Porto Business School

/ University of Porto

Social Network Analysis

Part 4 – Network Structure

Prof. Eduarda Mendes Rodrigues

Porto Business School
/ University of Porto

Part 4 - Network Structure

- Community structure
  - Strength of weak ties, community detection and betweenness centrality
  - Homophily, selection and social influence
  - Modularity, Graph partitioning methods
- Properties of real-world networks
  - Small world phenomenon: Clustering, Milgram's small world experiment
  - Structure and randomness, Small world models, Power-laws


Community Structure

Community

- Formed by individuals such that those within a group interact with each other more frequently than with those outside the group
  - a.k.a. group, cluster, cohesive subgroup, module

- **Community detection**: discovering groups of nodes in a network where the nodes' group memberships are not explicitly given

- Definition of a community may be subjective
  - densely-knit community (e.g. k-clique), each component is a community, etc.

Criteria to Identify Communities

Porto Business School / University of Porto

Main community detection approaches

- **Node-centric**: each node in a group satisfies certain properties
- **Group-centric**: the group, as a whole, has to satisfy certain properties regardless of the node-level properties
- **Network-centric**: partition the whole network into several disjoint sets
- **Hierarchy-centric**: construct a hierarchical structure of communities

Node-centric Community Detection

- Nodes satisfy different properties
    - Complete mutuality (cliques)
    - Reachability of members (e.g., k-clique, paths)
    - Node degrees
    - Relative frequency of ‘within community’ vs. ‘outside community’ links
- Commonly used in **traditional social network analysis**

Clique Percolation Method (CPM)

- Clique is a very strict definition!

- Normally use cliques as **a core or a seed** to find larger communities

- CPM is such a method to find **overlapping communities**

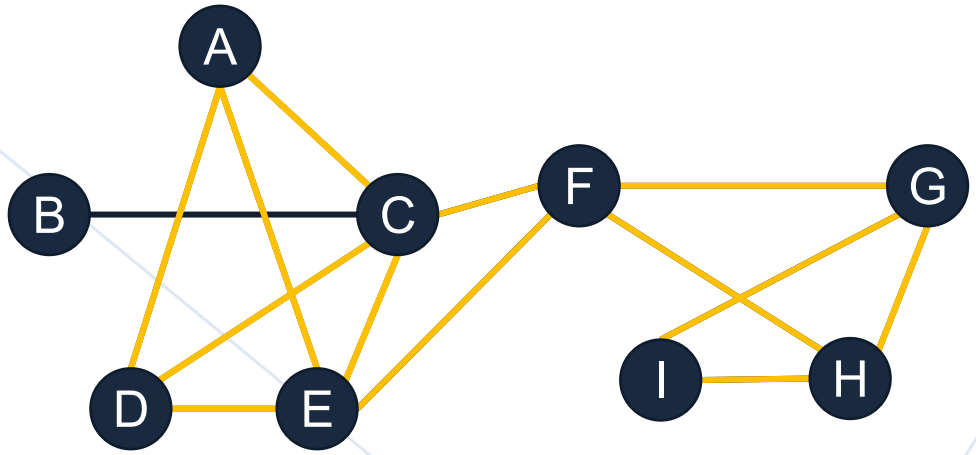
    - **Input:** a network and parameter k specifying the clique size

    - **Procedure:**
        - Find out all cliques of size k in a given network
        - Construct a **clique graph**. Two cliques are adjacent if they **share k-1 nodes**
        - Each **connected component** in the clique graph forms a community

CPM Example

- Given the network below and k=3, apply the CPM method and determine the resulting communities (clusters)
- How many K_3 cliques are there?

Exercise: How many K3 cliques are there?

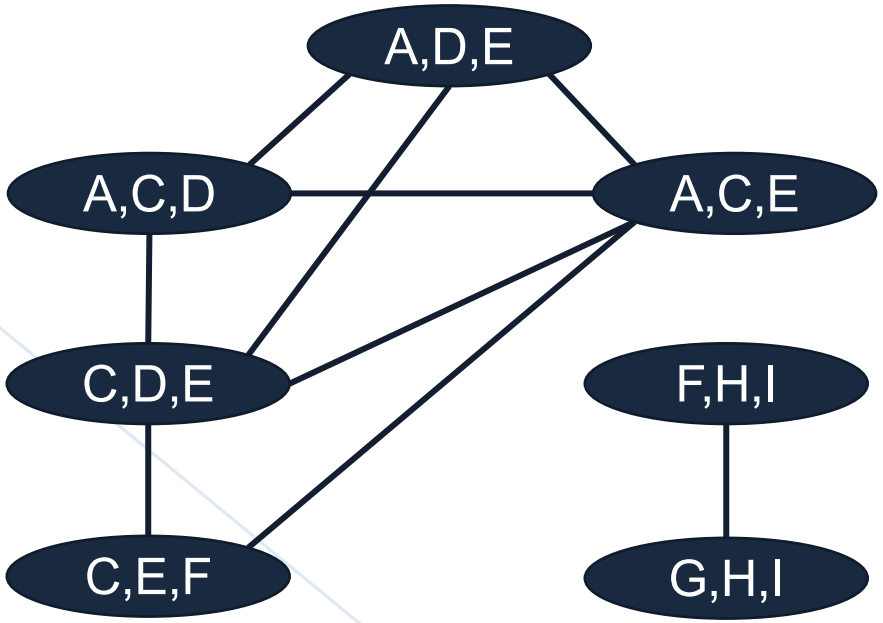


K_3

{A,D,E}
{A,C,D}
{A,C,E}
{C,D,E}
{C,E,F}
{F,G,H}
{G,H,I}

CPM Example

Exercise: Construct the clique graph. How many nodes and edges are there?

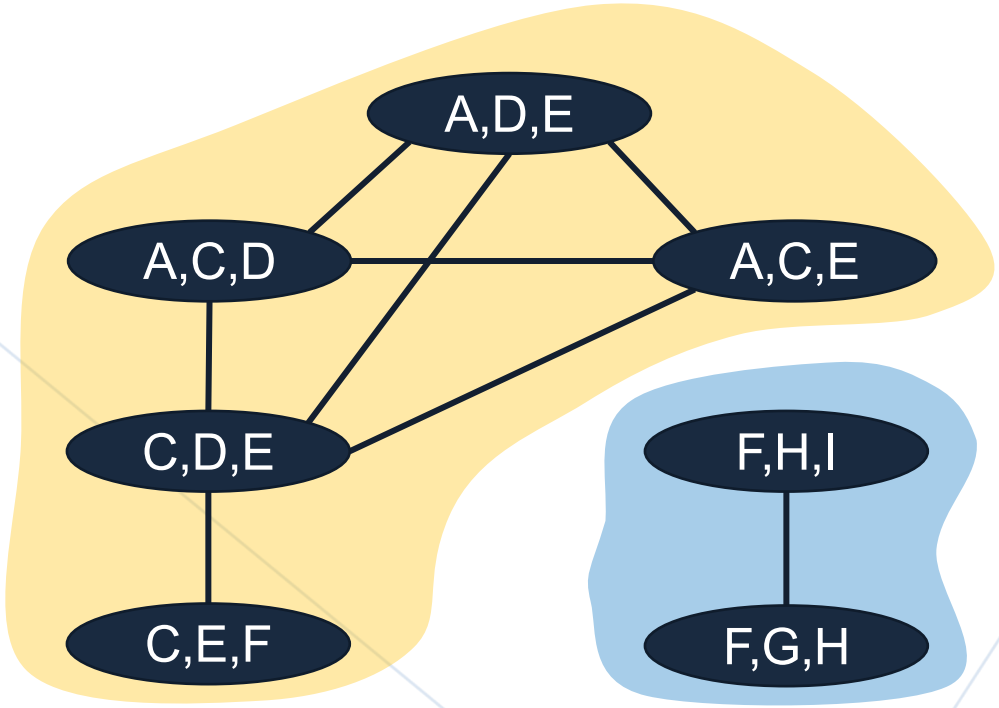


K_3

{A,D,E}
{A,C,D}
{A,C,E}
{C,D,E}
{C,E,F}
{F,G,H}
{G,H,I}

CPM Example

Exercise: How many communities are there?



Community 1

{ A,B,C,D,E,F }

Community 2

{ F,G,H,I }

Group-Centric Community Detection

Porto Business School / University of Porto

- The group-centric criterion requires the whole group to satisfy a certain condition
  - E.g., the group density being higher than a given threshold

- A subgraph $G(V, E)$ is a $\gamma-$dense quasi-clique if $\frac{2|E|}{|V|(|V|-1)} \geq \gamma$
  - I.e. 2x ratio of edges out of all possible edges needs to be at least equal to γ

- A similar strategy to that of cliques can be used
  - Sample a subgraph, and find a maximal $\gamma-$dense quasi-clique (say, of size $|V|$)
  - Remove nodes with degree **less than** the average degree

Network-Centric Community Detection

- **Goal:** partition nodes of a network into disjoint sets
- Network-centric criterion needs to consider the connections within a network at the global level
- Some approaches:
    - Clustering based on vertex similarity
    - Block model approximation
    - Modularity maximization
    - Spectral clustering
    - Etc.

