Digital Media and Social Networks

Kleomenis Katevas
k.katevas@qmul.ac.uk
https://minoskt.github.io

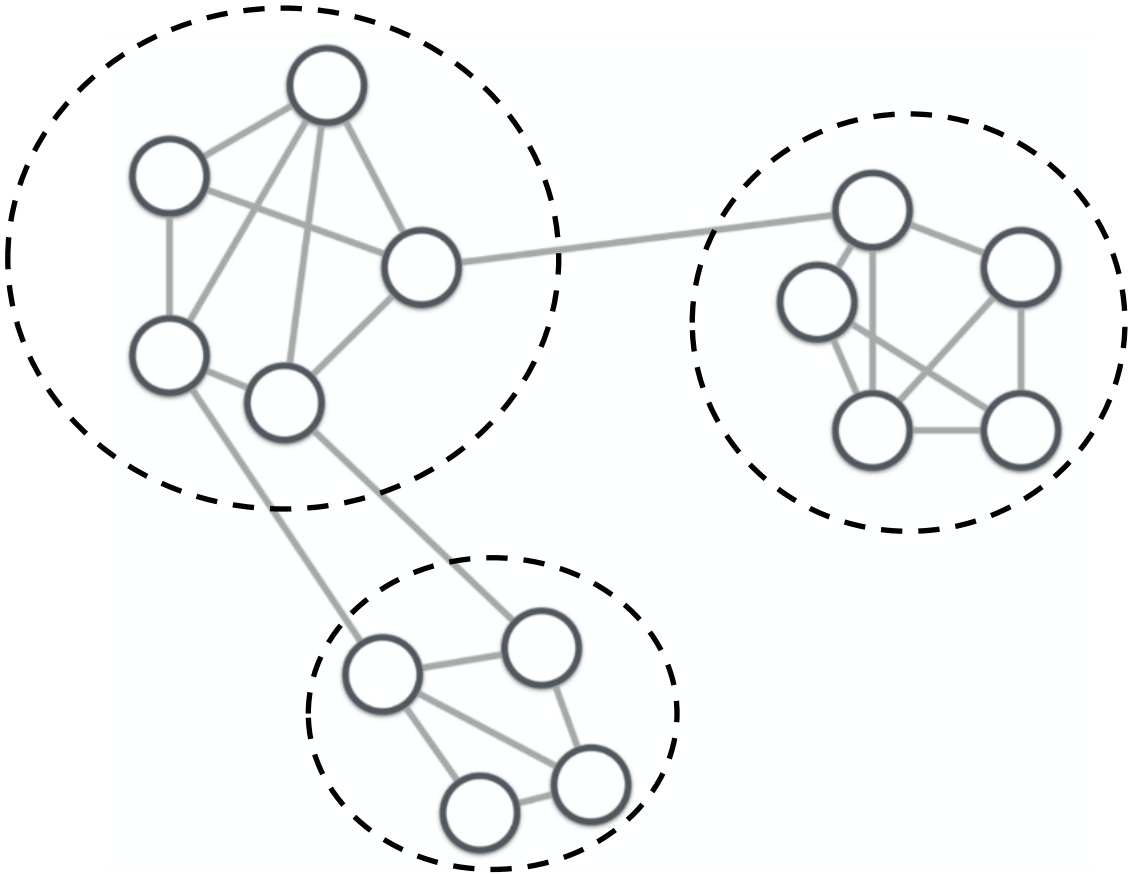
WEEK 5: COMMUNITY DETECTION AND OVERLAPPING COMMUNITIES

SOME SLIDES COPYRIGHT CECILIA MASCOLO (CAMBRIDGE) AND HAMED HADDADI (QMUL)

COMMUNITIES

- Weak ties (Lecture 2) seemed to bridge groups of tightly coupled nodes (communities)

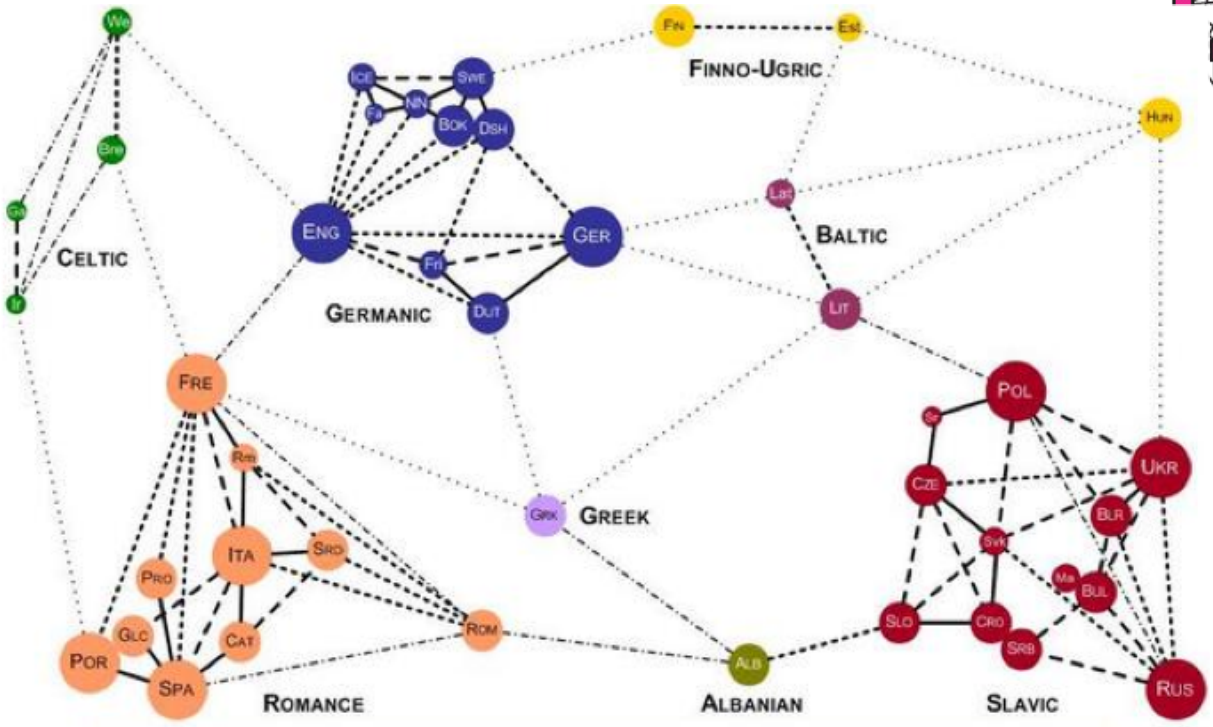
- How do we find these communities?

