On Syntactic Anonymity and Differential Privacy

Chris Clifton[#], Tamir Tassa[*]

[#]*Department of Computer Sciences, Purdue University*
clifton@cs.purdue.edu

[*]*Department of Mathematics and Computer Science, The Open University, Ra'anana, Israel*
tamirta@openu.ac.il

***Abstract*—Recently, there has been a growing debate over approaches for handling and analyzing private data. Research has identified issues with syntactic anonymity models. Differential privacy has been promoted as *the* answer to privacy-preserving data mining. We discuss here issues involved and criticisms of both approaches, and conclude that both have their place. We identify research directions that will enable greater access to data while improving privacy guarantees.**

I. Introduction

In recent years, there has been a tremendous growth in the amount of personal data that can be collected and analyzed. Data mining tools are increasingly being used to infer trends and patterns. Of particular interest are data containing structured information on individuals. However, the use of data containing personal information has to be restricted to protect individual privacy. Although identifying attributes like ID numbers and names can be removed from the data without affecting most data mining, sensitive information might still leak due to linking attacks that are based on the public attributes, a.k.a. *quasi-identifiers.* This has lead to two related research areas: privacy-preserving data mining [1] enables the learning and use of data mining models while controlling the disclosure of data about individuals; privacy-preserving data publishing focuses on anonymizing datasets, to allow data disclosure without violating privacy.

Probably the first formal mathematical model to achieve wide visibility in the computing research community was k-anonymity, proposed by Samarati and Sweeney [2], [3]. This model requires that each of the released records be indistinguishable from at least $k-1$ other records when projected on the quasi-identifier attributes. Several studies have pointed out weaknesses of the k-anonymity model and suggested stronger measures, e.g., ℓ-diversity [4], t-closeness [5], or β-likeness [6]. Other studies attempted to enhance the utility of such anonymized tables, e.g., [7], [8], [9], [10]. The models proposed in those studies are similar to k-anonymity in the sense that they achieve anonymity by means of (typically) generalize the database entries until some syntactic condition is met, so that the ability of an adversary to link a quasi-identifier tuple to sensitive values is restricted.

It has been shown that all of the models of so-called syntactic anonymity are susceptible to various attacks. The emergence of differential privacy [11], a rigorous notion of privacy based on adding noise to answers to queries on the data, has revolutionized the field of privacy-preserving data mining. This has been a catalyst in the observed decline in the support of syntactic anonymity. There seems to be a widespread belief that differential privacy and its offsprings are immune to those attacks, and that they render the syntactic models of anonymity obsolete. In this paper we discuss the problems with syntactic anonymity and argue that, while all those the problems are genuine, they can be addressed within the framework of syntactic anonymity. We further argue that differential privacy too is susceptible to attacks, as well as having other problems and (often unstated) assumptions that raise problems in practice.

While criticism of syntactic anonymity stems from its shortcomings in providing full privacy for the individuals whose data appear in the table, it is imperative also to discuss the second aspect of privacy-preserving data publishing: the utility of the sanitized data for legitimate (non-privacy-violating) purposes. As we see in the news on a regular basis (such as the changes to privacy policies and practices of Google and facebook), without regulation utility trumps privacy: if the choice is between a method that provides privacy but fails to adequately support data analysis, or sharing data at a greater risk to privacy, the choice will be to share the data. Another example comes from the U.S. HIPAA Privacy Rule [12], which provides a clear syntactic mechanism to anonymize data to meet legal standards. As for differentially private data, a study of its utility is still in order. Until it is clarified how useful it is for practitioners of data mining, differential privacy has still not reached the maturity to replace other existing models of privacy-preserving data mining.

II. PPDM and PPDP

There is a fundamental difference between the assumptions that underlie differential privacy and those that underlie syntactic privacy models. In fact, those two seemingly competing approaches play in different playgrounds.

Syntactic anonymity targets privacy-preserving data *publishing* (PPDP). A typical scenario of PPDP is that in which a hospital wishes to release data about its patients for public scrutiny of any type. The hospital possesses the data and is committed to the privacy of its patients. The goal is to *publish* the data in an anonymized manner without making any assumptions on the type of analysis and queries that will be executed on it. Once the data is published, it is available for any type of analysis.