Clustering based on Vertex Similarity

Porto Business School / University of Porto

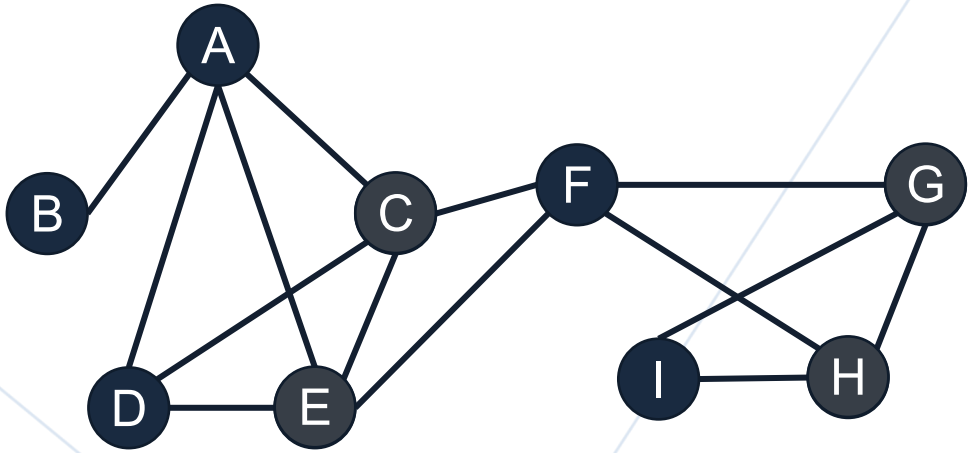
- Apply k-means or similarity-based clustering to nodes
  - Attribute based similarity
  - Ego-network structural similarity

- Vertex similarity is defined in terms of **the similarity of their neighborhood**

Clustering based on Vertex Similarity

- **Structural equivalence:** two nodes are structurally equivalent if and only if they are connecting to the same set of nodes, i.e. they can be swapped without modifying the overall structure of the network

Exercise: Can you find structurally equivalent nodes?



- Nodes C and E are structurally equivalent;
- So are nodes G and H.

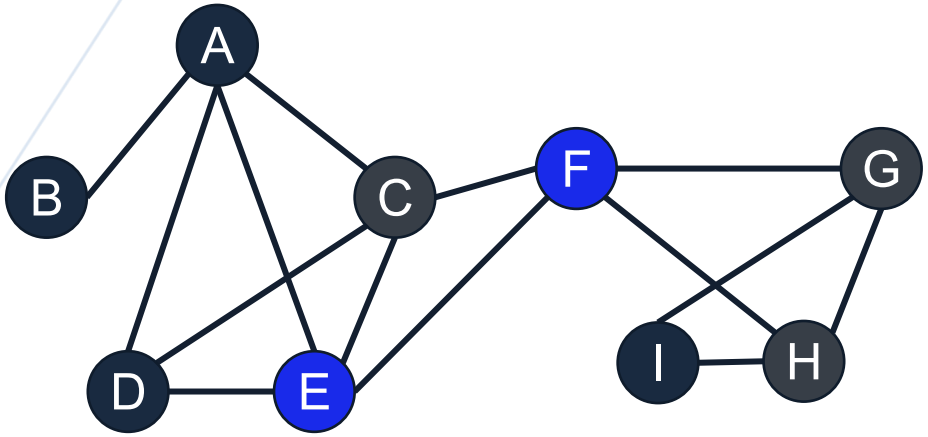
Structural equivalence is too restrictive for practical use!

Vertex Similarity

- Given two nodes v_i and v_j with a neighbourhood (i.e. set of adjacent nodes) N_i and N_j standard similarity measures can be applied
  - Jaccard Similarity: $Jac(v_i, v_j) = \frac{|N_i \cap N_j|}{|N_i \cup N_j|}$
  - Cosine similarity: $Cos(v_i, v_j) = \frac{|N_i \cap N_j|}{\sqrt{|N_i| \cdot |N_j|}}$

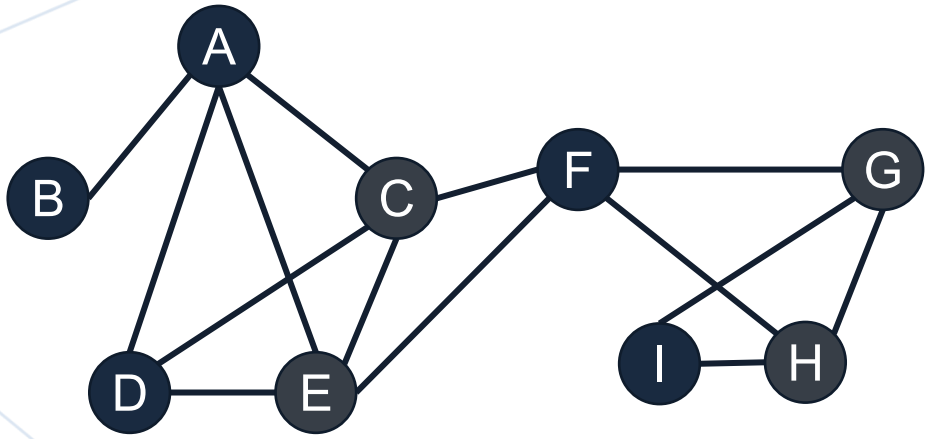
$$Jac(E,F) = \frac{|\{C\}|}{|\{A,C,D,E,F,G,H\}|} = 1/7$$

$$Cos(E,F) = \frac{1}{\sqrt{4 \cdot 4}} = 1/4$$



Block Models

- What is the ideal **block structure**?



Adjacency matrix

	A	B	C	D	E	F	G	H	I
A	-	1	1	1	1	0	0	0	0
B	1	-	0	0	0	0	0	0	0
C	1	0	-	1	1	1	0	0	0
D	1	0	1	-	1	0	0	0	0
E	1	0	1	1	-	1	0	0	0
F	0	0	1	0	1	-	1	1	0
G	0	0	0	0	0	1	-	1	1
H	0	0	0	0	0	1	1	-	1
I	0	0	0	0	0	0	1	1	-

Block Models

- The optimal solution S corresponds to the **top eigenvectors** of adjacency matrix

	A	B	C	D	E	F	G	H	I
A	-	1	1	1	1	0	0	0	0
B	1	-	0	0	0	0	0	0	0
C	1	0	-	1	1	1	0	0	0
D	1	0	1	-	1	0	0	0	0
E	1	0	1	1	-	1	0	0	0
F	0	0	1	0	1	-	1	1	0
G	0	0	0	0	0	1	-	1	1
H	0	0	0	0	0	1	1	-	1
I	0	0	0	0	0	0	1	1	-

$$min\|A - S\,\Sigma S^T\|^2$$

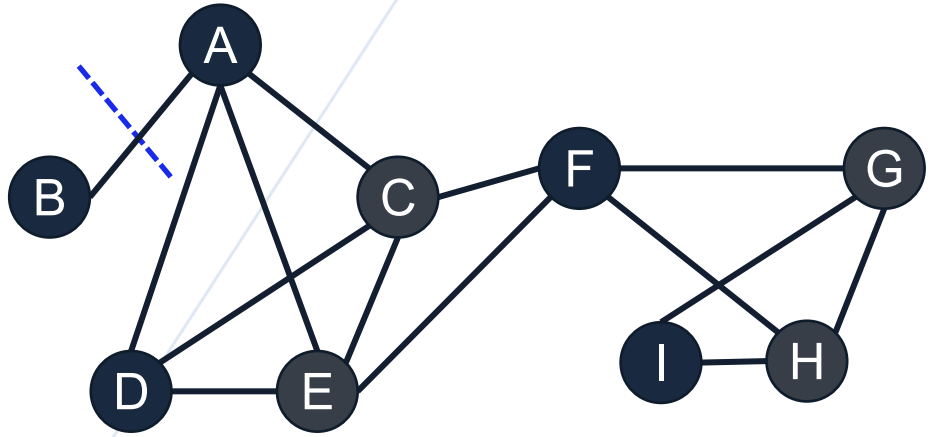
A: adjacency matrix
S: community membership index (numerical value)

Ideal block structure

-	1	1	1	1	0	0	0	0
1	-	1	1	1	0	0	0	0
1	1	-	1	1	0	0	0	0
1	1	1	-	1	0	0	0	0
1	1	1	1	-	0	0	0	0
0	0	0	0	0	-	1	1	1
0	0	0	0	0	1	-	1	1
0	0	0	0	0	1	1	-	1
0	0	0	0	0	1	1	1	-

Graph Cut

- Most interactions are within group whereas interactions across groups are few
- Community detection → minimum cut problem
- **Cut**: is a partition of vertices of a graph into two disjoint sets
- **Minimum cut problem** (Ahuja et al., 1993): find a graph partition such that the number of edges between the two sets is minimized



Ratio Cut & Normalized Cut

- **Drawback:** minimum cut often returns communities of **unbalanced** size, with one set being a singleton, e.g. node B

