AbuSniff: Automatic Detection and Defenses Against Abusive Facebook Friends

Sajedul Talukder
Florida Int'l University, Miami, USA
stalu001@fiu.edu

Bogdan Carbunar
Florida Int'l University, Miami, USA
carbunar@gmail.com

Abstract

Adversaries leverage social network friend relationships to collect sensitive data from users and target them with abuse that includes fake news, cyberbullying, malware, and propaganda. Case in point, 71 out of 80 user study participants had at least 1 Facebook friend with whom they never interact, either in Facebook or in real life, or whom they believe is likely to abuse their posted photos or status updates, or post offensive, false or malicious content. We introduce AbuSniff, a system that identifies Facebook friends perceived as strangers or abusive, and protects the user by unfriending, unfollowing, or restricting the access to information for such friends. We develop a questionnaire to detect perceived strangers and friend abuse. We introduce mutual Facebook activity features and show that they can train supervised learning algorithms to predict questionnaire responses.

We have evaluated AbuSniff through several user studies with a total of 263 participants from 25 countries. After answering the questionnaire, participants agreed to unfollow and restrict abusers in 91.6% and 90.9% of the cases respectively, and sandbox or unfriend non-abusive strangers in 92.45% of the cases. Without answering the questionnaire, participants agreed to take the AbuSniff suggested action against friends predicted to be strangers or abusive, in 78.2% of the cases. AbuSniff increased the participant self-reported willingness to reject invitations from strangers and abusers, their awareness of friend abuse implications and their perceived protection from friend abuse.

Introduction

Social networks provide an ideal platform for abuse, that includes the collection and misuse of private user information (Yang and Srinivasan 2014; Yates 2017; Kontaxis et al. 2011), cyberbullying (Singh et al. 2017; Kwan and Skoric 2013), and the distribution of offensive, misleading, false or malicious information (Cheng et al. 2017; Al-Shishani 2010; Weimann 2010; Aro 2016). The propensity of social networks towards such abuse has brought intense scrutiny and criticism from users, media, and politicians (Smith 2008; Angwin and Grassegger 2017; Shahani 2016; Lee 2017; Wakefield 2017).

Social networks like Facebook have made progress in raising user awareness to the dangers of making information public and the importance of deciding who can access it. However, many users still allow their Facebook friends to access their information, including timeline and news feed. This, coupled with the fact that people often have significantly more than 150 Facebook friends[1] – the maximum number of meaningful friend relationships that humans can manage (Dunbar 1992)) – suggests that Facebook users may still be vulnerable to attacks.

To evaluate user perception of exposure to abusive friend behaviors, we designed 2 user studies (total n = 80) where each participant had to evaluate 20 of their randomly selected Facebook friends. 65 of the 80 participants admitted to have at least 1 friend whom they perceived would abuse their status updates or pictures, and 60 of the participants had at least 1 friend whom they perceived would post abusive material (i.e., offensive, misleading, false or malicious). This is consistent with recent revelations of substantial abuse perpetrated through Facebook, including Cambridge Analytica's injection of content to change user perception (Ortutay and Jesdanun 2018), and Facebook's admission that in the past two years Russia-based operatives created 80,000 posts that have reached 126 million users in the US (BBC 2017a; Lee 2017).

Further, 55 of the 80 participants admitted to have at least 1 Facebook friend with whom they have never interacted, either online or in person. Such *stranger* friends could be bots (Varol et al. 2017) that passively collect sensitive user data and later use it against the user's best interest, as we also show through pilot study answers. This corroborates Facebook's recent estimate that 13% (i.e., 270 million) accounts are either bots or clones (Heath 2017). Stranger friends can use the collected data to infer other sensitive user information (Yang and Srinivasan 2014), identify "deep-seated underlying fears, concerns" by companies such as Cambridge Analytica (Lapowsky 2018), perform profile cloning (Kontaxis et al. 2011), sextorsion (Yates 2017), identity theft (Nosko, Wood, and Molema 2010), and spear phishing (Gao et al. 2010) attacks.

These studies signal the need for defenses against abusive and stranger friends, that include restricting the abusers' access to user information, unfollowing them and even unfriending - removing them from the friend list. When asked

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

directly, participants in our studies unfollowed and restricted access for abusers in 91.6% and 90.9% of the cases, respectively. When informed about the potential privacy risks posed by stranger friends, participants chose to unfriend or sandbox (block bi-directional communications with) such friends in 92.45% of the cases.

Contributions. We develop AbuSniff (Abuse from Social Network Friends), a system that evaluates, predicts and protects users against perceived friend abuse in Facebook. AbuSniff has the potential to mitigate the effects of abuse, and reduce its propagation through social networks and even its negative impact on social processes (e.g., electoral). We introduce the following contributions:

- Develop a friend abuse questionnaire that captures the user perception that a Facebook friend (1) is a stranger, (2) would publish abusive responses to pictures and status updates posted by the user, or (3) would publish and distribute offensive, misleading, false or potentially malicious information. Devise rules to convert identified abuse into defense actions.
- Propose the hypothesis that data recorded by Facebook can be used to predict the user perception of friend abuse. Introduce *mutual activity features* that quantify the Facebook recorded interactions between a user and her friend. Use supervised learning algorithms trained on these features to predict (1) user answers to the questionnaire, thus user perceived strangers and friend abuse and (2) the user willingness to take defensive actions against such friends.
- Implemented AbuSniff in Android (open source upon publication), and evaluated it through user studies with 263 participants from 25 countries and 6 continents.

