

大数据时代

网络科学助力社会科学研究





大数据时代

社会科学与网络挖掘 ——安全与隐私




主讲教师：沈浩 博士

中国传媒大学电视与新闻学院
中国传媒大学调查统计研究所
中国传媒大学数据挖掘研发中心

教授
副所长
主任





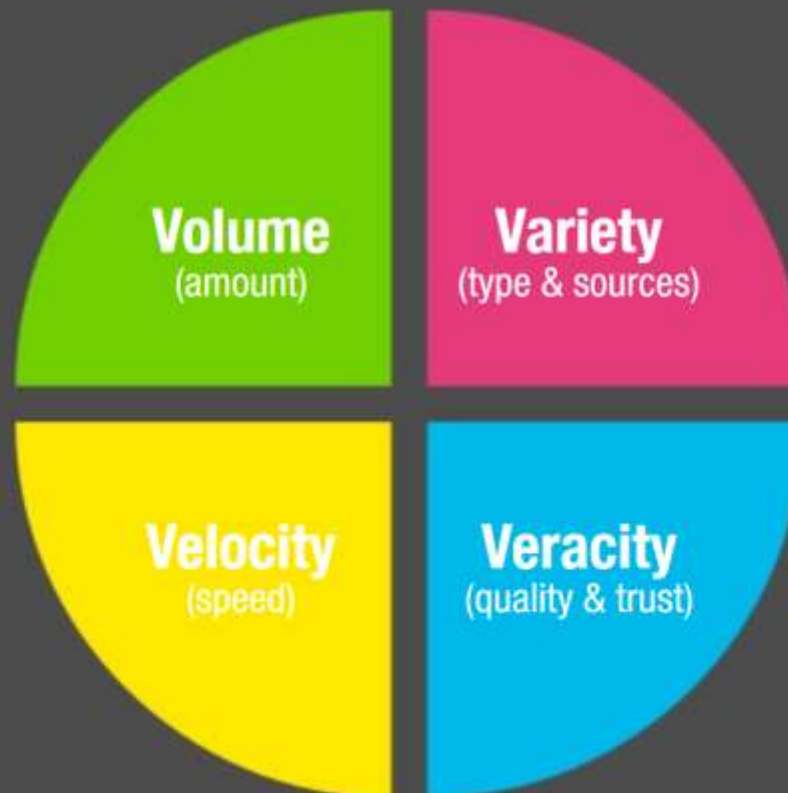
这是一个令人兴奋的时代，也是一个大数据的时代，网络科学让我们越来越多地从数据中观察到人类社会的复杂行为模式。以数据为基础的技术决定着人类的未来，但并非数据本身改变了我们的世界，起决定作用的是我们对可用知识的增加。

《暴发》正是让我们思考如何从大数据中塑造未来美好世界的正能量。



*According to Wikipedia: Big Data is defined as “data sets whose size is beyond the ability of commonly used software tools to **capture**, **manage**, and **process** the data within a **tolerable elapsed time**”*

The four Vs **OF BIG DATA**



多带来不同...

Big Data Is Only Getting Bigger

全媒时代信息就是选择



公开数据洞察...

重发现，轻抽样...



理论上再大的局部可能不如随机抽样有代表性

非结构化数据...

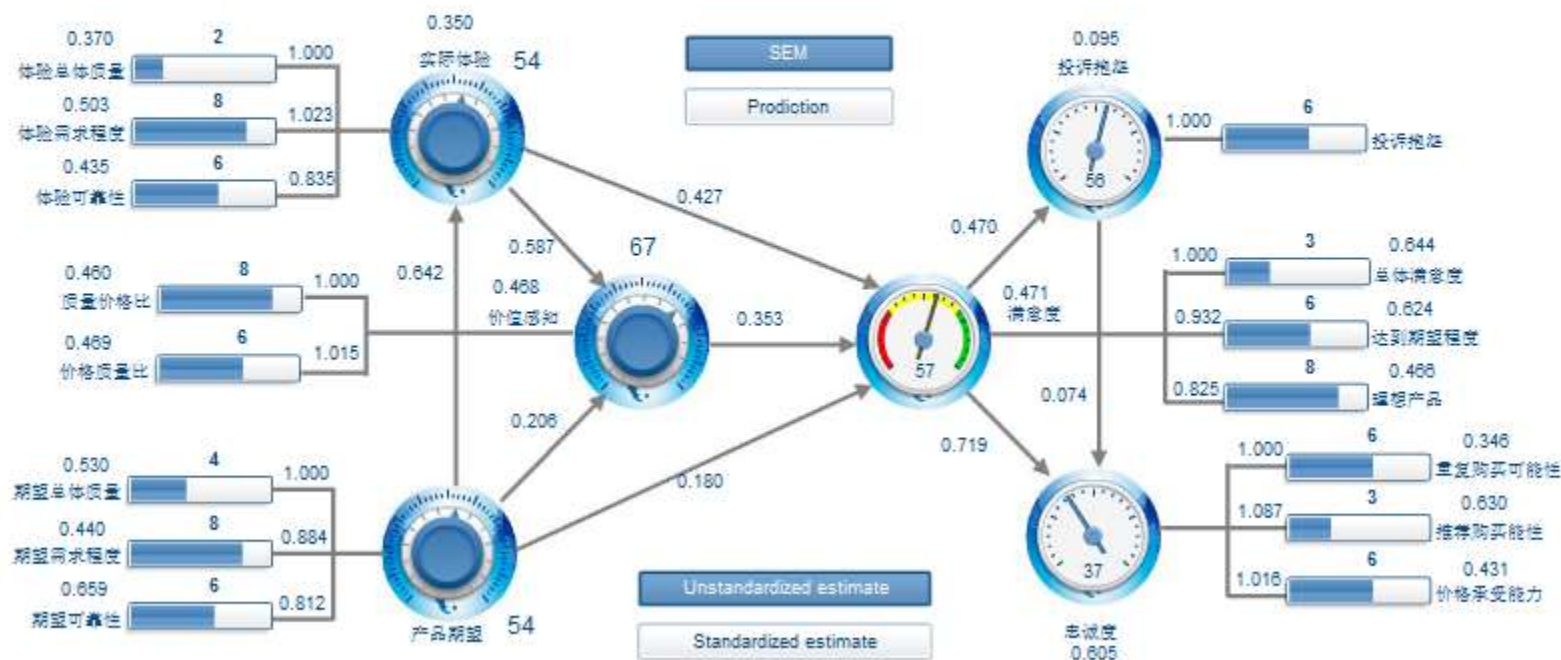


重关系不因果... 问什么不问为什么...

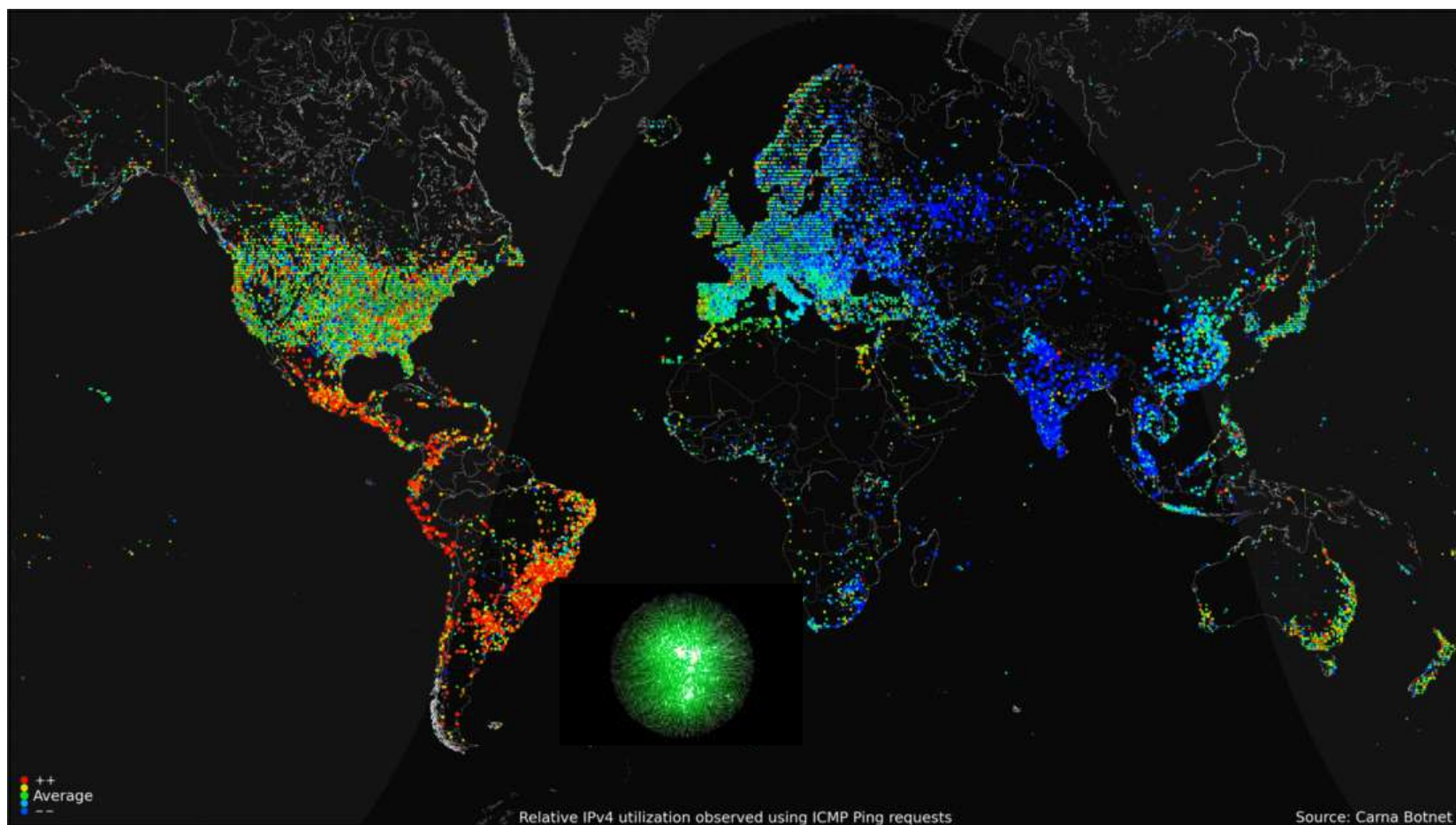


顾客满意度研究模型应用案例-AMOS

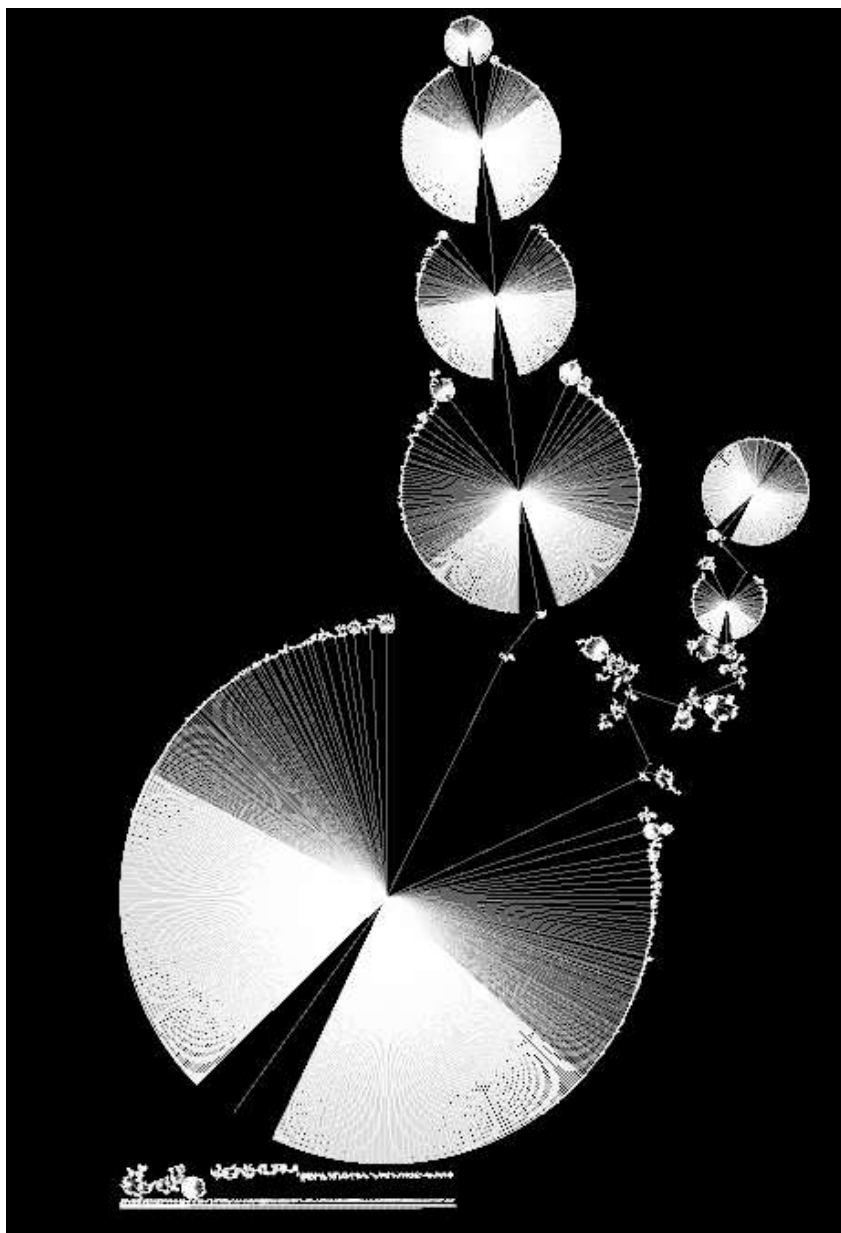
BI INSIGHT
博易智讯



小就是大，重大...



