UNIVERSITÀ DEGLI STUDI DI BARI ALDO MORO

Social Network Analysis

Slides partially based upon M. Tvesovat's slides http://www.slideshare.net/maksim2042

Outline

- Introduction
- Graph theory
- Small world
- Centrality
- Community detection

INTRODUCTION

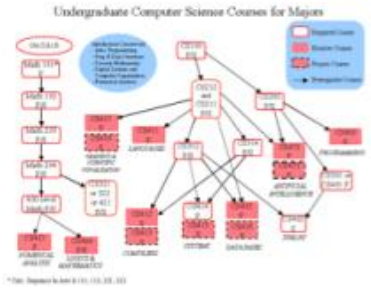
Prof. Filippo Lanubile

What is Network Analysis?

A quantitative studies of relationships (edges) among entities in the network (nodes)

(a) *Airline routes*

(b) *Subway map*



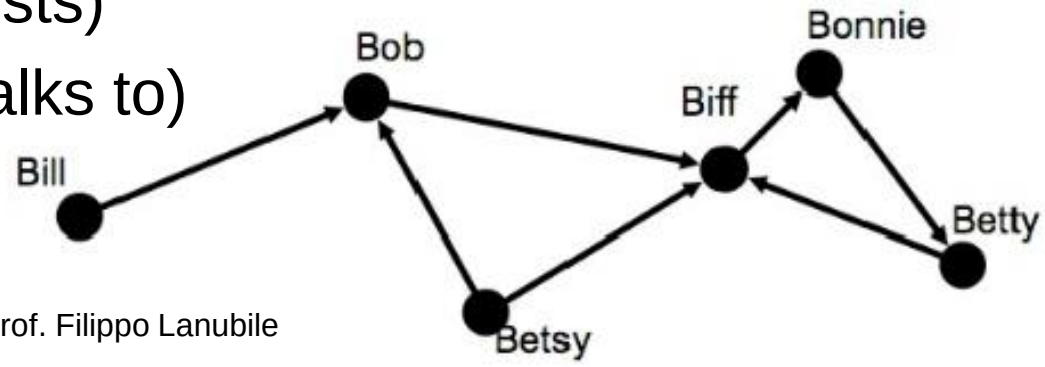
(c) *Flowchart of college courses*

(d) *Tank Street Bridge in Brisbane*

What is Social Network Analysis?

A quantitative studies of relationships (edges) among entities in the network (nodes) where

- Nodes represent *people* or *groups*
- Edges represent some form of *social interaction*
  - *Kinship* (Mother of, father of, sibling of)
  - *Role-based* (Boss of, teacher of, Friend of)
  - *Affective* (Likes, trusts)
  - *Communication* (Talks to)



Why Social Network Analysis?

- The explosion of research on large-scale networks in recent years has been fueled to a large extent by the increasing availability of large, detailed network datasets
  - Online social networking platforms
  - Instant messaging

Who-talks-to-Whom graphs

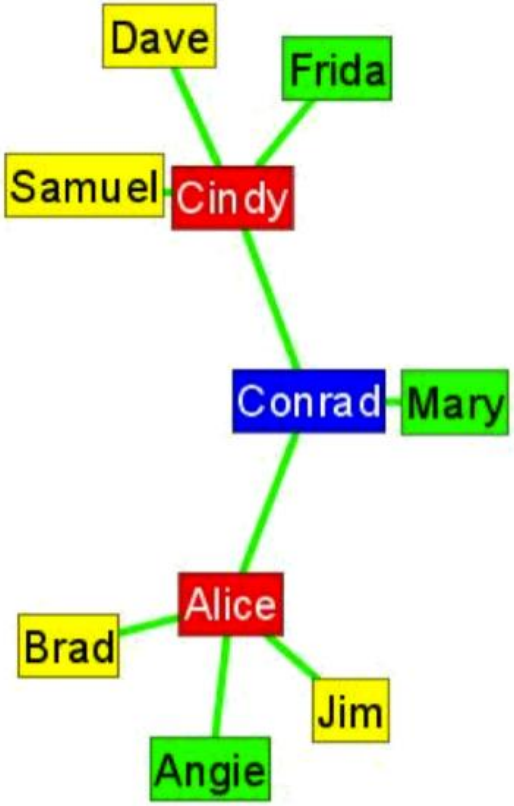
- Snapshot of a large community interactions over a given time span
  - IM graph
  - e-mail logs within a company or University
- Nodes represent customers, employees, or students of the organization that maintains the data
  - Privacy issues
- Who-transacts-with-whom: economic measurements for studying the structure of a market or financial community

Collaboration graphs

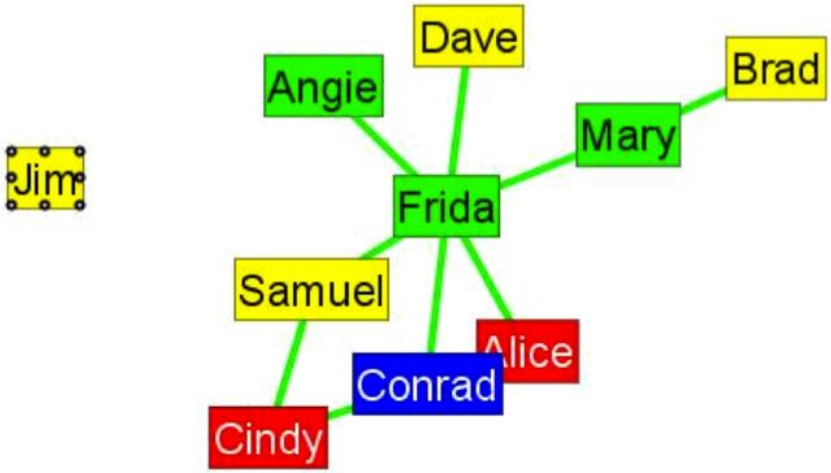
- Record who works with whom in a specific setting
  - co-authorships among scientists
  - co-appearance in movies by actors
  - Wikipedia collaboration graph on articles

