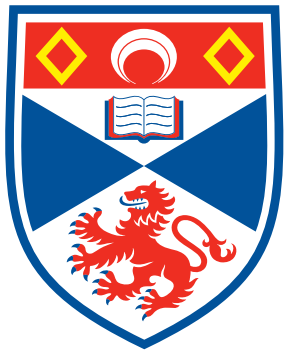
Protecting Online Privacy through Self-disclosure and User Identifiability

Sidi Zhan



University of
St Andrews

This thesis is submitted in partial fulfilment for the degree of

Master of Science by Research

at the University of St Andrews

June 2018

Abstract

As Web 2.0 technology thrives, Internet brings people social benefits apart from other advantages. More and more people disclose themselves online in exchange for better social support from other Internet users, which brings along privacy issues. Social Media users still pay insufficient attention to protect privacy although they are concerned about their privacy being intruded. This privacy paradox helps to form the research question in this thesis.

There are some studies focused on the measurement of social capital and social support during Social Media usage, and also corresponding online privacy issues. Most of them adopt social science questionnaires, online or offline, to measure social factors and privacy level. Measurements are investigated to quantify the extent of privacy risk. Among them, user identifiability is an accurate metric showing how easy the users can be identified through their online content. Currently, most of the user identification indicators only yielded through simple and explicit calculation. To fill the research gap of superficial measurement of user identifiability, and furthermore the insufficient evidence supporting whether over self-disclosure has impact on privacy risks, the research question of this thesis is proposed.

To answer that, experiments are conducted, where question author prediction is enhanced via equivalence class and the concept of user identifiability is redefined. It is found from the results that by 28 rounds of predictions, using top four feature sets and tree classifier, the accuracy can hit 76.0%. This shows that self-presentation features can indicate the extent of user identifiability. However, further results do not support any significant relationship between these two factors. The main contribution of this thesis would be that the study provides new metrics to represent self-disclosure and user identifiability, and thus be able to describe the extent of user identification covering user content generated by the users.

Declaration

Candidate's Declarations

I, Sidi Zhan, hereby certify that this thesis, which is approximately 23,200 words in length, has been written by me, that it is the record of work carried out by me and that it has not been submitted in any previous application for a higher degree.

I was admitted as a research student and as a candidate for the degree of Master of Science by Research in 27th August 2016; the higher study for which this is a record was carried out in the University of St Andrews between 2016 and 2017.

Date: 27 June, 2018 Signature of candidate: SIDI ZHAN

Supervisor's Declaration

I hereby certify that the candidate has fulfilled the conditions of the Resolution and Regulations appropriate for the degree of Master of Science (by Research) in the University of St Andrews and that the candidate is qualified to submit this thesis in application for that degree.

Date: 27 June, 2018 Signature of supervisor: TRISTAN HENDERSON

Contents

Chapter 1

Introduction

As Web 2.0 continues to flourish in this information era, social media's role in our daily life becomes more and more significant. Ever since the first social networking website, SixDegrees.com, launched in 1997[9], people never stop expanding their social net2working onto the Internet. The most popular social networking site in the world, Facebook, has 2,234 million active users (recorded in April 2018 [1]), which surpasses one third of current worldwide population. Each average day, 1.45 billion people log in to Facebook, and these daily active users contribute 510,000 comments, 293,000 statuses and 136,000 photos in every minute(assessed on May 13, 2018 [2]).

The high activity of online social networks shows that people are keen to present themselves on Internet via social interactions. In the process of releasing about their life, users post personal information intentionally or accidentally. Some personal information might be so sensitive that adverse parties can use the information to identify the subject user and conduct privacy harassment.

This chapter, as the beginning of the thesis, will present the social benefits and privacy risks of social networking sites usage, and propose research problem to place privacy issue in this scenario under scrutiny.

[1] https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/
[2] https://zephoria.com/top-15-valuable-facebook-statistics/

1.1 Social Usage of Internet

People interact with each other in their social network to satisfy their social need through getting social capital.

1.1.1 Social Capital and Social Support

The concept of social capital is referred to in paper [57, Ch3] as the accumulated resources derived from the relationships among people within specific social contexts or networks. One type of social capital that comes from weak ties like neighbors from heterogeneous network brings mainly information benefits, while the other type of social capital coming from strong ties like friends and family emphasises the emotional benefits of homogenouse networks [52]. These two types of social capital are called bridging social capital and bonding social capital, and this way to divide social capital is widely received among scholars. These form of social capital can be gained regardless real or virtual communities.

Social support is one form of social capital. It is described as "a manifestation of social capital" in paper[20]. Kaplan's definition of social support in 1977 [29] is rephrased by Thoits in 1982 [56] as the degree to which a person's basic social needs are gratified through interaction with others. Many researchers gives their categories of social support[15, 47, 16, 44]. Although these theories vary a lot, main components are informational support and emotional support, and the rest usually are esteemed support, tangible or instrumental support, and companionship.

1.1.2 Social Media, Online Social Network and Social Networking Sites

As social media become increasingly important in our daily life, more and more people expand their social interactions online to get social capital. Online social network is by its literal meaning the social network built upon online community. Social networking sites (SNS), also known as social network sites or social networking services, are web-based services that allow individuals to (1) create a profile; (2) add a list of friends; and (3) interact with friends with self-generated content [4].

On SNSs, bonding social capital grows fast though weak ties among online friends. Mostly information support and sometimes emotional support are exchanged via the interactions like posting, forwarding and replying. People also use social media to represent themselves in order to gain social support of self-esteem. Ellison et al. survey previous work on how Internet use influence social capital, and find among three types of conclusions, the use of Internet can generate, or reinforce social capital of users[22]. In a word, social media like SNSs provide a platform for Internet users to continue, expand or explore their social life exceeding the limitation of time and space.

1.2 Online Self-disclosure and Self-presentation

Social networking sites, as one important type of social media, provide arenas for users to present themselves and interact with each other online, which facilitate their construction of impression they want to show to their audience. Concepts of these behaviours are defined as self-disclosure and self-presentation.

1.2.1 Self-disclosure and Personal Information Disclosure

The process of "making the self known to" another person is called self-disclosure in Joinson et al.'s work from Trepte's book on online privacy[57, Ch4]. Joinson et al. also mention studies to support that both disclosing self and knowing others result in greater liking. Further more, deeper disclosure can help users to receive more personalised services. All these benefits encourage users to disclose themselves on the Internet.

1.2.2 Self-presentation and Impression Management

In both online and offline interpersonal communication, people present part of themselves to others to construct the impression that others might make of them, which is called self-presentation. Presented personal information can be individual tastes and likes, hobbies, physical appearance, or even names and addresses. Internet and Web 2.0 enable users to selectively reveal slices of their life, in order to display only the desired aspects of self-image in front of their observers. With the help of SNSs and domain-specific virtual communities, users gain

more control over their self-presentation behaviour. Kramer and Haferkamp [57, Ch10] review literature to elaborate the definition, usage and factors of online self-presentation in their work, and cover most of the topics related to it. They also compare concepts of self-disclosure and self-presentation, noting that self-presentation focuses more on the quality of information and the influence of the impression.

Social media not only serve as a platform for people to disclose themselves to a larger audience regardless of temporal-spatio limitation, it also allows users to manage their self-image through selective self-presentation. This can be achieved by concealing the aspects of self that users don't want others to know and only showing part of themselves, so that the degree and aspect of disclosure become controllable.

1.3 Online Privacy and User Identification

Along with benefits, online self-disclosure also brings potential threats of privacy. In this process, users reveal their personal information either intentionally or accidentally. Gross's study [23]shows in Facebook 90.8% of profiles contain an image, 87.8% of users reveal their birth date, 39.9% list a phone number (including 28.8% of profiles that contain a cellphone number), and 50.8% list their current residence.

Excessive personal information release might bring privacy concerns. Brandimarte et al. [11] shows the scenario of control paradox where people are more willing to disclose sensitive data if they perceive more control over the the release and access of the data. They conduct three user studies to examine how perceived control over personally identified information affect people's decision of revealing them, indicating that privacy concerns are affected by control over release of personal information. More self-disclosure brings more social support and more precise service, while it also brings potential privacy threats at the same time. This paradox urges people to identify the typical problems of online privacy, and come up with control strategies.

1.3.1 Potential Threats on Privacy

Personal data, which are any information related to an identified or identifiable natural person (definition given by 1995 EU Data Protection Directive[54]), includes personal attributes, writing style, network structure location, real name, birth date, age and so forth. Typically, privacy sensitive personal data leakage might cause privacy-related threats.

Possible harassments come from all kinds of reasons, influencing aspects from virtual world to reality. Firstly, individual features can be exposed not only through public user profile, but also the posts and blogs authored by the users or even their friends. Although individuals pay enough attention on concealing personal information, their friends on the same OSN will inadvertently reveal about them. The work of Lin et al [33] shows that friends' posts on a user's profile can inadvertently reveal details such as birthday (reveal rate of 87.0%) and educational background (91.7%), which is called friends annotation.

Secondly, for those anonymous users, their online identity can be disclosed under the assistance of released personal information, external name list, or social network structure.

Finally, real life identity could be detected even though users use multiple virtual identity accounts in those websites not requiring users to register by their real name. Real-time temporal-spatio status can be used to stalk users and conduct physical attacks.

1.3.2 User Identification

User identification is typically a behaviour of privacy threats. De-anonymising a user account, linking the account to other platforms, tracking down a user's real life identity, detecting the real author of a post, all of these could be considered as forms of user identification.

To solve online privacy problem, first of all, the causes or factors of privacy leakage need to be found out. Self-disclosure is one of the originalities of privacy issue in social media. The research topic of this thesis lies in this domain. To find out the relationship between self-disclosure and user identification, self-disclosure is represented by identifiable features, and the degree of identifiability is measured.

To elaborate the research context, a simulated SNS user, U_A, is taken as the focal user, and her ego-centric social network (egocentric social network, or ego-network is a network extracted from online social network, with one user as centre and all neighbours closely related to the user. The concept is adopted from [18], measured by number of followers and number of followees) is built online, as depicted in Figure 1.1.

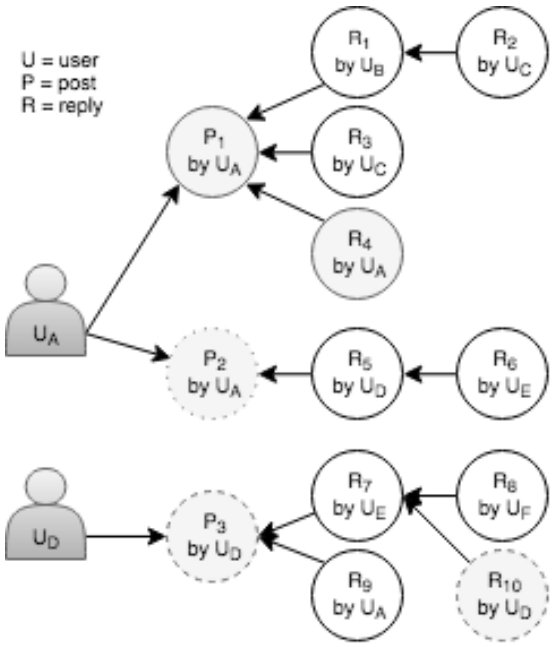


Figure 1.1: an example in SNS
Note: this is an example of U_A using SNS. If U_A's replies to other users' posts are ignored, P_1, R_4 and P_2 are contents that disclose about herself. P_3 and R_10 are U_D's self-disclosure.

This thesis will look at identification of the central user, U_A from Figure 1.1, based on the posts among the user and other users interacting with them, which are P_1, R_4 and P_2.

1.4 Structure of the Thesis

This thesis explores online privacy problem by finding its correlations with self-disclosure behaviours. The rest of the thesis will be organised in the following structure.

In the next Chapter 2, related works will be reviewed about quantifying social benefits and privacy risks online. It is found that few researchers have looked at the measurement of privacy,

as well as user identifiability, in a data analysis aspect. online privacy solutions will also be summarised, so that possible approaches according to the findings can be suggested.

The Methodology Chapter 3 mainly introduces the system of methods being used in the study. From proposing research question formally to evaluating the model, it will be elaborated that how the research gaps found in previous work will be filled. First of all, there is a whole section specifying the research topic, conveying the idea of research question and the hypotheses for it, and demonstrating how the hypotheses are going to be tested. Then, the reasoning and procedure of methods will be separately described in details.

Conclusions are then drawn and the future work explored after the results have been presented.

The main contributions of the thesis are as following:

1. it discusses online privacy and self-disclosure relationship in a broader context, by analysing large amount of posts in SNSs;
2. it finds systematic metrics for privacy risk in SNS context, and to describe the extent of user identification covering user content generated by the users;
3. and, it proposes suggestions for individuals and developers to protect online privacy in SNSs.

Chapter 2

Related Work

With the background of the thesis introduced in the last chapter, it is clear that social media users expose themselves in exchange of more specific social support, and this behaviour might also bring privacy risks. In this chapter, related studies in literature will be explored, first for measurements of social benefits and privacy risks, in order to find existence of social benefits - privacy risks trade-off; and then for measurement of user identification, to further represent the degree of privacy menace.

This study finds measurement of user identifiability, and uses it as an indicator for privacy risks in SNSs. In the search of privacy protection methods, the thesis introduces general scenario by listing different kinds of personal data protection solution, and then goes on to explore applicable approaches in SNS context, such as disclosure control with obfuscation.

Finally, from what is reviewed in literature, the research gaps to be tackled with will be pointed out.

2.1 Measuring Social Benefits and Privacy Risks in SNSs

Based on the two components of SNSs, being the users that constitute social networks and the content generated by them, the measurement in SNSs can be realised from investigating user profiles and relationships or mining the content. The measuring mechanisms might change due to various research goals.

2.1.1 Measurement of Social Capital and Social Support

The study of the metric of social capital and social support started in nineteen nineties. Researchers use questionnaires to assess the social benefits of participants by segmented categories. Even when scenario shift from real-world community to virtual communities on the Internet, the measuring approaches do not change too much.

Measurement of Social Capital Ellison et al.[21] adapt 19-item 5-point Likert scale measures of social capital and use them in Facebook users' bridging, bonding, and maintained (of their own creation) social capital. They also present measurements for Facebook usage, self-esteem and satisfaction of life, finding that social networking sites usage strongly connects with social capital, especially the bridging type. They exploit and develop these measures in their later studies[52, 22]. Shuai et al. [51] classify social capital in OSNs as strong and weak ties and measure them by ratio between the number of strong ties and weak ties. Apparently, they refer bonding social capital as strong tie, while bridging social capital as the weak one. They use these proxy to represent user behaviour of cyber-relationship addiction and information overload, and detect mental disorder in OSNs.

Measurement of Social Support Sarason et al. [47] propose the 27-item Social Support Questionnaire from previous works in 1983 to measure number of supporters and satisfactory of perceived support at the same time, and it later became a widely used psychometric questionnaire. Heitzmann and Kaplan[26] review 22 social support assessing scales and suggested researchers to choose a measure that matches their conceptualization. Drentea and Moren-Cross [20] study how mothers support each other online by observing an Internet mother site, and found two prevalent types of support, emotional support and instrumental support (i.e. information sharing).

Many researchers use social support as a proxy when they study social media's impact on physical or mental health and well-being. As early as 1999, Braithwaite et al. [10] propose a typology of social support in online settings to study their type, extent and pattern in computer-mediated context, and compared computer-mediated social support groups with non-mediated ones. Oh et al. [44] measure health-related social support (HRSS) in four dimensions, emotional, informational, tangible, and esteem, using an 18-item 7-point Likert scale measure of online

social support, and found nearly 40% users in Facebook had sought HRSS. Keating [30] studies computer-mediated support group and asks two human coders to code messages into three categories (Informational support, emotional support, and network support), using Cohen's Kappa for overall intercoder reliability. Zwaan et al. [58] use this typology in their conversational agents, providing users empathic virtual buddies as company.

There are computer-calculable instruments for social support in online communities. Some researchers find textual patterns of posts indicating social need and then use text mining approaches to measure social support. Wang et al. [63] use machine learning methods to classify posts in online health forum on five social support types (companionship, seeking informational support, seeking emotional support, providing informational support and providing emotional support), by training the model with manually annotated dataset. They capture writing styles or linguistic preferences of the posts by extracting basic features, lexical features, sentiment features, and topic features. Zhao et al. measure peer-supporting behaviour in online health communities as part of criteria to find influential users[65]. In their study, social support, especially emotional support, can be measured by users' emotional change, using sentiment analysis.

2.1.2 Relationship to Privacy Risks

The ways to quantify privacy danger level have also been studied, and then go further to look at the correlation between benefits and risks. Whether and how a certain privacy setting, like profile visibility control, segregating audience, content access restriction (or content distribution from the sender's point of view), or search engine searchability, is employed is one simple approach. This approach is dependent to the service provided by the SNS providers. Besides, the result of measurement is not scalable because the length of it is limited. Ellison et al. [22] question about negotiating privacy concerns and social capital needs in social media. They suggest types and audience of disclosure on SNSs complicates the relationship. They measure social capital gained, disclosure extent (number of friends) and privacy settings (the use of advanced privacy settings) through interviewing 299 Facebook users, and find that more advanced privacy settings and more friends indicates more social capital.

Many Information System researchers examined concept of privacy calculus to measure utility and risk, and their impact on decision of whether to disclose personal information. Privacy calculus [36] is a mechanism where the risks and utilities of privacy issue are traded off to make privacy-related decisions like, for instance, how people choose over coarse-grained or fine-grained location sharing options according to privacy risks and benefits[31]. Majumdar and Bose [36] summarise types of privacy benefits (financial rewards, personalization, and social adjustment) and risk (social, financial, time, psychological and physical) from literature, among which, social factors take both large part. They study potential privacy benefits and risk when employing the theory in the context of two information techniques, Bring Your Own Device (BYOD) and Internet of Things (IoT) by conducting interviews and focus group interviews.

Privacy risks measurement mechanisms in SNSs are decided by their types. How easy the user identity, both the account in online system or real life identity, is detected can also be a proxy of privacy risks measurement. elevant methods in paragraph 2.2. Vitak [60] find in study that privacy impacts disclosure and disclosure predicts bridging social capital in SNSs. The researcher measured disclosure by Wheeless and Grotz' (1976) General Disclosiveness Scale on the two subscales of amount and intended disclosure, bridging social capital by modified Williams' (2006) 10-item bridging social capital scale, privacy setting by the use of friend list, and privacy concerns by 10-item instrument on posting concerns.

2.2 Measuring Privacy Risks by User Identification

As discussed in previous section, threats of privacy in cyber communities can have various forms. Here the thesis mainly talks about privacy risks as personal information loss and the identification problem caused by it.

Researchers who are studying user identification sometimes place themselves as attackers or advisers, to quantify these risks. Four papers shown in Table 2.1 indicate that violating online privacy by re-identifying or de-anonymizing is a favoured choice for these researchers.

Table 2.1: papers on user identification risk

paper	source data	measurement	features
Becker [7]	Profile and posts of Facebook users and their social contacts	proportion of personal attributes related to participant users	attributes in users profile and their friends' posts
Narayanan and Shmatikov[41]	follow graph of Twitter, contact graph of Flickr and friend graph of Livejournal	re-identification rate	social graph structure, like nodes, edges and cliques
Ngoc et al. [42]	blog sentences, SNSs events and age statistics	information entropy difference	events and their properties in blog sentences
Barakat [5]	Facebook posts revealing location, and their corresponding likes, comments, tags, and replies.	threshold for dangerous posts	detection patterns like time and location prepositions, keywords following '@' symbol, specific verbs and other common phrases

Note: papers on user identification risk.

