DARC : Data Anonymization and Re-identification Challenge

Antoine Boutet, Mathieu Cunche, Sébastien Gambs, Benjamin Nguyen, Antoine Laurent

DARC : Data Anonymization and Re-identification Challenge

Antoine Boutet*, Mathieu Cunche*, Sébastien Gambs†, Antoine Laurent† and Benjamin Nguyen‡
*Univ Lyon, INSA Lyon, Inria, CITI
antoine.boutet@insa-lyon.fr, mathieu.cunche@insa-lyon.fr
†Université du Québec à Montréal
sebastien.gambs@uqam.ca, laurent.antoine@courrier.uqam.ca
‡INSA Centre Val de Loire, Laboratoire d'Informatique Fondamentale d'Orléans
benjamin.nguyen@insa-cvl.fr

***Résumé*—L'anonymisation est un des moyen de protection des données personnelles mis en avant par le nouveau règlement européens de protection des données personnelles (RGPD). Le projet DARC (Data Anonymization and Re-identification Challenge) est un challenge qui permet d'approcher le problème de l'anonymisation des données personnelles sous la forme d'un jeu. Ce challenge consiste dans un premier temps à anonymiser un jeu de données issu d'un site de vente en ligne, puis dans un second temps à essayer de ré-identifier les données anonymisées par les autres groupes. Un système de points évoluant positivement en fonction des ré-identifications effectués sur les autres groupes ou négativement si des ré-identifications sont effectués sur ses données alimente l'évolution du classement des groupes de manière interactif. Ce challenge d'anonymisation et de ré-identification de données s'est déroulé de Septembre à Novembre 2019 et a rassemblé trois formations d'étudiants d'INSA différents. Cette période a permis un approfondissement et une mise en pratique des enseignements au travers de la première phase d'anonymisation de données. Enfin, tous les étudiants ont été rassemblés sur deux jours sur le campus de l'INSA de Lyon pour effectuer en immersion la phase de ré-identification de données. Fort de cette première initiative réussie, le projet DARC est dorénavant mature pour ré-itérer l'expérience à plus large échelle avec un nombre plus important de formations impliquées.**

I. Introduction

Avec la nouvelle réglementation européenne relative à la protection des données personnelles (RGPD), le traitement des données à caractère personnel et la libre circulation de ces données son encadrés. En particulier, cette réglementation impose de rendre anonyme des données personnelles à des fins de traitement et d'échange. Une anonymisation consiste à supprimer toute possibilité de lier un individu (via par exemple son identifiant unique dans la base) à un ensemble de données personnelles (par exemple une liste de produits achetés). Plus concrètement, cela signifie que toutes les informations directement ou indirectement identifiantes doivent être supprimées ou modifiées, rendant impossible (ou tout du moins très difficile compte tenu de la connaissance de l'état de l'art) toute ré-identification des personnes. Avec une mise en application en 2018, cette réglementation a grandement développée l'activité de recherche et de développement ainsi que la mise en production de mécanisme d'anonymisation de données [1], [2], [3]. Pour plus de détails sur les techniques classiques d'anonymisations, le lecteur pourra se référer à l'article en français suivant [4].

Afin de sensibiliser et responsabiliser de manière ludique les futurs acteurs du numérique au sujet de la protection des données personnelles (objectifs inscrits dans la loi de refondation de l'école, l'Éducation aux Médias et à l'Information), nous avons expérimenté un challenge d'anonymisation et de ré-identification de données inter-INSA impliquant des étudiants de 4^{me} et 5^{me} année des départements Informatique de l'INSA de Lyon (IF), Télécommunications de l'INSA de Lyon (TC), et Sécurité et Technologies Informatiques de l'INSA Centre Val de Loire (STI). Ces étudiants suivaient tous une option en lien avec la cybersécurité et la protection des données personnelles dans leur formation respective.

