A Note on Data Privacy: Past, Present, and Future

Wentao Wu
Department of Computer Sciences
University of Wisconsin-Madison
wentaowu@cs.wisc.edu

The problem of data privacy has been studied for years. In this short paper, we provide a survey of important work in the history. Given that data privacy is still an active and evolving area, we do not try to be exhaustive in this paper. Rather, we hope the results surveyed in this paper can serve as a reference point for future research in this field.

1 k-Anonymity And Its Variants

The problem of data privacy perhaps goes back to 2000 when Rakesh Agrawal and Ramakrishnan Srikant published their seminal paper on privacy-preserving data mining [1]. The most influential data privacy model in the next few years was k-anonymity [2, 3]. Basically, k-anonymity captures the intuition of "syntactic indistinguishability": each tuple in the database is guaranteed to have at least k counterparts that have exactly the same values on the *quasi-identifiers*, i.e., attributes that can uniquely determine personal identity when combined together. Quasi-identifiers can be leveraged by *linkage attacks*, which join two different databases (about the same group of individuals) to associate sensitive attributes with the personal identifiers via the bridge of quasi-identifiers. Therefore, k-anonymity is an effective way of preventing linkage attacks. It then triggered extensive follow-up research that introduced various enhancements (e.g., l-diversity [4], m-invariance [5], t-closeness [6]) or extensions (e.g., the Mondrian model that extends k-anonymity to multi-dimensional space [7]).

The main problem of k-anonymity is that it is not clear what its privacy implication is. In other words, what kind of "privacy" guarantee it tries to offer? The line of k-anonymity research followed a "break-it-then-make-it" style: some new attack was identified that can break the state of the art (e.g., the minimality attack [8], the composition attack [9], and the attack based on deFinetti's theorem [10]), then a new, enhanced version of k-anonymity was proposed to resist this new attack. This endless loop came to an end when differential privacy arose, as we discuss in the next section.

2 Differential Privacy

Cynthia Dwork proposed a different privacy notion called differential privacy in 2006 [11]. The recent monograph [12] provides a comprehensive study of differential privacy. The original motivation of differential privacy was to provide a negative answer to the privacy notion based on *semantic security* [13, 14]: access to a statistical database should not enable one to learn anything about an individual that could not be learned without access. Put it another way, an outsider should learn the *same* information about any individual with/without access to the database. This impossibility result is due to the existence of *auxiliary information*. For example, suppose that knowing the height of a person compromises his/her privacy. A database that reports average heights of Americans looks completely harmless but will actually cause privacy leakage if an adversary has background knowledge such as "X's height is 2 inches above the average." But why is semantic security achievable in cryptosystems? This discrepancy is because of the different setting here. The database is publicly available, and there is no way to distinguish a normal user from an adversary: they see the same information disclosed by the database but they learn very different information because they have different pieces of background information. This obstacle is overcome in cryptosystems by using encryption/decryption: only trusted users have keys to decrypt the information; adversaries can only see encrypted text and therefore learn nothing. This formulation therefore leads to a natural question that was even untouched by k-anonymity work: what should be the goal of privacy protection? Or, what kind of "privacy" guarantee is achievable?

The success of differential privacy, in my option, is due to that it provides a reasonable answer to the above question, perhaps the first time in the history. In a nutshell, differential privacy formulates privacy as something based on the notion of *participation*: an outsider should learn the *same* information about any individual with/without his/her participation in the database. Note that this is weaker than the aforementioned privacy guarantee based on semantic security: an outsider can certainly learn more about an individual given access to the database, but what he learnt is independent of if that individual is in the database or not. Go back to our previous example on human heights: differential privacy cannot avoid this kind of privacy leakage — the average height is conceivably the same regardless of whether X is present in the database or not. However, by admitting that differential privacy does not try to cover all kinds of "privacy," it lays itself on a solid mathematical foundation by formulating its privacy guarantee in a probabilistic manner. A simple perturbation mechanism based on Laplacian noise was further proposed to ensure differential privacy [15]. There were later on other mechanisms such as the geometric mechanism [16] and the exponential mechanism [17], and there were also variants on relaxed versions of ϵ-differential privacy, prominently, (ϵ, δ)-differential privacy [18]. It remains an active area today on studying either differential privacy itself or its applications in other fields, notably, data mining and machine learning. The interaction between differential privacy and learning theory is an intriguing area that calls for future research.

