User identification and community exploration via mining big personal data in online platforms

Dissertation
for the award of the degree

Doctor of Philosophy (Ph.D.)
Division of Mathematics and Natural Sciences
of the Georg-August-Universität Göttingen

within the doctoral Program in Computer Science (PCS)
of the Georg-August University School of Science (GAUSS)

submitted by
Jiaquan Zhang

from Liaoning, China
Göttingen, 2021

Thesis Committee:

Prof. Dr. Xiaoming Fu
Institut für Informatik, Georg-August-Universität Göttingen

Prof. Dr. Stephan Herminghaus
Abteilung Dynamik komplexer Fluide, Max-Planck-Institut für Dynamik und Selbstorganisation

Members of the Examination Board:

Reviewer:
Prof. Dr. Xiaoming Fu
Institut für Informatik, Georg-August-Universität Göttingen

Second Reviewer:
Prof. Dr. Dieter Hogrefe
Institut für Informatik, Georg-August-Universität Göttingen

Further members of the Examination Board:
Prof. Dr. Carsten Damm
Institut für Informatik, Georg-August-Universität Göttingen

Prof. Dr. Lutz M. Kolbe
Fakultät für Wirtschaftswissenschaften, Georg-August-Universität Göttingen

Prof. Dr. Marcus Baum
Institut für Informatik, Georg-August-Universität Göttingen

Prof. Dr. Winfried Kurth
Institut für Informatik, Georg-August-Universität Göttingen

Date of the oral examination: 17. September 2021

Acknowledgement

"He said one day you will leave this world behind, so live a life you will remember". For four years, I was away from home and family, and the Covid-19 pandemic unexpectedly made things harder for all of us. The people I have met, the places I have visited, the papers I have read, the codes I have created, and every other thing that happened during those times are all memorable in this difficult but wonderful journey. I would like to thank all those people who helped me.

I would like to express my sincere gratitude to my supervisor, Prof. Dr. Xiaoming Fu, who gave me the opportunity of doctoral study, constantly encouraged me and guided me passing through many difficulties, and taught me how to do research. He not only delivers important characteristics of being a research to me, but also enlightens me of enjoying the life. His referrals created opportunities that I can collaborate with many excellent people from other institutions, where I learnt much about doing research from these senior researchers. Only with these helps could I finish my doctoral study and this thesis.

I also want to express my gratitude to my second supervisor, Prof. Dr. Stephan Herminghaus, who always delivered valuable suggestions for my researches, out of a busy schedule of directing the MPI of dynamics and self-organisation.

I would like to thank Prof. Dr. Dieter Hogrefe for reviewing my thesis. I also wish to thank Prof. Dr. Carsten Damm, Prof. Dr. Lutz M. Kolbe, Prof. Dr. Marcus Baum, and Prof. Dr. Winfried Kurth for serving as the examination board of my dissertation.

I wish to thank all my friends and colleagues in Computer Network Group and the institution of Computer Science, at the University of Goettingen, who supported and helped me in the past four years. In particular, I thank Tina Bockler, Heike Jachinke and Gunnar Krull for their help on technical issues and courses. I would thank Dr. David Koll, Dr. Tao Zhao, and Dr. Sakeer G Kulkarni for their kind guidance of getting me familiar with everything here. Many thanks to Prof. Jar-Der Luo and his team in Tsinghua University, with which I have a long time of collaboration in several projects.

I am also very gratitude to China Scholarship Council (CSC), and the European Union's Horizon 2020 research and innovation programme, for their financial support of my doctoral study.

Finally, I would like to thank my entire family for their unconditional support of my career for nearly ten years since I was away from home to enter the college. In particular, many thanks to my girlfriend, Ms Dan Lu, who also did four years of doctoral research in another country. We encouraged each other regularly and went through many difficult times, and we also acquire many fascinating experiences. Life is long, such four years of mutual supports will always motivate us.

Abstract

User-generated big data mining is vital important for large online platforms in terms of security, profits improvement, products recommendation and system management. Personal attributes recognition, user behavior prediction, user identification, and community detection are the most critical and interesting issues that remain as challenges in many real applications in terms of accuracy, efficiency and data security. For an online platform with tens of thousands of users, it is always vulnerable to malicious users who pose a threat to other innocent users and consume unnecessary resources, where accurate user identification is urgently required to prevent corresponding malicious attempts. Meanwhile, accurate prediction of user behavior will help large platforms provide satisfactory recommendations to users and efficiently allocate different amounts of resources to different users. In addition to individual identification, community exploration of large social networks that formed by online databases could also help managers gain knowledge of how a community evolves. And such large scale and diverse social networks can be used to validate network theories, which are previously developed from synthetic networks or small real networks. In this thesis, we study several specific cases to address some key challenges that remain in different types of large online platforms, such as user behavior prediction for cold-start users, privacy protection for user-generated data, and large scale and diverse social community analysis.

In the first case, as an emerging business, online education has attracted tens of thousands of users as it can provide diverse courses that can exactly satisfy whatever demands of the students. Due to the limitation of public-school systems, many students pursue private supplementary tutoring for improving their academic performance. Similar to online shopping platform, online education system is also a user-product based service, where users usually have to select and purchase the courses that meet their demands. It is important to construct a course recommendation and user behavior prediction system based on user attributes or user-generated data. Item recommendation in current online shopping systems is usually based on the interactions between users and products, since most of the personal attributes are unnecessary for online shopping services, and users often provide false information during registration. Therefore, it is not possible to recommend items based on personal attributes by exploiting the similarity of attributes among users, such as education level, age, school, gender, etc. Different from most online shopping platforms, online education platforms have access to a large number of credible personal attributes since accurate personal information is important in education service, and user behaviors could be predicted with just user attribute. Moreover, previous works on learning individual attributes are based primarily on panel survey data, which ensures

its credibility but lacks efficiency. Therefore, most works simply include hundreds or thousands of users in the study. With more than 200,000 anonymous K-12 students' 3-year learning data from one of the world's largest online extra-curricular education platforms, we uncover students' online learning behaviors and infer the impact of students' home location, family socioeconomic situation and attended school's reputation/rank on the students' private tutoring course participation and learning outcomes. Further analysis suggests that such impact may be largely attributed to the inequality of access to educational resources in different cities and the inequality in family socioeconomic status. Finally, we study the predictability of students' performance and behaviors using machine learning algorithms with different groups of features, showing students' online learning performance can be predicted based on personal attributes and user-generated data with MAE$< 10\%$.

As mentioned above, user attributes are usually fake information in most online platforms, and online platforms are usually vulnerable of malicious users. It is very important to identify the users or verify their attributes. Many researches have used user-generated mobile phone data (which includes sensitive information) to identify diverse user attributes, such as social economic status, ages, education level, professions, etc. Most of these approaches leverage original sensitive user data to build feature-rich models that take private information as input, such as exact locations, App usages and call detailed records. However, accessing users' mobile phone raw data may violate the more and more strict private data protection policies and regulations (e.g. GDPR). We observe that appropriate statistical methods can offer an effective means to eliminate private information and preserve personal characteristics, thus enabling the identification of the user attributes without privacy concern. Typically, identifying an unfamiliar caller's profession is important to protect citizens' personal safety and property. Due to limited data protection of various popular online services in some countries such as taxi hailing or takeouts ordering, many users nowadays encounter an increasing number of phone calls from strangers. The situation may be aggravated when criminals pretend to be such service delivery staff, bringing threats to the user individuals as well as the society. Additionally, more and more people suffer from excessive digital marketing and fraud phone calls because of personal information leakage. Therefore, a real time identification of unfamiliar caller is urgently needed. We explore the feasibility of user identification with privacy-preserved user-generated mobile, and we develop CPFinder, a system which implements automatic user identification callers on end devices. The system could mainly identify four categories of users: taxi drivers, delivery and takeouts staffs, telemarketers and fraudsters, and normal users (other professions). Our evaluation over an anonymized dataset of 1,282 users with a period of 3 months in Shanghai City shows that the CPFinder can achieve an accuracy of 75+% for multi-class classification and 92.35+% for binary classification.

In addition to the mining of personal attributes and behaviors, the community mining of a large group of people based on online big data also attracts lots of attention due to the accessibility of large-scale social network in online platforms. As one of the very important branches of social network, scientific collaboration network has been studied for decades as

online big publication databases are easy to access and many user attributes are available. Academic collaborations become regular and the connections among researchers become closer due to the prosperity of globalized academic communications. It has been found that many computer science conferences are closed communities in terms of the acceptance of newcomers' papers, especially are the well-regarded conferences [24]. However, an in-depth study on the difference in the closeness and structural features of different conferences and what caused these differences is still missing. Previous studies of coauthor networks did not adequately consider the central role of some authors in the publication venues, such as Program Committee (PC) chairs of the conferences. Such people could influence the evolutionary patterns of coauthor networks due to their authorities and trust for members to select accepted papers and their core positions in the community. Thus, in addition to the ratio of newcomers' papers it would be interesting if the PC chairs' relevant metrics could be quantified to measure the closure of a conference from the perspective of old authors' papers. Additionally, the analysis of the differences among different conferences in terms of the evolution of coauthor networks and degree of closeness may disclose the formation of closed communities. Therefore, we will introduce several different outcomes due to the various structural characteristics of several typical conferences. In this paper, using the DBLP dataset of computer science publications and a PC chair dataset, we show the evidence of the existence of strong and weak ties in coauthor networks and the PC chairs' influences are also confirmed to be related with the tie strength and network structural properties. Several PC chair relevant metrics based on coauthor networks are introduced to measure the closure and efficiency of a conference.

Contents

Chapter 1

Introduction

1.1 Personal data mining

Online personal big data mining and user data learning are key issues of managing a large online service platform. To efficiently allocate resources and identify malicious users, mining the correlations among the user attributes and user-generated data have been studied for decades [14, 7]. User attributes usually include intrinsic user profiles such as gender, age, personal interests, affiliations, as well as family economic status. While user-generated data usually include user activities during using apps or mobile devices such as interactions with items and geo-location information. For example, many online shopping systems will recommend items to consumers base on user-product interactions such as click, add to chart, buy, share, etc. And user Global Positioning System (GPS) information are occasionally used for predicting user actions or recommending trajectories, which is widely deployed in map applications and taxi-hailing services [3, 6]. Meanwhile, large online platforms usually suffer from malicious users who will bring threats and harasses to other normal users, and many robotic accounts will generate useless data and consume unnecessary computational resources. Therefore, identifying such abnormal users is critical for providing satisfactory services to the users, reducing cost expenses for the platform and preventing any form of malicious attempt. In addition to individual data mining, online user data also provide the possibility of mining dynamics of a community of people, where interactions and relations within a group of users worth exploration in case to learn the diverse evolutionary pattern in social communities. Knowing how a community evolves, and whether a community is reasonable and is developing in correct direction, is quite important for community managers because they can prevent the community from being controlled by a small group of people or being separated into fractured components. Mining large online personal and social big data could not only help companies create satisfactory services to costumers and detect abnormal users to protect the system, but also can provide visions of the dynamics of different types of communities.

Currently, personal data learning is mainly based on user-generated data to mine similarities between people with similar attributes or behaviors, where two different scenarios are usually

studied: 1) predicting user behavior based on user attributes; 2) identifying user attributes based on user-generated data. For the first scenario, user attributes are usually provided by users themselves and then directly used to predict user behavior, such as whether a user will buy a certain product or not. Most previous studies only consider raw user attributes without extending the exploration of additional information, such as retrieving an individual's spending power by crawling the house prices around a given address, and using commuting patterns between home and work locations to indicate a user's occupation. Collecting additional information usually improves the accuracy of predicting user behavior. In the second case of detecting malicious users, the data provided by the user is not credible because malicious accounts are usually registered with false information. The only available and trustworthy data is user-generated data, which usually contains the user's behavior and activities while using an application. Since user-generated data usually contains sensitive information about personal privacy, the processing of such data should comply with data protection regulations. Previous studies do not consider privacy issues and use raw plaintext data as input for identification, which may violate data protection regulations in some cases, especially when distributing data to the cloud. Therefore, using privacy-preserving data to identify user attributes is becoming increasingly important, and how to maximize the retention of valuable information while removing sensitive personal information is a key issue in solving the problem.

Most previous studies use user-generated data to learn user behavior and recommend items based on user-item interactions and similarities between users, while a few studies explore the possibility of using the user intrinsic attributes to predict user behaviors [83]. Actually, for most online platforms such as online shopping system and online social services, the user profiles provided by users are usually fake since user profiles are Non-essential in such platforms [111]. The few studies using reliable user profiles are mostly rely on panel survey data, where only hundreds of people are included [54]. Therefore, there is not enough credible data to allow the use of user profiles to learn about user attributes and user behavior in many existing online services. However, the emergence of online services for important applications such as health [117, 84] and education [110, 20, 54, 39] offers the possibility of collecting large amounts of user profile data. Users from these applications usually voluntarily provide credible personal data as accurate personal information is very important for acquiring high quality services. Meanwhile, credible personal information could be extended to many dimensions by adopting external data resource, which is infeasible for the previous online personal big data mining since there is a lack of untrustworthy information. As a result, mining large credible personal data to predict user behavior and user attributes is urgently required by many platforms.

In addition to mining user data for user behavior prediction, identifying user attributes and detecting anomalies is also a critical issue. Malicious users remain as serious threats for most online service platforms, and there are many researches attempt to detect abnormal users and behaviors base on user-generated data [29, 153, 102, 116]. Traditional methods only focus on how to improve the accuracy of malicious user detection and continuously adopting more and more user data into the model, less researches have considered reducing the data complexity

and protecting privacy of user-generated data. For mobile application, user-generated data typically contains user behavior within the platform and user locations recorded by devices, which usually include sensitive information of high privacy requirement. Especially for the studies that use mobile phone data, the data generated from mobile devices and collected by the telecommunication operators usually contains critical information of personal privacy, such as exact locations, specific apps usage and call records. The leakage of these categories of information could bring much threats for people since malicious attempts of using such sensitive personal information could easily trace a person's trajectory and precisely place advertisement to people during their use of specific apps. Therefore, many companies and institutions could only process such personal data within the server under highest security to prevent the leakage of sensitive personal information, which makes the whole system running inefficiently especially for the system serving users all over a country.

Besides mining online big personal data for individual behavior prediction and identification, online social big data mining is a big challenge for many social communities. In the study of mining online big social networks, many researches attempted to find the opinion leader in an unknown communities based on the social connections among the people or to find the optimized network structures by exploring different network structures [157, 131, 11]. However, most previous studies use anonymous data sets to construct the social networks since the data are mostly collected from social platforms, where the user data are usually under strict protections. Data collected from large online social platforms could enable exploring large scale of communities with tens and thousands of persons, but the network is only homogeneous as all the nodes are anonymous without any difference. There is a lack of investigation of some important roles in side a community. Therefore, most previous studies ignore some special roles in a community, whose impacts and influences actually are far more than those of others. In addition to global or local statistical metrics of a social collaboration network, structural statistics of some critical persons can be used to quantify the pros and cons of a network, such as openness and creativity. Fortunately, academic cooperation and publications provide a way of exploring large social networks that containing real name researchers. The co-author networks have been studied for years since the availability of large online publication platform such as DBLP for computer science. According to the DBLP statistics, currently there is an increase of nearly 50,000 publications every year in computer science society, and a total of 507,375 papers was published in the year 2020. Exploration of large online bibliography could not only help people learn the evolution pattern of large social communities, but also could analyze special authors as all the people can be identified with unique names without ambiguity.

This thesis focus on mining online big data from both personal level and community level, and it addresses several challenges of previous online data mining in specific scenarios and application by exploring the methodology of user behavior prediction, user identification and community structural pattern identification. Although, many previous studies attempt to continuously improve the accuracy of either user behavior prediction and user profile identification, but some challenges such as adopting trustworthy personal profiles and mining their relations

with user behavior, leveraging privacy-preserved user-generated data in user identification, and mining big evolutionary pattern of online large social collaboration communities. Specifically, the thesis studies three specific cases to address the above issues, respectively:

- Predict user behavior and user performance based on both user-generated data and trustworthy user profiles for online education platform. This work aims at mining the multi-dimension relations between student performance, courses participation and their personal profiles such as family status, school levels, grades, city levels, etc.
- Identify user professions based on privacy-preserved mobile phone data. This work aims at using statistical methods to eliminate sensitive information, such as exact coordinates, app preferences and CDR, from user-generated mobile phone data. Then leveraging privacy excluded data to identify user professions to protect users from malicious attempts.
- Identify evolutionary community patterns of academic collaboration networks in computer science society. This work aims at mining the different structural properties among different levels of conferences to find out the underlying evolutionary mechanism that cause the diversities between top level conferences and normal level conferences.

1.2 Challenges of data mining in big online platforms

1.2.1 User-level mining

For large online platforms, user-level data mining is very important for managing a highly efficient operating system and improving user experiences. Currently, most online platforms are focusing on two data mining issues: identify user behaviors and user identities. Accurate user behavior identification is usually used for product recommendation and for resources prediction, which leverages user historical behavioral records to predict next action of a user. For example, the key competence of online shopping platforms is recommendation algorithms, which could accurately identify user interests and recommend high probability purchase items to users. In another case, user identities identification is used to target abnormal users, including malicious users posing threat to other normal users and robotic account consuming unnecessary computational resources. Accurate identification of abnormal users could help reduce unnecessary expense for platforms and prevent normal users from malicious attempts. However, most platforms are suffering from absorbing a large number of new users, who have no historic records at all, in which case many user behavioral data prediction methods are not applicable as they are usually based on sufficient user historic records. These new users are often referred to as cold-start users, and users with a small number of records are also currently referred to as cold-start users in most studies, which attempt to provide satisfactory behavioral predictions with few records. While for predicting the behavior of pure cold-start users, existing studies try to divide people into groups based on their social connections. This

type of approach makes use of the assumption that a group of friends usually have similar interests and behave in a somewhat similar way when using online services. Currently, such approaches have not attain satisfying performances in terms of accuracy [114]. Besides user behavioral data, another data resource is user provided personal profile data including genders, ages, school, home address, professions etc. Such personal attributes can build up demographic information about the user and then be used to predict the user's behavior, where we believe in the assumption that people with similar demographic information are likely to behave in similar ways. However, since such personal profiles are not necessary in most online platforms and most users want their personal data not to be exposed, user profiles provided by user themselves are usually fake. As a result, the personal-level analysis between user behavioral data and user profile data is missing for most online big data mining.

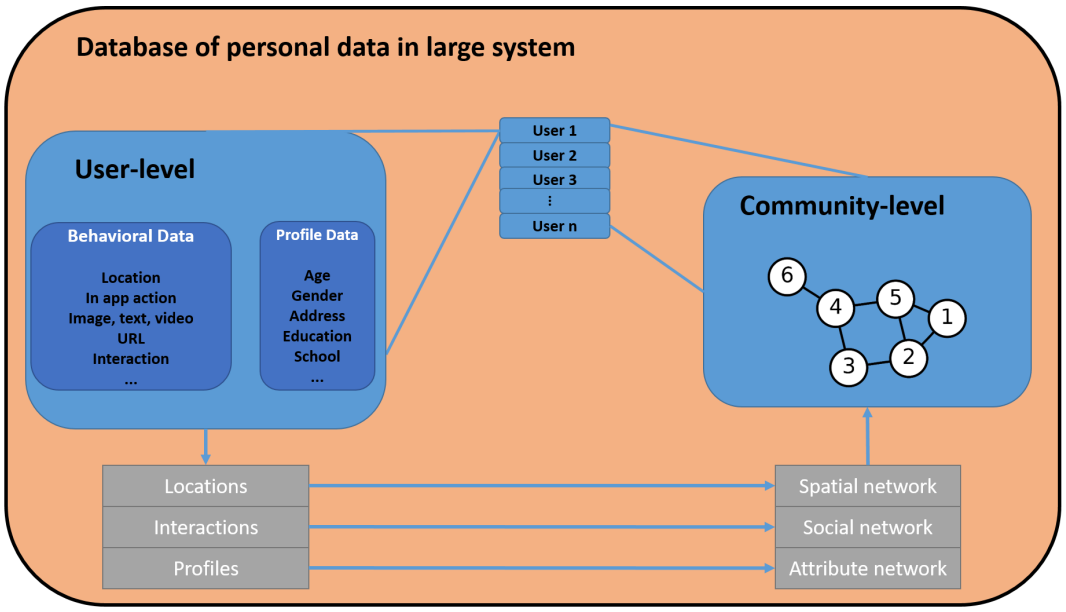


Figure 1.1: Databases of large personal data.

For the identification of anomalous users, existing approaches typically use machine learning methods based on rich features. These methods try to collect as much data as possible to build individual behavior patterns and work only for binary identification. In particular, cell phone caller identification is one of the most popular areas. However, the growth of online services based on mobile phones make people frequently receive calls from strangers such as taxi drivers and delivery men. This extends the threat from only fraudsters or telemarketers to criminals who will pretend to be cab drivers and delivery men, at which point the victim is in danger if they believe these false identities. As a result, the identification of mobile users should be extended to multiclass problems including different professions. Meanwhile, these feature-rich approaches require a large amount of data for each user and require large computational resources. In addition, user behavior data often contains sensitive personal information, such as location and application usage, the leakage of which would pose a serious risk to users. As a result, eliminating privacy and reducing data dimensions from mobile phone data while keep the valuable information for profession identification, is very important. The use of

privacy-preserving data to identify the occupations of multiple users can primarily help users identify malicious callers who provide false identities, which can then reduce the need for data storage considering data volume and data security.

1.2.2 Community-level mining

In addition to user-level data mining, community-level data mining is also an important issue of managing large online platforms [142, 2]. Understanding the structural patterns of large communities can help large system managers control the evolutionary direction of their communities and prevent them from being dominated by small local groups. Online social networks can be constructed by both user behavioral data and profile data, in terms of the type of connections. As shown in Figure. 1.1, there are various types of social networks formed by different types of connection. Typically, social networks are usually based on user interactions such as followers and followees, collaborations, and information exchange. Ties in spatial networks are formed by considering the visiting patterns of a pair of people. If two people frequently visit the same place at the same time, these two people may be connected with a high probability. While for attribute network based on user profile data, people in different units are connected if they preserve very similar pattern of demographic information. Existing data mining of community-level analysis typically use social connections as supplementary data resources for predicting user-level behaviors or overall trends of the entire community. None of the existing work compares different social network structures and explores the differences in group behavior between good or bad communities. In particular, scientific collaboration network based on co-authorship is one of the most important online social network, and naturally different journals or conferences form their own communities. Especially for the area of computer science, many ranking lists divide conferences into different levels according to their paper qualities. The studies of comparisons among different levels of conferences are missing, where the evolutionary patterns of different conferences, which maybe related to the cause of conference levels, is also lack of exploration.

1.3 Motivation

In this section, we list the the motivations of three works on online user big data mining in the dissertation.

1.3.1 Identify user behaviors in online education platforms

For large platforms with tens of millions or even millions of users, the critical issue of accurately predicting user needs and user behavior remains a key challenge. Especially for new users who do not have any historical information and only some untrustworthy personal profiles, predicting the behavior and needs of these new users is very difficult. This is because

most previous user data mining was mainly based on user-generated data and was based on the assumption that people who bought the same items might be interested in other similar items. These methods are not feasible for mining cold-start user data because they have no record of interaction with any items. To address the problem of predicting the behavior of cold-start users and their needs, some work [83, 71, 81] attempts to explore the relationship between user profiles and user-generated data. In contrast to mining the interaction between user-generated data and items, mining the relationship between user profiles and their interest in products does not rely exclusively on user-generated data. It aims to explore the possibility that people with very similar characteristics may have very similar behaviors and needs on online platforms. For example, people from residential areas with similar housing prices may have very similar spending levels, and people of the same age may have similar interests and product tastes. Therefore, by simply knowing as many user profiles as possible and constructing user demographics, it is possible to predict user behavior and needs by finding users whose demographics and historical interactions with the product are very similar.

However, people are often reluctant to provide the correct personal profile on online platforms because individuals have concerns about the security of their personal information. For example, users only need to provide their address to enjoy all the features of online shopping platforms, while for public social media, users do not actually need to provide any information but only an email for registration. Therefore, there is a lack of exploration on how to build user demographic information and use it to predict the behavior of cold-start users. Recently, new online services such as online education have grown significantly worldwide, especially after the explosion of Covid-19 its market size has almost doubled. The massive influx of students into online education platforms has created a huge challenge for online education providers in terms of human resources and hardware management. Faced with the large number of new users and the uncertainty of their needs, accurately predicting how many students will participate courses and how well they will perform is crucial for platforms to recruit enough teachers and tutors to deliver satisfactory course content and to prepare enough hardware resources to ensure good video streaming. Otherwise, insufficient human resources and hardware resources can lead to low quality courses and low resolution videos, which can make the platform lose many potential users. And excessive recruitment and hardware purchase may cause unnecessary expenses, which may reduce a lot of profiles.

In Chapter 3, a novel method of user behavior prediction on large online education platforms is introduced, which takes account the mining of the relations between user personal profiles and their performances, their participation in online courses.

1.3.2 Identify user professions based on privacy-preserved mobile phone data

Large web platforms often have malicious users who pose a threat to other normal users and platforms, and they consume unnecessary computational resources and generate much

useless data. To identify anomalous and bot accounts, many previous studies have leverage user-generated data and deploy machine learning methods in mining the differences between normal and malicious users. Typically, user-generated data contains diverse user activities and user interests during using an app or mobile devices, such as their interactions with items, their mobility patterns, their followings of other user accounts and some other temporal records [58, 140, 145]. Especially when considering the user data generated from mobile phone, we can build feature-rich demographic information for a user because the data collected from mobile phone contains much more useful information than an single app. Many previous researches [103, 97, 63] have proved that mobile phone usage is highly related to people's emotions and professions, that can be predicted or identified by mining user-generated mobile phone data. Meanwhile, there are some works try to detect fraud and malicious phone numbers by using mobile phone data including Call Detailed Record (CDR) data, network requests and application usage [115, 144].

However, serious user data leakage occasionally happened in the past several years. In 2019, data from 533 million people in 106 countries was stolen by hackers and was then published on a hacking forum. In late 2018, Google engineers discovered a software leak in the Google+ API used in the social media network. A total of 5 million users data was compromised, which caused a widely criticize of the level of consumer privacy within Google+ and the service was shutdown accordingly in the next year. Meanwhile, personal privacy and data protection regulations have been becoming more and more strict, where big personal data collection, sensitive personal data mining, data transmission and data storage should be well considered. Most previous studies that deal with big mobile phone data simply use the original form of the data, such as exact coordinates and their extends by exploring Point of Interest (POI) information around them, detailed call records about in and out calls of a personal number, and detailed apps usages records about when and how an app interact with users. Since these original form of sensitive personal data actually can be used to trace where users are and what users do in certain time period, it is very hazardous if such original data is intercepted by people with malicious attempts. Therefore, it is a big challenge for online platforms to secure user data especially when user-generated data are used for mining user behavior and user identification. If users could be identified based on privacy-preserved data or privacy-preserved data could achieve relative similar accuracy in user identification compared with original data, online platforms do not actually need to retain the original user data and the privacy-preserved data could even been distributed to the cloud and end devices for real time application.

In Chapter 4, a novel method of mobile phone user identification based on privacy-preserved data is introduced by a case study on identifying user professions. The methods actually address the challenges about how to deal with numerous personal data on large platforms, how to accurately and quickly identify a caller's profession, and how to use privacy-preserved data in mining user profiles.

1.3.3 Mining community difference based on collaboration networks

Online platforms not only need deep mining of personal data for user identification and user behavior prediction, but also they create another important research field of social network analysis. With the help of online social platforms, large scale of social interaction and collaboration networks are available by using user-generated data. Such online big data based large social networks could be used to valid several classical theorems of complex network [75, 67], which were proposed many years before and were only verified to be correct on small real social network. Large online social networks could also be used to detect community [3, 123, 25, 159], to find valuable and important person inside a large society [12, 40], to predict the link among people [105, 42, 113], as well as to predict some potential customers by exploring how the tie strength could help attract existing users' acquaintances.

Large online databases of scholarly publications offer the possibility of studying large networks of social collaboration, and studying strict temporal evolutionary patterns of such networks. As an online open source data, the Scientific Collaboration Network (SCN) also enables people to explore more information about each person as his/her name is necessary in the dataset, whereas on most previous online social platforms, people's information is anonymous for security reasons. As a result, SCN not only represent a classic type of large social network, but also create new dimensions of the measurement of social network by taking some personal information into consideration. Currently, the exploration of using personal information in large social network is still missing, and studying evolutionary patterns of large scale social networks is also missing due to the lack of temporal records of data.

The conferences of computer science strictly take place every year or every two years, which naturally show a temporal evolution pattern. In addition, the level of the conference is actually a strong indicator of the quality of the network structure. For example, top computer science conferences are considered as good social network because they produced many excellent papers, which means connections and collaborations among the researchers in top conferences are positively influences the evolution of the entire society of top conferences. Meanwhile, for computer science conferences, the Program Committee (PC) chair are usually prestigious researchers, who are in the core community of co-author network and are mainly responsible for the paper selection for conferences. Studying some metrics that are related with PC chairs of a conferences would present how a conference is influenced by these prestigious researchers.

In Chapter 5, we introduce the problem of mining community properties in computer science conferences, including how different levels of conferences evolve, and what are the differences in co-author networks. Meanwhile, several PC chair related metrics are studied to explore how prestigious researchers actually influence the society of a conference.

1.4 Contribution

In this section, we list the main contributions of three studies on the user and community data mining on large platforms in this dissertation.

1.4.1 Identify user behaviors in online education platforms

To address the problem of predicting the behavior of cold-start users on large online education platforms, we propose a method for building user demographic information using user profiles and user-generated data, which can be used to predict user performance and participation in courses either with and without the help of user history data. The demographic information includes gender, grade, school, home location, and family status, which are all trustworthy information that are important for online education to provide suitable services. In addition to the data provided by the users, we also extend the dimensions of the data by adopting online data resources, such as crawling house prices around a user's home location from real estate transaction websites to approximately denote his/her family status, using public city ranks to group the users into different groups according to the level of their cities, and computing distance between a user's home and school locations to denote the difference between home and school Socioeconomic status (SES). By aggregating user-generated data and user-provided profiles, the correlations among these parameters are analyzed in detail. In particular, user performances and participation of courses are found to be related to not only user historic data, but also user profiles. For example, users from different levels of cities show significantly different preferences of subjects in online education.

In this work, by employing a dataset from the world's largest online education platform, we successfully validate some hypothesis about the relation between student performance and student demographic information, that has not been considered before and are currently extreme important for online education platforms. Several dimensions of factors, such as city, grade, and school reputation are analyzed in detail to explore different user behaviors in different levels of above factors. Meanwhile, the user-generated data such as daily online time, historic participation of courses, subject preferences, in-class and after-class interactions with teachers, and number of terms since joining in the platform, are also proved to be correlated with user performance and participation. These correlations has been partly uncovered by some previous studies using panel survey data, which includes only hundreds or thousands of users to the maximum. To our knowledge, this is the first time that such correlations between user performance and user social profiles are studied based on a large dataset, which contains more than ten thousands users and is from the world's largest online education platform.

In addition, we explore the predictability of user behavior and participation in online education platforms by using either user profiles or user-generated data, and both of them. We divide several user features into three categories, that are user basic personal attribute, user

family and school SES, and user-generated data about previous course participation. We try to predict user performances on different courses they participate, the number of courses a user will participate, and the number of subjects a user will select, based on the three categories of features. And further experiments of using different combinations of the three categories of features to predict user behavior prove that not all the features positively contribute to the accuracy of prediction.

To sum up, the main contributions of this study could be listed as follows:

- A thorough analysis of 212,342 users from the world's largest K-12 online education;
- We build user demographic information for users on online education platforms by adopting several online data resources, which are more detailed than previous studies;
- Analysis on the relations between additional user features and user performances proves that student performance in online education are related to not only personal attribute, but also school and family SES.
- Conducted experiments of using different combinations of user features to predict user performance and user participation in online education courses, the results shows using machine learning algorithms students' online learning performance can be predicted with MAE<10%.

