Combatting The Challenges of Local Privacy for Distributional Semantics with Compression

Alexandra Schofield
Harvey Mudd College
xanda@cs.hmc.edu

Gregory Yauney
Cornell University
gyauney@cs.cornell.edu

David Mimno
Cornell University
mimno@cornell.edu

Abstract

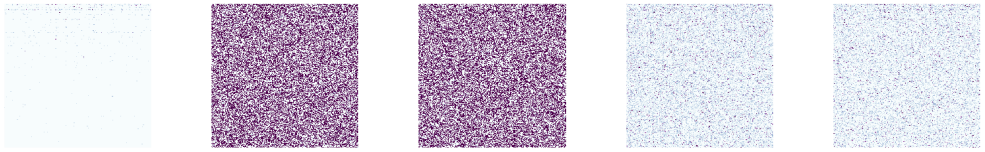
Traditional methods for adding locally private noise to bag-of-words features overwhelm the true signal in the text data, removing the properties of sparsity and non-negativity often relied upon by distributional semantic models. We argue the formulation of limited-precision local privacy, which guarantees privacy between documents of less than a user-specified maximum distance, is a more appropriate framework for bag-of-words features. To reduce the number of features to which we must add random noise, we also compress word features before adding noise, then decompress those features before model inference. We test randomized methods of aggregation as well as methods informed by distributional properties of words. Applying LDA and LSA to synthetic and real data, we show that these approaches produce distributional models closer to those in the original data.

1 Introduction

Text collections in bag-of-words format can surface very specific, unique phrases or associations that raise privacy concerns. An increasingly popular approach to protect data privacy works in the so-called *local model* of *differential privacy* (henceforth *local privacy*) [32, 12], which has recently been deployed by a number of commercial platforms [13, 31]. In local privacy, each user perturbs their own observation with random noise before sending it to a potentially untrusted aggregator, who will then extract useful information from the noisy data. Under local privacy, however, bag-of-words observations are difficult to privatize due to their high-dimensional features: the size of a vocabulary for a text dataset can be orders of magnitude larger than the number of words in the document and, according to Zipf's law, continues to grow as the number of documents increases. Additionally, words are bursty: if a word shows up at all in a document, it is likely to appear several times. Since standard local privacy requires any two documents to be indistinguishable after perturbation, such mechanisms add overwhelming amounts of random noise to all vocabulary features.

We offer three primary contributions. First, we detail how standard existing mechanisms of local differential privacy introduce prohibitive amounts of noise for distributional semantic models, making it difficult to retain the co-occurrence statistics valuable in data. Second, we justify the use of a recent generalization of local privacy, called *limited-precision local privacy* (LPLP) [26], to define our text privacy constraints. LPLP allows us to specify privacy at the level of text spans instead of documents, a more natural privacy guarantee for text that offers some insights in how to parameterize effectively for this privacy definition. Third, to more accurately retain co-occurrence information, we offer a modified version of classic differential privacy with compression [36]: instead of randomly combining documents, we use a low-rank projection to combine features. We show that this approach in conjunction with LPLP leads to released data that more closely preserves co-occurrence statistics of words in text. We measure this through comparisons of LDA [5] and LSA [11] models.

Submission to Workshop for Privacy in Machine Learning (PriML 2019), Vancouver, Canada. Do not distribute.



(a) Original data (b) Laplace (c) Geometric (d) Laplace, LPLP (e) Geometric, LPLP

Figure 1: A visualization of a portion of the Consumer Finance Protection Bureau dataset (a) as a frequency heatmap from 0 (lightest) to 5 or more (darkest). Rows are words in descending order by frequency, columns are documents. Data is processed with different randomized mechanisms under standard local differential privacy with $\epsilon = 2$ (b-c) and limited-precision local privacy (LPLP) with $\epsilon = 2, N = 2$. LPLP preserves more sparsity, it still provides sufficient random noise such that every randomized document shown contains many words not in the original.

