3-28-2019

Detection and Prevention of Abuse in Online Social Networks

Sajedul Karim Talukder
Florida International University, stalu001@fiu.edu

FLORIDA INTERNATIONAL UNIVERSITY

Miami, Florida

DETECTION AND PREVENTION OF ABUSE IN ONLINE SOCIAL NETWORKS

A dissertation submitted in partial fulfillment of the

requirements for the degree of

DOCTOR OF PHILOSOPHY

in

COMPUTER SCIENCE

by

Sajedul Karim Talukder

2019

To: Dean John L. Volakis
College of Engineering and Computing

This dissertation, written by Sajedul Karim Talukder, and entitled Detection and Prevention of Abuse in Online Social Networks, having been approved in respect to style and intellectual content, is referred to you for judgment.

We have read this dissertation and recommend that it be approved.

Sundaraja S. Iyengar

Deng Pan

Leonardo Bobadilla

B.M. Golam Kibria

Bogdan Carbunar, Major Professor

Date of Defense: March 28, 2019

The dissertation of Sajedul Karim Talukder is approved.

Dean John L. Volakis
College of Engineering and Computing

Andrès G. Gil
Vice President for Research and Economic Development
and Dean of the University Graduate School

Florida International University, 2019

DEDICATION

I dedicate this dissertation work to my beloved family, my wife, my son and especially my parents for their patience, love and support.

ACKNOWLEDGMENTS

It is the support from many people that brings me with the possibility to complete the dissertation for concluding my Ph.D. study. First and foremost, I would like to express my sincerest thanks and appreciation to my advisor Dr. Bogdan Carbunar, an eminent researcher in the community of security and privacy, for his inimitable support, meticulous guidance, insightful advice, and getting me immersed with an excellent research atmosphere. It is his great passion and endless patience that helped me to locate my research interest and develop the related research skills in the area of security and privacy.

Next, my thanks goes to my dissertation committee members: Dr. Sundaraja S. Iyengar, Dr. Deng Pan, Dr. Leonardo Bobadilla, and Dr. B. M. Golam Kibria, for their helpful advices, insightful comments and great support on my dissertation and future career plans.

Additionally, I am extremely grateful to my colleagues of Cyber Security and Privacy Research (CaSPR) Lab: Mahmudur, Mizanur, Mozhgan, Nestor and Ruben for their valuable insights and cooperation. I also like to thank our wonderful department staff: Olga Carbonell, Vanessa Cornwall, Ariana Taglioretti, Luis Rivera and others for their support.

Finally, I would like to express my utmost gratitude to my parents, M A Jalil Talukder and Sajeda Siddiqua, whose endless love and prayer are with me in whatever I pursue. My special gratitude goes to my beloved wife Dr. Nourin Sultana, who has made me happier, stronger, more empathetic and given me the cherished gift of her love, tears and support in my times of pain. I am also grateful to my dearest son Nushjad Karim Talukder for his unexpressible sacrifice during my Ph.D. journey. I am grateful to my beloved brother, Zahidur Rahim Talukder, who gave me inspiration at all times. Without the unlimited support from them, I would never go through any tough times in my life.

ABSTRACT OF THE DISSERTATION

DETECTION AND PREVENTION OF ABUSE IN ONLINE SOCIAL NETWORKS

by

Sajedul Karim Talukder

Florida International University, 2019

Miami, Florida

Professor Bogdan Carbunar, Major Professor

Adversaries leverage social networks to collect sensitive data about regular users and target them with abuse that includes fake news, cyberbullying, malware distribution, and propaganda. Such behavior is more effective when performed by the social network friends of victims. In two preliminary user studies we found that 71 out of 80 participants have at least 1 Facebook friend with whom (1) they never interact, either in Facebook or in real life, or whom they believe is (2) likely to abuse their posted photos or status updates, or (3) post offensive, false or malicious content. Such friend abuse is often considered to be outside the scope of online social network defenses. Several of our studies suggest that (1) perceived Facebook friend abuse as well as stranger friends are a significant problem; (2) users lack the knowledge or ability to address this problem themselves; and (3) when helped and educated, users are often willing to take defensive actions against abusive existing and pending friends, and strangers.

Motivated by the rich, private information of users that is available to the Facebook friends, often the entry point of this vulnerability is the pending friends. In an exploratory study with a number of participants, we found that participants not only tend to accept invitations from perfect strangers but can even invent a narrative of common background to motivate their choice. Further, based on our conjecture that Facebook's interface encourages users to accept pending friends, we develop new interfaces that seek to encourage users to explore the background of their pending friends and also to train them to avoid

suspicious friends. The efficacy and implementation simplicity of the proposed modifications suggest that Facebook's unwillingness to protect its users from abusive strangers is deliberate.

This dissertation explores the friend abuse problem in online social networks like Facebook. We introduce two novel approaches to prevent friend abuse problem in Facebook. (1) First, we introduce AbuSniff which can detect already existing abusive friends in Facebook, and prevent the abusive friend from doing abuse by taking some protective actions against them. (2) Second, we introduce FLock to address the problem of abuse prevention during the time of friend invitation: by educating and training the Facebook users about the abusive friend from the list of pending friend invitations, and introducing new User Interface to help users reject the potentially abusive friend invitation, thus protecting the user from abuse in advance.

TABLE OF CONTENTS

LIST OF TABLES

TABLE PAGE

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 Motivation

Social networks provide an ideal platform for abuse, that includes the collection and misuse of private user information [YS14, Yat17, KPIM11], cyberbullying [SRHF17, KS13], and the distribution of offensive, misleading, false or malicious information [CBDL17, AS10, Wei10, Aro16]. The propensity of social networks towards such abuse has brought intense scrutiny and criticism from users, media, and politicians [Smi08, AG17, Sha16, Lee17, Wak17].

Social networks like Facebook have made progress in raising user awareness to the dangers of making information public and the importance of deciding who can access it. However, many users still allow their Facebook friends to access their information, including timeline and news feed. This, coupled with the fact that people often have significantly more than 150 Facebook friends[1] – the maximum number of meaningful friend relationships that humans can manage [Dun92]) – suggests that Facebook users may still be vulnerable to attacks.

To evaluate user perception of exposure to abusive friend behaviors, we designed 2 user studies (total n = 80) where each participant had to evaluate 20 of their randomly selected Facebook friends. 65 of the 80 participants admitted to have at least 1 friend whom they perceived would abuse their status updates or pictures, and 60 of the participants had at least 1 friend whom they perceived would post abusive material (i.e., offensive, misleading, false or malicious). This is consistent with recent revelations of substantial abuse perpetrated through Facebook, including Cambridge Analytica's injection of content to change user perception [OJ18], and Facebook's admission that in the past two

[1] Participants in our studies had up to 4,880 friends, M=305.

years Russia-based operatives created 80,000 posts that have reached 126 million users in the US [BBC17a, Lee17].

Further, 55 of the 80 participants admitted to have at least 1 Facebook friend with whom they have never interacted, either online or in person. Such *stranger* friends could be bots [VFD+17] that passively collect sensitive user data and later use it against the user's best interest, as we also show through pilot study answers. This corroborates Facebook's recent estimate that 13% (i.e., 270 million) accounts are either bots or clones [Hea17]. Stranger friends can use the collected data to infer other sensitive user information [YS14], identify "deep-seated underlying fears, concerns" by companies such as Cambridge Analytica [Lap18], perform profile cloning [KPIM11], sextorsion [Yat17], identity theft [NWM10], and spear phishing [GHW+10] attacks.

These studies signal the need for defenses against abusive and stranger friends, that include restricting the abusers' access to user information, unfollowing them and even unfriending - removing them from the friend list. When asked directly, participants in our studies unfollowed and restricted access for abusers in 91.6% and 90.9% of the cases, respectively. When informed about the potential privacy risks posed by stranger friends, participants chose to unfriend or sandbox (block bi-directional communications with) such friends in 92.45% of the cases.

While Facebook has promised to take swift and decisive action to curtail attacks [Guy18, MZ.18], we observed few changes concerning the handling of friend relationships. This leaves spam friend requests as an alternative for attackers to gain access to personal data of Facebook users. To prove this point, in a first contribution, we show through user studies that little has changed since 2014 [RBJB14], and regular Facebook users are still vulnerable to spam friend requests. We further show that 8 out of 10 in-person interview participants that we recruited from random people in our city, and 26 out of 30 online participants who were recruited from the crowdsourcing website, have each accepted at

least 1 synthetic Facebook friend invitations from complete strangers that we fabricated, when asked to make a decision. There were 19 online participants who confirmed at least 1 such synthetic friend invitation even without inspecting the sender's profile.

We conjecture that the blame for confirming perfect strangers as friends, lays at least in part with the Facebook UI design for viewing and processing pending friends, that encourages users to quickly accept pending friends, especially if they are well-crafted, even though perfect strangers. We have developed new defensive user interfaces for processing pending invitations in Facebook, that encourage users to carefully inspect and question their friend invitations, and train users to avoid accepting invitations from strangers. From 3 user studies with 120 online participants, we found that when compared to a control, our solutions reduced confirmed synthetic friend invitations by up to 42.6%, reduced blind confirms by 21.9%, increased inspected synthetic friend profiles by 54.67%, and increased average decision times by 40.3%, respectively.

1.2 Problem Definition and Contributions

We observe that abuse detection and prevention mechanisms, as proposed in academic work and implemented in popular online social networks, are ineffective to ensure that the users are protected from the perceived friend abuse [Lee17, Lap18, AG17, Sha16]. In this thesis, we posit that friends and pending friends who are perceived as abusive by users, can be automatically detected, without human interaction. Further, we posit that tools can be developed to accurately predict the defenses against such friends, that users would be willing to execute.

This dissertation makes several contributions which are described in the following Sections from 1.2.1 to 1.2.8.

1.2.1 Contribution-1: Perceived Friend Abuse Questionnaire

In Section 1.1, we have introduced the idea of friend abuse problem in online social networks like Facebook. We have performed several user studies to confirm the widespread friend abuse problem in Facebook and developed a friend abuse questionnaire that captures the user perception that a Facebook friend is abusive or a stranger. Moreover, we have also devised rules to convert questionnaire answers into defense actions.

1.2.2 Contribution-2: Abuse and Defense Prediction

We have proposed and evaluated the hypothesis that data recorded by Facebook can be used to predict the user perception of friend abuse. In order to automate the implemented system, we have developed mutual activity features and used them to train supervised learning algorithms to predict user perceived friend abuse, and defensive actions that users are willing to take against the perceived perpetrators. We have also shown through our user studies, that the automatic abuse prediction and defense mechanism is efficient, practical and well-receieved by the real world Facebook users.

1.2.3 Contribution-3: Open Source AbuSniff as an Android App

We have implemented AbuSniff in Android platform as an Android app, and evaluated it through user studies with 263 participants from 25 countries and 6 continents. We have also made AbuSniff available to be downloaded from Google Play [Abu17a], and made it open source [Abu17b] so that anybody can contribute to the system and improve it further.

1.2.4 Contribution-4: Positive Influence on Participant Willingness

We have shown that the questionnaire based AbuSniff is significantly more efficient than control in terms of participant willingness to unfriend and restrict the abusive friends. We have also shown that the predictive AbuSniff can accurately predict questionnaire responses and the abusive friends that users prefer to ignore.

1.2.5 Contribution-5: Vulnerability on Stranger Friend Invitation

We conjecture that Facebook UI design for viewing and processing pending friends encourages users to quickly accept pending friends, especially if they are well-crafted, even though perfect strangers. We have confirmed through in-person and online studies, that Facebook encourages blind and rapid friend confirmations, including from perfect strangers.

1.2.6 Contribution-6: New Friend Request Interface for Facebook

We have introduced FLock, a new interface design for viewing and processing received friend invitations in Facebook, that encourages users to carefully investigate pending friends. We further introduced T-FLock, a system that trains users to be suspicious of friend spam and helps them to filter out the friend spam more effectively.

1.2.7 Contribution-7: Evaluation Scores

We have introduced scores to evaluate the ability of developed solutions to raise awareness and prevent friend spam attacks in Facebook. We have shown that careful changes to the Facebook pending friend UI can significantly improve the number of inspected profiles,

the number of blind confirms, the time to inspect and decide, and the rejection rate of strangers.

1.2.8 Contribution-8: User Decision Predictor

We have introduced and developed the first classifier to predict user decisions on pending invitations. We used k-fold cross-validation for time series for the classification. Without using the timing information, our Random Forest classifier achieved an aggregate F1 of over 65% and using the time information for an F1 of over 88%.

1.3 Related Publications

This dissertation has been written based on the following list of publications:

• Sajedul Talukder and Bogdan Carbunar. AbuSniff: Automatic Detection and Defenses Against Abusive Facebook Friends. In *Proceedings of the International AAAI Conference on Web and Social Media*, pages 385–394, 2018.

• Sajedul Talukder and Bogdan Carbunar. When Friend Becomes Abuser: Evidence of Friend Abuse in Facebook. In *Proceedings of 9th ACM Conference on Web Science (WebSci)*, 2017.

The following two manuscripts are currently under review:

• Sajedul Talukder and Bogdan Carbunar. A Study of Friend Abuse Perception in Facebook. *ACM Transactions on Social Computing (TSC)*, 2018. (In Submission).

• Sajedul Talukder, Nestor Hernandez, Mozhgan Azimpourkivi and Bogdan Carbunar. "I don't remember exactly, but ...": Vulnerabilities and Defenses to Facebook Strangers. In *The Fifteenth Symposium on Usable Privacy and Security (SOUPS)*, 2019. (In Sub-

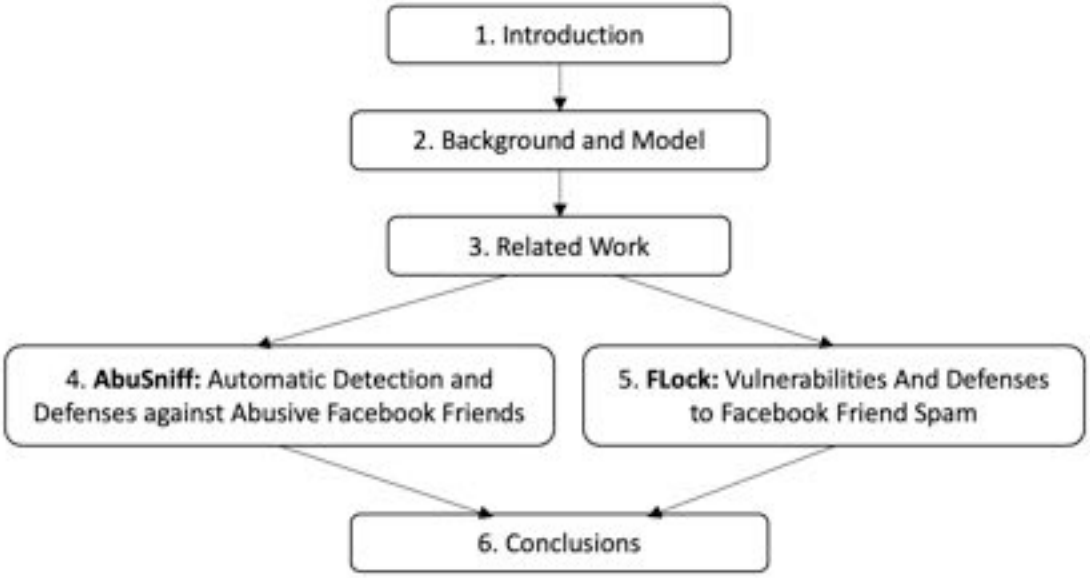


Figure 1.1: Organization of this dissertation with arrows indicating dependencies.

mission).

1.4 Organization of the Dissertation

We conclude this introductory chapter with a preview of the remainder of the dissertation. Chapter 2 presents the background and model of this dissertation. Chapter 3 presents the related work. Chapters 4 and 5 contain original contributions. Concluding remarks, open problems, and some potential avenues for future work appear in Chapter 6. The structure and dependencies between chapters are shown in Figure 1.1. The contributions of this dissertation are laid out in the following chapters as follows:

- **Chapter 2:** We discuss the background of the research of this dissertation. Section 2.1 provides the definition of the online social network. Section 2.2 discusses the history of the online social networks: CompuServe, Talkomatic, BBS, Six Degeres,

AOL, Friendster and Hi5. Section 2.3 describes the history of Facebook along with the terminologies used in Facebook. Section 2.4 presents the system model. Section 2.5 concludes the chapter by presenting the adversary model.

- **Chapter 3:** We present the literature related to this dissertation. Section 3.1 presents the literature for the cyber abuse problem. Section 3.2 presents the literature for the cyber abuse in social networks. Section 3.3 presents the literature for the cyber abuse detection and defenses. Section 3.4 presents the literature for the personalized privacy setting tools. Section 3.5 presents the literature for the industry developments. Section 3.6 presents the literature related to AbuSniff perspective. Section 3.7 presents the literature on the friend spam model. Finally, section 3.1 concludes the chapter by presenting the literature on the friend spam detection and prevention.
- **Chapter 4:** We present our work on automatically detecting and defensing against the abusive friends in Facebook. In Section 4.1, we present the introduction with the summary of contributions. In Section 4.2, we account for the restrictive actions against the Facebook friends. In Section 4.3 we present our research objectives. Section 4.4 discusses the questionnaire module, abuse inference engine and intervention module. Section 4.5 explains the predictive version of AbuSniff system with a detail discussion of the abuse prediction module and data collection module. Section 4.6 outlines the user study. Section 4.7 presents the results along with the summary of the findings. Section 4.8 discusses various aspects and limitations of the AbuSniff system. Finally, 4.9 concludes the chapter.
- **Chapter 5:** We present our work on the vulnerabilities of the Facebook stranger and spam friend invitations and their defenses. Section 5.1 presents the introduction with the summary of contributions. Section 5.2 presents the Control app, FLock system and T-FLock system. In Section 5.3, we present our methods along with

the evaluation scores and ethical considerations. Section 5.4 presents the results along with the comparisons of the Control, FLock and T-FLock, and prediction of the user decisions on the pending friends. Section 5.5 presents the study timelines from three different experiments. Section 5.6 discusses various aspects and presents some future recommendations. Section 5.7 accounts for the limitations. Finally, section 5.8 concludes the chapter. 5.9 contains the appendix that includes the questionnaire.

- **Chapter 6:** We conclude our discussion of this dissertation in this Chapter. Section 6.1 summarizes our work of the final contribution. Section 6.2 proposes some of the possible future research directions for our dissertaton work.

CHAPTER 2

BACKGROUND AND MODEL

2.1 Online Social Network Definition

There has been much controversy over the definition of an online social network and the scope of the definition. General people can easily spot and differentiate the social network from other networks, but establishing a formal definition is necessary in order to better understand it. According to Merriam-Webster, online social network is "an online service or site through which people create and maintain interpersonal relationships." [Web]. While this definition covers an important aspect of the online social network communication feature, a more broad definition encompassing various other communication aspects would be more accurate. We can define online social network as the digital communication platform which enables the users to create profile, search and connect with each other either individually or in groups, and share their thoughts, medias, and various other contents. One easy way to distinguish other forms of networks from online social networks is the social network users should have a public or a restricted profile and people should be able to search the profile if it is public. There should be some way to connect with the profile either through some indirect relationship (follow) or some direct relationship (friend).

Another important thing that comes out from this definition is that online social media should facilitate the user to communicate directly with their peers or connections, most often at real-time. This aspect of the definition differentiates the news websites, blogs and forums from the online social network. Although people can post content, comment and reply to the comments on those websites, these do not amount to a real-time direct communication. Also, another important point to be noted is that online social networks are often based on the content that their users generate, either through their communication or through their posting of contents.

2.2 History of Online Social Networks

Online social networks have become an integral part of people's life. It is simply unimaginable for many people what a world without the online social networks could be. Online social networks have dramatically changed the way we communicate with our families, friends and peers. we use online social networks to stay connected with our family members, to communicate with someone thousands of miles away, to get updates about the people we care and share our thoughts and updates with others. The whole world has come under a single umbrella and it is possible to search the huge database with billions of people to find the right people we look for. From the ancient times, people have always looked for ways to communicate with others, through pigeons, letters, telegraphs or telephones. Often these communications mediums were very time consuming compared to the modern times. Physical distance, geographical restrictions and weather conditions often hampered the older forms of communications. But the inquisitive nature of the human being led to the invention of internet in the 20th century, which radically changed the way people communicated. Thankfully, with the internet people experienced faster way to communicate irrespective of the geographical boundary. Internet enabled the people to think about smarter way of communication and invent the peer to peer communication like email and messaging application. The modern age took over the older internet communication and gave birth to the online social networks.

CompuServe. CompuServe was a file sharing service in the early internet ages in the 1970s where the users could also access news and events. At the beginning, it was a mainframe computer communication service mainly serving business-oriented solution. It had a chat application embedded in it which allowed the users to experience true interaction via exchanging messages with each other. The system also allowed the users to join thousands of discussion groups and forums and stay up-to-date with the current trends or

(a)

(b)

(c)

(d)

Figure 2.1: Screenshots of the older social networks: (a) CompuServe service. (b) Bulletin Board System. (c) Geocities. (d) Friendster.

subject matters. It was extremely popular and by the late 1980s, it was available in the public domain. Members of the forum were able to post questions or discussions, and fellow members could respond to that thread. This was a preliminary version of the concept of **"post"** like current day social networks where other users can respond by reacting or commenting.

Talkomatic. Talkomatic was the pioneer of the online chat system which was created in 1973 by Doug Brown and David R. Woolley. It was a multi-user system offering six different channels on the PLATO system that was located in the University of Illinios. The channels were similar to a "room" and each channels could support maximum five users. Talkomatic gained much popularity which was being used by a large online community.

This was one of the earliest online systems allowing **"online chatting"** that paved the way for the later modern online social networks.

BBS. Bulletin Board System, or in short BBS was the online meeting place in the late 1970s to 1980s and 1990s where the users were able to exchange messages with each other. The users could also connect to a central system using dial-up connection to the host machine which allowed them to download software, games and files. However, the connection to the BBS was limited to one person at a time. It was extremely popular to the hobbyists to discuss about various aspects of their projects. Since the connection was made over telephone lines, the cost of long-distance calling was high which often limited the system to local only users. Though the technology had various limitations, it was really popular among it's users which helped them to socialize virtually. However, the system was highly criticized for allowing the users to share virus code, illicit adult contents and various banned materials. Nonetheless, it was the beginning of the idea of current day **"timeline"**, where people can share their contents, ideas and files with other people.

Six Degrees. Six Degrees was launched in 1997 which is considered as one of the first true online social network sites. The site was based on the idea that every individual in the world is connected to each other with maximum six degrees of separation, or formally known as "six degrees of separation" theory. The site allowed the users to create individual profile, add friends the profile, search for other profiles and create groups. It was highly popular at that time which reached 3.5 million users. It introduced the concept of **"profile"** that is an integral part of current day social network like Facebook. However, there were complains against the social network that it was driving membership by allowing spam invitations. Eventually, the network was shut down in 2000 after being bought by YouthStream Media Networks in 1999.

AOL Instant Messenger. AOL Instant Messenger started it's journey in 1997. It was the first ever instant messaging service which allowed the users to chat with each other instantly. The users could also create profile which included a biography about the user. The users could also search for other profiles. It was one of the true predecessors of the modern online social networks with **"instant messaging"** capability.

Friendster. Friendster was launched in 2002 as a major competitor of Six Degrees. It was mainly intended to be a dating site but the users could create profile and add friends in their personal network. It also allowed the users to post photo, video and message and share with the connections. The connections could also post comments on the shared contents. Like the Six Degrees, it also shared a similar concept called "Circle of Friends". It was extremely popular and soon after it's start, it reached more than one hundred million users. Friendster was a true predecessor of the **"status update"** of the modern online social network like Facebook which allowed the posting of text, photo and video. However, despite it's popularity at that time, it failed to keep up with the more recent social networks and stopped it's services in 2015. It was officially closed in 2019.

Hi5. Hi5 was launched in 2003 and it soon established itself as one of the major online social networks. Like the other social networks, users can have profile in Hi5, but unlike other networks the friends are consist of direct friends (first degree), friends of friends (second degree) and friends of friends of friends (third degree). The profile can contain various personal information. It also allows the users to post status updates, upload photos and join various groups. It currently has around 60 million users in the world (mostly outside of US). This network first introduced the concept of **"public profile"**, where anyone can see the profile, and **"friends only profile"**, where only friends can see the profile.

Figure 2.2: Facebook.

2.3 Facebook

Facebook, the most influential and popular online social network in the history of mankind until now was founded in 2004 by Mark Zuckerberg, Andrew McCollom, Eduardo Saverin, Chris Hughes, and Dustin Moskovitz from a dorm room in Harvard. It first started as a campus-only networking site for the university students, which later transformed into a online social network for the general people. At the beginning, users were only allowed to join Facebook if some already existing Facebook user had sent them the invitation. However, as the popularity of Facebook grew, this restriction was removed and anyone could join Facebook without any exclusive invitation. It quickly surpassed all the other existing online social networks like MySpace, Hi5 and Friendster. Soon it became the most visited website in the world. It is now the most popular online social network in the US and in the whole world having more than 2.3 billion users. More than half of the users login to Facebook everyday and spend a substantial amount of time in Facebook. The network generates over $40 billion in revenue each year and values over $104 billion, making it one of the most important tech companies in the world.

2.3.1 Facebook Terminologies

We list the most common terminologies used in Facebook.

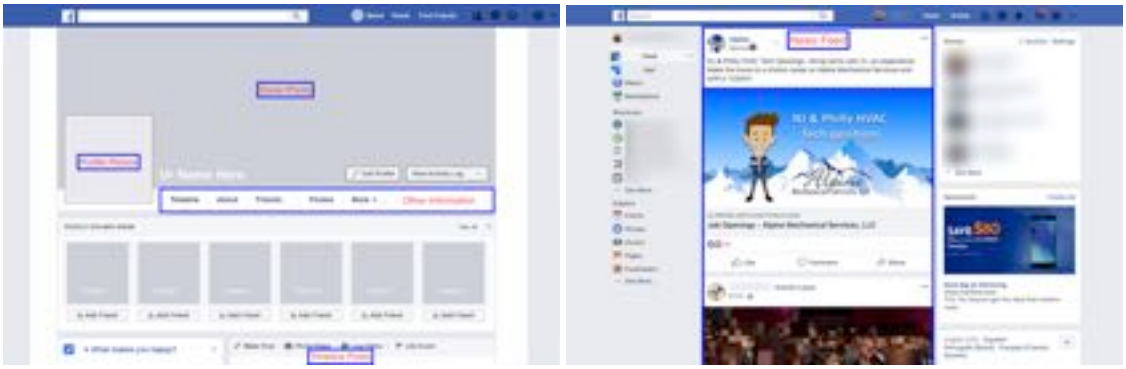
Profile. *Profile* in Facebook refers to the entire set of information that are saved in a Facebook account that contains the user photos, stories, videos, basic information, timeline, newsfeed, biography and all the Facebook friends. The profile can be private, locked or public.

Timeline. *Timeline* (a.k.a wall, or profile) is the place where the user can share her posts, photos, videos, check-ins, interests, event attendance and other activities (e.g., posting comments on a status update or picture, confirming a new friend, etc). These activities appear as *stories*, in reverse chronological order. The timeline also includes friend activities that directly concern the user, e.g., their comments, status updates, notes or pictures that reference or include the user. This sensitive information is accessible by default by the user's friends.

