Measuring Data Anonymization Quality

Elizabeth Witten

August 10, 2021

Abstract

As big data becomes more prominent, and the amount of data on any given individual—both voluntarily provided and inferred from other datasets—grows, privacy is a crucial concern. As a result, research into machine learning as it relates to data anonymization is also becoming more important. The goal of this study is to survey machine learning data anonymization studies to date. Specifically, we investigate how data anonymization is evaluated, and what makes an anonymization good, in terms of *utility retained* and *level of privacy protection*.

1 Introduction

People generate an incredible amount of data every day, just by living and going about their daily routines. According to a 2014 blog post, Facebook alone generates four petabyes of data every day (Wiener and Bronson, 2020). An individual may produce some of this data actively, e.g., by using an activity tracking service like a Fitbit or Apple Watch. However, an individual also produces a large portion of this data passively, often without realizing it. Golbeck and Mauriello (2016) found that Facebook users largely underestimated the amount of data that the Facebook app can access. Some of the data produced by an individual may be public. Voting records are an example of such public data, with who has access to that data and what detail is included varying by state (Underhill, 2019). Further, some of an individual's data may be available for use by private companies. In order to tailor customer experiences and increase sales, companies often track how an individual interacts with their website (Wharton School of the University of Pennsylvania, 2019).

All of this data can be very exciting for companies and researchers alike. As data mining techniques become more advanced, anyone who has access to the data may be able to discover hidden patterns and connections about an individual or a community as a whole. There are many beneficial uses for this kind of data. For example, Obenshain (2004) details some data mining uses in the healthcare industry. Companies may also find this data profitable—see Verbeke et al. (2012)—which is not intrinsically harmful. However, the privacy issues that result from the use of improperly anonymized data have the potential for great harm. Data mining allows others to discover, infer, or approximate information about an individual that that individual may not have intended to share, as demonstrated by Golbeck and Mauriello.

With these concerns in mind, machine learning as it relates to data anonymization has become an increasingly important research topic. There are two primary dimensions for measuring the quality of a data anonymization algorithm: *utility retained* and *level of privacy protection* (Duncan et al., 2001). Quantifying *utility retained* is highly subject to the specific usage of the dataset. However, as new data anonymization techniques emerge, it is important to have some common metrics with which to measure *level of privacy protection*.

Our contributions are: (1) a survey of prior data anonymization research with a focus on how those studies measured data anonymization quality, and (2) a short experiment demonstrating the use of data anonymization metrics.

The rest of this paper is organized as follows. Section 2 contains our survey of data anonymization quality metrics. Section 3 details our experiment. Section 4 gives our conclusions.

2 Survey of Data Anonymization Quality Metrics

2.1 Previous Surveys

This section surveys some previous surveys on data anonymization and privacy preservation research. In general, these surveys discuss *level of privacy protection* by identifying the specific vulnerabilities associated with each privacy technique or algorithm they surveyed. The metrics used for *utility retained* differ based on the type of data anonymization the survey is concerned with. For graph data, utility is measured with respect to how well graph properties—such as degree and path length—are retained, as well as how well the anonymized graph can be used for popular graph data applications—such as *Role eXtraction* and *Influence Maximization*. For differentially private algorithms, utility is discussed in terms of the amount of noise added to the data.

2.1.1 Anonymization of graph and social network data

Zhou et al. (2008) survey privacy preservation techniques for social network data. Social networks are graph models that represent individuals as vertices and the social relationships connecting them as edges. Zhou et al.'s first contribution is to analyze relational data privacy models and identify the places where those models fail for social network data. They identify and offer models of three major aspects of social network data privacy preservation: the *privacy information* at risk, the *background knowledge* that an adversary may use in an attack, and the *specific usage* that an anonymization method considers in order to maximize utility. Then, they map existing research into three corresponding dimensions to highlight the open problems regarding privacy preserving methods for social network data.

Zhou et al.'s second contribution is to systematically categorize social network anonymization approaches into two main categories: *clustering-based* approaches and *graph modification* approaches. Zhou et al. further break down the clustering-based approaches into *vertex clustering*, *edge clustering*, and *vertex and edge clustering* methods. They also break the graph modification approaches into *randomized graph modification* and *greedy graph modification* methods. For each approach, they detail the type of attack that approach is designed to mitigate, as well as what level of utility remains in the anonymized graph.

