NCHRP 08-36, Task 130

Inventory and Assessment of Methods for Making Collected Transportation Data Anonymous

Requested by:

American Association of State Highway and Transportation Officials (AASHTO)
Standing Committee on Planning

Prepared by:

William Bachman, Ph.D.
Tom Krenzke, M.S.
Jane Li, Ph.D.

Westat
Rockville, Maryland

Liisa Ecola
The RAND Corporation

September 2016

The information contained in this report was prepared as part of NCHRP Project 08-36, Task 130, National Cooperative Highway Research Program (NCHRP).

AASHTO
THE VOICE OF TRANSPORTATION

ACKNOWLEDGMENT

This study was conducted for the AASHTO Standing Committee on Planning, with funding provided through the National Cooperative Highway Research Program (NCHRP) Project 08-36, Research for the AASHTO Standing Committee on Planning. The NCHRP is supported by annual voluntary contributions from the state Departments of Transportation. Project 08-36 is intended to fund quick response studies on behalf of the Standing Committee on Planning. The report was prepared by William Bachman, Tom Krenzke, and Jane Li, of Westat; under the direction of Liisa Ecola, The RAND Corporation. The project was managed by Lawrence D. Goldstein, NCHRP Senior Program Officer.

DISCLAIMER

Table of Contents

Introduction

Transportation professionals are collecting or acquiring detailed spatial and temporal data from roadside sensors, vehicles, mobile phones, and navigation devices to support the research, analysis, and planning of transport systems. Agencies using this data are legally obligated to provide adequate protection of the rights and freedoms of individuals whose data provide intelligence on travel activity (Privacy Act of 1974). This document provides descriptions of data anonymization methods that offer privacy protection while maximizing the value of the information used in the design and development of efficient transportation systems.

Transportation agencies need data regarding the movement of goods and people to ensure that transportation investment decisions result in efficient, sustainable, and safe travel. The attention on efficiency of the transportation system requires accurate and precise information regarding travel behavior and system performance. Data collection technology advances resulted in opportunities for acquiring detailed travel information that can be used to support these refined data needs. As such, the detailed travel data may contain personally identifiable information (PII), which is protected by federal law as well as by professional and ethical standards. PII is not necessary to support the systems and methods needed for analysis and forecasting, and is simply a by-product of the data collection methods. Models and model-building efforts only quantify an individual's travel-making characteristics to understand travel behavior and build performance assessments of roads, transportation systems, neighborhoods, communities, and regions. The challenge for transportation agencies is finding the most reliable methods for protecting or removing PII while maintaining the value found in detailed travel data. In this report, we are specifically evaluating and describing methods for protecting PII found in data collected from sensor-based systems (Bluetooth®, Wi-Fi, radio frequency identification [RFID], and video-based license plate capture), and mobile devices using global positioning system (GPS) technology.

Most of the current PII protection laws/policies were originally developed to protect human subjects in health research. Over time, these laws/policies have been adopted across most federal agencies including the U.S. Department of Transportation. Given the specific transportation data context of this research, there are two primary protection objectives for implementing anonymization methods:

1. Preventing the identification or re-identification of the 18 "Safe Harbor" identifiers (HIPAA of 1996). Protection of identifiers and quasi-identifiers is of primary importance. Quasi-identifiers are those data elements that are not unique, but can be combined with other elements to identify a specific person or household. This includes spatial and temporal information that is intrinsic in the data collection technologies discussed in this research.
2. Preventing misuse of data that violates "informed consent" policies (U.S. vs. Jones, 2012, Title 45 CFR Part 46). While most anonymization methods are

designed to protect against unauthorized release of PII, there is also a concern regarding the misuse of data internal to a transportation organization.

What is Personally Identifiable Information (PII)?

PII, or, sensitive personal information (SPI), is defined by NIST Special Publication 800-122 (NIST, 2010) as:

"any information about an individual maintained by an agency, including (1) any information that can be used to distinguish or trace an individual's identity, such as name, social security number, date and place of birth, mother's maiden name, or biometric records; and (2) any other information that is linked or linkable to an individual, such as medical, educational, financial, and employment information."

Figure 1 sketches PII within a transportation planning or engineering context by showing the simple but important linkage between a person's identity and his or her technology (devices, vehicles, RFID cards, etc.). Knowing the identifiers of a person's technology can potentially be used to identify a specific person. Within the context of transportation data there is also a concern that data regarding travel can be used to derive an individual's identity even in the absence of other personal details (Zang, 2011; Hoh, 2010; Montjoye, 2013). Frequently, transportation data sources involve the use of data generated through personal technology or from a vehicle. In these cases, there is the potential for using technology or vehicle identifiers in conjunction with other information (i.e., product or vehicle registration documents). More importantly, there is the potential to use a simple sequence of coordinates representing a person's travel trace to indicate a home or other location that can be used to identify a specific person (Zang, 2011; Montjoye 2012). The anonymization of a person's location data is the biggest challenge facing transportation data collection methods.

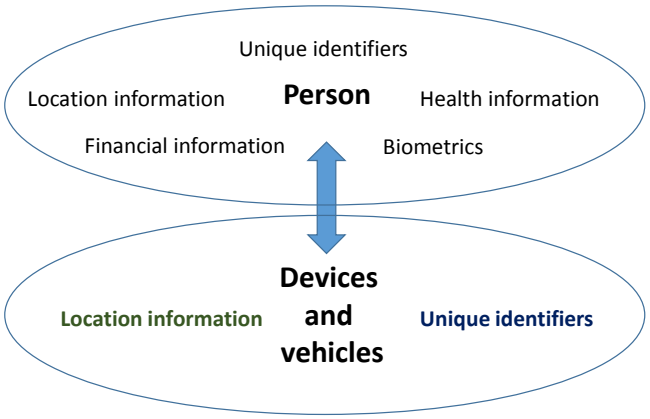

Figure 1 - Personally identifiable information – transportation industry focus

Legal Boundaries for Collecting and Using Transportation Data Containing PII

The legal boundaries that define the collection and use of PII by agencies stem from the Fourth Amendment and Title 5 of the U.S. Code (5 U.S.C. § 552). The code states that agencies must protect unlawful access of PII. This is commonly recognized to mean that possession of PII is not unlawful, but that the protection of unlawful access to PII is the responsibility of the agency. The challenge for agencies that knowingly or unknowingly collect PII is to ensure that adequate "protection of unlawful access" exists. Title 5 also authorizes the use of statistical methods for aggregating or otherwise protecting PII data when used "to support any research or statistical project."

In addition to the protection of existing PII, there are also legal issues related to data collection methods. Not many, however, are specifically related to transportation data. Table 1 provides a summary of key legal cases relevant to finding the balance between transportation data collection and privacy protection. A revelation in the U.S. cases are rulings about the "expectation of privacy" that apply in many situations regarding data from personal vehicles and personal devices. "Expectation of privacy" is used to describe the boundaries of what information is private (protected) and what is public (U.S. vs. Knotts, 1983). The term is loosely defined and varies with changing public attitudes, but should be considered carefully by transportation agencies as it has a legal context. The U.S. vs. Knotts case determined that someone traveling on a public road has no expectation of privacy, and their activity can therefore be observed and recorded. However, this was limited in a more recent 2012 case that suggested regular monitoring of a specific vehicle, regardless of its location on public roads, is considered a violation of "expectation of privacy" guidelines (U.S. vs. Jones, 2012). This limitation was expanded in a 2014 case in California to include data collected from cell phones without specific consent (Riley vs. California, 2014). For agencies looking for specific guidance on the boundaries of their data collection methods, there is no firm line. It is clear though that the trend in legal cases is that anything that collects detailed information about a specific person (vehicle or technology) without "informed consent" likely violates "expectation of privacy."