Power of SNA in Organizations: Informal Networks Matter

Formal Network



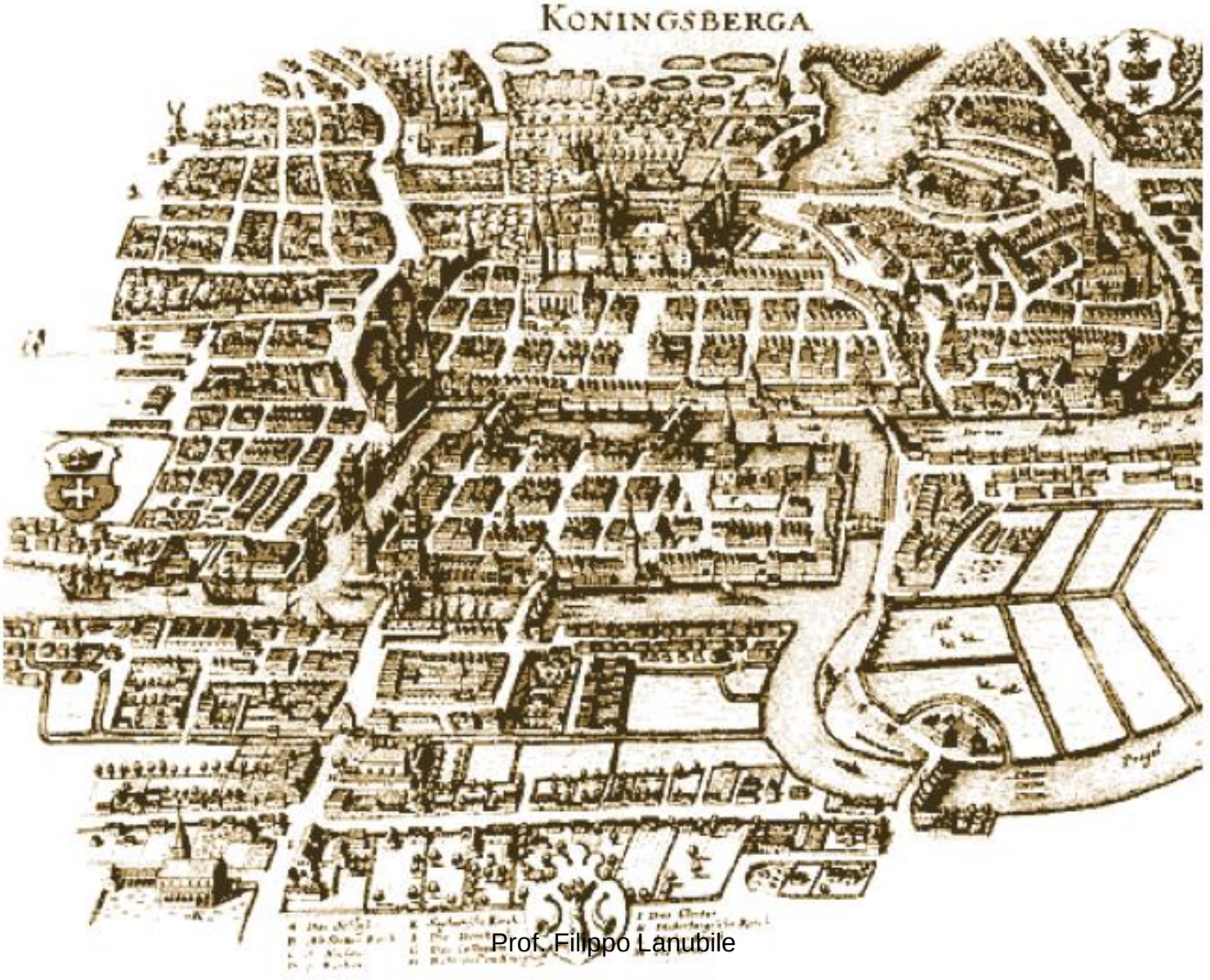
Informal Network



GRAPH THEORY

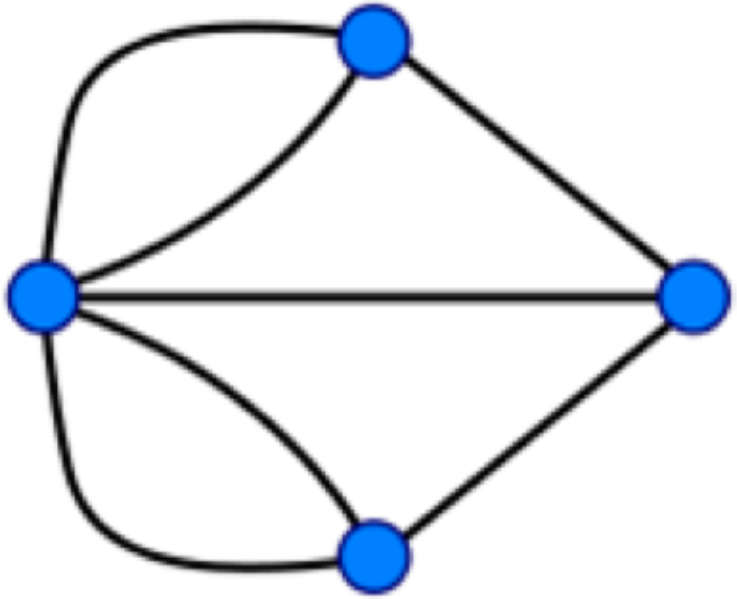
Prof. Filippo Lanubile

Seven bridges of Königsberg



Euler's formalization

- The islands and the two banks of the river are represented as graph nodes and bridges as edges



- Is it possible to traverse a graph without repeating any edges (but possibly repeating nodes) and returning to the starting point?

Graph definition

- Set of *nodes* (or *actors*), with certain pairs of these objects connected by *edges* (or *ties*)
- G = N, E
  - $N = \{n_1, n_2, \ldots, n_g\}$
  - $E = \{e_1, e_2, \ldots, e_l\}$

Directed and Undirected Ties

- Undirected Relations
  - Attended meeting with…
  - Friend of…
  - Communicated with…
- Directed Relations
  - Represent flows or subordination
  - Teacher of, Lends money to

Graph Density

- Defined as the ratio between *actual* and *possible* edges in the graph
  - Directed graphs

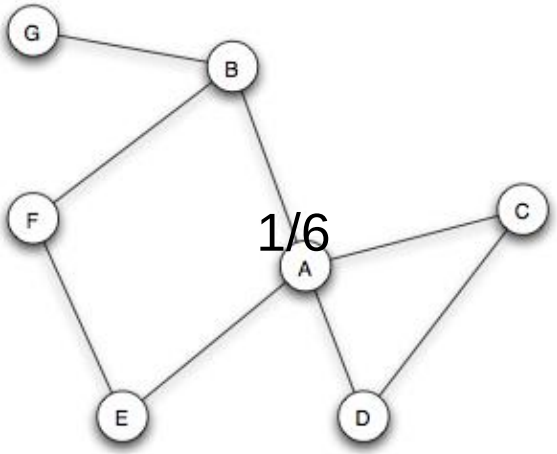
$$D = \frac{|E|}{|N|(|N|-1)}$$

  - Undirected graphs

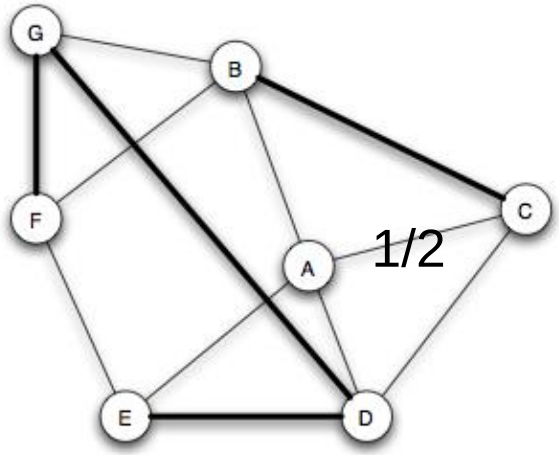
$$D = \frac{|E|}{|N|(|N|-1)/2} = \frac{2|E|}{|N|(|N|-1)}$$

Clustering Coefficient

- Clustering coefficient of a node quantifies how close its neighbors are to being a clique (complete graph)
  - The fraction of pairs of a node A's friends that are connected to each other by edges
    - 0 = none of the node's neighbors are connected with each other (e.g., center in a star network)
    - 1 = all of the node's neighbors are connected with each other



(a) *Before new edges form.*

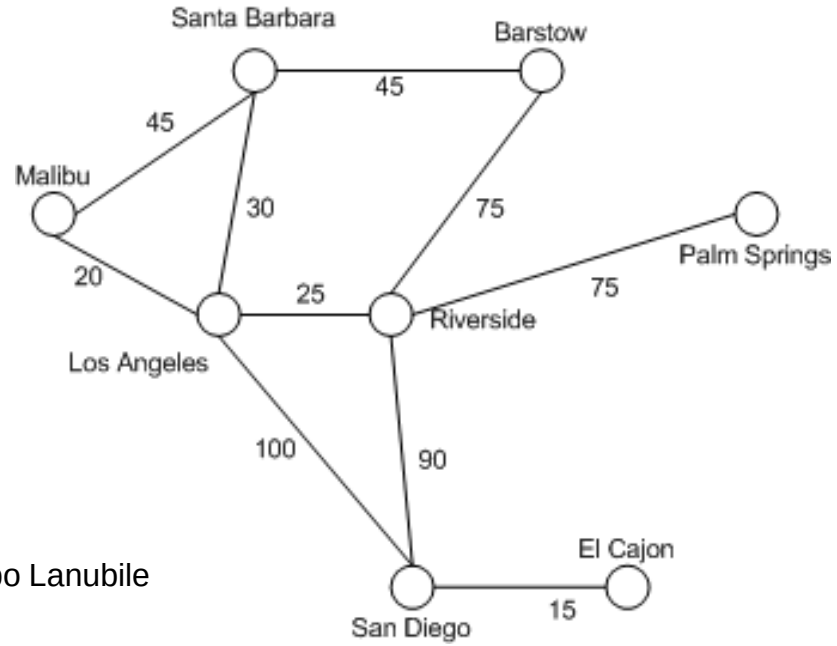


(b) *After new edges form.*

Tie Strength

- We can attach a weight to ties describing its ‘strength’, based on:
  - Frequency of interaction
  - Strength of relationship
  - Information capacity/bandwidth
  - Physical distance

