جامعة محمد الخامس بالرباط
Université Mohammed V de Rabat
École Nationale Supérieure d'Informatique et d'Analyse des Systèmes
Centre d'Études Doctorales en Sciences des Technologies de l'Information et de l'Ingénieur

AVIS DE SOUTENANCE DE THÈSE DE DOCTORAT

Monsieur Zakariae EL OUAZZANI

soutiendra publiquement sa thèse de Doctorat en Informatique
Le Samedi 19 Juin 2021 à 10h30 au Grand Amphi à l'ENSIAS

Intitulé de la thèse

ANONYMIZATION TECHNIQUES ENSURING PRIVACY IN E-HEALTH

Devant le Jury composé de :

Président :

Pr. Rachida AJHOUN, PES, ENSIAS, Université Mohammed V de Rabat

Directeur de thèse :

Pr. Hanan EL BAKKALI, PES, ENSIAS, Université Mohammed V de Rabat

Rapporteurs :

Pr. Abdelkrim HAQIQ, PES, FST, Université Hassan 1er de Settat

Pr. Nabil BENAMAR, PES, EST, Université Moulay Ismail de Meknès

Pr. Mostapha ZBAKH, PES, ENSIAS, Université Mohammed V de Rabat

Examinateurs :

Pr. An BRAEKEN, Associate Professor, Vrije Universiteit Brussel (VUB), Belgique

Pr. Abderrahmane NITAJ, Professeur HDR, Université de Caen Normandie, France

ANONYMIZATION TECHNIQUES ENSURING PRIVACY IN E-HEALTH

Abstract: Recently, collecting data has become crucial for most organizations due to the fast growth of data analytics tools that allow better use of the raw collected data ensuring higher added value and positive impact for these organizations. However, the explosive quantity of data that is collected might contain personally identifiable information (PII) that should be protected to be compliant with related laws and regulations. For example, in the health sector, no doubt that the use of recent Information and Communication Technologies (Cloud computing, Internet of Things, Big Data, Artificial Intelligence, ...) improve communication and access to the right information on the right time and guarantee a high quality of care to patients. However, the collected, stored and processed data by these technologies often include sensitive information that arise new security and privacy concerns. Many approaches and solutions are used to mitigate such issues. In particular, concerning privacy it is widely agreed that anonymization techniques are considered among the most efficient approaches. In this thesis, we first provide a new detailed classification of the most used cryptographic and non-cryptographic anonymization techniques ensuring privacy. Besides, we evaluate the presented techniques through data completeness, confidentiality and data accuracy criteria. Next, we focus more on three relevant anonymization techniques belonging to Generalization-based approaches that are: K-anonymity, L-diversity and T-closeness techniques. Our second contribution in this thesis concerns a novel way in applying K-anonymity principle for quasi-identifier (QI) attributes. In fact, unlike other works, we have used the principle of K-anonymity without specifying a prior value of the threshold K. Afterwards, we proposed an algorithm that deals with sensitive attributes by using the principle of L-diversity. This algorithm ensures privacy while reducing the correlation loss among attributes. However, L-diversity technique cannot resist

against the Similarity attack. That is why; we developed two main algorithms that test the degree of proximity for both numerical and categorical attributes. Besides, we have measured the information loss through a utility measurement called Normalized certainty penalty (NCP) before and after applying the anonymization process on categorical attributes. In fact, the combination of these proposed algorithms ensures privacy, preserves data utility and treats both QI and sensitive attributes.

Keywords: Privacy, Anonymization, Ehealth, K-anonymity, L-diversity, T-closeness, Similarity attack, NCP.

Résumé : Récemment, la collecte de données est devenue cruciale pour la plupart des organisations en raison de la croissance rapide des outils d'analyse de données qui permettent une meilleure utilisation des données brutes collectées garantissant une plus grande valeur ajoutée et un impact positif pour ces organisations. Cependant, la quantité explosive de données collectées peut contenir des informations personnellement identifiables (PII) qui doivent être protégées pour être conformes aux lois et réglementations connexes. Par exemple, dans le secteur de la santé, il est clair que l'utilisation des récentes Technologies de l'Information et de la Communication (Cloud computing, Internet des Objets, Big Data, Intelligence Artificielle, ...) améliore la communication et l'accès aux bonnes informations au bon moment et garantie une meilleure qualité des soins aux patients. Cependant, les données collectées, stockées et traitées par ces technologies incluent souvent des informations sensibles (ou "sensitive") qui soulèvent de nouveaux défis en matière de sécurité et de protection de la vie privée (ou "privacy"). De nombreuses approches et solutions sont utilisées pour atténuer ces problèmes. En particulier, en ce qui

concerne la protection de la vie privée (ou "privacy"), il est largement admis que les techniques d'anonymisation sont considérées parmi les approches les plus efficaces. Dans cette thèse, nous proposons tout d'abord une nouvelle classification détaillée des techniques les plus utilisées d'anonymisation cryptographiques et non cryptographiques garantissant la protection de la vie privée (ou "privacy"). En outre, nous évaluons les techniques présentées à travers des critères d'exhaustivité, de confidentialité et d'exactitude des données. Ensuite, nous nous concentrons davantage sur trois techniques d'anonymisation pertinentes appartenant à des approches basées sur la généralisation qui sont : "K-anonymity", "L-diversity" et "T-closeness". Notre deuxième contribution dans cette thèse concerne une nouvelle manière d'appliquer le principe de "K-anonymity" pour les attributs "quasi-identifier" (QI). En fait, contrairement à d'autres travaux, nous avons utilisé le principe de "K-anonymity" sans spécifier une valeur préalable au seuil K. Ensuite, nous avons proposé un algorithme qui traite les attributs sensibles (ou "sensitive") en utilisant le principe de "L-diversity". Cet algorithme garantit la protection de la vie privée (ou "privacy") tout en réduisant la perte de corrélation entre les attributs. Cependant, la technique "L-diversity" ne peut pas résister contre l'attaque de Similarité. C'est pourquoi ; nous avons développé deux principaux algorithmes qui testent le degré de proximité pour les attributs numériques et catégoriels. De plus, nous avons mesuré la perte d'information par une mesure d'utilité appelée pénalité de certitude normalisée (NCP) avant et après l'application du processus d'anonymisation sur les attributs catégoriels. En fait, la combinaison de ces algorithmes proposés garantit la protection de la vie privée (ou "privacy"), préserve l'utilité des données et traite à la fois les attributs QI et sensible (ou "sensitive").

Mots-clés : Protection de la vie privée "Privacy", Anonymisation, Santé-mobile, "K-anonymity", "L-diversity", "T-closeness", Attaque de Similarité, NCP.