Poster: Privacy Preserving Population Stratification for Collaborative Genomic Research

Leonard Dervishi*, Wenbiao Li*, Anisa Halimi†, Xiaoqian Jiang‡ Jaideep Vaidya§ and Erman Ayday*
*Case Western Reserve University, Cleveland, OH
†IBM Research, Dublin, Ireland
‡UTHealth, Houston, TX
§Rutgers University, Newark, NJ

***Abstract*—The rapid improvements in genomic sequencing technology have led to the proliferation of locally collected genomic datasets. Given the sensitivity of genomic data, it is crucial to conduct collaborative studies while preserving the privacy of the individuals. However, before starting any collaborative research effort, the quality of the data needs to be assessed. One of the essential steps of the quality control process is population stratification: identifying the presence of genetic difference in individuals due to subpopulations. One of the common methods used to group genomes of individuals based on ethnicity is principal component analysis (PCA). In this paper, we propose a framework to perform population stratification using PCA across multiple collaborators in a privacy-preserving way. In our proposed client-server-based scheme, we initially let the server train a global PCA model on a publicly available genomic dataset which contains individuals from multiple populations. The global PCA model is later used to reduce the dimensionality of the local data by each collaborator (client). After adding noise to achieve local differential privacy (LDP), the collaborators send metadata (in the form of their local PCA outputs) about their research datasets to the server, which then aligns the local PCA results to identify the genetic differences among collaborators' datasets. Our results on real genomic data show that the proposed framework can perform population stratification with high accuracy while preserving the privacy of the research participants.**

I. Introduction

Collaborative studies that pool large data from multiple sources are generally more powerful statistics, and as a result, are beneficial to all the participating parties. To ensure accurate outcomes of collaborative studies, the quality and consistency of the data among the datasets of the collaborators need to be ensured [1], [2]. A trivial solution is to pool the dataset of each party in a centralized environment and analyze the data. However, the dataset of each party might not always be shared, which could lead to sensitive information about the research participants being revealed.

There are known techniques to preserve privacy for genomic data processing such as meta-analysis, cryptographic solutions, and differential privacy-based solutions. In the meta-analysis, collaborators exchange aggregate statistics in order to obtain global statistics of a specific study. In cryptographic solutions, the collaborators are able to perform collaborative analysis of the encrypted data. Differential privacy-based solutions allow the collaborators to exchange perturbed data or statistics between them under some privacy guarantees. Each of these solutions comes with its drawbacks, as discussed below.

- Sharing aggregate statistics as part of meta-analysis still poses privacy risk for the dataset participants [3]. Although the aggregate statistics are anonymized, it is possible to re-identify individuals based on the information that is shared. This could lead to potential harm if the data is used for malicious purposes. It is important to be careful when sharing aggregate statistics and to make sure that only information that cannot be used to identify individuals is shared.
- Cryptographic techniques are often used to provide high privacy guarantees for data sets. However, these techniques are not applicable to large-scale data sets. This is because the computational resources required to implement cryptographic techniques on large data sets are prohibitive.
- Differential privacy-based solutions add a large amount of noise to the data they are protecting. This makes it difficult for an attacker to learn anything about the underlying data from the noise. However, it also makes it difficult for legitimate users of the data to get accurate results from queries.

Due to the aforementioned limitation, these technologies may not be able to provide ideal solutions to our problems in collaborative genomic research. One of the main quality control steps that need to be considered in collaborative studies is population stratification (i.e., population structure). This can occur when the genetic difference in case-control groups occurs due to differences in ancestry, rather than the association to the phenotype (trait) of interest. Previous genetic studies [4], [5], [6], [7] have been widely using principal component analysis (PCA) as a standard technique for population stratification. PCA is a popular technique used to analyze and identify the patterns and relationships within the data. PCA is also used in image compression [8], facial recognition [9], medical data correlation [10], covariance analysis in neuroscience [11], and quantitative finance [12]. In this work, we focus on privacy-preserving population stratification on collaborative studies using PCA.