Results. When using data we collected from 1,452 friend relationships (n=57), we found that supervised learning algorithms trained on AbuSniff's mutual activity features were able to predict the user answers to the questionnaire questions, with an F-measure ranging between 69.2% and 89.7%. Further, AbuSniff was able to predict the cases where the users chose to ignore the suggested defensive action against friends, with an F-Measure of 97.3%.

In a user study ($n = 40$) involving 1,200 Facebook friends, we found that without having to answer the questionnaire, participants accepted 78% of AbuSniff's recommendations for defensive actions against abusive friends and strangers. In another study ($n = 31$) AbuSniff increased participant self-reported willingness to reject invitations from perceived strangers and abusers, their awareness of friend abuse implications and perceived protection from friend abuse.

Background and Model

We briefly summarize the relevant features of Facebook. Facebook users form *friend* relationships. Each user has a *friend list* of other users with whom she has formed friend relationships. The *timeline* (a.k.a wall, or profile) is Facebook's central feature, the place where the user can share her 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, see e.g., (Madejski, Johnson, and Bellovin 2012). Further, a user's *news feed* shows stories created by her friends, groups, and subscribed events. Stories are sorted based on various features, e.g., post time and type, poster popularity.

Adversary Model

We consider adversaries who leverage the following mechanisms to perpetrate abuse through Facebook:

- **Privacy abuse**. Collect sensitive information (profiles, photos, friend lists, locations visited, opinions) posted by friends on their timelines or take screenshots of stories. The adversary can then use this data to infer more sensitive information (Yang and Srinivasan 2014), initiate sextorsion (Yates 2017), perform profile cloning (Kontaxis et al. 2011), identity theft (Nosko, Wood, and Molema 2010), and spear phishing (Gao et al. 2010) attacks. Facebook estimates that 13% (i.e., 270 million) of their accounts are either bots or clones (Heath 2017).
- **Timeline abuse**. Post abusive replies to stories (e.g., status updates, photos) posted by friends on their timeline. The abusive replies appear on the timeline of the victim, where the original stories were posted.
- **News-feed abuse**. The adversary posts abusive material on his timeline, which is then propagated to the news feed of his friends. Abusive information includes material perceived to be offensive, misleading, false, or malicious. Facebook revealed that Russia-based operatives created 80,000 posts that have reached 126 million US users (BBC 2017a; Lee 2017).

Restrictive Actions Against Friends

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

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

Research Objectives

We study several key questions about friend based abuse:

- **(RQ1)**: Are perceived strangers and friend abuse real problems in Facebook?
- **(RQ2)**: Are Facebook users willing to take defensive actions against abusive friends?
- **(RQ3)**: Does AbuSniff have an impact on the willingness of users to take defensive actions on Facebook friends, and is this willingness impacted by the type of abuse perpetrated by the friend and the suggested defensive action?

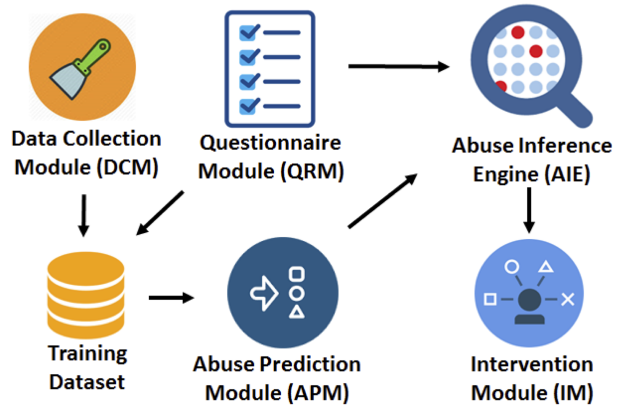

Data Collection Module (DCM)
Questionnaire Module (QRM)
Abuse Inference Engine (AIE)
Training Dataset
Abuse Prediction Module (APM)
Intervention Module (IM)

Figure 1: AbuSniff system architecture. The QRM module delivers the questionnaire. The DCM module collects the user responses and also Facebook data concerning the relationship with each friend. The APM module uses the collected data to predict the responses to the questionnaire. The AIE module uses the output of the QRM or APM to identify abusive friends, and the IM module asks the user to take a protective action against them.

- **(RQ4)**: Can AbuSniff predict abusive friends and the defenses that users are willing to take against such friends?
- **(RQ5)**: Does AbuSniff impact user awareness of stranger and abusive friends, and their perception of safety from such friends?

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

The AbuSniff System

We have designed the AbuSniff system to help us investigate these questions. AbuSniff is a mobile app that asks the user to login to her Facebook account. As illustrated in Figure 1, AbuSniff consists of modules to collect user responses and data, predict user responses, identify abusive friends, and recommend defensive actions. In the following, we describe each module.

The Questionnaire Module (QRM)

We have designed a questionnaire intended to capture the user perception of (potentially) abusive behaviors from friends in Facebook. Since Facebook users tend to have hundreds and even thousands of friends, we decided to present the questionnaire for each of only a randomly selected subset of the user's friends. One design goal was that the questions should help identify the perceived use of the abusive mechanisms listed in the adversary model. To ensure a simple navigation of the questionnaire, we further sought to fit all the questions on a single screen for a variety of popular smartphones. We have designed the questionnaire through an iterative process that included a focus group and a pilot study with 2 K-8 teachers, 1 psychologist, 8 students, 1 dentist and 1 homemaker (8 female and 5 male).