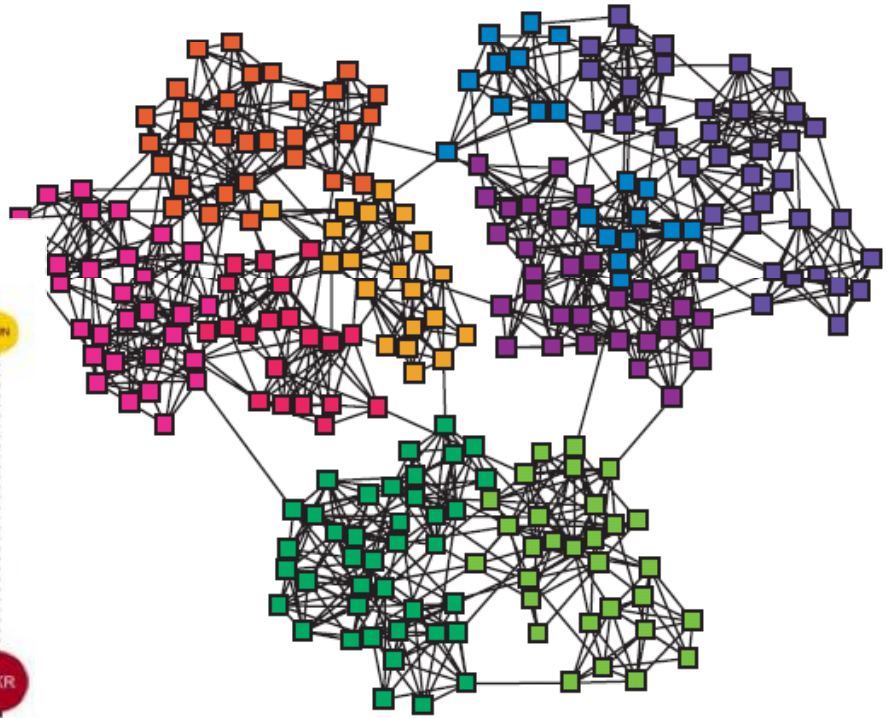


Not too hard…

But this???

WHAT IS A COMMUNITY?





WHY DO WE WANT TO FIND PARTITIONS/COMMUNITIES?

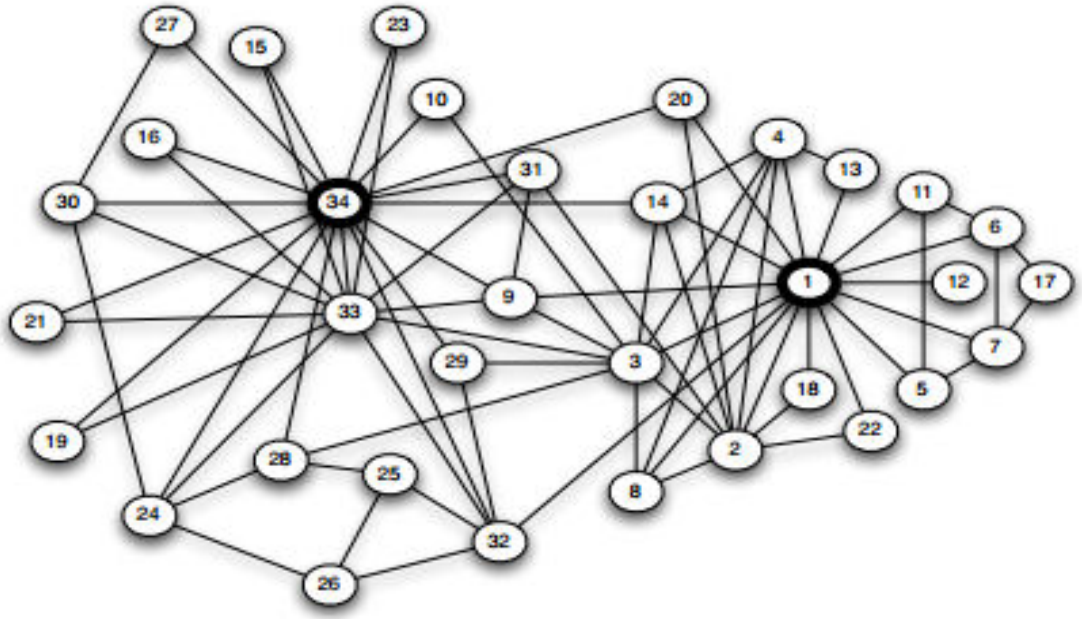
- Clustering web clients with similar interest or geographically near can improve performance
- Customers with similar interests could be clustered to help recommendation systems
- Clusters in large graphs can be used to create data structures for efficient storage of graph data to handle queries or path searches
- Study the relationship/mediation among nodes
  - Hierarchical organization study

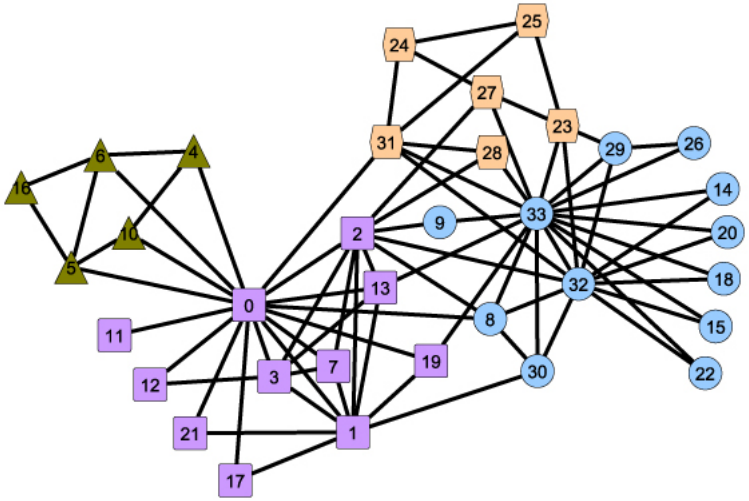
IN THIS LECTURE

- We will describe a Community Detection method based on betweenness centrality.
- We will describe the concept of Modularity and Modularity Optimization.
- We will describe methods for overlapping community detection.

EXAMPLE

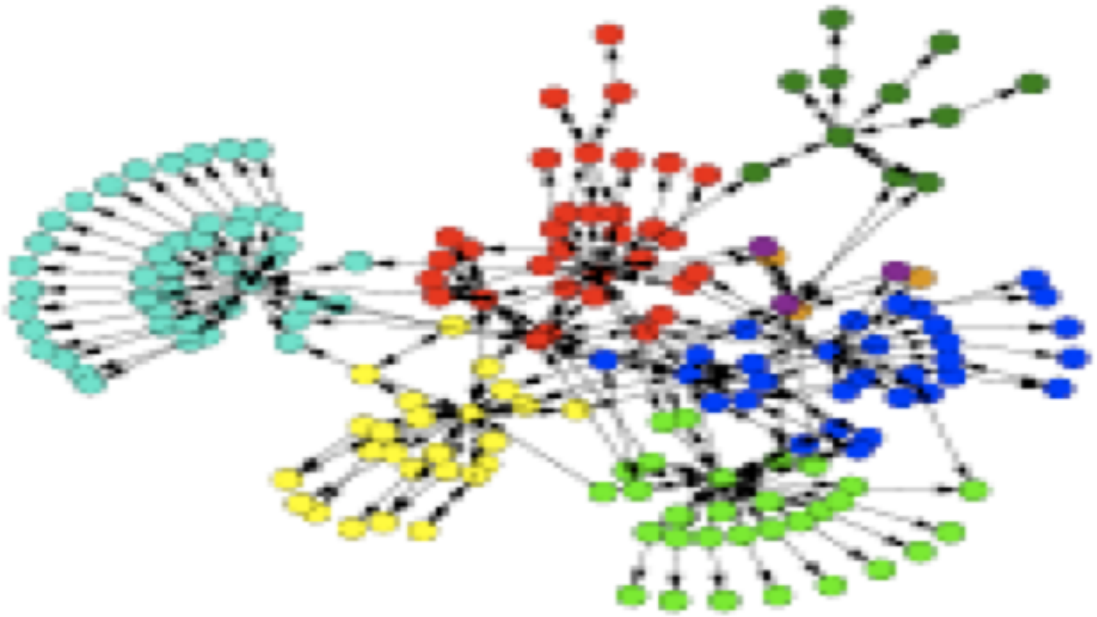
Zachary's Karate club: 34 members of a club over 3 years. Edges: interaction outside the club





EXAMPLE

WWW: pages and hyperlinks Identification of clusters can improve pageranking.



