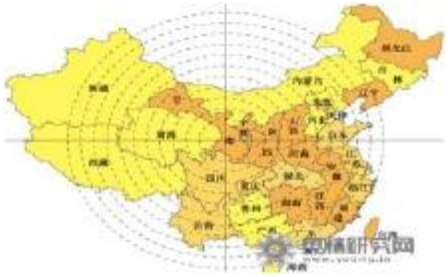
在线社会媒体计算的机遇与挑战

王　娜

深圳大学信息工程学院

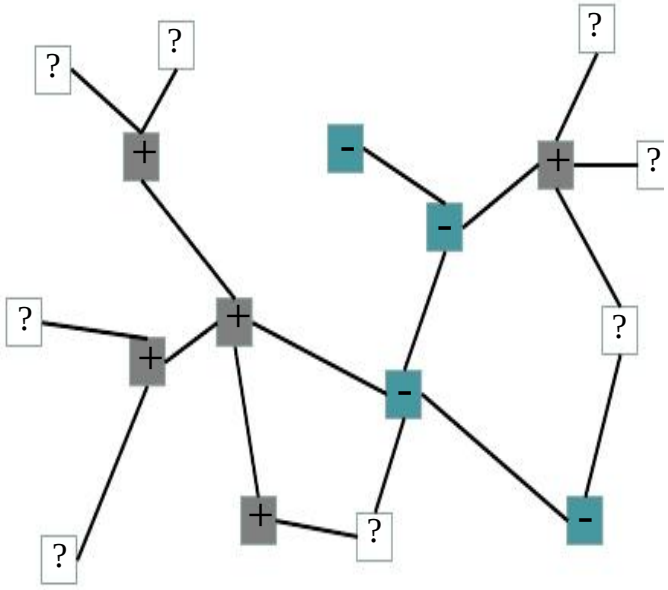
Social Networks

SN bridges our daily life and the **virtual** web space!

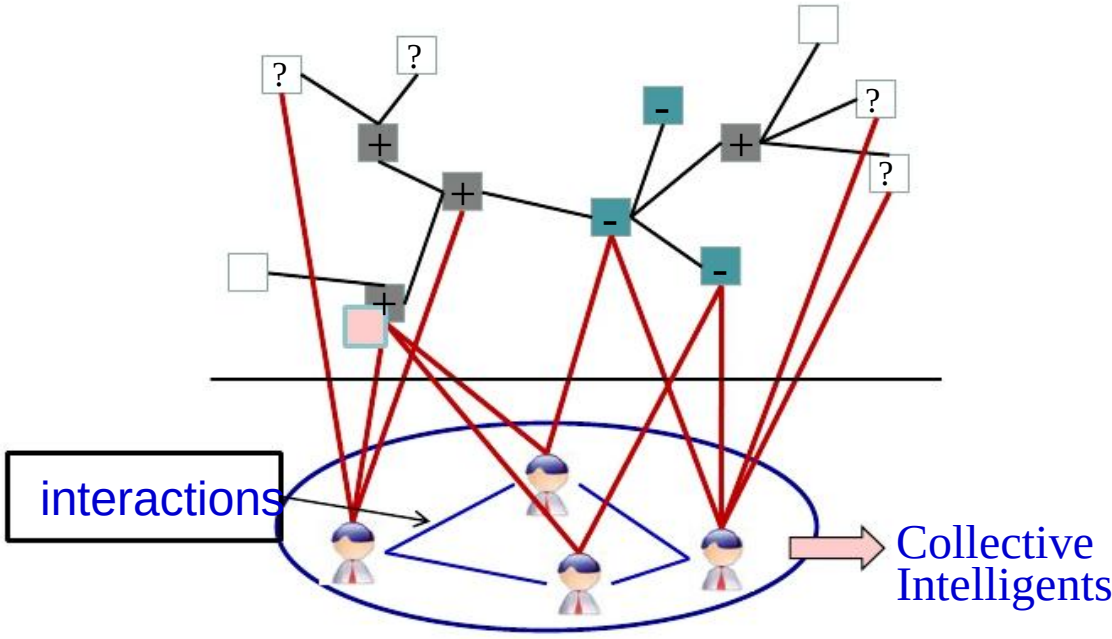


在线社会网络特征

- 多维网络（Multi-Dimensional Networks）
- 多模网络（Multi-Mode Networks）



Web 1.0



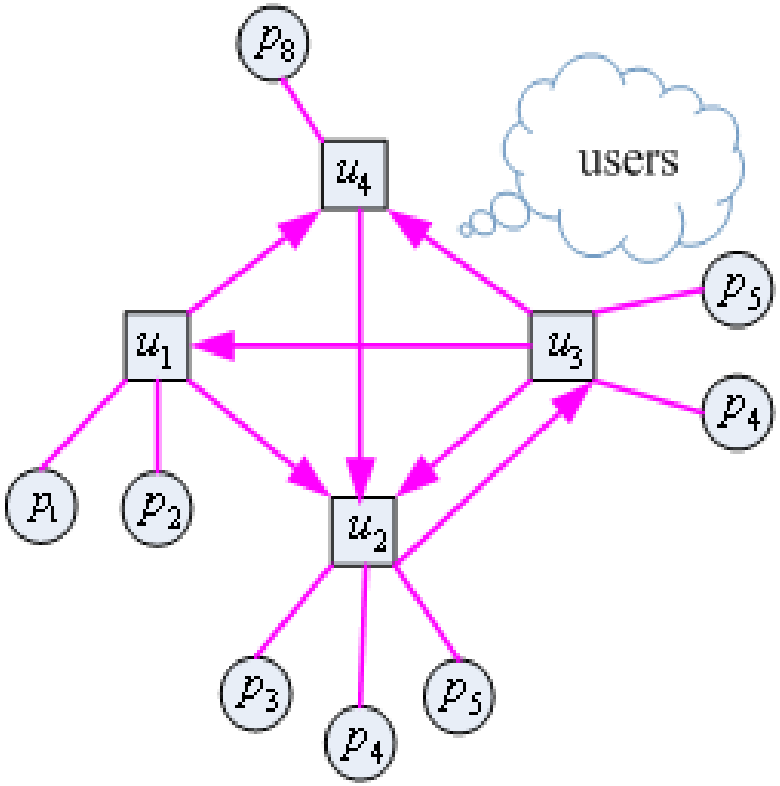
Social Web

在线社会网络大数据

- 大量社会媒体应用，产生了巨量数据
  - 250 million tweets per day
  - 3,000 photos in Flickr per minute
  - 153 million blogs posted per year

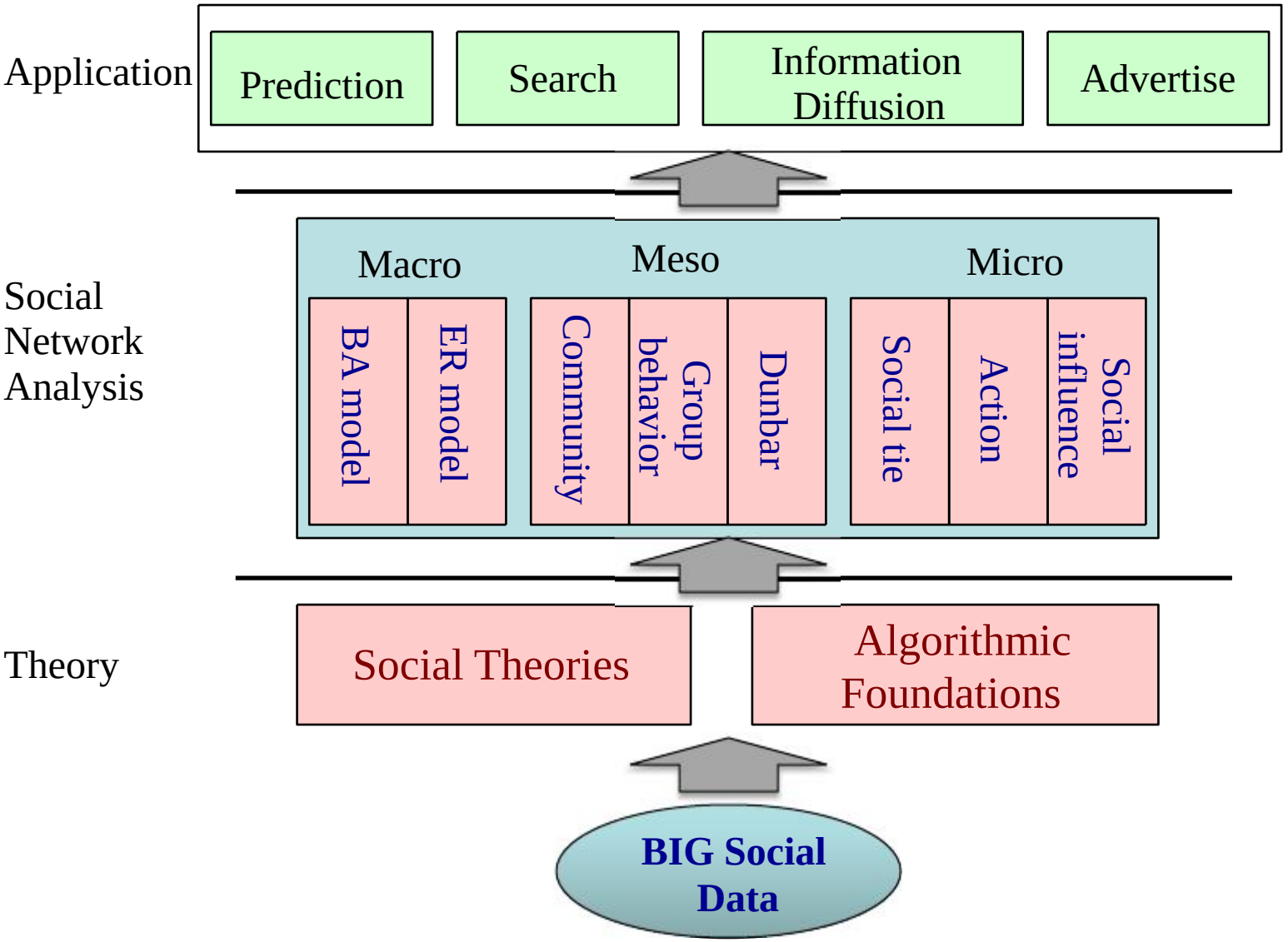
- 数据特点
  - 高维（high dimension）
  - 大规模（large-scale）
  - 链接的 （linked data）
  - 噪声（noisy）