Newsfeed. *Newsfeed* is a constantly updating place where the user receives the updates (likes, comments, attendance, interests, status updates, stories, events) of the friends that she follows. A user's *newsfeed* shows stories created by her friends, groups, liked pages, and subscribed events. Stories are sorted based on various features, e.g., post time and type, poster popularity.

Friend. *Friend* in Facebook is a form of relationship. Each user has a *friend list* of other users with whom she has formed friend relationships. To befriend someone, a user needs to send a friend invitation and be approved. When two users become friends, they both automatically *follow* each other as well. Adding someone as a friend in Facebook enables both the users to see each other profiles and interact.

Follow. *Follow* in Facebook is a type of relationship that allows the follower to subscribe to someone's public stories on Facebook and interact with the story (like or comment).



(a) (b)

Figure 2.3: Anonymized Facebook interfaces: (a) Timeline. (b) Newsfeed.

Following someone in Facebook automatically shows her updates in the newsfeed of the follower.

Post. *Post* in Facebook refers to the story that the user shares with her followers. Post may contain only text or include images and videos along with check-ins, tags, feelings and locations. User gets notification from Facebook when there happens any interaction with that post. Posts are stored in the timeline of the user's profile.

Tag. *Tag* is a type of connection in the Facebook post that can relate another person or page with the post by showing a link to that person or page in the post. If someone tags a user in a post, the user will get notification from Facebook and her name will appear in the post.

2.4 System Model

We consider a Facebook-like social network, where users form *friend* relationships, initiated when users send friend invitations to other Facebook users. Friend invitations are stored in a *pending friend list*. Such pending friends cannot access any of the non-public information of the user.

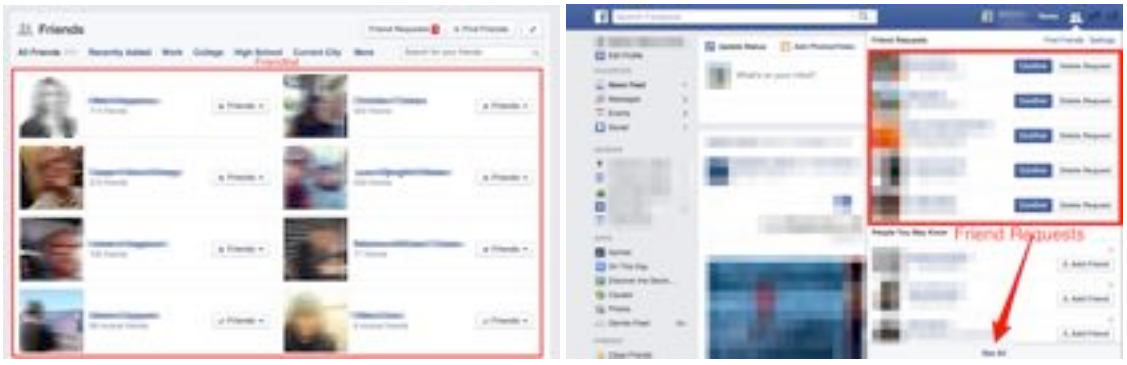
Users can inspect their pending friend lists in order to make a decision, and view the profile of any pending friend, by tapping on its entry in the pending friend list. If the account of the pending friend is public, the user will be able to access information that includes the profile photo, name, number of mutual friends, occupation, current city, country of origin, marital status, number of followers, education information, photos and recent timeline posts if any. If the profile of the pending friend is not public, the user can only see the profile photo, name, number of mutual friends, number of followers, and location.

The user can either *confirm* or *delete* a pending friend. Upon confirmation, the pending friend is moved to the user's friend list, where the friend will be able to access all the account information that the user shares with friends.

Each user has a *friend list* of other users with whom she has formed friend relationships. When two users become friends, they both automatically *follow* each other as well. Following someone lets the user receive that person's status updates show up in her *news feed*. A user's *news feed* shows stories created by her friends, groups, and subscribed events. Stories are sorted based on various features, e.g., post time and type, poster popularity.

The social network also allows users to simply follow instead of befriend another user. Following does not involve an invitation and approval process. Following someone other than a friend lets the user get that person's publicly shared status updates automatically show up in her news feed, but not vice versa.

The timeline of the user includes her updates, photos, check-ins, and other activities along with friend activities that directly concern the user, e.g., their comments, status updates, notes or pictures that reference or include the user. This sensitive information is accessible by default by the user's friends. While users can control with whom they share



(a) (b)

Figure 2.4: Anonymized Facebook interfaces: (a) Friendlist. (b) Friend Request.

each story, i.e., through the *audience selector* option, it is well known that they often use the default settings [MJB12, LGKM11, DLHH09].

2.5 Adversary Model

Friend spam. We consider an adversary who controls multiple social network accounts, a.k.a sockpuppets or Sybils, and uses them to send friend invitations to other, victim accounts. Such behavior is known as *friend spam* or *infiltration campaign* [RBJB14]. If a spam invitation is confirmed, the spammer can collect private information from the befriended victim [BBC17a, Yat17, YS14, NWM10, GHW+10], use it to infer other sensitive user information [YS14], identify "deep-seated underlying fears, concerns" by companies such as Cambridge Analytica [Lap18], inject content to change user perception [OJ18], distribute fake news, misinformation, propaganda and malware [BBC17a, Lee17, Sin14, Aro16, Wei10, AS10], and perform cyberbullying [WCZB+16, LLGL13, EN11], profile cloning [BSJ14], sextorsion [Yat17], identity theft [NWM10], and spear phishing [GHW+10] attacks.

We assume that the adversary can use bots to maintain his sockpuppet accounts and send friend spams. Facebook estimated that 13% (i.e., 270 million) of their accounts are either bots or clones [Hea17], and also admitted that Russia-based operatives created 80,000 posts that have reached 126 million users in the US [BBC17a, Lee17]. This suggests that friend spam, i.e., receiving friend invitations sent from adversary-controlled accounts, is a threat with substantial impact in social networks.

Privacy abuse. We consider adversaries who leverage the above communication mechanisms to send friend spam [GHW+10, SKV10, BHI+08, TN10] from fake, Sybil accounts to intended victims. Once victims accept friend invitations from such accounts, the attackers can collect sensitive information (profiles, photos, friend lists, locations visited, opinions) posted by friends on their timelines or take screenshots of stories. We focus on the following mechanisms to perpetrate abuse through Facebook friends: The adversary can collect personal information and then use this data to infer more sensitive information [YS14], initiate sextorsion [Yat17], perform profile cloning [KPIM11], identity theft [NWM10], and spear phishing [GHW+10] attacks.

Timeline abuse. The Facebook timeline is one of the social networking affordances that provides the dimensions of persistence, searchability, replicability and invisible audience, known to magnify bullying [Boy07]. Cyberbullying is defined as the use of the internet and devices to send or post text or images intended to hurt or embarrass another person, see e.g., [DJH+12]. In Facebook, adversarial users can post abusive replies to stories (e.g., status updates, photos) posted by friends on their own timeline. The abusive replies appear on the timeline of the victim, where the original stories were posted. Kwan et al. [KS13] have shown that the intensity of Facebook use and engagement in risky Facebook behaviors were related to Facebook victimization and bullying, respectively.

News feed abuse. The adversary posts abusive material on his timeline, which is then propagated to the news feed of his friends. Abusive information includes material per-

ceived to be offensive, misleading, false, or malicious. For instance, Facebook revealed that Russia-based operatives created over 80,000 posts that have reached 126 million US users [BBC17a,Lee17]. When studying the wall posts of 3.5 million Facebook users, Gao et al. [GHW+10] discovered more than 200K malicious posts with embedded URLs, with more than 70% pointing to phishing sites. Brown et al. [BHI+08] found that in Facebook, attackers could send context-aware email to approximately 85% of users, including those with private profiles.

This is consistent with findings that government sponsored and terrorist organizations use tweets and news feeds to distribute propaganda and misinformation [BBC17a, Lee17, Sin14, Aro16, Wei10, AS10].

The above mechanisms can be used to implement a wide range of cyber abuse. In this thesis, we focus on the perception of users concerning the willingness of their friends to perform such abuse through these Facebook mechanisms. In Section 4.4.1 we show examples of these abuse types, as real-life events reported by participants in a pilot study.

CHAPTER 3

RELATED WORK

The features provided by online services are known to influence abuse and generate negative socio-psychological effects [SRHF17]. Social networks in particular enable a diverse set of abusive behaviors, that include the adversarial collection and abuse of private information [BBC17a, Yat17, YS14, NWM10, GHW+10], cyberbullying [WCZB+16, LLGL13, EN11, QBC12], and the viral distribution of fake news, misinformation, propaganda and malware [BBC17a, Lee17, Sin14, Aro16, Wei10, AS10]. In the following we review literature on social networking abuse types relevant to this work, as well as on the automatic detection, prevention and mitigation of online abuse, and automated solutions for personalized privacy and security settings. In the following we review literature on social networking abuse types relevant to this thesis, as well as on the automatic detection, prevention and mitigation of online abuse, automated solutions for personalized privacy and security settings, friend spam models, and friend spam detection and prevention.

3.1 The cyber abuse problem

. The functionality provided by online services is known to influence abuse and generate negative socio-psychological effects [SRHF17]. Social networks in particular enable a diverse set of abusive behaviors, that include the adversarial collection and abuse of private information [BBC17a, Yat17, YS14, NWM10, GHW+10], cyberbullying [WCZB+16, LLGL13, EN11, QBC12], and the viral distribution of fake news, misinformation, propaganda and malware [BBC17a, Lee17, Sin14, Aro16, Wei10, AS10].

Wolford-Cleveng et al. [WCZB+16] have used the Partner Cyber Abuse Questionnaire and found a prevalence of 40% of victimization by cyber abuse among college students in dating relationships, with no differences in victimization of men and women. Landoll et al. [LLGL13] developed and used the Social Networking-Peer Experiences Question-

naire (SN-PEQ), and found that negative social networking experiences were associated with symptoms of social anxiety and depression, in adolescents and young adults. Elphinston and Noller [EN11] developed an 8 item questionnaire to explore the impact of involvement with Facebook on relationship satisfaction and found that Facebook intrusion was linked to relationship dissatisfaction via experiences of cognitive jealousy and surveillance behaviors.

3.2 Cyber abuse and social networks

. Cyber abuse perpetrated by friends has been considered in the past to be outside the scope of online social network defenses [SCM11]. For instance, Facebook's Immune System (FIS) [SCM11] states that "When two users are friends and the behavior of one is bothering another, ideally the two can resolve conflict [sic] without system involvement." However, [KSS+15] found that one of reason for people to fabricate, omit or alter the truth online is to avoid harassment or discrimination. Such self-censorship may defeat the "free flow of information" envisioned by Facebook [Zuc06].

Cho and Filippova [CF16] found that Facebook users apply a combination of collaborative, corrective and preventive strategies, along with information control, to address the privacy challenges encountered in their use of Facebook. Tamir and Eran [MT17] have shown that users with high perceived behavioral control are more susceptible to peer influence in adopting online privacy practices, while the intention to adopt privacy practices is correlated with the intention to influence others.

Wisniewski et al. [WIKP15] found that users whose privacy desires were met in social networks, reported higher levels of social connectedness than those who achieved less privacy than they desired. Matthews et al. [MOT+17] studied the digital privacy and security motivations, practices, and challenges of survivors of intimate partner abuse, and

conclude that the usability of and control over privacy and security functions need to be high priorities for technology creators. This research suggests that social networks would benefit and thus should be co-interested in providing the privacy levels and protections desired by their users.

3.3 Cyber abuse detection and defenses

. Quercia et al. [QBC12] found that the reasons for ending friend relations are similar in the real and online worlds, and conjectured that tools can be built to monitor online relations. Dinakar et al. [DJH+12] observed that by 2012, little attention had been devoted to automated social network solutions to detect bullying and provide interventions. To this end, they used datasets of manually annotated comments, NLP features, supervised learning, and reasoning technique, then introduced several intervention designs. Ashktorab and Vitak [AV16] conducted participatory design sessions with teenage participants to design, improve, and evaluate prototypes that address cyberbullying scenarios. They describe several design solutions proposed by the participants, and identified several subtypes of designs for the prevention of abuse, based on the perpetrator, the victim, and automated systems and bystanders.

Vitak and Kim [VK14] found that the risks perceived by US graduate students in their self-disclosure use of Facebook include fears of social rejection and of hurting other peoples' feelings, especially as posted information is persistent and shared with a diverse audience. Further, they observed that to mitigate risks, graduate students used a variety of risk management techniques that include limiting the recipients of posts, hiding friends from their news feed, and unfriending friends.

To address the challenge of obtaining large training corpora to detect abusive online content, Chandrasekharan et al. [CSSG17] proposed a Bag of Communities (BoC) tech-

nique that leverages large-scale, preexisting data from other Internet communities. They show that a dynamic BoC model achieves high accuracy after seeing 100,000 human-moderated posts, thus can be used to better deal with abusive behaviors. Further, Narwal et al. [NSL+17] proposed UnbiasedCrowd, an automated Twitter assistant that identifies text and visual bias, aggregates and presents evidence of bias to users, and enable activists to inform the public of bias, through bots.

Cao et al. [CSY+15] proposed to detect the fake accounts behind friend spam, by extending the Kernighan-Lin heuristic to partition the social graph into two regions, that minimize the aggregate acceptance rate of friend requests from one region to the other. Wu et al. [WSHY15] utilized posting relations between users and messages to combine social spammer and spam message detection. They extracted the social relations between users and the connections between messages, and used them as regularization terms over prediction results. Quercia and Hailes [QH10] leveraged information about friendly and suspicious devices that the user encounters in time, to decide if the user is the target of a friend spam attack. A different flavor of friend spam attacks was introduced by Huber et al. [HMW10], who exploit the unprotected communications between users in social networks.

AbuSniff focuses on the user perception of strangers friends, their automatic detection and defenses.

3.4 Personalized privacy setting tools

. Raber et al. [RLG16] introduced Privacy Wedges, a user interface that displays privacy settings for historical posts and enables users to meaningfully decide with whom to share social network posts. Liu et al. [LAS+16] proposed a mobile personalized privacy assistant and show that many of its recommendations on privacy settings for Android

devices were adopted by its users, and were perceived as useful and usable. Mondal et al. [MMG+16] have shown that tweet deletion was insecure in Twitter, and proposed an exposure control mechanism that eliminates information leakage via residual activities. Amos et al. [ARK15] proposed an agent that uses machine learning algorithms trained on the data collected from human participants, to detect and incriminate deceptive participants in forums and chat-rooms. While automatic, privacy setting assistants can improve the privacy and security of their users, one challenge is that not all the friends are equally trustworthy or suitable to see all the user's posts, while some may be even abusive or strangers.

3.5 Industry developments

. Industry backed technologies to solve the online abuse problem are emerging and are made publicly available. Notably, Jigsaw and Google have recently released Perspective [Per17], an API that provides developers with access to anti-harassment tools. Perspective uses machine learning to return a "toxicity" metric for any input sentence, that signals harassment, insults and abusive online speech [Gre17].

3.6 AbuSniff perspective

. We seek to develop tools that ensure that even "at risk" users are safe in Facebook. While automatic, privacy setting assistants [RLG16, LAS+16] can improve the privacy and security of their users, one challenge is that not all the friends are equally trustworthy or suitable to see all the user's posts, while some may be even abusive or strangers. Understanding and detecting user perception of abuse from friends and strangers is a first step in protecting social network users. AbuSniff is an automated victim-side approach [AV16]

to detect the user's abuse perception, that avoids the problems associated with accessing individual posts [DJH+12]. Similar to the experienced participants in the study of Vitak and Kim [VK14], our tools (1) limit the access to user data for friends perceived to be abusive, (2) hide friends perceived to post offensive, misleading, propaganda or malicious information from the news feed of the user, and (3) unfriend or sandbox friends who are perceived as strangers or who qualify for both points (1) and (2).

Unlike prior work on friend spam detection [CSY+15, WSHY15, QH10, HMW10], AbuSniff focuses on the user *perception* of strangers friends, and their automatic detection and defenses.

Further, AbuSniff can reduce the audience that needs to be considered by audience selector solutions, e.g., [RLG16], and can also be used in conjunction with tools that monitor individual social networking events and actions [DJH+12, Per17].

3.7 Friend spam models

. Previous work has explored influential factors in privacy-related behaviors of social network users. Dong et al. [DJK15] revealed that factors that impact privacy-related decisions in Google+ include the sensitivity of the shared data, the user's perceived trustworthiness of the requester, and the behavioral and psychological traits of the users. Toma [Tom14a] found that the factors that impact the evaluation of trustworthiness of other users in Facebook, include their number of friends, profile picture, number of "likes" received from other users, number of tagged images, and number of fields that were filled out in the profile. Hristova et al. [HMM14] have shown a greater profile similarity between accounts that have stronger friendship ties in Facebook. These studies suggest that people may be more willing to accept friend requests from accounts with trustworthy-looking profiles, and also similar to their own.

Rashtian et al. [RBJB14] used grounded theory to develop a friend request acceptance model that explains how various factors influence user behaviors. The identified factors include the profile picture and name, knowing the person in real life and having a common background, gender, and the number of mutual friends. Our interview study confirms that 5 years later, several of these factors remain decisive. In this paper, we instead study participants' behaviors as performed on both their actual pending friends, and on ground truth, synthetic strangers. We also design interfaces that encourage users to investigate their pending friend requests, and to identify and delete invitations received even from attractive strangers.

Rashtian et al. [RBJB14] further presented design guidelines to improve the design of the user interface for processing pending friends. The focus of our work is different, as our concern is to understand how modifications to the user interface, coupled with training, can improve user awareness when accepting friend requests in Facebook.

3.8 Friend spam detection and prevention

. Boshmaf et al. [BMBR11] and Potharaju et al. [PCNR10] were the first to demonstrate that large scale, and even automated friend spam campaigns are possible in Facebook. Potharaju et al. [PCNR10] showed that their friend spam experiments achieved an infiltration rate that ranged between 45% to 80%. Boshmaf et al. [BMBR11] developed adaptive social bots and have used them to spam and infiltrate Facebook accounts, with a success rate of up to 80%.

Sybil account detection solutions can help detect friend spam, e.g., by flagging invitations sent from detected Sybils, to be spam. Sybil detection work has often made the assumption that attackers can easily form social relationships between Sybil accounts they control, but find it hard to establish links to honest accounts [AAA+17, CSY+15,

YGKX10,YKGF08,TMLS09]. However, friend spam studies such as [RBJB14,BMBR11, PCNR10], have debunked this assumption in Facebook. Further, Yang et al. [YWW+14], showed that in Renren, Sybil accounts do not form tight-knit communities, and are well connected with honest users.

To decide if the user is the target of a friend spam attack, Quercia and Hailes [QH10] proposed to maintain information about friendly and suspicious devices that the user encounters in time.

To address this problem, Boshmaf et al. [BLS+15] leveraged the observation that spammers have no control over victim accounts. Thus, benign, victim accounts form a distinguishable classification category. They used this observation to develop a victim prediction classifier that uses features such as gender, number of friends, and time since last update. Boshmaf et al. [BLS+15] then introduced a fake social network account detection system that integrates information about predicted victims into social graph structures, and uses them to detect the fake accounts performing friend spam.

Cao et al. [CSY+15] proposed to detect fake accounts behind friend spam, by extending the Kernighan-Lin heuristic to partition the friend graph into two regions, that minimize the aggregate acceptance rate of friend requests from one region to the other. Wu et al. [WSHY15] utilized posting relations between users and messages to combine social spammer and spam message detection.

Talukder et al. introduced AbuSniff [TC18], a system that identifies Facebook friends perceived by users to be strangers or abusive, and protects the user by suggesting to unfriend, unfollow, or restrict access to information for such friends. Our focus is on pending friend invitations, where we lack a history of mutual social network activities, thus are unable to predict pending friends who would be perceived as abusive. Instead, we develop and evaluate techniques to improve user awareness and responses to friend spam, and build a classifier to predict user decisions.

CHAPTER 4

ABUSNIFF: AUTOMATIC DETECTION AND DEFENSES AGAINST ABUSIVE FACEBOOK FRIENDS

4.1 Introduction

Influential social networks like Facebook that encourage casual friendship relations, expose their users to a range of vulnerabilities and abuse, that include cyberbullying [SRHF17, KS13, KLLA12], the distribution of offensive, misleading, false or malicious information [CBDL17, AS10, Wei10, Aro16], and the collection and misuse of private information [YS14, Yat17, KPIM11]. Such abuse can be perpetrated by real friends, acquaintances, or even perfect strangers, since social network users often have significantly more than 150 friend relationships, which is the number of meaningful friend relationships that regular humans can manage [Dun92]). For instance, in [CRAD16], we observed that 75% of 68 participants declared not to remember at least one of their 20 randomly selected friends.

We conjecture that such attacks can occur through malicious users or even bots [Hea17], who infiltrate the friend lists of victims, then use social networking affordances, e.g., timelines and news feeds, to collect their sensitive information, and to actively abuse and manipulate them. Some of the "excess" friends may not be friends at all. For instance in [CRAD16], we observed that 75% of 68 participants declared not to remember at least one of their 20 randomly selected friends.

To further investigate this observation, we have evaluated the user perception of exposure to abusive friend behaviors, through 2 user studies (total n = 80) where each participant had to evaluate 20 of their randomly selected Facebook friends. 65 of the 80 participants admitted to have at least 1 friend whom they perceived would abuse their status updates or pictures, and 60 of the participants had at least 1 friend whom they per-

ceived would post abusive material (i.e., offensive, misleading, false or malicious). This is consistent with recent revelations of substantial abuse perpetrated through Facebook, including Cambridge Analytica's injection of content to change user perception [OJ18], and Facebook's admission that in the past two years Russia-based operatives created 80,000 posts that have reached 126 million users in the US [BBC17a, Lee17].

Further, 55 of the 80 participants admitted to have at least 1 Facebook friend with whom they have never interacted, either online or in person. Such *stranger* friends could be bots [VFD+17] that passively collect sensitive user data and later use it against the user's best interest, as we also show through pilot study answers. This corroborates Facebook's recent estimate that 13% (i.e., 270 million) accounts are either bots or clones [Hea17]. Stranger friends can use the collected data to infer other sensitive user information [YS14], identify "deep-seated underlying fears, concerns" by predators such as Cambridge Analytica [Lap18], perform profile cloning [KPIM11], sextorsion [Yat17], identity theft [NWM10], and spear phishing [GHW+10] attacks.

These results signal the need to develop defenses against abusive and stranger friends, that include restricting the abusers' access to user information, unfollowing them and even unfriending - removing them from the friend list. When asked directly, participants in our studies unfollowed and restricted access for abusers in 91.6% and 90.9% of the cases, respectively. When informed about the potential privacy risks posed by stranger friends, participants chose to unfriend or sandbox (block bi-directional communications with) such friends in 92.45% of the cases.

We leverage these findings to develop AbuSniff (Abuse from Social Network Friends), a system that evaluates, predicts and protects users against perceived friend abuse in Facebook. AbuSniff also builds on the findings of Grinberg et al. [GKAN17] that Facebook users are more likely to expect feedback on their posts from closer friends, and of

Toma [Tom14b] that Facebook profile cues (e.g., number of friends, tagged photographs, comments) can predict user perceived trustworthiness.

Specifically, we first develop a friend abuse questionnaire that captures the user perception that a Facebook friend (1) is a stranger, (2) would publish abusive responses to pictures and status updates posted by the user, or (3) would publish and distribute offensive, misleading, false or potentially malicious information. Further, we devise rules to convert identified abuse into defense actions.

We show that when using this questionnaire based version of AbuSniff, participants in our user studies were willing to restrict the access of friends whom they perceive would abuse their timeline, and were willing to unfollow friends whom they perceive would abuse them through their news feed. While participants were unwilling to unfriend strangers, they were more willing to sandbox them.

We also show that when compared to a control experiment performed with 27 participants, AbuSniff significantly increased participant willingness to unfriend (17% of cases vs. 1% in the control) and restrict friends (11% of cases vs. 2% in the control).

Further, in a user study with 40 participants, involving 1,200 Facebook friends, we found that without having to answer the questionnaire, participants accepted 78% of AbuSniff's recommendations for defensive actions against abusive friends and strangers. In a study with 62 participants, AbuSniff increased participant self-reported willingness to reject invitations from perceived strangers and abusers, their awareness of friend abuse implications and perceived protection from friend abuse.

To address the observed difficulty of answering a questionnaire for each Facebook friend, we propose and investigate the hypothesis that data recorded by Facebook can be used to predict the user perception of friend abuse. We introduce several *mutual activity features* that quantify the Facebook recorded interactions between a user and her friend. We then use supervised learning algorithms trained on these features to predict (1) user

answers to the questionnaire, thus user perceived strangers and friend abuse and (2) the user willingness to take defensive actions against such friends.

When using data we collected from 1,452 friend relationships (n=57), we found that supervised learning algorithms trained on AbuSniff's mutual activity features were able to predict the user answers to the questionnaire questions for friends (1) with whom the user never interacts in Facebook (F-measure = 86.6%), (2) with whom the user never interacts in real life (F-measure = 89.7%), as well as those whom the user perceives likely to (3) abuse their sensitive photos (F-measure = 75.7%), (4) abuse their status updates (F-measure 69.2%), and (5) post offensive, misleading, false, or malicious content (F-measure = 78.0%). Further, AbuSniff was able to predict the cases where the users chose to ignore the suggested defensive action against friends, with an F-Measure of 97.3%.

In this chapter we propose AbuSniff (Abuse from Social Network Friends), a system that leverages the above findings to evaluate, predict and protect users against perceived friend abuse in Facebook. AbuSniff has the potential to mitigate the effects of abuse, and reduce its propagation through social networks and its impact on social processes (including electoral). AbuSniff has the potential to mitigate the effects of friend based abuse, and reduce its propagation through social networks and even its negative impact on social processes (e.g., electoral). In summary, we introduce the following contributions:

- We develop a friend abuse questionnaire that captures the user perception that a Facebook friend (1) is a stranger, (2) would publish abusive responses to pictures and status updates posted by the user, or (3) would publish and distribute offensive, misleading, false or potentially malicious information. We introduce rules to convert identified abuse into defense actions.
- We conjecture that data recorded by Facebook can be used to predict the user perception of friend abuse. We introduce mutual activity features that quantify the Facebook recorded interactions between a user and her friend. We use supervised

learning algorithms trained on these features to predict (1) user answers to the questionnaire, thus user perceived strangers and friend abuse and (2) the user willingness to take defensive actions against such friends.

- We have implemented AbuSniff in Android, and evaluated it through user studies with 263 participants from 25 countries and 6 continents. AbuSniff can be downloaded from Google Play [Abu17a], and is open source [Abu17b].
- We show that that the questionnaire based AbuSniff is significantly more efficient than control in terms of participant willingness to unfriend and restrict friends. The predictive AbuSniff can accurately predict questionnaire responses and the abusive friends that users prefer to ignore.

4.2 Restrictive Actions Against Friends

AbuSniff leverages several defense mechanisms provided by Facebook to protect the user against strangers and abusive friends: **unfollow** – stories subsequently posted by the friend in his timeline no longer appear in the user's news feed, **restrict** – stories published by the user in her timeline no longer appear in the friend's news feed, and **unfriend** – remove the friend from the user's list of friends.

