Introduction to Social Networks

Advanced Computer Networks

Summer Semester 2014

Online Social Networks (OSNs)

- OSNs have become extremely popular within the last decade
  - E.g., Facebook: more than 1 billion users
- A giant pool of data, large-scale structures, communication, …
  - Research in this direction very interesting and multi-faceted:
    - What about data protection / privacy?
    - Arabic Spring: Political influence of OSNs?
    - Data propagation: How is information conveyed in OSNs?
    - Connections among people: useful for applications?

computer NETWORKS

Terminology

- Social Network
  - A network made up by a set of individuals interconnecting with each other basing on social relationships (such as friendships, partnerships, etc.)
- Entity: a basic unit of the network
- Link: interconnection between entities
- Behavior and dynamics
  - Each individual's actions have implicit consequences for the outcomes of everyone in the system
  - Individual actions are not in isolation: cause-effect
    - Changes in a product, a Web site, or a government program
    - The rich get richer; winners take all; small advantages are magnified to a critical mass; new ideas get attention that becomes viral

Network: Friendship

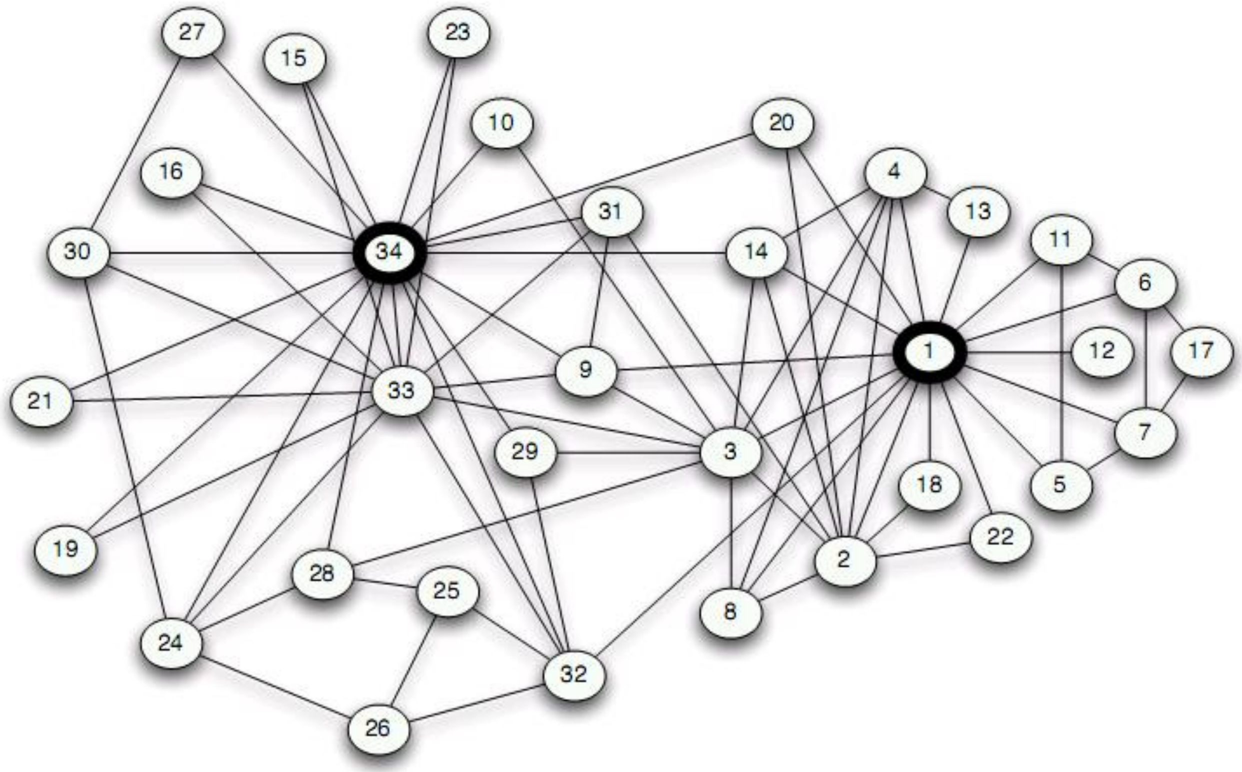
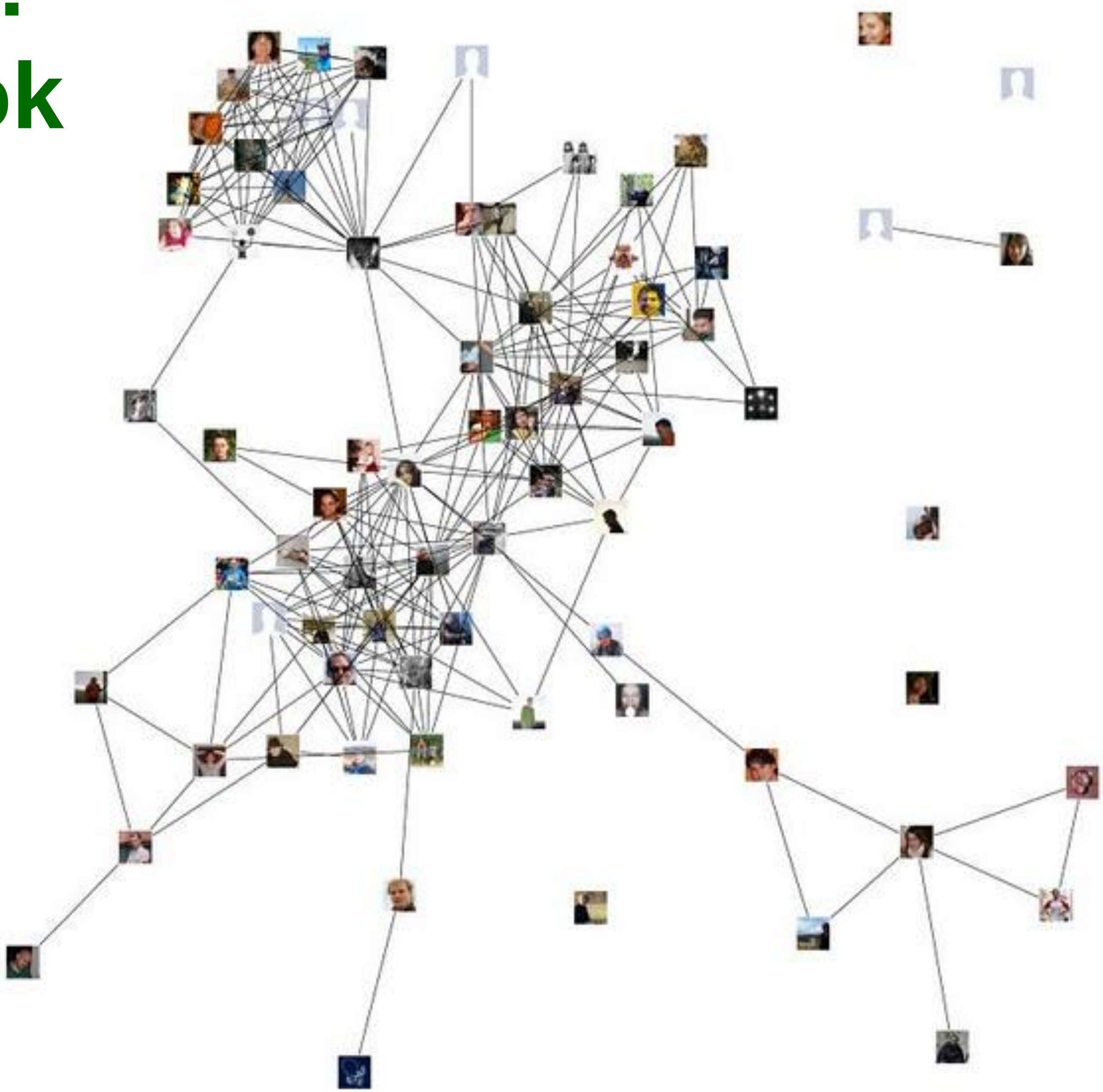


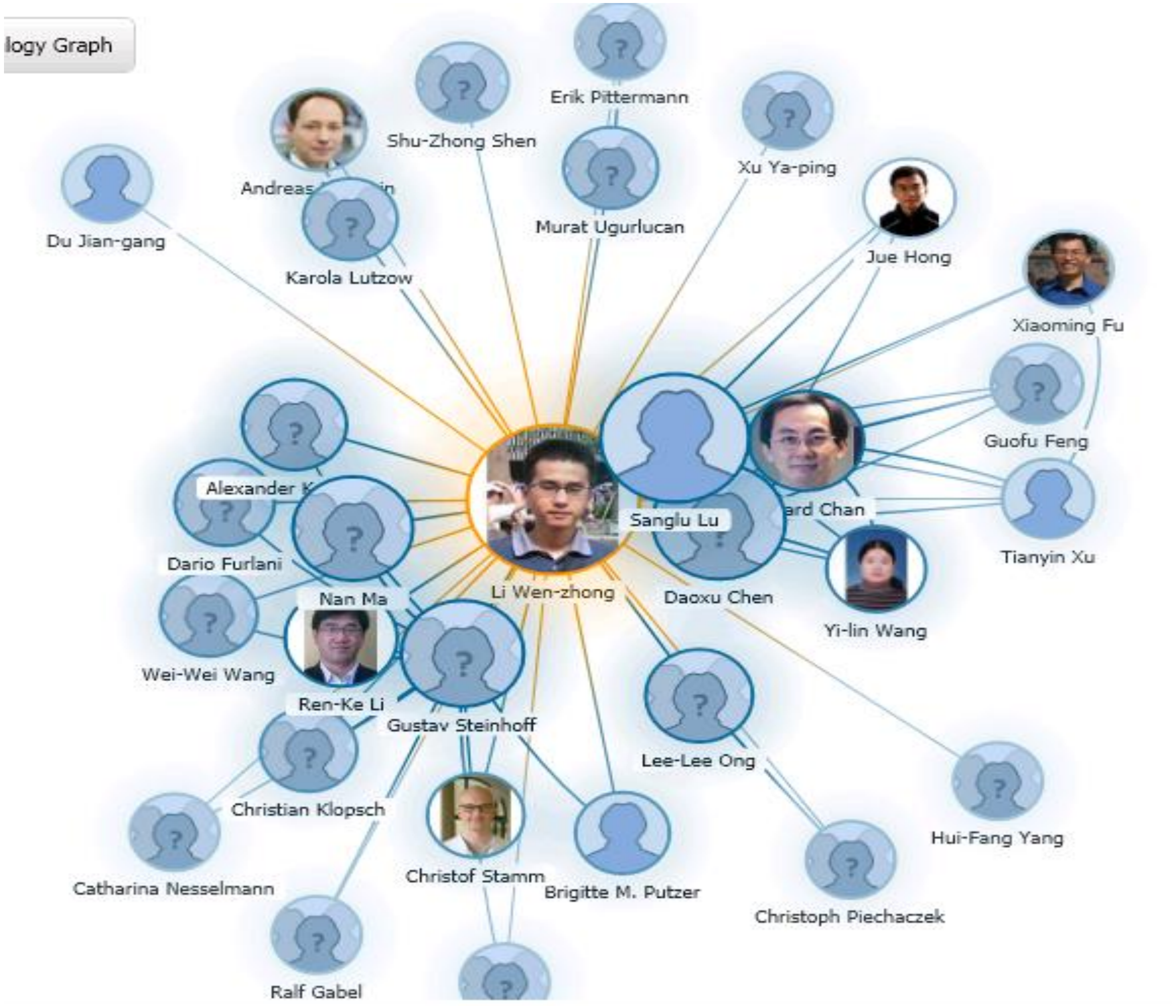
Figure 1.1: The social network of friendships within a 34-person karate club [421].