在线社会媒体—微博数据

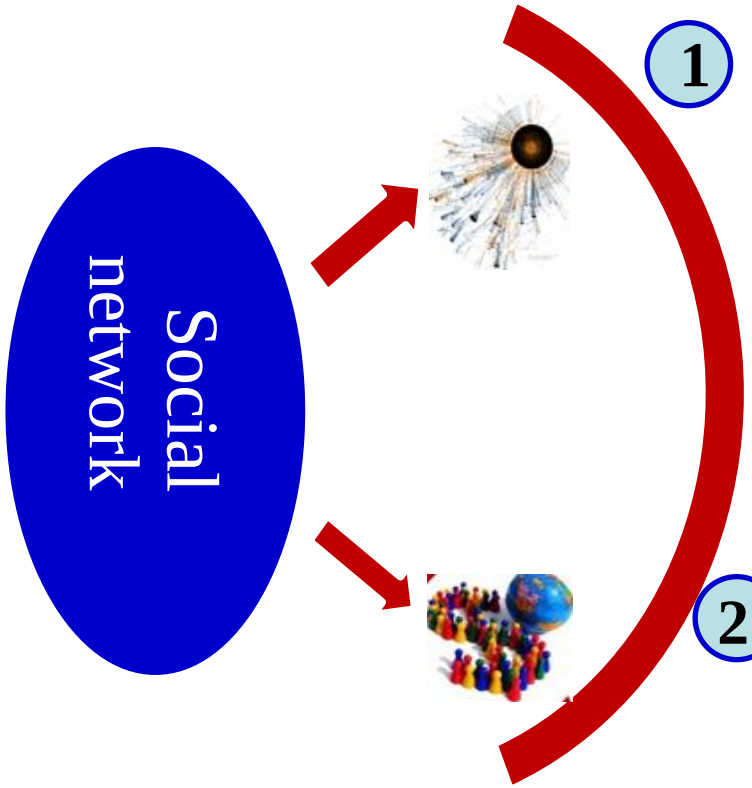


- 实体
  - 用户
  - 内容
- 关系
  - 用户—用户
  - 用户—内容
  - 内容—内容

Core Research in Social Network



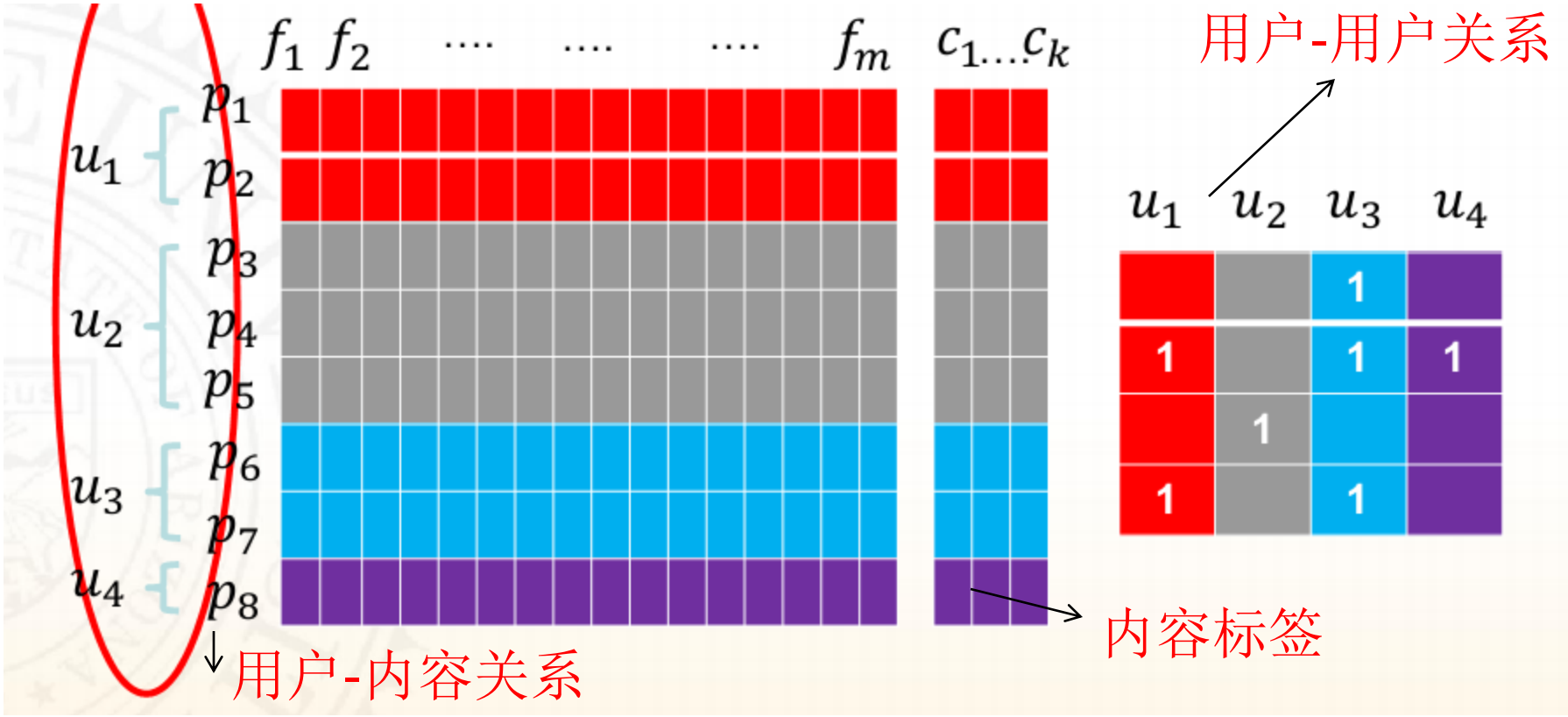
研究工作



博客分类
主题分类
话题识别

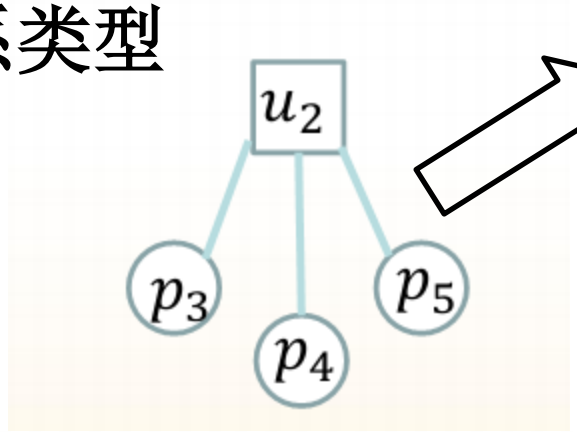
评分预测
信任预测

特征选择—链接数据



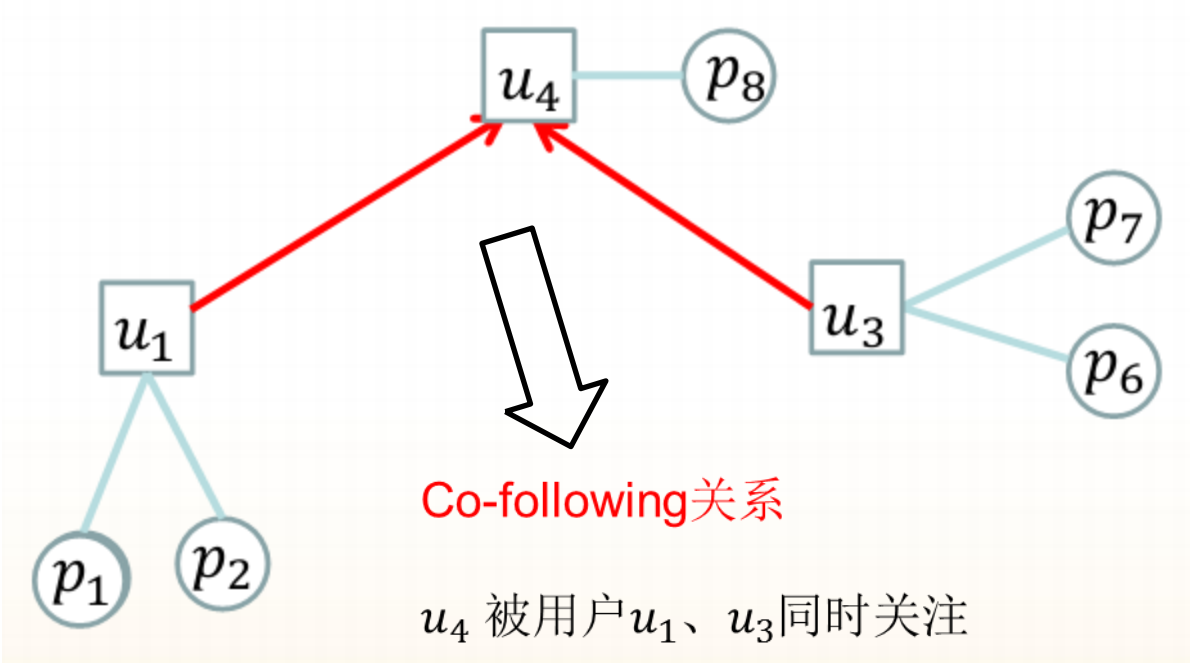
- 社会学原理(Social Theories)
  - 同质规则:认为两个有链接关系的对象倾向于具有相同或相似的类别
  - 合引规则:与同一个对象有链接关系的两个对象也倾向于具有相似的类别标签