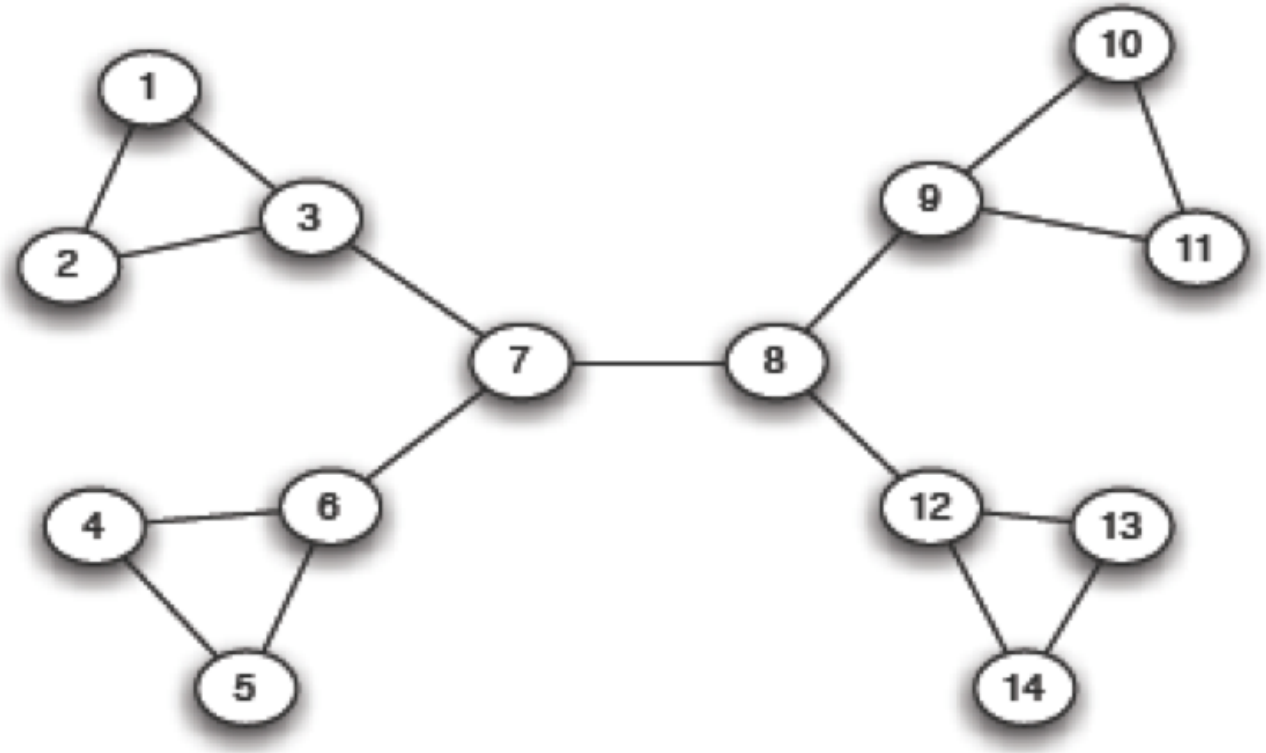
REMOVE WEAK TIES

- Local bridges connect weakly interacting parts of the network
- What if we have many bridges: which do we remove first? Or there might be no bridges.

- Note: **Without those bridges paths between nodes would be longer**

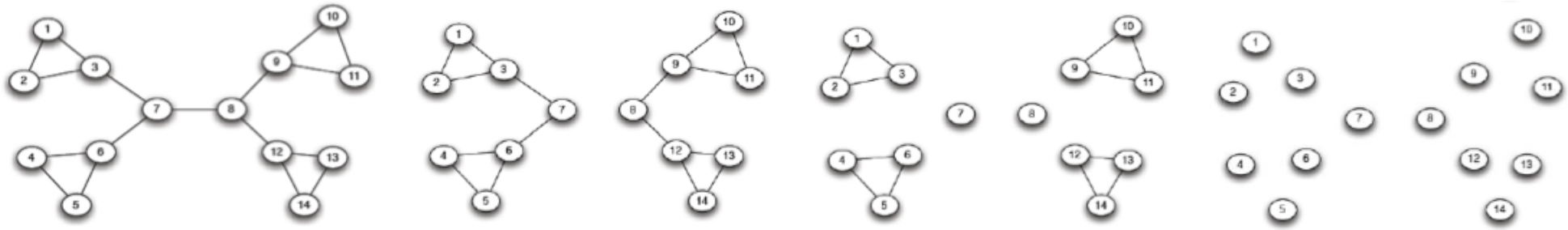
EDGE BETWEENNESS

Edge Betweenness: the number of shortest paths between pairs of nodes that run along the edge.



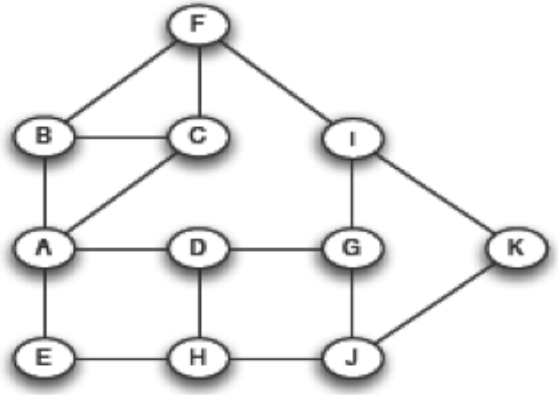
ALGORITHM OF GIRVAN-NEWMANN (PNAS 2002)

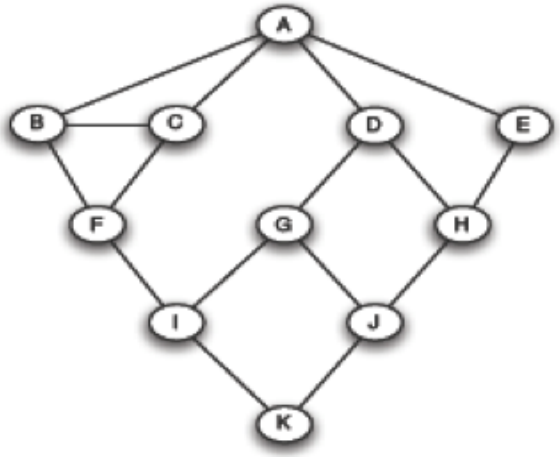
- Calculate the betweenness of all edges
- Cut the edge with highest betweenness
- Recalculate edge betweenness



