Session II

Community Detection and Centrality

Community Detection

Based on the slides for Data Driven SNA
by Kaltenbrunner & Gómez

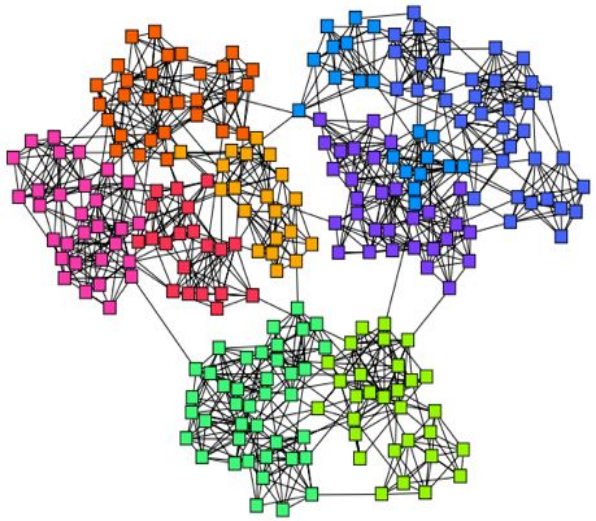
Community structure

Definition

Vertices in networks are often found to cluster into tightly-knit groups with a high density of within-group edges and a lower density of between-group edges.

Applications

- Identify groups of users who are more likely to interact with each other
- Identify customer with similar interests (purchasing history)
- Graph Compression
- Classification of vertices



Finding Communities

Given a graph G=<V,E>

- Community detection problem: find modules and their hierarchical organization

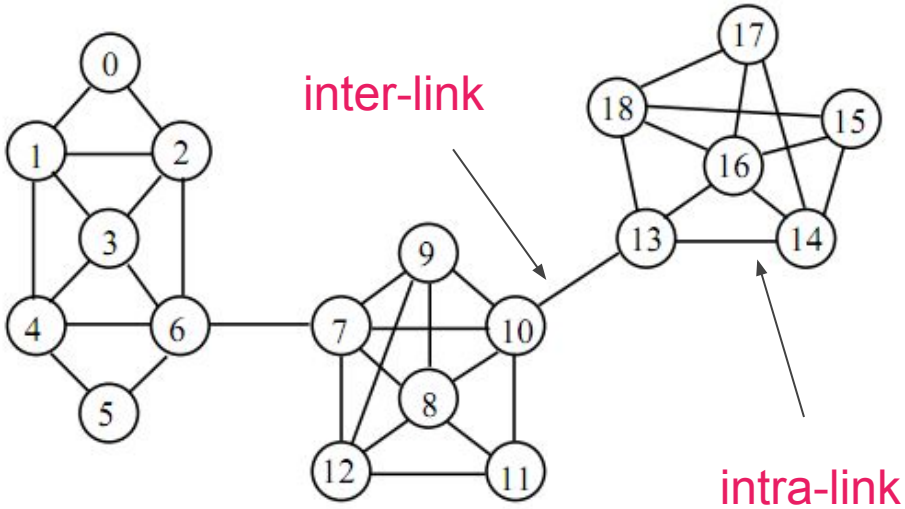
- What do we miss?
  - Define what is “a community”
  - Design algorithms that will find set of nodes which lead to “good communities”
  - Why just “good”??
  - Evaluate different results

Community

Definition

- There is no universally accepted definition of community

- Informally, a community C is a subset of nodes of V such that there are more edges inside the community than edges linking vertices of C with the rest of the graph



Community

Properties

Intra cluster density:

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

Inter cluster density:

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

$\partial_{ext}(C) << 2m/n(n-1) << \partial int(C)$

Community detection makes only sense on sparse graphs

Notation

V	set of nodes
E	set of edges
n	|V|
m	|E|
C	Subset of V
n_c	|C|

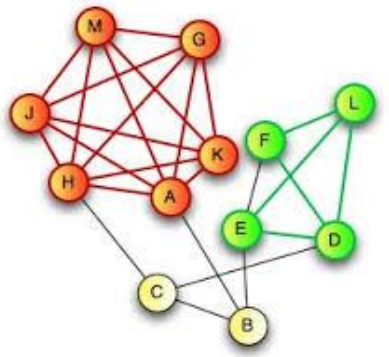
Local community: Clique

Definition

Subset of V such that all the vertices are adjacent to each other

Properties

- Triangles are really frequent in real networks
- Finding cliques in a graph is NP Complete
- However efficient algorithms for sparse graphs exist (e.g. Bron-Kerbosch algorithm)
- Too strict definition for Communities



Local community: Pseudo-clique

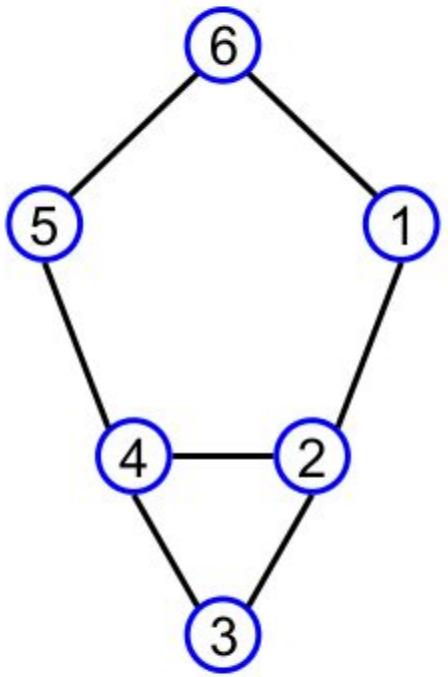
---

k-clique: Subset of vertices C: for every node i, j d(i,j)≤ k in G

k-club: Subset of vertices C: diam(G[C]) ≤ k

Exercise

- Find: 2-clique y 2-club
  (2,3,4 are 1-clique and 1-club)



Local community: Pseudo-clique

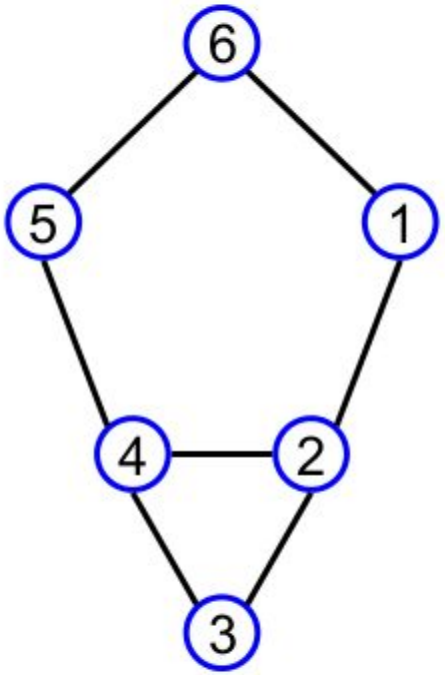
---

k-clique: Subset of vertices C: for every node i, j d(i,j)≤ k in G

k-club: Subset of vertices C: diam(G[C]) ≤ k

Exercise

- Find: 2-clique y 2-club
  (2,3,4 are 1-clique and 1-club)
  - 2-club:{1,2,4,5,6}
  - 2-clique:{1,2,3,4,5},{1,2,4,5,6}



Global community

Properties

- A graph has a community structure if it is different from a random graph
- A random graph is not expected to have any community structure: any two vertices have the same probability to be adjacent
- We can define a null model and use it to investigate whether a graph under consideration exhibits a community structure

Communities based on Similarity

- Community can be defined as a subset of vertices similar to each other.

- Structural equivalence: Nodes V_1 and V_2 are structurally equivalent if they are not adjacent but share the same neighbors.

$$d_{ij} = \sqrt{\sum_{k \neq i,j} (A_{ik} - A_{jk})^2},$$

- Overlap between the neighborhoods Γ(i) and Γ (j) of vertices i and j

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

- Pearson

$$C_{ij} = \frac{\sum_k (A_{ik} - \mu_i)(A_{jk} - \mu_j)}{n\sigma_i\sigma_j}, \quad \mu_i = (\textstyle\sum_j A_{ij})/n \quad \sigma_i = \sqrt{\textstyle\sum_j (A_{ij} - \mu_i)^2/n}.$$

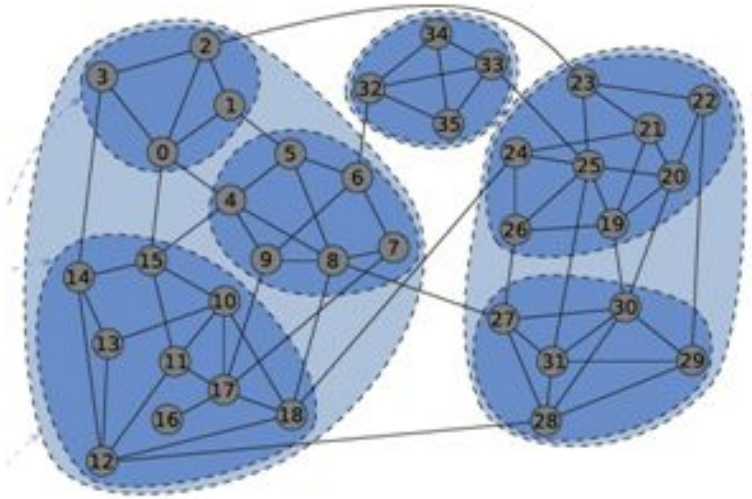
Graph partition

Definition

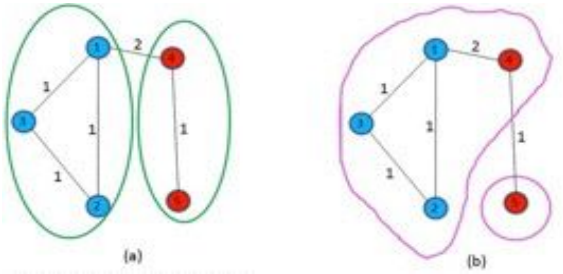
Division of a graph in clusters (communities), such that each vertex belongs to (at least)one cluster

Applications

- Hierarchical organization (communities can be embedded within other communities)
- Nodes can be shared between different communities (overlapping communities)



Comparing different partitions



- What is a good clustering?
- Answer is given by **quality function** that assigns a number to each partition of a graph
- We can rank partitions based on their score given by the quality function. A quality function Q is additive if there is an elementary function q such that, for any partition P of a graph

$$Q(\mathcal{P}) = \sum_{\mathcal{C} \in \mathcal{P}} q(\mathcal{C}),$$

- Performance measure P

$$P(\mathcal{P}) = \frac{|\{(i,j) \in E, C_i = C_j\}| + |\{(i,j) \notin E, C_i \neq C_j\}|}{n(n-1)/2}.$$

vertices belonging to the same community and connected by an edge
+ # vertices belonging to different communities and not connected by an edge.

Modularity

Definition

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

(density of edges in a subgraph vs density in a null model graph)

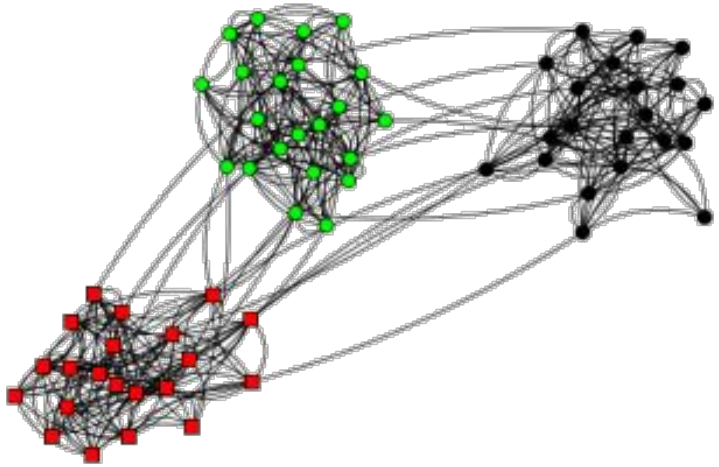
Applications

- The δ-function yields one if vertices i and j are in the same community, zero otherwise.
- P_{ij} represents the expected number of edges between vertices i and j in the null model (which is arbitrary)
- A_{ij} is the actual number of edges.

