Social and Technological Network Analysis

Lecture 4: Modularity and Overlapping Communities

Dr. Cecilia Mascolo

UNIVERSITY OF CAMBRIDGE

In This Lecture

- We will describe the concept of Modularity and Modularity Optimization.
- We will describe methods for overlapping community detection.

How do we know when to stop?

- We need to find a way to decide how many clusters there should be.
- How?

Modularity

- Perhaps a good measure of when to stop is when for each community the "cohesion" within the community is higher than outside…
- ***Q= (edges inside the community)- (expected number of edges inside the community)***

- Expected number of edges between nodes a and b:

$$\frac{k_a k_b}{2m}$$

m is the number of edges of the graph

Modularity (2)

- Number of edges **inside** a community:

$$\frac{1}{2}\sum_{a,b} A_{a,b}\delta(c_a, c_b)$$

- Where:
- $A_{a,b}$ is 1 if there is an edge a->b,
- $\delta(c_a , c_b)$ is the Kronecker Delta (1 if c_a is equal to c_b)

UNIVERSITY OF CAMBRIDGE

Modularity (3)

$$Q1 = \frac{1}{2}\sum_{a,b} A_{a,b}\delta(c_a, c_b) - \frac{1}{2}\sum_{a,b}\frac{k_a k_b}{2m}\delta(c_a, c_b)$$

$$Q1 = \frac{1}{2}\sum_{a,b}(A_{a,b} - \frac{k_a k_b}{2m})\delta(c_a, c_b)$$

$$Q = \frac{1}{2m}\sum_{a,b}(A_{a,b} - \frac{k_a k_b}{2m})\delta(c_a, c_b)$$

Fraction of edges over all edges m

Modularity (4)

- Modularity ranges from -1 to 1.
  - It is positive if the number of edges inside the group are more than the expected number.
  - Variation from 0 indicate difference with random case.

- Modularity can be used at each round of the Girvan-Newmann algorithm to check if it is time to stop. However the complexity of this is O (m^2n).
- Why don't we try to just maximize modularity?

Modularity Optimization

- Finding the configuration with maximum modularity in a graph is an NP complete problem.
- However there are good approximation algorithms.

UNIVERSITY OF CAMBRIDGE

Fast Modularity

- Start with a network of n communities of 1 node
- Calculate ΔQ for all possible community pairs
- Merge the pair of the largest increase in Q
- Repeat (2)&(3) until one community remains
- Cross cut the dendrogram where Q is maximum.
- This runs in O((m + n)n).

- A further optimization runs in O(m d logn) [d depth of dendrogram].
- Most networks are sparse so m~n and d~log n

Example of Dendrogram

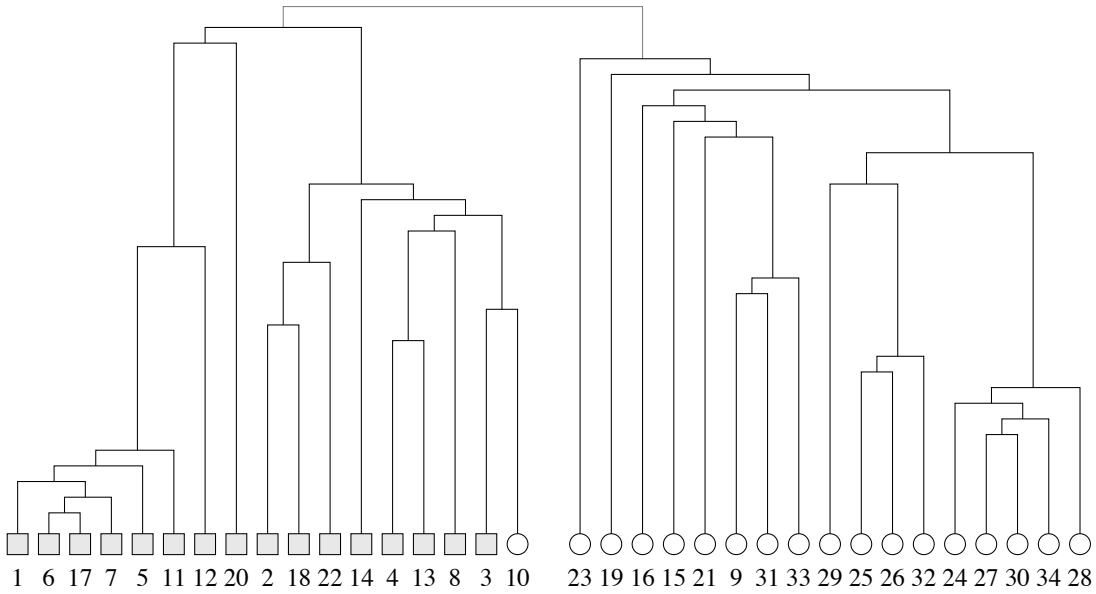


FIG. 2: Dendrogram of the communities found by our algorithm in the "karate club" network of Zachary [5, 17]. The shapes of the vertices represent the two groups into which the club split as the result of an internal dispute.

UNIVERSITY OF CAMBRIDGE

Application to Amazon Reccommedations

- Network of products.
- A link between product a and product b if b was frequently purchased by buyers of a.
- 200000 nodes and 2M edges.
- Max when 1684 communities
- Mean size of 243 products

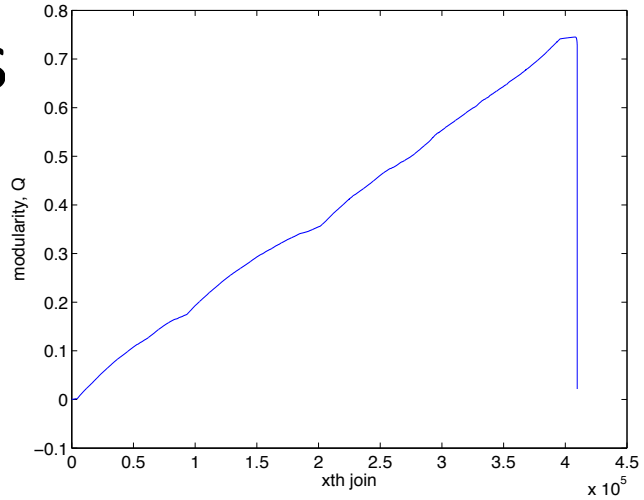


FIG. 1: The modularity Q over the course of the algorithm (the x axis shows the number of joins). Its maximum value is $Q = 0.745$, where the partition consists of 1684 communities.

