Preserving Privacy in Analyses of Textual Data

Tom Diethe
Amazon
tdiethe@amazon.com

Oluwaseyi Feyisetan
Amazon
sey@amazon.com

Borja Balle
Deep Mind
borja.balle@gmail.com

Thomas Drake
Amazon
draket@amazon.com

ABSTRACT

Amazon prides itself on being the most customer-centric company on earth. That means maintaining the highest possible standards of both security and privacy when dealing with customer data.

This month, at the ACM Web Search and Data Mining (WSDM) Conference, my colleagues and I will describe a way to protect privacy during large-scale analyses of textual data supplied by customers. Our method works by, essentially, re-phrasing the customer-supplied text and basing analysis on the new phrasing, rather than on the customers' own language.

CCS CONCEPTS

• **Security and privacy → Privacy protections**;

ACM Reference Format:
Tom Diethe, Oluwaseyi Feyisetan, Borja Balle, and Thomas Drake. 2020. Preserving Privacy in Analyses of Textual Data. In *Proceedings of Workshop on Privacy in Natural Language Processing (PrivateNLP '20).* Houston, TX, USA, 3 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

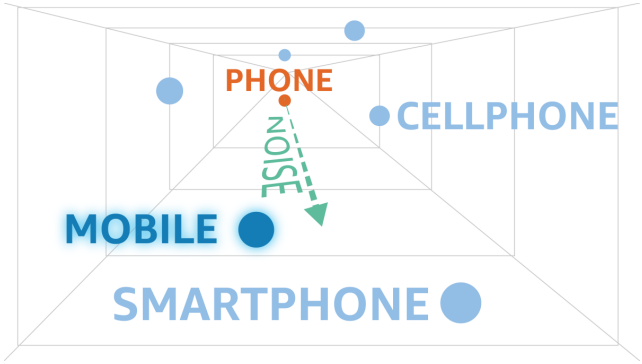

Figure 1: The researchers' technique adds noise (green) to the embedding of a word (orange) from a textual data set, producing a new point in the embedding space. Then it finds the valid embedding nearest that point - in this case, the embedding for the word 'mobile'. (STACY REILLY)

1 DIFFERENTIAL PRIVACY

Questions about data privacy are frequently met with the answer 'It's anonymized! Identifying features have been scrubbed!' However, studies such as this one from MIT show that attackers can de-anonymize data by correlating it with 'side information' from other data sources.

Differential privacy [2] is a way to calculate the probability that analysis of a data set will leak information about any individual in that data set. Within the differential-privacy framework, protecting privacy usually means adding noise to a data set, to make data related to specific individuals more difficult to trace. Adding noise often means a loss of accuracy in data analyses, and differential privacy also provides a way to quantify the trade-off between privacy and accuracy.

Let's say that you have a data set of cell phone location traces for a particular city, and you want to estimate the residents' average commute time. The data set contains (anonymized) information about specific individuals, but the analyst is interested only in an aggregate figure - 37 minutes, say.

Differential privacy provides a statistical assurance that the aggregate figure will not leak information about which individuals are in the data set. Say there are two data sets that are identical, except that one includes Alice's data and one doesn't. Differential privacy says that, given the result of an analysis - the aggregate figure - the probabilities that either of the two data sets was the basis of the analysis should be virtually identical.

Of course, the smaller the data set, the more difficult this standard is to meet. If the data set contains nine people with 15-minute commutes and one person, Bob, with a two-hour commute, the average commute time is very different for data sets that do and do not contain Bob. Someone with side information - that Bob frequently posts Instagram photos from a location two hours outside the city - could easily determine whether Bob is included in the data set.

Adding noise to the data can blur the distinctions between analyses performed on slightly different data sets, but it can also reduce the utility of the analyses. A very small data set might require the addition of so much noise that analyses become essentially meaningless. But the expectation is that as the size of the data set grows, the trade-off between utility and privacy becomes more manageable.

