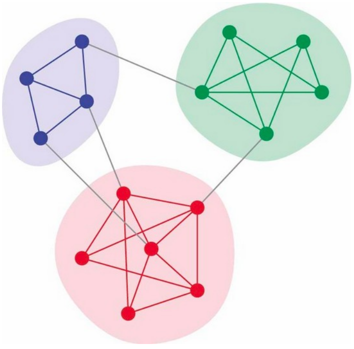
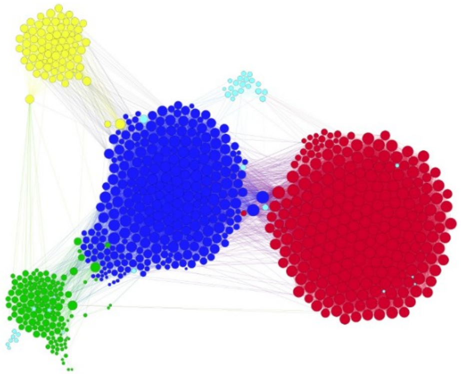
Introduction to the Analysis and Visualisation of Complex Networks

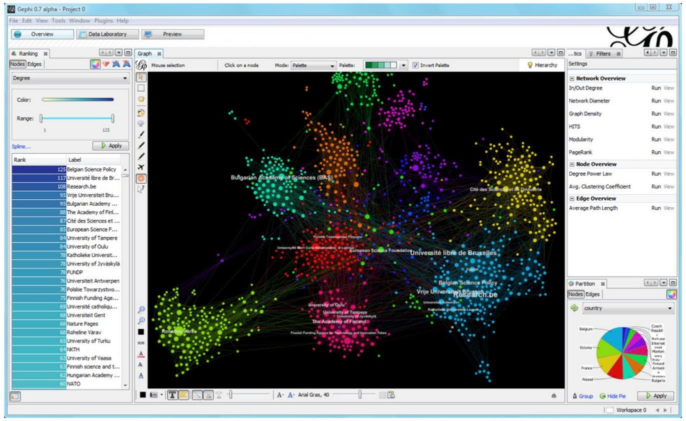
U.PORTO
FACULDADE DE CIÊNCIAS
UNIVERSIDADE DO PORTO

Pedro Ribeiro
(DCC/FCUP & CRACS/INESC-TEC)

CRACS
INESCTEC
ASSOCIATE LABORATORY
PORTUGAL







Rank	Label
125	Belgian Science Policy
117	Université libre de Br...
108	Research.be
95	Vrije Universiteit Bru...
95	Bulgarian Academy ...
88	The Academy of Finl...
87	Cité des Sciences et ...
85	European Science F...
84	University of Tampere
84	University of Oulu
78	Katholieke Universit...
78	University of Jyväskylä
78	FUNDP
76	Universiteit Antwerpen
76	Polskie Towarzystwo...
75	Finnish Funding Age...
69	Université catholiqu...
68	Universiteit Gent
68	Nature Pages
68	Roheline Värav
65	University of Turku
64	NKTH
63	University of Vaasa
63	Finnish science and t...
62	Hungarian Academy ...
60	NATO

Network Overview		
In/Out Degree	Run	View
Network Diameter	Run	View
Graph Density	Run	View
HITS	Run	View
Modularity	Run	View
PageRank	Run	View
Node Overview		
Degree Power Law	Run	View
Avg. Clustering Coefficient	Run	View
Edge Overview		
Average Path Length	Run	View

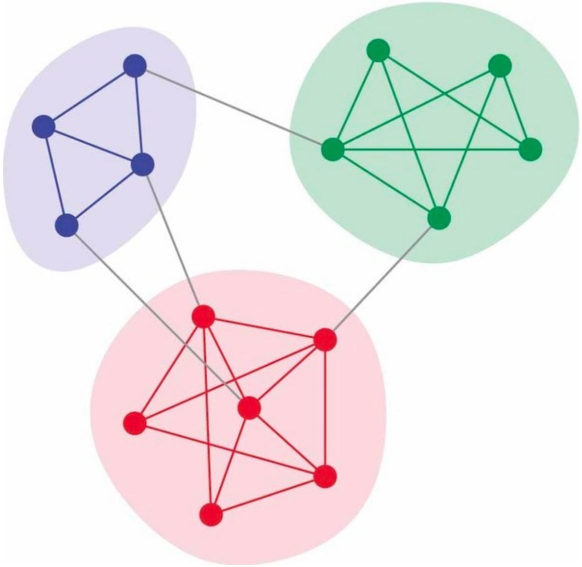
(this part includes some slides heavily based on material from Jure Leskovec and Lada Adamic @ Stanford University + Gonzalo Mateos @ Rochester University)

CLAD
Associação Portuguesa de Classificação e Análise de Dados

Community Structure

Network Communities

- Empirical data (and theory) supports the notion that networks are composed of **tightly connected sets of edges**

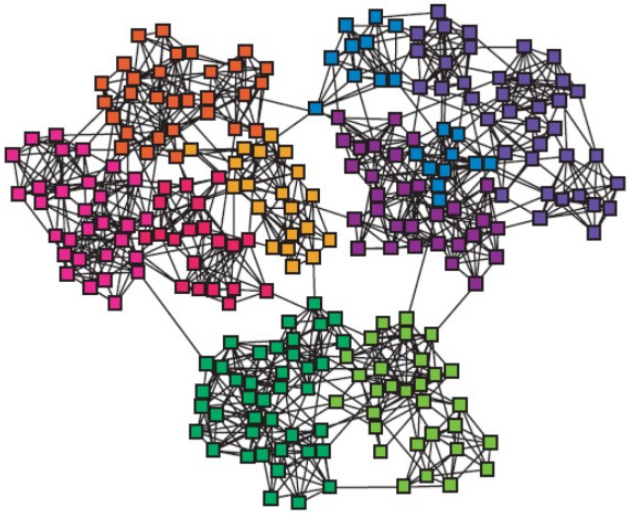


Communities, clusters, groups, modules

- **Network Communities**
  - Sets of nodes with **lots of internal** connections and **few external** ones (to the rest of the network)

Finding Network Communities

- How to **automatically** find such densely connected groups of nodes?

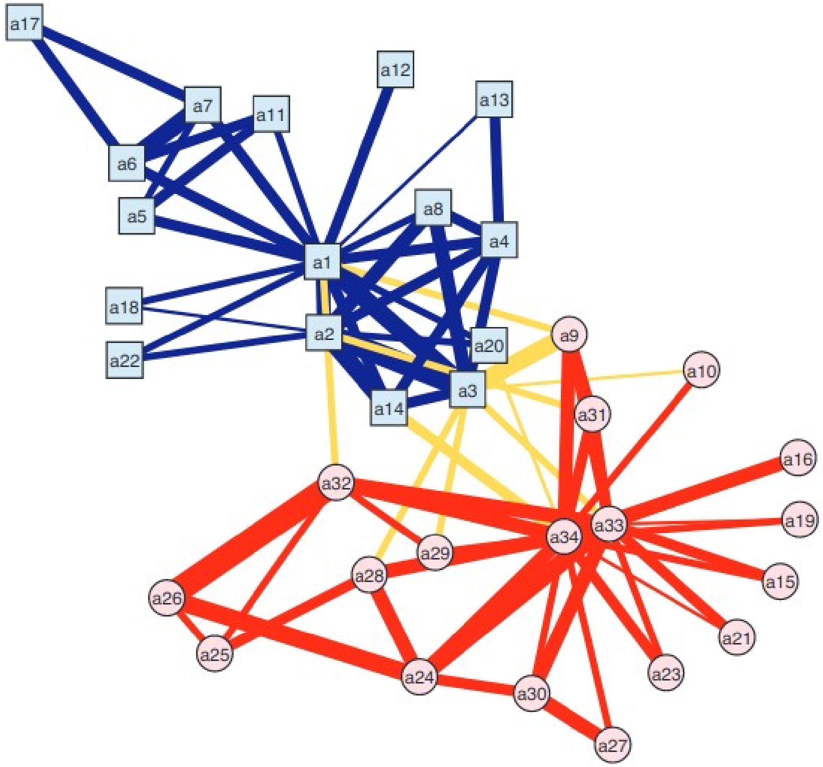


Communities, clusters, groups, modules

- Ideally, such discovered clusters would correspond to real groups. For example:

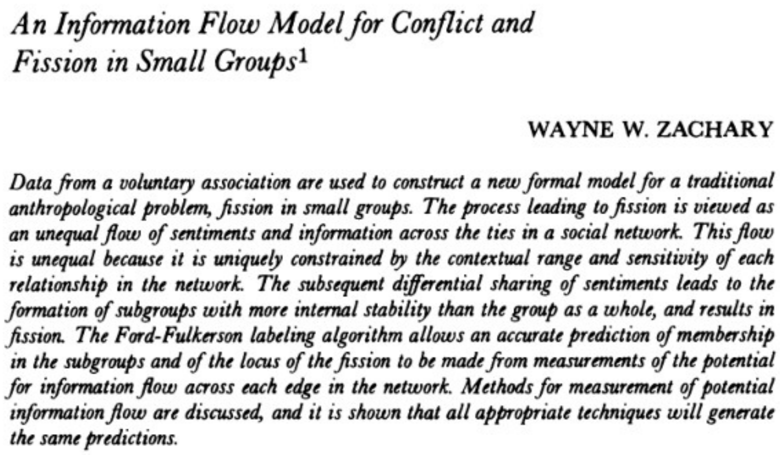
Zachary’s Karate Club

- Social interactions among members of a karate club in the 70s



- Zachary witnessed the club split in two during his study
  - Toy network, yet canonical for community detection algorithms
  - Offers “ground truth” community membership *(a rare luxury)*

Zachary’s Karate Club

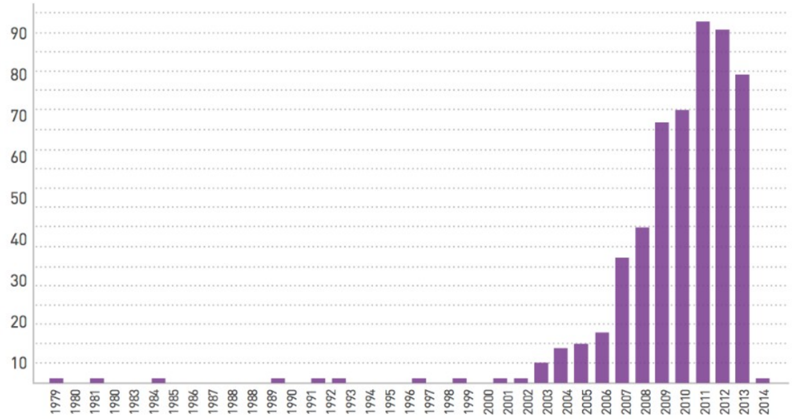


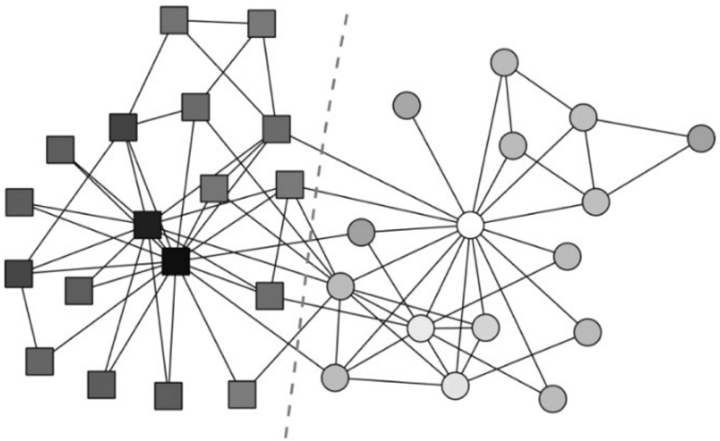
An Information Flow Model for Conflict and Fission in Small Groups[1]

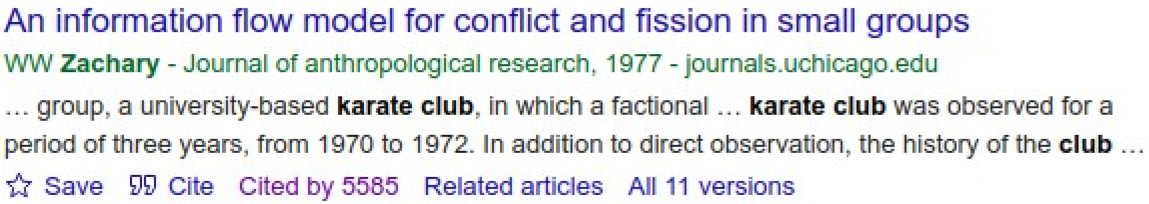
WAYNE W. ZACHARY

Data from a voluntary association are used to construct a new formal model for a traditional anthropological problem, fission in small groups. The process leading to fission is viewed as an unequal flow of sentiments and information across the ties in a social network. This flow is unequal because it is uniquely constrained by the contextual range and sensitivity of each relationship in the network. The subsequent differential sharing of sentiments leads to the formation of subgroups with more internal stability than the group as a whole, and results in fission. The Ford-Fulkerson labeling algorithm allows an accurate prediction of membership in the subgroups and of the locus of the fission to be made from measurements of the potential for information flow across each edge in the network. Methods for measurement of potential information flow are discussed, and it is shown that all appropriate techniques will generate the same predictions.

