Préservation de la vie privée

Pascal Poncelet

Ecole thématique BDA, Les Houches, 16-21 mai 2010

Laboratoire
d'Informatique
de Robotique
et de Microélectronique
de Montpellier
LIRMM

UNIVERSITÉ MONTPELLIER 2
SCIENCES ET TECHNIQUES

cnrs

Plan

- Protection de la vie privée ?
- K-anonymisation
- Fouille de données et Vie privée : un exemple
- Vers des fonctions d'oubli
- Quelques challenges

Plan

- Protection de la vie privée ?
- K-anonymisation
- Fouille de données et Vie privée : un exemple
- Vers des fonctions d'oubli
- Quelques challenges

De l'utilisation des sites de réseaux sociaux

- 72% des internautes sont inscrits sur au moins un réseau social (Sources Insites Consulting (mars 2010))
- 940 millions de personnes
- 72% sont inscrits sur au moins 2 plateformes. Moyenne de 2 connexions par jour sur les réseaux sociaux
- réseaux orientés Professionnels :
  - Nombre de connexions quotidiennes : 9
- Situation : Facebook (51%)

Myspace (20%), Twitter (17%)

- Temps passé : *(source Institut Nielsen, janvier 2010))*
- Augmentation de 82%

entre décembre 2008 à décembre 2009

- (Moyenne : 5h 35 minutes) France : 4h 04

De l'utilisation des sites de réseaux sociaux

- Sondage mené par Harris Interactive,
  - 45% des recruteurs Américains déclarent utiliser les sites de réseaux sociaux (Facebook, MySpace, Linkedin, Twitter, etc.) pour trouver des informations sur des candidats qui postulent à leurs offres d'emploi
  - 35% ont écarté des candidats en raisons ce que qu'ils ont trouvé :
    - 53 % publication du candidat de photos ou d'informations provocantes ou déplacées
    - 44 % parce que l'on voit les candidats buvant ou se droguant
    - 35 % parce qu'ils crachaient sur leurs anciens employeurs, leurs collègues ou leurs clients
    - 29 % parce qu'ils montraient un déficit de communication
    - 26 % parce qu'ils publiaient des propos discriminatoires
    - 24 % parce qu'ils mentaient sur leurs diplômes et
    - 20 % parce qu'ils ont publié des informations confidentielles sur leurs anciens employeurs
- Allemagne : 28% des employeurs (500 entreprises) utilisent Internet pour recueillir des informations dès le début du recrutement

Les amis de mes amis

- Entretien avec Alex Türk, président de la Cnil (Commission nationale de l'informatique et des liberté).
- « **Un de ses copains a pris la photo et l'a balancé sur le réseau social. C'est amusant. Quelques mois plus tard, il était candidat sur un poste et le recruteur lui a glissé sous les yeux la photo de ses fesses en lui demandant s'il était coutumier de ces pratiques »**. *Source (site Internet du quotidien La Provence)*

- « Oh mon dieu ! Je hais mon boulot » ajoutant que son responsable était « pervers » et qu'il ne lui donnait que « du travail de m… »
- …4 heures plus tard…
- « Tout d'abord arrêtez de vous flatter, cela ne fait que 5 mois que vous travaillez ici, n'avez pas remarqué que je suis gay. Ensuite le travail de m… comme vous dites est le travail pour lequel je vous paye […]. Vous semblez avoir oublié qu'il vous restait encore deux semaines de travail en période d'essai. **Ne prenez pas la peine de revenir demain**. »
- Son patron était en relation sur Facebook

Source Grande Bretagne - Août 2009

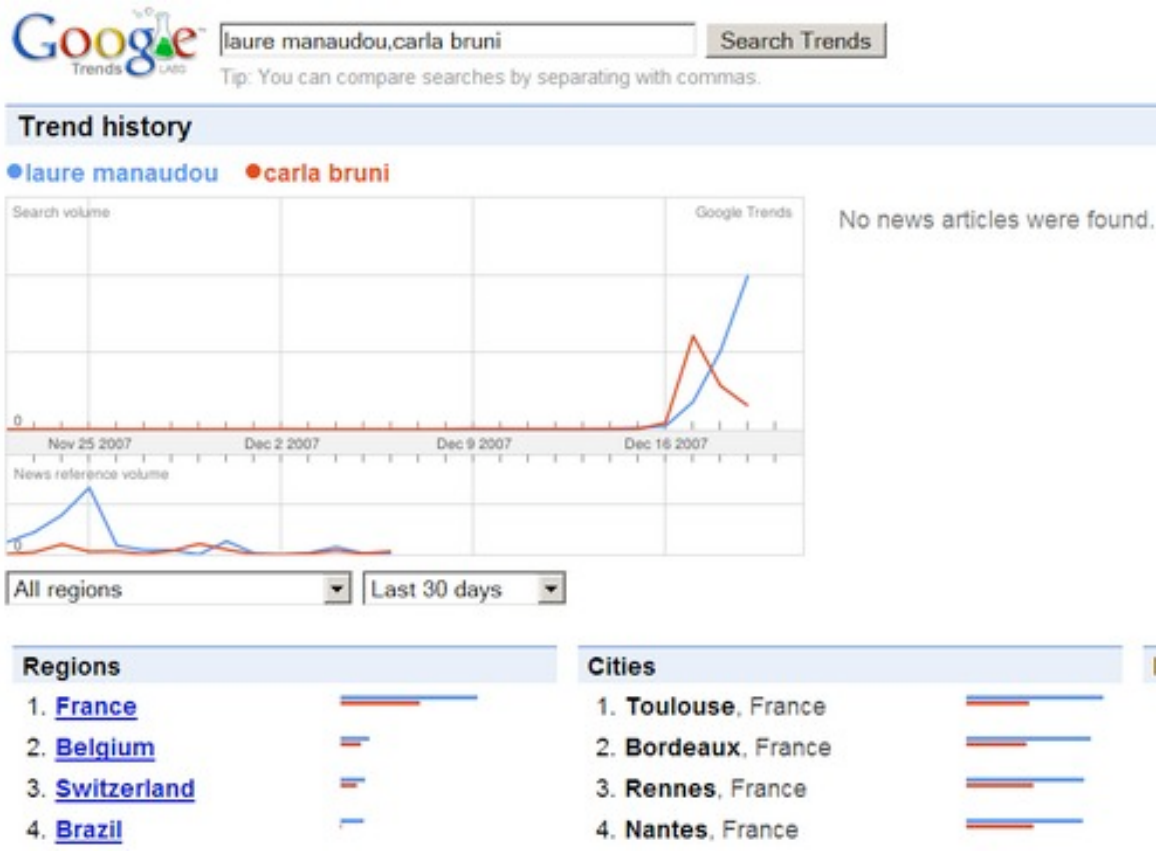
Notre responsabilité

- Expérience de l'éditeur britannique Sophos (2007)
- Création d'un compte Freddy Staur

