US perspectives on anonymization: operationalizing privacy

Hannah E. Rakoff
Volpe National Transportation Systems Center
U.S. Department of Transportation
31 August 2016
Objective: describe the current focus on expanding anonymization from an archive to an operationalized (real-time) context

Organization:
- Overview of the ITS JPO Data Program
- Focus of anonymization work:
  - To date: archive
  - Moving forward: archive + operationalizing privacy
    - ITS Privacy Module (algorithm)
    - ITS Privacy Sensitivity Model
- Video anonymization
Volpe, The National Transportation Systems Center

- Part of the U.S. Department of Transportation
- Based in Cambridge, Massachusetts near MIT
- 570 federal employees and 400 on-site contractors
- Fee-for-service; no direct appropriations

Advancing transportation innovation for the public good
Our Portfolio of Work

Volpe works on all modes of transportation.

- Creating aviation systems to alleviate congestion, improve safety, and mitigate environmental impacts
- Promoting crash avoidance and electronic systems safety and resilience
- Developing cybersecurity for the transportation enterprise
- Analyzing logistics and supply chain management
- Internationally recognized human factors research and development capabilities
- Supporting smart investments in the research, planning, development, testing, management, operations, and financing of transportation systems and agencies
- Measuring, analyzing, and modeling data to support energy consumption, air quality, safety, and noise policy
- Evaluating, testing, and engineering infrastructure
- Developing large-scale IT solutions to promote safety and efficiency through inspection and enforcement
Overview of the ITS JPO Data Program
The ITS Joint Program Office (JPO)

- Mission:
  Conduct research, development, and education activities to facilitate the adoption of information and communication technology to enable society to move more safely and efficiently.

- JPO strategic priorities (2015-2019) include
  - Realizing connected vehicle implementation
  - Advancing automation
  - Smart Cities
Data: the foundation of the ITS strategic plan

http://its.dot.gov/strategicplan/index.html
ITS Data Program vision

By 2019, State DOTs and MPOs will have access to low-cost, scalable, interoperable data management tools that can ingest new data sources and feed new applications in ways that protect the privacy of users while enabling on-demand data sharing at regional and national levels. This toolset will be tested and deployed within at least one region and support a third party ecosystem of applications and research.
Data Program approach

- Develop tools
  - Operational Data Environment (ODE)
  - ITS Privacy Module (algorithm)
  - ITS Privacy Sensitivity Model
  - ITS Data Policy Playbook

- Share data
  - Research Data Exchange (RDE)
  - National Strategy

- Enable use by early deployers
  - Connected vehicle (CV) pilots
  - Smart City Challenge
Focus of anonymization work to date: archive data
Anonymization for data archiving

- A focus of connected vehicle anonymization has been on archiving, particularly Basic Safety Messages (BSMs)
  - “Privacy by design,” but not when aggregated
  - Resource-intensive

- Recent milestone in archive data anonymization: posting excerpt of Safety Pilot Model Deployment data on the RDE
Moving forward:
archive + operationalizing privacy
New challenge: share operational (real-time) data that originates at state and local level

• The challenge:
  • How do you operationalize anonymization techniques for real-time data?
  • How do you transfer the capabilities to be administered at local levels (where data are generated)?

• Data Program philosophy:
  • Address technical and policy/institutional issues
  • Priorities and product development driven by user needs – user-centered design
  • Many Program products ultimately handed-off to user communities
ITS Privacy Module (algorithm)

- Initial focus is on BSMs
- Objective is to remove enough data to prevent derivation of a specific traveler’s origins and destination while maintaining the value of the data (trade-off)
- Approach focuses on truncating data (BSMs) at appropriate junction
ITS Privacy Sensitivity Model

- A tool for assessing the trade-off between data utility versus privacy protection
- Users can adjust privacy algorithm parameters and see the implications for the resulting (anonymized) data
ITS Data Policy Playbook

- Online collation of policies, principles and real-world examples of successful policies and practices, e.g.,
  - sample data retention policies
  - open source procurement language
  - guides to conducting privacy impact assessments and staying up to speed on current standards
- Intended audience is state and local agencies
- FY17 project – post initial version to support early deployments (e.g., CV Pilots)
- Hand off long-term maintenance to the deployment community
Video anonymization
Video anonymization

- One important part of USDOT’s overall data anonymization research
- Lack of anonymization limits data sharing
- Recent project – DOT Exploratory Advanced Research (EAR) project
  - Developed an algorithm that performs a one-way merge of the pixels with another face
Video anonymization example

- Automated system with manual QA

Tool is partially developed; application is the next step (pilots expected in 2017).
Contact us!

Ariel Gold, Data Program Manager, ITS Joint Program Office, USDOT
ariel.gold@dot.gov

Hannah Rakoff, Volpe Center, USDOT
hannah.rakoff@dot.gov

Matthew Burt, Volpe Center, USDOT
matthew.burt@dot.gov
Appendix A: the RDE
Research Data Exchange (RDE)

- USDOT has had archives for several years, e.g., RDE – [https://www.its-rde.net/](https://www.its-rde.net/)
- RDE provides a variety of data-related services that support the development, testing, and demonstration of multi-modal transportation mobility, weather, and environmental applications.
- Users can download data and documentation, collaborate with other users, and comment on data sets.
The RDE employs the concept of a Data Environment to structure the various data sets and has data from 7 US states

- RDE organizes data using a data environment / data set / data file hierarchy

- A Data Environment is a collection of data sets which were obtained under the same test / experiment

- Data Sets represent a logical arrangement of files that convey a common or unified concept of the data collection exercise

- Data sets contain Data Files that are archived collection of data (elements) and can be text, zip, binary, or other file types

Source: [http://its-ao.net/](http://its-ao.net/)
- California, Florida, Michigan, Minnesota, Oregon, Virginia, Washington
RDE screenshots
Appendix B: De-identification Algorithm*

*Slides developed by the Center for Trustworthy Embedded Systems, Oak Ridge National Labs
The Challenge

Remove enough personally identifiable information (PII) such that there is *no reasonable basis* to believe that the remaining information can be used to identify an individual.*

* NIST Guide to Protecting the Confidentiality of Personally Identifiable Information (PII)
Terms and Definitions

- **Original Trip** [●—●]: An original trip BSM sequence.
- **Subtrip** [●—●]: A sequence of consecutive trip points.
- **Critical Interval (CI)** [●—●]: A subtrip that *may* provide details useful for re-identification.
- **Privacy Interval (PI)** [●—●]: A subtrip that when removed introduces uncertainty.
  - **Goal**: Identify the collection of PIs that will make re-identification a significant challenge; measureable; *eliminate reasonable basis*
- **De-identified Trip (DI)** [●—●]: Original Trip with CIs and PIs removed
More about the CI and PI

- What is the **purpose of the steps** taken to achieve the above effect?

  - Critical Interval: the critical “thing” we are trying to protect (e.g., a specific location)
    - Removing this interval alone is insufficient.
  - Privacy Interval: the additional geopoints removed to introduce uncertainty.