Privacy and Confidentiality: Concepts & Challenges

Shawn Davis
Director of Digital Forensics – Edelson PC
Adjunct Industry Professor – Illinois Tech
Right of an individual to have control over how his or her personal information is collected, used, and/or disclosed
Core Security Principles
No agency shall disclose any record which is contained in a system of records by any means of communication to any person, or to another agency, except pursuant to a written request by, or with the prior written consent of, the individual to whom the record pertains, unless disclosure of the record would be—

(5) to a recipient who has provided the agency with advance adequate written assurance that the record will be used solely as a statistical research or reporting record, and the record is to be transferred in a form that is not individually identifiable;
Confidential Information Protection and Statistical Efficiency Act of 2002 (CIPSEA)

Data or information acquired by an agency under a pledge of confidentiality for exclusively statistical purposes shall not be disclosed by an agency in identifiable form, for any use other than an exclusively statistical purpose, except with the informed consent of the respondent.

(4) The term "identifiable form" means any representation of information that permits the identity of the respondent to whom the information applies to be reasonably inferred by either direct or indirect means.
Recommendations

1. Promote public trust in accuracy of information used to guide government decision-making

2. Establish a new transparency and accountability portal to ensure the public is notified about how confidential data is used

3. Develop a uniform process for external researchers to apply and qualify for secure access
Main Privacy Recommendations

1. Amend Privacy Act and CIPSEA to require departments to conduct risk assessments for public releases of de-identified confidential data.

2. Providing secure and restricted access to confidential data.

3. Adoption of cutting-edge technology for data security, integrity, and confidentiality.
**Personally Identifiable Information (PII)**

- **Direct Identifiers**
  - Name
  - Address
  - SSN
  - Phone Numbers
  - Biometrics
  - Email address
  - Account numbers
  - License numbers
  - Vehicle identifiers
  - Device identifiers
  - IP addresses
  - Photographs
  - URLs
  - Etc.

- **Indirect Identifiers**
  - Location
  - Ethnicity
  - Race
  - Religion
  - Age
  - Zip Code
  - DOB
  - Gender
  - Financial transactions
  - Place of birth
  - Medical information
  - Etc.
Figure 1: The Data Identifiability Spectrum.

“De-identified” Terminology Confusion

Dept of Education: “...removal of all personally identifiable information”

HIPAA: “...no reasonable basis to believe that the information can be used to identify an individual”

CCPA: “...cannot reasonably identify, relate to, describe, be capable of being associated with, or linked, directly or indirectly, to a particular customer”
## Data Category Examples

### Personally Identifiable Data (Absolute or High Risk)

<table>
<thead>
<tr>
<th>Record ID</th>
<th>Name</th>
<th>SSN</th>
<th>DOB</th>
<th>Gender</th>
<th>Income</th>
<th>Job</th>
<th>Height</th>
<th>Zip</th>
<th>Geolocation</th>
</tr>
</thead>
<tbody>
<tr>
<td>12548</td>
<td>Kat Robin</td>
<td>123-45-6789</td>
<td>01/02/60</td>
<td>F</td>
<td>60,000</td>
<td>Teacher</td>
<td>6’0”</td>
<td>60302</td>
<td>41.8869,-87.7801</td>
</tr>
<tr>
<td>12567</td>
<td>Jim Jones</td>
<td>987-65-4321</td>
<td>05/02/72</td>
<td>M</td>
<td>65,000</td>
<td>Teacher</td>
<td>5’11”</td>
<td>60646</td>
<td>41.9945,-87.5845</td>
</tr>
</tbody>
</table>

### Pseudonymized Data (Medium Risk)

<table>
<thead>
<tr>
<th>Record ID</th>
<th>Name</th>
<th>SSN</th>
<th>DOB</th>
<th>Gender</th>
<th>Income</th>
<th>Job</th>
<th>Height</th>
<th>Zip</th>
<th>Geolocation</th>
</tr>
</thead>
<tbody>
<tr>
<td>12548</td>
<td>sadfjkweSdi</td>
<td>29j2k34kjw</td>
<td>01/02/60</td>
<td>M</td>
<td>60,000</td>
<td>Teacher</td>
<td>6’0”</td>
<td>60302</td>
<td>41.8869,-87.7801</td>
</tr>
<tr>
<td>12567</td>
<td>Ekriudlkjwe</td>
<td>09dfgk3453</td>
<td>05/02/72</td>
<td>M</td>
<td>65,000</td>
<td>Teacher</td>
<td>5’11”</td>
<td>60646</td>
<td>41.9945,-87.5845</td>
</tr>
</tbody>
</table>

### De-identified Data (Low Risk)

<table>
<thead>
<tr>
<th>Record ID</th>
<th>Income</th>
<th>Job</th>
<th>Height</th>
</tr>
</thead>
<tbody>
<tr>
<td>12548</td>
<td>60,000</td>
<td>Teacher</td>
<td>6’0”</td>
</tr>
<tr>
<td>12567</td>
<td>65,000</td>
<td>Teacher</td>
<td>5’11”</td>
</tr>
</tbody>
</table>

