



中科院计算所
INSTITUTE OF COMPUTING TECHNOLOGY, CAS

关于社会信息网络模型及应用研究的 几点思考

程学旗, cxq@ict.ac.cn

中国科学院计算技术研究所

YOCSEF 2006.10.27

大纲

- 社会学与社会计算
- 社会信息网络的基本概念及其现实意义
- 社会信息网络模型研究
- 我们的工作
- 总结

社会学的基本范畴

- 奥古斯特·孔德 (Auguste Comte, 1798-1857) 创造了社会学这个词汇
 - 最初创立了“社会物理学 (social physics)”; 由于与他同时代的论敌也采用这一术语, 为了将自己的观点区别开来, 创造了“社会学 (sociology)” 词汇
 - 定义: 社会学是对人类社会, 尤其是现代的工业化体系的系统研究, 研究对象是人类自身的行为以及人类群体所表现出来的团体行为
- 当前社会学的研究范畴
 - 大众传媒与传播
 - 城市与城市空间, 人口增长与生态危机
 - 工作与经济生活
 - 阶级、社会分层与不平等; 贫困、福利与社会排斥
 - 犯罪与越轨行为; 种族、族群与移民
 - 文化; 变化; 日常生活; 性别与性; 身体社会学; 家庭; 现代组织; ; 政府与政治; 教育; 宗教

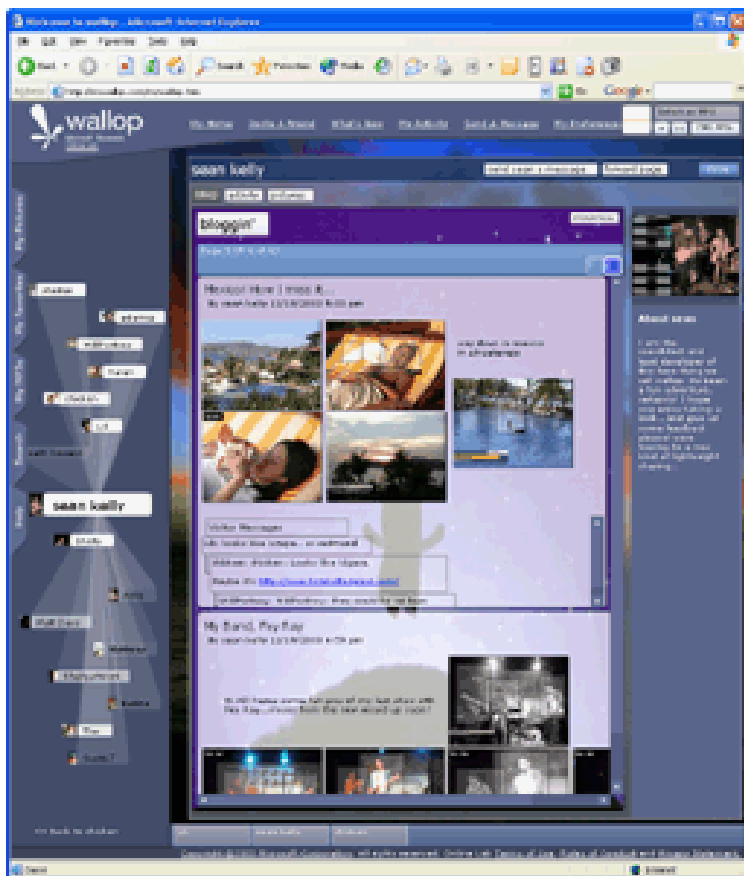


社会计算 (Social Computing)

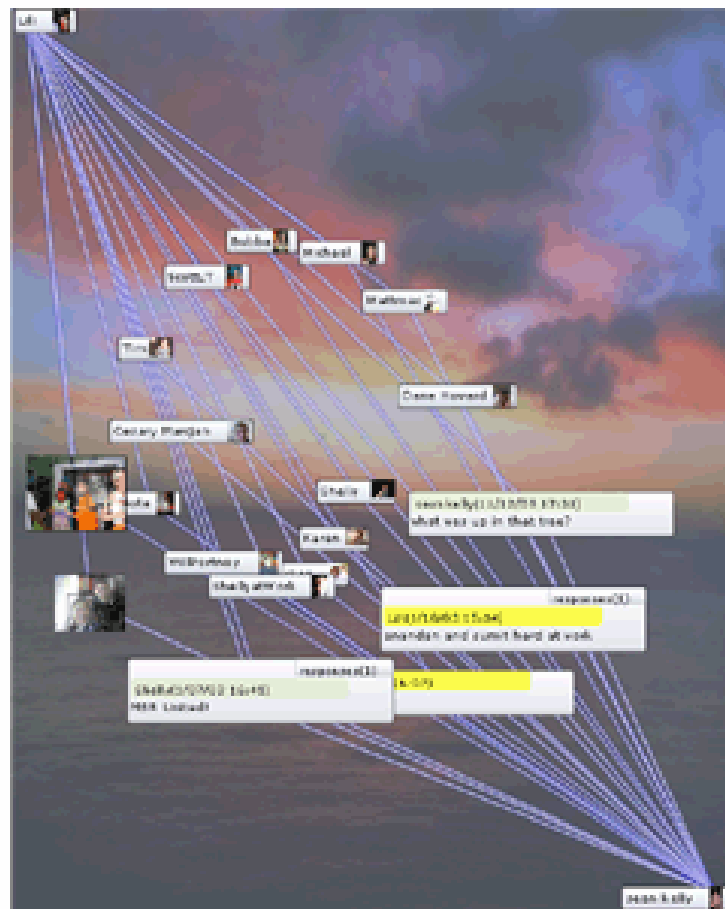
- 20世纪90年代中期，IT公司开始关注用户使用行为，纷纷成立了Social Computing Group，包括Microsoft, Intel, IBM和各大高校等。目的是希望借助软件工具实现人与人之间更好的交流与信息共享。
- 当前，IT领域所谓的社会计算 (Social Computing) 研究的主要是用于促进人机之间、组织机构之间互动的网络化的处理工具。目前有部分成熟的应用软件 (社会软件) 。如IBM开发的Babble软件，是一个类似聊天室的环境，它包含了一个被称作是社会代理 (Social Prosy) 的图形化结构，用于提供参与讨论者的相关信息。微软的一个研究项目Wallop，用于探索研究人们如何分享媒介并在社会性网络的环境下进行对话和交流。



社会计算 (续) :MS-SCG



Wallop



Sapphire



目前来看，社会计算的成果是社会软件

- 廉价而快捷的通讯技术和信息技术工具使得人类的沟通越来越简单
- 当前社会计算研究目的是使用信息技术工具，实现人与人之间更方便的社会性的交互和通讯
 - IM、论坛、Blogs、Wiki、... ..
 - del.icio.us、CiteULike



大纲

- 社会学与社会计算
- 社会信息网络的基本概念及其现实意义
- 社会信息网络模型研究
- 我们的工作
- 总结

现实世界中网络的分类

物理学家（**Newman***）将现实世界网络分为四类：
生物网络、技术网络、信息网络、社会网络

实际上，信息网络与社会网络已经融合在一起



信息网络趋向社会化 (Web2.0、P2P etc)

- 网络社区具有对应的社会群体代表性、社会关系的关联性

1. 陈水扁“废统” 关键词：陈水扁 终统 废统 民进党 台独 台湾 美国

===== 相关论坛帖子 =====

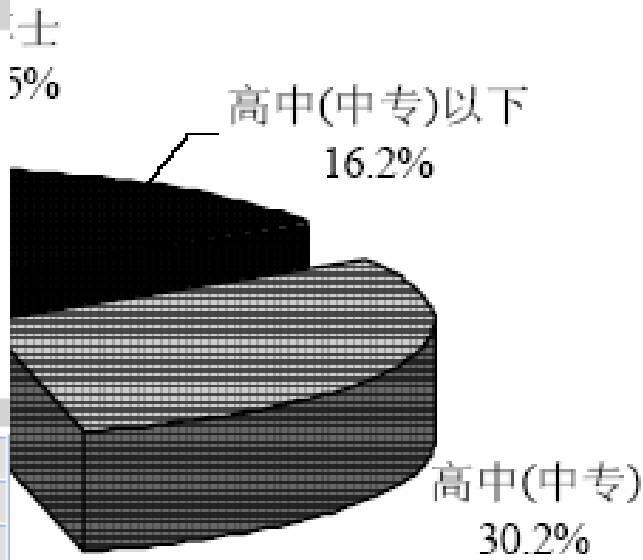
标题	来源论坛	发表时间
白忙活了一个月，阿扁公开澄清：国统..	凯迪社区	2006-3-14 23:49:00
“废统”到底是什么目的	搜户社区	2006-03-12 17:59:45
陈水扁“废统”要付出的代价	西陆论坛	2006-03-12 15:51:00
从对“废统”事件的应变看我国政府..	铁血论坛	2006-03-03 01:08:00
[杂谈]陈水扁废统了,诸位怎么看	天涯社区	2006-02-27 22:37:50

