# A Study of Friend Abuse Perception in Facebook

SAJEDUL TALUKDER, Edinboro University, USA
BOGDAN CARBUNAR, Florida Int'l University, USA

Social networks like Facebook provide functionality that can expose users to abuse perpetrated by their contacts. For instance, Facebook users can often access sensitive profile information and timeline posts of their friends, and also post abuse on the timeline and news feed of their friends. In this article we introduce AbuSniff, a system to identify Facebook friends perceived to be abusive or strangers, and protect the user by restricting the access to information for such friends. We develop a questionnaire to detect perceived strangers and friend abuse. We train supervised learning algorithms to predict questionnaire responses using features extracted from the mutual activities with Facebook friends. In our experiments, participants recruited from a crowdsourcing site agreed with 78% of the defense actions suggested by AbuSniff, without having to answer any questions about their friends. When compared to a control app, AbuSniff significantly increased the willingness of participants to take a defensive action against friends. AbuSniff also increased the participant self-reported willingness to reject friend invitations from strangers and abusers, their awareness of friend abuse implications and their perceived protection from friend abuse.

CCS Concepts: • **Security and privacy → Privacy protections**; *Social aspects of security and privacy*; *Spoofing attacks*;

Additional Key Words and Phrases: Social network friend abuse, friend spam, supervised detection

**ACM Reference Format:**
Sajedul Talukder and Bogdan Carbunar. 2020. A Study of Friend Abuse Perception in Facebook. *ACM Trans. Soc. Comput.* 1, 1, Article 1 (January 2020), 33 pages. https://doi.org/10.1145/3408040

## 1 INTRODUCTION

Influential social networks like Facebook encourage casual friendship relations. Social network users often have significantly more than 150 friends [1], which is the number of meaningful friend relations that a person can manage [26]). Past work has shown that adversaries, including bot-operated user accounts [71] [2], can establish friend relations with unsuspecting social network users, then expose them to vulnerabilities and abuse that include the collection and misuse of private information [24, 35, 59, 83, 84], identity theft [49] and spear phishing [28] attacks, the distribution of offensive, misleading, false or malicious information [2, 4, 19, 74], and cyber abuse that includes cyberstalking [25], doxing [24, 59], sextorsion [84] and cyberbullying [36, 37, 56]. High-profile cases of abuse perpetrated through Facebook include Cambridge Analytica's use of data collected from 87 million Facebook users [40] to identify "deep-seated underlying fears, concerns" [39] and to inject content to change user perception [50] and influence the outcome of elections [9, 10].

[1]The average number of friends per Facebook user is 338, while the median is 200 [58].
[2]Facebook estimated that 13% (i.e., 270 million) of their user accounts are either bots or clones [32].

Authors' addresses: Sajedul Talukder, Edinboro University, USA, stalukder@edinboro.edu; Bogdan Carbunar, Florida Int'l University, USA, carbunar@gmail.com.


2469-7818/2020/1-ART1 $15.00
https://doi.org/10.1145/3408040

In this article we focus on user perception of three types of abuse perpetrated through Facebook friend relations: (1) *active timeline abuse* consisting of abusive replies posted by friends to stories published by the user, (2) *active news feed abuse* consisting of offensive, misleading, false or potentially malicious information propagated by Facebook from the timeline of friends to the news feed of victims, and (3) *passive privacy abuse*, i.e., the surreptitious collection of sensitive account data by *stranger* friends, weak ties with whom the user has no interaction online and in real life.

We present a *friend abuse questionnaire* that investigates the user perception that a specific Facebook friend is a stranger or an active timeline or news feed abuser. We use this questionnaire to study the perception of abuse perpetrated through friend relations in Facebook, and also the willingness of users to accept defensive actions against friends that they perceive to be abusive.

Facebook users have been shown to rarely unfriend or block friends [78, 79]. To take steps toward addressing this problem, we introduce AbuSniff (Abuse from Social Network Friends), a system to help users better manage *interactional boundaries* [78]. AbuSniff leverages the friend abuse questionnaire to nudge users towards taking actions predicted to protect the user [7, 68, 73]: convert answers to the questionnaire to identify friends perceived by the user to be abusive, and suggest defenses to the user.

To address the cognitive load imposed by the task 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 *mutual activity features* that quantify the Facebook-recorded interactions between users and their friends. We use supervised learning algorithms trained on these features to predict user answers to the friend abuse questionnaire, thus identify user-perceived strangers and friend abuse.

Further, we use supervised learning to predict the user willingness to take suggested defensive actions. Specifically, we develop the *predictive* AbuSniff system that automatically identifies friends predicted to be perceived as abusive or strangers, and implements only the defense actions that it predicts that the users will approve.

**Findings**. In experiments where each of the 80 participants had to evaluate 20 of their randomly selected Facebook friends, 65 participants admitted to have at least one friend whom they perceived would abuse their status updates or pictures, and 60 of the participants had at least one friend whom they reported would post abusive material (i.e., offensive, misleading, false or malicious). 55 participants admitted to have at least one Facebook friend with whom they have never interacted in Facebook and in person. When asked directly, the participants unfollowed and restricted access for such friends 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 restrict bi-directional communications with such friends, in 92.45% of the cases.

When compared to a control app evaluated with 27 participants, AbuSniff increased participant willingness to unfriend and restrict the access of friends. In pre-test and post-test surveys with 62 participants, AbuSniff increased the self-reported willingness of participants to reject invitations from perceived strangers and abusers, their awareness of friend abuse implications and perceived protection from friend abuse.

Supervised learning algorithms trained on AbuSniff's mutual activity features and data that we collected from 1,452 friend relationships of 54 participants, were able to predict user answers to the questionnaire, with an F-measure ranging from 69.2% to 89.7%. In addition, AbuSniff predicted the cases where the users chose to ignore the suggested defensive action, with an F-Measure of 97.3%. In an experiment with 40 participants, involving 1,200 Facebook friends, we found that without having to answer the questionnaire, participants accepted 78% of the recommendations made by the predictive AbuSniff.

In summary, this article introduces the following contributions:

- Develop a friend abuse questionnaire to capture the user perception that a Facebook friend is a stranger or potentially abusive. Devise rules to convert answers into defense actions.
- Propose and evaluate the hypothesis that data recorded by Facebook can be used to predict the user perception of friend abuse. Develop features that can help predict user perceived friend abuse, and defensive actions that users are willing to take.
- Evaluate AbuSniff through experiments with 263 participants. AbuSniff can be downloaded from Google Play[3], and is open source[4].

## 2 MODEL AND OBJECTIVES

### 2.1 Social Network Abuse

The functionality provided by online services is known to influence abuse and generate negative socio-psychological effects [56]. Social networks, and Facebook in particular, implement affordances that provide the dimensions of persistence, searchability, replicability and invisible audience, known to magnify abuse [13]. In this section we describe the Facebook affordances that are the focus of our work, and explain the means through which they can be exploited to perpetrate abuse.

**Stranger Friends and Privacy Abuse**. Facebook users form *friend* relationships. Each user has a *friend list* of other users with whom she has formed friend relationships. To befriend someone, a user needs to first send a friend invitation, then be "confirmed". When two users become friends, they both automatically *follow* each other as well.

Friend relations can be exploited through fake friend invitations [15, 28, 62, 69] sent from sockpuppet accounts to intended victims. Patil [51] show that a significant percentage of social networking users are open to accepting friendship requests from strangers, especially when sent from accounts that have a profile photo of an attractive person of the opposite sex.

The reason for such invitations is that many users allow all their friends to access their profiles and timeline, and post to their timeline and news feed. Stutzman and Kramer-Duffield [64] show differences between the intended and expected audiences of such "friends-only" private profiles, and reveal that weak tie expectancy violations (i.e., weak tie friends reported to be an expected audience but not an intended audience) are associated with having a friends-only Facebook profile. Stutzman et al. [65] also show that the amount and scope of personal information that Facebook users revealed to their friends between 2005 and 2011, increased over time.

We define then *stranger friends* to be friend relations established with strangers, i.e., weak ties [16, 27] that an attacker can exploit. For instance, once a victim accepts a fake friend invite from an attacker-controlled sockpuppet account, the attacker can subsequently collect sensitive information (e.g., profiles, photos, friend lists, locations visited, opinions) posted on the victim's timeline, then perform attacks that include cyberstalking [25], doxing [24, 59], sextorsion [84], profile cloning [35], identity theft [49], and spear phishing [28] attacks.

**Timeline and its Abuse**. Facebook provides a *timeline* (a.k.a wall, or profile) for each user, the place where the user can share their updates, photos, check-ins, and other activities (e.g., posting comments on a status or picture of a friend, 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. While users can control with whom they share each story, i.e., through the *audience selector* option, it is well known that they often use the default settings [22, 44, 46]. Adversarial users can post abusive replies to

[3] AbuSniff app. https://goo.gl/LBWNWZ.
[4] AbuSniff source code. https://goo.gl/SZ7jrT.

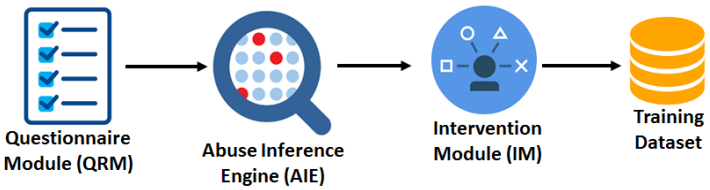



Fig. 1. Questionnaire based AbuSniff architecture. The QM module delivers the questionnaire. The AIE module uses the output of QM to identify abusive friends, and the IM module asks the user to take defensive actions. The output of these modules is stored for training, and is later used by the predictive AbuSniff (§ 4).

stories (e.g., status updates, photos) posted by their friends, on their own timeline. The abusive replies appear on the timeline of the victim friend, where the original stories were posted.

**News Feed and its Abuse**. After friending or following an account, the user will receive the updates posted on that account in their *news feed*. A user's news feed shows stories created by their friends, groups, and subscribed events. Stories are sorted based on various features, e.g., post time and type, and poster popularity. When a user follows someone other than a friend, the user receives that person's publicly shared status updates automatically, but not vice versa.

An adversary can exploit the relationship between his timeline and a friend's news feed to post abusive material on his timeline, which is then propagated to the news feed of his friends. Abusive information includes material perceived to be offensive, misleading, false, or malicious. For instance, Facebook revealed that Russia-based operatives created 80,000 posts that have reached 126 million US users [9, 41]. When studying the wall posts of 3.5 million Facebook users, Gao et al. [28] discovered more than 200K malicious posts with embedded URLs, with more than 70% pointing to phishing sites.