问题1：链接关系类型



Co-post 关系

p_3、p_4、 p_5由同一用户u_2发布



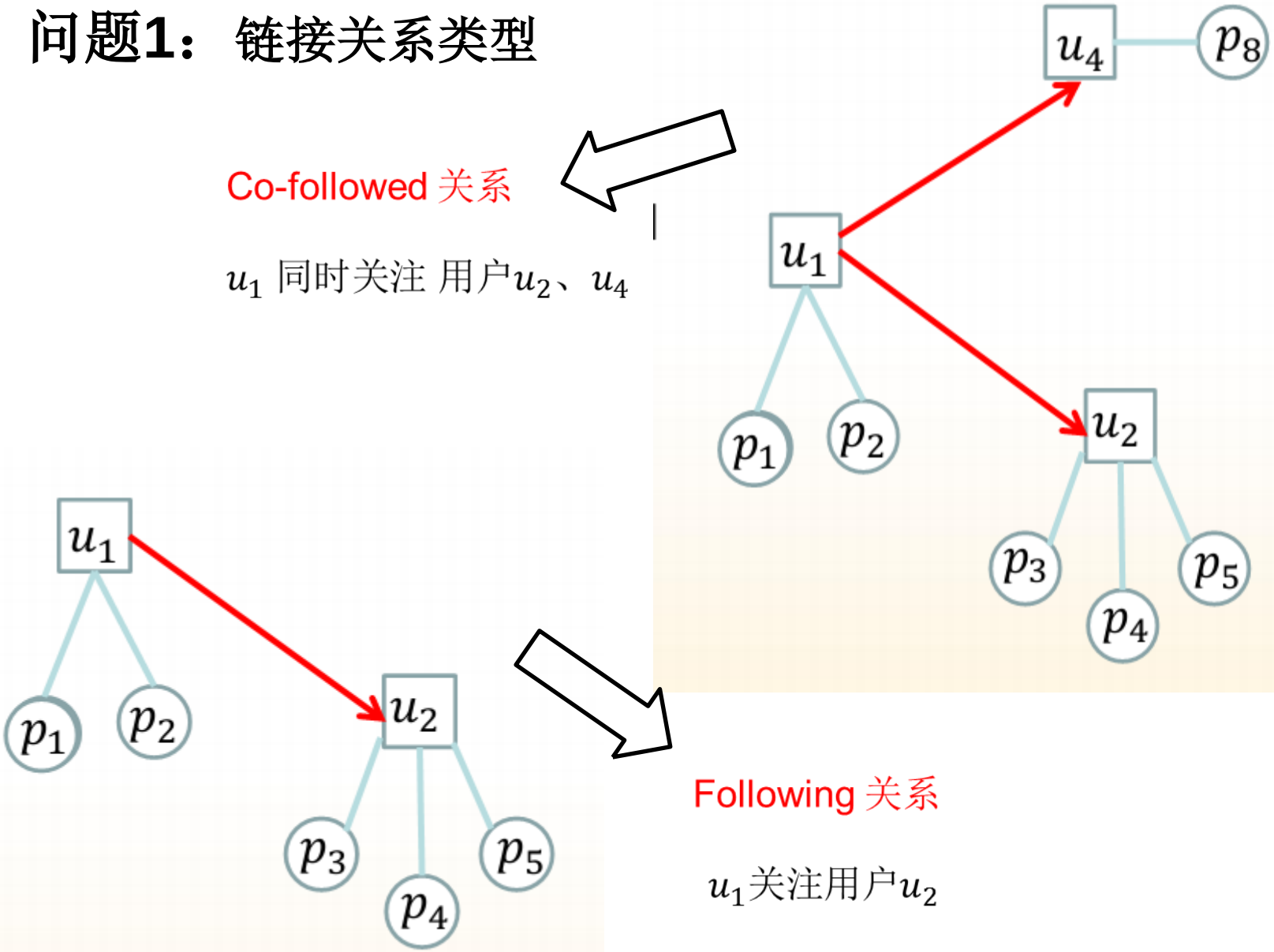
Co-following关系

u_4 被用户u_1、u_3同时关注

问题1：链接关系类型

Co-followed 关系

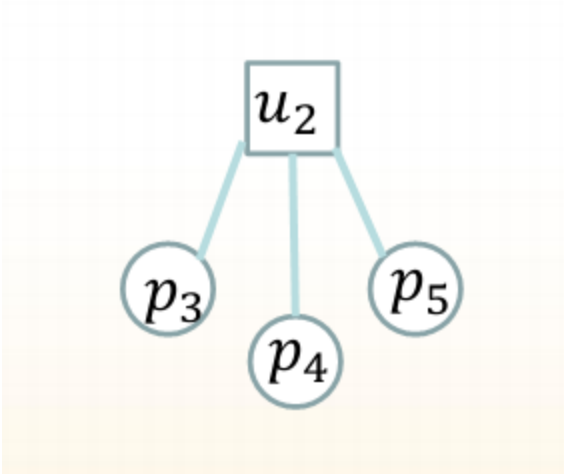
u_1 同时关注 用户u_2、u_4



Following 关系

u_1关注用户u_2

问题2：关系利用—Co-post



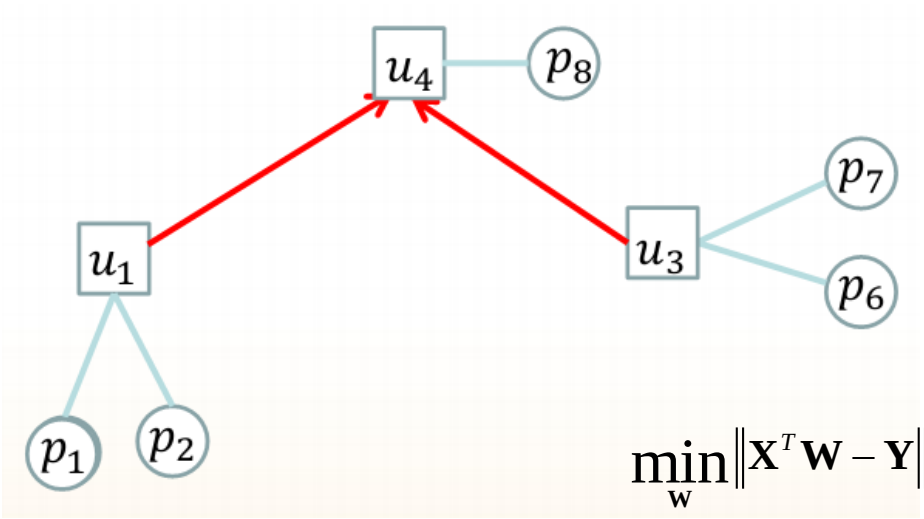
为用户内容标签预测的函数

$$\min_{\mathbf{W}} \left\| \mathbf{X}^T\mathbf{W} - \mathbf{Y} \right\|_F^2 + \alpha \left\| \mathbf{W} \right\|_{2,1} + \beta \sum_{u \in \mathbf{u}} \sum_{\mathbf{f}_i, \mathbf{f}_j \in \mathbf{F}_u} \left\| T(\mathbf{f}_i) - T(\mathbf{f}_j) \right\|_2^2$$

$$= \min_{\mathbf{W}} \left\| \mathbf{X}^T\mathbf{W} - \mathbf{Y} \right\|_F^2 + \alpha \left\| \mathbf{W} \right\|_{2,1} + \beta \sum_i \sum_j \mathbf{A}(\mathrm{i},\mathrm{j}) \left\| \mathbf{W}^T\mathbf{f}_i - \mathbf{W}^T\mathbf{f}_j \right\|_2^2$$

Co-post关系矩阵，如果内容p_i 、p_j 由同一个用户发布，那么$\mathbf{A}(i,j)=1$ ，否则为0.

问题2：关系利用—Co-following



$$\hat{T}(\mathrm{u}_k)=\frac{\sum_{\mathbf{f}_i\in\mathbf{F}_k}T(\mathrm{f}_i)}{|\mathbf{F}_k|}=\frac{\sum_{\mathbf{f}_i\in\mathbf{F}_k}\mathbf{W}^T\mathbf{f}_i}{|\mathbf{F}_k|}$$

定义用户主题兴趣：

$$\min_{\mathbf{W}}\left\|\mathbf{X}^T\mathbf{W}-\mathbf{Y}\right\|_F^2+\alpha\left\|\mathbf{W}\right\|_{2,1}+\beta\sum_{u_k\in\mathbf{u}}\sum_{\mathbf{u}_i,\mathbf{u}_j\in\mathbf{F}_k}\left\|\hat{T}(\mathrm{u}_i)-\hat{T}(\mathrm{u}_j)\right\|_2^2$$

$$=\min_{\mathbf{W}}\left\|\mathbf{X}^T\mathbf{W}-\mathbf{Y}\right\|_F^2+\alpha\left\|\mathbf{W}\right\|_{2,1}+\beta\sum_{ij}\mathbf{FI}(\mathrm{i},\mathrm{j})\left\|\mathbf{W}^T\mathbf{FH}(:,\mathrm{i})-\mathbf{W}^T\mathbf{FH}(:,\mathrm{j})\right\|_2^2$$

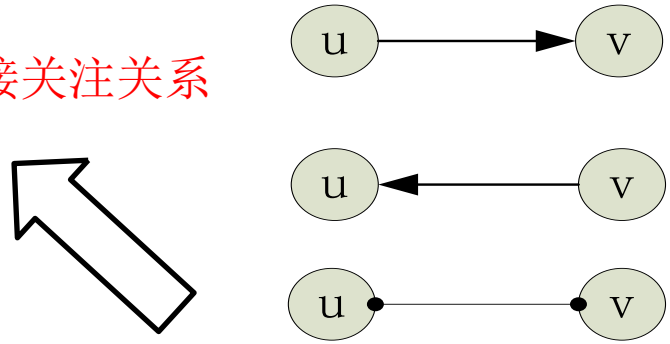
Co-following关系矩阵，如果用户 u_i 与 u_j

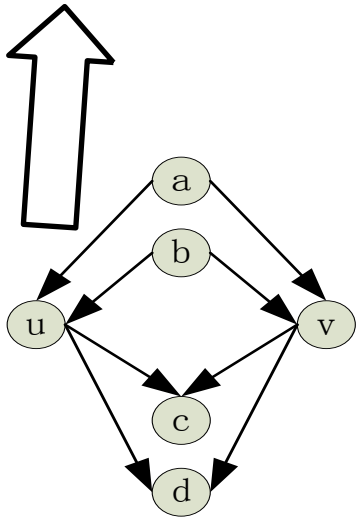
同时关注至少一个用户． $\mathbf{FI}(i,j)=1$

问题3：关系强度—局部相似性

用户间局部链接关系

用户间直接关注关系





(a) 传统的聚类假设　　(b) 链接结构假设

目标函数

用户局部关系相似指标：　(CN指标等)

$$score(u,v)=\left|\tau(u)\cap\tau(v)\right|$$

$$score(u,v)=\left|\tau(u)\cap\tau(v)\right|/\left|\tau(u)\cup\tau(v)\right|$$

$$score(x,y)=\frac{\left|\tau(x)\cap\tau(y)\right|}{\sqrt{k_x k_y}}$$

用户主题兴趣

$$\min_{\mathbf{W}}\left\|\mathbf{X}^T\mathbf{W}-\mathbf{Y}\right\|_F^2+\alpha\left\|\mathbf{W}\right\|_{2,1}+\beta\sum_{i,j}\mathbf{S}(i,j)\left\|\hat{T}(v_i)-\hat{T}(v_j)\right\|_2^2$$