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*) determine the user's frequency of interaction with the friend, on Facebook and in real life. The options are "Frequently", "Occasionally", "Not Anymore" (capturing the case of estranged friends), "Never" and "Don't Remember". We are particularly interested in the "Never" responses.

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

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

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

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

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

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

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

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

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

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

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

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

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

These examples show that privacy and security abuses occur in the real life interactions of Facebook users and their

	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. Like firewall filters, the first matching rule applies. $!A$ denotes any response different from A. NOP = no operation.

friends. The following AbuSniff modules seek to predict the user perception of abuse and convert it into defensive actions that users will consider appropriate.

The Abuse Prediction and Data Collection Modules

We investigate the ability of a supervised learning approach to provide an affirmative answer to question RQ4 (*can AbuSniff predict the abusive friends and the defenses that users are willing to take against such friends?*).

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

The abuse prediction module (APM) uses supervised learning algorithms trained on these features, and previously collected questionnaire responses and user decisions, to predict the user's answers to the QRM questionnaire and the user's reactions to suggested actions. We report accuracy results in the evaluation section.

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

The Abuse Inference Engine (AIE)

(Vitak and Kim 2014) found that to mitigate risks, prudent social network 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. AbuSniff seeks to provide similarly safe social interactions to regular social network users. To this end, the abuse inference engine (AIE) takes as input the responses collected by the QRM or predicted by the APM, and outputs suggested actions from the set { "unfriend", "unfollow", "restrict access", "sandbox", "ignore"}.

AIE uses the rules shown in Table 1, applied on a first match basis: rule r is evaluated only if all the rules 1 to r - 1 have failed. The first 15 rules detect restrictive actions; if none matches, the last rule decides that the friend is not abusive (i.e., ignore). Initially, we took a hard stance against abuse: a friend who scores negatively on any 2 out of the 5 questions (i.e., assigned "Never" in any of the 2 first questions, "Agree" in any of the last 3 questions) should be unfriended (rules 1-11). However, 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). We evaluate and adjust these rules in the evaluation section.

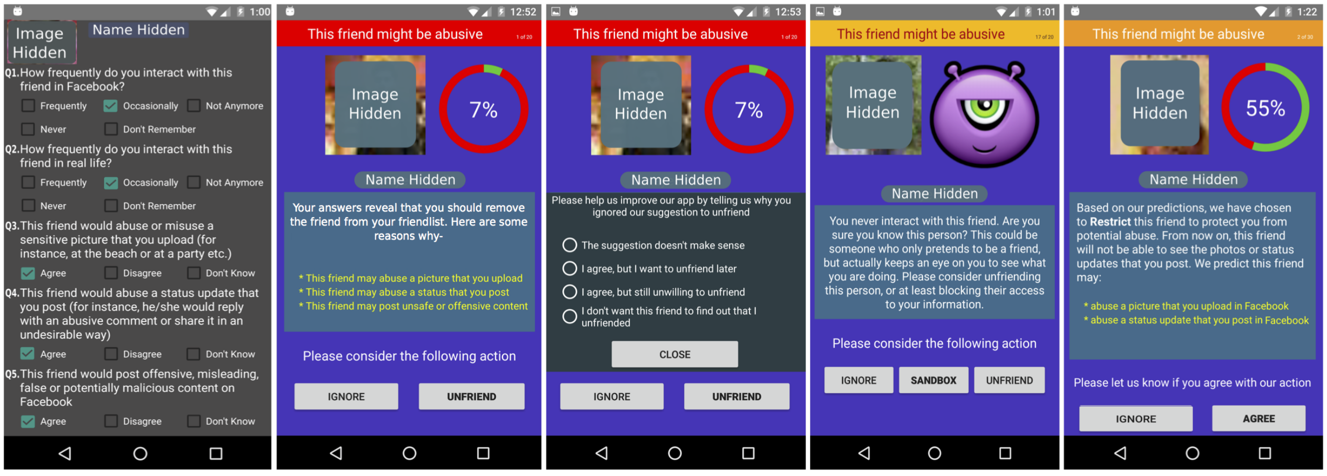
The Intervention Module (IM)

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

The user is offered the option to accept or ignore the suggestion. If the user chooses to ignore the suggestion, the IM module asks the user (through a PopupWindow) to provide a reason, see Figure 2(c). We have conducted a focus group with 20 participants in order to identify possible reasons for ignoring "unfriend" recommendations. They include "the suggestion does not make sense", "I agree, but I want to unfriend later", "I agree but I am still unwilling to unfriend", and "I don't want this friend to find out that I unfriended", see Figure 2(c). We did not include an open text box, as we did not expect that participants will type an answer on a mobile device.

The IM module educates users about the meaning and dangers of having a stranger as a friend, see Figure 2(d). It also offers the option to "sandbox" such friends. According to the rules of Table 1, IM also suggests unfollowing or restricting friends who are abusive in only one direction of their communications. Figure 2(e) shows a snapshot of the restrict screen, its meaning and reasons for selection.



Q1. How frequently do you interact with this friend in Facebook?

