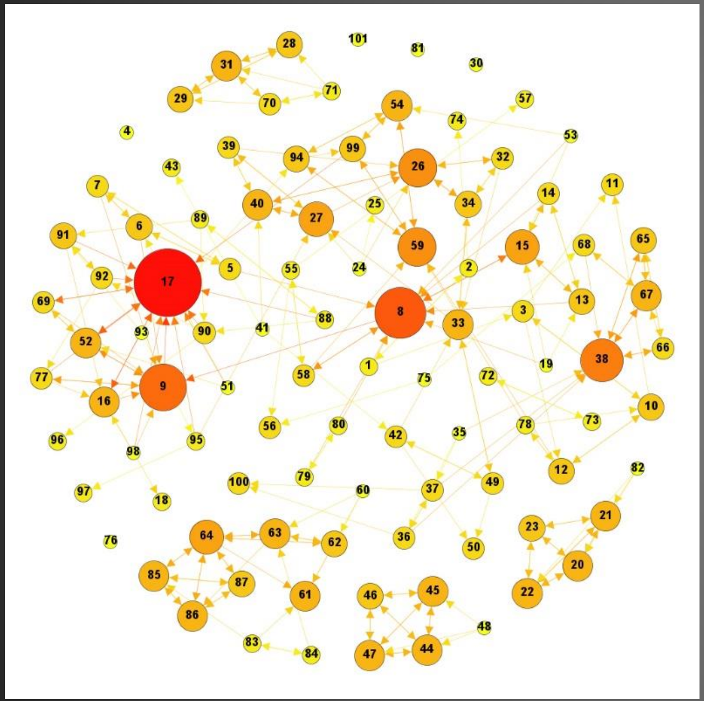
Chapter 4
Degree Correlations
& Community Structure

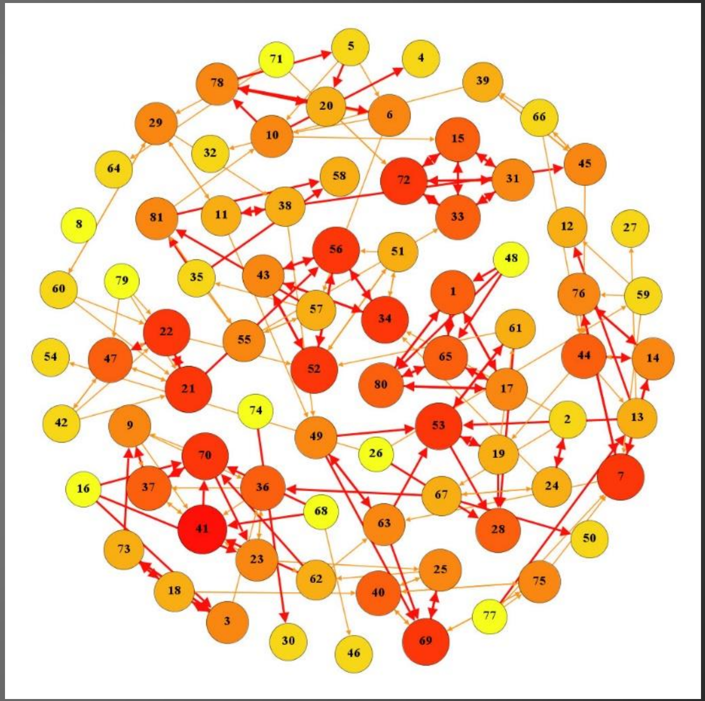
Xiaofan Wang
xfwang@sjtu.edu.cn

Your Directed Network

节点数 101；边数 242



节点数 81；边数 204

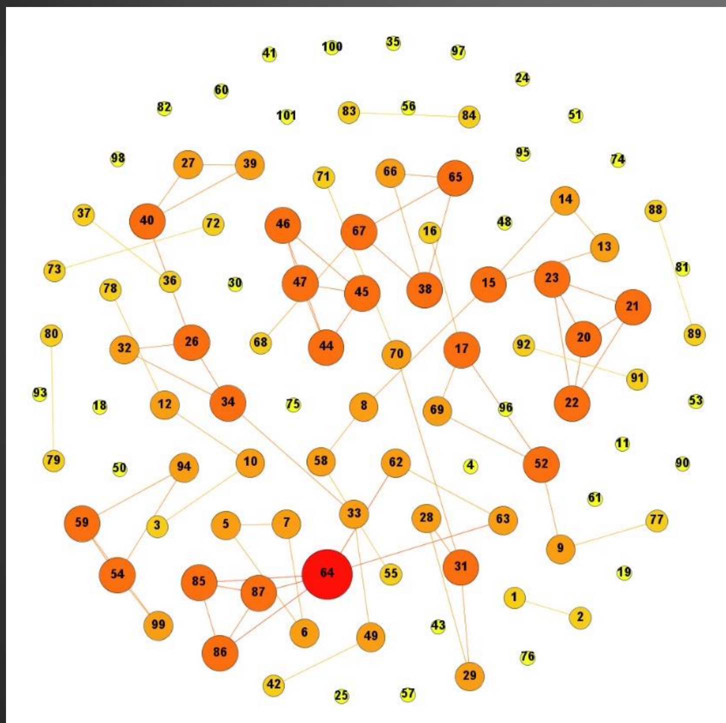


相同节点 72； 相同边数 106

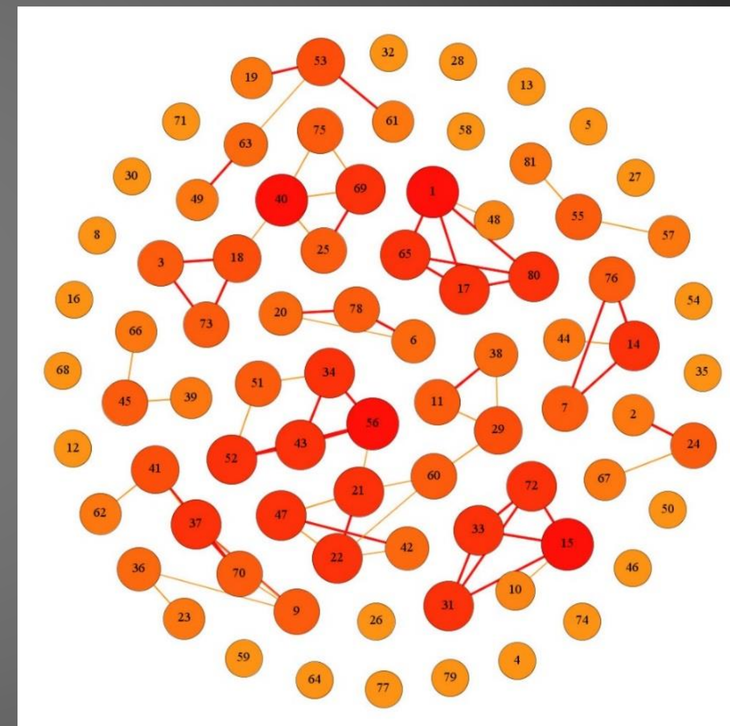
Xiaofan Wang

Your Undirected Network

节点数 101；边数 144



节点数 81；边数 121



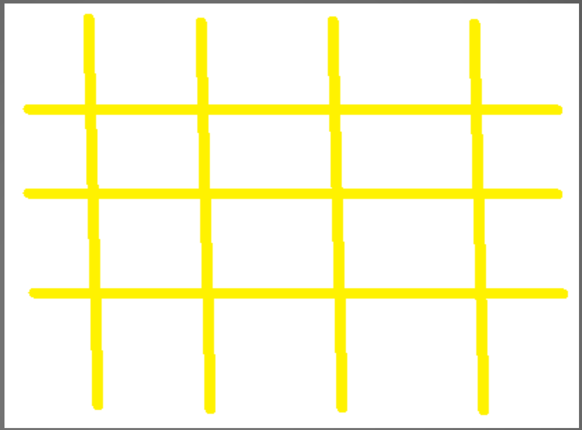
相同节点 72； 相同边数 73

  Xiaofan Wang

Quiz Q:
How can a network be constructed from these streets?

Nodes:

- **Street blocks**
- **Intersections**
- **Roads**

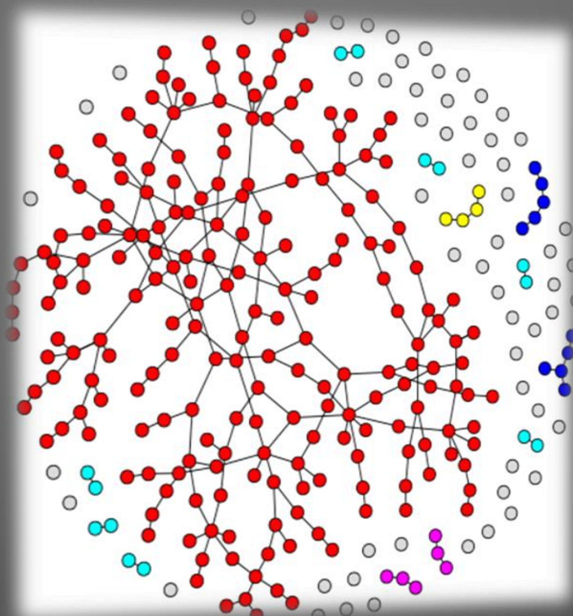


Edges: an edge is drawn between two nodes if they are

- adjacent
- directed connected by a segment of street with no intervening
- intersect

Connectivity Property of Real Networks

Many networks have a unique giant component

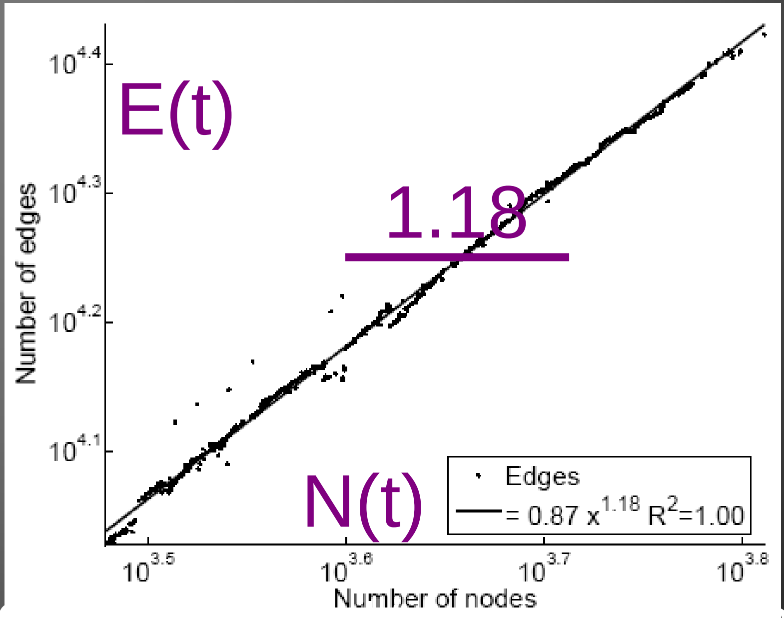


 Xiaofan Wang

Density Property of Real Networks

Many networks densify over time, but are still sparse

Autonomous Systems



 Xiaofan Wang

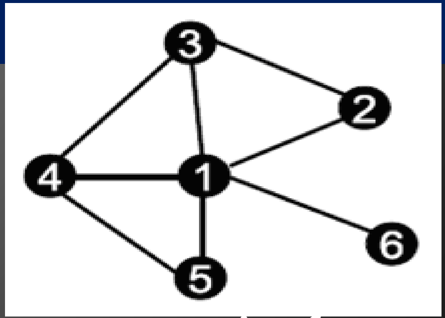
Small-World Property of Real Networks

- In most real networks, there are small distances between two randomly selected nodes.