Citation history
of the Zachary’s Karate club paper







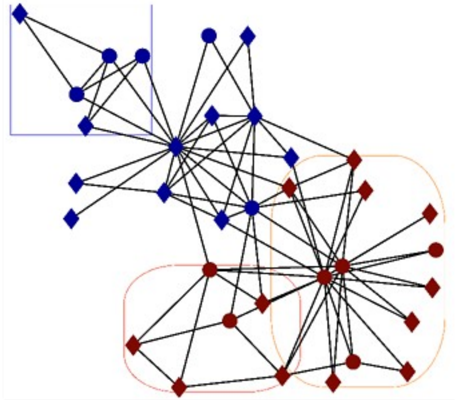
An information flow model for conflict and fission in small groups

WW **Zachary** - Journal of anthropological research, 1977 - journals.uchicago.edu

… group, a university-based **karate club**, in which a factional … **karate club** was observed for a period of three years, from 1970 to 1972. In addition to direct observation, the history of the **club** …

Zachary's Karate Club

The first scientist at any conference on networks who uses Zachary's karate club as an example is inducted into the Zachary Karate Club Club, and awarded a prize.

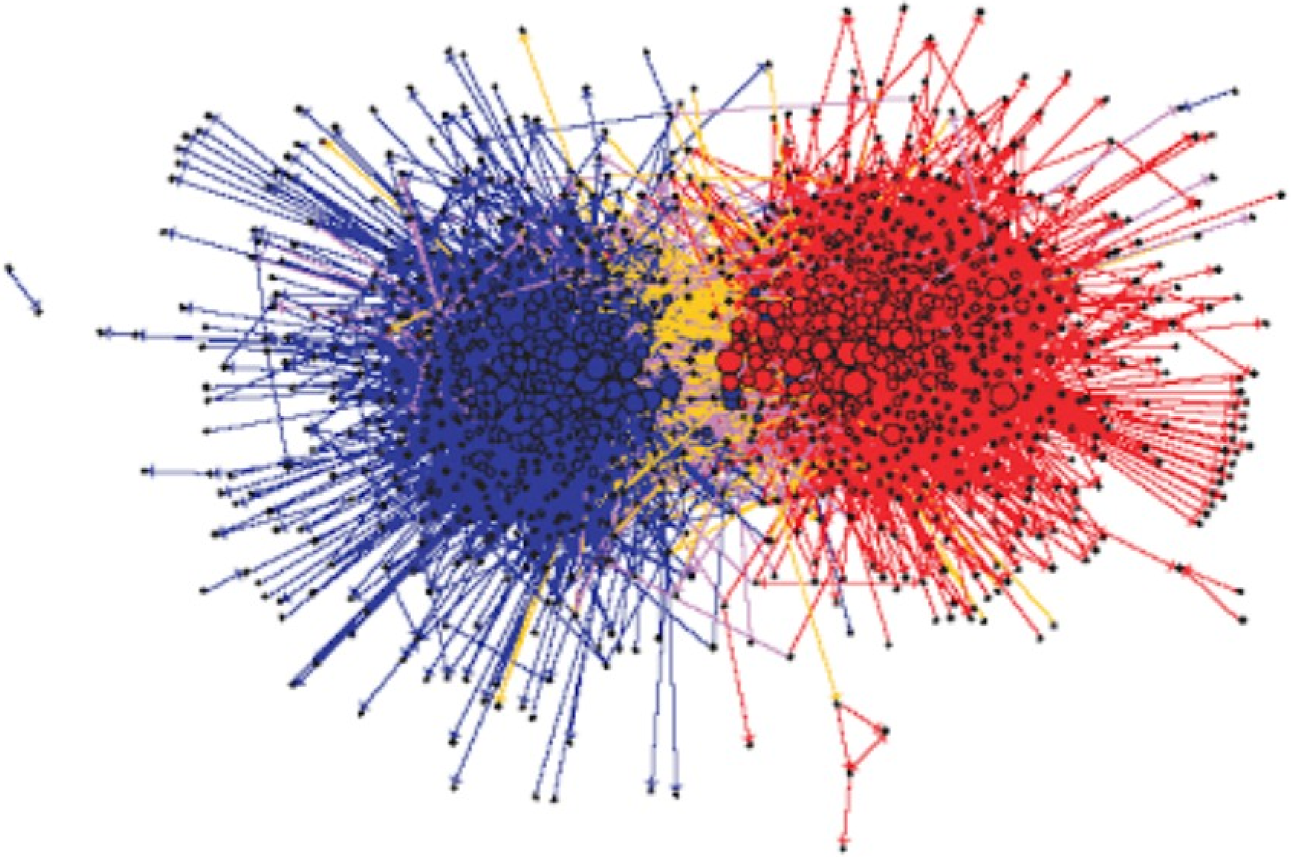


If you can't get it right on this network, then go home

https://networkkarate.tumblr.com/

Political Blogs

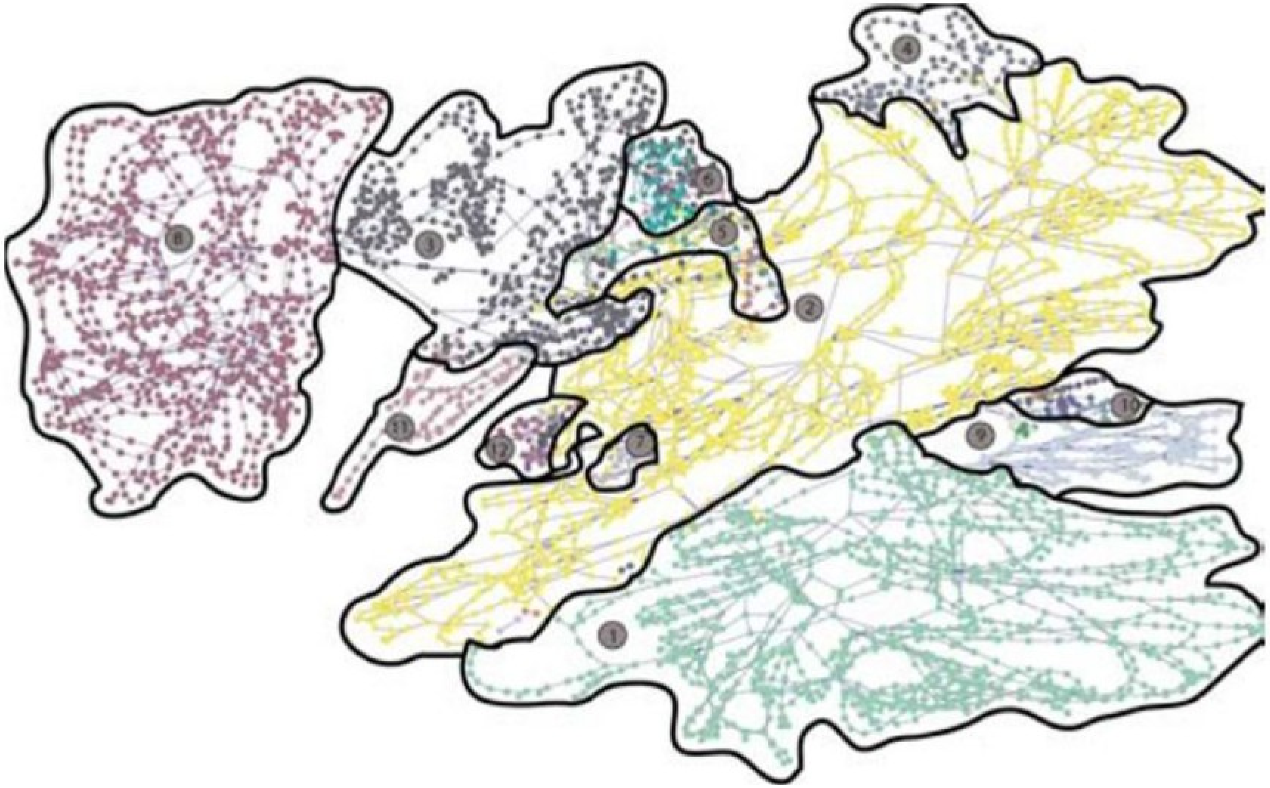
- The political blogosphere for the US 2004 presidential election



- Community structure of **liberal** and **conservative** blogs is apparent
  - People have a stronger tendency to interact with “equals”

Electrical Power Grid

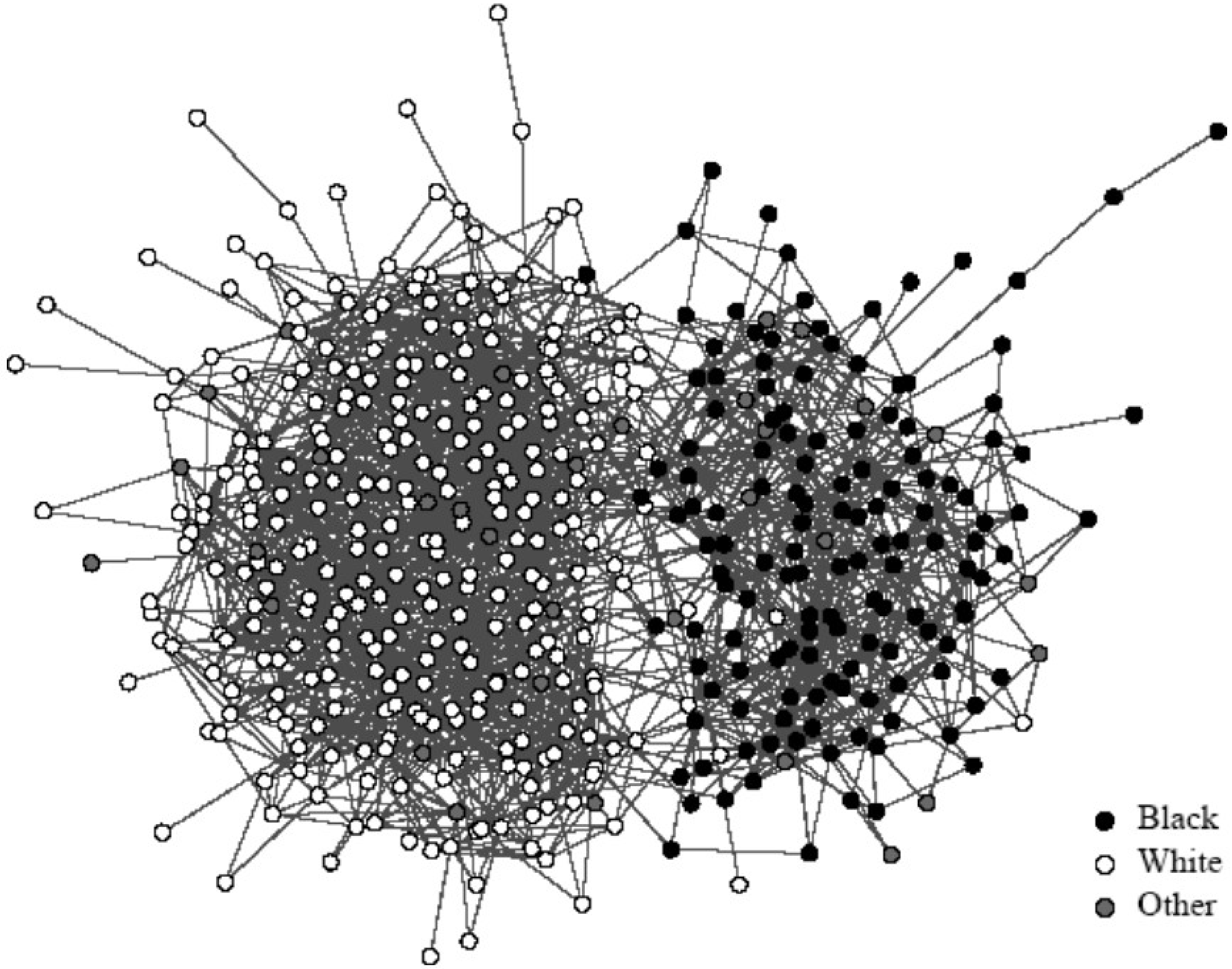
- Split power network into areas with minimum inter-area interactions



- Applications
  - Decide control areas for distributed power system state estimation
  - Parallel computation of power flow
  - Controlled islanding to prevent spreading of blackouts