共45012条帖子，分布在1208个论坛子.. [更多帖](#)

===== 相关新闻报道 =====

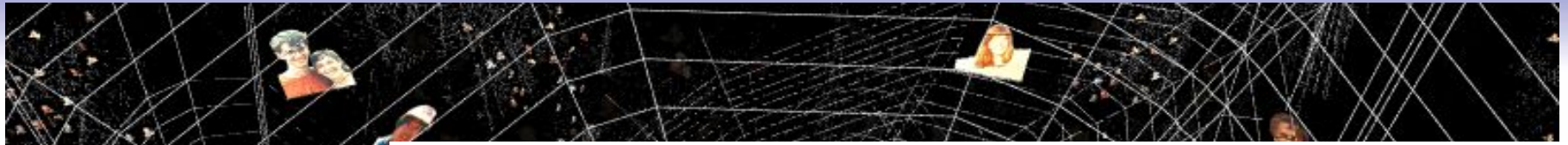
标题	来源站点	发表时间
世界广西同乡联谊会主席谴责陈水扁..	人民网	2006-3-16 08:18
宁被“罢免”都要“废统”	海峡导报	2006-3-19 16:02:05
马英九：陈水扁应说清楚国统纲领..	中国新闻网	2006-03-19 18:07
澳大利亚华人华侨社团谴责陈水扁..	国际在线	2006-03-04 13:22:36
日本华侨华人强烈谴责陈水扁“废统”..	北方网	2006-03-05 01:52

共3895条报道，分布在187个新闻网站 [更多报道...](#)

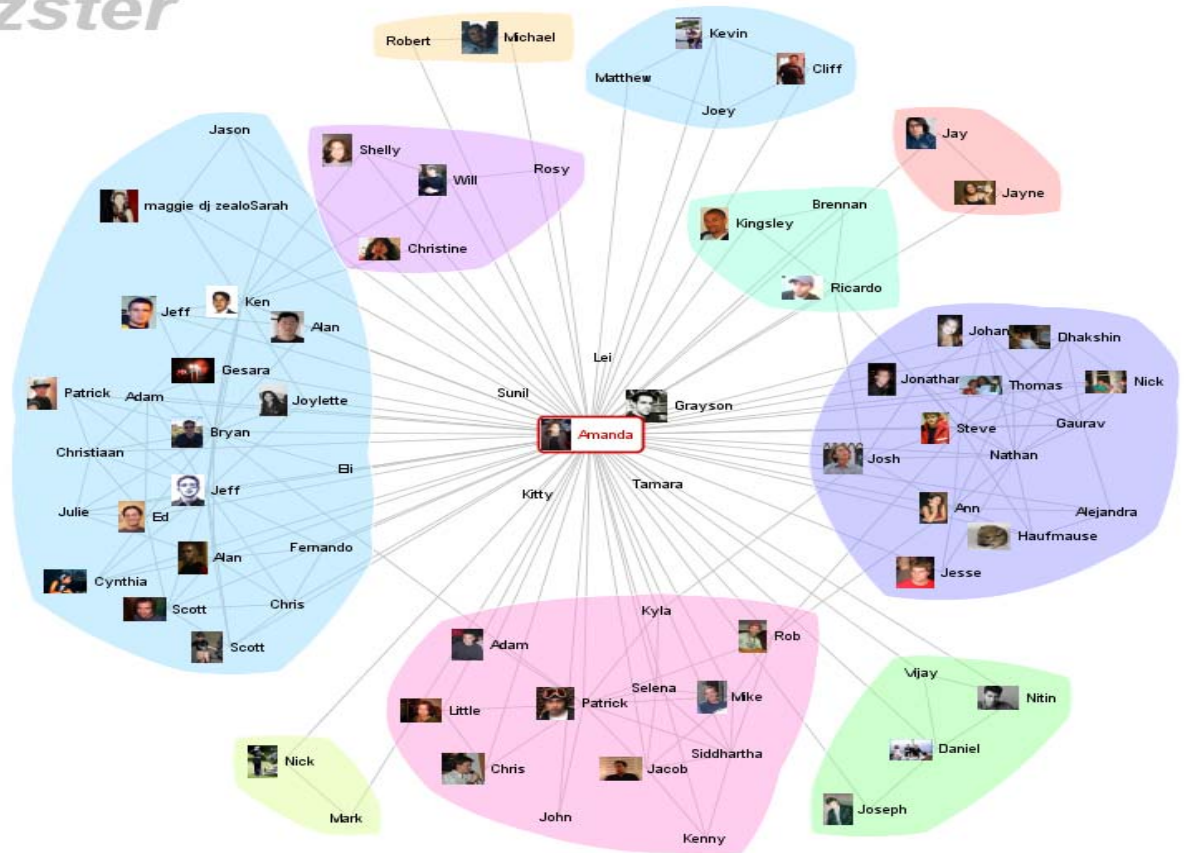


数据来源：中国互联网络信息中心 (CNNIC)

社会信息整体上的网络化



vizster



Acid rain

Water shortage!

How would you fix it?

'Storytime'

Saline water

Watersheds

What is hydrology?

Are raindrops
tear-shaped?



● 社会信息的网络化表示

- 社会组织与人与人之间的关系呈现出动态演化的复杂网络状态
- 人类活动以信息传播为表象，呈现动态群体的集聚性和关联性，包括个人、小群体、组织、行业等等之间的复杂网络关系

● 信息网络的社会化特性

- 在Web2.0时代，信息网络的动态性、交互性、参与性本身与社会化特征没有差别
- 无所不在的网络信息呈现特征空间的多维性(Rich- Dimensional FS)、涌现现象的大量产生（社区、话题）、话题与事件的快速扩散、内容之间动态关联等现象

不太严谨的定义：从可描述和可计算的角度来看，不同层次、不同颗粒度的信息社会和信息网络，可以统一表示为

“社会信息网络”

社会信息网络的表象特征

- 同时具有个性化与社会化特征
- 内容特征与结构特征的融合
- 存在大量的结构涌现、内容涌现现象
- 具有一定的传播规律
- 具有动态演化性质
-



举例：互联网对社会的“蝴蝶效应”

- 一个社会事件或者一条网络新闻通过互联网所产生的“蝴蝶效应”
 - “超级女声” 所产生信息涌现导致现实社会中“草根文化效应”
 - “虐猫事件”
 - 芙蓉姐姐
 - “馒头”、“无厘头”所产生的社会文化现象
- 一件网络事件所产生的社会效应
 - “孙志刚事件”对国家政策法规方面的影响
 - Q币、QQ挂机策略所导致的对社会其它行业的影响
 -

社会信息网络研究的现实需求

□ 规模与体系结构的挑战

- 社会信息的网络化、信息网络的社会化
- **Web2.0: Web**、邮件、博客、即时通信、短信、论坛、个人空间、**P2P**社区、... ..

□ 服务质量的要求

- 更高效、更个性的信息搜索
- 内容的深度挖掘

□ 效率与网络安全的需要

- 消息与病毒的传播与扩散

□ 社会安全的需要

- 网络行为的涌现现象判定、预测
- 多样化社区关系挖掘
- 宏观网络行为的规整与制导



共性问题

- ❑ 社会信息网络的基本要素与基本规律分析
- ❑ 社会信息网络的形式化表达与建模
- ❑ 影响网络宏观行为的微观规则是什么？
- ❑ 网络涌现现象的度量与分析

大纲

- 社会学与社会计算
- 社会信息网络的基本概念及其现实意义
- 社会信息网络模型研究
- 我们的工作
- 总结



早期社会学者的调研

- 早在半个多世纪以前社会学的研究就开始意识到了网络的意义：
 - 一个人的社会行为的倾向性，更多的由他的亲近的朋友和社会关系决定，而并非由于整个社会中的倾向性比例
- 社会网络分析的具体问题
 - 犯罪关系、选举活动的民意调查、谣言传播、社会营销、性关系

量化的社会网络建模

国外一些大学、研究机构进行的定量社会学
(**Quantitative Social Science**) 研究, 如:

- 哈佛大学的定量社会分析学院 (**The Institute for Quantitative Social Science at Harvard University, IQSS**)
- 美国华盛顿大学的社会统计学中心 (**Center for Statistics and the Social Sciences, CSSS**)
- 哥伦比亚大学的社会与经济政策研究院的计量社会科学研究课程 (**Quantitative Methods in the Social Sciences, QMSS**)