1.4.2 Identify user professions based on privacy-preserved mobile phone data

To address the problem of mining user-generated data with serious data privacy concern, we can actually use privacy-preserved data for identifying users or predict user behaviors. Instead of using original user-generated data directly into various models for leaning, using sensitive information eliminated data may significantly reduce the security issues of dealing with personal data. For example, to prevent malicious attempts of tracing user locations and trajectories from exact coordinates of user GPS data, we can do some statistical calculation on the location points of a user and generate several metrics (e.g., directional ranges, diameters) that can still represent the mobility pattern of the user, then only these statistical data are stored in the data bases. Meanwhile, using privacy preserved data could also reduce the dimensionality or complexity of the input data, because many data fields are not applicable and some data fields may be simplified by removing part of the factors. However, eliminating such sensitive information could absolutely reduce the quality of the data and bring challenge of keeping the accuracy of identification and prediction. Existing methods have not tackled such problem of using privacy-preserved data to identify user profession. In addition, eliminating private information by statistical methods and reducing the dimensionality of the data could also improve the efficiency of the learning and prediction processes and could reduce the network requirement for data transmission, as the size of input data decreases. Low dimension of input

data are more preferred in real-time system as the processing and transmission of such data do not consume much computational resources and network bandwidths.

In this work, we propose a system architecture to implement profession identification of mobile phone users based on privacy-preserved user-generated mobile phone data. The system also integrates a method for handling the elimination of private information from raw cell phone data, which includes three categories of information: location, app usage, and network usage. The system presents the full cycle of data procession from the data generation to end device application, in between which are raw private data storage, privacy preserve by statistical methods, data transmission from secure data bases to clouds, data request from end device, and etc. Since the prediction target is user profession, which is more related with user behavior in workdays, the data generated in holidays and weekends are all eliminated from the dataset because people's behavior during leisure time is irregular and personally motivated. In particular, we design directional ranges of a user's coordinates to present his/her mobility pattern. The output of the calculation of directional range is just several Standard Deviations (SD) with attenuation coefficient of the coordinates that are projected in several different directions, which exclude any exact location that can be used to trace where a user visits. For the information of apps, previous study attempts to find out user preferences on different apps or different types of apps, which need to know the exact names of the apps [133]. In addition, previous study also explores that on what time an app is used by a user, which could easily be exploited by malicious or marketing people for information steal and advertisement targeting. To prevent such risk of data security, instead of studying the user preferences of specific app names we attempt to anonymize the original information of different apps, i.e., we calculate the proportional preferences of all the apps according to their frequencies of use.

In conclusion, this study mainly has the following contributions:

- We propose a method of automatically identifying multiple caller's professions based on privacy-preserved mobile data, which exclude sensitive personal information to protect user from malicious attempts. And we present a system that includes the full cycle of using user-generated data to identify callers' professions.
- We conducted extensive experiments on a real data set from one of the largest telecommunication operators in China, the results show that the proposed methods could outperforms some state-of-the-art methods in terms of accuracy.
- We also analyze how different volumes of data actually influence the user identification performance, the results show that our method is robust to the size of input data with acceptable bias.
- We use statistical methods to significantly reduce the dimension of the input data without losing identification accuracy, whereas the computational efficiency could be improved.

1.4.3 Mining community difference based on collaboration networks

Motivated by the fast growth of scientific collaboration network, many studies we analyze a total of 334 conferences in computer science selected from one of the well-known conference rankings, CCF ranking [1] that rates the conferences into 3 levels: A, B and C. Firstly, we compare the evolutionary and structural differences in terms of coauthor networks. And try to trace the formation of weak ties and strong ties. Our results suggest that there are significant different structures among different levels of conferences, including the ratio of newcomers' papers, the community pattern, the connections inside the giant component. We also find top conferences are more likely to form a huge and fully connected communities, and the connections inside the giant component is also denser in top conferences, which means they are denser communities and have a high possibility to form elite small-world networks. To determine what makes the significant differences, we introduce several network based and PC chair based metrics to measure the relations among papers, authors and PC chairs. The PC chair related metrics also disclose the paper selection differences among the conferences, where several conferences are more likely to select papers (co-)authored by closely connected researchers. These quantified paper selection metrics provide useful information for authors to choose a conference that could most probably accept their papers, in addition to the common considerations. For the conference organizers, these metrics would be a warning for them to avoid the bad influences of high trust and adjust the policies to introduce more preconditions for the formation of weak ties to prevent the conferences from serious PC chairs' influences.

In summary, the main contributions of this work are:

- We analyze the structural differences among 334 different levels of computer science conferences regarding their co-author networks;
- We find that top level conferences are likely to form dense and large connected communities;
- We take a special group of identities, program committee chairs who are responsible for paper evaluation and selection, into analysis. And we design several PC chair relevant metrics to quantify their influences on chaired conferences;
- We adopt strong tie and weak tie theory to analysis hierarchical collaboration networks and find out diverse pattern that how PC chairs influence conferences.

1.5 Content Guide

This thesis includes contents of two published paper and one accepted papers.

[1] https://www.ccf.org.cn/en/Bulletin/2019-05-13/663884.shtml

- **Jiaquan Zhang**, Xiaoming Yao, and Xiaoming Fu. "Identifying unfamiliar callers' professions from privacy-preserving mobile phone data." 2020 16th International Conference on Mobility, Sensing and Networking (MSN). IEEE, 2020. [150].

- **Jiaquan Zhang**, Hui Chen, Xiaoming Yao, Xiaoming Fu. "CPFinder: Finding an Unknown Caller's Profession from Anonymized Mobile Phone Data" Digital Communications and Networks. [149]

- **Jiaquan Zhang**, Xin Gao, Hui Chen, Lei Chen, Liang Wang, Jar-Der Luo, Xiaoming Fu. "A Data-Driven Analysis of K-12 Students'Participation and Learning Performance on an Online Supplementary Learning Platform." 2021 30th International Conference on Computer Communications and Networks (ICCCN). IEEE, 2021.

- Chapter 1 introduces the background of this dissertation at the beginning. In Chpater 1.2, some basic challenges in big data mining in specific applications are discussed. Then the motivations of each work respectively are listed in Section 1.3. Section 1.4 summaries the contributions of each work. At last, Section 1.5 presents the content guide of this dissertation.

- Chapter 2 first describes general problems of user data mining in large online platforms and some remaining challenges. Section 2.2 review existing works in online education data mining and a specific challenge of addressing cold-start users. Section 2.3 briefly introduce malicious user detection and user identification in large online systems. In Section 2.4, community analysis of large online social network in academic society is introduce by a specific case in computer science conferences.

- Chapter 3 presents a novel user behavior prediction framework for large online education platform, which adopts various external data resources to build user demographic information and can predict user behavior without generate-data. And several hypothesises in terms of the relevance between user performances in online education and their social attributes are validated. In particular, Section 3.1 introduce the basic background of mining online private-tutoring data. A data collected from the world's largest online private education platform and statistical analysis on the dataset are introduced in Section 3.2. Section 3.4 present analysis of how various user attributes and user-generated data are correlated with their performances and course selections by regression methods. Finally, based on the feature regression, Section 3.5 explore using personal demographic information to predict user course selection and performances. Section 3.6 summarizes the chapter.

- Chapter 4 mainly focus on identifying mobile phone users' professions based on privacy-preserved data. Section 4.1 introduces the importance of accurate identification of mobile phone callers, and existing works on mobile phone data mining and user identification through user-generated data. Section 4.2 describes a problem state about how to identify

caller professions using user-generated data. In Section 4.3, we introduced the methods of eliminating sensitive information from raw mobile phone data and an overview of proposed CPFinder system. And in Section 4.4 the propose model is evaluated on a dataset from one of the largest Chinese telecommunication operator. Section 4.5 concludes this chapter.

- Chapter 5 studies community differences among different conferences in computer science. Different communities are found in different levels of conferences in terms of many classical network metrics and proposed PC chair relevant metrics. Section 5.1 introduces backgrounds and some basic notations of scientific collaboration communities, which in particular are co-author network. In Section 5.2, several classical metrics are compared among different levels of conferences and the results indicate significant differences among different conference levels. Section 5.3 conducts validation of PC influences on the communities by considering the relations among PC chairs and their collaborators. In Section 5.4, several PC chair relevant metrics are designed to measure the closeness of a conference from individual aspect. Finally, Section 5.5 summarize this chapter.

- Chapter 6 summarizes the contributions of this dissertation and provides plans for future work.

Chapter 2

Literature Review

Personal big data mining has been attracting the research society for decades since the availability of big personal data. Different types of platforms providing different services have diverse demands of processing user-generated data, such as item recommendation for online shopping system, user identification and community analysis for large online social platforms, and user behavior prediction for online services. Typically, machine learning methods such as Deep Neural Network (DNN), Recurrent Neural Network (RNN), and Convolution Neural Network (CNN), are used to mine feature-rich personal data to fulfill the aforementioned demands in large online platforms, and many modifications have been made to make these basic methods applicable in diverse scenarios.

In this Chapter, we first introduce basic personal big data mining issues, corresponding questions and solutions, and challenges in Section 2. Section 2.2 introduces a specific branch of personal data mining, which is user behavior prediction in online education platforms. Then in Section 2.3we introduce user identification issue based on user-generated data, which are typically referred to the data generated from mobile phones. At last, the community analysis of large online social network in academic society is introduce by a specific case in computer science conferences in Section 2.4.

Contents

2.1 Big data mining in large online platforms

Since the prosperity of diverse online services such as online shopping, online education, Online Social Network (OSN), and many other online content providers. Such online platforms usually have tens and millions of active and inactive users, and there are also considerable number of new users flooding into such platforms. As the world's largest online retailers, Amazon currently has near 300 million active users and maintains a growth rate of 10 million users per year. Especially due to the pandemic outbreak of Covid-19, many people are turning to online platforms for their daily consumption and essential resources. For example, the number of the users from online education platforms almost doubled in the year of 2020 compared to that in 2021, and Amazon delivered a record performance in 2020 with annual revenue up 38% to 386 billion dollars, a yearly increase of up to 84% net profit for the year 2020 as compared to last year. Faced with such a large amount of users and business requirements, online platforms have been working on mining user-generated big personal data to solve a variety of problems in large systems, which are key issues of providing satisfactory services to customers and managing systems in high efficiency with the lowest cost. For example, online platforms usually have to recommend products to users based on the interactions between users and items, in order to enhance user experience of browsing apps and to improve the sales [32, 127, 152]. And online platforms also have to predict user behaviors for estimating the data volume users will generate and their possible next action [82, 76, 137], according to which platforms could prepare sufficient computational resources in advance to ensure system uptime and early load resources for next action of a user to reduce latency. Second, large online platforms usually suffer from malicious users and robotic account, which pose threats to other normal users and consume unnecessary computation resources. User identification [48, 8] based on feature-rich personal data are usually applied to detect anormaly users and behaviors. Most studies attempt to adopt as much information as possible to improve the accuracy of anormaly detection, such as location with exact coordinate, interactions with items and other users, browsing history, incoming levels and other personal information, etc. In addition, large social networks formed by users from online platforms also attract many attentions since such large social networks pave the way of validating many social network theories, which are previously studied on small scale of networks formed by panel survey data. Opinion leader detection, link prediction, rumor propagation, and social community dynamics are several popular topics that depend on large online social networks, where many methods are developed to solve and improve the performance of above topics and OSN formed by real online data are usually used as an exemplary experimental targets. In particular, as one of the few large and open-source online database, scientific collaboration network is usually the primary exemplary target in many studies science the publication data is easy to access and users could be identified with their names. However, differences between communities have not been studied, as there are usually different rankings in terms of the quality of journals and conferences. Good conferences usually contain high quality publications, which may be the result of community patterns. High level conferences may form a community with high creativity due to the efficient communication

or collaboration pattern within the network, where excellent work and publications could be generated.

Personal data usually contain two categories of data resources: user-generated data and user profile data. User-generated data are generated by users when they use various services, which include locations, browsing histories, interactions, and many other text and image information. User profile data are user intrinsic information such as age, gender, profession, incoming level, education level and many other social attributes. Most of previous works focus on mining user-generated data, which are usually considered large and valuable. User profile data, on the other hand, usually provide insufficient value and in most cases are fake. As shown in Fig. 2.1, there are many patterns in terms of the data mining objectives and data categories. First, user behavior prediction and recommendation system typically target on predicting user behavioral data, and are also based on user historical behavioral records. Second, user identification aims to identity or learn personal attributes, which is also mainly based on user-generated data. Last, both user-generated data and user profiles could be used to create a variety of social networks based on different types of connections. As mentioned above, user profile data are infrequently used in personal data mining in online platforms as in most cases the information provided by user themselves are fake. For example, for many online services such as online shopping and online social platform, users are free to provide their personal information as such services do not requirement any authenticity of personal attributes. As result, only a few previous works tried to include personal profile data that are mainly collected from panel survey data. Then mapping such panel survey data to user-generated data and combine them together for user identification, user behavior prediction, and social network analysis.

However, many challenges remain in the above scenarios. First, for predicting behaviors of cold-start users, how to efficiently acquire sufficient and trustworthy data of personal attributes is still not well solved since the data provided by users in most online platforms are fake. One typical solution is finding a cold-star user's friends through social networks, who already generated enough data in the platform. Based on the assumption that close friends may have similar preference and behavior in online platform, actions of a cold-star user could be predicted approximately by data from his/her friend [34]. And a few works [4, 83] try to solve the prediction and recommendation for cold-start users combining sparse personal attributes with user-generated data into CNN models. As a fast growing industry, online education has attracted numerous users and online education platforms are faced with the challenge of how to predict student demand and performance for cold-start students. Many studies have explored the influences of online extra-curricular courses on student personal performance [54, 53], while less has studied on online private tutoring and how student's geo-location, family status, and degrees of their engagement affect learning achievements. Second, for user identification with user-generated data, the previous feature-rich methods usually take sensitive personal information such as locations, app records, call detailed records, and etc. Such personal data need to be stored in secure databases because leakage of such information bring serious threats to users. Faced with strict data protection regulations, online platforms

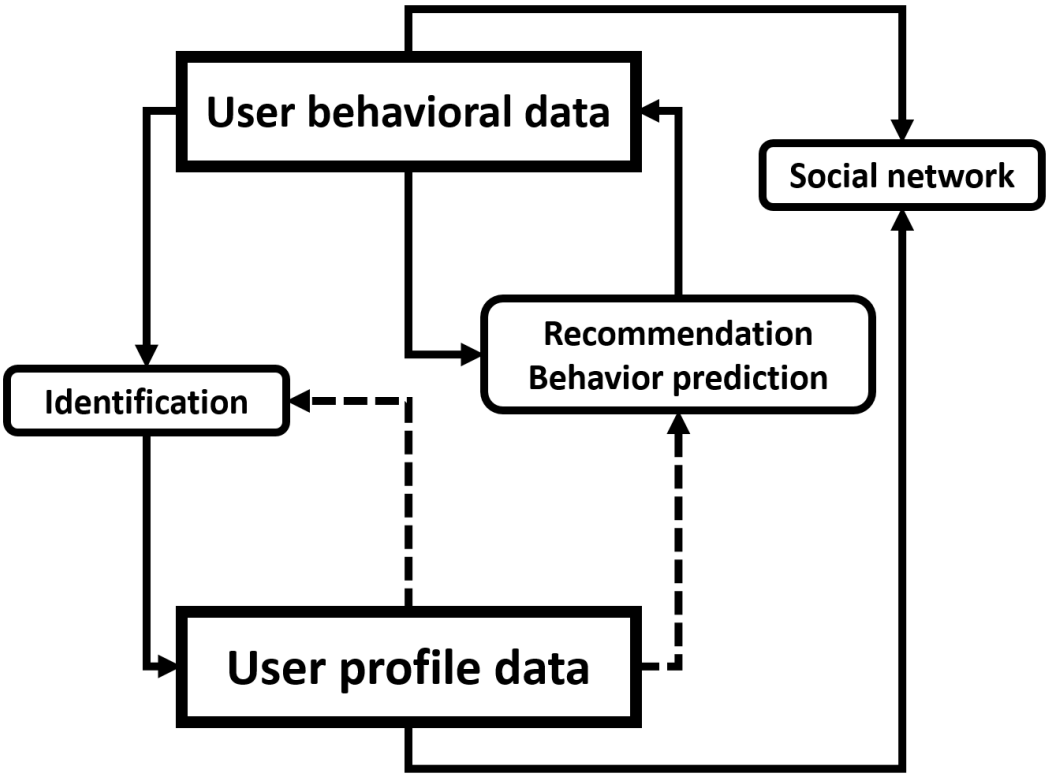


Figure 2.1: Overview of personal data mining.

have to seriously consider data protection during using user-generated data. Less studies have considered of using privacy-preserved data in user identification, i.e., eliminating privacy from original user-generated data and keeping the valuable information for modeling. In particular, malicious user detection of mobile phone numbers based on mobile phone data is one of the typical applications. Previous studies try to include as much information as possible for improving identification accuracy [99, 78, 141], without considerations of data privacy and efficiency. Third, for social network analysis on academic collaboration network, most previous works have not explore the differences among different scientific communities. Link prediction [143, 61], finding rising starts [30, 151], and both static and dynamic network structural pattern analysis [155, 156], are several hot topics regarding co-author network as experimental target. Moreover, as a special case, proceedings of conferences of computer science are usually divided into different levels according to their paper qualities. And different conferences could form their unique co-author networks, which may be correlated with their creativity. The exploration and comparison of different community patterns in different levels have not been studied, where community structure may present very different patterns in different conference levels.

2.2 User behavior prediction for online education platform

2.2.1 Behavior prediction for cold-start users

User behavior prediction based on user-generated data have been well studied by online platforms, which is usually used for item recommendation. By exploring interactions between items and users, By exploring the interaction between an item and a user, a recommendation system should predict whether a user will buy or view the item when it is displayed. Most of these approaches rely on the assumption that people buying similar goods may have similar interests, and a person's interactions with different items may also be quite similar [32, 127, 152]. In the most recent work of applying recommendation system in real platforms [83], Yifei develop a general and efficient recommendation model based on RNN, which could be applied in different use cases without time-consuming and labor-intensive training. The method combines several stat-of-the-art algorithms used in recommendation systems, such as RNN and meta-data learning. Using user-generated data for user behavior prediction has attain significant success as it is easy to build personal multi-dimension attribute based on historic data, and people usually have repeated actions periodically. One main challenge in user behavior prediction is dealing with cold-start scenario, where there is no or few historical interactions between a item with any user, or a user with any item. Meta-data learning [81, 83] has been proved to be an efficient way to deal with cold-start problem from model aspects. While content-aware methods [74, 135, 158], which try to exploit the auxiliary information form user connections and item relations, also attain success of cold-start recommendation from data aspect. However, none of these solutions consider to build user demographic information, which leverages the similarities of user social attributes, to address the cold-start problem in large online platform.

2.2.2 Mining online education data

Horton [59] classifies learning activities into three categories: 1) Absorb activities (presentations, readings, stories by a teacher, field trips etc); 2) Do activities (practice, discovery, games and simulations etc); and 3) Connect activities (guide learners to link what they are learning to prior learning and to situations where they apply current learning in subsequent courses or on their daily life). As an increasingly important way to improve academic achievements also through these learning activities, private tutoring differentiates itself from traditional public school learning (where theories and introductory information are provided which mainly constitute Absorb activities) by putting more emphasis on solving problems and practicing (which offers more Do and Connect activities) [21].

A number of work studies the impact of private tutoring on student's learning performance in different countries (e.g., Korea [110], Cambodia [21], Germany [53], China [54]). For example, Guo *et al.* [54] quantifies the impacts of private learning on students' learning outcomes based on the China Education Panel Survey, and finds that 1) subject-specific tutoring on literature and mathematics has more significant (and positive) impact on students' corresponding academic performance; 2) instead of improving students' general cognitive skills, private

tutoring improves mainly students' subject-specific knowledge and test skills; 3) private tutoring is more effective for female students, low-performing students, and students from better socioeconomic status (SES) families.

Wu *et al.* [139] find that there is no causal relationship between the family SES and student's academic performance; students' cognitive skills and psychological properties play more important role in their academic performance and students from families with high SES achieve better academic performance are likely due to their higher cognitive skills, family expectations and attitudes towards learning [18, 87]. Thus, improving low SES students' cognitive skills and psychological properties will help improving their academic performance. Matsuoka [86] indicates that students from high family SES and in higher ranking schools have more opportunities to participate in shadow education courses. The author uses the aggregation of all students' SES to compute the school SES and finds that there is a high correlation ($\beta = 0.73, p < 0.001$) between school SES and school ranking. However, obtaining the accurate SES information with household surveys and questionnaires is time-consuming and human-intensive. We note location-related information such as family housing price (which partially reflects the affordability of a family for children' private education) may provide a new means to estimate SES [36], which we leverage in this paper to understand the effects of SES and location on students' private tutoring participation and performance. Literatures [119, 20, 77] study how students' school attributes and family SES affect the decision to take private tutoring. Literatures [134, 56] also investigate the relationship between SES and academic achievements. While SES can be estimated through different means, e.g., with home location [36], the impact of SES on online private tutoring stays largely unknown. In summary, much research has examined how learning achievements are affected by private tutoring; less has studied on online private tutoring and how student's geo-location, family status and degrees of their engagement affect learning achievements. This paper aims to fill the gap with a data-driven approach.

2.3 User Identification for mobile phone users

Mobile phone data have received increasing attentions from data science researchers, especially in social computing areas [27, 88, 46, 26, 147, 60]. Some studies have analyzed the mobility patterns of populations on the city level [136]. They found people's moving distances follow a scale-free distribution. Public health [94] and regional air pollution [92] could also be inferred by mobile phone data. These studies analyzed group behaviors and the corresponding statistical parameters of certain districts. With the development and wide deployment of high-speed cellular networking technologies, users' data volume significantly increases, which makes it possible for researchers to study personal characteristics based on big data analysis [121, 28]. As a result, studies on profession, human activity, age, and gender prediction have become widespread. For example, mobile phone data are used to identify personalities and could implement multiclass classification [27, 88]. In particular, the usage of apps have become a main source feature for identifying personal characteristics [133, 28].

Besides phone-based data, these works also included other information, such as age and gender from survey or questionnaires. From these works' perspectives, simply relying on artificially collected data is inefficient and unsuitable for real-time identification.

There are also many studies on detecting harass or fraud phone callers using mobile phone data for run-time application, which are mostly considered as binary classification problems [66, 100, 78]. Furthermore, most of these studies used Call Detail Records (CDR) and Short Message Service (SMS) data to derive their predictive models [66, 63, 148]. CDR data are very useful for fraud phone detection because fraudsters significantly differ from normal users in terms of dialing properties. Meanwhile, content analysis based on natural language processing or artificial intelligence methods has recently attracted attention. [99, 154]. However, these kinds of data are insufficient for identifying strange callers' professions, which is a multiclass classification problem. People of different professions, such as delivery people, taxi drivers, and telemarketers, share similar dialing properties as they have to frequently communicate with customers. CDR data could also provide location information estimates based on the base stations' locations [46, 147, 63]. However there is a large bias in such locations because each tower is in a fixed position and with a coverage of up to 1-km radius in most urban areas. [80] shows GPS data could provide more precise location information, enabling people to build personal maps for mobility pattern analysis.

2.4 Analysis of scientific collaboration network from large online databases

With coauthor networks, researchers' attributes are no longer restricted within personal aspects, but generalized to social patterns containing many graph based metrics. People have used the graph based collaboration networks to find rising stars in the coming few years [30, 151], to predict personal academic progress such as h-index and number of publications [57, 37, 104], as well as to predict the links [143, 61, 125] or citations [146] among the authors. Despite personal behaviors studies, detecting communities concealed inside the extreme complex networks are also a hot topic including finding local communities, tracing communities' evolution [15, 155], network embedding [129, 160, 9] and hierarchical hybrid networks [156]. The scientific network based on DBLP data is used to test algorithms of social influence maximization [122]. A focus of these studies is on networks, including coauthor networks which regard authors as nodes and collaborations as edges, citation networks [35]which regard publications as nodes and citations, as well as mixed networks containing multiple different elements, especially on formulating network models to best characterize the academic society [91, 90].

Several platforms specifically developed for analyzing academic social networks become popular recently. For example, Aminer [130] from Tsinghua University and Acemap [128] from Shanghai Jiaotong University provide detailed and plentiful statistical indexes of a huge number of conferences and journals. Acemap further provides visible interfaces of the network structure of the academic collaborations and it supports querying of several network metrics. However,

these works do not investigate the differences among different conferences but only focus on the whole academic community constructed from the database, which is a homogeneous network. Recently, some works try to find out the diversity of academic collaboration networks that influences the papers [5]. They find there are strong correlations between the ethnic diversity and papers, as well as scientists. There are many natural ways of partitioning a researcher community of a certain area into pieces, such as by topics, by affiliations, by geographic boundaries, by conferences' levels, etc. Even though these factors have been considered for academic society analysis, but they are processed as a single node when constructing the networks, which could not provide in-depth knowledge of the unique network structure in each conference. The analysis of the statistics of some special nodes is a way to find out the diverse community patterns of different conferences.

Chapter 3

Mining user behaviors based on credible user profiles in online education

Due to the limitation of public school systems, many students pursue private supplementary tutoring for improving their academic performance. Different from public schools, the private online education provides diverse courses and satisfy differentiated demands of the students. Students' behavior and performance in online supplementary learning are relevant to not only personal attributes, but also some factors such as city levels, grades and family situation. Existing studies mostly rely on panel survey/questionnaire data and few studied online private tutoring. In this Chapter, with 212,342 anonymous From Kindergarten to 12th Grade (K-12) students' 3-year learning data from one of the world's largest online extra-curricular education platforms, we uncover students' online learning behaviors and infer the impact of students' home location, family socioeconomic situation and attended school's reputation/rank on the students' private tutoring course participation and learning outcomes. Further analysis suggests that such impact may be largely attributed to the inequality of access to educational resources in different cities and the inequality in family socioeconomic status. Finally, we study the predictability of students' performance and behaviors using machine learning algorithms with different groups of features, showing students' online learning performance can be predicted with Mean Absolute Error (MAE)$< 10\%$.

Contents

3.1 Introduction of online education data mining

Online education has significantly grew in last several years, especially in the areas where education competition is very severe. Due to the limitation of public school systems, many students pursue private supplementary learning for improving their academic performance, especially in east and southeast Asia [110, 20, 54] and US [39, 16]. According to the National Bureau of Statistics of China [89], 220 million students were enrolled in K-12 education in China in 2019, among which 150 million students (68.1%) enrolled in compulsory education (K1-K9). It is reported about 21.9% of K1-K6 (primary school) students, 36.8% of K7-K9 (junior high school) students and 57.8% of K10-K12 (high school) students in China took extra-curricular courses in 2017, in both on-site and online forms [33]. Faced with such a large amount of new users, online education platforms are challenged by how to prepare sufficient but not excessive computational resources to ensure the quality of video streaming, how to arrange different students into suitable classes according to their abilities, and how to recruit sufficient but not excessive teachers and tutors to ensure the quality of courses. To address such challenges in online education platforms, accurate user behavior prediction can help platforms estimate how many courses each user will participate and how much improvement each user will acquire after taking the online courses. In particular, how to accurately predict the demands for cold-start users, who have no historical user-generated data, is the biggest challenge.

Existing studies on private tutoring primarily focused on using panel survey/questionnaire data to understand the impact of its on-site form on academic performance [110, 21, 54, 53]. The outbreak of COVID-19 pandemic raises a new challenge for private tutoring. First, many private tutoring activities as well as public educational activities have to go online [17, 96]. Second, students with different socioeconomic background and from different cities may attend a same class, without the need to sit in a common physical room and with flexible tutoring hours and class composition [69]. Socioeconomic status (SES) as an important factor for studies on children's development, is typically measured by household income, education and occupational status, representing the economic position and social status (prestige) of a family in related to others. However, it has not reached a consensus when concerning different cultures and regions [64, 18]. [38] uses family pays for rent and housing as a proxy, which is also one of the important features to represent SES in this paper. With the increasing availability of computing devices and improved Internet connectivity, online private lessons and tutoring have been more and more popular nowadays. COVID-19 coincidentally increases the pace of wide adoption of online private tutoring.

We conjecture that factors like family inequality, location/regional inequality and students' engagement will affect the students' online extra-curricular course participation and learning performance, on which no consensus has reached in previous literatures [86]. Specifically, few studies consider the impact of the difference in family socioeconomic status (SES) on students' obtaining additional learning opportunities [118, 86]. Based on the behavior and performance data of the students in online learning courses, the online education institutions

could develop different advertising strategies for students in different regions, family and age groups. Meanwhile, the platforms could optimize their course structures and human resources according to the demands of students, in order to provide higher quality education and more cost-efficient services, where more students could benefit from.

With 11,392 anonymous K-12 students' data between January 2018 and December 2020 from one of the world's largest online supplementary education institutions, we quantitatively analyze these impacts and test the following hypotheses:

1. The more frequent interactions between a student and the teacher/tutor, the better learning performance;

2. Students with lower difference in housing prices between family and school locations will more likely choose more (and pay more) on private tutoring courses and get better learning performance;

3. Given a student's demographic information, family socioeconomic status (approximated by housing price levels), and learning behaviors, it is possible to predict the student's learning performance and course participation.

3.2 Data Statistics

3.2.1 Overview of Datasets

Our analysis is mainly based on three datasets: 1) learning data of 11,392 students with anonymous demographic and location information (from the online learning platform); 2) city ranking data; 3) students' family socioeconomic status (SES) data, using housing prices as a proxy.

The students' online learning dataset comprises three subgroups of information for each student: 1) demographic and location information such as city name, school's rough location, home's rough location, gender and grade/age; 2) registered course names and terms; 3) study records for each course, including in-class and after-class engagement/interactions with the teacher/tutor, ratio of correctly answered questions in in-class quiz. Note each academic calendar year is divided into 4 terms, namely spring semester, summer break, autumn semester, and winter break in China, each offering complete courses in different frequencies (e.g. once a week in semesters vs. 4-6 times a week during breaks). A student may choose up to one course for each subject during a term.

The influential Yicai Magazine ranks Chinese cities into 6 categories based on size and strength: Tier 1 (4 cities), New Tier 1 (15 cities), Tier 2 (30 cities), Tier 3 (70 cities), Tier 4 (90 cities) and others (below-Tier-4); we crawled the detailed list from `https://www.yicai.com/news/100648666.html`.

The housing prices near students' home locations are used as the proxy of the students' family SES. The housing price is divided into 13 categories with an interval of 10,000(Chinese Yuan (CNY)/m^2). The housing prices are crawled from the China Real Estate Association (`https://www.creprice.cn/`).

3.2.2 Statistics

The statistics are depicted in Table 3.1. Among all 11,392 students, 2.7% didn't provide gender information, and 49.8% of the remaining are males. 59.1%, 32.8% and 9.1% of the 11,392 students are enrolled respectively in primary schools, junior high schools and high schools. The students are distributed in Tier-1 (27.6%), New-Tier-1 (26.5%), Tier-2 (18.1%), Tier-3 (14.1%), Tier-4 (9.1%) and below-Tier-4 cities (4.6%). The housing prices of students' home locations range from 2,447 to 166,185 (CNY/m^2) (Mean=35,262, Standard Deviation (S.D.)=34,919), which act as a proxy for their families' SES [36]. The difference between students' home and school housing prices (Price$_{home}$−Price$_{school}$) ranges from $-154,695$ to 121,744 (CNY/m^2) (Mean=4,186, S.D.=21,945). Intuitively, except for those extremely high SES families, due to the affordability of expensive houses near the Chinese schools, the shorter commuting distance between a student's home and school locations, typically the higher SES is the student's family. Hence, the housing price difference between a student's home and school locations may additionally reflect the student's family SES. Therefore, we also compute the commuting distance between home and school locations (Mean= 29,266m, S.D.=136,950, Min=58, Max=2,903,401). As shown in Fig. 4(d), majority of students are living within $< 3km$ distance to their schools. The outstanding mean value here due to a small amount of students with very long distances between their home locations (e.g., in a different city) and school locations.

Among the 11,382 students, 8,211 have performance records with no missing values of other variables. The average number of questions a student answers in classes ('in-class interactions') is 8.76 (S.D.=0.11, Min=0, Max=137). The average online time is 93.44 minutes (Min=1.25, Max=296.13). The mean value of correct answer ratio ('ratio') in all classes is 21.74% (Min=0.41%, Max=100%). In the observed 3-year period, the average number of terms these students have attended online courses is 6.51 (Min=1, Max=21). For students' choice of courses/subjects, we denote the value as 1 if a student selects a certain course, 0 if not. Among all courses, Math, Chinese and English courses are most popular as 76%, 72% and 60% of students registered them respectively. The frequency of after-class interactions between students and teachers/tutors is also computed. The average interaction frequency (No$_{after_interactions}$) is 1,340 times (Min=0, Max=53,759).

