Privacy as a Service: Publishing Data and Models

Ashish Dandekar[1], Debabrota Basu[1], Thomas Kister[1], Geong Sen Poh[2], Jia Xu[2], and Stéphane Bressan[1]

[1] School of Computing, National University of Singapore, Singapore
(ashishdandekar,debabrota.basu)@u.nus.edu, (dcsktpf,steph)@nus.edu.sg
[2] NUS-Singtel Cyber Security R & D Lab, Singapore
(geongsen.poh,jia.xu)@singtel.com

Abstract. The main obstacle to the development of sustainable and productive ecosystems leveraging data is the unavailability of robust, reliable and convenient privacy management tools and services. We propose to demonstrate our Privacy-as-a-Service system and Liánchéng, the Cloud system that hosts it. We consider not only the publication of data but also that of models created by parametric and non-parametric statistical machine learning algorithms. We illustrate the construction and execution of privacy preserving workflows using real-world datasets.

1 Introduction

A Wired's online article titled "The Privacy Revolt: The Growing Demand For Privacy-as-a-Service" is asking every company the privacy question: "What are you doing to provide Privacy-as-a-Service?"[3]. Indeed, the main obstacle to the development of sustainable and productive ecosystems leveraging data, including data market places, recommendation systems and crowd sourcing systems, is the unavailability of robust, reliable and convenient privacy management tools and services. This entails developing privacy risk assessment and privacy preservation algorithms, and integrating them into a service architecture.

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. We refer to this extension as a *Privacy-as-a-Service* model. 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.

[3] https://www.wired.com/insights/2015/03/privacy-revolt-growing-demand-privacy-service/

2 Liánchéng: Workflow as a Service

Liánchéng is a private data sharing Cloud service. The processing of data in Liánchéng is programmable by means of a Workflow-as-a-Service model. Liánchéng is deployed on a hardware infrastructure consisting of 128 commodity servers.

Data Sharing. This aspect is reminiscent of services such as Dropbox[4]. 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. Liánchéng provides a Workflow-as-a-Service model and interface that offers a compromise between the traditional Infrastructure-as-a-Service (IaaS) and Platform-as-a-Service (PaaS) models. 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. The interaction and visualisation are reminiscent of that of Yahoo Pipes[5] and other graphical workflow design software.

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*. We provide statistical disclosure risk assessment techniques such as uniqueness, overlap [8], and more advanced techniques [7, 15] as well as privacy preservation mechanisms such as k-anonymity [14] and differential privacy [5]. In particular, we propose a set of machine learning algorithms offering differential privacy guarantees by means of the functional mechanism [16] and of functional perturbation [6].

Figure 1 shows a screenshot of the composition window of a Liánchéng workflow that applies a differentially private linear regression onto a dataset and uses the parameters of the model to generate a synthetic version of the provided dataset that preserves the targeted utility.

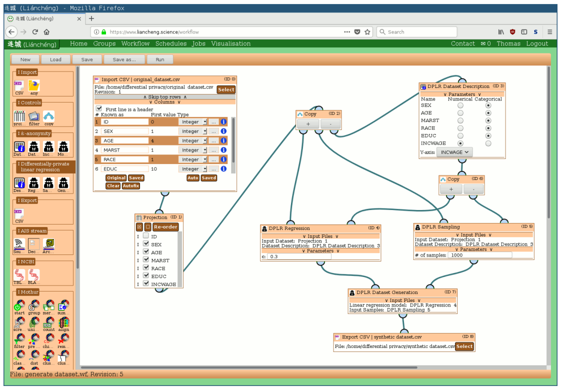
3 Private Data and Differentially Private Models

We consider both the publication of data and the publication of models created by the analysis of data by statistical machine learning algorithms. In both cases, there are risks of breach of privacy [2, 12].

In order to preserve privacy while publishing data, we use traditional anonymisation techniques such as k-anonymity [14], l-diversity [10] and t-closeness [9]. Alternatively, we generate fake but realistic datasets by using machine learning models that are trained on private datasets. We use machine learning algorithms

[4] https://www.dropbox.com

[5] http://radar.oreilly.com/2007/02/pipes-and-filters-for-the-inte.html


连城 (Liánchéng) - Mozilla Firefox
https://www.liancheng.science/workflow
Search
连城 (Liánchéng)
Home Groups Workflow Schedules Jobs Visualisation
Contact ✉ 0 Thomas Logout
New Load Save Save as... Run
Import
CSV any
Controls
proj... filter copy
k-anonymity
Dat... Dat... Inc... Mo...
Differentially-private linear regression
Des... Reg... Sa... Gen...
Export
CSV
AIS stream
Sou... Dec... Arc...
NCBI
TBL... BLA...
Mothur
start group mer... sum...
scre... uni... count align
filter pre... chi... rem...
clas... dist clus... clus...
Select
Original Saved Clear Autofix Auto Saved
Re-order
Copy (ID 2) + -
Copy (ID 6) + -

