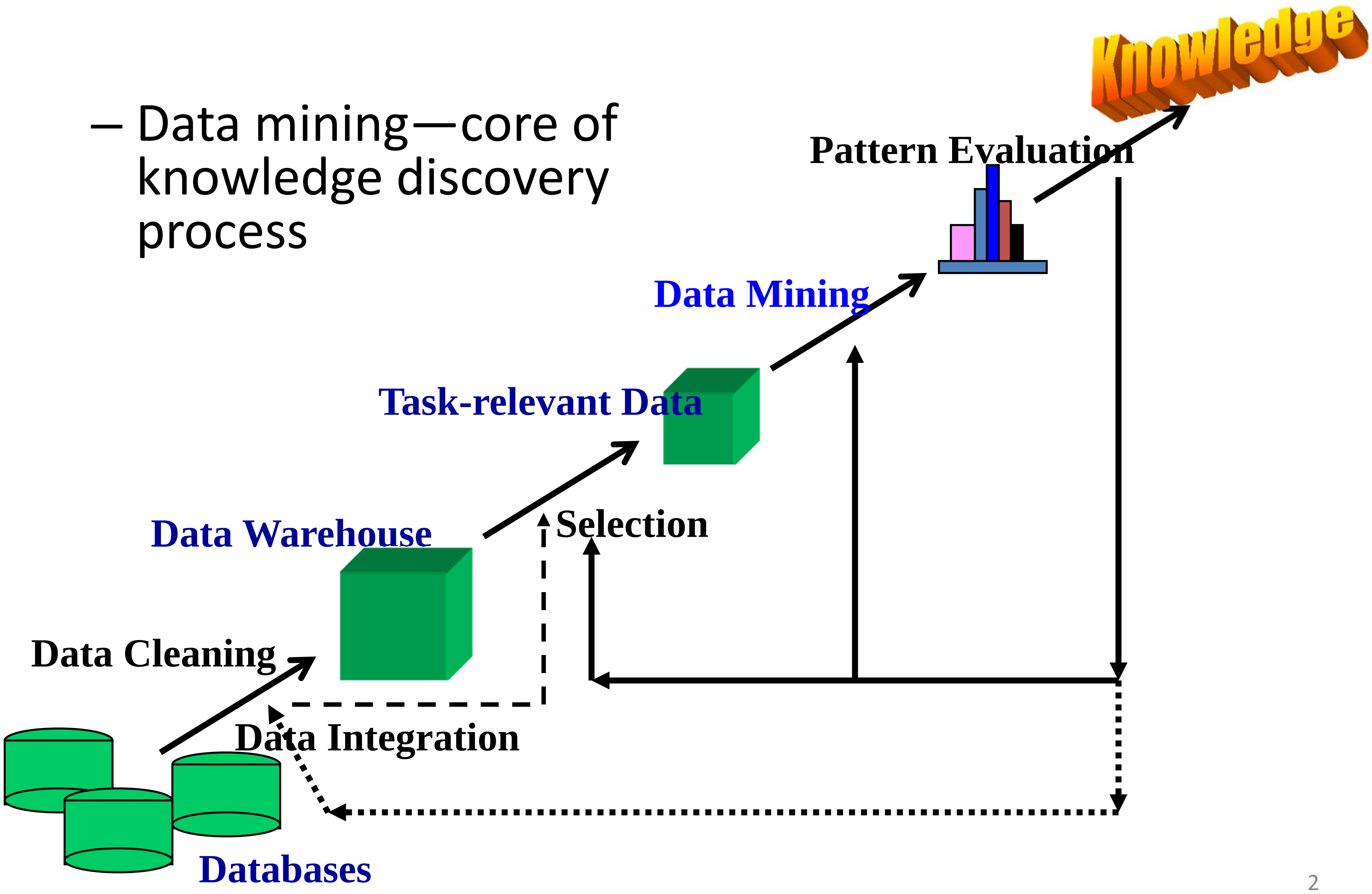
Social Big Data - Applications

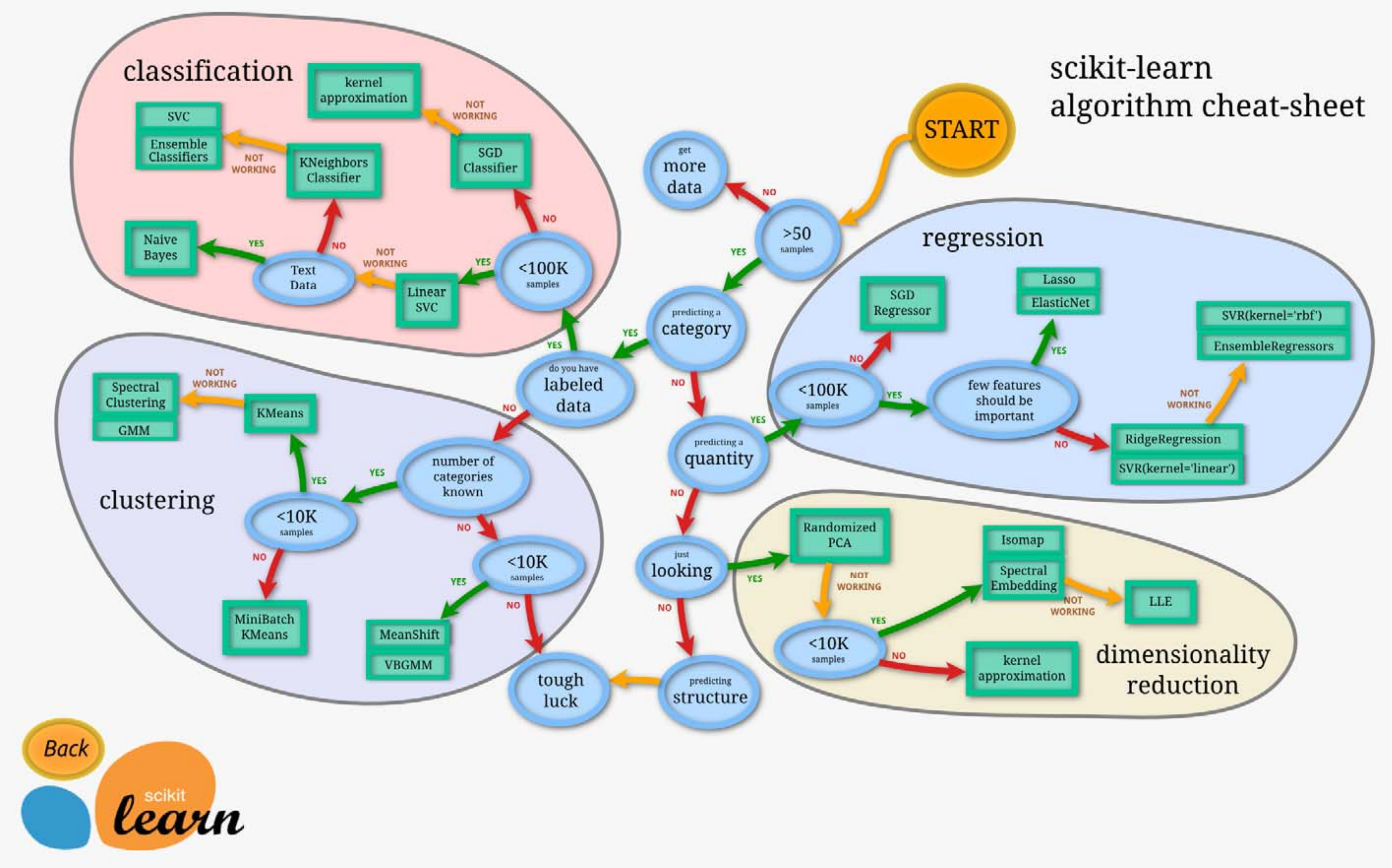
Dr. Hong Huang

Knowledge Discovery (KDD) Process

- Data mining—core of knowledge discovery process



Data mining algorithms



What can we do in social network?

- Community identification
- Influential user identification
- Link prediction
- Point of interest recommendation
- Disease prediction
- Crime prediction
- Event monitoring
- …

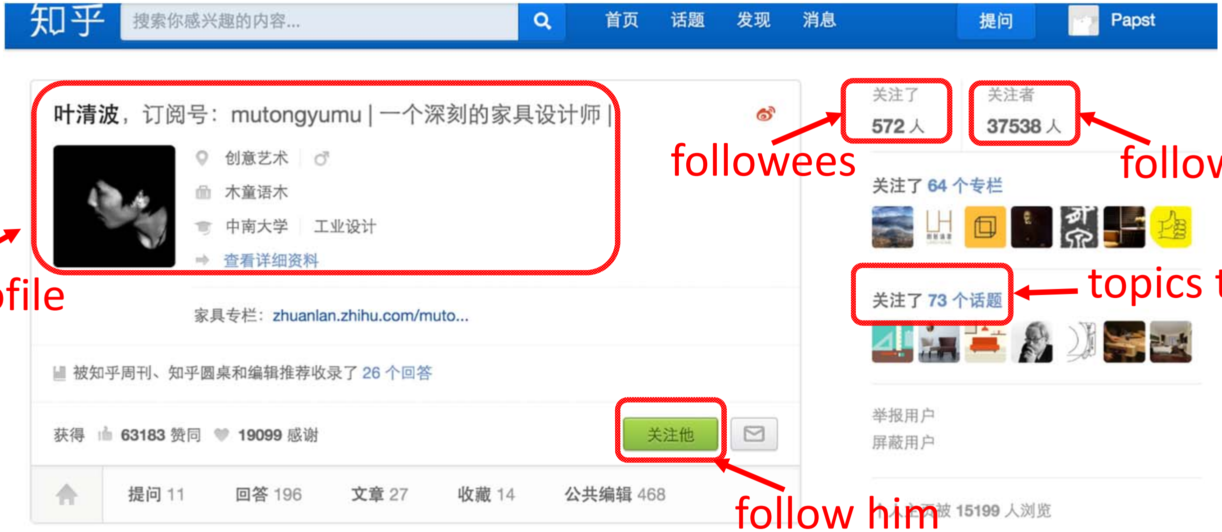
Opinion Leader Mining in CQA

Community question answering (CQA) sites

- What is CQA site
  - Allow users to answer the questions posted by other users
  - Give positive or negative judgments to answers provided by others via voting
  - Popular QA portals: Yahoo! Answers, Stack Overflow, Quora, Zhihu

An innovative CQA -- Zhihu

- What is Zhihu (Chinese Quora)
  - Traditional QA functions
    - Ask & answer questions
    - Vote answers
  - Social functions
    - Follow users
    - Follow topics and questions



叶清波，订阅号：mutongyumu | 一个深刻的家具设计师 |

创意艺术

木童语木

中南大学 | 工业设计

查看详细资料

家具专栏：zhuanlan.zhihu.com/muto...

