数据为王时代下的数据安全

张 敏

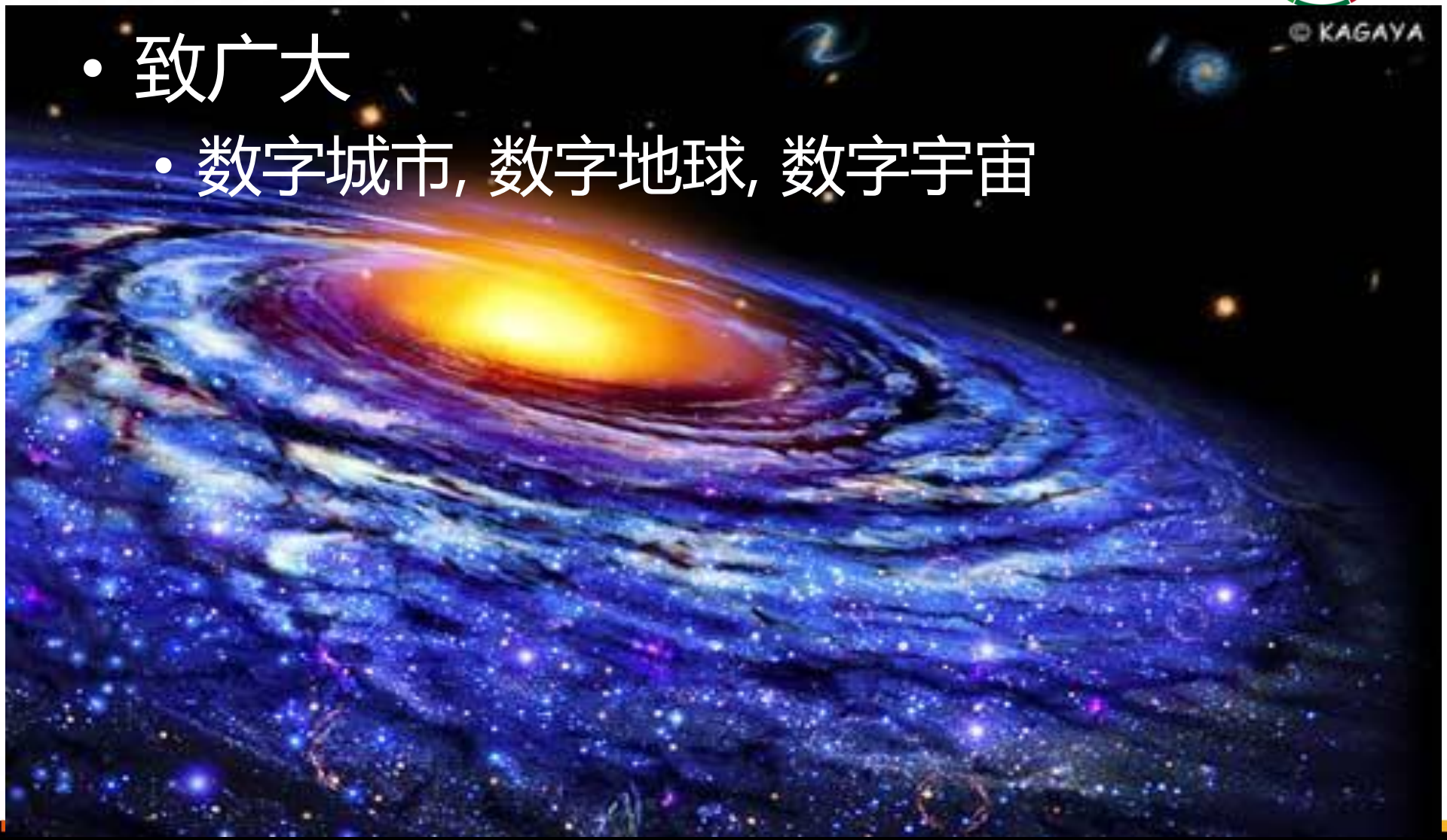
中国科学院软件研究所

2014年9月

内容提要

- 数据为王时代
- 安全与隐私问题
- 科研&技术进展
- 结束语

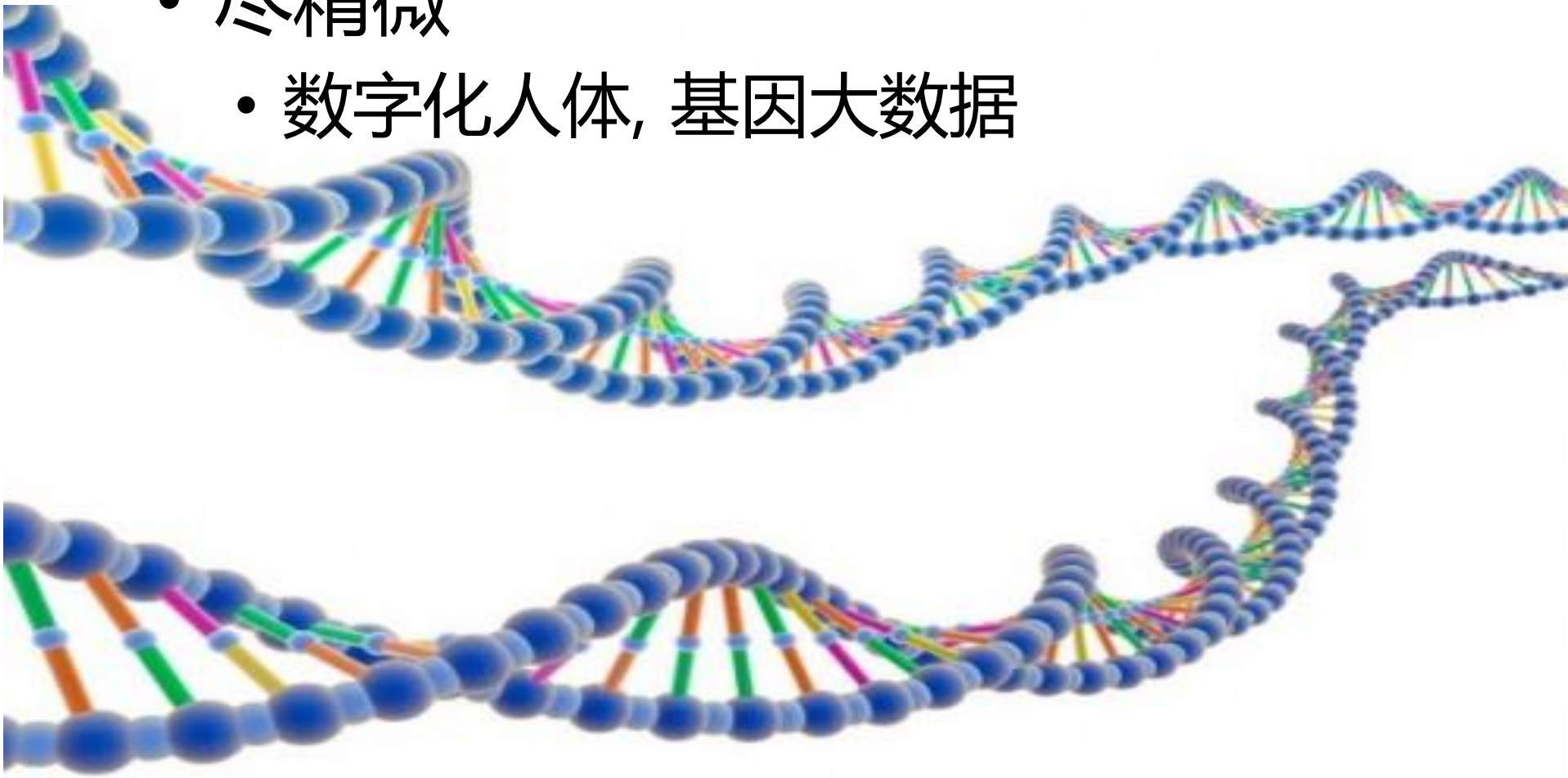
数据为王



- 致广大
  - 数字城市，数字地球，数字宇宙

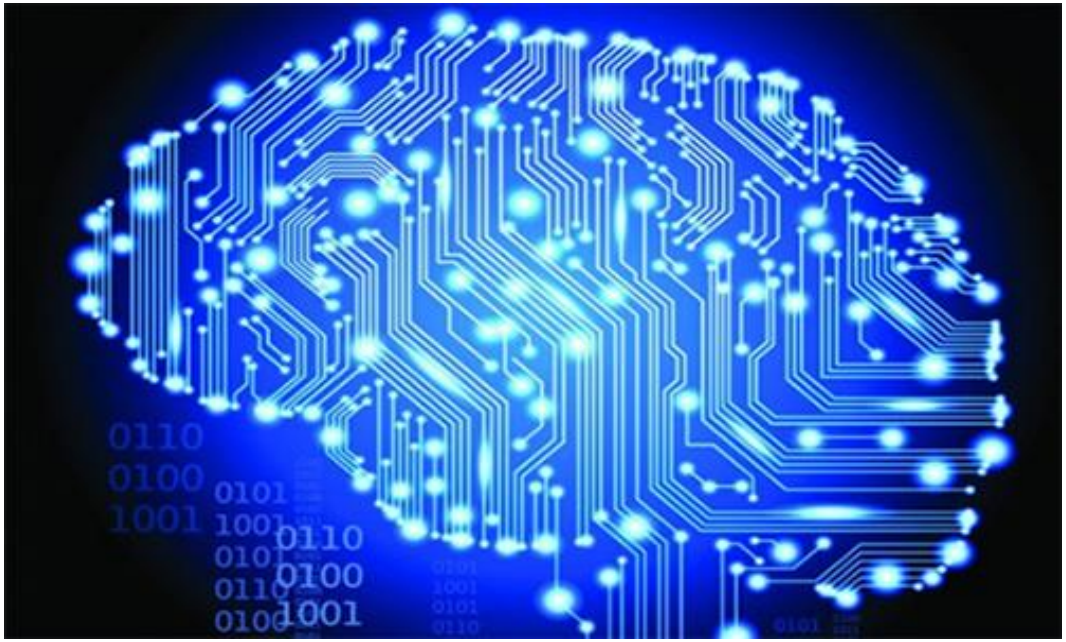
数据为王

- 尽精微
  - 数字化人体, 基因大数据



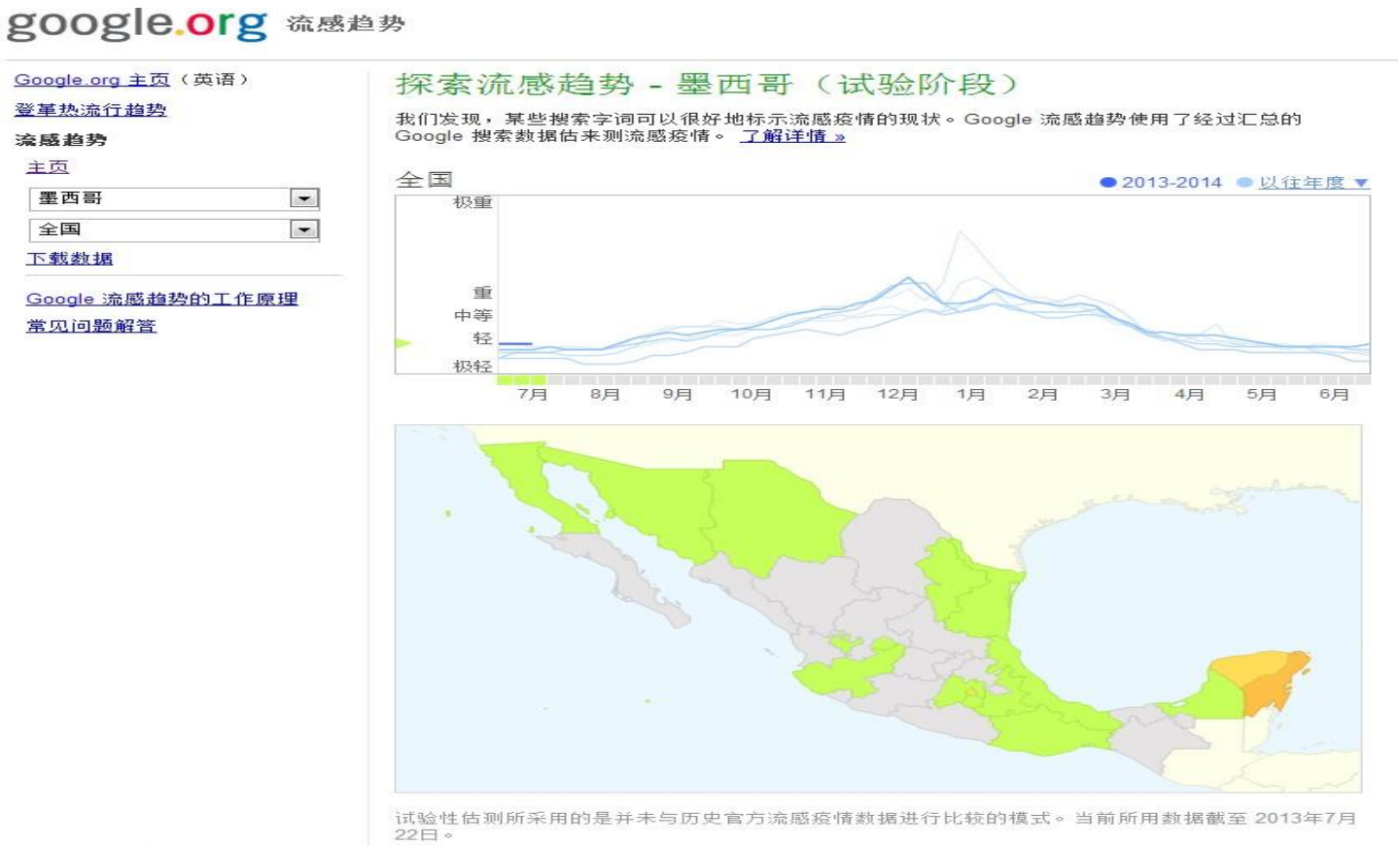
大数据分析

- 基于大样本空间发现统计规律
- 基于历史行为分析个体规律
- 多种信息源集成形成知识



1.发现统计规律

http://www.google.org/flutrends/



Detecting influenza epidemics using search engine query data

2.分析个体规律

Reliable data 40 days in advance

TempRisk is the first commercial application with scientifically validated statistical methods that analyze the risks for extreme heat and cold events up to 40 days in advance.

An independent approach

TempRisk provides independent, objective, and unbiased information that demonstrates a significant link between past global weather conditions and future extreme temperature events.

Breakthrough methodology

TempRisk's state of the art software tools allow you to make better weather driven decisions.

3.提升形成知识


Google 搜索前沿
首页 搜索原理 搜索技巧 功能 搜索故事 搜索乐园 帮助
达·芬奇
现在搜索某个人、地点或事物，
就会看到由"知识图"技术提供的新信息。
继续 ›


The Knowledge Graph
Learn more about one of the key breakthroughs behind the future of search.


