

Fouille de données et protection de la vie privée

Sébastien Gambs

Chercheur post-doctoral au LAAS-CNRS

Groupe : Tolérance aux fautes et Sûreté de
Fonctionnement informatique (TSF)

Superviseur : Yves Deswarte

28 novembre 2008

Introduction

Fouille de données et protection de la vie privée

Fouille de données et **protection de la vie privée** semblent *a priori* avoir des buts orthogonaux :

- ▶ la fouille de données s'intéresse à la *découverte de connaissances cachées dans les données* alors que
- ▶ la protection de la vie privée veut *préserver la confidentialité des données*.



Interrogation principale : comment extraire des connaissances utiles tout en préservant la confidentialité des données sensibles?

⇒ Le domaine de la **fouille de données préservant la confidentialité** (*privacy-preserving data mining* en anglais) essaye de répondre à cette question.

Scénario 1 : anonymisation de données

- ▶ Un important fournisseur d'accès veut rendre public les données d'accès de certains de ses clients afin d'offrir un jeu de données public à la communauté de fouille de données du Web.



Exemple de données révélées : requêtes personnelles des derniers mois.

- ▶ **Interrogation** : comment est ce que la compagnie peut anonymiser les données de telle manière à garantir à ses clients qu'aucune information sensible ne pourra être extraite à leurs propos?

Exemple de bris de vie privée

A Face Is Exposed for AOL Searcher No. 4417749

The New York Times

August 8, 2006

What Revealing Search Data Reveals

AOL posted, but later removed, a list of the Web search inquiries of 658,000 unnamed users on a new Web site for academic researchers. An interview with one of those unnamed users, Thelma Arnold, combined with her data reveal what she was searching for, why and on which Web sites.

A sample of Thelma Arnold's search data released by AOL

4417749	swing sets	2006-04-24	15:39:30	4	http://www.byoswingset.com
4417749	swing sets	2006-04-24	15:39:30	9	http://www.buychoice.com
4417749	swing sets	2006-04-24	15:39:30	10	http://www.creativeplaythings.com
4417749	swing sets	2006-04-24	15:39:30	5	http://www.childlife.com
4417749	swing sets	2006-04-24	15:39:30	6	http://www.planitplay.com
4417749	that do not shed	2006-04-28	9:55:54	2	http://www.gopetsamerica.com
4417749	dog who urinate on everything	2006-04-28	13:24:07	6	http://www.dogpaysia.com
4417749	walmart	2006-04-28	14:07:32	1	http://www.walmart.com
4417749	womens underwear	2006-04-28	14:12:28	10	http://www.bizrate.com
4417749	jcpenny	2006-04-28	14:16:05		
4417749	jcpenny	2006-04-28	14:16:49	1	http://www.jcpenny.com
4417749	tortois and turtles	2006-04-29	13:12:47		
4417749	manchester terrier	2006-05-02	9:05:31	1	http://www.manchesterterrier.com
4417749	delta	2006-05-02	11:49:26		
4417749	fingers going numb	2006-05-02	17:35:47		
4417749	dances by laura	2006-05-02	17:59:32		
4417749	dances by lori	2006-05-02	17:59:57		
4417749	single dances	2006-05-02	18:00:18	1	http://solosingles.com
4417749	single dances in atlanta	2006-05-02	18:01:13		
4417749	single dances in atlanta	2006-05-02	18:01:50		
4417749	dry mouth	2006-05-06	16:49:14	2	http://www.mayoclinic.com
4417749	dry mouth	2006-05-06	16:49:14	8	http://www.wrongdiagnosis.com
4417749	thyroid	2006-05-06	16:53:34		
4417749	thyroid	2006-05-06	16:55:44		
4417749	competitive market analysis of homes in lilburn	2006-05-14	12:14:52		
4417749	competitive market analysis of homes in lilburn	2006-05-14	12:16:17		
4417749	competitive market analysis of homes in lilburn	2006-05-14	12:16:43		

AOL posted, but later removed, a list of the Web search inquiries of 658,000 unnamed users on a new Web site for academic researchers. An interview with one of those unnamed users, Thelma Arnold, combined with her data reveal what she was searching for, why and on which Web sites.