□ Frequently ☑ Occasionally □ Not Anymore □ Never □ Don't Remember

Q2. How frequently do you interact with this friend in real life?

□ Frequently ☑ Occasionally □ Not Anymore □ Never □ Don't Remember

Q3. This friend would abuse or misuse a sensitive picture that you upload (for instance, at the beach or at a party etc.)

☑ Agree □ Disagree □ Don't Know

Q4. This friend would abuse a status update that you post (for instance, he/she would reply with an abusive comment or share it in an undesirable way)

☑ Agree □ Disagree □ Don't Know

Q5. This friend would post offensive, misleading, false or potentially malicious content on Facebook

☑ Agree □ Disagree □ Don't Know

(a)

Your answers reveal that you should remove the friend from your friendlist. Here are some reasons why-

* This friend may abuse a picture that you upload
* This friend may abuse a status that you post
* This friend may post unsafe or offensive content

Please consider the following action

(b)

Please help us improve our app by telling us why you ignored our suggestion to unfriend

○ The suggestion doesn't make sense
○ I agree, but I want to unfriend later
○ I agree, but still unwilling to unfriend
○ I don't want this friend to find out that I unfriended

(c)

You never interact with this friend. Are you sure you know this person? This could be someone who only pretends to be a friend, but actually keeps an eye on you to see what you are doing. Please consider unfriending this person, or at least blocking their access to your information.

Please consider the following action

(d)

Based on our predictions, we have chosen to **Restrict** this friend to protect you from potential abuse. From now on, this friend will not be able to see the photos or status updates that you post. We predict this friend may:

* abuse a picture that you upload in Facebook
* abuse a status update that you post in Facebook

Please let us know if you agree with our action

(e)

Figure 2: Anonymized screenshots of the Android AbuSniff app: (a) QRM questionnaire. The first two questions identify stranger friends, questions 3 and 4 identify perceived timeline abuse and question 5 identifies perceived news feed abuse. (b) The IM UI asking the user to unfriend an abusive friend also explains the reasons for the action, according to the questionnaire responses. (c) The IM UI asking the user to explain the reasons for the unwillingness to unfriend in the previous screen. (d) The "unfriend or sandbox" UI for privacy abuse: sandboxing isolates but does not unfriend or notify the friend. (e) The UI of the autonomous AbuSniff asking user confirmation to restrict the access of a friend predicted to be a timeline abuser.

User Study

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

We have recruited 325 participants from JobBoy (Job 2017), during 7 studies conducted between August 2016 and October 2017. The jobs we posted asked the participants to install AbuSniff from the Google Play store, use it to login to their Facebook accounts and follow the instructions on the screen. A participant who successfully completes the app, receives on the last screen a code required for payment. We have paid each participant $3, with a median job completion time of 928s (SD = 420s).

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

- **Attention-check screen**. To ensure that the participants pay attention and are able to understand and follow simple instructions in English, AbuSniff includes a standard attention-check screen at the beginning of the app.
- **Bogus friends**. To detect participants who answer questions at random, we used "bogus friends": 3 fake identities (2 female, 1 male) that we included at random positions in the AbuSniff questionnaire. 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 these mechanisms to discard 62 of the recruited 325 participants. The following results are shown over the remaining 263 participants. Figure 3 shows the distribution of the country of origin (left) and age (right), by gender, over these participants. The 151 male and 112 female participants are from 25 countries (top 5: US, Bangladesh, India, Nepal and UK) and 6 continents, and are between 18-52 years old (M = 23, SD = 7.22).

Ethical Data Collection

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

Results

Abuse Perception and Willingness to Defend

We developed 2 preliminary studies (n = 20 and n = 60) to evaluate the extent of the user perception of stranger friends and friend abuse in Facebook (RQ1) and the willingness of users to accept defensive actions against friends considered

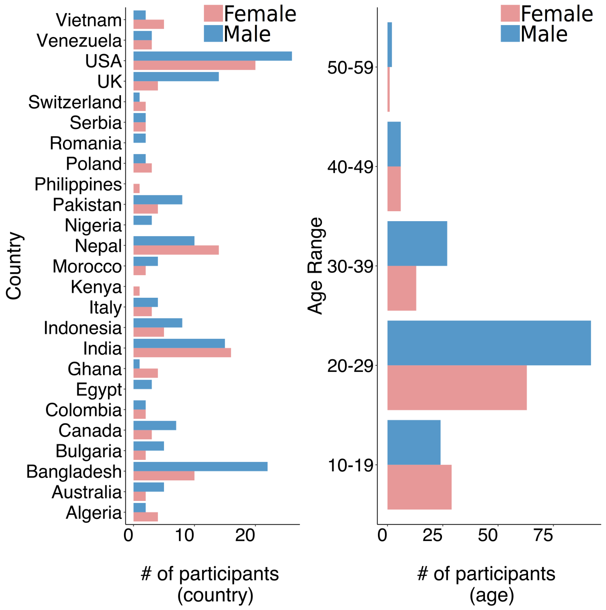


Figure 3: Participant demographics. (country) Distribution of the 25 countries of residence by gender. (age) Distribution of age range by gender.

