Privacy as a Service: Publishing Data and Models

Ashish Dandekar*, Debabrota Basu, Thomas Kister, Geong Poh Sen, Jia Xu, Stéphane Bressan
***ashishdandekar@u.nus.edu**
NUS-Singtel Cyber Security Research & Development Laboratory

INTRODUCTION

We demonstrate our *Privacy-as-a-Service (PaaS)* system and *Liánchéng*, our *Workflow-as-a-Service (WaaS)* cloud that hosts it.

Liánchéng is a data sharing cloud system that provides a graphical workflow language. We extend it by incorporating privacy risk assessment and privacy preservation operators. Privacy-as-a-Service provides operators to publish not only anonymised data but also models created by statistical machine learning with differential privacy guarantees. We illustrate the construction and execution of privacy preserving workflows in these Workflow-as-a-Service and Privacy-as-a-Service models and systems for publishing data and statistical machine learning models using a census dataset and a medical dataset.

LIÁNCHÉNG

- Liánchéng, a private data sharing cloud service, is deployed on a hardware infrastructure consisting of 128 commodity servers

- **Data Sharing**. This aspect is reminiscent of services such as *Dropbox.* Each Liánchéng user gets a private account on which she can upload, download, organise and manage her data. Liánchéng provides both a Web interface and a desktop computer synchronisation agent. The internal sharing mechanism (user-to-user) relies on access control lists on directories. Liánchéng also provides additional publishing mechanisms, such as public access through URLs, for files.

- **Workflow-as-a-Service**. While, on the one hand, *IaaS* is fully customisable, it requires computing skills and efforts that constitute unnecessary obstacles. On the other hand, *PaaS* often limits users to the options proposed and has insufficient programmability. Liánchéng realises the compromise by offering an interactive GUI-based workflow language and domain specific operators.

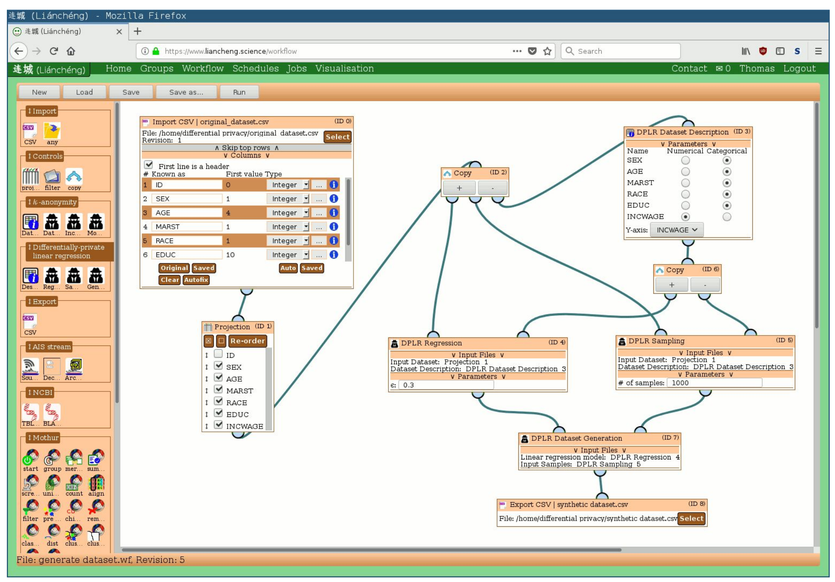
- A Liánchéng workflow is a directed acyclic graph whose vertices represent operators and whose edges represent data flow. An operator can have an arbitrary number of parameters and has at least one input or output interface.

PRIVACY AS A SERVICE

- We extend Liánchéng with disclosure risk assessment and privacy preservation operators. We refer to such a cloud system functionality as Privacy-as-a-Service.

- **Publishing Data**. We use traditional anonymisation techniques such as k-anonymity [1], l-diversity [2] and t-closeness [3]. Alternatively, we generate fake but realistic datasets by using machine learning models that are trained on private datasets. We use machine learning algorithms for Linear regression, Decision tree, Random forest, Neural network to generate fully as well as partially synthetic datasets [4]. We use statistical disclosure risk assessment techniques to assess the risk of disclosure of the synthetic datasets.