Ji et al. (2016) expands on Zhou et al.'s work, along with other graph data anonymization surveys. They survey graph data anonymization, de-anonymization, and de-anonymizability quantification techniques, including new anonymization techniques not present in prior surveys.

They classify existing graph data anonymization techniques into six categories: *naive ID removal*, *edge editing (EE) techniques*, *k-anonymity based techniques*, *aggregation / class / cluster based techniques*, *differential privacy (DP) based techniques*, and *random walk (RW) based schemes.* For each of these categories, they analyze utility performance with respect to 15 graph utility metrics and seven high-level application utility metrics, where graph utility refers to how the anonymized data preserves fundamental structural properties of the original graph and application utility refers to how the anonymized data preserves the usefulness of the data for actual applications, such as data mining.

They classify existing structure-based de-anonymization attacks into three categories: *seed-based*, *seed-free*, and *other* de-anonymization attacks. They analyze their performance with respect to metrics such as scalability, practicability, and robustness. They further analyze the vulnerability of the existing graph data anonymization techniques to the existing de-anonymization attacks, using the labels *vulnerable*, *conditionally vulnerable*, and *invulnerable*

Beigi and Liu more broadly survey research on user privacy in social media (Beigi and Liu, 2018, 2020). Their survey is not only concerned with *graph* data, but also *textual*, *spaciotemporal*, and *profile attribute* data. They discuss how the traditional privacy models for structured data—*k-anonymity*, *l-diversity*, *t-closeness*, and *differential privacy*—are adopted for use with unstructured user-generated social media data. They also review each model in terms of their anonymization algorithms and their privacy leakage attack vulnerabilities. They formally define two types of privacy leakage disclosures to cover most of the existing definitions in the literature: *identity disclosure attacks* and *attribute disclosure attacks.*

Beigi and Liu categorize relevant research into five groups: *social graphs and privacy*, *authors in social media and privacy*, *profile attributes and privacy*, *location and privacy*, and *recommendation systems and privacy.* For each category, they discuss existing attacks and proposed solutions.

In discussing works regarding authors in social media and privacy, they discuss *stylometry*, whereby an author of a text can be identified by writing style, citing Abbasi and Chen (2008) and Rao et al. (2000). They review identity disclosure attacks via stylometry, as well as anonymization techniques that can mitigate those attacks, citing Bowers et al. (2015) and Mack et al. (2015) (see 2.3.1), as well as Zhang et al. (2018) (see 2.2.1).

2.1.2 Differentially private machine learning algorithms

Ji et al. (2014) survey methods by which machine learning algorithms can be made differentially private. They begin with some introduction to differential privacy and continue on to categorize differentially private machine learning algorithms by the fundamental machine learning tasks they address. They discuss the different ways that prior papers have measured performance of those algorithms and have approached computing a privacy-preserving model. They also discuss differentially private data release mechanisms and compare their theoretical guarantees.

Ji et al. focus on studies that address how to train a differentially private model while minimizing the noise introduced. They summarize four main principles to achieving this. First, adding noise once is usually better than adding noise many times because the scale of noise is inversely proportional to the procedure's allocated budget. Second, lower global sensitivity leads to less noise. Third, using public datasets—often smaller subsets of larger private datasets—can reduce noise and even increase utility. Fourth, iterative noise addition may be preferred in scenarios where the sensitivity of output model parameters is very large.

They conclude by presenting some open questions, including how to incorporate public data, how to deal with missing data in private datasets, and whether, as the number of observed samples grows arbitrarily large, differentially private machine learning algorithms can be achieved at no cost to utility as compared to corresponding non-differentially private algorithms.

2.1.3 Privacy preservation techniques in big data analytics

Rao et al. (2018) survey privacy-preserving techniques in big data. They identify four major threats: *surveillance*, *disclosure*, *discrimination*, and *personal embarrassment and abuse*. They also list and discuss current privacy preserving techniques: *k-anonymity*, *l-diversity*, *t-closeness*, *randomization*, *data distribution*, *cryptographic techniques*, and *multidimensional sensitivity based anonymization (MDSBA)*. They discuss and compare the utility of these techniques in terms of *suitability for unstructured data*, *attribute preservation*, and *accuracy of results of data analytics*. They conclude their survey by addressing the fact that all the surveyed anonymization techniques are for structured data and that there is a need for a solution for unstructured data.