Differential privacy, on the other hand, typically targets privacy-preserving data *mining* (PPDM). In PPDM, as opposed to PPDP, the query that needs to be answered must be known prior to applying the privacy-preserving process. In the typical PPDM scenario, the data custodian maintains control of the data and does not publish it. Instead, the custodian responds to queries on the data, and ensures that the answers provided do not violate the privacy of the data subjects. In differential privacy this is typically achieved by adding noise to the data, and it is necessary to know the analysis to be performed in advance in order to calibrate the level of noise to the global sensitivity of the query and to the targeted differential privacy parameter ε [13]. While some differential privacy techniques (e.g., private histograms) are really intermediate analysis rather than a final data mining model, it is still necessary for the data custodian to know what analysis is intended to be performed.

In their criticism on syntactic models of privacy and defense of differential privacy, Narayanan and Shmatikov [14] state that PPDP is a bad idea and that only PPDM may provide sufficient privacy. They acknowledge the impracticality of that conclusion by adding that "this can be a hard pill to swallow, because it requires designing a programming interface for queries, budgeting for server resources, performing regular audits, and so forth." Hence, while interactive approaches do have some advantages in the privacy vs. utility tradeoff, their inherent limitations are such that PPDP is likely here to stay.

The comments in [14] also miss the point that differential privacy does not necessarily imply an interactive approach. Noise and syntactic generalization have in fact been combined to support real-world data publishing [15]. The definition of differential privacy supports a query such as "return the dataset D", requiring that the returned data have noise added (as with some public use microdata sets) to ensure that the information related to any individual is sufficiently hidden. While differentially private data publishing has been shown to be possible [16], [17], [18], [19], [20], [21], there has been little work to show that such an ε-differentially private dataset would be practical and useful.

Data publishing is a widespread practice (see, for example, public use microdata sets[1]); hence, it is important to develop appropriate techniques for PPDP. Fung et al. [22] argue that even if the data custodian knows in advance that data will be used for classification, it is not enough to just build and publish a classifier. First of all, even if the data custodian knows that the data will be used for classification, it may not know how the user may analyze the data. The user often has application-specific bias towards building the classifier. For example, some users prefer accuracy while others prefer interpretability, or some prefer recall while others prefer precision. In other cases, visualization or exploratory analysis of the data may guide the user toward the right approach to classification for their particular problem. Publishing the data provides the user a greater flexibility for data analysis. It should be noted that while data publishing techniques can be customized to provide better results for particular types of analysis [23], [24], [25], [26], data which is published towards a specific data mining goal can still be used for other data mining goals as well.

Mohammed et al. [27] also address the PPDP versus PPDM question. They provide additional arguments to support the necessity in publishing the data. First, the data custodian often has neither the expertise nor the interest in performing data mining. Second, it is unrealistic to assume that the data custodian could attend to repeated requests of the user to produce different types of statistical information and fine-tune the data mining results for research purposes.

In conclusion, PPDP is an essential paradigm that coexists alongside PPDM. Differential privacy is viable for PPDM, but it is still an open question if it can practically support PPDP. The syntactic notions of privacy are viable solutions for PPDP.

III. Criticisms of syntactic models of anonymity

Here we describe some of the main criticisms of syntactic models, and explain why they are a challenge for further research rather than a justified cause to abandon the models.

A. The deFinetti attack

The random worlds model [28] is commonly used to reason about attackers. According to that model, all tables with specific quasi-identifier values that are consistent with the published anonymized table are equally likely, and the adversary uses that assumption in order to draw from the anonymized table belief probabilities regarding the linkage between quasi-identifier tuples in the table and sensitive values. Based on that assumption, it is argued in [4] that anonymized tables that are ℓ-diverse prevent inferring belief probabilities that are larger than $1/\ell$.

In [29], Kifer showed that it is possible to extract from ℓ-diverse tables belief probabilities greater than $1/\ell$ by means of the so-called deFinetti attack. That attack uses the anonymized table in order to learn a classifier that, given the quasi-identifier tuple of an individual in the underlying population, is able to predict the corresponding sensitive value with probability greater than the intended $1/\ell$ bound.

