Collective Obfuscation and Crowdsourcing

Benjamin Laufer [1] Niko A. Grupen [1]

Abstract

Crowdsourcing technologies rely on groups of people to input information that may be critical for decision-making. This work examines obfuscation in the context of reporting technologies. We show that widespread use of reporting platforms comes with unique security and privacy implications, and introduce a threat model and corresponding taxonomy to outline some of the many attack vectors in this space. We then perform an empirical analysis of a dataset of call logs from a controversial, real-world reporting hotline and identify coordinated obfuscation strategies that are intended to hinder the platform's legitimacy. We propose a variety of statistical measures to quantify the strength of this obfuscation strategy with respect to the structural and semantic characteristics of the reporting attacks in our dataset.

1. Introduction

Crowdsourcing is a method of information retrieval that relies on the public to supply information to decision-makers including state authorities. In the public reporting domain, even established systems, such as emergency reporting (e.g. 911) and information hotlines (e.g. 311) are not exempt from problems of distrust (Rock, 2019; Sasson et al., 2015; Kessell et al., 2009; Smith & Holmes, 2003; Clark et al., 2020). It is no surprise then that more controversial platforms—including Victims of Immigration Crime Engagement (VOICE) (Kopan, 2017) and Texas Right to Life (McCammon, 2021)—are subjected to unintended use, spamming attacks, false reports, DDoS attacks, and more, which collectively render their crowdsourced information useless. These attacks, however, differ in nature from standard security and privacy attacks on technology platforms, as they represent a form of obfuscation; defined as "the deliberate addition of ambiguous, confusing, or misleading information to interfere with surveillance and data collection" (Brunton & Nissenbaum, 2013; 2015).

We posit that, unlike studies that characterize security with respect to the intended use of a technology, the reporting domain requires consideration of both harmful use and legitimate misuse (e.g. obfuscation). To this end, we introduce a parallel threat model describing both threats to the platform's legitimate functioning and platform-enabled violence. Inspired by Thomas et al. (2021), we propose a taxonomy of use and abuse in how people interact with reporting platforms, and discuss how legitimate responses to reporting platforms depend significantly on social context.

Finally, we study obfuscation in the context of a real-world collective spamming attack on the VOICE reporting system. Using open-source data from the VOICE system's call logs, we analyze both structural (e.g. report length) and semantic (e.g. sentence embedding distances) properties of spam and non-spam reports. We propose a variety of statistical measures to quantify the strength of an obfuscation strategy with respect to similarity and dissimilarity in the high-dimensional embedding spaces generated by neural network language models. Our analysis reveals a number of interesting insights for spamming techniques on reporting technologies: (i) Spam reports are longer than true reports (intending to waste operator time); (ii) Spam reports are semantically disparate, covering a wider swath of topics and clustering into more disjoint groups; (iii) Spam reports are thematically similar, suggesting that they may be from coordinated Internet campaigns. While many of the spam messages in the VOICE dataset were easily identifiable, future deception campaigns using advanced language models may pose more significant threats because they will more closely resemble the distribution of true reports.

2. Taxonomy of Threats

In this section, we report a taxonomy of threats that are mediated or enabled by reporting platforms. These threats are accumulated through systematically mapping the information flows that constitute reporting platforms (Figure 1).

[*]Equal contribution [1]College of Computing and Information Sciences, Cornell University. Correspondence to: Benjamin Laufer <bdl56@cornell.edu>, Niko Grupen <nag83@cornell.edu>.

Workshop on Disinformation Countermeasures and Machine Learning, International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022. Copyright 2022 by the author(s).

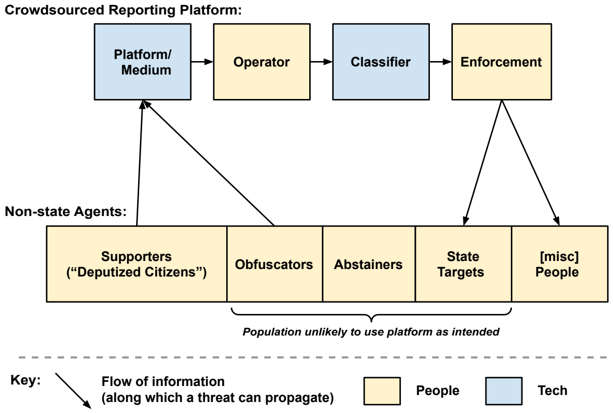
Crowdsourced Reporting Platform:
Platform/ Medium
Operator
Classifier
Enforcement
Non-state Agents:
Supporters ("Deputized Citizens")
Obfuscators
Abstainers
State Targets
[misc] People
Population unlikely to use platform as intended
Key:
Flow of information (along which a threat can propagate)
People
Tech

Figure 1. Diagram representing the flow of information through reporting platforms.

2.1. Deceptive Reporting

Submitting a deceptive report on a reporting platform entails fooling a call operator or spam classifier in order to shift the functionality and resource allocation associated with the platform. These reports may erode the reporting platform's legitimacy and functionality. Deceptive reports fall into two broad categories: false reporting aimed at harming or disempowering another private individual (described in Thomas et al. (2021)) and false reporting aimed at harming or disempowering the reporting platform itself.

