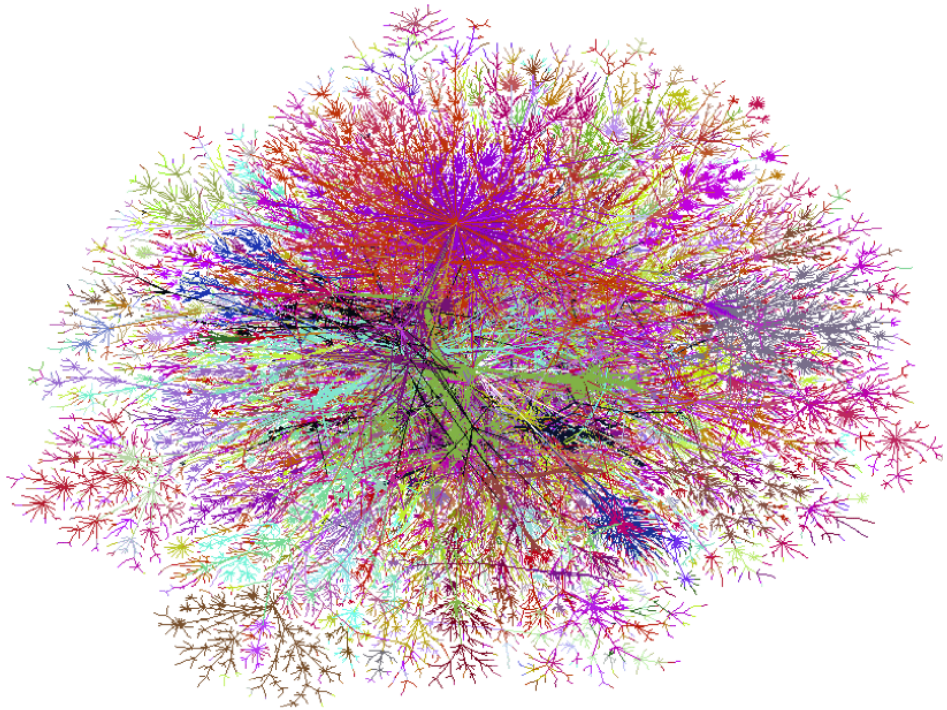
- friendship, collaborations, e-mail, etc.



Information networks

Nodes store information, links associate information

- citation networks, the web, p2p networks, etc.



Technological networks

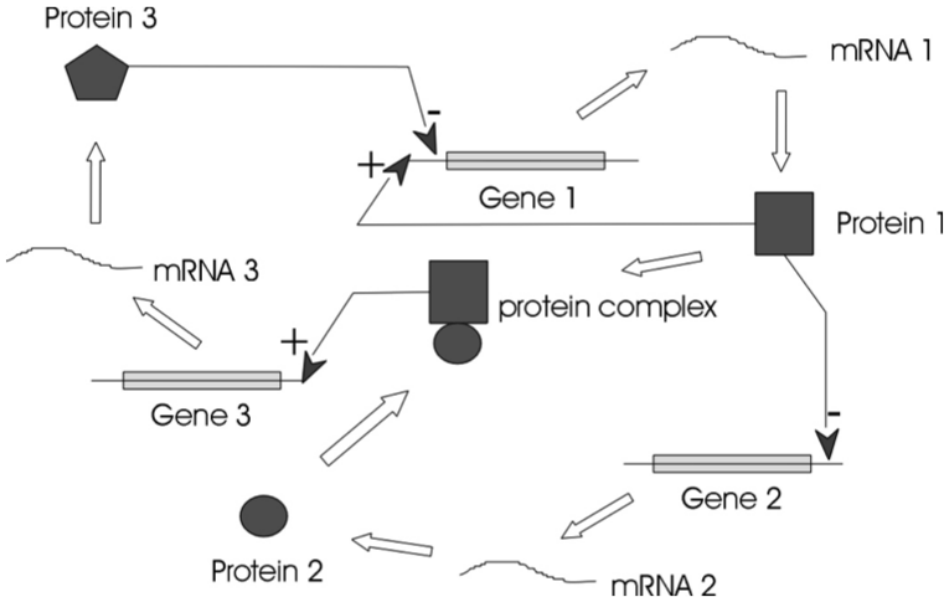
Man-built for the distribution of a commodity

- telephone networks, power grids, transportation networks, etc.

Biological networks

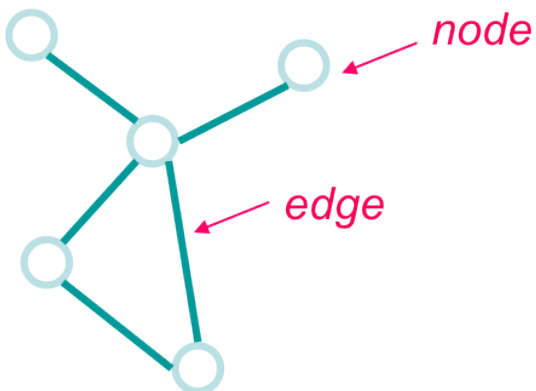
Represent biological systems

- protein-protein interaction networks, gene regulation networks, metabolic pathways, etc.



Representing networks

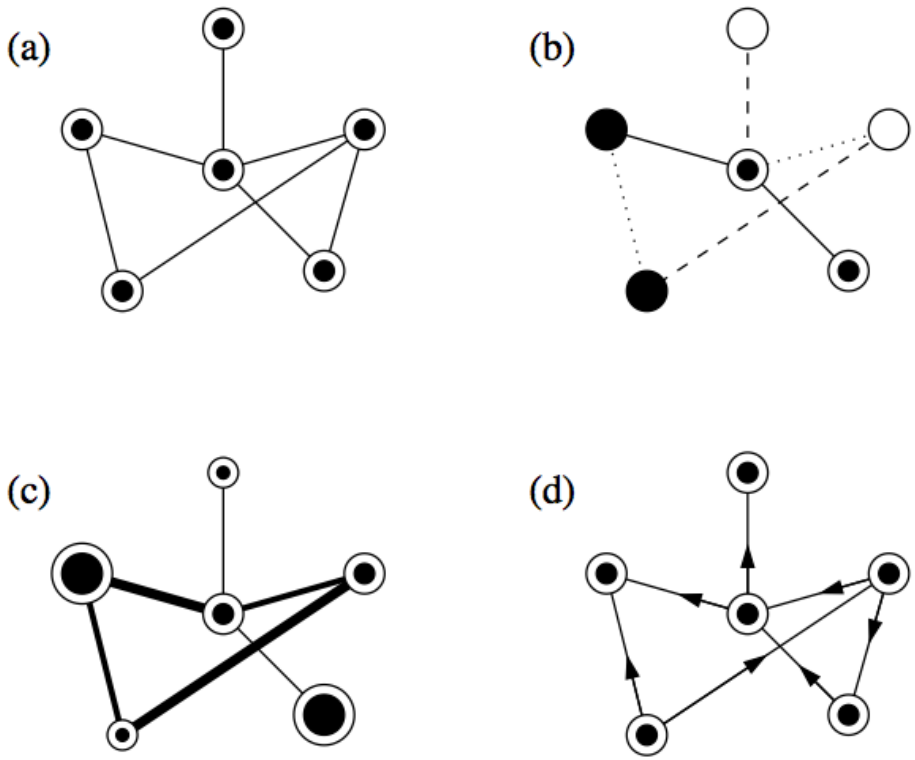
- Network $\equiv$ Graph
- Networks are just collections of “points” joined by “lines”



points	lines	
vertices	edges, arcs	math
nodes	links	computer science
sites	bonds	physics
actors	ties, relations	sociology

Types of networks

From [Newman 2003]



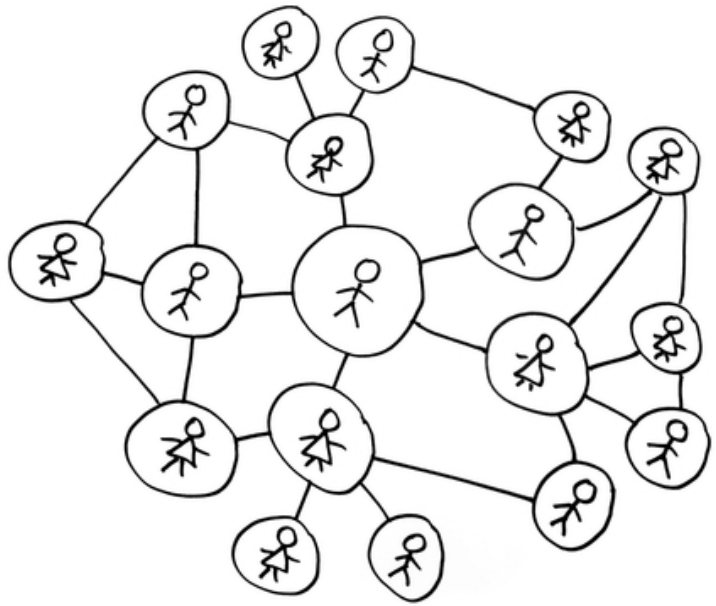
(a) unweighted, undirected

(b) discrete vertex and edge types, undirected

(c) varying vertex and edge weights, undirected

(d) directed

Three common properties



1. A friend of a friend is also frequently a friend
2. There are very short paths among most pairs of nodes

   "Only 6 hops separate any two people in the world"
3. Degree distribution follows a power law

1+2 is often called the small-world property.

