More privacy to formalize...

Simson L. Garfinkel
May 24, 2017

Privacy Semester at Simons Institute 2019 workshop
Abstract:

Differential privacy provides a formal definition of data privacy within a database, but experience has shown that it's hard to apply differential privacy beyond structured sets of tabular data and some limited graph databases. However, there are many kinds of information that require sharing and computation. Simple datatypes include time, geographical, and imagery information. How do you privatize a picture of a crowd? Today practitioners are at a loss for privatizing even many kinds of structured information, such as 3D models or genetic information. In the cybersecurity world, there is a need to privatize netflow data, cyber threat intelligence, and provenance. And then there's text. Even if the world of tabular databases, we still lack tools for applying differential privacy to high-dimensional data. Differential privacy doesn't seem to have a concept of group privacy. Finally, while differential privacy does give us tools for private data publishing, it is silent on the privacy of data users.

Simson Garfinkel will present a slide for each of these examples, discussing how it would be really neat to privatize this kind of data, but no recommendations on how to addresses these open problems.
"The Best API is no-API"

Differential privacy was created for interactive queries. Data scientists want to work with data.

We need better tools for creating high-dimensionality synthetic data.
Really, we need just one tool.

Any dataset → DP TOOL → DP Synthetic Dataset
Differential Privacy has been (mostly) focused on tabular data.

<table>
<thead>
<tr>
<th>President</th>
<th>Birth</th>
<th>Date of Inauguration</th>
<th>Age at Inauguration</th>
<th>Height (cm)</th>
</tr>
</thead>
<tbody>
<tr>
<td>George Washington</td>
<td>February 22, 1732</td>
<td>April 30, 1789</td>
<td>57 years, 67 days</td>
<td>188 cm</td>
</tr>
<tr>
<td>John Adams</td>
<td>October 30, 1735</td>
<td>March 4, 1797</td>
<td>61 years, 125 days</td>
<td>170 cm</td>
</tr>
<tr>
<td>Thomas Jefferson</td>
<td>April 13, 1743</td>
<td>March 4, 1801</td>
<td>57 years, 325 days</td>
<td>189 cm</td>
</tr>
<tr>
<td>James Madison</td>
<td>March 16, 1751</td>
<td>March 4, 1809</td>
<td>57 years, 353 days</td>
<td>163 cm</td>
</tr>
<tr>
<td>James Monroe</td>
<td>April 28, 1758</td>
<td>March 4, 1817</td>
<td>58 years, 310 days</td>
<td>183 cm</td>
</tr>
<tr>
<td>John Quincy Adams</td>
<td>July 11, 1767</td>
<td>March 4, 1825</td>
<td>57 years, 236 days</td>
<td>171 cm</td>
</tr>
<tr>
<td>Andrew Jackson</td>
<td>March 15, 1767</td>
<td>March 4, 1829</td>
<td>61 years, 354 days</td>
<td>185 cm</td>
</tr>
<tr>
<td>Martin Van Buren</td>
<td>December 5, 1782</td>
<td>March 4, 1837</td>
<td>54 years, 89 days</td>
<td>168 cm</td>
</tr>
</tbody>
</table>
While you are working on differential privacy, the world is pursuing de-identification.

De-identification is safe. De-identification is predictable. De-identification works.
Some existing US laws and regulations recognize/require de-identification

Educational records can be released if de-identified (FERPA).

Medical records can be released if de-identified (HIPAA)

Foodborne Illness Surveillance System allows public release of de-identified aggregate data.

Voluntary safety reports submitted to FAA can be released if the data they contain are de-identified.
De-identification is easy.

name, age, and home state of the child will be stated (e.g., Mike, age 7, Kentucky) unless the parent makes a hardcopy request to have additional information posted. More information about EPA's Children's Privacy Policy.

Introduction
At The Mark Travel Corporation, we believe that it is extremely important to protect the privacy of our customers. We are providing this privacy policy to help you better understand the ways in which your personal information is gathered and used on the www.marktravel.com website.

The Mark Travel Corporation’s affiliate sites gather, store, and process data for marketing purposes and will not sell that data. The Mark Travel Corporation uses customer information to create and distribute products, specials, promotions, and website features for our customers. The Mark Travel Corporation, and their affiliated technology and marketing agencies, and travel agency distributors (hereafter "Mark Travel") will not sell or share your personally identifiable information with any third party company or agency without your permission.

