基于数据立方体的方法[68]、基于多权重机制的方法[69,70]和基于学习模型的方法[71]等.然而,当属性个数较多时,即对于数据维度较高的数据集,现有的方法将遇到瓶颈[72,73],因此,数据的高维度给差分隐私技术带来极大的挑战.对于本地化差分隐私保护技术而言,高维数据不仅带来数据规模变大和信噪比降低两个方面的影响,而且还将增加通信代价.并且,通信代价根据不同的扰动机制随数据维度的增加呈线性增长或指数增长,而通信代价的增加将直接为本地化差分隐私技术的应用带来限制.

目前,针对高维数据的发布,主要是利用属性划分的思想,在满足本地化差分隐私的基础上,将高维数据的联合概率分布分解为多个低维的边缘概率分布的形式,以多个边缘概率通过某种推理机制近似估计联合概率分布.其中的关键步骤是对两两属性之间的关联性进行判断.当数据中存在 d 个属性时,对应的关联性存在 $\binom{d}{2}$ 种,这就意味着需要把有限的隐私预算进行 $\binom{d}{2}$ 次分割,高维数据下,这势必带来很大的噪声,使得推理结果的准确性大大降低.因此,还需要辅以其他的方式进行降维,如对属性进行聚类或分组等.目前仅有文献[47,48]讨论了本地化差分隐私下的高维数据发布问题,但其只考虑了数据收集者如何依据属性之间的相关性进行降维处理,因此高维数据下通信代价问题依旧存在.为了降低高维数据带来的通信代价,现有方法一般是在不同维度上利用采样技术进行降维[20,21],然而,采样技术依然不可避免地导致数据可用性的下降.因此,在该类问题上,还需考虑通信代价和数据可用性之间的平衡关系.

综上所述,我们认为,本地化差分隐私下的高维数据发布主要考虑 3 个方面的问题:(1) 如何在一定隐私预算内衡量属性之间的关联性,从而进行降维处理;(2) 如何设计推理模型,最小化边缘分布到联合分布的近似误差,提高数据可用性;(3) 如何控制高维数据在用户和数据收集者之间的通信代价.

6 结束语

大数据时代个人数据高度敏感,如何防止隐私信息泄露是当前面临的重大挑战.本地化差分隐私是继中心化差分隐私后新兴的隐私保护模型,其打破了中心化差分隐私中关于可信第三方数据收集者的假设,在用户端对数据进行隐私化处理.目前,本地化差分隐私保护技术是隐私保护领域的研究热点,本文对其研究成果进行总结和分析,综述了本地化差分隐私保护技术的研究现状,总结该技术在频数统计和均值统计中的应用,并进行实验特性分析.最后,本文就现有研究工作和现实需求进行探讨,结合二者提出未来研究挑战.总之,本地化差分隐私保护技术还是一个新兴研究领域,仍有诸多关键问题需要进行深入而细致的研究.

References:

[1] Dwork C. Differential privacy. In: Proc. of the ICALP. 2006. 1–12.

[2] Dwork C, Lei J. Differential privacy and robust statistics. In: Proc. of the 41st Annual ACM Symp. on Theory of Computing. ACM, 2009. 371–380. [doi: 10.1145/1536414.1536466]

[3] Smith A. Privacy-Preserving statistical estimation with optimal convergence rates. In: Proc. of the 43rd Annual ACM Symp. on Theory of Computing. ACM, 2011. 813–822. [doi: 10.1145/1993636.1993743]

[4] Samarati P, Sweeney L. Generalizing data to provide anonymity when disclosing information. PODS, 1998,98:188.

[5] Machanavajjhala A, Kifer D, Gehrke J, Kifer D, Venkitasubramaniam M. *l*-Diversity: Privacy beyond *k*-anonymity. ACM Trans. on Knowledge Discovery from Data (TKDD), 2007,1(1):3. [doi: 10.1109/ICDE.2006.1]

[6] Li N, Li T, Venkatasubramanian S. *t*-Closeness: Privacy beyond *k*-anonymity and *l*-diversity. In: Proc. of the 23rd IEEE Int'l Conf. on Data Engineering, ICDE 2007. IEEE, 2007. 106–115. [doi: 10.1109/ICDE.2007.367856]

[7] Kasiviswanathan SP, Lee HK, Nissim K, Raskhodnikova S, Smith A. What can we learn privately. In: Proc. of the 49th Annual IEEE Symp. on Foundations of Computer Science (FOCS). IEEE, 2008. 531–540.