Measuring the small-world phenomenon, I

- d_{ij} = length of the shortest path from i to j
- To discuss "every two people are 6 hops away" we use:
  - The diameter (max longest shortest-path distance) as

$$d = \max_{i,j} d_{ij}$$

  - The average shortest-path length as

$$l = \frac{2}{n\,(n+1)} \sum_{i>j} d_{ij}$$

  - The effective diameter as the d s.t. 95% of d_{ij} are $\leq d$

From [Newman 2003]

	network	type	n	m	z	ℓ	α	$C^{(1)}$	$C^{(2)}$	r	Ref(s).
social	film actors	undirected	449 913	25 516 482	113.43	3.48	2.3	0.20	0.78	0.208	20, 416
	company directors	undirected	7 673	55 392	14.44	4.60	–	0.59	0.88	0.276	105, 323
	math coauthorship	undirected	253 339	496 489	3.92	7.57	–	0.15	0.34	0.120	107, 182
	physics coauthorship	undirected	52 909	245 300	9.27	6.19	–	0.45	0.56	0.363	311, 313
	biology coauthorship	undirected	1 520 251	11 803 064	15.53	4.92	–	0.088	0.60	0.127	311, 313
	telephone call graph	undirected	47 000 000	80 000 000	3.16		2.1				8, 9
	email messages	directed	59 912	86 300	1.44	4.95	1.5/2.0		0.16		136
	email address books	directed	16 881	57 029	3.38	5.22	–	0.17	0.13	0.092	321
	student relationships	undirected	573	477	1.66	16.01	–	0.005	0.001	−0.029	45
	sexual contacts	undirected	2 810				3.2				265, 266
information	WWW `nd.edu`	directed	269 504	1 497 135	5.55	11.27	2.1/2.4	0.11	0.29	−0.067	14, 34
	WWW Altavista	directed	203 549 046	2 130 000 000	10.46	16.18	2.1/2.7				74
	citation network	directed	783 339	6 716 198	8.57		3.0/–				351
	Roget's Thesaurus	directed	1 022	5 103	4.99	4.87	–	0.13	0.15	0.157	244
	word co-occurrence	undirected	460 902	17 000 000	70.13		2.7		0.44		119, 157
technological	Internet	undirected	10 697	31 992	5.98	3.31	2.5	0.035	0.39	−0.189	86, 148
	power grid	undirected	4 941	6 594	2.67	18.99	–	0.10	0.080	−0.003	416
	train routes	undirected	587	19 603	66.79	2.16	–		0.69	−0.033	366
	software packages	directed	1 439	1 723	1.20	2.42	1.6/1.4	0.070	0.082	−0.016	318
	software classes	directed	1 377	2 213	1.61	1.51	–	0.033	0.012	−0.119	395
	electronic circuits	undirected	24 097	53 248	4.34	11.05	3.0	0.010	0.030	−0.154	155
	peer-to-peer network	undirected	880	1 296	1.47	4.28	2.1	0.012	0.011	−0.366	6, 354
biological	metabolic network	undirected	765	3 686	9.64	2.56	2.2	0.090	0.67	−0.240	214
	protein interactions	undirected	2 115	2 240	2.12	6.80	2.4	0.072	0.071	−0.156	212
	marine food web	directed	135	598	4.43	2.05	–	0.16	0.23	−0.263	204
	freshwater food web	directed	92	997	10.84	1.90	–	0.20	0.087	−0.326	272
	neural network	directed	307	2 359	7.68	3.97	–	0.18	0.28	−0.226	416, 421

z=avg degree; l=avg distance; α=exponent of degree powerlaw; C^1, C^2: clustering coefficients

Is this surprising?

Should we expect this in a random network?

It depends on what you mean by random network

The (basic) random graph model

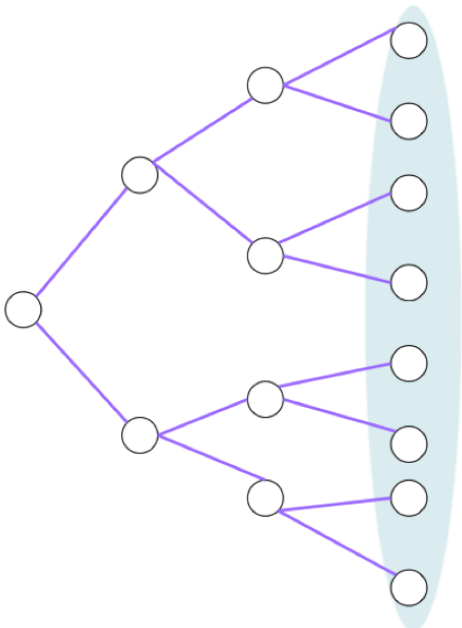
a.k.a. ER model

Basic $G_{n,p}$ Erdös-Rényi random graph model:

- parameter n is the number of vertices
- parameter p is s.t. $0 \leq p \leq 1$
- Generate and edge (i, j) independently at random with probability p

Measuring the diameter in ER networks

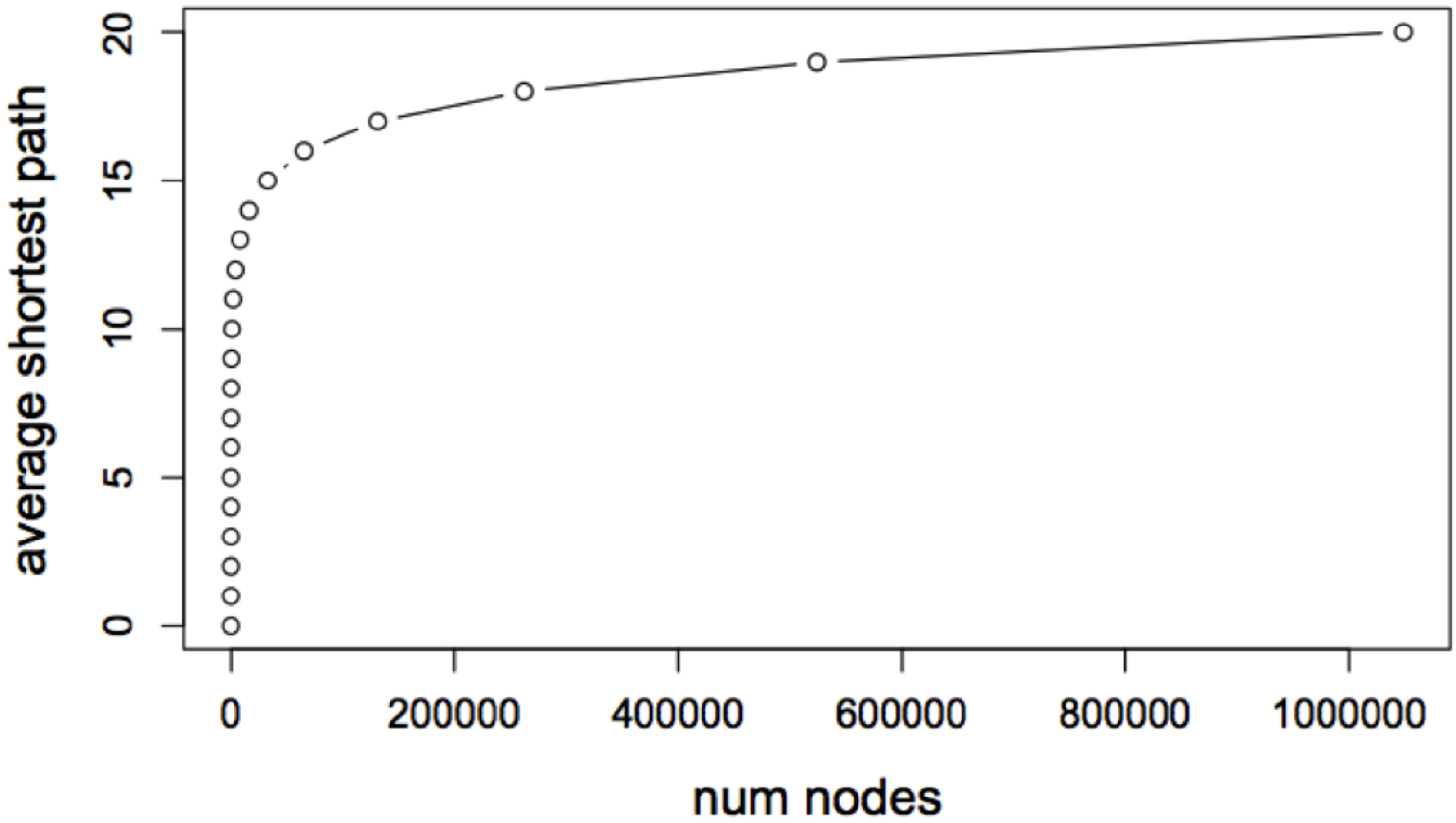
Want to show that the diameter in ER networks is small



- Let the average degree be z
- At distance l, can reach z^l nodes
- At distance $\frac{\log n}{\log z}$, reach all n nodes
- So, diameter is (roughly) $O(\log n)$

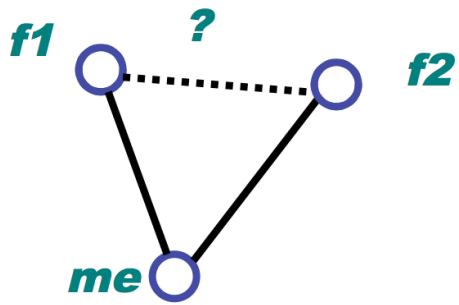
ER networks have small diameter

As shown by the following simulation



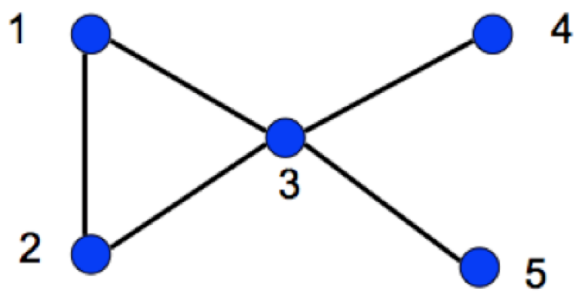
Measuring the small-world phenomenon, II

- To check whether “the friend of a friend is also frequently a friend”, we use:
  - The transitivity or clustering coefficient, which basically measures the probability that two of my friends are also friends



Global clustering coefficient

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



$$C = \frac{3 \times 1}{8} = 0.375$$

Local clustering coefficient

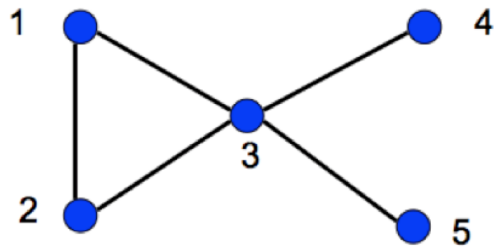
- For each vertex i, let n_i be the number of neighbors of i
- Let C_i be the fraction of pairs of neighbors that are connected within each other

$$C_i = \frac{\text{nr. of connections between } i\text{'s neighbors}}{\frac{1}{2} n_i \left(n_i - 1\right)}$$