Network: Facebook

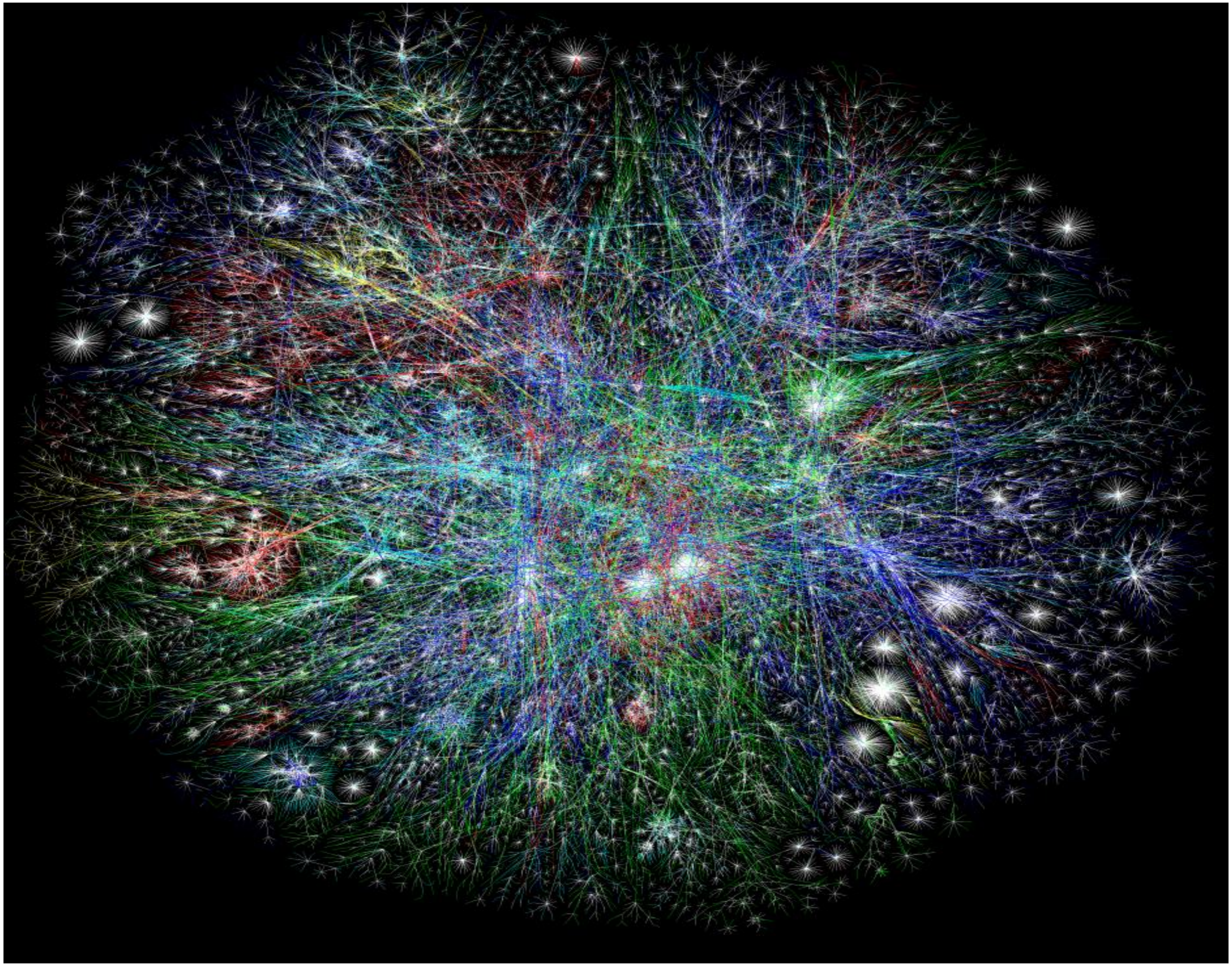


http://revolution-computing.typepad.com/.a/6a010534b1db25970b016760ccd666970b-pi

Network: Co-authorship



Network: Communication



computer NETWORKS

Network: Information

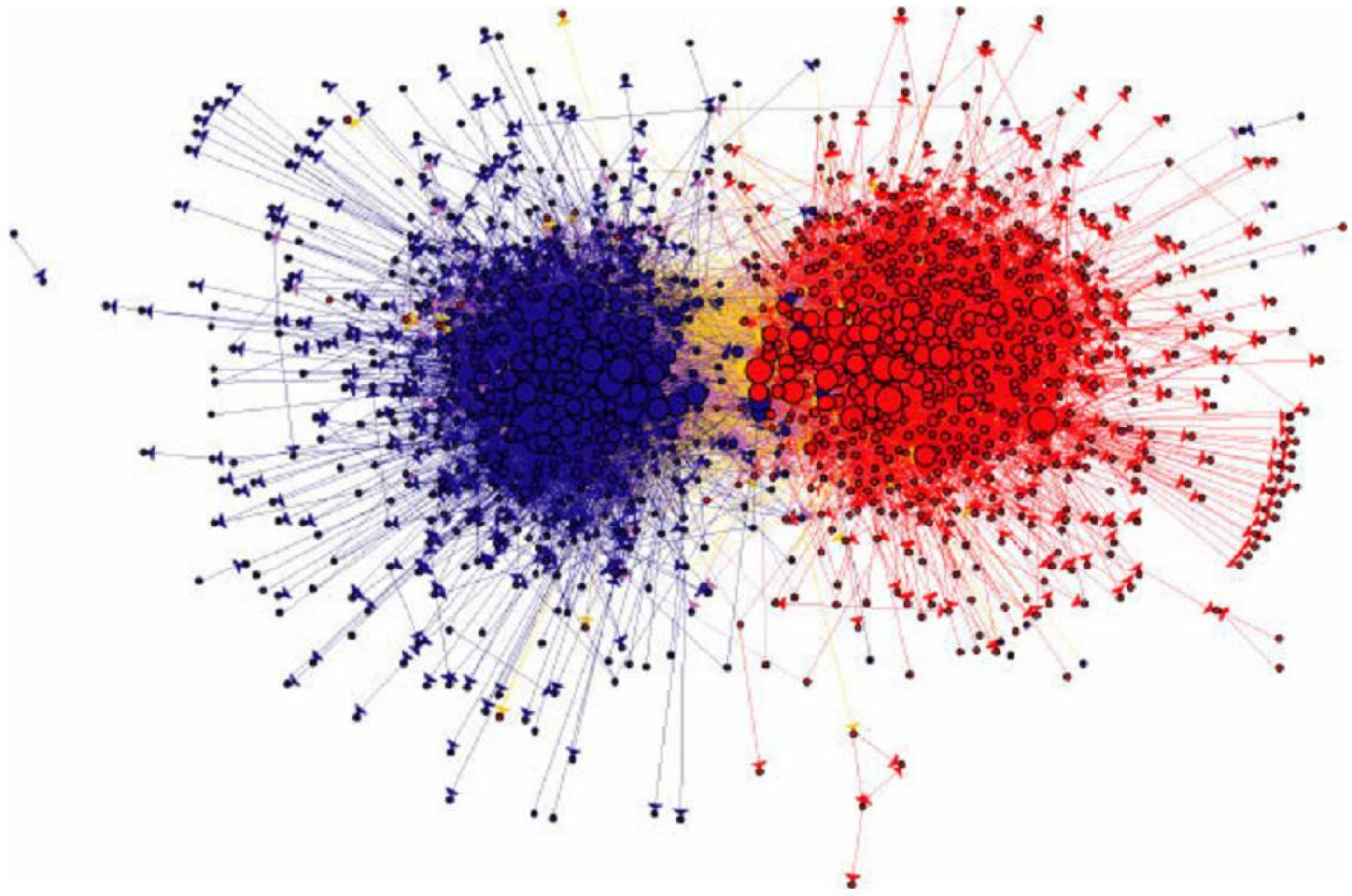
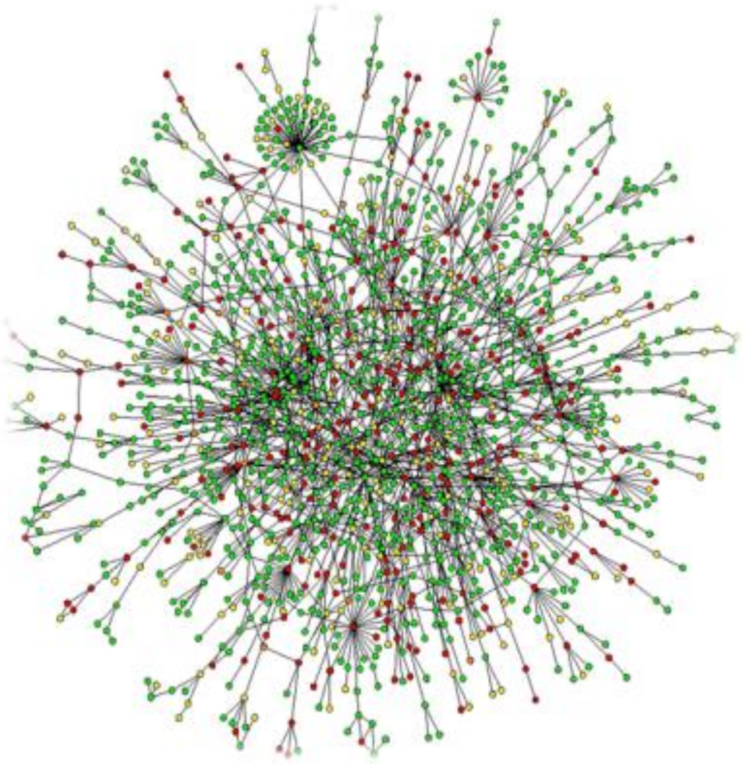


Figure 1.4: The links among Web pages can reveal densely-knit communities and prominent sites. In this case, the network structure of political blogs prior to the 2004 U.S. Presidential election reveals two natural and well-separated clusters [5]. (Image from http://www-personal.umich.edu/ ladamic/img/politicalblogs.jpg)

Network: Biological



Protein-Protein Interaction Networks:
Nodes: Proteins
Edges: 'physical' interactoins