Import CSV | original_dataset.csv (ID 0)
File: /home/differential privacy/original dataset.csv
Revision: 1
∧ Skip top rows ∧
∨ Columns ∨
☑ First line is a header

#	Known as	First value	Type
1	ID	0	Integer
2	SEX	1	Integer
3	AGE	4	Integer
4	MARST	1	Integer
5	RACE	1	Integer
6	EDUC	10	Integer

Projection (ID 1)
□ ID
☑ SEX
☑ AGE
☑ MARST
☑ RACE
☑ EDUC
☑ INCWAGE

DPLR Dataset Description (ID 3)
∨ Parameters ∨

Name	Numerical	Categorical
SEX	○	●
AGE	○	●
MARST	○	●
RACE	○	●
EDUC	○	●
INCWAGE	●	○

Y-axis: INCWAGE

DPLR Regression (ID 4)
∨ Input Files ∨
Input Dataset: Projection 1
Dataset Description: DPLR Dataset Description 3
∨ Parameters ∨
ϵ: 0.3

DPLR Sampling (ID 5)
∨ Input Files ∨
Input Dataset: Projection 1
Dataset Description: DPLR Dataset Description 3
∨ Parameters ∨
# of samples: 1000

DPLR Dataset Generation (ID 7)
∨ Input Files ∨
Linear regression model: DPLR Regression 4
Input Samples: DPLR Sampling 5

Export CSV | synthetic dataset.csv (ID 8)
File: /home/differential privacy/synthetic dataset.csv

File: generate dataset.wf, Revision: 5

Fig. 1. A Liánchéng workflow generating a differential privacy compliant dataset.

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.

In order to preserve privacy while publishing parametric models, we publish the parameters of statistical machine learning algorithms perturbed with the functional mechanism [16] with a differential privacy guarantee [3]. For non-parametric models, as they require to release the training dataset along with the parameters to compute the output [11], 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 [6] to provide differential privacy guarantees for non-parametric models that use kernels [13] such as Kernel density estimation, Kernel SVM and Gaussian process regression.

4 Demo scenario

We show experiments on the 2000 US census dataset [1] 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. We also show similar privacy evaluations on the New York hospital inpatient discharge dataset[6].

[6] https://health.data.ny.gov/Health/Hospital-Inpatient-Discharges-SPARCS-De-Identified/u4ud-w55t

Acknowledgement

This project is supported by the National Research Foundation, Singapore Prime Minister's Office, under its Campus for Research Excellence and Technological Enterprise (CREATE) programme and under its Corporate Laboratory@University Scheme between National University of Singapore and Singapore Telecommunications Ltd.

References

1. Minnesota population center. integrated public use microdata series – international: Version 5.0. https://international.ipums.org. (2009)
2. Regulation (eu) 2016/679 general data protection regulation (text with eea relevance). Official Journal of the European Union **L**(119), 1–88 (2016), https://eur-lex.europa.eu/eli/reg/2016/679/oj
3. Dandekar, A., Basu, D., Bressan, S.: Differential privacy for regularised linear regression. In: Database and Expert Systems Applications - 29th International Conference, DEXA 2018, Proceedings, Part II. pp. 483–491 (2018)
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. 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)
6. Hall, R., Rinaldo, A., Wasserman, L.: Differential privacy for functions and functional data. Journal of Machine Learning Research **14**(Feb), 703–727 (2013)
7. Heyrani-Nobari, G., Boucelma, O., Bressan, S.: Privacy and anonymization as a service: Pass. In: International Conference on Database Systems for Advanced Applications. pp. 392–395. Springer (2010)
8. Hundepool, A., Domingo-Ferrer, J., Franconi, L., Giessing, S., Nordholt, E.S., Spicer, K., De Wolf, P.P.: Statistical disclosure control. John Wiley & Sons (2012)
9. 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)
10. 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)
11. Murphy, K.P.: Machine Learning: A Probabilistic Perspective. The MIT Press (2012)
12. Shokri, R., Stronati, M., Song, C., Shmatikov, V.: Membership inference attacks against machine learning models. In: Security and Privacy (SP), 2017 IEEE Symposium on. pp. 3–18. IEEE (2017)
13. Smola, A.J., Schölkopf, B.: Learning with kernels, vol. 4. Citeseer (1998)
14. Sweeney, L.: k-anonymity: A model for protecting privacy. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems **10**(05), 557–570 (2002)
15. Zare-Mirakabad, M.R., Jantan, A., Bressan, S.: Privacy risk diagnosis: Mining l-diversity. In: International Conference on Database Systems for Advanced Applications. pp. 216–230. Springer (2009)
16. 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)