Le challenge consiste à anonymiser un jeu de données provenant d'un site de vente en ligne. Il s'effectue en deux phases. Durant la première phase, les étudiants, organisés en groupes de 4 ou 5 personnes, doivent développer un mécanisme pour anonymiser ces données. La modification d'un jeu de données à des fins de protection induit de manière inhérente une perte d'information [5], [6]. Par conséquent, un jeu de métriques est fournit afin d'évaluer l'information utile conservée (au travers de métriques cherchant à caractériser l'efficacité d'algorithmes d'IA classiques comme le clustering ou la classification), mais également le risque de réidentification en utilisant quelques approches très naïves (par exemple une réidentification basée sur la date et la quantité de produits achetés, sur le produit acheté et le prix, etc.). Ensuite, durant la seconde phase, les étudiants disposent du jeu de données anonymisé des autres groupes et doivent essayer de re-identifier les utilisateurs, c'est à dire les lier avec les utilisateurs du jeu de données d'origine [7], [8].

Ce projet a plusieurs objectifs. Le principal objectif est de former les étudiants aux enjeux sociétaux et éthiques associés à la révolution numérique et au traitement massif de données. Un second objectif est de stimuler les interactions et mutualiser des moyens entre les différents départements et centres INSA. Et enfin, le troisième objectif est de faire de la transmission de connaissance de manière ludique et d'améliorer la motivation des étudiants.

Ce challenge s'est déroulé sur le semestre d'Automne 2019 et a rassemblé trois formation d'étudiants issus d'INSA différents. La première phase d'anonymisation de données s'est étalée sur plusieurs semaines entre Septembre et Novembre. Tous les étudiants ont ensuite été rassemblés deux jours fin Novembre sur le campus de Lyon pour effectuer la phase de ré-identification en immersion. Le groupe d'étudiants finalistes issu de l'INSA Centre Val de Loire ainsi qu'un membre de la meilleure équipe de chacune des deux autres formations de Lyon, ont été invités à venir assister à la conférence PETS (Privacy Enhancing Technologies Symposium) 2020 et à présenter leur solution d'anonymisation et de ré-identification lors d'un workshop associé à cette conférence dédiée au challenge organisé en France.

Fort de cette première initiative réussie et d'un retour d'expérience positif, le projet DARC est dorénavant mature pour ré-itérer l'expérience à plus large échelle avec un nombre plus important de formations impliquées. Nous envisageons donc d'inclure des étudiants issus d'autres formations dans la prochaine édition du challenge.

II. Le Challenge DARC

Ce challenge a été conçu en collaboration avec l'UQAM et sera programmé lors de la conférence internationale Privacy Enhancing Technologies Symposium [1] organisée à Montréal en 2020. La réalisation d'une première expérimentation de ce challenge sur des étudiants avait pour but de finaliser et affiner les règles du jeu et d'avancer sur la plate-forme permettant la soumission de jeux de données anonymisés et le calcul des points pour le classement résultant de la ré-identification des autres groupes de manière automatique.

A. Objectif Pédagogique

D'un point de vue pédagogique, ce projet s'inscrit dans l'axe ludification et apprentissage par le jeu. Les étudiants manipulent et contextualisent les outils de protection des données personnelles et concepts théoriques associés au travers d'un cas d'usage concret. De plus, cette pédagogie active permet de générer des apprentissages à travers la réalisation d'un projet global, allant de l'anonymisation d'un jeu de données réel à la ré-identification de données anonymisées afin de mieux appréhender les risques. Enfin, l'intégration d'un élément de jeu dans le système de classement entre les groupes améliore l'investissement des étudiants et augmente leur motivation et l'interaction entre eux.

En plus d'apporter un caractère structurant entre les différents départements et différents centres INSA, ce projet permet également de sensibiliser les étudiants à la recherche scientifique et d'étendre leur réflexion au-delà de la simple synthèse bibliographique et de « toucher du doigt » des problématiques approfondies liées à l'anonymisation et la ré-identification de données personnelles.