[8] Duchi JC, Jordan MI, Wainwright MJ. Local privacy and statistical minimax rates. In: Proc. of the 54th Annual IEEE Symp. on Foundations of Computer Science (FOCS). IEEE, 2013. 429–438. [doi: 10.1109/FOCS.2013.53]

[9] Erlingsson Ú, Pihur V, Korolova A. Rappor: Randomized aggregatable privacy-preserving ordinal response. In: Proc. of the 2014 ACM SIGSAC Conf. on Computer and Communications Security. ACM, 2014. 1054–1067. [doi: 10.1145/2660267.2660348]

[10] Howe J. Crowdsourcing: How the Power of the Crowd is Driving the Future of Business. Random House, 2008.

[11] Li G, Wang J, Zheng Y, Franklin MJ. Crowdsourced data management: A survey. IEEE Trans. on Knowledge and Data Engineering, 2016,28(9):2296–2319. [doi: 10.1109/TKDE.2016.2535242]

[12] Wu S, Wang X, Wang S, Zhang Z, Tung AK. *K*-Anonymity for crowdsourcing database. IEEE Trans. on Knowledge and Data Engineering, 2014,26(9):2207–2221. [doi: 10.1109/TKDE.2013.93]

[13] Varshney LR, Vempaty A, Varshney PK. Assuring privacy and reliability in crowdsourcing with coding. In: Proc. of the Information Theory and Applications Workshop (ITA). IEEE, 2014. 1–6. [doi: 10.1109/ITA.2014.6804213]

[14] Mao J, Jain AK. Artificial neural networks for feature extraction and multivariate data projection. IEEE Trans. on Neural Networks, 1995,6(2):296–317. [doi: 10.1109/72.363467]

[15] Finn RL, Wright D, Friedewald M. Seven types of privacy. In: European Data Protection: Coming of Age. Springer Netherlands, 2013. 3–32.

[16] Erkin Z, Franz M, Guajardo J, Katzenbeisser S, Lagendijk I, Toft T. Privacy-Preserving face recognition. In: Proc. of the Int'l Symp. on Privacy Enhancing Technologies Symp. Berlin, Heidelberg: Springer-Verlag, 2009. 235–253.

[17] Qin Z, Yan J, Ren K, Chen CW, Wang C. Towards efficient privacy-preserving image feature extraction in cloud computing. In: Proc. of the 22nd ACM Int'l Conf. on Multimedia. ACM, 2014. 497–506. [doi: 10.1145/2647868.2654941]

[18] Ren K. Privacy-Preserving image processing in cloud computing. Chinese Journal of Network and Information Security, 2016,(1): 12–17 (in Chinese with English abstract).

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

[20] Qin Z, Yang Y, Yu T, Khalil I, Xiao X, Ren K. Heavy hitter estimation over set-valued data with local differential privacy. In: Proc. of the 2016 ACM SIGSAC Conf. on Computer and Communications Security. ACM, 2016. 192–203. [doi: 10.1145/2976749. 2978409]

[21] Bassily R, Smith A. Local, private, efficient protocols for succinct histograms. In: Proc. of the 47th Annual ACM on Symp. on Theory of Computing. ACM, 2015. 127–135. [doi: 10.1145/2746539.2746632]

[22] Kairouz P, Oh S, Viswanath P. Extremal mechanisms for local differential privacy. In: Advances in Neural Information Processing Systems. 2014. 2879–2887.

[23] Kairouz P, Bonawitz K, Ramage D. Discrete distribution estimation under local privacy. In: Proc. of the 33rd Int'l Conf. on Machine Learning. New York, 2016. 2436–2444.

[24] Duchi JC, Jordan MI, Wainwright MJ. Local privacy, data processing inequalities, and statistical minimax rates. arXiv Preprint arXiv:1302.3203, 2013.