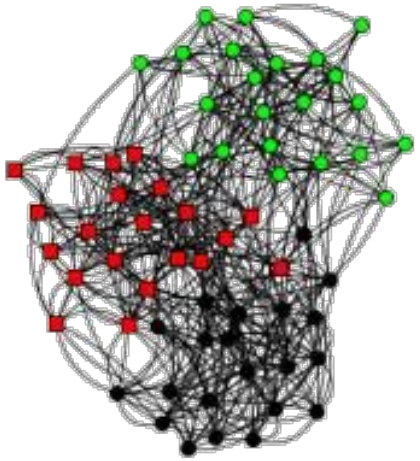
Modularity

Example

- In a random graph (Erdős-Rényi model), we expect that any possible partition would lead to Q=0
- Typically, in non-random graphs modularity takes values between 0.3 and 0.7.



Q = 0.60 clear community structure



Q = 0.37 fuzzy communities

Hierarchical Clustering

- Widely used in social network analysis
  - No need to specify the number of clusters
  - Graph may have a hierarchical structure
- Hierarchical Clustering aim at identifying groups of vertices with high similarity (not focusing on connectedness)
  - Define a similarity measure between vertices
  - Compute the n x n similarity matrix
  - **Agglomerative algorithms**: (bottom up) clusters are merged if their similarity if sufficiently high
  - **Divisive algorithms**: (top-down) clusters are iteratively split by removing edges connecting vertices with low similarity

Hierarchical Clustering: Merging Clusters

Merging Clusters criteria:

- Minimum od single linkage

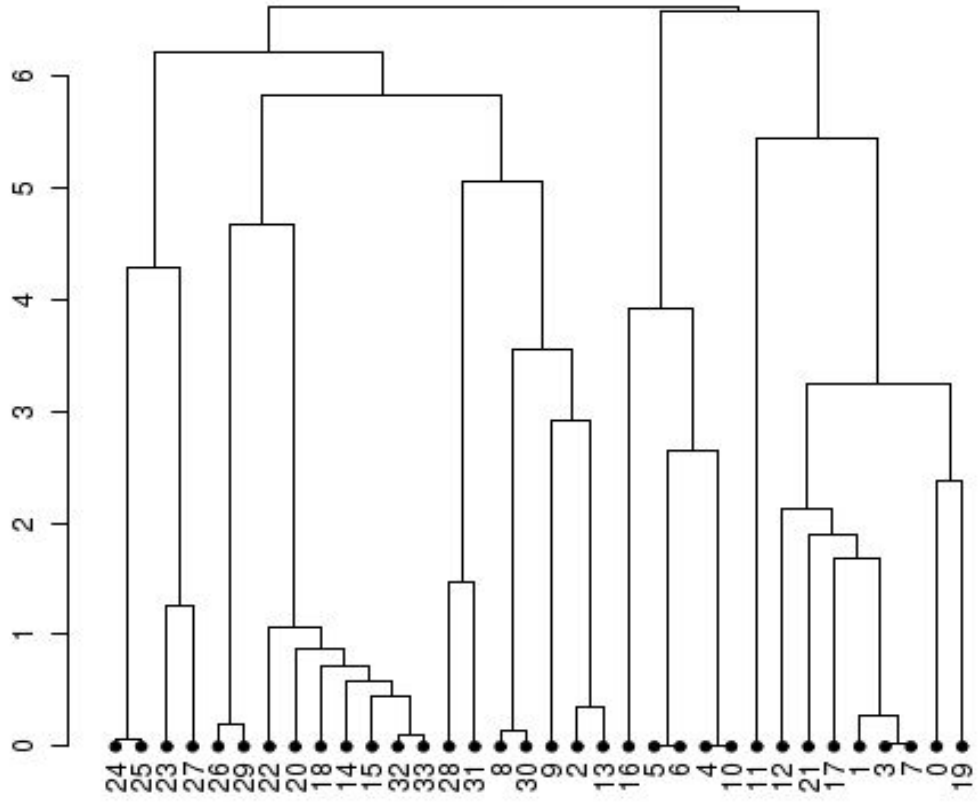
$$\max\{\, d(a,b) : a \in A,\, b \in B \,\}.$$

- Maximum or Complete linkage

$$\min\{\, d(a,b) : a \in A,\, b \in B \,\}.$$

- Average linkage

$$\frac{1}{|A||B|} \sum_{a \in A} \sum_{b \in B} d(a,b).$$



Drawback of the hierarchical procedure: it does not provide a way to discriminate which level better represents the community structure of the graph

Girvan-Newman Method

Definition [Girvan 2002]

Divisive method that detect edges that connect different communities and remove them until clusters are disconnected

Steps

1. Compute Edge centrality
2. Remove the edge with the highest centrality
3. Update Centralities
4. If number of edges |E|>0, go to step 2

Girvan-Newman Method

- Instead of trying to construct a measure which tells us which edges are most central to communities, we focus instead on those edges which are least central

- If a network contains communities or groups that are only loosely connected by a few inter-group edges, then all shortest paths between different communities must go along one of these few edges

Girvan-Newman Method

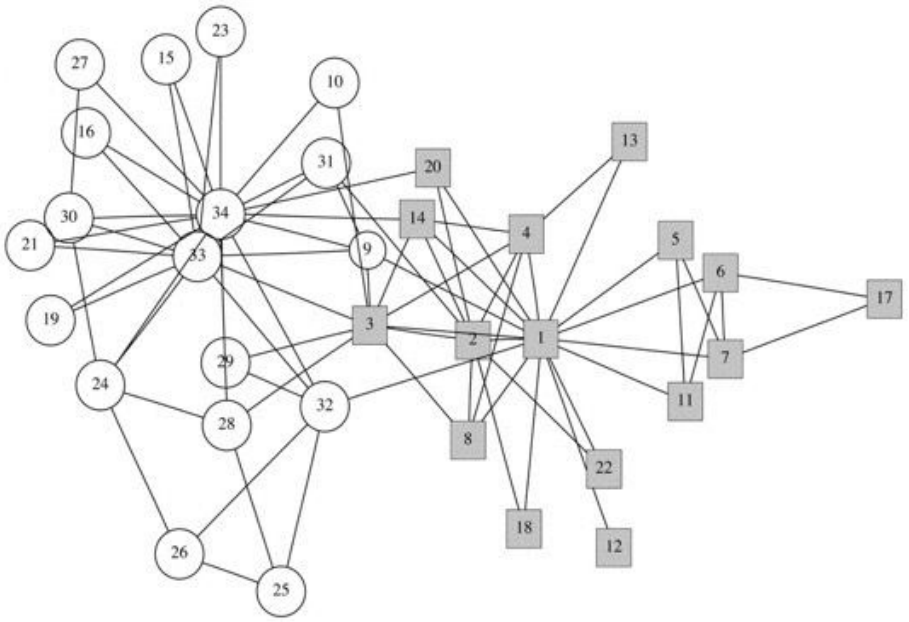
- Edge Betweenness O(mn)

  number of shortest path between all vertex pair that run along the considered edge

- The edges connecting communities will have high edge betweenness

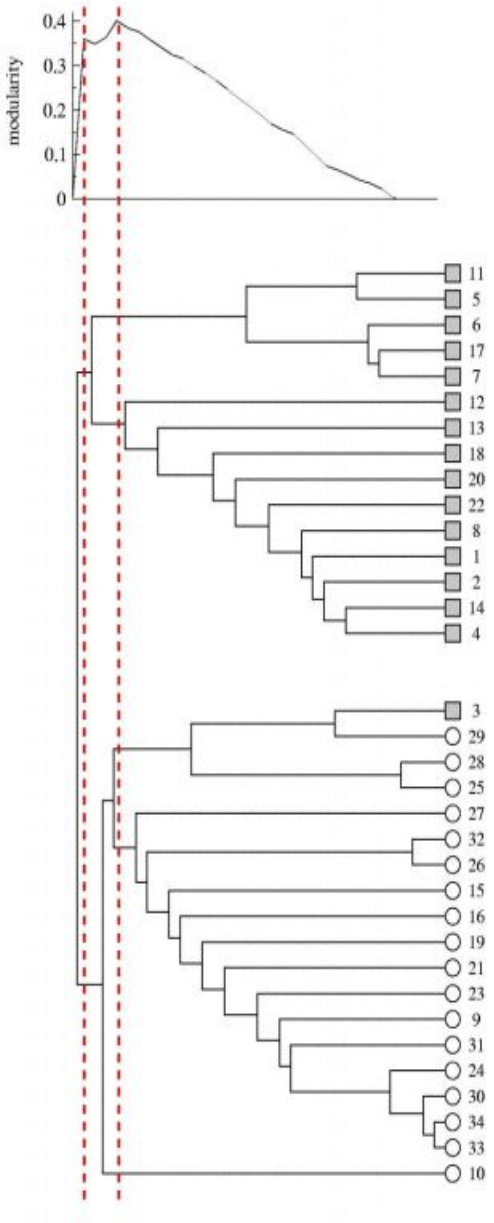
- Which partition is the best?
- Answer: Compute modularity

Girvan-Newman Method



Optimal community structure for Zachary's Karate club

Modularity without recalculation



Modularity optimization

If high modularity indicate goods partition, why not simply optimize Modularity over all partitions to find the best one?

$$Q = \sum_i (e_{ii} - a_i^2)$$

- e_{ii} is the fraction of edges in the network that connect vertices in the group i.
- a_i is the fraction of edges that connect vertices in the group i with every other group (including group i).

Answer: The search-space is exponential in |V|

Modularity optimization: Approximate Solution

Greedy algorithm [Newman 2004]

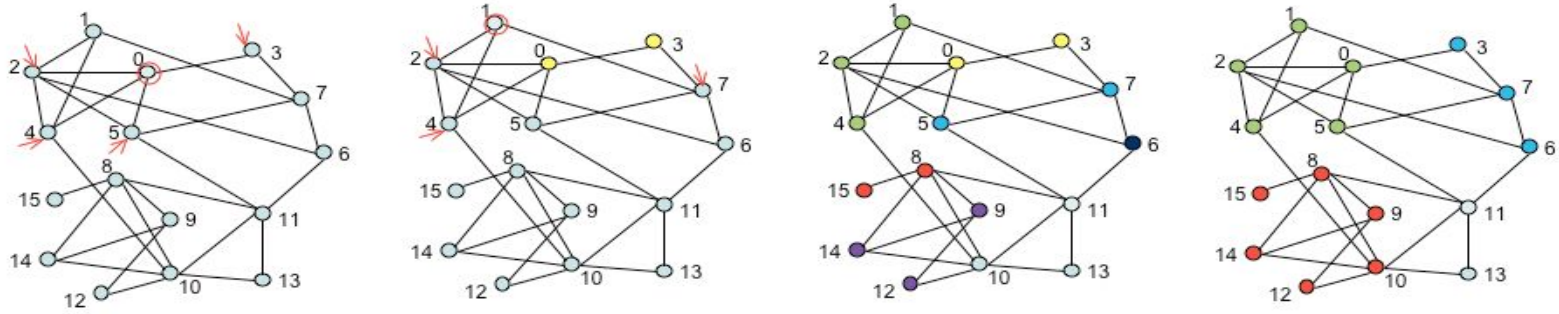
- **Agglomerative clustering**: we repeatedly join communities together in pairs, choosing at each step the join that results in the greatest increase (or smallest decrease) in Q

- Note: joining communities that are not connected cannot result in an increase in Q. => This limits the number of tentative joins to (m)

Louvain Method

Steps [Blondel 2008]

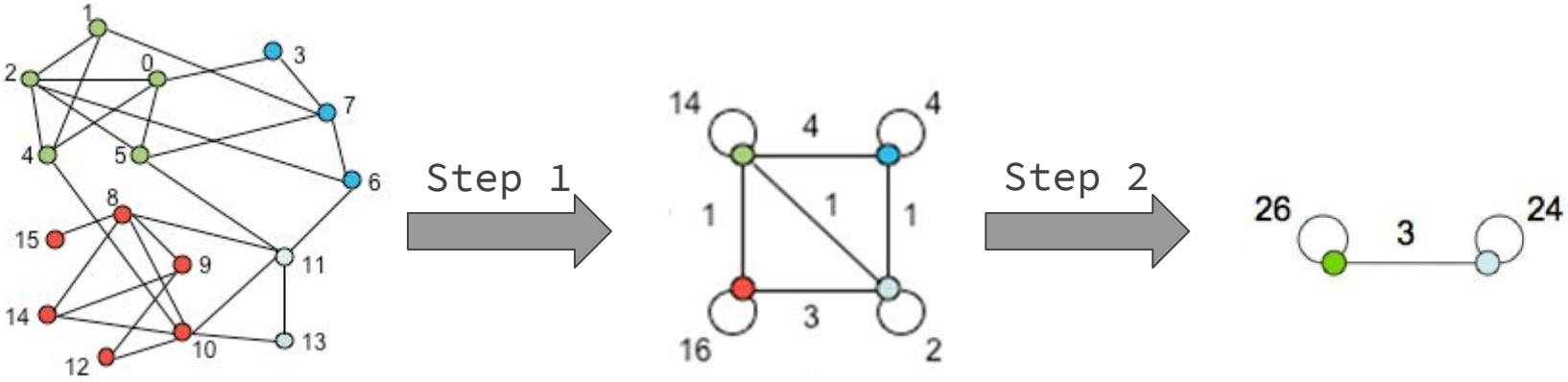
1. Initially, each node belongs to its own community (n nodes => n communities)
2. Pass through each node with a standard order. To each node, assign the community of their neighbor as long as this leads to an increase in modularity.
3. This step is repeated many times until a local modularity maximum is found.



