WWW2014勉強会 Session 6 (Security 2) 1本目

Monitoring Web Browsing Behavior with Differential Privacy

Authors: Liyue Fan, Luca Bonomi, Li Xiong and VaidySunderam (Emory University, GA, USA)

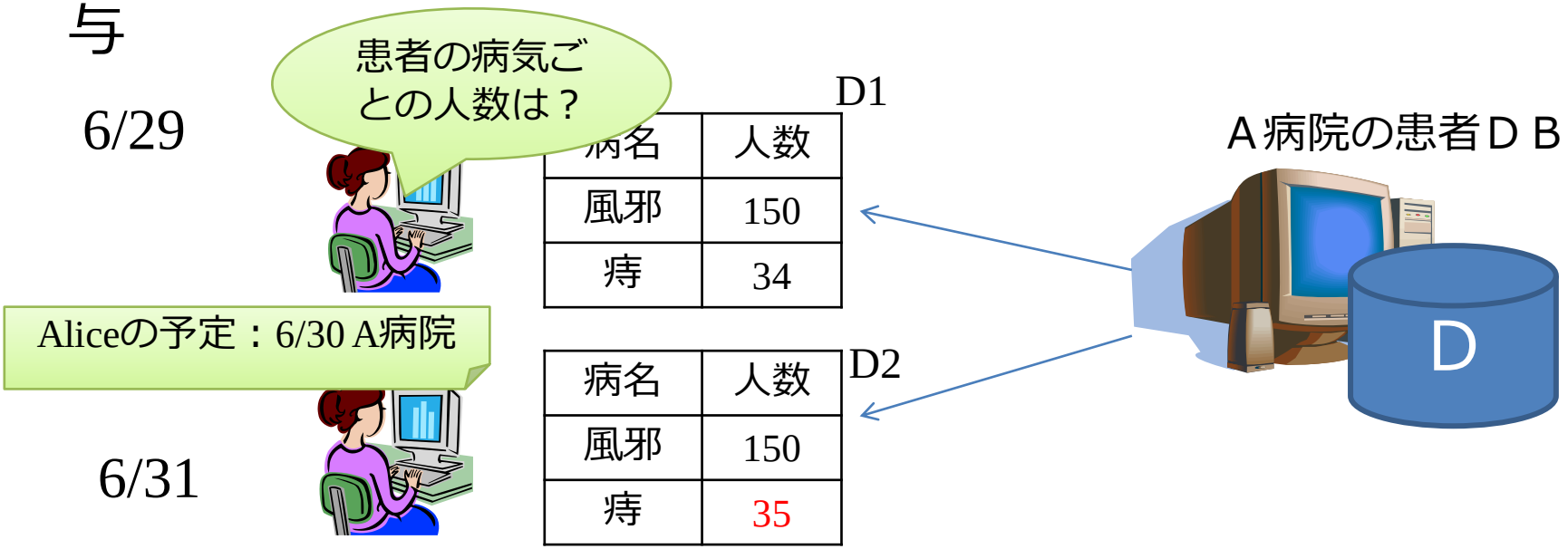
紹介：渡辺知恵美（筑波大）

概要

- **背景**
  - Webページのブラウジング履歴はトレンド分析など有用である一方プライバシに関する懸念がある
  - 事例：AOL data release (2006)
    - 検索履歴から利用者が特定された
- **アプローチ**
  - 各ページの時刻tにおけるアクセス数を集計して出力
  - 集計データに対して差分プライバシを適用
  - 匿名化されたデータから「本来の値」を予測して提供
    - カルファンフィルタを用いた見積もり
    - 匿名性を維持しつつutilityを向上させる

差分プライバシ

一つのデータベースから得られた複数の問合せ結果の差分から個人情報が判明しないように問合せ結果にノイズを付与



$$\Pr[A(D_1) = S] \leq e^{\varepsilon} \cdot \Pr[A(D_2) = S] + \delta$$

ラプラス分布を使ってノイズを付与する

$$\mathrm{A(D)} = g(D) + Lap\left(\frac{\Delta g}{\varepsilon}\right), \Delta g = \max_{D_1, D_2} ||g(D_1) - g(D_2)||$$

global sensitivity

State-Space Model (Univariate Time Series Approace)