- Change the objective function to minimize a function that **considers the community size**:

$$\textbf{Ratio Cut} = \frac{1}{k}\sum_{i=1}^{k}\frac{\text{cut}(C_i, \bar{C}i)}{|C_i|}$$

$$\textbf{Normalized Cut} = \frac{1}{k}\sum_{i=1}^{k}\frac{\text{cut}(C_i, \bar{C}i)}{\text{vol}(C_i)}$$

C_i: a community
$|C_i|$: number of nodes in C_i
$\text{vol}(C_i)$: sum of degrees in C_i

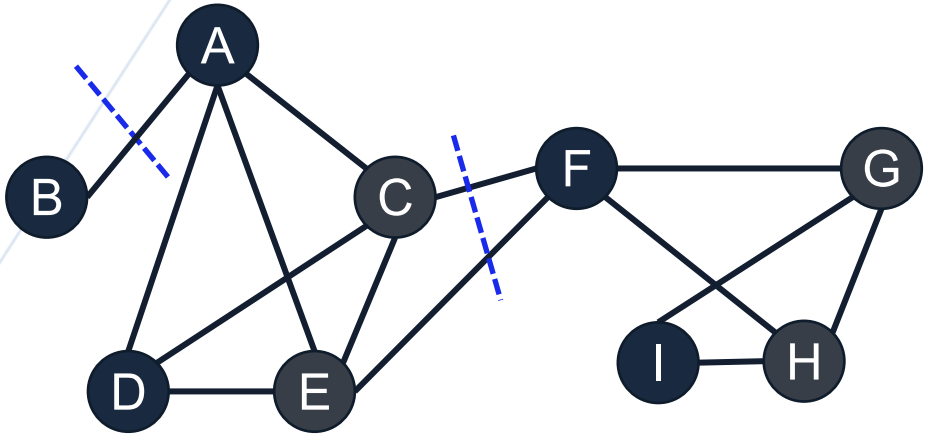
Ratio Cut & Normalized Cut Example

- Both ratio cut and normalized cut prefer a **balanced** partition

Exercise: Calculate the ratio cut and normalized cut of the partitions shown below.

- C_1= {B} and C_2={A,C,D,E,F,G,H,I} ?
  RC = 1/2 (1/1 + 1/8) = 0.56
  NC = 1/2 (1/1 + 1/27) = 0.52

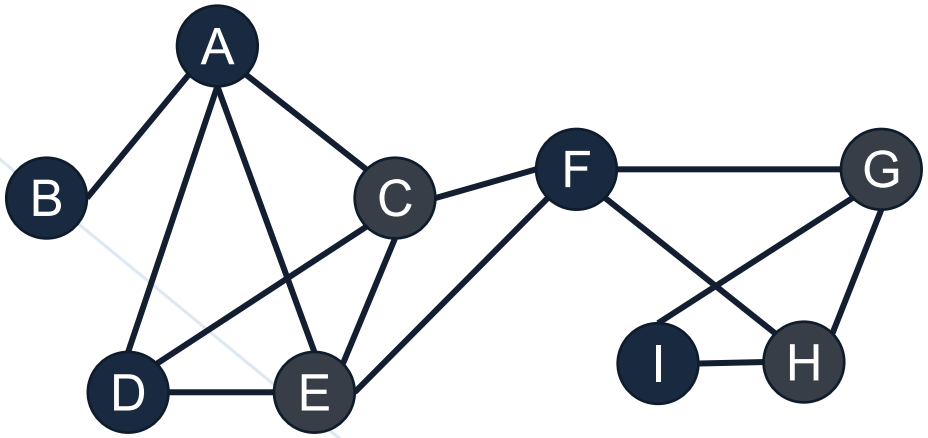
- C_1′={A,B,C,D,E} and C_2′={F,G,H,I} ?
  RC′ = 1/2 (2/5 + 2/4) = 0.45 < RC
  NC′ = 1/2 (2/16 + 2/12) = 0.15 < NC



Modularity Maximization

- **Modularity** measures the **strength of a community partition** by taking into account the degree distribution

- Given a network with *m* edges, the **expected number of edges** between two nodes with degrees d_i and d_j is $d_i d_j / 2m$



Exercise: What is the expected number of edges between nodes G and I?

3*2/ (2*14) = 3/14

Modularity Maximization

- Strength of a community: $S_l = \sum_{i \epsilon C_l, j \epsilon C_l} (A_{ij} - d_i d_j / 2m)$
- **Modularity** (Girvan and Newman, 2004): $Q = \frac{1}{2m} \sum_{l=1}^{k} S_l$
- A larger value indicates a good community structure
- Modularity Q is in the range [-1,1]
  - **Positive modularity** may indicate there is community structure in the network , i.e. the number of edges within cluster exceeds the number expected on the basis of chance
  - The higher the value, the closest the partition is to real communities (Q>0.3 usually indicates meaningful partition)

Hierarchy-Centric Community Detection

- **Goal:** build a hierarchical structure of communities based on network topology

- Allow the analysis of a network at **different resolutions / granularity levels**

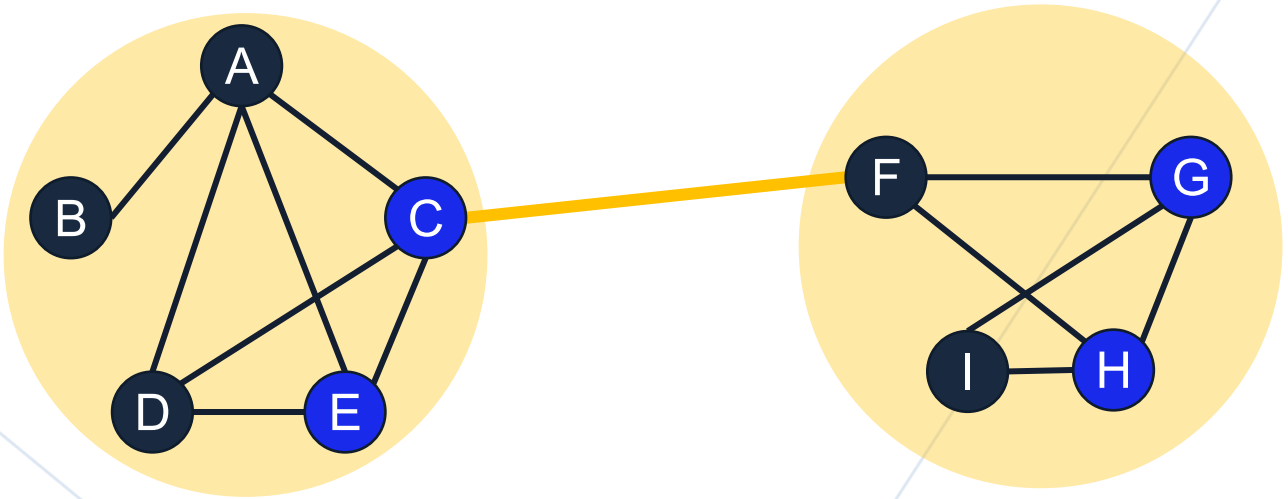
- Approaches:
  - DHC - Divisive Hierarchical Clustering (top-down)
  - AHC - Agglomerative Hierarchical Clustering (bottom-up)

Divisive Hierarchical Clustering (DHC)

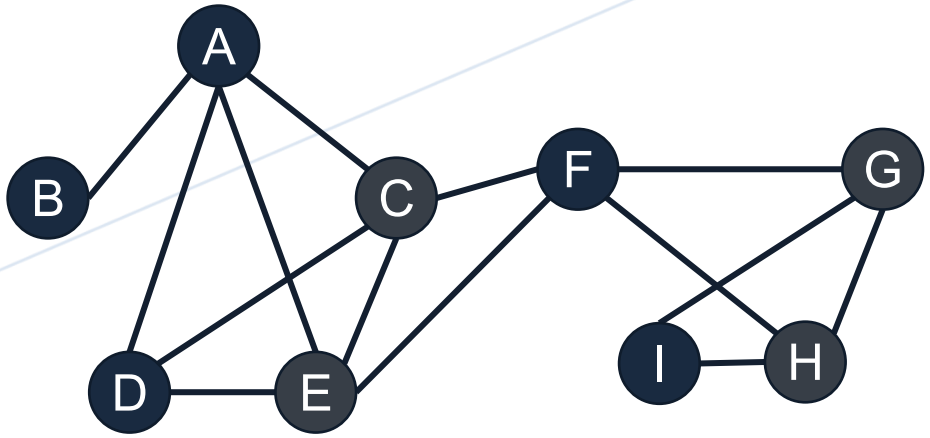
- Divisive clustering: partition nodes into several sets, each set is further divided into smaller ones
    - Network-centric partition can be applied for the partition
- Recursively remove the “weakest” link
    - Find the edge with the least strength
    - Remove the edge and update the corresponding strength of each edge
- Recursively apply the above two steps until a network is decomposed into desired number of connected components
- Each component forms a community

