

CAIM: Cerca i Anàlisi d'Informació Massiva

FB, Grau en Enginyeria Informàtica

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

Fall 2018

<http://www.cs.upc.edu/~caim>

7. Introduction to Network Analysis

Network Analysis, Part I

Today's contents

1 / 72

Examples of real networks

1. Examples of real networks
2. What do real networks look like?
 - real networks exhibit small diameter
 - .. and so does the Erdős-Rényi or random model
 - real networks have high clustering coefficient
 - .. and so does the Watts-Strogatz model
 - real networks' degree distribution follows a power-law
 - .. and so does the Barabasi-Albert or preferential attachment model

3 / 72

4 / 72

Social networks

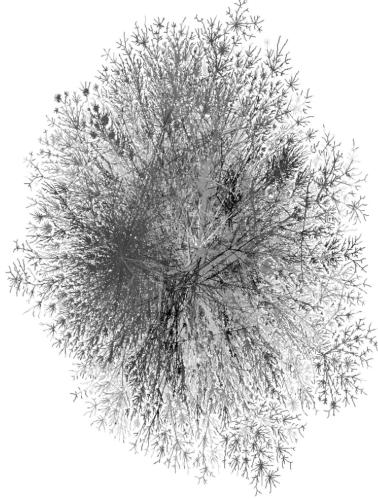
Links denote social “interactions”

- ▶ friendship, collaborations, e-mail, etc.



Information networks

- Nodes store information, links associate information
- ▶ citation networks, the web, p2p networks, etc.



5 / 72

Technological networks

Man-built for the distribution of a commodity

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

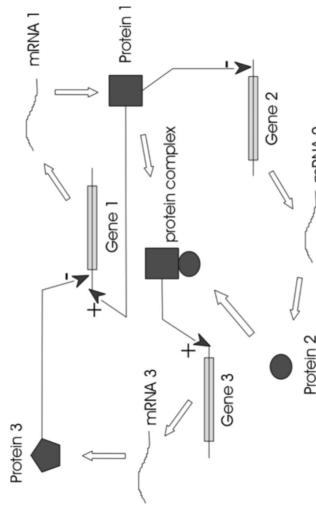


6 / 72

Biological networks

Represent biological systems

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



7 / 72

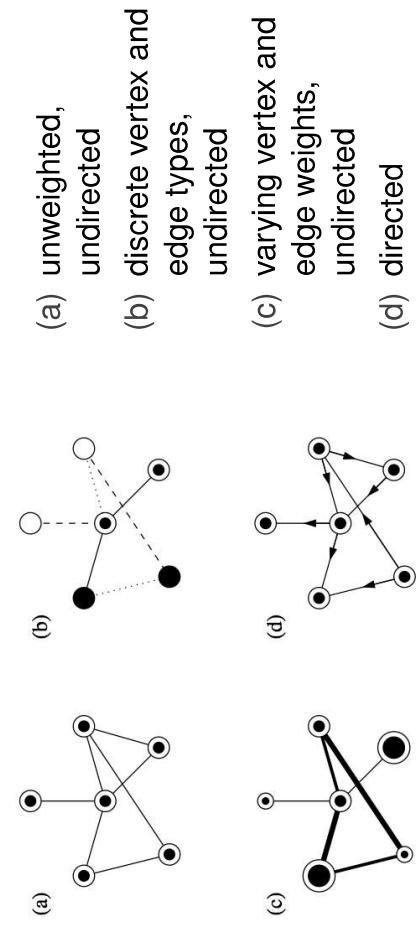
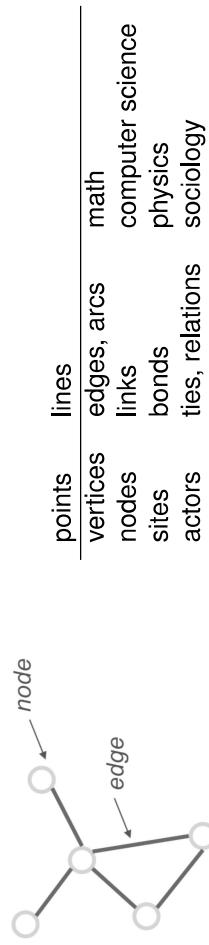
8 / 72

Representing networks

Types of networks

From [Newman, 2003]

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



Small-world phenomenon

9 / 72

- A friend of a friend is also frequently a friend
- Only 6 hops separate any two people in the world

Measuring the small-world phenomenon, I

- Let d_{ij} be the shortest-path distance between nodes i and j
- To check whether “any two nodes are within 6 hops”, we use:
 - The diameter (longest shortest-path distance) as

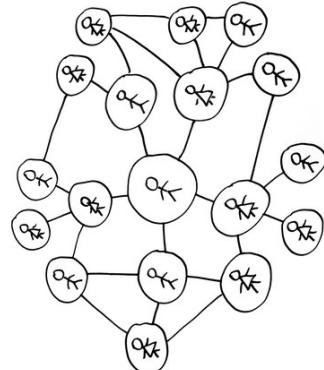
$$d = \max_{i,j} d_{ij}$$

- The average shortest-path length as

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

- The harmonic mean shortest-path length as

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



10 / 72

11 / 72

12 / 72

From [Newman, 2003]

But..

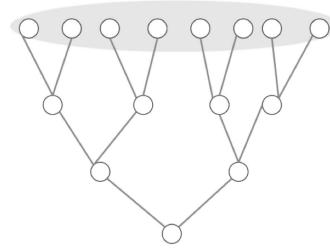
network	type	n	m	z	ℓ	α	$C^{(1)}$	$C^{(2)}$	r	Ref(s.)
film actors	undirected	449 913	25 516 482	113.43	3.48	2.3	0.20	0.78	0.208	20, 416
company directors	undirected	7 673	55 382	14.44	4.60	—	0.59	0.88	0.276	105, 323
math coauthorship	undirected	253 359	496 489	3.92	7.57	—	0.15	0.34	0.120	107, 182
physics 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	14.44	4.95	1.5/2.0	0.16	—	—	136
email addresses 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	—
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	—
Roger'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	—
Internet	undirected	10 697	31 992	5.98	—	2.5	0.035	0.39	-0.189	86, 148
power grid	undirected	4 941	6 594	2.67	18.89	—	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
metabolic networks	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	987	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

The (basic) random graph model a.k.a. ER model

13 / 72

Measuring the diameter in ER networks

Want to show that the diameter in ER networks is small



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

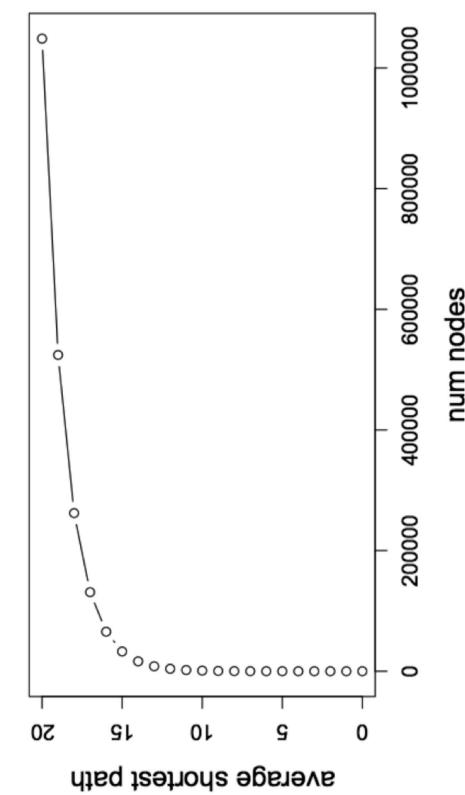
- ▶ parameter n is the number of vertices
- ▶ parameter p is s.t. $0 \leq p \leq 1$
- ▶ Generate and edge (i, j) independently at random with probability p
- ▶ Let the average degree be z
- ▶ At distance l , can reach z^l nodes
- ▶ At distance $\frac{\log n}{\log z}$, reach all n nodes
- ▶ So, diameter is (roughly) $O(\log n)$

14 / 72

ER networks have small diameter

As shown by the following simulation

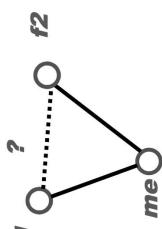
Measuring the small-world phenomenon, II