High School Students

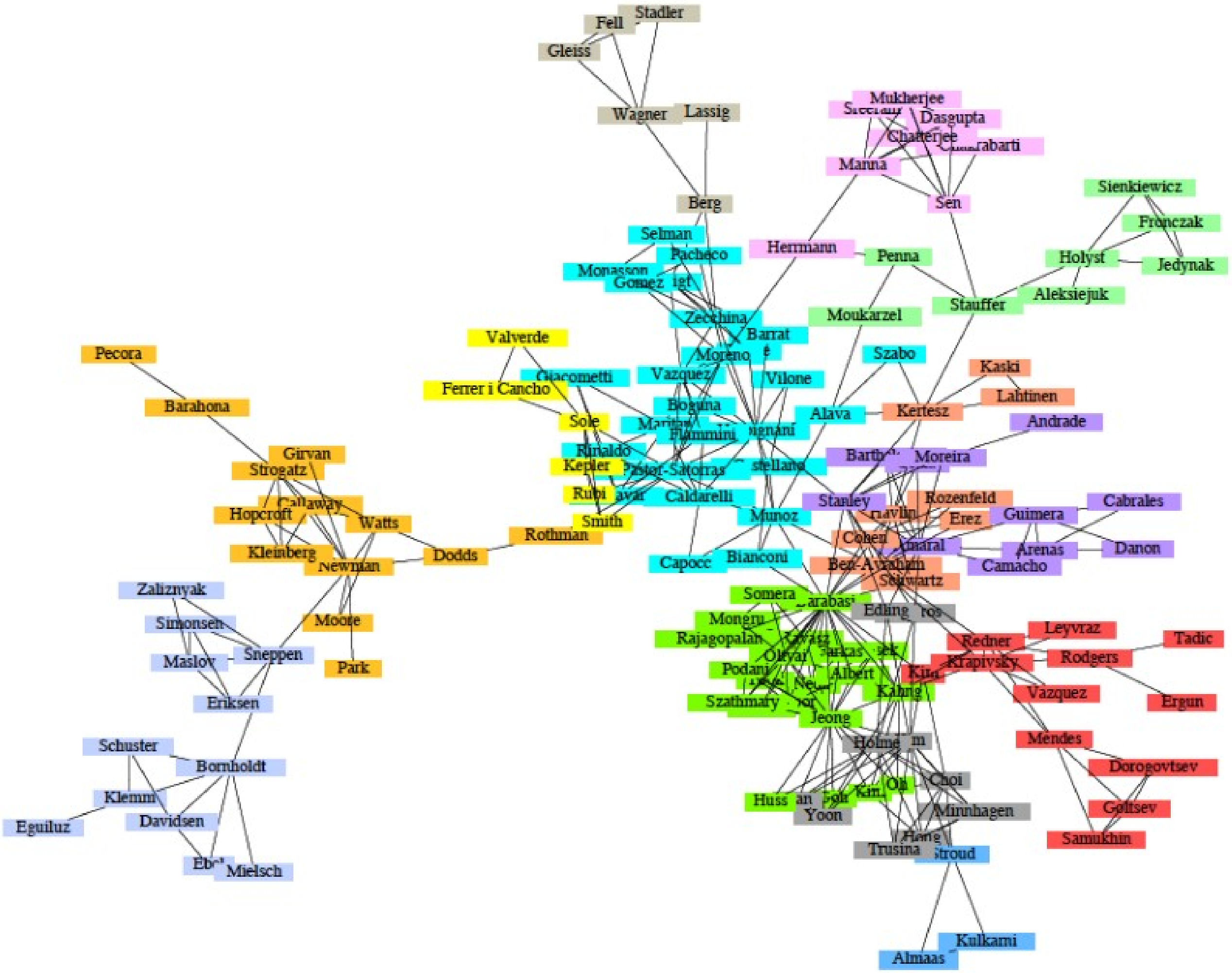
- Network of social interactions among high-school students



- Strong **assortative mixing**, with race as latent characteristic

Physicists working on Network Science

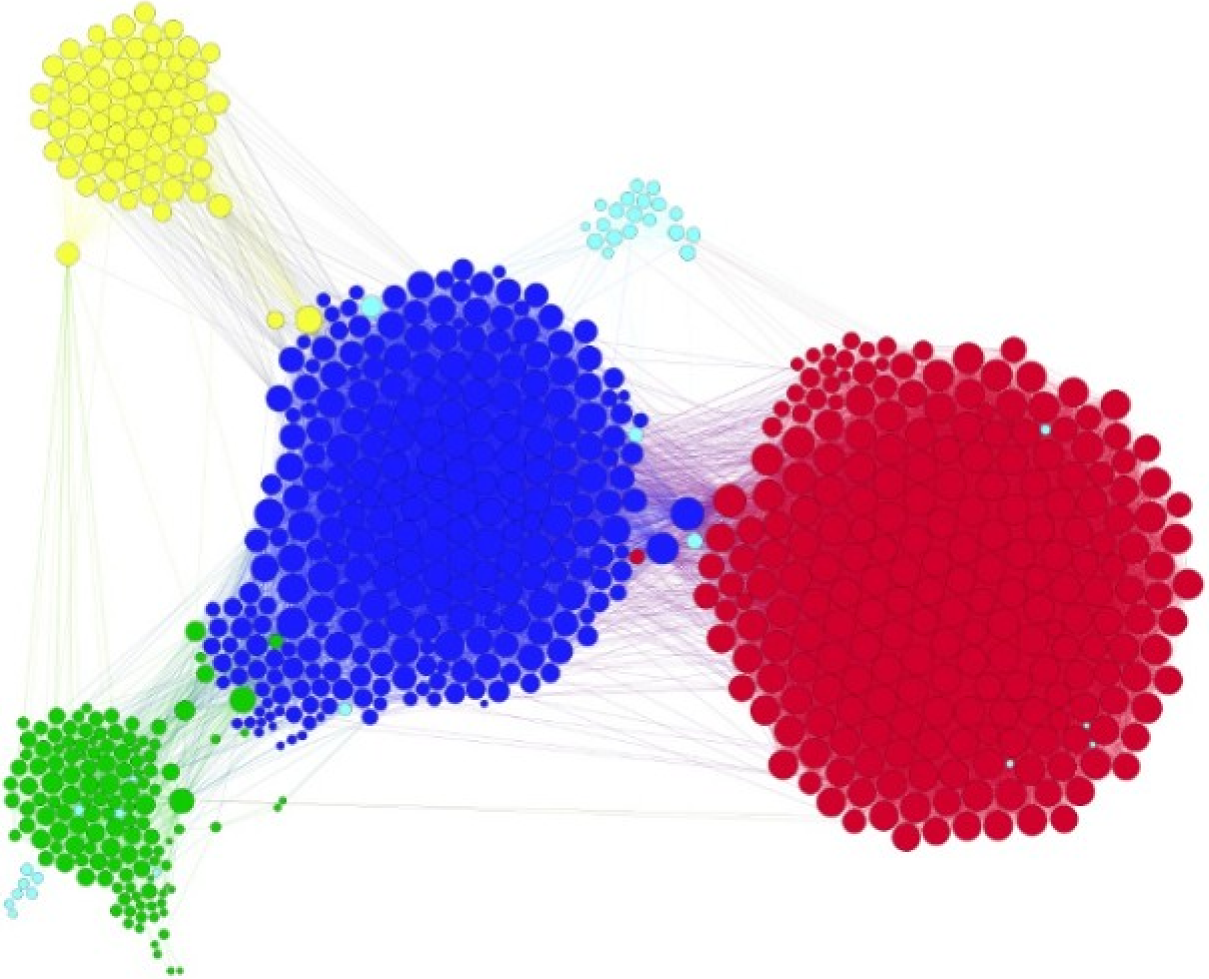
- Coauthorship network of physicists publishing networks' research



- Tightly-knit subgroups are evident from the network structure

Facebook Friendships

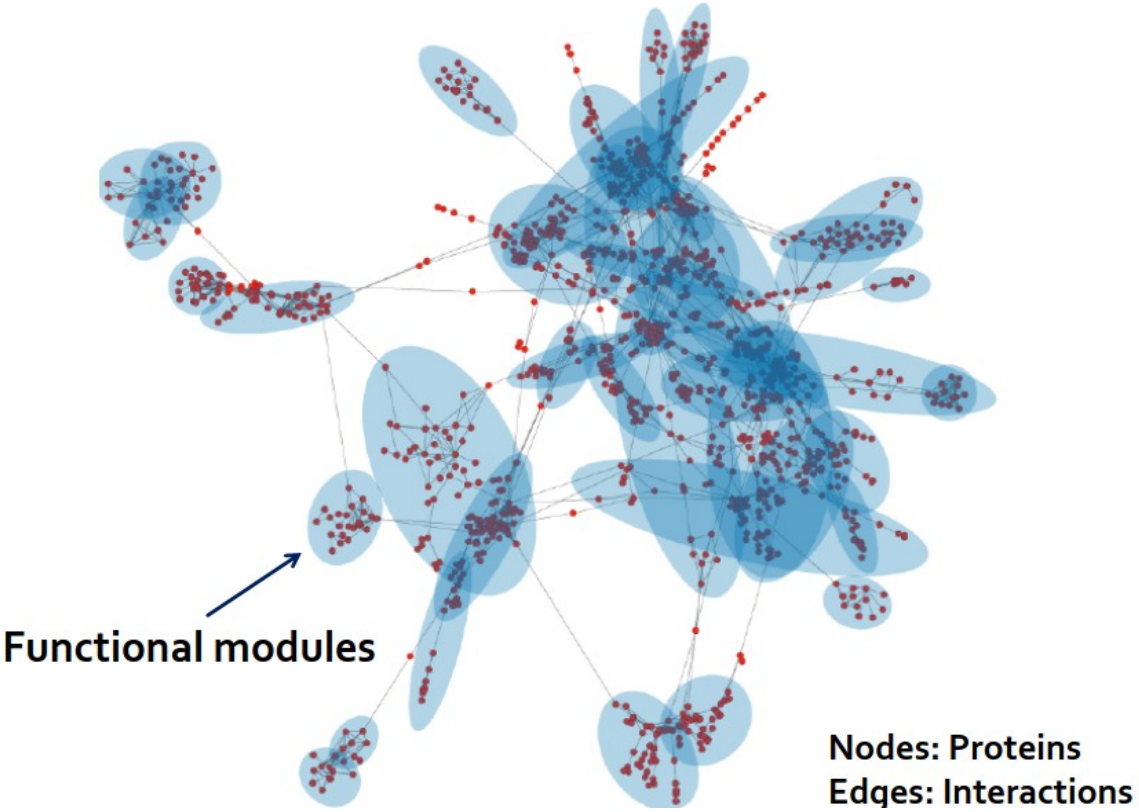
- Facebook egonet with 744 vertices and 30K edges



- Asked "ego" to identify social circles to which friends belong
  - Company, high-school, basketball club, squash club, family

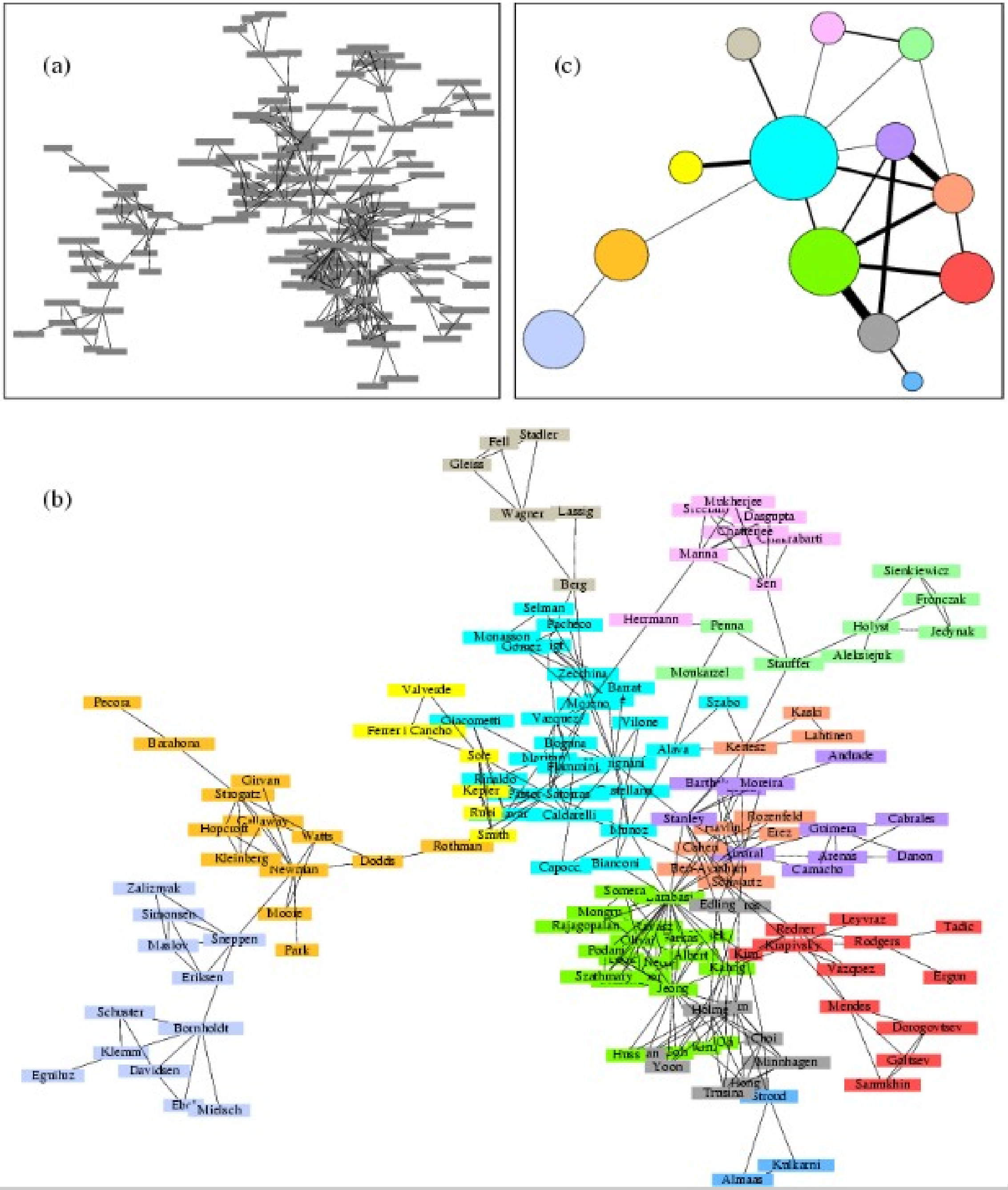
Why do it: gain understanding

- Gain understanding of networks
  - Discover communities in practice
  - Measure isolation of groups
  - Understand opinion dynamics / adoption
  - ...



