Privacy, Data Privacy and Differential Privacy

James Bailie

Statistics Department, Harvard University
jamesbailie@g.harvard.edu

HARVARD

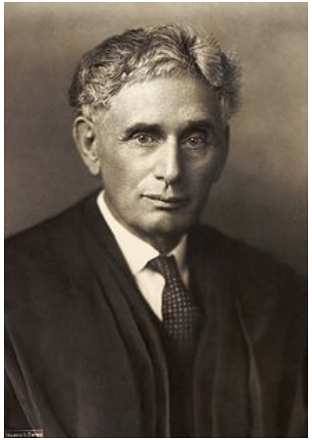
LAW REVIEW.

VOL. IV. DECEMBER 15, 1890. NO. 5.

THE RIGHT TO PRIVACY.

"It could be done only on principles of private justice, moral fitness, and public convenience, which, when applied to a new subject, make

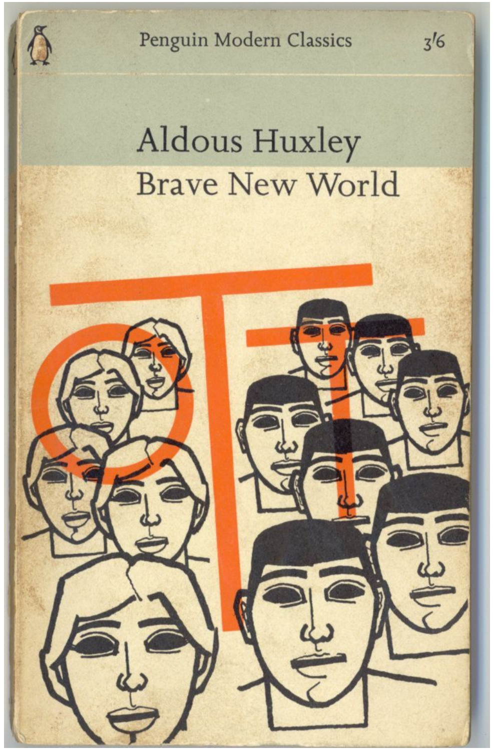
Samuel D. Warren II



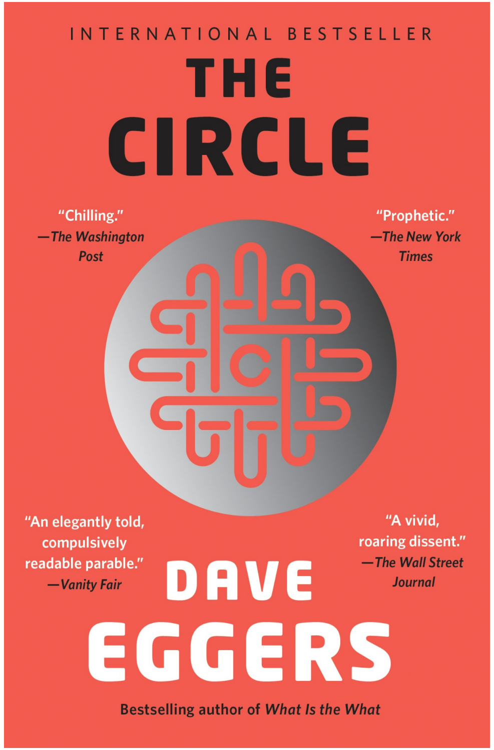
Louis Brandeis

The right to be ***let alone.***

“Nothing was your own except the few cubic centimeters in your skull.”



“I'm claiming the right to be unhappy.”


INTERNATIONAL BESTSELLER
THE CIRCLE
"Chilling."
—The Washington Post
"Prophetic."
—The New York Times
"An elegantly told, compulsively readable parable."
—Vanity Fair
"A vivid, roaring dissent."
—The Wall Street Journal
DAVE EGGERS
Bestselling author of What Is the What


A NETFLIX ORIGINAL SERIES
BLACK MIRROR
JUNE 5 | NETFLIX

Decisional Privacy

Informational Privacy

Decisional Privacy

Autonomy in making personal decisions, particularly regarding their body or actions within their home

Informational Privacy

Decisional Privacy

Autonomy in making personal decisions, particularly regarding their body or actions within their home

Informational Privacy

Control over one's personal information

(aka **Data Privacy**)

Differential Privacy (Dwork et al. 2006)

Differential Privacy (Dwork et al. 2006)