We also explore the correlation between each pair of variables. Fig. 3.1 illustrates the Pearson correlation between students' attributes, interaction frequency between students and teachers/teacher's assistants, course selection behaviors and students' performance. If there is no significant relationship between variables, the corresponding box will be blank. The blue

Table 3.1: Statistics

	min	max	mean	S.D.
$\text{No}_{interactions}$	0	137.00	8.76	11.02
$\text{No}_{after\_interactions}$	0	53,759	1,340	3,412
time (minutes)	1.25	296.13	93.44	33.51
ratio (%)	0.41	100.00	54.07	21.74
distance (meters)	58.29	2,903,401	29,266	136,950
Price_{home} (CNY)	2,447	166,185	35,262	34,919
$\text{Price}_{home}-\text{Price}_{school}$ (CNY)	-154,695	121,744	4,185	21,945
courses_per_term	0.71	8.75	1.68	0.79
paid_per_term (CNY)	-1,832	13,650	2,099	1,037
number_subjects	1.00	9.00	2.58	1.18
terms	1.00	21.00	6.51	3.40
English	0.00	1.00	0.60	0.49
Math	0.00	1.00	0.76	0.42
Chinese	0.00	1.00	0.72	0.45
Physics	0.00	1.00	0.20	0.40
Chemistry	0.00	1.00	0.11	0.31
Biology	0.00	1.00	0.03	0.17
Computer	0.00	1.00	0.13	0.34
Politics	0.00	1.00	0.00	0.05
History	0.00	1.00	0.00	0.06
Geography	0.00	1.00	0.01	0.08

circles represent a positive correlation while the red represents a negative correlation between them. The bigger the circle is, the higher Pearson coefficient value between them.

1) A strong positive correlation is shown between after-class interaction frequency and the amount of subjects, courses and tuition fee paid for each term on online courses, which means the students having more interactions with their teachers/tutor would select more subjects, more courses and spend more money on the extra-curricular education platform. However, the interaction frequency between students and teachers/tutors has no significant correlation with students' performance.

2) No significant correlation is shown between the housing price difference (or distance) and students' learning performance and purchasing behaviors, thus hypothesis 2 needs further analysis with regression models.

3) Concerning students' average online performances, students in higher grades who choose Math, Physics and Chemistry have higher in-class interaction frequencies, while students who choose English and Computer have shorter online time. The in-class quiz correctness has a strong negative correlation with selecting English, slightly negative correlation with selecting Math, subject number, course number and tuition fee, while having a positive correlation with in-class interaction frequency and online time. It means more active students tend to answer a higher percentage quiz. These results provide preliminary evidence for hypothesis 3.

More detailed regression analyses are given in the following sections to test the hypotheses proposed in Section 3.1.

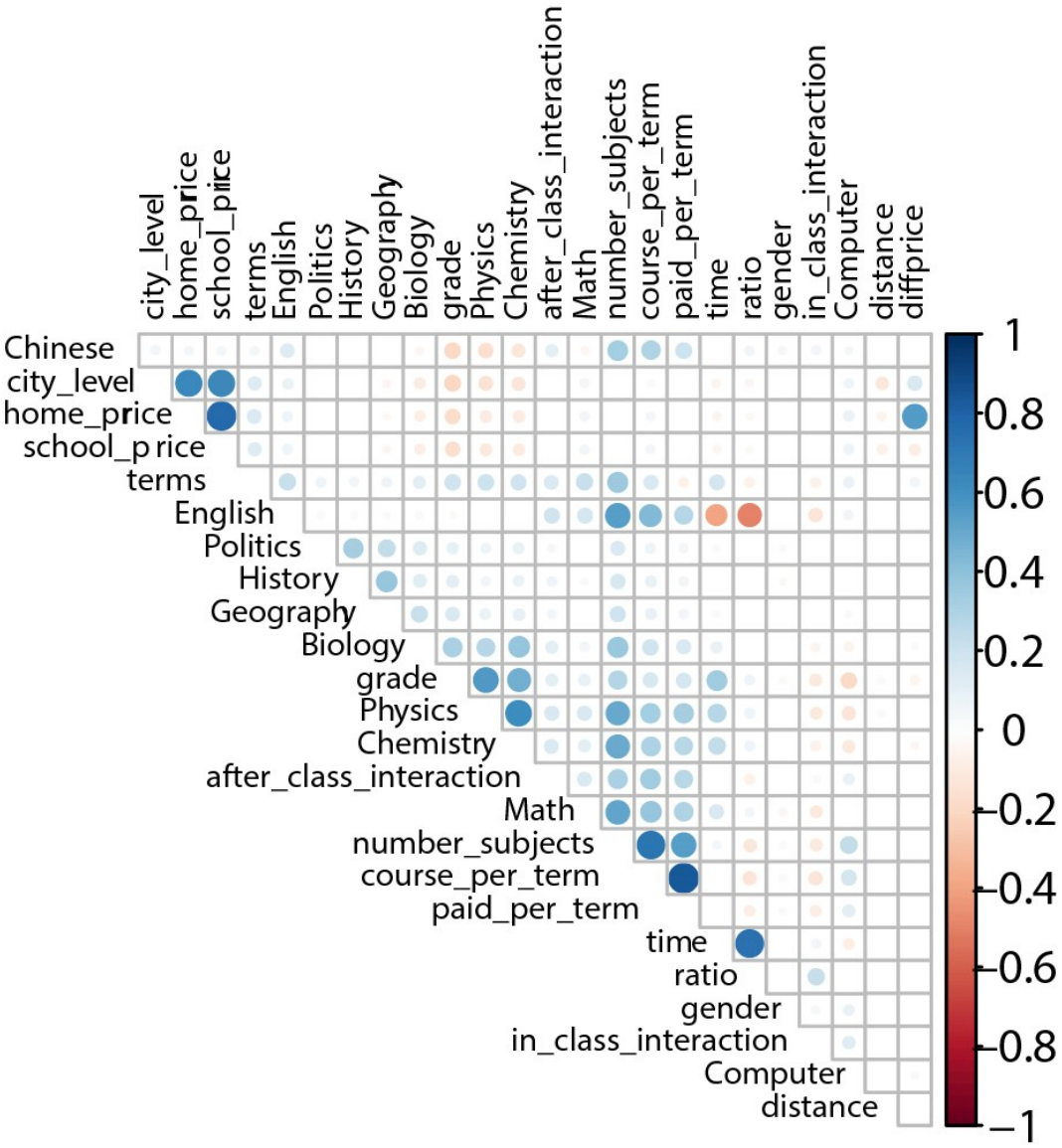
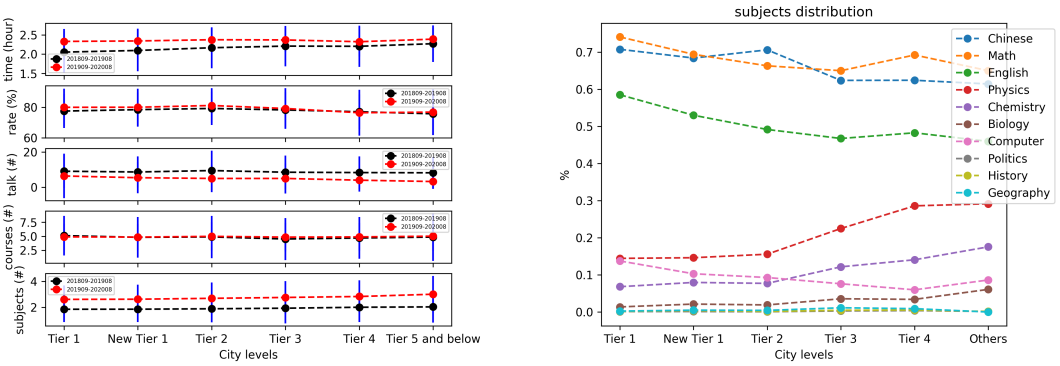


Figure 3.1: Correlation between variables

3.3 Student attributes and behaviors vs. learning performance and course selection

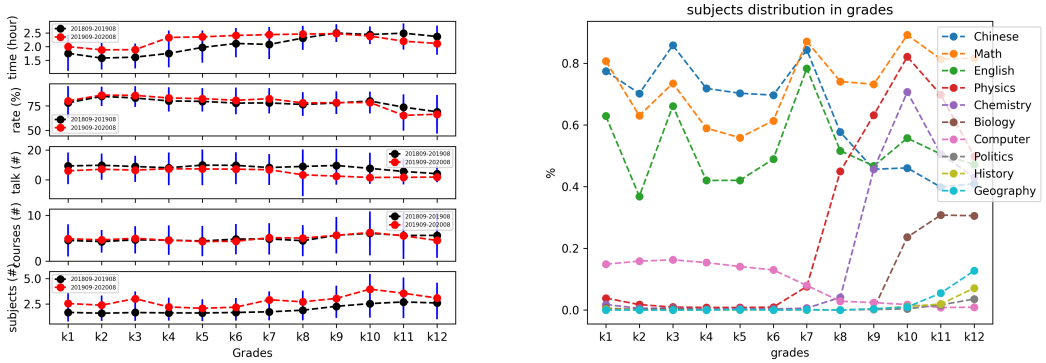
In this section, we study how the differences of students' demographic features and online learning behaviors (such as current grade a.k.a. age, city rank, family SES and geolocation information) may affect the course selection and learning performance. Specifically, we use five kinds of features that are relevant to in-class interactions and course selection:

- Online time: The average time a student spends in a class could indicate his/her persistence in classes.
- Frequency of in-class interactions: The average in-class interaction frequency indicates a student's engagement in class, the higher the more active the student is.
- Subject numbers: The number of subjects indicates a student's course diversity. Moreover, a detailed analysis of specific subjects is used to study the difference among different cities and grades.
- Course number per term: The average number of courses a student participates in each term, which strongly indicates how much effort a student puts on online courses.
- Quiz correctness ratio: This largely indicates a student's learning performance. Temporal analysis on ratio may also estimate learning effects.



(a) **City levels statistics** (b) **Subjects' popularity in different city levels**

Figure 3.2: Statistics of 5 features and subject selection in different city levels



(a) **Grades statistics** (b) **Subjects' popularity in different grades**

Figure 3.3: Statistics of 5 features and subjects selections in different grades

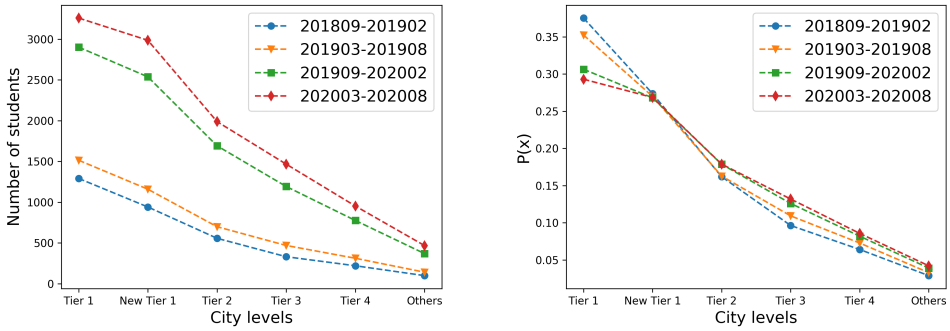
Different city levels: As shown in Fig. 3.2(a), the average time spent on online courses increases with the decrease of the city levels and finally leads to an average daily difference

about 0.5 hours. The correctness ratio does not present significant differences among the Tier-1, New Tier-1 and Tier-2 cities, but tends to decline after Tier-2 cities. Moreover, students from cities below Tier 4 averagely have fewer subjects and a lit bit more in-class interactions than students from higher level cities. For the statistics in terms of grades, the 4 features among the students in different grades show visible differences in Fig. 3.3(a). Overall, the time in each class, the correctness ratio and the number of subjects slightly increase with the increase of grades before K10, and tend to decrease between K10 and K12. But the interaction frequency in the classes does not differ much among different grades.

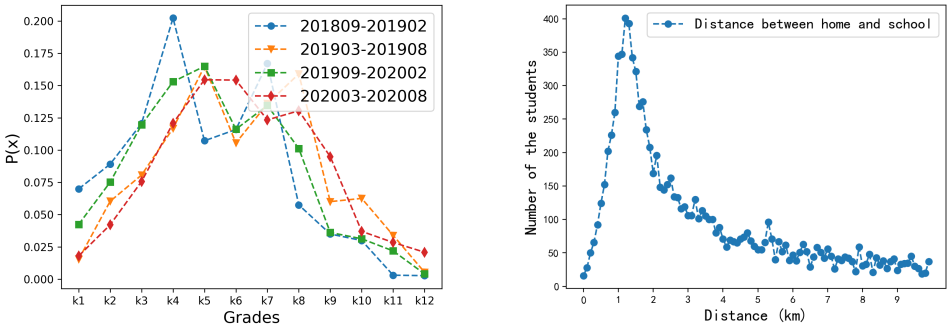
In addition to the average number of subjects that each student participates, the specific subjects each student selected differs in different city levels and grades, which can provide more information about different students' demands of extra-curricular education. Fig. 3.2(b) shows the percentages of students for each specific subject in different city levels. Typically, Math, English and Chinese as well as Computer courses are visibly becoming less popular with the decrease of the city levels. Reversely, the percentage of students who selected Physics, Biology and Chemistry courses slightly increases with the increase of the city levels. However, considering some subjects start in different grades (e.g., Chemistry only starts in K9), a lower percentage of the students does not indicate less importance or fewer students' attentions to the course.

Different grades: Fig. 3.3 shows the students' percentage of each subject in the dimension of grades. Overall, Math keeps the most importance and most popularity throughout the K-12 grades. Chinese has high importance in primary schools (having more students than Math) but turn to lose its importance after K7. An evolutionary pattern in terms of the importance is shown about English course as Math does, but less than that of Math. Chemistry and Physics start in junior high school and become more and more important until the beginning of high school, then start losing students in the following grades. Specially, K3, K7 and K10 are 3 grades that demands of the online education reach the peak. As K3, K7 and K10 are the first important years in primary school, junior high school and high school stages respectively in China, the increase of the demands of online education in these grades is likely due to the significant discontinuous articulation of knowledge, making them rely more on online learning.

Trend in student distribution: We study the overall trend from the perspective of the online participants' distribution in 2018-2020. Fig. 3.4(a) shows that the number of the students almost doubled in 2020 compared with that of 2019, while a trend can be seen that online education tends to attract more students from low-level cities. Currently, the linear descent of the student numbers from large cities to small cities is mainly due to the parents' higher SES and their attention on education. Parents from larger cities may have clearer understandings of the importance of education and have the ability to afford extra education expenses. Additionally, parents in small cities tend to trust more face-to-face education than online educations. A plausible response for online education institutions is to develop differentiated pricing strategies according to the overall economic status in different regions.



(a) **City level (number)** (b) **City level (ratio)**



(c) **Grades (ratio)** (d) **Home-school distances (ratio)**

Figure 3.4: Student distributions in city levels, grades and home-school distances

Fig. 3.4(c) and (d) show the distribution of the student numbers in different grades and different distances between home and school, respectively. The student numbers shows a Gaussian-like distribution with the highest demands between K4 and K7, and the distribution does not change too much over time. The learning content in lower primary grades is typically limited, while the high school students are mostly stressed by the learning content in their own school and have no time for extra online education. The distance between home and school locations shows a long tail distribution, where most students are close to their schools ($<$ 3km).

Our data analysis shows significant differences among the students in different city levels in terms of course selection and learning performance. Additionally, the student's family SES is a more precise factor of students' behavior in online education. The quiz correctness ratio slightly increases with the increase of student family SES, and the improvement of the correctness ratio after one year online learning also shows a positive relationship with family SES. For subject selection, the overall trend is similar with that of city level, with a few differences, such as Computer courses are mostly selected by students with high family SES (except the highest SES).

School ranking (Case in Beijing): To explore the relationship between school ranking and housing prices of students in Beijing according to the reputation and quality of these schools

reported by various sources (`http://www.51sxue.com/`, `http://www.xschu.com/` and `http://www.jzb.com/bbs/bj/`) with different ranks: high (1), middle (2), and low (3). The student distribution is 33.6% in rank 1, 31.2% in rank 2 and 35.3% in rank 3. We compute the Pearson correlation and find the school ranking highly correlates with school housing price ($\beta = .335, p < .001$) and family housing price ($\beta = .24, p < .001$). However, since currently we only get the school ranking in one Tier-1 city (Beijing), we don't explore this factor further in the studies in Sections V and VI.

Then we compute the mean value and standard deviation differences in students' performance and purchasing behaviors with different school ranking. As shown in Table 3.2, students in rank 1 take (and purchase) more private tutoring courses. Students in rank 2 have the longest average online time and the highest average ratio of correctness in quiz among the three groups. Students in rank 3 have the highest in-class interaction frequencies while the quiz correctness ratio is not as good as those from rank 2. Notably, although students in rank 1 were enrolled in more courses, they did worse than students from schools of lower ranks. We don't have these students' performance data in their schools, but since students in different areas and school rankings may have online courses in a class, the online performance can reflect the activity and learning outcomes of these students to some extent. In short, this evidence shows that students from higher ranked schools do not always achieve better learning performance or behave more actively than lower ranked schools. , which shows education inequality.

Table 3.2: Static statistics of students' performance and subject selection in different school levels

school rank		in-class interact	time(m)	ratio(%)	courses_ per_term	paid_per _term(CNY)
1	mean	7.45	4,871	47.11	1.71	2,367
	S.D.	7.63	2,004	24.72	0.79	1,292
2	mean	7.51	5,319	51.24	1.63	2,253
	S.D.	9.36	2,274	26.21	0.67	1,072
3	mean	8.43	4,994	47.85	1.64	2,286
	S.D.	11.16	1,984	25.72	0.73	1,170
all	mean	7.81	5,054	48.66	1.66	2,303
	S.D.	9.51	2,091	25.56	0.73	1,183

3.4 Feature Regression and Statistical Analysis

Ordinary Least Squares (OLS) regression analyses are used to examine how students' attributes, subject/course selection behavior and their family SES affect their course selection and performance. The regression results are shown in Table 3.3.

1) We find a significant positive effect of after-class frequencies between students and the teacher to in-class interaction ($\beta = .0886, p < .001$), which means the more interactions between students and teachers, the better in-class performance the students will get, **proving hypothesis 1 holds true**. This evidence was also found in [108].

2) The regression results show that there is no correlation between the higher difference in housing prices between a students' home and school locations and the students' performance. However, the higher difference in housing prices between a students' home and school locations, they tend to spend more money on private tutoring platforms ($\beta = .0284, p < .05$). Therefore, **hypothesis 2 is partially proven**. This also confirms the finding in [86] that family SES impacts students' participation in shadow education, which is similar to private tutoring courses in this paper.

3) Regression models 1-3 (Table 3.3) show that given a student's demographic information, family SES, and previous subject selection and in-class engagement, it's possible to predict the student's learning performance and course participation. The R^2 of ratio is 0.2925, online time is 0.3540. However, the R^2 of in-class interactions is 0.0665, which is not significant. This may be because some courses (e.g. Physics, History) do not provide many Question Answering (QA)s. Thus we use machine learning models (in Section 3.5) to further test hypothesis 3.

Table 3.3: Performance regression

Dependent variable	Ratio	Time	Interactions	No_subjects	Courses_per_term	Paid_per_term
control						
Grade	0.0617***	0.2650***	-0.0530***	-0.0016	0.0286**	0.0662***
Gender_Male	0.00001	-0.0092	0.0242*	0.0143	0.0047	0.0147
Family SES						
home_price	0.0026	-0.0113	0.0002	-0.0026	0.0364**	0.0503***
city level	0.0090	0.02543*	0.0042	-0.0004	0.0200	0.0284*
Diff (school SES, family SES)						
$\text{Price}_{home} - \text{Price}_{school}$	0.0024	-0.0040	-0.0158	0.0018	-0.0141	-0.0245*
Commuting distance						
Distance (family, school)	-0.0021	0.0041	0.0040	0.0012	-0.0045	0.0007
Course selection type						
English	-0.5111***	-0.4626***	-0.0924***	0.4083***	0.3342***	0.2368***
Math	0.1572***	0.1718***	-0.0200	0.3602***	0.2709***	0.2447***
Chinese	0.1804***	0.1210***	0.0901***	0.3749***	0.3063***	0.2469***
Physics	0.0588***	0.0731***	-0.0110	0.3264***	0.2416***	0.2904***
Chemistry	0.0415**	0.0429***	0.0301*	0.2580***	0.1688***	0.1136***
Biology	-0.0216*	-0.0337***	-0.0007	0.1397***	0.0432***	0.0304**
Computer	0.0751***	-0.0020	0.1301***	0.2859***	0.2075***	0.1645***
Politics	0.0001	-0.0005	0.0050	0.0389***	0.0090	0.0090
History	-0.00003	-0.0094	0.0070	0.0575***	0.0288***	0.0178
Geography	-0.0039	-0.0105	0.0057	0.0593***	0.0219**	0.0116
Historical behavior						
Duration of attendance (No_terms)	-0.0166	0.1659***	-0.0400**	0.0061***	-0.1243***	-0.3149***
Number of courses (per_term)	-0.0669***	-0.0053	-0.1619***	0.0163***		
Interactions (student, teacher/tutor)						
Avg interact frequency (student, teacher/tutor) (averaged by term)	-0.0176	-0.0186	0.0886***	0.0007	0.1355***	0.1158***
R^2	0.2925	0.3540	0.0665	0.9914	0.5333	0.4014

Table 3.4 shows that males have more in-class interactions ($\beta = 0.0242, p < .5$) but there is no significant difference in other performance between males and females, which shows different results with [54]. As proven by [43, 72] that males are not disadvantaged in students' learning achievements in Math and Science subjects, especially for participation rates which is similar to in-class interactions in this context. As for grades, the regressions show a similar finding.

We use group regression to further explore the effects of the difference among primary school, junior high school and high school. As shown in Table 3.4, high school students in highly developed cities have a higher correctness ratio and online time, while it is not found in

other two groups. Among the primary and middle school students, the higher their family house price is, the more money tends to be paid on the private tutoring platform, but it doesn't impact the purchasing behavior of high school students. As for after-class interactions between students and teachers/tutors, there are negative effects on the correctness ratio of primary school and junior high school students. This finding was also proved by [108] that the effects of negative relationships of students-teacher interaction and students' performance were stronger in primary than in secondary school. It is all positively correlated with purchasing behavior in all stages. As shown in Table 3.5, for primary school students, the larger distance between family and school, the more money is paid to the tutoring platform. And the difference in housing prices between family and school negatively affects the amount of tuition paid on tutoring courses. This result makes a complement for the findings in [86] that students with higher family SES will engage more in shadow education when taken their grades into consideration. Higher family SES has a stronger influence on students' selection behavior of private tutoring courses when they are in primary and junior high school stages. This study gives a more comprehensive perspective on family SES and the difference between family and school SES by considering the housing price, location and distance (housing price of students' family, commuting distance and difference in housing price between home and school). Moreover, we also select Beijing as an example to illustrate that there is a strong correlation between school housing price and their rankings, which gives a new enlightening for the measurement of school SES when there is no ground truth of school ranking in the whole country.

Table 3.4: Group regression of students' performance in three stages

	city level			home_price			after-class interactions		
Dependent Variable	primary	junior high	high	primary	junior high	high	primary	junior high	high
Rate	0.0036	0.0044	0.1136*	0.0074	-0.0186	-0.0286	-0.0439**	-0.0373*	0.0288
Time	0.0180	0.0185	0.1049*	-0.0088	0.0023	-0.0066	-0.0756***	-0.0155	-0.0650***
Interactions	-0.0021	-0.0216	-0.0544	-0.0048	0.0191	0.0623	0.0699**	0.1260***	0.1615***
paid_per_term	0.0127	0.0428*	0.0251	0.0958***	0.0687**	-0.0343	0.4652***	0.1285***	0.3443***

Table 3.5: Group regression of students' purchasing behavior in three stages

Independent variables	stages	paid_per_term
distance (home, school)	primary	0.0230*
	junior high	0.0022
	high	-0.0001
$\text{Price}_{home} - \text{Price}_{school}$	primary	-0.0359**
	junior high	-0.0406*
	high	0.0244

3.5 Prediction of Learning Performance and Course Selection

Based on the analysis above, we can conclude that the students' behavior and performance on online education platforms are related to many factors, such as grades, city levels, family SES, in-class interaction frequency, and so forth. Therefore, the students' behavior and performance could be predicted by several features. In addition to the statistical analysis of the students'

behavior in different cities, grades and SES, individual behavior prediction could provide more useful information for both students and online education platforms. If a student's behavior could be predicted, the online education providers could develop more precise advertisement placement to reduce cost and provide tailor-made course combinations to maximize the learning performance for each student. The students and the parents could decide how many courses they need to achieve the best learning improvement and how much efforts they should put in the courses (i.e. times and participation of the in-class interactivity). As a result, the study of predicting the students behaviors' on private online tutoring platforms is conducted to seek for possible scenarios that can be adapted into practicality. According to the analysis above, 3 metrics that are important for students' online learning behaviors are set for predictive targets, namely the ratio of correctness in in-class quiz, the number of courses bought in each term and selected subjects.

1. **Ratio**: Correctness ratio of in-class quiz;
2. **Daily online time**: The average time each student spends in online education every day;
3. **In-class interaction frequency**: Average number of interactions (Q&As) with teacher in a class;
4. **After-class interaction frequency**: Average number of after-class communications with teacher/tutor;
5. **Gender**: Whether the student is female or male;
6. **Grade**: The grade of a student by the end of year 2020;
7. **City level**: The city level (out of 6 categories) of a student's home location;
8. **Distance**: Distance between home and school locations;
9. **Home price**: Housing price around home location;
10. **Price difference (home-school)**: Housing price difference between a student's home and school locations;
11. **Terms**: The number of terms a student attends within the 3-year period;
12. **Courses per term**: Average number of courses per term;
13. **Subjects**: Total number of subjects;
14. **English course number**: Total number of all English courses a student participated on the online platform;
15. **Math course number**: Total number of online Math courses a student participated in;

16. **Chinese course number**: Total number of Chinese courses a student participated on the online platform.

For predicting these 3 metrics, 5 representative regression methods are used to test the prediction power of the given parameters: Gradient Boosting Regression (GBR), K-Nearest Regression (KNR), Random Forest Regression (RFR), Multi Layer Perceptron (MLP) regression, Adaptive Boosting (AdaBoost) [98] and TabNet [10]. All the boosting and bagging methods are set with a number of 500 estimators, the n_neighbors of KNR method is set to 10, while MLP has a hidden size of 100 layers and a max iteration of 500. To evaluate the performance of the prediction methods, the evaluation metric Mean Absolute Errors (MAE, see Eq. 3.1) is applied here as all the predictive targets are continuous values. All the features are divided into 3 clusters according to their attributes. Features 1) to 6) are relevant to students' basic attributes and combined as cluster C1; features 7) to 10) are combined as cluster C2 (related to family SES, approximated with home and school housing prices); and features 11) to 16) are combined as cluster C3 (about students' participation in the online education platforms). For each prediction target, the target itself is excluded from the inputs. Different combinations of the 3 clusters of features are tested to find out the optimal performance of each method. Except some stable embedding algorithms of bagging and boosting, other unstable methods are all averaged of 100 realizations.

$$\mathrm{MAE}(y, \hat{y}) = \frac{1}{n_{\text{samples}}} \sum_{i=0}^{n_{\text{samples}}-1} |y_i - \hat{y}_i| \,. \tag{3.1}$$

3.5.1 Prediction of Correctness Ratio (Learning Performance)

The correctness ratio of in-class quiz is a strong indicator for evaluating the performances of the students. If the online learning performance could be approximately predicted based on students' demographic attributes and family SES as well as previous course selection information, it may provide useful hints in advance for both the students and online education platforms. Table 3.6 shows the MAEs of 5 methods under 7 different feature combinations for predicting the ratio of correctness in in-class quiz. The best result is with MAE of 0.0877, when all the features are included in the input for the Adaboosting regressor, which means that all the features are useful when predicting the students' learning performance. Meanwhile, the combination of C1 and C3 features without C2 features could achieve a very close result to the best MAE of ratio prediction, which means family SES contributes the least for the prediction of students' performance on online learning platforms, while the other two clusters of features contribute the most for the accuracy of predicting the ratio of correctness.

Table 3.6: MAE of ratio prediction

	GBR	KNN	RFR	MLP	AdaBoost	TabNet
C1 + C2 + C3	0.12	0.1102	0.1296	0.1252	0.1005	**0.0877**
C1 + C2	0.1296	0.1278	0.1271	0.1047	0.111	**0.0969**
C1 + C3	0.1163	0.1108	0.1295	0.1051	0.099	0.0898
C2 + C3	0.1604	0.1201	0.1391	0.1191	0.1247	0.121
C1	0.1328	0.1247	0.1346	0.115	0.1134	0.0997
C2	0.215	0.1795	0.176	0.1752	0.1806	0.175
C3	0.1571	0.1167	0.1404	0.1174	0.1259	0.1162

3.5.2 Predicting of participation in courses

The prediction of the number of course per each term could help the platforms develop more adequate course structures and place the advertisements more precisely to the potential customers. Similarly, Table 3.7 shows the MAEs of predicting the average number of courses in a term with 5 algorithms under 7 different combinations of features. Unlike the results in ratio prediction, the best result in predicting the number of courses is obtained by the MLP regressor under the feature combination of C1 and C2, with MAE of 0.1363. Considering the average number of courses in the whole data set is 1.68, the error is less than 10% if compared with the mean value. However, the dataset does not contain information on some students' previous course selection. At this moment, the feature combination of only C1 and C2 is reasonable for predicting the number of courses, as the features in C3 contain much information about the course selection. Within such limitations, the best MAE achieved by the experiments without C3 features is 0.3688, also by using the MLP regressor. Moreover, the result demonstrates that the user profile data also contribute to the accuracy of predicting course participation.

Table 3.7: MAE of course numbers prediction

	GBR	KNN	RFR	MLP	AdaBoost	TabNet
C1 + C2 + C3	0.2894	0.2454	0.4369	0.1833	0.4555	0.1761
C1 + C2	0.498	0.426	0.4374	**0.3688**	0.396	0.4086
C1 + C3	0.2986	0.2419	0.4524	0.1866	0.4418	0.1495
C2 + C3	0.3144	0.1944	0.4788	**0.1363**	0.5002	0.1564
C1	0.5184	0.4311	0.4665	0.3788	0.3962	0.333
C2	0.7665	0.6286	0.6006	0.5988	0.6497	0.3995
C3	0.3003	0.1741	0.4653	0.1455	0.5024	0.1423

3.5.3 Prediction of subject numbers

The number of the subjects examines the diversity of a student's online learning demands. Different from correctness ratio and the number of the courses, the number of subjects is more personalized attributes, that has fairly little relations of the given features including the grade, city, genders and family SES. As the number of subjects is strongly related with the course selection, so the meaningful results in predicting the number of subjects should also exclude the features of cluster 3. As shown in Table 3.8, the best MAE in predicting the numbers of

subjects is 0.5929, which is produced by the MLP regressor. But if a students has some history data of course selection, the MAE will be improved a lot to 0.2016.

Table 3.8: MAE of subjects numbers prediction

	GBR	KNN	RFR	MLP	Adaboost	TabNet
C1 + C2 + C3	0.4609	0.5401	0.6695	0.2762	0.7408	**0.2016**
C1 + C2	0.7453	0.686	0.7018	0.5929	0.6435	0.7849
C1 + C3	0.4403	0.5364	0.6385	0.278	0.8016	0.2182
C2 + C3	0.4806	0.4672	0.6592	0.2582	0.7098	0.3014
C1	0.7461	0.6651	0.6918	0.5843	0.6997	0.6054
C2	1.1664	0.9593	0.9531	0.9518	0.9508	0.9622
C3	0.4893	0.4123	0.6661	0.2594	0.8383	0.2276

In addition, for the prediction of user subjects selection, we also examine the predictability of whether a user will choose courses of a specific subjects. Five different subjects are selected for experiments as they have sufficient number of users, while the other courses are ignored due to the lack of participation. For each subject, a new column indicating whether a user will choose this subject is generated and is used as prediction objective. As a binary prediction problem, instead of using MAE as evaluation metrics we adopt another more powerful metrics, F1-score, to evaluate the performance of predicting specific subject selection for individuals. In addition, because it is obvious that the three features in C3, English course number, Chinese course number, and Math course number, are highly related to individual selection of specific subject, we drop these three features in the experiments. For each subject and each combination of the features, we similarly applied the six algorithms into modeling and only the best one is shown in Table. 3.9. Similarly to above results, the best performance of predicting individual subject selection is using all the three categories of the features, which shows that user profile information could help improve the accuracy of user behavior prediction. In particular, using only C1 and C2 features, the F1-score is just a little bit worse than that of using all the three categories of feature, which means for cold-start users the individual selection of specific subjects could still be predicted.

