计算系统生物学

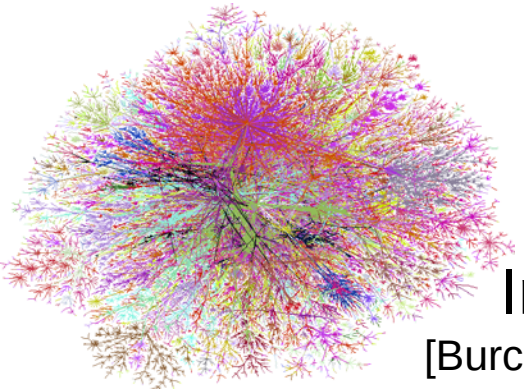
张世华

中国科学院数学与系统科学研究院

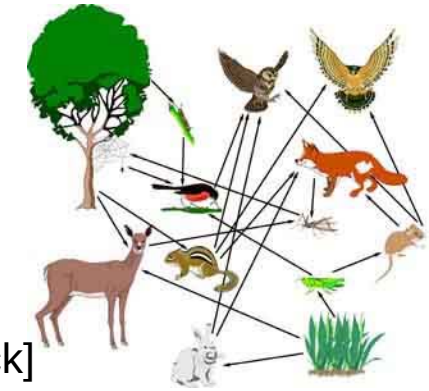
Modular structure of biological networks
Community structure of complex networks

Shihua Zhang

Networks as a universal language

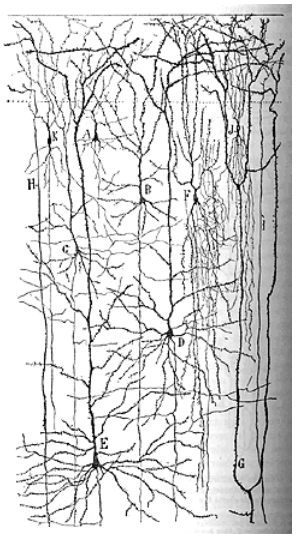


Internet
[Burch & Cheswick]

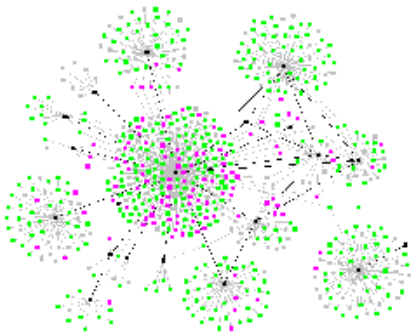


Food Web

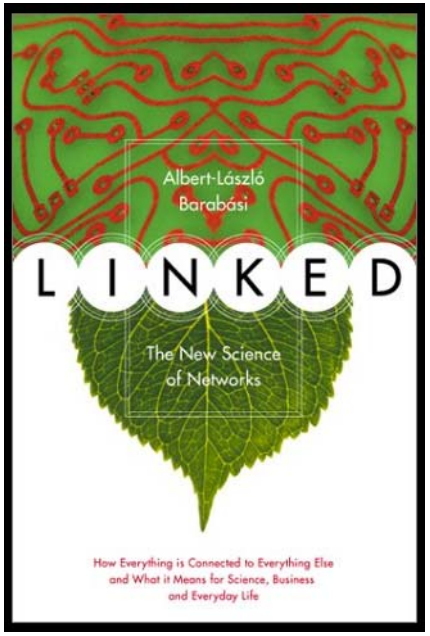
Electronic Circuit



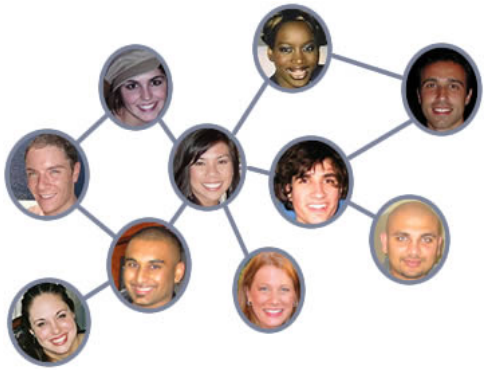
Neural Network
[Cajal]



Disease Spread
[Krebs]



Protein Interactions
[Barabasi]



Social Network

生物分子网络的模块结构——复杂网络的模块化性质

- 复杂网络包括生物网络中存在模块或者社区结构 **(Module or Community structure)** **(Girvan and Newman, PNAS，2002；Rives and Galitski，PNAS, 2003)**