Define the multiplicative distance $\textsc{Mult}(P, Q)$ between two distributions P and Q:

$$\textsc{Mult}(P, Q) = \sup_S \left| \ln \frac{P(S)}{Q(S)} \right|,$$

where $\frac{0}{0} := 1$.
Let $\Pr(\boldsymbol{X}|\boldsymbol{D})$ denote the probability distribution of the output $\boldsymbol{X}$ given the observed data $\boldsymbol{D}$.
For example, $\Pr(\boldsymbol{X}|\boldsymbol{D})$ is a Normal distribution centred at the observed-data estimate $q(\boldsymbol{D})$.
ϵ-differential privacy states that

$$\textsc{Mult}\Big(\Pr(\boldsymbol{X}|\boldsymbol{D}), \Pr(\boldsymbol{X}|\boldsymbol{D}') \Big) \leq \epsilon,$$

for all neighbouring datasets D, D'.

Differential Privacy (Dwork et al. 2006)

Neighbouring datasets usually mean: 1) delete/add one record/individual; 2) alter one record.

ε is called the 'privacy loss budget' but should be 'log of the privacy loss budget'

DP is a bound on the log-likelihood ratio, where the likelihood is of publishing **X** based on **D** versus **D'**.

Differential Privacy (Dwork et al. 2006)

Differential privacy limits the power of an α-level hypothesis test of $H_0 : D_i = s$ versus $H_1 : D_i = t$:

$$\text{Power} \leq \alpha e^{\epsilon}$$

(Wasserman & Zhou, 2010.)

Differential Privacy (Dwork et al. 2006)

Differential privacy limits the power of an α-level hypothesis test of $H_0 : D_i = s$ versus $H_1 : D_i = t$:

$$\text{Power} \leq \alpha e^{\epsilon}$$

(Wasserman & Zhou, 2010.)

Assumes that individual records are independent.

Differential Privacy (Dwork et al. 2006)

Differential privacy limits the power of an α-level hypothesis test of $H_0 : D_i = s$ versus $H_1 : D_i = t$:

$$\text{Power} \leq \alpha e^{\epsilon}$$

(Wasserman & Zhou, 2010.)

Assumes that individual records are independent.

If α = 0.05, non-trivial bound only when $\epsilon \leq \ln 20 \approx 3$.

Bayesian Interpretation

$\boldsymbol{X}$ satisfies ϵ-differential privacy if and only if for all priors π on $\boldsymbol{D}$,

$$\textsc{Mult}\left(\pi(D_i|\boldsymbol{X}, \boldsymbol{D}_{-i}), \pi(D_i|\boldsymbol{D}_{-i})\right) \leq \epsilon.$$

Worst-Case is Not Necessarily the Worst-Case

Consider $\boldsymbol{D} = \{0,1\}^n$ and $X = \sum_i D_i + \frac{1}{\epsilon}L$, where L is Laplace. Let

$$\pi\left(\sum_i D_i = 0\right) = \pi\left(\sum_i D_i = 1\right) = 0.5.$$

Suppose that $X = n+1$. Then

$$\pi\left(D_1 = 0 | X\right) = \pi\left(\sum_i D_i = 0 | X\right) = \frac{\exp\left(-[n+1]/\epsilon\right)}{\exp\left(-[n+1]/\epsilon\right) + \exp\left(-1/\epsilon\right)} \to 0,$$

as $n \to \infty$.

Statistical Data Privacy

*How to release informative statistics **X** without compromising privacy of respondents?*

NSOs have been aware of the privacy risks of publishing aggregate information **X** for at least 50 years (Dalenius, 1977)

Statistical Data Privacy

*How to release informative statistics **X** without compromising privacy of respondents?*

NSOs have been aware of the privacy risks of publishing aggregate information **X** for at least 50 years (Dalenius, 1977)

So why all the talk of differential privacy?

The Changing Privacy Landscape

ARTICLE

Revealing information while preserving privacy

Authors: Irit Dinur, Kobbi Nissim Authors Info & Claims

PODS '03: Proceedings of the twenty-second ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems • June 2003 • Pages 202–210 • https://doi-org.ezp-prod1.hul.harvard.edu/10.1145/773153.773173

Online: 09 June 2003 Publication History

The Changing Privacy Landscape

ARTICLE

Reveal

Authors:

PODS '03: Pr ...ems • June 2003 • Page

Online: 09 Ju

Robust De-anonymization of Large Sparse Datasets

Arvind Narayanan and Vitaly Shmatikov
The University of Texas at Austin

Abstract

We present a new class of statistical de-anonymization attacks against high-dimensional micro-data, such as individual preferences, recommendations, transaction records and so on. Our techniques are robust to perturbation in the data and tolerate some mistakes in the adversary's background knowledge.