- Finally, average C_i over all nodes i in the network

$$C = \frac{1}{n} \sum_i C_i$$

Local clustering coefficient example



- $C_1 = C_2 = 1/1$
- $C_3 = 1/6$
- $C_4 = C_5 = 0$
- $C = \frac{1}{5}(1 + 1 + 1/6) = 13/30 = 0.433$

From [Newman 2003]

	network	type	n	m	z	ℓ	α	$C^{(1)}$	$C^{(2)}$	r	Ref(s).
social	film actors	undirected	449 913	25 516 482	113.43	3.48	2.3	0.20	0.78	0.208	20, 416
	company directors	undirected	7 673	55 392	14.44	4.60	–	0.59	0.88	0.276	105, 323
	math coauthorship	undirected	253 339	496 489	3.92	7.57	–	0.15	0.34	0.120	107, 182
	physics coauthorship	undirected	52 909	245 300	9.27	6.19	–	0.45	0.56	0.363	311, 313
	biology coauthorship	undirected	1 520 251	11 803 064	15.53	4.92	–	0.088	0.60	0.127	311, 313
	telephone call graph	undirected	47 000 000	80 000 000	3.16		2.1				8, 9
	email messages	directed	59 912	86 300	1.44	4.95	1.5/2.0		0.16		136
	email address books	directed	16 881	57 029	3.38	5.22	–	0.17	0.13	0.092	321
	student relationships	undirected	573	477	1.66	16.01	–	0.005	0.001	−0.029	45
	sexual contacts	undirected	2 810				3.2				265, 266
information	WWW `nd.edu`	directed	269 504	1 497 135	5.55	11.27	2.1/2.4	0.11	0.29	−0.067	14, 34
	WWW Altavista	directed	203 549 046	2 130 000 000	10.46	16.18	2.1/2.7				74
	citation network	directed	783 339	6 716 198	8.57		3.0/–				351
	Roget's Thesaurus	directed	1 022	5 103	4.99	4.87	–	0.13	0.15	0.157	244
	word co-occurrence	undirected	460 902	17 000 000	70.13		2.7		0.44		119, 157
technological	Internet	undirected	10 697	31 992	5.98	3.31	2.5	0.035	0.39	−0.189	86, 148
	power grid	undirected	4 941	6 594	2.67	18.99	–	0.10	0.080	−0.003	416
	train routes	undirected	587	19 603	66.79	2.16	–		0.69	−0.033	366
	software packages	directed	1 439	1 723	1.20	2.42	1.6/1.4	0.070	0.082	−0.016	318
	software classes	directed	1 377	2 213	1.61	1.51	–	0.033	0.012	−0.119	395
	electronic circuits	undirected	24 097	53 248	4.34	11.05	3.0	0.010	0.030	−0.154	155
	peer-to-peer network	undirected	880	1 296	1.47	4.28	2.1	0.012	0.011	−0.366	6, 354
biological	metabolic network	undirected	765	3 686	9.64	2.56	2.2	0.090	0.67	−0.240	214
	protein interactions	undirected	2 115	2 240	2.12	6.80	2.4	0.072	0.071	−0.156	212
	marine food web	directed	135	598	4.43	2.05	–	0.16	0.23	−0.263	204
	freshwater food web	directed	92	997	10.84	1.90	–	0.20	0.087	−0.326	272
	neural network	directed	307	2 359	7.68	3.97	–	0.18	0.28	−0.226	416, 421

z=avg degree; l=avg distance; α=exponent of degree powerlaw; C^1, C^2: clustering coefficients

ER networks do not show transitivity

- In ER networks, $C = p$, since each edge is added independently
- in many real networks, $C \gg p$
- where p is estimated as $|E|/(n(n-1)/2)$

ER networks do not show transitivity

Table 1: Clustering coefficients, C, for a number of different networks; n is the number of node, z is the mean degree. Taken from [146].

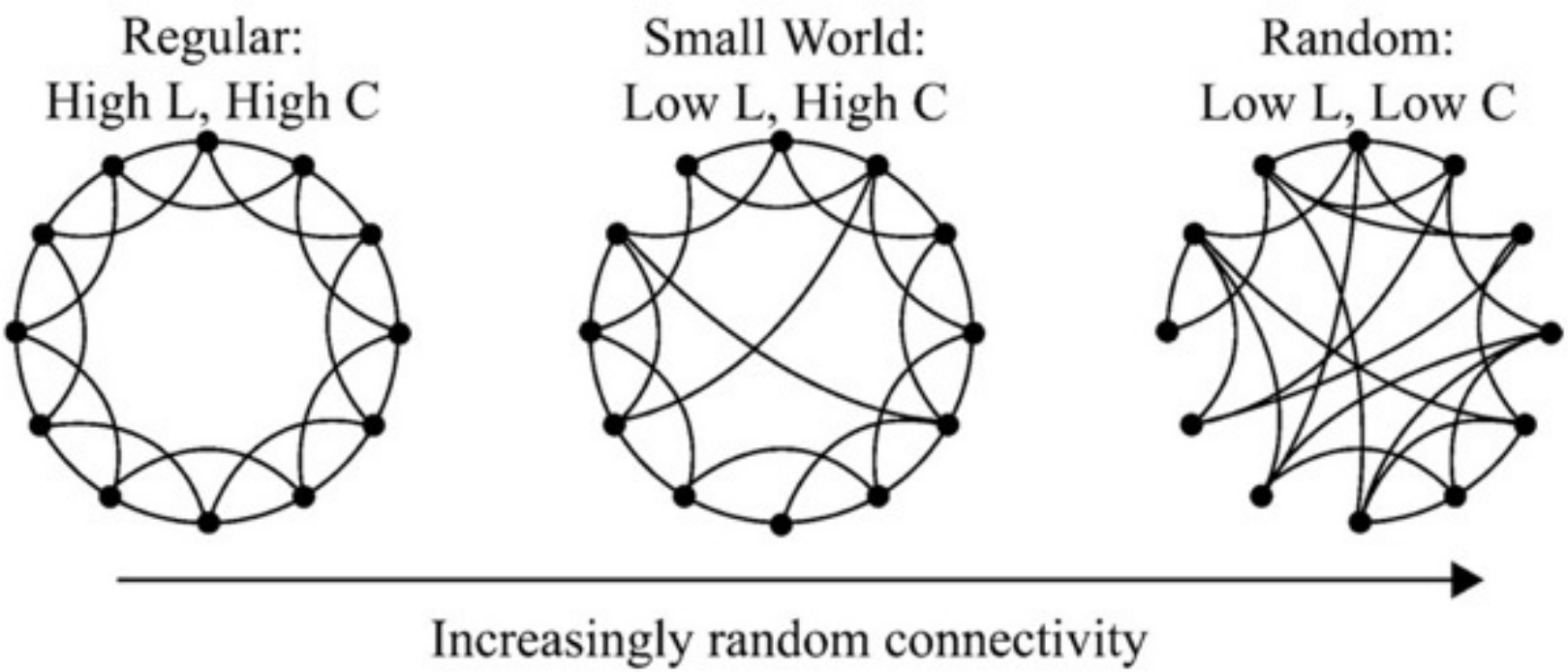
Network	n	z	C measured	C for random graph
Internet [153]	6,374	3.8	0.24	0.00060
World Wide Web (sites) [2]	153,127	35.2	0.11	0.00023
power grid [192]	4,941	2.7	0.080	0.00054
biology collaborations [140]	1,520,251	15.5	0.081	0.000010
mathematics collaborations [141]	253,339	3.9	0.15	0.000015
film actor collaborations [149]	449,913	113.4	0.20	0.00025
company directors [149]	7,673	14.4	0.59	0.0019
word co-occurrence [90]	460,902	70.1	0.44	0.00015
neural network [192]	282	14.0	0.28	0.049
metabolic network [69]	315	28.3	0.59	0.090
food web [138]	134	8.7	0.22	0.065

So ER networks do not have high clustering, but..

- Other “random network” models generate graphs with low diameter and high clustering coefficient
- The Watts-Strogatz model is an example

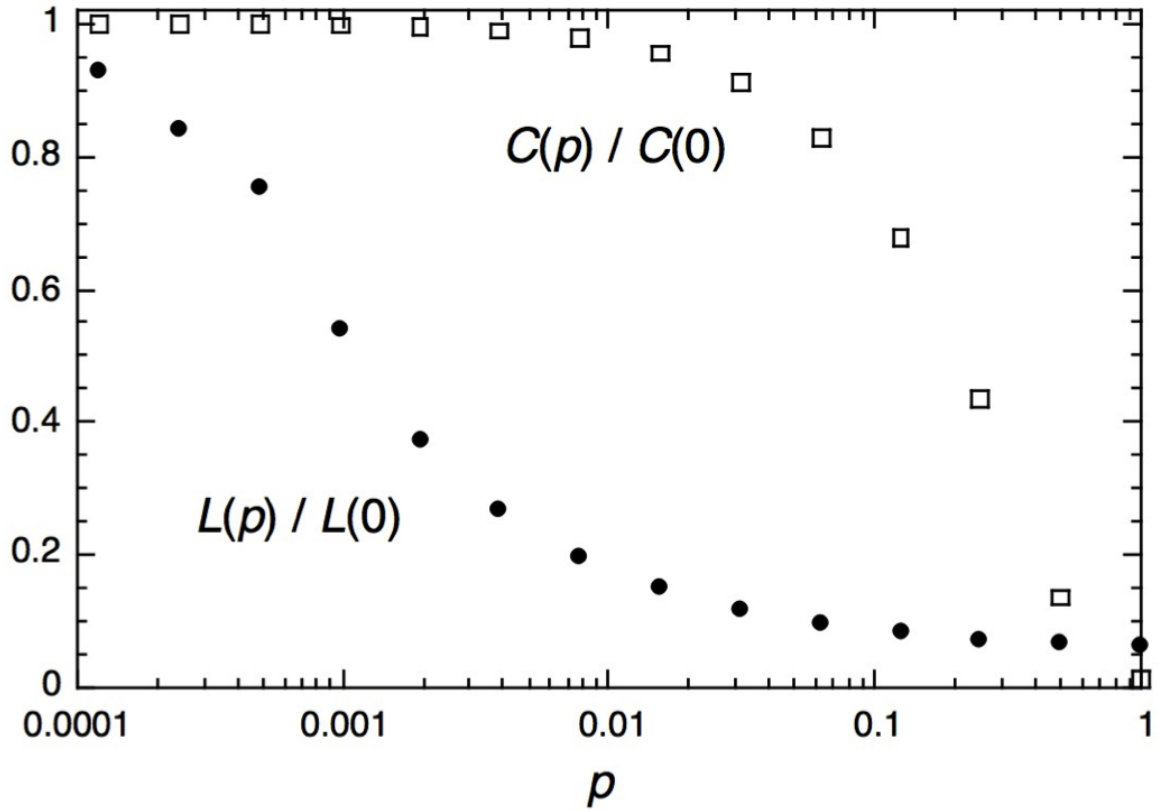
The Watts-Strogatz model

- Start with all n vertices arranged on a ring
- Each vertex has initially 4 connections to their closest nodes
- With probability p, rewire each local connection to a random vertex