Louvain Method

Folding

- Create new graph in which nodes correspond to the communities detected in the previous step.
- Edge weights between community nodes are defined by the number of inter-community edges.
- Folding ensures rapid decrease in the number of nodes that need to be examined and thus enables large-scale application of the method.



Louvain Method

Observations

- The output is also a hierarchy

- The method works for weighted graphs, and so modularity has to be generalized to

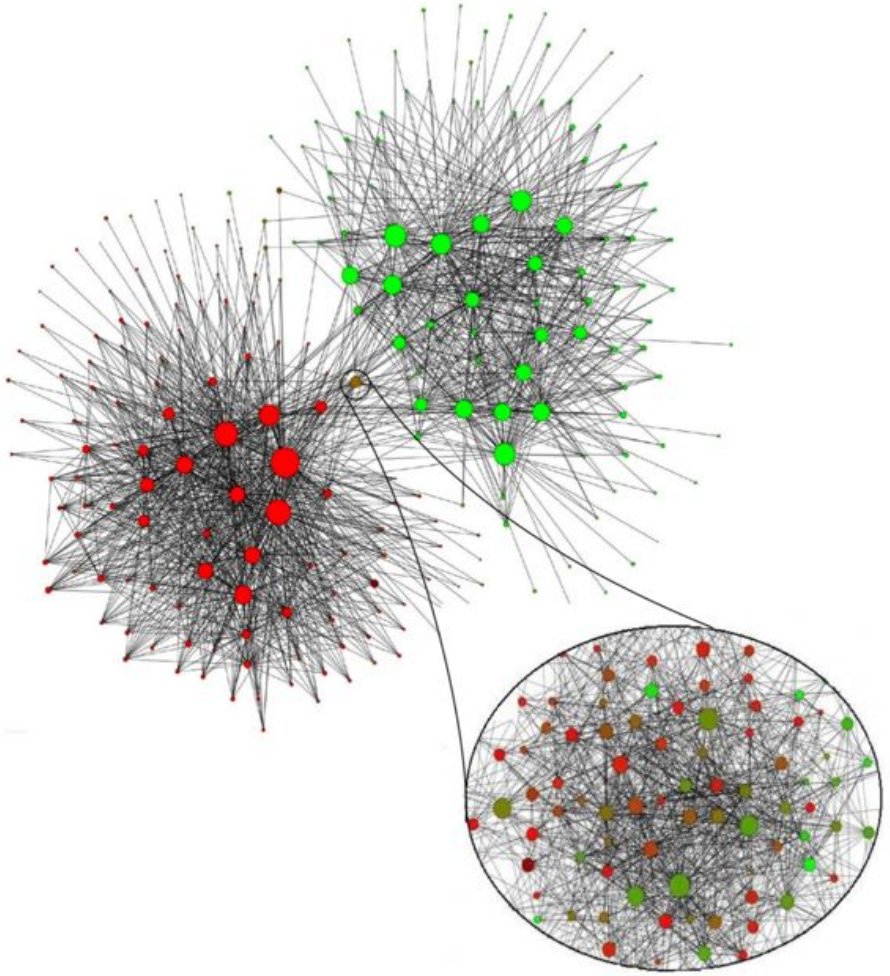
$$Q^w = \frac{1}{2W} \sum_{ij} \left(W_{ij} - \frac{s_i s_j}{2W} \right) \delta(C_i, C_j)$$

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

Louvain Method

Example

- Cell phone operator from Belgium
- 2.6 million customers
- 260 assemblages with over 100 customers, 36 with over 10,000
- 6 assemblage levels
- French and Dutch segments are almost independent



Conclusions

- Social networks are typically formed by communities of nodes.

- A community is a group of nodes with many edges between them and few edges with the rest of the nodes of the network.

- There are methods to detect communities, in this course we will use the **Louvain Method:**
  - Good results
  - Very fast

Centrality

Motivation

- People influence each other
- Interactions among individuals affect the thoughts, feelings and actions of others
- Can you measure the potential of a person in a social network to influence others?

Source: Mashable

Scores of influence

Core Concepts

Connecting Networks Can Only Help Your Score

We want to help you understand your influence wherever it may exist. We also understand, given the number of different networks out there, that it is nearly impossible for any person to be consistently effective across every network. Adding more networks helps us more accurately measure your influence and can only increase your Score.

Influence is Built Over Time

In most instances, your influence should not radically change from one day to the next. The Klout Score is based on a rolling 90-day window, with recent activity being weighted more than older activity. Being inactive over the weekend or taking a short break won't have a major impact on your Score, but if you're inactive for longer periods your Score will decrease gradually.

Influence is the Ability to Drive Action

It's great to have lots of connections, but what really matters is how people engage with the content you create. We believe it's better to have a small and engaged audience than a large network that doesn't act respond to your content.

Everyone has Klout

You are never penalized for connecting or engaging with someone with a low Klout Score. In fact, you are helping build their Klout Score. The more influential you are, the greater impact you have. All engagement positively contributes to your Score.

Klout is Constantly Evolving

The social Web is changing every day and the Klout Score will continue to evolve and improve. The best strategy for obtaining a high Klout Score is to simply create great content that your network wants to share and engage with.

Being Active is Different than Being Influential

Retweets, Likes, comments and other interactions on the social Web are all signals of influence. However, just looking at the count of these actions does not tell the whole story of a person's influence. It's important to look at how much content a person creates compared to the amount of engagement they generate.

KLOUT

http://klout.com

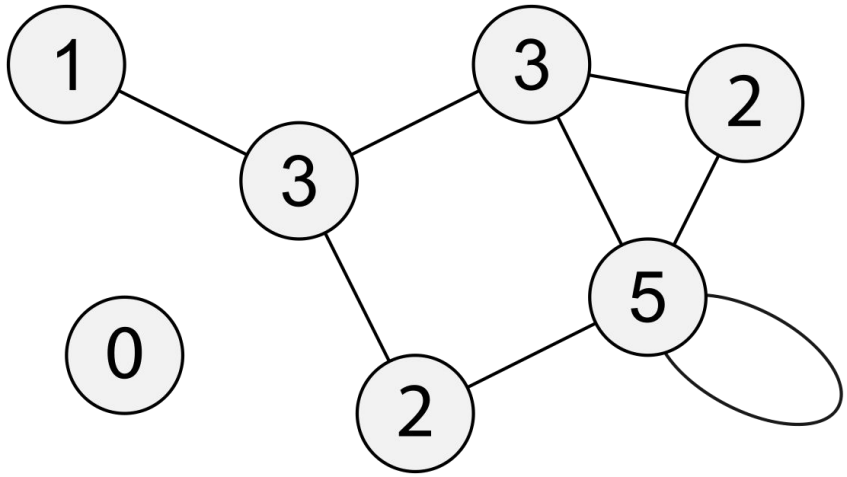
Degree centrality

Motivation

Identify the nodes with the highest number of links to other nodes

Method

- Node degree



Source: Wikipedia

Closeness centrality

Motivation

- In a diffusion model, it is often interpreted as the arrival time of something flowing through the network
- It measures the accessibility of one node to another.

Method

- It is the sum of the distances in a network from all the nodes in the network, where the distance from one node to another is defined as the length of the shortest path from one node to another.

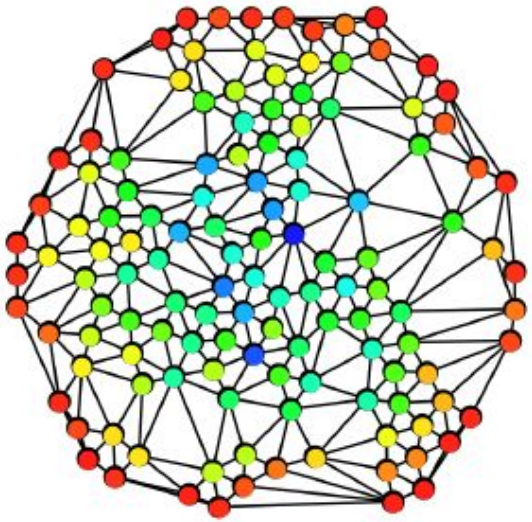
Betweenness centrality

Motivation

- Frequency that a node occurs on the shortest path between two others

Method

- Node i: $C_B(i)=\sum_{s\neq i\neq t\in V}\sigma_{st}(i)/\sigma_{st}$
- $\sigma_{st}(i)$ number of different shortest paths between nodes s,t
- σ_{st} number of different shortest paths between nodes s,t containing i

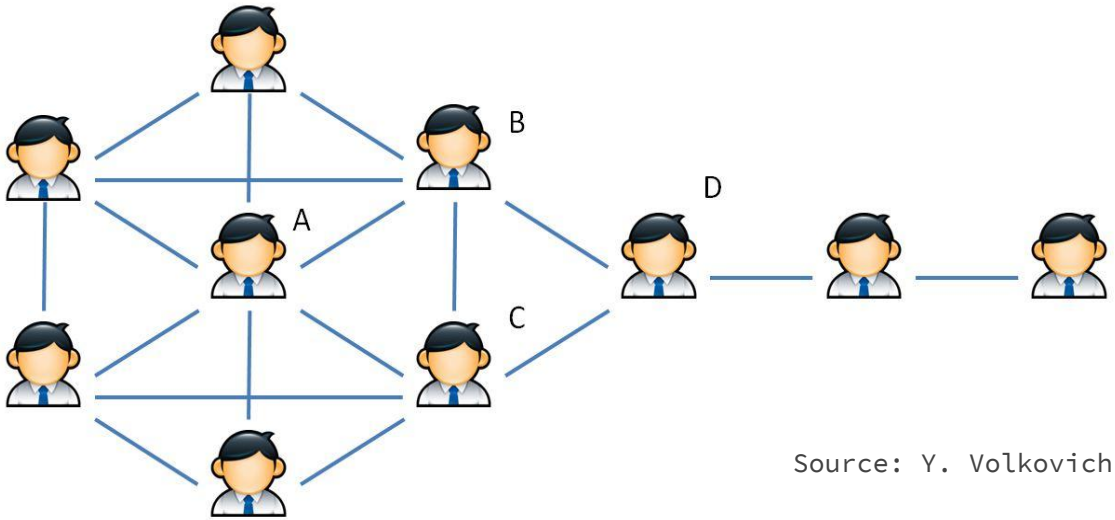


Source: Wikipedia

Comparison

Central nodes

- Degree centrality:
- Closeness centrality:
- Betweenness centrality:

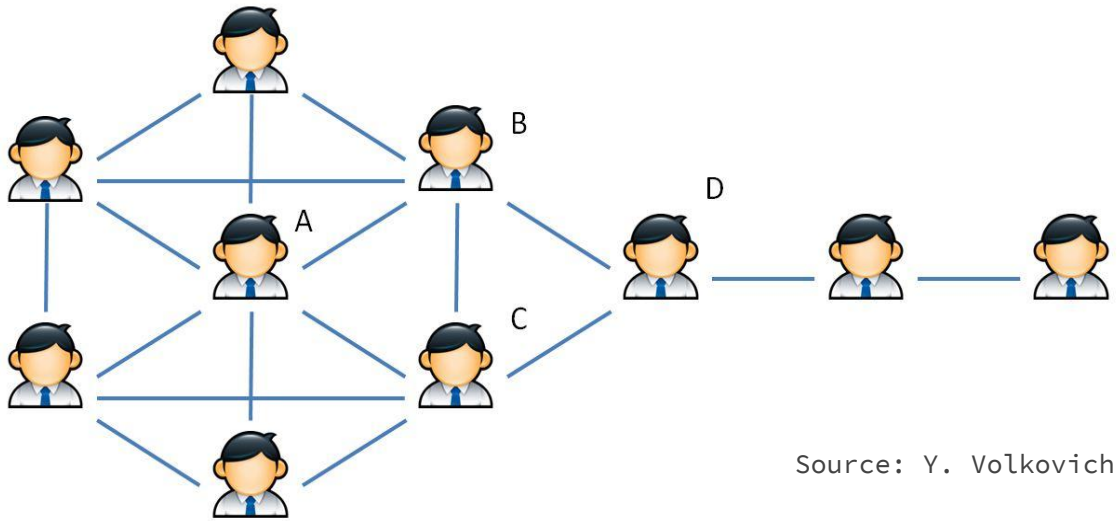


Source: Y. Volkovich

Comparison

Central nodes

- Degree centrality: USER A
- Closeness centrality: USERS B,C
- Betweenness centrality: USER D

