Vers une plus grande transparence du Web

Augustin Chaintreau, Guillaume Ducoffe, Roxana Geambasu, Mathias Lécuyer

Vers une plus grande transparence du Web †

A. Chaintreau[1] and G. Ducoffe[2,3] and R. Geambasu[1] and M. Lécuyer[1]

[1] *Columbia University, Computer Science Department, New York*
[2] *Univ. Nice Sophia Antipolis, CNRS, I3S, UMR 7271, 06900 Sophia Antipolis, France*
[3] *Inria, France*

De plus en plus les géants du Web (Amazon, Google et Twitter en tête) recourent à la manne des « Big data » : ils collectent une myriade de données qu'ils exploitent pour leurs algorithmes de recommandation personnalisée et leurs campagnes publicitaires. Pareilles méthodes peuvent considérablement améliorer les services rendus à leurs utilisateurs, mais leur opacité fait débat. En effet, il n'existe pas à ce jour d'outil suffisamment robuste qui puisse tracer sur le Web l'usage des données et des informations sur un utilisateur par des services en ligne.

Motivés par ce manque de transparence, nous avons développé un prototype du nom d'*XRay*, et qui peut prédire quelle donnée parmi toutes celles présentes dans un compte utilisateur est responsable de la réception d'une publicité. Dans cet article, nous présentons son principe ainsi que les résultats de nos premières expérimentations. Nous introduisons dans le même temps le tout premier modèle théorique pour le problème de la transparence du Web, et nous interprétons les performances d'Xray à la lumière de nos résultats obtenus dans ce modèle. En particulier, nous démontrons qu'un nombre $\theta(\log N)$ de comptes utilisateurs auxiliaires, remplis selon un procédé aléatoire, suffisent à déterminer quelle donnée parmi les N en présence a causé la réception d'une publicité. Nous aborderons brièvement les extensions possibles, et quelques problèmes ouverts.

Keywords: Sécurité de l'information (Privacy); Annonces publicitaires ciblées; Algorithmes d'apprentissage; Formes normales disjonctives monotones.

1 Introduction

Ces dernières décennies ont été marquées par l'essor des « Big data ». Quantités d'informations personnelles (allant de la localisation d'un utilisateur à l'historique de ses recherches, de ses courriels ou encore de ses photographies) sont à présent collectées et exploitées en permanence par les grands noms du Web. À terme, ces nouvelles pratiques offriront quantités d'avantages au quotidien. Sur le plan commercial, d'une part, puisqu'une analyse plus fine de la clientèle cible aide une entreprise à mieux positionner ses produits. Mais aussi, d'autre part, sur la qualité de service et de vie des utilisateurs. En effet, la collecte d'informations personnelles offre la possibilités pour des services en ligne de personnaliser leur offre, de prédire, et donc d'anticiper, les attentes de leurs utilisateurs. En termes d'utilité publique, les « Big data » sont d'ores et déjà un recours incontournable dont on se sert pour résoudre des problèmes de santé publique [SK13], de pollution [MPR09], voire d'endiguement de la criminalité [WGB12].

Cependant, l'engouement des entreprises pour nos données personnelles pose aussi la question de la *manière* dont elles les acquièrent, ainsi que de *l'usage* qu'elles en font. À l'heure actuelle, la collecte des « Big data » se fait encore dans la plus grande opacité. Ni les utilisateurs, ni les organisations étatiques ou de protection des utilisateurs, ne sont en mesure d'en connaître les détails. Sans surprise, il s'ensuit des dérives, observables, dont une utilisation de ces données pour de la discrimination à l'embauche [AF13], une différenciation des prix en fonction de l'utilisateur [HSL+14], etc... Dans cet article, nous nous intéressons à améliorer la transparence de ces pratiques.

† Ce travail est partiellement soutenu par les contrats DARPA FA8650-11-C-7190, NSF CNS-1351089 et CNS-1254035, Google, Microsoft. Les preuves omises peuvent être trouvées dans [LDL+14].

Motivations et objectifs. Plus précisément, nous posons les bases d'une algorithmique pour le problème de la transparence du Web. Étant données les publicités et autres recommandations personnalisées qu'un utilisateur reçoit, le problème est de calculer quelles sont parmi ses données personnelles celles qui en sont responsables. De cette façon, un utilisateur peut suivre la manière dont ses données sont échangées et exploitées par les différents services. Nos bases algorithmiques servent à concevoir de nouveaux outils pour résoudre ce problème. Néanmoins pour être utilisables en pratique, ces outils doivent satisfaire plusieurs contraintes.