Further, we introduce the **sandbox** defense option, a combination of unfollow and restrict: the user and her friend no longer receive stories published by the other. Unlike unfriending, sandboxing will not remove the user and her friend from each other's friend lists.

4.3 Research Objectives

In this chapter we focus on the following key questions that concern friend based abuse:

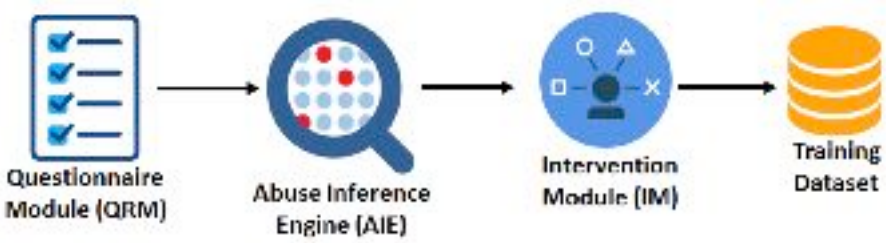


Figure 4.1: Questionnaire based AbuSniff architecture. The QRM module delivers the questionnaire. The AIE module uses the output of QRM to identify abusive friends, and the IM module asks the user to take a protective action against them. The output of the QRM, AIE and IM modules is stored for training, and is later used by the predictive AbuSniff (Section 4.5).

- **(RQ1)**: Are perceived strangers and friend abuse real problems in Facebook?
- **(RQ2)**: Are Facebook users willing to take defensive actions against abusive friends?
- **(RQ3-a)**: Does AbuSniff impact the willingness of users to take defensive actions on Facebook friends?
- **(RQ3-b)**: Is the willingness of users to take defensive actions impacted by the type of abuse perpetrated by the friend and the suggested defensive action?
- **(RQ4)**: Can AbuSniff predict abusive friends and the defenses that users are willing to take against such friends?
- **(RQ5)**: Does AbuSniff impact user awareness of stranger and abusive friends, and their perception of safety from such friends?

In order for AbuSniff to be relevant, RQ1, RQ2 and RQ5 need to have positive answers. The answer to RQ3-a and RQ3-b will impact the design of AbuSniff, while a positive answer to RQ4 will indicate that systems can be built to automatically defend users against friend abuse.

4.4 Questionnaire Based AbuSniff

We have designed the AbuSniff system to help us investigate these questions. AbuSniff is a mobile app that asks the user to login to her Facebook account. As illustrated in Figure 4.1, AbuSniff consists of modules to identify abusive friends and recommend defensive actions. The *questionnaire module* (QRM) delivers to the user a set of questions about each of her evaluation friends. The *abuse inference engine* (AIE) converts answers to the questionnaire into actions.

The *data collection module* (DCM) collects user responses as well as Facebook information about each of the user's evaluation friends. The *abuse prediction module* (APM) leverages data collected by DCM and supervised learning to predict the outcome of (and replace) the QRM and IM modules.

The *intervention module* (IM) displays the actions decided by the AIE and asks the user to confirm them. The output of the QRM, AIE and IM modules is stored in a training dataset, which is used by the predictive AbuSniff (see Section 4.5). In the following, we detail each module.

4.4.1 The Questionnaire Module (QRM)

We have designed a questionnaire intended to capture the user perception of (potentially) abusive behaviors from friends in Facebook. Since Facebook users tend to have hundreds and even thousands of friends, we decided to present the questionnaire for each of only a randomly selected subset of the user's friends. One design goal was that the questions should help identify the perceived use of the abusive mechanisms listed in the adversary model. To ensure a simple navigation of the questionnaire, we further sought to fit all the questions on a single screen for a variety of popular smartphones. We have designed the questionnaire through an iterative process that included a focus group and a pilot study

Figure 4.2: Anonymized screenshots of the Android AbuSniff app: (a) QRM questionnaire. The first two questions identify stranger friends, questions 3 and 4 identify perceived timeline abuse and question 5 identifies perceived news feed abuse. (b) The IM UI asking the user to unfriend an abusive friend also explains the reasons for the action, according to the questionnaire responses. (c) The IM UI asking the user to explain the reasons for the unwillingness to unfriend in the previous screen. (d) The "unfriend or sandbox" UI for privacy abuse: sandboxing isolates but does not unfriend or notify the friend. (e) The UI of the autonomous AbuSniff asking user confirmation to restrict the access of a friend predicted to be a timeline abuser.

with 2 K-8 teachers, 1 psychologist, 8 students, 1 dentist and 1 homemaker (8 female and 5 male).

Figure 4.2(a) shows a snapshot of the resulting questionnaire, that consists of 5 questions. The first two questions (Q1) (*How frequently do you interact with this friend in Facebook*) and (Q2) (*How frequently do you interact with this friend in real life*) determine the user's frequency of interaction with the friend, on Facebook and in real life. The options are "Frequently", "Occasionally", "Not Anymore" (capturing the case of estranged friends), "Never" and "Don't Remember". We are particularly interested in the "Never" responses.

After answering "Never" for Q1 for a friend, participants in the focus group explained that they have never initiated conversations with the friend and are either not aware of or interested in communications initiated by the friend, e.g.,

"I never did chat with him, he never commented on my photos or any shared thing. He never puts a like [sic]."

"I never like or comment on his post, I never chat with him. [..] Actually I do not notice if he likes my posts. But I do not do [sic] any interaction."

For question Q2, participants agreed that they have never met in real life friends for whom they answered "Never". Reasons for accepting the friend invitations from such friends include *"he is a friend of my friend and my friend met him in real life"*, and *"she is from my same [sic] college"*. This suggests that friends with whom the user has never interacted in Facebook and in real life, may be *strangers*. Such strangers may exploit Facebook affordances (e.g., claim college education) to befriend victims.

The next two questions identify perceived timeline abusers, i.e., (Q3) *This friend would abuse or misuse a sensitive picture that you upload* and (Q4) *This friend would abuse a status updated that you upload.* The possible responses are "Agree", "Disagree" and "Don't Know". After answering "Agree" for Q3, participants shared several stories of abuse, e.g.,

"Once this friend has downloaded my photo and then opened a fake Facebook account, like with that picture.", and

"This friend has posted a bad comment in one of my photos. That was my wedding photo. I felt so offended."

Participants who answered "Agree" for a friend on Q4 shared other stories of abuse, e.g.:

"This friend posted a bad comment on my post and from that post there was other bad stuff posted on my wall.", and

"Once I posted a sad status update because I was feeling frustrated. But this friend then posted a trolling comment on my post."

The last question (Q5) *This friend would post offensive, misleading, false or potentially malicious content on Facebook* identifies perceived news feed abusers. Stories shared by participants who answered "Agree" on Q5 include:

"This friend bothered friends by bad posts [..] The posts were against my own ideas [sic].", and

"I have often seen this friend sharing fake news. Sometimes she posts so much bogus stuff that my news feed gets flooded."

These examples show that privacy and security abuses occur in the real life interactions of Facebook users and their friends. The following AbuSniff modules seek to predict the user perception of abuse and convert it into defensive actions that users will consider appropriate.

4.4.2 The Abuse Inference Engine (AIE)

AbuSniff seeks to provide safe social interactions to regular social network users. To this end, the abuse inference engine (AIE) takes as input the responses collected by the QRM or predicted by the APM, and outputs suggested actions from the set { "unfriend", "unfollow", "restrict access", "sandbox", "ignore"}.

	Q1	Q2	Q3	Q4	Q5	Action
1	Never	Never	!Agree	!Agree	!Agree	Unfriend/ Sandbox
2	Never	Never	$*$	$*$	$*$	Unfriend
3	Never	!Never	Agree	Agree	Agree	Unfriend
4	!Never	Never	Agree	Agree	Agree	Unfriend
5	Never	!Never	Agree	!Agree	Agree	Unfriend
6	Never	!Never	!Agree	Agree	Agree	Unfriend
7	!Never	Never	Agree	!Agree	Agree	Unfriend
8	!Never	Never	!Agree	Agree	Agree	Unfriend
9	!Never	!Never	Agree	Agree	Agree	Unfriend
10	!Never	!Never	Agree	!Agree	Agree	Unfriend
11	!Never	!Never	!Agree	Agree	Agree	Unfriend
12	!Never	!Never	Agree	Agree	!Agree	Restrict
13	!Never	!Never	Agree	!Agree	!Agree	Restrict
14	!Never	!Never	!Agree	Agree	!Agree	Restrict
15	!Never	!Never	!Agree	!Agree	Agree	Unfollow
16	$*$	$*$	$*$	$*$	$*$	NOP

Table 4.1: Set of rules to convert questionnaire responses to defensive actions. Like firewall filters, the first matching rule applies. $!A$ denotes any response different from A. NOP = no operation.

Specifically, AbuSniff (1) limits the access to user data for friends perceived to be abusive, (2) hides posts from friends perceived to post offensive, misleading, propaganda or malicious information from the news feed of the user, and (3) unfriend or sandbox friends who are perceived as strangers or who qualify for both points (1) and (2).

We have used intuition to develop a preliminary set of rules for AIE, see Table 4.1. The rules are applied on a first match basis: rule r is evaluated only if all the rules 1 to r - 1 have failed.

Intuitively, the first 15 rules detect restrictive actions. For instance, rule 1 suggests that a stranger, non-abusive friend should be either unfriended or sandboxed. Rule 2 however applies to stranger friends who are perceived as also abusive: at least one of the questions Q3-Q5 has been answered with "Agree". Rule 2 suggests that such a friend should be unfriended without option for sandboxing. Rules 3-11 apply to non-stranger

friends who answered "Agree" on at least 2 of Q3-Q5. Since such friends are perceived as bi-directional abusers (i.e., capable to abuse at least some of the information posted by the user, and also to post abusive information on their own), the AIE module suggests unfriending, i.e., cutting bi-directional ties.

Initially, we considered a hard stance against abuse: a friend who scores negatively on any 2 out of the 5 questions (i.e., assigned "Never" in any of the 2 first questions, "Agree" in any of the last 3 questions) should be unfriended (rules 1-11). We have later relaxed rule 1, to also allow sanboxing of such friends.

AIE outputs less restrictive actions against friends with whom the user has interacted both in Facebook and in real life, and is either only a timeline abuser (restrict, rules 12-14) or only a news feed abuser (unfollow, rule 15). If none of the first 15 rules matches, the last rule decides that the friend is not abusive (i.e., ignore). We evaluate and adjust these rules in the evaluation section.

4.4.3 The Intervention Module (IM)

To help us answer the key research questions RQ2, RQ3-a and RQ3-b we have designed a user interface that asks the user to take a defensive action against each friend detected as abusive by the AIE module. The action, i.e., unfriend, restrict, unfollow, is determined according to the rule matched in Table 4.1.

Figure 4.2(b) shows a snapshot of the "unfriend" recommendation. The UI further educates the user on the meaning of the action, and lists the reasons for the suggestion, based on the questionnaire responses that have matched the rule, see Figure 4.2(a).

The user is offered the option to accept or ignore the suggestion. If the user chooses to ignore the suggestion, the IM module asks the user (through a PopupWindow) to provide a reason, see Figure 4.2(c). We have conducted a focus group with 20 participants in order

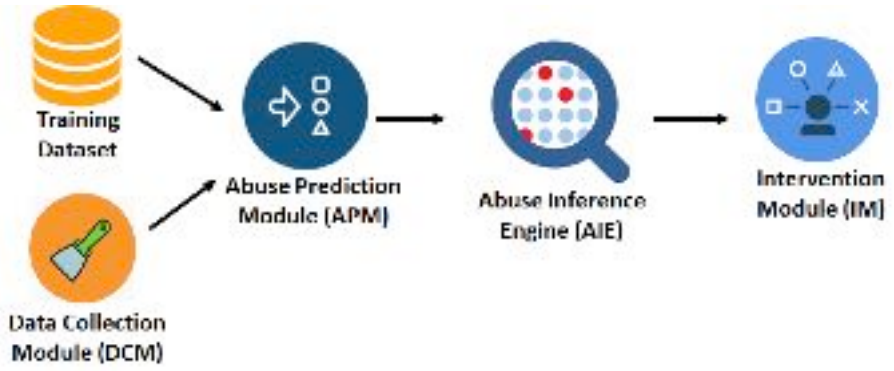


Figure 4.3: Predictive AbuSniff system architecture. The DCM module collects Facebook data concerning the relationship between the user and each friend. The APM module uses this data, and training data collected by the questionnaire based AbuSniff (Section 4.4), to predict the user responses to the questionnaire. The AIE module is inherited from the questionnaire based AbuSniff, but uses the output of APM instead of QRM to identify abusive friends. The optional IM module asks the user to confirm the predicted action against detected abusive friends.

to identify possible reasons for ignoring "unfriend" recommendations. They include "the suggestion does not make sense", "I agree, but I want to unfriend later", "I agree but I am still unwilling to unfriend", and "I don't want this friend to find out that I unfriended", see Figure 4.2(c).

The IM module educates users about the meaning and dangers of having a stranger as a friend, see Figure 4.2(d). It also offers the option to "sandbox" such friends. According to the rules of Table 4.1, IM also suggests unfollowing or restricting friends who are abusive in only one direction of their communications. Figure 4.2(e) shows a snapshot of the restrict screen, its meaning and reasons for selection.

4.5 Predictive AbuSniff

The questionnaire based AbuSniff requires the user to manually evaluate each friend. However, the average number of friends per Facebook user is 338, while the median is 200 [Smi18]. To address this problem, we propose to use a supervised learning approach,

to automatically predict the friends perceived to be abusive and the defenses that users are willing to implement against them. However, for this approach to work, we need to provide an affirmative answer to question RQ4 (*can AbuSniff predict the abusive friends and the defenses that users are willing to take against such friends?*).

To this end, we introduce the predictive AbuSniff system, see Figure 4.3. The predictive AbuSniff replaces the QRM module with an Abuse Prediction Module (APM). APM uses training data collected through the questionnaire based AbuSniff (see Figure 4.1 and Section 4.7.3) and data collected by the *data collection module* (DCM), to predict the outcome of the QRM module. In the following we detail the APM and DCM modules.

4.5.1 The Abuse Prediction Module

We introduce several *mutual activity* features, based on the Facebook data shared by a user U and a friend F, and use them to evaluate the ability to predict questionnaire responses and user decisions. Specifically, the features are (1) **mutual post count**: the number of stories posted by either U or F, on which the other has posted a comment, (2) **common photo count**: the number of photos in which both U and F are tagged together, (3) **mutual friend count**: the number of common friends of U and F, (4,5) **same current city and hometown**: boolean values that are true when U and F live in the same city and are from the same place, (6,7) **common study and work count**: the total number of places where U and F have studied and were employed together, respectively.

The abuse prediction module (APM) uses supervised learning algorithms trained on these features, and previously collected questionnaire responses and user decisions, to predict the user's answers to the QRM questionnaire and the user's reactions to suggested actions. Specifically, for each user U and friend F, APM stores a tuple that consists of (1) the mutual activity feature values of U and F, (2) U's responses to the 5 questions for

F, and (3) the suggested action for F – ignore (safe friend), unfriend, unfollow, restrict, sandbox, and (4) the action taken by U for F.

The intuition behind using these features, is that the quantity and quality of shared Facebook activities with a friend may determine whether the user has interacted with that friend in Facebook and in real life. In Section 4.7.3 we evaluate the ability of these features to predict questionnaire answers and user decisions on suggested actions.

4.5.2 The Data Collection Module (DCM)

The Data Collection Module (DCM) collects Facebook data from the user and her evaluation friends, as well as user provided input (e.g., responses from the QRM, choices from the IM) and timing information. AbuSniff uses this data to make local decisions and partially reports it to our server for evaluation purposes. In Section 4.6.1 we discuss ethical considerations of the data collection process.

This task is made challenging by the restrictions imposed by Facebook in the Graph API v2.0 and above. This policy enables an installed app to collect data from the user's Facebook account, including gender, birthday and the current city, but prevents the app to even retrieve the full list of a user's friends or their Facebook ids. The Facebook friends API endpoint returns information only from the friends who are using the same app (i.e., AbuSniff) and who have specifically granted permission for the app to see their data using the *user_friends* permission. This is done in order to protect the privacy of users, a big step forward from early day Facebook policies that enabled crawlers to collect detailed data of millions of users.

To address this challenge, we have leveraged the observation that Facebook's app policy allows JavaScript injection into the HTML source page itself. We have then dynamically created different Facebook URLs for the information that we seek to collect.

The URLs are loaded in WebView, an embedded browser wrapper around the WebKit rendering engine, which can be used to display web pages inside Android applications. We then developed JavaScript code that fetches the HTML source contents of the Facebook page into a Java string. AbuSniff injects this code into the WebView at runtime, i.e., through the webview.loadUrl ("javascript:*code*") method, where "*code*" is the Javascript code.

We use this process to extract the features of the APM module: retrieve first the user's *Friends* page, that contains the Facebook ids of the friends, then for each evaluation friend, collect their *About* and *Mutual Friendship* pages. We use regular expressions to extract the data required to build the APM features. This process is fast: each page takes approximately 1.5s to fetch and process, for around 3s per friend. AbuSniff performs this process in the background, e.g., while the user is answering the QRM questionnaire.

4.6 User Study

We have conducted several user studies to answer our key research questions. In the following we describe the participant recruitment procedure, the experiment design, and techniques we used to ensure data quality.

We expect a positive answer for RQ1 (*is friend abuse a problem in Facebook?*, as previous studies have shown that Facebook users are aware of friend spam (see e.g., [CRAD16]).

We have devised a suite of user experiences that attempt to convince users to protect themselves against dangerous friends. We have performed user studies to evaluate and improve on the willingness of users to take defensive actions against abusive friends. We observed that users are not only unaware of the risks undertaken by keeping potentially abusive friends, but, if inappropriately approached, may also be unwilling to defend against such friends. This reveals the importance of identifying not only convincing

Figure 4.4: More (anonymized) screenshots: (a) Definitions of the suggested actions. (b) Attention check screen. (c-e) Pre-test survey questions. (f) Control study screenshot for suggested actions for sample selected friend. (g) Bogus friend example. (h-j) Post-test survey questions.

reasons but also the appropriate defensive actions against the different types of abusive friends.

We have recruited 325 participants from JobBoy [Job17], during 7 studies conducted between August 2016 and October 2017. The jobs we posted asked the participants to install AbuSniff from the Google Play store, use it to login to their Facebook accounts and follow the instructions on the screen. A participant who successfully completes the app, receives on the last screen a code required for payment.

In the first session we recruited 20 participants over 25 days, 60 in session 2 over 45 days, 95 in session 3 over 52 days, and 52 in session 4 over 15 days. The sessions were set around 3 weeks apart. Second, we have recruited 98 participants between June and

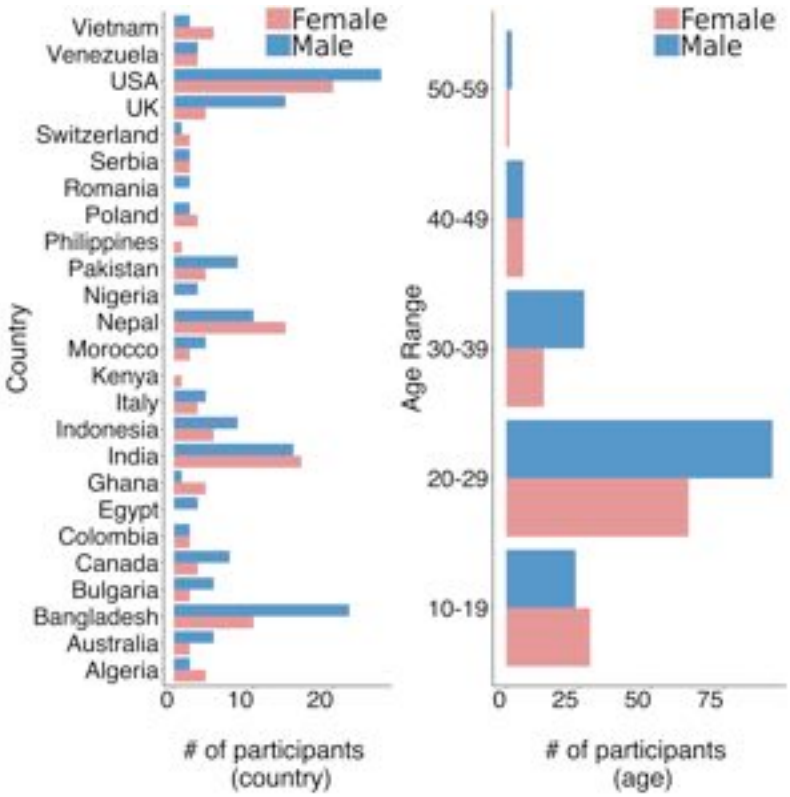


Figure 4.5: Participant demographics. (country) Distribution of the 25 countries of residence by gender. US, Bangladesh and India are the top 3 countries of residence. (age) Distribution of age range by gender. A majority of the 151 male and 112 female participants are 20-29 years old; 15 are 40-59 years old.

July 2017, during 3 additional sessions, as follows. In a control session we recruited 30 participants over 6 days, in a pre-test survey study we recruited 36 participants over 8 days, and in a post-test study we recruited 32 participants over 9 days.

We have paid each participant $3, with a median job completion time of 928s (SD = 420s).

We have only recruited participants who have at least 30 Facebook friends, had access to an Android device, and were at least 18 years old. Further, we have used the following mechanisms to ensure the quality of the data collected.

• **Attention-check screen**. To ensure that the participants pay attention and are able to understand and follow simple instructions in English, AbuSniff includes a standard

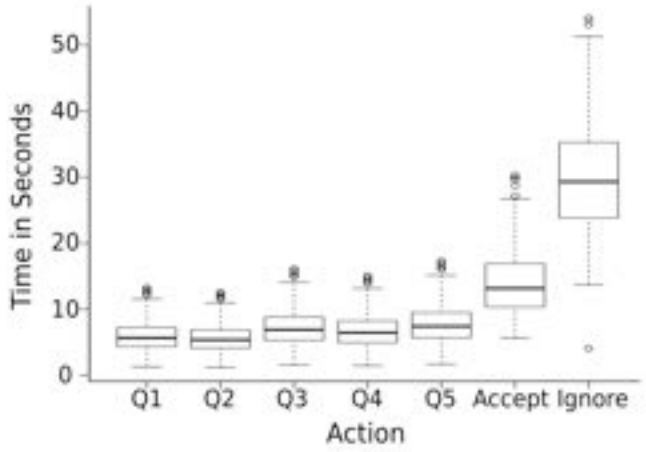


Figure 4.6: Distribution of times taken by participants in the first 2 user studies to answer the questionnaire, and decide whether to accept or ignore an action suggested against an abusive friend. The median time to answer any of the 5 questions exceeds 4.11s, with a maximum time of 17.29s. Participants have taken significantly more time to ignore a suggestion (M = 29.30s, SD = 9.86) than to accept it (M = 13.14s, SD = 4.71). The times suggest deliberation, not random choices.

attention-check screen at the beginning of the app, (see Figure 4.4(b) for a snapshot). Specifically, the screen states that "AbuSniff will help us understand how people perceive their friends in Facebook and how they make decisions about their friends. For instance, we are interested whether you take the time to read the directions. To show that you have read the instructions, please ignore the question below and just choose "none of the above" option. Please check all the words that describe how you feel now". The last of the 10 options is "none of the above". Less than 10% of the participants in our studies have failed this test. We have discarded all the data from these participants.

• **Bogus friends**. To detect participants who answer questions at random, we used "bogus friends": 3 fake identities (2 female, 1 male) that we included at random positions in the AbuSniff questionnaire (see Figure 4.4(g) for a bogus friend screenshot). We have discarded the data from participants who answered Q1 and Q2 for the bogus friends, in any other way than "Never" or "Don't Remember".

Study	Questions	Task	Modules	User #	Days	Results
1	RQ1, RQ2, RQ3-a	Explore problem	QRM, DCM, AIE, IM	20	25	Participants willing to defend against timeline and news feed abuse but unwilling to unfriend strangers
2	RQ1, RQ2, RQ3-b	Sandbox effect	QRM, DCM, AIE, IM	60	45	Participants willing to sandbox strangers
3	RQ3-a	Control study	simplified IM	27	6	Without questionnaire and explanations, participants are seldom willing to defend against a friend
4	RQ4	Data collection & Prediction eval.	QRM, DCM, AIE, IM	54	52	Mutual activity features accurate in predicting answers to Q1-Q4 and the "ignore" action but less accurate in predicting Q5
5	RQ4	Prediction eval.	DCM, APM, AIE, IM	40	15	78% of AbuSniff suggestions accepted by participants
6 & 7	RQ5	Pre-test vs. AbuSniff+Post-test	QRM, DCM, AIE, IM	31, 31	8,9	AbuSniff increases participant (1) willingness to reject invitations from strangers and abusers, (2) understanding of abuse and (3) perceived safety

Table 4.2: Summary of user studies: The research questions investigated, the task performed, the modules used, the number of participants, the duration in days of the study, and results.

• **Timing information**. We have measured the time taken by participants to answer each questionnaire question and to make a decision on whether to accept or ignore the suggested action. We have discarded data from participants whose average response time was below 3s. Figure 4.6 shows the distribution of the response times over the remaining participants. For the remaining participants, the total time taken to answer the 5 questions for a friend ranged between 7s and 74s (M = 32s, SD = 12.046). The time taken to make a decision ranged between 4 to 54s (M = 15s, SD = 8.960). Participants took significantly longer to ignore a suggestion (M = 29.30s, SD = 9.86) than to accept it (M = 13.14s, SD = 4.71). These numbers suggest that participants have carefully considered AbuSniff's questions and suggestions, and have not randomly browsed through the app.

We have used these mechanisms to discard 62 of the recruited 325 participants. The following results are shown over the remaining 263 participants. Figure 4.5 shows the distribution of the country of origin (left) and age (right), by gender, over these participants. The 151 male and 112 female participants are from 25 countries (top 5: US, Bangladesh, India, Nepal and UK) and 6 continents, and are between 18-52 years old (M = 23, SD = 7.22).

Table 4.2 summarizes our user studies, including the research questions they investigate, the task they perform, the AbuSniff modules they use, the number of participants

(after discarding the above 62), the number of days over which the experiment took place and the results which we further detail in the following section.

4.6.1 Ethical Data Collection

We have developed our protocols to interact with participants and collect data in an ethical, IRB-approved manner (Approval #: IRB-16-0329-CR01). The 54 participants from whose friends we collected mutual activity features, were made aware and approved of this data collection step. We have collected minimalistic Facebook data about only their investigated friend relationships. Specifically, we have only collected the counts of common friends, posted items, studies and workplaces, and boolean values for the same current city and hometown, but not the values of these fields. Further, we have only collected anonymized data, and the automated AbuSniff version *never* sends this data from the user's mobile device. AbuSniff only uses the data to make two predictions (the type of abuse and whether the user will take the suggested action, then erases the collected Facebook data.