We apply our de-anonymization methodology to the Netflix Prize dataset, which contains anonymous movie ratings of 500,000 subscribers of Netflix, the world's largest online movie rental service. We demonstrate

and sparsity. Each record contains many attributes (*i.e.*, columns in a database schema), which can be viewed as dimensions. Sparsity means that for the average record, there are no "similar" records in the multi-dimensional space defined by the attributes. This sparsity is empirically well-established [7, 4, 19] and related to the "fat tail" phenomenon: individual transaction and preference records tend to include statistically rare attributes.

Our contributions. Our first contribution is a formal model for privacy breaches in anonymized micro-data (section 3). We present two definitions, one based on the

The Changing Privacy Landscape

ARTICLE

Reveal

Authors:

PODS '03: Pr

2003 • Page

Online: 09 Ju

461

We prese
anonymizatio
micro-data, s
dations, tran
are robust to
mistakes in th
We apply
Netflix Prize
ratings of 50
largest onlin

practice

DOI:10.1145/3287287

Article development led by acmqueue
queue.acm.org

These attacks on statistical databases are no longer a theoretical danger.

BY SIMSON GARFINKEL, JOHN M. ABOWD, AND CHRISTIAN MARTINDALE

Understanding Database Reconstruction Attacks on Public Data

IN 2020, THE U.S. Census Bureau will conduct the

so the reconstruction no longer results in the original data. This has implications for the 2020 census.

The goal of the census is to count every person once, and only once, and in the correct place. The results are used to fulfill the Constitutional requirement to apportion the seats in the U.S. House of Representatives among the states according to their respective numbers.

In addition to this primary purpose of the decennial census, the U.S. Congress has mandated many other uses for the data. For example, the U.S. Department of Justice uses block-by-block counts by race for enforcing the Voting Rights Act. More generally, the results of the decennial census, combined with other data, are used to help distribute more than $675 billion in federal funds to states and local organizations.

Beyond collecting and distributing data on U.S. citizens, the Census Bureau is also charged with protecting the privacy and confidentiality of survey responses. All census publications must uphold the confidentiality standard specified by Title 13, Section 9 of the U.S. Code, which states that Census Bureau publications are prohibited from identifying "the data furnished by any particular establishment or individual." This section prohibits the Census

y attributes (*i.e.*,
can be viewed as
average record,
ulti-dimensional
parsity is empir-
elated to the "fat
n and preference
attributes.

ition is a formal
ized micro-data
one based on the

ems • June

PDF

The Changing Privacy Landscape

ARTICLE

Rev

Authors

PODS '0

2003 •

Online:

461

• June

PDF

Journal of Privacy and Confidentiality
Vol. 10 (1) 2020

TPDP 2018

Submitted Jan 2019
Published Jan 2020

LINEAR PROGRAM RECONSTRUCTION IN PRACTICE

ALONI COHEN AND KOBBI NISSIM

Boston University
e-mail address: aloni@bu.edu

Department of Computer Science, Georgetown University
e-mail address: kobbi.nissim@georgetown.edu

ABSTRACT. We briefly report on a successful linear program reconstruction attack performed on a production statistical queries system and using a real dataset. The attack was deployed in test environment in the course of the Aircloak Challenge bug bounty

IN 2020, THE U.S. Census Bureau will conduct the

identifying "the data furnished by any particular establishment or individual." This section prohibits the Census

The Changing Privacy Landscape

Journal of
Vol. 10 (1
an 2019
an 2020

ARTICLE

Rev

Authors

PODS '0
2003 •
• June

Online:

461

PDF

Bos
e-m

Dep
e-m

Exposed! A Survey of Attacks on Private Data

Cynthia Dwork,[1] Adam Smith,[2] Thomas Steinke,[3] and Jonathan Ullman[4]

[1]Microsoft Research, Mountain View, California 94043; email: dwork@microsoft.com

[2]Department of Computer Science and Engineering, Pennsylvania State University, State College, Pennsylvania 16802; email: asmith@psu.edu

[3]John A. Paulson School of Engineering and Applied Sciences, Harvard University, Cambridge, Massachusetts 02138; email: tsteinke@seas.harvard.edu

[4]College of Computer and Information Science, Northeastern University, Boston, Massachusetts 02115; email: jullman@ccs.neu.edu

