Privacy Breaches in Privacy-Preserving Data Mining

Johannes Gehrke
Department of Computer Science
Cornell University

Joint work with Sasha Evfimievski (Cornell), Ramakrishnan Srikant (IBM), and Rakesh Agrawal (IBM)

Motivation: Information Spheres

Local information sphere

- Within each organization
- Continuously process distributed high-speed distributed data streams
- Online evaluation of thousands of triggers
- Storage/archival, data provenance of all data is important
- One view: The "real-time" enterprise

Global information sphere

- Between organizations
- Share data in a privacy-preserving way

Global Information Sphere

Distributed privacy-preserving information integration and mining

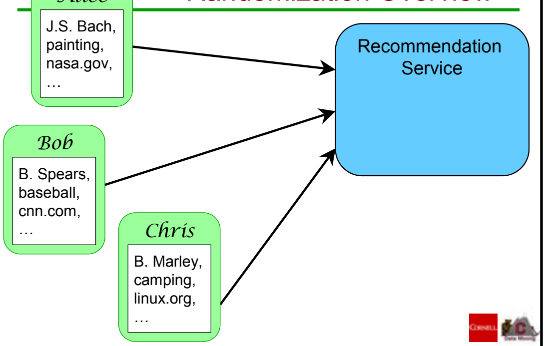
Technical challenges:

- Collaboration of different distributed parties without revealing private data

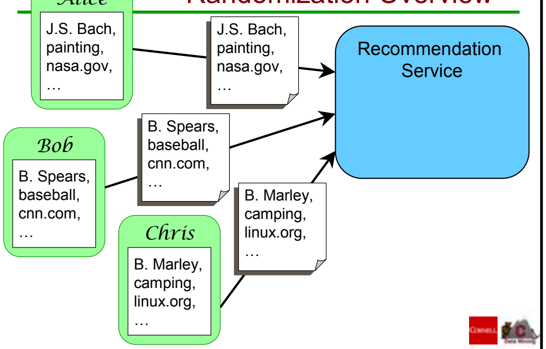
Data Mining and Privacy

- The primary task in data mining: Develop models about aggregated data.
- Can we develop accurate models without access to precise information in individual data records?

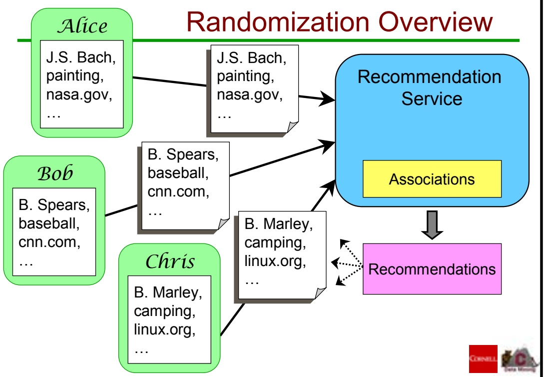
Randomization Overview



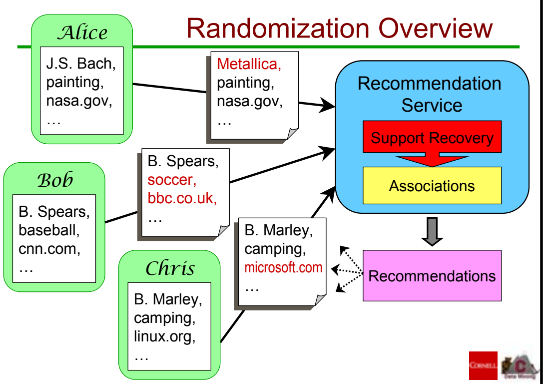
Randomization Overview



Randomization Overview



Randomization Overview



Associations Recap

- A transaction t is a set of items (e.g. books)
- All transactions form a set T of transactions
- Any itemset A has support s in T if

$$s = \text{supp}(A) = \frac{\#\{t \in T \mid A \subseteq t\}}{|T|}$$

- Itemset A is frequent if $s \geq s_{min}$
- If $A \subseteq B$, then supp (A) $\geq$ supp (B).

Associations Recap

- A transaction t is a set of items (e.g. books)
- All transactions form a set T of transactions
- Any itemset A has support s in T if

$$s = \text{supp}\,(A) = \frac{\#\{t \in T \mid A \subseteq t\}}{|T|}$$

- Itemset A is frequent if $s \geq$ smin
- If $A \subseteq B$, then supp $(A) \geq$ supp (B).
- Example:
  - 20% transactions contain X,
  - 5% transactions contain X and Y;
  - Then: confidence of "$X \Rightarrow Y$" is 5/20 = 0.25 = 25%.