列奥纳多·达·芬奇
列奥纳多·达·芬奇，又译达文西，全名列奥纳多·迪·瑟皮耶罗·达·芬奇，是意大利文艺复兴时期的一个博学者：除了是画家，他还是雕刻家、建筑师、音乐家、数学家、工程师、发明家、解剖学家、地质学家、制图师，植物学家和作家。他的天赋或许比同时期的其他人物都高，这使他成为文艺复兴时期人文主义的代表人物，也使得他成为文艺复兴时期典型的艺术家，也是历史上最著名的画家之一，与米开朗基罗和拉斐尔并称文艺复兴三杰。小行星3000被命名为"列奥纳多"以表示人们对列奥纳多·达·芬奇的尊敬和怀念。
维基百科
生于: 1452 年 4 月 15 日, 芬奇
搜索结果中发现更多
蒙娜丽莎 1506 年
最后的晚餐 1498 年
维特鲁威人
岩间圣母 1486 年
抱银鼠的女子 1490 年
用户还搜索了
米开朗基罗
拉斐尔·圣齐奥
巴勃罗·毕卡索
文森特·梵高
皮埃尔-奥古斯特·雷诺阿
反馈/详情

安全与隐私问题

引子

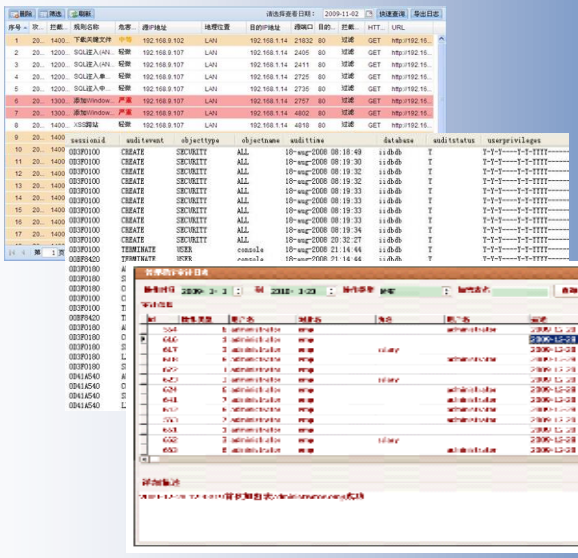
光顾着看围脖留言忘记给老爸指路！都开到中关村了：（爸爸开始叨叨我说导航吧

@王珞丹V：爸爸送我和小6去给＜无人驾驶＞配音的路上 原文转发(154) | 原文评论(310)


TARGET


BEST BUY

另一个例子



通过电子邮件、社交网络扫描
发现员工异常，提醒企业防止其泄密

社交网络中的隐私信息

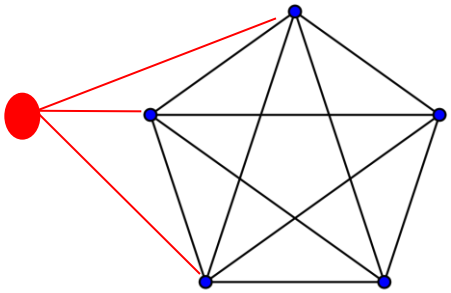
1. 身份匿名

去匿名化De-anonymity

- 基于特定模式精确匹配
- 基于种子匹配
- 基于相似度的匹配

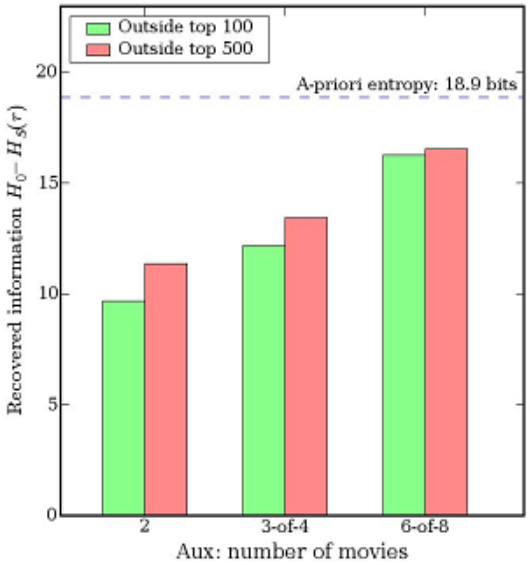
IMDB影评库

节点重识别攻击



Backstrom等人对攻击的攻击方式进行了划分，认为攻击者可通过主动或被动方式生成识别度高的社交结构，并与攻击目标连接，从而实现在匿名后的图中重新识别攻击目标的目的。

Backstrom, L.etc. ‘Wherefore art thou r3579x? Anonymized social networks, hidden patterns, and structural steganography’ , World Wide Web 2007



Narayanan等人利用多个其他社交网络的信息作为背景知识，识别出攻击目标发布的匿名图中的某些特定节点作为种子节点，利用种子节点，进一步实现其邻居节点的识别。

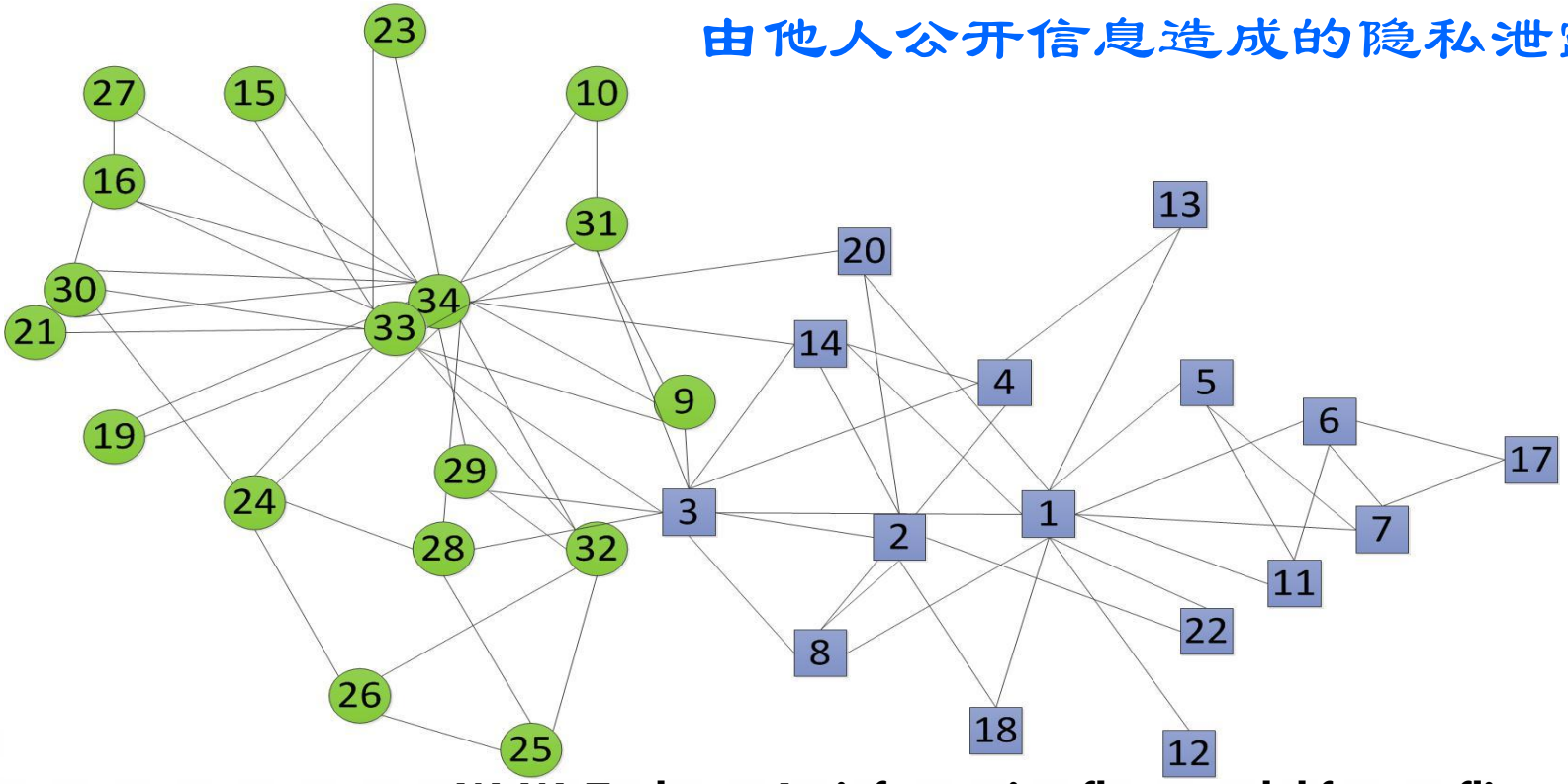
Narayanan A, Shmatikov V. Robust De-anonymization of Large Sparse Datasets2008

