"Help, I've Been Hacked!": Insights from a Corpus of User-Reported Cyber Victimization Cases on Twitter

Margaret Gratian[1], Darshan Bhansali[1], Michel Cukier[1], and Josiah Dykstra[2]
[1]University of Maryland, College Park
{mgratian, dbhansali, mcukier}@umd.edu
[2]U.S Department of Defense
jdykstra@ltsnet.net

Understanding users is critical to develop secure IT systems and effective cybersecurity solutions. Unfortunately, research that analyzes user cybersecurity behaviors and experiences is limited. Motivated by work in public health that relies on Twitter user-reported health conditions, we explore using Twitter to understand user victimization experiences. Observing that users often self-report victimization (Help, I've been hacked!), we construct a corpus of 2,910 Tweets representing user reports of victimization experiences. We label these reports with information about the affected device or account and the associated consequences, as expressed by the users themselves. We begin to uncover trends in compromise in addition to startling attitudes and behaviors that security practitioners must contend with when developing cybersecurity solutions. To the best of our knowledge, this is one of the first papers to construct and analyze a dataset of cyber victimization user reports.

INTRODUCTION

Understanding how users think, behave, and interact with technology is critical to develop secure IT systems and effective cybersecurity solutions for their operation. For example, studies to evaluate users have found correlations between demographics and a user's likelihood of clicking phishing links, visiting malicious websites, and sharing passwords (Sheng, Holbrook, Kumaraguru, Cranor, & Downs, 2010; Parrish, Bailey, & Courtney, 2009; Darwish, El Zarka, & Aloul, 2012; Mohebzada, El Zarka, Bhojani, & Darwish, 2012; Gratian, Bandi, Cukier, Dykstra, & Ginther, 2018; Levesque, Nsiempba, Fernandez, Chiasson, & Somayaji, 2013).

Most user-focused studies are conducted through surveys or controlled-experiments. Though the value of this work is indisputable, surveys and controlled experiments can be resource intensive, limited in the breadth of a population they cover, prone to response bias, and sometimes unable to capture authentic or unexpected user behaviors (Coppersmith, Dredze, & Harman, 2014; Eichstaedt et al., 2015).

Additionally, due to the pace at which technology platforms and cyber threats emerge, there is a need for new studies and sources of data to understand users and their interactions with technology. How do we broaden our understanding of user cybersecurity? How do we enable rapid discovery of unknown unknowns? This paper explores opportunities to answer these questions using social media data.

Twitter data is already used in the field of cybersecurity to monitor emerging threats. For example, (Sabottke, Suciu, & Dumitras, 2015) presented one of the first Twitter-based methods for early detection of software exploits by monitoring Twitter for the keyword "CVE." Similarly, (Mittal, Kumar Das, Mulwad, Joshi, & Finin, 2016) developed a framework to issue vulnerability alerts to users by searching Twitter for keywords such as "XSS," "spoofing," and "buffer."

Researchers in the field of public health have also turned to Twitter, relying on user-reported information to study health conditions in populations. For example, Achrekar et al. (2011) tracked the spread of influenza via user reports on Twitter of the flu ("I got flu*"*) and obtained a 0.985 Pearson Correlation Coefficient with the U.S. Centers for Disease Control's flu predictions. Golbeck (2018) identified users beginning an Alcoholics Anonymous program ("first AA meeting") and identified users who either maintained sobriety ("I've officially been sober for 4 months") or relapsed ("Taking 5 shots of vodka after I left work tonight was not a good idea"). Using these data, the author predicted alcoholism recovery rate to be 22.8%, a result comparable to the U.S. National Institutes of Health's prediction of 18.2%. Finally, Coppersmith et al. (2014) identified user self-reported mental health issues ("I've been diagnosed with depression") and uncovered previously unknown linguistic signals of mental health disorders.

Motivated by these studies, we investigate the feasibility of using Twitter to understand user cybersecurity experiences. We hypothesize that user-reported cyber victimization experiences will provide an opportunity to develop new insights into users and ultimately may enable active discovery of the threats facing users. We collect 2,910 cases of user self-reports of cyber victimization (*Help, I've been hacked!*) during the period of January 1, 2018 through March 13, 2018. We identify the devices, accounts, and consequences associated with victimization, as explained by users themselves. We also identify users who have been victimized repeatedly. Initial analysis reveals Twitter is a key source to understand user cybersecurity experiences: users openly and explicitly Tweet about everything from illegal downloading habits to compromised corporate infrastructures.

