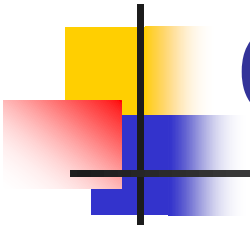
Algorithms and Social Network Analysis

吳邦一（Bang Ye Wu ）
中正資工系



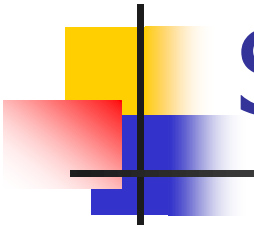
Outlines

- Social networks
- Research issues
  - Finding network structure
  - Features computation and analysis
  - Building model --- Random graphs
- Further topics

SOCIAL NETWORKS

Social network

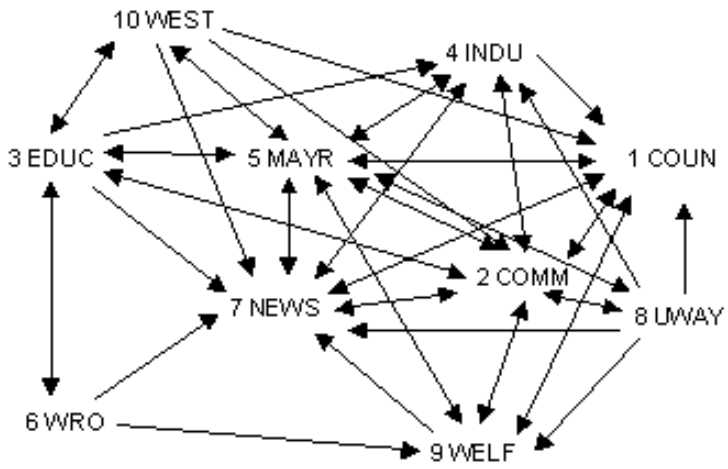
- Wiki: A **social network** is a social structure made of individuals (or organizations) called "nodes," which are tied (connected) by one or more specific types of interdependency,
  - such as friendship, kinship, financial exchange, dislike, sexual relationships, or relationships of beliefs, knowledge or prestige. (From Wiki)



Social network

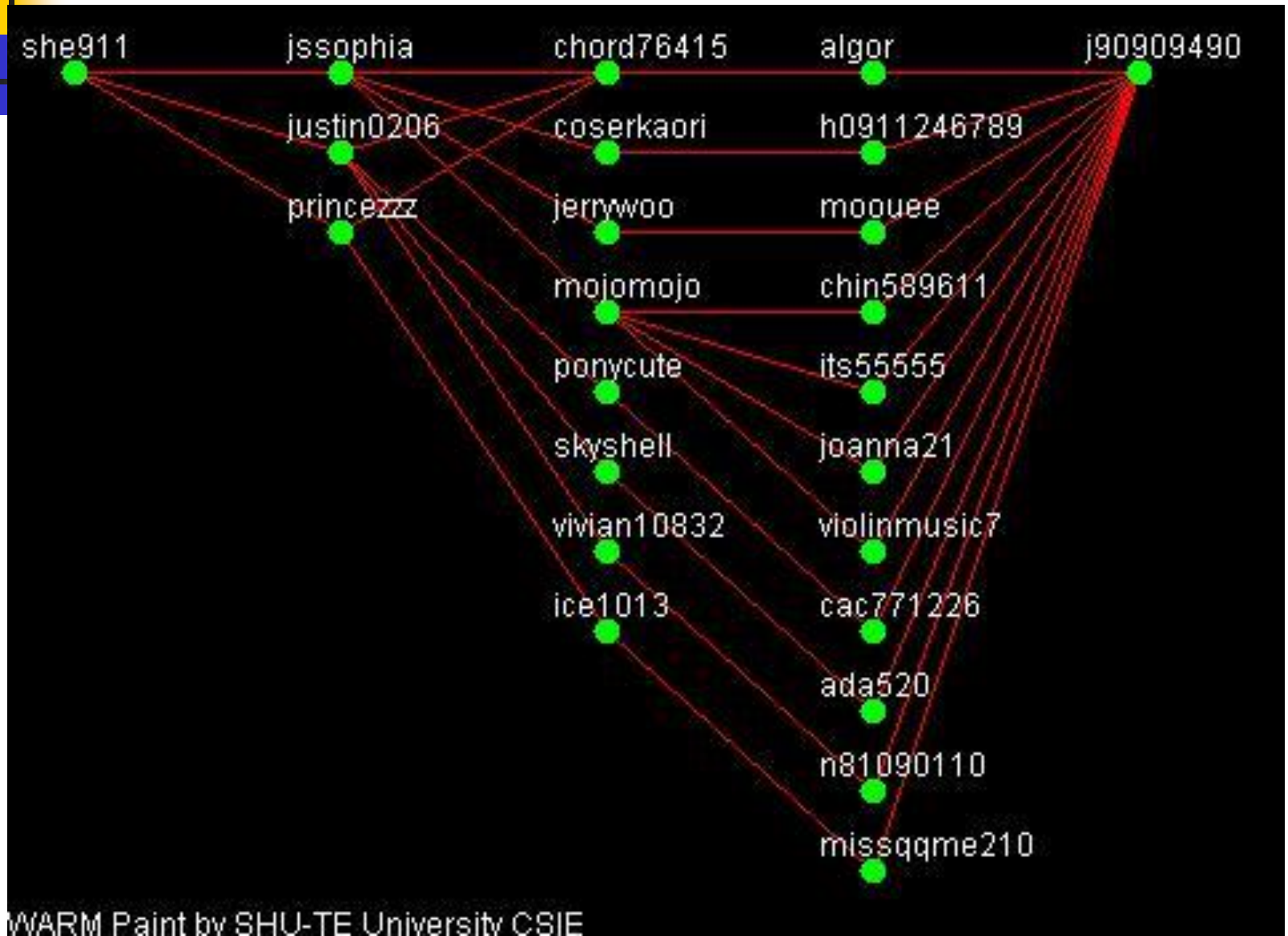
- Node: actor (people, group, organization)
- Arc (edge) : social relation tie, such as friend, collaboration, message transmission…
  - Directed or undirected (bidirectional or unidirectional)
- An example: Friend network
  - Node: people
  - Arc: friend relationship
  - In blogosphere: a node is a blog

Social network Example



- Information flow network
  - Node: organization
  - Edge: information flow

All shortest paths between SHE and Jolin on Wretch (無名 Blog, Taiwan)



Social networks--- Example

- Co-author network of scientists (Erdos number/distance)

MathSciNet® Mathematical Reviews on the Web

Search MSC | **Collaboration Distance** | Current Journals | Current Publications

MR Erdos Number = 3

Bang Ye Wu	coauthored with	Arie Tamir	MR2377190
Arie Tamir	coauthored with	Alexander Gersh Felzenbaum	MR0545729 (81g:90064)
Alexander Gersh Felzenbaum	coauthored with	Paul Erdős[1]	MR0942479 (89i:11004)

Change First Author | **Change Second Author** | **New Search**

MR Collaboration Distance = 5

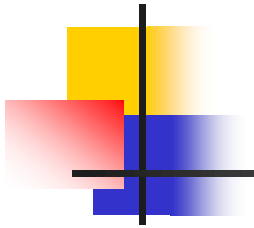
Bang Ye Wu	coauthored with	Arie Tamir	MR2377190
Arie Tamir	coauthored with	Wolfgang W. Bein	MR0777206 (86c:90039)
Wolfgang W. Bein	coauthored with	Alan J. Hoffman	MR1247604 (94k:90039)
Alan J. Hoffman	coauthored with	Ernst Gabor Straus	MR0080072 (18,185d)
Ernst Gabor Straus	coauthored with	Albert Einstein	MR0012947 (7,87j)

