COVID Information Commons (CIC) Research Lightning Talk

Transcript of a Presentation by Murat Kantarcioglu (University of Texas at Dallas), October 16, 2020

Title: RAPID: Collaborative: A Privacy Risk Assessment Framework for Person-Level Data Sharing During Pandemics

Murat Kantarcioglu CIC Database Profile

NSF Award #: 2029661

YouTube Recording with Slides

October 2020 CIC Webinar Information

Transcript Editor: Julie Meunier

Transcript

Slide 1

Bonjour, je pense que c'est bon après-midi pour certains. Je m'appelle Murat Kantarcioglu. Je suis professeur d'informatique à l'université du Texas à Dallas. Il s'agit de notre travail RAPID sur l'évaluation des risques d'atteinte à la vie privée dans le cadre du partage de données en cas de pandémie, avec le professeur Brad Malin de la Vanderbilt Medical School.

Slide 2

Comme beaucoup d'entre vous le savent, à mesure que le COVID-19 progresse, de nouvelles questions se posent quant à savoir s'il est lié à la race, si le fait de fumer, par exemple, augmente le risque de contracter le COVID-19, si les résultats sont mauvais, etc. Par conséquent, pour répondre à ce type de questions, qu'il s'agisse de la relation entre la race et la mortalité due au COVID-19 ou d'autres facteurs tels que le tabagisme ou la pression artérielle, il est clair que nous devons collecter davantage de données. Il est clair que nous devons collecter davantage de données et, bien entendu, les analyser.

Slide 3

Si vous regardez cet article de Science publié en juillet, vous verrez que la Californie ne voulait pas partager - les représentants de l'État de Californie ne voulaient pas partager les détails des données COVID-19 avec certains chercheurs, ce qui a entravé certaines recherches.

Slide 4

Dans un sens, ils ont raison, car on sait que les données partagées peuvent être facilement réidentifiées. Il a été démontré il y a de nombreuses années que si vous combinez différents ensembles de données, par exemple, si vous avez des données COVID-19 qui indiquent le code postal, l'âge et le sexe, il y a un risque plus élevé de les relier à des listes d'enregistrement des eaux et d'identifier les personnes qui se

trouvent dans les données COVID-19 sur les patients. Bien entendu, ce problème est connu depuis longtemps, il n'est donc pas nouveau et la compréhension de ces risques a fait l'objet de recherches actives au cours des vingt dernières années. Il n'en reste pas moins que COVID-19 a changé beaucoup de choses.

Slide 5

L'une des choses est que nous disposons maintenant, en particulier avec les médias sociaux, de différentes sources de réidentification. Voici un exemple tiré du New York Times, qui a dressé le profil de nombreux patients décédés à cause de la COVID-19 dans leurs récits de vie.

Slide 6

Donc, si vous revenez à ce scénario d'attaque, vous pouvez maintenant disposer d'une source supplémentaire où vous savez que certaines personnes sont décédées et leurs informations et vous pouvez l'utiliser pour supprimer les cibles et réexécuter les attaques d'identification.

Slide 7

Par conséquent, ce RAPID tente vraiment de comprendre les défis uniques liés au partage des données sur la pandémie, en particulier les risques de réidentification. Nous nous rendons compte que, contrairement aux travaux antérieurs, les attributs qui doivent être partagés, tels que le tabagisme ou la race, peuvent changer avec le temps. Nous devons donc vraiment comprendre cela. De même, le nombre de cas évolue, ce qui signifie que certains risques pour la vie privée peuvent changer au fil du temps. Nous devons donc réellement mettre à jour les risques pour la vie privée au fur et à mesure que de nouvelles sources de données sont partagées avec d'autres. Ainsi, l'évolution rapide et incertaine du partage des données sur les pandémies a un impact sur les cadres d'évaluation des risques, et c'est pour cette raison que nous développons et mettons à jour les modèles existants.

Slide 8

Notre objectif est donc de déterminer quelle est la meilleure utilité que nous puissions obtenir compte tenu des objectifs de confidentialité des données, c'est-à-dire, dans le cas présent, de veiller à ce que le risque de réidentification soit faible, en d'autres termes, qu'il soit très difficile de réidentifier des personnes. Par ailleurs, quelle est la meilleure protection possible de la vie privée compte tenu des objectifs d'utilité des données ? Encore une fois, notre objectif est de fournir des informations aussi précises que possible. La deuxième option entre en jeu parce qu'il y a cet argument selon lequel, vous savez, nous devrions vraiment nous concentrer sur la lutte contre la maladie, de sorte que la confidentialité des données peut être secondaire. Cependant, nos recherches antérieures ont montré qu'en fonction de vos objectifs en matière de partage des données, différentes alternatives peuvent avoir des résultats différents en matière de protection de la vie privée. Il serait donc important de comprendre cela. Bien entendu, la principale orientation serait de déterminer l'utilité des données et la manière de la définir, ce qui fait l'objet de nos recherches en cours.

Slide 9

Pour conclure, nous avons construit nos propres travaux antérieurs sur l'analyse du risque d'identification et nous mettons à jour et développons des techniques basées sur ce défi unique de l'information variable dans le temps et changeante en raison de la pandémie. Bientôt, nous l'espérons,

dans un mois environ, nous disposerons d'un document sur les politiques générales de partage des données pour guider les responsables de la santé, comme les données qui peuvent être partagées avec un faible risque, les options à plus haut risque et leur impact. N'hésitez donc pas à nous contacter si vous avez des inquiétudes concernant le partage des données et la confidentialité des données relatives à la pandémie. Je vais donc m'arrêter là, et je pense que je vais dépasser un peu le temps imparti. Je vous remercie de votre attention.