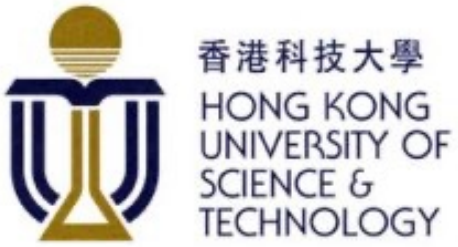
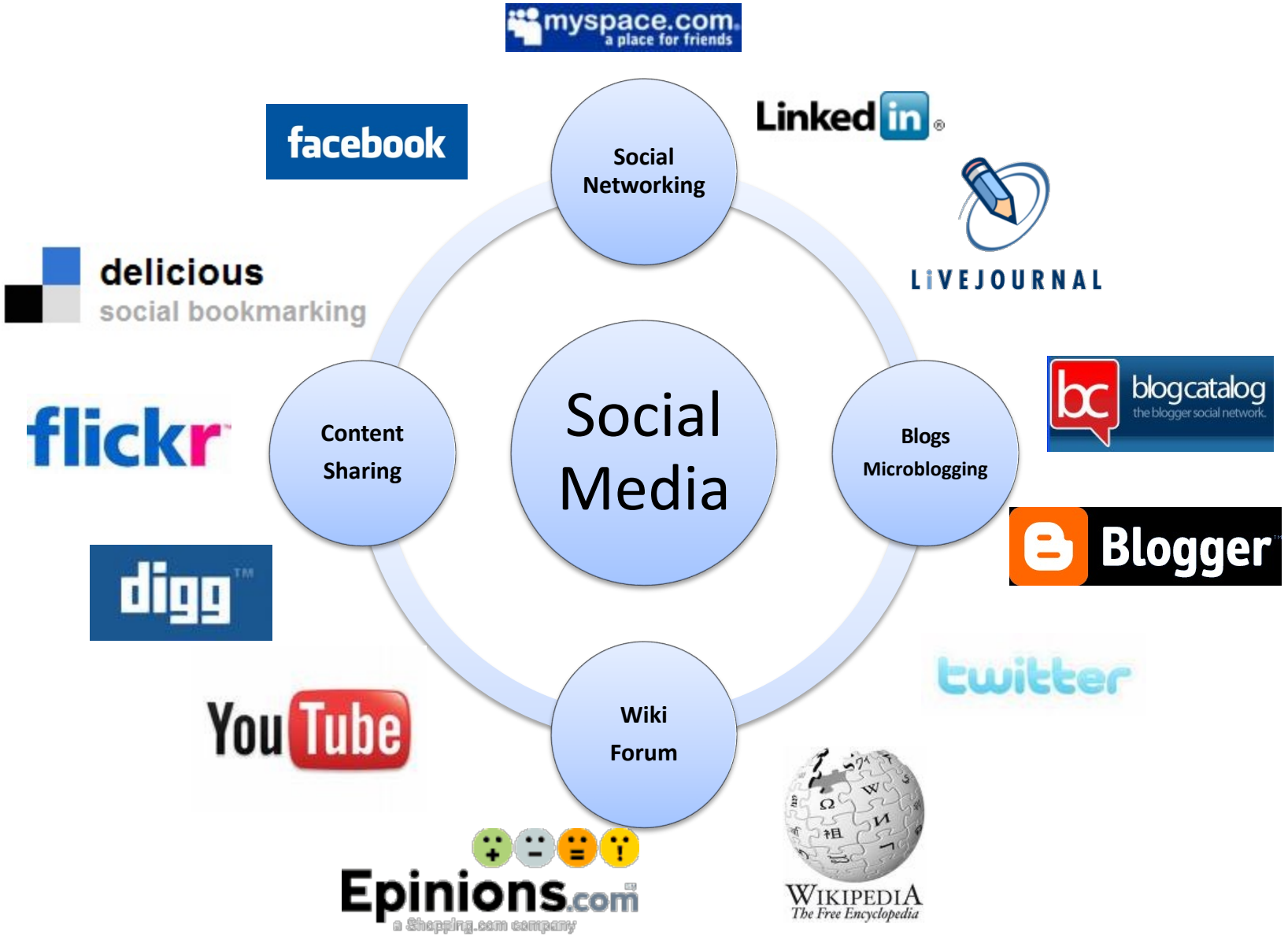
Social Networks and Social Media



Social Media: Many-to-Many



Characteristics of Social Media

- "Consumers" become "Producers"
- Rich User Interaction
- User-Generated Contents
- Collaborative environment
- Collective Wisdom
- Long Tail

Broadcast Media
Filter, then Publish

Social Media
Publish, then Filter

Top 20 Websites at USA

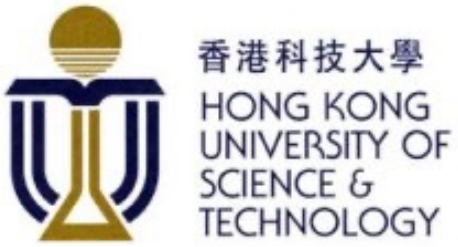
1	Google.com	11	Blogger.com
2	Facebook.com	12	msn.com
3	Yahoo.com	13	Myspace.com
4	YouTube.com	14	Go.com
5	Amazon.com	15	Bing.com
6	Wikipedia.org	16	AOL.com
7	Craigslist.org	17	LinkedIn.com
8	Twitter.com	18	CNN.com
9	Ebay.com	19	Espn.go.com
10	Live.com	20	Wordpress.com

40% of websites are social media sites

What is Social Network and Social Media

- Social Network
  - The networks formed by individuals
- Social Media
  - social network + media
    - media = content of twitter, tag, videos, photos ....

Statistical Properties of Social Networks



Why do statistics

- To understand the networks
  - Understand their topology and measure their properties
  - Study their evolution and dynamics
  - Create realistic models
  - Create algorithms that make use of the network structure

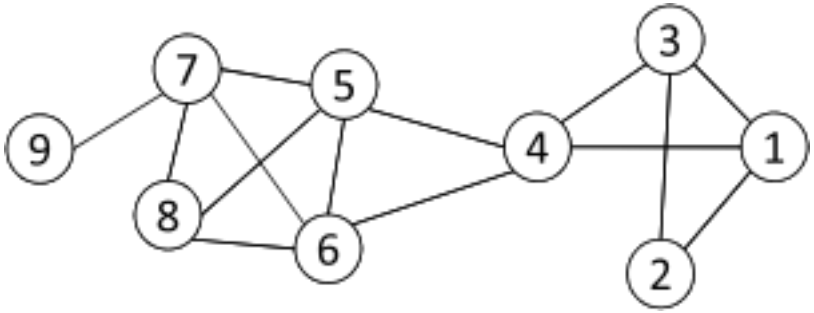
Interesting Questions: demonstration

- Some interesting questions
  - What do social networks look like, on a large scale?
  - How do networks behave over time?
  - How do the non-giant weakly connected components behave over time?
  - What distributions and patterns do weighted graphs maintain?

Networks and Representation

Social Network: A social structure made of nodes (individuals or organizations) and edges that connect nodes in various relationships like friendship, kinship etc.

- Graph Representation



- Matrix Representation

Node	1	2	3	4	5	6	7	8	9
1	-	1	1	1	0	0	0	0	0
2	1	-	1	0	0	0	0	0	0
3	1	1	-	1	0	0	0	0	0
4	1	0	1	-	1	1	0	0	0
5	0	0	0	1	-	1	1	1	0
6	0	0	0	1	1	-	1	1	0
7	0	0	0	0	1	1	-	1	1
8	0	0	0	0	1	1	1	-	0
9	0	0	0	0	0	0	1	0	-

Basic Concepts

- A: the adjacency matrix
- V: the set of nodes
- E: the set of edges
- v_i: a node v_i
- $e(v_i, v_j)$: an edge between node v_i and v_j
- N_i: the neighborhood of node v_i
- d_i: the degree of node v_i
- geodesic: a shortest path between two nodes
  - geodesic distance

Statistical Properties

- Static analysis
  - Static snapshots of graphs
- Dynamic analysis
  - A series of snapshots of graphs

Some famous properties

1. ‘Small-world’ phenomenon
   – An Experiment by Milgram (1967)
   - Asked randomly chosen “starters” to forward a letter to the target
   - Name, address, and some personal information were provided for the target person
   - The participants could only forward a letter to a single person that he/she knew on a first name basis
   - Goal: To advance the letter to the target as quickly as possible

The Milgram Experiment (Wikipedia)

Detailed procedure

1. Milgram typically chose individuals in the U.S. cities of Omaha, Nebraska and Wichita, Kansas to be the starting points and Boston, Massachusetts to be the end point of a chain of correspondence
   - because they were thought to represent a great distance in the United States, both socially and geographically.
2. Information packets were initially sent to "randomly" selected individuals in Omaha or Wichita. They included letters, which detailed the study's purpose, and basic information about a target contact person in Boston.
   - It additionally contained a roster on which they could write their own name, as well as business reply cards that were pre-addressed to Harvard.

The Milgram Experiment (cont.)

3. Upon receiving the invitation to participate, the recipient was asked whether he or she personally knew the contact person described in the letter.
   - If so, the person was to forward the letter directly to that person. For the purposes of this study, knowing someone "personally" was defined as knowing them on a first-name basis.
4. In the more likely case that the person did not personally know the target, then the person was to think of a friend or relative they know personally that is more likely to know the target.
   - A postcard was also mailed to the researchers at Harvard so that they could track the chain's progression toward the target.

The Milgram Experiment

5. When and if the package eventually reached the contact person in Boston, the researchers could examine the roster to count the number of times it had been forwarded from person to person.
   - Additionally, for packages that never reached the destination, the incoming postcards helped identify the break point in the chain.

Result of the Experiment

- However, a significant problem was that often people refused to pass the letter forward, and thus the chain never reached its destination.
- In one case, 232 of the 296 letters never reached the destination.[3]
- However, 64 of the letters eventually did reach the target contact.
- Among these chains, the average path length fell around 5.5 or six.

Some famous properties con't

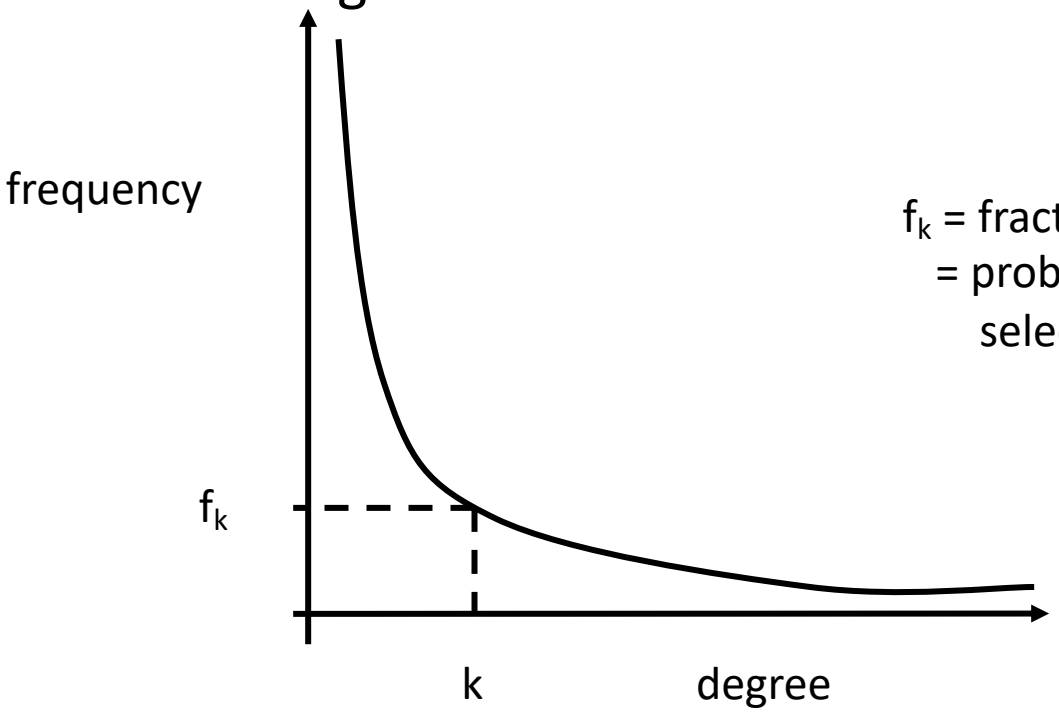
1. 'Small-world' phenomenon
   - Property
     - Any two people can be connected within 6 hops

       six degrees of separation

   - Verified on a planetary-scale IM network of 180 million users (Leskovec and Horvitz 2008)
     - The average path length is 6.6

Some famous properties con't

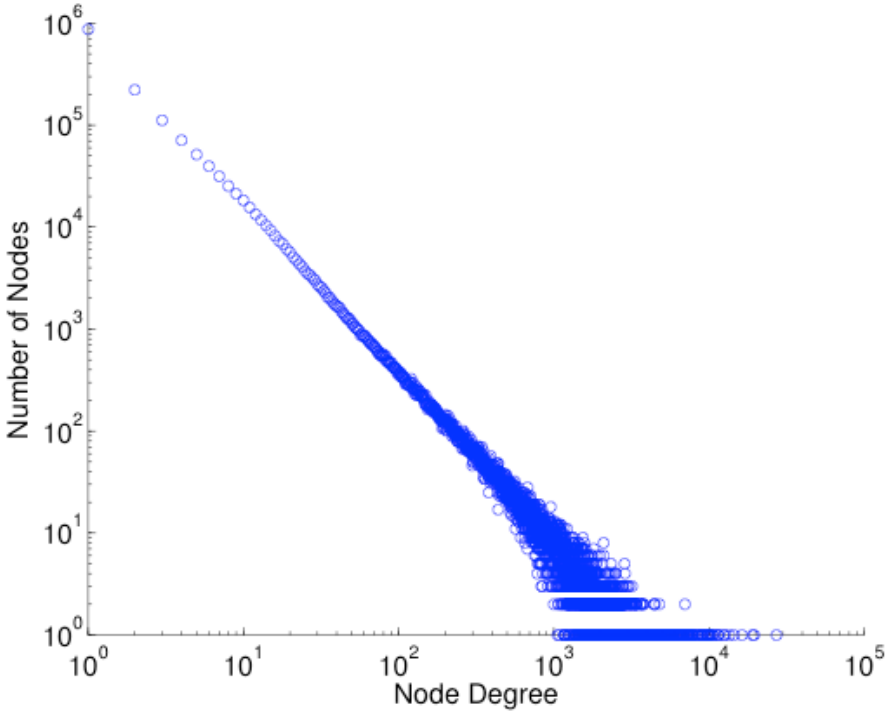
2. 'Power-law' degree distributions
   - Degree distributions