Weighted graph



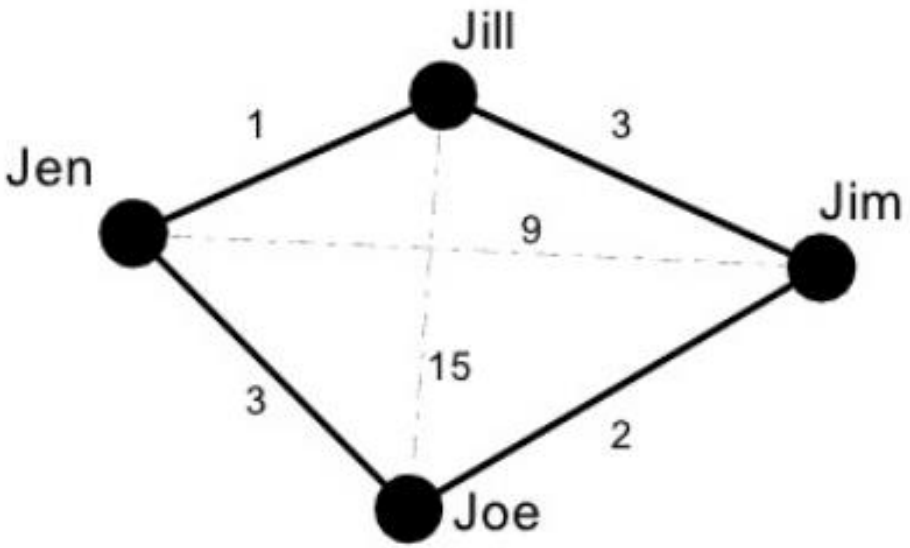
Adjacency Matrices

Friendship

	Jim	Jill	Jen	Joe
Jim	-	1	0	1
Jill	1	-	1	0
Jen	0	1	-	1
Joe	1	0	1	-

Proximity

	Jim	Jill	Jen	Joe
Jim	-	3	9	2
Jill	3	-	1	15
Jen	9	1	-	3
Joe	2	15	3	-

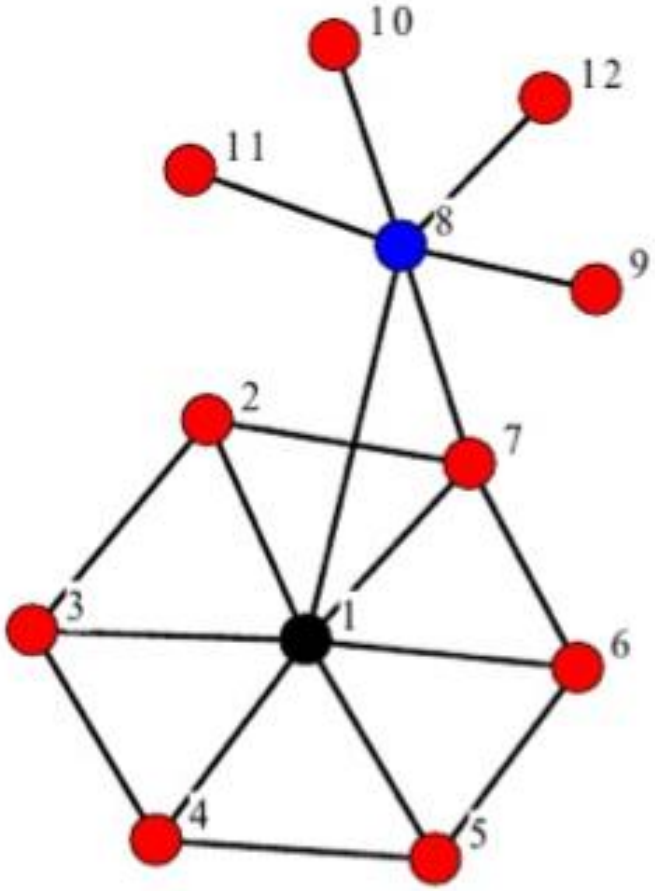


Connectivity

- **Walk**: a sequence of nodes such as each consecutive pair in the sequence is connected by an edge
  - **Open walk**: the starting and ending nodes are different
  - **Closed walk**: same starting and ending node
- **Trail**: walk with no repeated lines
- **Path**: walk with no repeated nodes
- **Cycle**: path with a ring structure

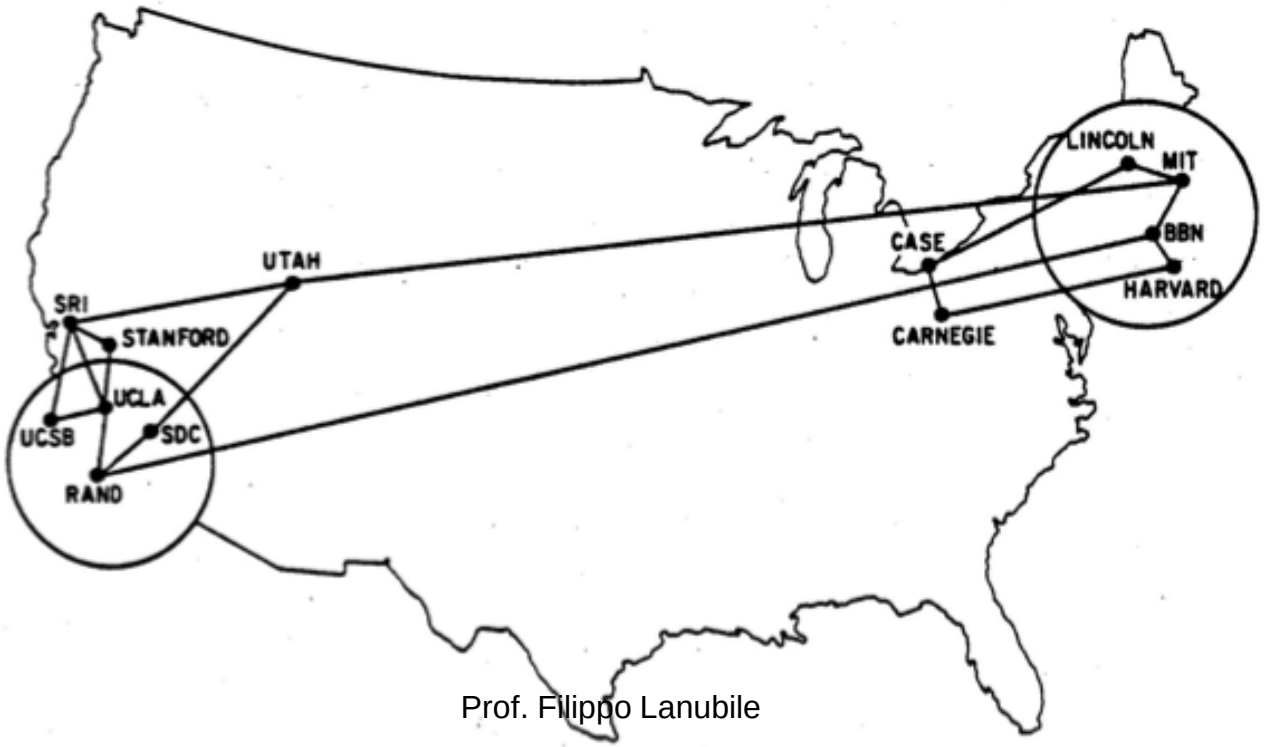
Path Length and Distance

- Length of a path: number of links
- Distance (or ‘geodesic’): length of the shortest path between two nodes
- Graph diameter: longest geodesic between any two nodes



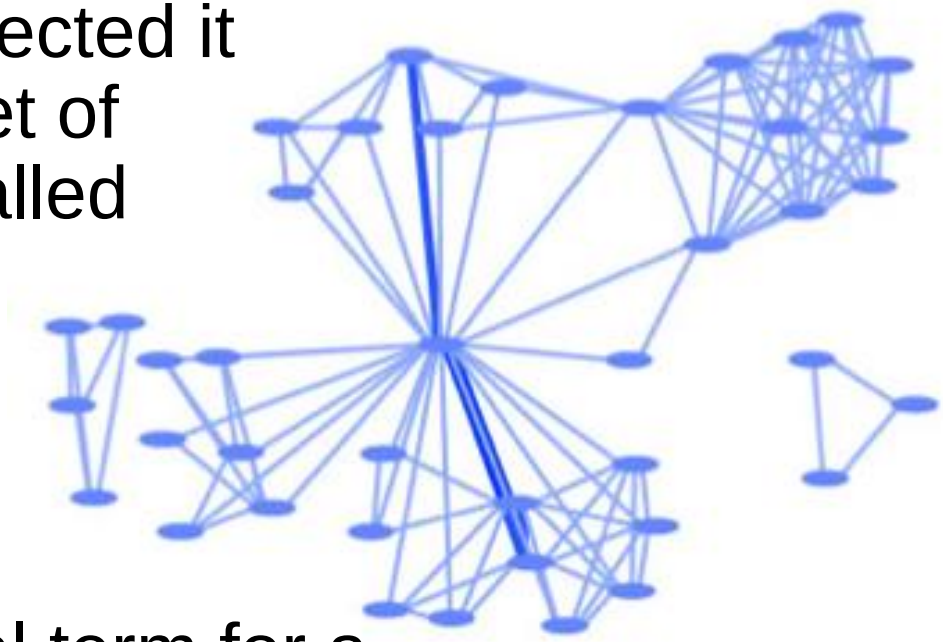
Connectivity

- A graph is connected if for every pair of nodes, there is a path between them
  - Path: walk with no repeated nodes



