第一讲 绪论

(一) 什么是生物信息学/系统生物学？

(二) 生物信息学的一些基本概念

(三) 生物信息学/系统生物学的一些基本问题

(四) 复杂网络与系统生物学

(五) 结论

3.1 单核苷酸 (A-T-G-C) 多态性 的研究

3.2 蛋白质结构预测问题

Protein Structure Prediction

传统的Bioinformatics包括了Sequencing（测序）和 Alignment（比对）

3.3 基因调控网络的推断

Gene Network inferring

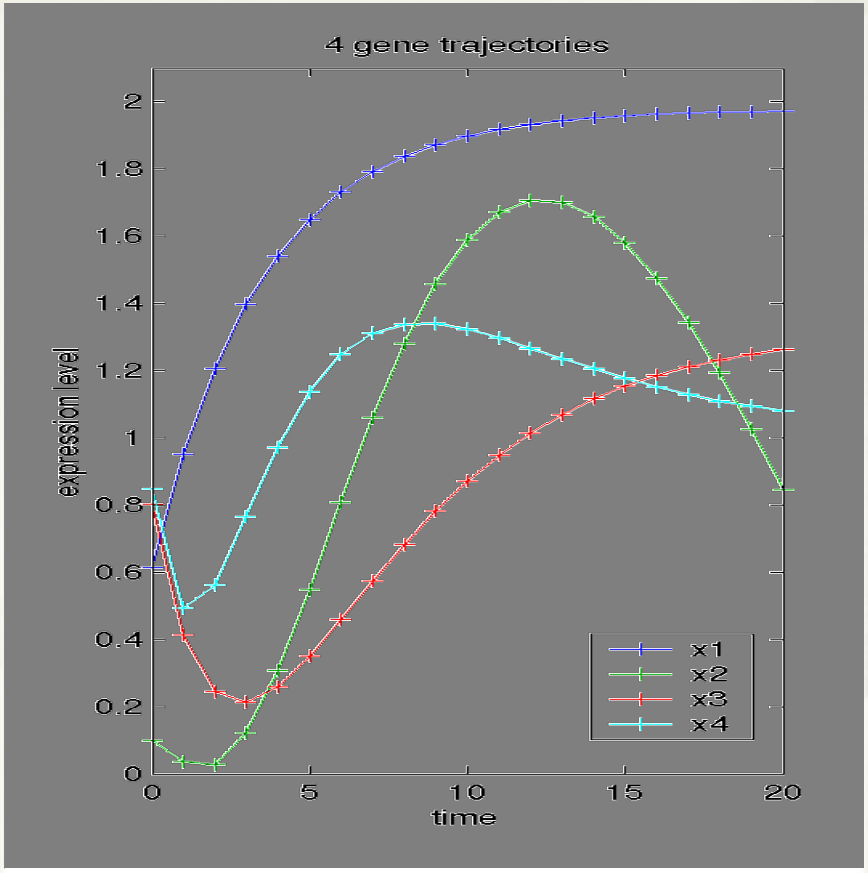
3.3 基因调控网络的推断

Gene Network inferring

什么是基因调控网络？（生物问题的抽象）

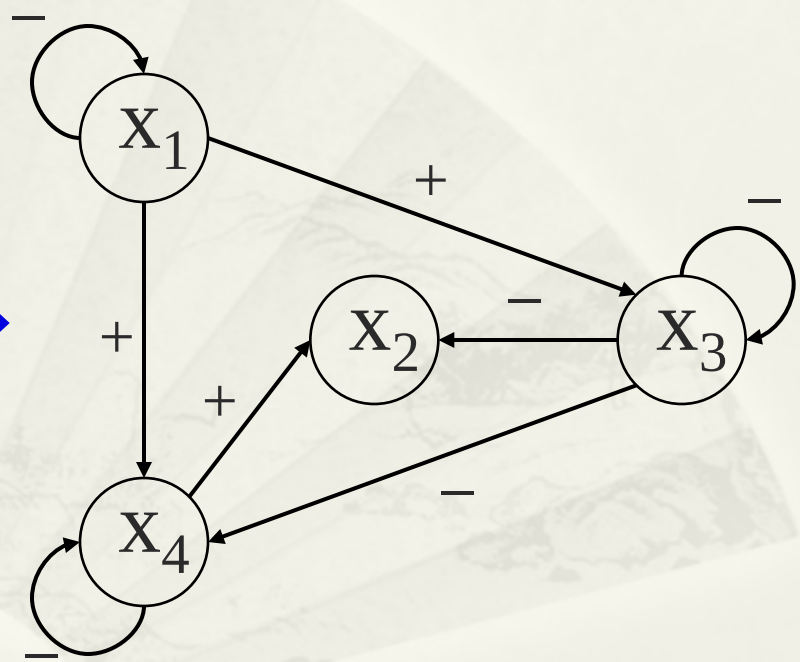
- 一个有向赋权网络:
  - 节点：基因
  - 边：表示基因之间的直接或者间接调控关系，基因A对基因B的调控关系指基因A的状态可以影响基因B的状态
  - 权：（权值大于 0 或者小于 0），调控作用可以是激活，或者是压抑

基因调控网络的推断



Time series

?



Gene network

用时变 n-维向量 $X(t)$ 表示基因组在时间$t_1, t_2, ...t_m$的基因表达量, n 是基因组内基因的个数。

用 $n \times n$ 矩阵 A 表示基因两两之间的调控关系：

$a_{ij} = 1$　　表示基因激活基因

$a_{ij} = -1$　　表示基因抑制基因

$a_{ij} = 0$　　表示基因和不相关

基因的时间序列表达数据

给定 n 个基因的的 m 个时间点的芯片数据实质上就是如下一个矩阵X，

x_{ij} = *第 i 个基因在第 j 时刻芯片测量得到的表达值*

$$\mathbf{X} = \begin{pmatrix} x_{11} & \cdots & x_{1j} & \cdots & x_{1n} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{i1} & \cdots & x_{ij} & \cdots & x_{in} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{m1} & \cdots & x_{mj} & \cdots & x_{mn} \end{pmatrix}$$

行代表第i个时刻点

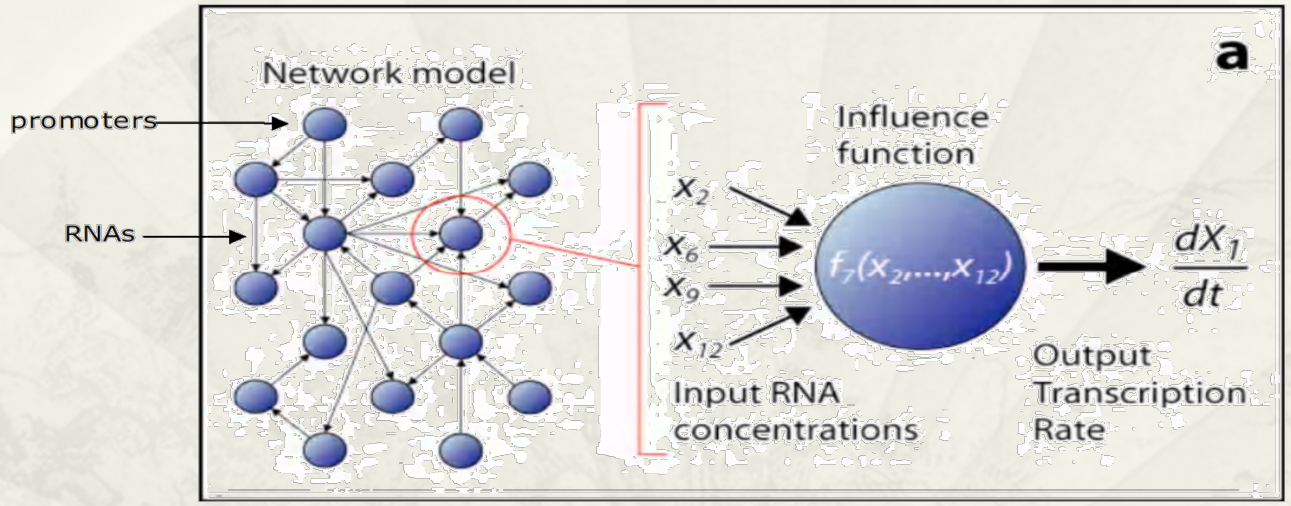
各个基因的表达值

列代表第 j 个基因的所有时刻的表达值

基因调控的微分方程模型

通常利用线性微分方程模型来建模基因调控关系:

$$dx_i(t)/dt = a_0 + a_{i,1}x_1(t) + a_{i,2}x_2(t) + \ldots + a_{i,n}x_n(t)$$



$$dX_1/dt = f_7(X_2,\ldots) = a_2 X_2 + a_6 X_6 + a_9 X_9 + a_{12} X_{12}$$

例如：基因1的表达值的变化率是对它有调控作用的基因2，6，9，13的表达值x_2, x_6。x_9, x_{12}的线性组合

数学表达

- 反向工程（Reverse-Engineering）

$$\dot{X} = J_{n\times n} X + B + \varepsilon$$
$$X(1), \cdots, X(m) \Rightarrow J_{n\times n}$$
$$X \in \mathrm{R}^n, \qquad m << n$$

- 已知X, B, 求矩阵 J

$$J_{n\times n} X_{n\times m} = \dot{X}_{n\times m} - B_{n\times m}$$

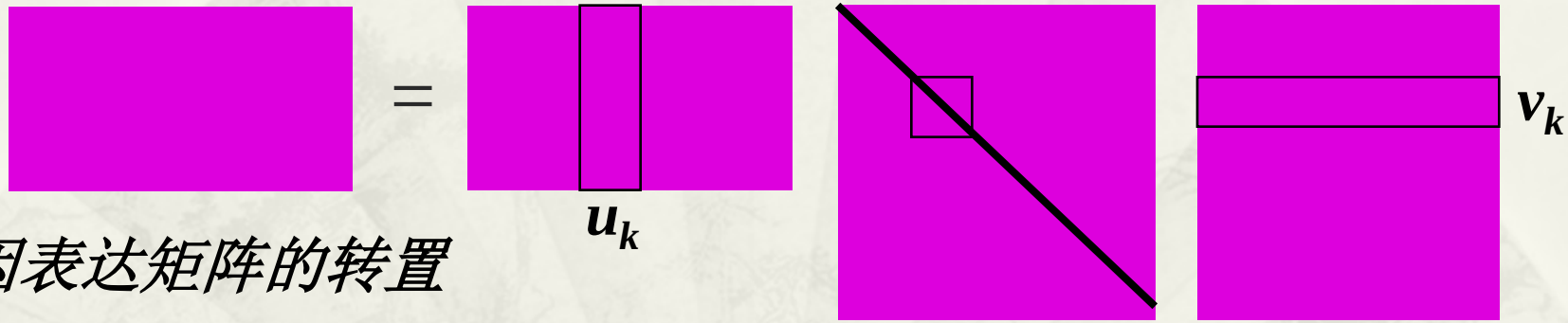
- m>n 时为最小二乘拟合问题，容易求解。但m<<n方程组有无穷多解。

维数问题

- 维数问题（Dimension Problem）：生物实验提供的数据的特点是时间点个数<<变量的个数。
- 例如在酵母中，最多可测量的时间点 m(约为20)<<酵母中基因的个数 n(约为6000) 。
- 上述网络推断问题从数学上是不定的，即有无穷多个网络结构可以拟合出实验观测到的数据。
- 关键问题：利用优化技术克服维数问题探求基因调控网络的拟最优结构。

奇异值分解（SVD）

SVD分解： $X^T_{m\times n} = U_{m\times n}E_{n\times n}V^T_{n\times n} \quad (m << n)$



基因表达矩阵的转置

奇异值

特解： $\hat{J} = (\dot{X} - B)UE^{-1}V^T$

通解表达

- SVD 解是最小二乘意义下的特解

$$\hat{J} = \arg\min \|JX + B - \dot{X}\|_2$$

- 通解表达 (General solution)

$$J = \hat{J} + YV^T$$

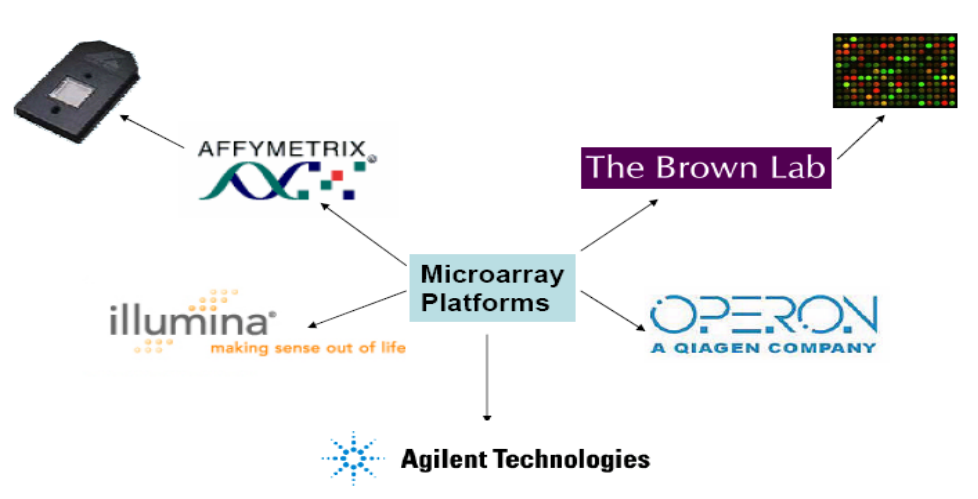
- Y 作为优化变量可以寻找最能解释生物数据的的通解

$$Y = \begin{bmatrix} y_{11} & y_{12} & \cdots & y_{1l} & 0.0 & \cdots & 0.0 \\ y_{21} & y_{22} & \cdots & y_{2l} & 0.0 & \cdots & 0.0 \\ \cdots & & \cdots & & & & \cdots \\ y_{n1} & y_{n2} & \cdots & y_{nl} & 0.0 & \cdots & 0.0 \end{bmatrix}$$

l 为基因表达矩阵中非零奇异值的个数

虽然每个时间序列只有有限的几个时间点，但是却有成千上万的时间序列

- *众多的基因芯片平台来测量基因表达*



- *公开数据库中的基因表达数据每年翻三番*

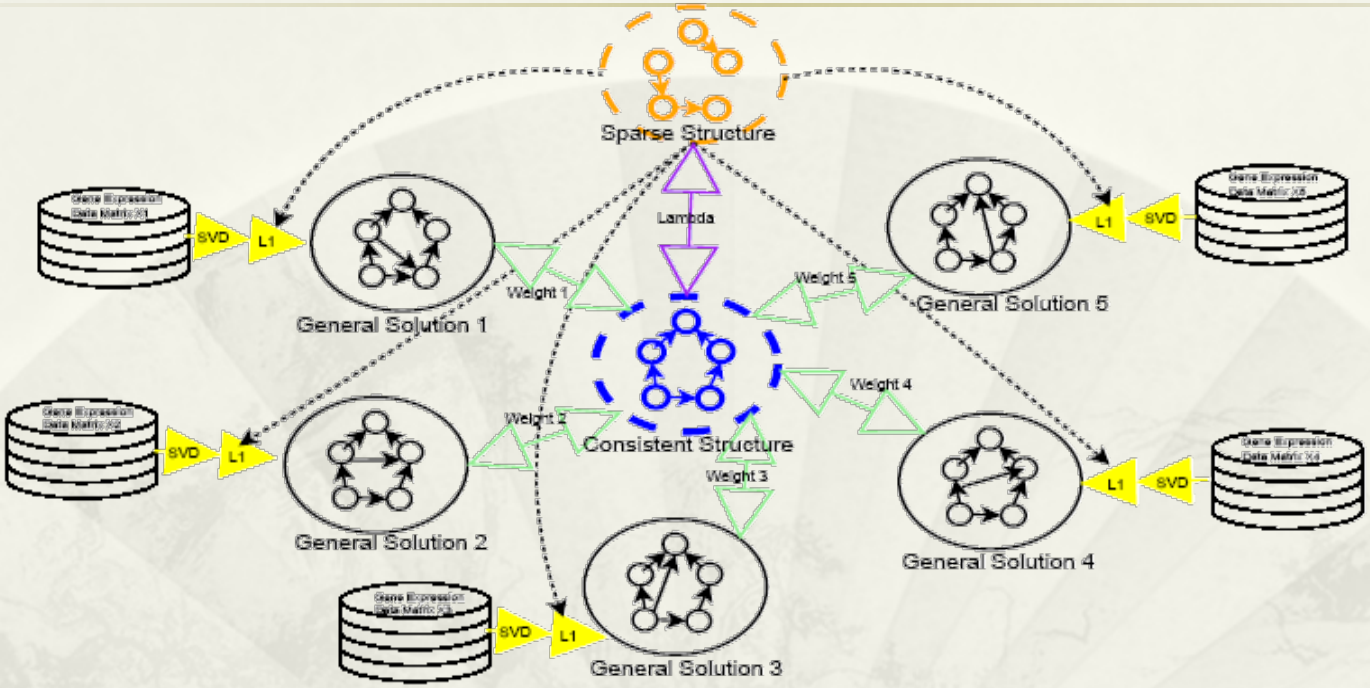
- NCBI Gene Expression Omnibus

GEO Gene Expression Omnibus 137231 experiments

- EBI Array Express

EMBL-EBI 55228 experiments

集成多个基因表达数据集



***Yong Wang**, Trupti Joshi, **Xiang-Sun Zhang**, Dong Xu, and Luonan Chen. Inferring gene regulatory networks from multiple microarray datasets, **Bioinformatics**, 22, 2413-2420, 2006.*