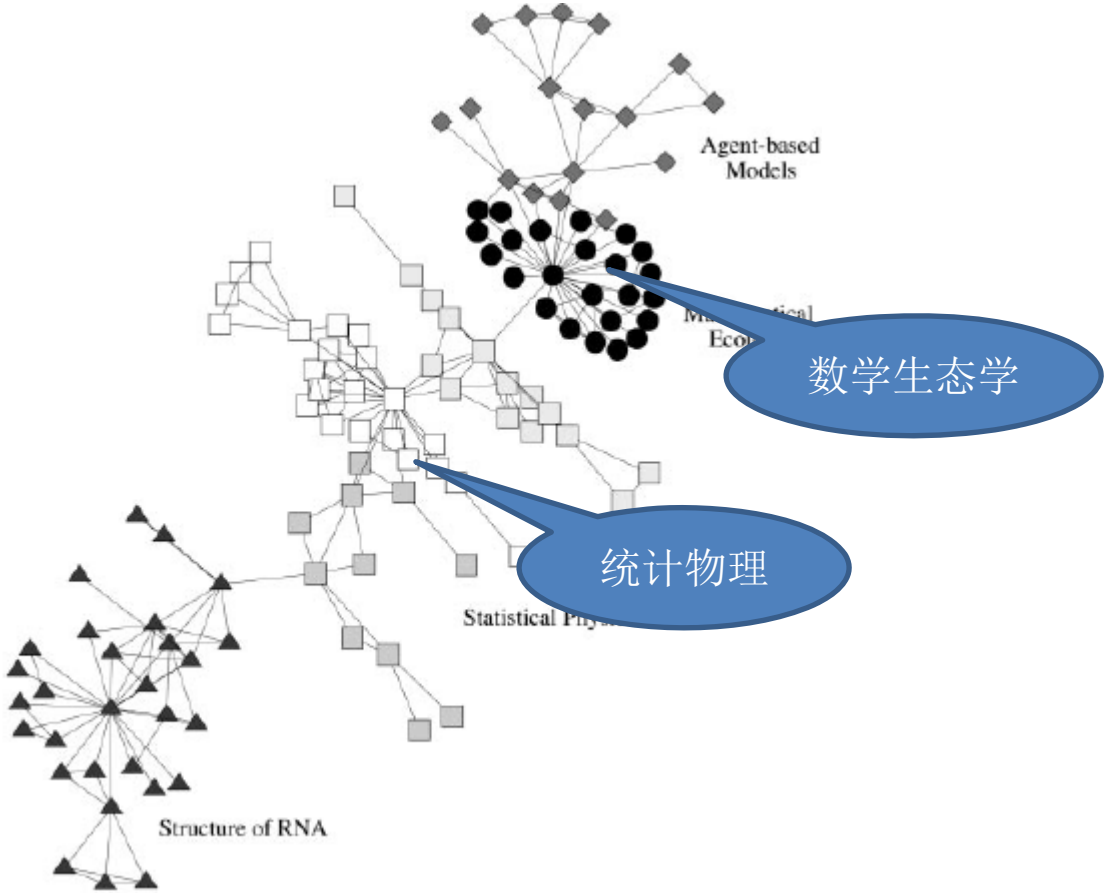
- 模块或者社区定义为网络中内部连接稠密，与外部连接稀疏的节点的集合 **(Radicchi *et. al.* PNAS, 2004).**
- 数学表述:

$$\sum_{i \in V} k_i^{\text{in}}(V) > \sum_{i \in V} k_i^{\text{out}}(V)$$

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

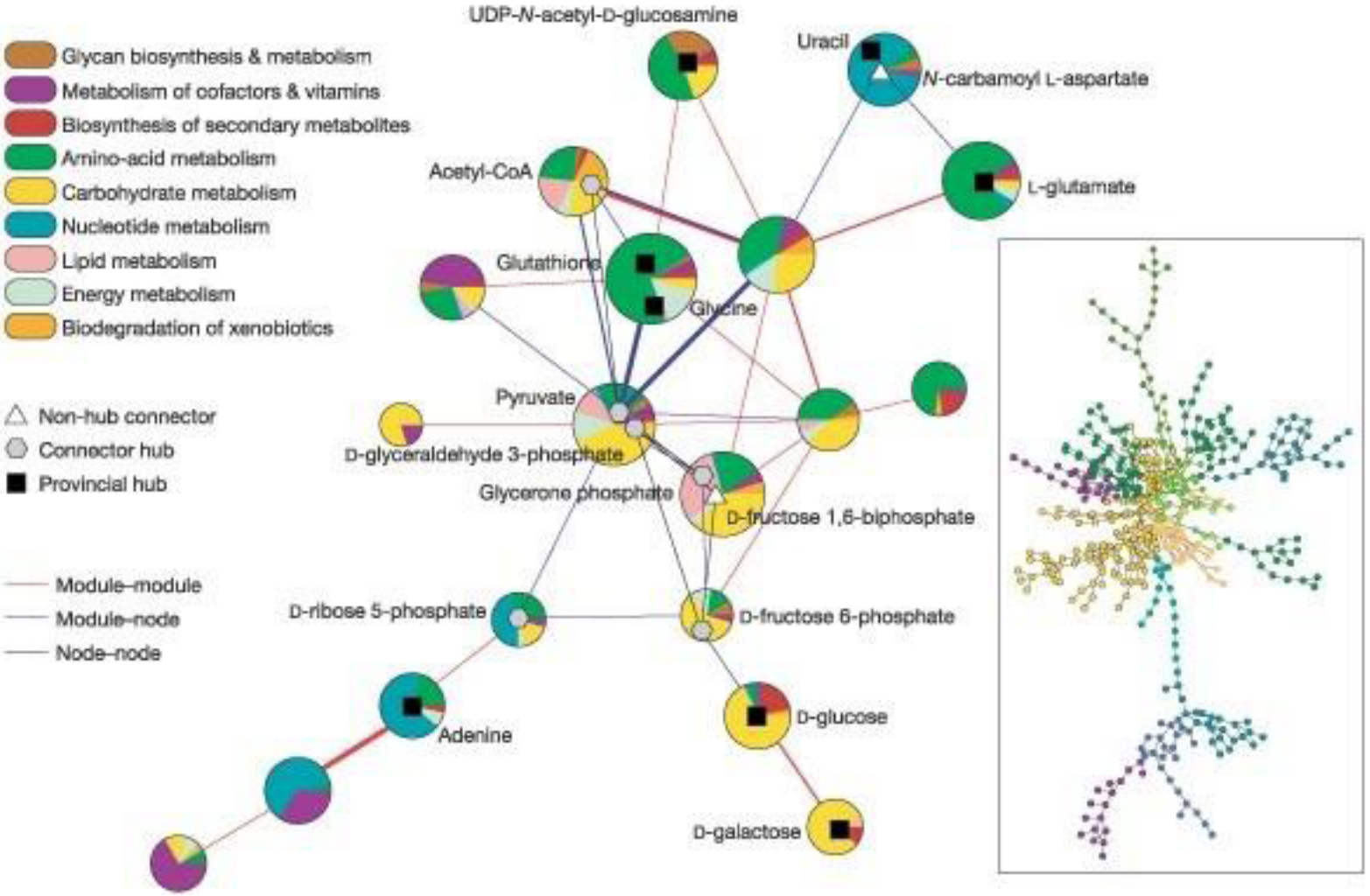
社团(模块)结构识别的重要性

- 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



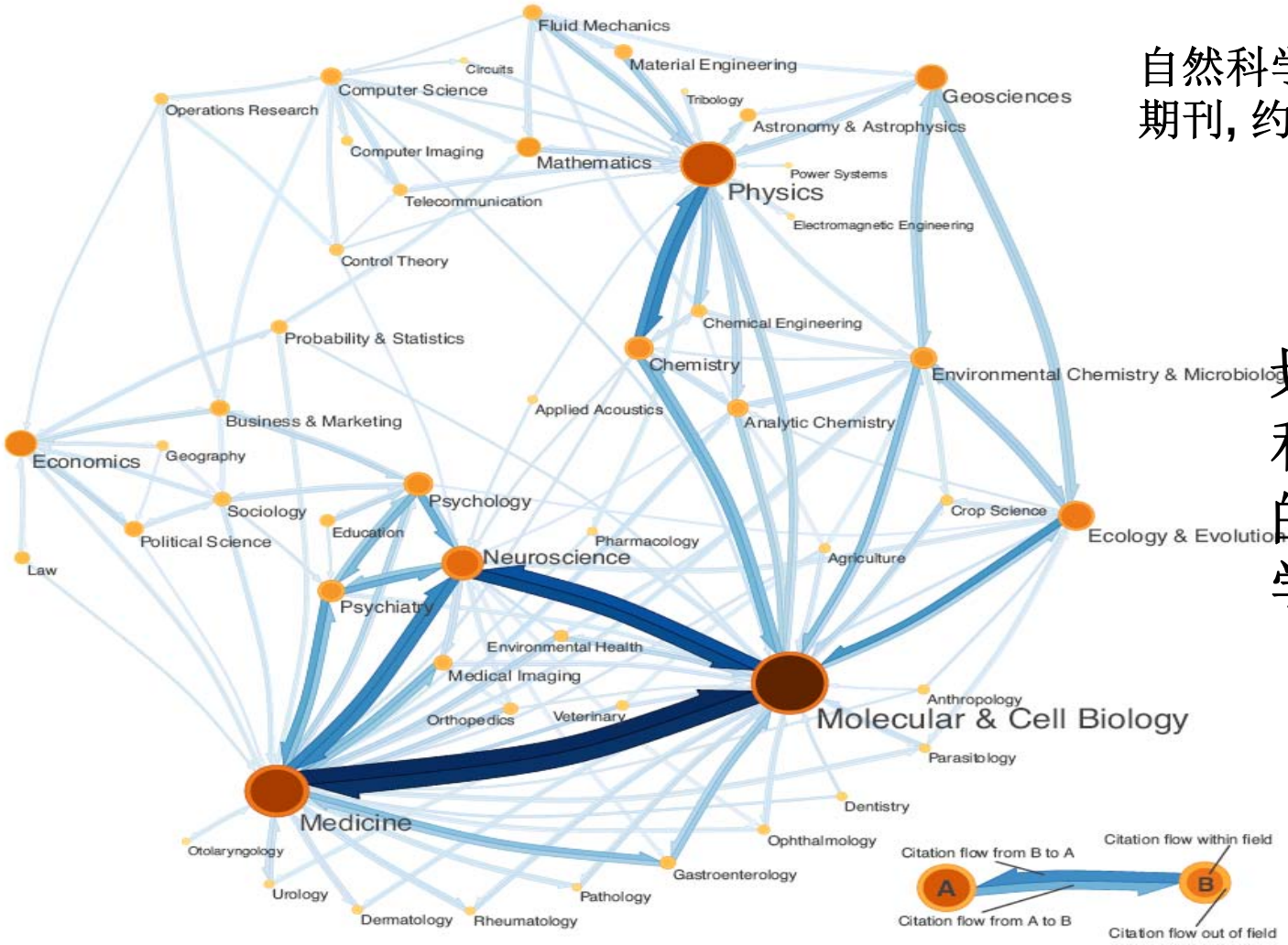
圣塔菲研究所的科学家合作网：模块代表从事相似领域研究的科学家集合

社团(模块)结构识别的重要性



Guimera and Amaral, *Nature*, 2005

社团(模块)结构识别的重要性



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

Rosvall and Bergstrom, PNAS, 2007

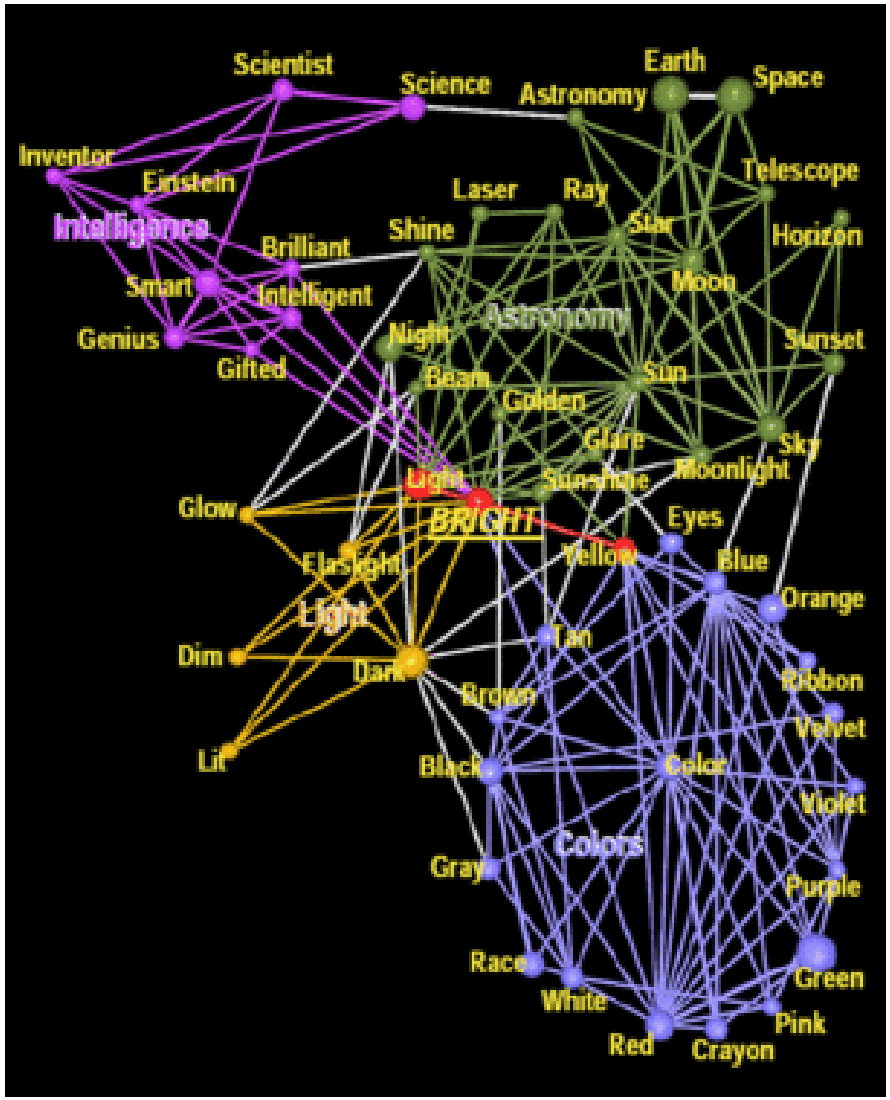
这一课题引起广泛关注！

- Girvan, M, Newman, M., ***PNAS,*** 2002
- Ravasz, E, Somera, A, Mongru, D, et al., ***Science***, 2002
- Radicchi, F, Castellano, C, Cecconi, F., ***PNAS,*** 2004
- Guimera, R, Mossa, S, Turtschi, A., ***PNAS***, 2005
- Guimera, R, Amaral, L., ***Nature***, 2005
- Palla *et al.,* ***Nature***, 2005
- Newman, M., ***PNAS,*** 2006
- Rosvall, M, Bergstrom, C., ***PNAS***, 2007
- Fortunato, S, Barthelemy, M., ***PNAS***, 2007
- Weinan, E, Li, T, Vanden-Eijnden, E., ***PNAS***, 2008
- Rosvall, M, Bergstrom, C., ***PNAS***, 2008
- Peter J. Mucha, *et al.,* ***Science,*** 2010
- Ahn, Y.Y, Bagrow, J.P. & Lehmann, S., ***Nature***, 2010

社团结构探索方法概论

Santo Fortunato, Physics Reports, 2010.

PPI网络的模块结构到复杂网络的社团结构(2005)