Edge Betweenness

- The strength of a tie can be measured by **edge betweenness** measure, which quantifies the number of shortest paths that pass along with the edge
- The edge with higher betweenness tends to be the **bridge** between two communities



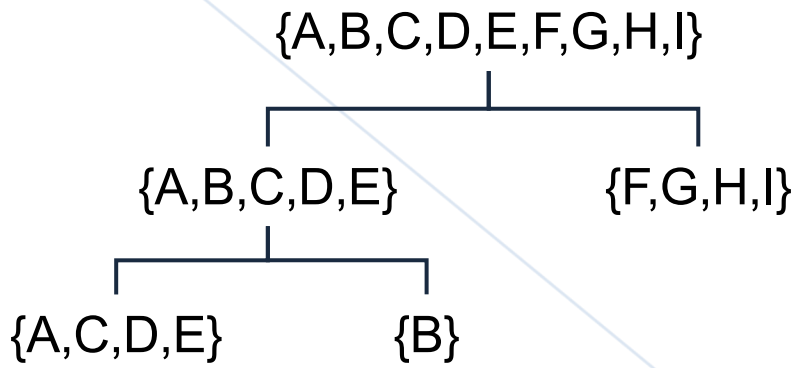
DHC based on Edge Betweenness



Initial edge betweenness

	A	B	C	D	E	F	G	H	I
A	0	8	6	2	6	0	0	0	0
B	8	0	0	0	0	0	0	0	0
C	6	0	0	3	1	10	0	0	0
D	2	0	3	0	3	0	0	0	0
E	6	0	1	3	0	10	0	0	0
F	0	0	10	0	10	0	9	9	0
G	0	0	0	0	0	9	0	1	4
H	0	0	0	0	0	9	1	0	4
I	0	0	0	0	0	0	4	4	0

DHC dendogram

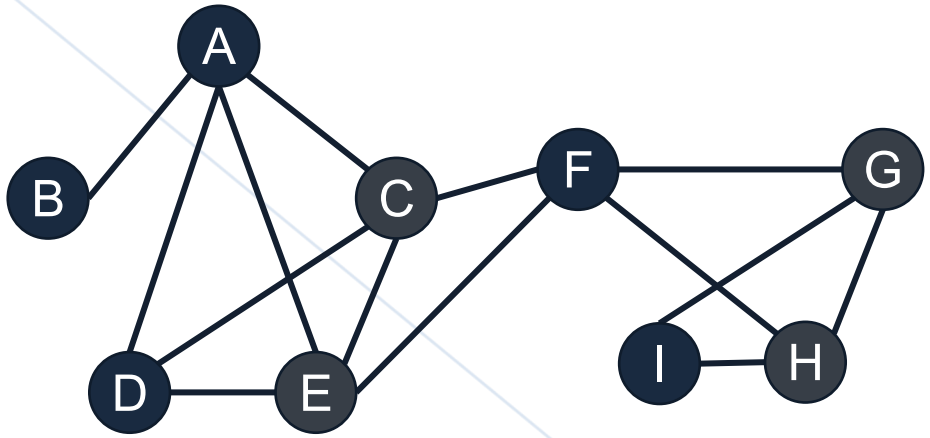


After removing e(C, F), the betweenness of e(E, F) becomes 20, which is the highest

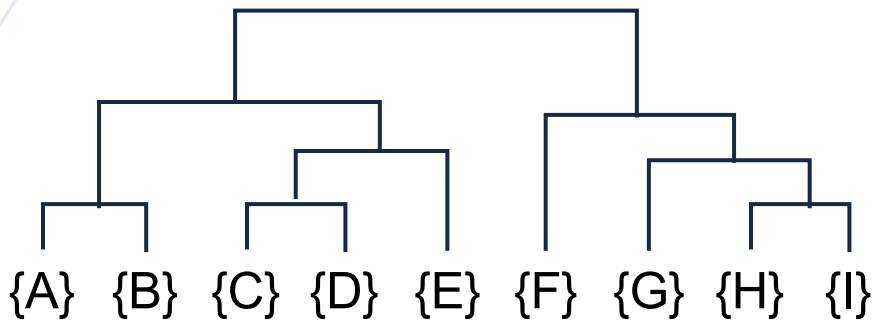
After removing e(E, F), the edge e(A, B), has the highest betweenness value 4, and should be removed

Agglomerative Hierarchical Clustering

- Initialize each node as a community
- Merge communities successively into larger communities following a certain criterion
  - E.g., based on modularity increase



AHC dendogram (modularity criteria)



Agglomerative Hierarchical Clustering

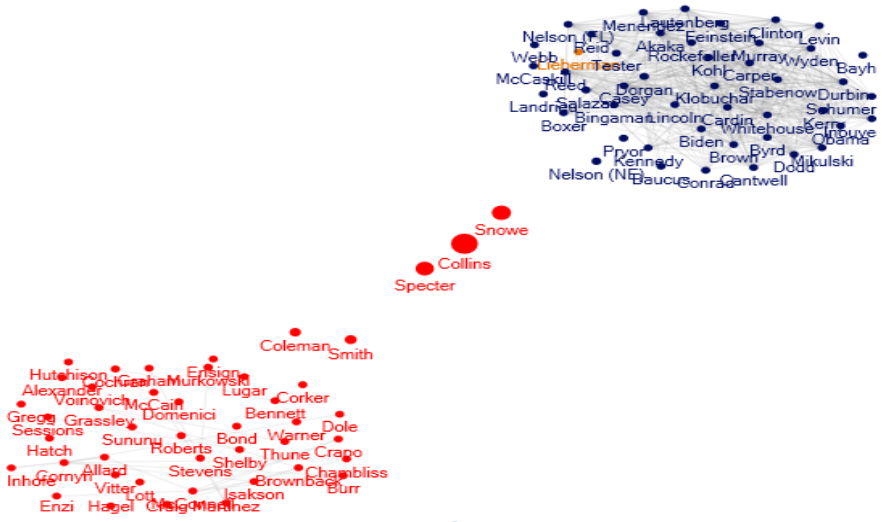
- **Louvain method** (Blondel et al., 2008) is a greedy optimization method that performs AHC
  - **Step 1:** minimizes modularity in a local way - starting with each node as one community, assesses the increase/decrease in modularity by joining other nodes
  - **Step 2:** aggregates nodes belonging to the same community and then represents each such community as a “meta-node”
  - Then repeat both steps until modularity is maximized
- This method exhibits good performance with large networks, but is order-sensitive

Applications of Community Detection

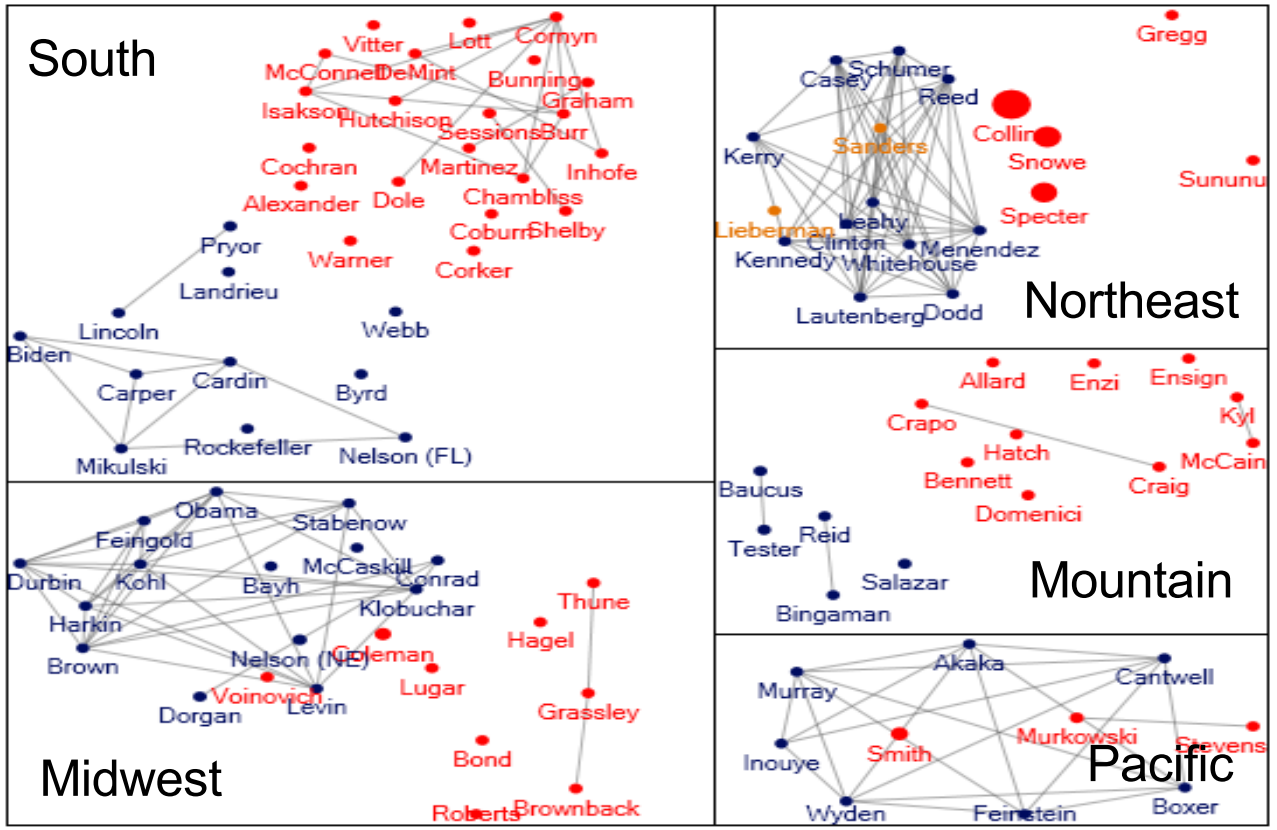
- Polarity analysis on a given topic (e.g. communities “for” and “against” some legislation, product, etc.)
- Consumer segmentation
- Information diffusion
- Recommendation systems
- Functional modules in biological networks
- Etc.