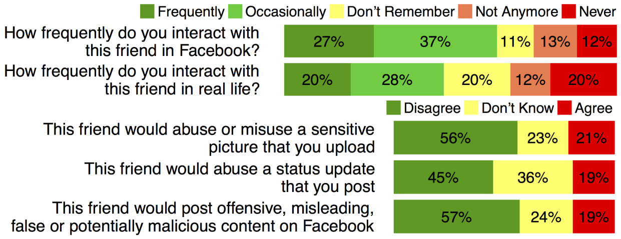


Figure 4: Distribution of responses for the friend abuse questionnaire over 1,600 Facebook friend relationships. The red sections correspond to potential strangers or abusive friends.

to be abusive (RQ2 and RQ3). To this end, AbuSniff used the QRM, DCM, AIE and IM modules, see Figure 1. Further, AbuSniff randomly selected 20 Facebook friends of 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.

Figure 4 shows the distribution of the responses for each of the 5 questions from the 1,600 friend relationships (20 from each participant). Further, 64 of the 80 participants stated that they have at least one friend with whom they have never interacted in Facebook, while 73 of the participants had at least one friend with whom they have never interacted in real life. 68 of the participants had at least 1 friend whom they perceived would abuse their photos, 62 of the participants have at least 1 friend who would abuse their status updates, and 62 have at least 1 friend who would post abusive content.

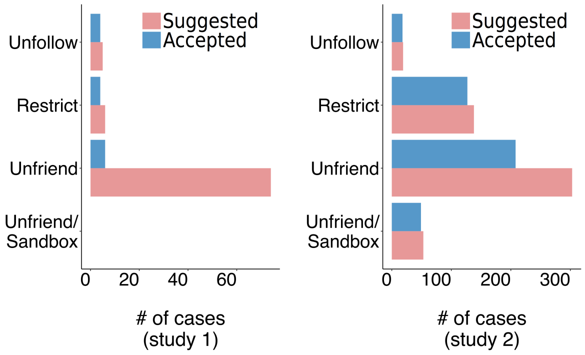


Figure 5: Recommendation vs. acceptance in study 1 (n = 20) and study 2 (n = 60). **The "sandbox" option and user education were effective**: 92% of the suggested "unfriend or sandbox" suggestions were approved by participants.

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

Willingness to Defend Against Abuse. In the first of the above 2 studies (n = 20, 400 investigated friend relationships), AbuSniff identified 85 abusive friend relations. Of these, AbuSniff recommended 74 to unfriend, 6 to restrict and 5 to unfollow. The results are summarized in Figure 5(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. In 55 of the 68 (74 - 6) unfriended friends, the participants believed that our warning was correct. However, they refused to unfriend because either they were not ready to unfriend them at that time (18 of the 55 cases), they still wanted to keep those abusive friends (11 cases), or they were afraid that this action will be observable by the abuser (26 cases).

The "Sandbox" Effect. To address the fear of being observed by the unfriended friend, we have relaxed rule 1 in Table 1, to give the user the option to either sandbox or unfriend a non-abusive stranger. A sandboxed friend can no longer harm the user, as all Facebook communication lines are interrupted. Sandboxing achieves this without severing the friend link, thus is not observable by the friend. Further, we have modified AbuSniff's UI to educate the user through a description of the harm that strangers can perform, and of the defenses that the user can take against such friends. Figure 2(d)) shows a snapshot of the modified UI screen that offers the sandbox alternative to unfriending strangers.

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

Table 2: Precision, recall and F-measure of APM for questions Q1 (RF), Q2 (RF), Q3 (DT), Q4 (DT) and Q5 (RF).

The second user study described above ($n = 60$) evaluated the updated AbuSniff, that identified a total of 513 abusive friend relations. Figure 5(study 2) shows that AbuSniff recommended 303 to unfriend, 53 to unfriend or sandbox, 138 to restrict and 19 to unfollow. Consistent with the first study, 18 of the 19 unfollow and 127 of the 138 restrict suggestions were accepted. In contrast to the first study, 49 of the 53 "unfriend or sandbox" suggestions were accepted. In addition, 208 of 303 "pure" unfriend recommendations were accepted, again a significant improvement over the first study (6 out of 74). Only 5 out of the 95 ignored unfriend recommendations were due to the participant not believing our recommendation.

Efficacy of Abuse Prediction

To answer key question RQ4, in a fourth study we used AbuSniff to collect a subset of Facebook data from 1,452 friend relationships of 54 participants. We have computed the 7 mutual activity features of the 54 participants and the 1,452 friends, and used 10-fold cross validation to evaluate the ability of the abuse prediction module (APM) to predict questionnaire responses and user defense decisions.

As shown in Figure 4, the distribution of the answers to the 5 questions of the questionnaire was not balanced. To address this imbalance, we have duplicated tuples from the minority classes up to the the number of the majority class. We have ensured that duplicates appear in the same fold, to prevent testing on trained tuples. We have used Weka 3.8.1 (Wek 2017) to test several supervised learning algorithms, including Random Forest (RF), Decision Trees (DT), SVM, PART, MultiClassClassifier, SimpleLogistic, K-Nearest Neighbors (KNN) and Naive Bayes, but report only the best performing algorithm.