Rao et al. also propose a data lake privacy preservation model. In this model, the data lake holds raw data from different sources. A machine learning algorithm can dynamically identify sensitive attributes in the data. Once the sensitive attributes have been removed, the data is able to be published.

2.2 Data Anonymization Research with Social Media Data

2.2.1 Privacy-preserving social media data outsourcing

Zhang et al. (2018) discuss and propose a model for the novel problem of privacy-preserving social media data outsourcing. The model of this problem consists of three parties: *social media users*, *social media data service providers (DSPs)*, and *data consumers*. The social media users use the social media to generate and share textual data that can be private or public. The DSP is a social media provider or a third-party data company partnering with a social media provider who hosts and provides access to the social media data. The DSP can outsource the data to data consumers in accordance with privacy policies and agreements. The data consumers are those who send and pay for requests specifying query conditions to the DSP, who then sends the data in user-keyword format. The data consumers cannot obtain intact social media data without going through the DSP.

They propose a novel differentially private mechanism for social media data outsourcing that maintains high utility by introducing less noise than the Laplacian mechanism. To combat the curse of dimensionality, they define a new metric called ϵ-text indistinguishability, which asserts that the more similar two user text vectors are, the more non-distinguishable they are after transformation, and vice versa.

Zhang et al. apply their mechanism to a real-world dataset, and evaluate its performance with regard to privacy and usefulness, defense against user-linkage attacks, and utility on a typical ML task. In evaluating privacy and usefulness, they show that their proposed mechanism can tolerate an arbitrary number of dimensions, meaning that ϵ-text indistinguishability successfully breaks the curse of dimensionality. Their experimental results are able to demonstrate a 64.1% reduction of privacy leakage via inference attacks with a 1.61% reduction of utility in terms of classification accuracy.

2.3 Research on Author De-Anonymization

2.3.1 Mitigating de-anonymization attacks via iterative language translation

Mack et al. (2015) review and expand on Bowers et al. (2015)'s work on the *iterative language translation (ILT)* technique for mitigating de-anonymization attacks via author identification systems (AISs). AISs enable an attacker to identify the author based on writing characteristics of the seemingly anonymized text.

Bowers et al. demonstrated that ILT is effective at concealing an author's writing style. To demonstrate this, they iteratively translated an English text into a foreign language (e.g. Spanish, Chinese, Arabic) and back into English for one, two, and three iterations. They then compared the effectiveness of ILT against two well-known AISs.

Because the AISs used in Bowers et al. (2015) were relatively weak, Mack et al. developed four stronger AISs, against which they observed the effectiveness of ILT. They begin by defining four baseline AISs, two of which are the AISs used by Bowers et al.: Uni-Gram, O. de Vel et. al., Hybrid-I (combining the feature sets of Uni-Gram and O. de Vel et al.), and Hybrid-II (based on the AIS proposed by Narayanan et al. (2012)). They then introduce the concept of genetic and evolutionary feature selection (GEFeS). GEFeS is a feature selection technique that is based on simulated evolution and used to evolve feature masks (FMs). The AISs extract feature vectors representing writing style from author samples, and the feature masks "mask out" the non-salient features in a feature vector in an effort to discover high-performing sub-feature sets. Mack et al. applied GEFeS to the four baseline AISs in order to produce four stronger AISs.

Mack et al. used those eight AISs in experiments to determine the effectiveness of ILT against stronger AISs. Their dataset consisted of English language blog text samples from 1000 authors on a wide variety of blog sites. For each author, the associated sample was split into two-paragraph sub-samples, the first of which they placed into the probe set, and the rest of which they placed into the gallery set. Experiment I was "English-to-English", in which the associated probe and gallery instances were the original English text. This experiment was used to determine the relative strengths of the eight AISs in terms of author identification accuracy, and they found that the Uni-Gram AIS was the strongest baseline AIS, while the Hybrid-II+GEFeS AIS was the strongest overall. Experiment II was the application of ILT, in which all of the gallery instances were translated into Spanish, Chinese, or Arabic, and then translated back to English. This process was repeated for one, two, and three iterations for each of the three languages. This experiment was used to determine the effectiveness of ILT in concealing an author's identity. The results of this experiment indicate that ILT significantly improves level of privacy protection—measured by accuracy of the Baseline+GEFeS AISs—with the first iteration causing the largest drop in author identification rate. There was no measure of utility for this experiment.

