11-785 PROJECT PROPOSAL: SCRUBBING PRIVACY SENSITIVE ENTITIES FROM TEXT MESSAGES

Fanglin Chen
fanglin@cmu.edu

Zhangsihao Yang
zhangsiy@andrew.cmu.edu

Haolian Jiang
haolianj@andrew.cmu.edu

Lan Zou
lzou1@andrew.cmu.edu

1 INTRODUCTION

Text messaging (e.g. email messages, instant messages, and SMS) offers us an intimate, connected and always-on medium to communicate to the rest of the world. Numerous text messaging apps continuously generate text messages on the personal devices and the cloud, containing rich personal information of users. However, these text messages, which are expected to be only shared between the two ends of the conversation, are increasingly likely to be seen by the eyes of the crowd: we share access our personal emails to the crowd to reduce our information overload (Swaminathan et al. (2017); Kokkalis et al. (2013; 2017)); we share pieces of our instant messages to the crowd seeking witty responses for encounters in our private life [1], and we even text to the crowd workers directly to offload many complicated and time-consuming tasks (such as trip planning, grocery shopping, etc) (Lasecki et al. (2013b)).

Crowdsourcing technologies have evolved a lot to be more responsive and efficient to process human text messages. For example, there are services such as Scale [2] to enable service providers to get quick and high-quality crowd results using a single API. Covert human workforces have always been a crucial component of creating and maintaining AI-driven services [3]. This little transparency into what is the audience of the sent text messages (e.g, an computational algorithm or a crowd worker) introduces great privacy risks to the end users, which can cause severe personal harm, ranging from mild social embarrassment to severe financial loss and/or security risk. In this work, we introduce DeepScrub - a scrubbing technique (shown in Figure 1) that can identify and replace sensitive contents from text messages, and evaluate the effectiveness of the technique to be integrated into common crowdsourcing tasks that are related to personal text messages.

2 PROOF OF CONCEPT

The baseline method will separate identification and replacement tasks using two models. For identification algorithm, we will investigate first is to implement is an ANN-based entity recognition model. Specifically, it relies on long short-term memory (LSTM). The NER engine's ANN contains three layers:

- Character-enhanced token-embedding layer

[1] http://www.getmagic.com/

[2] https://www.scaleapi.com/

[3] https://techcrunch.com/2015/08/26/facebook-is-adding-a-personal-assistant-called-m-to-your-messenger-app

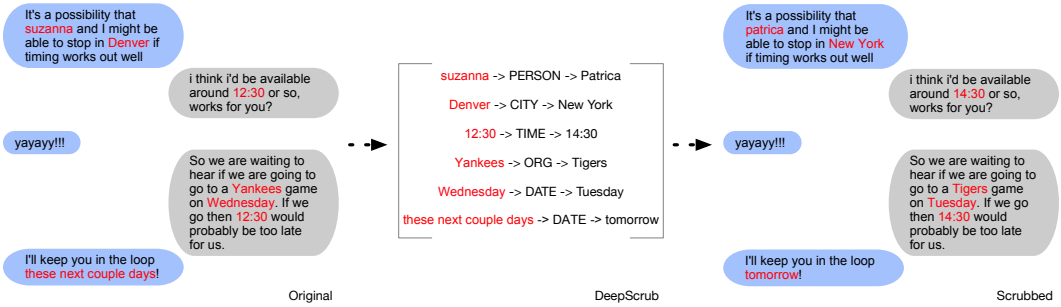

Figure 1: DeepScrub identifies the sensitive data elements in the text messages using a deep NER trained in mixed message corpus, and replaces the sensitive elements with semantic similar texts.

- Label prediction layer
- Label sequence optimization layer

The text replacement algorithm will be based on a hybrid method of template-based text randomizer and word-embedding-based similar word retrieval.

We will also study ways to frame the scrubbing problem as an end-to-end method so that there are no separated algorithms executing by steps. Specifically, we will explore the Transformer model Vaswani et al. (2017) and frame the scrubbing procedure as a sequence to sequence problem.

3 Final Goals & Evaluation

Apart from detection accuracy, in DeepScrub, we implemented four different policies for replacing potentially sensitive text to explore the trade-off while designing a privacy-friendly crowd sourcing pipeline to deal with personal text messages:

- Replacing every privacy component with *XXXX*.
- Replacing specific instance with generic category name, such all persons' names by the word *PERSON*, all addresses and locations by the word *Place* and so on.
- Replacing with generic category name as the previous policy, but separate instances will be identified by numbers. For e.g., if there are two persons referred in a text, they will be replaced by *PERSON1* and *PERSON2* respectively.
- Similar to the previous policy, but additionally the gender of a person will be retained.