If you have any comments or questions regarding this privacy statement, please contact Mark Travel at 414-228-7472.

Information Mark Travel Collects
Mark Travel website collects customer information in a number of ways: when you sign up to receive emails, when you fill out a Customer Service request, when you send a information request to a travel agent via their agency profile—and when you forward an email to a friend.
More privacy to formalize

1. Understanding de-identification (formally).
   • Drop-in replacement for de-identification.
Imagery...
Multimedia de-identification / redaction is an area of growing concern.

A primary interest is public release of police body cameras:


Other uses:

- Scientific research; privacy preserving surveillance; data retention
Most research has focused on faces and license plates

- Google’s Street View — 90% of faces; 95% of license plates

“Large-scale Privacy Protection in Google Street View,” Frome et al, 2009
De-identifying photographs and video

Key challenges:
• What to remove?
• Usefulness of de-identified imagery
• Evaluation
Face encryption

By using encryption, a distorted video that unauthorised viewers cannot visualise is obtained. Only users who have the proper key for decryption can visualise it. Through the inverse operation and the used key, the ciphered data can be deciphered in order to retrieve the original images. This is named conditional access through encryption. Encryption methods operate over the whole frame or a delimited region of all the video frames. Although such methods do not provide a balance between privacy and intelligibility, they enable to perform data analysis over unprotected data once authorisation and required permissions have been granted. In such cases, privacy would be protected until access to raw data is eventually requested.

Generally naive video encryption algorithms have treated the compressed video bitstream as text data, therefore encrypting the entire video bitstream. Hence, commonly used encryption algorithms such as Data Encryption Standard (DES), Rivest's Cipher (RC5), Advanced Encryption Standard (AES), Rivest, Shamir and Adleman (RSA) and so on, have been used. These algorithms guarantee the highest security level but, unfortunately, they are not suitable for real-time video encryption because they are very time consuming (Yang et al., 2004; Pande & Zambreno, 2013). Due to this, selective encryption algorithms have been proposed (Spanos & Maples, 1995). These algorithms keep using text-based encryption but encrypt only a selected part of the video bitstream so as to get real-time encryption. Other encryption algorithms have also been proposed for real-time encryption, namely light-weight encryption algorithms (Zeng & Lei, 2003). These algorithms are suitable for real-time applications because, when encrypting, they use a simple XOR cipher or only encrypt some bits of the video bitstream. Thereby, they are much faster than the first ones. Finally, there are methods based on scrambling (Tang, 1996). Traditional scrambling methods modify an analogue video signal like those found on closed-circuit television cameras to make it unintelligible. However, with the proliferation of digital video cameras, scrambling techniques are also applied to digital videos in the field of

Figure 4: Two examples of an encrypted image where the face of the person is considered the sensitive region. Reprinted from Boult (2005).
Regarding video inpainting, some of the first straightforward approaches tried to apply image inpainting methods to individual images of the underlying video data. However, they did not take full advantage of the temporal correlation of video sequences. Due to this, the previous methods are often modified in order to be adapted to sequences of images, reconstructing a given frame by interpolating missing parts from adjacent frames. These methods can be classified into patch-based methods (Shiratori et al., 2006; Wexler et al., 2007; Patwardhan et al., 2007; Ghanbari & Soryani, 2011) and object-based methods (Zhang et al., 2005b; Cheung et al., 2006; Vijay Venkatesh et al., 2009). As patch-based methods are unable to perform both spatial and temporal aspects simultaneously, object-based methods were introduced in order to overcome these constraints.

3.4.5. Visual abstraction / object replacement

Object replacement involves the substitution of objects (or persons) appearing in an image or video by a visual abstraction (or visual model) that protects the privacy of an individual while enabling activity awareness. As far as we know, the term 'visual abstraction' was early coined by Chinomi et al. (2008)
Obscuring with synthetic faces: preserves context, prevents automated identification

These techniques can preserve:
- Gender
- Race
- Age

Effectiveness:
- Stops automated face identification.
- Humans can still identify people they know.
Figure 6: An example of a people removal method where the person has been manually selected in the real image (a), and then automatically removed in the second image (b) by filling the region concerning the person using an exemplar-based image inpainting method. Reprinted from Criminisi et al. (2004).