AMERICAN MATHEMATICAL SOCIETY
MathSciNet® *Mathematical Reviews* on the Web

MR Collaboration Distance = 5

Chien-Ping Chang	coauthored with	Lih-Hsing Hsu	MR1657345 (99g:68013)
Lih-Hsing Hsu	coauthored with	Sheng Bau	MR2154708 (2006a:05088)
Sheng Bau	coauthored with	Peter Dankelmann	MR2007647 (2004g:05113)
Peter Dankelmann	coauthored with	Ramamoorthi Ravi	MR2032568 (2004j:68142)
Ramamoorthi Ravi	coauthored with	Bang Ye Wu	MR1642908

Change First Author | Change Second Author | New Search



CATEGORIES AND ONLINE SOCIAL NETWORKS

Categories of Social networks

- Single relation
  - One-mode (relation between member):
    - Classmate, club members, colleague, like or dislike, coauthor, friendship
    - General graph, directed or undirected
  - Two-mode (group-member relation):
    - BBS board vs. author
    - Fans
    - teacher-student
    - Bipartite graph
  - Mixed mode
    - Blog with fans of celebogs, FB with fans
- Multiple relation
  - With more than one relation
  - Graph with edges of different colors

Categories of Social networks (II)

- Levels of Measurement of ties:
  - Binary
    - is this person a close friend or not
    - Yes or no; 1 or 0; a tie exists or not
  - Signed
    - whether you like, dislike, or don't care.
    - + , -, or 0
  - Valued
    - "rank the three people on this list in order of who you like most, next most, and least.“
      - Ordinal data
    - on a scale from minus one hundred to plus one hundred - where minus 100 means you hate this person, zero means you feel neutral, and plus 100 means you love this person - how do you feel about..."

Categories of Social networks (III)

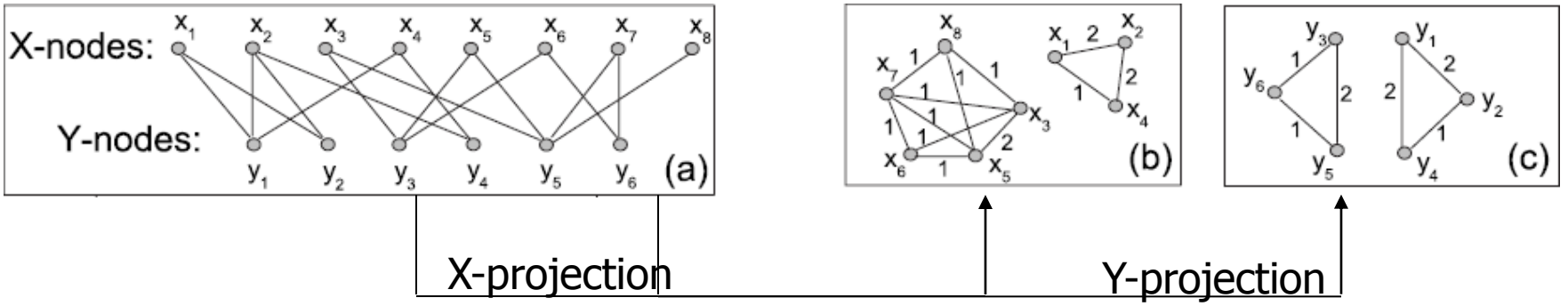
- Directed or "Bonded" Ties
  - Directed tie (arc)
    - Each person (ego) is being asked about ties or relations that they themselves direct toward others (alters).
    - arrows that have arrowheads
  - Bonded tie (undirected, bidirectional, edge)
    - A and B are having a conversation
    - line segments without arrowhead

Social networks on Internet

- One-mode
  - WWW hyperlink relation (Directed)
  - Blog friend relation (Directed?)
  - Facebook (Undirected?)
  - Instant messenger (Undirected)
- Two-mode
  - BBS
  - Celeblog (Blog fans relation)

2-mode (bipartite) network

- Group-member relation
  - PPT authors and boards
- One-mode projection
  - Using num of common neighbors as the strength (similarity) of tie



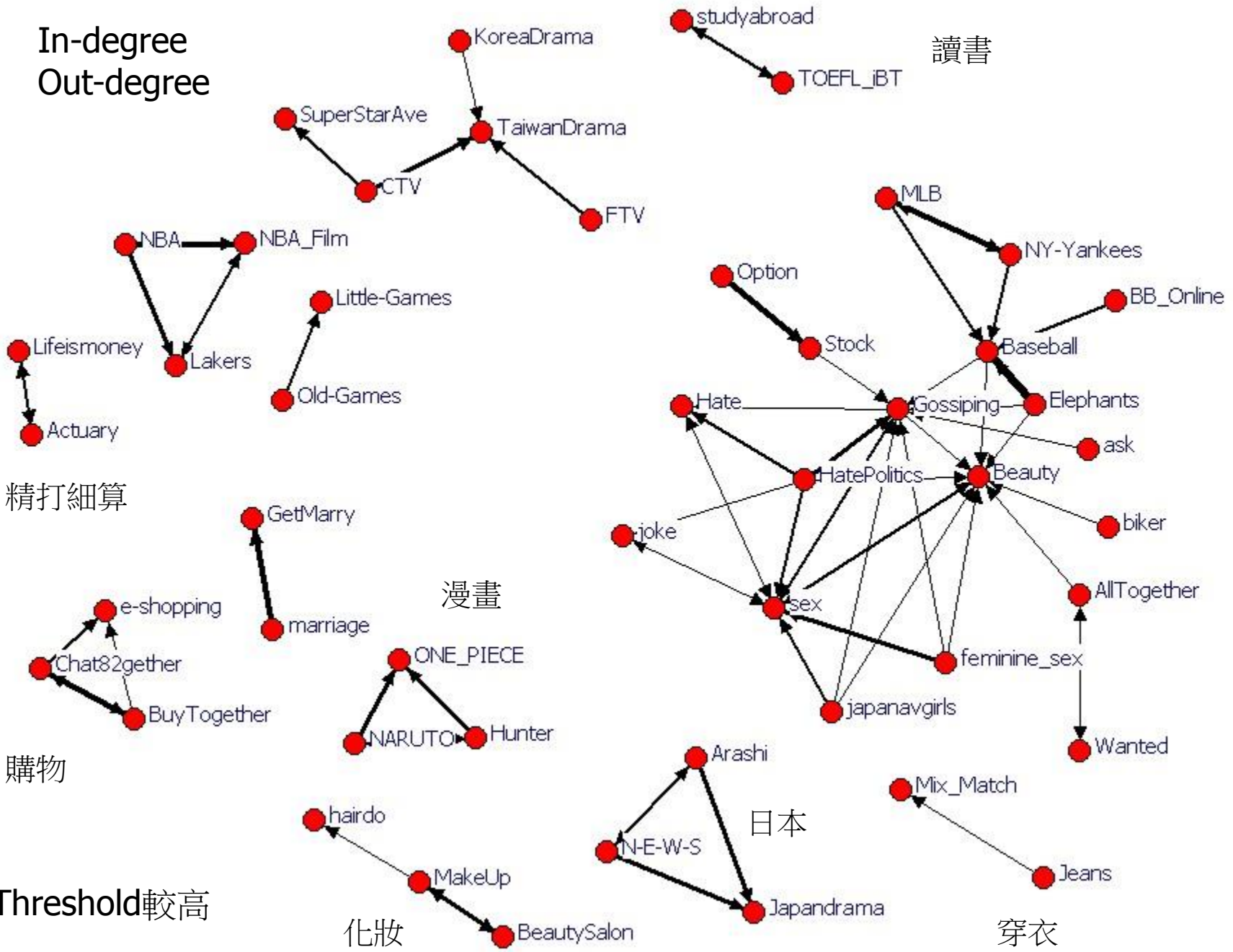
One-mode projection

- Used in personal recommendation and predictions
  - Friend recommendation in FB
- Simple but Loss information
  - Similarity in the other part is not considered
  - Only one-step neighborhood