- D'abord, ils doivent pouvoir *passer à l'échelle*, s'entend s'exécuter en temps et en espace polynomial en le nombre de données personnelles à traiter.
- Par ailleurs, il faut que, sous certaines hypothèses réalistes, leurs réponses (c'est-à-dire, la donnée expliquant pourquoi une publicité a été observée) soient *correctes* avec forte probabilité.
- Enfin on demande à ce que les outils développés puissent être appliqués à l'audit d'un grand nombre de services différents (par exemple, Gmail, Youtube, Amazon et Facebook) sans changement conceptuel ni de code notables.

État de l'art. La plupart des outils existants à ce jour ne satisfont pas l'ensemble de ces contraintes. En général, ils sont spécifiques à l'audit d'un unique service, ou bien d'un seul usage par les services des données personnelles de leurs utilisateurs [MGEL12, XMD+14]. Nous souhaitons, au contraire, pouvoir repérer *n'importe quel usage* de ces données, et ce pour le plus grand nombre possible de services en ligne. De plus, la plupart des outils proposés retournent des réponses qui sont *invérifiables*. À l'exception notable de [DTD14], il n'existe pas de théorie sous-jacente pour le problème de la transparence du Web sur laquelle on pourrait s'appuyer pour valider les résultats obtenus par des expérimentations. Le modèle décrit dans [DTD14] est en fait assez proche de notre propre modèle théorique (que nous introduirons dans la suite de ce papier), toutefois il repose sur des hypothèses qui nous paraissent trop faibles pour obtenir des algorithmes à la fois *certifiés corrects* (avec forte probabilité) et *efficaces* en pratique.

Nos contributions. Dans cet article nous présentons un nouvel outil, du nom d'*Xray*, pour le problème de la transparence du Web. Informellement, les résultats retournés par Xray sont obtenus en comparant les publicités reçues par un compte utilisateur à celles reçues par un petit nombre de comptes utilisateurs auxiliaires, remplis d'une partie des données personnelles de l'utilisateur selon un procédé aléatoire. Contrairement à ses concurrents directs, nous montrons que Xray satisfait à toutes les contraintes que nous avons exposées ci-dessus. Le code de l'outil est disponible en accès *open source* [xRaa]. Nous présentons également les résultats obtenus lors de nos expérimentations avec le tout premier prototype d'Xray. En particulier, nous observons qu'il passe remarquablement bien à l'échelle. En effet, il ne nécessite qu'un nombre de comptes auxiliaires *logarithmique* en le nombre de données pour corréler chaque publicité reçue à la donnée personnelle qu'elle cible. Motivés par les bons résultats obtenus en pratique, nous introduisons finalement un nouveau modèle théorique pour la transparence du Web (pour ainsi dire, le premier modèle pour ce problème[‡]), et nous nous en servons pour valider théoriquement les performances de notre prototype.

2 Présentation d'Xray

Hypothèses. Xray repose sur une vision simplifiée des services Web. Chaque service est modélisé par une *boîte noire*. Le service prend en entrée des données sur un utilisateur, dont nous ne différencions pas la nature (courriels, images, recherches et autres sont traitées de façon identique). Il retourne une sortie (publicité, vidéo, recommandation, etc ...), qu'on supposera observable. Xray part de l'hypothèse que parmi les sorties observées, certaines l'ont été en raison de la présence d'une donnée particulière. Il prend en entrée les données et les sorties des services Web, et il associe à chaque sortie la donnée la plus susceptible de lui être corrélée.

[‡] Le modèle théorique de [DTD14] a été introduit indépendamment du nôtre.

Architecture. La corrélation entre les sorties et les données est calculée selon une méthode d'inférence Bayésienne. Nous renvoyons à notre article [LDL$^+$14] pour plus de détails. Intuitivement, si la sortie O_j est observée à cause de la présence de la donnée $\mathcal{D}_i$ dans le compte utilisateur, alors la plupart des comptes utilisateurs contenant $\mathcal{D}_i$ devraient également observer O_j, et presque aucun de ceux qui ne contiennent pas $\mathcal{D}_i$ devraient observer O_j. L'algorithme au coeur d'Xray formalise cette intuition en un test probabiliste, qu'il applique à tous les couples $(\mathcal{D}_i, O_j)$. Finalement, il associe à chaque sortie O_j la donnée $\mathcal{D}_i$ telle que le score obtenu au test par le couple $(\mathcal{D}_i, O_j)$ est maximisé.
Afin de pouvoir appliquer le test probabiliste au coeur d'Xray, nous créons des comptes utilisateurs auxiliaires qui sont chacun remplis par le prototype avec une fraction des données de l'utilisateur. Chaque donnée est placée indépendamment des autres, et avec la même probabilité, dans chacun des comptes. Les sorties observées par les comptes auxiliaires sont ensuite comparées aux sorties observées pour le compte utilisateur principal.