Table 1 - Legal cases of note regarding transportation data and privacy

Case	Core issue	Value
Katz vs. U.S. (1967)	Right to privacy when in a public space	Helped define "expectation of privacy" guidelines
U.S. vs. Knotts (1983)	"Expectation of privacy" when traveling on public roads and the role of electronic devices for tracking movement	"Person traveling in an automobile on public thoroughfares has no reasonable expectation of privacy in his movements from one place to another," including stops made along the way
U.S. vs. Jones of 2012	Placement of a tracking device (GPS) on a vehicle to monitor movement	Defined limits to U.S. vs. Knotts by saying that surveillance and monitoring **did** violate "expectation of privacy" rules. Further, it also indicated that placing a device on a vehicle without knowledge of the owner is considered trespassing
Riley v. California (2014)	Access to cell phone data	Data from a mobile device is likely considered private information and not accessible without consent
Google Street View (2011)	Remote Collection of Wi-Fi data	Google was fined for remotely retrieving Wi-Fi signal details while collecting data from their Street View vehicles
TomTom (2011)	Use of an individual's travel data from a navigation device vs. aggregated travel data	In the Netherlands, TomTom was determined to be in violation of privacy laws because individual trace data, while cleaned of private data, can be connected to other data to determine private details. It was also determined that aggregated summaries of data collected from individuals does not violate privacy law

While legal cases are important to understand, the public perceptions of privacy protection can be just as important and much harder to define. The Location Privacy Protection Act of 2001 and the Wireless Privacy Protection Act of 2003 were proposed bills that were never enacted. These efforts would have specified stricter controls on location data protection from personal devices such as smartphones and navigation systems. While not enacted, the bills recognize a public perception issue regarding the protection of such data and its use without consent.

A transportation agency may be legally protected regarding data collection and use but this may not protect them from public relations issues. Even minor concerns over public perception can generate costly delays or create concerns for elected officials. Public perceptions evolve with time making it difficult to define rigid privacy protection and data management policies for those wishing to use transportation data. Further

exasperating the understanding of public perception of privacy rights is the issue of social networking where individuals openly share many details of their private lives but also voice concerns when their data are stored and analyzed. While public perception is a moving target, there are research reports by the Pew Research Center and others that evaluate the limits and intensity of privacy concerns. A Pew research study was summarized at the 2015 Transportation Research Board Conference by Lee Rainie in a presentation titled "Americans" views about privacy in the network age." Pew suggested that public opinion regarding data privacy is highly context dependent and variable over time. This uncertainty regarding acceptable practice suggests that agencies should take a conservative approach when determining data collection methods, sources, and data management practices.

There are several opportunities for protecting or eliminating PII during the typical lifecycle of data collection to reporting. Figure 2 - Opportunities for applying PII protection techniques shows a generic data flow and highlights three stages where PII data anonymization methods can occur.

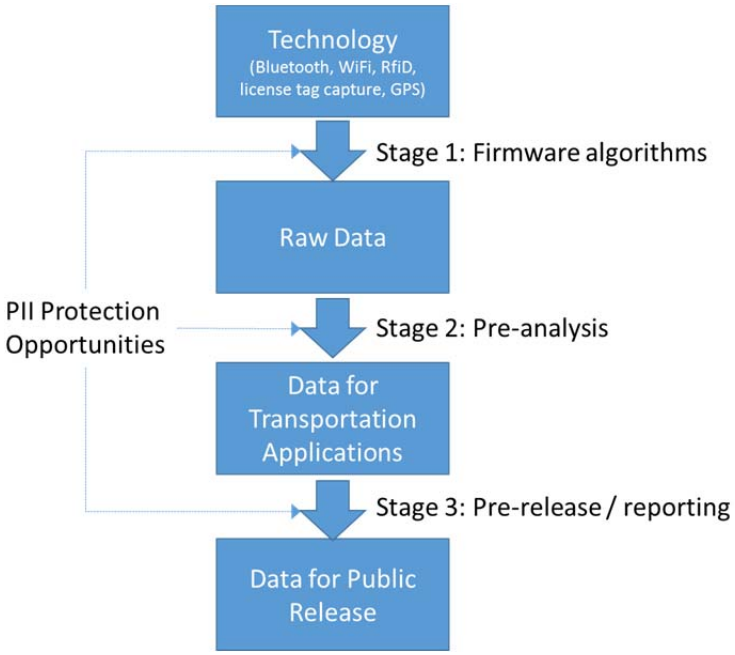

Figure 2 - Opportunities for applying PII protection techniques

In stage 1, field-collected data are transferred to data storage such as a central database server. A firmware method at this stage prevents PII from ever being stored at an agency. The biggest risk of implementing at this stage is that all connection to the original raw data is lost. Given that data collection costs are typically high, many agencies wish to maintain the full raw data to support validation efforts and secondary uses. Stage 2 anonymization can occur before stored raw data are used in analysis or modeling

software. The advantage of processing raw data at this stage is that the original raw data can be maintained but the exposure of PII is limited to those that have rights to the original database. Analysts, modelers, and business partners only have access to the anonymized data. Stage 2 anonymization methods are the primary focus of this research. In the third stage, data are prepared for public release in reports, and available to those requesting data through freedom of information act requests. Waiting until this stage allows analysts access to raw details (including PII), but removes those details prior to public use.

Anonymization Methods

PII protection methods can be divided into policy/procedural methods and technical anonymization methods. The focus of this research is on technical anonymization methods. However, discussion of the technical anonymization methods assumes that agencies avoid collecting or acquiring data that monitors an individual's travel patterns without their express consent (U.S. vs. Jones, 2012, Title 45 CFR Part 46). Identifying the travel behavior of individuals, families, and communities is an important part of travel forecasting. However, accessing an individual's travel patterns without his or her knowledge or consent is likely a legal violation of privacy even if used for the greater good of society. Despite software or hardware license agreements, data purchased from third-party sources could include trace data from individuals who did not expressly authorize use by a transportation agency. Agencies purchasing third party data must therefore have a policy to avoid this data or request a Stage 1 (see Figure 2) anonymization method from their data vendor that removes the ability to construct origin/destination details for an individual.

Transportation agencies should also implement robust data security procedures that prevent unauthorized access, as protecting PII is a legal responsibility of the agency. Appropriate levels of data encryption (128 bit AES encryption) that prevent unauthorized data access should be developed and maintained. While this issue is not the focus of this research, it is a vital element of legal and perceived PII protection.

Transportation Data Collection Technologies

Stationary sensor data can be single observations or based on multiple observations. Single observations are typically collected in support of measuring or monitoring account activity, such as for transit and tolling operations that record events for the purpose of charging user fees. Single events can be tied directly to account information that can also provide account home address information and be used for transit/toll market analysis. Privacy protection risks for single data point observations of data are primarily associated with secondary uses of the data and the potential for using unique identifiers to link to other data sources.

Multiple sensor-based data points indicate that movement occurred, and this has value to agencies interested in the distance, speed, and path of that movement. Transportation system operational performance analysis efforts frequently use sensor-based systems to derive travel conditions. Performance monitoring is becoming increasingly important to support active traffic management (incident response, event management), before/after investment analysis, before/after operational analysis, and general performance reporting (FHWA, 2014). Supporting these efforts requires enough observations to represent travel conditions experienced by travelers and it requires enough spatial and temporal resolution to provide actionable intelligence.

There are several sensor-based technologies that fit this category:

Bluetooth® is a short-range wireless communication protocol that facilitates communication between electronic devices and follows standards set by IEEE 802. Many modern vehicles use this technology to allow connection between a driver's mobile phone and the vehicle to allow hands-free communication. This technology is also used for wireless headsets or ear pieces that allow users hands-free communication. Most modern devices now have Bluetooth as a standard capability, and emit this signal even when not in use as long as the signal is active. These type of devices have a typical range of 10-12 meters but it can be more or less depending on the manufacturer's design.

Private companies are making use of the Bluetooth signals for transportation data collection. Roadside Bluetooth sensors can be placed over or alongside travel lanes and can pick up the signals from active Bluetooth devices within vehicles as they pass by. Each device has a unique Median Access Control (MAC) address that can be identified by these roadside devices. Because MAC address is unique, two roadside devices (or a network of devices) can be used to identify the time and location that each unique device passed by (Wasson, 2008). Analyzing the data from multiple devices allows transportation agencies to gather estimated travel time between sensors. While there can be lost signals, turning vehicles, or stopped vehicles, there are typically enough good matches on major roads to estimate average travel time. A network of three or more sensors is commonly used to capture average travel times across multiple points (Filgueiras, 2013).

