Protéger la confidentialité de données statistiques

Anne-Sophie Charest,
Professeure adjointe, Département de mathématiques et statistique
Université Laval

Colloque 631 - Méthodes statistiques et statistiques publiques
Congrès de l'ACFAS, Québec, 6 mai 2013

PLAN

- Le problème
- Quelques solutions
- La confidentialié différentielle

Deux objectifs contradictoires

Promesse de confidentialité

Vos réponses seront conservées confidentielles et ne seront utilisées qu'à des fins statistiques.

Deux objectifs contradictoires

Promesse de confidentialité

Vos réponses seront conservées confidentielles et ne seront utilisées qu'à des fins statistiques.

Utilité des données

- Des rapports sont publiés à partir des données récoltées.
- Parfois, certaines données sont accessibles aux chercheurs ou à la population en général.

Deux objectifs contradictoires

Promesse de confidentialité	Utilité des données
Vos réponses seront conservées confidentielles et ne seront utilisées qu'à des fins statistiques.	- Des rapports sont publiés à partir des données récoltées. - Parfois, certaines données sont accessibles aux chercheurs ou à la population en général.

Comment peut-on s'assurer de respecter notre promesse de confidentialité tout en utilisant les données collectées?

Solution Intuitive:

Anonymisation des données

i.e. enlever toutes les variables qui pourraient permettre d'identifier directement le répondant (nom, NAS, adresse, …)

Solution Intuitive:

Anonymisation des données

i.e. enlever toutes les variables qui pourraient permettre d'identifier directement le répondant (nom, NAS, adresse, …)

Ce n'est malheureusement pas suffisant…

Example – Concours Netflix

Lesbian Sues Netflix Amid Privacy Concerns

1:31 PM - December 18, 2009 - By Jane McEntegart - Source : Tom's Guide US

A lesbian is suing Netflix amid concerns that the company did not do enough to ensure user data would remain anonymous once released and made available to the public.

Wired reports that the mother of two is suing the movie rental company alleging Netflix made it possible for her to be "outed" by disclosing insufficiently anonymous information about nearly half-a-million customers. The information was disclosed as part of the company's bid to find a more reliable recommendation system for customers.

Zoom

When Netflix released the 100 million movie ratings, along with the date of the rating the company assigned a unique ID number to the subscriber, and the movie information. However, according to Wired, two Texas University students quickly identified a number of Netflix subscribers by comparing their “anonymous” reviews in the data to ones posted on IMDb.

Conséquence:

3/12/2010 @ 12:35PM | 1,417 views

Netflix Settles Privacy Lawsuit, Cancels Prize Sequel

Taylor Buley, Contributor

+ Comment now

On Friday, Netflix announced on its corporate blog that it has settled a lawsuit related to its Netflix Prize, a $1 million contest that challenged machine learning experts to use Netflix's data to produce better recommendations than the movie giant could serve up themselves.

The lawsuit called attention to academic research that suggests that Netflix indirectly exposed the movie preferences of its users by publishing anonymized customer data. In the suit, plaintiff Paul Navarro and others sought an injunction preventing Netflix from going through the so-called "Netflix Prize II," a follow-up challenge that Netflix promised would offer up

Deuxième essai:

Ne publier que des données agrégées

Deuxième essai:

Ne publier que des données agrégées

Encore une fois, ce n'est pas suffisant.

Deuxième essai:

Ne publier que des données agrégées

Encore une fois, ce n'est pas suffisant.

Pensez par exemple à un tableau avec une catégorie complètement vide, ou à une cellule avec un seul individu.

Deuxième essai:

Ne publier que des données agrégées

Encore une fois, ce n'est pas suffisant.

Pensez par exemple à un tableau avec une catégorie complètement vide, ou à une cellule avec un seul individu.