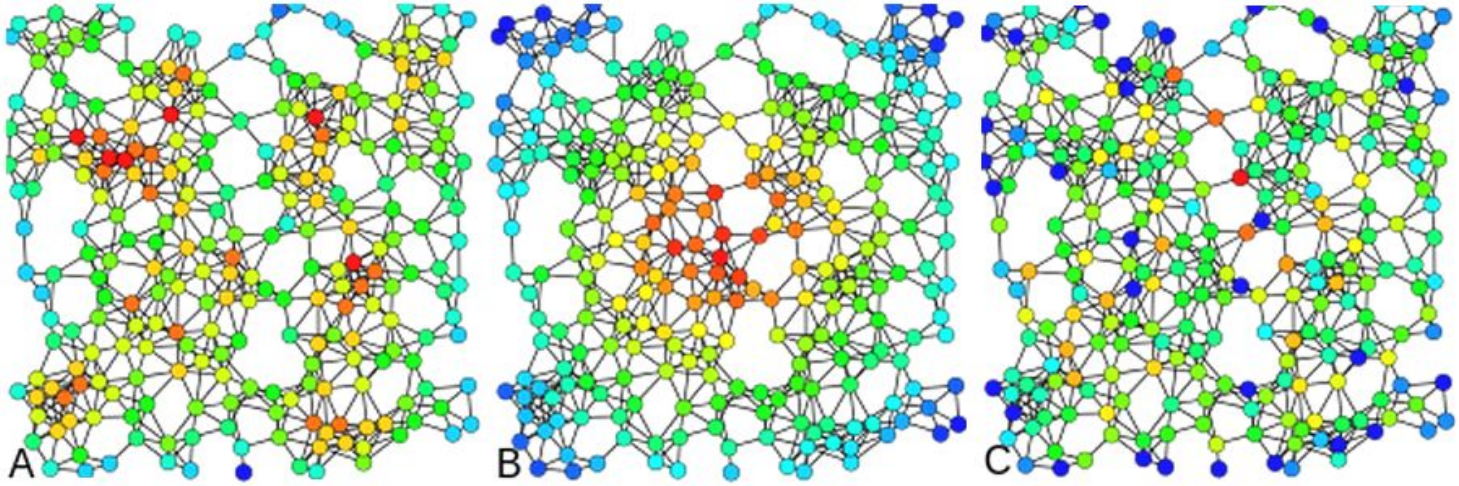


Source: Y. Volkovich

Comparison

Central nodes

- Degree centrality: Graph A
- Closeness centrality: Graph B
- Betweenness centrality: Graph C



Source: Wikipedia

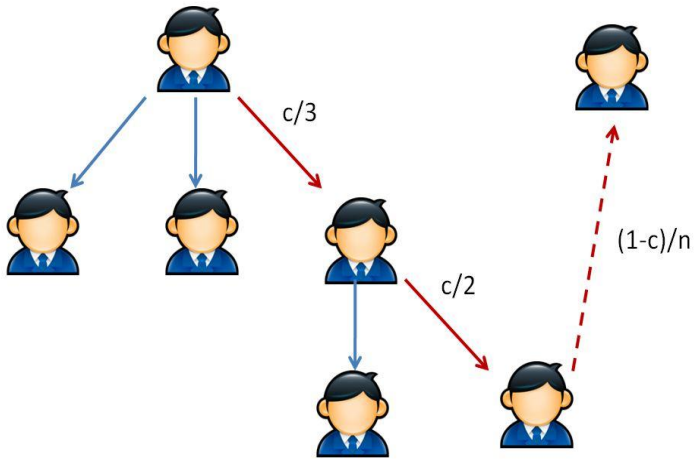
Pagerank [Brin 1998]

Motivation

- Google-defined popularity metrics for web ranking
- A random walk is simulated where at each step a jump is made to a random node with a probability (1-c)

$$PR^*(i) = c \sum_{j \to i} \frac{1}{d_j^*} PR^*(j) + \frac{1-c}{N^*},$$

- $PR^*(i)$ PageRank
- d^*j Outdegree node j
- N^* Number of nodes



Source: Y. Volkovich

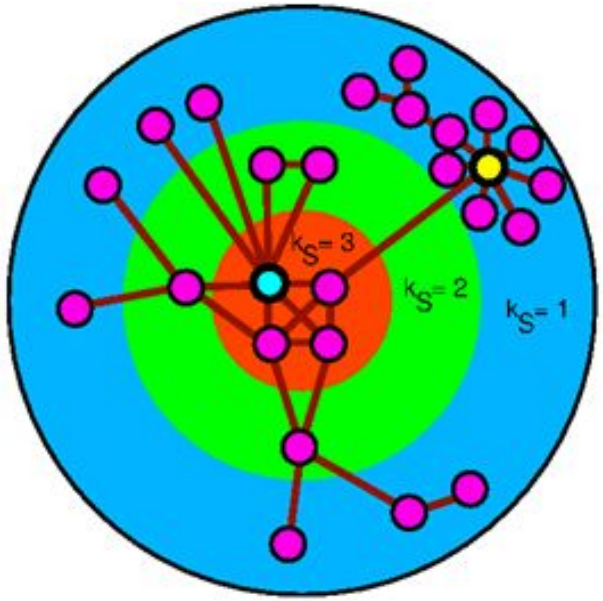
K-core decomposition

Motivation

- Detect nodes that are globally efficient to infect other nodes
- Discard local hubs (with many isolated contacts)

Method

- Larger sub-graph where each node has at least k direct neighbours



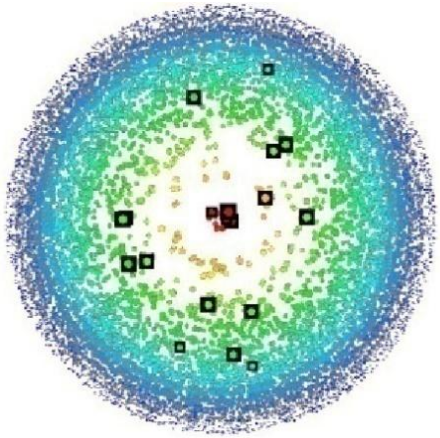
Source: Wikipedia

K-core decomposition

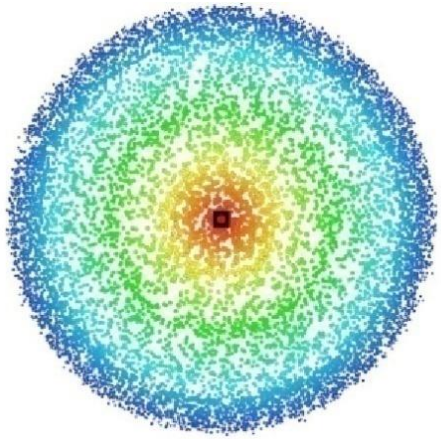
K-index

- Maximum k-shell that a node belongs to contacts)

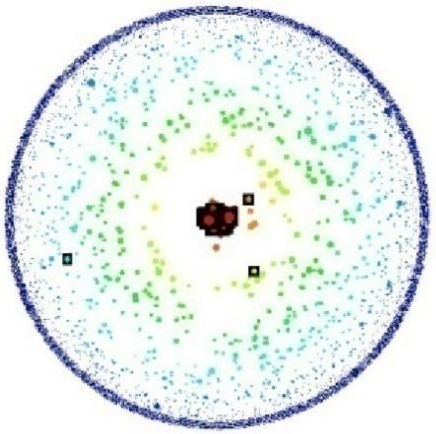
Examples



Physicists



Actors



Mails

Conclusions

Duncan Watts. Challenging the influential hypothesis

- The detection of influencers always happens a posteriori
- Influence might be based non-repeatable anecdotal data
- Influence might occur by accident
- Anyone can be influential
- Someone can be influential on one issue but not on another
- Influence exploitation probably leads to loss of influence

In short...

- There are nodes with more potential for influence than others
- But there's no guarantee they will exploit their capabilities

References

- Blondel, V. D., Guillaume, J. L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. Journal of Statistical Mechanics: Theory and Experiment, 2008(10), P10008.
- Brin, S., & Page, L. (1998). The anatomy of a large-scale hypertextual web search engine. Computer networks and ISDN systems, 30(1-7), 107-117.
- Girvan, M., & Newman, M. E. (2002). Community structure in social and biological networks. Proceedings of the national academy of sciences, 99(12), 7821-7826.
- Newman, M. E. (2004). Fast algorithm for detecting community structure in networks. Physical review E, 69(6), 066133.

Case-study: Barcelona en Comú

Aragón, P., Gallego, H., Laniado, D., Volkovich, Y., & Kaltenbrunner, A. (2017). Online network organization of Barcelona en Comú, an emergent movement-party. Computational Social Networks. doi:10.1186/s40649-017-0044-4

https://computationalsocialnetworks.springeropen.com/articles/10.1186/s40649-017-0044-4

Movement organization

Networked social movement: Networked in multiple forms (multimodal, on/offline, across platforms) without a central node, and with a decentered structure

(Castells, 2013)

Change from logic of collective action to a logic of **connective action**

(Bennett & Segerberg, 2013)

15M

#SPANISH REVOLUTION

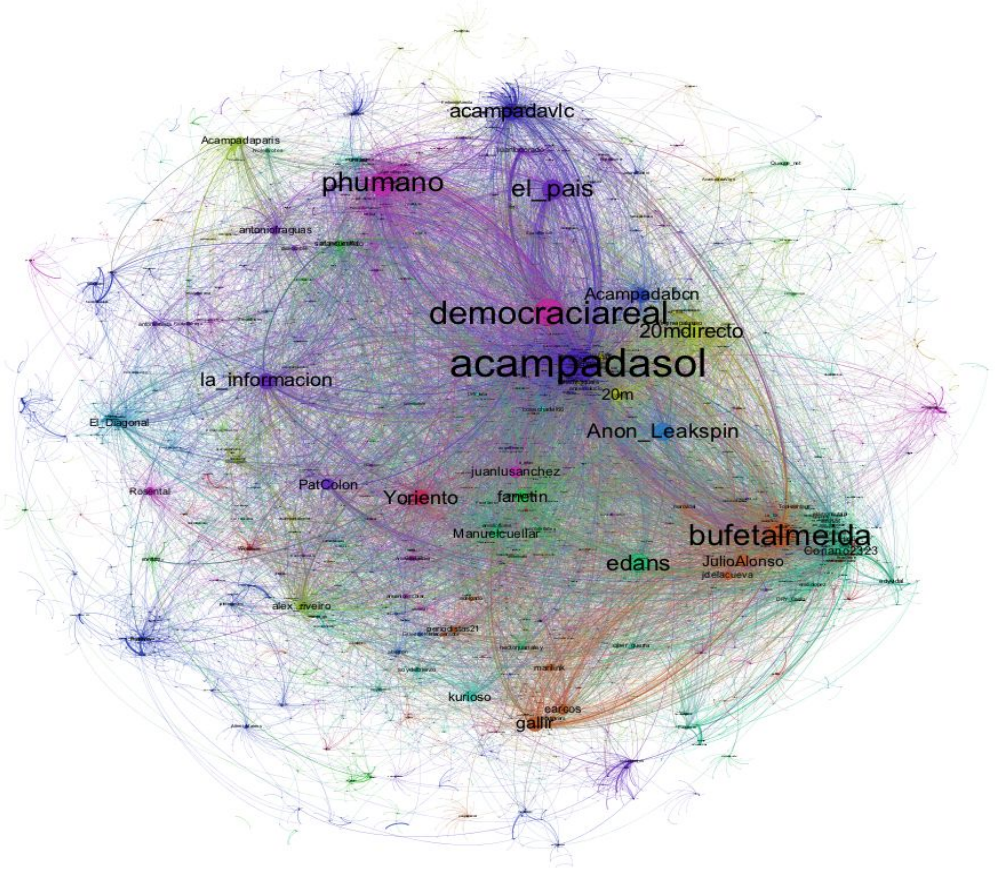
Movement networks

“Decentralized structure, based on coalitions of smaller organizations”