4.7 Results

In Section 4.7.1 we first use the data collected in the first 2 user studies to evaluate the perception of friend abuse in Facebook. We then use the data from the first user study to evaluate the willingness of users to defend against friends that they perceive as abusive. We use the data from the second user study to evaluate the impact of the "sandbox" option on the willingness of participants to take a weaker but still effective defensive measure against stranger friends. In Section 4.7.2 we compare AbuSniff against a control experiment in order to understand if AbuSniff had an effect on the willingness of users to take defensive actions for friends, by comparing it against a control study. In Section 4.7.3

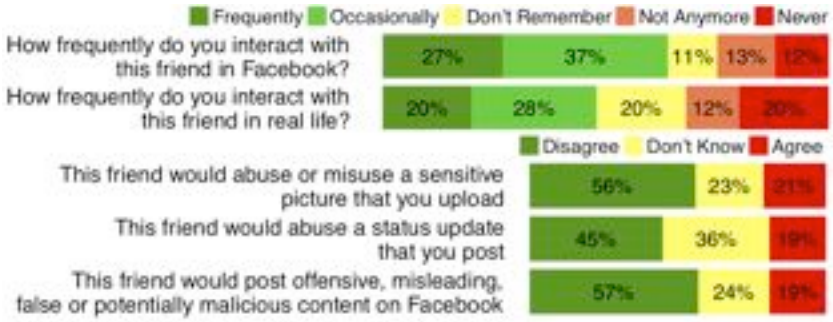


Figure 4.7: Distribution of responses for the friend abuse questionnaire over 1,600 Facebook friend relationships. From top to bottom: Q1: frequency of Facebook interaction, Q2: frequency of real world interaction, Q3: friend would abuse posted sensitive picture, Q4: friend would abuse status update post, and Q5: friend would post offensive, misleading, false or potentially malicious content. The red sections correspond to potential strangers or abusive friends.

we use the collected user data to evaluate the ability of the APM module to predict the questionnaire answers and defense actions. In Section 4.7.4 we evaluate the accuracy of an automated AbuSniff in a separate user study. In Section 4.7.5 we evaluate the user perception of the impact of AbuSniff.

4.7.1 Studies 1 & 2: Abuse Perception and Willingness to Defend

We developed 2 preliminary studies (n = 20 and n = 60) to evaluate the extent of the user perception of stranger friends and friend abuse in Facebook (RQ1) and the willingness of users to accept defensive actions against friends considered to be abusive (RQ2, RQ3-a and RQ3-b). To this end, we used the questionnaire based AbuSniff of Section 4.4: The AbuSniff app randomly selected 20 Facebook friends for each participant, asked the participant to answer the questionnaire for each friend, then asked the participant to take a defensive action against the friends detected to be abusive, or provide a reason for ignoring the suggested action. AbuSniff then collected and sent to our server, the questionnaire answers for the 20 friends, the decisions taken for the abusive friends, and the reasons provided for ignoring the suggestions.

Figure 4.7 shows the distribution of the responses for each of the 5 questions from the 1,600 friend relationships (20 from each participant). The top bar shows that in 12% of the 1,600 friend relationships, the participants stated that they have never interacted with that friend in Facebook. Further, 64 of the 80 participants stated that they have at least one friend with whom they have never interacted in Facebook.

The second bar from the top shows that in 20% of the 1,600 friend relationships, the participants stated that they have never interacted with the corresponding friend in real life. 73 of the participants had at least one friend with whom they have never interacted in real life .

In 21% of the 1,600 friend relationships, participants stated that the queried friend would abuse a photo they post (third bar), in 19% of the cases they admit the friend would abuse their status updates (fourth bar), while in 19% of the cases, they admit that the friend would post offensive, misleading, false or potentially malicious content (bottom bar). 68 of the participants had at least 1 friend whom they perceived would abuse their photos, 62 of the participants have at least 1 friend who would abuse their status updates, and 62 have at least 1 friend who would post abusive content.

Gender and age impact. In terms of having at least 1 friend perceived as abusive, Chi-square tests revealed no significant difference between genders on any of the 5 questions. Similarly, Chi-square tests revealed no significant differences between the age groups of under 30 years old and above 30 years old participants (61 vs 19 participants), on questions 1, 2 and 4. However, participants under 30 are significantly more likely (χ^2 = 4.417, df = 1, p = 0.03) to have at least 1 friend whom they perceive would abuse a photo they post, than participants over 30 (52 out of 61 vs 12 out of 19). Younger participants were also more likely to answer that they have at least 1 friend who would post offensive, misleading, false or potentially malicious content (50 out of 61 vs 10 out of 19, χ^2 = 6.64, df = 1, p = 0.01).

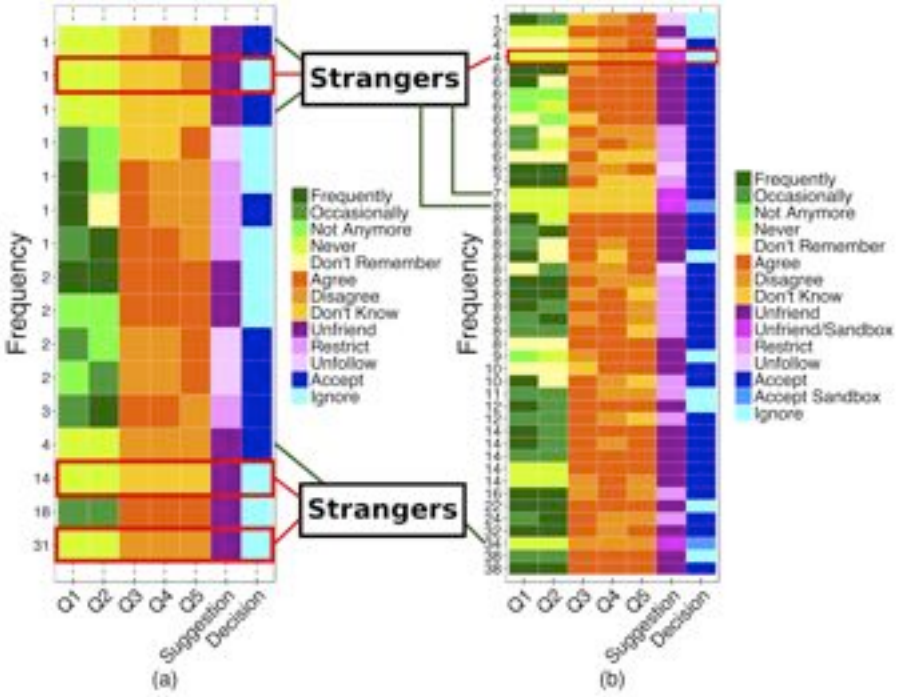


Figure 4.8: Heatmap of frequent questionnaire responses and corresponding user decisions, for (a) 1st user study (n = 20) and (b) 2nd user study (n = 60). A row on the y axis shows the number of friend relationships that have the questionnaire response, recommendation and user decision pattern shown on the x axis. The responses for each question, the recommendations, and user decisions are shown with different colors, see legend. (a) In the 1st study, 52 out of 85 abusive relationships were perceived to be stranger friends. Of these 52 cases, participants have been unwilling to unfriend 46 of the corresponding stranger friends (red border rectangles). (b) In the 2nd study, 53 out of 513 abusive relationships were perceived as strangers. Of these 53 cases, in only 4 cases (red border rectangles) the participants have been unwilling to either unfriend or sandbox the corresponding stranger friends. The contrast to the 1st study suggests participant preference to sandbox vs. unfriend stranger friends.

Willingness to defend against abuse. We now investigate RQ2 (*are users willing to take defensive actions against abusive friends?*) and RQ3-b (*is the willingness of users to take defensive actions impacted by the type of abuse perpetrated by the friend and the type of suggested defensive action?*).

In the first of the above 2 studies (n = 20, 400 investigated friend relationships), AbuSniff identified 85 abusive friend relations. Of these, AbuSniff recommended 74 to unfriend, 6 to restrict and 5 to unfollow. Figure 4.8(a) shows a detailed view (heatmap)

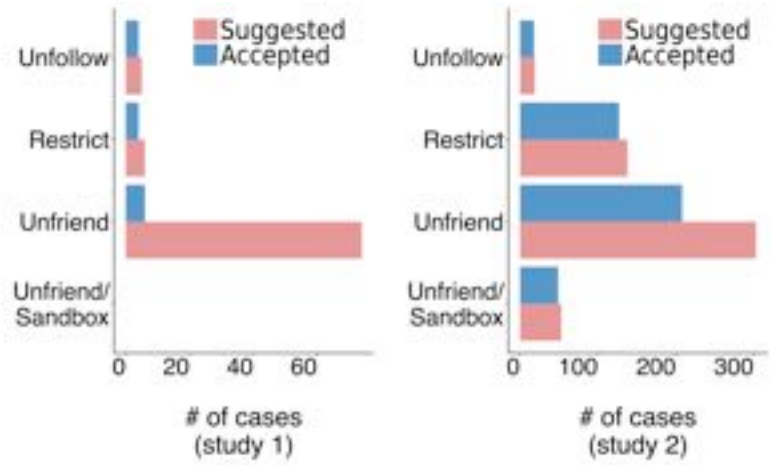


Figure 4.9: Comparison of recommendation vs. acceptance in study 1 (n = 20) vs. study 2 (n = 60). In study 1, 8% of the recommended "unfriend" actions were accepted. The undefined "unfriend or sandbox" option is shown for alignment. The "sandbox" option and user education were effective: in study 2, 92% of the suggested "unfriend or sandbox" suggestions were approved by participants.

of these 85 abusive relationships, along with the frequently occurring (shown on y axis) patterns of questionnaire answers, suggested actions and user decisions (shown on the x axis). Of particular interest is the fact that 52 out of 85 abusive relationships were perceived to be non-abusive stranger friends, and, of these 52 cases, participants have been unwilling to unfriend 46 of the corresponding stranger friends (red border rectangles). The small approval rate of the suggestion to unfriend strangers (11.5%) motivates the second study.

The results of this study are further summarized in Figure 4.9(a). 4 out of the 6 recommended restrict friends were restricted, and 4 out of the recommended 5 were unfollowed. However, only 6 out of 74 recommended unfriend were unfriended.

The "Sandbox" effect. To address the fear of being observed by the unfriended friend, we have relaxed rule 1 in Table 4.1, to give the user the option to either sandbox or unfriend a non-abusive stranger. A sandboxed friend can no longer harm the user, as all Facebook communication lines are interrupted. Sandboxing achieves this without sever-

ing the friend link, thus is not observable by the friend. However, if the stranger exhibits any sort of abusive behavior (timeline or news feed abuse) AIE recommends that the friend should be unfriended (rule 2). Rule 2 is evaluated only if rule 1 fails (see Table 4.1).

Further, we have modified AbuSniff's UI to educate the user through a description of the harm that strangers can perform, and of the defenses that the user can take against such friends. Figure 4.2(d)) shows a snapshot of the modified UI screen that offers the sandbox alternative to unfriending strangers.

The second user study described above (n = 60) evaluated the updated AbuSniff, that identified a total of 513 abusive friend relations. Figure 4.8(b) provides a detailed view of the user responses, recommended actions and user decisions. For instance, it shows that 53 friend relations are considered non-abusive strangers. Of these 53 cases, participants were unwilling to unfriend or sandbox only 4 of the corresponding stranger friends. The contrast to the 1st study is significant: there, participants took a defensive action for a non-abusive stranger friend in only 11.5% of the cases, whereas in the second study participants took an action in 92.4% of the cases. This contrast suggests participant preference to sandbox stranger friends.

Figure 4.9(study 2) further shows that AbuSniff recommended 303 friend relations to unfriend, 53 to unfriend or sandbox, 138 to restrict and 19 to unfollow. Consistent with the first study, 18 of the 19 unfollow and 127 of the 138 restrict suggestions were accepted. In contrast to the first study, 49 of the 53 "unfriend or sandbox" suggestions were accepted. In addition, 208 of 303 "pure" unfriend recommendations were accepted, again a significant improvement over the first study (6 out of 74). These results suggest that the explanation of the harm is effective in raising user awareness, and that user awareness of the harm converts into more restrictive actions.

Reasons to ignore recommended action	S1	S2
Recommendation does not make sense	9	1
Not ready to take action at that time	12	2
Agree but still want to keep stranger friend	9	1
Afraid that action will be observable	16	0
Total Cases	46	4

Reasons to ignore recommended action	S1	S2
Recommendation does not make sense	4	6
Not ready to take action at that time	6	48
Agree but still want to keep abusive friend	3	9
Afraid that action will be observable	9	32
Total Cases	22	95

Table 4.3: Comparison of reasons to ignore AbuSniff suggested action in user study 1 (n=20) and user study 2 (n = 60) for (left) non-abusive strangers, where the suggestion was "unfriend" in the first study but "sandbox or unfriend" in the second study, and (right) abusive non-strangers, where the suggestion was "unfriend" in both studies. The addition of the "sandbox" option for non-abusive strangers had a significant effect on participant perception of suitability of suggestion, readiness for action, and fear of observability.

Reasons to ignore recommendations. As mentioned above, in the first study, participants ignored recommended defenses in 46 out of 52 non-abusive strangers cases, while in the second study, participants ignored such recommendations in only 4 out of 53 non-abusive strangers cases. Table 4.3(left) summarizes and compares the reasons given by the participants in the two studies, to ignore the suggested recommendations.

We observe that the addition of the "sandbox" option for non-abusive strangers had a significant effect on participant perception of suitability of the AbuSniff suggestion: 9 participants believed that AbuSniff's recommendation does not make sense in the first study, compared to only 1 in the second study. It further had a significant effect on participant readiness for the suggested defense: only 3 participants in the second study were either not ready for the action or wanted to keep the stranger friend, compared to 21 in the first study. Notably, none of the participants in the second study were afraid that the action will be observable by the friend, a steep decrease from 16 participants in the first study.

Table 4.3(right) further compares the reasons chosen by participants in the two studies, to ignore the recommended defense action of unfriending abusive non-stranger friends. These are friends with whom the participants had interacted in Facebook and/or real life, and who are perceived as potentially "bi-directional" (timeline and news feed) abusive.

We observe that 22 such friends were identified in the first study, and 95 such cases were found in the second study. We observe that in only 4 and 6 cases in the two studies, participants believe that the recommendation does not make sense. In 9 cases in the first study and 57 in the second study, participants agree with the suggested defense but are either not ready to take it, or would prefer to keep the friend. Further, in almost 1 third of the cases in both studies, the participants did not take the action due to fear of observability. These numbers are consistent with the "unfriend" suggestion for non-abusive stranger friends in the first study (16 out of 46 cases), see Table 4.3(left) and suggest that these participants may prefer to sandbox even abusive non-stranger friends.

In 13 of the 68 ignored cases, the participants believed that AbuSniff's unfriend recommendation does not make sense. In these cases, AbuSniff suggested to unfriend due to the fact that the friend was a stranger (the participant never interacted with that friend either in Facebook or real life) In 55 (68 - 13) of the 68 ignored unfriend cases, the participants believed that our warning was correct. However, they refused to unfriend because either they were not ready to unfriend them at that time (18 of the 55 cases), they still wanted to keep those abusive friends (11 cases), or they were afraid that this action will be observable by the abuser (26 cases).

Only 5 out of the 95 ignored unfriend recommendations were due to the participant not believing our recommendation

In addition, in 52% of the cases the participant was not ready to unfriend at that time, in 10% of the cases the participant still wanted to keep the abusive friend and in 36% of the cases, the participant did not want the friend to find out that he/she was unfriended. We conjecture that we could achieve increased performance, if similarly, AbuSniff offered users the option to sandbox (or at least restrict/unfollow) even actively abusive friends.

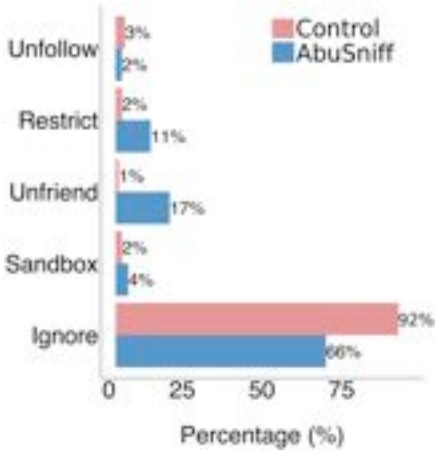


Figure 4.10: Questionnaire based AbuSniff vs. control experiment. AbuSniff had a significant impact on the willingness of participants to unfriend (17% vs. 1%) and restrict (11% vs. 2%) friends.

4.7.2 Study 3: Questionnaire Based AbuSniff vs. Control

In order to understand if AbuSniff had an effect on the willingness of users to take defensive actions on Facebook friends (key question RQ3-a), we have designed a control app, that first explains each user action ("unfriend", "unfollow", "sandbox", "restrict" and "ignore", see Figure 4.4(a), then, for each of 30 randomly selected Facebook friends of the user, asks the user to take one of these actions for the friend. (see Figure 4.4(f) for a snapshot). The control app does not require the user to answer a questionnaire and the user is not provided with a motivation for taking an action for the friend. We performed a control study with this app on 27 participants.

Figure 4.10 compares the results of our second user study with the questionnaire based AbuSniff, against the control, in terms of the percentage of actions taken by the participants. For AbuSniff, the "Ignore" bar shows not only the recommended actions that were ignored by the participants, but also the much larger number of relationships that were not identified by AbuSniff as problematic (abusive or strangers).

We found that in the control experiment, participants did not take a restrictive action in 92% of cases. In contrast, with AbuSniff, friend relationships (including "safe" ones)

were ignored in only 66% of the cases. We observed significant differences for the "unfriend" and "restrict" options, that were chosen in only 1% and 2% of cases respectively during the control experiment, but in 17% and 11% of the cases respectively during the AbuSniff experiment.

4.7.3 Study 4: Efficacy of Abuse Prediction

We investigate now key question RQ4 (*Can AbuSniff predict abusive friends and the defenses that users are willing to take against such friends*). For this, we performed a fourth study where we have used the questionnaire based AbuSniff to collect a subset of Facebook data from 1,452 friend relationships of 54 participants.

Feature correlation investigation. First, we investigate the intuition behind the features we used in the APM module (see Section 4.5.1). For that, we have performed multinomial logistic regression (MLR) analysis using SPSS to find out whether the mutual activity features (number of mutual posts, common photos, mutual friends, common studies, common workplaces, living in the same current city, and having the same hometown) are good predictors for the actions decided by AIE. Multinominal logistic regression is appropriate since the outcome variable is categorical with more than two categories and the predictors are of either continuous or nominal. We used the 7 mutual activity features (5 continuous, 2 categorical) as the independent variables and the AIE decision with 5 categories (Unfriend, Sandbox/Unfriend, Restrict, Unfollow, Safe) as the dependent variable.

Model fit statistics indicate a good fit, i.e., χ^2 (28) = 385.037, p ¡ 0.05 which confirms our model predicts significantly better, or more accurately, than the null model. Table 4.4 shows the output from the likelihood ratio test that checks the contribution of each effect/feature to the model. For each effect, the -2 log-likelihood is computed for the reduced model; that is, a model without the effect. If the significance of the test is small

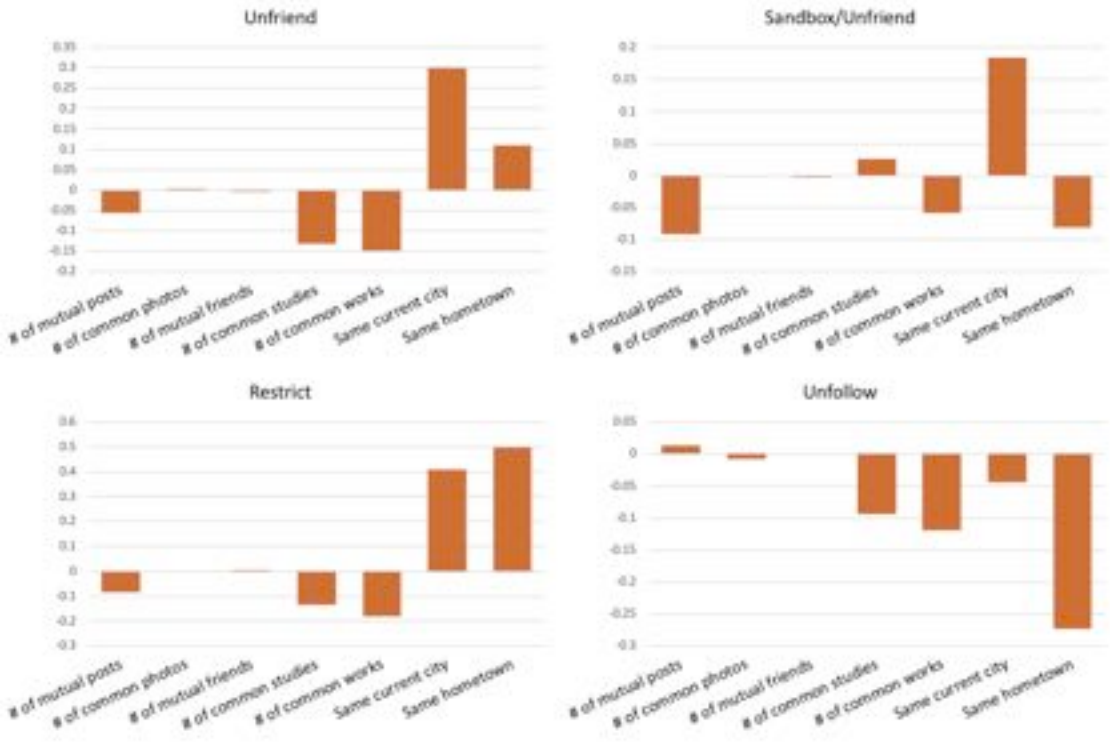


Figure 4.11: Multinomial logistic regression (MLR) correlations between mutual activity features and AIE decision for each abuse category. Coefficients for the mutual activity features are plotted as Sign(C_f)*Log(1+Abs(C_f)), where C_f denotes the actual co-efficient. For the Same current city and same hometown features, we have analyzed the values of [Same current city=No] and [Same hometown=No]. The same current city, the same hometown, the number of common workplaces, the number of common studies and the number of mutual posts have the highest impact on all of the AIE decisions.

(less than 0.05) then the feature contributes to the model. The chi-square statistic is the difference between the -2 log-likelihoods of the reduced model and the final model. 6 of the 7 features show statistically significant overall association with the AIE decision. The only exception is the "number of common photos" feature.

Figure 4.11 plots the value of each coefficient in the model with its respective sign, to show if the features are positively or negatively correlated with the AIE decisions. For convenience of showing the results, we plot the co-efficients for the mutual activity features as Sign(C_f)*Log(1+Abs(C_f)), where C_f denotes the actual co-efficient value. A larger (absolute) coefficient means that the corresponding feature has more impact on the

APM Feature	χ^2	Significance
Number of mutual posts	39.652	0.000
Number of common photos	1.938	0.747
Number of mutual friends	38.739	0.000
Number of common studies	14.442	0.006
Number of common workplaces	31.942	0.000
Same current city	55.134	0.000
Same hometown	41.555	0.000

Table 4.4: Independent variables that have statistically significant overall association with the AIE decision. The results are significant (at $p < 0.05$) between each feature and the AIE decision except for the number of common photos.

prediction. It shows that according to the MLR analysis, the same current city, the same hometown, the number of common workplaces, the number of common studies and the number of mutual posts have the highest impact on all of the AIE decisions.

Experiment details. We have used 10-fold cross validation to evaluate the ability of the abuse prediction module (APM) to predict questionnaire responses and user defense decisions. For this, we have computed the 7 mutual activity features of the 54 participants and the 1,452 friends. We have generated a dataset of 1,452 tuples, one for each friend relationship. Each data tuple corresponds to a user U and friend F, and consists of (1) the mutual activity feature values of U and F, (2) U's responses to the 5 questions for F, and (3) the suggested action for F: ignore (safe friend), unfriend, unfollow, restrict, sandbox, and (4) the action taken by U for F. We have divided this dataset into 10 folds of 145 tuples each, selected randomly, and used 10-fold cross validation to evaluate several supervised learning algorithms. That is, in each of 10 experiments, we used 9 folds to train and 1 to test. We used the features of each tuple in the 9 folds to separately train supervised learning algorithms for each of the 5 questionnaire questions and for the user decision. Then, for each tuple in the remaining fold, APM uses the trained algorithms to predict the answers to the 5 questions and the user decision. We report averages of the prediction accuracy over the 10 experiments.

As shown in Figure 4.7, the distribution of the answers to the 5 questions of the questionnaire was not balanced. To address this imbalance, we have duplicated tuples from the minority classes up to the the number of the majority class. We have ensured that duplicates appear in the same fold, to prevent testing on trained tuples.

We have used Weka 3.8.1 [Wek17] to evaluate several supervised learning algorithms, including Random Forest (RF), Decision Trees (DT), SVM, PART, MultiClassClassifier, SimpleLogistic, K-Nearest Neighbors (KNN) and Naive Bayes, but report only the best performing algorithm.

Predicting questionnaire answers. We now evaluate the ability of the APM module of the predictive AbuSniff to predict the user's questionnaire answers. Table 4.5 shows the precision, recall and F-measure achieved by the best performing supervised learning algorithm for each of the questionnaire questions (Q1-Q5).

For question Q1, the Random Forest (RF) classifier achieved the best performance. We have used 1 class for each of the 5 possible responses. Table 4.5 (top section) shows the classification results of RF for each class and as a weighted aggregate. APM with RF predicts the "Never" response with precision 81.8% and recall 92%, for a F-measure of 86.6%. RF also achieves the best performance for question Q2, with an overall F-measure of 90% (see second section of Table 4.5). APM with Random Forest is able to predict the "Never" response for a friend, with precision 86.5% and recall 93.2%.

For Q3, APM achieved the best performance when using the Decision Tree (DT) classifier (see third section of Table 4.5), with an average F-Measure of 79.3%. The DT classifier also achieved the best results for Q4 (the fourth section of Table 4.5), with an average F-Measure is 80.4%. For Q5, APM achieved best performance with RF, see Table 4.5 (fifth section), with an F-Measure for the news feed abuse indicator ("Agree" response) of 78%.

Question	Precision	Recall	F-Measure	Class
	0.983	1.000	0.992	Frequently
	0.928	0.897	0.912	Occasionally
Q1	0.962	0.797	0.872	Not Anymore
(RF)	**0.818**	**0.920**	**0.866**	**Never**
	0.934	0.898	0.916	Don't Remember
	0.917	0.914	0.914	Weighted Avg.
	0.966	0.905	0.934	Frequently
	0.893	0.869	0.881	Occasionally
Q2	0.893	0.877	0.885	Not Anymore
(RF)	**0.865**	**0.932**	**0.897**	**Never**
	0.907	0.911	0.909	Don't Remember
	0.902	0.900	0.900	Weighted Avg.
	0.725	**0.792**	**0.757**	**Agree**
Q3	0.820	0.793	0.806	Disagree
(DT)	0.810	0.791	0.800	Don't Know
	0.794	0.792	0.793	Avg.
	0.662	**0.725**	**0.692**	**Agree**
Q4	0.791	0.778	0.785	Disagree
(DT)	0.857	0.844	0.851	Don't Know
	0.805	0.803	0.804	Avg.
	0.794	**0.765**	**0.780**	**Agree**
Q5	0.837	0.845	0.841	Disagree
(RF)	0.830	0.842	0.836	Don't Know
	0.824	0.824	0.824	Avg.

Table 4.5: Precision, recall and F-measure of APM for questions Q1 (RF), Q2 (RF), Q3 (DT), Q4 (DT) and Q5 (RF). (Question 1) APM with Random Forest (RF) predicts the "Never" response with precision 81.8% and recall 92% (Kappa statistic = 0.88). (Question 2) APM with Random Forest (RF) predicts the "Never" response with precision 86.5% and recall 93.2% (Kappa statistic = 0.86). (Question 3) APM with Decision Tree predicts the abuse indicator "Agree" response with F-Measure of 75.7% (Kappa statistic = 0.68). (Question 4) APM with Decision Tree achieves F-Measure of 69.2% when predicting the abuse indicator "Agree" (Kappa statistic = 0.67). (Question 5) APM with Random Forest has the F-Measure for "Agree" of 78% (Kappa statistic = 0.73).