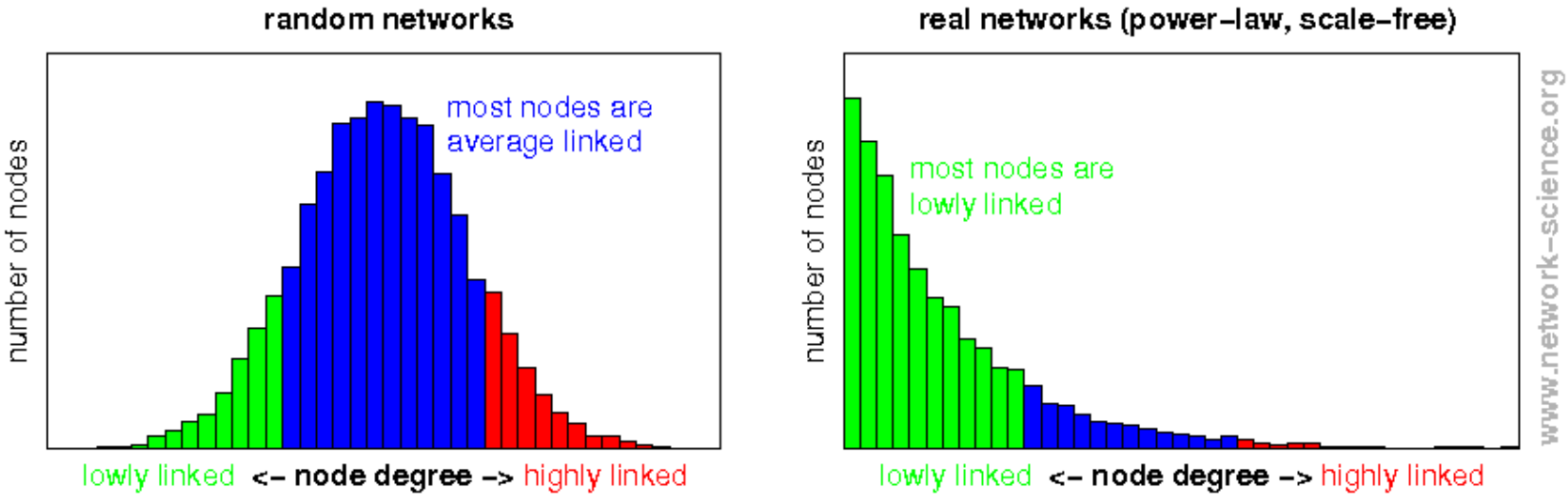
The Watts-Strogatz model

For an appropriate value of $p \approx 0.01$ (1%), the model achieves high clustering and small diameter



Degree distribution

Histogram of nr of nodes having a particular degree



$f_k =$ fraction of nodes of degree k

Degree distribution

The degree distribution of most real-world networks follows a power-law distribution

$$f_k = ck^{-\alpha}$$

- "heavy-tail" distribution, implies existence of hubs
- hubs are nodes with very high degree

Scale-free or scale-invariant

Networks with power-law degree distribution are often called scale-free or scale-invariant.

- D is scale-invariant if $D(\lambda x) = f(\lambda)D(x)$
- True for powerlaw degree distribution (x = #links)
- For non-powerlaws, the $f(\lambda)$ instead depends on x

- This means no characteristic scale or “units of measure”

For “growing” networks, it implies that the statistics remain similar as the network grows - fractality etc.

ER Random networks are not scale-free!

For ER random networks, the degree distribution follows the binomial distribution (or Poisson if n is large)

$$f_k = \binom{n}{k} p^k (1-p)^{(n-k)} \approx \frac{z^k e^{-z}}{k!}$$

- Where $z = p(n-1)$ is the mean degree
- Probability of nodes with very large degree becomes exponentially small
- Maximum degree is $pn + O(\sqrt{(pn)})$ with high probability
- so no hubs

So ER networks are not scale-free, but. . .

- One can build models of “random graph” that do
- Barabasi-Albert “preferential attachment”

Preferential attachment

- “Rich get richer” dynamics
  - The more someone has, the more she is likely to have
- Examples
  - the more friends you have, the easier it is to make new ones
  - the more business a firm has, the easier it is to win more
  - the more people there are at a restaurant, the more who want to go

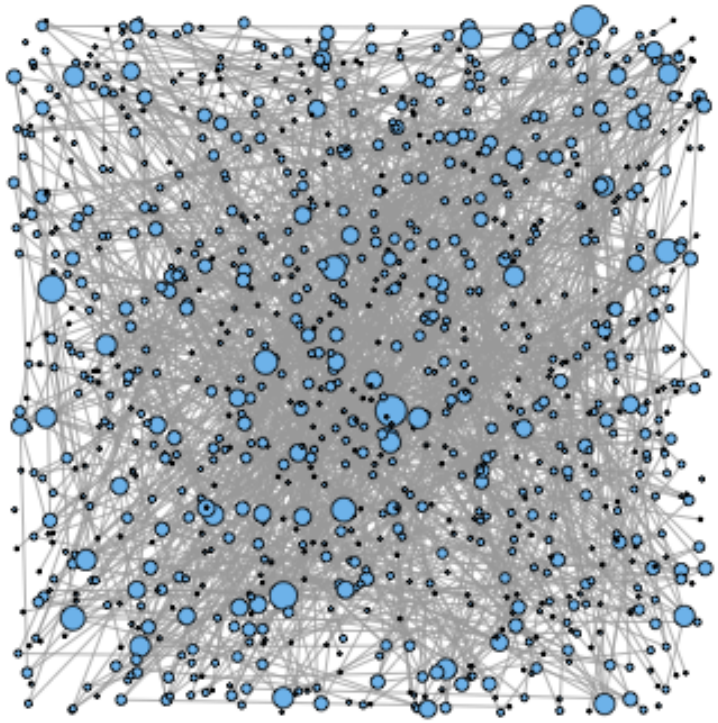
Barabási-Albert model

From [Barabasi 1999]

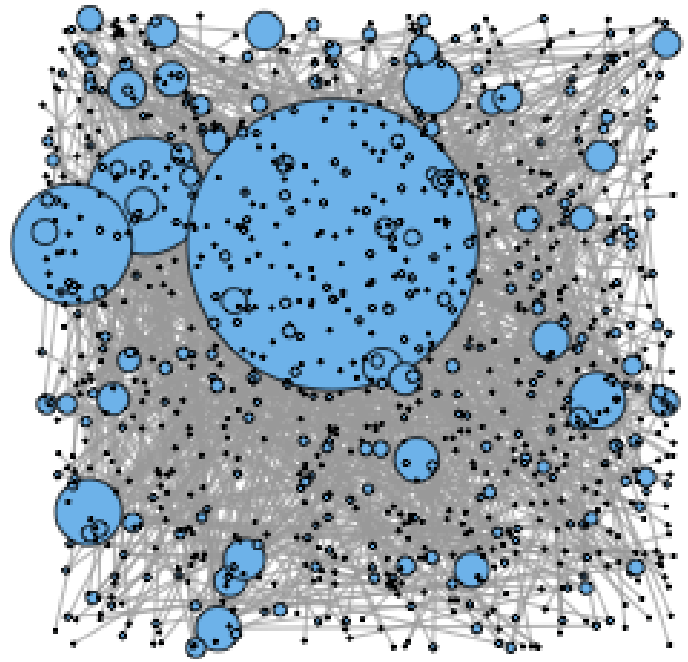
- "Growth" model
  - The model controls how a network grows over time
- Uses preferential attachment as a guide to grow the network
  - new nodes prefer to attach to well-connected nodes
- (Simplified) process:
  - the process starts with some initial subgraph
  - each new node comes in with m edges
  - probability of connecting to existing node i is proportional to i's degree
  - results in a power-law degree distribution with exponent $\alpha = 3$

ER vs. BA

Experiment with 1000 nodes, 999 edges ($m_0 = 1$ in BA model).



