Enseignement de la sécurité des données individuelles en STID

Paul-Marie Grollemund, Clément Jacq, Kevin Thiry-Atighehchi

Université Clermont-Auvergne, IUT Clermont Auvergne, Campus Simone Veil, Aurillac, France

***Résumé*—Le département STID (Statistique et Informatique Décisionnelle) d'Aurillac a ouvert en 2019. La spécificité de cette formation est une adaptation locale du programme pédagogique national à hauteur de 30% pour y enseigner la cybersécurité, avec la volonté de couvrir un spectre assez large. Au regard du programme initial, la formation se prête bien à l'enseignement de la sécurité des microdonnées. Ce document en présente une partie du programme et des éléments contextuels.**

I. Introduction

La formation STID propose aux étudiants d'acquérir les compétences essentielles pour la gestion informatique des données, leur traitement statistique et l'informatique décisionnelle. Les diplômés sont capables de collecter des données et contrôler leur qualité, de les structurer dans des bases de données relationnelles, puis d'administrer ces base de données. Dans leur démarche de statisticiens, ils savent extraire et présenter des informations pertinentes, effectuer une analyse statistique et communiquer des résultats. Dans le domaine de l'informatique décisionnelle, ils participent à la mise en place et à l'exploitation de systèmes d'information décisionnels et peuvent concevoir des indicateurs de performance ou des tableaux de bord. Par ailleurs, les diplômés du département STID d'Aurillac disposent de compétences en cybersécurité leur permettant de mieux appréhender la réglementation entourant les données personnelles. Ils connaissent les risques, les bonnes pratiques et des outils pour sécuriser un système d'information.

Un des atouts de cette formation est de former des étudiants de premier cycle à la cybersécurité, permettant d'aboder des notions disciplinaires élémentaires (mathématiques, statistiques, informatique, droit et économie) sous le prisme de la cybersécurité. L'objectif est de donner aux étudiants l'opportunité de devenir rapidement des techniciens de la cybersécurité ou de pouvoir accéder à des formations de deuxième cycle plus poussées autour de la cybersécurité.

Afin de former des étudiants avec les objectifs susmentionés, l'équipe pédagogique colore la formation de plusieurs manières : 1) des conférences sont organisées avec des invités spécialistes de la cybersécurité, 2) les contenus pédagogiques sont illustrés dans des contextes pratiques de cybersécurité et 3) des projets importants et techniques sont mis en place afin de scénariser l'application de notions de disciplines différentes dans le contexte de la cybersécurité.

Le propos de cet article est justement de présenter les principaux projets menés avec les étudiants autour de la cybersécurité. La section II décrit des projets relatifs à la sécurité des bases de données, traitant de contrôle d'accès, d'anonymisation des données ou encore de l'utilisation d'une ETL. Un autre projet est présenté en section III, traitant sous des prismes disciplinaires multiples de la possibilité de conserver le secret médical dans un contexte épidémique.

II. Sécurité des bases de données

A. Contrôle d'accès

Les méthodes de contrôle d'accès discrétionnaire, d'accès obligatoire et d'accès basé sur les rôles sont introduites dans le cadre des systèmes de gestion de base de données (SGBD). Les objectifs sont l'acquisition des connaissances générales et des connaissances pratiques en langage SQL pour créer des rôles, octroyer et révoquer des privilèges ou des rôles. Nous proposons des exercices avec des bases de données relationnelles où une réfléxion est menée sur les droits d'accès à définir en respectant le principe du moindre privilège, et comment mettre en œuvre les contrôles d'accès. Par exemple, dans une base de données médicale, il est demandé aux étudiants de mettre en place un système RBAC en idenfiant les rôles et en les hiérarchisant par généralisation/spécialisation. Il est ensuite demandé d'écrire les commandes pour : 1) créer les rôles et les hiérarchiser , 2) assigner les rôles aux utilisateurs, et 3) octroyer des privilèges aux rôles.