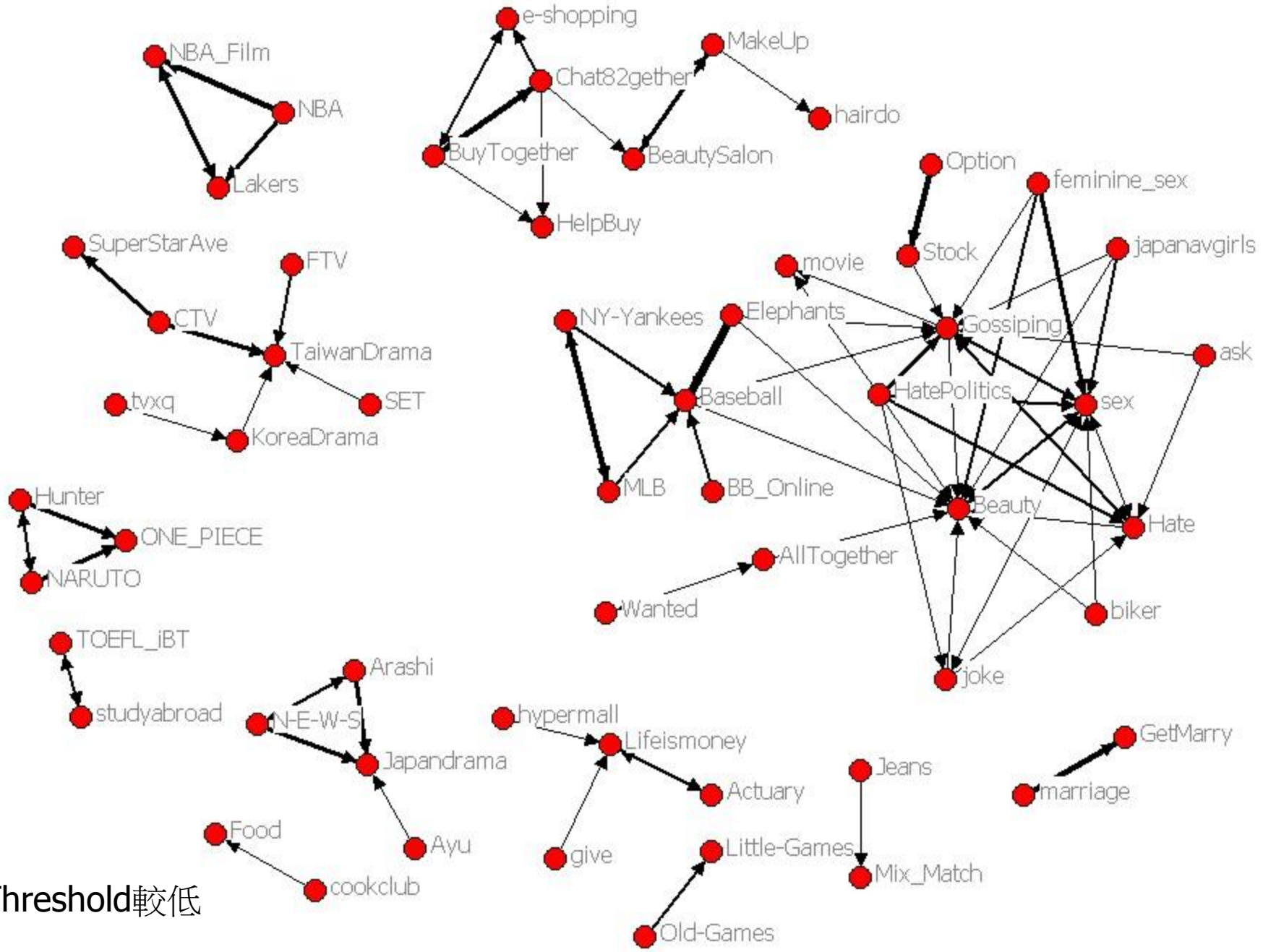
Example of 1-mode projection

- The relations between PTT boards
  - We analyze the authors for each board of PTT
  - 找出各版之間共同的帳號
  - 分析版與版的關連性強弱
  - 在118版中留下關聯性較強的版
  - A->B: 意味著在A版的人也有到B版的傾向
  - 線條粗細表示關聯強弱

In-degree
Out-degree



Threshold較高



Threshold較低

Mixed Social networks on Internet

- FaceBook, Twitter, line, Instagram
  - friendship and fans
  - information transmission and friendship
- YouTube, YouKu
  - people->film, people->people

Social networks on Internet

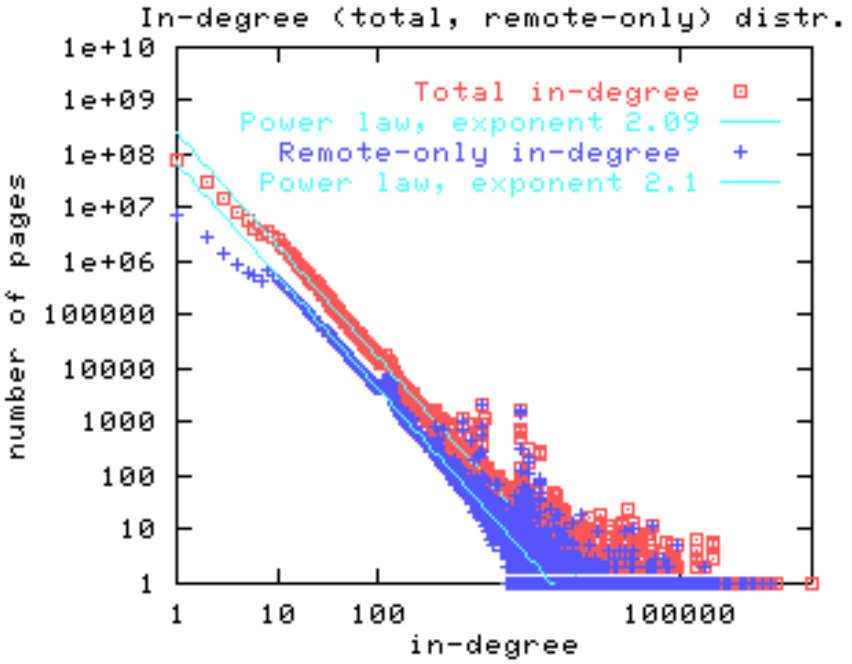
- Vs. traditional
  - Huge
  - Unreliable links
    - False positive (link to a beauty by a friend link)
    - False negative (lose of links by degree restriction or for some other reason)
  - Duplicated nodes
    - One may have several accounts

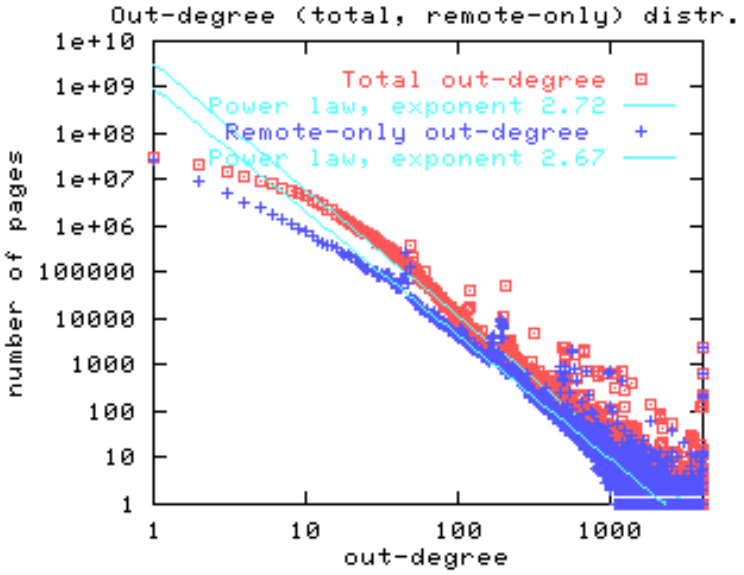
Social networks on Internet

- Typical features
  - Power-law and fat-tail degree distribution
    - Hub effect
  - Sparse
  - A large component
  - Small diameter (small world phenomenon, 6 degrees of separation)
    - Effective diameter: most of the nodes can be reached within this distance