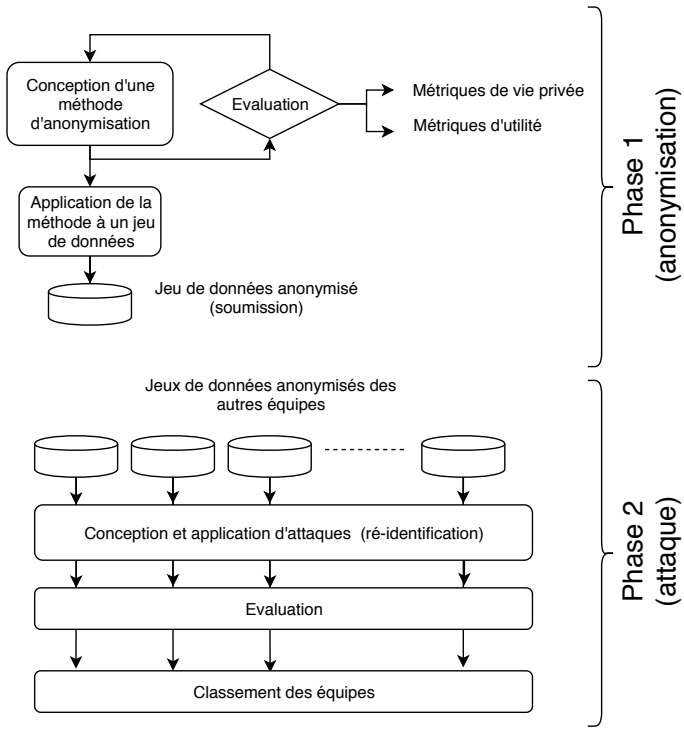

FIGURE 1. Principe de fonctionnement du challenge en deux phases 1) Anonymisation d'un jeu de données 2) Attaque de ré-identification sur les jeux de données anonymisés des autres équipes.

B. Déroulement du Challenge

Le challenge d'anonymisation et de ré-identification de données DARC s'est déroulé sur le semestre d'Automne 2019. Les spécialités des trois promotions participantes étaient toutes centrées sur la cybersécurité et la protection des données personnelles. Les premiers cours sur la protection des données personnelles et l'éthique du numérique ont commencés fin Septembre. Le challenge DARC, quant à lui, a débuté en Octobre et s'est achevé fin Novembre.

Les étudiants se sont organisés en groupe de quatre à six personnes. Le jeu de données utilisé dans ce challenge provenait d'un site de vente en ligne [9] et est publiquement accessible [2]. Ce challenge s'est déroulé en deux phases : la phase d'anonymisation de données et la phase de ré-identification (voir Figure 1). L'ensemble des règles du challenge ont été rassemblées dans un document remis à l'ensemble des participants.

Durant la première phase, chaque groupe devait développer un mécanisme pour anonymiser le jeu de données. Afin d'évaluer l'impact de leur mécanisme d'anonymisation à la fois sur la perte d'information utile ainsi que sur la protection des données, les règles contenaient un set de métriques d'utilité et de vie privée. Les métriques d'utilité correspondaient à des mesures permettant d'analyser les données telles que des

1. https://petsymposium.org

2. https://archive.ics.uci.edu/ml/datasets/Online+Retail

mesures de similarité inter objets, la moyenne des différences de date d'achat et de quantité ou le nombre d'entrées du jeu de données supprimées. De leur côté, les métriques de vie privée correspondaient à un taux de ré-identification réalisé en fonction de certains critères tels que la date ou le prix par exemple. Ces deuxièmes métriques étaient basées sur des calculs relativement naïfs, ainsi un très bon score sur une telle métrique n'était pas la garantie d'avoir une excellente sécurité au final. Toutefois un mauvais score à ces métriques était rédhibitoire car représentait clairement une faiblesse dans le processus d'anonymisation.