研究的问题是现实存在的社会问题, 但更多的还是采用传统的数学和统计学方法, 构建数学模型或是数学方程式 (主要是微分方程), 然后与现实数据进行拟合。可以在此模型的基础上进行短期社会行为进行预测。但是, 微分方程模型能够解释的现象与规律非常有限



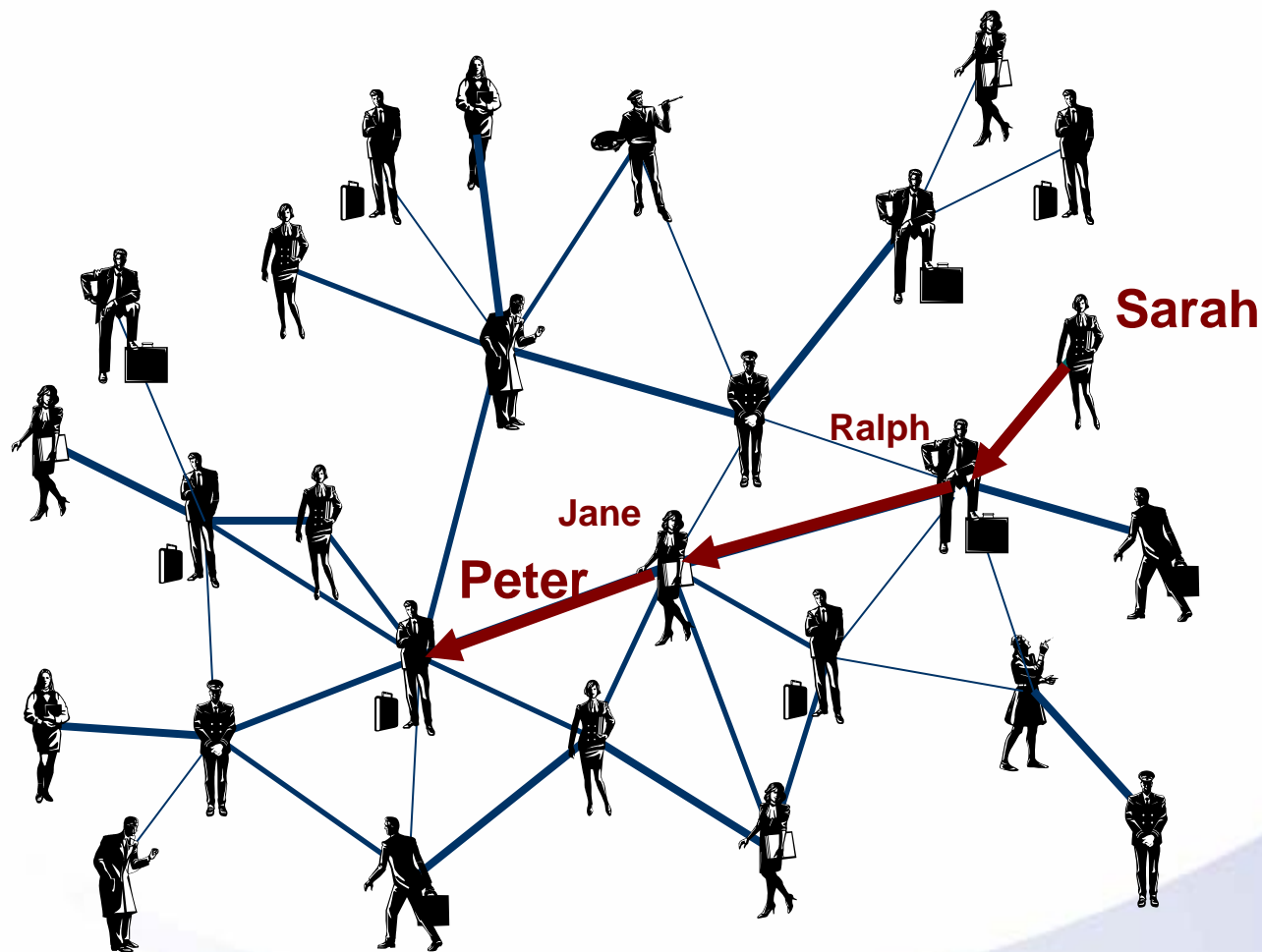
基于图论的现实网络研究

● 基本属性:

- 节点、连边与连接度
- 网络直径
- 聚集系数
- 度分布
- 层次结构
- 连通性等



对小世界现象的观察

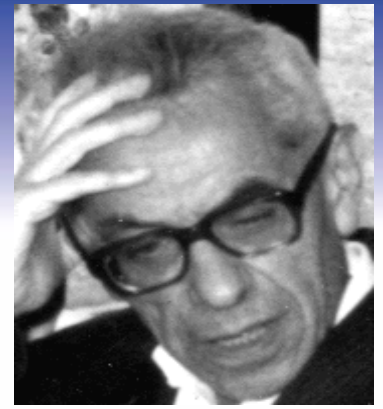


Society:
Six degrees
S. Milgram 1967
F. Karinthy 1929

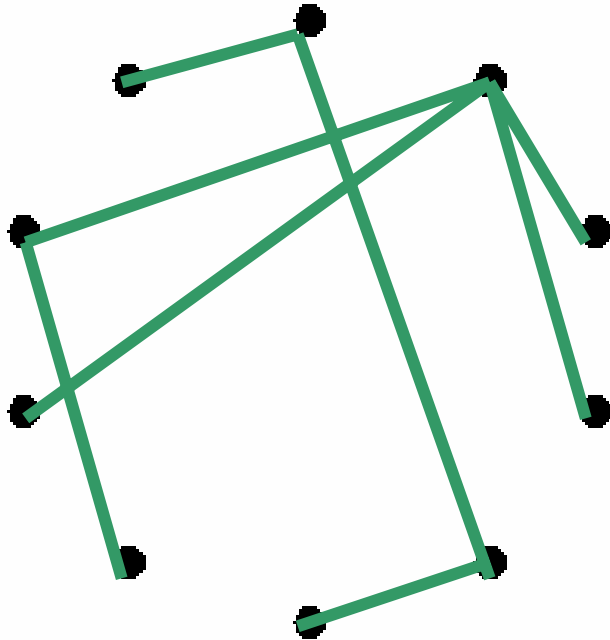
WWW:
19 degrees
Albert *et al.* 1999

随机网络模型

Erdős-Rényi model (1960)



Pál Erdős
(1913-1996)