Giant Components

- If a graph is not connected it breaks apart into a set of connected ‘pieces’ called components



- Giant components:
  a deliberately informal term for a connected component that contains a significant fraction of all the nodes

SMALL WORLD

Prof. Filippo Lanubile

The Small World Phenomenon

- When a network contains a giant component, it almost always contains only one
- In the global friendship network, the argument explaining why you belong to a giant component asserts that:
  - You have paths of friends connecting you to a large fraction of the world population
  - These paths are surprisingly short

Six Degree of Separation

- A pop-cultural mantra originated from a study by Milgram in the 1960s
- Idea: people are really connected in the global friendship network by short chains of friends
  - Task: forwarding a letter to a “target” person through known people
  - Letters as ‘tracers’ of the path to destination: he didn’t know the graph describing the social network of friendship of people involved in the experiment
  - The median length of the 64 letter forwarding chains was six

Social Networks are ‘small worlds’

- Current general consensus due to empirical studies in several domains where the full graph structure is known
  - Paul Erdös experiment’ (Graham, 1979), studying collaboration networks within professional communities
  - Instant messaging (Leskovec & Horvitz, 2008) studying communication among people using Microsoft Messenger

Erdös number

- A mathematician who published 1500 papers
- Collaboration graph: nodes are mathematicians, edges represent co-autorship
- Erdös number: a mathematician's distance from him to Erdös in this graph
  - most mathematicians have Erdös n. of at most 5
  - extending the graph to all the sciences most scientists have very small Erdös n.: Einstein's is 2, Fermi's is 3, Chomsky's is 4
  - The world of science is truly a small one

Ronald L. Graham. On properties of a well-known graph, or, What is your Ramsey number? Annals of the New York Academy of Sciences, 328(1):166-172, June 1979

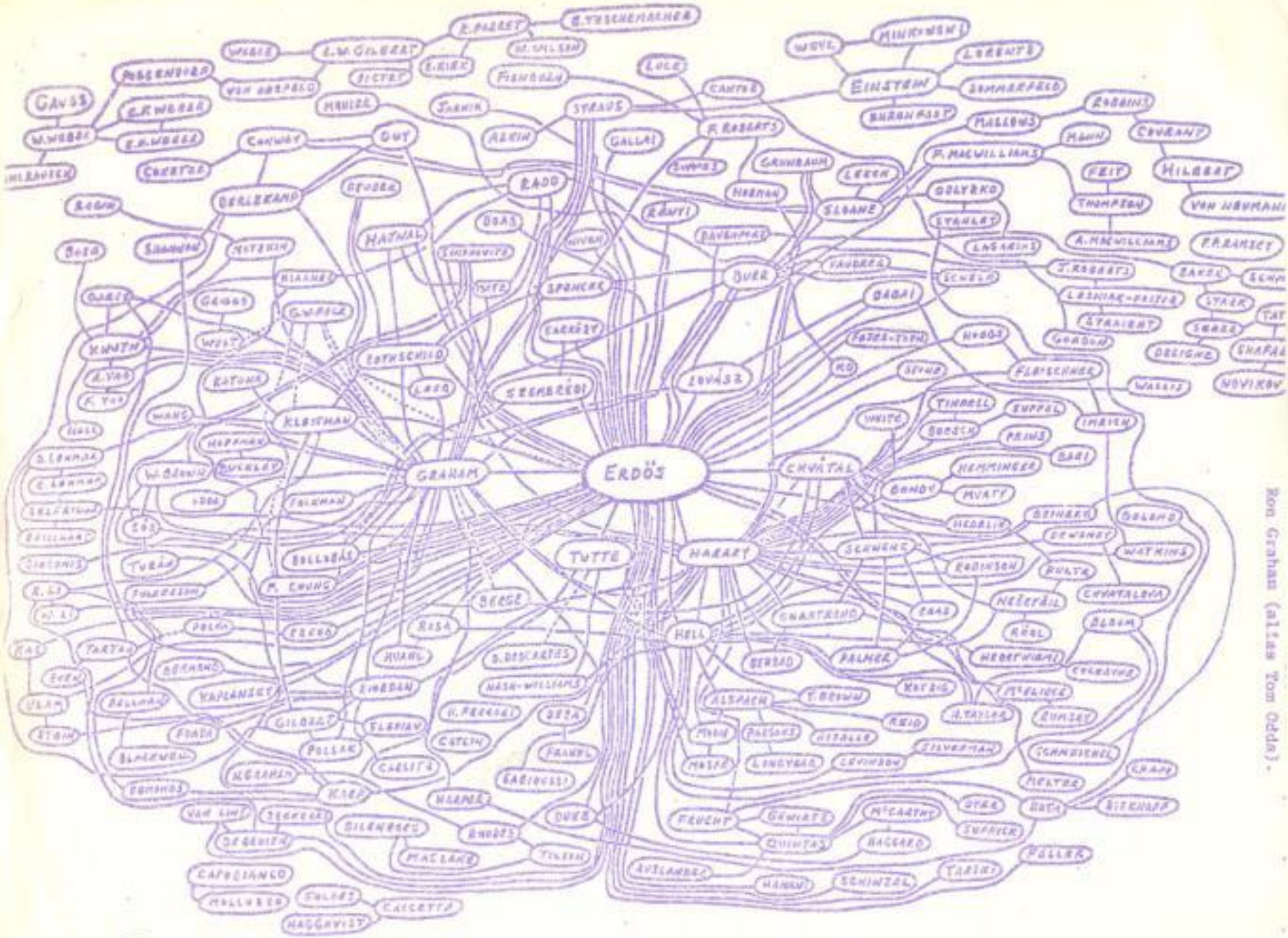


Figure 1

To appear in Topics in Graph Theory (F. Harary, ed.) New York Academy of Sciences (1979).

IM study

- The graph:
  - 240 million active user accounts (nodes) on Microsoft Instant Messenger
  - each node corresponds to a user
  - edge = two-way conversation at any point during a month-long observation period
- Results:
  - One giant component containing almost all of the nodes, the distances within it were very small
  - The distances closely corresponded to the numbers from Milgram experiment (average distance = 6.6, median = 7)

Jure Leskovec and Eric Horvitz. Worldwide buzz: Planetary-scale views on an instant- messaging network. In Proc. 17th International World Wide Web Conference, 2008.

Who has the power?

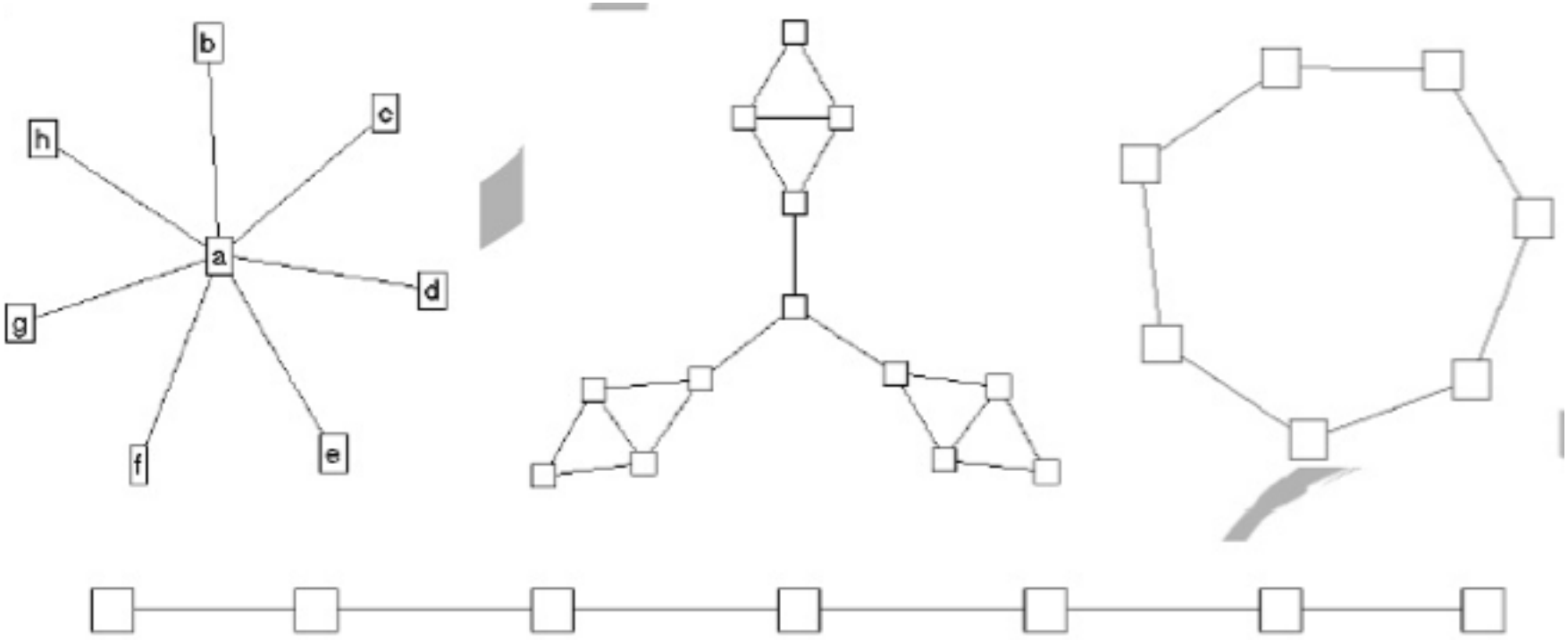
CENTRALITY

What is Centrality?