L'ensemble des métriques étaient implémentées via des scripts python. Grâce à ces métriques, les étudiants ont pu donc itérativement améliorer leur mécanisme de protection. Cette première phase s'est achevée mi Novembre.

FIGURE 2. Des étudiants sur le campus de l'INSA de Lyon durant la seconde phase dédiée à la ré-identification des jeux de données anonymisés des autres groupes.

Les étudiants se sont ensuite préparés pour la second phase consistant à dé-anonymiser le jeu de données protégés des autres groupes (Figure 2). Pour cela, ils ont préparé quelques attaques en utilisant leurs propres données anonymisées. Cette analyse sur leur propre jeu de données anonymisés leur a également permis d'améliorer leur solution de protection. A l'issu de cette période, chaque groupe a sélectionné deux jeux de données anonymisés fournissant le meilleur compromis entre utilité et vie privée pour le reste de la seconde phase. Cette seconde phase s'est clôturée les 25 et 26 Novembre sur le campus de Lyon avec l'ensemble des participants par deux sessions de ré-identification en immersion. Durant ces deux sessions, les étudiants se confrontaient de manière interactive et ludique aux autres groupes en exploitant les attaques conçues. Plus précisément, chaque groupe avait accès aux jeux de données anonymisés des autre groupes et essayait de ré-identifier le plus d'utilisateurs, c'est à dire à identifier l'association entre identifiant d'utilisateur d'une entrée anonymisée avec l'identificant d'une entrée du jeu de données d'origine.

Le système de classement élaboré à l'origine calculait un score en fonction des métriques liées au jeu de données anonymisés et des ré-identifications réussies entre les groupes. Plus spécifiquement, un groupe marque des points si il ré-identifie correctement des utilisateurs d'un autre groupe, ce dernier en perd par la même occasion. L'objectif de ce système de classement est d'engendrer une interaction entre les groupes où le plus haut classé est le plus susceptible d'être attaqué par les autres groupes. Les étudiants ont donc dû s'adapter et adopter une bonne stratégie afin de ne pas se faire doubler par les autres groupes concurrents.

C. Groupes Finalistes

Le groupe d'étudiants finalistes issu de l'INSA Centre Val de Loire a été invité à venir assister à la conférence PETS et à présenter un poster décrivant leur solution d'anonymisation et de ré-identification lors d'un workshop associé au challenge. Un membre de la meilleure équipe de chacune des deux autres formations de Lyon est également invité à assister à la conférence PETS et au workshop associé au challenge. De plus, tous les étudiants sont conviés à participer à la compétition organisée dans le cadre de la conférence PETS.

D. Retours d'Expérience

Malgré l'indisponibilité de la plate-forme permettant le recueil des jeux de données anonymisés (Phase 1) ainsi que les attaques de ré-identification avec calcul automatique des scores (Phase 2), cette première initiative a été très bien accueillie par les étudiants. Cette plate-forme devait être hébergée sur CrowdAI [3] et demandait un temps d'intégration de la part de l'équipe d'administration de la plate-forme pour installer le support nécessaire aux calculs automatiques des métriques de références liées à la Phase 1 et au classement interactif de la Phase 2. Malheureusement, l'équipe d'administration n'a pas pu réaliser cette intégration avant la date de notre rassemblement sur le campus de Lyon pour effectuer la Phase 2. Cette indisponibilité nous a imposé des tâches d'évaluation manuelles afin de construire le classement interactif durant l'évènement (affiché via un document partagé sur le web). A l'avenir, afin d'éviter une dépendance à un service tiers, nous considérons le développement de notre propre solution de gestion du challenge.

La deuxième difficulté est liée à la participation de différentes formations avec des contraintes d'emplois du temps légèrement différentes. Ces contraintes ont impacté le quota d'heures dédiées au challenge qui était différent entre formations, ainsi que la difficulté de trouver un créneau pour rassembler les étudiants pour la Phase 2 qui avait la contrainte de durer plusieurs jours.