Why do it: visualize

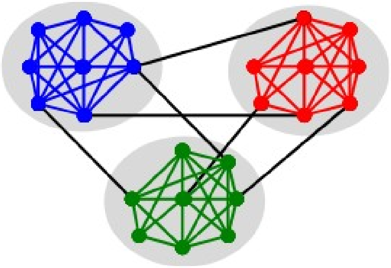
- Communities help to “aggregate” network data



Community Discovery

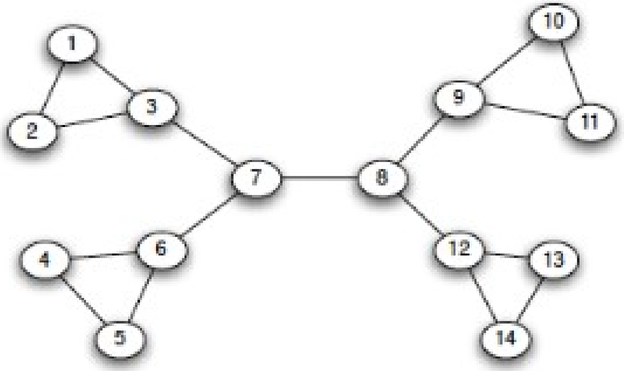
- Community discovery is a challenging computational problem
  - No consensus on the structural definition of community
  - Node subset selection often intractable
  - Lack of ground-truth for validation

- Number and sizes of groups most often unspecified in community detection
  - Identify the *natural fault lines* along which a network separates

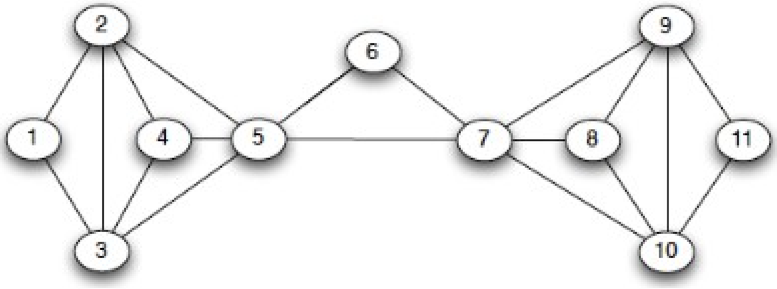


Bridges and Communities

- Local bridges connect weakly interacting parts of the network



- What about removing those to reveal communities?

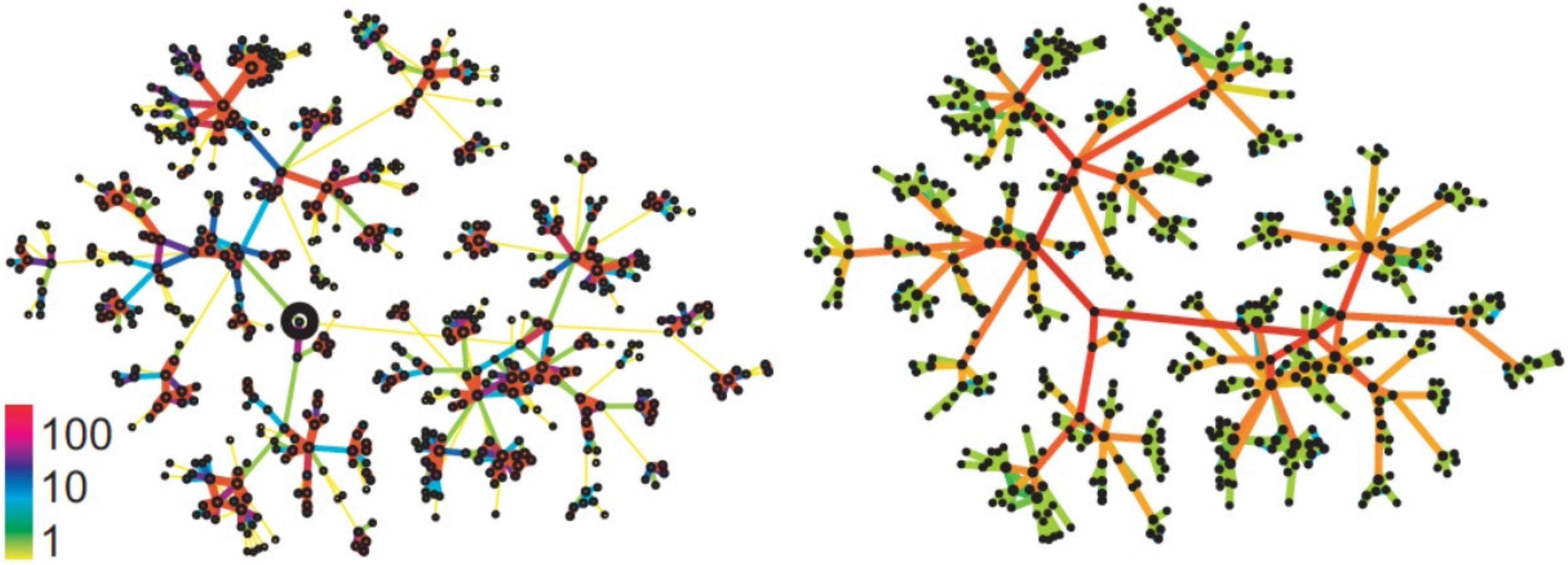


- Some challenges
  - Multiple local bridges. Some better that others? Which one first?
  - There might be no local bridge, yet an apparent natural division

Edge Betweenness Centrality

- Idea: high edge betweenness centrality to identify weak ties
  - High $c_{Be}(e)$ edges carry large traffic volume over shortest paths
  - Position at the interface between tightly-knit groups

- Ex: cell-phone network with colored edge strength and betwenness



Edge strength

Edge betweenness

Girvan-Newman's method

- Girvan-Newmann's method extremely simple conceptually

  ⇒ Find and remove "spanning links" between cohesive subgroups

- **Algorithm:** Repeat until there are no edges left

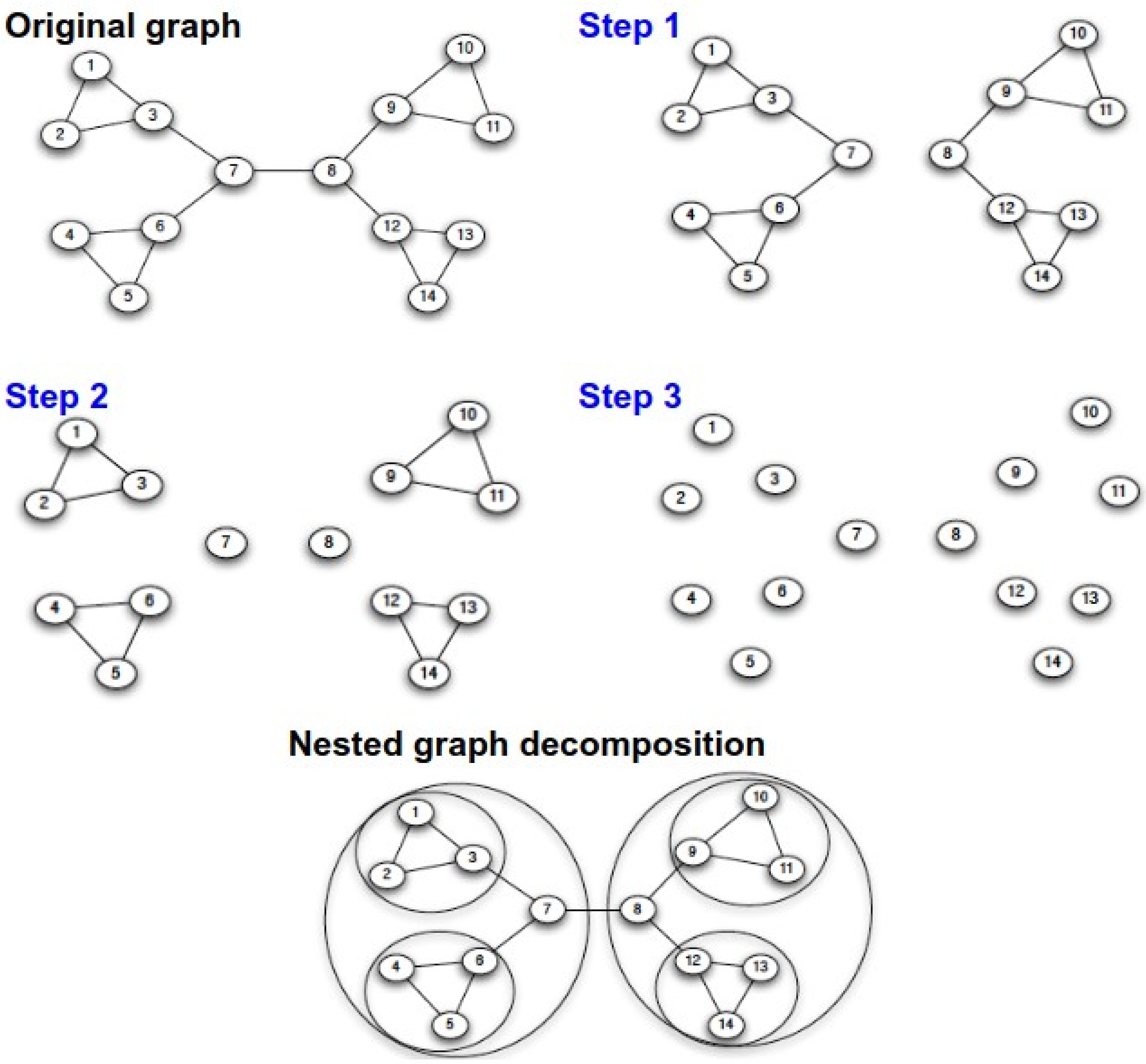
  ⇒ Calculate the betweenness centrality $c_{Be}(e)$ of all edges

  ⇒ Remove edge(s) with highest $c_{Be}(e)$

- Connected components are the communities identified
  - Divisive method: network falls apart into pieces as we go
  - Nested partition: larger communities potentially host denser groups
  - Recompute edge betweenness in $O(N_v N_e)$-time per step

- M. Girvan and M. Newman, "Community structure in social and biological networks," *PNAS*, vol. 99, pp. 7821-7826, 2002

Girvan-Newman in action



Original graph

Step 1

Step 2

Step 3

Nested graph decomposition

Hierarchical Clustering

- Greedy approach to iteratively modify successive candidate partitions
  - Agglomerative: successive coarsening of partitions through merging
  - Divisive: successive refinement of partitions through splitting
- Per step, partitions are modified in a way that minimizes a cost
  - Measures of (dis)similarity x_{ij} between pairs of vertices v_i and v_j
  - Ex: Euclidean distance dissimilarity

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

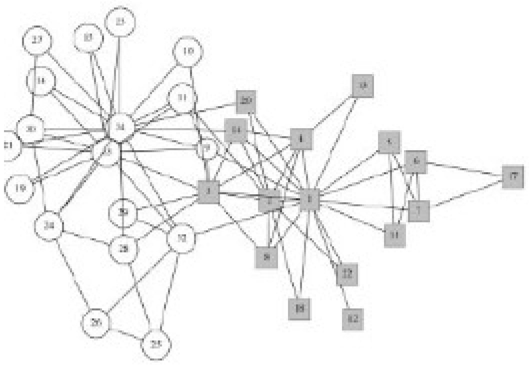
- Method returns an entire hierarchy of nested partitions of the graph

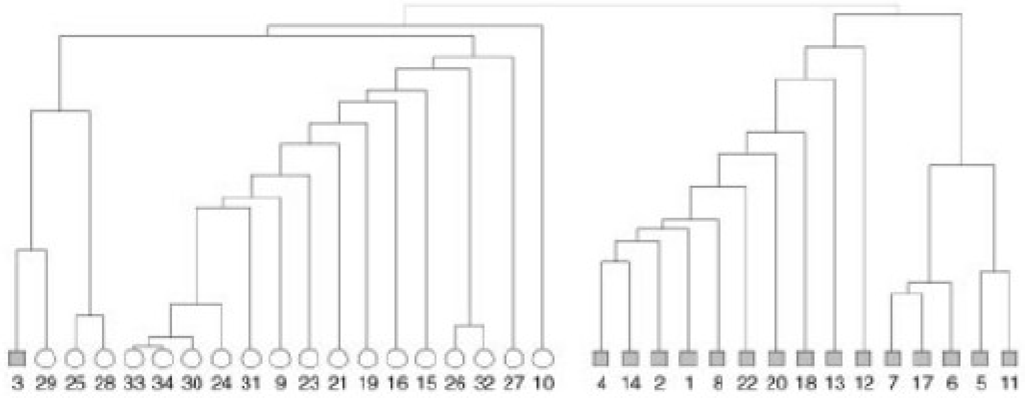
  $\Rightarrow$ Can range fully from $\{\{v_1\}, \ldots, \{v_{N_v}\}\}$ to V