### 2.2 Research Objectives

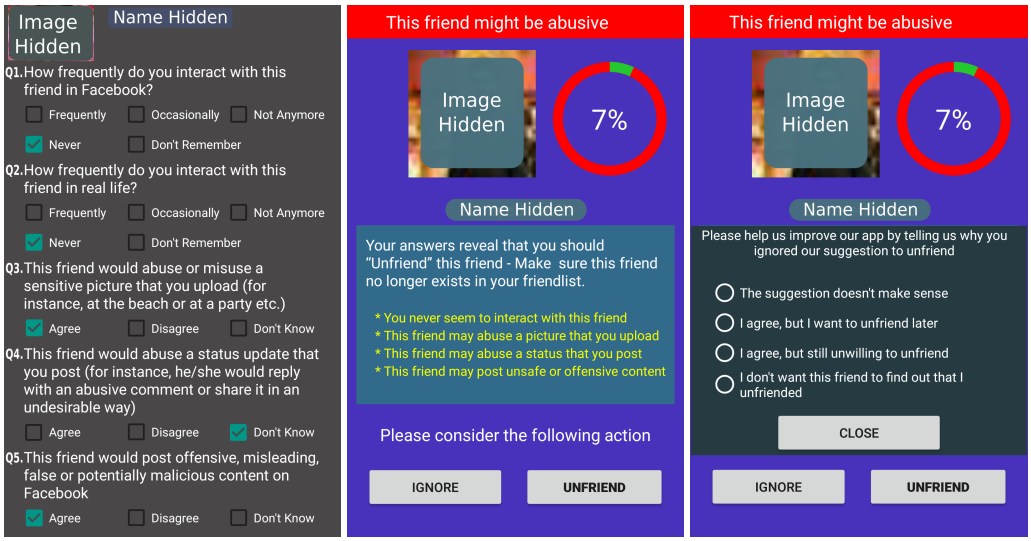
In this article we investigate the following research questions on social network friend abuse:

- **(RQ1)**: Facebook users are aware of abuse perpetrated by some of their friends.
- **(RQ2)**: Facebook users are willing to take defensive actions against friends that they perceive to be potentially abusive.
- **(RQ3)**: Tools can be designed to improve the willingness of users to take defensive actions on perceived abusive Facebook friends. The willingness of users to take defensive actions is impacted by the type of abuse perpetrated by the friend and the type of suggested defense.
- **(RQ4)**: Tools can be designed to predict the friends that users perceive to be abusive, and the defenses that users are willing to take for such friends.
- **(RQ5)**: Tools can be designed to improve user awareness of stranger and abusive friends, and their perception of safety from such friends.

Developing tools to help users detect and defend against abuse perpetrated through Facebook can help address concerns associated with the use of social networks. For instance, Wisniewski et al. [75] 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. We note that an end-goal of this research is to pave the way towards an automatic, user-transparent management of privacy settings on a per-friend basis.

## 3 QUESTIONNAIRE BASED ABUSNIFF

We have designed the AbuSniff system to help us investigate the research questions of § 2.2. AbuSniff is a mobile app that asks the user to login to their Facebook account. As illustrated in




(a) (b) (c)

Fig. 2. Anonymized screenshots of the Android AbuSniff app: (a) QM 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 UI of the Intervention Module (IM) 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. The circled percentage in the upper right corner is the job progress.

Figure 1, AbuSniff consists of modules to identify abusive friends and recommend defensive actions. The *questionnaire module* (QM) delivers to the user a set of questions about each of their evaluation friends. The *abuse inference engine* (AIE) converts answers to the questionnaire into actions. The *intervention module* (IM) displays the actions decided by the AIE and asks the user to confirm them. In the following, we detail each module.

### 3.1 Questionnaire Module (QM)

We have designed a questionnaire intended to investigate 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 only a randomly selected subset of the user's friends. We designed the questions to help identify types of abuse described in § 2.1. To ensure a simple navigation of the questionnaire, all the questions fit on a single screen for a variety of popular smartphones.

Figure 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*) investigate the user's frequency of interaction with the friend, in 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.

The next two questions investigate 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 update that you upload*). The possible responses are "Agree", "Disagree" and "Don't Know". Question (Q5)


(a) (b) (c)

Fig. 3. (a) The "unfriend or sandbox" UI for privacy abuse: sandboxing isolates but does not unfriend or notify the friend. (b) The UI of the autonomous AbuSniff asking user confirmation to restrict the access of a friend predicted to be a timeline abuser. The circled percentage in the upper right corner is the job progress, e.g., 55% means that the user has processed 11 out of 20 friends. (c) AbuSniff app screen shown to unfollow a friend detected to be a news feed abuser according to the questionnaire responses.

(*This friend would post offensive, misleading, false or potentially malicious content on Facebook*), investigates perceived news feed abusers.

### 3.2 Abuse Inference Engine (AIE)

To nudge users toward implementing safer social interactions, we leverage 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 their 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 their friend no longer receive stories published by the other. Unlike unfriending, sandboxing will not remove the user and their friend from each other's friend lists.

The sandboxing vs. unfriending options are meant to address the difference between passive and active abuse described in § 2.1, i.e., to sandbox stranger friends who are abusive but unfriend those who are also actively abusive.

The abuse inference engine (AIE) builds on these protective options. It takes as input the responses collected by the QM or predicted by the APM module (§ 4.1), and outputs suggested actions from the set {"unfriend", "unfollow", "restrict access", "sandbox", "ignore"}. That is, 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) unfriends or sandboxes friends who are perceived as strangers or who qualify for both points (1) and (2).

| | 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 1. Set of rules to convert questionnaire responses to defensive actions. Similar to firewall filters, the first matching rule applies. !$A$ denotes any response different from $A$. NOP = no operation.

We introduce first an intuitive approach based on the set of rules of Table 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 two 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 unfriending friends who were assigned "Never" in any of the two first questions, and "Agree" in any of the last three questions), see 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).

In § 7 we evaluate and adjust the AIE rules. In § 4 we develop a supervised learning based approach to predict the defensive actions that users would in fact agree to implement.

### 3.3 Intervention Module (IM)

To help us answer the key research questions RQ2 and RQ3 we have designed the Intervention Module (IM) as 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 1.

Figure 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 2(a).

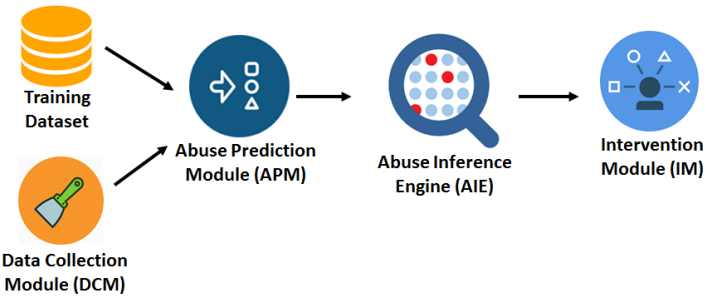



Fig. 4. Predictive AbuSniff system architecture. The DCM collects Facebook data concerning the relationship between the user and each friend. The APM uses this data, and training data collected by the questionnaire based AbuSniff (§ 3), to predict the user responses to the questionnaire. The AIE is inherited from the questionnaire based AbuSniff, but uses the output of APM instead of QM to identify abusive friends. The optional IM asks the user to confirm the predicted action for detected abusive friends.

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 specify the reason, see Figure 2(c). In the first experiment of § 7 ($n$ = 20) we have included several suggestions, and also an input text field for the participants to type their reason. The final options are (1) "the suggestion does not make sense", (2) "I agree, but I want to unfriend later", (3) "I agree but I am still unwilling to unfriend", and (4) "I don't want this friend to find out that I unfriended", see Figure 2(c).

For detected stranger friends, the IM module educates users about the meaning and dangers of having such a friend, see Figure 3(a). It also offers the option to "sandbox" such friends. Following the rules of Table 1, IM also suggests unfollowing or restricting friends who are abusive in only one direction of their communications. Figure 3(b) shows a snapshot of the restrict screen, its meaning and reasons for selection. Figure 3(c) shows a snapshot of the screen shown to nudge users to unfollow news feed abusers.

## 4 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 [58]. Further, we lack confidence that defenses inferred by the AIE module would be accepted by the user. To address these problems, 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.

To this end, we introduce the predictive AbuSniff system illustrated in Figure 4. The predictive AbuSniff replaces the QM module with an Abuse Prediction Module (APM). APM uses training data collected through the questionnaire based AbuSniff (see Figure 1 and § 9) and data collected by the *data collection module* (DCM), to predict the outcome of the QM module. In the following we detail the APM and DCM modules.

### 4.1 Abuse Prediction Module (APM)

We introduce several *mutual activity* features based on the Facebook data shared by a user $U$ and a friend $F$. We use these features 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) **same current city**: boolean value that is true when $U$ and $F$ live in the same city, (5) **same hometown**: boolean value that is true when $U$ and $F$ are from the same

place, (6) **common education count**: the total number of educational institutions that both $U$ and $F$ have attended, and (7) **common workplace count**: the total number of places where $U$ and $F$ were both employed.

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 QM 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 of the UM module for $F$, (3) the AIE-suggested action for $F$, i.e., ignore (safe friend), unfriend, unfollow, restrict, sandbox, and (4) the action taken by $U$ for $F$.

Previous work has shown that information shared by social network users can predict their tie strength [8, 29, 30]. We conjecture that shared Facebook activities with a friend may further predict user perception of the friend's ability to perform abuse, whether the user has interacted with that friend in Facebook and in real life, and the willingness of the user to take a defensive action for such a friend. In § 9 we evaluate this conjecture, i.e., the ability of the above features to predict questionnaire answers and user decisions on suggested actions.

### 4.2 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 QM, 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 § 6.1 we discuss ethical considerations of the data collection process.

Data collection 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 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 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 through the webview.loadUrl ("javascript:*code*") method: "*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, accounting for a total of around 3s per friend. AbuSniff performs this process in the background, e.g., while the user is answering the QM questionnaire.

## 5 QUALITATIVE INVESTIGATION

In this section we discuss the design and findings from a qualitative investigation about user perception of abuse for Facebook friends.

**Methods**. We conducted a qualitative, in-person investigation with 13 participants consisting of two K-8 teachers, one psychologist, eight students, one dentist and one homemaker (eight female

and five male). We posted flyers to recruit the participants from our city, and also invited people from our community to participate in the study.

We asked each participant to install and run the questionnaire based AbuSniff. If the participant did not have an Android device, we have provided a lab device for this purpose. After answering the questionnaire for 20 randomly selected Facebook friends, we asked the participant to allow us to revisit the answers together. We then asked participants to explain their reasons for answering questions in a manner indicative of abuse. We have asked permission from each participant, to record their responses using a voice recorder.

Participation took an average of 28 minutes. We paid each participant $10.

**Findings**. After answering "Never" for Q1 (*How frequently do you interact with this friend in Facebook*) for a friend, 3 participants 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]."* (P4, 22 year old female undergraduate student), and

*"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."* (P12, 31 year old female dentist)

For question Q2 (*How frequently do you interact with this friend in real life*), 4 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"* (P11, 34 year old male psychologist), and

*"She is from my same [sic] college"* (P7, 24 year old female undergraduate student).

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.