- x_k:**時刻kのアクセス数ベクトル**

$$x_k = \{x_k^1, \dots x_k^i, \dots, x_k^m\}$$

$$x_{k+1} = Ax_k + \omega_k \quad (5)$$

$$\omega_k \sim f_\omega(\cdot) \quad (6)$$

- z_k **：DPで匿名化された時刻kの アクセス数ベクトル**

$$z_k = Hx_k + \nu_k \quad (7)$$

$$\nu_k \sim f_\nu(\cdot) \quad (8)$$

$$\nu_k^i \sim Lap(0, \frac{l_{max}}{\alpha}) \quad (12)$$

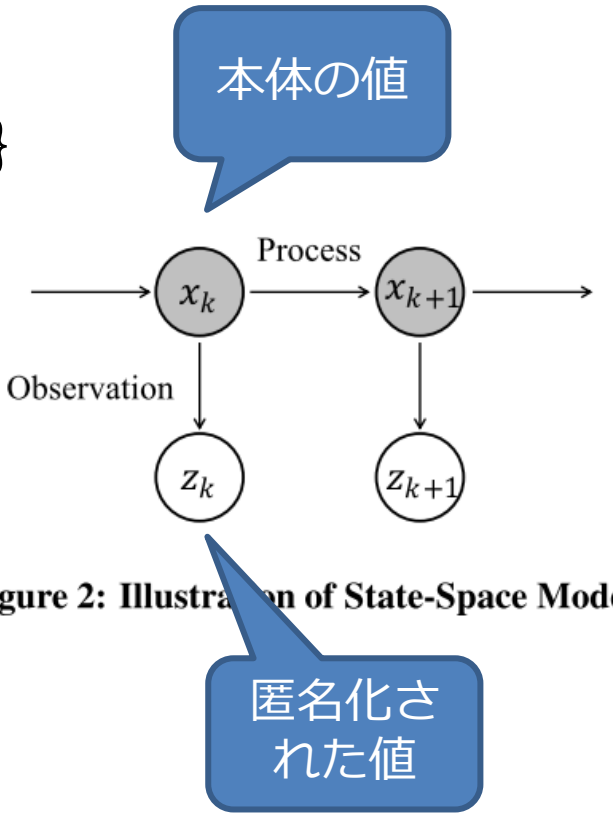


Figure 2: Illustra n of State-Space Model

先行研究：
Uniable Time Series Approach

- **CIKM12で提案した手法**
- **DPで匿名化した後、そのデータを用いて元の値をカルマンフィルタで推測してリリース**

Algorithm 3 Univariate Time-Series Algorithm(k)

Input: Raw counts $\{x_k^i, \text{for } i = 1, \ldots, m\}$, privacy budget α
Output: Private, released counts $\{r_k^i, \text{for } i = 1, \ldots, m\}$

1: **for** $i = 1, \ldots, m$, **do**
2: $prior \leftarrow$ **Prediction**(i, k)
3: $z_k^i \leftarrow$ perturb x_k^i by $Lap(\frac{l_{max}}{\alpha})$
4: $posterior \leftarrow$ **Correction**(i, k)
5: $r_k^i \leftarrow posterior$

事前推測値

Algorithm 1 Prediction(i,k)

Input: Previous posterior estimate x_{k-1}^i
Output: Prior estimate $\hat{x}_k^{i-}$

1: $\hat{x}_k^{i-} = \hat{x}_{k-1}^i$
2: $P_k^{i-} = P_{k-1}^i + Q^i$

ひとつ前の推測値

Algorithm 2 Correction(i,k)

Input: Perturbed count z_k^i
Output: Posterior estimate $\hat{x}_k^i$

1: $K_k^i = P_k^{i-}(P_k^{i-} + R)^{-1}$
2: $\hat{x}_k^i = \hat{x}_k^{i-} + K_k^i(z_k^i - \hat{x}_k^{i-})$
3: $P_k^i = (1 - K_k^i)P_k^{i-}$

匿名化された値

事後推測値

カルマンフィルタによる見積もり

提案手法：
Multivariate Time Series Approach

- **ページiからページjのアクセス遷移を考慮**
- **State-Spaceモデルにマルコフ連鎖を導入**
  - $p_{i,j}$はtraining dataから生成

$$\mathbf{X_{k+1}} = \mathbf{M}\mathbf{X_k} + \omega_\mathbf{k} \quad (17)$$

$$\omega_\mathbf{k} \sim \mathbb{N}(\mathbf{0}, \mathbf{Q}) \quad (18)$$

$$\mathbf{M} = \begin{pmatrix} p_{1,1} & p_{1,2} & \cdots \\ p_{2,1} & p_{2,2} & \cdots \\ \vdots & \vdots & \ddots \end{pmatrix} \quad (19)$$

