Reasoning about Unstructured Data De-Identification

Patricia Thaine and Gerald Penn, University of Toronto

Patricia Thaine

Patricia Thaine is a Computer Science PhD Candidate at the University of Toronto and a Postgraduate Affiliate at the Vector Institute doing research on privacy-preserving natural language processing, with a focus on applied cryptography. Her research interests also include computational methods for lost language decipherment. She is a recipient of the NSERC Postgraduate Scholarship, the RBC Graduate Fellowship, the Beatrice “Trixie” Worsley Graduate Scholarship in Computer Science, and the Ontario Graduate Scholarship. She has eight years of research and software development experience, including at the McGill Language Development Lab, the University of Toronto's Computational Linguistics Lab, the University of Toronto's Department of Linguistics, and the Public Health Agency of Canada.
Patricia is the Co-Founder and CEO of Private AI, a Toronto- and Berlin-based startup creating a suite of privacy tools that make it easy to comply with data protection regulations, mitigate cybersecurity threats, and maintain customer trust. She is also a member of the Board of Directors of Equity Showcase, one of Canada's oldest not-for-profit charitable organizations.

Email: pthaine@cs.toronto.edu
Phone number: 647-609-1217
Mailing Address: 2117-35 Charles St W, Toronto, ON, M4Y 1R6

Professor Gerald Penn

Gerald Penn is a Professor of Computer Science at the University of Toronto, where he studies spoken language processing and computational linguistics. He has over 100 publications, with the top one accruing 1,581 citations. He is a senior member of IEEE and AAAI, and a past recipient of the Ontario Early Researcher Award. His lab revolutionized speech recognition with its work on neural networks, which received the IEEE Signal Processing Society's Best Paper Award. He has led numerous research projects, including ones funded by Avaya, Bell Canada, CAE, the Connaught Fund, Microsoft, NSERC, the German Ministry for Training and Research, SMART Technologies, the U.S. Army and the U.S. Office of the Director of National Intelligence. Gerald has also worked at Bell Labs and NASA.

Email: gpenn@cs.toronto.edu
Phone number: 647-201-9845

Abstract

We frame the problem of de-identifying unstructured text within the greater landscape of privacy enhancing technologies. We then cover what sort of background knowledge can be gained from only stylistic information about a written document and how we can use research on authorship attribution and author profiling to improve our understanding about the sorts of inferences that can be made from an otherwise de-identified text. Finally, we provide a risk score for determining the likelihood that a message will be attributed to a particular author within a dataset using only author profiling tools.

Keywords (4-6)
anonymization, de-identification, authorship attribution, author profiling, unstructured data, risk

INTRODUCTION

When it comes to developing privacy-preserving tools, there is no one-size-fits-all solution. Every scenario requires careful consideration of the kind of privacy guarantees one wants to make and the kinds of task one wants to achieve. The balance between privacy and utility must be weighed on a case-by-case basis to determine what kind of technology or combination of technologies are best to adapt to privacy legislation requirements, user expectations, employee trust, and data security guarantees. Combining these considerations with Artificial Intelligence (AI) is tricky, as privacy-preserving AI is a fairly new sub-field. We discuss some of the current techniques available to preserve privacy in natural language processing tasks, expand on data de-identification as a technique and the controversies it has faced, and finally explore how to reason about data de-identification in the case of unstructured data, as opposed to its more common application to structured (esp. medical) datasets.

PRIVACY ENHANCING TECHNOLOGIES

Homomorphic Encryption

Homomorphic encryption allows for computations to be performed directly on encrypted data, without needing to decrypt it. One scenario to which homomorphic encryption is ideally suited is when there are computations that must be performed on the cloud which cannot be performed on-device, usually because of resource scarcity, such as low memory or compute power. Alternatively, one might want to keep documents in the cloud and ensure that no one but the owner of the documents can ever search through or decrypt them. While many homomorphic encryption schemes are quantum-safe and can be combined with private information retrieval algorithms to ensure maximal privacy for the data owner, they do have limitations in terms of higher computational cost, availability of information for debugging, and easy information sharing.

Secure Multiparty Computation

Secure Multiparty Computation (MPC) allows for two or more parties to collaborate on computing a result. Neither should know the input of the other, but all parties should know the outcome of the computation. A great example of MPC in practice is [2] where the researchers made it possible to access genomic data from different hospitals in order to make genomic diagnoses. MPC is often combined with homomorphic encryption in order to improve communication costs. One major limitation of MPC is the fact that changing the algorithm sometimes requires changing the entire circuit underlying the cryptographic protocol.