Network Name	N	L	$\langle k \rangle$	$\langle d \rangle$	d_{max}	$\frac{\log N}{\log \langle k \rangle}$
Internet	192,244	609,066	6.34	6.98	26	6.59
WWW	325,729	1,497,134	4.60	11.27	93	8.32
Power Grid	4,941	6,594	2.67	18.99	46	8.66
Mobile Phone Calls	36,595	91,826	2.51	11.72	39	11.42
Email	57,194	103,731	1.81	5.88	18	18.4
Science Collaboration	23,133	186,936	8.08	5.35	15	4.81
Actor Network	212,250	3,054,278	28.78	-	-	-
Citation Network	449,673	4,707,958	10.47	11.21	42	5.55
E Coli Metabolism	1,039	5,802	5.84	2.98	8	4.04
Yeast Protein Interactions	2,018	2,930	2.90	5.61	14	7.14

Clustering Property of Real Networks

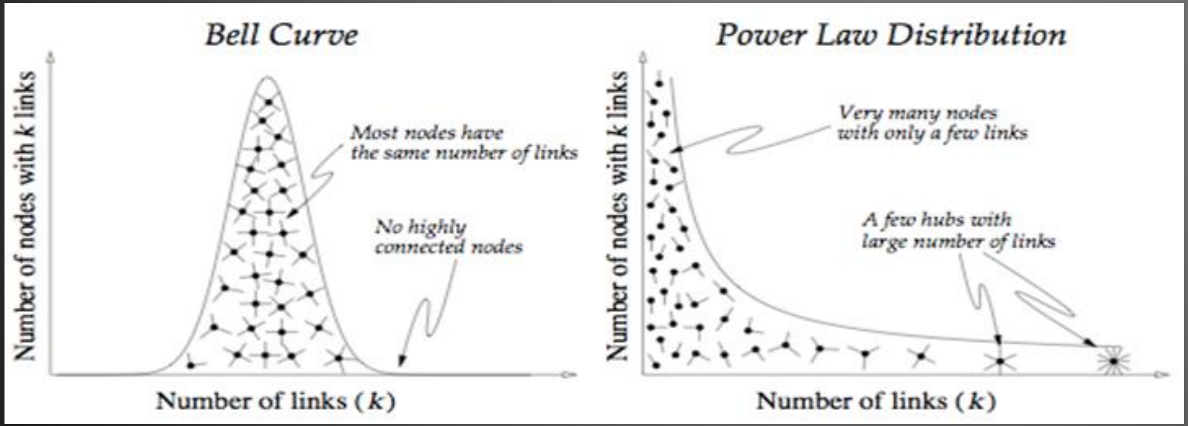
- Many real networks have a much higher clustering coefficient than expected for a completely random network of same no. of nodes and links
- High-degree nodes tend to have a smaller clustering coefficient than low-degree nodes.

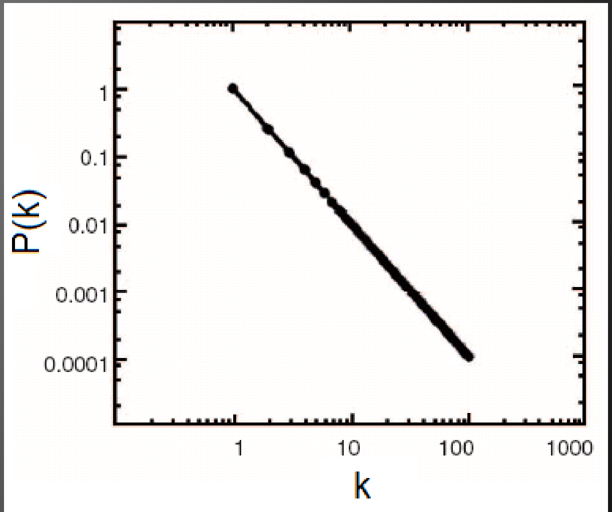


 Xiaofan Wang

Scale-Free Feature of Real Networks

- Many real networks are scale-free in the sense that the degree distribution deviates significantly from the Poisson distribution





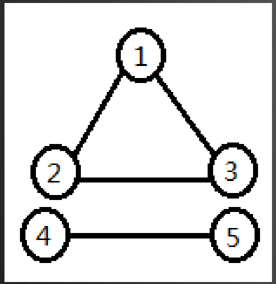
 Xiaofan Wang

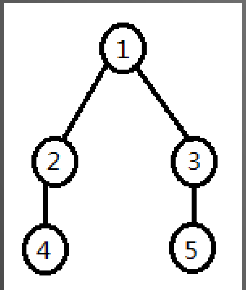
Towards High-Order Degree Distribution

Average Degree $<k>=2M/N$

Degree Distribution $P(k)=n(k)/N$ $\langle k\rangle=\sum_{k=0}^{\infty}kP(k)$

How many nodes have degree k? What's the maximum degree?





We need more properties to characterize a network

Ex. Whom do u contact with?

 Xiaofan Wang

Joint Probability Distribution

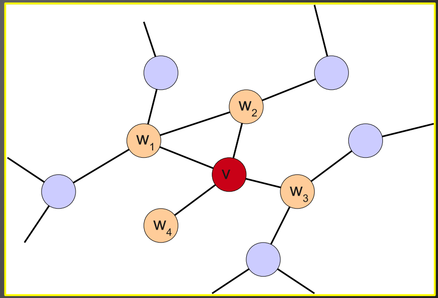
Average Degree $$<k>=2M/N$$

Degree Distribution $$P(k)=n(k)/N \qquad \langle k\rangle=\sum_{k=0}^{\infty}kP(k)$$

Joint Probability Distribution: prob. to find a node with degree j and k at the two ends of a randomly selected link

$$P(j,k)=\frac{m(j,k)}{2M} \quad (j\neq k) \qquad P(j,j)=\frac{m(j,j)}{M}$$

$$P(j,k)=\frac{m(j,k)\mu(j,k)}{2M}$$



 Xiaofan Wang

Degree Correlation

$$P(j,k)=P(k,j) \qquad \sum_{j,k=k_{\min}}^{k_{\max}} P(j,k)=1 \qquad P_n(k)=\sum_{j=k_{\min}}^{k_{\max}} P(j,k)$$

Excess Degree Dis. $P_n(k)$: prob. to have a degree k node at the end of a link

$$p_k \triangleq P(k) \qquad q_k \triangleq P_n(k) \qquad e_{jk} \triangleq P(j,k)$$

$$p_k=\frac{\langle k\rangle}{k}\sum_{j=k_{\min}}^{k_{\max}} e_{jk}=\frac{\langle k\rangle}{k}q_k$$

If the network has no degree correlations (neutral):

$$e_{jk}=q_j q_k$$

M. E. J. Newman, Phys. Rev. Lett. 89, 208701 (2002)

 Xiaofan Wang

Assortativity & Disassortativity

Assortative: hubs tend to link to each other.

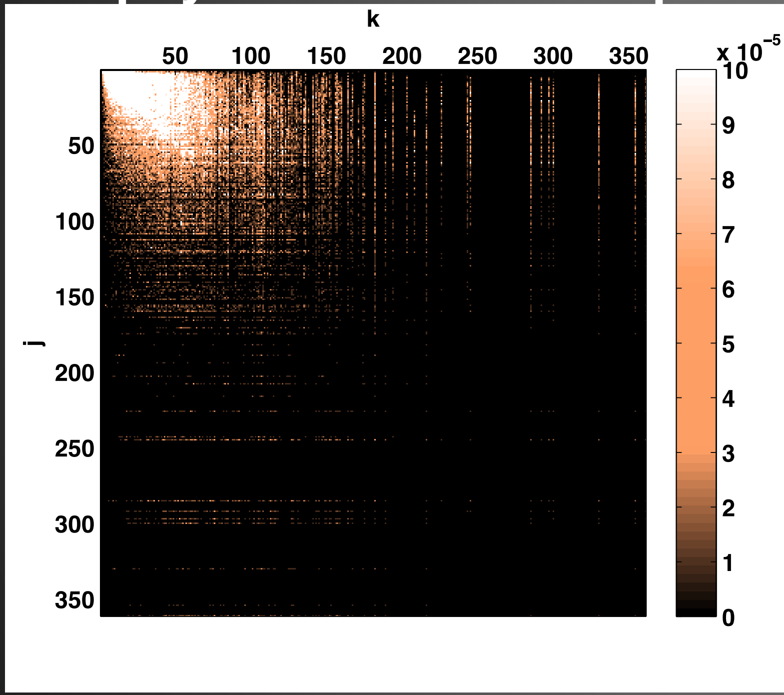
Neutral

Disassortative: Hubs tend to avoid linking to each other.

 Xiaofan Wang

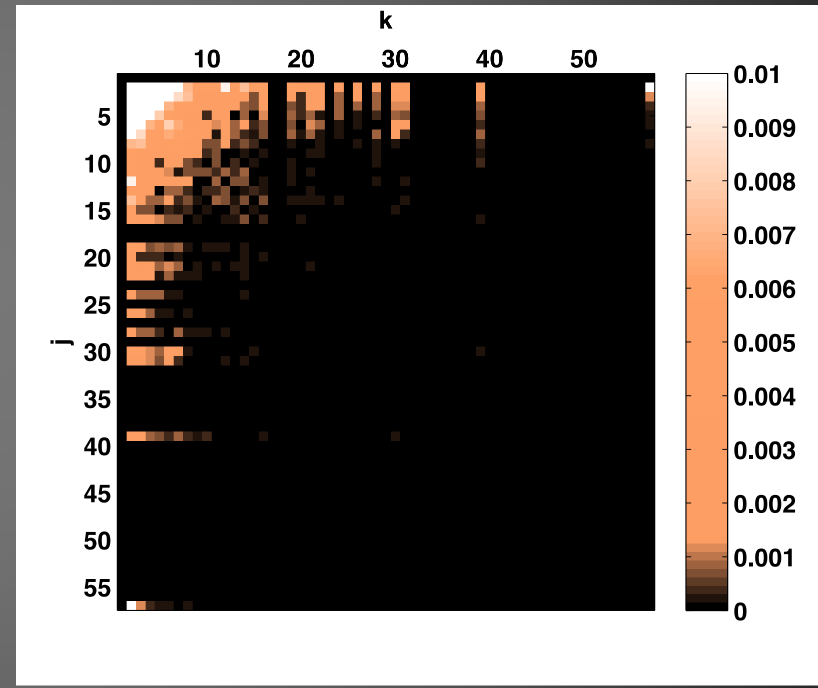
Assortativity: Full Statistical Description

Astrophysics co-authorship network



e_{jk}

Yeast PPI



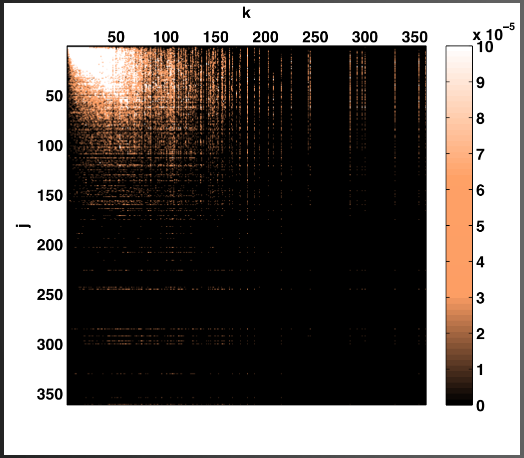
Assortative

Disassortative

 Xiaofan Wang

Problem with Full Statistical Description

(1) Difficult to extract information from a visual inspection of a matrix.