最优化模型

每个单数据集可以得到一个通解表达，代表与这个数据集相容的所有网络结构

集成多个数据集的目标是构建一个与各个数据集尽量相容的聚合的(aggregate)网络结构

$$\min_{Y,J} \sum_{k=1}^{N} \sum_{i=1}^{n} \sum_{j=1}^{n} [\omega^k |J_{ij} - J_{ij}^k| + \lambda |J_{ij}|]$$

目标 1：聚合的网络同各个数据集产生的网络之间尽量相容

目标 2：所得的生物网络结构是稀疏的

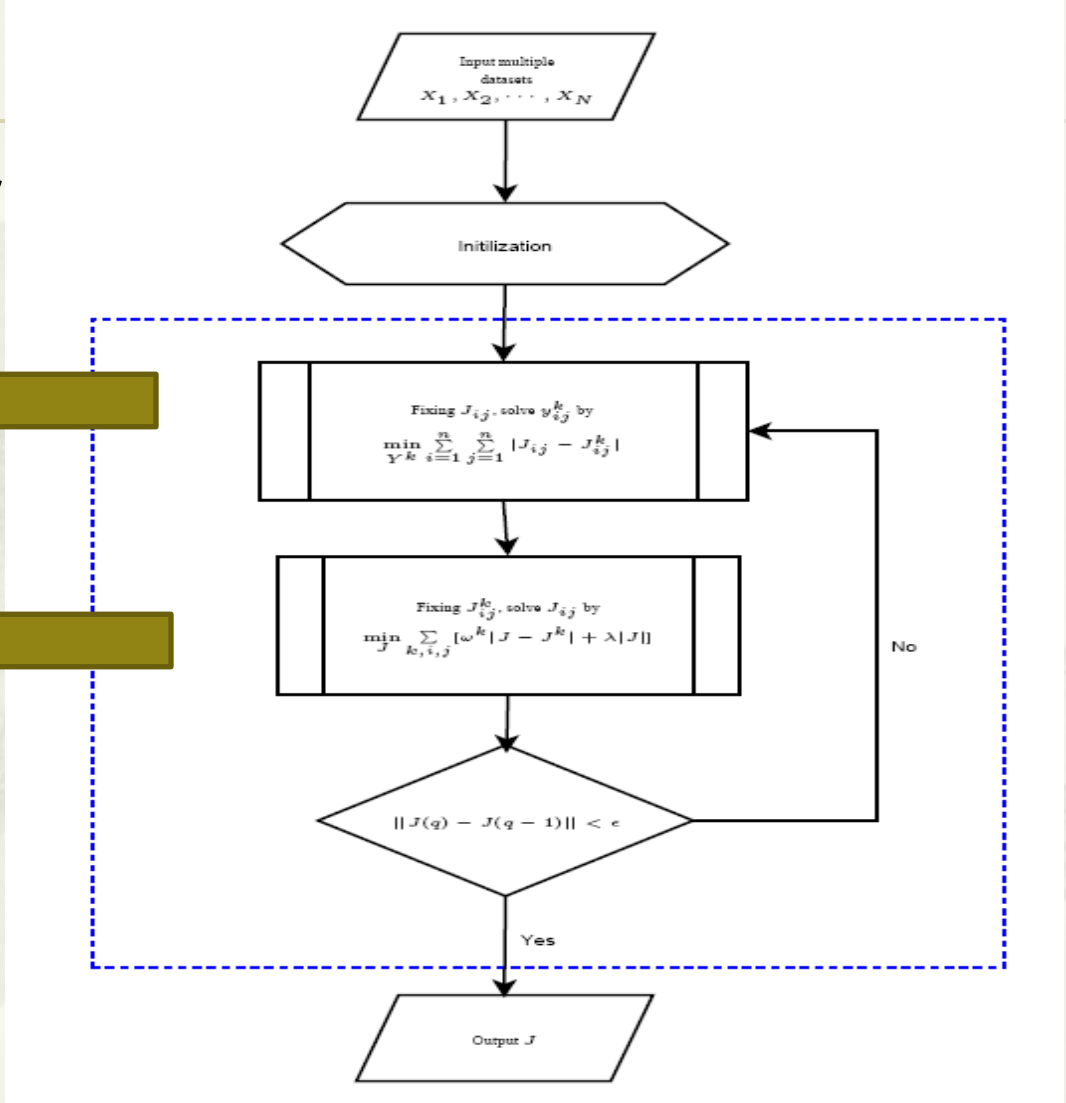
多目标转化为单目标：引入一个参数来权衡这两个优化目标

分解算法

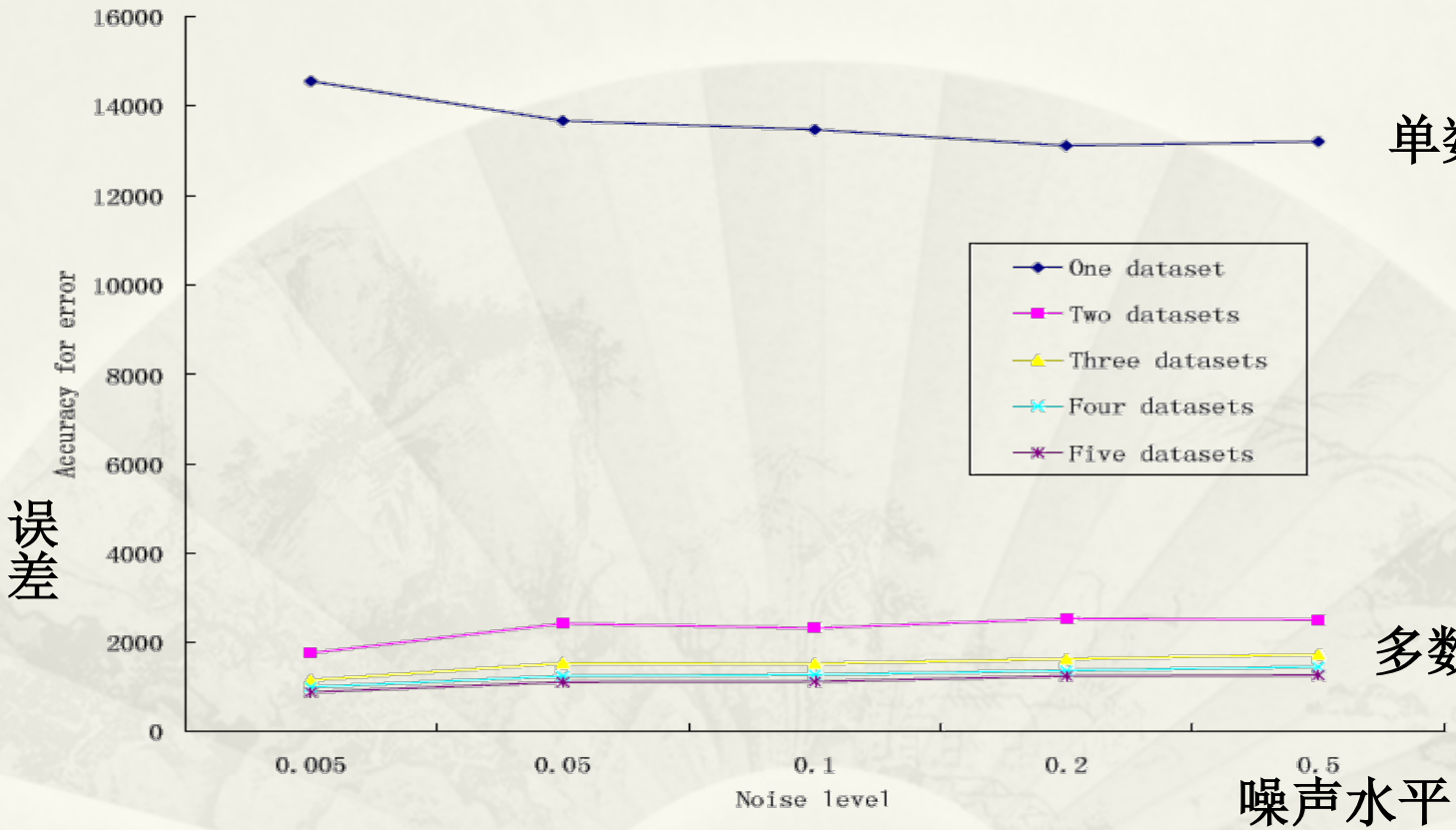
上述问题可以转化为大规模的线性规划问题

固定J，求Y

固定Y，求J



多个数据集显著提高精度



误差

单数据集

多数据集

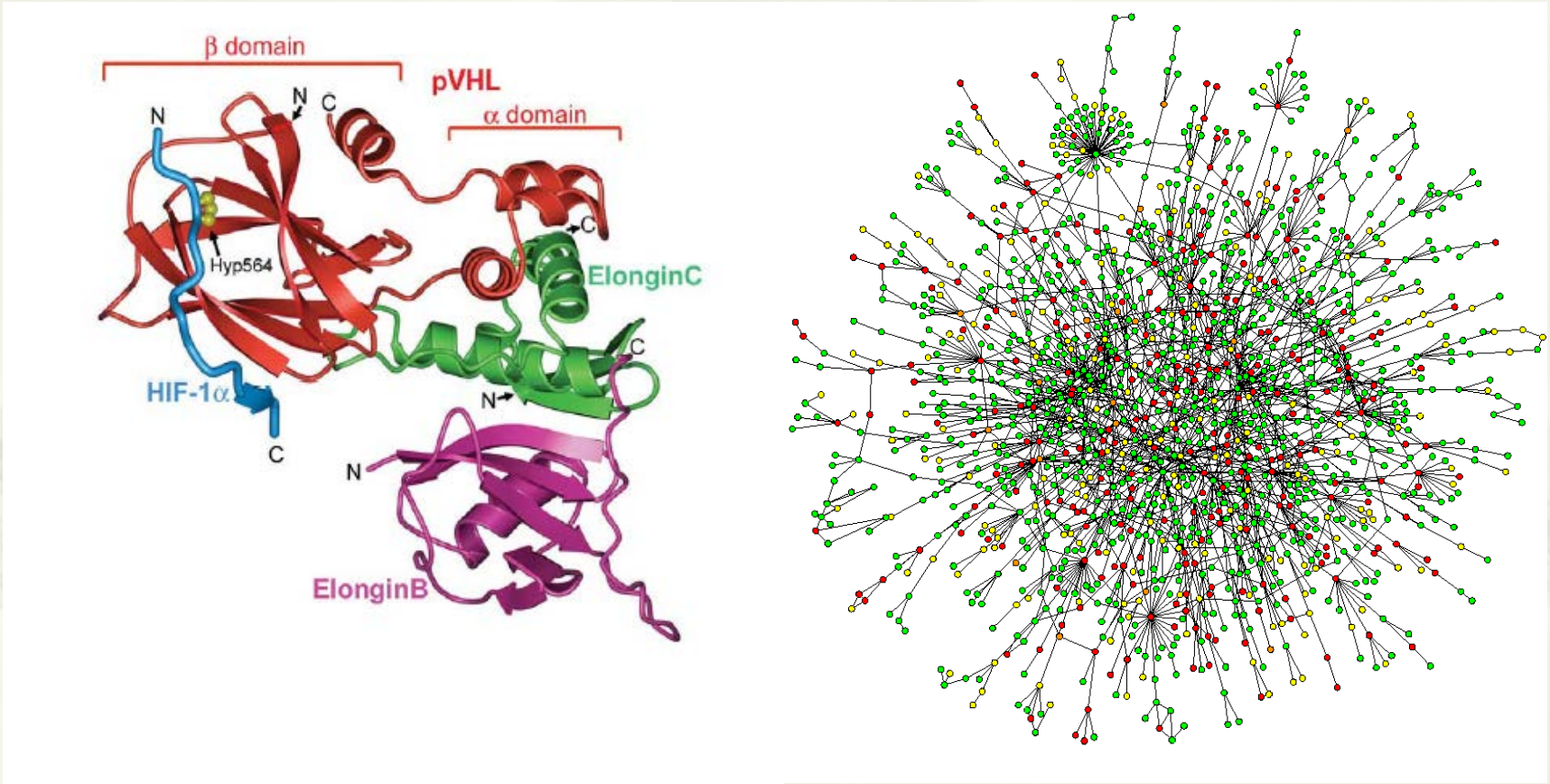
噪声水平

3.4 蛋白质相互作用网络

Protein-Protein Interaction Network

(PPI-Network)