Expérimentations. Nous avons évalué les résulats obtenus par Xray sur des expériences conduites sur Gmail, Amazon et Youtube. En particulier, nous avons mesuré la *fiabilité* des réponses retournées par le prototype, et l'évolution du nombre de comptes nécessaires à créer afin de maintenir un niveau de fiabilité acceptable. Pour les expériences menées sur Amazon et Youtube, nous avons mesuré la fiabilité des réponses obtenues en les comparant à la « vérité-terrain » (*ground-truth*) fournie par ces services. En revanche, pour Gmail, nous avons dû nous-mêmes établir (à la main) la vérité-terrain. Afin de mesurer l'erreur qui en a résulté, nous avons créé des campagnes publicitaires, via Google AdWords, sur des thèmes ultra-spécifiques (tels que la poésie chaldaïque). Leur grande spécifité nous a permis de mieux contrôler quels étaient les mots-clés responsables de leur observation.

Thème	**Publicités ciblées**	**Xray Scores**	**Accounts Displays**
Alzheimer	Black Mold Allergy Symptoms ? Expert to remove Black Mold.	0.99, 0.05	9/9, 61/198
	Adult Assisted Living. Affordable Assisted Living.	0.99, 0.99	8/8, 12/14
Dépression	Shamanic healing over the phone.	0.99, 0.99	16/16, 117/117
	Text Coach - Get the girl you want and Desire.	0.93, 0.04	7/7, 31/276
Dettes	Take a New Toyota Test Drive, Get a $50 Gift Card On The Spot.	0.99, 0.9	7/7, 58/65
	Great Credit Cards Search. Apply for VISA, MasterCard...	0.99, 0.0	9/9, 151/2358
	Stop Creditor Harassment, End the Harassing Calls.	0.99, 0.96	8/8, 256/373

FIG. 1: **Exemples de publicités ciblées.**

Nous avons observé qu'en moyenne, Xray atteint un niveau de précision de l'ordre de $85-90\%$ en n'utilisant qu'un nombre de comptes auxiliaires *logarithmique* en le nombre de données dans le compte utilisateur principal. Pour l'anecdote, nous présentons quelques-unes des corrélations calculées par notre algorithme, en Figure 1, parmi les plus troublantes. En effet, bon nombre de publicités ciblent des mots-clés ultra-sensibles tels que « maladie », « dette » ou encore « dépression ». De plus amples détails, sur Xray et sur nos expérimentations, sont disponibles sur le site du projet [xRab].

3 Un modèle théorique pour la Transparence du Web

Confortés par nos bons résultats expérimentaux, nous souhaiterions pouvoir *certifier* correctes les réponses retournées par Xray (avec forte probabilité), sous des hypothèses simples et réalistes. Pour ce faire, nous introduisons un modèle théorique pour le problème de la transparence du Web. Ce modèle n'est pas spécifique à Xray, il est utilisable dans l'étude de ses concurrents directs.

Formalisation du problème. Nous considérons les N données dans un compte utilisateur comme l'ensemble canonique $[N] = \{1, \ldots, N\}$ des N premiers entiers. À chaque sortie observée O_j, nous associons une fonction $f_j : \mathcal{P}([N]) \to \{0,1\}$. Intuitivement, l'argument de la fonction représente le sous-ensemble

des données dans un compte auxiliaire, et sa valeur (booléenne) indique si la sortie O_j a été observée par ce compte. Nous supposons la fonction f_j *monotone* : pour tous sous-ensembles $\mathcal{C} \subseteq \mathcal{C}' \subseteq [N]$, on a que $f_j(\mathcal{C}) \leq f_j(\mathcal{C}')$. Afin d'éviter les cas triviaux, nous supposons aussi que la fonction f_j vérifie que $f_j(\mathcal{C}) \neq f_j(\mathcal{C}')$ pour (au moins) un couple de sous-ensembles $\mathcal{C}, \mathcal{C}' \subseteq [N]$.
En pratique, la fonction f_j n'est pas connue. Mais nous pouvons l'approcher grâce aux observations expérimentales que nous collectons pour chaque compte auxiliaire. Formellement, nous supposons l'existence d'un oracle qui pour tout sous-ensemble $\mathcal{C}$ retourne la valeur $f(\mathcal{C})$, avec une certaine probabilité de se tromper. Soient $p_{\text{in}}, p_{\text{out}}$ les probabilités respectives pour l'oracle de retourner 1 quand $f(\mathcal{C}) = 1$ et $f(\mathcal{C}) = 0$. Nous faisons l'hypothèse additionnelle que $p_{\text{in}} > p_{\text{out}}$.