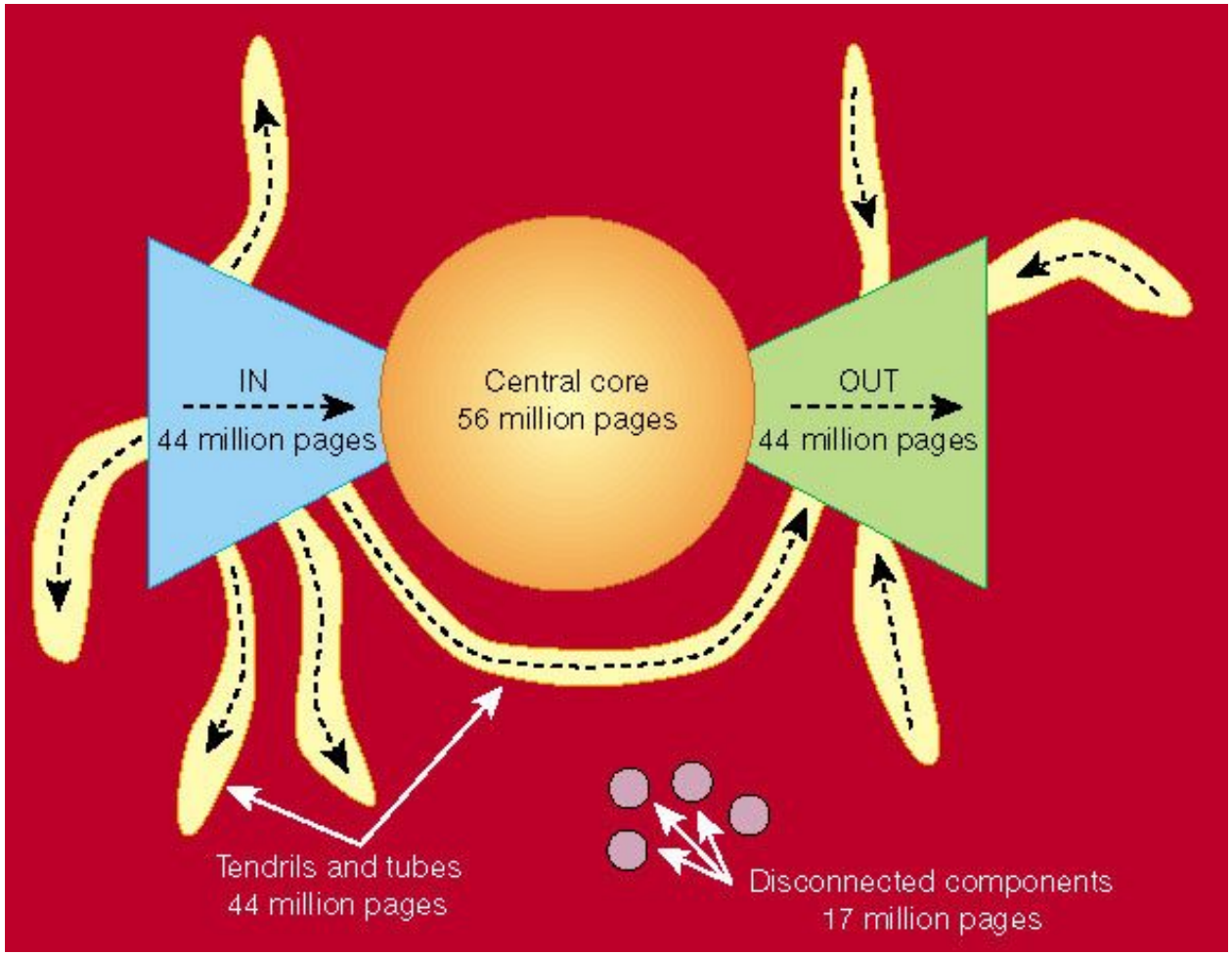
random



preferential attachment

The Web

. . . is different. “Bowtie” structure



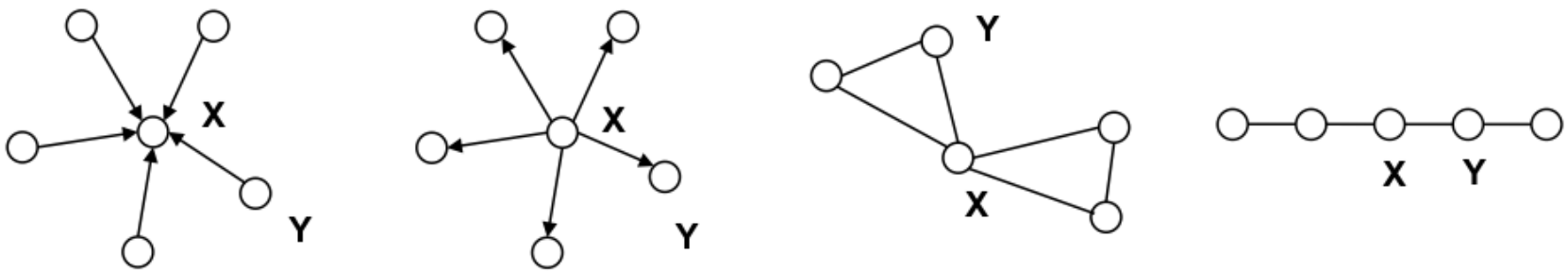
[The web is a bow tie. Nature 405, 113 (2000) doi:10.1038/35012155]
`https://en.wikipedia.org/wiki/Topology_of_the_World_Wide_Web`

`http://cs.wellesley.edu/~pmetaxas/Why_Is_the_Shape_of_the_Web_a_Bowtie.pdf`

Centrality in Networks

Centrality is a node's measure w.r.t. others

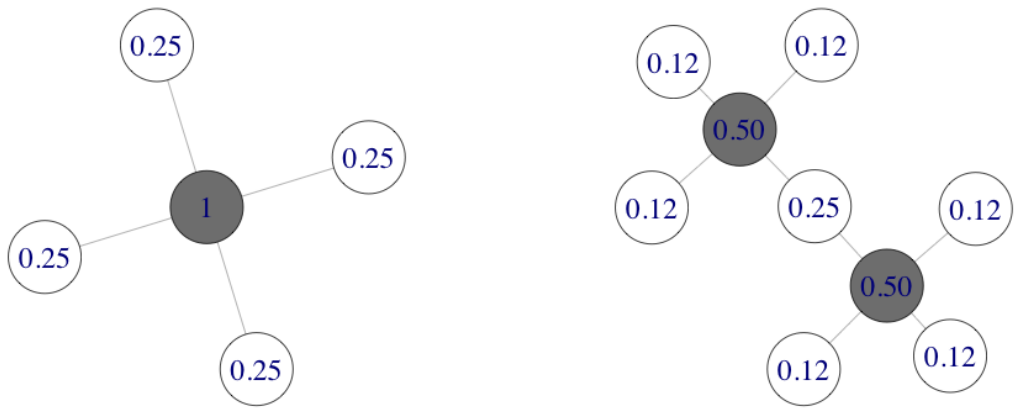
- A central node is *important* and/or *powerful*
- A central node has an *influential position in the network*
- A central node has an *advantageous position in the network*



Degree centrality

Power through connections

First approximation: Centrality $\simeq$ number of connections
Normalize by maximum possible number of connections to put it in [0,1]

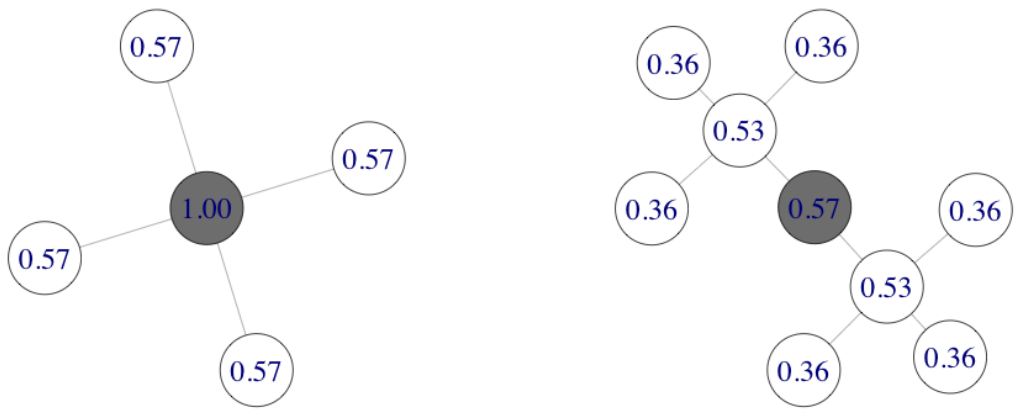


But look at these examples, does degree centrality look OK to you?

Closeness centrality

Power through proximity to others

$$closeness\_centrality(i) \stackrel{def}{=} \left(\frac{\sum_{j \neq i} d(i,j)}{n-1} \right)^{-1} = \frac{n-1}{\sum_{j \neq i} d(i,j)}$$



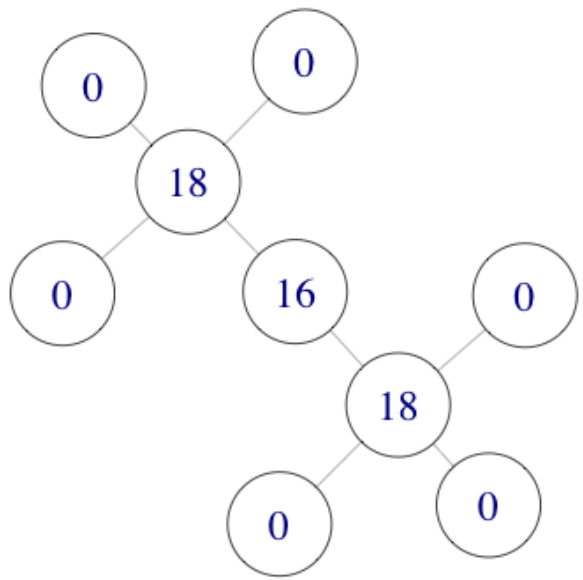
Here, what matters is to be close to everybody else, i.e., to be easily reachable or have the power to quickly reach others.