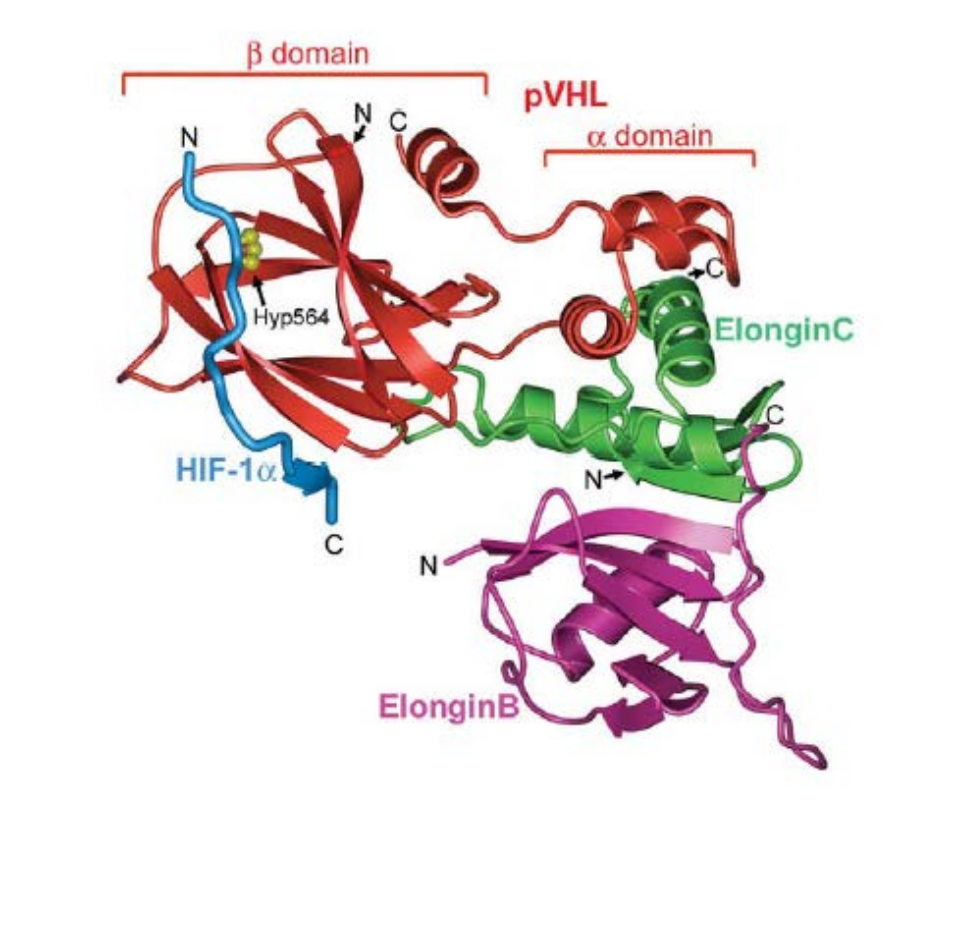
蛋白质相互作用网络的基本概念



基于domain信息PPI的预测
PPI prediction based domain information

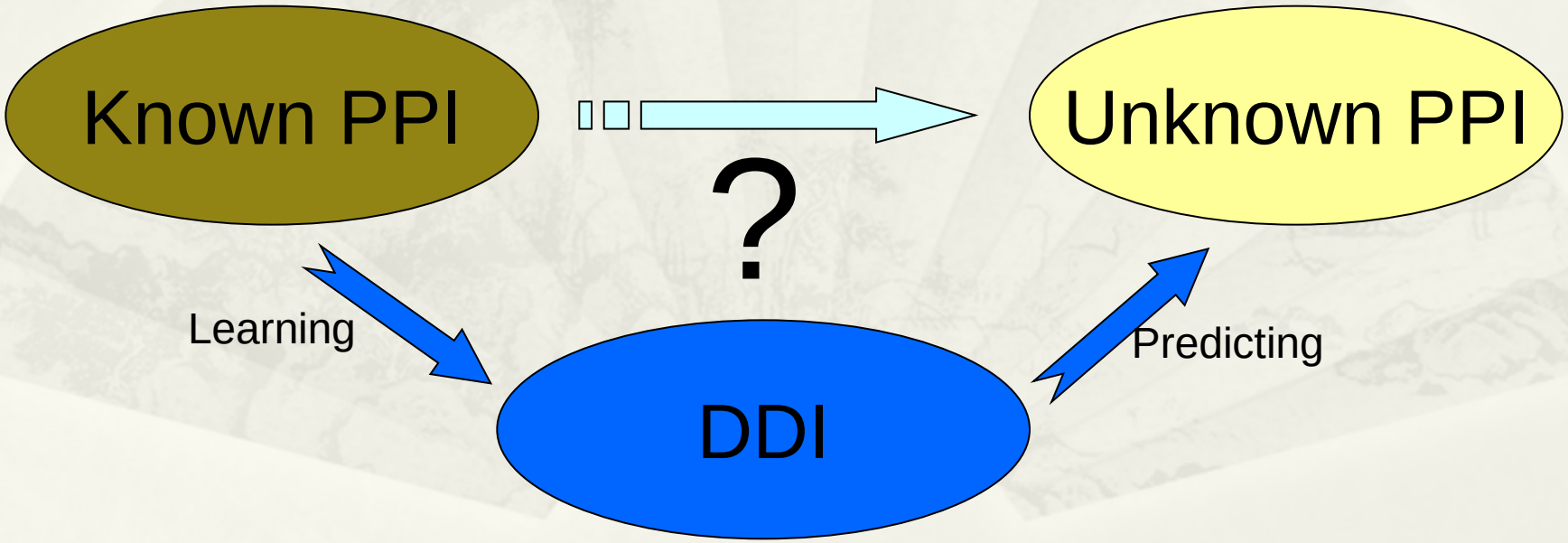
* Domain（域）：蛋白质的一个组成部分，它的空间结构可独立于其他氨基酸折叠而成，具有自己的功能
* 一个蛋白质中的domain组合在一起确定该蛋白质的总的功能

蛋白质相互作用网络的基本概念(continued)



蛋白质相互作用网络的基本概念(continued)

* 基本假设: 两个蛋白质相互作用，至少各有一个domain相互作用(domain-domain interaction, DDI)



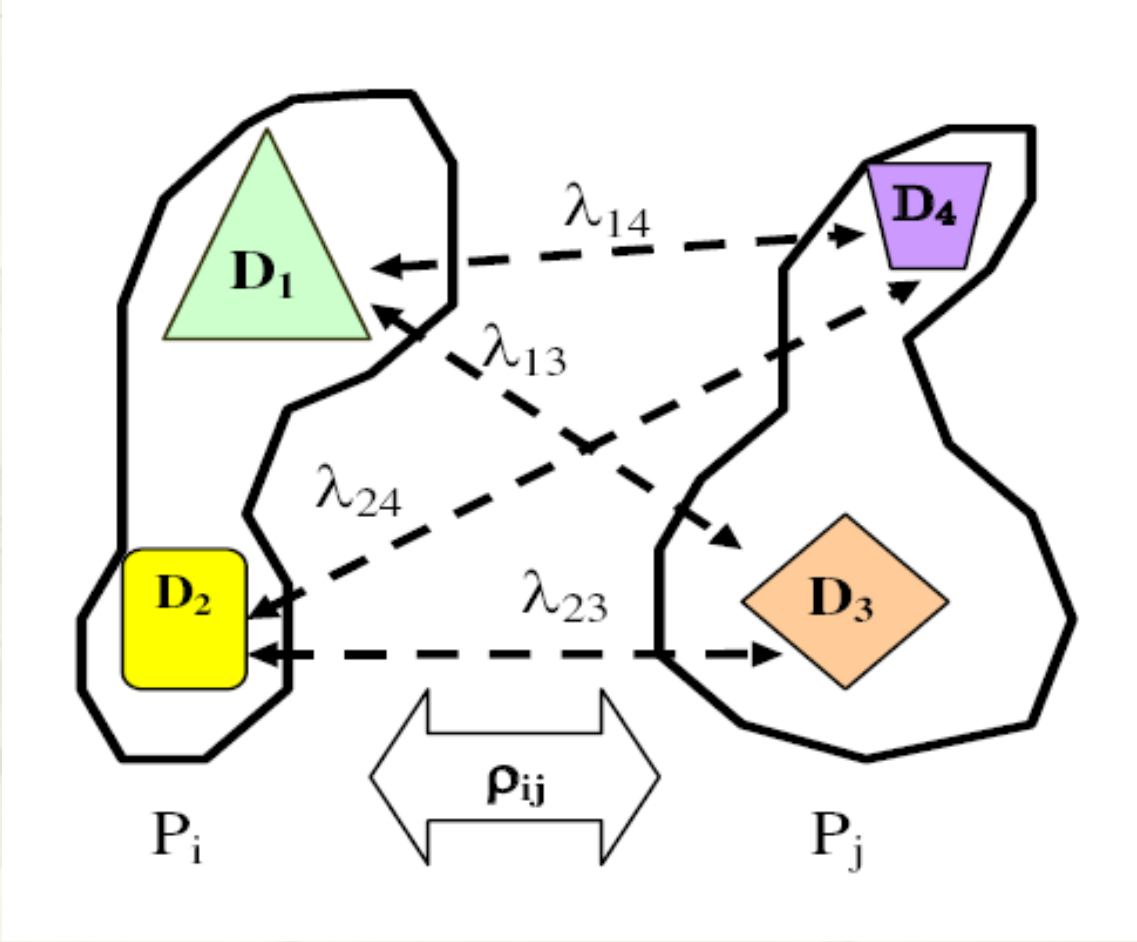
Notations

- 蛋白质组: $P = \{ P_1, P_2, \cdots, P_N \}$
- Domain 组: $D = \{ D_1, D_2, \cdots, D_M \}$
- 一个蛋白质内的domain: $P_i = \{ D^i_1, D^i_2, \cdots, D^i_n \}$
- 一对蛋白质（i, j）: $P_{ij} = (P_i, P_j)$
- 一对 domain （m, n）: $D_{mn} = (D_m, D_n)$
- $P_{ij} = \{ D_{mn} \mid D_m \in P_i, D_n \in P_j \}$ 也用来记蛋白质对(i, j)相应的所有domain对

Notations (continued)

- PPI度量
  - 蛋白质对 (i, j) 相交：$p_{ij} = 1$
  - 蛋白质对 (i, j)不 相交：$p_{ij} = 0$
  - 蛋白质对 (i, j) 相交的概率： ρ_{ij}
- DDI度量
  - domain对 (m, n) 相交： $d_{mn} = 1$
  - domain对 (m, n) 不相交： $d_{mn} = 0$
  - domain对 (m, n) 相交的概率： λ_{mn}

An example



蛋白质相互作用网络预测的概率模型

- 基本假定:
  - 每一 DDI 是相互独立的
  - 两个蛋白质相互作用，至少有一个 DDI
- 基本问题:
  - 从PPI观察值来推导λ_{mn} 和 ρ_{ij} 的研究
  - λ_{mn} 和 ρ_{ij} 的互导

概率模型A. 从PPI观察值来推导λ_{mn} 和和ρ_{ij} 的研究:

对 λ_{mn} 的估计— Association Method

$$\lambda_{mn} = \frac{\sum_{\substack{D_{mn} \in P_{ij} \\ P_{ij} \in \mathcal{P}_{train}}} o_{ij}}{\sum_{\substack{D_{mn} \in P_{ij} \\ P_{ij} \in \mathcal{P}_{train}}} 1}$$

where o_{ij}=0/1 is an observation which denotes whether P_i and $P_j$have interaction.

- Sprinzak, E. and Margalit, H. (2001) Correlated sequence-signatures as markers of protein-protein interactions. *Journal of Molecular Biology*, 311, 681-692.

A. 从PPI观察值来推导λ_{mn} 和 ρ_{ij} 的研究: 对 ρ_{ij}的估计

—— Expectation-Maximization (EM) method

$$\begin{aligned}\Pr(o_{ij}=1) &= \Pr(o_{ij}=1, p_{ij}=1) + \Pr(o_{ij}=1, p_{ij}=0) \\ &= \Pr(o_{ij}=1|p_{ij}=1)\Pr(p_{ij}=1) \\ &\quad + \Pr(o_{ij}=1|p_{ij}=0)\Pr(p_{ij}=0) \\ &= \Pr(p_{ij}=1)(1-fn) + (1-\Pr(p_{ij}=1))fp\end{aligned}$$

where

$$\text{false positive rate } fp = \Pr(o_{ij}=1|p_{ij}=0)$$
$$\text{false negative rate } fn = \Pr(o_{ij}=0|p_{ij}=1)$$

Likelihood function:

$$L = \prod (\Pr(o_{ij}=1))^{o_{ij}} (1-\Pr(o_{ij}=1))^{1-o_{ij}}$$

- **Deng, M. *et al.* (2002) Inferring domain-domain interactions from protein-protein interactions. *Genome Research*, 12, 1540-1548.**

概率模型B.　λ_{mn} 和 ρ_{ij} 的互导：　$\lambda_{mn} \rightarrow \rho_{ij}$

知道了DDI的概率λ_{mn}，可以推出PPI的概率ρ_{ij}

PPI的概率 = 1 – 所有DDI不相交的概率

= 1 – Π{ 1 – 每个DDI相交的概率}，即

$$\Pr(p_{ij}=1) = 1 - \prod_{D_{mn} \in P_{ij}} (1 - \Pr(d_{mn}=1))$$

亦即

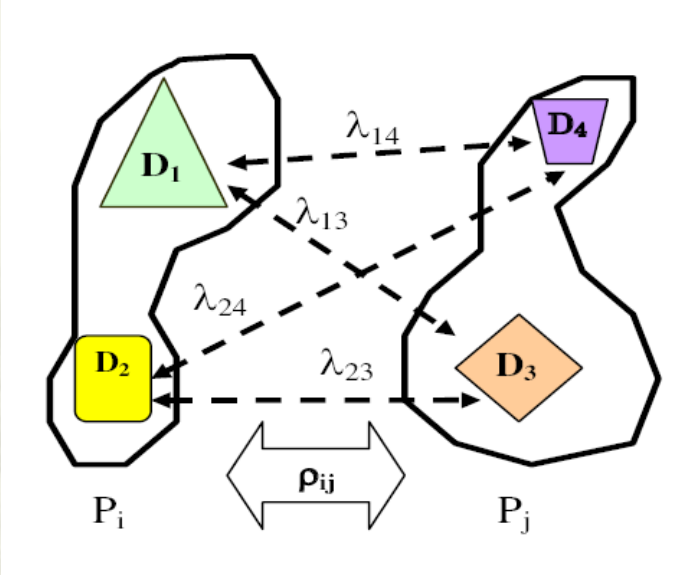
$$\rho_{ij} = 1 - \Pi\{1 - \lambda_{mn}\}$$

概率模型B. λ_{mn} 和 ρ_{ij} 的互导：$\rho_{ij} \rightarrow \lambda_{mn}$ —— ASNM method

$$\lambda_{mn} = \frac{\sum_{\substack{D_{mn} \in P_{ij} \\ P_{ij} \in \mathcal{P}_{train}}} \rho_{ij}}{\sum_{\substack{D_{mn} \in P_{ij} \\ P_{ij} \in \mathcal{P}_{train}}} 1}$$

- Hayashida, M. and Ueda, N. (2004) A simple method for inferring strengths of protein-protein interactions. *Genome Informatics*, 15(1), 56-68.

ASNM方法的问题是对所有的ρ_{ij}一视同仁，其实要加权处理



- Assume ρ_{ij}=0.1, then λ_{13} =λ_{14}= λ_{23} = λ_{24} =0.1 according to the ASNM method; So Pr(p_{ij} =1) = 0.3439 from the prediction formula.

- According to the APM method, λ_{13} =λ_{14}= λ_{23} = λ_{24} =1 – (1- 0.1)$^{1/4}$. So Pr(p_{ij} =1) = 0.1 that is consistent with the experimental ratio ρ_{ij}=0.1.

概率模型B. λ_{mn} 和 ρ_{ij} 的互导：

$\rho_{ij} \rightarrow \lambda_{mn}$ —— APM method

* An association probabilistic method (APM):

$$\lambda_{mn} = \frac{\sum_{\substack{D_{mn} \in P_{ij} \\ P_{ij} \in \mathcal{P}_{train}}} \left(1 - (1 - \rho_{ij})^{\frac{1}{|P_{ij}|}}\right)}{\sum_{\substack{D_{mn} \in P_{ij} \\ P_{ij} \in \mathcal{P}_{train}}} 1}$$

where $|P_{ij}|$ represents the number of domain pairs in P_{ij}.