Human brain has between
10-100 billion neurons

Commonalities among Networks

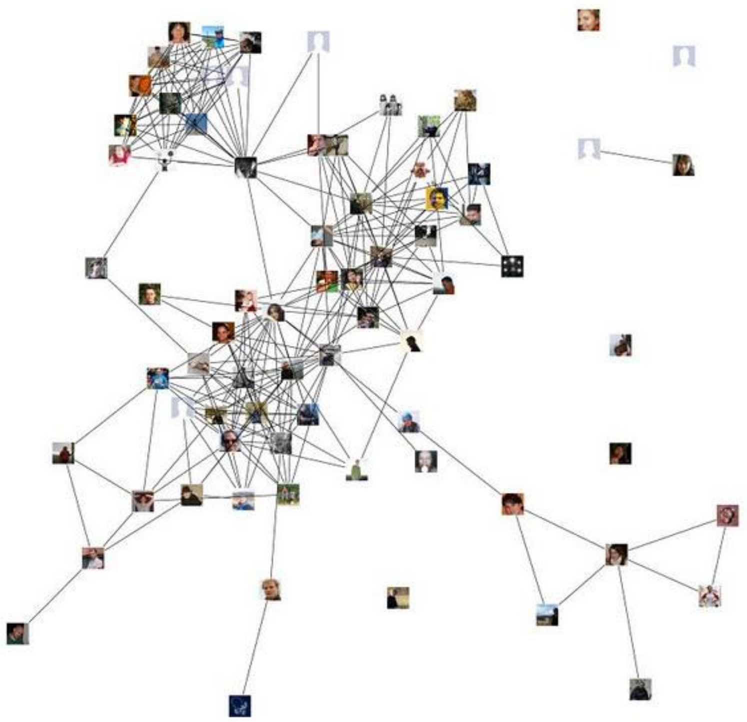
- A network that defines the interactions between the components
  - Seems random, but displays signatures of order and self-organization
- Characteristics
  - Virtual: it does not physically exist
  - Complex: it consists of a large scale number of nodes
  - Grouping: it forms communities due to different interests
  - Dynamic: its structure is evolving over time

computer NETWORKS

Social Network Analysis

Research Questions

- Structure and evolution
  - What is the structure of a network?
  - Why and how has this structure evolved?
- Processes and dynamics
  - Social networks provide "skeleton" for spreading of information, behavior, …



Methods

- Empirical:
  - Study network data to find organizational principles
- Mathematical models:
  - Probabilistic, graph theory
- Algorithms
  - Algorithms for analyzing graphs

Targets of this part of the lecture

- Learn about patterns and statistical properties of network data
- Understand why social networks are organized the way they are (prediction)

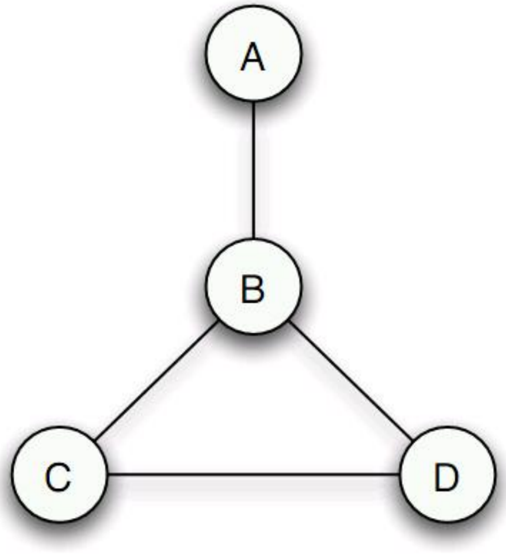
Research Work Done by the Lab

- SOUP: An Online Social Network By The People, For The People
- How do Online Social Networks grow?
- Mining Triadic Closure Patterns in Social Networks
- Multi-Objective Data Placement for Multi-Cloud Socially Aware Services
- Cuckoo: Scaling Microblogging Services with Divergent Traffic Demands
- LENS: Leveraging Social Networking and Trust to Prevent Spam Transmission
- Exploring Regional and Global Population Growth in Online Social Networks
- On the Effectiveness of OSN based Sybil defenses
- …

Modeling Social Networks

Network as a graph

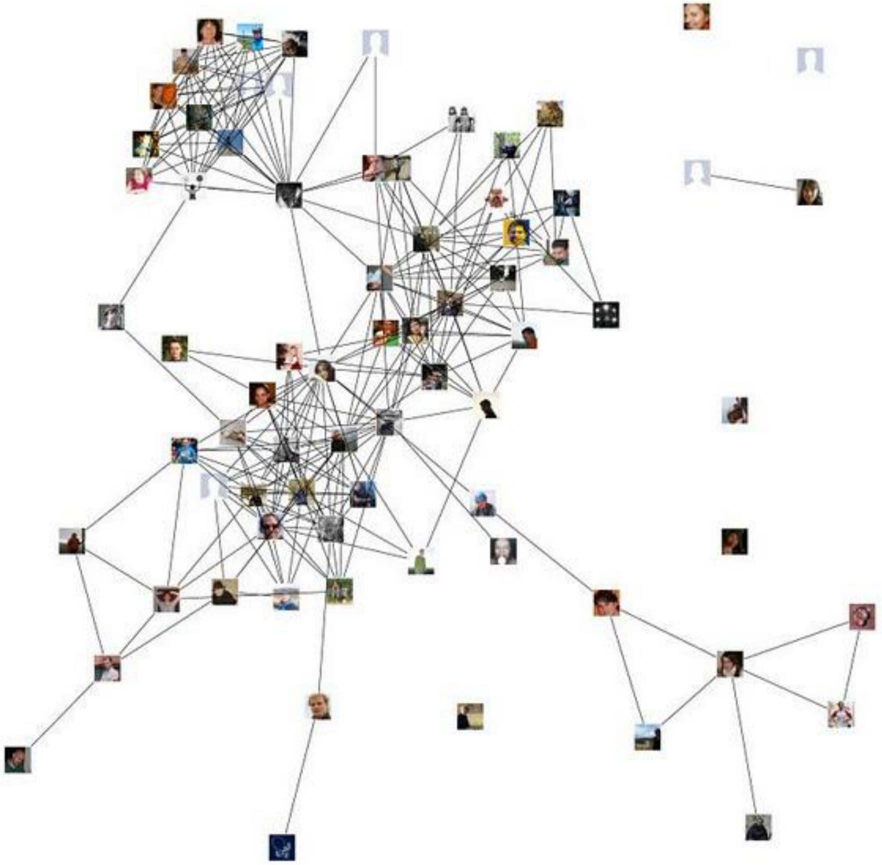
- A network can be represented by a graph mathematically
- Node: an object in the network
- Edge: a link between objects
- Neighbors: nodes connected by edg



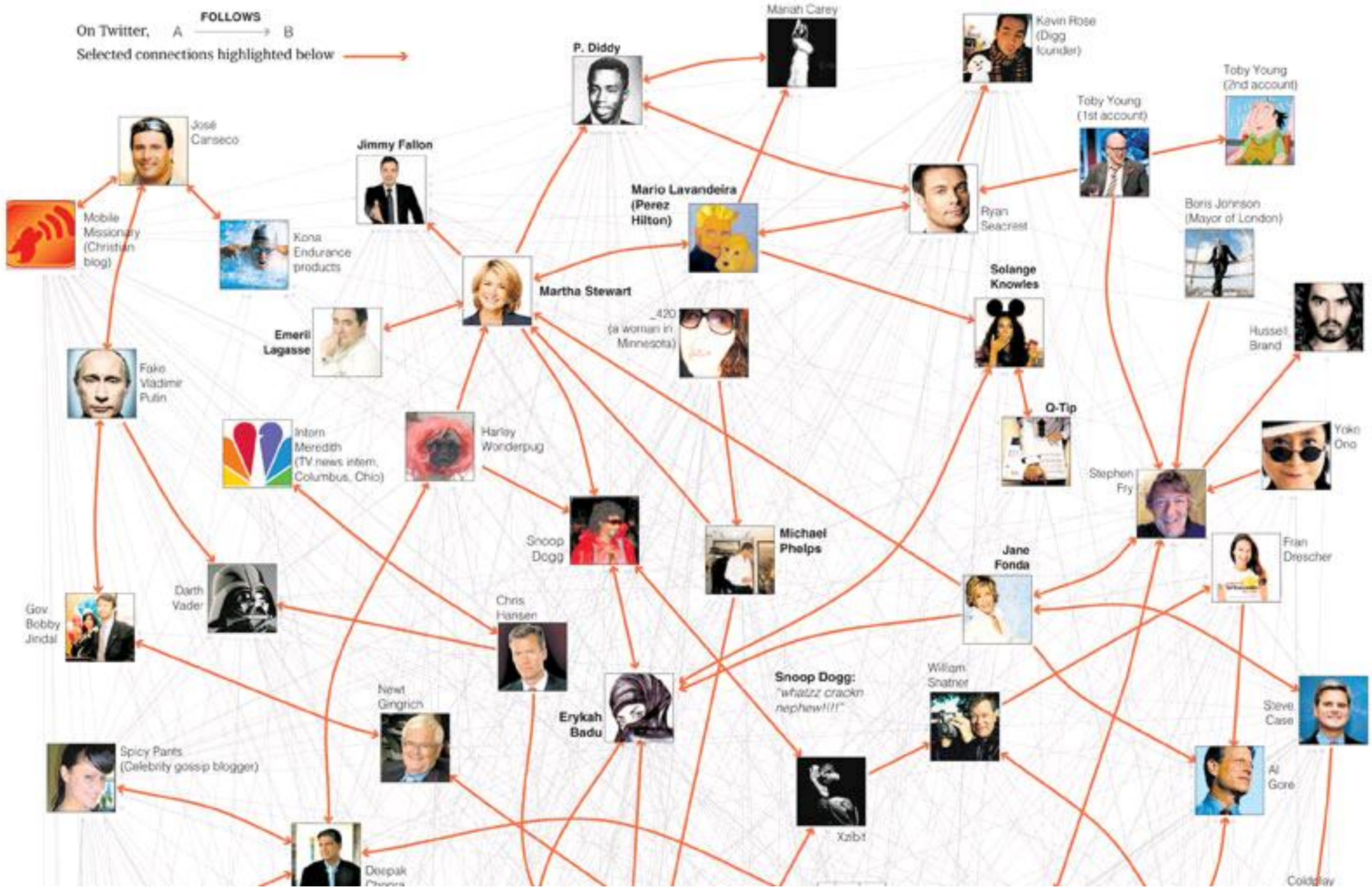
(a) *A graph on 4 nodes.*

Undirected Graph

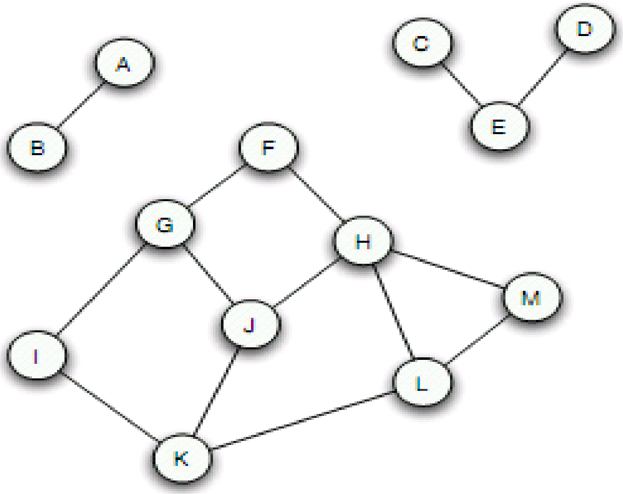
- Facebook friendship network



Directed Graph



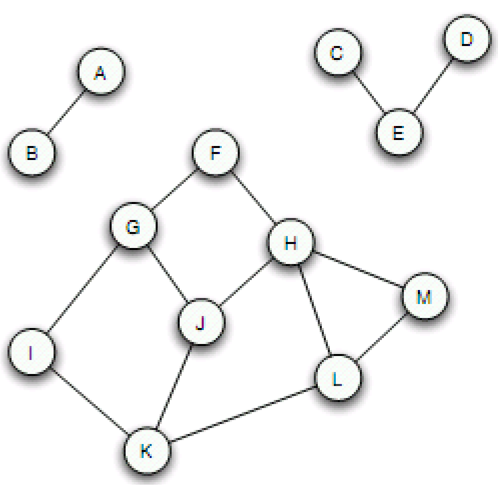
- Path: a sequence of interconnected nodes
- Cycle: a path, the first and last nodes are the same, but other nodes are distinct.
- Connectivity:
  - A graph is connected if for every pair of nodes, there is a path between them



computer NETWORKS

Components

- If a graph is not connected, it breaks apart into several connected subgraphs
- A connected component is a subset of the nodes such that (i) every node in the subset has a path to every other; and (ii) the subset is not part of some larger set with the property that every node can reach every other



