基于复杂回复网络的影响因子研究

2012级　李锡涵　竺可桢学院　计算机科学与技术学院

2012级　陈凡　竺可桢学院　数学科学学院

导师：　金小刚　副教授　计算机科学与技术学院

2015年11月28日

背景

- 什么是回复网络？
  - 回复网络（Reply Network）是指基于用户之间的相互回复而形成的网络，以网络论坛/贴吧为典型代表，与以微博/Twitter为代表的转发网络相比具有许多独特的性质。
  - 主题帖与回复
  - 版面与社群
- 问题
  - 回复网络有哪些基本性质？
  - 如何度量回复网络之间的用户关系，“谁与我相似，在哪个方面相似”？
  - 回复网络随着时间的变化而如何发展？

数据集

- 基于某中型规模论坛从2004年到2014年的原始数据库
  - 69个版面（N=69）
  - 6731位用户（M=6731）
  - 57053篇主题，366060篇文章
- 对数据库进行Dump操作，储存为SQL文件并导入本地（图1）
- 编写Python程序，从原始数据库中抽取并整理以下信息
  - 对于每个回复，记录所属的版面、内容、发帖者、回复用户对象
  - 对于每个用户$i\left(1 \le i \le M\right)$，记录所有回复所属的版面、内容和总量$n_i$；特别地，统计每个用户在每个版面$j(1 \le j \le N)$的发帖总数$b_{ij}$，以及他所参与的版面总数$\left|b_i\right| = \#\left\{b_{ij} > 0 | 1 \le j \le N\right\}$

数据集

- 编写Python程序，统计基本信息
  - 帖子总量，每月帖子数量等（图2）
- 编写Python程序，建立回复网络
  - 使用networkx工具包，建立基础网络模型
  - 对于每个回复，在发表回复的用户u与回复对象v之间建立连边
    - 如果指定了回复对象v，则边（u,v）的权值+1
    - 否则将u与主题帖发布者p之间的权值+1
  - 使用matplotlab绘制网络图像（图3，图4）
  - 导出回复网络为标准网络文件（.net）

数据集

- 编写Python程序，对回复网络的节点度数进行统计
- 幂律分布现象（图5，图6）

影响因子：
用户相似度的度量与计算

- 结合回复网络特点，对用户之间的相似度进行度量
- 我们提出三个相似度维度
  - 行为风格相似性
  - 兴趣相似性
  - 交际结构相似性

1、行为风格相似性

- 通过发帖数/版面数衡量用户活跃水平，通过用户在各版面活动的方差衡量用户活动均匀程度，两者作为一个二维向量取夹角。
- 设n_i为每个用户的发帖数量，$|b_i|$为每个用户活动版面的数量，我们定义一个**描述活跃水平的统计量μ_i**以及**描述用户活动均匀程度的统计量${\sigma_i}^2$**，即

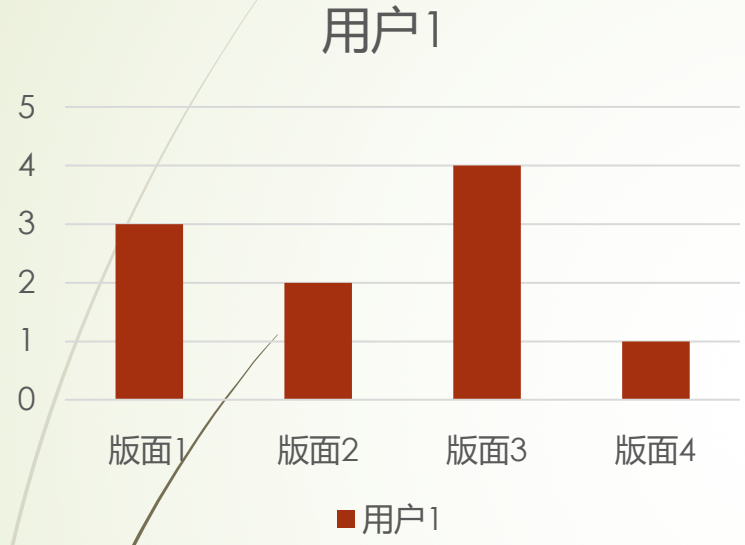
$$\mu_i = \frac{n_i}{|b_i|},$$

$$\sigma_i{}^2 = \frac{\sum_{j=1}^{|b_i|}(b_{ij}-\mu)^2}{|b_i|}, \sigma_i = \sqrt{\sigma_i{}^2}$$

- 行为风格相似性r_{ij}^S为：

$$r_{ij}^B = \frac{1}{2} + \frac{1}{2\pi}\cos^{-1}\frac{\mu_i\mu_j + \sigma_i\sigma_j}{\sqrt{\mu_i{}^2 + \sigma_i{}^2}\sqrt{\mu_j{}^2 + \sigma_j{}^2}}$$

行为风格相似性：举例



活跃水平$\mu_1 = \frac{n_1}{|b_1|} = \frac{3+2+4+1}{4} = 2.5$

活动均匀程度${\sigma_1}^2 = \frac{\sum_{j=1}^{|b_i|}(b_{ij}-\mu)^2}{|b_i|} = \frac{0.25+0.25+2.25+2.25}{4} = 1.25$



$\mu_2 = \frac{n_1}{|b_1|} = \frac{12}{8} = 1.5$

${\sigma_2}^2 = \frac{\sum_{j=1}^{|b_i|}(b_{ij}-\mu)^2}{|b_i|} = \frac{0.25\times 8}{8} = 0.25$

2、兴趣相似性

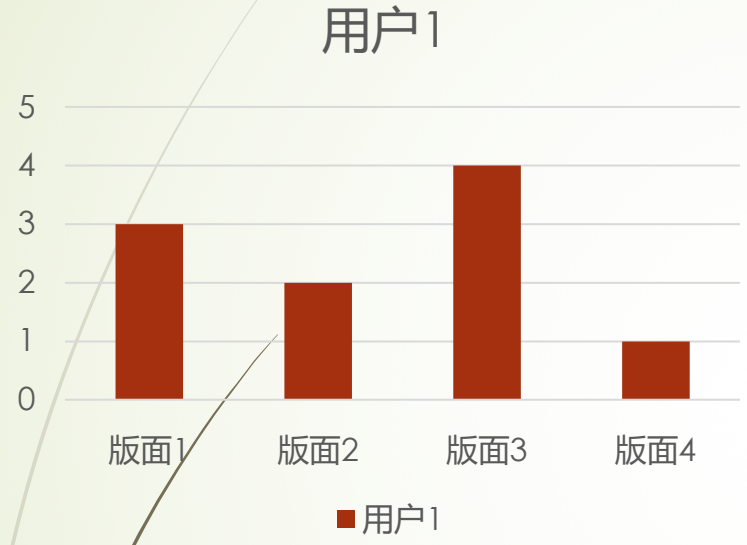
- 对用户在各版面的发帖数进行统计形成一个N维向量，求夹角作为相似度。
- 兴趣分布向量I_i：

$$I_i = \frac{(b_{i1}, b_{i2}, \dots, b_{iN})}{\sum_{j=1}^{N} b_{ij}} = (I_{i1}, I_{i2}, \dots, I_{iN})$$