Why the search

"I was thinking about my grandchildren"

"I was looking for some."

"A woman was in the [public] bathroom crying. She was going through a divorce. I thought there was a place called 'Dances by Lori,' for singles."

"I wanted to find out what my house was worth."



Erik S. Lesser for The New York Times

Thelma Arnold's identity was betrayed by AOL records of her Web searches, like ones for her dog, Dudley, who clearly has a problem.

(Extrait d'un article du New York Times paru le 6 août 2006)

Scénario 2 : calcul de statistiques jointes

- ▶ Différentes agences gouvernementales (par exemple l'agence du revenu, l'office de la santé publique et le ministère de la justice) souhaitent calculer et rendre public des statistiques portant sur l'ensemble de la population.
- ▶ Les contraintes juridiques interdisent de communiquer des informations sur un individu précis, même à une autre agence gouvernementale.
- ▶ **Interrogation** : comment les agences peuvent-elles calculer ces statistiques de manière relativement précise tout en protégeant la vie privée des citoyens?

Scénario 3 : apprentissage distribué

- ▶ Soit deux compagnies de bioinformatique : Alice Corporation et Bob Trust.



- ▶ Chaque compagnie possède une base de données gigantesque constituée de mesures collectées à partir d'expériences effectuées dans leurs laboratoires.
- ▶ Les deux sont prêtes à coopérer pour réaliser une tâche d'apprentissage d'intérêt commun mais ...
- ▶ aucune ne souhaite communiquer sa base de données en clair.
- ▶ **Interrogation** : comment peuvent-ils atteindre ce but sans divulguer aucune information non nécessaire?

Fouille de données préservant la confidentialité

Approches basées sur la perturbation des données

Méthodes de randomisation

Calcul multiparti sécuritaire

Conclusion et perspectives futures

Fouille de données préservant la confidentialité

Fouille de données préservant la confidentialité

La notion de **protection de la vie privée** est

- ▶ *difficile à formaliser et quantifier*
- ▶ *et dépendante du contexte.*

La **fouille de données préservant la confidentialité** étudie :

- ▶ *comment les algorithmes de fouille de données affecte la protection de la vie privée et,*
- ▶ *essaye de trouver et d'analyser des algorithmes qui protège la confidentialité des données.*

Origine du domaine

Deux articles adoptant des approches très différentes ont consacré le terme en 2000 :

Privacy-Preserving Data Mining

Rakesh Agrawal



Ramakrishnan Srikant



Privacy Preserving Data Mining

Yehuda Lindell



Benny Pinkas



(Faites attention à la subtile différence entre les deux titres :-))

Classification des algorithmes préservant la confidentialité

Principales dimensions utilisées pour classer les algorithmes :

1. Distribution des données :

- ▶ Dans les mains d'une seule entité.
- ▶ Les attributs d'un enregistrement particulier sont partagés entre différents sites (*partitionnement vertical*).
- ▶ Plusieurs bases de données sont situés à différents endroits (*partitionnement horizontal*).

2. Algorithme de fouille de données utilisé.

3. Technique de protection de la vie privée utilisée :

- ▶ Approches basées sur la perturbation des données.
- ▶ Méthodes de randomisation.
- ▶ Calcul multiparti sécuritaire.

Approches basées sur la perturbation des données

Approches basées sur la perturbation des données

Idee principale : *modifier les valeurs des attributs sensibles afin de préserver la confidentialité des données.*

Exemples de modifications :

- ▶ *Altérer la valeur d'un attribut* en le perturbant (ABMIV¹ 99) ou en le remplaçant par "?" (valeur inconnue) (Chang et Moskowitz 00).
- ▶ *Échanger les valeurs d'un attribut* entre deux enregistrements différents (Fienberg et McIntyre 04).
- ▶ *Utiliser une granularité plus grossière* en fusionnant plusieurs valeurs possibles d'un attribut en une seule (Chang et Moskowitz 00).