The PII data found within Bluetooth systems is very limited. The MAC address that is used cannot easily be traced to an individual. That identifier can be used to assess the manufacturer of the device but there is no central registry of MAC address owners. The MAC address can be considered a unique identifier as its observed position in space and time can possibly be used to identify an individual using deductive reasoning. A sensor location at a specific address or along very low volume roads could lead to assumptions about a person's identity. Additionally, long-term tracking of MAC addresses can lead to the establishment of travel profiles by MAC address (recurrent travel times and locations from the same device). While this still does not easily reveal a person's identity, the possibility of inference goes up as more data by unique MAC address is archived (Liebig, 2012). The monitoring of individual MAC addresses over time could likely be considered a violation of "expectation of privacy" rules.

Wi-Fi sensors are frequently combined with Bluetooth sensors as they operate under the same basic principles. Bluetooth devices actively search for a compatible device and wait for a response. Wi-Fi is passive in that it just "listens" for devices that are searching for Wi-Fi signals. The combined sensor approach improves the "hit" rate significantly, allowing more devices to be detected, and thereby increasing the pool of samples identified as traveling between two points. Just as in Bluetooth, the Wi-Fi reads the MAC address and records the unique identifier along with the time and sensor location. Because the resulting data are almost exactly the same as that of Bluetooth, Wi-Fi collected data shares the same PII issues discussed regarding Bluetooth.

License plate tags on vehicles are unique within their state and have been used for many years by transportation agencies to identify travel time between two roadside observation points. The data capture methods rely on observing license plate information, typically through specialized video equipment and software. The data recorded include the license plate identifier (tag), location, and time. As with other roadside sensors, the data are compared to other data collection locations to determine travel time between two points. In many instances, the license plate is matched to vehicle registration data to identify the owner's name, address, and phone number. Survey post cards or phone interviews are sometimes sent to these specific individuals to gather travel details such as trip purpose, trip origin, and trip destination.

In addition to the PII issues for other roadside sensors, license plate captured data also have PII associated with accessing vehicle registration data. These individual and household details are clearly considered PII and therefore protected. The Drivers Privacy Protection Act (DPPA) enacted in 1994 prohibits the release or use of PII by any state department of motor vehicles unless specific consent is provided (DPPA, 1994). Use of the registration data is specifically allowed for a number of reasons including research and surveys. It should be noted that if data are stored in a database for transportation planning purposes, protection of the PII remains in accordance with the DPPA as well as the broad PII protection statutes.

Radio Frequency Identification (RFID) technology is frequently used by transportation agencies for toll and transit fares. Subscribers use RFID cards to access transportation facilities. When an RFID card is read, an event is recorded that includes a time, location, and account number. The account number can be directly connected to a customer database that contains private information including names, home addresses, financial information, and other identifiers and quasi-identifiers. This technology is also used by private companies for similar financial purposes. Secondary use of this information for traffic engineering and transportation planning are common (NCHRP W174, 2011).

A primary privacy issue that can arise with sensor data is the level of consent. Informed consent, where the user authorizes the data collection for a specific purpose, is not typically pursued for data used in the analysis or travel behavior or system performance. General consent is sometimes provided through device or software license agreements, but this does not typically meet institutional review board (IRB) definitions.

It is conceivable that a single identifier from someone's personal technology is observed over time to the extent that someone's travel habits and behavior can be analyzed in the absence of their consent or knowledge, and, based on unprotected identifiers or data analysis, their identity can be revealed. This level of detail could possibly be interpreted as a legal violation of privacy even if the data were collected entirely on public access facilities.

GPS-capable technology is prevalent across a wide range of in-vehicle and handheld/wearable devices. GPS-enabled devices can estimate their position on the earth's surface and generate these estimates in rapid succession, allowing for movement to be tracked and recorded. The number of GPS-reliant applications has exploded in recent years, particularly in the transportation area, as they can be used for personal navigation and fleet monitoring. Unlike roadside sensors, GPS data are gathered and stored on the device, or broadcast real-time to a server. Accessing this detailed travel data requires data communication directly to the device or from transferring archived data files. For transportation agencies, GPS data are either collected by an agency through the distribution of hardware or software to individuals (recruited public or installed in fleet vehicles), or purchased through a third party source.

Due to its potential for accurate and precise positioning, there is a general PII risk based on location and time. An analysis of GPS points can possibly reveal a person's identity, home location, work/school location, and typical travel patterns. Other PII risks come from the potential linkage to other databases and violation of "expectation of privacy" guidelines regarding monitoring in the absence of consent.

For the sake of evaluating specific PII protection risks, GPS data are grouped into the following categories:

GPS data collected for household travel or activity surveys: GPS is now commonly used as part of many household travel or activity surveys that are conducted to support the development of regional travel models and policy evaluation (NCHRP 775, 2014). GPS data regarding a household member's travel is collected either through a device placed in a vehicle, a device carried by the household member, or through an application run on the household member's personal smartphone. In all of these cases, GPS data is recorded for a designated period of time and used to identify travel origins, travel destinations, and travel paths. For GPS data from devices carried by an individual, the data can also be used to identify travel mode and travel paths away from roads. Since the data is collected from recruited participants, an agency has participant-specific consent that eliminates many of the legal data usage concerns. However, an agency using this data is still obligated to protect the PII from unlawful data access and is limited to uses provided in the specific consent language. Data that are also collected outside of a mutually agreed date and time range may also be protected and unusable due to lack of consent issues.

GPS data collected by probe vehicles: GPS collected from probe vehicles can either come from recruited private citizens, data collection contractors, or employees, and the vehicles used can be privately owned or owned by the agency. GPS data collected as a

paid employee or contractor must be limited to the agreed upon work activities (National Workrights Institute, 2009). Any data collected outside of designated work schedules is considered private information and subject to monitoring without consent rules.

GPS data purchased from private company: Several private companies are re-selling GPS trace data collected from numerous sources as an aid to transportation agencies for performance evaluation of transportation networks (NCHRP 775, 2014). The quality, extent, and level of PII protection provided by the data re-seller can vary significantly. Consent for the data use and re-selling is generated from license and use agreements accompanying hardware and software used by individuals. This consent is general in nature and not specific to a particular use or agency.

Anonymization Methods

Anonymization methods for these fixed and mobile data sources target four primary objectives for transportation data gathered and used by agencies without specific consent:

1. Protect PII direct identifiers (names, IDs, addresses, etc.)
2. Protect pseudo-identifiers and non-specific attributes that can be combined to identify PII (including position information that can be used to derive physical address information)
3. Protect sequences of travel origins and destinations that reveal a behavioral pattern location that can identify PII
4. Protect micro-scale movement traces that can reveal a behavioral pattern or location that can identify PII

"Safe Harbor" Method

The "Safe Harbor" method comes directly from the Health Insurance Portability and Accountability Act of 1996 (HIPAA) Privacy Rule §164.514(b) and is the simple removal of 18 specific identifiers from a database.[1] This method, while targeted for organizations that maintain health information of individuals, has been regarded by many public agencies as the definitive list of identifiers that should be protected. These identifiers are:

1. Names
2. All geographic subdivisions smaller than a state, including street address, city, county, precinct, zip code, and their equivalent geocodes, except for the initial three digits of the zip code if, according to the current publicly available data from the Bureau of the Census:

[1] http://www.hhs.gov/ocr/privacy/hipaa/understanding/coveredentities/De-identification/guidance.html#safeharborguidance

   a. The geographic unit formed by combining all zip codes with the same three initial digits contains more than 20,000 people; and
   b. The initial three digits of a zip code for all such geographic units containing 20,000 or fewer people is changed to 000

3. All elements of dates (except year) that are directly related to an individual, including birth date, admission date, discharge date, death date, and all ages over 89 and all elements of dates (including year) indicative of such age, except that such ages and elements may be aggregated into a single category of age 90 or older
4. Telephone numbers
5. Fax numbers
6. Email addresses
7. Social security numbers
8. Medical record numbers
9. Health plan beneficiary numbers
10. Account numbers
11. Certificate/license numbers
12. Vehicle identifiers and serial numbers, including license plate numbers
13. Device identifiers and serial numbers
14. Web Universal Resource Locators (URLs)
15. Internet Protocol (IP) addresses
16. Biometric identifiers, including finger and voice prints
17. Full-face photographs and any comparable images
18. Any other unique identifying number, characteristic, or code, except as permitted by paragraph (c) (assignment of a unique code that can be used to re-identify records at a later date).