Table 3.9: F1-score of subjects selection prediction

	Chinese	Math	English	Physics	Chemistry
C1 + C2 + C3	0.9274	0.9537	0.9374	0.8745	0.8355
C1 + C2	0.9134	0.9340	0.9252	0.8622	0.6875
C1 + C3	0.9178	0.9449	0.9279	0.8739	0.7076
C2 + C3	0.9096	0.9317	0.8934	0.6216	0.6854
C1	0.9109	0.9347	0.9272	0.8588	0.6700
C2	0.9137	0.9298	0.8583	0.7855	0.7328
C3	0.9151	0.9359	0.8972	0.5799	0.7689

3.6 Chapter Summary

This paper studies students' learning behaviors and performance on an online private supplementary education platform. We specifically focus on the effects of inequality from different aspects (family SES and difference between family and school SES, gender, city level, etc.) on the performance and courses' selection behaviors. Our study exploits a new angle by using housing price, location and distance (housing price of students' family, commuting distance and difference in housing prices between home and school) to represent the family SES and school attributes. Furthermore, our analyses show although the higher family SES and higher difference in family and school SES significantly influence students' monetary spending on private tutoring courses, the effect diminishes with the increase of grades. Also, these inequalities don't affect students' online engagement and performance significantly. As for gender, except for class participation, there is no significant difference in class performance and course selection behaviors for males and females. We also use machine learning to examine the predictability of learning performance and course selection.

Our analysis of SES and learning performance is based on some simplifications which may lead to inaccuracy. We plan to conduct online user questionnaires like what we have done in [45] for collecting more ground truth data from students and teachers to better assess student's academic achievements and their influencing factors including more accurate measures of family SES. Furthermore, we are currently studying the impact of the COVID-19 pandemic on different types of students.

Chapter 4

Mining user behaviors based on privacy-preserved mobile phone data

Identifying an unfamiliar caller's profession is important to protect citizens' personal safety and property. Due to limited data protection of various popular online services in some countries such as taxi hailing or takeouts ordering, many users nowadays encounter an increasing number of phone calls from strangers. The situation may be aggravated when criminals pretend to be such service delivery staff, bringing threats to the user individuals as well as the society. Additionally, more and more people suffer from excessive digital marketing and fraud phone calls because of personal information leakage. However, previous works on malicious call detection only focused on binary classification, which do not work for identification of multiple professions. We observed that web service requests issued from users' mobile phones may exhibit their Apps preferences, spatial and temporal patterns, and other profession related information. This offers researchers and engineers a hint to identify unfamiliar callers. In fact, some previous works already leveraged raw data from mobile phones (which includes sensitive information) for personality studies. However, accessing users' mobile phone raw data may violate the more and more strict private General Data Protection Regulation (GDPR). We observe that appropriate statistical methods can offer an effective means to eliminate private information and preserve personal characteristics, thus enabling the identification of the types of mobile phone callers without privacy concern.

In this chapter, we develop CPFinder, a system which exploits privacy-preserving mobile data to automatically identify the callers who are divided into four categories of users: taxi drivers, delivery and takeouts staffs, telemarketers and fraudsters, and normal users (other professions). Our evaluation over an anonymized dataset of 1,282 users with a period of 3 months in Shanghai City shows that the CPFinder can achieve an accuracy of 75+% for multi-class classification and 92.35+% for binary classification.

Contents

4.1 Mobile phone user identification

With the widespread use of online services, such as online shopping, food delivery ordering, and taxi hailing, there is a significant annual growth of platform-to-consumer delivery (8.2%) [101], restaurant-to-consumer delivery (6.8%) [106], ride and taxi hailing (17.5%) [107]. As a result, numerous people frequently receive unfamiliar phone calls from delivery people and taxi drivers. Moreover, many people are suffering from digital marketing and fraud phone calls due to the leakage of personal information from phone number related services. Criminals may pretend to be delivery men and contact victims by making phone calls, which could cause personal safety crisis and property loss. Similarly, fraud phone and telemarketing calls cause millions of financial loss yearly [1]. Therefore, it is critical for the mobile phone users to recognize the professions of callers to protect themselves from potential dangers and losses. The phone identification systems, such as Baidu phone number labeling [1] and 360 phone number query [2], offer an effective way of solving the aforementioned problems by retrieving information of each phone call accumulated before. Originally developed for only identifying fraud phone calls, the phone identification systems were generalized to cover new phone user types, such as delivery persons and taxi drivers since the emergence of smart phones. However, the currently off-the-shelf phone identification systems face the following challenges, especially concerning efficiency and accuracy. First, these phone identification systems are based on manual and subjective reports and annotations from end users. As shown in Fig. 4.1, a phone number is labeled as advertisement purposes and has been reported by nine users, as shown in Baidu search. However, owing to the laziness and unwillingness of many end users, the annotation rate is very low (0.154) and a high percentage of incorrect labels are generated [79]. Similar observations are also made in Google search. The incorrect labels could bring much inconvenience for normal users, e.g., their phone calls may be refused by others if they are mislabeled as fraud phones. To make the service more accurate and avoid wrong information due to incorrect labels, only the phone numbers with sufficient reported times could be released to the public by phone labeling services providers. Therefore, there will be a long gap for releasing the label information since a phone number was first labeled. To overcome these problems and improve the performance of the phone labeling systems, we develop *CPFinder*, a new machine learning-based system to automatically identify phone labels by leveraging only anonymized mobile phone cellular data without the need of any manual interventions.

Smart phones exchange a large amount of data traffic and application (app) request data across the cellular base station, where large records of Uniform Resource Locator (URL) data are generated by users. Such requests could provide lots of useful information, enabling researchers and engineers to study people's daily behaviors and living habits, which are highly related to people's occupation [126, 27, 88, 46, 133]. However, privacy preservation remains a critical issue for handling personal data [65, 124]. Most personality studies use raw mobile phone data, including sensitive personal information such as locations, application logs, and

[1] https://haoma.baidu.com/mark
[2] https://chaxun.360.cn/chaxun/tel

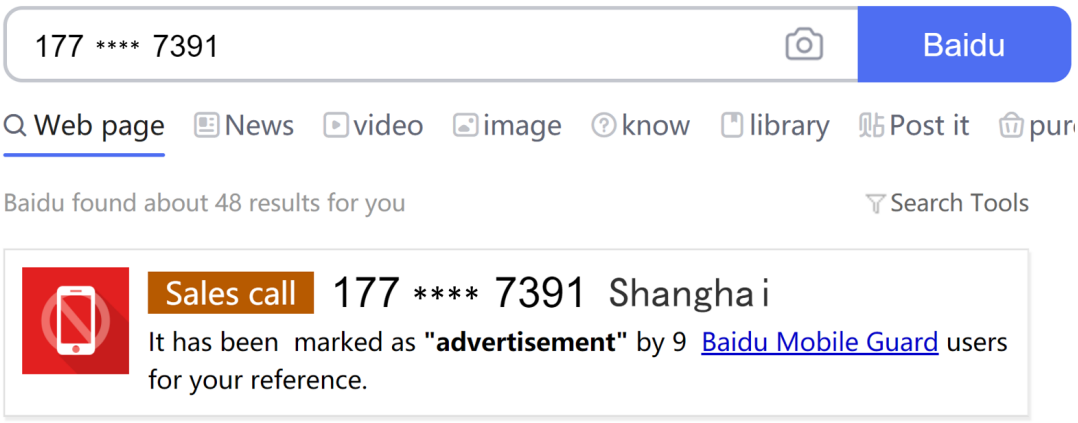


Figure 4.1: Baidu phone label search result.

charging records. Often, more detailed personal data from surveys are included as well. The major privacy concern in these works is believed to be addressed by encrypting personal IDs, such that the owners of the data are claimed to be protected. However, a user could be re-identified even though the labels are invisible [124, 133] provided the input data remains unchanged. Personal locations, application usages, and activities could be used to trace and (re-)identify a person.

As personal data protection policies are becoming more strict, using sensitive information for personality identification may violate various data protection regulations. For example, General Data Protection Regulation (GDPR Art. 71) mentioned that private data, such as personal preferences or interests, reliability or behaviour and locations should be processed by appropriate mathematical or statistical procedures for privacy concern. Compared with where people visit (exact coordinates) and what apps people use (app logs), how people move and how people use smart phones are less sensitive. Typically, these statistical parameters could not be used to reversely trace any personal information and will not violate privacy protection regulations. As GDPR recital 71 recommends [49], appropriate statistical procedures should be used while profiling personal data to prevent potential risks of retrieving personal interest and privacy.

In this work, based on our analysis on an anonymous dataset from one of the largest telecommunication operators in China, we develop CPFinder, a machine learning-based intelligent phone user profession identification framework, which can identify phone labels among four categories, including normal users, taxi drivers, takeouts and delivery, fraud, and telemarketing. According to the information provided by various apps' network traffic and the professional attributes of each category, we develop comprehensive features via statistical analysis, and some nonsensitive parameters based on privacy eliminated data are created to model users' patterns. We show that only with the mobile phone user data of a single day, the profession category of a user could be identified with a high accuracy. The key highlights of the proposed framework are as follows:

- It uses privacy-preserved personal data as the inputs for user identification, which enables end-device data processing and complies with data protection regulations.
- It contains multiclass classification module, which complements previous binary classification methods. Binary classification is also embedded in the system, and outperforms some previous works [1, 126].
- It significantly reduces the number of features and the requirement of data size, which helps improve the efficiency to 10 times compared with feature-rich methods.

4.2 Problem statement

In this paper, we develop a method to allow mobile phone users to identify a caller's profession immediately when they receive a call from an unknown number. The data used for identification are primarily the web service requests issued from the devices and recorded by telecommunication operators, who could push the identification results to end-users when necessary. Suppose there is a set of mobile phone users U, each user has a phone number p_i, a label of profession l_i, and mobile phone records R_i, where $U = \{[p_1,\ l_1,\ R_1],\ [p_2,\ l_2,\ R_2],\ ...,\ [p_n,\ l_n,\ R_n]\}$. The mobile phone records are web services requests, that $R_i = \{[time,\ request],\ [time,\ request],\ ...\}$. $Time$ is in the form of *YYYYMMDDHHmmSS* with a precision of seconds, and the request is always formed as $www.xxx.com/...longitude = xx\&latitude = xx/....$ The problem is using the data of R_i to identify the label of profession l_i without any external data resource. Currently, four categories of professions are included, where $l_i \in \{Normal,\ Driver,\ Delivery,\ Harass\}$.

Base on the information of web services requests, three kinds of highly profession-related and sensitive information-eliminated features are constructed: mobility pattern, request volume, and apps preferences.

4.2.1 Mobility pattern

A user's mobility pattern characterizes the user's overall movements throughout a day. A statistical parameter, Standard Deviation (SD) is applied to describe mobility. Because the location records are not always complete, using statistical parameter could maximally preserve mobility characteristics. The four categories of professions have very different moving ranges in every single working day.

4.2.2 Request volume

Request volume tells the number of requests in different time, which could indicate people's activeness. The more requests, the more active a user is. Different professions have significantly different active time. Request volume is different from data volume, as one request of video

may generate much more data transmission than several requests of pure text. Therefore, the number of requests is much more suitable to characterize activeness.

4.2.3 App preferences

App preferences indicate how frequently each App is used. The apps here are web services extracted from the main domain names of the requests because the requests could be generated by browsers or many other apps. The main domain name is extracted by eliminating the prefix and suffix representing the same institution (i.e., diditaxi out of www.poiservice.diditaxt.com). Therefore, two different domains, www.common.diditaxi.com.cn, and www.poiservice.diditaxi.com are referred to an identical main domain name, which indicate a root related web service. Each profession has its own several frequently used services, which results in its unique apps preferences pattern.

4.2.4 Time span of data

The time span of the data is also considered to be a factor for modeling. Each of the above mentioned features could be computed using the data of an arbitrary number of days, but at least one day is needed. Generally, if more days are included in the data, a better identification performance will be achieved. The problem is how much the accuracy improves with the increasing of time length of data. It addresses the question of the identifying efficiency, which cares about if the data of one single day is sufficient to identify a caller's profession. The overall privacy preservation processing for CPFinder is shown in Fig. 4.2.

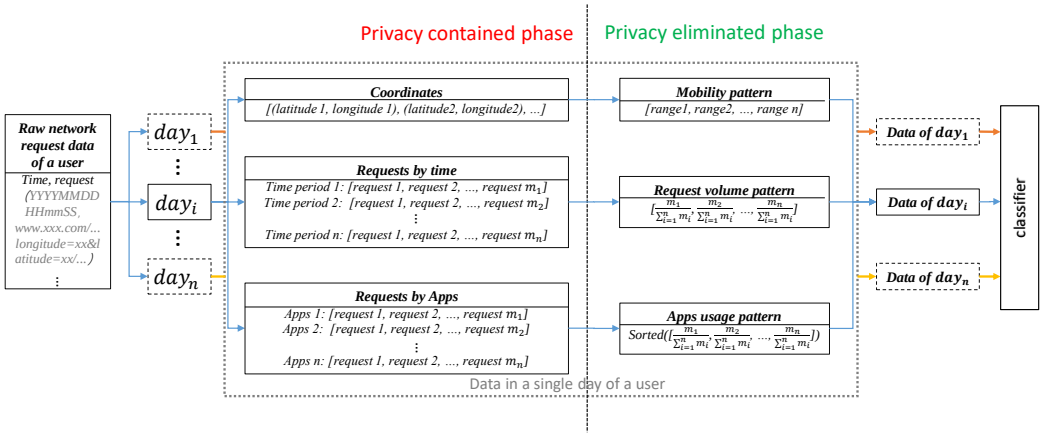


Figure 4.2: CPFinder privacy preservation process.

4.3 Mobile phone users identification framework

4.3.1 Mobility pattern

Researchers found that mobility trajectories could be useful in identifying personalities [133, 147]. In these works, users' exact coordinates are used to locate their work and residential places. Exact coordinates could provide the possibilities of tracing where users are and thus may face the personal data protection issues. However, using some statistical parameters to characterize people's mobility pattern instead of using exact locations could avoid violating any data protection policies. To ensure data privacy protection, any information about how people move cannot be used to trace personal location.

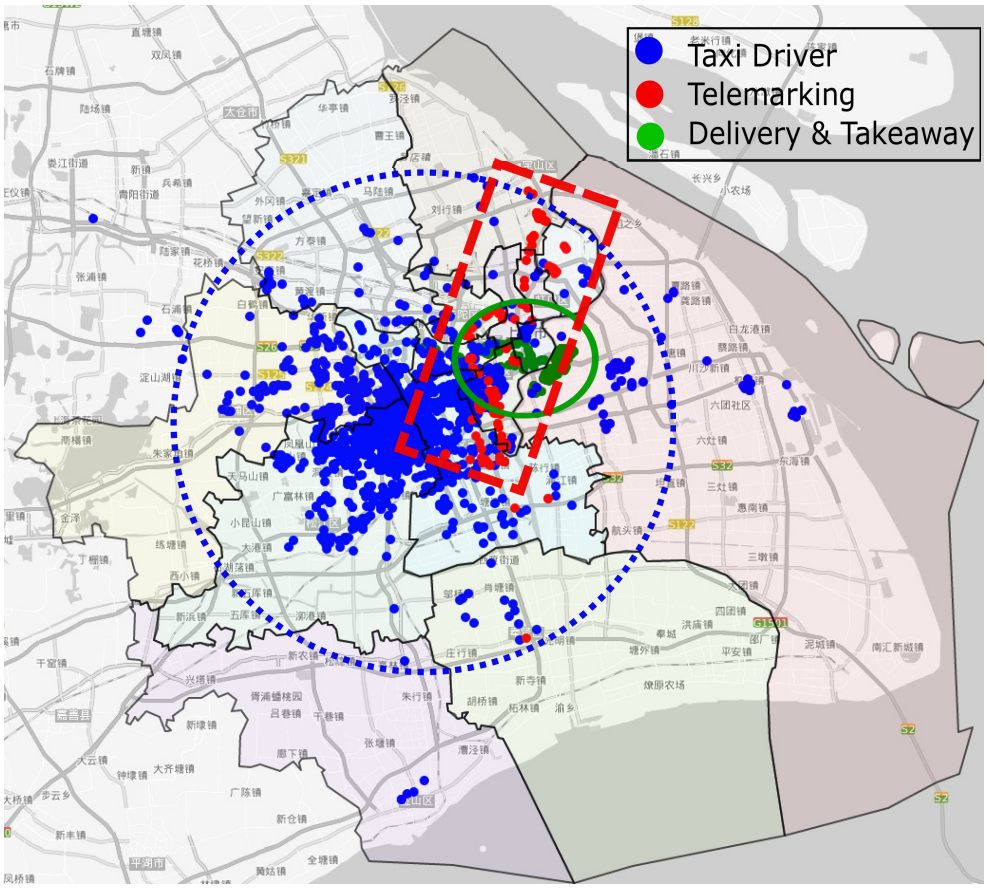
Our analysis shows mobility pattern is strongly related to users' professions, especially on workdays. This coincides with the following intuitive scenario: during holidays, people's activities are self-related with high randomness, which means it is hard to find a common mobility pattern for people with an identical label if considering their mobility in the leisure time. For this reason, we first study users' mobility patterns of active time (6:00–24:00) on workdays. As shown in Fig. 4.3(a), three typical users with different labels are presented with their locations in one week. Not surprisingly, each category of users has its unique mobility pattern: taxi drivers cover a large area and have large ranges in all directions, takeouts and delivery staffs mostly serve in certain regions with limited areas, whereas telemarketing and fraud phone callers mostly show a strip-like commuting pattern (which indicates their commutes between their residences and offices). Normal users are excluded in the figure, because normal users' mobility patterns are not similar and are of large bias among different entities.

Base on the above analysis , we propose an SD-based parameter to denote the mobility patterns of different professions. The parameter contains the ranges of 12 evenly distributed directions on the flat. For a given set of N points, a list containing frequencies of each point V, a certain direction d (angled with x axis of θ), and a hyperparameter α to adjust the logarithmic calculation of frequency, the range of the points in direction d is as follows:

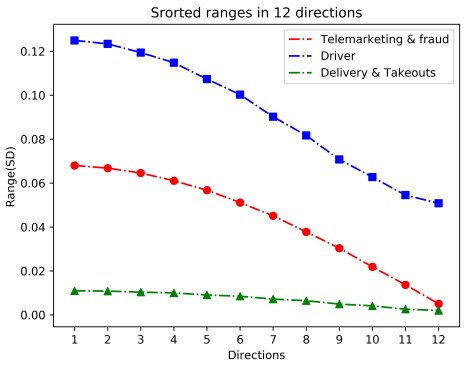
$$R_\theta{}^2 = \frac{1}{N} \cdot \sum_{1}^{N} \Big((x_i \cos\theta + y_i \sin\theta - \overline{x\cos\theta + y\sin\theta})^2 \cdot (1 + \alpha ln(V_i)) \Big) \tag{4.1}$$

where N is the number of points, x_i and y_i are the longitude and latitude of each point. The hyperparameter α is a real number between 0 and 1, which can be tuned to improve the prediction accuracy by reducing the bias cause by some point with extra large frequencies.

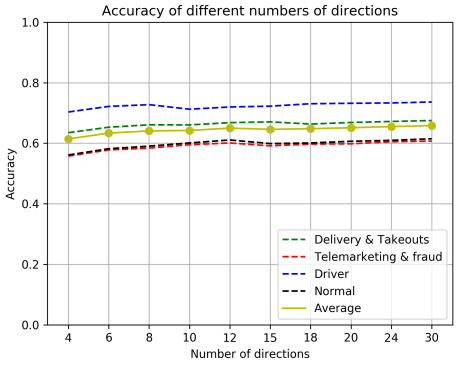
For each user, his/her daily mobility pattern is characterized by a vector, containing ranges of several different but evenly distributed directions. In addition, the range vector is sorted from maximum to minimum to eliminate the directional bias. For a certain pattern, its sorted range vector remains the same no matter how it rotates. As shown in Fig. 4.3(b), the sorted ranges of 12 directions significantly differ among three professions. Drivers and telemarketers have a decreasing trend of the sorted ranges, but driver's range is much larger than telemarketing and fraud users. Delivery staff have a more balanced pattern, where there are rare differences among



(a) Locations of three different professions



(b) Sorted standard deviations of 12 directions



(c) Accuracy of different numbers of directions

Figure 4.3: Users' mobility patterns in active time of workdays and its statistical results.

all directions. For the normal users, their patterns are somewhere between drivers and delivery people, and no unified pattern could be found because their trajectories vary among different people. However, they are different from the other professions in terms of the following features. The number of directions used for characterizing a person's mobility pattern needs investigation in terms of both accuracy and efficiency. As shown in Fig. 4.3(c), when increasing the number of directions, the accuracy remains almost unchanged after more than 12 directions, which means using more than 12 directions will only increase the computational complexity without bringing any improvement on the performance.

4.3.2 User data volume in different time periods

Data volume can indicate the frequency of using mobile phones. The higher the data volume, the more network requests are generated, which also means the mobile phone is more frequently used. The data volumes in different time slices are also related to professions. Each profession has its unique pattern of active duration. As shown in Fig. 4.4, a day is divided into 6 equal time slices: 6:00–9:00, 9:00–12:00, 12:00–15:00, 15:00–18:00, 18:00–21:00, 21:00–24:00. Data between 0:00 am and 6:00 am are excluded, because there are too few records during this period. Since some apps may automatically generate network requests when executing in background and this kind of execution usually generates a large amount of repeated requests within a second. These large amounts of repeated requests, which are generated within a short time period, could significantly interfere with the analysis of data usage in different time periods. Therefore, for each time stamp, if a request has many duplicated records, the duplicated records will be deleted. Fig. 4.4 shows the distribution of number of network requests for all professions. Drivers and telemarketers share a very similar data usage pattern; they are more active than the other two professions after 18:00. For delivery staff, their data volume is somehow evenly distributed in different time slices and is more stable than the others. For normal users, the data volume distribution is slightly different from the delivery people's. For each user, his/her data volume pattern in a single day is created by calculating the daily network request distributions, which is a vector of six elements. The ith element is given by $v_i = n_i/n_{total}$, where n_i is the number of network requests in time slice and n_{total} is the number of requests in an entire day.

4.3.3 Apps preference distribution

Mobile phone apps usage plays an important role in personality studies [27, 88, 133, 93]. These studies attempted to figure out what apps people use, what categories of apps each person prefers to use, as well as how someone use different apps. Apps log containing what apps are used, as well as when and how long an App is used, could easily be leveraged by others to forward advertisements to certain devices. For privacy concern, plaintext of apps names should not be contained in any parameters. In addition, the exact apps' names could not be directly used as inputs unless they are transferred as numeric values, e.g., mapping each App to an integer. Moreover, as there are thousands of apps used by people, it is very complex to create users' apps preference patterns when considering the exact apps. Futher, as there are

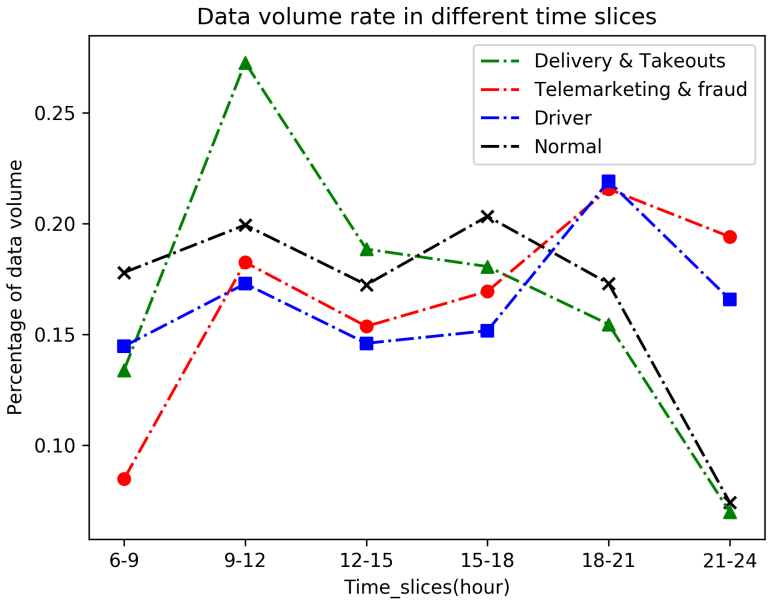
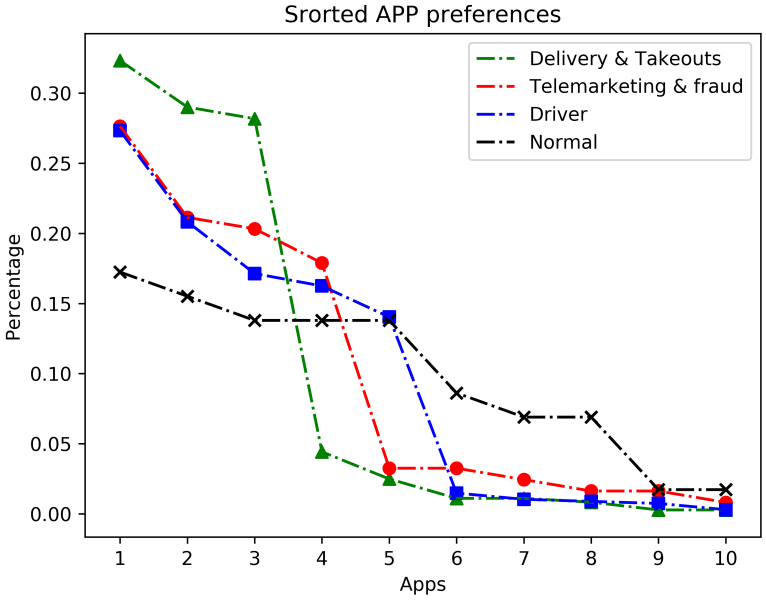


Figure 4.4: Data volume distribution in different time slices.

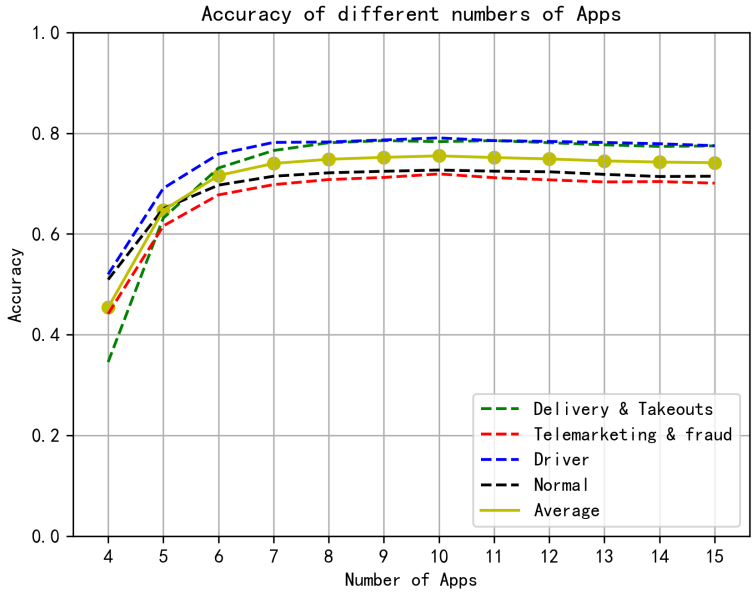
new apps released frequently, it is impossible to maintain the mapping list when considering the new apps. A more efficient and more accurate way to characterize users' App preferences is finding the distribution of 10 mostly used apps, regardless of the exact host names. Each network request is related to a web service, which could be found from the host. Fig. 4.5 shows the normalized daily apps usage distribution (sorted from the most frequently used to the least frequently used) of four categories. Except normal users, other professions have a steep descend in their apps usage distributions, which means they have several frequently used apps or services. In particular, the delivery people have the least frequently used apps or services. Meanwhile, the normal users show a more balanced distribution. The best number of most frequently used apps is also investigated in Fig. 4.5(b). With more than 10 apps, the percentages of the four categories are all extremely low and close to each other. As a result, it is unnecessary to include more than 10 apps. Moreover, the accuracy tends to decrease slightly when including more than 10 apps.

4.3.4 CPFinder system implementation

Based on the above analysis , the users' data will not contain sensitive personal information after privacy-preserving processing. The privacy preserved data could then be stored out of the centralized secure sever and even be transmitted to end devices. Previous methods use original data that contain sensitive personal information to build features, resulting in the limitations that the entire procedure (including both training and identifying phases) should be operated within a secured server. Then, only the identification results will be directly transmitted to the end users. Considering that there are numerous cell phone calls every second, generating users' data and identifying the labels for all phone calls is a big challenge for the server.



(a) Sorted usage rate of 10 most frequently used apps



(b) Accuracy of different numbers of apps

Figure 4.5: Apps preference.

However, if the identification task could be pushed to be done in each end device, the workload of the centralized server could be substantially reduced in terms of both computational amount and storage. Transmitting the input data from a secure server to end devices is feasible only when the input data do not contain any sensitive personal information, in case of not violating some data protection policies and regulations. Distributing the privacy-preserved data is acceptable as the data cannot be used for any malicious purpose. As shown in Fig. 4.6, the private user-generated data, which are stored in the centralized secure sever, are first being processed to eliminate sensitive information (by Fig.4.2) and exported as privacy-preserved data. Then, the data could be distributed to servers with lower security levels or even unsecured servers. The servers could train a model with the privacy-preserved data periodically, and the identification models in the end devices could be kept updated when a newly trained model is available. When a device receives a call from an unfamiliar number, the serve could transmit the corresponding data of the number to the device simultaneously for identification.

The CPFinder system is also flexible and efficient, as it does not require excessive computational and network resources. For the secure database maintaining original user data, the main computation is processing the user data into privacy-preserved form and transmitting the processed data to distributed insecure clouds. The processing of the user data only has to traverse each user just once, the time complexity is linear. For each user, the additional data outputs the module is only 28 float numbers and a phone number, which are much smaller than the size of original data and are portable for transmission. For the clouds that holds non-private user data, the main computation is training the models and transmitting the identification models to the end devices. As the above tasks only have to be done periodically, they do not require powerful CPUs or large bandwidths. The only real-time application is when an end device receives an unfamiliar phone call, the end device has to request the data of the specific users from the clouds and identify the label. However, the data query only needs one hash, map and the main body of transmission only contains 28 float numbers.

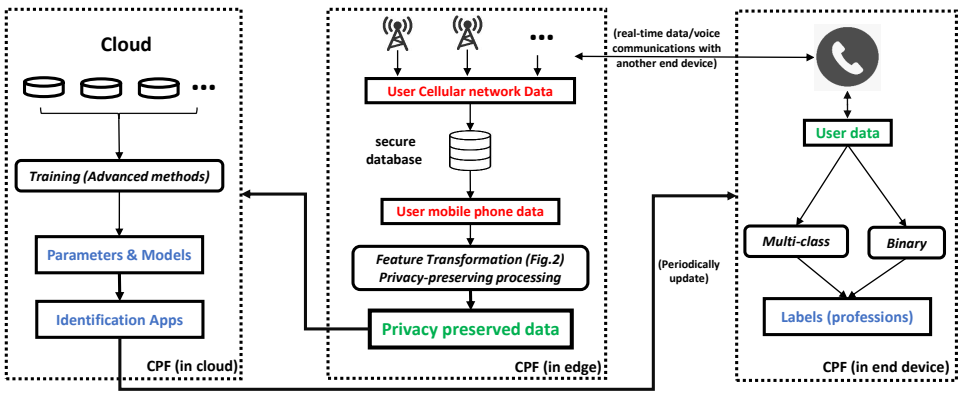


Figure 4.6: CPFinder implementation: an overview.