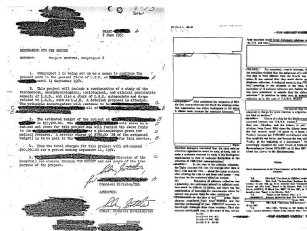
¹Attalah, Bertino, Elmagarmid, Ibrahim et Verykios

Sanitisation

Sanitisation : processus qui accroît l'incertitude dans les données afin de protéger la vie privée.

⇒ Compromis inhérent entre le niveau de protection de la vie privée et l'utilité de la base de données "sanitisée".

Exemple typique d'utilisation : rendre public des données.



Exemples extraits de l'entrée "sanitization" sous Wikipedia (octobre 08)

Idee : protéger son anonymat en se fondant dans la foule



Photo de groupe de Charles Bennett pour la conférence QIP 2007

k-anonymité

- ▶ Dans de nombreuses situations, trouver la manière optimale de “sanitiser” les données est un problème NP-ardu.
- ▶ **Exemple de méthode de sanitisation avec garanties :**
k-anonymité (Sweeney 02).
- ▶ **Idée principale :** protéger la vie privée d'un individu en le faisant *se fondre dans la “foule”*.
- ▶ **k-anonymisation :** processus qui construit une base de données (par suppression et généralisation) dans laquelle chaque enregistrement est indistinguishable d'au moins $k - 1$ autres enregistrements.
- ▶ **Garantie :** aucun individu ne peut être ciblé avec probabilité supérieure à $\frac{1}{k-1}$, même pour un adversaire disposant d'information auxiliaire.

Illustration du processus de *k*-anonymisation

	Non-Sensitive			Sensitive
	Zip Code	Age	Nationality	Condition
1	13053	28	Russian	Heart Disease
2	13068	29	American	Heart Disease
3	13068	21	Japanese	Viral Infection
4	13053	23	American	Viral Infection
5	14853	50	Indian	Cancer
6	14853	55	Russian	Heart Disease
7	14850	47	American	Viral Infection
8	14850	49	American	Viral Infection
9	13053	31	American	Cancer
10	13053	37	Indian	Cancer
11	13068	36	Japanese	Cancer
12	13068	35	American	Cancer

Figure 1. Inpatient Microdata

	Non-Sensitive			Sensitive
	Zip Code	Age	Nationality	Condition
1	130**	< 30	*	Heart Disease
2	130**	< 30	*	Heart Disease
3	130**	< 30	*	Viral Infection
4	130**	< 30	*	Viral Infection
5	1485*	≥ 40	*	Cancer
6	1485*	≥ 40	*	Heart Disease
7	1485*	≥ 40	*	Viral Infection
8	1485*	≥ 40	*	Viral Infection
9	130**	3*	*	Cancer
10	130**	3*	*	Cancer
11	130**	3*	*	Cancer
12	130**	3*	*	Cancer

Figure 2. 4-anonymous Inpatient Microdata

Exemple extrait de l'article “*l*-diversity: privacy beyond *k*-anonymity”
 (Machanavajjhala, Gehrke, Kifer et Venkatasubramanian 07)

Limites de la k -anonymité

Vulnérabilité de la k -anonymité à certaines attaques telles que :

- ▶ **Attaque basée sur l'homogénéité** : si tous les individus d'un groupe partagent la même valeur pour un attribut sensible
⇒ la valeur de l'attribut d'un individu n'est plus protégé si on peut identifier son groupe.
- ▶ **Attaque basée sur des connaissances *a priori*** : l'adversaire peut avoir des connaissances *a priori* lui permettant d'attaquer la confidentialité.

Exemples de connaissances potentielles :

- ▶ présence d'un individu parmi les données anonymisées,
- ▶ connaissance partielle de ses attributs (sensibles ou non),
- ▶ connaissance de la distribution des attributs (sensibles et non-sensibles) parmi la population.

Autres métriques de protection de la vie privée

- ▶ l -diversité (MKG² 07) : maintenir de la diversité dans chaque groupe au niveau des valeurs possibles des attributs sensibles.
- ▶ Peut-être instanciée par une mesure basée sur l'entropie.
- ▶ Prémunit contre les attaques basées sur l'homogénéité et certaines autres attaques.
- ▶ t -proximité (t -closeness en anglais) (LLV³ 07) : la distribution des attributs dans chaque groupe doit être proche de celle de la population globale.
- ▶ t est un seuil à ne pas dépasser pour la proximité entre les distributions.
- ▶ **Remarque** : ces deux méthodes augmentent la protection de la vie privée mais sacrifient potentiellement de l'utilité.