Connect with probability p

$$p=1/6$$

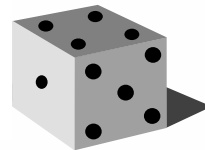
$$N=10$$

$$\langle k \rangle$$

Poisson distribution

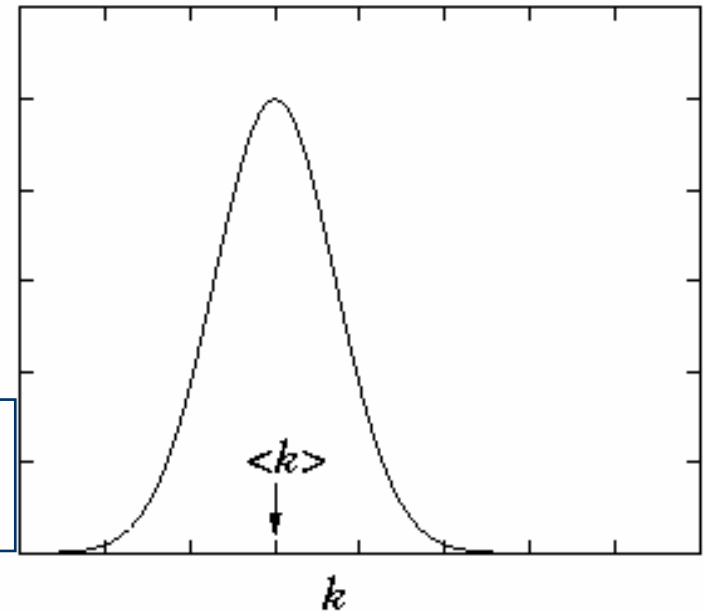
- Democratic

- Random



$P(k)$

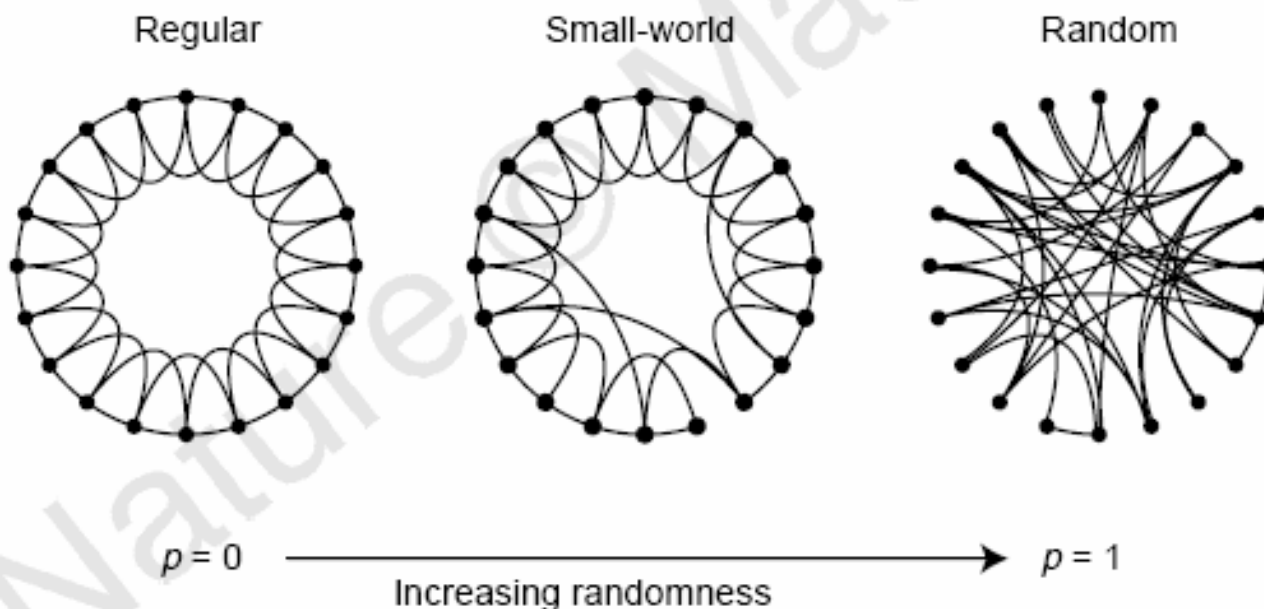
$$P(k) \sim e^{-\langle k \rangle} \frac{\langle k \rangle^k}{k!}$$



小世界网络模型

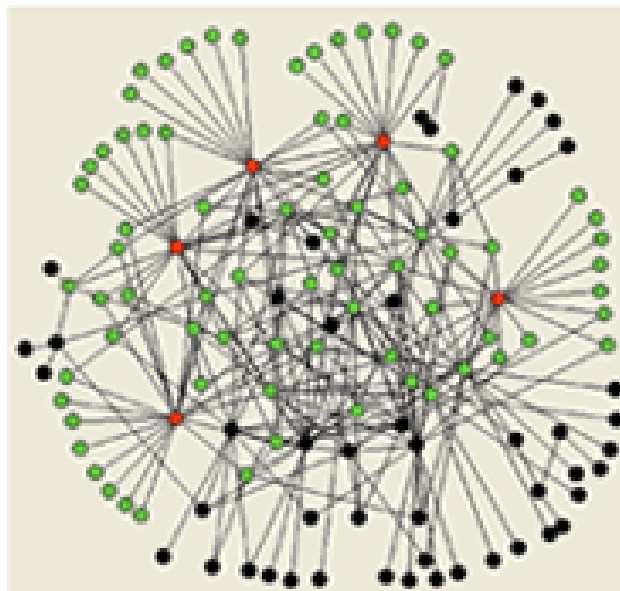
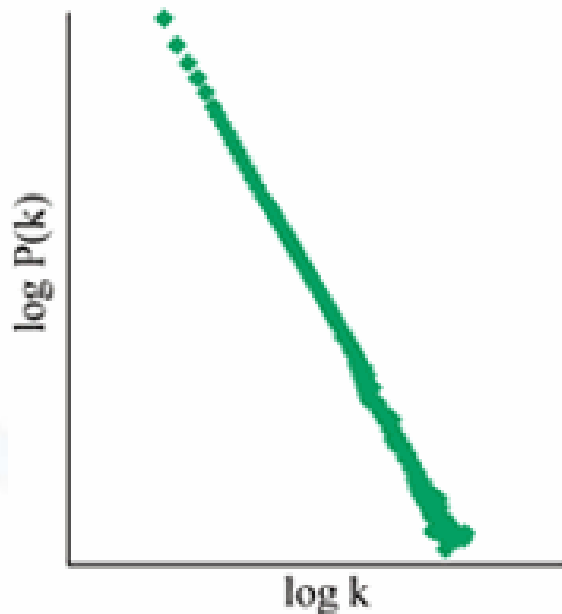
● Watts & Strogatz 1998

- 通常来说，现实网络的连接拓扑结构应该是既非完全规则也非完全随机。
- D.J. Watts



自由标度网络模型

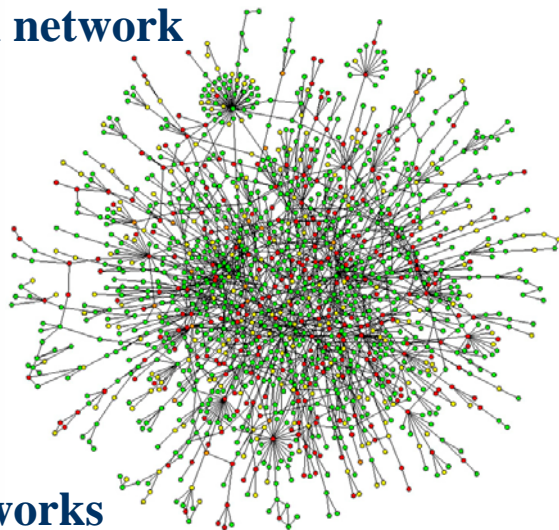
- A-L. Barabasi & R. Albert 1999
 - 度分布的幂率特性
 - 连边倾向的生长原则



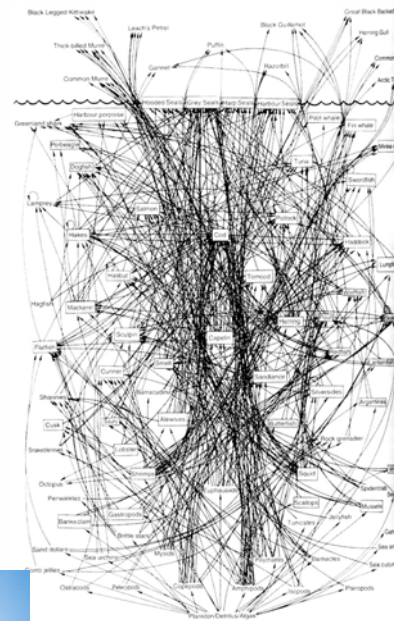
Power-law everywhere

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

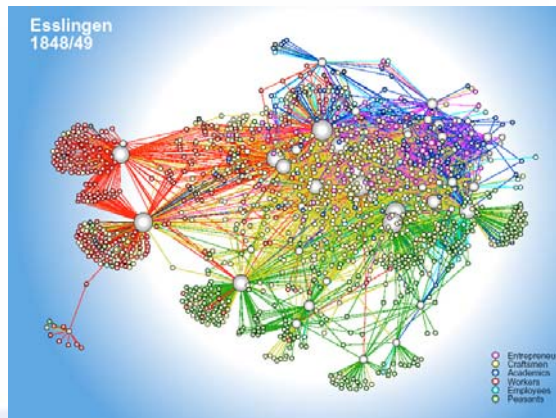
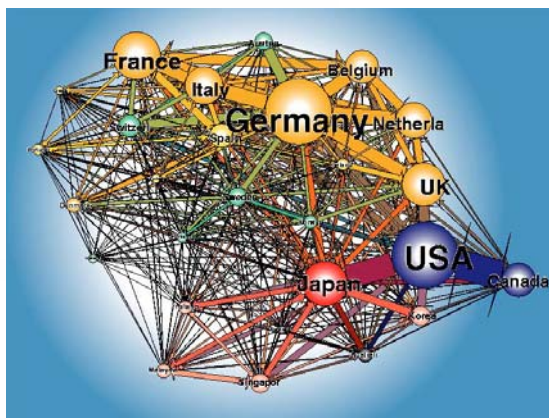
Yeast protein network