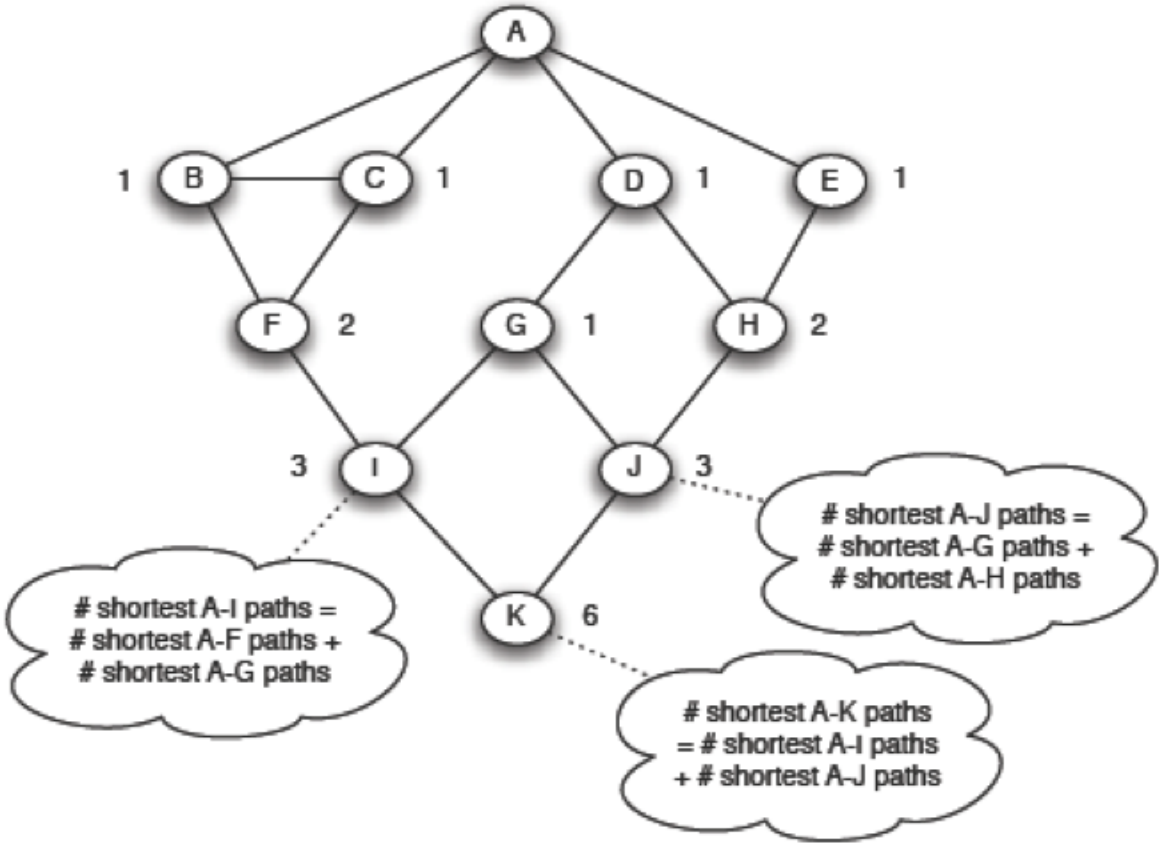
HOW IS BETWEENNESS COMPUTED?

- Calculate the shortest paths from node A
  - BFS search from A.
  - Determine number of shortest paths from A to each node
  - Based on these numbers, determine the amount of flow from A to all other nodes that uses each edge.

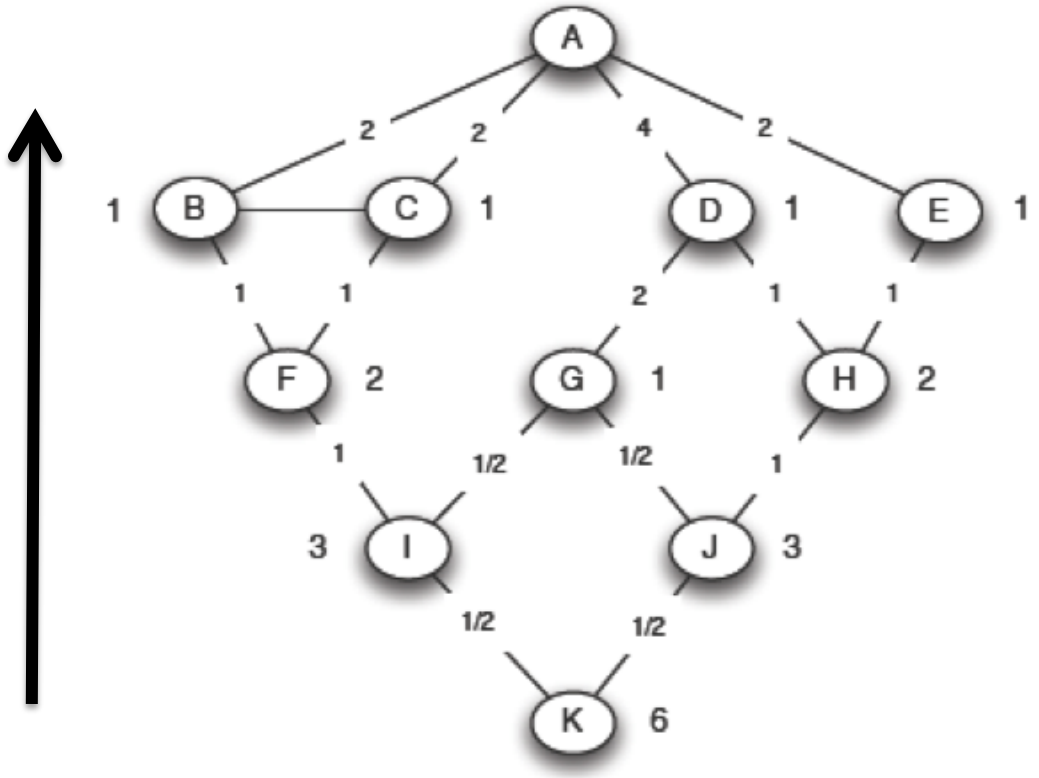




CALCULATING NUMBER OF SHORTEST PATHS



CALCULATING FLOWS



When we get to a node X in the breadth-first search structure, working up from the bottom, we add up all the flow arriving from edges directly below X, plus 1 for the flow destined for X itself. We then divide this up over the edges leading upward from X, in proportion to the number of shortest paths coming through each.

CALCULATING EDGE BETWEENNESS

- Build one of these graphs for each node in the graph
- Sum the values on the edges on each graph to obtain the edge betweenness

EDGE: DELETION
WHEN DO WE STOP?

- How do we know when to stop?

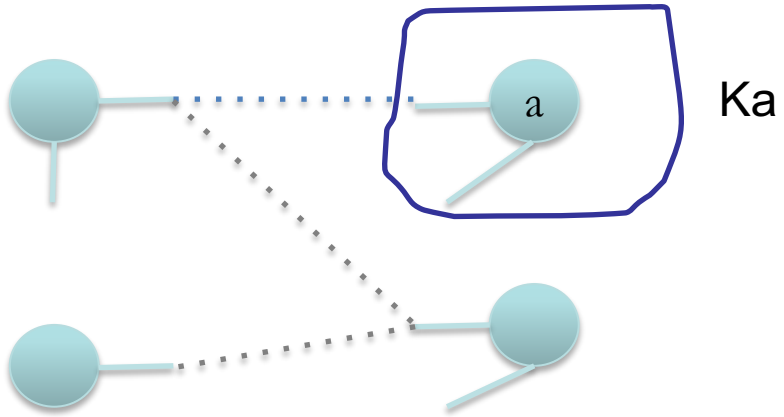
- When X communities have been detected?
- When the level of cohesion inside a community has reached Y?