These policies are increasingly closer to the original text in terms of information content. We conducted a separate experiment to evaluate how much information is retained by measuring how well people understand texts that were scrubbed using different policies. Since there is a possibility of making inference attacks easier using these contextual cues left by such less aggressive replacement policy, in this experiment we also quantified the probability of inferring person identities.

There is a growing popularity of automated systems that can read and understand text, as well as answer questions based on the information provided in the text. However, if the text and/or questions are complex enough that requires human intelligence, they are redirected to crowd workers. Here our potential attacker is a crowd worker, who is not familiar in person with any of the people involved or referred in the text, but trying to infer their identities. Again the privacy-utility trade-off is to remove all identifying data but keep enough information so that the worker can answer the question. We consider two broad categories of questions: asking about overall meaning or purpose of the text, and asking about any specific set of entities referred in the text. An example will best illustrate this scenario. Consider a text message "Alice, can you meet me at McDonalds around 6 p.m.?". The topic of this message is *meeting request*, and there are four entities, namely the sender, Alice, McDonalds, and 6 p.m. So the questions involving entities might be "Who sent this message?", "Where should I meet with Alice?" and so on. We imagine a crowd working infrastructure where texts will be scrubbed prior to sending to a worker. It is able to understand the question and instruct

the scrubber about which type of data should and should not be scrubbed. We also restrict each question to ask about only one entity, and each subsequent question will be redirected to a different crowd worker. When the requester asks a question about an entity *e*, the scrubber will be instructed not to scrub any entity with the same category as *e* (e.g. do not remove any place entity if the question is "Where should I meet with Alice?"). This ensures that a worker with malicious intention either will learn only the least amount of information that is absolutely required for the desired purpose. We evaluate how well answering such questions are possible when the text is scrubbed with different replacement policies.
In this experiment we evaluate the risk of identification in scenario 1 , and the understand-ability and question answering ability in both scenarios for different scrubbing policies. We selected four broader topic (t1,t2,t3,t4) and sampled ten messages pertaining to each topic from our dataset. These messages were then scrubbed using the four replacement policies, and used in this experiment. To measure identification risk, we asked the participants to infer the relationship between the receiver and the sender. We assume that when the attacker personally knows the receiver, then correct guess about the sender (or any other person referred in the message) will cause an identity leak. To measure the utility of a scrubbed text, we asked five questions:

- What is the topic of this message?
- How many people are referred in the message body?
- Gender of the people referred
- In case of meeting/shopping/lunch/dinner, what is the (generic) location, e.g. restaurant/mall.
- In case of task request, what is the task
- In case of information: is it good news or bad news.

4 Related Work

4.1 Privacy in Crowd-powered Systems

Crowd-powered systems often deal with user-generated personal data in various forms, from text (Swaminathan et al. (2017); Lasecki et al. (2013b)), audio (Lasecki et al. (2012)) to images (Noronha et al. (2011); Burton et al. (2012)). Many of these tasks process potentially sensitive data or make important decisions. For those that run in real time Lasecki et al. (2011), there is greater likelihood for the system to mistakenly leak sensitive information. Even worker activities are monitored Kokkalis et al. (2013), users have very short reaction time to act on the privacy intrusion. Due to ethical, institutional, financial or legal reasons to preserve private information embedded in tasks Brinkman (2013), researchers have proposed various kinds of techniques to make the crowd-sourcing tasks more privacy friendly. Imagery data usually has nice property of keeping most scene-related information with low fidelity, so it is possible for crowd-sourcing systems to obfuscate the images to minimize identity disclosure while still support crowd workers to exact key behaviors (Lasecki et al. (2013a; 2015)). However, blurring or pixelating text regions will completely jeopardize the understability of the texts. Text-oriented crowd-powered systems commonly include facilities for accountability and access control by either recruiting private crowds Bernstein et al. (2015) or only make a subset of the textual data accessible to workers Kokkalis et al. (2013). Another effective method for reducing privacy risks is to identify and then remove or replace the sensitive data elements before presenting the text to the crowd-workers Jozefowicz et al. (2016). However, currently such systems can redact private information effectively only if the data structure can be clearly specified (e.g., phone numbers, license plate numbers, etc.). *DeepScrub* introduces finer-grained protection than parsimonious access control to the level of word phrases and has the potential to keep semantic integrity after scrubbing sensitive texts.