De-anonymizing social networks. Security and Privacy, Security and Privacy， 2009

2.属性匿名推测

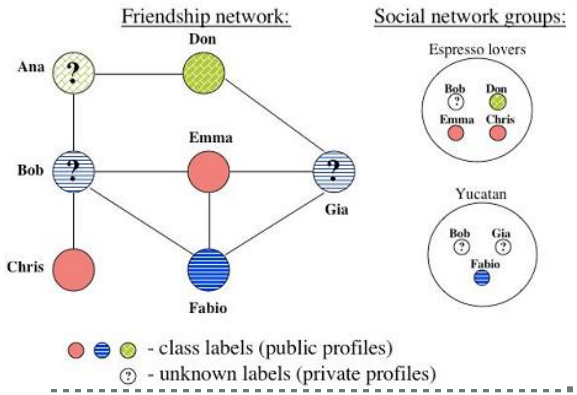
群组倾向性预测

由他人公开信息造成的隐私泄露



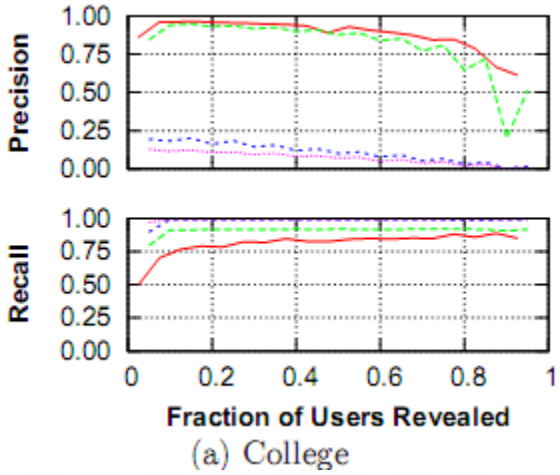
W. W. Zachary, An information flow model for conflict and fission in small groups, Journal of Anthropological Research

属性重识别攻击



Zheleva等人研究发现，参与同一小组的用户倾向于具有相似的属性。并可利用用户的群组标签对用户可能具有的属性进行预测。

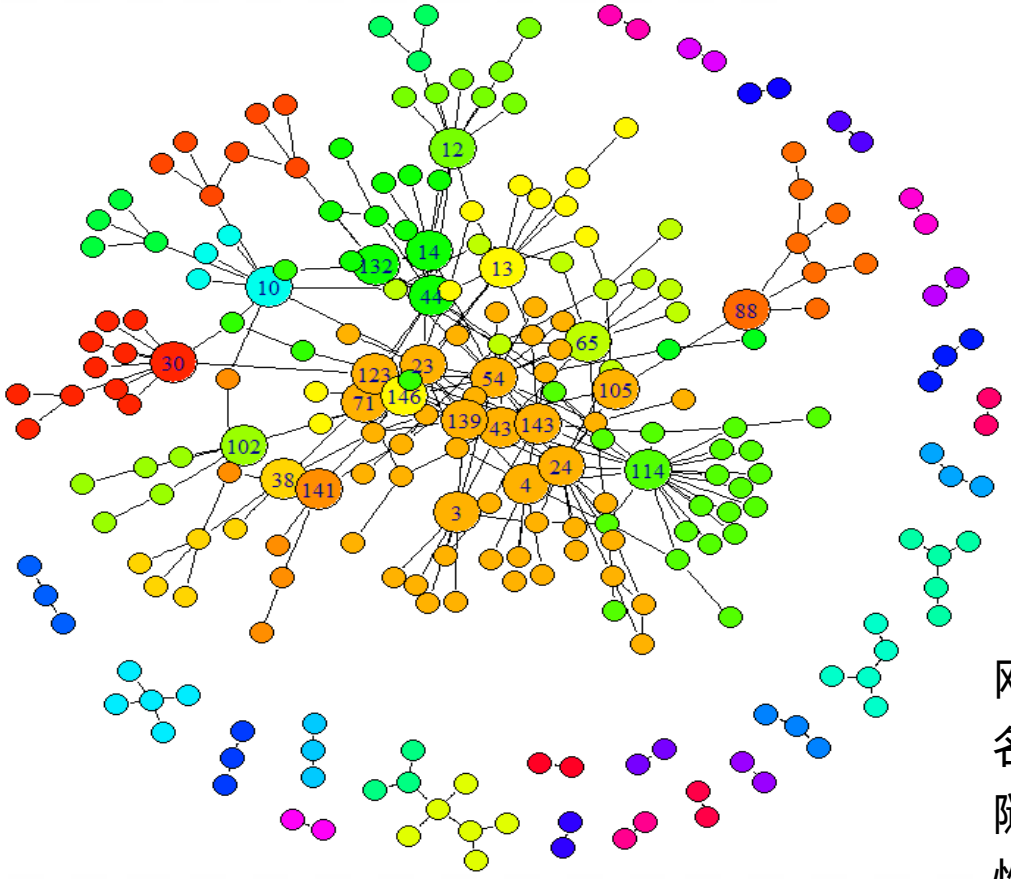
Zheleva E, Getoor L. To join or not to join: the illusion of privacy in social networks with mixed public and private user profiles. WWW 2009



Mislove等人研究发现，用户可能与其好友具有类似的属性。可以通过好友的公开信息对用户未公开的信息进行推测。

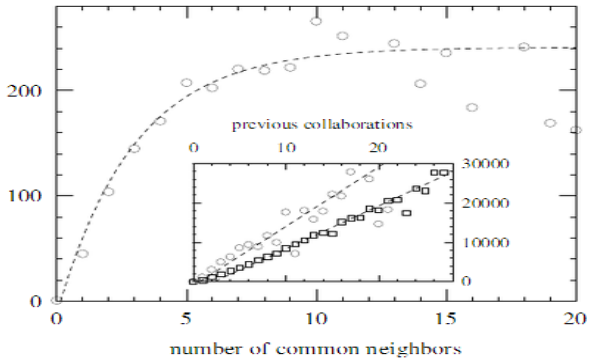
Mislove A, Viswanath B, Gummadi K P, et al. You are who you know: inferring user profiles in online social networks.In Proceedings of the third ACM international conference on Web search and data mining. ACM, 2010: 251-260

