Online Social Network Based Information Disclosure Analysis

by

LI Yan

Submitted to School of Information Systems in partial fulfillment of the requirements for the Degree of Doctor of Philosophy in Information Systems

Dissertation Committee:

Yingjiu LI (Supervisor / Chair)
Associate Professor of Information Systems
Singapore Management University

Robert DENG Huijie (Co-supervisor)
Professor of Information Systems
Singapore Management University

Xuhua DING
Associate Professor of Information Systems
Singapore Management University

Tieyan LI
Security Expert of Security and Privacy Lab
Huawei Technologies Co., Ltd.

Singapore Management University
2014

Online Social Network Based Information Disclosure Analysis

LI Yan

Abstract

In recent years, online social network services (OSNs) have gained wide adoption and become one of the major platforms for social interactions, such as building up relationship, sharing personal experiences, and providing other services. A huge number of users spend a large amount of their time in online social network sites, such as Facebook, Twitter, Google+, etc. These sites allow the users to express themselves by creating their personal profile pages online. On the profile pages, the users can publish various personal information such as name, age, current location, activity, photos, etc. Sharing the personal information can motivate the interaction among the users and their friends. However, the personal information shared by users in OSNs can disclose the private information about these users and cause privacy and security issues. This dissertation focuses on investigating the leakage of privacy and the disclosure of face biometrics due to sharing personal information in OSNs.

The first work in this dissertation investigates the effectiveness of privacy control mechanisms against privacy leakage from the perspective of information flow. These privacy control mechanisms have been deployed in popular OSNs for users to determine who can view their personal information. Our analysis reveals that the existing privacy control mechanisms do not protect the flow of personal information effectively. By examining representative OSNs including Facebook, Google+, and Twitter, we discover a series of privacy exploits. We find that most of these exploits are inherent due to the conflicts between privacy control and OSN functionalities. The conflicts reveal that the effectiveness of privacy control may not be guaranteed as most OSN users expect. We provide remedies for OSN users to mitigate the

risk of involuntary information leakage in OSNs. Finally, we discuss the costs and implications of resolving the privacy exploits.

In addition to the privacy leakage, sharing personal information in OSNs can disclose users' face biometrics and compromise the security of systems, such as face authentication, which rely on the face biometrics. In the second work, we investigate the threats against real-world face authentication systems due to the face biometrics disclosed in OSNs. We make the first attempt to quantitatively measure the threat of OSN-based facial disclosure (OSNFD). We examine real-world face-authentication systems designed for both smartphones, tablets, and laptops. Interestingly, our results find that the percentage of vulnerable images that can be used for spoofing attacks is moderate, but the percentage of vulnerable users that are subject to spoofing attacks is high. The difference between the face authentication systems designed for smartphones/tablets and laptops is also significant. In our user study, the average percentage of vulnerable users is 64% for laptop-based systems, and 93% for smartphone/tablet-based systems. This evidence suggests that face authentication may not be suitable to use as an authentication factor, as its confidentiality has been significantly compromised due to OSNFD. In order to understand more detailed characteristics of OSNFD, we further develop a risk estimation tool based on logistic regression to extract key attributes affecting the success rate of spoofing attacks. The OSN users can use this tool to calculate risk scores for their shared images so as to increase their awareness of OSNFD.

This dissertation makes contributions on understanding the potential risks of private information disclosure in OSNs. On one hand, we analyze the underlying reasons which make the privacy control deployed in OSNs vulnerable against privacy leakage. On the other hand, we reveal that the face biometrics can be disclosed in OSNs and compromise the security of face authentication systems.

Table of Contents

List of Figures

List of Tables

Acknowledgments

I would like to thank Associate Professor Yingjiu LI, Professor Robert DENG, Associate Professor Xuhua DING, and Doctor Tieyan LI for their guidance in completing my dissertation.

I also thank my friends Qiang YAN, LO Swee Won, Shaoying CAI, Jilian ZHANG, Freddy CHUA, and Ke XU for the research collaboration, their friendship, and their encouragement.

Finally, I would like to thank Yanming TANG, Benjamin LI, Sujun SONG, and Qingbin LI, who are my family and always supporting me and encouraging me with their best wishes.

Dedication

I dedicate my dissertation work to Yanming TANG and Benjamin Z. LI for your love and encouragement.

Chapter 1

Introduction

Online Social Network Services (OSNs) is a platform for social interactions, such as building up relationship, sharing personal experiences, and providing other services. A typical OSN consists of each user's profile, his/her social links, and various additional services. Early OSNs, such as Classmate.com [17], simply brought users together in chatting rooms and encouraged them to share their information via personal webpages. Then new generation of OSNs has begun to flourish since 2000. These new OSNs develop more advanced features for users to find and manage friends, and share information. Now Facebook [21], Google+ [26], Twitter [68], and LinkedIn [50] have become the largest OSNs in the world.

About 82% online population use at least one OSN such as Facebook, Google+, Twitter, etc [5]. Via OSNs, massive amount of personal data, such as personal images and interests, is published online and accessed by users from all over the world. According to a recent report by Facebook, averages 350 million personal images are published by users on Facebook every day. The wide adoption of OSNs raises concerns about private information disclosure due to personal data shared online. The disclosure of the private information poses threats to privacy and security and may eventually cause severe impact on people's daily life, such as breaking relationship, losing job, and resulting public embarrassment [9, 12].

As OSNs become a landmine for privacy and security issues, the debate on

these issues has been opened for over a decade. Prior research shows that the information disclosed in OSNs can leak user privacy and threaten security systems [80, 41, 14, 7, 3, 31]. For example, seemingly harmless data, such as personal interests and shopping patterns, could leak sensitive private information including sexual preference [36]. To prevent information disclosure, privacy control mechanisms are deployed by OSNs to allow users to control who can access their information. Also, significant research efforts have been made to improve security and usability of the privacy control [11, 73, 76, 24]. However, the private information can still be disclosed even if privacy control mechanisms are properly deployed and configured. This raises questions why privacy control is vulnerable against the information disclosure in OSNs and what potential threats can be caused by the information disclosure.

This dissertation investigates the effectiveness of privacy control against information disclosure in OSNs and the threat of OSN-based face biometric disclosure. We first analyze the underline reasons that make the privacy control in OSNs vulnerable to the information disclosure, and then study the OSN-based face biometric disclosure threat against real-world face authentication systems.

1.1 Analyzing OSN-based Privacy Leakage

The first work in this dissertation reveals the underlying reasons that make the privacy control vulnerable against privacy leakage. As online Social Network services (OSNs) become an essential element in modern life for human beings to stay connected to each other, people are publishing various personal data and exchanging information with their friends in OSNs. Although most OSNs deploy privacy control mechanisms to prevent unauthorized access to the personal data, it is still possible to infer such data from *publicly* shared information as shown in prior research [80, 41, 14, 7]. Thus it raises a question how effective the existing privacy control mechanisms are against privacy leakage in OSNs.

To answer the above question, we investigate the problem of *privacy leakage under privacy control* (PLPC). PLPC refers to private information leakage even when privacy rules are properly configured and enforced. Instead of focusing on new attacks, we analyze the underlying reasons that make privacy control vulnerable from the perspective of information flow. Based on the analysis, we inspect representative real-world OSNs including Facebook, Google+, and Twitter. Our analysis reveals that the existing privacy control mechanisms do not protect the flow of personal information effectively. Privacy exploits and their corresponding attacks are identified in the above OSNs.

According to our analysis, most of the privacy exploits are caused by the conflicts between privacy control and essential OSN functionalities. Therefore, the effectiveness of privacy control may not be guaranteed even if it is technically achievable. We analyze the feasibility of our identified attacks through user study. Suggestions are provided for users to minimize the risk of involuntary information leakage when sharing private personal information in OSNs. We further discuss the costs and implications of resolving these privacy exploits.

1.2 Understanding OSN-based Facial Disclosure

As numerous personal data, especially personal images, are being published in OSNs such as Facebook, Google+, and Instagram, users' biometrics information, such as face biometrics, can be disclosed in OSNs. The disclosed face biometrics can further lead to security issues to the systems relying on the face biometrics, such as face authentication systems. The second work in this dissertation investigates the threat of face biometrics disclosure.

The OSN images chosen and published by users usually contain facial images where the users' faces can be clearly seen. The large base number indicates that these shared personal images could become an abundant resource for potential attackers to exploit, which introduces the threat of OSN-based facial disclosure (OS-

NFD). OSNFD may have a significant impact on the current face authentication systems which have been widely available on all kinds of consumer-level computing devices such as smartphones, tablets, and laptops with built-in camera capability.

In this study, we make the first attempt to quantitatively measure the threat of OSNFD against real-world face authentication systems for smartphones, tablets, and laptops. Our study collects users' facial images published in OSNs and uses them to simulate the spoofing attacks against these systems. Our study indicates that face authentication may not be suitable to use as an authentication factor. Although the percentage of vulnerable images that can be used for spoofing attacks is moderate, the percentage of vulnerable users that are subject to spoofing attacks is high. On average, the percentage of vulnerable users is 64% for laptop-based systems, and 93% for smartphone/tablet-based systems. OSNFD would compromise the confidence of face authentication significantly.

In order to understand more detailed characteristics of OSNFD, we propose a risk estimation tool. The risk estimation tool can help users estimate the risk of an uploaded image to face authentication and make them aware of the threat of OSNFD.

1.3 Contributions and Organization

To summarize, the following contributions have been made in this dissertation:

- We investigate the interaction between privacy control and information flow in OSNs. We show that the conflict between privacy control and essential OSN functionalities restricts the effectiveness of privacy control in OSNs. We identify privacy exploits for current privacy control mechanisms in typical OSNs, including Facebook, Google+, and Twitter. Based on these privacy exploits, we introduce a series of attacks for adversaries with different capabilities to obtain private personal information. We investigate the necessary

conditions for protecting against privacy leakage due to the discovered exploits and attacks. We provide suggestions for users to minimize the risk of privacy leakage in OSNs. We also analyze the costs and implications of resolving discovered exploits. While it is possible to fix the exploits due to implementation defects, it is not easy to eliminate the inherent exploits due to the conflicts between privacy control and the functionalities. These conflicts reveal that the effectiveness of privacy control may not be guaranteed as most OSN users expect.

- We investigate the threat of OSN-based face disclosure (OSNFD) against face authentication. Our results suggest that face authentication may not be suitable to use as an authentication factor, as its confidentiality has been significantly compromised by OSNFD. We make the first attempt to quantitatively measure the threat of OSNFD by testing real-world face authentication systems designed for smartphones, tablets, and laptops. We also build a dataset containing important image attributes that significantly affect the success rate of spoofing attacks. These attributes are common in real-life photos but rarely used in prior controlled study on face authentication [16, 30]. We use logistic regression to extract key attributes that affect the success rate of spoofing attacks. These attributes are further used to develop a risk estimation tool to help users measure the risk score of uploading images to OSNs.

The reminder of this dissertation is organized as follows: Chapter 2 is a literature review which examines closely related research on information disclosure in OSNs. Chapter 3 investigates the OSN-based privacy leakage under privacy control. Chapter 4 studies the OSN-based facial disclosure threat against face authentication systems. Finally, Chapter 5 summarizes the contributions of this dissertation.

Chapter 2

Literature Review

Due to wide adoption of OSNs, the privacy and security problems caused by OSNs have attracted strong interest among researchers. We summarize the closely related research work from the following aspects: attacks to privacy, privacy settings, access control models, face recognition, spoofing attack to face authentication, and liveness detection.

In OSNs, the users' privacy leakage is a major concern. The attack techniques against privacy proposed in prior literature mainly focus on inferring users' identity [6] and other personal information [80, 7, 14] from public information shared in OSNs. Zheleva et al. [80] proposed a classification-based approach to infer users' undisclosed personal particulars from their social relationships and group information which are publicly shared. Chaabane et al. [14] proposed to infer users' undisclosed personal particulars from public shared interests and public personal particulars of other users who have similar interests. Balduzzi et al. [7] utilized email addresses as unique identifiers to identify and link user profiles across several popular OSNs. Since users' information may be shared publicly in an OSN but not be shared in another OSN, certain hidden information can be revealed by combining public information collected from different OSNs. The effectiveness of these attacks largely depends on the quality of public information, which can be affected due to users' awareness of privacy concerns. As reported in [14], only 18% of Face-

book users now publicly share their social relationships and 2% of Facebook users publicly share their dates of birth. Thus it is more realistic to analyze the threats caused by more powerful adversaries or insiders as in our analysis.

The threat of privacy leakage caused by insiders is also mentioned by Johnson et al. [41]. They investigated users' privacy concerns on Facebook and discovered that the privacy control mechanisms in existing OSNs help users manage outsider threats effectively but cannot mitigate insider threats because users often wrongly include inappropriate audiences as members of their friend network. Wang et al. [73] analyzed reasons why users wrongly configure privacy settings and provided suggestions for users to avoid such mistakes. To help users handle complex privacy policy management, Cheek et al. [15] proposed two approaches using clustering techniques to assist users in grouping friends and setting appropriate privacy rules. However, as shown in our work, privacy leakage could still happen even if a user correctly configures his privacy settings due to the exploits caused by inherent conflicts between privacy control and OSN functionalities.

Some researchers addressed the privacy control problem in traditional access control modeling. Several models [24, 11] are established to provide more flexible and fine-grained control so as to increase the expressive power of privacy control models. Nevertheless, this is not sufficient to guarantee effective privacy protection. From our analysis on information flows, OSN functionalities may be affected by privacy control. On the other hand, a more complex privacy control model increases users' burden on configuring privacy rules.

One of the exploits found in our work (Exploit 5) is also mentioned in previous research on resolving privacy conflicts in collaborative data sharing. Wishart et al. [76] and Hu et al. [37] analyzed co-owned information disclosure due to conflicts of privacy rules set by multiple owners. They also introduced a negotiation mechanism to seek a balance between the risk of privacy leakage and the benefit of data sharing. Compared to them, our work investigates a broader range of privacy threats in OSNs, discovers the underlying conflicts between privacy control

and social/business values of OSNs, and analyzes the difficulty in resolving these conflicts, which have not been addressed in previous works.

Besides privacy leakage, the security problems caused by OSNs become another concern, among which the disclosure of face biometrics is a typical example and may significantly threaten face authentication systems. In face authentication, face recognition is a core module for matching the face biometrics. Holistic approaches and local landmark based approaches are the two major types of popular face recognition algorithms [1, 79]. The holistic approaches, such as PCA-based algorithms and LDA-based algorithms, use the whole face region as input. Local landmark based approaches extract local facial landmarks such as eyes, nose, mouth, etc and feed locations and local statistics of these local facial landmarks into a structure classifier. As an important application of face recognition, face authentication validates a claimed identity based on comparison between a facial image and an enrolled facial image and determines either accepting or rejecting the claimed identity [53]. Trewin et al. [67] show that the face authentication is faster and causes lower interruption of user memory recall task than voice, gesture, and typical password entry. Another advantage of face authentication is that it provides stronger defense against repudiation than token based authentication and password based authentication [55]. Besides face authentication, face identification is another application of face recognition, which compare a facial image with multiple registered users and identifies the user in the facial images. The face identification can cause privacy leakage in OSNs due to the identifiable personal images published in OSNs [3, 29]. Compared to their work, our study focuses on investigating the impact of the shared personal images that can be used to attack face authentication systems.

It is a well-known fact that face authentication is subject to spoofing attacks. An attacker can pass the authentication by displaying images or videos of a legitimate user in hard copy or on the screen [8]. But it is generally believed sufficiently secure as an authentication factor for common access protection, as an adversary

usually has to be physically proximate to a victim in order to collected required face biometrics. Our findings indicate that this belief is not valid as the emergence of OSNFD. Face biometrics can now be disclosed in large scale and acquired by a remote adversary.

Liveness detection is the major countermeasure designed to mitigate the risk of spoofing attacks. Interaction based approach, multi-modal based approach, and motion based approach are three popular types of liveness detection [56, 42, 4]. Interaction based approaches require real-time responses from claimants, including eye blink, head rotation, facial expression, etc. However, these approaches can be bypassed with one or two images [59]. Multi-modal based approaches take face biometric and other biometrics into consideration together such as voice, facial thermogram, etc [56]. The multi-modal based approaches require additional hardware and specific environment. Motion based approaches are based on the detection of involuntary motions of a 3D face, such as involuntary rotation of head [42]. The approaches require high quality images captured with ideal lighting condition. Compared to these approaches, our estimation tool addresses this problem from a different perspective. Since OSNFD significantly compromise the confidentiality of face authentication, our tool is designed to increase the users' awareness before they publish their personal images so as to reduce the number of exploitable images available to an adversary.

Chapter 3

Analyzing Privacy Leakage under Privacy Control in Online Social Networks

3.1 Introduction

This chapter investigates the effectiveness of privacy control mechanisms against privacy leakage in online social networks. According to a recent report, about 82% online population use at least one OSN such as Facebook, Google+, Twitter, and LinkedIn, which facilitates building relationship, sharing personal experiences, and providing other services [5]. Via OSNs, massive amount of personal data is published online and accessed by users from all over the world. Prior research [80, 41, 14, 7] shows that it is possible to infer undisclosed personal data from *publicly* shared information. Nonetheless, the availability and quality of the public data causing privacy leakage are decreasing due to the following reasons: 1) privacy control mechanisms have become the standard feature of OSNs and keep evolving. 2) the percentage of users who choose *not* to publicly share information is also increasing [14]. In this tendency, it seems that privacy leakage could be *prevented* as increasingly comprehensive privacy control is in place. However, this

may not be achievable according to our findings.

Instead of focusing on new attacks, we investigate the problem of *privacy leakage under privacy control* (PLPC). PLPC refers to private information leakage even if privacy rules are properly configured and enforced. For example, Facebook allows its users to control over who can view their friend lists on Facebook. Alice, who has Bob in her friend list on Facebook, may not allow Bob to view her complete friend list. As an essential functionality, Facebook recommends to Bob a list of users, called "*people you may know*", to help Bob make more friends. This list is usually compiled by enumerating the friends of Bob's friends on Facebook, which includes Alice's friends. Even though Alice doesn't allow Bob to view her friend list, Alice's friend list could be leaked as recommendation to Bob by Facebook.

We investigate the underlying reasons that make privacy control vulnerable from the perspective of information flow. We start with categorizing the personal information of an OSN user into three *attribute sets* according to *who the user is*, *whom the user knows*, and *what the user does*, respectively. We model the information flow between these attribute sets and examine the functionalities which control the flow. We inspect representative real-world OSNs including Facebook, Google+, and Twitter, where privacy exploits and their corresponding attacks are identified.

Our analysis reveals that most of the privacy exploits are inherent due to the underlying conflicts between privacy control and essential OSN functionalities. The recommendation feature for social relationship is a typical example, where it helps expanding a user's social network but it may also conflict with other users' privacy concerns for hiding their social relationships. Therefore, the effectiveness of privacy control may not be guaranteed even if it is technically achievable. We investigate necessary conditions for protecting against privacy leakage due to the discovered exploits and attacks. Based on the necessary conditions, we provide suggestions for users to minimize the risk of involuntary information leakage when sharing private personal information in OSNs.

We analyze the feasibility of our identified attacks through user study, in which

we investigate participants' usage, knowledge, and privacy attitudes towards Facebook, Google+, and Twitter. Based on the collected data, we evaluate the feasibility of leaking the private information of these participants. We further discuss the costs and implications of resolving these privacy exploits.

We summarize the contributions of this paper as follows:

- We investigate the interaction between privacy control and information flow in OSNs. We show that the conflict between privacy control and essential OSN functionalities restricts the effectiveness of privacy control in OSNs.

- We identify privacy exploits for current privacy control mechanisms in typical OSNs, including Facebook, Google+, and Twitter. Based on these privacy exploits, we introduce a series of attacks for adversaries with different capabilities to obtain private personal information.

- We investigate necessary conditions for protecting against privacy leakage due to the discovered exploits and attacks. We provide suggestions for users to minimize the risk of privacy leakage in OSNs. We also analyze the costs and implications of resolving discovered exploits. While it is possible to fix the exploits due to implementation defects, it is not easy to eliminate the inherent exploits due to the conflicts between privacy control and the functionalities. These conflicts reveal that the effectiveness of privacy control may not be guaranteed as most OSN users expect.

The rest of this paper is organized as follows: Section 3.2 provides background information about OSNs. Section 3.3 presents our threat model and assumptions. Section 3.4 models information flows between attribute sets in OSNs. Section 3.5 presents discovered exploits, attacks, and mitigations for the exploits. Section 3.6 analyzes the feasibility of the attacks. Section 3.7 discusses the implications of our findings.

3.2 Background

In a typical OSN, Alice owns a space which consists of a *profile page* and a *feed page* for publishing Alice's personal information and receiving other users' personal information, respectively. Alice's profile page displays Alice's personal information, which can be viewed by others. Alice's feed page displays other users' personal information which Alice would like to keep up with. The personal information in a user's profile page can be categorized into three *attribute sets*: a) personal particular set (PP set), b) social relationship set (SR set), and c) social activity set (SA set), according to who the user is, whom the user interact with, and what the user does, respectively. We show corresponding personal information and attribute sets on Facebook, Google+, and Twitter in Table 3.1.

Alice's PP set describes persistent facts about Alice in an OSN, such as gender, date of birth, and race, which usually do not change frequently. Alice's SR set records her social relationships in an OSN, which consist of an *incoming list* and an *outgoing list*. The incoming list consists of the users who include Alice as their friends while the outgoing list consists of the users whom Alice includes as her friends. In particular, on Google+, the incoming list and the outgoing list correspond to "have you in circles" and "your circles", respectively. On Twitter, the incoming list and the outgoing list correspond to "following" and "follower", respectively. The social relationships in certain OSNs are mutual. For example, on Facebook, if Alice is a friend of Bob, Bob is also a friend of Alice. In such a case, a user's incoming list and outgoing list are the same, which are called friend list. Lastly, Alice's SA set describes Alice's social activities in her daily life. The SA set includes status messages, photos, links, videos, etc.