2.1.1. Weaponizing Reports

Deceptive reporting to intimidate, disempower or harm typically involves a private-citizen attacker whose goal is to harm a private-citizen target. In a phenomenon known as "SWATing", individuals sometimes use false reports to provoke emergency services to locate and confront a target. If a caller falsely reports a serious crime (for exmaple a bomb threat), use-of-force may lead to fatality (Krebs, 2019). Attackers who weaponize reports have goals ranging from intimidation to coercion to direct physical harm. Depending on the purpose and functionality of a reporting system, weaponized reports can intimidate targets or convey a threat.

2.1.2. Poison Reports

Poison reports aim to harm platforms rather than private individuals. Analogous to cache poisoning (Son & Shmatikov, 2010), poison reports are a particular type of spam that relies on deceiving an operator or spam classifier. By planting false reports that are believed to be true by the platform and its operators, poison data may significantly harm the platform's ability to distinguish true reports (Vincent et al., 2021). These attacks are particularly harmful in large (variegated) quantities, or when a false positive report classification incurs large costs for the platform operator.

2.1.3. Exaggerated Reports

These reports come from people who intend to use the reporting platform for its express purposes, but wish to receive priority. Participants may behave strategically, knowing that their case may be prioritized if they exaggerate the urgency of their report. This can take the form of an emergency declaration, use of inflammatory or urgent words, pleading, or fabricating aspects of an account.

2.2. Abnormal Reporting

Abnormal reports contain information that was not expected or designed to be processed by the platform. These reports need not be deceptive or even malicious, but can be considered "threats" because they harm the functioning of a reporting system. The mechanisms through which they cause harm include creating many meaningless reports that need to be sifted through, wasting phone operators' time, creating backlogs, and generally adding friction to the reporting process. Abnormal reports span a wide number of different categories; we list a few below.

2.2.1. Opinions

Crowdsourced reporting platforms, especially non-emergency platforms, can be inundated with calls from people who simply wish to state sentiments and opinions. For example, people may call to state their approval or disapproval of a platform's existence. In this case, the 'attacker' is an individual user, and the target is the reporting platform and its operators who read or listen to reports.

2.2.2. Trolling

We define trolling as the use of "inflammatory, insincere, digressive, extraneous, or off-topic messages" (Wikipedia) to provoke or manipulate. Troll behavior is distinct from deceptive reporting—although troll reports are often false, they are not deceptive nor do they expect to successfully trick an operator or platform into reallocating resources. Attackers target an operator or platform and might be capable of coordinating large-scale attacks on platforms (Birkbak, 2018; Navarro-Carrillo et al., 2021).

2.2.3. Threats

These reports are often politically-motivated reports by individuals who believe the platform is harmful or should not exist. Their goal is to intimidate, coerce, or threaten the humans who are operating and maintaining the reporting system. Attacks range from accusatory political sentiments (e.g., telling an operator that they should be ashamed of their work) to direct threats (e.g., attempting to use an operator's personal information to intimidate).

2.2.4. Hate Speech and Profanity

The use of corrosive, belligerent, and hateful language can occur on any platform that is not censored. These reports need not be relevant to the platform's intended purpose, and hateful language may not be directed at anyone in particular. There may be no real target or goal (from a security per-spective) associated with these attacks, but they may harm a platform's functioning and so should be taken seriously.

2.2.5. Accidental Calls

Many people call a hotline by accident. Emergency hotlines are inundated with accidental calls, such as 'butt dials' or mistakenly entered area codes or phone numbers (Friedman & Albo, 2017). These calls are not malicious in nature but need to be efficiently triaged to prioritize intended calls.

2.3. Overloading

Overloading attacks take aim at a system's ability to field large numbers of reports or traffic. Overloading attacks specific to reporting platforms include report spamming, raiding or brigading, and distributed denial of service (DDoS). For a general overview of overloading see Thomas et al. (2021).

2.3.1. Report Spam

In this context, we define spamming as friction-generating calls whose purpose is solely to use platform resources. Report spam can include troll reports, but may even include silent calls, music, pre-recorded messages, or operator-operator routing calls (where two operators are put in touch by an intermediary user to sow confusion).

2.3.2. Raiding or Brigading

These are instances where people coordinate to overwhelm a feed, platform, or comment section to target an individual or group. In reporting software, raiding and brigading may take the form of script-based reports, large-scale spam 'dumps', or coordinated trolling.

2.3.3. Distributed Denial of Service (DDoS)

DDoS attacks make a platform useless by jamming communication channels so that nobody can participate in information reporting.

2.4. Information Leakage

Crowdsourced enforcement mechanisms can enable interpersonal harm, even when they are intended by the system's designer. In a number of cases including immigration enforcement, policing, hotlines enable people to report one another and leak private information that could be used to incarcerate, interrogate, and search. Reporting platforms can be a means by which people leak citizenship documentation status, criminal histories, or private sexual and health information—all instances of information leakages.

2.5. Coercion

Depending on the decision motivating a crowdsourced information retrieval system, the platforms may enable coercive leverage between people. 911 systems enable people to call the police on anybody on the street or in their lives. This ability injects state force into interpersonal relationships, which can lead to threats and coercion between people. Somebody may threaten to call the police on somebody else, and this type of threat can be harmful and manipulative.