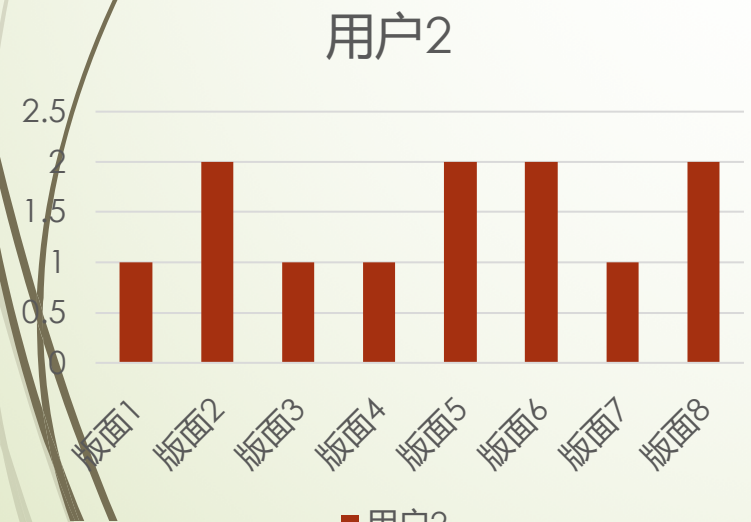
- 兴趣相似度r_{ij}^I为：

$$r_{ij}^I = \frac{\sum_{k=1}^{N}(I_{ik} - \overline{I_i})(I_{jk} - \overline{I_j})}{\sqrt{\sum_k (I_{ik} - \overline{I_i})^2}\sqrt{\sum_k \left(I_{jk} - \overline{I_j}\right)^2}}.$$

兴趣相似性：举例



兴趣分布向量I_1
$=(3，2，4，1，0，0，0，0)$



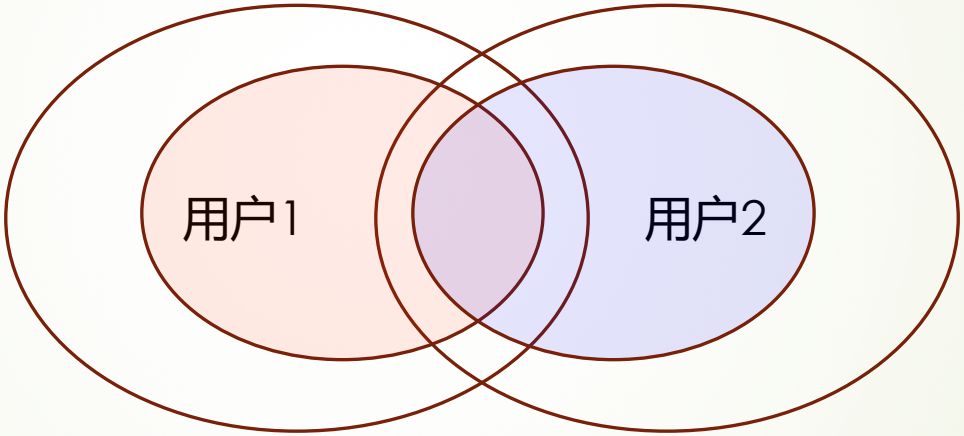
$I_2=(1，2，1，1，2，2，1，2)$

3、交际结构相似性

- 求两个用户共同好友的交集用户数，除以共同好友的并集用户数，作为交际结构相似性。
- 我们要求两个节点i和j的共同一级好友k的熟悉程度（熟络度，用回复边权c_{ij}表达）必须具有一定程度ε的接近，才有可能两个用户的社交网络交际结构相似性（Topological Similar）有所贡献，若记$TS_{ij} = \#\left\{\left(k \,\middle|\, \left|\frac{c_{ik}}{\sum_k c_{ik}} - \frac{c_{jk}}{\sum_k c_{jk}}\right| < \varepsilon\right)\right\}$，$TF_{ij} = \#\{k \mid c_{ik} > 0, c_{jk} > 0\}$.
- 交际结构相似度为：

$$r_{ij}^{T} = \frac{TS_{ij}}{TF_{ij}}.$$

交际结构相似性：举例



计算：简易社会网络分析软件“SNAKit”

- 使用Java开发
- 具有下列功能
  - 读入先前使用Python处理并生成的网络文件
  - 对于每个用户，可以计算
    - 用户与所选择用户的行为风格相似性/兴趣相似性/交际结构相似性（降序排列）
    - 用户度中心性
    - 用户加权度中心性
    - 总发帖版面数
    - 总发帖数
    - 总回复数
    - 与选择用户的交互次数
    - ……
  - 其他功能（后面）
    - 计算影响因子权值矩阵
    - 重建随机图