2.2.1 Identification Approaches and Metrics

User profile revealing lots of private demographic attributes is a breeding ground of identification attacks. Chen [13] use two game-theoretic methods to demonstrate in his doctoral dissertation that online social network users tend to reveal profile attributes that have a larger impact on the privacy as well as the social capital. In Gross and their colleagues user study on Facebook[23], 45.8% of users list birthday, gender, and current residence. This shows high possibility of re-identification by demographic attributes revealed in user profile by user themselves.

Privacy-sensitive phrases in user content are time bombs that are easily disregarded. Metrics for privacy risk can be calculated via privacy-sensitive phrases. Barakat and Magel [5] detect key phrases by patterns like time and location prepositions, keywords following '@' symbol, specific verbs and other common phrases in social networking sites to recognise dangerous posts. The false alarms rate, that is the proportion of 'dangerous posts', of their awareness system reaches 11%. On social media like blog, Ngoc and their colleagues [42] present a privacy metric based on probability and entropy theory . They suggest that adversaries in SNSs will ask questions about their interesting events, for example, the university of a user, and find answer in the user's blogs. They infer events and their properties in each blog sentence, e.g. named entities for locations and buildings of the university, and calculate information entropy, or cross-sentence

entropy if possible. The larger the information entropy grosses after a blog sentence is analysis, the greater privacy risk is. Selfies can also be counted as a piece of graphic private information, as face recognition technology develops rapidly in this era.

Social network structure adds to uncontrollable factors for the individual to be identified by other users in the network. Gross and their colleagues [23] employ a user's friends networks to verify whether the user is honest with user profile, because their interactions can help to check obviously erroneous information. Becker [7] suggests that even if a user control the access of his or her profile, privacy information indicated by the profile can still be inferred by his or her friends. So the researcher proposes an approach to measure privacy risk by looking at private attributes inferred by user's friends, and give several suggestions about how to manage friend lists to reduce these threats. Some researchers also measure the possibility of identification by social-network graph structure. Narayanan and Shmatikov [41] locate user in social graph by mapping the sanitized graph released by service provider after anonymisation and adversary graph crawled by unintended parties.

2.2.2 User Identifiability: levels of user identification risks

As personal information is referred here as "personally identified or identifiable information" (The EU Data Protection Directive of 1995 explained in paper [48]), any information that is clearly related to an individual not others or that can be used to identify the individual can be considered as personal information. An identifiable person is defined as "one who can be identified, directly or indirectly, in particular by reference to an identification number or to one or more factors specific to his physical, physiological, mental, economic, cultural or social identity" in the Directive. According to that, the concept of identifiability is extent of identifying a user while user identifiability is to describe how identifiable the user is with all the information related to them.

Categorizing degree of identification into simple levels is the most straightforward approach. It is often used to measure the identifiability of piece of data containing personal information. Gross et al. [23] proposed the concept of data indentifiability as how identifiable or granular the

provided data is, and did the categorization for identifiability of profile attributes like name (real name, partial name or fake name), birthdate (year, month, day), or profile image (identifiable, semi-identifiable, group image, joke image). Shen et al. [50] adapt an approach to measure identifiability of search query by entropy from information theory. They propose the four levels of privacy protection in personalised search, namely pseudo identity, group identity, no identity, and no personal information, by looking at whether the real identity or information need for re-identification is from one single user or shared among a group of users. These can be also considered as four levels of identifiability. Brandimarte et al. [11] measure identifiability by number of published uncertain publication condition, demographic items. The paper simply assumes that more private information released, higher identifiability would be. They conduct three user studies to examine how perceived control over personally identified information affect people's decision of revealing them.

Identifiability of users as a group is usually calculated when one user can be either identified or not identified. Narayanan et al. [40] produce high levels of identifiability in their experiment where they use stylometric features to identify the simulated anonymous posts (test posts) to the author with her other recognised posts (training posts): in 20% trials, three test posts (excerpted from blogs) are enough to predict one true blog out of 100,000 blogs. To measure individual, apart from using boolean indicator for the state of identified or not, numeric scores are also used. Iwaihara et al.[28] propose three level of user identifiability (identity factor, IF): let IF = 6 when the individual can be easily identified, IF=3 when the individual can be identified by a certain effort, and IF =1 otherwise.

2.3 Personal Data Protection Solutions

In the domain of online privacy, personal data protection has been an umbrella topic that never fall out of the spotlight. Personal data, like emails, photos, medical records, invoices, bills, payments, certificates, phone calls[61], are the main part of information loss causing privacy risk. UK (Data Protection Act 1998[2]), EU(Data Protection Directive 95/46/EC[54], General Data Protection Regulation[55]) and US(Principles for Providing and Using Personal Information[1])

public their own personal data protection regulations. Here the thesis discusses using personal data protection framework to reduce chance of user identification.

2.3.1 Local Managers and Personal Data Stores

As the prevalence of IoT, ubiquitous and pervasive systems, personal data can be generated by mobile devices, and all kinds of sensors any time and anywhere. It is necessary for data subject to use centralised data manager to store the sensor data and negotiate with service providers. Local data managers and personal data stores (PDS) or personal data vaults (PDV) are good solutions in this scenario.

PDVLoc [39] is a controller for location data sharing. It is a typical local data manager as the gateway between mobile devices and cloud services. Databox [17] is a model to manage personal data and exploit the economic value of them. It integrates the data stores, applications and data controller together, allowing data subjects to add the service providers they want and manage data by service level agreement at user level. Mydex [46] provides personal data store whose function is to negotiate privacy policy on data type, access requests, purpose, duration and sharing parties. Personal Data Lake [61] are proposed together with a new semantic serialization format to better describe properties of data to be pulled.

Components of the aforementioned framework usually contains a) data subject, service providers and data manager, b) data stores and their interfaces, c) access controller and data manager[38].

2.3.2 Access Management Agents and Privacy Query in Database

However, some data providers generate or collect data themselves, they will usually who want to get access to their resources.

Access Management Agents Access control in SNSs can also be implemented by service provider, by using friend lists and setting access policy. Tags are allocated to posts and audience

before the links are built[25]. Researchers mainly focus on recommending access control policies.

Privacy Query in Database Some researchers [32] mainly develop private information retrieval technology for database, with queries or protocols.

2.3.3 Data Sharing Protocols Among Service Providers

For those data already flowing into the Internet, their safety largely depends on those web service providers that exchange them. To cater for web services significant in quantity and variety, comprehensive between-services protocols are made and observed by those players in the field, protocols being OAuth 2.0[1], OpenID Connect[2], and UMA[3], to name but a few.

2.3.4 Encryption System for Data Transition

Encryption methods are used between end users and services or at database, with or without keys. Guha et al. propose NOYB to protect SNS users a steganography encryption method to protect privacy data by dividing them into atoms and apply dictionary to them[24].

Devices to protect personal data can be service providers that store users data, the protocol controlling the flow of data, or an encryption system applying locks or masks to the data.

2.4 Self-disclosure Control

The protection of personal data should be considered in whichever digital or online environment, but when it comes to SNSs, the release of personal data is not explicit as that of structured data. Data subjects might consciously or unconsciously disclosure their private information when they interact with each other online, or present themselves in the profile. This might cause the discover of either virtual or real-life identity. How to make users aware of their behaviours

[1] https://oauth.net/2/
[2] http://openid.net/connect/
[3] https://docs.kantarainitiative.org/uma/rec-uma-core.html

and help them to control excessive self-disclosure without compromising the service they receive is the main topic for researchers in the domain of social media self-disclosure control.

2.4.1 User Identity Protection

Protection Actions Taken by Data Holders For data holders, if SNS service providers want to public their data for research use, governmental surveillance, or commercial operations, the best way to protect their users privacy is to anonymise data dumps before releasing them[6].

In this sense, the privacy of data subjects from structured or semi-structured datasets can be protected by blocking identity sensitive attributes, decreasing uniqueness of attributes, and disconnecting the datasets from external name lists. This could be achieved by k-anonymity, a protection model proposed by Sweeney[53], making at least every k individuals in a dataset indistinct from each other in identity by modifying their information.

Approaches on User Side Social media users as personal data generators should also be involved in privacy protection process. Many tools or recommendations help users to do this.

Generalizing named entities in SNSs like geographic locations, ages, colleges or companies is a common approach. Nguyen-Son et al. [43] develop an algorithm to anonymise information that can identify the related users in SNSs, for instance, replacing the name of school with name of university or the city of university. The synonyms of the information, called fingerprint in their work, could also be used to detect a discloser of the information.

For those social networks sharing spatio-temporal information, which are derived from Location Sharing Services (LSS) or Location-Based Services (LBS) [45, Ch9], generalizing geographical information is also a useful technique to protect privacy.

Prevention of privacy leakage should also be conducted in the aspect of social network structure, that is connection between users like friends or friend list.

For privacy leaks caused by friend annotation, Becker [7] provides three heuristic solutions after measuring the privacy risks by inferring user's missing attributes from user's friends: removing random friends, friends with most attributes, friends with most common friends. Although it sounds impractical to take the researcher's advices.

In summary, disclosure control in social media could be generalization (replacing sensitive phrases with a more ambiguous synonym) and suppression (removing sensitive phrases). Private information revealing the identity of users can be personal-specific, spatio-temporal or textual. The idea of making data more ambiguous by adding noise can also be referred as obfuscation, which will be discussed in the next section.

2.4.2 Obfusctation Approaches

Obfuscation [12] is a privacy protection mechanism being used on personal data to prevent discrimination, prevent web search engine from profiling users by their queries, deceive predicting algorithms. The objective of this mechanism is to remove attributes from data that can help to reveal user's personal information against low-level adversaries like eye-witnessing, but might be not efficient for large scale privacy attackers like learning machines.

Obfuscation is widely adopted to anonymise public data dumps before them are released[64], to pre-process data generated by the sensors in smart and wearable devices or applications [66, 62], or to confuse user profiling attempt conducted by server providers like web search engine (WSE)[50, 27, 59]. Interpolating fake content into the real ones are common way to implement obfuscation tools.

2.5 Research Gaps

The benefits users gain from SNSs using can be measured as social capital or social support, while privacy risks coming along can also be measured. Most of the studies reported a positive relationship between social benefits and privacy risks. In fact, users disclosure more about themselves, the more precise they are identified by potential friends and helpers (also the ill-intended parties), and thus they will be faced with more benefits and risks at the same time. However, self-disclosure is more than realizing personal attributes, any identifiable information, implicit or explicit, can increase privacy risks. Few researchers studied the relationship between self-disclosure and privacy risks from a thorough and quantified aspect.

Identifiability, as the extent of being identified or identifiable, can serve to measure personal data containing private information or their subject users. Privacy sensitive profile attributes or phrases can be used to identify a user by their textual posts. Profile images can also do the job. The network of friends sometimes add up to the chance of the user being identified. From these metrics it can be concluded that the studies in this field investigate more about data identifiability than user identifiability, as personal information are more easy to calculate. They gave only simple and surface grade of identifiability, and most of them are restricted by the private items used in measurement. A more statistic metric system for user identifiability is still in need.

Knowing the metrics of privacy violation, to find out the ways to protect privacy is of importance. One big category of privacy protection solution is to protect personal data, so that privacy leakage from its source can be prevented. The protection of personal data can be achieved by adding data manager in personal data store locally or on the cloud, using third party agents or privacy query to access private data, or obeying data sharing protocols among service providers.

In the scenario of social network sites, privacy protection can be better achieved by controlling self-disclosure of the users, as SNS users are given much larger power to generate and manage their content. Most of the privacy protection approaches under this circumstance can be classified as obfuscation approaches. By leaving out sensitive personal attributes in user profile or posts, by generalizing specific named entities like temporal-spatio information or company department, noises are interpolated into original data flow in the scale remaining honest and consistent.

Many researchers have looked at finding methods to solve identification problem, but no one has tested the rationale of using those methods. How can self-disclosure behaviour in SNSs influences user's chance of being identified? Does it make sense to reduce the extent of self-disclosure, in order to preserve users' privacy online? Experiment need to be conducted to collect evidence and support this statement before exploring any privacy protection methods based on it. In next chapter, the thesis will discusses the design of experiment.

Chapter 3

Experiment Methodology

The goal of this study is to find how online self-disclosure influences the extent of user being identified. To achieve this goal, an example is firstly used from a community Q&A platform to elaborate the problem being targeted at. From the example, the relationship between users and posts are modelled. Within the research context, two main factors, user self-disclosure and identifiability, are clarified. After that, the thesis will bring up research question on relationship of self-disclosure and user identification, and propose two hypotheses to test whether self-disclosure matters and how it affects user identification. By summarising the process of the experiment, the thesis introduces entities and concepts to be used. Finally, it gives the strategies to test the hypotheses, so that the research question can be furthermore answered according to results, and thus final goal can be achieved.

3.1 Problem Statement

3.1.1 Research Context

Among social media, community-based social question and answering portals (henceforth CQA for Community Question Answering) are platforms where users exchange their knowledge by asking and answering questions. Typical CQA sites include first CQA site Naver Knowledge

iN[1], first English CQA site Yahoo! Answers[2],[49] and Stack Exchange[3], Quora[4]. As a category of social media services, CQA sites allow users to interact with their fellows by behaviours other than asking and answering questions, like voting and commenting[34]. However, they are distinct in their own way. Compared to services for social networking, such as Facebook[5] and Linkedin[6], recommending users to use their real name and share personal profile, CQA sites encourage users. Unlike expert question answering sites like justanswer.com and now defunct Google Answers where roles of users are separated and only vetted experts can answer the question, CQA communities provide all users chances to be the question proposer and answerer. It is CQA's complexity, subjectiveness and yet enclosure of personal profile[35] makes it our choice of research domain.

Take a query thread on Stack Exchange[7] for example, the screen-shots of all its posts, one question and two answers, are shown in Figure 3.1. In Stack Exchange, just as other CQA sites, there are users U asking questions Q, and other users interacting them by answers A and comments C all of this forming the query threads QT.

Figure 3.2. One user U_A ask a question under the topic domain of apple, about installing dual-boot system on her Mac Mini. She also attaches four tags to her question and submits one comments under it. The question receives two up-votes by viewers. Then user U_B and U_C give their answers respectively. U_C's answer receives one up-votes. Also, he comments on the U_B's answer. In this example, the asker U_A interact with the two answerers and potential visitors by posting one question and one comment. Although there are actions can be taken for safety, such as she does not reveal her personal information like real name, she can submit this question anonymously, and her question can be released in an anonymous data dump; her activity on the sites can still be used by adversaries, either ill-intentioned or unintended, to reveal her identity.

[1] kin.naver.com
[2] answers.yahoo.com
[3] www.stackexchange.com
[4] www.quora.com
[5] www.facebook.com
[6] www.linkedin.com
[7] visited on 21 Dec, 2017, from https://apple.stackexchange.com/questions/120311/can-i-dual-boot-mac-mini-with-windows-8

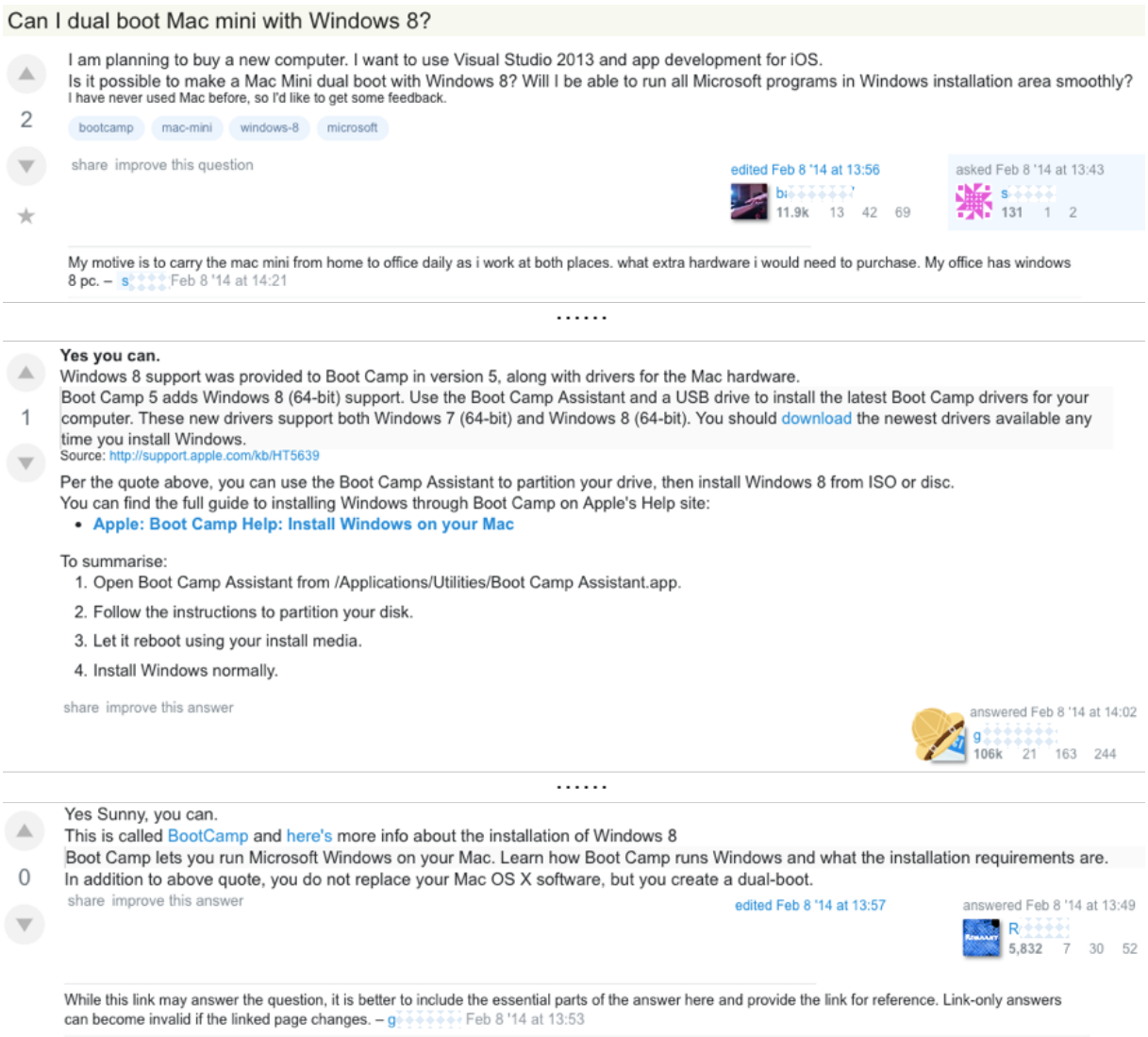


Figure 3.1: screenshot of one Stack Exchange query thread
Note: a screen shot of a query thread on stackexchange.

For instance, once her other questions are found under her name, her anonymous posts under the same pseudo name can be examined together to predict her as the author, and thus identify her.