4.4 Evaluation results and discussions

4.4.1 Dataset

The dataset of user mobile cellular data comes from a major telecommunication operator in China. Each record contains three fields: phone number, time stamp, network request. The data contain 1,282 users in Shanghai City and cover a period from November 1, 2016 to February 16, 2017. In the total 108 days, call records during the 35 Chinese public holidays are excluded due to their different nature from ordinary days. For the phone label data, if a phone number is labeled, the annotation information could be easily found by searching the phone number via Baidu search engine (Fig. 4.2). An annotation is the detailed description of the phone number, which directly indicates the profession of the user. In addition, each labeled number would be assigned with a root annotation, which is formatted textual data containing several different types of professions. Originally, the online datasets typically contains the following annotations: fraud, harass, illegal, insurance, finance management, intermediary, recruiting hunter, takeouts, delivery, driver, and customer service. These root annotations could be divided into three categories: taxi drivers, delivery, harass. Delivery includes express delivery and takeouts delivery, whereas harass is a general designation of harass, fraud, and telemarketing. For numbers that are not annotated, they are regarded as normal users. The label crawling is done by the telecommunication operator within their internal server for privacy concerns, and the phone numbers are replaced by the categories of professionals for later processing.

Table 4.1: Phone labels distribution within the data set

Group Number	Group Description	Group description	Number of users
1	Normal	Normal users	591
2	Drivers	Taxi drivers	176
3	Delivery	Delivery & Takeouts men	183
4	Harass	Telemarketing & Fraud people	332

Although there are only 1,282 users, each user has 73 days of data and could generate up to 73 different instances. However, the instances of a single user are only used in the training or testing phase because instances of a same user may show significant similarities and will seriously impact the performance of the models. If a testing instance belongs to a user, whose other instances were used in the training phase, the testing instance will be correctly identified with a high probability. We seek to find a general pattern for individuals of the same profession, rather than particular individuals. As a result, applying the instances of an identical person both in the training and testing phases will not make much sense, and a better testing data selection scheme should be based on persons rather than instances. Several regression and classification algorithms are applied to evaluate the performance of the model. Each algorithm is fine-pretested to present its best performance. Tab. 4.1 shows the distribution of user quantities in different categories. Considering the unbalanced distribution of the user quantities, 50 users are randomly selected from each category; a total of 200 users are formed for testing.

4.4.2 Evaluation results: multiclass classification

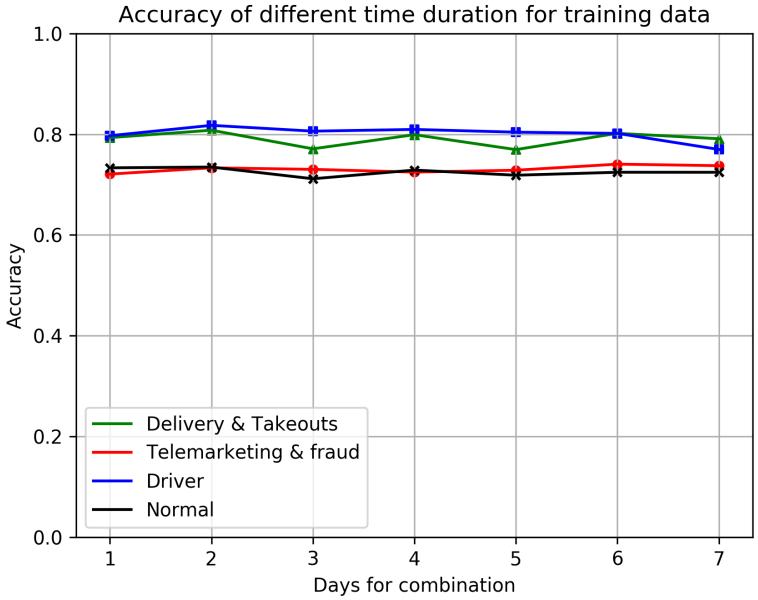
To test the performance of our method, we adopt several classic and state-of-the-art machine learning algorithms, e.g., Logistic Regression (LR), K-Nearest Neighbors (KNN), Support Vector Machine (SVM) with Radial Basis Function (RBF) kernel, Random Forest (RF), AdaBoost, and Neural Network (NN). Tab. 4.2 shows the identification accuracy of all algorithms, accuracy of each category of people is also presented. Each result is an average of 100 experiments with cross validations. Except the LR, other classification algorithms perform very similarly. Random forest could achieve the highest overall accuracy of 75.64%. The taxi drivers, delivery and takeouts, are more accurately identified, whereas the identification accuracy for the other two groups are slightly lower. The result shows our approach outperforms the performance of the approach presented in [126], whose accuracy is 70.4% for unemployment identification (and lower for exact profession identification).

Table 4.2: Identification accuracy of different categories and different methods

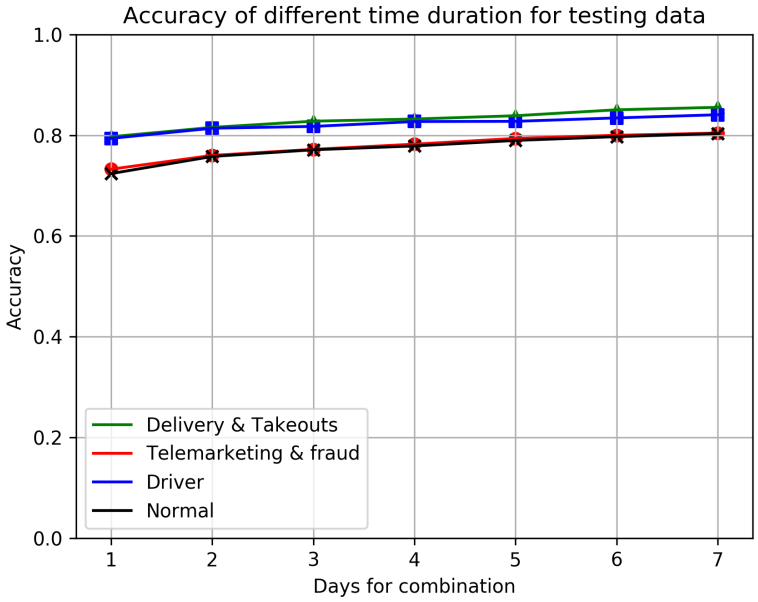
	Normal	Driver	Delivery	Harass	Overall
LR	0.5677	0.6284	0.6257	0.6113	0.6083
KNN	0.6970	0.7771	0.7705	0.6864	0.7328
RF	0.7300	0.7912	0.7884	0.7160	**0.7564**
SVM-RBF	0.7062	0.7625	0.7633	0.7112	0.7358
ADABOOST	0.6865	0.7404	0.7510	0.6888	0.7167
NN (MLP)	0.7216	0.7505	0.7366	0.7142	0.7307

However, because a user's data on each single day is not always complete, merging the data of several days together may influence the accuracy. Therefore, we also investigate how time duration impacts the performance of the model by combining data from different numbers of days. The combination is not the average of the statistical parameters from different days, but is the one-time calculation of all data within these days. For example, the mobility pattern of a two-day combined data is the range of all locations during the two days. We study the impact of combining data in the training phase, while the test data remains the duration of one day. Reversely, combining testing data while the training data remains uncombination is also investigated. Fig. 4.7 shows the identification accuracy of combining training data over different time lengths.

The identification performance does not change too much when the number of days increases for data combination, which means combining data for training phase makes no much sense. However, the accuracy slightly improves when the number of days for data combination increases. The improvement is not significant even when comparing the accuracy of combining 7 days of data together with non-combination data. In general, we can see from our data that using the data of a single day is sufficient to identify a strange caller's profession with a good performance, even though adding data of more days could somehow improve the accuracy.



(a) Training data combination



(b) Testing data combination

Figure 4.7: Accuracy of different time duration for training (a) and testing (b) data.

The latitudes and longitudes extracted from mobile cellular network requests could indicate the exact positions of the users, but the dataset could not fully contain all locations or trajectories of a user because not each network request contains location related information. Generally, as users could control the access of location services for each APP, and most people prefer to turn off the location services for most apps except map-related applications such as ride-hailing applications (e.g., Uber and DiDi). The incomplete location information brings the bias between the actual and recorded patterns of a user, which turns out to influence the performance of the proposed method. This is why the accuracy slightly improves with increasing the number of days for data combination. In a long period of nearly 3 months, some users may change their professions. These changes negatively impact the accuracy of identifying callers' professions, and it is hard to verify how serious is the impact because how and when people changes their professions are not recorded.

Moreover, the deviation of personalities between individuals of the same label also negatively influences the performance of the model. This challenge holds true for most personality-related studies, especially for multiclass classification problems [126, 27, 88]. Overall, CPFinder achieves relatively good performances based on such limited information and outperforms some previous methods.

4.4.3 Evaluation results: binary identification

Mobile phones users typically have some prior knowledge of incoming phone calls. If a person orders a taxi and receives a call later from unfamiliar phone numbers, he/she just needs some information about whether the caller is a driver from the taxi company that he just ordered. In addition, there is another scenario, that users may need identification quickly upon receiving a call and knowing some basic information from the caller, e.g., the caller states as a driver or delivery staff. In such scenarios, a binary identification rather than multiclass classes classification can adequately meet the demands of providing information to identify the callers. Therefore, for a specific category of callers, a binary model regarding the other categories as negative labels can be created to serve as special binary identification system, which typically requires quick and accurate identification. At the moment, each category should have its own model and trained individually with the data. For demonstrating the advantage of the method, we compare the results with the work of using CDR data to identify harass calls (DeMalC) [79] and using real-time mobile phone data to prediction unemployment and professions (TRP) [126]. Based on the results in [79], we adopt exactly the same algorithms, that is GBDT [98] for training the models. All metrics are compared, including precision, recall, F1-score and overall accuracy, between our methods and others'. As there are no other metrics except the accuracy in TRP, only the overall accuracy of unemployment prediction is compared here.

To better demonstrate the comparison between our method and some previous studies, several evaluation metrics will be introduced. For the binary classification problem, the classification results consist of four classes, based on ground truth labels and predicted labels.

True Positive (TP) is the number of correctly classified positive samples, while False Negative (FN) is the number of misclassified positive samples. True Negatives (TN) is the number of correctly classified negative samples while False Positive (FP) is the number of misclassified samples. The classified results in a confusion matrix for binary classification problem is shown in Table 4.3.

Table 4.3: Confusion matrix for the binary classification results

	Predicted Positive	Predicted Negative
Actually Positive	True Positives (TP)	False Negatives (FN)
Actually Negative	False Positives (FP)	True Negatives (TN)

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{4.2}$$

$$Error = \frac{FN + FP}{TP + TN + FP + FN} = 1 - Accuracy \tag{4.3}$$

The very Naive metrics used to evaluate prediction performance are accuracy and error, whose summation is strictly one. However, they can not precisely evaluate the classification performance in variety conditions, such as when the data is imbalanced. As a complement to accuracy, there are several metrics that are much more powerful than it, such as Precision, Recall, and F1-score:

$$Precision = \frac{TP}{TP + FP} \tag{4.4}$$

$$Recall = \frac{TP}{TP + FN} \tag{4.5}$$

$$F1 - score = \frac{2 \cdot precision \cdot recall}{precision + recall} = \frac{TP}{TP + \frac{1}{2}(FP + FN)} \tag{4.6}$$

First, CPFinder outperforms the DeMalC method when used for identifying harass calls. DeMalC is based on CDR data and some explicit information about mobile devices and calls, including IP address, base station tower locations, and call frequencies. As mentioned above, the data used for the inputs of this method contain much plaintext of sensitive personal information, which could be used to trace a person's trajectory or retrieve personal behavior. Moreover, using such data requires the approvals of the data owners. Although our method outperforms DeMalC for identifying harass phone calls with just around 0.01 accuracy improvement, it contains much less features and variables: CPFinder contains only 3 features and 28 variables in total, whereas DeMalC needs 7 features and up to 190 variables. Obviously, less variables could lead to higher efficiency in terms of shorter training time and lower identifying speed,

as well as the faster transmission of the data. Our method also adopts the privacy-preserved features, hence the input data could be transmitted to end devices or third-party servers to further speed up the computational efficiency. The performances of two other categories are better than the category of harass and could achieve an accuracy more than 0.93 for driver identification; TRP also uses CDR data to generate feature-rich models for predicting personal employment status and professions. The best result TRP could achieve for predicting binary status (employed or unemployed) is 0.704. Moreover, the overall average accuracy of predicting personal professions is 0.675, which is worse than our original model used for multiclass identification.

Both DeMalC and TRP contain a large amount of features (up to 160) and variables, which requires high computational complexity. As is well known, the time complexity of decision tree is $O(m)$, whereas for XGboost algorithm, it is $O(m^2)$, where m is the number of the attributes. As a result, our method will be more than six times faster than DeMalC and more than five times faster than TRP if applying algorithms with linear time complexity, and the time efficiency will be quadratically increased if algorithms with $O(m^2)$ time complexity are applied in the model.

In addition to the overall accuracy, the precision of our method also outperforms the counterpart methods. The overall precision of our methods of all the three categories is around 0.85, where left a false positive rate of around 0.15. The false positive rate means for each positive identification, it is actually not a positive sample with a probability of 15%. In whatever conditions, misidentifying negative to positive is a big problem, e.g., identifying normal phone numbers as harass phone calls may make user miss important calls. However, reversely identifying harass phone calls as normal calls is somehow less severe than missing important calls. Currently, the methods used in the paper are optimized by maximizing overall accuracy. If precision is more preferable than accuracy, the optimization function could be changed to maximize precision, where the false positive rate will decrease but the overall accuracy will also decrease. Therefore, there is a consideration of the balance between precision and accuracy.

Table 4.4: Experimental results of binary classification

	Precision	Recall	F1-score	Accuracy	Features
Driver (GBDT)	0.8672	0.8012	0.8329	**0.9351**	
Harass (GBDT)	0.8338	0.7935	0.8132	**0.9287**	28
Delivery (GBDT)	0.8476	0.7790	0.8119	**0.9235**	
DeMalC (GBDT)[79]	0.8395	0.7521	0.7934	0.9186	190
TRP (DNN) [126]	-	-	-	0.704	160

4.4.4 Capability of privacy-preserving

We compare our approach with other works (e.g., DeMalC and TRP) in terms of privacy preservation of user data. In addition to the advantage of reducing the dimensions of the input features and computational complexity, adopting privacy-preserved data in the training

phase could also bring other benefits, such as enabling end devices identification and looser requirements for data arrangements.

Prevent re-identification of personal data

According to the study in [133], mobile phone data could be used to identify a person. In other words, a person could be uniquely re-identified in the crowd by the apps he/she uses and the locations he/she visits. However, the identification requires investigating a few frequently used apps by knowing their names and how they operated in the devices. In terms of locations, with additional information of coordinates, a person is entirely exposed to the public if such kind of data are disclosed. Therefore, using original sensitive personal data brings a lot of risks on people's privacy, even in the training phase. So obviously, the original personal data should not be distributed outside the secure data centers to end devices. However, privacy-preserved data do not support identifying exactly who a person is, but only some non-private attributes, e.g., professions. DeMaLC requires city level locations of the data, whereas TRP requires some information of a person's home district and charge amount of each account. Consequently, the attributes of the two models are not in compliance of data protections when the data are distributed.

Privacy considerations on data acquisition and processing

According to various data protection regulations and policies (e.g., Art. 5 GDPR), collecting personal data for any purpose of processing should be conducted under the consent of the data owners. Consequently, the data used in the methods of DeMalC and TRP, as well as the data in many studies that use personal mobile phone data for various purpose, are mostly declared to be used under the agreement of the data owners. Although the data used in the case study in this paper also complies with the consent of the users, our model could be further generalized to the data without the users' additional agreement. Because users' requests with encryption through the network are public to all people, our approach just needs the information of some frequently used apps without knowing their exact names. For example, the name of apps could be hashed and captured by the network edge devices, the same hash value indicates a unique app and the data is somehow public. As a consequence, the users' preferences of apps could be collected and processed in a privacy-preserving manner without violating data privacy regulations. In contrast, as the name of the apps could not be reversely retrieved from the encrypted requests, the results obtained in previous methods relying on data containing plaintext of personal data, are not easily reproducible, since they are not applicable with privacy-preserved data.

4.5 Chapter Summary

We present CPFinder, a new methodological framework and system for identifying a mobile phone caller's profession using mobile cellular data, i.e., users' cellular network traffic recorded by telecommunication operators. Three kinds of privacy-preserving features are

constructed based on the information provided by the cellular data: mobility pattern, data volume distribution, and apps preferences. All features are formed by several statistical parameters, which exclude sensitive information such as coordinates and apps log. These private information could be used to reversely trace people's locations and living habits, thus eliminating the private information, even for input data, could avoid violating any data privacy regulation.

We apply several state-of-the-art classification and regression algorithms in our model; the experimental results of a dataset containing 1,282 users in Shanghai City prove that CPFinder could achieve an accuracy of 0.7564 (against the alternative methods that achieve accuracies up to 0.7358) when identifying four different categories of callers: normal users (other professions), drivers, delivery and harass. If binary classification is applied to the same dataset, CPFinder achieves an accuracy of 0.923+, outperforming two existing approaches with accuracies of 0.704 (TRP) and 0.9186 (DeMalC).

Combining data of different time lengths is also investigated for both training and test phases to study how the amount of data impact the performance of identification. The result demonstrates that data of a single day are sufficient for identifying a caller's profession, even though the accuracy slightly improves when using combined data of several days. Again, the overall performance outperforms some previous studies in terms of both accuracy and efficiency.

Chapter 5

Mining community differences in computer science conference

Academic collaborations become regular and the connections among researchers become closer due to the prosperity of globalized academic communications. It has been found that many Computer Science (CS) conferences are closed communities in terms of the acceptance of newcomers' papers, especially are the well-regarded conferences [24]. However, an in-depth study on the difference in the closeness and structural features of different conferences and what caused these differences is still missing. Previous studies of coauthor networks did not adequately consider the central role of some authors in the publication venues, such as PC chairs of the conferences. Such authors could influence the evolutionary patterns of coauthor networks due to their authorities and trust for members to select accepted papers and their core positions in the community. Thus, in addition to the ratio of newcomers' papers it would be interesting if the PC chairs' relevant metrics could be quantified to measure the closure of a conference from the perspective of old authors' papers. Additionally, the analysis of the differences among different conferences in terms of the evolution of coauthor networks and degree of closeness may disclose the formation of closed communities.

In this chapter, using the Digital Bibliography & Library Project (DBLP) dataset of computer science publications and a PC chair dataset, we show the evidence of the existence of strong and weak ties in coauthor networks and the PC chairs' influences are also confirmed to be related with the tie strength and network structural properties. Several PC chair relevant metrics based on coauthor networks are introduced to measure the closure and efficiency of a conference.

Contents

5.1 Community mining in computer science society

It has been reported that computer science conferences are closed communities as they have very low ratio of newcomers' papers every year, especially are the top level conferences [24]. However, using newcomers' papers alone to characterizing the closure of a conference is not sufficient, it would be interesting if more metrics such as structural properties of the collaboration network, relations between the existing authors' papers and the communities, as well as PC chair relevant metrics are considered. Adopting more metrics to characterizing the closure of a conference could help distinguish different degrees of closed communities among different conferences. Moreover, temporal analysis of these metrics may provide the perspective that what factors lead to different structural properties in co-author networks. With the rapid growth of scientific collaborations due to academic globalization, the prosperity of co-authored publications makes it possible to access large digital library data and analyze huge researcher collaboration networks. The structure of relations among researchers is becoming more and more complicated and it is important to know how the academic communities evolve and what affects the paper acceptance in addition to the quality alone. Typically, when researchers attempt to submit their works, especially when considering papers that may be of marginal quality, they may evaluate the possibilities of getting them accepted among different conference or journal options. The typical factors of such evaluation include conference's ranking (e.g., China Computer Federation (CCF) ranking [1], CORE ranking [2], CSranking [3]), acceptance rate, the fitness of the topics and the paper style preferences of the conferences, which are easily justifiable for authors. Now, the closure of the conference can be a new and even powerful factor.

However, as an important branch of social networks, the coauthor networks formed by scientific collaborations preserve the patterns that the connections among the people could influence personal behaviors [109]. Previous studies show that researchers tend to cite papers from their collaborators [13]. Similarly, in addition to the quality of the paper and the factors about conferences, will the collaboration ties among the authors could also influence the paper acceptance? Especially, when a special role (i.e., program committee chairs, who are mostly in the core positions of the community and with a high level of authorities to decide the lists of accepted paper) are taken into considerations, the statistical analysis of coauthor networks are no longer restrained within the homogeneous nodes. Intuitively, PC chairs could influence the evolution of the coauthor networks and bring subsequent effects, however to the best of our knowledge, such influence has not been quantitatively studied before.

Moreover, strong tie and weak tie theory [52, 85] has been developed for almost 30 years which studies the strength of the interactions between two nodes. But few studies explore the

[1] https://www.ccf.org.cn/en/Bulletin/2019-05-13/663884.shtml
[2] http://portal.core.edu.au/conf-ranks/
[3] http://csrankings.org/

applicability of these theories in the paper selection process. Can weak ties exert strength in selecting papers and benefit the whole community? What role do weak ties and strong ties play in the evolution of community structures? Reviewing the weak tie theory, Granovetter[52] indicated when people acted like a bridge to span different separate groups and there was a "forbidden triad" (not all members connect each other) in the network when weak ties exist. Thus, there is a situation that the weak tie could not require an actual link between the nodes. Predicting the probability of the presence of a link between the nodes that are not connected is a kind of detecting invisible weak ties problem. A classic weak tie definition is that if two unconnected nodes are both directly connected to a same node, there will be a greater-than-chance probability that these two nodes will be connected after enough time. Therefore, this kind of connection can bring heterogeneous diverse information and broaden the size of communities. After that, Burt[22] proposed the structural hole theory to illustrate that the existence of weak ties can strongly boost the network effectiveness and efficiency, which can be called brokers. They are not constrained by certain connections and have a high betweenness. Existing work shown that a long-range tie which span large social networks and is "assumed to be weak, composed of sporadic and emotionally distant relationship, can strength the diffusion and social integration of network[95]. In coauthor networks, despite the relations of collaborations, the paper selection relations between PC chairs and the papers they select may present some weak tie patterns because most of the papers do not contain a author who is directly connected to the chair in terms of coauthor relations. They must have some connections which are indirectly have coauthor-ship with them. Different PC chairs have different preference to maintain weak ties of different proportions. If these potential weak ties are proved to exist, then the existence of PC chairs' influences on the paper selection is also confirmed. It will also be very interesting to detect how the influences change with the weak tie strengths. The strong ties, are obviously the direct connections between two authors that jointly published a paper. Previous studies [70] highlighted that strong tie - a kind of emotional friendliness between leaders and their employees within the organization had significant benefits to companies. In this situation, the strong ties between chairs and their collaborators could also have different strengths in terms of their positions in the network and their connections with other researchers, not just the times of collaborations.That is we should not neglect the fact that there is a PC chair ego-centered network structure exist and the different intimacy between chairs and other authors. Existing researches [62] [31][41][44]have explored the dynamic process of trust-building in academic communities' partnership sustainability in health research. However, these works only focus on the building and maintaining of the trust relationship and the positive benefits of trust-building and pay little attention to the weak tie effects to the communities. [47] proposed that strong ties can promote the emergence of elites in cooperation networks. Also, researches proved that there are negative effects of a high trust network that is composed of strong ties merely. Members are closely connected and easier to form an information cocoon to harm innovation and efficiency [73][112]. On the other hand, if the communities have a very low level of trust or no trust, the efficiency of information dissemination in this community will be extremely low and the cooperation between members will be hard. Thus, a polycentric small-world network with moderate trust is more efficient

and processes a higher ability to develop and update knowledge[68]. Therefore, predicting and describing the evolution of weak ties and strong ties is helpful for us to better grasp the dynamic evolution process of academic communities.

As a fast evolving researcher area, the computer science society pays more attention on the conferences than journals. There are tens of broader fields in computer science, each having tens to hundreds of known international conferences today. According to a statistics of DBLP [132], a total number of 5,295 conferences are included in the database till the year of 2020. Meanwhile, the yearly published number and total number of conference and workshop papers has reached 216,299 and 2,736,712 in 2020 (in contrast to 109,911 and in 2010), more than doubled in last 10 years. And the increasing trend will most likely continue according to the records of last decade. Notably, the quality among these conferences differs significantly. As a result, many institutions and organizations have different rating standards that divide conferences into different levels according to their publications' quality for their recruitment, promotion, budget allocation, or university/departmental ranking purposes. A rough consensus in the computer science community is that high level conferences are relatively harder to enter because of their rigorous requirements on the quality of the accepted papers. However, despite the quality differences, other factors that differentiate conferences have not been investigated, such as the collaborations among the researchers, the communities within related conferences, as well as the evolutionary patterns of the networks. So far analyzing the differences among different conferences in the view of coauthor network is missing. A deeper analysis of the complex network based differences among conferences could provide useful hints for both authors and conference organizers. The authors could consider more dimensions of factors when choosing a proper conference to submit their works, while the organizers could trace the evolution of communities to prevent the conformation of locally dense groups and provide a fair environment for the broader research community.

The scientific reputation, citation index, H5-index, the quality of accepted papers and keynote speakers are typical factors when rating the level of a conference. High quality publications, prestigious committee members and chairs, high citation index, are common advantages of top level conferences. Despite these obvious and well known factors, the analysis on some "hidden" features of conferences which drive the evolution of conferences may also help authors and organizers to make appropriate decisions. For example, an interesting aspect is the patterns of communities inside a given conference and how they interact with each other. If the coauthor network of a conference has a large size of Giant Component (GC), it is critical to find out how this strong connectivity is formed. The collaboration patterns among authors and the paper selection process could both affect the evolution of the coauthor networks and result in different network structures. The influences could be quantified by some network based metrics, which could clearly disclose the particularity of selecting accepted papers. And when taking PC chairs into consideration, if the paper selection of a conference is proved to have strong relations with the PC chairs and the community pattern, its overall acceptance rate of the papers may be no longer that important to some authors. For example, a stable ratio of papers from the chairs,

the chairs' former collaborators or authors that are already inside the core community will also decrease the acceptance rate for those authors who are far way from the giant component in the conference.

In this paper, we analyze a total of 334 conferences in computer science selected from one of the well-known conference rankings, CCF ranking that rates the conferences into 3 levels: A, B and C. Firstly, we compare the evolutionary and structural differences in terms of coauthor networks. And try to trace the formation of weak ties and strong ties. Our results suggest that there are significant different structures among different levels of conferences, including the ratio of newcomers' papers, the community pattern, the connections inside the giant component. We also find top conferences are more likely to form a huge and fully connected communities, and the connections inside the giant component is also denser in top conferences, which means they are denser communities and have a high possibility to form elite small-world networks. To determine what makes the significant differences, we introduce several network based and PC chair based metrics to measure the relations among papers, authors and PC chairs. The PC chair related metrics also disclose the paper selection differences among the conferences, where several conferences are more likely to select papers (co-)authored by closely connected researchers. These quantified paper selection metrics provide useful information for authors to choose a conference that could most probably accept their papers, in addition to the common considerations. For the conference organizers, these metrics would be a warning for them to avoid the bad influences of high trust and adjust the policies to introduce more preconditions for the formation of weak ties to prevent the conferences from serious PC chairs' influences.

5.2 Is the CS conferences a closed or an elite small-world network community

5.2.1 Newcomers' papers

Researches[24] used DBLP data to prove that computer science conferences are closed communities as the acceptance of the newcomers' papers (all the authors of a paper are new to a conference) is very low among CS conferences, especially in top level and well-regarded conferences. However, they only analyze the index of newcomers' papers and based on static data in 2017 and are there more forms of community exist? Here we borrow the notion of the ratio of newcomers' papers and check how closed the CS conferences are in a dynamic perspective, as well as the differences among different levels of conference.In the meantime, this paper tries to introduce different perspectives to distinguish whether the community is closed or not.

Based on the different levels of trust between members and the composition of strong and weak ties in communities, we define three kinds of communities. The first is that there are only strong ties with high density and few weak ties in the community, which is called a closed network. A closed community formed by some researchers who have a close cooperative

relationship and high trust within the group and the conference has a relatively low acceptance rate. The second kind of community is that there are a few strong ties but many weak ties and long-distance bridges[138] in the network. The average path length of the network is low and there are structural holes [22] and elites (such as PC chairs) in the network. We define this kind of network as an elite small-world network community. Although this kind of conference community has a low acceptance rate, the quality of papers may higher than the closed communities. And because of the existence of weak ties, the size of the network continues to expand. The third is called an amorphous community which is composed of many unconnected nodes and a lack of trust in the network. Therefore, there is no stable community structure in this kind of network. Such communities most likely appeared in the early stage of the formation of the first two kinds of communities.

Here, a well-known conference ranking system defined by CCF is applied here in order to check the differences among different conference levels. For each conference, its yearly ratio of the papers from newcomers are calculated and sorted by time, which naturally started from 1 because all the papers are new to the conference when it is firstly held. Similarly, only the papers from main tracks are counted since short papers, posters and demos are typically easier to be accepted. Additionally, other community based metrics such as giant component and average degree are used to demonstrate the different closures of the conferences in detail.

As shown in Fig.5.1, since the start of a conference, its acceptance of newcomers' papers is really high regardless of its level. All conferences show a highly similar evolution of the ratios of newcomers' papers in the first several years (i.e. logarithmically decrease over the time). Later, the ratio of the newcomers' papers of level C conferences stop decreasing and keep a stable ratio around 0.3, but for higher level (Level A and B) conferences the ratios of newcomers' papers continue decreasing at low speed. Specially, level A conferences even reach a very low ratio of newcomers' papers that is less than 0.1, which indicates the threshold is higher than other conferences and maybe a serious closed with high trust community or an elite small-world network community with moderate trust is forming among new researchers. The result is partly consistent with the analysis in [24] that the median value of the ratios of the newcomers' in top conference is 14%. However, we need to compute other network properties and based on papers' quality to figure out which kinds of communities these conferences are.

5.2.2 Co-author network, Giant component, Average degree

In addition, in order to explore whether there is a closed community or an elite small-world network forming in higher level conferences, we analyze other structural metrics of authors' collaborations pattern, that is giant component and average degree. For each conference, the authors who ever published any paper could form a co-authorship network. The networks only contain the collaborations with this conference without any outside relation. The network structures may not be similar among different conferences or even differ significantly among

Table 5.1: Coauthor network relevant metrics

	Names	Descriptions	Symbols
Network metrics	Giant Component	The largest connected sub graph of a network	p_{gc}
	Average Degree	Average number of links per node	$< k >$
	Shortest Path Length	Shortest hops between a pair of nodes (v1, v2)	$Distance(v1, v2)$
	Average Shortest Path Length	Average shortest path between each pair of nodes	-
	External-internal index	density of largest component/ density of entire community	EI
	Q value	cluster coefficient (CC)/ average path length (PL)	Q
PC chair metrics	Paper-chair distance	Average reciprocal of all distances between authors and chair	$Rd_{paper-chair}$
	Paper-community distance	# of the authors already in the giant component for a paper	$d_{paper-community}$
	Chair associated paper index	Probability of a paper contains 1+ collaborator of the chair	P_{cap}
	Chair paper index	Probability of a paper contains the chair him/herself	p_{cp}
	Chair betweenness centrality	Node importance (betweenness) in the network	$c_B(chair)$

them. The network structure could be characterized by several metrics, and the metrics could provide some hints on how conferences select submissions, what is the relation between authors and committees and what makes the conferences different from each other. To present the structural characteristics of the conferences that could best uncover the differences among them, we firstly check several community related parameters to find out how the communities look like and evolve.We aim to find out the diverse community structures of different conferences in computer science, and quantify the differences by both traditional metrics used in graph theory and newly designed metrics that take PC chairs into considerations. The following are the notations used in this paper and their explanations.

Giant component size, average shortest path length and average degree of nodes are basic metrics of graph theory, which could characterize the structures of a complex network. The size of the giant component measures the connectivity of a network, a fully connected network's p_{gc} equals to 1. The average degree and average shortest path length could jointly indicate the density and internal communicative pattern. In coauthor network, the low path length and high degree are resulted from a densely connected community, where the collaborations among the authors are very frequent. These traditional metrics are implemented by Networkx [55], a famous network analyzing library.