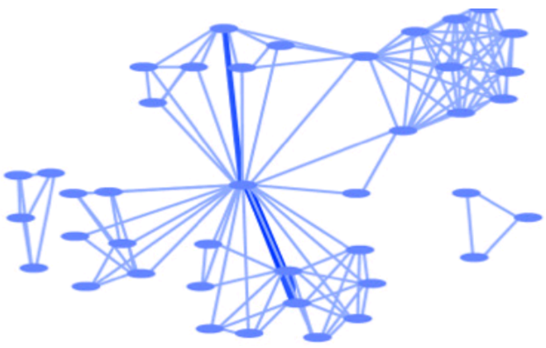
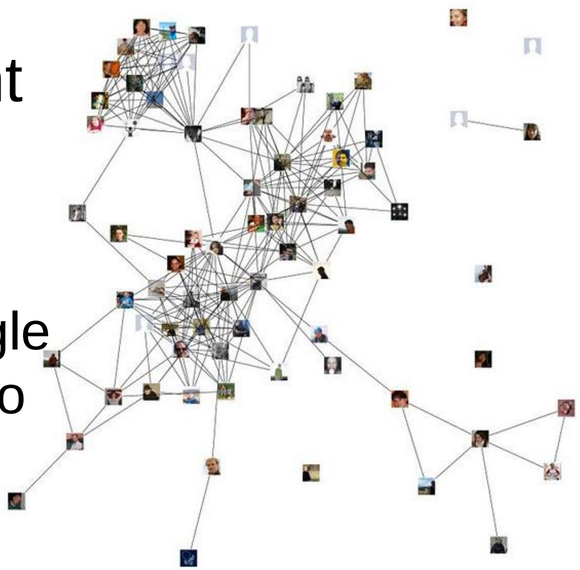


Figure 2.6: The collaboration graph of the biological research center *Structural Genomics of Pathogenic Protozoa (SGPP)* [134], which consists of three distinct connected components. This graph was part of a comparative study of the collaboration patterns graphs of nine research centers supported by NIH's Protein Structure Initiative; SGPP was an intermediate case between centers whose collaboration graph was connected and those for which it was fragmented into many small components.

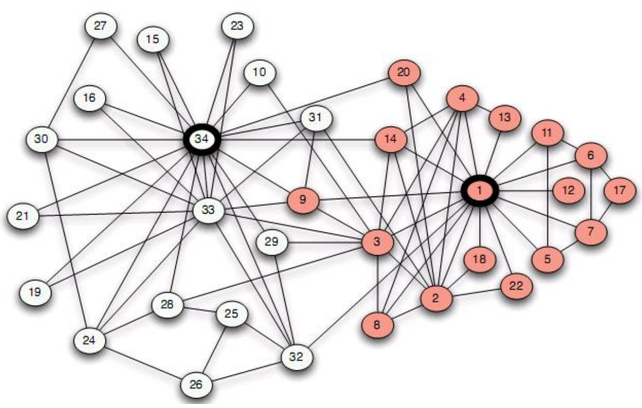
Example: Giant Component

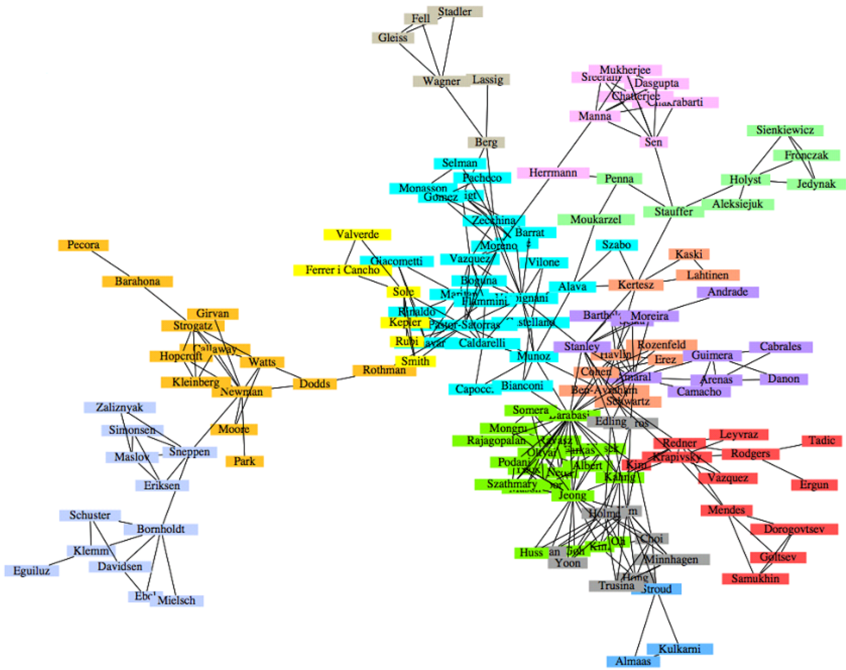
- Is the global friendship network connected?
  - Not necessary, some nodes may have no friends
  - Large complex networks often have a giant component, a connected component that contains a significant fraction of all the nodes
  - Why only one?
    - If there are two, there must not be a single connecting link between nodes in the two components, which is unlikely.



Community

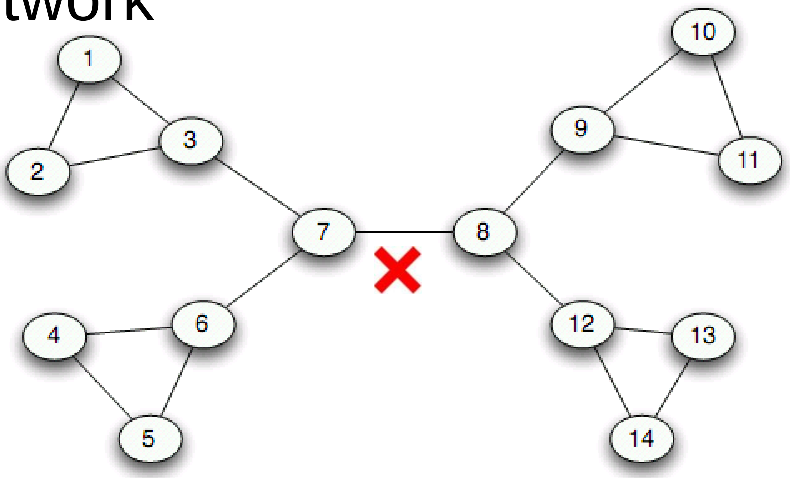
- Social networks tend to group into clusters due to different interests
- Communities
  - Sets of nodes with lots of connections inside and few to outside





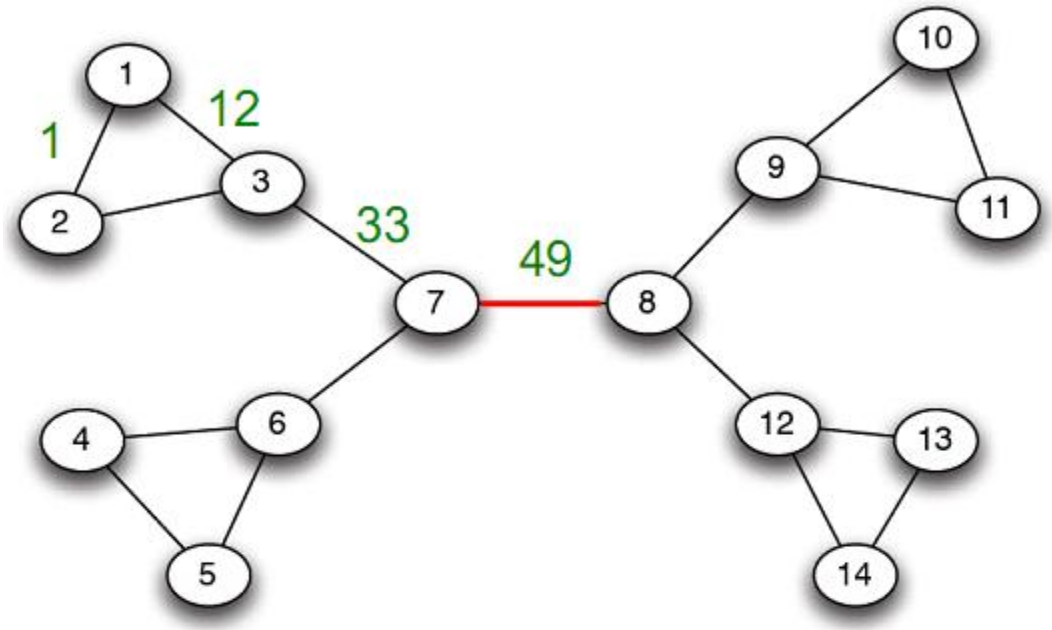
Community Detection

- How to divide a network into communities?
  - By observation?
  - Automatically?
- A possible idea
  - Finding the most important edges to divide the network
  - Imagine traffic flows in the network



Most important edges?