To the best of our knowledge, only one other paper has explored user-reported cases of victimization. Zangerle et al. (2014) collected Tweets containing the keywords '"hacked" or "compromised" and "account"' and trained a Support Vector Machine to classify Tweets with different types of user responses to victimization. Our work is distinct for two reasons. First, we conduct a thorough manual review process to validate

the authenticity of user-reports. Second, our scope is much broader, as we explore a variety of different user experiences on multiple online platforms; Zangerle et al. only retain victimization cases on Twitter and their analysis is limited to identifying the percent of users who create a new account in response to victimization and the percent of users who either apologize or state they have been hacked.

The contributions of our work are as follows:

- We present one of the first papers that uses Twitter as a source of unsolicited user feedback on user cybersecurity experiences and illustrate the viability of using these data to understand users.
- We uncover victimization trends and user attitudes that can help cybersecurity and human factors practitioners understand the factors they must contend with when developing secure IT systems and cybersecurity policies.

METHOD

Data Collection

To construct the dataset, we followed a similar methodology to the work discussed in the Introduction, searching for victimization-related keywords across all public, English Tweets over the period of January 1, 2018 through March 13, 2018. We used both the Twitter search API and the search page on Twitter's site. We incrementally constructed our dataset using a Grounded Theory Sampling approach (Foley & Timonen, 2015). We identified a few Tweets that demonstrated self-reports of victimization, (e.g., *Don't click the link, my account was hacked*), and gradually identified new cases and keywords to search as we learned more about disclosure habits.

Data Cleaning

Two reviewers manually reviewed all Tweets returned by the search and either labeled a result as a true case of victimization or discarded it. We only retained Tweets if they provided some proof of victimization, (e.g., through a written description of the incident, requests for company or customer support assistance, or screenshots of the hacked account). We only retained Tweets if they were self-reports of victimization.

Initial data collection and cleaning resulted in 3,398 users. Since determining if a Tweet represented a true case of victimization was a subjective decision, this set of Tweets was then split across four reviewers, who again reviewed each Tweet and made an independent assessment of its authenticity as a true case of victimization.

If a reviewer had any doubt about the authenticity of the self-report, either due to a lack of specifics about the event, grammatical errors making the Tweet difficult to understand (a common problem), contradictions or discrepancies created by the user when explaining the problem, or the context surrounding the Tweet when viewed directly on Twitter's site, the Tweet was discarded from the sample. While general approaches to reviewing the Tweets were discussed among the reviewers, each reviewer made a determination about whether to keep or discard a Tweet independently. This process left us with 2,910 Tweets we deemed authentic. This was intended as a feasibility study, but in future work we will expand the review process to include adding reviewers and computing inter-rater reliability.

Data Labeling

Since the goal of our study was to investigate the feasibility of using Twitter to understand user cybersecurity experiences, our analysis focused on answering three high level questions: *What device or account was affected, what were the associated consequences,* and *has this user been victimized before?* To answer these questions, reviewers were each given a fourth of the corpus to label with account, device, consequence, and repeat incident details, if applicable. All initial labels were verified by two reviewers.

Device. If a user specifically named a compromised device type in their self-report of victimization, reviewers added a label indicating the device type. Device labeling resulted in six labels: 1) computer, which included desktops and laptops, 2) phone, 3) USB, 4) iPad, 5) iPod, and 6) memory card.

Account. Similarly, for the 'affected account' category, reviewers added details about the compromised account if the user explicitly named it. Account labeling resulted in 223 unique accounts. To help us understand the nature of these 223 accounts, we also sorted these labels into 17 high-level categories, which included social media and chat applications (e.g., Facebook, Twitter), online games (e.g., Roblox, Fortnite), media and entertainment (e.g., Spotify, Netflix), retail (e.g., eBay), and so on. Table I contains the full list of categories.