Differential Privacy

Differential privacy [3] allows for generalizations to be made about a population without revealing information that is unique to an individual within that population – be it when querying a dataset or when training a statistical algorithm such as a neural network [4]. It excels at protecting the privacy of a neural network's training data, but it is not effective at, say, extracting data in order to debug software or at making very specific inferences about an individual or about uncommon data points.

Data De-identification

Data de-identification has often been used to make datasets public to either researchers or the general population. The goal behind data de-identification is often not merely to comply with the removal of personally identifiable information, but also to hide the relationship between individuals and their sensitive data (e.g., disease), while allowing enough information (e.g., state and age range) to be available so that some usable conclusions might be drawn about a population [5]. Guarantees that can be made about de-identified data are based on empirical analysis and the statistical information available about relevant populations.

MAKING SENSE OF IT ALL

1000 to 3000 author tasks might seem large scale, but these numbers come nowhere close to the over 4.4 billion Internet users (statistic from January 2019 [19]). So the question remains: if we were to remove all personal identifiers and quasi-identifiers from text before they are posted online, including user IP addresses, email addresses, names, locations, etc., how would that affect the likelihood that a text might be traced back to an author if one were to conduct a stylistic analysis thereof.

We show the results of these calculations in Table 1.

Language	**# speakers on the Internet**[Error! Bookmark not defined.] **(s)**	**% gender accuracy (g)**	**% native language variety accuracy (l)**	**# speakers accurately profiled ($s \times g \times l$)**	**% speakers accurately profiled ($g \times l$)**
English	1,105,919,154	82.3%	89.8%	817,333,974	73.91%
Spanish	344,448,932	83.2%	96.2%	275,691,414	80.04%
Arabic	226,595,470	80%	83.1%	150,640,668	66.48%
Portuguese	171,583,044	84.5%	98.1%	142,232,906	82.89%

Table 1: Proportion of English-, Spanish-, Arabic-, and Portuguese-speaking Internet users who would be correctly profiled by gender and language variety, based on results from the PAN 2017 shared task (a significantly smaller dataset!).

The prospect of accurately identifying 1,38 billion people is rather disconcerting. But now let us look at these impressive numbers in context. Suppose WhatsApp were to profile its users using only stylistic information about their messages. How much would that narrow down the potential author of a message? Table 2 is based on the number of WhatsApp users per country (September, 2019) [21].

Country	**Number of Users (approx.)**[Error! Bookmark not defined.] **(u)**	**# correctly identified: language variety & gender ($u \times g \times l$)**	**# women within correctly identified group ($u \times g \times l \times p_w$)**	**# men within correctly identified group (assuming same proportion of men (p_m) as reported for entire country) ($u \times g \times l \times p_m$)**
Brazil	99,000,000	82,061,100	42,425,589 [22]	39,635,511 [22]
United States	68,100,000	50,332,710	24,663,028 [23]	25,669,682 [23]
Mexico	57,200,000	45,782,880	23,599,423 [24]	22,183,457 [24]

Spain	30,500,000	24,412,200	12,445,077 [25]	11,967,123 [25]
United Kingdom	27,600,000	20,399,160	10,331,603 [26]	10,067,557 [26]

Table 2: Estimated number of WhatsApp users that would be correctly profiled. Note that we assume same proportion of women (p_w) as reported for entire country.

Those fairly high author profiling accuracies suddenly seem less threatening. Making sense of how authorship attribution accuracies might generalize is a little trickier. We would need more information about how the task accuracy decreases as the number of authors increases while the amount of text to train with remains steady. Though close to doing so, [11] does not give us that information.

MEASURING RISK

Calculating a re-identification risk is a much more complicated task than calculating the likelihood that a user has been correctly profiled on all fronts. For one, a risk score should be based on the number of users that satisfy each possible profile combination. In an extreme case, suppose that there were only one Internet user. No matter their profile, we know with certainty that they are the author of the any message we are trying to attribute. Now suppose we have a total of two users, whose demographic information we know and whose messages we are trying to link back to one of them through author profiling.