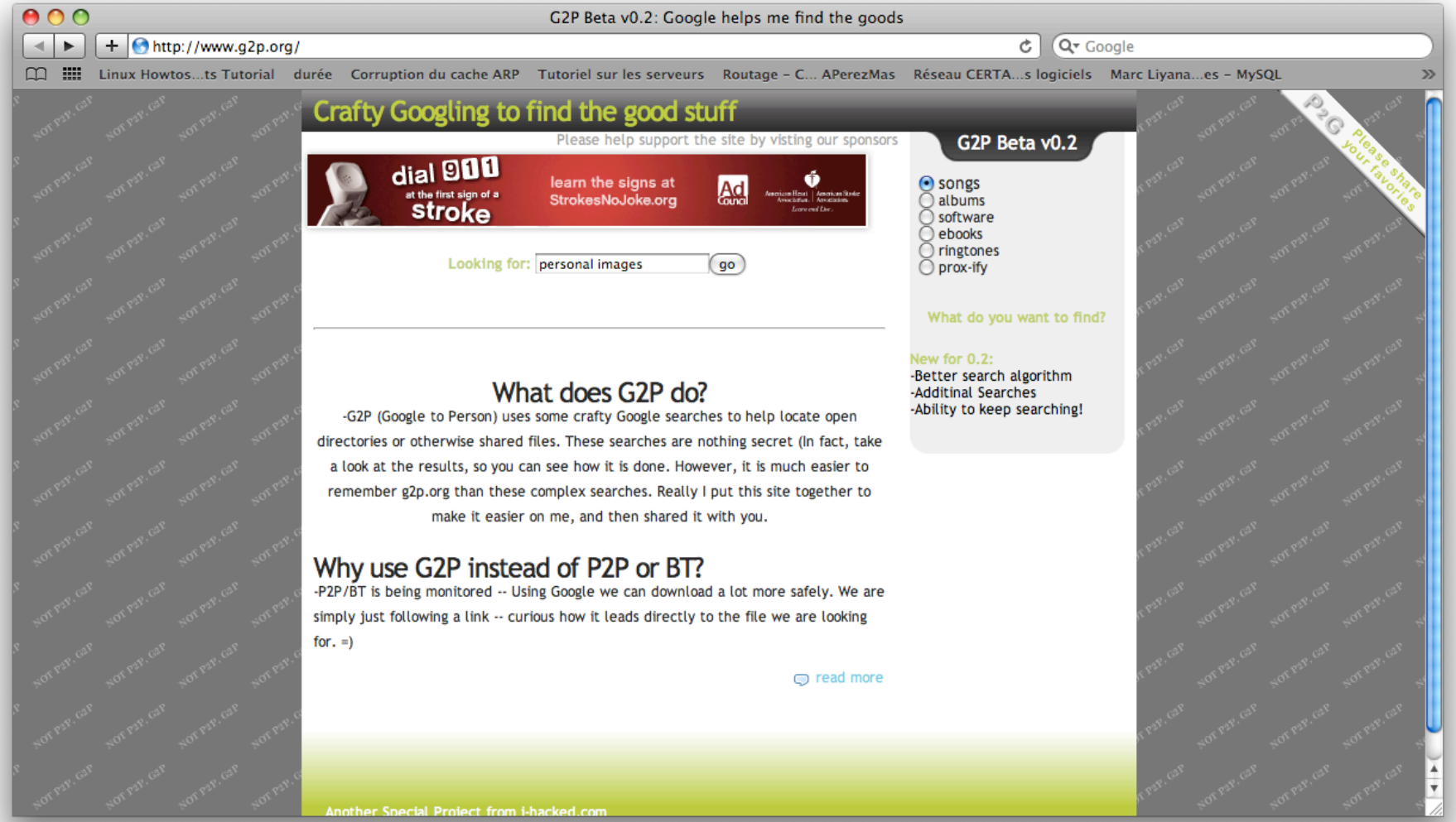
- Envoi de Friends à un échantillon de 200 personnes sur FaceBook
- **87** personnes ont répondu en donnant accès à des photos de familles, des informations sur leur goûts, le nom de leur compagnon, compagne, (le nom de jeune fille de leur mère) leur CV ….

Les moteurs de recherche

- Les photos de Laure Manaudou - décembre 2007



Difficile ?

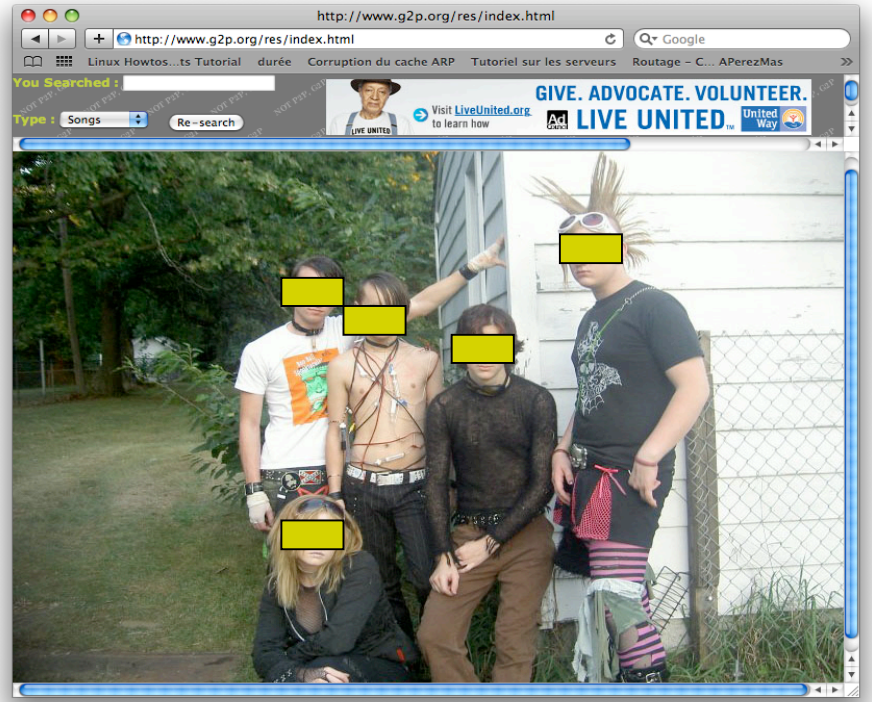

G2P Beta v0.2: Google helps me find the goods
http://www.g2p.org/
Google
Linux Howtos...ts Tutorial
durée
Corruption du cache ARP
Tutoriel sur les serveurs
Routage - C... APerezMas
Réseau CERTA...s logiciels
Marc Liyana...es - MySQL
Crafty Googling to find the good stuff
Please help support the site by visting our sponsors
dial 911 at the first sign of a stroke
learn the signs at StrokesNoJoke.org
Ad Council
Looking for:
personal images
go
G2P Beta v0.2
songs
albums
software
ebooks
ringtones
prox-ify
What do you want to find?
New for 0.2:
-Better search algorithm
-Additinal Searches
-Ability to keep searching!
Please share your favories
What does G2P do?
-G2P (Google to Person) uses some crafty Google searches to help locate open directories or otherwise shared files. These searches are nothing secret (in fact, take a look at the results, so you can see how it is done. However, it is much easier to remember g2p.org than these complex searches. Really I put this site together to make it easier on me, and then shared it with you.
Why use G2P instead of P2P or BT?
-P2P/BT is being monitored -- Using Google we can download a lot more safely. We are simply just following a link -- curious how it leads directly to the file we are looking for. =)
read more
Another Special Project from i-hacked.com

Non - google requêtes complexes

La requête google :

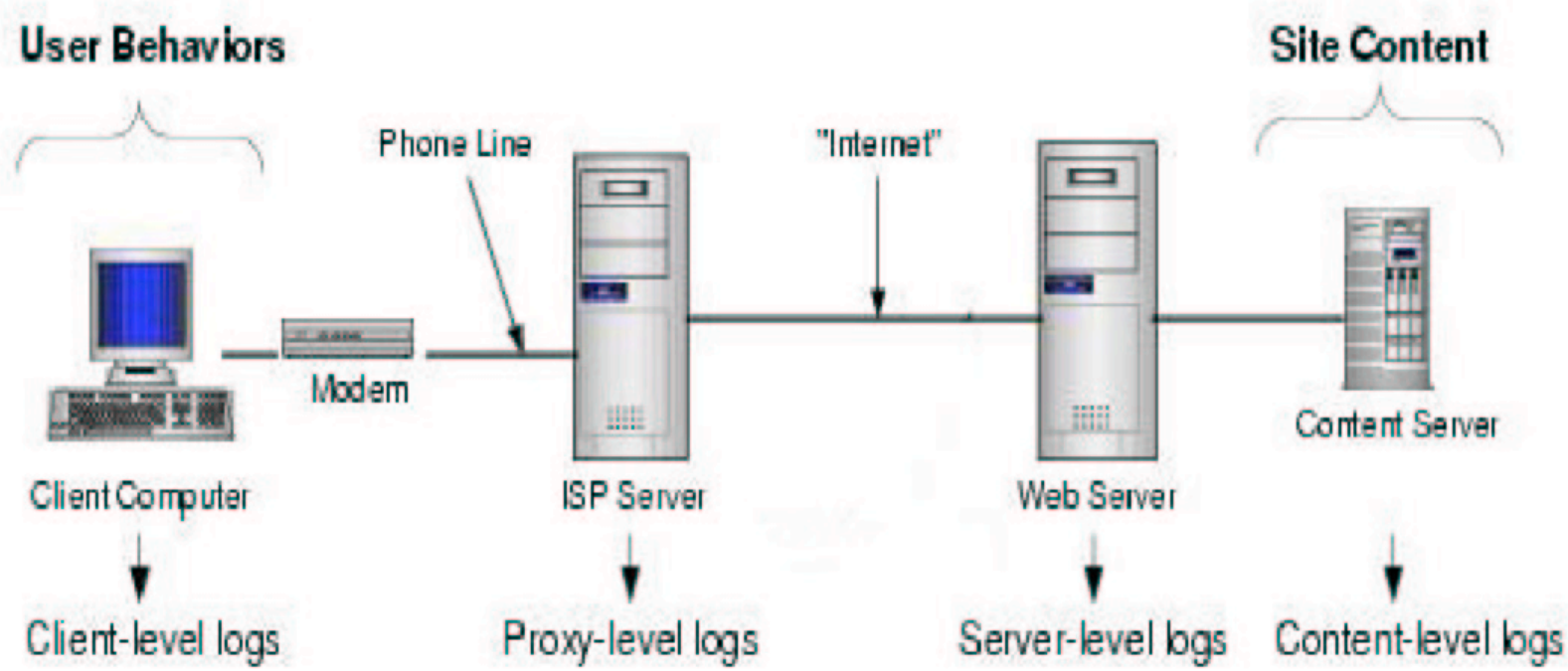
```
intitle:index.of +"Last modified "
   +"Parent directory " +(XXXXXXXXX)
   +(jpeg) +"" -htm -html -php -asp
```