(a) Real image

(b) Modified image
1. Understanding de-identification (formally).
   • Drop-in replacement for de-identification.

2. Imagery
Geospatial information

Everything happens somewhere.

Some locations can be highly identifying
  • A farm house on the prairie.

Some locations are not identifying in 2D but highly identifying in 3D:
  • A 1-bedroom apartment in a high-rise apartment building.

Some locations are identifying if you have temporal information:
  • The speaker’s podium at a public event.

And most locations aren’t identifying at all.
Noise infusion doesn't work for geospatial


More privacy to formalize...

1. Understanding de-identification (formally).
   - Drop-in replacement for de-identification
2. Imagery
3. Geospatial information
### Time, time series and synthetic data

<table>
<thead>
<tr>
<th>President</th>
<th>Birth</th>
<th>Date of Inauguration</th>
<th>Age at Inauguration</th>
<th>Height (cm)</th>
</tr>
</thead>
<tbody>
<tr>
<td>George Washington</td>
<td>February 22, 1732</td>
<td>April 30, 1789</td>
<td>57 years, 67 days</td>
<td>188 cm</td>
</tr>
<tr>
<td>John Adams</td>
<td>October 30, 1735</td>
<td>March 4, 1797</td>
<td>61 years, 125 days</td>
<td>170 cm</td>
</tr>
<tr>
<td>Thomas Jefferson</td>
<td>April 13, 1743</td>
<td>March 4, 1801</td>
<td>57 years, 325 days</td>
<td>189 cm</td>
</tr>
<tr>
<td>James Madison</td>
<td>March 16, 1751</td>
<td>March 4, 1809</td>
<td>57 years, 353 days</td>
<td>163 cm</td>
</tr>
<tr>
<td>James Monroe</td>
<td>April 28, 1758</td>
<td>March 4, 1817</td>
<td>58 years, 310 days</td>
<td>183 cm</td>
</tr>
<tr>
<td>John Quincy Adams</td>
<td>July 11, 1767</td>
<td>March 4, 1825</td>
<td>57 years, 236 days</td>
<td>171 cm</td>
</tr>
<tr>
<td>Andrew Jackson</td>
<td>March 15, 1767</td>
<td>March 4, 1829</td>
<td>61 years, 354 days</td>
<td>185 cm</td>
</tr>
<tr>
<td>Martin Van Buren</td>
<td>December 5, 1782</td>
<td>March 4, 1837</td>
<td>54 years, 89 days</td>
<td>168 cm</td>
</tr>
</tbody>
</table>
How do we create formally private synthetic data with persistent identifiers from year-to-year?
Synthetic datasets ... how do we find them?

Synthetic population housing and person files for the United States
• https://zenodo.org/record/556121
• http://doi.org/10.5281/zenodo.556121

Synthetic Survey of Income and Program Participation:

Synthetic Longitudinal Business Database:
• https://www.census.gov/ces/dataproducts/synlbd/

Virtual RDC@Cornell:
• https://www2.vrdc.cornell.edu/news/synthetic-data-server/
More privacy to formalize

1. Understanding de-identification (formally).
   - Drop-in replacement for de-identification
2. Imagery
3. Geospatial information
4. Time and time Series
Teenager finds sperm donor dad on internet

Ian Sample, science correspondent
@iansample
Wednesday 2 November 2005 20.42 EST

Using nothing more than a swab of saliva and the internet, a 15-year-old boy has tracked down his anonymous sperm donor father, according to details released today.

By sending a swab taken from the inside of his cheek to a genetic testing service, the teenager was able to use genealogy websites to trace his father's lineage for men with a matching Y-chromosome, which is passed down from father to son.
Huntington Disease: A Case Study Describing the Complexities and Nuances of Predictive Testing of Monozygotic Twins

Audrey Heimler¹,³ and Andrea Zanko²

When a candidate for predictive testing for the Huntington disease gene is a monozygotic twin, confidentiality of the co-twin’s diagnosis and autonomy of participation are among the critical genetic counseling issues. Predictive testing can proceed when twins voluntarily and simultaneously request counseling and evaluation in an HD testing program. This case describes a young man referred for predictive testing to an HD testing site on the East Coast of the United States. Family history revealed a twin brother of unknown zygosity who resided on the West Coast of the United States. The genetic counselors on opposite coasts collaborated to provide genetic counseling and evaluation for voluntary, informed predictive testing of the twins, protecting their rights while observing national protocol guidelines.