- **L. Chen, L.Y. Wu, Y. Wang, X.S. Zhang. Inferring protein interactions from experimental data by association probabilistic method.** ***Proteins: Structure, Function, and Bioinformatics,*** **2006.**

蛋白质相互作用网络预测的优化模型
—— *Parsimony Model* （PM)

基本思想：

* 两个蛋白质相接， 在各自的domain组中仅有少数几对domain对相接；

* 对于已出现在多对PPI中的某一对DDI，一旦出现在另一蛋白质对的domain对中，很有可能成为这一对蛋白质相接的媒介

蛋白质相互作用网络预测的优化模型

—— *Parsimony Model* (continued)

- Parsimony Method (PM): 我们用尽可能少的DDI对来“解释”被观察到的PPI对.
- 这可以用以下的整数线性规划 (ILP) 来描述：

$$\text{Minimize}_{\{d_{mn}\}} \quad \sum_{D_{mn} \in \mathcal{D}} d_{mn}$$

$$\text{Subject to:} \quad \sum_{D_{mn} \in P_{ij}} d_{mn} \geq 1 \quad \text{for each } P_{ij} \in \mathcal{P}$$

$$d_{mn} \in \{0, 1\} \quad \text{for all } m, n.$$

所得到的解称为节俭的DDI组

蛋白质相互作用网络预测的优化模型

—— *Parsimony Model* (continued)

* d_p 称为有效信息利用度，是指所有被观察到的PPI对中具有有效信息的利用部分的比例。以上的ILP可以改进为

$$\text{Minimize}_{\{d_{mn},e_{ij}\}} \quad \sum_{D_{mn}\in\mathcal{D}} d_{mn}$$

$$\text{Subject to:} \quad \sum_{D_{mn}\in P_{ij}} d_{mn} + e_{ij} \geq 1 \quad \text{for each } P_{ij} \in \mathcal{P}$$

$$\sum_{P_{ij}\in\mathcal{P}} e_{ij} \leq (1 - d_p) \cdot |\mathcal{P}|$$

$$d_{mn} \in \{0,1\} \quad \text{for all } m,n$$

$$e_{ij} \in \{0,1\} \quad \text{for all } P_{ij} \in \mathcal{P}.$$

* 改进后的模型称为 ILP_pd.

蛋白质相互作用网络预测的优化模型

—— *Parsimony Model* (continued)

* 我们还可以利用已知的负例子集（a gold-standard negative data set）N，即由这些蛋白在细胞中的定位信息可知它们不会相接，按假定它们的domain也不会相接。此时模型为

$$\text{Minimize}_{\{d_{mn},e_{ij}\}} \sum_{D_{mn}\in\mathcal{D}} d_{mn} + \sum_{D_{mn}\in N_{ij}} d_{mn}$$

$$\text{Subject to:} \sum_{D_{mn}\in P_{ij}} d_{mn} \geq 1 \quad \text{for each } P_{ij} \in \mathcal{P}$$

$$d_{mn} \in \{0,1\} \quad \text{for all } m,n$$

* 记这一模型为 ILP_neg.

蛋白质相互作用网络预测的优化模型— *Parsimony Model* (continued)

—— computational complexity

- 基本模型 ILP 是图论中有名的 “Hitting Set problem”， 是 NP-hard problem.
- 我们直接解这些问题的线性松弛问题，分别记为 LP, LP_pd, LP_neg
- LP_pd 和 LP_neg 还可以结合成为 LP_neg_pd.

蛋白质相互作用网络预测的优化模型

— *Parsimony Model* (continued)

得到一组节俭的 DDI 组以后，我们用以下公式来预测蛋白质对 Pij 是否相互作用：

$$\text{If} \sum_{D_{mn} \in P_{ij}} d_{mn} \geq \rho, \text{then } p_{ij} = 1, \text{else } p_{ij} = 0,$$

此处 ρ **可以设为** 1 或任意小于 1 的正数.

（四）

复杂网络与系统生物学

4.1 什么是复杂网络(特指尺度无关网络,Scale-free network)?

一大类处于确定性网络和随机网络之间的大型网络

一般复杂网络(尺度无关网络)

- 许多复杂系统可以用网络进行表示
- 社会网络：科学合作网、食物网络和运输网络等
- 科学技术网络：因特网，万维网，软件相关网络等
- 生物分子网络：蛋白质相互作用网络，基因调控网络，新陈代谢网络等
- 有趣的是这些网络中的大部分都具有Scale-free网的一些拓扑性质.

Scale-free网的一些拓扑性质

- 小世界(Small world)
- 尺度无关 (Scale free)
- 聚类特性 (Clustering); 模块结构(Module); 社团结构(Community Structure)
- ……

A. Small World

- 对一个有 n 个点的网，每两点之间的距离满足:

$$\max_{(i,j)} short(v_i, v_j) \sim \ln n$$

称为 Logarithemic path length.

- local clustering coefficient (LCC) was introduced to measure if a network is a small world network (D.J. Watts, S.Strogatz, *Nature*, 1998)

B. Scale-free network

* 对一个大的网络, 记p(k)为度为k的点的比例, 则

$$p(k) \sim k^{-\gamma}$$

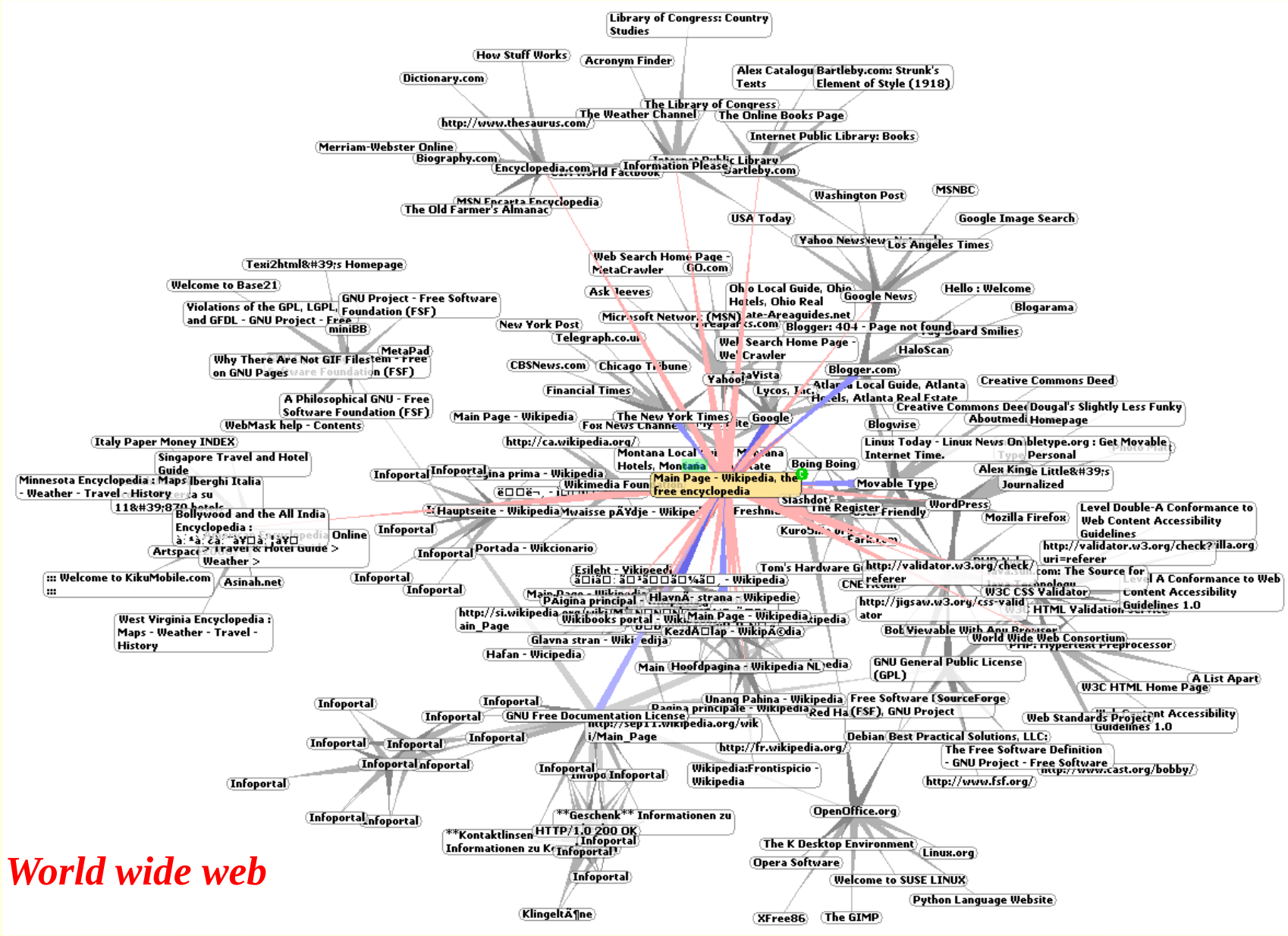
这一公式同 n 无关，称为 Power law. γ为幂指数，对于实际的网络，取值在（2,4）的范围内。

$$2.1 \leq \gamma \leq 4$$

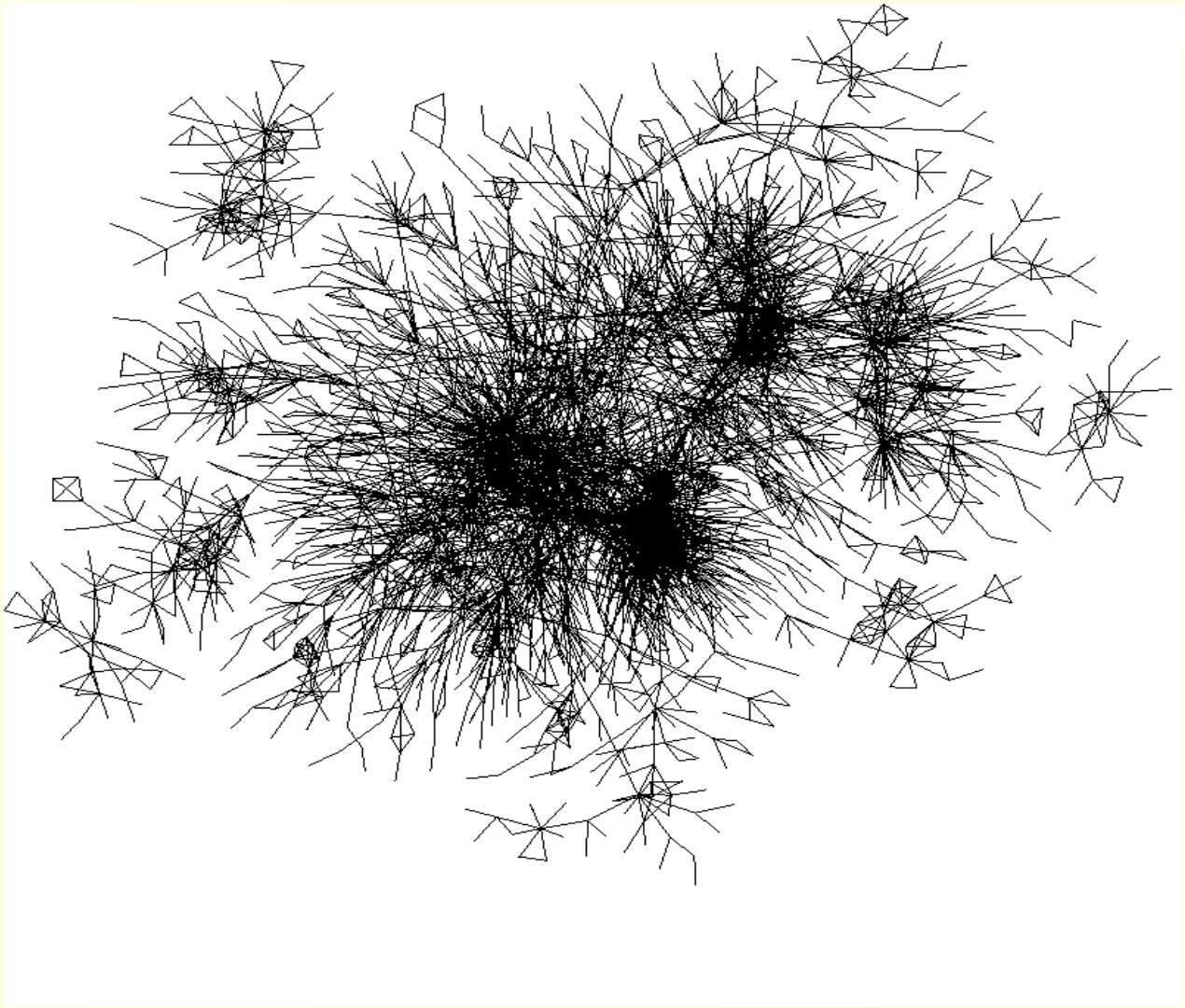
A.L.Barabasi, R.Albert, *Science*, 1999

一些有名网络例子

- 酵母细胞中的蛋白质相互作用网络 (A.-L. Barabási, NATURE REVIEWS GENETICS, 2004)
- Jeong H, Mason SP, Barabási AL, Oltvai ZN (2001) Lethality and centrality in protein networks. NATURE, 411(6833): 41–4
- Haiyuan Yu et al. High-Quality Binary Protein Interaction Map of the Yeast Interactome Network, SCIENCE, 2008
- Jean-Franc¸ois Rual et al, Towards a proteome-scale map of the human protein–protein interaction network, NATURE, Vol 437|20 October 2005
- Yong-Yeol Ahn et al. Link communities reveal multiscale complexity in networks, NATURE, Volume: 466, Pages: 761–764, (05 August 2010)
- Gergely Palla, Imre Derényi, Illés Farkas and Tamás Vicsek.
  Uncovering the overlapping community structure of complex networks in nature and society, NATURE, 435, 814-818(9 June 2005)
- Li et al A map of the interactome network of the metazoan C. elegans.
  SCIENCE. 2004, 303(5657):540-3.