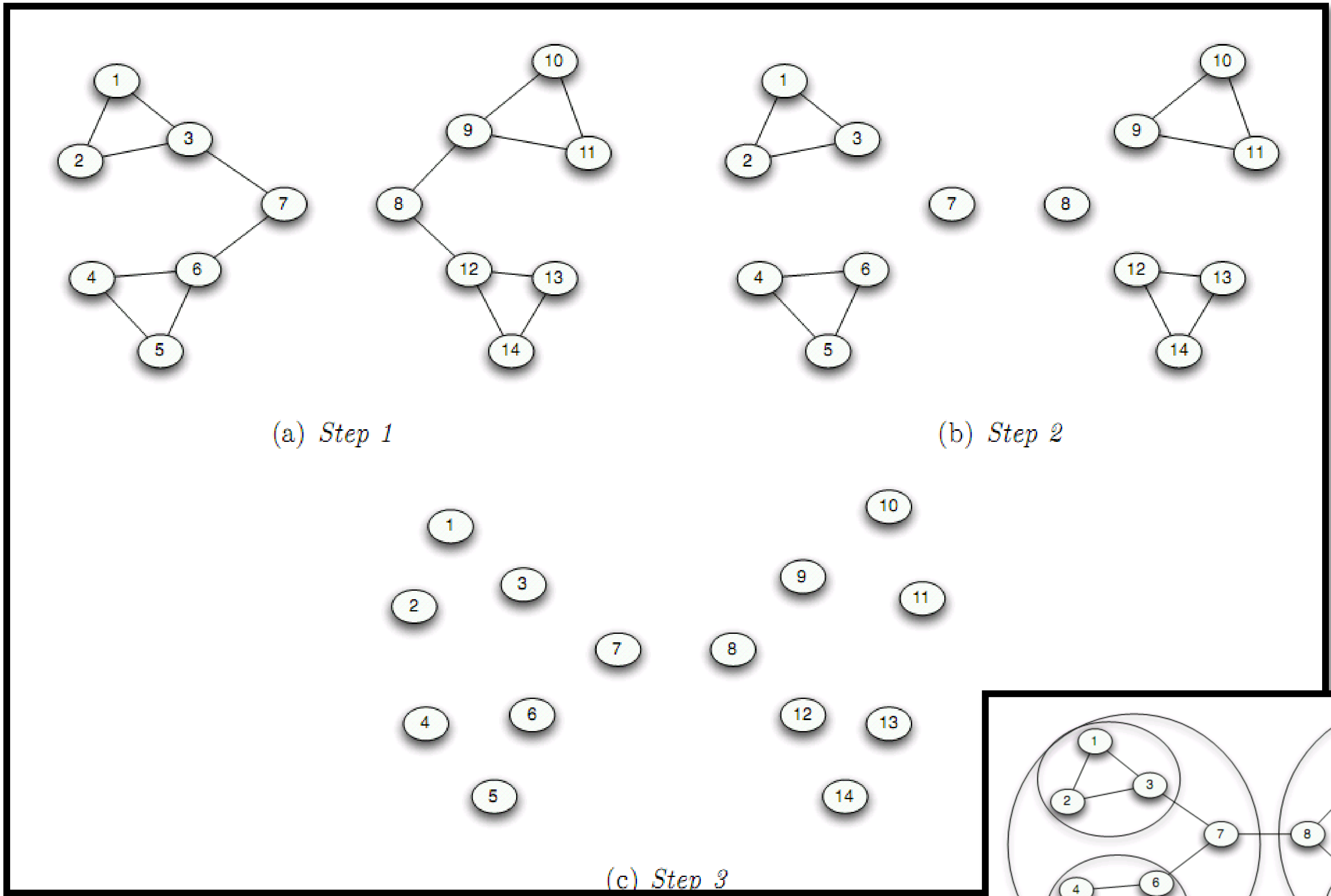
- Metric: Betweenness
  - The number of shortest paths passing through the edge



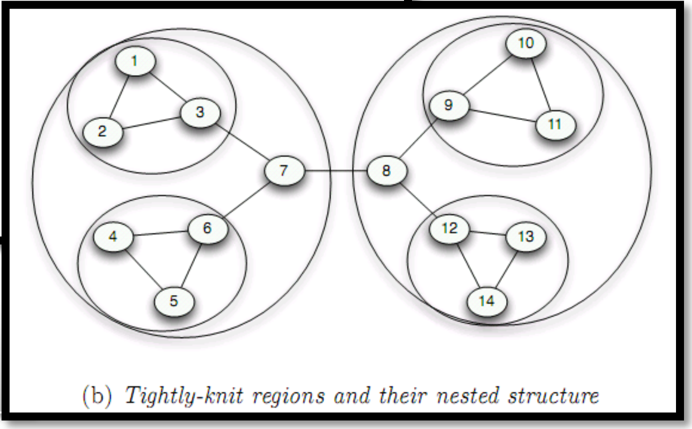
Girvan-Newman Algorithm

- Process
  - 1. Calculate betweenness of each edge
  - 2. remove edges with highest betweenness
  - 3. repeat 1,2 until the number of communities reaches a threshold or no edges are left

Example



(a) *Step 1* (b) *Step 2* (c) *Step 3*



(b) *Tightly-knit regions and their nested structure*

- ➢ Works for undirected unweighted graph
- ➢ Gives a hierarchical decomposition of the network

Louvain Method

- Iterative approach based on modularity:
  - Concentration of nodes within modules compared to random distribution of links
- Idea:
  - Start with small (1-node) communities
  - Optimize modularity on small level
  - Aggregate nodes in the same community and build a new network existing of aggregate nodes
  - Repeat until no modularity gain is possible
- Can also provide hierarchical structure
- Seems to run in O(n log n)

Example

Blondel et al.: Fast unfolding of communities in large networks, Journal of Statistical Mechanics, 2008

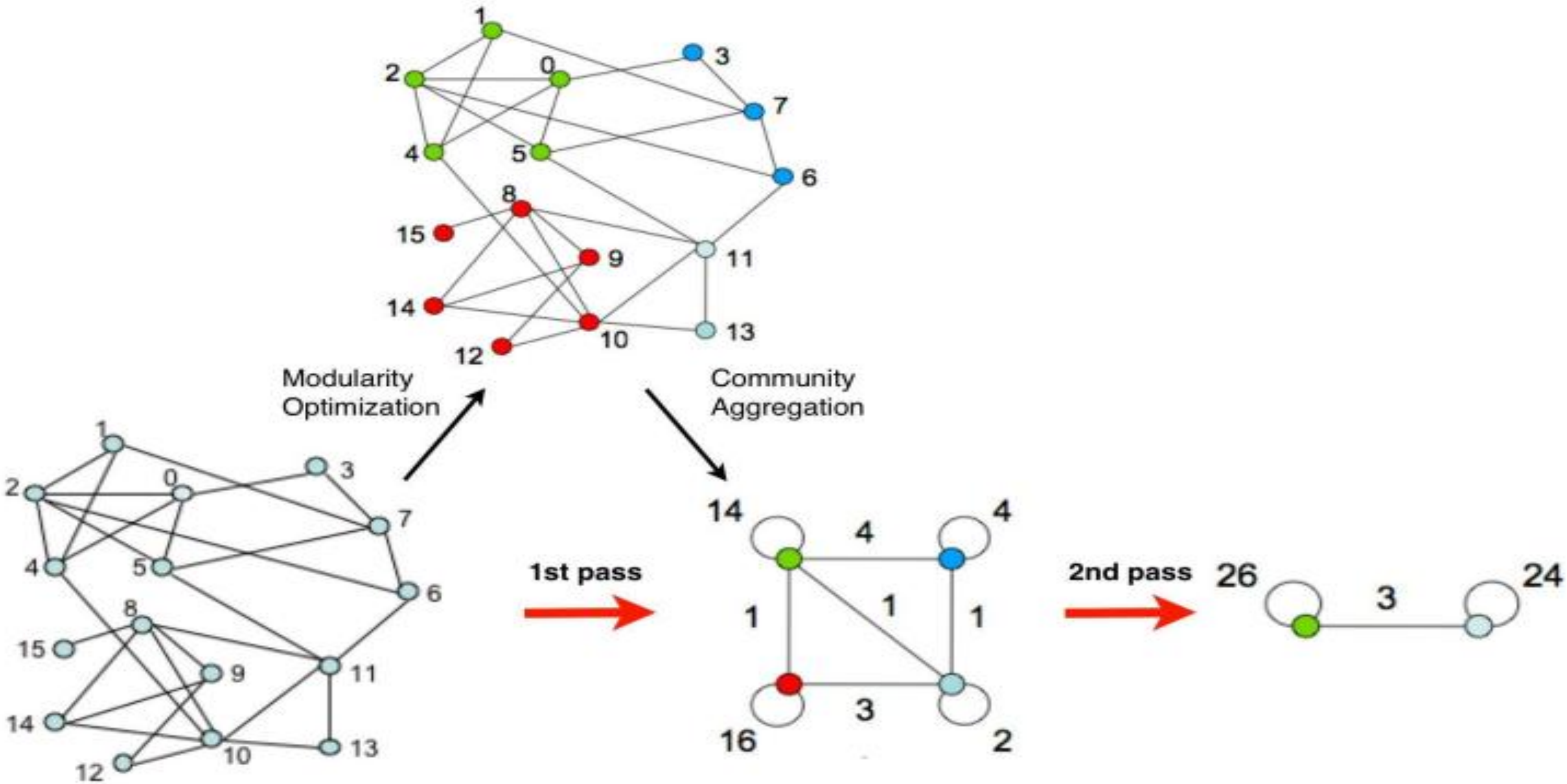
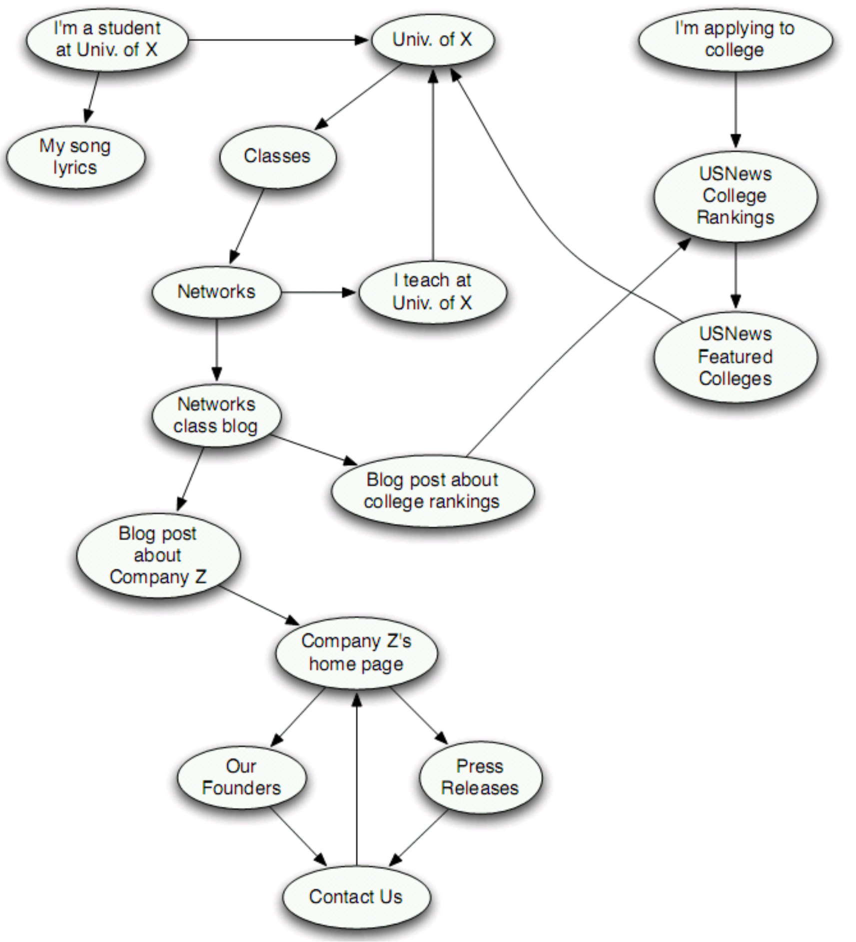


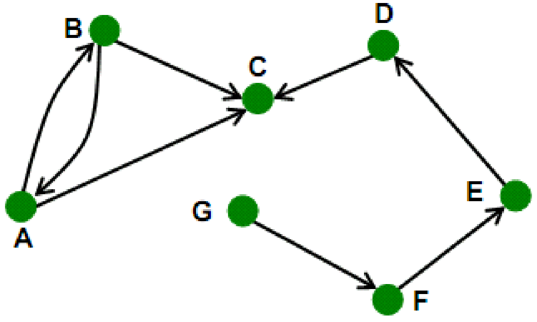
Figure 1. Visualization of the steps of our algorithm. Each pass is made of two phases: one where modularity is optimized by allowing only local changes of communities; one where the found communities are aggregated in order to build a new network of communities. The passes are repeated iteratively until no increase of modularity is possible.