Applications of Community Detection

- Multi-faceted Analysis with Group-In-a-Box Layout

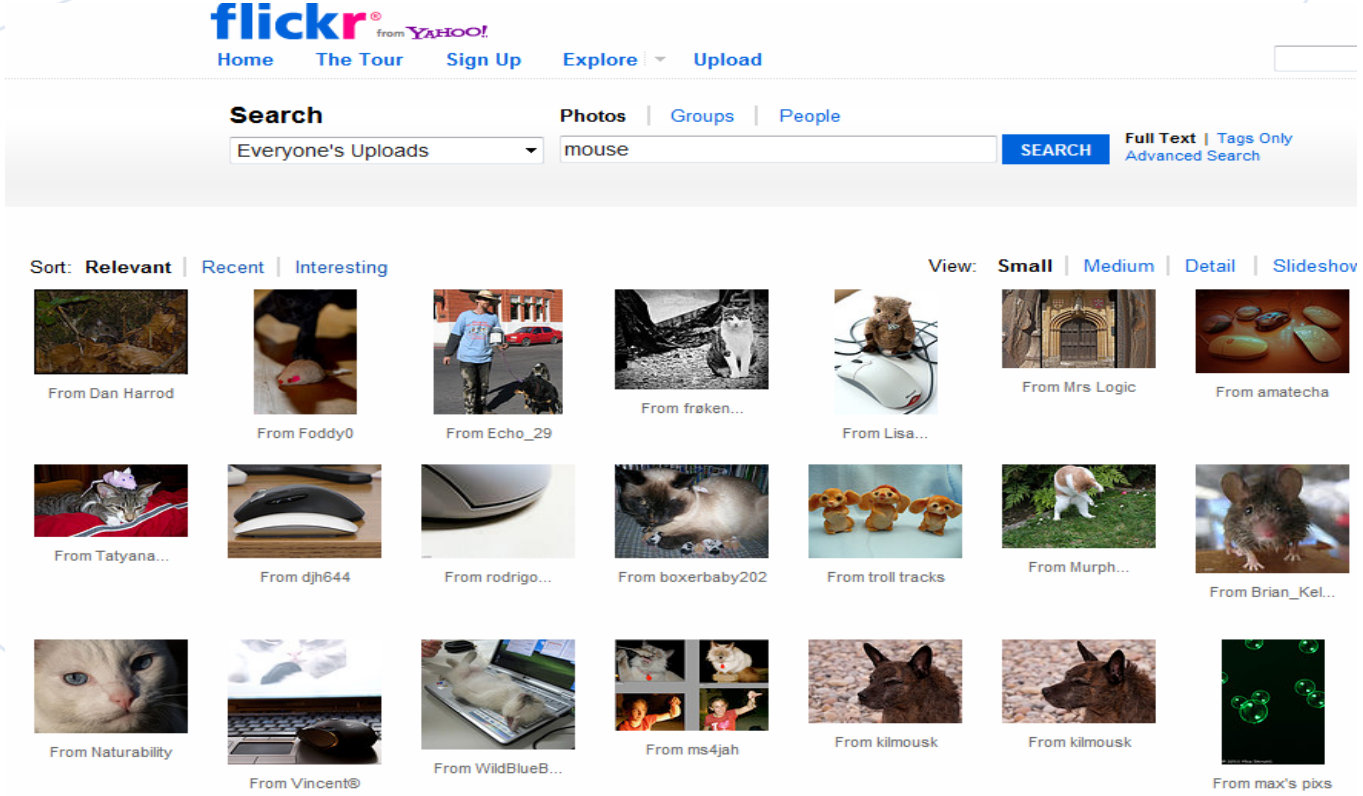


US Senators Votting Patterns



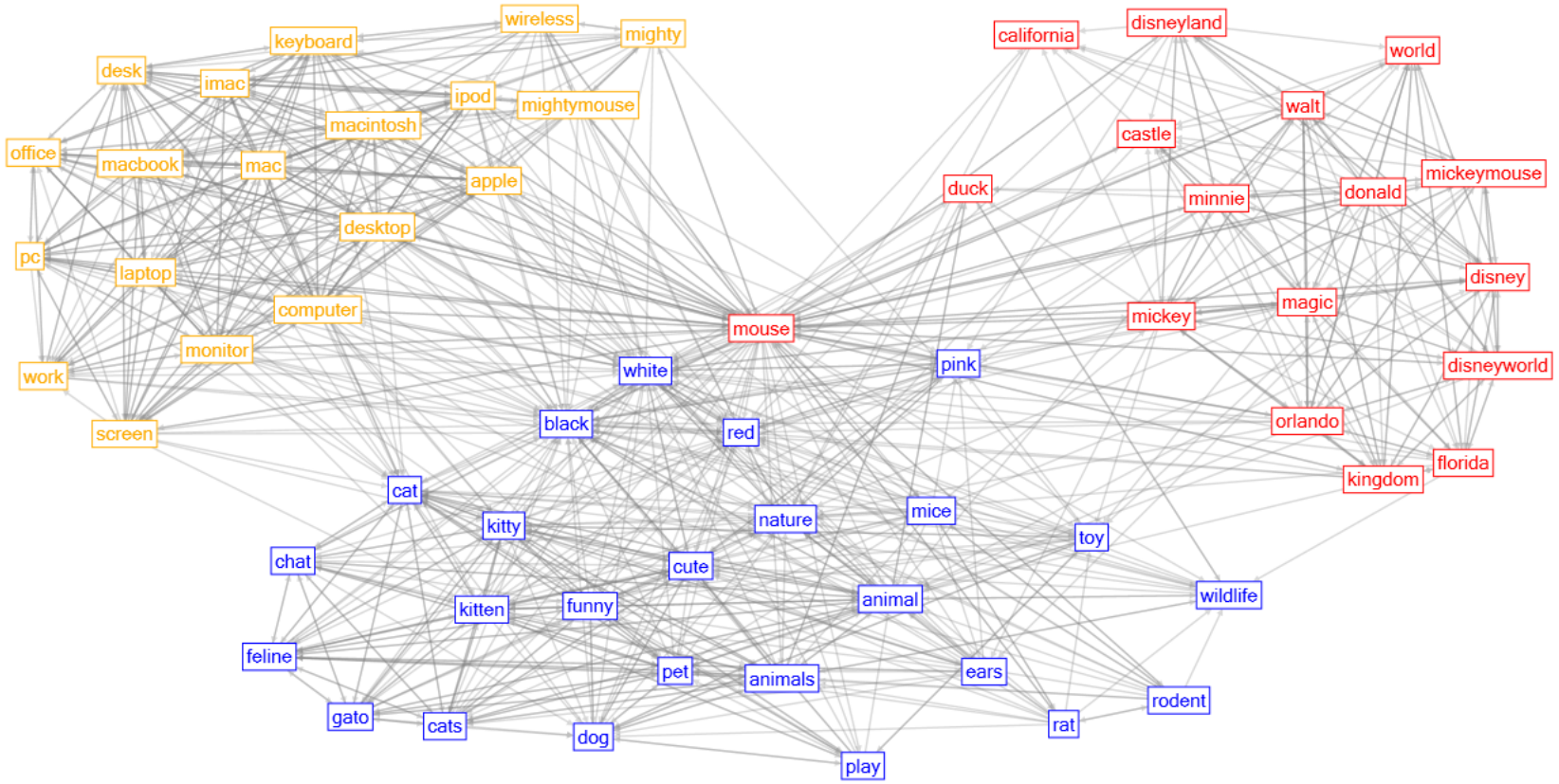
Applications of Community Detection

- Word Sense Disambiguation: Flickr photos with “Mouse” tag



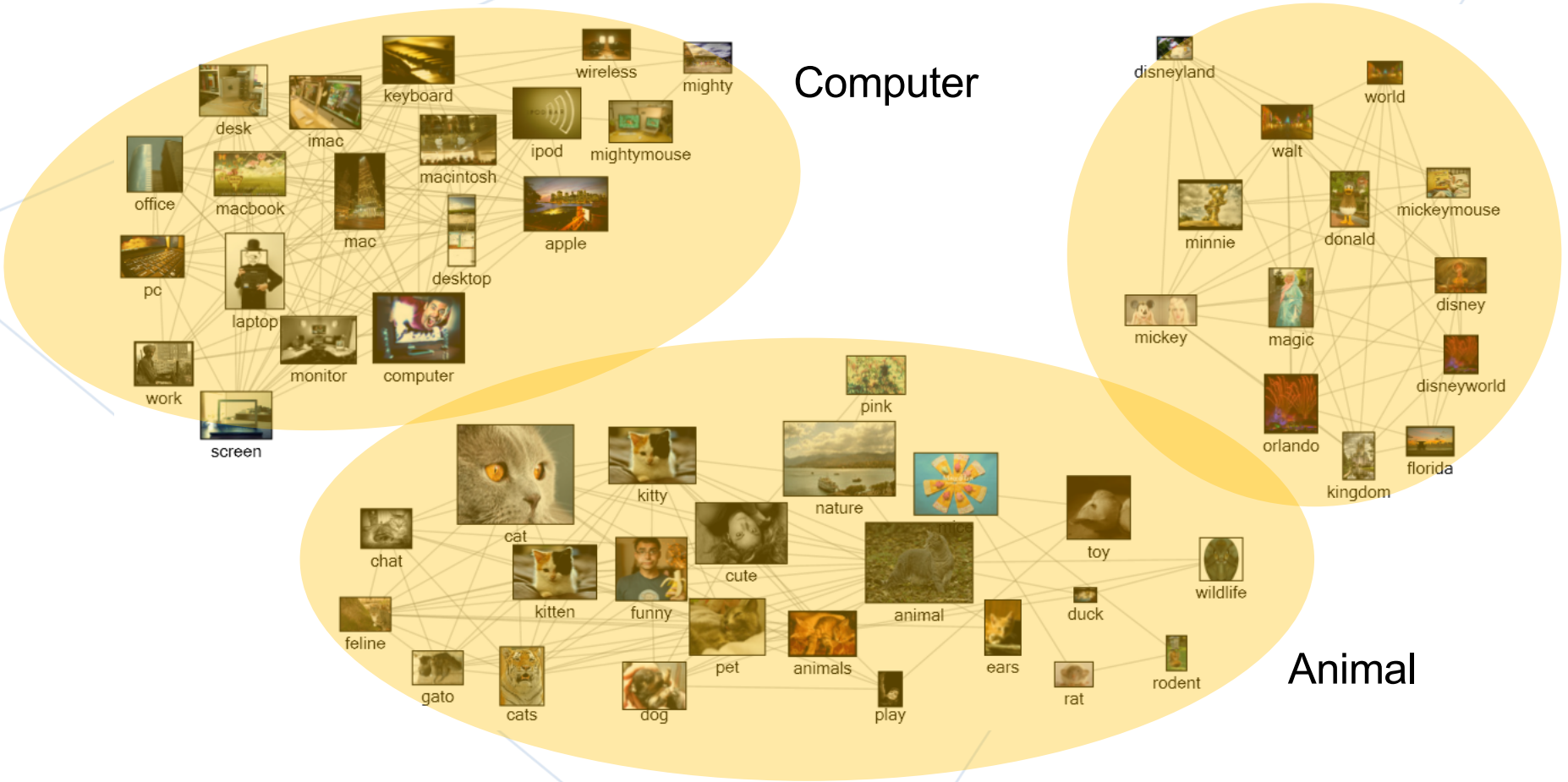
Applications of Community Detection

- Word Sense Disambiguation: Flickr photos with “Mouse” tag



Applications of Community Detection

- Word Sense Disambiguation: Flickr photos with “Mouse” tag



Try it out in R: Community Detection

```
library(igraph)

# Load karate club network

karate <- graph.famous("Zachary")
plot(karate, layout=layout.fruchterman.reingold)

# Edge betweenness (Newman and Girvan 2004)

c <- edge.betweenness.community(karate)
modularity(c)
membership(c)

# plot communities with shading
plot (c, karate)

# plot communities using just node colors
plot(karate, vertex.color=membership(c))
```

Find communities in the Zachary Karate Club network using different methods

Try it out in R: Community Detection

Porto Business School / University of Porto

```
# Modularity optimization (Clauset et al. 2005)

c <- fastgreedy.community(karate)
modularity(c)

# plot dendogram
dendPlot(c)