Consequences. As before, reviewers assigned labels only if users made explicit comments about a consequence. For example, *@AskPlaystation my account's been hacked will you please help they changed my email :(* was labeled with 'altered settings,' while *I spent $100 to fix my computer because I accidentally downloaded a virus* was labeled with 'financial loss.' This process resulted in 13 consequence categories listed in Table II. Tweets were labeled with multiple consequences if multiple consequences were reported.

Prior Victimization. Users were labeled as 'repeat victims' if they made a comment about being victimized 'again' or indicated the number of times they've been victimized.

RESULTS

Affected Devices

Of our 2,910 cases, 472 specified an affected device. The majority of users experienced victimization on a computer (277 total) or phone (189 total). Two users experienced a compromised USB; two a compromised iPad; one a compromised iPod; and one a compromised memory card.

Affected Accounts

Table I presents the 17 high-level categories that accounts were grouped into, their frequency of occurrence, and what percent of the total number of reported affected accounts they

represent. Table III presents an overview of all the affected accounts that were reported by users with a frequency greater than 10.

Social media and chat applications were the most commonly occurring type of account for which users self-reported compromise. It is unclear if this is because users are simply more likely to discuss social media victimization on social media or because social media sites actually experience a higher volume of user exploitation. Both are valid possibilities. Research has shown that the click through rate for phishing links is higher on social media than in email and that 8% of all URLs posted on Twitter are actually spam, phishing, or malicious links (Grier, Thomas, Paxson, & Zhang, 2010). And since users likely share similar networks of friends on different services, it makes sense they would post about compromise experienced on one site on another site: *Hello friends! My GroupMe was hacked so don't click anything that was sent.*

Looking at specific accounts, Twitter had the highest number of cases. Of the 382 Twitter cases, 102 were posts reaching out to Twitter (e.g., *@Twitter why is my account hacked pls help*) and 56 were posts reaching out to Twitter Support (e.g., *@twittersupport I've been hacked, how can you help me?*). Though almost half of all Twitter cases attempted to reach out to Twitter for help, many did not seem optimistic about the possibility of recovering an account or receiving help from Twitter: *Friends: someone seems to have hacked my account so I'm using my secondary once since I have very little faith in Twitter's interest in solving my problem.*

The next highest occurring account type was Roblox, a massive, multiplayer, online gaming platform. Mainly marketed towards children and teenagers, the game has over 64 million active players per month, as of November 2017 (Bort, 2017). Similar to Twitter, we observed many users reaching out to Roblox or to well-known game moderators and developers for help: *@Roblox help please someone hacked my account.*

Some account types were surprising: Dominos, an American pizza chain, occurred with a frequency of 6; users reported hacked accounts resulting in unauthorized pizza orders and use of account rewards points and free pizza offers.

We were also able to observe a common attack vector against users during our time window of data collection. For example, of the 76 reported cases of compromised GroupMe accounts, 72 users reported their accounts were used to send "Happy New Year" messages containing links to weight loss pills or requests for bitcoin. This emphasizes the opportunity to use Twitter for real-time detection and analysis of the threats. If security operations center personnel are actively monitoring user reports, seeing posts such as this could prompt blacklisting of malicious applications or websites before others are affected.

Consequences

Of our 2,910 cases, 1,884 specified a total of 2,220 consequences. Table II presents an overview of the major categories of consequences reported by users and the frequency and percentage with which they occurred. Loss of account access was the most frequently reported: *@Microsoft my account was suspended since it was hacked and now I am locked out of all my services.* Altered setting was the next most frequently reported consequence: *@DavorCoinHELP somebody hacked my account and turned on two factor authentication.* Spam – *My Twitter was hacked don't try to buy fake Ray Bans from the link I sent* – and financial loss – *@AirBnBHelp my account was hacked for $1800 in fraudulent charges* – were also frequently reported.