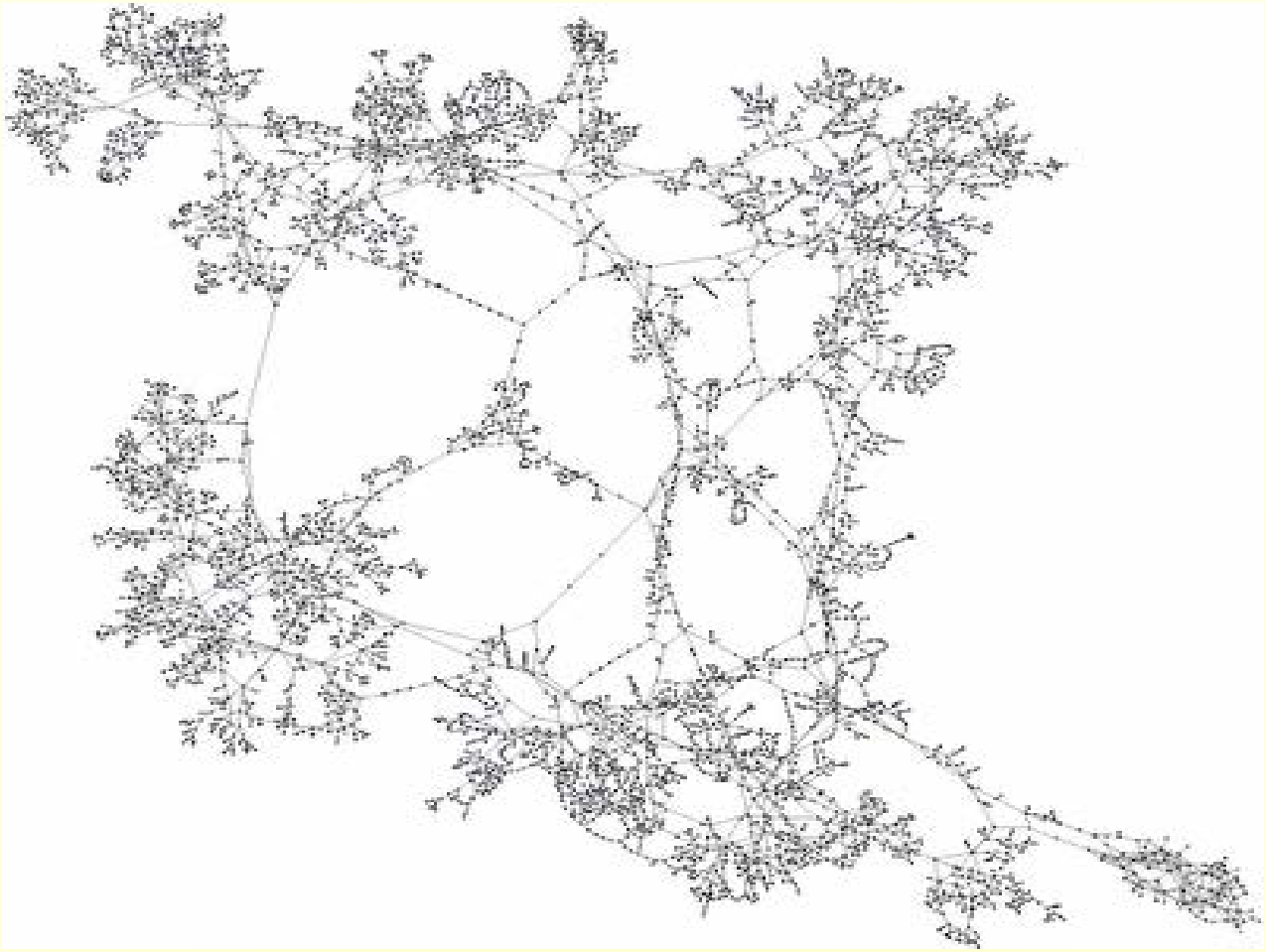


World wide web

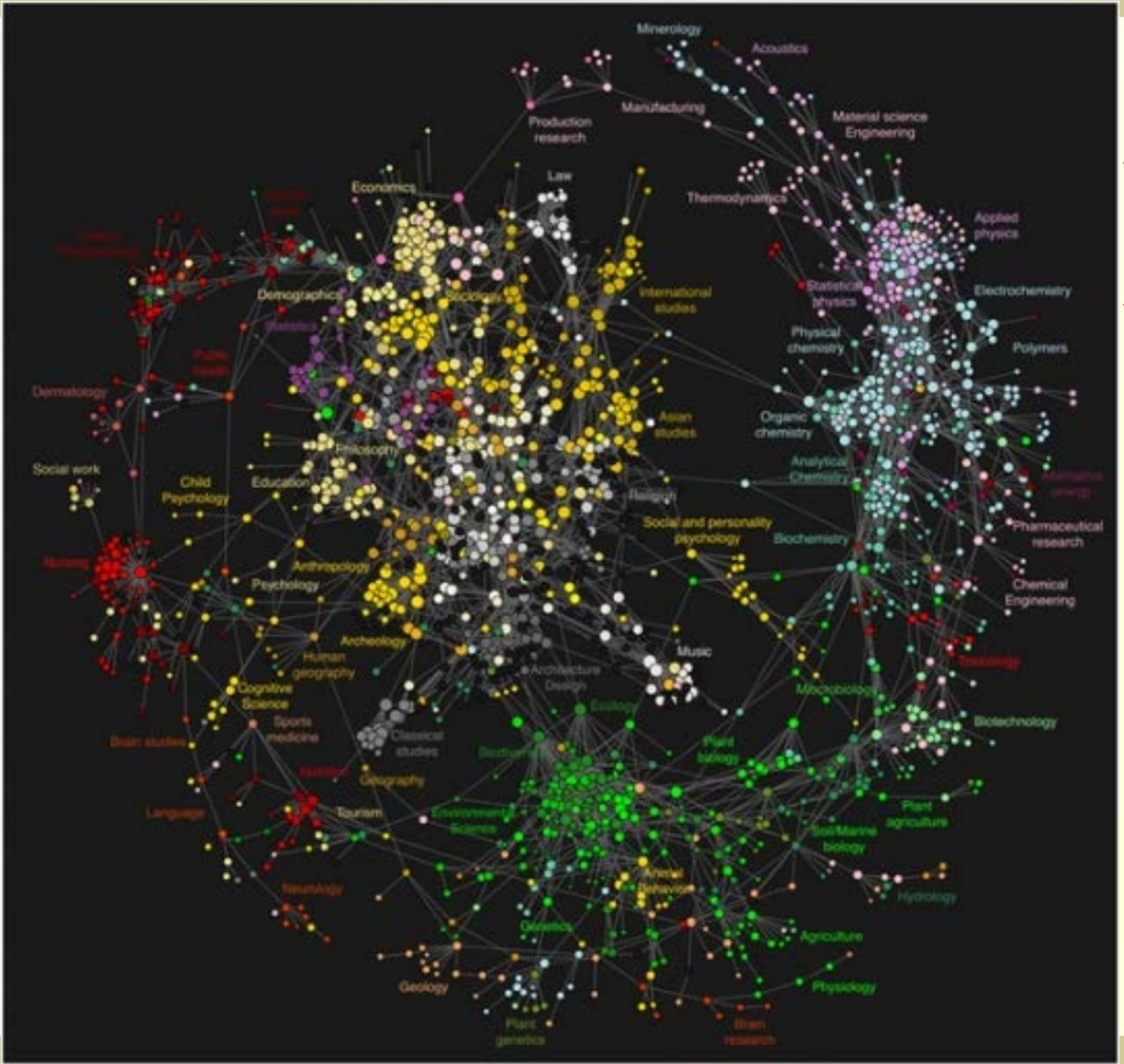


The collaboration graph of movie actors

Each of these graphs represents a different, and disjoint, 10k vertex induced subgraph fo the IMDB graph (specifically, the largest connected component of such a graph). The IMDB graph consists of actors, where two actors are connected if they were in a common movie. The data is courtesy of IMDB.



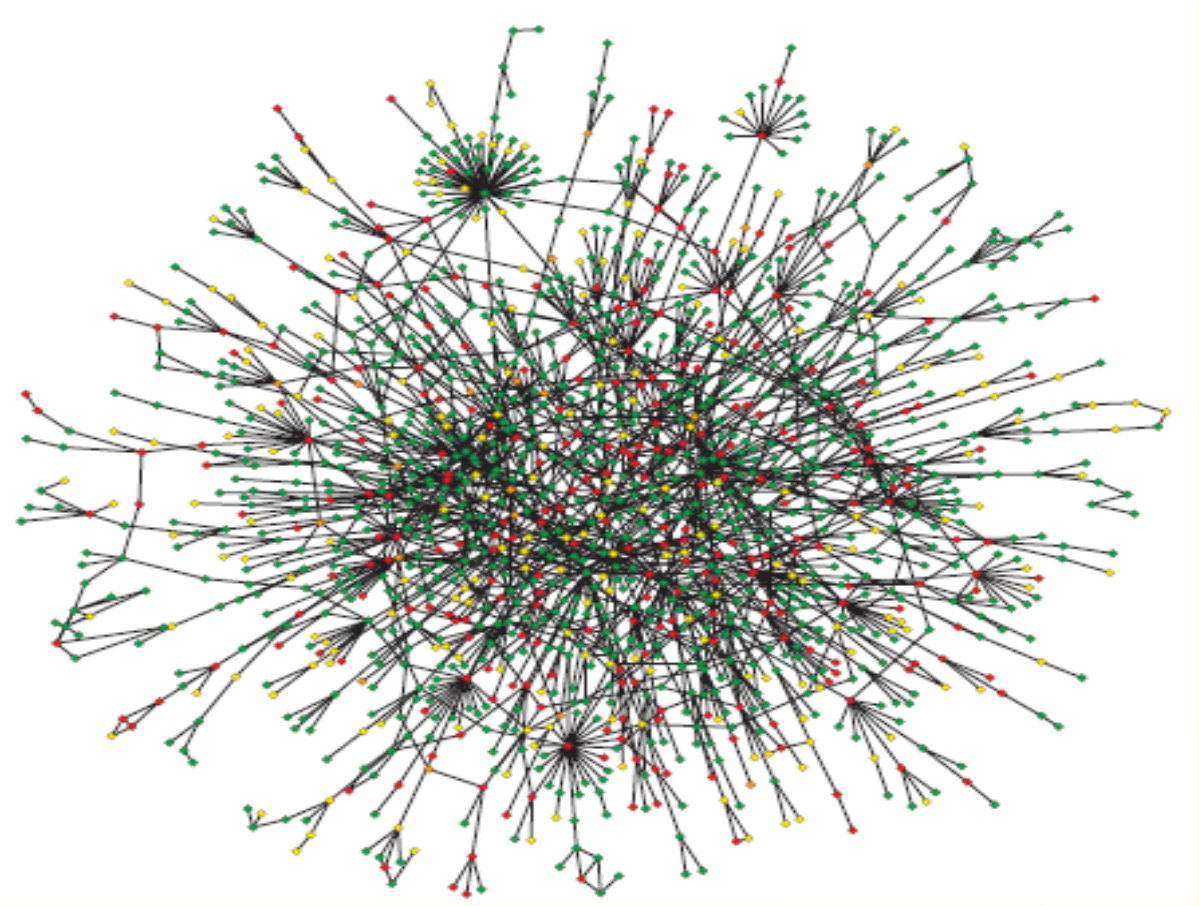
Electrical power grid



citation patterns of the scientific publications

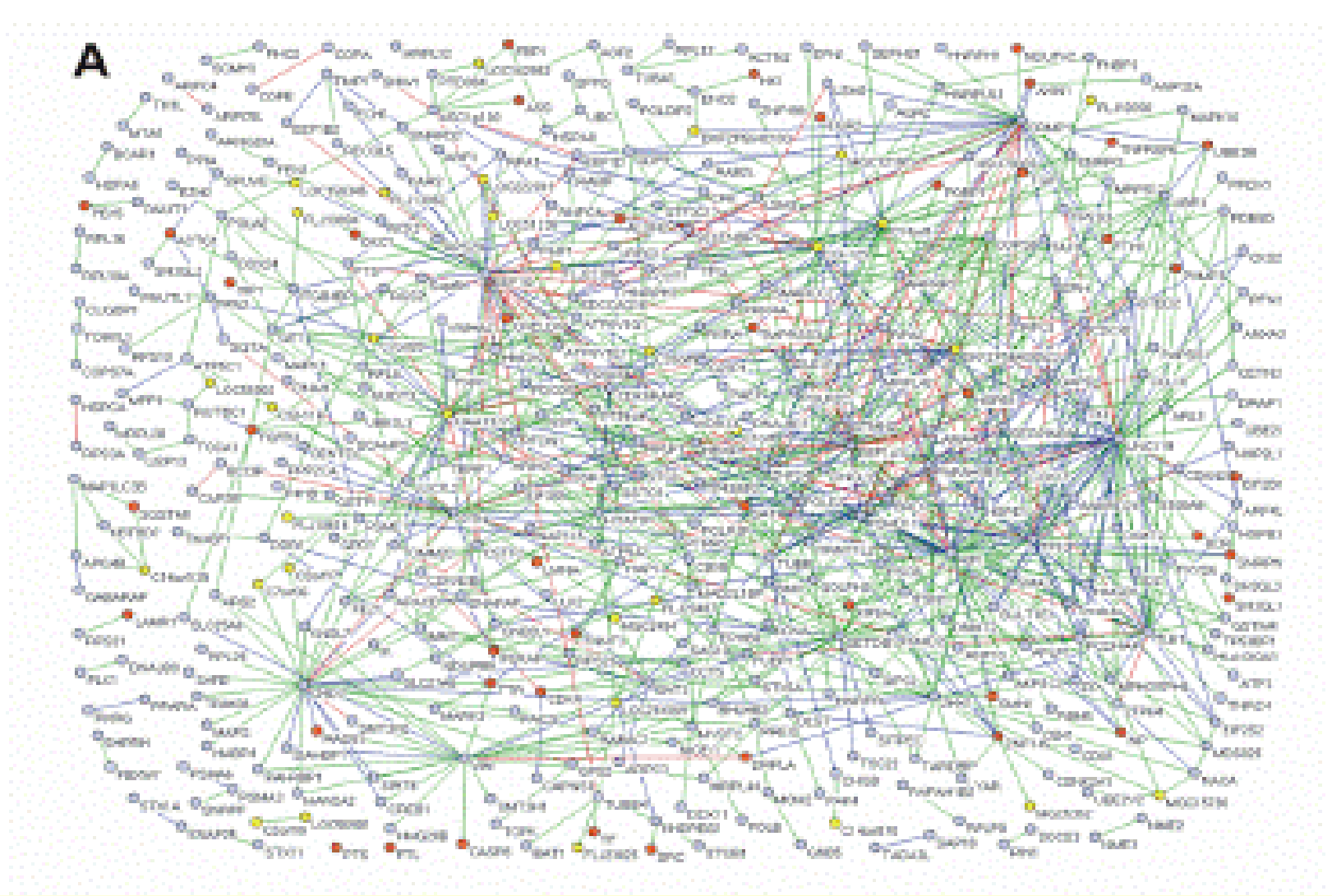
Networks of science have been created from citation data to visualize the structure of scientific activity.

复杂网络的典型代表：生物分子网络



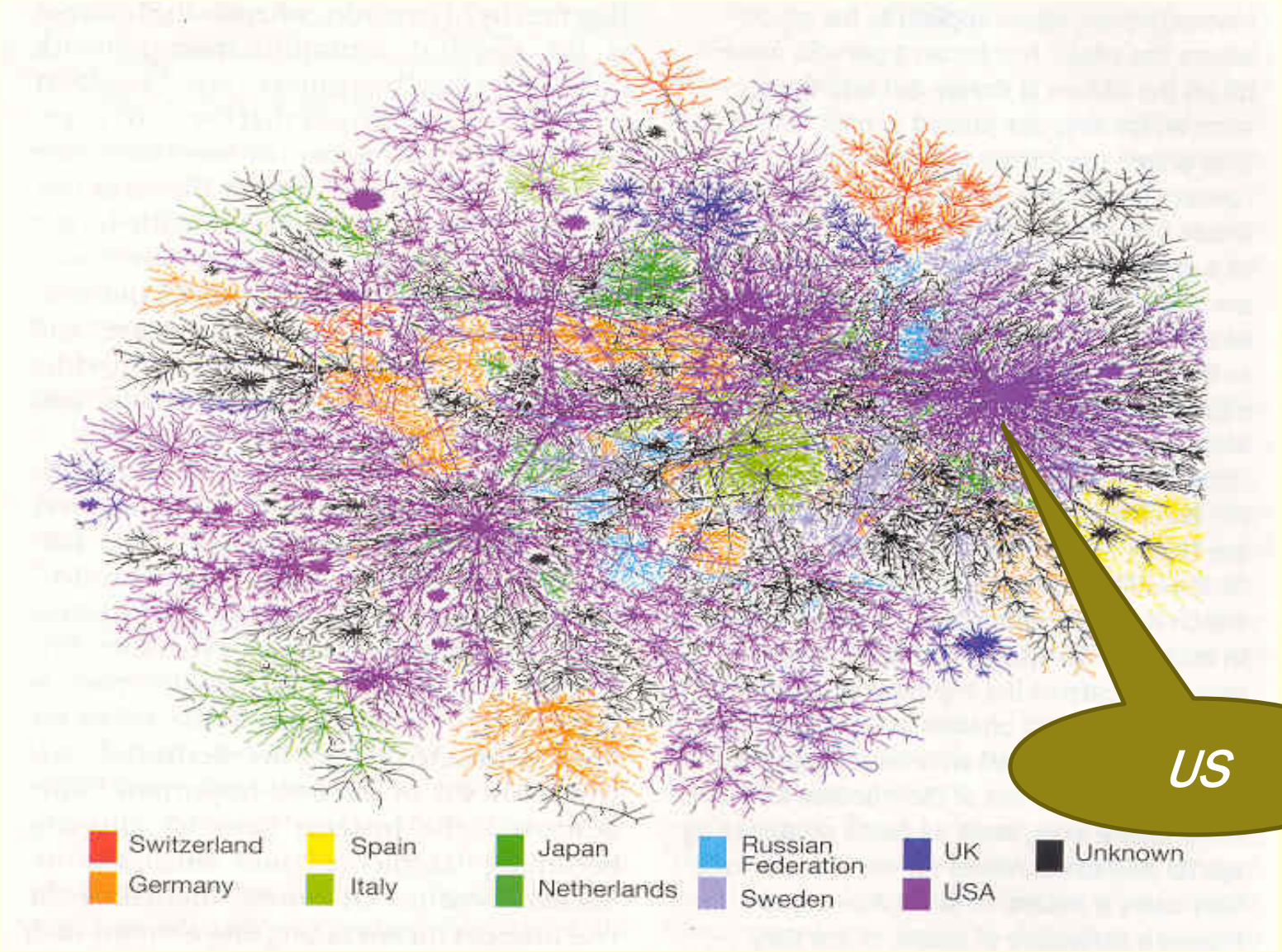
酵母细胞中的蛋白质相互作用网络 (A.-L. Barabási, NATURE REVIEWS GENETICS, 2004)