17 / 72

Global clustering coefficient

- 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



18 / 72

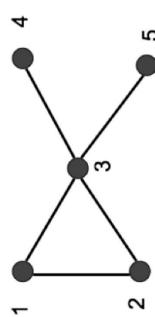
Local clustering coefficient

- 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

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

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

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



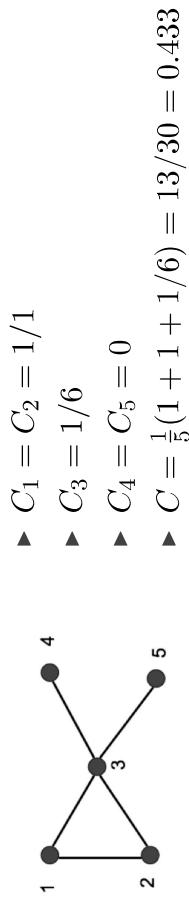
19 / 72

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

20 / 72

Local clustering coefficient example

From [Newman, 2003]



ER networks do not show transitivity

21 / 72

ER networks do not show transitivity

22 / 72

network	type	n	m	\bar{e}	α	$C^{(1)}$	$C^{(2)}$	r	Ref(s.)
film actors	undirected	449,913	25,516,482	113,43	3.48	2.3	0.78	0.208	20, 416
company directors	undirected	7,673	55,382	14,44	4.60	—	0.59	0.88	0.276
math coauthorship	undirected	253,359	496,489	3,92	7,57	—	0.15	0.34	0.120
physics coauthorship	undirected	52,909	245,300	9,27	6,19	—	0.45	0.56	0.363
biology coauthorship	undirected	1,520,251	11,803,064	15,53	4,92	—	0.088	0.60	0.127
telephone call graph	undirected	47,000,000	80,000,000	3,16	—	2.1	—	—	311, 313
email messages	directed	59,912	86,300	1,44	4,95	1.5/2.0	0.16	0.44	8, 9
email address books	directed	16,881	57,029	3,38	5,22	—	0.17	0.13	0.092
student relationships	undirected	573	477	1,66	16,01	—	0.005	0.001	-0.029
sexual contacts	undirected	2,810	—	—	—	3.2	—	—	45
WWW nd.edu	directed	269,504	1,497,135	5,55	11,27	2.1/2.4	0.11	0.29	-0.067
WWW Altavista	directed	203,549,046	2,130,000,000	10,46	16,18	2.1/2.7	—	—	14, 34
citation network	directed	783,359	67,716,198	8,57	3.0/-	—	0.13	0.15	74
Roget's Thesaurus	directed	1,022	5,103	4,99	4,87	—	—	—	351
word co-occurrence	undirected	460,902	17,000,000	70,13	—	2.7	0.44	0.157	244
Internet	undirected	10,697	31,992	5,98	3,31	2.5	0.035	0.39	86, 148
power grid	undirected	4,941	6,594	2,67	18,99	—	0.10	0.080	-0.003
train routes	undirected	587	19,603	66,79	2,16	—	0.69	0.033	416
software packages	directed	1,439	1,723	1,20	2,42	1.6/1.4	0.070	0.082	-0.016
software classes	directed	1,377	2,213	1,61	1,51	—	0.033	0.012	318
electronic circuits	undirected	24,087	53,248	4,34	11,05	3.0	0.010	0.030	-0.119
peer-to-peer network	undirected	880	1,296	1,47	4,28	2.1	0.012	0.011	-0.366
metabolic network	undirected	765	3,686	9,64	2,56	2.2	0.090	0.077	6, 354
protein interactions	undirected	2,115	2,240	2,12	6,80	2.4	0.072	0,071	-0.240
marine food web	directed	135	598	4,43	2,05	—	0.16	0.23	-0.263
freshwater food web	directed	92	997	10,84	1,90	—	0.20	0.087	-0.326
neural network	directed	307	2,359	7,68	3,97	—	0.18	0.28	-0.226

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]	253,339	3.9	0.15	0.000015
company directors [149]	449,913	113.4	0.20	0.00025
word co-occurrence [90]	7,673	14.4	0.59	0.0019
neural network [192]	460,902	70.1	0.44	0.00015
metabolic network [69]	282	14.0	0.28	0.049
food web [138]	315	28.3	0.59	0.090
	134	8.7	0.22	0.065

► $C = p$, since edges are added independently

► Given a graph with n nodes and e edges, we can “estimate” p as

$$\hat{p} = \frac{e}{1/2 n (n - 1)}$$

► We say that clustering is high if $C \gg \hat{p}$

► Hence, ER networks do not have high clustering coefficient since for them $C \approx \hat{p}$

23 / 72

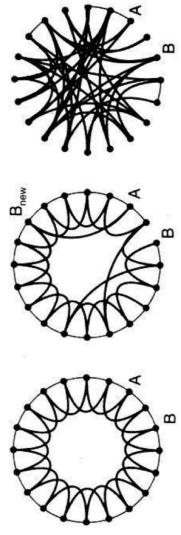
24 / 72

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

The Watts-Strogatz model, I

From [Watts and Strogatz, 1998]

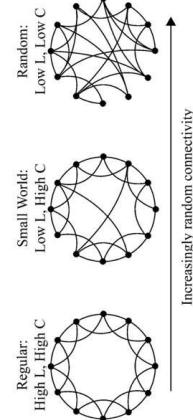
- Can we mimic this phenomenon in simulated networks (“models”), while keeping the diameter small?
- The answer is YES!



25 / 72

The Watts-Strogatz model, II

- Start with all n vertices arranged on a ring
- Each vertex has initially 4 connections to their closest nodes
 - mimics local or geographical connectivity
- With probability p , rewire each local connection to a random vertex
 - $p = 0$ high clustering, high diameter
 - $p = 1$ low clustering, low diameter (ER model)
- What happens in between?
 - As we increase p from 0 to 1
 - Fast decrease of mean distance
 - Slow decrease in clustering



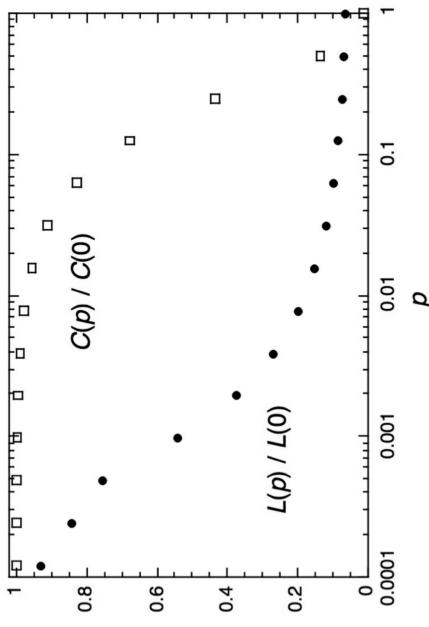
Reconciling two observations from real networks:

- High clustering: my friend's friends are also my friends
- small diameter

26 / 72

The Watts-Strogatz model, III

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

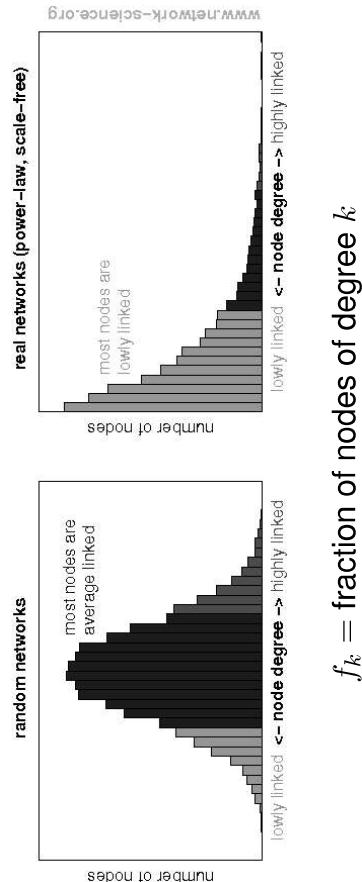


27 / 72

Degree distribution

Scale-free networks

Histogram of nr of nodes having a particular degree



f_k = fraction of nodes of degree k

29 / 72

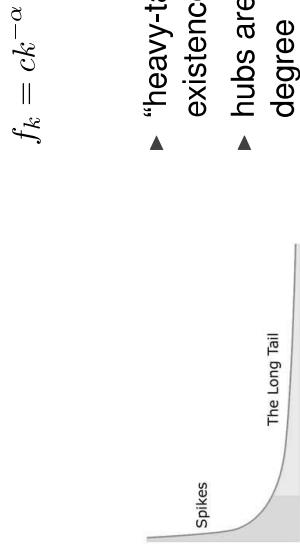
Random networks are not scale-free!