When we examined how consequences related to account types, we found that social media victims experienced every type of consequence except device damage, with 30% reporting a social media account was used to send spam and 21% reporting they lost access to their account. For online games, 29% of users experienced account loss, 21% experienced data loss, usually in the form of in-game items, and 20% experienced altered settings on their accounts, such as username or password changes. Thirty eight percent of those who reported a compromised media or entertainment account experienced altered settings and 32% reported account loss. Nine percent observed unauthorized activity on their account, which often entailed the attacker using the entertainment service: *whoever hacked into my @Spotify account, your music taste is awful. I wish @SpotifyCares would let me see your IP address or login info.* For victimization reported regarding technology and electronics companies, almost all consequence categories were reported except reputational loss and professional issues. Sixty seven percent of users who reported victimization involving cryptocurrency experienced financial loss.

We also observed nine users who reported victimization that led to professional issues varying from compromised business accounts to damaged workplace infrastructures: *Apparently my computer had a virus so when I connected it to the WIFI at work it wrecked the whole server.*

Repeat Victims

In our sample, 99 users reported repeat victimization. Some users made generic comments about repeat victimization: *To everyone who got a DM from me with a Instagram link, ignore. I got hacked again.* But others revealed that victimization was a common occurrence: *[Facebook] is hacked daily...@facebook you need to do something about this. This is the 13th time this year...*

Criminology research suggests victimization is concentrated on a few repeated cases and prior victimization is often an indicator of enhanced risk of future victimization (Polvi, Looman, Humphries, & Pease, 1991). About 3% of our sample reported repeat victimization; the number is likely higher as not all users can be expected to self-report victimization. Regardless, our data may be useful to evaluate hypotheses on repeat victimization in the cyber domain.

Additional Observations

Data collection provided insight into the poor security decisions and risks some users choose to take online. Users are regularly reminded not to share passwords and yet our data collection phase revealed that these practices still occur. For example, we observed many users who claimed to have "been hacked," when in reality a friend of the user had temporary

access to the user's account or device and posted as the user. Many of these occurrences appeared to be good-natured and were viewed as humorous by the user: *lol I left my account logged in and @*<username> *hacked me!* However, some users experienced the more sinister consequences of shared accounts and passwords: *@AirennorGAMES Hi Airennor I was hacked by 18 people because someone said my password in one of their videos and I lost all my things...*

User reports may also be useful for understanding how and why users fall for victimization, as we observed several phishing victims who seemed to recall when they were compromised: *I received a DM from a trustworthy account with a link. I clicked it but it was some sort of hack.*

We observed some users who not only rejected common security advice but also actively chose to risk viruses or hijacked accounts. Typically, this behavior was reported in relation to accessing online content freely or illegally: *99% sure I gave my computer a virus trying to watch the show from a sketchy site*. Even more troubling is we observed several users who demonstrated this attitude while at work: *Got a virus on my computer at work trying to click a link to take a quiz...how was I supposed to resist that clickbait though?*

Users willing to risk illegal downloading sometimes expressed amusement at the outcome: *I tried to download an apk file to get Spotify for free and now I have a phone virus LMAO.* While it was sometimes difficult to tell if humor was being used in a self-report sarcastically, seriously, or as a coping mechanism, it was nonetheless expressed quite often: *Hahaha I haven't been able to access my Twitter since early Feb because I was hacked...*

Many users expressed attitudes we expected to see, such as fear and distress: *My computer was hacked by a virus...I miss everything and I'm scared and depressed now.* Some even reported having nightmares about viruses: *I've been having the same nightmare over and over where my computer gets a horrible virus and does something personal and its creeping me out.* Note that we did not perform automated sentiment analysis to assess emotional responses; this is considered an area of future work.

Online gamers in particular expressed desperation to recover hijacked accounts, which may be due to the fact that the population skews younger than the population on other services and many people take their online games quite seriously. This was often evident in game-victimization related posts: *Roblox I have been hacked please help me...I am crying so much please help me....* Desperation was also evident in the content users were willing to share to recover an account: *My account has been hacked and they changed my phone number so I can't get back in please help. My username was:* <username> *and my password was*: <password>. For the game Roblox, we observed a total of nine users who publicly shared their usernames and passwords on Twitter when requesting help for hacked accounts.