Classified As					
Unfriend	**Sandbox**	**Restrict**	**Unfollow**	**Ignore**	**Decision**
882	13	10	13	3	**Unfriend**
103	27	1	1	3	Sandbox
77	1	6	0	1	Restrict
79	3	0	6	0	Unfollow
5	0	0	0	**218**	**Ignore**

Table 4.6: APM confusion matrix for predicting user decisions. The rows show participant decisions, the columns show APM predictions during the experiment. AbuSniff will leverage APM's high precision (96.9%) and recall (97.8%) for the "ignore" action, to decide which abusive friends to ignore.

We observe a higher F-measure in predicting answers to the questions that suggest stranger friends (Q1 and Q2) than in predicting answers to the questions that suggest abuse (Q3-Q5). This is not surprising, as the mutual activity features are more likely to predict online and real life closeness.

Predicting the user decision. We have evaluated the ability of APM to predict the defense action that the user agrees to implement, according to the 5 possible classes: "unfriend", "restrict", "unfollow", "sandbox", and "ignore". APM achieved the best performance with the RF classifier. Table 4.6 shows the confusion matrix for APM with RF, over the 10-fold cross validation performed on the 1,452 friend instances. APM's overall F-Measure is 73.2%. The APM's precision, recall and F-Measure for the "unfriend" option are 77.0%, 95.8% and 85.3% respectively. However, APM achieved an F-measure of 97.3% when predicting the "ignore" option. We emphasize the importance of this result, as AbuSniff uses APM's predictions to decide which friends to recommend for the user to defend against.

Feature rank. Since we have only used 7 features, we have not selected or reduced the feature set. The most informative features in terms of information gain were consistently among the mutual post count, mutual friend count and mutual photo count; same hometown and common study count were the least informative features. We found correlations

between the common photo count and mutual post count (Pearson correlation coefficient of 0.65), mutual friend count and mutual photo count (Pearson correlation coefficient of 0.57), and mutual post count and mutual friend count (Pearson correlation coefficient of 0.45). The rest of the features had insignificant positive or negative correlations.

4.7.4 Study 5: Predictive AbuSniff in the Wild

As described in Section 4.5, the predictive AbuSniff system replaces the questionnaire delivery module (QRM) with the abuse prediction module (APM). It then asks the user to either accept or ignore the APM predicted defense action, shown only for the friends for whom the APM predicts that the user will defend against.

We have recruited 49 participants to evaluate their reaction to the predictions of the APM module. We have discarded 9 participants who failed the data quality verification tests previously described. Of the 1,200 friend relationships investigated for the remaining 40 participants (30 friends per participant), APM automatically labeled 403 as potentially abusive. AbuSniff predicted that 359 of these will be approved by the participants, i.e., 41 unfollow, 30 restrict, 137 unfriend and 151 sandbox. AbuSniff displayed only these suggestions to the respective participants. All the unfollow and 29 of the 30 restrict suggestions were accepted by the participants. 119 of the suggested sandbox relationships and 92 of the suggested unfriend relationships were accepted. Thus, overall, the 40 participants accepted 78% of AbuSniff's suggestions.

4.7.5 Studies 6 & 7: Impact of AbuSniff

We have designed 2 user studies to evaluate the impact of AbuSniff on (1) the willingness of participants to ignore new friend invitations based on their perception of the prospective

friend being a stranger or an abuser, on (2) participant awareness of and perception of safety from friend abuse, and (3) their willingness to screen other friends.

For this, we have designed a pre-study survey that consists of 3 Likert items: (I1) "When I receive a friend invitation in Facebook, I reject it if I have never interacted with that person in real life or online", (I2) "When I receive a friend invitation in Facebook, I reject it if I think that the person would abuse my photos or status updates in Facebook, and (I3) " When I receive a friend invitation in Facebook, I reject it if I think that the person would post abusive material (offensive, misleading, false or potentially malicious)." We performed a pre-test only study with 31 participants, where we have delivered (only) this survey. Figure 4.4(c-e) shows screenshots of this survey.

Further, we have designed a post-study survey that, in addition to the above 3 items, includes the following 3 Likert-scored statements: (I4) "After completing AbuSniff, I feel more aware of the implications of friend abuse in Facebook", (I5) "After completing AbuSniff, I feel more protected from abuse from Facebook friends", and (I6) "I will go to my friend list and evaluate my other friends to defend against those I feel could be abusive". Figure 4.4(h-j) shows snapshots of these questions. In a post-test study with a different set of 31 participants, we asked them to first run the questionnaire based AbuSniff, then answer the post-study survey.

Figure 4.12(a) compares the user responses in the pre-test (top) and post-test (bottom) experiments, for each of the first 3 Likert items. In the pre-test experiment, the user responses are balanced between agree, neutral and disagree, and there are no strong agree and strong disagree responses. In contrast, after running AbuSniff (i.e., in the post-test experiment), significantly more participants either strongly agree or agree on all 3 items. Specifically, for (I1), 14 out of 31 participants Strongly Agree or Agree that they would always reject a pending friend with whom they have never interacted, while 9 Disagree. Only 1 participant strongly disagreed. 19 participants strongly agree or agree with (I2),

(a)

(b)

Figure 4.12: (a) The impact of the questionnaire based AbuSniff on (I1), (I2) and (I3). For each question, top bar shows pre-test and bottom bar shows post-test results. In the post-test, significantly more participants tend to strongly agree or agree that they would reject new friend invitations based on lack of interaction or perceived timeline or news feed abuse, when compared to the pre-test. (b) Post-test results for (I4), (I5) and (I6). 23 out of 31 participants perceived that AbuSniff improved their understanding of abuse, more than half perceived that AbuSniff has impacted and improved their safety, and more than half agreed to continue the process on other friends.

and only 4 disagree. Finally, 17 participants strongly agree or agree with (I3), and 5 disagree.

Figure 4.12(b) shows the participant responses to only the 3 new post-test Likert items. 23 out of 31 participants strongly agree or agree that after running AbuSniff they feel more aware of the implications of friend abuse; only 1 disagreed. 19 participants strongly agree or agree that after running AbuSniff they feel more protected from friend abuse; 4 participants disagree. 20 participants strongly agree or agree that they would revisit their other friends after running AbuSniff, and only 3 disagree, 1 strongly disagrees.

4.7.6 Summary of Findings

- **Significant perception of friend abuse**. We observed significant user perception of potentially abusive Facebook friend connections. Male and female participants did not exhibit significant differences. Younger participants were more vulnerable to timeline and news feed abuse than older ones, but not to strangers.
- **Sandbox preferred over unfriend**. We observed a mixed response for RQ2, thus a positive answer for RQ3-b. Specifically, Participants tended to accept suggestions to unfollow news feed abusers (22 out of 24 cases), restrict timeline abusers (131 out of 144 cases), and were more willing to sandbox than unfriend perceived non-abusive strangers.
- **AbuSniff better than control**. When compared to the control group, AbuSniff has significantly increased the willingness of participants to unfriend and restrict friends.
- **AbuSniff can predict friend abuse and willingness to defend**. APM is better at predicting user responses to Q1 and Q2 than to Q3-Q5, and is able to accurately predict the user choice to ignore the suggested defense action. However, APM has

a lower accuracy when predicting the specific defense action that the user will be willing to take.

- **Life imitates art**. In real life studies, AbuSniff had similar accuracy with cross-validation experiments.
- **AbuSniff influences participants**. AbuSniff increased participant self-reported willingness to reject invitations from strangers and abusers, their awareness of abuse implications and perceived protection from friend abuse.

4.8 Discussion and Limitations

We have explored the perception of friend abuse in Facebook, and the willingness of users to take radical actions against friends that they perceive to be abusive or strangers. Unlike previous work that detects, prevents and mitigates cyberbullying in teenagers [DJH+12, AV16], we have focused on general abuse, as perceived by the general Facebook population (equipped with Android devices). We have developed AbuSniff, an automated victim-side tool [AV16] that avoids the problems of accessing individual posts [DJH+12], to instead detect friends perceived to be strangers or abusive.

AbuSniff differs from prior work on cyber abuse and victimization, e.g., [WCZB+16, LLGL13] in that we (1) focus on specific types of abuse perpetrated through Facebook, i.e., timeline and news feed abuse, and stranger friends, (2) investigate abuse perception from individual friends and not general exposure, (3) seek to automatically detect abuse perception and (4) provide a first line of suitable defenses against abuse for Facebook users who are unlikely to know and trust all their friends. We performed studies with 263 participants from 25 countries and 6 continents. We acknowledge a common crowdsourcing worker background between participants.

AbuSniff reduces the *attack surface* of its users, by reducing the number of, or isolating friends predicted to be perceived as potential attack vectors. AbuSniff can reduce the audience that needs to be considered by audience selector solutions, e.g., [RLG16], and can be used in conjunction with tools that monitor social networking events [DJH+12, Per17].

We expect AbuSniff to have more impact for users who have significantly more than 150 friends, the maximum number of meaningful friend relationships that humans can manage [Dun92]. We note that false positives, while being a nuisance, can be fixed by reinstating removed or restricted friends. However, false negatives (keeping abusive and stranger friends) can harm the user and even influence the outcome of elections [BBC17a, BBC17b].

Online relationships and loose tie friends. Social networks like Facebook encourage online relationships (people never met in real life) and loose ties (users keeping up to date with the posts of others, without bi-directional communication). AbuSniff defines and detects "strangers" as friends with whom the user has no online and real world communications. Thus, since "keeping up to date" is considered interaction, AbuSniff does not detect and suggest removing strictly online relationships or loose tie friends.

Ability of questionnaire to identify perceived abuse. Questions 3, 4 and 5 in the questionnaire explicitly evaluate participant perception on the potential for abuse of their friends, i.e., abusing status updates or pictures that they post (questions 3 and 4), and friends posting abusive information on their news feed (question 5). However, questions 1 and 2 identify friends with whom the user has never interacted both in real life and online. Not all such "strangers" may be truly abusive, but simply random people befriended online. Our second user study (see Table 4.2) reveals that indeed, few participants agreed to completely remove such friends, thus they are likely to seldom perceive such friends as being abusive. However, our third study has shown that participants had a much higher

likelihood to "sandbox", i.e., isolate such friends. Even if only a small percentage of those friends are a true threat, the participants in our user study had up to 4,880 friends (median of 305 friends). Since any stranger could be an "attack vector", sandboxing or unfriending strangers can reduce the user's "attack surface", and protect from both mishandling of private, sensitive information, and from attacks such as spear phishing and malware distribution.

Further, in the last user study, 23 participants strongly agreed or agreed that after running the questionnaire based AbuSniff they felt more aware of the implications of friend abuse; only 1 disagreed, none strongly disagreed. This suggests that participants perceived that AbuSniff had an impact on their understanding of abuse.

Prediction accuracy. The APM features extracted from mutual Facebook activities are less effective in predicting the user responses to Q3-Q5. This is not surprising, as we have trained APM on relationship closeness features. We note that the choice of features was due to our need to respect Facebook's terms of service. Access to more information, e.g., stories on which friends posted replies and the friend replies, and abuse detection APIs [Per17] can improve APM's prediction performance. We emphasize however that AbuSniff had an F-Measure of 97.3% when predicting the "ignore" action. Thus, we see potential for improvement and also promise for the feasibility of developing fully automated abuse detection and defense solutions for social networks.

Due to Facebook restrictions, the mechanisms we developed to automatically detect abuse perception are not yet mature, e.g., as they cannot accurately detect perceived news feed abuse. Abuse prediction at the level of individual posts is likely to be much more accurate, especially with the advent of industry APIs for abuse detection [Per17]. However, Facebook's decision to ignore the automatic abuse detection problem and its policies that prevent other parties from accessing the data required to solve it, are significant road blocks toward eliminating friend abuse.

The autonomous AbuSniff had lower performance when predicting the answer to question 5 (*This friend would post offensive, misleading, false or potentially malicious content on Facebook*)), vs. questions 1-4, i.e., an F-Measure of 82.7%. Further, while AbuSniff was able to accurately predict the abusive friends whom a user would choose to defend against vs. ignore, it was not able to accurately predict the exact defensive action that the user would approve. We need to investigate additional features that can boost the accuracy of classifiers. For instance, develop and adapt techniques (e.g., [NTT+16]) that automatically detect abusive posts of evaluated friends, and use the absolute and relative counts of abusive posts as additional APM features.

Validity of AbuSniff recommendations. Participants took significantly longer time to ignore a suggestion (M = 29.30s, SD = 9.86) than to accept it (M = 13.14s, SD = 4.71s), see Figure 4.6 in Section 4.6. This suggests that decisions to ignore recommendations were not taken randomly, participants took the time to process this decision. This implies the quality of AbuSniff recommendations, as obviously incorrect recommendations would have been quickly ignored.

Keeping friends perceived to be abusive. The above discussion may also suggest that some participants had stronger reasons for keeping abusive or stranger friends. In the first AbuSniff study, for 11 of the 68 unfriended friend cases, the participants believed that our warning was correct, but still wanted to keep those friends. One reason for this may be that the participant had reasons to make him or her abusive toward that friend. As mentioned by Dinakar et al. [DJH+12], determining the victim and the perpetrator in an interaction is not an easy task, as victims may also retaliate thus become perpetrators. We leave this investigation for future work, but note that AbuSniff is a victim-side abuse prevention tool, thus may protect the friends if they installed AbuSniff.

"Stranger" friends. Our interest in "stranger" friends is motivated by the fact that participants in our study had up to 4,880 friends (median of 303 friends). This is in line with

Facebook stats, whose current median number of friends per user is 200. Our studies have shown that participants have high numbers of perceived stranger friends. Even if only 1% of those friends are truly abusive, they could launch damaging attacks that include cyberbullying (e.g., outing), identity theft, profile cloning and spear phishing. We emphasize that strangers are just that: friends whom the user perceives now to have never interacted with in real life and online. Since any stranger could be an "attack vector", from a security and privacy perspective, it makes sense to sandbox or remove those friends, thus minimize the user's "attack surface".

Implementation of the suggested actions. AbuSniff does not implement the actions taken by users against their friends. While programatically implementing unfriend, unfollow or restrict actions is possible, one of our goals was for AbuSniff to be minimally invasive. Since participants in our studies has up to 4,880 friends (median of 303), taking an action against a subset of 20-30 friends is unlikely to provide real protection. However, one goal of AbuSniff is to also make users aware of the implications of potential friend abuse. Further, we note that we have not made participants aware of the choice of not implementing their actions. Since we allowed a crowdsourcing worker to participate in at most one of our studies, this design decision is unlikely to have impacted their behavior.

Friend evaluation limitations. We chose to evaluate 20 to 30 friends per participant. A larger number may increase participant fatigue or boredom when answering the questionnaire, thus reduce the quality of the data, AbuSniff's ability to make predictions, and our ability to generalize results. More studies are needed to find the optimal number of evaluated friends per participant, and whether it should be a function of the participant background, e.g., friend count, age, gender.

Further, we note that we have evaluated AbuSniff only on Facebook and make no claims on the applicability of our results to other social networks. More work is needed to identify other relevant forms of abuse (e.g., abusive comments on Instagram photos,

"unprofessional" remarks on LinkedIn, real-life cyberbullying), study their prevalence, and investigate the ability of tools similar to AbuSniff, to educate and protect users.

4.9 Conclusions

We have introduced and studied AbuSniff, the first friend abuse detection and defense system for Facebook. We have developed a compact "stranger and abuse" detection questionnaire. We have introduced and studied rules to convert questionnaire answers to defensive actions. We have shown that supervised learning algorithms can use social networking based features to predict questionnaire answers and defense choices. AbuSniff increased participant willingness to reject invitations from perceived strangers and abusers, as well as awareness of friend abuse implications and perceived protection from friend abuse.

CHAPTER 5

FLOCK: VULNERABILITIES AND DEFENSES TO FRIEND SPAM

5.1 Introduction

Adversarial organizations and governments have used social networks such as Facebook and Twitter to collect and infer private and sensitive information from users [Lap18], inject content to change user perception [OJ18], and distribute fake news, misinformation, propaganda and malware [BBC17a, Lee17, Sin14, Aro16, Wei10, AS10].

Friend relationships are one gateway for such attacks. This is because in sites like Facebook, where users tend to disclose honest self-representations [JGB+18], their data is often shared by default with their friends. Following recent scandals [Lap18, OJ18], Facebook has promised to take swift and decisive action to curtail attacks [Guy18, MZ.18]. However, we observed few changes concerning the friend invitation and acceptance process.

In fact, we show through user studies that little has changed over the years, see e.g., [RBJB14, BMBR11, PCNR10], and Facebook users with diverse backgrounds and mobile devices, are still vulnerable to *friend spam*, invitations received from accounts controlled by strangers. For this, we developed a Facebook-like interface to ask participants to evaluate pending friends. Since we lack ground truth data for evaluation, we invented it. Specifically, we fabricated *synthetic pending friends*, i.e., profiles that simulate spam friend requests, and mixed 5 of them randomly with the actual pending friends, i.e., those who actually sent a friend invitation . We found that 8 out of 10 in-person interview participants, and 25 out of 30 online participants, have each confirmed at least 1 synthetic friend. 19 of the 30 online participants confirmed at least 1 synthetic friend blindly, i.e., without inspecting its profile; 4 participants did not inspect any of the 5 synthetic friends.

4 out of the 10 interview participants went as far as to invent narratives of common background with synthetic pending friends, where none can exist. However, asking participants questions about common background with pending friends, decreased their tendency to confirm strangers: after acknowledging that they do not know them, participants confirmed 35% fewer synthetic friends.

These observations suggest that the blame for confirming perfect strangers as friends, lays at least in part with the Facebook UI design, that encourages users to confirm pending friends without even inspecting their profiles.

To take steps toward addressing this problem, we first propose FLock, an alternative UI design for processing pending friend invitations in Facebook. Our goal is to encourage users to investigate their pending friends, increase their decision time, and reduce their impulse to confirm them as friends. Further, we leverage the synthetic pending friend concept as an opportunity to educate users about the dangers of quickly and greedily accepting friend invitations. To this end, we introduce T-FLock, a FLock extension that trains users to spot and delete synthetic friends, and raises awareness to the importance of carefully considering each pending friend.

To help determine user vulnerability to friend spam without requiring user interaction, we investigate correlations between several features extracted from user accounts and friend invitations, and develop a classifier to predict user decisions on pending friend invites.

Results. In studies with 120 online participants (30 control, 30 FLock, and 60 T-FLock), we found that when compared to a control Facebook-like interface, T-FLock and FLock (1) reduced the percentage of confirmed synthetic friends to 8% and 24.6% respectively, from 50.6%, (2) increased the number of inspected synthetic friend profiles to 82% and 39.33% from 27.33%, (3) decreased the percentage of blind confirms from 55.9% to 37% and 34% for synthetic friends, and (4) increased the time to make a decision for a pending

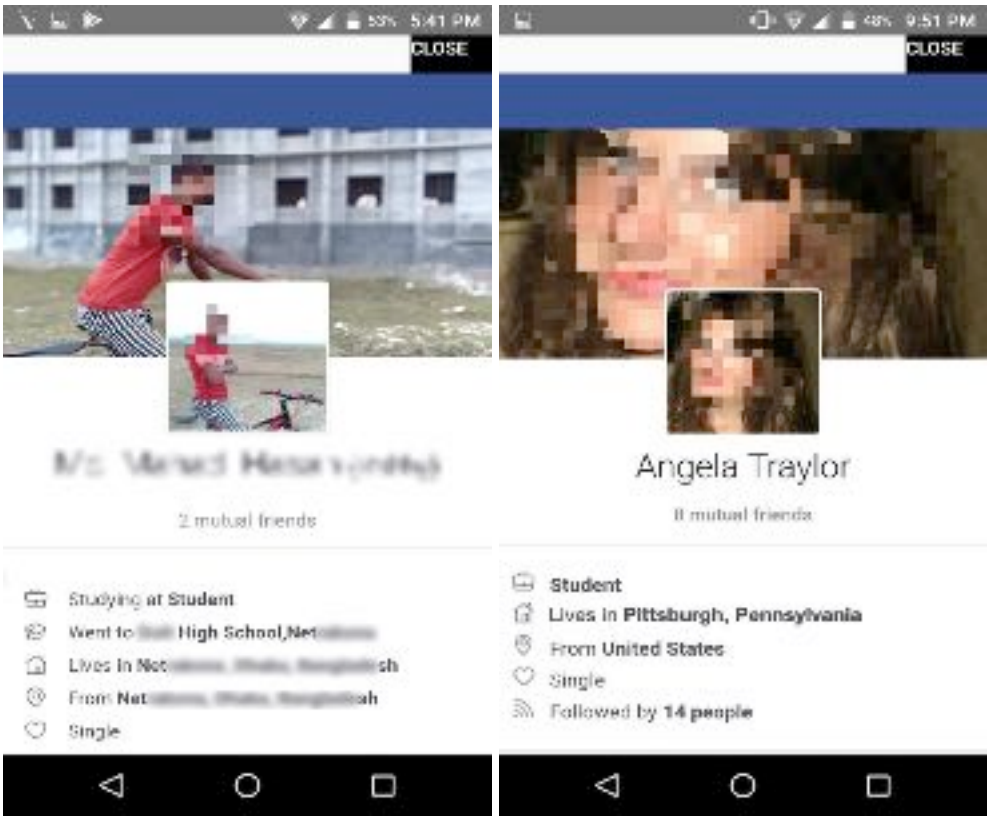
friend by 40.4% and 29.8% respectively, and even more for synthetic friends - by 75.7% and 40.5% respectively.

We found that user decisions were not influenced by the age, gender and education of the participants, or by the age and price of the devices from which the decisions were made. However, even when trained with only 5 past decisions per user, our decision predictor achieved an F1 score of 83.43%. In summary, we introduce the following contributions:

- **Investigate Facebook pending friend UI**. Confirm through in-person and online studies, that Facebook encourages blind and rapid friend confirmations, including from perfect strangers [§ 5.4.1].
- **New designs**. Introduce FLock, a new design to view and process received friend invitations, that encourages users to carefully investigate pending friends, and T-FLock, a system that further trains users to be suspicious of friend spam [§ 5.2].
- **Predict user decisions**. Develop the first classifier to predict user decisions on pending invitations [§ 5.4.4].
- **Evaluation scores and results**. Introduce scores to evaluate the performance of designs for processing pending friends [§ 5.3.3]. Show that design changes to the Facebook UI significantly reduce vulnerability to friend spam attacks [§ 5.4.3].

5.2 Systems

We designed several mobile apps to investigate pending friend vulnerabilities, and mitigating solutions. All apps require the users to login through their Facebook account, and fetch all the pending friends, the total number of existing friends, and 2 randomly selected existing friends. In addition, the apps have the following common features:



(a) (b)

Figure 5.1: Anonymized screenshots of profiles of (a) actual pending friend (of one of the authors), and (b) synthetic friend, emulating the profile page of an actual, but private pending friend. Developed apps display such information when the user taps on the profile photo or name of a pending friend.

- **Attention check screen**. Before login, the apps display an instructional manipulation check [OMD09], to verify that the participant understands English, reads all the text, understands and follows simple instructions, see Figure 5.2(a).
- **App tutorial**. Following login, while the app loads Facebook data of the user's pending friends, the app displays a tutorial screen, see Figures 5.2(c) and 5.4(c).
- **Synthetic pending friends**. To evaluate the willingness of users to accept perfect strangers, the apps randomly mix among the real pending friends, 5 synthetic pending friends, i.e., fake profiles (3 female, 2 male if the user is male; 3 male, 2 female if the user is female). We personalized synthetic friends to be from the same country and live in the same city as the user, and have a randomly chosen number of mutual friends. We created a personalized list of names [Beh] and profile photos for the synthetic friends, for each country of the participants, see Figure 5.1 for anonymized screenshots. We chose the number of followers of a synthetic friend, to be a random value between zero and half of the user's number of friends. Further, we chose identity-neutral images (e.g., nature photos) to display as part of the synthetic friend's profile.
- **Existing friends**. To evaluate participant attitudes toward existing friends, the apps also randomly select and mix with the pending and synthetic friends, 2 *existing friends*, from the user's friend list.
- **Profile inspection**. Similar to Facebook, when users tap on any pending friend, they are shown a screen where they can explore the profile of the pending friend. For real pending friends, the apps show the same information that is shown by Facebook (see § **??**). For synthetic friends, the apps show only the minimalistic information mentioned above. We used javascript injection in the webview, to modify

the html code of the profile page, and hide all the links that take the users to other pages.

- **Skip button**. In addition to the Facebook standard "Confirm" and "Delete" buttons, our apps include a "Skip" button, for the case where the user does not feel comfortable making a decision during the study.
- **Time recording**. The apps log the time of each user action, e.g., profile view and decision (confirm vs. delete vs. skip) taps.
- **Post-study questionnaire**. After processing pending friends, we ask the participants to answer several multiple-choice questions about their age, gender, occupation, highest level of education, the age of their Facebook account, frequency of use, and frequency of received friend invites, see Appendix 5.9.1.
- **Payment code**. The final screen of the apps displays a code, which the user needs to prove completion of the app experience and redeem their payment.

Rationale for synthetic and existing friends. We use synthetic pending friends to evaluate design performance on ground truth data: ideally, participants should not confirm any synthetic pending friend. The reason for showing existing friends is to understand if participants are unhappy with past friend choices, and would like to unfriend some of their existing friends.

In the following, we describe additional elements of each of the interfaces that we designed.

5.2.1 Control App

We have designed a control app, that seeks to emulate the Facebook interface for processing pending friends. The control app, (see Figure 5.3(a) for a snapshot), shows the mix of real pending, 5 synthetic pending and 2 existing friends, in random order, on a

single screen. When the user taps the "Proceed" button at the bottom of the screen, the app displays the next screen only if the user has made a decision for each entry in the list. Otherwise, the app changes the current screen to include an additional, "Skip" button for each shown pending friend, see Figure 5.2(d). The app also pops-up a message asking the user to make a decision for each pending friend. The reasons for this design are that (1) we want the initial app to look as close to Facebook's, however, we also want (2) participants to make an explicit decision for each friend, but (3) without forcing them to decide between "Confirm" and "Delete".

As mentioned above, the user can inspect the profile of each pending friend by tapping on the profile photo or name of the friend.