To enable users protect their personal information in the three attribute sets, most OSNs provide privacy control, by which users may set up certain *privacy rules* to control the disclosure of their personal information. Given a piece of personal information, the privacy rules specify who can/cannot view the information. A

privacy rule usually contains two types of lists, *white list*, and *black list*. A white list specifies who can view the information while a black list specifies who cannot view the information. A white/black list could be local or global. If a white/black list is local, this list takes effect on specific information only (e.g. an activity, age information, or gender information). If a white/black list is global, this list takes effect on all information in a user's profile page. For example, if Alice wants to share a status with all her friends except Bob, Alice may use a local white list which includes all Alice's friends, as well as a local black list which includes Bob only. If Alice doesn't want to share any information with Bob, she may use a global black list which includes Bob.

To help users share their personal information and interact with each other, most OSNs provide four basic functionalities including PUB, REC, TAG, and PUSH. The first three functionalities, PUB, REC, and TAG, mainly affect the personal information displayed in a user's profile page, while the last functionality PUSH makes some other users' personal information appear in the user's feed page. These basic functionalities are described as follows. We exclude any other functionalities which are not relevant to our findings.

Alice can use PUB functionality to share her personal information with other users. As shown in Figure 3.1(a), PUB displays Alice's personal information in her profile page. Other users may view Alice's personal information in Alice's profile page.

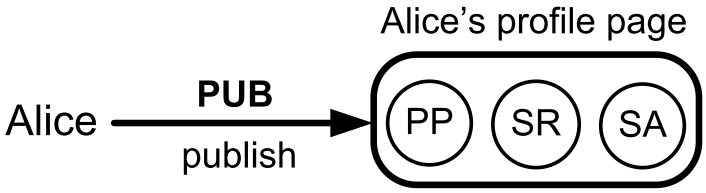
To help Alice make more friends in an OSN, REC is an essential functionality by which the OSN recommends to Alice a list of users that Alice may include in her SR set. The list of recommended users is composed based on the social relationships of the users in Alice's SR set. Considering an example shown in Figure 3.1(b), Alice's SR set consists of Bob while Bob's SR set consists of Alice, Carl, Derek, and Eliza. After Alice logs into her space, REC automatically recommends Carl, Derek, and Eliza to Alice who may update her SR set. If Alice intends to include Carl in her SR set, Alice may need Carl's approval depending on OSN implementations. Upon

Table 3.1: Types of Personal Information on Facebook, Google+, and Twitter

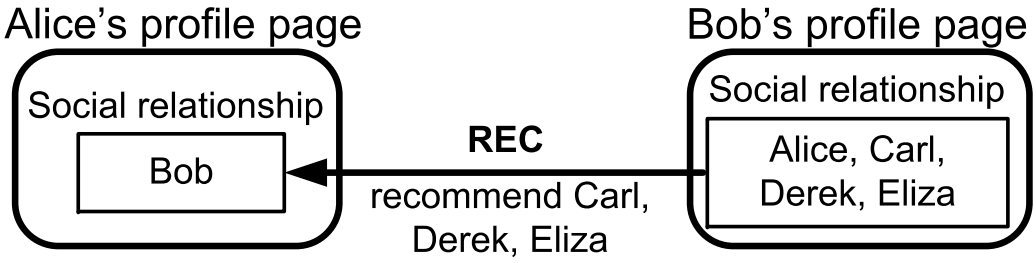
Acronym	Attribute set	Facebook	Google+	Twitter
PP	Personal Particulars	Current city, hometown, sex, birthday, relationship status, employer, college/university, high school, religion, political views, music, books, movies, emails, address, city, zip	Taglines, introduction, bragging rights, occupation, employment, education, places lived, home phone, relationship, gender	Name, location, bio, website
SR	Social Relationship (incoming list, outgoing list)	Friends, friends	Have you in circles, your circles	Following, follower
SA	Social Activities	Status message, photo, link, video, comments, like	Post, photo, comments, link, video, plus 1's	Tweets

approval if needed, Alice can include Carl in her SR set. At the same time, Alice is automatically included in Carl's SR set. In particular, on Facebook, if Alice intends to include Carl in her SR set, Alice needs to get Carl's approval. Upon approval, Alice includes Carl in her friend list. Meanwhile, Facebook automatically includes Alice in Carl's friend list. On Google+, Alice can include Carl in her outgoing list without Carl's approval. Then Google+ automatically includes Alice in Carl's incoming list. On Twitter, if Alice intends to include Carl in her SR set, Alice may need Carl's approval depending on Carl' option whether his approval is required. Upon approval if required, Alice includes Carl in her incoming list. Then Twitter includes Alice in Carl's outgoing list automatically.

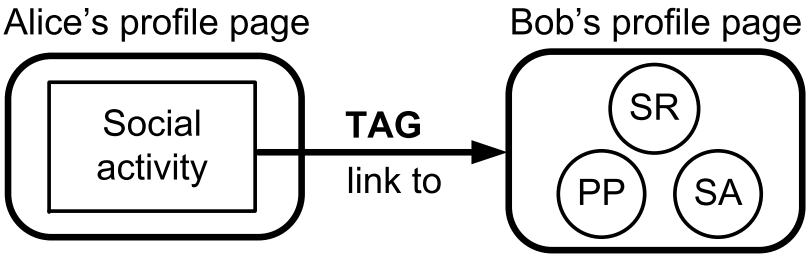
To motivate users' interactions, TAG functionality allows a user to mention an-



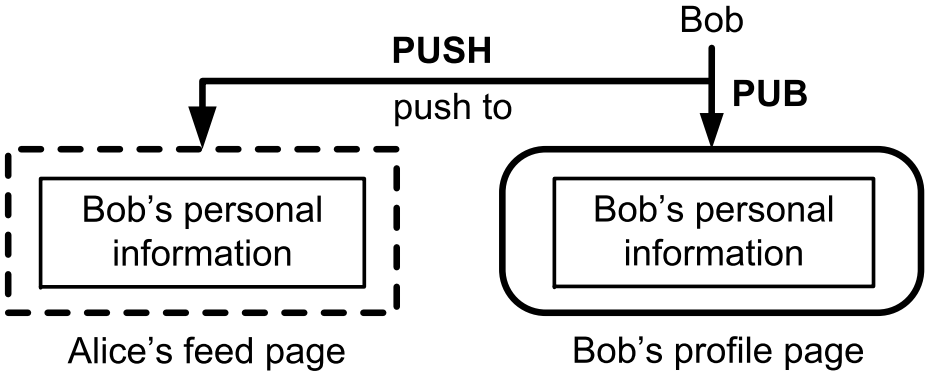
(a) Alice publishes her personal information



(b) Bob's social relationships are recommended to Alice



(c) Alice tags Bob in her social activity



(d) Bob's personal information is pushed to Alice's feed page when Bob publishes his personal information

Figure 3.1: Basic functionalities in OSNs

other user's name in his/her social activities when the user publishes social activities in his/her profile page. In Figure 3.1(c), when Alice publishes a social activity in her profile page, she can mention Bob in the social activity via TAG, which provides a link to Bob's profile page (shown as a HTML hyperlink).

For the convenience of keeping up with the personal information published by other users, OSNs provides *feed page* for users. Considering an example in which Alice intends to keep up with Bob, Alice can subscribe to Bob, and Alice is called Bob's *subscriber*. As Bob's subscriber, Alice is included in Bob's SR set. In particular, on Facebook, a user's subscribers are usually his/her "friends". On Google+, a user's subscribers are usually the users in his/her outgoing list, i.e. "your circles". On Twitter, a user's subscribers are usually the users in his/her incoming list, i.e. "follower". Figure 3.1(d) shows that when Bob updates his personal information via PUB and allows Alice to view the updated personal information, a copy of the updated personal information is automatically pushed to Alice's feed page via PUSH. Then, Alice can view Bob's updated personal information both in her feed page and in Bob's profile page.

3.3 Threat Model

The problem of PLPC investigates privacy leakage in a system where privacy control is enforced. Given a privacy control mechanism, *PLPC* examines whether a user's private personal information is leaked *even if* the user properly configures privacy rules to protect the corresponding information.

The problem of PLPC in OSNs involves two parties, *distributor* and *receiver*. A user who publishes and shares his/her personal information is a *distributor* while the user whom the personal information is shared with is a *receiver*. An *adversary* is a receiver who intends to learn a distributor's information that is not shared with him. Correspondingly, the target distributor is referred to as *victim*.

Prior research [80, 14, 7] mainly focuses the inference of undisclosed user information from their *publicly* shared information. Since the effectiveness of these inference techniques will be hampered by increasing user awareness of privacy concern [14], we further include *insiders* in our analysis. The adversaries have the incentive to register as OSN users so that they may directly access a victim's private

personal information or infer the victim's private personal information from other users connected with the victim in OSNs.

The capabilities of an adversary can be characterized according to two factors. The first factor is the distance between adversary and victim. According to privacy rules available in existing OSNs, a distributor usually chooses specific receivers to share her information based on the distance between the distributor and the receivers. Therefore, we classify an adversary's capability based on his distance to a victim. Considering the social network as a directed graph, the distance between two users can be measured by the number of hops in the shortest connected path between the two users. An n-hop adversary can be defined such that the length of the shortest connected path from victim to adversary is n hops. We consider the following three types of adversaries in our discussion, 1-hop adversary, 2-hop adversary, and k-hop adversary, where $k > 2$. On Facebook, they correspond to Friend-only, Friend-of-Friend, and Public, respectively. On Google+, they correspond to Your-circles, Extended-circles, and Public, respectively. For ease of readability, we use *friend*, *friend of friend*, and *stranger* to represent 1-hop adversary, 2-hop adversary, and k-hop adversary (where $k > 2$) adversaries respectively: 1) If an adversary is a friend of a victim, he is stored in the outgoing list in the victim SR set. The adversary can view the victim's information that is shared with her friends, friends of friends, or all receivers in an OSN. However, the adversary cannot view the information that is not shared with any receivers (e.g. the "only me" option on Facebook). 2) If an adversary is a friend of friend, he can view the victim's information shared with her friend-of-friends or all receivers. However, the adversary cannot view any information that is shared with friends only, or any information that is not shared with any receivers. 3) If an adversary is a stranger, he can access the victim's information that is shared with all receivers. However, the adversary cannot view any information which is shared with friends of friends and friends.

Besides the above restrictions, an adversary cannot view a victim's personal information if the adversary is included in the victim's black lists (e.g. "except" or

"block" option on Facebook, and "block" option on Google+).

An adversary may have prior knowledge about a victim. We will specify the exact requirement of such prior knowledge for different attacks in Section 3.5.

Since a user may use multiple OSNs, it is possible for an adversary to infer the user's private data by collecting and analyzing the information shared in different OSNs. We exclude social engineering attacks where a victim is deceived to disclose her private information voluntarily. We also exclude privacy leakage caused by improper privacy settings. These two cases cannot be addressed completely by any technical measures alone.

3.4 Information Flows Between Attribute Sets in Profile Pages

In this section, we examine explicit and implicit information flows in OSNs. These information flows could leak users' private information to an adversary even after the users have properly configured the privacy rules to protect their information.

As analyzed in Section 3.2, the personal information shared in a user's profile page can be categorized into three attribute sets including PP set, SR set, and SA set, which are illustrated as circles in Figure 3.2. The attribute sets of multiple users are connected within an OSN, where personal information may explicitly flow from a profile page to another profile page via *inter-profile functionalities*, including REC (recommending) and TAG (tagging), as represented by solid arrows and rectangles in Figure 3.2. It is also possible to access a user's personal information in PP set and SR set via implicit information flows marked by dashed arrows. The details about these information flows are described below.

The first explicit flow is caused by REC, as shown in arrow (1) in Figure 3.2. REC recommends to an OSN user Bob a list of users according to the social relationships of the users included in Bob's SR set. Therefore, the undisclosed users

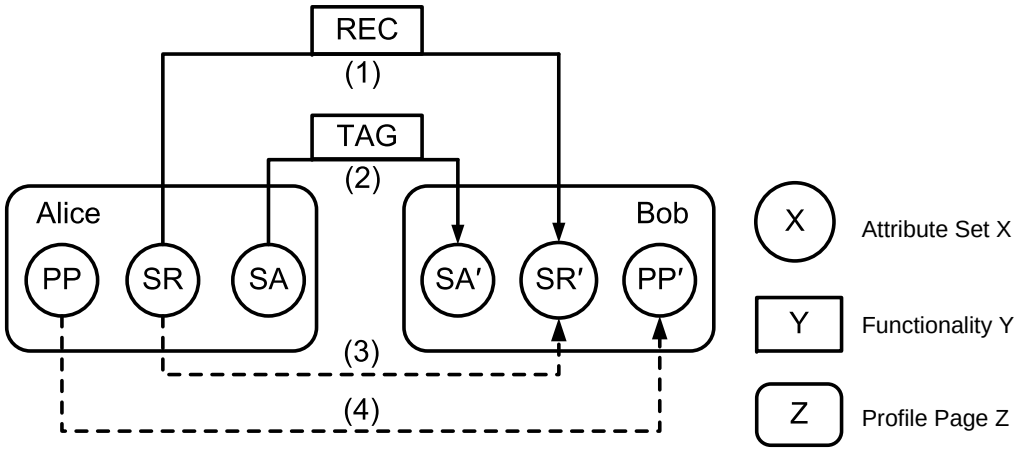


Figure 3.2: Information flows between attribute sets

included in Alice's SR may be recommended to Bob via REC, if Bob is connected with Alice.

The second explicit flow caused by TAG is shown in arrow (2) in Figure 3.2. A typical OSN user may mention the names of other users in a social activity in SA set in his/her profile page via TAG, which creates explicit links connecting SA sets within different profile pages.

The third flow is an implicit flow caused by the design of information storage for SR sets, which is shown in arrow (3) in Figure 3.2. A user's SR set stores his/her social relationships as connections. From the perspective of information flow, a *connection* is a directional relationship between two users, including a *distributor* and his/her *1-hop receiver*, i.e., *friend*. The direction of a connection represents the direction of information flow. Correspondingly, Alice's SR set consists of an *incoming list* and an *outgoing list* as defined in Section 3.2. For each user u_i in Alice's incoming list, there is a connection from u_i to Alice. For each user u_o in Alice's outgoing list, there a connection from Alice to u_o. Alice can receive information distributed from the users in her incoming list, and distribute her information to the users in her outgoing list. Given a connection from Alice to Bob, Bob is included in the outgoing list in Alice's SR set. Meanwhile Alice is included in the incoming list in Bob's SR set. The social relationships in certain OSNs such as Facebook are

mutual. Such mutual relationship can be considered as a pair of connections linking two users with opposite directions, similar to replacing a bidirectional edge with two equivalent unidirectional edges.

The fourth flow is an implicit flow related to PP set, which is shown as the arrow (4) in Figure 3.2. Due to the homophily effect [52, 13], a user is more willing to connect with the users with similar personal particulars compared to other users with different personal particulars. This tendency can be used to link PP sets of multiple users. For example, colleagues working in the same department are often friends with each other on Facebook.

In addition to the above information flows, an OSN user may simultaneously use multiple OSNs, and thus create other information flows connecting the attribute sets of the same user across different OSNs.

It is difficult to prevent privacy leakage from all these information flows. A user may be able to prevent privacy leakage caused by explicit information flows by carefully using corresponding functionalities, as these flows are materialized only when inter-profile functionalities are used. However, it is difficult to avoid privacy leakage due to implicit information flows, as they are caused by inherent correlations among the information shared in OSNs. In fact, all these four information flows illustrated in Figure 3.2 correspond to inherent exploits, which will be analyzed in Section 3.5 and 3.7. The existence of these information flows introduces a large attack surface for an adversary to access undisclosed personal information if any of these flows is not properly protected. The existing privacy control mechanisms [11, 24] regarding data access within a profile page are not sufficient to prevent against privacy leakage. However, the full coverage of privacy control may not be feasible as it conflicts with social/business values of OSNs as analyzed in Section 3.7.

In this paper, we focus on the information flows from the attribute sets in a profile page to the attribute sets in another profile page, which may lead to privacy leakage even if users properly configure their privacy rules. There may exist other exploitable information flows leading to privacy leakage, which are left as our future

work.

3.5 Exploits, Attacks, And Mitigations

In this section, we analyze the exploits and attacks which may lead to privacy leakage in existing OSNs even if privacy controls are enforced. We organize the exploits and attacks according to their targets, which could be a victim's PP set, SR set, and SA set. We also investigate necessary conditions regarding prevention of privacy leakage due to the identified exploits and attacks. Based on these necessary conditions, we provide suggestions on mitigating the corresponding exploits and attacks. All of our findings have been verified in real-world settings on Facebook, Google+, and Twitter[1].

3.5.1 PP Set

A user's PP set describes persistent facts about who the user is. The undisclosed information in PP set protected by existing privacy control mechanisms can be inferred by the following inherent exploits, namely *inferable personal particular* and *cross-site incompatibility*.

Inferable Personal Particular

Human beings are more likely to interact with others who share the same or similar personal particulars (such as race, organization, and education) [52, 13, 36]. This phenomenon is called homophily. Due to homophily [52, 13], users are connected with those who have similar personal particulars at higher rate than with those who have dissimilar personal particulars. This causes an inherent exploit named *inferable personal particulars*, which corresponds to the information flow shown as dashed arrow (4) in Figure 3.2.

[1] All of our experiments were conducted from September, 2011 to September, 2012

Exploit 1: *If most of a victim's friends have common or similar personal particulars (such as employer information), it could be inferred that the victim may have the same or similar personal particulars.*

An adversary may use Exploit 1 to obtain undisclosed personal particulars in a victim's PP set. The following is a typical attack on Facebook.

Attack 1: Considering a scenario on Facebook shown in Figure 3.3, where Bob, Carl, Derek, and some other users are Alice's friends, and Bob is a friend of Carl, Derek, and most of Alice's friends (Note that in Figure 3.3, a solid arrow connects from a distributor to a friend of the distributor). Alice publishes her employer information "*XXX Agency*" in her PP set and allows Carl and Derek only to view her employer information. However, most of Alice's friends may publish their employer information and allow their friends to view this information due to different perceptions in privacy protection. In this setting, Bob can collect the employer information of Alice's friends and infer that Alice's employer is "*XXX Agency*" with high probability.

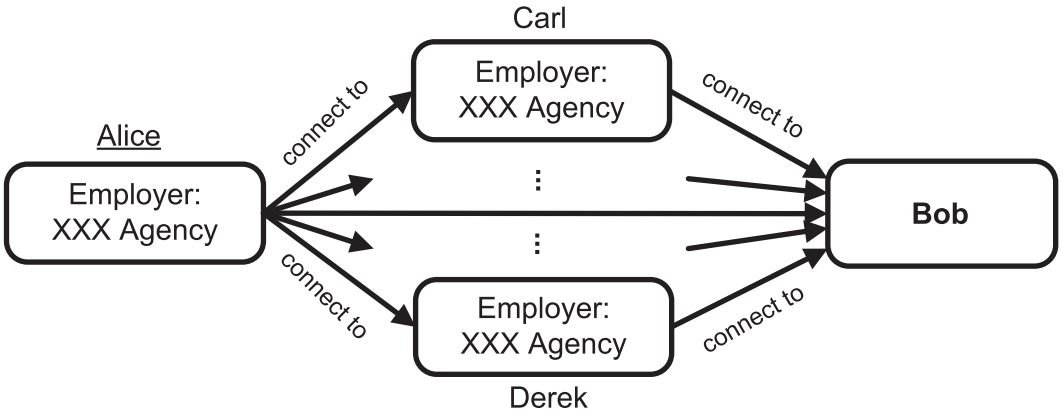


Figure 3.3: Alice and most of her friends have common personal particulars (e.g. employer information)

The above attack works on Facebook, Google+, and Twitter. The attack can be performed by any adversary who has two types of knowledge. The first type of knowledge includes a large portion of users stored in the victim's SR set. The second type of knowledge includes the personal particulars of these users. To prevent

against privacy leakage due to Exploit 1, the following necessary condition should be satisfied

Necessary Condition 1: *Given a subset* $U = \{u_1, u_2, ..., u_n\}$ *of a victim* v*'s* SR *set in an OSN and personal particular value* pp_{u_i} *(*$pp_{u_i} \neq$ null*) of each receiver* $u_i \in U$ *which are obtained by an adversary, there exists at least one personal particular value* pp *such that* $|U_{pp}| \geq |U_v|$ *and* $pp \neq pp_v$ *where* pp_v *is the victim's personal particular value and* $U_{pp} = \{u_i | (u_i \in U) \wedge (pp_{u_i} = pp)\}$ *and* $U_v = \{u_j | (u_j \in U) \wedge (pp_{u_j} = pp_v)\}$.

Proof. The input of an adversary includes two types of knowledge about a victim: a subset $U = \{u_1, u_2, ..., u_n\}$ of a victim v's SR set in an OSN, and personal particular value pp_{u_i} ($pp_{u_i} \neq$ null) of each receiver $u_i \in U$. The adversary may infer the victim's personal particular pp_v ($pp_v \neq$ null) by calculating the common personal particular value shared by most of the victim's friends with Algorithm 1.

Algorithm 1 Infer Personal Particular

Require: $U = \{u_1, u_2, ..., u_n\}$; $pp_{u_1}, pp_{u_2}, ..., pp_{u_n}$;
Ensure: pp_{infer}
1: compute $PP = \{pp_1, pp_2, ..., pp_m\}$ from pp_{u_i} for all $i \in \{1, 2, ..., n\}$
2: **for all** $j \in \{1, 2, ..., m\}$ **do**
3: calculate $U_{pp_j} \subseteq U$ such that for all $u \in U_{pp_j}$, $pp_u = pp_j$
4: **end for**
5: **if** there exists U_{pp_t} such that $|U_{pp_t}| > |U_{pp_s}|$ for all $s \in \{1, 2, ..., m\}$ and $t \neq s$ **then**
6: **return** personal particular value pp_t
7: **else**
8: **return** null
9: **end if**

Given the inputs, if Algorithm 1 returns a value pp_{infer} which is equal to the victim's personal particular pp_v, then the victim's personal particular information is leaked to the adversary.

□

To satisfy Necessary Condition 1, the following mitigations are suggested.

Mitigation 1: *If a victim publishes information in her* PP *set and allows a set of receivers to view the information, the privacy rules chosen by the victim should be propagated to all users in the victim's* SR *set who have similar or common information in their* PP *sets.*

Mitigation 2: *A victim should intentionally set up a certain number of connections with other users who have different personal particulars.*

Cross-site incompatibility

If a user publishes personal information in multiple OSNs, she may employ different privacy control rules provided by different OSNs. This causes an inherent exploit named *cross-site incompatibility*.