2.6. Surveillance

Reporting systems turn everyday citizens into deputies whose responsibility is to report information that may be relevant to state law enforcement, regulatory services, or resource allocation. These systems raise two potential threats in the category of surveillance, which we refer to as state surveillance and interpersonal surveillance.

2.6.1. State Surveillance

Reporting platforms can enable state surveillance by expanding the information available to law enforcement. This can provide warrant for searching, spying, and arresting people. Further, anonymity guarantees in a variety of platforms can be violated if law enforcement has a reasonable cause.

2.6.2. Interpersonal Surveillance

Reporting platforms rely on people to observe and provide information to platform operators. In the case of law enforcement, citizens feel they have a responsibility to inform authorities of emergency situations. Citizens may take this responsibility too far, and feel emboldened to surveil neighbors and community members who they deem suspect. The same is true for immigration hotlines, abortion hotlines, and even whistleblower and other crowdsourced information hotlines. Such snooping behavior can be a threat to privacy.

3. Case Study: VOICE Logs

In this section, we analyze a dataset of 5164 publicly-available call logs from the VOICE reporting system in 2017, which exhibits some phenomena described above. We focus on the following questions: (i) What structural qualities (e.g. report length) differentiate spam from non-spam interactions? (ii) Do spam reports cover a wider range of topics than non-spam reports? (iii) Are spam reports semantically similar to other spam reports? We address each of these questions in the subsections below. The structural analysis can be found in Appendix C.

3.1. Semantic Analysis

To capture the semantic content of the reports, we generate sentence-level embeddings for each report using Sentence-BERT (Reimers & Gurevych, 2019). Sentence embeddings map text sentences into a high-dimensional vector space in which semantically similar sentences are close, and unrelated sentences are far apart. Such representations are effective for downstream textual tasks like clustering and semantic search (Conneau & Kiela, 2018). Here, we leverage this high-dimensional semantic space to examine the similarities and dissimilarities of spam vs. non-spam reports by clustering their respective embeddings and measuring text similarity with a variety of metrics that we introduce to examine the underlying report distributions.

3.1.1. Clustering Analysis

First, we perform hierarchical clustering over the set of report embeddings. Formally, given a distribution of report sentences S, we pass each sentence $s_i \in S$ through an embedding-generating function $f_\theta : S \to \mathbb{R}^d$ (with parameters θ), which produces a d-dimensional embedding vector $v_i \in \mathbb{R}^d$. Following the method introduced by Grootendorst (2020), we cluster the embeddings by preojecting each v_i onto a lower-dimensional manifold using UMAP reduction (McInnes et al., 2018), then performing density-based clustering with HDBSCAN (Campello et al., 2013). The optimal clustering yields a set of topics T for the reports, which are shown in Figure 2. From the y-axis of Figure 2, we see a number of obviously spam clusters that correspond to attack vectors from our taxonomy; such as abnormal reporting—e.g. false reports describing UFOs and extraterrestrials (trolling), politically-charged commentary (opinions), and operator harassment (threats, profanity)—and overloading—e.g. staying silent or playing music (report spam). This analysis suggests that a large number of spam calls can be accurately identified by topic alone. We therefore conjecture that a cursory human review of the clusters produced yields spam/non-spam labels that are highly accurate at detecting semantically obvious, non-deceptive spam attacks. We supplement this analysis by clustering spam and non-spam reports separately in Appendix B.

3.1.2. Semantic Similarity Analysis

We also quantify the similarity of spam vs. non-spam reports as a function of distance in semantic space. Given spam embeddings $v^{spam} = \{v_1^{spam}, v_2^{spam}, \ldots, v_n^{spam}\}$, we define the following three distance measures:

Definition 3.1. Within-category distance D_{WC} is the average cosine distance between each spam embedding v_i^{spam} and the mean spam embedding $\bar{v}^{spam}$:

$$D_{WC}(v^{spam}) = \frac{1}{n}\sum_{i}^{n} 1 - \frac{v_i^{spam} \cdot \bar{v}^{spam}}{||v_i^{spam}||_2 ||\bar{v}^{spam}||_2}$$

Intuitively, $D_{WC}(v^{spam})$ measures the extent to which spam reports resemble each other. Due to the varied topics of spam, as outlined in Section 3.1.1, we expect there to be a greater difference between the semantic content of spam than there is for non-spam. We therefore hypothesize that $D_{WC}(v^{spam}) > D_{WC}(v^{non\text{-}spam})$, where $D_{WC}(v^{non\text{-}spam})$ is the average distance for non-spam embeddings $v^{non\text{-}spam}$.