Broder et al., Computer networks, 2000

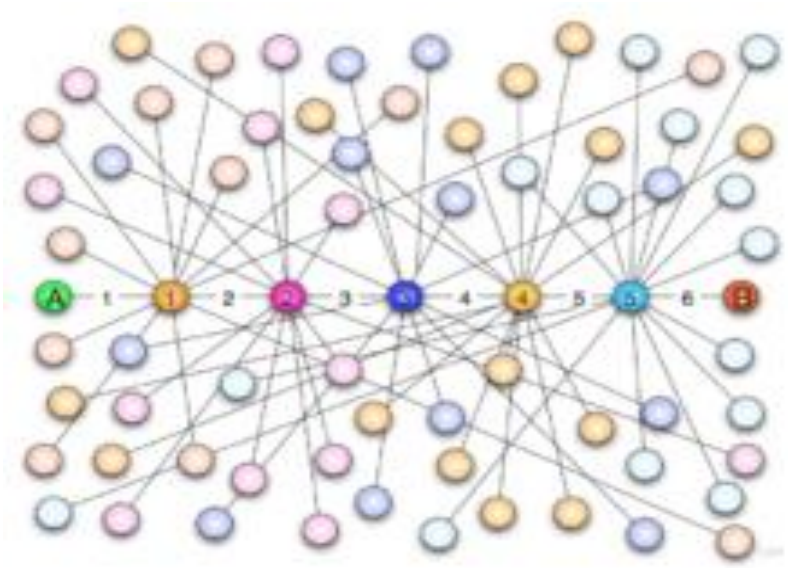
WWW Structure, degree distribution





Using crawler to retrieve 200 million pages and 1.5 billion links

Small world



- **Social network analysis** views social relationships in terms of network theory consisting of *nodes* and *ties*.
  - Example: six degrees of separation

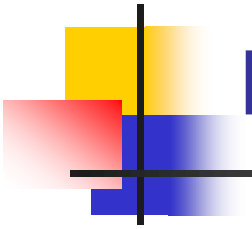
Facebook (November 2011, 721 million users with 69 billion friendship links): average distance=4.74, 99.91% of Facebook users in a large connected component.
The distance had decreased to 3.57 in February 2016, when it had 1.6 billion users (about 22% of world population).

RESEARCH ISSUES AND IMPORTANT METRICS

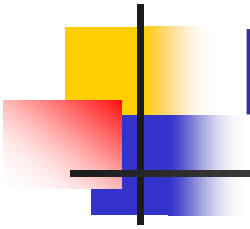
Research issues in social network analysis

- Finding network structure
  - Crawler—entire network or sampling
  - Latent relation mining (predicting) (Data Mining)
- Features computation and analysis
  - Important metrics (approximately or exactly)
    - Diameter, degree distribution, and other distribution
    - Centrality: degree, betweenness, closeness, and others
  - Sub-structure analysis
    - Community detection, clustering
  - Other applications
    - Influence analysis, personal recommendation, social trust

Research issues in social network analysis



- Building model --- Random graphs
  - Mathematical or computational
  - Generative models
    - Simulating the growth
  - Configuration models
    - Generating random graphs satisfying some given parameters

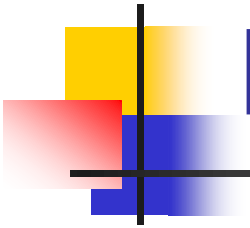


Finding network structure

- Crawling and sampling
  - Find the all relations by crawler, or sometimes we only want to sample them
  - Snowball sampling
    - Start from seeds and BFS
  - Induced subgraph
    - Guess the nodes and check the edges

Limitations and weaknesses

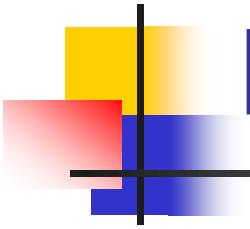
- Snowball method
  - Actors who are not connected (i.e. "isolates") are not located
  - may tend to overstate the "connectedness" and "solidarity" of populations of actors.
  - Missing neighbors for actors on the boundary
- Induced subgraph
  - Sampling rate should not be too small
  - Missing high-degree nodes loses important message
- Other methods: Forest fire, random walk, and some mixed method
  - A node with high degree has more prob. to be sampled.



Ego-centric networks

- Ego-centric network (ego only)
  - Ego: the node we want to study
  - Alter of ego: the neighbors of ego
  - Only the information on ego's connections to alters -- but not information on the connections among those alters.
- Ego-centric network with alter connections
  - The subgraph induced by N[v], i.e., the closed neighborhood of ego

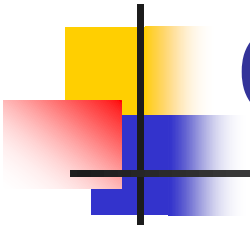
Finding network structure



- Relation prediction
  - Content analysis
    - AI methods
  - Structure analysis
    - Similarity
    - The simplest way: co-friends.
    - Two actors with many co-friends tend to be friends
  - Mixed method

Balance theory

- People tend to maintain balanced relationship.
  - Reciprocity: bidirectional tie, i.e., symmetry
  - Transitivity: a friend's friend
- A real social network (friendship)
  - Relatively high reciprocity
  - Relatively high transitivity



Clustering Coefficient

- Used to estimate the cohesion of the neighborhood of a node.
- For a node v,
  - Prob((u,w) in E | u,w in N(v))
- For a (sub)graph,
  - Prob((u,w) in E | (u,v) and (v,w) in E)
- By balance theory, a well developed social network has a larger CC.
  - A important feature of social networks

TYPICAL EXAMPLES

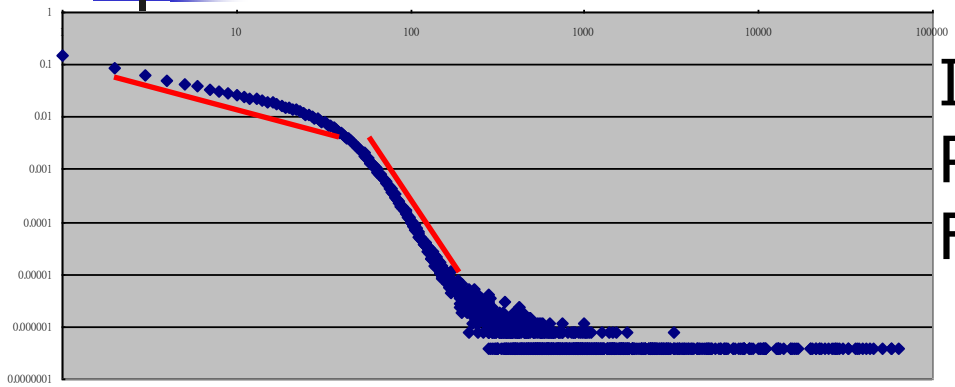
幾個Blog system的規模

Blog	節點數	邊數	雙向邊數	
Wretch	2,807,082	41,538,959	10,295,446	
Yam	173,633	1,412,626	446,299	
Pixnet	47,092	13,849	4,522	
Xuite	62,257	159,891	15,826	