上帝的指纹



可算计个人...



越来越个性化，意味着越来越社会化

可预知社会...

数据挖掘

文本挖掘

LBS和二维码

数据可视化

统计分析

情感分析



个性化推荐

商业智能

自然语言处理NLP

意见挖掘

语义分析

内存计算



社会科学是研究人的

不同的**学科**不是研究不同的
问题，而是从不同的**角度**研
究同一个问题！



三大社会科学理论

1. 突变理论 (catastrophe theory)

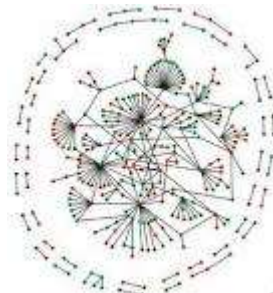
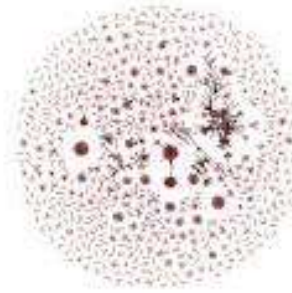
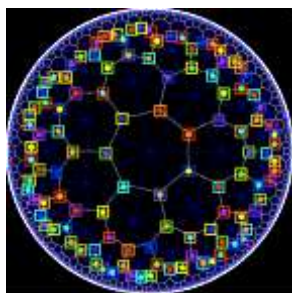
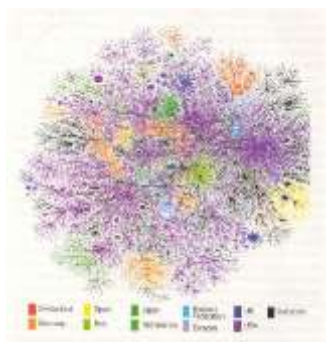
为人们理解微小作用导致社会突然变化的机理开拓了道路。

2. 混沌理论 (Chaos theory)

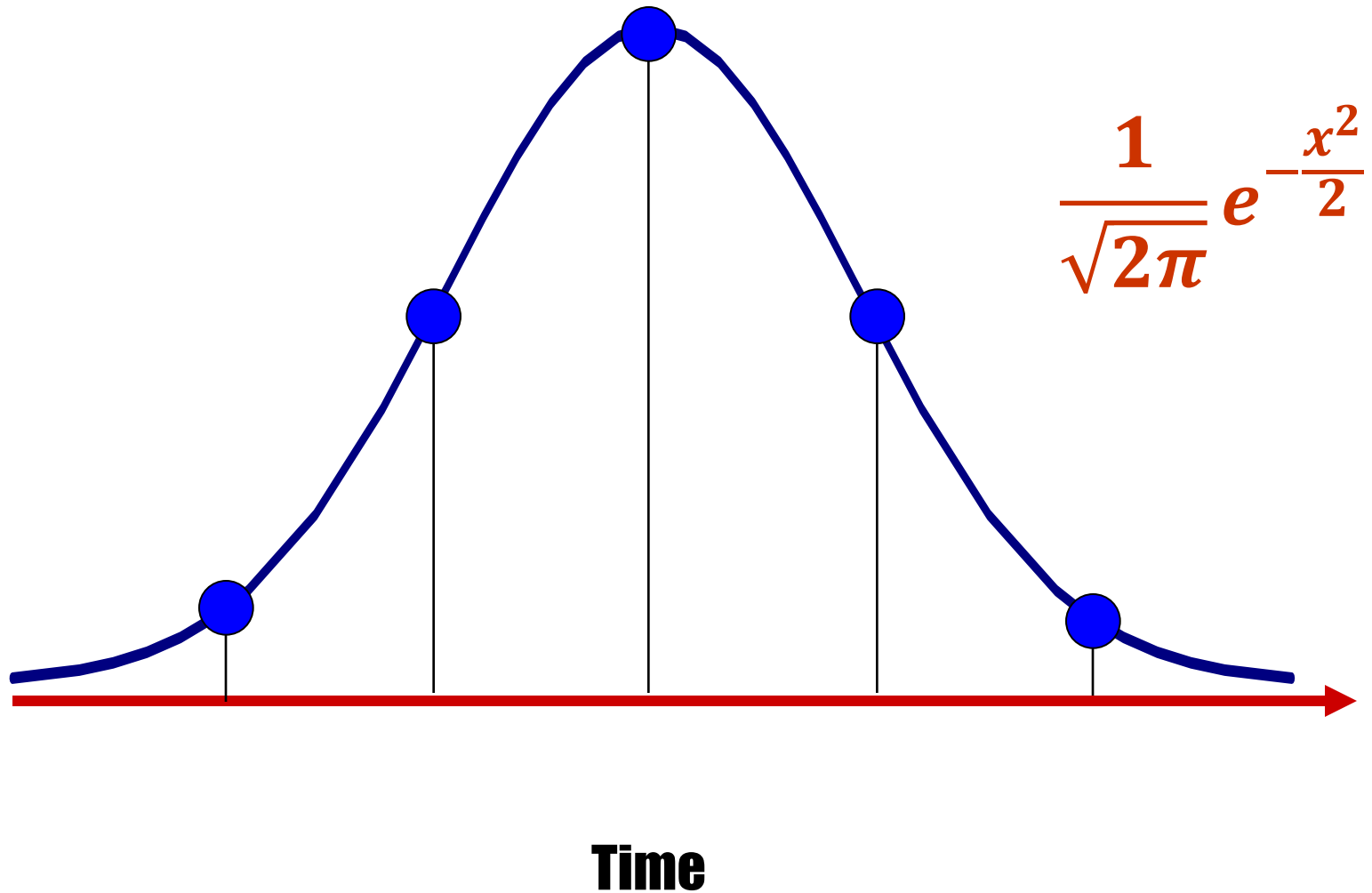
复杂而不断变化的系统，即使其初始状态是详尽了解的，也会迅速进入无法精确预知的状态。

3. 复杂性理论 (Complexity theory)

