Bachelor Thesis Proposal - Private mouse and keyboard behavioral data

Hossam Elfar
Supervised by: Mayar Elfares and Guanhua Zhang

January 2023

1 Introduction

Our ability to answer important questions about human behavior becomes difficult due to limited access to personal behavioral data. This data is usually distributed among several parties (individuals or organizations) that do not allow the sharing of data due to privacy concerns. Hence, remote data science can be a promising solution. It uses a whole host of servers to answer these questions while ensuring that the data is protected and secure using privacy-enhancing methods (e.g. differential privacy). Therefore, in this project, we try to answer the question of how can behavior scientists perform remote and reliable data science without actually seeing the personal data.

As an example of privacy-sensitive data, we will investigate a behavioural mouse and keyboard dataset. Differential privacy (DP) [1, 2] allows the public sharing of information about a dataset by computing insights of groups within the dataset while preserving the individuals' privacy. In other words, DP allows data scientists to train behaviour models without collecting the raw inputs from users. The challenge is to enable behavior scientists to train models on mouse and keyboard data that they can't see using differential privacy, hence a tool for remote behavior computation is needed.

2 Related work

Privacy of behavioural dataset : Studying human behaviour is of particular interest within the field of human-computer interaction (HCI). Behavioural datasets collected from these interactions can be privacy-sensitive, hence, techniques to protect individuals' privacy against unwanted inferences are required. A number of prior works have focused on protecting users' data [3] where they derived different approaches (non-deterministic and deterministic methods) for behavioural data anonymization, The non-deterministic methods rely on randomness in their transformation, which can yield different results for the same input and deterministic methods always give the same result.

Mouse and Keyboard dataset : Keystroke dynamics and mouse movements are effective behavioural modalities which can be useful for data analysis, as they can contain important data which can be useful as biometrics in active authentication and predicting user's intents [7] or performing different tasks [5]. **Privacy on Mouse and keyboard dataset** : The privacy of collection and transmission of keyboard and mouse data did not receive much attention and such data collection may compromise the privacy of users (e.g. during the process of active authentication). The only existing method used for behavioural data is based on anonymization [6]. Two approaches based on where the data anonymization is performed were proposed, either on the client side or on the server side. If the data anonymization issue is handled on the client side, then the data can be sent to the server directly without any identifier. If the processing is done on the server side, the data transmission security needs to be guaranteed. However, de-anonymisation attacks [4] can be performed to leak the behavioural data.

3 Key novelty and contributions

Our key contributions are two-fold: We propose the first differential privacy approach for mouse and keyboard datasets that (1) allows data owners to upload their datasets on network nodes while preserving the data privacy and (2) allows remote processing of mouse and keyboard behavioural data by performing remote data analysis on the network node where data scientists can log into the network and get the required data to run their machine learning models.

4 Method

Dataset: Mouse and keyboard datasets are highly sensitive and can include personally sensitive information such as username, password, and social security number. However, the privacy of the collection of keyboard and mouse data was under-studied in the literature. The standard methods suffer from a variety of vulnerabilities such as masquerading and potential system compromise, which can allow the attacker to steal important personal and organizational information as well as the identity of the victim. This makes mouse and keyboard data different from other private behavioural data.
We will implement a remote data science approach for a private mouse and keyboard behaviour dataset that enables data scientists to train models on the privately-uploaded mouse and keyboard data. We will study this method on the BEHACOM Dataset which is a dataset modelling users' behaviour in computers. This sample is captured from twelve users interacting with their computers for fifty-five days and performing different tasks using keyboard and mouse. The data collection component runs on each personal computer and gathers the raw data generated by each individual interacting with his/her device. Another proposed dataset is the Four HCI Tasks Dataset . This dataset contains keystroke

and mouse input from 46 users performing 5 different tasks.
Launch Domain node: In this step, we will use PySyft [8] which is an open-source multi-language library enabling secure and private machine learning, to launch the domain that acts like a server to host our private behaviour dataset. Then, we will use HAGrid tool to launch a domain node .
Data preparation: After the last step, the domain node will be ready to host data but before we do that data preparation is needed. Data preparation is particularly important for remote data science and machine learning. This is because the data scientist does not inherently have the ability to freely check how clean the data is and whether it needs pre-processing. As such, in this project, the responsibility falls onto the data owner to ensure the data is clean, annotated, and usable. This is implemented through data acquisition, quality check, data annotation and converting the data to a PyGrid-compatible format.
Differential privacy: Behaviour mouse and keyboard data are protected via differential privacy by adding a certain calculated Gaussian noise (quantified measure of privacy loss) and an upper bound measure according to the meta-data (minimum value, maximum value and entities) of the dataset.
Data scientists: Data scientists can log into the network, and get a privacy budget (The privacy budget represents how much noise the data scientist can remove from a dataset when accessing it). Domains will set a privacy budget per data scientist and this is done by making a request which will allow them to study the dataset remotely and select a dataset on the domain node to study and then perform their analysis on the selected data.