- Who is more important in this network?
  - At the individual level, one dimension of position in the network can be captured through centrality
- We want to identify which nodes are in the 'center' of the network
  - In terms of power, influence, or other individual characteristics of people (based on their connection patterns)

Centrality in Social Networks

Intuitively, we want a method that allows us to distinguish ‘important’ actors



Measures of node centrality

Undirected graph

- Degree centrality
- Closeness centrality
- Betweenness centrality
- Eigenvector centrality

Directed graph

- Page rank
- Hubs and authorities

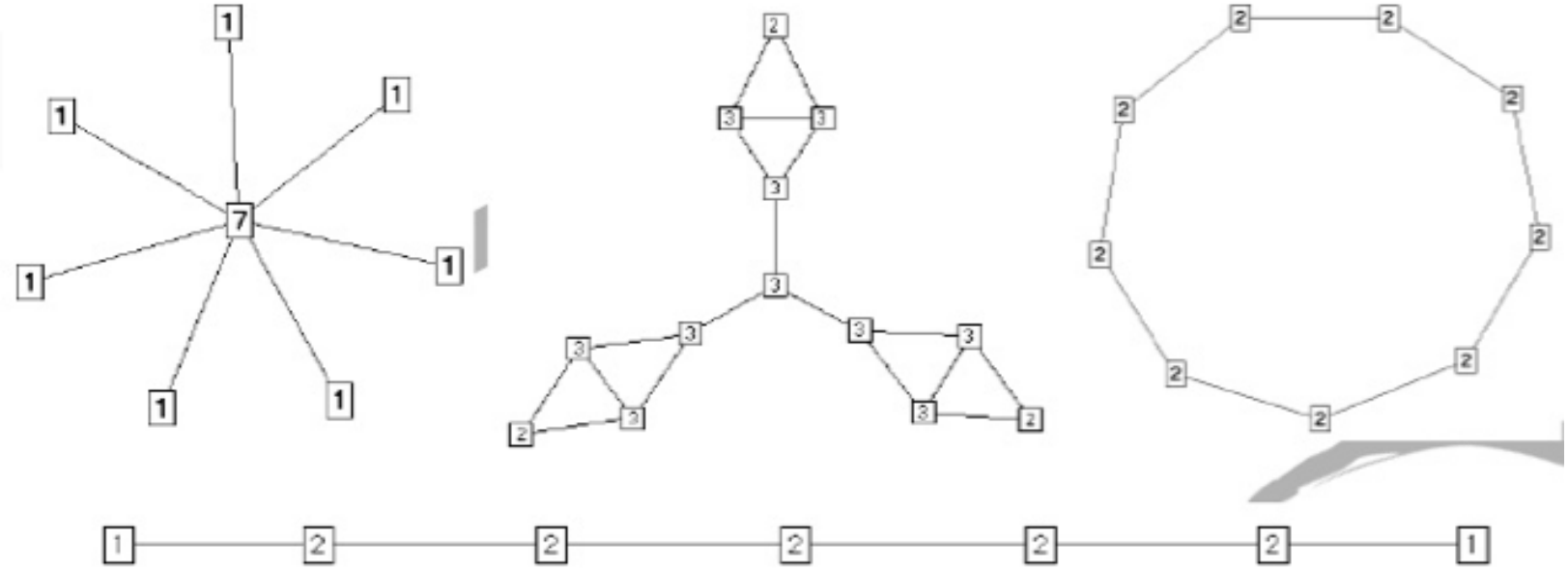
Degree Centrality: Find the 'Celebrities'

- The most intuitive notion of centrality focuses on degree: the actor with most ties is the most important

$$C_D(v) = \deg(v)$$

- In undirected graphs:
  - **Node degree:** number of lines that connect it to other nodes
- In directed graphs:
  - **Node indegree**: number of incoming edges
  - **Node outdegree**: number of outgoing edges

Degree Centrality



Closeness Centrality: Find the Gossipmongers

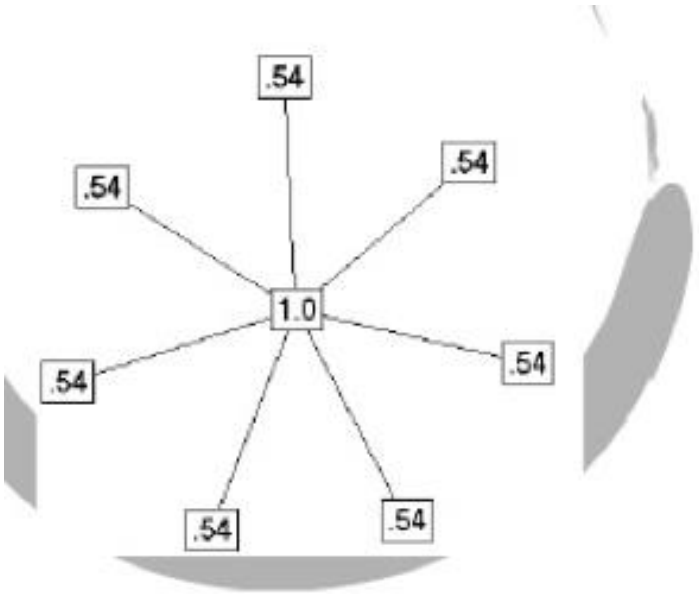
- An actor is considered important if he is relatively close to all other actors
- Closeness is based on the inverse of the distance of each actor to every other actor in the network

$$C(x) = \frac{1}{\sum_y d(y,x)}$$

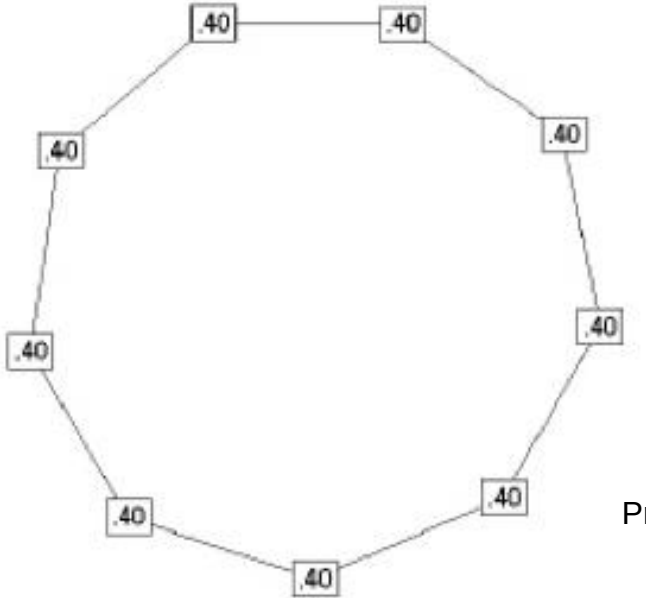
- Standardized form:

$$C(x) = \frac{N}{\sum_y d(y,x)}.$$

Closeness Centrality

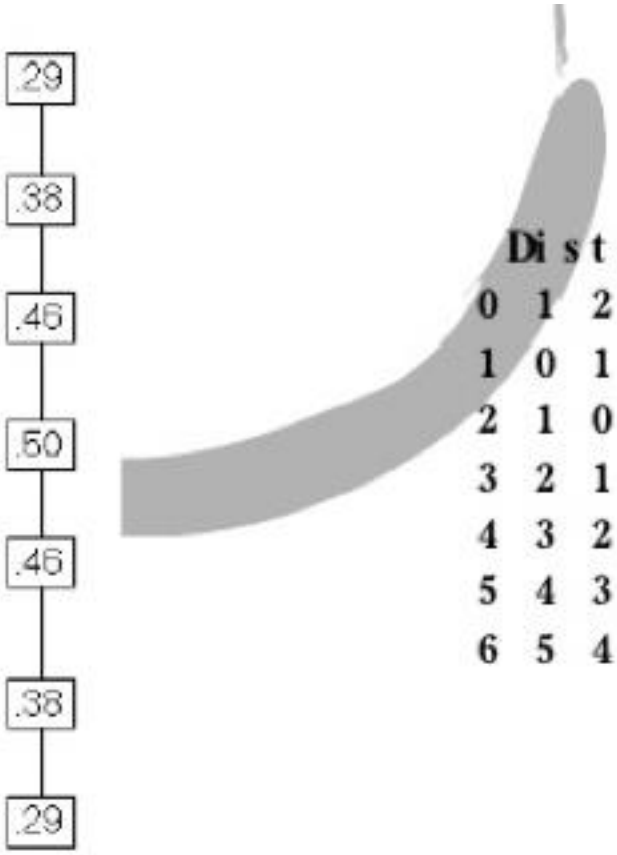


Distance								Closeness
0	1	1	1	1	1	1	1	.143
1	0	2	2	2	2	2	2	.077
1	2	0	2	2	2	2	2	.077
1	2	2	0	2	2	2	2	.077
1	2	2	2	0	2	2	2	.077
1	2	2	2	2	0	2	2	.077
1	2	2	2	2	2	0	2	.077
1	2	2	2	2	2	2	0	.077



Distance									Closeness
0	1	2	3	4	4	3	2	1	.050
1	0	1	2	3	4	4	3	2	.050
2	1	0	1	2	3	4	4	3	.050
3	2	1	0	1	2	3	4	4	.050
4	3	2	1	0	1	2	3	4	.050
4	4	3	2	1	0	1	2	3	.050
3	4	4	3	2	1	0	1	2	.050
2	3	4	4	3	2	1	0	1	.050
1	2	3	4	4	3	2	1	0	.050

Closeness Centrality



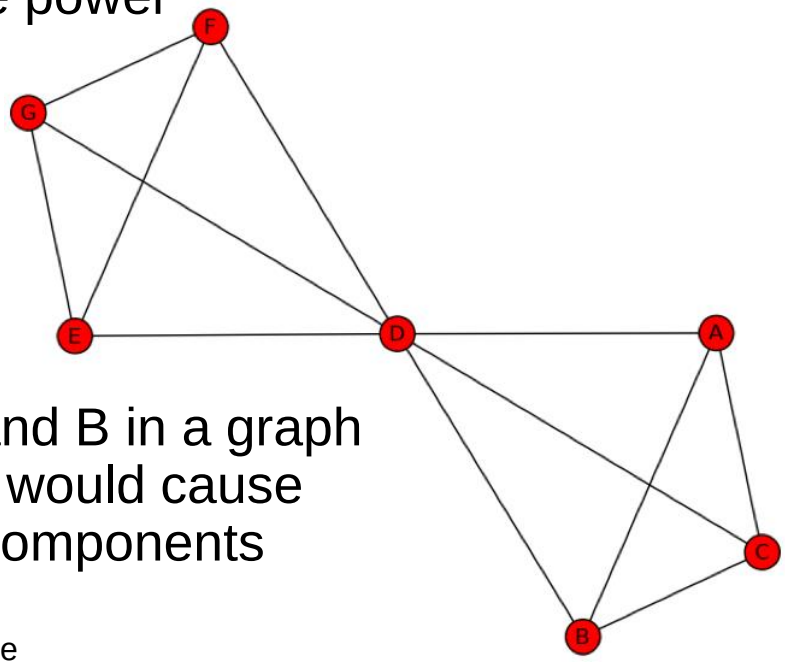
Distance							Closeness
0	1	2	3	4	5	6	.048
1	0	1	2	3	4	5	.063
2	1	0	1	2	3	4	.077
3	2	1	0	1	2	3	.083
4	3	2	1	0	1	2	.077
5	4	3	2	1	0	1	.063
6	5	4	3	2	1	0	.048

Interpreting Closeness Centrality

- Ability to move information from one side of the network to another (i.e., gossip) may be defined in terms of distance to others (or the inverse of it, closeness) and can define a person's role in the network.
- The *horizon of observability*- that is, the ability to see into the network- is about 2 levels, meaning that a node has almost no insight into what is happening 3 or more steps away

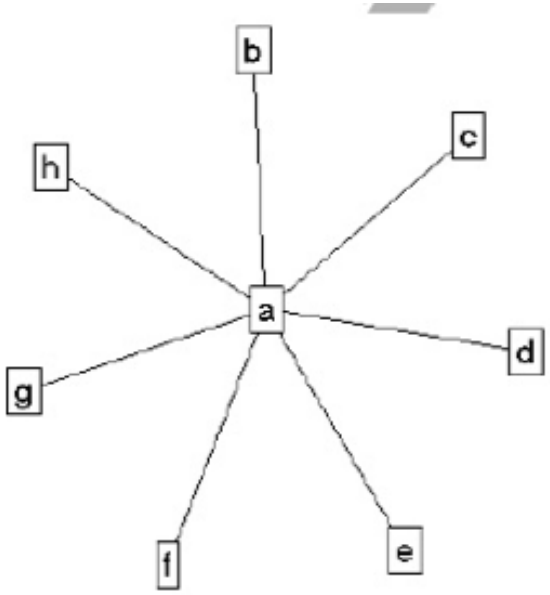
Betweenness Centrality: Find the Communication Bottlenecks and/or Community Bridges

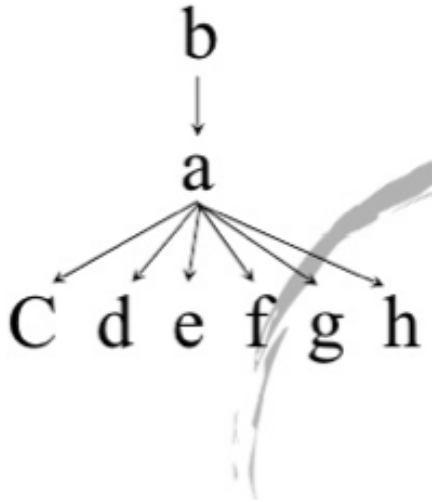
- Model based on communication flow
  - A person who lies on communication paths can control communication flow, and thus is important
    - node D is in a position of some power
  - Betweenness centrality is also able to identify people that act as bridges between two or more communities
    - An edge joining two nodes A and B in a graph is a bridge if deleting the edge would cause A and B to lie in two different components



Betweenness Centrality

- Betweenness centrality quantifies the number of times a node acts as a bridge along the shortest path between two other nodes



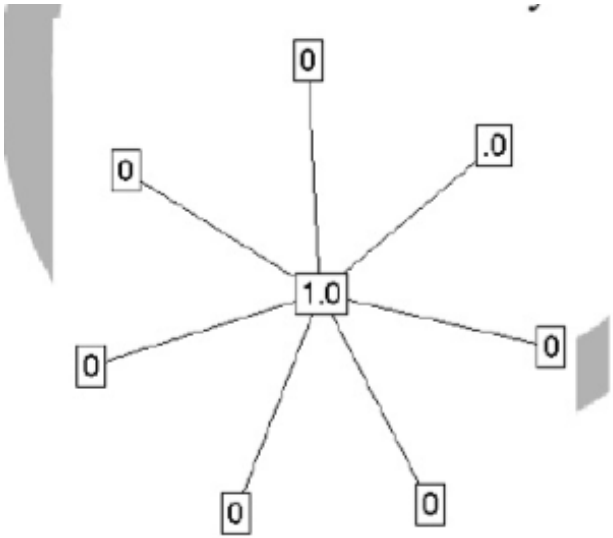


Betweenness Centrality

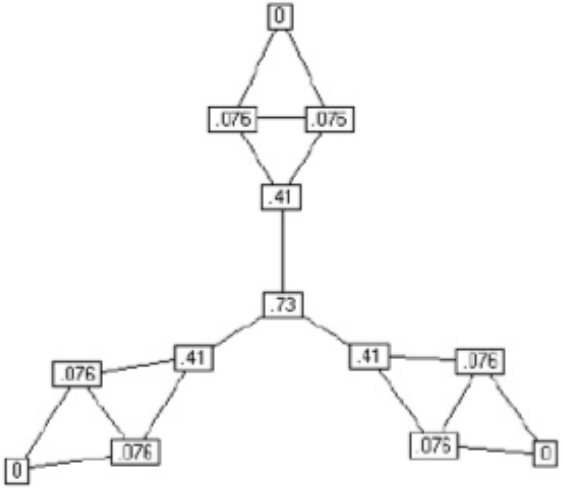
- For each pair of vertices (*s*,*t*), compute the shortest paths between them
- For each pair of vertices (*s*,*t*), determine the fraction of shortest paths that pass through the vertex in question (here, vertex *v*)
- Sum this fraction over all pairs of vertices (*s*,*t*)