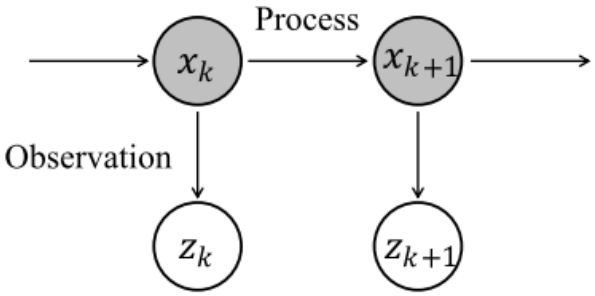


Figure 2: Illustration of State-Space Model

（再掲）

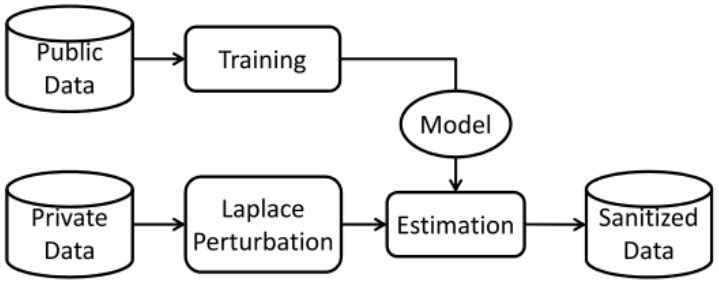


Figure 3: Proposed Framework

Experiments

- **平均相対誤差, top-K mininの精度等を比較**
  - LPA: DPを適用（Baseline）
  - U-KF, M-KF : 提案手法（Univariate, Multivariate）
  - DFT : 既存手法[Rastogi, SIGMOD2010]

M-KFはBaselineより10倍精度が高く既存手法より良い

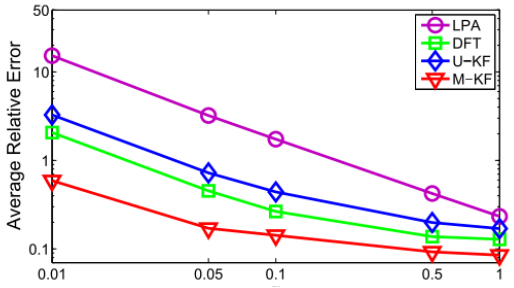


Figure 4: Comparison ol

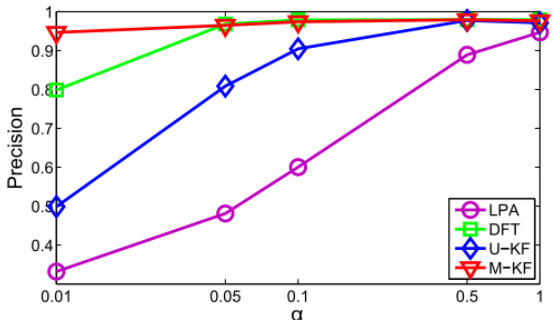


Figure 5: Comparison of top-K mining

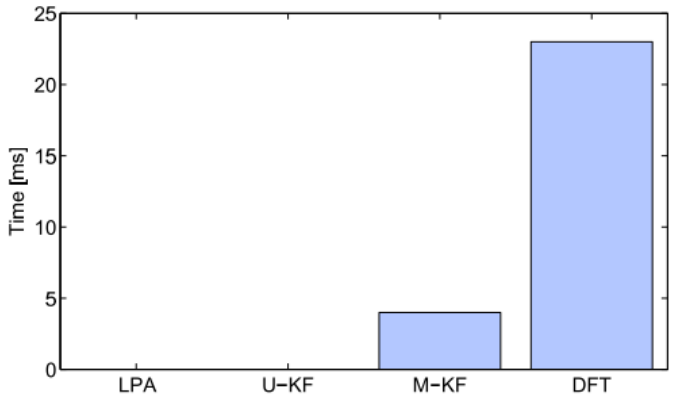


Figure 7: Comparison of runtime performance

M-KFはDFTより実行速度が速い