XXXXMyBestFriendsXXXXXX.jpg

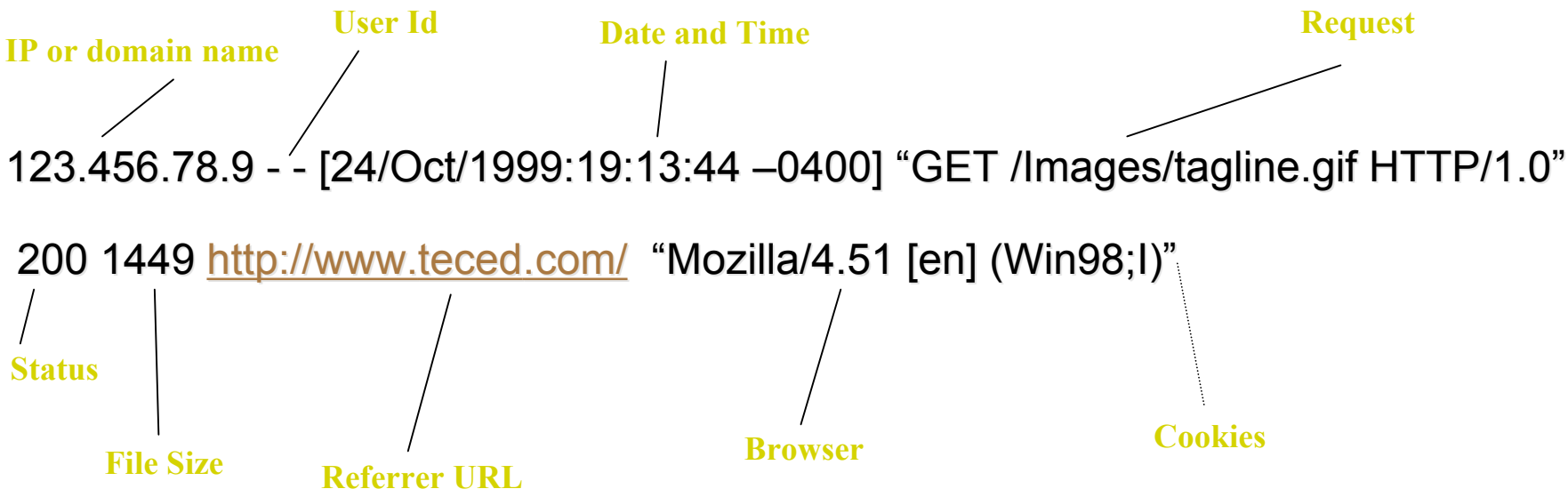


Log ou Logs ?

Information sur les chemins de navigation dans les fichiers logs



Web logs + ?



Bases de données des achats
Bases de données des partenaires
Géolocalisation
Cookies

Une vrai valeur commerciale

- Décembre 2007, (Google, Microsoft, MySpace, AOL et Yahoo!), ont enregistré 336 milliards de données personnelles
- Yahoo! a récolté 110 milliards de transmissions de données, soit en moyenne 811 (1.700 avec l'ensemble de ses partenaires) informations pour chaque internaute ayant visité un de ses sites durant cette période.
- 110 milliards de données personnelles en un mois !
- Dresser un portrait-robot fiable de l'internaute consommateur
- De 10 à 50 euros !!

Tout s’achète

- Site de ventes en ligne sur les clients intéressés par la voyance
- Nom, prénom, adresse, numéro de CB
- 1 euro par personne

- A essayer :)

Les bases clients protégées ?

- Janvier 2009 : 400 000 fiches du fournisseur d'accès à Internet Orange laissées en libre accès sur Internet via une faille de sécurité
- Octobre 2008 : 30 millions de données de Deutche Telekom (avec numéros de CB)
- Août 2008 : les données bancaires d'un million de clients en vente sur eBay (pour 44 euros)
- Janvier 2009 : 4 millions de comptes visités par des hackers sur Monster
- Mars 2010 : Fichier SNCF (1 adresse et coordonnées d'un voyageur 8 à 20 euros)

De l'anonymisation

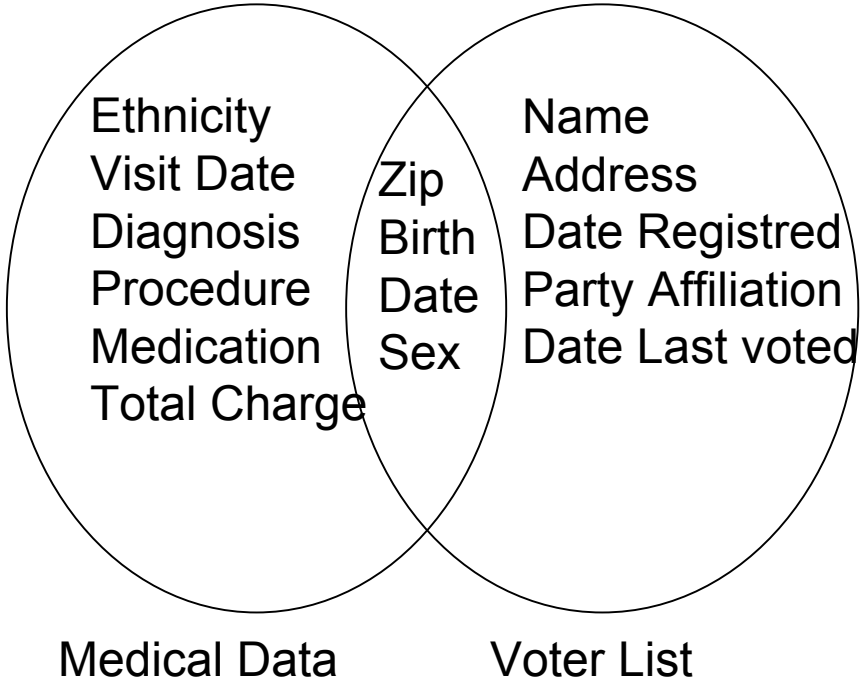
- Expérience d'AOL en 2006
- Une liste de 20 millions de recherche d'internautes mis en ligne après avoir été anonymisées
- No. 4417749 a effectué de nombreuses recherches sur « un homme célibataire de 60 ans » et « des informations sur un chien qui urine partout »
- En recherchant, localisation (Lilburn, Ga), vue d'un lac, …
- Thelma Arnold, a 62-year-old veuve qui vie à Lilburn, Georgie

De l'anonymisation

- Fichier anonymisé des soins de santé des fonctionnaires de l'état du Massachusetts mis en ligne (L. Sweeney, 1997)
- La liste électorale de Cambrige, MA (53 805 inscrits)

- 69 % d'enregistrements uniques par rapport à code postal, date de naissance

Dossier médical du gouverneur du Massachusetts



Plan

- Protection de la vie privée ?
- K-anonymisation
- Fouille de données et Vie privée : un exemple
- Vers des fonctions d'oubli
- Quelques challenges

Pourquoi la k-anonymisation ?

Table médicale : anonyme

SSN	Name	Race	DOB	Sex	Zip	Marital	Heath
		Asian	09/07/64	F	22030	Widow	Obesity
		Black	05/14/61	M	22030	Married	Obesity
		White	05/08/61	M	22030	Married	Chest pain
		White	09/15/61	M	22031	Married	**Aids**

Table de la liste des votants : publiques

Name	Address	City	Zip	DOB	Sex	Party
….	….	….	….	….	….	….
John DOE	900 Market St.	New York	22031	09/15/61	M	Democrat

John Doe qui habite
New York a le sida !!!!

