Demonstration & Course assignment

Advanced Topics in Mobile Communications (AToMIC): Social Network in Mobile Big Data

Summer Semester 2016

M.Sc. Tao Zhao

Ph.D. student

Outline

- Demonstration
- Course assignment

Demonstration

What can we do in social network?

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

Community question & answer (Q&A) sites

- What is community Q&A site (CQA)
  - 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
  - An innovative CQA that is more professional than traditional CQA sites and takes special efforts in improving content quality
  - Draw the participation of both a rapidly growing user population and specific domain experts
  - Another similar site: Quora

Main features of Zhihu

- Ask & answer questions
- Vote answers
- Follow users
- Follow topics & questions

New features


知乎
搜索你感兴趣的内容...
首页
话题
发现
消息
提问
Papst
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
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？
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⊙)
2 条评论
分享
邀请回答
举报
235 个回答
按投票排序
FanN，公众号 电影细节控（dianyingxijiekong）
喵呜的小圈、周末、h11 等人赞同
可能剧透 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

What can we do with the dataset of Zhihu

- Many interesting things can be done
  - Characterize user activity in Zhihu
    - Distribution analysis
    - Correlation analysis
  - Identify influential users (opinion leaders) in some specific topic

Data collection

- Gather Zhihu dataset
  - Collect a set of 105118 users in Zhihu through a web-based crawler from February 2016
- Each user data contains
  - Follower and followee lists
  - Answer and question information, containing the number of answers and questions, topic tags, the number of received votes and thanks