2 Prior Methods

There are two primary models of differential privacy. In the *central model*, a trusted curator has direct access to raw private data (without any noise) and releases the output of a differentially private algorithm run on that data. Recent work on differentially private inference of models in machine learning and NLP often uses this model [23, 20, 37, 8]. In contrast, the *local model* [32], or *randomized mechanism* of privacy, adds noise to data rows before they are sent to a central data curator, without knowledge of other data. Since the released data already satisfies differential privacy, any computation or processing on those data can be performed without further privacy cost.

Definition 1 (Local privacy) *Consider a database D with rows in $\mathbb{R}^m$. A randomized mechanism R is* ϵ-locally private *if, for all pairs of possible rows $y, y' \in \mathbb{R}^m$, and a set of possible outputs $S \subset \mathbb{R}^m$,*

$$\Pr[R(y) \in S] \leq e^{\epsilon} \cdot \Pr[R(y') \in S].$$

The amount of randomness introduced in central differential privacy is determined by the "sensitivity" of a feature, defined here as the largest difference summed across all features between any two documents in the corpus. In related problems such as histograms and contingency tables often constrain total corpus sensitivity based on the idea of only modifying one count by one unit in a large collection of counts, requiring relatively little total noise to produce a meaningful privacy guarantee [**?** 3, 18]. However, in this setting, not only does the count data have large dimensions and difficult properties of sparsity, but the difference between neighboring datasets must be defined instead by an upper bound on the number of times a single word would show up in a document. The Laplace [12] and geometric [14] mechanisms both add random noise to produce impractically dense data in Figure 1b and 1c from the original data in Figure 1a. Further, random noise explodes document lengths, reaching an average length of one thousand times the original average document length with Laplace noise enforcing a per-feature privacy budget of $\epsilon = 2$.

An alternate technique specifically targeted for sparse data release subject to privacy constrains is the data sketch [1, 3, 19], inspired by the Johnston-Lindenstrauss lemma [6, 17], which adds noise to each document by randomly projecting the data into a "sketch" and then estimating the true data from the sketch. However, this method, too, is far too dense and noisy to recover meaningful co-occurrence. Further, the complexity of this algorithm scales inversely with the level of privacy desired: if we need little privacy, storing the random projection is impractically memory- and time-intensive. All three approaches conceal term frequencies and variances, with no meaningful overlap of high-probability terms between topics learned on the original data and on the private data.

3 Privacy Definition

Standard local differential privacy in the context of documents aims to add sufficient random noise to make it impossible to distinguish whether a particular document was in the collection or not. For two dominant domains in private text analysis, medicine [10, 15, 24, 25, 27, 34] and search query analysis [9, 16, 22, 28, 29, 30, 35], this level of privacy appears appropriate: identifying a record's identity may leak sensitive information about an individual. However, there are many cases where

this level of anonymity is not required to protect what is sensitive about the underlying data, such as text under copyright or anonymous papers under submission to a conference. Though a paper title may reveal a broader idea, the real content that must be kept private is more likely specific portions of text describing the unique contents of that paper. One objective in text reconstruction prevention, then, is to treat small passages as the object of privacy, not documents.

A recent alternate framework for this is *limited-precision local privacy* [26].

Definition 2 (Limited-precision local privacy) *Consider a database D with rows in $\mathbb{R}^m$. A randomized mechanism R is* (N, ϵ)-limited-precision locally private *if, for all pairs of possible rows $y, y' \in \mathbb{R}^m$ with ℓ_1 difference $\|y - y'\|_1 \leq N$, and a set of possible outputs $S \subset \mathbb{R}^m$,*

$$\Pr[R(y) \in S] \leq e^{\epsilon} \cdot \Pr[R(y') \in S].$$

This formulation allows us to obtain a concrete privacy guarantee, related to portions of documents instead of their full length, without adding prohibitive levels of noise. However, to provide an LPLP guarantee still requires the addition of dense random noise to every vocabulary feature in every document. As shown in Figures 1d and 1e, the resulting output is still quite dense.