被知乎周刊、知乎圆桌和编辑推荐收录了 26 个回答

获得 63183 赞同 19099 感谢

关注他

提问 11 回答 196 文章 27 收藏 14 公共编辑 468

关注了 572 人

关注者 37538 人

关注了 64 个专栏

关注了 73 个话题

举报用户
屏蔽用户

个人主页被 15199 人浏览

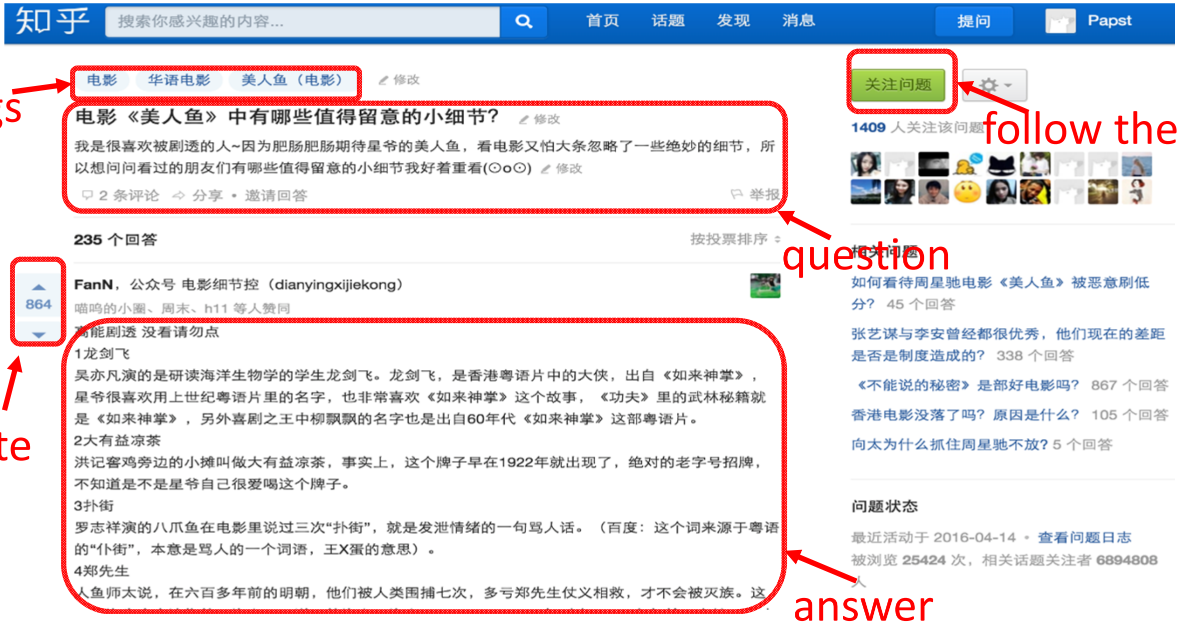
user profile

followees

followers

topics the user follows

follow him



电影 华语电影 美人鱼（电影） 修改

电影《美人鱼》中有哪些值得留意的小细节？ 修改

我是很喜欢被剧透的人~因为肥肠肥肠期待星爷的美人鱼，看电影又怕大条忽略了一些绝妙的细节，所以想问问看过的朋友们有哪些值得留意的小细节我好着重看(⊙o⊙) 修改

2 条评论 分享 • 邀请回答 举报

235 个回答 按投票排序

864

FanN，公众号 电影细节控（dianyingxijiekong）

喵呜的小圈、周末、h11 等人赞同

高能剧透 没看请勿点

1龙剑飞

吴亦凡演的是研读海洋生物学的学生龙剑飞。龙剑飞，是香港粤语片中的大侠，出自《如来神掌》，星爷很喜欢用上世纪粤语片里的名字，也非常喜欢《如来神掌》这个故事，《功夫》里的武林秘籍就是《如来神掌》，另外喜剧之王中柳飘飘的名字也是出自60年代《如来神掌》这部粤语片。

2大有益凉茶

洪记窖鸡旁边的小摊叫做大有益凉茶，事实上，这个牌子早在1922年就出现了，绝对的老字号招牌，不知道是不是星爷自己很爱喝这个牌子。

3扑街

罗志祥演的八爪鱼在电影里说过三次“扑街”，就是发泄情绪的一句骂人话。（百度：这个词来源于粤语的“仆街”，本意是骂人的一个词语，王X蛋的意思）。

4郑先生

人鱼师太说，在六百多年前的明朝，他们被人类围捕七次，多亏郑先生仗义相救，才不会被灭族。这

关注问题

1409 人关注该问题

相关问题

如何看待周星驰电影《美人鱼》被恶意刷低分？ 45 个回答

张艺谋与李安曾经都很优秀，他们现在的差距是否是制度造成的？ 338 个回答

《不能说的秘密》是部好电影吗？ 867 个回答

香港电影没落了吗？原因是什么？ 105 个回答

向太为什么抓住周星驰不放？5 个回答

问题状态

最近活动于 2016-04-14 • 查看问题日志

被浏览 25424 次，相关话题关注者 6894808 人

topic tags

follow the question

question

vote

answer

Opinion leader identification

- What is opinion leader
  - Give their influential comments and opinions, put forward guiding ideas, agitate and guide the public to understand social problems[1]
- Topical opinion leader in Zhihu
  - Give authoritative and influential answers, comments and other activities in some topic area
  - Play an important role in promoting formation and management of online public opinion and knowledge base

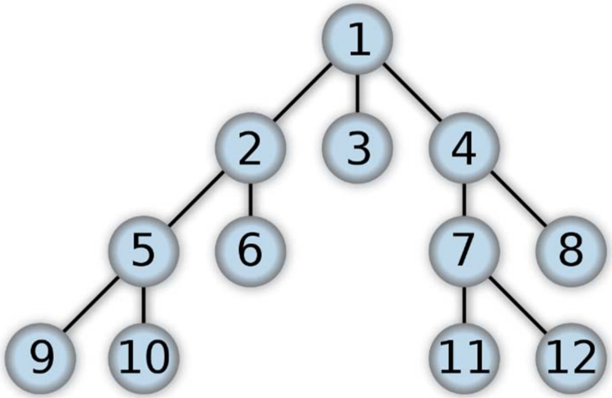
[1] Lazarsfeld, P.F., Berelson, B., Gaudet, H.: The People's Choice: How the Voter Makes up His Mind in a Presidential Campaign. New York: Columbia University Press, (1948)

Benefits of opinion leader identification

- Users: realize public opinion and authoritative knowledge, get specific answers efficiently
- Zhihu: invite them to attend public activities(e.g., editing, publication) to attract more users
- Marketing: influence customer opinions on products and services
- Government: realize, guide and interfere public opinion on the internet

Dataset collection

- We gathered the Zhihu dataset through a web-based crawler from March to June in 2016
  - Start the crawler using a set of 10 popular Zhihu users from a webpage
  - The crawler follows a **Breadth First Search** (BFS) pattern through the follower and followee links of each user.



BFS

Dataset

- Each user data contains
  - user ID, the lists of the user’s followers and followees, the user’s answers and questions posted
- Each answer/question data contains
  - its topic and the number of received votes

Total number of users	**1,411,669**
Total number of questions	701,982
Total number of answers	4,047,183
Total number of topics	160,664

Data analysis tools

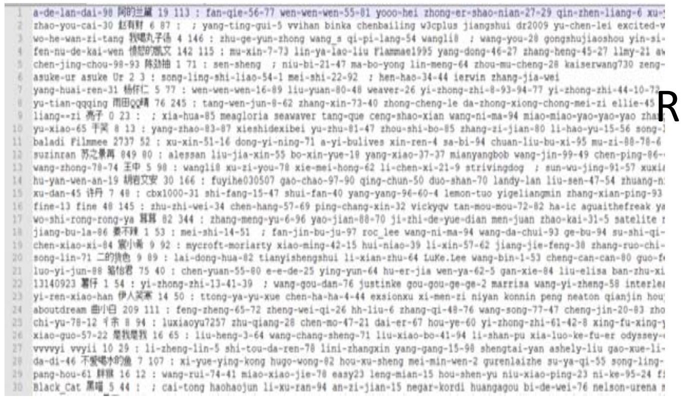
- Depending on needs, some data analysis tools are as follows
  - MATLAB, or its open-source alternatives, Scilab and GNU Octave (great at dealing with numbers)
  - Python with libraries like Numpy, Scipy and Matplotlib (great for general purpose data analysis- particularly good at interacting with other tools)
  - R (Great for statistics)
- Use Python to process the Zhihu dataset

Python libraries for data science

- **NumPy**
  - the foundational library for scientific computing in Python, and many of the libraries on this list use NumPy arrays as their basic inputs and outputs. In short, NumPy introduces objects for multidimensional arrays and matrices, as well as routines that allow developers to perform advanced mathematical and statistical functions on those arrays with as little code as possible.
- **SciPy**
  - builds on NumPy by adding a collection of algorithms and high-level commands for manipulating and visualizing data. This package includes functions for computing integrals numerically, solving differential equations, optimization, and more.
- **Pandas**
  - adds data structures and tools that are designed for practical data analysis in finance, statistics, social sciences, and engineering. Pandas works well with incomplete, messy, and unlabeled data (i.e., the kind of data you're likely to encounter in the real world), and provides tools for shaping, merging, reshaping, and slicing datasets.
- **scikit-learn**
  - builds on NumPy and SciPy by adding a set of algorithms for common machine learning and data mining tasks, including clustering, regression, and classification. As a library, scikit-learn has a lot going for it. Its tools are well-documented and its contributors include many machine learning experts. What's more, it's a very curated library, meaning developers won't have to choose between different versions of the same algorithm. Its power and ease of use make it popular with a lot of data-heavy startups, including Evernote, OKCupid, Spotify, and Birchbox.
- **NetworkX**
  - allows you to create and analyze graphs and networks. It's designed to work with both standard and nonstandard data formats, which makes it especially efficient and scalable. All this makes NetworkX especially well suited to analyzing complex social networks.