Choose 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

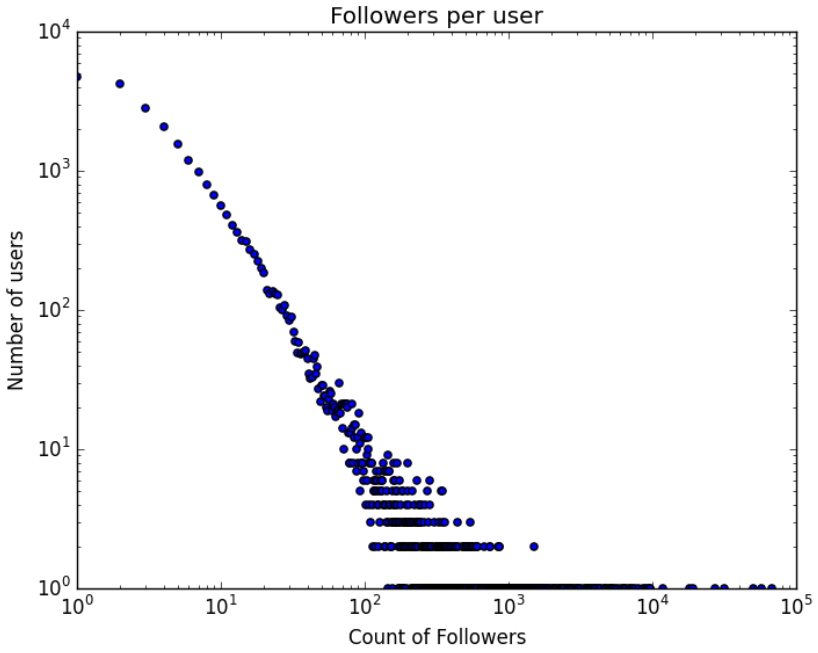
Basic statistics of dataset

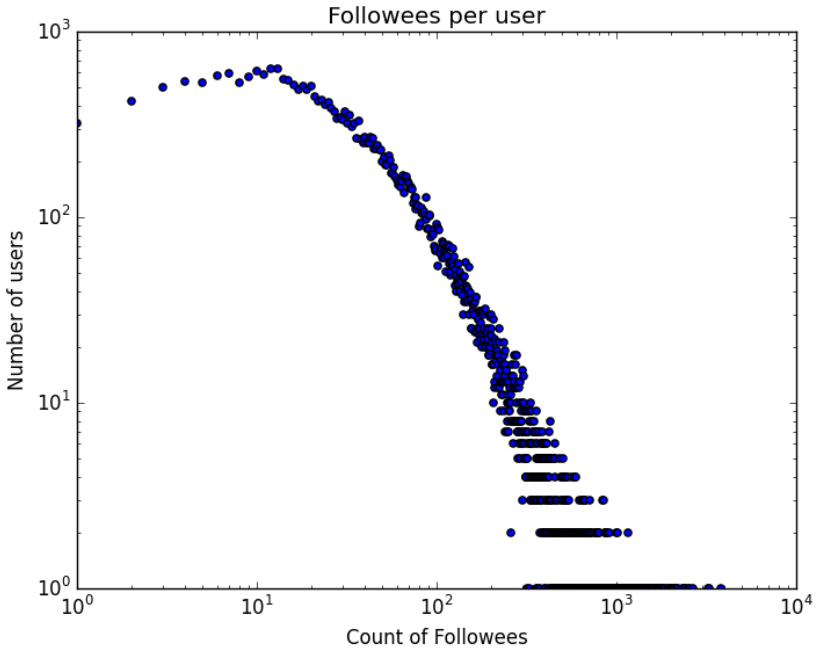
Number of total users	105118
Average number of followers per user	12.2
Average number of followees per user	44.2
Average number of questions per user	0.5
Average number of answers per user	2.9
Average number of votes per user	34.8
Average number of thanks per user	8.1