(González-Bailón et al, 2011)

“Decentralized organization, without leaders or stable representatives”

(Aragón et al, 2015)



RT network of the 15M movement
May 15-22, 2011 (Aragón et al, 2015)

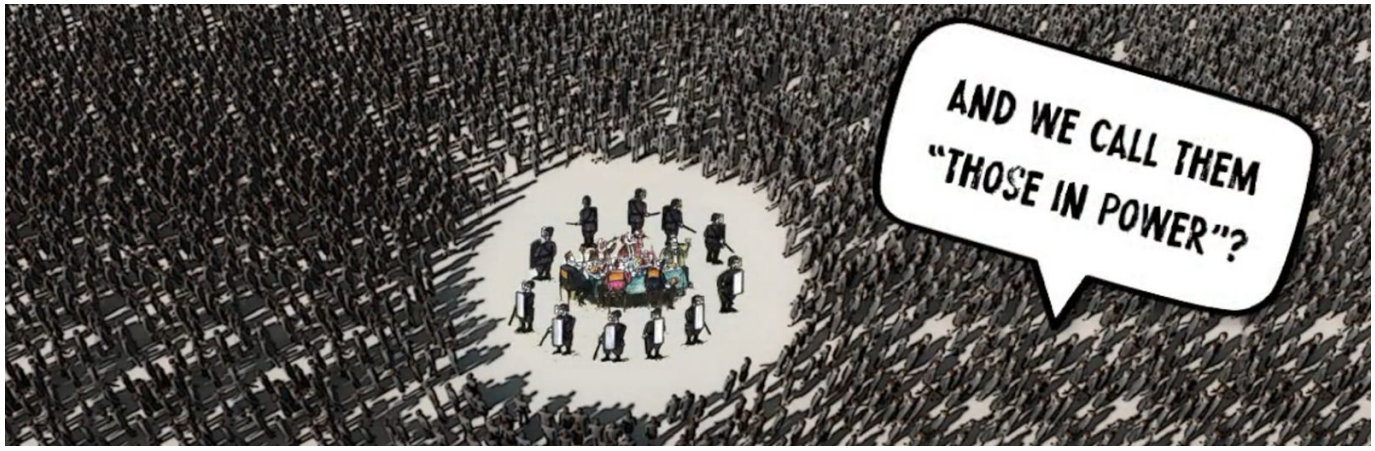
Party organization

Iron Law of Oligarchy: Political parties, like any complex organization, self-generate an elite ("Who says organization, says oligarchy")

(Michels, 1915)

Elite theory: Small minorities (elites) hold the most power in political processes

(Pareto et al, 1935; Mosca, 1939; Mills 1999)

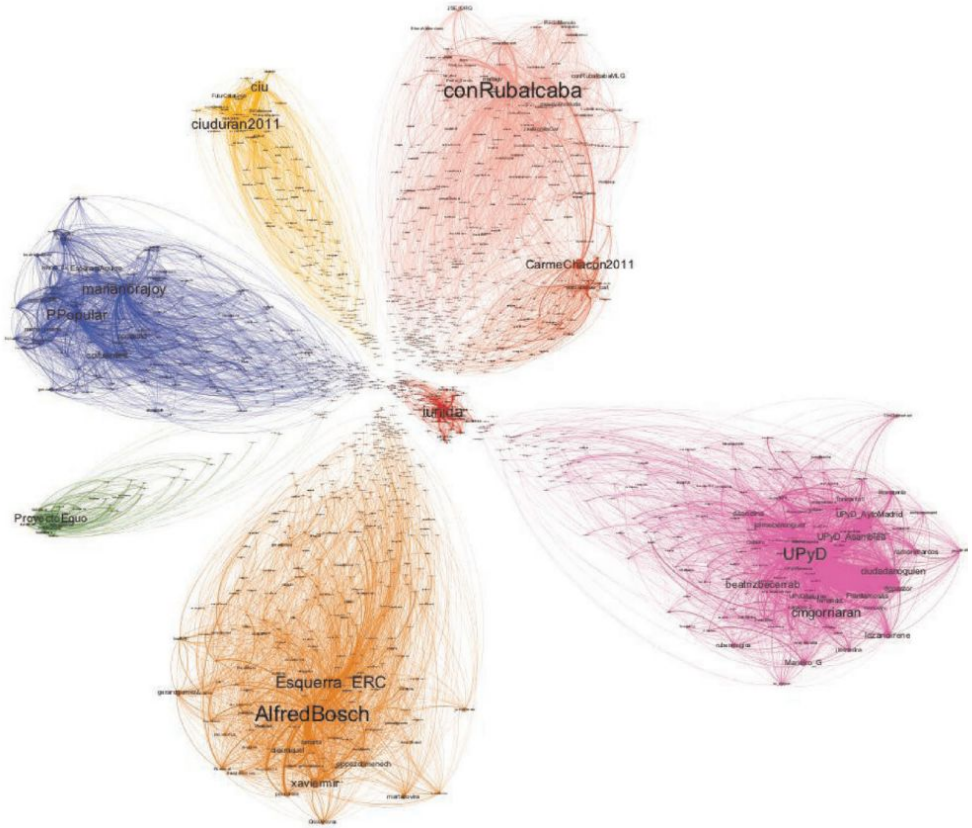


Party networks

The Twitter party networks in the 2011 Spanish election presented:

- Isolated clusters for each party.
- Minor and new parties were more clustered and better connected.
- Every party cluster was **strongly centralized** around candidate and/or party profiles.

(Aragón et al, 2013)



RT network of political parties in the 2011 Spanish election (Aragón et al, 2013)

Spanish Local Elections 2015

Grassroots parties emerged from the 15M movement:

- Barcelona en Comú
- Ahora Madrid
- Zaragoza en Común
- Marea Atlántica
- Compostela Aberta
- Por Cádiz Sí se puede
- Guanyem Badalona en Comú

Research question

Assuming that:

- **Barcelona en Comú** emerged from the 15M movement
- the **15M movement** followed a decentralized structure

Has Barcelona en Comú…

- preserved a decentralized structure?
- adopted a conventional centralized organization?

Research question

"Political parties share some interesting patterns of behavior, but also exhibit some unique and interesting idiosyncrasies" (e.g. tagging practice of politicians)

(Lietz et al, 2015)

Political Party / Coalition	Party account(s)	Candidate account
CiU - Convergència i Unió	@CDCBarcelona @uniobcn	@xaviertrias
PSC - Partit dels Socialistes de Catalunya	@pscbarcelona	@jaumecollboni
PP - Partit Popular de Catalunya	@PPBarcelona_	@albertofdezxbcn
BeC - Barcelona en Comú	@bcnencomu @icveuiaBCN @Podem_BCN @Equobcn @pconstituentBCN	@AdaColau
ERC - Esquerra Republicana de Catalunya	@ERCbcn	@AlfredBosch
Cs - Ciudadanos	@Cs_BCNA	@CarinaMejias
CUP - Capgirem Barcelona	@CapgiremBCN @CUPBarcelona	@MJLecha

Sampling criteria based on candidate and party accounts:

373 818 RTs

RT network

- 6 492 nodes
- 16 775 edges

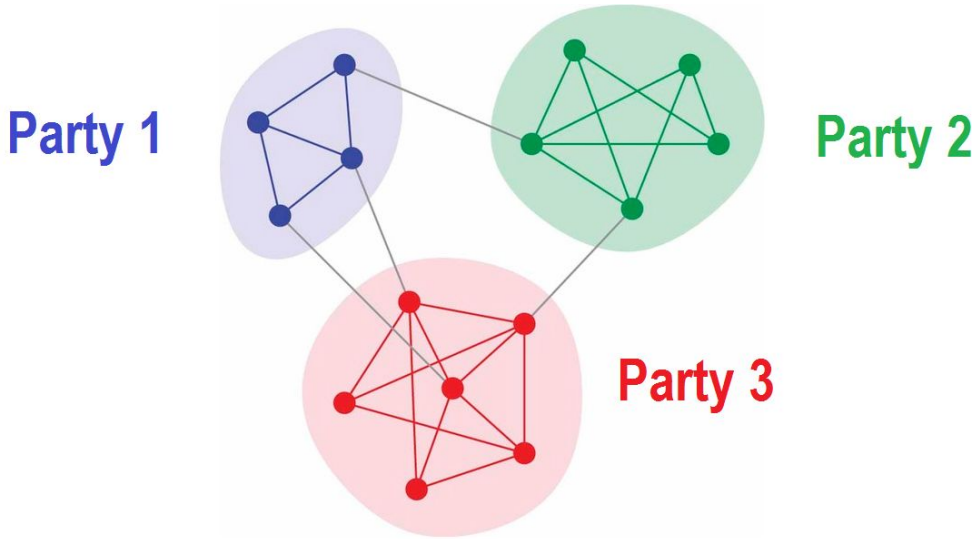
Methodology

Community detection

Identify the organization of nodes in clusters: political party networks.

Cluster characterization

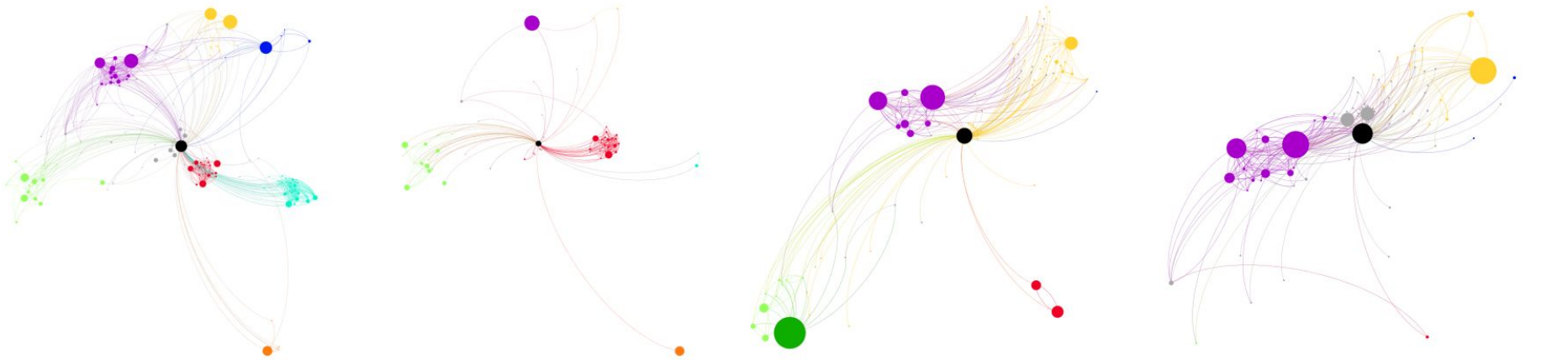
Characterize the topology of the intra-network of each cluster.