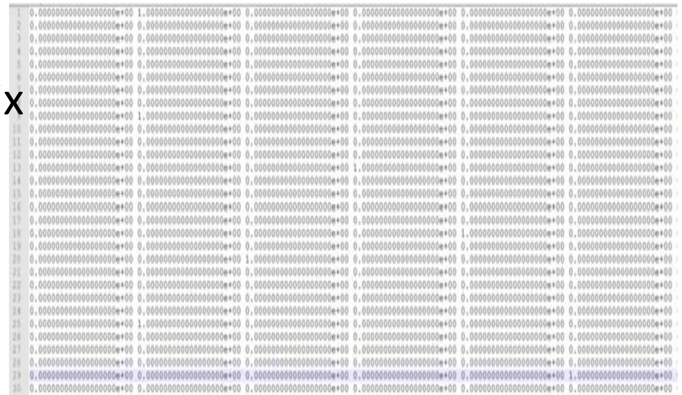
Data preprocessing

- Data cleaning
  - Remove incomplete user data
- Data transformation



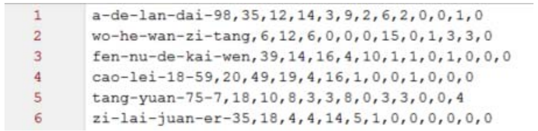
```
a-de-lan-dai-98 阿的兰黛 19 113 : fan-qie-56-77 wen-wen-wen-55-81 yooo-hei zhong-er-shao-nian-27-29 qin-zhen-liang-6 xu-
zhao-you-cai-30 赵有财 6 87 :  ; yang-ting-gui-5 vvihan binka chenbailing w3cplus jiangshui dr2009 yu-chen-lei excited-v
wo-he-wan-zi-tang 我喝丸子汤 4 146 : zhu-ge-yun-zhong wang_s qi-pi-lang-54 wangli8  ; wang-you-28 gongzhujiaoshou yin-si-
fen-nu-de-kai-wen 愤怒的凯文 142 115 : mu-xin-7-73 lin-ya-lao-liu Flammae1995 yang-dong-46-27 zhang-heng-45-27 llmy-21 aw
chen-jing-chou-98-93 陈劲抽 1 71 : sen-sheng  ; niu-bi-21-47 ma-bo-yong lin-meng-64 zhou-mu-cheng-28 kaiserwang730 zeng-
asuke-ur asuke Ur 2 3 : song-ling-shi-liao-54-1 mei-shi-22-92  ; hen-hao-34-44 ierwin zhang-jia-wei
yang-huai-ren-31 杨怀仁 5 77 : wen-wen-wen-16-89 liu-yuan-80-48 weaver-26 yi-zhong-zhi-8-93-94-77 yi-zhong-zhi-44-10-72
yu-tian-qqing 雨田QQ晴 76 245 : tang-wen-jun-8-62 zhang-xin-73-40 zhong-cheng-le da-zhong-xiong-chong-mei-zi ellie-45
liang--zi 亮子 0 23 :  ; xia-hua-85 meagloria seawaver tang-que ceng-shao-xian wang-ni-ma-94 miao-miao-yao-yao-yao zhan
yu-xiao-65 于笑 8 13 : yang-zhao-83-87 xieshidexibei yu-zhu-81-47 zhou-shi-bo-85 zhang-zi-jian-80 li-hao-yu-15-56 song-l
baladi Filmmee 2737 52 : xu-xin-51-16 dong-yi-ning-71 a-yi-bulives xin-ren-4 sa-bi-94 chuan-liu-bu-xi-95 mu-zi-88-78-6
suzinran 苏之景再 849 80 : alessan liu-jia-xin-55 bo-xin-yue-18 yang-xiao-37-37 mianyangbob wang-jin-99-49 chen-ping-86-
wang-zhong-78-74 王中 5 98 : wangli8 xu-zi-you-78 xie-mei-hong-62 li-chen-xi-21-9 strivingdog  ; sun-wu-jing-91-57 xuxia
hu-yan-wen-an-19 胡岩文安 30 166 : fuyihe030507 gao-chao-97-90 qing-chun-50 duo-shan-70 landy-lan liu-sen-47-54 zhuang-n
xu-dan-45 许丹 7 48 : cbx1000-31 shi-fang-15-47 shui-fan-40 yang-yang-96-60-4 lemon-tuo yigeliangmin zhang-xian-ping-93
fine-13 fine 48 145 : zhu-zhi-wei-34 chen-hang-57-69 ping-chang-xin-32 vickyqw tan-mou-mou-72-82 ha-ic aguaithefreak ya
wo-shi-rong-rong-ya 茸茸 82 344 : zhang-meng-yu-6-96 yao-jian-88-70 ji-zhi-de-yue-dian men-juan zhao-kai-31-5 satelite r
jiang-bu-la-86 姜不辣 1 53 : mei-shi-14-51  ; fan-jin-bu-ju-97 roc_lee wang-ni-ma-94 wang-da-chui-93 ge-bu-94 su-shi-qi-
chen-xiao-xi-84 紫小希 9 92 : mycroft-moriarty xiao-ming-42-15 hui-niao-39 li-xin-57-62 jiang-jie-feng-38 zhang-ruo-chi-
song-lin-71 二的执色 9 89 : lai-dong-hua-82 tianyishengshui li-xian-zhu-64 LuKe.Lee wang-bin-1-53 cheng-can-can-80 guo-f
luo-yi-jun-88 骆怡君 75 40 : chen-yuan-55-80 e-e-de-25 ying-yun-64 hu-er-jia wen-ya-62-5 gan-xie-84 liu-elisa ban-zhu-xi
13140923 薯仔 1 54 : yi-zhong-zhi-13-41-39  ; wang-gou-dan-76 justinke gou-gou-ge-ge-2 marrisa wang-yi-zheng-58 interlea
yi-ren-xiao-han 伊人笑寒 14 50 : ttong-ya-yu-xue chen-ha-ha-4-44 exsionxu xi-men-zi niyan konnin peng neaton qianjin hou
aboutdream 曲小白 209 111 : feng-zheng-65-72 zheng-wei-qi-26 hh-liu-6 zhang-qi-48-76 wang-song-77-47 cheng-jin-20-83 zho
chi-yu-78-12 千赤 8 94 : luxiaoyu7257 zhu-qiang-28 chen-mo-47-21 dai-er-67 hou-ye-60 yi-zhong-zhi-61-42-8 xing-fu-xing-y
xiao-guo-57-22 是我是我 16 65 : liu-heng-3-64 wang-chang-sheng-71 liu-xiao-bo-41-94 li-shan-pu xia-luo-ke-fu-er odyssey-
vvvvyi vvyii 10 29 : li-zheng-lin-5 shi-tou-da-ren-78 lini-zhangxin yang-gang-15-98 shengtai-yan ashely-liu gao-xue-li-
da-di-46 不爱喝水的鱼 7 107 : xi-yue-ying-kong hugo-wong-82 hou-xu-sheng mei-min-wen-2 gurenlaizhe su-ya-qi-55 song-ling-
pang-hou-61 胖猴 16 12 : wang-rui-74-41 miao-xiao-jie-78 easy23 leng-mian-15 hou-shen-yu niu-xiao-ping-23 ni-ke-95-24 fi
Black_Cat 黑喵 5 44 :  ; cai-tong haohaojun li-xu-ran-94 an-zi-jian-15 negar-kordi huangagou bi-de-wei-76 nelson-urena r
```