Exploit 2: *Personal information could be inferred in multiple OSNs if it is protected by incompatible privacy rules in different OSNs.*

The incompatibility of privacy rules in different OSNs is due to: 1) inconsistent privacy rules in different OSNs, 2) different social relationships in different OSNs, and 3) different privacy control mechanisms in different OSNs (e.g. different privacy control granularities). Due to Exploit 2, an adversary may obtain a victim's personal particulars which are hidden from the adversary in one OSN but are shared with the adversary in another OSN. The following is an exemplary attack on Facebook and Google+.

Attack 2: Bob is Alice's friend on both Google+ and Facebook. On Google+, Alice publishes her gender information in her PP set and shares this information with some friends but not including Bob. On Facebook, Alice publishes her gender information and allows all users to view this information because Facebook allows her to share it with either all users or no users. Comparing Alice's personal information published on Facebook and Google+, Bob is able to know Alice's gender published on Facebook which is not supposed to be viewed by Bob on Google+.

Any adversary can perform this attack to infer personal information in a victim's

PP set from multiple OSNs. This exploit can also be used to infer undisclosed information in SR set and SA set. To prevent privacy leakage due to Exploit 2, the following necessary condition needs to be satisfied.

Necessary Condition 2: *Given a set of privacy rules* $\mathsf{PR} = \{pr_1, pr_2, ..., pr_n\}$ *and* $pr_i = (wl_i, bl_i)$ *where* pr_i *is the privacy rule for a victim's personal particular published in* OSN_i*,* wl_i *is a set of all receivers in a white list, and* bl_i *is a set of all receivers in a black list for* $i \in \{1, 2, ..., n\}$*, the following condition holds: for any* $i, j \in \{1, 2, ..., n\}$*,* $wl_i \setminus bl_i = wl_j \setminus bl_j$.[2]

Proof. A victim uses the privacy rules $pr_1, pr_2, ..., pr_n$ to protect her personal particular published in $OSN_1, OSN_2, ..., OSN_n$ respectively where each privacy rule $pr_i = (wl_i, bl_i)$ contains a white list wl_i and a black list bl_i. Assuming there are two privacy rules pr_t and pr_j such that $wl_t \setminus bl_t \neq wl_j \setminus bl_j)$ where $t, j \in \{1, 2, ..., n\}$ and $t \neq j$, we have $U_{diff} = (wl_t \setminus bl_t) \setminus (wl_j \setminus bl_j) \neq \emptyset$. If an adversary $adv \in U_{diff}$, then the victim's personal information is leaked to the adversary although the information is supposed to be hidden from the adversary by pr_j on OSN_j.

□

To satisfy Necessary Condition 2, the following mitigation strategies can be applied.

Mitigation 3: *A victim should share her personal information with the same users in all OSNs.*

Mitigation 4: *If different OSNs provide incompatible privacy control on certain personal information, a victim should choose a privacy rule for this information under two requirements: 1) the privacy rule can be enforced in all OSNs; 2) the privacy rule is at least as rigid as the privacy rules which the victim intends to choose in any OSNs.*

[2]Given a privacy rule $pr = \{wl, bl\}$ with a white list wl and a black list bl, only the receivers who are in white list and are not in black list (i.e. any reciever $u \in wl \setminus bl$) are allowed to view the protected information.

3.5.2 SR Set

A user's SR set records social relationships regarding whom the user knows. The undisclosed information in SR set protected by existing privacy control mechanisms can be inferred by two inherent exploits, namely *inferable social relationship* and *unregulated relationship recommendation*.

Inferable Social Relationship

OSNs provide SR set for a user to store the lists of the users who have connections with him/her. If there exists a connection from Alice to Carl, then Carl is recorded in the outgoing list in Alice's SR set while Alice is recorded in the incoming list in Carl's SR set. The connection between Alice and Carl is stored in both Alice's SR set and Carl's SR set. This causes an inherent exploit named *inferable social relationship*, which corresponds to the information flow shown as dashed arrow (3) in Figure 3.2.

Exploit 3: *Each social relationship in a victim's* SR *set indicates a connection between the victim and another user* u*. User* u*'s* SR *set also stores a copy of this relationship for the same connection. The social relationship in the victim's* SR *set can be inferred from the* SR *set of another user who is in the victim's* SR *set.*

An adversary may use Exploit 3 to obtain undisclosed social relationships in a victim's SR set, which is shown in the following exemplary attack on Facebook.

Attack 3: Figure 3.4 shows a scenario on Facebook, where Bob is a stranger to Alice, and Carl is Alice's friend. Alice shares her SR set with a user group including Carl. Bob guesses Carl may be connected with Alice, but cannot confirm this by viewing Alice's SR set as it is protected against him (who is a stranger to Alice). However, Carl shares his SR set to the public due to different concerns in privacy protection. Seeing Alice in Carl's SR set, Bob infers that Carl is Alice's friend.

Although the adversary is assumed to be a stranger in the above attack, any adversary with stronger capabilities can utilize Exploit 3 to perform the attack as

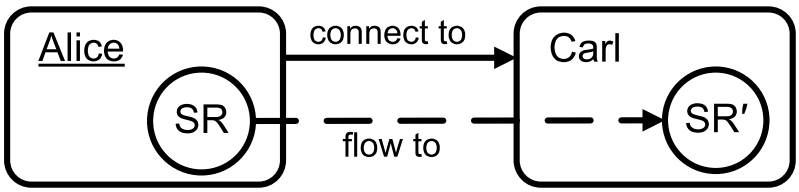


Figure 3.4: Alice's social relationships flow to Carl's SR set

long as he has two types of knowledge: 1) a list of users in the victim's SR set; 2) social relationships in these users' SR sets. This attack could be a stepping stone for an adversary to infiltrate a victim's social network. Once the adversary discovers a victim's friends and establishes connections with them, he becomes a friend of the victim's friends. After that, he has a higher probability to be accepted as the victim's friend, as they have common friends [75]. To prevent privacy leakage caused by Exploit 3, the following necessary condition should be satisfied.

Necessary Condition 3: *Given a victim v's privacy rule $pr_v = (wl_v, bl_v)$ for her* SR *set, a set of all users $U = \{u_1, u_2, ..., u_n\}$ included in the victim's* SR *set in an OSN, and a set of privacy rules* $\mathsf{PR} = \{pr_1, pr_2, ..., pr_n\}$ *where each $pr_i = (wl_i, bl_i)$ is the privacy rule for u_i's* SR *set with white list wl_i and black list bl_i, the following condition holds: for all $i \in \{1, 2, ..., n\}$, $wl_i \setminus bl_i \subseteq wl_v \setminus bl_v$.*

Proof. A victim v sets the privacy rule $pr_v = (wl_v, bl_v)$ for her SR set with white list wl_v and black list bl_v. The victim's SR includes a set of users $U = \{u_1, u_2, ..., u_n\}$. Each user u_i sets the privacy rule $pr_i = (wl_i, bl_i)$ for his/her SR set with white list wl_i and black list bl_i for all $i \in \{1, 2, ..., n\}$. Assuming an adversary adv is not in $wl_v \setminus bl_v$, the adversary is not allowed to view any relationships in the victim's SR set. If there is a privacy rule pr_t such that $wl_t \setminus bl_t$ is not a subset of $wl_v \setminus bl_v$ and $t \in \{1, 2, ..., n\}$, then we have $U_{diff} = (wl_t \setminus bl_t) \setminus (wl_v \setminus bl_v) \neq \emptyset$. Assuming $adv \in U_{diff}$, then the relationship between user u_t and victim v is known by adversary adv although the information in the victim's SR set should be hidden from adv by pr_v.

□

To satisfy Necessary Condition 3, the following mitigation strategy can be ap-

plied.

Mitigation 5: *Let* $U = \{u_1, u_2, ..., u_m\}$ *denote the set of users in a victim's* SR *set. If the victim shares her* SR *set with a set of receivers, then each user* $u_i \in U$ *should share the social relationship between the user and the victim in the user's* SR *set with the same set of receivers only. Since most of existing OSNs use coarse-grained privacy rules to protect social relationships in* SR *set, all users in the victim's* SR *set should share their whole* SR *sets with the same set of receivers chosen by the victim in order to prevent privacy leakage.*

Unregulated Relationship Recommendation

To help a user build more connections, most OSNs provide REC functionality to automatically recommend a list of other users whom this user may know. The recommendation list is usually calculated based on the relationships in SR set but not regulated by the privacy rules chosen by the users in the recommendation list. This causes an inherent exploit named *unregulated relationship recommendation*, which corresponds to the information flow shown as solid arrow (1) in Figure 3.2.

Exploit 4: *All social relationships recorded in a victim's* SR *set could be automatically recommended by* REC *to all users in the victim's* SR *set, irrespective of whether or not the victim uses any privacy rules to protect her* SR *set.*

An adversary may use Exploit 4 to obtain undisclosed social relationships in a victim's SR set, which is shown in the following attack on Facebook.

Attack 4: On Facebook, Bob is a friend of Alice, but not in a user group named `Close_Friends`. Alice shares her SR set with `Close_Friends` only. Although Bob is not allowed to view Alice's social relationships in her SR set, such information is automatically recommended by REC to Bob as "*users he may know*". If Bob is connected with Alice only, the recommendation list consists of the social relationships in Alice's SR set only.

The recommendation list generated by REC may be affected by other factors

such as personal particulars and interests, which may bring noise in social relationships. To minimize such noise, Bob could temporarily delete all his personal particulars and stay connected with the victim only.

The attack may happen on both Facebook and Google+ as long as an adversary is a *friend* of a victim. There is no prior knowledge required for this attack. The attack on Google+ is similar to the attack on Facebook but with a slight difference. On Facebook, the adversary cannot be connected with the victim unless the victim agrees since the relationship is mutual. By contrast, the adversary can set up a connection with the victim on Google+ without getting approval from the victim because the connection is unidirectional. This may make it easier for the adversary to obtain social relationships in the victim's SR set via REC.

We have reported Exploit 4 to Facebook and got confirmation from them. Exploit 4 occurs because REC functionality is implemented in a separate system not regulated by privacy control of Facebook. To prevent privacy leakage due to Exploit 4, the following necessary condition should be satisfied.

Necessary Condition 4: *Given a privacy rule $pr = (wl, bl)$ with white list wl and black list bl for a victim's* SR *set in an OSN and a set of all users U included in the* SR *set, the following condition holds: $U \subseteq wl \setminus bl$.*

Proof. A victim sets a privacy rule $pr_v = (wl_v, bl_v)$ for her SR set with white list wl_v and black list bl_v. The victim's SR includes a set of users $U = \{u_1, u_2, ..., u_n\}$. Assuming that U is not a subset of $wl_v \setminus bl_v$, then we have $U_{diff} = U \setminus (wl_v \setminus bl_v) \neq \emptyset$. If adversary $adv \in U_{diff}$, then REC functionality recommends almost all users in U to adv. Note that these users should be hidden from adv by privacy rule pr_v because adv is not in $wl_v \setminus bl_v$.

□

To satisfy Necessary Condition 4, the following mitigation strategy can be applied.

Mitigation 6: *Let $U = \{u_1, u_2, ..., u_m\}$ denote the set of users in a victim's* SR *set.*

If the victim shares her SR *set with a set of users* $U' \subseteq U$ *only, the victim should remove any users in* $U \setminus U'$ *from her* SR *set in order to mitigate privacy leakage caused by* REC.

3.5.3 SA Set

A user's SA set contains social activities about what the user does. The undisclosed information in SA set protected by existing privacy control mechanisms can be inferred due to the following inherent exploits and implementation defects, including *inferable social activity*, *ineffective rule update*, and *invalid hiding list*.

Inferable Social Activity

If two users are connected in OSNs, a user's name can be mentioned by the other in a social activity via TAG such that this social activity provides a link to the profile page of the mentioned user. Such links create correlations among all the users involved in the same activity. This causes an inherent exploit named *inferable social activity*, which corresponds to the information flow shown as solid arrow (2) in Figure 3.2.

Exploit 5: *If a victim's friend uses* TAG *to mention the victim in a social activity published by the victim's friend, it implies that the victim may also attend the activity, which is indicated by the link created by* TAG *pointing to the victim's profile page. Although this activity may involve the victim, the visibility of this activity is solely determined by the privacy rules specified by the victim's friend who publishes the activity, which is out of the control of the victim.*

An adversary may use Exploit 5 to obtain undisclosed social activities in a victim's SA set, which is shown in the following attack on Facebook.

Attack 5: Figure 3.5 shows a scenario on Facebook, where Bob and Carl are Alice's friends, and Bob is Carl's friend. Alice publishes a social activity in her SA set regarding a party which Carl and she attended together and she allows Carl only to

view this social activity. However, Carl publishes the same social activity in his SA set and mentions Alice via TAG. Due to different concerns in privacy protection, Carl allows all his friends to view this social activity. By viewing Carl's social activity, Bob can infer that Alice attended this party.

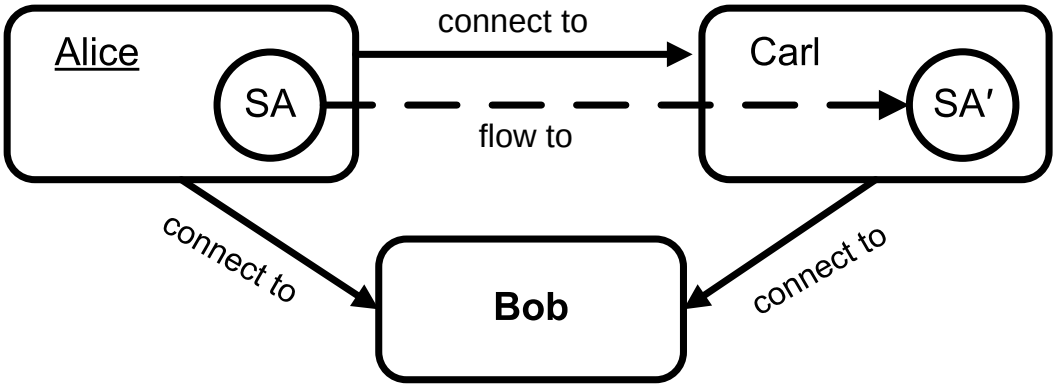


Figure 3.5: Alice's social activities flow to Carl's SA set

This attack works on Facebook, Google+, and Twitter. Any adversary can perform this attack if he knows the social activities published by the victim's friends pointing to the victim via TAG. To prevent privacy leakage due to Exploit 5, the following necessary condition should be satisfied.

Necessary Condition 5: *Given a privacy rule* $pr_u = (wl_u, bl_u)$ *for an activity where a victim* v *is tagged by her friend* u *in an OSN and* v*'s intended privacy rule* $pr_v = (wl_v, bl_v)$ *for the activity, the following condition holds:* $wl_u \setminus bl_u \subseteq wl_v \setminus bl_v$.

Proof. Given a privacy rule $pr_u = (wl_u, bl_u)$ for an activity with white list wl_u and black list bl_u where victim v is mentioned by her friend u, any receivers in $wl_u \setminus bl_u$ are allowed to view the activity. We assume that v's intended privacy rule for the activity is $pr_v = (wl_v, bl_v)$ with white list wl_v and black list bl_v. If $wl_u \setminus bl_u$ is not a subset of $wl_v \setminus bl_v$, then we have $U_{diff} = (wl_u \setminus bl_u) \setminus (wl_v \setminus bl_v) \neq \emptyset$. Assuming $adv \in U_{diff}$, then adv can obtain the activity published by u although the victim's privacy rule pr_v prevents adv from viewing the activity.

□

To satisfy Necessary Condition 5, the following mitigation strategy can be applied.

Mitigation 7: *If a victim is mentioned in a social activity in another user's* SA *via* TAG*, the victim should be able to specify additional privacy rules to address her privacy concerns even when the social activity is not in her profile page.*

Ineffective Rule Update

It is common in OSNs that users regret sharing their social activities with wrong audience. Typical reasons include being in state of high emotion or under influence of alcohol [73]. It is necessary to allow users to correct their mistakes by revoking the access rights of those unwanted audience. Once the access right of viewing a particular social activity is revoked, a receiver should not be able to view the activity protected by the updated privacy rule. On Facebook, a user can remove a receiver from the local white list specifying who is allowed to view a social activity or add the receiver to the local black list for the activity. Google+ and Twitter currently do not provide local black lists for individual social activities. A user may remove a receiver from the white list or from a user group if the user group is used to specify the scope of the white list (e.g. sharing a social activity within a *circle* on Google+). However, if a user's social activity has been pushed to her subscribers' feed pages, the update of privacy rules on Google+ and Twitter does not apply to this social activity in feed pages. This causes an implementation defect named *ineffective rule update*.

Exploit 6: *Once a victim publishes a social activity, the social activity is immediately pushed to the feed pages of the victim's subscribers who are allowed to view the social activity according to the victim's privacy rule. Later, even after the victim changes the privacy rule for this activity to disallow a subscriber to view this activity, the social activity still appears in this subscriber's feed pages on Google+ and Twitter. The current implementation of Google+ and Twitter enforces a privacy rule*

only when a social activity is published and pushed to corresponding subscribers' feed pages. Updated privacy rules are not applied to the activities which have already been pushed to feed pages (see Figure 3.6).

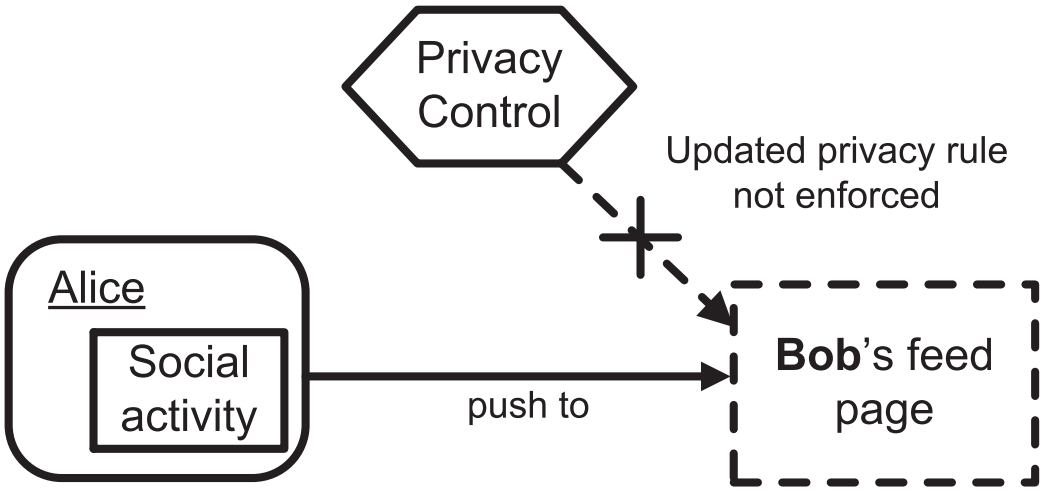


Figure 3.6: Privacy control doesn't enforce the updated privacy rule to a social activity that has been pushed to a feed page.

An adversary may use Exploit 6 to obtain undisclosed social activities in a victim's SA set without the victim's awareness. Below shows a typical attack on Google+.

Attack 6: On Google+, Bob is Alice's friend and subscriber. Alice publishes a social activity and allows her friends in group `Classmate` only to view the activity. Alice assigned Bob to the group `Classmate` by mistake and realized this mistake after publishing the activity. Then, Alice removed Bob from the group. However, Bob can still view this social activity as it has already been pushed to his feed page.

The above attack can happen on Google+ and Twitter. To perform the attack, an adversary should be the victim's *friend* and *subscriber*. The attack doesn't work on Facebook as privacy control in Facebook always actively examines whether privacy rule for a social activity is updated. If a privacy rule is updated, the privacy control is immediately applied to the social activity in corresponding feed pages. Consequently, the social activity is removed from the feed pages. To prevent this attack in certain OSNs such as Google+ and Twitter, the following mitigation strategy can be

applied.

Mitigation 8: *If a victim mistakenly shares a social activity with an unintended receiver, instead of changing the privacy rules, the victim should delete the social activity as soon as possible so that the social activity is removed from all feed pages.*

Note that Mitigation 8 is not effective unless the deletion of the social activity takes place before an adversary views the social activity. If the adversary views the social activity before it is deleted, the adversary could keep a copy of this activity, which cannot be prevented.

Invalid Hiding List

To support flexible privacy control, many OSNs enable users to use black lists so as to hide information from specific receivers. On Facebook, a local black list is called *hiding list*. Using hiding list, a user may apply fine-grained privacy control on various types of personal information. However, the hiding lists take no effect except for the user's friends. This causes an implementation defect named *invalid hiding list*.

Exploit 7: *In certain OSN, a victim may include some of her friends in hiding lists to protect her personal information. However, when a friend breaks his relationship with the victim, the OSN automatically removes him from the hiding lists as the friend relationship terminates. Releasing from hiding lists, this former friend is allowed to view the victim's protected information if he is not restricted by other privacy rules.*

The implementation defect behind this exploit creates a false impression on the effectiveness of hiding lists. An adversary may use Exploit 7 to obtain undisclosed social activities in a victim's SA set without the victim's awareness. A typical attack on Facebook is given below.

Attack 7: On Facebook, Bob and Carl are Alice's friends. Bob is Carl's friend, which means Bob is also a friend of Alice's friend. Alice publishes a social activity

which allows her friends and her friends-of-friends to view, except that Bob is added to the hiding list of this activity. Although Bob cannot view this activity under the current privacy rule, he can break his connection with Alice. Then, he is automatically removed from the hiding list. After that, Bob is able to view the undisclosed activity since he is a friend of Alice's friend.

Note that this attack does not work on Google+ and Twitter because their current privacy control mechanisms do not support any local black lists. Also note Exploit 7 can be exploited to target at not only SA set, but also PP set and SR set.

We have reported Exploit 7 to Facebook and received a confirmation from them[3]. To prevent this attack in affected OSNs such as Facebook, the following mitigation strategy can be applied.