Ces contraintes n'ont également pas rendu possible l'organisation de la Phase 2 en mode hackathon intégrant une nuit à l'évènement sur le campus de Lyon.

3. https://www.crowdai.org

III. Conclusion

Afin de sensibiliser et responsabiliser de manière ludique les futurs acteurs du numérique au sujet de la protection des données personnelles, nous avons expérimenté un challenge d'anonymisation et de ré-identification de données inter INSA.

Ce challenge consistait dans un premier temps à anonymiser un jeux de données provenant d'un site de vente en ligne. Ensuite, durant une seconde phase, les étudiants devaient attaquer le jeu de données anonymisés des autres groupes afin de re-identifier les données en liant les utilisateurs du jeu de données anonymisés aux données d'origine.

Fort de cette première initiative réussie et d'un retour d'expérience positif, le projet DARC est dorénavant mature pour ré-itérer l'expérience à plus large échelle avec un nombre plus important de formations impliquées.

Remerciements

Ce challenge a été financé en partie par le Groupe INSA, l'INSA de Lyon, l'INSA Centre Val de Loire et par le Projet IDEXLYON de l'Université de Lyon dans le cadre du Programme Investissements d'Avenir (ANR-16-IDEX-0005).

Références

[1] A. Boutet and S. Gambs, "Demo : Inspect what your location history reveals about you Raising user awareness on privacy threats associated with disclosing his location data," in *The 28th ACM International Conference on Information and Knowledge Management (CIKM 2019)*, Nov. 2019.

[2] V. Primault, A. Boutet, S. Ben Mokhtar, and L. Brunie, "The Long Road to Computational Location Privacy : A Survey," *Communications Surveys and Tutorials, IEEE Communications Society*, p. 1, 2018.

[3] R. Pires, D. Goltzsche, S. Ben Mokhtar, S. Bouchenak, A. Boutet, P. Felber, R. Kapitza, M. Pasin, and V. Schiavoni, "CYCLOSA : Decentralizing Private Web Search Through SGX-Based Browser Extensions," in *The 38th IEEE International Conference on Distributed Computing Systems (ICDCS 2018)*, Jul. 2018, pp. 467–477.

[4] B. Nguyen and C. Castelluccia, "Techniques d'anonymisation tabulaire : Concepts et mise en oeuvre," *Bulletin 1024*, no. 15, pp. 1–21, 2020, à paraître.

[5] S. Cerf, V. Primault, A. Boutet, S. Ben Mokhtar, R. Birke, S. Bouchenak, L. Y. Chen, N. Marchand, and B. Robu, "PULP : Achieving Privacy and Utility Trade-off in User Mobility Data," in *The 36th IEEE International Symposium on Reliable Distributed Systems (SRDS 2017)*, Sep. 2017, pp. 164–173.

[6] V. Primault, A. Boutet, S. Ben Mokhtar, and L. Brunie, "Adaptive Location Privacy with ALP," in *The 35th Symposium on Reliable Distributed Systems (SRDS 2016)*, Sep. 2016.

[7] S. Gambs, M.-O. Killijian, and M. N. del Prado Cortez, "De-anonymization attack on geolocated data," *Journal of Computer and System Sciences*, vol. 80, no. 8, pp. 1597–1614, 2014.

[8] A. Petit, T. Cerqueus, A. Boutet, S. Ben Mokhtar, D. Coquil, L. Brunie, and H. Kosch, "SimAttack : private web search under fire," *Journal of Internet Services and Applications*, p. 17, Mar. 2016.

[9] D. Chen, S. L. Sain, and K. Guo, "Data mining for the online retail industry : A case study of rfm model-based customer segmentation using data mining," *Journal of Database Marketing & Customer Strategy Management*, vol. 19, no. 3, pp. 197–208, Sep 2012.