Distribution analysis

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

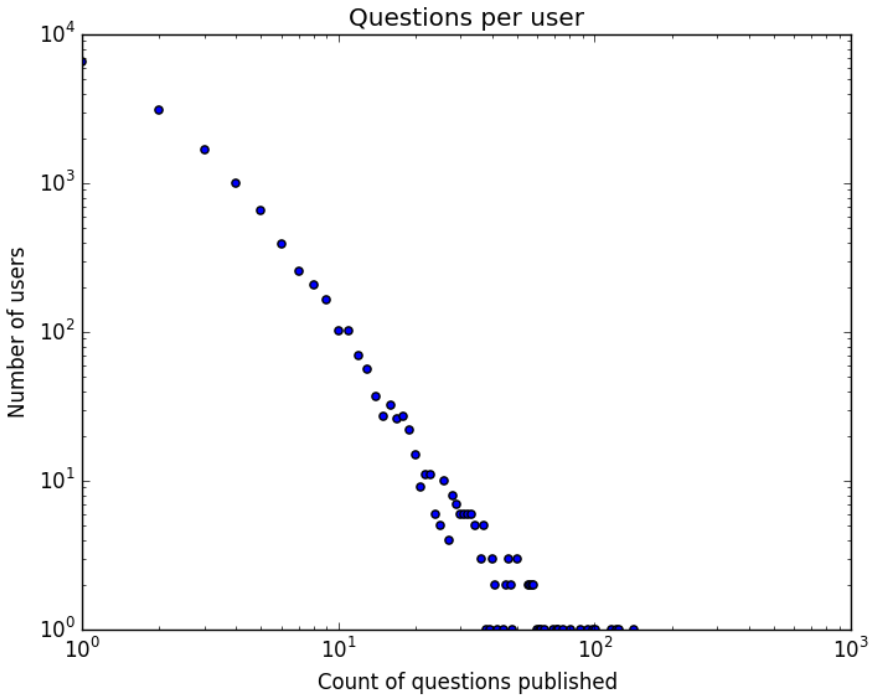
Distribution of follower & followee

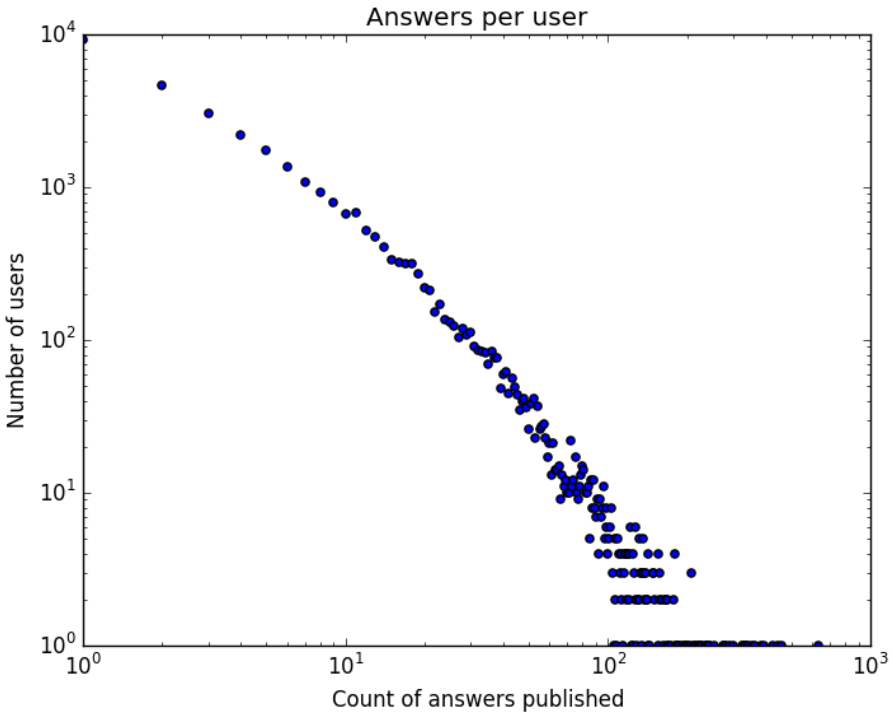




About 37% of users have no follower and 0.03% do not follow anyone, while 93% of users have less than 10 followers and 99.8% have less than 100 followees