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 3: 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.**

Predicting questionnaire answers. Table 2 shows the precision, recall and F-measure achieved by the best performing supervised learning algorithm for each of the questionnaire questions (Q1-Q5). The RF classifier achieved the best F-measure for questions Q1, Q2 and Q5, while the DT classifier achieved the best F-measure for Q3 and Q4. We observe a higher F-measure in predicting answers to the questions that suggest stranger friends (Q1 and Q2) than in predicting answers to the questions that suggest abuse (Q3-Q5). This is not surprising, as the mutual activity features are more likely to predict online and real life closeness.

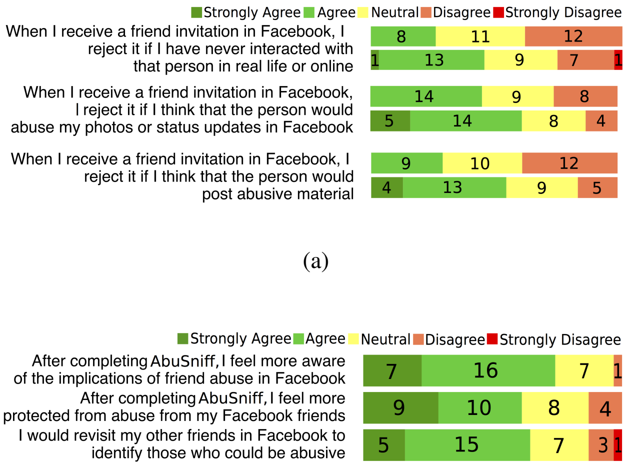
Predicting the user decision. We have evaluated the ability of APM to predict the defense action that the user agrees to implement, according to the 5 possible classes: "unfriend", "restrict", "unfollow", "sandbox", and "ignore". APM achieved the best performance with the RF classifier. Table 3 shows the confusion matrix for APM with RF, over the 10-fold cross validation performed on the 1,452 friend instances. While the overall F-Measure is 73.2%, APM achieved an F-measure of 97.3% when predicting the "ignore" option.

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; same hometown and common study count were the least informative features. We found correlations between the common photo count and mutual post count (Pearson correlation coefficient of 0.65), mutual friend count and mutual photo count (Pearson correlation coefficient of 0.57), and mutual post count and mutual friend count (Pearson correlation coefficient of 0.45). The rest of the features had insignificant positive or negative correlations.

AbuSniff in the Wild

To evaluate the autonomous AbuSniff live, on real users, we have replaced the questionnaire delivery module (QRM) with the abuse prediction module (APM). AbuSniff then asks the user to either accept or ignore the APM predicted defense action, only for the friends for whom the APM predicts that the user will defend against.

We have recruited 49 participants to evaluate their reaction to the predictions of the APM module. We have discarded 9 participants who failed the data quality verification tests previously described. Of the 1,200 friend relationships



(a)

(b)

Figure 6: (a) AbuSniff impact on (I1), (I2) and (I3). For each question, top bar shows pre-test and bottom bar shows post-test results. In the post-test, significantly more participants tend to strongly agree or agree that they would reject new friend invitations based on lack of interaction or perceived timeline or news feed abuse, when compared to the pre-test. (b) Post-test results for (I4), (I5) and (I6).

investigated for the remaining 40 participants (30 friends per participant), APM automatically labeled 403 as potentially abusive. AbuSniff predicted that 359 of these will be approved by the participants, i.e., 41 unfollow, 30 restrict, 137 unfriend and 151 sandbox. 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.

Impact of AbuSniff

In the last 2 user studies we have evaluated the impact of AbuSniff on (1) the willingness of participants to ignore new friend invitations based on their perception of the prospective friend being a stranger or an abuser, on (2) participant awareness of and perception of safety from friend abuse, and (3) their willingness to screen other friends.

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

Further, we have designed a post-study survey that consists of the above 3 items, plus the following 3 Likert items: (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". In a post-test study with a different set of 31 participants, we asked them to first run the questionnaire based AbuSniff version, then answer the post-study survey.

Figure 6(a) compares the user responses in the pre-test (top) and post-test (bottom) for each of the first 3 Likert items. In the pre-test, 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, significantly more participants either strongly agree or agree on all 3 items.

Figure 6(b) shows the participant responses to only the 3 new post-test Likert items. 23 out of 31 participants strongly agree or agree that after running AbuSniff they feel more aware of the implications of friend abuse; only 1 disagreed. 19 participants strongly agree or agree that after running AbuSniff they feel more protected from friend abuse; 4 participants disagree. 20 participants strongly agree or agree that they would revisit their other friends after running AbuSniff, and only 3 disagree, 1 strongly disagrees.

Discussion and Limitations

AbuSniff differs from prior work on cyber abuse and victimization, e.g., (Wolford-Cleveng et al. 2016; Landoll, La Greca, and Lai 2013) in that we (1) focus on specific types of abuse perpetrated through Facebook, i.e., timeline and news-feed abuse, and stranger friends, (2) investigate abuse perception from individual friends and not general exposure, (3) seek to automatically detect abuse perception and (4) provide a first line of suitable defenses against abuse for Facebook users who are unlikely to know and trust all their friends. We performed studies with 263 participants from 25 countries and 6 continents. We acknowledge a common crowdsourcing worker background between participants.

AbuSniff reduces the *attack surface* of its users, by reducing the number of, or isolating friends predicted to be perceived as potential attack vectors. AbuSniff can reduce the audience that needs to be considered by audience selector solutions, e.g., (Raber, Luca, and Graus 2016), and can be used in conjunction with tools that monitor social networking events (Dinakar et al. 2012; Per 2017).