For 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
 - so no hubs

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



30 / 72

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

- Can we obtain scale-free simulated networks?
- The answer is YES!

31 / 72

Preferential attachment

Barabási-Albert model

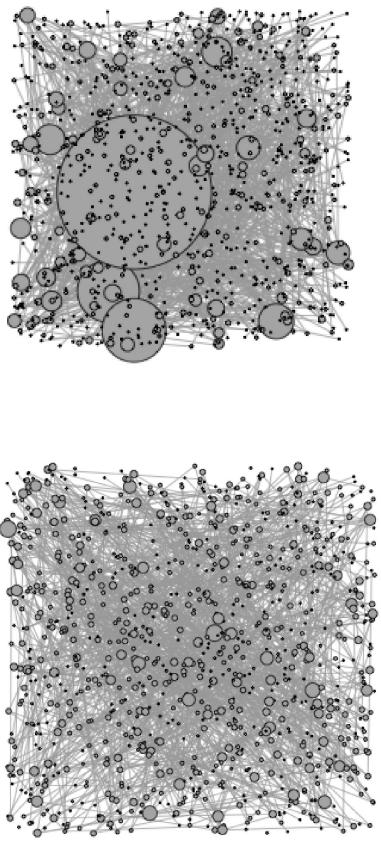
From [Barabási and Albert, 1999]

- ▶ “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
- ▶ “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 $\alpha = 3$

34 / 72

ER vs. BA

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



random

preferential attachment

phenomenon	real networks	ER	WS	BA
small diameter	yes	yes	yes	yes
high clustering	yes	no	yes	yes ¹
scale-free	yes	no	no	yes

¹clustering coefficient is higher than in random networks, but not as high as for example in W/S networks

35 / 72

In summary..

33 / 72

Network Analysis, Part II

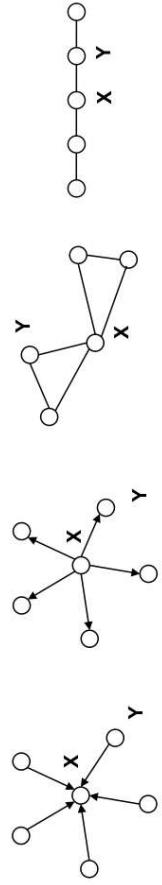
Today's contents

Centrality in Networks

1. Centrality
 - Degree centrality
 - Closeness centrality
 - Betweenness centrality
2. Community finding algorithms
 - Hierarchical clustering
 - Agglomerative
 - Girvan-Newman
 - Modularity maximization: Louvain method

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*

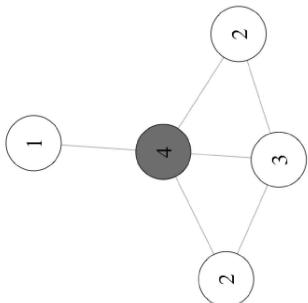


37 / 72

Degree centrality

Power through connections

$$\text{degree_centrality}(i) \stackrel{\text{def}}{=} k(i)$$

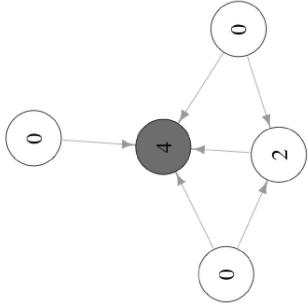


38 / 72

Degree centrality

Power through connections

$$\text{in_degree_centrality}(i) \stackrel{\text{def}}{=} k_{in}(i)$$



39 / 72

40 / 72

Degree centrality

Power through connections

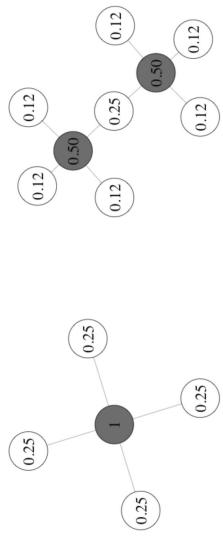
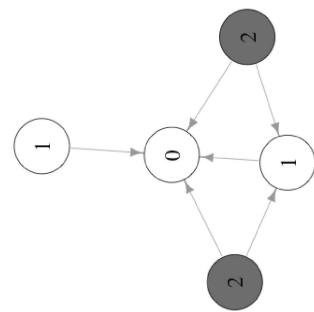
Degree centrality

Power through connections

Degree centrality

Power through connections

$out_degree_centrality(i) \stackrel{def}{=} k_{out}(i)$



By the way, there is a *normalized* version which divides the centrality of each degree by the maximum centrality value possible, i.e. $n - 1$ (so values are all between 0 and 1).

42 / 72

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

41 / 72

Closeness centrality

Power through proximity to others

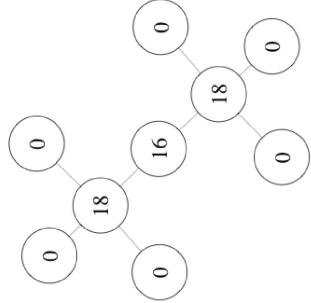
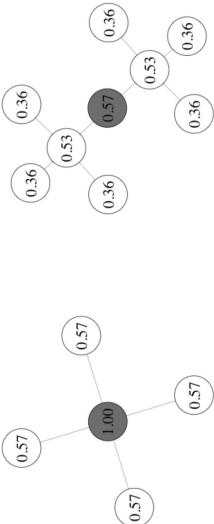
Betweenness centrality

Power through brokerage

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

- so it is essential in passing information through the network

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



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

43 / 72

Betweenness centrality

Power through brokerage

Betweenness centrality

Examples (non-normalized)

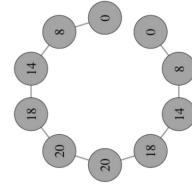
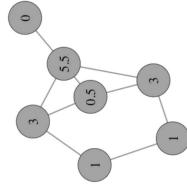
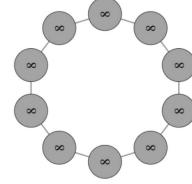
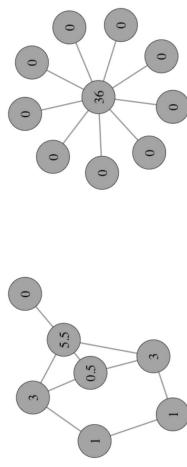
$$\text{betweenness_centrality}(i) \stackrel{\text{def}}{=} \sum_{j < k} \frac{g_{jk}(i)}{g_{jk}}$$

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:

$$\text{norm_betweenness_centrality}(i) \stackrel{\text{def}}{=} \frac{\text{betweenness_centrality}(i)}{\binom{n-1}{2}}$$

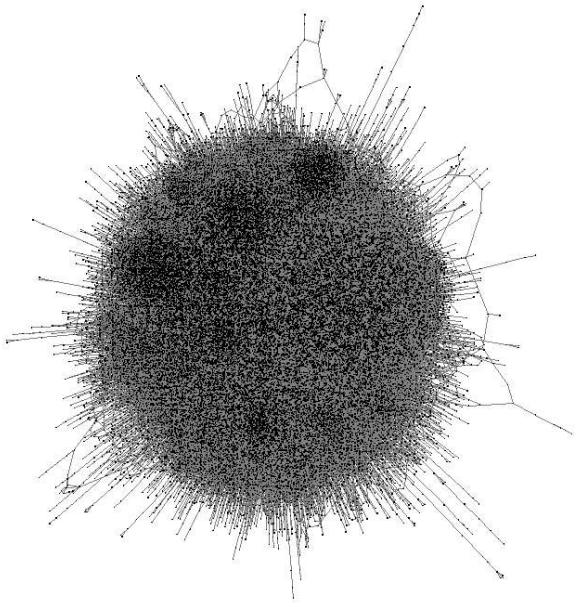
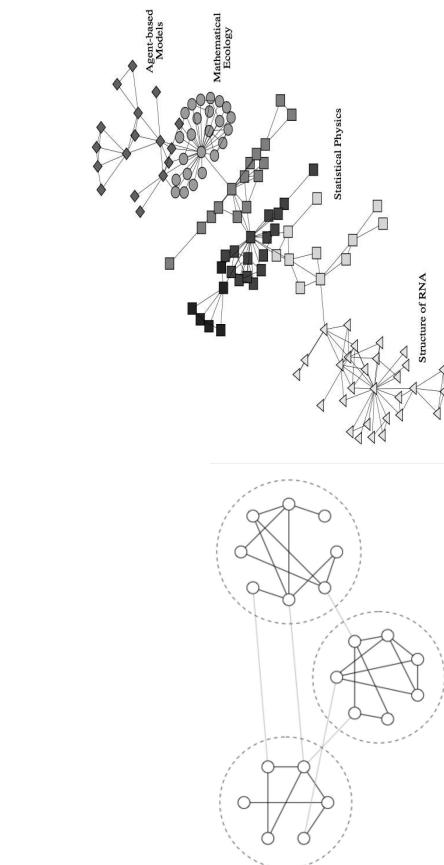


What is community structure?