4 Horizontal Compression

We next propose a mechanism for operationalizing limited-precision local privacy through compression, an easy way to retain information about large-scale correlation in the data. In our case, we add noise to a compressed representation of the data and then reverse the compression, so that noise will be distributed over multiple elements of the original data. This approach, which we call *horizontal compression*, effectively leverages the *hashing trick* used in nonprivate scenarios to compress sparse feature representations by combining them in a lower-dimensional space [2, 33]. We show this general approach to be private in the appendix.

To create compressed features while maintaining positive integer counts, we group vocabulary terms from the original M-dimensional feature space to sum into K compressed features. We then add noise to the compressed features entry-wise privacy budget using the geometric mechanism [14] under limited-precision local privacy guarantees with an increased per-entry privacy budget $\epsilon' = M\epsilon/K$ corresponding to our K/M-compression. Finally, we reverse this public compression: if private feature k with original feature set F_k has count $\tilde{n}_{dk}$ in document d, we can release an estimate of the total private corpus count of feature $f \in F_k$, $\tilde{c}_f$, using the geometric mechanism. We use this to resample counts of each true feature $\hat{n}_{df}$ by sampling from a multinomial across F_k with priors given by $\tilde{c}_f$: $\vec{n}_{d,F_k} \sim \text{Multi}(\hat{n}_{dk}, \langle \frac{\tilde{c}_f}{\sum_{f' \in F_k} c_{f'}} \rangle)$. This final operation spends a small additional quantity of privacy budget scaling with the original number of features.

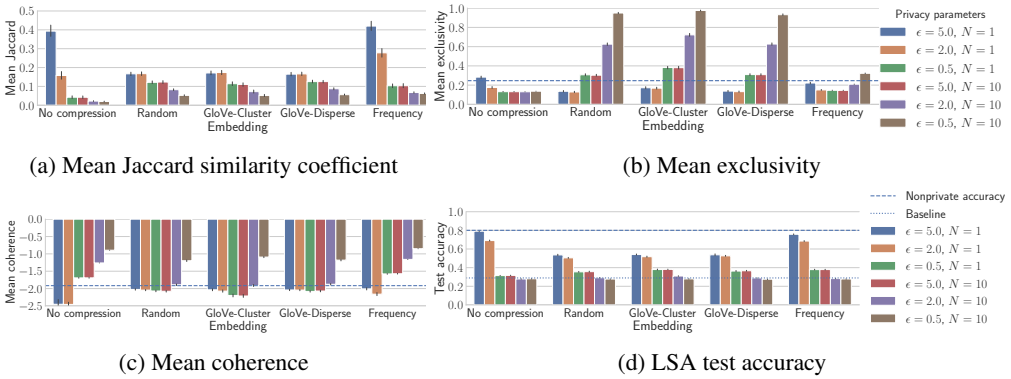
We test four different approaches to assigning N original features to K compressed indices:

- RANDOM: assign features uniformly at random to indices. This is similar to existing methods that combine records instead of features through compression [36].
- FREQUENCY: sort features by decreasing frequency using differential privacy guarantees for histograms, then assign the ith original feature round-robin to $k = i(\mod K)$.
- GLOVE-CLUSTER: Using K-means with the Hungarian algorithm, learn K same-sized clusters of features, making one compressed feature per cluster.
- GLOVE-DISPERSE: as above, but learn $N \mod K$ same-sized clusters, distributing each feature in a cluster to a different compressed index.

In each approach, we can release the projection we used to perform this transformation, as it was either produced using public data or differentially private computations.

5 Experiments

We experimentally validate the efficacy of horizontal compression on LDA model inference [5] and LSA document embedding classification [11] using both synthetic data generated from LDA generative models and real data from a U.S. database of consumer complaints [7]. We compressed from 100 to 10 dimensions for synthetic data and 6800 to 100 dimensions for the real data.