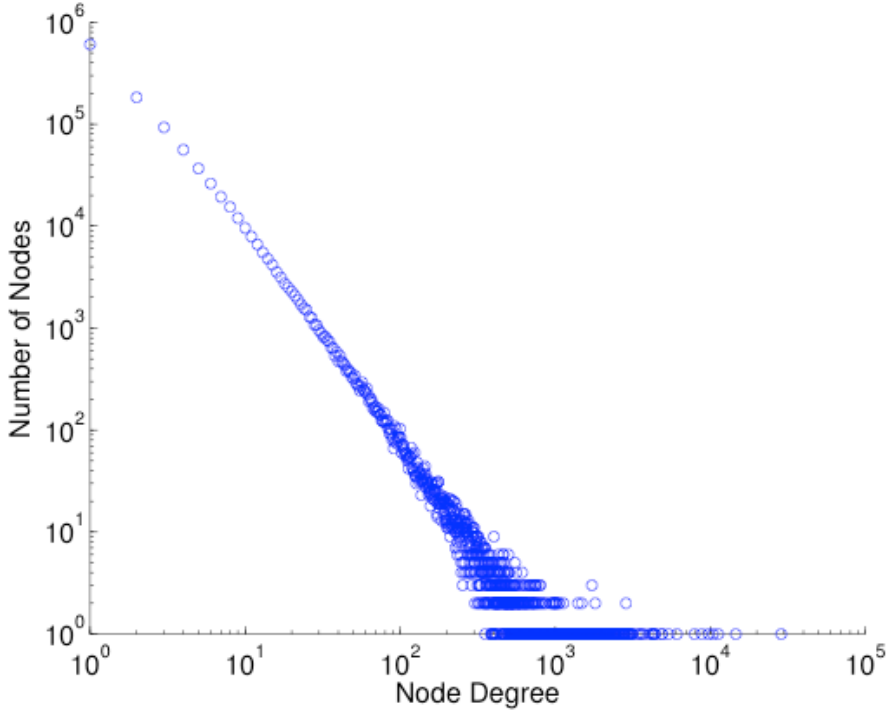
f_k = fraction of nodes with degree k
= probability of a randomly selected node to have degree k

log-log plot

- Power law distribution becomes a straight line if plot in a log-log scale



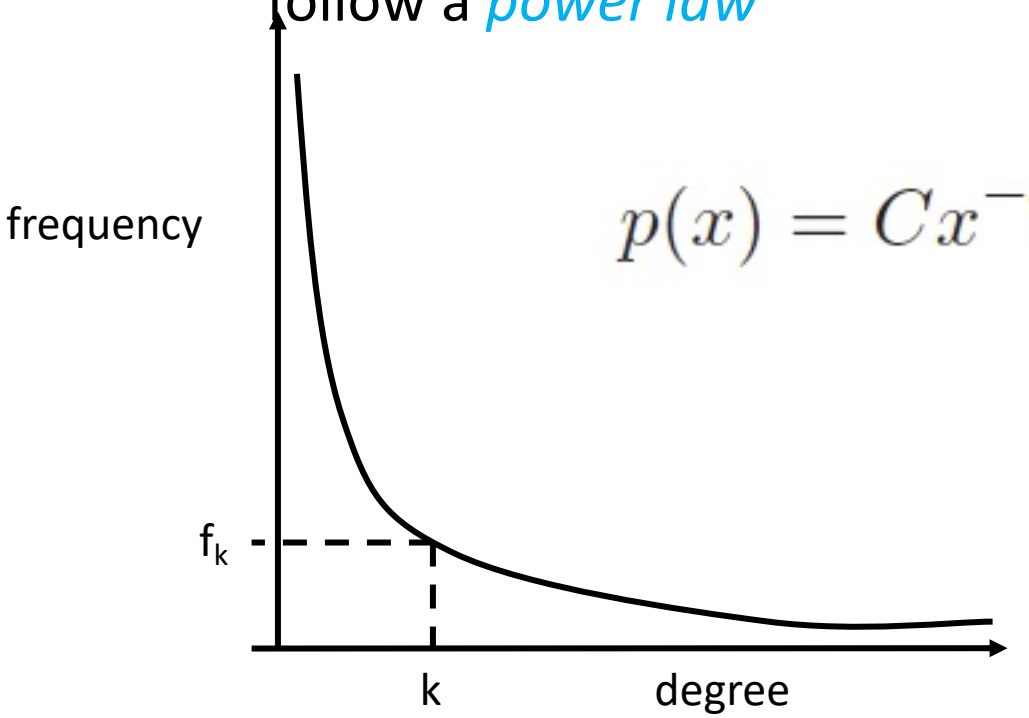
Friendship Network in Flickr



Friendship Network in YouTube

Some famous properties con't

2. 'Power-law' degree distributions
   - The degree distributions of most real-life networks follow a *power law*



$$p(x) = Cx^{-\alpha}, \quad x \geq x_{min}, \quad \alpha > 1$$

Some other properties

3. Triangle Power Law

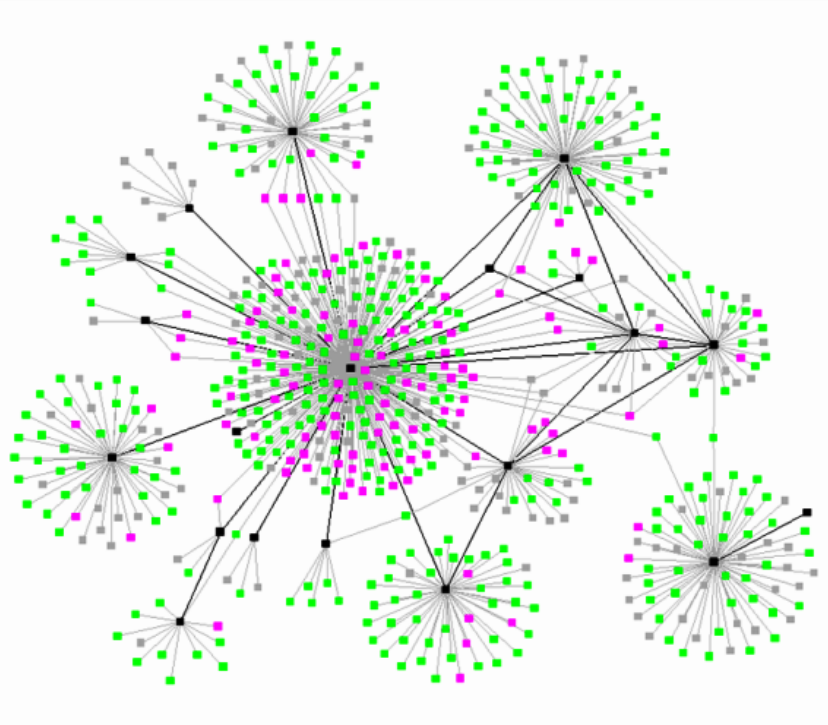
$$f(\Delta) \propto \Delta^{\alpha} \qquad \alpha < 0$$

Some other properties

4. Eigenvalue Power Law
   - The 20 or so largest eigenvalues of the adjacency matrix are power law distributed
   - This is consequence of the “Degree Power Law”

Some other properties

5. Community Structure
   - Social networks are modular
     - i.e nodes form communities

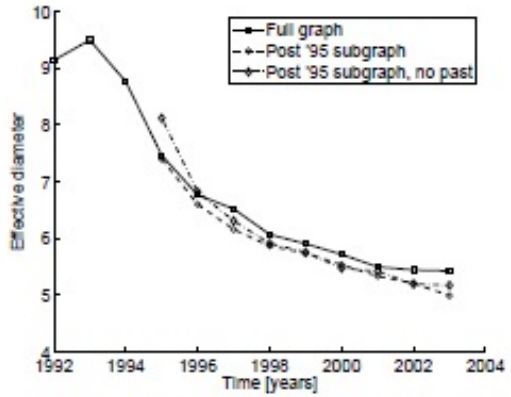


Statistical Properties

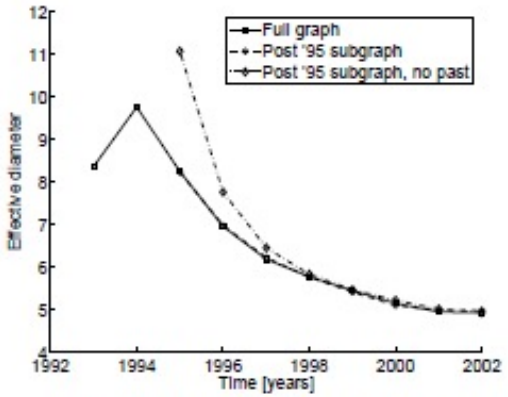
- Static analysis
  - Static snapshots of graphs
- Dynamic analysis
  - A series of snapshots of graphs

Properties

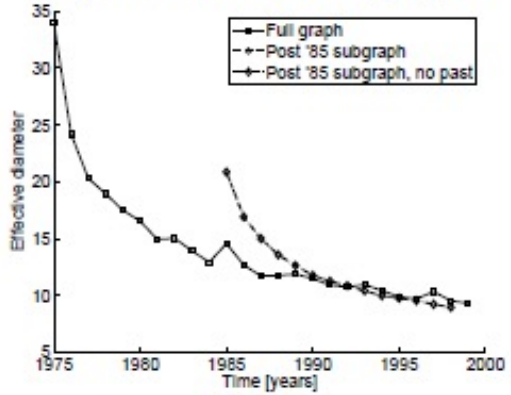
1. Shrinking Diameter



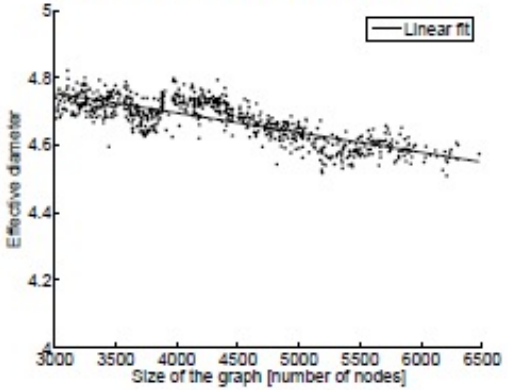
(a) arXiv citation graph



(b) Affiliation network



(c) Patents

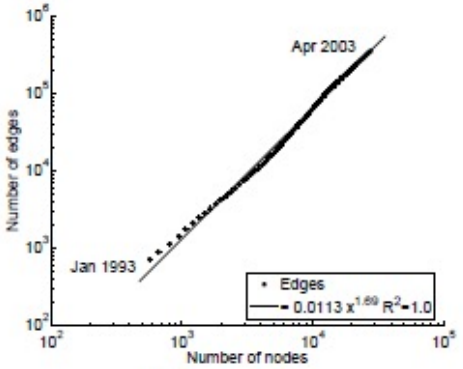


(d) AS

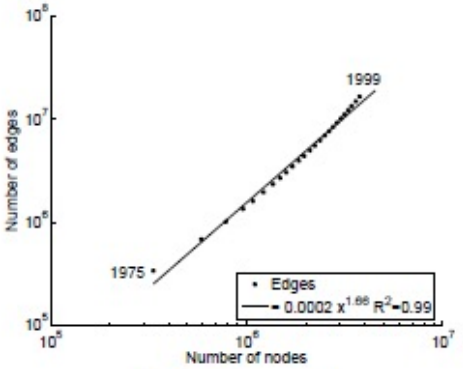
Properties con't

2. Densification Power Law (DPL)
   - E(t): the number of edges
   - N(t): the number of nodes

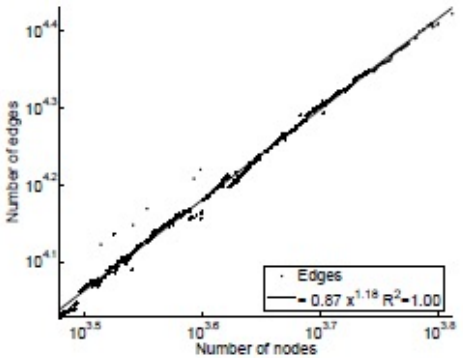
$$E(t) \propto N(t)^{\beta}$$

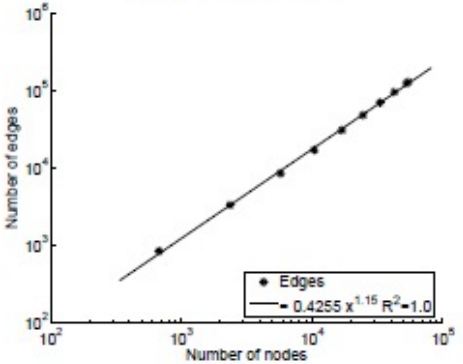


(a) arXiv



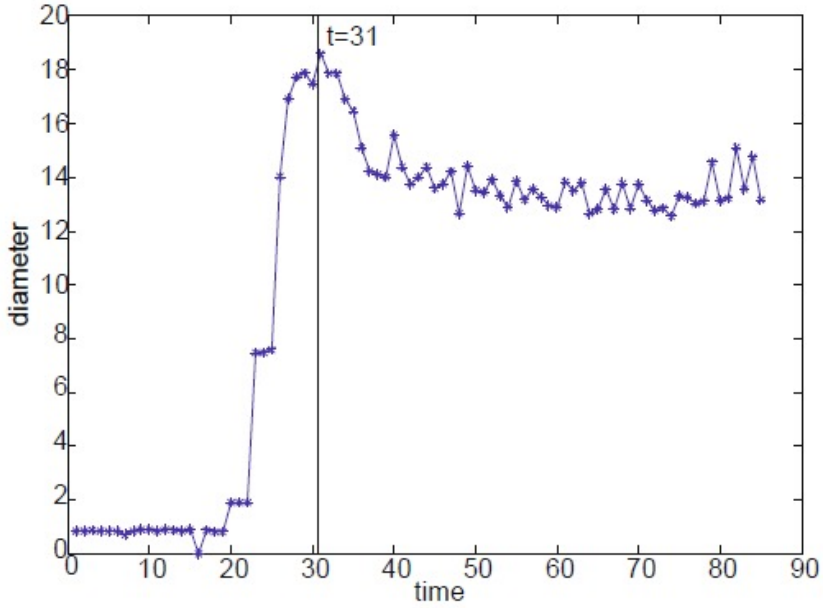
(b) Patents





Properties con't

3. Diameter-plot and Gelling point
   - Graph forming
     - Establishment period
     - Stable period



(a) Diameter(t)

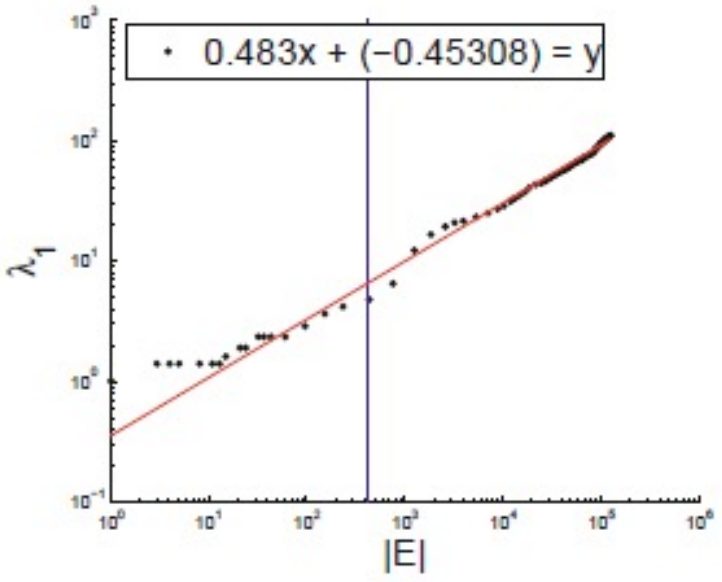
Properties con’t

4. Constant/Oscillating Connected Components (CC)
   - Largest: constant
   - Second/Third: oscillation

Properties con't

5. Principal eigenvalue over time
   - E(t): the number of edges
   - $\lambda_1(t)$ the largest eigenvalue

$$\lambda_1(t) \propto E(t)^{\alpha} \qquad \alpha \leq 0.5$$



(b) Blog Network

Conclusion

- Usefulness of the statistical properties
  - Understanding human behaviors
  - Anomalous graphs/subgraphs detection
  - Identifying authorities and search algorithms
  - Prepare resources based on the prediction
  - …

Statistical Properties of Social Networks

香港科技大學
HONG KONG UNIVERSITY OF SCIENCE & TECHNOLOGY

Why do statistics

- To understand the networks
  - Understand their topology and measure their properties
  - Study their evolution and dynamics
  - Create realistic models
  - Create algorithms that make use of the network structure

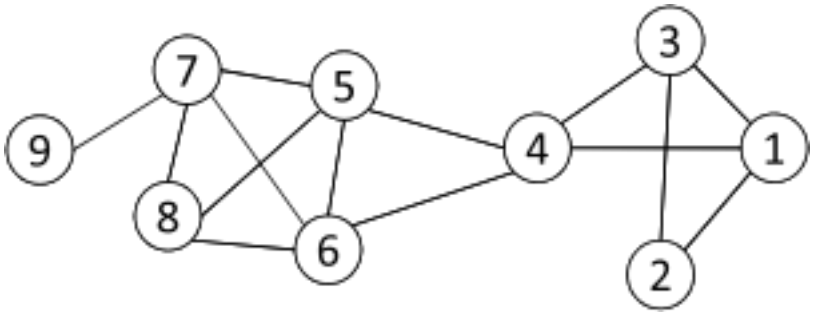
Interesting Questions: demonstration

- Some interesting questions
  - What do social networks look like, on a large scale?
  - How do networks behave over time?
  - How do the non-giant weakly connected components behave over time?
  - What distributions and patterns do weighted graphs maintain?

Networks and Representation

Social Network: A social structure made of nodes (individuals or organizations) and edges that connect nodes in various relationships like friendship, kinship etc.