(2) Requires a large number of elements to inspect:

$$\frac{k_{\max}\left(k_{\max}-1\right)}{2}-1-k_{\max}$$

Nr. of independent elements

Undirected network: k_{max} x k_{max} matrix

$$\sum_{j,k} e_{jk} = 1 \qquad \sum_{j=1,k_{\max}} e_{jk} = q_k$$

Constraints

We need to find a way to reduce the information contained in e_{jk}

M. E. J. Newman, Phys. Rev. Lett. 89, 208701 (2002)

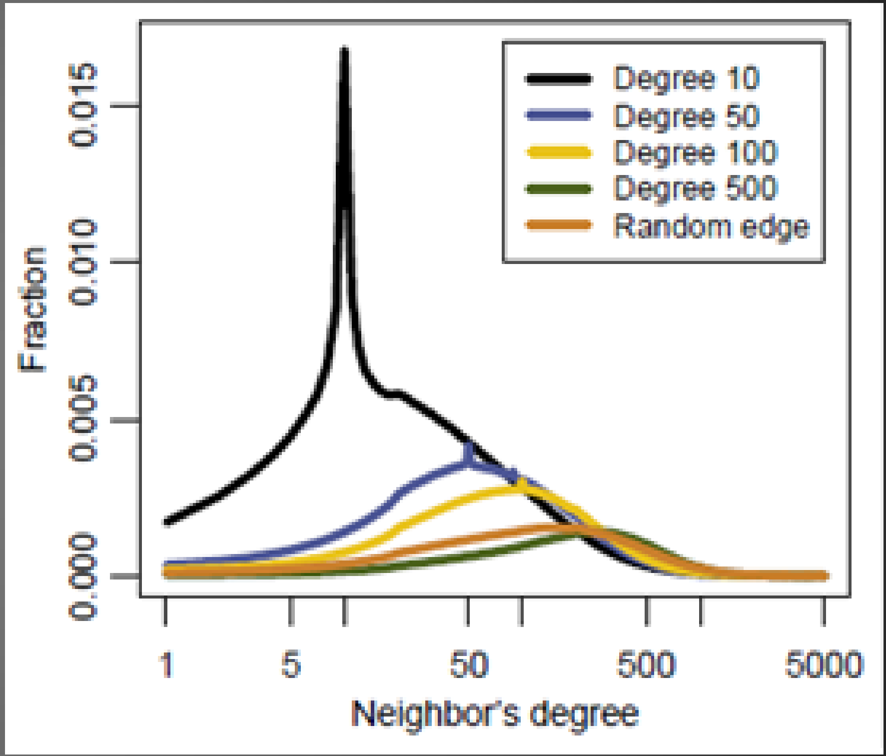
 Xiaofan Wang

Conditional Probability Distribution

$$P_c(k'|k) = \frac{P(k',k)}{P_n(k)}$$

If the network is neutral:

$$P_c(k'|k) = P_n(k') = \frac{k'P(k')}{\langle k \rangle}$$



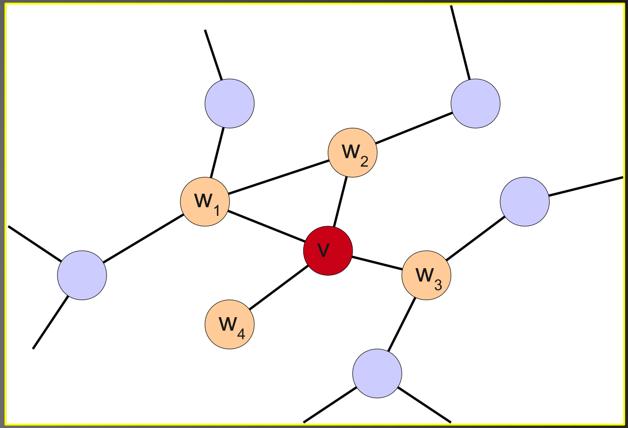
Facebook

 Xiaofan Wang

Excess Average Degree

Average Next Neighbor Degree

$$\langle k_{nn}\rangle_i = \frac{1}{k_i}\sum_{j=1}^{k_i} k_{i_j} \qquad \langle k_{nn}\rangle(k) = \frac{1}{i_k}\sum_{i=1}^{i_k}\langle k_{nn}\rangle_{v_i}$$



$$\langle k_{nn}\rangle(k) = \sum_{k'} k' P(k' \mid k) = \frac{1}{q_k}\sum_{k'} k' e_{kk'}$$

If the network is neutral, $k_{nn}(k)$ is independent of k:

$$\langle k_{nn}\rangle_v = \frac{4+3+3+1}{4}$$

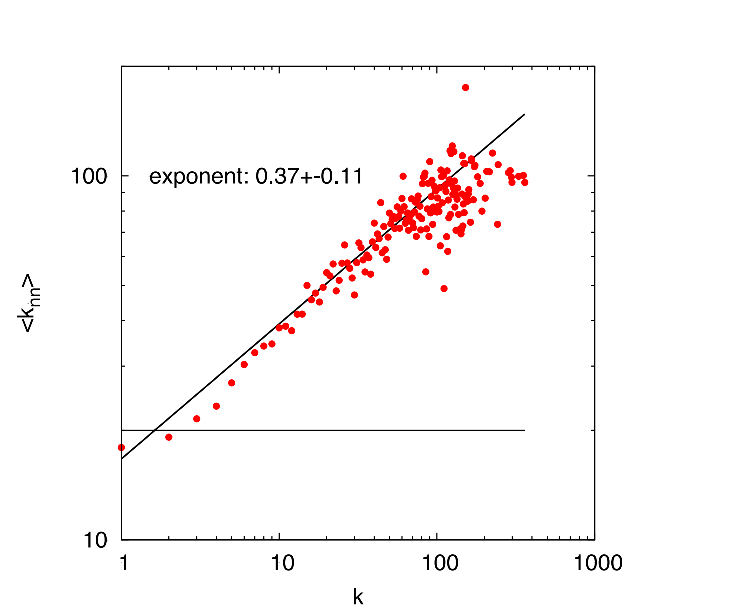
$$k_{nn}(k) = \frac{\sum_{k'} k' q_k q_{k'}}{q_k} = \sum_{k'} k' q_{k'} = \sum_{k'} k' \frac{k' p(k')}{<k>} = \frac{\langle k^2\rangle}{\langle k\rangle}$$

R. Pastor-Satorras, A. Vázquez, A. Vespignani, Phys. Rev. E 65, 066130 (2001)

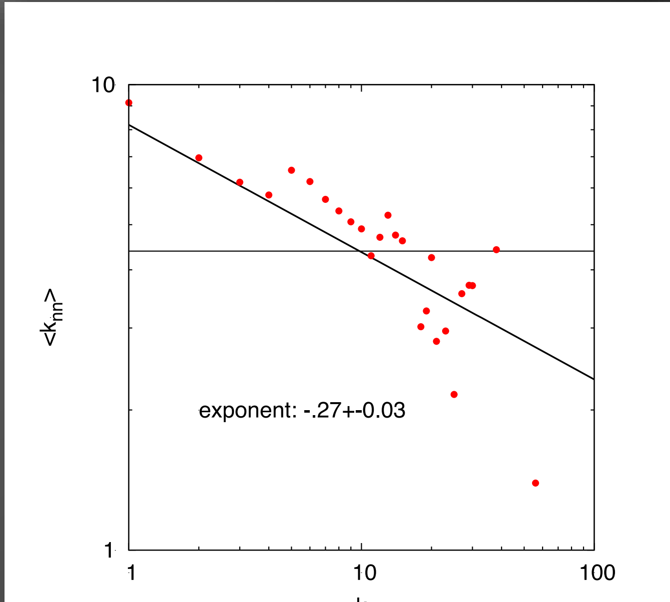
 Xiaofan Wang

Assotativity: Excess Average Degree

$$\langle k_{nn} \rangle (k)$$



Astrophysics co-authorship network

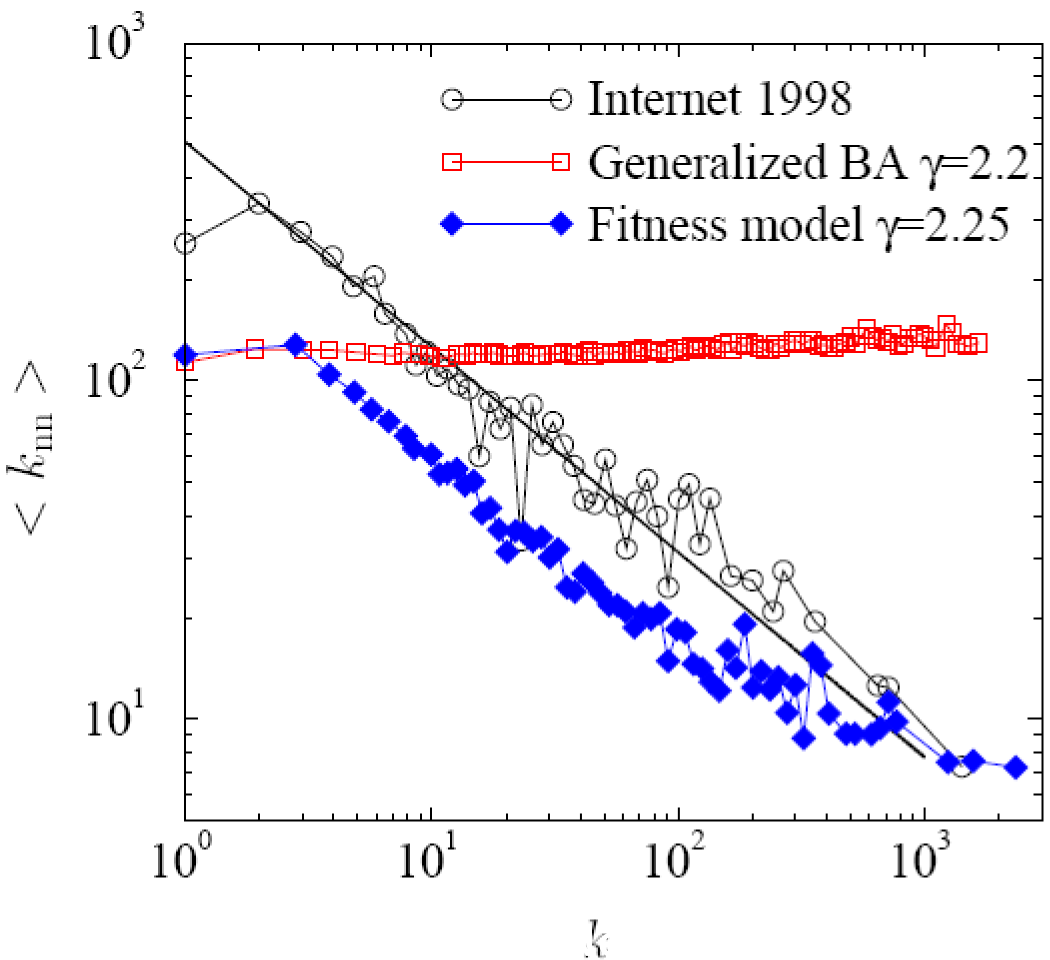


Yeast PPI

Assortative **Disassortative**

Dynamical and correlation properties of the Internet

Romualdo Pastor-Satorras,[1] Alexei Vázquez,[2] and Alessandro Vespignani[3]

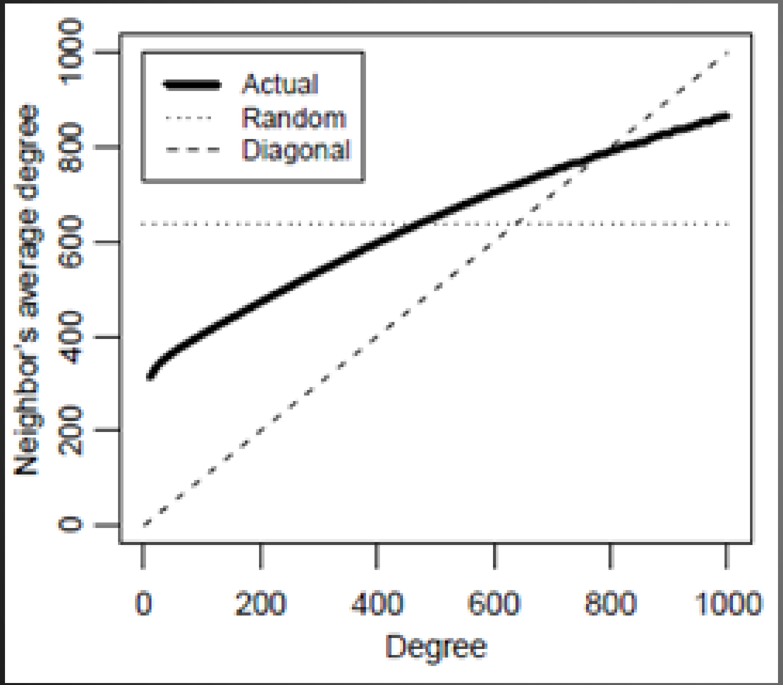


Phys. Rev. Lett. 87, 258701 (2001)

Facebook

Why your friends have more friends than you do?