UNIVERSITY OF CAMBRIDGE

Amazon: Top Communities (87% of nodes)

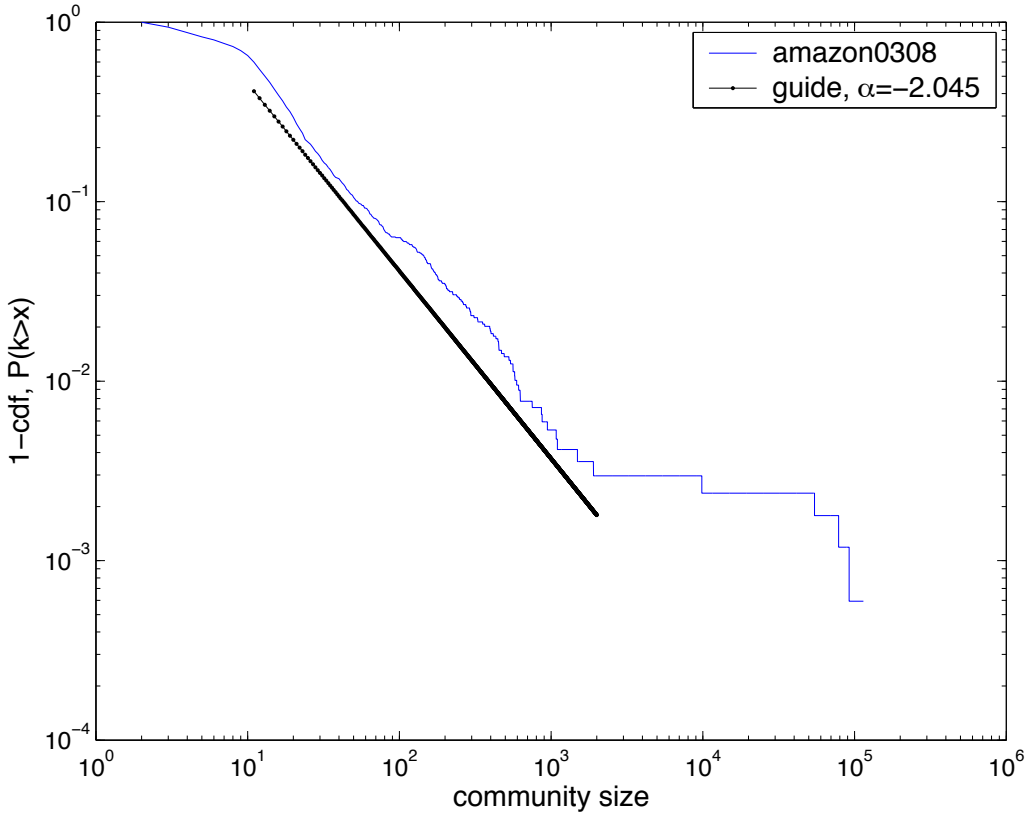
Rank	Size	Description
1	114538	General interest: politics; art/literature; general fiction; human nature; technical books; how things, people, computers, societies work, etc.
2	92276	The arts: videos, books, DVDs about the creative and performing arts
3	78661	Hobbies and interests I: self-help; self-education; popular science fiction, popular fantasy; leisure; etc.
4	54582	Hobbies and interests II: adventure books; video games/comics; some sports; some humor; some classic fiction; some western religious material; etc.
5	9872	classical music and related items
6	1904	children's videos, movies, music and books
7	1493	church/religious music; African-descent cultural books; homoerotic imagery
8	1101	pop horror; mystery/adventure fiction
9	1083	jazz; orchestral music; easy listening
10	947	engineering; practical fashion

TABLE I: The 10 largest communities in the Amazon.com network, which account for 87% of the vertices in the network.

UNIVERSITY OF CAMBRIDGE

Amazon: Community Size Distribution

- A power law distribution of community size
- (more on power laws in later lectures)



UNIVERSITY OF CAMBRIDGE

Limitations of Modularity

- Modularity is not a perfect measures
- It appears to depend on the number of links in the network (L).
- Problems for modules with a number of internal links of the order of √2L or smaller.

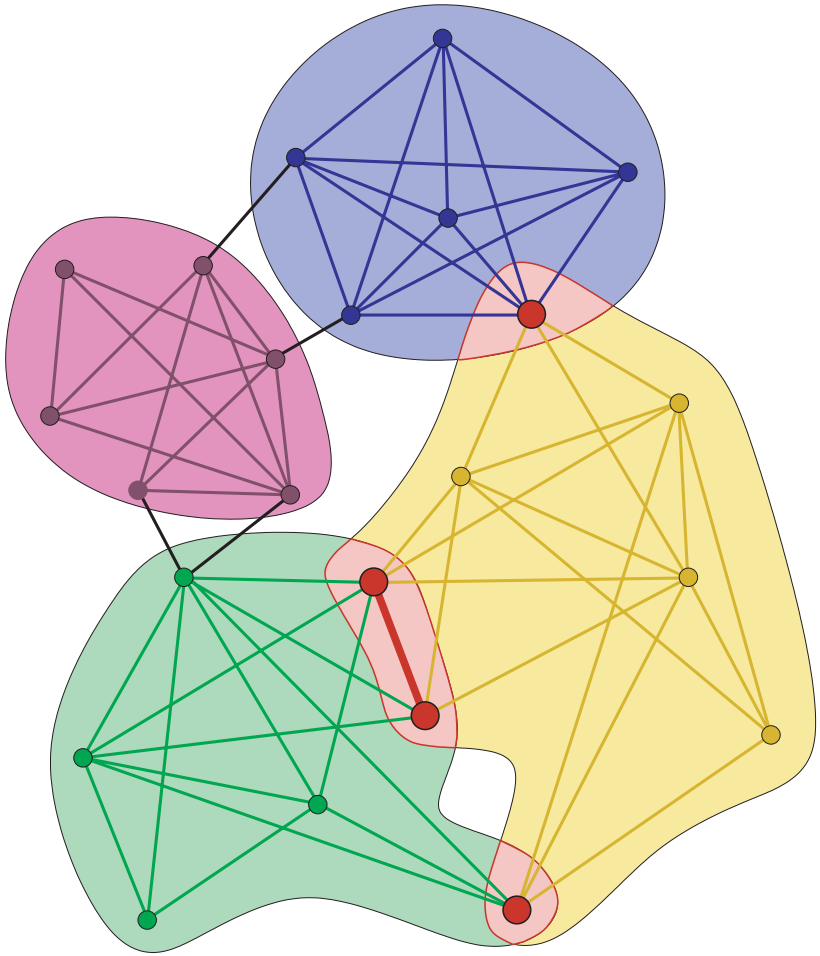
- Intuition: modularity depends on links of a community to the "outside", ie the rest of the network.