Food web



Economic Networks

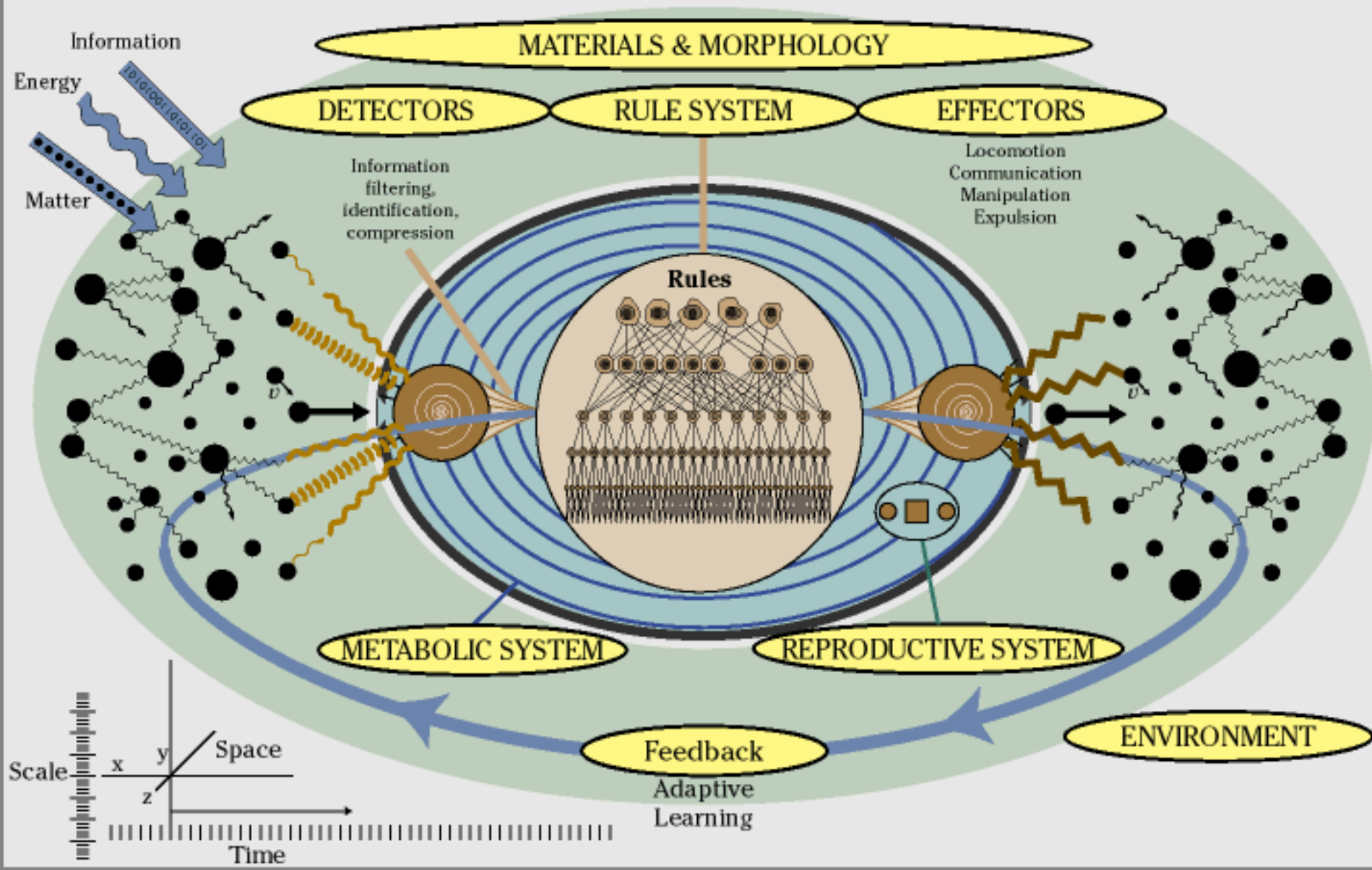


Social networks

Complex Adaptive System

- Hidden Order (John Holland 1994, SFI)
 - 适应性造就复杂性
- CAS与计算机模拟
 - 刺激-反应模型、适应度的确认与修改、新规则的产生机制
 - SWARM (Santa Fe Institute)

Complex Adaptive System Model



其它相关模型

- 信息网络的复杂系统建模及相关计算
 - 基于链接结构的信息网络拓扑分析 (Kleinberg, Brin & Page, 等)
 -
- 复杂动力学模型 (Dynamics)



大纲

- 社会学与社会计算
- 社会信息网络的基本概念及其现实意义
- 社会信息网络模型研究
- 我们的工作
- 总结

研究动机与研究内容

- 社会信息网络是一个复杂的巨大规模系统
- 互联网挖掘与搜索的传统思维是集中式、深度内容计算，其策略类似拿着放大镜到大海里寻针
- 物理学家的跨越式思维：
 - 从牛顿力学到热力学
 - 从微观要素的度量到宏观要素的度量
- 当信息规模上升到一定程度之后，社会安全、内容搜索以及交互式新型信息共享与信息服务应该有更加有效的表示与计算模式

我们目前研究的落脚点（方向）：

1. 复杂信息网络的基本特征与模型
2. 面向社会计算的体系结构与系统
3. 网络关系挖掘与多维度特征融合计算

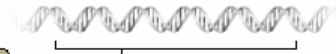
更直接动机：网络计算与内容计算的融合

Humans have only about three times as many genes as the fly,

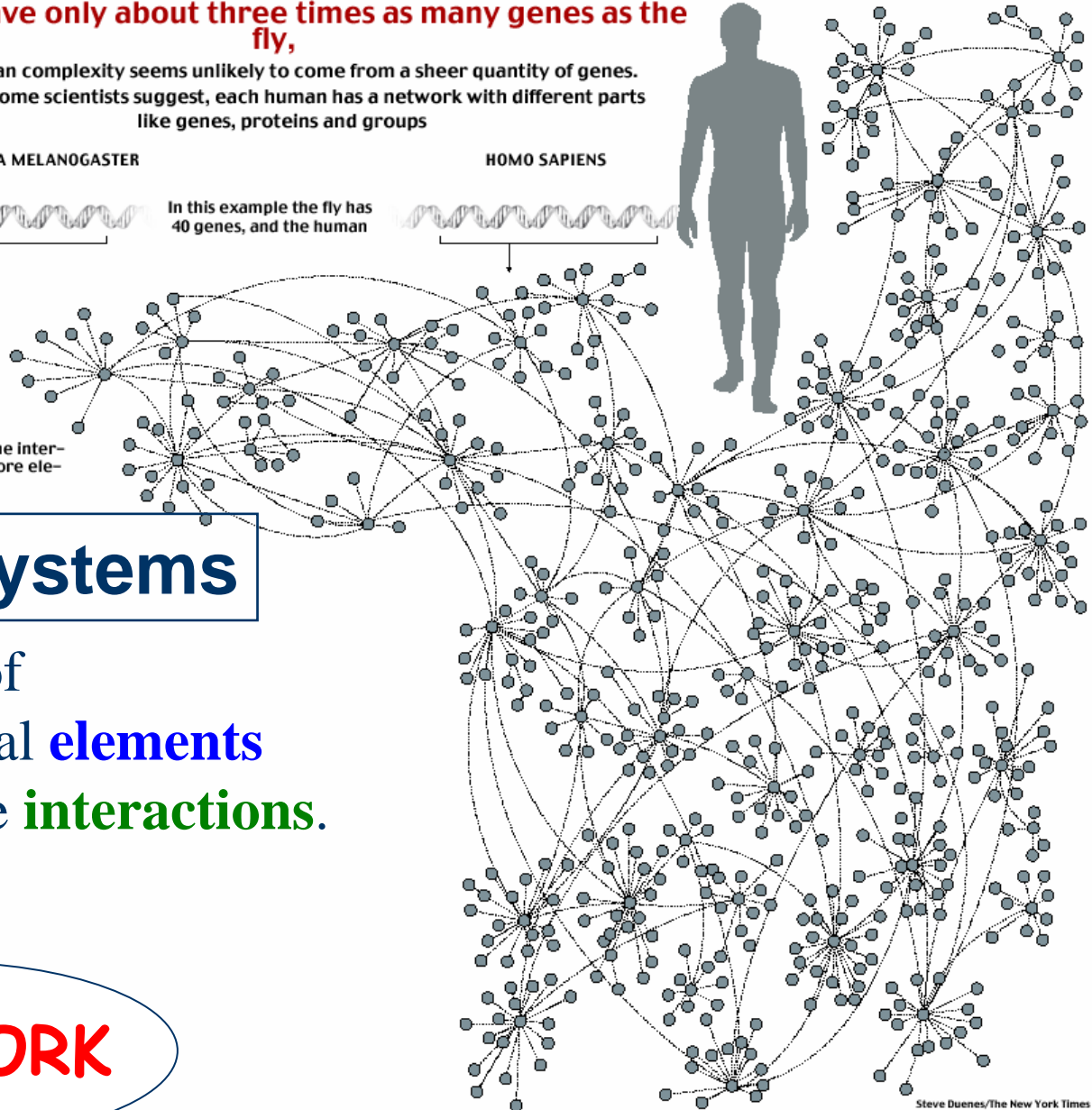
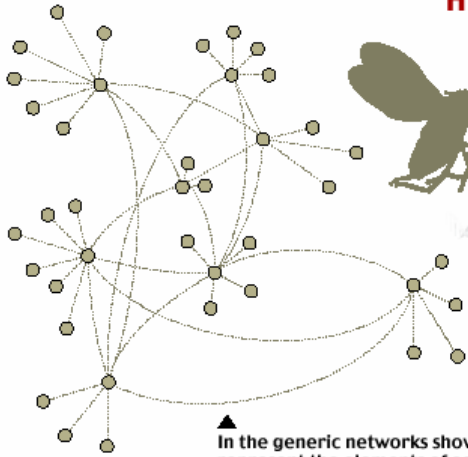
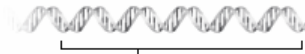
so human complexity seems unlikely to come from a sheer quantity of genes. Rather, some scientists suggest, each human has a network with different parts like genes, proteins and groups

DROSOPHILA MELANOGASTER
(Fruit fly)

HOMO SAPIENS



In this example the fly has 40 genes, and the human



▲ In the generic networks shown, the points represent the elements of each organism's genetic network, and the dotted lines show the interactions between them. Humans have many more ele-

Sources: Dr. Albert-László Barabási, University of Notre Dame; Science; Celera Genomics

Complex systems

Made of many non-identical **elements** connected by diverse **interactions**.