(a) Mean Jaccard similarity coefficient

(b) Mean exclusivity

(c) Mean coherence

(d) LSA test accuracy

Figure 2: Evaluations of LDA and LSA models of real data. Baselines for nonprivate data for exclusivity and coherence of LDA topics are 0.2642 and -1.9167, respectively. Precision, recall, and F_1 score are similar to accuracy for LSA classification. Error bars give 95% confidence intervals across five initializations per combination of privacy level and compression method.

Evaluations. We use several metrics to evaluate LDA models of private data with respect to models of the true data. First, we compare the top 20 terms of private and nonprivate topics to find the Jaccard similarity between the closest pairing of corresponding topics. Second, we study the *exclusivity* of each topic, measured as how unique top terms are to a given topic as compared to other topics in the same model [4]. Third, in the real data, we consider topic coherence [21] to measure how well the top words of a topic actually adhere to a shared subject in meaning space. For latent semantic analysis (LSA) embeddings [11], we apply five-fold cross-validation to compute reconstruction error of the true text and classification accuracy of the data into their product categories.

Results. For weaker privacy guarantee $\epsilon = 5$ with $N = 10$ for the compressed data, the Jaccard similarity for neither the RANDOM (0.281) nor FREQUENCY (0.310) approaches matches that achieved with the same privacy budget on the uncompressed data with $N = 1$ (0.408). However, comparing the stronger privacy level $\epsilon = 0.5, N = 10$ from the uncompressed data with $\epsilon = 0.5, N = 1$ compressed data, both RANDOM (0.201) and FREQUENCY (0.301) outperform the uncompressed model (0.108) in terms of topic Jaccard similarity. The frequency-based approach consistently outperforms the random approach for these and other privacy parameter settings on synthetic data. An additional positive feature of compression is retained sparsity of the data: while document lengths increase, they remain closer to the same average length.

In Figure 2 and 2d, we see that stronger privacy guarantees often benefit from using compression even before acknowledging that compressed methods spend a tenth of the total privacy budget. Our results, however, do not suggest a single optimal method: random projections or frequency-distributed features actually often fare as well or better than compression using a pre-trained embedding. The intuition behind this relates to what we observe in the synthetic data: there is a benefit to avoiding grouping together high-frequency features. As random methods are likely to distribute frequent original features among the compressed features, these produce better reconstructions of the original data. However, we found in Figure 2c that the models using embedding-based methods retain comparable coherence scores with other compressive methods.

6 Discussion

In our work, we describe why limited-precision local privacy may provide more suitable privacy guarantees for bag-of-words features. Further, we show that in cases with more conservative privacy guarantees, using horizontal compression can preserve sparsity and therefore improve the quality of distributional semantic applications. We demonstrate several promising compression approaches leveraging distributional properties of the text. In the future, we hope to combine these approaches with existing mechanisms for privacy in histogram-like data [3, 18] to produce more-optimal privacy guarantees for limited noise.

7 Acknowledgments

The authors would like to thank the reviewers for their helpful comments, as well as Rishi Bommasani, Aaron Schein, Hanna Wallach, and Steven Wu. This work was funded through the National Defense Science and Engineering Graduate Fellowship Program (NDSEG) supported by the Department of Defense, NSF grants #1526155 and #1652536, and a Sloan Fellowship.

References

[1] C. Aggarwal and P. Yu. On Privacy-Preservation of Text and Sparse Binary Data with Sketches. In *Proceedings of the 2007 SIAM International Conference on Data Mining*, Proceedings, pages 57–67. Society for Industrial and Applied Mathematics, 2007. DOI: 10.1137/1.9781611972771.6.

[2] J. Attenberg, K. Weinberger, A. Dasgupta, A. Smola, and M. Zinkevich. Collaborative email-spam filtering with the hashing trick. In *Proceedings of the Sixth Conference on Email and Anti-Spam*, 2009.