There are three arguments why that attack is not a solid argument to abandon syntactic privacy models in favor of differential privacy. First, while Kifer showed that the effectiveness of the attack reduces with ℓ and its computational complexity grows dramatically, Cormode [30] found that even for small values of ℓ, the effectiveness of the attack diminishes substantially when the size of the ℓ-diverse blocks (k) grows. Second, Cormode showed that the deFinetti attack provides similar sensitive inference accuracy for both differentially private data and ℓ-diverse data. The third argument is more fundamental. The deFinetti attack relies on building a classifier based on the entire database. The question is whether the inference of a general behavior of the population in order to draw belief probabilities on individuals in that population constitutes a breach of privacy; differential privacy explicitly allows learning general behavior as long as it is not dependent on a single individual. To answer this question positively for

[1] http://www.census.gov/main/www/pums.html

an attack on privacy, the success of the attack when launched against records that *are* part of the table should be significantly higher than its success against records that *are not* part of the table. We are not aware of such a comparison for the deFinetti attack. Moreover, recent work showed experimentally that a classifier-type attack poses the same risk of sensitive inference for records in the training set as well as for records outside the training set [9]. This finding supports our claim that such an attack should not be regarded as a breach of privacy. It can only be regarded as a successful learning of the behavior of the general population, which is the raison d'être of any data publishing.

B. Minimality attacks

The minimality attack [31] exploits the knowledge of the anonymization algorithm in order to infer properties of the original data and, consequently, of individuals. An anonymized view of the original data induces a set of "possible worlds" for what the original data might have been. The knowledge of the anonymization algorithm and its decision making process enables, sometimes, to eliminate some of the possible worlds and, by thus, to increase the belief of the attacker in certain events to a level that is inconsistent with the desired privacy requirements. Cormode et al. [32] identified three safeguards that render syntactic anonymity algorithms immune against such attacks: randomness, a high degree of symmetry in the grouping of records, and decoupling of the quasi-identifiers and the sensitive attribute.

C. The curse of dimensionality

Aggarwal [33] showed that when the number of quasi-identifiers is large, most of the table entries have to be suppressed in order to achieve k-anonymity. Due to this so-called "curse of dimensionality", applying k-anonymity on high-dimensional data would significantly degrade the data quality. This is an essential problem, but it may be addressed within the framework of syntactic privacy.

As in real-life privacy attacks, it can be very difficult for an adversary to acquire complete background information on target individuals, Mohammed et al. [27] suggested the LKC-privacy model for anonymizing high-dimensional data. LKC-privacy bounds the probability of a successful identity linkage to be at most $1/K$ and the probability of a successful attribute linkage to be at most $1/C$, provided that the adversary's prior knowledge does not exceed L quasi-identifiers. Their experiments showed that this privacy notion can effectively retain the essential information in anonymous data needed for data analysis.

In that context, it is important to understand that not all non-sensitive attributes should be automatically classified as quasi-identifiers. The data custodian should assess the chances of an adversary to get hold of each of the attributes in the data schema. If the chance of an adversary to get hold of some attribute is smaller than the chance of acquiring the sensitive data, then there is no need to relate to such an attribute as a quasi-identifier.

Another important observation that mitigates the curse of dimensionality is that not all quasi-identifiers are needed for every data sharing purpose. Hence, instead of a single publication of the entire high-dimensional dataset, it is expected that the data will be published in several releases, where each release is an anonymization of a lower dimensional dataset which is a projection of the original dataset onto a subset of the attributes. Algorithms for anonymizing datasets that are released in this manner were proposed in [34], [35], [36].

IV. Criticisms of differential privacy

It is important to be clear about the claims of differential privacy. Differential privacy bounds the impact an individual has on the outcome (data mining model, or published dataset.) The main premise is that if knowledge can be gained without an individual's data, then that individual's privacy is not violated – even if the knowledge can be used to learn private information about the individual. This means that certain types of background knowledge (e.g., how far an individual deviates from the mean) can be used with a differentially private result to learn specific values about the individual without violating differential privacy; the promise of differential privacy is (by design) not absolute secrecy. Many of the criticisms of both syntactic anonymity and differential privacy (such as some background knowledge attacks) presume any disclosure of information about an individual is a violation; this cannot be achieved without entirely foregoing data utility. That said, differential privacy is a strong notion of privacy – but it still suffers from a number of practical problems and limitations.