NETWORK

现阶段工作进展

- 社会信息网络的特征发现与建模
- 社区发现与网络关系挖掘
- 自组织的轻量级网络操作系统LIOS
- 相关应用：互联网搜索与挖掘、信息安全、社会信息分析等



因特网的基本特征分析：具体问题的提出

- 人们通过观察发现如下的现象
 - 从基于超链的拓扑结构来看，**Web**网络具有一般复杂网络所具有的共同特性：自由标度（**scale-free**）、聚团性(**Clustering**)和结构的层次性(**hierarchy**)。但是层次结构与功能（内容）的关系还非常不清楚（**Barabasi 2005**）
 - 宏观上，**Web**网络中，通过超链接形成的物理上紧密连接网页“往往”与某个主题相关。由此产生了类似**PageRank**、**HITS**等方法。问题是，研究发现类似的模型其参数没有规律性，大部分情况下效果很差。（**SIGIR 02**）
- 问题：因特网的内容与结构到底是什么样的关系？有没有可量化特征？



统计发现：拓扑结构与内容具有一定相关性

- 结构属性与内容属性是信息网络中两类截然不同而又相互依存的基本要素
- 微观粒度下的结构聚团性与内容聚团性之间存在指数比例关系



信息网络建模：问题的提出

● 相关问题

- 信息网络拓扑结构的自由标度等特性以及结构与内容之间稳定的映射关系特征是如何产生的？
- 微观要素是如何影响宏观行为的？
- 为什么信息网络宏观粒度的连边密集社区“往往”是内容相关的？

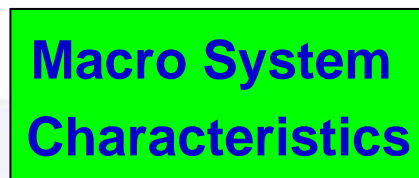
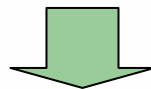
● 研究思路

- 给出“内在规律”的假设 → 建立模型 → 理论与实验验证



More about the correlation between the content and the structure of an information network

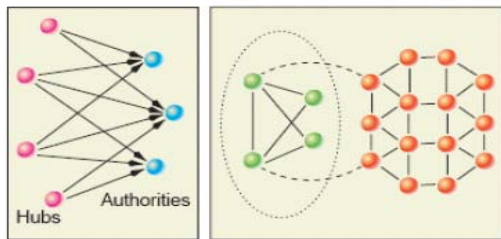
- **Why topological structure reflects content distributions?**
 - From structure to content? (Link analysis?)
 - From content to structure? (Semantic Web?)



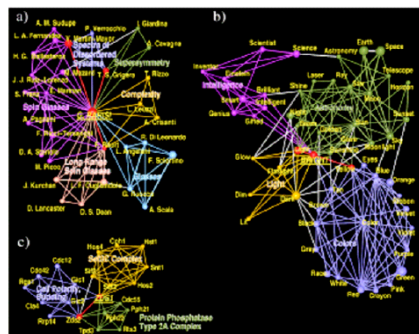
Thus : Pageranking etc should be modified to be more reasonable and effective

社区分析的现状：连边密度社区

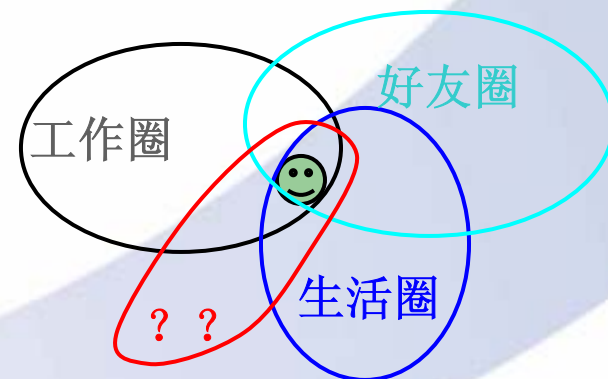
- 社区定义：网络中物理连边稠密的节点子集为一个社区
- 潜在的前提：连边最密集必然意味着功能最相关
- 优点：计算方法相对简单
- 主要缺点：
 - 忽略了真实社区之间存在的多样化的判定尺度和交叉重叠的关系
 - 只能处理静态网络关系（网络快照：snapshot）



Link density community
Kleinberg etc, Science 294 (2001)



K-clique community
Palla etc, natural (2005)



Message transferring density Community

Delta-Closure Community

- 多维度（尺度）、动态消息传播密集社区的定义

- Transferring Probability :

$$A_{n*n}$$

- Random walk:

$$P = P(D_1) \times A^1 + P(D_2) \times A^2 + \dots + P(D_k) \times A^k$$

- Delta-closure:

$$\forall n_i \in S_\delta, \sum_{j=1}^K P_{ij} \geq \delta (i \neq j)$$

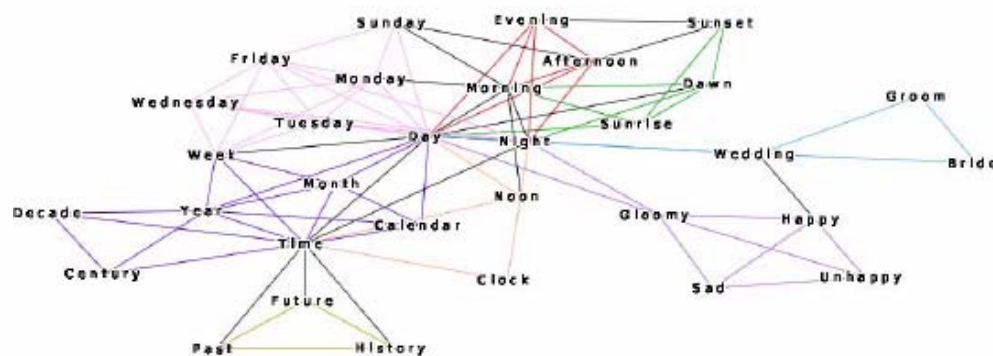
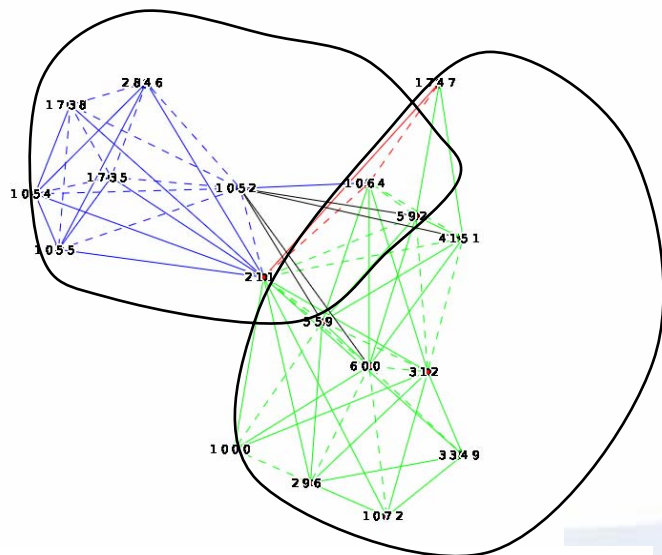
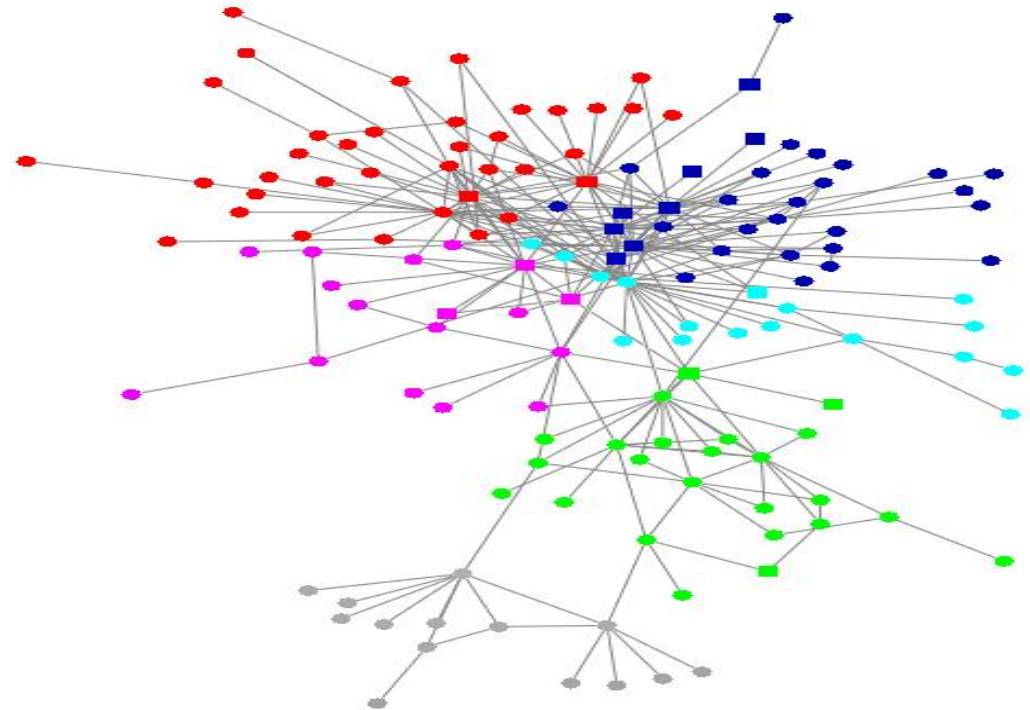