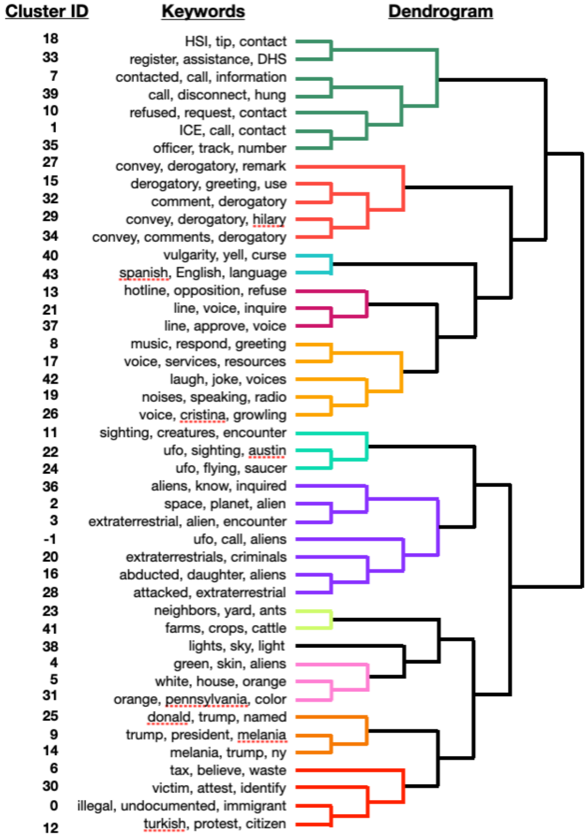
Cluster ID
Keywords
Dendrogram
18 HSI, tip, contact
33 register, assistance, DHS
7 contacted, call, information
39 call, disconnect, hung
10 refused, request, contact
1 ICE, call, contact
35 officer, track, number
27 convey, derogatory, remark
15 derogatory, greeting, use
32 comment, derogatory
29 convey, derogatory, hilary
34 convey, comments, derogatory
40 vulgarity, yell, curse
43 spanish, English, language
13 hotline, opposition, refuse
21 line, voice, inquire
37 line, approve, voice
8 music, respond, greeting
17 voice, services, resources
42 laugh, joke, voices
19 noises, speaking, radio
26 voice, cristina, growling
11 sighting, creatures, encounter
22 ufo, sighting, austin
24 ufo, flying, saucer
36 aliens, know, inquired
2 space, planet, alien
3 extraterrestrial, alien, encounter
-1 ufo, call, aliens
20 extraterrestrials, criminals
16 abducted, daughter, aliens
28 attacked, extraterrestrial
23 neighbors, yard, ants
41 farms, crops, cattle
38 lights, sky, light
4 green, skin, aliens
5 white, house, orange
31 orange, pennsylvania, color
25 donald, trump, named
9 trump, president, melania
14 melania, trump, ny
6 tax, believe, waste
30 victim, attest, identify
0 illegal, undocumented, immigrant
12 turkish, protest, citizen

Figure 2. Hierarchical clustering of VOICE logs. The first seven clusters (top, colored green) appear to be true reports, whereas many of the others fall into threat categories including opinions (e.g. cluster 13) and obvious spam (e.g. references to UFO's).

Definition 3.2. Distance from the mean non-spam report D_{NSR} is the average cosine distance between each spam embedding v_i^{spam} and the mean non-spam embedding $\bar{v}^{non\text{-}spam}$:

$$D_{NSR}(v^{spam}) = \frac{1}{n}\sum_{i}^{n} 1 - \frac{v_i^{spam} \cdot \bar{v}^{non\text{-}spam}}{||v_i^{spam}||_2 ||\bar{v}^{non\text{-}spam}||_2}$$

Intuitively, $D_{NSR}(v^{spam})$ measures the extent to which spam reports resemble non-spam reports. If

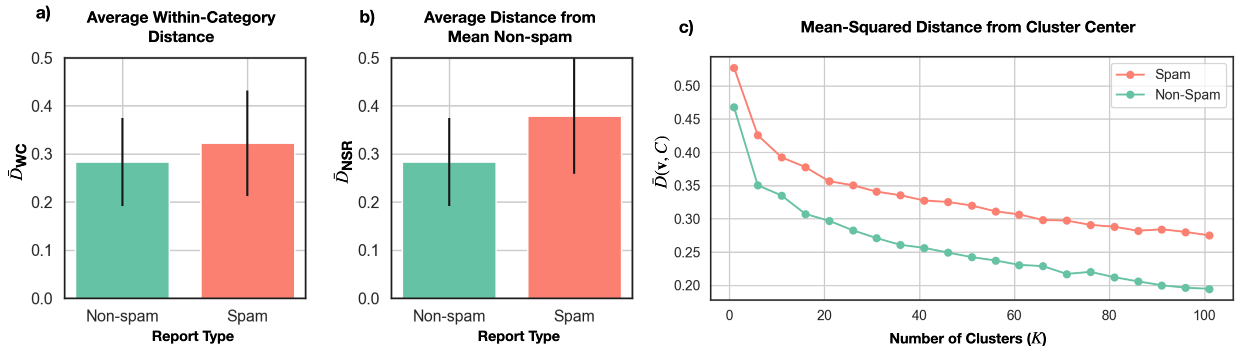

Figure 3. a) Average within-category distance. Spam (red) reports display less semantic similarity than non-spam (green) reports, but not considerably less. Despite covering a wide range of topics, spammers are relatively within each topic. b) Average distance from the mean non-spam embedding. Spam are further from the mean non-spam vector than non-spam, indicating that spam reports are not well-disguised. c) Distance of spam vs. non-spam samples from their cluster centers as a function of the number of clusters K .