Mitigation 9: *A victim should avoid using hiding lists when protecting personal information. Instead, a victim may use white lists or global black lists in forming privacy rules.*

3.6 Feasibility Analysis of the Attacks

The personal information in OSNs could be leaked to adversaries who acquire necessary capabilities to perform the attacks, which have been discussed in Section 3.5. The success of the attacks can be affected by users' behaviors in OSNs. To evaluate the feasibility of these attacks, we conducted an online survey and collected users' usage data on Facebook, Google+, and Twitter. In this section, we first describe the design of the online survey. We then present the demographic data collected in the survey. Based on the survey results, we analyze how widely users' personal information in OSNs could be leaked to adversaries through the corresponding attacks.

[3]Exploit 7 has been fixed by Facebook in 2013.

3.6.1 Methodology

The participants to our online survey are mainly recruited from undergraduate students in our university. We mainly focus on young students in our survey because they are active users of OSNs. Our study shows that they are particularly vulnerable to the privacy attacks. Each participant uses at least one OSN among Facebook, Google+, and Twitter.

The survey questionnaire consists of four sections including 37 questions in total. In the first section, we gave an initial set of demographic questions and a set of general questions such as participants' awareness on privacy and what OSNs (i.e. Facebook, Google+, and Twitter) they use. All the participants need to answer the questions in the first section. In the following three sections, questions about participants' knowledge and privacy attitude towards Facebook, Google+, and Twitter are raised, respectively. Each participant only needs to answer the questions which are relevant to them in these three sections.

3.6.2 Demographics

There are 97 participants in total, among which 60 participants reported being male, and 37 reported female. Our participants' age ranges from 18 to 31, with an average of 22.7.

All of the 97 participants are Facebook users, among whom 95 participants have been using Facebook for more than 1 year, and 2 have been using Facebook for less than 1 month. About a half participants (41/97) are Google+ users, among whom 23 participants have been using Google+ for more than 1 year, 13 have been using Google+ for about 1 month - 1 year, and 5 have been using Google+ for less than 1 month. Similarly, about a half participants (40/97) are Twitter users, among whom 36 participants have been using Twitter for more than 1 year, 3 have been using Twitter for about 1 month - 1 year, and 1 has been using Twitter for less than 1 month.

3.6.3 Attacks to PP Set

To obtain the undisclosed personal information in a victim's PP set, adversaries could exploit the inferable personal particulars and cross-site incompatibility to launch two corresponding attacks as discussed below.

Inferable Personal Particulars

As discussed in Section 3.5.1, due to inferable personal particular (Exploit 1), a victim and most of his/her friends may share common or similar personal particulars. Our study results show that 71% of the Facebook users are connected with their classmates on Facebook; 78% of the Google+ users are connected with their classmates on Google+; and 73% of the Twitter users are connected with their classmates on Twitter.

Via Exploit 1, an adversary could perform Attack 1 and infer a victim's personal particular from the personal particulars shared by most of her friends. To perform Attack 1, two types of knowledge are required: a large portion of users stored in the victim's SR set and their personal particulars.

The protection of the victim's SR set could help prevent the adversary from obtaining the victim's relationships. Unfortunately, our study shows that 22% of the Facebook users, 39% of the Google+ users, and 35% of the Twitter users choose the "*Public*" privacy rule or the default privacy rule[4] for their social relationships, which means that these users share their social relationships with the public. Moreover, the OSNs users may connect to strangers. According to our study, 60% of the Facebook users, 27% of the Google+ users, and 30% of the Twitter users have set up connections with strangers, which leave their SR set information vulnerable to Exploit 4 (unregulated relationship recommendation) as discussed in Section 3.5.2.

The privacy rules for personal particulars of the victim's friends can be set to prevent the adversary from obtaining the second type of knowledge required in

[4]Facebook, Google+, and Twitter set "*Public*" as default privacy rule for the SR set of each user

Attack 1. However, the victim's personal particulars can be exposed to threats if his/her friends publicly share their personal particulars. In our study, 43% of the Facebook users, 44% of the Google+ users, and 48% of the Twitter users share their personal particular publicly because they choose the "*Public*" privacy rule or the default privacy rule[5].

Cross-site Incompatibility

Users may use multiple OSNs at the same time. According to our survey, 54 out of 97 participants use at least two OSNs as shown in Figure 3.7. And 27 participants publish their posts in more than one OSN at the same time as shown in Figure 3.8. If a user publishes personal information in multiple OSNs, he/she may set different privacy control rules vulnerable to Exploit 2, i.e. cross-site incompatibility.

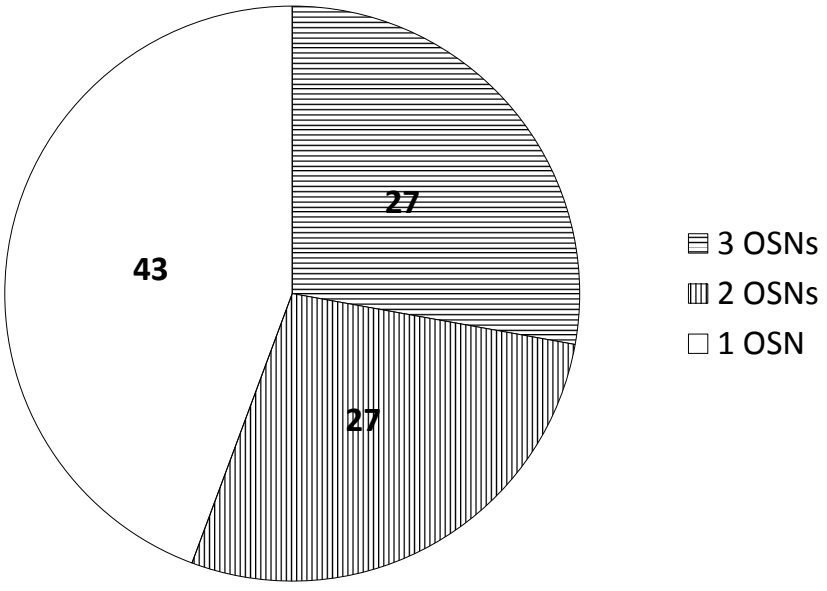


Figure 3.7: Participants' usage of multiple OSNs

Due to Exploit 2, an adversary can perform Attack 2 if the victim shares her

[5]Facebook, Google+, and Twitter set "*Public*" as the default privacy rule for each user's personal particulars such as "university" information

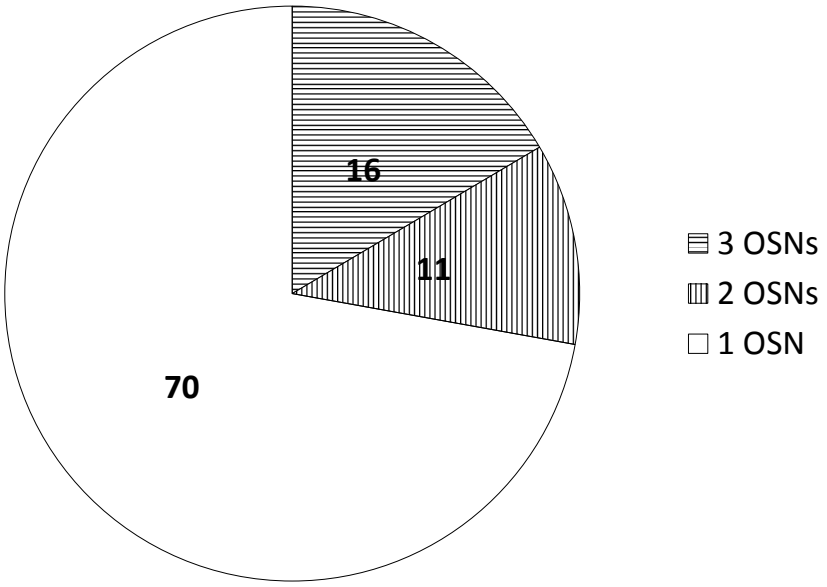


Figure 3.8: Participants' publishing posts in multiple OSNs

personal information with the adversary in any OSN site. This attack is due to three reasons.

The first reason is that users employ inconsistent privacy rules in different OSNs. The results of our study show that 27 out of 97 participants use inconsistent privacy rules to protect their gender information, 25 participants use inconsistent privacy rules to protect their university information, and 21 participants use inconsistent privacy rules to protect their political view information.

The second reason is that users maintain different social relationships in different OSNs. According to the study, 59 out of 97 participants reported that their social relationships on Facebook, Google+, and Twitter are different. Therefore, even though users protect their information by the same privacy rules on multiple OSNs, an adversary can still obtain their information if he can exploit this vulnerability.

The third reason is the difference between privacy control mechanisms in different OSNs. The protection of gender information is a typical example which is discussed in Section 3.5.1.

3.6.4 Attacks to SR Set

Adversaries could obtain social relationships in a victim's SR set through two exploits, which are inferable social relationship and unregulated recommendation.

Inferable Social Relationship

Inferable social relationship (Exploit 3) is caused by the storage format of social relationships in SR set as explained in Section 3.5.2. If two users set up a relationship with each other, then each of them stores a copy of the relationship in his/her SR set and choose a privacy rule to protect his/her SR set.

Via Exploit 3, an adversary could perform Attack 3 given two types of knowledge, including a list of users in the victim's SR set and the social relationships in these users' SR set. Therefore, the protection of the social relationships in the victim's SR set depends on the privacy rules for the SR sets of the users in the victim's SR set. Unfortunately, as mentioned in 3.6.3, 22% of the Facebook users, 39% of the Google+ users, and 35% of the Twitters share their SR sets publicly. These users reveal social relationships with their friends publicly regardless of the privacy rules for their friends' SR sets.

Unregulated Relationship Recommendation

REC functionality helps users establish more social relationships. According to our study, 71 out 97 Facebook users, 21 out of 41 Google+ users, and 17 out of 40 Twitter users have used REC functionality in OSNs. Unregulated relationship recommendation (Exploit 4) could leak all social relationships in a user's SR set due to automatic relationship recommendation of REC.

By Exploit 4, an adversary can perform Attack 4 to obtain all social relationships in a victim's SR set on Facebook or Google+ if the adversary manages to become a "friend" of the victim.

As shown in Figure 3.9, 4% of the Facebook users and 7% of the Google+

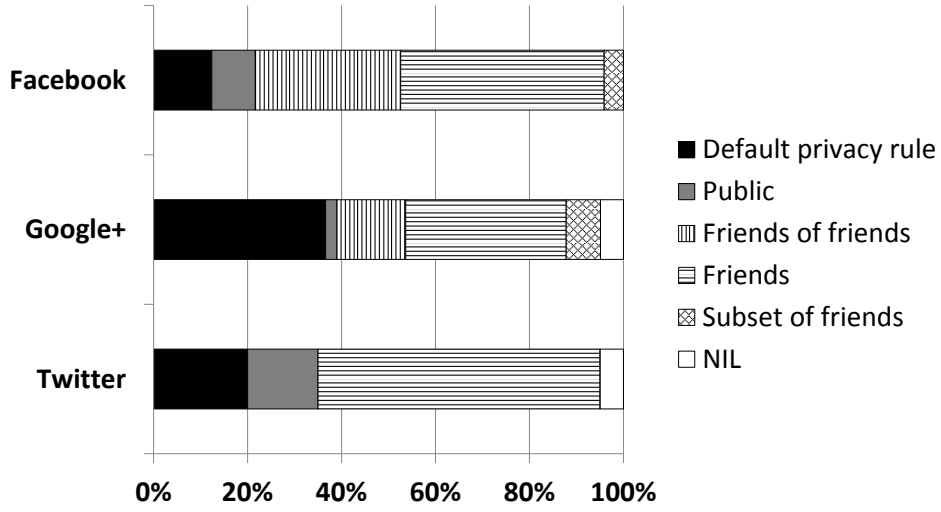


Figure 3.9: Privacy rules for participants' SR sets in OSNs

users choose to share their SR set with a proper subset of their friends[6]. Exploit 4 explicitly violates these users' privacy rules.

Although most of the Facebook users and the Google+ users share their SR sets with friends, friends of friends, or public, their selection of privacy rules may contradict their privacy attitude.

In Figure 3.9, 53% of the Facebook users share their SR sets with friends of friends or publicly[7]. Among the Facebook users who share their SR sets with friends of friends or public, 88% of them address concerns about their social relationships being revealed to others whom they don't know.

Among the Google+ users, 36% of them share their SR sets with friends of friends or the public. However, 71% of the Google+ users who share their SR sets with friends of friends or the public are not willing to reveal their social relationships to strangers.

As shown in our survey, 43% of the Facebook users and 20% of the Google+ users have concerns about revealing their social relationships to strangers but ever

[6] An empty subset corresponds to the privacy rule "*Only me*".

[7] The "*Public*" privacy rule and the default privacy rule lead to sharing SR set publicly

including strangers to their SR sets. This may leak the users' social relationships to the strangers irrespective of any privacy rules chosen to protect their SR sets.

3.6.5 Attacks to SA Set

To obtain social activity information in a victim's SA set, adversaries could perform 3 attacks due to three exploits including inferable social activity, ineffective rule update, and invalid hiding list.

Inferable Social Activity

In OSNs, if a user is mentioned in his/her friends' social activity via TAG, the privacy rule for the activity is determined by the friends and out of this user's control. This leads to inferable social activity (Exploit 5).

Via Exploit 5, an adversary may infer a victim's social activities from the victim's friends' SA set. As shown in Figure 3.10, 99% of the Facebook users, 44% of the Google+ users, and 78% of the Twitter users have experience of being tagged in activities. On the other hand, 36% of the Facebook users, 34% of the Google+ users, and 40% of the Twitter users have concerns about being tagged in certain activities published by their friends without any negotiations. Since their friends determine the visibility of the activities, these users can inform their friends of their concerns. Our results show that 82% of the Facebook users, 73% of the Google+ users, and 73% of the Twitter users will inform their friends of their concerns if they don't agree on being tagged by their friends. The rest of them keep silent even though their privacy could be violated.

Ineffective Rule Update

As discussed in Section 3.5.3, if a user changes his/her privacy rules for social activities, the updated privacy rules do not apply to the activities which have been pushed to the feed pages of the user's subscribers. This is named as ineffective rule

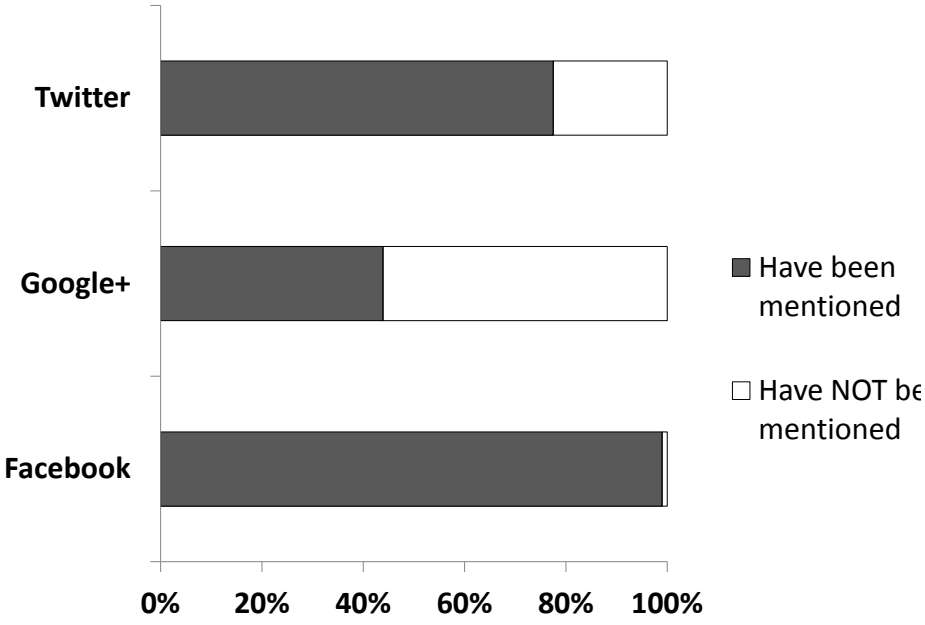


Figure 3.10: Participants being mentioned in OSNs

update (Exploit 6).

Via Exploit 6, an adversary could perform Attack 6 on Google+ and Twitter and obtain a victim's activities which are shared with the adversary before privacy rules update.

Changing privacy rules may occur if users regret publishing their activities. According to our study, 15% of Google+ users and 15% of Twitter users have experience of regretting publishing their posts. As shown in Figure 3.11, 20% of the Google+ users choose to change their privacy rules if they regret sharing activities, while 38% of the Twitter users choose to change their privacy rules by turning on the protect my tweets option if they regret sharing such activities.

To mitigate Exploit 6, users may delete the activities they regret sharing as soon as possible. We found that 61% of the Google+ users and 23% of the Twitter users choose to do so.

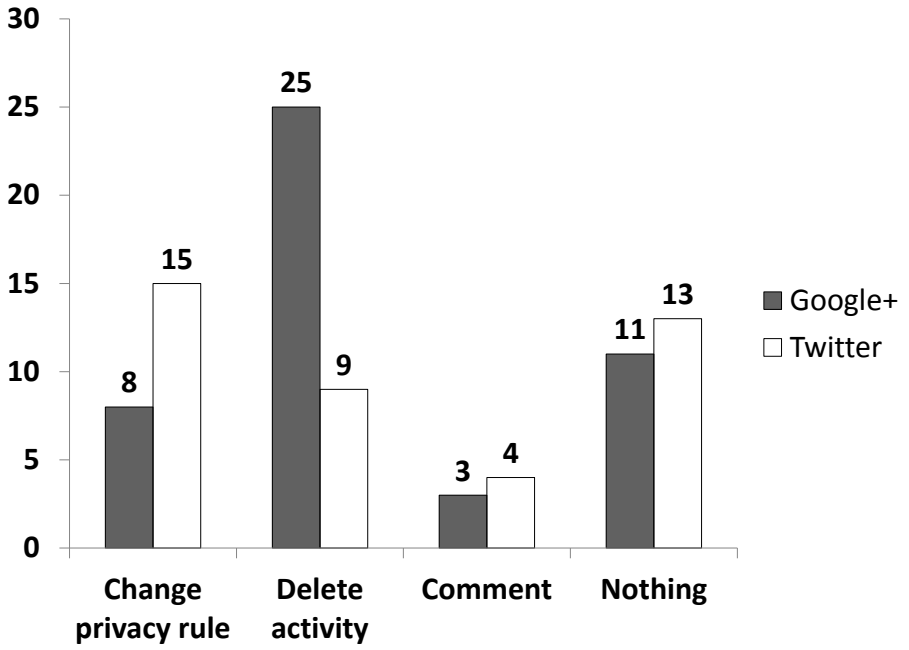


Figure 3.11: Participants' actions if regretting sharing activities

Invalid Hiding List

On Facebook, if a user protects his/her social activity by using a hiding list including the user's friends, these friends will be automatically removed from the hiding list after they terminate their relationships with the user. This is referred to as the invalid hiding list (Exploit 7).

Via Exploit 7, an adversary could perform Attack 7 to obtain a victim's social activities if the victim uses the "*friends of friends*" privacy rule with a hiding list containing the adversary. Our study shows that 54% of the Facebook users have ever used the "*friends of friends*" privacy rule with a hiding list that includes their friends when they publish activities. To evaluate the awareness of the risks caused by using the invalid hiding list, we summarized participants' confidence level regarding whether their activities are hidden from their friends who are included in their hiding lists on Facebook. As shown in Figure 3.12, 31% (30 out of 97) of the

Facebook users feel confident in the effectiveness of the hiding list on Facebook. If attack 7 happens, these participants may misunderstand the validity of the hiding lists and still believe that their activities are hidden from their friends included in the hiding lists.

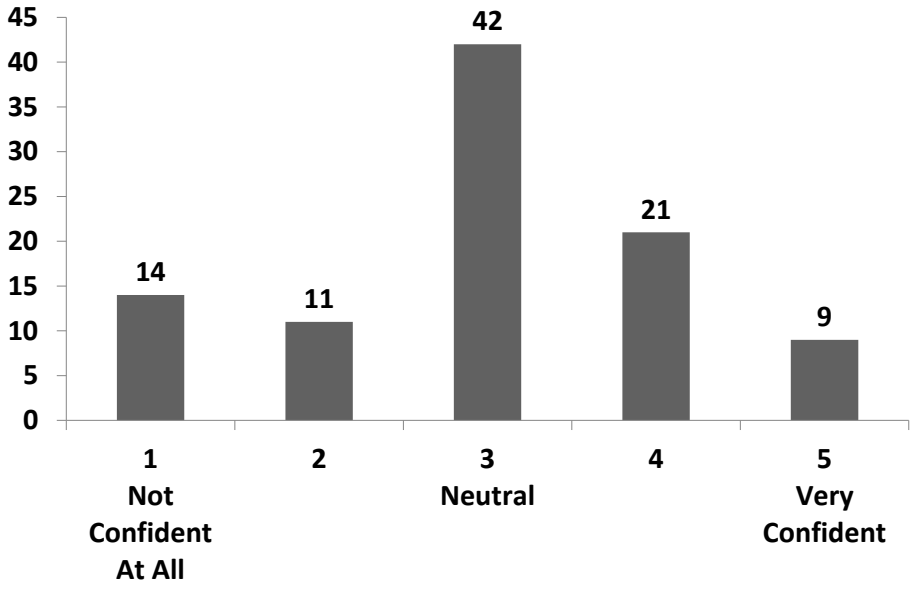


Figure 3.12: Users' confidence in validity of Facebook hiding list

3.7 Implications of Our Findings

On the surface, our exploits are caused by the inconsistencies between privacy control and functionalities of OSN. In fact, these inconsistencies reflect the conflicts between people's intention on privacy protection and social/business values of OSNs. We discuss the implications of these conflicts in this section.

Most of the functionalities involved in our exploits are essential in OSNs. These functionalities deal with personal particulars, social relationships, and social activities. While the social values of these functionalities should be preserved from a user's perspective, they are restricted due to privacy controls.

First, exhibiting personal particulars is an important feature for social recognition. Most OSNs encourage users to share genuine information about their personal particulars in order to foster trust and respect in OSNs [22]. This would help users discover new relationships with those who have similar interests. This is explained by the homophily theory [52, 13], which states that a human being is more willing to interact with others who have similar personal particulars such as race, organization, and education. Meanwhile, the implicit connections among users may be exploited to infer undisclosed personal particulars. To mitigate this threat, **mitigations** 1 and 2 require users to connect with other dissimilar users which they may not even like.