- Graph Representation



- Matrix Representation

Node	1	2	3	4	5	6	7	8	9
1	-	1	1	1	0	0	0	0	0
2	1	-	1	0	0	0	0	0	0
3	1	1	-	1	0	0	0	0	0
4	1	0	1	-	1	1	0	0	0
5	0	0	0	1	-	1	1	1	0
6	0	0	0	1	1	-	1	1	0
7	0	0	0	0	1	1	-	1	1
8	0	0	0	0	1	1	1	-	0
9	0	0	0	0	0	0	1	0	-

Basic Concepts

- A: the adjacency matrix
- V: the set of nodes
- E: the set of edges
- v_i: a node v_i
- $e(v_i, v_j)$: an edge between node v_i and v_j
- N_i: the neighborhood of node v_i
- d_i: the degree of node v_i
- geodesic: a shortest path between two nodes
  - geodesic distance

Statistical Properties

- Static analysis
  - Static snapshots of graphs
- Dynamic analysis
  - A series of snapshots of graphs

Some famous properties

1. ‘Small-world’ phenomenon
   - An Experiment by Milgram (1967)
     - Asked randomly chosen “starters” to forward a letter to the target
     - Name, address, and some personal information were provided for the target person
     - The participants could only forward a letter to a single person that he/she knew on a first name basis
     - Goal: To advance the letter to the target as quickly as possible

The Milgram Experiment (Wikipedia)

Detailed procedure

1. Milgram typically chose individuals in the U.S. cities of Omaha, Nebraska and Wichita, Kansas to be the starting points and Boston, Massachusetts to be the end point of a chain of correspondence
   - because they were thought to represent a great distance in the United States, both socially and geographically.
2. Information packets were initially sent to "randomly" selected individuals in Omaha or Wichita. They included letters, which detailed the study's purpose, and basic information about a target contact person in Boston.
   - It additionally contained a roster on which they could write their own name, as well as business reply cards that were pre-addressed to Harvard.

The Milgram Experiment (cont.)

3. Upon receiving the invitation to participate, the recipient was asked whether he or she personally knew the contact person described in the letter.
   - If so, the person was to forward the letter directly to that person. For the purposes of this study, knowing someone "personally" was defined as knowing them on a first-name basis.
4. In the more likely case that the person did not personally know the target, then the person was to think of a friend or relative they know personally that is more likely to know the target.
   - A postcard was also mailed to the researchers at Harvard so that they could track the chain's progression toward the target.

The Milgram Experiment

5. When and if the package eventually reached the contact person in Boston, the researchers could examine the roster to count the number of times it had been forwarded from person to person.
   - Additionally, for packages that never reached the destination, the incoming postcards helped identify the break point in the chain.

Result of the Experiment

- However, a significant problem was that often people refused to pass the letter forward, and thus the chain never reached its destination.
- In one case, 232 of the 296 letters never reached the destination.[3]
- However, 64 of the letters eventually did reach the target contact.
- Among these chains, the average path length fell around 5.5 or six.

Some famous properties con't

1. 'Small-world' phenomenon
   - Property
     - Any two people can be connected within 6 hops

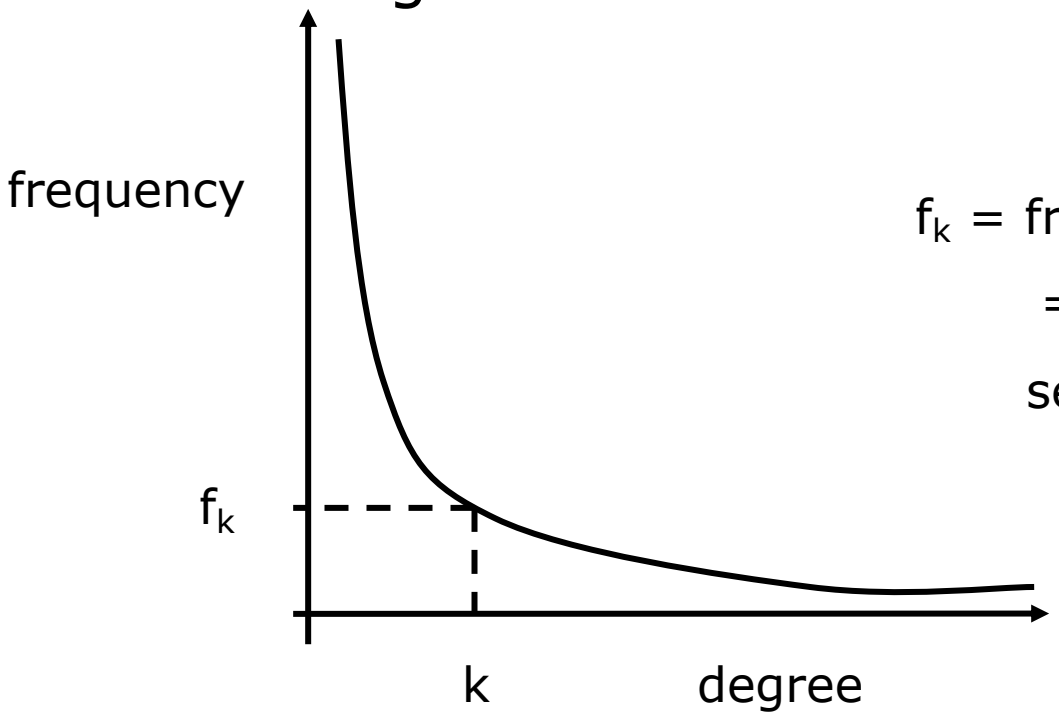
six degrees of separation

   - Verified on a planetary-scale IM network of 180 million users (Leskovec and Horvitz 2008)
     - The average path length is 6.6

Some famous properties con't

2. 'Power-law' degree distributions

- Degree distributions

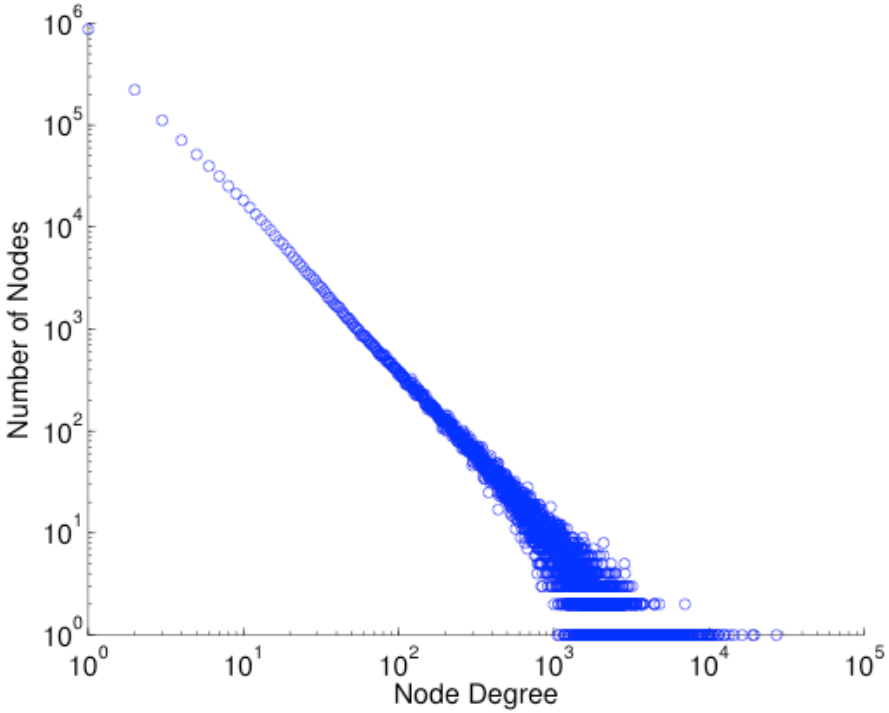


f_k = fraction of nodes with degree k

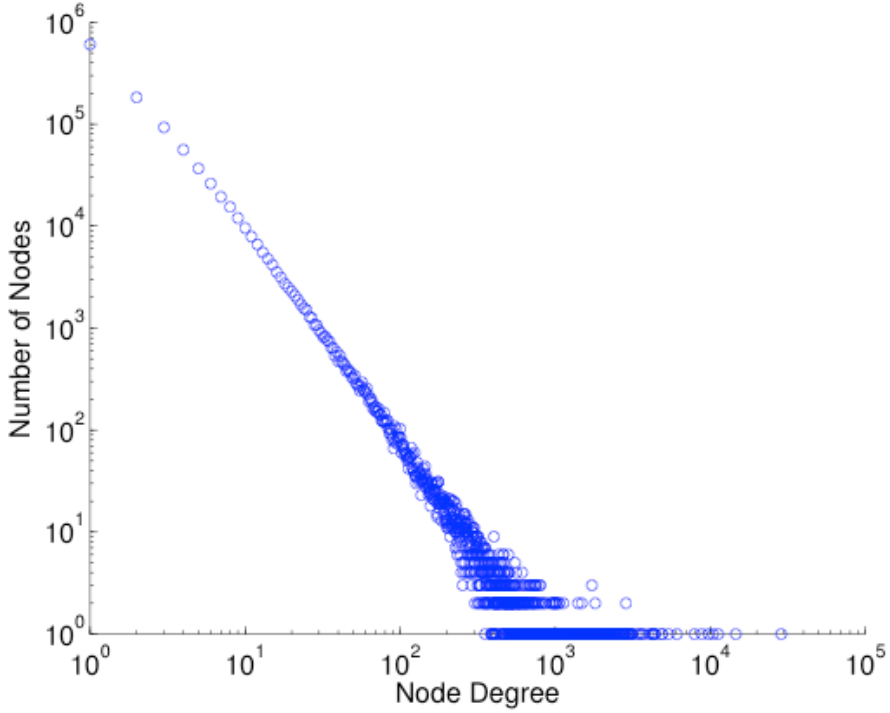
= probability of a randomly selected node to have degree k

log-log plot

- Power law distribution becomes a straight line if plot in a log-log scale



Friendship Network in Flickr

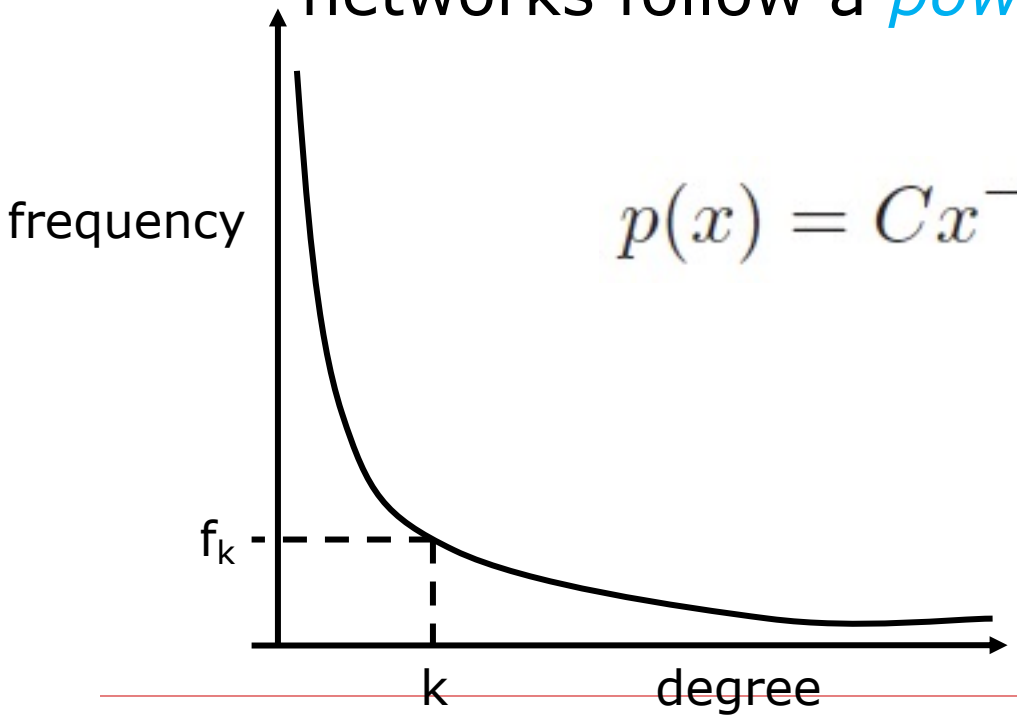


Friendship Network in YouTube

Some famous properties con't

2. 'Power-law' degree distributions

- The degree distributions of most real-life networks follow a *power law*



$$p(x) = Cx^{-\alpha}, \quad x \geq x_{min}, \quad \alpha > 1$$

Some other properties

3. Triangle Power Law

$$f(\Delta) \propto \Delta^{\alpha} \qquad \alpha < 0$$

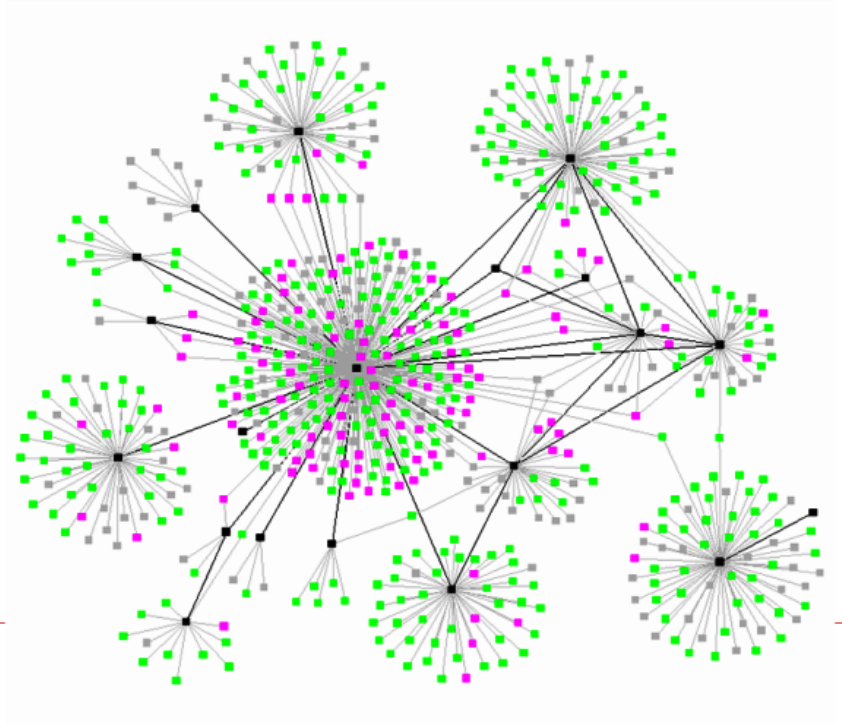
Some other properties

4. Eigenvalue Power Law
   - The 20 or so largest eigenvalues of the adjacency matrix are power law distributed
   - This is consequence of the “Degree Power Law”

Some other properties

5. Community Structure

- Social networks are modular
  - i.e nodes form communities

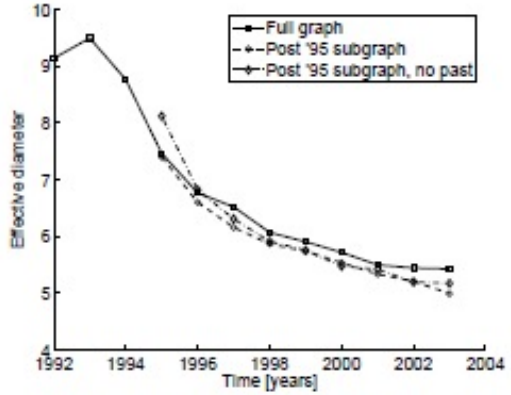


Statistical Properties

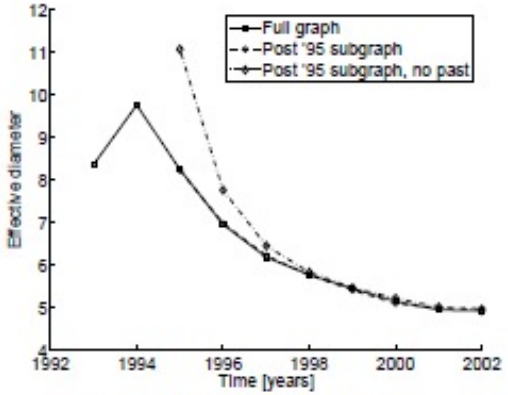
- Static analysis
  - Static snapshots of graphs
- Dynamic analysis
  - A series of snapshots of graphs