²Machanavajjhala, Gehrke, Kifer et Venkatasubramanian

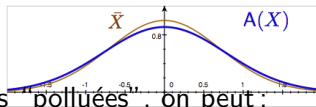
³Li, Li et Venkatasubramanian

Méthodes de randomisation

Méthodes de randomisation

Randomisation : ajout de bruit indépendant (comme gaussien ou uniforme) aux valeurs des attributs.

But : cacher les valeurs spécifiques des attributs tout en préservant la distribution jointe des données.



À partir des données polluées, on peut :

- ▶ Reconstruire la distribution originale des données par un algorithme du type Expectation-Maximization (Agrawal et Aggarwal 01).
- ▶ Apprendre directement sur les données bruitées (Agrawal et Srikant 00).

Remarque : proche en esprit de l'approche par perturbation.

Méthodes de randomisation (suite)

Contexte d'utilisation : particulièrement adaptées aux cas où :

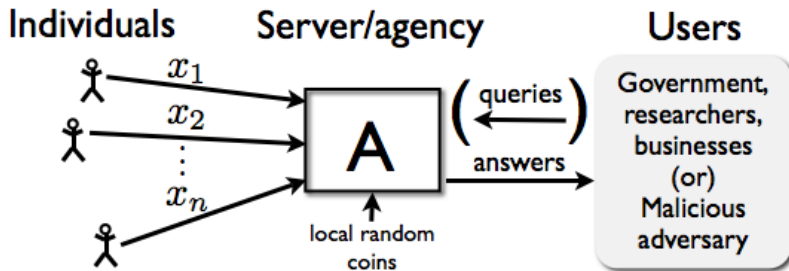
- ▶ les données sont distribuées entre plusieurs participants,
- ▶ ceux-ci sont prêts à envoyer une version randomisée de leurs données à un tiers parti en qui on a une confiance limitée (*semi-trusted party* en anglais).

⇒ Le tiers parti réalise l'algorithme d'apprentissage et publie ensuite les résultats.

Compromis inhérent entre la préservation de la confidentialité et la précision du modèle appris paramétrable par :

- ▶ *l'intensité* et
- ▶ *le type de bruit*.

Modèle possible pour les méthodes de randomisation



Extrait d'un tutoriel de Adam Smith sur la protection de la vie privée dans les bases de données (mars 2008)

Entropie conditionnelle

Entropie conditionnelle (Agrawal et Aggarwal 01) :

- ▶ Mesure provenant de théorie de l'information.
- ▶ Représente l'information mutuelle partagée entre l'ensemble de données originel et la version randomisée.
- ▶ **Intuitivement** : combien d'information la version randomisée révèle sur l'ensemble originel.
- ▶ Information mutuelle basse
 - ⇒ haut niveau global de préservation de la vie privée
 - ⇒ bas niveau de précision pour l'apprentissage

Brèche de confidentialité

Brèche de confidentialité (Evfimievski 02) : changement important de confiance concernant la valeur possible d'un attribut d'un enregistrement particulier.

Exemple :

- ▶ Alice sait que son voisin Bob se trouve dans un ensemble de données particulier et qu'elle présuppose que Bob a un salaire modeste.
- ▶ Après avoir vu la version randomisée des données, elle pense avoir identifier l'enregistrement de Bob avec une probabilité non-négligeable.
- ▶ Elle apprend que celui-ci est en fait dans une tranche de salaire élevé.
- ▶ ⇒ **Brèche de confidentialité!!!**

Brèche de confidentialité (suite)