UNIVERSITY OF CAMBRIDGE

S. Fortunato, S. Barthelemy. Resolution limit in community detection. Proc. Natl. Acad. Sci., 2007.

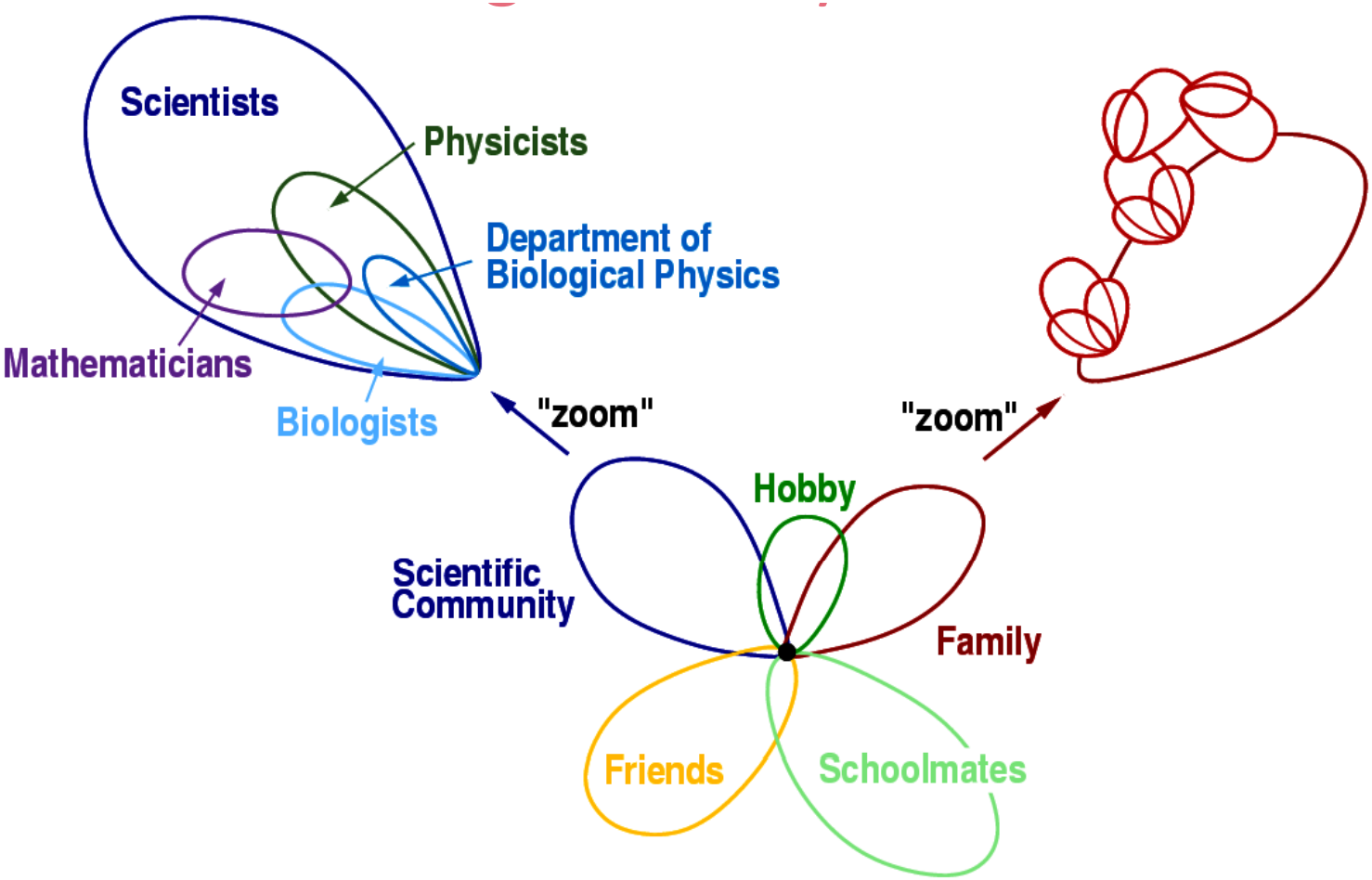
Overlapping Communities

- Community membership could overlap: a node could be part of more than 1 community.



UNIVERSITY OF CAMBRIDGE

Nodes can belong to more than 1 social circle!

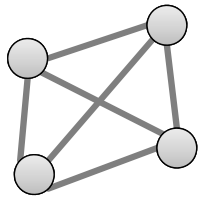


UNIVERSITY OF CAMBRIDGE

Clique Percolation Method: the idea (Palla 2005)

- Two nodes belong to the same community if they can be connected through adjacent *k*-cliques.
- A k-clique is a fully connected graph of k nodes.
- K-cliques are adjacent if they have k-1 overlapping nodes.
- K-clique community: nodes which can be reached through a sequence of adjacent k-cliques.