To better understand our research objects, simulate one more question posted by U_A and one more user U_D, and portrait the heterogeneous CQA network[14] in Figure 3.3.U_A posts two questions, and sets up two query threads. For the first question, it gets two answers by U_B and U_C, and two comments by U_C and herself. Another user U_B also posts a question, and one of his two answers are contributed by U_A. To better address the issue, this study only analyses self-disclosure in relevant posts under U_A's query thread and submitted by her, i.e. Q_1, C_2 and

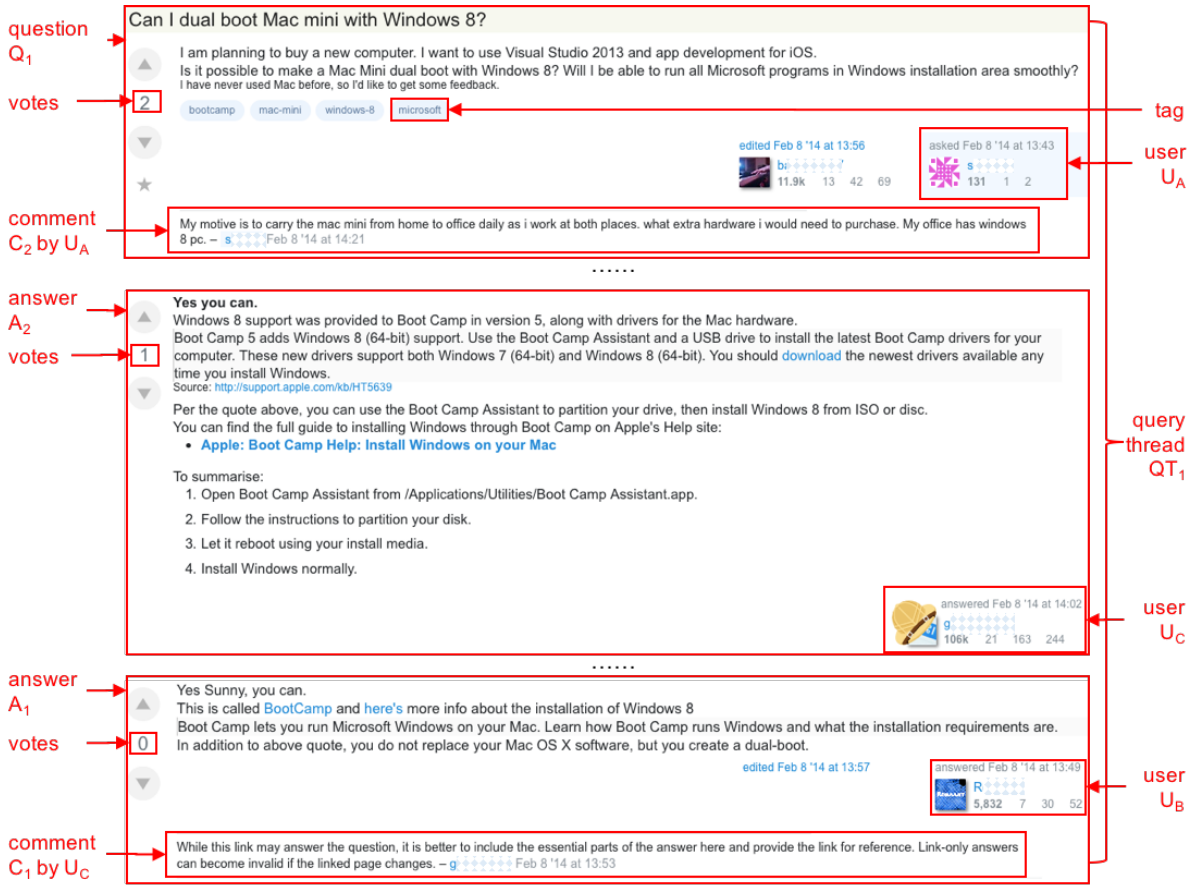


Figure 3.2: annotated screenshot of one Stack Exchange query thread
Note: posts in the order of creation time, are $Q_1 \rightarrow A_1 \rightarrow C_1 \rightarrow A_2 \rightarrow C_2$.

Q_2, as depicted in Figure 3.3. Those consist of all her two questions and her comment on the first question.

3.1.2 Research Thesis

From the network model, factors can be clearly extracted and researched in the study, self-disclosure and identifiability. To quantify the extent of self-disclosure and user identifiability, the thesis specifies their concept and propose measurements of them in CQA.

Self-disclosure Applying aforementioned concept of self-disclosure, self-disclosure behaviour in CQA users is recognised as bringing up a query thread by asking questions, or as replying to the query thread by providing answers and comments. The extent of it is measured by counting number of posts under their query threads. In the example of U_A and her fellows, U_A disclose her

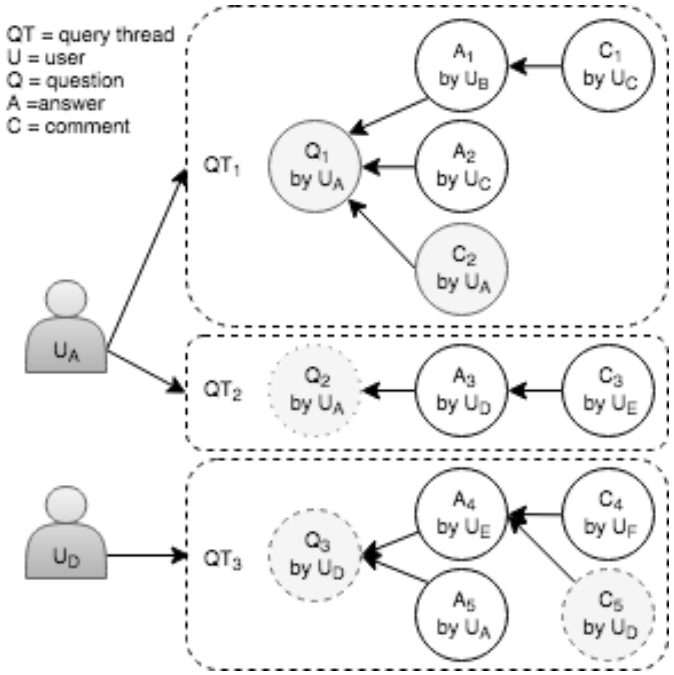


Figure 3.3: Synthetic example of Heterogeneous CQA Network
Note: this is an example of U_A using CQA. Only the grey-backgrounded circles under U_A's query threads are her relevant posts. From top to bottom, Q_1 and C_2 are relevant posts for the first question Q_1 (or query thread QT_1), Q_2 (i.e. the question itself) is relevant for the second question Q_2, and Q_3 and C_5 are relevant posts by U_D for the third question Q_3.

interests in computer operating system by asking about the choice between Windows and Mac OS. It can also be assumed that her style of making decision is to list her plans and enquire the feasibility. All these implicit information can be indicated from one question and two comments released by herself under the first query thread in Figure 3.3.

Identifiability Questions are posted by different users on CQA. In this experiment, actual author of a question is predicted by analysing posts under the corresponding query thread. If the true author is successfully found out for the anonymous question, then it can be said that the focal user is identified. In that context, identifiability is a measure of the extent to which a user is identified by her self-disclosure. This value is calculated over all the questions asked by the user, by considering her posts under these questions' query threads. Take a look at U_A and U_D in Figure 3.3, the identifiability of U_A is judged by the process in which the experiment assumes all three questions are anonymous. The posts considered are Q_1, C_2, Q_2, Q_3 and C_5. Separately, the experiment selects Q_1 and C_2 for Q_1, Q_2 for Q_2, and Q_3 and C_5 for the Q_3. Then by predicting U_A as the author for the first and the second question, the experiment figures out

proper metrics to measure her identifiability in this CQA. Those metrics are formulated from relationships between statistic values like equivalence class size, number of hit, and so on, which will be discussed in later sections.

The following question is proposed to inspire our research interest in CQA.

RQ. Does self-disclosure positively affect the extent to which users are identified?

To answer this research question,one hypothesis is proposed to test whether users' self-disclosure will help to identify themselves, and the other hypothesis based on the first one to test in what kind of tendency does self-disclosure affect identifiability.

H1. Users' self-disclosure in SNSs makes them identifiable.

This hypothesis describes the influence of user self-disclosure on identifiability. Therefore, it can be tested by analysing how content authors are identified by the information they disclose about themselves. This information contains not only personal information, which is explicit attributes, but also other implicit indicators that can be extracted from the content and their meta-data. The experiment will mainly look at the extent of identification and how significant it is when H1 is tested.

H2. The more users disclose themselves in SNSs, the more identifiable they are.

Hypothesis H2 discusses the relationship between user activeness and their identifiability in a dynamic way. It states that users' activism in disclosing themselves will promote their risk of being identified. To test H2, mechanism needs to be found to represent the variety of users online self-disclosure as well as the corresponding tendency of the identification risk.

3.1.3 Overview of Methodologies

Before a series of methodologies are presented, an overview of experiment process is given.

Experiment procedures Firstly, the experiment chooses CQA as the circumstance where our research question is inspected. It is a type of social media where users communicate with each other by asking and answering questions, as well as commenting those questions and answers under same query threads. The theme of query threads, content of all kinds of posts, and the network of replies form the overview of users' self-disclosure and interactions for us. Our research objects are users U and certain posts P under the same query threads QT, including the question Q, its answers A by the same author, comments C of the question by the same author, and comments of all the answers by the same author. These posts are called as relevant posts P (posts for short) to the focal user U. The relationships of the entities discussed in this experiment are described in the following formulas. $A \rightarrow Q, C \rightarrow Q, C \rightarrow A, P = Q \| A \| C, QT \ni P$, *where* $\|$ *means or, and* $\rightarrow$ *means dependent on.*

Secondly, feature sets (denoted as FS) about self-disclosure are extracted from relevant posts, and those feature sets that can actually contribute to identification are selected. The experiment reduces the size of input data by filtering every single feature (denoted as F) in the combination of those feature sets. After preparing the input data, classification method is used to estimate whether the user would be true author for the given question, by giving the user and the question a score (denoted as S). Scores for all question and user pairs constitute the score matrix. Sample and class label are the denotions for question and their author.

Thirdly, in order to calculate the identifiability more precisely, the experiment predicts more class labels so that recall rate will be increased. To achieve this, the experiment generates equivalence class (denoted as EC) in which predicted labels are regarded as equally likely to be true label, using some algorithms and setting different threshold values.

Finally, the experiment measures user identifiability and evaluate the model performance, after the prediction. All the notations of entities to be used to describe the methodology are listed in the Table 3.1.

Definition of equivalence class In the experiment to test hypothesis, he likelihood of a user to be the actual author of a question is predicted by giving the user-question pair a score. This score is usually calculated by calibrated estimators, the estimators that predict labels to be true label by giving each label a decimal probability. It can also suggest the confidence of the prediction.

Table 3.1: notations of methodology

notation	*entity*	*meaning*
U	user	The subject user in Q&A system, and he or she have submitted some questions.
QT	query thread	A query thread lead by a question, and it might have several answers and comments.
Q	question	A question in Q&A system, it is the head of a query thread, and it will always be the first post in that query thread.
FS	feature set	A feature set is a group of features extracted from original data according to one of their properties.
F	feature	A feature is a value in a feature set.
S	score	Score of a question-user pair, indicating the probability of the user to be the true author of the question.
EC	equivalence class	A group of users that predicted to be true author of the sample question.

Note: This notations might occur in formulas and algorithms of methodology.

The score is normalized between 0.0 and 1.0. The higher it is, the more likely the predicted user is true author as the most highly-scored one. To widen the prediction scope, the experiment continues to select more candidate users with close scores and announce that they are equally likely to be the true author. In logistic description, a set of criteria is made to mark the score of all users above the plateau as 1.0 and the score of the rest as 0.0. The class or cluster containing all these 1-scored users to a certain question is called as the equivalence class **EC** of that question.

In a nutshell, a sample question's equivalence class is a group of users that are predicted to have equal possibility to be its true author.

Hypotheses testing process H1 will be tested when the metrics about accuracy of the prediction are calculated. However, to test H2, it is necessary to exhibit the change of these metrics and the impact of self-disclosure extent. The experiment includes one post under the query thread into sub-dataset at each round of experiment, and then predicts authors from these posts, and finally calculates the metrics. The posts are ordered by their creation time chronologically, and they are all composed by the querier. The number of round indicates how many posts can be used as the resources which features are extracted from. However, as users will post different numbers of posts under their query threads, the actual number of posts might be smaller than number of experiment round. The minimum number of posts is set to control the quality of query thread when the dataset being used in the experiment is generated.

The above procedures, together with evaluation, make up the solution to the research question (shown in Figure 3.4), and answer the hypotheses one by one.

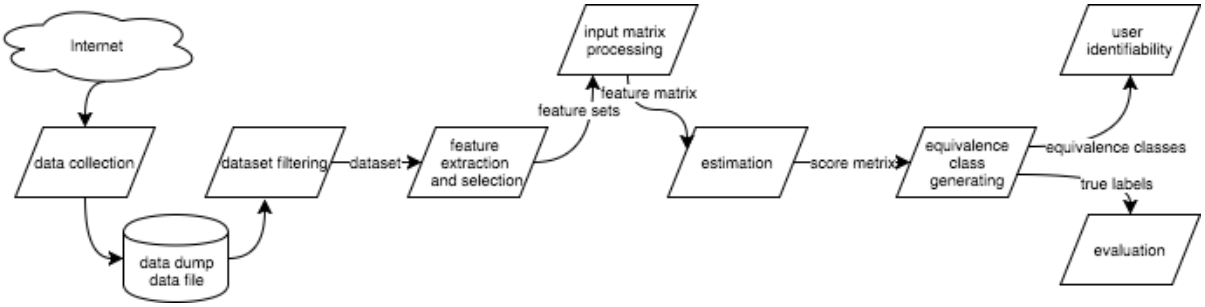


Figure 3.4: framework

Note: In each phases of experiment, there are several algorithms to choose from. The best combination is selected by comparing the devices in literature or doing local optimal test at each phase.

In the following sections, procedures of the experiment will be introduced in sequential order, with reasons of the selected methods given.

3.2 Experiment Setting

Once the main methodology of identification process is determined, the thesis sets out to go through the experiment framework.

3.2.1 Environment

Here it will be introduced running environment and programming language, software and hardware. The experiment uses a laptop with the configuration of Intel Core i5, 1.6 GHz, 8 GB, to run the experiment on the small dataset at model building stage, and a desktop with Intel Core i5-34705, 2.90GHz, 8 GB to run the experiment on the large dataset. When coding the model, python and its machine learning toolkits scikit-learn[8] are mainly used to implement the functional modules in experiment model.

[8]http://scikit-learn.org/stable/

3.3 Data collection and filtering

3.3.1 Dataset selection

There are two strategies being used for obtaining datasets, finding open data dumps from CQAs or finding Q&A related datasets from notable open dataset providers.

CQA sites comparison To find datasets to be used in the study, data dumps of CQA sites have been looked at, as well as several open machine learning or language study dataset providers. The availability and quality of open datasets are two basic conditions being relied on. Quality of datasets can be assessed by checking the existence of the entities desired for the study, depicting their relationships into a heterogeneous network as the example in Figure 3.3, and if conditions warranted, checking the inter-entity ratio to further guarantee the quality of dataset.

First, the experiment looks at CQA services like Quora, Yahoo! Answers, Stack Exchange, and Ask Reddit, and assesses them by making a comparison on whether they contain questions and users, whether they have released their data dumps and some criteria for their available data dumps. Then, it also searches for datasets in open dataset providers like Stanford Large Network Dataset Collection[9], Arizona State University SNS datasets[10], UCI Machine Learning Repository[11] and Yahoo Webscope Program[12], with keywords “question”, “answer”, “ask”, and “qa”, “q&a”. This action does not bring remarkable result. Table 3.2 gives a comparison on what is gotten.

As shown in Table 3.2, only the data dumps of Stack Exchange and Stack Overflow satisfy the criteria of identifiable authors and textual feedback. Since Stack Exchange provides more

[9] https://snap.stanford.edu/data/
[10] http://socialcomputing.asu.edu/pages/datasets
[11] https://archive.ics.uci.edu/ml/datasets.html
[12] https://webscope.sandbox.yahoo.com/
[13] answers.yahoo.com
[14] L6 - Yahoo! Answers Comprehensive Questions and Answers version 1.0, https://webscope.sandbox.yahoo.com/catalog.php?datatype=l
[15] www.quora.com
[16] stackexchange.com
[17] https://archive.org/details/stackexchange
[18] www.reddit.com /r/AskReddit
[19] https://snap.stanford.edu/data/web-Reddit.html

Table 3.2: comparison of datasets

CQA	*organization*	*data dump*	*release time*	*format*	*identifiable author*	*visitor feedback*
Yahoo! Answers [13]	A question has tags, a title and a content, and several answers. It can be followed. Answers can be commented on, and voted up or down.	a language dataset [14] on The Yahoo Webscope Program.	2007	all the questions and their answers, with best answers marked	No	No
Quora [15]	A user asks a question, the question has multiple tags (topics), a title and a content. Other users give their answers with content. Questions and answers can be up-voted but cannot be commented.	No public datasets available	Na	Na	Na	Na
Stack Exchange [16]	A question belongs to one community, and has several tags. Other users give answers to the question. It can be followed. Both question and answer can be commented and voted.	data dumps by communities can be found at their archive [17].	2017	XML files with users, questions, and their answers and comments	Yes	Yes
		Stack Overflow data dump	2011	like Stack Exchange data dump but much smaller	Yes	Yes
Ask Reddit [18]	A question is posted as the head of thread under a topic, and it can be answered in the form of comments. Answers can be replied, and replies can be replied too. Both answers and questions can be voted up or down.	a user has shared dataset with all the comments in 2015 [19]	2013	csv files on submissions with image, 132,308 textual submissions	Incomplete	Yes

Note: four CQAs are compared in this table.

up-to-date and diverse data dumps than Stack Overflow, it is chosen to generate the experiment dataset.

Stack exchange Now let us take a closer look at the dataset and corresponding CQA having been chosen. Stack Exchange is a typical Q&A system focused on various communities (i.e. topic domains), 3.7 million questions and 5.0 million users[20]. As described in comparative table Table 3.2, the data dump of Stack Exchange has all the entities and relationships needed, including users, questions, answers and comments under the same query thread. This information is stored in a pack of comma-separated text files (affixed with csv), from which the network of users and their posts can be easily extracted.

[20]statistics by 2015, from https://stackexchange.com/about

3.3.2 Data hygiene

Filtering rule Since the entity network generated from row dataset is large and sparse, it is better to filter the dataset and make it suitable for experiment. Several data filtering criteria are set according to experiment operations.

Rule I Firstly, it should be guaranteed that under each query thread at least four answers or comments generated by the same author of the question can be found. By doing so, there will be at least five relevant posts per query thread for the focal user, and a relative steady increase of number of posts can also be ensured, when one post is added into dataset at each round of experiment.

Rule II Secondly, because of the five-fold cross-validation method being used in calibrated estimating process, it is important that every one-fifth subset includes cases for all users, so the quantitative rules are set that each users in the dataset should composes at least five questions that meet the Rule I. Full justification to this will be given in the section on estimation.