After answering "Agree" for Q3 (*this friend would abuse or misuse a sensitive picture that you upload*) , 4 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."* (P2, 27 year old female graduate student), and

*"This friend has posted a bad comment in one of my photos. That was my wedding photo. I felt so offended."* (P13, 46 year old female homemaker)

We note that P2 identified the friend who cloned her account, as being also a "stranger" (answered "Never" to both Q1 and Q2). This suggests that users may perceive some strangers to be not only passive, but also active abusers.

Three participants who answered "Agree" for a friend on question Q4 (*This friend would abuse a status updated that you upload*), 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."* (P6, 26 year old male graduate student), and

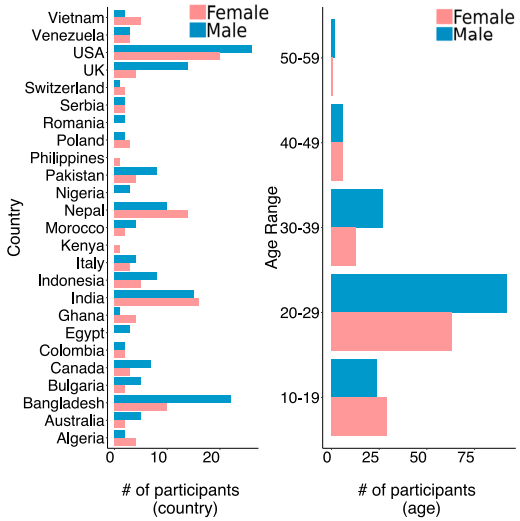



Fig. 5. Demographics over the 263 participants in all experiments. (country) Distribution of the 25 countries of residence by gender. US, Bangladesh and India are the top three countries. (age) Distribution of age range by gender. A majority of the 151 male and 112 female participants are 20-29 years old; 15 are 40-59 years old.

*"Once I posted a sad status update because I was feeling frustrated. But this friend then posted a trolling comment on my post."* (P5, 19 year old male undergraduate student)

Stories shared by the 5 participants who answered "Agree" on question Q5 (*This friend would post offensive, misleading, false or potentially malicious content on Facebook*) include:

*"This friend bothered friends by bad posts [..] The posts were against my own ideas [sic]."* (P1, 30 year old female graduate student), and

*"I have often seen this friend sharing fake news. Sometimes she posts so much bogus stuff that my news feed gets flooded."* (P3, 32 year old male graduate student)

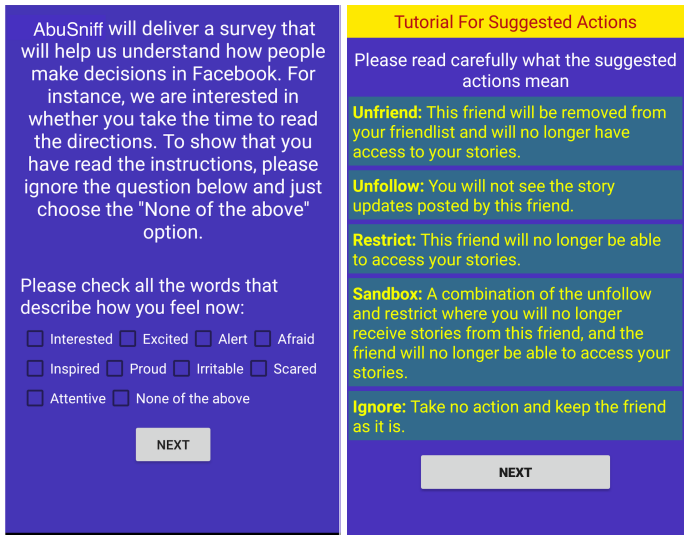
We conclude that Facebook users can experience and are aware of abuse perpetrated by their social network friends, providing evidence toward research question RQ1 (§ 2.2). We further evaluate RQ1 through a quantitative experiment in § 7. In the following we discuss several experiments that we conducted to evaluate AbuSniff. We first describe the common methods of these experiments, then detail the methods and finding of individual experiments.

## 6 METHODS: ONLINE EXPERIMENTS

We have conducted several experiments to investigate the research questions introduced in § 2.2. In the following, we first describe the participant selection procedure, including techniques we used to ensure data quality. In subsequent sections we provide specific methods used for each experiment, along with results.

We have recruited 325 participants (August 2016 to October 2017), by posting jobs on JobBoy[5] asking participants to install AbuSniff from the Google Play store, use it to login to their Facebook

[5]http://www.jobboy.com/.




(a) (b)

Fig. 6. (a) Attention check screen. (b) Definitions of the suggested actions, in both AbuSniff and control apps.

accounts and follow the instructions on the screen. The job specified that participants need to have at least 30 Facebook friends, have access to an Android device, and be at least 18 years old.

We have paid each crowdsourced participant $3, for a median job completion time of 928s (SD = 420s). This was done through a code shown to the participant on the last screen of the AbuSniff app. In addition to discarding data from participants with fewer than 30 Facebook friends, we have also used the following mechanisms to eliminate low quality data.

- **Attention-check screen**. The AbuSniff app presents a tutorial of the actions that are later suggested, see Figure 6(b). To ensure that the participants pay attention and are able to understand and follow simple instructions in English, AbuSniff includes a standard attention-check screen at the beginning of the app, see Figure 6(a) for a snapshot: "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 experiments 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": three fake identities (two female, one male) that we included in the AbuSniff questionnaire at random positions. 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".
- **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.

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

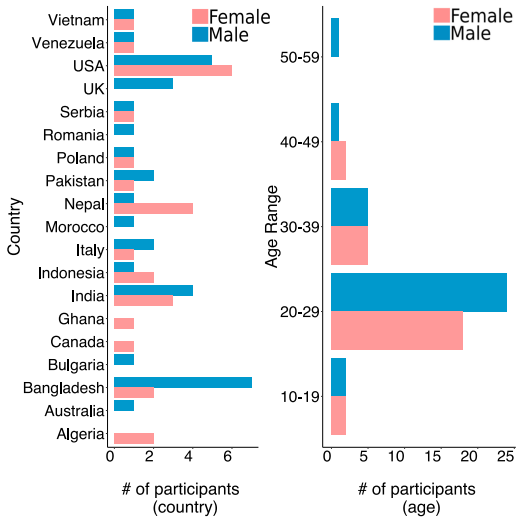



Fig. 7. Demographics over participants in the second experiment ($n$ = 60). (country) Distribution of the 19 countries of residence by gender. US, Bangladesh and India are the top three countries of residence. (age) Distribution of age range by gender. A majority of the 33 male and 27 female participants are 20-29 years old.

### 6.1 Ethical Considerations

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.

## 7 EFFECTS OF SANDBOXING

### 7.1 Methods

We performed two online experiments ($n$ = 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 and RQ3). Figure 7 shows the distribution of the country of origin and age, by gender, over the participants in the second experiment ($n$ = 60).

In the first experiment we used the questionnaire based AbuSniff of § 3. 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 collected the questionnaire answers for the 20 friends, the decisions taken for the abusive friends, and the reasons provided for ignoring the suggestions.

In the second experiment we have evaluated a version of AbuSniff that relaxed rule 1 in Table 1, to give the user the option to either *sandbox* or unfriend a non-abusive stranger. A sandboxed friend

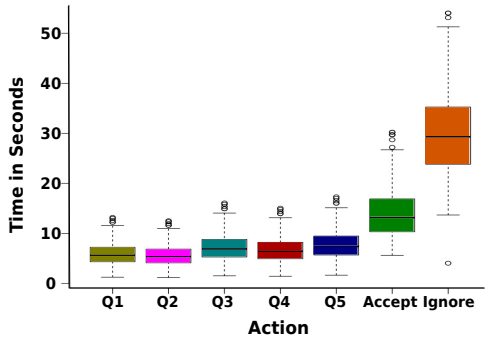



Fig. 8. Distribution of times taken by participants in the first two experiments to answer each question in the questionnaire, and decide whether to accept or ignore an action suggested against an abusive friend. The median time to answer any of the five questions exceeds 4.11s, with a maximum time of 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.

can no longer harm the user, as all Facebook communication lines are interrupted. Sandboxing achieves this without severing 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 1). We have modified the UI of AbuSniff to justify this choice, through a description of the harm that strangers can perform and the defenses that the user can take for such friends. Figure 3(a) shows a snapshot of the modified UI screen that offers the sandbox alternative to unfriending strangers.

### 7.2 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. Figure 8 shows the distribution of the response times over the participants whose average response time exceeded 3s. For these participants, the total time taken to answer the five 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), ($t(597)=6.205$, $p<.001$). These numbers suggest that participants have carefully considered AbuSniff's questions and suggestions, and have not randomly browsed through the app.

**Abuse Perception**. Figure 9 shows the distribution of the responses for each of the five AbuSniff questions, from the 1,600 friend relationships (20 from each of 80 participants) queried in the two experiments. The top bar shows that in 12% of the 1,600 friend relations 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,

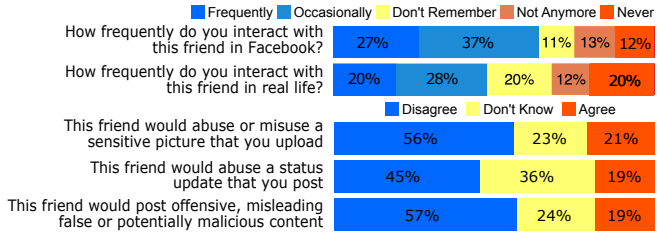



Fig. 9. Distribution of responses for the friend abuse questionnaire over 1,600 Facebook friend relationships. From top to bottom: Q1: frequency of Facebook interaction, Q2: frequency of real world interaction, Q3: friend would abuse posted sensitive picture, Q4: friend would abuse status update post, and Q5: friend would post offensive, misleading, false or potentially malicious content. The orange sections correspond to potential strangers or abusive friends.

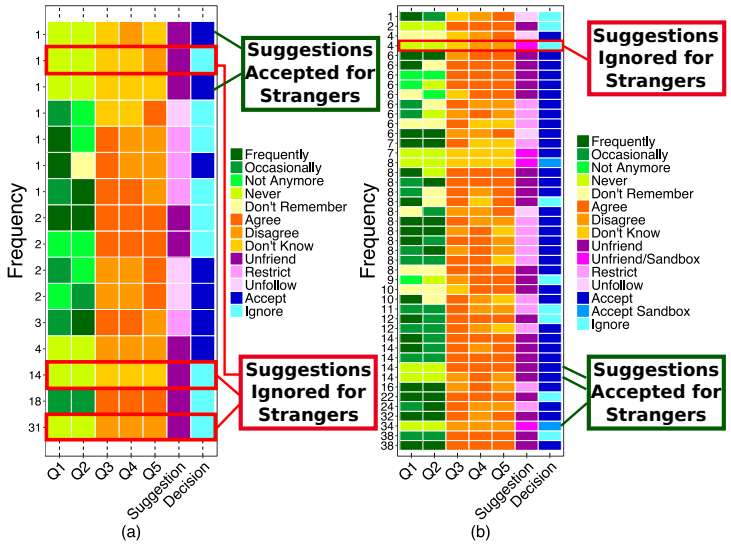



Fig. 10. Color-coded plot of frequent questionnaire responses and corresponding user decisions, for (a) the first experiment ($n$ = 20) and (b) the second experiment ($n$ = 60). A row on the $y$ axis shows the frequency of friend relationships that have the questionnaire response, recommendation and user decision pattern shown on the $x$ axis. The questions are, 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. We observe participant preference to sandbox vs. unfriend stranger friends, which results in a reduction in the number of suggestions ignored for strangers.

misleading, false, or potentially malicious content (bottom bar). 68 of the participants had at least one friend whom they perceived would abuse their photos, 62 of the participants have at least one friend who would abuse their status updates, and 62 have at least one friend who would post abusive content.