3.关系匿名推测



边匿名猜测攻击：社交网络中群组的存在，使得用户之间的匿名联系仍有可能推测出来。简单边匿名、随机边匿名方案匿名效果不理想，可用性差。

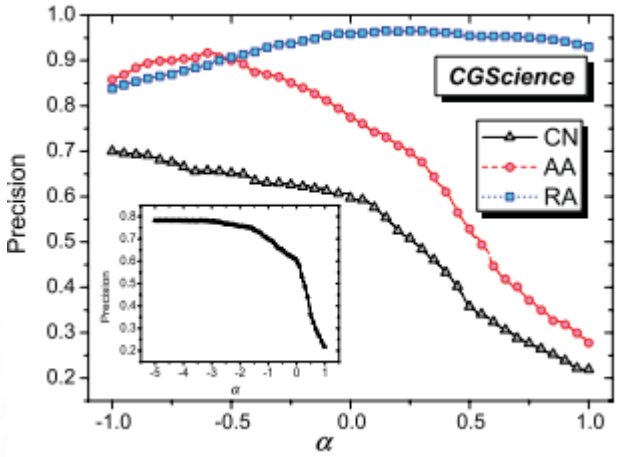
连接关系重识别攻击



Newman等人发现，两个用户间的共同朋友越多，两者间具有连接关系的可能性越大。提出根据共同朋友预测连接关系的模型

Adamic等人分析了节点间共同朋友的度数与节点间建立连接可行性之间的关系，Zhou等人建立了资源分配模型对节点间信息流动进行分析，并预测连接关系。

Zhou T.etc. Predicting missing links via local informatio]. The European Physical Journal B, 2009, 71(4): 623-630.



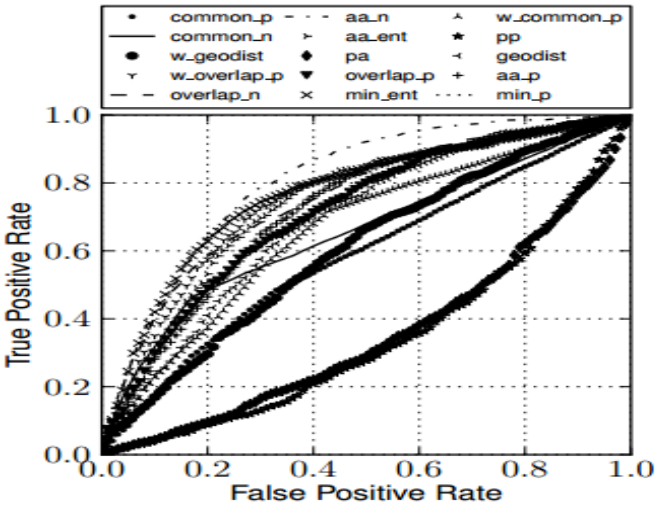
Zhou等人发现，在某些社交网络图中，与共同朋友间具有弱连接的两者，更容易形成朋友关系。据此提出了基于弱连接的朋友关系预测。

Lü L, Zhou T. Link prediction in weighted networks: The role of weak ties[J]. EPL (Europhysics Letters), 2010, 89(1): 18001.

4.位置隐私


@杨承平V：【深圳高二女生外出遇害 疑为玩微博定位惹祸上身！】宝安职业技术学校高二女生赖曾裕童@Mysshl 12日晚失踪，13日尸体被发现。鉴定系他杀，警方正全力侦破此案。她之前经常在微博晒自拍照和定位，在此提醒广大网友，在微博上保护好自己的隐私，避免过多定点自拍，女性尤甚，转发周知@袁裕来律师 @何兵
收起 | 查看大图 | 向左转 | 向右转
复习 @港湾茶餐廳 http://t.cn/zjBysET
广东省.深圳市.宝安区.电信路 - 显示地图
1月11日15:33 来自Instagram | 举报

位置-社交关系攻击



(c) Place-social

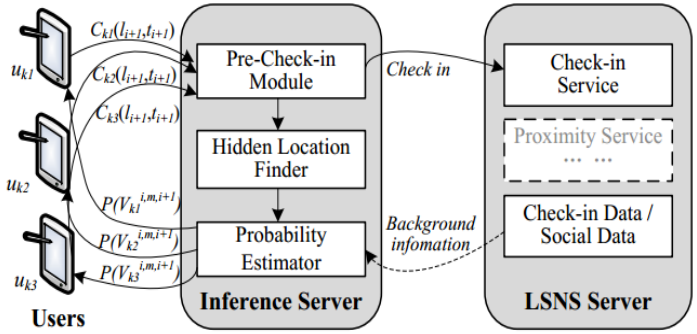


Fig. 4. System Architecture

Salvatore Scellato等人在社交网络中使用社交关系和用户的签到历史信息，对稀疏的预测空间进行压缩，通过机器学习的方法对用户的社交关系进行预测，并取得了良好的预测效果。

Exploiting place features in link prediction on location-based social networks 2011

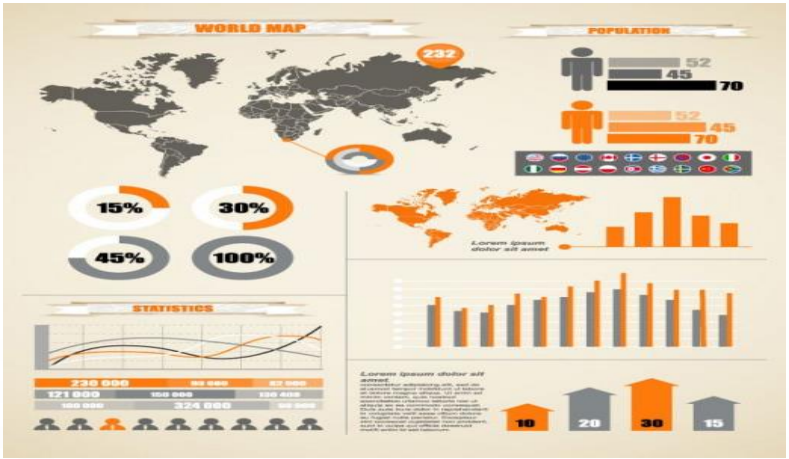
作者Huo Zheng等人提出一种安全的社交网络签到系统框架，该系统以用户之间的社交关系、用户的签到历史轨迹信息以及地理位置信息为背景，对用户可能去过的位置进行预测，并将预测的位置反馈给用户，询问用户是否真实签到。

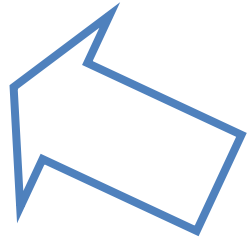
Feel Free to Check-in_Privacy alert against Hidden Location Inference Attacks in GeoSNs 2013

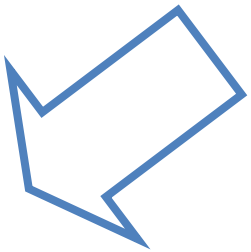
5.数据隐私

加密数据使用


合计：3422892
外科手术类
总计：1804125
非外科手术类
总计：1618767
整形手术在中国


WORLD MAP
POPULATION
STATISTICS
15%
30%
45%
100%





科研&技术进展

- 大数据访问控制
- 密文检索

基于风险的访问控制

访问需求无法明确预知

2011年， Wang等人指出，实际上用户对数据的访问需求往往无法明确预知。例如，医疗系统中很难确定医护人员为了治疗需要获取病患的哪些信息。针对该问题提出了基于风险的访问控制：根据资源内容和访问用户的属性对资源进行风险估计，设定风险阈值，控制用户内访问任意资源，只要所访问的数据风险总和小于风险阈值。对于新添加的资源，可以利用机器学的方法对其进行风险评估，形成一种自动的访问控制。