45 / 72

Why is community structure important?

46 / 72



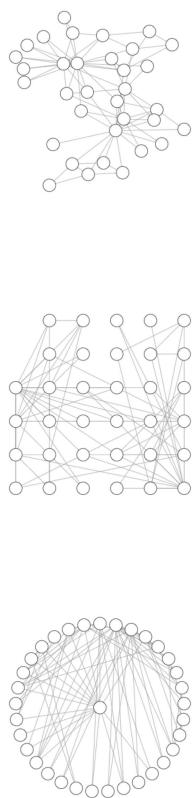
47 / 72

.. but don't trust visual perception

it is best to use objective algorithms

Main idea

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

49 / 72

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.

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

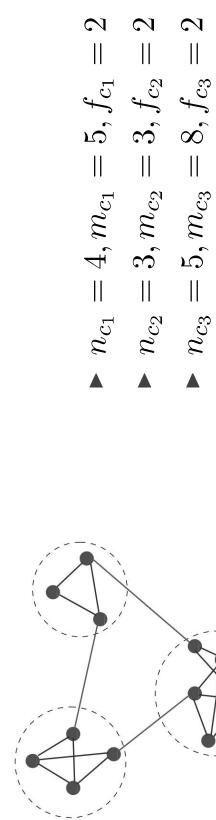
50 / 72

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:

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



51 / 72

52 / 72

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 with the same degree distribution

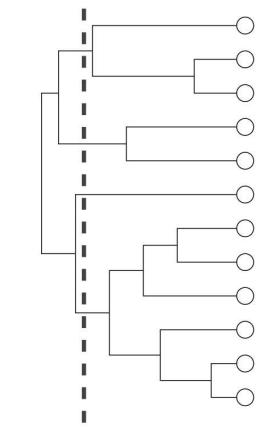
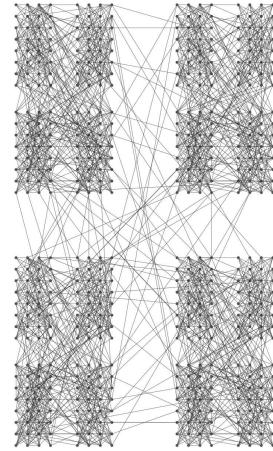
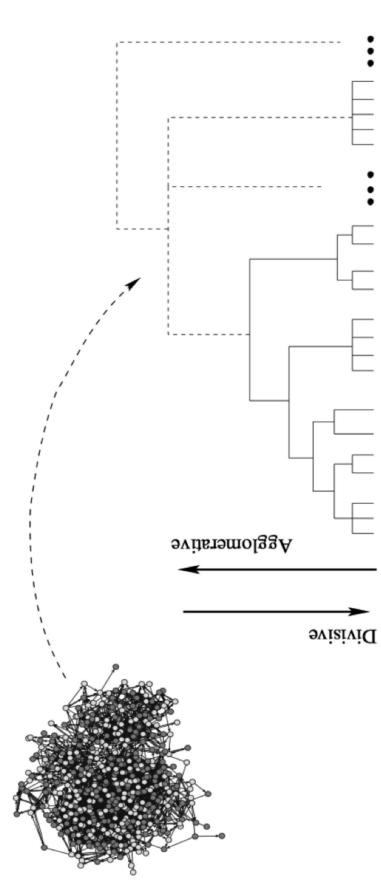
$$\frac{1}{4m} (m_c - E[m_c])$$

Hierarchical clustering

From hairball to *dendrogram*

54 / 72

Suitable if input network has hierarchical structure



55 / 72

Methods we will cover

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

56 / 72

Agglomerative hierarchical clustering [Newman, 2010]