Scenario 1 (same gender, same language variety):

User 1: (Male, Canadian English)
User 2: (Male, Canadian English)

Risk of attributing a message to the correct author, considering that author profiling techniques are useless here: 50%

Scenario 2 (different gender, same language variety):

User 1: (Male, Canadian English)
User 2: (Female, Canadian English)

Risk of attributing a message to the correct author using gender profiling techniques: 82.3%

This scoring system can be combined with information about the likelihood of an author of a message being identified correctly by author profiling tools. For this purpose, we introduce the concept of *distinctive features*, which will denote features which make one entry distinctive from others in a dataset. For example, if an app has four users; namely three male speakers of Australian English (M, AE) and one female speaker of Australian English (F, AE), then the dataset containing user profiles has one distinctive feature (gender). These features can be independent (like gender and language variety, as can be seen from the marginal distributions in Table 3) or dependent. In general, the author profiling features may be dependent and of varying accuracies relative to their outcomes. We can then compute marginal distributions of accuracy for $\boldsymbol{X}$, over each of its subvectors as defined by selection matrices, $\boldsymbol{S_X} = \{i_1, i_2, \dots, i_k\}$, as follows:

$$p_{S_X}(\boldsymbol{X}) = \sum_{j_i : i \notin \boldsymbol{S_X}} p(x_{1j_1}, x_{2j_2}, \ldots, x_{nj_n}) .$$

We write $i \notin \boldsymbol{S_X}$ exactly when $\boldsymbol{S_X}$ does not select the i^{th} profiling feature. We can denote the equivalence classes over the author profiling features selected by $\boldsymbol{S_X}$ as $\boldsymbol{E_{S_X}}$ and the size of the smallest of those equivalence classes as $\boldsymbol{e_{S_X}}$.

CONCLUSION

We discussed some of techniques available to preserve privacy in natural language processing tasks, expanded on data de-identification as a technique and the controversies it has faced, and explored data de-identification when used on unstructured data as opposed to structured datasets. As a result of our exploration, we proposed a risk score specifically meant for calculating the probability of a user being identified given an author profiling analysis. We hope this risk score can assist experts in determining the re-identification risk of unstructured documents. We expect that it can be enhanced with notions borrowed from Bayes-Optimal Privacy, ℓ-diversity, and t-closeness.

[1] Pathak, M., Portelo, J., Raj, B. and Trancoso, I. "Privacy-Preserving Speaker Authentication," in *Information Security*, vol. 7483, D. Gollmann and F. C. Freiling, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 1–22.

[2] Jagadeesh, K. A., Wu, D. J., Birgmeier, J. A., Boneh, D. and Bejerano, G. "Deriving genomic diagnoses without revealing patient genomes," *Science*, vol. 357, no. 6352, pp. 692–695, Aug. 2017, doi: 10.1126/science.aam9710.

[3] Dwork, C., Smith, A., Steinke, T. and Ullman, J. "Exposed! A Survey of Attacks on Private Data," *Annu. Rev. Stat. Its Appl.*, vol. 4, no. 1, pp. 61–84, Mar. 2017, doi: 10.1146/annurev-statistics-060116-054123.

[4] Song, S., Chaudhuri, K. and Sarwate, A. D. "Stochastic gradient descent with differentially private updates," in *2013 IEEE Global Conference on Signal and Information Processing*, 2013, pp. 245–248, doi: 10.1109/GlobalSIP.2013.6736861.

[5] Wong, R. C.-W., Fu, A. W.-C., Wang, K. and Pei, J. "Minimality Attack in Privacy Preserving Data Publishing," p. 12.

[6] El Emam, K., Jonker, E., Arbuckle, L. and Malin, B. "A Systematic Review of Re-Identification Attacks on Health Data," *PLoS ONE*, vol. 6, no. 12, p. e28071, Dec. 2011, doi: 10.1371/journal.pone.0028071.

[7] `Guidance Regarding Methods for De-identification of Protected Health Information in Accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule,' available at https://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/index.html#safeharborguidance, last accessed on 19/06/2020.

[8] Machanavajjhala, A., Kifer, D., Gehrke, J. and Venkitasubramaniam, M. "*L* -diversity: Privacy beyond *k* -anonymity," *ACM Trans. Knowl. Discov. Data*, vol. 1, no. 1, pp. 3-es, Mar. 2007, doi: 10.1145/1217299.1217302.

[9] Allison, B. and Guthrie, L. "Authorship Attribution of E-Mail: Comparing Classifiers Over a New Corpus for Evaluation," *LREC*, 2008.

[10] Qian, T., Liu, B., Chen, L. and Peng, Z. "Tri-Training for Authorship Attribution with Limited Training Data," in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Baltimore, Maryland, 2014, pp. 345–351, doi: 10.3115/v1/P14-2057. https://pan.webis.de/clef17/pan17-web/author-profiling.html, last accessed on 19/06/2020.