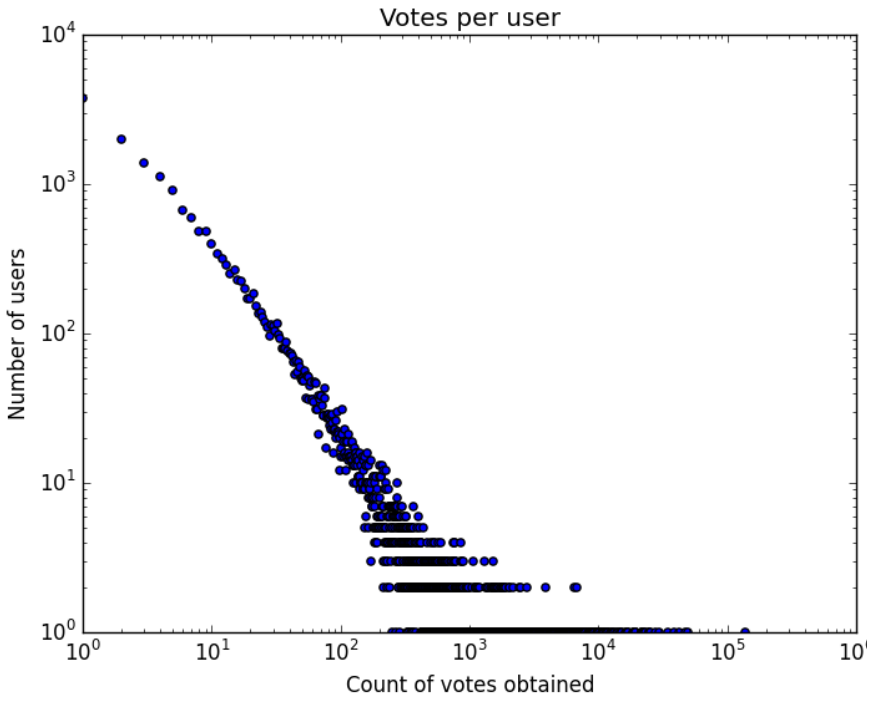
Distribution of question & answer

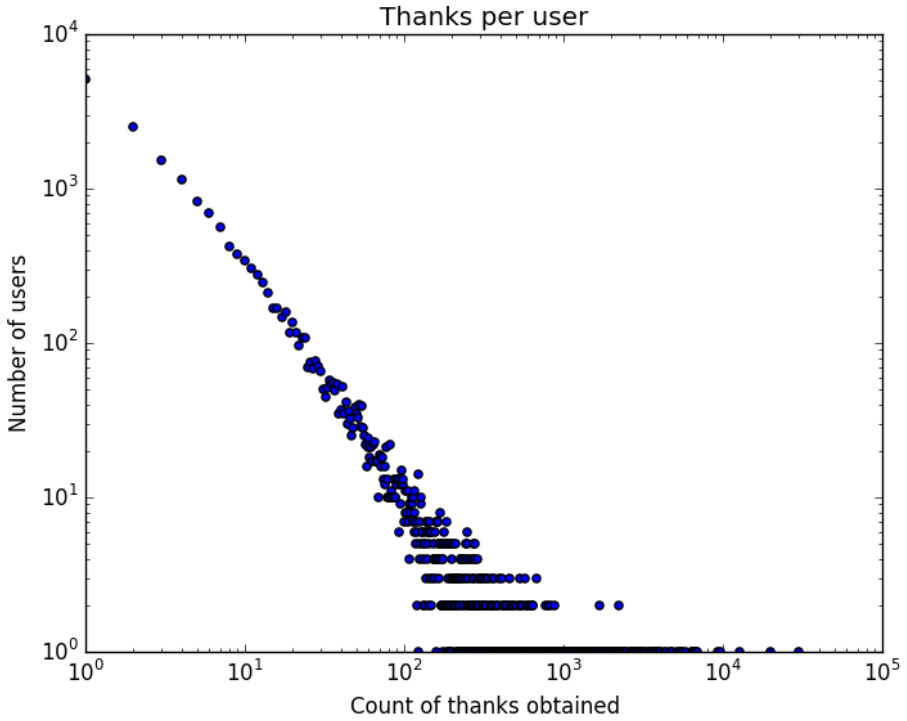




About 80% of users do not ask any questions and 69% of users give no answer, while 99% of users ask less than 10 questions and 93% of users give less than 10 answers

Distribution of vote & thank





About 82% of users do not get any thanks and 79% get no vote, while 95% of users get less than 10 thanks and 90% get less than 10 votes

All the distributions follow power law distribution

Correlation analysis

- Investigate which factors influence the number of followers per user
  - Leverage Pearson correlation coefficient to measure correlation between the number of followers per user and another factor (the number of answers, votes and questions)
    - Pearson product-moment correlation coefficient is a measure of the linear correlation between two variables
      - 0.00-0.19: “very weak”
      - 0.20-0.39: “weak”
      - 0.40-0.59: “moderate”
      - 0.60-0.79: “strong”
      - 0.80-1.0: “very strong”