$\langle k_{nn} \rangle(k)$

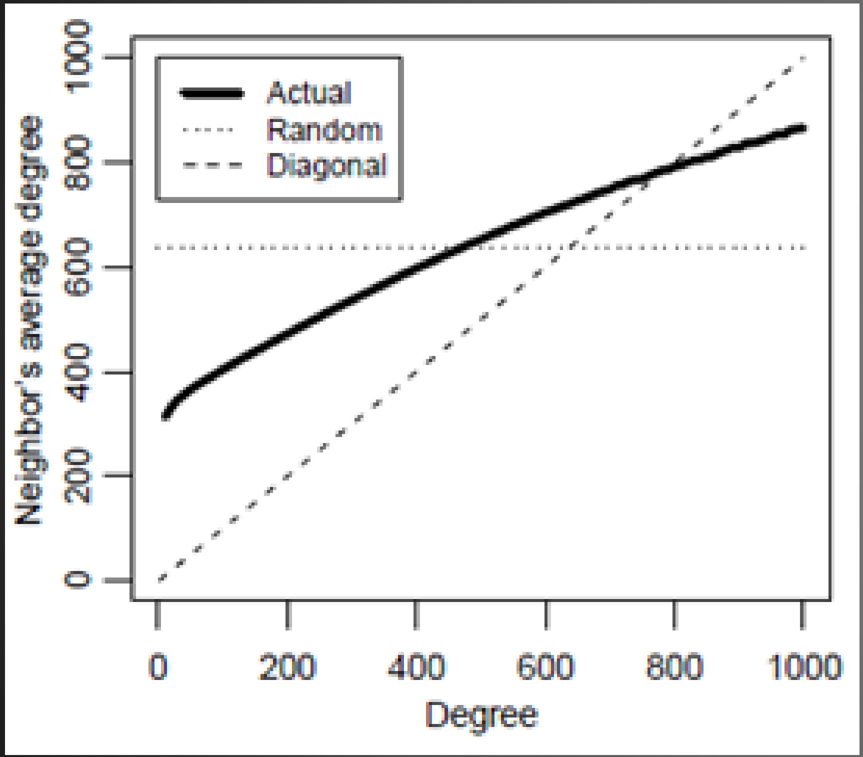


- ◆平均而言，除非你的朋友数超过700，否则你的朋友比你拥有更多的朋友!
- ◆约92%的Facebook用户的朋友数都小于700
- ◆绝大部分用户都会觉得自己的朋友比自己拥有更多的朋友!

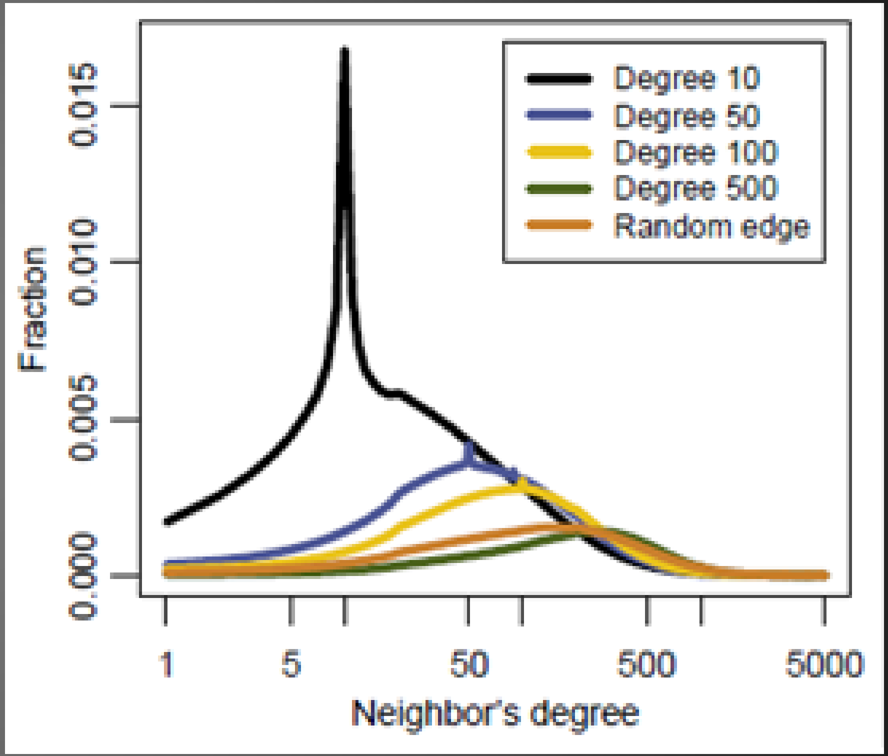
友谊悖论（Friendship paradox）！

Facebook

$\langle k_{nn} \rangle (k)$



$P_c(k'|k)$



<k_{nn}>(k) is a *k*-dependent function, hence it has much fewer parameters,

and it is easier to interpret/read.

 Xiaofan Wang

Pearson-Correlation Coefficient

$$q_k \triangleq P_n(k) \qquad e_{jk} \triangleq P(j,k)$$

If there are degree correlations, e_{jk} will differ from $q_j q_k$.

The magnitude of the correlation is captured by

$$\langle jk \rangle - \langle j \rangle \langle k \rangle = \sum_{j,k} jk(e_{jk} - q_j q_k)$$

- *positive* for *assortative* networks,
- *zero* for *neutral* networks,
- *negative* for *dissasortative* networks

To compare different networks, we should normalize it

M. E. J. Newman, Phys. Rev. Lett. 89, 208701 (2002)

 Xiaofan Wang

Pearson-Correlation Coefficient

$$\langle jk \rangle - \langle j \rangle \langle k \rangle = \sum_{j,k} jk(e_{jk} - q_j q_k)$$

normalize it with its maximum value; the maximum is reached for a *perfectly assortative network*, i.e. $e_{jk}=q_k\delta_{jk}$

$$\sigma_q^2 = \max \sum_{jk} jk(e_{jk} - q_j q_k) = \sum_{jk} jk(q_k \delta_{jk} - q_j q_k) = \sum_k k^2 q_k^2 - \left[\sum_k k q_k\right]^2$$

$$r = \frac{\sum_{jk} jk(e_{jk} - q_j q_k)}{\sigma_q^2}$$

$$-1 \le r \le 1$$

$r \le 0$ *disassortative*

$r = 0$ *neutral*

$r \ge 0$ *assortative*

M. E. J. Newman, Phys. Rev. Lett. 89, 208701 (2002)

 Xiaofan Wang

VOLUME 89, NUMBER 20 PHYSICAL REVIEW LETTERS 11 NOVEMBER 2002

Assortative Mixing in Networks

M. E. J. Newman

Department of Physics, University of Michigan, Ann Arbor, MI 48109–1120
Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, New Mexico 87501

(Received 20 May 2002; published 28 October 2002)

A network is said to show assortative mixing if the nodes in the network that have many connections tend to be connected to other nodes with many connections. Here we measure mixing patterns in a variety of networks and find that social networks are mostly assortatively mixed, but that technological and biological networks tend to be disassortative. We propose a model of an assortatively mixed network, which we study both analytically and numerically. Within this model we find that networks percolate more easily if they are assortative and that they are also more robust to vertex removal.

the normalized correlation function is

$$r = \frac{1}{\sigma_q^2} \sum_{jk} jk(e_{jk} - q_j q_k), \tag{3}$$

which is simply the Pearson correlation coefficient of the degrees at either ends of an edge and lies in the range $-1 \leq r \leq 1$. For the practical purpose of evaluating r on

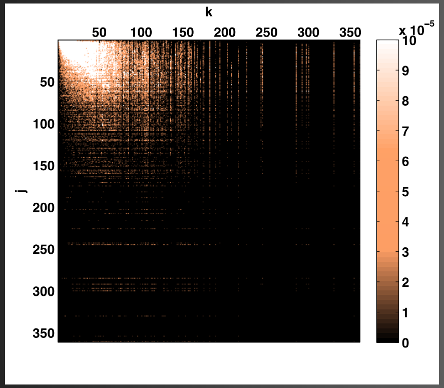
Social networks are *assortative*

Network	n	r
Physics coauthorship (a)	52 909	0.363
Biology coauthorship (a)	1 520 251	0.127
Mathematics coauthorship (b)	253 339	0.120
Film actor collaborations (c)	449 913	0.208
Company directors (d)	7 673	0.276
Internet (e)	10 697	−0.189
World-Wide Web (f)	269 504	−0.065
Protein interactions (g)	2 115	−0.156
Neural network (h)	307	−0.163
Marine food web (i)	134	−0.247
Freshwater food web (j)	92	−0.276
Random graph (u)		0
Callaway *et al.* (v)		$\delta/(1+2\delta)$
Barabási and Albert (w)		0

Biological, technological networks are *disassortative*

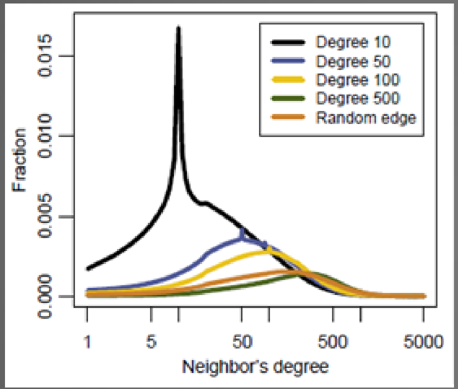
Summary: Degree Correlation

Joint Probability Distribution



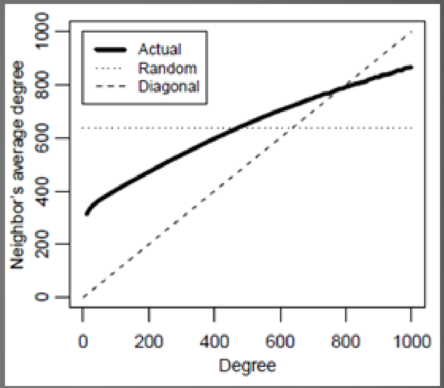
e_{jk}

Conditional Probability Distribution



$P_c(k'|k)$

Excess Average Degree



$\langle k_{nn} \rangle(k)$

Assortativity Coefficient

$$r = \frac{\sum_{jk} jk(e_{jk} - q_j q_k)}{\sigma_q^2}$$

 Xiaofan Wang

Quiz Q:

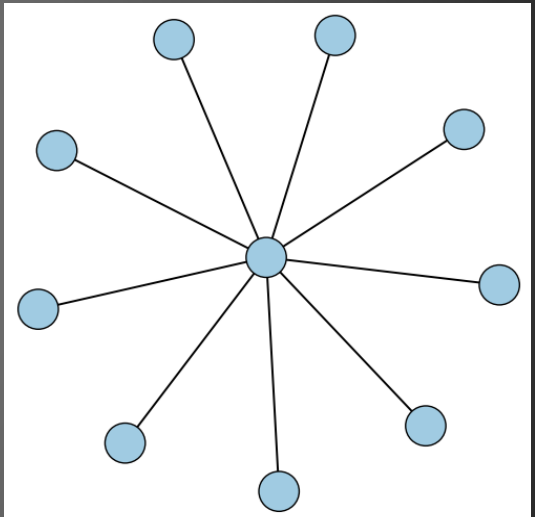
Assortativity of a Star Network (N>>1)

Degree Dis $p_k \triangleq P(k)$

Excess Degree Dis $q_k \triangleq P_n(k) = \frac{k}{<k>} p_k$

Joint Prob. Dis $e_{jk} \triangleq P(j,k)$

$$P(j,k) = \frac{m(j,k)}{2M} \quad (j \neq k)$$



Assortativity Coefficient

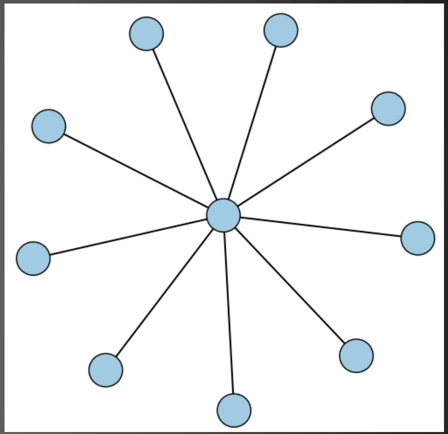
$$r = \frac{\sum_{jk} jk(e_{jk} - q_j q_k)}{\sigma_q^2}$$

 Xiaofan Wang

Quiz Q: Assortativity of a Star Network (N>>1)

Degree Dis

$$P(k)=\begin{cases}\frac{N-1}{N} & \text{for } k=1\\ \frac{1}{N} & \text{for } k=N-1\\ 0 & \text{otherwise}\end{cases}$$



Excess Degree Dis

$$\langle k\rangle=\sum_{k=1}^{\infty} kP(k)=1\times\frac{N-1}{N}+(N-1)\times\frac{1}{N}=2\frac{N-1}{N}\approx 2.$$

$$q(k)=\frac{k}{\langle k\rangle}P(k)$$

$$q(k)=\begin{cases}\frac{1}{2} & \text{for } k=N-1\\ \frac{1}{2} & \text{for } k=1\\ 0 & \text{otherwise}\end{cases}$$

Joint Prob. Dis

$$e_{1,N-1}=e_{N-1,1}=1/2.$$

Assortativity Coefficient

$$r=\frac{\sum_{jk} jk(e_{jk}-q_j q_k)}{\sigma_q^2}\rightarrow -1$$

 Xiaofan Wang

School integration and friendship segregation in America

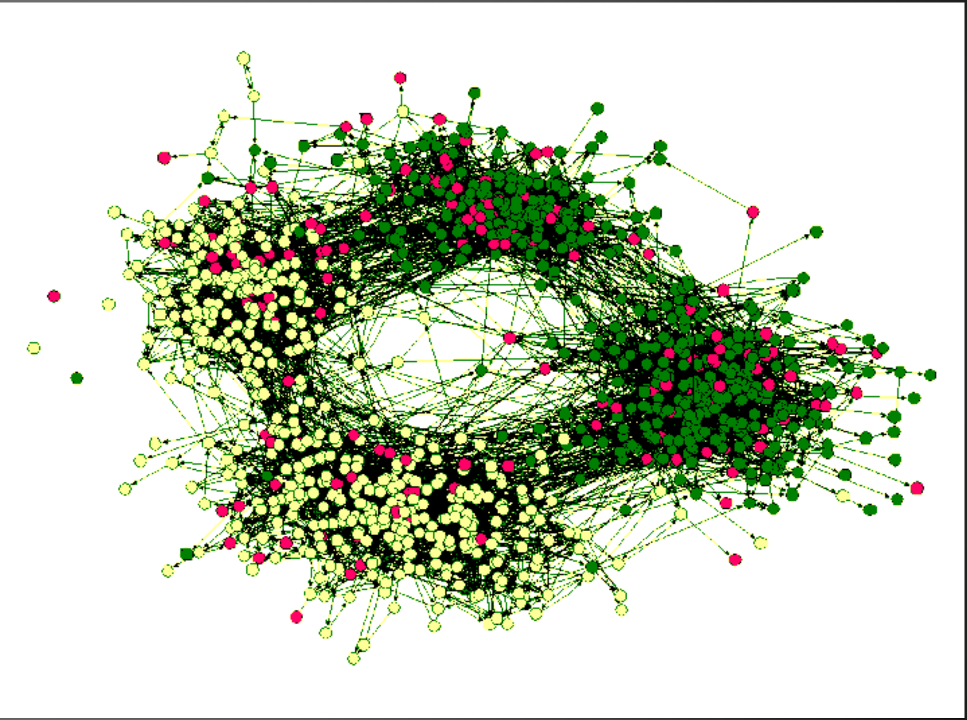
RACE M J. Amer. J. of Sociology, 2001

Race:

Left (white) to right (black)

Grade:

Up (junior) to down (senior)

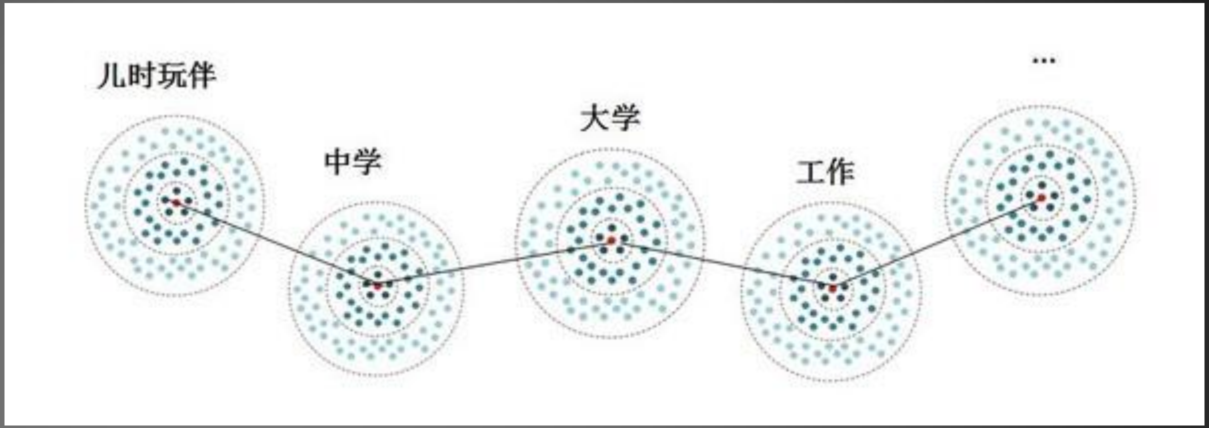


中学生朋友关系网络

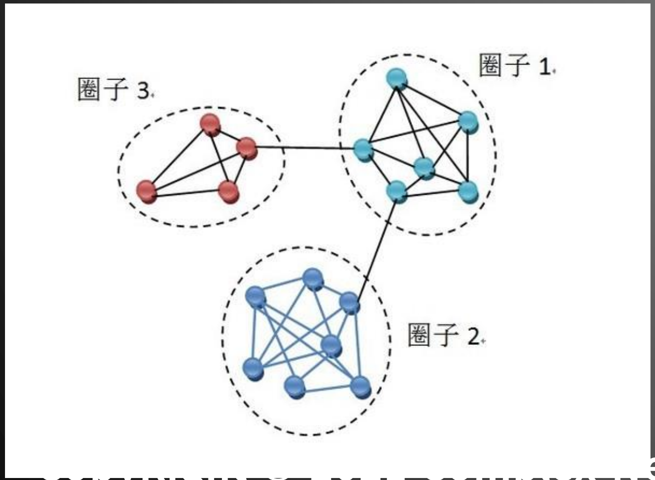
 Xiaofan Wang

QQ圈子

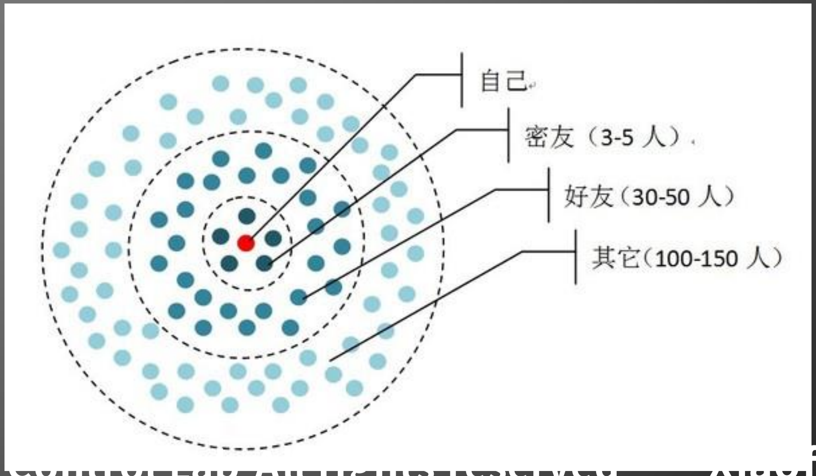
QQ圈子
社交关系
我是Q主好友
我是Q主
我和其他人都是
Q主好友的好友的...(拓宽社交关系)
PCHOME
WWW.PCHOME.NET



儿时玩伴
中学
大学
工作
...