Relationship matrix



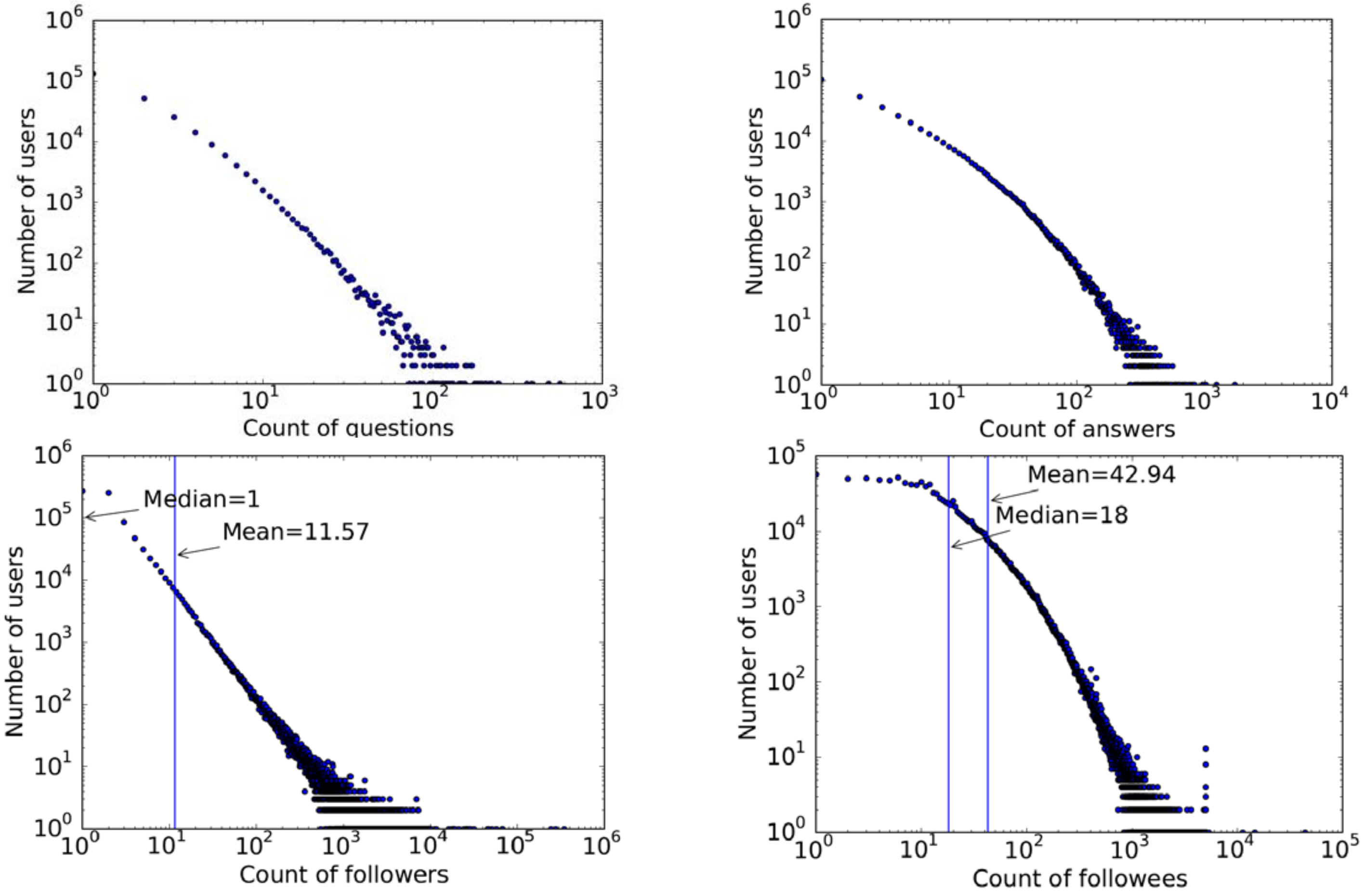
```
0.000000000000000000e+00 1.000000000000000000e+00 0.000000000000000000e+00 0.000000000000000000e+00 0.000000000000000000e+00 0.000000000000000000e+00
0.000000000000000000e+00 0.000000000000000000e+00 0.000000000000000000e+00 0.000000000000000000e+00 0.000000000000000000e+00 0.000000000000000000e+00
0.000000000000000000e+00 0.000000000000000000e+00 0.000000000000000000e+00 0.000000000000000000e+00 0.000000000000000000e+00 0.000000000000000000e+00
0.000000000000000000e+00 0.000000000000000000e+00 0.000000000000000000e+00 0.000000000000000000e+00 0.000000000000000000e+00 0.000000000000000000e+00
0.000000000000000000e+00 0.000000000000000000e+00 0.000000000000000000e+00 0.000000000000000000e+00 0.000000000000000000e+00 0.000000000000000000e+00
0.000000000000000000e+00 0.000000000000000000e+00 0.000000000000000000e+00 0.000000000000000000e+00 0.000000000000000000e+00 0.000000000000000000e+00
0.000000000000000000e+00 0.000000000000000000e+00 0.000000000000000000e+00 0.000000000000000000e+00 0.000000000000000000e+00 0.000000000000000000e+00
0.000000000000000000e+00 0.000000000000000000e+00 0.000000000000000000e+00 0.000000000000000000e+00 0.000000000000000000e+00 0.000000000000000000e+00
0.000000000000000000e+00 1.000000000000000000e+00 0.000000000000000000e+00 0.000000000000000000e+00 0.000000000000000000e+00 0.000000000000000000e+00
0.000000000000000000e+00 0.000000000000000000e+00 0.000000000000000000e+00 0.000000000000000000e+00 0.000000000000000000e+00 0.000000000000000000e+00
0.000000000000000000e+00 0.000000000000000000e+00 0.000000000000000000e+00 0.000000000000000000e+00 0.000000000000000000e+00 0.000000000000000000e+00
0.000000000000000000e+00 0.000000000000000000e+00 0.000000000000000000e+00 0.000000000000000000e+00 0.000000000000000000e+00 0.000000000000000000e+00
0.000000000000000000e+00 0.000000000000000000e+00 0.000000000000000000e+00 1.000000000000000000e+00 0.000000000000000000e+00 0.000000000000000000e+00
0.000000000000000000e+00 0.000000000000000000e+00 0.000000000000000000e+00 0.000000000000000000e+00 0.000000000000000000e+00 0.000000000000000000e+00
0.000000000000000000e+00 0.000000000000000000e+00 0.000000000000000000e+00 0.000000000000000000e+00 0.000000000000000000e+00 0.000000000000000000e+00
0.000000000000000000e+00 0.000000000000000000e+00 0.000000000000000000e+00 0.000000000000000000e+00 0.000000000000000000e+00 0.000000000000000000e+00
0.000000000000000000e+00 0.000000000000000000e+00 0.000000000000000000e+00 0.000000000000000000e+00 0.000000000000000000e+00 0.000000000000000000e+00
0.000000000000000000e+00 0.000000000000000000e+00 0.000000000000000000e+00 0.000000000000000000e+00 1.000000000000000000e+00 0.000000000000000000e+00
0.000000000000000000e+00 0.000000000000000000e+00 0.000000000000000000e+00 0.000000000000000000e+00 0.000000000000000000e+00 0.000000000000000000e+00
0.000000000000000000e+00 0.000000000000000000e+00 1.000000000000000000e+00 0.000000000000000000e+00 0.000000000000000000e+00 0.000000000000000000e+00
0.000000000000000000e+00 0.000000000000000000e+00 0.000000000000000000e+00 0.000000000000000000e+00 0.000000000000000000e+00 0.000000000000000000e+00
0.000000000000000000e+00 0.000000000000000000e+00 0.000000000000000000e+00 0.000000000000000000e+00 0.000000000000000000e+00 0.000000000000000000e+00
0.000000000000000000e+00 0.000000000000000000e+00 0.000000000000000000e+00 0.000000000000000000e+00 0.000000000000000000e+00 0.000000000000000000e+00
0.000000000000000000e+00 0.000000000000000000e+00 0.000000000000000000e+00 0.000000000000000000e+00 0.000000000000000000e+00 0.000000000000000000e+00
0.000000000000000000e+00 1.000000000000000000e+00 0.000000000000000000e+00 0.000000000000000000e+00 0.000000000000000000e+00 0.000000000000000000e+00
0.000000000000000000e+00 0.000000000000000000e+00 0.000000000000000000e+00 0.000000000000000000e+00 0.000000000000000000e+00 0.000000000000000000e+00
0.000000000000000000e+00 0.000000000000000000e+00 0.000000000000000000e+00 0.000000000000000000e+00 0.000000000000000000e+00 0.000000000000000000e+00
0.000000000000000000e+00 0.000000000000000000e+00 0.000000000000000000e+00 0.000000000000000000e+00 0.000000000000000000e+00 0.000000000000000000e+00
0.000000000000000000e+00 0.000000000000000000e+00 0.000000000000000000e+00 0.000000000000000000e+00 0.000000000000000000e+00 1.000000000000000000e+00
0.000000000000000000e+00 0.000000000000000000e+00 0.000000000000000000e+00 0.000000000000000000e+00 0.000000000000000000e+00 0.000000000000000000e+00
```

Topic vector



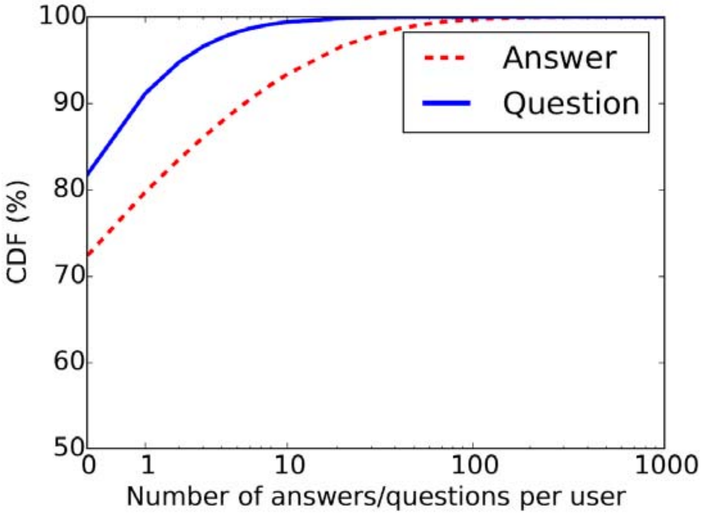
```
a-de-lan-dai-98,35,12,14,3,9,2,6,2,0,0,1,0
wo-he-wan-zi-tang,6,12,6,0,0,0,15,0,1,3,3,0
fen-nu-de-kai-wen,39,14,16,4,10,1,1,0,1,0,0,0
cao-lei-18-59,20,49,19,4,16,1,0,0,1,0,0,0
tang-yuan-75-7,18,10,8,3,3,8,0,3,3,0,0,4
zi-lai-juan-er-35,18,4,4,14,5,1,0,0,0,0,0,0
```

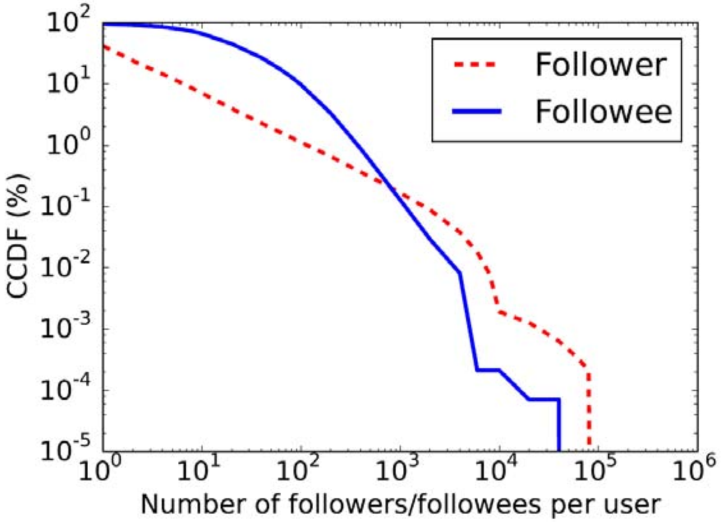
Initial analysis (1/2)



- Power law distribution
  - It has been identified in social science
  - It means that a small portion have extremely high degree while most have low degree

Initial analysis (2/2)





- CDF (Cumulative distribution function) & CCDF (Complementary Cumulative Distribution Function)
  - 81% of the users did not ask any question and 72% of users did not publish any answer
  - about 38% of users have no follower and more than 99% of users have followees

Methods on opinion leader identification

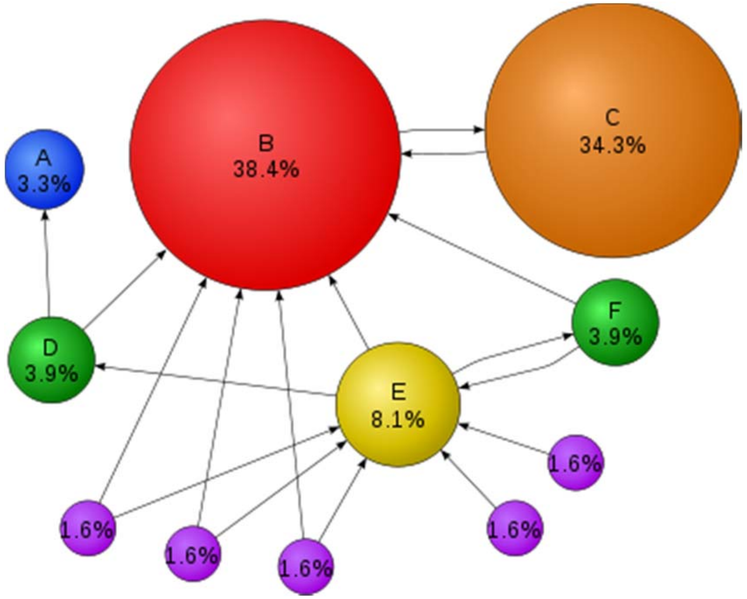
- Opinion leader identification is a ranking problem
  - PageRank (Larry Page), HITS (Jon Kleinberg)