[25] Wainwright MJ, Jordan MI, Duchi JC. Privacy aware learning. In: Advances in Neural Information Processing Systems. 2012. 1430–1438.

[26] Nguyên TT, Xiao X, Yang Y, Hui SC, Shin H, Shin J. Collecting and analyzing data from smart device users with local differential privacy. arXiv Preprint arXiv:1606.05053, 2016.

[27] McSherry FD. Privacy integrated queries: An extensible platform for privacy-preserving data analysis. In: Proc. of the 2009 ACM SIGMOD Int'l Conf. on Management of Data. ACM, 2009. 19–30. [doi: 10.1145/1559845.1559850]

[28] Dwork C, McSherry F, Nissim K, Smith A. Calibrating noise to sensitivity in private data analysis. In: Proc. of the Theory of Cryptography Conf. Berlin, Heidelberg: Springer-Verlag, 2006. 265–284.

[29] McSherry F, Talwar K. Mechanism design via differential privacy. In: Proc. of the 48th Annual IEEE Symp. on Foundations of Computer Science (FOCS). IEEE, 2007. 94–103. [doi: 10.1109/FOCS.2007.66]

[30] Wang Q, Zhang Y, Lu X, Wang Z, Qin Z, Ren, K. Real-Time and spatio-temporal crowd-sourced social network data publishing with differential privacy. IEEE Trans. on Dependable and Secure Computing, 2016. [doi: 10.1109/TDSC.2016.2599873]

[31] Zhang X, Chen R, Xu J, Meng X, Xie Y. Towards accurate histogram publication under differential privacy. In: Proc. of the 2014 SIAM Int'l Conf. on Data Mining. Society for Industrial and Applied Mathematics, 2014. 587–595.

[32] Su S, Tang P, Cheng X, Chen R, Wu Z. Differentially private multi-party high-dimensional data publishing. In: Proc. of the 32nd IEEE Int'l Conf. on Data Engineering (ICDE). IEEE, 2016. 205–216. [doi: 10.1109/ICDE.2016.7498241]

[33] Yaroslavtsev G, Cormode G, Procopiuc CM, Srivastava D. Accurate and efficient private release of datacubes and contingency tables. In: Proc. of the 29th IEEE Int'l Conf. on Data Engineering (ICDE). IEEE, 2013. 745–756. [doi: 10.1109/ICDE.2013.6544871]

[34] Zhang T, Zhu Q. Dynamic differential privacy for ADMM-based distributed classification learning. IEEE Trans. on Information Forensics and Security, 2017,12(1):172–187. [doi: 10.1109/TIFS.2016.2607691]

[35] Abadi M, Chu A, Goodfellow I, McMahan H B, Mironov I, Talwar K, Zhang L. Deep learning with differential privacy. In: Proc. of the 2016 ACM SIGSAC Conf. on Computer and Communications Security. ACM, 2016. 308–318. [doi: 10.1145/2976749.2978318]

[36] Bun M, Steinke T, Ullman J. Make up your mind: The price of online queries in differential privacy. In: Proc. of the 28th Annual ACM-SIAM Symp. on Discrete Algorithms. Society for Industrial and Applied Mathematics, 2017. 1306–1325.

[37] Yuan G, Yang Y, Zhang Z, Hao Z. Convex optimization for linear query processing under approximate differential privacy. In: Proc. of the 22nd ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. ACM, 2016. 2005–2014. [doi: 10.1145/2939672.2939818].

[38] Chen R, Li H, Qin AK, Kasiviswanathan SP, Jin H. Private spatial data aggregation in the local setting. In: Proc. of the 32nd IEEE Int'l Conf. on Data Engineering (ICDE). IEEE, 2016. 289–300. [doi: 10.1109/ICDE.2016.7498248]

[39] Zhang X, Meng X. Differential privacy in data publication and analysis. Chinese Journal of Computers, 2014,(4):927–949 (in Chinese with English abstract).