The Structure of the Web

- Web as a directed graph
  - Nodes: pages
  - Edges: hyperlinks (directed)
- Question: What does Web look like at a global level?
  - Giant component?
  - Small Communities?



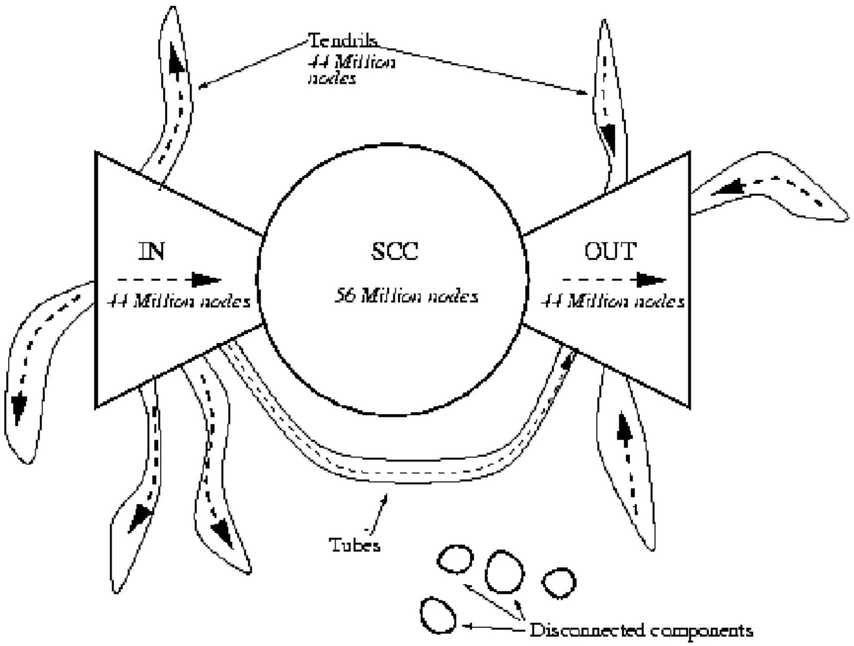
Directed Graph



- Path: directional
- Strong connectivity
  - A directed graph is strongly connected if there is a path from every node to every other node
- Strongly connected component (SCC)
  - A subset of the nodes that (i) every node in the subset has a path to every other; and (ii) the subset is not part of some larger set with the property that every node can reach every other.

The Bow-Tie Structure of the Web

- 250 million pages, 1.5 billion links (1999)
- A giant SCC (56 million nodes)
- IN set (44 million nodes)
  - Nodes that can reach the giant SCC but cannot be reached from it
- OUT set (44 million nodes)
  - Nodes that can be reached from the giant SCC but cannot reach it
- Tendrils (44 million nodes)
  - The nodes reachable from IN that cannot reach the giant SCC
  - The nodes that can reach OUT but cannot be reached from the giant SCC.
- Tubes
  - The nodes reachable from IN to OUT
- Disconnected



Analysing Social Networks

Online Social Network Analysis

- A huge research field
- Analysis of…
  - …structure (social graphs, discussed before)
  - …content
  - …much more

- In this lecture:
  - Structure analysis

Online Social Network Analysis

- Mislove et. al: *Measurement and Analysis of Online Social Networks* (IMC’07)
  - One of the most cited papers wrt. OSN analysis (>1600 citations)

- Gilbert and Karahalios: *Predicting tie strength with social media* (CHI’09)
  - Best paper at SIGCHI 2009

Online Social Network Analysis

- ***Mislove et. al: Measurement and Analysis of Online Social Networks (IMC'07)***
- Large-scale study of 11.3 million users (entities) and 328 million links in four different OSNs
- Methodology: Crawling
  - Back in 2007: efficient
  - Nowadays: often inefficient because of increasing privacy settings in OSNs
    - Content often not publicly accessible (think of your own Facebook accounts, where this hopefully holds)

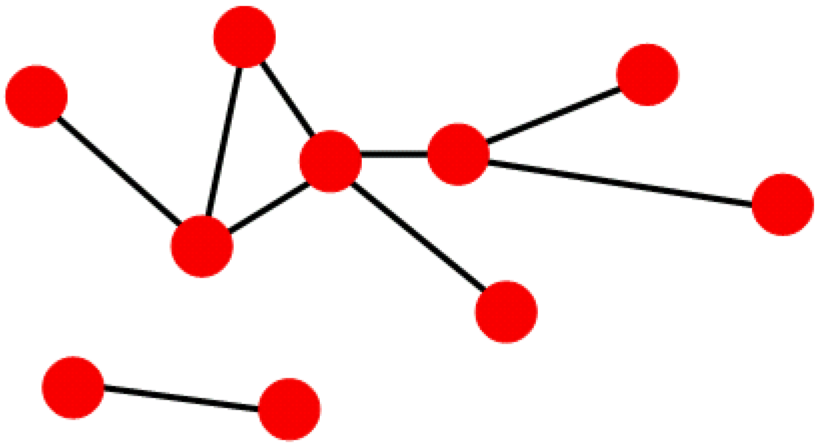
computer NETWORKS

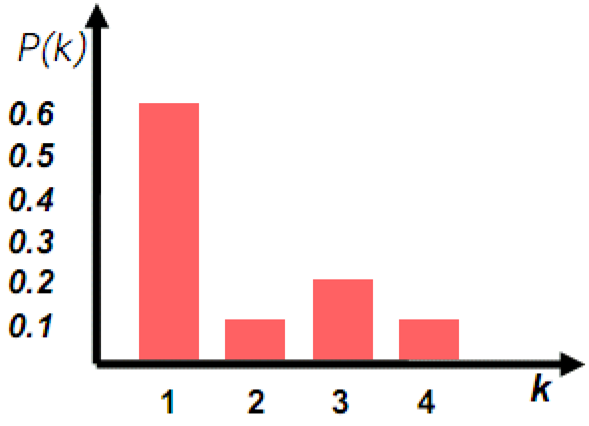
Online Social Network Analysis

- Most important results:
  - 1) Entity/user degrees follow a power-law distribution

Power-Law Distribution?

- Degree distribution P(k)
  - Probability that a randomly chosen nodes has degree k
  - N_k: number of nodes with degree k
  - $P(k)=N_k/N$

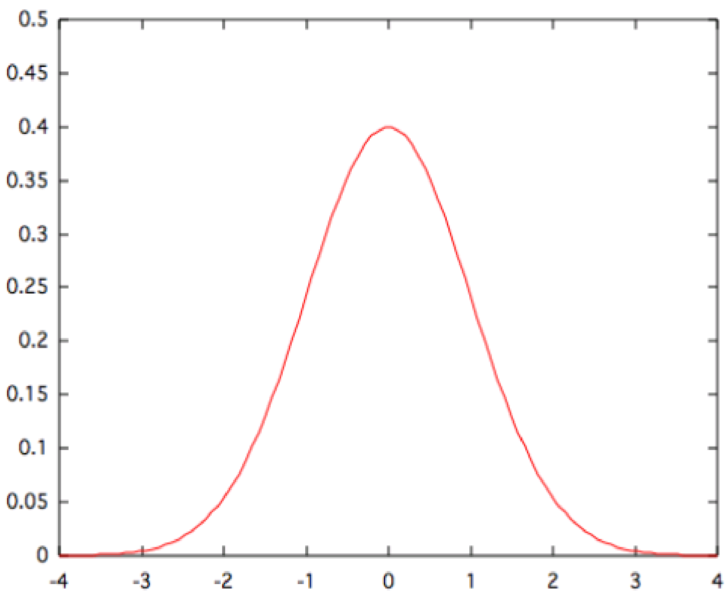




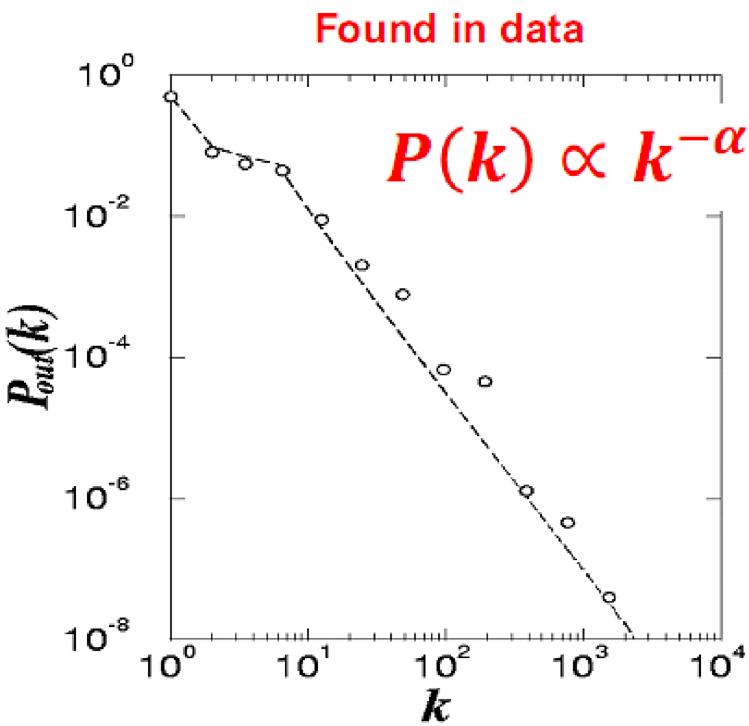
- Popularity of nodes in social networks
  - Imbalance
  - 20% of web pages receives 80% visits
  - Celebrities in Twitter have millions of fans
  - A few rich people own a large amount of wealth
- Node degree distribution
  - What fraction of all nodes have degree k? P(k)=?
  - Normal distribution? – for random graph
    P(k)=exponential function of -k