Properties

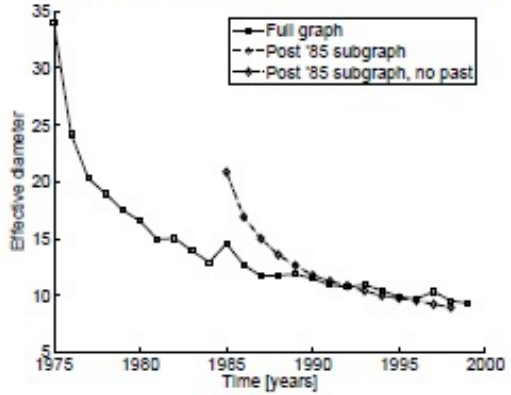
1. Shrinking Diameter



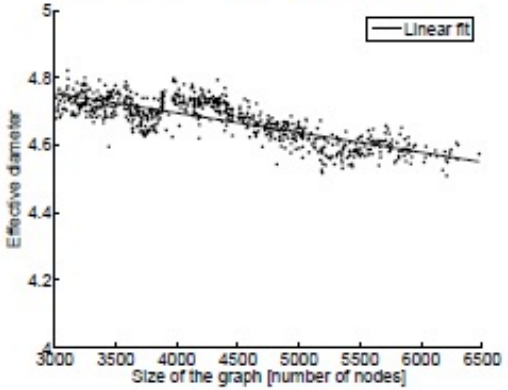
(a) arXiv citation graph



(b) Affiliation network



(c) Patents



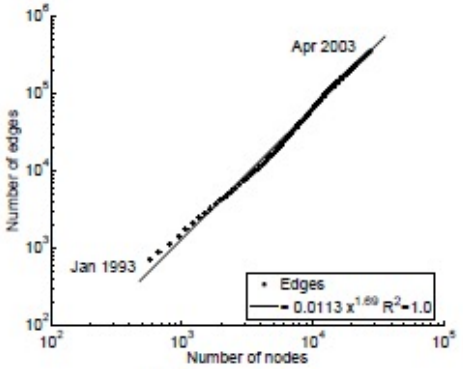
(d) AS

Properties con't

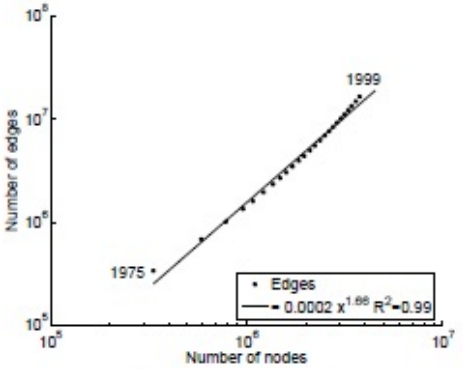
2. Densification Power Law (DPL)

- E(t): the number of edges
- N(t): the number of nodes

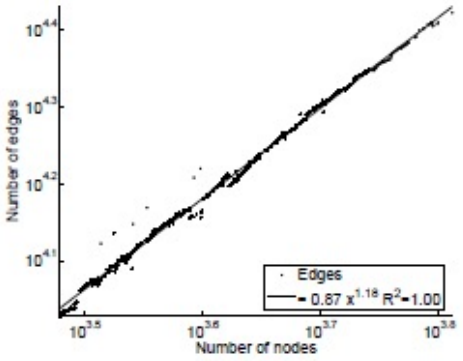
$$E(t) \propto N(t)^{\beta}$$

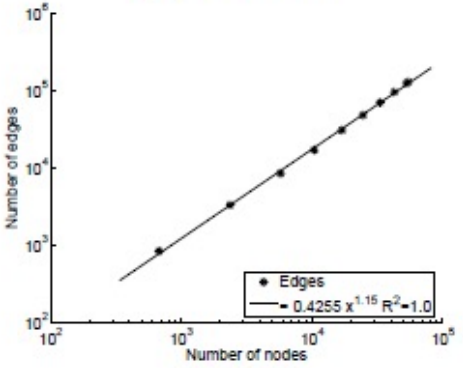


(a) arXiv



(b) Patents

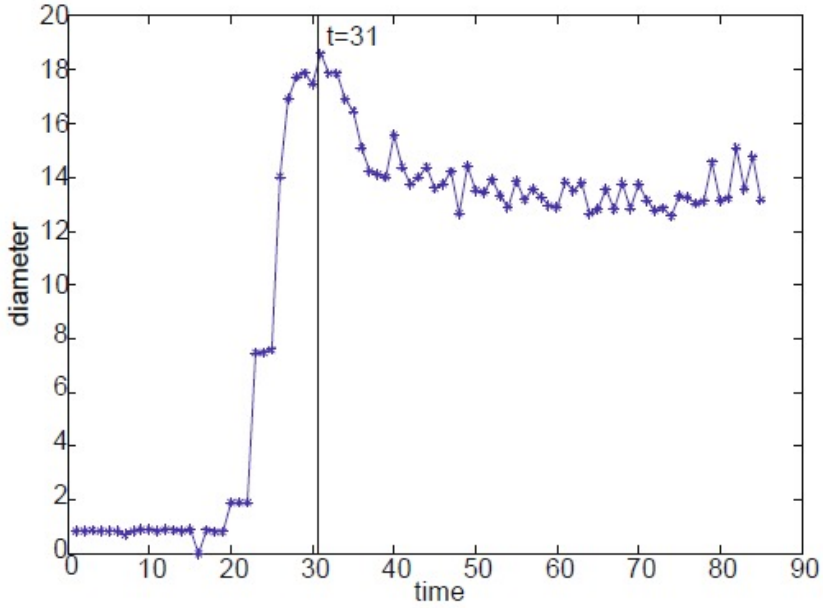




Properties con't

3. Diameter-plot and Gelling point

- Graph forming
  - Establishment period
  - Stable period



(a) Diameter(t)

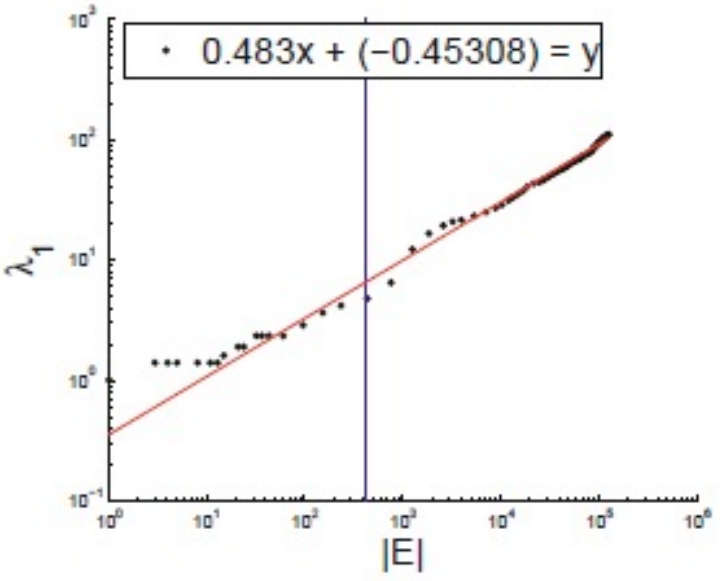
Properties con’t

4. Constant/Oscillating Connected Components (CC)
   - Largest: constant
   - Second/Third: oscillation

Properties con't

5. Principal eigenvalue over time
   - E(t): the number of edges
   - $\lambda_1(t)$: the largest eigenvalue

$$\lambda_1(t) \propto E(t)^{\alpha} \qquad \alpha \leq 0.5$$



(b) Blog Network

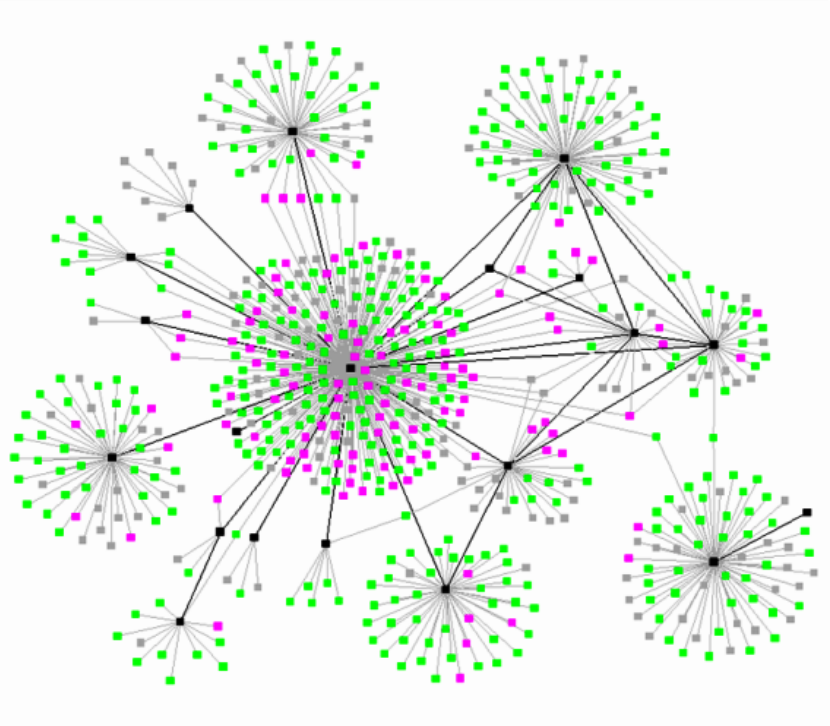
Conclusion

- Usefulness of the statistical properties
  - Understanding human behaviors
  - Anomalous graphs/subgraphs detection
  - Identifying authorities and search algorithms
  - Prepare resources based on the prediction
  - …

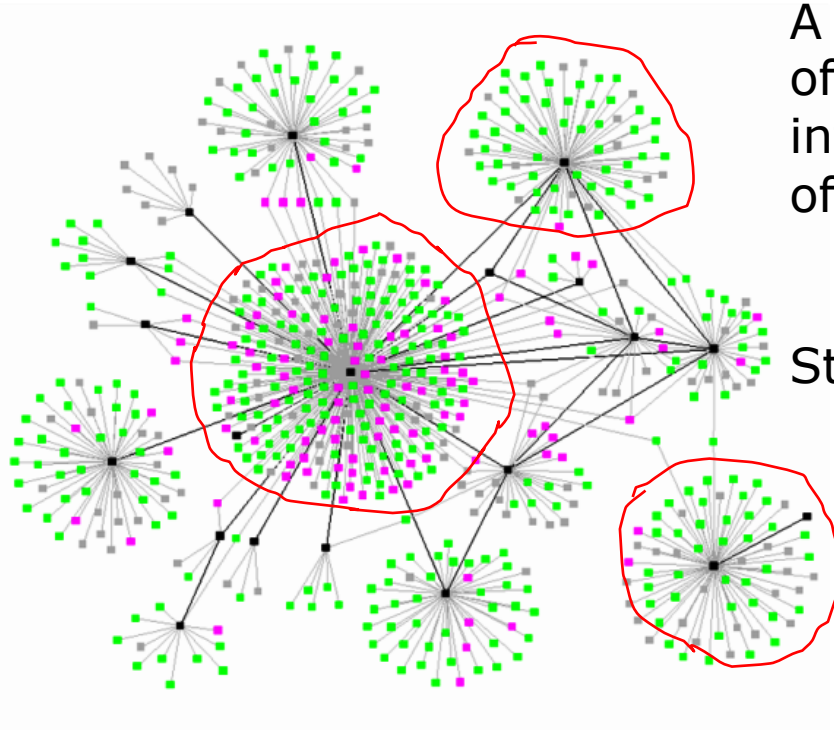
Community Discovery in Social Networks

香港科技大學
HONG KONG UNIVERSITY OF SCIENCE & TECHNOLOGY

What is community



What is community



A ***community*** is a group of nodes with greater ties internally than to the rest of the network

Strictest: clique

Why community exists?

- Why communities in social media?
  - Human beings are social
  - Easy-to-use social media allows people to extend their social life in unprecedented ways
  - Difficult to meet friends in the physical world, but much easier to find friend online with similar interests
  - Interactions between nodes can help determine communities

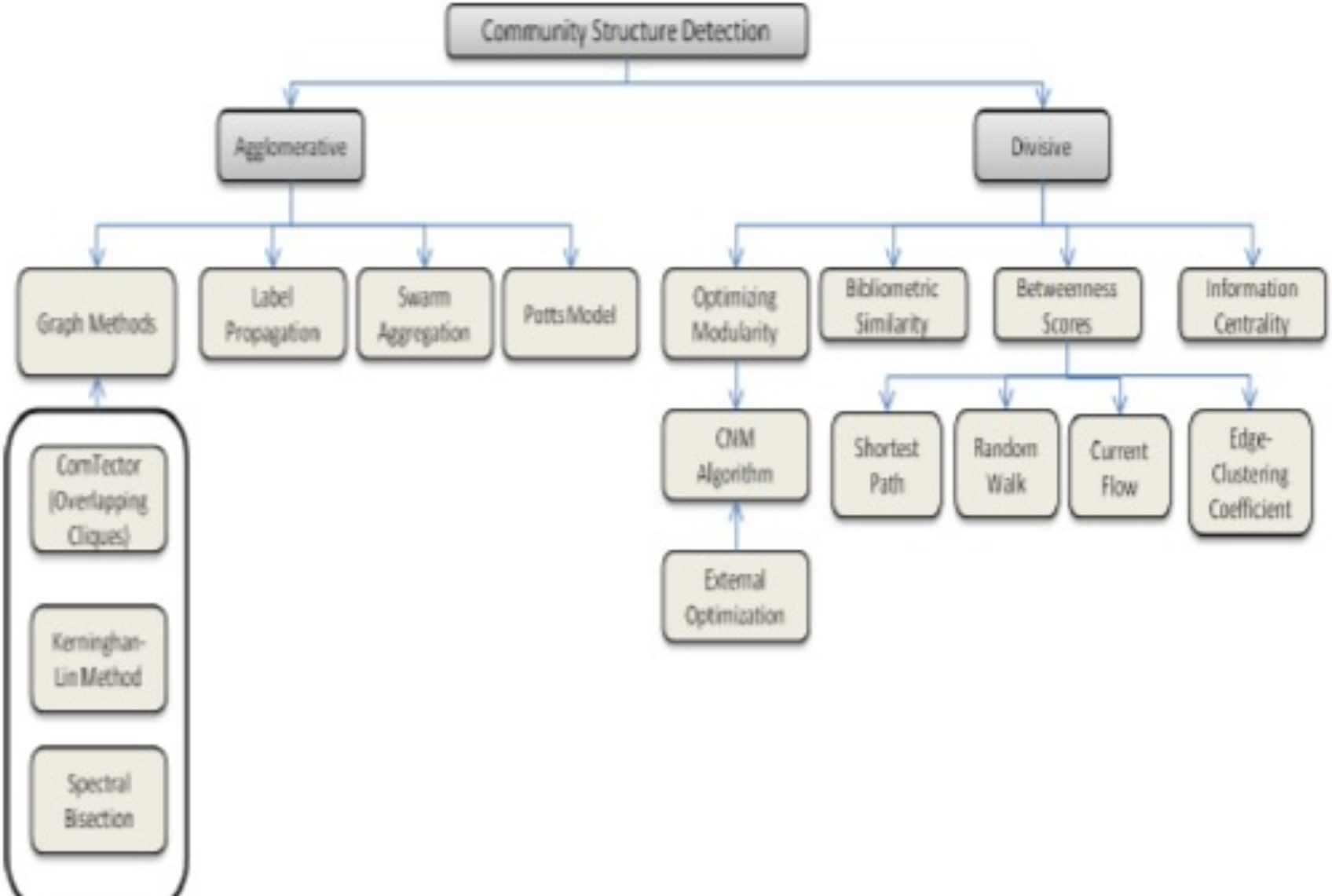
Why do community discovery

- Why?
  - Understand and underpin social structure and its evolution
  - Study the social behavior of human/animals
  - Set proxy caches
  - Detect link farms
  - Do personalized recommendation
  - Enable efficient message routing and posting in mobile ad-hoc networks
  - Enable efficient customer feedback
  - …

Discovering Algorithms

- Outline
  - Overview
  - Quality Estimation
  - Some Algorithms

Overview


Community Structure Detection
Agglomerative
Divisive
Graph Methods
Label Propagation
Swarm Aggregation
Potts Model
Optimizing Modularity
Bibliometric Similarity
Betweenness Scores
Information Centrality
CNM Algorithm
External Optimization
Shortest Path
Random Walk
Current Flow
Edge-Clustering Coefficient
ComTector (Overlapping Cliques)
Kerninghan-Lin Method
Spectral Bisection

Quality Estimation

- How to measure the discovering results?
  - Normalized Cut $\frac{\sum_{i\in S, j\in \bar{S}} A(i,j)}{\sum_{i\in S} \deg ree(i)} + \frac{\sum_{i\in S, j\in \bar{S}} A(i,j)}{\sum_{i\in \bar{S}} \deg ree(i)}$
  - Conductance $\frac{\sum_{i\in S, j\in \bar{S}} A(i,j)}{\min(\sum_{i\in S} \deg ree(i), \sum_{i\in \bar{S}} \deg ree(i))}$
  - Kernighan-Lin (KL) objective

$$\sum_{i\neq j} A(V_i, V_j) \; with \; |V_1| \equiv |V_2| \equiv \ldots \equiv |V_k|$$

Quality Estimation: Modularity

Q(division) = #(internal edges) - E(#(internal edges) in a RANDOM graph with same node degrees)

Trivial division: all vertices in one group ==> Q(trivial division) = 0

$$Q = \frac{1}{2m} \sum_{\ell=1}^{k} \sum_{i \in C_\ell, j \in C_\ell} (A_{ij} - d_i d_j / 2m)$$

m is the number of edges in the network

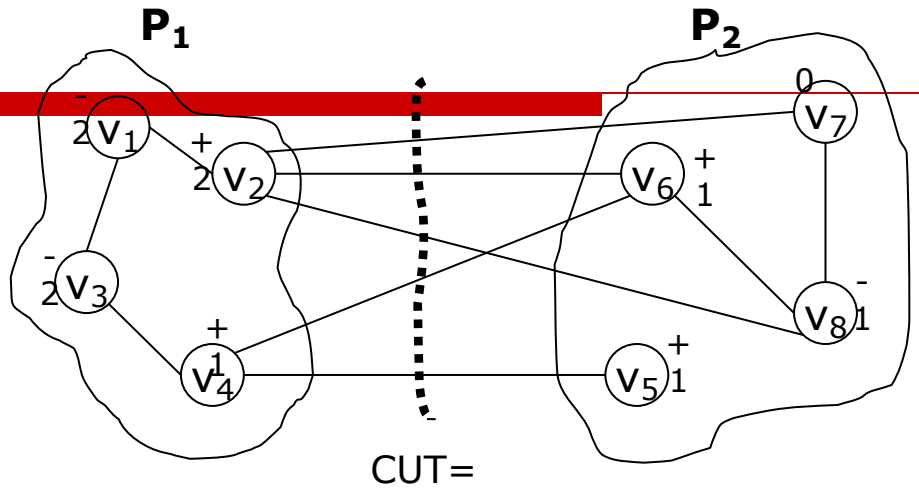
...