$D_{NSR}(v^{spam}) \quad D_{WC}(v^{non\text{-}spam})$, we can say that the generated spam reports are well-disguised semantically amongst non-spam reports. For spam VOICE reports, which cover a much wider qualitative range of semantic content than non-spam reports in Figure 2, we expect $D_{NSR}(v^{spam}) > D_{WC}(v^{non\text{-}spam})$.

Definition 3.3. Let C^{spam} be the set of K disjoint clusters (with centroids $^{spam}_{j}$) computed by the K-means algorithm (Hartigan & Wong, 1979) over embeddings v^{spam}. We define the mean-squared distance of each sample from its cluster centroid, $D(v^{spam};\ ^{spam}_{j})$, as:

$$D(v^{spam};\ ^{spam}_{j}) = \frac{1}{n}\overset{n}{\underset{i}{X}} jjv^{spam}_{i} \quad ^{spam}_{j}jj^{2}$$

Intuitively, $D(v^{spam};\ ^{spam}_{j})$ measures the internal coherence of the clusters generated for spam reports. Given the large variety of spam observed, we expect that it will be much more difficult to cluster coherently than non-spam, so we hypothesize that $D(v^{spam};\ ^{spam}_{j}) > D(v^{non\text{-}spam};\ ^{non\text{-}spam}_{j})$. Finally, because this distance requires selecting K as an input, we compute distances for each K value in the range [1; 100] to get a holistic view of the "clusterability" of spam vs. non-spam reports.

The results for each of our distance measures are shown in Figure 3. The results confirm our hypothesis that spam reports are less well-structured semantically than non-spam reports. This finding is shown in Figure 3b, which shows that spam report embeddings are a considerable distance further from the mean non-spam embedding than are non-spam reports. This indicates that the spam reports are not very well-disguised among the non-spam reports. Moreover, in Figure 3c, the distance from spam embeddings to their cluster centers is greater than that of non-spam embeddings for every value of K . This suggests our results are robust to number-of-clusters: non-spam reports can be clustered more coherently than spam reports. The results for average within-category distance (Figure 3a) are somewhat surprising, however. Though distance within spam reports is greater than distance within non-spam reports (as predicted), the disparity between the two is small. This suggests that despite covering a much wider range of semantic topics than non-spam reports (as outlined in Figures 3b and 3c), the text within each topic tend to be relatively similar. This is an interesting insight into the behavior of spammers of the VOICE platform; and possibly of spammers in a larger context. The spammers seem to agree upon a set of coordinated spam campaigns that use semantically similar verbiage—a testament of spammers' ability to coordinate attacks.

4. Conclusion

This work has initiated a discussion on the security and privacy implications of crowdsourcing and reporting technologies. We have outlined the web of actors, attackers, and victims that are involved in reporting and introduced a unique parallel threat model for this setting. We also examined a case of attacks in the context of a real-world collective spamming attack on the VOICE reporting system. Our analysis suggests that while spam attacks can be harmful to a platform, not all attacks are equal: those that better resemble the set of true reports may be more harmful to an agency's ability to classify true reports from false reports. Humor, troll behavior, and easily-identifiable spam, while more likely to go viral on social media, may be less effective if the goal is obfuscation. As attacker capabilities improve, we posit that deception spam campaigns will pose a significant threat to crowdsourcing technologies. Further research is needed to see how large-scale campaigns, especially those that leverage language models like GPT-3, might be able to submit large numbers of reports that resemble the set of true reports and do not use easily identifiable language.

References

Barocas, S. and Levy, K. Privacy dependencies. Wash. L. Rev., 95:555, 2020.

Bhatti, F., Shah, M. A., Maple, C., and Islam, S. U. A novel internet of things-enabled accident detection and reporting system for smart city environments. Sensors, 19(9):2071, 2019.

Birkbak, A. Into the wild online: Learning from internet trolls, 2018. URL https://www.firstmonday.org/ojs/index.php/fm/article/view/8297.

Braithwaite, J., Westbrook, M., and Travaglia, J. Attitudes toward the large-scale implementation of an incident reporting system. International journal for quality in health care, 20(3):184–191, 2008.

Braithwaite, J., Westbrook, M. T., Travaglia, J. F., and Hughes, C. Cultural and associated enablers of, and barriers to, adverse incident reporting. BMJ Quality & Safety, 19(3):229–233, 2010.

Brunton, F. and Nissenbaum, H. Political and ethical perspectives on data obfuscation. In Privacy, due process and the computational turn, pp. 185–209. Routledge, 2013.

Brunton, F. and Nissenbaum, H. Obfuscation: A user's guide for privacy and protest. Mit Press, 2015.

Campello, R. J., Moulavi, D., and Sander, J. Density-based clustering based on hierarchical density estimates. In Pacific-Asia conference on knowledge discovery and data mining, pp. 160–172. Springer, 2013.

Clark, B. Y., Brudney, J. L., Jang, S.-G., and Davy, B. Do advanced information technologies produce equitable government responses in coproduction: an examination of 311 systems in 15 us cities. The American review of public administration, 50(3):315–327, 2020.

Conneau, A. and Kiela, D. Senteval: An evaluation toolkit for universal sentence representations. arXiv preprint arXiv:1803.05449, 2018.