- There is no prescriptive way for every case
- There are also many other ways of detecting communities

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 for a random graph with same node degree distribution as the given network)***

MODULARITY ON A RANDOMIZED GRAPH CALCULATION



The expected number of edges in the randomized version of the graph where nodes are rewired:

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

m is the number of edges of the graph = ½ sum(ki)

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)

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.
  - But, even these start failing as the graph gets larger.

EXAMPLE OF DENDROGRAM

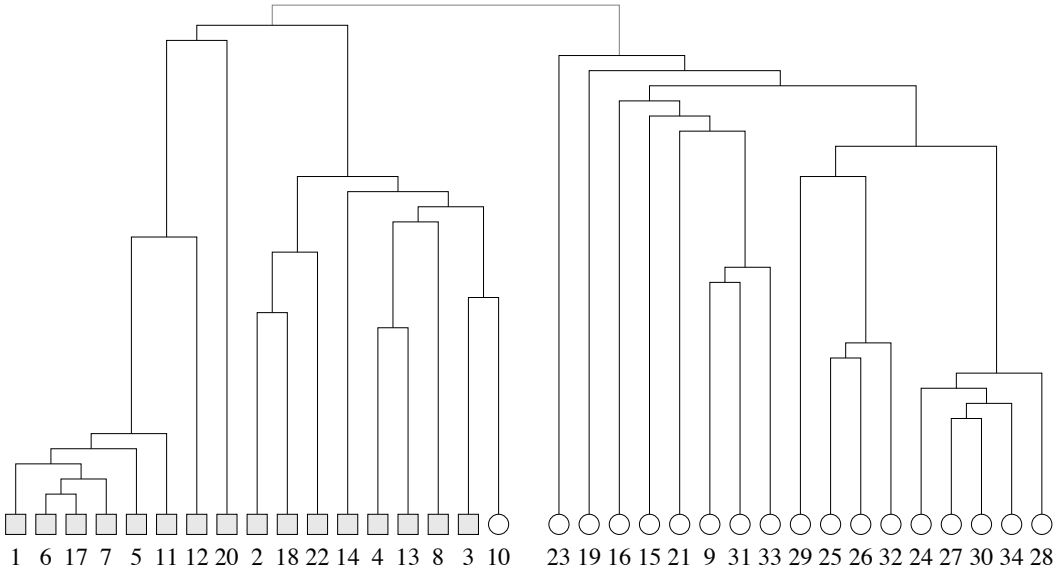


FIG. 2: Dendrogram of the communities found by our algorithm in the "karate club" network of Zachary [5, 17]. The shapes of the vertices represent the two groups into which the club split as the result of an internal dispute.

APPLICATION TO AMAZON RECOMMEDATIONS

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

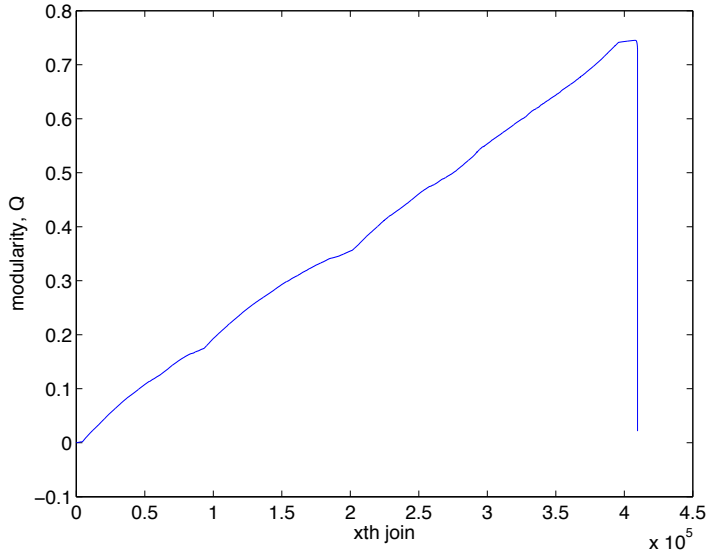


FIG. 1: The modularity Q over the course of the algorithm (the x axis shows the number of joins). Its maximum value is $Q = 0.745$, where the partition consists of 1684 communities.

AMAZON: TOP COMMUNITIES (87% OF NODES)

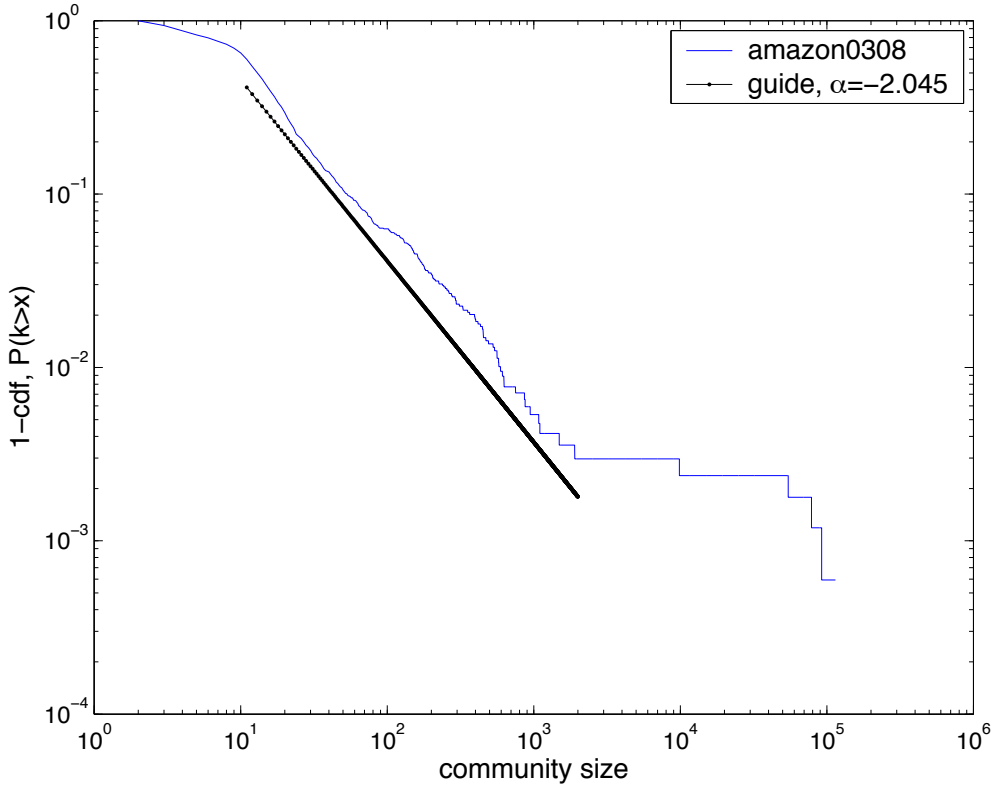
Rank	Size	Description
1	114538	General interest: politics; art/literature; general fiction; human nature; technical books; how things, people, computers, societies work, etc.
2	92276	The arts: videos, books, DVDs about the creative and performing arts
3	78661	Hobbies and interests I: self-help; self-education; 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.