As for the PC chair related data, paper-chair distance measures the relation between the authors and the chairs. For a conference in a specific year, there will be an entire coauthor network for the conferences authors, and there will be a shortest path between each author and the chair if they are connected. The chairs' preferences of selecting papers from their personal communities could be quantified by these distances. Chair associated paper index ignores the distances between the chair and the authors, but counts the number of papers coauthored by at least one direct collaborator of the chair. The detailed description of these chair relevant metrics is given in Section 6. The PC chair data is collected from the official websites of the conferences and matching operations are applied to solve the inconsistency between the DBLP and the web data. Finally, a set of 3,697 PC chairs of different conferences (level A and level B) and different years is constructed.

A graph has two phases in terms of connectivity, fully connected or not. If a graph is not fully connected, it means there are pairs of nodes that no paths could connect them. As a result, the whole network is decomposed into several connected components.

The most important parameter is the distribution of the size of the giant isolated component. The larger the giant component without any connections with other small components are, the more centralized the network is. If the distribution of the size of the giant component is even, the network is decentralized and may be dominated by several different communities, where checks and balances exist. Even better, if there are a small number of weak ties between different components, the different components are not completely isolated. To present the size of the giant component of each conference, we construct each of them a coauthor network including the authors since the start of the conference till the year 2020. As shown in Figure 5.2, a total of 334 conferences within three levels are extracted and their sizes of the giant component of their coauthor networks in the year 2020 are sorted from the least to the largest.

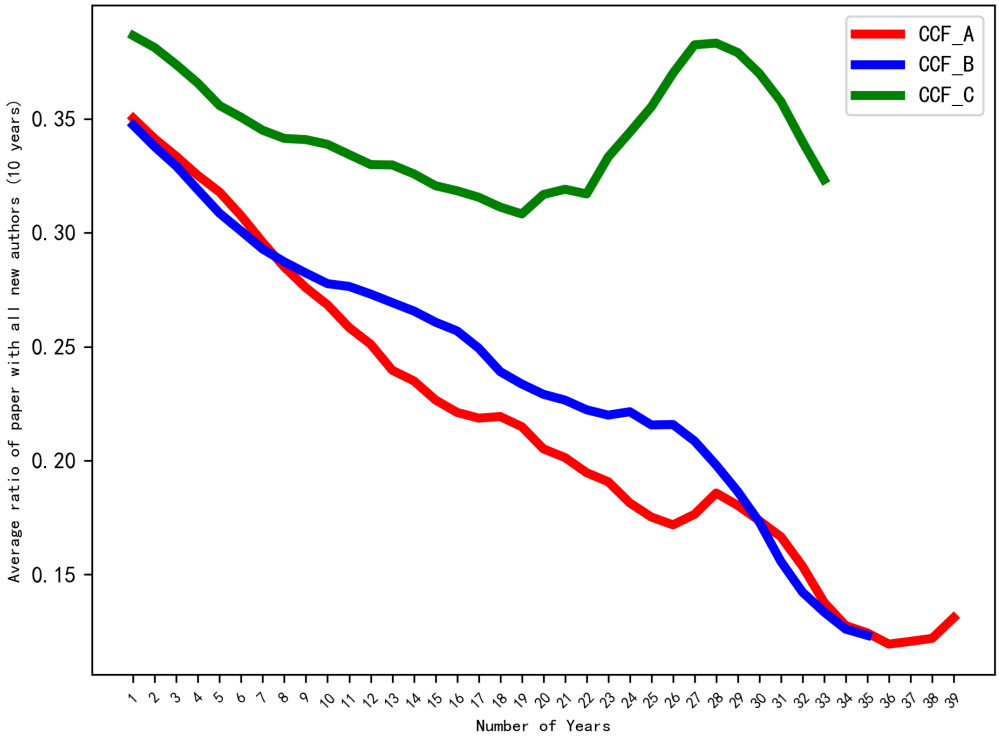


Figure 5.1: Ratio of newcomers' papers evolves with time for different levels of conference.

Similar to the ratio of newcomers' papers, almost all CCF-A conferences have a large size of giant component. The average size of giant components by conference levels shown in the subfigure in Figure 5.1 proves that higher level conferences are more likely to form a big community in terms of collaborations among the authors. In addition to the giant component, the average degree of each network shows similar pattern to giant component size, where the number of connections among the authors increase with the conference level. The large

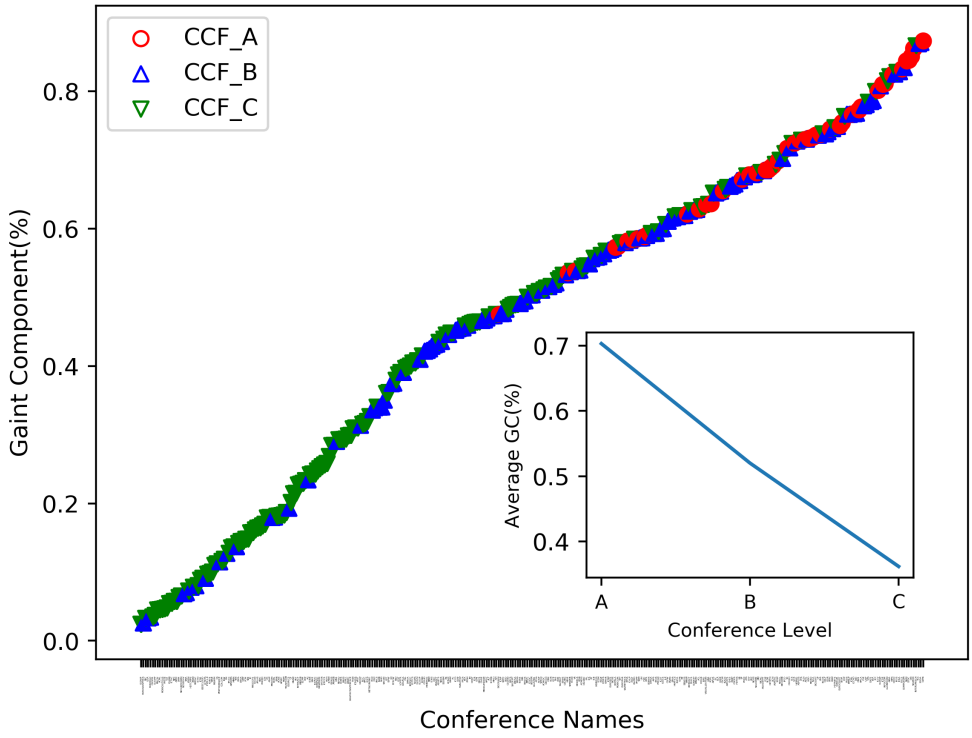


Figure 5.2: The giant component size of 334 conferences in the year of 2020.

size of a giant component and dense connections among authors indicate a tight and strong communities inside the top level conferences. To understand how the communities are formed in top conferences, we need to analyze the evolution process of the conferences' author networks. Fortunately, conferences have a strict temporal stream since they are typically organised every year or once another year. In addition to the static networks, Figure 5.4 shows the average size of giant component of the conferences of a same level from the year of 1981. At the beginning there are not visible differences among different levels of conferences, but the communities grow differently with the time. Top conferences formed their communities rapidly after early 1990s and the growth of the giant components is almost linear with a stable speed. However, for other levels of conferences, their sizes of giant component increased almost linearly with time and grew fast in last 3 decades. It is easy to imagine that there is a threshold of the giant component for each conference (i.e. less than 1) and the trend of the increasing size of GC will decelerate in the future. But currently, averagely the GC of the conferences will still increase in the coming a few years based on the data end in 2020. Figure 5.5 shows the evolutionary process of average degree for different levels of conferences. Similar with the evolutionary pattern of GC, the average degree of top conferences increases quickly than that of lower level conferences. We also fit a straight line for each level of conferences, which shows that top conferences significantly have formed densely connected communities and will become denser in the future.

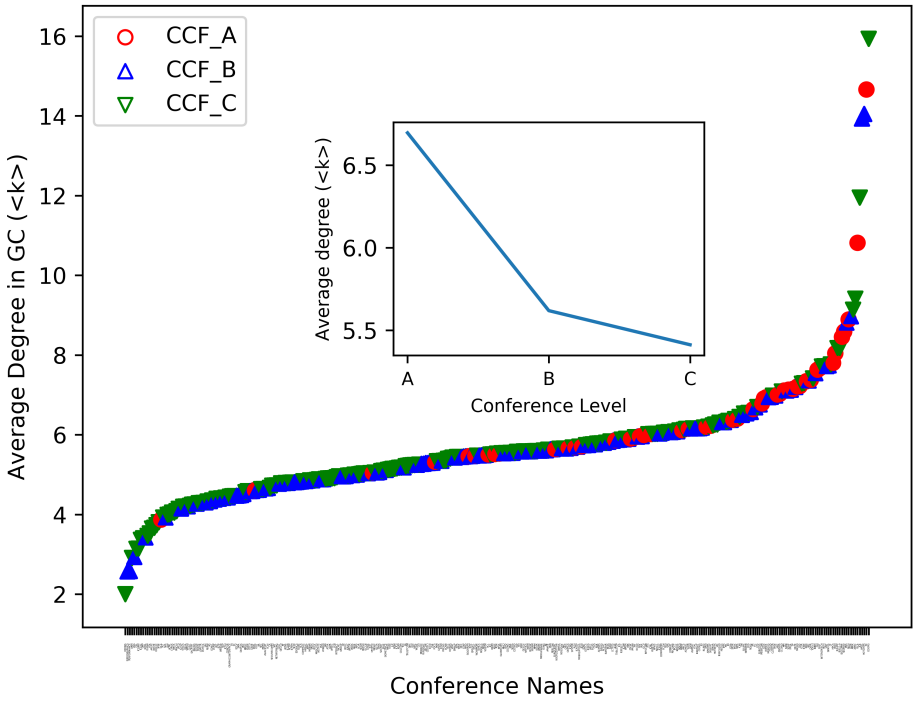


Figure 5.3: The average degree of the giant component in 334 conferences in the year of 2020.

5.2.3 Shortest path length, small world, strong and weak ties

The average degree and average shortest path length could jointly indicate the density and internal communicative pattern. In the coauthor network, the low path length results from a densely connected community with more weak ties and long-distance bridges exist, where the efficiency of information transfer among the authors is very high. In this kind of community, members can communicate with each other through limited links. In addition, two small-world network metrics are also calculated to measure the connections among different local communities, which are External-internal index (E-I index) and Q value:

$$E - I = \frac{\text{Density of max component}}{\text{Density of network}} \tag{5.1}$$

$$Q = \frac{\text{Cluster Coefficient}}{\text{Average Shortest Path}} \tag{5.2}$$

$$\text{Cluster Coefficient for a node} u = c_u = \frac{2T(u)}{deg(u)(deg(u)-1)}, \tag{5.3}$$

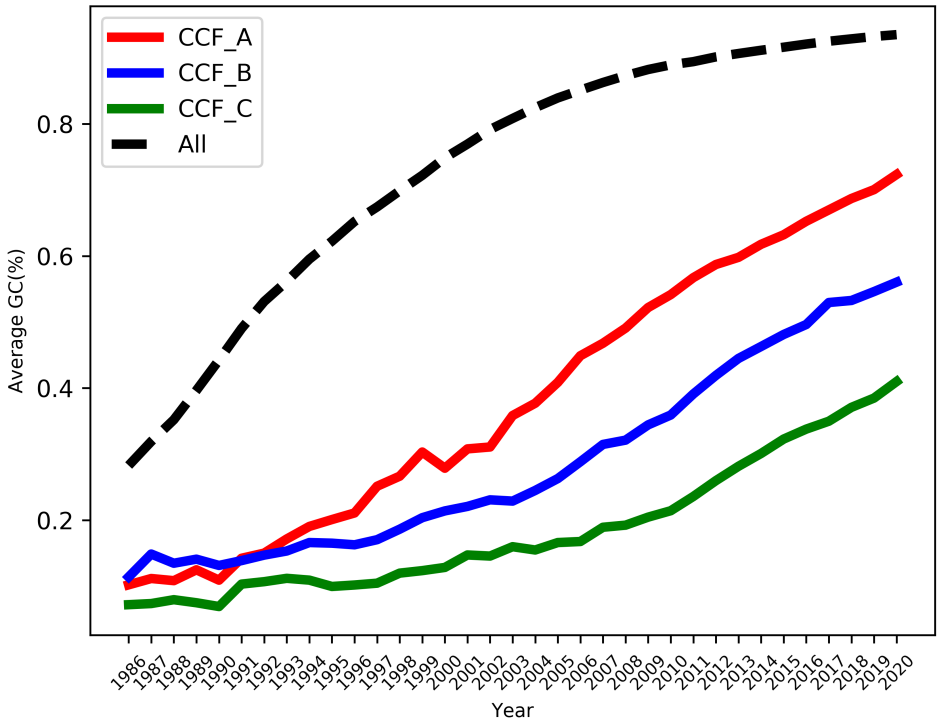


Figure 5.4: Giant component size evolves with time for different levels of conference.

where $T(u)$ is the number of triangles through node u and $deg(u)$ is the degree of u.

$$\text{Average Shortest Path} = \sum_{s,t \in V} \frac{d(s,t)}{n(n-1)}, \tag{5.4}$$

where V is the set of nodes in network, $d(s,t)$ is the shortest path from s to t, and n is the number of nodes in network.

Therefore, an elite small-world network community has a low path length value although the network size is big. On the contrary, the closed network consists of strong ties with higher path length value when the network size is big or a low value in a small community. Therefore, we need to compute the proportion of weak and strong ties to determine which type of community is. In Table 5.2, several classic network structural metrics are calculated and averaged by conference levels. All these metrics could jointly describe the network structural properties and a comparison among them could help explore the structural difference among different levels of conferences. First, for GC and degree, top conferences have large giant components and densely connected communities, which means communications among authors in top conferences are very efficient and effective as most pairs of authors could reach each other and each author has many authors with others. Then for average shortest path in the giant component, level A conferences and level C conferences share a similar value while level B conferences are averagely larger. For EI index, there are significant differences among different

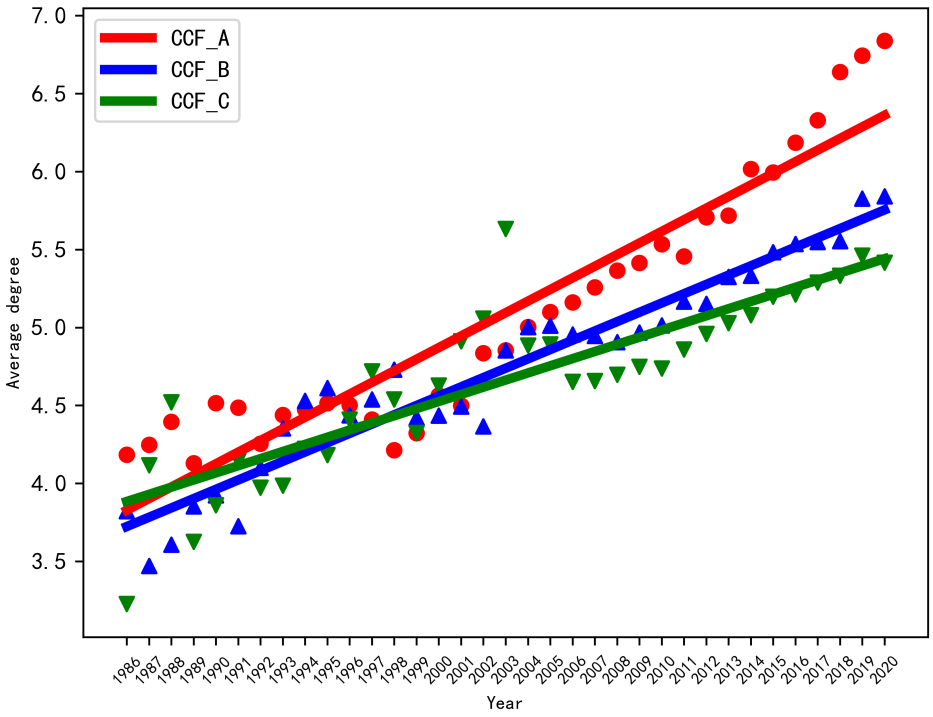


Figure 5.5: Average degree evolves with time for different levels of conference.

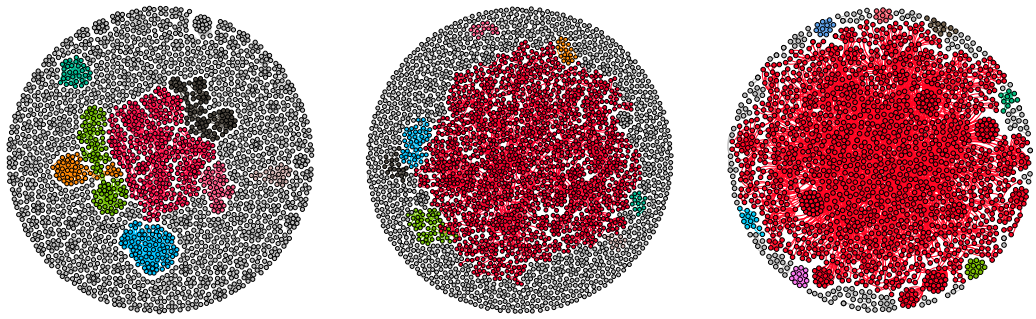
levels of conferences. Top level conferences have larger and denser communities overall, but connections within the core community are much thinner than at lower-level conferences, implying that lower level conferences are preliminary driven by a small group of people. In addition, the proportion of strong connections supports this in another way, namely the frequency of collaboration. The connections in core communities in lower level conferences are not only dense, but also highly weighted. Finally, there are not significant differences in different conference levels in terms of the Q value, meaning that the small world pattern is similar among these conferences.

Table 5.2: Network metrics for different levels of conferences

	GC	Degree	Shortest Path	EI index	Q value	Strong Tie
CCF-A	0.7031	6.6963	5.9876	1.6986	0.1323	0.00267
CCF-B	0.5200	5.6190	6.2360	4.6683	0.132	0.00329
CCF-C	0.3617	5.4123	6.0478	8.7030	0.152	0.00966

5.2.4 Relations between newcomers' papers and giant component

New author with a paper accepted by top conferences will get the opportunity to further collaborate with the researchers who are already inside the giant component. When comparing the evolutionary patterns of giant component (Fig.5.4) and newcomers' papers(Fig.5.1), it is easily seen that level A and level B conferences share a very similar pattern in terms of the evolution of newcomers' papers, but level A conferences are more likely to form a larger giant component in terms of the collaborations among the authors. Similar ratio of newcomers' papers but significant different size of core community could only appear when the new authors are later connected to the giant component of the co-author network after their papers are accepted. The formation of such communities in top conferences may be explained by different reasons: 1) researchers with publications in top conferences, in addition to the good quality of their submissions, may have some connections with the existing communities. This can be explained by signal theory [120], which indicates that trust can be bulit by reputation signal and multiple collaborations. 2) with papers accepted by the top conferences, people could have more chances to collaborate with prestigious researchers who will connect them to the core communities, which can be explained by the relational embeddedness theory [52].



(a) **ASAP** (n = 2335, gc = 0.1272, $< k >_{gc}$=5.1143) (b) **ICNP** (n = 2635, gc = 0.5089, $< k >_{gc}$=5.2549) (c) **NSDI** (n = 1694, gc = 0.8583, $< k >_{gc}$=8.9772)

Figure 5.6: Coauthor networks of 3 conferences in year 2020. The 8 largest components are marked by colors (red for giant component).

In Fig. 5.6, 3 different conferences with very different sizes of giant components are selected to show their entire coauthor network structures by a view of graph. It is clearly seen that the community patterns differ a lot among them in terms of both the component distribution and the connections inside the giant component. For the ASAP conference which has a low size of giant component, its entire society of the authors is decomposed into several balanced groups regardless of the dispersed and isolated authors. Although the giant component seems to have a little bit more authors than the second giant component, the third giant component and so on,

these strong components are not much different with each other. Therefore, there are few bridge nodes between different communities which means few weak ties and strong ties exist.

For ICNP, which has a middle level size of giant component with a p_{gc} near 0.5, its giant component already dominates the entire coauthor networks. Even the size of the second largest component is far smaller than the giant component and the core community of this conferences is already formed and starts growing, but there are still not too many connections inside the core component, which is a sparely connected network. Thus, few strong ties exist in the biggest component. For NDSI, which just became a CCF-A level conference two years before has a very large giant component 0.8583, as well as a densely connected core community with an average degree of 8.9772. Moreover, there are some obvious sub communities in the giant components that indicate different groups of researchers who frequently collaborate and publish papers in this conference.That is, the combination of weak and strong ties can facilitate communication between and within the components.

5.3 Validation of PC chairs' prestige by community metrics and tie strengths

Program committee chairs are typically the persons playing a decisive role in selecting accepted papers for the conference. Regardless of PC chairs' special role in the conferences, they are also ordinary researchers inside the whole scientific community. It is interesting to find out the behavior patterns of the PC chairs in order to figure out how the PC chairs influence the communities of the conferences, as well as their personal change from being a conference chair. To the consensus, PC chairs of computer science conferences (especially for top level conferences) are one of the most prestigious researchers in the community. Typically, they are in the core locations or communities in either co-author networks or citation networks of the entire computer science society. Additionally, PC chairs will form an ego-centered network which can be depicted by the different levels of intimacy and distance between them and their community members. The connections involve trust, power, and mutual obligation [19][23][51], which shape acceptance behaviors and co-authorship between them. Therefore, the closure of a conferences can be additionally evaluated by the relations between chairs and authors. For example, if the papers selected by chairs are very close to the chairs in terms of the collaboration distances, the conferences could be regarded as closed communities or elite small-world network communities. Similarly, we need to combine density, network size, numbers of strong and weak ties, and betweenness to figure out these two kinds of communities. However, compared with PC chairs' metrics in the entire society, the metrics in specific conferences that they ever chaired could be more precisely to calculate their relations with the papers and authors. In this section, several metrics are carried out to verify that the chairs are already inside the core communities or in the key locations of the conferences the time they were elected as

chairs. Overall, a set of 3697 PC chairs from level A and level B conferences are constructed to complete the following analysis. [4]

5.3.1 PC chairs from conferences' own core community

Betweenness centrality and closeness centrality are the metrics that can precisely represent the importance of a node. The larger betweenness and larger closeness of a node are both indicating a core location in the community. In specific, betweenness indicates that the PC chairs act as a bridge and span more different groups, possessing more structural hole benefits such as obtain and control more diverse recourses and information [22]. To measure whether a chair is selected from the core locations of the conference's co-author network, a comparison of the betweenness and closeness between the chairs and the average value of all the authors in the co-author network at the time the chairs are elected is a straightforward and effective method. When chairs are not in the co-author network of the conference they chair, that means currently they have no publications in the conferences before, the betweenness and closeness are set to 0 (minimal value). Here, we leverage the normalized versions of betweenness and closeness in order to eliminate the bias caused by the scale of the network and fairly compare the differences among conferences.

For a specific node v in a network, its betweenness centrality is:

$$C_B(v) = \frac{2}{(N-1)(N-2)} \sum_{s,t \in V} \frac{\sigma(s,t|v)}{\sigma(s,t)}, \tag{5.5}$$

where V is the set of nodes, N is the number of the nodes, $\sigma(s,t)$ is the number of shortest (s,t)-paths, and $\sigma(s,t|v)$ is the number of those paths passing through some node v other than s,t. If $s=t$, $\sigma(s,t)=1$, and if $v \in s,t$, $\sigma(s,t|v)=0$.

And its closeness centrality is:

$$C_C(v) = \frac{n-1}{N-1} \frac{n-1}{\sum_{v=1}^{n-1} d(v,u)}, \tag{5.6}$$

where n is the number of the reachable nodes from v, N is the number of the nodes, $d(v,u)$ is the shortest-path distance between v and u.

Table.5.3 presents comparisons of the both betweenness centrality and closeness centrality between PC chairs and average values of the other authors. It is obviously seen that PC chairs have significant larger centrality than the average centrality of other authors in the network at the same time, which demonstrate that PC chairs are mostly selected from core positions of bridging and central role in the co-author network at the time. Moreover, chairs of level A conferences are more close to the center of the community than those of level B as they have larger betweenness and closeness centrality. It indicates that long-distance bridges exist in the

[4] The data and codes could be found here: `https://github.com/zhangjq1014/DBLP-data`

network. If the network contains a few strong ties and many weak ties and low density, we can distinguish this kind of network as an elite small-world network community.

Table 5.3: PC chairs' high centrality.

Conference level	Betweenness		Closeness	
	Chair	Average	Chair	Average
CCF-A	0.008701	0.000903	0.08689	0.05052
CCF-B	0.007297	0.000973	0.06609	0.03987

5.3.2 Paper-chair tie strength of different distances

Weak tie theory depicts one kind of relationships that potentially exists between two people [52], which means two nodes in a social network may know each other and convey information even though there is no visible edge between them. To validate the weak tie existence and to prove there are PC chairs' influence on paper selections from a statistical aspect, we investigate the relations between each PC chair and the accepted papers from the conference in the year he/she chaired. The minimal distance between a paper and its corresponding chair is defined as the minimal shortest path length among all the authors of the paper. The reason of using minimal distance instead of averaged distance is that we believe the closest relation dominates the influences from the chair to the paper. Firstly, we check the distribution of the paper-chair distances of two levels of conferences separately, level A and level B. The level C conferences are not considered in this part because most of them have not formed a well-organized community. Moreover, it is hard to collect the PC chairs information from the C level conferences because many of their past conferences are not accessible through the website. For each chair, the naive paper-chair distance distribution is the ratio of the numbers of papers in all distances. Suppose there are N papers in a conference in a specific year, each accepted paper $AP_i \in \{AP_1, AP2, ..., AP_N\}$ has k_i authors, the distances between each paper and a chair is

$$D(chair, AP_i) = \min\{Distance(author_j, chair)\ \forall author_j \in AP_i)\}.$$

Then the distribution of the paper distances for a chair is:

$$p(k) = \frac{1}{N}\sum_{i=1}^{N} D(chair, AP_i) = k \tag{5.7}$$

For better visualization here the distances that are greater than 10 are combined as a single data point named >10, which also includes the distance of ∞. Figure 5.7 shows that most papers have a distance of 3 to the chair, and the number of papers with a distances greater than 10 is also considerable. The increase after 10 is because there are many papers without even one author who has a path to the chair in the coauthor network. And there are small but visible differences between the two levels of conferences. The A level conferences have a shorter distances, which means there are more weak ties and more efficient propagation tracks between

the members in the community and chairs (chairs directly connect a small number of people, but they can indirectly influence a large community through weak ties). Thus, it is the reason why the top level conferences form a relative larger giant component. The large connected component is formed by selecting the papers from chairs close collaborators. The distribution of the absolute numbers of papers in terms of paper-chair distances does not show a tie strength related pattern. The distribution is most likely influenced by the statistical distribution of author-chair distances that overall the authors are mostly likely to have a distance of 3 to the chairs that makes the distribution of paper-chair distances the similar statistics.

To eliminate the statistical bias cause by the intrinsic network statistics in terms of the distances, the simple paper-chair distribution should be adjusted and normalized according to the author-chair distribution. For a specific chair, the author-chair distribution is calculated by counting the number of authors of different distances in the whole coauthor network that contains the 334 selected conferences. A key point here is the distribution of author-chair distances used for adjusting the distribution of paper-author distances is strongly aware of temporal series. That is when adjusting the distribution of paper-chair distances for a chair in a specific year, the author-chair statistics from this chair in exactly the same year is used. As a result, the paper-chair distance distribution multiplied by the author-chair distribution in the same year could accurately indicate the relation between tie strength and paper selection. Similarly, the distribution of author-chair distances is:

$$p_{author}(k) = \frac{1}{n}\sum_{i=1}^{n} Distance(chair, author_i) = k, \tag{5.8}$$

where n is the number of the authors in a specific year.

Next, the adjusted distribution of paper-chair distances is:

$$p_{adjust}(k) = p_(k) \times \frac{1}{p_{author}(k)} \tag{5.9}$$

Finally, the adjusted distribution is normalized to make the summation of all elements equals to 1. Obviously, the distribution becomes more different if the absolute numbers are normalized by multiplying a ratio of the distribution of author-chair distances. Even thought there is a turning point for the data for those papers with a distance of 1 to the chair, the whole pattern of this normalized paper-chair distance distribution proves that the strong and weak ties do exist and the constraint of strong ties decreases when the shortest path distance grows. Because if there is no bias when chairs selecting the papers, the normalized distribution should be even that all distances should have an identical probability. This also discloses that tie strength in the coauthor networks does influence the selection of the papers, that the closer authors connect to the chairs, the more probably their papers could be accepted. However, this tie strength influenced paper selection pattern does not indicate a subjective factor or objective factor. It could be caused by many reasons, e.g., the chairs are more unsuspecting of the works from people they know well, or the people around the chair are doing excellent works that in the

lead, or the chairs just want to pick the papers from their own community, or whatever possible scenarios. It is impossible to figure out what is exactly the reason because reevaluating the paper selection is not feasible. But as long as the chairs' influences exist, it is worth knowing how the chairs influence the paper selection. And learning such influences may help people know the reasons that cause the closed communities in conferences.

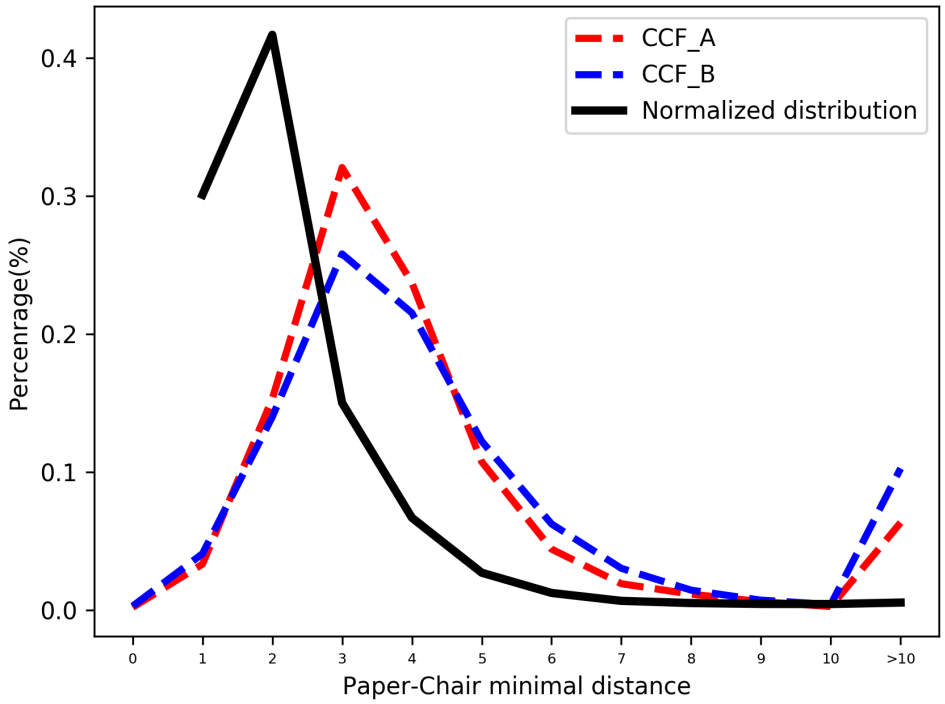


Figure 5.7: Distribution of the minimal distance between each paper and its corresponding chair, and a normalized distribution.