Datasets generating There are in total 170 communities (or sites of topic domain) in Stack Exchange, for instance, Stack Overflow is the biggest community for programmers, with fifteen million questions and twenty-three million answers accumulated in over nine years[21]. Several strategies are tried to extract entities from communities to form dataset. A pilot dataset is first generated to build the network of relationships between entities. Then, datasets are generated from multiple communities or one single community, and are used in model building and coding process. Decision is finally made on an experimental use dataset generated from the data dump of a intermediate-sized community “serverfault" (with query threads on the topic of server fault). It has 253,000 questions, 421,000 answers, 341,000 users and 78% questions having been answered. Reasons are, a) this can help to centralise the questions in one community and exclude the impact of topic domain on users, b) the dataset meets quantitative rules well. Demographic characteristics of the new dataset are summarised in Table 3.3.

Figure 3.5 gives a clear view of the frequency distribution of two ratios. It can be found from sub-graph (a) that most of query threads have less than ten posts, and from sub-graph (b) that

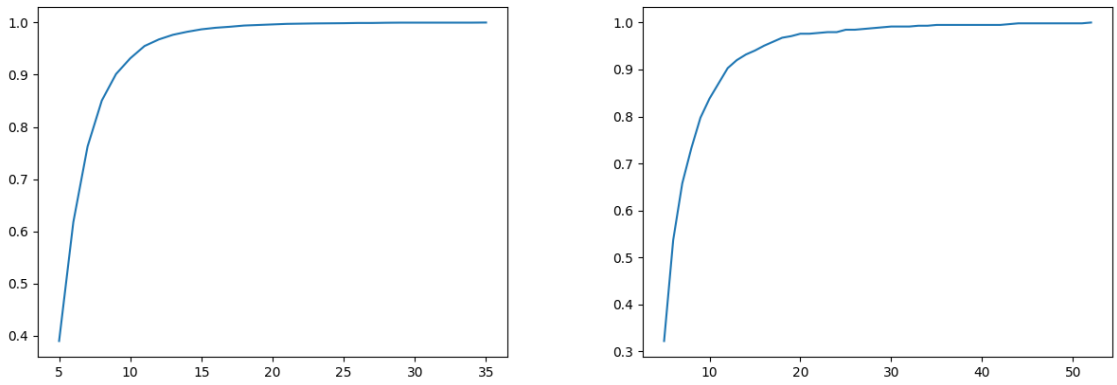
[21] statistics by 30 November 2017, from stackexchange.com/sites

Table 3.3: demographics of the original and filtered dataset

dataset	*original*	*filtered*
data sources	All from serverfault community	Part from serverfault community
restriction	-	users with ≥5 questions, query thread with ≥5 posts
users	9628	587
questions	108740	4682
answers and comments	401631+813288	26796
tags	3536	1696
posts per query thread	-	min = 5, max = 35
questions per user	-	min = 5, max = 52

Note: the experiment dataset is filtered from original dataset, and significantly refine the dataset.

most of users have fifteen qualified questions. The sub-dataset of this dataset with only questions is adopted to select feature sets and estimators.



(a) query thread (question) vs. posts

(b) users vs. questions

Figure 3.5: Cummulated frequency distribution of quantitative relationship between entitites
Note: This figure shows quantitative relationship. It can be indicated that how many posts by the same author that can be found under a query thread, and how many questions a user can ask.

3.3.3 Round of prediction

To test the relationship between the extent of self-disclosure, i.e. the number of posts user proposes, and her identifiability, multiple rounds of prediction with different sizes of dataset are conducted in this experiment. In each round, a subset of original dataset, which contains all posts, is generated from the first to the current, in each user's list of related posts. To generate

one subset per round of prediction, all related posts in one query thread for a certain user is first ordered by their creation time, from early to late. Then, in each round, one post are put in order into the dataset. Dataset for the first round includes all first related posts of all users, i.e. the questions in query threads. In the second round, sequentially-created related posts are appended to the previous dataset to generate the dataset for this round, so each user will have two related posts in the second dataset. So as the dataset for the third and forth and later rounds. Due to **Rule I** in data hygiene process, all users are guaranteed to have five related posts in the fifth round, but some users might not have more than five related posts. Therefore, after the fifth round, it is only appended the next related posts of those who have that much posts under their name. Take the entity network of U_A and U_D in figure 3.3 for example, posts in the dataset for the first round are Q_1, Q_2 and Q_3; while the dataset for next round has Q_1, Q_2 and Q_3 from the previous dataset and also has C_2 and C_3. There is no a third round of prediction, because all query threads have no more than two related posts. In this experiment, the numbers of posts by round are shown in figure 3.6.

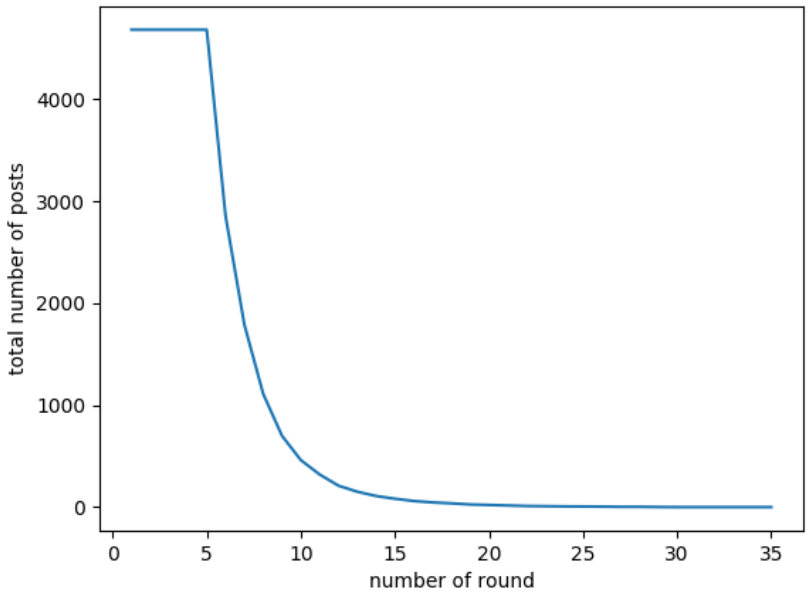


Figure 3.6: size of dataset each round

Note: this figure shows how many posts are there in each dataset. It can be inferred from the figure that there are at most 35 related posts in each query thread, and number of query threads, or questions, are over 4000.

3.4 Feature extraction and selection

Through feature extraction and selection, dataset is prepared as feature matrix that can be input into estimation model.

3.4.1 Feature extraction

In classification model, samples need to be coded by their features before they comprise feature matrix and enter estimators. In this case, properties of query threads related to user self-disclosure is used to extract corresponding groups of features, or feature sets. It should be avoided to built user profile by extracting their personal information like real name, date of birth, gender and so on. Instead, their characteristics should be indicated merely from the content of what they post, and the patterns of their activities. Eight feature sets are extracted from data set in three aspects, namely temporal usage patterns, textual patterns, and quality assessment.

Topic distribution Because of the homogeneity characteristic of cyber communities (that is, online users interact more frequently with friends of the common interests than other users)[37], users tend to present themselves through interests or hobbies. Topics extracted from user generated textual content thus can be used to describe a user in their interests to some extent. The idea of topic frequency distribution is realised by Latent Dirichlet Allocation (LDA), as Blei et al. present in their paper [8]. In the context of text modelling, LDA is used to calculate probabilities of a set of topics for each document. Documents are made up of words, and topics are represented as distribution of significant vocabulary. Thus, LDA builds a three-level hierarchical Bayesian model with words, topics and documents.

In this case, topic distribution is extracted from documents generated by concatenating relevant posts like questions, answers and comments, as feature set. Words are selected into vocabulary by filtering stop-words (words with high frequency or less meaning like we, are, from, etc.), and select 50 the most high frequency topics into feature set.

Stylometry Stylometry is a statistic description of writing style[3]. The feature set to it is shared by every textual document. It can be quantified by calculating symbols and words. For instance, what is used is the length of a word, a sentence or a paragraph; the count of typical

words, capital letters, long words or acronyms; the frequency of symbols, or punctuation like comma and dot; the syntactic structure of sentence and so on. Narayanan et al. extract writing style features and use them to identify author for blogs[40]. The top ten stylometric features are adopted in this thesis by information gain from their study, including frequency or count of punctuation symbols, words fitting certain patterns, and part of speech entities. Those features are:

1. Frequency of ';
2. Number of characters;
3. Frequency of words with only first letter uppercase;
4. Number of words;
5. Frequency of (NP, PRP) (noun phrase containing a personal pronoun);
6. Frequency of .;
7. Frequency of all lowercase words;
8. Frequency of (NP, NNP) (noun phrase containing a singular proper noun);
9. Frequency of all uppercase words;
10. Frequency of ,.

Tags and tag clustering Considering that each tag is used by merely four questions on average, it is suggested to make the feature set of tags denser. Following are the approaches being tried.

Firstly, the tags are clustered by tag pattern. A tag is a word (e.g. "ajax"), or a phrase of word or number linked with hyphens (e.g. "access-control-list", "500-error"). When all the tags are sorted by alphabetical order, it is found that some of them are both morphologically similar and semantically similar. For instance, "apache-1.3", "apache-2.2" and "apache-2.4" all begin with "apache" and they are different versions of Apache HTTP Server. The maximum common beginning of several tags is called a stem. If several tags beginning with same word(s)

are founded, they will be replaced by word(s), and the substitute will be the cluster label if the tags are clustered. On this base, a tag can be reduced into its stem, and then cluster tags into their stem by string similarity of their stems. For instance, if words are mapped like "hyper-v", "hyper-v-server-2008-r2", "hyper-v-server-2012", "hyper-v-server-2012-r2", and "hyper-v-server-2016" onto "hyper-v", the approach can reduce the number of tags by 4, and at the same time increase the density of tag usage.

Secondly, it is worth trying to treat tags selection as a feature dimension reduction problem. Only tag feature are used to calculate the prediction accuracy, with estimators that can rank the importance of each feature. Then, tags are selected by kicking out the ones with low importance, and keeping an eye on their performance by repeating the prediction every time a larger feature set is selected.

The last approach is to select tags by using frequency. Some tags are selected by hundreds of questions, such as the most popular "linux" is attached to 689 questions, while some are rarely chosen, for instance "screen-resolution" is used for only one question and there are 628 more tags as such. It is assumed that the more questions a tag is attached to, the less helpful it is to distinguish those questions. For users, if they use a common tag, they might be less identifiable by their questions than when they choose a rarely used tag. So tag popularity might also affect the uniqueness of a question and thus the identifiability of the author. To find out the mechanism for this, two feature sets are generated by excerpting and removing the top 100 frequently used tags.

Note that however the size of tags is reduced, it will compromise the performance of tag feature set. So the principle of doing so is to preserve the prediction preference. As the execution time and memory cost impact little on the results, all tags are kept while the experiment is running to test hypotheses.

Time slot or span Temporal patterns are extracted from two attributes of posts related to time, viz., creation time and last active time. The time slot distribution of creation time shows when the user usually uses CQA sites. The distribution has eight time intervals, 0-6, 7-11, 12-16, 17,

18, 19, 20 and 21-23, setted to even the frequency of each slot. The response time span is the difference between the last active time of last two posts. When each new related post is added into dataset, the response time span will be averaged among all the related posts under the same query thread. This shows the how frequent a user check and response their question. The last temporal feature set, active time span, is the difference between creation time of the first post, usually the question, and last active time of the last post. It keeps being refreshed as new related posts join the old dataset. It tells that how long a user will maintain their question.

Mean of score and count of comments The rest two feature sets are related to posts' quality assessed by other users. Score of a question or an answer is the sum of up or down vote by viewers. If a viewer agrees a post, then they will give an up-vote (score 1) to that post, if not, a down-vote (score -1) will be given. Count of comments of a question or an answer shows how many visitors are willing to reply to that post. It is also an indicator of popularity of the posts. This value is the mean of all related posts under a query thread currently in the dataset. All the eight feature sets being considered in experiment are listed in Table 3.4.

Table 3.4: feature sets

feature set	**feature set**	**description**	**number of features**
F_{st}	Stylometric patterns	10 writing style patterns of text content	10
F_{tp}	topic distribution	distribution of topics, i.e. keywords tuple distribution	50
F_{tg}	tag distribution	tags attached by authors when they ask questions	1696
F_{sc}	mean of score	score given by viewers, only for questions and answers	1
F_{cm}	# comments (count of comments)	count the number of comments under a question or an answer	1
F_{ct}	creation time distribution	distribution of post creation hour, to track the active time span	8
F_{rt}	response time span	the time span between last two posts' last active time	1
F_{at}	active time span	the user active time span from post creation to last edition	1

Note: the table shows eight feature sets extracted from original dataset, before they are selected and combined into dataset that can be used in this experiment.

3.4.2 Feature selection

Before data are fed into the training model, they need to be refined first. Because of the different levels of contribution from different feature sets to model performance, it is necessary

to find out how each feature set contribute to identification accuracy. By doing this, model performance can also be promoted, that is to a) increase the accuracy of prediction, b) reduce running cost by replacing with a smaller dataset. When different combinations of feature sets are tried and compared, it is founded that feature sets have low overall accuracy, and yet for a single feature set, their accuracy varies from each other's greatly. So it is reasonable to select those feature sets with accuracy high and growing as number of posts increases as the number of posts in dataset becomes larger. The total number of combination of feature sets that is needed to evaluate, also the number of times to train predictors, is $8+1=9$, where 8 single feature sets and 1 combination of the selected feature sets are counted. All of the eight feature sets extracted in the previous work are investigated here, producing the results in Section 4.1.1.

3.5 Input matrix pre-processing

After a feature matrix is generated from the feature sets being selected, the size of the matrix are reduced. However, this combination of feature sets can have large variation. Therefore, a last step is taken before it is used, which makes each element in the input matrix fluctuates in a unified range.

Normalization Normalizing a column can be done by converting it into a new column within the range of [0,1]. It can be implemented by scaling all the values within a column c using formula $c := \frac{c-min(c)}{max(c)-min(c)}$. One shortcoming of normalization is that if the data matrix are processed by column, it will diminish the sense of feature sets with multiple columns. Take topic distribution for an instance. Twenty topics are chosen and each document's distribution are computed upon them. These twenty features of a documents sum up to one, and their sum might be changed after being normalized. The best way is to normalize those columns with maximum values exceeding 1, and leave aside those feature sets representing probability (so their features add up to 1).

Dimension reduction Apart from feature set selection, the number of features can also be reduced by examining the contribution of a single feature, i.e. a column in the input matrix. For features that have few correlations between themselves, principal component analysis can

be deployed to assemble original features into less new features by linearly combining them. Estimators that grade each feature with a score of importance are also used to select the features with high importance. It is also useful to delete some meaningless columns, such as all-zero columns. Reduction at this stage will be witnessed to remain or even cut down the hit rate. However, this operation can still be adopted, if the objective is to shorten the processing time when the performance does not change significantly.

3.6 Estimation

After the feature matrix is prepared, they are used to calculate the probability of a certain user to be the true author of a certain question, in order to generate equivalence class as described in Section 3.7. This process is called estimation, just to differ from the whole process of true author prediction. Them are put into estimator that marks scores to demonstrate the probability, which is regarded as a machine learning question. Therefore, it is necessary to compare several proper estimators, and find the best matched to learn the marking process here.

Note that the operation of estimation is differentiated from prediction in this study, because of the specific procedure setting designed for the experiment. Firstly, machine learning models are used to estimate the scores of probability for all users, and then an equivalence class are predicted containing the subset of users that are equally likely to be the actual author.

3.6.1 Criteria of question-user scores

The score is referred to decide whether a user should be put into equivalence class or not, so the score needs to meet several standards. First of all, it should lie between 0 and 1, as it represents the estimation of probability that a user is true author of a question. Secondly, it should be variable enough to help us distinguish between predicted authors and non-authors.

3.6.2 Calibrated estimators

Estimators are in need to give distinguishable scores, in order to generate equivalence class by finding users equally possible to be true authors. Estimators are inspected, by comparing their characteristics that are useful in estimation. First thing to do is to check learning and estimating

algorithm of the estimator. Next thing to be examined is the occasion in which the estimator are chosen. To obtain the best performance of the model, the parameters should also be tuned. Some of them do not have parameters related to the input data, while some of them have multiple parameters that might change the results a lot. Then their performance is evaluated by using simplified dataset to train and test them, to get the efficiency in accuracy and spatial-temporal cost. The dataset used here is generated when first post of every query is extracted. It has 4,682 samples, each sample having 1,768 features. Five-fold cross validation is used to yield results. Finally reference to which the implementation of the estimator can be found is given.

This process is comparable to calibrated classification, where estimators predict true labels by giving scores of probability. Although some of the classifiers only indicate a sample belongs to a class or not, rather than telling how likely the sample belongs to the class. Therefore, the aforementioned criteria are used to compare and choose the best one from eight classifiers, Naive Bayesian Classifier, Logistic Regression Classifier, Linear Discriminant Analysis Classifier, k-Nearest Neighbours Classifier, Decision Tree Classifier, Extra-tree Classifier, Multi-layer Perceptron Classifier, and Support Vector Machine Classifier, listed in Table 3.5.

Table 3.5: estimators and comparison

estimator	**algorithm**	**using scenario**	**parameters**	**accuracy (%)**	**cost**	**available**
Naive Bayesian Classifier	use Bayes' theorem to estimate probability of a class from priori probability of the class and posteriori probability of feature given the class	assume that every pair of features is independent	there is no parameters to tune	1.9	11.5 s	sklearn .naive_ bayes .GaussianNB
Logistic Regression Classifier	use sigmoid function to replace independent variables in linear regression estimator's cost function, and generalise it into a one-vs-all classifier	classification problem in hyperspace	minimizing methods like gradient descent and number of iteration, penalizing methods and weight (to regularise model and prevent overfitting)	3.1	8.9 min	sklearn .linear_ model .LogisticRegression

Linear Discriminant Analysis	use Bayes' theorem to find decision surface, and use Gaussian distribution to model the posteriori probability of features conditioned on class	when linear boundaries can be used to divide the hyper points into classes, for more flexible boundaries, we try its counterpart Quadratic Discriminant Analysis	shrinkage, store covariance, and solver	7.0	2.3 min	sklearn .discriminant_ analysis .LinearDiscriminantAnalysis
k-Nearest Neighbours	use focal point's k nearest neighbours in the training set to predict the class of focal point	it is an instance-based method, and it is fast learner	the number of neighbours, the method to calculate distance	0.2	1.2 s	sklearn .neighbors .KNeighborsClassifier
Decision Tree Classifier	split the tree by one feature to partition data into classes, the path to each leaf is one decision rule for a class	reuse features	criterion to measure the quality of a split, depth of tree	-	-	sklearn .tree .DecisionTreeClassifier
Extra-trees Classifier	Extremely randomized trees method is based on decision tree, this algorithm varies conditions randomly at split, and classifies through branches.	multi-classification problems, generating prediction probability and feature importance	number of estimation trees	8.6	1.8 min	sklearn .ensemble .ExtraTreesClassifier
Multilayer Perceptron Classifier	neural network that use layers of neurons to weave input and output, the activate mapping between layers is logistic regression	complicated relationship between input and output	number of layers and number of neurons in them, methods of learning parameters, learning rate	0.9	2.5 min	sklearn .neural_ network .MLPClassifier
Support Vector Machine	find decision surface by support vectors	high dimensional space	kernel functions for decision function, class weight	1.1	2.4 min	sklearn .svm .SVC

Note: feature sets we extracted from dataset, before we select them by their contribution to the user identification. Efficiency is evaluated in estimation accuracy, and running time, so we use the unit percentage, second, or minute.