K-anonymisation

- Latanya Sweeney (Canergie Mellon University)
- Si l'information de chaque personne contenue dans la base ne peut pas être distinguée d'au moins k-1 personnes dont les informations apparaissent dans la base
- Exemple de k-anonymité :
  - essai d'identification d'une personne avec comme seules informations sa date de naissance et son sexe. Il y a k-personnes qui vérifient ces propriétés dans la base

Objectifs

- Limiter la capacité à pouvoir lier les informations restantes à d'autres informations externes

- Besoin d'identifier tous les attributs des données intimes pouvant être utilisés pour faire des jointures

- Attention :
- Nom, prenom, identification, pas suffisant !
- 87% de la population des Etats Unis peut être indentifée par *code postal, date de naissance et sexe* (c.f. données de recensement)

Quasi-identifier

- Ensemble minimal d'attributs dans la relation pouvant être liés avec des infos externes pour re-identifier les enregistrements des individus.
- Plus formellement :
- *Etant donné une population d'entité U et une relation d'entités spécifiques T(A1, ..An), fc:U→ T et fg:T→U' /U⊆ U'.*

  Un quasi-identifiant de T noté Q_T est un ensemble d'attributs {Ai, …Aj}⊆{{A1, …An}, ∃pi∈U fg(fc(pi)[Q_T])=pi

- V: relation spécifique des votants
- Quasi-indentifiant Q_V de V : {*name, address, zip, birthday, genre*}
- Lien avec table médicale: {*DoB, ZIP, sexe*} ⊆ Q_V

- Mais : {*name, address*} ⊆ Q_V car les attributs peuvent aussi apparaître comme informations externes et utilisés pour les jointures

Notion d'ensemble fréquent

Hospital Patient Data

Birthdate	Sex	Zipcode	Disease
1/21/76	Male	53715	Flu
4/13/86	Female	53715	Hepatitis
2/28/76	Male	53703	Brochitis
1/21/76	Male	53703	Broken Arm
4/13/86	Female	53706	Sprained Ankle
2/28/76	Female	53706	Hang Nail

```
Select count(*)
    from patients
    group by sex, zipcode
```

K-anonymisation

- Une relation est dite k-anonyme si chaque count dans l'ensemble fréquent est ≥ k
- Un tuple dans une relation ne peut pas être distingué de au moins K autres lignes

- Lemme :
  - Soit $RT(A_1, ..A_n)$ une relation et $QI_{RT} = (A_i, ..A_j)$ l'ensemble de quasi-indentifier associé à RT, $A_i, ..A_j \subseteq A_1, .. A_n$ et RT satisfait la k-anonymity. Alors chaque séquence de valeur dans RT[Ax] apparaît avec au moins k occurrences dans $RT[QI_{RT}]$ pour x=i..j

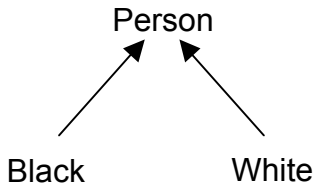
K-anonymité, K=2, QI={Race, Birth, Gender, Zip}

Race	Birth	Gender	ZIP	Problem
Black	1965	m	0214*	short breath
Black	1965	m	0214*	chest pain
Black	1965	f	0213*	hypertension
Black	1965	f	0213*	hypertension
Black	1964	f	0213*	obesity
Black	1964	f	0213*	chest pain
White	1964	m	0213*	chest pain
White	1964	m	0213*	obesity
White	1964	m	0213*	short breath
White	1967	m	0213*	chest pain
White	1967	m	0213*	chest pain

Pour chaque tuple contenu dans T , les valeurs du tuple comprenant le QI apparaît au moins 2 fois dans T. t1[QIT]= t2[QIT] t3[QIT]=t4[QIT]…..

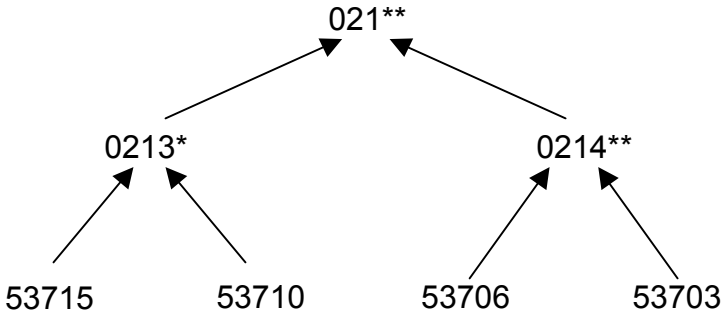
Chaque valeur apparaissant dans une valeur associée aux attributs de QI dans T apparaît au moins k fois. |T[Race=« black »] | = 6, |T[Birth=« 1967 »] | = 2,…

Prise en compte de hiérarchie de généralisation



Race E_0	ZIP Z_0
Black	02138
Black	02139
Black	02141
Black	02142
White	02138
White	02139
White	02141
White	02142

Private Table



E1 = {Person}

↑

E0 = {Black, White}

Z2 = {021**}

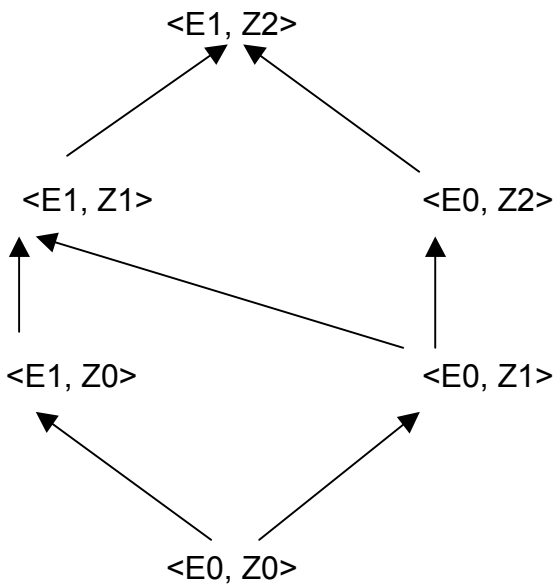
↑

Z1 = {0213*, 0214*}

↑

Z0 = {02138, 02139, 02141, 02142}

Et donc du treillis de spécialisation/généralisation



E1 = {Person}

↑

E0 = {Black, White}

Z2 = {021**}

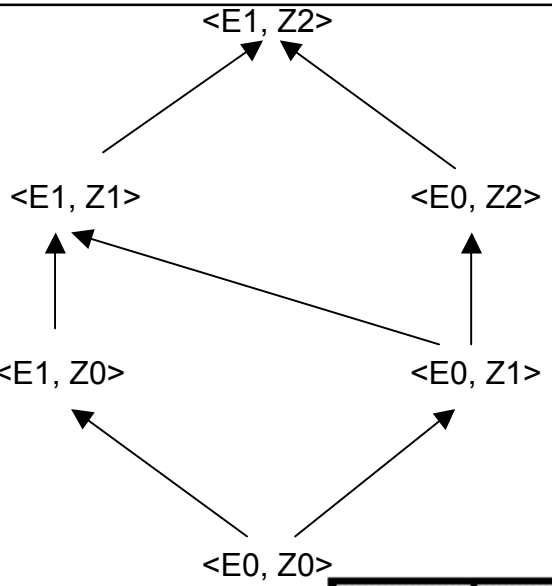
↑

Z1 = {0213*, 0214*}

↑

Z0 = {02138, 02139, 02141, 02142}

Généralisation des relations



Race E_0	ZIP Z_0
Black	02138
Black	02139
Black	02141
Black	02142
White	02138
White	02139
White	02141
White	02142

PT

Race E_1	ZIP Z_0
Person	02138
Person	02139
Person	02141
Person	02142
Person	02138
Person	02139
Person	02141
Person	02142

$GT_{[1,0]}$

Race E_0	ZIP Z_1
Black	0213*
Black	0213*
Black	0214*
Black	0214*
White	0213*
White	0213*
White	0214*
White	0214*

$GT_{[0,1]}$

Race E_0	ZIP Z_2
Black	021**
Black	021**
Black	021**
Black	021**
White	021**
White	021**
White	021**
White	021**