5.3.3 Collaborator-chair tie strength

In addition to the paper-chair distances, the tie strength between PC chairs and their collaborators is also an important parameter that could validate the existence of PC chairs' influences on the conferences. Especially when removing the chair from the network, the changes of their collaborators' connections with other researchers could provide an insight knowledge of the tie strength between chairs and their collaborators. For a conference, if a chair's collaborators' connections with the other people change a lot after eliminating the chair from the coauthor network, the collaborator-chair tie strength effect of this chair is very high that this chair significantly influences the network structure of the conference. To measure the change of the connections, the average shortest path length from a single source to the other nodes is used. For each chair and a conference he/she ever chaired, each of the chair's collaborators has an averaged shortest path length to the other nodes in the giant component. Then the changes of

these average shortest path lengths from the chair's collaborators are a strong indicator of the influences from the chair. Suppose that for a conferences there are N people including a PC chair in the giant component, which is a connected networks. And if the chair has a number of C collaborators, then the average shortest path length of all this chair's collaborators is defined as $LC(chair)$:

$$LC(chair) = \frac{1}{C}\sum_{i=1}^{C}\frac{1}{N-2}\sum_{j=1}^{N-2} Distance(i,j) \tag{5.10}$$

However, after eliminating the node of chair from the network, the calculation of the $LC(chair)$ is a little bit different. Because removing the chair from the network may cause the decomposition of the network, some pairs of nodes are not reachable by each other. To keep a consistence of the $LC(chair)$ before and after removing the chair from the network, we use the maximal shortest path length of the original network to indicate the length between two unreachable nodes after removing the chair. Then the $LC(chair)$ after removing the chair is :

$$LC_{after}(chair) = \frac{1}{C}\sum_{i=1}^{C}\frac{1}{N-2}\sum_{j=1}^{N-2} f(i,j) \tag{5.11}$$

$$f(i,j) = \begin{cases} \text{max length}, & \text{if } \nexists \text{ Distance(i,j)} \\ Distance(i,j), & \text{else} \end{cases}$$

.

Then the change of the LC is:

$$\Delta LC_(chair) = LC_{after}(chair) - LC(chair) \tag{5.12}$$

Figure 5.8 shows a sorted results of the ΔLC of the chairs, as well as a base line. The base line is a result from a set of randomly selected people who have exactly the same degree distribution with the chairs. The same degree distribution could guarantee that the selected people are at the similar positions of the network as the chairs are and could eliminate the bias. Obviously, removing the chairs could influence the connections of their collaborators, in other words without the chairs their collaborators may not be connected to the community of the conferences.

Also, in an elite small-world network community, few high-betweenness members connect and they all have their own ego-centered network. If we delete these nodes (e.g. PC chairs), the connectivity of the network will decrease seriously. As shown in Figure 6, some elites who connect with PC chairs in the community.

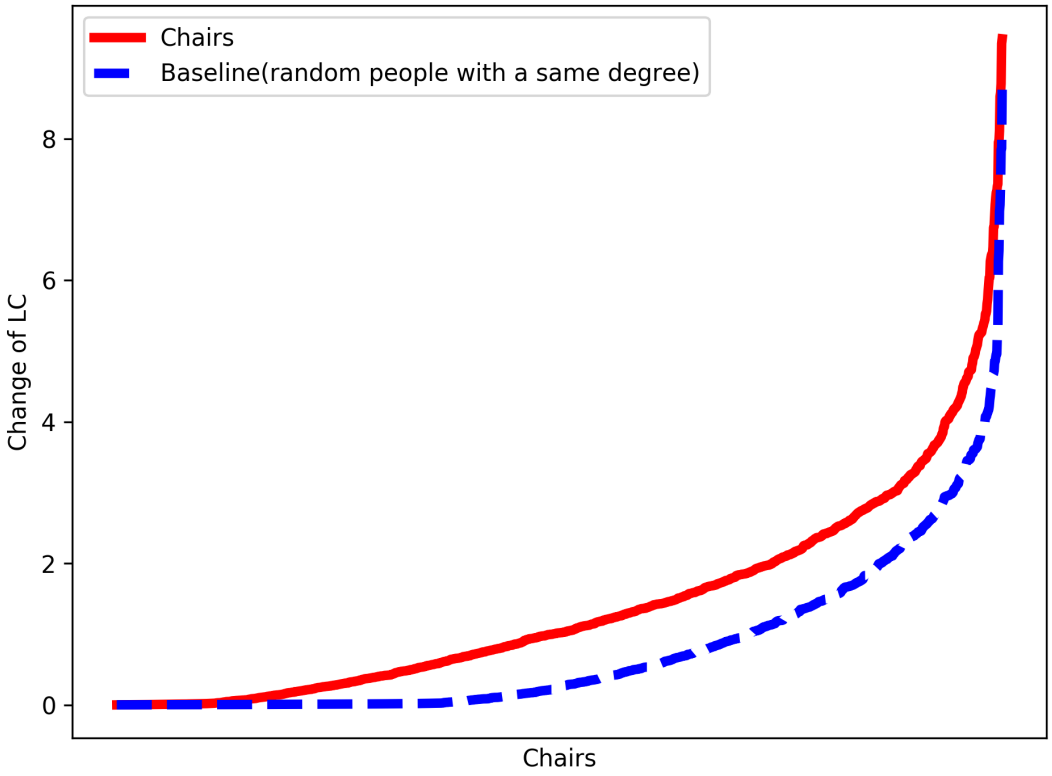


Figure 5.8: The change of the average shortest path length of the chairs' collaborators.

5.4 Quantifying the closure of the conference by PC chair aware metrics

Based on the analysis above, most PC chairs are the people who are already in the core communities when they are elected as chairs of a conference. It is partly consistent with the consensus that PC chairs are prestigious researchers in computer science communities. Here we use several PC chairs' network related metrics to quantitatively presents the centrality of the chairs and find out that PC chairs' of top conferences are likely to have higher centrality. Therefore, in addition to the newcomers' papers, the relation between the old authors' papers their corresponding chairs could quantify the closure of a conference from another perspective. Especially at the time when the chairs are already in the central positions of the community. None of the previous works on scientific collaboration networks have ever taken the PC chairs' influences into consideration. In this section, several metrics, which are related to PC chairs, are investigated to introduce the differences among conferences. No explicit researchers data is presented here, but the statistical parameters of the conference level are examined instead.

5.4.1 PC chairs' connections with their former collaborators

As it is well-known, people are embedding in different type of social networks which influence their behaviors and interaction pattern [50]. The scientific collaboration society is not different from any other social network; it is often the case that the PC chairs have to process the submissions from their close collaborators. For some submissions of the same quality, the same fitness for the topic and the same level of reviewers' rating, it is really hard to make the decisions with absolute objective evaluations. We conjecture that personal social connections may be additional factors for deciding which submissions should be accepted. This kind of social connections impacted decisions are not only personal behaviors, it could also influence the conferences and bring some continuous effects for the communities (e.g. the community structures, behaviors and cultures). It is easy to imagine that if chairs frequently pick the submissions from their previous collaborators in the conference, the community will be closer and the network density will be increase. Therefore, the number of the accepted papers that contain authors having close collaborations with the PC chairs is used to quantify the influences. Here the collaborations between authors and chairs are constrained within a single conference itself, because it is impossible to find the collaborations through the entire coauthor network for all publications (not computer science alone). A paper is defined as a associated paper if it has at least one author that ever co-authored a paper with the chair, the relation between a paper and chair $r_{paper-chair}$ is either 0 or 1:

$$r_{paper-chair}(p,c) = \begin{cases} 1, & \text{if } \exists \text{ Distance(author, chair)=1} \\ 0, & \text{else} \end{cases}$$

.

Suppose that a conference has N accepted papers and in a period of Y years, $\{X_1, X_2, ..., X_Y\}$ is a set of accepted papers organized by year, that $X_i \in \{X_1, X_2, ..., X_Y\}$ is the set of accepted papers in year i, and X_i contains N_i papers. Each paper $x_{ij} \in X_i$ has a number of k_i authors. $\{PC_1, PC_2, ..., PC_Y\}$ is a set of PC chairs organized by year, that $PC_i \in \{PC_1, PC_2, ..., PC_Y\}$ is the PC chair in year i.The average ratio of the accepted papers that with authors having collaborations with the PC chairs for each conference is defined as:

$$p_{cap} = \frac{1}{Y}\sum_{i=1}^{Y}\frac{1}{N_i}\sum_{j=1}^{len(N_i)} r_{paper-chair}(x_{ij}, PC_i) \tag{5.13}$$

,

The averaged proportion of associated accepted papers of some A and B conferences in CCF recommendations are shown in Table.5.4. For each conference, the results are the average P_{cap} of the data since the year the conference started. Unlike the intrinsic metrics of the network such as average degree, giant component size and ratio of newcomers' papers, average P_{cap} of

Table 5.4: Statistics of papers from chairs' collaborators.

Conference level	p_{cap}			
	Avearage	Max	Min	Std
CCF-A	0.04506	0.1558	0.007946	0.02972
CCF-B	0.05375	0.3715	0.007192	0.05788

B conferences is higher than that of A conference. As a result, P_{cap} actually can measure the closure of a conference from the perspective of PC chairs. The max P_{cap} is even large as 0.3715 in B conferences, which means a conference averagely has up to 37.15% papers from chairs' collaborators. Overall, the average P_{cap} of most of the conferences ranges between 0 and 0.1, while the five unique conferences with relative higher proportion of associated accepted papers. These five conferences (TCC, COMPGEOM, NSDI, ICDT, CRYPTO) are significantly different from the other conferences as their PC chairs like to select the papers authored by their previous collaborators. Moreover, the P_{cap} of TCC is even much higher than that of other four conferences, and is more than a double of the P_{cap} of the second conference. If we check some other structural metrics of the network, the proportion of associated accepted papers P_{cap} is not related to any of them. It means the p_{cap} is only PC chair based parameter that indicates the chairs' behaviors and dose not affect the communities' pattern. The reason why a conference keeps an enthusiasm in selecting PC chair associated papers is likely due to the fact that there is a dense and internal group of researchers dominating the operations. However, this kind of community has nothing to do with the quality of the accepted paper because it is possible that some densely connected communities are doing the most advanced works and create high quality products.The p_{cap} is only considered under the circumstance that a borderline paper tries to find a suitable conference to submit, when the authors should consider their connections with the communities of the conferences and the PC chairs. Choosing the conferences with lower p_{cap} may probably increase the possibility of being accepted. It is assuredly that a submission with a high quality will be accepted regardless of the communities and the chairs of the conferences.

To further test whether the quality of these accepted papers, we collect the real citation numbers of them in three typical conferences over years.

5.4.2 Distance between paper and chair

p_{cap} actually measures the direct connections between PC chairs and the authors. Even though it could largely distinguish the conferences in terms of PC chair influences, it is still limited because of considering just single author connections with the chairs. A more general metric is average distance of accepted papers that takes all the authors within the paper into consideration, that is $d_{paper-chair}$. Suppose that a conference has N accepted papers and in a period of Y years, $\{X_1, X_2, ..., X_Y\}$ is a set of accepted papers organized by year, that $X_i \in \{X_1, X_2, ..., X_Y\}$ is the set of accepted papers in year i, and X_i contains K authors. $\{C_1, C_2, ..., C_Y\}$ is a set of giant component organized by year, that $C_i \in \{C_1, C_2, ..., C_Y\}$

is the giant component in year i. $\{PC_1, PC_2, ..., PC_Y\}$ is a set of PC chairs organized by year, that $PC_i \in \{PC_1, PC_2, ..., PC_Y\}$ is the PC chair in year i. Then for each author k in a accepted paper x_{ij} in year i, the reciprocal of distances between the authors and the PC chair for a paper is $d_{i,j,k} = Distance_(author_k, PC_i)\ in\ C_i$. Then the reciprocal averaged distance between accepted papers and PC chairs is:

$$Rd_{paper-chair} = \frac{1}{N}\sum_{i=1}^{Y}\sum_{j=1}^{len(N_i)}\sum_{k=1}^{len(N_{ij})} f(d_{ijk}) \tag{5.14}$$

. Where

$$f(d_{ijk}) = \begin{cases} 0, & \text{if } d_{ijk} = \infty \\ 2, & \text{if } d_{ijk} = 0 \\ \frac{1}{d_{ijk}}, & \text{else} \end{cases}$$

.

Table 5.5: Statistics of average distances between chairs and papers.

Conference level	$Rd_{paper-chair}$			
	Avearage	Max	Min	Std
CCF-A	0.2269	0.3466	0.09124	0.05173
CCF-B	0.2074	0.4856	0.07959	0.07444

As a complement to p_{cap}, the average distance between paper and chairs $d_{paper-chair}$ takes the connections that are more than one hop distance into considerations, which could be used to evaluate the PC chairs' influences of selecting accepted papers from closely connected authors. The reciprocal of the distance is used for computing the metric, for the reason that if two nodes are not connected whose shortest path is ∞, reciprocal could make the value back to 0. And when distance equals to 0 that means a paper containing the PC chair exists, the reciprocal of this condition is defined as 2. As a result, the higher $Rd_{paper-chair}$ means that the PC chairs prefer more on the closely connected authors. As shown in Table.5.5, different from the metric of p_{cap}, A conferences have less average distances between authors and chairs. It means that A conferences are statistically prefer the papers with the authors who are close to the chairs, in terms of the average collaboration distances of all the authors within a paper. However, the conference TCC (Theory of Cryptography Conference) is still significantly more severe than the others in terms of the distance between chair and papers, while the following-up conferences are different from the top 5 conferences in terms of p_{cap} .

5.4.3 Distance between paper and community

In addition to the above PC chair included metrics, another similar metric examining the evolutionary relations between authors and the giant component $I_{paper-community}$ is used to find out that how much the PC chairs prefer selecting authors who are already inside the giant community, or probably the unique collaboration pattern of each conference in terms of authors. $I_{paper-community}$ measures the distribution of the numbers of authors who are already

inside the giant component (i.e., having an accepted paper). Suppose that a conference has N accepted papers and in a period of Y years, $\{X_1, X_2, ..., X_Y\}$ is a set of accepted papers organized by year, that $X_i \in \{X_1, X_2, ..., X_Y\}$ is the set of accepted papers in year i, and X_i contains K authors. $\{C_1, C_2, ..., C_Y\}$ is a set of giant component organized by year, that $C_i \in \{C_1, C_2, ..., C_Y\}$ is the giant component in year i. $\{PC_1, PC_2, ..., PC_Y\}$ is a set of PC chairs organized by year, that $PC_i \in \{PC_1, PC_2, ..., PC_Y\}$ is the PC chair in year i. Let $d_{paper-community}$ denote the number of authors that are already in the giant component for paper N_{ij}, that $d_{paper-community} = \#\ of\ authors\ in\ C_i$. Then the preference for authors already in giant component is:

$$I_{paper-community} = \frac{1}{N} \sum_{i=0}^{Y} \sum_{j=0}^{len(N_i)} d_{paper-community}(X_{ij}) \tag{5.15}$$

.

$I_{paper-community}$ also provides a view that how the conferences deal with the submissions from different kinds of author combinations. The higher the $I_{paper-community}$, the more likely a conference prefers to accept submissions that contain more authors that are already inside the giant component. Naturally, higher $I_{paper-community}$ would lead to a larger size of a giant component which means a more self-community selection oriented community. Unlike the ratio of newcomers' papers, this metric considers all the papers regardless of newcomers' papers or existing authors' paper, and precisely quantify the relations between a paper and the giant community by examining every author in each paper. As shown in Table.5.6, CCF-A conferences averagely show a larger $I_{paper-community}$ than that of CCF-B conferences, which is inconsistent with the hypothesis that top conferences are more self-community selection oriented communities. This phenomenon results from the trust within the community to a great extent. However, if the community lacks weak ties or indirect connections between nodes in a long run, the community will become more and more closed. Therefore, preferring papers with more authors who are already in the giant community and lacking weak ties simultaneously is one of the causations of a closed community.

Table 5.6: Statistics of average number of authors in GC.

Conference level	$I_{paper-community}$			
	Avearage	Max	Min	Std
CCF-A	0.9443	1.8935	0.3170	0.3610
CCF-B	0.5744	1.5221	0.02734	0.3501

5.4.4 How these PC chair relevant metrics be useful and discussions

Even the PC chairs metrics are used to measure the self-community selection degree of a conference, the proportion of weak and strong ties in the community will decide whether the community to develop into a closed community or an elite small-world network community.

Table 5.7: Highly closed conferences based on 5 metrics in the recent 5 venues.

Conference	gc	$1-p_{new}$	$Rd_{paper-chair}$	p_{cap}	$I_{paper-community}$	CCF ranking
TCC	0.8538	0.9167	1.0	1.0	0.6408	B
SIGCOMM	0.6530	0.9809	0.6779	0.3593	0.9055	A
IMC	0.7541	0.9133	0.6791	0.3327	0.8720	B
NSDI	0.8564	0.9039	0.7202	0.4511	0.6895	A
FSE	0.7584	0.9538	0.7117	0.4681	0.6533	B
SC	0.6923	0.9501	0.5138	0.0875	1.0	A
MICRO	0.7069	0.9765	0.5821	0.2712	0.7811	A
CRYPTO	0.6721	0.9238	0.7933	0.4762	0.5628	A
ISCA	0.7004	0.9720	0.6264	0.1986	0.7939	A
COMPGEOM	0.8350	0.9424	0.6700	0.3481	0.6200	B
CVPR	0.8447	1.0	0.5073	0.0716	0.8963	A
SIGMOD	0.8465	0.9418	0.5894	0.1725	0.7814	A
ICDT	0.5881	0.8686	0.7712	0.5179	0.4335	B
UIST	0.7526	0.9305	0.5645	0.2507	0.6882	B
HPCA	0.7539	0.9605	0.5964	0.2148	0.6969	A
INFOCOM	0.7967	0.9937	0.5592	0.0948	0.8058	A
ICDE	0.8091	0.8955	0.5401	0.1090	0.7332	A
CHI	0.8226	0.9694	0.4174	0.0528	0.7854	A
ICSE	0.6512	0.9761	0.5248	0.1085	0.7156	A
PODS	0.7461	0.9300	0.6904	0.2297	0.5893	B

Table 5.8: Highly closed conferences based on 5 metrics since the start.

Conference	gc	$1-p_{new}$	$Rd_{paper-chair}$	p_{cap}	$I_{paper-community}$	CCF ranking
TCC	0.7240	0.8633	1.0	1.0	0.7095	B
NSDI	0.7206	0.8983	0.7344	0.4382	0.8500	A
COMPGEOM	0.7133	0.9150	0.7107	0.4574	0.8156	B
SIGMOD	0.6896	0.9320	0.5502	0.1830	0.9509	A
IMC	0.5946	0.8709	0.6321	0.2631	0.8699	B
CRYPTO	0.4796	0.8052	0.6401	0.3762	0.5526	A
ICDE	0.5850	0.9195	0.4830	0.1001	0.8133	A
STOC	0.7029	0.8917	0.6673	0.1545	0.7481	A
COLT	0.5922	0.8539	0.7117	0.2870	0.6122	B
PODS	0.5425	0.8529	0.6648	0.2263	0.6726	B
FOCS	0.7034	0.8648	0.6338	0.2049	0.6895	A
ICDT	0.3725	0.7343	0.7119	0.4442	0.4401	B
FSE	0.5199	0.7824	0.6134	0.3164	0.5598	B
SODA	0.6680	0.9243	0.6222	0.1550	0.6998	B
CVPR	0.5518	1.0	0.3943	0.0729	0.7717	A
SC	0.4120	0.7314	0.4270	0.1062	0.7346	A
SIGCOMM	0.3191	0.7615	0.5476	0.2211	0.6061	A
SENSYS	0.4638	0.7541	0.4599	0.1868	0.6222	B
ISCA	0.3327	0.8023	0.5655	0.2033	0.5911	A
UIST	0.4532	0.8422	0.4587	0.2085	0.5761	B

When PC chairs maintain too many strong relationships and prefer to select the submissions from their own community, it is difficult for them to find other excellent papers. The trust level of the community is too high and the community size is hard to grow, thus it tends to be a closed one. On the opposite, if the PC chairs connect a lot of members through indirect links, they are more likely to break the information cocoons and community size will maintain sustainable growth.

The acceptance of a paper is not only related to PC chairs, but also related to many other factors. The reason for using PC chairs relevant metrics is that they are highly temporal in nature. PC chairs are rarely repeated in consecutive years, the metrics measuring the relations between the chairs and the authors could preserve more dynamic characteristics of a conference's community. In addition to the characterization of community patterns, these metrics could provide some other valuable information for the reevaluation of the possibility that a paper to be accepted. Although the quality of the paper is the primary determinant of acceptance, it is very common that many papers of a similar quality compete for a limited number of slots. Finally, some of them are rejected and it is not only because of their quality. The question is the quality of a paper is very subjective and hard to quantify. And there is no data of rejected papers, it is impossible to figure out what are the main factors that make a paper to win the competition among the papers with almost the same quality. However, the quantified metrics that are relevant with PC chairs provide useful information in terms of evaluating the actual possibility of paper acceptance, especially when a conference keeps a stable and significant PC chair influences. It turns out that in addition to the factors of quality and conferences' overall paper acceptance, these PC chair relevant metrics could help the authors adjust the general acceptance rates to their unique values, according to their connections to the PC chairs and their positions inside the community. In a nutshell, for a certain conference,the actual acceptance rate is different for different authors, people could compute their own acceptances based on the proposed PC chair relevant metrics.

As for the conference organizers, these PC chair relevant metrics could help them to evaluate the behavior of the program committees and the paper selection process. When serious PC influences are detected, they should make quick responses such as limiting the number of the papers from PC chairs, bringing more PC members from other relevant conferences and double-checking the papers that are close to the chairs in terms of collaboration distance, to prevent the conference from being dominated by several powerful groups and closed community. In a long run, they should encourage PC chairs to build more weak ties instead of strong ties merely to avoid the formation of closed communities. However, having with high PC chairs' influences are not always a bad situation, because it is possible that a researcher network centered by PC chairs of a conference is leading the scientific progress in their respective research areas and the papers they produced are containing high quality works. These excellent papers deserve to be accepted even though the authors are associated with PC chairs. As a result, the PC chair relevant metrics are not used to determine whether a conferences is bad or not, but just an implication of a quantified paper selection patterns that take PC chairs into consideration.

Whether a highly PC chair influenced conference is biased from a fairness point of view requires more subjective judgements, which could not be simply indicated by numbers. Nevertheless, as the influences for the authors are fixed and can be quantified, these metrics provide each author a different actual acceptance rate by taking their positions in the community and their relations with chairs into considerations.

Table 5.7 and Table 5.8 shows some conferences that have highly self-community selection oriented communities, in terms of the data since the start of a conference and the data in last 5 venues, respectively[5]. A total of 20 conferences are presented here. Each PC chair relevant metric is normalized to a proportion of its maximum. Then the table is sorted according to the summation of all the relevant metrics. The ranking of the closure of the conference are different in terms of different periods. For example, SIGCOMM is in the 17th place in the rankings of using all the data since its start, but it rises to the second in the past 5 years. Especially, its ratio of newcomers' papers drops from 0.3325 in full-time data to just around 0.08 in the past 5 year (here the original result is used, not the normalized version in the table).

5.5 Chapter Summary

In this paper, we analyze the community differences among different conferences in computer science society, which has not been investigated before. Based on the weak tie and structural hole theories, we proposed an elite small-world network community to distinguish one kind of network structure form from a single judgment in a closed network among the conference levels. Furthermore, trust theory is discussed deeply in three kinds of communities (closed community, elite small-world community and amorphous community) in this paper. We also conclude that a moderate trust with a decentralized, sparely, large network which called elite small-world network benefits more to the development and innovation of the conference. The result from a total of 334 conferences in the CCF ranking list supports that top conferences are statistically more likely to form a densely and largely connected coauthor network as they have a low ratio of newcomers papers. Moreover, new authors in top conferences are more likely to further collaborate with the authors in the core community and are more likely to form an elite small-world network community because of the existence of weak ties and high-betweenness elites. The coauthor network of each conference preserves clearly a unique pattern in terms of its giant component. Overall, tie strength influenced paper selection pattern is disclosed and the weak tie is proved to exist in coauthor networks. In addition to the classic metrics, several PC chair relevant metrics that quantify their influences on the closures of conferences are proposed to offer more insights into how each conference evolves. A total of 3,697 chairs from 93 CCF-A and CCF-B ranked conferences are used to analyze the PC chair influences from both evolutionary and static phases. The metrics measuring the relations among the chairs, the authors and the conference communities could effectively disclose the interactions among them

[5]The full list of the 93 conferences can be found in: `https://github.com/zhangjq1014/DBLP-data/blob/main/closure%20ranking.csv`

and some conferences with high PC chair influences could be identified base on the proposed metrics.

Chapter 6

Conclusion and Future Work

This chapter summaries this dissertation and provides an outlook for the future work.

6.1 Conclusion

This dissertation focuses on addressing some challenges that remain in personal data mining in some specific applications in online platforms, and propose corresponding solutions to solve the problems. In particular, the problem of user behavior prediction in large online education platform especially for cold-start users, user identification for mobile phone system using privacy-preserved data, and mining different community patterns in different levels of computer science conferences.

Over the past few years, online education platforms have attained tremendous success, and many of the same challenges arise when operating large online education platforms with tens of thousands of users. Especially for for k-12 users, how extracurricular course influence different students performances and course participation is vital important for both students and platforms. Trough studying a dataset from the world's largest extracurricular K-12 education, the correlations between many user-generated data and user profiles data hold true. User profile data is not only provided by the users themselves, but also extended by employing external data sources based on trusted user-generated data. E.g., such as crawling housing prices around a user's home location and use the prices to approximately indicate his/her family socioeconomic status, crawling school reputations to indicate the educational quality a user attains in public school, and using city levels to group users. It is found that user behavior in k-12 online education is significantly correlated with such user profile information, which could be used to solve the problem of user behavior prediction or item recommendation for cold-start users who have no user-generated data at all. We divide user demographic information into three different categories and try to predict user performance, course participation, and subject selection based on different combinations of the feature categories. The experimental results show that user profiles data could contribute to the accuracy of user behavior prediction. In particular, for the prediction of user test scores, the best accuracy is obtained using all the features, with a bias

less than 0.09, while the MAE is just reduced by less than 0.01 when user profile data alone is used without user history. For the prediction of course participation and subject selection, user profiles also slightly improve the prediction accuracy.

In the second work, we propose a model to identify user profession in mobile phone system, using privacy-preserved user generated mobile phone data. Accurate identification of a caller's profession is vital important for protecting mobile users when they use a variety number of online services and they have to engage with strangers. Most of the existing studies on mobile phone user identification have incorporated as much original data as possible into the modeling, ignoring the consideration of operational efficiency and data security problems. Using privacy-preserved data to identify users can not only solve the problem of data security by reducing requirements for secure databases, but also can improve computation efficiency for real time applications by reducing the dimensionality of the input data. To eliminate the sensitive personal information from original data and reduce the dimensionality of the input features, our model transmit the original mobile phone data into three categories of data and , such as mobility, app usage, and amount of data generated per unit of time. For location records, we project a user's coordinates of arbitrary number of days into 12 different directions and calculate directional bias by adopting Standard Deviation, an attenuation coefficient, and natural logarithmic function to reduce the impact of large frequency of a point after projection. For usage of different apps, we use hash functions to encrypt the exact names of applications and design a sorted distribution of several most frequently used apps, which could preserve the apps usage pattern and exclude sensitive information. The experimental results on a dataset from telecommunication operators show that our methods could achieve high accuracy in multiclass identification, and outperforms some existing methods in binary identification. In addition, our method takes more than ten times less features, which could reduce the computational time linearly for $O(n)$ complexity algorithms, and exponentially for algorithms with $O(n^2)$ and higher exponents time complexity.

Finally, we study a typical online social network, scientific collaboration network created by co-authorship, in computer science conferences. First, through analyzing many classical network metrics we find a more closed and densely connected community in top level conferences. In particular, the average giant component size is more than 0.7 for top top conferences, which means for each conference nearly 70% of authors are connected. Similarly, for average degree, ratio of new papers, and average shortest path, top conferences all show a significant transcendence over lower level conferences. Furthermore, the evolutionary analysis of the above metrics shows a growing trend, which implies that the computer science community is going to be more closed in the future, and the collaboration will become denser. In particular, top conferences have significantly formed very closed and dense communities and will significantly widen the gap with lower-level conferences in terms of GC, average degree, average shortest path, Q value, E-I index, and new author rate, according to the evolutionary pattern in the past decades. For the first time, we built a dataset containing information of PC chairs from 93 conferences, and we validated and quantified the impact of PC chairs on community formation

in a conference. In addition, we designed several metrics related to PC chairs, considering the link between PC chairs and the papers of the year they chaired. Finally, we seek to access the closeness of the community in a conference from multidimensional analysis taking both the classic metrics and proposed PC chair metrics into account.

6.2 Future Work

Although, our methods have somehow filled in some gaps in personal data mining in several specific applications and achieved promising performance compared existing works, there are still some issues to be improved in the future.

For mining user behaviors in online education platform, since there is no ground truth data about students' public school performances, the comparison between private extra-curricular courses and public school courses is missing. So as the question that how extra-curricular courses be as a complement to compulsory education is also missing due the lack of data from questionnaires. In the next step, we will send questionnaires to the users in this study, especially to the users in k9 and k12 who have already attended the official entrance examination, collecting their performances in two different time before and after taking the online courses. Then the impact of extracurricular online education will be quantified by comparing their performances in two time slots. As the dataset actually contains a special circumstance, Covid-19 outbreak, we will analyze how Covid-19 outbreak impacts users participation and performance in online education platforms and its different impacts on users of different demographic information. At last, our current study is limited within just one online education platform which is lack of generalization. And for the prediction accuracy, since currently many SES information is approximated by online external data resources, we believe that using data from questionnaires will help improve prediction accuracy.

For identifying mobile phone callers using privacy preserved data, currently we are seeking to collaborate with other telecommunication operators to acquire additional data from different platforms and different regions, in order to verify the performance of our proposed methods in a various of environments. Currently, our method takes only network requests data into account, which could be only acquired from telecommunication operator. Many mobile phone itself generated data is ignored, such as data from diverse sensors, user click behavior, and user operations within a application. We believe that these additional data could be used to build multidimentional features for a user and could significantly improve the accuracy of identifying a caller's profession. However, acquiring such data needs cooperation from several independent companies or institutions, such as device manufacturer, telecommunication operations, and various application providers. Previously, such different institutions seldom collaborate with each other for data mining issues, especially for big and sensitive personal data sharing. With our approach, maybe many different institutions holding different data resources could share privacy-preserved but identifier-retained data. We are going to collaborate with one of the largest Internet Company to acquire some privacy-preserved data within a few

specific applications and try to use the data to improve the identification accuracy, especially for multiclass identification.

For the analysis of coauthor networks in computer science conferences, we will refer to many other ranking lists instead of CCF ranking alone. Taking different ranking lists into account, especially the lists that are ranked by numeric factors and have detailed continuous ranks, will help more on exploring the cause of the formation of diverse communities with different closeness. In addition, we will explore more community differences in networks formed by other type of connections, e.g., citation network, affiliation network, and possible peer review network. These different networks could be coupled together and many new metrics could be designed for measuring the community patterns.

Bibliography

[1] *A Growing Threat to Your Finances: Cell-Phone Account Fraud*. `https://www.consumerreports.org/scams-fraud/cell-phone-account-fraud`.

[2] Divyakant Agrawal, Ceren Budak, Amr El Abbadi, Theodore Georgiou, and Xifeng Yan. "Big data in online social networks: user interaction analysis to model user behavior in social networks". In: *International Workshop on Databases in Networked Information Systems*. Springer. 2014, pp. 1–16.

[3] Alexandre Alahi, Kratarth Goel, Vignesh Ramanathan, et al. "Social lstm: Human trajectory prediction in crowded spaces". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 961–971.

[4] Mohammed F Alhamid, Majdi Rawashdeh, Haiwei Dong, M Anwar Hossain, and Abdulmotaleb El Saddik. "Exploring latent preferences for context-aware personalized recommendation systems". In: *IEEE Transactions on Human-Machine Systems* 46.4 (2016), pp. 615–623.

[5] Bedoor K AlShebli, Talal Rahwan, and Wei Lee Woon. "The preeminence of ethnic diversity in scientific collaboration". In: *Nature communications* 9.1 (2018), pp. 1–10.

[6] Florent Altché and Arnaud de La Fortelle. "An LSTM network for highway trajectory prediction". In: *2017 IEEE 20th International Conference on Intelligent Transportation Systems (ITSC)*. IEEE. 2017, pp. 353–359.

[7] Xavier Amatriain. "Mining large streams of user data for personalized recommendations". In: *ACM SIGKDD Explorations Newsletter* 14.2 (2013), pp. 37–48.

[8] Ketan Anand, Jay Kumar, and Kunal Anand. "Anomaly detection in online social network: A survey". In: *2017 International Conference on Inventive Communication and Computational Technologies (ICICCT)*. IEEE. 2017, pp. 456–459.

[9] Akash Anil, Uppinder Chugh, and Sanasam Ranbir Singh. "On applying meta-path for network embedding in mining heterogeneous DBLP network". In: *International Conference on Pattern Recognition and Machine Intelligence*. Springer. 2019, pp. 249–257.