From the table it can be indicated that classifiers have different learning and predicting mechanisms, which decide their using scenario. Different parameter settings will also affect the result, so their prediction accuracy scores are compared after the optimal parameters are tuned and gotten. Due to the differences in learning algorithm and number of parameters, their expense of time varies a lot, from 1.2 seconds to 8.9 minutes per fold in five-fold cross validation. K-nearest Neighbours classifier is the most fast classifier among the seven recorded

classifiers, because they predict the class label of a data point by its labelled neighbours. While the slowest model, Logistic Regression classifier, need to tune as many binary classifiers as the number of classes in dataset. As prediction accuracy is placed at a higher position than time consumption, estimators that can give an accurate prediction and the output scores should be calibrated enough for next step is chosen, as plotted in Figure 3.7. All eight classifiers above can be found implemented in python's machine learning library scikit-learn.

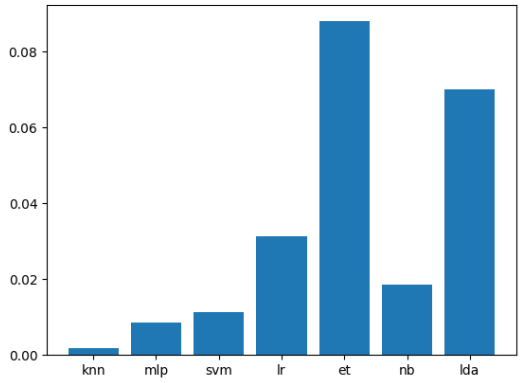


Figure 3.7: prediction accuracy of estimators in their best performance
Note: y-axis is percentage. knn is K Nearest Neighbours, mlp is Multi-layer Perceptron, svm is Support Vector Machine, lr is Logistic Regression, et is Extra-trees, nb is Naive Bayes, and lda is Linear Discriminant Analysis.

The classifier finally being decided on is Extra-trees Classifier, an ensemble estimator of several randomized decision trees, for the classifier satisfies both high hit rate and variant estimating scores. An additional advantage for tree based learners is, that they also calculate feature importance, which makes it also a good choice in feature selection process. Why it can achieve the highest accuracy and how to tune its parameters later will be discussed later.

3.6.3 Selected estimator and its parameters

Extremely Randomized Trees The Extra-tree Classifier, also Extremely Randomized Trees classifier, used here is derived from Decision Tree classifier.

Decision Tree is a non-parametric learner. Researchers compare the segregation of dataset by their class as building a tree. All instances in dataset are started in same class as the root node of the tree. Then, they are split into two classes by setting a threshold for one feature, just as a tree

is split into two branches. To find the threshold and the standard feature, methods are adopted to measure the efficiency of the split. Possible measures are the Gini index and cross-entropy of classes. The threshold that can reduce these index most is considered to be the best. When the leaf node is reached, one decision rule is gotten for a class, that is, the path from root to the leaf node. Note that one class might have more than one decision rule.

If a random subset of all features is selected when the optimal split is chosen, the variance of the model will be reduced as well as raise the bias. Nevertheless, the influence of the latter outweighs the former one. That is called randomized trees or random forest. If threshold is set for the feature randomly on the base of random forest, the variance can be further reduced, and thus model will suit dataset better. That is called extremely randomized trees. By choosing from subsets of features and threshold of features, the classifier ensembles multiple trees into a forest. Ensemble models from basic learners can reduce the variance, although sometimes they will also increase the bias in the meanwhile.

Parameters Parameters that can be tuned are maximum number of features and number of estimators. Experiments tell us, when maximum number of features equals to square root of feature number, the classifier can reach its best performance. As for number of estimators, i.e. the number of trees, the more trees are, the better, yet it will take more time to compute. The accuracy is calculated over the primitive test set, by setting an arithmetic progression for tree amount. As the parameter tuning experiment shows, when there are over a thousand features, three hundred trees can be used to achieve the best performance.

3.6.4 Learning and estimating

The estimator is trained and tested by cross validation, that is to stratify the dataset into five folds, and then pick out four subsets to form the training set and the rest one to be test set. To guarantee each class label in the test set has appeared in training set at least once, it need to be guaranteed that all the questions in each fold have true authors that complete the whole user set. That is one of the quantitative rules, each user has at least five questions. Next, the model is trained with training set and is used to mark the probability scores for the potential authors of each question in the test set. After five rounds of validation, the scores are generated for each

question and user pair. It can also be calculated the five accuracy rates, the mean of which are used to measure the performance of the model.

After estimating scores, methods introduced in Section 3.7 are used to predict true authors and put them into equivalence class.

3.7 Prediction

More than one user is predicted as likely author for a sample question and they are put into equivalence class.

3.7.1 Equivalence class generation

Users are selected into equivalence class by their scores. To make the scores distinguishable between users is the goal of estimation process, which will be discussed in details in the later sections. Assuming that the matrix of scores is given, how to generate equivalence class can also affect prediction results. Here different algorithms are investigated to pick out the high-marked users, by controlling the threshold of score set or the size of equivalence class.

Jump points A jump point in an ordered sequence is defined as the two adjacent elements pair that has largest difference between them. This is used as the threshold to pick out all the users ranked higher than or equal to the jump point. The steps of generating equivalence class by jump points are as follows. First, the users are sorted by their scores in descendent order. Secondly, the difference between each two adjacent scores are calculated. Next step is to select the score, from which the highest difference is subtracted, i.e. the minuend of the highest difference, as jump points. Finally, all the users ranked higher of the jump point are chosen, as well as jump point itself into equivalence class. If there are two or more highest differences, only their highest minuend is picked as jump point, so that the size of equivalence class are be able to remain as small as possible. This choosing process are repeated to generate the equivalence class for every test question. Other than highest difference, other thresholds are also tried like mean, median, mode, and 95th percentile. Take mean for instance, the users are sorted in the same way

as original jump point method, and then calculate average score of the sequence as threshold. In this experiment, jump points with highest difference works best among all threshold strategies to generate a equivalence class with most suitable size.

Highest 90% score Compared with jump point algorithm, this method uses scores directly to select competent users into equivalence class. The distribution of scores is checked first, and it is found that when the percentile is set to be 90%, equivalence classes are relatively compact in their size. Hence, all users between 90% and 100% are selected.

Top 95% users The above two methods decide whether a user belongs to the question's equivalence class or not by comparing user score to a certain threshold. Apart from that, a loose frame can also be set for the size of equivalence class, that is, controlling the number of selected users. As usual, the first thing to do is to rank the users by their possibilities to be the real author. Next step is to select top 95% users. The percentage can change according to the total size of user set. The size can also be loosen by allowing those users closely following the last candidate in the list into equivalence class.

3.7.2 Methods selection

The evaluation of these algorithms requires the consideration of the goal of generating equivalence class. Equivalence classes are generated to allow predictors to choose from a set of candidates instead of one single candidate as the possible author for the query thread. This modification can increase the chance to hit the true author, however, the scope will be large and vague. So it is hoped that the equivalence class will be representative enough so that the users in the class is significantly more likely to be true author than those lying outside the class. This brings about the first criterion, the scores of predicted users are distinguishable enough. The second criterion should restrict the size of equivalence class, the distribution of which should not be too uneven, nor too flat.

3.8 User identifiability measurement

After predicting the users to be the true author for a sample question, user identifiability is calculated to measure the extent to which the user who actually posts the question is identified. Multiple ways are tried to calculate identifiability, because this concept relates to more than one factors. Between true author user and her questions, there is a one-to-many relationship, and it is the same between question and its predicted author users. When a equivalence class with one or more predicted users for a question is generated, two factors are introduced, the **size** of equivalence class and the **rank** of each user in the class according to their scores. If the true author is not included in the equivalence class, then her rank would be positive infinity. In addition, different summarizing measures will also lead to different results. For instance, what can be used are the size itself or a binary indicator to represent whether the equivalence class predicts only one candidate or not. When identifiability of one user is estimated, the value is averaged over all her questions. When the identifiability of the whole user group is concluded, it will be averaged over all users in the group.

Looking back at the example of U_A using CQA in Figure 3.3, identifiability of this user is calculated using her query threads, which are the first and the second query threads. As there are two authors in this example, the maximum of equivalence class size is 2. Suppose for the first question, an equivalence class is generated with U_A and U_D in it, and U_A is ranked higher than U_D. Then the two factors induced from the experiment results of the first question are, 2 for equivalence class size and 1 for true author rank.

3.8.1 Metrics

Metrics are derived from the two basic factors mentioned above, including the size of equivalence class and the rank of intended user in equivalence class. One indicator is used to give 1 if the equivalence class size is one and 0 if not, and other indicator to give 1 if the rank of the user is finite, i.e. the user is in the equivalence class, and 0 if not. Four metrics that comprise of the three factors are presented as below. They are uniqueness, exactness, inclusion and closeness. For any new annotations, Tl is the abbreviation for true label, which is actual user who ask the question.

Uniqueness is the number of *EC* who have only one element. When the question is narrowed down that whether the only prediction is the true label to a polar question, i.e. a yes-no question, to which there is only a positive or negative answer, the concept of uniqueness are brought up. Compared to the work of study[19] where an individual is identified by only one mobile trace of spatial-temporal points, uniqueness is calculated as an indicator when the set of similar traces is sized to 1. Uniqueness can be calculated in the following formulas.

$$F(u) = \frac{1}{\|Q_u\|} \sum^{q \in Q_u} U(q)$$

$$U(q) = \begin{cases} 1 & , \left\|EC_q\right\| = 1 \\ 0 & , \left\|EC_q\right\| \neq 1 \end{cases} \tag{3.1}$$

where u is one of the focal users, Q_u is set of all questions queried by u and q is one of those questions, EC_q is a list of predicted users that are equally possible to be the actual querier of q. $U(\cdot)$ is an indicator function to check whether the size of EC_q is 1. $F(\cdot)$ calculates user uniqueness from all her questions, and its mean over all users is one of the measures of user identifiability.

Exactness is the number of *EC* who have only one element and that element is the very class label of the true user. Exactness requires that the only element in *EC* is exactly the Tl, so it is more restricted than uniqueness.

$$F(u) = \frac{1}{\|Q_u\|} \sum^{q \in Q_u} A(q)$$

$$A(q) = \begin{cases} 1 & , EC_q = u \\ 0 & , EC_q \neq u \end{cases} \tag{3.2}$$

where $A(\cdot)$ is an indicator function to check whether predicted EC_q contains only one label and that label is the Tl of q, and $F(\cdot)$ here produces the exactness of the focal user.

Inclusion is the number of *EC* who contain the true user who actually submits the question. When predictor cannot decide on only one Tl, threshold is loosened to cover more possible Tl.

Predicting more possible users increases the size of EC, as well as the chance to hit on the true user. Inclusion is much looser than uniqueness and exactness.

$$F(u) = \frac{1}{\|Q_u\|} \sum^{q \in Q_u} I(q)$$

$$I(q) = \begin{cases} 1 & , u \in EC_q \\ 0 & , u \notin EC_q \end{cases} \tag{3.3}$$

where $I(\cdot)$ is an indicator function for inclusion, while $F(\cdot)$ here counts its proportion and gets the inclusion for the user u.

Closeness is the reciprocal mean of the sum of true class rank and equivalence class size. It take both size of EC and position of the Tl in the predicted EC into consideration. Because the most confident prediction is the top predicted label in EC, if labels in EC is ranked, then how close the hit label is to the top one is expected. Considering the case when the Tl is not near the first place, but still in the predicted EC no matter how small it is, the size of EC is involved in to show that relatively precise prediction can still be given here. Closeness punishes on both size and rank.

$$F(u) = \frac{1}{\|Q_u\|} \sum^{q \in Q_u} C(q)$$

$$C(q) = \begin{cases} \frac{2}{R(u) + \|EC_q\|} & , u \in EC_q \\ 0 & , u \notin EC_q \end{cases} \tag{3.4}$$

where $R(\cdot)$ is the rank of u in each EC_q, if hit, and $F(\cdot)$ averages it for all questions by the user u to generate closeness.

3.8.2 Testing hypotheses

Higher user identifiability indicates a stronger impact of user's self-disclosure on privacy leakage. It is recommended to look at identifiability over the whole user set to support hypothesis **H1**. However, what is lacked is the baseline value to judge whether the prediction is significant enough to identify users in the dataset. So only the percentage and extent of user being identified are calculated. Using this kind of metrics to test **H2** is much more explicit. The change of total

identifiability can be drawn when the number of posts is controlled to increase by one post at each time. The tendency of the value is examined, and see whether the curve is going up.

3.9 Accuracy evaluation

For a question whose true author is needed to be identified, all the users in its equivalence class are the candidates nominated by the estimator. When the prediction process is evaluated as a whole, from estimating the probability to generating equivalence class, the true author is compared with a bunch of predicted users in the equivalence class.

3.9.1 Equivalence class and user identifiability

When the methodology is introduced, the first to be presented is new approaches for author identification problem, predicting more possible authors for one question into a equivalence class and measure the corresponding performance by four user identification metrics. Smaller equivalence class size leads to a better predicting accuracy, given that all other conditions, like whether the true author is included, remain unchanged. As for user identifiability, uniqueness, exactness, inclusion and closeness, the higher they are, the better the prediction model is.

3.9.2 Accuracy of prediction

All in all, predicting true author for anonymous question is considered as classification problem. Therefore, there are traditional classification metrics for us to choose from. Most of the metrics are borrowed from information retrieval problem. For binary classification, the number of instances as positive or negative is listed according to their actual class or predicted class in confusion matrix (shown in Table 3.6). If a sample falls in the cell of “true positive", it means that the model predicts the sample to be positive, i.e. the sample belongs to the class, and because in reality it belongs to the class, so the predicted positive is true. Samples in “false positive" cell also have similar meaning: the prediction that they belong to the class is positive, but they are actually not, so the prediction is false. The cells “false negative" and “true negative" follow the same interpretation.

Table 3.6: confusion matrix

prediction \ truth	**positive**	**negative**
positive	true positive (TP)	false positive (FP)
negative	false negative (FN)	true negative (TN)

Note: this is confusion matrix for measuring error of supervised classification problem.

From confusion matrix showed in Table 3.6, induct Precision and Recall can be inducted. Also, from Precision and Recall, F-measure can be generated. Once the rank of candidate users in predicted equivalence class has been known, Mean Reciprocal Rank (MRR)[14] can be calculated.

- **Precision** is the proportion of those samples correctly predicted as positive in all of those predicted as positive. It is calculated by $Precision = \frac{TP}{TP+FP}$. The higher it is, the more actually positive samples in the positive prediction.

- **Recall** is, compared with precision, the correct rate of samples correctly predicted as positive. The formula $Recall = \frac{TP}{TP+FN}$ is used to calculate its value. The higher it is, the more positive samples it will cover.

- **F-measure** is generated from precision and recall, by $F\text{-}measure = \frac{Precision \times Recall}{Precision+Recall}$. The statistics of TP, FP and FN are considered this time, and make the measurement more thorough.

- **Mean Reciprocal Rank** is commonly used in recommendation problem, because it take the rank of prediction, i.e. the how confident the predictor is about the predicted class label, into consideration. Its formula is $MRR = \sum \frac{1}{Rank}$.

3.9.3 Tendency of metrics when extent of self-disclosure increases

To answer the research question, and especially test the Hypothesis H2, it needs to be found out the influence of the extent of self-disclosure on the extent of user identifiability. So it will also be examined the tendency of metrics that measure identification when number of posts in query thread increases. In this experiment, one post is put in the dataset at a time, and the identification is evaluated over each dataset. Therefore, it is necessary to observe the tendency of

user identifiability, as well as metrics of prediction accuracy that are discussed in this section. One approach is to calculate derivatives and prove that the relationship is (monotone) increasing. In this study, to avoid further, unnecessary calculation, after variables are plotted, researchers use mainly bare eyes to find out the trend of the curves.

This chapter gives an insight of the methodology system, from which experiment results will be generated. To recap, Stack Exchange is selected at first as research sample for CQA sites, the users and their posts are filtered from “Serverfault" community into experimental data repository. Then, the datasets are represented as input matrix by feature extraction and data pre-processing. Prediction is conducted by first using estimators to score each question-user pair and next selecting highly-scored users into the equivalence class of a certain question. Finally, user identifiability and prediction accuracy are calculated to evaluate the experiment from various aspects. In the next chapter, it will be discussed the results and how they support the research question to be solved.

Chapter 4

Results & Discussion

In the last chapter, process of experiment is presented in details. In this chapter, results are presented and explained from the experiment to find the optimal combination of methods and then test hypothesis H1 and hypothesis H2 respectively.

Firstly, because that not all feature sets are necessary, some of them are assessed and selected into feature matrix, to make sure what being fed into estimators does help to recognise true author for sample questions. The assessment is completed by using single feature set as feature matrix and predicting the tops-scored user to be true author. The accuracy, proportion of successfully assigned questions, serves as the metric. The other method to be assessed and selected is the strategy to choose users into equivalence class. It also use accuracy as metrics. After building the complete model for experiment, the final results of total accuracy and identifiability over all rounds of prediction will be produced. Then, these results are able to be used to test hypothesis **H1**: predictability of the authorship, and hypothesis **H2**: user identifiability has a (positive) correlation with self-disclosure.

After all results being provided, findings about hypotheses testing are discussed in the subsequent sections. The significance of what have been found will be firstly discussed from an overall viewpoint, and then they are used to solve research questions being proposed at the beginning of thesis, and finally give suggestions to solve the privacy concerns out of the support of those findings.