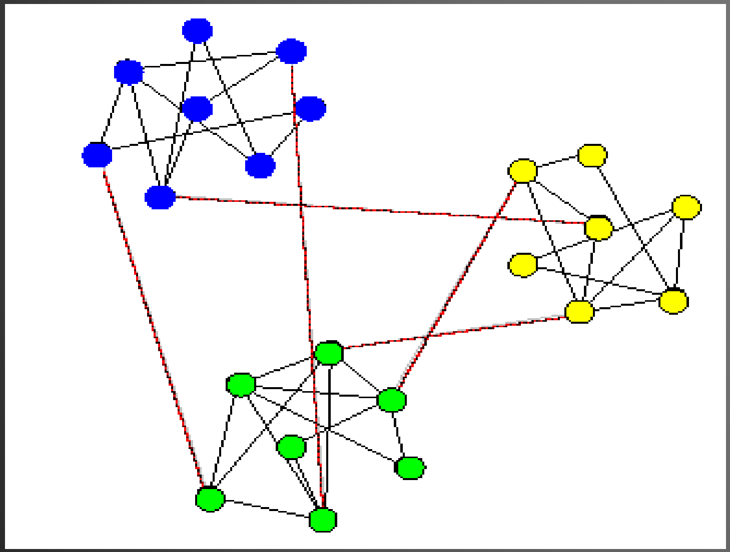
圈子 3
圈子 1
圈子 2

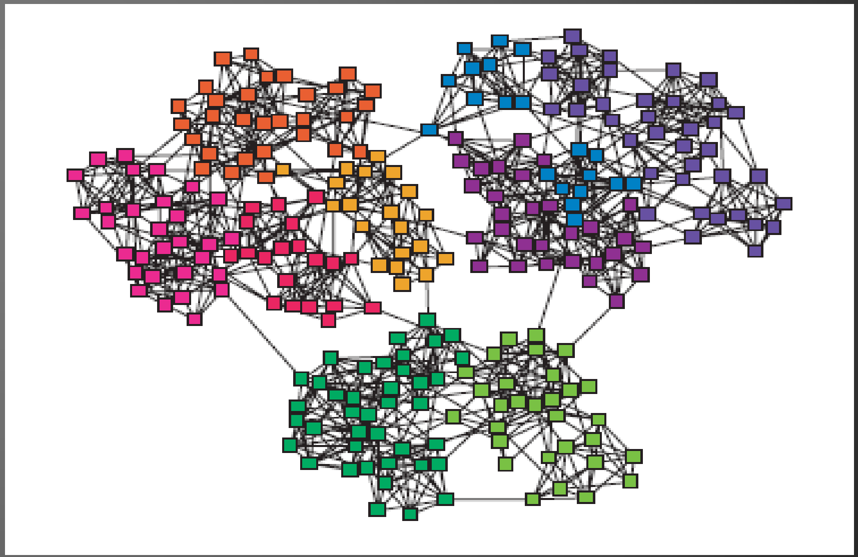


自己
密友（3-5 人）
好友（30-50 人）
其它（100-150 人）

Community Structure in Complex Networks

- **Communities:** groups of nodes between which links are sparse but within which links are dense.

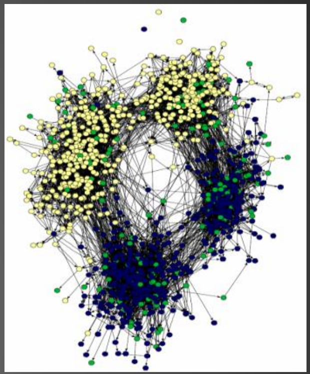


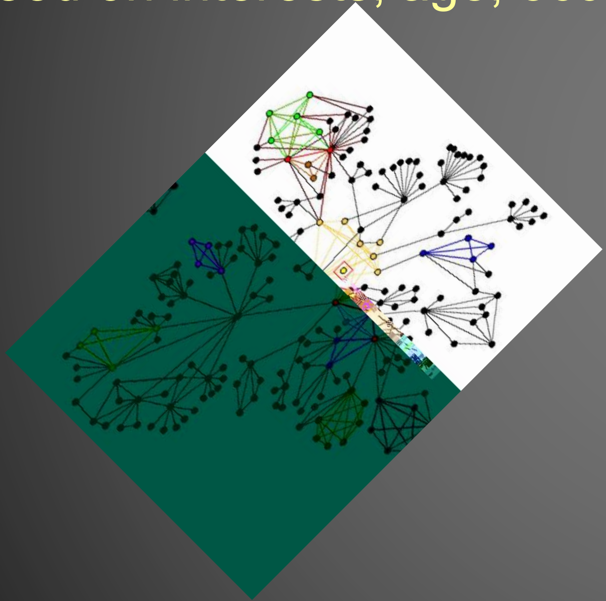


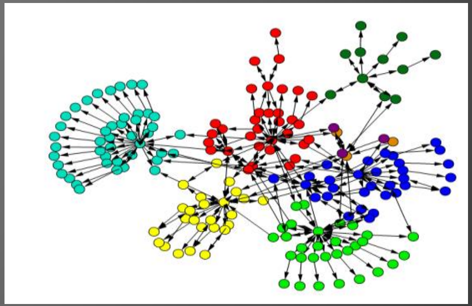
 Xiaofan Wang

Community Structure in Complex Networks

- Social networks: People naturally divide into groups based on interests, age, occupation,…







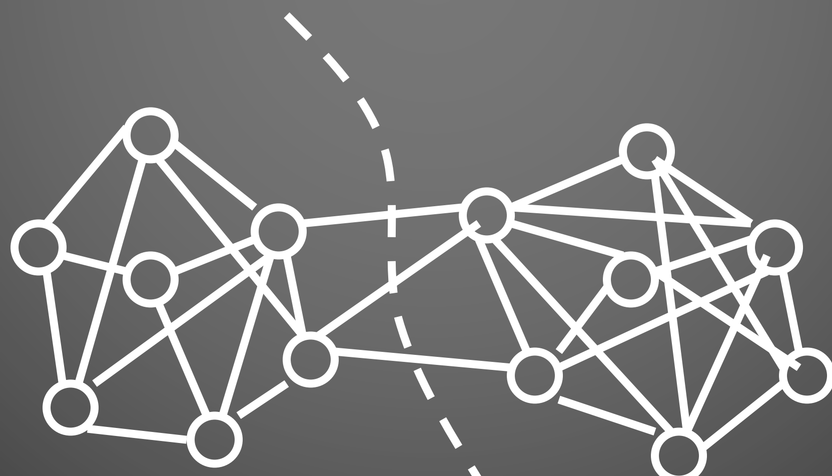
- WWW: subject matters of webpages

- Metabolic: functional units

 Xiaofan Wang

Community Detection

- Many real networks have a natural community structure
- We want to discover this structure rather than impose a certain size of community or fix the number of communities
- Without “looking”, can we discover community structure in an automated way?



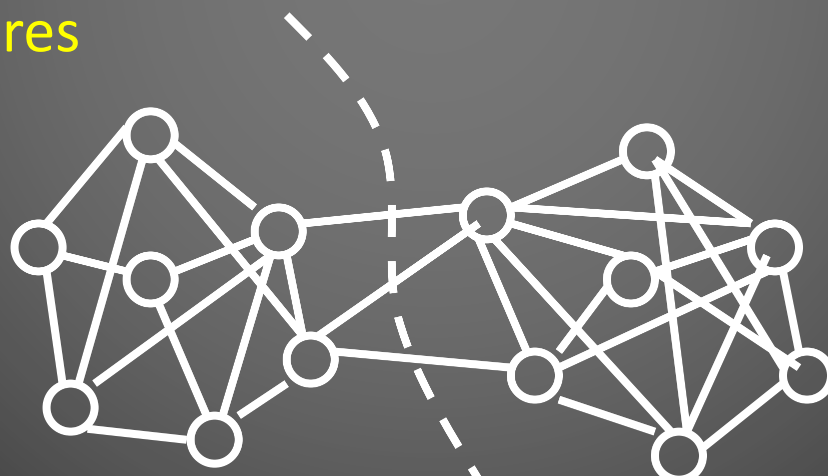
Community Detection: Which one is the best?

- Which is the best algorithm to characterize networks of known community structure?

  Benchmarks

- How to evaluate algorithm performance when the community structure is unknown?