Hierarchical Clustering and Dendograms

- Hierarchical partitions often represented with a dendrogram
- Shows groups found in the network at all algorithmic steps

  ⇒ Split the network at different resolutions
- Ex: Girvan-Newman's algorithm for the Zachary's karate club



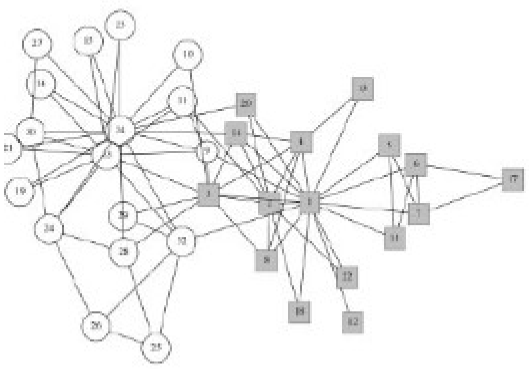


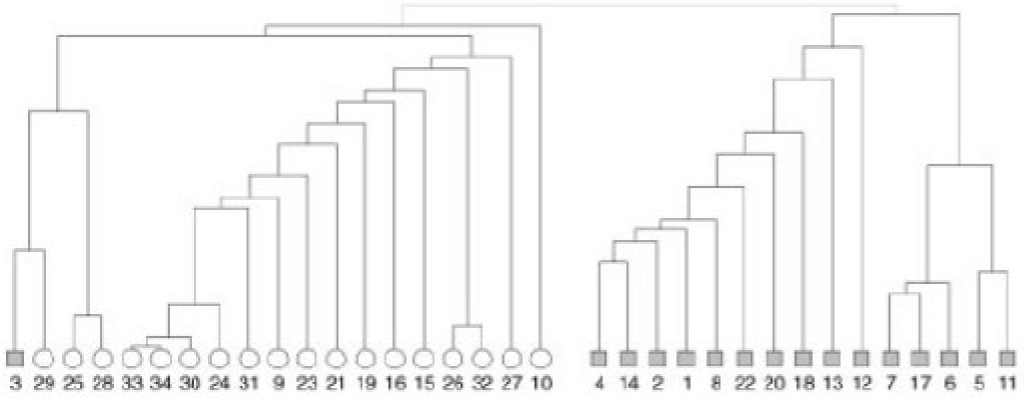
- Q: Which of the divisions is the most useful/optimal in some sense?
- A: Need to define metrics of graph clustering quality

Hierarchical Clustering and Dendograms

- Hierarchical partitions often represented with a dendrogram
- Shows groups found in the network at all algorithmic steps

  ⇒ Split the network at different resolutions
- Ex: Girvan-Newman's algorithm for the Zachary's karate club





- Q: Which of the divisions is the most useful/optimal in some sense?
- A: Need to define metrics of graph clustering quality

Modularity

- Size of communities typically unknown ⇒ Identify automatically
- Modularity measures how well a network is partitioned in communities
  - Intuition: density of edges in communities higher than expected
- Consider a graph G and a partition into groups $s \in S$. Modularity:

$$Q(G,S) \propto \sum_{s \in S} [(\#\text{ of edges within group } s) - \mathbb{E}[\#\text{ of such edges}]]$$

- Formally, after normalization such that $Q(G,S) \in [-1,1]$

$$Q(G,S) = \frac{1}{2N_e} \sum_{s \in S} \sum_{i,j \in s} \left[A_{ij} - \frac{d_i d_j}{2N_e} \right]$$

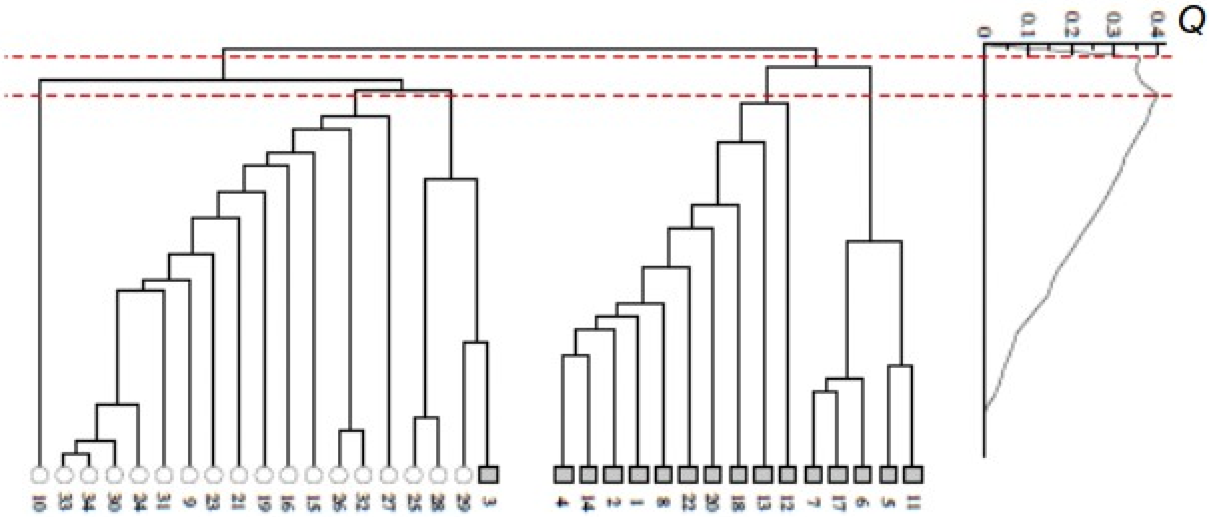
⇒ Null model: randomize edges, preserving degree distribution

Assessing clustering quality

- Can evaluate the modularity of each partition in a dendrogram

  ⇒ Maximum value gives the "best" community structure

- Ex: Girvan-Newman's algorithm for the Zachary's karate club



- Q: Why not optimize $Q(G, S)$ directly over possible partitions S?

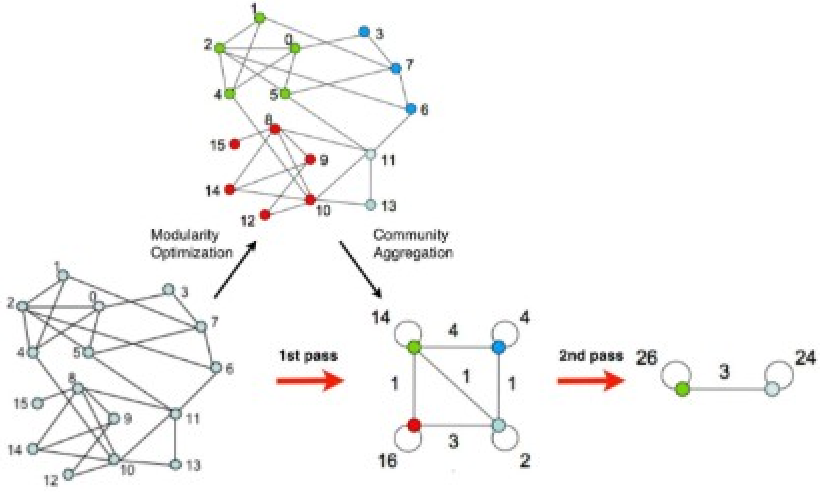
Louvain Algorithm

- **Greedy algorithm** for community detection
  - O(n log n) run time
- Supports weighted graphs
- Provides hierarchical partitions
- Widely utilized to **study large networks** because:
  - Fast
  - Rapid convergence properties
  - High modularity output (i.e., “better communities”)

Louvain Algorithm Overview

- Louvain algorithm **greedily** **maximizes** modularity
- **Each pass is made of 2 phases:**
  - **Phase 1:** Modularity is optimized by allowing only local changes of communities
  - **Phase 2:** The identified communities are aggregated in order to build a new network of communities
  - **Goto Phase 1**

The passes are repeated **iteratively** until no increase of modularity is possible!



Louvain Algorithm: 1^{st} Phase

- Put each node in a graph into a distinct community (one node per community)
- For each node *i*, the algorithm performs two calculations:
  - Compute the modularity gain (ΔQ) when putting node i into the community of some neighbor j
  - Move i to a community of node j that yields the largest gain ΔQ
- The loop runs until no movement yields a gain

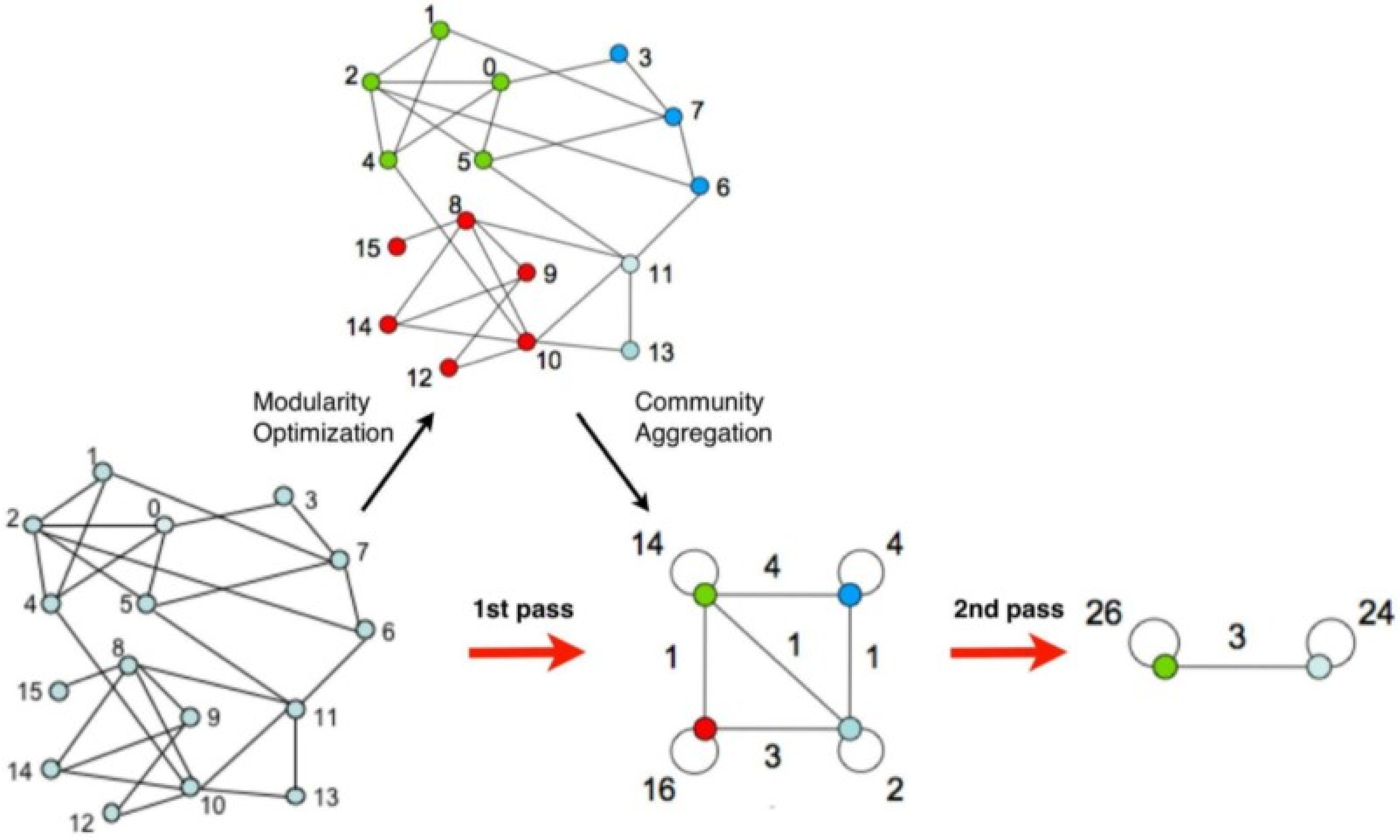
Louvain Algorithm: 2^{nd} Phase

- The partitions obtained in the first phase are contracted into **super-nodes**, and the network is created accordingly
  - Super-nodes are connected if there is at least one edge between nodes of the corresponding partitions
  - The weight of the edge between the two super-nodes is the sum of the weights from all edges between their corresponding partitions
- **The loop runs until the community configuration does not change anymore**