AMAZON: COMMUNITY SIZE DISTRIBUTION

A power law distribution of community size

(more on power laws
in later lectures)



LIMITATIONS OF MODULARITY

Modularity is not a perfect measures

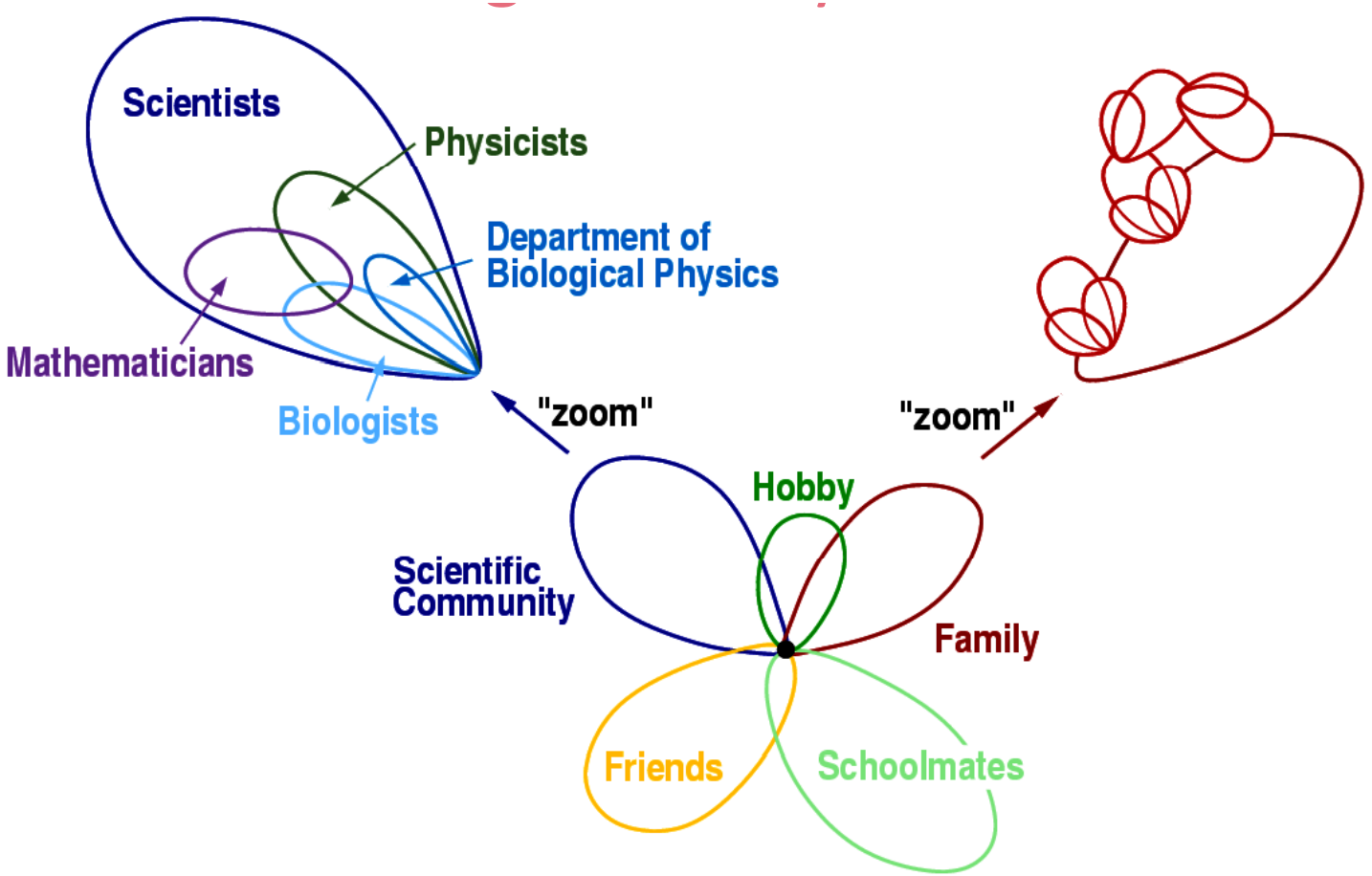
It appears to depend on the number of links in the network (L).

Problems for modules with a number of internal links of the order of $\sqrt{2L}$ or smaller.

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

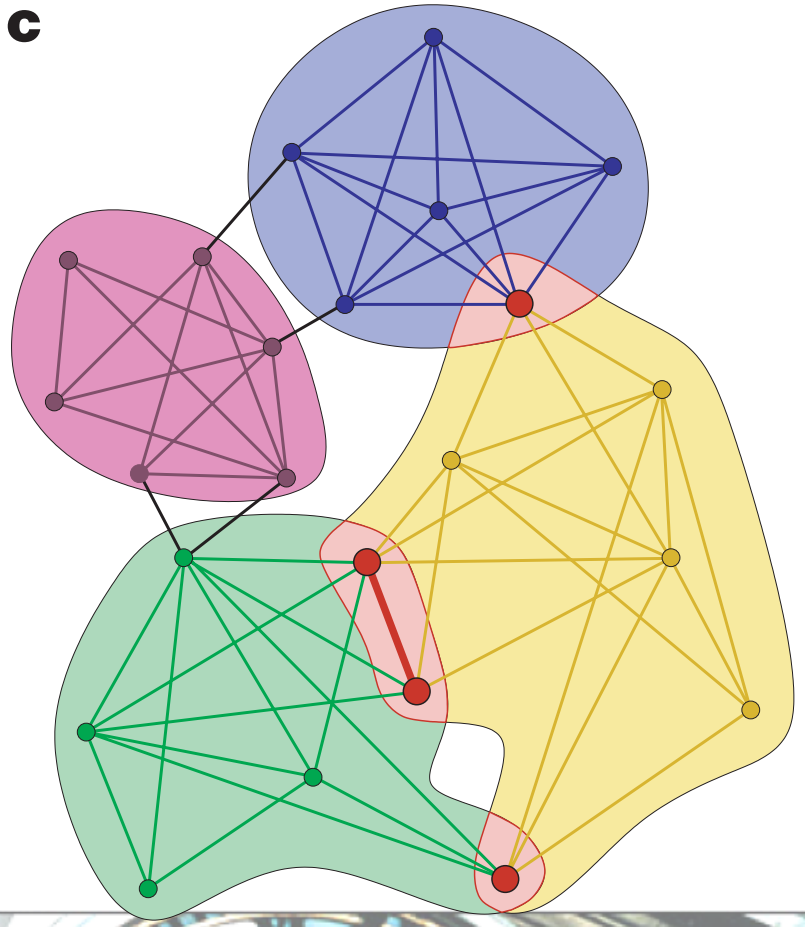
S. Fortunato, S. Barthelemy. Resolution limit in community detection. Proc. Natl. Acad. Sci., 2007.

NODES CAN BELONG TO MORE THAN 1 SOCIAL CIRCLE!