- A typical ranking method – PageRank
  - An algorithm used by Google Search to rank websites in their search engine results
  - It works by counting the number and quality of links to a page to determine a rough estimate of how important the website is

PageRank

- used by Google Search to rank websites in their search engine results (Larry Page)

PageRank works by counting the number and quality of links to a page to determine a rough estimate of how important the website is. The underlying assumption is that more important websites are likely to receive more links from other websites



Original PageRank algorithm

- PR(A) = (1-d) + d (PR(T1)/C(T1) + ... + +PR(Tn)/C(Tn))
- Where:
  - PR(A) is the PageRank of page A
  - PR(Ti) is the PageRank of pages Ti which link to page A
  - C(Ti) is the number of outbound links on page Ti
  - d is a damping factor which can be set between 0 and 1

A simple example of PageRank

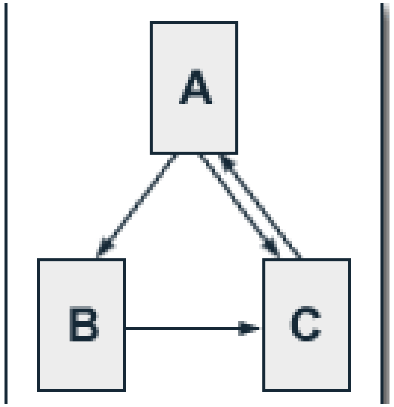
- We regard a small web consisting of three pages A, B and C
  - Page A links to the pages B and C, page B links to page C and page C links to page A
  - The damping factor d is usually set to 0.85, but to keep the calculation simple we set it to 0.5
- Calculating PageRank

PR(A) = 0.5 + 0.5 PR(C)

PR(B) = 0.5 + 0.5 (PR(A) / 2)

PR(C) = 0.5 + 0.5 (PR(A) / 2 + PR(B))

We get the following PageRank values for the single pages:



PR(A) = 14/13 = 1.07692308

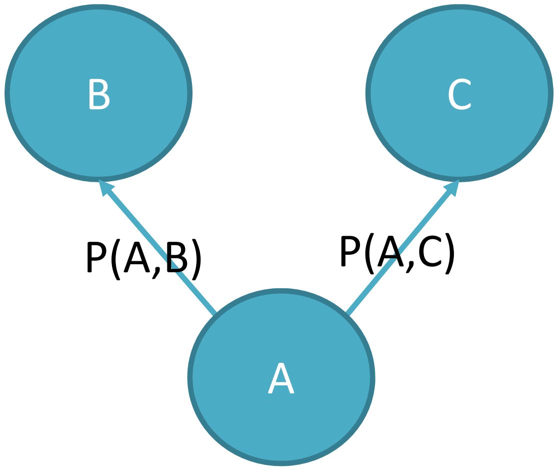
PR(B) = 10/13 = 0.76923077

PR(C) = 15/13 = 1.15384615

The sum of all pages' PageRanks is 3 and thus equals the total number of web pages

Topical opinion leader identification in Zhihu

- Our method considers three aspects
  - Social network structure
    - Based on PageRank
  - Topical interest similarity
    - A user's influence on each follower depends on the topic interest similarity between them
  - Knowledge authority
    - The higher authority (the number of votes) a user has, the higher probability he has to impact his followers



- Set link weight between users according to the topical interest similarity and knowledge authority

Results and evaluation

Topic	**Top 5 opinion leaders in each topic**
Movie	xiepanda, liuniandate, vikinglau, WxzxzW, chen-yao-39-75
Psychology	xiepanda, liuniandate, WxzxzW, zhang-xiao-wei-23, yezhuang
Travel	WxzxzW, chico-62, xu-wen-39, li-zhi-qiang-peter, qiu-shi-19-94
Food	xiepanda, anshi, weijiali, ji-li-ji-li, liuniandate
Fitness	WxzxzW, chico-62, xiepanda, summer.li, guo-fu-lin
Internet	xiepanda, WxzxzW, liuniandate, big_caaat, 8king
Fashion	WxzxzW, 8king, sickberry, liuniandate, xiepanda
Pioneer	wangxing, zhou-kui, xiepanda, liuniandate, dreamcog
Design	WxzxzW, 8king, xiepanda, soulchef, xiaoxiao
Finance	xiepanda, liuniandate, WxzxzW, ji-li-ji-li, big_caaat

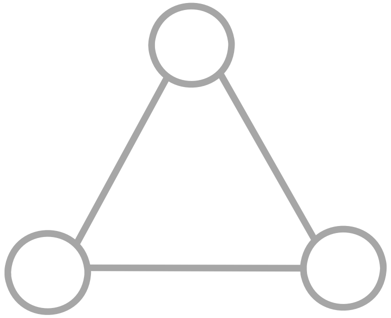
- Overall evaluation
  - They always published lots of topic-related posts and received a great number of votes, and have a great number of followers including some important followers

Topical opinion leader examples

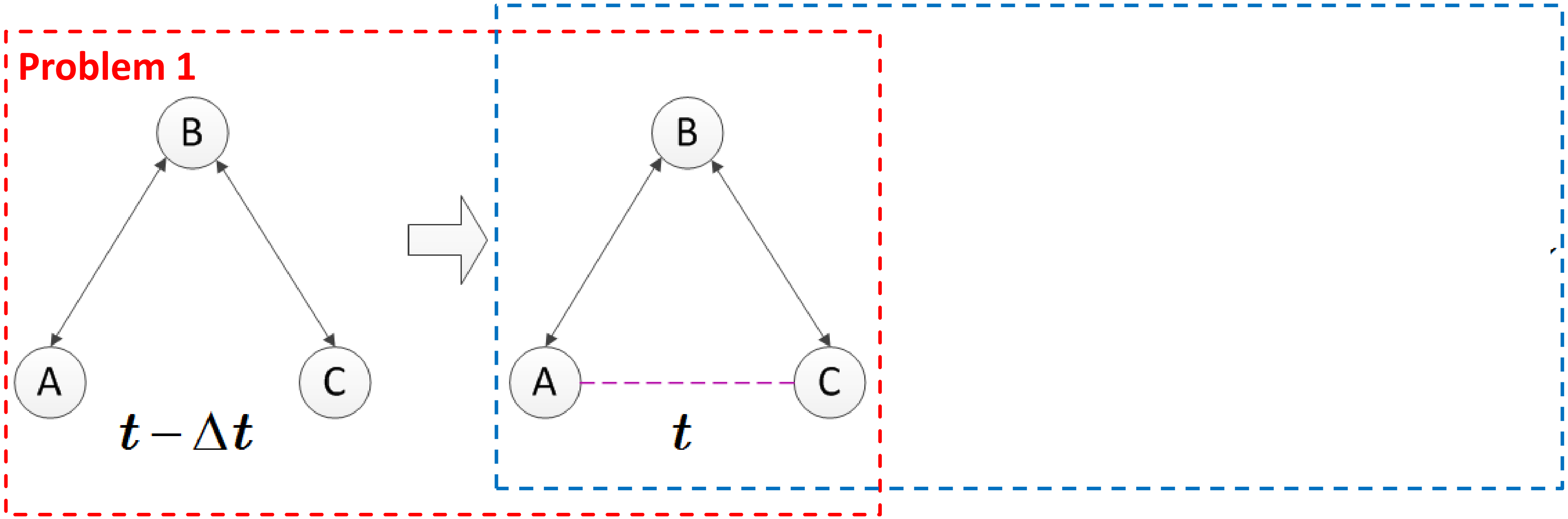
- “xiepanda”, “liuniandate"and “WxzxzW"are identified among the top 5 topical opinion leaders in most topics
  - This kind of users are so-called **cewebrity**, who often posts a large number of content about various topics to acquire fame on the Internet so that they have more than 200K followers including many important ones.
- “yezhuang" is identified as a opinion leader in psychology
  - He is a psychology trainer, whose posts are all related to psychology
  - His answers received 458 of average vote count. He is also followed by more than 40K users including some top-ranked ones
- Most of pioneer-related top 5 opinion leaders are successful company founders in real life
  - “wangxing” posted mainly about Pioneer and has 61,268 followers including a few of influential users “yuyue-51”, “zhou-kui”, and “GavinQi”. He founded many popular websites such as Meituan, Fanfou and Renren

Triadic Closure and Its Influence in Social Networks

– A Microscopic View of Network Structural Dynamics

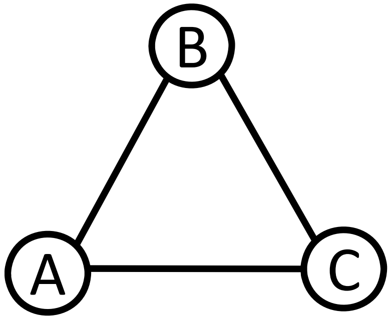


Two Issues in Triadic Closure Process

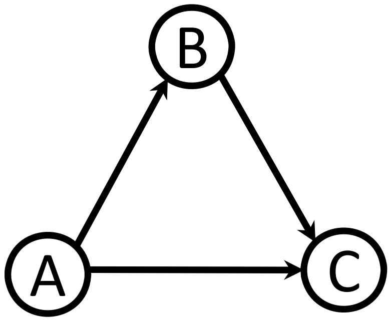


- Problems:
  - 1: will open triad be closed at time t?
  - 2: at $t+\Delta t$, will the tie AB become stronger or weaker? Also BC?

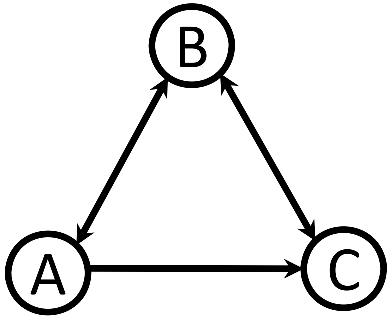
Related Work



Undirected [1]



Directed – Case 1 [2]



Directed – Case 2 [3]

[1] Zignani, M., et.al. Link and triadic closure delay: Temporal metrics for social network dynamics. In ICWSM'14.