Human protein interaction network



Ulrich Stelal et al. A human protein-protein interaction network: A resource for annotating the proteome, ***Cell, Vol. 122, 957–968, September 23, 2005***

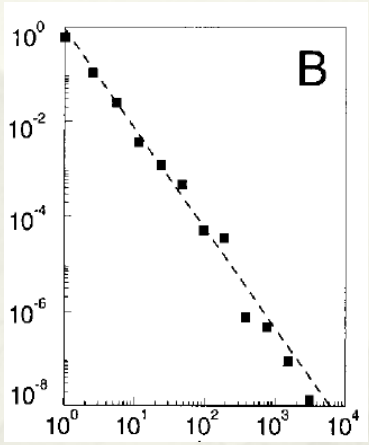
The network of Autonomous systems (ASs): an IP network



Power Law

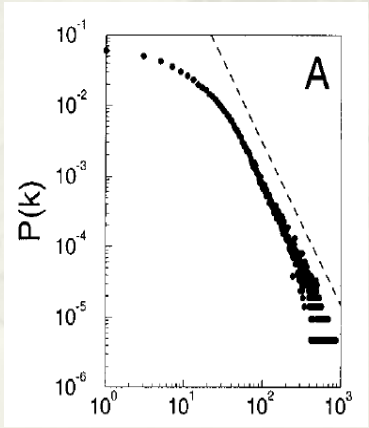
- World wide web (www) 有10亿个节点

$$\gamma_{www} = 2.1 \pm 0.1$$



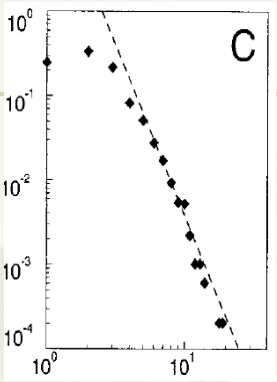
- The collaboration graph of movie actors:

$$\gamma_{actor} = 2.3 \pm 0.1$$



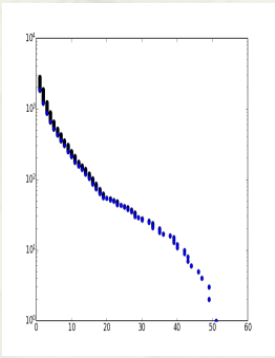
* Electrical power grid of the western United States:

$$\gamma_{power} = 4$$



* The citation patterns of the scientific publications:

$$\gamma_{cite} = 3$$



复杂网络在OMICS研究中的重要性

- Cellular networks are scale-free
- High clustering in cellular networks
- Modules (motifs) are elementary units of cellular networks

NATURE REVIEWS | GENETICS VOLUME 5 | FEBRUARY 2004 | 101

- “We find that scale-free networks describe the metabolic networks in all （43）organisms in all three domains of life: Archae（太古生物）, Bacterium（细菌）, Eukaryote （真核生物）, indicating the generic nature of this structural organization.”

Nature 407-MetabolicNet

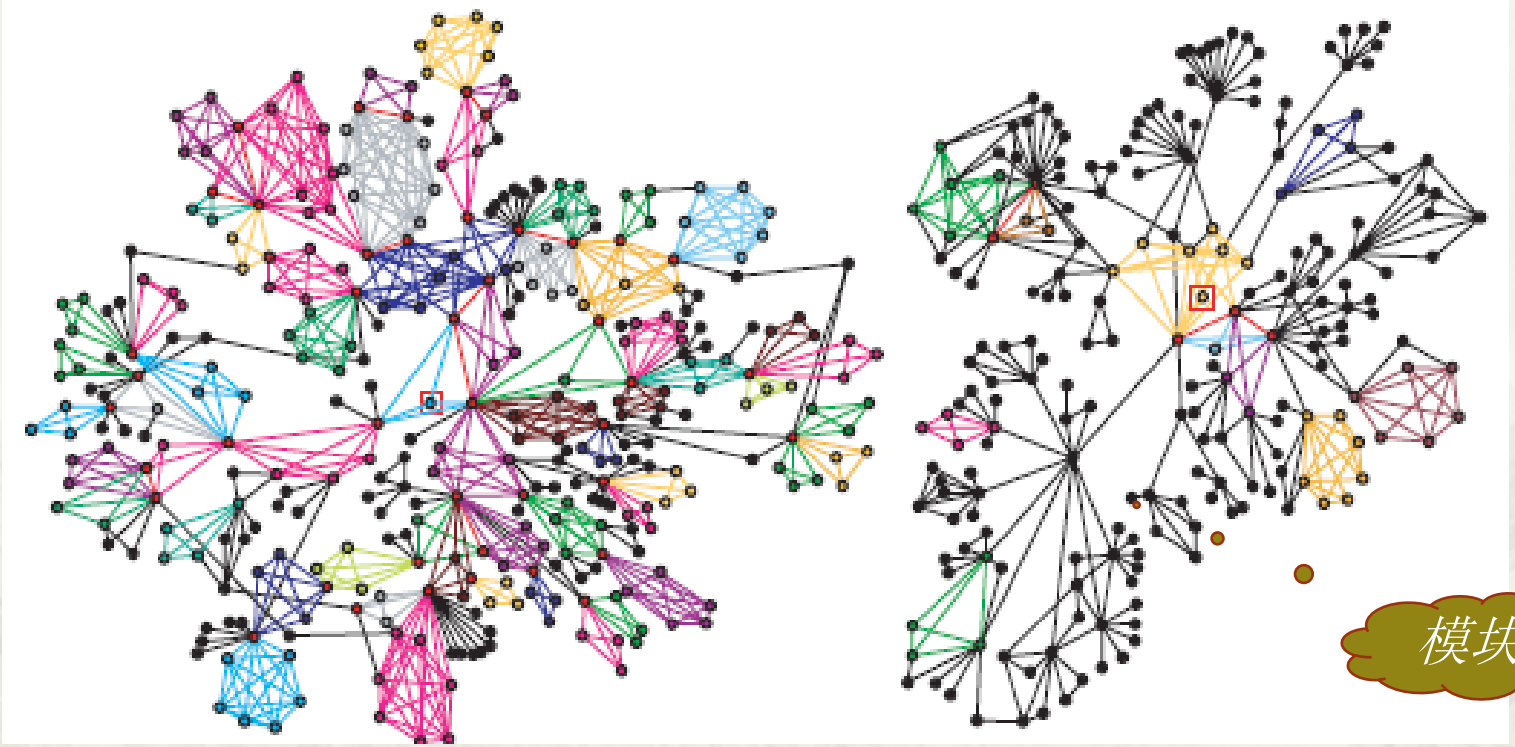
C. Community Structure

- 复杂网络中存在模块或者社区结构 (Module or Community structure)
- 模块或者社区定义为网络中内部连接稠密，与外部连接稀疏的节点的集合 (Filippo Radicchi et. al. *PNAS*, Vol.101, No.9, *2658-2663,* 2004).
- 数学表述: $$\sum_{i \in V} k_i^{\text{in}}(V) > \sum_{i \in V} k_i^{\text{out}}(V)$$

其中V是子图，K是顶点的度。即子图 *V*是模块的条件是模块内顶点的内部连边的度值之和大于模块内顶点的外部连边的度值之和。

PNAS — Proc. Natl. Acad. Sci. USA 美国科学院院刊

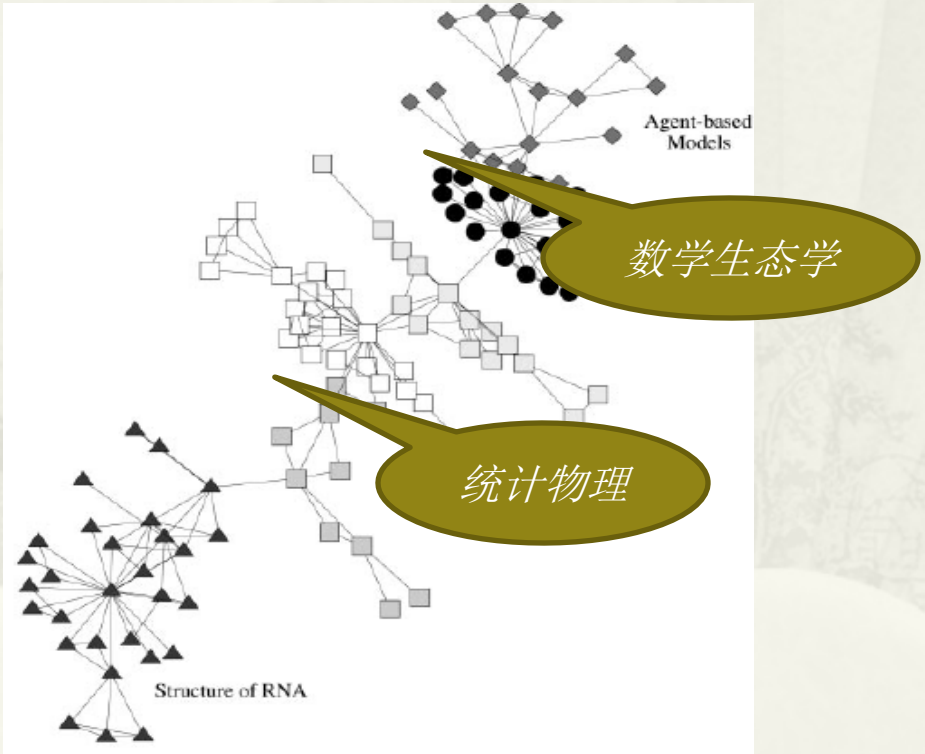
复杂网络的模块性质



发表文章合作网　　电话网络

模块划分的重要性

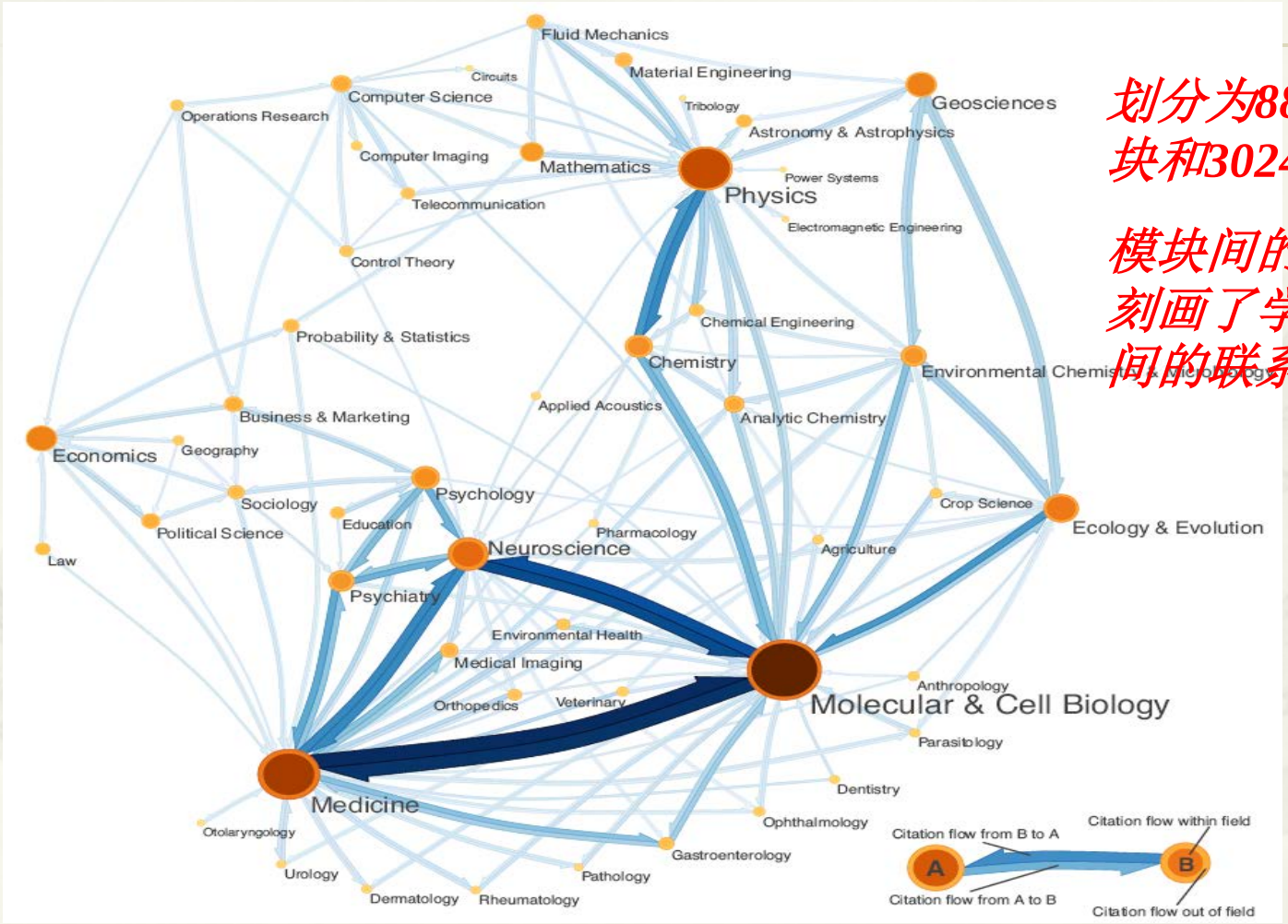
- 许多复杂网络共有的性质。
- 研究模块结构有助于研究整个网络的结构和功能



圣塔菲研究所的科学家合作网：模块代表从事相似领域研究的科学家集合

Martin Rosvall, Carl T. Bergstrom, PNAS, vol. 105, no.4. 1118-1123, 2007

自然科学论文引用网络：6128期刊, 约600万次引用，



划分为88个模块和3024条

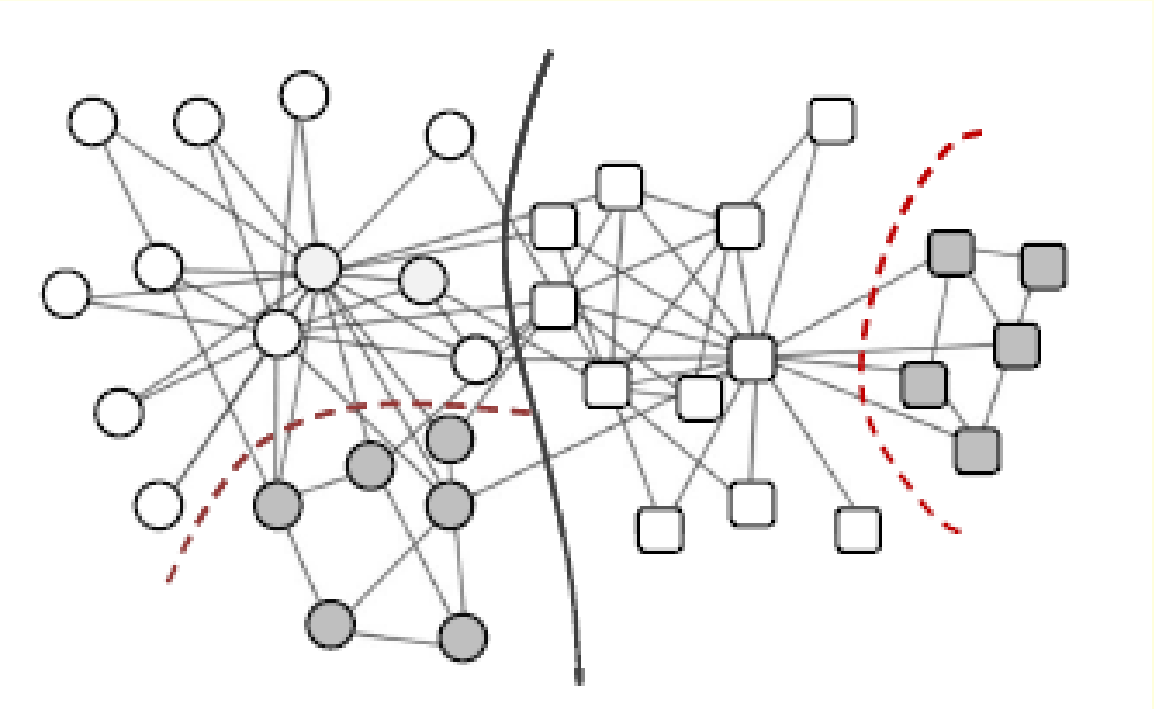
模块间的连接，刻画了学科之间的联系

Modularity: where subsystems are physically or functionally insulated so that failure in one module does not spread to other parts and lead to system-wide catastrophe.