OVERLAPPING COMMUNITIES

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



LINK COMMUNITIES (NATURE 2010)

Communities in networks often overlap such that nodes simultaneously belong to several groups! Instead of finding overlapping node communities, reinvent:

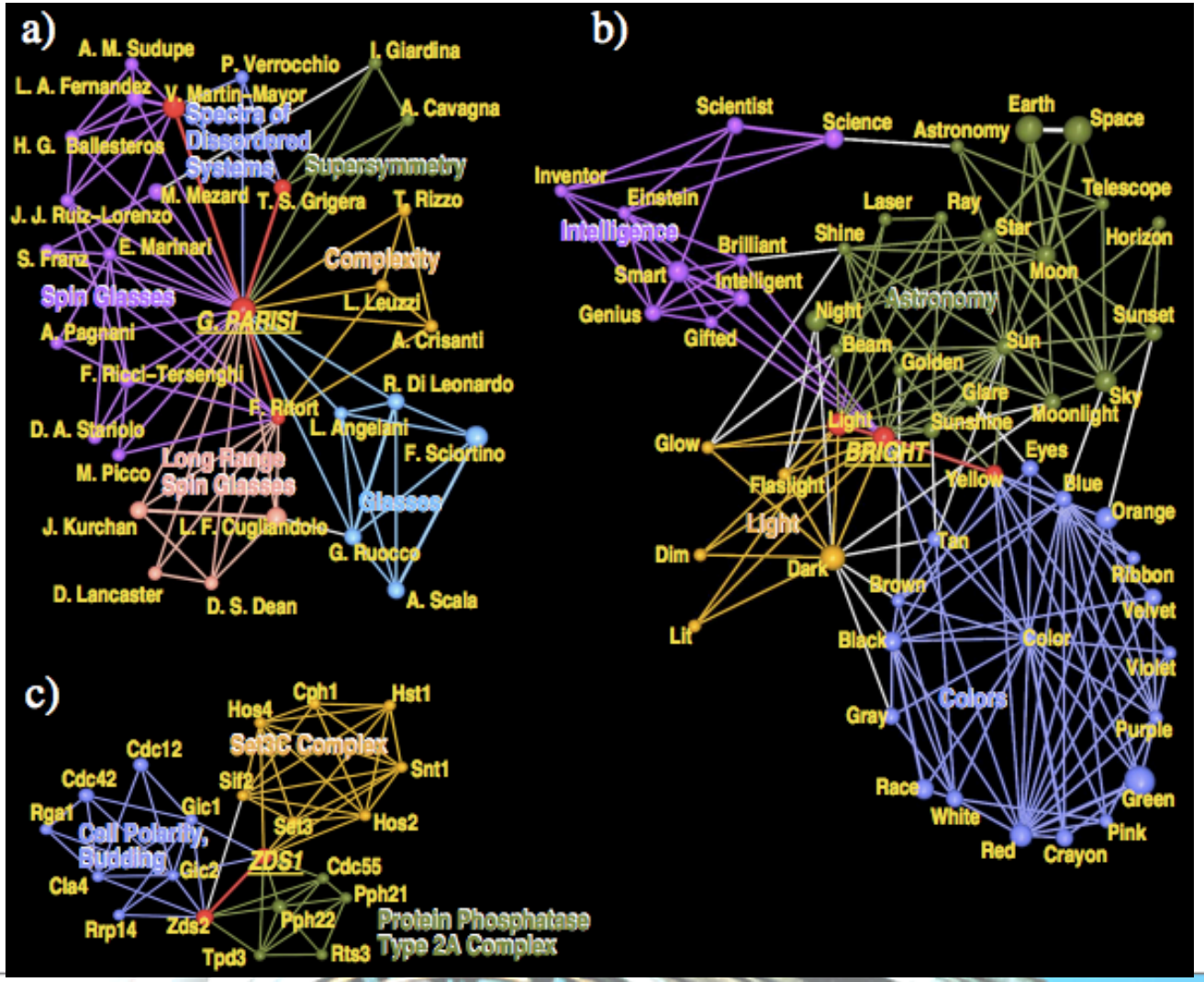
- communities as groups of links rather than nodes

Link communities reveal overlap and hierarchical organisation in networks to be two aspects of the same phenomenon.

APPLICATION

Overlapping networks:

1) Parisi's coauthorship networks
2) Networks of "bright" in the word association network
3) Protein to protein interaction network



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.