软件界面

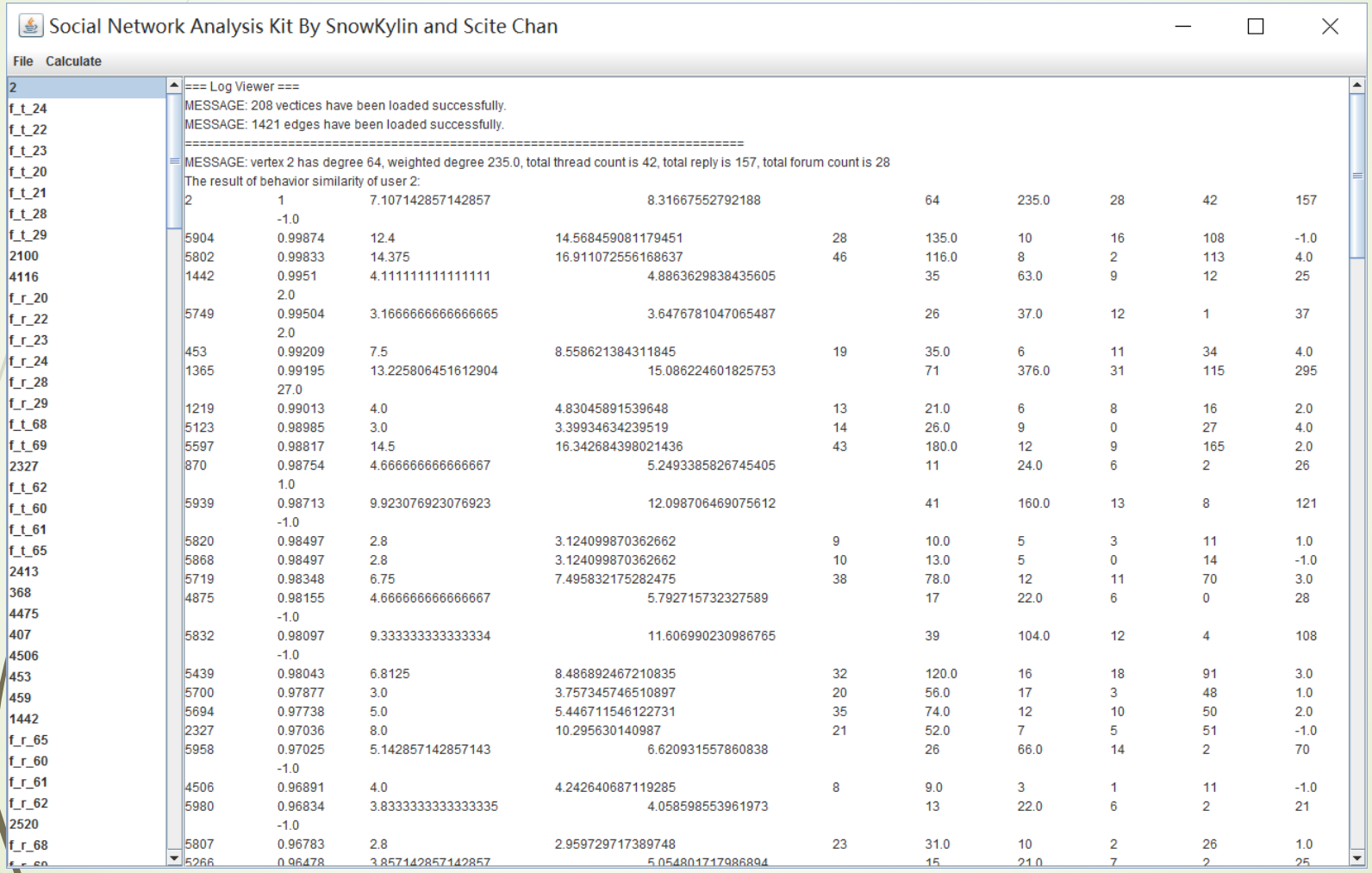


图7：软件界面示意图

软件使用案例分析

例如：我们想研究2013年时，ID为2的用户与其他用户的相似度，找出和该用户行为风格、兴趣、交际结构较相似的用户。
操作：载入2013年网络文件，侧边栏选择ID为2的用户。

行为风格相似性

用户ID	行为风格相似性（降序排列）	μ_i	σ_i^2	度	加权度（加边的权值）	总版面活跃数	总帖子数	总回复数	与所选用户边权值
2	1	7.107143	8.316676	64	235	28	42	157	-1
5904	0.99874	12.4	14.56846	28	135	10	16	108	-1
5802	0.99833	14.375	16.91107	46	116	8	2	113	4
1442	0.9951	4.111111	4.886363	35	63	9	12	25	2
5749	0.99504	3.166667	3.647678	26	37	12	1	37	2
453	0.99209	7.5	8.558621	19	35	6	11	34	4
1365	0.99195	13.22581	15.08622	71	376	31	115	295	27
1219	0.99013	4	4.830459	13	21	6	8	16	2
5123	0.98985	3	3.399346	14	26	9	0	27	4
5597	0.98817	14.5	16.34268	43	180	12	9	165	2

兴趣相似性

用户ID	兴趣相似性（降序排列）	度	加权度（加边的权值）	总版面活跃数	总帖子数	总回复数	与所选用户边权值
2	1	64	235	28	42	157	-1
5430	0.85507	37	131	22	24	71	21
5123	0.81815	14	26	9	0	27	4
1365	0.80396	71	376	31	115	295	27
1022	0.78433	41	91	18	24	41	5
4851	0.77942	31	82	15	20	61	7
599	0.77585	60	212	27	12	201	17
5996	0.77378	36	85	8	5	20	2
5670	0.76719	82	547	29	73	399	27
5936	0.7613	23	54	9	8	18	-1

软件使用案例分析

交际结构相似性							
用户ID	交际结构相似性（降序排列）	度	加权度（加边的权值）	总版面活跃数	总帖子数	总回复数	与所选用户边权值
2	1	64	235	28	42	157	-1
1365	1	71	376	31	115	295	27
5670	1	82	547	29	73	399	27
3696	1	62	187	16	20	177	-1
3508	0.9776	57	130	14	15	83	3
599	0.96906	60	212	27	12	201	17
4855	0.95028	60	116	21	16	108	7
5760	0.93905	46	137	18	17	92	1
5430	0.93264	37	131	22	24	71	21
5690	0.93264	37	78	12	21	60	4