- 基于**k-clique**渗透理论

  overlapping community structure

- 可是当时三个物种的PPI网络中，Yeast 相对比较稠密

  Fly 非常稀疏，几乎没有什么clique

Palla *et al.*, Nature, 2005

Fuzzy/overlapping社团结构识别

基本想法：

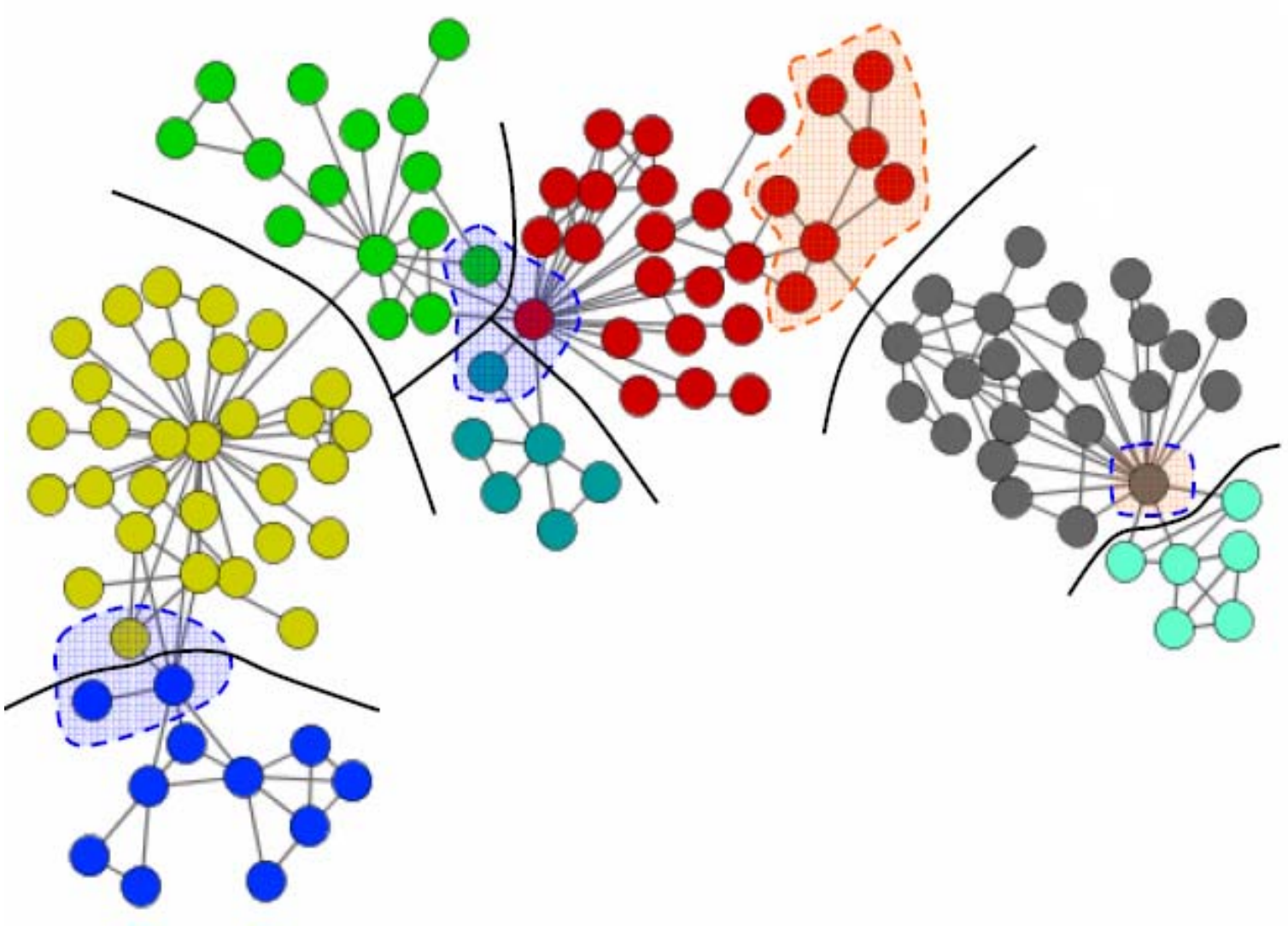
1. 推广Modularity指标如Q，到Fuzzy划分的评价；

2. 考虑传统的Fuzzy clustering方法。

S. Zhang, R.S. Wang, and X.-S Zhang. Physica A, 2007.

***S. Zhang, R.S. Wang, and X.-S Zhang. Physical Review E,* 2007.**

算例——科学合作网络



Graph kernels, hierarchical clustering, and network community structure: experiments and comparative analysis

S. Zhang, X.M. Ning, and X.-S Zhang. EPJB, 2007.

衡量网络模块化的指标Q值

- 设网络为 **N=(V,E), P_k = {(V_1, E_1), …, (V_k, E_k)}** 为一个分划。**L(V_i, V_j) =|E_{ij}|, i in V_i, j in V_j.**

- **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:

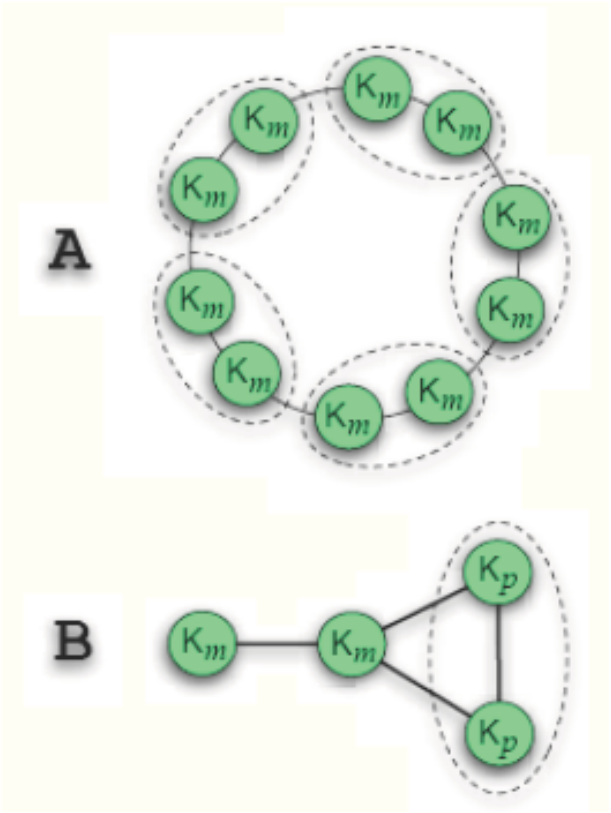
$$\max_k \overline{Q_k} = \max_k \max_{\sum_{i=1}^{k}|V_i|=n} \sum_{i=1}^{k} Q_i$$

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

- **Resolution limit** 现象

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, PNAS, 2007.