PrivateNLP '20, February 7, 2020, Houston, TX, USA

2 PRIVACY IN THE SPACE OF WORD EMBEDDINGS

In the field of natural-language processing, a word embedding is a mapping from the space of words into a vector space, i.e., the space of real numbers. Often, this mapping depends on the frequency with which words co-occur with each other, so that related words tend to cluster near each other in the space:

So how can we go about preserving privacy in such spaces? One possibility is to modify the original text such that its author cannot be identified, but the semantics are preserved. This means adding noise in the space of word embeddings. The result is sort of like a game of Mad Libs, where certain words are removed from a sentence and replaced with others.

While we can apply standard differential privacy in the space of word embeddings, doing so would lead to poor performance. Differential privacy requires that any data point in a data set can be replaced by any other, without an appreciable effect on the results of aggregate analyses. But we want to cast a narrower net, replacing a given data point only with one that lies near it in the semantic space. Hence we consider a more general definition known as 'metric' differential privacy [1].

3 METRIC DIFFERENTIAL PRIVACY

I said that differential privacy requires that the probabilities that a statistic is derived from either of two data sets be virtually identical. But what does 'virtually' mean? With differential privacy, the allowable difference between the probabilities is controlled by a parameter, epsilon, which the analyst must determine in advance. With metric differential privacy, the parameter is epsilon times the distance between the two data sets, according to some distance metric: the more similar the data sets are, the harder they must be to distinguish.

Initially, metric differential privacy was an attempt to extend the principle of differential privacy to location data. Protecting privacy means adding noise, but ideally, the noise should be added in a way that preserves aggregate statistics. With location data, that means overwriting particular locations with locations that aren't too far away. Hence the need for a distance metric.

The application to embedded linguistic data should be clear. But there's a subtle difference. With location data, adding noise to a location always produces a valid location - a point somewhere on the earth's surface. Adding noise to a word embedding produces a new point in the representational space, but it's probably not the location of a valid word embedding. So once we've identified such a point, we perform a search to find the nearest valid embedding. Sometimes the nearest valid embedding will be the original word itself; in that case, the original word is not overwritten.

In our paper, we analyze the privacy implications of different choices of epsilon value. In particular, we consider, for a given epsilon value, the likelihood that any given word in a string of words will be overwritten and the number of semantically related words that fall within a fixed distance of each word in the embedding space. This enables us to make some initial arguments about what practical epsilon values might be.

4 HYPERBOLIC SPACE

In November 2019, at the IEEE International Conference on Data Mining (ICDM), we presented a paper [4] that, although it appeared first, is in fact a follow-up to our WSDM paper [3]. In that paper, we describe an extension of our work on metric differential privacy to hyperbolic space.

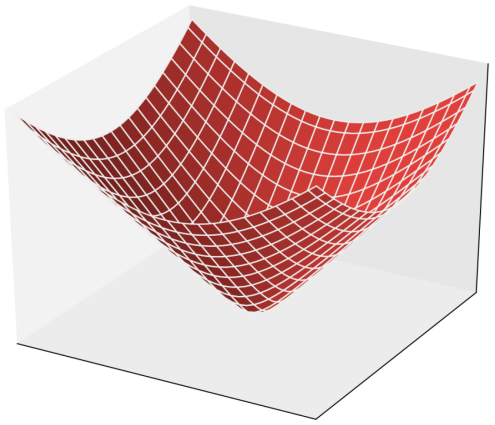

Figure 2: A two-dimensional hyperboloid

The word-embedding space we describe in the WSDM paper is the standard Euclidean space. A two-dimensional Euclidean space is a plane. A two-dimensional hyperbolic space, by contrast, is curved.

In hyperbolic space, as in Euclidean space, distance between embeddings indicates semantic similarity. But hyperbolic spaces have an additional degree of representational capacity: the different curvature of the space at different locations can indicate where embeddings fall in a semantic hierarchy [5].