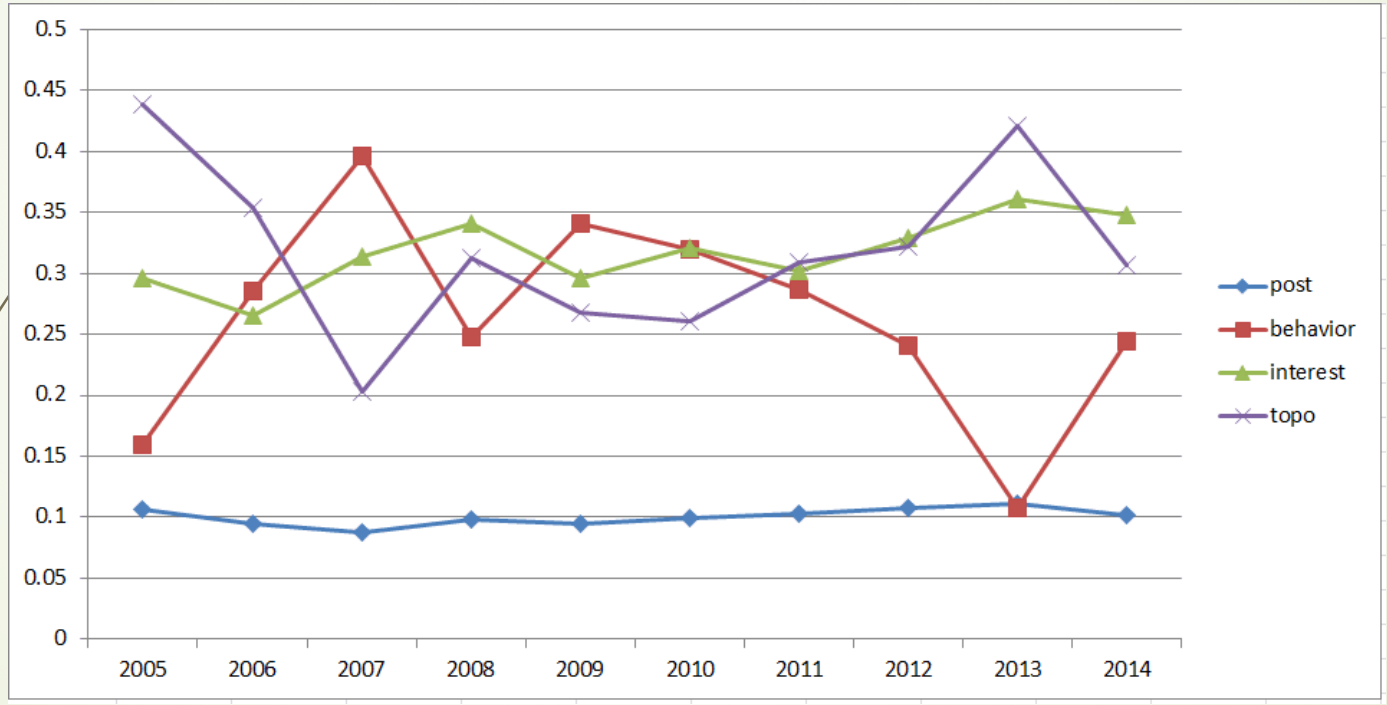
对三种相似度的权重计算

- 我们认为，随着在线社交网络的发展，三种相似度的权值（或重要度）会发生变化。
  - 例如，我们可以猜想，在社交网络发展的初期，一群相互熟识的人们组成小的社群，此时交际结构的相似性会居于主导地位。随着时间的发展，用户的兴趣和行为风格在其活动中得以显现，兴趣相似性和行为风格相似性的重要性逐渐提升。
- 我们希望计算，在社会网络的演化中，这三种相似度的权值（或重要度）如何变化。

对三种相似度的权重计算

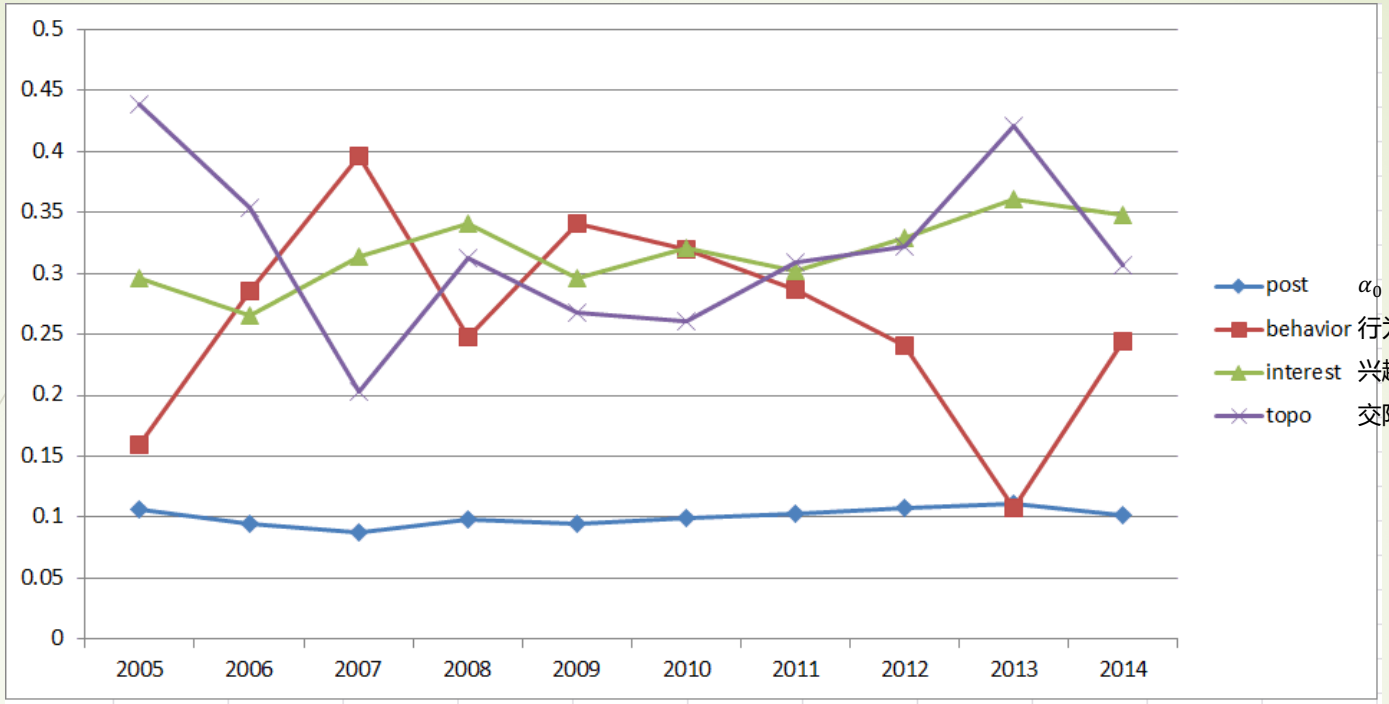
- 回顾我们之前建立的回复网络G
  - 节点为用户
  - 边为用户之间的回复关系（带权无向图，权值越高相互回复越频繁）
- 我们根据popularity×similarity原理，对用户之间的回复关系进行预测
  - $x_{ij} = \left(n_i \cdot n_j\right)^{\alpha_0} \cdot (r_{ij}^1)^{\alpha_1} \cdot (r_{ij}^2)^{\alpha_2} \cdot (r_{ij}^3)^{\alpha_3}$
  - 得到新图G'（α_0，α_1，α_2，α_3），节点为用户，边为x_{ij}
  - 对原图G进行拟合，求出拟合情况最好的$\alpha_0 \sim \alpha_3$
- 以年为单位，对每年的回复网络求$\alpha_0 \sim \alpha_3$