图5.11 Citeseer中的两个论文社区

word association中与词汇day相关的社区

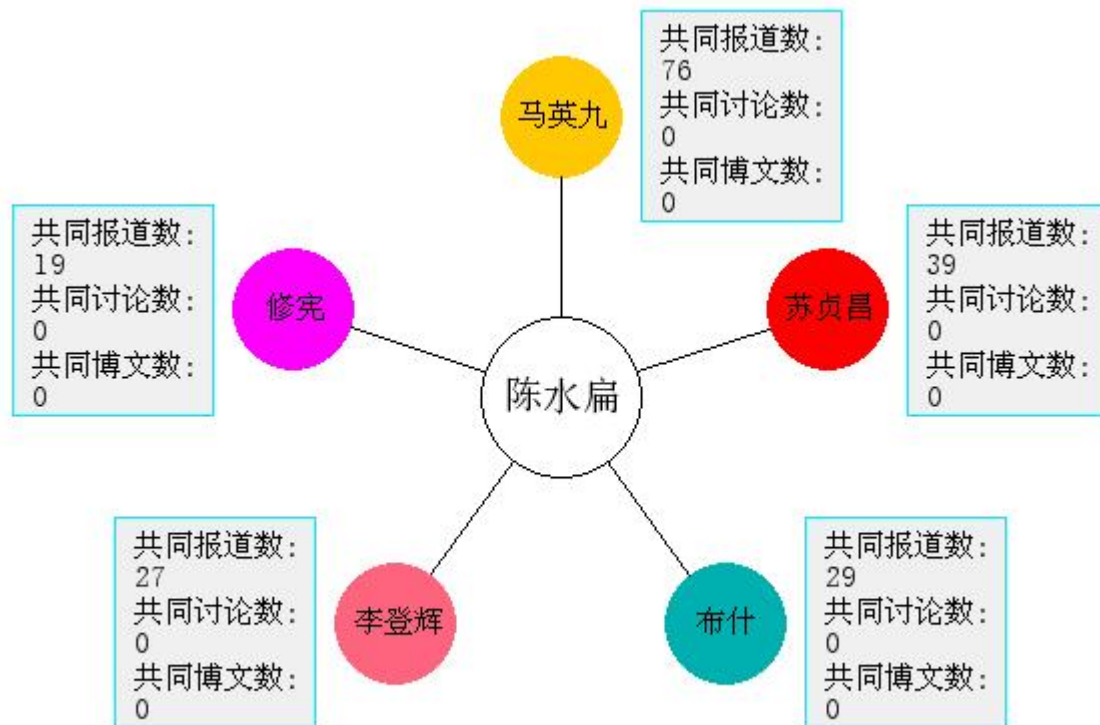
Expert Finding in the Internet Forums

- 从邮件列表、BBS这样的Internet论坛中寻找专家用户：
 - 以咨询为主的网络中，专家通常更多的为非专家用户提供咨询答疑，而较少与其他专家讨论
 - Disassortative mixing
 - 专家往往分散在多个社区中，且在社区中成为中心人物。