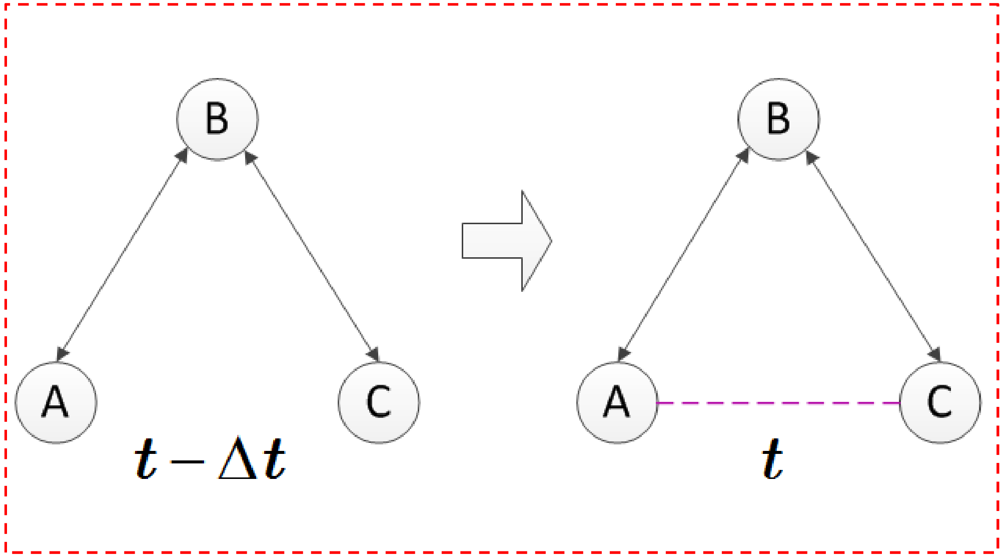
[2] Romero, D. M. and Kleinberg, J. The directed closure process in hybrid social-information networks, with an analysis of link formation on twitter. Stat, 2010.

[3] Lou, T et.al. Learning to predict reciprocity and triadic closure in social networks. TKDD, 2013.

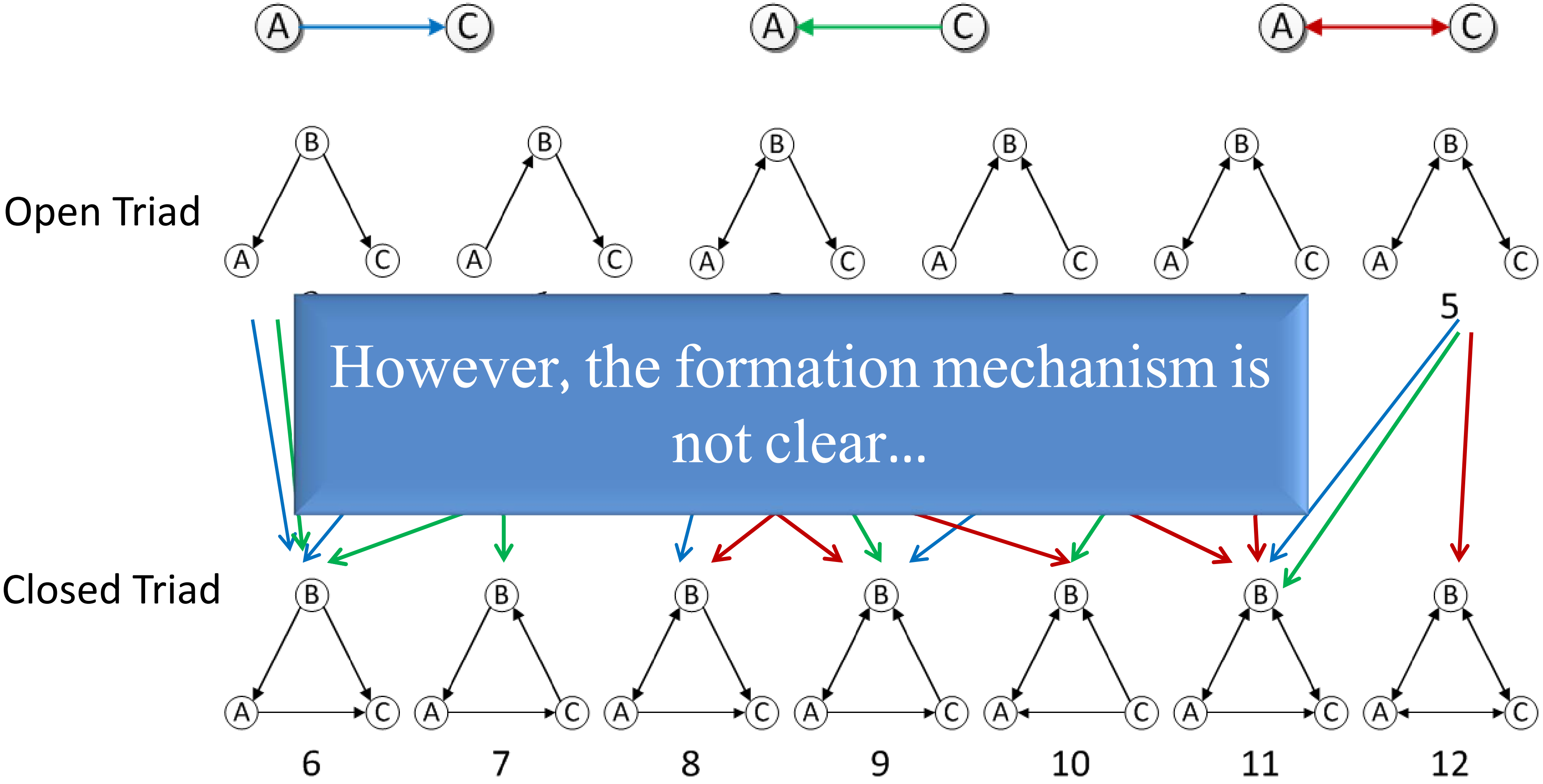
Triad: A group of three people

WHY TRIADS? TRIADIC CLOSURES?

What are underlying factors that trigger triadic closure?



Open Triad to Triadic Closure



However, the formation mechanism is not clear…

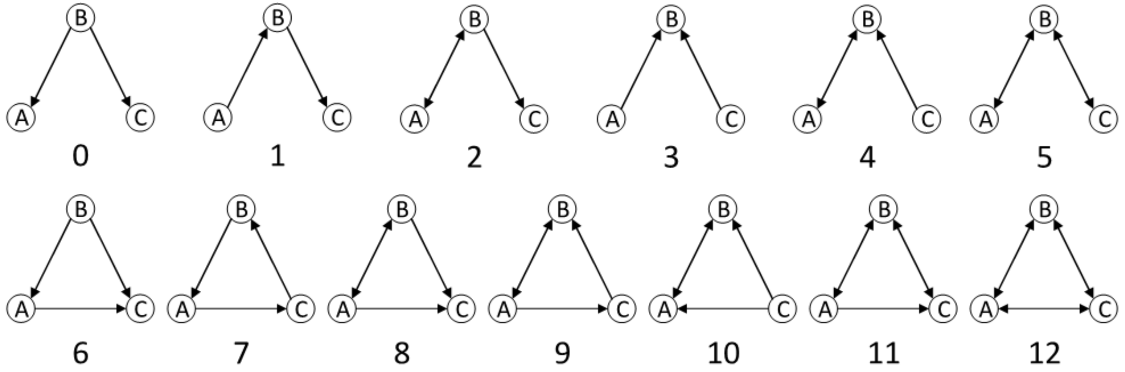
Milo R, Itzkovitz S, Kashtan N, et al.. Superfamilies of evolved and designed networks. Science, 2004

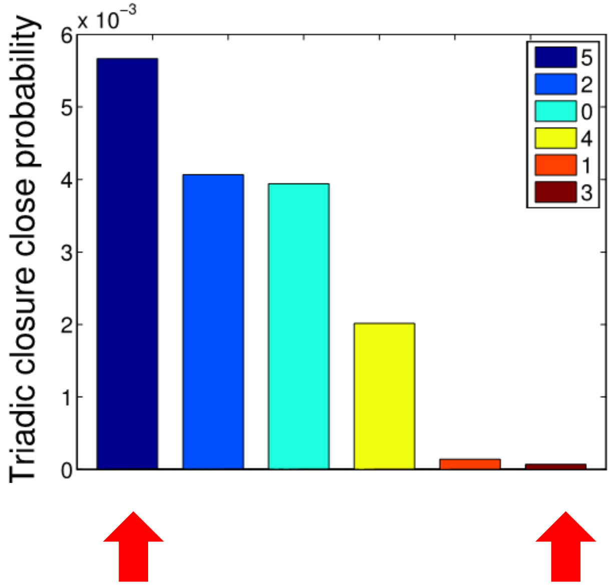
Look into one social network…

- Time span: Sep 29th, 2012 – Oct 29th, 2012
- 1.7 million nodes
- 400 million following links
- 200 out-degree per user
- 360 thousand new links
- 746 thousand newly formed closed triads per day

新浪微博
weibo.com

Observation - Network Topology

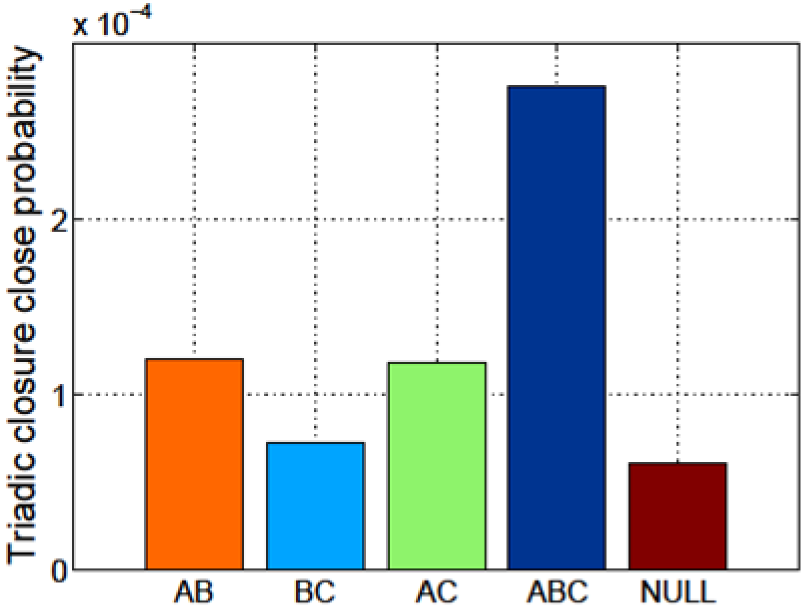




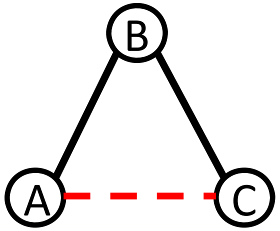
Y-axis: probability that each open triad forms triadic closures