B. Anonymisation des données

Après avoir rappelé le cadre réglementaire (RGPD et loi Informatique et Libertés) abordé dans le cours de droit, nous évoquons avec des exemples les risques en matières d'anonymisation, selon les trois critères relevés par le groupe de travail G29 : l'individualisation, la corrélation et l'inférence.

Une des premières notions abordées est celle de l'inférence dans les bases de données statistiques. Dans ces bases de données, les informations ne sont accessibles que par des requêtes statistiques, des fonctions d'aggrégation du type somme, moyenne, comptage, écart-type, maximum, etc. Au travers d'exemples, il est montré aux étudiants qu'un problème de divulgation inférencielle peut se produire dans des bases de données en apparence sécurisée, à savoir qu'il est possible d'accéder à des informations sensibles (individuelles) à partir d'informations non sensibles. Avec des exemples d'attaques du type traqueur (tracker) individuel et traqueur généralisé, nos étudiants comprennent que de simples contraintes sur le nombre d'enregistrements qui doit correspondre à une condition logique d'une requête ne sont pas des contremesures suffisantes pour se prémunir d'attaques par inférence. En parcourant d'autres mécanismes de l'état de l'art prévenant ou contrôlant l'inférence, l'objectif est de leur montrer qu'il peut être difficile d'obtenir une bonne confidentialité des données tout en permettant de calculer des résumés statistiques utiles.

Concernant le traitement des données à des fin de publication, la première technique abordée est celle de la pseudonymisation afin d'enlever à un jeu de données tout attribut directement identifiant. Quelques attaques bien connues sont citées pour montrer que cette tehcnique d'anonymisation ne suffit pas à se prémunir des risques énoncés par le G29 (exemples du prix Netflix, et du recoupement par Latanya

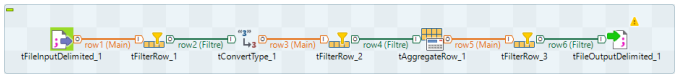

FIGURE 1. Exemple de job Talend effectuant un simple nettoyage de données (éliminations des valeurs exotiques, consolidation des doublons)

Sweeney d'une base de données médicale pseudonymisée avec une liste électorale). Les techniques par généralisation préservant du risque d'individualisation (k-anonymat [1] et l-diversité [2]) sont abordées en travaux dirigés sur de petits jeux de données fictifs, et sur de gros jeux de données dans le cadre de travaux pratiques sur le logiciel Talend présentés dans la section suivante. Enfin, la confidentialité différentielle [3] est abordée au premier semestre dans un module de méthodes d'enquêtes pour éviter les sources d'erreur [1] puis au troisième semestre dans une étude de cas sur le secret médical dans un contexte épidémique.

C. Utilisation d'un ETL pour l'anonymisation des données

Le module de deuxième année Système d'information décisionnel (SID) a deux objectifs principaux. D'une part, il décrit le fonctionnement de la chaine décisionnelle, dont il recontextualise les composantes, et les contraintes logistiques inhérentes à leurs interactions. D'autre part, il approfondit l'étape d'extraction et d'harmonisation des données en introduisant les ETL (Extract, Load, Transform), des outils logiciels conçus pour cette étape. C'est dans le cadre de ce second objectif que l'équipe pédagogique a choisi d'introduire le logiciel Talend, un outil ETL freemium très polyvalent dont la pratique occupe la majeure partie des heures du module SID. Talend est un outil visuel dans lequel la création d'une séquence de nettoyage des données (job) est construite en reliant graphiquement des composants paramétrables effectuant des opérations simples (agrégation, filtrage, jointure, conversion). Un des projets proposés pour l'apprentissage du logiciel concerne l'anonymisation des données. L'étude se concentre principalement sur deux des critères classiques : la k-anonymité et la l-diversité.