在大量元体 (agent) 各自按照不多的几条简单规则相互作用时，如何从中产生出秩序与稳定。

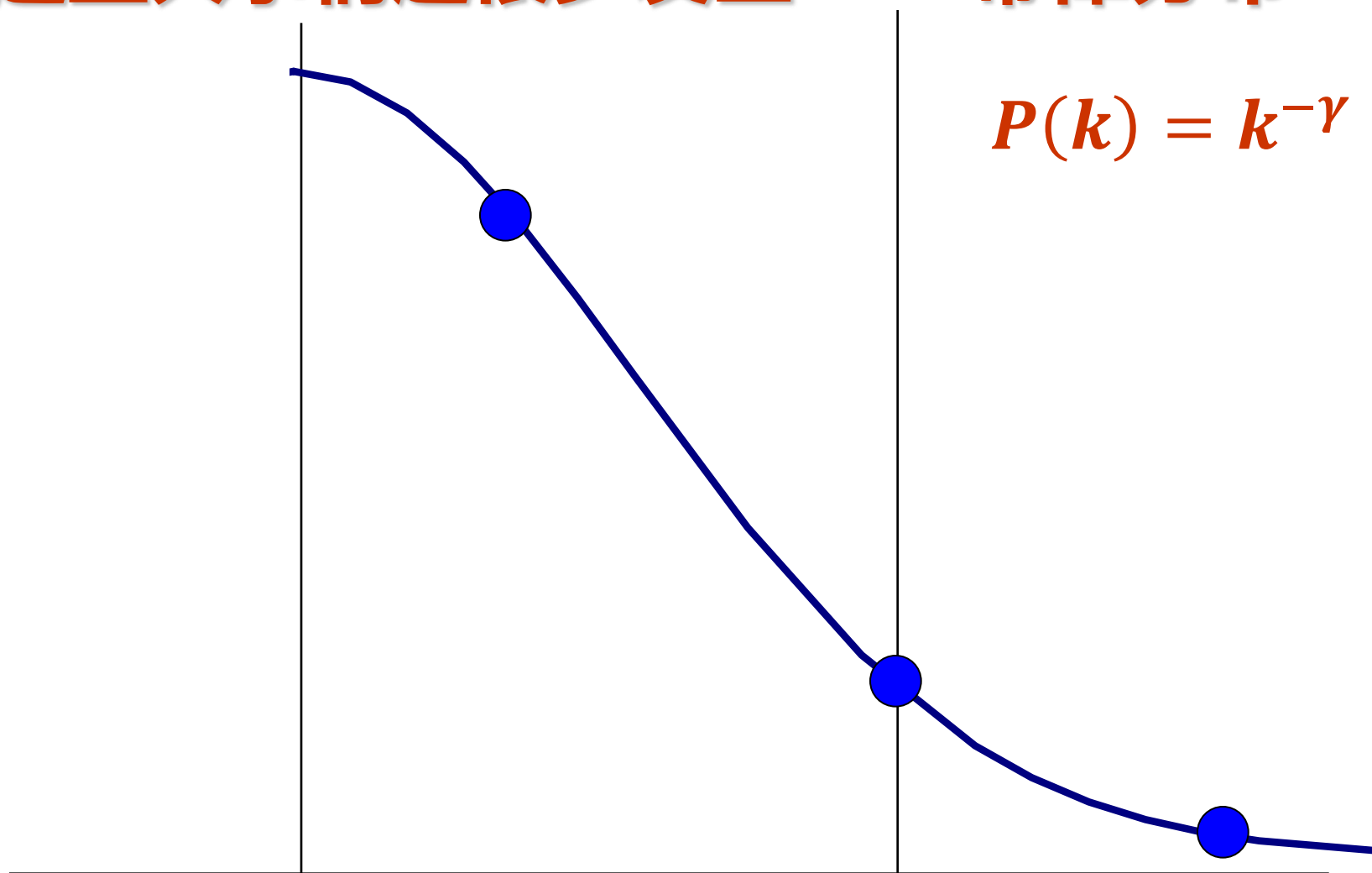


构想未来的工具



越重大事情越很少发生——幂律分布

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





越分享越懂得欣赏

The more TV gets personal
The more TV will become social



广播APP应用秀



技术是文化的表现，
是文化得以创造和表达的方法，
技术是手段也是目的。



社会化媒体是技术也是平台！

社会化媒体是技术，它拓展了我们与社会其他人联系的能力。





用户为王、数据为王、关系为王



我们生活在一个关系的社会

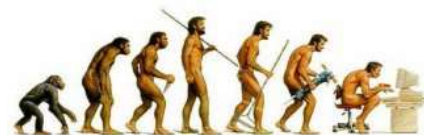
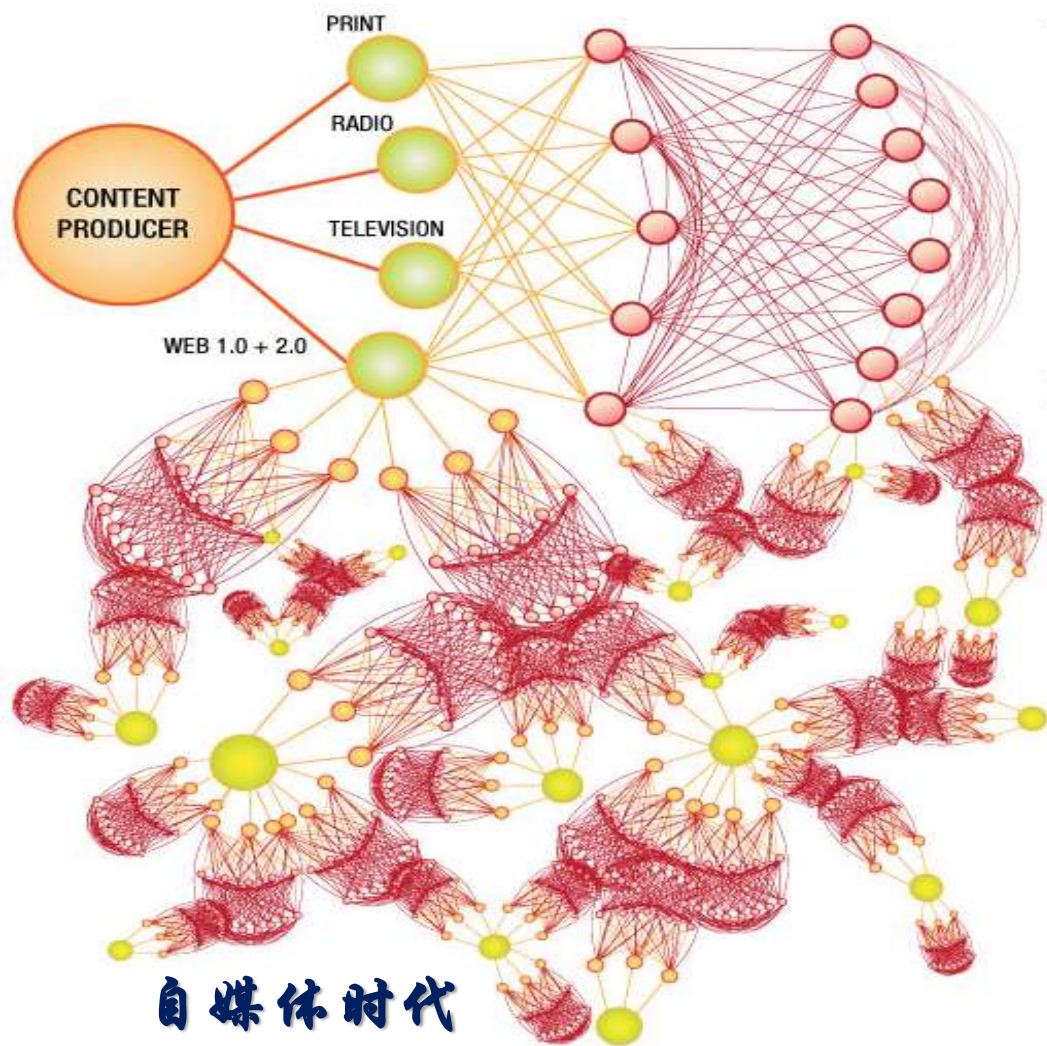


微博——重塑社会关系总和

不在于你知道什么，在于你认识谁！



社交媒体——重塑人的社会关系总和



关系——社会网络——社会结构



派系、凝聚子群体、成分、孤立点、结构洞、社会资本

从定性到定量，从属性到关系

今天关注：

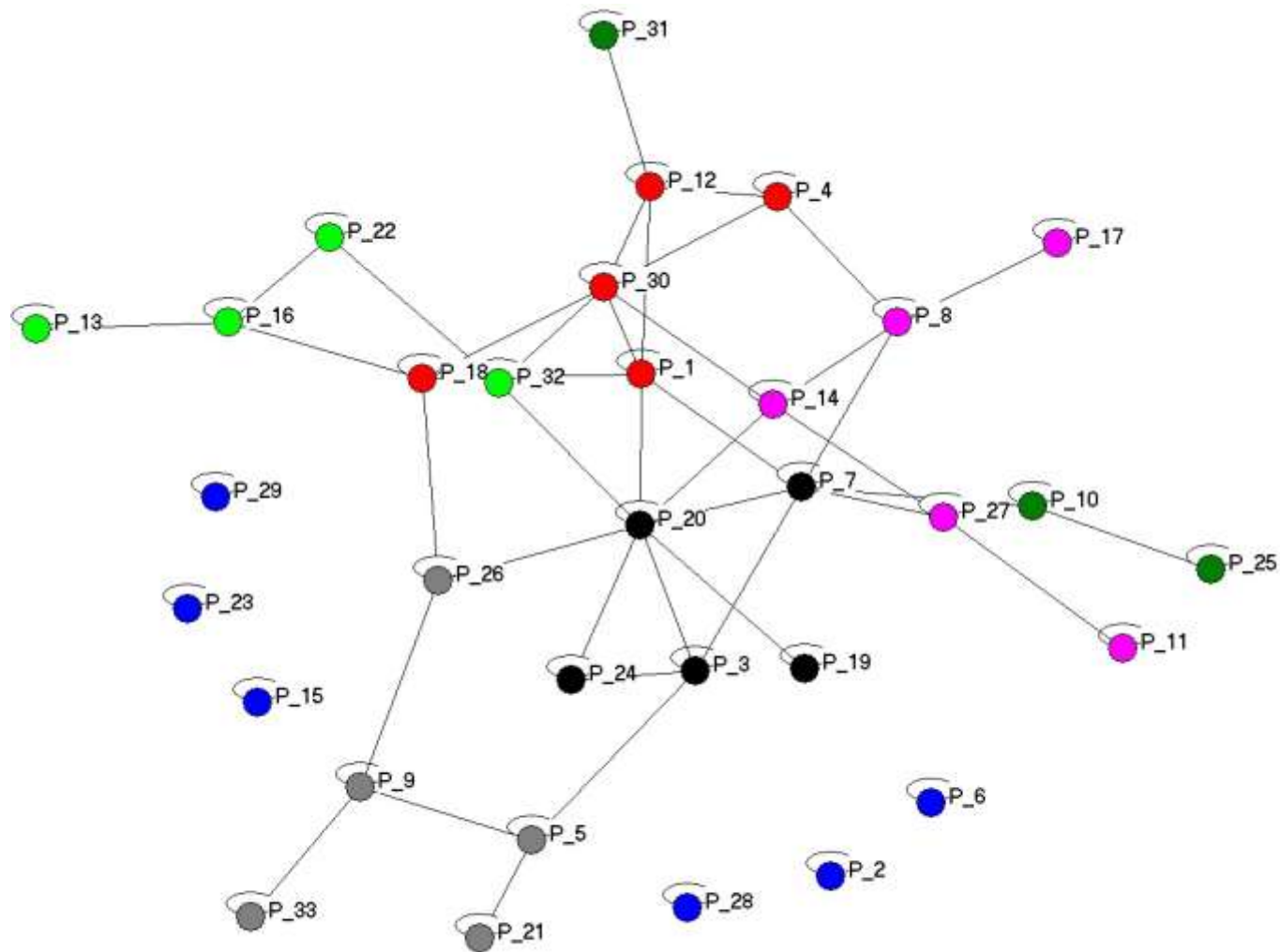


骇客帝国——矩阵-Matrix

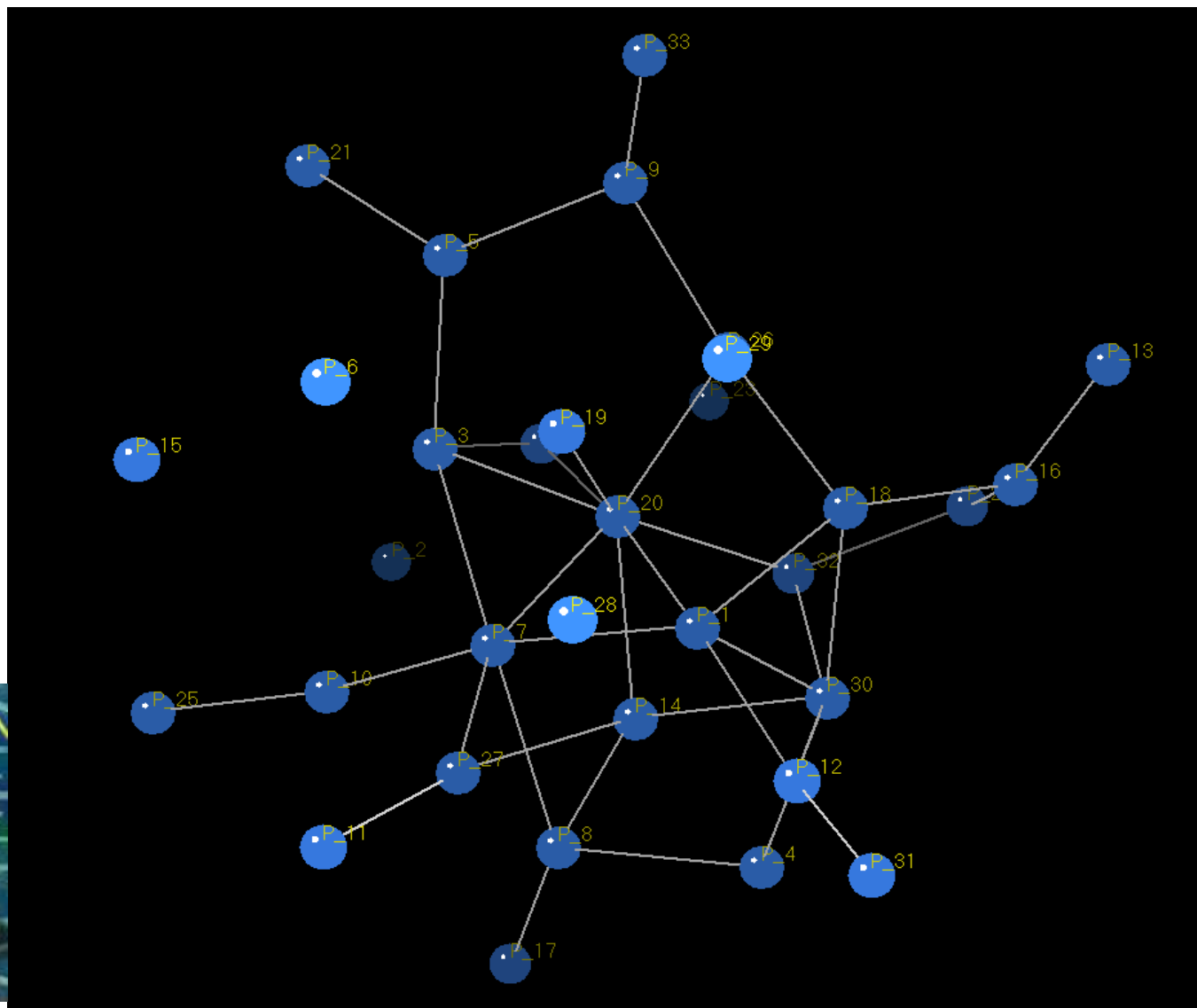
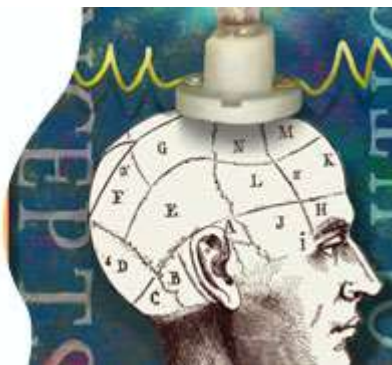
信息就是矩阵