5.2.2 FLock

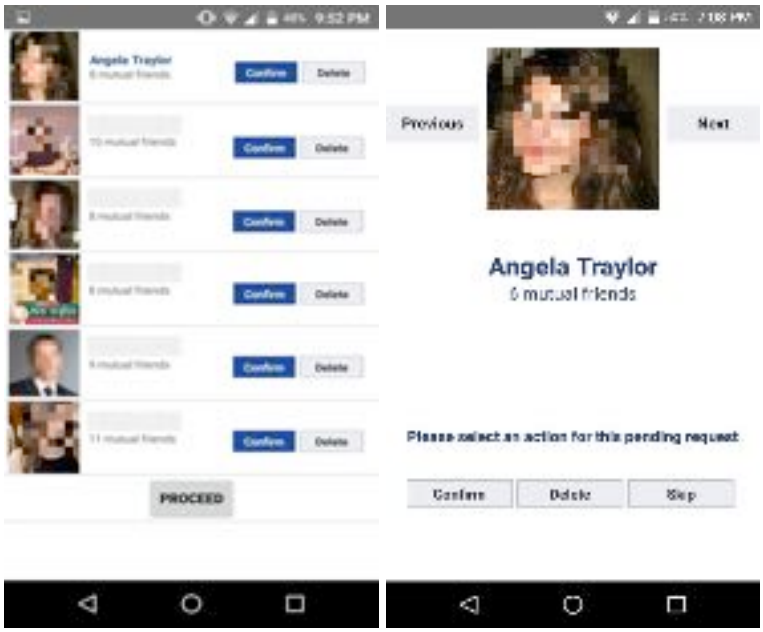
We conjectured that the Facebook UI (see Figure 5.3 and Figure 5.4(a) for snapshots) implicitly encourages users to accept pending friends. This includes the emphasized, blue-colored "Confirm" button vs. the gray "Delete" button, and the crammed listing of all pending friends in a single screen. Participants in the studies of Rashtian et al. [RBJB14] complained about unclear small profile photos.

We designed FLock (Friend Lock) to address these issues, see Figure 5.3(b) for an illustrative snapshot. First, to remove the clutter and reduce cognitive load, FLock displays each pending friend in a single screen, with a large, centrally-placed profile photo. Second, we transform the "Confirm" button into an inhibitive attractor [BLKC$^+$13], by displaying it in the same gray color of the "Delete" button. The users can navigate their pending friend list using buttons on the sides of the profile photo, see Figure 5.3(b). Since we want participants in our studies to make an explicit decision for each pending friend, we activate the "Next" button only after the user makes a decision. We do not force par-

(a) (b)

(c) (d)

Figure 5.2: Anonymized screenshots used in our systems: (a) Attention check screen. (b) Post-study questionnaire. (c) Tutorial screen for the skip button, in the control app. (d) New interface with "Skip" buttons, in the control app, shown when the user taps "Proceed" without processing all friends. Screens (a) and (b) are shown in all the apps.



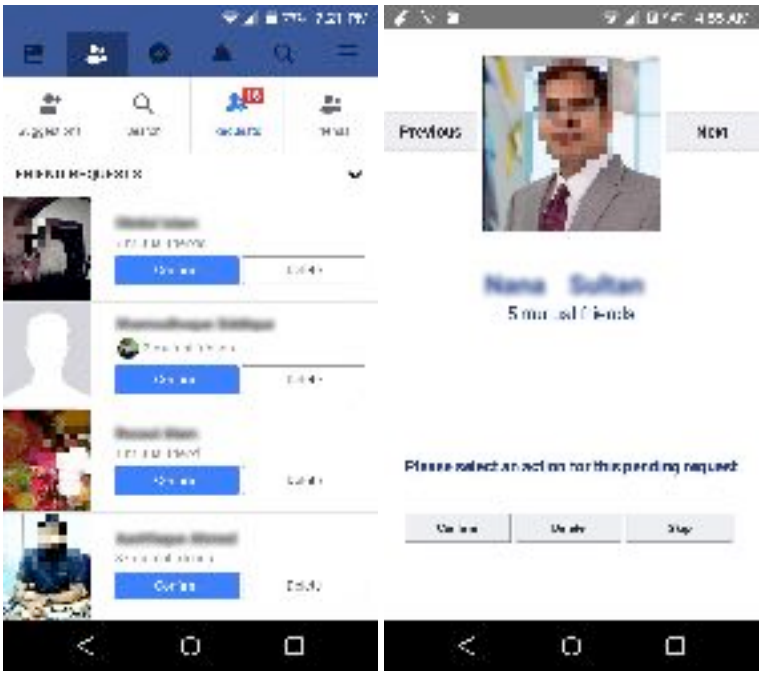
(a) (b)

Figure 5.3: Anonymized screenshots of the friend request interface of the (a) control app, where the first on the list is a synthetic pending friend, and (b) FLock app, showing the decision page for the same synthetic pending friend.

ticipants to choose between "Confirm" and "Delete", but also include the "Skip" option, shown as a button in the same gray color.

Figures 5.4(a) and 5.4(b) compare the Facebook friend request UI with the FLock UI. In FLock, each pending friend is displayed on a different screen, with a larger profile photo and name fonts. The decision buttons also include the "Skip" option, but, unlike Facebook, are shown in the same color, i.e., the "Confirm" button is transformed into an inhibitive attractor. Figure 5.4(d) shows a snapshot of the pop-up that asks the user to pick a reason for deleting an existing friend, that was presented as a pending friend.

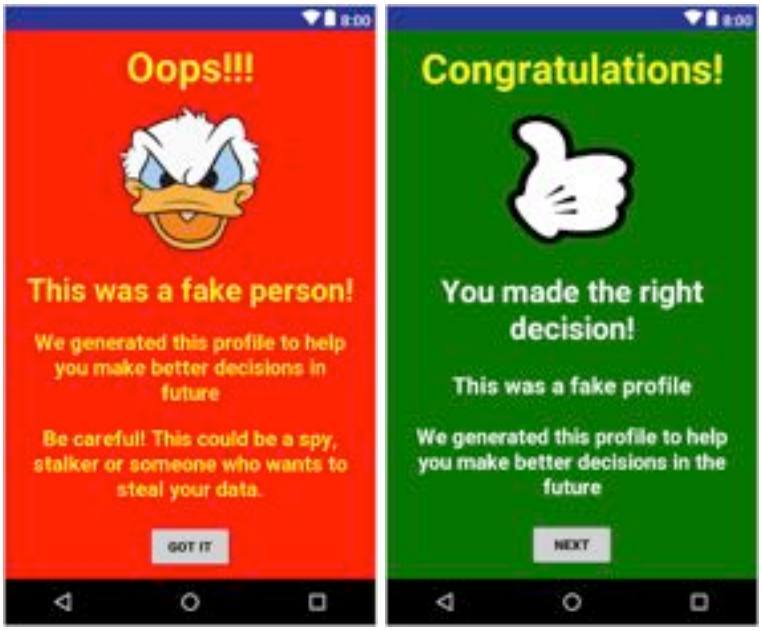
Further, to understand the reasons for which users delete existing friends, presented as pending friends, FLock pops-up a multiple-choice question asking for the reason to delete such a friend, immediately after taking the action. The choices were (a) "This person is already my friend", (b) "I don't know this person", (c) "I don't trust this person", (d) "I



(a) (b)

(c) (d)

Figure 5.4: (a) Genuine Facebook friend request interface, shown for comparison side-by-side with (b) FLock friend request interface. (c) Tutorial screen shown in FLock and T-FLock. (d) Screenshot of pop-up asking the user to pick a reason for deleting an existing friend, that was presented as a pending friend.



(a) (b)

Figure 5.5: Screenshots of the Android T-FLock app shown during the training session: (a) Warning screen (red background) for confirming a synthetic friend. (b) Congratulation screen (green background) for deleting a synthetic friend.

don't want this person to be my friend anymore" and, (e) "Type other reason", along with a text input field where the users can type the reason, see Figure 5.4(d).

5.2.3 T-FLock

Previous work, e.g. [RBJB14, BMBR11] provides early evidence that carefully crafted accounts have a high success rate of befriending strangers. We conjecture that Facebook users can be trained to spot strangers, inhibit their impulse to quickly confirm friend requests, and spend more time investigating them.

To evaluate this hypothesis, we have developed T-FLock (Trained FLock), an FLock extension that consists of two parts. In the first, training part, T-FLock displays three synthetic friends in succession, using the same UI of Figure 5.3(b). For each such synthetic friend, if the user chooses to confirm, T-FLock immediately displays the warning screen

of Figure 5.5(a); if the user chooses to delete, T-FLock shows the congratulatory dialogue screen of Figure 5.5(b).

In the second, evaluation part, T-FLock behaves like FLock, i.e., it displays the participant's actual pending friends, mixed with 5 additional synthetic (i.e., different from the training ones) and 2 existing friends as before, in random order. T-FLock no longer shows the training warning or congratulation screens at this stage, in order to be able to gauge the effects of training on the decisions made for the later synthetic friends.

Figure 5.2(a) and Figure 5.2(b) show screenshots of the attention check screen, and the post-study questionnaire, that we display in the control, FLock and T-FLock apps. Figure 5.2(c) shows the tutorial screen for the control app, and below it, Figure 5.4(c) shows the tutorial screen that we display in FLock and T-FLock. Figure 5.2(d) shows the control app interface that we display when the user taps the "Proceed" button before processing all pending invitations. It contains a new "Skip" button added for each pending invitation, and also a pop-up message at the bottom.

5.2.4 Collected Data

Our systems collect from each participant, the anonymized participant id, the total number of friends, gender, country of origin - selected from a drop-down menu, answers to the post-study questionnaire, mobile device model, and, for each processed pending friend, its type (pending, synthetic, existing), decisions taken and timing data.

5.3 Methods

We have used in-person interviews to investigate user perception and reaction to pending friends. Further, we have used remote online participants to evaluate the impact of our solutions. We now detail both methods.

5.3.1 In-Person Exploratory Interviews

We conducted semi-structured, in-person interviews with 10 participants that we recruited from various locations in our city (e.g., supermarkets, grocery stores, restaurants, social events etc.). Each participant signed a consent form.

Each interview took place in two rounds. In round 1, we used a FLock-like app (but without confirm, delete and skip buttons) that displays each of the pending, synthetic and existing friends, in random order, one by one, on the screen. We asked each participant to make a decision for each friend (either confirm or delete) and recorded their answers.

In round 2, we revisited each of the pending friends, and asked the participant 21 open-ended questions, intended to capture the user perception of their pending friends (see Appendix 5.9.1). We have designed the questionnaire through an iterative process that included a focus group with 5 participants. The open-ended questions concern the relationship of the participant with each of the shown "pending friends", and include (1) 13 interaction type and frequency questions (e.g., *"do you know this person?"*, *"do you remember when and where you met this person for the first time?"*, (2) 4 perception based questions, e.g., *"do you have any negative feelings about this person"*, *"do you trust this person?"*, and (3) 4 common background questions, e.g., *"have you ever studied/worked together?"*.

After asking these questions for each pending friend, we asked the participant to make a decision again. We audio recorded the entire interview with the consent of the participants, and transcribed their responses at a later time.

After round 2, we asked each participant to answer the post-questionnaire described in Section 5.2, on paper, concerning their demographics, occupation, education, Facebook activity levels, see also Appendix 5.9.1.

The participants took an average 38 minutes (M = 35, SD = 8.43) to complete the interview, and we paid each $15.

5.3.2 Crowdsourced User Studies

We have conducted 3 user studies, with the control, FLock and T-FLock apps, with a total of 145 remote participants, recruited from JobBoy [Job17], between September 2018 and February 2019. The jobs we posted asked the participants to install our app (control, FLock or T-FLock) from the Google Play store, use it to login to their Facebook accounts and follow the instructions on the screen. We have only recruited participants who have at least 30 Facebook friends and 1 pending friend, had access to an Android device, and were at least 18 years old. We applied a cryptographic one-way function to the Facebook account ids of the participants, and used the anonymized ids to ensure that each human participated only once in all 3 studies.

To ensure the quality of the data collected, we discarded the data of participants who failed the attention check or declared an incorrect location, i.e., inconsistent with both the Facebook listed location and the device IP address (see § 5.2.4). We note that we do not exfiltrate or store Facebook listed locations or device IPs, see § 5.2.4. In total, we have discarded data from 25 participants: 12 of the 42 recruited participants in the control study, 5 out of 35 in the FLock study, and 8 out of 68 in the T-FLock study.

We have paid each participant, including the ones whose data we discarded, $3, for an average job completion time of 11.18 minutes (M = 11.58, SD = 3.18).

5.3.3 Evaluation Scores

We introduce the following scores to evaluate user behaviors concerning their pending friends, in the control study, and the impact of FLock and T-FLock:

- **Inspected profiles**. The number of pending friend profiles inspected.
- **Blind confirmations**. The number of pending friends confirmed without prior profile inspection.

- **Decision split**. The number of pending friends confirmed, deleted and skipped.
- **Timing**. The time spent inspecting a profile, and making a decision.

Ideally, we want to see a reduction in the number of confirmed synthetic friends, and blind decisions, and an increase in the number of pending profiles inspected, and in the time to inspect profiles and make decisions.

5.3.4 Ethical Considerations

We have developed our protocols to interact with participants and collect data in an ethical, IRB-approved manner. We did not store the Facebook data (e.g., profile picture, name, number of mutual friends, see § **??** and § 5.2) that we extracted from the pending friends and 2 existing friends of participants. We only displayed this information for the participant to see. Further, we did not exfiltrate and store Facebook listed locations of the participants or their IP addresses, but only the user-selected country of location, if consistent. Since we do not preserve this information, its handling does not fall within the PII definition of NIST SP 800-122. Under GDPR, the use of information without context, e.g., name or personal identification number, is not considered to be "personal information". We communicated our data collection process during recruitment and in the consent form, and we recruited only consenting participants. We paid all the recruited participants, regardless of them passing the attention check screen.

5.4 Results

We now present the results of our studies. We use a significant level of $\alpha = 0.05$ in all the statistical tests.

Age		Gender		Geographical Region		Occupation		Education	
Category	Percentage	Category	Percentage	Category	Percentage	Category	Percentage	Category	Percentage
18-24	49.16%	Male	69.16%	Africa	3.33%	Unemployed	9.16%	Elementary	5.0%
25-34	35.83%	Female	30.83%	Asia	63.3%	Homemaker	9.16%	High school	30.83%
35-44	10.83%			Europe	9.16%	Retired	0.83%	College	10.0%
45-54	3.33%			Middle East	5.0%	Student	41.66%	Bachelors	45.0%
55-64	0.83%			North America	13.33%	Employed	39.16%	Graduate	9.16%
				South America	4.16%				
				Other	1.66%				
Account Age		**Friend Count**		**Pending Friend Count**		**Account Usage Frequency**		**Frequency Received Invitations**	
Category	Percentage	Category	Percentage	Category	Percentage	Category	Percentage	Category	Percentage
¡1 year	0%	¡=35	2.50%	¡=5	26.66%	Every hour	24.16%	Almost everyday	0%
1-2 years	10.83%	36-75	7.50%	6-10	13.33%	Several times a day	31.66%	Once in every few days	1.66%
3-5 years	50.0%	76-150	10.00%	11-14	15.0%	Once a day	7.50%	Once a week	36.66%
6-10 years	20.0%	151-250	18.33%	15-18	20.0%	Several times a week	10.83%	Once a month	35.83%
¿10 years	19.16%	251-450	24.16%	¿18	25.0%	Once a week	8.33%	Occasionally	25.83%
		451-700	9.16%			Occasionally	17.50%	Rarely	0%
		¿700	28.33%			Rarely	0%		

Table 5.1: Breakdown of post-study questionnaire answers over 120 online participants.

5.4.1 In-Person Exploratory Interviews

We discuss first the participants, then qualitative and quantitative findings from our exploratory study.

Participants. The 6 male and 4 female participants were between 19-46 years old (M = 28.5, SD = 8.64) and had a diverse background: 1 unemployed, 1 housewife, 1 grocery store owner, 1 high school teacher, 1 engineer, 1 medical resident, and 4 students (bachelors to graduate degree). The participants had between 46 and 1,838 Facebook friends with an average of 488 friends (M = 413, SD = 524.56).

Round 1. We found that 40% (i.e., 20 of the total 50) of the synthetic pending friends shown to the 10 participants, were confirmed in the first round of the experiment. Surprisingly, in 11 of these cases, the participants confirmed such synthetic friends even though they later said that they did not trust the account owner, had a negative feeling about them, or thought they would be abusive in the future. 7 of the 10 participants have confirmed at least one synthetic pending friend, with one participant confirming 4 synthetic pending friends.

In addition, 5 participants did not confirm a total of 7 of their actual pending friend invitations.

Reasons to confirm "synthetics". In the open-ended questionnaire, 4 out of 10 participants **invented stories** to support their desire to confirm synthetic pending friends.

In 5 out of the above 20 synthetic accepted cases, the participants claimed that they either met the synthetic pending friends in real life or they don't remember if they met.

In 4 other cases, the participants claimed that they either attended the same school or lived in the same city. Participant P3 (42 years old immigrant male and grocery store owner with 125 friends in Facebook) said about a synthetic friend:

"She looks like I have seen her in my store before. I don't remember exactly when."

P9 (19 years old female, college student with 624 friends in Facebook) said about another synthetic friend:

"I think I saw this boy in my school. I don't exactly remember, but his face looks so familiar."

P5 (46 years old male, high school teacher with 81 friends in Facebook) said about another synthetic friend:

"She might be one of my students."

Other participants were primarily concerned with **growing their friend list**. For instance, P10 (23 years old male unemployed but college graduate with 1,883 friends in Facebook) said, about two synthetic pending friends:

"I spend a lot of time in Facebook, so I want to grow my friends list. I don't care much when accepting. I have nothing to hide. I will accept", and

"[..] does not look like a bad guy. I will accept. My friend list is growing so fast!"

Similar to Rashtian et al. [RBJB14], we found that Facebook users may be more vulnerable to requests from **perceived attractive accounts**. For instance P4 (22 years old female undergraduate student, with 448 Facebook friends) said about a synthetic friend:

"I don't remember meeting him, but I would not mind accepting him. Looks like a gentle guy."

Curiosity seems to also be a reason. For a synthetic friend, P9 said:

"Honestly saying, this guy looks very handsome. I don't know why he sent me a request but I want to find out more about him."

P1 (30 years old male grad student with 378 friends) said:

"She is an attractive girl. I think I could accept her because she is really pretty. Even though I feel like there is no reason this girl would send me a friend request, I still want to accept this request. I will delete this friend if I find it to be fake."

We found that sometimes the Facebook users accept a request to find out about the person in more detail although they may delete the request at a later time. For instance, P1 said *"I want to accept her first and then check her profile. I mean I want to see where she is from. I will delete this friend if I find it to be fake."*

The **quality of the accounts** counts. For instance, P3 said:

"This is a beautiful woman. Also the name seems Bangladeshi. Maybe she is someone from my community. I think I will decide later",

while P9 said:

"She has many mutual friends with me. I think she might be someone from my school."

Similar to Zou et al. [ZMMS18], we found **perceived harmlessness** is also a factor. For instance, P6 said:

"I feel like she is an innocent girl. May be she is trying to get advice from me by connecting in Facebook. I don't think this girl can do anything bad to me."

Round 2: Impact of questionnaire. We observed that when we asked the open-ended questions during round 2, it changed the participant opinions and decisions. During round

2, only 13 of the 50 synthetic friends were confirmed, down by 35% from the 20 in the 1st round. Only 5 participants confirmed at least one synthetic friend. For instance, participant P5 said:

"At first I thought I will accept her. But now I think I should first send her a message to ask who she is. I am now thinking of some potential abuse after answering the questions."

The questionnaire also had an impact on the number of actual pending friends confirmed in the second round: 11 out of the 29 total actual pending friends (vs. 7 in round 1) were rejected by the participants. For instance, participant P1 explained about a real pending friend:

"He is my ex-girlfriend's uncle. He is a professor. I was confused whether I should accept him. After answering the questions, I now feel that this friend will not add anything to my Facebook experience. So I don't want to accept him now."

In summary, we observed participant vulnerability to synthetic pending friends, being mitigated by a careful assessment of past interactions, common background and perception of each pending friend.

5.4.2 Confirming Actual Invites

Participants in our in-person exploratory study reported several reasons to confirm invitations received from actual pending friends.

Real-life relationship. P8 (28 years old male engineer with 659 Facebook friends) said about an actual pending friend:

"He is my childhood friend. We used to meet a lot when we were younger. I like him and will definitely accept."

P7 (35 years old female housewife with 173 Facebook friends) said:

"She is my neighbor. We have a good family relation. Our kids are also friends."

P3 said about another pending friend:

"He is my nephew. I love him very much. I would definitely accept him."

Professional relationship: P8 (28 years old male engineer with 659 Facebook friends) said about a pending friend:

"He is also an engineer. We met when we worked on a project. I will accept him."

Peer-pressure. P6 (29 years old male, medical resident, with 512 Facebook friends) said about a real pending friend:

"She is my colleague and a senior doctor. I am a little bit scared of her. I will definitely accept because it will send a very bad message if I do not accept. I don't want to risk that."

5.4.3 T-FLock and FLock vs. Control

In the following, we first describe the study participants, then compare the performance of T-FLock and FLock against the control app, on the evaluation scores of Section 5.3.3.

Participants. Table 5.1 summarizes the answers of the 120 participants in the post-study questionnaire. Most participants were young, with only 4.16% being more than 45 years old (M = 25, SD = 7.56), 69.16% male and 30.83% female, representing 5 continents (but 63.33% from Asia) (top 5 countries: Bangladesh 25.0%, India 22.50%, USA 11.66%, Nepal 5.83% and Pakistan 5.0%). The participants had diverse occupations, and education levels, with high-school and bachelors degrees being most frequent (30.83% and 45.0% respectively). 50% of the participants' account were between 3 and 5 years old, while 19.16% were over 10 years.

Participant-declared frequency of Facebook account use was also diverse, ranging from frequent (every hour 24.16%, several times a day 31.66%, once a day 7.50%), to less frequent (several times a week 10.83%, once a week 8.33% and occasionally 17.50%). None of the participants declared to rarely access their accounts. Most participants declared that they received friend invitations either once a week (36.66%), once a month (35.83%), or occasionally (25.83%).

The participants had an average of 12 pending friends (M = 13, SD = 6.71). 27 participants had each 20 or more pending friends. The participants had between 30 and 2,720 Facebook friends, with a average of 574 friends (M = 323, SD = 588.84).

Decision split. Figure 5.6 compares the percentage of synthetic, pending and existing friends that were confirmed, deleted or skipped, in the control, FLock and T-FLock studies. In the control study, we have shown a total of 613 pending friend invitations, (including 403 real pending, 150 synthetic pending, and 60 existing friends) to the 30 participants. Out of the 150 synthetic pending invitations, 76 were confirmed, 59 were deleted and 15 were skipped. This 50.6% confirmation rate exceeds the rate in the exploratory

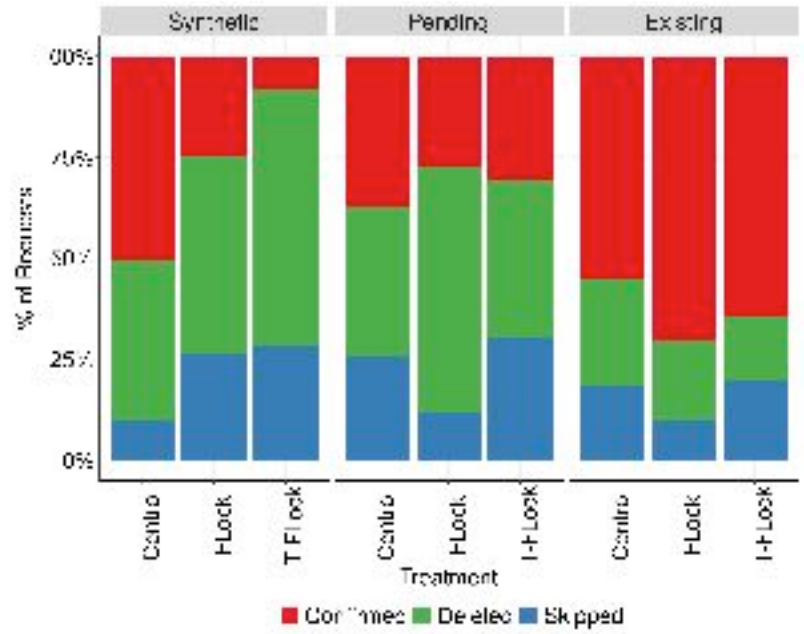
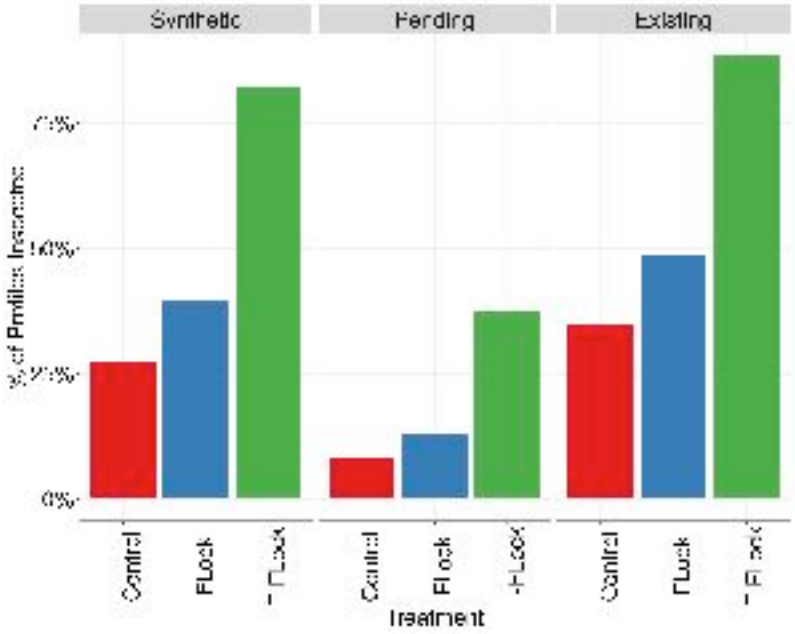


Figure 5.6: Per-treatment distribution of confirmed, deleted, and skipped requests. FLock reduces the percentage of confirmed synthetic friends from 50% (control) to 25%, while T-FLock further reduces it to 8.8%. FLock and T-FLock also reduce the percentage of actual pending friends confirmed, but slightly increase the percentage of confirmed existing friends.

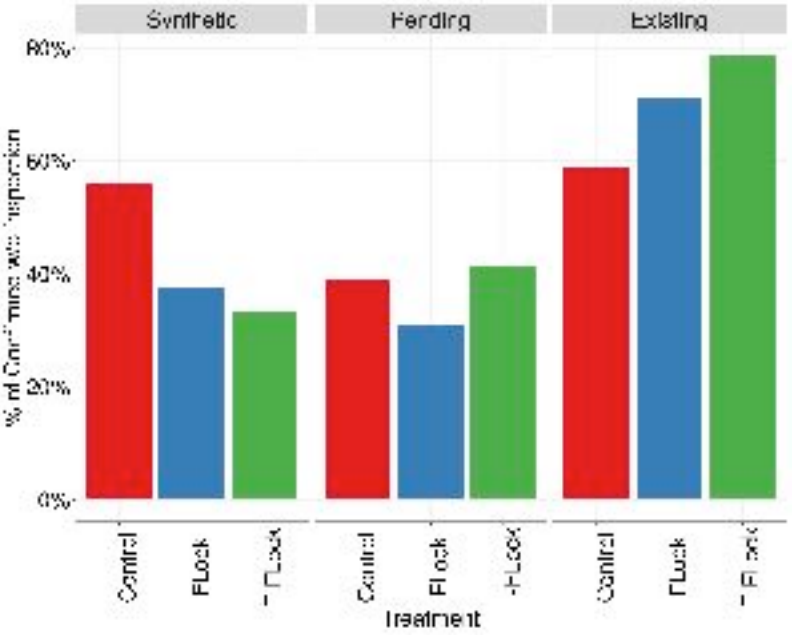
study. More importantly, 26 of the 30 participants confirmed at least one synthetic friend; 2 participants confirmed all the 5 synthetic friends. The average number of synthetic friends accepted is 2.53 (M=3, SD = 1.43).