结果与讨论：
回复网络随时间的演化

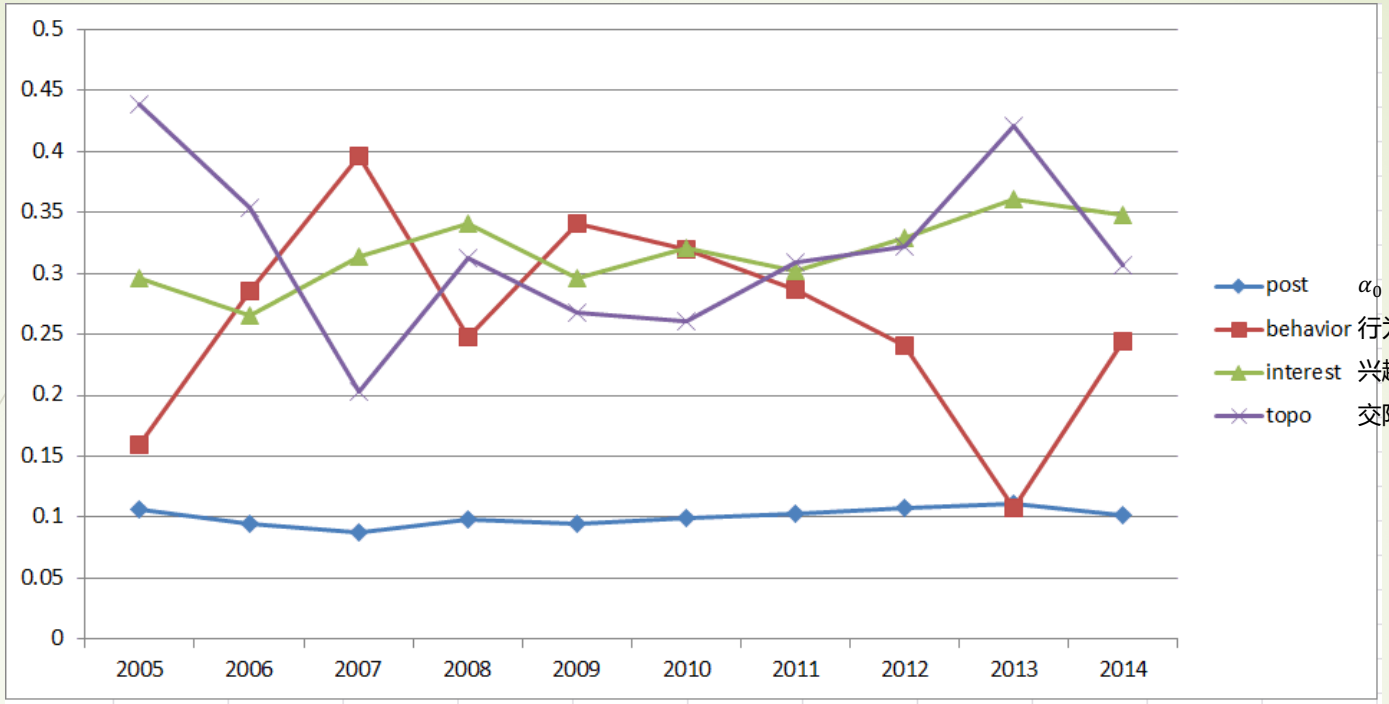


Post: α_0
behavior: 行为风格相似度（α_1）
interest: 兴趣相似度（α_2）
topo: 交际结构相似度（α_3）

图8：回复网络随时间的演化
（2005-2014）



时间	事件	结果
2005	论坛建立初期，相互熟识的用户群结成初始社群	交际结构相似度处于主导地位
2007	论坛出现了版面的解散和合并，部分交际结构紧密的用户群被强行打散	交际结构相似性跌落到峰谷
2013	论坛遭遇黑客攻击，老用户大量流失，同时新用户进入	行为风格相似度跌落到峰谷，交际结构重新占据主导地位，与用户数较少的2005年类似



我们注意到：
兴趣相似度在回复网络随时间的演化过程中始终保持稳定，受回复网络内部发生的时间影响不明显，且有逐渐缓慢上升的趋势。而行为风格相似度和交际结构相似度则会由于回复网络内部发生的事件而发生较大变化。

参考文献

- van Steen M. Graph Theory and Complex Networks: An Introduction (2011), 285 pages[R]. ISBN 978-90-815406-1-2.
- Kouznetsov M T A. Social Network Analysis for Startups[J]. 2011.
- Papadopoulos F, Kitsak M, Serrano M Á, et al. Popularity versus similarity in growing networks[J]. Nature, 2012, 489(7417): 537-540.
- 伊斯利, 克莱因伯格, 李晓明, 等. 网络, 群体与市场: 揭示高度互联世界的行为原理与效应机制[M]. 清华大学出版社, 2011.
- 陈为, 张嵩, 鲁爱东. 数据可视化的基本原理与方法[J]. 2013.