Wang, Q., & Jin, H. Quantified risk-adaptive access control for patient privacy protection in health information systems. ASIACCS2011

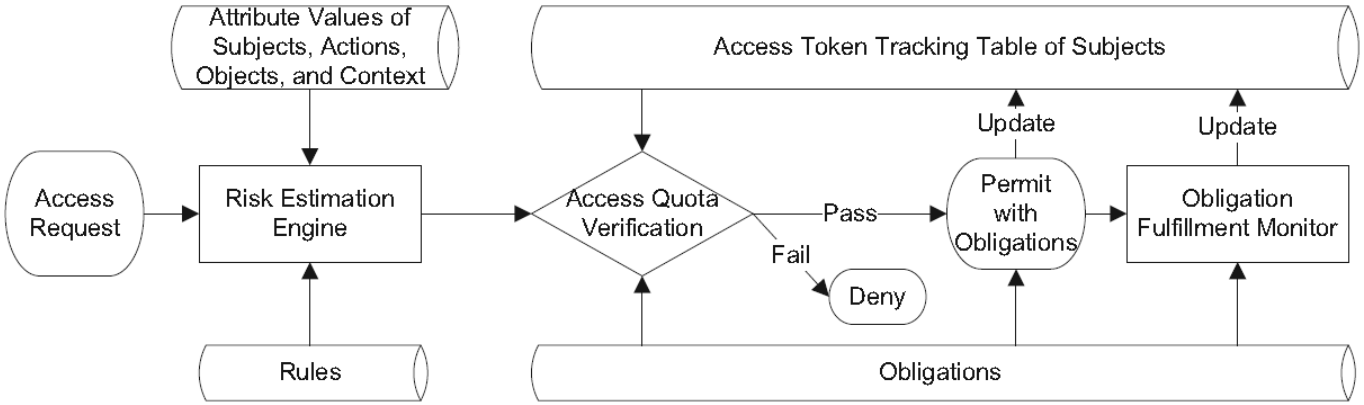


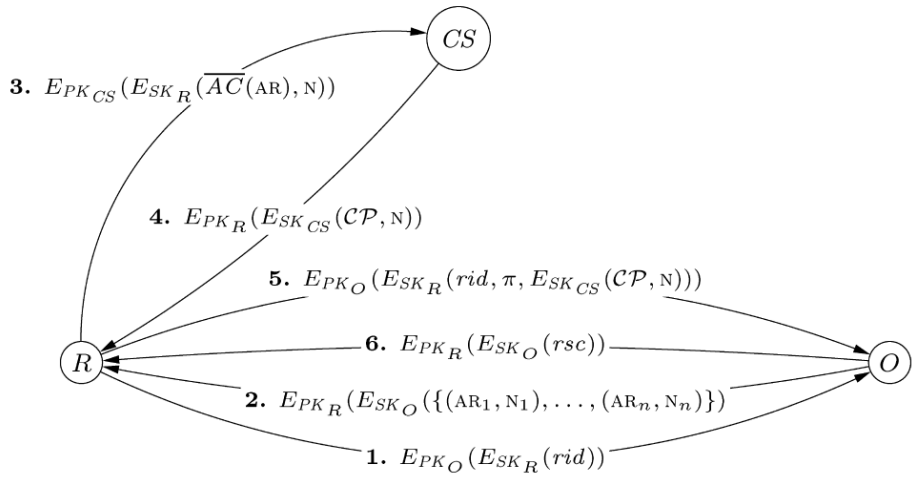
Figure 6: A General Method to Control the Overall Damage

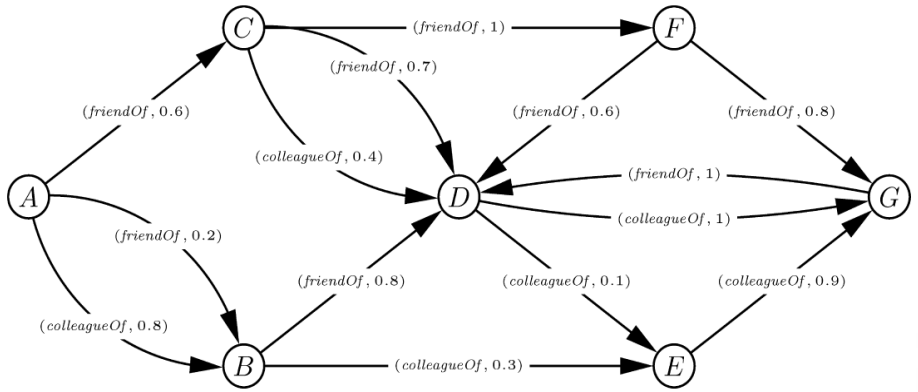
社交网络中的访问控制

1.基于关系结构的访问控制

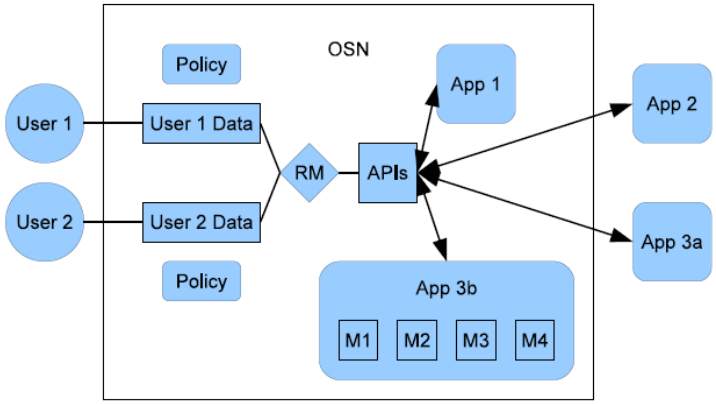
2009年， Carminati等人讨论了社交网络访问控制的需求，给出了一种半分散的结构实现访问控制。文中提出用直接好友之间的关系(朋友、同学、亲友等)和信任程度以及与间接好友的路径深度的计算其对间接好友的信任，作为授权访问的依据。同时给出了一个基于上述方法的访问控制协议。

Carminati, B., Ferrari, E., & Perego, A. (2009). Enforcing access control in web-based social networks. *TISSEC09*





2. 隐私保护的访问控制框架



2013年， Cheng等人提出一种用以保护用户隐私的在线社交网络(OSN)的访问控制框架。提出一种内外区分的社交网络体系结构：将第三方应用分成社交网络内外不同模块，并对对内部组件的交互进行细分，确保隐私数据相关模块只能在内部运行，控制隐私数据不流向外部。数据按照其敏感与否以及第三方应用对其需求程度分成四类，采用不用的访问控制策略。

Cheng, Y., Park, J., & Sandhu, R. Preserving user privacy from third-party applications in online social networks. WWW *companion* 2013

Table 2: Strategy

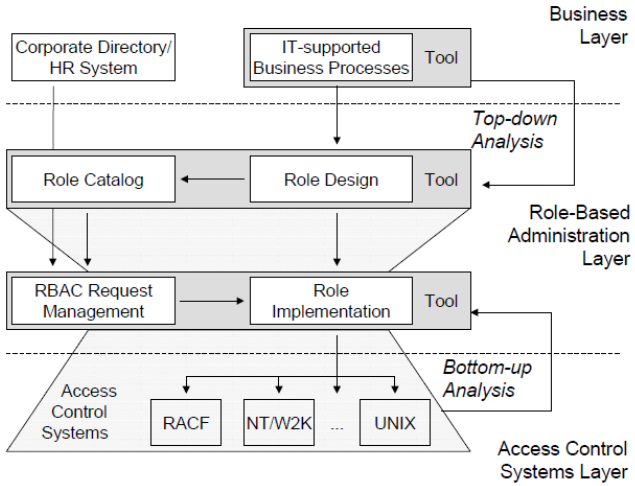
Data Classification	Strategy
unnecessary & private	do not permit
unnecessary & non-sensitive	user's choice
essential & non-sensitive	transmittable outside of OSN
essential & private	processable within OSN

Table 3: Module Types in Internal Components

	OSN provided	3rd-party provided
Communication w/ system calls	M1	M2
Communication w/ non-private data	M3	M4

角色挖掘

1. 基本角色挖掘问题



2003年，Kuhlmann等人首次提出了角色挖掘的概念。利用k-means聚类算法和数据挖掘技术，对权限进行聚类、从已存在的权限分配中寻找角色。并且经验性的给出了利用数据挖掘工具进行角色挖掘的7个基本步骤。

Kuhlmann, M., Shohat, D., & Schimpf, G. Role mining-revealing business roles for security administration using data mining technology. SACMAT03.