2007/12

Till 2010/3, about 5M nodes and 90M links

Degree 分佈

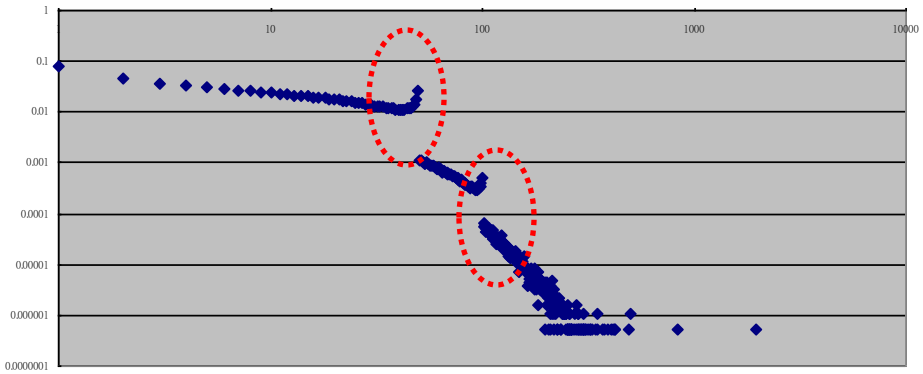


In log-log scale
Power-law distribution (2 stages)
Fat-tail

In-degree

Out-degree restriction for three kinds of members

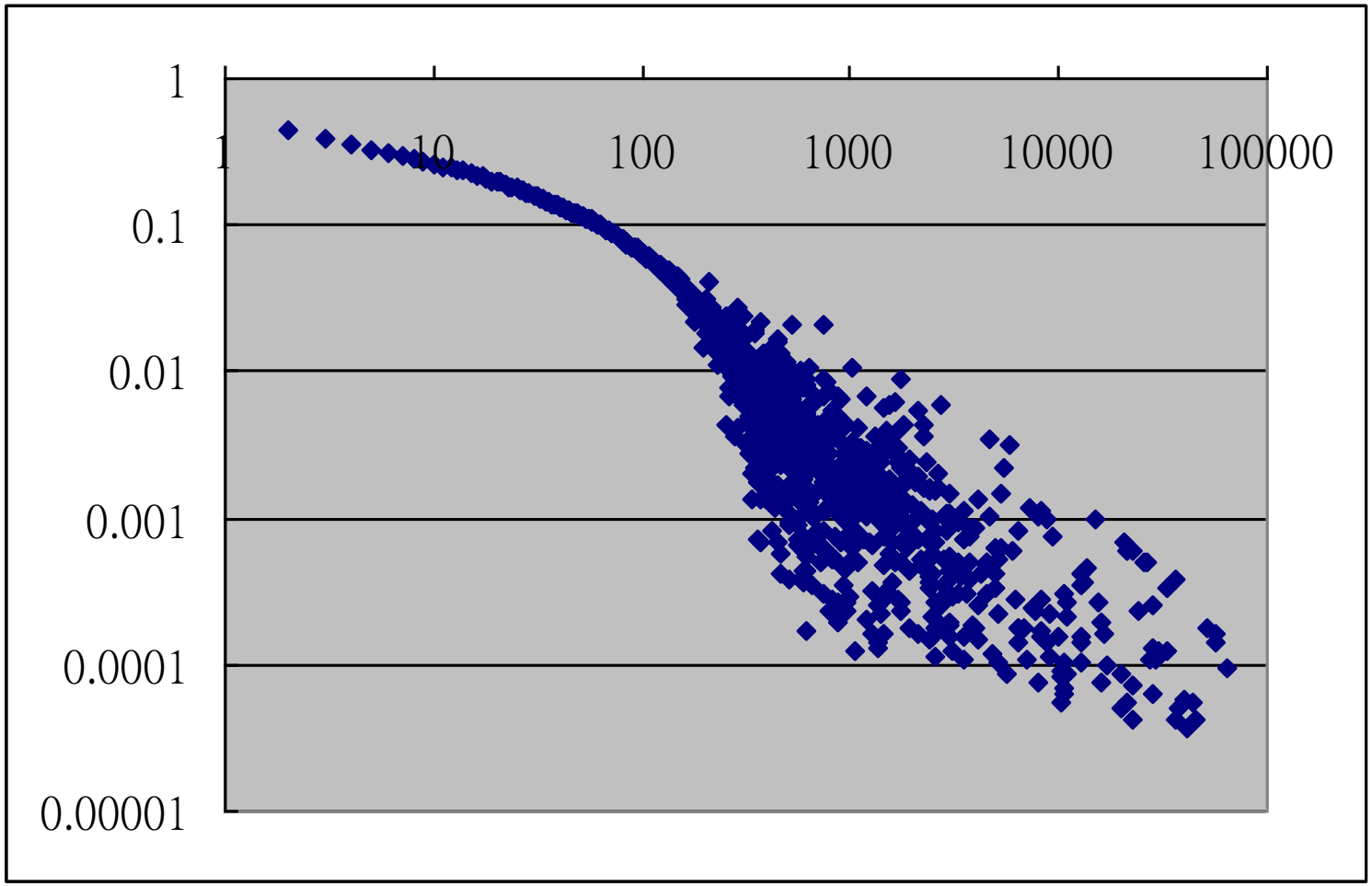
$$p_k \propto k^{-r}, \text{typically} 2 \le r \le 3$$



out-degree

Clustering 係數

Degree為k者其好友之間有關聯之機率



Results of South Korea 2007

From WWW2007

Set #	I	II	III	IV
	Cyworld	Testimonial(Cy)	MySpace	orkut
sampling ratio p	100%	0.77%	$\sim 0.08\%$	$\sim 0.30\%$
no. of nodes N	12,048,186	92,257	100,000	100,000
no. of edges L	190,589,667	703,835	6,854,231	1,511,117
mean degree $\langle k \rangle$	31.6	15.3	137.1	30.2
avg. clustering coefficient C	0.16	0.32	0.26	0.31
assortativity r	−0.13	0.43	0.02	0.31
estimated degree of separation ℓ	3.2	7.1	2.7	3.8