2.3.2 Neural Generative Models for Synthetic Text Data

Ororbia II et al. (2016) present a method for generating synthetic versions of Twitter data using a novel neural architecture to protect authors from stylometric re-identification attacks. They compare their method to redaction and iterative translation with regard to risk and utility, and their experimental results indicate an improved risk-utility trade-off when using the neural models. Further, this risk-utility trade-off can be managed via a straightforward privacy tuning parameter.

For each user in the attack dataset, they trained a classifier to predict that user from the other users in the dataset. They used a variety of feature sets and four classifiers per feature set: regularized least squares (RLSC), support vector machines (SVM), naive Bayes (NB), and k-nearest neighbors (KNN). Then, they tested each classifier-feature set combination against the release dataset. From these results, they measured identification risk as the percentage of users with correct matches in the top x most likely users.

They measured utility using four measures. The first two measures—average user uni-gram and bi-gram cosine similarity between the baseline and altered datasets—are general utility measures where high similarity implies better utility. The last two measures—performance on a classification task and a sentiment analysis task—are model-specific utility measures where similarity of task outcomes implies better utility.

3 Data Anonymization Quality Experiment

After investigating the *level of privacy protection* and *utility retained* metrics used in prior studies, we conducted a short experiment demonstrating the use of data anonymization quality metrics—specifically *level of privacy protection*. The structure for this experiment was drawn from the research surveyed in Section 2, in particular the re-identification attack model in Section 2.3.2 by Ororbia II et al.

The purpose of this experiment is to create a baseline for measuring data anonymization quality. Using a powerful model such as those found on Hugging Face[1], data anonymization quality metrics can be integrated into the anonymization process.

3.1 Data

We drew our data from a set of tweets written from 2012-2013 by Twitter users located around Rochester, New York. The original dataset consisted of 7,679,355 tweets, which we filtered to tweets by users who had at least 10 tweets in the dataset. The resulting dataset consisted of 668,552 tweets by 6,109 users. We then split the dataset into test, dev, and train sets as follows. We first grouped all tweets by user. Then, for each group of tweets, 25% (rounded down) were assigned to the test set, 25% (rounded down) were assigned to to the dev set, and the remaining tweets were assigned to the train set. This ensures that

[1] `https://huggingface.co/`

the users were distributed proportionally. In the end, the test set contained 164,776 tweets, the dev set contained 164,776 tweets, and the train set contained 339,000 tweets.

3.2 Methods

As a measure of anonymization quality, we trained a classifier for each unique user in our dataset to predict whether a given tweet was authored by that user, resulting in 6,109 classifiers. Specifically, for users $U = \{u_1, u_2, ..., u_k\}$, we trained binary classifiers $C = \{c_1, c_2, ..., c_k\}$ so that, given a tweet written by u_i, c_j classifies the tweet in the positive class if it predicts that $i = j$.

To implement training, we chose to use the multinomial naive Bayes classifier using the `sklearn` library (Pedregosa et al., 2011). Naive Bayes classifiers are simple supervised learning algorithms based on applying Bayes' theorem with the naive assumption that each feature is independent of every other feature given the class variable. Multinomial naive Bayes is an implementation of the naive Bayes algorithm for multinomially distributed data, and is one of two naive Bayes classifiers primarily use for text classification. The data is represented as word count vectors or tf-idf vectors. While simple, naive Bayes are known to be decent classifiers. However, it should be noted that the naive Bayes probability estimation is poor. This is due to the fact that its loss function does not penalize inaccurate probability estimation as long as the maximum probability is assigned to the correct class (Zhang, 2004). We chose to use multinomial naive Bayes classifiers because its `partial_fit` method enabled us to perform out-of-core model fitting that was advantageous on lower-memory machines.

Our feature set consisted of uni-grams, bi-grams, and tri-grams limited to the top 4000 features ordered by term frequency. We used the `CountVectorizer` from `sklearn` to convert the tweet content to a count matrix. We then used the `TfidfTransformer` to transform that count matrix into a normalized tf-idf representation.

Using these classifiers, we can integrate *level of privacy protection* metrics into an anonymization algorithm. The code for this experiment, as well as a prototype anonymization script using the uncased BERT base model[2] from Hugging Face, can be found in our GitHub repository[3].