Example

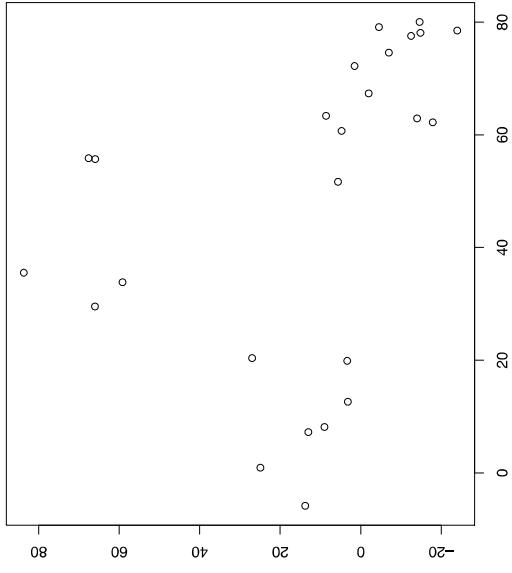
Ingredients

- 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

Data



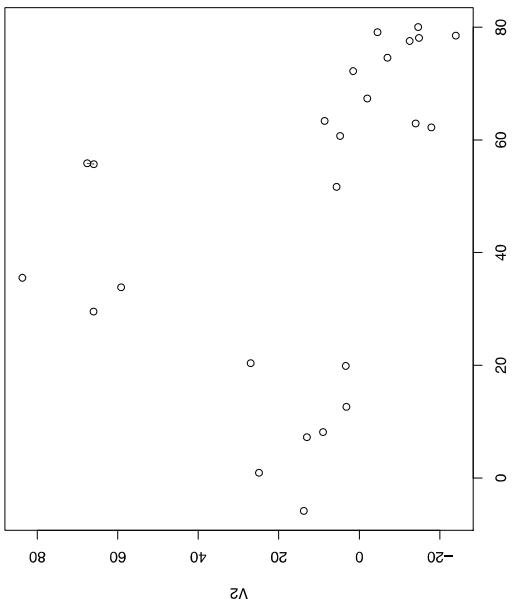
57 / 72

Example

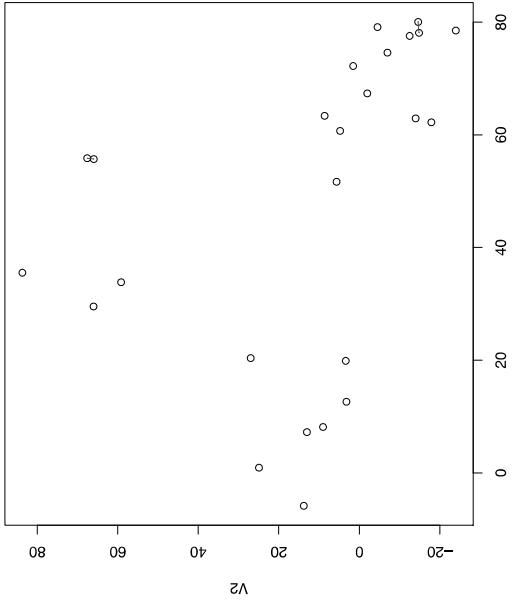
Example

5 / 21

iteration 001



iteration 002



5 / 21

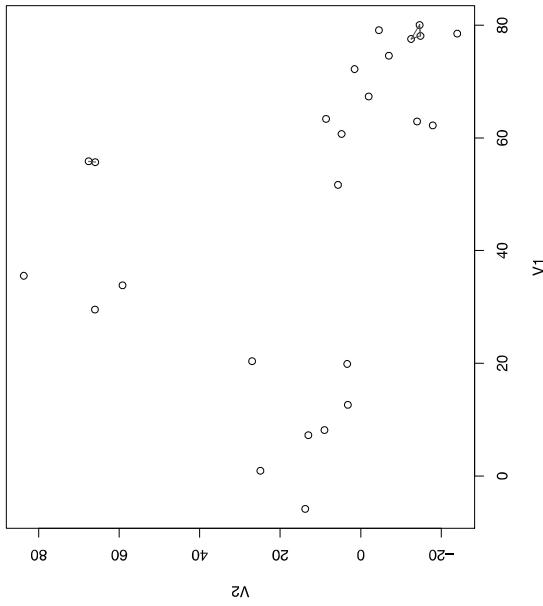
D. Blei
Clustering 02

5 / 21

D. Blei
Clustering 02

Example

iteration 003

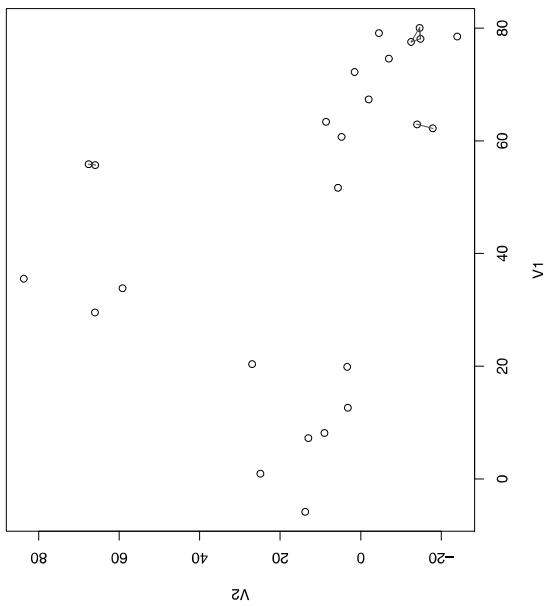


5 / 21

D. Blei Clustering 02

Example

iteration 004

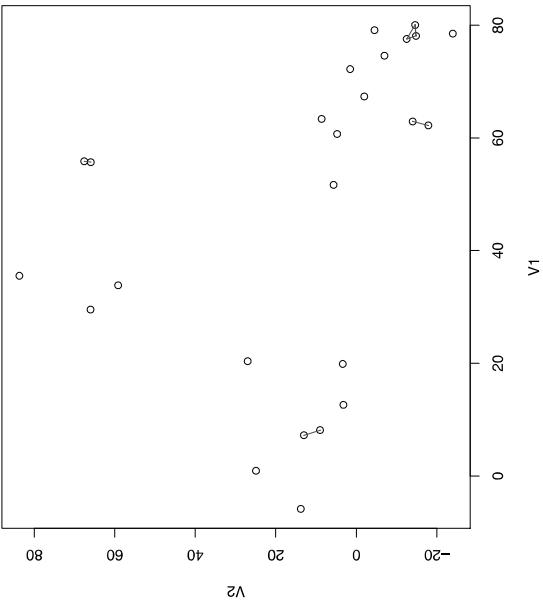


5 / 21

D. Blei Clustering 02

Example

iteration 005

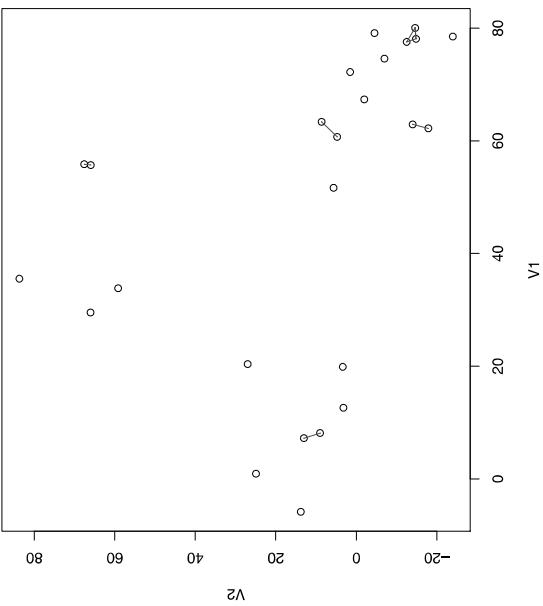


5 / 21

D. Blei Clustering 02

Example

iteration 006

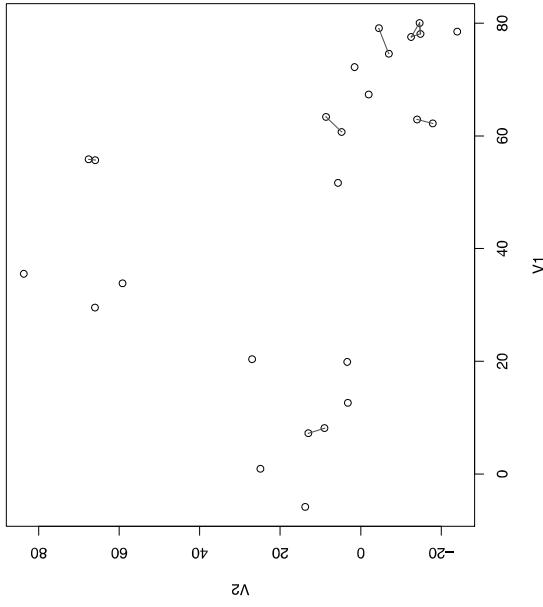


D. Blei Clustering 02

Example

Example

iteration 007

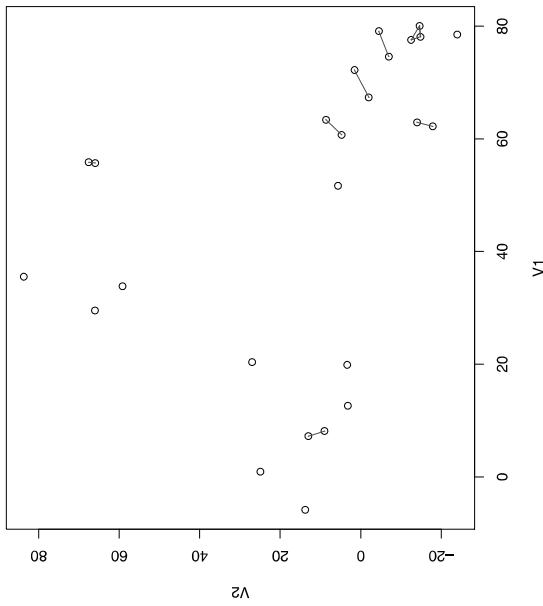


5 / 21

D. Blei Clustering 02

Example