Correlation analysis

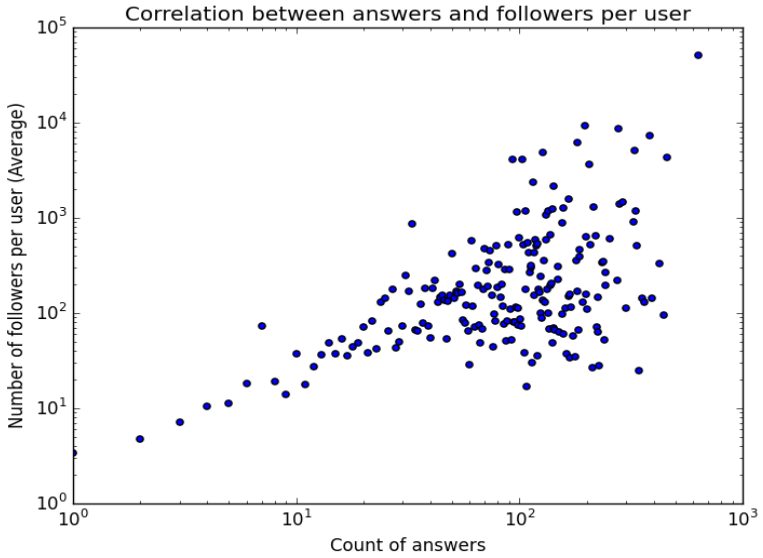


Fig. 1

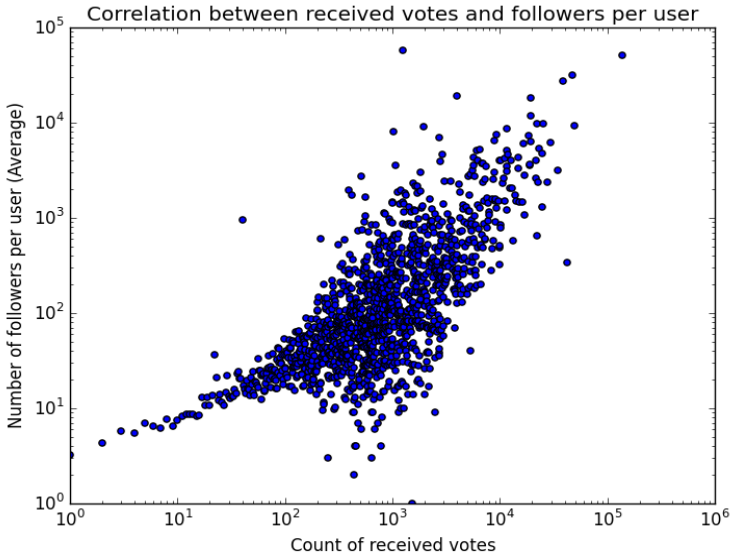


Fig. 2

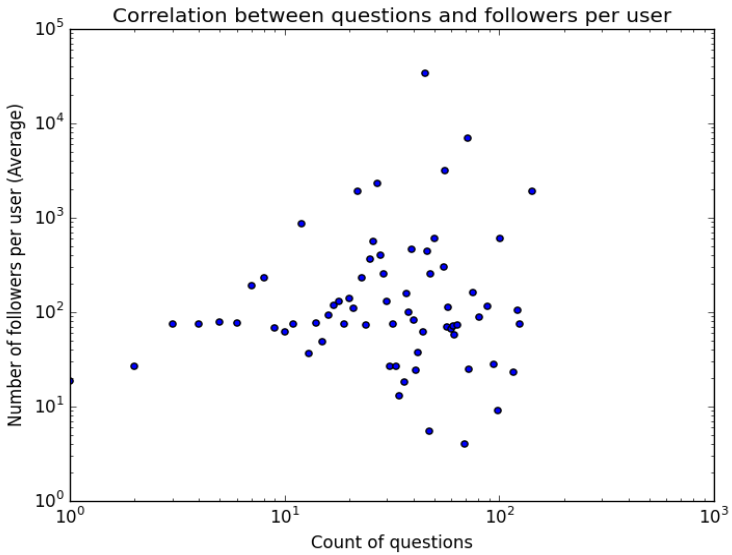


Fig. 3

Correlation analysis

	Pearson correlation coefficient
Count of answers	0.45 (moderate)
Count of votes	0.65 (strong)
Count of questions	0.05 (very weak)

- Preliminary conclusion
  - The majority of Zhihu users attract followers by contributing a large number of high-quality answers
  - Asking more questions cannot help attract more followers

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]
- Define 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

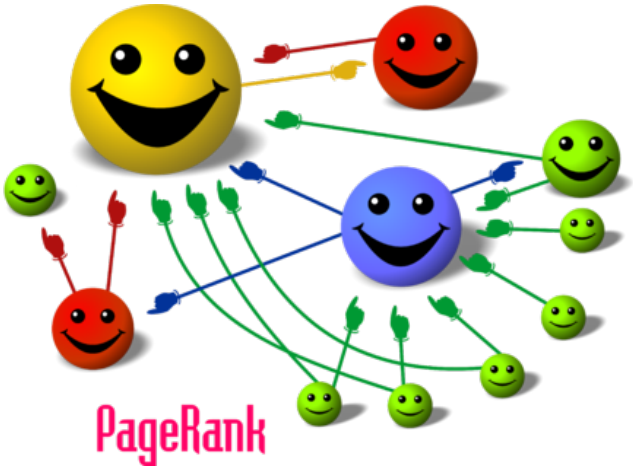
[1] Lazarsfeld, P.F., Berelson, B., Gaudet, H.: The People's Choice: How the Voter Makes up His Mind in a Presidential Campaign. New York: Columbia University Press, (1948)

Why to identify opinion leader

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

How to identify opinion leader

- Opinion leader identification is a ranking problem
  - PageRank, HITS
- Based on 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



How to identify opinion leader

- Extend PageRank through considering expertise authority in some specific topic
  - Form a directed graph with the users(nodes) and the "following" relationships among them
    - Choose a set of 8558 active users who publish at least 10 posts (Data reduction)
  - Calculate expertise authority in each topic for each node
    - Consider the number of answers and received votes for each user
  - Set link weight according to the expertise authority
    - The higher expertise authority the user has, the more important he is
  - Identify opinion leaders in each topic based on PageRank

Results

Topic #	Opinion leaders (Top 5)
0	anshi, chen-yao-39-75, a-xu-6, xiong-xiong-xiong-xiong-xiong-xiong-xiong-xiong, zhouyao
1	chen-yao-39-75, anshi, xiong-xiong-xiong-xiong-xiong-xiong-xiong, tan-hao-tommy, maji
2	chen-yao-39-75, anshi, polyhedron, ju-xuan-ya, yezhuang
3	xiong-xiong-xiong-xiong-xiong-xiong-xiong, maji, chen-yao-39-75, tan-hao-tommy, a-xu-6
4	chen-yao-39-75, a-xu-6, xiong-xiong-xiong-xiong-xiong-xiong-xiong, zhouyao, maji
5	chen-yao-39-75, baladi, yyss2037, yang-shuo, yuanxiafeel