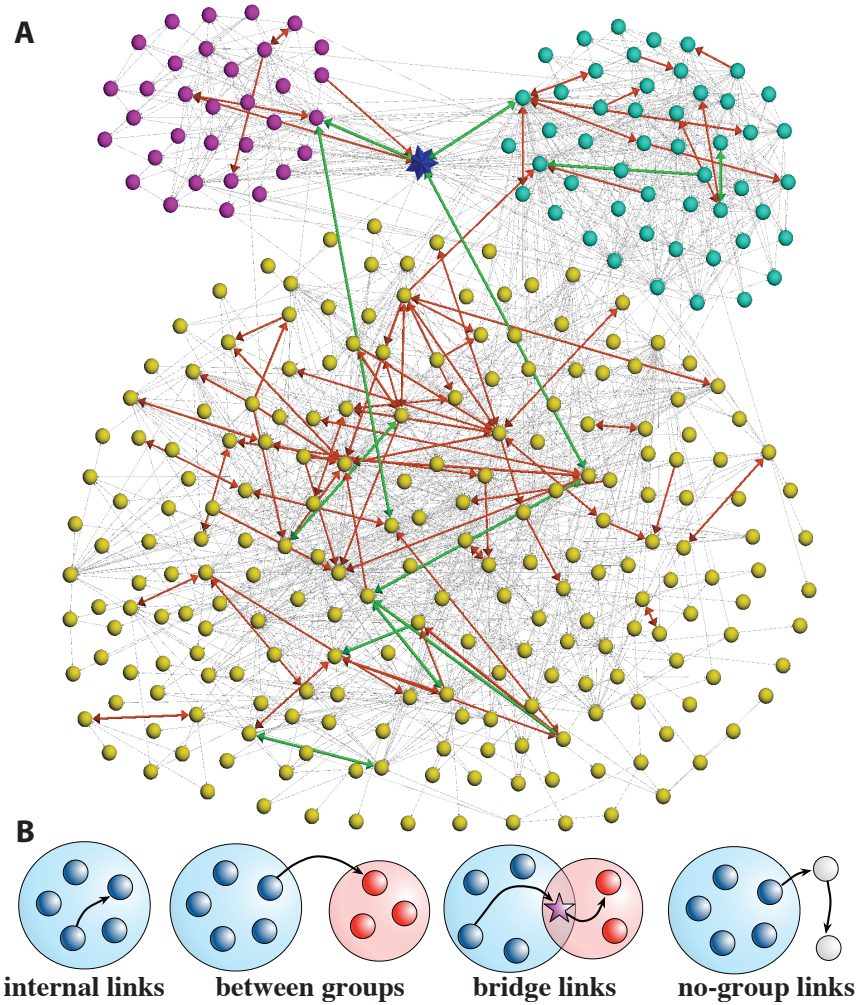
SAMPLE

Gray: followers

Red: mentions

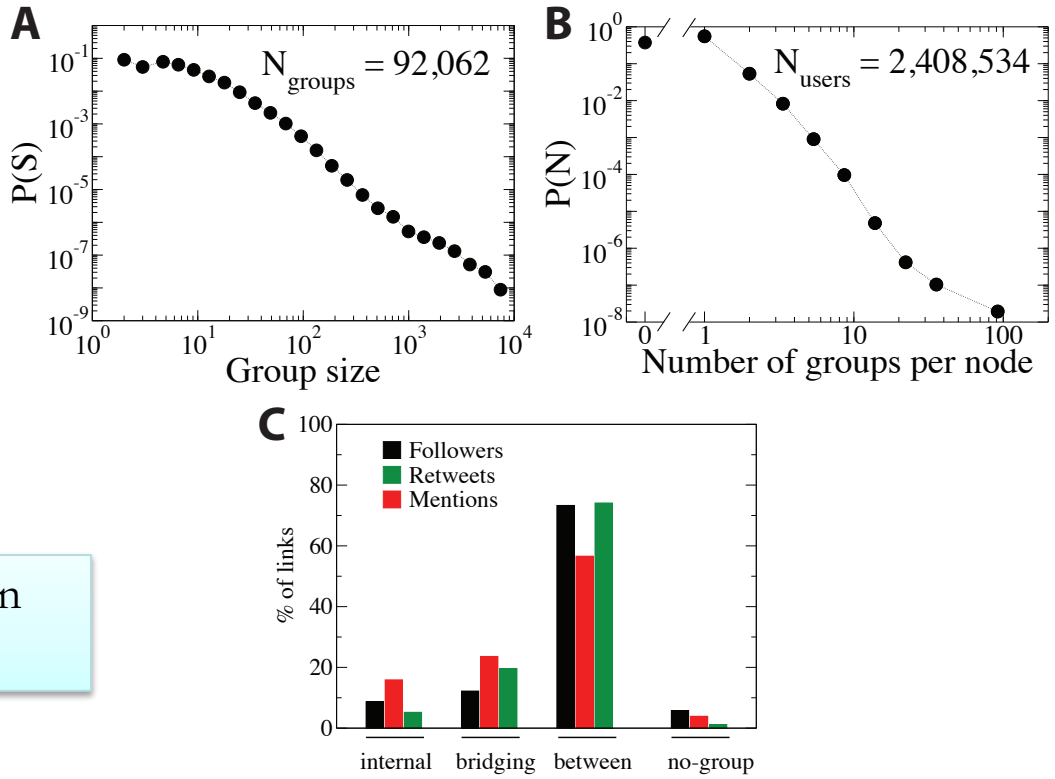
Green: retweet

3 groups, one users between groups.



SOME STATISTICS

92,000 groups
Largest group: 10,000 users



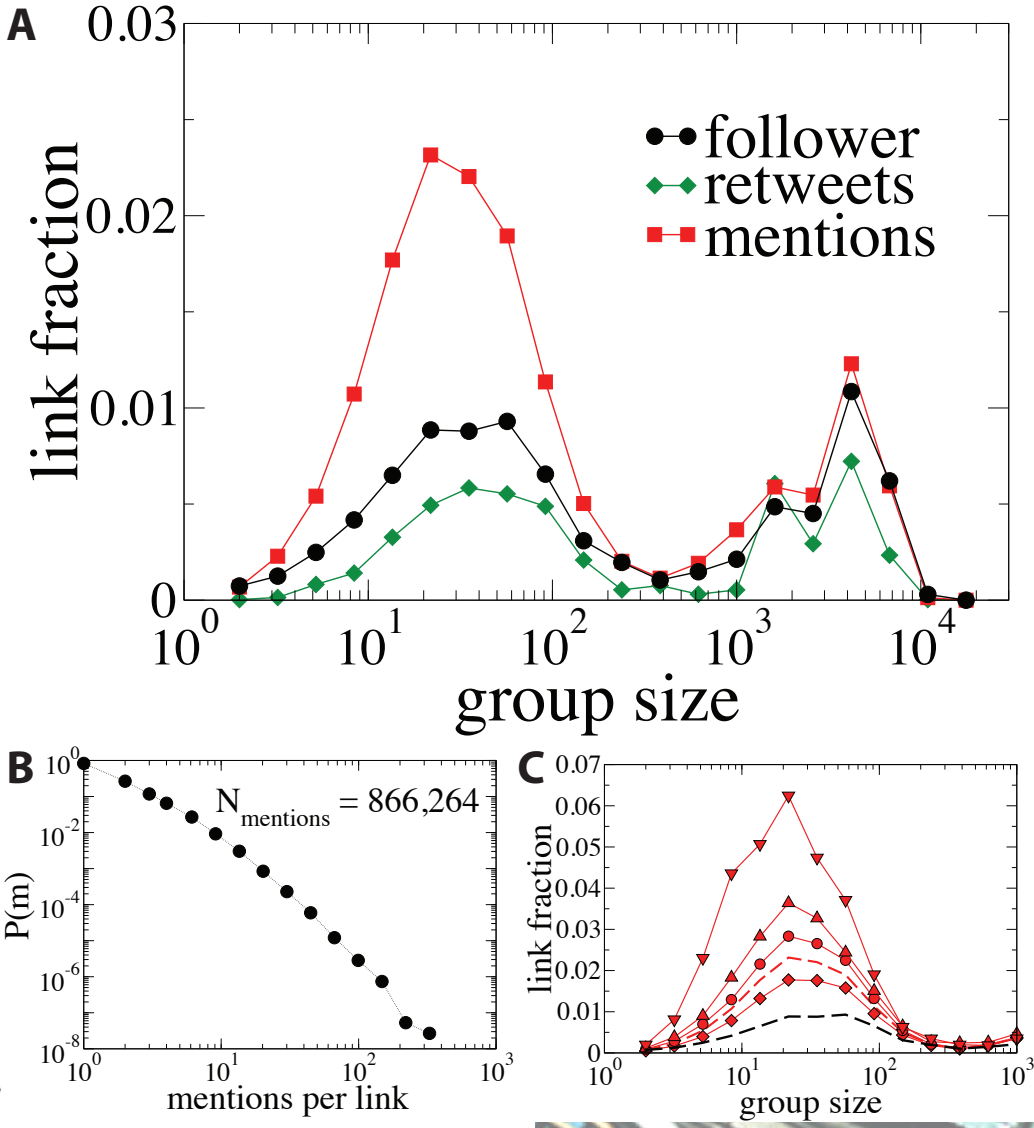
Mentions are double the followers in internal and bridging

INTERNAL LINKS

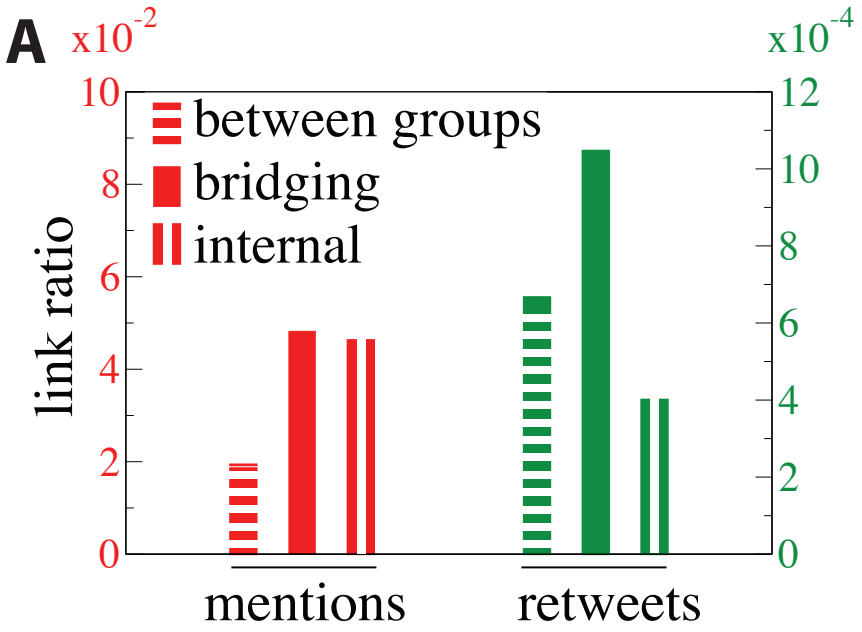
Internal mentions are more than follower links with groups around 100.

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).



BRIDGE LINKS



Bridges have similar mentions internally
But they attract more retweets

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

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

- M. Girvan and M. E. J. Newman. **Community structure in social and biological networks** Proc. Natl. Acad. Sci. USA, 99(12):7821–7826, June 2002.
- 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).
- 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.**