The Problem

- How to randomize transactions so that
  - we can find frequent itemsets
  - while preserving privacy at transaction level?

Talk Outline

- Problem Definition
- Uniform Randomization and Privacy Breaches
- Cut-and-Paste Randomization
- Experimental Evaluation
- Generalized Privacy Breaches

Uniform Randomization

- Given a transaction,
  - keep item with 20% probability,
  - replace with a new random item with 80% probability.

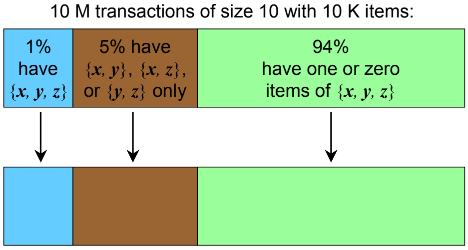
Example: {x, y, z}

10 M transactions of size 10 with 10 K items:

1% have $\{x, y, z\}$	5% have $\{x, y\}$, $\{x, z\}$, or $\{y, z\}$ only	94% have one or zero items of $\{x, y, z\}$

Example: {x, y, z}

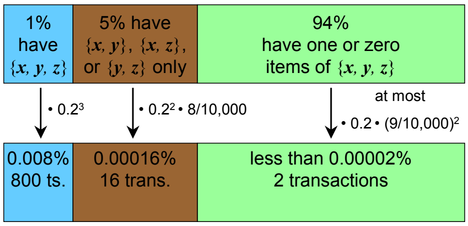
10 M transactions of size 10 with 10 K items:



Uniform randomization: How many have $\{x, y, z\}$?

Example: {x, y, z}

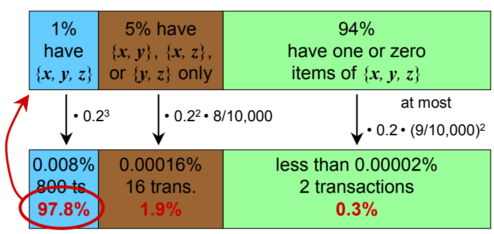
10 M transactions of size 10 with 10 K items:



Uniform randomization: How many have {x, y, z} ?

Example: {x, y, z}

10 M transactions of size 10 with 10 K items:



Uniform randomization: How many have {x, y, z} ?

Example: {x, y, z}

- Given nothing, we have only 1% probability that {x, y, z} occurs in the original transaction
- Given {x, y, z} in the randomized transaction, we have about 98% certainty of {x, y, z} in the original one.
- This is what we call a privacy breach.
- Uniform randomization preserves privacy "on average," but not "in the worst case."

Privacy Breaches

- Suppose:
  - t is an original transaction;
  - t′ is the corresponding randomized transaction;
  - A is a (frequent) itemset.
- Definition: Itemset A causes a privacy breach of level ρ (e.g. 50%) if, for some item z ∈ A,

$$\Pr\left[z \in t \mid A \subseteq t'\right] \geq \rho$$

  - Assumption: no external information besides t′.

Talk Outline

- Problem Definition
- Uniform Randomization and Privacy Breaches
- Cut-and-Paste Randomization
- Experimental Evaluation
- Generalized Privacy Breaches

Our Solution

"Where does a wise man hide a leaf? In the forest.
But what does he do if there is no forest?"
"He grows a forest to hide it in."

G.K. Chesterton

- Insert many false items into each transaction
- Hide true itemsets among false ones
- Can we still find frequent itemsets while having sufficient privacy?

Definition of cut-and-paste

- Given transaction t of size m, construct t':

t = $a, b, c, u, v, w, x, y, z$

t' =

Definition of cut-and-paste

- Given transaction t of size m, construct t':
  - Choose a number j between 0 and K_m (cutoff);

t = $a, b, c, u, v, w, x, y, z$

t' = $\boxed{}$

$j = 4$

Definition of cut-and-paste

- Given transaction t of size m, construct t':
  - Choose a number j between 0 and K_m (cutoff);
  - Include j items of t into t';

t = $a, b, c, u, v, w, x, y, z$

t' = b, v, x, z

$j = 4$

Definition of cut-and-paste

- Given transaction t of size m, construct t':
  - Choose a number j between 0 and K_m (cutoff);
  - Include j items of t into t';
  - Each other item is included into t' with probability p_m.

The choice of K_m and p_m is based on the desired level of privacy.

t = *a, b, c, u, v, w, x, y, z*

t' = *b, v, x, z* | *d, e, g, h, l, m, n, p, s, …*

$j = 4$

Partial Supports

To recover original support of an itemset, we need randomized supports of its subsets.