Community detection

First result with the Louvain Method (Blondel et al, 2008):

- Eight major clusters (seven parties)
- Every cluster contains some media accounts: media build weak ties
- Analysis of the ego-network of relevant media accounts:
  - Public TV account retweeted by users from every cluster
  - Private media mostly retweeted by users from like-minded parties



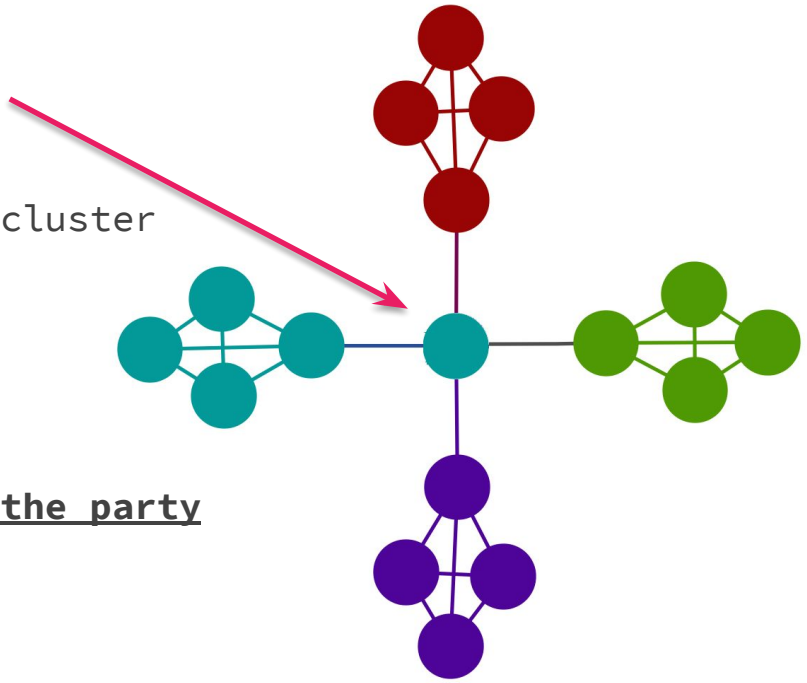
(a) @btvnoticies (public media) (b) @elpaiscat (private progressive media) (c) @arapolitica (private Catalan nationalist media) (d) @naciodigital (private Catalan nationalist media)

Community detection

Every cluster contains some media accounts

Each execution produces different results:

Some media do not always belong to the same cluster



We want the real intra-network structure of the party

Confident version of the Louvain Method

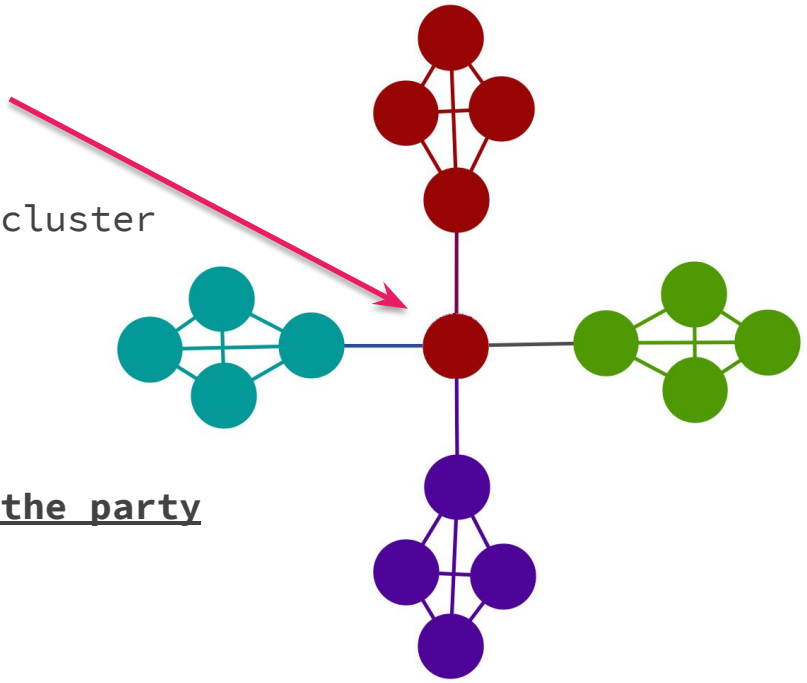
- Multiple executions (N=100)
- Identify each cluster by the most relevant user (PageRank)
- Just consider nodes that appear in the same cluster many times (1-ε=0.95)

Community detection

Every cluster contains some media accounts

Each execution produces different results:

Some media do not always belong to the same cluster



We want the real intra-network structure of the party

Confident version of the Louvain Method

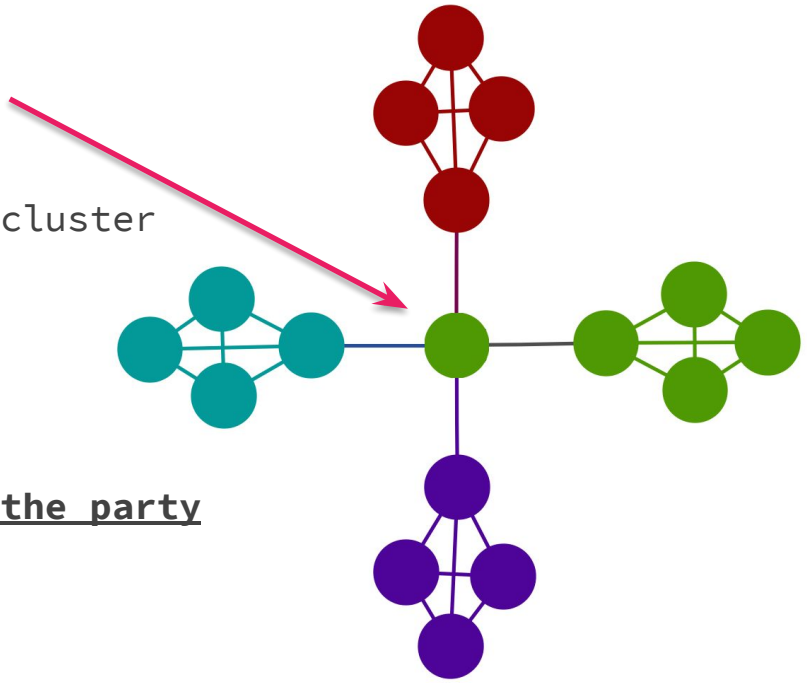
- Multiple executions (N=100)
- Identify each cluster by the most relevant user (PageRank)
- Just consider nodes that appear in the same cluster many times (1-ε=0.95)

Community detection

Every cluster contains some media accounts

Each execution produces different results:

Some media do not always belong to the same cluster



We want the real intra-network structure of the party

Confident version of the Louvain Method

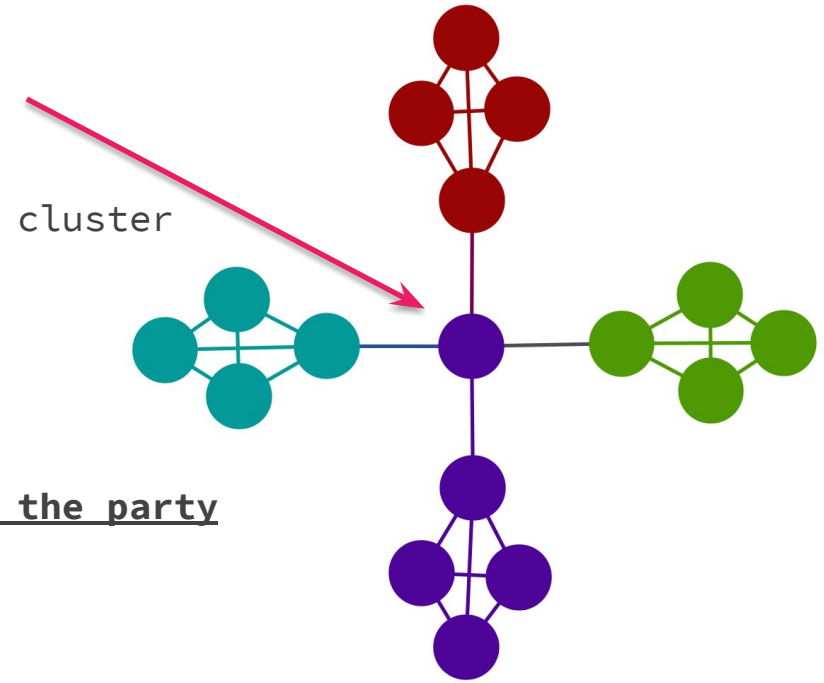
- Multiple executions (N=100)
- Identify each cluster by the most relevant user (PageRank)
- Just consider nodes that appear in the same cluster many times (1-ε=0.95)

Community detection

Every cluster contains some media accounts

Each execution produces different results:

Some media do not always belong to the same cluster



We want the real intra-network structure of the party

Confident version of the Louvain Method

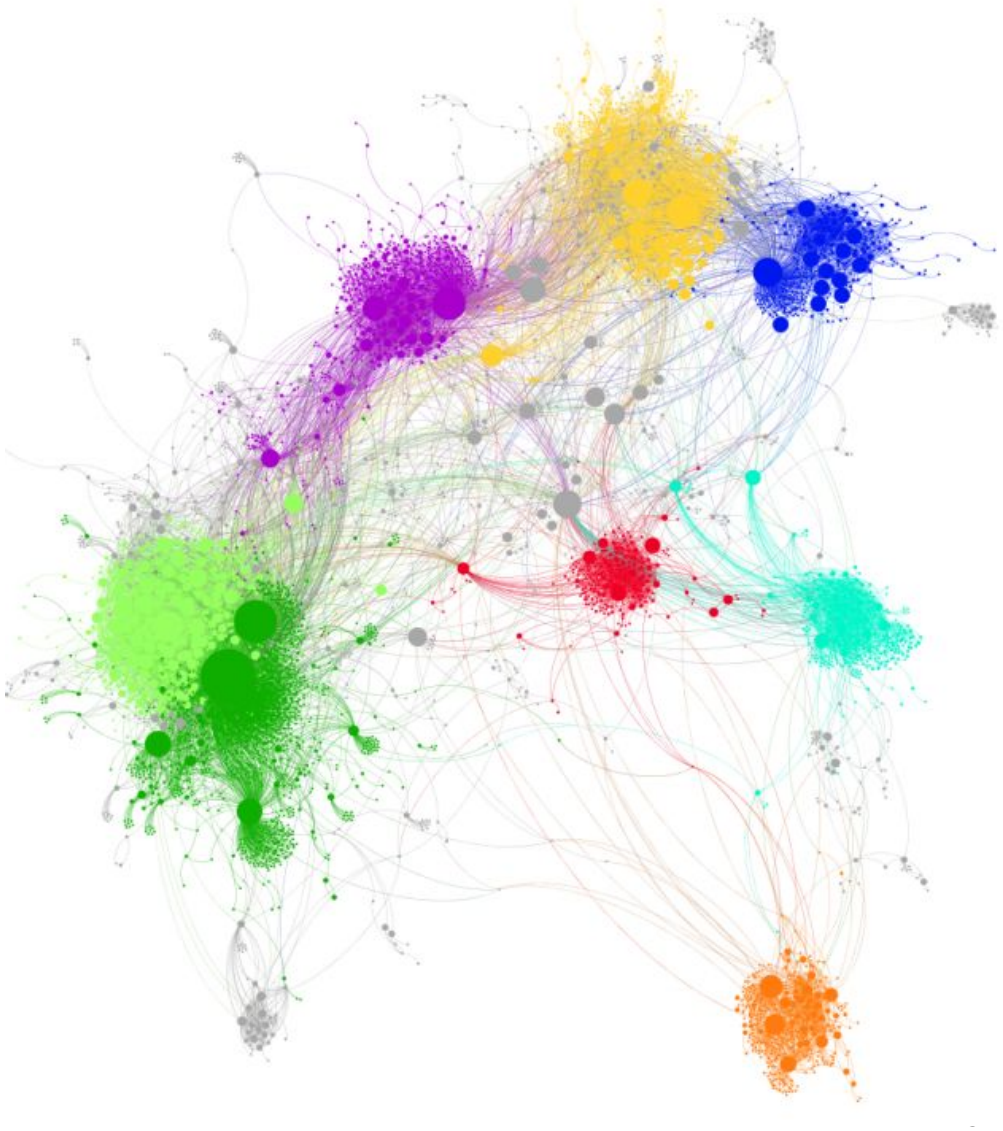
- Multiple executions (N=100)
- Identify each cluster by the most relevant user (PageRank)
- Just consider nodes that appear in the same cluster many times (1-ε=0.95)

Community detection

Results with the **extended version of the Louvain Method**

Constant presence of eight major clusters (seven parties) along the 100 executions:

- Most media accounts do not appear in major clusters
- Two clusters for Barcelona en Comú