II. Proposed Framework

We consider an environment in which multiple collaborators (researchers) are present whose aim is to determine whether the individuals that are present in their local datasets belong to

multiple populations and a centralized entity (server) which is used to capture the population structure and assign the population cluster to each data point (individual). Our main goals are to obtain an accurate PCA outcome of the federated dataset of all researchers and to preserve the privacy of the samples (data contributors) in researchers' datasets. The challenge lies in the fact that due to privacy concerns, researchers cannot send their local datasets to the server (or to each other) for a combined PCA. One trivial way to perform combined PCA across different researchers is to let each researcher conduct PCA on their local datasets, and then send the PCA results to the server. However, combining independent PCA results at the server to identify the relative distance between the data points is not possible.

Thus, we propose a privacy-preserving collaborative PCA scheme between researchers and the server. The server initially trains a PCA model using a publicly available genomic dataset which contains individuals of various populations. Next, the trained PCA model is sent to the researchers and the researchers conduct their local PCA using this model. As a result, each researcher obtains the projection of individuals in their datasets with respect to the trained PCA model. To achieve local differential privacy (ϵ-LDP, where ϵ is the privacy parameter), and hence to protect the privacy of the records in their local datasets, researchers add Laplacian noise (i.e., noise based on Laplacian distribution) to each sample and send some metadata, which contains the PCA data points (principal components) and the hashed IDs of each sample, to the server. The server then calculates the coordinates for all the users coming from multiple researchers, combines the PCA results coming from multiple researchers, and identifies the population substructure. Finally, it sends back to each researcher the hashed user IDs and the label of the population cluster they belong to.

III. EVALUATION

We quantify the privacy risk of sharing such metadata. We consider membership inference risk as the main vulnerability and we show that the privacy risk introduced by our framework at the server is below the baseline risk of sharing summary statistics as part of genetic studies, which is allowed by a lot of institutions, including NIH [13]. We evaluate our framework in terms of both utility and privacy. From the obtained results, we observe that the accuracy of correctly identifying population clusters increases when the server trains the PCA model on a public genomic dataset consisting of various populations. We also monitor the effect of the number of populations that are present in the researchers' datasets for various ϵ values. Our framework achieves a precision of 0.89 and a recall of 0.88 for $\epsilon = 3$ when the researchers' datasets contain individuals from multiple populations. For the same setup, it achieves a membership inference power of 0.19 (which is a significantly low value indicating high privacy).

The proposed framework achieves a higher power (a power of 0.39) when the researchers' datasets contain only 1 population, which is still lower than the power due to sharing the GWAS statistics (the baseline risk). Furthermore, we also explore the effect of using different numbers of principal components (dimensions). We notice that when researchers use PCA to transform their data to more than 3 dimensions (i.e., providing more than 3 principal components), the utility increases immensely, while we observe only a slight increase in the power of membership inference.

IV. CONCLUSION

In this work, we have proposed a novel and effective privacy-preserving framework which partitions populations in collaborative studies. We have shown that the proposed framework identifies with high accuracy, precision, and recall the genetic differences among collaborators' datasets while preserving the privacy of the research participants. This work will enable researchers to conduct collaborative research with high quality data while ensuring that the privacy of the research participants is preserved.

ACKNOWLEDGMENT

The work was partly supported by the National Library of Medicine of the National Institutes of Health under Award Number R01LM013429 and by the National Science Foundation (NSF) under grant numbers 2141622, 2050410, 2200255, and OAC-2112606.

REFERENCES

[1] S. Turner et al., "Quality control procedures for genome-wide association studies," Curr. Protoc. Hum. Genet., vol. Chapter 1, p. Unit1.19, Jan. 2011.

[2] R. L. Zuvich et al., "Pitfalls of merging GWAS data: lessons learned in the eMERGE network and quality control procedures to maintain high data quality," Genet. Epidemiol., vol. 35, no. 8, pp. 887–898, Dec. 2011.

[3] E. A. Zerhouni and E. G. Nabel, "Protecting aggregate genomic data," Science, vol. 322, no. 5898, pp. 44–44, 2008.