	P_1	P_2	P_3	P_4	P_5	P_6	P_7	P_8	P_9	P_10	P_11	P_12	P_13	P_14	P_15	P_16	P_17	P_18	P_19	P_20	P_21	P_22	P_23	P_24	P_25	P_26	P_27	P_28	P_29	P_30	P_31	P_32	P_33	
P_1	109	11	16	16	15	13	23	20	13	13	17	22	21	21	10	14	18	25	15	27	19	20	13	12	18	20	19	15	12	22	16	20	9	
P_2	11	93	18	16	15	15	17	13	14	16	14	14	11	15	13	15	15	12	11	19	14	15	14	17	14	12	19	10	12	21	16	17	10	
P_3	16	18	114	18	28	13	23	20	18	20	18	9	13	17	16	16	20	13	13	26	21	15	13	24	19	19	17	20	12	20	17	21	17	
P_4	16	16	18	107	15	21	14	23	16	18	17	22	14	15	13	18	21	14	16	12	16	15	17	12	16	16	19	14	16	18	23	20	18	12
P_5	15	15	28	15	106	17	10	16	22	16	12	14	12	19	16	17	14	18	10	21	23	14	17	17	16	15	20	16	11	17	21	19	17	
P_6	13	15	13	21	17	90	16	20	8	17	10	13	14	18	10	13	16	11	15	20	15	14	7	13	15	14	12	9	9	19	17	13	13	
P_7	23	17	23	14	10	16	108	23	6	28	2											8	15	19	14	15	24	18	14	21	12	20	12	
P_8	20	13	20	23	16	20	23	106	16	16	1											10	14	17	15	15	17	12	14	21	16	18	14	
P_9	13	14	18	16	22	8	6	16	97	10	1											19	12	17	20	23	16	14	12	8	13	17	22	
P_10	13	16	20	18	16	17	28	16	10	103	1											18	12	16	22	21	18	15	16	17	17	17	17	
P_11	17	14	18	17	12	10	20	19	15	11	9											13	12	15	14	20	25	15	13	15	17	16	9	
P_12	22	14	9	22	14	13	12	15	15	12	1											18	16	11	16	16	18	16	16	24	24	19	10	
P_13	21	11	13	14	12	14	14	20	15	13	1											9	18	15	17	14	14	14	12	20	18	20	14	
P_14	21	15	17	15	19	18	21	22	19	12	1											14	13	11	17	17	30	14	13	27	17	17	18	
P_15	10	13	16	13	16	10	11	9	12	11	1											13	15	7	14	15	13	10	18	15	20	15	17	
P_16	14	15	16	18	17	13	16	18	14	16	1											23	17	14	12	15	13	17	17	16	15	16	15	
P_17	18	15	20	21	14	16	20	23	10	13	2											18	9	12	12	16	15	19	20	20	16	19	14	
P_18	25	12	13	14	18	11	16	10	17	16	1											17	15	12	16	25	18	14	15	24	14	19	14	
P_19	15	11	13	16	10	15	17	15	12	12	1											16	12	13	18	16	14	10	17	15	16	16	13	
P_20	27	19	26	12	21	20	24	16	19	20	2											20	12	23	19	31	16	20	19	19	17	23	15	
P_21	19	14	21	16	23	15	17	13	21	19												20	19	17	17	19	14	17	17	19	14	19	15	
P_22	20	15	15	15	14	14	8	10	19	18	1											00	16	18	10	17	15	11	11	17	18	22	16	
P_23	13	14	13	17	17	7	15	14	12	12	1											16	91	13	19	17	8	12	13	18	14	21	12	
P_24	12	17	24	12	17	13	19	17	17	16	1											18	13	91	19	14	13	7	9	12	10	16	10	
P_25	18	14	19	16	16	15	14	15	20	22	1											10	19	19	100	15	14	17	16	7	15	17	10	
P_26	20	12	19	19	15	14	15	15	23	21	2											17	17	14	15	111	17	19	17	16	18	15	13	
P_27	19	19	17	14	20	12	24	17	16	18	2											15	8	13	14	17	108	20	17	18	20	19	13	
P_28	15	10	20	16	16	9	18	12	14	15	15	16	14	14	10	17	19	14	10	20	17	11	12	7	17	19	20	95	12	19	17	18	12	
P_29	12	12	12	18	11	9	14	14	12	16	13	16	12	13	18	17	20	15	17	19	17	11	13	9	16	17	17	12	89	13	11	14	5	
P_30	22	21	20	23	17	19	21	21	8	17	15	24	20	27	15	16	20	24	15	19	19	17	17	18	12	7	16	18	19	13	116	16	22	19
P_31	16	16	17	20	21	17	12	16	13	17	17	24	18	17	20	15	16	14	16	17	14	18	14	10	15	18	20	17	11	16	104	11	17	
P_32	20	17	21	18	19	13	20	18	17	17	16	19	20	17	15	16	19	19	16	23	19	22	21	16	17	15	19	18	14	22	11	115	21	
P_33	9	10	17	12	17	13	12	14	22	17	9	10	14	18	17	15	14	14	13	15	15	16	12	10	10	13	13	12	5	19	17	21	89	