3 Beyond Differential Privacy

There is a recent move towards addressing privacy issues that are beyond the framework of differential privacy. Notably, there is a line of work trying to model privacy leakage with Bayesian formulations by capturing the difference between the prior and posterior learnt from the database [19, 20, 21, 22, 23, 24]. While this seems to be interesting work, it somehow falls back to the question that perplexed the privacy research community for years before the emergence of differential privacy: modeling "privacy" by modeling background knowledge (i.e., prior knowledge). There were indeed some attempts even in the k-anonymity era towards this direction. For example, Rastogi, Suciu, and Hong [25] showed that there is some kind of dichotomy based on how much prior knowledge an adversary possesses: if the prior belief about some individual is large then there is no useful anonymization algorithm; if the prior belief is bounded for all individuals then there exists an algorithm that is both private and useful. There were also proposals based on using formal, logic-based languages to represent background knowledge, such as worst-case background knowledge [26] and privacy skylines [27, 28]. Tailored privacy mechanisms were then designed to ensure privacy under the assumption that an adversary could only have background knowledge expressible in these languages. However, this effort towards modeling background knowledge seems having limited success. Indeed, one important motivation of differential privacy is right the difficulty in formally modeling and reasoning background knowledge. The recent move to modeling prior/posterior with Bayesian approaches therefore might be just another round of repeating history. Nonetheless, there is still some interesting work in characterizing the "semantics" of differential privacy from a Bayesian perspective [29]. It basically says that, under ϵ-differential privacy, the posteriors over two neighboring databases are close, regardless of the prior owned by an adversary. However, this does not necessarily hold if only (ϵ, δ)-differential privacy is achieved.

Finally, while there has been tons of theoretic work on differential privacy, its application in the practice is still limited [30, 31]. There was work in both the database and the theory research communities targeting at various data publishing tasks using differential privacy, including contingency tables [32], data cubes [33], histograms [34], boolean conjunctions [35], among others. Recently, Google reported that they have developed a differential-privacy based software called RAPPOR [36] that collected sensitive information from Chrome users, which seems to be the first reported application of differential privacy in industrial products. RAPPOR used localized rather than centralized mechanism, though [12]. On the other hand, there is also a recent line of study that tries to combine differential privacy with k-anonymity as practical privacy solutions (e.g., [37, 38, 39]). This interaction between old privacy techniques and differential privacy deserves further research effort. The recent paper [40] is an excellent survey on comparing k-anonymity and differential privacy from a practitioner's perspective.

References

[1] R. Agrawal and R. Srikant, “Privacy-preserving data mining,” in *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data, May 16-18, 2000, Dallas, Texas, USA.*, 2000, pp. 439–450.

[2] L. Sweeney, “k-anonymity: A model for protecting privacy,” *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 10, no. 5, pp. 557–570, 2002.

[3] ——, “Achieving k-anonymity privacy protection using generalization and suppression,” *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 10, no. 5, pp. 571–588, 2002.

[4] A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam, “l-diversity: Privacy beyond k-anonymity,” in *ICDE*, 2006, p. 24.

[5] X. Xiao and Y. Tao, “M-invariance: towards privacy preserving re-publication of dynamic datasets,” in *SIGMOD Conference*, 2007, pp. 689–700.

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

[7] K. LeFevre, D. J. DeWitt, and R. Ramakrishnan, “Mondrian multidimensional k-anonymity,” in *ICDE*, 2006, p. 25.

[8] R. C. Wong, A. W. Fu, K. Wang, and J. Pei, “Minimality attack in privacy preserving data publishing,” in *Proceedings of the 33rd International Conference on Very Large Data Bases, University of Vienna, Austria, September 23-27, 2007*, 2007, pp. 543–554.

[9] S. R. Ganta, S. P. Kasiviswanathan, and A. Smith, “Composition attacks and auxiliary information in data privacy,” in *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Las Vegas, Nevada, USA, August 24-27, 2008*, 2008, pp. 265–273.

[10] D. Kifer, “Attacks on privacy and definetti’s theorem,” in *Proceedings of the ACM SIGMOD International Conference on Management of Data, SIGMOD 2009, Providence, Rhode Island, USA, June 29 - July 2, 2009*, 2009, pp. 127–138.

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

[12] C. Dwork and A. Roth, “The algorithmic foundations of differential privacy,” *Foundations and Trends in Theoretical Computer Science*, vol. 9, no. 3-4, pp. 211–407, 2014.

[13] T. Dalenius, “Towards a methodology for statistical disclosure control,” *Statistik Tidskrift*, vol. 15, no. 429-444, pp. 2–1, 1977.

[14] S. Goldwasser and S. Micali, “Probabilistic encryption,” *J. Comput. Syst. Sci.*, vol. 28, no. 2, pp. 270–299, 1984.

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

[16] A. Ghosh, T. Roughgarden, and M. Sundararajan, “Universally utility-maximizing privacy mechanisms,” *SIAM J. Comput.*, vol. 41, no. 6, pp. 1673–1693, 2012.

[17] F. McSherry and K. Talwar, “Mechanism design via differential privacy,” in *48th Annual IEEE Symposium on Foundations of Computer Science (FOCS 2007), October 20-23, 2007, Providence, RI, USA, Proceedings*, 2007, pp. 94–103.

[18] C. Dwork, K. Kenthapadi, F. McSherry, I. Mironov, and M. Naor, “Our data, ourselves: Privacy via distributed noise generation,” in *Advances in Cryptology - EUROCRYPT 2006, 25th Annual International Conference on the Theory and Applications of Cryptographic Techniques, St. Petersburg, Russia, May 28 - June 1, 2006, Proceedings*, 2006, pp. 486–503.