Structuré autour d'un schéma fixé de base de données inspiré d'une base de données médicale, le projet se déroule en trois actes. Dans un premier temps, les étudiants doivent produire un job talend permettant de calculer le degré de vérification des critères d'anonymisation d'une base de données. Dans un second temps, ils doivent construire plusieurs jobs effectuant des opérations d'anonymisation classiques telles que le shuffling, la pseudonymisation ou la généralisation, et ce en utilisant que les fonctionnalités gratuites du logiciel. Enfin, les étudiants doivent choisir une ou plusieurs combinaisons des jobs précédemment créés dans l'optique d'augmenter les degrés de k-anonymité et de l-diversité de plusieurs bases de données tout en justifiant que leurs traitements permettent toujours une étude pertinente des données. Il leur faut parfois revoir les méthodes d'anonymisation implémentées, trop naïves ou trop peu flexibles pour produire des résultats positifs. En plus des aspects techniques de la manipulation de Talend, le projet est conçu pour que les étudiants se rendent compte des difficulté d'application des diverses techniques d'anonymisation face au contenu des bases de données.

1. La technique de la réponse aléatoire est employée pour obtenir des réponses fiables mais en partie bruitées à des questions sensibles portant sur l'illégalité, l'immoralité, ou une caractéristique *a priori* inavouable.

III. LE SECRET MÉDICAL DANS UN CONTEXTE ÉPIDÉMIQUE

Dans le cadre de la formation, le programme propose la possibiltié de mener avec les étudiants un travail concret, au travers du module "Etude de cas en statistique et informatique décisionnelle". Ce module a pour vocation de mettre en concert des intervenants de plusieurs disciplines au service de l'étude d'une question concrète et réaliste. La question abordée avec les étudiants est celle du secret médical dans un contexte épidémique.

A. Les objectifs

Cette étude a pour objectif d'aiguiser le sens critique des étudiants concernant la nécessité de la protection des données personnelles, ainsi que plus généralement de devélopper leurs capacités à comprendre et à modéliser une situation concrète.

Concernant la protection des données personnelles, les étudiants sont introduits à la confidentialité différentielle, qui est une approche à utiliser dans le cadre de questionnaire contenant des questions sensibles (*i.e.* auxquelles on serait tenté de ne pas répondre honnêtement). De plus, il s'agit d'une occasion supplémentaire d'aborder avec les étudiants de ce cursus les termes du RGPD. Un axe important du RGPD indique qu'il est recommande de ne pas demander plus d'informations que nécessaire, relativement aux besoins de la situation qui sous-tendent la demande sur ces données personnelles. Le cas concret étudié dans ce travail permet aux étudiants d'apprécier cette ligne floue que définit le RGPD. Cela permet aux étudiants de se rendre compte que des approches pourraient être mises en place pour gérer une crise sanitaire, tout en permettant de conserver le plus possible le secret médical et la confidentialité des informations personnelles.

Le sens critique à développer dans ce cas concerne la levée du secret médical, qui de prime abord se justifie par la nécessité de devoir gérer une épidémie d'un point de vue national. La réflexion à mener ici est de comprendre les arguments quant à la levée de ce secret médical, ainsi que de comprendre d'un point de vue du droit qu'est-ce qui justifie de le lever, et ce en notant ce qui pourraient être les raisons de ne pas le lever. Un premier constat serait que le lever du secret médical convient à un cas de force majeure, et surtout s'il n'y a pas d'autres recours à mettre en place afin de gérer efficacement une crise sanitaire. Cependant, la question à poser est de savoir s'il n'y aurait pas d'autres leviers qui pourraient être adaptés à la gestion de la situation, tout en conservant la confidentialité des données personnelles. Même si ces leviers pourraient être moins efficaces, leur alignement avec les recommandations du RGPD en font des approches à étudier. Le propos de ce sujet est entre autres de fournir un contexte suffisamment simplifié de sorte à pouvoir réfléchir en ces termes sur la levée du secret médical et de pouvoir comprendre qu'il ne s'agit pas d'un mécanisme automatique.