Optimizing any of these objectives is NP-Hard

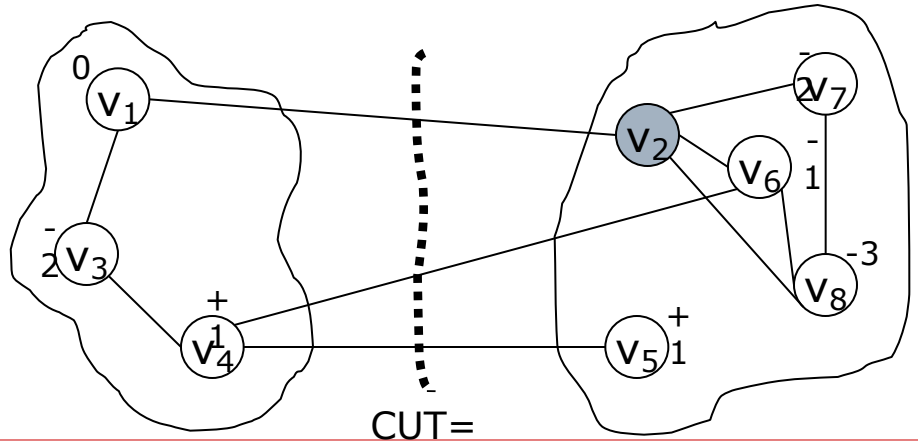
Discovering Algorithms

- Outline
  - Overview
  - Quality Estimation
  - Some Algorithms

The Kernighan-Lin (KL) algorithm

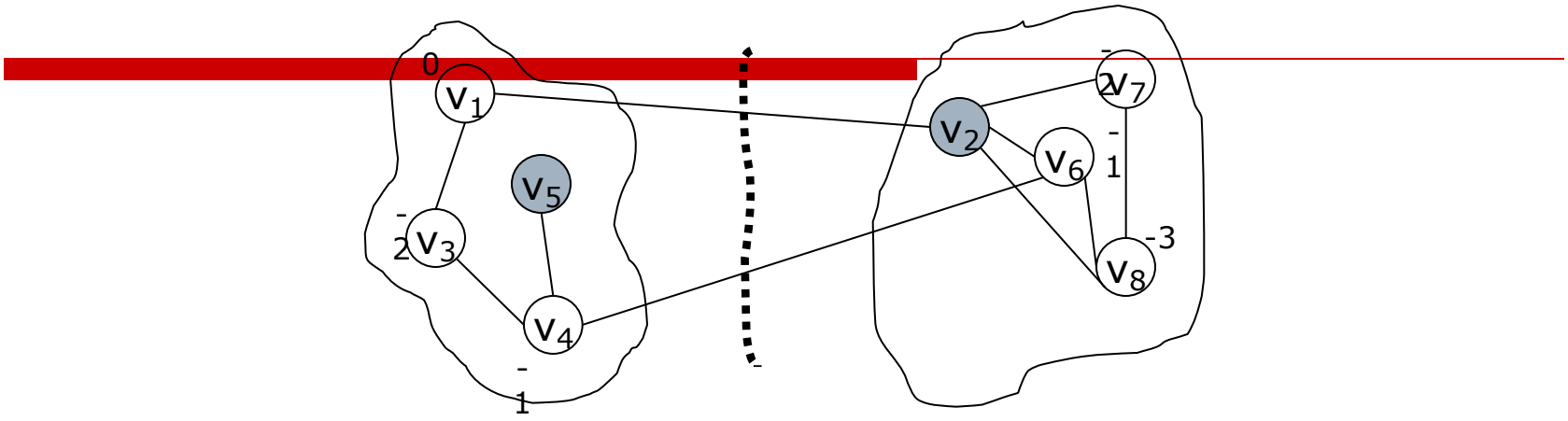


IF V_2 MOVES GAIN=2 and TOT_GAIN=2



IF V_5 MOVES GAIN=1 and TOT_GAIN=3

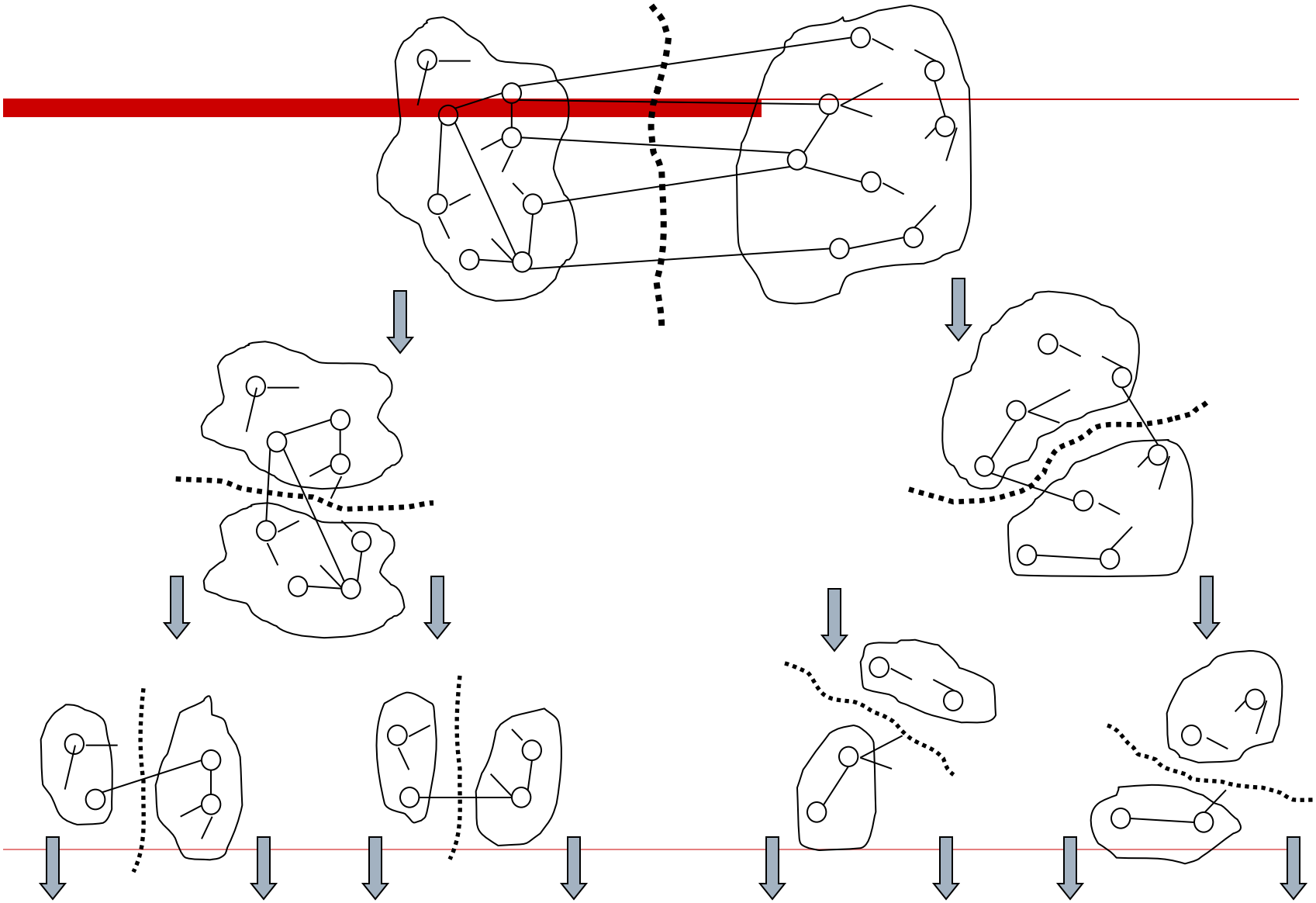
The Kernighan-Lin (KL) algorithm con't



IF V1 MOVES GAIN=0 and TOT_GAIN=3

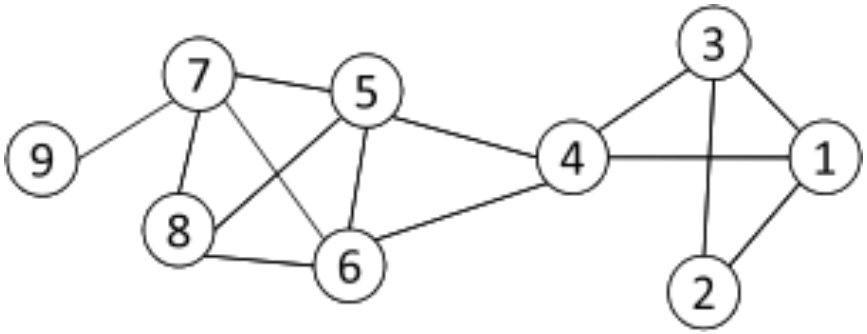
CUT=
2

The Kernighan-Lin (KL) algorithm con't



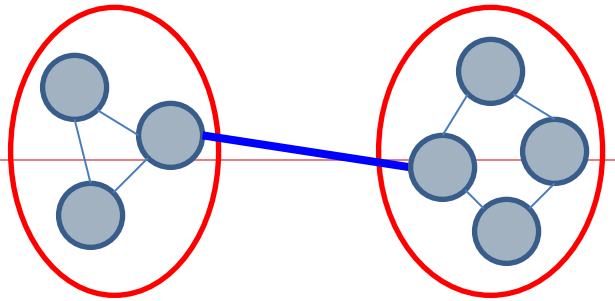
Edge Betweenness Method

- The strength of a tie can be measured by *edge betweenness*
- *Edge betweenness*: the number of shortest paths that pass along with the edge



The ***edge betweenness*** of e(1, 2) is 4 (=6/2 + 1), as all the shortest paths from 2 to {4, 5, 6, 7, 8, 9} have to either pass e(1, 2) or e(2, 3), and e(1,2) is the shortest path between 1 and 2

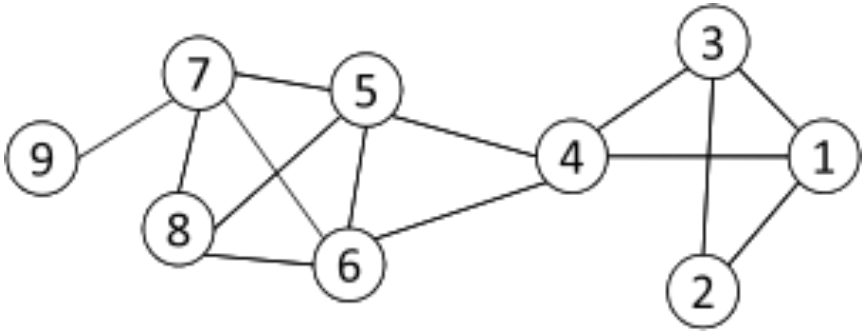
- The edge with higher betweenness tends to be the bridge between two communities.



Edge Betweenness Method

- Basic idea
  1. Calculate betweenness score for all edges
  2. Find the edge with the highest score and remove it from the network
  3. Recalculate betweenness for all remaining edges
  4. Repeat from step 2

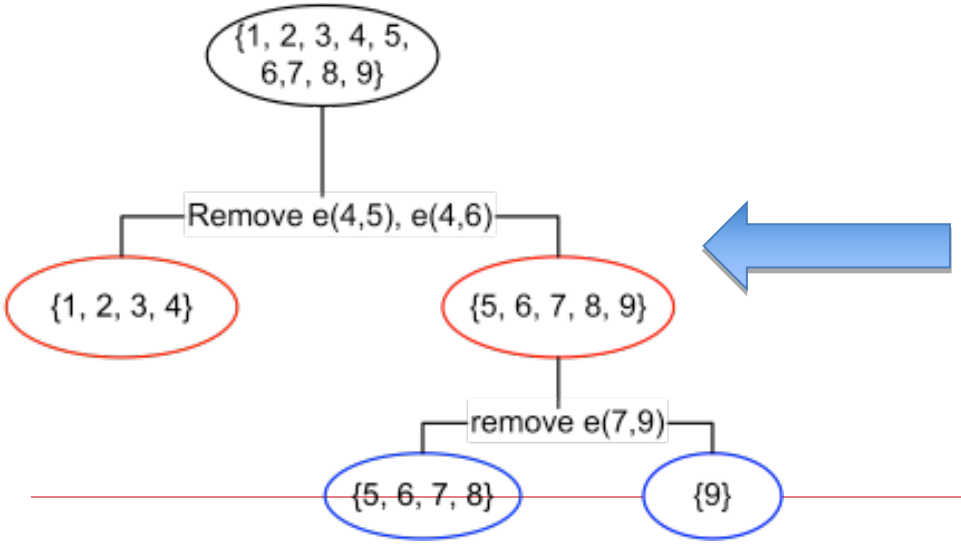
Edge Betweenness Method con't



Initial betweenness value

Table 3.3: Edge Betweenness

	1	2	3	4	5	6	7	8	9
1	0	4	1	9	0	0	0	0	0
2	4	0	4	0	0	0	0	0	0
3	1	4	0	9	0	0	0	0	0
4	9	0	9	0	10	10	0	0	0
5	0	0	0	10	0	1	6	3	0
6	0	0	0	10	1	0	6	3	0
7	0	0	0	0	6	6	0	2	8
8	0	0	0	0	3	3	2	0	0
9	0	0	0	0	0	0	8	0	0



After remove e(4,5), the betweenness of e(4, 6) becomes 20, which is the largest;

After remove e(4,6), the edge e(7,9) has the largest betweenness value 4, and should be removed.