Guess: Normal distribution

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$



Real network: Power-law



- Power-law distribution
  - Let f(k) be the fraction of items have value k

$$f(k) = zk^{-\alpha}$$

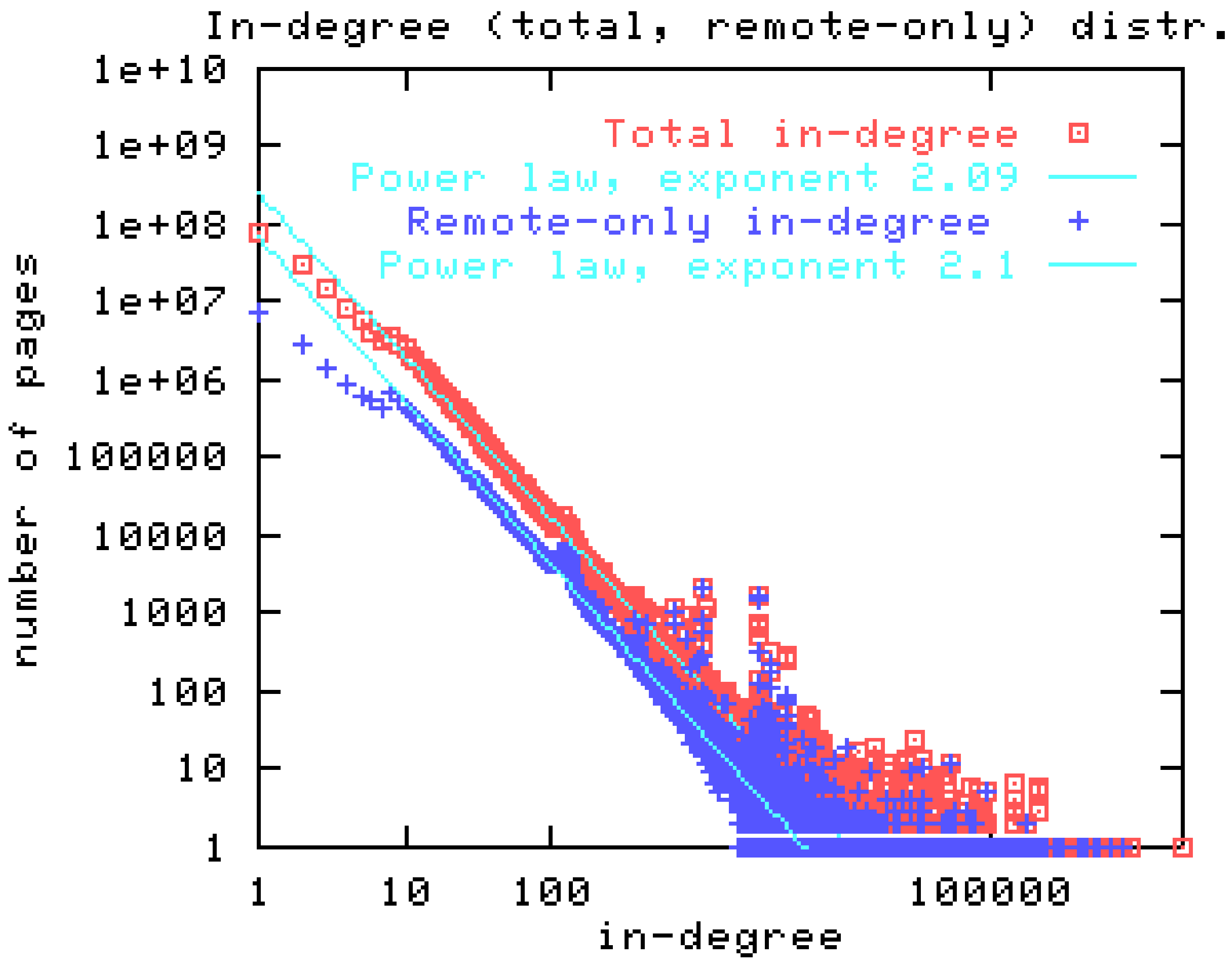
where α and z are constants

- Taking logarithms of both sides

$$log f(k) = log z - \alpha log k$$

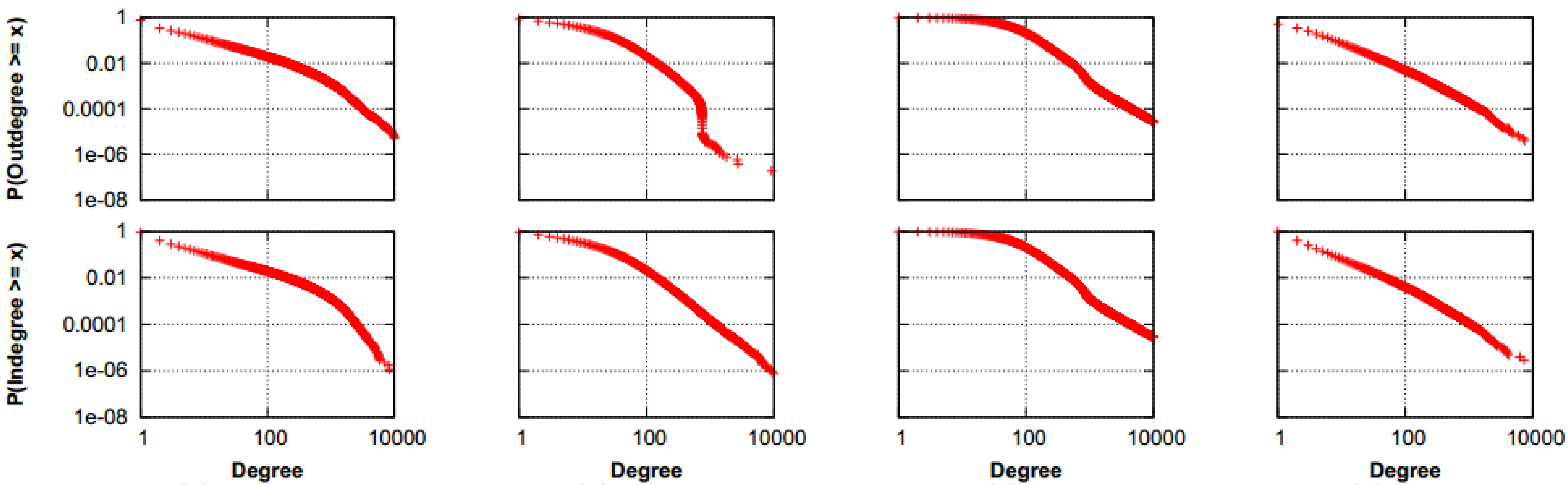
- Testing for power-law distribution
  - If we draw k and f(k) in “log-log” scale, it shows a straight line

Node Degree of Websites



Online Social Network Analysis

- Most important results:
  - 1) Entity/user degrees follow a power-law distribution
    - Power-law was observed in offline social networks, Mislove et al. confirm it for online social networks
    - Power-law: Few entities have very large degree, most entities have low degree
    - In the figures: CDF - consider the scale on the y-axis!



(a) Flickr (b) LiveJournal (c) Orkut (d) YouTube

Online Social Network Analysis

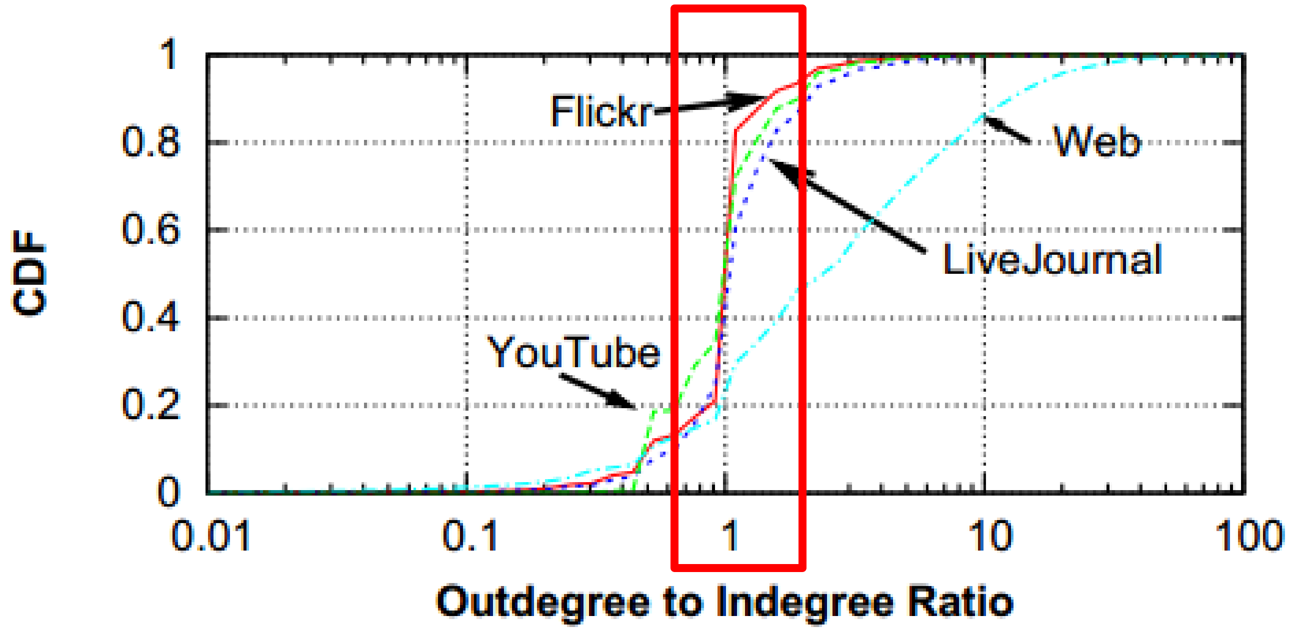
- Most important results:
  - 2) Correlation between in-degree and out-degree
    - Link symmetry: If a link for a source to a destination (i.e., between two entities) is established, the destination often also establishes a link back to the source in directed graphs (e.g. Twitter following)

Network	Outdegree α	Outdegree D	Indegree α	Indegree D
Web [12]	2.67	-	2.09	-
Flickr	1.74	0.0575	1.78	0.0278
LiveJournal	1.59	0.0783	1.65	0.1037
Orkut	1.50	0.6319	1.50	0.6203
YouTube	1.63	0.1314	1.99	0.0094

Table 2: Power-law coefficient estimates (α) and corresponding Kolmogorov-Smirnov goodness-of-fit metrics (D). The Flickr, LiveJournal, and YouTube networks are well approximated by a power-law.

Online Social Network Analysis

- Most important results:
  - 2) Correlation between in-degree and out-degree
    - Figure: for most users (~60%) ratio of out degree to in degree is ~1



Online Social Network Analysis

- Most important results:
  - 3) Confirmation of the Small-World-Phenomenon (SWP):
    - SWP: Everybody is connected to everybody within 6 hops on average (see “six degrees of separation”)
    - For OSNs: between 4 and 6 hops (see table)!

Network	Avg. Path Len.	Radius	Diameter
Web [12]	16.12	475	905
Flickr	5.67	13	27
LiveJournal	5.88	12	20
Orkut	4.25	6	9
YouTube	5.10	13	21

Online Social Network Analysis

- ***Gilbert and Karahalios: Predicting tie strength with social media (CHI'09)***

- An analysis of the strength of the links between entities
  - E.g., how much trust does a Facebook friendship imply?

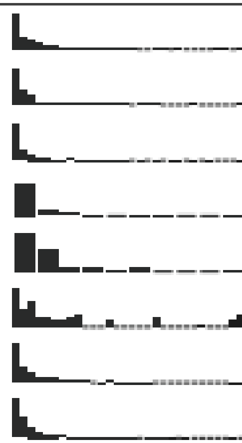
- Definition:
  - "The strength of a tie *(link)* is a (probably linear) combination of the amount of time, the emotional intensity, the intimacy (mutual confiding), and the reciprocal services which characterize the tie."