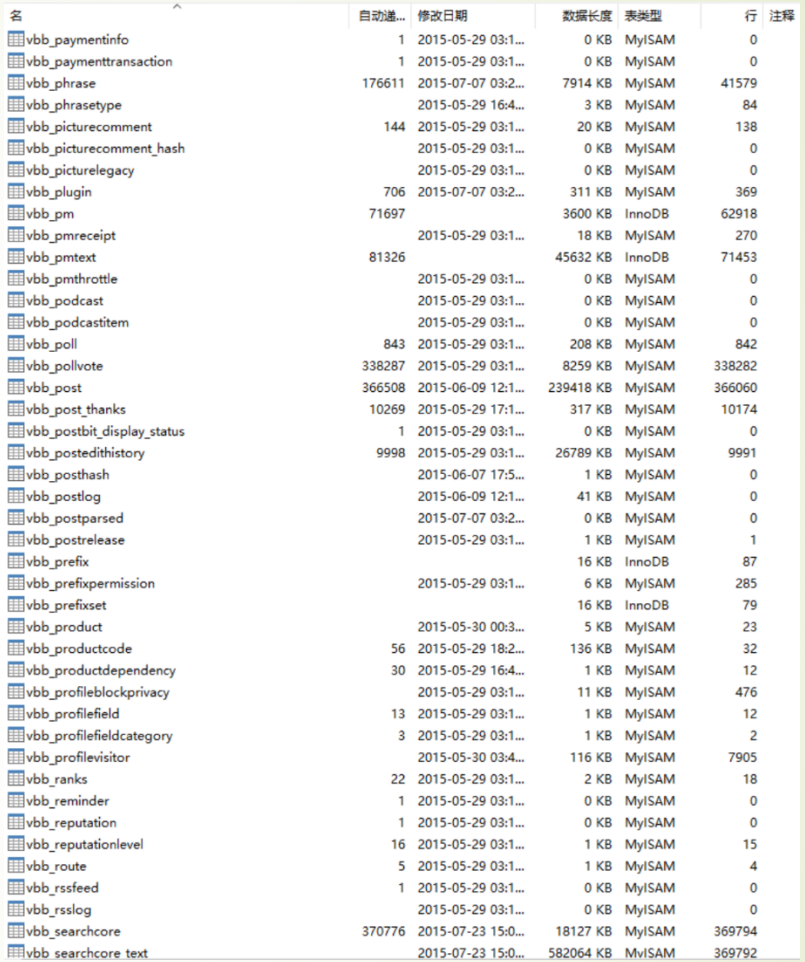
谢谢！

2012级　李锡涵　竺可桢学院　　计算机科学与技术学院
2012级　陈凡　　竺可桢学院　　数学科学学院

图1

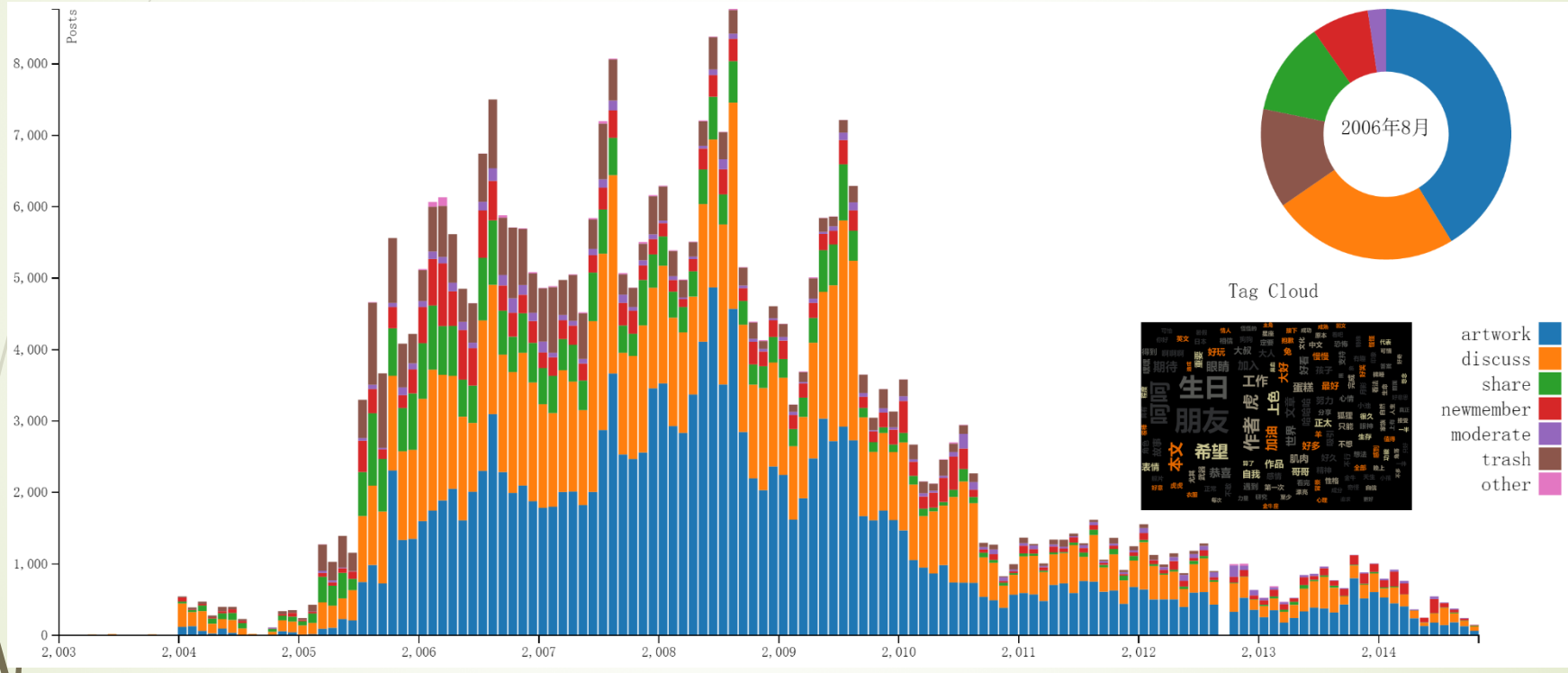
对数据库进行Dump操作，储存为SQL文件并导入本地

返回



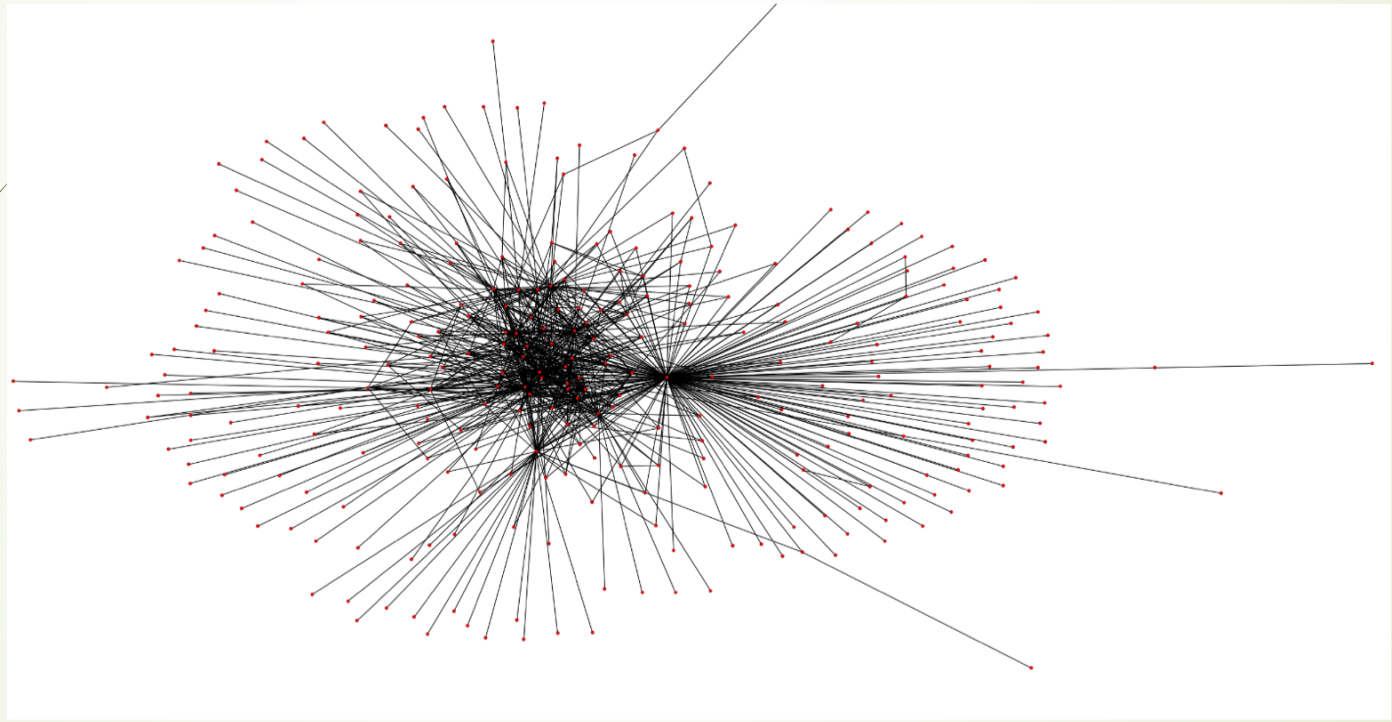
名	自动递...	修改日期	数据长度	表类型	行	注释
vbb_paymentinfo	1	2015-05-29 03:1...	0 KB	MyISAM	0	
vbb_paymenttransaction	1	2015-05-29 03:1...	0 KB	MyISAM	0	
vbb_phrase	176611	2015-07-07 03:2...	7914 KB	MyISAM	41579	
vbb_phrasetype		2015-05-29 16:4...	3 KB	MyISAM	84	
vbb_picturecomment	144	2015-05-29 03:1...	20 KB	MyISAM	138	
vbb_picturecomment_hash		2015-05-29 03:1...	0 KB	MyISAM	0	
vbb_picturelegacy		2015-05-29 03:1...	0 KB	MyISAM	0	
vbb_plugin	706	2015-07-07 03:2...	311 KB	MyISAM	369	
vbb_pm	71697		3600 KB	InnoDB	62918	
vbb_pmreceipt		2015-05-29 03:1...	18 KB	MyISAM	270	
vbb_pmtext	81326		45632 KB	InnoDB	71453	
vbb_pmthrottle		2015-05-29 03:1...	0 KB	MyISAM	0	
vbb_podcast		2015-05-29 03:1...	0 KB	MyISAM	0	
vbb_podcastitem		2015-05-29 03:1...	0 KB	MyISAM	0	
vbb_poll	843	2015-05-29 03:1...	208 KB	MyISAM	842	
vbb_pollvote	338287	2015-05-29 03:1...	8259 KB	MyISAM	338282	
vbb_post	366508	2015-06-09 12:1...	239418 KB	MyISAM	366060	
vbb_post_thanks	10269	2015-05-29 17:1...	317 KB	MyISAM	10174	
vbb_postbit_display_status	1	2015-05-29 03:1...	0 KB	MyISAM	0	