Idea: progressively removing edges with the highest betweenness

Modularity Matrix

Q = ∑(Aij - ki*kj/M | i,j in the same group)

- Modularity matrix:

$$B = A - \mathbf{d}\mathbf{d}^T/2m \quad (B_{ij} = A_{ij} - d_i d_j/2m)$$

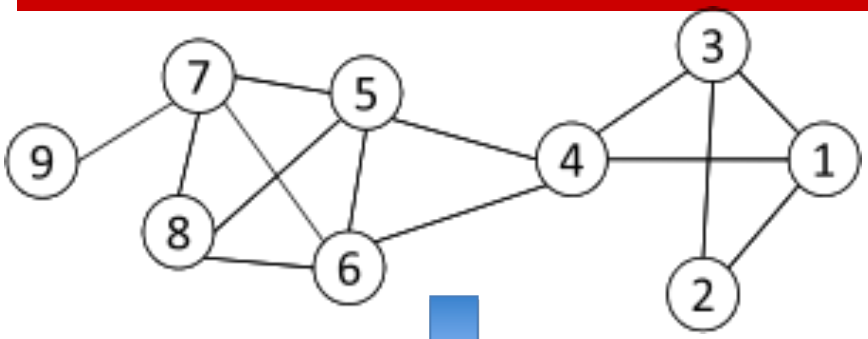
- Modularity maximization can be reformulated as

$$\max Q = \frac{1}{2m} Tr(S^T B S) \quad s.t.\ S^T S = I_k$$

- Optimal solution: top eigenvectors of the modularity matrix
- Apply k-means to S as a post-processing step to obtain community partition

$$Q = \frac{1}{2m} \sum_{\ell=1}^{k} \sum_{i \in C_\ell, j \in C_\ell} (A_{ij} - d_i d_j/2m)$$

Modularity Maximization Example



Two Communities:

{1, 2, 3, 4} and {5, 6, 7, 8, 9}

k-means

$$B = \begin{bmatrix} -0.32 & 0.79 & 0.68 & 0.57 & -0.43 & -0.43 & -0.43 & -0.32 & -0.11 \\ 0.79 & -0.14 & 0.79 & -0.29 & -0.29 & -0.29 & -0.29 & -0.21 & -0.07 \\ 0.68 & 0.79 & -0.32 & 0.57 & -0.43 & -0.43 & -0.43 & -0.32 & -0.11 \\ 0.57 & -0.29 & 0.57 & -0.57 & 0.43 & 0.43 & -0.57 & -0.43 & -0.14 \\ -0.43 & -0.29 & -0.43 & 0.43 & -0.57 & 0.43 & 0.43 & 0.57 & -0.14 \\ -0.43 & -0.29 & -0.43 & 0.43 & 0.43 & -0.57 & 0.43 & 0.57 & -0.14 \\ -0.43 & -0.29 & -0.43 & -0.57 & 0.43 & 0.43 & -0.57 & 0.57 & 0.86 \\ -0.32 & -0.21 & -0.32 & -0.43 & 0.57 & 0.57 & 0.57 & -0.32 & -0.11 \\ -0.11 & -0.07 & -0.11 & -0.14 & -0.14 & -0.14 & 0.86 & -0.11 & -0.04 \end{bmatrix}$$

Modularity Matrix

0.4384	-0.2709
0.3809	0.2671
0.4384	-0.2709
0.1716	0.6063
-0.2861	-0.3487
-0.2861	-0.3487
-0.3754	0.3355
-0.3421	0.1855
-0.1396	-0.1552

Spectral Clustering

- Both ratio cut and normalized cut can be reformulated as

$$\min_{S \in \{0,1\}^{n \times k}} Tr(S^T \widetilde{L} S)$$

- Where $\widetilde{L} = \begin{cases} D - A & \text{graph Laplacian for ratio cut} \\ I - D^{-1/2} A D^{-1/2} & \text{normalized graph Laplacian} \end{cases}$

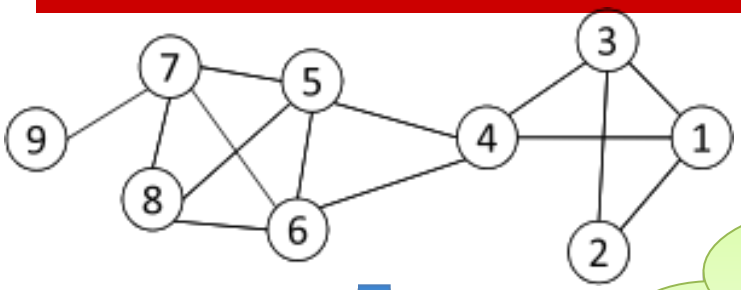
$D = diag(d_1, d_2, \cdots, d_n)$ A diagonal matrix of degrees

- Spectral relaxation:

$$\min_{S} Tr(S^T \widetilde{L} S) \quad s.t.\ S^T S = I_k$$

- Optimal solution: top eigenvectors with the smallest eigenvalues

Spectral Clustering Example



Two communities:

{1, 2, 3, 4} and {5, 6, 7, 8, 9}

The 1st eigenvector means all nodes belong to the same cluster, no use

k-means

$$D = diag(3, 2, 3, 4, 4, 4, 4, 3, 2)$$

$$\widetilde{L} = D - A = \begin{bmatrix} 3 & -1 & -1 & -1 & 0 & 0 & 0 & 0 & 0 \\ -1 & 2 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ -1 & -1 & 3 & -1 & 0 & 0 & 0 & 0 & 0 \\ -1 & 0 & -1 & 4 & -1 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 & 4 & -1 & -1 & -1 & 0 \\ 0 & 0 & 0 & -1 & -1 & 4 & -1 & -1 & 0 \\ 0 & 0 & 0 & 0 & -1 & -1 & 4 & -1 & -1 \\ 0 & 0 & 0 & 0 & -1 & -1 & -1 & 3 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 1 \end{bmatrix} \Rightarrow S = \begin{bmatrix} 0.33 & -0.38 \\ 0.33 & -0.48 \\ 0.33 & -0.38 \\ 0.33 & -0.12 \\ 0.33 & 0.16 \\ 0.33 & 0.16 \\ 0.33 & 0.30 \\ 0.33 & 0.24 \\ 0.33 & 0.51 \end{bmatrix}$$

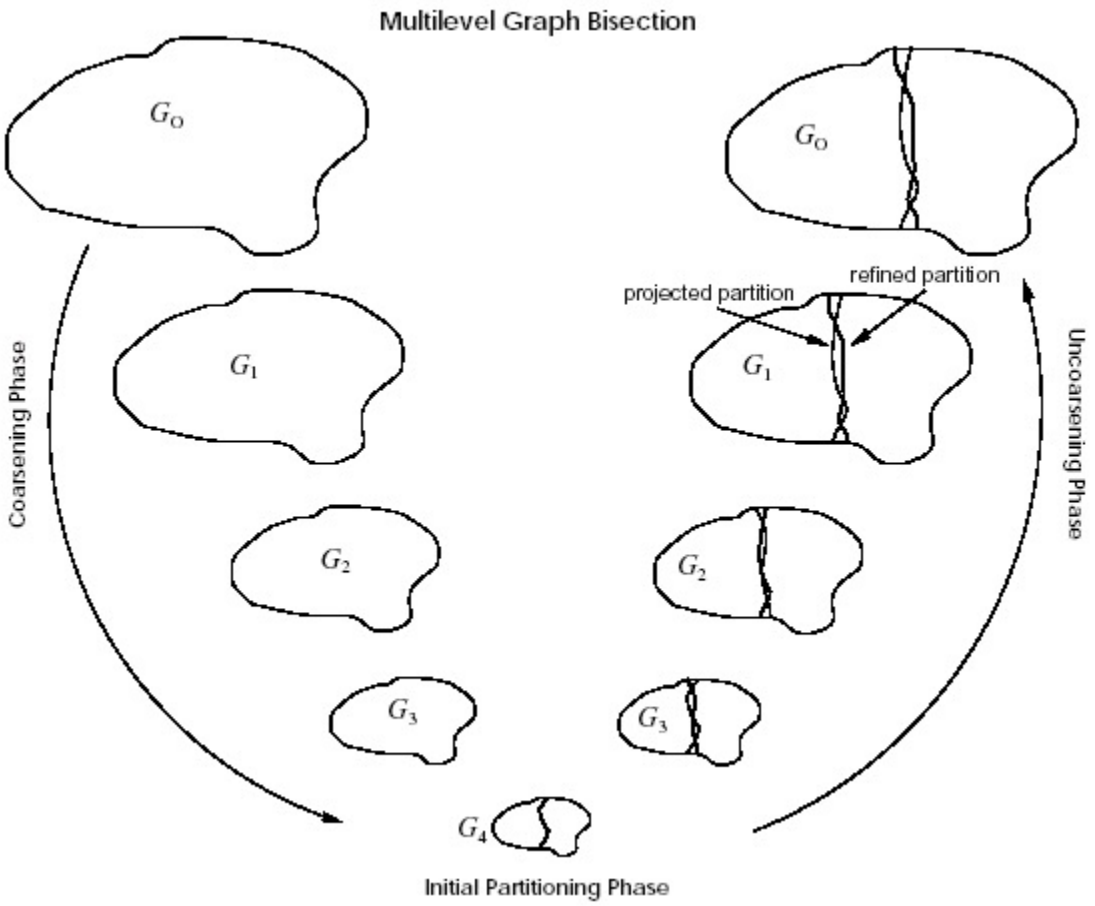
Centered matrix

Multi-level Graph Partitioning

- Logic flow
  1. Produce a smaller graph that is similar to the original graph
  2. A partitioning of the coarsest graph is performed.
  3. the partitioning of the coarser graph is projected back to the original graph. The partition is further refined.

Multi-level Graph Partitioning

- *3 Phases*
  - *Coarsen*
  - *Partition*
  - *Uncoarse*



Coarsening Phase

- A coarser graph can be obtained by collapsing adjacent vertices
  - Matching, Maximal Matching
- Different Ways to Coarsen
  - Random Matching (RM)
  - Heavy Edge Matching (HEM)
  - Light Edge Matching (LEM)
  - Heavy Clique Matching (HCM)

Other methods

- Markov Clustering
- Ratio Cut & Normalized Cut
- Local Graph Clustering
- Flow-Based Post-Processing method
- Shingling method
- …

Other works

- Community Discovery in Dynamic Networks
  - How should community discovery algorithms be modified to dynamic networks?
  - How do communities get formed?
  - How persistent and stable are communities and their members?
  - How do they evolve over time?
- Community discovery in Heterogeneous Networks
- Coupling Content Relationship Information for Community Discovery

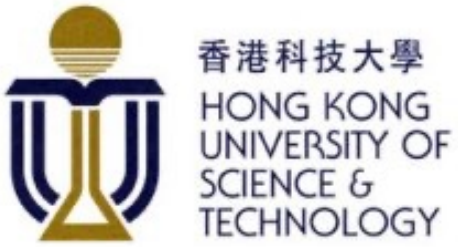
Issues

- Issues
  - Scalable Algorithms
  - Visualization of Communities and their Evolution
  - Incorporating Domain Knowledge
  - Ranking and Summarization in Community

Reference

1. Chapter 3, Community Detection and Mining in Social Media. Lei Tang and Huan Liu, Morgan & Claypool, September, 2010.
2. http://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=1&ved=0CFYQFjAA&url=http%3A%2F%2Fdelab.csd.auth.gr%2Fcourses%2Fc_mmdb%2Fmmdb-2011-2012-metis.ppt&ei=SEIjUJa8OKm0iQf0vYGIBg&usg=AFQjCNEP5NPt_GFIpnTQXye8l3Fzc8EHAg

Link Prediction in Social Networks

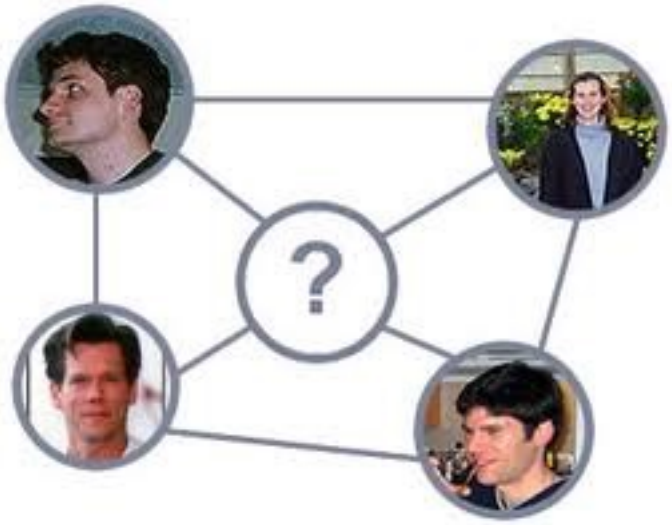


Outline

- Link Prediction Problems
  - Social Network
  - Recommender system
- Algorithms of Link Prediction
  - Supervised Methods
  - Collaborative Filtering
- Recommender System and The Netflixprize
- References

Link Prediction Problems

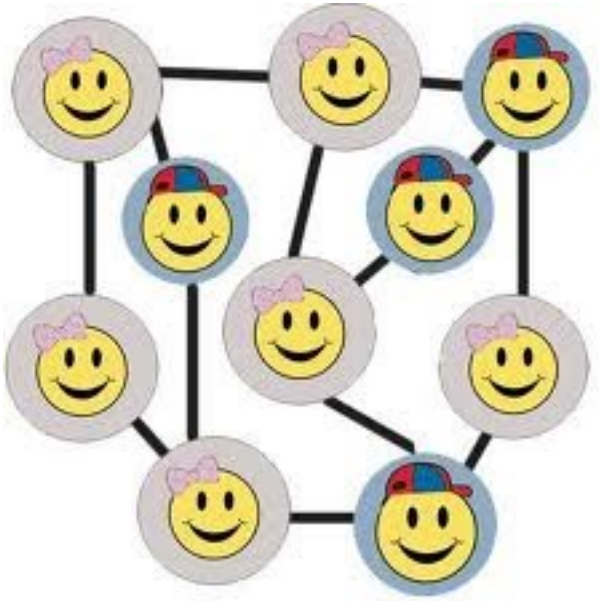
- Link Prediction is the task to predict the missing links in graphs.



- Applications
  - Social Network
  - Recommender systems

Links in Social Networks

- A **social network** is a social structure of people, linked(directly or indirectly) to each other through a common relation or interest
- **Links in Social network**
  - Like, dislike
  - Friends, classmates, etc.



Link Prediction in Social Networks

- Given a social network with an incomplete set of social links between a complete set of users, predict the unobserved social links
- Given a social network at time t predict the social link between actors at time t+1

(Source: Freeman, 2000)

Link Prediction in Recommender Systems

- Recommender Systems

Link Prediction in Recommender Systems

- Users and items form a bipartite-graph
- Predict links between users and items

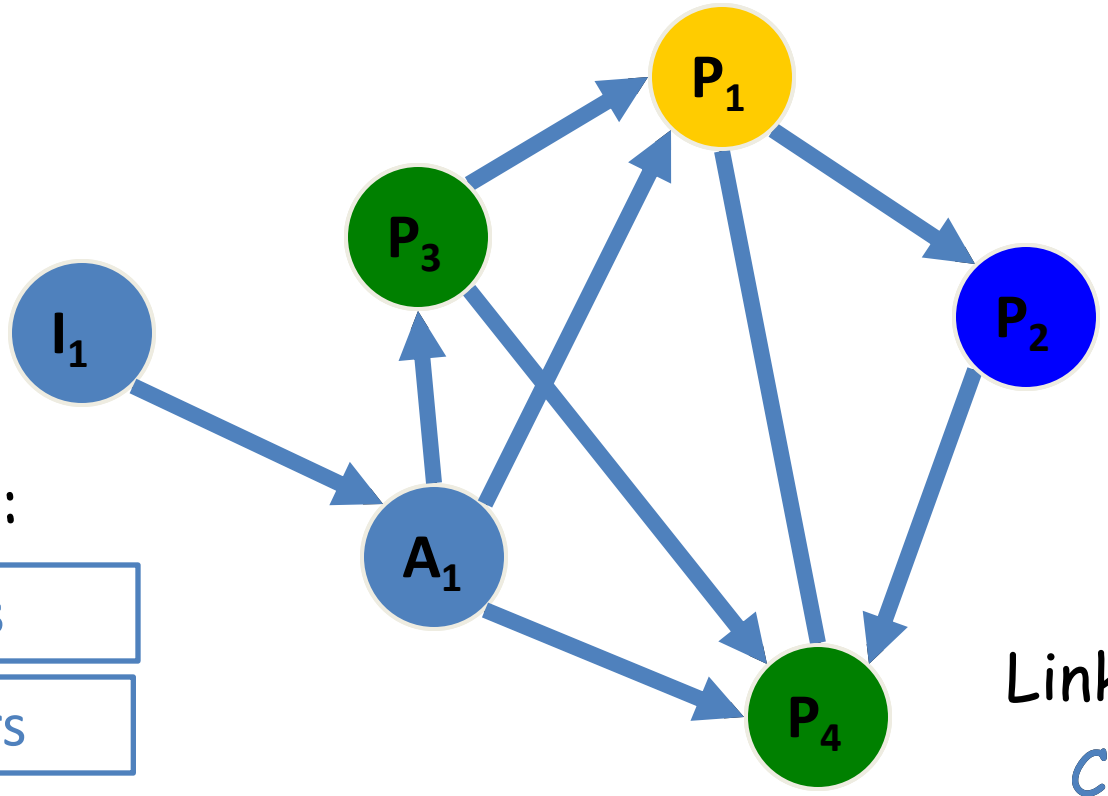
Predicting Link Existence

- Predicting whether a link exists between two items
  - **web**: predict whether there will be a link between two pages
  - **cite**: predicting whether a paper will cite another paper
  - **epi**: predicting who a patient's contacts are
- Predicting whether a link exists between items and users

Everyday Examples of Link Prediction/Collaborative Filtering...