Théorème 1 *S'il existe $\mathcal{D}_i$ telle que $\bigcap_{\mathcal{C}|f_j(\mathcal{C})=1} = \{\mathcal{D}_i\}$, alors la donnée $\mathcal{D}_i$ peut être détectée et retournée avec forte probabilité, en temps linéaire $O(N)$ et en $O(\log N)$ requêtes à l'oracle.*

La preuve du Théorème 1 est constructive. Nous en avons déduit un nouvel algorithme pour le problème de la transparence du Web (`Set Intersection Algorithm`). Empiriquement, l'algorithme Bayésien décrit dans la section précédente arrive souvent au même résultat d'inférence et certains cas simples permettent de le prouver théoriquement. Le Théorème 1 nous permet donc de certifier que les résultats de notre prototype sont corrects (avec forte probabilité).

4 Conclusion

Nous avons développé un nouvel outil pour le problème de la transparence du Web, et qui permet aux utilisateurs une analyse fine de l'exploitation de leurs données personnelles par les services en ligne. Cette analyse repose sur une toute nouvelle théorie de la transparence du Web, que nous introduisons dans ce papier. Nous l'utilisons pour prouver que, sous certaines hypothèses simples et réalistes, les corrélations détectées par Xray sont correctes avec forte probabilité. Notre travail en cours étend nos résultats à des campagnes publicitaire plus complexes, où l'observation d'une sortie peut dépendre de la présence de *plusieurs* données dans un compte utilisateur, voire d'une combinaison de ces données parmi plusieurs combinaisons possibles. À terme, nous souhaiterions remplacer la phase, très coûteuse, de création de comptes auxiliaires par un algorithme de collaboration distribué entre plusieurs comptes utilisateurs.

Références

[AF13] Alessandro Acquisti and Christina M Fong. An experiment in hiring discrimination via online social networks. *Available at SSRN 2031979*, 2013.

[DTD14] Amit Datta, Michael Carl Tschantz, and Anupam Datta. Automated Experiments on Ad Privacy Settings : A Tale of Opacity, Choice, and Discrimination. *arXiv.org*, August 2014.

[HSL+14] Aniko Hannak, Gary Soeller, David Lazer, Alan Mislove, and Christo Wilson. Measuring Price Discrimination and Steering on E-commerce Web Sites. In *IMC '14 : Proceedings of the 2014 Conference on Internet Measurement Conference*. ACM Request Permissions, November 2014.

[LDL+14] Mathias Lecuyer, Guillaume Ducoffe, Francis Lan, Andrei Papancea, Theofilos Petsios, Riley Spahn, Augustin Chaintreau, and Roxana Geambasu. XRay : Enhancing the Web's Transparency with Differential Correlation . In *23rd USENIX Security Symposium (USENIX Security 14)*, San Diego, CA, 2014. USENIX Association.

[MGEL12] Jakub Mikians, László Gyarmati, Vijay Erramilli, and Nikolaos Laoutaris. Detecting price and search discrimination on the internet. In *Proceedings of the 11th ACM Workshop on Hot Topics in Networks*, pages 79–84, 2012.

[MPR09] Deepak Merugu, Balaji S Prabhakar, and N S Rama. An incentive mechanism for decongesting the roads : A pilot program in Bangalore. 2009.

[SK13] Adam Sadilek and Henry Kautz. Modeling the impact of lifestyle on health at scale. In *Proceedings of the Sixth ACM International Conference on Web Search and Data Mining*, 2013.

[WGB12] Xiaofeng Wang, MatthewS. Gerber, and DonaldE. Brown. Automatic crime prediction using events extracted from twitter posts. In ShanchiehJay Yang, ArielM. Greenberg, and Mica Endsley, editors, *Social Computing, Behavioral - Cultural Modeling and Prediction*, volume 7227 of *Lecture Notes in Computer Science*, pages 231–238. 2012.

[XMD+14] Xinyu Xing, Wei Meng, Dan Doozan, Nick Feamster, Wenke Lee, and Alex C Snoeren. Exposing Inconsistent Web Search Results with Bobble. *Passive and Active Measurements Conference*, 2014.

[xRaa] `https://github.com/matlecu/xray/`.

[xRab] `http://xray.cs.columbia.edu/`.