UNIVERSITY OF CAMBRIDGE



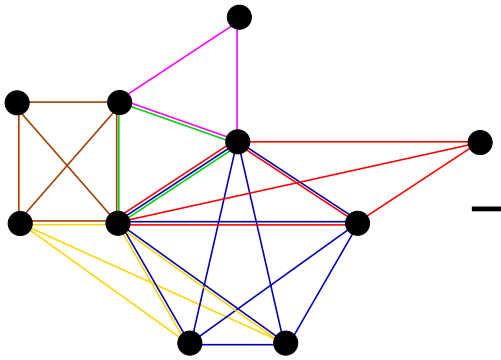
4-clique

Clique Percolation Method: The algorithm

- Find the maximal cliques
  - A maximal clique is a clique that cannot be extended by including one more adjacent vertex
  - This is complex but real networks are relatively sparse.
- Build clique overlap matrix
  - Each clique is a node
  - Connect two cliques if they overlap in at least *k-1* nodes
- Communities:
  - Connected components of the clique overlap matrix

Example

Maximal cliques



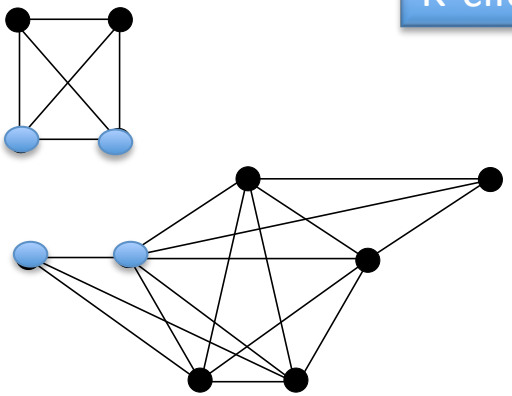
	5	3	2	1	3	1
	3	4	2	1	1	1
	2	2	3	2	1	2
	1	1	2	3	0	1
	3	1	1	0	4	2
	1	1	2	1	2	4

Overlap Matrix: elements are n. of overlapping nodes

k=4

Erase elements less than 4 on diagonal and less than 3 elsewhere

	1	1	0	0	1	0
	1	1	0	0	0	0
	0	0	0	0	0	0
	0	0	0	0	0	0
	1	0	0	0	1	0
	0	0	0	0	0	1



K-cliques

UNIVERSITY OF CAMBRIDGE

Application

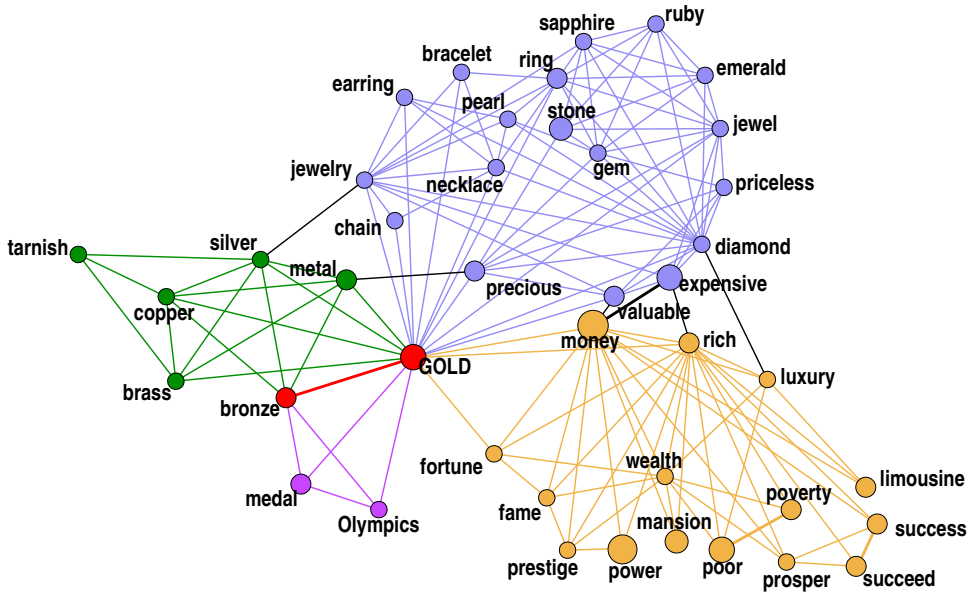
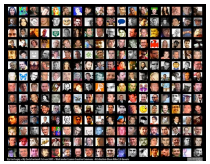
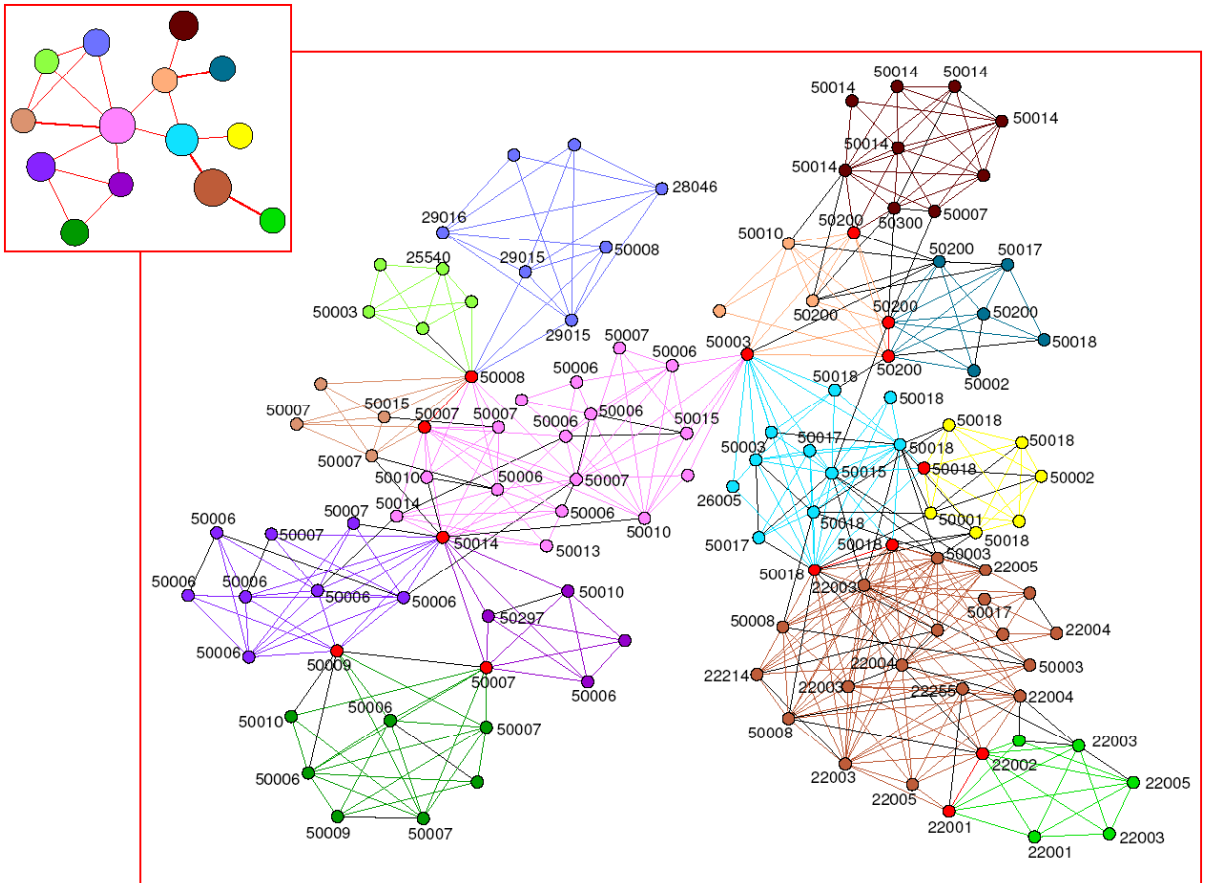