Table 1: Summary of data sets from online social networking services

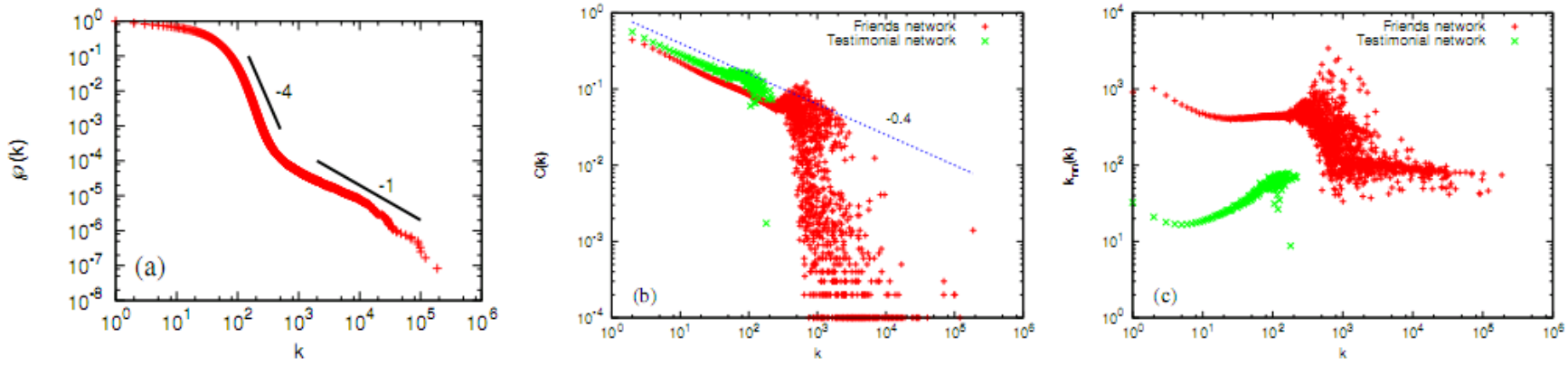
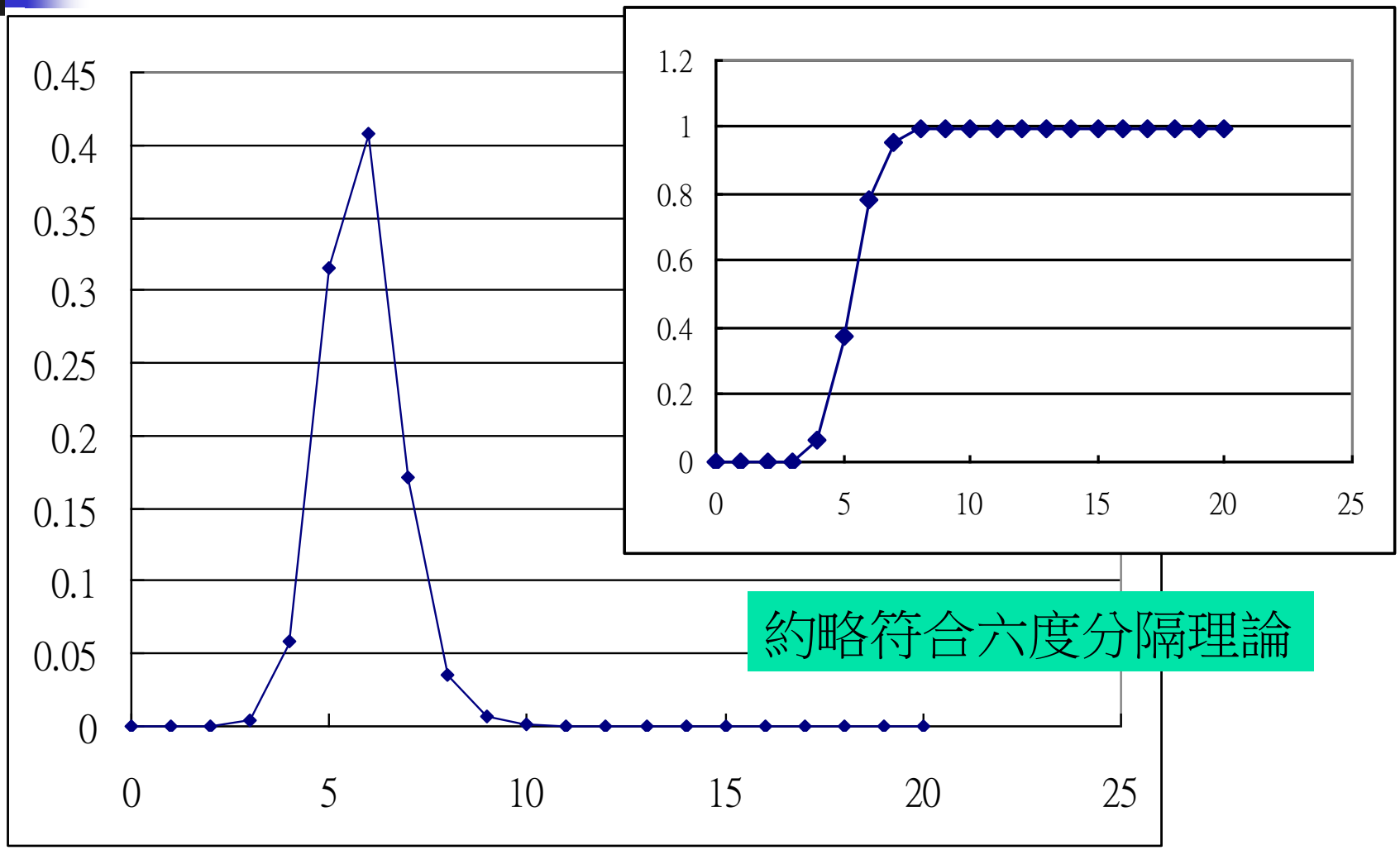
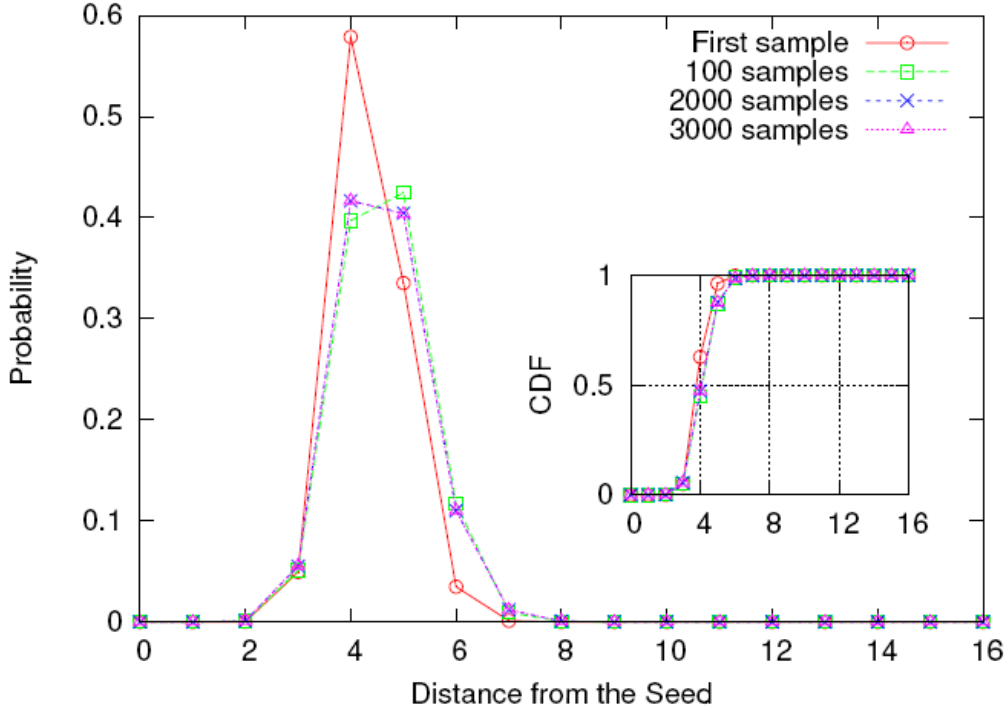


Figure 1: Topological characteristics of the complete Cyworld network

Average distance



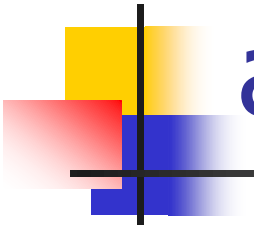
Distances in CyWorld



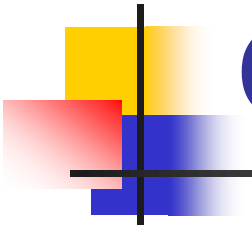
From WWW2007

Figure 2: Probability distribution function of distances between nodes

Features computation and analysis

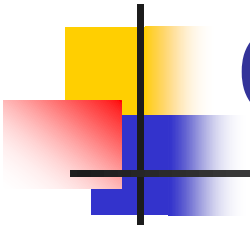


- Important metrics (approximately or exactly)
  - Diameter, degree distribution, and other distribution
  - Centrality: degree, betweenness, closeness, and others
- Sub-structure analysis
  - Community detection, clustering
- Other applications
  - Influence analysis, personal recommendation, social trust



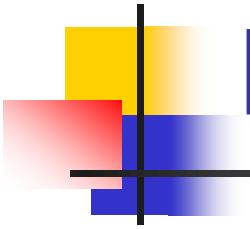
Centrality

- In SNA, centrality = importance of actor
- Ego's power is alter's dependence
  - Or ego’s power comes from alter’s powers
- Several measurements
  - Degree:
  - Closeness: total distance to others
  - Betweenness: how many pairs the node (edge) is at the geodesic path of them
  - Flow: How critical is it to the maximum flow