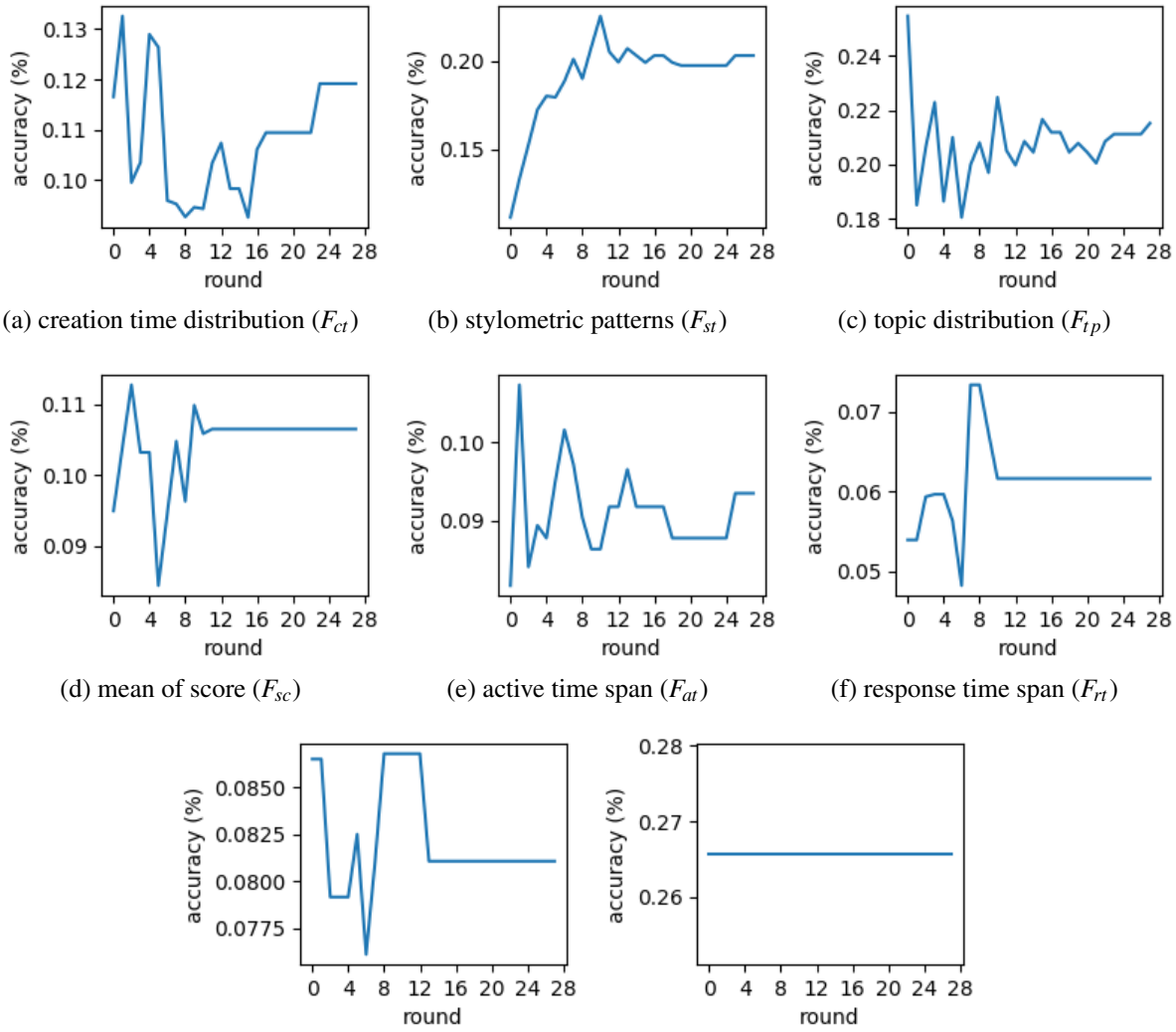
4.1 Methods selection

In this section, intermediate results are checked to select methodologies and finally assemble the framework of experiment. Before prediction with the best-performing model, the accuracy is gauged in the percentage of successful suggestions by the highest-scored user to each question. The combination of feature sets and algorithms to generate equivalence class are decided by this accuracy.

4.1.1 Feature sets selection and combination

Single feature set performance After extracting eight feature sets as described in Section3.4, the feature selection process is simplified as only looking at single feature set accuracy. Firstly, all the feature sets are extracted from the same dataset for every round, and the accuracies to them in each round are plotted in Figure 4.1. The dataset being used here contains 30 users, which means the stochastic probability of a question to be successfully assigned to its true author among these users are 3.33%. Any feature set that gives an accuracy higher than this value could be considered helpful to question author identification. There are at most 28 related posts in each sample question, that is to say there are in total 28 rounds of prediction as one new related post is added for each sample question at one round. Note that number of round starts from 0, so Round 0 is actually the 1st round, i.e. and there is only one related post in each query thread being put in the dataset for the first round, and there are 28 rounds in total as the last round is Round 27.

Here are the observations that can be gotten from the figure. The feature set with highest accuracy is **tag distribution (F_{tg})** in Sub-figure 4.1(h), with a constant value of 26.6%. Tags are an attribute that only belongs to question, i.e. the first post from a query thread to be put into dataset, so tag distribution feature set is only extracted at the first round of prediction. Therefore, the performance of this feature set stays the same during all rounds of prediction. One horizontal line with slope of zero can be observed in the graph. Every question will have one or more tags, and question authors choose them as keywords for the query of questions, so tags are highly correlated to the content of questions they are attached to. The fact that this feature set F_{tg} performs the best in author prediction tells that it contributes most to identification of users through their questions and that the combination of tags is special for each question.



(a) creation time distribution (F_{ct}) (b) stylometric patterns (F_{st}) (c) topic distribution (F_{tp})

(d) mean of score (F_{sc}) (e) active time span (F_{at}) (f) response time span (F_{rt})

(g) # comments (F_{cm}) (h) tag distribution (F_{tg})

Figure 4.1: accuracy of each feature set over round of experiment
Note: precision of using single feature set. Feature sets are F_{ct}, F_{st}, F_{tp}, F_{sc}, F_{at}, F_{rt}, F_{cm}, and F_{tg}.

In Sub-figure 4.1(c), **topic distribution** (F_{tp}) starts at the highest point 25% and then drops down to the lowest point 18% at Round 6, and finally become stable on 22%. At the first round, there is only one question as related post for each sample query thread, and questions are usually completely accounted and with a concise title. It is reasonable that topic distribution can help to recognise authors for questions most efficiently in the first round. As more related posts being added into dataset, answers and comments bring new topics, and thus the success rate of prediction fluctuates in the following several rounds, and then become more and more stable till the final round. Due to the knowledge exchange intention of CQA services, query threads

are mostly topic-concentrated and they serve a good way to profile their authors' interests in study or at work. It thus can be expected how much topic-related attributes contribute to author identification.

The third most important feature set is also textual attributes, which is **stylometric patterns (F_{st})** in Sub-figure 4.1(b) with an accuracy reaching 20% at last half of total rounds. it starts from 10% and peaks at 22% on around 11th round. The calculation of stylometrics is finding writing style habits of users in 10 aspects, from punctuation choosing to sentence length preference. The more posts are included in the dataset, the more representative the statistic patterns are. This could explain why the graph of this feature set shows a steady and huge growth at the first ten rounds (by at most 12%), compared to those of the rest of feature sets. Also, the convergence of it can be explained by that there are less new related posts to be included into dataset at later rounds. Writing style, as a kind of non-private (not releasing privacy) but somehow characteristic qualities, make users more identifiable by the content they generated in Social Media.

Creation time distribution (F_{ct}) is the first non-textual feature set that has high accuracy. The graph of it in Sub-figure 4.1(a) is a curve concaving upwards, in spite of the fluctuation at the beginning. It reaches its peak, 13% proportion of correct prediction, at Round 2, falls by 4% until Round 8, and then climbs up to 12% for the rest of rounds. As each round gives only one creation time slot extracted from the newly-added related post to update the distribution, the undulation, or the large wave, seems inevitable. The accuracy of this feature set shows, that what time in the day does the user usually ask questions does differ among individuals.

The rest of temporal feature sets are less indicative and yet still offer accuracy higher than 3.33%, namely **active time span (F_{at})** and **response time span (F_{rt})**.

The graph of active time span (F_{at}) in Sub-figure 4.1(e) slightly fluctuate around 9%. The time span from when user ask the question to when user finally response to the query thread varies a lot. It can be less than one hour, or as long as several years. However, it is also unpredictable among the questions by the same user. Therefore, it is weak to serve as prediction feature of question author.

The curve in Sub-figure 4.1(f) is the accuracy of **response time span (F_{rt})**, which jumps from 5% to 7% and finally rests at 6%. How frequent does a user check their query thread can be represented by the time span between the creation time of the last post and the new post. It is one of the using habit. But just like active time span F_{at}, sometimes it can be unpredictable, and even varies in the same query thread between different posts. It is the least significant feature set for prediction.

Finally, the two feature sets that represent posts quality, **mean of score (F_{sc}) and # comments (F_{cm})**, turns out to work as expected as well.

Mean of score (F_{sc}) is shown in Sub-figure 4.1(d). Its graph waves in a small range of 1%, which is second to the flat line of tag distribution (F_{tg}). The score is marked by query visitors, and it can indicate the quality of the post, well-received or doubtful. If users keep their quality consistent, this could be an indicator for authorship prediction too.

Number of comments (F_{cm}), plotted in Sub-figure 4.1(g), is around 8%. The feature set counts how many times do users reply for the questions or answers. This shows the popularity of the posts, and can be used for identification.

All the curves are over 3.33%, the lowest of which is that of response time span F_{rt} with 5% as its minimum accuracy. They are all in some degree indicative to user prediction. Most of graphs end their dramatic change around Round 8. In the meanwhile, demographics of the dataset show that all of questions have at least 8 related posts in their query threads, and that most of questions have exacted 8 related posts. In the later rounds, there are less and less new related posts being added into dataset, and this might count for mild changes. Another reason for the convergence is the accumulation of statistic patterns as more related posts are added to dataset.

Optimal combination of feature sets It is noticed that there are four feature sets having an accuracy steady and higher than 10%, taking up half of the total feature sets, so these four feature sets are combined, namely creation time distribution (F_{ct}), stylometric patterns (F_{st}), topic distribution (F_{tp}), and tag distribution (F_{tg}), and check their accuracy in each round of prediction, and plot them in Figure 4.2.

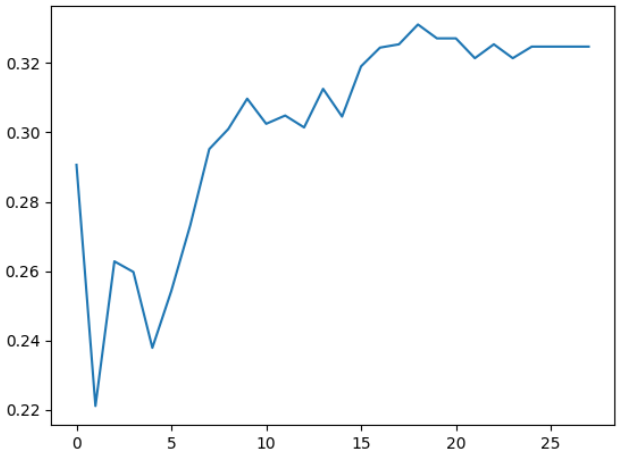
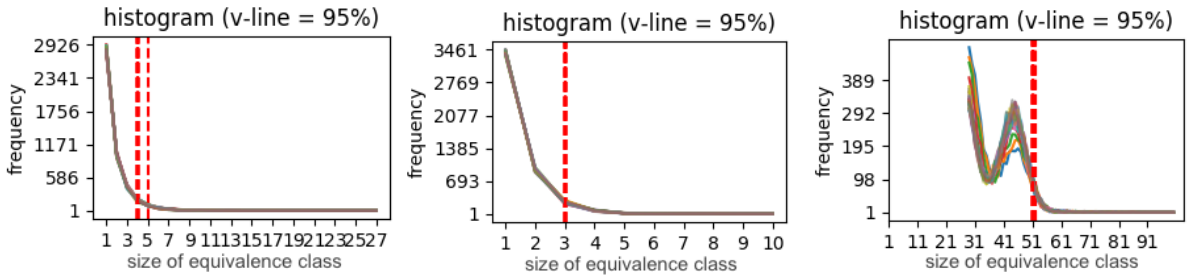


Figure 4.2: accuracy of feature set combination over round of experiment
Note: the diagram shows accuracy of using multiple feature sets. The selected feature sets are F_{ct}, F_{st}, F_{tp} and F_{tg}.

4.1.2 Equivalence class methods selection

Next, algorithms that are used to choose users into equivalence class are also decided by their accuracy performance. For each sample question, users are ordered by their scores to the question from high to low, and then those ranking higher than threshold are predicted as potential authors and put into the equivalence class. Larger size of equivalence class can raise the chance including the true author and thus increase the true positive cases (predict the actual author to be author), but also the false positive (predict non-author users to be author). To make the prediction hit the real author yet nominate a small number of users, one optimal algorithm is chosen that can generate equivalence class with appropriate size. Figure 4.3 shows the distribution of equivalence class size using the three algorithms, testing over all the subsets of the experimental dataset.

Methods selection There are 587 users and 4682 questions in the dataset being used in this section, which means the maximum size of equivalence class (x-axis) can be 587 and the highest frequency (y-axis) is 4682. Sub-figure 4.3(a) the distribution using jump points tells that more equivalence classes have a small size, and that although the maximum size is 27, 95% of total equivalence classes have a size no more than 5. There are 2962 equivalence classes have only one predicted candidate for true author, while the number of equivalence classes whose size is more than 9 is near to 1. If top 10% scored users are selected into equivalence class, a size distribution



(a) jump points (b) 90% score (c) 5% number

Figure 4.3: distribution of equivalence class size using three generating algorithms (frequency vs. equivalence class size)
Note: a red dashed vertical line indicate where the 95% of samples' equivalence class size is in one subset of experimental dataset.

will be gotten as shown in Sub-figure 4.3(b). All equivalence classes have no more than 10 users and 95% of them have merely 3 users or less. Those whose size is only 1 user take up more than three quarters of equivalence classes. Sub-figure 4.3(c) shows that picking users whose score exceed 5% of users score, there will be 95% of all 4682 questions whose equivalence class size is between 21 and 51. There could be an equivalence class of a question with over a hundred of users. Considering total number of possible users are 587, the equivalence class predicts one possible label out of every five class label.

In conclusion, too many equivalence classes have only one user as their prediction will not show the benefits of using equivalence class to enlarge the prediction scope. However, too large the size of equivalence class is will also "overfit" the prediction model. Algorithm using 95% score has too many one-sized equivalence classes while algorithm using 5% number brings large equivalence classes. Therefore, algorithm is chosen using jump points method with mean as threshold.

Rationality of equivalence class Next, the usage of equivalence class will be assessed by checking on two questions' equivalence class, and find correlations of users in the same equivalence class by inspecting their own query threads. The two questions are the 4th question and the 68th question in sample set. Their equivalence classes are of the same size, each having three users, but question no.4 have no true author in its equivalence class while question no.68 successfully predict the true author. That makes these two questions comparable for finding why

users are in the same equivalence class or why not.

The 4th question After prediction, an equivalence class containing three recommended users is generated for the question. However, among those recommendations, there is no hit the true author. The question is about anti-virus in small workgroup. It has 4 comments posted by the question author and all of them are composed within 2 days. One question and four comments are the related posts where feature matrix is extracted for prediction. When looking at the true author missing in equivalence class, they also start 4 query threads other than the one including question no.4. The rest 4 query threads are about VMWare, IIS and Google Webmaster Tools, each having one question, and four or five comments. The three recommended users have their own query threads.

If the three users are checked in equivalence class, the user winning highest score, 0.029, authors nine query threads, all having less than six comments. They are about error in OpenBSD, Windows Vista and VMWare, Exchange mailbox and so on. The time span spent on each question by the author varies from several hours to several months. The user that is given second highest score, 0.026, has 11 questions each with less than five comments. They are about domain WSUS, samba, SCSI, Hyper-V error, and other topics under server virtualization. One of the questions is given all the comments in two days, while one of them can be as long as two years. The last user in equivalence class wins 0.024 point, with 10 questions, and most of them have only four comments. They are about CISCO, Microsoft Forefront, RDP certificate, and process by SID. The total active time of each query thread can be less than 1 hour, or longer than 2 months.

From the comparison of the true author and predicted false author, there are some observations found. Firstly, although all the questions are from server fault community, they still have closer topic distribution. This can also be shown by tags attached to their questions. Secondly, the dataset filter rules restrict the number of related posts under the same query thread to be more than five. So the minimum combination of question and comment under the same query thread would be 1 question and 4 comments. These predicted users all give around 4 comments to most of their questions or answers to the questions.

The 68th question The second question to be looked at also having three users in the equivalence class, among which the highest-scored user, scored 0.036, fits the true author of the question. The focal question is about comparing virtualization machine tools (Hyper-V and Virtual PC), and it has 15 comments composed within 1 day as its related posts. The hit user has 24 more questions except the focal question. They have around 6 and 7 comments within several days and they cover the topics of IIS redirection and question routing, SQLServer, and website mapping. Another user that is given the same score as the hit user has 7 questions, each with around six comments. They are mostly about Hyper-V, or about SQL, Amazon EC, account migration, and WAN. The overall active time of this query thread is up to 10 days. The last user is scored 0.031 and has 8 questions. Some of them have ten or more comments, posted within several days, and most are about Windows Server, CNAME and DNS, firewall, group policy and Virtualization with Hyper-V.

Comparing the 4th sample question with the 68th by their selected users and their query threads, users in the equivalence class of 4th question has longer active time and less related posts, posts in different equivalence classes are about VMWare or Hyper-V. Therefore, it can be concluded that users in the same equivalence class have stronger correlation with each other in some aspects. In that case, the usage of equivalence class is helpful to group users by their relevance to the true user of a certain question. By doing so, more accurate true author suggestions can be recommended and factors causing identification of the author can be found more efficiently.

4.1.3 Workflow formation

By this point all the crucial parts of the experiment have been determined, accuracies are generated using the selected dataset, extract the four feature sets and combine them into one input matrix in each round of prediction, and use method of jump point to generate equivalence class, and eventually predict true author for the given anonymous question.

4.2 Hypotheses Testing

4.2.1 Accuracy of Author Detection for Question

To answer whether self-disclosure information can identify content author in CQA environment, accuracy, i.e. the rate of correctly predict true authors for questions, is calculated over various datasets and by several classifiers. The calculation of accuracy in this case is to work out the rate when a equivalence class given by predictor contains the true user of a question. For datasets, four dataset are taken to cover the factors of a) number of communities being used to generate the dataset, b) the size of dataset. Datasets are either generated from the intermediate “serverfault" community or other nine smaller communities ("3dprinting", "ai", "beer", "bricks", "coffee", "computergraphics", "crafts", "ebooks" and "esperanto"). There are also different quantitative criteria for datasets on the number of related posts of a question (in a query thread), which decides the number of competent users in the datasets. This action is intended to show whether the diversity of topic domains and the quantitative assurance will affect accuracy performance. The four datasets used here are:

- Nine community large (NCL): from 9 small communities this dataset is generated, where users have at least 5 questions and those questions have at least 4 related posts. It has 116 users and 705 questions, and the maximum round of prediction (i.e. the number of related posts under one query thread) is 8.

- Nine community small (NCS): it is generated under the same condition as dataset NCL, but with only 10 qualified users, and therefore 64 questions, the maximum round of prediction is also 8.

- One community large (OCL): from 1 intermediate community this dataset is generated, where users have at least 5 questions and those questions have at least 5 related posts. It has 587 users and 4685 questions, the maximum round of prediction is 35.

- One community small (OCS): it is generated under the same condition as dataset OCL. It has users have at least 6 questions and those questions have at least 8 related posts. It has 30 users and 224 questions, the maximum round of prediction is 28.

Their accuracy performance over various estimating models is show in Table 4.1.

Table 4.1: Accuracy of six estimating models over four datasets

dataset	lda	**et**	svm	mlp	knn	nb
NCL	0.01	**0.46**	0.1	0.16	0.11	0.39
NCS	0.11	**0.67**	0.34	0.41	0.42	0.55
OCL	0.02	**0.17**	0.03	0.16	0.05	0.1
OCS	0.34	***0.76***	0.31	0.04	0.09	0.3

Note the abbreviated classifier names: lda for Linear Discriminant Analysis, et for Extreme Random Trees, svm for Support Vector Machine, mlp for Multi-layer Perceptron, knn for k-Nearest Neighbours, and nb for Naive Bayesian classifier.

From the table it can be inferred that Extremely Random Trees classifier, the estimator selected in the last chapter, always keeps best performance among the models, which is in accordance of what is in Figure 3.7.