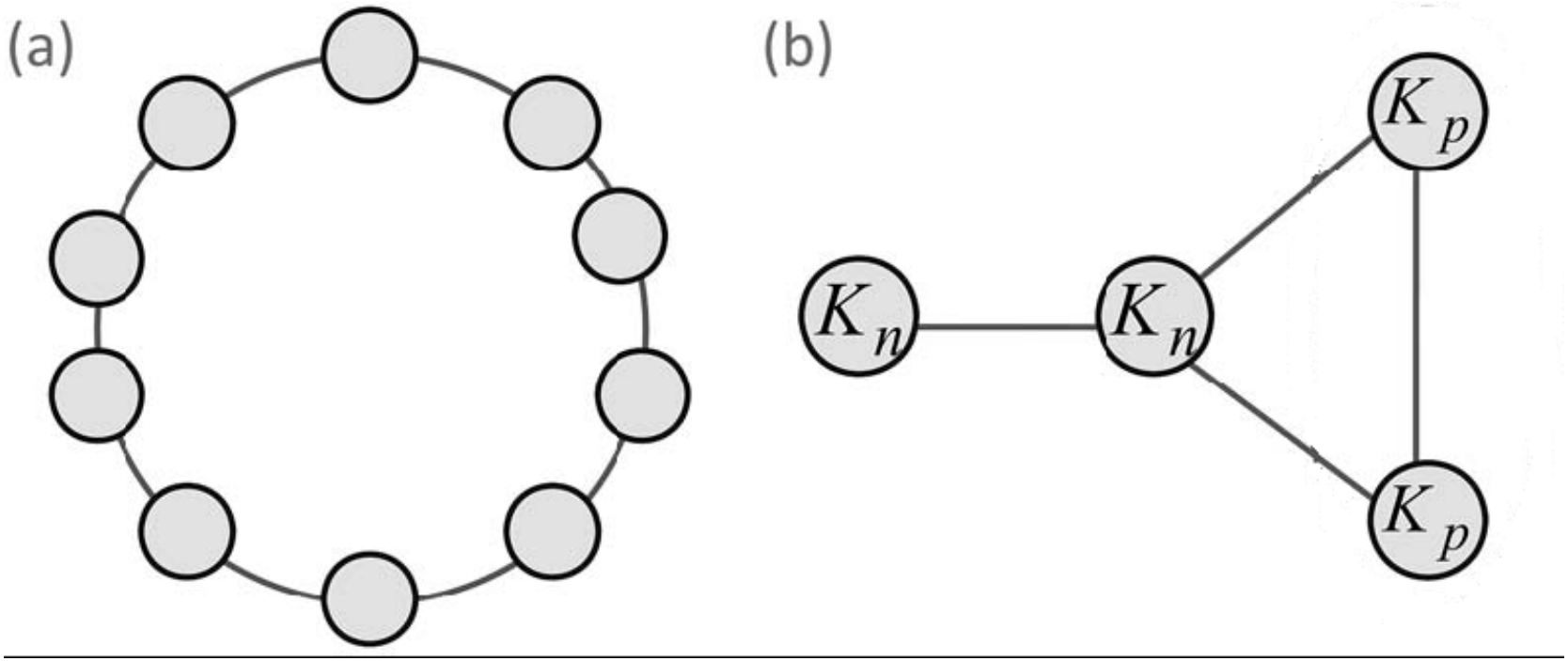
提出新的模块化指标D值

- 模块化密度函数 **D:**

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

- 一个整数规划模型求解：

Z. Li*, S. Zhang*, R.S. Wang, X.-S. Zhang, L. Chen.
***Physical Review E*, 77, 036109, 2008**



D值克服了Q值存在的 resolution limit 问题

TABLE I. Benchmark performance for symmetric and asymmetric group detection measured as fraction of correct assignments, averaged over 100 network realizations with the standard deviation in parentheses.

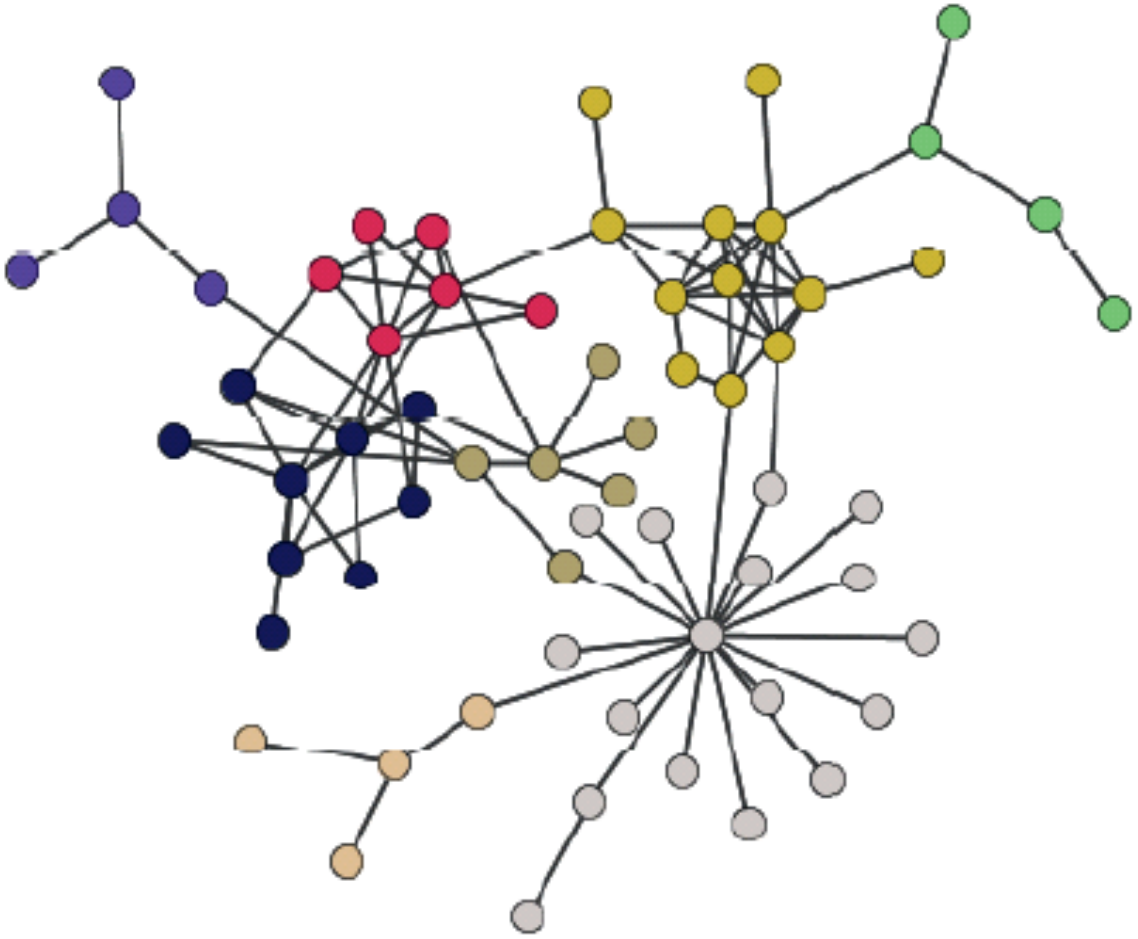
Group	k_{out}	Compression	Q	D value
Symm.	6	0.99 (0.01)	0.99 (0.01)	0.99 (0.01)
	7	0.97 (0.02)	0.97 (0.02)	0.97 (0.02)
	8	0.87 (0.08)	0.89 (0.05)	0.91 (0.03)
Node asymm.	6	0.99 (0.01)	0.85 (0.04)	0.99 (0.01)
	7	0.96 (0.04)	0.80 (0.03)	0.98 (0.02)
	8	0.82 (0.10)	0.74 (0.05)	0.94 (0.03)
Link asymm.	2	1.00 (0.00)	1.00 (0.01)	1.00 (0.00)
	3	1.00 (0.00)	0.96 (0.03)	1.00 (0.00)
	4	1.00 (0.01)	0.74 (0.10)	0.99 (0.01)

多尺度度量

$$D_\lambda = \sum_{i=1}^{m} \frac{2\lambda L(V_i, V_i) - 2(1-\lambda) L(V_i, \bar{V}_i)}{|V_i|}.$$