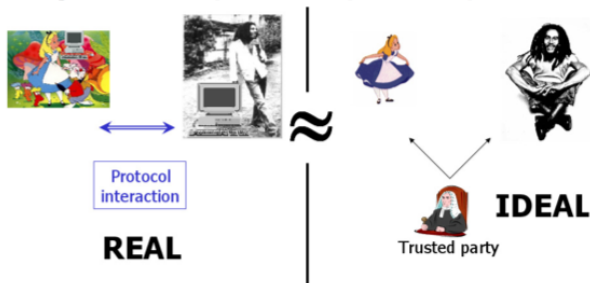
- ▶ **Difficulté**: modéliser les connaissances *a priori* de l'adversaire.
- ▶ **Solution possible**: une méthode limitant les brèches de confidentialité ne présupposant aucune connaissance de la distribution des données (Evfimievski, Gerhke et Srikant 03).
- ▶ Information mutuelle basse \nRightarrow risque faible de brèche de confidentialité
- ▶ L'information mutuelle est une *mesure globale* de protection de la confidentialité.

Calcul multiparti sécuritaire

Calcul multiparti sécuritaire

Calcul multiparti sécuritaire: branche de la cryptographie qui s'occupe de la réalisation sécuritaire de tâches distribuées.

Tâche typique: calculer une certaine fonction $f(x, y)$, où x est l'entrée du participant A et y l'entrée du participant B.



Paradigme cryptographique: un protocole est sécuritaire si les participants n'apprennent rien de plus que la sortie de la fonction f .

Modèles de sécurité

Honnête mais curieux (parfois appelé *passif* ou *semi-honnête*) :

- ▶ les participants suivent les directives du protocole mais ...
- ▶ enregistrent toutes les communications échangées pour en extraire le maximum d'information.
- ▶ Modélise bien les situations où les participants sont prêts à coopérer pour atteindre un but commun mais ne souhaitent pas communiquer directement leurs ensembles de données.
- ▶ Presque toujours considéré en fouille de données préservant la confidentialité.

Autres modèles :

- ▶ Participants **malicieux** pouvant tricher durant l'exécution du protocole.
- ▶ Adversaire ayant accès à un ordinateur quantique.

Théorèmes généraux

Résultat général (CCD⁴ 88, BGW⁵ 88): n'importe quelle fonction f peut être implémentée de manière inconditionnellement sécuritaire (dans le sens de la théorie de l'information) pourvu qu'au moins une certaine proportion des participants soient honnêtes.

Commentaire: bien qu'universel cette méthode générique peut être inefficace quand :

- ▶ la fonction f est complexe,
- ▶ la taille des entrées est importante (ce qui est typiquement le cas en fouille de données).

⁴Chaum, Crépeau et Damgard

⁵Ben-Or, Goldwasser et Wigderson

Mesures de complexité et d'utilité

Deux mesures de complexité :

- ▶ **Complexité de communication** : nombre de bits échangés durant le protocole.
- ▶ **Complexité calculatoire** : temps de calcul requis localement par chaque participant pendant l'exécution du protocole.

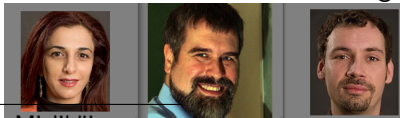
Mesure d'utilité :

- ▶ **Erreur de généralisation** : représente l'erreur que fera le classifieur dans le futur sur des cas non-rencontrés auparavant.
- ▶ S'estime par l'erreur obtenue sur un ensemble de test.
- ▶ **Important** : toujours comparer l'utilité de la version préservant la confidentialité d'un algorithme à celle de sa version standard.

Algorithmes d'apprentissage préservant la confidentialité

Implémentations d'algorithmes d'apprentissage préservant la confidentialité :

- ▶ (Approximation) de ID3 (Lindell et Pinkas 00).
- ▶ Réseaux de neurones artificiels (Chang et Lu 01).
- ▶ Classifieur de Bayes naïf (Kantarcioglu et Vaidya 04).
- ▶ k -moyennes (Kruger, Jha et McDaniel 05).
- ▶ Machines à vecteurs de support (LLM⁶ 06).
- ▶ **Algorithme de boosting**, travail conjoint avec :
Esma Aïmeur, Gilles Brassard et Balázs Kégl.



⁶Laur, Lipmaa et Mielikäinen

AdaBoost

Adaptive Boosting (Freund et Schapire 97) : algorithme de boosting qui fait partie de l'état de l'art (prix Gödel en 2003).