The following data obscuring techniques can apply to these cases:

- **Deletion** – values are removed
- **Hiding** – values are replaced with constants
- **Hashing** – values are translated into a code
- **Permutation –** values are given unique new values
- **Shift** – numerical values are given a fixed offset
- **Enumeration –** numerical values are changed but maintain order
- **Truncation** – values are shortened to obscure detail

The practice of removing/obscuring identifiers offers minimal protection and has been shown that de-identified data can be re-identified (Denning, 1982; Duncan, 1991; Sweeney, 2000; Elliott, 2001; Sweeney, 2002). This "Safe Harbor" method also requires removal of anything that provides "actual knowledge" of an individual. "Actual knowledge" refers to any information that can be combined with other information to reveal an individual. This directive is less certain and is the main focus of many of the technical solutions offered in this review.

Removal, masking, or suppression of these identifiers also does not address the particular location issues facing transportation data collection methods (Krumm, 2007; Hoh, 2006). Precise coordinate information is vital to the value of the data as well as the source of privacy concerns.

K-anonymity

K-anonymity (Sweeney, 2002) is a recognized statistical disclosure control standard for de-identifying data, and expands on the protections of the simple removal of "Safe Harbor" identifiers. The core concept of k-anonymity is that identifiers and quasi-identifiers within a database or relational database can be tested to ensure that no combination of data elements is unique to a single record, and therefore a single individual cannot be identified. When unique records are identifiable, cell values are combined to the point where unique identification is no longer possible. For example, if a database of individuals includes a 5-digit zip code variable for both home and work locations and analysis shows that there are circumstances where there are unique combinations, then the 5-digit zip can be altered to a 3-digit zip for one or both values. K-anonymity is then tested to ensure that the resulting values are no longer unique within a minimal tolerance. In essence, the results rely on "safety in numbers" and that records in the released database are indistinguishable from k-1 records. Accepted minimum sample sizes can vary and typically are increased with the level of sensitivity in the data.

This technique and its many derivations have been widely used, particularly in health and demographic datasets. In cases of transportation data, the problem sets become more complicated but the protection concepts of k-anonymity can still hold true. Pensa (2008) and Negiz (2009) both showed that the k-anonymity approach could be applied to sequences of point locations such as might be found in sensor-based data products or coarse GPS traces. Points along the multiple vectors can be assumed to be the same if they fall within a certain tolerance. K-anonymity of the vectors would be satisfied when every generalized point-to-point vector has enough records to prevent unique record identification.

One example of using k-anonymity for detecting high-risk situations from combining variables comes from the Census Transportation Planning Products (CTPP). For the CTPP that was based on the 2000 Census, the Census Bureau required at least three cases (k=3) in a table cell (table includes demographics, means of transportation, and location variables) before an estimate was reported for that cell. The application of this rule resulted in a sizable amount of data suppression. Since the application of the Rule of Three with the 2006-2010 American Community Survey (ACS) would lead to

substantially more suppression due to its smaller sample size, the National Cooperative Highway Research Program (NCHRP) commissioned a research study (NCHRP 08-79) into approaches that would retain the necessary variables to support the desired tables at the traffic analysis zone (TAZ) level.

L-diversity

L-diversity is an approach to counter known flaws in the k-anonymity concepts (Machanavajjhala, 2007). A homogeneity attack and background attack are situations where a person seeking to identify someone from a k-anonymized dataset can use other knowledge about a person or event to reduce the possible candidates within the data. The proposed L-diversity approach expands on k-anonymity by requiring that any combination of quasi-identifiers have defined variability within each set. The resulting combinations thereby prevent external knowledge from reducing the possible record values. This approach was presented as an improved method for cloaking identities in a mobile environment (Bamba, 2008; Dewri, 2011).

T-closeness

T-closeness expands on the k-anonymity grouping and the l-diversity requirement by adding an additional condition that each grouping must have a distribution of grouped values similar to the universe of records (Li, 2007). This additional variation within a grouping prevents skewed samples from being de-identified.

CTPP - MACH perturbation approach

The methods for tabular data considered in the NCHRP 08-79 study were of two broad types. Post-tabular methods use cell suppression or modified table values after the tables are generated from the microdata (Hundepool et al., 2012), while pre-tabular methods apply treatments to the original microdata to produce a perturbed dataset from which all the tables are produced. The post-tabular approaches, other than cell suppression that were considered, included adding noise to the counts in tables, controlled rounding of table counts that ensure internal cell counts sum to the original published marginal counts, controlled tabular adjustment (Dandekar, 2004), and the approach adopted for the Longitudinal Employment Household Dynamics (LEHD) OnTheMap system (Machanavajjhala et al., 2008; Abowd et al., 2009).

While the post-tabular approaches have some advantages, it was decided to adopt a pre-tabular approach for the 2006-2010 CTPP. This choice was based on criteria relating to the impact on disclosure risk and utility (e.g., maintaining multivariate associations), operational practicability, applicability with a variety of types of variables (mainly demographics) and estimates, ability to facilitate variance estimation, and ability to provide consistent results within the set of CTPP tables. Of particular importance is the fact that a pre-tabular approach can be used in an already established production framework to produce multiple tables that have consistency in the margins and, therefore, are additive as tables are aggregated. Krenzke et al. (2011, 2013) provide a detailed

review of the pre-tabular approaches considered for the CTPP, including a thorough evaluation of select synthetic and perturbation approaches. The pre-tabular perturbation approaches used for the 2006-2010 CTPP generated an ACS dataset that comprised a mixture of original ACS data and randomly perturbed values. This dataset was then used to generate all CTPP tables.The core of the pre-tabular perturbation approaches for CTPP is a model-assisted constrained hotdeck (MACH) procedure. The objective of the approach is to change the value of the published categories by only one or two categories by changing the value of the underlying version of the variable, and still retain multivariate associations among the variables.

While these methods were not targeted for location data, the complexity of the approach necessary to achieve objectives while preserving privacy exemplify expected solutions that may need to be targeted for specific applications. "Privacy-by-design" is an approach that considers both the application and the data structures, and is likely necessary to achieve maximum data utility (Cavoukian, 2009).

The following privacy protection methods are specifically targeted for spatial and temporal data and considered with respect to the transportation fixed and mobile datasets.

Geographic Masking (Armstrong, 1999)

Armstrong proposed a variety of methods for masking location data for privacy protection while maintaining the ability to conduct spatial analysis (Armstrong, 1999). These methods involve masking entities within a zone through affine transformations, random perturbation, point aggregation, and nearest neighbor assessments. Each technique offers strengths and weaknesses for various spatial analyses. For transportation data, there may be some value in these approaches for masking transaction or travel activity locations within zones (such as TAZs). Clifton 2014 provides a more thorough review of these techniques.

- **Affine transformations –** Values are manipulated geometrically such that the original positions of the points are altered but the spatial relationships between the points are maintained. This also alters any defined relationship between other geographic entities and has limited value for transportation datasets.
- **Random Perturbation –** Values are randomly moved within a limited area such that the original locations are altered but the general position and relative position are limited. This also alters the time/distance calculations for sequential points, as well as the spatial connection to other transportation datasets.
- **Aggregation –** Location points are "blurred" and snapped to some aggregation point with an appropriate scale for analysis. Resulting points are in the general area of their true location but do not reveal their true location. Careful consideration must be given to the scale of the aggregation. Aggregation points maintain an approximation of their true position and therefore can potentially relate with other transportation features.

Spatial and Temporal Cloaking (M. Gruteser and D. Grunwald, 2003)

Gruteser and Grunwald demonstrated a spatial and temporal cloaking approach for on-road positional that uses estimated traffic volumes by time of day to determine when a k-anonymity threshold has been achieved. They also proposed a quadtree algorithm to generate the results. This approach was targeted specifically for Location Based Services (LBS) but may have merit in transportation planning and engineering applications, particularly in their approach for defining k-anonymity using standard traffic data.