We expect AbuSniff to have more impact for users who have significantly more than 150 friends, the maximum number of meaningful friend relationships that humans can manage (Dunbar 1992). We note that false positives, while being a nuisance, can be fixed by reinstating removed or restricted friends. However, false negatives (keeping abusive and stranger friends) can harm the user and even influence the outcome of elections (BBC 2017a; 2017b).

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

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. The choice of features was needed 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 (Per 2017) can improve APM's prediction performance. We observe that AbuSniff had an F-Measure of 97.3% when predicting the "ignore" action.

Keeping friends perceived to be abusive. In the first study, for 11 of the 68 unfriended friend cases, the participants believed that our warning was correct, but still wanted to keep those friends. One reason may be that the participant had reasons to make him or her abusive toward that friend. We leave this investigation for future work, but note that AbuSniff may protect the friends if they installed AbuSniff.

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. More studies are needed to find the optimal number of evaluated friends per participant, and whether it should be a function of the participant background, e.g., friend count, age, gender.

Related Work

The features provided by online services are known to influence abuse and generate negative socio-psychological effects (Singh et al. 2017). Social networks in particular enable a diverse set of abusive behaviors, that include the adversarial collection and abuse of private information (BBC 2017a; Yates 2017; Yang and Srinivasan 2014; Nosko, Wood, and Molema 2010; Gao et al. 2010), cyberbullying (Wolford-Cleveng et al. 2016; Landoll, La Greca, and Lai 2013; Elphinston and Noller 2011; Quercia, Bodaghi, and Crowcroft 2012), and the viral distribution of fake news, misinformation, propaganda and malware (BBC 2017a; Lee 2017; Sindelar 2014; Aro 2016; Weimann 2010; Al-Shishani 2010).

(Cao et al. 2015) 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. 2015) utilized posting relations between users and messages to combine social spammer and spam message detection. (Quercia and Hailes 2010) maintain information about friendly and suspicious devices that the user encounters in time, to decide if the user is the target of a friend spam attack. AbuSniff focuses on the user perception of strangers friends, their automatic detection and defenses.

(Wolford-Cleveng et al. 2016) have used the Partner Cyber Abuse Questionnaire and found a prevalence of 40% of victimization by cyber abuse among college students in dating relationships, with no differences in victimization of men and women. (Landoll, La Greca, and Lai 2013) developed the Social Networking-Peer Experiences Questionnaire (SN-PEQ) and used it to study cyber victimization in adolescents and young adults. They found that negative social networking experiences were associated with symptoms of social anxiety and depression. (Elphinston and Noller 2011) developed an 8 item questionnaire to explore the impact of involvement with Facebook on relationship satisfaction and found that Facebook intrusion was linked to relationship dissatisfaction via experiences of cognitive jealousy and surveillance behaviors. AbuSniff can help detect and protect against such behaviors.

(Quercia, Bodaghi, and Crowcroft 2012) found that the reasons for ending friend relations are similar in the real and online worlds, and conjectured that tools can be built to monitor online relations. To detect cyberbullying, (Dinakar et al. 2012) used datasets of manually annotated comments, NLP features, supervised learning, and reasoning technique, then proposed several intervention designs. (Ashktorab and Vitak 2016) developed and conducted participatory design sessions with teenage participants to design, improve, and evaluate prototypes that address cyberbullying scenarios. (Narwal et al. 2017) introduced an automated Twitter assistant that identifies text and visual bias, aggregates and presents evidence of bias to users, and enable activists to inform the public of bias, through bots.

(Kwak, Chun, and Moon 2011) found that in Twitter, Korean users tended to unfollow people who posted many tweets per time unit, about uninteresting topics, including details of their lives. (Xu et al. 2013) found that unfollow tends to be reciprocal. This relatively harmless tit-for-tat behavior may explain the willingness of participants in our studies to unfollow abusive friends. (Kwak, Moon, and Lee 2012) found that users who receive acknowledgments from others are less likely to unfollow them. Future work may compare the willingness of a user to unfollow Facebook friends who posted general abuse vs. abuse personally targeted to the user.

Conclusions

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

Acknowledgments

We thank Mozhgan Azimpourkivi and Debra Davis for early discussions. This research was gracefully supported by NSF grants SES-1450619, CNS-1527153, and CNS-1526494 and by the Florida Center for Cybersecurity.

References

Al-Shishani, M. B. 2010. Taking al-qaeda's jihad to facebook. *The Jamestown Foundation: Terrorism Monitor* 8(5):3.

Angwin, J., and Grassegger, H. 2017. Facebooks secret censorship rules protect white men from hate speech but not black children. [CNBC Tech] tinyurl.com/y7ncjgqx.

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

Ashktorab, Z., and Vitak, J. 2016. Designing cyberbullying mitigation and prevention solutions through participatory design with teenagers. In *Proceedings of CHI*.

BBC. 2017a. Russia-linked posts ŕeached 126m facebook users in us.́ [BBC News] tinyurl.com/y93ylwdw.

BBC. 2017b. Theresa May accuses Vladimir Putin of election meddling. [BBC Politics] tinyurl.com/y8d2pwmy.

Cao, Q.; Sirivianos, M.; Yang, X.; ; and Munagala, K. 2015. Combating Friend Spam Using Social Rejections. In *Proceedings of the IEEE ICDCS*.