用户关系强度矩阵

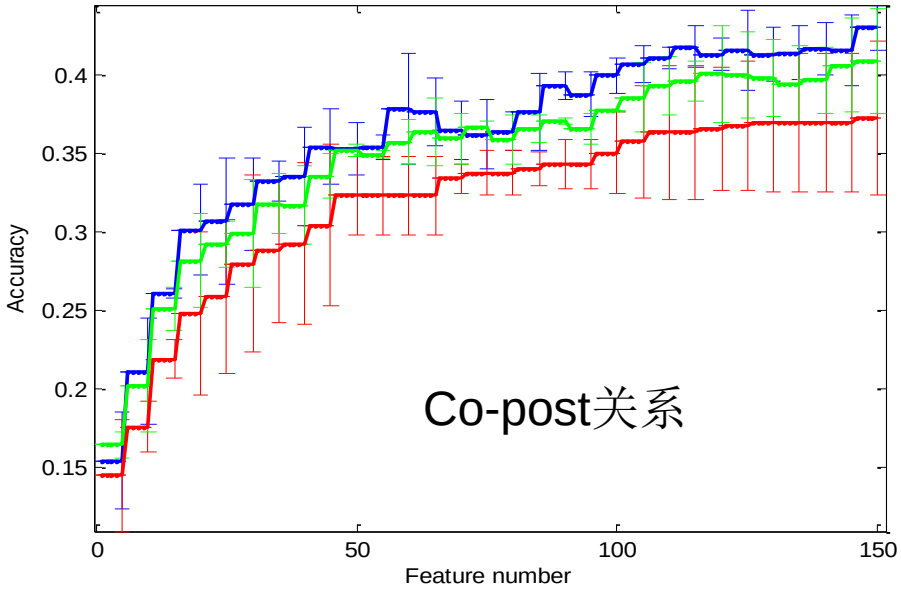
数据平台

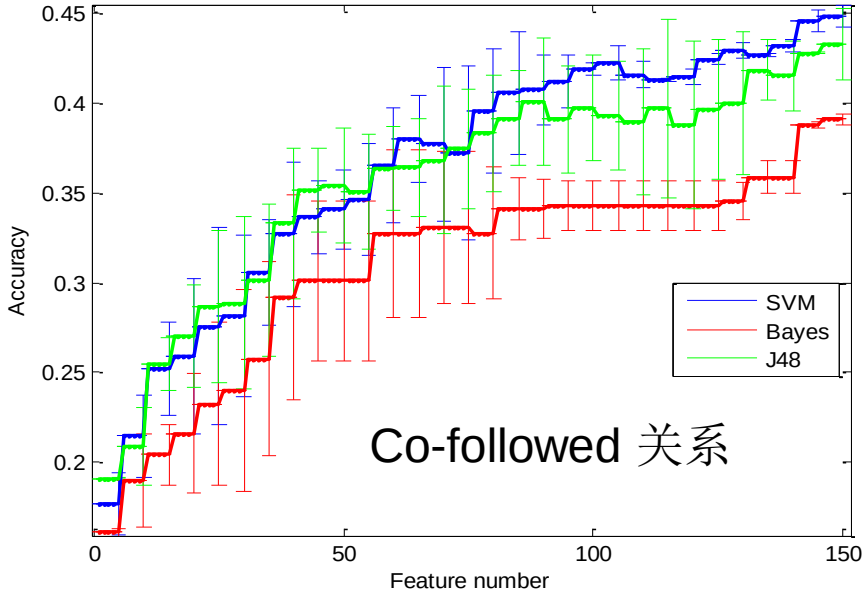
BlogCatalog数据集是国外真实的一个博客平台，用户可以在预定义好的类别标签下记录并发表他们的博文。这些博文内容都经过了停用词的移除等相关预处理，去除了那些极不相关的特征词得到博文内容的属性特征。

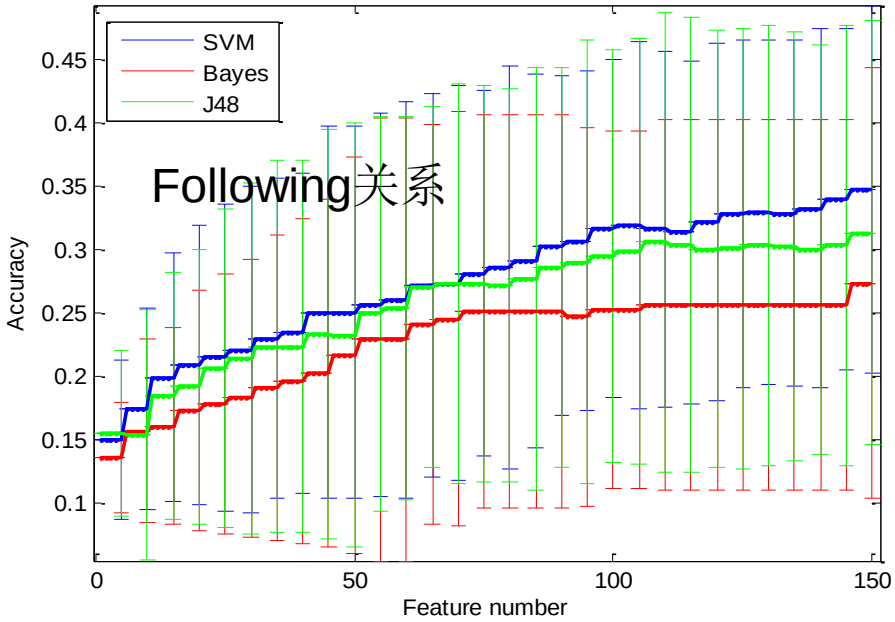
Table: Statistics of the BlogCatalog Dataset

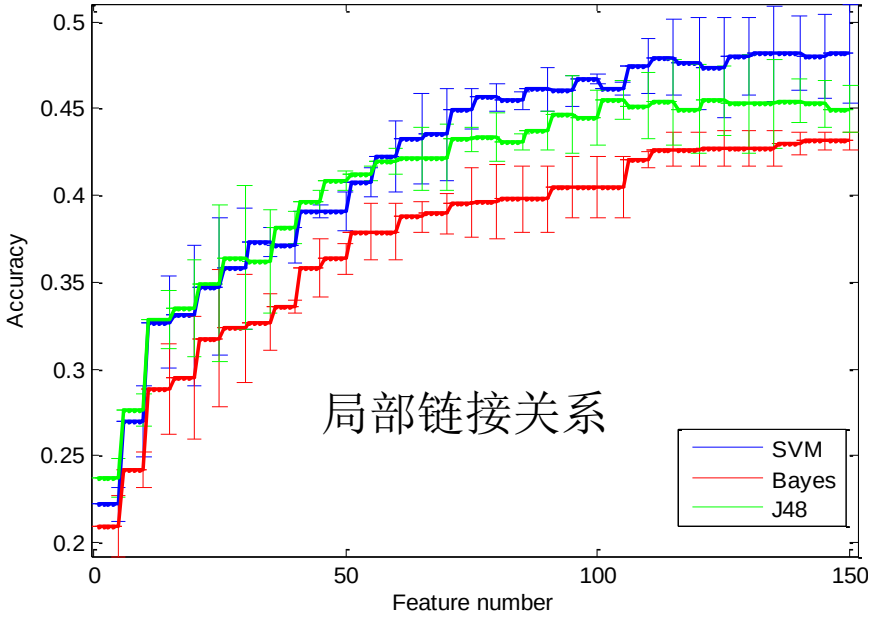
# of Users	2242
# of Posts	3000
# Features after TFIDF	6000
# Following Relations	55356
#of Classed	15
Max#Followers	80
Min#Followers	1