3.3 Results

We evaluated the quality of the user classifiers by plotting Receiver Operating Characteristic (ROC) (figure 1) and Precision-Recall (figure 2) curves using the un-anonymized test dataset.

For the ROC curve, we plot both the macro-average and the micro-average curves. For the precision-recall curve, we plot the micro-average curve. The macro-average treats all classes equally by first computing the curves for each user classifier and then plotting the average. The micro-average aggregates the contributions of all the classes and then computes the curve. In this way, classes that make up a higher proportion of the data contribute more to the final average.

[2] `https://huggingface.co/bert-base-uncased`

[3] `https://github.com/eaw8044/data-anonymization-quality`

The ROC curves both have a fairly high area under the curve—0.95 for the micro-average curve and 0.88 for the macro-average curve—indicating that the user classifiers are skillful. The precision-recall curve, however, indicates a very low skill for most thresholds.

This is likely due to the fact that the class distribution is highly imbalanced. There is a much higher occurrence of the negative class—that tweet is not by the user associated with the classifier—than the positive class. Therefore, the ROC curves are likely over-optimistic since that measurement of skill includes the many true negative predictions. On the other hand, the precision-recall curve measures the classifiers' skill at predicting just the positive classes, indicating that the user classifiers are not skilled at correctly identifying tweet authorship.

This low skill is likely due to the simple features used in our experiment. Feature sets like those used in Narayanan et al. (2012) would likely result in better classifiers. Another potential factor contributing to the low skill precision-recall curve is the fact that, as mentioned above, naive Bayes classifiers have largely inaccurate probability estimates, which we used to calculate the precision-recall curve (Zhang, 2004).

Figure 1: Receiver Operating Characteristic (ROC) curves for the user classifiers on the non-anonymized test data

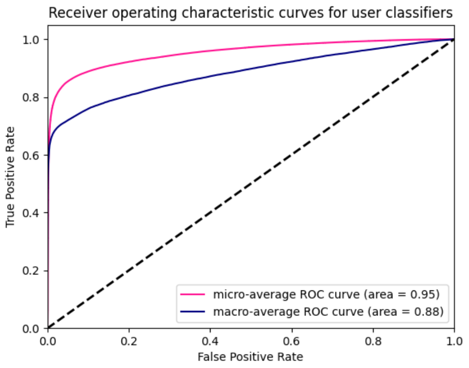
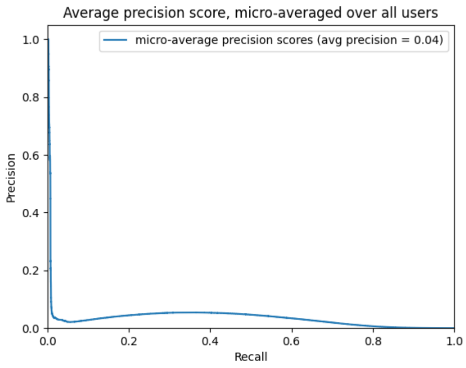


Figure 2: Precision-Recall curves for the user classifiers on the non-anonymized test data



4 Conclusion

In a world where the collection and use of people's data is present in almost every aspect of online life, there are great opportunities for researchers to use that data. Data mining techniques allow researchers to find patterns that were previously invisible. However, users of a website or application—Twitter, for example—may not intend to share information beyond the surface level content they share. Therefore, to allow for more responsible use of such data, data anonymization is an important research topic.

In the first part of this paper, we surveyed prior research of machine learning as it relates to data anonymization. We specifically were interested in detailing how those papers measured the quality of a data anonymization algorithm, in terms of *utility retained* and *level of privacy protection*. We found that previous surveys discuss *level of privacy protection* by identifying specific vulnerabilities associated with each privacy technique they surveyed. Since the surveys covered a range of data anonymization types—graph anonymization, differential privacy, unstructured data, author de-anonymization—there was no unified way to measure *utility retained*. For the other research papers we surveyed, we found that the most common way to measure *level of privacy protection* was to simulate one or more attacks and measure the success of those attacks against the model they proposed. The most common way to measure *utility retained* was to perform a relevant machine learning task (such as classification) and compare the performance of the anonymized dataset against the original dataset.