Closeness and Distance

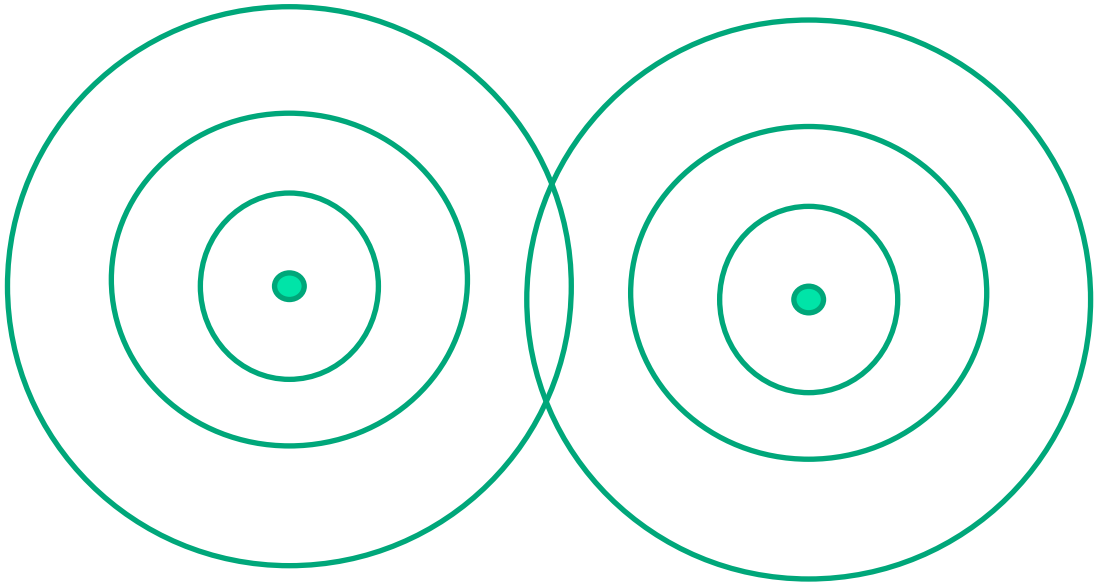
- Finding shortest st-path or single source all destinations
  - BFS or Dijkstra, unweighted or weighted
  - Intractable for huge networks, but no better available in the sense of worst case
- Several algorithms with better practical performances



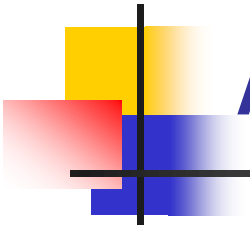
How to find all shortest paths

- It is easy by BFS (breadth first search)
  - Easy?
  - 吃飯是小事, 但是十億人吃飯就是件大事
  - Wretch has roughly 5M nodes and 80M edges. Typically $N_i=30N_{i-1}$
  - Most of the nodes has distance 5 or 6
    - A simple BFS almost search the whole graph

2-way BFS

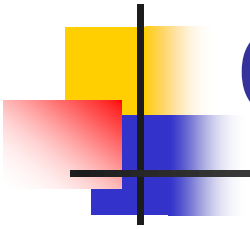


- We need to search roughly $2*30^3$ nodes
- 2-way Dijkstra can also used in weighted graphs



ALT

- ALT algorithm:
  - A* search, Landmarks and Triangle inequality. Basically, a prune Dijkstra.
  - Good performance, especially on geometric data, such as road map. Also tested on social networks.



Computing the diameter

- iFUB algorithm:
  - iterative Fringe Upper Bound
- Bounding-Diameter algorithm
  - Using pruning method
- 4-sweep algorithm and its modification
  - Using 4 BFS to find the approximate diameter

Experimental results

Non-Social Network graphs

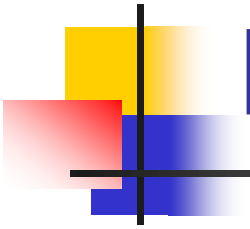
Graphs	nodes	edges	Avg. degree
inet	1,719,037	11,095,298	6.45
ip	2,250,498	19,394,216	8.61
facebook	4,039	88,234	21.85
Slashdot0902	82,168	948,464	11.54
Slashdot0811	77,360	905,468	11.70
Wiki-Vote	7,115	103,689	14.57
soc-Epinions1	75,879	508,837	6.71
soc-pokec-relationships	1,632,803	30,622,564	18.75
soc-LiveJournal1	4,847,571	68,993,773	14.23
amazon	334,863	925,872	2.77
dblp	317,080	1,049,866	3.31
orkut	3,072,441	117,185,083	38.14
youtube	1,134,890	2,987,624	2.63
road net-TX	1,393,383	3,843,320	2.76
road net-CA	1,971,281	5,533,214	2.81
road net-PA	1,090,920	3,083,796	2.83

Experiment result for modified iFUB with deleting methods

		The number of BFS executions						
Graph	exact	i+HD	BD	iC	iC+BD1	iC+BD2	iC+BD3	iC+BD4
INET	5	699164	578998	699164	466314	450517	528543	280021
IP	5	289771	421	289771	90749	89999	129434	75455
facebook	8	143	4	207	158	158	36	3
Slashdot0902	13	5	11	4	4	4	6	6
Slashdot0811	12	4	8	3	3	3	4	4
Wiki-Vote	7	14	15	14	14	14	12	2
Epinions1	15	8	9	8	8	8	14	14
pokec	14	2	6	2	2	2	2	2
LiveJournal	20	6	8	6	6	6	9	7
amazon	47	52	14	52	14	14	14	6
DBLP	23	9	31	15	7	7	7	11
orkut	10	39295	70	99	98	98	91	94
Youtube	24	12	3	12	12	12	5	18
road-net TX	1064	589878	95	544280	1843	1759	2923	103
road-net PA	794	263606	59	535961	17347	17418	13190	88
road-net CA	865	354382	152	354382	1727	1694	2815	143

Good for most social networks

- TC Lin and BY Wu, 2016

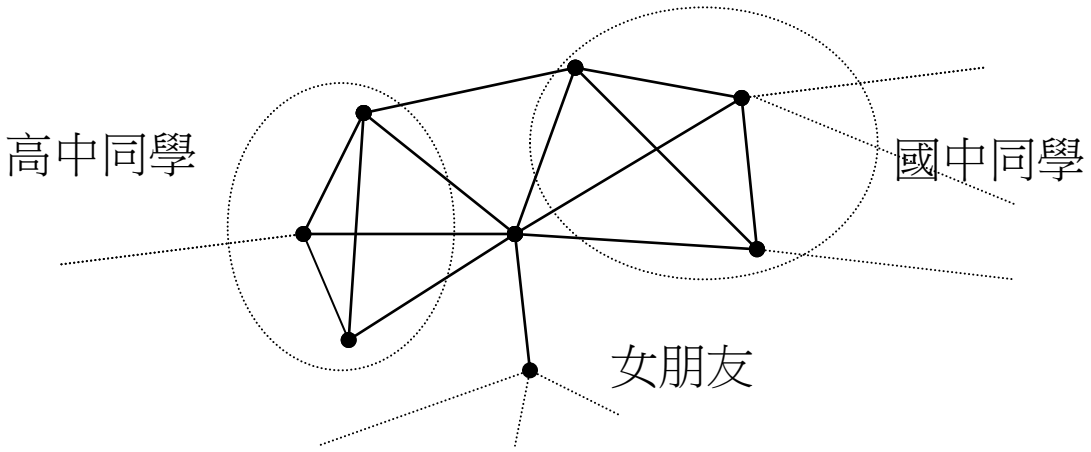


Betweenness centrality

- Betweenness indicates the importance of an actor in message flow
- Brandes algorithm
  - Two-pass BFS (or Dijkstra)
  - O(mn), but too costly for large networks
- Approximation or modified definition

Community detection

- 好友群(for friend network)
  - 一群彼此聯繫緊密的朋友，她們內部聯繫緊密，而對外部的人的連結相對較疏
  - 一個人的好友通常可以分成若干群