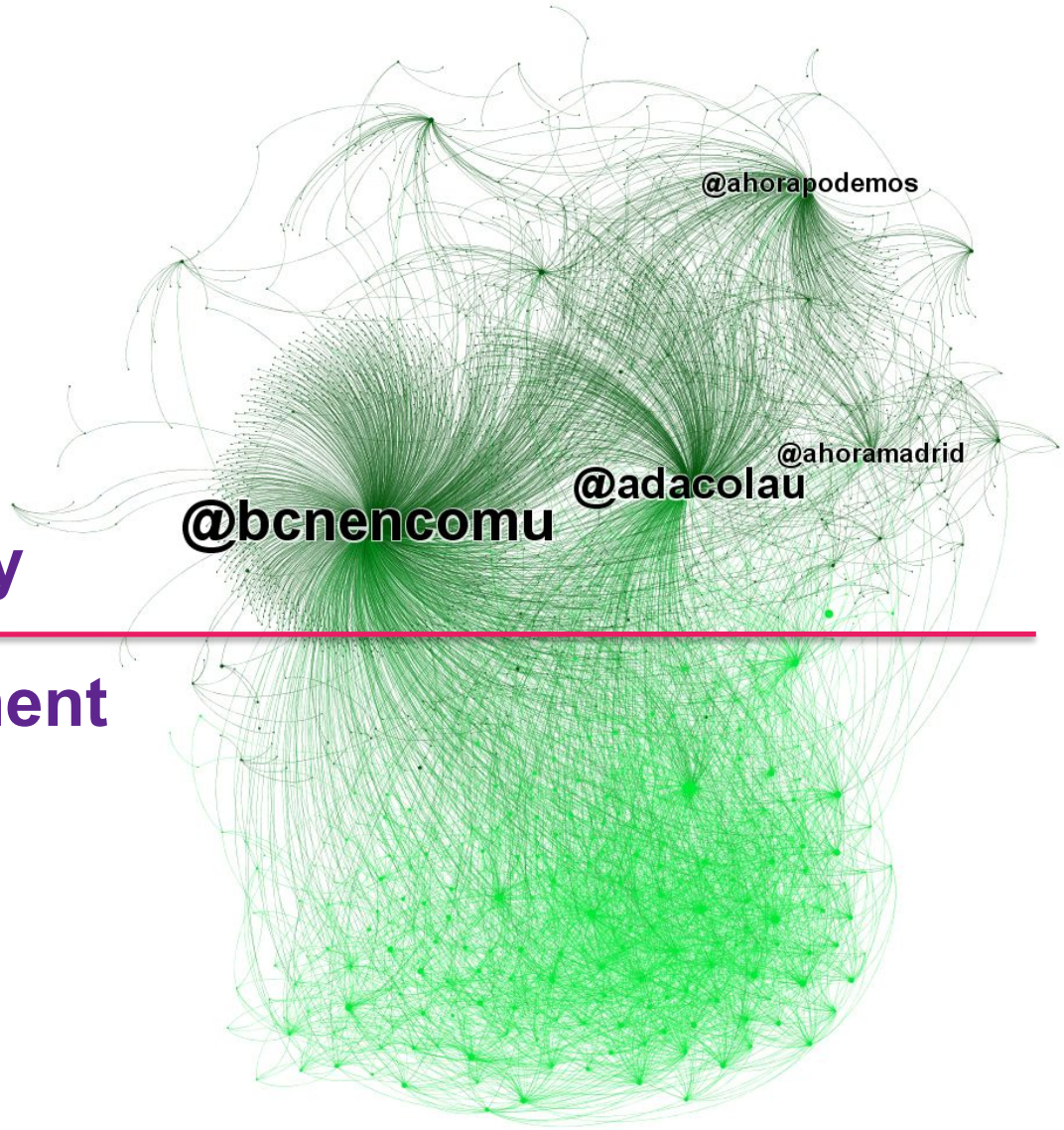
Community detection

Results with the **extended version of the Louvain Method**

Constant presence of eight major clusters (seven parties) along the 100 executions:

- Most media accounts do not appear in major clusters
- Two clusters for Barcelona en Comú

Community detection



Party

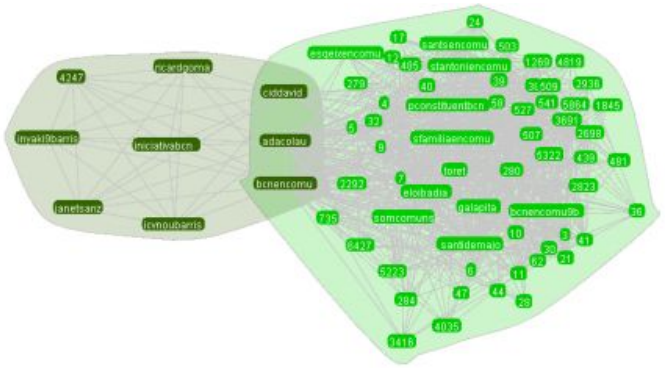
Movement

Community detection

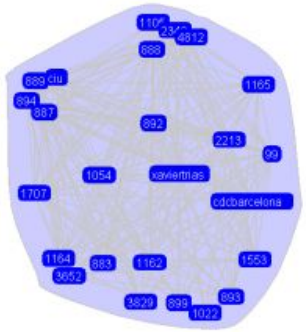
Clique Percolation Model

Similar results but…

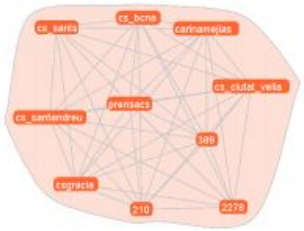
- CPM is O(exp(n))
  (NP-complete problem)
- CPM is not sensitive to different sizes and structures of parties
- K-cliques are only the core of the structure of party networks
  (the periphery is relevant)



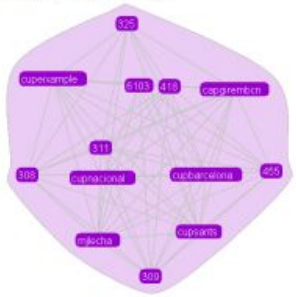
(a) BeC-p (left) & BeC-p (right) (k=9).



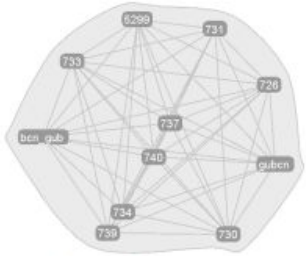
(b) CiU ($k = 9$).



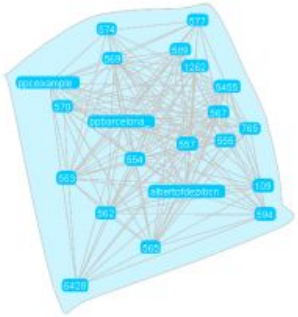
(c) Cs ($k = 9$).



(d) CUP ($k = 9$).

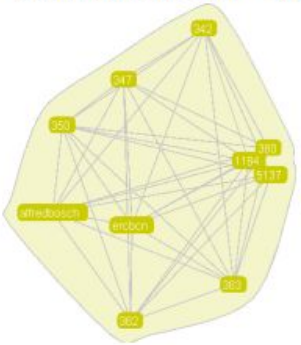


(e) Guardia Urbana ($k = 9$).



(f) PP (k=9).

(g) PSC ($k = 8$).



(h) ERC ($k = 7$).

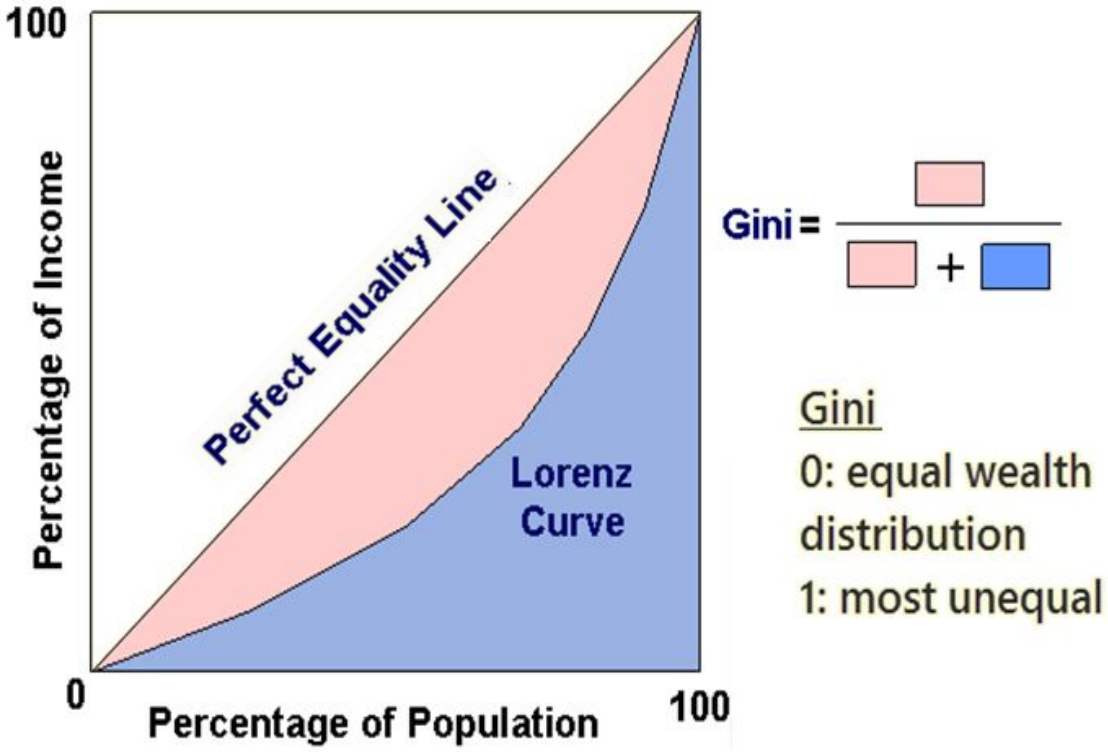
Cluster characterization

Inspired by the social dimensions of García et al. (2015):

- Hierarchical structure
  In-degree centralization → Gini coefficient of the in-degree distribution
- Small world phenomenon (f.k.a. information efficiency)
  Avg. path length + Clustering coefficient
- Coreness (f.k.a. social resilience)
  Maximal k-core → Distribution of k-indices