Second, maintaining and expanding social relationships is one of the major benefits of OSNs. As socially-oriented beings, humans have a desire to stay connected so that they have a sense of communion with others [62]. This desire is addressed in OSNs with the relationship list and the recommendation function. Although the public display of a user's relationship list may disclose certain private information, it also helps build more connections in OSNs. If a user's profile contains a large number of connections, it brings satisfactory social recognition for the user [33]. The recommendation function further makes it easier to establish new connections based on relationship lists and other information. This is especially important for new users to make friends in OSNs. The current recommendation function operates according to the small-world theory [75], which states that two connected users are likely to have common friends who have not yet recorded in their current relationship lists. This function can also be exploited by an adversary to enumerate all social relationships of a victim. To mitigate the privacy leakage about social relationships, a user may use **mitigations** 5 and 6. The consequences of applying these mitigation strategies are: 1) If a user sets up a strict privacy rule on his relationship list, this rule should propagate to all users in his relationship list. 2) The effectiveness of the recommendation function would be significantly influenced by such mitigations.

Third, sharing social activities is an important part of human social life. Human beings are curious about what happen around them. They would like to under-

stand the surrounding environment by knowing how other people behave, think, and feel [58]. OSNs enable users to receive the activities published by other users to cure such curiosity. On the other hand, users who publish activities feel rewarded due to attentions of other users, which is usually interpreted as a sign for social recognition [35]. Since a social activity usually involves multiple users, sharing this activity may conflict with the privacy concern of these users. In order to mitigate this threat, the scope of privacy control in OSNs should be extended as mentioned in **mitigation** 7, which enforces privacy control to an activity no matter who publishes it. However, this may frustrate users who intend to share that activity, and might be difficult to achieve due to the incompatibility among privacy control mechanisms in different OSNs. As suggested in **mitigations** 3 and 4, a user may choose a strict privacy rule so as to achieve his privacy objective. However, this may significantly restrict the sharing nature of OSNs.

While OSN users are concerned with the social values of OSN functionalities, OSN service providers are more concerned with business values. As a company, the first priority of an OSN service provider is to generate revenue. However, most existing OSN service providers do not charge their users. As Andrew Lewis pointed out, "If you're not paying for something, you're not the customer; you're the product being sold." This is exactly what OSN service providers do, monetizing user-generated contents by maintaining an OSN-based ecosystem. As one of the most successful OSN-based business models, targeted advertising [65], usually demands large number of connected individuals and high quality of personal information such as location and photos to differentiate each individuals [22, 10, 18, 77, 19]. Thus, an OSN service provider has strong incentive to encourage users to generate and to share personal information, and also to connect to more users so as to attract more people to join in the OSNs [36].

The business values could be significantly degraded by users' privacy concerns. Though users are willing to expose their personal information to intended audience as analyzed above, these social values may be corrupted if the shared information is

revealed to the unwanted audience. For example, if an intimate message is disclosed to a person who is not supposed to know, the sender of the message may even suffer from negative emotion. Therefore, the privacy concerns limit the spread of user-generated contents. Almost all mitigations discussed in the paper add additional restrictions on user generated contents published or shared in OSNs.

These inconsistencies may explain why the effectiveness of privacy control is limited in existing OSNs. This limitation will not be easily resolved if the social and business factors behind are not dealt properly.

Chapter 4

Understanding OSN-Based Facial Disclosure against Face Authentication Systems

4.1 Introduction

This chapter analyze the online social network based facial disclosure threats against face authentication systems. As the platforms for experience sharing and social interaction, numerous personal images are being published in OSNs such as Facebook, Google+, and Instagram at every moment. According to a recent report by Facebook, 350 million personal images are published by users on Facebook every day [72]. It is very likely that these images contain facial images where the users' faces can be clearly seen. The large base number indicates that these shared personal images could become an abundant resource for potential attackers to exploit, which introduces the threat of OSN-based facial disclosure (OSNFD).

OSNFD may have a significant impact on the current face authentication systems, which is one of promising biometrics-based user authentication mechanisms. Face authentication have been widely available on all kinds of consumer-level computing devices such as smartphones, tablets, and laptops with built-in camera ca-

pability. Popular face authentication systems include Face Unlock [27], Facelock Pro [23], and Visidon [70] on smartphones/tablets, Veriface [47], Luxand Blink [51], and FastAccess [71] on laptops. These systems provide attractive alternatives of legacy passwords, as face authentication requires zero memory efforts from users and usually has higher entropy than legacy password as users tend to choose easy-to-guess passwords [55]. Previously, the major obstacle for an adversary to compromise face authentication is that physical proximity is required to capture a victim's facial images. However, this is no longer necessary since the appearance of OSNFD. OSNFD provides abundant exploitable resources affecting the applicability of face authentication as it compromises its confidentiality, which is one of fundamental requirements for authentication [32, 40]. The facial images used for face authentication are no longer secrets and can be disclosed in large scale due to OSNFD.

In this paper, we make the first attempt to provide a quantitative measurement on the threat of OSNFD against face authentication. We investigate real-world face-authentication systems designed for both smartphones, tablets, and laptops. These systems recognize users by analyzing facial images captured by built-in cameras. Our study collects users' facial images published in OSNs and uses them to simulate the spoofing attacks against these systems. Since all target systems including Google's Face Unlock [27, 23, 70, 47, 51, 71] are closed-source and do not provide any programmable testing interfaces, enormous efforts are made for image collection and testing. We also build a dataset containing important image attributes that are common in real-life photos but rarely used in prior controlled study on face authentication [16, 30].

Our study reveals interesting results indicating that face authentication may not be suitable to use as an authentication factor. Although the percentage of vulnerable images that can be used for spoofing attacks is moderate, the percentage of vulnerable users that are subject to spoofing attacks is high. On average, the percentage of vulnerable users is 64% for laptop-based systems, and 93% for smartphone/tablet-

based systems. Our results also show the difference between systems designed for smartphones/tablets and laptops, as smartphones/tablets have to be accessible in more varied environments. Further investigation shows the quality of images is a more important factor affecting the success rate of spoofing attacks compared to quantity. A user who uploads a few clear facial images is more vulnerable than another user who uploads much more facial images of lower quality due to makeup, illumination, or other negative effects. All these findings show that OSNFD has significantly compromised the confidentiality of face authentication.

In order to understand more detailed characteristics of OSNFD, we further develop a risk estimation tool based on our dataset. Logistic regression is used to extract key attributes affecting the success rate of spoofing attacks. It achieves a precision of 81%, a recall of 83%, and an F1 score of 82% on average. It can help users evaluate the risk of uploading an image by calculating a risk score based on the extracted attributes, which makes them aware of the threat of OSNFD.

The contributions of this paper are summarized as follows:

- We investigate the threat of OSN-based face disclosure (OSNFD) against face authentication. Our results suggest that face authentication may not be suitable to use as an authentication factor, as its confidentiality has been significantly compromised by OSNFD.

- We make the first attempt to quantitatively measure the threat of OSNFD by testing real-world face authentication systems designed for smartphones, tablets, and laptops. We also build a dataset containing important image attributes that significantly affect the success rate of spoofing attacks. These attributes are common in real-life photos but rarely used in prior controlled study on face authentication [16, 30].

- We use logistic regression to extract key attributes that affect the success rate of spoofing attacks. These attributes are further used to develop a risk estimation tool to help users measure the risk score of uploading images to OSNs.

4.2 Preliminaries

4.2.1 Face Authentication

Face authentication is a biometrics-based user authentication mechanism, which verifies a user's identity by using information extracted from the user's facial features. As illustrated in Figure 4.1, a typical face authentication system uses a camera to capture the user's facial image/video as input, and then verifies it with enrolled biometric information for the claimed identity. The objective of a face authentication system is to recognize a user as long as the input is collected from the legitimate user, while rejecting the inputs from all other users.

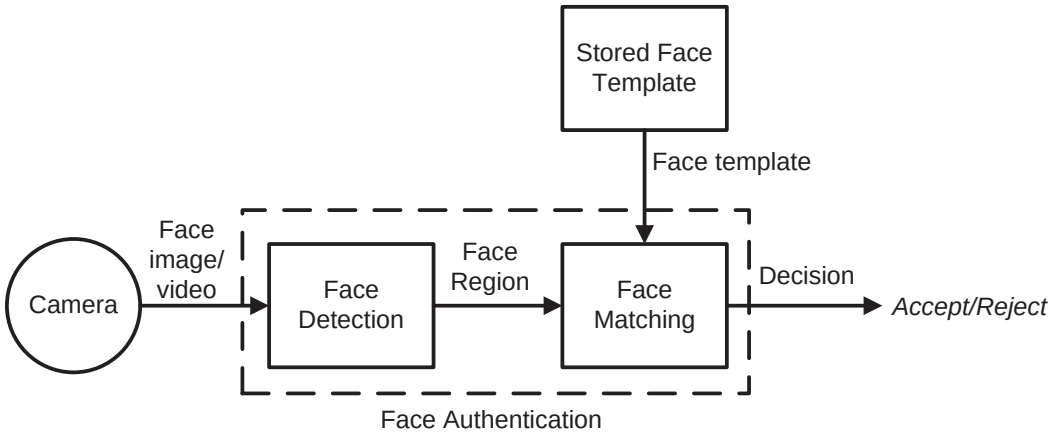


Figure 4.1: Work flow of a typical face authentication system

Two key modules are involved in this verification process. The first module is the face detection module, which identifies the face region and removes irrelevant information of an image. The processed image is then passed to the next module named face matching. This module computes a similarity score for the input image based on an enrolled face template containing key features which can be used to distinguish a user from other users and imposters. Different algorithms may be used for these two modules, but all face authentication systems generally have these two modules and follow this work flow. In the end, a face authentication system outputs the final decision (i.e. accepting or rejecting a claim) according to whether or not the similarity score is higher than a matching threshold. This threshold is

carefully chosen so as to achieve a proper balance between false rejection rate and false acceptance rate.

4.2.2 OSN-based Facial Disclosure and Threat Model

The OSN-based facial disclosure (OSNFD) addresses the issue when users' face biometrics are involuntarily disclosed by sharing personal images in OSNs. These disclosed face biometrics would raise security risks against face authentication systems.

It is a well-known limitation of face authentication that it is subject to spoofing attacks based on captured face biometrics, where an adversary attempts to circumvent user authentication by replaying a victim's facial images/videos collected at an early time. As shown in Figure 4.1, a face authentication system is not expected to tell whether an input image is from a live user or from a captured image/video, as they are all valid inputs from a legitimate user collected at different times. Nevertheless, the impact of these attacks was believed to be limited due to the requirement that an adversary had to be physically close to a victim in order to collect the required information. Therefore, it is generally considered sufficiently secure as an authentication factor for common access protection [8], as we observe that many face authentication systems [27, 23, 70, 47, 51, 71] such as Google's Face Unlock and Lenovo's Veriface, are widely available on all kinds of consumer-level computing devices. Considering its zero-memory requirement, it does provide an attractive alternative for legacy passwords.

However, this belief may be questionable since OSNFD becomes a common phenomenon. OSNFD supplies an adversary with abundant facial images to exploit and makes large-scale identity theft possible for those who use face authentication. Our work investigates the OSNFD threat and quantitatively measures its impacts. We consider OSNFD-based attacks where an adversary attempts to forge a valid input from image resources disclosed from OSNFD so as to pass face authentication.

Our study focuses on image-based attacks unless explicitly mentioned.

The OSNFD threat may be mitigated with liveness detection technologies, which rely on extra information sources or heuristic algorithms to distinguish a live user from a captured image/video. All the existing sophisticated liveness detection technologies associate with considerable costs, which will be explained later in Section 4.5.2. This may explain that only weak liveness detection technologies are currently deployed on the face authentication systems designed for consumer-level computing devices [56, 42]. For example, eye blinking detection is a common heuristic used by many face authentication systems [27, 70, 56] including Google's Face Unlock; however, it can be easily bypassed using two facial images as demonstrated in [59]. Similar tricks can also apply to other weak liveness detection mechanisms such as head rotation detection [56, 59]. Even worse is that the existing liveness detection mechanisms are disabled by default in most popular face authentication systems [27, 23, 70, 47, 51, 71], as they may have negative impacts on accessibility.

4.3 Data Collection and Empirical Analysis

In order to quantitatively measure the impacts of OSNFD, we conduct a user study to collect real personal images that have been shared in OSNs. The collected images are used to test against real-world face-authentication systems chosen from the most popular face authentication products in terms of user base [69, 28]. This section describes the detailed process of data collection and the results of our empirical analysis. We use the following classifications in our discussion.

First, we classify the security settings of a face authentication system into *low* and *high*. Most of face authentication products [27, 23, 70, 47, 51, 71] provide very limited choices on security settings that generally affect the recognition threshold used in the face matching module. For example, Google's Face Unlock [27] does not provide any option for users to adjust its security strength. Most of our tested

products [23, 70, 47, 51, 71] only have two options for users, labeled as "high accessibility" (i.e. low security) and "high security". Only Lenovo's Veriface [47] provides a scrollbar for users to adjust its security strength from the lowest to the highest. Therefore, we use "low" to indicate that a target system enforces the weakest security protection, and use "high" to indicate the strongest security protection achievable to the system.

Second, we classify face authentication systems into *mobile* and *traditional*. A system is labeled as mobile if it is used for smartphones or tablets, while a traditional system is used for laptops or desktops. A mobile system is usually more tolerant to varied environments, as it should be accessible no matter where a user uses the device. Laptops is considered as traditional as it is not expected to be used from anywhere at any time like what users expect smartphones and tablets.

Third, we classify users into different groups according to the pattern of their sharing behaviors. As observed in our study, it is quite common that a user tends to upload edited images where facial landmarks are significant changed to create better visual appeal. Therefore, it is also an important factor that needs to be considered.

These classifications represent three major factors that affect the effectiveness of OSNFD-based attacks, which are security settings, target platforms, and user behaviors, respectively. We use them as controlled parameters to evaluate the severity of OSNFD, and more sophisticated statistical analysis will be given in the next section to identify the key attributes that can be used to mitigate the OSNFD threat.

4.3.1 Data Collection

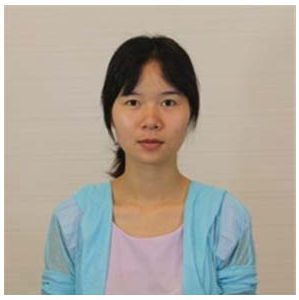
There are 74 participants involved in our study including 36 males and 38 females with age range between 19 and 35. Most of these participants are students in our university. Each participant is paid with 10 dollars as compensation. The study is conducted in a quiet room. The study consists of three parts. In the first part, we ask each participant to select and download 20 *facial* images published within the last

12 months in popular OSNs such as Facebook, Google+, Instagram, etc. To ensure that these images are from OSNs, the participants need to randomly choose 5 – 8 images and show us the same images in their OSN profile pages. A facial image is defined as an image where a participant's face can be seen. But the participant's face may be affected by many negative effects such as blur, occlusion (e.g. covered by a sunglasses), head rotation (e.g. non-frontal head pose). All these effects will be examined in our study.

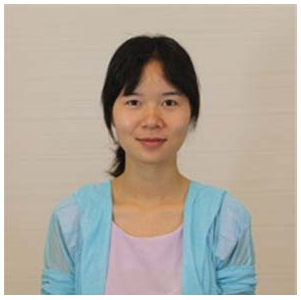
In the second part, we capture the participant's facial images with 35 controlled head poses and 5 facial expressions using a Canon EOS 60D (18.0-megapixel DSLR CMOS camera), as shown in Figure 4.2 and Figure 4.3 respectively. The resulting images are 5184×3456 in size with inner pupil distance of the subjects typically exceeding 400 pixels. 35 controlled head poses are specified by both horizontal and vertical rotation. Rotation angles are represented as (rot_H, rot_V) where rot_H corresponds to the angle of horizontal rotation while rot_V corresponds to the angle of vertical rotation. The value range of rot_H contains $0°$, $10°$ to left/right, $20°$ to left-/right, $30°$ to left/right while the value range of rot_V contains $0°$, $10°$ to up/down, $20°$ to up/down. We choose these boundary values according to the common restriction of existing face authentication systems [1], where a participant should not pass user authentication if rot_H exceeds $30°$ or rot_V exceeds $20°$ degrees. On the other hand, 5 facial expressions include neutral expression, smile without showing teeth, smile showing teeth, closed eyes, and open mouth. Continuous lighting system is used to eliminate the shadow on the participants' faces, as shown in Figure 4.4. To help the participants concentrate and reduce stress, the above process for each participant is completed strictly within 35 minutes and accompanied with music. Each participant have 1-2 minutes break for every 10 minutes and is offered beverage.

We use a helmet equipped with a gyroscope to control head rotation of the participants. The use of gyroscope has advantages over the other approaches, which includes attaining theoretical accuracy of less than 1 degree, ignoring the head position, measuring only orientation, not affected by metallic interference [54]. For

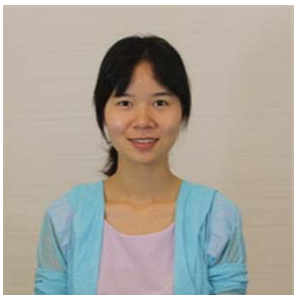
Figure 4.2: Sample images of 35 head poses (Courtesy of Lizi Liao from Singapore Management University)



Neutral

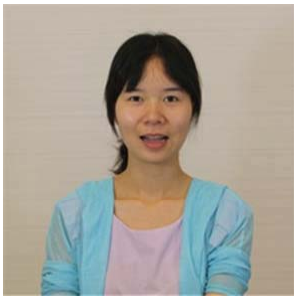


Smile without showing teeth



Smile showing teeth

Closed eyes



Open mouth

Figure 4.3: Sample images of 5 facial expressions

Figure 4.4: Continuous lighting systems

each head pose, we firstly ask the participants to face to the DSLR camera and help them adjust their heads to frontal position in the way similar to [30]. Then the participants rotate their heads to the required angles with help of the gyroscope. The gyroscope generates real-time rotation angles and broadcasts them via WiFi, as shown in Figure 4.5. This rotation information will be received and displayed on an iPad screen, and shown to the participants. Thirdly, we ask the participants to hold their head poses and one of our researchers then removes the helmet gently and quickly in order to avoid movement of the heads during helmet removal. After that, the images of each head pose are captured immediately.

In the final part, the participant will be asked to fill in a questionnaire for col-

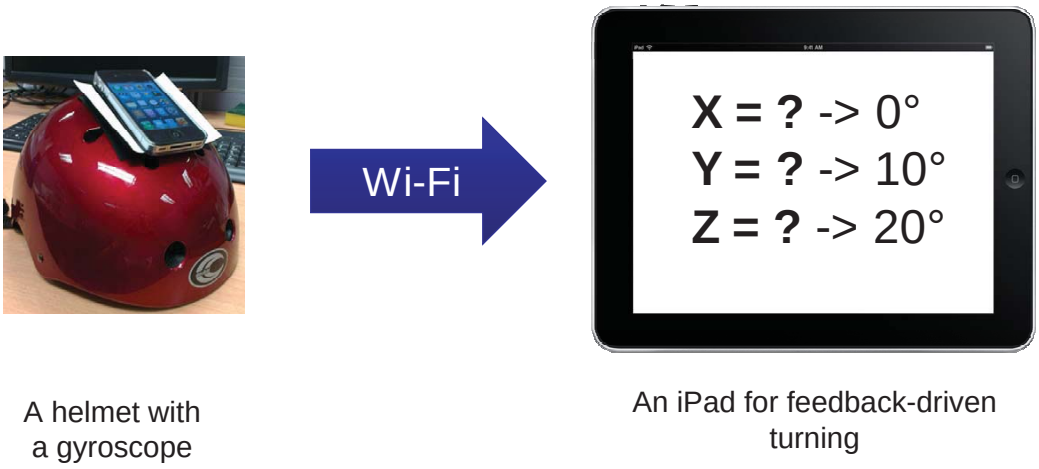


Figure 4.5: Rotation angles generated by gyroscope on helmet are displayed on iPad

lecting the participant's attitudes towards usage of face authentication systems and sharing behaviors in OSNs.

4.3.2 Empirical Results

Based on the collected images, we inspect the realistic threat of OSNFD against the latest version of popular real-world face authentication systems. We use the common experiment procedure similar to prior work [45, 16], which is described as follows: The frontal image is first used to enroll each participant into a face authentication system. Then we use a participant's own OSN images to test whether it can be used to log in a target face authentication system for his/her own account. The participant's OSN images are displayed on an LCD screen with resolution 1600×900 pixels. We fix the location of both the LCD screen and the camera of the face authentication systems. The camera is placed to face the LCD screen. We adjust the scales of the images in order to improve the focus of the camera. The result whether a target system can be spoofed by an OSN image will be recorded for each system and each image.

Our analysis uses two basic metrics, namely *vulnerable images* and *vulnerable users*. A vulnerable image, denoted by $VulImage$, is defined as a facial image

Table 4.1: Overall percentage of $VulImage$ and $VulUser$

	$VulImage\%$	$VulUser\%$
Face Unlock	45%	86%
Facelock Pro	46%	96%
Visidon	68%	97%
Veriface	27%	73%
Luxand Blink	20%	41%
FaceAccess	33%	80%
Average	39%	77%

which is wrongly accepted as a genuine user by a face authentication system during user authentication and therefore enables an adversary to circumvent the face authentication system. A vulnerable user, denoted by $VulUser$, is a user enrolled in a face authentication system who has at least one vulnerable image published in OSNs.

Table 4.1 shows that the face authentication systems are vulnerable to the OSNFD in general. On average, 39% of the OSN images and 77% of the participants are vulnerable. Among popular face authentication systems, Visidon is more vulnerable in low security level, for which 68% of the images and 97% of the participants are vulnerable. Especially for Google's Face Unlock that comes as a built-in feature of all Android-based systems whose version is higher than 4.0 [27], 45% of the OSN images and 86% of the participants are vulnerable.