- Given an itemset A of size k and transaction size m,
- A vector of <u>partial supports</u> of A is

$$\vec{s} = (s_0, s_1, \ldots, s_k), \text{ where}$$

$$s_l = \frac{1}{|T|} \cdot \#\{t \in T \mid \#(t \cap A) = l\}$$

  - Here s_k is the same as the support of A.
  - Randomized partial supports are denoted by $\vec{s}\,'$.

Transition Matrix

- Let $k = |A|$, $m = |t|$.
- <u>Transition matrix</u> $P = P(k, m)$ connects randomized partial supports with original ones:

$$\mathrm{E}\,\vec{s}\,' = P \cdot \vec{s}, \text{ where}$$

$$P_{l',l} = \Pr\left[\#(t' \cap A) = l' \mid \#(t \cap A) = l\right]$$

- Randomized supports are distributed as a sum of multinomial distributions.

The Unbiased Estimators

- Given randomized partial supports, we can estimate original partial supports:

$$\vec{s}_{\text{est}} = Q \cdot \vec{s}\,', \text{ where } \quad Q = P^{-1}$$

- Covariance matrix for this estimator:

$$\text{Cov}\, \vec{s}_{\text{est}} = \frac{1}{|T|} \sum_{l=0}^{k} s_l \cdot Q\, D[l]\, Q^T,$$

$$\text{where } \quad D[l]_{i,j} = P_{i,l} \cdot \delta_{i=j} - P_{i,l} \cdot P_{j,l}$$

- To estimate it, substitute s_l with $(s_{\text{est}})_l$.
  - Special case: estimators for support and its variance

Class of Randomizations

- Our analysis works for any randomization that satisfies two properties:
  - A per-transaction randomization applies the same procedure to each transaction, using no information about other transactions;
  - An item-invariant randomization does not depend on any ordering or naming of items.
- Both uniform and cut-and-paste randomizations satisfy these two properties.

Apriori

Let k = 1, candidate sets = all 1-itemsets.

Repeat:

1. Count support for all candidate sets
2. Output the candidate sets with support ≥ s_{min}
3. New candidate sets = all (k + 1)-itemsets s.t. all their k-subsets are candidate sets with support ≥ smin
4. Let k = k + 1

Stop when there are no more candidate sets.

The Modified Apriori

Let $k = 1$, candidate sets = all 1-itemsets.

Repeat:

1. Estimate support and variance (σ^2) for all candidate sets
2. Output the candidate sets with support $\geq s_{\min}$
3. New candidate sets = all $(k + 1)$-itemsets s.t. all their k-subsets are candidate sets with support $\geq s_{\min} - \sigma$
4. Let $k = k + 1$

Stop when there are no more candidate sets, or the estimator's precision becomes unsatisfactory.

Privacy Breach Analysis

- How many added items are enough to protect privacy?
  - Have to satisfy $\Pr[z \in t \mid A \subseteq t'] < \rho$ ($\Leftrightarrow$ no privacy breaches)
  - Select parameters so that it holds for all itemsets.
  - Use formula $s_l^+ = \Pr[\#(t \cap A) = l, z \in t]$, $s_0^+ = 0$, k=|A|, $P_{l',l} = \Pr[\#(t' \cap A) = l' \mid \#(t \cap A) = l]$

$$\Pr[z \in t \mid A \subseteq t'] = \sum_{l=0}^{k} s_l^+ \cdot P_{k,l} \bigg/ \sum_{l=0}^{k} s_l \cdot P_{k,l}$$

- Parameters are to be selected in advance!
  - Construct a privacy-challenging test: an itemset such that all subsets have maximum possible support.
  - Need to know maximal support of an itemset for each size.

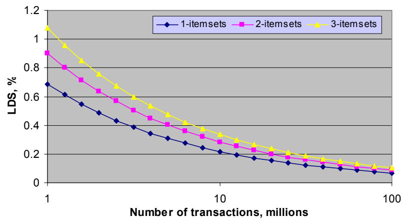
Pros and Cons

- Strength: Graceful tradeoff between precision and privacy
  - Adjust privacy breach level
  - A small relaxation of privacy restrictions results in a small increase in precision of estimators.
- Weakness: No firm guarantee against breaches
  - Is the "privacy-challenging test" challenging enough?
  - Solution: Amplification.
- Weakness: We still need to know something about the prior distribution
  - The definition of breaches needs adjustment
  - Solution: Amplification.
- Weakness: The server has to do a lot more work
  - Can we compress long transactions?
  - Solution: Use error-correcting codes