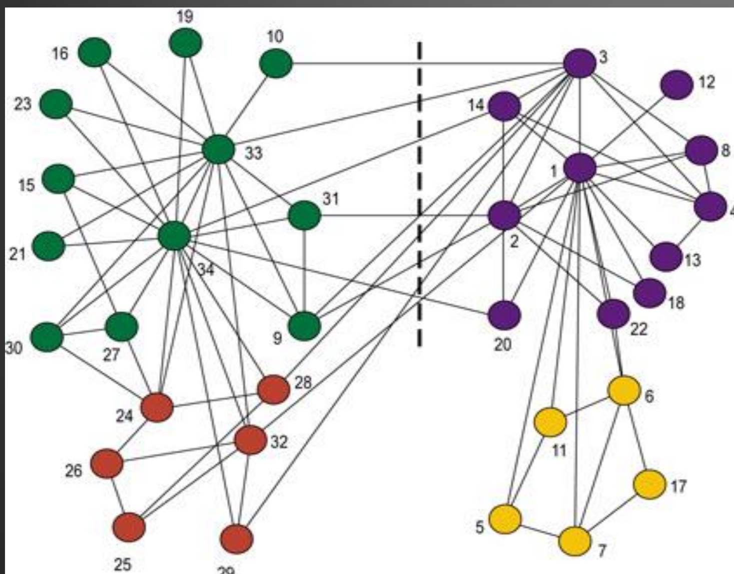
  Quality measures



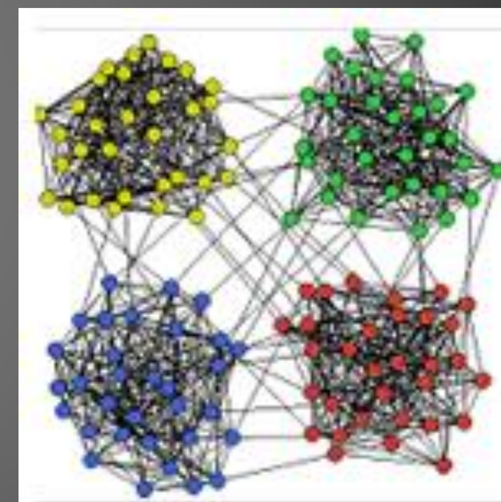
 Xiaofan Wang

Benchmark Examples

- An real network:
- Zachary's Karate Network

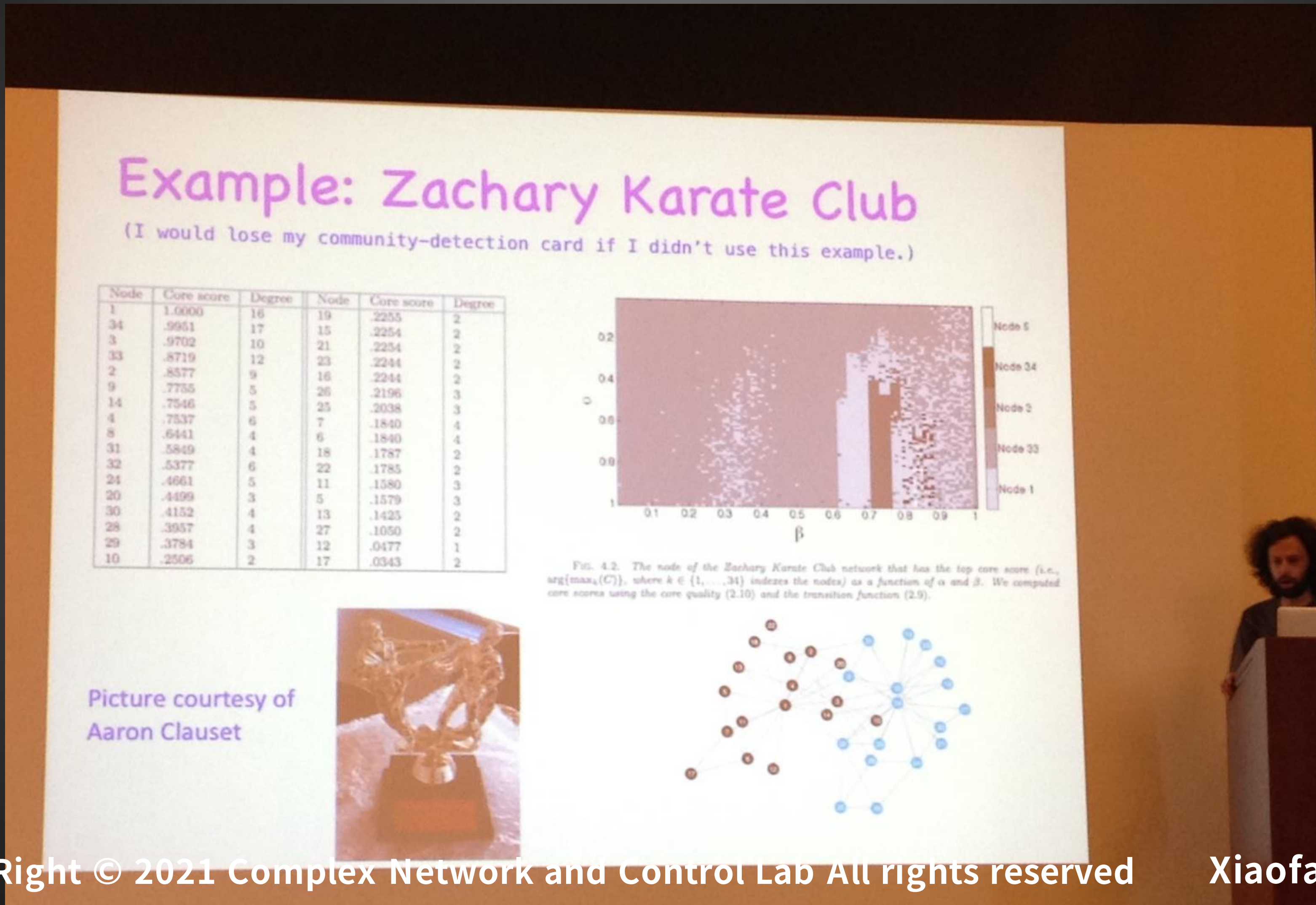


- An artificial network:
- Planted l-partition model



W. W. Zachary, *J. Anthropological Research* **33**, 452-473 (1977).

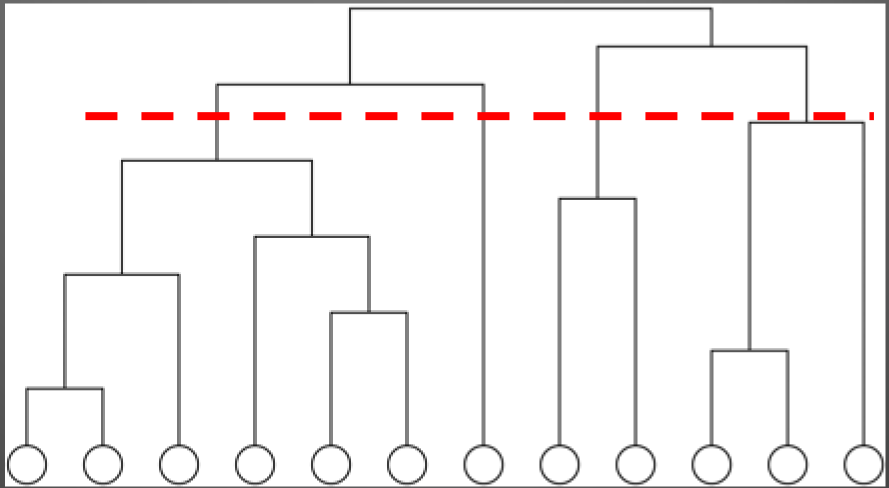
I would lose my community-detection card if I didn't use this example



Node	Core score	Degree	Node	Core score	Degree
1	1.0000	16	19	.2255	2
34	.9951	17	15	.2254	2
3	.9702	10	21	.2254	2
33	.8719	12	23	.2244	2
2	.8577	9	16	.2244	2
9	.7755	5	26	.2196	3
14	.7546	5	25	.2038	3
4	.7537	6	7	.1840	4
8	.6441	4	6	.1840	4
31	.5849	4	18	.1787	2
32	.5377	6	22	.1785	2
24	.4661	5	11	.1580	3
20	.4499	3	5	.1579	3
30	.4152	4	13	.1425	2
28	.3957	4	27	.1050	2
29	.3784	3	12	.0477	1
10	.2506	2	17	.0343	2

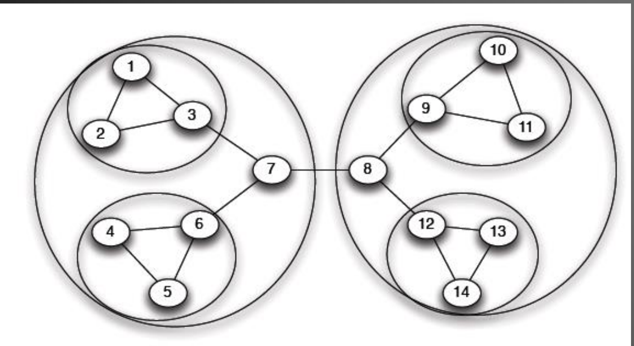
Hierarchical Clustering

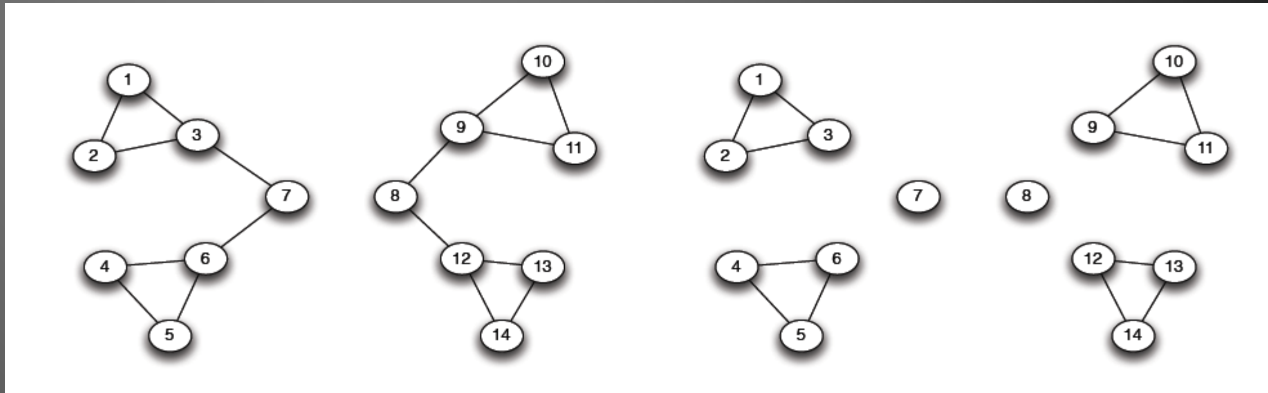
- Calculate a "weight" (e.g., no. of node-independent paths) for every pair of vertices, which represents how closely connected the vertices are
- Start with all n vertices disconnected, add edges between pairs one by one in order of decreasing weight
- Result: nested components, where one can take a 'slice' at any level of the tree



Betweenness Clustering

1. Calculate the betweenness for all edges in the network.
2. Remove the edge with the highest betweenness.
3. Recalculate betweennesses for all edges affected by the removal. (very expensive)
4. Repeat from step 2 until no edges remain.