- **Publishing Models.** We publish the parameters of statistical machine learning algorithms perturbed with the functional mechanism [5] with a differential privacy guarantee [6]. For non-parametric models, as they require to release the training dataset along with the parameters to compute the output, we release a non-parametric model as a service wherein the training data and model parameters reside at the server and users send their queries to get the answers. We use functional perturbation [7] to provide differential privacy guarantees for non-parametric models that use kernels such as Kernel density estimation, Kernel SVM and Gaussian process regression.



PRIVACY AS A SERVICE

- We extend Liánchéng with disclosure risk assessment and privacy preservation operators. We refer to such a cloud system functionality as Privacy-as-a-Service.

- **Publishing Data**. We use traditional anonymisation techniques such as k-anonymity [1], l-diversity [2] and t-closeness [3]. Alternatively, we generate fake but realistic datasets by using machine learning models that are trained on private datasets. We use machine learning algorithms for Linear regression, Decision tree, Random forest, Neural network to generate fully as well as partially synthetic datasets [4]. We use statistical disclosure risk assessment techniques to assess the risk of disclosure of the synthetic datasets.

- **Publishing Models.** We publish the parameters of statistical machine learning algorithms perturbed with the functional mechanism [5] with a differential privacy guarantee [6]. For non-parametric models, as they require to release the training dataset along with the parameters to compute the output, we release a non-parametric model as a service wherein the training data and model parameters reside at the server and users send their queries to get the answers. We use functional perturbation [7] to provide differential privacy guarantees for non-parametric models that use kernels such as Kernel density estimation, Kernel SVM and Gaussian process regression.

DEMO SCENARIO

We show experiments on the 2000 US census dataset [8] that consists of 1% sample of the original census data. We select 212, 605 records, corresponding to heads of the households, and 6 attributes, namely, *Age, Gender, Race, Marital Status, Education, Income*.

We start by uploading the data into Liánchéng. We initiate the workflow with a filtering operator for data cleaning. We further extend the workflow by adding different operators. For instance, we use Linear regression operator to fit a regression model on a selection of attributes. We show the use of a trained model to synthetically generate a sensitive attribute such as Income in the dataset. We show the the application of the functional mechanism operator to release the model with differential privacy guarantees. For non-parametric models, we show the application of functional perturbation operator. We use different workflows to compare the effectiveness of differentially private machine learning algorithms with their non-private counterparts.

ACKNOWLEDGEMENT

This research is supported by the National Research Foundation, Prime Minister's Office, Singapore under its Corporate Laboratory@University Scheme, National University of Singapore, and Singapore Telecommunications Ltd.

REFERENCES

[1] Sweeney, L.: k-anonymity: A model for protecting privacy. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems 10 (05), 557–570 (2002)
[2] Machanavajjhala, A., Kifer, D., Gehrke, J., Venkitasubramaniam, M.: L-diversity: Privacy beyond k-anonymity. ACM Transactions on Knowledge Discovery from Data (TKDD) 1 (1), 3 (2007)
[3] Li, N., Li, T., Venkatasubramanian, S.: t-closeness: Privacy beyond k-anonymity and l-diversity. In: Data Engineering, 2007. ICDE 2007. IEEE 23rd International Conference on. pp. 106–115. IEEE (2007)
[4] Dandekar, A., Zen, R.A.M., Bressan, S.: A comparative study of synthetic dataset generation techniques. In: Database and Expert Systems Applications - 29th International Conference, DEXA 2018, Proceedings, Part II. pp. 387–395 (2018)
[5] Zhang, J., Zhang, Z., Xiao, X., Yang, Y., Winslett, M.: Functional mechanism: regression analysis under differential privacy. Proceedings of the VLDB Endowment 5(11), 1364–1375 (2012)
[6] Dwork, C., Roth, A., et al.: The algorithmic foundations of differential privacy. Foundations and Trends in Theoretical Computer Science 9(3-4), 211-407 (2014)
[7] Hall, R., Rinaldo, A., Wasserman, L.: Differential privacy for functions and functional data. Journal of Machine Learning Research 14 (Feb), 703–727 (2013)
[8] Minnesota population center. Integrated public use microdata series - international: Version 5.0. https://international.ipums.org. (2009)