Figure 5: The k-clique communities of the word gold in the South Florida Free Association norm list for $w^* = 0.025$ and $k = 4$. The purple community is related to Olympic medals, the green one consists of metals, the blue one can be associated to jewels and finally the yellow community is related to welfare.

Application: Phone Call Network

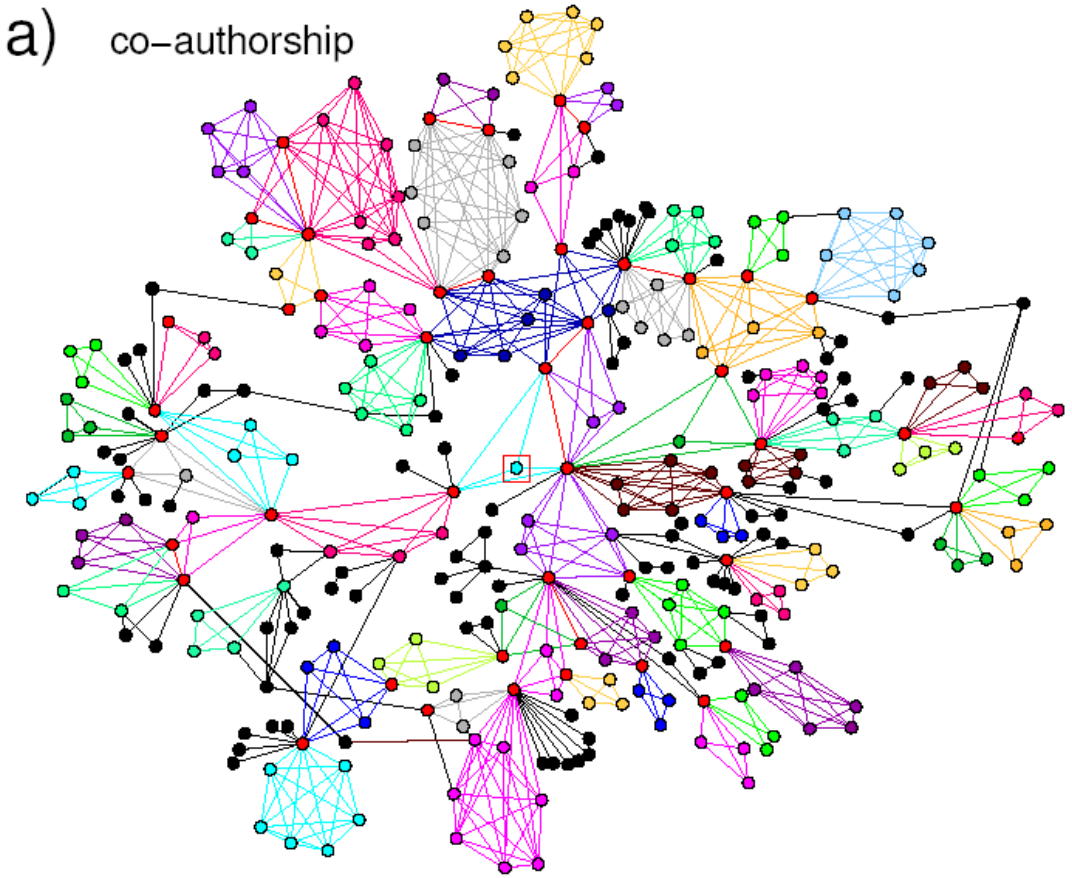




UNIVERSITY OF CAMBRIDGE

From Leskovec

Application: co-authorship network



UNIVERSITY OF CAMBRIDGE

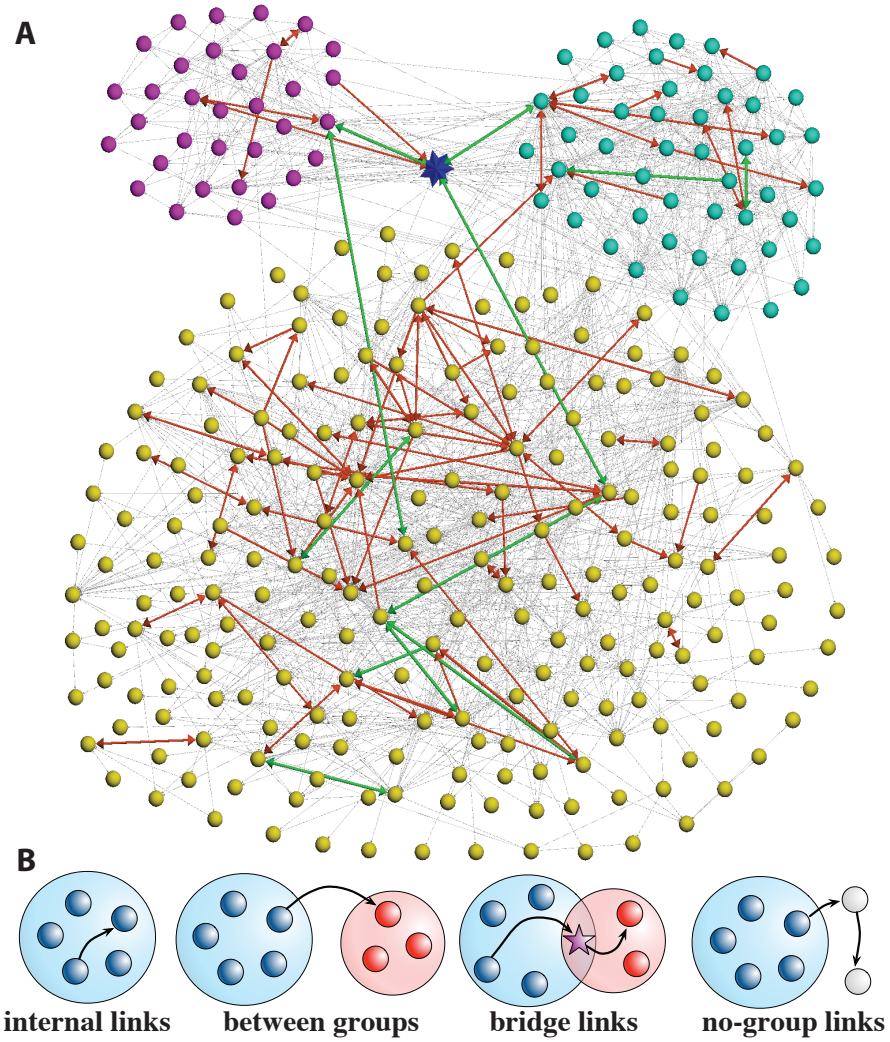
Community Detection and Weak Ties

- Twitter was analyzed trying to identify if the static network of followers gives information about the dynamics of retweeting and mentioning.
- Dataset: follower network (undirected), 2M users, and network of tweets, mention and retweets for 1 month.
- Some community detection methods are used to find clusters in the follower network.

UNIVERSITY OF CAMBRIDGE

Sample

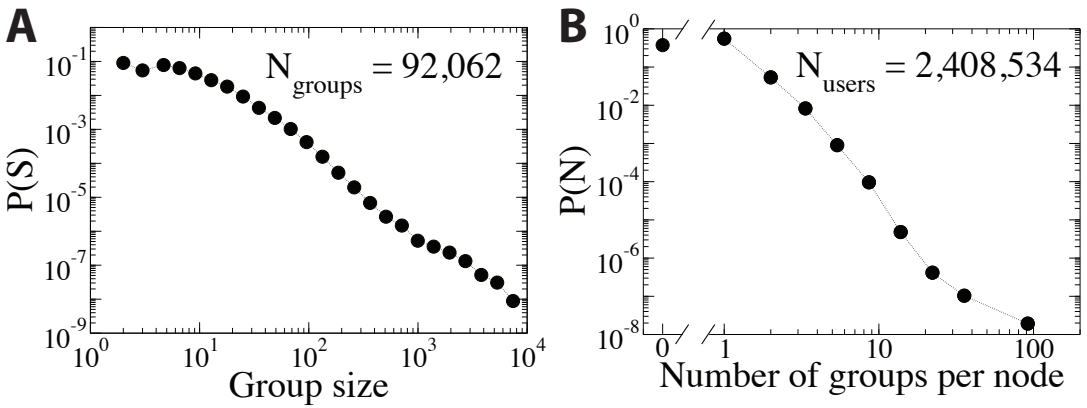
- Gray: followers
- Red: mentions
- Green: retweet
- 3 groups, one users between groups.