Ou encore à des données beaucoup plus compliquées…

Example - Étude d'association pangénomique (GWAS)

RESEARCH ARTICLE

Resolving Individuals Contributing Trace Amounts of DNA to Highly Complex Mixtures Using High-Density SNP Genotyping Microarrays

Nils Homer[1,2], Szabolcs Szelinger[1], Margot Redman[1], David Duggan[1], Waibhav Tembe[1], Jill Muehling[1], John V. Pearson[1], Dietrich A. Stephan[1], Stanley F. Nelson[2], David W. Craig[1]*

1 Translational Genomics Research Institute (TGen), Phoenix, Arizona, United States of America, **2** University of California Los Angeles, Los Angeles, California, United States of America

Abstract

We use high-density single nucleotide polymorphism (SNP) genotyping microarrays to demonstrate the ability to accurately and robustly determine whether individuals are in a complex genomic DNA mixture. We first develop a theoretical framework for detecting an individual's presence within a mixture, then show, through simulations, the limits associated with our method, and finally demonstrate experimentally the identification of the presence of genomic DNA of specific individuals within a series of highly complex genomic mixtures, including mixtures where an individual contributes less than 0.1% of the total genomic DNA. These findings shift the perceived utility of SNPs for identifying individual trace contributors within a forensics mixture, and suggest future research efforts into assessing the viability of previously sub-optimal DNA sources due to sample contamination. These findings also suggest that composite statistics across cohorts, such as allele frequency or genotype counts, do not mask identity within genome-wide association studies. The implications of these findings are discussed.

Suite à la parution de cet article en 2008, le National Institute of Health (NIH) a enlevé toutes les données génomiques agrégées de son site internet.

Qu’est-ce qu’on fait alors?

- Limiter l'accès aux données.
  - Seulement aux chercheurs d’une institution reconnue avec un projet de recherche sérieux qui nécessite les données
  - Établir une sorte de contrat qui interdit d’essayer d’identifier les individus

Qu'est-ce qu'on fait alors?

- Limiter l'accès aux données.
  - Seulement aux chercheurs d'une institution reconnue avec un projet de recherche sérieux qui nécessite les données
  - Établir une sorte de contrat qui interdit d'essayer d'identifier les individus

- *Confidentialiser* les données agrégées (tableaux) ou sorties statistiques
  - Avant la publication
  - À l'aide d'un logiciel statistique qui donne accès aux données

Qu'est-ce qu'on fait alors?

- Limiter l'accès aux données.
  - Seulement aux chercheurs d'une institution reconnue avec un projet de recherche sérieux qui nécessite les données
  - Établir une sorte de contrat qui interdit d'essayer d'identifier les individus
- *Confidentialiser* les données agrégées (tableaux) ou sorties statistiques
  - Avant la publication
  - À l'aide d'un logiciel statistique qui donne accès aux données
- *Confidentialiser* les données avant de les partager.

Qu'est-ce qu'on fait alors?

- Limiter l'accès aux données.
  - Seulement aux chercheurs d'une institution reconnue avec un projet de recherche sérieux qui nécessite les données
  - Établir une sorte de contrat qui interdit d'essayer d'identifier les individus

- *Confidentialiser* les données agrégées (tableaux) ou sorties statistiques
  - Avant la publication
  - À l'aide d'un logiciel statistique qui donne accès aux données

- *Confidentialiser* les données avant de les partager.

Souvent, on fait un mélange de toutes ces méthodes.

Les statisticiens tentent de maximiser l'utilité tout en minimisant le risque de divulgation.

Confidentialiser?

- Méthodes de réduction (non-perturbatives)
  - Masquer la valeur de certaines cellules dans un tableau de fréquences
  - Enlever certaines variables pour certains ou tous les individus
  - Partager seulement un échantillon des données
  - Combiner certaines catégories pour une variable catégorique

Confidentialiser?