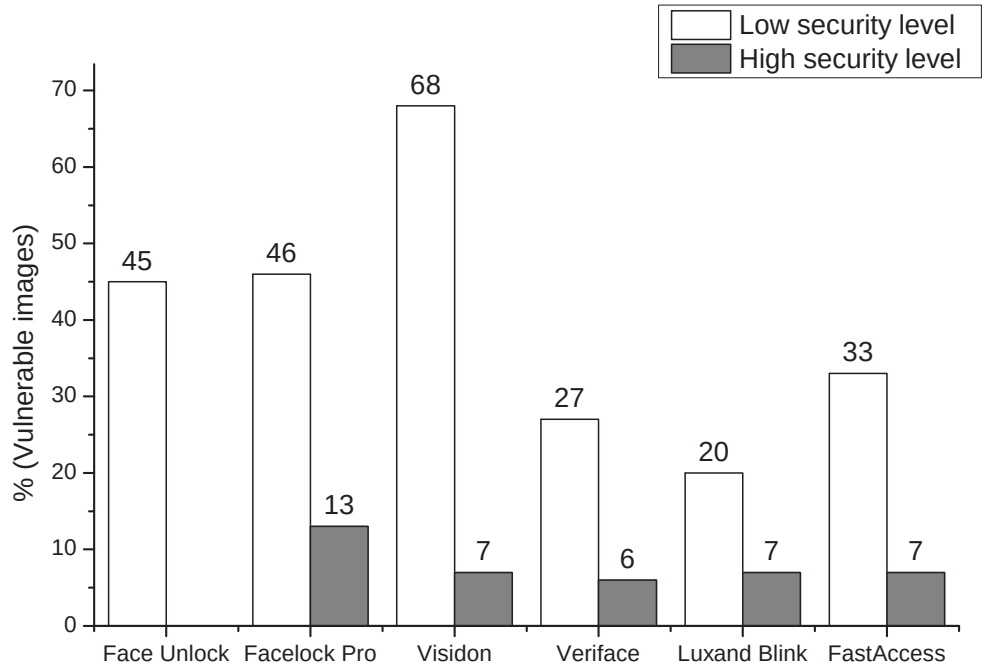
Although the percentage of vulnerable images is moderate, the quantity of the vulnerable images is large due to the huge amount of images in OSNs. These large amount of vulnerable images create resources online for potential attacks. Even worse, users share their personal images with their friends in OSNs, most of them tend to publish the images where the users' faces can be clearly viewed for easier recognition. Consequently, the percentage of vulnerable users would be high as observed in our study. The following subsections will further analyze the detailed characteristics of these vulnerable images and users from three major perspectives, security settings, target platforms, and user behaviors.

Impacts of Security Settings

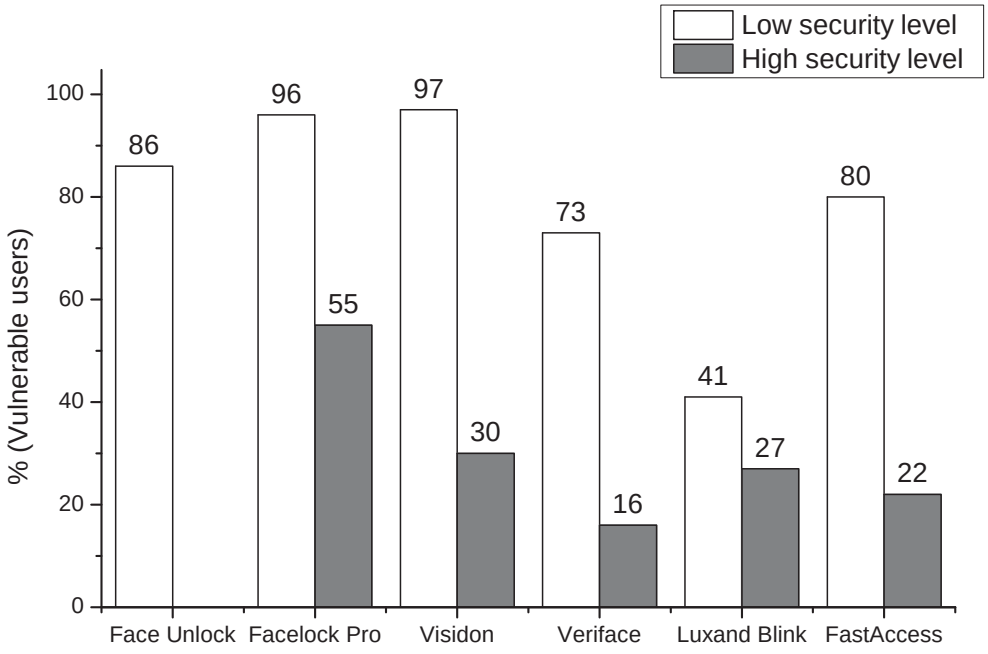
Security settings specify the security strength of a face authentication system against potential attacks. As previously explained, most of face authentication products [10, 8, 34, 23, 24, 35] provide very limited choices on security level. So we focus our analysis on lowest and highest security level that can be provided by each system, which are denoted as low security and high security, respectively. Since there is only one security level in Face Unlock and the observed security strength of Face Unlock is comparable to the other systems in low security level, we classify its security level as low. As expected, Figure 4.6 shows that the face authentication systems in low security level are facing more severe OSNFD threat than those in high security level. On average, 40% of the images and 79% of the participants are vulnerable for the face authentication systems in low security level while 8% of the images and 30% of the participants are vulnerable for the face authentication systems in high security level.

The change of security settings generally affects the recognition threshold in the face matching module. As the security level is raised, the recognition threshold becomes higher which imposes more restrictions for matching between login facial image and pre-stored facial image. Therefore the face authentication imposes more rigid restrictions on the login facial image. The major restrictions observed in our study are head pose and lighting condition.

For head pose, we use *acceptable head pose range* to measure the tolerance of a face authentication system on head pose variations. It describes the head rotation range of head poses with which at least 50% of the participants successfully log in the face authentication systems. In these tests, we use participant's frontal image for enrollment and use the images collected with controlled head poses as test inputs (i.e. login images). Figure 4.7 shows the average results computed from all tested systems, where each closed curve corresponds to the acceptable head pose range. The results for each individual system that indicate the difference between high



(a)



(b)

Figure 4.6: Percentage of $VulImage$ and $VulUser$ in different security levels

security and low security are similar to Figure 4.7, which are not shown.

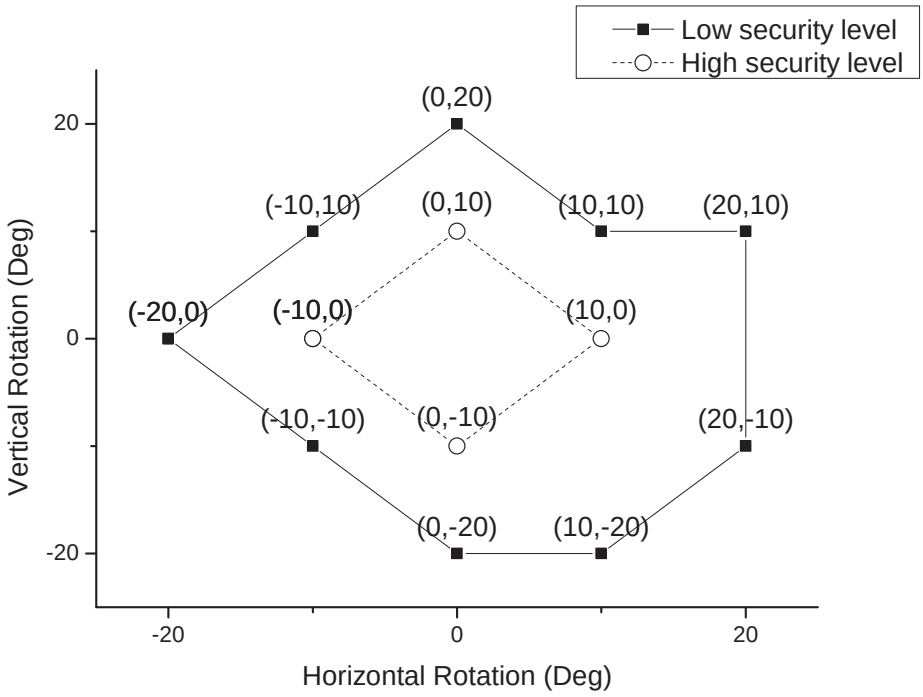


Figure 4.7: Tolerance of the rotation range of head pose

For lighting condition, we further classify it into different types of illumination and low lighting [25, 79, 43]. The face authentication systems in low security level are observed to have higher tolerance for variation of lighting conditions than the systems in high security level. In our study, illumination is observed in 27% (394 out of 1440) of the OSN images while low lighting is observed in 18% (266 out of 1440) of the OSN images. On average, 81% of the OSN images with illumination and 79% of the OSN images with low lighting cannot be used to log in the face authentication systems in low security level while 96% of the OSN images with illumination and 94% of the OSN images with low lighting cannot be used to log in the systems in high security level.

On the other hand, a face authentication system in low security level has higher tolerance for varied login environments, which is necessary for the system to be usable in the complex environments. As a tradeoff for higher security strength, the false rejection rates in high security level may be significantly increased. As shown

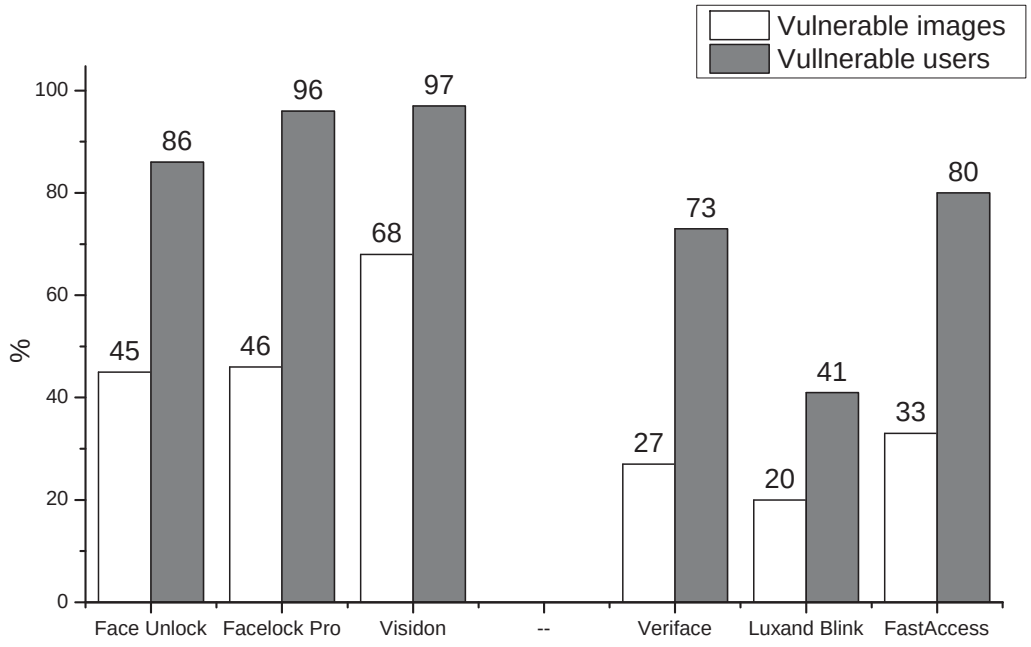
in the follow-up experiment described in Section 4.5.1, the false rejection rate could be as high as 85%. This will cause a significant concern on the accessibility. From our questionnaire on user perception, 70% of the participants think it is important to successfully log in their smartphones, tablets, or laptops at the time they want to use. If the face authentication system is not always functional, 67% of the participants give up using the system which causes the serious accessibility problem to their devices. This may also explain why the popular face authentication systems always use low security level by default.

Impacts of Target Platforms

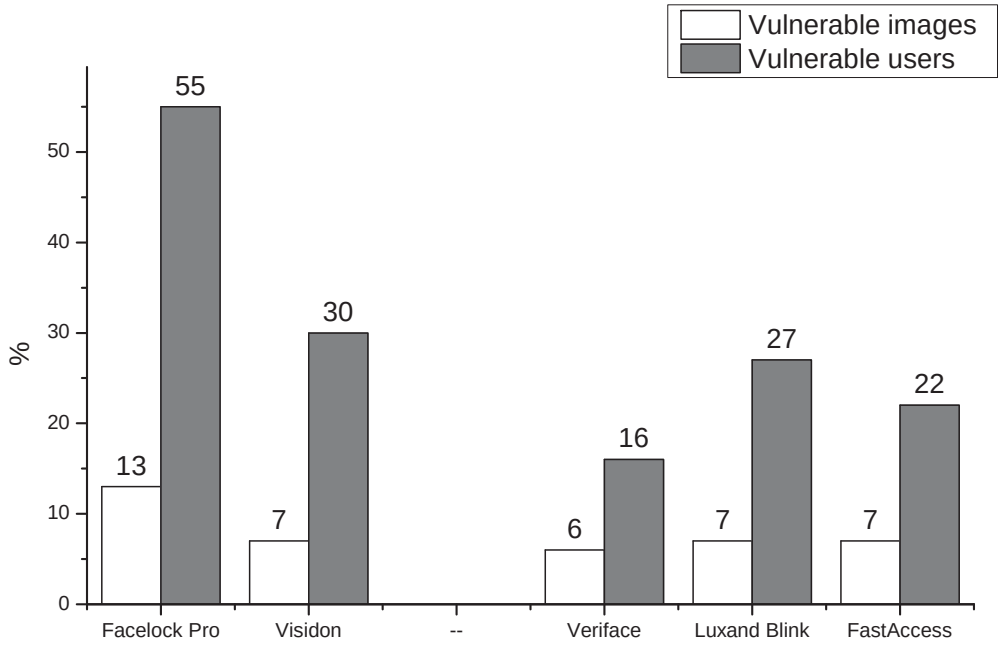
The target platform of a face authentication system imposes the platform-specific requirements on both security and usability. In our tested systems, Face Unlock, Facelock Pro, and Visidon are targeting for mobile platform, while Veriface, Luxand Blink, and FastAccess are targeting for traditional platform.

Figure 4.8 shows that the OSNFD threat for mobile platform is generally more severe than the OSNFD threat for traditional platform. On average, in low security level, 53% of the images and 93% of the participants are vulnerable for the face authentication systems on mobile platform while 27% of the images and 64% of the participants are vulnerable for the systems on traditional platform. In high security level, 10% of the images and 43% of the participants are vulnerable for the face authentication systems on mobile platform while 7% of the images and 22% of the participants are vulnerable for the face authentication systems on traditional platform.

These results clearly show the difference caused by platform-specific requirements. Compared to a traditional system, a mobile system is usually designed to be more robust and more tolerant to varied environments such as outdoor environment in order to meet accessibility expectation by users. Meanwhile it leads to the more severe OSNFD threat for mobile platform based systems. This difference is confirmed by the results of our questionnaire, which shows that 91% of the participants



(a) Low security level



(b) High security level

Figure 4.8: Difference in $VulImage$ and $VulUser$ between systems targeting for mobile platform and traditional platform.

believe that it is important to log in smartphones or tablets in both indoor and outdoor environment while only 36% of the participants think it is important to log in laptops in both indoor and outdoor environment.

This difference is also revealed in our tests on head pose and lighting condition. Figure 4.9 shows the face authentication systems targeting for mobile platform have higher tolerance for variations of the head poses than the systems targeting for traditional platform.

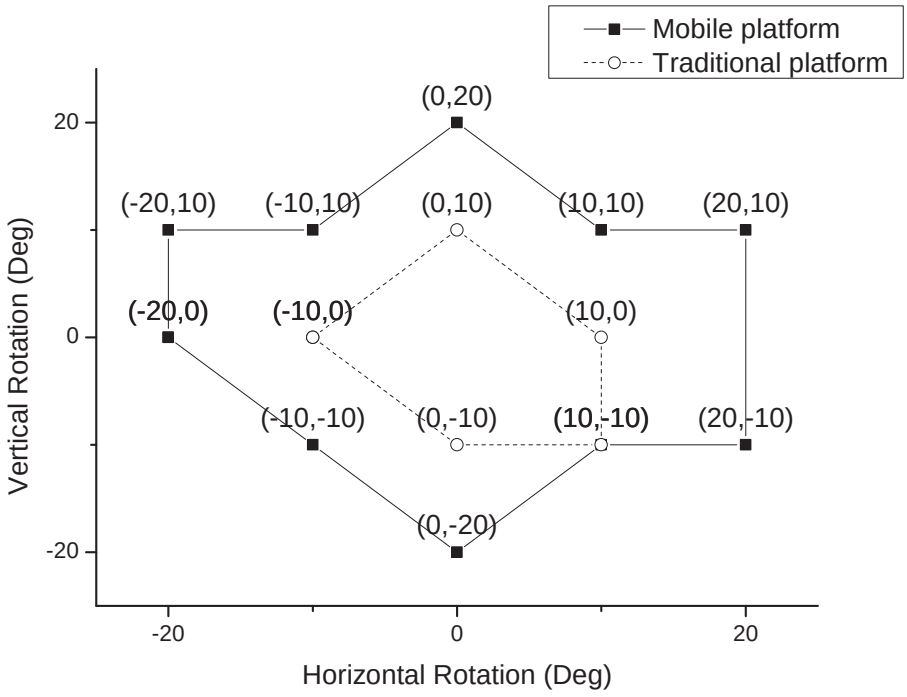


Figure 4.9: Difference in the tolerance of the rotation range of head pose.

Our tests on lighting conditions further show the face authentication systems targeting for mobile platform are more tolerant to variations of the lighting conditions. In our study, 81% of the OSN images with illumination and 77% of the OSN images with low lighting cannot be used to log in the face authentication systems targeting for mobile platform, while these rates increase to 96% for the images with illumination and 96% for the images with low lighting on traditional platform.

Impacts of User Behaviors

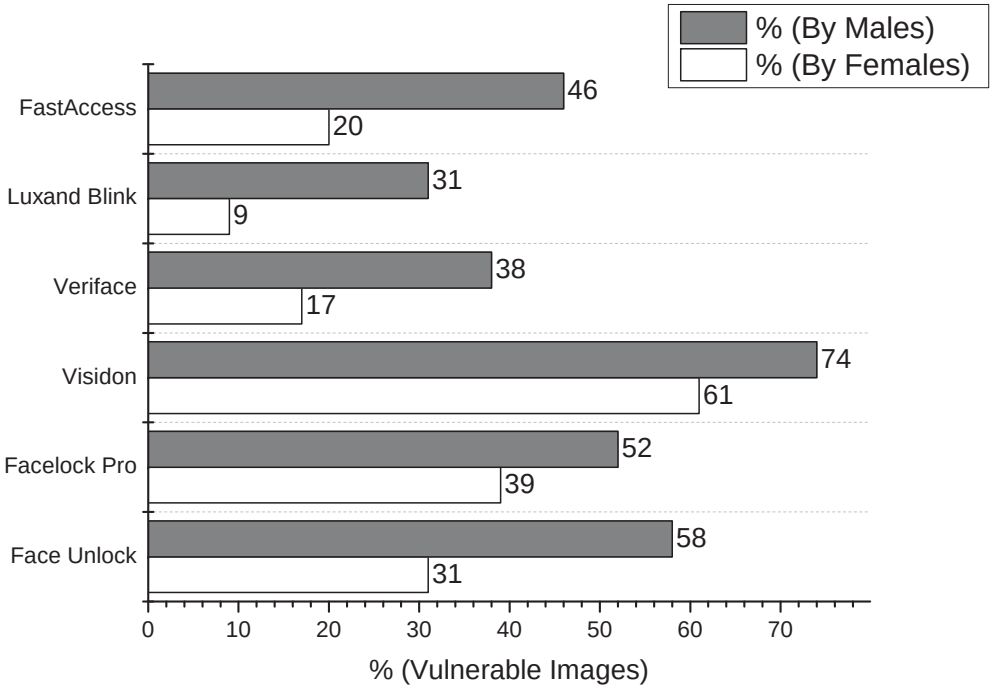
The difference in user behavior is another major factor influencing the quality of shared images that decides whether these images can be eventually used for successful OSNFD-based attacks. Our study reveals that the participants who publish more facial images in OSNs are not necessarily more vulnerable than those who publish less facial images in OSNs. In fact, the OSNFD threat is more severe among the participants who publish facial images with higher quality in OSNs.

To illustrate the impact of user behaviors, we use the different sharing behaviors and the different OSNFD threat between females and males as example. In our study, female participants are reported to publish facial images in OSNs more frequently than male participants in general. On average, each of the female participants publishes 65 facial images per year while each of the male participants publishes 34 facial images per year. However, the OSNFD threat for the females is less severe than that for the males, as shown in Figure 4.10 and Figure 4.11.

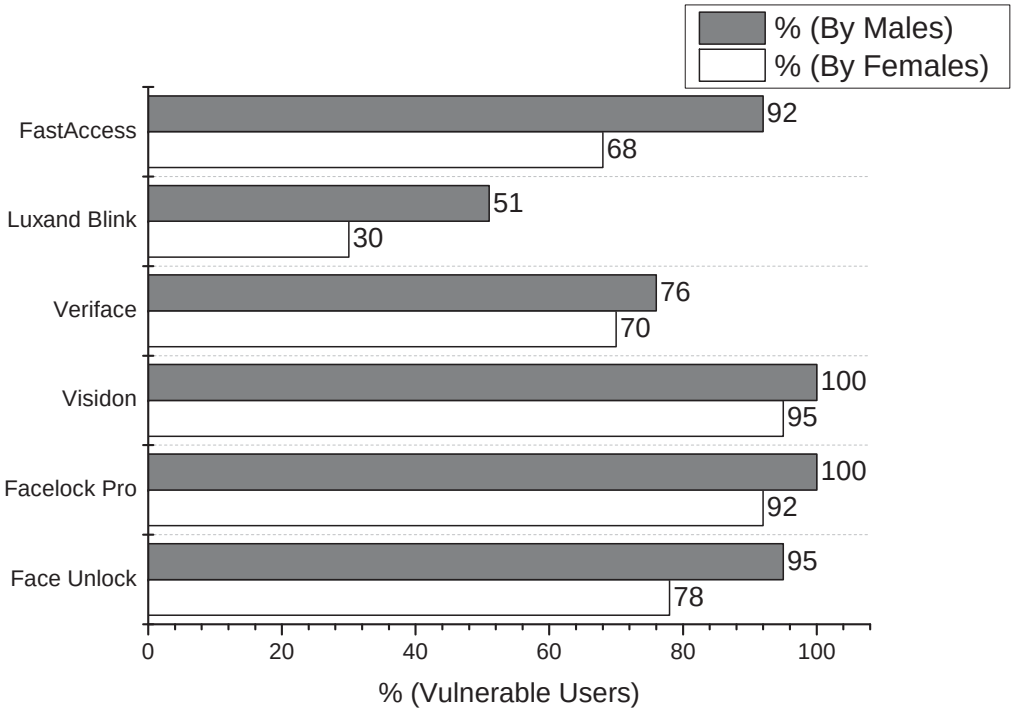
This can be explained by the lower quality of the OSN images published by the females. We find that the female participants are more likely to publish blurred images, edited images, or images with their makeup, as shown in Figure 4.12. The blur, edit, and makeup can degrade the quality of an image and therefore lead to the difficulty in face recognition [38, 20]. In our study, 12% of the OSN images suffer from these negative effects. Among these low quality images, 61% are published by the females while only 39% of the images are published by the males. All of these blurred, makeup, or edited images fail to pass at least one face authentication system.