**Gender and age impact**. In terms of having at least one friend perceived as abusive, Chi-square tests revealed no significant difference between genders on any of the five questions. Similarly, Chi-square tests revealed no significant differences between the age groups of under 30 year old and above 30 year old participants (61 vs 19 participants), on questions 1, 2 and 4. However, participants under 30 are significantly more likely ($\chi^2$ = 4.417, df = 1, p = 0.03) to have at least one friend whom

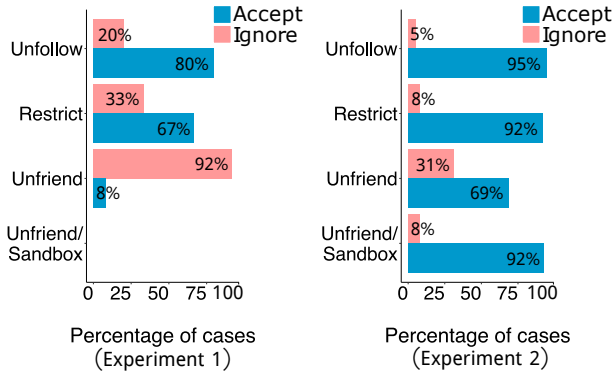



Fig. 11. Comparison of recommendation vs. acceptance in the exploratory experiments ($n$ = 20 and $n$ = 60). The values on the x axis are percentages. In the first experiment, 8% of the recommended "unfriend" actions were accepted. The undefined "unfriend or sandbox" option is shown for alignment. **The "sandbox" option and user education were effective**: in the second experiment, 92% of the suggested "unfriend or sandbox" suggestions were approved by participants.

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 one friend who would post offensive, misleading, false or potentially malicious content (50 out of 61 vs 10 out of 19, $\chi^2$ = 6.64, df = 1, p = 0.01).

**Willingness to Defend Against Abuse**. In the first experiment, out of 400 investigated friend relations AbuSniff identified 85 abusive or non-abusive stranger friends. Of these, AbuSniff recommended 74 to unfriend, 6 to restrict and 5 to unfollow. Figure 10(a) shows a detailed view 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). A total of 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 experiment.

The results of this experiment are further summarized in Figure 11(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**. The second experiment ($n$ = 60) evaluated the AbuSniff version updated to also offer sandboxing options for stranger friends. In this experiment, participants declared a total of 513 of the evaluated friends to be either abusive or non-abusive strangers. Figure 10(b) provides a detailed view of the user responses, recommended actions and user decisions. It shows that 53 friend relations were considered non-abusive strangers, and participants were unwilling to unfriend or sandbox only four of the corresponding stranger friends. In the first experiment participants took a defensive action for a non-abusive stranger friend in 11.5% of the cases (6 out of 52), whereas in the second experiment participants took an action in 92.4% of the cases (49 out of 53), (p<0.00001 with Fisher's exact test). This contrast suggests participant preference to sandbox stranger friends.

In the second experiment, AbuSniff recommended 303 friend relations to unfriend, 53 to unfriend or sandbox, 138 to restrict and 19 to unfollow. Figure 11 (experiment 2) shows for each of these types of recommendations, the percentage that was accepted and the percentage that was ignored.

| Non-abusive strangers<br>Reasons to ignore recommended action | S1 | S2 |
|---|---|---|
| Recommendation does not make sense | 19.5% | 25% |
| Not ready to take action at that time | 26% | 50% |
| Agree but still want to keep stranger friend | 19.5% | 25% |
| Afraid that action will be observable | 35% | 0% |
| **Total Cases** | 46 | 4 |

| Abusive non-strangers<br>Reasons to ignore recommended action | S1 | S2 |
|---|---|---|
| Recommendation does not make sense | 18% | 6% |
| Not ready to take action at that time | 27% | 51% |
| Agree but still want to keep abusive friend | 14% | 9% |
| Afraid that action will be observable | 41% | 34% |
| **Total Cases** | 22 | 95 |

Table 2. Comparison of reasons to ignore the AbuSniff suggested action in the first experiment ($n$=20) and second experiment ($n$ = 60). (Left) non-abusive strangers, where the suggestion was "unfriend" in the first experiment but "sandbox or unfriend" in the second experiment. (Right) abusive non-strangers, where the suggestion was "unfriend" in both experiments.

Consistent with the first experiment, a large percentage of unfollow (18 out of 19) and restrict suggestions (127 out of 138) were accepted.

However, in the second experiment 49 out of 53 "unfriend or sandbox" suggestions were accepted, and 208 of 303 unfriend recommendations were accepted. This is an improvement over the first experiment where only 6 unfriend recommendations were accepted 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 Recommendations**. Table 2(left) summarizes and compares the reasons given by the participants in the two experiments, to ignore the suggested recommendations for non-abusive stranger friends. We observe that when we added the "sandbox" option for non-abusive strangers, only one participant believed that AbuSniff's recommendation does not make sense, compared to 9 participants in the first experiment. Further, only three participants in the second experiment were either not ready for the action or wanted to keep the stranger friend, compared to 21 in the first experiment. Notably, none of the participants in the second experiment were afraid that the action will be observable by the friend, a steep decrease from 16 participants in the first experiment.

Table 2 (right) further compares the reasons chosen by participants in the two experiments, 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 experiment vs. 95 were found in the second one. In only 4 and 6 cases in the two experiments respectively, participants believed that the recommendation does not make sense. In 9 cases in the first vs. 57 in the second experiment, participants agree with the suggested defense but are either not ready to take it, or would prefer to keep the friend. Further, in almost one third of the cases in both experiments, 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 experiment (16 out of 46 cases), see Table 2(left) and suggest that these participants may prefer to sandbox even abusive non-stranger friends.

## 8 CONTROL EXPERIMENT

**Methods**. In order to understand if AbuSniff has an effect on the willingness of users to take defensive actions on Facebook friends (RQ3, § 2.2), we have designed a control app. Similar to AbuSniff, the control app first explains each user action ("unfriend", "unfollow", "sandbox", "restrict" and "ignore", see Figure 6(b). 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 12 for a snapshot). The control app emulates the color scheme of Facebook, which includes the colors of the buttons and of the background.

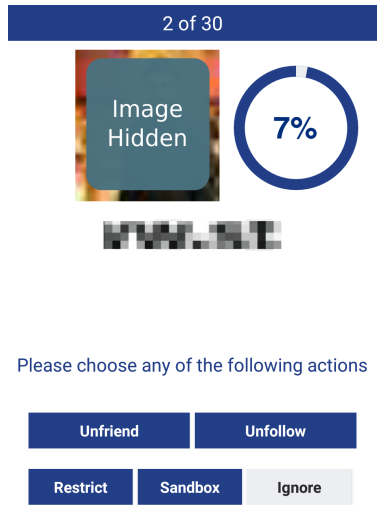



Fig. 12. Control app screen shown for a randomly selected friend, asking the user to choose an action. Control app highlights the defense actions.

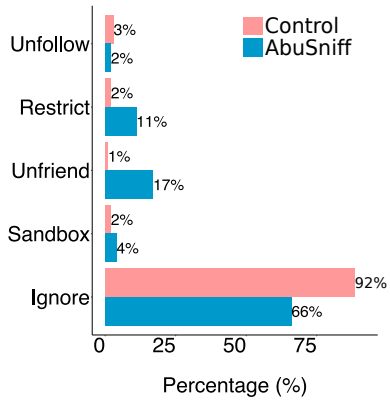



Fig. 13. Questionnaire based AbuSniff vs. control app. AbuSniff participants were more open to unfriend (17% vs. 1%) and restrict (11% vs. 2%) friends, than control participants.

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 have conducted an online, control condition experiment with this app, with 27 crowdsourced participants.

**Results**. Figure 13 compares the performance of the questionnaire based AbuSniff in the second experiment from § 7 against the control app, 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 condition, participants did not take a restrictive action in 92% of cases. In contrast, in AbuSniff, friend relationships were ignored in only 66% of the cases, where we included in the count also the friends that were perceived to be safe. We observed differences for the unfriend option that was chosen in 1% of cases during the control experiment, but in 17% of the

cases during the AbuSniff experiment. We also observed differences for the "restrict" option that was chosen in 2% of cases during the control vs. 11% of the cases during the AbuSniff experiment.
**Summary of Findings**. We observed significant user perception of potentially abusive Facebook friend connections, confirming research question RQ1 (§ 2.2). 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. Further, we observed a mixed response for RQ2, thus a positive answer for RQ3. Specifically, participants tended to accept suggestions to unfollow news feed abusers, restrict timeline abusers, and were more willing to sandbox than unfriend perceived non-abusive strangers.

## 9 EVALUATION OF ABUSNIFF PREDICTIONS

As described in § 4, the predictive AbuSniff system replaces the questionnaire delivery module (QM) with the abuse prediction module (APM): For each friend predicted to be perceived as abusive, and for whom the user is predicted to take the suggested defense action, the predictive AbuSniff asks the user to either accept or ignore the action.

### 9.1 Methods

To investigate question RQ4 (§ 2.2), we performed two experiments. In the first experiment we collect data required to train the predictive AbuSniff and use cross-validation to evaluate the offline accuracy of its predictions. In the second experiment we evaluate the trained AbuSniff on live participants. In the following we first describe the methods we used for this evaluation, then present our results.
**Cross-validation of Predictive AbuSniff**. We first performed an experiment with 54 crowdsourced participants, which we asked to install and run the questionnaire based AbuSniff app. We have collected mutual activity Facebook data from 1,452 friend relationships of the 54 participants, associated to their AbuSniff questionnaire answers and defense decisions.

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 five 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 one to test. We used the features of each tuple in the 9 folds to separately train supervised learning algorithms for each of the five 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 five questions and the user decision. We report averages of the prediction accuracy over the 10 experiments.

As shown in Figure 9, the distribution of the answers to the five 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[6] to evaluate several supervised learning algorithms, including Random Forest (RF), Decision Tree (DT), SVM, PART, MultiClassClassifier, SimpleLogistic, K-Nearest Neighbors (KNN) and Naive Bayes, but report only the best performing algorithm.

[6]https://www.cs.waikato.ac.nz/ml/weka/.