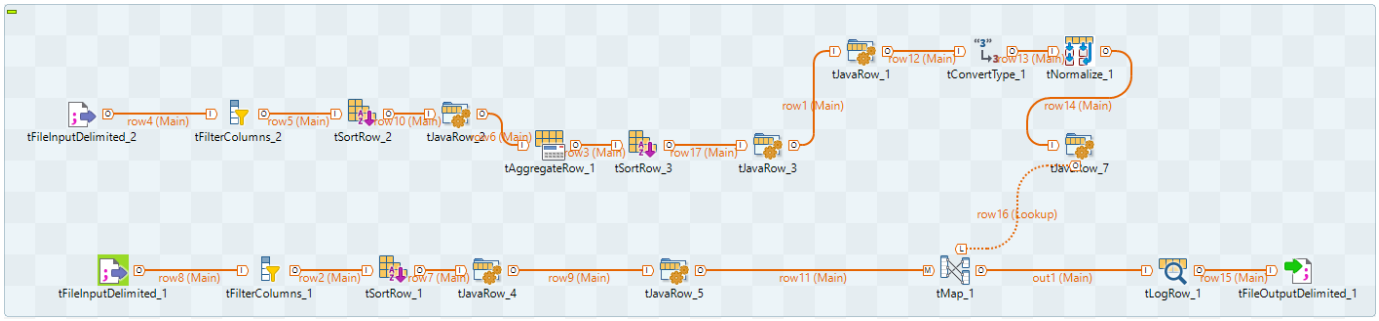

FIGURE 2. Job final produit par un étudiant dans le but d'augmenter le degré de l-diversité d'une table

Pour les aspects de modélisation, ce sujet permet d'introduire aux étudiants 1) les propriétés statistiques des estimateurs sur les résultats d'une enquêtes incluant de la confidentialité différentielle, 2) les outils de modélisation mathématique pour traiter de l'évolution d'une épidémie, et 3) le moyen de prendre en compte une incertitude avec une approche probabiliste.

B. Les étapes de travail avec les étudiants

1) Mise en place et compréhension de la confidentialité différentielle: Afin de compléter ce que les étudiants connaissent de la confidentialité différentielle, cette étude de cas comporte deux volets concernant cette méthode. Le premier est la mise en place d'un questionnaire numérique incluant des notions de confidentialité différentielle, et le second consiste à étudier des propriétés statistiques des estimateurs à utiliser pour analyser les données résultantes de ce type de questionnaire.

Pour la mise en place, les étudiants doivent implémenter une application Rshiny correspondant à un questionnaire de santé basé sur une approche de confidentialité différentielle. L'utilisation de Rshiny permet d'obtenir aisément et assez rapidement un rendu correct pour une application, avec une minimum d'effort de programmation. De plus, Rshiny permet d'inclure des éléments aléatoires dans l'application (relatifs à la confidentialité différentielle) assez simplement grâce à du code R. Ainsi, lorsque l'application s'exécute pour ouvrir un questionnaire vide (voir la figure III-B1), cela commence par déterminer des réponses structurées pour les questions sensibles, de sorte à ce que de manière aléatoire : soit le champ de la réponse soit bloqué avec une réponse par défaut déterminée aléatoirement, soit un champ libre avec une mention "Répondez honnêtement à la question". Coder cette application permet aux étudiants de se rendre compte qu'il n'est pas particulièrement complexe de mettre en place et de diffuser à grande échelle un questionnaire numérique permettant d'obtenir des informations concernant l'état de santé des citoyens.

Pour le second volet, la spécificité de la confidentialité différentielle permet d'aborder des notions standards, mais avec des résultats originaux. En particulier, il est possible d'aborder avec les étudiants l'espérance de l'estimateur à utiliser habituellement pour traiter des résultats d'un questionnaire. Du fait de l'utilisation de la confidentialité différentielle, l'estimateur classique est biaisé, puisqu'il y a une partie inconnue

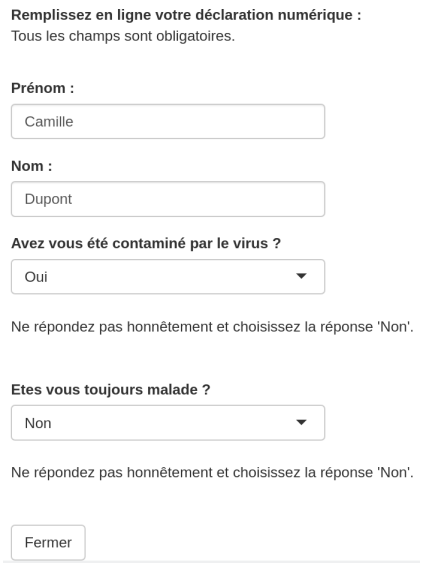

FIGURE 3. Production rapide avec R Shiny d'un formulaire d'enquête en ligne. La réponse est forcée ou suggérée par l'utilisation d'un mécanisme de confidentialité différentielle.