It can also be observed that the OCS dataset, having single community, high quality and small size, gives the top accuracy. Among all four datasets, OCL contains the largest sample, and in later part of experiment it will be found that OCL has an overall smaller equivalence class size than OCS. Considering that dataset affects little on the performance of the chosen classifier, i.e. Extremely Random Trees, dataset OCL is still chosen to complete this study.

Take a look at the accuracy given by using OCS dataset and Extremely Random Threes classifier, 76% of total 224 questions are assigned with correct authors. The equivalence class size is restricted to 1 here, that means 1 user can be successfully identified out of 30 users by 5 or more query threads under each of their names.

In conclusion, optimal combination of feature sets, i.e. creation time distribution (F_{ct}), stylometric patterns (F_{st}), topic distribution (F_{tp}), and tag distribution (F_{tg}), extracted from refined dataset performs well in identifying authors for anonymous questions, through the tuned ensemble tree classifiers. This is supportive for hypothesis H1, which means self-disclosure information can identify content author.

4.2.2 Tendency Verses Number of Round

Tendency in this section means the dynamic relationship between two factors, the synonyms of it being trend of decreasing or increasing, gradient of a curve or slope of a line if the graphs of factors are plotted. Adopting all the methods selected in method selection Section 4.1, and using the dataset and estimating model chosen in accuracy Section 4.2.1, the identifiability evaluation results are now generated for each round of prediction, so that the dynamic relationship between identifiability and self-disclosure can b answered.

Tendency of equivalence class size With other conditions being the same, larger size of equivalence class often results in lower predicting accuracy. When the distribution of equivalence class size is plotted, we expect the size of equivalence class to decrease when number of posts goes up, according to hypothesis **H2**. The shape of distribution (uniform, skewed, bell-shaped, or U-shaped) can also help to evaluate the equivalence class generating algorithms, given that size of equivalence class can serve as an indicator of prediction performance and that the optimal size has been decided in the previous experimental process.

For instance, if the graph is uniform, i.e. a horizontal line, it means the size of equivalence class will not change as the number of round increases, and thus how many related posts a user discloses for a question will not influence the identifiability. However, if there is a curve skewed to the right in the graph, it means before a certain limit, the more related posts are disclosed, the more users are predicted into the equivalence class, and therefore, the more possible those predictions hit the true author. Moreover, if the curve is skewed to the left, there would be an opposite explanation. A bell-shaped curve indicates that the intermediate number of related posts disclosed in the dataset helps to predict more users as elements of equivalence class, and the earlier or later rounds generate smaller equivalence class. If the assess of equivalence class size shows the larger the better, then the number of round enhances the prediction first and suppress it later. A U-shaped curve, or a well curve is the inversion of bell-shaped curve, so the situation of it would be opposite. Therefore, one of the curves is plotted, which is the central tendency of the size when round, i.e. the number of posts that is included in the experiment dataset from the same query thread, goes up.

The total number of user is 587 and there will always be predicted users in one equivalence

class, so the size of equivalence class is between 1 and 537. Graph in Sub-figure 4.3(a) shows that he distribution of equivalence class size is extremely skewed on the small end and that the vertical lines of 95% number of equivalence classes can be drawn left to the point when equivalence class size equals to 5, which means most of equivalence classes have a size smaller than 5 predicted users. The mean of equivalence class size is plotted in Figure 4.5, where horizontal axis is for number of round (starting from zero) and the vertical axis is for mean of equivalence class size. A slightly declining curve can be found, and it can be inferred that the average equivalence class size is between 1.72 and 1.80. That means that average questions will have one or two users predicted as their true author, which is sufficient to recognise the right user from total 30 users. It also means that as prediction round goes up, questions will have smaller equivalence classes.

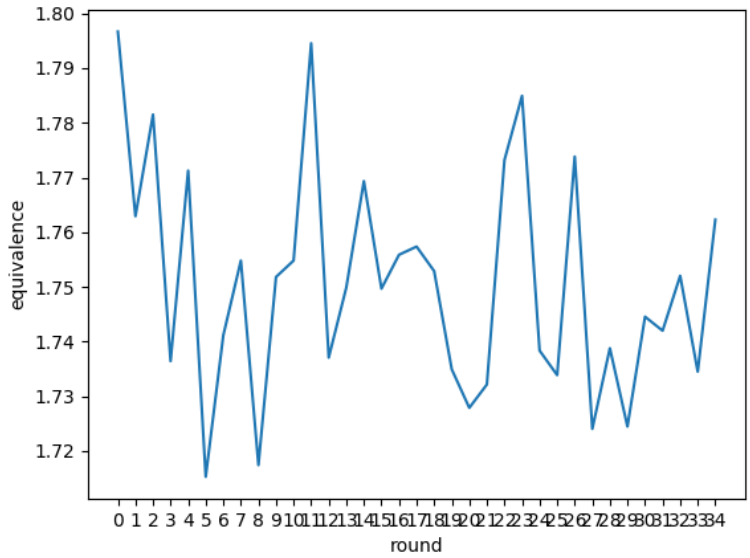


Figure 4.4: tendency of equivalence class size along number of round (mean of size vs. round)
Note: We present size of equivalence class for all samples from one round in six different ways. This figure shows the trend of the most expected one among all the estimator-equalizer combinations, a.k.a. probability estimator - equivalence class selector pair.

Tendency of identifiability Identifiability assesses the extent of users being correctly identified as the author of their questions, so the expecting tendency of it is to go up along the number of round. Figure 4.5 shows the graph of four indicators of identifiability, namely uniqueness, exactness, inclusion and closeness (introduced in Section 3.8 of last chapter). From uniqueness and inclusion in the graph it can be inferred that the majority of questions only have one user in

their equivalence classes, but less than 10% of questions include true author in their equivalence class. For the rest two identifiability indicators, the exactness and closeness, their insignificance shows Those equivalence classes sized one can barely hit the true author.

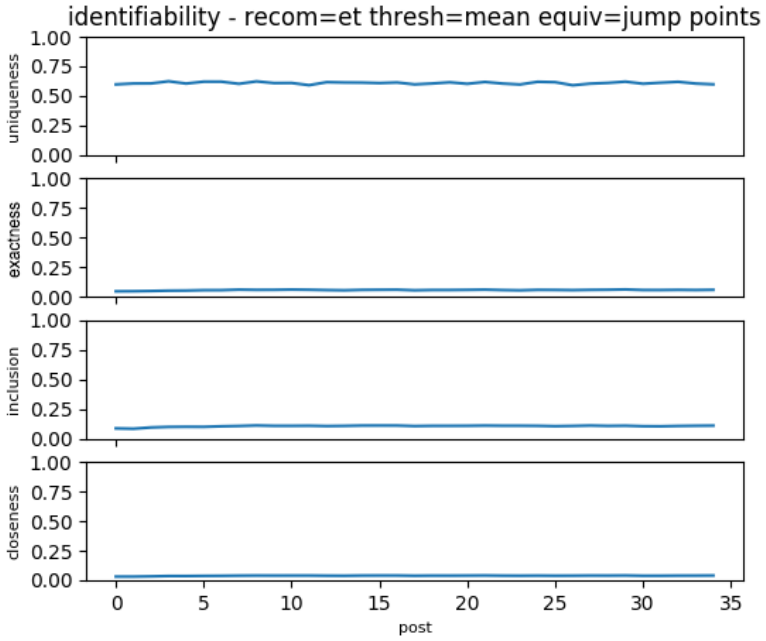


Figure 4.5: tendency of identifiability along number of round (identifiability vs. round)
Note: This figure shows the most expected trend (up-going) among all the estimator and equivalence class generating algorithm combinations.

Tendency of traditional IR metrics When evaluation strategies is introduced in Methodology Chapter 3, four metrics are adopted from Information Retrieval and Recommending Systems, including Precision, Recall rate, F-measure score and MRR. As the higher they are the better the prediction is, four up-climbing curves along the round of prediction are expected here. Now they are calculated by round and plotted in Figure 4.6. All the four metrics stay below 20%, however, they slightly go up at the beginning and remain flat in the rest of rounds. Considering the need to choose from 587 users to predict one true author for each question, and the average number of user chosen is 1.6-1.8, these metrics are in a sense supportive for the possibility of using what is disclosed in user generated content to identify content authors. However, it is hard to find evidence for the relationships between amount of disclosed content and extent of identification.

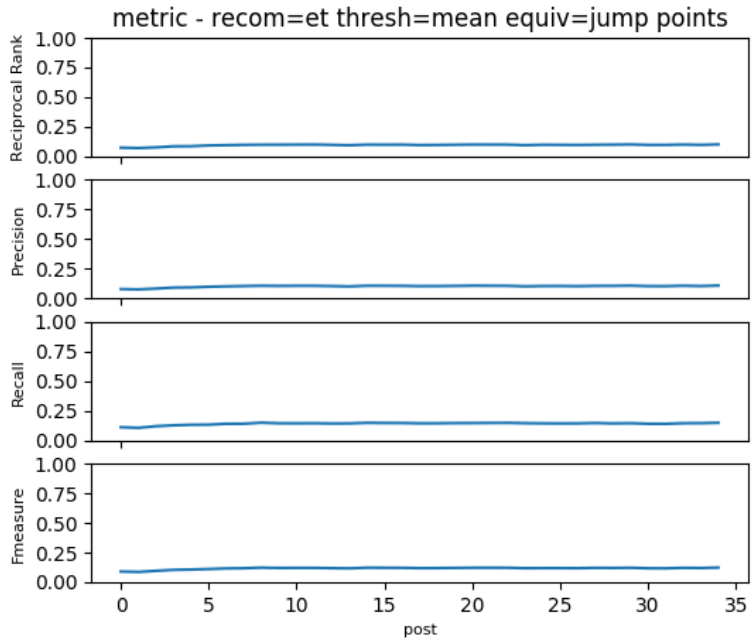


Figure 4.6: tendency of metrics along number of round (metrics vs. round)
Note: The ideal tendency for them is slightly going up.

The figures in this section are either too flat or swaying too unpredictable around a certain baseline value. Therefore they are not supportive enough for us to find out the relationships between degree of disclosure by quantity and that of identifiability.

4.2.3 Tendency Refinement by Slope

Attempt of figuring out whether self-disclosure and identifiability is still half way, for that the degree of self-disclosure have been quantitatively represented by number of related posts for the question to be tested, and the degree of identifiability by metrics calculated from equivalence class size and hit, but the expected evidence still cannot be found to support the existence of the relationships. When tendencies of the curves are hard to recognise by eyes, the value of their slope can be calculated. The linear regression is used to find the slope for mean of equivalence class size, four identifiability metrics, and four conventional IR metrics (take six decimal places).

- Slope of the curve of equivalence class: mean of size = -0.000512
- Slope of the curve of identifiability: Uniqueness = -0.000078, Inclusion = 0.000356,

Closeness = 0.000142, Exactness = 0.000196

- Slope of the curve of IR and RS metrics: Precision = 0.000473, Recall = 0.000490, F-measure = 0.000602, MRR = 0.000543

Noting that all of them have 4 as their order of magnitude, excepting Uniqueness smaller than this scale (its order of magnitude is 5), they are inspected in the scale of 1.0×10^{-4} and ignore Uniqueness. The first observation is, size of equivalence class has a negative slope, which means it is negatively related to the number of related posts with their authors true known. The second is that all of the rest are coefficients of identifiability metrics or IR metrics, and their positiveness indicates that the accuracy of author identification becomes higher as the users disclose more about themselves by exposing more posts (either answers or comments) under the questions they propose.

Although the above two observation seem in accordance of hypothesis **H2**, they are negligible considering their scale, 1.0×10^{-4}. Therefore, it can be concluded that the evidences that can support for **H2** are not significant. In other words, the extent of user self-disclosure measured by number of posts related to users cannot determine the difficulty of user being identified.

In the following sections, the results will still be discussed but from an overall prospect, and hence the research question will be solved.

4.3 Significance of Results

In Social Media services like Community Question Answering sites, users disclose themselves in exchange of emotion and knowledge support from others. The high rate of questions being given with true author indicates that when users interact with other users by posting in online community, even they don't reveal their personal information, or uncover their online or real-life identity, there are still chances for them to be related to anonymous posts that are actually composed by them.

The clues of this kind of disclosure can be figured out by generating different feature sets from various attributes user content and them comparing impact of them in author identification. It is found that temporal using patterns like does a user usually ask or answer a question in the morning, afternoon, evening or night; or like how long does it usually takes for a user to response to other users, are to some extent different from person to person but are not unique to users. Textual patterns, for example the writing style, the common covered topics, or the tags that serve to specify topic domains of a question, contributes more to the prediction of true author. Later in this chapter, solutions aiming at these features will be discussed.

However, according to the testing process for hypothesis **H2**, it still cannot be told in which direction user self-disclosure affects identifiability. When the degree of self-disclosure is represented by the number of related posts that user releases under the query threads to their questions, the factoid (simulated) accumulation of number of related posts in dataset does not significantly raise the extent of identifiability of the user.

This might due to the mistake that the number of related posts should not be used as proxy for extent of disclosure at first place. Indeed, the more user release related posts under a query thread, the more they are prone to disclose about themselves. However, not all feature statistic can be achieved by related posts accumulation. For example, temporal using patterns can be summarised by calculating time span mean or time slot distribution from posts. The extracted feature set will be more representative as more posts are added into dataset. However, if the semantic patterns are to be calculated, for instance what kinds of answer would the query owner accept or doubt and how they do that, it may be necessary to consider all posts and build the structure of the query thread first.

Another possible reason is the limited consideration of types of feature sets. Before extraction of feature sets, prospective feature sets are enumerated by common sense or experience. For instance, topic distribution is a popular way to analyse textual content, topic distribution is extracted from related posts of each query thread, and their textual attributes are concatenated like title, body and content. Another example is that all three temporal feature sets are extracted using time stamp attributes, because these attributes are obviously writing as post properties in data dump files.

Nevertheless, more implicit features might be missed out if the users and their content are inspected further, and one of the obstacles for considering them is the complexity of extracting them from dataset. Take comments by users under their query threads for an example, these are all potential indicators, to name but a few: 1) how frequently a user responds to others' answers by comments, 2) the habit of appreciating for help or pointing out the flaws of the given suggestions when user comments on answers.

The reflection on potential improvement in feature sets to make the experiment capable of **H2** testing is in next chapter.

4.4 Privacy Protection Inspirations

The research question is proposed in this thesis in the concern of privacy risks in social media by user identification. In the following part, several strategies to alleviate this concern are suggested from what have been found. For feature sets being detected as causes to the risk of user being identified by online content they generated, the study will go on to tackle them individually in the removal of personal attributes. Four feature sets are selected by from eight due to their accuracy in the results chapter.

The outstanding of tag distribution (F_{tg}), having an accuracy of 25%, suggests that although all the questions are from the same community, which means they are all under the topic domain of "server fault", they still can be uniquely marked by tags attached to them. Tags are keywords for the topics the question covers, so they help a lot for visitors to search for questions by keywords. The frequency distribution of tags also tells the interest on study or work of the question querier. Thus, the personal sense of it is inevitable in some ways. However, one can reduce this uniqueness by avoiding compounding new words or phrases and use general, separated terms.

Topic distribution (F_{tp}), are also important attributes to identify users, the accuracy of which reaches 21%. The frequency of topics covered in whatever posts, questions, answers or comments, also suggest user interests as tags do. For instance, if a user starts to use virtualization in their small company, there might be a bunch of questions about comparison between virtualization

techniques first, and questions on configuration of hyper-visor later. Unlike tags, it is hard to blur topics by using alternative expressions. But they can diverse topics by asking more questions in other domains. Since this action takes extra labour, especially when users are not interested in such questions, in latter section it will be discussed the employment of obfuscation tools that help to submit questions by other users in the same equivalence class.

The last textual feature set is stylometric patterns (F_{st}) , which successfully predict authors for 20% questions. Habits on writing differs from person to person, and are usually stable for individuals. The influence of writing style on the conveyance of objective information is softer than other textual attributes like topic distribution and tag distribution. For instance, no meaning would be greatly compromised if the author break all long sentences in their questions into shorter ones. Therefore, the suggestion for keeping identity-unknown in social media is to re-edit the post before release, and avoid the habitual punctuation, wording, or structure patterns.

The most important feature set among temporal attributes is creation time distribution (F_{ct}), with the highest prediction success rate of 11%. Creation time might tell that the user often ask questions on workstation building at work and ask about parenting at home. Users can be recognised by their asking, answering or commenting routines. The solution is to break such habits intentionally.

It will be briefly discussed in the section of future work of the next chapter how the removal of these factors is simulated.

Chapter 5

Conclusions & Future Work

The overarching goal of the work is to help user protect their online privacy by suggesting users on which activity patterns on CQA sites will make them identifiable, by showing the predictability of feature sets and the trend of the correlation between user's self-disclosure and their identifiability. In the final chapter of the thesis, it will be concluded what have been discussed in the thesis so far, and what modifications could be made in the future.

5.1 Summary

In this thesis, self-disclosure behaviours and privacy risks are discussed in the usage of social media, like social networking sites and community question answering services. From which we proposed the research question of protecting online privacy by instructing self-disclosure in social media.

The thesis also reviewed the literature and find measurement methods of social benefits and privacy risks in various criteria. What have been decided are to inspect the experimental scenario to select personal-specific attributes, and thus extracted feature sets from these attributes. To define own metrics of user identifiability is chosen after works have been researched that regarding user identifiability as one way of privacy risk quantification.

In order to widen the scale of prediction, equivalence class is defined and three different algorithms are provided to generate equivalence class, and the optimal one is picked by their

pre-assessed predicting performance. After all details of the experiment workflow determined, results are produced to pave the way for analysis and hypotheses testing.

Through the study it have been observed that to get more social benefits, users are tend to expose more about their personal information, however, this can results in higher risks of being identified. Approaches are also proposed in the thesis to show that the degree of users' self-disclosure and identification behaviours in CQA can be represented, and that feature sets about the textual using habits of users can best summarise user self-disclosure. The experiment that repeats multiple rounds shows that self-disclosure helps to identify the users by the questions they ask, however, evidences that support the positive relationship between extend of self-disclosure and identifiability are rarely observed.

5.2 Future work

Although what is discussed in the thesis can serve as a first step to investigate the relationship between self-disclosure and user identifiability in social media, and finally propose strategies for online privacy protection, there are still space to improve in this study, due to the limitation of reached resources, immature experiment design, or the thoughtlessness of testing and solving process. Apart from enhancing the existing study, further steps could also be done to explore the solutions of online privacy issue according to what have been found here.

5.2.1 Refining Factors Representation

Take a look at two factors aimed to find relationship, self-disclosure and user identifiability, the insignificant evidence for hypothesis testing might be due to the way the two variables are represented. In this study, self-disclosure is represented with feature sets extracted from personal attributes like temporal CQA services using routine, posting habits, content quality assessed by audience, and so forth. Some of these features need statistical calculation, but most of them are explicit. The next step is to try more implicit representation, looking at how question asker response to different answers. The extent of self-disclosure can also be more than number of posts that the features are extracted from, because sometimes when more posts are used

the population is approximated more closely, instead of the disclosure degree being increased directly. Therefore, finding a representative featuring approach should be the first problem when the work is continued.

5.2.2 Improving Experiment Design

In the experiment designing process, the strategies to choose best dataset and best combination of feature sets are simple and brutal. For dataset, sample data are always necessary during the building of experiment, for parameters need to be tuned in several processes. The ideal way is to first create a small and clean dataset from data dump and use it as sample to select the phase-wise options like estimators and equivalence class generating algorithms.