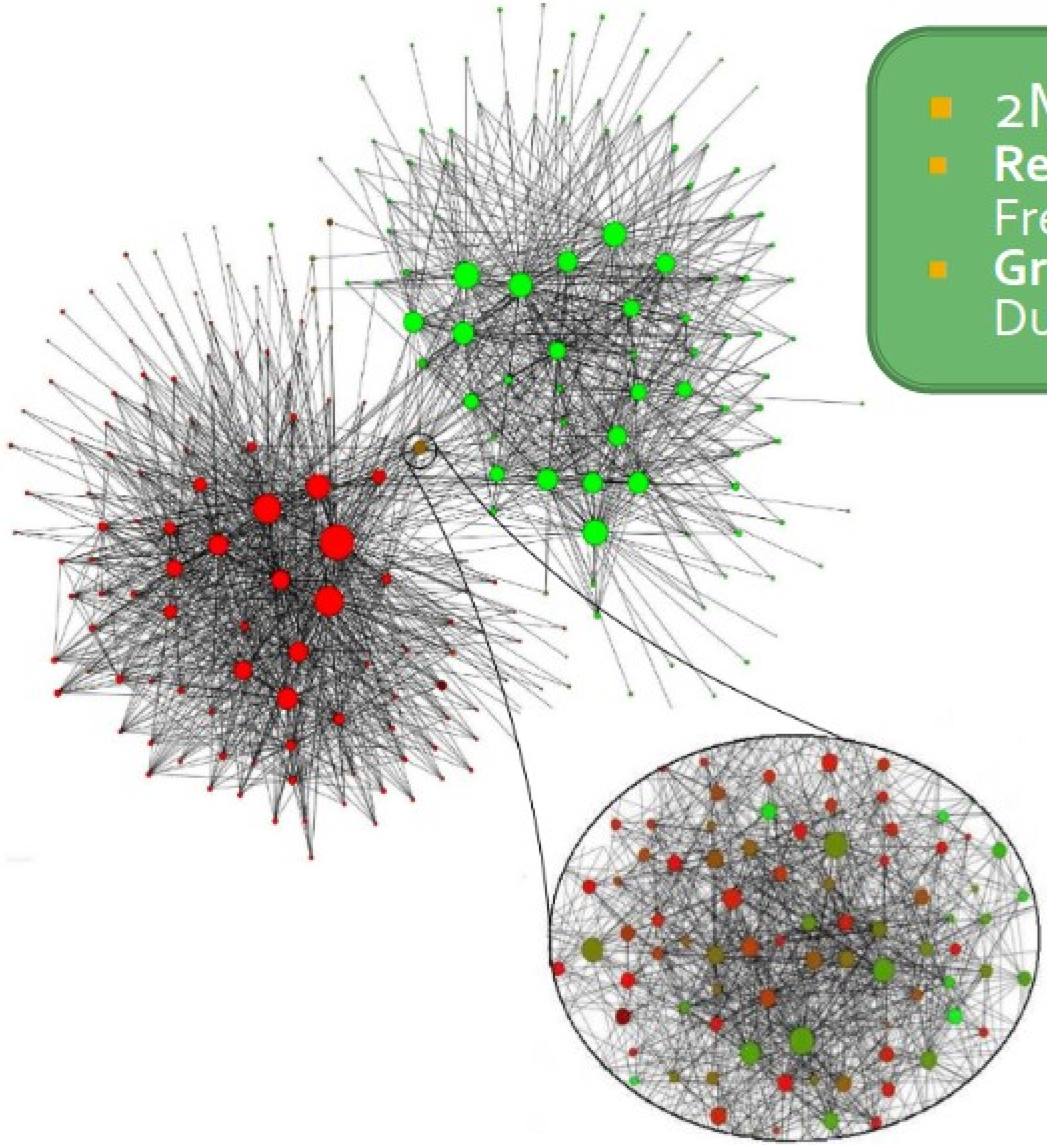
Louvain Algorithm: 2nd Phase

- The partitions obtained in the first phase are contracted into **super-nodes**, and the network is created accordingly
  - Super-nodes are connected if there is at least one edge between nodes of the corresponding partitions
  - The weight of the edge between the two super-nodes is the sum of the weights from all edges between their corresponding partitions
- **The loop runs until the community configuration does not change anymore**

Louvain Algorithm: 2nd Phase

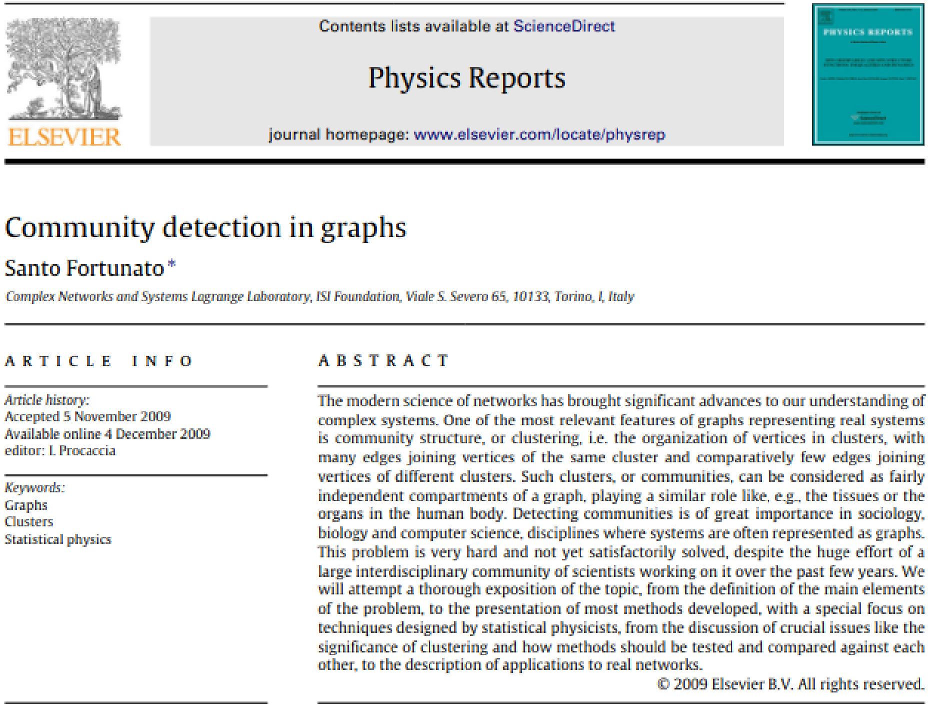


Belgian Phone Network



- 2M nodes
- **Red nodes**: French speakers
- **Green nodes**: Dutch speakers

Community Discovery Algorithms



Contents lists available at ScienceDirect

Physics Reports

journal homepage: www.elsevier.com/locate/physrep

ELSEVIER

Community detection in graphs

Santo Fortunato*

Complex Networks and Systems Lagrange Laboratory, ISI Foundation, Viale S. Severo 65, 10133, Torino, I, Italy

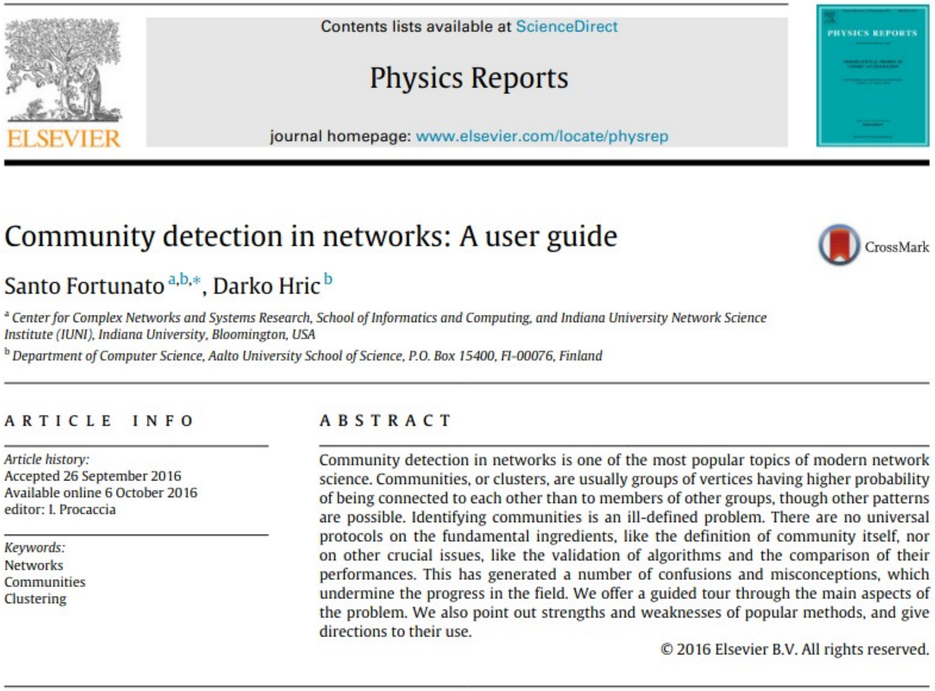
ARTICLE INFO

Article history:
Accepted 5 November 2009
Available online 4 December 2009
editor: I. Procaccia

Keywords:
Graphs
Clusters
Statistical physics

ABSTRACT

The modern science of networks has brought significant advances to our understanding of complex systems. One of the most relevant features of graphs representing real systems is community structure, or clustering, i.e. the organization of vertices in clusters, with many edges joining vertices of the same cluster and comparatively few edges joining vertices of different clusters. Such clusters, or communities, can be considered as fairly independent compartments of a graph, playing a similar role like, e.g., the tissues or the organs in the human body. Detecting communities is of great importance in sociology, biology and computer science, disciplines where systems are often represented as graphs. This problem is very hard and not yet satisfactorily solved, despite the huge effort of a large interdisciplinary community of scientists working on it over the past few years. We will attempt a thorough exposition of the topic, from the definition of the main elements of the problem, to the presentation of most methods developed, with a special focus on techniques designed by statistical physicists, from the discussion of crucial issues like the significance of clustering and how methods should be tested and compared against each other, to the description of applications to real networks.



Contents lists available at ScienceDirect

Physics Reports

journal homepage: www.elsevier.com/locate/physrep

ELSEVIER

Community detection in networks: A user guide

CrossMark

Santo Fortunato[a,b,*], Darko Hric[b]

[a] *Center for Complex Networks and Systems Research, School of Informatics and Computing, and Indiana University Network Science Institute (IUNI), Indiana University, Bloomington, USA*

[b] *Department of Computer Science, Aalto University School of Science, P.O. Box 15400, FI-00076, Finland*

ARTICLE INFO

Article history:
Accepted 26 September 2016
Available online 6 October 2016
editor: I. Procaccia

Keywords:
Networks
Communities
Clustering

ABSTRACT

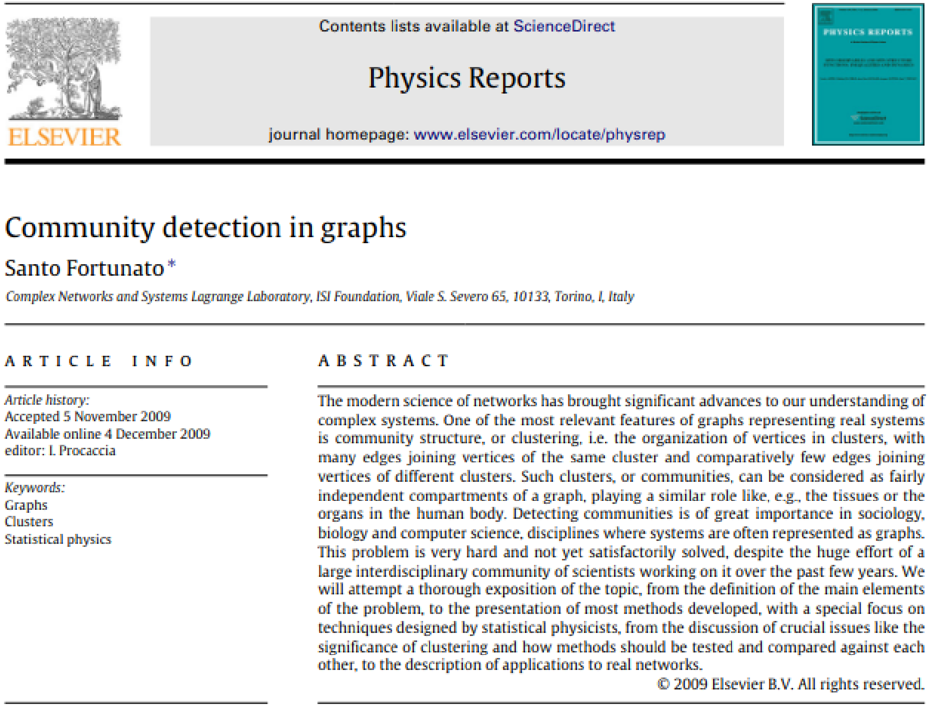
Community detection in networks is one of the most popular topics of modern network science. Communities, or clusters, are usually groups of vertices having higher probability of being connected to each other than to members of other groups, though other patterns are possible. Identifying communities is an ill-defined problem. There are no universal protocols on the fundamental ingredients, like the definition of community itself, nor on other crucial issues, like the validation of algorithms and the comparison of their performances. This has generated a number of confusions and misconceptions, which undermine the progress in the field. We offer a guided tour through the main aspects of the problem. We also point out strengths and weaknesses of popular methods, and give directions to their use.

https://doi.org/10.1016/j.physrep.2009.11.002

https://doi.org/10.1016/j.physrep.2016.09.002

There are many possible algorithms and definitions

Community Discovery Algorithms



Contents lists available at ScienceDirect

Physics Reports

journal homepage: www.elsevier.com/locate/physrep

Community detection in graphs

Santo Fortunato*

Complex Networks and Systems Lagrange Laboratory, ISI Foundation, Viale S. Severo 65, 10133, Torino, I, Italy

ARTICLE INFO

Article history:
Accepted 5 November 2009
Available online 4 December 2009
editor: I. Procaccia