vbb_postedithistory	9998	2015-05-29 03:1...	26789 KB	MyISAM	9991	
vbb_posthash		2015-06-07 17:5...	1 KB	MyISAM	0	
vbb_postlog		2015-06-09 12:1...	41 KB	MyISAM	0	
vbb_postparsed		2015-07-07 03:2...	0 KB	MyISAM	0	
vbb_postrelease		2015-05-29 03:1...	1 KB	MyISAM	1	
vbb_prefix			16 KB	InnoDB	87	
vbb_prefixpermission		2015-05-29 03:1...	6 KB	MyISAM	285	
vbb_prefixset			16 KB	InnoDB	79	
vbb_product		2015-05-30 00:3...	5 KB	MyISAM	23	
vbb_productcode	56	2015-05-29 18:2...	136 KB	MyISAM	32	
vbb_productdependency	30	2015-05-29 16:4...	1 KB	MyISAM	12	
vbb_profileblockprivacy		2015-05-29 03:1...	11 KB	MyISAM	476	
vbb_profilefield	13	2015-05-29 03:1...	1 KB	MyISAM	12	
vbb_profilefieldcategory	3	2015-05-29 03:1...	1 KB	MyISAM	2	
vbb_profilevisitor		2015-05-30 03:4...	116 KB	MyISAM	7905	
vbb_ranks	22	2015-05-29 03:1...	2 KB	MyISAM	18	
vbb_reminder	1	2015-05-29 03:1...	0 KB	MyISAM	0	
vbb_reputation	1	2015-05-29 03:1...	0 KB	MyISAM	0	
vbb_reputationlevel	16	2015-05-29 03:1...	1 KB	MyISAM	15	
vbb_route	5	2015-05-29 03:1...	1 KB	MyISAM	4	
vbb_rssfeed	1	2015-05-29 03:1...	0 KB	MyISAM	0	
vbb_rsslog		2015-05-29 03:1...	0 KB	MyISAM	0	
vbb_searchcore	370776	2015-07-23 15:0...	18127 KB	MyISAM	369794	
vbb_searchcore_text		2015-07-23 15:0...	582064 KB	MyISAM	369792	