A. Computing global sensitivity

Computing a realistic bound on the global sensitivity of multidimensional queries requires a very complex analysis of the domain of all possible tuples in the multidimensional space. For example, assessing the global sensitivity of queries that relate height and weight, based only on the ranges of each of those attributes, without taking into consideration their correlation, may give unreasonably high values; specifically, even though a typical range of heights includes the height 2 meters, and a typical range of weights includes the weight 3 kilograms, it would be devastating to add noise for calculating the body mass index for protecting against the possibility that the database includes a person with height 2 meters and weight 3 kilograms. Unrealistic sensitivity values give excessive noise, resulting in little utility from a differentially privacy result.

While specific types of queries may be amenable to specific techniques that do not pose these issues (e.g., the previously mentioned histogram queries), in general computing a global sensitivity that both guarantees privacy and provides usable levels of noise is a difficult task.

B. Non-compact uncertainty

Another problem with the applicability of differential privacy is the inherent uncertainty in the answer. In disciplines such as biostatistics or biomedical research, it is imperative to have known bounds on the value of the original data [37]. This

is the case with syntactic anonymization models, in which data is generalized according to accepted generalization rules. This is not the case with perturbation models in which the correlation between the original and perturbed data is probabilistic. Because of those reasons, syntactic privacy models, such as k-anonymity, are still perceived by practitioners as sufficient for mitigating risk in the real world while maximizing utility, and real life applications still utilize them for sanitizing data (see [38], [37]).

In addition to the inherent uncertainty in the answer, the quality of results obtained from a differentially private mechanism can vary greatly. Many of the positive results have been obtained using histogram-style queries on Boolean data. However, the differentially private mechanism of adding Laplacian noise can significantly alter the answer. An example is provided in [39]: a differentially private query for the mean income of a single U.S. county, with $\varepsilon = 0.25$ (resp. $\varepsilon = 1.0$), deviates from the true value by \$10,000 or less only 3% (resp. 12%) of the time! This can be extremely misleading, given that the true value is \$16,708. (This is a real-world example of a query with high-income outliers that cause high global sensitivity. In methods of syntactic anonymity, such outliers may only have local effect on records that were grouped with them in the same anonymity block.)

Wasserman and Zhou [40] show similar results for the differentially private histogram method of [13]; substantial error arises with smaller sample sizes. They also formally analyzed such accuracy variation for the exponential mechanism of [41]. They showed that the accuracy is linked to the rate at which the empirical distribution concentrates around the true distribution.

C. How to set ε?

The U.S. HIPAA "Safe Harbor" rules [12] specify legally acceptable syntactic anonymizations. Certain types of identifying information must be removed, and dates and locations have to be generalized to some extent: locations must be generalized into geographic units that have at least 20,000 residents; date of birth must be rounded up to the year of birth only (unless the age is 90 years or more, in which case wider ranges are required). A simple "back of the envelope" calculation yields the level k of anonymity that those rules induce. In differential privacy, on the other hand, little has been done to address the practically essential question of how to set the privacy parameter ε. While the definition of differential privacy clearly addresses the issue of identification (if it is hard to determine whether an individual is in the database, it is certainly hard to identify that individual's record in the database), the way in which ε affects the ability to identify an individual is not as clear. The parameter ε in ε-differential privacy is not a measure of privacy in the normal sense: it bounds the impact an individual has on the result, not what is disclosed about an individual. Queries that specifically ask information about an individual, e.g. "Is Bob in the database", are an exception. In such queries, ε directly relates to disclosure of information on that particular individual. However, for queries that ask more general properties of the data, the impact of ε on identifying an individual is less clear. As shown in [42], for a given setting of ε, the confidence an adversary can have that a specific individual is in the database can change depending on the query, values in the data, and even on values not in the data. While ε-differential privacy does adjust for changes in values in both the data and values outside the dataset (for example, both are incorporated in the calculation of the query's global sensitivity for queries that are based on a numeric function of the data values), this is not a direct measure of what is revealed about an individual.

This may not be an insurmountable problem; a *differential identifiability* approach that has much in common with differential privacy is given in [43]. In differential identifiability, the adversary model is essentially the same as in differential privacy. The key distinction is that the parametrization of the noise to be added is based on the posterior confidence an adversary, knowing the value of ε, can have about the inclusion of any specific individual in the database. This mechanism allows calibrating the added noise to enforce identifiability requirements such as those derived from the HIPAA safe harbor rules. Having said that, there are limitations and assumptions in the adversary model, such as the assumption of a uniform prior adversary belief in the presence of individuals in the database, that demand further research.