Observation - Demography

Location

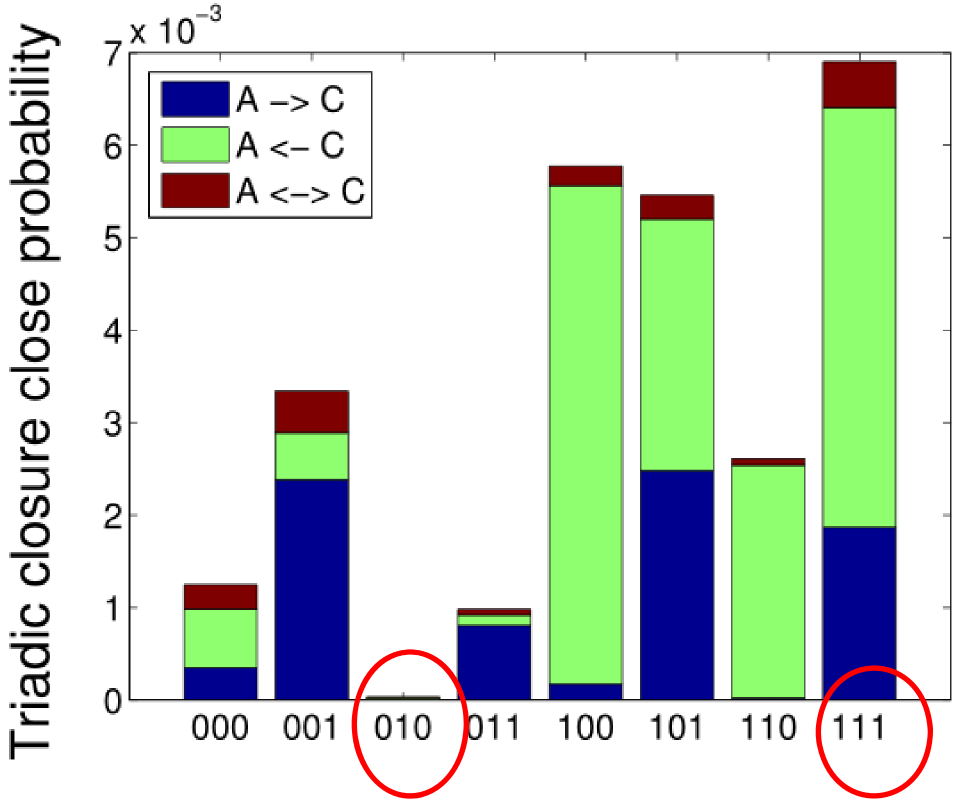


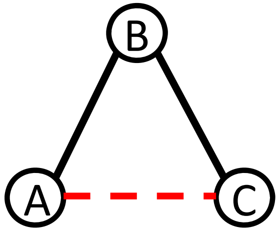
Gender



AB means A and B are from the same city

Observation – Opinion Leader

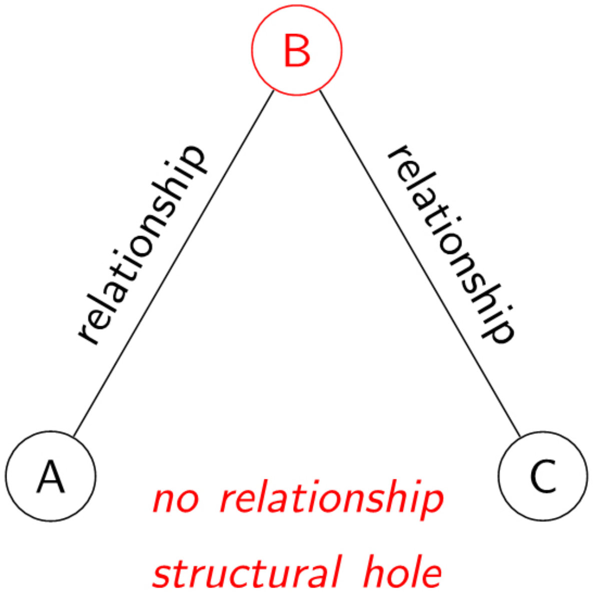


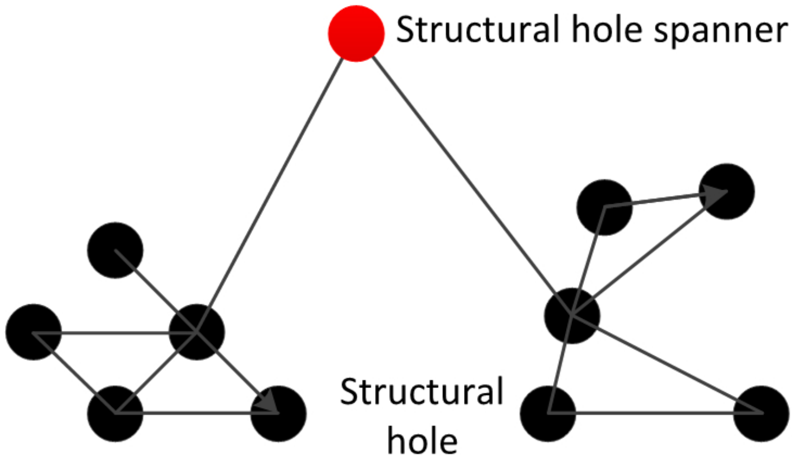


0—ordinary user
1—opinion leader

Structural Hole

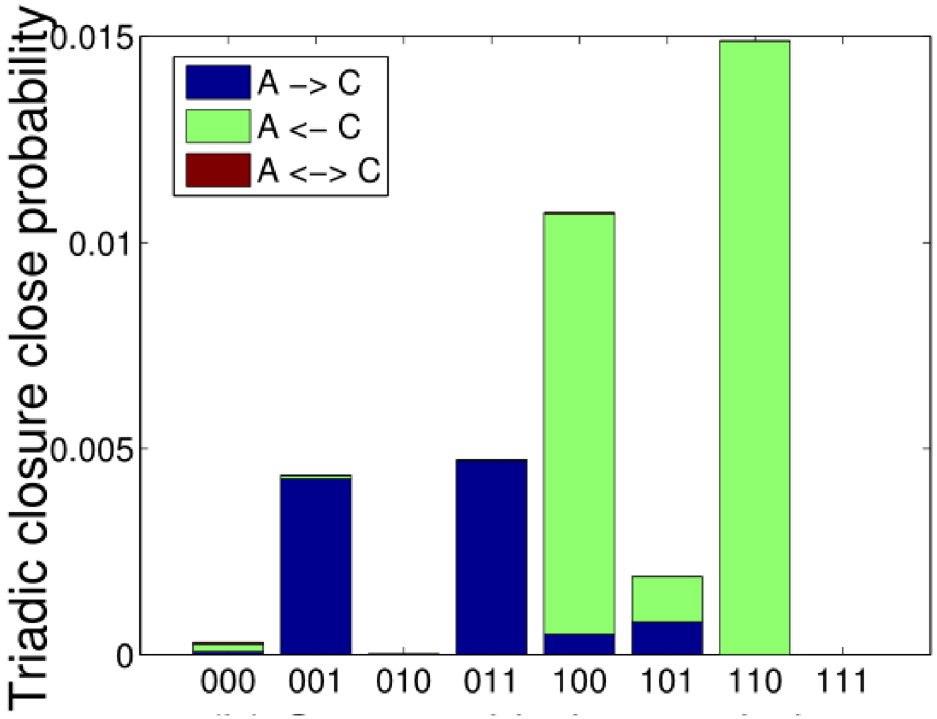
- **Structural hole:** two separate clusters possess non-redundant information.

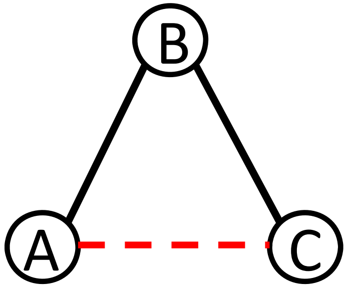




Lou T, Tang J. Mining structural hole spanners through information diffusion in social networks, www2013

Observation - Structural Hole

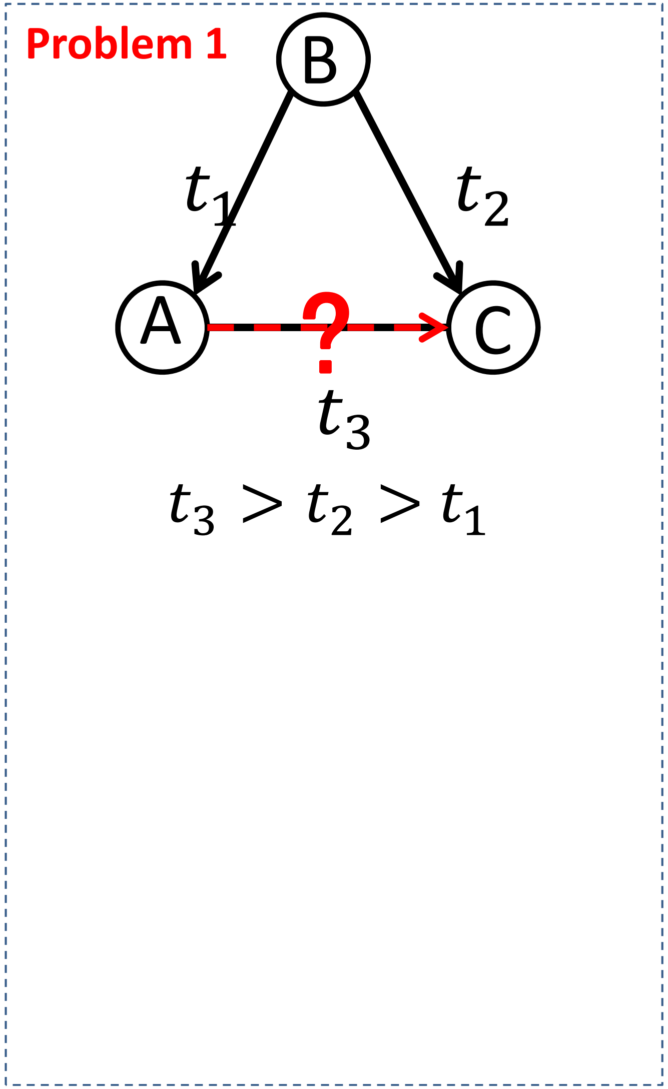




0—ordinary user

1—structural hole spanner

Problem Formalization



- Given a network $G^t = (V, E, X, Y)$,
  - X features defined for candidate triads
  - Y whether an open triad become closed or not (Problem 1)
- **Goal**: Predict the formation of triadic closure given learned features and the network