So, for instance, the embeddings of the words 'ibuprofen', 'medication', and 'drug' may lie near each other in the space, but their positions along the curve indicate which of them are more specific terms and which more general. This allows us to ensure that we are substituting more general terms for more specific ones, which makes personal data harder to extract.

In experiments, we applied the same metric-differential-privacy framework to hyperbolic spaces that we had applied to Euclidean space and observed 20-fold greater guarantees on expected privacy in the worst case.

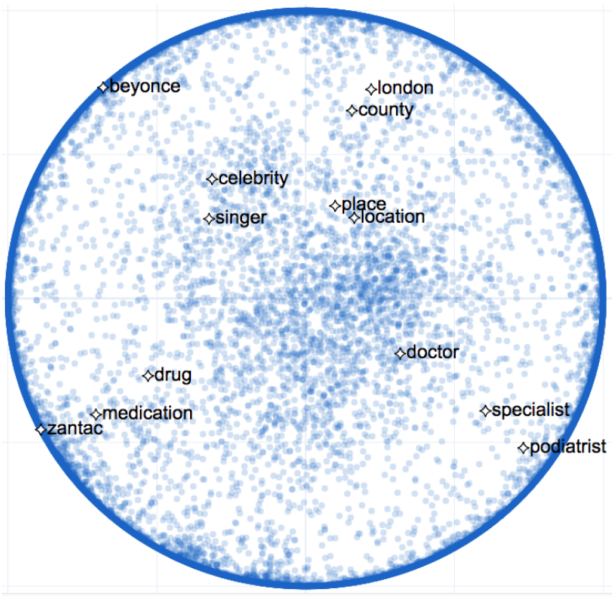

Figure 3: A two-dimensional projection of word embeddings in a hyperbolic space. More-general concepts cluster toward the center, more specific concepts toward the edges.

5 BIOGRAPHY

Dr. Tom Diethe is an Applied Science Manager in Amazon Research, Cambridge UK. Tom is also an Honorary Research Fellow at the University of Bristol. Tom was formerly a Research Fellow for the "SPHERE" Interdisciplinary Research Collaboration, which is designing a platform for eHealth in a smart-home context. This platform is currently being deployed into homes throughout Bristol.

Tom specializes in probabilistic methods for machine learning, applications to digital healthcare, and privacy enhancing technologies. He has a Ph.D. in Machine Learning applied to multivariate signal processing from UCL, and was employed by Microsoft Research Cambridge where he co-authored a book titled 'Model-Based Machine Learning.' He also has significant industrial experience, with positions at QinetiQ and the British Medical Journal. He is a fellow of the Royal Statistical Society and a member of the IEEE Signal Processing Society.

REFERENCES

[1] Konstantinos Chatzikokolakis, Miguel E Andrés, Nicolás Emilio Bordenabe, and Catuscia Palamidessi. 2013. Broadening the scope of differential privacy using metrics. In *International Symposium on Privacy Enhancing Technologies Symposium*. Springer, 82–102.

[2] Cynthia Dwork. 2008. Differential privacy: A survey of results. In *International conference on theory and applications of models of computation*. Springer, 1–19.

[3] Oluwaseyi Feyisetan, Borja Balle, Thomas Drake, and Tom Diethe. 2020. Privacy- and Utility- Preserving Textual Analysis via Calibrated Multivariate Perturbations. In *Proceedings of the 13th International Conference on Web Search and Data Mining*.

[4] Oluwaseyi Feyisetan, Tom Diethe, and Thomas Drake. 2019. Leveraging Hierarchical Representations for Preserving Privacy and Utility in Text. In *IEEE International Conference on Data Mining (ICDM)*.

[5] Maximillian Nickel and Douwe Kiela. 2017. Poincaré embeddings for learning hierarchical representations. In *Advances in Neural Information Processing Systems*. 6338–6347.

A version of this first appeared on the Amazon science blog at:
https://www.amazon.science/blog/preserving-privacy-in-analyses-of-textual-data