D指标在生物网络上的应用

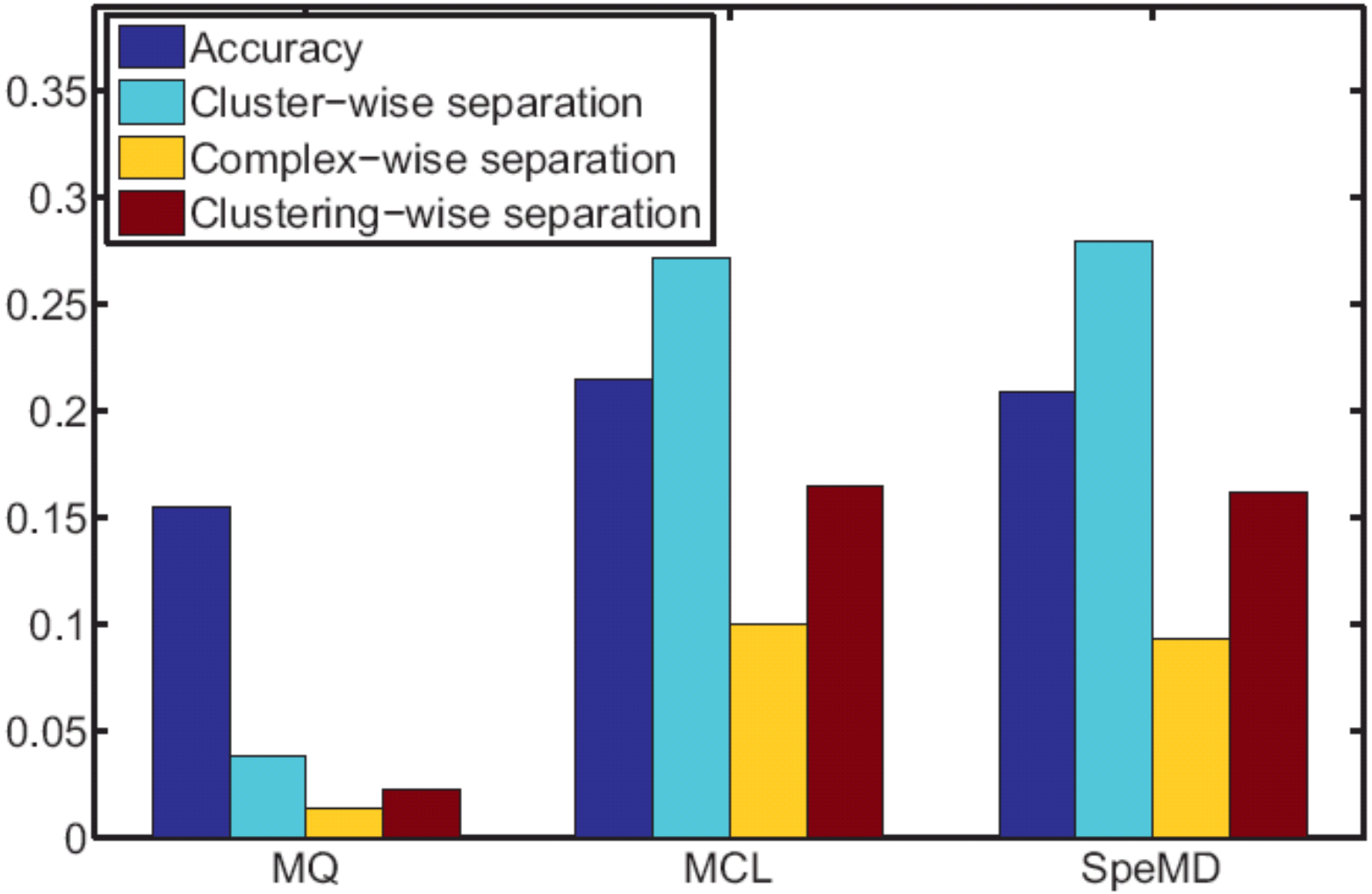
针对D指标



解！

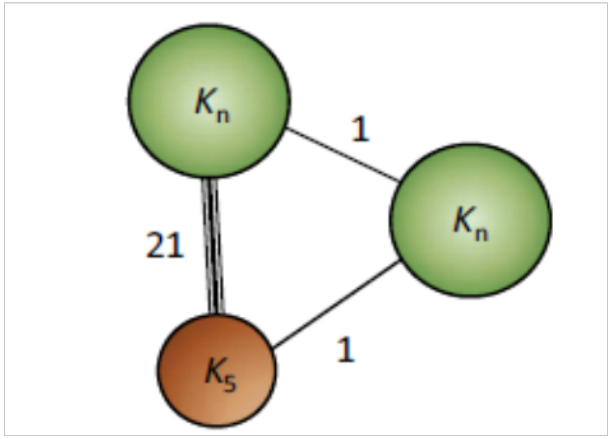
S. Zhang, X.M. Ning, C. Ding and X.-S Zhang. BMC Systems Biology (2010)

生物复合物作为benchmark的比较



进一步Prof. Zhang运用运筹学的方法分析了错分现象---Misidentification

- 用Q或D作优化可能得到不满足定义的模块



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.

X.S. Zhang, R.S. Wang, Y. Wang, J. G. Wang, Y. Q Qiu, L. Wang, and L.Chen. Europhysics Letters (EPL), 2009.

被评为 **EPL 2009 best paper**

- 解析解表明，对这两个经典的算例，的现象产生，所以Q 和D均只Q和D都有**Resolution limit**和**Misidentification**是近似的定量评估函数。

- 网络社团划分的问题可以用一个优化问题来精确描述，我们证明了这一模型是**NP-hard**的。

- 我们相信用优化理论可以彻底解决网络社团划分的问题。网络科学是运筹学的下一个热点。

为了解决这些问题

- 提出一个新的 **OR** 模型和相应的算法，这一算法不会产生**resolution limit** 和 **mis-identification** 现象

X.S. Zhang, Z. Li, R.S. Wang, Y. Wang.
Journal of Combinatorial Optimization, 2011.