Annu. Rev. Stat. Appl. 2017. 4:61–84

First published online as a Review in Advance on December 21, 2016

Keywords

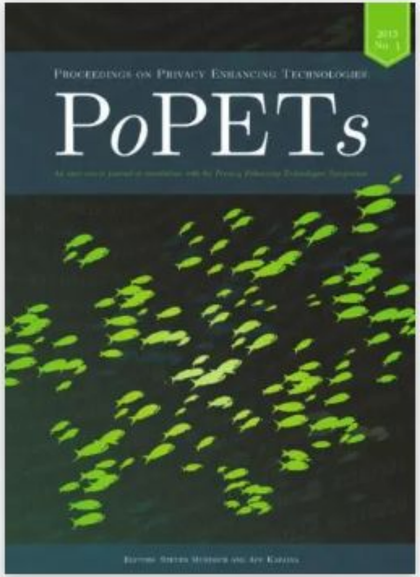
privacy, privacy attacks, re-identification, reconstruction attacks, tracing

IN 2020, THE U.S. Census Bureau will conduct the

identifying "the data furnished by any particular establishment or individual." This section prohibits the Census

The Changing Privacy Landscape

Journal of ... an 2019



Proceedings on Privacy

Open Access

Averaging Attacks on Bounded Noise-based Disclosure Control Algorithms

Hassan Jameel Asghar and Dali Kaafar

Published Online: 08 May 2020

Volume & Issue: Volume 2020 (2020) - Issue 2 (April 2020)

Page range: 358 - 378

Received: 31 Aug 2019

Accepted: 16 Dec 2019

DOI: https://doi.org/10.2478/popets-2020-0031

First published online as a Review in Advance on December 21, 2016

Keywords

privacy, privacy attacks, re-identification, reconstruction attacks, tracing

IN 2020, THE U.S. Census Bureau will conduct the

identifying "the data furnished by any particular establishment or individual." This section prohibits the Census

But…

But…

2001: “advances in technology endanger[ed] our privacy in ways never before imagined” (Garfinkel)

But…

2001: "advances in technology endanger[ed] our privacy in ways never before imagined" (Garfinkel)

1964: privacy was "evaporating [and] under assault from many directions" (Packard) – so much so, in fact, that "we [were standing] on the threshold of what might be called the Age of the Goldfish Bowl" (Brenton)

But…

2001: “advances in technology endanger[ed] our privacy in ways never before imagined” (Garfinkel)

1964: privacy was “evaporating [and] under assault from many directions” (Packard) – so much so, in fact, that “we [were standing] on the threshold of what might be called the Age of the Goldfish Bowl” (Brenton)

1890: “Numerous mechanical devices threaten to make good the prediction that ‘what is whispered in the closet shall be proclaimed from the house-tops’” (Warren & Brandeis)

Some interdisciplinary conundrums

- How do we measure an individual’s valuation of their private information?

Some interdisciplinary conundrums

- How do we measure an individual’s valuation of their private information?
- What are the social norms for privacy?

Some interdisciplinary conundrums

- How do we measure an individual's valuation of their private information?
- What are the social norms for privacy?
- How do we trade off the value of these social norms against the social value of releasing high quality information?

Some interdisciplinary conundrums

- How do we measure an individual's valuation of their private information?
- What are the social norms for privacy?
- How do we trade off the value of these social norms against the social value of releasing high quality information?
- How do we get informed consent? How do we articulate the privacy risks while still emphasising the net positive benefit to responding to a survey?

Some interdisciplinary conundrums

- How do we measure an individual's valuation of their private information?
- What are the social norms for privacy?
- How do we trade off the value of these social norms against the social value of releasing high quality information?
- How do we get informed consent? How do we articulate the privacy risks while still emphasising the net positive benefit to responding to a survey?
- What data is identifying, what data is identifiable and what data constitutes an identity?

Some interdisciplinary conundrums

- How do we measure an individual's valuation of their private information?
- What are the social norms for privacy?
- How do we trade off the value of these social norms against the social value of releasing high quality information?
- How do we get informed consent? How do we articulate the privacy risks while still emphasising the net positive benefit to responding to a survey?
- What data is identifying, what data is identifiable and what data constitutes an identity?
- What data is publicly available and how may this change the privacy assessment?

Contact Details:

James Bailie

Statistics Department, Harvard University

jamesbailie@g.harvard.edu

jameshbailie.github.io/