Fakhraei, S., Foulds, J., Shashanka, M., and Getoor, L. Collective spammer detection in evolving multi-relational social networks. In Proceedings of the 21th acm sigkdd international conference on knowledge discovery and data mining, pp. 1769–1778, 2015.

Freed, D., Palmer, J., Minchala, D., Levy, K., Ristenpart, T., and Dell, N. "a stalker's paradise" how intimate partner abusers exploit technology. In Proceedings of the 2018 CHI conference on human factors in computing systems, pp. 1–13, 2018.

Friedman, B. D. and Albo, M. J. Punishing members of disadvantaged minority groups for calling 911. Policing and Race in America: Economic, Political, and Social Dynamics. New Brunswick: Lexington Books, pp. 141–162, 2017.

Grootendorst, M. Bertopic: Leveraging bert and c-tf-idf to create easily interpretable topics., 2020. URL https://doi.org/10.5281/zenodo.4381785.

Hartigan, J. A. and Wong, M. A. Algorithm as 136: A k-means clustering algorithm. Journal of the royal statistical society. series c (applied statistics), 28(1):100–108, 1979.

Kaghazgaran, P., Caverlee, J., and Squicciarini, A. Combating crowdsourced review manipulators: A neighborhood-based approach. In Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining, pp. 306–314, 2018.

Keats Citron, D. Sexual privacy. Yale LJ, 128:1870, 2018.

Kessell, E. R., Alvidrez, J., McConnell, W. A., and Shumway, M. Effect of racial and ethnic composition of neighborhoods in san francisco on rates of mental health-related 911 calls. Psychiatric Services, 60(10):1376–1378, 2009.

Kopan, T. What is voice? trump highlights crimes by undocumented immigrants. CNN, 2017. URL https://www.cnn.com/2017/02/28/politics/donald-trump-voice-victim-reporting/index.html.

Krebs, B. Man behind fatal 'swatting' gets 20 years, 2019. URL https://krebsonsecurity.com/2019/03/man-behind-fatal-swatting-gets-20-years/.

McCammon, S. What the texas abortion ban does — and what it means for other states. NPR, 2021. URL https://www.npr.org/2021/09/01/1033202132/texas-abortion-ban-what-happens-next.

McInnes, L., Healy, J., and Melville, J. Umap: Uniform manifold approximation and projection for dimension reduction. arXiv preprint arXiv:1802.03426, 2018.

Navarro-Carrillo, G., Torres-Marín, J., and Carretero-Dios, H. Do trolls just want to have fun? assessing the role of humor-related traits in online trolling behavior. Computers in Human Behavior, 114:106551, 2021.

Özkul, M. and Çapuni, I. Police-less multi-party traffic violation detection and reporting system with privacy preservation. IET Intelligent Transport Systems, 12(5):351–358, 2018.

Rayana, S. and Akoglu, L. Collective opinion spam detection: Bridging review networks and metadata. In Proceedings of the 21th acm sigkdd international conference on knowledge discovery and data mining, pp. 985–994, 2015.

Reimers, N. and Gurevych, I. Sentence-bert: Sentence embeddings using siamese bert-networks. arXiv preprint arXiv:1908.10084, 2019.

Rock, J. One call away: 911 abuse as a weapon against minorities. FAU undergraduate law journal, 1:160–160, 2019.

Sasson, C., Haukoos, J. S., Ben-Youssef, L., Ramirez, L., Bull, S., Eigel, B., Magid, D. J., and Padilla, R. Barriers to calling 911 and learning and performing cardiopulmonary resuscitation for residents of primarily latino, high-risk neighborhoods in denver, colorado. Annals of emergency medicine, 65(5):545–552, 2015.

Smith, B. W. and Holmes, M. D. Community accountability, minority threat, and police brutality: An examination of civil rights criminal complaints. Criminology, 41(4): 1035–1064, 2003.

Son, S. and Shmatikov, V. The hitchhiker's guide to dns cache poisoning. In International Conference on Security and Privacy in Communication Systems, pp. 466–483. Springer, 2010.

Sveen, F. O., Sarriegi, J. M., Rich, E., and Gonzalez, J. J. Toward viable information security reporting systems. Information Management & Computer Security, 2007.

Thomas, K., Akhawe, D., Bailey, M., Boneh, D., Bursztein, E., Consolvo, S., Dell, N., Durumeric, Z., Kelley, P. G., Kumar, D., et al. Sok: Hate, harassment, and the changing landscape of online abuse. IEEE Symposium on Security & Privacy, 2021.

Vincent, N., Li, H., Tilly, N., Chancellor, S., and Hecht, B. Data leverage: A framework for empowering the public in its relationship with technology companies. In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, pp. 215–227, 2021.

Wikipedia. Internet troll. URL https://en.wikipedia.org/wiki/Internet_troll.

Zou, S., Xi, J., Wang, S., Lu, Y., and Xu, G. Reportcoin: A novel blockchain-based incentive anonymous reporting system. IEEE Access, 7:65544–65559, 2019.