[3] R. Bassily and A. Smith. Local, private, efficient protocols for succinct histograms. In *Proceedings of the forty-seventh annual ACM symposium on Theory of computing*, pages 127–135. ACM, 2015.

[4] J. Bischof and E. M. Airoldi. Summarizing topical content with word frequency and exclusivity. In *Proceedings of the 29th International Conference on Machine Learning*, pages 201–208, 2012.

[5] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent Dirichlet allocation. *The Journal of Machine Learning Research*, 3:993–1022, 2003.

[6] J. Blocki, A. Blum, A. Datta, and O. Sheffet. The Johnson-Lindenstrauss transform itself preserves differential privacy. In *Foundations of Computer Science (FOCS), 2012 IEEE 53rd Annual Symposium*, pages 410–419. IEEE, 2012.

[7] CFPB. Consumer complaint database. *ConsumerFinance.gov*, 2018.

[8] K. Chaudhuri, A. D. Sarwate, and K. Sinha. A near-optimal algorithm for differentially-private principal components. *The Journal of Machine Learning Research*, 14(1):2905–2943, 2013.

[9] X. Cheng, S. Su, S. Xu, P. Tang, and Z. Li. Differentially private maximal frequent sequence mining. *Computers & Security*, 55(Supplement C):175–192, 2015.

[10] F. K. Dankar and K. El Emam. The application of differential privacy to health data. In *Proceedings of the 2012 Joint EDBT/ICDT Workshops*, EDBT-ICDT '12, pages 158–166, New York, NY, USA, 2012. ACM.

[11] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman. Indexing by latent semantic analysis. *Journal of the American society for information science*, 41(6):391–407, 1990.

[12] C. Dwork, F. McSherry, K. Nissim, and A. Smith. Calibrating noise to sensitivity in private data analysis. In *Proceedings of the 3rd Theory of Cryptography Conference*, volume 3876, pages 265–284, 2006.

[13] Ú. Erlingsson, V. Pihur, and A. Korolova. RAPPOR: randomized aggregatable privacy-preserving ordinal response. In *Proceedings of the ACM SIGSAC Conference on Computer and Communications Security*, pages 1054–1067, 2014.

[14] A. Ghosh, T. Roughgarden, and M. Sundararajan. Universally utility-maximizing privacy mechanisms. *SIAM Journal on Computing*, 41(6):1673–1693, 2012.

[15] A. Gkoulalas-Divanis, G. Loukides, and J. Sun. Publishing data from electronic health records while preserving privacy: A survey of algorithms. *Journal of Biomedical Informatics*, 50(Supplement C):4–19, 2014.

[16] Y. Hong, J. Vaidya, H. Lu, P. Karras, and S. Goel. Collaborative search log sanitization: Toward differential privacy and boosted utility. *IEEE Transactions on Dependable and Secure Computing*, 12(5):504–518, 2015.

[17] K. Kenthapadi, A. Korolova, I. Mironov, and N. Mishra. Privacy via the Johnson-Lindenstrauss transform. *Journal of Privacy and Confidentiality*, 5(1):39–71, 2013.

[18] B. Li, V. Karwa, A. Slavković, and R. C. Steorts. A privacy preserving algorithm to release sparse high-dimensional histograms. *Journal of Privacy and Confidentiality*, 8(1), 2018.

[19] K. Liu, H. Kargupta, and J. Ryan. Random projection-based multiplicative data perturbation for privacy preserving distributed data mining. *IEEE Transactions on Knowledge and Data Engineering*, 18(1):92–106, 2006.

[20] H. B. McMahan, D. Ramage, K. Talwar, and L. Zhang. Learning differentially private language models. In *Proceedings of the 6th International Conference on Learning Representations*, 2018.

[21] D. Mimno, H. Wallach, E. Talley, M. Leenders, and A. McCallum. Optimizing semantic coherence in topic models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 262–272. Association for Computational Linguistics, 2011-07.