Lowest Discoverable Support

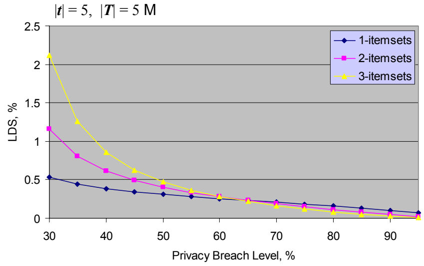
- LDS is s.t., when predicted, it is 4σ away from zero.
- Roughly, LDS is proportional to $1/\sqrt{|T|}$

$|t| = 5,\ \rho = 50\%$



LDS vs. Breach Level

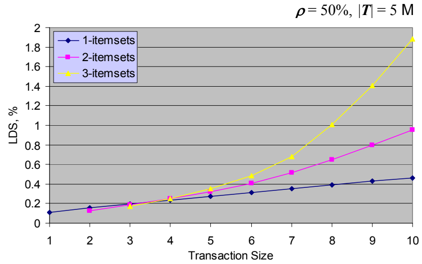
$|t| = 5,\ |T| = 5$ M



- Reminder: breach level is the limit on $\Pr[z \in t \mid A \subseteq t']$

LDS vs. Transaction Size

$\rho = 50\%,\ |T| = 5$ M



- Very long transactions cannot be used for prediction

Talk Outline

- Problem Definition
- Uniform Randomization and Privacy Breaches
- Cut-and-Paste Randomization
- Experimental Evaluation
- Generalized Privacy Breaches

Real datasets: soccer, mailorder

- Soccer is the clickstream log of WorldCup'98 web site, split into sessions of HTML requests.
  - 11 K items (HTMLs), 6.5 M transactions
  - Available at http://www.acm.org/sigcomm/ITA/
- Mailorder is a purchase dataset from a certain on-line store
  - Products are replaced with their categories
  - 96 items (categories), 2.9 M transactions

A small fraction of transactions are discarded as too long.

- longer than 10 (for soccer) or 7 (for mailorder)

Modified Apriori on Real Data

Breach level = 50%. Inserted 20-50% items to each transaction.

Soccer:

$s_{\min} = 0.2\%$

$\sigma \approx 0.07\%$ for 3-itemsets

Itemset Size	True Itemsets	True Positives	False Drops	False Positives
1	266	254	12	31
2	217	195	22	45
3	48	43	5	26

Mailorder:

$s_{\min} = 0.2\%$

$\sigma \approx 0.05\%$ for 3-itemsets

Itemset Size	True Itemsets	True Positives	False Drops	False Positives
1	65	65	0	0
2	228	212	16	28
3	22	18	4	5

False Drops False Positives

Soccer

Pred. supp%, when true supp ≥ 0.2%

Size	< 0.1	0.1-0.15	0.15-0.2	≥0.2
1	0	2	10	254
2	0	5	17	195
3	0	1	4	43

True supp%, when pred. supp ≥ 0.2%

Size	< 0.1	0.1-0.15	0.15-0.2	≥0.2
1	0	7	24	254
2	7	10	28	195
3	5	13	8	43

Mailorder

Pred. supp%, when true supp ≥ 0.2%

Size	< 0.1	0.1-0.15	0.15-0.2	≥0.2
1	0	0	0	65
2	0	1	15	212
3	0	1	3	18

True supp%, when pred. supp ≥ 0.2%

Size	< 0.1	0.1-0.15	0.15-0.2	≥0.2
1	0	0	0	65
2	0	0	28	212
3	1	2	2	18

Actual Privacy Breaches

- Verified actual privacy breach levels
- The breach probabilities are counted in the datasets for frequent and near-frequent itemsets.
- If maximum supports were estimated correctly, even worst-case breach levels fluctuated around 50%
  - At most 53.2% for soccer,
  - At most 55.4% for mailorder.

Talk Outline

- Problem Definition
- Uniform Randomization and Privacy Breaches
- Cut-and-Paste Randomization
- Experimental Evaluation
- General Privacy Breaches

Classes of Privacy Breaches: Example

- Assume that private information is a single item $x \in \{0,\ldots, 1000\}$. Chosen such that
  - P[X=0]=0.01
  - P[X=k]=0.00099, k=1,…,1000
- We would like randomize x by replacing it with y=R(x)
- Three example randomization operators:
  - R1(x)=x with 20% probability, uniform random choice otherwise
  - R2(x)=x + e (mod 1001), where e chosen uniformly at random in {-100,…,100}
  - R3(x) = R2(x) with 20% probability, uniform random choice otherwise

Example (Contd.)