[40] Kifer D, Machanavajjhala A. No free lunch in data privacy. In: Proc. of the 2011 ACM SIGMOD Int'l Conf. on Management of data. ACM, 2011. 193–204. [doi: 10.1145/1989323.1989345]

[41] Yang B, Sato I, Nakagawa H. Bayesian differential privacy on correlated data. In: Proc. of the 2015 ACM SIGMOD Int'l Conf. on Management of Data. ACM, 2015. 747–762. [doi: 10.1145/2723372.2747643]

[42] Xiong S, Sarwate AD, Mandayam NB. Randomized requantization with local differential privacy. In: Proc. of the 2016 IEEE Int'l Conf. on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2016. 2189–2193. [doi: 10.1109/ICASSP.2016.7472065]

[43] Sarwate AD, Sankar L. A rate-distortion perspective on local differential privacy. 2014. 903–908. [doi: 10.1109/ALLERTON.2014.7028550]

[44] Wang S, Huang L, Wang P, Nie Y, Xu H, Yang W, Li X, Qiao C. Mutual information optimally local private discrete distribution estimation. arXiv Preprint arXiv:1607.08025, 2016.

[45] Ye M, Barg A. Optimal schemes for discrete distribution estimation under locally differential privacy. arXiv Preprint arXiv:1702.00610, 2017.

[46] Fanti G, Pihur V, Erlingsson Ú. Building a RAPPOR with the unknown: Privacy-Preserving learning of associations and data dictionaries. Proc. on Privacy Enhancing Technologies, 2016,2016(3):41–61. [doi: 10.1515/popets-2016-0015]

[47] Ren X, Yu CM, Yu W, Yang S, Yang X, McCann JA, Yu PS. LoPub: High-Dimensional crowdsourced data publication with local differential privacy. IEEE Trans. on Dependable and Secure Computing, 2018,13(9):2151–2166. [doi: 10.1109/TIFS.2018.2812146]

[48] Ren X, Yu CM, Yu W, Yang S, Yang X, McCann J. High-Dimensional crowdsourced data distribution estimation with local privacy. In: Proc. of the 2016 IEEE Int'l Conf. on Computer and Information Technology (CIT). IEEE, 2016. 226–233. [doi: 10.1109/CIT.2016.57]

[49] Zhang J, Cormode G, Procopiuc CM, Srivastava D, Xiao X. Privbayes: Private data release via bayesian networks. In: Proc. of the 2014 ACM SIGMOD Int'l Conf. on Management of Data. ACM, 2014. 1423–1434. [doi: 10.1145/2588555.2588573]

[50] Qardaji W, Yang W, Li N. Priview: Practical differentially private release of marginal contingency tables. In: Proc. of the 2014 ACM SIGMOD Int'l Conf. on Management of Data. ACM, 2014. 1435–1446. [doi: 10.1145/2588555.2588575]

[51] Kellaris G, Papadopoulos S. Practical differential privacy via grouping and smoothing. Proc. of the VLDB Endowment, 2013,6(5):301–312. [doi: 10.14778/2535573.2488337]

[52] Day WY, Li N. Differentially private publishing of high-dimensional data using sensitivity control. In: Proc. of the 10th ACM Symp on Information, Computer and Communications Security. ACM, 2015. 451–462. [doi: 10.1145/2714576.2714621]

[53] Xu J, Zhang Z, Xiao X, Yang Y, Yu G, Winslett M. Differentially private histogram publication. The VLDB Journal, 2013,22(6): 797–822. [doi: 10.1007/s00778-013-0309-y]

[54] Zhang J, Cormode G, Procopiuc C M, Srivastava D, Xiao, X. Private release of graph statistics using ladder functions. In: Proc. of the 2015 ACM SIGMOD Int'l Conf. on Management of Data. ACM, 2015. 731–745.

[55] Day WY, Li N, Lyu M. Publishing graph degree distribution with node differential privacy. In: Proc. of the 2016 Int'l Conf. on Management of Data. ACM, 2016. 123–138. [doi: 10.1145/2723372.2737785]

[56] Bloom BH. Space/Time trade-offs in hash coding with allowable errors. Communications of the ACM, 1970,13(7):422–426. [doi: 10.1145/362686.362692]

[57] Tibshirani R. Regression shrinkage and selection via the Lasso. Journal of the Royal Statistical Society (Series B-Methodological), 1996, 267–288.