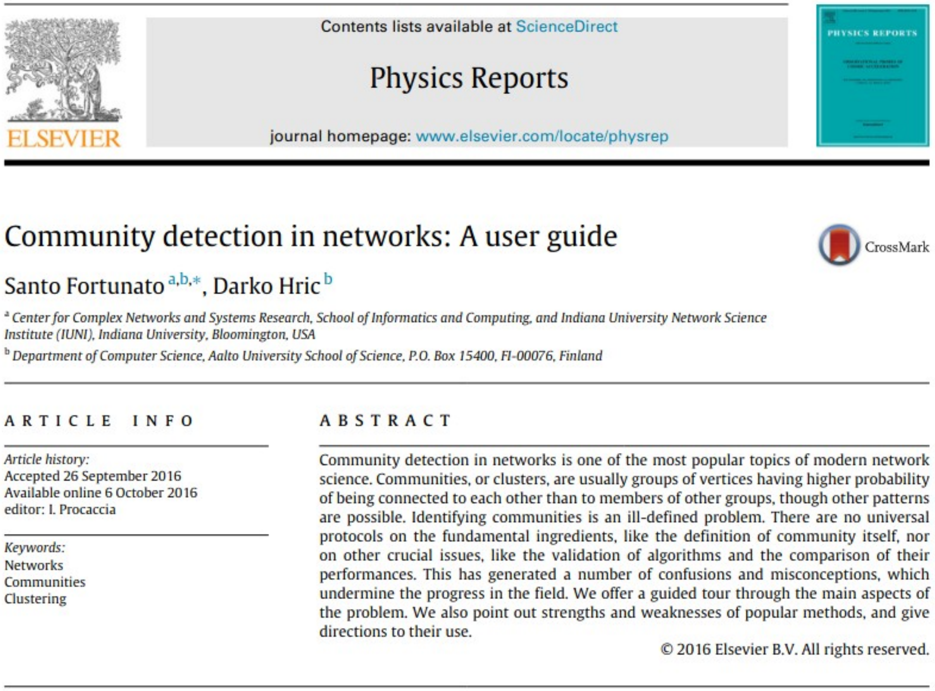
Keywords:
Graphs
Clusters
Statistical physics

ABSTRACT

The modern science of networks has brought significant advances to our understanding of complex systems. One of the most relevant features of graphs representing real systems is community structure, or clustering, i.e. the organization of vertices in clusters, with many edges joining vertices of the same cluster and comparatively few edges joining vertices of different clusters. Such clusters, or communities, can be considered as fairly independent compartments of a graph, playing a similar role like, e.g., the tissues or the organs in the human body. Detecting communities is of great importance in sociology, biology and computer science, disciplines where systems are often represented as graphs. This problem is very hard and not yet satisfactorily solved, despite the huge effort of a large interdisciplinary community of scientists working on it over the past few years. We will attempt a thorough exposition of the topic, from the definition of the main elements of the problem, to the presentation of most methods developed, with a special focus on techniques designed by statistical physicists, from the discussion of crucial issues like the significance of clustering and how methods should be tested and compared against each other, to the description of applications to real networks.

© 2009 Elsevier B.V. All rights reserved.

https://doi.org/10.1016/j.physrep.2009.11.002



Contents lists available at ScienceDirect

Physics Reports

journal homepage: www.elsevier.com/locate/physrep

Community detection in networks: A user guide

CrossMark

Santo Fortunato[a,b,*], Darko Hric[b]

[a] *Center for Complex Networks and Systems Research, School of Informatics and Computing, and Indiana University Network Science Institute (IUNI), Indiana University, Bloomington, USA*

[b] *Department of Computer Science, Aalto University School of Science, P.O. Box 15400, FI-00076, Finland*

ARTICLE INFO

Article history:
Accepted 26 September 2016
Available online 6 October 2016
editor: I. Procaccia

Keywords:
Networks
Communities
Clustering

ABSTRACT

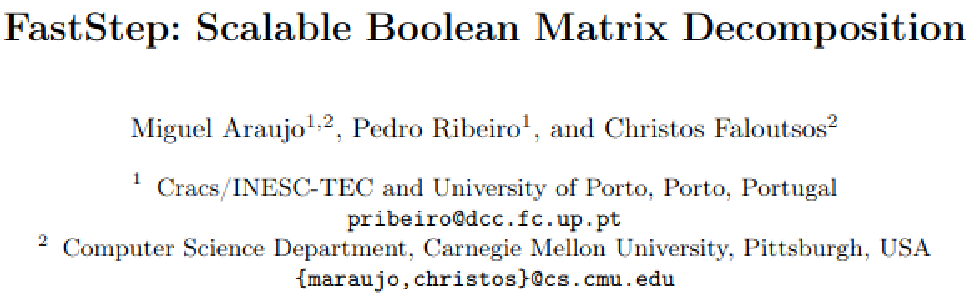
Community detection in networks is one of the most popular topics of modern network science. Communities, or clusters, are usually groups of vertices having higher probability of being connected to each other than to members of other groups, though other patterns are possible. Identifying communities is an ill-defined problem. There are no universal protocols on the fundamental ingredients, like the definition of community itself, nor on other crucial issues, like the validation of algorithms and the comparison of their performances. This has generated a number of confusions and misconceptions, which undermine the progress in the field. We offer a guided tour through the main aspects of the problem. We also point out strengths and weaknesses of popular methods, and give directions to their use.

© 2016 Elsevier B.V. All rights reserved.

https://doi.org/10.1016/j.physrep.2016.09.002

There are many possible algorithms and definitions

A paper on community discovery



FastStep: Scalable Boolean Matrix Decomposition

Miguel Araujo[1,2], Pedro Ribeiro[1], and Christos Faloutsos[2]

[1] Cracs/INESC-TEC and University of Porto, Porto, Portugal
pribeiro@dcc.fc.up.pt
[2] Computer Science Department, Carnegie Mellon University, Pittsburgh, USA
{maraujo,christos}@cs.cmu.edu

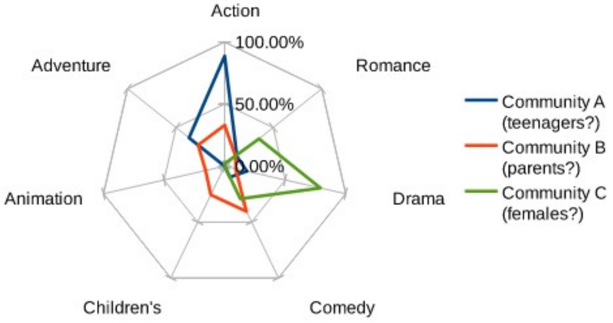
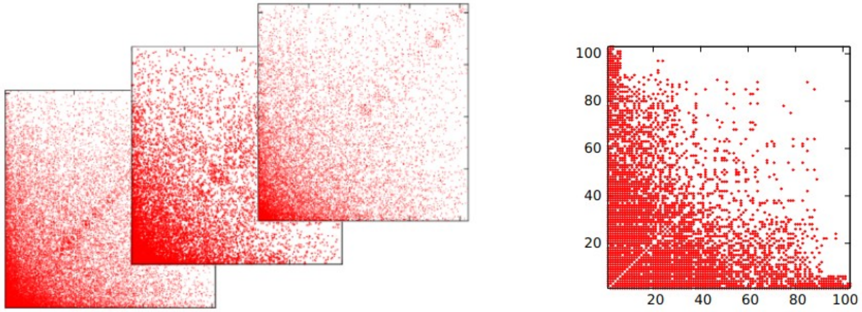


Fig. 5: **MovieLens genre separation**



(a) **DBLP real communities** - PAKDD, KDD and VLDB.

(b) **FastStep community** - A can airports.

Fig. 1: **Realistic hyperbolic structure -** Adjacency Matrices of real co nities in DBLP and a community found by FastStep.

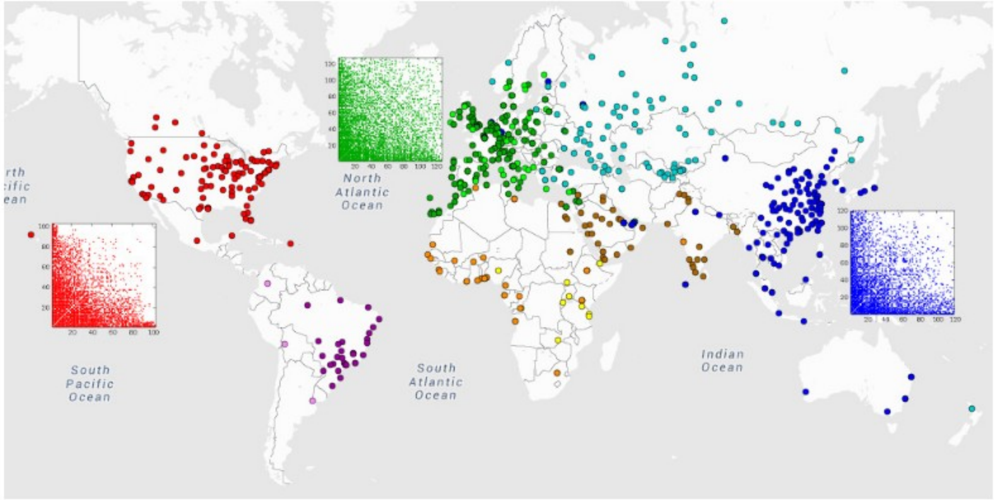


Fig. 2: **Intuitive non-block communities** - Communities automatically found in the Airports dataset from flight records. (best viewed in color)

Network Visualization and Exploration with Gephi

Why Visualization?

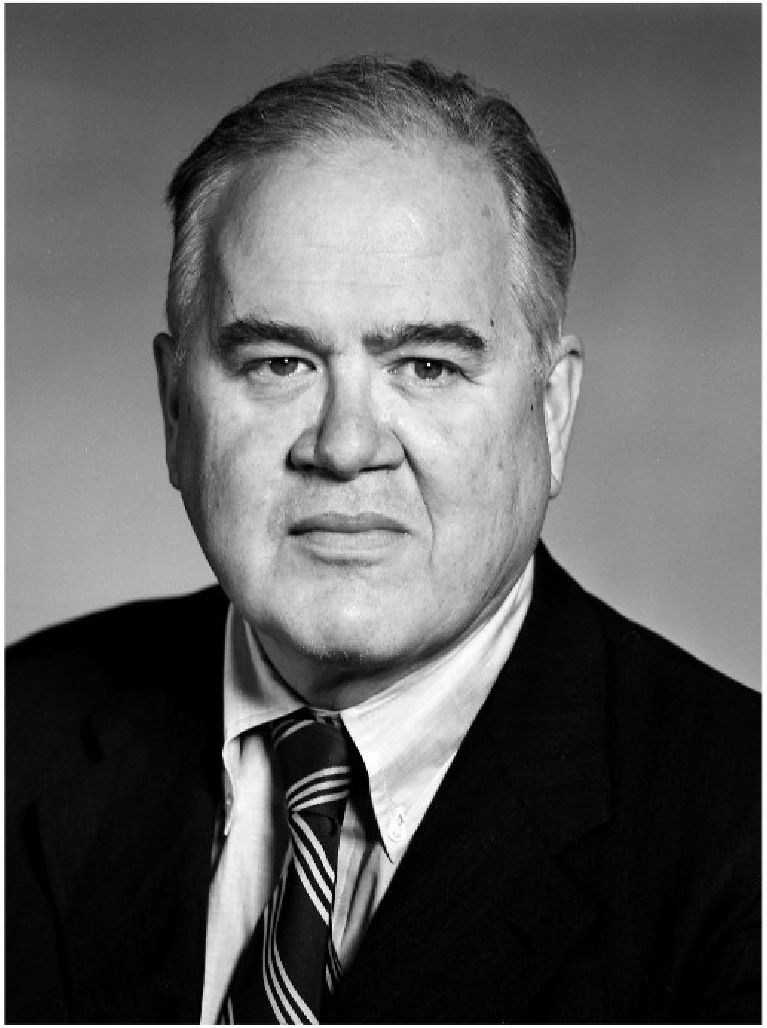


Photo Credit: Princeton University, Robert Matthews

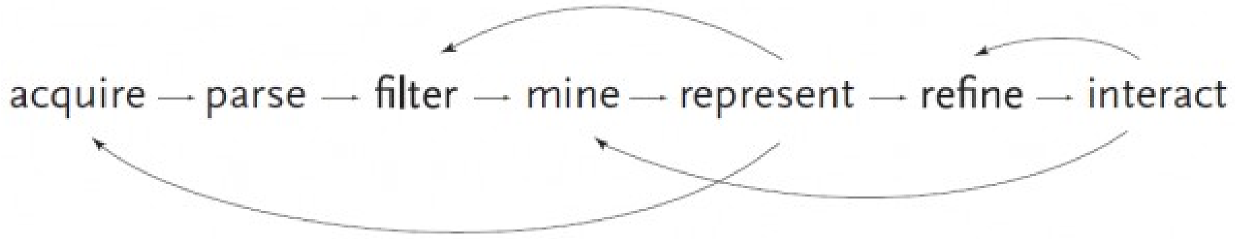
John W. Tukey

“The greatest value of a picture is when it forces to notice what we never expected to see”

Exploratory Data Analysis

- **Visualization alone is not enough**
  - Part of a larger process to extract insight
- **Data process chain**

acquire	parse	filter	mine	represent	refine	interact
live or changing data sources	modular parsers for new data sources	automation of tedious manual processes modify filter in real-time	modify parameters of statistical methods in real-time	rapid prototyping and iteration juxtapose large amounts of data try multiple representations	change design rules without manual redesign computation as its own "medium"	smooth transition between states to maintain context additional information as viewpoint shifts

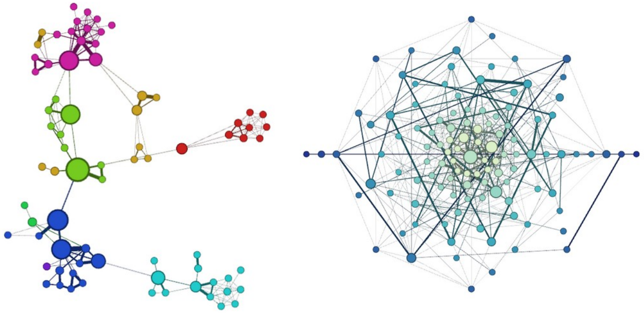

acquire → parse → filter → mine → represent → refine → interact

Non-linear
Trial and Error!