$GT_{[0,2]}$

Race E_1	ZIP Z_1
Person	0213*
Person	0213*
Person	0214*
Person	0214*
Person	0213*
Person	0213*
Person	0214*
Person	0214*

$GT_{[1,1]}$

Attaques contre les relations k-anonymes

Race	ZIP
Asian	02138
Asian	02139
Asian	02141
Asian	02142
Black	02138
Black	02139
Black	02141
Black	02142
White	02138
White	02139
White	02141
White	02142

PT

Race	ZIP
Person	02138
Person	02139
Person	02141
Person	02142
Person	02138
Person	02139
Person	02141
Person	02142
Person	02138
Person	02139
Person	02141
Person	02142

GT1

Race	ZIP
Asian	02130
Asian	02130
Asian	02140
Asian	02140
Black	02130
Black	02130
Black	02140
Black	02140
White	02130
White	02130
White	02140
White	02140

GT2

Ordre d'apparition des tuples dans les relations générées

Solution : trié de manière différentes les relations

Attaques avec les relations k-anonymes complémentaires

Race	BirthDate	Gender	ZIP	Problem
black	9/20/1965	male	02141	short of breath
black	2/14/1965	male	02141	chest pain
black	10/23/1965	female	02138	painful eye
black	8/24/1965	female	02138	wheezing
black	11/7/1964	female	02138	obesity
black	12/1/1964	female	02138	chest pain
white	10/23/1964	male	02138	short of breath
white	3/15/1965	female	02139	hypertension
white	8/13/1964	male	02139	obesity
white	5/5/1964	male	02139	fever
white	2/13/1967	male	02138	vomiting
white	3/21/1967	male	02138	back pain

PT

Race	BirthDate	Gender	ZIP	Problem
black	1965	male	02141	short of breath
black	1965	male	02141	chest pain
person	1965	female	0213*	painful eye
person	1965	female	0213*	wheezing
black	1964	female	02138	obesity
black	1964	female	02138	chest pain
white	1964	male	0213*	short of breath
person	1965	female	0213*	hypertension
white	1964	male	0213*	obesity
white	1964	male	0213*	fever
white	1967	male	02138	vomiting
white	1967	male	02138	back pain

GT1

Race	BirthDate	Gender	ZIP	Problem
black	1965	male	02141	short of breath
black	1965	male	02141	chest pain
black	1965	female	02138	painful eye
black	1965	female	02138	wheezing
black	1964	female	02138	obesity
black	1964	female	02138	chest pain
white	1964	male	02138	short of breath
white	1965	female	02139	hypertension
white	1964	male	02139	obesity
white	1964	male	02139	fever
white	1967	male	02138	vomiting
white	1967	male	02138	back pain

LT

Jointure sur problem

Race	BirthDate	Gender	ZIP	Problem
black	1965	male	02141	short of breath
black	1965	male	02141	chest pain
black	1965	female	02138	painful eye
black	1965	female	02138	wheezing
black	1964	female	02138	obesity
black	1964	female	02138	chest pain
white	1960-69	male	02138	short of breath
white	1960-69	human	02139	hypertension
white	1960-69	human	02139	obesity
white	1960-69	human	02139	fever
white	1960-69	male	02138	vomiting
white	1960-69	male	02138	back pain

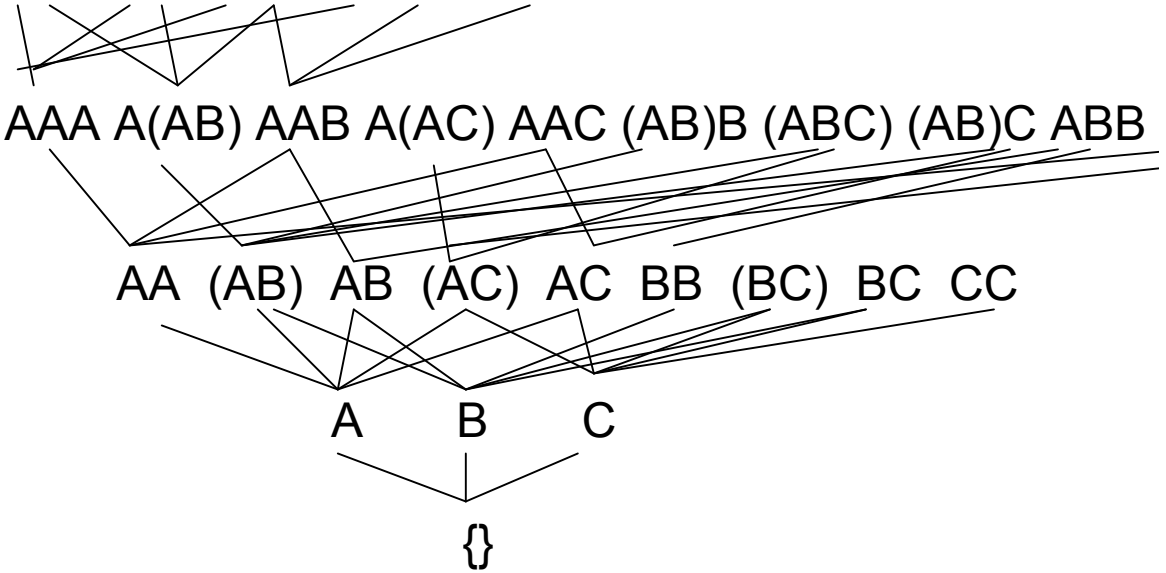
GT3

Plan

- Protection de la vie privée ?
- K-anonymisation
- Fouille de données et Vie privée : un exemple
- Vers des fonctions d'oubli
- Challenges

Problématique des motifs séquentiels

Trans. ID	Items
1	(A) (A,B,D)
2	(A) (C,D) (A)
3	(A) (B,D) (C) (B)
4	(A) (B) (C, D) (A)



Itemsets : A, B ou A, B, C

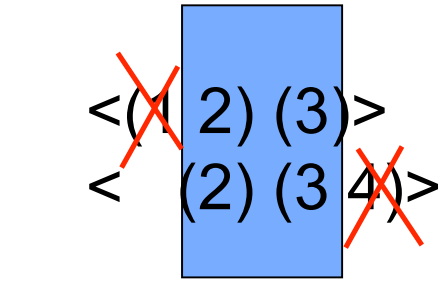
Séquence : ⟨(A) (C,D) (A) ⟩

Support pour une séquence : Supp (⟨(A) (C,D) (A) ⟩)=2

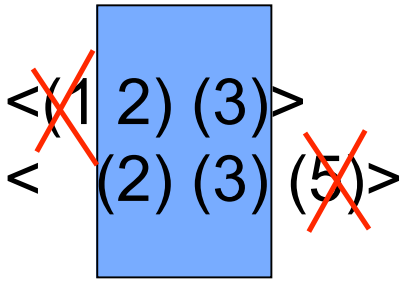
Motifs séquentiels fréquents (minSupp=50%) : ⟨(A) (A)⟩,
⟨(A) (B,D) ⟩, ⟨(A) (C,D) ⟩

Génération des candidats

- S-Extension : ajout d'une séquence
- I-Extension : ajout d'un itemset



<(1 2) (3 4)>
I-Extension



<(1 2) (3) (5)>
S-Extension

SPAM

- Utilisation de bitmaps pour rechercher les motifs fréquents
- Hypothèse : la base tient toujours en mémoire
- On construit d'un arbre lexicographique contenant toutes les branches possibles – élimination des branches en fonction du support (cf. espace de recherche)
- Nouvelle représentation des données

SPAM (cont.)

- Représentation verticale des données

$C1 = <(1)_3\ (1)_5>$

		(1)
C1	**T1**	0
	T2	0
	T3	1
	T4	0
	T5	1

- S-Extension
- I-Extension

SPAM (cont.)