- Méthodes de réduction (non-perturbatives)
  - Masquer la valeur de certaines cellules dans un tableau de fréquences
  - Enlever certaines variables pour certains ou tous les individus
  - Partager seulement un échantillon des données
  - Combiner certaines catégories pour une variable catégorique

- Méthodes perturbatives
  - Arrondir à la hausse ou à la baisse les valeurs extrêmes de certaines variables
  - Échanger les valeurs de certaines variables entre des répondants
  - Ajouter du bruit aléatoire aux données
  - Arrondir les fréquences dans un tableau
  - Créer des jeux de données complètement synthétiques

Le travail du statisticien

- Développer des méthodes pour publication d'information obtenue à partir de données confidentielles

Le travail du statisticien

- Développer des méthodes pour publication d'information obtenue à partir de données confidentielles
- Mesurer le risque de bris de la promesse de confidentialité

Le travail du statisticien

- Développer des méthodes pour publication d'information obtenue à partir de données confidentielles
- Mesurer le risque de bris de la promesse de confidentialité
- Évaluer l'effet de la protection des données sur la qualité des analyses statistiques

Le travail du statisticien

- Développer des méthodes pour publication d'information obtenue à partir de données confidentielles
- Mesurer le risque de bris de la promesse de confidentialité
- Évaluer l'effet de la protection des données sur la qualité des analyses statistiques
- Développer des méthodes pour analyser les données publiées si elles ont été modifiées pour protéger la confidentialité

Qu’est-ce qu’on promet exactement?

- Très difficile de prédire quelle information pourrait causer du tort au répondant si elle était rendue publique.

Qu'est-ce qu'on promet exactement?

- Très difficile de prédire quelle information pourrait causer du tort au répondant si elle était rendue publique.

- Dalenius 1977: Soit D_K la valeur de la caractéristique D pour l'individu K.

  « Si la publication d'une statistique S permet de déterminer la valeur D_K plus précisément que sans accès à S, alors une divulgation a eu lieu »

Qu'est-ce qu'on promet exactement?

- Très difficile de prédire quelle information pourrait causer du tort au répondant si elle était rendue publique.

- Dalenius 1977: Soit D_K la valeur de la caractéristique D pour l'individu K.

  « Si la publication d'une statistique S permet de déterminer la valeur D_K plus précisément que sans accès à S, alors une divulgation a eu lieu »

- L'inférence elle-même serait ainsi une divulgation!

Un compromis intéressant: La confidentialité différentielle

- Promettre aux répondants qu'une tierce personne ne pourra rien apprendre de plus sur eux **qu'ils acceptent de participer à l'enquête que s'ils refusent.**

Un compromis intéressant: La confidentialité différentielle

- Promettre aux répondants qu'une tierce personne ne pourra rien apprendre de plus sur eux **qu'ils acceptent de participer à l'enquête que s'ils refusent.**

- Protection **rigoureusement mesurable** de la confidentialité des données.

Un compromis intéressant: La confidentialité différentielle

- Promettre aux répondants qu'une tierce personne ne pourra rien apprendre de plus sur eux **qu'ils acceptent de participer à l'enquête que s'ils refusent.**

- Protection **rigoureusement mesurable** de la confidentialité des données.

- Le mécanisme de protection est complètement public.

Un compromis intéressant: La confidentialité différentielle

- Promettre aux répondants qu'une tierce personne ne pourra rien apprendre de plus sur eux **qu'ils acceptent de participer à l'enquête que s'ils refusent.**
- Protection **rigoureusement mesurable** de la confidentialité des données.
- Le mécanisme de protection est complètement public.
- Protège la confidentialité même si l'adversaire a accès à des données auxiliaires sur les mêmes individus.

