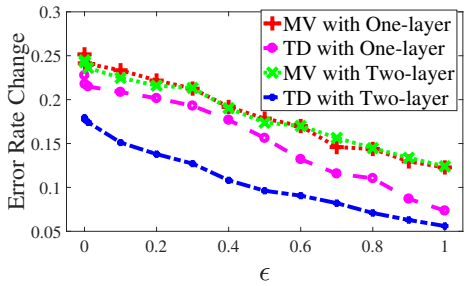

Figure 4: Utility-Privacy Trade-off on Duchenne Smile Dataset

8 CONCLUSIONS

Crowdsourcing has been successfully applied to solve many challenging question answering tasks. However, individual users may have the privacy concern when sharing their sensitive answers. Motivated by this strong need, we propose efficient and effective two-layer mechanism for crowdsourced question answering, which allows users to randomly perturb their answers and then conduct truth discovery on the perturbed answers. Theoretical analysis proves that the two-layer mechanism provides the same level of privacy guarantee as the one-layer mechanism. Furthermore, we theoretically show that good utility can be guaranteed by the two-layer mechanism even with strong privacy constraints. This benefit is brought by the fact that the two-layer mechanism fully utilizes the properties of truth discovery which automatically estimates user quality to derive aggregated answers. The advantage of the proposed two-layer mechanism is confirmed by the experimental results on two real-world datasets. With our developed privacy-preserving mechanism, we can greatly broaden the application domain of truth discovery and enable tasks that would otherwise be infeasible due to privacy concerns.

9 ACKNOWLEDGMENTS

This work was sponsored in part by US National Science Foundation under grant IIS-1553411, CNS-1742845, CNS-1652503 and CNS-1737590. The views and conclusions contained in this paper are those of the authors and should not be interpreted as representing any funding agency.