- Search engine
- Shopping
- Reading
- Social
- ....
- **Common insight**: personal tastes are *correlated*:
  - If Alice and Bob both like X and Alice likes Y then Bob is more likely to like Y
  - especially (perhaps) if Bob knows Alice

Example: Linked Bibliographic Data



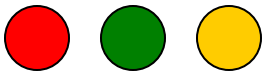
Objects:

- Papers
- Authors
- Institutions

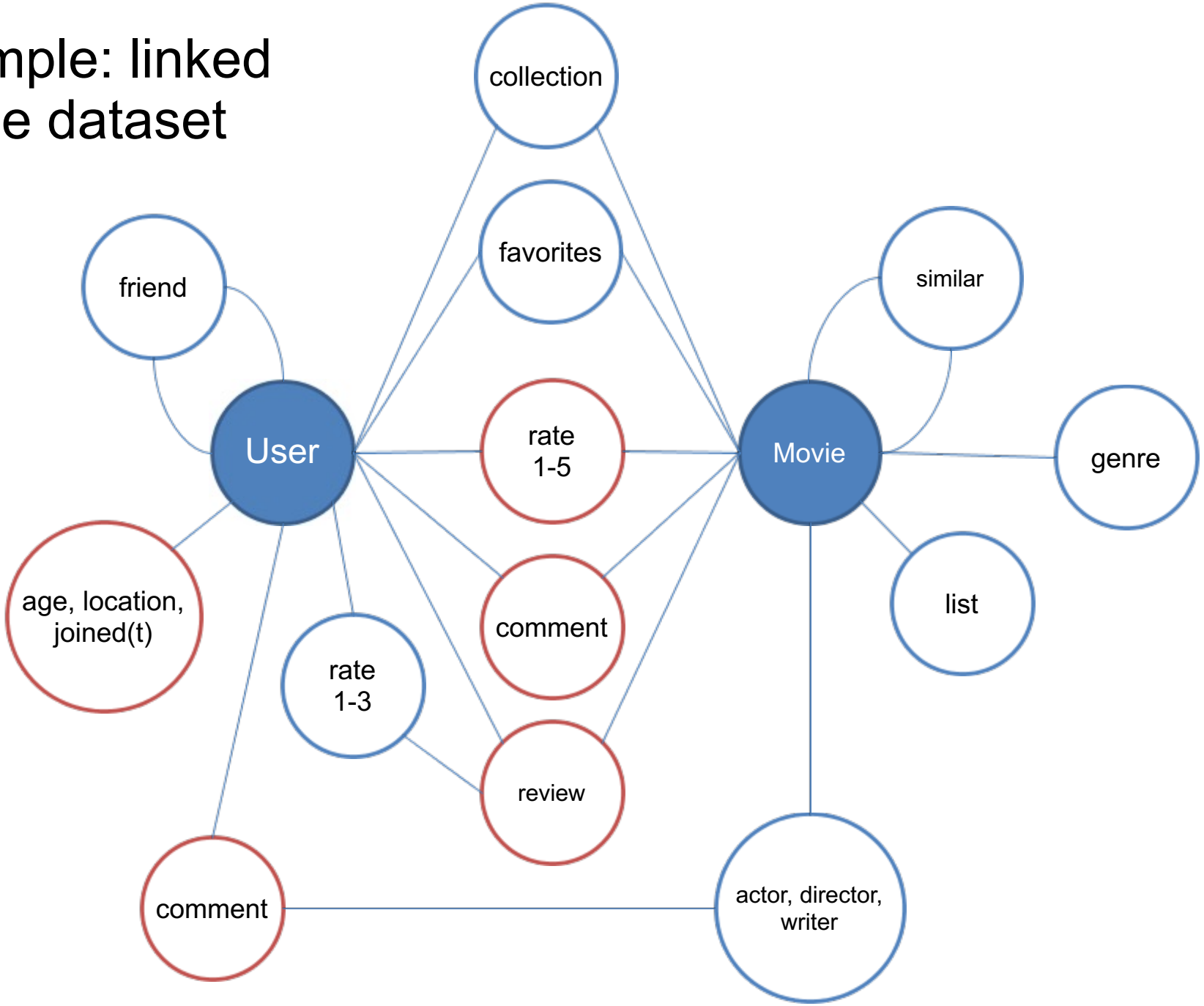
Links:

- Citation
- Co-Citation
- Author-of
- Author-affiliation

Attributes:



Example: linked movie dataset

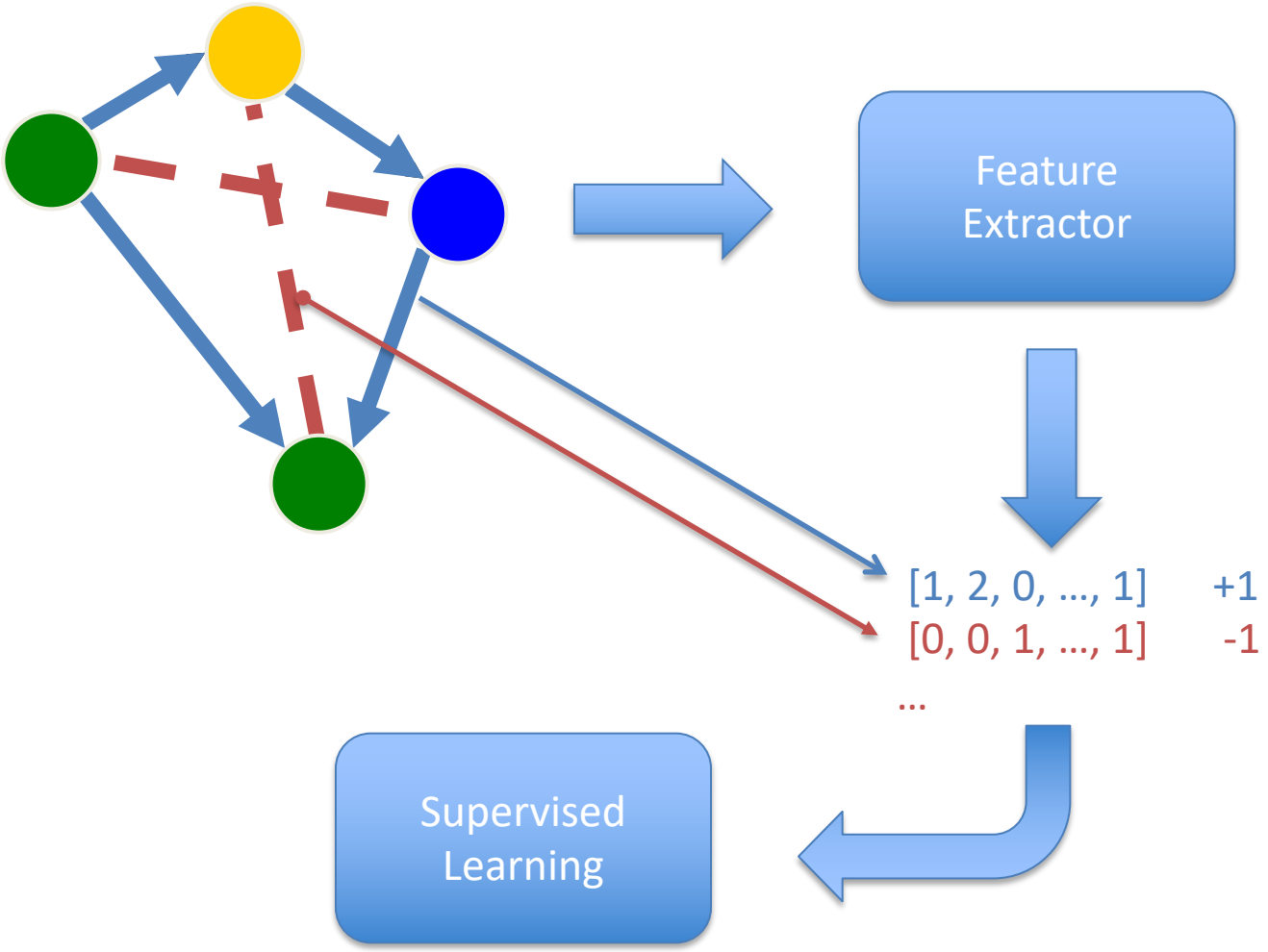


How to do link prediction?

amazon.com

How can you do recommendation based on this item?

Link Prediction using supervised learning methods



Supervised Learning Methods [Liben-Nowell and Kleinberg, 2003]

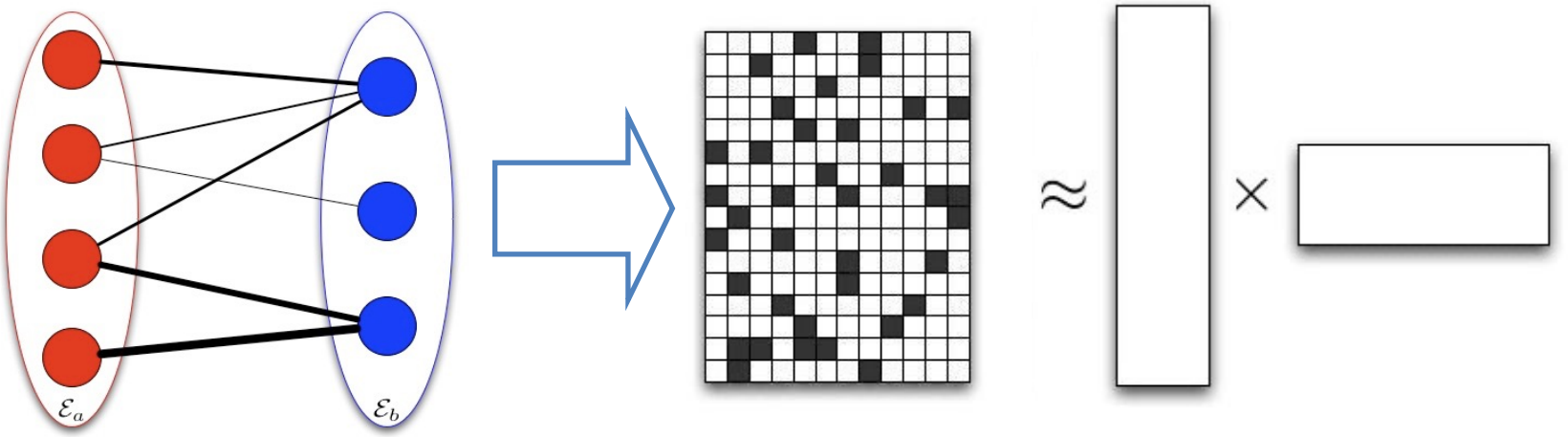
- Link prediction as a means to gauge the usefulness of a model
- Proximity Features: Common Neighbors, Katz, Jaccard, etc
- No single predictor consistently outperforms the others

supervised learning methods [Hasan et al, 2006]

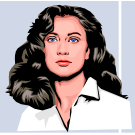
- Citation Network (BIOBASE, DBLP)
- Use machine learning algorithms to predict future co-authorship (decision tree, k-NN, multilayer perceptron, SVM, RBF network)
- Identify a group of features that are most helpful in prediction
- Best Predictor Features: Keyword Match count, Sum of neighbors, Sum of Papers, Shortest Distance

Link Prediction using Collaborative Filtering

- Find the background model that can generate the link data



Link Prediction using Collaborative Filtering

	Item 1	Item 2	Item 3	Item 4	Item 5
User 1 	8	1	?	2	7
User 2 	2	?	5	7	5
User 3 	5	4	7	4	7
User 4 	7	1	7	3	8
User 5 	1	7	4	6	?
User 6 	8	3	8	3	7

Challenges in Link Prediction

- Data!!!
- Cold Start Problem
- Sparsity Problem

Link Prediction using Collaborative Filtering

- Memory-based Approach
  - User-base approach [Twitter]
  - item-base approach [Amazon & Youtube]
- Model-based Approach
  - Latent Factor Model [Google News]

- Hybrid Approach

Memory-based Approach

- Few modeling assumptions
- Few tuning parameters to learn
- Easy to explain to users
  - Dear Amazon.com Customer, We've noticed that customers who have purchased or rated *How Does the Show Go On: An Introduction to the Theater* by Thomas Schumacher have also purchased *Princess Protection Program #1: A Royal Makeover (Disney Early Readers)*.

Algorithms: User-Based Algorithms (Breese et al, UAI98)

- $v_{i,j}$= vote of user *i* on item *j*
- I_i = items for which user *i* has voted
- Mean vote for *i* is

$$\overline{v}_i = \frac{1}{|I_i|} \sum_{j \in I_i} v_{i,j}$$

- Predicted vote for "active user" *a* is weighted sum

$$p_{a,j} = \overline{v}_a + \kappa \sum_{i=1}^{n} w(a,i)(v_{i,j} - \overline{v}_i)$$

normalizer

weights of *n* similar users

Algorithms: User-Based Algorithms (Breese et al, UAI98)

- K-nearest neighbor

$$w(a,i) = \begin{cases} 1 & \text{if } i \in \text{neighbors}(a) \\ 0 & \text{else} \end{cases}$$

- Pearson correlation coefficient (Resnick ’94, Grouplens):

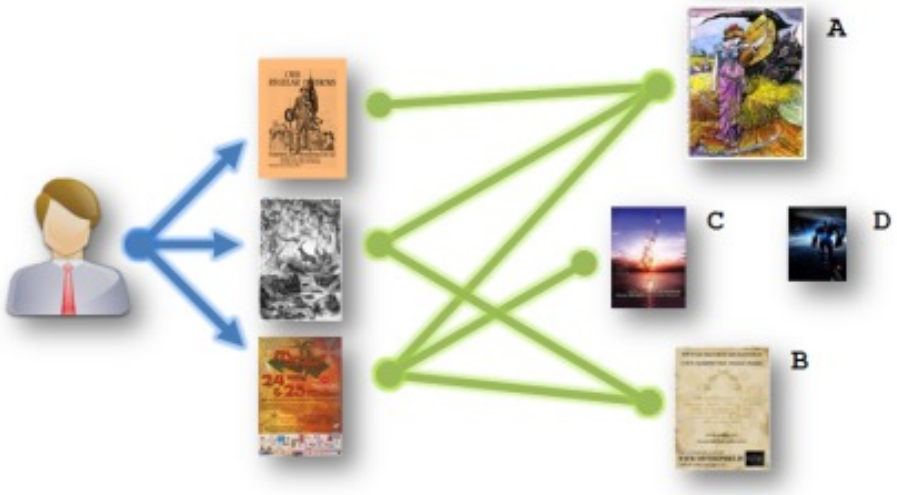
$$w(a,i) = \frac{\sum_j (v_{a,j} - \overline{v}_a)(v_{i,j} - \overline{v}_i)}{\sqrt{\sum_j (v_{a,j} - \overline{v}_a)^2 \sum_j (v_{i,j} - \overline{v}_i)^2}}$$

- Cosine distance (from IR)

$$w(a,i) = \sum_j \frac{v_{a,j}}{\sqrt{\sum_{k \in I_a} v_{a,k}^2}} \frac{v_{i,j}}{\sqrt{\sum_{k \in I_i} v_{i,k}^2}}$$

Algorithm: Amazon's Method

- Item-based Approach
  - Similar with user-based approach but is on the item side



Item-based CF Example: infer (user 1, item 3)

	Item 1	Item 2	Item 3	Item 4	Item 5
User 1 	8	1	?	2	7
User 2 	2	?	5	7	5
User 3 	5	4	7	4	7
User 4 	7	1	7	3	8
User 5 	1	7	4	6	?
User 6 	8	3	8	3	7

How to Calculate Similarity (Item 3 and Item 5)?

	Item 1	Item 2	Item 3	Item 4	Item 5
User 1 	8	1	?	2	7
User 2 	2	?	5	7	5
User 3 	5	4	7	4	7
User 4 	7	1	7	3	8
User 5 	1	7	4	6	?
User 6 	8	3	8	3	7

Similarity between Items

Item 3	Item 4	Item 5
?	2	7
5	7	5
7	4	7
7	3	8
4	6	?
8	3	7

- How similar are items 3 and 5?
  - How to calculate their similarity?