There are also some privacy preserving crowd-sourcing workflows aiming to propose privacy friendly features that can generalize to a wide range of crowd-sourcing tasks. For example, CrowdMask Kaur et al. (2017) allows images with potentially sensitive content to be masked by appearing in progressively larger, more identifiable segments, and masking portions of the image as soon as a risk is identified. This masking technique can potentially map to the domain of textual data when workers see words or phrases that might constitute sensitive information, they should be able to

click them and filter them out. WearMail Swaminathan et al. (2017) ask crowd workers to help retrieve pieces of personal data from private emails by requesting workers to generate examples of the queried subject without looking into the email bodies at all. We see *DeepScrub* as a general privacy preserving operation in any text message related crowd-sourcing tasks.

4.2 Handling Privacy in Large Scale Text Datasets

Data Loss Protection (DLP) is a set of techniques for enterprise corporations to protect corporate information on storage, network traffics and end point devices. The three main methods of current DLP products (e.g., Google Data Loss Protection API dlp, Spirion spi, etc.) are regular expressions, keywords and hashing, which work very well on medical and financial records because of their structures. However it can miss confidential information if it is reformatted or rephrased for different contexts such as email, instant messages or social network posts. To better provide privacy protection, new DLP protection techniques Hart et al. (2011) takes a more conservative approach to classify document or sentences into sensitive or not.

Early practices of privacy protection on large text datasets starts from anonymizing the parties who generate or own the data, however it is soon discovered that anonymization is not sufficient because malicious attackers can possibly identify a person using not directly identifiable data nyt. K-anonymity Sweeney (2002), L-diversity Machanavajjhala et al. (2006) later provide easy solutions to avoid de-anoymization of the datasets if a set of the so-called quasi-identifier attributes can be learned as prior knowledge. These two techniques and their variants (LeFevre et al. (2005); Zhou & Pei (2011)) have been treated as the golden practices for systems and even affect the government policy for years Benitez & Malin (2010). However, this anonymization is also proven to be insufficient because it fails to deal with new attack models which leverage auxiliary information Narayanan & Shmatikov (2008). Differential privacy Dwork et al. (2014) is a major step in the right direction. Instead of the unattainable goal of deidentifying the data, it formally defines what it means for a computation to be privacy-preserving Narayanan & Shmatikov (2010).

Our work, on the other hand, focuses on a different setting of text privacy. Instead of dealing large-scale text datasets, we look at attackers who are only accessible to a small amount of texts which are opportunistically shared in scenarios such as user-generated smartphone notifications and text-message-related crowd-sourcing tasks. By characterizing sensitive information available in text messages, we are proposing sentence and phrase level sensitivity understanding so future researchers may also make use of it.

5 Datasets

We find little has been studied extensively in prior work to characterize sensitive data from text messages in detail. To better understand what kinds of sensitive data should a scrubbing system identify, we examine the sensitive data categories in three different text message datasets, collected from email apps, instant messaging apps, and public tweets because they are generally considered top three text messaging usage platforms [4].

Enron Emails: As one of the earliest large-scale email datasets, Enron email corpus was released to the public and has been used for many researchers to study natural language understanding and social science problems ever since. According to the case studies from EDRM and NUIX nui; edr, more than 10,000 items of privacy-sensitive information was identified in the whole email dataset of 500,000+ emails from 150 senior Enron officials. According to Nuix's case study nui, top sensitive data categories contain a lot of information directly related to someone's identity, such as personal contact details (6237), social security numbers (572), dates of birth (292) and credit card numbers (60). However, the case study was mainly focused on the strict PIIs.

Student Instant Messages: Since instant messaging are becoming more frequently used than email apps on the mobile platform, and there is a lack of the instant messaging dataset publicly available, we conducted a small-scale collection of instant messages to see if there are different patterns in the sensitive data, and have the deeper understanding of sensitive data. We recruited 17 participants

[4] Mobile Messaging and Social Media 2015
http://www.pewinternet.org/2015/08/19/mobile-messaging-and-social-media-2015/

(7 males and 10 females, ages 18-45, mean age 25.2, SD=5.48) from a private U.S. university. All participants used English as the main language on their smartphones and used Android phones running Android 4.4 or above as their main mobile phones. All of our participants also self-identified as active smartphone users. During the course of the study, we found that our participants used their smartphones for an average of 4.2 hours per day, with the least active participant spending an average of 1.1 hours per day and the most active one spending on average 7.7 hours per day. Each participant signed the informed consent form and received a compensation of $50 USD. Participants were then asked to install a custom Android app that we built to track the instant messages for 2 weeks.