Given	X=0	X not in {200,…,800}
Nothing	1%	40.5%
R1(x)=0	71.6%	83.0
R2(x)=0	4.8%	100%
R3(x)=0	2.9%	70.8%

Recall:

- R1(x)=x with 20% probability, uniform random choice otherwise
- R2(x)=x + e (mod 1001), where e chosen uniformly at random in {-100,…,100}
- R3(x) = R2(x) with 20% probability, uniform random choice otherwise

Two Kinds of Breaches

- Property $P(t)$ was unlikely, but becomes likely once we see t'
  - Example: $X=0$ was 1% likely, but becomes 71.6% likely given that R1(X)=0.
- Property $P(t)$ was uncertain, but becomes virtually certain once we see t'
  - Example: $X \notin \{200,\ldots,1000\}$ was 40.5% likely, but becomes 100% likely given that R2(X)=0.
  - Can think of it inversely: $X \in \{200,\ldots,1000\}$ was 59.5% likely, but becomes only 0% likely given that R2(X)=0.

Definition of General Breach

- Suppose we randomize $t \sim \tau$ into $R(t) = t'$, $0 < \rho_1 << \rho_2 < 1$ are two probabilities;
- We say that there is an upward (straight) privacy breach from ρ_1 to ρ_2 if, for some property $P(t)$,

$$\Pr[P(t)] \le \rho_1, \quad \Pr[P(t) \mid t'] \ge \rho_2$$

- We say that there is a downward (inverse) privacy breach from ρ_2 to ρ_1 if, for some property $P(t)$,

$$\Pr[P(t)] \ge \rho_2, \quad \Pr[P(t) \mid t'] \le \rho_1$$

- For instance, we may have $\rho_1 = 5\%$ and $\rho_2 = 50\%$.

Limiting General Breaches

Suppose that $\rho_2 = \gamma \cdot \rho_1$.

- To prevent all possible upward breaches, it is sufficient to have

$$\forall t, \forall t' : \frac{\Pr[t \mid R(t) = t']}{\Pr[t]} \le \gamma$$

- To prevent all possible downward breaches, it is sufficient to have

$$\forall t, \forall t' : \frac{1}{\gamma} \le \frac{\Pr[t \mid R(t) = t']}{\Pr[t]}$$

- We call a privacy breach that violates one of the above a γ-privacy breach.

Limiting General Breaches (Contd.)

- Thus to prevent all possible γ-privacy breaches, we need to have

$$\forall t, \forall t' : \frac{1}{\gamma} \le \frac{\Pr[t \mid R(t) = t']}{\Pr[t]} \le \gamma$$

Amplification

- Inequality $\forall t, \forall t': \frac{1}{\gamma} \leq \frac{\Pr[t \mid R(t)=t']}{\Pr[t]} \leq \gamma$ sounds good, but…
  - There are way too many possibilities for t to check.
  - We do not know $\Pr[t]$ in advance! What to do?
- Amplification Theorem:
  Revealing R(t) will cause neither an upward nor downward γ-privacy breach if the following condition is satisfied:

$$\frac{\rho 2}{\rho 1} \cdot \frac{1-\rho 1}{1-\rho 2} \leq \gamma$$

Summary

- Privacy breaches: Provided a solution for controlling general breaches
- Algorithm for discovering associations in randomized data
- Validated on real-life datasets
- Can find associations while preserving privacy at the level of individual transactions
- Opens lots of interesting issues.

Ongoing Work and Open Problems

Ongoing work:

- Compression of long transactions
- More sophisticated notions of privacy
- Other data mining models
- Privacy-preserving information integration across different relations and organizations
- Usage of cryptographic techniques

Publications in ACM SIGKDD 2002

[ESA+02] A. Evfimievski, R. Srikant, R. Agrawal, and J. Gehrke. Privacy-Preserving Association Rule Mining.

[DG02] A. Dobra and J. Gehrke. Scalable Regression Tree Construction.

[DGS02] S. Ben-David, J. Gehrke, and R. Schuller. Learning From Multiple Heterogeneous Sources.

[AGYF02] J. Ayres, J. Gehrke, T. Yiu, and J. Flannick. SPAM: Mining Sequential Pattern Using Bitmaps.

[BGK+02] C. Bucila, J. Gehrke, D. Kifer, and W. White. DualMiner: A Dual Pruning Algorithm for Mining with Constraints

More work recently accepted at PODS 2003 and SIGMOD 2003.

Questions?

http://www.cs.cornell.edu/johannes