Systems Biology: A Brief Overview
Hiroaki Kitano, *Science* 2002, vol. 295

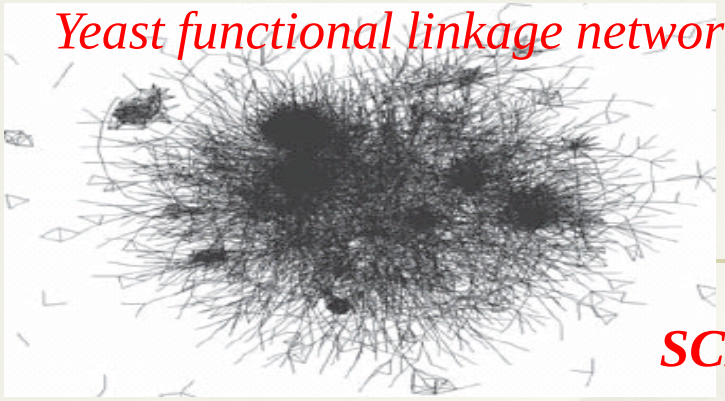
一个社会网络的例子

*W. W. Zachary, An information flow model for conflict and fission in small groups, Journal of Anthropological Research **33**, 452-473 1977*

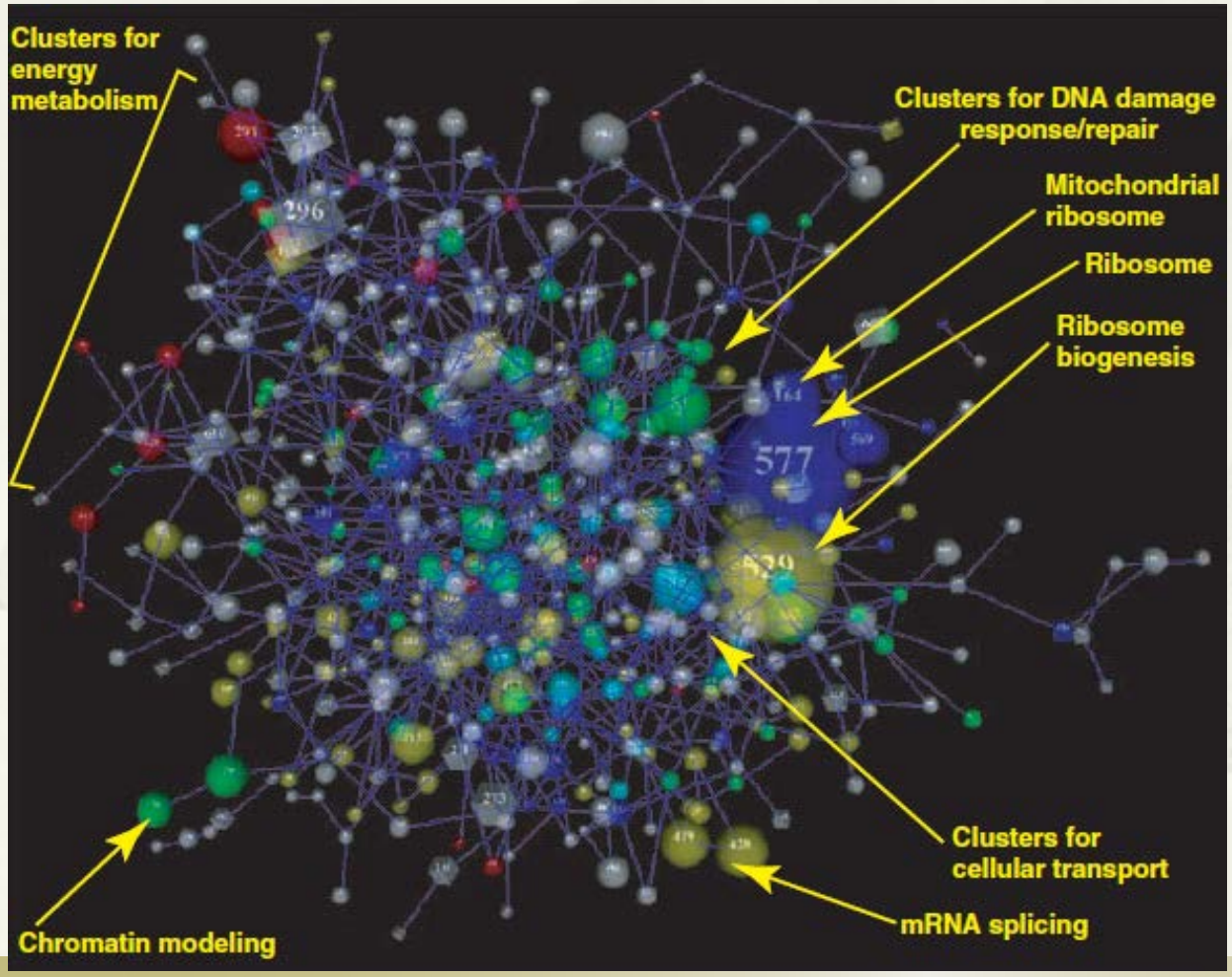


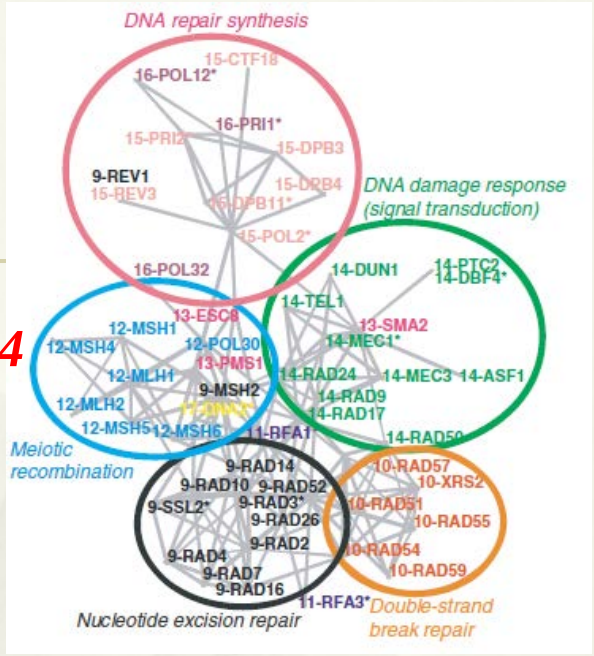
- *1970年美国大学里的一个空手道俱乐部关系网络：节点是其34名成员，边是他们两年间的友谊关系，边数为78。俱乐部里的矛盾导致其分裂为两个小的俱乐部。问题是能否用网络的模块结构来重现这个过程？*
- *它是模块探测研究中的经典例子。*

Yeast functional linkage network



SCIENCE Vol 306(26) 2004





DNA damage module

564 modules are shown

Connected by the 950

strongest inter-module

linkages.

Problem Definition (continued)

* Given a network/graph $N = (V, E)$, partition N into several sub-networks which satisfy module conditions

* As we said before, a popular (descriptive but not mathematical) module definition is

  The nodes in a module are densely linked but nodes in different modules are sparsely linked

Filippo Radicchi et. al. *Proc. Natl. Acad. Sci. USA (PNAS)*, Vol.101, No.9, *2658-2663,* 2004

* Mathematically, let

$$d_i = d_i^{in} + d_i^{out}$$

then the condition for a sub-network $N_k = (V_k, E_k)$ being a module is

$$\sum_{i \in V_k} d_i^{in} - \sum_{i \in V_k} d_i^{out} > 0$$

which is usually called a weak definition.

Filippo Radicchi et. al. *Proc. Natl. Acad. Sci. USA (PNAS),* Vol.101, No.9, *2658-2663,* 2004

Model/Complexity

Problem definition: *Given a network, the community identification problem is to partition the network into as many non-overlapping subnetworks as possible such that each subnetwork satisfies a given community definition.*

In the following of this paper, we use the weak community definition (1) to set up a mathematical model. For other community definitions there would be a similar framework. Let n be the number of nodes in the network and it is also the maximum number of possible communities, and L be the number of edges in the network. z_{lk} denotes whether the edge e_l belongs to the k-th community, $l = 1, 2, \ldots, L, k = 1, 2, \ldots, n$, where $e_l = (v_i, v_j)$ represents the edge connecting the nodes v_i and v_j. Let x_{ik} be a binary variable indicating whether the node v_i belongs to community k. Then the relations between z_{lk} and x_{ik}, x_{jk} can be described as:

$$z_{lk} \leq x_{ik} \quad \text{and} \quad z_{lk} \leq x_{jk}$$

which indicate that if one adjacent node of an edge is not in community k, then this edge definitely does not belong to the community. We use

$$x_{ik} + x_{jk} - 1 \leq z_{lk}$$

to indicate that if both v_i and v_j are in community k, then the edge connecting these two nodes must be in the community. Let y_k be a binary variable denoting whether the k-th community is empty. $y_k = 0$ if and only if the k-th community has no nodes, so

$$y_k \leq \sum_{i=1}^{n} x_{ik} \leq n y_k.$$

The weak community definition condition when $y_k = 1$ can be formulated as:

$$2\sum_{l=1}^{L} z_{lk} > \sum_{j=1}^{n}\sum_{i=1}^{n} x_{ik} a_{ij} - 2\sum_{l=1}^{L} z_{lk}. \tag{5}$$

To incorporate the case when $y_k = 0$, we restate the inequality (5) as follows,

$$2\sum_{l=1}^{L} z_{lk} \geq \sum_{j=1}^{n}\sum_{i=1}^{n} x_{ik} a_{ij} - 2\sum_{l=1}^{L} z_{lk} + y_k. \tag{6}$$

问题可以写成以下的线性整数规划：

$$
\begin{aligned}
&\max \sum_{k=1}^{n} y_k \\
&s.t. \sum_{k=1}^{n} x_{ik} = 1 \quad i = 1, 2, \cdots, n \\
&z_{l,k} \le x_{ik} \\
&z_{l,k} \le x_{jk} \\
&x_{ik} + x_{jk} - 1 \le z_{l,k} \\
&\sum_{i=1}^{n} x_{ik} \ge y_k \\
&\sum_{i=1}^{n} x_{ik} \le M y_k \\
&2\sum_{l=1}^{L} z_{lk} \ge \sum_{j=1}^{n}\sum_{i=1}^{n} x_{ik} a_{ij} - 2\sum_{l=1}^{L} z_{lk} + y_k \\
&x_{ik} \in \{0, 1\}, y_k \in \{0, 1\}, z_{lk} \in \{0, 1\} \\
&i = 1, 2, \cdots, n, k = 1, 2, \cdots, n, l = 1, 2, \cdots, L
\end{aligned}
\tag{6}
$$

3 NP-completeness of the community identification problem

Let $G = (V, E)$ be an undirected graph, and $d_G(v) = |\{u \in V | (u, v) \in E\}|$ denote the degree of the node v in G. A graph G is a cubic graph if $d_G(v) = 3$ for every $v \in V$. Any subset of vertices $S \subseteq V$ creates a cut of G, which is denoted by $C(S, \bar{S}) = \{(u, v) | u \in S, v \in V \setminus S\}$. The size of $C(S, \bar{S})$ is defined as $L_G(S, \bar{S}) = |C(S, \bar{S})|$ and $d_G(S) = \sum_{v \in S} d_G(v) = L_G(S, S) + L_G(S, \bar{S})$ denotes the sum of degrees of the nodes in the subset $S \subseteq V$. To prove the NP-completeness of the problem, we first prove the NP-completeness of a simplest case, that is, partition a network into two subnetworks such that each subnetwork satisfies the weak community definition, which we call as the qualified cut problem. The corresponding decision version of this problem can be formulated as follows:

The Qualified Cut Problem

Instance: An undirected graph $G = (V, E)$.

Question: Is there a subset $S \subset V$ such that $L_G(S, S) > L_G(S, \bar{S})$ and $L_G(\bar{S}, \bar{S}) > L_G(\bar{S}, S)$

Then we show that any instance of the maximum cut problem for cubic graph, which has been proved to be NP-hard (Alimonti and Kann 2000), can be transformed into a qualified cut problem in polynomial time, and the solution of the maximum cut problem for cubic graph exactly corresponds to that of the qualified cut problem. We note that the qualified cut problem is a special case of the decision version for the conductance problem, which has been often stated to be an NP-complete problem in the literature (Šíma and Schaeffer 2006). Thus we can borrow some ideas from the NP-completeness proof from Šíma and Schaeffer (2006) and Shi and Malik (2000). The detailed NP-completeness proof of our qualified cut problem is as follows.

Maximum Cut for Cubic Graph (Max Cut-3)
Instance: A cubic graph $G = (V, E)$ and a positive integer a.
Question: Is there a cut $(B, \bar{B})$, such that $L_G(B, \bar{B}) > a$?

Theorem 1 *The Qualified Cut Problem is NP-complete.*

- There are many heuristic methods to solve this problem.

- Among them a popular method to partition a network into module structure is to define a measure (a quantity) for a given partition, then optimize the measure to find a proper partition.

- The fist measure is called *modularity*. It is defined by Newman and Girvan (*Physical Review E*, 2004)

Modularity *Q* quantitatively evaluates a partition

* Newman and Girvan (*Physical Review E*, 2004) gives a quantitative measure *Q*

$$Q(N_1, \cdots, N_k) = \sum_{i=1}^{k} \left[\frac{|E_i|}{|E|} - \left(\frac{d_i}{2|E|} \right)^2 \right]$$

* where *N1, …, Nk* is a partition of *N*, the term under the summation is the number of edges falling within *Ni* minus the expected number in an equivalent network with edges placed at random.