Tweet Dialogues: We are also interested in understanding the extend of privacy sensitive information in public text messages. We therefore use the Microsoft Twitter Chat corpus Sordoni et al. (2015) which contains 4,232 three-step conversational snippets extracted from Twitter logs.

To annotate the dataset, we will ask crowd workers on Mechanical Turk to annotate the Enron emails and Tweet dialogues and ask a group of 5 internal annotators to annotate the Student Instant Messages to protect participant privacy. Crowd workers need to submit their worker id in order to proceed to the task. In the public annotation task for crowd workers, each task equally composes of three separate texts. Two are dialogues coming from the Tweet dialogues, and one is a complete email thread from Enron emails. They are instructed to identify all parts of texts which are sensitive data which are defined as any information that can be leveraged to identify, contact, locate or harm an individual in the context. We also provide concrete examples of different sensitive data categories on the left of the screen. In each task, by selecting any parts of the texts, a crowd worker will be prompted with a panel with the defined sensitive data categories. From here the crowd worker can annotate the desired category. We have direct feedback of what have been annotated per category, and if a mistake is made, they can dismiss the annotation by clicking the highlighted part of texts and select the Undo button. If they think none of the categories can apply, they can add a new category. We offer this option to look for qualitative insights to provide potential finer sensitive data granularity. Each crowd worker will be compensated with 5 different annotation tasks in each HIT. To ensure the annotation quality, we have also set up a warm up annotation task for the crowd worker, and if a submitted result is not satisfactory (below 80% alignment of the ground truth), the worker will be given another warm-up task to learn more about the task until their submitted result are good enough. The task assignment is in control of the backend server so that each crowd worker will not annotate one task twice.

Since the crowd annotations are potentially noise, we will leverage data programming method Ratner et al. (2016) to generate probabilistic training datasets.

REFERENCES

Google data loss prevention (dlp) api. https://cloud.google.com/dlp/. Accessed: 2010-09-30.

Edrm enron dataset. http://www.edrm.net/resources/data-sets/edrm-enron-email-data-set/. Accessed: 2010-09-30.

Nuix enron dataset. https://www.nuix.com/sites/default/files/IRMS_Bulletin_Removing_PII_from_Enron_Data_Set_Case_Study.pdf/. Accessed: 2010-09-30.

A face is exposed for aol searcher no. 4417749. https://query.nytimes.com/gst/abstract.html?res=9E0CE3DD1F3FF93AA3575BC0A9609C8B63. Accessed: 2010-09-30.

Spirion data security software - dlp solutions. https://www.spirion.com/. Accessed: 2010-09-30.

Kathleen Benitez and Bradley Malin. Evaluating re-identification risks with respect to the hipaa privacy rule. *Journal of the American Medical Informatics Association*, 17(2):169–177, 2010.

Michael S Bernstein, Greg Little, Robert C Miller, Björn Hartmann, Mark S Ackerman, David R Karger, David Crowell, and Katrina Panovich. Soylent: a word processor with a crowd inside. *Communications of the ACM*, 58(8):85–94, 2015.

Bo Brinkman. An analysis of student privacy rights in the use of plagiarism detection systems. *Science and engineering ethics*, 19(3):1255–1266, 2013.

Michele A Burton, Erin Brady, Robin Brewer, Callie Neylan, Jeffrey P Bigham, and Amy Hurst. Crowdsourcing subjective fashion advice using vizwiz: challenges and opportunities. In *Proceedings of the 14th international ACM SIGACCESS conference on Computers and accessibility*, pp. 135–142. ACM, 2012.

Cynthia Dwork, Aaron Roth, et al. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3–4):211–407, 2014.

Michael Hart, Pratyusa Manadhata, and Rob Johnson. Text classification for data loss prevention. In *Privacy Enhancing Technologies*, pp. 18–37. Springer, 2011.

Rafal Jozefowicz, Oriol Vinyals, Mike Schuster, Noam Shazeer, and Yonghui Wu. Exploring the limits of language modeling. *arXiv preprint arXiv:1602.02410*, 2016.

Harmanpreet Kaur, Mitchell Gordon, Yiwei Yang, Jeffrey P Bigham, Jaime Teevan, Ece Kamar, and Walter S Lasecki. Crowdmask: Using crowds to preserve privacy in crowd-powered systems via progressive filtering. 2017.

