Poster: MBeacon: Privacy-preserving Beacons for DNA Methylation Data

Inken Hagestedt*, Yang Zhang*, Pascal Berrang*, Mathias Humbert†,
Haixu Tang‡, Xiaofeng Wang‡, Michael Backes§
*CISPA, firstname.lastname@cispa.saarland
†Swiss Data Science Center, mathias.humbert@epfl.ch
‡Indiana University Bloomington, {hatang, xw7}@indiana.edu
§CISPA Helmholtz Center i.G., backes@cispa.saarland

***Abstract*—The advancement of molecular profiling techniques fuels biomedical research with a large quantity of data. To facilitate biomedical data sharing, the Global Alliance for Genomics and Health established the Beacon system, a search engine designed to help researchers to find datasets of interest. While the current Beacon system only supports genomic data, other types of biomedical data, such as DNA methylation, are also essential for advancing our understanding of the field. In this paper, we propose the first Beacon system to share DNA methylation data, namely the MBeacon system. Since the current genomic Beacon is vulnerable to membership inference attacks, and DNA methylation data is highly privacy-sensitive, we take a privacy-by-design approach to construct MBeacon. First, we demonstrate the privacy threat, by proposing a membership inference attack tailored specifically to methylation-based Beacons. Our experimental results show that 100 queries on the MBeacon are sufficient to achieve a successful attack with AUC above 0.9. Then, we propose a privacy-preserving mechanism and show, by simulating realistic adversaries and legitimate researchers, that membership inference attacks can be successfully prevented with AUC dropping to 0.5 without harming researchers' utility. We further implement a fully functional prototype of MBeacon that we will make available to the research community in order to facilitate data sharing.**

I. Introduction

Data sharing is essential for the development of biomedical research. However, large-scale data sharing is limited in its success, primarily due to privacy concerns. Aiming for a responsible genomic data sharing solution, the Global Alliance for Genomics and Health established the Beacon system,[1] a search engine providing information about biomedical data from all institutions being part of the system. Each institution establishes its own *Beacon* with its dataset. The Beacon only supports one type of query: whether the Beacon's dataset contains a specified nucleotide at a given position and chromosome. The response is a "Yes" for those institutions' Beacons that possess such a data record and "No" otherwise.

Currently, the Beacon project only supports genomic data, while other types of biomedical data are also essential for biomedical study. DNA methylation is one of the most important epigenetic elements and is very influential to human health: for instance, anomalous changes in the DNA methylation patterns are frequently observed in cancer [1]. Therefore, there exists a huge demand for methylation data sharing.

In this paper, we construct the first Beacon system for sharing DNA methylation data, namely, the MBeacon system.

Recently, researchers have shown that the current genomic Beacon is vulnerable to membership inference attacks [2]–[5]. By inferring whether her victim is part of the database, the attacker can infer sensitive attributes that are published as meta-information about the database, e.g., that it contains samples from patients with a specific disease or phenotype. In addition, the authors of [6], [7] have demonstrated the severe privacy risks stemming from sharing DNA methylation data. Therefore, to construct the MBeacon system, we follow a privacy-by-design approach.

II. Attack

The first step towards a privacy-preserving MBeacon is to evaluate the privacy threat of membership inference attacks against a non privacy-preserving MBeacon. Since existing attacks on Beacon-like systems are tailored to genomic data only, we first design a membership inference attack suitable for DNA methylation data. Our membership inference attack relies on the likelihood-ratio test. To estimate the probabilities of the Beacon answering "No" resp. "Yes", we rely on the normal distribution calibrated to mean and standard deviation of the general population's methylation values.

We empirically evaluate our attack on several unprotected MBeacons composed of various methylation datasets and show that the attack achieves a superior performance. For instance, the simulated attacker can achieve an AUC value (area under the ROC curve) of over 0.9, for just 100 queries submitted to a MBeacon. The results fully demonstrate the privacy threat of the Beacon system for methylation data.

III. Defense

We propose a defense mechanism that could be implemented and deployed jointly with the novel MBeacon system.

Our defense mechanism, namely SVT$^{\mathbf{2}}$, is a variant of the sparse vector technique in differential privacy. Since the main challenge of differential privacy is to scale the noise in a utility-preserving manner, SVT$^{\mathbf{2}}$ utilizes again the background knowledge of means and standard deviations of the

[1] https://beacon-network.org/

general population being available: Only answers from the MBeacon that deviate from what one would expect from the background knowledge are treated highly privacy sensitive. Since this is the case only for a minority of queries, the total amount of highly privacy-relevant answers can be bounded and noise is calibrated with respect to this bound, also called the privacy budget. Additionally, we introduce a k-anonymity style threshold to further reduce the amount of highly privacy relevant cases and, in consequence, the total amount of noise to be added in each computation. We prove that SVT^2 is differentially private.

IV. Utility Metrics

The goal of the MBeacon system is to facilitate data sharing in biomedical research. Therefore, the main users of MBeacons are researchers who want to discover the institutions that possess data of their interests.