D. Independence assumption

Differential privacy makes some hidden assumptions that are not necessary in syntactic models. One such assumption is that individuals are independent. The problem becomes quite apparent with relational learning, where values of one individual can influence what is learned about another. When one individual can influence another (most obvious with social networks), what does it mean to calculate the sensitivity, or impact that one individual may have on the query's result? Suppose, for example, that we want to predict election results in a differentially private manner. While removing one individual from the dataset would seem to change only one vote, the effect on the prediction made by a relational learner may be significantly larger, depending on the social role of that individual. Indeed, removing an organizational leader from the database may change the prediction regarding the votes of many in that organization (just as removing one person from the real-world might change the voting results significantly). Hence, the social dependencies cannot be ignored as they may cause nearly unbounded (and very difficult to calculate) changes in the query's outcome. Achieving meaningful results from differential privacy may require assumptions on the model for data generation [44], new ways of defining what constitutes information about a single individual [45], or even entirely new privacy definitions [46], [47].

Syntactic models avoid this problem since in such models all individuals are anonymized, and the method of anonymization is independent of the social relations between the individuals.

E. Immunity to background knowledge

One of the main claims of differential privacy is that it is immune to attacks based on the adversary's background knowledge. In some cases this claim is not as strong as it might appear. An example is given in [39]: given relative background knowledge (an individual earns $5M more than the U.S. average), a differentially private query for the needed information (U.S. average income) can return quite accurate results – essentially violating the privacy of the rich individual. Hence, some background knowledge may allow an adversary to learn information on one individual from a differentially private answer that is computed from the values of other individuals.

V. Summary and conclusions

This study examined two types of privacy models: syntactic models of anonymity and differential privacy. The syntactic models are designed for privacy-preserving data *publishing* while differential privacy is typically applicable for privacy-preserving data *mining*. Hence, one approach cannot replace the other, and they both have a place alongside the other.

Next, we discussed criticisms of syntactic anonymization models and explained why none is a show stopper. Then, we proceeded to point out issues that need to be resolved with differential privacy. Our conclusion is that while differential privacy is a valuable weapon in the fight to both maintain privacy and foster use of data, it is not the universal answer. It provides a way to deal with a previously unanswered question in privacy-preserving data mining: how to ensure that the model developed does not inherently violate privacy of the individuals in the training data? While there are still issues related to both privacy and utility to be resolved, the basic concept is a strong one. At the same time, privacy-preserving data publishing remains a pertinent and essential notion. While privacy advocates may not like it, societal practice (and laws such as HIPAA and those mandated by [48]) recognize that the right to privacy must be balanced against the public good. Syntactic models substantially reduce privacy risk compared to a release of actual data values, and provide guarantees on the correctness of analysis of the anonymized data. In many cases, this is preferable to many noise addition techniques.

It should be clarified that the two paradigms are not necessarily exclusive: recent work by Li, Qardaji, and Su suggests a link [16]. By first randomly selecting a subset of the data, and then applying k-anonymization, they show that the resulting syntactic anonymization can be made consistent with (ε, δ)-differential privacy. A key point is that the k-anonymization algorithm must introduce some random variability in the anonymization *process* (as recommended by [32], see Section III-B). In particular, the generalization function must be developed using an ε-differentially private mechanism. They do require a slight relaxation of the background knowledge available to the attacker (see more details in [16]). A research challenge for PPDP is privacy definitions with adversary models that capture issues such as data correlation and inherently control potential real-world problems such as the deFinetti and minimality attacks.

In conclusion, in both paradigms, the issues raised should be viewed as opportunities for future research, rather than a call for abandoning one approach or the other. Advances in both paradigms are needed to ensure that the future provides reasonable protections on privacy as well as supporting legitimate learning from the ever-increasing data about us.

References

[1] R. Agrawal and R. Srikant, "Privacy-preserving data mining," in *SIGMOD Conference*, 2000, pp. 439–450.

[2] P. Samarati, "Protecting respondents' identities in microdata release," *IEEE TKDE*, vol. 13, pp. 1010–1027, 2001.