结论—BlogCatalog

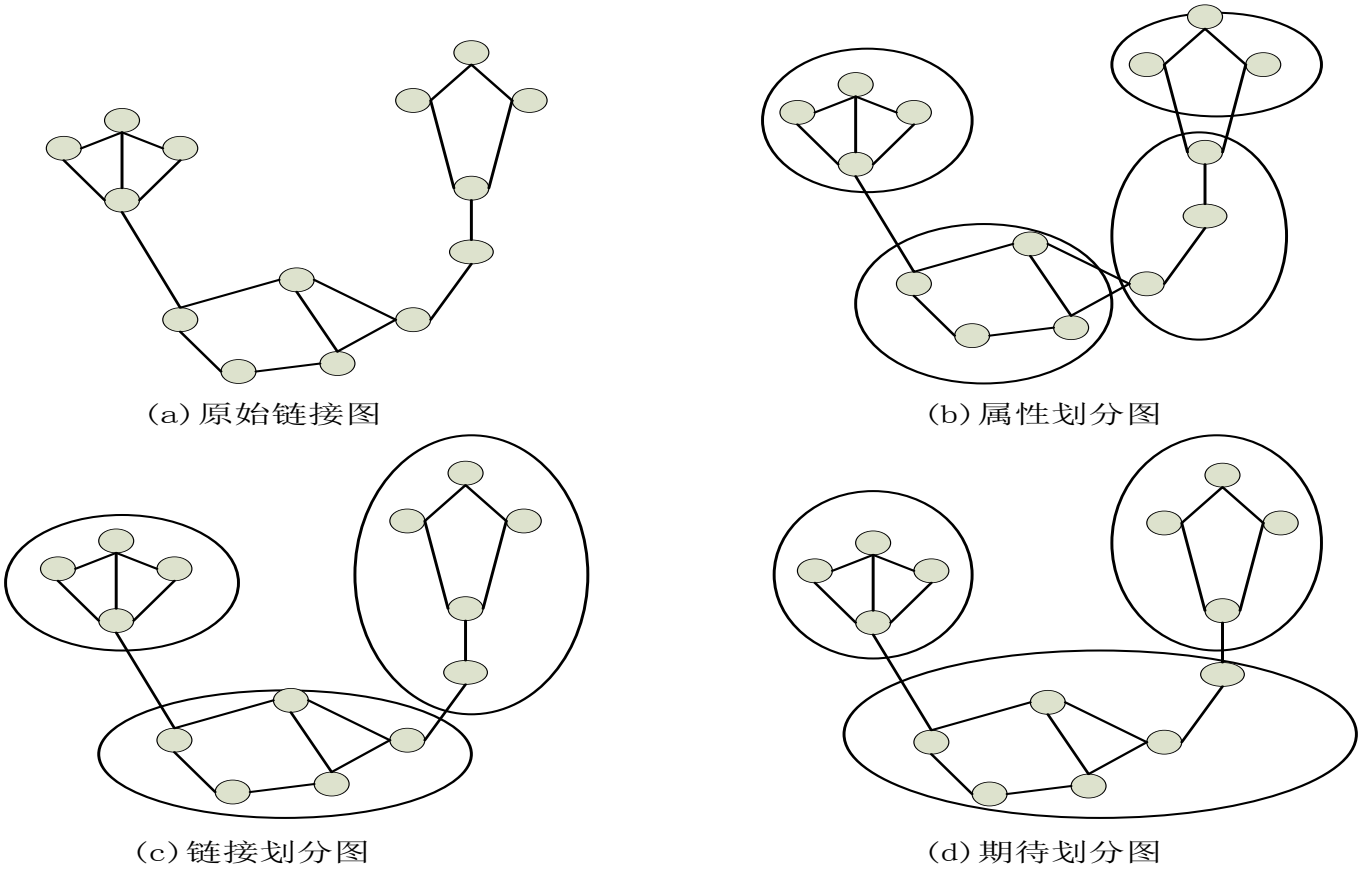






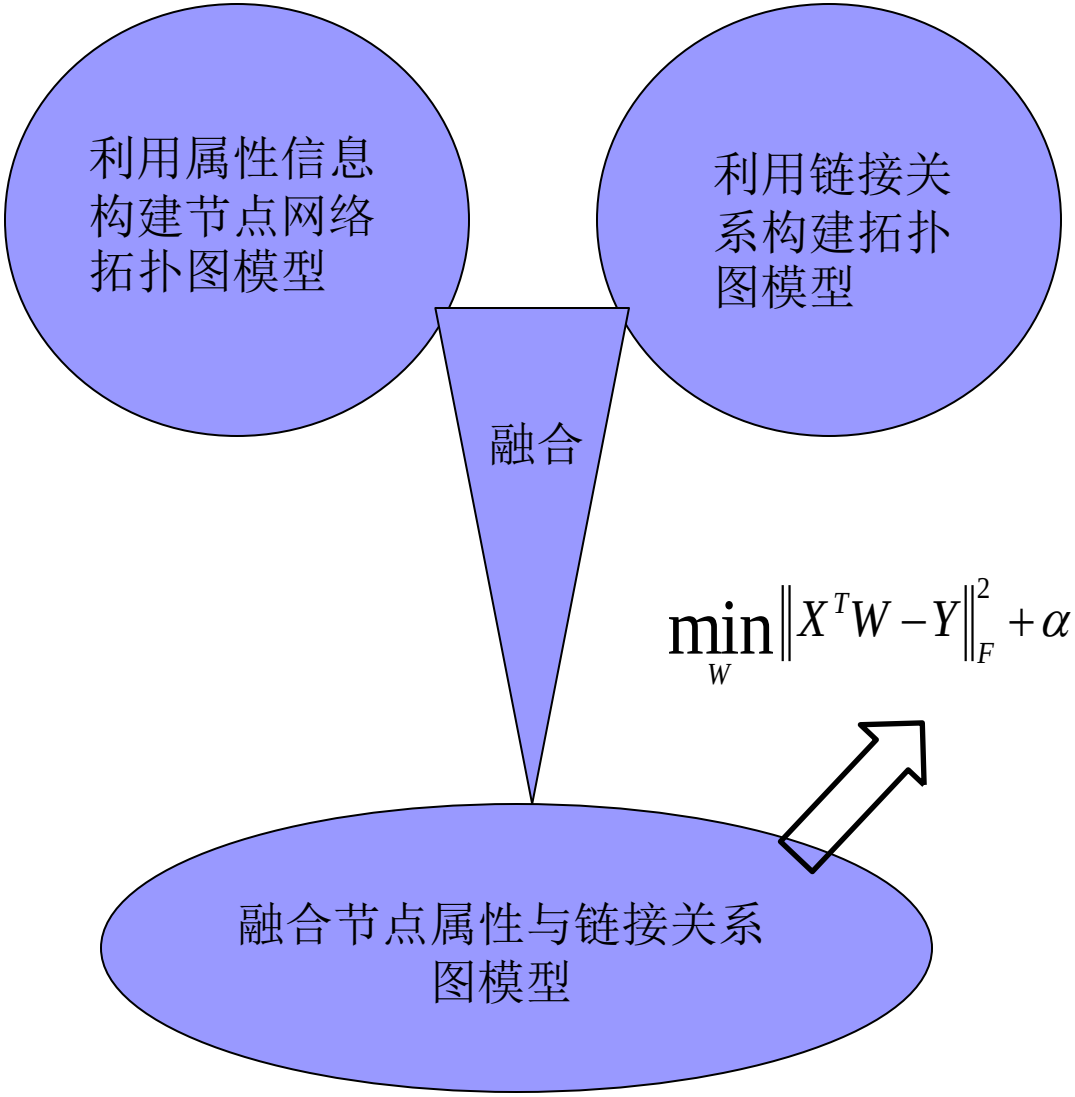


问题4：异质节点属性和链接关系利用



社区划分示意图

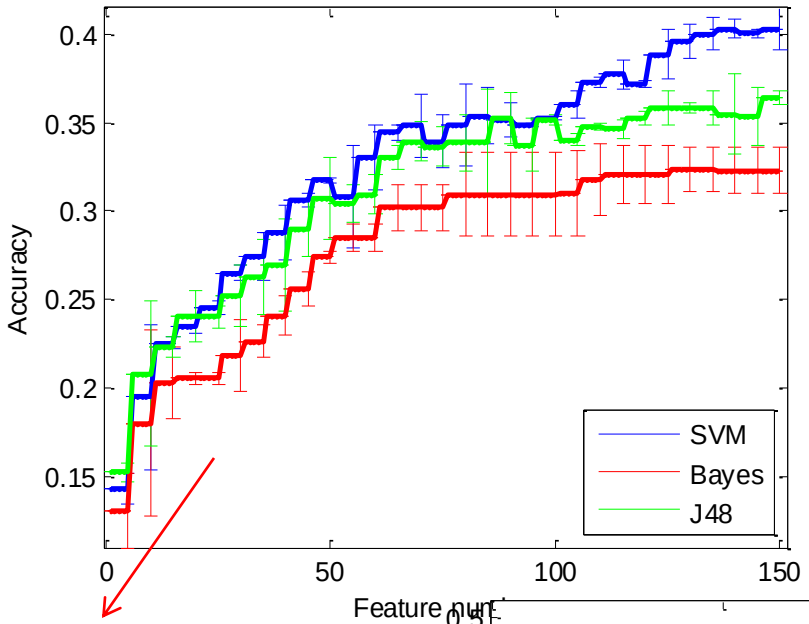
问题4：异质节点属性和链接关系利用



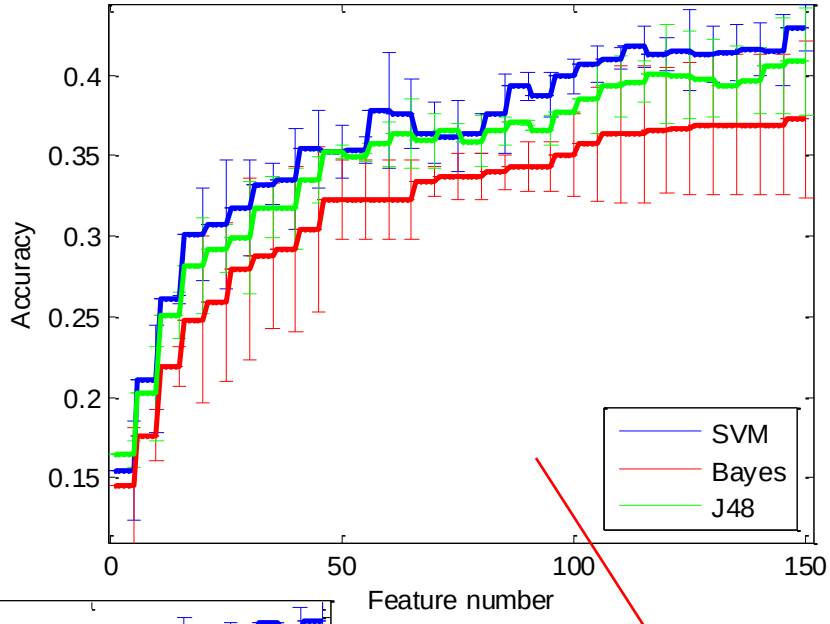
用户内容-内容相似关系强度

$$\min_{W} \left\| X^T W - Y \right\|_F^2 + \alpha \left\| W \right\|_{2,1} + \beta \sum_{i,j=1}^{N} \mathbf{M}(i,j) \left\| W^T \mathbf{x}_i - W^T \mathbf{x}_j \right\|_2^2$$