[22] H. Pang, X. Ding, and X. Xiao. Embellishing text search queries to protect user privacy. *Proc. VLDB Endow.*, 3(1-2):598–607, 2010.

[23] M. Park, J. R. Foulds, K. Chaudhuri, and M. Welling. Private topic modeling. In *Proceedings of the NeurIPS Workshop on Private Multi-Party Machine Learning*, 2016.

[24] M. K. Ross, W. Wei, and L. Ohno-Machado. "big data" and the electronic health record. *Yearb Med Inform*, 9(1):97–104, 2014.

[25] S. Scardapane, R. Altilio, V. Ciccarelli, A. Uncini, and M. Panella. Privacy-preserving data mining for distributed medical scenarios. In *Multidisciplinary Approaches to Neural Computing*, Smart Innovation, Systems and Technologies, pages 119–128. Springer, Cham, 2018. DOI: 10.1007/978-3-319-56904-8_12.

[26] A. Schein, Z. S. Wu, A. Schofield, M. Zhou, and H. Wallach. Locally private bayesian inference for count models. *arXiv preprint arXiv:1803.08471*, 2018.

[27] A. Stubbs and Ö. Uzuner. De-identification of medical records through annotation. In *Handbook of Linguistic Annotation*, pages 1433–1459. Springer, 2017.

[28] W. Sun, B. Wang, N. Cao, M. Li, W. Lou, Y. T. Hou, and H. Li. Privacy-preserving multi-keyword text search in the cloud supporting similarity-based ranking. In *Proceedings of the 8th ACM SIGSAC Symposium on Information, Computer and Communications Security*, ASIA CCS '13, pages 71–82, New York, NY, USA, 2013. ACM.

[29] W. Sun, B. Wang, N. Cao, M. Li, W. Lou, Y. T. Hou, and H. Li. Verifiable privacy-preserving multi-keyword text search in the cloud supporting similarity-based ranking. *IEEE Transactions on Parallel and Distributed Systems*, 25(11):3025–3035, 2014.

[30] Y. Tang and L. Liu. Privacy-preserving multi-keyword search in information networks. *IEEE Transactions on Knowledge and Data Engineering*, 27(9):2424–2437, 2015.

[31] A. G. Thakurta, A. H. Vyrros, U. S. Vaishampayan, G. Kapoor, J. Freudiger, V. R. Sridhar, and D. Davidson. Learning new words, 2017. US Patent US9594741B1.

[32] S. L. Warner. Randomized response: A survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association*, 60(309):63–69, 1965.

[33] K. Weinberger, A. Dasgupta, J. Langford, A. Smola, and J. Attenberg. Feature hashing for large scale multitask learning. In *Proceedings of the 26th International Conference on Machine Learning*, pages 1113–1120. ACM, 2009.

[34] Y. Xiao, J. Gardner, and L. Xiong. DPCube: Releasing differentially private data cubes for health information. In *2012 IEEE 28th International Conference on Data Engineering*, pages 1305–1308, 2012.

[35] S. Xu, S. Su, X. Cheng, Z. Li, and L. Xiong. Differentially private frequent sequence mining via sampling-based candidate pruning. In *2015 IEEE 31st International Conference on Data Engineering*, pages 1035–1046, 2015.

[36] S. Zhou, K. Ligett, and L. Wasserman. Differential privacy with compression. In *Information Theory, 2009. ISIT 2009. IEEE International Symposium on*, pages 2718–2722. IEEE, 2009.

[37] T. Zhu, G. Li, W. Zhou, P. Xiong, and C. Yuan. Privacy-preserving topic model for tagging recommender systems. *Knowledge and Information Systems*, 46(1):33–58, 2016.