- I-Extension : AND
- Exemple : recherche du candidat (1,2)

(1)
0
0
1
0
1

AND

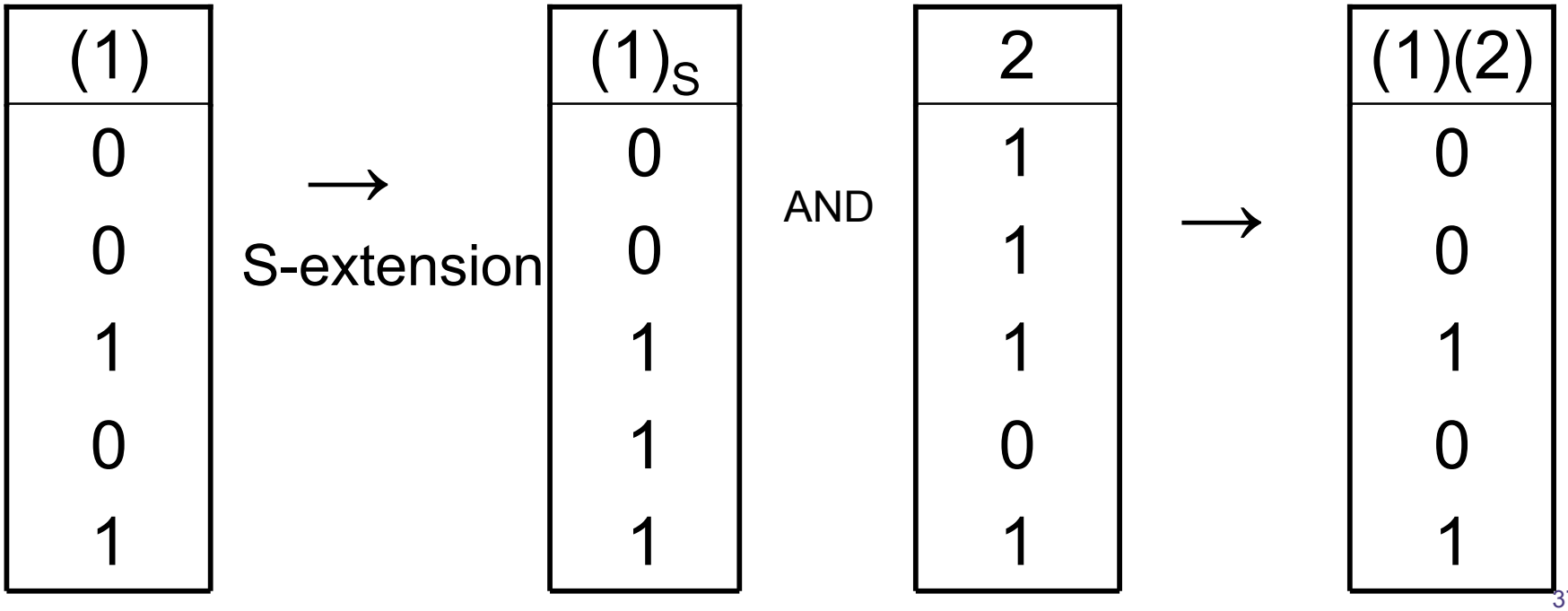
2
1
1
1
0
1

→

(1,2)
0
0
1
0
1

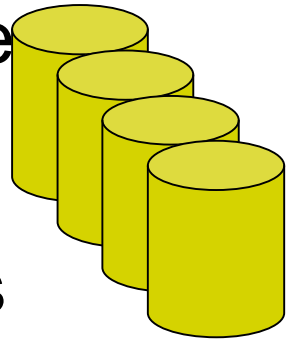
SPAM (cont.)

- S-Extension : un bitmap transformé + AND
- Exemple : recherche du candidat (1) (2)



Motifs séquentiels & BD distribuées

- Un ensemble de bases de donnée
  $DB = DB_1 \cup DB_2 \ldots DB_n$



- Extraire les séquences fréquentes
  Pour c1 : Alice $(1)_1$, Bob $(2)_2$, Carol $(7)_4$, Alice $(3)_5$

CID	Alice	Bob	Carol
1	$(1)_1$ $(3)_5$	$(2)_2$	$(7)_4$
2	$(2)_4$	$(1)_3$	$(3)_6$
3	$(2)_6$ $(3)_7$		$(1)_2$ $(7)_3$

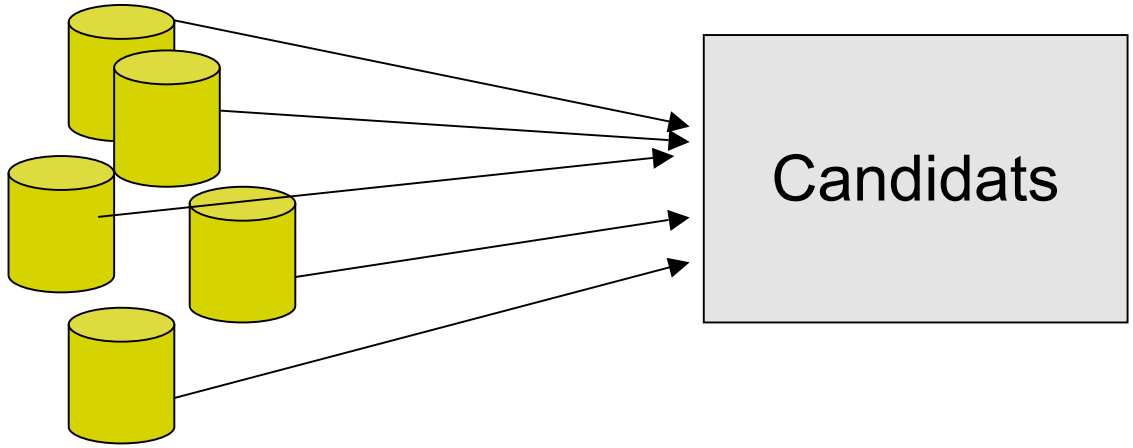
Motifs séquentiels collaboratifs

- Une représentation verticale des items en fonction des bases

		V_1^1	…	V_1^D
C1	T1	0		1
	T2	0		0
	T3	1		0
	T4	1		1
C2	T1	0		1
	T2	1		0
	T3	0		0
	T4	0		1

Motifs séquentiels collaboratifs (cont.)

- Hypothèse : un algorithme de génération-vérification est disponible
- Génération : combiner les k-1 séquences fréquentes pour générer des candidats de taille k
- Améliorer la phase de vérification



Motifs séquentiels collaboratifs (cont.)

- Exemple : Recherche de (1) (2) dans DB_1, DB_2

  - Demander à DB_1 et DB_2 leur vecteur correspondant à l'item spécifique (1) (V_1^1 et V_1^2)
  - Faire un *OR* logique entre V_1^1 et V_1^2
  - Utiliser la S-extension de SPAM (fonction F)
  - Demander à DB_1 et DB_2 le vecteur pour (2) (V_2^1 et V_2^2)
  - Faire un *OR* logique entre V_2^1 et V_2^2
  - Appliquer un *AND* entre les deux (fonction G)
  - Convertir le bitmap en entier et compter le nombre de 1 (fonction $\sum$)

Motifs séquentiels collaboratifs (cont.)

		V_1^1
C1	T1	0
	T2	0
	T3	1
	T4	0
C2	T1	0
	T2	1
	T3	0
	T4	0

OR

		V_1^D
C1	T1	0
	T2	1
	T3	1
	T4	0
C2	T1	1
	T2	0
	T3	0
	T4	0

$Z1 = f(V_1^1 \text{ OR } V_1^2)$

$\longrightarrow$ S-Extension

		Z1
C1	T1	0
	T2	0
	T3	1
	T4	1
C2	T1	0
	T2	1
	T3	1
	T4	1

		Z1
C1	T1	0
	T2	0
	T3	1
	T4	1
C2	T1	0
	T2	1
	T3	1
	T4	1

AND

		Z2 = V21 OR V22
C1	T1	0
	T2	0
	T3	0
	T4	1
C2	T1	1
	T2	1
	T3	1
	T4	1

$\longrightarrow$ G=(Z1 AND Z2)

	Z3
C1	1
C2	1

$\xrightarrow[\sum]{}$ 2

Préservation de la vie privée

- Une contrainte forte :

Alice, Bob et Carol ne peuvent pas donner d'information sur le contenu de leur base

- Problème : Evaluation d'une séquence candidate
  - Est-ce que l'item 1 pour le client 1 appartient à la base de Carol ?

Préservation de la vie privée (cont.)