iteration 008

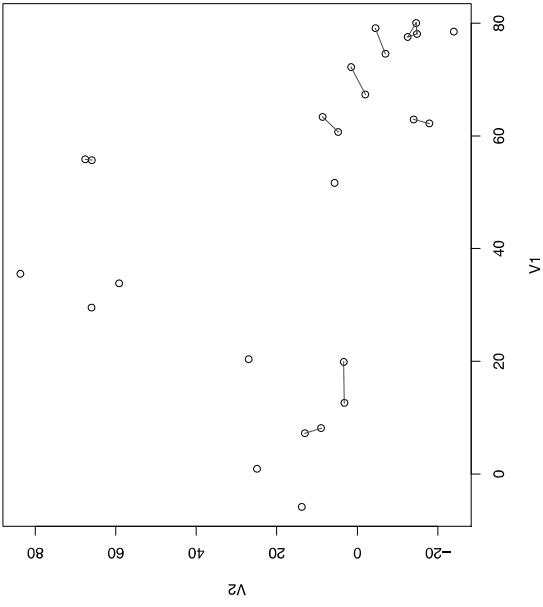


5 / 21

D. Blei Clustering 02

Example

iteration 009

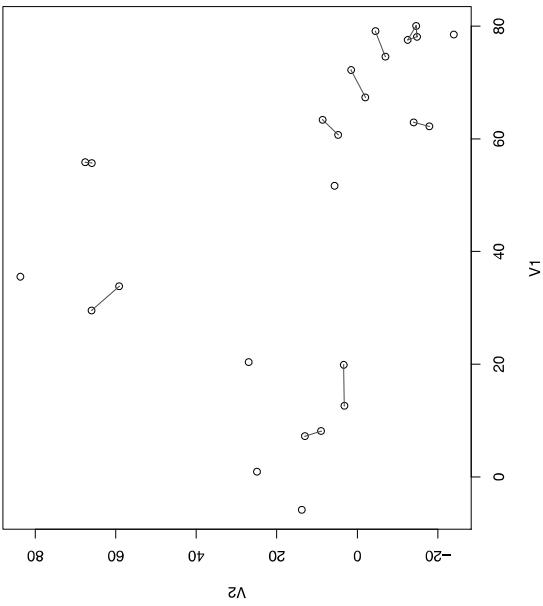


5 / 21

D. Blei Clustering 02

Example

iteration 010

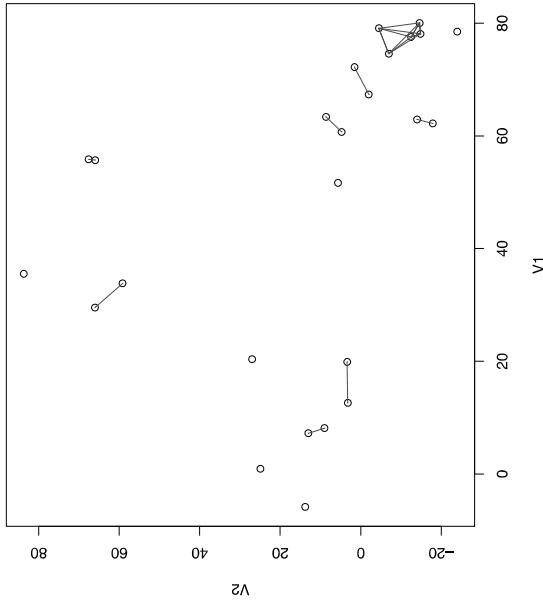


D. Blei Clustering 02

5 / 21

Example

iteration 011

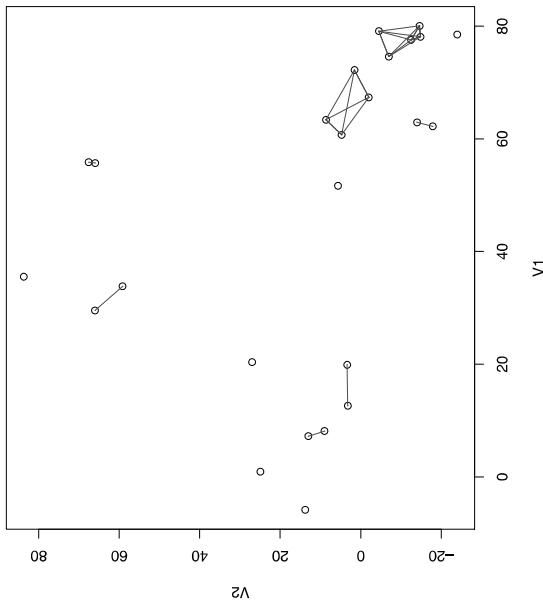


5 / 21

D. Blei Clustering 02

Example

iteration 012

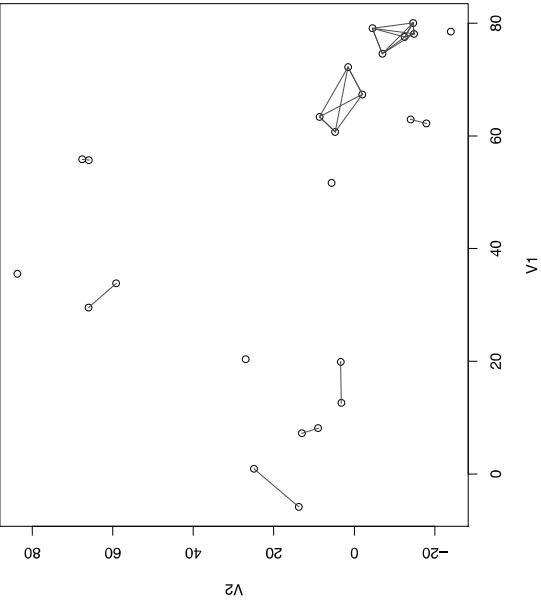


5 / 21

D. Blei Clustering 02

Example

iteration 013

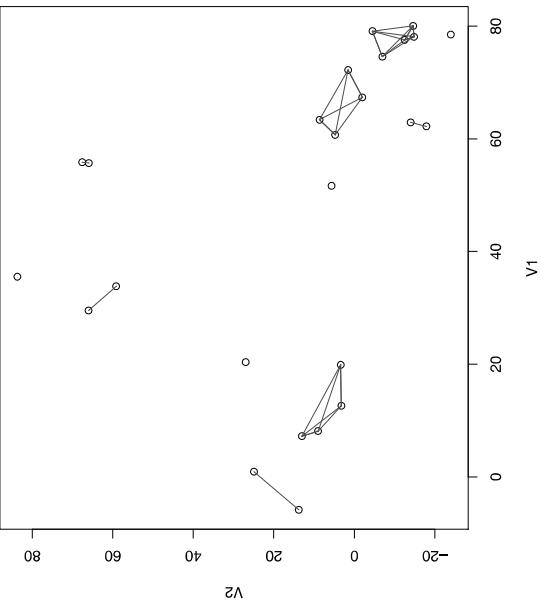


5 / 21

D. Blei Clustering 02

Example

iteration 014

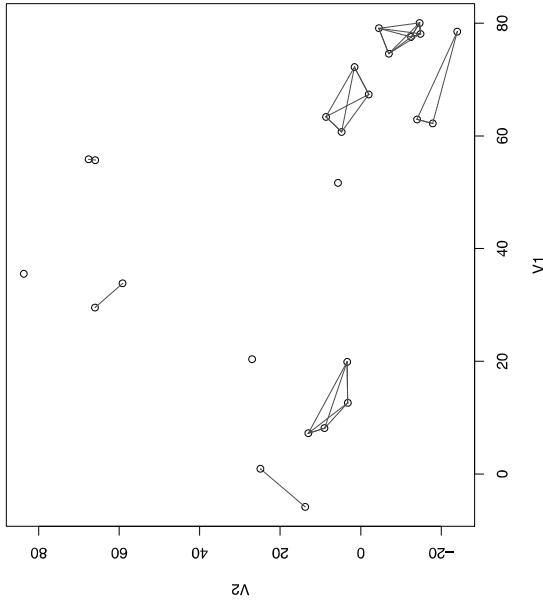


5 / 21

D. Blei Clustering 02

Example

iteration 015

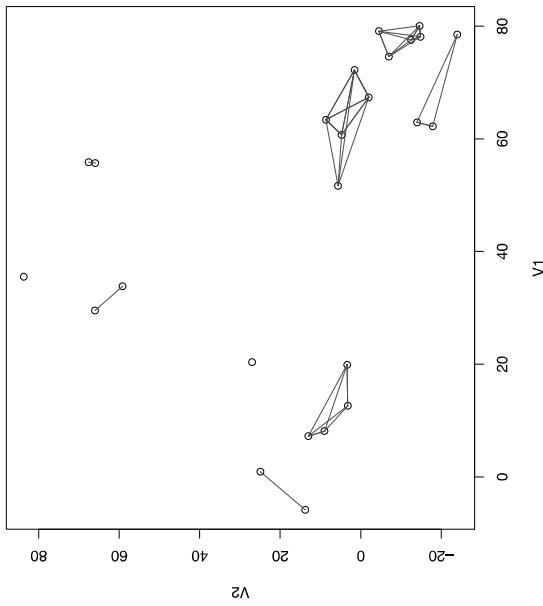


5 / 21

D. Blei Clustering 02

Example

iteration 016

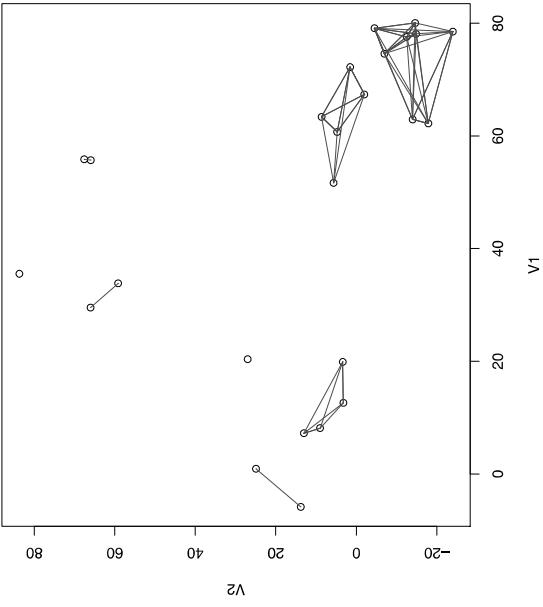


5 / 21

D. Blei Clustering 02

Example

iteration 017