In the second part of this paper, we performed a short experiment based on surveyed research looking at how user classifiers could be used as a measure of *level of privacy protection*. This experiment serves as a baseline for future work investigating data anonymization metrics. Future research could look at incorporating both measures of data anonymization quality into an anonymization algorithm to even better tailor the anonymization to the machine learning task it will be used for.

5 Acknowledgements

I'd like to thank my research advisor Dr. Christopher Homan for his guidance and support on this project.

References

Abbasi, A. and H. Chen
2008. Writeprints: A stylometric approach to identity-level identification and similarity detection in cyberspace. *ACM Transactions on Information Systems (TOIS)*, 26(2):1–29.

Beigi, G. and H. Liu
2018. Privacy in social media: Identification, mitigation and applications. *CoRR*, abs/1808.02191.

Beigi, G. and H. Liu
2020. A survey on privacy in social media: Identification, mitigation, and applications. *ACM Transactions on Data Science*, 1(1).

Bowers, J., H. Williams, G. Dozier, and R. Williams
2015. Mitigation deanonymization attacks via language translation for anonymous social networks. In *Proc. the 7th International Conference on Machine Learning and Computing*.

Duncan, G., S. Keller-McNulty, and S. Stokes
2001. Disclosure risk vs. data utility: the ru confidentiality map. Technical report, Technical Report. Durham: US National Institute of Statistical Sciences.

Golbeck, J. and M. L. Mauriello
2016. User perception of facebook app data access: A comparison of methods and privacy concerns. *Future Internet*, 8(2).

Ji, S., P. Mittal, and R. Beyah
2016. Graph data anonymization, de-anonymization attacks, and de-anonymizability quantification: A survey. *IEEE Communications Surveys & Tutorials*, 19(2):1305–1326.

Ji, Z., Z. C. Lipton, and C. Elkan
2014. Differential privacy and machine learning: a survey and review. *arXiv preprint arXiv:1412.7584*.

Mack, N., J. Bowers, H. Williams, G. Dozier, and J. Shelton
2015. The best way to a strong defense is a strong offense: Mitigating deanonymization attacks via iterative language translation. *International Journal of Machine Learning and Computing*, 5(5):409.

Narayanan, A., H. Paskov, N. Z. Gong, J. Bethencourt, E. Stefanov, E. C. R. Shin, and D. Song
2012. On the feasibility of internet-scale author identification. In *2012 IEEE Symposium on Security and Privacy*, Pp. 300–314. IEEE.

Obenshain, M. K.
2004. Application of data mining techniques to healthcare data. *Infection Control & Hospital Epidemiology*, 25(8):690–695.

Ororbia II, A. G., F. Linder, and J. Snoke
2016. Privacy protection for natural language: Neural generative models for synthetic text data. *CoRR*, abs/1606.01151.

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay
2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Rao, J. R., P. Rohatgi, et al.
2000. Can pseudonymity really guarantee privacy? In *USENIX Security Symposium*, Pp. 85–96.

Rao, P. R. M., S. M. Krishna, and A. S. Kumar
2018. Privacy preservation techniques in big data analytics: a survey. *Journal of Big Data*, 5(1):1–12.

Underhill, W.
2019. https://www.ncsl.org/research/elections-and-campaigns/access-to-and-use-of-voter-registration-lists.aspx.

Verbeke, W., K. Dejaeger, D. Martens, J. Hur, and B. Baesens
2012. New insights into churn prediction in the telecommunication sector: A profit driven data mining approach. *European journal of operational research*, 218(1):211–229.

Wharton School of the University of Pennsylvania
2019. Your data is shared and sold...what's being done about it? https://knowledge.wharton.upenn.edu/article/data-shared-sold-whats-done/.

Wiener, J. and N. Bronson
2020. Facebook's top open data problems.

Zhang, H.
2004. The optimality of naive bayes. *AA*, 1(2):3.

Zhang, J., J. Sun, R. Zhang, Y. Zhang, and X. Hu
2018. Privacy-preserving social media data outsourcing. In *IEEE INFOCOM 2018 - IEEE Conference on Computer Communications*, Pp. 1106–1114.

Zhou, B., J. Pei, and W. Luk
2008. A brief survey on anonymization techniques for privacy preserving publishing of social network data. *ACM Sigkdd Explorations Newsletter*, 10(2):12–22.