REFERENCES

[1] Shipra Agrawal, Jayant R Haritsa, and B Aditya Prakash. 2009. FRAPP: a framework for high-accuracy privacy-preserving mining. *Data Mining and Knowledge Discovery* 18, 1 (2009), 101–139.
[2] Yoram Bachrach, Tom Minka, John Guiver, and Thore Graepel. 2012. How to Grade a Test without Knowing the Answers – A Bayesian Graphical Model for Adaptive Crowdsourcing and Aptitude Testing. In *Proc. of ICML*. 255–262.
[3] Anirban Basu, Jaideep Vaidya, Juan Camilo Corena, Shinsaku Kiyomoto, Stephen Marsh, Guibing Guo, Jie Zhang, and Yutaka Miyake. 2014. Opinions of people: factoring in privacy and trust. *ACM SIGAPP Applied Computing Review* 14, 3 (2014), 7–21.
[4] Elisa Bertino, Beng Chin Ooi, Yanjiang Yang, and Robert H Deng. 2005. Privacy and ownership preserving of outsourced medical data. In *Proc. of ICDE*. 521–532.
[5] Arijit Chaudhuri and Rahul Mukerjee. 1988. *Randomized response: Theory and techniques*. Marcel Dekker New York.
[6] Chris Clifton, Murat Kantarcioglu, AnHai Doan, Gunther Schadow, Jaideep Vaidya, Ahmed Elmagarmid, and Dan Suciu. 2004. Privacy-preserving data integration and sharing. In *Proc. of ACM SIGMOD workshop*. 19–26.
[7] Alexander Philip Dawid and Allan M Skene. 1979. Maximum likelihood estimation of observer error-rates using the EM algorithm. *Applied statistics* (1979), 20–28.
[8] Xin Luna Dong, Laure Berti-Equille, and Divesh Srivastava. 2009. Integrating Conflicting Data: The Role of Source Dependence. *PVLDB* 2, 1 (2009), 550–561.
[9] John C Duchi, Michael I Jordan, and Martin J Wainwright. 2013. Local privacy and statistical minimax rates. In *Proc. of FOCS*. 429–438.
[10] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. 2006. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography*. 265–284.
[11] Cynthia Dwork and Aaron Roth. 2014. The algorithmic foundations of differential privacy. (2014).
[12] Úlfar Erlingsson, Aleksandra Korolova, and Vasyl Pihur. 2014. RAPPOR: Randomized Aggregatable Privacy-Preserving Ordinal Response. In *Proc. of CCS*. 1054–1067.
[13] Alban Galland, Serge Abiteboul, Amélie Marian, and Pierre Senellart. 2010. Corroborating Information from Disagreeing Views. In *Proc. of WSDM*. 131–140.
[14] Haibo Hu, Jianliang Xu, Sai Tung On, Jing Du, and Joseph Kee-Yin Ng. 2010. Privacy-aware location data publishing. *ACM Transactions on Database Systems (TODS)* 35, 3 (2010), 18.
[15] Peter Kairouz, Sewoong Oh, and Pramod Viswanath. 2014. Extremal Mechanisms for Local Differential Privacy. In *NIPS*. 2879–2887.
[16] Hiroshi Kajino, Hiromi Arai, and Hisashi Kashima. 2014. Preserving worker privacy in crowdsourcing. *Data Mining and Knowledge Discovery* 28, 5-6 (2014), 1314–1335.
[17] Daniel Kifer and Ashwin Machanavajjhala. 2012. A rigorous and customizable framework for privacy. In *Proc. of PODS*. 77–88.
[18] Qi Li, Yaliang Li, Jing Gao, Lu Su, Bo Zhao, Demirbas Murat, Wei Fan, and Jiawei Han. 2015. A Confidence-Aware Approach for Truth Discovery on Long-Tail Data. *PVLDB* 8, 4 (2015), 425–436.
[19] Qi Li, Yaliang Li, Jing Gao, Bo Zhao, Wei Fan, and Jiawei Han. 2014. Resolving Conflicts in Heterogeneous Data by Truth Discovery and Source Reliability Estimation. In *Proc. of SIGMOD*. 1187–1198.
[20] Xian Li, Xin Luna Dong, Kenneth B. Lyons, Weiyi Meng, and Divesh Srivastava. 2012. Truth Finding on the Deep Web: Is the Problem Solved? *PVLDB* 6, 2 (2012), 97–108.
[21] Yaliang Li, Jing Gao, Chuishi Meng, Qi Li, Lu Su, Bo Zhao, Wei Fan, and Jiawei Han. 2015. A Survey on Truth Discovery. *ACM SIGKDD Explorations Newsletter* 17, 2 (2015), 1–16.
[22] Fenglong Ma, Yaliang Li, Qi Li, Minghui Qiu, Jing Gao, Shi Zhi, Lu Su, Bo Zhao, Heng Ji, and Jiawei Han. 2015. FaitCrowd: Fine Grained Truth Discovery for Crowdsourced Data Aggregation. In *Proc. of KDD*. 745–754.
[23] Chenglin Miao, Wenjun Jiang, Lu Su, Yaliang Li, Suxin Guo, Zhan Qin, Houping Xiao, Jing Gao, and Kui Ren. 2015. Cloud-enabled privacy-preserving truth discovery in crowd sensing systems. In *Proc. of SenSys*. 183–196.
[24] Chenglin Miao, Lu Su, Wenjun Jiang, Yaliang Li, and Miaomiao Tian. 2017. A Lightweight Privacy-Preserving Truth Discovery Framework for Mobile Crowd Sensing Systems. In *Proc. of INFOCOM*. 1539–1547.
[25] Liam O' Neill, Franklin Dexter, and Nan Zhang. 2016. The Risks to Patient Privacy from Publishing Data from Clinical Anesthesia Studies. *Anesthesia & Analgesia* 122, 6 (2016), 2017–2027.
[26] Vibhor Rastogi and Suman Nath. 2010. Differentially private aggregation of distributed time-series with transformation and encryption. In *Proc. of SIGMOD*. 735–746.
[27] Elaine Shi, T-H Hubert Chan, Eleanor G Rieffel, Richard Chow, and Dawn Song. 2011. Privacy-Preserving Aggregation of Time-Series Data.. In *Proc. of NDSS*.
[28] R. Snow, B. O'Connor, D. Jurafsky, and A. Ng. 2008. Cheap and Fast - But is it Good? Evaluating Non-Expert Annotations for Natural Language Tasks. In *Proc. of EMNLP'08*. 254–263.
[29] Hien To, Gabriel Ghinita, and Cyrus Shahabi. 2014. A Framework for Protecting Worker Location Privacy in Spatial Crowdsourcing. *PVLDB* 7, 10 (2014), 919–930.
[30] J. Whitehill, P. Ruvolo, T. Wu, J. Bergsma, and J. Movellan. 2009. Whose Vote Should Count More: Optimal Integration of Labelers of Unknown Expertise. In *NIPS*. 2035–2043.
[31] Xiaokui Xiao, Yufei Tao, and Minghua Chen. 2009. Optimal random perturbation at multiple privacy levels. *PVLDB* 2, 1 (2009), 814–825.
[32] Xiaoxin Yin, Jiawei Han, and Philip S. Yu. 2007. Truth discovery with multiple conflicting information providers on the web. In *Proc. of KDD*. 1048–1052.
[33] Ye Zhang, Wai-Kit Wong, Siu-Ming Yiu, Nikos Mamoulis, and David W Cheung. 2013. Lightweight privacy-preserving peer-to-peer data integration. In *PVLDB*, Vol. 6. 157–168.
[34] Yifeng Zheng, Huayi Duan, Xingliang Yuan, and Cong Wang. 2017. Privacy-Aware and Efficient Mobile Crowdsensing with Truth Discovery. *IEEE Transactions on Dependable and Secure Computing* (2017).
[35] Yudian Zheng, Guoliang Li, Yuanbing Li, Caihua Shan, and Reynold Cheng. 2017. Truth inference in crowdsourcing: is the problem solved? *PVLDB* 10, 5 (2017), 541–552.
[36] D. Zhou, J. C. Platt, S. Basu, and Y. Mao. 2012. Learning from the Wisdom of Crowds by Minimax Entropy. In *NIPS*. 2204–2212.