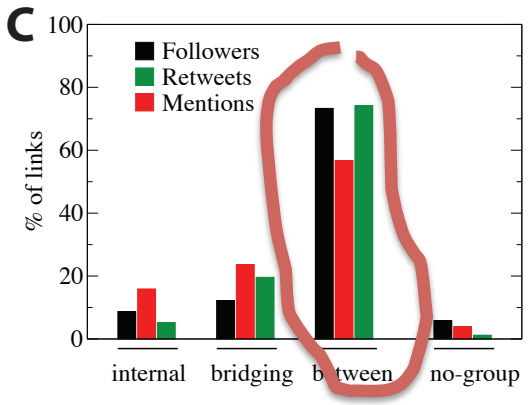


UNIVERSITY OF CAMBRIDGE

Some statistics

92,000 groups
Largest group: 10,000 users
37% users: no group





Mentions are double the followers in internal and bridging

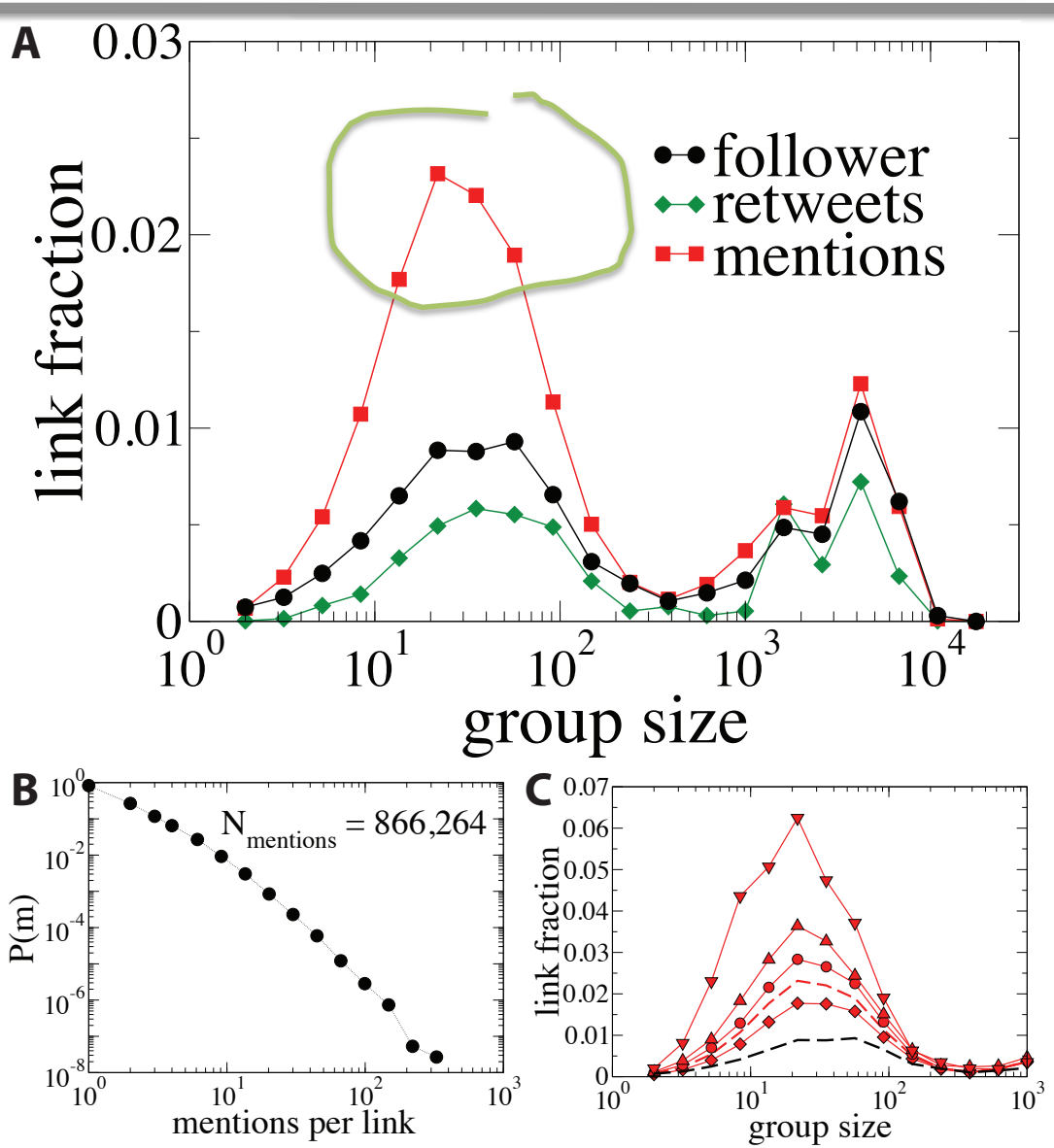
UNIVERSITY OF CAMBRIDGE

Internal Links

Internal mentions are more than follower links with groups around 100.

The distribution of mentions over links is quite wide

C: The dashed curves are the total for the follower network (black) and for the links with mentions (red). Others (from bottom to top): fractions of links