Discovering Information-Leaking Samples and Features

Hsiang Hsu, Shahab Asoodeh, and Flavio P. Calmon

John A. Paulson School of Engineering and Applied Science, Harvard University

Background

- Given a target set of private attributes, samples and features within a dataset
  $\Rightarrow$ **leak different levels of private information**
  - Not all Tweets equally reveal a users political preference
  - Not all pixels in face images equally disclose emotion
- A natural, yet mostly overlooked, first step in designing **context-aware privacy mechanisms**
  - Information-theoretic privacy, e.g., the privacy funnel
  - Generative Adversarial Privacy (GAP)
- Compared with uniformly adding perturbations $\Rightarrow$ Utility $\uparrow$ and interpretability $\uparrow$
- **Discovering samples or features which leak information about correlated private data**

Information Density

- An information-theoretic quantity $\Rightarrow$ Sample-wise non-linear correlation measure
- Known as Point Mutual Information (PMI) in NLP literature
- Setup:
  - A dataset $\mathcal{D} = \{(\mathbf{s}_n, \mathbf{x}_n)\}_{n=1}^N$, drawn i.i.d. from $P_{S,X}$
  - $\mathbf{s}_n \in \mathcal{S} = \mathbb{R}^m$: the n^{th} private attribute (e.g. binary emotion labels)
  - $\mathbf{x}_n \in \mathcal{X} = \mathbb{R}^k$: data sample (e.g. a face image)
  - $\mathbf{x}_n^j$: the j^{th} feature (i.e., coordinate) of $\mathbf{x}_n$ ($j \in \{1, \dots, k\}$)
- Definition:

  The information density of the n^{th} sample
  $$i(\mathbf{s}_n; \mathbf{x}_n) \triangleq \log \frac{P_{S,X}(\mathbf{s}_n; \mathbf{x}_n)}{P_S(\mathbf{s}_n) P_X(\mathbf{x}_n)} = \log \frac{P_{S|X}(\mathbf{s}_n|\mathbf{x}_n)}{P_S(\mathbf{s}_n)}$$

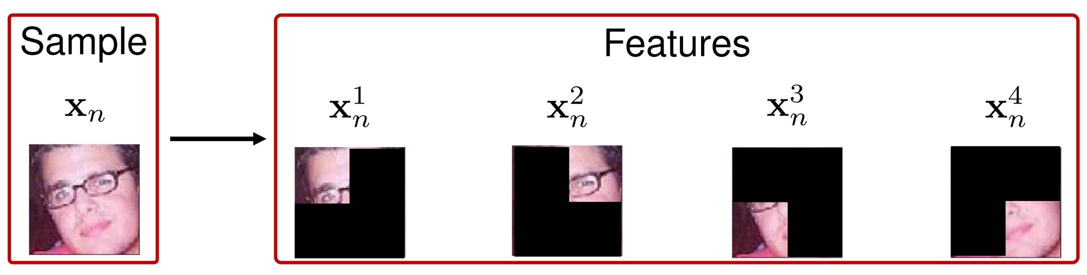
  The information density of the j^{th} feature of the n^{th} sample
  $$i(\mathbf{s}_n; \mathbf{x}_n^j) \triangleq \log \frac{P_{S,X}(\mathbf{s}_n; \mathbf{x}_n^j)}{P_S(\mathbf{s}_n) P_X(\mathbf{x}_n^j)} = \log \frac{P_{S|X}(\mathbf{s}_n|\mathbf{x}_n^j)}{P_S(\mathbf{s}_n)}$$
- $|i(\mathbf{s}_n; \mathbf{x}_n)|$ evaluates the change of belief about $\mathbf{s}_n$ upon observing $\mathbf{x}_n$
  - $\mathbf{s}_n \perp \mathbf{x}_n \Rightarrow P_{S,X}(\mathbf{s}_n; \mathbf{x}_n) \approx P_S(\mathbf{s}_n) P_X(\mathbf{x}_n) \Rightarrow |i(\mathbf{s}_n; \mathbf{x}_n)| \approx 0$
  - $\mathbf{s}_n$ and $\mathbf{x}_n$ are highly correlated $\Rightarrow |i(\mathbf{s}_n; \mathbf{x}_n)|$ bounded away from 0
  - A score for identifying information-leaking samples and features
- Widely used in outlier detection, transfer learning, generative adversarial nets, etc.
- The expected information density is equal to the mutual information, i.e., $\mathbb{E}_{P_{S,X}} i(S; X) = I(S; X)$