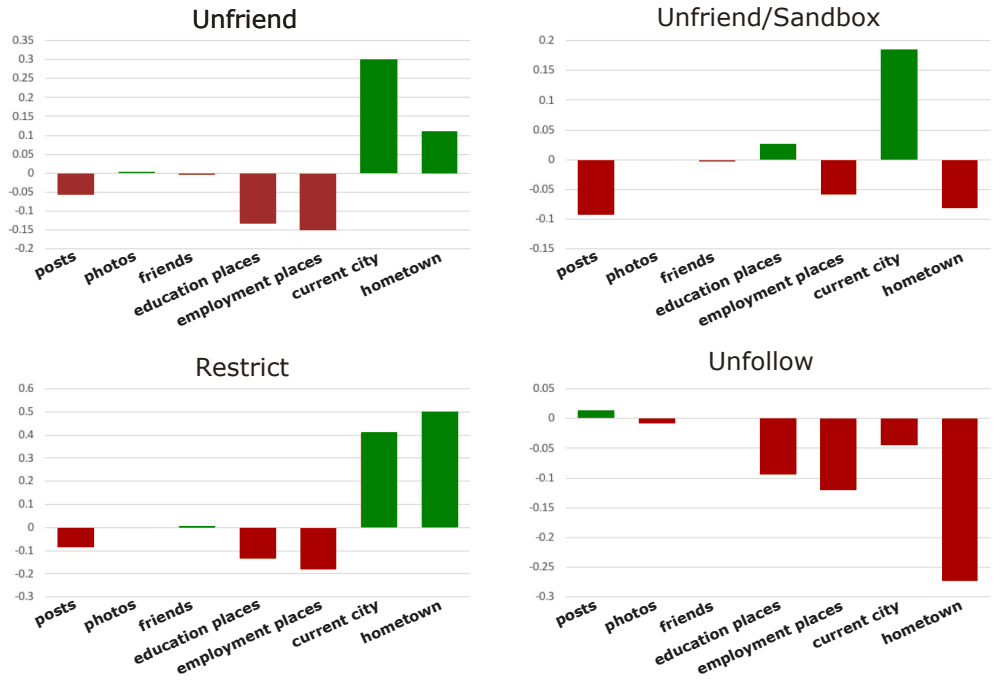



Fig. 14. Multinomial logistic regression (MLR) correlations between mutual activity features and AIE decision for each abuse category. Coefficients for the mutual activity features are plotted as Sign($C_f$)*Log(1+Abs($C_f$)), where $C_f$ denotes the actual co-efficient. For the same current city and 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 common education count, the common workplace count, and the number of common posts, have the highest impact on all of the AIE decisions.

**Evaluation of Predictive AbuSniff in the Wild**. We performed an online experiment with 40 crowdsourced participants (49 recruited, 9 discarded for failing the data quality verification tests of § 6). We asked these participants to install and run the predictive AbuSniff app. The predictive AbuSniff does not ask the participant to answer the questionnaire, but asks the participants to make defense decisions only for the friends predicted to be perceived to be abusive and for whom AbuSniff predicts that the participant will agree with the suggestion.

### 9.2 Results

**Feature Correlation Investigation**. We first performed a multinomial logistic regression (MLR) analysis using SPSS to find out whether the mutual activity features of the APM module (§ 4.1) are good predictors for the defense actions recommended by AIE.

We used the 7 mutual activity features (five continuous, two categorical) as the independent variables, and the AIE decision with five categories (Unfriend, Sandbox/Unfriend, Restrict, Unfollow, Safe) as the dependent variable.

Model fit statistics indicate a good fit, i.e., $\chi^2$ (28) = 385.037, p < 0.05 which confirms our model predicts significantly better, or more accurately, than the null model. Figure 14 plots the value of each coefficient in the model with its respective sign, and shows whether the features are positively or negatively correlated with the AIE decisions. For convenience, 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

| 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 3. Precision, recall and F-measure of APM for questions Q1-Q5 (RF: Random Forest, DT: Decision Tree). Q1: *How frequently do you interact with this friend in Facebook*, Q2: *How frequently do you interact with this friend in real life*, Q3: *This friend would abuse or misuse a sensitive picture that you upload* and Q4: *This friend would abuse a status update that you upload*, Q5: *This friend would post offensive, misleading, false or potentially malicious content on Facebook.*

value. A larger (absolute) coefficient means that the corresponding feature has more impact on the prediction. The plot shows that according to the MLR analysis, the same current city, the same hometown, the common workplace count, the common education count, and the number of common posts have the highest impact on all of the AIE decisions.

**Predicting Questionnaire Answers**. Table 3 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 one class for each of the five possible responses. Table 3 (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% (Kappa statistic = 0.88). RF also achieves the best performance for question Q2, with an overall F-measure of 90% (see second section of Table 3). APM with Random Forest is able to predict the "Never" response for a friend, with precision 86.5% and recall 93.2% (Kappa statistic = 0.86).

For Q3, APM achieved the best performance when using the Decision Tree (DT) classifier (see third section of Table 3), with an average F-Measure of 79.3% (Kappa statistic = 0.68). The DT classifier also achieved the best results for Q4 (the fourth section of Table 3), with an average F-Measure of 80.4% (Kappa statistic = 0.67). For Q5, APM achieved best performance with RF, see

| 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. 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.**

Table 3 (fifth section), with an F-Measure for the news feed abuse indicator ("Agree" response) of 78% (Kappa statistic = 0.73).

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 suggests that 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 five possible classes: "unfriend", "restrict", "unfollow", "sandbox", and "ignore". APM achieved the best performance with the RF classifier. Table 4 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: AbuSniff uses APM's predictions to decide which friends to recommend for the user to defend against.

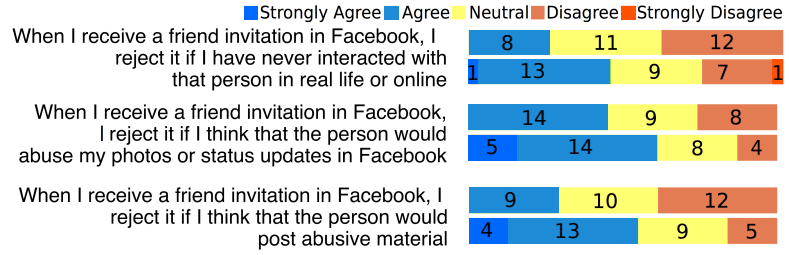
**Feature Rank**. The most informative features in terms of information gain were consistently among the mutual post count, mutual friend count and mutual photo count; the same hometown and common education count were the least informative features. We found correlations between the common photo count 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.

**Predictive AbuSniff in the Wild**. In the second experiment with 40 participants (§ 9.1) and 1,200 friend relationships investigated, (30 friends per participant), the APM module of the predictive AbuSniff automatically labeled 403 friends 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.

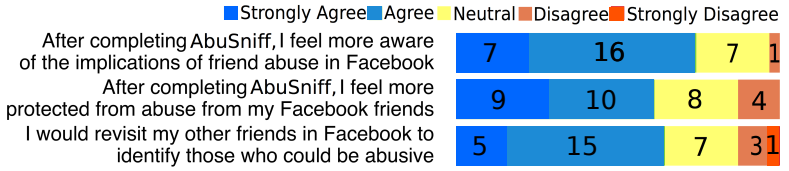
**Summary of Findings**. We observe the ability of AbuSniff to predict friend abuse and the user willingness to adopt defenses. Further, when evaluated with real users, the trained, predictive AbuSniff had performance similar to an offline cross-validation experiment. This provides encouraging evidence suggesting a positive answer to research question RQ4 (§ 2.2).

## 10 PRE-TEST AND POST-TEST EXPERIMENTS

To evaluate the impact of AbuSniff we have designed pre-test and post-test surveys. We describe first our methods then present our findings.




(a)




(b)

Fig. 15. (a) Results 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, 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.

### 10.1 Methods

We first designed a pre-test survey that consists of three 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)."*

Further, we have designed a post-test survey that, in addition to the above three items, includes the following three 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"*.

We conducted a pre-test experiment with 31 participants, where we asked them to answer only the pre-test survey. We then conducted a post-test experiment with a different set of 31 participants, where we asked to first run the questionnaire based AbuSniff, then answer the post-test survey. We did not collect any training data during the pre-test and post-test experiments.

### 10.2 Results

Figure 15(a) compares the user responses in the pre-test (top) and post-test (bottom) surveys, for each of the first three 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), more participants either strongly agree or agree on all three 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 one participant strongly disagrees. 19 participants strongly agree or agree with (I2), and only 4 disagree. Finally, 17 participants strongly agree or agree with (I3), and 5 disagree.

Figure 15(b) shows the participant responses to the three 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 one disagrees. 19 participants strongly agree or agree that after running AbuSniff they feel more protected from friend abuse; four participants disagree. 20 participants strongly agree or agree that they would revisit their other friends after running AbuSniff. Only three disagree, and one strongly disagrees. This experiment suggests a positive answer to research question RQ5 (§ 2.2).

## 11 DISCUSSION

We have explored the perception of friend abuse in Facebook, and the willingness of users to take defensive actions against friends that they perceive to be passively or actively abusive. We have focused on abuse perpetrated through Facebook friend relationships, the timeline and the news feed affordances. We have investigated the perception of abuse perpetrated by individual, specific friends, and not perception of general exposure to abuse. We developed an automated, predictive tool to detect and defend against perceived abuse, and provide a transparent, first line of defense against abuse, for Facebook users who are unlikely to know and trust all their friends.

AbuSniff reduces the *attack surface* of its users, by unfriending or restricting communications with 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., [53], and can be used in conjunction with tools that monitor social networking events [21, 23].

**Ability of Questionnaire to Identify Perceived Abuse**. Questions Q3, Q4 and Q5 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 weak ties, i.e., random people befriended online. The first experiment in § 7 reveals that indeed, few participants agreed to unfriend such friends, thus they are likely to seldom perceive such friends as being abusive. However, the second experiment § 7) shows that participants had a much higher likelihood to "sandbox", i.e., isolate such 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 our pre-test and post-test surveys, 23 participants strongly agreed or agreed that after running the questionnaire based AbuSniff, they felt more aware of the implications of friend abuse; only one disagreed, none strongly disagreed.

**Benefits of Weak Ties**. Weak ties in social networks can be beneficial. For instance, Burke and Kraut [16] report that bridging social capital after losing a job, comes from both strong and weak ties. This was also reported to be the case by Ellison et al. [27], in Facebook. However, perhaps unsurprisingly, Burke and Kraut [16] also report that communication with strong ties is a better predictor of finding employment within three months. AbuSniff recommends severing ties only with friends perceived to be strangers, or abusive through timeline and news feed communications. AbuSniff does not recommend unfriending friends with whom the user has communicated at any time, unless those communications were perceived to be abusive.

In § 5 we provide example explanations given by participants in a qualitative investigation that we conducted with 13 in-person participants, after answering the questionnaire in a manner

indicative of passive or active abuse. While these explanation indeed suggest abuse, a larger study is needed to understand the perceived value of these relationships, and reasons why Facebook users choose to maintain them.