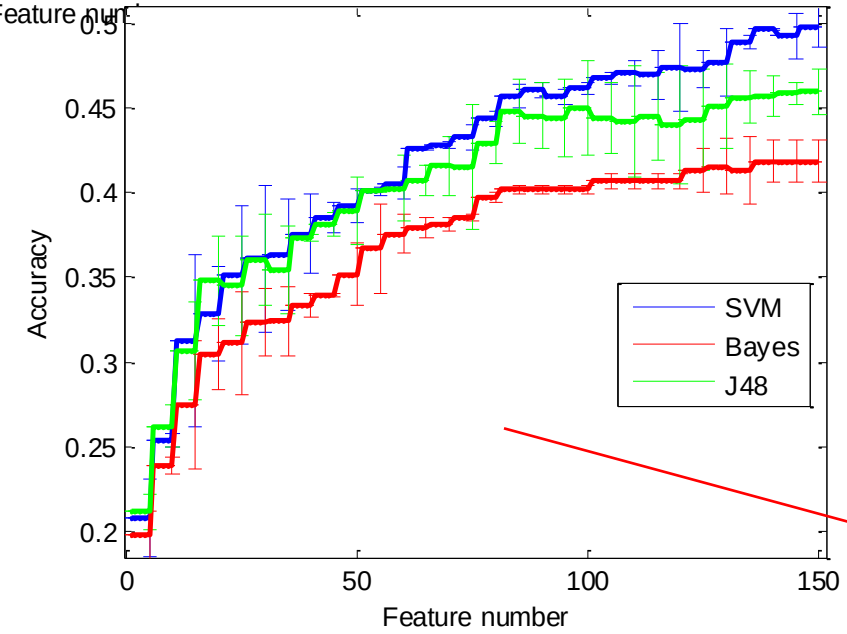
结论—BlogCatalog



只利用属性信息

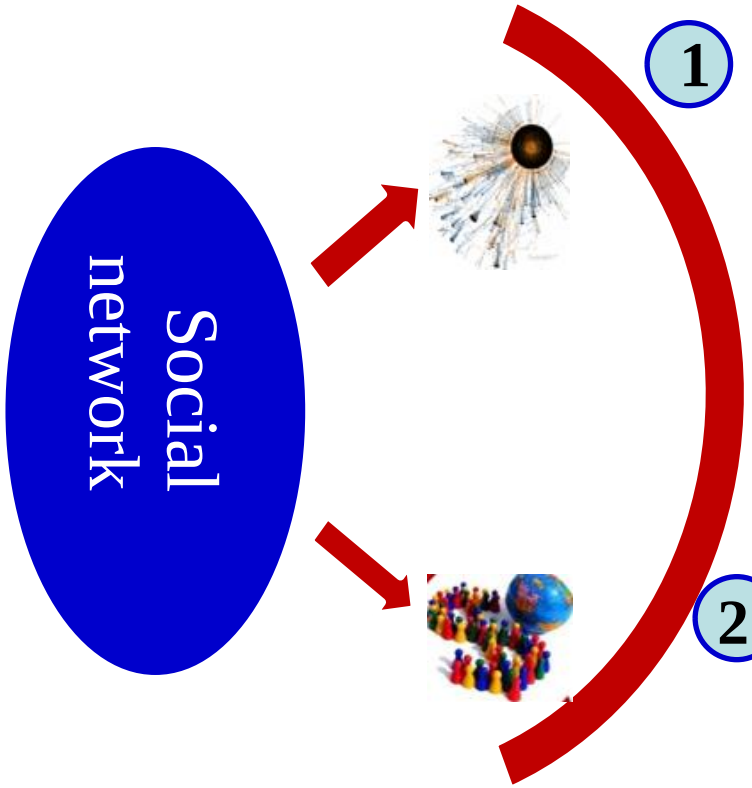


只利用链接关系



同时利用节点属性和链接关系

研究工作



博客分类
主题分类
话题识别

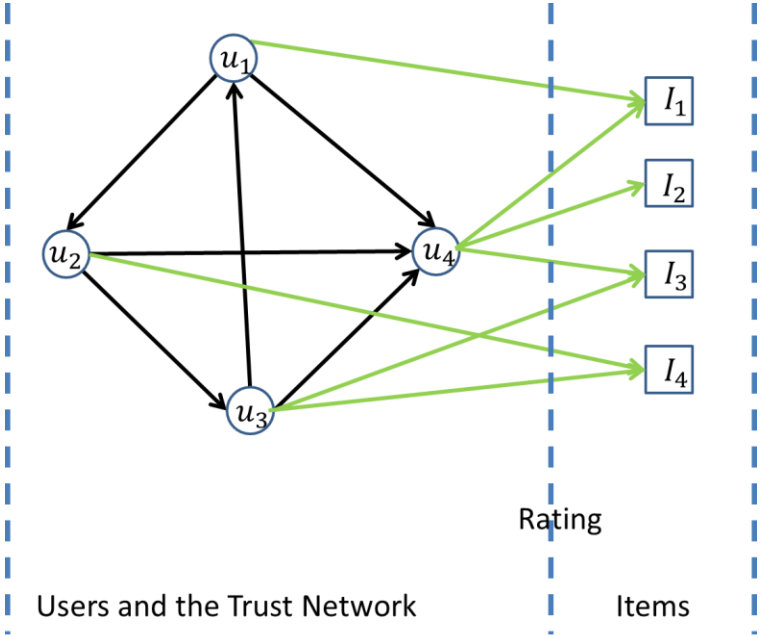
评分预测
信任预测

信任计算

- 匿名性:允许用户匿名注册，一个用户可以拥有几个ID号。
- 开放性：任何人在任何时间、任何地点都可以自由地加入或离开。
- 用户生成内容：数据中包含了大量反映网民兴趣爱好、观点和价值观的信息。

- 由于互联网的开放性和匿名性,如何给在线社会网络中的用户提供合理的信任计算机制成为一个急需解决的问题。
- 作为一个新的商业渠道,面向实际应用,如何利用大量的用户生成内容对在线社会网络进行挖掘分析从而取得更好的经济效益和社会效益也成为一个备受关注的焦点问题。
- 这两方面直接关系到在线社会网络的安全性和实用性,因此对其研究既具有理论价值又具有实际意义