$$C_B(v) = \sum_{s \neq v \neq t \in V} \frac{\sigma_{st}(v)}{\sigma_{st}}$$

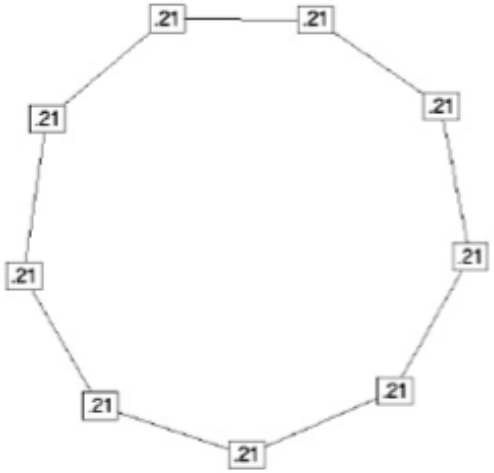
Betweenness Centralization



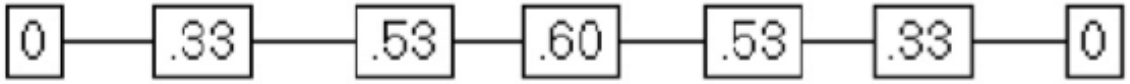
Centralization: 1.0



Centralization: .59



Centralization: 0



Centralization: .31

Comparison

- Generally the different centrality types will be positively correlated
- When they are not/low correlated, it probably tells us something interesting about the network

Combined interpretation

	Low Degree	Low Closeness	Low Betweenness
High Degree		Embedded in cluster that is far from the rest of the network	Ego's connections are redundant - communication bypasses him/her
High Closeness	Key player tied to important important/active alters		Probably multiple paths in the network, ego is near many people, but so are many others
High Betweenness	Ego's few ties are crucial for network flow	Very rare cell. Would mean that ego monopolizes the ties from a small number of people to many others.	

FINDING COHESIVE SUBGROUPS IN NETWORK DATA

COMMUNITY DETECTION

Prof. Filippo Lanubile

Goal

- Find a meaningful way to separate larger networks into groups
- Meaningful =
  - Reduce overlap
  - Locate cohesive groups

Component and Subgraphs

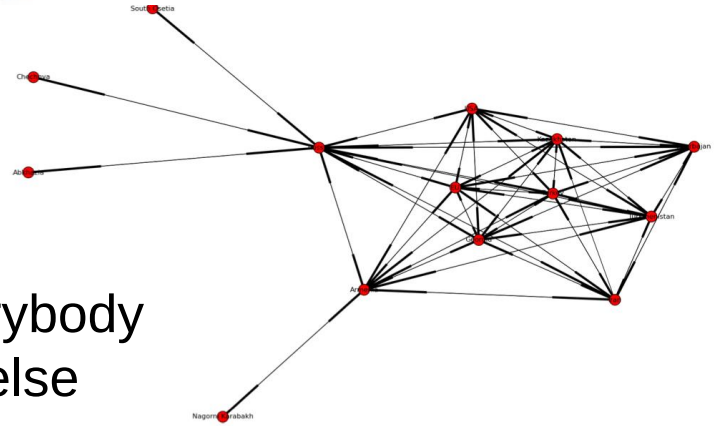
- *Subgraph*: subset of the nodes of a network and all of the edges linking these nodes
  - Any group of nodes can form a subgraph
- *Component*: portion of the network that are disconnected from each other
  - Montecchi and Capuleti before Romeo and Juliet met

Cliques

- Maximal, complete subgraphs

$$\forall u, v \in S, \exists (u, v) \in E$$

- Properties
  - Maximum density (1.0)
  - Minimum distances (all 1)
    - e.g., a group of people where everybody is connected directly to everyone else



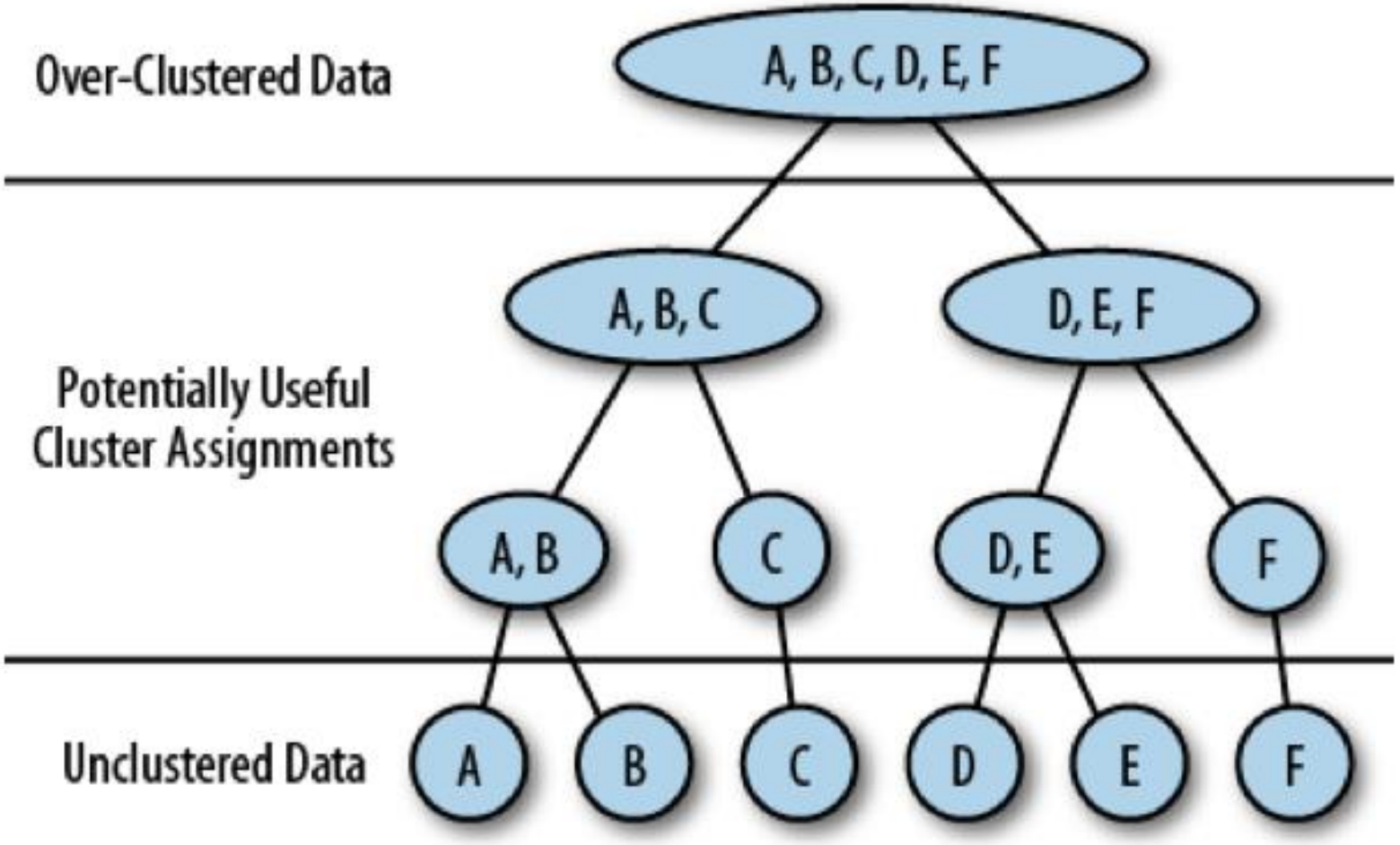
- The strict clique definition may be too strong
  - N-cliques: N stands for the length of the path allowed to make a connection to all other members
    - e.g., N-cliques for N=2