During data collection, we also observed misconceptions about cybersecurity propagated on Twitter. We observed several users who seemed to believe that avoiding viruses simply requires common sense: *If you need anti-virus software you probably shouldn't be using a computer.* Others still believe that it is not possible to get viruses on phones or Apple devices: *iPhone is impervious to viruses.*

Lastly, our dataset hinted at problems in how account recovery procedures are designed and implemented: *@LinkedInHelp I can't follow up on the case I made because obviously I don't have access to my hacked account. The link you sent when I created the ticket about my hacked account requires log in. VERY SMART YOU GUYS.* We observed many users expressing frustration at the lack of real customer service support: *@instagram my account's been hacked twice in 24 hours. I've tried to call but your voicemail clearly says you don't speak to people. I already changed my password 3 times. Any help?* Others expressed frustration at how long it takes to recover an account: *My Facebook account was hacked January 1st...almost 1.5 months passed and I didn't get any solution.* While most services received criticism for how they handle compromise, one company regularly received praise: *my @Netflix account was hacked. The nice guy from the help desk took less than two minutes to recover it. Great service!* ☺

DISCUSSION

There is a wealth of information contained in self-reports of users. A six-week period of data collection provided us with hundreds of examples of victimization. Our data provided insight into the accounts where users most frequently report compromise, hinting at the possibility of communities or platforms that are especially vulnerable. Our data also provided insight into how users react to compromise and the resulting consequences. In turn, user reactions to compromise alerted us to flaws in how fraud/abuse detection and account recovery are handled on various platforms. Understanding the accounts that experience elevated levels of compromise and the experiences and reactions of users can help security researchers and developers chose how to focus limited resources. Lastly, our work revealed the possibility of using Twitter data to monitor emerging threats, similar to work that monitors Twitter for new software vulnerabilities and exploits.

There are certainly some challenges and limitations associated with using Twitter data. While we had multiple reviewers assess the user reports, there may still be subjectivity in the final dataset. Additionally, we do not know how representative our dataset is of the general population. However, real victimization data are generally difficult to obtain, as organizations keep a close hold on such information for privacy or security reasons. Though there are certainly still privacy concerns associated with using social media data to study users' cybersecurity experiences, turning to Twitter to understand users expands the data available to security researchers and is worthwhile as long as the benefits to users outweigh the risks.

Ultimately, we hope our work motivates interest in the security and human factors communities to explore unsolicited user feedback on user cybersecurity experiences. In future work, we plan to conduct a deeper analysis of user reports and explore techniques to automate data collection using insights from this study into user-reported experiences.

ACKNOWLEDGEMENTS

The authors acknowledge the support of the U.S. Department of Defense under contract H98230-19-D-0003. The views and conclusions expressed in this paper are those of the authors, and do not necessarily represent those of the Department of Defense or U.S. Federal Government.

REFERENCES

Sheng, S., Holbrook, M., Kumaraguru, P., Cranor, L. F., & Downs, J. (2010, April). Who falls for phish?: a demographic analysis of phishing susceptibility and effectiveness of interventions. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 373-382). ACM.

Parrish Jr, J. L., Bailey, J. L., & Courtney, J. F. (2009). A personality based model for determining susceptibility to phishing attacks. *Little Rock: University of Arkansas*, 285-296.

Darwish, A., El Zarka, A., & Aloul, F. (2012, December). Towards understanding phishing victims' profile. In *2012 International Conference on Computer Systems and Industrial Informatics* (pp. 1-5). IEEE.

Mohebzada, J. G., El Zarka, A., BHojani, A. H., & Darwish, A. (2012, March). Phishing in a university community: Two large scale phishing experiments. In *2012 International Conference on Innovations in Information Technology (IIT)* (pp. 249-254). IEEE.

.Gratian, M., Bandi, S., Cukier, M., Dykstra, J., & Ginther, A. (2018). Correlating human traits and cyber security behavior intentions. *Computers & Security*, *73*, 345-358.

Lalonde Levesque, F., Nsiempba, J., Fernandez, J. M., Chiasson, S., & Somayaji, A. (2013, November). A clinical study of risk factors related to malware infections. In *Proceedings of the 2013 ACM SIGSAC conference on Computer & communications security* (pp. 97-108). ACM.

Coppersmith, G., Dredze, M., & Harman, C. (2014). Quantifying mental health signals in Twitter. In *Proceedings of the workshop on computational linguistics and clinical psychology: From linguistic signal to clinical reality* (pp. 51-60).