# Louvain method (Blondel et al. 2008)

c <- multilevel.community(karate)
modularity(c)
plot (c, karate)

# Leading eigenvector (Newman, 2006)

c <- leading.eigenvector.community(karate)
plot (c, karate)
```

Find communities in the Zachary Karate Club network using different methods

Properties of Real-world Networks

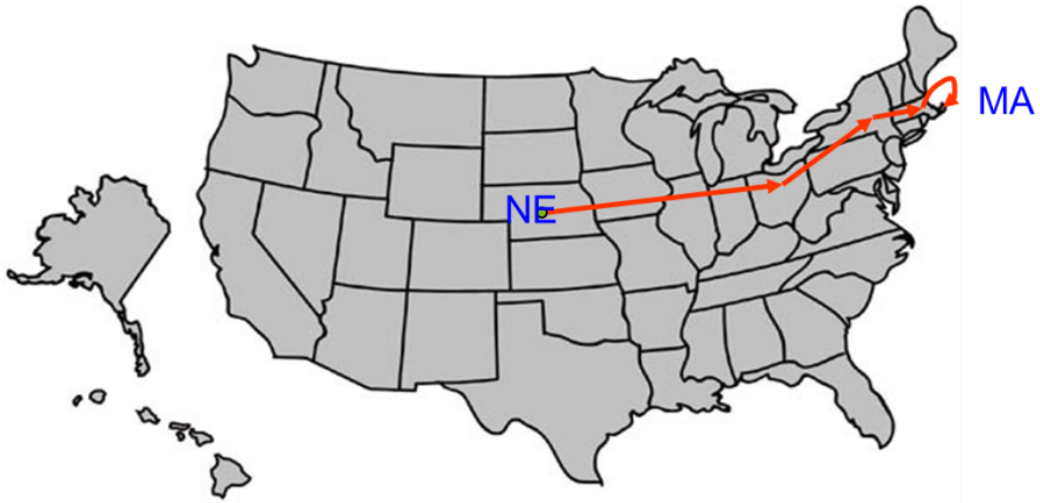
Characteristics of Real-world Networks

- Real-world networks are non-random and present characteristic properties
  - Small-world phenomenon
  - Clustering
  - Community structure
  - Power-law distributions
  - Resilience

Small World Phenomenon

- Milgram's Experiment

- **Instructions:** given a target individual (stockbroker in Boston, MA), pass the message to a person you correspond with who is "closest" to the target
- **Outcome:** 20% of initiated chains reached target with an average chain length = **6.5**



"Six degrees of separation"

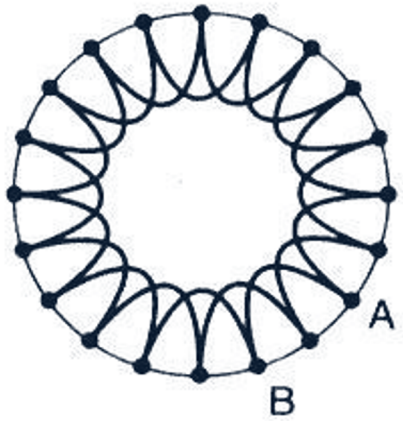
Small World Phenomenon

- Same pattern:
  - High clustering: $C_{network} \gg C_{random\ graph}$
  - Low average shortest path: $l_{network} \approx \ln(N)$
- Observed in real world networks :
  - neural network of C. elegans
  - semantic networks of languages
  - movie co-stars network
  - food webs, etc.