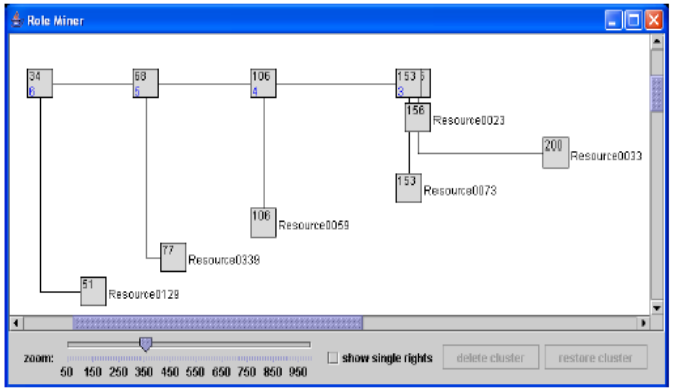


Figure 2: The cluster hierarchy view in ORCA

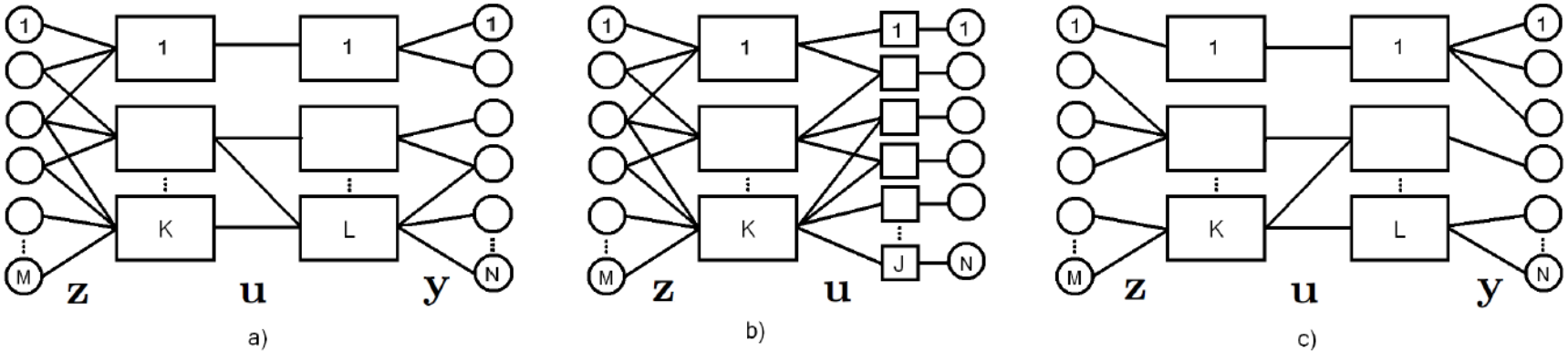
2003年， Schlegelmilch和Steffens对角色挖掘算法展开研究。提出了一种利用合成聚类进行角色挖掘的方法和ORCA角色挖掘工具。该算法对权限进行层次聚类：从初始权限集合 $S=\{\{p_1\},\ldots,\{p_n\}\}$迭代的计算相同用户数目最多的权限集合$S_i,S_j \in S$，将其聚类。最终形成一个树形结构的角色层级。

Schlegelmilch, J., & Steffens, U. Role mining with ORCA. SACMAT05.

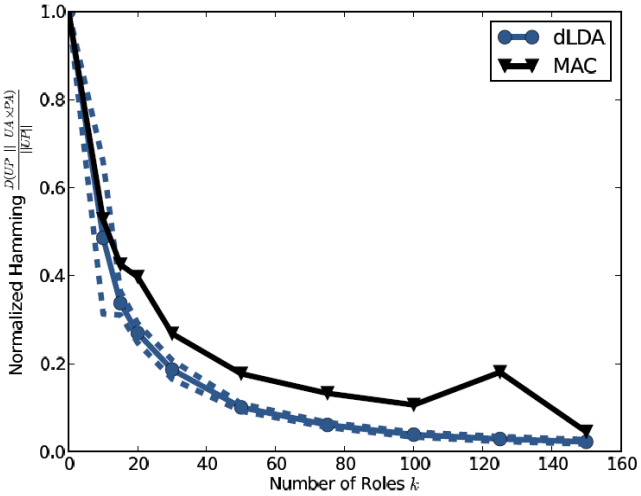
2. 角色挖掘概率模型

2008年，Frank等人指出传统组合模型的两大缺陷：1.原始数据重的错误会保留到生成的RBAC中；2.生成的角色难以解释。因此提出了用统计的方法构建角色挖掘概率模型。该模型将对称的角色拆分为业务角色和技术角色，利用最大似然原则推测未知的**z**,**y**,**u**矩阵使得**x**的似然性最大。由于采用统计的手段，当数据样本越大时，对模型的建立有很好的促进。

Frank, Mario, David Basin, and Joachim M. Buhmann. "A class of probabilistic models for role engineering." CCS08



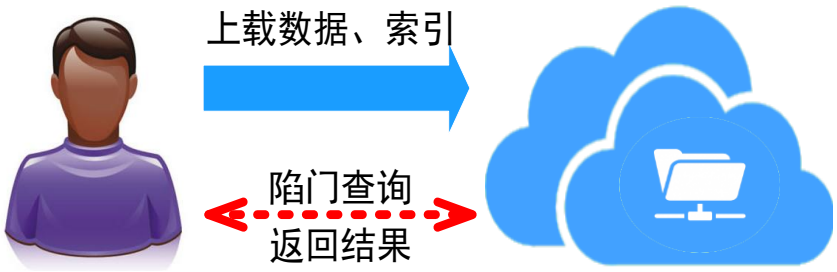
3. 利用多元数据的角色挖掘



2012年， Molloy等人提出利用除了用户和权限的对应关系之外的其他数据进行数据挖掘的想法（generative role mining）。采用数据挖掘和机器学习技术，从用户的访问日志、以及访问资源的属性等多元化的信息为权限分配权重，建立角色生成模型。

Molloy, I., Park, Y., & Chari, S. Generative models for access control policies: applications to role mining over logs with attribution. SACMAT12