with: 1 non-reciprocated mention (diamonds), 3 mentions (circles), 6 mentions (triangle up) and more than 6 reciprocated mentions (triangle down).

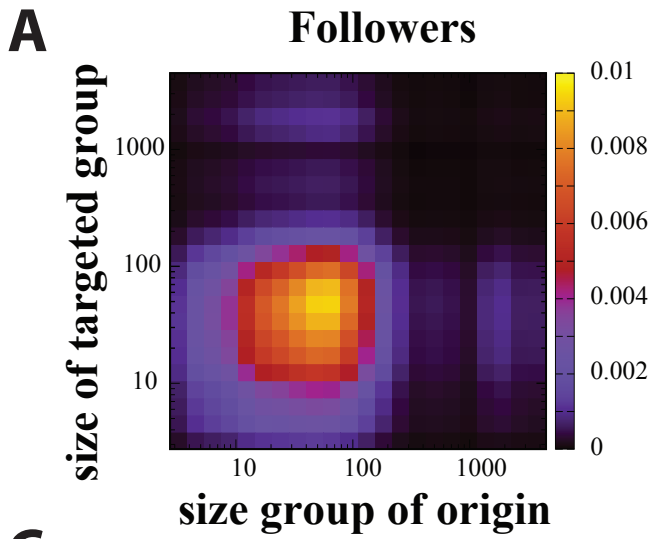


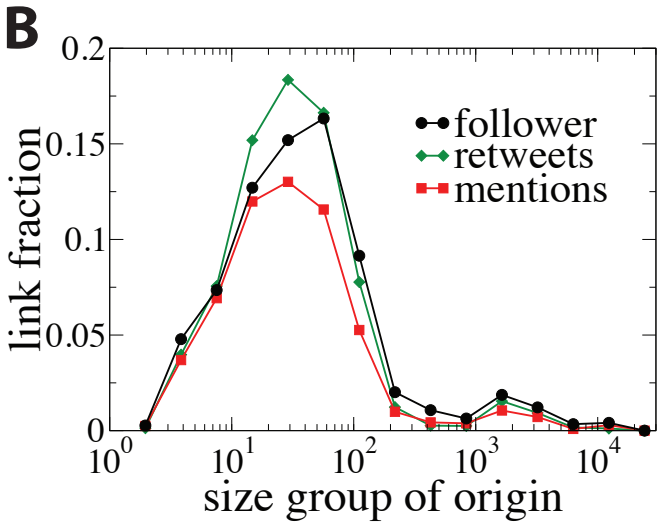
Links between groups

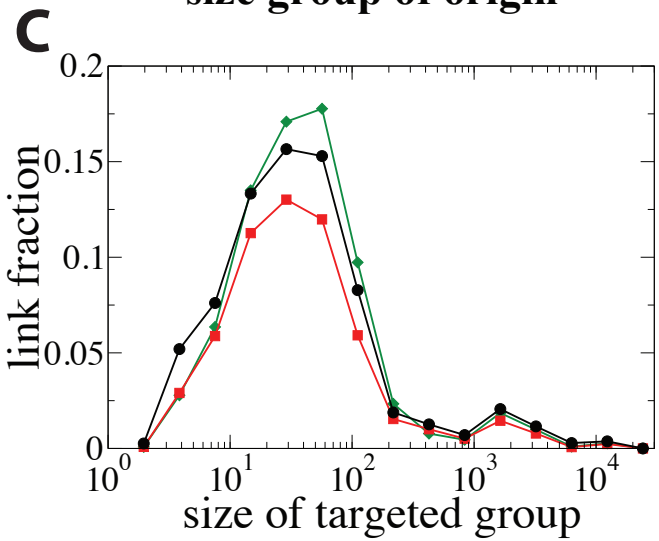
Occur between groups of <200 nodes

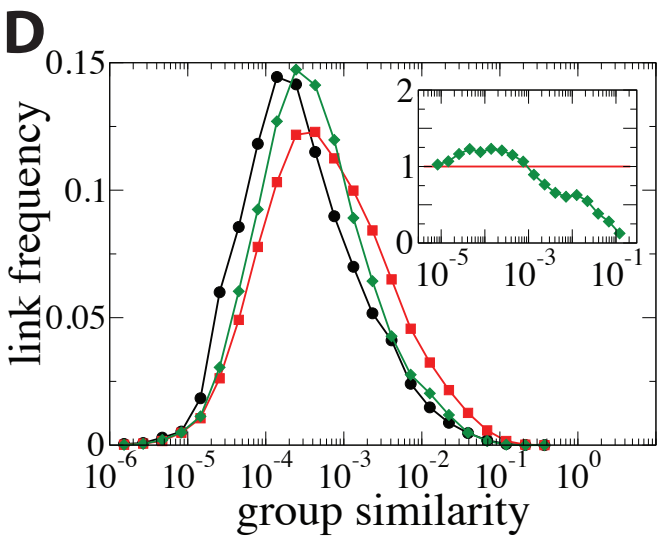
$$sim(A,B) = \frac{|\bigcap linksAandB|}{|\bigcup linksAandB|}$$

Retweets seem to occur more between groups than within! Weak ties!!!!! Retweets also seem to happen **between** less similar groups!