Hierarchical Clustering

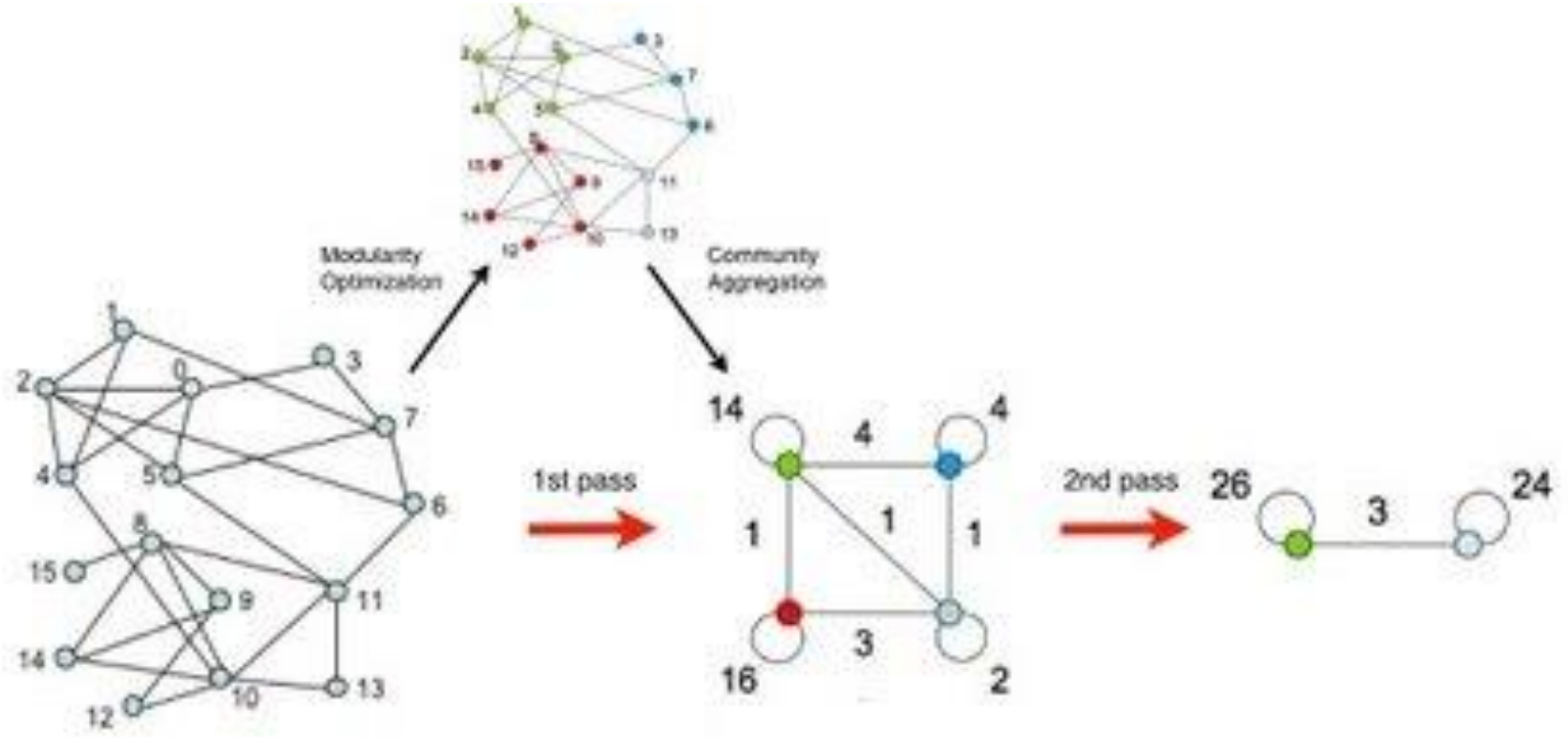
- Using an NxN distance or similarity matrix
- Can use multiple distance metrics
  - Graph distance – binary or weighted
  - Euclidean distance
    - Similarity of relational vectors
  - Similarity matrix

Hierarchical Clustering

Algorithm based on the notion of distance between nodes/clusters



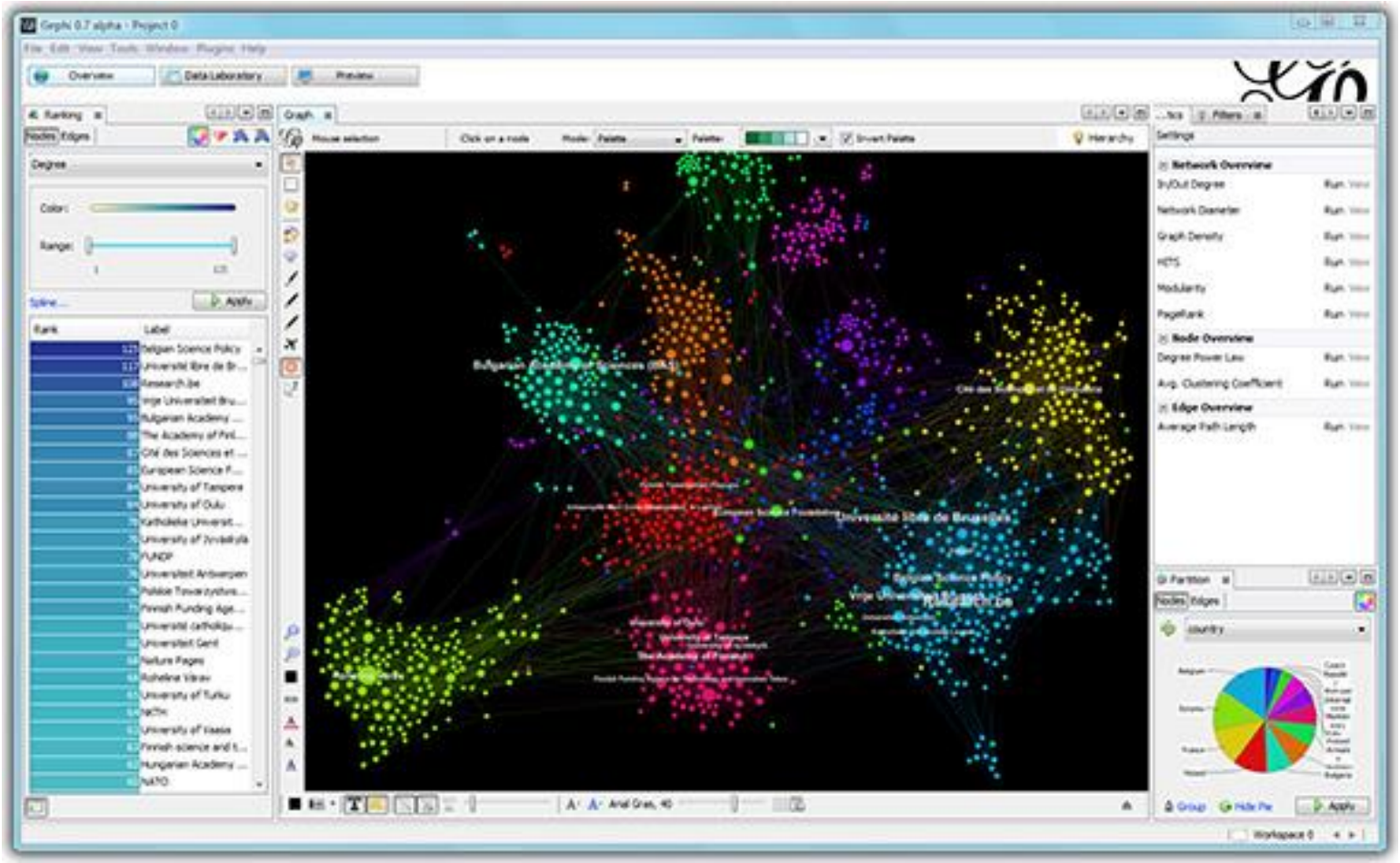
Finding communities in large networks: Louvain method



V.D. Blondel et al. Fast unfolding of communities in large networks. Journal of Statistical Mechanics: Theory and Experiment, 2008

Gephi: open source network visualization platform

https://gephi.org/



References

- Social Network Analysis for Startups (M. Tvesovat and A. Kounznetsov)
  - http://mediashow.ru/sites/default/files/books/2011/11/social.network.analysis.for_.startups.1449306462.pdf
- Networks, Crowds, and Markets: Reasoning about a Highly Connected World (D. Easley and J. Kleinberg)
  - http://www.cs.cornell.edu/home/kleinber/networks-book/
- N. Novielli, S. Marczak. “Social Network Analysis for Global Software Engineering: Exploring Developer Relationships from a Fine-Grained Perspective“, tutorial at the 8th IEEE Int. Conf. on Global Software Engineering Workshops, 2013
  - http://www.slideshare.net/nolli82/social-network-analysis-for-global-software-engineering-exploring-relationships-from-a-finegrained-level-icgse-2013