Similarity between items

Item 3	Item 5
?	7
5	5
7	7
7	8
4	?
8	7

- Only consider users who have rated both items
- For each user:
  Calculate difference in ratings for the two items
- Take the average of this difference over the users

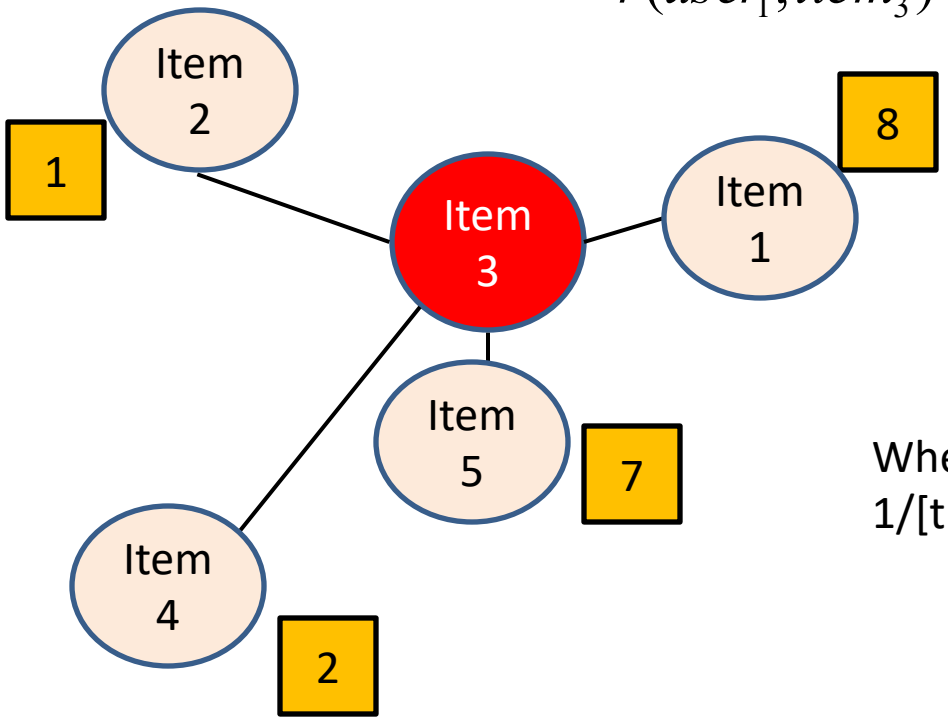
sim(item 3, item 5) = cosine((5, 7, 7), (5, 7, 8))

$= (5*5 + 7*7 + 7*8)/(\text{sqrt}(5^2+7^2+7^2)* \text{sqrt}(5^2+7^2+8^2))$

- Can also use Pearson Correlation Coefficients as in user-based approaches

Prediction: Calculating ranking

r(user1,item3)



$$r(user_1, item_3) = \alpha * \{r(user_1, item_1) sim(item_1, item_3) + r(user_1, item_2) sim(item_2, item_3) + r(user_1, item_4) sim(item_4, item_3) + r(user_1, item_5) sim(item_5, item_3)\}$$

Where α is a normalization factor, which is 1/[the sum of all sim($item_i$,$item_3$)].

Netflixprize

"We're quite curious, really. To the tune of one million dollars." – Netflix Prize rules

- Goal to improve on Netflix's existing movie recommendation technology
- Contest began October 2, 2006
- Prize
  - Based on reduction in root mean squared error (RMSE) on test data
  - $1,000,000 grand prize for 10% drop
  - Or, $50,000 progress for best result each year

Data Details

- Training data
  - 100 million ratings (from 1 to 5 stars)
  - 6 years (2000-2005)
  - 480,000 users
  - 17,770 “movies”
- Test data
  - Last few ratings of each user
  - Split as shown on next slide

Data about the Movies

Most Loved Movies	Avg rating	Count
The Shawshank Redemption	4.593	137812
Lord of the Rings :The Return of the King	4.545	133597
The Green Mile	4.306	180883
Lord of the Rings :The Two Towers	4.460	150676
Finding Nemo	4.415	139050
Raiders of the Lost Ark	4.504	117456

Most Rated Movies
Miss Congeniality
Independence Day
The Patriot
The Day After Tomorrow
Pretty Woman
Pirates of the Caribbean

Highest Variance
The Royal Tenenbaums
Lost In Translation
Pearl Harbor
Miss Congeniality
Napolean Dynamite
Fahrenheit 9/11

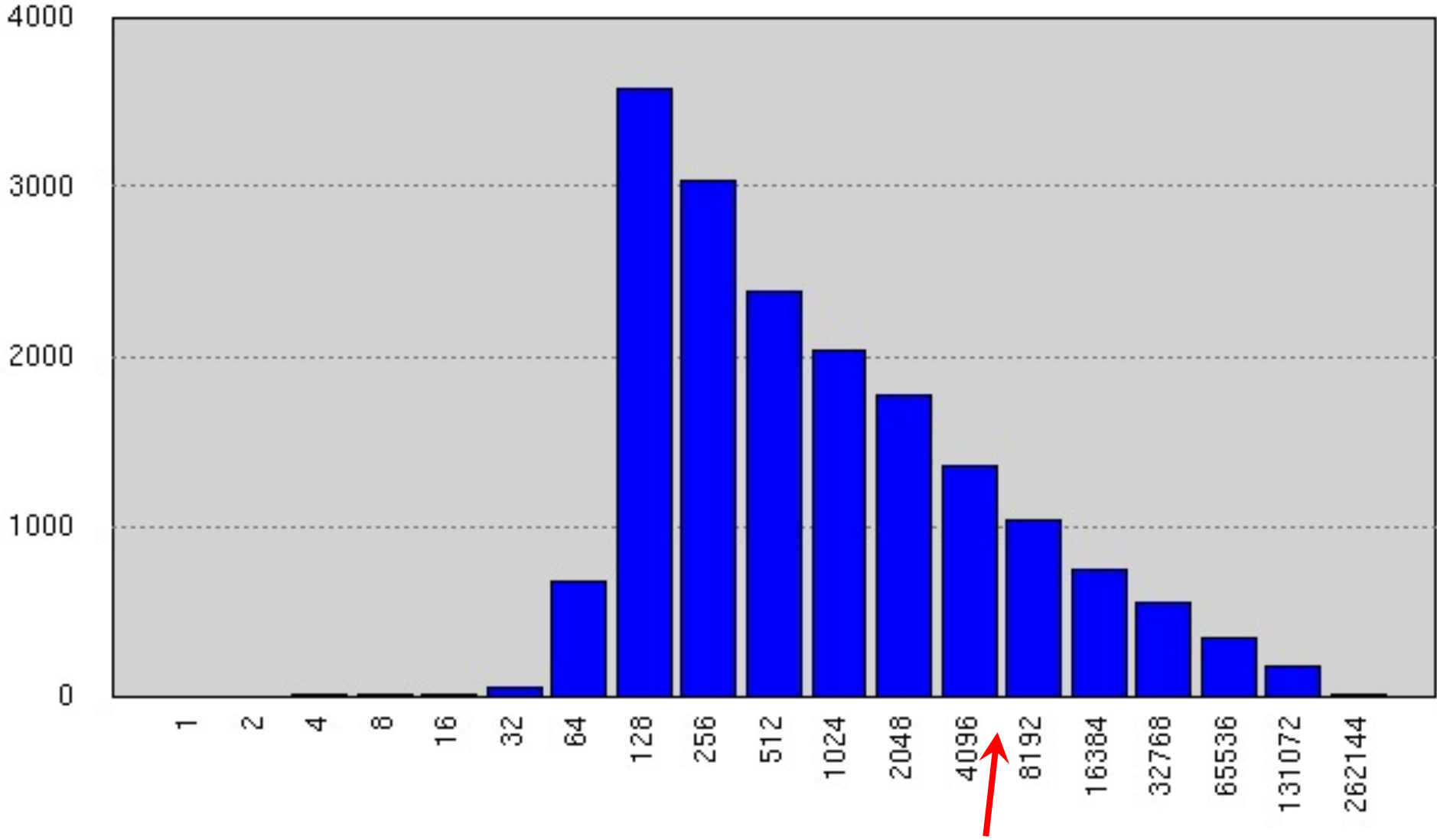
Major Challenges

1. Size of data
   - Places premium on efficient algorithms
   - Stretched memory limits of standard PCs
2. 99% of data are missing
   - Eliminates many standard prediction methods
   - Certainly *not* missing at random
3. Training and test data differ systematically
   - Test ratings are later
   - Test cases are spread uniformly across users

Major Challenges (cont.)

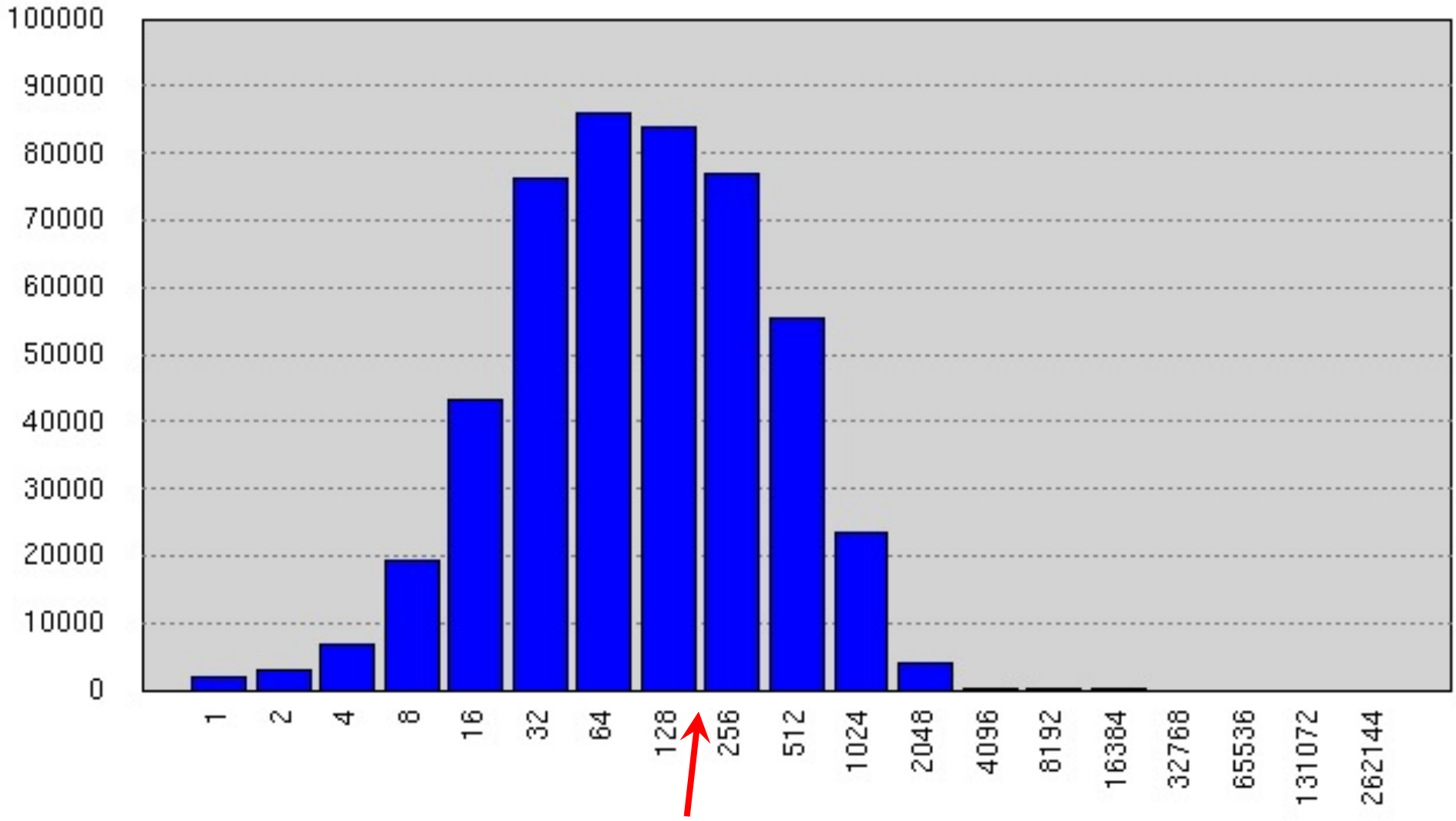
4. Countless factors may affect ratings
   - Genre, movie/TV series/other
   - Style of action, dialogue, plot, music et al.
   - Director, actors
   - Rater's mood
5. Large imbalance in training data
   - Number of ratings per user or movie varies by several orders of magnitude
   - Information to estimate individual parameters varies widely

Ratings per Movie in Training Data



Avg #ratings/movie: 5627

Ratings per User in Training Data



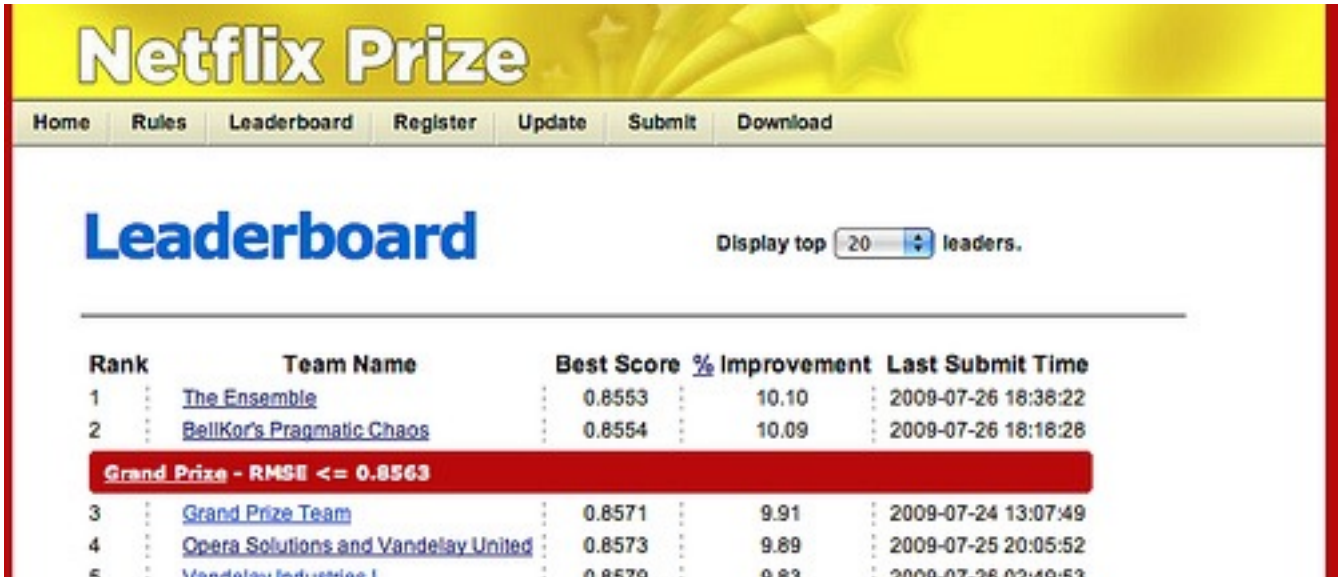
Avg #ratings/user: 208

The Fundamental Challenge

- How can we estimate as much signal as possible where there are sufficient data, without over fitting where data are scarce?

Test Set Results

- The Ensemble: 0.856714
- BellKor's Pragmatic Theory: 0.856704
- Both scores round to 0.8567
- Tie breaker is submission date/time



Lessons from Netflixprize

- Lesson #1: Data >> Models
- Lesson #2: The Power of Regularized SVD Fit by Gradient Descent
- Lesson #3: The Wisdom of Crowds (of Models)

References

- Koren, Yehuda. “Factorization meets the neighborhood: a multifaceted collaborative filtering model.” In *Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, 426–434. ACM, 2008. http://portal.acm.org/citation.cfm?id=1401890.1401944

- Koren, Yehuda. “Collaborative filtering with temporal dynamics.” *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD ’09* (2009): 447. http://portal.acm.org/citation.cfm?doid=1557019.1557072.

- Das, A.S., M. Datar, A. Garg, and S. Rajaram. “Google news personalization: scalable online collaborative filtering.” In *Proceedings of the 16th international conference on World Wide Web*, 271–280. ACM New York, NY, USA, 2007. http://portal.acm.org/citation.cfm?id=1242610.

- Linden, G., B. Smith, and J. York. “Amazon.com recommendations: item-to-item collaborative filtering.” *IEEE Internet Computing* 7, no. 1 (January 2003): 76-80. http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=1167344.

- Davidson, James, Benjamin Liebald, and Taylor Van Vleet. “The YouTube Video Recommendation System.” *Design* (2010): 293-296.