    [Granovetter 1973]

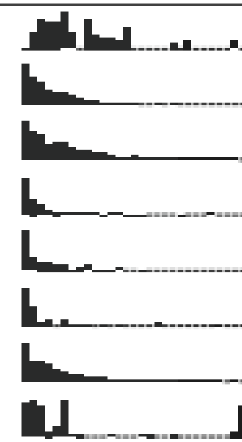
computer NETWORKS

Online Social Network Analysis

- Main result:
  - A list of variables to determine/predict the real strength of the representative tie of an offline relationship (e.g., a Facebook friendship between work colleagues) in an OSN
- Figure: distribution of occurrence of the variable between two entities
  - e.g., most users exchange few wall words with each other (long tail of the distribution) but there were two users who exchanged 9549 wall words (max) with each other



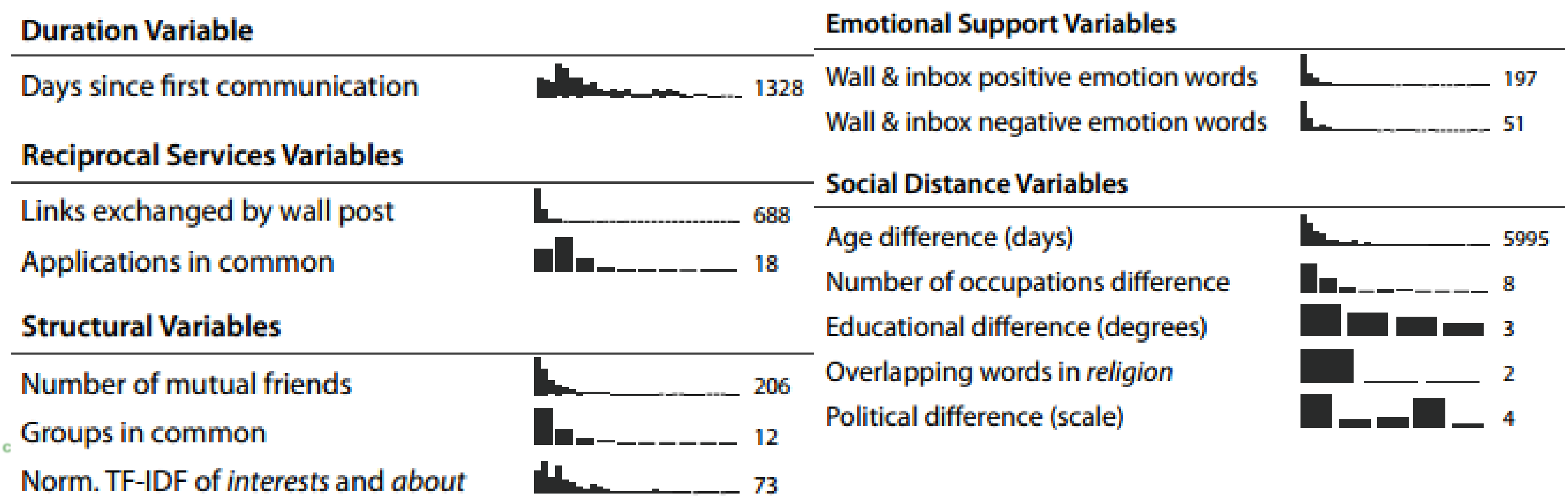
Predictive Intensity Variables	Distribution	Max
Wall words exchanged		9549
Participant-initiated wall posts		55
Friend-initiated wall posts		47
Inbox messages exchanged		9
Inbox thread depth		31
Participant's status updates		80
Friend's status updates		200
Friend's photo comments		1352



Intimacy Variables	Distribution	Max
Participant's number of friends		729
Friend's number of friends		2050
Days since last communication		1115
Wall intimacy words		148
Inbox intimacy words		137
Appearances together in photo		73
Participant's appearances in photo		897
Distance between hometowns (mi)		8182
Friend's relationship status	6% engaged, 32% married, 30% single, 30% in relationship	

Online Social Network Analysis

- Main result:
  - A list of variables to determine/predict the real strength of the representative tie of an offline relationship (e.g., a Facebook friendship between work colleagues) in an OSN



Variable	Max
Duration Variable	
Days since first communication	1328
Reciprocal Services Variables	
Links exchanged by wall post	688
Applications in common	18
Structural Variables	
Number of mutual friends	206
Groups in common	12
Norm. TF-IDF of *interests* and *about*	73
Emotional Support Variables	
Wall & inbox positive emotion words	197
Wall & inbox negative emotion words	51
Social Distance Variables	
Age difference (days)	5995
Number of occupations difference	8
Educational difference (degrees)	3
Overlapping words in *religion*	2
Political difference (scale)	4

Online Social Network Analysis

- Out of these variables, build a linear model to predict the strength of a relation

$$s_i = \alpha + \beta R_i + \gamma D_i + N(i) + \epsilon_i$$

$$N(i) = \lambda_0 \mu_M + \lambda_1 med_M + \sum_{k=2}^{4} \sum_{s \in M} \lambda_k (s - \mu_M)^k + \lambda_5 min_M + \lambda_6 max_M$$

$$M = \{s_j : j \text{ and } i \text{ are mutual friends}\}$$

- S_i = tie strength with i-th friend
- R_i = vector of predictive variables (seen on slide before)

Online Social Network Analysis

- Out of these variables, build a linear model to predict the strength of a relation

$$s_i = \alpha + \beta R_i + \gamma D_i + N(i) + \epsilon_i$$

$$N(i) = \lambda_0 \mu_M + \lambda_1 med_M + \sum_{k=2}^{4} \sum_{s \in M} \lambda_k (s - \mu_M)^k + \lambda_5 min_M + \lambda_6 max_M$$

$$M = \{s_j : j \text{ and } i \text{ are mutual friends}\}$$

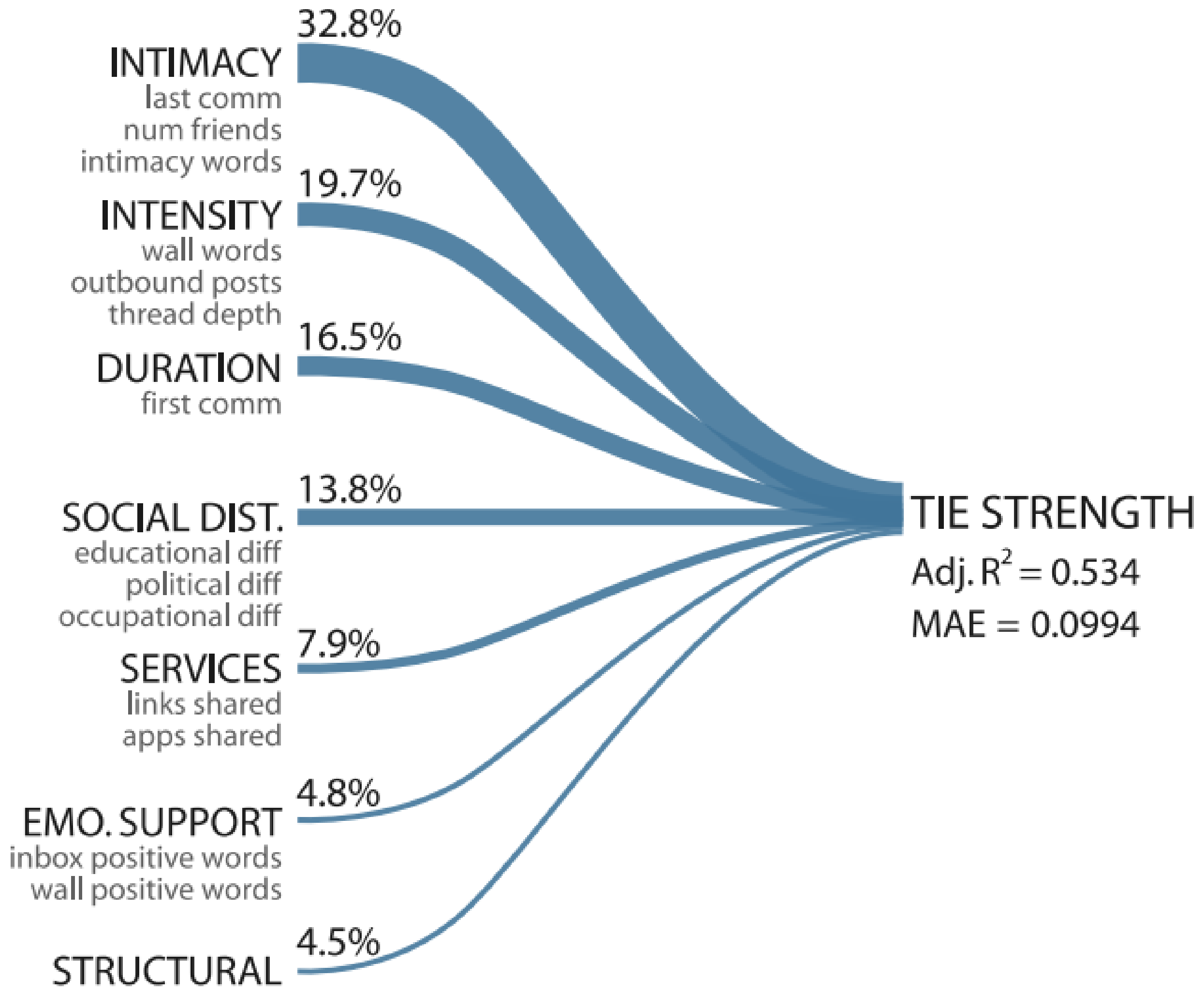
- D_i = represents the pairwise interactions of the top 13 variables
- N_i = an expression of structural strength due to the graph structure (consider the tie strength between mutual friends as well)
- ϵ_i = error term

Online Social Network Analysis

- To validate the model, ask participants of the study (from whose data the variables were derived) to rate the strength of their ties and compare the answer to the result of the model
- Result: The model predicts the correct tie strength with ~90% accuracy

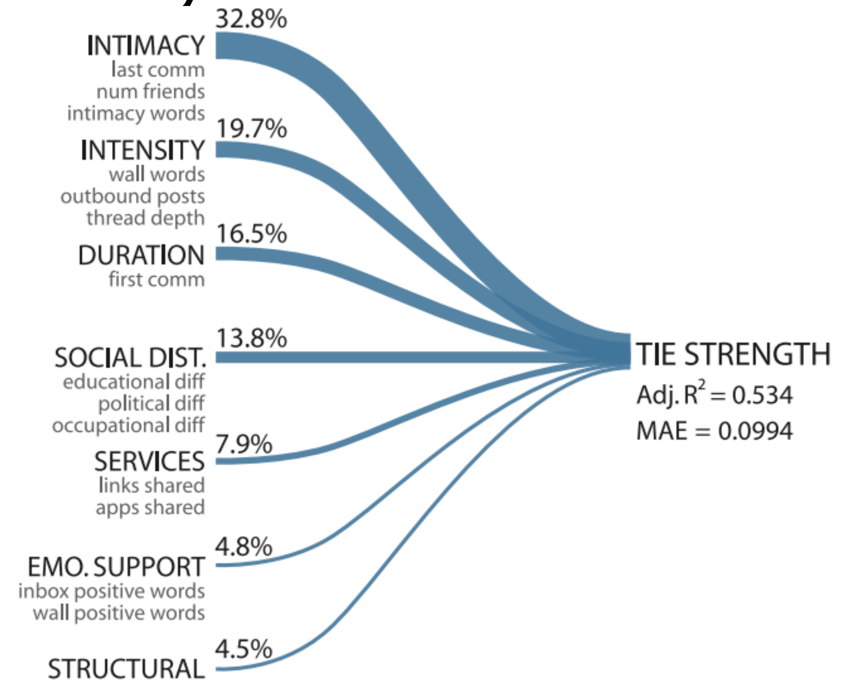
Online Social Network Analysis

- So…what are the factors contributing to tie strength?



Online Social Network Analysis

- Most importantly: the existence of the link itself only contributes 4.5% (structural)
  - Has impact on research that assumed that the mere existence of a link can substantiate a strong tie
- Instead, intimacy and intensity contribute over 50% to the tie strength



computer NETWORKS

Conclusion (1/2)

- Introduction to the structure and modeling of social networks
  - Directed and undirected graphs in different networks
  - (Giant) Connected components
  - Communities and their detection

- We have seen some analysis on OSNs – there is a lot more research going on!

Conclusion (2/2)

- Next week: Last lecture! (Also on OSNs)

- Remember to register for the exam, which will be held on July 24th!