Enabling Trade-offs in Privacy and Utility in Genomic Data Beacons and Summary Statistics*

✉ Rajagopal Venkatesaramani[1], Zhiyu Wan[2], Bradley A. Malin[2], and ✉ Yevgeniy Vorobeychik[1]

[1] Washington University in St. Louis, St. Louis MO 63130
✉ Corresponding Authors: {rajagopal, yvorobeychik}@wustl.edu
[2] Vanderbilt University Medical Center, Nashville TN 37203
zhiyu.wan@vanderbilt.edu, b.malin@vumc.org

The increased sharing of genomic data over the past decade has sparked acute privacy concerns. To address these concerns, data custodians often resort to only releasing summary-level information, such as alternate allele frequencies (AAF) for single nucleotide variants (SNVs) in a given dataset, which was initially thought to sufficiently protect individual privacy. A notable example is the Beacon service [3] introduced by the Global Alliance for Genomics and Health (GA4GH), which is a web service that only allows queries about the presence or absence of alternate alleles in a dataset. However, research [5, 6] has shown that even such summary statistics are vulnerable to membership inference attacks. These attacks leverage likelihood ratio test (LRT) scores, whereby an attacker, given a target genome, computes the log-likelihood that the target individual was part of the dataset D over which the summary statistics were computed, compared to the likelihood that they were not. Typically, the attacker identifies a threshold θ and claims that the target individual was part of the dataset (i.e., $i \in D$) if the target's LRT score lies below θ. Membership inference allows the attacker to potentially infer sensitive information about the individual such as underlying medical conditions, based on dataset metadata. Approaches used to prevent such attacks typically involve adding noise to the summary information using heuristics or differentially-private mechanisms (flipping bits in the case of Beacons, and adding real-valued noise to AAF releases), or suppressing information release for a subset of SNVs (masking).

In this work, we consider both forms of summary statistics releases: a) the Beacon service, where the system's responses are binary indicators of the presence of certain alleles, and b) the release of alternate allele frequencies for a given set of single nucleotide variants (SNVs). We present a rigorous optimization framework that enables the data custodian to combine the addition of noise (either binary or real-valued, depending on the release) with masking and finely tune the privacy-utility tradeoff as desired, in contrast to state-of-the-art methods which rely on only one mode of data obfuscation. Following the model presented in [8], we further consider two attack models: a) a *fixed-threshold* model where the attacker identifies an LRT score threshold *a priori*, using either simulated or external data according to a maximum allowable false positive rate, and b) an *adaptive-threshold* model where the attacker identifies a threshold which best separates the LRT scores of individuals in the dataset from those who are not *after* the defense is implemented. We capture the latter as the threshold being set to the mean of the lowest K percentile of LRT scores for a set of individuals not in the dataset D. Our goal is to solve an optimization problem capturing the privacy-utility tradeoff, taking into account the relative impact of adding noise to masking SNVs on utility, as well as the relative importance of utility versus privacy using exogenous parameters. We call this the Summary Stats Privacy Problem (SSPP).

Solving the SSPP problem optimally has exponential worst-case runtime. Therefore, to approximately solve SSPP, we propose highly scalable algorithms which leverage greedy heuristics (which we call Soft Privacy Greedy (SPG)). In the case of Beacons, we compute the average marginal contribution of flipping each SNV, as well as the average marginal impact of masking each SNV on the LRT score, normalized by the relative costs of flipping and masking (user specifies the cost of flipping as α, and the cost of masking is $1 - \alpha$). We then rank the SNVs in decreasing order of their marginal contributions, and greedily either flip or mask SNVs until a desired level of privacy is achieved (using a user-specified parameter, w). We assume that the subset of flipped SNVs and the subset of masked SNVs are disjoint. In the case of AAF releases, we propose a greedy heuristic that alternates between masking SNVs, and adding real-valued noise to the remaining release using a differentially-private Laplace mechanism.

* We acknowledge support of this work by the National Institutes of Health (NIH) under grant RM1HG009034 and the National Science Foundation (NSF) CAREER award program under grant IIS-1905558.

Our experiments were conducted on a dataset based on the 1000 Genomes Project [1] made available as part of the 2016 iDash workshop on Privacy and Security [7], consisting of 800 individuals and over 1.3 million SNVs. We compare our approach to state-of-the-art baseline approaches, including differentially private mechanisms [2], randomized noise on rare alleles [4], strategic flipping using differential discriminative power [9], linkage-equilibrium based suppression [5], and prior greedy methods [8]. We also evaluate performance against a more powerful adversary that attempts to infer missing or flipped SNVs using linkage disequilibrium as a metric for correlation between SNVs. Our approach scales easily to 1.3 million SNVs, and the results presented in Fig. 1 demonstrate that our approach outperforms prior art in both utility and privacy.