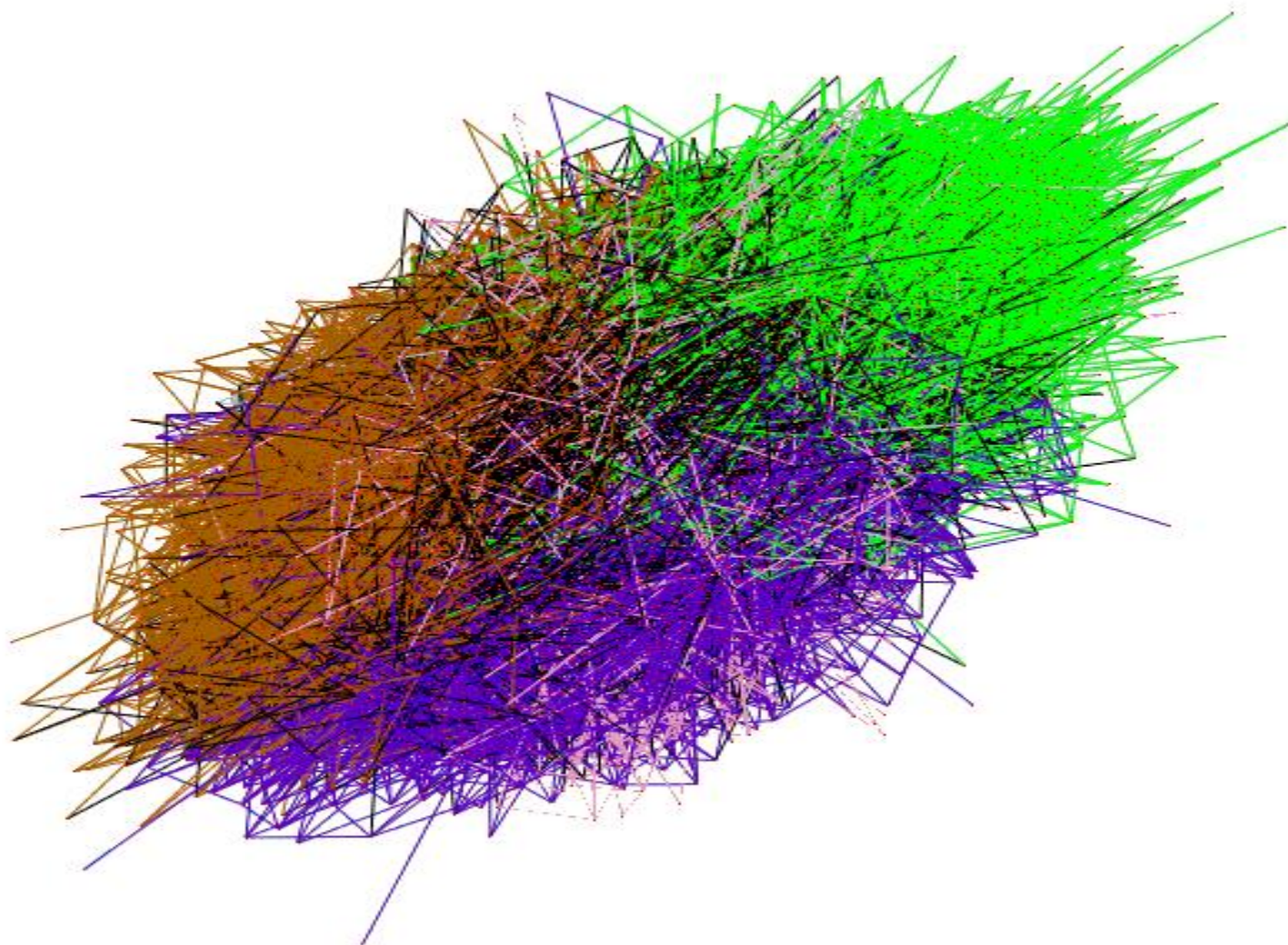


网络热点人物关系网络

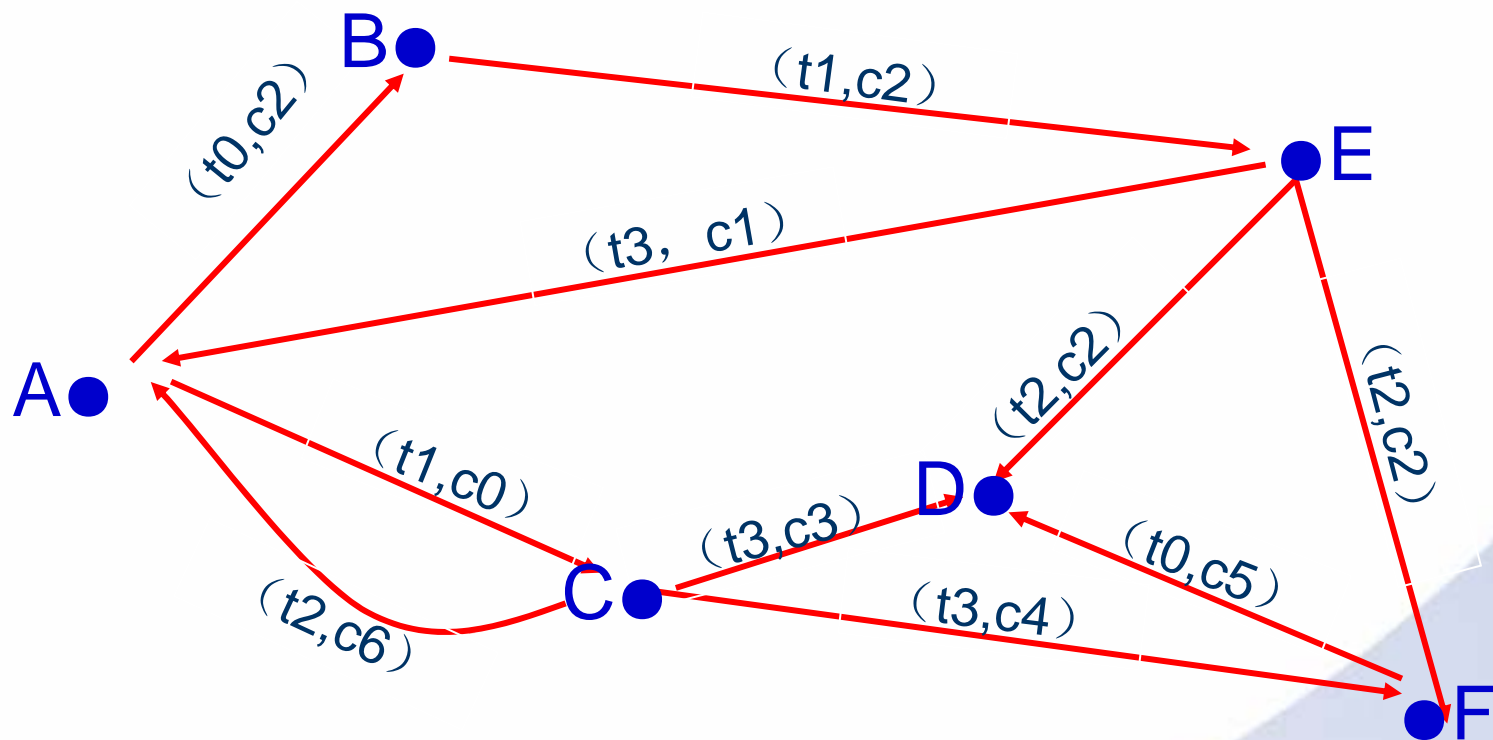
● 根据新闻报道共现关系挖掘



词连通图聚类呈现结果



垃圾短信扩散网络



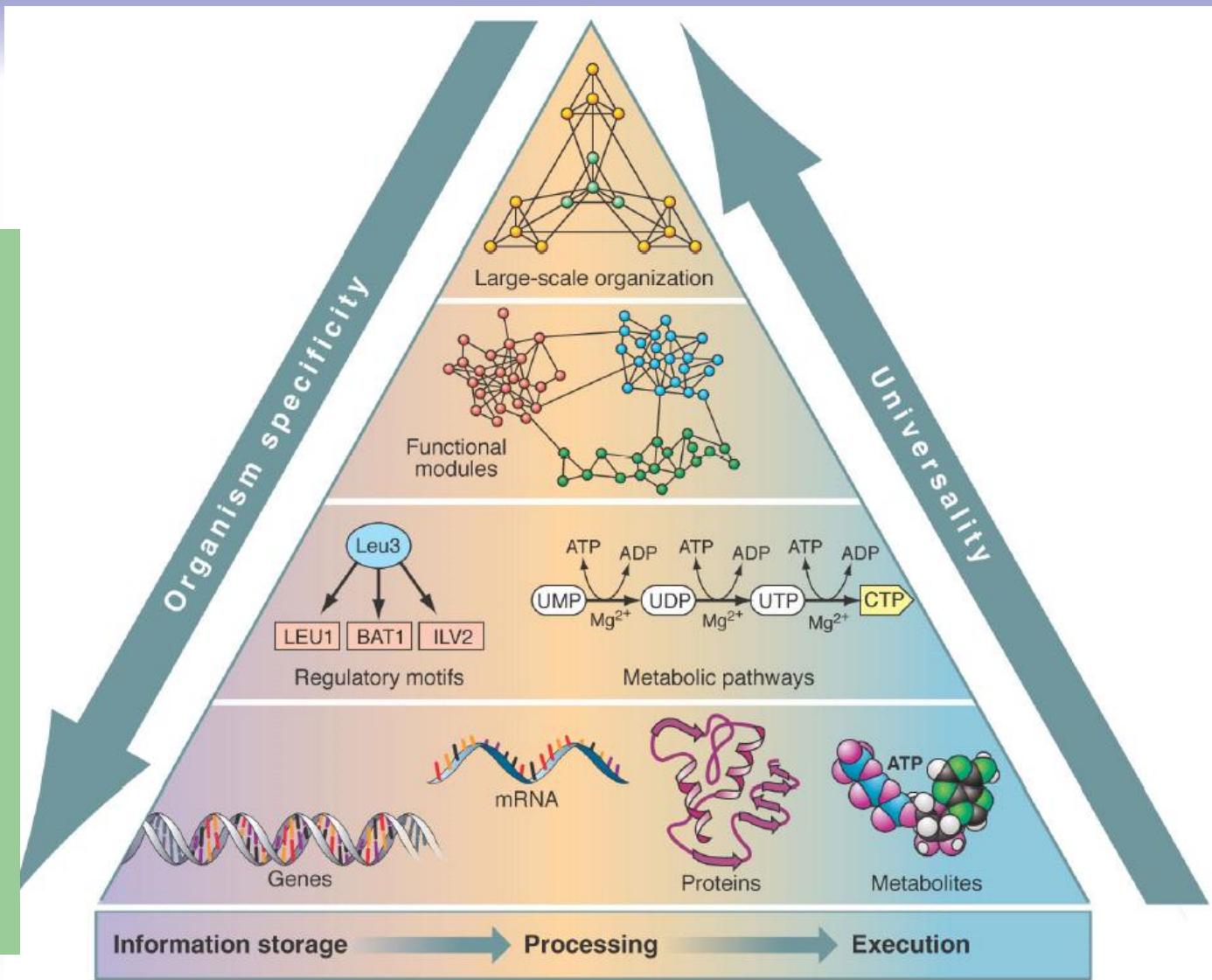
垃圾短信扩散网络示意图

总结

- 信息科学与社会科学的融合
 - 关注重大社会问题
 - 关注信息科学与社会科学交叉研究
 - 从人类社会和自然界的构造规律中学习，提升信息系统的计算能力
- 未来计算可能社会计算，即网络计算与智能计算的融合
- 社会信息网络对信息技术有新的启示

Life's Complexity Pyramid

复杂的系统有令人惊讶的
简单而又完美的特性



Z.N. Oltvai and A.-L. B. *Science*, 2002



中国科学院计算技术研究所
INSTITUTE OF COMPUTING TECHNOLOGY, CHINESE ACADEMY OF SCIENCES

**Now we are close to knowing just about everything
there is to know about the pieces.**

**But we are as far as we have ever been from
understanding nature as a whole**

Barabasi, 2005

谢谢!



中国科学院计算技术研究所
INSTITUTE OF COMPUTING TECHNOLOGY, CHINESE ACADEMY OF SCIENCES