### Anonymized Data [Summary example] (Zero Risk)

<table>
<thead>
<tr>
<th>Count</th>
<th>Income</th>
<th>Job</th>
<th>Height</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>60-65k</td>
<td>Teacher</td>
<td>5’11” to 6’0”</td>
</tr>
</tbody>
</table>
Re-identification Risks – Combination of Indirect Identifiers

<table>
<thead>
<tr>
<th>Record ID</th>
<th>Name</th>
<th>SSN</th>
<th>DOB</th>
<th>Gender</th>
<th>Income</th>
<th>Job</th>
<th>Height</th>
<th>Zip</th>
<th>Geolocation</th>
</tr>
</thead>
<tbody>
<tr>
<td>12548</td>
<td>sadfjkwESdi</td>
<td>5484548</td>
<td>01/02/60</td>
<td>M</td>
<td>60,000</td>
<td>Teacher</td>
<td>6’0”</td>
<td>60302</td>
<td>41.8869,-87.7801</td>
</tr>
<tr>
<td>Pseudonym</td>
<td>Pseudonym</td>
<td>Indirect</td>
<td>Indirect</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Indirect</td>
<td>Indirect</td>
</tr>
</tbody>
</table>

“It was found that 87% (216 million of 248 million) of the population in the United States had reported characteristics that likely made them unique based only on {5-digit ZIP, gender, date of birth}.”

Re-identification Risks: Linkage Attack Using Indirect Identifiers

MIT Grad Student Latanya Sweeney re-identified MA Gov. Weld from two datasets:

1. Insurance reimbursement records
2. Voter registration list

Other Real World Linkage Attack Examples

Netflix Prize: Researchers matched de-identified Netflix viewing data with IMDB data.

Medical Tests: 5-7 lab results enough to identify a patient from de-identified biomedical research database.

Credit Card Transactions: Uniquely identified 90% of people in de-identified credit card transactions from four distinct transactions in space and time.

Taxi Ride Data: NYC de-identified dataset of 173 million taxi rides and ride timestamps, but left 32-bit code (that could be easily converted to a taxi medallion number). Individuals re-identified from photos of themselves and taxi medallion along with photo timestamp.
## Re-identification Risks - Geolocation

<table>
<thead>
<tr>
<th>Record ID</th>
<th>Name</th>
<th>SSN</th>
<th>DOB</th>
<th>Gender</th>
<th>Income</th>
<th>Job</th>
<th>Height</th>
<th>Zip</th>
<th>Geolocation</th>
</tr>
</thead>
<tbody>
<tr>
<td>12548</td>
<td>sadfjkwESdi</td>
<td>5484548</td>
<td>01/02/60</td>
<td>M</td>
<td>60,000</td>
<td>Teacher</td>
<td>6’0”</td>
<td>60302</td>
<td>41.8869,-87.7801</td>
</tr>
<tr>
<td>Pseudonym</td>
<td>Pseudonym</td>
<td>Indirect</td>
<td>Indirect</td>
<td></td>
<td>Indirect</td>
<td></td>
<td></td>
<td>Indirect</td>
<td></td>
</tr>
</tbody>
</table>

Spokeo. Know More.

- Name: [201 South Blvd, Oak Park, IL](https://www.google.com/maps/place/201+South+Blvd,+Oak+Park,+IL+)
- Email: [ ]
- Phone: [ ]
- Address: [201 South Blvd, Oak Park, IL]

**CURRENT RESIDENTS**

- **PERSON**: [Redacted]
- **LOCATION**: 201 South Blvd
  - Oak Park, IL 60302
NYT 2019 Review of Geolocation DB – 12 Mil Phones

“We were able to track smartphones in nearly every major government building and facility in Washington.”

“We could follow them back to homes and, ultimately, their owners’ true identities.”
# Re-identification Risks – Matching Device Metadata

<table>
<thead>
<tr>
<th>Data Type</th>
<th>Value</th>
<th>Data Type</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Geolocation</td>
<td>41.9.., -88.1..</td>
<td>Geolocation</td>
<td>41.9.., -88.1..</td>
</tr>
<tr>
<td>MAC</td>
<td>43:df:23...</td>
<td>MAC</td>
<td>43:df:23...</td>
</tr>
<tr>
<td>IP Address</td>
<td>74.26...</td>
<td>IP Address</td>
<td>74.26...</td>
</tr>
<tr>
<td>User Agent</td>
<td>Macintosh/ Safari...</td>
<td>User Agent</td>
<td>Macintosh/Safari...</td>
</tr>
</tbody>
</table>

| First Name      | Kat                       | Cookie          | kskjeris3243              |
| Last Name       | Rob                       | Referrers       | nbc.com/news... yelp.com/review... amazon.com/productpage... |
| Address         | 201 S. Blvd               |                 |                           |
| City            | Oak Park                  |                 |                           |
| State           | IL                        |                 |                           |
| Email           | random@gmail.com          |                 |                           |