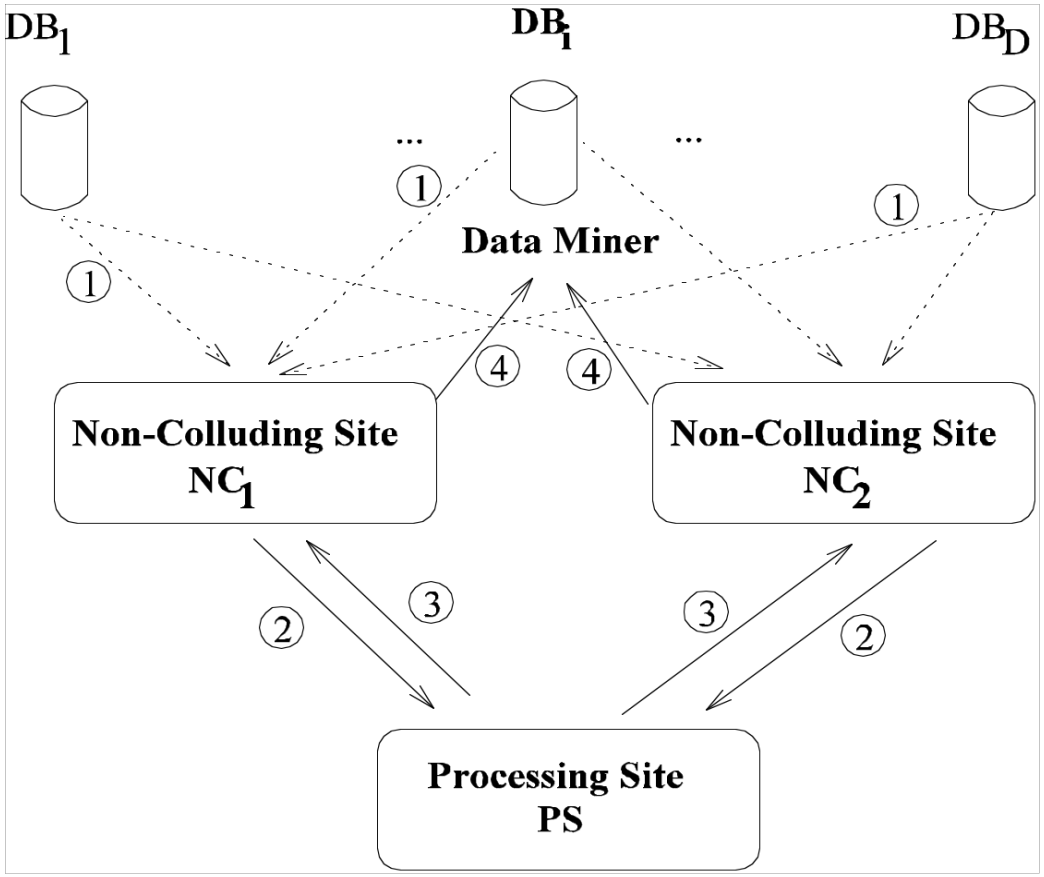
- Collaboratif ?
- Considérer de nouvelles fonctions sûres (AND^S, OR^S, G^S, F^S, $\sum^S$)
- Réaliser la vérification sans fournir d'information des bases de données sources (Alice, Bob, Carol)
- Une nouvelle architecture
  - Trois sites ne collaborant pas (semi honnêtes)

Préservation de la vie privée (cont.)

Site « Data Miner »
Réalise les opérations de fouille

Sites « Non Colluding »
Réalisent les opérations sûres

Site « Processing »
Calcul des fonctions



Préservation de la vie privée (cont.)

- Etape de prétraitements
- Ajout de faux clients dans les bases de données sources
  - Idée : on augmente le nombre de clients pour éviter qu'une partie puisse obtenir le bon résultat par rapport au support
- Permuter la liste de clients
  - Idée : éviter pour un site de savoir de quel client il traite (cf K-anonymous)
- Comptage du support des clients ajoutés
  - Idée : supprimer le bruit au final

Préservation de la vie privée (cont.)

- Rappel :

XOR	1	0
1	0	1
0	1	0

Préservation de la vie privée (cont.)

- Envoi des données des BD sources aux sites « Non Colluding »
- NC_1 et NC_2 doivent avoir le minimum d'information
- Pour chaque base DB_i, pour chaque item it, générer un vecteur de bits aléatoirement de la même taille que le vecteur it (R_{DBi})
- $Z_{DBi} = V_{it}$ XOR R_{DBi}
- Envoyer Z_{DBi} à NC_1 et R_{DBi} à NC_2 (et vice versa)
- NC_1 et NC_2 : R_{DBi} ou vecteurs XOR-isés

Préservation de la vie privée (cont.)

- Un exemple : le protocole AND^S

- Entrée : $(X^+, Y^+ \mid X^-, Y^-)$ sont des bits tels que X^+ et Y^+ appartiennent à NC_1 et X^- et Y^- appartiennent à NC_2

- Sortie : $(A^R|B^R)$ sont tels que :

  A^R *XOR* $B^R = (X^+ XOR\ X^-)$ *AND* $(Y^+ XOR\ Y^-)$

Préservation de la vie privée (cont.)

- NC_1 et NC_2 génèrent mutuellement et s'échangent des nombres aléatoires R_A, R'_A, R_B et R'_B tels que : $X^{+'} = X^+ \mathit{XOR}\ R_A$, $Y^{+'} = Y^+ \mathit{XOR}\ R'_A$, $X^{-'} = X^- \mathit{XOR}\ R_B$ et $Y^{-'} = Y^- \mathit{XOR}\ R'_B$
- NC1 envoie $X^{+'}$ et $Y^{+'}$ à PS
- NC2 envoie $X^{-'}$ et $Y^{-'}$ à PS
- PS calcule : $C^+ = X^{+'} \mathit{AND}\ Y^{-'}$ et $C^- = X^{-'} \mathit{AND}\ Y^{+'}$ ainsi qu'un nombre aléatoire R_{PS}

Préservation de la vie privée (cont.)

- PS envoie $A'_{PS} = C^{+}$ *XOR* R_{PS} à NC_1 et $B'_{PS} = C^{-}$ *XOR* R_{PS} à N_{C2}
- NC_1 calcule $A^R = A'_{PS}$ *XOR* $(X^{+}$ *AND* $R'_B)$ *XOR* $(Y^{+}$ *AND* $R_B)$ *XOR* $(X^{+}$ *AND* $Y^{+})$ *XOR* $(R_B$ *AND* $R'_A)$
- NC_2 calcule $B^R = B'_{PS}$ *XOR* $(X^{-}$ *AND* $R'_A)$ *XOR* $(Y^{-}$ *AND* $R_A)$ *XOR* $(X^{-}$ *AND* $Y^{-})$ *XOR* $(R_A$ *AND* $R'_B)$
- Résultat final (AR|BR) tels que A^R *XOR* $B^R = (X^{+}$*XOR* $X^{-})$ *AND* $(Y^{+}$ *XOR* $Y^{-})$

Préservation de la vie privée (cont.)

- Résultat final
- $A^R \; XOR \; B^R = (X^+ \; AND \; R'_B) \; XOR \; (Y^+ \; AND \; R_B) \; XOR \; (X^+ \; AND \; Y^+) \; XOR \; (R_B \; AND \; R'_A) \; XOR \; (X^- \; AND \; R'_A) \; XOR \; (Y^- \; AND \; R_A) \; XOR \; (X^- \; AND \; Y^-) \; XOR \; (R_A \; AND \; R'_B) \; XOR \; (X^- \; AND \; R'_B) \; XOR \; (Y^+ \; AND \; R_B) \; XOR \; (X^- \; AND \; R'_A) \; XOR \; (Y^- \; AND \; R_A) \; XOR \; (X^- \; AND \; Y^-) \; XOR \; (X^- \; AND \; Y^-) \; XOR \; (R_A \; AND \; R'_B) \; XOR \; (R_B \; AND \; R'_A) \; XOR \; R_{PS} \; XOR \; R_{PS}$

Préservation de la vie privée (cont.)

- Propriété du XOR :

- R *XOR* R = 0

Préservation de la vie privée (cont.)

- $A^R \; XOR \; B^R = (X^+ \; AND \; R'_B) \; XOR \; (Y^+ \; AND \; R_B) \; XOR \; (X^+ \; AND \; Y^+) \; XOR \; (R_B \; AND \; R'_A) \; XOR \; (X^- \; AND \; R'_A) \; XOR \; (Y^- \; AND \; R_A) \; XOR \; (X^- \; AND \; Y^-) \; XOR \; (R_A \; AND \; R'_B) \; XOR \; (X^- \; AND \; R'_B) \; XOR \; (Y^+ \; AND \; R_B) \; XOR \; (X^- \; AND \; R'_A) \; XOR \; (Y^- \; AND \; R_A) \; XOR \; (X^- \; AND \; Y^-) \; XOR \; (X^- \; AND \; Y^-) \; XOR \; (R_A \; AND \; R'_B) \; XOR \; (R_B \; AND \; R'_A) \; XOR \; R_{PS} \; XOR \; R_{PS}$

Préservation de la vie privée (cont.)