[19] D. Kifer and B. Lin, “Towards an axiomatization of statistical privacy and utility,” in *Proceedings of the Twenty-Ninth ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems, PODS 2010, June 6-11, 2010, Indianapolis, Indiana, USA*, 2010, pp. 147–158.

[20] D. Kifer and A. Machanavajjhala, “No free lunch in data privacy,” in *SIGMOD Conference*, 2011, pp. 193–204.

[21] ——, “A rigorous and customizable framework for privacy,” in *Proceedings of the 31st ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems, PODS 2012, Scottsdale, AZ, USA, May 20-24, 2012*, 2012, pp. 77–88.

[22] D. Kifer and B.-R. Lin, “An axiomatic view of statistical privacy and utility,” *Journal of Privacy and Confidentiality*, vol. 4, no. 1, p. 2, 2012.

[23] B. Lin and D. Kifer, “Information preservation in statistical privacy and bayesian estimation of unattributed histograms,” in *Proceedings of the ACM SIGMOD International Conference on Management of Data, SIGMOD 2013, New York, NY, USA, June 22-27, 2013*, 2013, pp. 677–688.

[24] B.-R. Lin and D. Kifer, “Towards a systematic analysis of privacy definitions,” *Journal of Privacy and Confidentiality*, vol. 5, no. 2, p. 2, 2014.

[25] V. Rastogi, S. Hong, and D. Suciu, “The boundary between privacy and utility in data publishing,” in *Proceedings of the 33rd International Conference on Very Large Data Bases, University of Vienna, Austria, September 23-27, 2007*, 2007, pp. 531–542.

[26] D. J. Martin, D. Kifer, A. Machanavajjhala, J. Gehrke, and J. Y. Halpern, “Worst-case background knowledge for privacy-preserving data publishing,” in *Proceedings of the 23rd International Conference on Data Engineering, ICDE 2007, The Marmara Hotel, Istanbul, Turkey, April 15-20, 2007*, 2007, pp. 126–135.

[27] B.-C. Chen, R. Ramakrishnan, and K. LeFevre, “Privacy skyline: Privacy with multidimensional adversarial knowledge,” in *VLDB*, 2007, pp. 770–781.

[28] B. Chen, K. LeFevre, and R. Ramakrishnan, “Adversarial-knowledge dimensions in data privacy,” *VLDB J.*, vol. 18, no. 2, pp. 429–467, 2009.

[29] S. P. Kasiviswanathan and A. Smith, “On the ‘semantics’ of differential privacy: A bayesian formulation,” *CoRR*, vol. abs/0803.3946, 2008.

[30] C. Dwork, “Differential privacy: A survey of results,” in *Theory and Applications of Models of Computation, 5th International Conference, TAMC 2008, Xi’an, China, April 25-29, 2008. Proceedings*, 2008, pp. 1–19.

[31] C. Dwork and R. Pottenger, “Toward practicing privacy,” *JAMIA*, vol. 20, no. 1, pp. 102–108, 2013.

[32] B. Barak, K. Chaudhuri, C. Dwork, S. Kale, F. McSherry, and K. Talwar, “Privacy, accuracy, and consistency too: a holistic solution to contingency table release,” in *Proceedings of the Twenty-Sixth ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems, June 11-13, 2007, Beijing, China*, 2007, pp. 273–282.

[33] B. Ding, M. Winslett, J. Han, and Z. Li, “Differentially private data cubes: optimizing noise sources and consistency,” in *SIGMOD Conference*, 2011, pp. 217–228.

[34] J. Thaler, J. Ullman, and S. P. Vadhan, “Faster algorithms for privately releasing marginals,” in *ICALP (1)*, 2012, pp. 810–821.

[35] A. Gupta, M. Hardt, A. Roth, and J. Ullman, “Privately releasing conjunctions and the statistical query barrier,” *SIAM J. Comput.*, vol. 42, no. 4, pp. 1494–1520, 2013.

[36] Ú. Erlingsson, V. Pihur, and A. Korolova, “RAPPOR: randomized aggregatable privacy-preserving ordinal response,” in *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security, Scottsdale, AZ, USA, November 3-7, 2014*, 2014, pp. 1054–1067.

[37] N. Li, W. H. Qardaji, and D. Su, “On sampling, anonymization, and differential privacy or, *k*-anonymization meets differential privacy,” in *7th ACM Symposium on Information, Compuer and Communications Security, ASIACCS ’12, Seoul, Korea, May 2-4, 2012*, 2012, pp. 32–33.

[38] G. Kellaris and S. Papadopoulos, “Practical differential privacy via grouping and smoothing,” *PVLDB*, vol. 6, no. 5, pp. 301–312, 2013.

[39] J. Soria-Comas, J. Domingo-Ferrer, D. Sanchez, and S. Martinez, “Enhancing data utility in differential privacy via microaggregation-based k-anonymity,” *The VLDB Journal*, vol. 23, no. 5, pp. 771–794, 2014.

[40] C. Clifton and T. Tassa, “On syntactic anonymity and differential privacy,” in *ICDE Workshops*, 2013, pp. 88–93.