Community detection

- Cohesive group
  - Clique: not practical
  - Several relaxations
    - K-clique: the distance on original graph $\leqq$ k
    - K-club: subgraph of diameter k
    - K-plex: each connected to all but at most k nodes
    - …
  - Max Clique and Clique Partition are NP-hard
    - Also for many clique-like problem, e.q., Densest k-subgraph
    - And hard to approximate

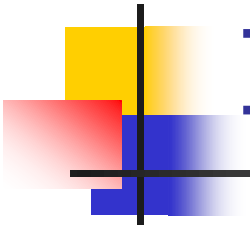
DIFFERENT APPROACHES FOR DEFINING SUBSTRUCTURES AND GROUPS

Substructure Name	Description
Clique	Actors who have all possible ties among themselves
N-Clique	Actors are connected to every member of the group at a maximum distance of N
N-Clans	N-Cliques that all paths among members occur by the way of other members of N-Clique
K-Plex	Clique in which actors have ties to all but k of members of the group
K-Core	Actors are connected to k of members of the group
Component	Parts of sociogram that are connected within bu disconnected with other components
Cut Points	Nodes which if removed, the structure becomes divided into un-connected systems
Block	The divisions into which cutpoints divide a graph
Lambda Set	Set of actors who if disconnected, would most greatly disrupt the flow among all of the actors

Community detection (partition)

- Partitioning the nodes into several groups
- Variants
  - Overlap or non-overlap
  - Fixed or variable num of groups
- Measurement the goodness of clustering: modularity (by Newman 2006), good for sparse graphs.
- Finding the max modularity is also NP-hard, but existing approximation algorithms

$$Q = \frac{1}{2m} \sum_{ij} \left[A_{ij} - \frac{k_i k_j}{2m} \right] \delta(c_i, c_j),$$



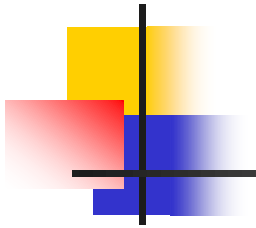
Influence issue

- How an idea (rumor) can be spread over almost the whole network?
  - Depending on the models?
    - How one be influenced by the other
    - How one will send a message heard from others
- Maximum coverage problem
  - Domination set and similar problems

An simple mathematic result

- Each person who has heard the idea communicates it to neighbors with independent probability *r*
  - The expected number who hear it is r(k-1).
  - The average number of people a person passes the idea on to, also called the *basic reproductive number* R_0, is

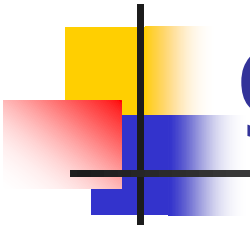
$$R_0 = r\frac{\sum_i k_i(k_i-1)}{\sum_i k_i} = r\frac{\langle k^2\rangle - \langle k\rangle}{\langle k\rangle}.$$



- If $R_0 > 1$, then the number of people hearing the idea grows as it gets passed around and it will take off exponentially. If $R_0 < 1$ then the idea will die.
- We have a phase transition for the spread of the idea: it spreads if and only if

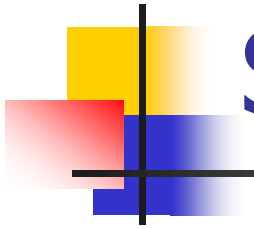
$$r > \frac{\langle k \rangle}{\langle k^2 \rangle - \langle k \rangle}.$$

Building the models – random graphs



- Used to explain how the network grows and predict.
  - Predict the behavior of actors.
  - Generate random graph for experiments.
- Generate random graphs with some given features.

Some Popular random graphs



- Bernoulli random graph
  - Each edge has i.i.d. probability
  - Easy, simple but not social network
    - Deg distribution: binomial (approximately Poisson)
- Preferential attachment
  - Rich-get-richer
- Random graphs with any given degree distribution
- Usually, model generates graphs with some, but not all, features
  - Used to explain the reasons
  - Practical networks are much more complex

Further issue

- 現實生活中的社會網路是各種relation的混合體而更加的複雜
- 理論的模型都簡化了非常多的考量
  - Type of flow
- 其他重要研究主題
  - Social trust, privacy and security, visualization (Drawing)

Different network flows

- Centrality and community should not only consider shortest path
- But the influence depends on how the flow process
  - Rumors are spread not necessarily along the shortest paths.

Typology of flow processes
-classification

1. Transmission dimension

   parallel duplication, serial duplication, and transfer

2. Trajectory dimension

   geodesic path, path, trails, and walks

Table 1
Typology of flow processes

	Parallel duplication	Serial duplication	Transfer
Geodesics	<No process>	Mitotic reproduction	Package delivery
Paths	Internet name-server	Viral infection	Mooch
Trails	E-mail broadcast	Gossip	Used goods
Walks	Attitude influencing	Emotional support	Money exchange

Borgatti, Social Networks 2005

Social trust

- FaceBook 是個混合體
  - 有單向的粉絲，有朋友
  - 有真實的訊息，也有更多的搞笑KUSO
  - 訊息的轉傳，幾乎是不需要任何代價
- 如何判斷資訊的正確性
  - 沒有人去抱怨金庸的小說串改歷史，因為擺明是小說
  - 但網路的資訊往往是片段的，也多半沒有註明出處，甚至難以查證。謠言傳了一年半載還在傳的比比皆是

Trust propagation

- 信任的人所信任的也是信任的
- 類似網頁的重要性，因此有些研究採取PageRank-like的方式來計算在社會網路中的Trust問題
  - Time consuming
  - Implementation可能還有諸多問題

網路的假資料

- 匿名與實名制
  - 網路的發展與安全必成相互矛盾的兩件事
  - 如何取得平衡或是發展出其他方式
    - 不知道

Privacy and Anonymity

- 發覺隱私與預測是同一件事
  - 給定一些資料，經過分析計算，預測
    - 你跟誰會成為好朋友 ⇔ 你跟誰其實是好朋友
    - 你對什麼有興趣 ⇔ 你其實有什麼興趣
  - 研究的題目通常是預測
  - 反方的研究題目通常是隱私保護

發布資料與隱私保護

- 資料的公開是為研究
- 與隱私是相互矛盾的兩件事
- 拿掉資料的標籤並不足夠
- 唯一性與獨特性的資料容易被攻擊

SNA的匿名性研究

- 在盡量不變動網路的某些特性目標下，將一個社會網路修改成滿足某種匿名性。
  - degree變動最小，滿足degree k-anonymity
    - 每種出現的degree都至少出現k個
  - 可能希望同時滿足closeness, betweenness或其他特性

資訊暗黑再度降臨，
不同於隻手遮天的曾經，
這次老大哥就是你我自己，
與那自詡主宰的網路之神。

資訊氾濫圖文橫陳，
一次次地按讚之中，
遞移與平衡理論之宿命，
依你情之所鍾，
想看的喜歡的，
眼前鋪陳；
逆耳的苦澀的，
消失無蹤；
逐步失去的是不同，
慢慢遠離的是多樣性，
繽紛的世界，各自單調的聲音。

懶人包與偽裝的鄉民，
以為的資訊其實是行銷的專門，
人們呀，
發文為的是影響而不是分享；
閱讀為的是分界而不是理解；
轉傳並未辨別是假是真。

這是資訊最爆炸的年代，
也是資訊最黑暗的年代，
不是網的網，
提供人們，
最缺乏資訊的資訊。

Life finds the way

- 這不是個預言，只是指出可能將要面對的問題
- 人類社會是個有機體，應該會找到適應與面對Internet Social network的方法
  - Life finds the way.
  - 資訊爆炸，來得快去的也快，歷史的教訓，忘的可能也很快
- 網路霸凌也是一個更早更需要面對的問題
  - 鍵盤真的會殺人

Thank you

網路多酸民
正義會殺人
風狂得自在
心靜幡無痕