Thresholded Information Density Estimator

- $i(\mathbf{s}_n; \mathbf{x}_n)$ is unbounded
  $\Rightarrow$ estimating the information density from samples is hard in sample complexity
- Plug-in estimators perform poorly unless adequate parametric models are assumed (e.g., linear, kernel, or exponential family models)
- No need to precisely estimate information density in our privacy setup
  $\Rightarrow$ Only need to know which samples or features have $|i(\mathbf{s}_n; \mathbf{x}_n)|$ higher than a given threshold ϵ
  $\Rightarrow$ A much easier estimation problem: **thresholded information density estimation**
- Variational representation of f-divergences
  - f: a convex function with $f(1) = 0$, $f^*(t) \triangleq \sup_{x \in \mathbb{R}} \{xt - f(t)\}$: the Fenchel convex conjugate of f
  - $D_f(P\|Q) \triangleq \mathbb{E}_Q f\left(\frac{P}{Q}\right) = \sup_{g: \mathcal{X} \to \mathbb{R}} \mathbb{E}_P[g(X)] - \mathbb{E}_Q[f^*(g(X))] \Rightarrow g^* = \partial f(\frac{P}{Q})$
- Donsker-Varadhan (DV) representation of KL Divergence ($f(t) = t \log t$)
  - $I(S; X) = D(P_{S,X} \| P_S P_X) = \sup_{g: \mathcal{S} \times \mathcal{X} \to \mathbb{R}} \mathbb{E}_{P_{S,X}}[g(S, X)] - \log \mathbb{E}_{P_S P_X}[e^{g(S,X)}] \Rightarrow g^*(s, x) = i(s; x)$
  - Estimating information density is equivalent to solving the functional optimization problem
  - Search space in is unconstrained $\Rightarrow$ unsolvable
- Thresholded Information Density Estimator (TIDE)
  - **Restricted** g to $\mathcal{G}(\Theta)$: continuous functions g_θ $\begin{cases} \text{Bounded by } M \\ \text{Parameterized by } \theta \text{ in a compact domain } \Theta \subset \mathbb{R}^d \end{cases}$
  - TIDE: $\hat{g}_n(s, x) = \operatorname{argmax}_{g_\theta \in \mathcal{G}(\Theta)} \mathbb{E}_{P_{S_n, X_n}}[g_\theta(S, X)] - \log \mathbb{E}_{P_{S_n} P_{X_n}}[e^{g_\theta(S,X)}]$
- Consistency
  - TIDE: extremum estimators of the form $\hat{a} = \operatorname{argmax}_{a \in \mathcal{A}} \Lambda_n(a)$
  - $\Lambda_n(a)$: objective function, $\mathcal{A}$: parameter space
  - **Newey-McFadden Lemma** $\Rightarrow$ Consistency of extremum estimators $\Rightarrow$ Consistency of TIDE
  - (i) compact $\mathcal{A}$ (ii) $\exists \Lambda(a)$ such that $\Lambda_n(a) \xrightarrow{p} \Lambda(a)$ (iii) $\Lambda(a)$ is continuous with unique maximum
- Sample Complexity
  - Assuming g is L-Lipschitz with respect to θ, $|\Theta| \leq C$
  - $n = O(\frac{M^2 d (\log(LC) - \log \eta + M)}{\eta^2}) \Rightarrow$ for all s, x, $\Pr\{|\hat{g}_n(s, x) - g^*(s, x)| \leq \eta\} \geq 1 - e^{-M}$
- Implementation
  - Consider functions representable by a feed-forward deep neural network (clipping outputs to $[-M, M]$)
  - Outputs the thresholded information density of samples $|i(\mathbf{s}_n; \mathbf{x}_n)| \leq M$ and of features $|i(\mathbf{s}_n; \mathbf{x}_n^j)| \leq M$