Nicolas Kokkalis, Thomas Köhn, Carl Pfeiffer, Dima Chornyi, Michael S Bernstein, and Scott R Klemmer. Emailvalet: Managing email overload through private, accountable crowdsourcing. In *Proceedings of the 2013 conference on Computer supported cooperative work*, pp. 1291–1300. ACM, 2013.

Nicolas Kokkalis, Chengdiao Fan, Johannes Roith, Michael S Bernstein, and Scott Klemmer. Myriadhub: Efficiently scaling personalized email conversations with valet crowdsourcing. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pp. 73–84. ACM, 2017.

Walter Lasecki, Christopher Miller, Adam Sadilek, Andrew Abumoussa, Donato Borrello, Raja Kushalnagar, and Jeffrey Bigham. Real-time captioning by groups of non-experts. In *Proceedings of the 25th annual ACM symposium on User interface software and technology*, pp. 23–34. ACM, 2012.

Walter S Lasecki, Kyle I Murray, Samuel White, Robert C Miller, and Jeffrey P Bigham. Real-time crowd control of existing interfaces. In *Proceedings of the 24th annual ACM symposium on User interface software and technology*, pp. 23–32. ACM, 2011.

Walter S Lasecki, Young Chol Song, Henry Kautz, and Jeffrey P Bigham. Real-time crowd labeling for deployable activity recognition. In *Proceedings of the 2013 conference on Computer supported cooperative work*, pp. 1203–1212. ACM, 2013a.

Walter S Lasecki, Rachel Wesley, Jeffrey Nichols, Anand Kulkarni, James F Allen, and Jeffrey P Bigham. Chorus: a crowd-powered conversational assistant. In *Proceedings of the 26th annual ACM symposium on User interface software and technology*, pp. 151–162. ACM, 2013b.

Walter S Lasecki, Mitchell Gordon, Winnie Leung, Ellen Lim, Jeffrey P Bigham, and Steven P Dow. Exploring privacy and accuracy trade-offs in crowdsourced behavioral video coding. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pp. 1945–1954. ACM, 2015.

Kristen LeFevre, David J DeWitt, and Raghu Ramakrishnan. Incognito: Efficient full-domain k-anonymity. In *Proceedings of the 2005 ACM SIGMOD international conference on Management of data*, pp. 49–60. ACM, 2005.

Ashwin Machanavajjhala, Johannes Gehrke, Daniel Kifer, and Muthuramakrishnan Venkitasubramaniam. l-diversity: Privacy beyond k-anonymity. In *Data Engineering, 2006. ICDE'06. Proceedings of the 22nd International Conference on*, pp. 24–24. IEEE, 2006.

Arvind Narayanan and Vitaly Shmatikov. Robust de-anonymization of large sparse datasets. In *Security and Privacy, 2008. SP 2008. IEEE Symposium on*, pp. 111–125. IEEE, 2008.

Arvind Narayanan and Vitaly Shmatikov. Myths and fallacies of personally identifiable information. *Communications of the ACM*, 53(6):24–26, 2010.

Jon Noronha, Eric Hysen, Haoqi Zhang, and Krzysztof Z Gajos. Platemate: crowdsourcing nutritional analysis from food photographs. In *Proceedings of the 24th annual ACM symposium on User interface software and technology*, pp. 1–12. ACM, 2011.

Alexander J Ratner, Christopher M De Sa, Sen Wu, Daniel Selsam, and Christopher Ré. Data programming: Creating large training sets, quickly. In *Advances in Neural Information Processing Systems*, pp. 3567–3575, 2016.

Alessandro Sordoni, Michel Galley, Michael Auli, Chris Brockett, Yangfeng Ji, Margaret Mitchell, Jian-Yun Nie, Jianfeng Gao, and Bill Dolan. A neural network approach to context-sensitive generation of conversational responses. *arXiv preprint arXiv:1506.06714*, 2015.

Saiganesh Swaminathan, Raymond Fok, Fanglin Chen, Ting-Hao Kenneth Huang, Irene Lin, Rohan Jadvani, Walter S Lasecki, and Jeffrey P Bigham. Wearmail: On-the-go access to information in your email with a privacy-preserving human computation workflow. 2017.

Latanya Sweeney. k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(05):557–570, 2002.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, pp. 6000–6010, 2017.

Bin Zhou and Jian Pei. The k-anonymity and l-diversity approaches for privacy preservation in social networks against neighborhood attacks. *Knowledge and Information Systems*, 28(1):47–77, 2011.