定义可以用一个整数线性规划来描述

$$\max \sum_{k=1}^{n} y_k$$

$$s.t. \sum_{k=1}^{n} x_{ik} = 1 \quad i = 1, 2, \cdots, n$$

$$z_{l,k} \leq x_{ik}$$

$$z_{l,k} \leq x_{jk}$$

$$x_{ik} + x_{jk} - 1 \leq z_{l,k}$$

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

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

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

$$x_{ik} \in \{0,1\}, y_k \in \{0,1\}, z_{lk} \in \{0,1\}$$

$$i = 1, 2, \cdots, n, k = 1, 2, \cdots, n, l = 1, 2, \cdots, L$$

- 我们证明了这个模型是 **NP-hard .**

困惑依旧

- 社团结构识别问题为什么成为这么热闹的研究领域？
- 单纯从算法的角度是否有进一步的必要？
- 我们是否可能找到一个万能有效的方法/指标？
- 出路在那里？

进一步的问题/方向？

- 社团结构的演化？

  比如：时序网络中社团结构的识别？

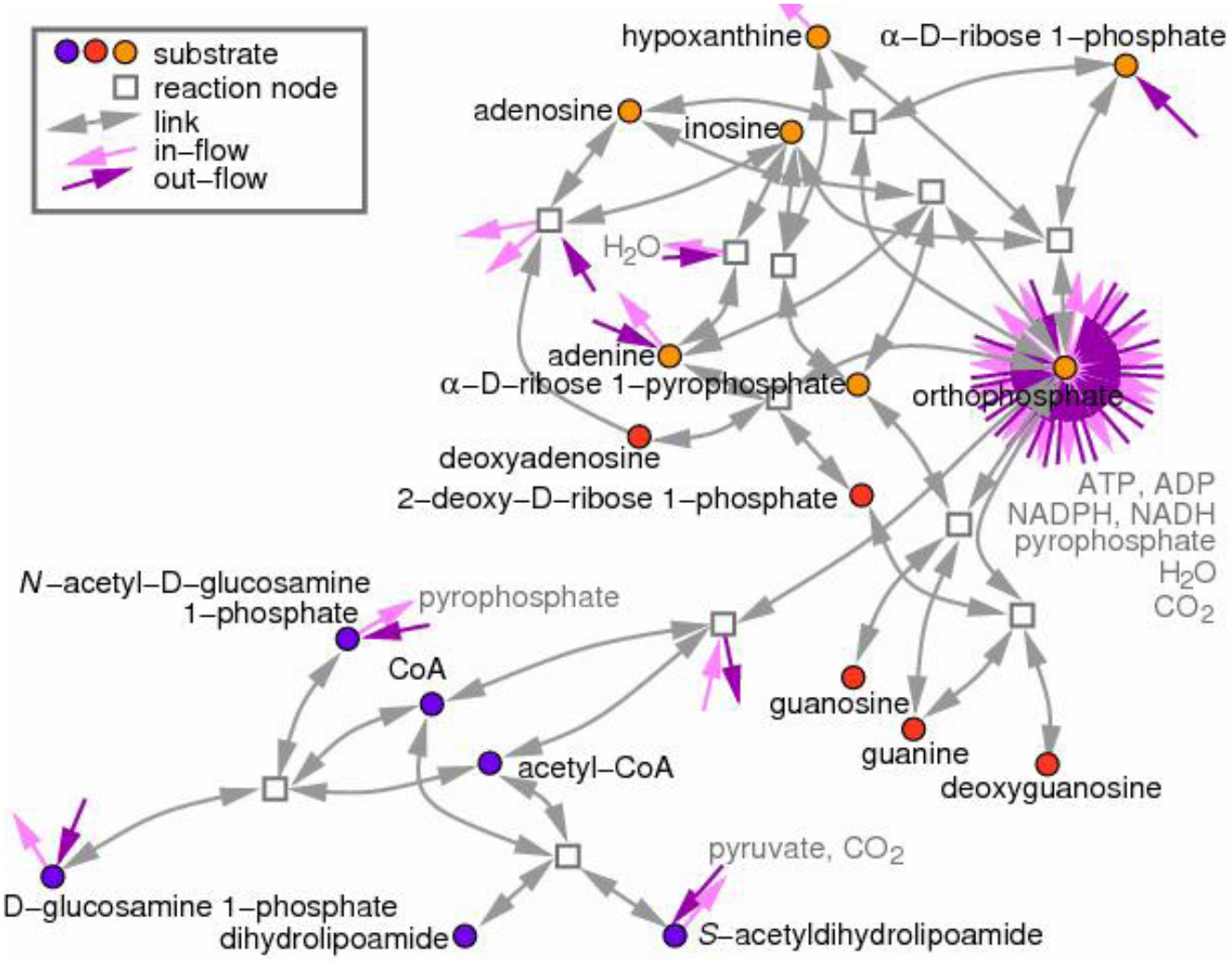
  。 。 。

- 社团结构对其他属性的影响？

  比如：疾病/谣言/信息传播的影响？

- 有向网络社团结构的定义、识别，是否可能？

一个有向网络


substrate
reaction node
link
in–flow
out–flow
hypoxanthine
α–D–ribose 1–phosphate
adenosine
inosine
H2O
adenine
α–D–ribose 1–pyrophosphate
orthophosphate
deoxyadenosine
2–deoxy–D–ribose 1–phosphate
ATP, ADP
NADPH, NADH
pyrophosphate
H2O
CO2
N–acetyl–D–glucosamine
1–phosphate
pyrophosphate
CoA
guanosine
guanine
deoxyguanosine
acetyl–CoA
pyruvate, CO2
D–glucosamine 1–phosphate
dihydrolipoamide
S–acetyldihydrolipoamide

谢谢大家！

欢迎访问 **ZHANGroup,**

http://zhangroup.aporc.org