Result analysis

- The results are reasonable
  - "chen-yao-39-75" is among the top-5 opinion leaders in all the six topics
    - He gives many answers and receives a large number of votes in each topic. He has the highest number of followers, including some influential ones like "tan-hao-tommy", "baladi"
  - "anshi" is identified as an opinion leader in topic 0, topic 1, and topic 2
    - He mostly answers about these three topics and gets more than 10000 votes in these topics. He has the second highest number of followers, including "a-xu-6" and "yezhuang"
  - "polyhedron" is an opinion leader in topic 2
    - He often answers about topic 2. He has much fewer followers than "chen-yao-39-75" and "anshi" but has many influential followers in topic 2, including "anshi" and "ju-xuan-ya"

Course assignment

Course assignment

- Each group choose one topic (due next Tuesday)
  - Group size = 2 students
- Each group will show a demo (June 25th) (20%)
- Each group will give a final presentation (first two Fridays of July) (40%)
  - Comprehensive survey + final experiment results
- Each group will submit a final report (end of September) (40%)

Topic	Description	Dataset
Influential user identification	The project is to identify influential users based on users' features	Twitter http://snap.stanford.edu/data/egonets-Twitter.html
Community detection	The project is to cluster different communities based on topics	Facebook http://snap.stanford.edu/data/egonets-Facebook.html
Point-of-Interest recommendation	The project is to make point-of-interest(POI) recommendation based on social influence and check-ins	Gowalla http://snap.stanford.edu/data/loc-gowalla.html
Link prediction and friend recommendation	The project is to make friend recommendation based on social networks and check-ins	Brightkite http://snap.stanford.edu/data/loc-brightkite.html
Analysis of individual activity and mobile pattern	The project is to give a detailed analysis of individual activity and mobile pattern based on everyday life tracks.	Social Evolution Dataset http://realitycommons.media.mit.edu/socialevolution4.html

Course link:
https://wiki.net.informatik.uni-goettingen.de/wiki/Advanced_Topics_in_Mobile_Communications_(AToMIC):_Social_Network_in_Mobile_Big_Data_(Summer_2016)

Topic	Literature
Influential user identification	[1] Guille A, Hacid H, Favre C, et al. Information diffusion in online social networks: A survey[J]. ACM SIGMOD Record, 2013, 42(2): 17-28. [2] Weng J, Lim E P, Jiang J, et al. Twitterrank: finding topic-sensitive influential twitterers[C]//Proceedings of the third ACM international conference on Web search and data mining. ACM, 2010: 261-270.
Community detection	[3] Xie J, Kelley S, Szymanski B K. Overlapping community detection in networks: The state-of-the-art and comparative study[J]. Acm computing surveys (csur), 2013, 45(4): 43. [4] Du N, Wu B, Pei X, et al. Community detection in large-scale social networks[C]//Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis. ACM, 2007: 16-25.
Point-of-Interest recommendation	[5] Wang H, Terrovitis M, Mamoulis N. Location recommendation in location-based social networks using user check-in data[C]//Proceedings of the 21st ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems. ACM, 2013: 374-383. [6] Yuan Q, Cong G, Ma Z, et al. Time-aware point-of-interest recommendation[C]//Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval. ACM, 2013: 363-372.
Link prediction and friend recommendation	[7] Al Hasan M, Zaki M J. A survey of link prediction in social networks[M]//Social network data analytics. Springer US, 2011: 243-275. [8] Scellato, Salvatore, Anastasios Noulas, and Cecilia Mascolo. "Exploiting place features in link prediction on location-based social networks." 17th ACM SIGKDD. 2011.
Analysis of individual activity and mobile pattern	[9] Li N, Chen G. Analysis of a location-based social network[C]//Computational Science and Engineering, 2009. CSE'09. International Conference on. IEEE, 2009, 4: 263-270. [10] Sun, Yeran, and Ming Li. "Investigation of Travel and Activity Patterns Using Location-based Social Network Data: A Case Study of Active Mobile Social Media Users." ISPRS International Journal of Geo-Information 4.3 (2015): 1512-1529.

Thank you!

Email: Tao.Zhao@cs.uni-goettingen.de