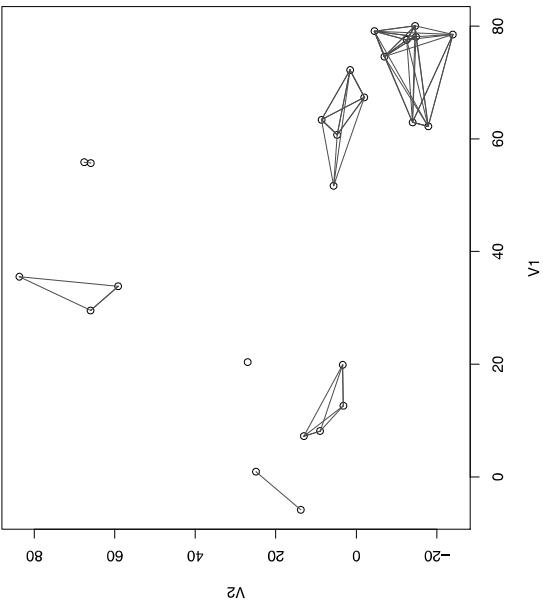


5 / 21

D. Blei Clustering 02

Example

iteration 018

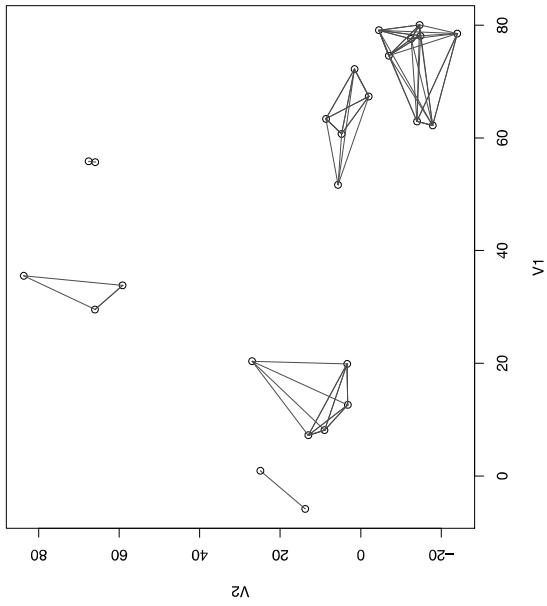


5 / 21

D. Blei Clustering 02

Example

iteration 019

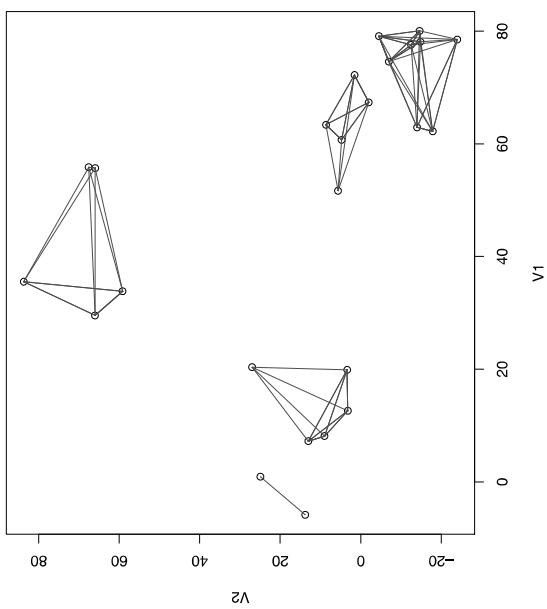


5 / 21

D. Blei Clustering 02

Example

iteration 020

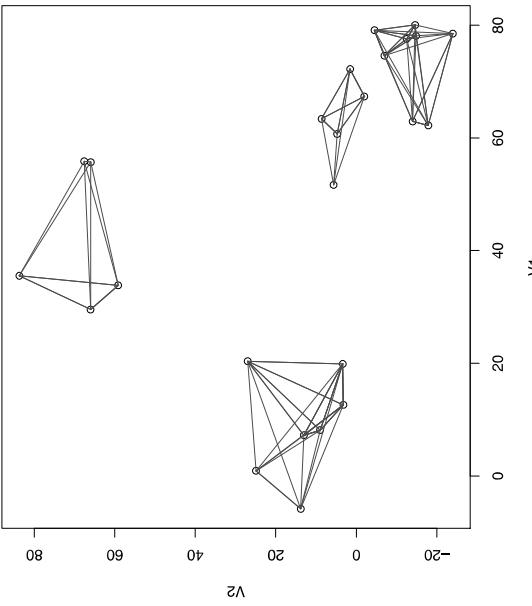


5 / 21

D. Blei Clustering 02

Example

iteration 021

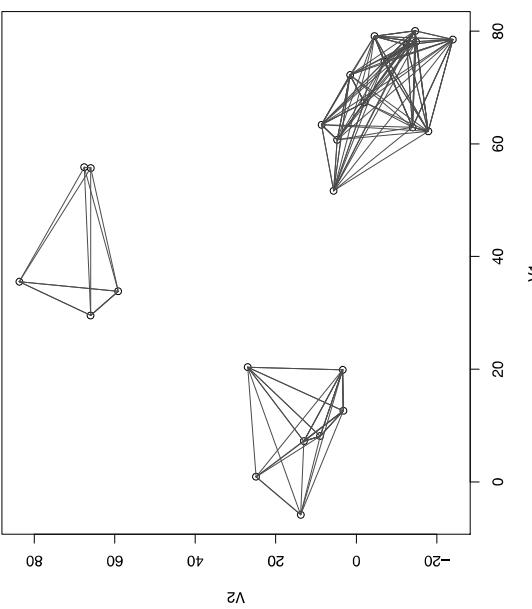


5 / 21

D. Blei Clustering 02

Example

iteration 022



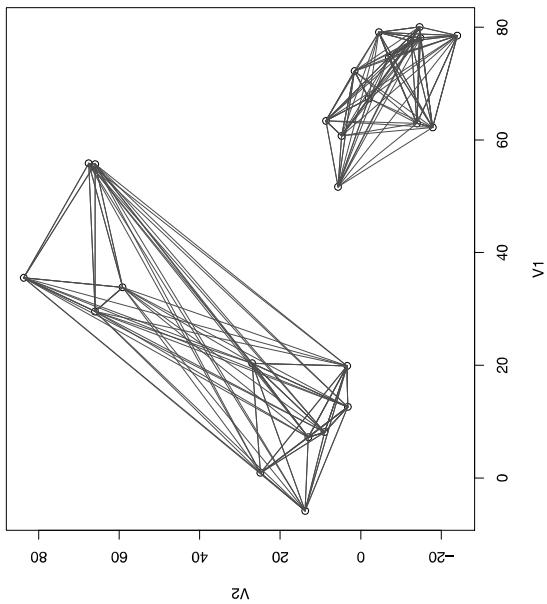
5 / 21

D. Blei Clustering 02

Example

Example

iteration 023



Similarity measures w_{ij} for nodes I

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

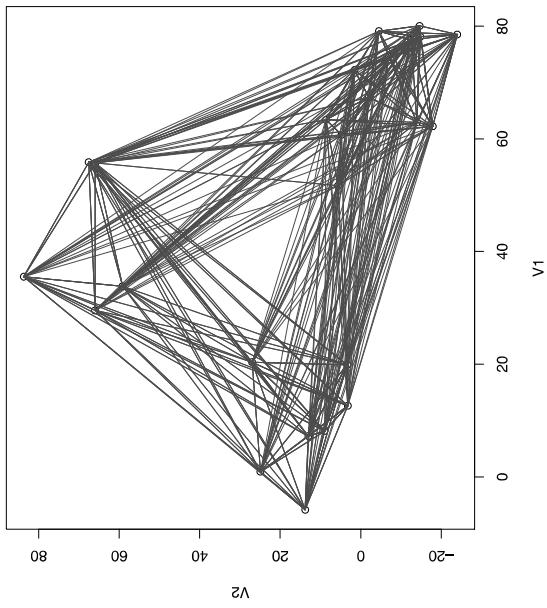
Cosine similarity:²

$$w_{ij} = \frac{\sum_k A_{ik} A_{kj}}{\sqrt{\sum_k A_{ik}^2} \sqrt{\sum_k A_{kj}^2}} = \frac{n_{ij}}{\sqrt{k_i k_j}}$$

where:

- $n_{ij} = |\Gamma(i) \cap \Gamma(j)| = \sum_k A_{ik} A_{kj}$, and
- $k_i = \sum_k A_{ik}$ is the degree of node i

iteration 024



5 / 21

D. Blei Clustering 02

Similarity measures w_{ij} for nodes II

- **Euclidean distance:** (or rather Hamming distance since \mathcal{A} is binary)

$$d_{ij} = \sum_k (A_{ik} - A_{jk})^2$$

Normalized Euclidean distance:³

$$d_{ij} = \frac{\sum_k (A_{ik} - A_{jk})^2}{k_i + k_j} = 1 - 2 \frac{n_{ij}}{k_i + k_j}$$

Pearson correlation coefficient

$$r_{ij} = \frac{cov(A_i, A_j)}{\sigma_i \sigma_j} = \frac{\sum_k (A_{ik} - \mu_i)(A_{jk} - \mu_j)}{n \sigma_i \sigma_j}$$

where $\mu_i = \frac{1}{n} \sum_k A_{ik}$ and $\sigma_i = \sqrt{\frac{1}{n} \sum_k (A_{ik} - \mu_i)^2}$

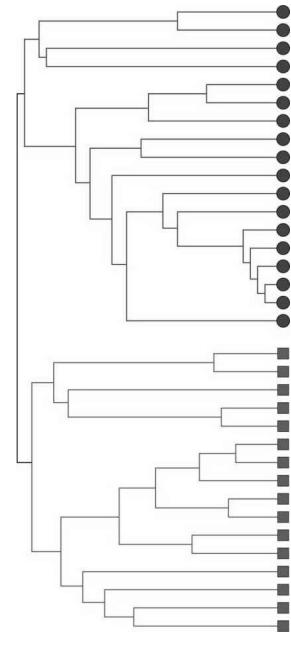
-
- From the equation $xy = |\mathbf{x}||\mathbf{y}| \cos \theta$
 - Uses the idea that the maximum value of d_{ij} is when there are no common neighbors and then $d_{ij} = k_i + k_j$

Similarity measures for sets of nodes

Agglomerative hierarchical clustering on Zachary's network

Using average linkage

- 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: $s_{XY} = \frac{\sum_{x \in X, y \in Y} s_{xy}}{|X| \times |Y|}$



The Girvan-Newman algorithm

A *divisive* hierarchical algorithm [Girvan and Newman, 2002]

60 / 72

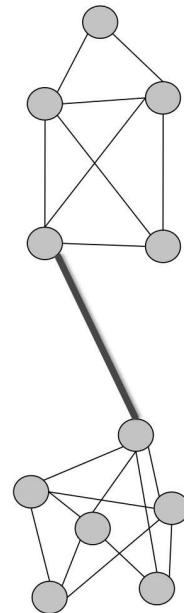
The Girvan-Newman algorithm

Pseudocode

Edge betweenness

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

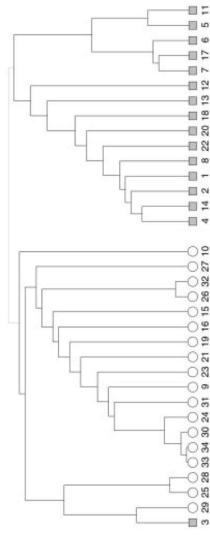
It uses the idea that “bridges” between communities must have high edge betweenness



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 dendrogram



61 / 72

62 / 72

63 / 72

Definition of modularity [Newman, 2010]

Using a *null* model

How do we compute P_{ij} ?

Using the “configuration” null model

Random graphs are not expected to have community structure,
so we will use them as null models

$$Q = (\text{nr. of intra-cluster communities}) - (\text{expected nr of edges})$$

In particular:

$$Q = \frac{1}{2m} \sum_{ij} (A_{ij} - P_{ij}) \delta(C_i, C_j)$$

where P_{ij} is the expected number of edges between nodes i and j under the null model, C_i is the community of vertex i , and $\delta(C_i, C_j) = 1$ if $C_i = C_j$ and 0 otherwise.

Properties of modularity

$$Q = \frac{1}{2m} \sum_{ij} \left(A_{ij} - \frac{k_i k_j}{2m} \right) \delta(C_i, C_j)$$

- Q depends on nodes in the same clusters only
- Larger modularity means better communities (better than random intra-cluster density)
- $Q \leq \frac{1}{2m} \sum_{ij} A_{ij} \delta(C_i, C_j) \leq \frac{1}{2m} \sum_{ij} A_{ij} \leq 1$
- Q may take negative values
 - partitions with large negative Q implies existence of cluster with small internal edge density and large inter-community edges

The “configuration” random graph model chooses a graph with the same degree distribution as the original graph uniformly at random.

- Let us compute P_{ij}
- There are $2m$ stubs or half-edges available in the configuration model
- Let p_i be the probability of picking at random a stub incident with i

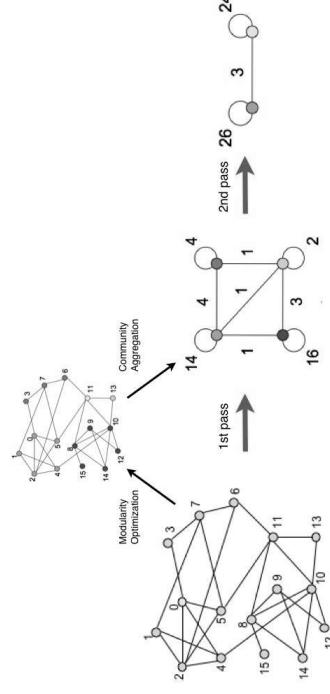
$$p_i = \frac{k_i}{2m}$$

- The probability of connecting i to j is then $p_i p_j = \frac{k_i k_j}{4m^2}$
- And so $P_{ij} = 2m p_i p_j = \frac{k_i k_j}{2m}$

64 / 72

The Louvain method [Blondel et al., 2008]

Considered state-of-the-art



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

66 / 72

67 / 72

The Louvain method

Phase 1: optimizing modularity

The Louvain method

Phase 2: agglomerating clusters to form new network

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 choose to place i into community of neighbor that leads to highest modularity gain
3. Repeat until no improvement can be done

The Louvain method

Observations

- The output is also a hierarchy
- Works for weighted graphs, and so modularity has to be generalized to

$$Q^w = \frac{1}{2W} \sum_{ij} \left(W_{ij} - \frac{s_i s_j}{2W} \right) \delta(C_i, C_j)$$

where W_{ij} is the weight of undirected edge (i, j) ,
 $W = \sum_{ij} W_{ij}$ and $s_i = \sum_k W_{ik}$.

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)

68 / 72

References |

- Observations
- Barabási, A.-L. and Albert, R. (1999). Emergence of scaling in random networks. *science*, 286(5439):509–512.
 - Blondel, V. D., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E. (2008). Fast unfolding of community hierarchies in large networks. *Networks*, pages 1–6.
 - Girvan, M. and Newman, M. E. J. (2002). Community structure in social and biological networks. *Proceedings of the National Academy of Sciences of the United States of America*, 99:7821–7826.
 - Newman, M. (2010). *Networks: An Introduction*. Oxford University Press, USA, 2010 edition.

69 / 72

70 / 72

71 / 72

References ||

- Newman, M. E. (2003).
The structure and function of complex networks.
SIAM review, 45(2):167–256.
- Watts, D. J. and Strogatz, S. H. (1998).
Collective dynamics of small-world networks.
nature, 393(6684):440–442.