4.4 Statistical Analysis and Risk Estimation

Although the OSNFD threat is significant as shown in the previous section, we observe the effectiveness of OSNFD-based attacks may be significantly reduced by

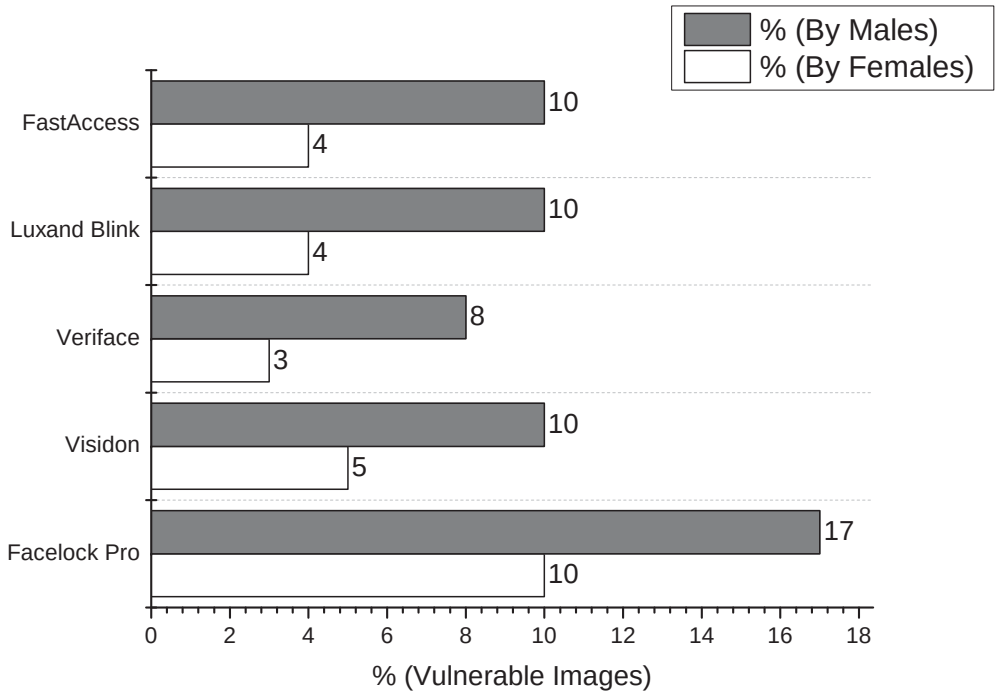


(a) $VulImage\%$ in low security level

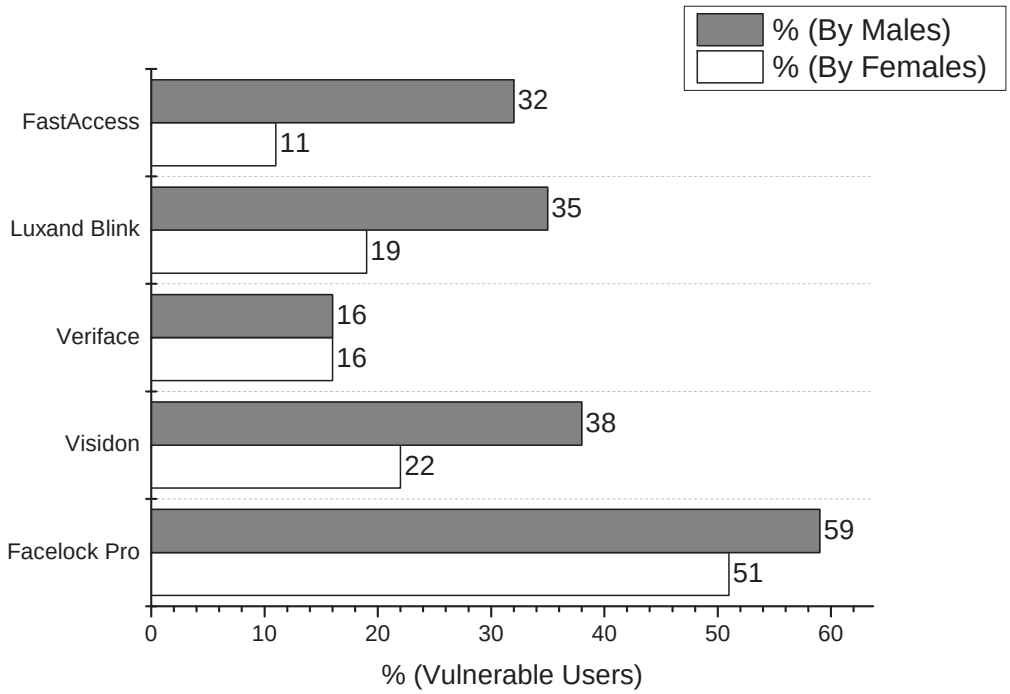


(b) $VulUser\%$ in low security level

Figure 4.10: Difference in $VulImage$ and $VulUser$ between females and males configured in low security level



(a) $VulImage\%$ in high security level



(b) $VulUser\%$ in high security level

Figure 4.11: Difference in $VulImage$ and $VulUser$ between females and males configured in high security level

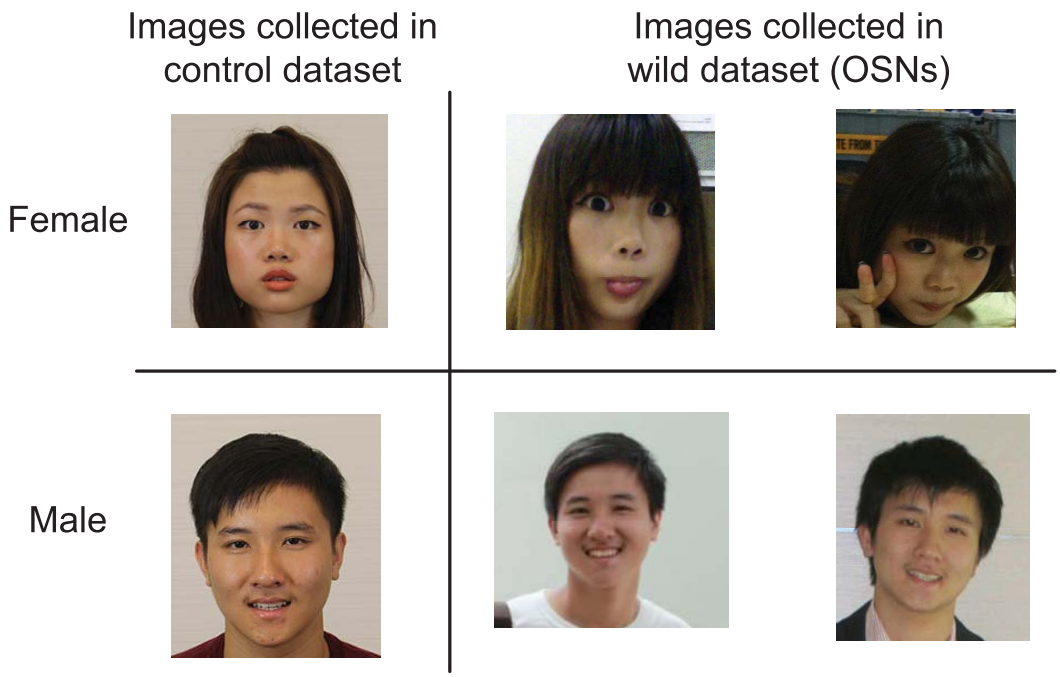


Figure 4.12: Sample images of female and male collected in controlled dataset and wild dataset

manipulating certain attributes of facial images. In this section, we extract these key attributes via statistical analysis and use them to develop an estimation tool for end users to calculate the risk of their shared images.

4.4.1 Key Attributes Affecting OSNFD-Based Attacks

From the theoretical perspective, there are still many challenges for face recognition algorithms. These challenges also become key attributes that limit the effectiveness of OSNFD-based attacks. The common attributes addressed in the prior study [1] include head pose, lighting condition, facial expression, facial occlusion, and image resolution. Beside these traditional attributes, we also observe blur, facial makeup, and editing (using Photoshop-like software) as the extra key attributes which often appear in the real world images shared in OSNs, though they are usually not considered in the controlled settings of traditional study on face authentication. We describe the details of these key attributes as follows.

Head pose is a prominent challenge to face recognition. The performance of

face recognition algorithms in face authentication can be significantly affected if the head pose in a login image and the head pose in the pre-stored facial image are different [79]. The affecting variations of a head pose mainly include two out-of-plane rotations, namely horizontal rotation and vertical rotation [54].

Lighting condition is another prominent challenge in the realm of face recognition. The variation of lighting conditions mainly includes illumination and low lighting [25, 79, 43]. The illumination is mainly caused when direct light shoots on the 3D structure of a face and strong shadows can be casted which diminish facial features [25, 79]. The illumination can be classified into side illumination and top/bottom illumination [25]. Low lighting is another negative lighting condition, which usually happens when a facial image is taken in dim environment or with extreme bright background. The low lighting may diminish facial features since the luminance in face region is too low for face recognition algorithms to recognize [43].

Facial expression such as smile, surprise, etc, can change face geometry and therefore affect the performance of face recognition algorithms [1]. The common facial expressions include neutral expression, smile without showing teeth, smile showing teeth, closed eyes, open mouth, and other expressions.

Facial occlusion often happens in real world due to additional accessories on face, such as sunglasses, scarf, hands on face, etc. The occlusion can result in the failure of face appearance representation or imprecise facial feature searching and localization, and therefore have negative influence on the performance of face recognition algorithms. The common facial occlusions include forehead occlusion, eyebrow occlusion, eye occlusion, cheek occlusion, and mouth occlusion [1].

The **resolution** of an image can affect accuracy of facial landmark localization and therefore influence the performance of face recognition algorithms. As the resolution of face images decreases, the performance of the face recognition algorithms drops [79].

The **blur** in a facial image causes difficulty in accurate localization of edges of

facial region and facial landmarks (i.e. eyes, nose, mouth, etc) by face recognition algorithms and therefore harms the performance of the algorithms.

Facial **makeup** can substantially change the appearance of a face and facial landmarks, such as the alternations of perceived facial shape, nose shape, location of eyebrows, etc. These alternations by the facial makeup, especially by non-permanent facial makeup, challenge face recognition significantly [20].

The **editing** of an image introduces noise pixels and change the appearance of the face in the image [2, 20]. Face recognition algorithms can be affected by these noises and appearance changes due to the edited image.

All these attributes significantly degrade the image quality and therefore lead to the failure of OSNFD-based attacks. They are used as input parameters to build our risk estimation tool in the next section.

4.4.2 Risk Estimation Model

We use binomial logistic regression [34] to model the impact of the key attributes introduced in the previous subsection. The notions of these attributes are defined in Table 4.2. Then the key attributes of each image can be represented by an input parameter vector, denoted as $V = (rot_H, rot_V, ill_{sd}, ill_{tb}, dm, bg, FEx_n, FEx_s, FEx_{st}, FEx_{ce}, FEx_m, FEx_{other}, Occ_{fh}, Occ_{eb}, Occ_{eye}, Occ_{chk}, Occ_{mh}, res, blur, mk, ed)$.

For the output, we assign an OSN image to either a positive class or a negative class. The positive class means the image can be used to pass the login of a specific face authentication system, otherwise the image will be in the negative class.

Binomial logistic regression is a classic probabilistic classification model [34], which accepts multiple predictor variables as inputs, and predicts the outcome for a dependent variable which has only two possible types, such as “positive” vs “negative”. Thus it is a proper tool to calculate the probability of an image assigned to

Table 4.2: Parameters related to the key attributes

Attribute	**Parameter**	**Notation**
Head pose	Horizontal rotation	rot_H
	Vertical rotation	rot_V
Lighting condition	Side illumination	ill_{sd}
	Top/bottom illumination	ill_{tb}
	Dimness	dm
	Bright background	bg
Facial expression	Neutral	FEx_n
	Smile without showing teeth	FEx_s
	Smile showing teeth	FEx_{st}
	Closed eyes	FEx_{ce}
	Open mouth	FEx_m
	Other expressions	FEx_{other}
Facial occlusion	Occluded forehead	Occ_{fh}
	Occluded eyebrow	Occ_{eb}
	Occluded eye	Occ_{eye}
	Occluded Cheek	Occ_{chk}
	Occluded mouth	Occ_{mh}
Resolution	Resolution	res
Blur	Blur	$blur$
Facial makeup	makeup	mk
Edit	Edit	ed

the positive class based on the key attributes extracted from an OSN image. Given a parameter vector V_i of a facial image i and a face authentication system in a security level, the regression function is

$$\ln\left(p_i/(1-p_i)\right) = \beta_0 + \beta_1 v_1 + \cdots + \beta_m v_m \tag{4.1}$$

where p_i is the probability that an image i is assigned to the positive class, v is a parameter in V_i, and β is a regression coefficient. The risk score of the facial image i is the value of p_i. The facial image i is assigned to the positive class if $p_i \geq 0.5$. Otherwise, i is assigned to the negative class. The correctness of these assignments is verified with the ground truth data collected from the previous empirical analysis.

For each combination of face authentication system and its security level, we examine the model fitting of binomial logistic regression and the significance of the

parameters by using the real world OSN images and run binomial logistic regression on SAS software [61]. The likelihood ratio test and wald statistic [34] for all the face authentication systems are smaller than 0.0001.

Our statistical analysis shows the most influential attributes are resolution res, occluded eye Occ_{eye}, makeup mk, and illumination ill_{sd}. Resolution res has positive impact on the risk of OSNFD. It is because higher resolution contributes to more accurate facial landmark localization and results in better performance of face recognition and increases the risk of OSNFD. The occluded eye Occ_{eye}, makeup mk, and illumination ill_{sd} have negative impact and lower the risk of OSNFD. In particular, the occluded eye leads to decrease in the performance of face recognition algorithms, as accurate localization of eyes is important for the alignment process in all major face recognition algorithms [1]. Makeup can significantly change the appearance of the face and the facial landmarks and therefore lowers the performance of face recognition. The illumination is a prominent attribute which causes difficulty in face recognition since it diminishes facial features.

The parameters related to other attributes, including head pose and facial expression, are generally not statistically significant. Among the collected OSN images, the variations of head pose and facial expression are limited since users are usually cooperative when these images are captured and tend to publish the images from which they are easily recognized. As observed in our study, the head poses in most OSN images are within the acceptable head pose ranges of the face authentication systems, which causes the insignificance due to lack of samples with extreme head pose. On the other hand, facial expressions observed in most OSN images are only mild-mannered expressions including neutral expression, smile without showing teeth, smile showing teeth, closed eyes, open mouth. These common expressions do not have significant impact as they have been well handled in current face recognition algorithms [1]. Other extreme facial expressions, such as making faces, do significantly affect the face recognition, but they are observed in only 5% of the OSN images.

4.4.3 Model Evaluation

To evaluate the performance of the proposed risk estimation tool, we use cross-validation method. In each round, for each of the face authentication systems in a specific security level, we randomly choose 80% of the OSN images to train the model and use the risk estimation tool to automatically classify the rest of the images. The above process is repeated by 10 rounds. The performance is measured by standard classification evaluation metrics, including precision, recall, and F1 score [60].

Precision is defined as the percentage of the true positive images among the images assigned to the positive class by the risk estimation tool, which can be calculated by $tp/(tp + fp)$ where tp is the number of true positive images and fp is the number of false positive images. Recall is defined as the percentage of the true positive images detected by the risk estimation tool among the positive images in ground truth, which can be calculated by $tp/(tp + fn)$ where tp is the number of true positive images and fn is the number of false negative images. F1 score considers both the precision and the recall, which can be calculated by $\mathsf{F1} = 2 \times \mathsf{precision} \times \mathsf{recall}/(\mathsf{precision} + \mathsf{recall})$.

Table 4.3 shows the performance evaluation metrics of the risk estimation tool. On average, the risk estimation tool achieves a precision of 81%, a recall of 83%, and an F1 score of 82%. The performance evaluation indicates that the risk estimation tool detects most of the vulnerable images which can lead to successful OSNFD-based attacks if these images are published in OSNs.

4.5 Discussion

4.5.1 Tradeoff between Security and Accessibility

Clear tradeoffs between security and accessibility can be observed in our tested systems, which are decided by security settings and target platforms as analyzed

Table 4.3: Effectiveness of our risk estimation tool

System	Security level	Precision	Recall	F1 score
Face Unlock	N/A	73%	77%	75%
Facelock Pro	Low	70%	69%	69%
	High	81%	75%	78%
Visidon	Low	79%	90%	84%
	High	86%	92%	89%
Veriface	Low	79%	68%	73%
	High	90%	98%	94%
Luxand Blink	Low	84%	87%	85%
	High	87%	90%	88%
FaceAccess	Low	77%	67%	72%
	High	89%	95%	92%
Average	N/A	81%	83%	82%

in Section 4.3.2. The increasing security strength inevitably decreases the accessibility. We conduct a follow-up experiment to collect quantitative evidence for the impact of these tradeoffs.

20 participants from the main user study are invited for this follow-up study. The participants need to enroll their faces in the 6 face authentication systems in low/high security level in a meeting room with normal lighting, respectively. To mimic the different login environment, the experiments are conducted between 2pm-4pm in a sunny day at four fixed indoor/outdoor locations, including 1) a meeting room with normal lighting condition, 2) a meeting room with dim lighting condition, 3) outdoor ground in the sunshine, and 4) shelter of building. This setting simulates a situation when a user registers in one place, but tries to access the system in many other places. The participants are asked to login by using each face authentication systems. In this experiment, there are no OSN images, but only live legitimate users who attempt to access a face authentication system. Each participant has at most three attempts for each login before we record it as a false rejection.

Table 4.4 shows the false rejection rates of the face authentication systems in low security level are lower than those of the face authentication systems in high security level in overall. Moreover, the face authentication systems on mobile platform have lower false rejection rates than those on traditional platform. The highest

Table 4.4: Significant increase in false rejection rates when using high security level settings. The increments of false rejection rates are more significant for traditional platform-based systems (the last three systems).

System	Security level	Room+ normal lighting	Room+ dim lighting	Outdoor ground	Shelter
Face Unlock	N/A	0%	5%	10%	0%
Facelock Pro	Low	0%	10%	10%	0%
	High	0%	45%	60%	25%
Visidon	Low	0%	5%	5%	0%
	High	5%	55%	65%	50%
Veriface	Low	0%	25%	35%	20%
	High	10%	60%	85%	60%
Luxand Blink	Low	0%	30%	50%	45%
	High	5%	55%	70%	55%
FastAccess	Low	0%	15%	30%	15%
	High	5%	55%	65%	55%

observed false rejection rate is 85% for Veriface in high security level. This accessibility degradation could be a disaster for end users. In our questionnaire, 91% of the participants believe that it is important to log in smartphones and tablets in both indoor and outdoor environments, while 36% of the participants think that it is important to log in laptops in both indoor and outdoor environments. If a face authentication system is set to high security level in order to mitigate the OSNFD threat, the system will be less tolerant for complex environments and violate the users' need of accessibility.

4.5.2 Costs of Liveness Detection

Liveness detection could be a mitigation for OSNFD-based attacks, which is designed to distinguish between a live face and a facial image in front of the camera. The most common liveness detection mechanisms deployed on popular face authentication systems are eye-blinking and head rotation detection, as they have the advantages of no additional hardware support, requiring moderate image quality, and involving relatively low usability cost. This is important to all consumer-level products that are price-sensitive and accessibility-first. However, the security strength

of the two mechanisms is weak. They can be easily bypassed with one or two pre-catched images as shown in [59]. The practicality of these attacks is also verified by our experiments.

Besides these two simple mechanisms, several sophisticated liveness detection techniques have been proposed for face authentication. However, all of them are associated with considerable costs as shown in Table 4.5 [56]. Their costs include requiring additional hardware, high quality images, ideal environment that are usually not universally available, and high user collaborations that may cause inconvenience. This indicates they may not be suitable for consumer-level face authentication systems. It still remains a challenge to deploy reliable and practical liveness detection in face authentication systems that can be used by the public.

Table 4.5: Costs associated with existing liveness detection mechanisms for face authentication. * sign indicates a requirement involves a significant cost for end users or device manufacturers.

Liveness detection	**Image quality**	**Additional hardware**	**Usability cost**
Eye blinking	Low	No	Low
Mouth movement	Middle	No	Middle
Degradation	High*	No	Low
Head movement	High*	No	Middle
Facial expressions	High*	No	Middle
Facial thermogram	N/A	Yes*	Low
Multi-modal	N/A	Yes*	Middle/High*
Facial vein map	N/A	Yes*	Middle
Interactive response	N/A	Yes*	High*

4.5.3 Implications of Our Findings

Face authentication does provide an attractive alternative of user authentication for its non-intrusive and zero-memory procedure. However, the appearance of OSNFD brings a significant threat to question the practicality of face authentication as a usable authentication factor. Nowadays, a huge amount of personal facial images/videos have been published in OSNs that can be accessible to potential adversaries without the previously required physical proximity. Therefore, face biomet-

rics can now be disclosed in large scale and acquired by adversaries remotely. Face biometrics are no longer secrets only owned by the users and can be disclosed to anyone who has access to victim's personal images shared in OSNs.