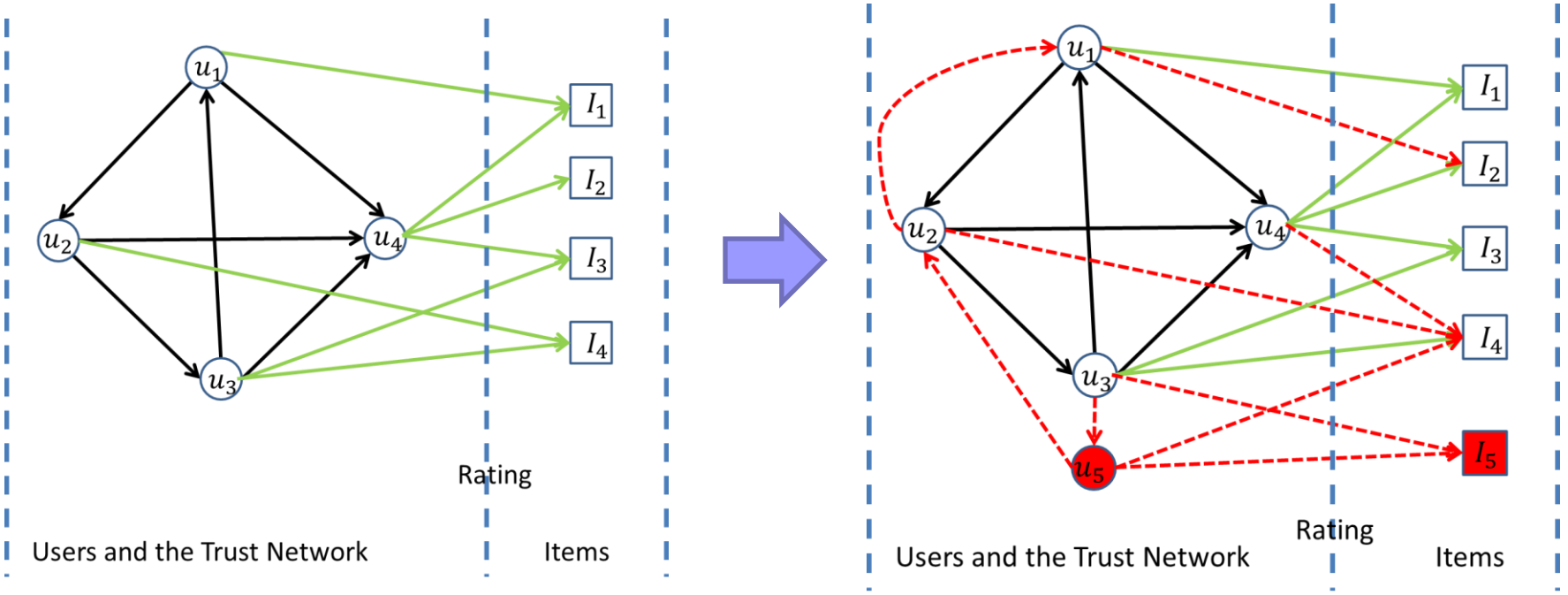
信任计算_动态演化



time t

Epinions在线评分网络

信任计算_动态演化

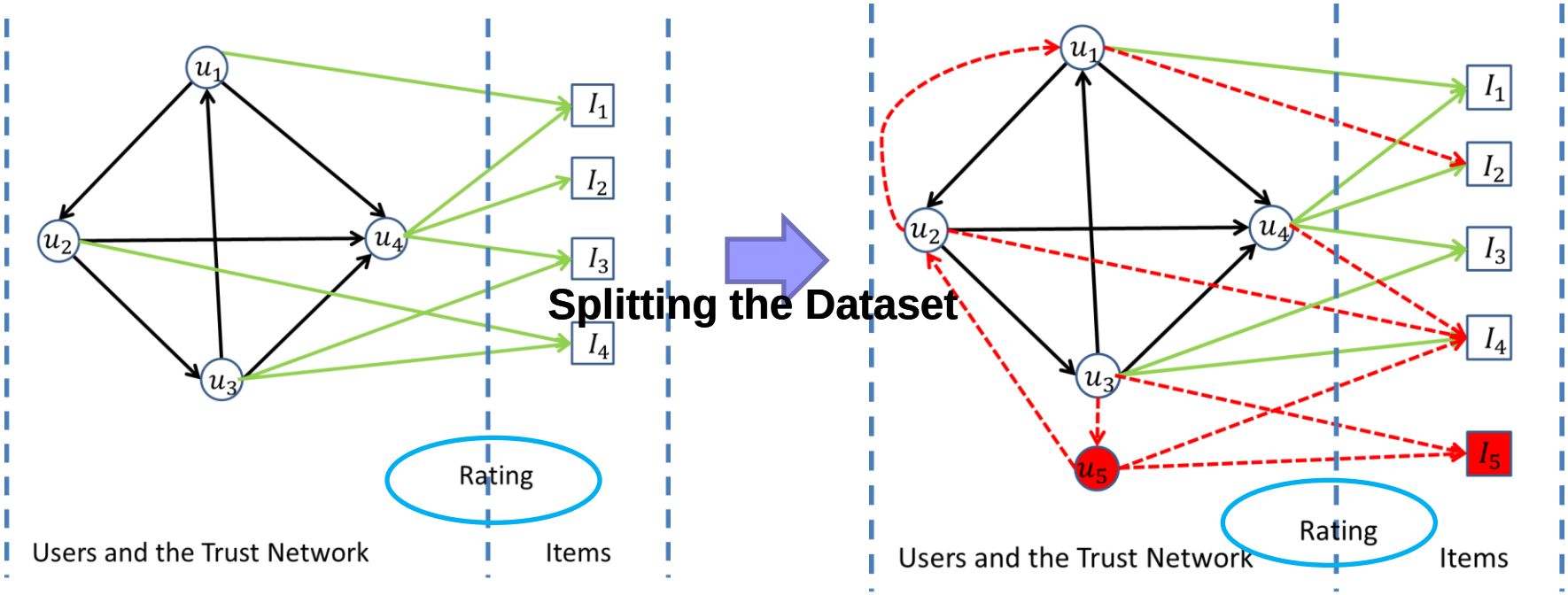


time t

time t+1

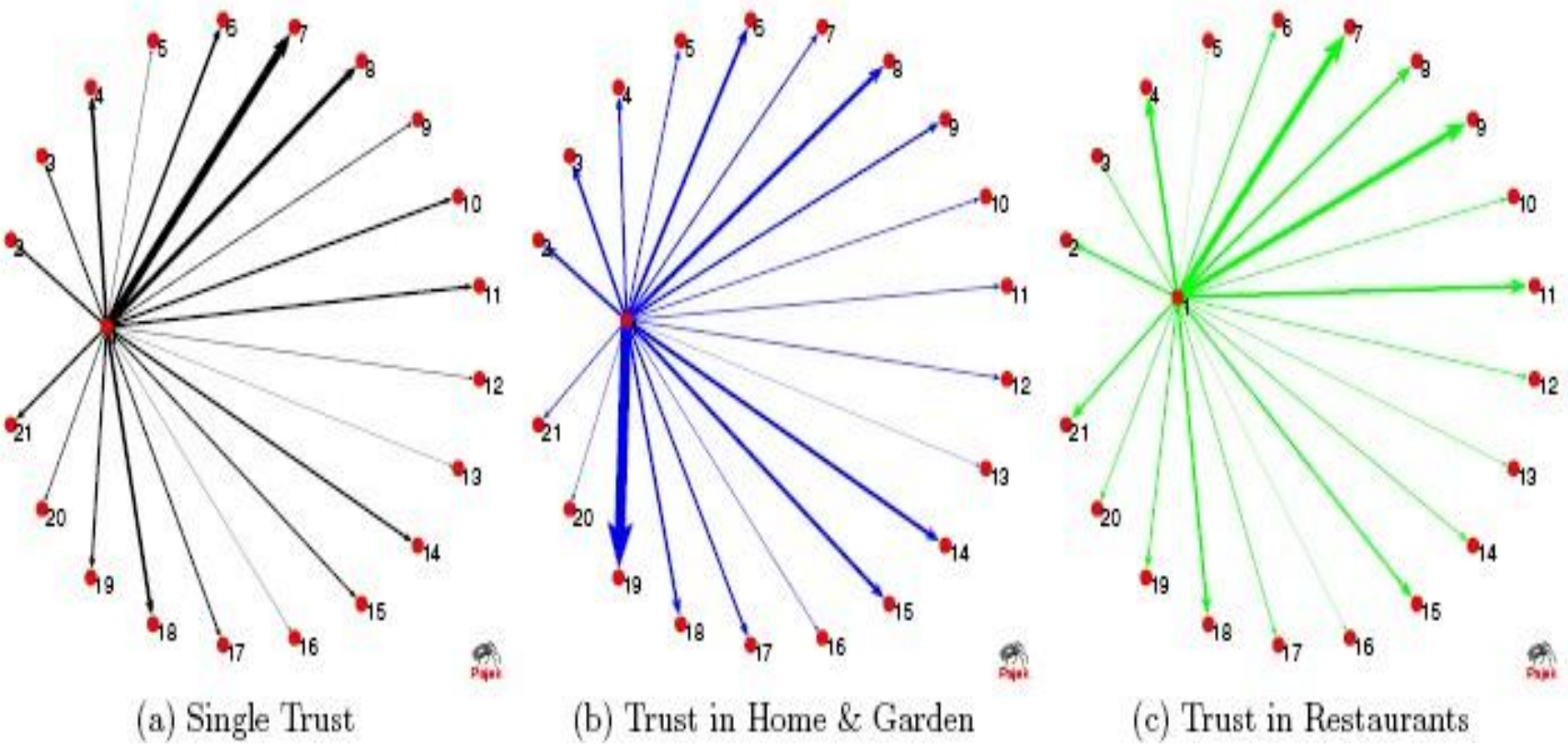
Epinions在线评分网络

信任计算_动态演化



Epinions在线评分网络

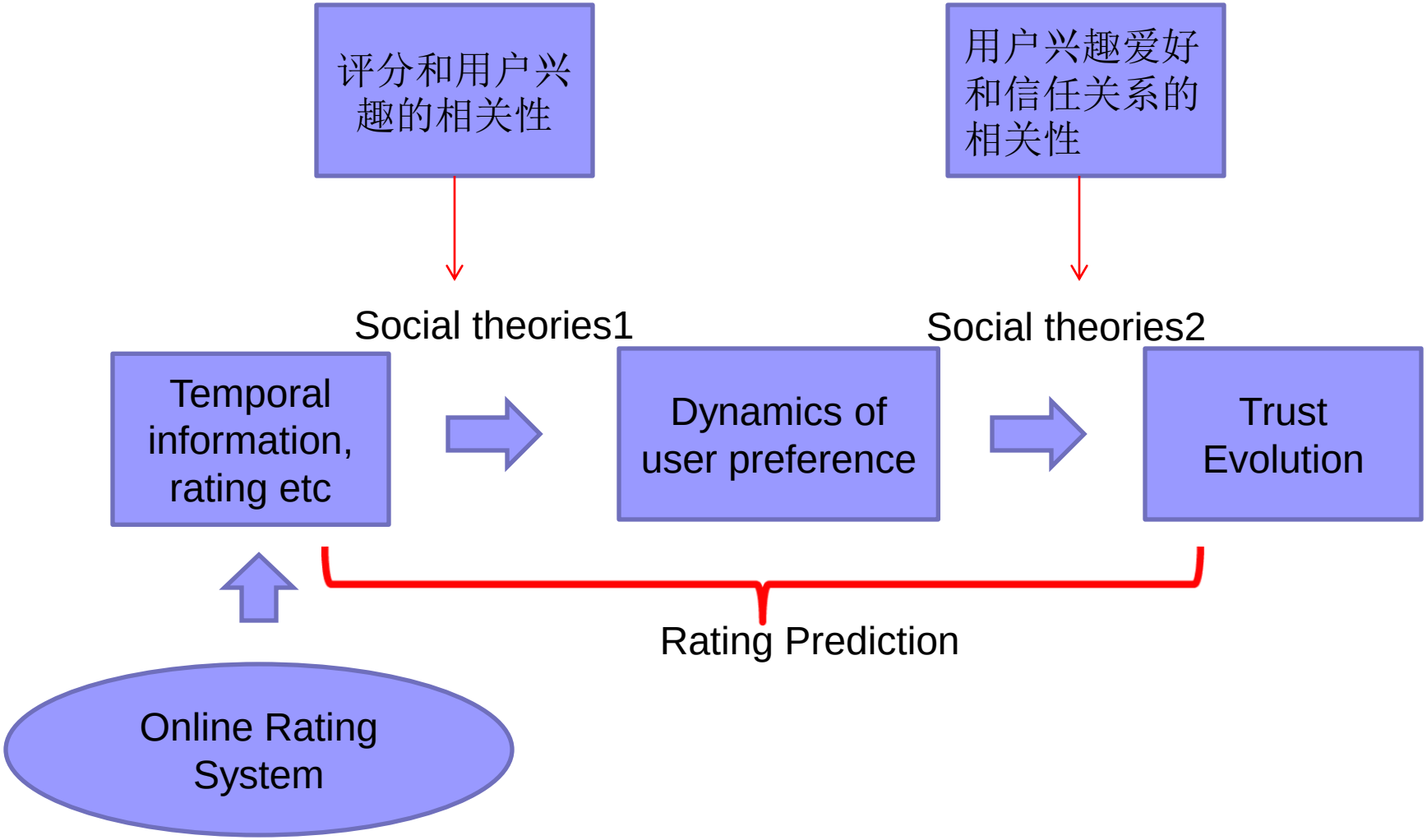
信任计算_多主题



(a) Single Trust　(b) Trust in Home & Garden　(c) Trust in Restaurants

Epinions在线评分网络

信任计算思想


评分和用户兴趣的相关性
用户兴趣爱好和信任关系的相关性
Social theories1
Social theories2
Temporal information, rating etc
Dynamics of user preference
Trust Evolution
Rating Prediction
Online Rating System

信任计算模型

$$\min_{\mathbf{p}_i^t \geq 0,\ \mathbf{q}_j \geq 0,\ \eta_i \geq 0} \sum_{t=1}^{T} \sum_{(i,j)\in\mathcal{O}_t} \Big(r_{ij}^t - \alpha \sum_{k=1}^{K} \mathbf{q}_j(k)\mathbf{p}_i^t(k) -$$

$$(1-\alpha)\frac{\sum_{v\in N_i^t}\sum_{k=1}^{K} f(\mathbf{w}^\top \mathbf{s}_{ivk}^t + b_i)\mathbf{q}_j(k) r_{vj}}{\sum_{v\in N_i^t}\sum_{k=1}^{K} f(\mathbf{w}^\top \mathbf{s}_{ivk}^t + b_i)\mathbf{q}_j(k)}\Big)^2$$

$$+ \beta\Big(\sum_{u_i\in\mathcal{U}_T}\sum_{t=t_{u_i}}^{T} \|\mathbf{p}_i^t\|_2^2 + \sum_{I_j\in\mathcal{I}_T} \|\mathbf{q}_j\|_2^2$$

$$+ \|\mathbf{w}\|_2^2 + \sum_{u_i\in\mathcal{U}_T} \|b_i\|_2^2 + \sum_{u_i\in\mathcal{U}_T} \|\eta_i\|_2^2 \Big)$$

$$+ \lambda \sum_{u_i\in\mathcal{U}_T}\sum_{t=t_{u_i}+1}^{T}\sum_{k=1}^{K} c\Big(\mathbf{p}_i^t(k) - \mathbf{p}_i^{t-1}(k)\Big)$$

Part 1

Part 2

Part 3

Part 4

Part 1： Modeling Rating via User Preference

- 用户兴趣爱好和商品特征相关性

$$\hat{r}_{ij}^{t} = \mathbf{q}_j^{\top} \mathbf{p}_i^{t} = \sum_{k=1}^{K} \mathbf{q}_j(k) \mathbf{p}_i^{t}(k)$$

- p_i^t表示用户i在时间t时的兴趣爱好，q_j 表示商品j的特征，k为商品分类

Part 2: Modeling Rating via Trust Network

- 利用信任网络计算商品评分

$$\hat{r}_{ij}^{t} = \frac{\sum_{v \in N_i^t} \sum_{k=1}^{K} w_{ivk}^{t} \mathbf{q}_j(k) r_{vj}}{\sum_{v \in N_i^t} \sum_{k=1}^{K} w_{ivk}^{t} \mathbf{q}_j(k)}$$

表示用户i和用户v在主题分类k下的信任值大小

早期评分的影响力会随着时间而衰减

$$r_{vj} = e^{-\eta_i (t - t_{vj})} r_{vj}^{t_{vj}}$$

Part 3: Modeling Trust and User preference

- 建模信任网络和用户兴趣爱好的相关性

$$w_{ivk}^{t} = f(\mathbf{w}^{\top}\mathbf{s}_{ivk}^{t} + b_i),$$

s_{ivk}^{t} 是用户间兴趣爱好相似矩阵，b_i 是用户偏差。

Part 4: Modeling Change of User Preference

- 建模用户兴趣爱好的变化

$$\lambda \sum_{u_i \in \mathcal{U}_T} \sum_{t=t_{u_i}+1}^{T} \sum_{k=1}^{K} c\left(\mathbf{p}_i^t(k) - \mathbf{p}_i^{t-1}(k)\right)$$

c 函数描述用户兴趣爱好是如何变化, λ控制变化速度

实验数据

- Epinions
  - 商品评价网站

# of Users	22,166
# of Items	296,277
# of Categories	27
# of Ratings	922,267
# of Links	355,813
First Rating on	Jul 05 1999
Last Rating on	May 08 2011
Trust Network Density	0.0014
Clustering Coefficient	0.1518