[10] Sercan O Arık and Tomas Pfister. "Tabnet: Attentive interpretable tabular learning". In: *arXiv* (2020).

[11] Seyed Mojtaba Hosseini Bamakan, Ildar Nurgaliev, and Qiang Qu. "Opinion leader detection: A methodological review". In: *Expert Systems with Applications* 115 (2019), pp. 200–222.

[12] Rashmi Dutta Baruah and Plamen Angelov. "Evolving social network analysis: A case study on mobile phone data". In: *2012 IEEE Conference on Evolving and Adaptive Intelligent Systems*. IEEE. 2012, pp. 114–120.

[13] Steven Bethard and Dan Jurafsky. "Who should I cite: learning literature search models from citation behavior". In: *Proceedings of the 19th ACM international conference on Information and knowledge management*. 2010, pp. 609–618.

[14] Jose Borges and Mark Levene. "Data mining of user navigation patterns". In: *International workshop on web usage analysis and user profiling*. Springer. 1999, pp. 92–112.

[15] Katy Börner, Jeegar T Maru, and Robert L Goldstone. “The simultaneous evolution of author and paper networks”. In: *Proceedings of the National Academy of Sciences* 101.suppl 1 (2004), pp. 5266–5273.

[16] Jered Borup, Charles R Graham, and Andrea Velasquez. “Technology-mediated caring: Building relationships between students and instructors in online K-12 learning environments”. In: *Emotion and school: Understanding how the hidden curriculum influences relationships, leadership, teaching, and learning*. Emerald Group Publishing Limited, 2013.

[17] A. Bozkurt et al. “A global outlook to the interruption of education due to COVID-19 Pandemic: Navigating in a time of uncertainty and crisis”. In: *Asian Journal of Distance Education* 15.1 (2020), pp. 1–126.

[18] R. H. Bradley and R. F. Corwyn. “Socioeconomic status and child development”. In: *Annual Review of Psychology* 21.3 (2002), pp. 371–399.

[19] Daniel J. Brass. “Being in the Right Place: A Structural Analysis of Individual Influence in an Organization”. In: *Administrative Science Quarterly* 29.4 (1984), pp. 518–539.

[20] Mark Bray, Shengli Zhan, Chad Lykins, Dan Wang, and Ora Kwo. “Differentiated demand for private supplementary tutoring: Patterns and implications in Hong Kong secondary education”. In: *Economics of Education Review* 38 (2014), pp. 24–37.

[21] William C Brehm and Iveta Silova. “Hidden privatization of public education in Cambodia: Equity implications of private tutoring”. In: *Journal for Educational Research Online* 6.1 (2014), pp. 94–116.

[22] R. S. Burt. *Structural Holes*. Cambridge: Harvard University Press, 1992.

[23] Ronald S. Burt and Katarzyna Burzynska. “Chinese Entrepreneurs, Social Networks, and Guanxi”. In: *Management and Organization Review* 13.2 (2017), pp. 221–260.

[24] Jordi Cabot, Javier Luis Cánovas Izquierdo, and Valerio Cosentino. “Are CS conferences (too) closed communities?” In: *Communications of the ACM* 61.10 (2018), pp. 32–34.

[25] Nanang Cahyana and Rinaldi Munir. “Community and important actors analysis with different keywords in social network”. In: *2017 3rd International Conference on Science in Information Technology (ICSITech)*. IEEE. 2017, pp. 498–502.

[26] Wang Chen, Qiang Gao, and Huagang Xiong. “Temporal Predictability of Online Behavior in Foursquare”. In: *Entropy* 18 (2016), p. 296.

[27] Gokul Chittaranjan, Jan Blom, and Daniel Gatica-Perez. “Mining large-scale smartphone data for personality studies”. In: *Personal and Ubiquitous Computing* 17.3 (2013), pp. 433–450.

[28] Gokul Chittaranjan, Jan Blom, and Daniel Gatica-Perez. “Who’s who with big-five: Analyzing and classifying personality traits with smartphones”. In: *2011 15th Annual international symposium on wearable computers*. IEEE. 2011, pp. 29–36.

[29] Xinyi Cui, Qingshan Liu, Mingchen Gao, and Dimitris N Metaxas. “Abnormal detection using interaction energy potentials”. In: *CVPR 2011*. IEEE. 2011, pp. 3161–3167.

[30] Ali Daud, Naif Radi Aljohani, Rabeeh Ayaz Abbasi, et al. “Finding rising stars in co-author networks via weighted mutual influence”. In: *Proceedings of the 26th International Conference on World Wide Web Companion*. 2017, pp. 33–41.

[31] G. Dave, L. Frerichs, J. Jones, et al. “Conceptualizing trust in community-academic research partnerships using concept mapping approach: A multi-CTSA study”. In: *Evaluation and Program Planning* (2017), pp. 70–78.

[32] James Davidson, Benjamin Liebald, Junning Liu, et al. “The YouTube video recommendation system”. In: *Proceedings of the fourth ACM conference on Recommender systems*. 2010, pp. 293–296.

[33] Deloitte China. *A new era of education: China education development report 2018*. 2018.

[34] Shuiguang Deng, Longtao Huang, and Guandong Xu. “Social network-based service recommendation with trust enhancement”. In: *Expert Systems with Applications* 41.18 (2014), pp. 8075–8084.

[35] Sitaram Devarakonda, Dmitriy Korobskiy, Tandy Warnow, and George Chacko. “Viewing computer science through citation analysis: Salton and Bergmark Redux”. In: *Scientometrics* 125.1 (2020), pp. 271–287.

[36] S. Ding, X. Gao, Y. Dong, Y. Tong, and X. Fu. “Estimating Multiple Socioeconomic Attributes via Home Location: A Case Study in China”. In: *Journal of Social Computing* 2.1 (2021), pp. 71–88.

[37] Yuxiao Dong, Reid A Johnson, and Nitesh V Chawla. “Will this paper increase your h-index? Scientific impact prediction”. In: *Proceedings of the eighth ACM international conference on web search and data mining*. 2015, pp. 149–158.

[38] D. R. Entwislea and M. A. Nan. “Some Practical Guidelines for Measuring Youth’s Race/Ethnicity and Socioeconomic Status”. In: *Child Development* 65.6 (1994).

[39] Evergreen Education Group. *Keeping pace with K-12 online learning*. 2016.

[40] Aftab Farooq, Gulraiz Javaid Joyia, Muhammad Uzair, and Usman Akram. “Detection of influential nodes using social networks analysis based on network metrics”. In: *2018 International Conference on Computing, Mathematics and Engineering Technologies (iCoMET)*. IEEE. 2018, pp. 1–6.

[41] Mariela Fernandez, Rasul A. Mowatt, Kimberly J. Shinew, Monika Stodolska, and William Stewart. “Going the Extra Mile: Building Trust and Collaborative Relationships with Study Participants”. In: *Leisure Sciences* 0.0 (2020), pp. 1–18.

[42] Michael Fire, Lena Tenenboim, Ofrit Lesser, et al. “Link prediction in social networks using computationally efficient topological features”. In: *2011 IEEE third international conference on privacy, security, risk and trust and 2011 IEEE third international conference on social computing*. IEEE. 2011, pp. 73–80.

[43] H. J. Forgasz and S. Griffith. “Computer algebra system calculators: Gender issues and teachers’ expectations”. In: *Australian Senior Mathematics Journal* 20.2 (2006), pp. 18–30.

[44] L. Frerichs, M. Kim, G. Dave, et al. “Stakeholder Perspectives on Creating and Maintaining Trust in Community–Academic Research Partnerships”. In: *Health Education & Behavior* 44.1 (2017), pp. 182–191.

[45] Xiaoming Fu, Hong Huang, Xiang-Yang Li, Haisheng Tan, and Jie Tang. “A comparative analysis of school pupils’ daily habits in Germany and China”. In: *IEEE INFOCOM-HotPOST’18*.

[46] Lorenzo Gabrielli, Barbara Furletti, Roberto Trasarti, Fosca Giannotti, and Dino Pedreschi. “City users’ classification with mobile phone data”. In: *2015 IEEE International Conference on Big Data (Big Data)*. IEEE. 2015, pp. 1007–1012.

[47] E. Gallo, Y. E. Riyanto, T. H. Teh, and N. Roy. “Strong links promote the emergence of cooperative elites”. In: *Scientific Reports* 9.1 (2019).

[48] Yang Gao, Yan Ma, and Dandan Li. “Anomaly detection of malicious users’ behaviors for web applications based on web logs”. In: *2017 IEEE 17th International Conference on Communication Technology (ICCT)*. IEEE. 2017, pp. 1352–1355.

[49] *GDPR recital 71*. `https://gdpr-info.eu/recitals/no-71/`.

[50] M. Granovetter. *Economic Action and Social Structure: The Problem of Embeddedness*. Readings in Economic Sociology, 2008.

[51] M. Granovetter. *Society and Economy—Framework and Principles*. Cambridge: Harvard University Press, 2017.

[52] Mark Granovetter. “The strength of weak ties: A network theory revisited”. In: *Sociological theory* (1983), pp. 201–233.

[53] Karin Guill, Oliver Lüdtke, and Olaf Köller. “Assessing the instructional quality of private tutoring and its effects on student outcomes: Analyses from the German National Educational Panel Study”. In: *British Journal of Educational Psychology* 90.2 (2020), pp. 282–300.

[54] Yuhe Guo, Qihui Chen, Shengying Zhai, and Chunchen Pei. “Does private tutoring improve student learning in China? Evidence from the China Education Panel Survey”. In: *Asia & the Pacific Policy Studies* 7.3 (2020), pp. 322–343.

[55] Aric Hagberg, Pieter Swart, and Daniel S Chult. *Exploring network structure, dynamics, and function using NetworkX*. Tech. rep. Los Alamos National Lab.(LANL), Los Alamos, NM (United States), 2008.

[56] Michael R. Harwell and Qian Zhao. “An empirical example of capturing the impact of SES on student achievement using path analysis”. In: *International Journal of Educational Research* 105.101715 (2021).

[57] Jorge E Hirsch. “An index to quantify an individual’s scientific research output”. In: *Proceedings of the National academy of Sciences* 102.46 (2005), pp. 16569–16572.

[58] Taekeun Hong, Chang Choi, and Juhyun Shin. “CNN-based malicious user detection in social networks”. In: *Concurrency and Computation: Practice and Experience* 30.2 (2018), e4163.

[59] William Horton. *E-learning by design*. Wiley, 2011.

[60] Eaman Jahani, Pål Sundsøy, Johannes Bjelland, Linus Bengtsson, Yves-Alexandre de Montjoye, et al. “Improving official statistics in emerging markets using machine learning and mobile phone data”. In: *EPJ Data Science* 6.1 (2017), p. 3.

[61] Yifan Jia and Li Qu. “Improve the Performance of link prediction methods in citation network by using H-Index”. In: *2016 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery (CyberC)*. IEEE. 2016, pp. 220–223.

[62] Justin, Jagosh, Paula, et al. “A realist evaluation of community-based participatory research: partnership synergy, trust building and related ripple effects”. In: *BMC Public Health* 15 (2015), p. 725.

[63] Chaogui Kang, Song Gao, Xing Lin, et al. “Analyzing and geo-visualizing individual human mobility patterns using mobile call records”. In: *2010 18th international conference on geoinformatics*. IEEE. 2010, pp. 1–7.

[64] White R. Karl. “The relation between socioeconomic status and academic achievement.” In: *Psychological Bulletin* 91.3 (1982), pp. 229–248.

[65] Krishnaram Kenthapadi, Ilya Mironov, and Abhradeep Guha Thakurta. “Privacy-preserving data mining in industry”. In: *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*. 2019, pp. 840–841.

[66] Muhammad Usman Khan and Shoab Ahmed Khan. “Social networks identification and analysis using call detail records”. In: *Proceedings of the 2nd International Conference on Interaction Sciences: Information Technology, Culture and Human*. 2009, pp. 192–196.

[67] MM Abo Khedra, AA Abd EL-Aziz, and Hesham A Hefny. “Social Network Analysis through Big Data Platform Review”. In: *2019 International Conference on Computer and Information Sciences (ICCIS)*. IEEE. 2019, pp. 1–5.

[68] M. Kilduff, A. Mehra, D. A. Gioia, and S. Borgatti. *Brokering Trust to Enhance Leadership: A Self-Monitoring Approach to Leadership Emergence*. Springer International Publishing, 2017.

[69] Csaba Komlo and Lajos Kis-Toth. “Virtual and on-line classrooms of e-learning”. In: *IEEE 63rd Annual Conference International Council for Education Media (ICEM 2013)*.

[70] D. Krackhardt. *The Strength of Strong Ties: The Importance of Philos in Organizations*. Networks in the Knowledge Economy, 1992.

[71] Xuan Nhat Lam, Thuc Vu, Trong Duc Le, and Anh Duc Duong. "Addressing cold-start problem in recommendation systems". In: *Proceedings of the 2nd international conference on Ubiquitous information management and communication*. 2008, pp. 208–211.

[72] Stephen Lamb. "Gender differences in mathematics participation in Australian schools: Some relationships with social class and school policy". In: *British Educational Research Journal* 22.2 (1996), pp. 223–223.

[73] C. W. Langfred. "Too Much of a Good Thing? Negative Effects of High Trust and Individual Autonomy in Self-Managing Teams". In: *Academy of Management Journal* 47.3 (2004), pp. 385–399.

[74] Duc-Trong Le, Yuan Fang, and Hady W Lauw. "Modeling sequential preferences with dynamic user and context factors". In: *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer. 2016, pp. 145–161.

[75] Hui Li. "Centrality analysis of online social network big data". In: *2018 IEEE 3rd International Conference on Big Data Analysis (ICBDA)*. IEEE. 2018, pp. 38–42.

[76] Jiachen Li, Hengbo Ma, and Masayoshi Tomizuka. "Interaction-aware multi-agent tracking and probabilistic behavior prediction via adversarial learning". In: *2019 international conference on robotics and automation (ICRA)*. IEEE. 2019, pp. 6658–6664.

[77] Shifeng Li, Qiongying Xu, and Ruixue Xia. "Relationship between SES and academic achievement of junior high school students in China: The mediating effect of self-concept". In: *Frontiers in Psychology* 10 (2020), p. 2513.

[78] Xin Li, Yiqun Liu, Min Zhang, and Shaoping Ma. "Fraudulent Support Telephone Number Identification Based on Co-Occurrence Information on the Web". In: *AAAI*. 2014.

[79] Yuhong Li, Dongmei Hou, Aimin Pan, and Zhiguo Gong. "Demalc: A feature-rich machine learning framework for malicious call detection". In: *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*. 2017, pp. 1559–1567.

[80] Lin Liao, Donald J. Patterson, Dieter Fox, and Henry Kautz. "Building personal maps from GPS data." In: *Annals of the New York Academy of Sciences* 1093 (2006), pp. 249–65.

[81] Yuanfu Lu, Yuan Fang, and Chuan Shi. "Meta-learning on heterogeneous information networks for cold-start recommendation". In: *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2020, pp. 1563–1573.

[82] Xiong Luo, Changwei Jiang, Weiping Wang, et al. "User behavior prediction in social networks using weighted extreme learning machine with distribution optimization". In: *Future Generation Computer Systems* 93 (2019), pp. 1023–1035.

[83] Yifei Ma, Balakrishnan Narayanaswamy, Haibin Lin, and Hao Ding. "Temporal-Contextual Recommendation in Real-Time". In: *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2020, pp. 2291–2299.

[84] Nicole Martinez-Martin, Thomas R Insel, Paul Dagum, Henry T Greely, and Mildred K Cho. "Data mining for health: staking out the ethical territory of digital phenotyping". In: *NPJ digital medicine* 1.1 (2018), pp. 1–5.

[85] Alexandra R Maryanski. "African ape social structure: Is there strength in weak ties?" In: *Social Networks* 9.3 (1987), pp. 191–215.

[86] Ryoji Matsuoka. "Inequality in shadow education participation in an egalitarian compulsory education system". In: *Comparative Education Review* 62.4 (2018), pp. 565–586.

[87] Ryoji Matsuoka. "School socioeconomic compositional effect on shadow education participation: Evidence from Japan". In: *British Journal of Sociology of Education* 36.2 (2015), pp. 270–290.

[88] Yves-Alexandre de Montjoye, Jordi Quoidbach, Florent Robic, and Alex Sandy Pentland. "Predicting personality using novel mobile phone-based metrics". In: *International conference on social computing, behavioral-cultural modeling, and prediction*. Springer. 2013, pp. 48–55.

[89] National Bureau of Statistics of China. "China Statistical Yearbook 2020". In: (2020).

[90] Mark EJ Newman. "Coauthorship networks and patterns of scientific collaboration". In: *Proceedings of the national academy of sciences* 101.suppl 1 (2004), pp. 5200–5205.

[91] Mark EJ Newman. "The structure of scientific collaboration networks". In: *Proceedings of the national academy of sciences* 98.2 (2001), pp. 404–409.

[92] MM Nyhan, I Kloog, R Britter, C Ratti, and P Koutrakis. "Quantifying population exposure to air pollution using individual mobility patterns inferred from mobile phone data". In: *Journal of exposure science & environmental epidemiology* 29.2 (2019), p. 238.

[93] Rodrigo de Oliveira, Alexandros Karatzoglou, Pedro Concejero Cerezo, Ana Armenta Lopez de Vicuña, and Nuria Oliver. "Towards a psychographic user model from mobile phone usage". In: *CHI'11 Extended Abstracts on Human Factors in Computing Systems*. 2011, pp. 2191–2196.

[94] Nuria Oliver, Bruno Lepri, Harald Sterly, et al. "Mobile phone data for informing public health actions across the COVID-19 pandemic life cycle". In: *Science Advances* 6.23 (2020), eabc0764.

[95] P. S. Park, J. E. Blumenstock, and M. W. Macy. "The strength of long-range ties in population-scale social networks". In: *Science* 362 (2018).

[96] Sudhaman Parthasarathy and San Murugesan. "Overnight Transformation To Online Education Due to the COVID-19 Pandemic: Lessons learned". In: *ACM eLearn Magazine* 2020.9 (2020).

[97] Md Salik Parwez, Danda B Rawat, and Moses Garuba. "Big data analytics for user-activity analysis and user-anomaly detection in mobile wireless network". In: *IEEE Transactions on Industrial Informatics* 13.4 (2017), pp. 2058–2065.

[98] F. Pedregosa, G. Varoquaux, et al. "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.

[99] Lu Peng and Rongheng Lin. "Fraud phone calls analysis based on label propagation community detection algorithm". In: *2018 IEEE World Congress on Services (SERVICES)*. IEEE. 2018, pp. 23–24.

[100] Lu Peng and Rongheng Lin. "Fraud Phone Calls Analysis Based on Label Propagation Community Detection Algorithm". In: *2018 IEEE World Congress on Services (SERVICES)* (2018), pp. 23–24.

[101] *Platform-to-Consumer Delivery - worldwide | Statista Market Forecast*. `https://www.statista.com/outlook/376/100/platform-to-consumer-delivery/worldwide`.

[102] Zhijin Qin, Yue Gao, and Mark D Plumbley. "Malicious user detection based on low-rank matrix completion in wideband spectrum sensing". In: *IEEE Transactions on Signal Processing* 66.1 (2017), pp. 5–17.

[103] Reza Rawassizadeh, Chelsea Dobbins, Mohammad Akbari, and Michael Pazzani. "Indexing multivariate mobile data through spatio-temporal event detection and clustering". In: *Sensors* 19.3 (2019), p. 448.

[104] Xiang Ren, Jialu Liu, Xiao Yu, et al. "Cluscite: Effective citation recommendation by information network-based clustering". In: *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. 2014, pp. 821–830.

[105] John Resig, Santosh Dawara, Christopher M Homan, and Ankur Teredesai. "Extracting social networks from instant messaging populations". In: *Proc. of ACM SIGKDD*. Citeseer. 2004, pp. 22–25.

[106] *Restaurant-to-Consumer Delivery - worldwide | Statista Market Forecast*. https://www.statista.com/outlook/375/100/restaurant-to-consumer-delivery/worldwide.

[107] *Ride-Hailing & Taxi - worldwide | Statista Market Forecast*. https://www.statista.com/outlook/368/100/ride-hailing-taxi/worldwide.

[108] Debora L. Roorda, Helma M. Y. Koomen, Jantine L. Spilt, and Frans J. Oort. "The Influence of Affective Teacher-Student Relationships on Students' School Engagement and Achievement: A Meta-Analytic Approach". In: *Review of Educational Research* 81.4 (2011), pp. 493–529.

[109] Ryan A Rossi, Brian Gallagher, Jennifer Neville, and Keith Henderson. "Modeling dynamic behavior in large evolving graphs". In: *Proceedings of the sixth ACM international conference on Web search and data mining*. 2013, pp. 667–676.

[110] Deockhyun Ryu and Changhui Kang. "Do Private Tutoring Expenditures Raise Academic Performance? Evidence from Middle School Students in South Korea". In: *Asian Economic Journal* 27.1 (2013), pp. 59–83.

[111] Somya Ranjan Sahoo and Brij B Gupta. "Fake profile detection in multimedia big data on online social networks". In: *International Journal of Information and Computer Security* 12.2-3 (2020), pp. 303–331.

[112] P. A. Saparito. "The central role of calculus-based trust and relational trust in bank-small firm relationships". In: *Academy of Management Proceedings & Membership Directory* 47.1 (2002), pp. 400–410.

[113] Salvatore Scellato, Anastasios Noulas, and Cecilia Mascolo. "Exploiting place features in link prediction on location-based social networks". In: *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*. 2011, pp. 1046–1054.

[114] Suvash Sedhain, Scott Sanner, Darius Braziunas, Lexing Xie, and Jordan Christensen. "Social collaborative filtering for cold-start recommendations". In: *Proceedings of the 8th ACM Conference on Recommender systems*. 2014, pp. 345–348.

[115] VV Sergeev, IM Gorbchenko, and VV Safronov. "Comparative analysis of fraud detection systems by phone number". In: *Journal of Physics: Conference Series*. Vol. 1679. 5. IOP Publishing. 2020, p. 052003.

[116] Edoardo Serra, Anu Shrestha, Francesca Spezzano, and Anna Squicciarini. "DeepTrust: An automatic framework to detect trustworthy users in opinion-based systems". In: *Proceedings of the Tenth ACM Conference on Data and Application Security and Privacy*. 2020, pp. 29–38.

[117] Katrina J Serrano, Kisha I Coa, Mandi Yu, Dana L Wolff-Hughes, and Audie A Atienza. "Characterizing user engagement with health app data: a data mining approach". In: *Translational behavioral medicine* 7.2 (2017), pp. 277–285.

[118] Emer Smyth. "Buying your way into college? Private tuition and the transition to higher education in Ireland". In: *Oxford Review of Education* 35.1 (2009), pp. 1–22.

[119] Kyoung-Oh Song, Hyun-Jeong Park, and Kyong-Ah Sang. "A cross-national analysis of the student-and school-level factors affecting the demand for private tutoring". In: *Asia Pacific Education Review* 14.2 (2013), pp. 125–139.

[120] A. Michael Spence. *Market Signaling: Informational Transfer in Hiring and Related Screening Processes*. Cambridge: Harvard University Press, 1974.

[121] Kristin Brooke Stecher and Scott Counts. "Spontaneous Inference of Personality Traits and Effects on Memory for Online Profiles". In: *ICWSM*. 2008.

[122] Ana-Andreea Stoica, Jessy Xinyi Han, and Augustin Chaintreau. "Seeding network influence in biased networks and the benefits of diversity". In: *Proceedings of The Web Conference 2020*. 2020, pp. 2089–2098.

[123] Victor Ströele, Fernanda Campos, José Maria N David, et al. "Data abstraction and centrality measures to scientific social network analysis". In: *2017 IEEE 21st International Conference on Computer Supported Cooperative Work in Design (CSCWD)*. IEEE. 2017, pp. 281–286.

[124] Du Su, Hieu Tri Huynh, Ziao Chen, Yi Lu, and Wenmiao Lu. "Re-identification Attack to Privacy-Preserving Data Analysis with Noisy Sample-Mean". In: *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2020, pp. 1045–1053.

[125] Yizhou Sun, Rick Barber, Manish Gupta, Charu C Aggarwal, and Jiawei Han. "Co-author relationship prediction in heterogeneous bibliographic networks". In: *2011 International Conference on Advances in Social Networks Analysis and Mining*. IEEE. 2011, pp. 121–128.

[126] Pål Sundsøy, Johannes Bjelland, Bjørn-Atle Reme, et al. "Towards real-time prediction of unemployment and profession". In: *International Conference on Social Informatics*. Springer. 2017, pp. 14–23.

[127] Huiyi Tan, Junfei Guo, and Yong Li. "E-learning recommendation system". In: *2008 International conference on computer science and software engineering*. Vol. 5. IEEE. 2008, pp. 430–433.

[128] Zhaowei Tan, Changfeng Liu, Yuning Mao, et al. "AceMap: A novel approach towards displaying relationship among academic literatures". In: *Proceedings of the 25th international conference companion on world wide web*. 2016, pp. 437–442.

[129] Jian Tang, Meng Qu, Mingzhe Wang, et al. "Line: Large-scale information network embedding". In: *Proceedings of the 24th international conference on world wide web*. 2015, pp. 1067–1077.

[130] Jie Tang. "AMiner: Toward understanding big scholar data". In: *Proceedings of the ninth ACM international conference on web search and data mining*. 2016, pp. 467–467.

[131] Jie Tang. "Computational models for social network analysis: A brief survey". In: *Proceedings of the 26th international conference on world wide web companion*. 2017, pp. 921–925.

[132] DBLP Team. *DBLP Statistics*. `https://dblp.org/statistics/index.html`. 2020.

[133] Zhen Tu, Runtong Li, Yong Li, et al. "Your apps give you away: distinguishing mobile users by their app usage fingerprints". In: *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2.3 (2018), pp. 1–23.

[134] Reyn van Ewijk and Peter Sleegers. "The effect of peer socioeconomic status on student achievement: A meta-analysis". In: *Educational Research Review* 5.2 (2010), pp. 134–150.

[135] Maksims Volkovs, Guang Wei Yu, and Tomi Poutanen. "DropoutNet: Addressing Cold Start in Recommender Systems." In: *NIPS*. 2017, pp. 4957–4966.

[136] Fei Wang, Wei Chen, Ye Zhao, et al. "Adaptively exploring population mobility patterns in flow visualization". In: *IEEE Transactions on Intelligent Transportation Systems* 18.8 (2017), pp. 2250–2259.

[137] Wen Wang, Wei Zhang, Shukai Liu, et al. "Beyond clicks: Modeling multi-relational item graph for session-based target behavior prediction". In: *Proceedings of The Web Conference 2020*. 2020, pp. 3056–3062.

[138] S. Watts D.and Strogatz. "Collective dynamics of 'small-world' networks". In: *Nature* 393 (1998), 440–442.

[139] Lili Wu, Dajun Zhang, Gang Cheng, Tianqiang Hu, and Detlef H Rost. "Parental emotional warmth and psychological Suzhi as mediators between socioeconomic status and problem behaviours in Chinese children". In: *Children and Youth Services Review* 59 (2015), pp. 132–138.

[140] Xiaocan Wu, Yu-E Sun, Yang Du, et al. "An Efficient Malicious User Detection Mechanism for Crowdsensing System". In: *International Conference on Wireless Algorithms, Systems, and Applications*. Springer. 2020, pp. 507–519.

[141] Jian Xing, Miao Yu, Shupeng Wang, Yaru Zhang, and Yu Ding. "Automated Fraudulent Phone Call Recognition through Deep Learning". In: *Wireless Communications and Mobile Computing* 2020 (2020).

[142] Xingwei Yang, Rhonda McEwen, Liza Robee Ong, and Morteza Zihayat. "A big data analytics framework for detecting user-level depression from social networks". In: *International Journal of Information Management* 54 (2020), p. 102141.

[143] Yabing Yao, Ruisheng Zhang, Fan Yang, et al. "Link prediction in complex networks based on the interactions among paths". In: *Physica A: Statistical Mechanics and its Applications* 510 (2018), pp. 52–67.

[144] Josh Jia-Ching Ying, Ji Zhang, Che-Wei Huang, Kuan-Ta Chen, and Vincent S Tseng. "FrauDetector+ An Incremental Graph-Mining Approach for Efficient Fraudulent Phone Call Detection". In: *ACM Transactions on Knowledge Discovery from Data (TKDD)* 12.6 (2018), pp. 1–35.

[145] Jiahui Yu, Kun Wang, Peng Li, et al. "Efficient trustworthiness management for malicious user detection in big data collection". In: *IEEE Transactions on Big Data* (2017).

[146] Xiao Yu, Quanquan Gu, Mianwei Zhou, and Jiawei Han. "Citation prediction in heterogeneous bibliographic networks". In: *Proceedings of the 2012 SIAM International Conference on Data Mining*. SIAM. 2012, pp. 1119–1130.

[147] Yinan Yu, Hailiang Chen, Baojun Ma, and Benjamin P. C. Yen. "Utilizing Geospatial Information in Cellular Data Usage for Key Location Prediction". In: *Proceedings of the 51st Hawaii International Conference on System Sciences*. 2018.

[148] Yahan Yuan, Ke Ji, Runyuan Sun, et al. "An Integration Method Of Classifiers For Abnormal Phone Detection". In: *2019 6th International Conference on Behavioral, Economic and Socio-Cultural Computing (BESC)*. IEEE. 2019, pp. 1–6.

[149] Jiaquan Zhang, Hui Chen, Xiaoming Yao, and Xiaoming Fu. "CPFinder: Finding an Unknown Caller's Profession from Anonymized Mobile Phone Data". In: *Digital Communications and Networks* (2021).

[150] Jiaquan Zhang, Xiaoming Yao, and Xiaoming Fu. "Identifying unfamiliar callers' professions from privacy-preserving mobile phone data". In: *2020 16th International Conference on Mobility, Sensing and Networking (MSN)*. IEEE. 2020, pp. 524–530.

[151] Jun Zhang, Feng Xia, Wei Wang, et al. "Cocarank: A collaboration caliber-based method for finding academic rising stars". In: *Proceedings of the 25th International Conference Companion on World Wide Web*. 2016, pp. 395–400.

[152] Yongfeng Zhang, Xu Chen, Qingyao Ai, Liu Yang, and W Bruce Croft. "Towards conversational search and recommendation: System ask, user respond". In: *Proceedings of the 27th acm international conference on information and knowledge management*. 2018, pp. 177–186.

[153] Song Zhanwei and Liu Zenghui. "Abnormal detection method of industrial control system based on behavior model". In: *Computers & Security* 84 (2019), pp. 166–178.

[154] Qianqian Zhao, Kai Chen, Tongxin Li, Yi Yang, and XiaoFeng Wang. "Detecting telecommunication fraud by understanding the contents of a call". In: *Cybersecurity* 1.1 (2018), pp. 1–12.

[155] Le-kui Zhou, Yang Yang, Xiang Ren, Fei Wu, and Yueting Zhuang. "Dynamic Network Embedding by Modeling Triadic Closure Process." In: *AAAI*. 2018, pp. 571–578.

[156] Xiaokang Zhou, Wei Liang, I Kevin, Kai Wang, and Laurence T Yang. "Deep correlation mining based on hierarchical hybrid networks for heterogeneous big data recommendations". In: *IEEE Transactions on Computational Social Systems* (2020).

[157] Jinghua Zhu, Yong Liu, and Xuming Yin. "A new structure-hole-based algorithm for influence maximization in large online social networks". In: *IEEE Access* 5 (2017), pp. 23405–23412.

[158] Yu Zhu, Jinghao Lin, Shibi He, et al. "Addressing the item cold-start problem by attribute-driven active learning". In: *IEEE Transactions on Knowledge and Data Engineering* 32.4 (2019), pp. 631–644.

[159] Xianglin Zuo, Bo Yang, and Wanli Zuo. "Exploring uncertainty methods for centrality analysis in social networks". In: *2017 IEEE International Conference on Data Mining Workshops (ICDMW)*. IEEE. 2017, pp. 163–169.

[160] Yuan Zuo, Guannan Liu, Hao Lin, et al. "Embedding temporal network via neighborhood formation". In: *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2018, pp. 2857–2866.

List of Acronyms

List of Figures

List of Tables