In order to quantify the impact of the proposed privacy-preserving mechanism on the real-world utility of our MBeacon system, we introduce a new utility metric simulating the behavior of a legitimate researcher. Concretely, we simulate researchers knowing 5 patients of a specific disease that query either a MBeacon containing only patients from a different disease or a MBeacon containing few patients of the disease of interest and the majority of a different disease. The simulated researcher's goal is to correctly identify the former as not interesting for her research and the later as interesting. To directly compare researchers' and attackers' performance, we simulate two types of attackers. The first attacker, referred to as "full" attacker, tries to infer whether her victim is in the MBeacon, not knowing whether the victim is from the minority or the majority disease and which of the two different MBeacons she is querying. Additionally, we simulate the "best" attacker, that is guaranteed to get a victim from the majority disease and queries the MBeacon containing only patients from this disease. This attacker gives an upper bound on the privacy threat.

V. Defense Evaluation

Through extensive experiments using 8 different methylation datasets, simulating researchers' behavior along with attackers', we evaluate the performance of our privacy-preserving MBeacon. Our results show that the privacy loss with regard to membership inference attacks can be minimized while the researchers' utility still remains high. In particular, for carefully chosen privacy parameters, it is possible to decrease the attacker's performance to random guessing (AUC close to 0.5) while preserving a high utility for the researcher (AUC $>$ 0.9). Figure 1 shows the results of one experiment using data from two different brain tissues for the simulation.

Furthermore, we conduct an extensive evaluation of privacy parameters for SVT^2 and provide the necessary tools for an institution to tune these parameters to their needs. In addition, we have implemented a fully-functional prototype of the MBeacon system and will make it available to the research community.

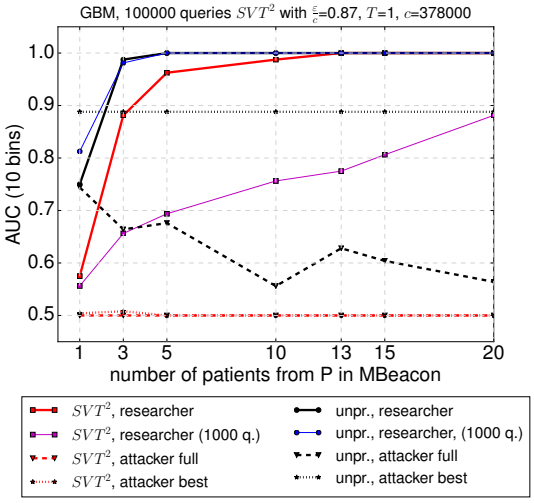

Fig. 1. Comparison of researchers' and attackers' performances in unprotected MBeacon (black, "unpr.") and protected MBeacon (red) using Glioblastoma data as majority of the patients and Ependomyoma as minority (P), allowing using up to 100,000 queries. Additionally, we plot the researchers' performances for 1,000 queries in blue (unprotected) and magenta (protected). AUCs with values smaller than 0.5 are displayed as 0.5.

References

[1] M. Esteller and J. G. Herman, "Cancer as an Epigenetic Disease: DNA Methylation and Chromatin Alterations in Human Tumours," *The Journal of Pathology*, vol. 196, no. 1, pp. 1–7, 2002.

[2] S. S. Shringarpure and C. D. Bustamante, "Privacy risks from genomic data-sharing beacons," *The American Journal of Human Genetics*, vol. 97, no. 5, pp. 631–646, 2015.

[3] J. L. Raisaro, F. Tramèr, Z. Ji, D. Bu, Y. Zhao, K. Carey, D. Lloyd, H. Sofia, D. Baker, P. Flicek *et al.*, "Addressing beacon re-identification attacks: quantification and mitigation of privacy risks," *Journal of the American Medical Informatics Association*, p. ocw167, 2017.

[4] M. M. Al Aziz, R. Ghasemi, M. Waliullah, and N. Mohammed, "Aftermath of bustamante attack on genomic beacon service," *BMC medical genomics*, vol. 10, no. 2, p. 43, 2017.

[5] Z. Wan, Y. Vorobeychik, M. Kantarcioglu, and B. Malin, "Controlling the signal: Practical privacy protection of genomic data sharing through beacon services," *BMC medical genomics*, vol. 10, no. 2, p. 39, 2017.

[6] M. Backes, P. Berrang, M. Bieg, R. Eils, C. Herrmann, M. Humbert, and I. Lehmann, "Identifying Personal DNA Methylation Profiles by Genotype Inference," in *Proceedings of the 38th IEEE Symposium on Security and Privacy (S&P)*. IEEE, 2017, pp. 957–976.

[7] P. Berrang, M. Humbert, Y. Zhang, I. Lehmann, R. Eils, and M. Backes, "Dissecting Privacy Risks in Biomedical Data," in *Proceedings of the 3rd IEEE European Symposium on Security and Privacy (Euro S&P)*. IEEE, 2018.