In the section of feature selection 3.4.2, only the prediction success rate of individual feature set is compared and then the top four are picked as feature combination, which only considers the local optimum. The most thorough approach is to all combination of feature sets, but it will be effort-taking since to traverse all possible combination of eight feature sets in need to try the prediction $C_8^1 + C_8^2 + ... + C_8^8 = 324$ times. However, more simplified and time-saving alternatives do exist. For instance, it could start from input matrix only contain one feature set, and choose the best performing one as current feature set combination. Next, try each feature set from not-chosen feature sets with current feature combination, and choose the optimal new combination. Repeat the addition and prediction process until reach the desired result, then the maximum total prediction experiment number would be only $1 + 2 + ... + 8 = 36$.

5.2.3 Exploring Obfuscation Solutions

In the future, solutions to privacy issue on SNSs should also be considered after the factors of risky self-disclosure are detected. Other than removing attributes related to significant feature sets mentioned in Section 4.4, obfuscation is also one alternative, which has been introduced briefly in the chapter of related work 2.4.2. Obfuscating user's social media using patterns here is to make users generated data less person-specific, in order to reduce the chance of private information like real name and profile information, without doing harm to the service provided to users.

Group identity in the four privacy protection level proposed by Shen et al. [50] can be considered as a way to realised k-anonymity. Privacy protection levels can also be applied in SNS user identification, for that to profile users their generated content is also required. If equivalence class is compared as a group of user, authorship detection as user profiling, then using equivalence class to confuse users' identity can be regarded as group identity level privacy protection issue. This is also an implementation of k-anonymity[53], the method introduced by Sweeney in 2002 to prevent structured personal data released by public data holder from telling at least $k-1$ data subjects apart.

Considering that users in the same equivalence class is obscured to be true question author for predictor, this similarity can be used to share the question authorship between its true author and other users in the same equivalence class. The significance of contextual attributes has been discussed, therefore it is also possible to make user identity more ambiguous by mixing up their content. By combining those users' similar question, the answer and comment flow are concentrated in the same question, and thus make the original questions' author less distinguishable.

The importance of choosing the proper equivalence class size is brought up here again, because there should be enough users in the same equivalence class to mask the true user, while too many users to look at will cost running resources, and also make the shared content less specific. By doing so, the utility of the query thread, i.e. the effectiveness of problem solving process should not be compromised too heavily. Therefore, questions looking for advices catering personal preference are not recommended to use this approach.

References

[1] Principles for Providing and Using Personal Information, 1933. Online at https://aspe.hhs.gov/report/options-promoting-privacy-national-information-infrastructure/2-iitf-privacy-principles.

[2] Data Protection Act 1998 (DPA), 1998. Online at http://www.legislation.gov.uk/ukpga/1998/29/contents.

[3] A. Abbasi and H. Chen. Writeprints: A Stylometric Approach to Identity-Level Identification and Similarity Detection in Cyberspace. *ACM Transactions on Information Systems*, 26(2):1–29, mar 2008. doi:10.1145/1344411.1344413.

[4] R. Barakat. *Automated framework to improve users' awareness and categorize friends on online social networks*. PhD thesis, North Dakota State University, 2015. Online at https://search.proquest.com/docview/1746683609?accountid=8312.

[5] R. Barakat and K. Magel. Automated Framework to Improve Users ' Awareness on Online Social Networks. pages 428–433, 2016.

[6] R. J. Bayardo and R. Agrawal. Data privacy through optimal k-anonymization. *Proceedings - International Conference on Data Engineering*, (Icde):217–228, 2005. doi:10.1109/ICDE.2005.42.

[7] J. L. Becker. *Measuring privacy risk in online social networks*. PhD thesis, University of California, Davis, 2009. Online at https://search.proquest.com/docview/304853196?accountid=8312.

[8] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3(4-5):993–1022, 2003. doi:10.1162/jmlr.2003.3.4-5.993.

[9] d. m. Boyd and N. B. Ellison. Social Network Sites: Definition, History, and Scholarship. *Journal of Computer-Mediated Communication*, 13(1):210–230, oct 2007. doi:10.1111/j.1083-6101.2007.00393.x.

[10] D. O. Braithwaite, V. R. Waldron, and J. Finn. Communication of Social Support in Computer- Mediated Groups for People With Disabilities. *Health communication*, 11(2):97–121, 1999. doi:10.1207/s15327027hc1102.

[11] L. Brandimarte, A. Acquisti, and G. Loewenstein. Misplaced Confidences : Privacy and the Control Paradox. *Social Psychological and Personality Science*, 4(3):340–347, 2013. doi:10.1177/1948550612455931.

[12] F. Brunton and H. Nissenbaum. International Workshop on Obfuscation: Science, Technology, and Theory. Technical report, New York University, 2017. doi:10.1029/EO063i039p00797.

[13] J. Chen. *MODELING PROFILE-ATTRIBUTE DISCLOSURE IN ONLINE SOCIAL NETWORKS FROM A GAME THEORETIC PERSPECTIVE*. PhD thesis, COLLEGE OF ENGINEERING AND SCIENCE LOUISIANA TECH UNIVERSITY, 2014.

[14] Z. Chen, C. Zhang, Z. Zhao, and D. Cai. Question Retrieval for Community-based Question Answering via Heterogeneous Network Integration Learning. 2016. doi:10.475/123.

[15] S. Cobb. Social support as a moderator of life stress. *Psychosomatic medicine*, 38(5):300–314, 1976. doi:10.1097/00006842-197609000-00003.

[16] S. Cohen and T. A. Wills. Stress, Social Support, and the Buffering Hypothesis. *Psychological Bulletin*, 98(2):310–357, 1985. doi:10.1037/0033-2909.98.2.310.

[17] A. Crabtree, T. Lodge, J. Colley, C. Greenhalgh, R. Mortier, and H. Haddadi. Enabling the New Economic Actor: Data protection, the digital economy and the Databox. *Personal and Ubiquitous Computing*, 20(6):1–11, 2016. doi:10.1007/s00779-016-0939-3.

[18] M. De Choudhury, S. Counts, and E. Horvitz. Predicting postpartum changes in emotion and behavior via social media. *Proceedings of the ACM Annual Conference on Human Factors in Computing Systems (CHI)*, pages 3267–3276, 2013. doi:10.1145/2470654.2466447.

[19] Y.-A. de Montjoye, C. A. Hidalgo, M. Verleysen, and V. D. Blondel. Unique in the Crowd: The privacy bounds of human mobility. *Scientific reports*, 3:1376, 2013. doi:10.1038/srep01376.

[20] P. Drentea and J. L. Moren-Cross. Social capital and social support on the web: The case of an internet mother site. *Sociology of Health and Illness*, 27(7):920–943, 2005. doi:10.1111/j.1467-9566.2005.00464.x.

[21] N. B. Ellison, C. Steinfield, and C. Lampe. The benefits of facebook "friends:" Social capital and college students' use of online social network sites. *Journal of Computer-Mediated Communication*, 12(4):1143–1168, 2007. doi:10.1111/j.1083-6101.2007.00367.x.

[22] N. B. Ellison, J. Vitak, C. Steinfield, R. Gray, and C. Lampe. Negotiating Privacy Concerns and Social Capital Needs in a Social Media Environment. In *Privacy Online*, volume 15, chapter 3, pages 19–32. Springer Berlin Heidelberg, Berlin, Heidelberg, 2011. doi:10.1007/978-3-642-21521-6_3.

[23] R. Gross, A. Acquisti, and H. J. H. Iii. Information Revelation and Privacy in Online Social Networks. pages 71–80, 2005.

[24] S. Guha, K. Tang, and P. Francis. NOYB: Privacy in Online Social Networks. *Workshop on Online Social Networks (WOSN)*, pages 49–54, 2008. doi:10.1145/1397735.1397747.

[25] M. Hart, R. Johnson, and A. Stent. More Content - Less Control : Access Control in the Web 2 . 0. *Control*, pages 1–3, 2006.

[26] C. A. Heitzmann and R. M. Kaplan. Assessment of methods for measuring social support. *Health Psychology*, 7(1):75–109, 1988. doi:10.1037/0278-6133.7.1.75.

[27] D. C. Howe and H. Nissenbaum. TrackMeNot: Resisting Surveillance in Web Search. *Lessons from the Identity Trail: Anonymity, Privacy and Identity in a Networked Society*, pages 417–436, 2009.

[28] M. Iwaihara, KoheiMurakami, G.-J. Ahn, and M. Yoshikawa. Risk Evaluation for Personal Identity Management Based on Privacy Attribute Ontology. pages 183–198, 2008.

[29] B. H. Kaplan, J. C. Cassel, and S. Gore. Social Support and Health. *Medical Care*, 15(5):47–58, 1977.

[30] D. M. Keating. Spirituality and Support: A Descriptive Analysis of Online Social Support for Depression. *Journal of Religion and Health*, 52(November):1014–1028, 2013. doi:10.1007/s10943-012-9577-x.

[31] B. P. Knijnenburg and A. Kobsa. Preference-based Location Sharing : Are More Privacy Options Really Better? pages 2667–2676, 2013.

[32] Z. Kwecka, W. Buchanan, B. Schafer, and J. Rauhofer. âĂŸâĂŸI am Spartacus": privacy enhancing technologies, collaborative obfuscation and privacy as a public good. *Artif Intell Law*, pages 113–139, 2014. doi:10.1007/s10506-014-9155-5.

[33] P.-c. Lin and P.-y. Lin. Unintentional and Involuntary Personal Information Leakage on Facebook from User Interactions. *KSII Transactions on Internet and Information Systems*, 10(8):3301–3318, aug 2016. doi:10.3837/tiis.2016.07.024.

[34] T. Liu, W. N. Zhang, L. Cao, and Y. Zhang. Question popularity analysis and prediction in community question answering services. *PLoS ONE*, 9(5):1–12, 2014. doi:10.1371/journal.pone.0085236.

[35] Y. Liu, J. Bian, and E. Agichtein. Predicting information seeker satisfaction in community question answering. *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR '08*, (Section 2):483, 2008. doi:10.1145/1390334.1390417.

[36] A. Majumdar and I. Bose. Privacy Calculus Theory and Its Applicability for Emerging Technologies. volume 258, pages 191–195. 2016. doi:10.1007/978-3-319-45408-5_20.

[37] M. McPherson, L. Smith-Lovin, and J. M. Cook. Birds of a Feather: Homophily in Social Networks. *Annual Review of Sociology*, 27(1):415–444, aug 2001. doi:10.1146/annurev.soc.27.1.415.

[38] M. Mun, S. Hao, N. Mishra, K. Shilton, J. Burke, D. Estrin, M. Hansen, and R. Govindan. Personal data vaults: a locus of control for personal data streams. *Proceedings of the 6th International COnference on - Co-NEXT '10*, page 1, 2010. doi:10.1145/1921168.1921191.

[39] M. Y. Mun, D. H. Kim, K. Shilton, D. Estrin, M. Hansen, and R. Govindan. PDVLoc: A Personal Data Vault for Controlled Location Data Sharing. *ACM Trans. Sen. Netw.*, 10(4):58:1—-58:29, 2014. doi:10.1145/2523820.

[40] A. Narayanan, N. Z. Gong, and D. Song. On the Feasibility of Internet-Scale Author Identification. 2012. doi:10.1109/SP.2012.46.

[41] A. Narayanan and V. Shmatikov. De-anonymizing social networks. *Proceedings - IEEE Symposium on Security and Privacy*, pages 173–187, 2009. doi:10.1109/SP.2009.22.

[42] T. H. Ngoc, I. Echizen, K. Komei, and H. Yoshiura. New Approach to Quantification of Privacy on Social Network Sites. *Advanced Information Networking and Applications (AINA), 2010 24th IEEE International Conference on*, pages 556–564, 2010. doi:10.1109/AINA.2010.118.

[43] H. Q. Nguyen-Son, Q. B. Nguyen, M. T. Tran, D. T. Nguyen, H. Yoshiura, and I. Echizen. Automatic anonymization of natural languages texts posted on social networking services and automatic detection of disclosure. *Proceedings - 2012 7th International Conference on Availability, Reliability and Security, ARES 2012*, pages 358–364, 2012. doi:10.1109/ARES.2012.18.

[44] H. J. Oh, C. Lauckner, J. Boehmer, R. Fewins-Bliss, and K. Li. Facebooking for health: An examination into the solicitation and effects of health-related social support on social networking sites. *Computers in Human Behavior*, 29(5):2072–2080, 2013. doi:10.1016/j.chb.2013.04.017.

[45] M. F. Palmer. *Virtual Communities, Social Networks and Collaboration*, volume 15 of *Annals of Information Systems*. Springer New York, New York, NY, 2012. doi:10.1007/978-1-4614-3634-8.

[46] E. Papadopoulou, A. Stobart, N. K. Taylor, and M. H. Williams. Enabling Data Subjects to Remain Data Owners. In *Smart Innovation, Systems and Technologies*, volume 38, pages 239–248. 2015. doi:10.1007/978-3-319-19728-9_20.

[47] I. G. Sarason, H. M. Levine, R. B. Basham, and B. R. Sarason. Assessing Social Support: The Social Support Questionnaire. *Journal of Personality and Social Psychology*, 44(1):127–139, 1983. doi:10.1037/0022-3514.44.1.127.

[48] P. M. Schwartz and D. J. Solove. Reconciling Personal Information in the United States and European Union. *Clr*, 102(4):877–916, 2014. doi:10.2139/ssrn.2271442.

[49] C. Shah and J. Pomerantz. Evaluating and predicting answer quality in community QA. *Proceeding of the 33rd international ACM SIGIR conference on Research and development in information retrieval - SIGIR '10*, (March 2008):411, 2010. doi:10.1145/1835449.1835518.

[50] X. Shen, B. Tan, and C. Zhai. Privacy protection in personalized search. *ACM SIGIR Forum*, 41(1):4–17, 2007. doi:10.1145/1273221.1273222.

[51] H.-h. Shuai, C.-y. Shen, D.-n. Yang, Y.-f. Lan, W.-c. Lee, P. S. Yu, and M.-s. Chen. Mining Online Social Data for Detecting Social Network Mental Disorders. pages 275–285, 2016. doi:10.1145/2872427.2882996.

[52] C. Steinfield, N. B. Ellison, and C. Lampe. Social capital, self-esteem, and use of online social network sites: A longitudinal analysis. *Journal of Applied Developmental Psychology*, 29(6):434–445, 2008. doi:10.1016/j.appdev.2008.07.002.

[53] L. Sweeney. k- ANONYMITY: A MODEL FOR PROTECTING PRIVACY. *International Journal of Uncertainty, Puzziness and Knowledge-Based Systems*, 10(5):557–570, 2002. doi:10.1142/S0218488502001648.

[54] THE EUROPEAN PARLIAMENT AND THE COUNCIL OF THE EUROPEAN UNION. Data Protection Directive, 1995. Online at http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:31995L0046:en:HTML.

[55] THE EUROPEAN PARLIAMENT AND THE COUNCIL OF THE EUROPEAN UNION. General Data Protection Regulation, 2016. Online at http://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32016R0679{&}qid=1475818206471{&}from=en.

[56] P. A. Thoits. Conceptual , Methodological , and Theoretical Problems in Studying Social Support as a Buffer Against Life Stress. *Journal of Health and Social Behavior*, 23(2):145–159, 1982.

[57] S. Trepte and L. Reinecke. *Privacy Online Perspectives on Privacy and Self-Disclosure in the Social Web*. Springer Berlin Heidelberg, Berlin, Heidelberg, 2011. doi:10.1007/978-3-642-21521-6.

[58] J. M. van der Zwaan, V. Dignum, and C. M. Jonker. Social Support Strategies for Embodied Conversational Agents. pages 134–147. 2014. doi:10.1007/978-3-319-12973-0_8.

[59] A. Viejo and J. Castellà-roca. Using social networks to distort users ' profiles generated by web search engines. *Computer Networks*, 54(9):1343–1357, 2010. doi:10.1016/j.comnet.2009.11.003.

[60] J. Vitak. The Impact of Context Collapse and Privacy on Social Network Site Disclosures. *Journal of Broadcasting & Electronic Media*, 56(4):451–470, 2012. doi:10.1080/08838151.2012.732140.

[61] C. Walker and H. Alrehamy. Personal Data Lake with Data Gravity Pull. *Proceedings - 2015 IEEE 5th International Conference on Big Data and Cloud Computing, BDCloud 2015*, pages 160–167, 2015. doi:10.1109/BDCloud.2015.62.

[62] J. Wang, N. Wang, and H. Jin. Context Matters?: How Adding the Obfuscation Option Affects End Users' Data Disclosure Decisions. In *Proceedings of the 21st International Conference on Intelligent User Interfaces - IUI '16*, pages 299–304, New York, New York, USA, 2016. ACM Press. doi:10.1145/2856767.2856817.

[63] X. Wang, K. Zhao, and N. Street. Social Support and User Engagement in Online Health Communities. *International Conference for Smart Health (ICSH) 2014*, pages 97–110, 2014.

[64] F. Zhang, V. E. Lee, and R. Jin. k -CoRating : Filling Up Data to Obtain Privacy and Utility. 2014.

[65] K. Zhao, J. Yen, G. Greer, B. Qiu, P. Mitra, and K. Portier. Finding influential users of online health communities: a new metric based on sentiment influence. *Journal of the American Medical Informatics Association*, 21(e2):e212–e218, 2014. doi:10.1136/amiajnl-2013-002282.

[66] Y. Zhao, J. Ye, and T. Henderson. Privacy-aware Location Privacy Preference Recommendations. *Proceedings of the 11th International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services*, 2014. doi:10.4108/icst.mobiquitous.2014.258017.

Glossary

social capital Social capital is the accumulated resources derived from the relationships among people within a specific social context or network.. 7, 8, 90

social support Social support is usually defined as the existence or availability of people on whom we can rely, people who let us know that they care about, value, and love us.[47]. 8, 90

social network Social networks include people communicating each others and their links within these relationships, and support is the content of interaction.[29]. 7, 90

online social network Social network in virtual communities.. 8, 90

social networking site acronym as SNS, also social network site, social networking service. 8, 90

Community Question Answering acronym as CQA, also online question and answering system, question and answer site,Q&A system. 90

self-disclosure Self-disclosure is where user reveal her explicit or implicit personal information in her generated contents. 9, 30, 90

self-presentation Self-presentation is that people present part of themselves to others to construct the impression that others might make of them.. 9, 90

identifiability A user's identifiability is the chance that we successfully identify her as true author of her questions. 31, 90

equivalence class For a question that need us to predict its author, its equivalence class is a set contains and only contains predicted users. 33, 90

feature In the context of this study, a feature is one single value in the feature set, such as the frequency of the top topic in topic distribution. 33, 41, 90

feature set In the context of this study, a feature set is a set of features extracted from original data according to one property, for instance, distribution of ten most frequent topics is a feature set with ten scalar elements. 33, 41, 90

estimation In the process of predicting users to be true author, we first estimate a probability for each question-user pair. 46, 90

estimator An estimator is the machine learning model that we use to calculate the probability from feature matrix. 46, 90

prediction In the process of predicting users to be true author, we predict true author and select them into equivalence class, by setting selection criteria, i.e. equivalence class generation algorithms, to user estimated probabilities. 46, 90