* Step 1: Fix k $(k = 1, \ldots n)$, $N_1 \cup \ldots \cup N_k = N$ compute

$$\max_{N_1,\ldots N_k} Q(N_1,\ldots,N_k)$$

* Step 2: Compute

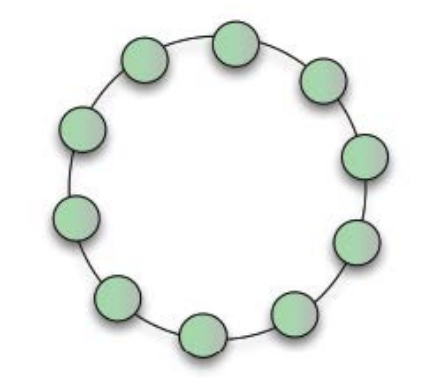
$$\max_{k\in\{1,\ldots n\}} \max_{N_1,\ldots N_k} Q(N_1,\ldots,N_k)$$

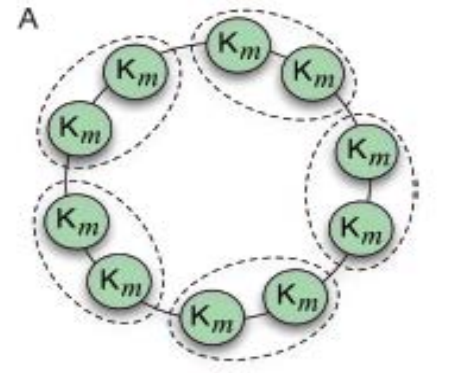
This is a NP-hard problem, then heuristic algorithms including simulation annealing, genetic algorithm are generally used (Newman, *PNAS*, 2006; Guimera, Nature, 2005).

Problems raised:

* Resolution Limit
* Mis-identification

Resolution Limit : Modularity Q fails to identify correct module structure in some case





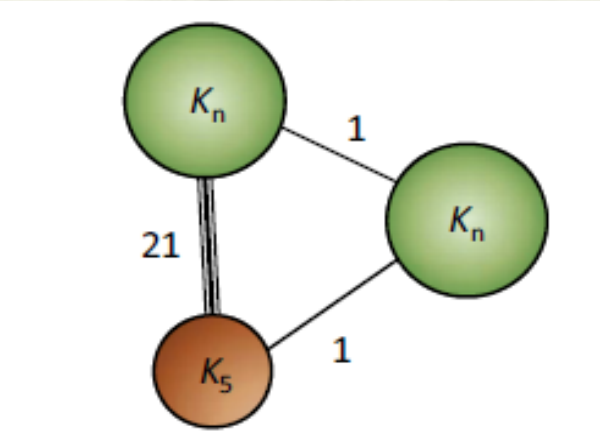
Left: a graph called a ring of cliques , that is, two neighboring cliques are connected by single links.

Right: when the number of cliques is larger than , the modularity optimization gives a partition where two cliques are combined into one community. This phenomenon is called resolution limit.

Fortunato & Barthelemy, *Proc. Natl. Acad. Sci.* (2007)

Misidentification

- Some derived communities do not satisfy the weak community definition or other community definition we can find.



Q partitions the network into three communities (two Kn and one K5) when n>=16 (respectively, n>=21), in which K_5 is a sub-graph violating all reasonable community definition.

Xiang-Sun Zhang, Rui-Sheng Wang, Yong Wang, Ji-Guang Wang, Yu-Qing Qiu, Lin Wang, and Luonan Chen. Modularity optimization in community detection of complex networks.
***Europhysics Letters*, 87, 38002, 2009.**

We suggested a new quantitative measure

* Modularity Density D:

$$D(N_1, \cdots, N_K) = \sum_{i=1}^{K} \left(\frac{2|E_i|}{|V_i|} - \frac{|\bar{E}_i|}{|V_i|} \right)$$

Modularity density D overcomes "resolution limit" problem in the case of the ring of L cliques

Zhenping Li, Shihua Zhang, Rui-Sheng Wang, Xiang-Sun Zhang, Luonan Chen, Quantitative function for community detection.
Physical Review E, 77, 036109, 2008

衡量网络模块化的指标Q值

- Newman 和 Girvan (*Physical Review E*, 2004) 提出一种衡量网络社区结构的指标 *Q* 值

$$Q(P_k) = \sum_{c=1}^{k} \left[\frac{L(V_c, V_c)}{L(V, V)} - \left(\frac{L(V_c, V)}{L(V, V)} \right)^2 \right]$$

指标*Q*的问题 (Resolution limit)

Fortunato and Barthélemy, *PNAS*, 2007

- 利用Q划分网络的计算步骤:

  1. 固定要分成的模块数 k, 将网络 N 分成 k 块 $N_1, ..., N_k$, 使 $Q_k = Q(N_1) + ... + Q(N_k)$ 最大
  2. 对 $k = 1, ..., n$, 求 k^* 使 Q_{k^*} 最大

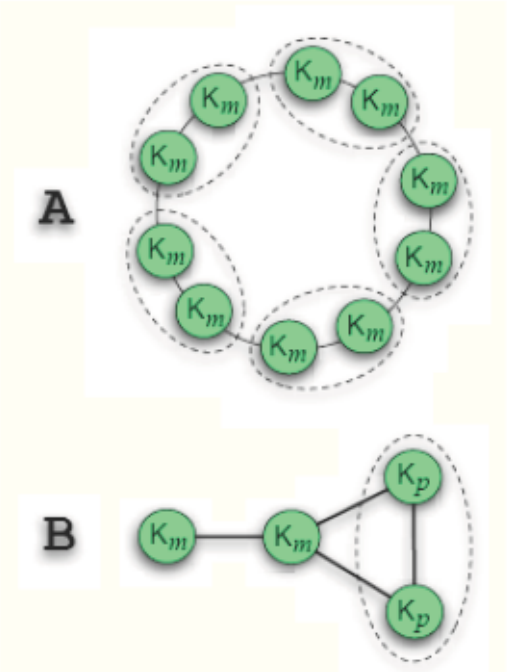
- 目前很大一部分模块探测的方法集中于利用各种启发式算法来极大化Q值，例如模拟退火、遗传算法等(Newman, *PNAS*, 2006; Guimera, Nature, 2005).

- Q 值依赖于网络的规模，及网络的总边数.

- 无法正确识别一些明显的模块，例如一个完全子图

极端例子：ring of cliques

Problems, or not?



Modules indistingisable via Optimization of modularity

$$l_S < 2l_R^{min} = \sqrt{2L}.$$

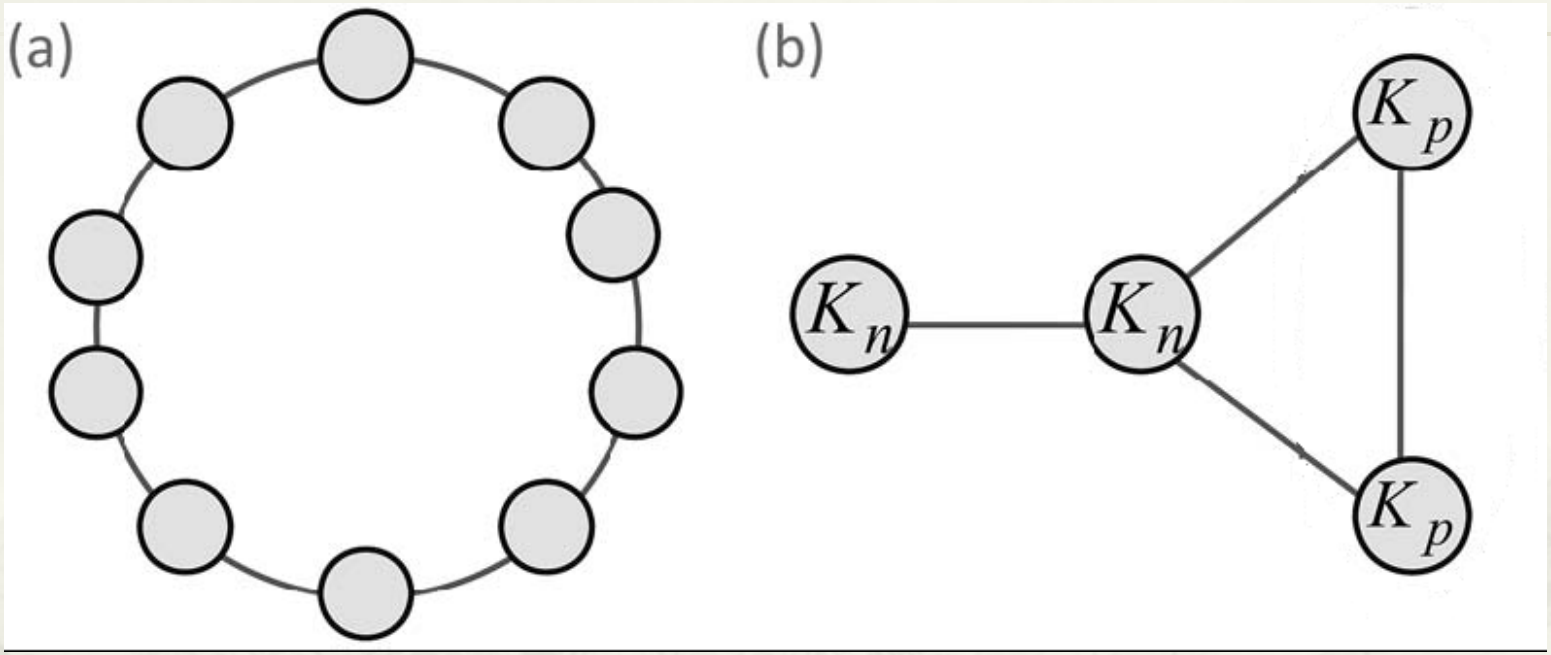
Fortunato & Barthelemy, Proc. Natl. Acad. Sci. USA 104 (1), 36-41 (2007)

提出新的模块化指标D值

- 模块化密度函数 D:

$$D(P_k) = \sum_{c=1}^{k} \frac{L(V_c, V_c) - L(V_c, \overline{V}_c)}{|V_c|}$$

Zhenping Li, Shihua Zhang, Rui-Sheng Wang, Xiang-Sun Zhang, Luonan Chen, Quantitative function for community detection. Physical Review E, 77, 036109, 2008



D值克服了Q值存在的 resolution limit 问题

结果

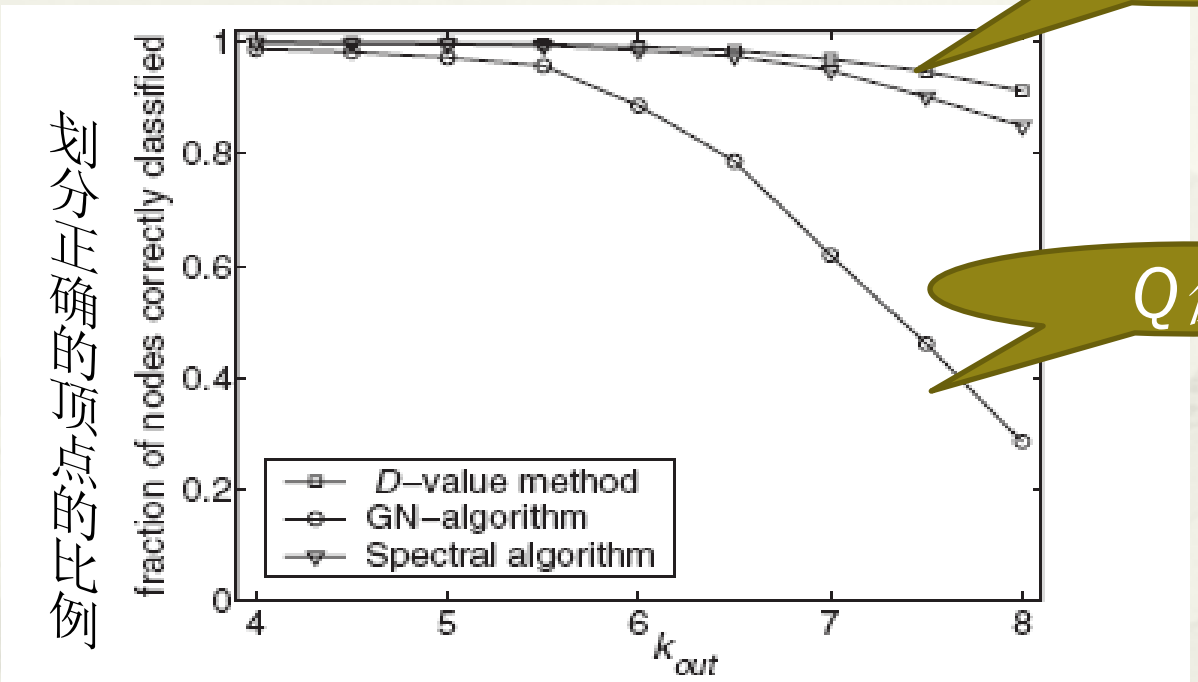


FIG. 2. Test of various methods on computer-generated networks with known community structures. It is a plot of the fraction of nodes correctly classified with respect to k_{out}. Each point is an average over 100 realizations of the networks.

(五)

结　　论

OR在Bioinfomatics研究中的作用

- 优化原则在生物进化中的体现：
  - 俭简原则(Parsimony, combinatorial optimization)在进化树形成、单核苷酸多态性等研究中的应用
  - 能量函数概念在蛋白质折叠、蛋白质比对中的应用（Minimization, continuous optimization)
- 优化理论和算法在生物数据挖掘中的重要性：
  - 计算复杂性理论的指导做作用
  - DP、MP、LP、ILP、Graph Theory、GA、BB、NN、SVM等在生物研究中的频繁使用

EWG CBBM - The Operational Research in Computational Biology, Bioinformatics and Medicine Working Group of EURO

- The goal of *EWG CBBM* is to promote and to facilitate communication links among European (and other) researchers working in areas of operational research in computational biology, bioinformatics and Medicine.
- *EWG CBBM* was established with numerous founding members at a satellite meeting of the *EURO XXI 2006 Conference* in Iceland, There after *EWG CBBM* is organized / co-organized in numerous workshop on Workshop on Networks in Computational Biology in Ankara, Turkey (September 10-12, 2006).
- Last year in Prague, we celebrated our first annual meeting by organizing Workshop on OR in Computational Biology, Bioinformatics and Medicine in Prague, Czech Republic, July 2007.

中国运筹学会计算系统生物学分会

Computational Systems Biology Society

within

Operations Research Society of China

Trends in Commercial Bioinformatics

BIOTECHNOLOGY REVIEW *13 MARCH 2000*

- The spectacular rise of the commercial genomics industry has created a commercial market for bioinformatics software, hardware and services.
- By some estimates, the total market for bioinformatics tools and services, including custom databases, could exceed $2.0 billion within five years.
- In our opinion, bioinformatics technology will become an increasingly important competitive differentiator for public and private life science companies going forward.
- Bioinformatics is becoming a directly investible theme. By our estimation, there are now more than 50 companies which offer bioinformatics products and services. Most of these are private companies, but we would not be surprised to see a number of the more mature players go public in the next 12 months.

结论（continue)

- 生物信息学（Bioinformatics）为应用数学的理论研究和实际应用打开了一扇通向新探索和新创造机会的大门。
- 随着人们对生物信息属性的了解越来越深入、信息越来越翔实，确定型方法逐渐受到重视，成为传统上一直占主导地位的概率、统计方法的有力补充。

Thank you!

若要得到我们的研究工作的更多的信息，请使用以下网址

http://zhangroup.aporc.org