Eichstaedt, J. C., Schwartz, H. A., Kern, M. L., Park, G., Labarthe, D. R., Merchant, R. M., & Weeg, C. (2015). Psychological language on Twitter predicts county-level heart disease mortality. *Psychological science*, *26*(2), 159-169.

Sabottke, Carl, Suciu, Octavian & Dumitras, Tudor. (2015). Vulnerability Disclosure in the Age of Social Media: Exploiting Twitter for Predicting Real-World Exploits. In *Proceedings of the 24th USENIX Security Symposium (*pp. 1041-1056). USENIX.

Mittal, Sudip, Kumar Das, Prajit, Mulwad, Varish, Joshi, Anupam, & Finin, Tim. (2016). Cybertwitter: Using twitter to generate alerts for cybersecurity threats and vulnerabilities. In *Proceedings of the 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*. (pp. 860-867). IEEE.

Achrekar, H., Gandhe, A., Lazarus, R., Yu, S. H., & Liu, B. (2011, April). Predicting flu trends using twitter data. In *2011 IEEE conference on computer communications workshops (INFOCOM WKSHPS)* (pp. 702-707). IEEE.

Golbeck, J. (2018, July). Predicting Alcoholism Recovery from Twitter. In International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction and Behavior Representation in Modeling and Simulation (pp. 243-252). Springer, Cham.

Zangerle, E., & Specht, G. (2014, March). Sorry, I was hacked: a classification of compromised twitter accounts. In *Proceedings of the 29th Annual ACM Symposium on Applied Computing* (pp. 587-593). ACM.

Grier, C., Thomas, K., Paxson, V., & Zhang, M. (2010, October). @ spam: the underground on 140 characters or less. In *Proceedings of the 17th ACM conference on Computer and communications security* (pp. 27-37). ACM.

Foley, G., & Timonen, V. (2015). Using grounded theory method to capture and analyze health care experiences. *Health services research*, *50*(4), 1195-1210.

Bort, J. (2017, November*). 50 startups that will boom in 2018, according to VCs.* Retrieved from http://www.businessinsider.com/50-startups-to-boom-in-2018-according-to-vcs-2017-11.

Polvi, N., Looman, T., Humphries, C., & Pease, K. (1991). The time course of repeat burglary victimization. *The British Journal of Criminology*, *31*(4), 411-414.

TABLES

TABLE I. CATEGORIES OF AFFECTED ACCOUNTS

Category	Frequency	Percent of Total
Social media and chat	992	42.6
Online games	743	31.6
Media and entertainment	130	5.5
Technology and electronics	118	5.0
Cryptocurrency	72	3.1
Email	62	2.6
Ride sharing	45	1.9
Payment and money transfer	26	1.1
Traditional banks and financial services	40	1.7
E-commerce, delivery, and retail	39	1.7
Housing and hospitality	31	1.3
Website and file hosting	13	0.6
Online dating	5	0.2
Miscellaneous	5	0.2
Airlines	4	0.2
School systems	3	0.1
Online betting	3	0.1

TABLE II. CONSEQUENCES OF COMPROMISE

Consequence	Frequency	Percent of total
Account loss	592	26.7
Altered settings	406	18.3
Spam	359	16.2
Financial loss	267	12.0
Data loss	247	11.1
Unauthorized account activity	87	3.9
Access vector	65	2.9
Reputation loss	61	2.7
Device damage	59	2.7
Professional issues	29	1.3
Identity theft	22	1.0
Threats	17	0.8
Miscellaneous	9	0.4

TABLE III. AFFECTED ACCOUNT OVERVIEW

Service/ Platform Name	Frequency	Service/ Platform Name	Frequency
Twitter	382	Netflix	30
Roblox	258	Steam	27
Instagram	177	Amazon	27
PlayStation	161	Airbnb	26
Facebook	160	Mojang	25
Spotify	90	Gmail	23
GroupMe	76	Discord	21
Jagex	73	Apple	21
Snapchat	70	Youtube	20
EA	60	PayPal	18
Uber	44	eBay	15
Rockstar Games	36	Xbox	14
Email	35	Bizzard	13
Linkedin	34	Binance	12
Yahoo	33	Google	11