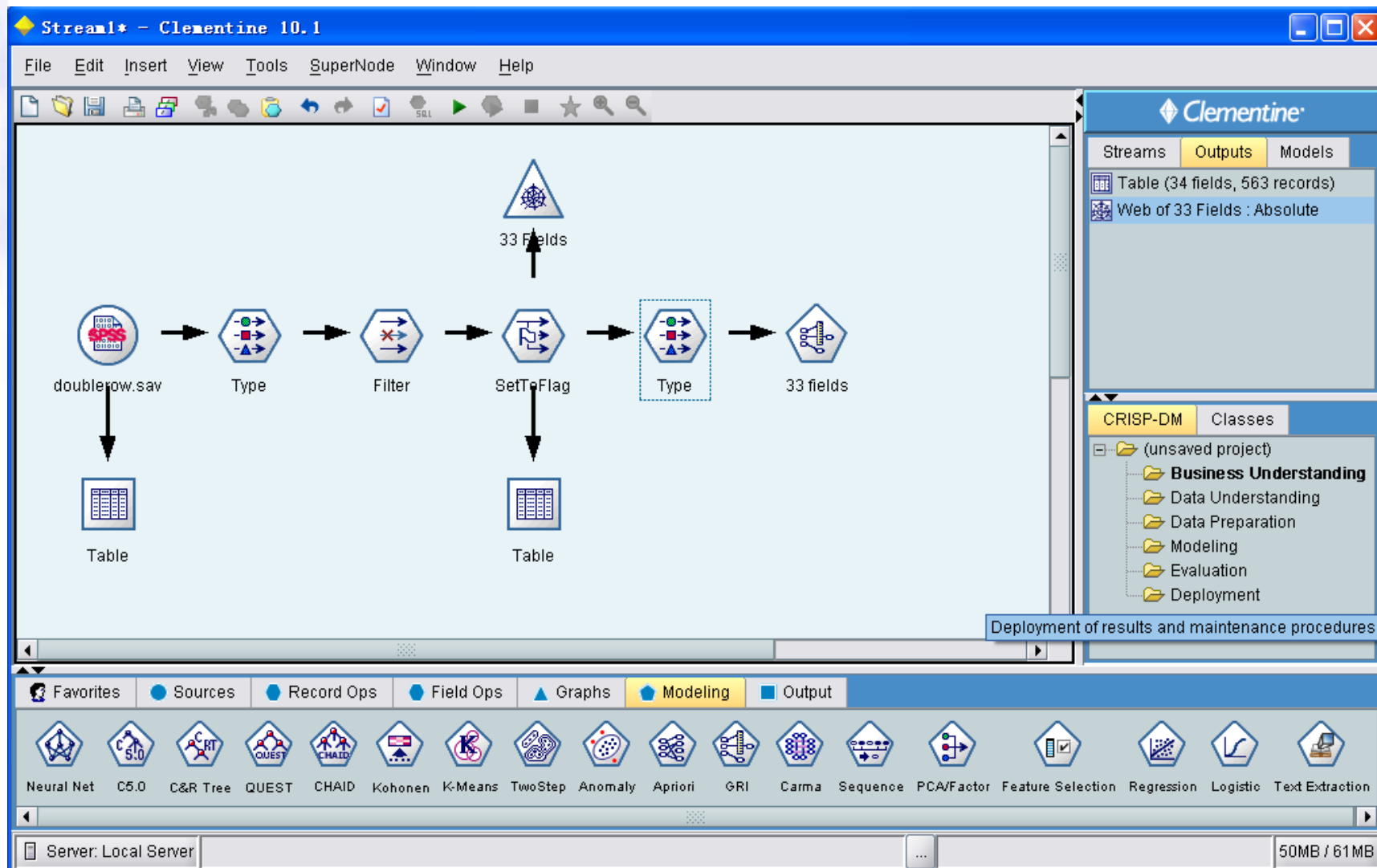




关系——DNA——结构



数据挖掘——挖掘关联规则



丰富的数据挖掘算法

决策树模型



C5.0



C&RT



QUEST



CHAID

聚类模型



Kohonen



K-Means



TwoStep



异常

关联分析模型



Apriori



GRI



Carma



序列

回归模型



回归



Logistic



Cox



Generalized
Linear

其他模型



神经网络



特征选择



主成分/因子分析



Bayes Net



SVM



Time Series



RFM Aggregate



RFM Analysis



Decision List



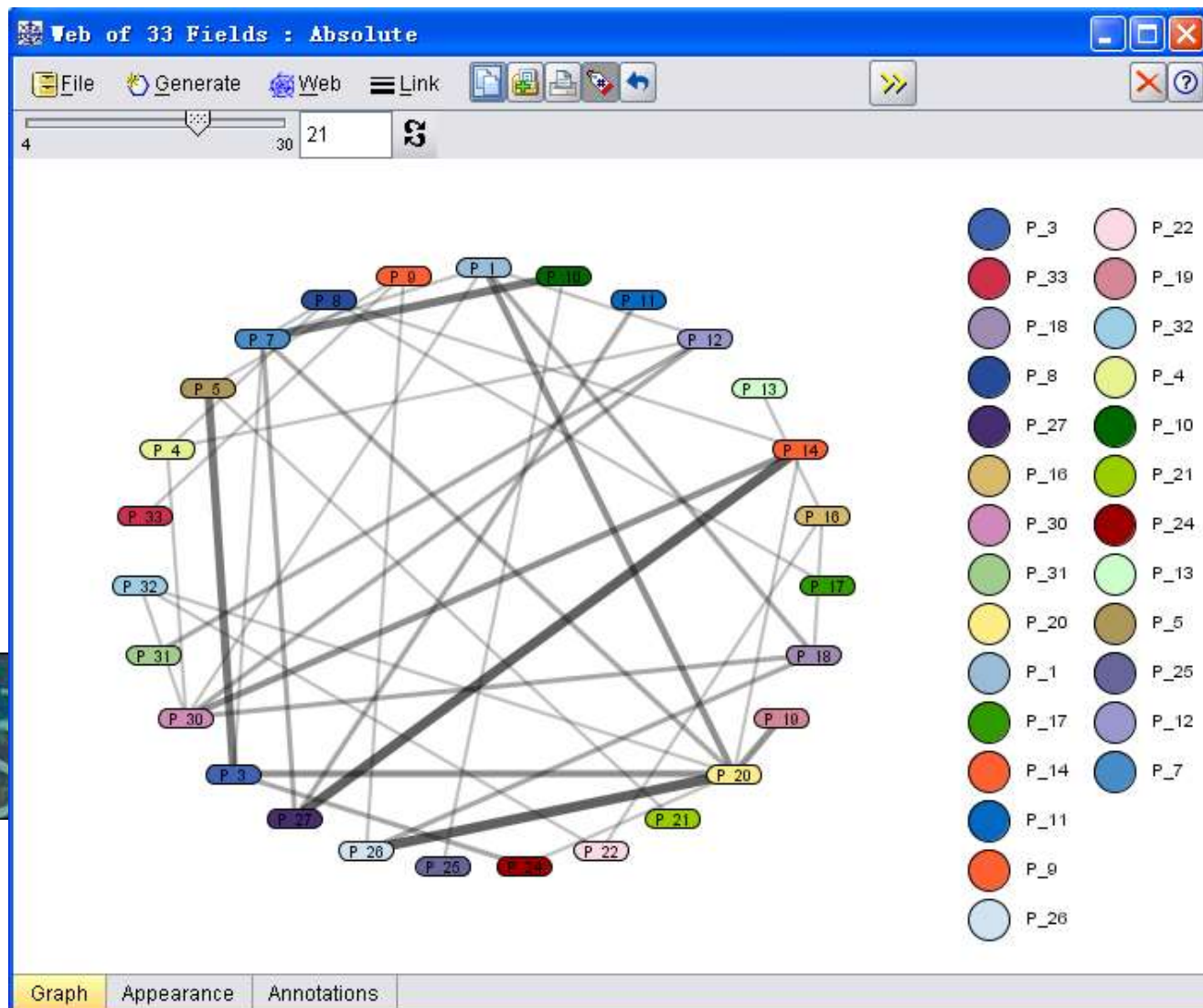
Binary Classifier



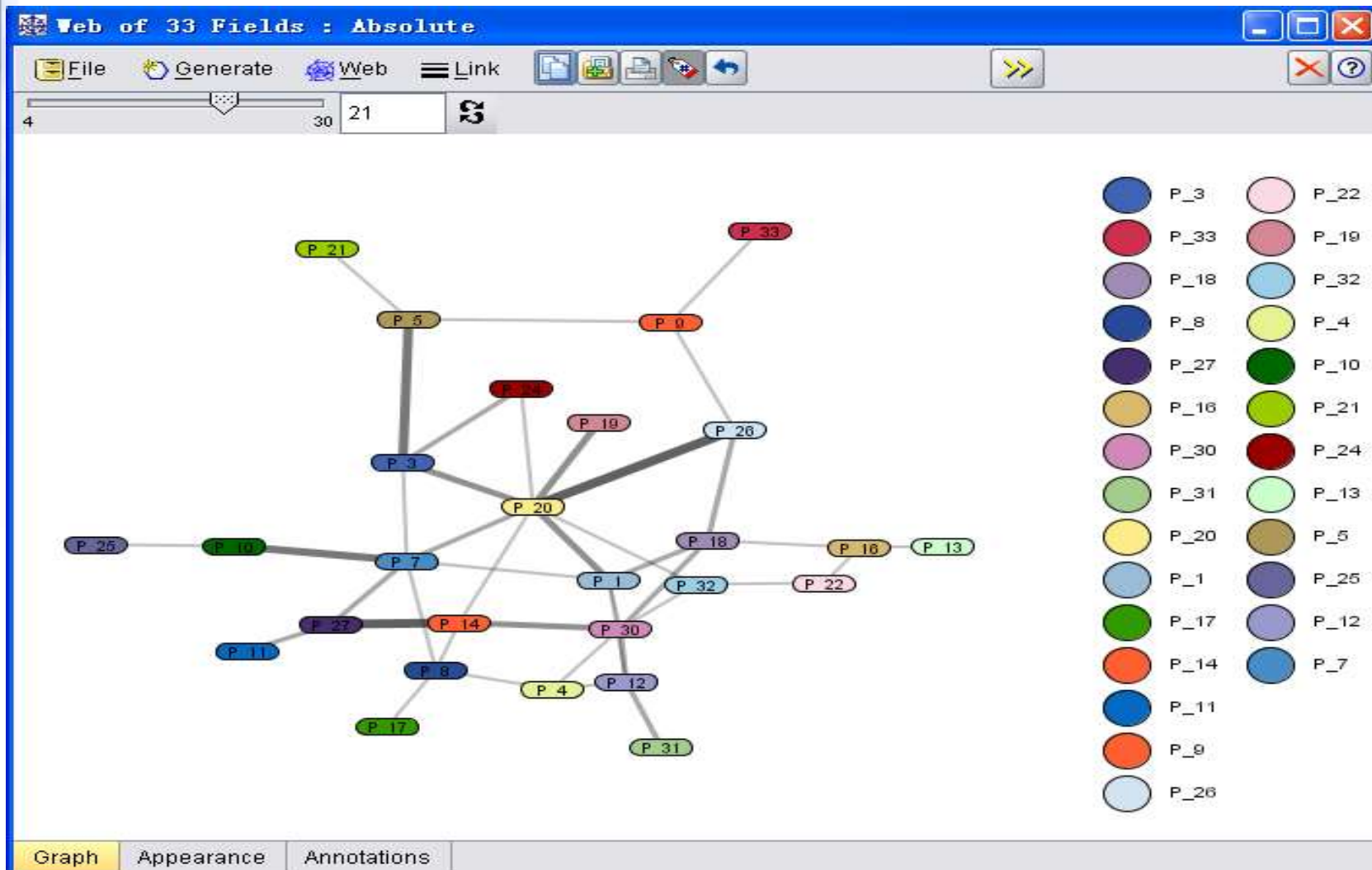
Numeric Predictor

关系——Web网络分析

关系的强弱



关系——Web网络分析



社会计算

微博抓取

文本存储结构

NLP分词技术

词性抽取

聚类与相关

语料库

可视化

社会计算

词频统计

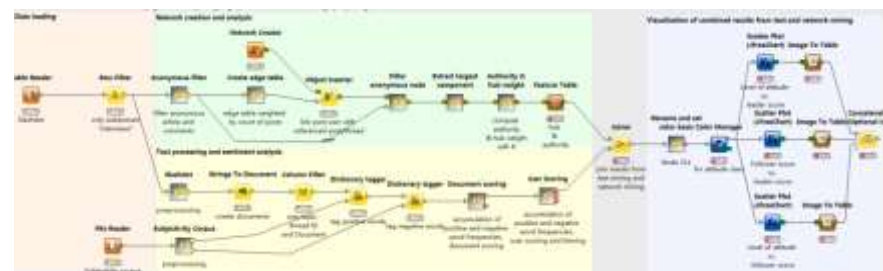
预测与判断

社会网络

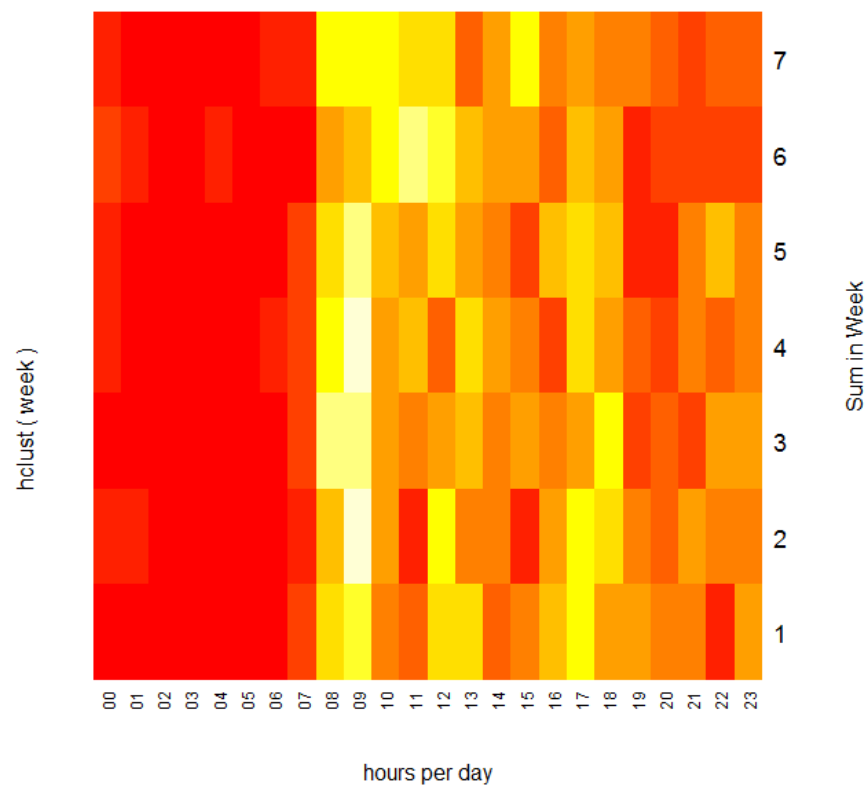
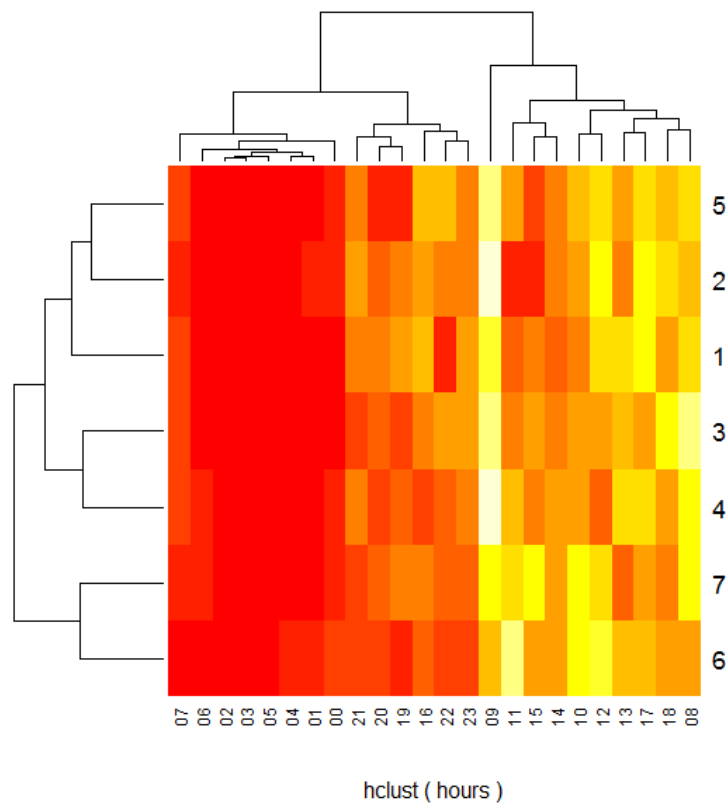
XML