Cheng, J.; Bernstein, M. S.; Danescu-Niculescu-Mizil, C.; and Leskovec, J. 2017. Anyone can become a troll: Causes of trolling behavior in online discussions. In *ACM CSCW*.

Dinakar, K.; Jones, B.; Havasi, C.; Lieberman, H.; and Picard, R. 2012. Common sense reasoning for detection, prevention, and mitigation of cyberbullying. *ACM TiiS*.

Dunbar, R. I. 1992. Neocortex size as a constraint on group size in primates. *Journal of human evolution* 22(6):469–493.

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

Gao, H.; Hu, J.; Wilson, C.; Li, Z.; Chen, Y.; and Zhao, B. Y. 2010. Detecting and characterizing social spam campaigns. In *Proceedings of the ACM IMC*.

Heath, A. 2017. Facebook quietly updated two key numbers about its user base. [Business Insider] tinyurl.com/y76s8rvs.

2017. JobBoy. http://www.jobboy.com/.

Kontaxis, G.; Polakis, I.; Ioannidis, S.; and Markatos, E. P. 2011. Detecting social network profile cloning. In *PercomW*.

Kwak, H.; Chun, H.; and Moon, S. 2011. Fragile Online Relationship: A First Look at Unfollow Dynamics in Twitter. In *Proceedings of ACM CHI*, 1091–1100.

Kwak, H.; Moon, S. B.; and Lee, W. 2012. More of a receiver than a giver: why do people unfollow in Twitter? In *Proceedings of the AAAI ICWSM*.

Kwan, G. C. E., and Skoric, M. M. 2013. Facebook bullying: An extension of battles in school. *Computers in Human Behavior* 29(1):16–25.

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

Lapowsky, I. 2018. Cambridge Analytica Execs Caught Discussing Extorsion and Fake News. [Wired] https://tinyurl.com/yaagbe9h.

Lee, D. 2017. Facebook, Twitter and Google berated by senators on Russia. [BBC Technology] tinyurl.com/ybmd55js.

Madejski, M.; Johnson, M.; and Bellovin, S. M. 2012. A study of privacy settings errors in an online social network. In *Proceedings of PERCOM Workshops*.

Narwal, V.; Salih, M. H.; Lopez, J. A.; Ortega, A.; O'Donovan, J.; Höllerer, T.; and Savage, S. 2017. Automated assistants to identify and prompt action on visual news bias. In *ACM CHI*.

Nosko, A.; Wood, E.; and Molema, S. 2010. All about me: Disclosure in online social networking profiles: The case of facebook. *Computers in Human Behavior* 26(3).

Ortutay, B., and Jesdanun, A. 2018. How Facebook likes could profile voters for manipulation. [ABC News] https://tinyurl.com/yaaf3lws.

2017. What if technology could help improve conversations online? https://www.perspectiveapi.com/.

Quercia, D., and Hailes, S. 2010. Sybil attacks against mobile users: friends and foes to the rescue. In *IEEE INFOCOM*.

Quercia, D.; Bodaghi, M.; and Crowcroft, J. 2012. Loosing "friends" on Facebook. In *ACM WebSci*.

Raber, F.; Luca, A. D.; and Graus, M. 2016. Privacy wedges: Area-based audience selection for social network posts. In *Proceedings of SOUPS*.

Shahani, A. 2016. From Hate Speech To Fake News: The Content Crisis Facing Mark Zuckerberg. [NPR] tinyurl.com/ycxmke8r.

Sindelar, D. 2014. The kremlin's troll army. *The Atlantic* 12.

Singh, V. K.; Radford, M. L.; Huang, Q.; and Furrer, S. 2017. "They basically like destroyed the school one day": On Newer App Features and Cyberbullying in Schools. In *ACM CSCW*.

Smith, M. 2008. Facebook Abuse: Is Blocking People Enough? tinyurl.com/yajt93gh.

Varol, O.; Ferrara, E.; Davis, C. A.; Menczer, F.; and Flammini, A. 2017. Online human-bot interactions: Detection, estimation, and characterization. In *Proceedings of the AAAI ICWSM*.

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

Wakefield, J. 2017. Facebook and Twitter could face 'online abuse levy'. [BBC Technology] tinyurl.com/ycob8538.

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

2017. Weka. tinyurl.com/36z952.

Wolford-Cleveng, C.; Zapor, H.; Brasfield, H.; Febres, J.; Elmquist, J.; Brem, M.; Shorey, R. C.; and Stuart, G. L. 2016. An examination of the partner cyber abuse questionnaire in a college student sample. *Psychology of violence* 6(1):156.

Wu, F.; Shu, J.; Huang, Y.; and Yuan, Z. 2015. Social spammer and spam message co-detection in microblogging with social context regularization. In *Proceedings of the ACM CIKM*.

Xu, B.; Huang, Y.; Kwak, H.; and Contractor, N. 2013. Structures of broken ties: Exploring unfollow behavior on twitter. In *Proceedings of ACM CSCW*, 871–876.

Yang, C., and Srinivasan, P. 2014. Translating surveys to surveillance on social media: methodological challenges & solutions. In *Proceedings of WebSci*.

Yates, J. 2017. From temptation to sextortion Inside the fake Facebook profile industry. [Radio Canada] tinyurl.com/ycf5md7t.