[3] P. Samarati and L. Sweeney, "Generalizing data to provide anonymity when disclosing information (abstract)," in *PODS*, 1998, p. 188.

[4] A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam, "ℓ-diversity: Privacy beyond k-anonymity," *ACM Transactons on Knowledge Discovery from Data*, vol. 1, p. 3, 2007.

[5] N. Li, T. Li, and S. Venkatasubramanian, "t-closeness: Privacy beyond k-anonymity and ℓ-diversity," in *ICDE*, 2007, pp. 106–115.

[6] J. Cao and P. Karras, "Publishing microdata with a robust privacy guarantee," *PVLDB*, vol. 5, pp. 1388–1399, 2012.

[7] A. Gionis, A. Mazza, and T. Tassa, "k-anonymization revisited," in *ICDE*, 2008, pp. 744–753.

[8] T. Tassa, A. Mazza, and A. Gionis, "k-concealment: An alternative model of k-type anonymity," *Transactions on Data Privacy*, vol. 5, pp. 189–222, 2012.

[9] M. Last, T. Tassa, A. Zhmudyak, and E. Shmueli, "Improving accuracy of classification models induced from anonymized datasets," *Submitted*.

[10] W. K. Wong, N. Mamoulis, and D. W.-L. Cheung, "Non-homogeneous generalization in privacy preserving data publishing," in *SIGMOD Conference*, 2010, pp. 747–758.

[11] C. Dwork, "Differential privacy," in *ICALP (2)*, 2006, pp. 1–12.

[12] "Standard for privacy of individually identifiable health information," *Federal Register*, vol. 67, no. 157, pp. 53 181–53 273, Aug. 14 2002. [Online]. Available: http://www.hhs.gov/ocr/privacy/hipaa/administrative/privacyrule/index.html

[13] C. Dwork, F. McSherry, K. Nissim, and A. Smith, "Calibrating noise to sensitivity in private data analysis," in *TCC*, 2006, pp. 265–284.

[14] A. Narayanan and V. Shmatikov, "Myths and fallacies of personally identifiable information," *Comm. of the ACM*, vol. 53, pp. 24–26, 2010.

[15] R. Moore, "Controlled data-swapping techniques for masking public use microdata sets," *Statistical Research Division Report Series RR 96-04*, U.S. Bureau of the Census, Washington DC, 1996.

[16] N. Li, W. Qardaji, and D. Su, "On sampling, anonymization, and differential privacy: Or, k-anonymization meets differential privacy," in *7th ACM Symposium on Information, Computer and Communications Security (ASIACCS'2012)*, Seoul, Korea, May 2-4 2012.

[17] A. Blum, K. Ligett, and A. Roth, "A learning theory approach to non-interactive database privacy," in *STOC*, 2008, pp. 609–618.

[18] C. Dwork, M. Naor, O. Reingold, G. N. Rothblum, and S. P. Vadhan, "On the complexity of differentially private data release: efficient algorithms and hardness results," in *STOC*, 2009, pp. 381–390.

[19] M. Hardt and G. N. Rothblum, "A multiplicative weights mechanism for privacy-preserving data analysis," in *FOCS*, 2010, pp. 61–70.

[20] A. Roth and T. Roughgarden, "Interactive privacy via the median mechanism," in *STOC*, 2010, pp. 765–774.

[21] R. Chen, N. Mohammed, B. C. M. Fung, B. C. Desai, and L. Xiong, "Publishing set-valued data via differential privacy," *PVLDB*, vol. 4, no. 11, pp. 1087–1098, 2011.

[22] B. C. M. Fung, K. Wang, and P. S. Yu, "Anonymizing classification data for privacy preservation," *IEEE TKDE*, vol. 19, pp. 711–725, 2007.

[23] V. Iyengar, "Transforming data to satisfy privacy constraints," in *Proc., the Eigth ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining*, 2002, pp. 279–288. [Online]. Available: http://doi.acm.org/10.1145/775047.775089

[24] M. Nergiz and C. Clifton, "Thoughts on k-anonymization," *Data Knowl. Eng.*, vol. 63, no. 3, pp. 622–645, 2007.

[25] J. Goldberger and T. Tassa, "Efficient anonymizations with enhanced utility," *TDP*, vol. 3, pp. 149–175, 2010.