Définition (Dwork et al., 2006)

Une fonction randomisée κ garantie la confidentialité différentielle de niveau ϵ si et seulement si pour tous jeux de données voisins D_1 et D_2 et pour tout $S \in Image(\kappa)$,

$$e^{-\epsilon} \leq \frac{\Pr(\kappa(D_1) \in S)}{\Pr(\kappa(D_2) \in S)} \leq e^{\epsilon}$$

Définition (Dwork et al., 2006)

Une fonction randomisée κ garantie la confidentialité différentielle de niveau ϵ si et seulement si pour tous jeux de données voisins D_1 et D_2 et pour tout $S \in Image(\kappa)$,

$$e^{-\epsilon} \leq \frac{\Pr(\kappa(D_1) \in S)}{\Pr(\kappa(D_2) \in S)} \leq e^{\epsilon}$$

- Deux jeux de données sont voisins s'ils diffèrent par un répondant.
- Une plus petite valeur de implique une confidentialité accrue.
- Cette définition est valide pour tout type de fonction
  (e.g. moyenne, coefficients de régression, jeu de données synthétiques)
- Un individu peut prétendre que sa réponse avait n'importe quelle valeur.

Exemple – données discrètes

Probabilités de transition pour un méchanisme de publication du nombre de succès observés sur une variable binaire pour 5 individus

Valeur réelle

	$\tilde{Y}=0$	$\tilde{Y}=1$	$\tilde{Y}=2$	$\tilde{Y}=3$	$\tilde{Y}=4$	$\tilde{Y}=5$
Y=0	0.7160	0.1884	0.0667	0.0222	0.0060	0.0009
Y=1	0.3391	0.2999	0.1994	0.1074	0.0439	0.0107
Y=2	0.1396	0.2327	0.2507	0.2051	0.1253	0.0465
Y=3	0.0465	0.1253	0.2051	0.2507	0.2327	0.1397
Y=4	0.0107	0.0439	0.1074	.01995	0.2999	0.3391
Y=5	0.0009	0.0060	0.0222	0.0665	0.1884	0.7160

Valeur synthétique

Exemple – données discrètes

Probabilités de transition pour un méchanisme de publication du nombre de succès observés sur une variable binaire pour 5 individus

Valeur réelle

	$\tilde{Y}=0$	$\tilde{Y}=1$	$\tilde{Y}=2$	$\tilde{Y}=3$	$\tilde{Y}=4$	$\tilde{Y}=5$
Y=0	0.7160	0.1884	0.0667	0.0222	0.0060	0.0009
Y=1	0.3391	0.2999	0.1994	0.1074	0.0439	0.0107
Y=2	0.1396	0.2327	0.2507	0.2051	0.1253	0.0465
Y=3	0.0465	0.1253	0.2051	0.2507	0.2327	0.1397
Y=4	0.0107	0.0439	0.1074	.01995	0.2999	0.3391
Y=5	0.0009	0.0060	0.0222	0.0665	0.1884	0.7160

Valeur synthétique

Ici, $\epsilon \approx \log\left(\frac{0.0107}{0.0009}\right) = 2.39.$ (ratio d'environ 11)

Plusieurs sujets de recherche

- Création de jeux de données synthétiques
- Analyse de jeux de données synthétiques
- Sélection de modèle sous la contrainte de confidentialité différentielle
- Choix du niveau de confidentialité (ε)
- Relaxation de la définition de la confidentialité différentielle

Questions?

Commentaires?

anne-sophie.charest@mat.ulaval.ca

Pour plus d'informations…

Charest, A-S. How Can We Analyze Differentially-Private Synthetic Datasets?, *Journal of Privacy and Condentiality,* 2(2):21{33 (2011).

Charest, A-S. *Creation and Analysis of Differentially-Private Synthetic Datasets*, Thesis, Carnegie Mellon University, May 2012.

Dalenius, T. Towards a Methodology for Statistical Disclosure Control. *Statistik Tidskrift*, 15:429-444 (1977).

Dwork, C. Differential Privacy. *Automata, languages and programming*, 1-12 (2006).

Machanavajjhala, A., Kifer, D., Abowd, J., Gehrke, J., and Vilhuber, L. *Privacy: Theory Meets Practice on the Map*. In Data Engineering, 2008. ICDE 2008. IEEE 24th International Conference on, 277-286 (2008).