Philosophie de boosting : créer un classifieur efficace en combinant itérativement plusieurs classifieurs faibles.

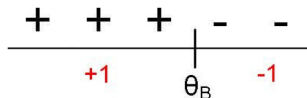
Condition d'apprentissage faible : la prédiction d'un classifieur faible peut être à *peine meilleure qu'une prédiction aléatoire*.

Propriétés intéressantes d'AdaBoost :

- ▶ Diminue l'erreur d'entraînement *exponentiellement rapidement avec le nombre d'itérations*.
- ▶ Peut continuer à *faire décroître l'erreur de test même lorsque l'erreur d'entraînement a atteint zéro*.

Étiquettes de décisions

- ▶ Famille de classifieurs faibles.
 - ▶ Arbre de décision avec seulement une racine et deux feuilles.
 - ▶ Peut conduire à un classifieur final très performant lorsqu'utiliser en conjonction avec AdaBoost.
 - ▶ Peut être décrit comme une règle telle que :
Si la valeur de l'attribut est inférieure à un seuil
 - ▶ alors l'objet appartient à la classe C1.
 - ▶ sinon l'objet appartient à la classe C2.
- ou graphiquement comme :



BiBoost et MultBoost

- ▶ **Contexte d'apprentissage** : les données sont partagées (*horizontalement*) entre plusieurs participants honnêtes-mais-curieux.
- ▶ **But** : construire un classifieur de type boosting de manière distribuée et préservant la confidentialité des données.
- ▶ **Modèle de communication** :
 - ▶ Canal privé entre chaque paire de participants.
 - ▶ Canal de diffusion authentifié.
- ▶ Deux algorithmes :
 - ▶ **BiBoost** (*Bipartite Boosting*).



- ▶ **MultBoost** (*Multiparty Boosting*), pour un nombre de participants $m > 2$

Utilité de MultBoost sur Pendigits

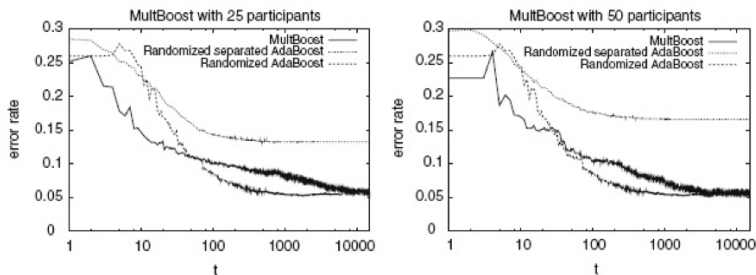


Fig. 8 Comparison of randomized ADABOOST, randomized separated ADABOOST, and MULTBOOST with 25 and 50 participants on the pendigits data set during 15,000 iterations

Complexité de MultBoost

Complexité de communication : $\Theta(Tm^2 \log k)$.

Complexité de calcul (avec étiquettes de décision) :
 $\Theta(T(dn + em))$.

Paramètres de l'algorithme :

- ▶ m le nombre de participants,
- ▶ k la taille de la famille de classifieurs faibles utilisée,
- ▶ T le nombre d'itérations de l'algorithme de boosting,
- ▶ n , le nombre de points de données,
- ▶ d le nombre d'attributs (soit la dimension) et
- ▶ e le temps requis pour faire une encryption.

Sécurité cryptographique de MultBoost

- ▶ **Menace spécifique au cas multiparti** : une *collusion* peut survenir si plusieurs participants mettent en commun les informations qu'ils ont accumulé durant l'exécution du protocole.
⇒ peut amener à un bris total de confidentialité
- ▶ Utilisation d'un **canal de diffusion anonyme** pour éviter la tracabilité de l'origine des classifieurs faibles.
- ▶ Un **protocole sécuritaire de calcul de somme** permet de calculer l'erreur global du classifieur généré à chaque itération sans avoir à divulguer les erreurs individuelles.
- ▶ Le reste des informations échangées est directement inclus dans la description du classifieur final.