- Epinions分为11个时间戳

eTrust提高信任相关应用的性能

- 评分预测 （Rating Prediction ）

	$\mathcal{K}$	$\mathcal{N}$	$\mathcal{K}+\mathcal{N}$
Mean	1.1054	1.1562	1.1106
NN	1.1092	1.1566	1.1148
MF	1.0804	1.1472	1.0872
MF+NN	1.0675	1.1392	1.0747
mTrust	1.0566	1.1375	1.0646
eTrust	1.0299	1.0783	1.0347

*K*原有的无变化评分系统，
N有新产品或者新用户加入的评分演化系统(10.06%)

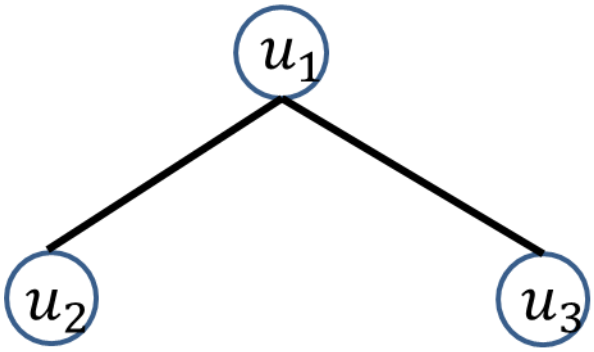
- 信任预测（Trust Prediction）

	$\mathcal{E}(\%)$	$\mathcal{N}(\%)$	$\mathcal{E}+\mathcal{N}(\%)$
Simi	48.41	28.94	43.93
TP	45.47	N.A.	35.01
Simi+TP	50.19	28.94	45.31
eTrust	55.07	33.83	50.18

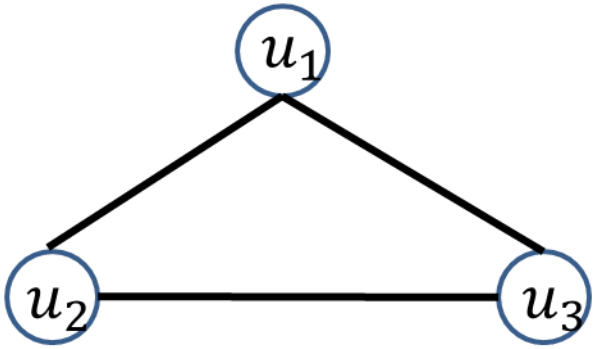
*E*现有用户的信任关系建立，
N新加入新用户后的信任关系预测(23.51%)

结论

- 开放三角（ an open triad）的演化速度是闭合三角（a closed triad）的 6.12 倍

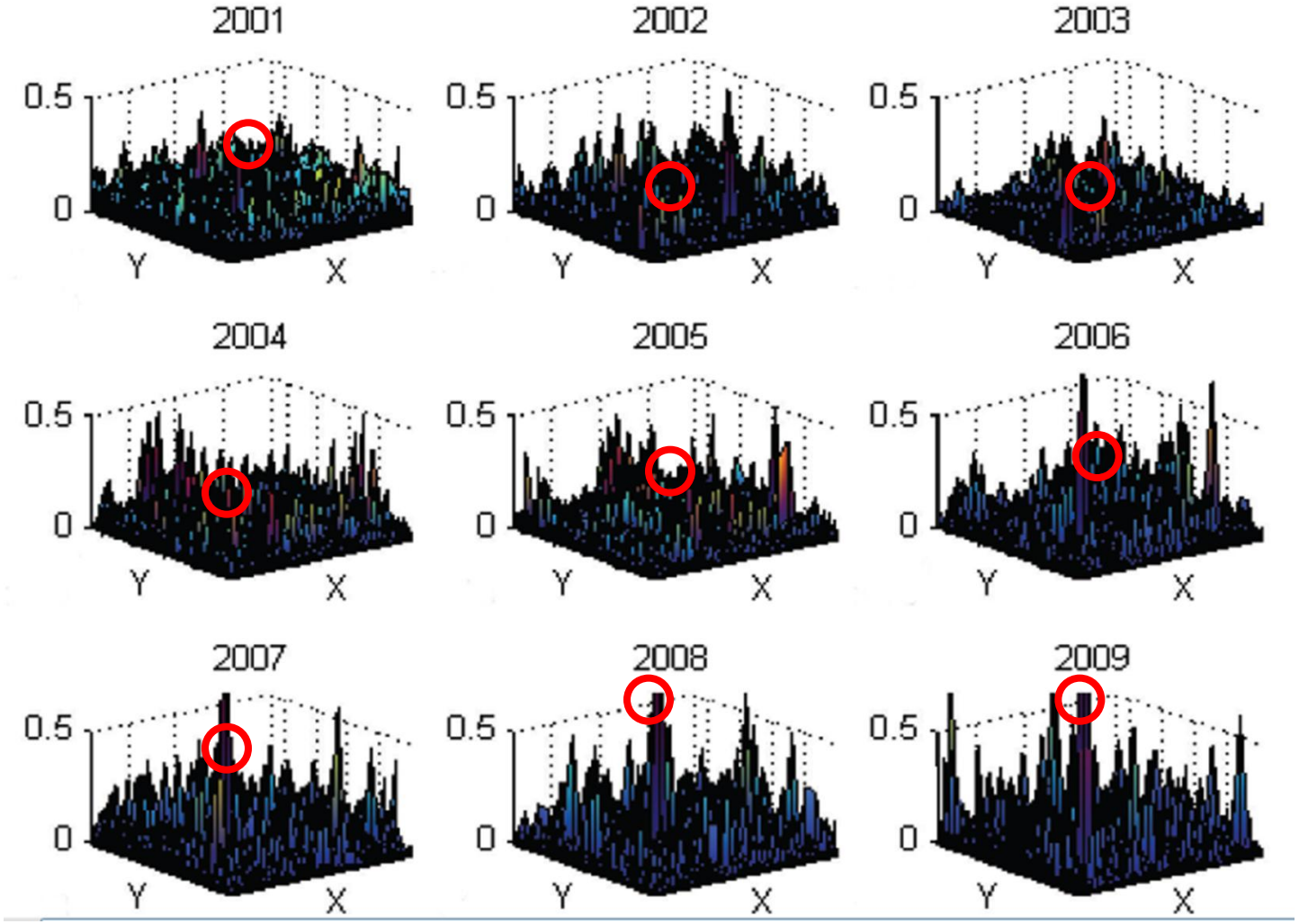


Open Triad

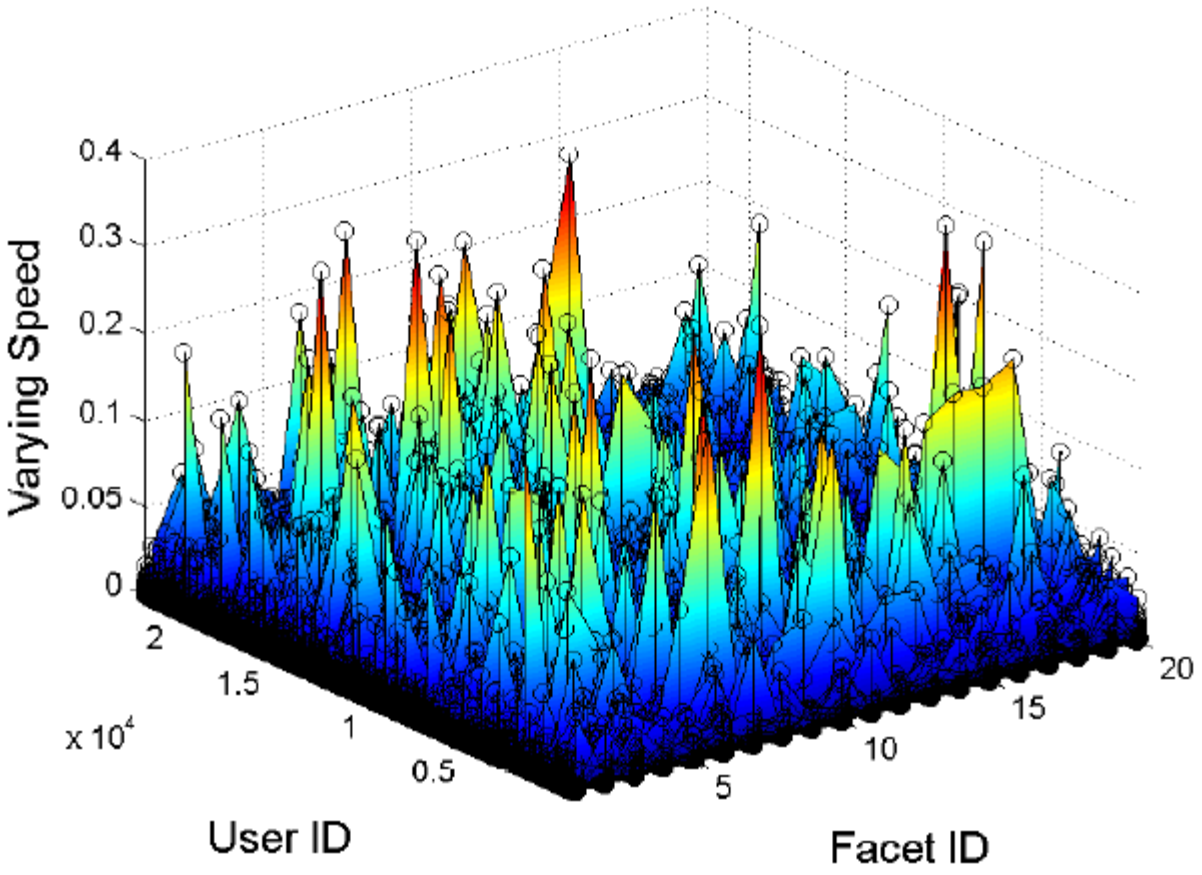


Closed Triad

用户兴趣爱好随时间变化图



用户和商品的变化速率图



未来工作

- 寻找更多的实际应用
  - Ranking evolution
  - Recommendation systems
  - Helpfulness prediction

SiteReco™

智能商品推荐

谢 谢 ！