[4] O. Hanotte, D. G. Bradley, J. W. Ochieng, Y. Verjee, E. W. Hill, and J. E. O. Rege, "African pastoralism: genetic imprints of origins and migrations," Science, vol. 296, no. 5566, pp. 336–339, Apr. 2002.

[5] A. L. Price, N. J. Patterson, R. M. Plenge, M. E. Weinblatt, N. A. Shadick, and D. Reich, "Principal components analysis corrects for stratification in genome-wide association studies," Nat. Genet., vol. 38, no. 8, pp. 904–909, Aug. 2006.

[6] N. Patterson, A. L. Price, and D. Reich, "Population structure and eigenanalysis," PLoS Genet., vol. 2, no. 12, p. e190, Dec. 2006.

[7] S. Moss, "Faculty Opinions recommendation of An African origin for the intimate association between humans and Helicobacter pylori," Faculty Opinions – Post-Publication Peer Review of the Biomedical Literature. 2007. doi: 10.3410/f.1065791.518721.

[8] V. Gaidhane, V. Singh, and M. Kumar, "Image Compression Using PCA and Improved Technique with MLP Neural Network," in 2010 International Conference on Advances in Recent Technologies in Communication and Computing, Oct. 2010, pp. 106–110.

[9] R. Gottumukkal and V. K. Asari, "An improved face recognition technique based on modular PCA approach," Pattern Recognition Letters, vol. 25, no. 4. pp. 429–436, 2004. doi: 10.1016/j.patrec.2003.11.005.

[10] N. A. Qureshi et al., "Application of principal component analysis (PCA) to medical data," Indian J. Sci. Technol., vol. 10, no. 20, pp. 1–9, Feb. 2017.

[11] K. Polat and S. Güneş, "An expert system approach based on principal component analysis and adaptive neuro-fuzzy inference system to diagnosis of diabetes disease," Digit. Signal Process., vol. 17, no. 4, pp. 702–710, Jul. 2007.

[12] H. Yu, R. Chen, and G. Zhang, "A SVM Stock Selection Model within PCA," Procedia Comput. Sci., vol. 31, pp. 406–412, Jan. 2014.

[13] "NOT-OD-19-023: Update to NIH Management of Genomic Summary Results Access." https://grants.nih.gov/grants/guide/notice-files/NOT-OD-19-023.html (accessed Apr. 08, 2022).

Poster: Privacy Preserving Population Stratification for Collaborative Genomic Research

Leonard Dervishi✵, Wenbiao Li✵, Anisa Halimi†, Xiaoqian Jiang‡ Jaideep Vaidya§ and Erman Ayday✵
✵Case Western Reserve University, Cleveland, OH
†IBM Research, Dublin, Ireland
‡UTHealth, Houston, TX
§Rutgers University, Newark, NJ

Introduction

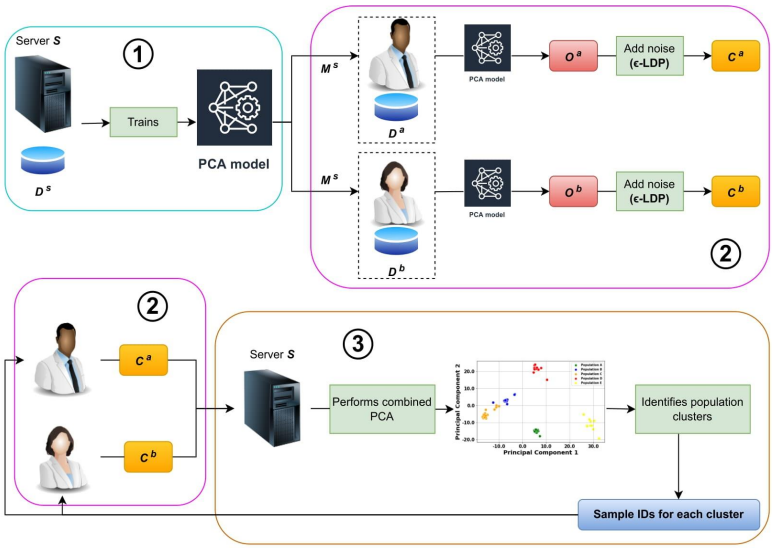
- Collaborative studies of local genomic datasets are increasingly used in genomic sequencing.
- Quality control needs to be assessed on the local data.
- Privacy preserving is crucial due to the sensitivity of genomic data.