- Le résultat final :

$$A^R \; XOR \; B^R = (X^+ XOR \; X^-) \; AND \; (Y^+ XOR \; Y^-)$$

Plan

- Protection de la vie privée ?
- K-anonymisation
- Fouille de données et Vie privée : un exemple
- Vers des fonctions d'oubli
- Challenges

La tendance

- Raccourcissement de la durée de conservation des données personnelles (directive Européenne du 15 mars 2006)
- Conservation de 6 à 24 mois
- En 2008, le G29 (homologues européens de la CNIL) : préconisation 6 mois
- Microsoft (Adresses IP de Bing) : 18->6 mois Mais pas les cookies
- Google : 9 mois (cookies mais adresses IP partiellement effacées
- Yahoo! et AOL : 13 mois (cookies, partie de l'adresse IP, navigations)

Et pourquoi pas des données agrégées ?

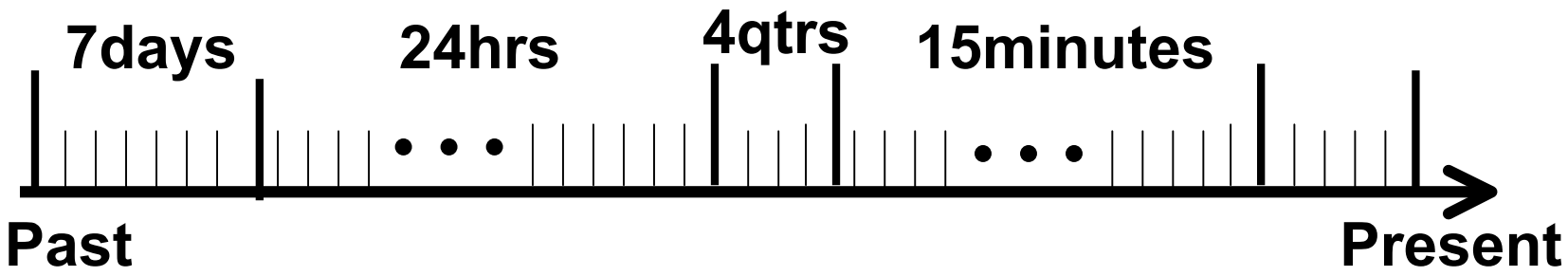
- Effacer les données
  - un jour :)
  - Une estampille temporelle associée aux données (e.g. données de logs)
- Effacer les données : gain important pour la préservation de la vie privée
- Mais perte d'informations intéressantes

Et pourquoi pas des données agrégées ?

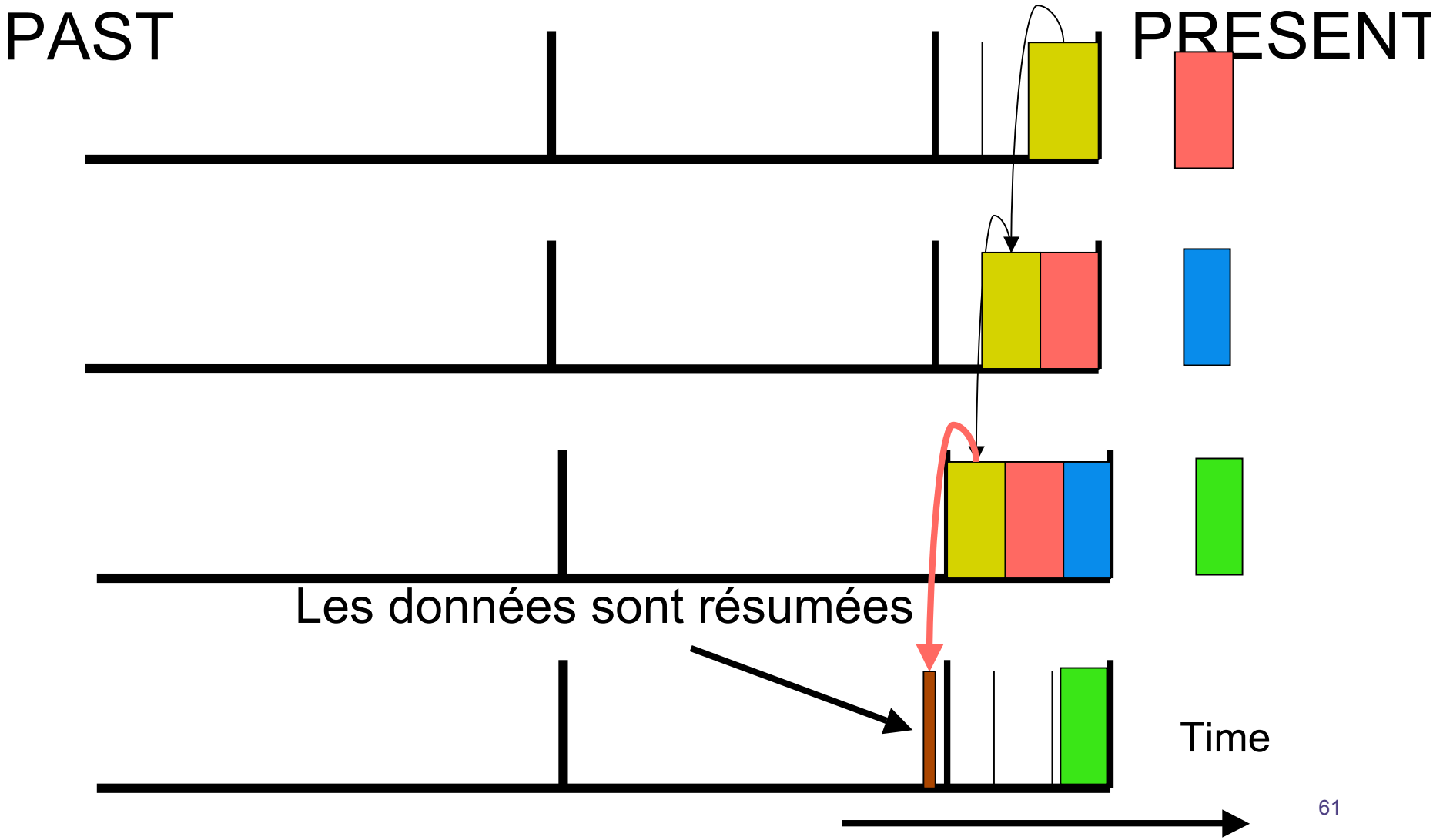
- Il est impossible de sauvegarder tout l'historique aussi à cause d'une capacité de stockage limitée
- Nous avons besoin de compresser les données sans perdre d'information
- Les données récentes sont généralement plus utiles et importantes que les données historiques
- Les gens sont souvent intéressés par les changement récents avec un fin niveau de détail et par les anciens changements avec un niveau de détail plus grossier
  - Problématique similaire aux flots de données !!

Les modèles de tilted time windows

- Les données récentes sont enregistrées et vues avec une plus fine granularité que les anciennes
- Quand la fenêtre de temps est atteinte, les données à fine granularité sont résumées et propagées à une granularité moins fine.
- Les fenêtres sont maintenues automatiquement



Tilted Time Windows : Principe



Plan

- Protection de la vie privée ?
- K-anonymisation
- Fouille de données et Vie privée : un exemple
- Vers des fonctions d'oubli
- Quelques challenges

Recherche : de nombreux défis à relever

- Données textuelles : aucunes méthodes de modification ne marchent
- Données mobiles
  - Des données issues de capteurs (eg. GPS, puces RFID)
  - Foursquare : « *dire où l'on se trouve chaque fois que l'on se déplace* »
  - Projet Européen GeoPKDD : modifier les données de manière à ce que chaque trajectoire ne soit pas distinguable de k autres trajectoires

De nouveaux défis sociétaux

- Données médicales
  - Forte demande de eSanté
  - Les données sortent des hôpitaux
  - Conséquences ?
- Problème sociétal
  - Est il encore possible de préserver sa vie privée ?
  - Pessimiste : trop de facilité pour récupérer les données - nous donnons nos propres données
  - Optimiste : stocker l'information longtemps coûte cher