KEY WORDS: Huntington disease; predictive testing; twins; confidentiality; autonomy.

Monozygotic twins discordant for Huntington disease after 7 years.

Friedman JH, Trieschmann ME, Myers RH, Fernandez HH.

BACKGROUND:
Huntington disease (HD) has only rarely been identified in identical twins. All described twins have had disease onset within 1 year of each other, suggesting that disease onset is determined solely by genetic influences.

OBJECTIVE:
To describe a unique set of monozygotic twins in whom clinical HD onset is at least 7 years apart.

DESIGN:
A 71-year-old woman was diagnosed as having HD based on medical history, physical examination results consistent with HD, and a CAG trinucleotide repeat number of 39 in the HD gene on chromosome 4. Her onset was 6 years earlier. Her genetically confirmed identical twin, carrying the same number of CAG repeats, was neurologically healthy when examined the next year. Only the HD-manifest twin had chronic bronchitis, rheumatoid arthritis, type 2 diabetes mellitus, and chronic anemia. Both had hypertension.

CONCLUSIONS:
To our knowledge, this is the first report of monozygotic twins discordant for HD by more than 2 years. The onset of HD symptoms in a patient with 39 triplet repeats at least 7 years earlier than her identical twin suggests the possibility that the disease may be initiated (or delayed) by environmental factors. We have identified increased cigarette use and longer exposure to various industrial toxins as potential explanations for the earlier onset in one twin.

PMID: 15956172 DOI: 10.1001/archneur.62.6.995

Should All Ashkenazi Women Get Tested for BRCA Gene Mutations?

By Karen Iris Tucker

September 12, 2014  Thinkstock

Is it time to offer BRCA1 and BRCA2 testing to all Jewish women?

K.A. Metcalfe, RN PhD,† A. Eisen, MD,‡§ J. Lerner-Ellis, PhD,† and S.A. Narod, MD†

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4530819/
More privacy to formalize

1. Understanding de-identification (formally).
   - Drop-in replacement for de-identification

2. Imagery

3. Geospatial information

4. Time and time series

5. Genomic information
Challenges:

- Finding the direct identifiers
- Not removing important medical information like eponyms. (e.g. “Addison’s Disease”)

NL Approaches:

- Rule-based (e.g. regex)
- Statistical machine learning.

Several evaluations.

Success rate ≈ 95%
1. Understanding de-identification (formally).
   • Drop-in replacement for de-identification

2. Imagery

3. Geospatial information

4. Time and time series

5. Genomic information

6. Narrative text
Modeling of administrative controls. Formalizing use limitations.
People who model re-identification risk take into account the ability, resources and motivation of the data intruder.

**General public** — anyone who has access to the data.

**Expert** — A computer scientist skilled in re-identification.

**Insider** — A member of the organization that produced the dataset.

**Insider recipient** — A member of the organization that received the data and has more background information than the general public.

**Information broker** — An organization that systematically collects both identified and de-identified information to re-identify.

**Nosy Neighbor** — Friend or family member with specific info.
Cryptography's success required moving beyond perfect secrecy.

- Diffie-Hellman
- RSA
- DES & 3DES
- Certificates & PKI
- PGP
- S/MIME

Key Escrow
Identity-Based Encryption
Password reset by email
Secure web mail
Security questions
More privacy to formalize

1. Understanding de-identification (formally).
   • Drop-in replacement for de-identification
2. Imagery
3. Geospatial information
4. Time and time series
5. Genomic information
6. Narrative text
7. Attacks and controls
TIME'S UP!
More privacy to formalize

1. Understanding de-identification (formally).
   - Drop-in replacement for de-identification
2. Imagery
3. Geospatial information
4. Time and time series
5. Genomic information
6. Narrative text
7. Attacks and controls

Thank you.
- These slides available at: https://simson.net/ref/2017/2017-05-23_Formal_Privacy.pdf