**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 [21] 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.

**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 8. This suggests that decisions to ignore recommendations were not taken randomly, and participants took the time to process this decision. We believe that 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 experiment of § 7, for 11 of the 68 unfriended friend cases, the participants believed that our warning was correct, but still wanted to keep those friends. We conjecture that this may be because the participant had reasons that would make him or her abusive toward that friend. As mentioned by Dinakar et al. [23], determining the victim and the perpetrator in an interaction is not an easy task, as victims may also retaliate thus become perpetrators. We note that AbuSniff is a victim-side abuse prevention tool, thus may protect these friends if they installed AbuSniff.

**Stranger Friends**. Our interest in stranger friends is motivated by the fact that participants in our online experiments 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. We have shown that participants tend to have high numbers of perceived stranger friends. Some of those friends could launch damaging attacks that include cyberbullying (e.g., outing), identity theft, profile cloning and spear phishing. Since any stranger could be an "attack vector", from a security and privacy perspective, it makes sense to sandbox or remove such friends, and minimize the user's "attack surface".

## 12 RELATED WORK

Cyber abuse perpetrated by friends has been considered in the past to be outside the scope of online social network defenses [61]. For instance, Facebook's Immune System (FIS) [61] 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, Wisniewski et al. [76] report forms of user withdrawal from social network interactions, that include self-censorship, detachment, and retreat. Further, Van Kleek et al. [34] found that one reason for people to fabricate, omit or alter the truth online is to avoid harassment or discrimination. While arguably healthy, withdrawal strategies may defeat the "free flow of information" envisioned by Facebook [85].

This article extends the conference version [67] with novel experiments that include a comparison of AbuSniff with a control app (§ 8), an analysis of frequent questionnaire responses, suggested defense actions and corresponding user decisions (§ 7.2), an analysis of participant reasons to ignore AbuSniff recommendations (§ 7.2), an analysis of the time taken by participants to interact

with various elements of the AbuSniff app (§ 7), and a statistical analysis of the quality of the Abuse Prediction Module features as Abuse Inference Engine decision predictors (§ 9).

We organize the remainder of this section into work related to abuse questionnaires, abuse detection and abuse defenses.

### 12.1 Related Work on Cyber Abuse Questionnaires

Our timeline and news feed abuse questions are inspired by existing cyber abuse questionnaires [11, 12, 45, 66, 70, 80]. For instance, our timeline abuse questions build on the partner cyber abuse questionnaire (PCAQ) developed by Wolford et al. [80], that includes questions on whether the partner wrote negative material on the social network. We are also influenced by the cyber dating abuse questionnaire (CDAQ) of Borrajo et al. [11, 12], in particular their question about comments received on the social network wall.

Our news feed abuse question also builds on (1) the PCAQ of Wolford et al. [80], i.e., their questions about the partner sending angry or insulting text or emails, (2) the CDAQ of Borrajo et al. [11, 12], i.e., their questions on posting media with the intent to insult or humiliate, and (3) the questionnaire of Machimbarrena et al. [45] on partners sending threatening or insulting messages.

We note however that partner cyber abuse questionnaires [11, 12, 45, 70, 80] focus on internet and social network interactions between partners who either live together or share login information on social networks. We also note that general cyberbullying questionnaires, e.g., [47, 57, 60], focus on the user perception of general exposure to, or participation in cyberbullying, and do not identify specific perpetrators or victims.

In contrast, we leverage our ability to use identifying information for Facebook friends, to design the AbuSniff questionnaire to be answered not for one partner, but for multiple social networking friends specifically identified by their names and profile photos. AbuSniff presents to each participant, names and photos of friends targeted by the questionnaire, and does not rely on a generic "partner" or "friend" denomination. Further, our additional requirements that the questionnaire is to be shown for multiple friends and to be delivered on a smartphone, impact the number of unique questions that we can include.

### 12.2 Related Work on Cyber Abuse Detection

**Predicting Tie Strength**. The abuse prediction module (APM) of AbuSniff is related to work on predicting tie strength, a concept introduced by Granovetter [30]. Granovetter [30] argued that the strength of the tie between two individuals varies directly with the size of the overlap between their networks. Since then, more features were proposed that impact tie strength in social network interactions. For instance, Banks and Wu [8] used the intensity of interactions with a friend, as the number of initiated conversations, received wall posts, and photo tagging, which are a subset of APM's features. Banks and Wu [8] used the inferred measure of the intensity of interactions with a friend to decide privacy settings on bidirectional data-sharing with the friend.

Several of the mutual activity features of AbuSniff are similar to features introduced by Gilbert and Karahalios [29] to predict tie strength with friends. Our novel features are whether the user and friend live in the same city or come from the same hometown, and the number of places where they studied and worked together. We conjecture that using other features proposed by Gilbert and Karahalios [29] could improve the accuracy of AbuSniff's abuse prediction module. However, several of these features invade user privacy and we are reluctant to access sensitive user information.

**Detection of Friend and Message Spam**. Friend spam detection solutions [17, 33, 52, 81] attempt to identify fake, sockpuppet accounts in social networks. For instance, Cao et al. [17] 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. [81] utilize posting relations between users and messages to combine social spammer and spam message detection. AbuSniff focuses instead on the user *perception* of passively or actively abusive friends, their detection and defenses. However, we note that AbuSniff can be used in conjunction with such solutions to nudge users to unfriend fake, sockpuppet friends.

**Recommender Systems**. Recommender systems [1, 54] exploit knowledge of past user decisions to suggest information relevant to the user goals. Collaborative filtering uses past decisions made by similar users, while content based filtering uses past decisions made by the same user, to recommend new decisions. Recommender system techniques could be used to recommend defenses against detected abusive friends. Since Facebook users have been shown to rarely take advantage of friend blocking mechanisms provided by social networks [78, 79], further investigations are needed to determine the impact of potentially replicating past mistakes made by others, or by the same user.

**Detection of Abuse Instances**. Comment-level techniques to detect abuse, e.g., [18, 23, 48], are orthogonal to, and can be used in conjunction with AbuSniff, to improve abuse detection. For instance, to better detect abusive behaviors, AbuSniff could use the Bag of Communities (BoC) techniques of Chandrasekharan et al. [18], that leverage large-scale data collected from Internet communities. AbuSniff could also use the NLP, supervised learning and reasoning technique of Dinakar et al. [23], or the "toxicity" metric returned for any input sentence by the Perspective API [21], that signals harassment, insults and abusive online speech [31]. Comment-level abuse detection can enable a more accurate identification of specific abuse instances, then convert them into features to predict abuse perception.

**Privacy and Interpersonal Boundary Regulation**. In seminal work, Altman [3] presented privacy as a process of interpersonal boundary regulation. Wisniewski et al. [78] introduce a taxonomy of interpersonal boundaries used by social network users to manage their privacy preferences. Wisniewski et al. [76] document coping behaviors of social network users, to maintain or recover interpersonal boundaries. Notably, while users rarely take advantage of friend blocking mechanisms provided by social networks [78, 79], Wisniewski et al. [76] report that users resort to other solutions (e.g., pseudonymous accounts, using other people's accounts) to achieve similar functionality. AbuSniff can be viewed as an automated tool to manage *interactional boundaries* [78], e.g., to unfollow, restrict, block/sandbox, or unfriend a friend predicted to be perceived as abusive.

Xu et al. [82] analyze insights gained from the Facebook news feed outcry [14] to propose hypotheses about the interplay of privacy concerns and behavioral responses, perceived information control, trust in social network providers and trust in friends. Shi et al. [55] show that violations or changes of contexts, actors, attributes and transmission principles in the friendship pages of Facebook, can result in privacy concerns, while some users asked for more privacy enhancing features, e.g., to opt-out or turn-off such pages. By reducing the information flow between users and predicted abusive friends, AbuSniff implicitly reduces the privacy concerns associated with their corresponding friendship and news feed pages.

The study of Lampinen et al. [38] on coping with social network disclosure by friends (e.g., posting of photos) reveals the importance of collaborative strategies in regulating boundaries. We note that proactive and reactive strategies require substantial human effort, and were shown by Lampinen et al. [38] to be error-prone, especially when disclosure is performed by a weak tie or a stranger friend. In an effort to reduce human effort, we take steps toward automatic, user-transparent detection of perceived friend abuse and regulation of boundaries with predicted perpetrators.

### 12.3 Related Work on Cyber Abuse Defense Strategies and Intervention

Ashktorab and Vitak [5] 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. AbuSniff is an automated *victim-side approach* [5] to detect the user's abuse perception, that avoids the problems associated with accessing individual posts [23].

Vitak and Kim [72] found that to mitigate risks, experienced Facebook users, i.e., 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. Similar to the strategies employed by the experienced participants in the study of Vitak and Kim [72], AbuSniff (1) limits the access to user data for friends perceived to be abusive, (2) hides friends perceived to post offensive, misleading, propaganda or malicious information from the news feed of the user, and (3) unfriends or sandboxes friends who are perceived to be strangers, or who qualify for both points (1) and (2).

Cho and Filippova [20] report 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. Wisniewski et al. [77] show that users can have different privacy management strategies, as well as different awareness levels to the available privacy features, that demonstrate the need to tailor privacy education and nudging, to the end-user. To improve the user ability to manage privacy, Lipford et al. [42] introduce a new interface for managing privacy settings in Facebook, focused around an audience point of view. Raber et al. [53] introduce a user interface that displays privacy settings for historical posts and enables users to meaningfully decide with whom to share social network posts.

The questionnaire based AbuSniff can be viewed as a tool to nudge users towards more privacy-preserving actions [7, 68, 73]. We note however that changing privacy settings alone, is not sufficient to prevent the friend abuse we consider in this article. Setting privacy on a per-friend basis does not scale well: participants recruited in our experiments had an average of 303 friends, and a maximum of 4,880 friends. The task of deciding privacy settings for each post and friend is likely to impose an insurmountable cognitive load on users. AbuSniff takes first steps toward user-transparent, per-friend privacy settings, perhaps similar to spam e-mail filters. In this respect, AbuSniff is also similar to the mobile personalized privacy assistant developed by Liu et al. [43] that recommends privacy setting in Android, and by the agent developed by Amos et al. [6] that uses supervised learning to detect and incriminate deceptive participants in forums and chat-rooms.

## 13 LIMITATIONS

**Understanding of Suggestions**. Each participant in our experiments was presented with a tutorial that explains the meaning of each of the defense actions that may be suggested by AbuSniff, before starting the experiment. However, we did not test the understanding of the participants, of the meaning of the suggested actions, e.g., that they understand the difference between "unfriend" and "sandbox" options. We leave for future work an investigation of interfaces that, for instance, once the user checks the "sandbox" option, display a message informing the user that this will simultaneously "unfollow" and "restrict" the friend. Alternatively, an interface that will replace "sandbox" with "unfollow" and "restrict" boxes, and will allow the user to check multiple boxes.