CliqueCloak Algorithm (B. Gedik and L. Liu, 2005)

The CliqueCloak algorithm is targeted for individuals wishing to suppress information broadcast to organizations. This application is not ideally targeted for the specific applications being evaluated, but it presents some approaches and concepts that may prove useful in combination with others. The approach defines a spatial box around an individual point to achieve k-anonymity in combination with the suppression of other attributes.

Path Confusion (Hoh and Gruteser, 2005)

Hoh and Gruteser developed an anonymization approach that is specifically targeted for GPS vehicle traces in support of traffic engineering and transportation planning applications. Their approach processes trip points for a large vehicle trajectory dataset and identifies locations where two or more paths cross or meet in close proximity. At these points, the paths are perturbed such that one trace is no longer distinguishable from another, in essence, eliminating the possibility of identifying a unique trace. This approach requires a significant sample size and trace density to be effective.

Location Suppression (Terrovitis, 2008)

Terrovitis proposed a method for privacy protection of transaction record databases that include location information (Terrovitis, 2008). He demonstrated a location suppression algorithm that assumes that an adversary holds partial information regarding an individual's movement. The algorithm iteratively suppresses locations until a privacy constraint is achieved. This approach has value in particular for applications in RFID where account information may be known by more than one organization.

Never Walk Alone (Abul, 2008)

Abul developed an approach for anonymizing data from moving objects called "Never Walk Alone" (NWA). This approach relies on the positional uncertainty of a point in moving objects data to define a cylindrical error range. Other traces with overlapping error ranges are all combined to satisfy a k-anonymity principle. The solution has particular value for traces of data where the resolution of the data are fairly coarse, allowing for a cylinder definition that is large enough to find other trace matches to satisfy the k-anonymity goal.

Time to Confusion (Hoh, 2010)

Hoh later proposed an expansion of the path confusion termed "time to confusion." This approach assesses the maximum time that a user (GPS trace) can be followed before privacy rules are violated. The approach also includes a frequent trip end filtering process so that GPS traces do not reveal home/work/other activity locations that can be used to reveal identity. The time-to-confusion metric quantifies tracking risk and is used to apply a path cloaking algorithm that removes traces in low density areas and prevents re-identification based on speed/direction logical extensions. The solution is designed for large datasets of GPS traces suitable for traffic engineering and transportation planning applications.

Mix-Zone Algorithm (Dahl et al., 2010; Carianha et al., 2011)

Mix-zone approaches are designed for individual vehicles that are attempting to protect their identity from systems. This differs from the objective of this research but the methods can still be evaluated for relevance. In mix-zone algorithms, unique positions are maintained, but identifiers are provided pseudonyms within a zone or ring that includes a sufficient number of other vehicles. This prevents systems from tracking movement for a single vehicle but still allowing for use of the individual point data. This technique has been widely published in intelligent transportation system (ITS) research related to connected vehicle systems.

Virtual Trip Lines (Ban, 2012)

Virtual Trip Lines (VTL) are a method for applying privacy protection principles specifically for traffic engineering studies. The concepts presented offer techniques for designing traffic data locations that maximize data utility for some purpose but minimize the privacy exposure of the data by only collecting information needed. Further, the approach provides standard cloaking methods within those minimal datasets. This discussion has value to the research due to its dialogue of data collection restriction for privacy protection.

VTL Zone System (Sun, 2013)

Sun et al. provide a research paper with an excellent review of the state of practice in privacy protection for transportation data, and proposes a VTL zone approach that addresses many of the specific location privacy concerns while maximizing data utility specifically for traffic engineering. In the paper, various technical approaches are outlined and criticized as being flawed for transportation engineering purposes. Methods for location perturbation (Gedik, 2005), sample reduction (Tang, 2006), location hiding (Hoh 2010), and dummy traces (Nergiz, 2009) all degrade usefulness. The VTL zone system allows microscale detail with zones or areas of concern (signalized intersections, congested area, etc.). Identifiers are anonymized between zones so that individual traces within a zone cannot be matched with traces in other zones even if they are the same vehicle. Further, a random sampling of traces within a zone can be implemented for

additional protection. This approach maximizes the value of second-by-second GPS traces necessary for traffic applications.

Discussion of Privacy Protection Algorithms

The discussion of privacy protection algorithms in this section is intended to provide practitioners with targeted and detailed descriptions of application and value. The discussions are based on protection objectives that address both intruder risks and accidental disclosure. These objectives are consistent with most PII protection scenarios facing transportation agencies; the same list is in the previous section:

1. Protect PII direct identifiers (names, IDs, addresses, etc.)
2. Protect pseudo-identifiers and non-specific attributes that can be combined to identify PII (including position information that can be used to derive physical address information)
3. Protect sequences of travel origins and destinations that reveal a behavioral pattern or location that can identify PII
4. Protect micro-scale movement traces that can reveal a behavioral pattern or location that can identify PII

These discussions and examples are targeted for specific use scenarios. The many PII protection techniques published in literature are designed to address specific conditions, and few can be considered for broad application across mobile or fixed-sensor transportation data. In many cases, however, transportation agencies gather data for a specific purpose but wish to maintain an "anonymized" dataset to serve multiple secondary purposes or have available for future distribution. In these cases, the intended end use is unknown, and therefore, a specific algorithm designed to maximize utility is not possible. Generalized methods such as this restrict the full data utility but are useful for owners of raw private data who wish to have a shareable dataset for those conducting research or secondary analyses. This situation is discussed in the methods below given that there is no one-stop PII algorithm. Agencies attempting this solution are encouraged to be familiar with the PII objectives used in this review, the Safe Harbor variables, and the principles of *k*-anonymity.

Sample Data

The sample data used for the illustrating anonymization methods comes from the Northeast Ohio Areawide Coordinating Agency (NOACA) GPS-based travel survey conducted in 2012. The GPS data was used to develop multiple sample datasets for exploring both fixed and mobile data sources. GPS data for all study trips that passed near the intersection of Lee Road and Van Aken Boulevard were selected and included in the sample dataset. The resulting dataset includes 1,573 unique trips. Figure 3 shows a sample GPS trip from the NOACA data and the key intersection that was used to select trips for exploring anonymization methods in this research.

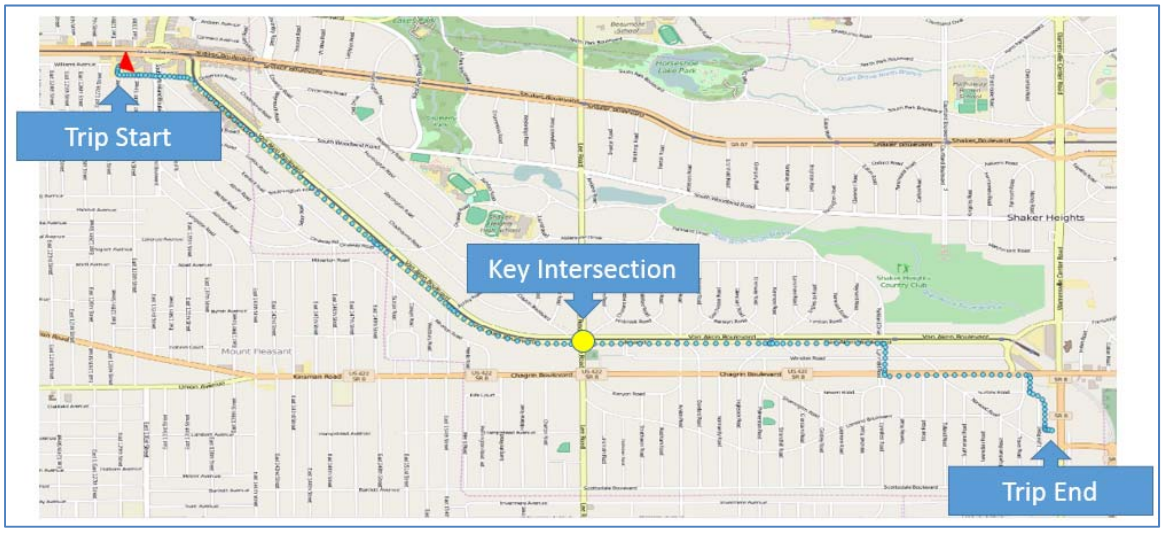

Figure 3 - Sample GPS trip and key intersection

Trips within this dataset are composed of individual GPS points, typically collected every 3 seconds during the travel. The fields for each data point include a unique person ID, trip ID, longitude, latitude, date-time, speed. No other trip metadata regarding trip purpose or other information is included in the dataset. The simple trip trace, as evident in Figure 3, includes a trip start and trip end that must be considered protected in the absence of informed consent. The ability of the various algorithms to protect this information, as well as prevent the isolation of a single identifiable trip trace, will be a focus of the methods evaluation. The start date and start times of the sample trips have been altered for the evaluation to ensure enough temporal trip overlap to properly test the various methods.