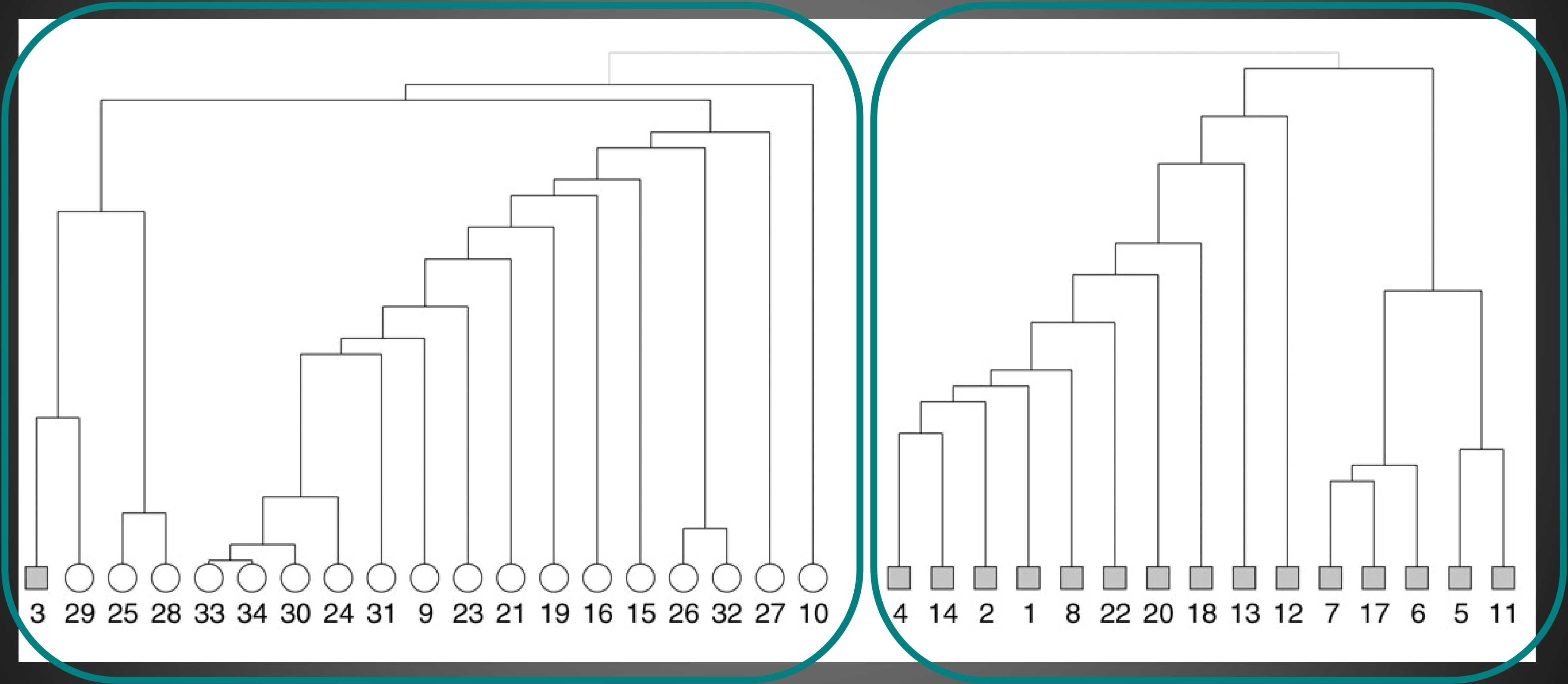


(a) *Step 1* (b) *Step 2*

Girvan & Newman, PNAS June 11, 2002 99(12):7821-7826

 Xiaofan Wang

Betweenness clustering algorithm & the karate club data set

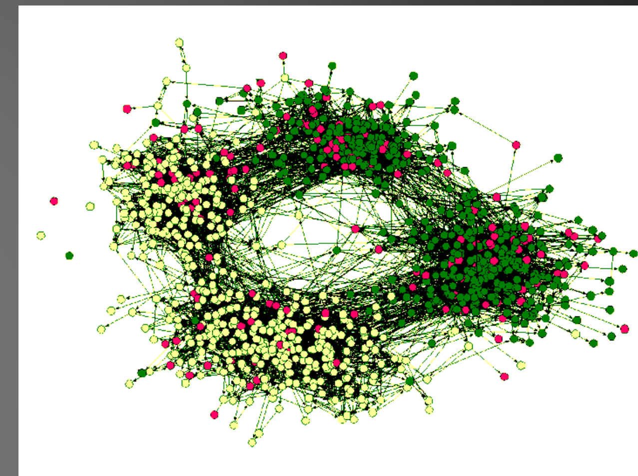


Girvan & Newman, PNAS June 11, 2002 99(12):7821-7826

 Xiaofan Wang

Modularity

$$Q = \frac{Q_{real} - Q_{null}}{M}$$



$$Q_{real} = \frac{1}{2}\sum_{ij} a_{ij}\,\delta(C_i, C_j)$$

$$Q_{null} = \frac{1}{2}\sum_{ij} p_{ij}\,\delta(C_i, C_j)$$

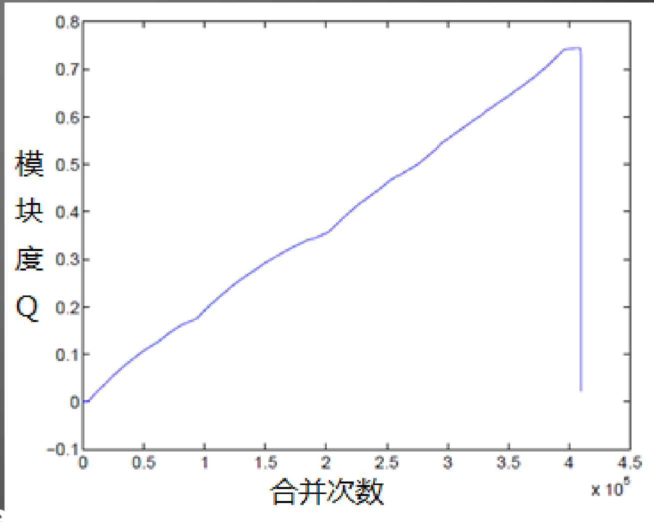
$$p_{ij} = k_i k_j / (2M)$$

Compare to a null model with the same degree distribution

A. Clauset, M. E. J. Newman, & C. Moore, Finding community structure in very large networks, Phys. Rev. E 70, 066111 (2004).

Community Detection based on Modularity Optimization

- Start with all vertices as isolates
- follow a greedy strategy (or simulated annealing…) :
  - successively join clusters with the greatest increase ΔQ in modularity
  - stop when the maximum possible ΔQ <= 0 from joining any two

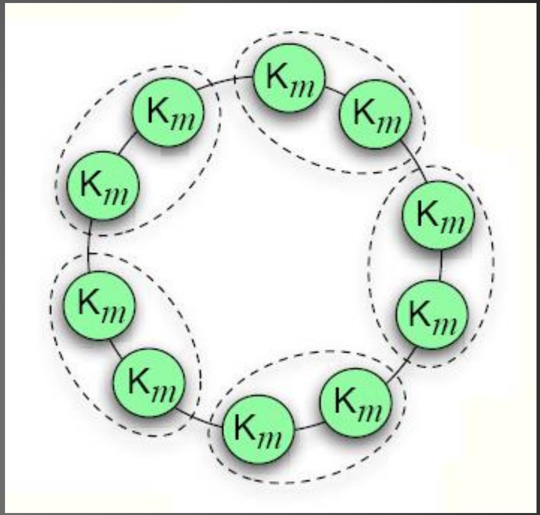


Resolution Limit in Community Detection

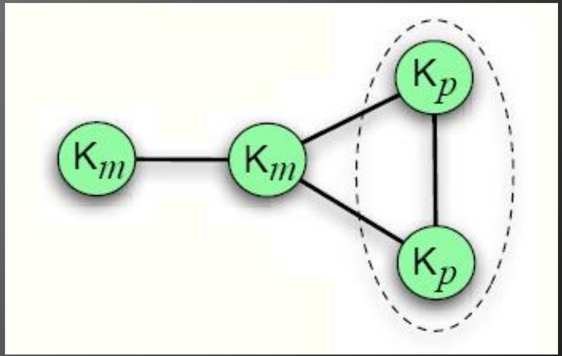
$$Q_{single} = 1 - \frac{2}{m(m-1)+2} - \frac{1}{n}$$

$$Q_{pairs} = 1 - \frac{1}{m(m-1)+2} - \frac{2}{n}$$

$$Q_{single} > Q_{pairs} \Leftrightarrow m(m-1)+2 > n$$

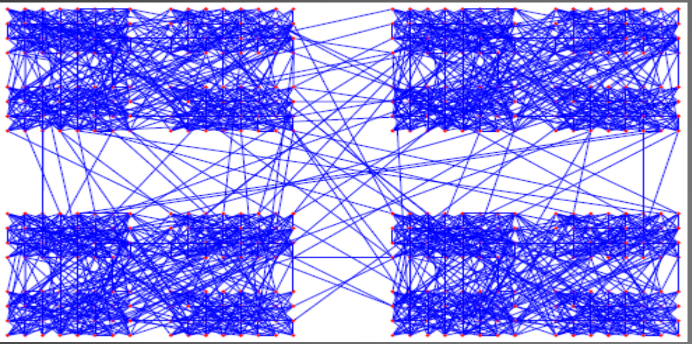


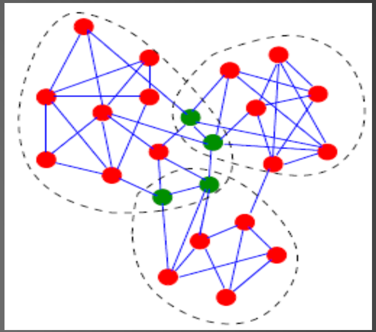
p=5, m=20. The maximal Q corresponds to the partition in which the two smaller cliques are merged



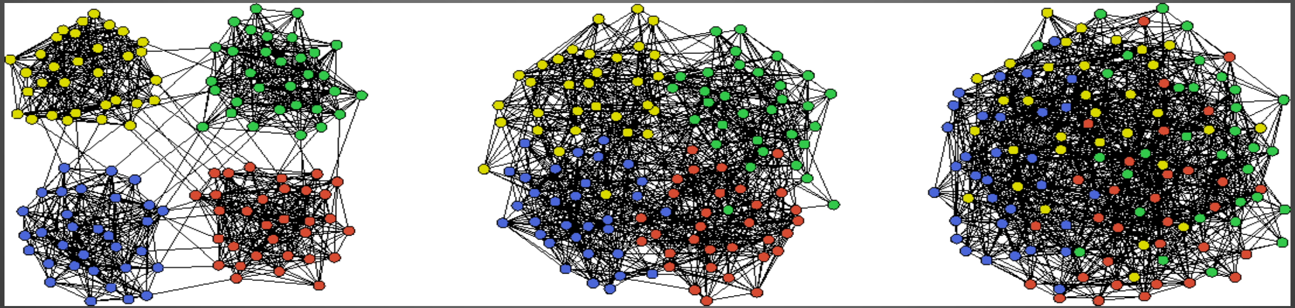
FORTUNATO S, BARTHÉLEMY M. Proc. Natl. Acad. Sci. USA, 2007, 104(1): 36-41.

Detecting Community Structure: Challenges



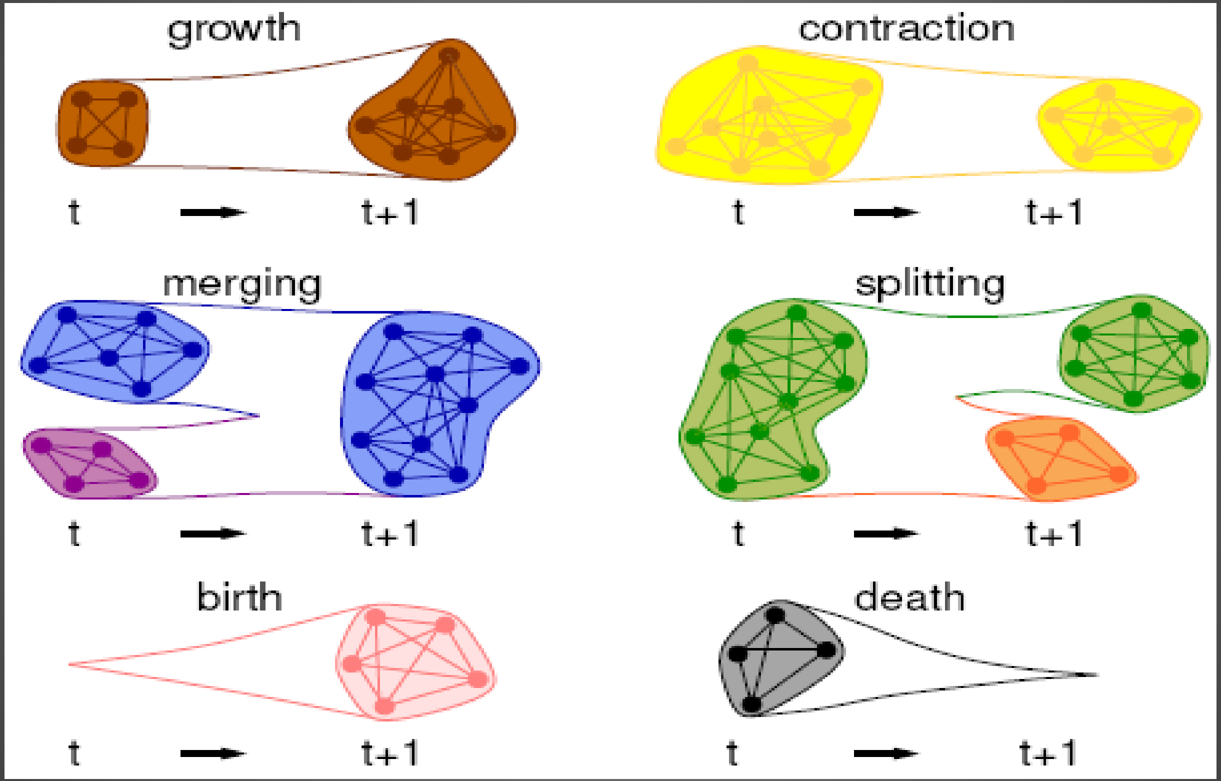


- Hierarchical
- Overlapping



 Xiaofan Wang

Detecting Community Structure: Challenges



- Evolution, Emergence

 Xiaofan Wang

网络拓扑性质分析

网络有多大？连通性

网络有多小？平均距离

网络有多紧？聚类系数

网络有多密？连边密度

网络有多匀？度分布

网络有多配？同配性

网络有多分？社团结构

思考题：
情绪传染 Emotion Contagion

- 假设你是人人的研究人员，你可以经公司允许在人人上做实验以验证情绪是如何在人们之间传播的。
- 例如：如果一个人看到更多正面或者负面的帖子，是否自己也会变得更为正面或者负面？
- 请问你应该如何设计实验？

A. D. I. Kramer et al., Experimental evidence of massive-scale emotional contagion through social networks, PNAS, 111(24), 2014

- *We show, via a massive (N = 689,003) experiment on Facebook, that emotional states can be transferred to others via emotional contagion, leading people to experience the same emotions without their awareness.*
- *We provide experimental evidence that emotional contagion occurs without direct interaction between people (exposure to a friend expressing an emotion is sufficient), and in the complete absence of nonverbal cues.*

How Does Facebook Choose What To Show In News Feed?

News Feed Visibility =* I x P x C x T x R

(Interest x Post x Creator x Type x Recency)

Interest: Interest of the user in the creator

Post: This post's performance amongst other users

Creator: Performance of past posts by the content creator amongst other users

Type: Type of post (status, photo, link) user prefers

Recency: How new is the post

* This is a simplified equation. Facebook also looks at roughly 100,000 other high-personalized factors when determining what's shown.

 Xiaofan Wang

Thank You!

Xiaofan Wang

Shanghai Jiao Tong University

xfwang@sjtu.edu.cn