**Generalization of Results**. Our online participant recruitment process is biased, since we have recruited participants who have Facebook accounts and at least 30 friends, and have an Android device. Further, a majority of our participants were 20-29 years old, which differs for instance from the distribution of U.S. Facebook users. Our results may thus not be representative of the entire Facebook population. Also, we have evaluated AbuSniff only on Facebook and make no claims on the applicability of our results to other social networks.

Further, we have only collected occupation information from participants in our qualitative investigation, but not from participants in our online experiments.

**Cultural Variation**. Our experiments were performed with a diverse set of participants from 25 countries and 6 continents. The diverse cultural background of the participants suggests different notions and perceptions of abuse in general, and abuse from Facebook friends in particular. We consider an exploration of the perception of Facebook friend abuse of participants from different cultures, to be outside the scope of this work, especially as we acknowledge the small number of participants recruited from several countries.

**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 experiments 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.

**Comparison of AbuSniff and Control**. The design colors of the AbuSniff and control apps differ: the background color is white in the control vs. blue in the AbuSniff app; the buttons in the control app emulate the blue vs. gray colors of the Facebook interface. The reason for this is that we attempted to emulate in the control app, the general Facebook mobile app interface. Further, for the AbuSniff app we have collected feedback from 20 randomly selected friends per participant. However, in the control experiment we collected participant feedback from 30 randomly selected friends per participant. We did this in order to compensate for the smaller number of participants in the control experiment, and enable us to compare more balanced sets of participant decisions. We have not however factored in participant fatigue associated with processing 10 additional friends.

The assignment to the AbuSniff and control experiments was done serially and not randomized. However, there was no overlap between the participants in any experiments, and the recruitment process was identical. Figures 5 and 7 show similar demographics between the set of all the participants and only the participants in the above mentioned AbuSniff experiments. Thus, we expect no significant differences between the participants in the two experiments.

**Multiple Accounts**. While we made sure that all the participants in our experiments had "well-formed" accounts (i.e., with at least 30 friends), we have not considered the case of Facebook users who maintain multiple profiles (i.e., user accounts). Stutzman and Hartzog [63] report that the maintenance of multiple profiles is motivated by privacy, identity, utility, and propriety factors. We conjecture future interest in comparing perceived exposure to friend abuse in multiple accounts of the same participant.

**Pre-Test and Post-Test During Control**. We did not include pre-test and post-test surveys in the control experiment, thus cannot compare the impact of the questionnaire-based AbuSniff with the impact of the control app.

## 14 CONCLUSIONS

Social networks provide users with access to sensitive data and communication channels of their friends. In this article we have studied the user perception of abuse perpetrated by Facebook friends. We have introduced AbuSniff, a friend abuse detection and protection system for Facebook. We have developed a questionnaire to study perception of abuse for a specific friend, and rules to convert answers to defensive actions.

We report user perception of relationships with potentially abusive Facebook friends, with no significant differences between male and female participants. We have shown that AbuSniff is more efficient than a control app, in terms of participant willingness to unfriend and restrict friends. Further, we have shown that supervised learning algorithms can use features extracted from mutual social networking activities, 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.

AbuSniff can be viewed as a tool that nudges social network users towards managing their interactional boundaries, and takes steps toward minimizing the cognitive load imposed on users through user-transparent, per-friend privacy settings. This suggests that AbuSniff can be used in conjunction with other defenses employed by social networks, to help prevent abuse that includes cyber stalking, cyberbullying and the distribution of fake news.

## 15 ACKNOWLEDGMENTS

We thank the reviewers for their constructive feedback. We thank Mozhgan Azimpourkivi and Debra Davis for early discussions. This research was supported in part by NSF grants CNS-2013671, CNS-1840714 and CNS-1527153.

## REFERENCES

[1] Gediminas Adomavicius and Alexander Tuzhilin. 2005. Toward the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions. *IEEE Trans. on Knowl. and Data Eng.* 17, 6 (June 2005), 734–749.
[2] Murad Batal Al-Shishani. 2010. Taking al-Qaeda's Jihad to Facebook. *The Jamestown Foundation: Terrorism Monitor* 8, 5 (2010), 3.
[3] Irwin Altman. 1975. The Environment and Social Behavior: Privacy, Personal Space, Territory, and Crowding. (1975).
[4] Jessikka Aro. 2016. The Cyberspace War: Propaganda and Trolling as Warfare Tools. *European View* 15, 1 (2016).
[5] Zahra Ashktorab and Jessica Vitak. 2016. Designing Cyberbullying Mitigation and Prevention Solutions Through Participatory Design With Teenagers. In *Proceedings of CHI*.
[6] Amos Azaria, Ariella Richardson, and Sarit Kraus. 2015. An Agent for Deception Detection in Discussion Based Environments. In *Proceedings of CSCW*.
[7] Rebecca Balebako, Pedro G Leon, Hazim Almuhimedi, Patrick Gage Kelley, Jonathan Mugan, Alessandro Acquisti, Lorrie Faith Cranor, and Norman Sadeh. 2011. Nudging Users Towards Privacy on Mobile Devices. In *Proceedings of the CHI Workshop on Persuasion, Nudge, Influence and Coercion*. 193–201.
[8] Lerone Banks and Shyhtsun Felix Wu. 2009. All Friends Are Not Created Equal: An Interaction Intensity Based Approach to Privacy in Online Social Networks. In *Proceedings of the International Conference on Computational Science and Engineering*, Vol. 4. IEEE, 970–974.
[9] BBC. 2017. Russia-linked posts 'reached 126m Facebook users in US'. BBC News, https://goo.gl/2qy5et.
[10] BBC. 2017. Theresa May accuses Vladimir Putin of election meddling. https://goo.gl/EtMSRF.
[11] E Borrajo, M Gámez-Guadix, and E Calvete. 2015. Cyber Dating Abuse: Prevalence, Context, and Relationship with Offline Dating Aggression. *Psychological Reports* 116, 2 (2015), 565–585.
[12] Erika Borrajo, Manuel Gámez-Guadix, Noemí Pereda, and Esther Calvete. 2015. The Development and Validation of the Cyber Dating Abuse Questionnaire Among Young Couple. *Computers in Human Behavior* 48 (2015), 358 – 365.
[13] Danah Boyd. 2007. 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* (2007), 119–142.
[14] Danah Boyd. 2008. Facebook's Privacy Trainwreck: Exposure, Invasion, and Social Convergence. *Convergence* 14, 1 (2008), 13–20.
[15] Garrett Brown, Travis Howe, Micheal Ihbe, Atul Prakash, and Kevin Borders. 2008. Social Networks and Context-Aware Spam. In *Proceedings of the ACM conference on Computer supported cooperative work*.
[16] Moira Burke and Robert Kraut. 2013. Using Facebook After Losing a Job: Differential Benefits of Strong and Weak Ties. In *Proceedings of the 2013 conference on Computer supported cooperative work*. ACM, 1419–1430.
[17] Qiang Cao, Michael Sirivianos, Xiaowei Yang, and Kamesh Munagala. 2015. Combating Friend Spam Using Social Rejections. In *2015 IEEE 35th International Conference on Distributed Computing Systems*. IEEE, 235–244.
[18] Eshwar Chandrasekharan, Mattia Samory, Anirudh Srinivasan, and Eric Gilbert. 2017. The Bag of Communities: Identifying Abusive Behavior Online with Preexisting Internet Data. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. ACM, 3175–3187.
[19] Justin Cheng, Michael Bernstein, Cristian Danescu-Niculescu-Mizil, and Jure Leskovec. 2017. Anyone Can Become a Troll: Causes of Trolling Behavior in Online Discussions. In *Proceedings of the 2017 ACM conference on computer supported cooperative work and social computing*. ACM, 1217–1230.
[20] Hichang Cho and Anna Filippova. 2016. Networked Privacy Management in Facebook: A Mixed-Methods and Multinational Study. In *Proceedings of the 19th ACM CSCW*.
[21] Jared Cohen. 2017. What if technology could help improve conversations online? https://www.perspectiveapi.com/.
[22] Bernhard Debatin, Jennette P Lovejoy, Ann-Kathrin Horn, and Brittany N Hughes. 2009. Facebook and online privacy: Attitudes, behaviors, and unintended consequences. *Journal of Computer-Mediated Communication* 15, 1 (2009), 83–108.