Raising the security level of face authentication systems could mitigate the OSNFD threat by scarifying the accessibility, which leads to the inconvenience for legitimate users. Liveness detection is another major countermeasure to mitigate the spoofing attack against the face authentication systems. Unfortunately, existing liveness detection techniques available on consumer-level computing devices can be easily circumvented by one or two images. More reliable liveness detection like multi-modal mechanisms usually relies on using additional authentication factor (e.g. another biometrics such as voice and fingerprint). This introduces another liveness detection problem for the additional authentication factor, which may not be reliable. For example, voice and fingerprint can also be spoofed. Even worse, more serious privacy concerns will rise if a system requires to collect many biometrics information from a user [78], which may eventually cause the rejection of the liveness detection mechanism.

As the emergence of OSNFD, the face biometrics is losing confidentiality which is one of the fundamental requirements for a usable authentication factor. Moreover, the existing liveness detection techniques are either too weak to defend against the OSNFD or too difficult to be deployed on the consumer-level devices. All these findings suggest that face authentication may not be a proper authentication factor unless we can resolve the discovered problems.

4.5.4 Limitations

Ecological validity is a challenge to any user study. Like most prior research [29, 66, 3], our study only recruits students in university. These participants are more active in using consumer-level computing devices and sharing images in OSNs. Thus the evaluation of the OSNFD may vary with other populations.

In the user study design, it is still a challenge to collect facial images with precisely controlled head poses [54]. Like the prior head pose data sets [30, 46, 63], the accuracy of the head poses in our data set may be affected by the poor ability of the participant to accurately direct his/her head, the unconscious movement of human beings and limit of resources. In another experiment of examining the false rejection rates of the face authentication systems, we choose 4 locations to mimic different login environments in daily life. Since it is impossible for all the participants to do the tests at the same time and at the same physical positions, the background of image inputs captured by the camera may change.

Another challenge in our study is to accurately estimate parameters [44] such as head pose, illumination, and makeup in our collected OSN dataset. Since the accuracy of automatic labeling tools is limited [1, 57], we manually label the OSN images with the help of automatic tools and follow the similar validating methodology used in prior study [44, 39, 74]. For each OSN image, we estimate the head pose with typical head pose estimation algorithms including POSIT and LGBP [54]. And we manually validate the estimation of the head pose by comparison between the OSN image and the participant's images with controlled head poses. We manually label the parameters related to lighting conditions according to the shadow and histogram of face region similar to the approaches in [44, 39]. The parameters related to facial expressions are label by comparing the OSN image with the images captured in our user study, which is similar to [44, 39]. We use popular face detection software Picasa to mark the face region with a rectangle in the image and calculate the resolution of the face region. The parameters related to the attributes of blur, makeup, and edit are labeled in the way similar to [44, 39, 74].

It is also possible to further improve our risk estimation tool. To our best knowledge, our work is the first attempt to semi-automatically detect the vulnerable images that can be used to attack face authentication. Our current risk estimation tool can serve as a baseline for future improvement by refining the key parameters and the statistical model. It is also valuable to incorporate automatic high accuracy la-

beling for those hard-to-label attributes like illumination and facial makeup, once the ongoing research [54, 25, 20] resolves these challenges.

Chapter 5

Dissertation Conclusion and Future Work

5.1 Summary of Contribution

This dissertation makes contributions on analyzing privacy leakage under privacy control in OSNs and understanding OSN-based facial disclosure threats against real-world face authentication systems.

Our first work investigated privacy leakage under privacy control in online social networks. Our analysis showed that privacy leakage could still happen even after users correctly configure their privacy settings. We examined real-world OSNs including Facebook, Google+, and Twitter, and discovered the exploits which lead to privacy leakage. Based on the findings, a series of attacks were introduced for adversaries with different capabilities to learn undisclosed personal information. We analyzed necessary conditions and provided suggestions for users to mitigate privacy leakage in OSNs. We conducted a user study to evaluate the feasibility of the attacks. In the end, we discussed the implications of resolving privacy leakage in OSNs. The partial results of this work have been published in Proceedings of the 7th International Conference on Network and System Security (NSS 2013) [48]. And this work has been submitted to a security journal at the time when this dissertation

was submitted.

In the second work, we investigated the threat of OSN-based facial disclosure (OSNFD) against some real-world face authentication systems. We made the first attempt to provide a quantitative measurement on the threat of OSNFD against face authentication. Our results show that the face authentication systems are vulnerable to OSNFD-based attacks. We analyzed the characteristics of these attacks from three major perspectives including security settings, target platforms and user behavior. The key attributes of the OSNFD were further extracted to develop a risk estimation tool that can help users understand the risks associated with their personal images shared in OSNs. Our work made the first step in systematically understanding the OSNFD. Quantitative evidence indicates that face authentication may not be a proper authentication factor as the confidentiality of face biometrics has been significantly compromised by OSNFD. This work has been accepted and will be published in Proceedings of the 9th ACM Symposium on Information, Computer and Communications Security (ASIACCS 2014) [49].

5.2 Future Direction

As OSNs become landmine for information disclosure, the privacy and security problems due to the information disclosure in OSNs have attracted strong interest among researchers. Many research works show that the private information can be disclosed via the publicly shared information in OSNs due to inference attacks by malicious adversaries and wrong configuration of privacy control by users [80, 41, 14, 7, 48]. The disclosed information directly leaks users' privacy. Among the disclosed information, the disclosed face biometrics information can further threaten the security of face authentication systems, which has been proved by our second work [49]. Thus it is urgent to look for solutions to address the above problems. Although many efforts are made to design more powerful privacy control models and mechanisms for OSNs, none of them are accepted and used by

existing OSN users and OSN service providers [24, 11, 37, 76]. The underlying reason is that the inconsistencies between privacy control and OSN functionality make privacy control in OSNs vulnerable even when the privacy control is properly configured. These inconsistencies reflect the conflicts between users' privacy intention and social/business values of OSNs. The social and business factors behind make it difficult to completely resolve the information disclosure in OSNs.

Although it is difficult to completely avoid the information disclosure in OSNs, we can still mitigate risks of the privacy leakage and the threat of face biometrics disclosure in OSNs in the following two ways: (1) improving the usability of privacy control; (2) designing reliable and practical liveness detection for face authentication. The details of the mitigations are discussed below.

The privacy control mechanisms in OSNs enable users to determine who can view their information. Users may wrongly choose privacy control rules for their published information by mistake. It is important to improve the usability of the existing privacy control in OSNs in order to help users be aware of the privacy control rules chosen by mistake. Wang et al show that wrong privacy control rules are chosen by users due to various reasons including misunderstanding privacy rules provided by multiple OSNs and publishing information in a state of high emotion [73].

On one hand, the privacy control rules provided by different OSNs may be incompatible as analyzed in our first work in Chapter 3. An automatic tool for detecting incompatibility of privacy control rules can be used to remind a user of the potential disclosure of his/her information when he/she is publishing personal information in multiple OSNs. The automatic incompatibility detection tool needs to compare the privacy control rules for the same information published in multiple OSNs and report the inconsistent privacy control rules.

On the other hand, when users are in a state of high emotion, they are more likely to wrongly choose privacy control rules which may be different from the privacy control rules they usually choose for the same or similar information published in OSNs. The above problem can be mitigated by using an automatic policy

recommendation mechanism. The automatic policy recommendation mechanism is supposed to recommend privacy control rules based on the content of the published information and the users' historical privacy control rules for a similar content. The policy prediction mechanism for text-based content made the first attempt to address the above problem [64]. However, it remains challenging if the content of published information is images, videos, or short text.

The OSN-based face biometrics disclosure causes serious threat against face authentication systems, as revealed in Chapter 4 in this dissertation. Liveness detection could be a mitigation for the threat, which is designed to distinguish between a live face and a facial image in front of the camera. The most common liveness detection mechanisms deployed on popular face authentication systems are eye-blinking and head rotation detection, which can be easily bypassed with one or two pre-catched images according to our experiments. Although sophisticated liveness detection techniques have been proposed, all of them are associated with considerable costs include requiring additional hardware, high quality images, ideal environment, and considerable user collaborations. Due to these limitations, it still remains a challenge to deploy reliable and practical liveness detection in face authentication systems for consumer-level computing devices.

Bibliography

[1] A. F. Abate, M. Nappi, D. Riccio, and G. Sabatino. 2d and 3d face recognition: A survey. *Pattern Recognition Letters*, 28(14):1885–1906, 2007.

[2] M. Abdel-Mottaleb and M. H. Mahoor. Assessment of blurring and facial expression effects on facial image recognition. In *Advances in Biometrics*, pages 12–18, 2005.

[3] A. Acquisti, R. Gross, and F. Stutzman. Faces of facebook: Privacy in the age of augmented reality. *BlackHat USA*, 2011.

[4] A. Anjos and S. Marcel. Counter-measures to photo attacks in face recognition: a public database and a baseline. In *Biometrics (IJCB), 2011 International Joint Conference on*, pages 1–7, 2011.

[5] C. Aquino. http://blog.comscore.com/2012/01/its _a_social_world.html.

[6] L. Backstrom, C. Dwork, and J. Kleinberg. Wherefore art thou r3579x?: anonymized social networks, hidden patterns, and structural steganography. In *Proceedings of the 16th international conference on World Wide Web*, pages 181–190, 2007.

[7] M. Balduzzi, C. Platzer, T. Holz, E. Kirda, D. Balzarotti, and C. Kruegel. Abusing social networks for automated user profiling. In *Proceedings of the 13th international conference on Recent advances in intrusion detection*, pages 422–441, 2010.

[8] B. Biggio, Z. Akhtar, G. Fumera, G. Marcialis, and F. Roli. Security evaluation of biometric authentication systems under real spoofing attacks. *Biometrics, IET*, 1:11–24, 2012.

[9] C. BOYLE. http://articles.nydailynews.com/2010-06-19/news/27067639_1_pierogi-facebook-page-team-mascots.

[10] W. F. Brown. The determination of factors influencing brand choice. *Journal of Marketing*, 14(5):699–706, 1950.

[11] B. Carminati, E. Ferrari, R. Heatherly, M. Kantarcioglu, and B. Thuraisingham. A semantic web based framework for social network access control. In *Proceedings of the 14th ACM symposium on Access control models and technologies*, pages 177–186, 2009.

[12] CBSNews. http://www.cbsnews.com/2100-3445_162-7323148.html.

[13] D. Centola, J. C. Gonzalez-Avella, V. M. Eguiluz, and M. S. Miguel. Homophily, cultural drift, and the co-evolution of cultural groups. *The Journal of Conflict Resolution*, 51(6):905–929, 2007.

[14] A. Chaabane, G. Acs, and M. A. Kaafar. You are what you like! information leakage through users interests. In *Proceedings of the 19th annual network & distributed system security symposium*, 2012.

[15] G. P. Cheek and M. Shehab. Policy-by-example for online social networks. In *Proceedings of the 17th ACM symposium on Access Control Models and Technologies*, SACMAT '12, pages 23–32, 2012.

[16] I. Chingovska, A. Anjos, and S. Marcel. On the effectiveness of local binary patterns in face anti-spoofing. In *Biometrics Special Interest Group (BIOSIG), 2012 BIOSIG - Proceedings of the International Conference of the*, pages 1–7, 2012.

[17] Classmates.com. http://www.classmates.com/.

[18] T. E. Coffin. A pioneering experiment in assessing advertising effectiveness. *Journal of Marketing*, 27(3):1–10, 1963.

[19] P. J. DANAHER and G. W. MULLARKEY. Factors affecting online advertising recall: A study of students. *Journal of Advertising Research*, 43(3):252–267, 2003.

[20] A. Dantcheva, C. Chen, and A. Ross. Can facial cosmetics affect the matching accuracy of face recognition systems? In *Biometrics: Theory, Applications and Systems (BTAS), 2012 IEEE Fifth International Conference on*, pages 391–398, 2012.

[21] Facebook. https://www.facebook.com/.

[22] FaceBook. http://sec.gov/Archives/edgar/data/ 1326801/000119312512034517/d287954ds1.htm.

[23] Facelock.mobi. http://www.facelock.mobi/facelock-for-apps.

[24] P. Fong, M. Anwar, and Z. Zhao. A privacy preservation model for facebook-style social network systems. In *Computer Security – ESORICS 2009*, volume 5789, pages 303–320, 2009.

[25] A. S. Georghiades, P. N. Belhumeur, and D. J. Kriegman. From few to many: Illumination cone models for face recognition under variable lighting and pose. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 23(6):643–660, 2001.

[26] Google. https://plus.google.com/.

[27] Google. http://www.android.com/about/ice-cream-sandwich/.

[28] Google. https://play.google.com/store/apps?hl=en.

[29] R. Gross and A. Acquisti. Information revelation and privacy in online social networks. In *Proceedings of the 2005 ACM workshop on Privacy in the electronic society*, pages 71–80, 2007.

[30] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker. Multi-pie. *Image and Vision Computing*, 28(5):807–813, 2010.

[31] P. Gupta, S. Gottipati, J. Jiang, and D. Gao. Your love is public now: Questioning the use of personal information in authentication. In *Proceedings of the 8th ACM SIGSAC Symposium on Information, Computer and Communications Security*, pages 49–60, 2013.

[32] A. J. Harris and D. C. Yen. Biometric authentication: assuring access to information. *Information Management & Computer Security*, 10(1):12–19, 2002.

[33] E. Hei-man TS. An ethnography of social network in cyberspace: The facebook phenomenon. *The Hong Kong Anthropologist*, 2:53–77, 2008.

[34] D. W. Hosmer Jr, S. Lemeshow, and R. X. Sturdivant. *Applied logistic regression*. Wiley. com, 2013.

[35] R. L. HOTZ. Science reveals why we brag so much. http://online.wsj.com/article/SB10001424052702304451104577390392329291890.html?mod=googlenews_wsj.

[36] W. House. http://www.whitehouse.gov/sites/default/files/microsites/ostp/PCAST/pcast_big_data_and_privacy_-_may_2014.pdf.

[37] H. Hu, G.-J. Ahn, and J. Jorgensen. Detecting and resolving privacy conflicts for collaborative data sharing in online social networks. In *Proceedings of the 27th Annual Computer Security Applications Conference*, pages 103–112, 2011.

[38] F. Hua, P. Johnson, N. Sazonova, P. Lopez-Meyer, and S. Schuckers. Impact of out-of-focus blur on face recognition performance based on modular transfer function. In *Biometrics (ICB), 2012 5th IAPR International Conference on*, pages 85–90, 2012.

[39] G. B. Huang, M. Mattar, T. Berg, E. Learned-Miller, et al. Labeled faces in the wild: A database forstudying face recognition in unconstrained environments. In *Workshop on Faces in'Real-Life'Images: Detection, Alignment, and Recognition*, 2008.

[40] A. K. Jain, A. Ross, and S. Pankanti. Biometrics: A tool for information security. *Trans. Info. For. Sec.*, 1(2):125–143, 2006.

[41] M. Johnson, S. Egelman, and S. M. Bellovin. Facebook and privacy: it's complicated. In *Proceedings of the Eighth Symposium on Usable Privacy and Security*, pages 9:1–9:15, 2012.

[42] K. Kollreider, H. Fronthaler, and J. Bigun. Non-intrusive liveness detection by face images. *Image and Vision Computing*, 27(3):233–244, 2009.

[43] S. G. Kong, J. Heo, B. R. Abidi, J. Paik, and M. A. Abidi. Recent advances in visual and infrared face recognitionła review. *Computer Vision and Image Understanding*, 97(1):103–135, 2005.

[44] N. Kumar, A. C. Berg, P. N. Belhumeur, and S. K. Nayar. Attribute and simile classifiers for face verification. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 365–372, 2009.

[45] I. Lab. https://www.idiap.ch/dataset/replayattack.

[46] K.-C. Lee, J. Ho, and D. J. Kriegman. Acquiring linear subspaces for face recognition under variable lighting. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 27(5):684–698, 2005.

[47] Lenovo. http://en.wikipedia.org/wiki/VeriFace.

[48] Y. Li, Y. Li, Q. Yan, and R. Deng. Think twice before you share: Analyzing privacy leakage under privacy control in online social networks. In *Network and System Security*, Lecture Notes in Computer Science, pages 671–677, 2013.

[49] Y. Li, K. Xu, Q. Yan, Y. Li, and R. Deng. Understanding osn-based facial disclosure against face authentication systems. In *roceedings of the 9th ACM SIGSAC Symposium on Information, Computer and Communications Security*, 2014.

[50] LinkedIn. https://www.linkedin.com/.

[51] Luxand. http://www.luxand.com/.

[52] M. McPherson, L. Smith-Lovin, and J. M. Cook. Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, 27:415–444, 2001.

[53] H. Moon and P. J. Phillips. The feret verification testing protocol for face recognition algorithms. In *Automatic Face and Gesture Recognition, 1998. Proceedings. Third IEEE International Conference on*, pages 48–53, 1998.

[54] E. Murphy-Chutorian and M. M. Trivedi. Head pose estimation in computer vision: A survey. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 31(4):607–626, 2009.

[55] L. O'Gorman. Comparing passwords, tokens, and biometrics for user authentication. *Proceedings of the IEEE*, 91(12):2021–2040, 2003.

[56] G. Pan, L. Sun, Z. Wu, and S. Lao. Eyeblink-based anti-spoofing in face recognition from a generic webcamera. In *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, pages 1–8, 2007.

[57] P. J. Phillips, P. J. Flynn, T. Scruggs, K. W. Bowyer, J. Chang, K. Hoffman, J. Marques, J. Min, and W. Worek. Overview of the face recognition grand challenge. In *Computer vision and pattern recognition, 2005. CVPR 2005. IEEE computer society conference on*, volume 1, pages 947–954, 2005.

[58] B. Renner. Curiosity about people: The development of a social curiosity measure in adults. *Journal of Personality Assessment*, 87(3):305–316, 2006.

[59] J. Rice. http://www.androidpolice.com/2012/08/03/android-jelly-beans-face-unlock-liveness-check-circumvented-with-simple-photo-editing/.

[60] C. J. V. Rijsbergen. *Information Retrieval*. Butterworth-Heinemann, 1979.

[61] SAS. http://www.sas.com/.

[62] K. M. Sheldon and B. A. Bettencourt. Psychological need-satisfaction and subjective well-being within social groups. *British Journal of Social Psychology*, 41(1):25–38, 2002.

[63] T. Sim, S. Baker, and M. Bsat. The cmu pose, illumination, and expression database. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 25(12):1615–1618, 2003.

[64] A. Sinha, Y. Li, and L. Bauer. What you want is not what you get: Predicting sharing policies for text-based content on facebook. In *Proceedings of the 2013 ACM Workshop on Artificial Intelligence and Security*, pages 13–24, 2013.

[65] T. J. Spaulding. How can virtual communities create value for business? *Electronic Commerce Research and Applications*, pages 38 – 49, 2010.

[66] F. Stutzman, R. Gross, and A. Acquisti. Silent listeners: The evolution of privacy and disclosure on facebook. *Journal of Privacy and Confidentiality*, 4(2):2, 2013.

[67] S. Trewin, C. Swart, L. Koved, J. Martino, K. Singh, and S. Ben-David. Biometric authentication on a mobile device: a study of user effort, error and task disruption. In *Proceedings of the 28th Annual Computer Security Applications Conference*, pages 159–168, 2012.

[68] Twitter. https://twitter.com/.

[69] VagueWare.com. http://www.vagueware.com/top-globally-popular-face-recognition-software/.

[70] Visidon. http://www.visidon.fi/en/Home.

[71] S. Vision. http://www.sensiblevision.com/en-us/home.aspx.

[72] K. Wagner. http://mashable.com/2013/09/16/facebook-photo-uploads/.

[73] Y. Wang, G. Norcie, S. Komanduri, A. Acquisti, P. G. Leon, and L. F. Cranor. "i regretted the minute i pressed share": a qualitative study of regrets on facebook. In *Proceedings of the Seventh Symposium on Usable Privacy and Security*, pages 10:1–10:16, 2011.

[74] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *Image Processing, IEEE Transactions on*, 13(4):600–612, 2004.

[75] D. J. Watts. *Small worlds: the dynamics of networks between order and randomness*. Princeton University Press, 1999.

[76] R. Wishart, D. Corapi, S. Marinovic, and M. Sloman. Collaborative privacy policy authoring in a social networking context. In *Proceedings of the 2010 IEEE International Symposium on Policies for Distributed Systems and Networks*, pages 1–8, 2010.

[77] L. D. Wolin and P. KorgaonkarBhat. Web advertising: gender differences in beliefs, attitudes and behavior. *Internet Research*, 13(5):375–385, 2003.

[78] J. D. Woodward. Biometrics: Privacy's foe or privacy's friend? *Proceedings of the IEEE*, 85(9):1480–1492, 1997.

[79] W. Zhao, R. Chellappa, P. J. Phillips, and A. Rosenfeld. Face recognition: A literature survey. *Acm Computing Surveys (CSUR)*, 35(4):399–458, 2003.

[80] E. Zheleva and L. Getoor. To join or not to join: the illusion of privacy in social networks with mixed public and private user profiles. In *Proceedings of the 18th international conference on World wide web*, pages 531–540, 2009.

List of Publications

Conference Papers

Yan Li, Ke Xu, Qiang Yan, Yingjiu Li, and Robert H. Deng. Understanding OSN-Based Facial Disclosure Against Face Authentication Systems. In *Proceedings of the 9th ACM Symposium on Information, Computer and Communications Security*, Japan, 2014.

Yan Li, Yingjiu Li, Qiang Yan, and Robert H. Deng. Think Twice before You Share: Analyzing Privacy Leakage under Privacy Control in Online Social Networks. In *Proceedings of the 7th International Conference on Network and System Security*, Spain, 2013.

Arunesh Sinha, **Yan Li**, and Lujo Bauer. What you want is not what you get: Predicting sharing policies for text-based content on Facebook. In *Proceedings of the 6th ACM Workshop on Security and Artificial Intelligence*, Germany, 2013.

Journal Paper Submission

Yan Li, Yingjiu Li, Qiang Yan, and Robert H. Deng. Privacy Leakage Analysis in Online Social Networks. Submitted to Journal of Computers & Security, Elsevier.