规则与模型

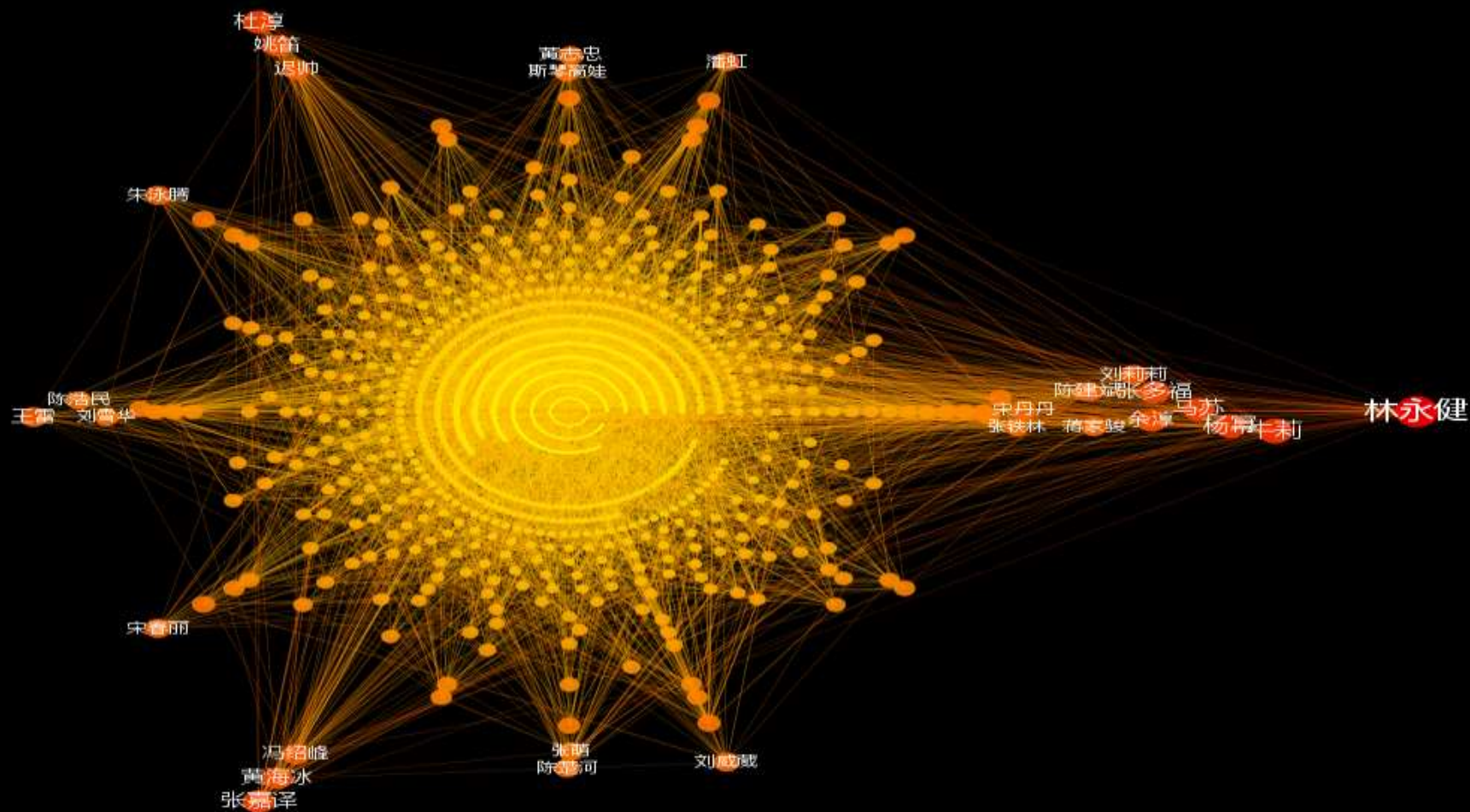
复杂网络



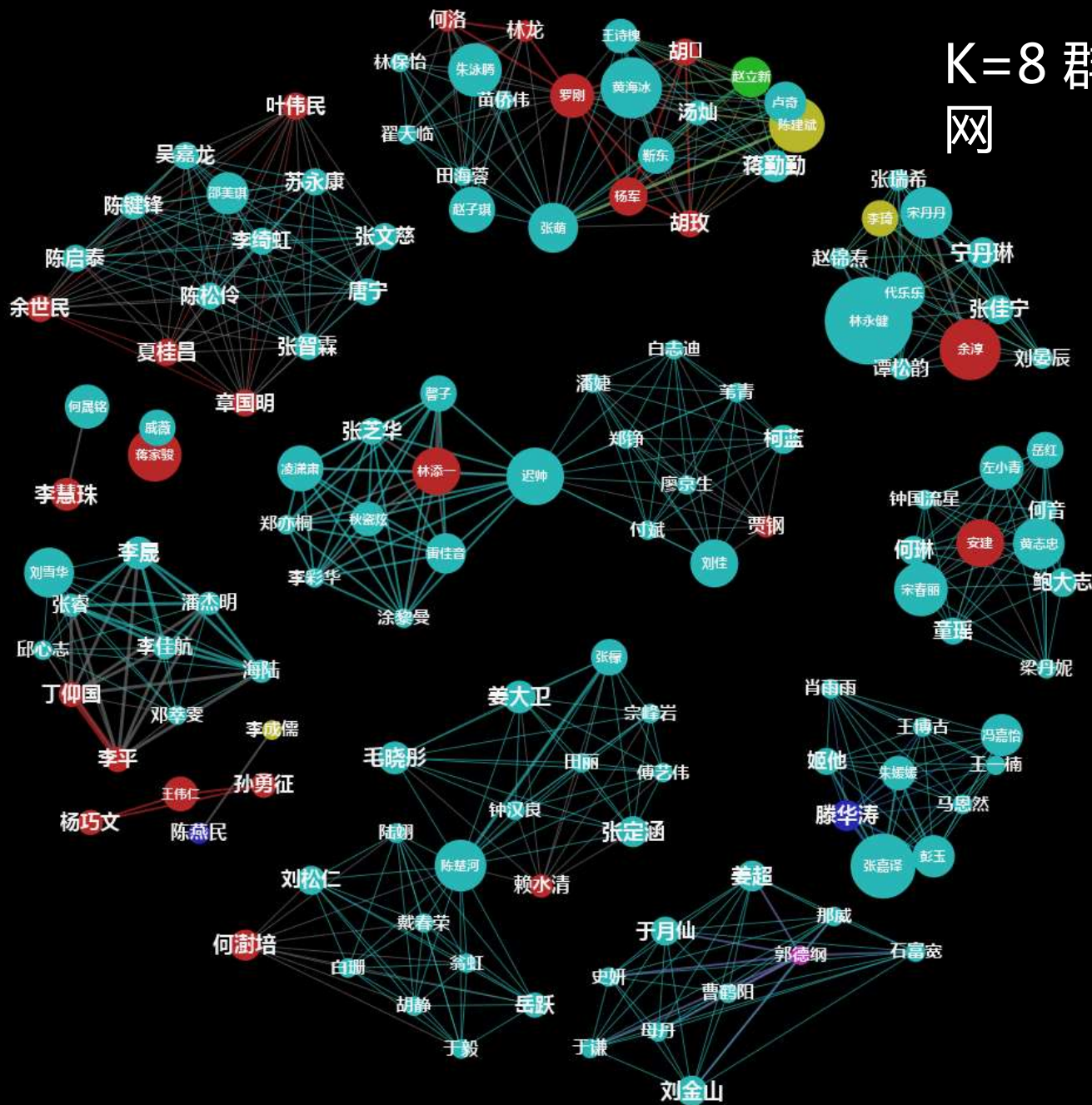
社交行为模式挖掘



黄金档电视剧 导演、编剧和主演的合作网



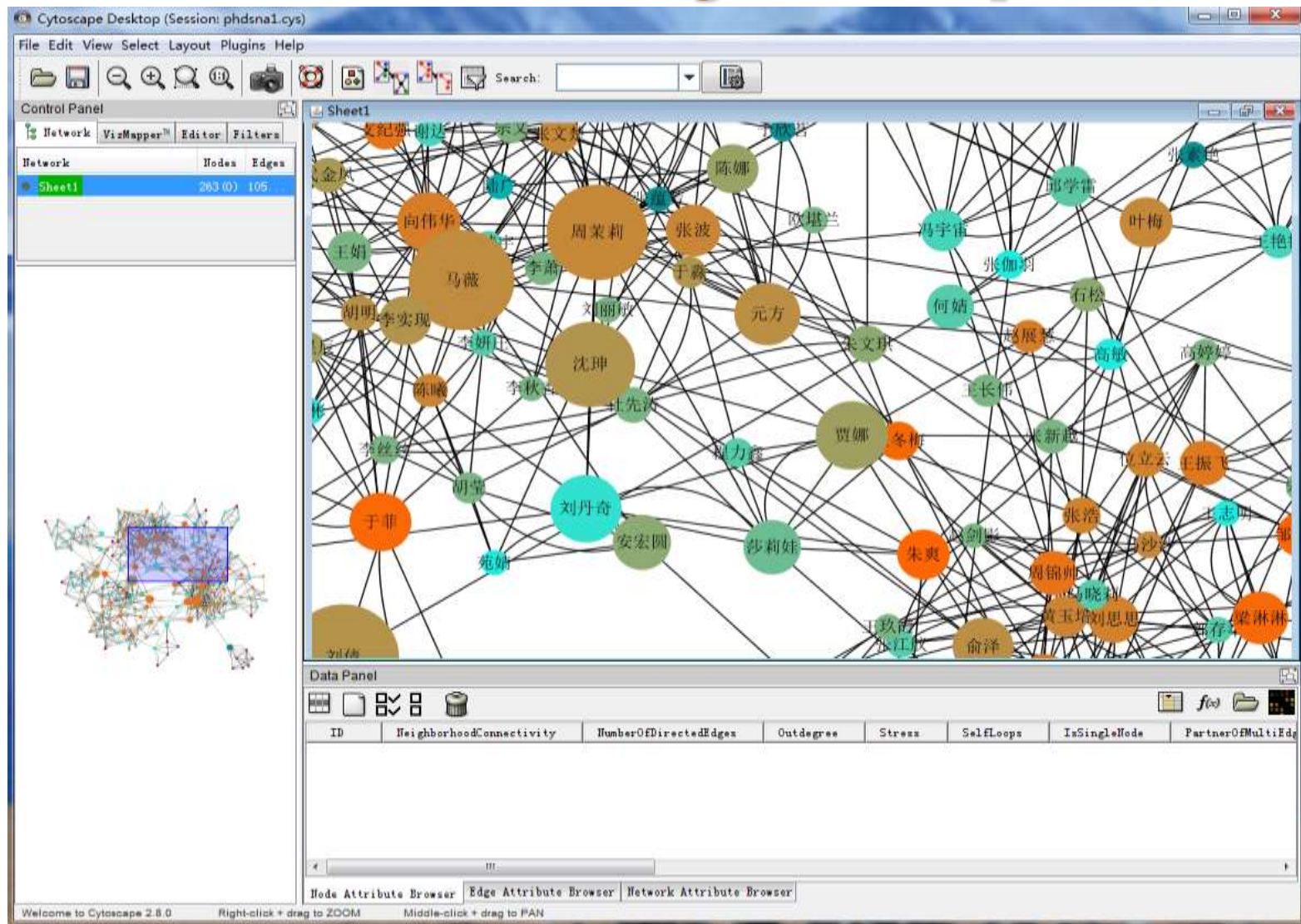
K=8 群体网



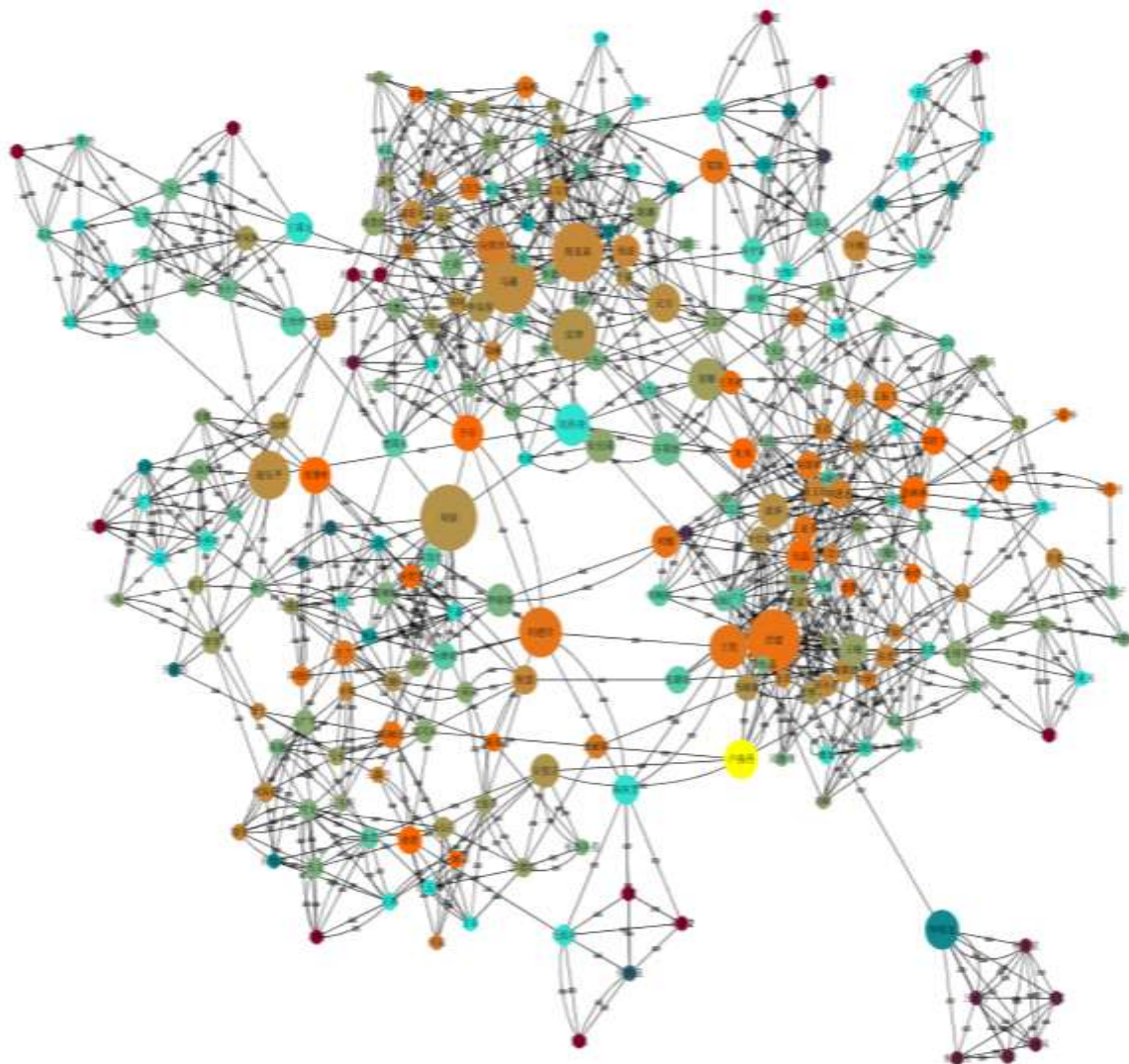


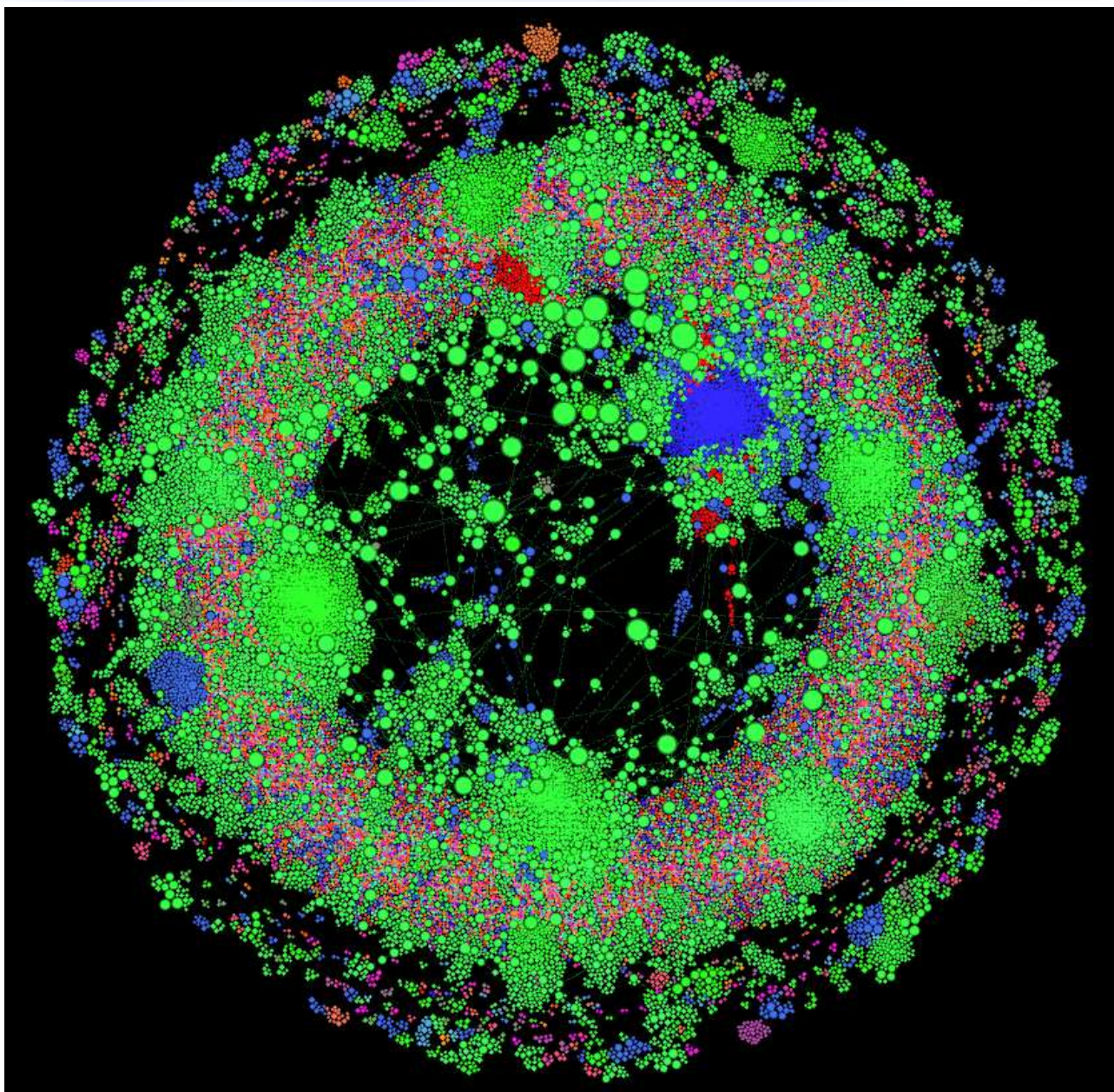


网络分析工具——Cytoscape

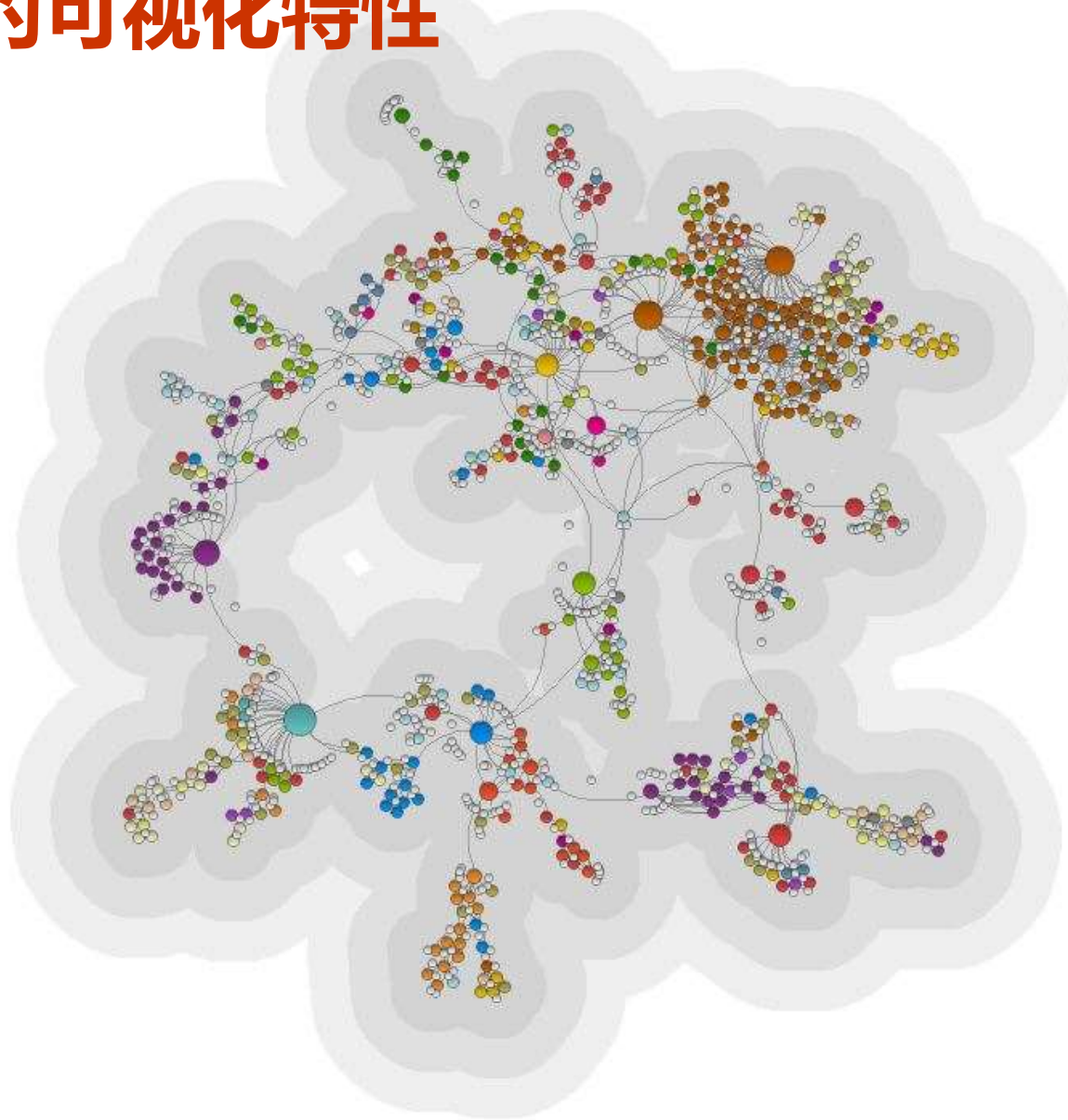


信息传播的关系网





网络的可视化特性





追踪流行病的扩散——病毒式营销