[58] Blum A, Ligett K, Roth A. A learning theory approach to noninteractive database privacy. Journal of the ACM (JACM), 2013,60(2): 12. [doi: 10.1145/2450142.2450148]

[59] Hsu J, Khanna S, Roth A. Distributed private heavy hitters. Automata, Languages, and Programming, 2012, 461–472. [doi: 10. 1007/978-3-642-31594-7_39]

[60] Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society (Series B-Methodological), 1977, 1–38.

[61] Chen R, Xiao Q, Zhang Y, Xu J. Differentially private high-dimensional data publication via sampling-based inference. In: Proc. of the 21st ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. ACM, 2015. 129–138. [doi: 10.1145/2783258. 2783379]

[62] Boyd S, Vandenberghe L. Convex Optimization. Cambridge University Press, 2004.

[63] Newey WK, McFadden D. Large sample estimation and hypothesis testing. Handbook of Econometrics, 1994,4:2111–2245.

[64] Meng X, Zhang X. Big data privacy management. Journal of Computer Research and Development, 2016,52(2):265–281 (in Chinese with English abstract).

[65] Viswanath B, Kiciman E, Saroiu S. Keeping information safe from social networking apps. In: Proc. of the 2012 ACM Workshop on Online Social Networks. ACM, 2012. 49–54. [doi: 10.1145/2342549.2342561]

[66] Karwa V, Raskhodnikova S, Smith A, Yaroslavtsev G. Private analysis of graph structure. ACM Trans. on Database Systems (TODS), 2014,39(3):22. [doi: 10.1145/2611523]

[67] Hay M, Li C, Miklau G, Jensen D. Accurate estimation of the degree distribution of private networks. In: Proc. of the 9th IEEE Int'l Conf. on Data Mining, ICDM 2009. IEEE, 2009. 169–178. [doi: 10.1109/ICDM.2009.11]

[68] Ding B, Winslett M, Han J, Li Z. Differentially private data cubes: Optimizing noise sources and consistency. In: Proc. of the 2011 ACM SIGMOD Int'l Conf. on Management of Data. ACM, 2011. 217–228. [doi: 10.1145/1989323.1989347]

[69] Hardt M, Rothblum GN. A multiplicative weights mechanism for privacy-preserving data analysis. In: Proc. of the 51st Annual IEEE Symp. on Foundations of Computer Science (FOCS). IEEE, 2010. 61–70. [doi: 10.1109/FOCS.2010.85]

[70] Hardt M, Ligett K, McSherry F. A simple and practical algorithm for differentially private data release. In: Advances in Neural Information Processing Systems. 2012. 2339–2347.

[71] Thaler J, Ullman J, Vadhan S. Faster algorithms for privately releasing marginals. In: Proc. of the Int'l Colloquium on Automata, Languages, and Programming. Berlin, Heidelberg: Springer-Verlag, 2012. 810–821.

[72] Aggarwal CC. On randomization, public information and the curse of dimensionality. In: Proc. of the 23rd IEEE Int'l Conf. on Data Engineering, ICDE 2007. IEEE, 2007. 136–145. [doi: 10.1109/ICDE.2007.367859]

[73] Aggarwal CC. Privacy and the dimensionality curse. Privacy-Preserving Data Mining, 2008, 433–460. [doi: 10.1007/978-0-387-70992-5_18]

附中文参考文献:

[18] 任奎.云计算中图像数据处理的隐私保护.网络与信息安全学报,2016,(1):12–17.

[39] 张啸剑,孟小峰.面向数据发布和分析的差分隐私保护.计算机学报,2014,(4):927–949.

[64] 孟小峰,张啸剑.大数据隐私管理.计算机研究与发展,2016,52(2):265–281.

叶青青(1992－),女,福建宁德人,博士生,CCF 学生会员,主要研究领域为隐私保护.

朱敏杰(1993－),女,硕士生,CCF 学生会员,主要研究领域为隐私保护.

孟小峰(1964－),男,博士,教授,博士生导师,CCF 会士,主要研究领域为 Web 数据管理,移动数据管理,云数据管理,隐私保护.

霍峥(1982－),女,博士,讲师,CCF 专业会员,主要研究领域为位置及轨迹隐私保护技术,移动对象数据库管理.