密文检索



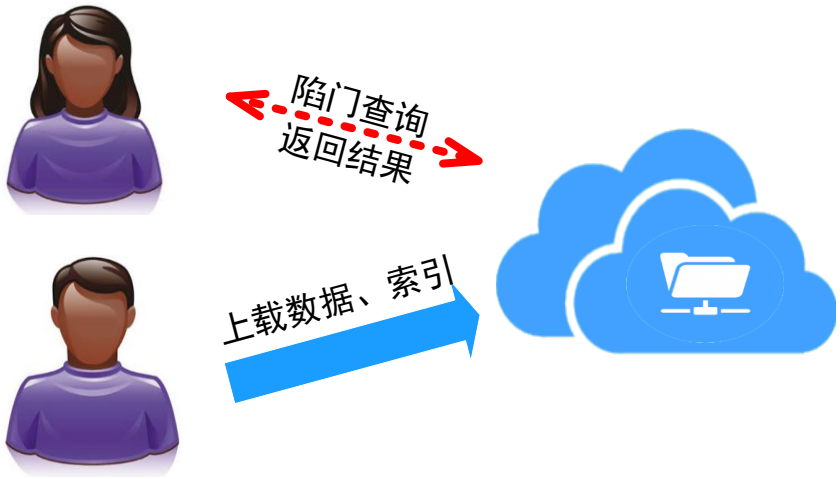
- 用户既是密文数据生成者，又是密文信息检索者
- 用户端计算关键词陷门
- 除极少数全文扫描类方案[song00]外，服务器端维护密文索引
- 服务器端根据索引返回与陷门相关的文档编号集合

索引构造方式

- 关键词集合式
- 加密链表式
- 位图索引式
- 多项式判定式

对称可搜索加密（SSE）

非对称可搜索加密（ASE）



- 密文数据生成者与密文数据检索者并非同一人
- 数据生成者利用接收者的公钥加密数据，同时生成关键词陷门
- 数据接收者利用自己的私钥解密计算出关键词陷门

ASE单关键字检索： [BDOP 04]

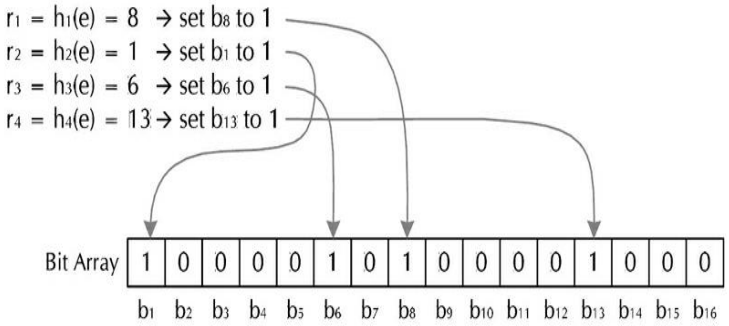
用公钥对数据内容M及关键字集合$\{W_1,\ldots,W_m\}$进行加密

用私钥生成某关键字W_j的检索陷门

服务器根据陷门与密文索引判断关键字集合中是否包含关键字W_j

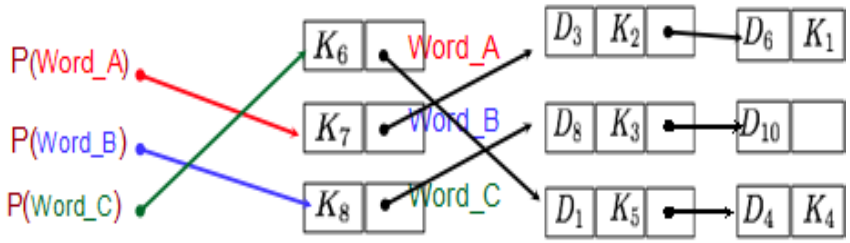
$$E_{A_{pub}}(M) \parallel \mathrm{PEKS}(A_{pub}, W_1) \parallel \cdots \parallel \mathrm{PEKS}(A_{pub}, W_m)$$

1. SSE密文索引构造



“关键词集合式”索引：[GOH 03]

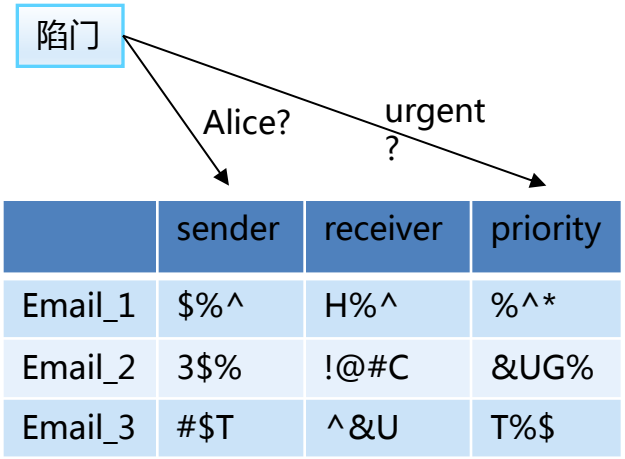
- 基于Bloom Filter 生成关键词码字;多关键词插入同一个BF中
- 随机插入一定的1补齐; 依次扫描所有文档的BF，看码字是否存在，
- 判断关键词是否在该文档中



“加密链表式”索引 [RJSR 06]

- 为每个关键词建立一个链表，链表内容加密，密钥保存在前一个节点中；链表的表头保存在数组中；数组内容加密，加密密钥由F（关键词）=K计算得出；通过随机置换函数P将关键词对应到数组下标；

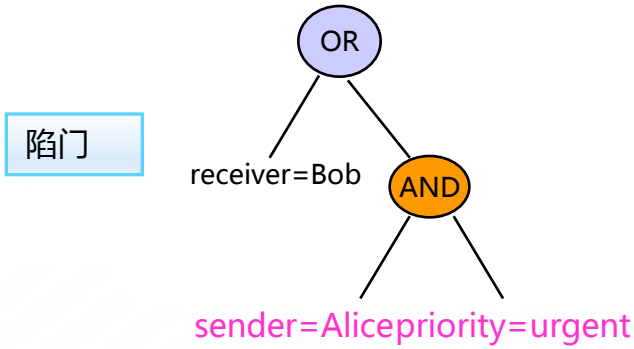
2. ASE密文索引构造



	sender	receiver	priority
Email_1	$%^	H%^	%^*
Email_2	3$%	!@#C	&UG%
Email_3	#$T	^&U	T%$

支持连接关键字检索：[HL 07]

每份数据具有相同的关键字域，利用公钥对关键字域中的关键字进行加密；利用私钥生成多个关键字的检索陷门，并指定要检索的关键字域；服务器返回密文关键字域中满足检索条件的数据



支持复杂逻辑结构的关键字检索

[LZDLC13]

利用接收者的公钥对关键字域中的关键字进行加密，利用私钥生成复杂逻辑结构的关键字检索陷门，逻辑词包括AND、OR等

3. Secure-Aware Query Processing

对于“安全相关”或者“隐私相关”的查询进行特殊处理，防止查询结果泄露隐私信息

针对用户查询模式的隐私保护，防止了解用户的真实意图，例如构造“ghost query”的方法[Pang ICDE'12]

针对数据关系运算结果的隐私保护，例如防止聚集信息泄露(Aggregate Suppression)的方法 [Zhang SIGMOD'12]

针对以视图表达的访问控制策略，Guarnieri 等给出了Boolean Query优化算法分析[VLDB'14]

结束语

- 如果说互联网时代人们的隐私受到了威胁，那么大数据时代无疑加深了这种威胁
  - 前者涉及特定的隐私信息；后者是对用户的全景洞察
- 需要从国家与社会层面限定互联网企业对用户隐私信息的收集与使用，从根源上解决问题
  - 用户隐私信息是互联网企业核心竞争力之一，不会主动放弃
- 从技术角度实现大数据隐私保护十分必要
  - 匿名保护
  - 访问控制

谢 谢!