A. Additional Related Work

Novel crowdsourced reporting systems have been proposed in many domains by computer scientists. Papers have suggested and analyzed healthcare self-reporting technologies (Braithwaite et al., 2010; 2008), blockchain approaches to anonymous reporting (Zou et al., 2019), decentralized traffic and accident reporting (Özkul & Çapuni, 2018; Bhatti et al., 2019), and a variety of updates to crime reporting technologies. Studies have attempted to use simulation to see how robust reporting platforms are to scaling reports (Sveen et al., 2007), however few or no measurement studies have attempted to formally report on the unique security and privacy challenges that these systems face. As crowdsourced sensing and reporting systems become a prominent way for policy-makers to represent constituents and respond to urgent needs, these systems need to be scrutinized from a security perspective to identify the numerous ways that they may enable harms.

Broadly, we leverage the concept of obfuscation conceived in Brunton & Nissenbaum (2013) and further developed in Brunton & Nissenbaum (2015). The book on the subject discusses a relatively long list of examples of obfuscation in the wild, and does not include reporting mechanisms. We therefore extend the theoretical work on obfuscation to a new domain, and test some empirical questions related to effective and ineffective obfuscation strategies. The reporting technology domain is unique because it involves technology-enabled privacy violations that are ultimately carried out by citizens, rather than the police or state actors. This private arena of privacy violations has been theorized from legal and sociological perspectives, for example in work on privacy contingencies (Barocas & Levy, 2020) and intimate partner privacy (Keats Citron, 2018; Freed et al., 2018). The moral complexity of this domain–including the questions about harmful use and legitimate misuse of platforms–requires that we extend the taxonomy work in Thomas et al. (2021) to specifically understand use and abuse behaviors in the context of reporting.

Our taxonomy and threat modelling contributions build on computer security research that systematically analyzes violence including hate and harassment mediated over the internet (Thomas et al., 2021) and other technologies (Freed et al., 2018). Our contributions related to the VOICE logs, which aim to characterize broad-participation obfuscation techniques, is most similar to spam detection literature, especially (Rayana & Akoglu, 2015; Fakhraei et al., 2015; Kaghazgaran et al., 2018). These studies tend to focus on rating reviews which include a numerical parameter. They also often explicitly model social networks to understand the spread of information or opinions. We are less interested in how obfuscation campaigns spread, and more interested in the security implications of these campaigns and what makes them detectable and/or powerful.

B. Clustering Analysis Continued

In addition to the corpus-wide clustering analysis outlined in Section 3.1.1, we perform an additional clustering analysis after manually separating the sets of spam and true reports.

Formally, given a , we pass each sentence $s_i \in S$ through the embedding-generating function $f : S \to \mathbb{R}^d$ (with parameters), which produces a corresponding d-dimensional embedding vector $v_i \in \mathbb{R}^d$. We separate the distribution of report sentences S into distributions for spam (S_{spam}) and non-spam ($S_{non\text{-}spam}$) and cluster each individually. We again follow the clustering method introduced by Grootendorst (2020), which yields a set of topics T_s and T_m for spam reports and non-spam reports, respectively.

Our hypothesis is that spam reports will cover a much broader set of conversational topics than non-spam reports – i.e. $|T_{spam}| > |T_{non\text{-}spam}|$. This stems from the observation that spam responses may be incited by any of the sub-categories of deceptive and abnormal reporting (as outlined in Sections 3.1 and 3.2), whereas valid reports are constrained to the few topics for which the platform was originally intended.

Results of this clustering are shown in Figure 4. We find that the optimal number of clusters, as dictated by HDBSCAN, are $|T_{spam}| = 34$ and $|T_{non\text{-}spam}| = 13$, indicating a much wider range of dialogue in spam reports. Similar to Figure 2, we see a large portion of spam reports on the y-axis of Figure 4a dedicated to abnormal reporting—e.g. obviously false reports describing UFO and extraterrestrial encounters (trolling), politically-charged commentary (opinions), and operator harassment (threats, profanity)—and overloading—e.g. staying silent or playing music (report spam). Non-spam reports, on the other hand, consist mostly of information gathering requests and reports of possible tips / witness accounts related to the original use-case of the VOICE system. Altogether, these results confirm our hypothesis from Section 3.1.1 that spam reports cover a much broader set of conversational topics than non-spam reports.