Betweenness centrality

Power through brokerage

A node is important if it lies in many shortest-paths

- so it is essential in passing information through the network



Betweenness centrality

Power through brokerage

$$betweenness\_centrality(i) \stackrel{def}{=} \sum_{j<k} \frac{g_{jk}(i)}{g_{jk}}$$

Where

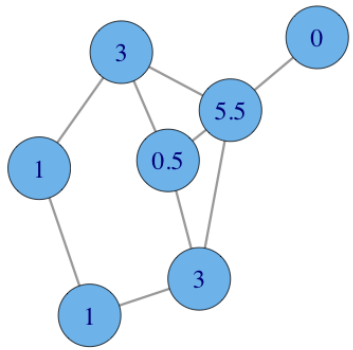
- g_{jk} is the number of shortest-paths between j and k, and
- $g_{jk}(i)$ is the number of shortest-paths through i

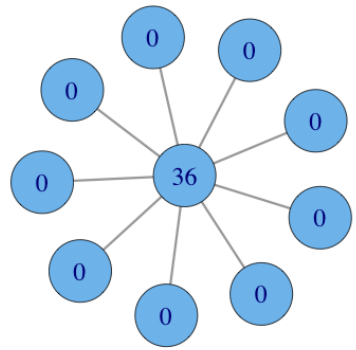
Oftentimes it is normalized:

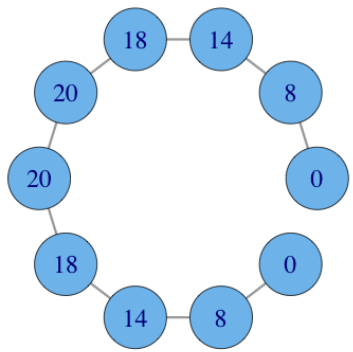
$$norm\_betweenness\_centrality(i) \stackrel{def}{=} \frac{betweenness\_centrality(i)}{\binom{n-1}{2}}$$

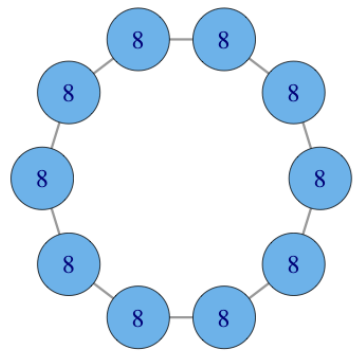
Betweenness centrality

Examples (non-normalized)

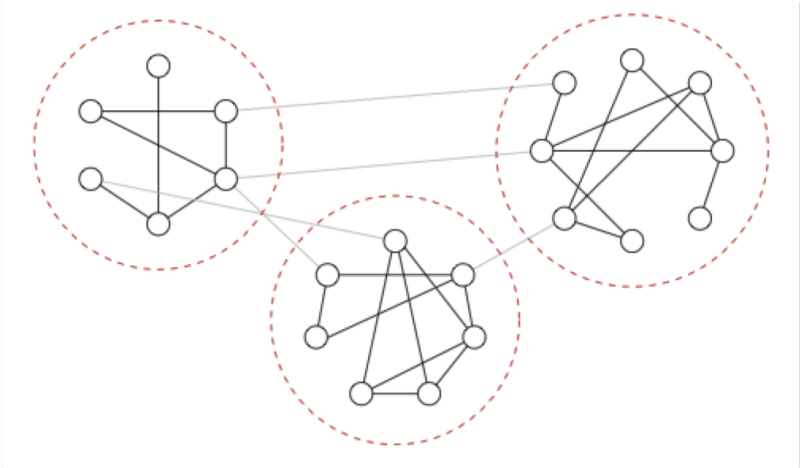


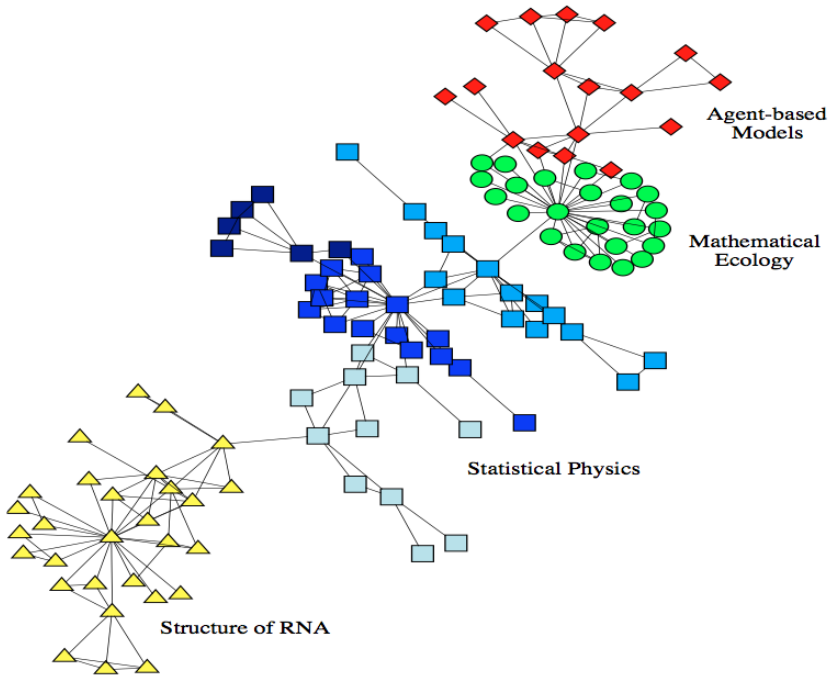






Communities





What are communities?

A community is *dense* in the inside but *sparse* w.r.t. the outside

No universal definition! But some ideas are:

- A community should be *densely connected*
- A community should be *well-separated* from the rest of the network
- Members of a community should be *more similar* among themselves than with the rest

Most common

nr. of intra-cluster edges $>$ nr. of inter-cluster edges

Some definitions

Let $G = (V, E)$ be a network with $|V| = n$ nodes and $|E| = m$ edges. Let C be a subset of nodes in the network (a "cluster" or "community") of size $|C| = n_c$. Then

- *intra-cluster density*:

$$\delta_{int}(C) = \frac{\text{nr. internal edges of } C}{n_c(n_c - 1)/2}$$

- *inter-cluster density*:

$$\delta_{ext}(C) = \frac{\text{nr. inter-cluster edges of } C}{n_c(n - n_c)}$$

A community should have $\delta_{int}(C) > \delta(G)$, where $\delta(G)$ is the average edge density of the whole graph G, i.e.

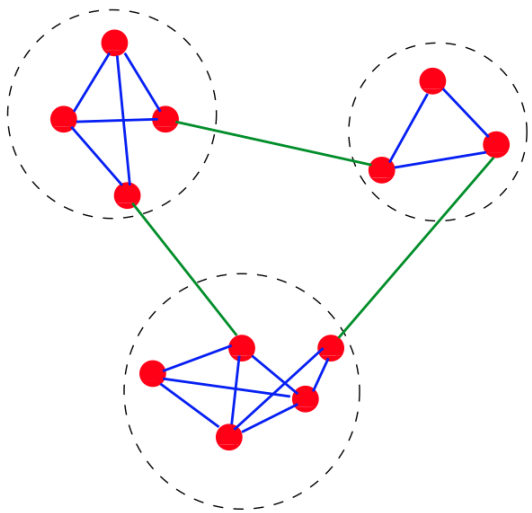
$$\delta(G) = \frac{\text{nr. edges in } G}{n(n-1)/2}$$

Most algorithms search for tradeoffs between *large* $\delta_{int}(C)$ and *small* $\delta_{ext}(C)$

- e.g. optimizing $\sum_C \delta_{int}(C) - \delta_{ext}(C)$ over all communities C

Define further:

- $m_c =$ nr. edges within cluster $C = |\{(u, v)|u, v \in C\}|$
- $f_c =$ nr. edges in the frontier of $C = |\{(u, v)|u \in C, v \notin C\}|$



- $n_{c_1} = 4, m_{c_1} = 5, f_{c_1} = 2$
- $n_{c_2} = 3, m_{c_2} = 3, f_{c_2} = 2$
- $n_{c_3} = 5, m_{c_3} = 8, f_{c_3} = 2$

Community quality criteria

- **conductance**: fraction of edges leaving the cluster $\frac{f_c}{2m_c+f_c}$
- **expansion**: nr of edges per node leaving the cluster $\frac{f_c}{n_c}$
- **internal density**: a.k.a. "intra-cluster density" $\frac{m_c}{n_c(n_c-1)/2}$
- **cut ratio**: a.k.a. "inter-cluster density" $\frac{f_c}{n_c(n-n_c)}$
- **modularity**: difference between nr. of edges in C and the expected nr. of edges $E[m_c]$ of a random graph (notion of "random graph" is a parameter)

$$\frac{1}{4m}(m_c - E[m_c])$$

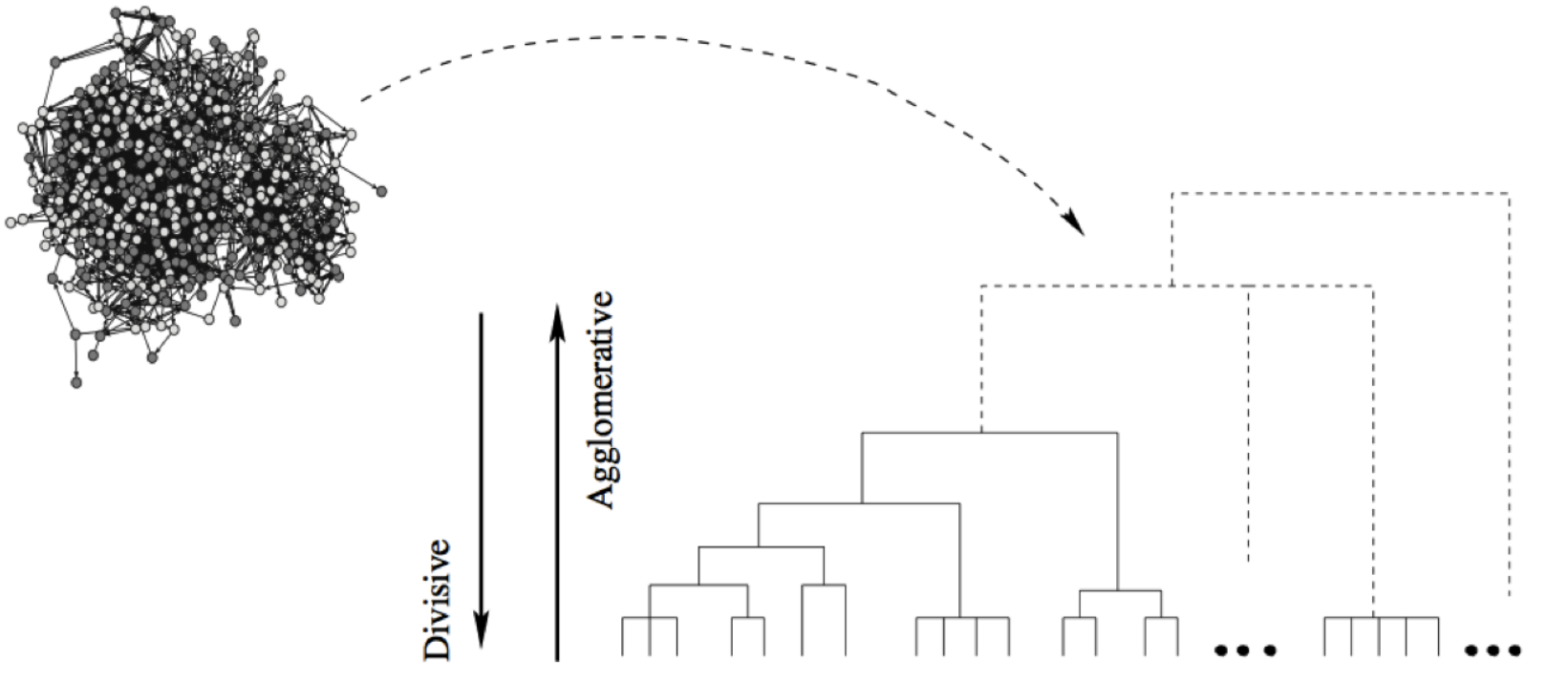
(often used: random graph with the same degree distribution)