des données qui est totalement aléatoire. Un intérêt pour les étudiants est que de débiaiser cet estimateur correspond à un mécanisme qu'ils connaissent déjà, au travers du débiaisement de l'estimateur de la variance, mais en se faisant avec des calculs différents et plus complexes. Pour finir, l'étude de la variance de cet estimateur permet de voir dans un cas simple et inhabituel ce type de calculs, et ce qui peut de plus donner lieu à une étude par simulation afin de vérifier les résultats théoriques.

2) Modélisation d'une épidémie avec le modèle SIR: Pour modéliser la dynamique de l'épidémie, une introduction aux systèmes d'équations différentielles et, en particulier au modèle SIR, est dispensée aux étudiants. Ce modèle est à la fois abordable pour des étudiants d'un point de vue mathématique, mais il est de plus flexible et assez facilement interprétable. Cela permet en particulier de pouvoir aborder avec les étudiants comment de nouveaux compartiments et de nouveaux flux peuvent être incluent dans ce type d'approche, et comment ceux-là peuvent être définis de sorte à modéliser des aspects particuliers d'une épidémie (vaccination, perte d'immunité, sensibilité suivant des facteurs socio-démographiques, ou autres).

3) Décision autour d'une campagne de vaccination: Pour prendre une décision concernant le fait de réaliser ou non une campagne de vaccination, et si oui avec quelle niveau d'intensité, dans un contexte épidémique décrit par un modèle de type SIR, nous traitons avec les étudiants de la modélisation de ce que pourrait être un "coût sociétal et moral" d'une épidémie sur une période donnée. Pour modéliser ce coût, qui ne contient pas que des aspects financiers, les étudiants ont pu inclure des facteurs de coût comme le coût moral du à des décès, le coût de la prise en charge de personnes infectées, le coût du vaccin en lui-même (recherche, production, logistique), ou encore le coût associé à une perte de chance causée par le faite d'être infecté. Bien que la modélisation de ce coût soit simpliste et abstraite, cela permet aux étudiants d'évaluer les différents impacts d'une épidémie et de sa gestion sur la société et les citoyens. De plus, le calcul pratique du coût considéré dans ce travail correspond à fonction non-monotone du niveau d'intensité de la campagne de vaccination, ce qui permet d'appréhender qu'un optimum du coût peut exister en réalisant un compromis entre plusieurs sources de coûr et, autrement dit, qu'une campagne de vaccination peut être une outils dont l'impact est à controler.