Limite du paradigme cryptographique

- ▶ **Paradigme cryptographique** : un protocole calculant une fonction f de manière sécuritaire ne doit pas révéler plus d'information que la sortie de la fonction.
- ▶ **Problème** : il est possible que la sortie de la fonction elle-même révèle beaucoup d'information.
- ▶ **Exemple** : pour calculer sécuritairement (dans le sens cryptographique) un classifieur de type k -plus proches voisins, il suffit de révéler publiquement la description de votre ensemble de données.

Préservation de la confidentialité du classifieur final

- ▶ **Question ouverte fondamentale** : comment quantifier l'information révélée par la description d'un classifieur?
- ▶ **Borne supérieure sur l'information révélée** : taille de la description du classifieur mais ...
- ▶ cela ne dit rien à propos de la “qualité” de l'information révélée.
- ▶ **Intuitivement** : les classifieurs “opaques” (tel que les réseaux de neurones) semblent révéler moins d'information que les classifieurs “transparentes” (comme les k -plus proches voisins ou les machines à vecteurs de support).
- ▶ Approches possibles pour limiter la divulgation d'information :
 - ▶ Avoir un paramètre qui contrôle la complexité du modèle.
 - ▶ Injecter de l'aléatoire dans l'algorithme.

Propriétés des étiquettes de décisions

Propriétés intéressantes des étiquettes de décision du point de vue protection de la confidentialité :

- ▶ Leurs descriptions requièrent seulement un nombre constant de bits $\Theta(1)$.
- ▶ Il est difficile de trouver comment les projections de deux classifieurs faibles différents sont connectés s'ils ont été choisis par un mécanisme de randomisation.
 \Rightarrow le nombre d'ensembles de données consistant avec p étiquettes de décision en d dimension est de l'ordre de $(p!)^{d-1}$
- ▶ Il est facile de partitionner les étiquettes de décision en sous-ensembles, possiblement selon ses propres critères de protection de la vie privée.

Conclusion et perspectives futures

Perspectives futures

- ▶ **Développer** un cadre formel et des techniques permettant de *quantifier la quantité d'information contenue dans la description de f sur l'ensemble de données originel*, et pas seulement l'information révélée durant l'exécution du protocole (contrairement au paradigme habituel).
- ▶ **Explorer** l'utilisation de *versions approximatives* et/ou de la *randomisation* à l'intérieur même des algorithmes d'apprentissage afin de préserver la confidentialité.
- ▶ **Combiner** plusieurs approches.
Exemple: perturbation + calcul multiparti sécuritaire.
- ▶ **Développer** des variantes efficaces des algorithmes d'apprentissage dans le modèle où les participants peuvent être malicieux.

Travail en cours : comment gérer les participants utilisant des ensembles de données totalement artificiels?

- ▶ **Problème fondamental** : comment empêcher un participant de donner en entrée au protocole un ensemble de données totalement artificiel?
- ▶ **Modèle de sécurité considéré** : les participants sont potentiellement malicieux et peuvent dévier de l'exécution du protocole.
- ▶ Dans ce cas, ils ont une probabilité non-négligeable de déclencher une alarme publique et le protocole se termine.
- ▶ Un participant honnête peut aussi décider de terminer le protocole plus tôt s'il considère qu'il n'a plus aucun gain à espérer.

Travail en cours : comment gérer les participants utilisant des ensembles de données totalement artificiels? (suite)

- ▶ **Initialisation** : chaque participant s'engage (*commitment* en anglais) sur une version cryptée de son ensemble de données.
- ▶ **Révélation partielle d'information** : un protocole de transfert inconscient de type *k-out-of-n* chaînes de bits est conduit entre chaque paire de participants :
 - ▶ qui révèle la description de *k* points de données parmi les *n* sans que
 - ▶ le possesseur de l'ensemble de données apprennent lesquels.
- ▶ Un **test statistique sur l'information révélée** sert à décider si l'ensemble de données semble contenir de l'information utile.
- ▶ Si c'est le cas l'apprentissage se poursuit normalement sur les ensembles "engagés",
- ▶ sinon le protocole avorte.

C'est la fin !

Merci pour votre attention.
Questions ?