A Proof of horizontal compression

Theorem 1 (Horizontal Compression) *Consider a database D with m columns, a public projection matrix A of dimension $m \times k$ with $k < m$ and inverse A^{-1}, and a differentially private data release mechanism M that uses elementwise noise with per-entry privacy budget ϵ to perturb the input database before releasing it. In this case, $M(DA, \epsilon)A^{-1}$ produces an ϵ-differentially private release of D.*

Proof. Consider two datasets D and D' with n rows and m columns differing in one row i:

$$\forall j \neq i, 1 \leq i, j \leq n, d_j = d_i.$$

We want to show that the ratio of the probability of an output C given input D is less than $m \cdot \exp e$ times the probability of the same output given input D', or the additive sum of the privacy budget for each column of the differing document:

$$\frac{P(M(DA, \epsilon)A^{-1} = C)}{M(D'A, \epsilon)A^{-1} = C)} \leq m \cdot \exp \epsilon.$$

Because this is a deterministic linear transformation that is nonprivate, we can move to before this post-processing step:

$$\frac{P(M(DA, \epsilon)A^{-1} = C)}{M(D'A, \epsilon)A^{-1} = C)} = \frac{P(M(DA, \epsilon) = CA)}{M(D'A, \epsilon) = CA)}.$$

We then can factorize these probabilities, ignoring rows where D and D' (and thus their corresponding probabilities) are the same and replacing CA with $\tilde{C}$:

$$\frac{P(M(DA, \epsilon) = \tilde{C})}{M(D'A, \epsilon) = \tilde{C})} = \frac{P(M(d_i A, \epsilon) = \tilde{c_i})}{P(M(d_i' A, \epsilon) = \tilde{c_i})}.$$

If M is a valid differentially private data release mechanism, as previously established, we already know that for any pair of rows and resulting output, the above ratio should be less than $\exp \epsilon$ by definition.

$$\frac{P(M(d_i A, \epsilon) = \tilde{c_i})}{P(M(d_i' A, \epsilon) = \tilde{c_i})} \leq k \exp \epsilon \leq m \exp \epsilon.$$

We thus end our proof, showing that as long as $k \leq m$, this projection preserves ϵ-differential privacy.

B Example topics

Embedding	ϵ	N	Jaccard	Top 20 words in private topic	Top 20 words in corresponding nonprivate topic
No compression	5	1	0.6000	**wells fargo mortgage property** loan **which** hsbc **documents who where never case time trust foreclosure** their **any all** transfer its	**wells fargo** american express bank **mortgage case** amex **documents where all** new **trust time which who any never foreclosure property**
Random	5	1	0.2121	**payment** has but **account** times how been will **payments late** being never accounts now **which month made** reported funds told	**payment payments late** due **made** make **month account** would pay amount paid had past time monthly days were months **which**
Frequency	5	1	0.6000	**information report verified experian verify** their **dispute reporting fcra verification accounts transunion these are disputed** provide **has** under any law	**information report verified experian** bankruptcy **verify reporting** equifax **transunion fcra accounts has verification are these** file **dispute** investigation **disputed** record
No compression	$\frac{1}{2}$	10	0.0256	outgoing reassured speaks reactivated capacity **month** storing manipulating xxxx ceo attended studying faxes promptly responsibly tear delinquencies balanced excuses present	balance interest amount paid statement pay full payment charges off due charged account which charge made bmo **month** billing statements
Random	$\frac{1}{2}$	10	0.0811	**could** these **informed** america him return based part delinquent disputing gone repeated arrangements agents **after** buying discharged expired know only	had would were did about told time **after** there what when **informed** which contacted **could** been should them never being
Frequency	$\frac{1}{2}$	10	0.0811	personal breach like customers returned still collect noticed **bureau** every sent over sending months federal **were** prove score cash **remove**	report inquiries inquiry transunion removed hard **were remove** companies would contacted requested did these **bureau** inquires are reporting union all

Table 1: Comparison of private and nonprivate topics. Pairs of topics were matched according to highest Jaccard similarity coefficients, and matches with the 10^{th}-highest Jaccards are shown. Words in bold are shared between both topics. An extended version of this table is available in the appendix.