Images: Ben Fry, 2004

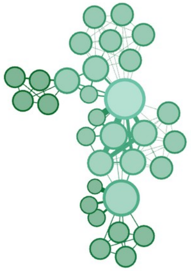
Exploring a Network

- **1) See the network**
  - Draw using a certain layout, .



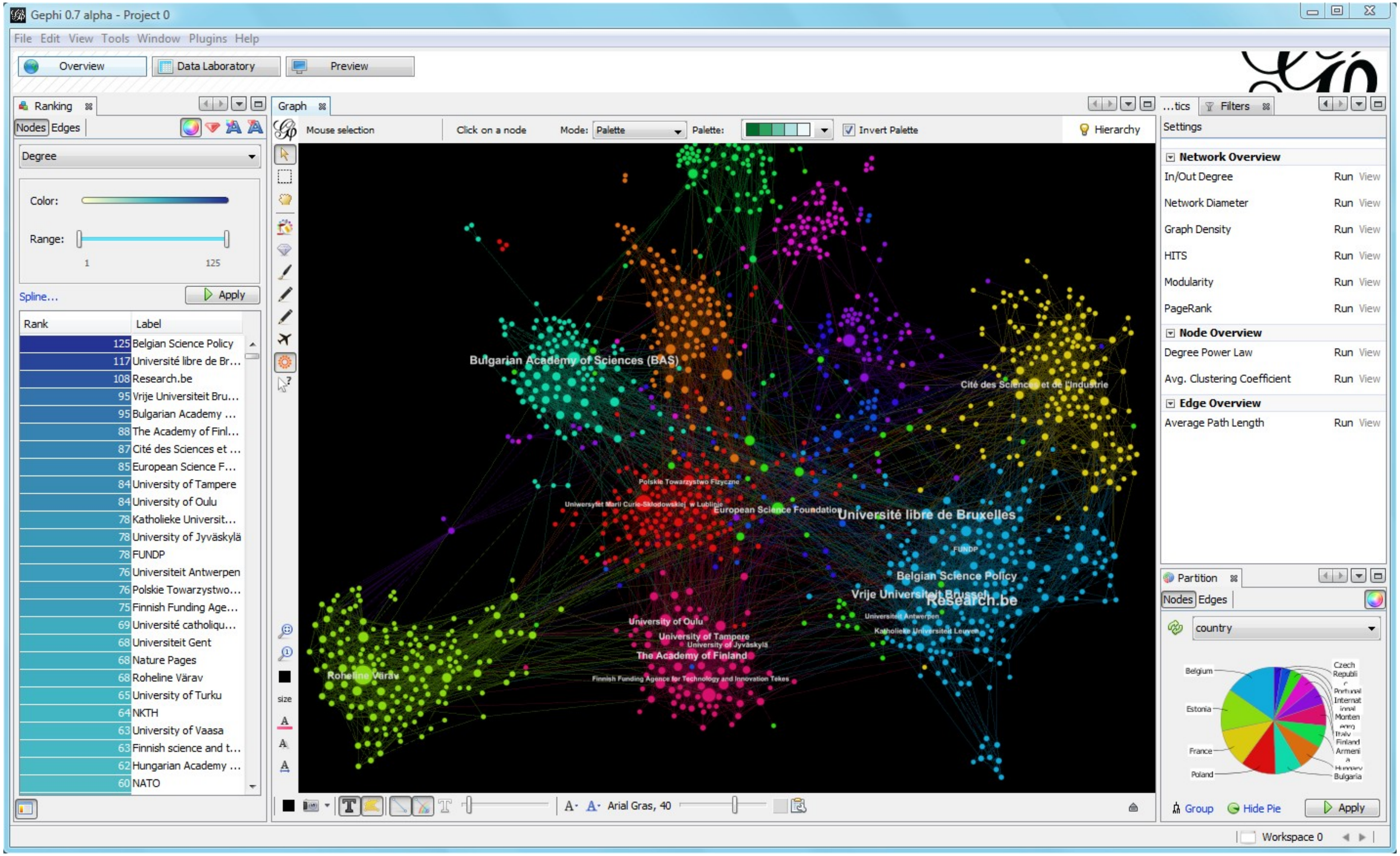
- **2) Interact in real time**
  - Group, filter, compute metrics,

- **3) Build a visual language**
  - Size of nodes, thickness of edge colors, ...



Exploring Graphs

- **Today we are going to use** Gephi
  - Open-Source Network **Analysis** and **Visualization** Platform (written in Java)



Rank	Label
125	Belgian Science Policy
117	Université libre de Br...
108	Research.be
95	Vrije Universiteit Bru...
95	Bulgarian Academy ...
88	The Academy of Finl...
87	Cité des Sciences et ...
85	European Science F...
84	University of Tampere
84	University of Oulu
78	Katholieke Universit...
78	University of Jyväskylä
78	FUNDP
76	Universiteit Antwerpen
76	Polskie Towarzystwo...
75	Finnish Funding Age...
69	Université catholiqu...
68	Universiteit Gent
68	Nature Pages
68	Roheline Värav
65	University of Turku
64	NKTH
63	University of Vaasa
63	Finnish science and t...
62	Hungarian Academy ...
60	NATO

Why Gephi?

- **Because it has a large community**
- **Because it has history (and will continue to have)**
  - Started at 1998
  - Maintained by a consortium (long-term vision)
- **Because it is extensible with plugins**
  - Gephi marketplace

Gephi

- **Limitations:**
  - Still in beta version so there are a few rough edges
  - Not prepared to handle very large networks (depending on the infrastructure – RAM mostly - can manage up to 100 000 nodes)

- **There are other options:**
  - The main concepts and ideas we will show can be used on any other visualization tool

Gravisto
Graph Visualization Toolkit

Pajek
and
Pajek-XXL
Programs for Analysis
and Visualization
of Very Large Networks
Reference Manual
List of commands with short explanation

Cytoscape

NODEXL

LINKURIOUS
VISUALIZE GRAPH DATA EASILY

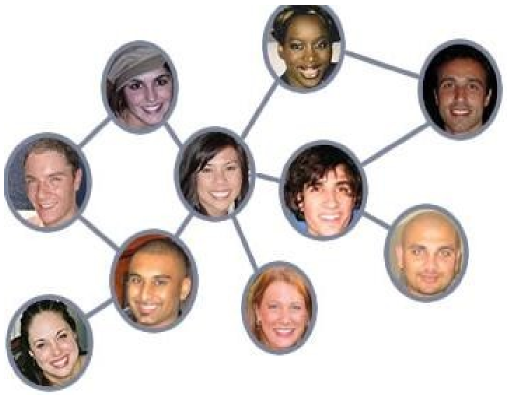
Goals of this activity

- **Consolidate the main concepts** and techniques learned by **performing a network analysis** of a real world network

- Specific Goals:
  - Perform an empirical analysis of the network
  - Loading Networks (opening, importing raw data, ...)
  - Computing Metrics (centralities, degrees, distances, communities, ...)
  - Filtering (main operators, selecting, ranges, combining, ...)
  - Create a clear and simple to understand visualisation of the network (color or size of the nodes and edges according to a metric or partition, ...)

Facebook Network

- My own facebook (ego) network – facebook.gephi
  - Nodes are users and links represent friendships (undirected graph)
  - Ego Network: all nodes connected to me and their connections (without myself)
  - Collected automatically (there used to be plugins for that)
  - 356 persons, 4,365 connections



DEMO!

Flights Network

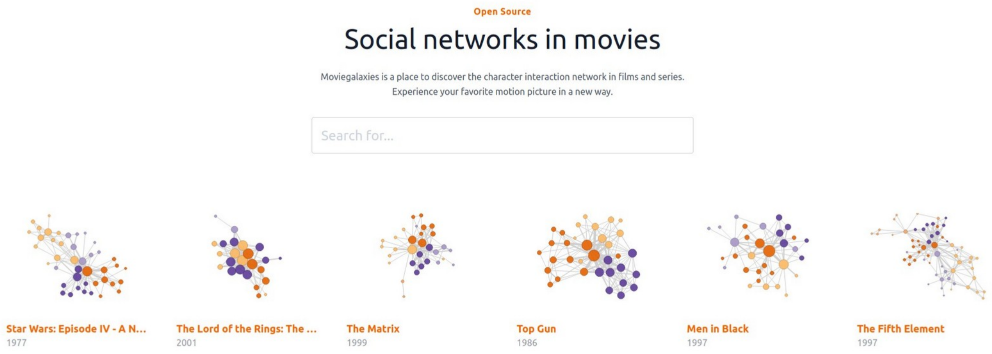
- Flights Data (OpenFights) – airports.csv and routes.csv
  - http://openflights.org/data.html
  - Compiled (also) by Open Flights website users
  - 3,154 airports, 66,500 routes from 538 airlines
  - Made for showing GeoLayout

OpenFlights.org

DEMO!

More Resources

- More datasets to toy with
  - Movie Galaxies (5 known movies in course resources)
    [Blade Runner] [Pulp Fiction] [The Godfather II] [Starwars IV] [LOTR: Return of the King]

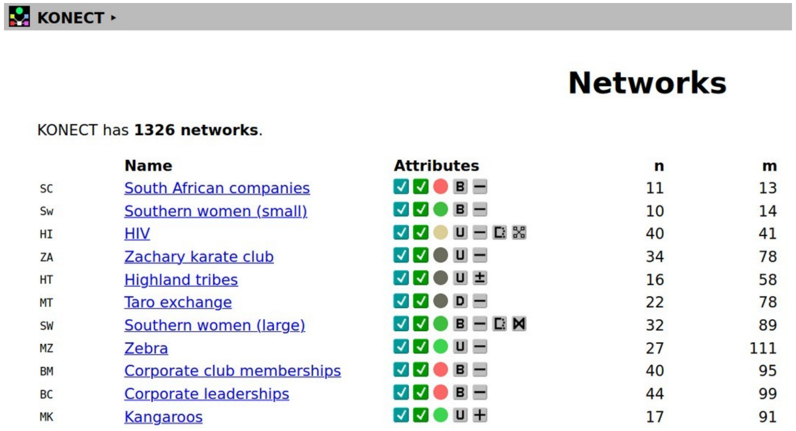


Open Source

Social networks in movies

Moviegalaxies is a place to discover the character interaction network in films and series.
Experience your favorite motion picture in a new way.

  - Konect project:
    http://konect.cc/
  - Network Data repository:
    https://networkrepository.com/



Networks

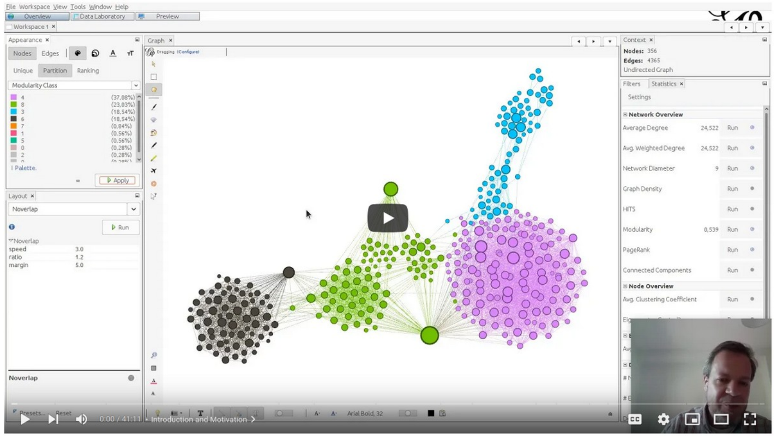
KONECT has **1326 networks**.

	Name	Attributes	n	m
SC	South African companies	✓ ✓ ● B −	11	13
Sw	Southern women (small)	✓ ✓ ● B −	10	14
HI	HIV	✓ ✓ ● U − ◫ ⁜	40	41
ZA	Zachary karate club	✓ ✓ ● U −	34	78
HT	Highland tribes	✓ ✓ ● U ±	16	58
MT	Taro exchange	✓ ✓ ● D −	22	78
SW	Southern women (large)	✓ ✓ ● B − ◫ ⋈	32	89
MZ	Zebra	✓ ✓ ● U −	27	111
BM	Corporate club memberships	✓ ✓ ● B −	40	95
BC	Corporate leaderships	✓ ✓ ● B −	44	99
MK	Kangaroos	✓ ✓ ● U +	17	91

More Resources

- Tutorial in Video (by myself)

  https://youtu.be/rnCTpzY2xUM



- Other Tutorials
  - https://gephi.org/users/

Official Tutorials

Gephi is really easy to handle if you learn the basics. Let's follow these tutorials to quickly manage the main features!

- How to Import Spreadsheet (Excel) Data // video
- How to Import Dynamic Data

  - https://github.com/kateto/Gephi-0.9-Tutorial/
  - http://www.martingrandjean.ch/gephi-introduction/