[26] K. LeFevre, D. J. DeWitt, and R. Ramakrishnan, "Workload-aware anonymization techniques for large-scale datasets," *ACM Trans. Database Syst.*, vol. 33, no. 3, pp. 17:1–17:47, 2008. [Online]. Available: http://doi.acm.org/10.1145/1386118.1386123
[27] N. Mohammed, B. C. M. Fung, P. C. K. Hung, and C. kwong Lee, "Anonymizing healthcare data: a case study on the blood transfusion service," in *KDD*, 2009, pp. 1285–1294.
[28] F. Bacchus, A. J. Grove, D. Koller, and J. Y. Halpern, "From statistics to beliefs," in *AAAI*, 1992, pp. 602–608.
[29] D. Kifer, "Attacks on privacy and definetti's theorem," in *SIGMOD Conference*, 2009, pp. 127–138.
[30] G. Cormode, "Personal privacy vs population privacy: learning to attack anonymization," in *KDD*, 2011, pp. 1253–1261.
[31] R. C.-W. Wong, A. W.-C. Fu, K. Wang, and J. Pei, "Minimality attack in privacy preserving data publishing," in *VLDB*, 2007, pp. 543–554.
[32] G. Cormode, N. Li, T. Li, and D. Srivastava, "Minimizing minimality and maximizing utility: Analyzing method-based attacks on anonymized data," *PVLDB*, vol. 3, pp. 1045–1056, 2010.
[33] C. C. Aggarwal, "On k-anonymity and the curse of dimensionality," in *VLDB*, 2005, pp. 901–909.
[34] E. Shmueli and T. Tassa, "Privacy by diversity in sequential releases of databases," *Submitted*.
[35] E. Shmueli, T. Tassa, R. Wasserstein, B. Shapira, and L. Rokach, "Limiting disclosure of sensitive data in sequential releases of databases," *Inf. Sci.*, vol. 191, pp. 98–127, 2012.
[36] K. Wang and B. Fung, "Anonymizing sequential release," in *KDD*, 2006, pp. 414–423.
[37] K. E. Emam, F. K. Dankar, R. Issa, E. Jonker, D. Amyot, E. Cogo, J.-P. Corriveau, M. Walker, S. Chowdhury, R. Vaillancourt, T. Roffey, and J. Bottomley, "A globally optimal k-anonymity method for the de-identification of health information," *Journal of the American Medical Informatics Association*, vol. 16, pp. 670–682, 2009.
[38] K. E. Emam and F. Dankar, "Protecting privacy using k-anonymity," *Journal of the American Medical Informatics Association*, vol. 15, pp. 627 – 637, 2008.
[39] R. Sarathy and K. Muralidhar, "Evaluating laplace noise addition to satisfy differential privacy for numeric data," *Transactions on Data Privacy*, vol. 4, pp. 1–17, 2011.
[40] L. Wasserman and S. Zhou, "A statistical framework for differential privacy," *J. Amer. Stat. Assoc.*, vol. 105, pp. 375–389, 2010.
[41] F. McSherry and K. Talwar, "Mechanism design via differential privacy," in *FOCS*, 2007, pp. 94–103.
[42] J. Lee and C. Clifton, "How much is enough? choosing ϵ for differential privacy," in *ISC*, 2011, pp. 325–340.
[43] ——, "Differential identifiability," in *KDD*, 2012, pp. 1041–1049.
[44] D. Kifer and A. Machanavajjhala, "No free lunch in data privacy," in *SIGMOD*, 2011, pp. 193–204. [Online]. Available: http://doi.acm.org/10.1145/1989323.1989345
[45] C. Task and C. Clifton, "A guide to differential privacy theory in social network analysis," in *ASONAM*, 2012.
[46] J. Gehrke, E. Lui, and R. Pass, "Towards privacy for social networks: A zero-knowledge based definition of privacy," in *Theory of Cryptography Conference*, 2011, pp. 432–449. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-19571-6_26
[47] D. Kifer and A. Machanavajjhala, "A rigorous and customizable framework for privacy," in *PODS*, Scottsdale, Arizona, May 21-23 2012, pp. 77–88. [Online]. Available: http://doi.acm.org/10.1145/2213556.2213571
[48] "Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data," *Official Journal of the European Communities*, vol. I, no. 281, pp. 31–50, 1995.