Further, out of the 403 real pending invitations, 151 were confirmed, 149 were deleted and 103 were skipped for later. We observe a higher rate of confirmation for the synthetic pending invitations than the actual pending invitations (50% vs 37%). Finally, out of the 60 existing friends, 33 were confirmed, 16 were deleted and 11 were skipped.

In the FLock app study, we showed a total 475 pending friend invitations to 30 participants, that include 265 actual pending friends, 150 synthetic pending friends and 60 existing friends. Out of the 150 synthetic pending invitations, 37 were confirmed, 73 were deleted, and 40 were skipped. The average number of synthetic friends accepted is 2.47 (M=1, SD = 1.23). Out of the 265 actual pending invitations, 72 were confirmed,



(a)



(b)

Figure 5.7: (a) Per-treatment distribution of profiles inspected by type of friend request. FLock and T-FLock significantly increase the percentage of all of synthetic, actual pending and existing friends whose profiles were inspected. (b) Per-treatment distribution of profiles confirmed without inspection. While during control, 55.9% of synthetic friends were confirmed blindly, this is significantly reduced to 37% and 34% by FLock and T-FLock respectively.

161 were deleted and 32 were skipped for later. Out of the 60 existing friends, 42 were confirmed, 12 were deleted and 6 were skipped.

In the T-FLock app study, we showed 1,193 friend invitations to the 60 participants: 773 actual pending friends, 300 synthetic pending friends and 120 existing friends. Out of the 300 synthetic pending invitations, only 24 were confirmed, 190 were deleted, and 86 were skipped. On average 0.3 (M=0, SD = 0.91) synthetic friends were accepted. Out of the 773 actual pending invitations, 237 were confirmed, 302 were deleted and 234 were skipped for later. Lastly, out of the 120 existing friends, 77 were confirmed, 19 were deleted and 24 were skipped.

We observe thus a significantly lower rate of confirmation (24.7%) in the FLock study, for the synthetic pending invitations, than the control experiment (50.6%). A two-proportion one sided z-test ($H_0 : p_{cb} = p_{fb}$ vs. $H_a : p_{cb} > p_{fb}$ where p_{cb} and p_{fb} are the proportion of synthetic friends confirmed under the control and FLock treatment respectively) produced a $p - value = 0.00$ ($Z = 4.64$), suggesting H_a, i.e., that the proportion of synthetic pending friends confirmed is statistically significantly lower for the FLock group. Even further, the portion of accepted synthetic friend requests in the T-FLock (8%) study was significantly lower than that of FLock (24.7%) study ($Z = 4.86$, $p - value = 0.00$).

In addition, we observe that in the FLock and T-FLock studies, participants confirmed fewer actual pending friend invitations, compared to the control study: 30.66% in T-FLock and 27.16% in FLock vs. 37.46% in the control study. A two-proportion one sided z-test revealed the portion of confirmed actual pending friends in the FLock study is significantly lower than that of control study ($Z = 2.78$, $p - value = 0.00$), and the portion of confirmed actual pending friends in the T-FLock study is significantly lower than that of the control study ($Z = 2.35$, $p - value = 0.00$). However, this test did not

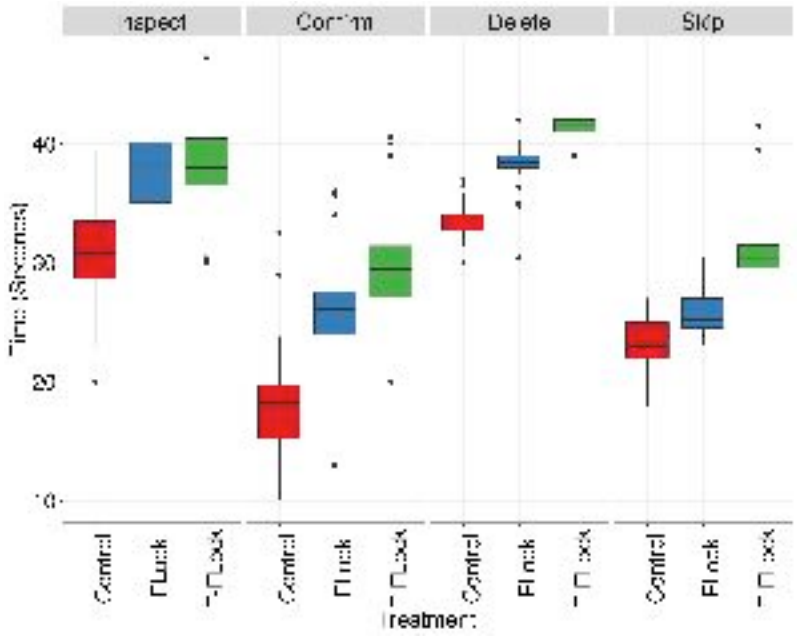
show any significant difference between the portion of confirmed actual pending friend request in FLock and T-FLock.

In addition, in all the studies, a majority of the existing friends presented as pending friends, were accepted by the participants: 55%, 70% and 64.16% respectively for control, FLock and T-FLock studies.

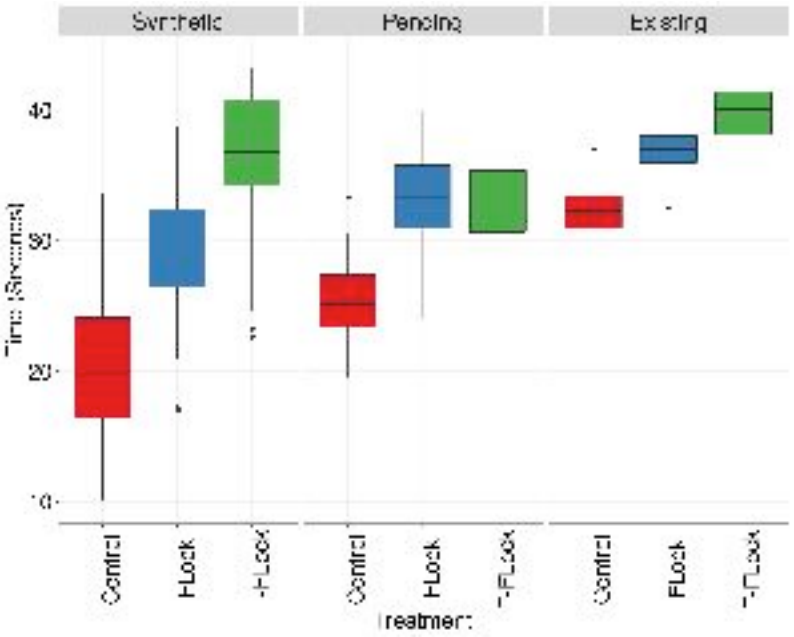
Profile inspection. Figure 5.7(a) compares the percentage of profiles inspected, grouped by the type of friend request, in the control, FLock and T-FLock studies. In the control study, only 94 of the 613 pending friend profiles were inspected by 28 participants. 2 participants did not inspect any profile. Participants inspected 27.33% of the synthetic, 7.94% of the actual pending, and 35% of the existing friend invitations.

Figure 5.8(a) (left most column) compares the average time taken by the control, FLock and T-FLock participants to inspect profile of pending friends. The control study participants took an average of 20.11s (M = 21, SD = 4.04) to inspect the profile of an actual pending friend, 31.95s (M = 33, SD = 5.44) for a synthetic friend, and 43s (M = 43, SD = 2.52) for an existing friend. Thus, the participants took more time to inspect the profile of a synthetic pending friend than an actual pending friend ($t = 11.33$, $p - value = 0.00$). We conjecture that this occurs because the participants were more interested in the synthetic pending friends, or they had inspected some of their actual pending friends at an earlier time. Further, we conjecture that the participants took the most time to inspect the profile of their existing friends presented as pending, because such friends have longer Facebook timelines, which take longer to inspect.

In the FLock study, 122 of the 475 pending friend profiles were inspected by the 30 participants, in an average of 37.46s (M=37.5, SD=7.91). Participants inspected 39.33%, 48.33%, and 12.45% of synthetic, existing and pending friend requests respectively. Similar to the control study, a majority of the actual pending friend requests were not inspected. However, compared to the control study, significantly more synthetic profiles



(a)



(b)

Figure 5.8: (a) Per-treatment average time taken by the participants to make a decision, per decision type. FLock increased the average confirmation time by 39.6% compared to the control, while T-FLock increased it by 57.8%. (b) Per-treatment average time to make a final decision (confirm, delete, or skip) per type of pending friend. Compared to the control study, FLock increased the participant decision time for synthetic pending friends by 40.5%, while T-FLock increased it by 75.7%.

were inspected by participants in the FLock study (59 vs 41, $Z = 2.20, p-value = 0.01$). In addition, the FLock study participants took more time on average to inspect the profile of an actual pending friends (26.81s vs 20.11s, $t = 8.17$, $p - value = 0.00$), a synthetic pending friend (38.24s vs 31.95s, $t = 7.21$, $p - value = 0.00$) and an existing Facebook friend (47.93s vs 43s, $t = 6.6$, $p - value = 0.00$).

In the T-FLock study, 643 profiles out of 1,193 were inspected by the 60 participants. Particularly, 82%, 88.33% and 37.52% of the synthetic, existing and pending friend requests were inspected respectively. Two-sample proportion z-test shows that the percentage of inspected profiles of each friend request type in T-FLock is significantly larger than that of control and FLock studies. In addition, compared to the FLock study, participants in T-FLock took significantly longer time on average to inspect each type of profile: actual pending friends (31.14s vs 26.81s, $t = 8.8$, $p - value = 0.00$); synthetic pending friends (41.14s vs 38.24s, $t = 6.3$, $p - value = 0.00$); real pending friends (49.99s vs 47.93s, $t = 2.5$, $p - value = 0.00$).

Blind confirmations. Figure 5.7(b) shows the percentage of pending friends (by type) who were confirmed blindly, i.e., without prior profile inspection. During the control study, 55.96%, 39.08% and 59.97% of synthetic, actual pending and real friend requests that were confirmed by participants, were decided blindly (i.e. without prior inspection). We observe that the number of blindly confirmed synthetic and real pending friends are significantly higher than that of actual pending friends ($z = 3.13$ and $z = 2.40$ respectively with $p - value = 0.00$).

In the FLock study, 37.36%, 31.03% and 70.97% of confirmed synthetic, actual pending and existing friends were decided blindly. We use a two-proportion one sided z-test to compare the portion of blindly confirmed synthetic profiles. Compared to the control study, the portion of synthetic friends confirmed blindly was significantly reduced (37.36% vs 55.96%, $z = 2.62$, $p - value = 0.00$).

Finally, in the T-FLock study, 33.33%, 41.41% and 78.57% of the confirmations for synthetic, actual pending and real friends profiles were decided blindly. Compared to the participants in the control study, T-FLock significantly reduced the portion of blindly confirmed synthetic profiles (33.33% vs 55.96%, $z = 2.72$, $p - value = 0.00$).

Decision timing. We have also analyzed the time taken by the participants to make a decision for a friend invitation. We computed this time as the difference between the decision time and the time of the previous event (e.g., return from profile inspection, decision on previous pending friend).

Figure 5.8(a) compares the average time taken by the control, FLock and T-FLock participants to make a decision, per decision type. In the control study, the participants took an average 24.77s (M = 24, SD = 8.45) to make a decision for a friend request. They took an average of 18.47s (M = 18, SD = 6.85) to confirm a friend invitation and 22.57s (M = 23, SD = 4.0) to skip, but significantly longer, i.e., 33.36s (M = 34, SD = 3.04) to delete a friend invitation ($p - value = 0.00$).

The FLock participants took on average, 39.6% more time to confirm, 16.1% more time to skip, and 13.8% more time to delete pending friends, when compared to the control participants. A substantial increase was further induced by T-FLock, where participants took on average, 57.8% more time to confirm, 34.7% more time to skip and 24.1% more time to delete a pending friend, than in the control study.

Figure 5.8(b) compares the average time taken by the control, FLock and T-FLock participants to make a decision for each type of friend request (synthetic, actual pending and existing). A Welchs t-test shows that, in the control study, the average time to make a decision for a synthetic pending friend, 20.49s (M = 14, SD = 10.79), was significantly shorter than the 32.20s (M = 32, SD = 2.64) for existing friends ($t = 12.34$, $p - value = 0.00$) and also than the 25.27s (M = 24, SD = 7.00) for actual pending friends ($t = 5.02$,

Variable 1	Variable 2	Control	FLock	T-FLock
Age	# delete	-0.02	0.02	0.09
Device release price	# confirm synthetic	-0.17	-0.09	0.08
Usage frequency	# skip	-0.33*	0.25	0.09
Usage frequency	Frequency of receiving friend request	0.38*	0.38*	0.28*
Usage frequency	# confirm synthetic	0.06	0.31*	-0.08
# friends	# delete synthetic	0.4	0.32*	0.07
# friends	# confirm synthetic	-0.13	-0.35*	0.08
Avg time to inspect & decide	# confirm synthetic	-0.44*	-0.31*	-0.42*
Avg time to inspect & decide synthetic	# delete synthetic	0.63*	0.39*	0.50*

Table 5.2: Kendall tau-b correlation coefficient between different variables in control, FLock and T-FLock. * indicates statistically significant result at $\alpha = 0.05$.

$p - value = 0.00$). This occurs even though the participants took more time to inspect the profiles of synthetic than of real pending friends (see Profile inspection above).

In the FLock study, this relationship was preserved. However, compared to the control study, the participants took significantly ($p-value = 0.00$) more time to make a decision for an actual pending friend (33.04s vs 25.27s, $t = 14.0$), synthetic pending friend (28.79s vs 20.49s, $t = 6.67$) and existing friend (36.81s vs 32.2s, $t = 9.15$). In the T-FLock study, we found further improvement: when compared to the FLock study, the T-FLock participants took significantly ($p - value = 0.00$) more time to make a decision for a synthetic friends (36.01s vs 28.79s, $t = 7.31$) and existing friends (39.78s vs 36.81s, $t = 6.88$). However, they took almost the same time to decide about actual pending friend requests (33.51s vs 33.04s, $t = 0.95$, $p - value = 0.34$).

In figure 5.10, we include and discuss the timelines (i.e., the times when each participant open or closed a pending friend profile, or made a decision) of 1 participant from each of the control, FLock and T-FLock studies, who had the largest number of pending friends in each study (20 actual pending, 5 synthetic and 2 existing).

5.4.4 Predicting User Decisions

We first study associations between several features extracted from user accounts and friend invitations, then zoom-in into device model impact on user decisions, then present results over a classifier that we developed to predict user decisions.

Correlation analysis. A Fisher exact test did not show any significant relation between the gender of the participants and their decision to confirm at least one synthetic friend request in the control, FLock, and T-FLock studies ($p-value = 1, 0.38$, and 0.72 respectively). Similarly, a χ^2-test did not reveal any relation between the education level of the participants and their decision to confirm at least one synthetic friend request (p-values for control, FLock and T-FLock are 0.73, 0.81, 0.51 respectively). Further, we computed Kendall tau-b correlation coefficients (see Table 5.2, first row) and found no significant correlation between the age of the participants and the number of deleted or confirmed friend requests.

Table 5.2 further shows other Kendall tau-b correlation coefficients for several observations in the control, FLock and T-FLock studies. Perhaps unsurprisingly, the declared frequency of Facebook use is strongly positively correlated with the declared frequency of receiving friend requests in the control, FLock and T-FLock studies. The declared frequency of use is negatively correlated to the number of skips in the control study, but positively correlated to the number of skips and confirmed synthetics in the FLock study.

In the FLock study, the participant number of friends was strongly correlated with the number of synthetic friends deleted (positive) and confirmed (negative).

In the 3 studies, the number of confirmed synthetic friends significantly decreased as the participants took more time to inspect and make decision about the friend requests. Similarly, the number of deleted synthetic friend requests was strongly positively correlated with the decision and inspection time. This provides evidence that participant en-

Variable 1	Variable 2	Control	FLock	T-FLock
Age	# delete	-0.02	0.02	0.09
Usage frequency	# confirm	0.14	-0.04	-0.08
Usage frequency	# skip	-0.33*	0.25	0.09
Usage frequency	# delete	0.11	0.08	-0.14
Usage frequency	# delete synthetic	0.15	0.21	-0.04
Usage frequency	Frequency of receiving friend request	0.38*	0.38*	0.28*
Usage frequency	# confirm synthetic	0.06	0.31*	-0.08
# friends	# delete synthetic	0.4	0.32*	0.07
# friends	# confirm synthetic	-0.13	-0.35*	0.08
# confirm synthetic	Device release price	-0.17	-0.09	0.08
Declared frequency of received invites	# confirm	0.03	-0.09	-0.04
Avg time to inspect profile	# confirm	-0.10	-0.12	0.20
Avg time to inspect profile	# confirm synthetic	-0.03	0.19	-0.04
Avg time to inspect profile	# delete	-0.01	-0.11	-0.18*
Avg time to inspect & decide	# confirm synthetic	-0.44*	-0.31*	-0.42*
Avg time to inspect & decide synthetic	# delete synthetic	0.63*	0.39*	0.50*

Table 5.3: Kendall tau-b correlation coefficient between different variables in control, FLock and T-FLock. * indicates statistically significant result at $\alpha = 0.05$.

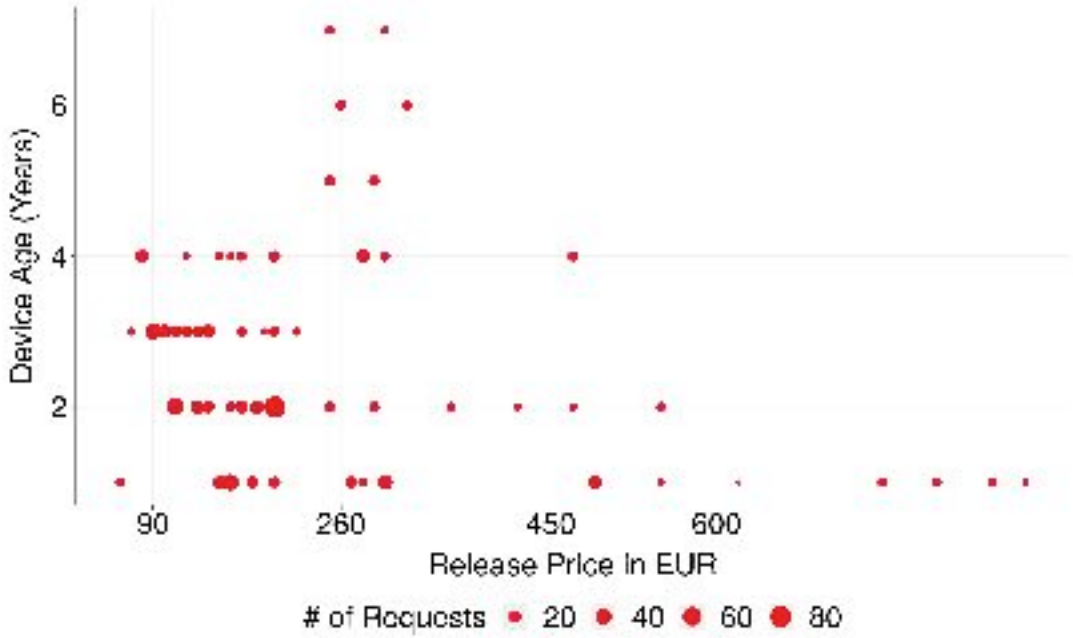


Figure 5.9: Scatter plot of device release price (EUR) vs. model age (years) at the time of study, for each of 756 confirmed requests processed in the three online user studies, from 67 unique device models. Most devices are old and low-end (32.27%) or middle-aged and low-end (30.55%).

gagement in the decision making process, decreases friend spam vulnerability. Table 5.3 in the appendix includes more studied correlations.

Impact of device models. We investigate correlations between participant decisions for pending friend invitations, and the device model age and release price, as perhaps indicators of social status. The 120 participants in our studies have used 67 unique Android device models (Samsung Galaxy J7 Pro, Xiaomi Mi A1, Samsung Galaxy S6, Lava Iris 820, and Gionee A1 Lite were the most popular), released between 2012 and 2018. We investigated relationships between the device model price and age, and the tendency of users to confirm, delete or skip friend requests. Figure 5.9 shows the scatterplot of the device model age at the time of the study and release price (EUR), over 756 confirmed friend invitations across all three user studies. We obtained similar plots for the delete and confirm decisions.

We categorize devices into (1) low, mid, or high-end, if their release price is in the range 0-260, 260-450, and $>$ 450 EUR, respectively [Tri17], and (2) fresh ($<$ 1 year), middle-aged (1-2 years), and old ($>$ 2 years). 72.62% of invitations shown in Figure 5.9 were processed from low-end, 16.73% from mid-end, and 10.64% from high-end devices; 43.8% were from old and 24.94% from new models. We found no statistically significant correlation between the number of confirmed requests and the device release price or age (see also Table 5.2).

The predictor. We have developed a classifier to predict user decisions on friend invitations. Such a classifier can help (1) determine users who are vulnerable to friend spam, without asking them to make decisions and (2) predict which friend invitations the user will confirm, as soon as they are received, and before being shown to the user. Thus, we consider a scenario where the classifier is trained on past user decisions.

Training set tuples for the classifier consist then of a user and a pending friend. For each such tuple, we extracted features from the available data: (1) the user declared gender, age, sharing preferences, frequency of invitation reception and account use, region, occupation, education, account age, (2) extracted number of friends, number pending invitations, device price and age, (3) the pending friend type (synthetic, pending, existing) and the order of the pending friend in the interface, and (4) the percentage of past confirm, delete and skip decisions. For a user and test pending friend, the classifier outputs one of 3 classes, confirm, delete or skip.

We used k-fold cross-validation for time series [AC10], where we train on the first k decisions of each user, and test on the rest of the invitations. k is between 1 and the maximum number of pending friends per participant. We have used only tuples from the control study, where participants had the unadulterated, Facebook-like view of the friend invitation interface, i.e., 613 tuples from 30 participants.

Algorithm	k	Precision	Recall	F1
GBM	**5**	**84.19%**	**82.68%**	**83.43%**
	16	**90.68%**	**90.68%**	**90.68%**
RF	**5**	**83.19%**	**81.64%**	**82.41%**
	26	94.23%	92.31%	93.26%
SVM	5	71.98%	71.71%	71.84%
	15	82.16%	80.00%	81.07%
KNN	5	54.93%	49.68%	52.17%
	26	82.05%	76.92%	79.40%

Table 5.4: Precision, Recall, and F1 measure for different machine learning classifiers as a function of the history length k. Second row shows the k for which the performance was maximum. GBM and RF achieve an F1 of 83.43% and 82.41% respectively considering only 5 previous friend requests, while GBM further improves F1 up to 90.68% when $k = 16$.

Table 5.4 shows the performance of the predictor when using 4 supervised learning classifiers. Gradient Boosting Machine (GBM) outperformed RF, SVM, and KNN, achieving an F1 of 83.43% and 90.68% when trained on only the first 5 and 16 friend invitation decisions from each user, respectively.

The top 5 most impactful features where the order of the pending friend in the interface, the percentage of past confirm, delete and skip decisions, the friend type and the user's friend count.

5.5 Study Timelines

We picked 1 participant from each of the control, FLock and T-FLock studies, who had the largest number of pending friends in each study (20 actual pending, 5 synthetic and 2 existing). Figure 5.10 compares their timelines, i.e., the times when each participant open or closed a pending friend profile, or made a decision. We observe that the control participant (Figure 5.10(a)) confirmed 4 of the 5 synthetic friends without inspection, inspected only 2 of the friends (1 synthetic and 1 pending) and took 669s to process all the pending friends (avg. 24.77s per friend). The FLock participant (Figure 5.10(b))

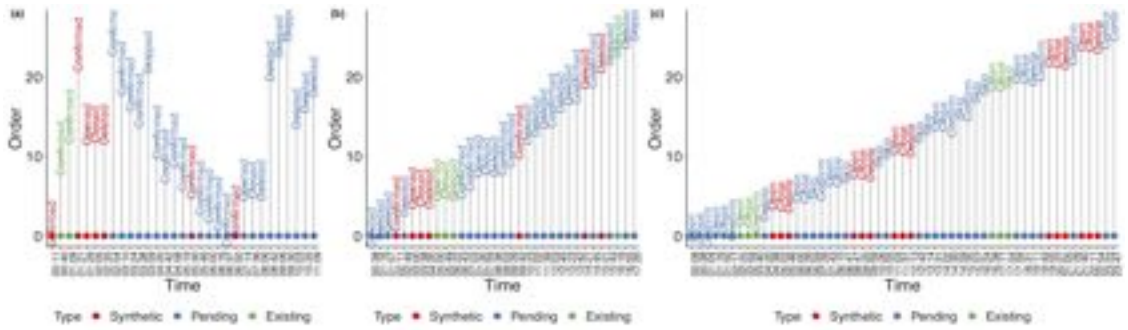


Figure 5.10: Participant timeline of the control study: (a) The timeline plot of the participant having maximum number of total pending friends in the control experiment. (b) The timeline plot of the participant having maximum number of total pending friends in the FLock experiment. (c) The timeline plot of the participant having maximum number of total pending friends in the T-FLock experiment. All of the 4 evaluation scores improved from control to FLock and FLock to T-FLock.

inspected then deleted 1 synthetic friend, deleted 1 synthetic friend without inspection, but accepted 3 synthetic friends without inspection. The participant also inspected 1 pending friend and 1 existing friend, an improvement from the previous participant. For the same number of processed pending friends, the participant also took 1,040s to process all the pending friends (avg. 38.51s per friend), 55% longer than the control participant.

In contrast, the T-FLock participant (Figure 5.10(c)) inspected then deleted all of the synthetic friends, inspected 6 of the 20 actual pending friends, and both existing friends. We also observe a substantially longer time, 1,427s, that the participant took to process the same number of pending friends as the control and FLock participants (52.85s per friend).

5.6 Discussion

Relevance of our work. 78.4% of the participants in our studies had more than 150 friends (and up to 2,720), which is the average number of stable relationships that humans can maintain [Dun92]. This suggests that the more friends a user has (over 150),

the higher the chance that the user's existing friends, and the friend invites she receives, contain spam.

We believe that T-FLock's substantial reduction in the total number of confirmed synthetic friends and in the number of participants who confirmed at least 1 such friend, is essential in protecting against friend spam and subsequent attacks (§ **??**). We further emphasize the importance of T-FLock's increase in the number of inspected pending friend profiles, and in the time taken to inspect them and make a decision. This reveals that users can be trained to carefully consider each pending friend. This is confirmed by our interview study, where we found that taking the time to consider common background, can encourage users to even question invitations received from people with whom they have intricate common background.

Deleting existing friends. We asked the FLock participants for their reasons to delete 8 of their existing friends, that we presented as pending friends. In four cases, the participants recognized that this person was already their friend and hence they chose to delete the invitation. However, in two cases, the participants mentioned that they did not want that person to be their friend anymore. In one case, the participant reported that she did not even know that person, while in the last case, the participant reported that the request was from a fake profile.

This emphasizes the importance of properly vetting pending friends, and suggests the need for solutions that continue this vetting process for existing friends. We argue thus that contrary to Facebook's mantra of connecting people at any cost [Fac18], carefully vetting online friends can reduce friend spam vulnerabilities, and increase user well-being [DVK15, FE16, CFL+16].