To illustrate the protection methods for Bluetooth, Wi-Fi, and other fixed-sensor systems, a network of device locations was developed for the study area along several of the roads where GPS trips are evident. Simulated sensor locations were virtually placed within a GIS environment at intersections along varying classifications of roads, including a few lightly traveled local roads. Having sensors along lightly traveled roads, while not typically useful for applied traffic measurement, will help this research effort in illustrating the effectiveness of algorithms in low volume conditions. Figure 4 shows the location of the 70 sensors.

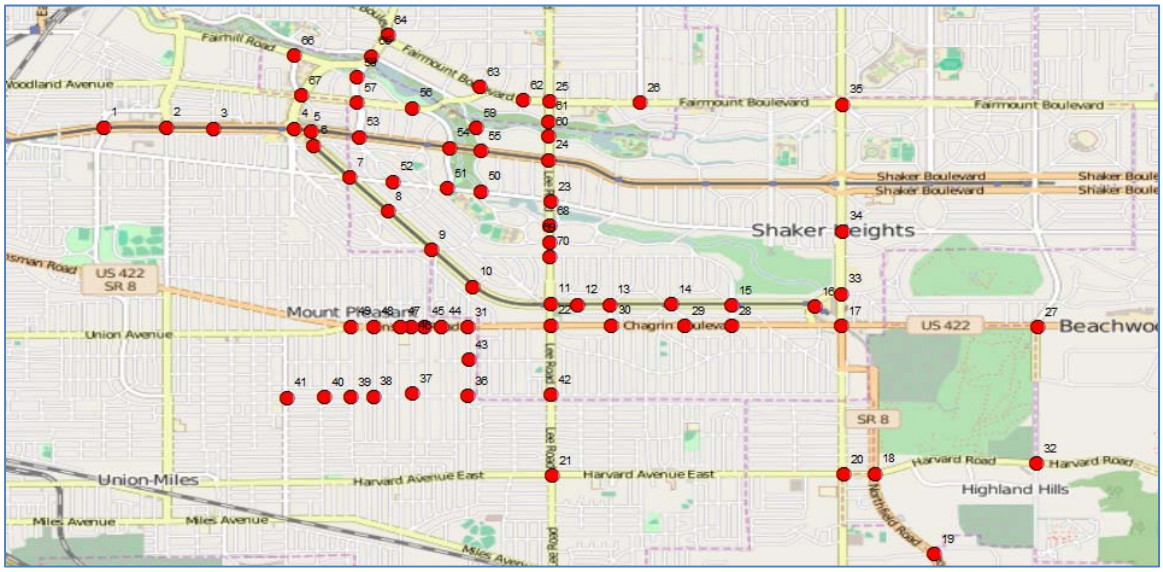

Figure 4 - Network of fixed sensors

After comparing the GPS travel data with the sensor network, an "observation" dataset is created that identifies the location and time when a person (device) passes within 100 feet of a sensor. The resulting dataset includes a sensor ID, person-device ID, and a timestamp. A sensor match table is also created that records travel between two sensors by the same person-device ID. That table includes a person-device ID, "from" sensor ID, "to" sensor ID, "from" observed time, "to" observed time, distance between sensors, and the estimated average speed between sensors. These two tables represent information that can be retrieved from a sensor-based system excluding connection to other datasets or subject to any data cleaning algorithms.

To support the evaluation of origin-destination based sensors, such as transit fare cards or other RFID tags, every GPS trip end (origin and destination) was assigned a U.S. Census Block. The resulting OD dataset includes a person-device ID, trip start time, origin zone, trip end time, and destination zone.

The resulting set of sample tables includes those shown in Table 2:

Table 2 - Data tables used for evaluation

Table	Description	Records	Fields
GPSData	GPS points from 1,574 trips	368,656	personID GPStripID longitude latitude timestamp speed
Sensor_Obs	List of observations by sensor	9,458	sensorID personID timestamp
Sensor_Match	List of trip segment matches by sensor pair	9,065	uniqueID fromBTID toBTID startTime endTime distance(miles) speed(mph)
OD	List of trip origin and destination pairs	1,574	personID startTime originBlock endTime destinationBlock

Safe Harbor Identifiers

Agencies understand and frequently implement the basic privacy protection method of removing obvious identifiers in shared datasets such as names, social security numbers, account numbers, and financial information. The 18 Safe Harbor identifiers expand on this approach by providing agencies with a comprehensive list of items that should be considered protected. These 18 types of identifiers should not be shared unless there is specific consent. Additionally, agencies should carefully consider what elements need to be shared internally to their organization to minimize exposure of private information.

One of the biggest challenges that transportation agencies face regarding the 18 Safe Harbor variables comes from the need to protect location information: "All geographic subdivisions smaller than a state, including street address, city, county, precinct, ZIP code, and their equivalent geocodes...." Location information is intrinsic to

transportation data, and its removal from datasets greatly limits the data's utility. For this reason, agencies should first ensure that consent is achieved for the specific agencies and uses of the data. In the absence of specific consent, or when the need to include location information for geographic regions with populations less than 20,000 is necessary, technical anonymization methods, such as those discussed in this document, should be employed.

Consider Pseudo-Identifiers

Pseudo-identifiers are data attributes that alone do not identify a specific individual, but when combined with other non-specific attributes, they result in the unique identification of an individual. For example, consider a travel survey database where personal identifiers have been removed, but other aggregate fields exist. There may be 100+ records showing participants who live in a particular zone, or have a family size of four, or an income range of $50,000-$75,000 per year. But if you combine those conditions, you could potentially limit matches to one or two participants. In this case, the fields of zone, family size, and income are pseudo-identifiers. Identification and treatment of pseudo-identifiers can be managed through k-anonymity principles described in the next step.

Datasets that are generated for distribution to serve a specific secondary purpose, such as releasing GPS-based travel survey data for traffic speed analysis, can limit pseudo-identifiers by only providing the necessary variables required.

K-anonymity / L-diversity / T-closeness

K-anonymity is the basic principle measure of the uniqueness of a record. Statistical disclosure control (SDC) methods, such as aggregation (coarsening), have the principle aim of reducing the uniqueness of a single record to the point where it is indistinguishable from k-1 other records. Relating to a location-based sensor or GPS data, as mentioned in Bamba et al. (2008), a user is considered location k-anonymous if and only if the location information is indistinguishable from the location information of at least k-1 other users. Machanavajjhala et al. (2007) argue that k-anonymity is deficient and cannot prevent attribute disclosure, that is, there may be at least *k* cases with the same values among indirect identifers, but they all might have the same sensitive value, so the sensitive values need to be diverse. Therefore, it is l-diverse if there are at least *l* cases of a sensitive variable (e.g., static locations or symbolic addresses such as a church, restaurant, or doctor's office). It is t-close if the difference between the distribution of a sensitive attribute in the particular subgroup and the distribution for all subgroups together is less than *t*.