Experiments and Discussions

- GENKI-$4k$ Smiling Dataset
  - 2400 images for training and 600 for testing
  - $\mathbf{x}_n$: 64×64-pixels face images, $\mathbf{s}_n = \begin{cases} 1 & \text{if similing} \\ 0 & \text{otherwise} \end{cases}$
  - TIDE (VGG-16 CNN) achieves $I(S; X) = 0.594$ bits
  - Image features: 2×2-pixel patches
  - Hide the private information of emotion
  - Preserving other useful information irrelevant of smiling, e.g., gender

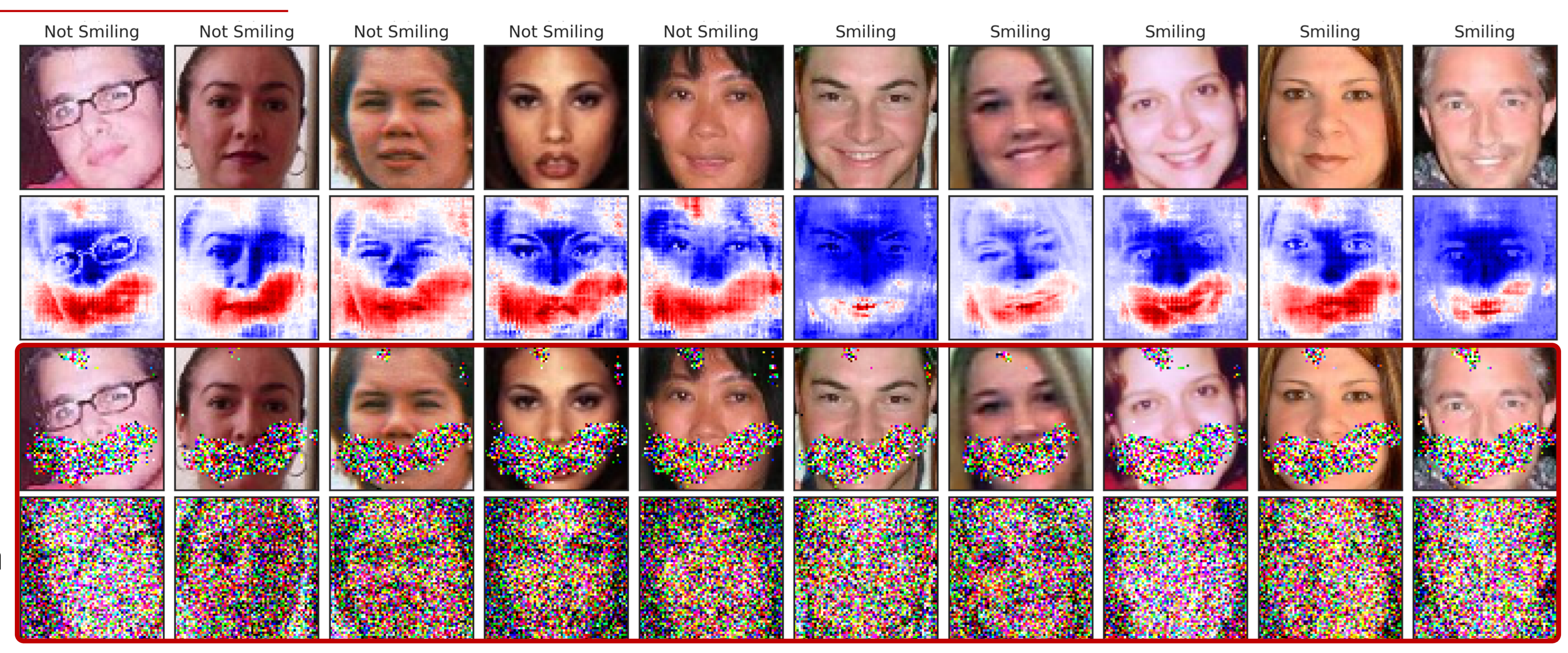


Original Images

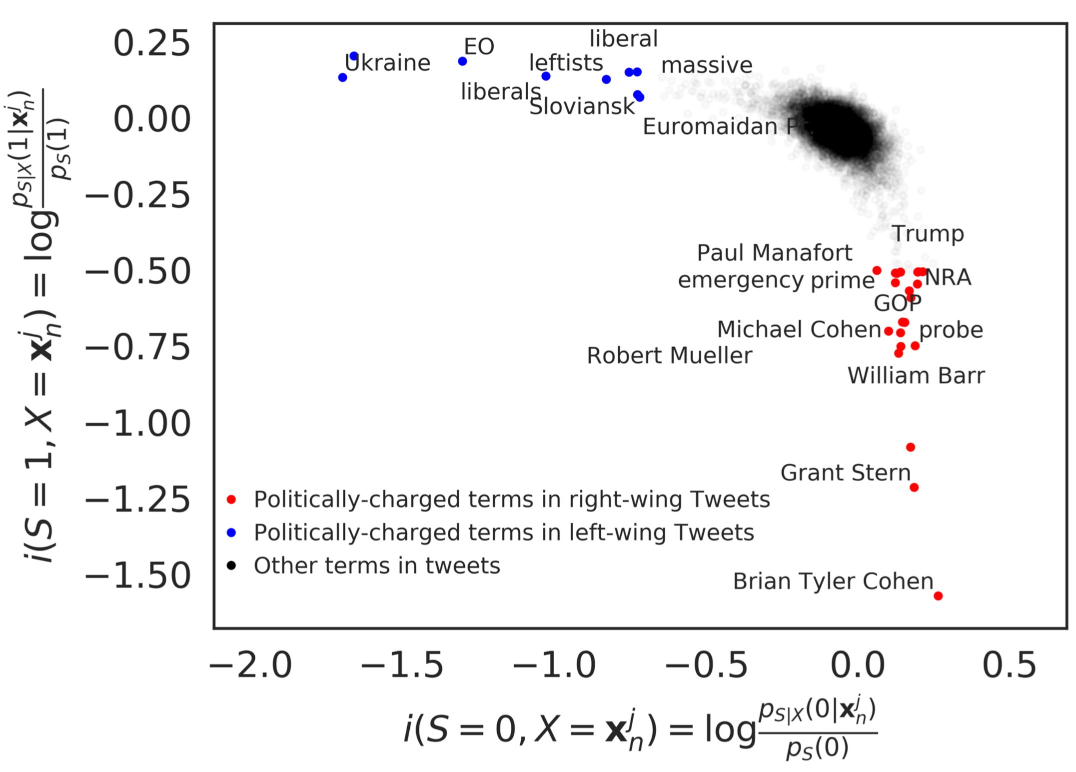
Information-leaking score information density $i(\mathbf{s}_n; \mathbf{x}_n^j)$

Adding local Gaussian noise to patches with $i(\mathbf{s}_n; \mathbf{x}_n^j) \geq 0.9$

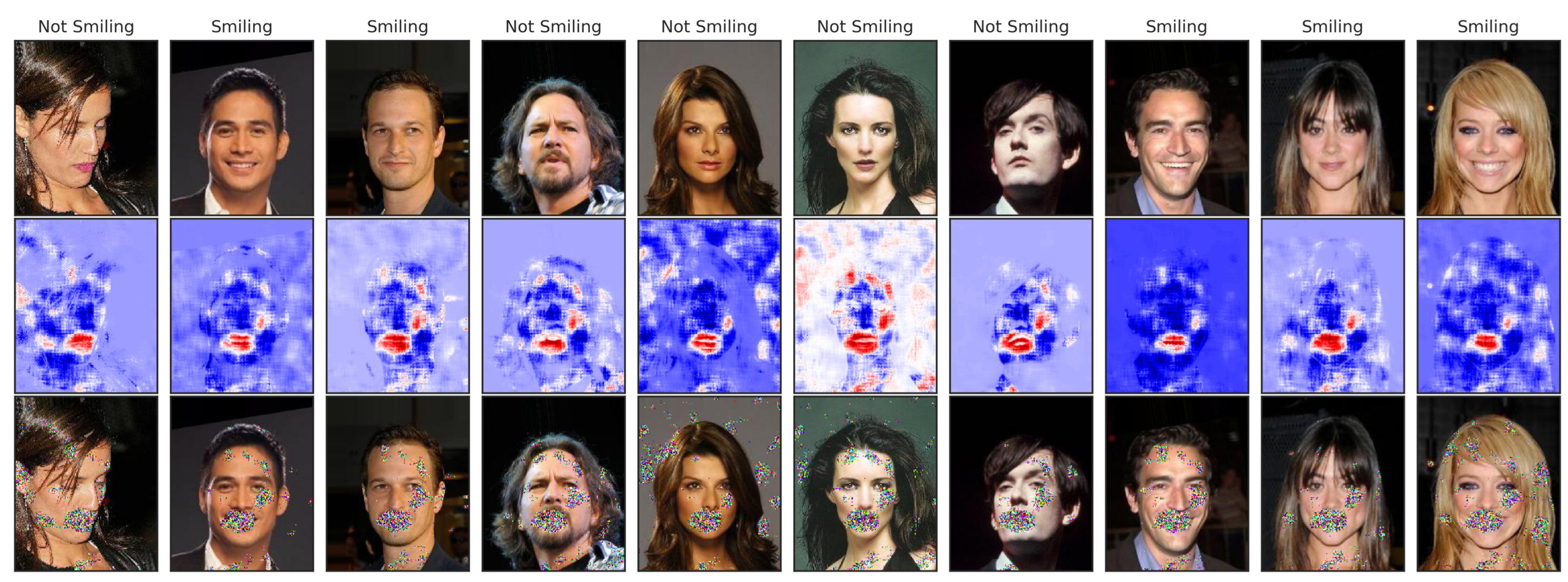
Adding indiscriminate Gaussian noise to the image sample