$$\varphi = logP(Y|X, G)$$

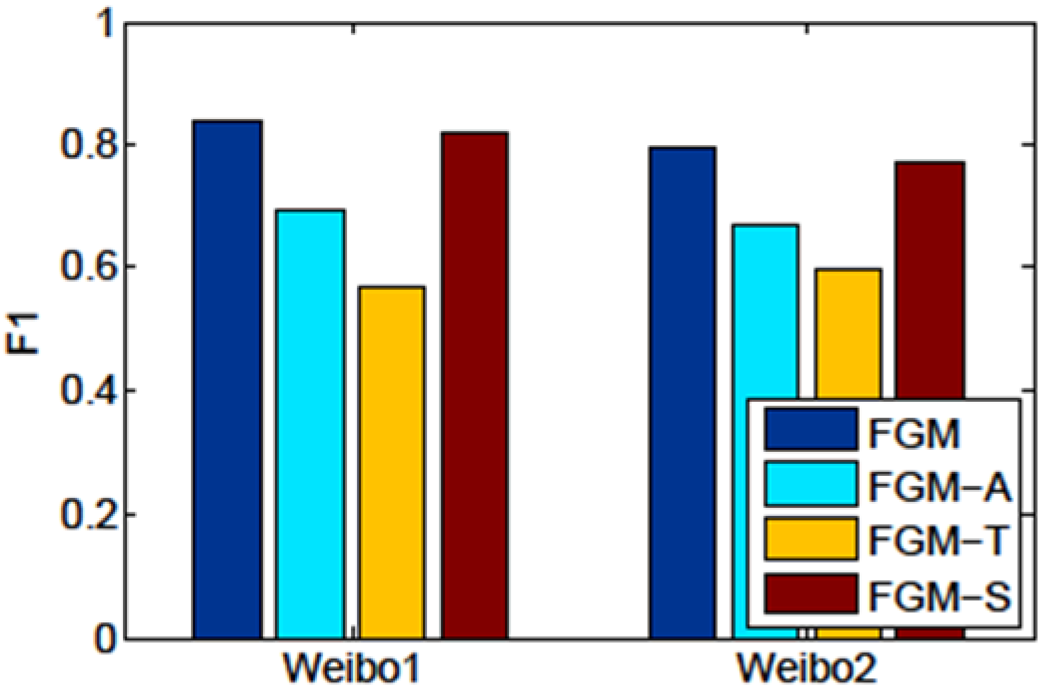
Experiments & Results

- Datasets: Weibo
- 50% as a training set, 50% as a testing set

Problem 1

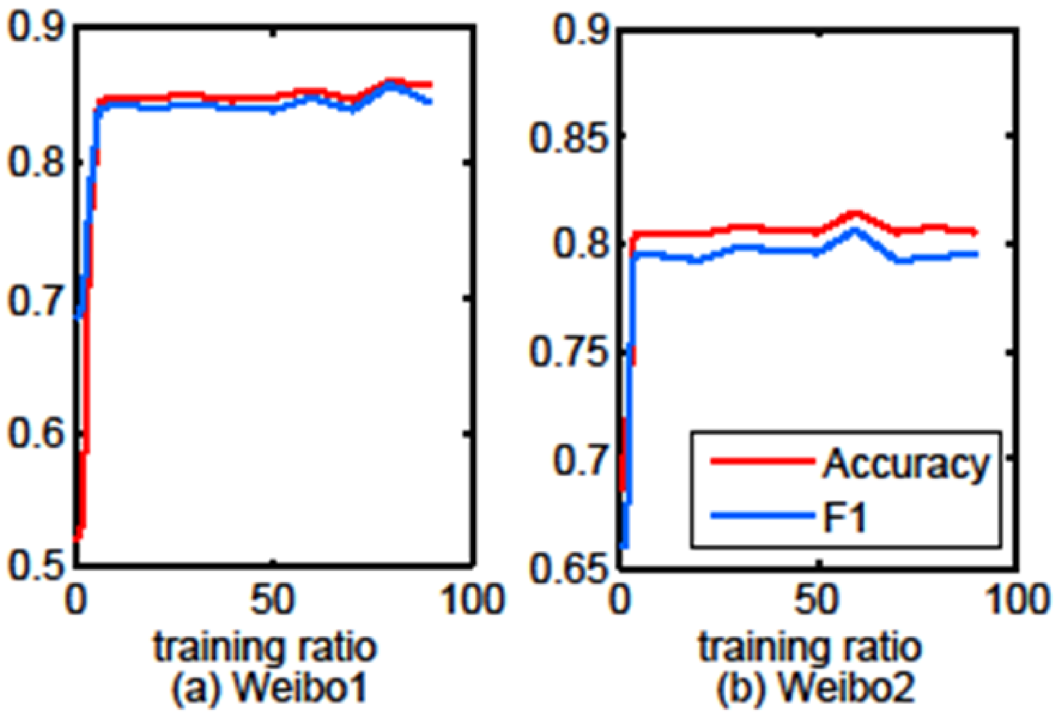
Algorithm	Precision	Recall	F1	Accuracy
SVM	0.890	0.844	0.866	0.882
Logistic	0.882	0.913	0.897	0.885
Our	**0.901**	**0.953**	**0.926**	**0.931**

Factor Contributions



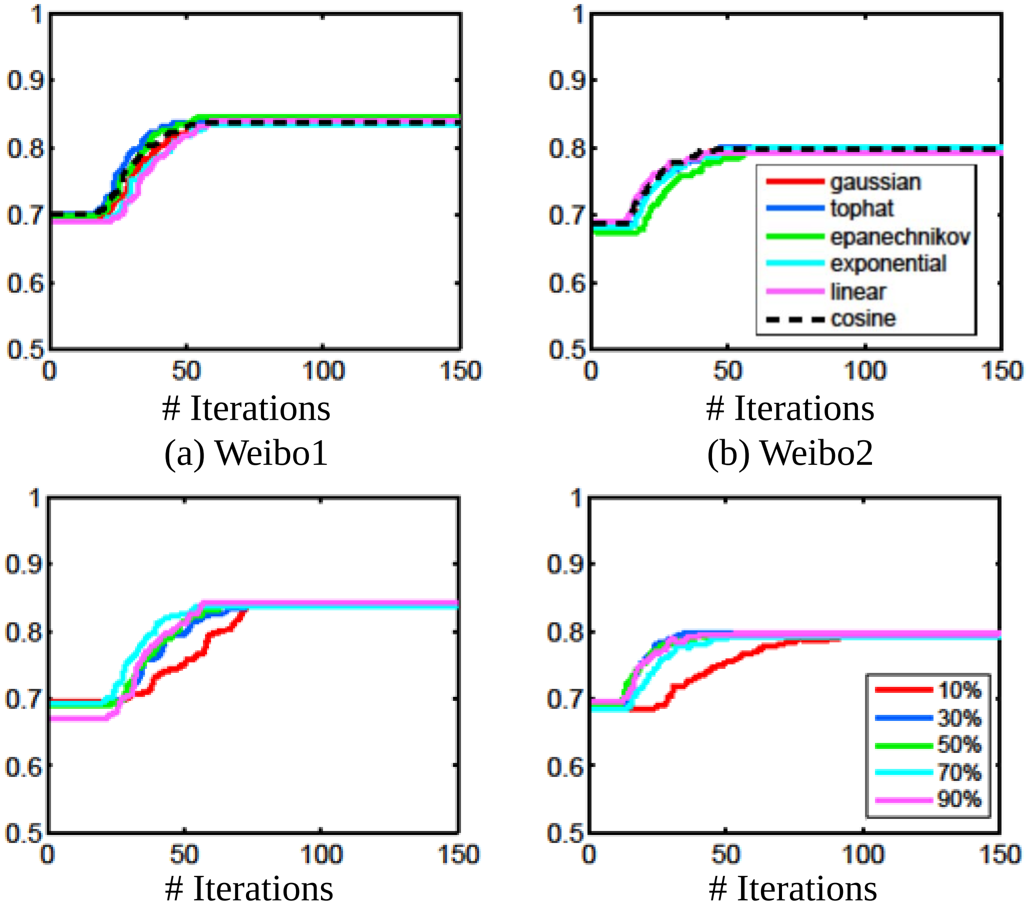
Temporal factor > Attribute factor > Social factor

Performance - Training Data Ratio



Robust to the size of training dataset

Performance - Convergence



(a) Weibo1

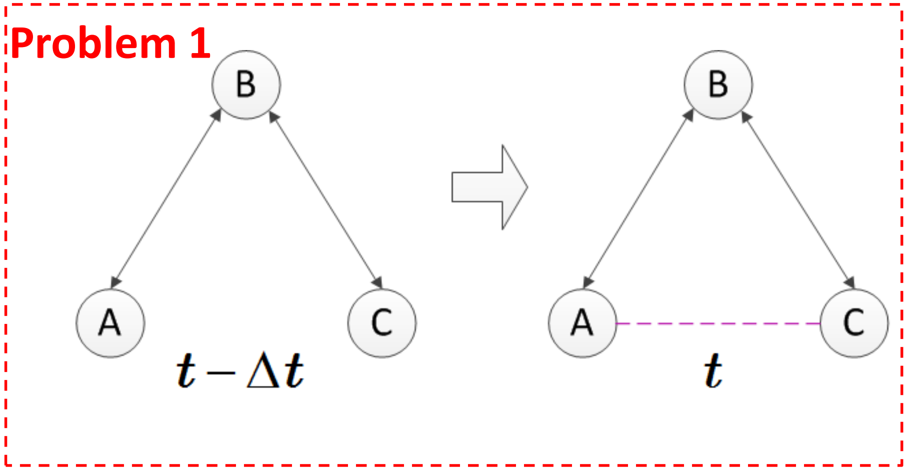
(b) Weibo2

(c) Weibo1

(d) Weibo2

# iterations < 50

Conclusion



- Studying triadic closure process
- Uncovering underlying factors that trigger triadic closure
- Proposing efficient models to predict triadic

Announcement

- Next class on 8th June will start at 10:30am and end at 12:00.