The k-anonymity, l-diversity, and t-closeness approaches are used to identify high risk scenarios (either from single identifiers or combinations of pseudo-identifiers) that are in the data. Once the high risk scenarios are identified, the typical steps are to redact information from the file (sometimes referred to as data suppression), combine or

aggregate the data (sometimes referred to as data coarsening or cloaking), or treat the data by perturbing, subsampling, or creating synthetic model-based data, all via a random mechanism. In any case, the quality of service (QoS) or data utility is reduced in order to maintain confidentiality. In the case of data suppression or coarsening, information is lost. In the case of perturbation, or generating synthetic data, information is changed and some distinct information is blurred. Reducing disclosure risk and the retaining QoS is the underlying objective.

A high value of *k* offers more protection but reduces the granularity of the data. Defining the ideal value of *k* is dependent upon the intended use of the final data as well as the potential risk scenarios. In the absence of a full risk assessment (many risks are unknown), agencies vary on the defining of a minimum standard for *k*. Relating to the CTPP, k=3 is used to identify risk, whereas there are examples of $k \geq 5$ elsewhere (FCSM [2005], NCHS [2004]). In light of applying l-diversity and t-closeness, the value of *k* would need to be increased to allow for minimal requirements on *l*. In any case, the idea is to create geographic ranges such that any subgroup defined has at least k-cases.

Consider mobility traces, which are records of locations and times that vehicles visit. In practice, researchers have formed grids to start, and combine the grids to meet the thresholds relating to *k* and *l*. We are not aware of an application of t-closeness in this context.

In any application, subgroups need to be defined. In our test data, subgroups could be defined as mobility traces. Relating to identifying high risk cases in OD data, k-anonymity would ensure there are at least *k* cases in all specific traffic OD flows. Any flows with less than k-cases would be considered high risk. This is the approach used in the CTPP project. In addition, key trip ends (work, school, etc.) can be indicated on the file to help address l-diversity. To address the high-risk scenarios, statistical disclosure control treatments are applied, such as discussed above.

With sensor-based observation data, as illustrated in our test data Sensor_Obs and Sensor_Match, and GPS data, specific traces need to have at least k-cases. There are several aggregation possibilities that can be explored to ensure that all records meet the *k*-anonymity threshold. For example, the sensor data can be transformed to an OD file. The data coarsening requires changes to start and end times, and re-calculations of distance and speed. For GPS data, with data recorded every 3 seconds, suppressing data points is an option, however, it would not provide much privacy protection. Transforming the GPS data into an OD structured file is also an option, depending on the analyst's needs.

Figure 5 shows the Sensor_Match data where records with less than six observations are suppressed and excluded from further analysis or distribution. Agencies wishing to use data such as this typically exclude records with low sample sizes due to statistical validity concerns. However, the privacy of data traces should also be considered, particularly in low traffic volume locations where observed travel could possibly be connected to an individual regularly traveling in the area.

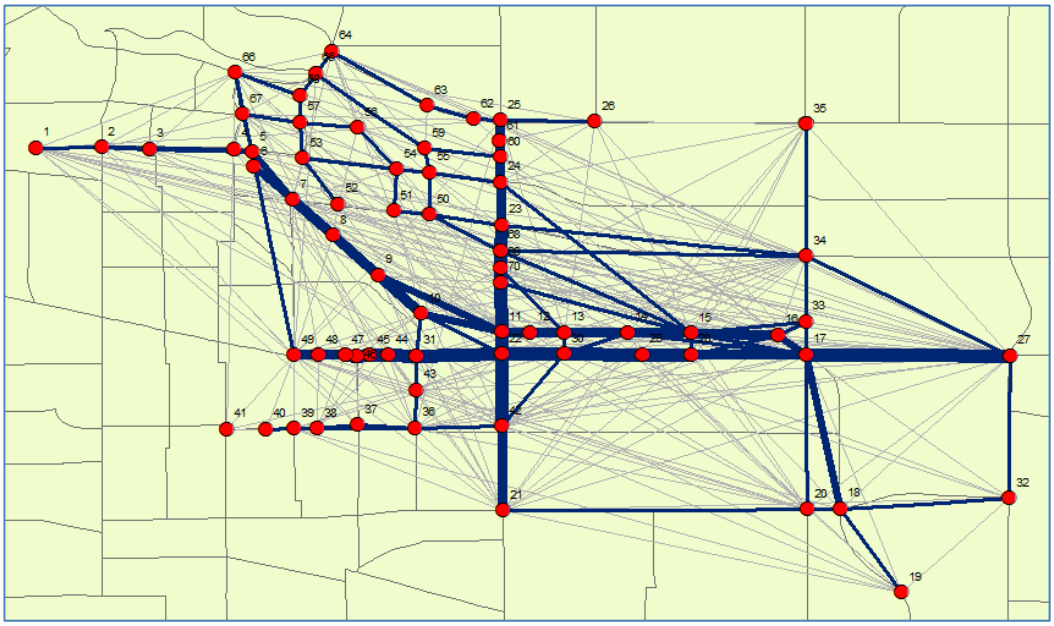

Figure 5 - Observed sensor to sensor travel with frequencies < 6 removed

Bamba et al. (2008) promote a PrivacyGrid system that attempts to capture the user's privacy requirements. It is a framework for supporting anonymous location-based queries in mobile information delivery systems. This appears to be an online system, with queries and real-time confidentiality edits. The basic concept begins with the following steps:

1. Identify and define the universe of discourse, which is a map defining the boundaries by latitude and longitude coordinates.
2. Define a grid and its grid cells within the universe.
3. Define the position of a moving object in the universe.
4. Define the current grid cell of the moving object.
5. Capture the user privacy preference profile.

Then, the risk of the moving object is determined. To provide risk protection, different algorithms were discussed and evaluated, such as "bottom-up grid cloaking" and "hybrid cloaking." The PrivacyGrid system was evaluated Bamba et al., and determined to arrive at close-to-optimum k-anonymity, and considers l-diversity. It may be worth pursuing the system further to evaluate and apply.

Montjoye et al. (2013) used coarsened trace data that had hourly time values and location information comprised typically of 10 to 20 city blocks. With only four randomly chosen observations of an individual, they could uniquely identify 95 percent of the 1.5 million individuals in the file. This illustrates the challenge, and that treatments other than coarsening and suppressing data are necessary to retain some of the usefulness of the data. Further, Dewri (2011) makes a claim that forming geographic ranges may actually help attackers be more confident about the geographic location, and therefore, the author explored an algorithm to perturb location points using a "bucket" of *k*

individuals that are close in proximity to the true location point. Similar treatments, as discussed later, such as path confusion and virtual trip lines, allow for some of the details of the data to be released due to the uncertainty introduced in the mobility traces.

Location Suppression

Location suppression as described by Terrovitis 2008 has value for scenarios where trip traces/vectors are identifiable from a dataset. This algorithm balances the utility of the sequential trip data and the potential exposure of PII. Terrovitis describes this method using credit card transaction data that can be used to build individual credit card owner's sequential patterns of activity. This same approach has relevance to other datasets where trip point or trip end spatial and temporal details are known. For example, consider two origin-destination study example scenarios:

1. The sensor-match table as described in Table 2 includes trip records with sequential travel observations at roadside sensors scattered through a network. The dataset reveals the individual travel patterns of drivers.
2. The OD pairs table as described in Table 2 includes trip records with sequential trip end information developed from RFID cards.

Both of these datasets reveal PII in the absence of informed consent by the fact that they monitor an individual's travel behavior. Additionally, PII can be revealed when the raw data are accessed by a group or individual who combine the data with other datasets or specific knowledge of a user.

The utility of the scenario datasets is the ability of the data to reveal OD travel statistics that can be used to evaluate performance and support planning models. The sensor-match table can be used to identify facility-specific travel metrics for operational assessments. Neither of these uses of the data requires PII to be effective. In fact, individual traces are less useful than a large number of observations, which would be more conducive to estimating travel conditions.

The location suppression approach follows the same concepts of k-anonymity and l-diversity, but applied to spatiotemporal datasets where location becomes a pseudo-identifier. The following example describes the various ways the data can be parsed.

Sensor Match Example

The sensor match database includes records that can be sorted by UniqueID to reveal a sequential series of trajectories. For example:

2000285_1: 30->29->28->16->36->43->…….

2004977_1: 22->30->29->28->16->33->…….