Friend spam detection. Given access to relevant information, e.g., the history of friend invitations sent and received by an account, (including their timing), the history of changes to the account, the IP addresses and devices used to login to the account, the nature of the

account's posts, the account age and level of activity, we conjecture that one could predict if (1) a user account was used to send friend spam, (2) a user has common background with a pending friend, and (3) it is safe to confirm that pending friend. A focus group we organized with 6 participants, proposed using this information to (1) provide more effective attractors, e.g., highlight the "Delete" button for suspicious pending friends, (2) pop-up a menu of friend lists where the pending friend could be added, if the user decides to confirm, including a list for "unsafe" friends, with restricted permissions to the user's information, or even (3) a *spam folder* to store the suspicious pending friend invitations.

In summary, instead of forensically detecting friend spam, in this chapter we leverage the observation that victims are the first line of defense, and develop online techniques that involve the user in the spam detection process, at the time when the spam attack takes place. Our approach seeks to encourage users to carefully evaluate their pending friends, and trains them to avoid accepting strangers.

5.7 Limitations

Our exploratory study was performed on a small set of participants, all recruited from a single US city. Our crowdsourced studies were performed with a biased population that consisted of participants with access to mobile devices, were mostly young (49.2% were 18-24 years old), had more than 30 Facebook friends and 1 pending friend, and understood English. Thus, we do not claim that our results generalize to the entire Facebook population. However, the participants had a diverse education and occupational background, and the results of the exploratory study were confirmed by the control study, even though performed on different population samples. Further, Redmiles et al. [RKM19] revealed that regarding security and privacy experiences, 18-29 year-old participants re-

cruited from Amazon Mechanical Turk more closely matched the U.S. population than did participants from a census-representative web-panel sample.

We did not implement participant decisions to confirm or delete actual pending friends. Instead, we notified participants at the end of the study (in the payment code screen), that we did not implement their decisions, and recommended them to visit their Facebook accounts to implement these decisions. The reason for this is the high latency of the operation, which would have made our systems unresponsive, affecting user experience and thus the quality of our results.

In the control study, we found 8 instances (2 pending, 5 synthetic, 1 existing) where a single profile was inspected twice, by 4 different participants. In FLock, we found 4 such instances, all for synthetic pending friends, from 4 different participants. We conjecture, but did not seek participant feedback, that such "profile double-checks" for synthetic friends expose the psychological conflict presented by well-crafted stranger accounts. We observe then that among the 60 participants in the T-FLock study, there was only 1 profile double-check, for an actual pending friend.

5.8 Conclusions

In this chapter we have shown, through exploratory interviews and a control study with online participants, that Facebook users continue to be vulnerable to friend spam attacks. We have developed new designs of interfaces to process pending invitations. We have shown, through studies with online participants, that when compared to a control system, our designs significantly reduce the number of synthetic friends confirmed, and increase the number of profiles inspected and time taken to decide. We reveal potential to predict user decisions on pending invitations, before the user makes them.

5.9 Appendix

5.9.1 The Questionnaire

Interview Study Questionnaire

A. Interaction

- Do you know this person?
- Have you ever met this person in real life?
- Do you remember where you met this person for the first time?
- Do you remember when was that?
- How frequently do you meet this person?
- Do you remember the last time when you met this person?
- Where did this happen?
- When did this happen?
- Do you know this person online? For instance, have you ever interacted with this person on email, chat, SMS, Skype, phone, forum, or newsgroup? Have you visited their webpage if they have one?
- Do you remember the first time when you had such an interaction with this person?
- How frequently do you interact with this person?
- How recent was your last interaction?
- How did you interact?

B. Perception

- Do you have any negative feelings about this person?
- Do you know if this person was ever abusive toward you or others?
- Do you think this person could be abusive? If yes, in which way?
- Do you trust this person?

C. Common Background

- Do you share any common interest?
- Have you ever lived in the same city?
- Have you ever attended the same school?
- Have you ever worked together?

Post-Study Questionnaire

(a) What is your age?

- 18-24 years old
- 25-34 years old
- 35-44 years old
- 45-54 years old
- 55-64 years old
- Other:

(b) What is your gender?

- Female
- Male
- Other:

(c) What is your occupation?

- Employed
- Student
- Homemaker
- Retired
- Unemployed
- Other:

(d) What is your highest level of education?

- Elementary
- High school
- College
- Bachelors
- Graduate
- Other:

(e) For how long have you had your Facebook account?

- Less than a year

- 1-2 years
- 3-5 years
- 6-10 years
- More than 10 years

(f) How frequently do you use Facebook?

- Every hour
- Several times a day
- Once a day
- Several times a week
- Once a week
- Occasionally
- Rarely

(g) How often do you receive friend requests in Facebook?

- Almost everyday
- At least once in every few days
- At least once a week
- At least once a month
- Occasionally
- Rarely

CHAPTER 6

CONCLUSIONS

6.1 Summary

Detection of friend abuse and prevention is challenging and exhaustive. In this dissertation, we have introduced and studied AbuSniff, the first friend abuse detection and defense system for Facebook. We have developed a compact stranger and abuse detection questionnaire. We have introduced and studied rules to convert questionnaire answers to defensive actions.

While participants tended to agree to defend against perceived timeline and news-feed abusers, they needed more information and flexible options in order to agree to defend against strangers. Further, we have shown that on data we collected, AbuSniff can predict Facebook friends who are perceived as abusive or strangers, and against whom users are willing to take a defensive action.

We have introduced social networking based features and shown that supervised learning algorithms can use them to accurately predict users questionnaire answers and defense choices. We have implemented the AbuSniff system and investigated user behaviors through 7 studies with 263 participants. AbuSniff increased participant willingness to reject invitations from perceived strangers and abusers, as well as awareness of friend abuse implications and perceived protection from friend abuse.

We have shown, through exploratory interviews and a control study with online participants, that Facebook users continue to be vulnerable to friend spam attacks. We have developed FLock and T-FLock, new designs of user interfaces to process pending invitations. We have shown, through multiple user studies with online participants, that when compared to a control system, they significantly reduce the number of synthetic friends confirmed, and increase the number of profiles inspected and time taken to decide.

6.2 Future Work

The research described in this dissertation can be improved in three possible directions. Firstly, we plan to continue the exploration of user perception of abuse for *pending* friends, from whom the user has received a friend invitation but has not yet accepted. we have done some preliminary work, but we can aim to automatically detect the potentially abusive pending friend in Facebook. This is a more difficult problem, as we lack access to a common history between the user and the pending friend. However, while users may find it hard to detect abuse, we also conjecture that users will be more willing to closely scrutinize pending friends and ignore invitations from perceived strangers. Secondly, another future work direction is to customize AbuSniff for known vulnerable populations. Participatory design studies [AV16] could help identify approaches suitable for adolescents, including questions to identify cyberbullying perception and personalized defense actions. Further, we plan to investigate techniques to identify abusive posts (that contain harassment, bullying, links to malware [NTT+16, DJH+12]), without breaking the terms of service of social networking sites. One direction we plan to investigate is to identify privacy preserving techniques to extract computational linguistic features from posts, without direct access to the posts. Finally, we can extend our research ideas to other online social networks like Twitter, LinkedIn, Instagram, Tumblr, SnapChat etc. Since the abuse problem is widespread due to the unchecked access of the user data to the connections and friends in those social networks, it is highly likely that the same solutions can be applied in those networks.

BIBLIOGRAPHY

[AAA+17] Muhammad Al-Qurishi, Mabrook Al-Rakhami, Atif Alamri, Majed A. Al-Rubaian, Sk. Md. Mizanur Rahman, and M. Shamim Hossain. Sybil Defense Techniques in Online Social Networks: A Survey. *IEEE Access*, 5:1200–1219, 2017.

[Abu17a] AbuSniff app. `https://goo.gl/LBWNWZ`, 2017.

[Abu17b] AbuSniff source code. `https://goo.gl/SZ7jrT`, 2017.

[AC10] Sylvain Arlot and Alain Celisse. A Survey of Cross-validation Procedures for Model Selection. *Statist. Surv.*, 4:40–79, 2010.

[AG17] Julia Angwin and Hannes Grassegger. Facebooks secret censorship rules protect white men from hate speech but not black children. CNBC Tech, `https://goo.gl/8kZhgD`, 2017.

[ARK15] Amos Azaria, Ariella Richardson, and Sarit Kraus. An agent for deception detection in discussion based environments. In *Proceedings of CSCW*, 2015.

[Aro16] Jessikka Aro. The cyberspace war: propaganda and trolling as warfare tools. *European View*, 15(1), 2016.

[AS10] Murad Batal Al-Shishani. Taking al-qaeda's jihad to facebook. *The Jamestown Foundation: Terrorism Monitor*, 8(5):3, 2010.

[AV16] Zahra Ashktorab and Jessica Vitak. Designing cyberbullying mitigation and prevention solutions through participatory design with teenagers. In *Proceedings of CHI*, 2016.

[BBC17a] BBC. Russia-linked posts 'reached 126m Facebook users in US'. BBC News, `https://goo.gl/2qy5et`, 2017.

[BBC17b] BBC. Theresa May accuses Vladimir Putin of election meddling. `https://goo.gl/EtMSRF`, 2017.

[Beh] Behind The Name: Random Name Generator. `https://www.behindthename.com/random/`.

[BHI+08] Garrett Brown, Travis Howe, Micheal Ihbe, Atul Prakash, and Kevin Borders. Social networks and context-aware spam. In *Proceedings of the ACM CSCW*, 2008.

[BLKC+13] Cristian Bravo-Lillo, Saranga Komanduri, Lorrie Faith Cranor, Robert W. Reeder, Manya Sleeper, Julie Downs, and Stuart Schechter. Your Attention Please: Designing Security-decision UIs to Make Genuine Risks Harder to Ignore. In *Proceedings of the Ninth Symposium on Usable Privacy and Security*, SOUPS '13, pages 6:1–6:12, New York, NY, USA, 2013. ACM.

[BLS+15] Yazan Boshmaf, Dionysios Logothetis, Georgos Siganos, Jorge Lería, José Lorenzo, Matei Ripeanu, and Konstantin Beznosov. Integro: Leveraging Victim Prediction for Robust Fake Account Detection in OSNs. In *Proceedings of the Annual Network and Distributed System Security Symposium*, 2015.

[BMBR11] Yazan Boshmaf, Ildar Muslukhov, Konstantin Beznosov, and Matei Ripeanu. The Socialbot Network: When Bots Socialize for Fame and Money. In *Proceedings of the 27th Annual Computer Security Applications Conference*, ACSAC '11, pages 93–102, New York, NY, USA, 2011. ACM.

[Boy07] Danah Boyd. Why youth (heart) social network sites: The role of networked publics in teenage social life. *MacArthur foundation series on digital learning–Youth, identity, and digital media volume*, pages 119–142, 2007.

[BSJ14] Piotr Bródka, Mateusz Sobas, and Henric Johnson. Profile Cloning Detection in Social Networks. In *European Network Intelligence Conference*, pages 63–68, 2014.

[CBDL17] Justin Cheng, Michael S. Bernstein, Cristian Danescu-Niculescu-Mizil, and Jure Leskovec. Anyone can become a troll: Causes of trolling behavior in online discussions. In *ACM CSCW*, 2017.

[CF16] Hichang Cho and Anna Filippova. Networked privacy management in facebook: A mixed-methods and multinational study. In *Proceedings of the 19th ACM CSCW*, 2016.

[CFL+16] Wu Chen, Cui-Ying Fan, Qin-Xue Liu, Zong-Kui Zhou, and Xiao-Chun Xie. Passive Social Network Site Use and Subjective Well-being: A Moderated Mediation Model. *Computers in Human Behavior*, 64:507–514, 2016.

[CRAD16] Bogdan Carbunar, Mizanur Rahman, Mozhgan Azimpourkivi, and Debra Davis. GeoPal: Friend Spam Detection in Social Networks Using Private Location Proofs. In *Proceedings of the IEEE International Conference on Sensing, Communication and Networking (SECON)*, 2016.

[CSSG17] Eshwar Chandrasekharan, Mattia Samory, Anirudh Srinivasan, and Eric Gilbert. The bag of communities: Identifying abusive behavior online with preexisting internet data. In *Proceedings of CHI*, 2017.

[CSY+15] Qiang Cao, Michael Sirivianos, Xiaowei Yang, , and Kamesh Munagala. Combating Friend Spam Using Social Rejections. In *Proceedings of the IEEE ICDCS*, 2015.

[DJH+12] Karthik Dinakar, Birago Jones, Catherine Havasi, Henry Lieberman, and Rosalind Picard. Common sense reasoning for detection, prevention, and mitigation of cyberbullying. *ACM TiiS*, 2012.

[DJK15] Cailing Dong, Hongxia Jin, and Bart P Knijnenburg. Predicting Privacy Behavior on Online Social Networks. In *Proceedings of the International AAAI Conference on Web and Social Media*, pages 91–100, 2015.

[DLHH09] Bernhard Debatin, Jennette P Lovejoy, Ann-Kathrin Horn, and Brittany N Hughes. Facebook and online privacy: Attitudes, behaviors, and unintended consequences. *Journal of Computer-Mediated Communication*, 15(1):83–108, 2009.

[Dun92] Robin IM Dunbar. Neocortex size as a constraint on group size in primates. *Journal of human evolution*, 22(6):469–493, 1992.

[DVK15] Dian A De Vries and Rinaldo Kühne. Facebook and Self-perception: Individual Susceptibility to Negative Social Comparison on Facebook. *Personality and Individual Differences*, 86:217–221, 2015.

[EN11] Rachel A Elphinston and Patricia Noller. Time to face it! facebook intrusion and the implications for romantic jealousy and relationship satisfaction. *Cyberpsychology, Behavior, and Social Networking*, 14(11):631–635, 2011.

[Fac18] Growth At Any Cost: Top Facebook Executive Defended Data Collection In 2016 Memo And Warned That Facebook Could Get People Killed. [BuzzFeed News] `https://tinyurl.com/y3bp8x99`, March 2018.

[FE16] Eline Frison and Steven Eggermont. "Harder, better, faster, stronger": negative comparison on Facebook and Adolescents' life satisfaction are reciprocally related. *Cyberpsychology, Behavior, and Social Networking*, 19(3):158–164, 2016.

[GHW+10] Hongyu Gao, Jun Hu, Christo Wilson, Zhichun Li, Yan Chen, and Ben Y Zhao. Detecting and characterizing social spam campaigns. In *Proceedings of the ACM IMC*, 2010.

[GKAN17] Nir Grinberg, Shankar Kalyanaraman, Lada A. Adamic, and Mor Naaman. Understanding feedback expectations on facebook. In *Proceedings of the ACM CSCW*, 2017.

[Gre17] Andy Greenberg. Now anyone can deploy google's troll-fighting ai. Wired Magazine, 2017.

[Guy18] Jessica Guynn. Facebook's Mark Zuckerberg has promised to protect user privacy before. Will this time be different? [USA Today], `https://tinyurl.com/yxhuhqkj`, April 2018.

[Hea17] Alex Heath. Facebook quietly updated two key numbers about its user base. Business Insider, `https://goo.gl/LCLfBx`, 2017.

[HMM14] Desislava Hristova, Mirco Musolesi, and Cecilia Mascolo. Keep Your Friends Close and Your Facebook Friends Closer: A Multiplex Network Approach to the Analysis of Offline and Online Social Ties. In *Proceedings of the International AAAI Conference on Web and Social Media*, 2014.

[HMW10] Markus Huber, Martin Mulazzani, and Edgar Weippl. Who on earth is mr. cypher: automated friend injection attacks on social networking sites. In *Security and Privacy–Silver Linings in the Cloud*, pages 80–89. Springer, 2010.

[JGB+18] Kokil Jaidka, Sharath Chandra Guntuku, Anneke Buffone, H Andrew Schwartz, and Lyle Ungar. Facebook vs. Twitter: Differences in Self-disclosure and Trait Prediction. In *Proceedings of the International AAAI Conference on Web and Social Media*, 2018.

[Job17] JobBoy. `http://www.jobboy.com/`, 2017.

[KLLA12] Robin M Kowalski, Susan P Limber, Sue Limber, and Patricia W Agatston. *Cyberbullying: Bullying in the digital age*. Wiley & Sons, 2012.

[KPIM11] Georgios Kontaxis, Iasonas Polakis, Sotiris Ioannidis, and Evangelos P Markatos. Detecting social network profile cloning. In *PercomW*, 2011.

[KS13] Grace Chi En Kwan and Marko M Skoric. Facebook bullying: An extension of battles in school. *Computers in Human Behavior*, 29(1):16–25, 2013.

[KSS+15] Max Van Kleek, Daniel A. Smith, Nigel R. Shadbolt, Dave Murray-Rust, and Amy Guy. Self curation, social partitioning, escaping from prejudice and harassment: The many dimensions of lying online. In *Proceedings of the ACM WWW*, 2015.

[Lap18] Issie Lapowsky. Cambridge Analytica Execs Caught Discussing Extorsion and Fake News. Wired, `https://goo.gl/5dBtty`, 2018.

[LAS+16] Bin Liu, Mads Schaarup Andersen, Florian Schaub, Hazim Almuhimedi, Shikun (Aerin) Zhang, Norman Sadeh, Yuvraj Agarwal, and Alessandro Acquisti. Follow my recommendations: A personalized privacy assistant for mobile app permissions. In *Proceedings of SOUPS*, 2016.

[Lee17] David Lee. Facebook, Twitter and Google berated by senators on Russia. `https://goo.gl/288SmQ`, 2017.

[LGKM11] Yabing Liu, Krishna P Gummadi, Balachander Krishnamurthy, and Alan Mislove. Analyzing facebook privacy settings: user expectations vs. reality. In *Proceedings of the ACM IMC*, 2011.

[LLGL13] Ryan R Landoll, Annette M La Greca, and Betty S Lai. Aversive peer experiences on social networking sites: Development of the social networking-peer experiences questionnaire (sn-peq). *Journal of Research on Adolescence*, 23(4):695–705, 2013.

[MJB12] Michelle Madejski, Maritza Johnson, and Steven M Bellovin. A study of privacy settings errors in an online social network. In *Proceedings of PERCOM Workshops*, 2012.

[MMG+16] Mainack Mondal, Johnnatan Messias, Saptarshi Ghosh, Krishna P. Gummadi, and Aniket Kate. Forgetting in social media: Understanding and controlling longitudinal exposure of socially shared data. In *Proceedings of SOUPS*, 2016.

[MOT+17] Tara Matthews, Kathleen O'Leary, Anna Turner, Manya Sleeper, Jill Palzkill Woelfer, Martin Shelton, Cori Manthorne, Elizabeth F.

Churchill, and Sunny Consolvo. Stories from survivors: Privacy and security practices when coping with intimate partner abuse. In *Proceedings of CHI*, 2017.

[MT17] Tamir Mendel and Eran Toch. Susceptibility to social influence of privacy behaviors: Peer versus authoritative sources. In *Proceedings of CSCW*, 2017.

[MZ.18] Mark Zuckerberg on Facebook's 2018: We've changed, we promise. CNET, `https://tinyurl.com/yamv8zx8`, December 2018.

[NSL+17] Vishwajeet Narwal, Mohamed Hashim Salih, Jose Angel Lopez, Angel Ortega, John O'Donovan, Tobias Höllerer, and Saiph Savage. Automated assistants to identify and prompt action on visual news bias. In *ACM CHI*, 2017.

[NTT+16] Chikashi Nobata, Joel R. Tetreault, Achint Thomas, Yashar Mehdad, and Yi Chang. Abusive language detection in online user content. In *Proceedings of the ACM WWW*, 2016.

[NWM10] Amanda Nosko, Eileen Wood, and Seija Molema. All about me: Disclosure in online social networking profiles: The case of facebook. *Computers in Human Behavior*, 26(3), 2010.

[OJ18] Barbara Ortutay and Anick Jesdanun. How Facebook likes could profile voters for manipulation. ABC News, `https://goo.gl/eD6Ap3`, 2018.

[OMD09] Daniel M Oppenheimer, Tom Meyvis, and Nicolas Davidenko. Instructional Manipulation Checks: Detecting Satisficing to Increase Statistical Power. *Journal of Experimental Social Psychology*, 45(4):867–872, 2009.

[PCNR10] Rahul Potharaju, Bogdan Carbunar, and Cristina Nita-Rotaru. ifriendu: leveraging 3-cliques to enhance infiltration attacks in online social networks, 2010.

[Per17] What if technology could help improve conversations online? `https://www.perspectiveapi.com/`, 2017.

[QBC12] Daniele Quercia, Mansoureh Bodaghi, and Jon Crowcroft. Loosing "friends" on Facebook. In *ACM WebSci*, 2012.

[QH10] Daniele Quercia and Stephen Hailes. Sybil attacks against mobile users: friends and foes to the rescue. In *IEEE INFOCOM*, 2010.

[RBJB14] Hootan Rashtian, Yazan Boshmaf, Pooya Jaferian, and Konstantin Beznosov. To Befriend Or Not? A Model of Friend Request Acceptance on Facebook. In *Proceedings of the Symposium on Usable Privacy and Security*, pages 285–300, 2014.

[RKM19] Elissa M Redmiles, Sean Kross, and Michelle L Mazurek. How Well Do My Results Generalize? Comparing Security and Privacy Survey Results from MTurk, Web, and Telephone Samples. In *Proceedings of the 40th IEEE Symposium on Security and Privacy*, 2019.

[RLG16] Frederic Raber, Alexander De Luca, and Moritz Graus. Privacy wedges: Area-based audience selection for social network posts. In *Proceedings of SOUPS*, 2016.

[SCM11] Tao Stein, Erdong Chen, and Karan Mangla. Facebook Immune System. In *Proceedings of the 4th Workshop on Social Network Systems*, page 8. ACM, 2011.

[Sha16] Aarti Shahani. From Hate Speech To Fake News: The Content Crisis Facing Mark Zuckerberg. NPR,`https://goo.gl/kN1qQZ`, 2016.

[Sin14] Daisy Sindelar. The kremlin's troll army. *The Atlantic*, 12, 2014.

[SKV10] Gianluca Stringhini, Christopher Kruegel, and Giovanni Vigna. Detecting spammers on social networks. In *Proceedings of the 26th Annual Computer Security Applications Conference*, pages 1–9. ACM, 2010.

[Smi08] Mari Smith. Facebook Abuse: Is Blocking People Enough? `https://goo.gl/5T1Bzd`, 2008.

[Smi18] Kit Smith. 47 Incredible Facebook Statistics and Facts. `http://tinyurl.com/y3lvcjj8`, 2018.

[SRHF17] Vivek K. Singh, Marie L. Radford, Qianjia Huang, and Susan Furrer. "They basically like destroyed the school one day": On Newer App Features and Cyberbullying in Schools. In *ACM CSCW*, 2017.

[TC17] Sajedul Talukder and Bogdan Carbunar. When Friend Becomes Abuser: Evidence of Friend Abuse in Facebook. In *Proceedings of 9th ACM Conference on Web Science (WebSci)*, 2017.

[TC18] Sajedul Talukder and Bogdan Carbunar. AbuSniff: Automatic Detection and Defenses Against Abusive Facebook Friends. In *Proceedings of the International AAAI Conference on Web and Social Media*, pages 385–394, 2018.

[TMLS09] Nguyen Tran, Bonan Min, Jinyang Li, and Lakshminarayanan Subramanian. Sybil-resilient Online Content Voting. In *Proceedings of the USENIX Symposium on Networked Systems Design and Implementation*, pages 15–28, 2009.

[TN10] Kurt Thomas and David M. Nicol. The koobface botnet and the rise of social malware. In *5th International Conference on Malicious and Unwanted Software, MALWARE*, pages 63–70, 2010.

[Tom14a] Catalina L. Toma. Counting on Friends: Cues to Perceived Trustworthiness in Facebook Profiles. In *Proceedings of the International AAAI Conference on Weblogs and Social Media*, 2014.

[Tom14b] Catalina L Toma. Counting on friends: Cues to perceived trustworthiness in facebook profiles. In *ICWSM*, 2014.

[Tri17] Robert Triggs. Flagship? Mid-range? Budget? Find the Best Phone for You. [AndroidAuthority], `http://tinyurl.com/ybo3errt`, 2017.

[VFD+17] Onur Varol, Emilio Ferrara, Clayton A Davis, Filippo Menczer, and Alessandro Flammini. Online human-bot interactions: Detection, estimation, and characterization. In *Proceedings of ICWSM*, 2017.

[VK14] Jessica Vitak and Jinyoung Kim. You can't block people offline": examining how Facebook's affordances shape the disclosure process. In *Proceedings of CSCW*, 2014.

[Wak17] Jane Wakefield. Facebook and Twitter could face 'online abuse levy'. BBC Technology, `https://goo.gl/TmvNND`, 2017.

[WCZB+16] Caitlin Wolford-Cleveng, Heather Zapor, Hope Brasfield, Jeniimarie Febres, JoAnna Elmquist, Meagan Brem, Ryan C Shorey, and Gregory L

Stuart. An examination of the partner cyber abuse questionnaire in a college student sample. *Psychology of violence*, 6(1):156, 2016.

[Web] Merriam-Webster. `http://tinyurl.com/y87jgph7`.

[Wei10] Gabriel Weimann. Terror on facebook, twitter, and youtube. *The Brown Journal of World Affairs*, 16(2), 2010.

[Wek17] Weka. `https://www.cs.waikato.ac.nz/ml/weka/`, 2017.

[WIKP15] Pamela J. Wisniewski, A. K. M. Najmul Islam, Bart P. Knijnenburg, and Sameer Patil. Give social network users the privacy they want. In *Proceedings of CSCW*, 2015.

[WSHY15] Fangzhao Wu, Jinyun Shu, Yongfeng Huang, and Zhigang Yuan. Social spammer and spam message co-detection in microblogging with social context regularization. In *Proceedings of the ACM CIKM*, 2015.

[Yat17] Jeff Yates. From temptation to sextortion Inside the fake Facebook profile industry. Radio Canada, `https://goo.gl/KLH7DB`, 2017.

[YGKX10] Haifeng Yu, Phillip B. Gibbons, Michael Kaminsky, and Feng Xiao. SybilLimit: A Near-optimal Social Network Defense Against Sybil Attacks. *IEEE/ACM Trans. Netw.*, 18(3):885–898, June 2010.

[YKGF08] Haifeng Yu, Michael Kaminsky, Phillip B. Gibbons, and Abraham D. Flaxman. SybilGuard: Defending Against Sybil Attacks via Social Networks. *IEEE/ACM Trans. Netw.*, 16(3):576–589, June 2008.

[YS14] Chao Yang and Padmini Srinivasan. Translating surveys to surveillance on social media: methodological challenges & solutions. In *Proceedings of WebSci*, 2014.

[YWW+14] Zhi Yang, Christo Wilson, Xiao Wang, Tingting Gao, Ben Y Zhao, and Yafei Dai. Uncovering Social Network Sybils in the Wild. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 8(1):2, 2014.

[ZMMS18] Yixin Zou, Abraham H. Mhaidli, Austin McCall, and Florian Schaub. "I've Got Nothing to Lose": Consumers' Risk Perceptions and Protective Actions after the Equifax Data Breach. In *Fourteenth Symposium on Us-*

able Privacy and Security (SOUPS 2018), pages 197–216, Baltimore, MD, 2018. USENIX Association.

[Zuc06] Mark Zuckerberg. An Open Letter from Mark Zuckerberg, 2006.