[23] Karthik Dinakar, Birago Jones, Catherine Havasi, Henry Lieberman, and Rosalind Picard. 2012. Common sense reasoning for detection, prevention, and mitigation of cyberbullying. *ACM TiiS* (2012).
[24] David M Douglas. 2016. Doxing: a conceptual analysis. *Ethics and information technology* 18, 3 (2016), 199–210.
[25] Harald Dreßing, Josef Bailer, Anne Anders, Henriette Wagner, and Christine Gallas. 2014. Cyberstalking in a large sample of social network users: Prevalence, characteristics, and impact upon victims. *Cyberpsychology, Behavior, and Social Networking* 17, 2 (2014), 61–67.
[26] Robin IM Dunbar. 1992. Neocortex size as a constraint on group size in primates. *Journal of human evolution* 22, 6 (1992), 469–493.
[27] Nicole B Ellison, Charles Steinfield, and Cliff Lampe. 2007. The Benefits of Facebook "Friends:" Social Capital and College Students' Use of Online Social Network Sites. *Journal of computer-mediated communication* 12, 4 (2007), 1143–1168.
[28] Hongyu Gao, Jun Hu, Christo Wilson, Zhichun Li, Yan Chen, and Ben Y Zhao. 2010. Detecting and Characterizing Social Spam Campaigns. In *Proceedings of the 10th ACM SIGCOMM conference on Internet measurement*. ACM, 35–47.
[29] Eric Gilbert and Karrie Karahalios. 2009. Predicting Tie Strength with Social Media. In *Proceedings of the SIGCHI conference on human factors in computing systems*. ACM, 211–220.
[30] Mark S Granovetter. 1973. The Strength of Weak Ties. *Amer. J. Sociology* 78 (1973), 1360–1380. Issue 6.
[31] Andy Greenberg. 2017. Now Anyone Can Deploy Google's Troll-Fighting AI. Wired Magazine.
[32] Alex Heath. 2017. Facebook Quietly Updated Two Key Numbers About Its User Base. Business Insider, https://goo.gl/LCLfBx.
[33] Markus Huber, Martin Mulazzani, and Edgar Weippl. 2010. Who on Earth Is Mr. Cypher: Automated Friend Injection Attacks on Social Networking Sites. In *Security and Privacy–Silver Linings in the Cloud*. Springer, 80–89.
[34] Max Van Kleek, Daniel A. Smith, Nigel R. Shadbolt, Dave Murray-Rust, and Amy Guy. 2015. Self Curation, Social Partitioning, Escaping from Prejudice and Harassment: The Many Dimensions of Lying Online. In *Proceedings of the ACM WWW*.
[35] Georgios Kontaxis, Iasonas Polakis, Sotiris Ioannidis, and Evangelos P Markatos. 2011. Detecting Social Network Profile Cloning. In *2011 IEEE international conference on pervasive computing and communications workshops (PERCOM Workshops)*. IEEE, 295–300.
[36] Robin M Kowalski, Susan P Limber, Sue Limber, and Patricia W Agatston. 2012. *Cyberbullying: Bullying in the Digital Age*. Wiley & Sons.
[37] Grace Chi En Kwan and Marko M Skoric. 2013. Facebook bullying: An extension of battles in school. *Computers in Human Behavior* 29, 1 (2013), 16–25.
[38] Airi Lampinen, Vilma Lehtinen, Asko Lehmuskallio, and Sakari Tamminen. 2011. We're in It Together: Interpersonal Management of Disclosure in Social Network Services. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 3217–3226.
[39] Issie Lapowsky. 2018. 2019 State of Malware. Wired, https://goo.gl/5dBtty.
[40] Issie Lapowsky. 2018. Facebook Exposed 87 Million Users To Cambridge Analytica. Wired, https://www.wired.com/story/facebook-exposed-87-million-users-to-cambridge-analytica/.
[41] David Lee. 2017. Facebook, Twitter and Google berated by senators on Russia. https://goo.gl/288SmQ.
[42] Heather Richter Lipford, Andrew Besmer, and Jason Watson. 2008. Understanding Privacy Settings in Facebook with an Audience View. In *Proceedings of the 1st Conference on Usability, Psychology, and Security (UPSEC'08)*. USENIX Association, 2:1–2:8.
[43] Bin Liu, Mads Schaarup Andersen, Florian Schaub, Hazim Almuhimedi, Shikun (Aerin) Zhang, Norman Sadeh, Yuvraj Agarwal, and Alessandro Acquisti. 2016. Follow My Recommendations: A Personalized Privacy Assistant for Mobile App Permissions. In *Proceedings of SOUPS*.
[44] Yabing Liu, Krishna P Gummadi, Balachander Krishnamurthy, and Alan Mislove. 2011. Analyzing facebook privacy settings: user expectations vs. reality. In *Proceedings of the ACM IMC*.
[45] Juan Machimbarrena, Esther Calvete, Liria Fernández-González, Aitor Álvarez-Bardón, Lourdes Álvarez-Fernández, and Joaquín González-Cabrera. 2018. Internet Risks: An Overview of Victimization in Cyberbullying, Cyber Dating Abuse, Sexting, Online Grooming and Problematic Internet Use. *International journal of environmental research and public health* 15, 11 (2018), 2471.
[46] Michelle Madejski, Maritza Johnson, and Steven M Bellovin. 2012. A study of privacy settings errors in an online social network. In *Proceedings of PERCOM Workshops*.
[47] Faye Mishna, Charlene Cook, Tahany Gadalla, Joanne Daciuk, and Steven Solomon. 2010. Cyber Bullying Behaviors Among Middle and High School Students. *American Journal of Orthopsychiatry* 80, 3 (2010), 362–374. https://doi.org/10.1111/j.1939-0025.2010.01040.x
[48] Vishwajeet Narwal, Mohamed Hashim Salih, Jose Angel Lopez, Angel Ortega, John O'Donovan, Tobias Höllerer, and Saiph Savage. 2017. Automated Assistants to Identify and Prompt Action on Visual News Bias. In *ACM CHI*.

[49] Amanda Nosko, Eileen Wood, and Seija Molema. 2010. All About Me: Disclosure in Online Social Networking Profiles: The Case of Facebook. *Computers in Human Behavior* 26, 3 (2010).
[50] Barbara Ortutay and Anick Jesdanun. 2018. How Facebook Likes Could Profile Voters for Manipulation. ABC News, https://goo.gl/eD6Ap3.
[51] Sameer Patil. 2012. Will You Be My Friend?: Responses to Friendship Requests from Strangers. In *Proceedings of the iConference*. ACM, 634–635.
[52] Daniele Quercia and Stephen Hailes. 2010. Sybil Attacks Against Mobile Users: Friends and Foes to the Rescue. In *2010 Proceedings IEEE INFOCOM*. IEEE, 1–5.
[53] Frederic Raber, Alexander De Luca, and Moritz Graus. 2016. Privacy Wedges: Area-Based Audience Selection for Social Network Posts. In *Proceedings of SOUPS*.
[54] Paul Resnick and Hal R. Varian. 1997. Recommender Systems. *Commun. ACM* 40, 3 (March 1997), 56–58.
[55] Pan Shi, Heng Xu, and Yunan Chen. 2013. Using Contextual Integrity to Examine Interpersonal Information Boundary on Social Network Sites. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 35–38.
[56] Vivek K. Singh, Marie L. Radford, Qianjia Huang, and Susan Furrer. 2017. "They basically like destroyed the school one day": On Newer App Features and Cyberbullying in Schools. In *ACM CSCW*.
[57] Robert Slonje and Peter K Smith. 2008. Cyberbullying: Another Main Type of Bullying? *Scandinavian journal of psychology* 49, 2 (2008), 147–154.
[58] Kit Smith. 2018. 47 Incredible Facebook Statistics and Facts. https://www.brandwatch.com/blog/47-facebook-statistics/.
[59] Peter Snyder, Periwinkle Doerfler, Chris Kanich, and Damon McCoy. 2017. Fifteen minutes of unwanted fame: Detecting and characterizing doxing. In *Proceedings of the 2017 Internet Measurement Conference*. 432–444.
[60] Mona E Solberg and Dan Olweus. 2003. Prevalence Estimation of School Bullying with the Olweus Bully/Victim Questionnaire. *Aggressive Behavior: Official Journal of the International Society for Research on Aggression* 29, 3 (2003), 239–268.
[61] Tao Stein, Erdong Chen, and Karan Mangla. 2011. Facebook Immune System. In *Proceedings of the 4th Workshop on Social Network Systems*. ACM, 8.
[62] Gianluca Stringhini, Christopher Kruegel, and Giovanni Vigna. 2010. Detecting Spammers on Social Networks. In *Proceedings of the 26th Annual Computer Security Applications Conference*. ACM, 1–9.
[63] Frederic Stutzman and Woodrow Hartzog. 2012. Boundary Regulation in Social Media. In *Proceedings of the ACM Conference on Computer Supported Cooperative Work*. ACM, 769–778.
[64] Fred Stutzman and Jacob Kramer-Duffield. 2010. Friends Only: Examining a Privacy-Enhancing Behavior in Facebook. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 1553–1562.
[65] Frederic D Stutzman, Ralph Gross, and Alessandro Acquisti. 2013. Silent Listeners: The Evolution of Privacy and Disclosure on Facebook. *Journal of Privacy and Confidentiality* 4, 2 (2013), 2.
[66] Sajedul Talukder and Bogdan Carbunar. 2017. When Friend Becomes Abuser: Evidence of Friend Abuse in Facebook. In *Proceedings of the 9th ACM Conference on Web Science (WebSci '17)*. ACM, New York, NY, USA. https://doi.org/10.1145/3091478.3098869
[67] Sajedul Talukder and Bogdan Carbunar. 2018. AbuSniff: Automatic Detection and Defenses Against Abusive Facebook Friends. In *Twelfth International AAAI Conference on Web and Social Media*.
[68] Richard H Thaler and Cass R Sunstein. 2009. *Nudge: Improving Decisions About Health, Wealth, and Happiness*. Penguin.
[69] Kurt Thomas and David M. Nicol. 2010. The Koobface Botnet and the Rise of Social Malware. In *5th International Conference on Malicious and Unwanted Software, MALWARE*. 63–70.
[70] Joris Van Ouytsel, Koen Ponnet, and Michel Walrave. 2017. Cyber Dating Abuse: Investigating Digital Monitoring Behaviors Among Adolescents from a Social Learning Perspective. *Journal of interpersonal violence* (2017), 0886260517719538.
[71] Onur Varol, Emilio Ferrara, Clayton A Davis, Filippo Menczer, and Alessandro Flammini. 2017. Online human-bot interactions: Detection, estimation, and characterization. In *Proceedings of ICWSM*.
[72] Jessica Vitak and Jinyoung Kim. 2014. You can't block people offline": examining how Facebook's affordances shape the disclosure process. In *Proceedings of CSCW*.
[73] Yang Wang, Pedro Giovanni Leon, Kevin Scott, Xiaoxuan Chen, Alessandro Acquisti, and Lorrie Faith Cranor. 2013. Privacy Nudges for Social Media: An Exploratory Facebook Study. In *Proceedings of the 22nd International Conference on World Wide Web*. ACM, 763–770.
[74] Gabriel Weimann. 2010. Terror on Facebook, Twitter, and YouTube. *The Brown Journal of World Affairs* 16, 2 (2010).
[75] Pamela Wisniewski, AKM Islam, Bart P Knijnenburg, and Sameer Patil. 2015. Give Social Network Users the Privacy They Want. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*. ACM, 1427–1441.
[76] Pamela Wisniewski, Heather Lipford, and David Wilson. 2012. Fighting for My Space: Coping Mechanisms for Sns Boundary Regulation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 609–618.

[77] Pamela J Wisniewski, Bart P Knijnenburg, and Heather Richter Lipford. 2017. Making Privacy Personal: Profiling Social Network Users to Inform Privacy Education and Nudging. *International Journal of Human-Computer Studies* 98 (2017), 95–108.

[78] Pamela J Wisniewski, AKM Najmul Islam, Heather Richter Lipford, and David C Wilso. 2016. Framing and Measuring Multi-dimensional Interpersonal Privacy Preferences of Social Networking Site Users. *Communications of the Association for information systems* 38, 1 (2016).

[79] Pamela Karr Wisniewski, David C. Wilson, and Heather Richter Lipford. 2011. A New Social Order: Mechanisms for Social Network Site Boundary Regulation. In *Proceedings of the 17th Americas Conference on Information Systems*.

[80] Caitlin Wolford-Clevenger, Heather Zapor, Hope Brasfield, Jeniimarie Febres, JoAnna Elmquist, Meagan Brem, Ryan C Shorey, and Gregory L Stuart. 2016. An Examination of the Partner Cyber Abuse Questionnaire in a College Student Sample. *Psychology of Violence* 6, 1 (2016), 156.

[81] Fangzhao Wu, Jinyun Shu, Yongfeng Huang, and Zhigang Yuan. 2015. Social Spammer and Spam Message Co-Detection in Microblogging with Social Context Regularization. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*. ACM, 1601–1610.

[82] Heng Xu, Rachida Parks, Chao-Hsien Chu, and Xiaolong (Luke) Zhang. 2010. Information Disclosure and Online Social Networks: From the Case of Facebook News Feed Controversy to a Theoretical Understanding. In *Proceedings of the 16th Americas Conference on Information Systems*.

[83] Chao Yang and Padmini Srinivasan. 2014. Translating Surveys to Surveillance on Social Media: Methodological Challenges & Solutions. In *Proceedings of the 2014 ACM conference on Web science*. ACM, 4–12.

[84] Jeff Yates. 2017. From Temptation to Sextortion Inside the Fake Facebook Profile Industry. Radio Canada, http://ici.radio-canada.ca/special/sextorsion/en.

[85] Mark Zuckerberg. 2006. An Open Letter from Mark Zuckerberg.