- Information-Leaking Terms in Tweets
  - $\mathbf{x}_n$: 75946 Tweets from more than 20 online publishers (CNN, Bloomberg, New York Times)
  - $\mathbf{s}_n$: Political preference
    $\mathbf{s}_n = 0$: right-wing
    $\mathbf{s}_n = 1$: left-wing
  - Term frequency and bag-of-words (BoW) model $\Rightarrow$ 24657 terms
  - $I(S; X) = 0.645$ bits
  - Right-wing politics: "Grand Old Party", "National Rifle Association"
  - Left-wing politics: "Europe", "liberal(s)"



- Celebrity Attributes (CelebA) Dataset
  - 202599 high-resolution images
  - $\mathbf{x}_n$: 218×178-pixel celebrity faces
  - $\mathbf{s}_n = \begin{cases} 1 & \text{if similing} \\ 0 & \text{otherwise} \end{cases}$
  - TIDE (VGG-16 CNN) achieves $I(S; X) = 0.967$ bits
  - Image features: 2×2-pixel patches
  - Gaussian noise to patches with $i(\mathbf{s}_n; \mathbf{x}_n^j) \geq 0.74$
  - Perserve other information, e.g., hair color, gender



Selected Reference

- S. Liu, A. Takeda, T. Suzuki, and K. Fukumizu, Trimmed density ratio estimation, in Proc. of Advances in Neural Information Processing Systems (NeurIPS), 2017.
- Y. Polyanskiy, H. V. Poor, and S. Verdú, Channel coding rate in the finite blocklength regime, IEEE Transactions on Information Theory, vol. 56, no. 5, pp. 23072359, 2010.
- M. Sugiyama, T. Suzuki, and T. Kanamori, Density ratio estimation in machine learning. Cambridge University Press, 2012.
- X. Nguyen, M. J. Wainwright, and M. I. Jordan, Estimating divergence functionals and the likelihood ratio by convex risk minimization, IEEE Transactions on Information Theory, vol. 56, no. 11, pp. 58475861, 2010.
- I. Belghazi, S. Rajeswar, A. Baratin, R. D. Hjelm, and A. Courville, MINE: mutual information neural estimation, arXiv preprint arXiv:1801.04062, 2018.
- W. K. Newey and D. McFadden, Large sample estimation and hypothesis testing, Handbook of econometrics, vol. 4, pp. 21112245, 1994.

Remarks

- Limitation - two key assumptions
  - Knowing *a priori* private attributes that we wish to hide (e.g., political preference)
  - A reference dataset from which we can train machine learning models
- Future Directions
  - Privacy-assuring mechanisms beyond the indiscriminate (uniform) addition of noise
  - Optimal perturbations or randomization based on the TIDE

Contact

Hsiang Hsu

Extended Abstract

Extended Paper