黑色是传染源或临床确诊感

染者，**粉红色**的潜在传染源，

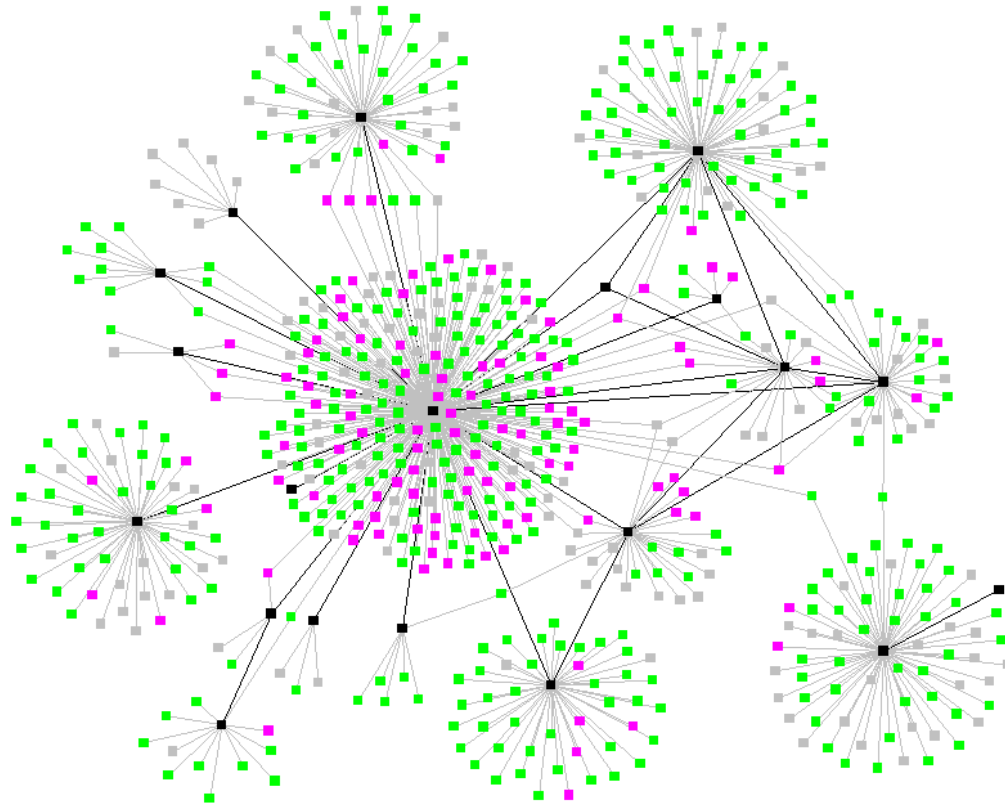
表示暴露在易感环境感染他人，

是没有确诊的感染疾病、**绿色**

代表暴露的人无感染和不是传

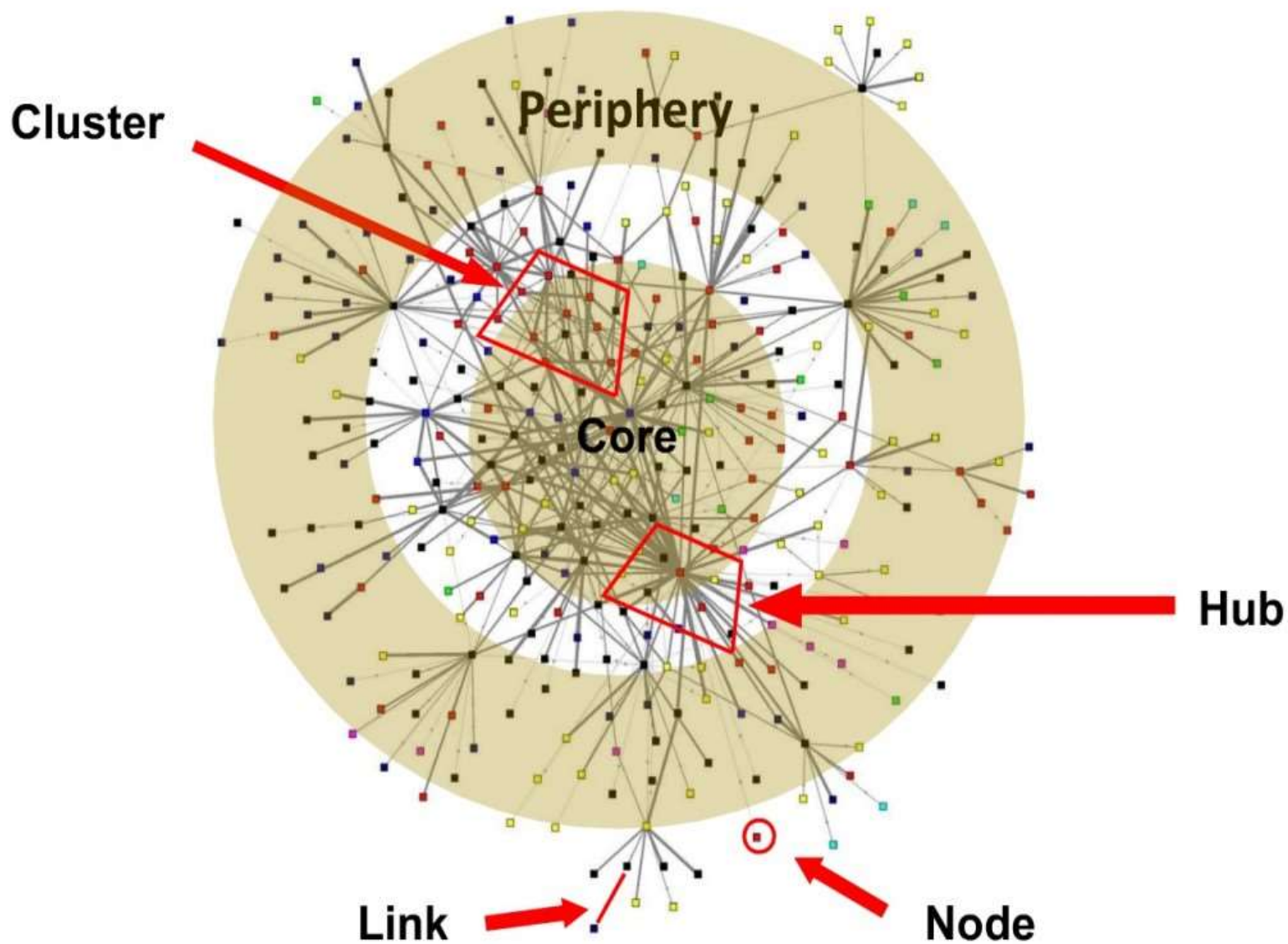
染性的。感染状况是未知的为

灰色的节点。



Source: Valdis Krebs, <http://www.orgnet.com/contagion.html>

结构主义思想





关系强度

强关系带来信任，

弱关系带来信息的传递！

发现意见领袖



基本概念：中心性测量



威望领袖
(Prestige leaders)

意见中介者
(Opinion brokers)