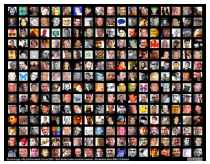


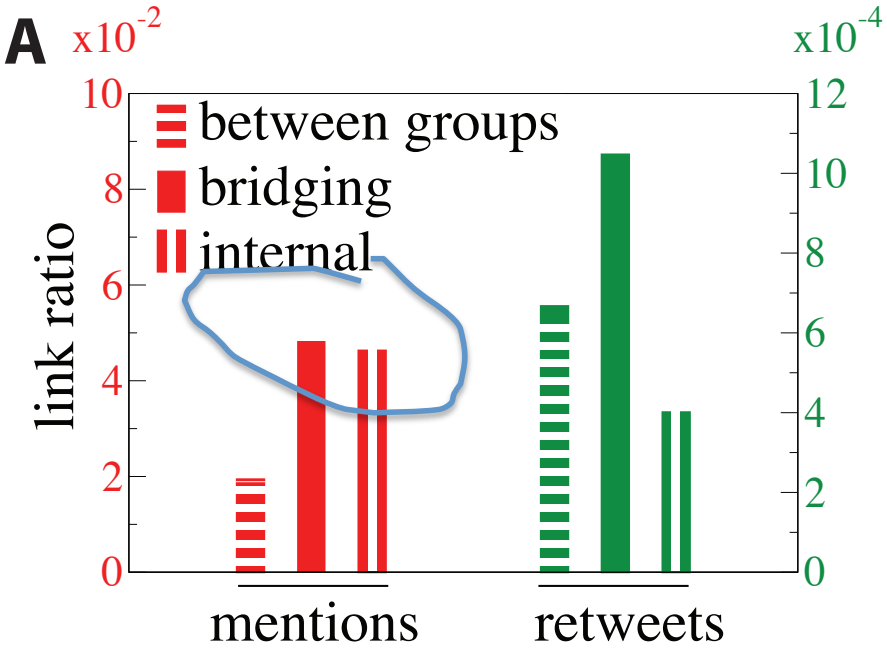


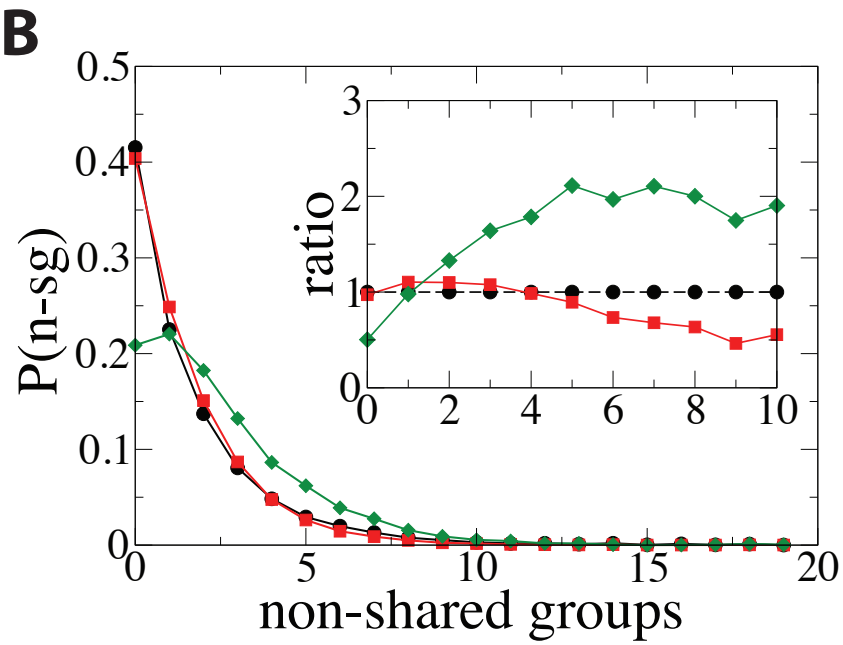


UNIVERSITY OF CAMBRIDGE

Bridge Links







Retweets on a bridge increase with the number of groups assigned to the bridging nodes

UNIVERSITY OF CAMBRIDGE

Discussion on findings

- There seems to be a correlation with the role of weak ties and the clustering done on the followers network
- Weak ties seem to be carrier of information (retweets) while internal group links seem to be more about mentions and communication

UNIVERSITY OF CAMBRIDGE

Summary

- We have discussed modularity based community detection as well as overlapping community detection.
- Many methods exist…

- We have shown cluster and weak ties analysis on an online social network dataset.

References

- S. Fortunato. **Community detection in graphs**, Arxiv 2009.
- Michelle Girvan and Mark E. J. Newman. **Community structure in social and biological networks.** Proc. Natl. Acad. Sci. USA, 99(12):7821–7826, June 2002.
- M.E.J. Newman, M. Girvan. **Finding and evaluating community structure in networks.** Phys. Rev. E 69, 026113, 2004.
- A. Clauset, M.E.J. Newman, C. Moore. **Finding community structure in very large networks**. Phys. Rev. E 70, 066111, 2004.
- G. Palla, I. Derényi, I. Farkas, and T. Vicsek. **Uncovering the overlapping community structure of complex networks in nature and society**. *Nature.* **435**, 814-818 (2005).
- A. Lancichinetti, F. Radicchi,3 J. Ramasco, S. Fortunato. **Finding statistically significant communities in networks.** PLOS One 2011; 6(4). (not discussed in class).
- P. Grabowicz, J. Ramasco, E. Moro, J. Pujol, V.. Eguiluz. **Social features of online networks: the strength of weak ties in online social media. arXiv: 1107.4009. July 2011.**

UNIVERSITY OF CAMBRIDGE