5 Intended outcomes

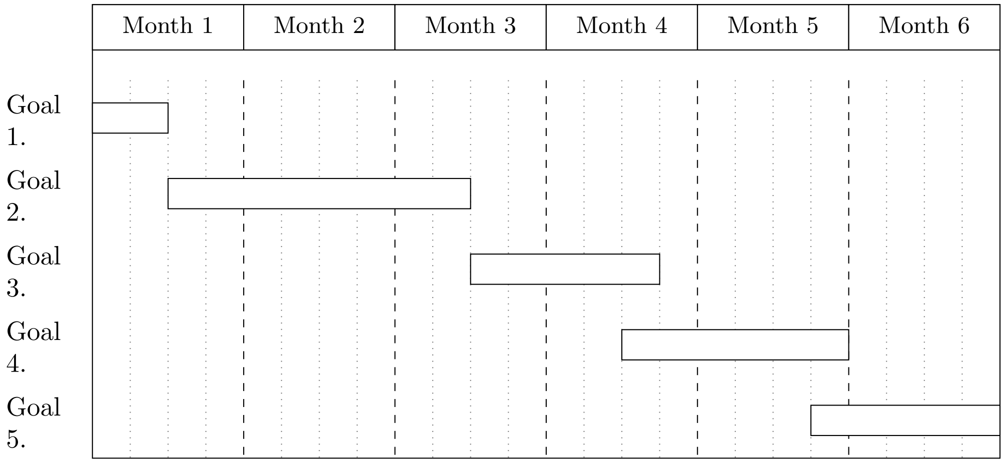
Mandatory goals:

1. Deploy a domain node using HAGrid
2. Deploy a network node that collects data from different domain nodes and handles the network requests using PySyft and PyGrid
3. Allowing data owners to upload datasets to domain nodes.
4. Obfuscating the data once uploaded via differential privacy
5. Allow data scientists to log into the network, get a privacy budget, and run machine learning models.

Optional goals:

1. Building novel deep learning models for behavioural mouse and keyboard data.

6 Schedule with milestones



References

[1] Dwork, C. Differential privacy: A survey of results. In *International conference on theory and applications of models of computation* (2008), Springer, pp. 1–19.

[2] Dwork, C., Roth, A., et al. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science 9*, 3–4 (2014), 211–407.

[3] Hanisch, S., Arias-Cabarcos, P., Parra-Arnau, J., and Strufe, T. Privacy-protecting techniques for behavioral data: A survey. *arXiv preprint arXiv:2109.04120* (2021).

[4] Narayanan, A., Shi, E., and Rubinstein, B. I. Link prediction by de-anonymization: How we won the kaggle social network challenge. In *The 2011 International Joint Conference on Neural Networks* (2011), IEEE, pp. 1825–1834.

[5] Sánchez, P. M. S., Valero, J. M. J., Zago, M., Celdrán, A. H., Maimó, L. F., Bernal, E. L., Bernal, S. L., Valverde, J. M., Nespoli, P., Galindo, J. P., et al. Behacom-a dataset modelling users' behaviour in computers. *Data in brief 31* (2020), 105767.

[6] Sun, Y., and Upadhyaya, S. Secure and privacy preserving data processing support for active authentication. *Information Systems Frontiers 17*, 5 (2015), 1007–1015.

[7] Zhang, G., Hindennach, S., Leusmann, J., Bühler, F., Steuerlein, B., Mayer, S., Bâce, M., and Bulling, A. Predicting next actions and latent intents during text formatting.

[8] Ziller, A., Trask, A., Lopardo, A., Szymkow, B., Wagner, B., Bluemke, E., Nounahon, J.-M., Passerat-Palmbach, J., Prakash, K., Rose, N., et al. Pysyft: A library for easy federated learning. In *Federated Learning Systems*. Springer, 2021, pp. 111–139.