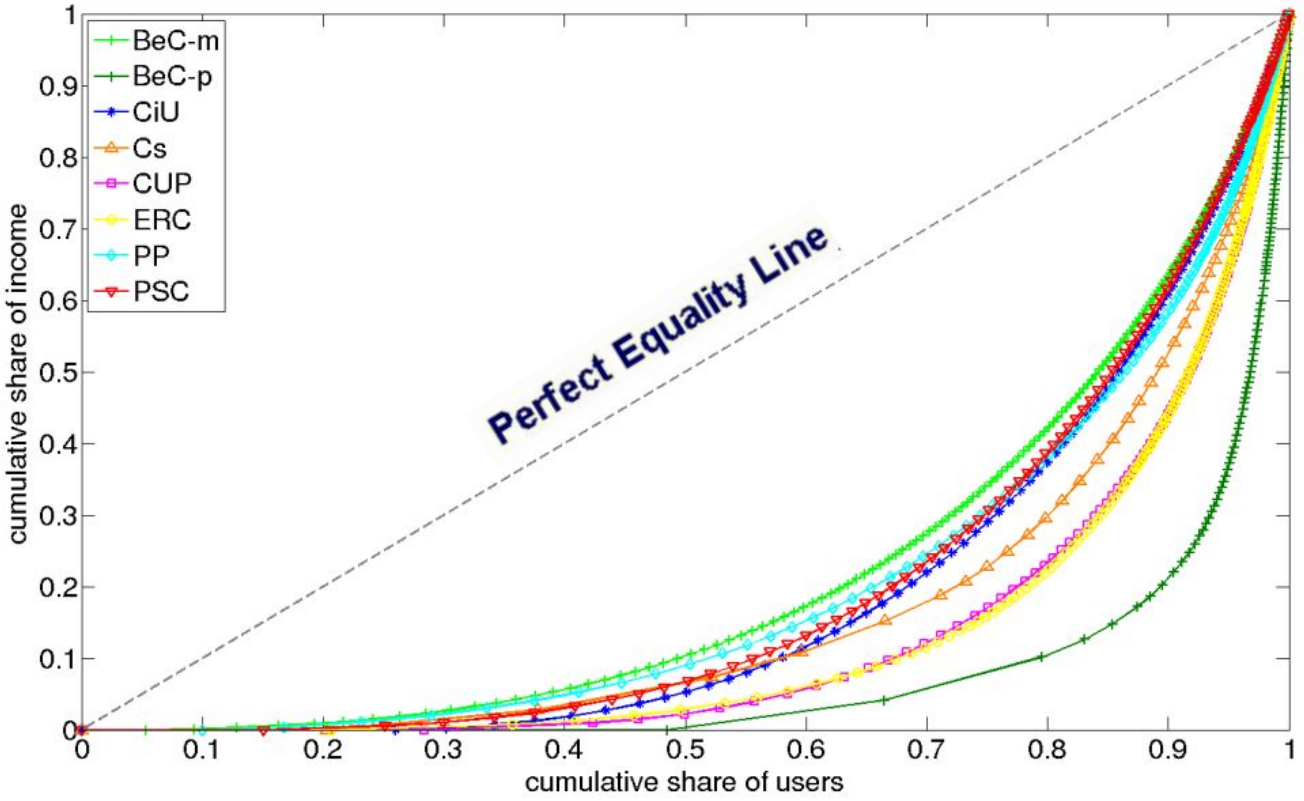
Hierarchical structure

Gini coefficient of the in-degree distribution



Hierarchical structure

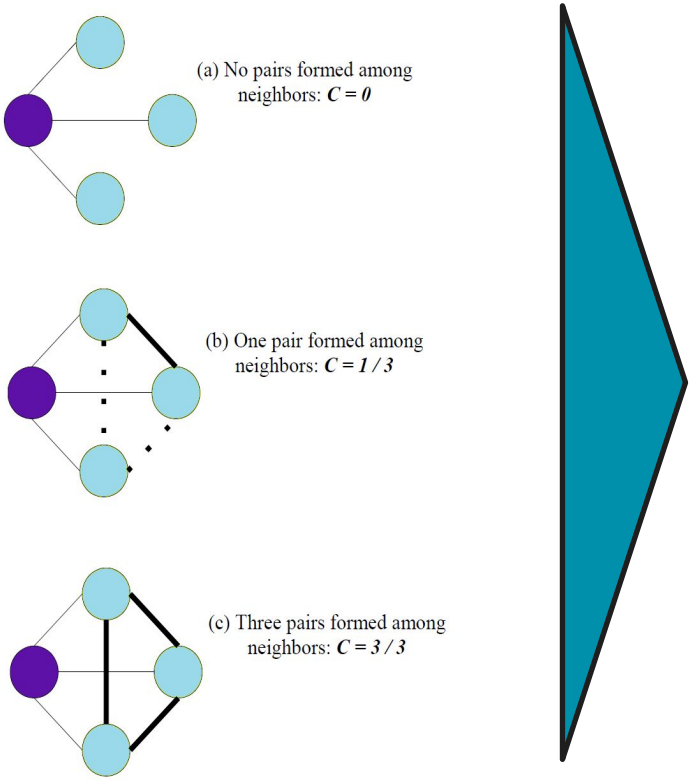
Gini coefficient of the in-degree distribution

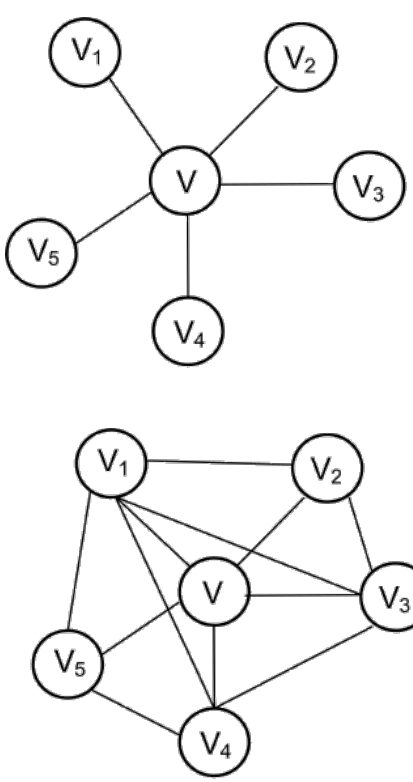


Cluster	G_{in}
BeC-p	0.995
Cs	0.964
ERC	0.954
CUP	0.953
CiU	0.893
PP	0.876
PSC	0.818
BeC-m	0.811

Small world phenomenon

Clustering coefficient



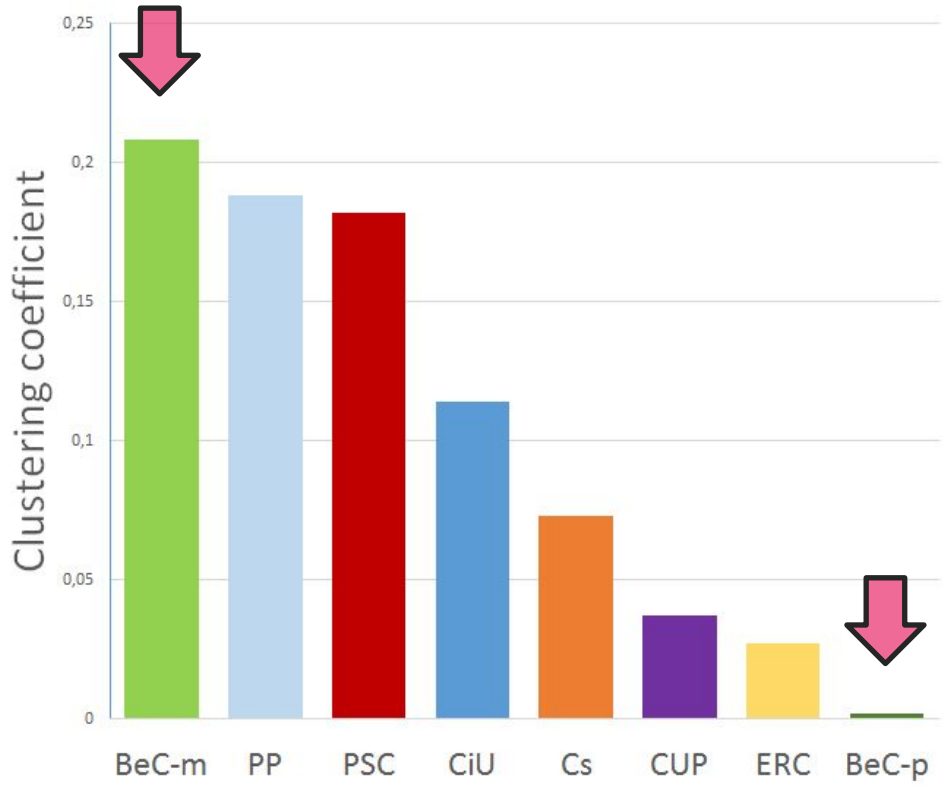


LOW

HIGH

Small world phenomenon

Clustering coefficient

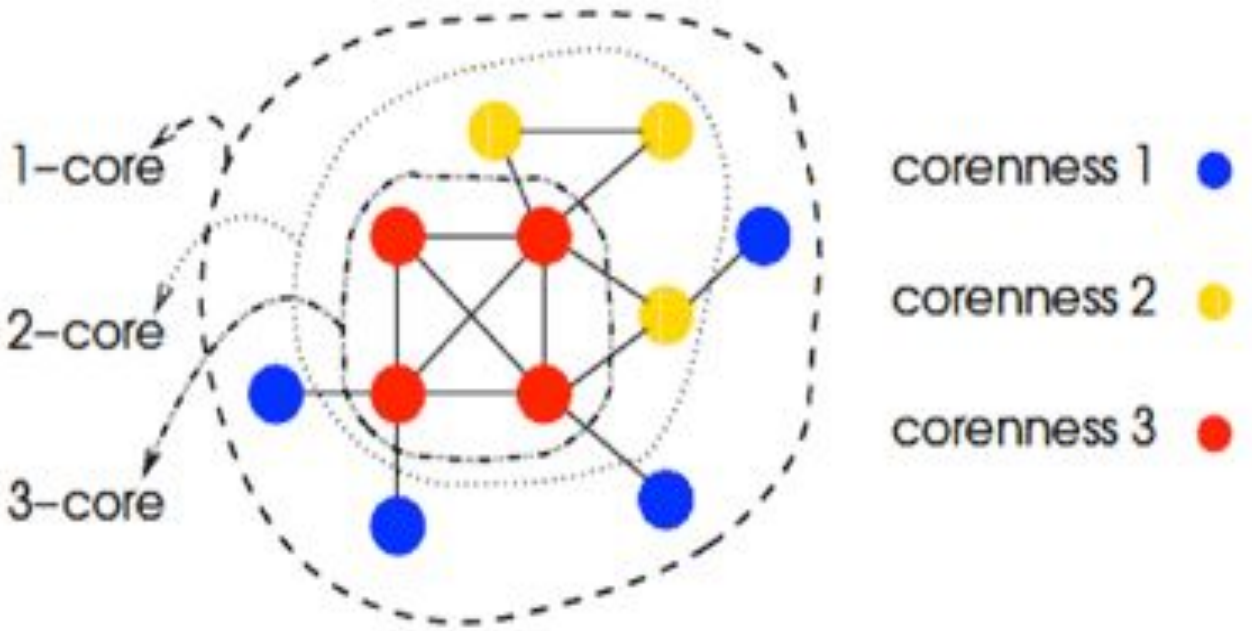


Number of nodes (N) and edges (E), clustering coefficient (Cl) and average path length (l) of each cluster.

Cluster	N	E	Cl	l
BeC-m	427	2 431	0.208	3.35
PP	301	1 163	0.188	2.73
PSC	211	810	0.182	2.29
CiU	337	1 003	0.114	4.66
Cs	352	832	0.073	2.57
CUP	635	1 422	0.037	2.57
ERC	866	1 899	0.027	5.43
BeC-p	1 844	2 427	0.002	2.48

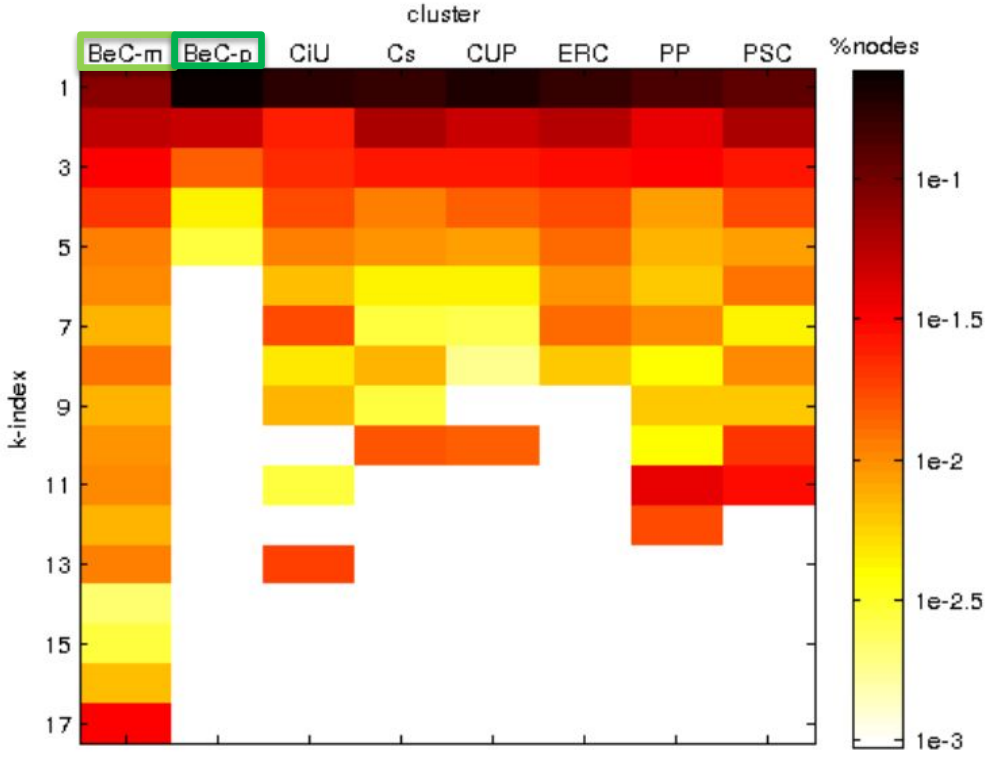
Coreness

K-core decomposition



Coreness

K-core decomposition



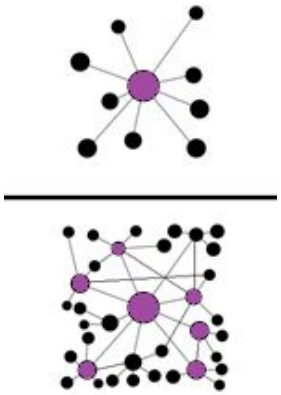
Maximal and average k-index (standard deviation in parentheses) for the intra-network of each cluster.

cluster	k_{max}	k_{avg}
BeC-m	17	5.90 (5.46)
PP	12	4.02 (3.99)
PSC	11	3.85 (3.55)
CiU	13	3.10 (3.44)
ERC	8	2.25 (1.85)
Cs	10	2.42 (2.42)
CUP	10	2.19 (2.22)
BeC-p	5	1.33 (0.71)

Conclusions

For Barcelona en Comú, two paradigms co-exist:

- A **centralized** and low resilient party cluster
- A **decentralized** and resilient movement cluster



Polarized scenario likes previous studies of election campaigns on Twitter

- Data preparation process accentuated the polarization effect

Media accounts build weak ties between clusters

- Public media became more plural than private media