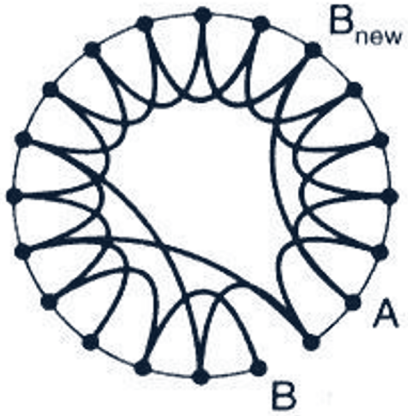
Watts-Strogatz Small World Model

- This model reconciles two observations (Watts and Strogatz, 1998):
  - **High clustering** (my friends' friends tend to be my friends)
  - **Short average paths** (low degree of separation)

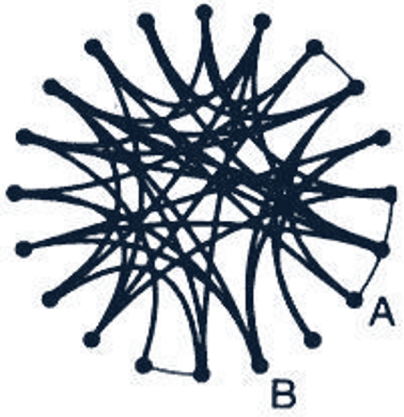
High clustering coefficient



Regular



Small world

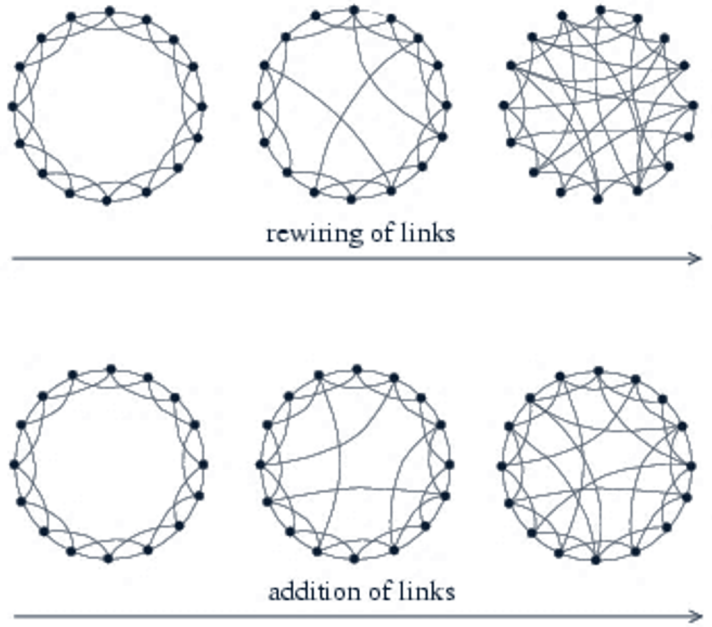


Random

Short average path length

Watts-Strogatz Small World Model

- Generative model:
  - Start with a regular network
  - Select a fraction *p* of edges and rewire one of their endpoints
  - Add a fraction *p* of additional edges leaving underlying lattice intact
- Disallow self-edges and disallow multiple edges



Watts-Strogatz Small World Model

- Each node has K>=4 nearest neighbours (local)

- Tunable: vary the probability ***p*** of rewiring any given edge
  - **small *p***: regular lattice
  - **large *p***: classical random graph

Try it out in R: Watts-Strogatz Model

Porto Business School / University of Porto

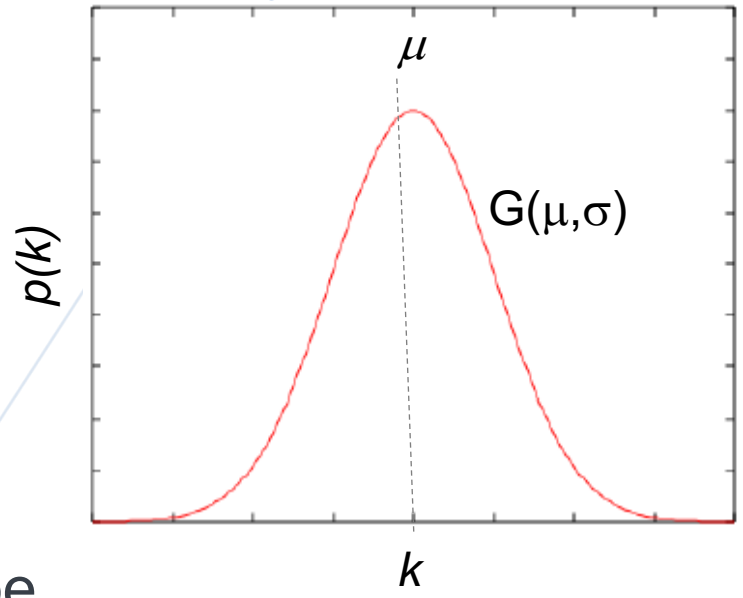
```
library(igraph)

# Generate small world

dim  <- 1    # dimension of the starting lattice
size <- 10   # size of the lattice along each dimension
nei  <- 3    # neighborhood within which the vertices of the lattice will be connected
p    <- 0.05 # rewiring probability

g <- watts.strogatz.game(dim, size, nei, p)
average.path.length(g)
transitivity(g,  type="average")

plot(g, layout=layout.circle)
```

Small World vs. Real-world Social Networks

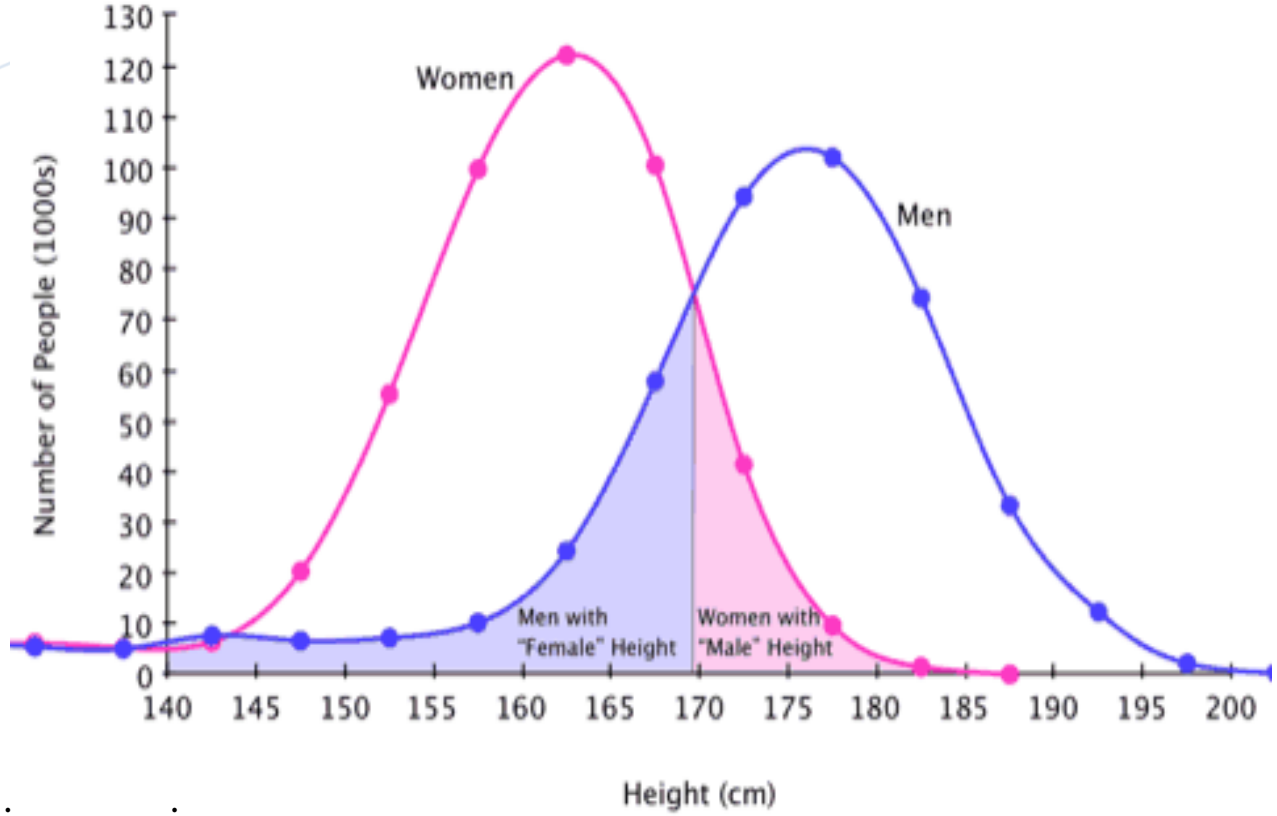
- What features of real social networks are missing from the small world model?
  - Long range links not as likely as short range ones
  - Hierarchical structure / communities
  - Hubs

Network Distributions

- Social networks are characterized by extreme imbalances – few people are extremely connected while others only connect to their inner social circles
- How is connectivity distributed?
- Assuming each node connects to other nodes **at random**
  - the number of in-links of a given node is the sum of many random quantities → in-links are expected to be **normally distributed**



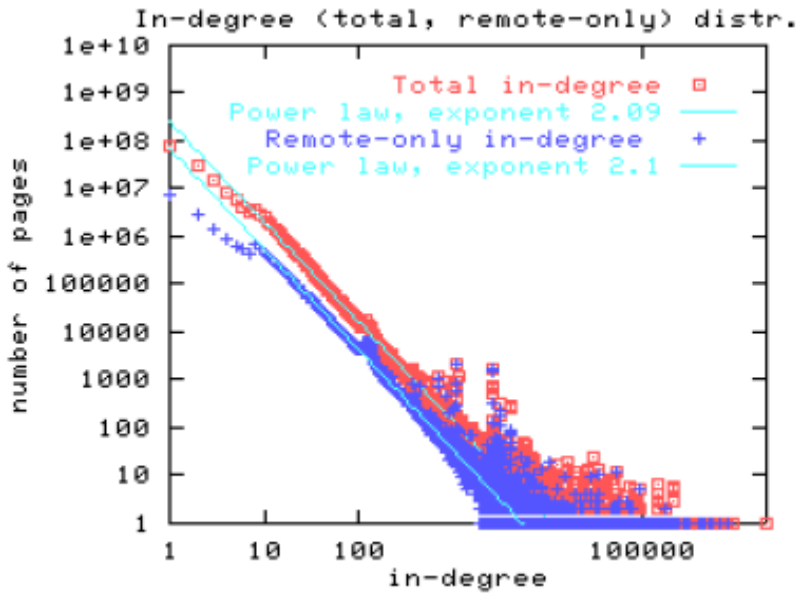
Normal Distribution of Human Heights



Random sample of 1 million american
Source: http://sugarandslugs.wordpress.com/2011/02/13/sex-differences/