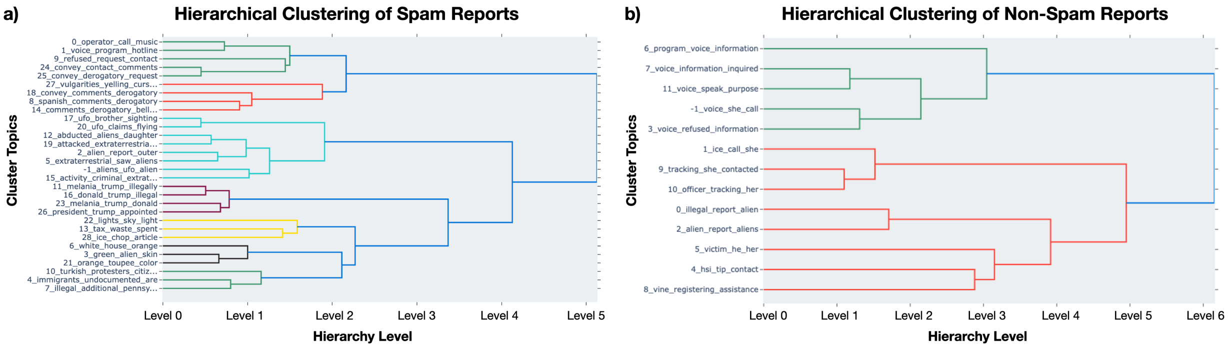
a)
Hierarchical Clustering of Spam Reports
Cluster Topics
0_operator_call_music
1_voice_program_hotline
9_refused_request_contact
24_convey_contact_comments
25_convey_derogatory_request
27_vulgarities_yelling_curs...
18_convey_comments_derogatory
8_spanish_comments_derogatory
14_comments_derogatory_bell...
17_ufo_brother_sighting
20_ufo_claims_flying
12_abducted_aliens_daughter
19_attacked_extraterrestria...
2_alien_report_outer
5_extraterrestrial_saw_aliens
-1_aliens_ufo_alien
15_activity_criminal_extrat...
11_melania_trump_illegally
16_donald_trump_illegal
23_melania_trump_donald
26_president_trump_appointed
22_lights_sky_light
13_tax_waste_spent
28_ice_chop_article
6_white_house_orange
3_green_alien_skin
21_orange_toupee_color
10_turkish_protesters_citiz...
4_immigrants_undocumented_are
7_illegal_additional_pennsy...
Level 0
Level 1
Level 2
Level 3
Level 4
Level 5
Hierarchy Level
b)
Hierarchical Clustering of Non-Spam Reports
Cluster Topics
6_program_voice_information
7_voice_information_inquired
11_voice_speak_purpose
-1_voice_she_call
3_voice_refused_information
1_ice_call_she
9_tracking_she_contacted
10_officer_tracking_her
0_illegal_report_alien
2_alien_report_aliens
5_victim_he_her
4_hsi_tip_contact
8_vine_registering_assistance
Level 0
Level 1
Level 2
Level 3
Level 4
Level 5
Level 6
Hierarchy Level

Figure 4. a) Visualization of the 34 clusters from the set of all spam topics are shown along with their hierarchy. Spam topics include everything from operator harassment (e.g. derogatory comments) to policitally-charged commentary (e.g tax spending, comments about the president) to obviously false reports (e.g. mentions of extraterrestrials). b) Visualization of the 13 clusters from the set of all non-spam topics. These clusters are more relevant to the intended use of the platform (e.g. information gathering).

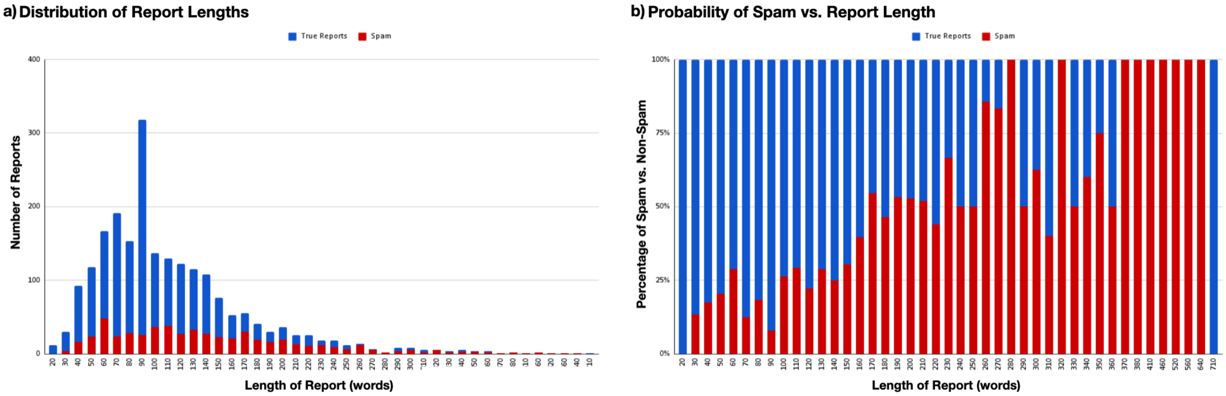

Figure 5. Analysis of report length for spam (red) vs. non-spam (blue) reports. There exists a positive correlation between report length and the probability that a report is spam.

C. Structural Analysis of Reports

As a first test to compare spam and non-spam activity, we examine the structure (rather than content) of the reports. Given that valid users interact with the platform for legitimate purposes, we expect that non-spam will be more concise and to-the-point than spam. To examine this hypothesis, we measure each report's length in words and compare the relative word counts of spam and non-spam reports.

The results of this study are shown in Figure 5. In Figure 5a, we find that, below 150 words, the number of non-spam reports far outweighs the number spam reports. When the number of words in the report exceeds 150, we find significantly fewer non-spam reports and a larger proportion of spam. This finding is re-iterated in 5b, which shows the probability of a report being spam vs. non-spam, given its length. Again we see a stark shift in the percentage of spam reports when the word count exceed 150 words. Altogether, these results suggest that, though it is certainly possible for a shorter length report to be spam, the likelihood of a report being spam grows considerably with the length of the report.