Existing Solution and Drawbacks

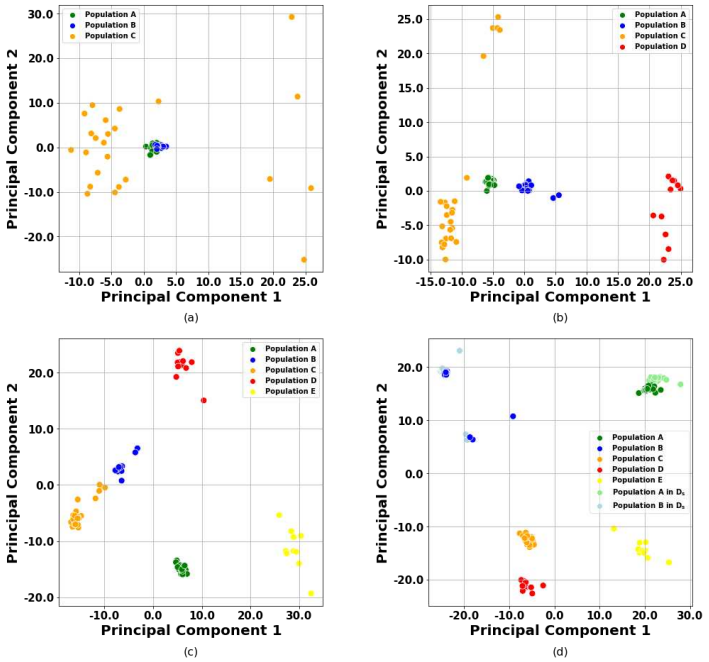
- Sharing aggregate statistics as part of meta-analysis still poses privacy risk for the participants [1].
- Cryptographic techniques are not applicable for large-scale datasets
- Differential privacy-based solution hinder legitimate users obtain accurate results

Proposed Framework

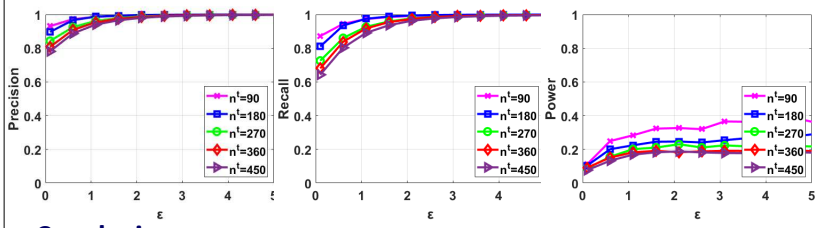
1. The server trains the PCA model [2] and sends it to each researcher.
2. Each researcher uses the trained PCA model to transform their original local dataset. Next, they noise in order to achieve Ɛ-LDP. Then, they send the metadata back to the server.
3. The server combines the PCA results and classifies users by population clusters. Finally, the server sends the respective sample IDs along with the corresponding population cluster to each researcher and the total number of individuals that each cluster contains in the federated dataset.



PCA Plots



Evaluation



Conclusion

In this work, we have proposed a novel and effective privacy-preserving framework which partitions populations in collaborative studies. We have obtained that the proposed framework identifies with high performance metrics while preserving the privacy of the research participants with low power values.

Acknowledgement

The work was partly supported by the National Library of Medicine of the National Institutes of Health under Award Number R01LM013429 and by the National Science Foundation (NSF) under grant numbers 2141622, 2050410, 2200255, and OAC-2112606.

References

1. Elias A Zerhouni and Elizabeth G Nabel. Protecting aggregate genomic data. Science, 322(5898):44–44, 2008.
2. Namrata Vaswani, YuejieChi, and Thierry Bouwmans. Rethinking pcafor modern data sets: Theory, algorithms, and applications [scanning the issue]. Proceedings of the IEEE, 106(8):1274–1276, 2018.