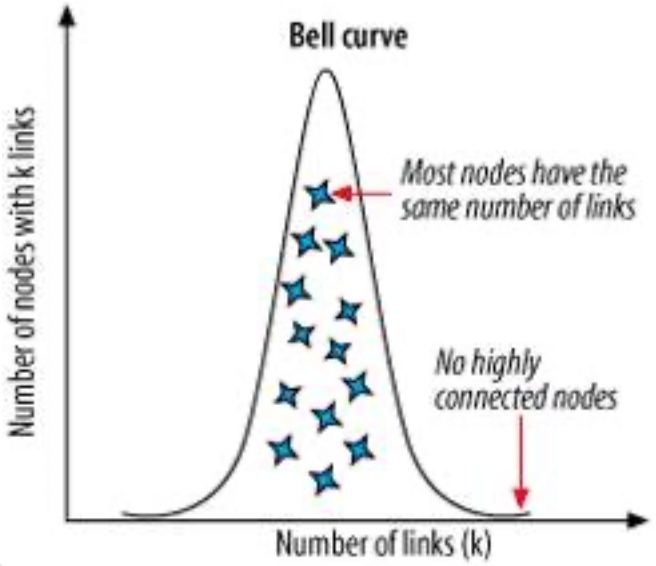
Power Laws

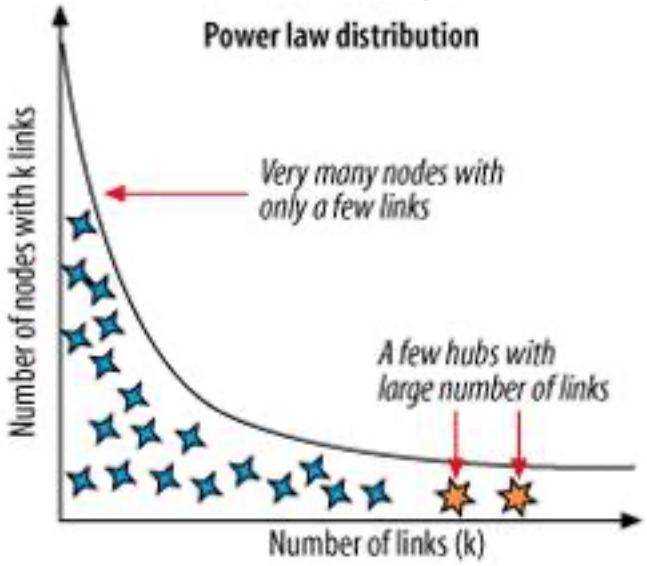
- But, in real networks degree distributions (or popularity measures) follow a power-law $p(k) \propto k^{-\alpha}$



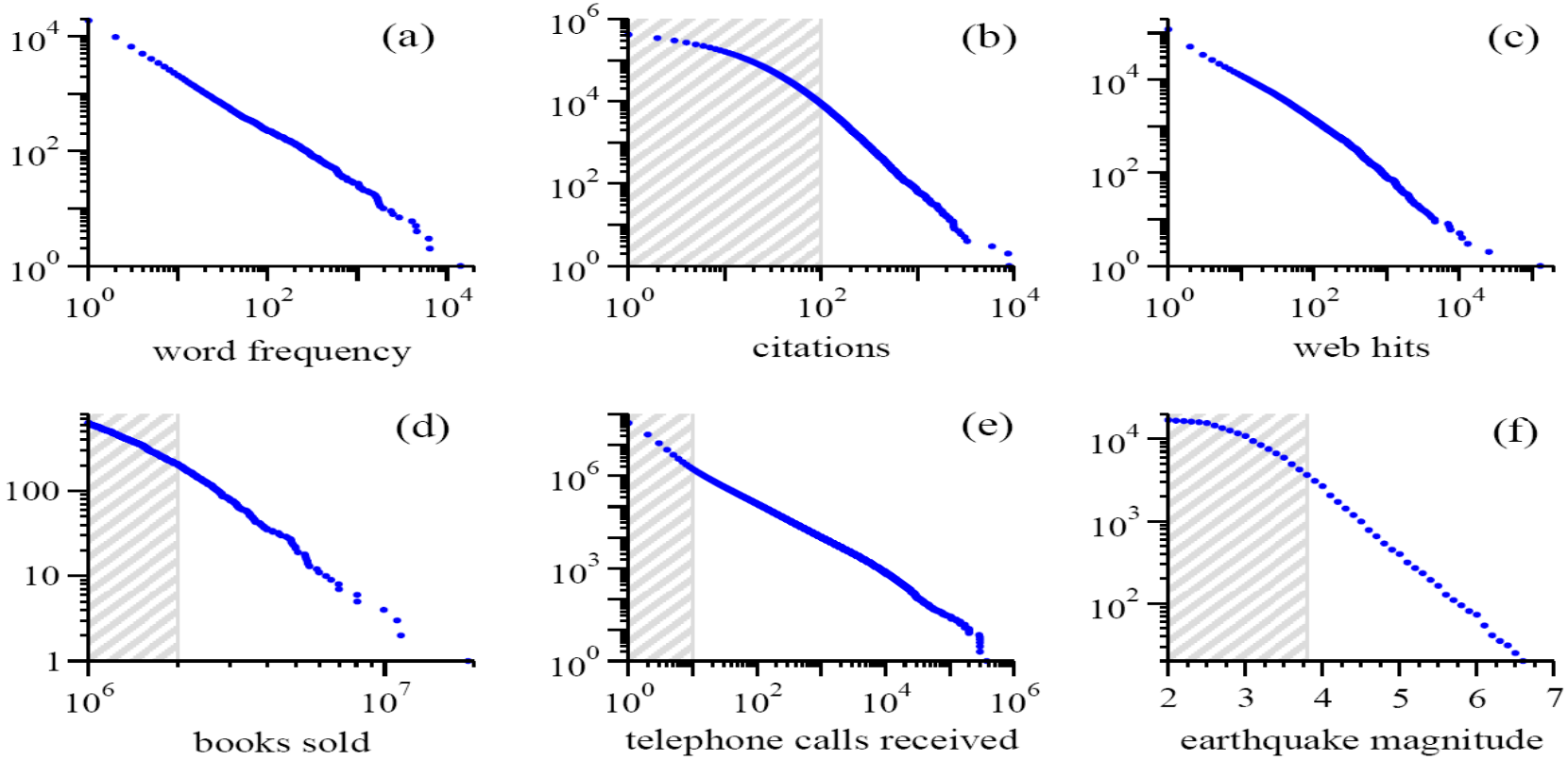
- Take a network and plot a histogram of $p(k)$ vs. k
- Plot the same data on **log-log** axis

Normal vs. Power Law distribution





Power Laws are Everywhere: Why?



Source: MEJ Newman (2005)

Coming up next

- Network dynamics!
  - How do networks evolve?
  - Do the general laws of evolution explain the power laws?
  - And more…

- Lecturer: Dra. Márcia Oliveira

References

Publications

- Ahuja, R. K., Magnanti, T. L., and Orlin, J. B. (1993). Networks Flows: Theory, Algorithms, and Applications. Prentice Hall, New Jersey, USA.
- Girvan, M. and Newman, M. E. J. (2002). Community structure in social and biological networks. Proceedings of the National Academy of Sciences of the United States of America, 99(12):78217826.
- Blondel, V., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E. (2008). Fast unfolding of communities in large networks. Journal of Statistical Mechanics: Theory and Experiment, 2008(10):P10008.
- Newman, M. E. J. and Girvan, M. (2004). Finding and evaluating community structure in networks. Physical Review E, 69(2):026113.
- Clauset, A., Newman, M. E. J., and Moore, C. (2004). Finding community structure in very large networks. Physical Review E, 70(6):066111.

References

Porto Business School / University of Porto

Publications

- Erdös, P. and Rényi, A. (1961). On the evolution of random graphs. Bull. Inst. Internat. Statist., 38(4):343347.
- Barabási, A.-L. and Albert, R. (1999). Emergence of scaling in random networks.Science,286(5439):50951.
- MEJ Newman (2005), ’Power laws, Pareto distributions and Zipf’s law’, Contemporary Physics 46, 323–351
- Mendes Rodrigues, E., Milic-Frayling, N., Smith, M., Shneiderman, B., Hensen, D., Group-In-a-Box Layout for Multi-faceted Analysis of Communities, IEEE SocialCom 2011, Oct. 2011.

Make change happen