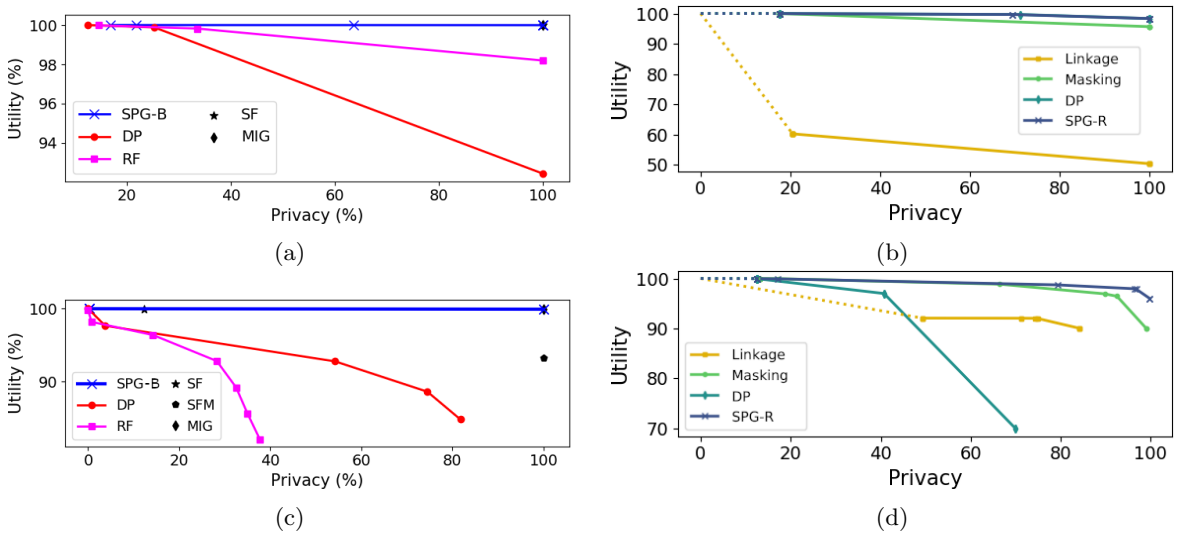

Fig. 1: Utility-Privacy plots for the various attack models. a) Fixed threshold model, Beacon release, $\theta = -250$, baselines flip SNVs, b) Fixed threshold model, AAF release, $\theta = 0$, c) Adaptive threshold model, Beacon release, $K = 10$, baselines flip SNVs, d) Adaptive threshold model, AAF release, $K = 10$.

References

1. 1000 Genomes Project Consortium, et al.: A global reference for human genetic variation. Nature **526**(7571), 68 (2015)
2. Cho, H., Simmons, S., Kim, R., Berger, B.: Privacy-preserving biomedical database queries with optimal privacy-utility trade-offs. Cell Systems **10**(5), 408–416 (2020)
3. Fiume, M., Cupak, M., Keenan, S., Rambla, J., de la Torre, S., Dyke, S.O., Brookes, A.J., Carey, K., Lloyd, D., Goodhand, P., et al.: Federated discovery and sharing of genomic data using beacons. Nature biotechnology **37**(3), 220–224 (2019)
4. Raisaro, J.L., Tramer, F., Ji, Z., Bu, D., Zhao, Y., Carey, K., Lloyd, D., Sofia, H., Baker, D., Flicek, P., et al.: Addressing beacon re-identification attacks: quantification and mitigation of privacy risks. Journal of the American Medical Informatics Association **24**(4), 799–805 (2017)
5. Sankararaman, S., Obozinski, G., Jordan, M.I., Halperin, E.: Genomic privacy and limits of individual detection in a pool. Nature genetics **41**(9), 965–967 (2009)
6. Shringarpure, S.S., Bustamante, C.D.: Privacy risks from genomic data-sharing beacons. The American Journal of Human Genetics **97**(5), 631–646 (2015)
7. Tang, H., Wang, X., Wang, S., Jiang, X.: idash privacy and security workshop (2016), http://www.humangenomeprivacy.org/2016/
8. Venkatesaramani, R., Wan, Z., Malin, B.A., Vorobeychik, Y.: Defending against membership inference attacks on beacon services. arXiv preprint arXiv:2112.13301 (2021)
9. Wan, Z., Vorobeychik, Y., Kantarcioglu, M., Malin, B.: Controlling the signal: Practical privacy protection of genomic data sharing through beacon services. BMC Medical Genomics **10**(2), 87–100 (2017)