Methods we will cover

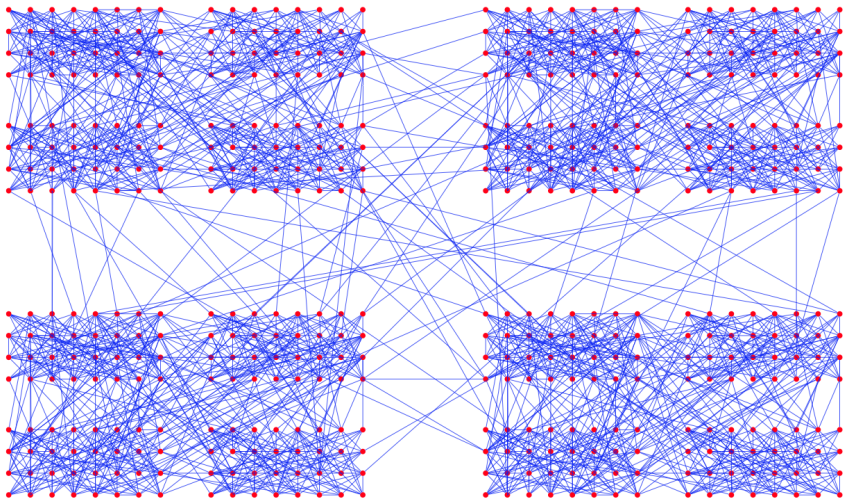
- Hierarchical clustering
  - Agglomerative
  - Divisive (Girvan-Newman algorithm)
- Modularity maximization algorithms
  - Louvain method

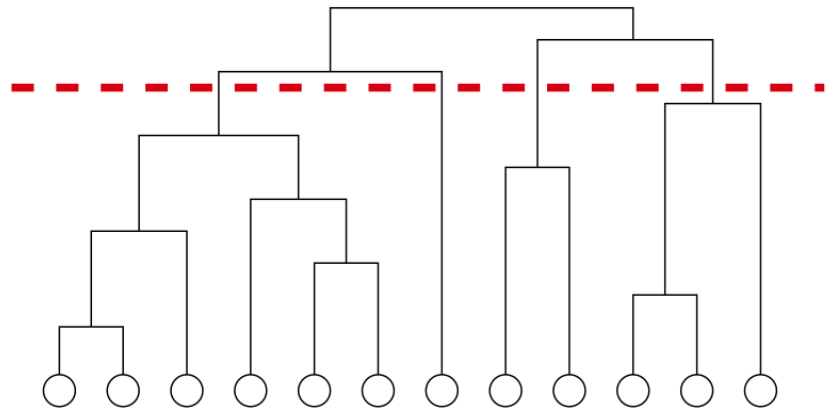
Hierarchical clustering

From hairball to *dendogram*



Suitable if input network has hierarchical structure





Agglomerative hierarchical clustering [Newman 2010]

Ingredients

- Similarity measure between nodes
- Similarity measure between *sets of nodes*

Pseudocode

1. Assign each node to its own cluster
2. Find the cluster pair with highest similarity and join them together into a cluster
3. Compute new similarities between new joined cluster and others
4. Go to step 2 until all nodes form a single cluster

Similarity measures w_{ij} for nodes

Let $\mathbf{A}$ be the adjacency matrix of the network, i.e. $A_{ij} = 1$ if $(i, j) \in E$ and 0 otherwise.

- **Jaccard index:**

$$w_{ij} = \frac{|\Gamma(i) \cap \Gamma(j)|}{|\Gamma(i) \cup \Gamma(j)|}$$

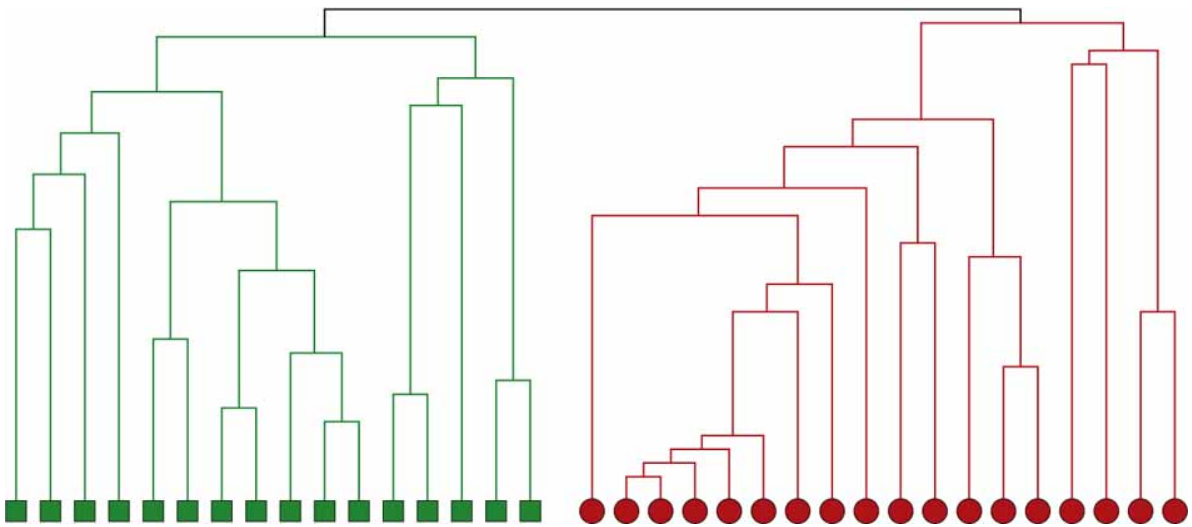
  where $\Gamma(i)$ is the set of neighbors of node i
- Cosine similarity, Hamming distance, Pearson correlation...

Similarity measures for sets of nodes

- Single linkage: $s_{XY} = \max_{x \in X, y \in Y} s_{xy}$
- Complete linkage: $s_{XY} = \min_{x \in X, y \in Y} s_{xy}$
- Average linkage: $s_{XY} = \frac{\sum_{x \in X, y \in Y} s_{xy}}{|X| \times |Y|}$

Agglomerative hierarchical clustering on Zachary's network

Using average linkage



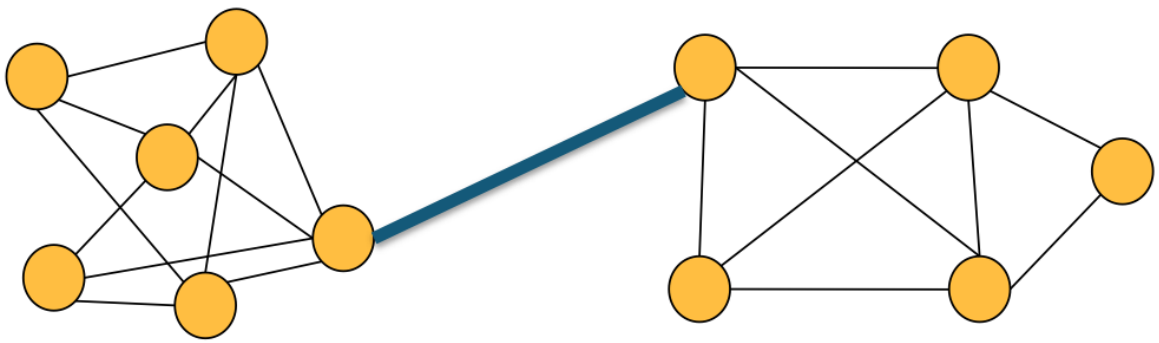
The Girvan-Newman algorithm

A *divisive* hierarchical algorithm [Girvan 2002]

Edge betweenness

The betweenness of an edge is the nr. of shortest-paths in the network that pass through that edge

It uses the idea that “bridges” between communities must have high edge betweenness

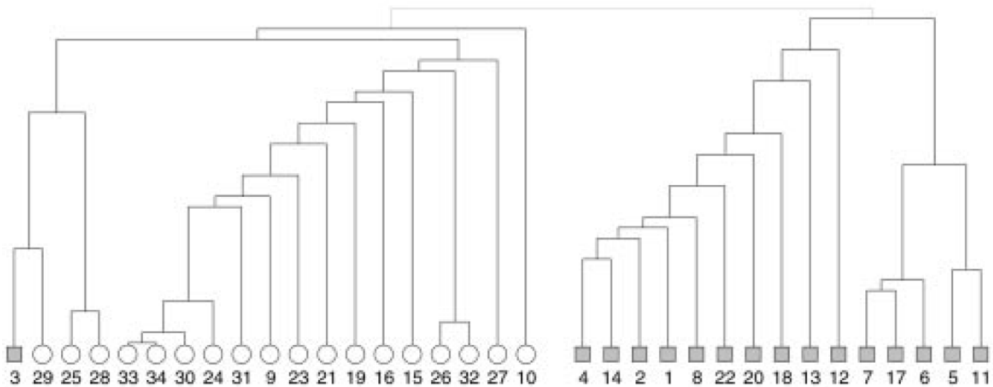


The Girvan-Newman algorithm

Pseudocode

1. Compute betweenness for all edges in the network
2. Remove the edge with highest betweenness
3. Go to step 1 until no edges left

Result is a dendogram



Definition of modularity [Newman 2010]

Random graphs are not expected to have community structure, so we will use them as null models.
The modularity of a decomposition in communities $c_1, \dots c_k$:

$$Q(\{c_1, \dots, c_k\}) = \frac{1}{2m}(m_i - E[m_i])$$

where m_i is number of edges in c_i, $E[m_i]$ the number of edges in c_i in the null model, and $m = \sum_i m_i = |E|$.