图2：论坛发帖数随时间的变化关系统计图



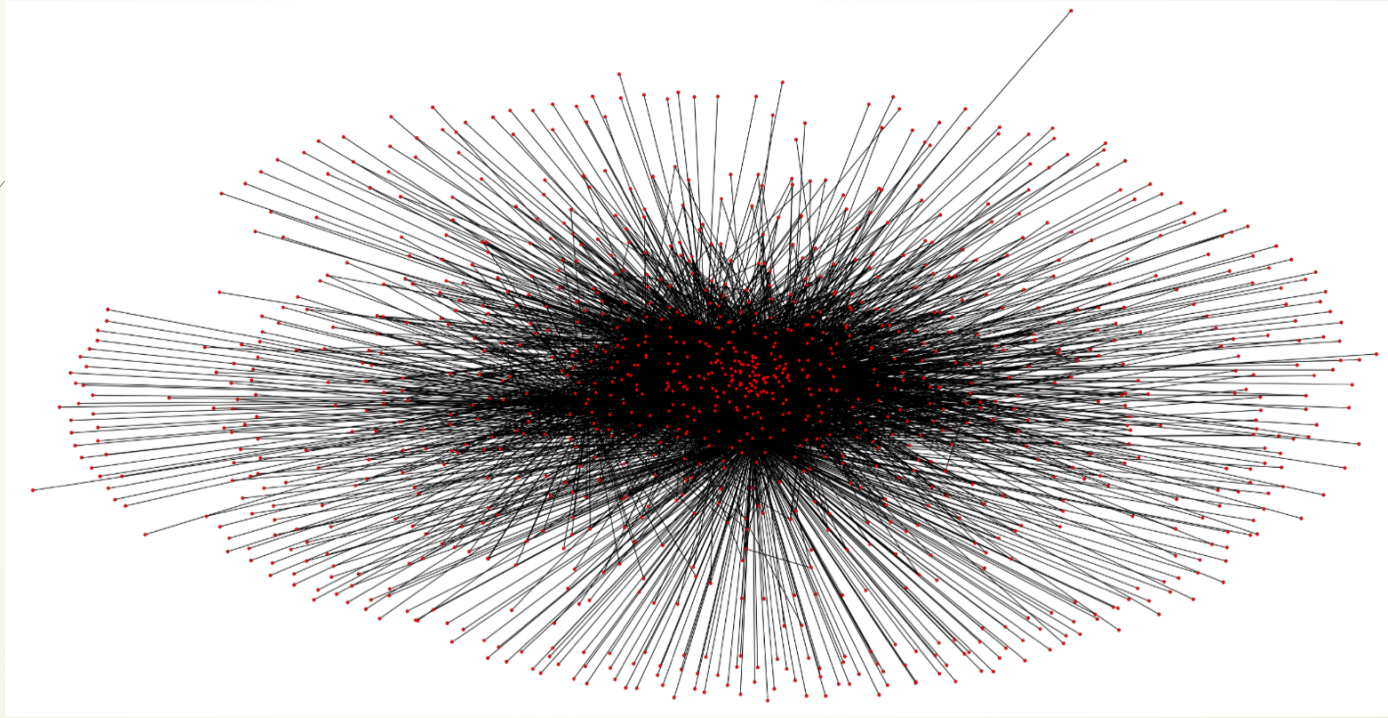
返回

图3：2005年1月至3月论坛用户回复网络



返回

图4：2005年全年用户回复网络



返回

图5、图6

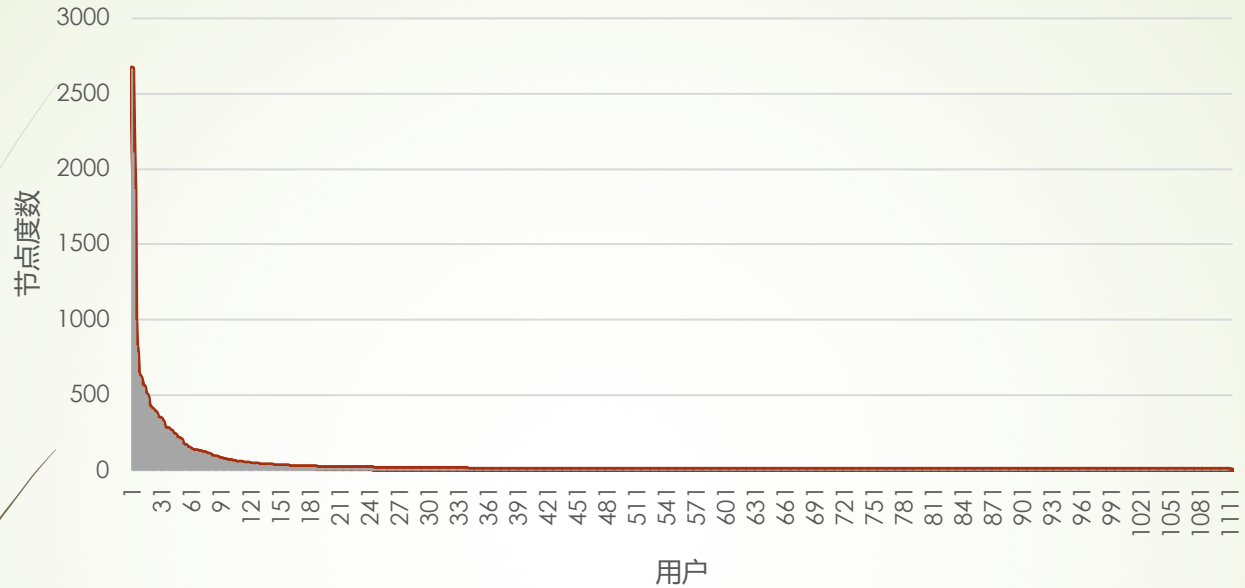


图5：2005年全年回复网络节点度分布

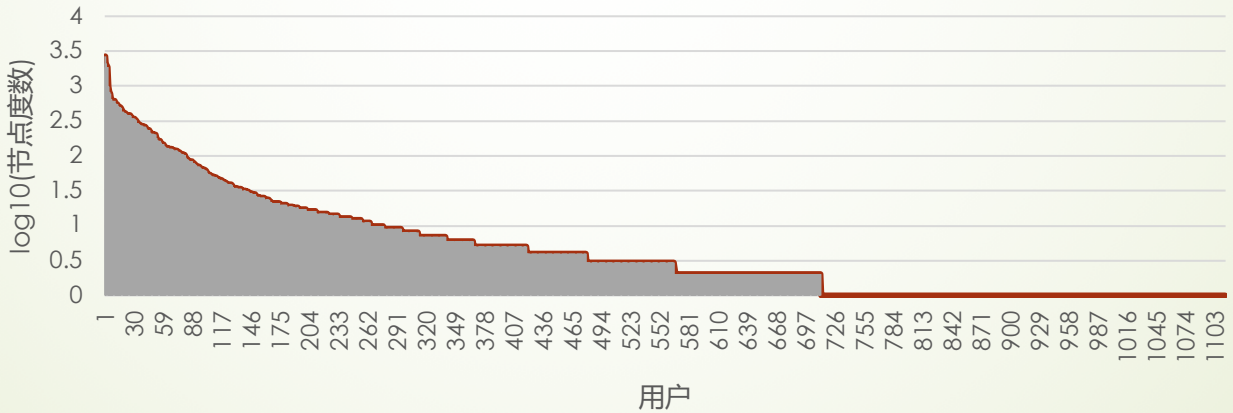


图6：节点度数取对数后的2005年全年回复网络节点度分布

返回