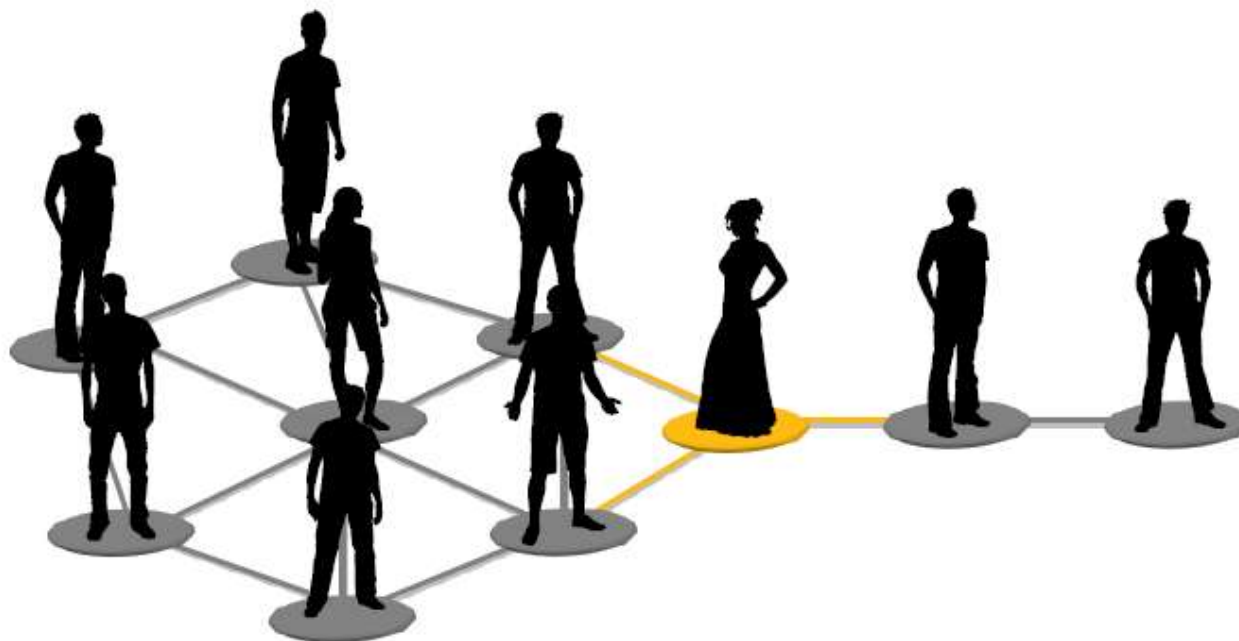
意见领袖
(Opinion leaders)

中心性测量：度中心性-Degree



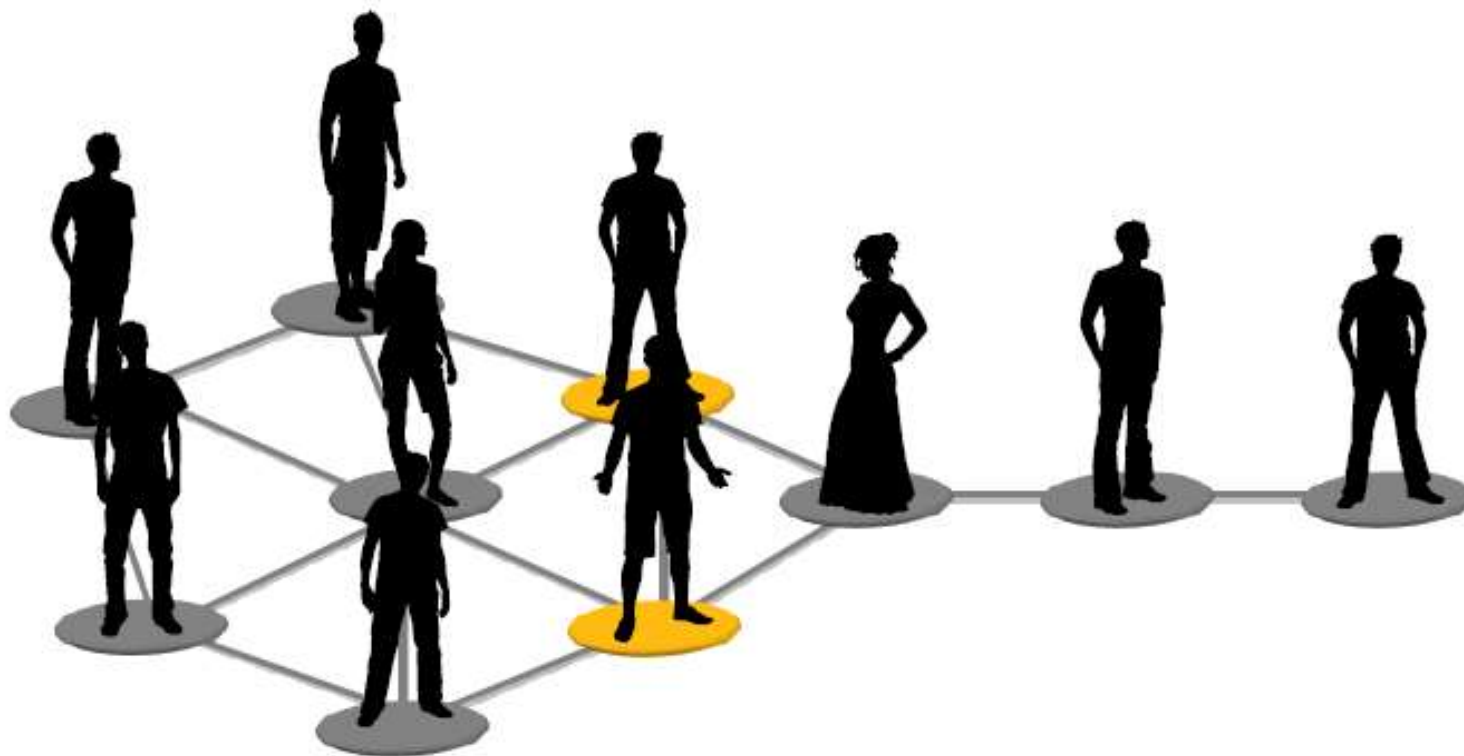
度中心性Degrees: 表示节点的链接数量
(出度与入度中心性)

中心性测量：中介中心性-Betweenness



中介中心性Betweenness: 在路径上能够到达其它节点的度量

中心性测量：接近中心性-Closeness



接近中心性Closeness: 有能力在最短路径到达其它点的节点度量

社会化营销

社会化媒体是一种重要的营销工具，它是企业发布信息和影响消费者，并收集反馈信息与之互动的重要渠道。如何从海量的关系数据中发现有价值的信息、建立精准营销的目标客户、分析客户价值模型是很多企业关注的问题。

企业对社会化媒体的认知和投入，将催生新媒介形态与产品营销思路。

网络分析、文本挖掘和意见挖掘



Data Insight

文本聚类分类
KOL意见领袖 PMML模型与云端部署
API接口 情感词典建设
归并文本挖掘与网络挖掘

规则建模推荐算法

RoambiAPP移动应用

网络抓数据

MYSQL和HADOOP存储

用户词典构建 cytoScape可视化分析 Xcelsius仪表盘
GEPHI动态可视化分析

API插件和接口程序
TABLEAU可视化分析
情感分析

Khime和R语言挖掘

R语言与分词
网络分析





Xcelsius
动态仪表盘

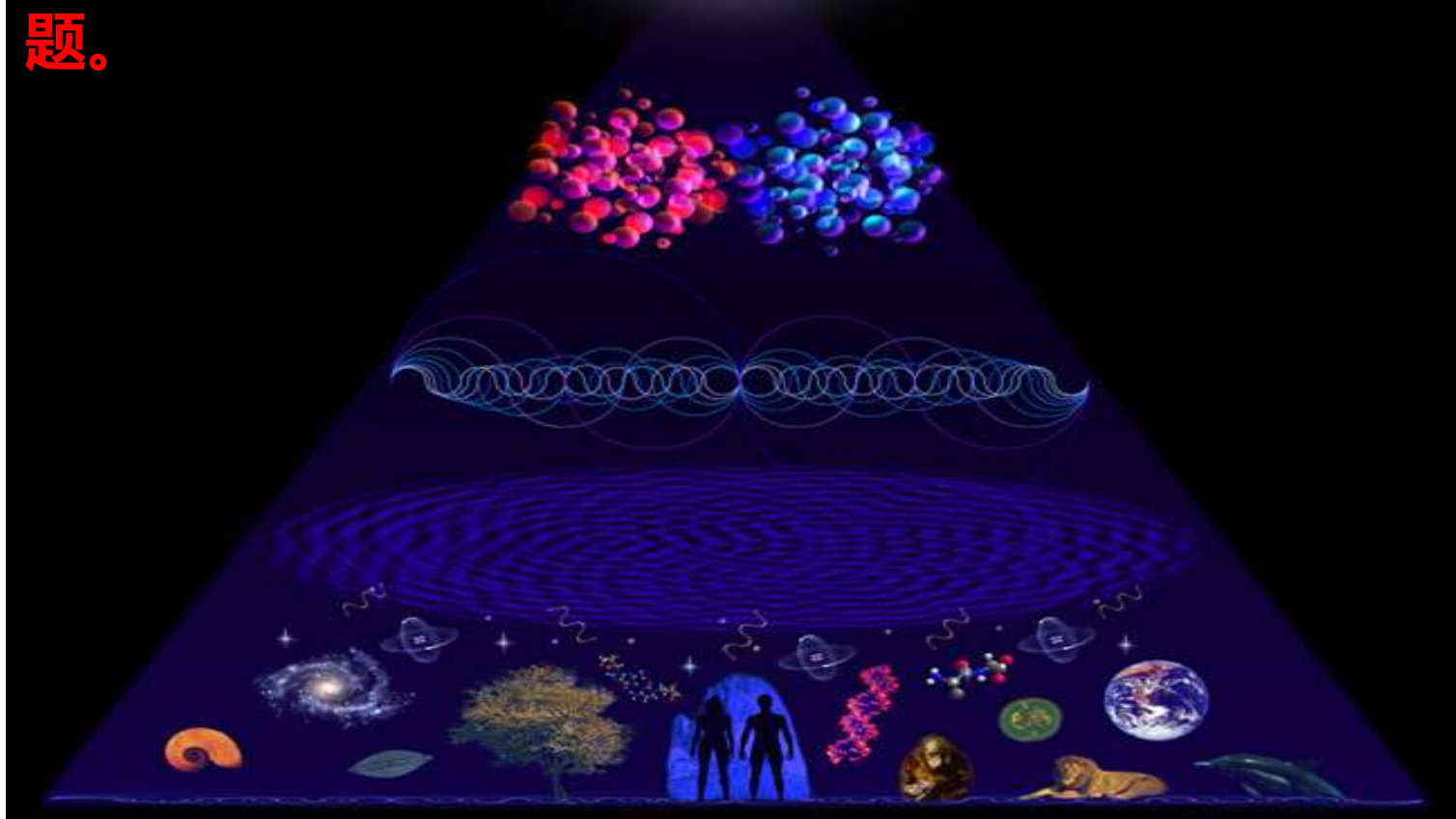
Tableau
在线分析

RoamBI
移动应用



预知社会

不管个人的偏好或思想是什么，个人行为如何加总而为集体行为的方式却是不相关的另外一个复杂问题。



网络安全与个人隐私

每个人都有自己的容忍的限度，
个人隐私：不同的人可能有不同的
理解！

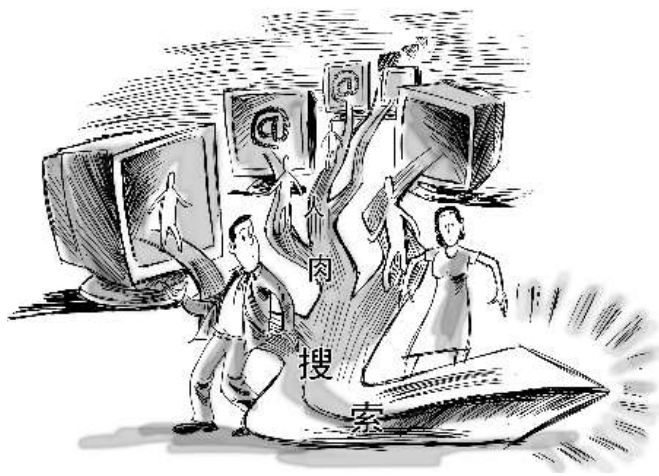


绝对安全、
相对隐私



个人隐私

孤立于社会或公众的注意之外，免受打扰！



不被公开或不受大众控制，只属于个人！

个人隐私是社会问题

社会生活中，每个人必须受到社会规范和制度控制。在一个相互联系的社会，完全自由的那个人是不存在的：一个人的自由或许是对另一个人的压迫。



每个公民必须服从一套简单的规则，这些规则必须被强制执行，特别是对政府、企业和个体。且公民要树立隐私保护意识。

个人隐私：共性或个性

一些人不愿意让别人知道
自己的任何信息！



一些人却希望自己生活的
每一个细节都展示给
世界！

网络科学的角色











**网络科学技术在定义隐私、
保护隐私和侵犯隐私方面都扮
演者重要角色！**





社会科学的研究春天来了！









欢迎光临@沈浩老师的博客


沈浩老师的博客  档案  微博  日志  相册  视频  分享  



 Data Insight

 Data Insight Presentation

 首页  置顶日志  Data Insight  图片浏览  推荐好书

 使用此博客主题