For example, for ER null model, $E[m_i] = p|c_i|(|c_i| - 1)/2$.
"Configuration model": random graph with same degree distribution
(`https://en.wikipedia.org/wiki/Configuration_model`)

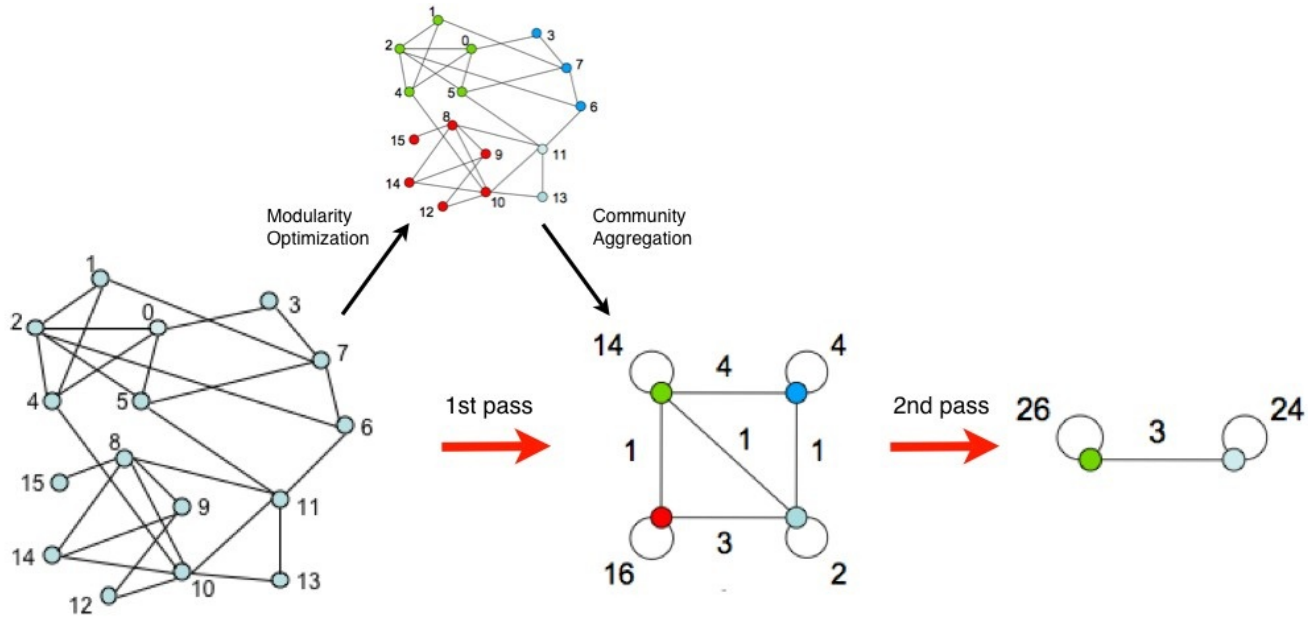
$$Q = \frac{1}{2m} \sum_{i \neq j} \left(A_{ij} - \frac{deg(i) \cdot deg(j)}{2m} \right) \delta(c(i), c(j))$$

($\delta(c(i), c(j)) = 1$ if i and j in same community, 0 otherwise)

The Louvain method [Blondel 2008]

Considered state-of-the-art

A heuristic to maximize modularity efficiently.



Pseudocode

1. Repeat until local optimum reached
   - 1.1 Phase 1: partition network greedily using modularity
   - 1.2 Phase 2: agglomerate found clusters into new nodes

The Louvain method

Phase 1: optimizing modularity

Pseudocode for phase 1

1. Assign a different community to each node
2. For each node i
   - For each neighbor j of i, consider removing i from its community and placing it to j's community
   - Greedily chose to place i into community of neighbor that leads to highest modularity gain
3. Repeat until no improvement can be done

Notes: One loop over i in arbitrary order - greedy!

The Louvain method

Phase 2: agglomerating clusters to form new network

Pseudocode for phase 2

1. Let each community C_i form a new node i
2. Let the edges between new nodes i and j be the sum of edges between nodes in C_i and C_j in the previous graph (notice there are self-loops)

The Louvain method

Observations

- Most of the time spent in first greedy phase, node level (95%?).
- For graphs with community structure, time empirically $O(n \log(\text{avg degree})) = O(n \log n)$.
- The output is also a hierarchy.
- Define a weighted version of modularity, then Louvain still applies.

Other stuff in community finding

- Overlapping communities
- Spectral decomposition (SVD...) approaches
- Link prediction (should (i, j) be an edge?)

Spreading in networks

Just a taster. . .

`http://web.stanford.edu/class/cs224w`

Jiahao Chen, Epidemics on Small-World Networks, 2005

. . .

Spreading in networks

Cascading behaviors:

- biological epidemics
- diffusion of innovation
- technology failures
- rumors, news
- recommendations
- viral marketing

Spreading in networks

Two distinct phenomena:

- Nodes make "intelligent" decisions based on expected benefits. There follow strategies
- Epidemic spreading: No decision making. More random.

Decision making

A node observes decisions of its neighbors and makes its own decision

Example: Adopting technology. I win by adopting same technology as my friends.

- "If more than >50% of my friends adopt Whatsapp vs. Skype, I adopt Whatsapp"
- "If more than >70% of my friends join the revolution, I join the revolution... with probability 30%"

Question: Who starts the cascades? Usually a little core of influencers
Uncertainty: Causality. Not knowing the reasons that move people.

Epidemics

When there is no decision-making, no strategy.

Simple contagion model:

- An infected person infects every person s/he meets with probability q, independently
- Initially: an infected person enters an otherwise healthy population.

Questions: How far does the epidemic spread? Does it become a pandemic? Does it die out?

Simple contagion model of epidemics

Let's analyze in an Erdös/Rényi-like model of interaction.

Each person meets $d = pn$ other random people.

Simple contagion model of epidemics

d-ary tree
$p_\ell =$ Probability that at least one node at depth ℓ is infected

If $\lim_{\ell \to \infty} p_\ell = 0$, the epidemics dies out
If $\lim_{\ell \to \infty} p_\ell = 1$, the epidemics remains

Also interesting: Fraction of nodes infected at depth ℓ

Simple contagion model of epidemics

$p_\ell =$ Probability that at least one node at depth ℓ is infected

- $p_0 = 1$ by hypothesis
- No infected node at level 1 with probability $(1-q)^d$
- $p_1 = 1 - (1-q)^d$
- No infected node at level ℓ with probability $(1 - q \cdot p_{\ell-1})^d$
- $p_\ell = 1 - (1 - qp_{\ell-1})^d$

$p_\ell =$ Probability that at least one node at depth ℓ is infected
$p_\ell = 1 - (1 - q \cdot p_{\ell-1})^d$

Claim:

$$\lim_{\ell \to \infty} p_\ell = \begin{cases} 0 & \text{if } q \cdot d < 1 \\ 1 & \text{if } q \cdot d \geq 1 \end{cases}$$

Hint of proof: study fixpoint $p = 1 - (1 - q \cdot p)^d$, etc.

Contagion or reproduction number

Contagion number or reproduction number $R_0 = qd$

In random networks, fate depends exclusively on R_0.

- If $R_0 < 1$, infection dies out
- If $R_0 > 1$, epidemic never dies and spreads exponentially

Epidemics prevention: Reduce R_0

- Reduce d: Isolate infected people
- Reduce q: Better hygienic measures

In small-world networks

Network topology modifies R_0. In particular, degree distribution.

- Super-spreaders: Nodes of large degree
- E.g. in STD, very promiscuous people
- E.g. hospitals, hotels, schools, planes...
- Target of prevention measures

https://en.wikipedia.org/wiki/Super-spreader
https://en.wikipedia.org/wiki/Timeline_of_the_SARS_outbreak
"The next pandemic, explained", Netflix.
Klovdahl, A. S. Social networks and the spread of infectious diseases: the AIDS example. Soc. Sci. Med. 21, 1985, 1203-1216.

In small-world networks

Besides dying-out and pandemics...

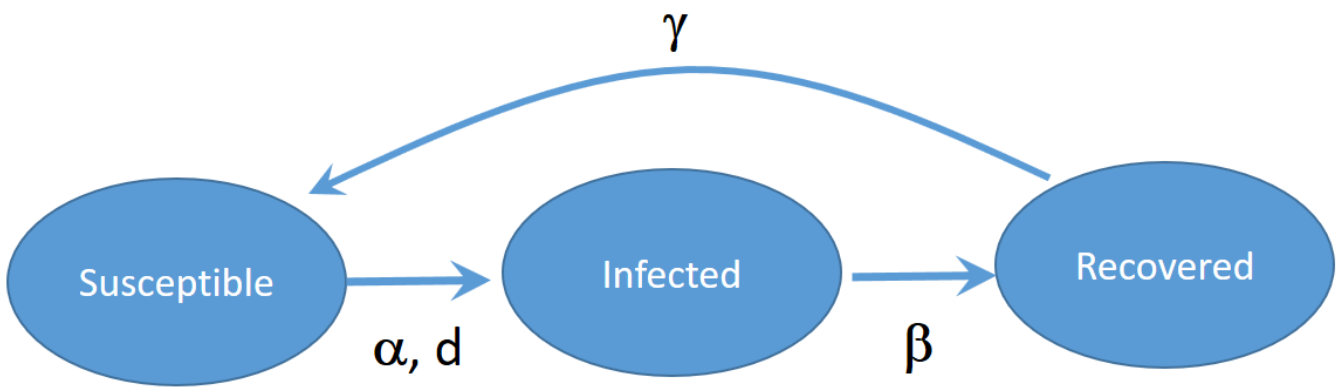
Third option:

- Clusters of localized infections that arise and disappear
- Infection does not die out, but there is no exponential infection
- Communities

Other models of infection

Approached by Markov chains or differential equations

- SIR model: People are Susceptible, Infected, or Recovered
- Susceptible people become infected
- Infected but surviving people may become immunized (forever, or temporarily)
- Extension: People die, people are born



Two algorithmic problems - combinatorial optimization

Given a network, info on spreading mechanisms, and budget k:

How to detect outbreaks soon

Choose a subset S of nodes, $|S| \leq k$, "watchers", that minimizes the time-to-detection of a large cascade

How to maximize spread

Choose a subset S of nodes, $|S| \leq k$, "influencers" or "seeds", so that cascades initiating from them spread maximally (become viral)