4) Modélisation l'incertitude due à la confidentialité différentielle: Du faite de l'utilisation de la confidentialité différentielle, et des estimations qui en découlent, on peut obtenir des résultats entâchés d'une incertitude particulière. Non seulement il y a une incertitude inhérente à toute procédure statistique, qui découle d'un échantillonnage de la population cible, mais en plus une autre source d'incertitude découle de la procédure de confidentialité différentielle qui peut s'assimiler à du bruitage d'une partie des données. Afin de prendre en compte cette seconde incertitude de l'état de la population concernant la maladie, il est proposé aux étudiants de procéder à une phase de simulation supplémentaire dans la modélisation de l'évolution de l'épidémie. Pour ce qui suit, on note I la part réelle (mais inconnue) de la population qui est infectée et $\tilde{I}$ l'estimation de celle-ci obtenue grâce un questionnaire utilisant la confidentialité différentielle. On peut estimer la variance de cette estimateur $\hat{\sigma}^2_{\tilde{I}}$ et faire une approximation de la loi de l'estimateur $\tilde{I}$ par une loi Normale, centrée en I et de variance $\hat{\sigma}^2_{\tilde{I}}$. Cela permet de justifier qu'on puisse simuler numériquement une valeur potentielle de I à partir de cette approximation gaussienne, ce qui donne une des valeurs initiales pour le modèle SIR. En réalisant des simulations similaires pour les autres sous-populations concernées (S et R), on peut obtenir de quoi simuler une trajectoire de l'évolution de l'épidémie. En répétant cela une multitude de fois, pour de nouvelles simulations gaussiennes, on peut obtenir des trajectoires potentielles de l'évolution de l'épidémie, représentant ce que pourrait être l'évolution de l'épidémie conditionnellement à notre incertitude des valeus initiales des sous-populations d'intérêt. Pour le formuler plus directement, il s'agit d'utiliser une approche de Monte Carlo, afin de pouvoir appréhender des quantités d'intérêt comme le temps de retour à un niveau donné de l'épidémie, ou la quantité attendue de personnes à décéder, ou autres. Tout ces quantités peuvent être à la source d'une prise de décision concernant la gestion d'une crise sanitaire : "Doit-on confiner ? Combien de temps ? Comment ? Doit-on vacciner ? Qui ? A quel intensité ?". Cela permet de donner aux étudiants l'intuition de ce à quoi peut correspondre une prise de décision dans un contexte d'incertitude.

IV. Conclusion

Afin de traiter de la sécurité des données et de notions de cybersécurité, nous avons complété le programme avec des projets innovants, qui prennent un appui cohérent sur les connaissances et les compétences acquises par les étudiants durant ce parcours.

Le projet sur l'outil Talend a permis aux étudiants d'implémenter de façon originale les techniques d'anonymisation classiques. Concernant le projet relatif à l'étude d'une épidémie en accord avec la préservation du secret médical, nous avons introduit et mis en pratique dans un contexte complexe une méthode de confidentialtié différentielle permettant de traiter de la préservation des données personnelles. Cela a été l'occasion de traiter de nombreuses notions et compétences annexes, que celles-ci soient relatives aux mathématiques, au droit, à la médecine ou à des approches stochastiques.

L'ensemble de ces projets permettent de mettre en lumière l'adéquation de ce parcours multi-compétences (informatique et statistique) avec des notions pointues en cybersécurité. Au travers des développements proposés par l'équipe pédagogique du département STID d'Aurillac, nous avons donné du liant aux différentes disciplines du département, ainsi que rendu plus concrètes et accessibles les notions vues en cours et les notions hors-programme de cybersécurité.

Relativement au travail proposé aux étudiants, une suite pertinente serait d'étendre les travaux pratiques introductifs à d'autres algorithmes de confidentialité différentielle. Notamment en proposant aux étudiants d'implémenter d'autres approches pour déterminer une réponse aléatoire, toujours dans le cas d'une variable catégorielle, puis de comparer les effets qu'induisent ces processus aléatoires instanciés avec différents paramètres sur l'intervalle de confiance des estimateurs. Les étudiants pourront réfléchir à d'autres indicateurs que celui de la précision des résultats, en particulier l'effet sur le déni possible d'un individu de sa propriété altérée par ces méthodes, ou encore l'effet des paramètres choisis sur le niveau de protection qui devrait alors être contrebalancé avec la précision des résultats.

Reconnaissance

L'équipe pédagogique du département STID d'Aurillac remercie Béatrice Collay, Michel Misson et Pascal Lafourcade pour leurs implications importantes dans l'émergence de ce département.

Références

[1] Latanya Sweeney. k-anonymity : A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(05) :557–570, 2002.

[2] Ashwin Machanavajjhala, Daniel Kifer, Johannes Gehrke, and Muthuramakrishnan Venkitasubramaniam. l-diversity : Privacy beyond k-anonymity. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 1(1) :3–es, 2007.

[3] Cynthia Dwork, Aaron Roth, et al. The algorithmic foundations of differential privacy. *Found. Trends Theor. Comput. Sci.*, 9 (3-4) :211–407, 2014.