Extracting each individual trip segment (i.e., 30->29) results in 9,064 records, and 324 of these records (3.6%) have less than three observations. Each trajectory also has a start time, travel time, and speed. Restricting observations to time period summaries (i.e.,

AM, PM) as typically needed in developing road performance metrics, would restrict the data and increase the likelihood of a trip segment having less than three observations. Further qualifiers on the data (time, day of week, month, etc.) would result in fewer observations and increase the potential that unique records could be identified.

The Terrovitis algorithm expands on this concept and works with the full travel sequences and iteratively suppresses data until a privacy threshold is passed. Terrovitis includes an additional factor in his algorithm that considers an adversary that has additional knowledge of the sum of the data points. As described in our scenarios, it is assumed that all trip ends or trip segment ends are known to a potential adversary (instead of a subset). Applying the Terrovitis approach to this dataset will result in common sequences of trip segments and isolate the uncommon, which could be removed to protect PII in accordance with *k*-anonymity principles. In the two-record example below, the bold trip segments are equivalent:

2000285_1: **30->29->28->16**->36->43->…….

2004977_1: 22->**30->29->28->16**->33->…….

Eliminating the unique segments results in:

2000285_1: **30->29->28->16**

2004977_1: **30->29->28->16**

Which could then be part of a shared database as the pattern is indistinguishable when unique identifiers are removed. In this particular set of sample data, the sequence of 30->29->28->16 occurred 20 times.

When to apply the Terrovitis algorithm

The Terrovitis approach is targeted for situations where an organization is attempting to anonymize trajectory data (sequences of trip end locations) when a potential third dataset exists with some knowledge about the users. The value of the algorithm for Bluetooth or RFID data is in the identification of common sequences of observations that could be used to evaluate OD travel times or corridor travel times. One approach would be to select sequences of trip ends that hold particular value for an end user, such as performance statistics for a particular signalized travel path. Once the select sequence of trip ends or trip segments is established, the database could then be queried for this sequence. As long as the select set has *k*-1 matches, the results can be added to an analysis database. Following this process for a full set of selected trip end or trip segment sequences results in the establishment of a database that is both anonymized as well as useful for the specific user objectives.

The full and original Terrovitis algorithm can be found in Appendix A.

Time to Confusion

Some papers are written to protect users' traces and identities in real-time, however, we assume that we are discussing these approaches in terms of a centralized data source, where the locations of other users are known to the data producer.

As previously mentioned, employing a technique that loses information (cloaking, suppression) may not provide enough protection, or there may be too much aggregation or suppression to yield usable data for researchers and analysts. Hoh et al. (2010) discuss that the data coarsening approaches (spatial cloaking) fail to provide usable data, for example, probe deployments of 2,000 to 5,500 vehicles are needed for spatial accuracy to remain over 1,000 meters to reach the k=3 requirement.

Furthermore, setting up cloaking boxes (aggregating spatial areas) to satisfy k-anonymity thresholds does not prevent an attacker from tracing a vehicle across the cloaking boxes, through tying together the longitudinal data on the vehicle. Random subsampling (suppression) of traces or data points reduces risk, but does not provide a risk level value, and leaves those in low traffic areas at risk. Longer sampling intervals do not necessarily decrease the likelihood of home identification. The effectiveness of the approach is related to the length of the trace.

To help retain information, a path confusion perturbation algorithm by Hoh and Gruteser (2005) reduces risk from multi-target tracking (MTT). MTT is used not necessarily as an attacking algorithm, but rather to recreate most likely paths based on general assumptions of users' movements in the tracking system communities. The path confusion algorithm tries to out-think the MTT attack by confusing the attacker by perturbing trajectories at points where they spatially meet. That is, the traces are switched where paths are crossed, or where the traces are close in distance between "users" and going in the same direction.

A path cloaking algorithm was created by Hoh et al. (2010) to take into account time-to-confusion to allow for outlier paths to be identified and treated for confidentiality purposes. When tracing time increases, the probability of sensitive destinations and the probability of identification increases. Therefore, the "uncertainty-aware path cloaking" (UAPC) approach includes a time-to-confusion metric, and strives to anonymize location traces with emphasis on traces at high risk of target tracking and home identification. The time-to-confusion is the tracking time between two data points where the track to the next sample could not be identified with near certainty. The algorithm minimizes home identifications while limiting the maximum time-to-confusion. The intrigue is that the developers claim to offer guaranteed protection for users that move into the low-density areas.

Risk and data quality impacts

Privacy impact measures include the percentage of occurrences of crossing segments within the radii R, or include showing how long an adversary can correctly follow users' traces. The UAPC takes into consideration measures of uncertainty via various tracking

algorithms and models. The uncertainty of home identification can be measured by clustering the end points of trips and computing a probability of re-identification using the distances of the end points from the centroid of the cluster.

Quality impact measures such as the distance between original and perturbed trajectories are needed. Measures that take into account distance and speed are needed. In terms of distance, the mean location error provides a useful utility measure. To reduce the impact, a radius can be used to constrain the perturbations within the same area. Useful metrics explored by Hoh et al. (2010) include the maximum and median time-to-confusion, road coverage, and home identification rate (done through manually determining the homes for specific traces). The relative road coverage is computed as the fraction of road segments where speed updates were received within a given time interval, weighted by traffic volume of the segments, relative to the road coverage in the original dataset.

Minimum conditions

A limitation of the initial path confusion approach is that it requires more traffic to allow for the perturbation to take place, and take place in a way that will not damage the data. The path confusion approach needs enough traffic (authors suggest at least 10 vehicles) at time *t*, and therefore the approach does not protect in low-density scenarios. It also relies on the frequency of parallel segments to crossing segments. The approach also can produce artificial intersections and unlikely location samples. Lastly, the algorithm studies are too complex to be administered in a large dataset with many vehicles.

By design, the more recent UAPC algorithm identifies the low traffic areas, especially home residential areas, and does not release them. The authors also provide an option to modify the algorithm to strengthen protection from revealing sensitive locations (hospitals) by ensuring a level of confusion takes place. The UAPC algorithm takes in GPS data for a specific time interval *t*, and the data producers set the maximum time-to-confusion allowed, and the uncertainty threshold.

Process for applications

The earlier path confusion approach could be implemented in various ways on GPS and sensor data. One approach may be to

1. Create segments, which are areas where two paths "meet," to apply the perturbation approach. The segments have a minimum number of consecutive location samples. The paths must lie near each other and may cross or run parallel.
2. Select segments of paths to apply the algorithm. The amount of computation can be reduced by processing short two-user segments.
3. Create a candidate list of users based on risk criteria. Users that do not cross paths are typically at higher risk.

4. Identify pairs of users to apply the path confusion algorithm by identifying pairs of traces that are close in distance.

The more recently developed UAPC algorithm also applies to GPS data, and sensor data. The steps begin with predicting the position of each vehicle based on prior information. Vehicles are identified and released if they already are less than the maximum time to confusion. Other vehicles are identified and released if their tracking uncertainty is higher than the allowed threshold. For the revealed samples, vehicle information is updated for the point of last confusion and the last visible GPS sample. Therefore, the UAPC algorithm only reveals location samples when the time since the last point of confusion is less than the maximum specified time to confusion, or the tracking uncertainty is above the specified threshold.

Example

Consider the GPSData sample data file that includes 1,574 trips represented as sequences of records containing coordinates. These data were originally collected as part of a travel behavior study and contain full location details for participants, including trip ends such as home and work. If an agency wished to release the data in support of a secondary purpose such as building road segment average speeds, there should be an effort to ensure that private location details (home, work, school, etc.) are protected as well as sequences of travel, which could be used to identify a unique pattern and therefore a unique person.

The Time to Confusion approach to this issue is to identify sequences of GPS data that can be linked simply based on their position, speed, and heading. Those sequences that surpass a user-defined distance are considered unique and excluded from the protected output database. In this example case, trip traces that are indistinguishable from at least two other traces are removed within an intersection-to-intersection segment. Figure 6 shows a portion of the GPSData sample with trip end points highlighted as green circles. The elements of privacy concern are the trip ends and GPS traces that traverse segments where 2 or fewer vehicles traveled.

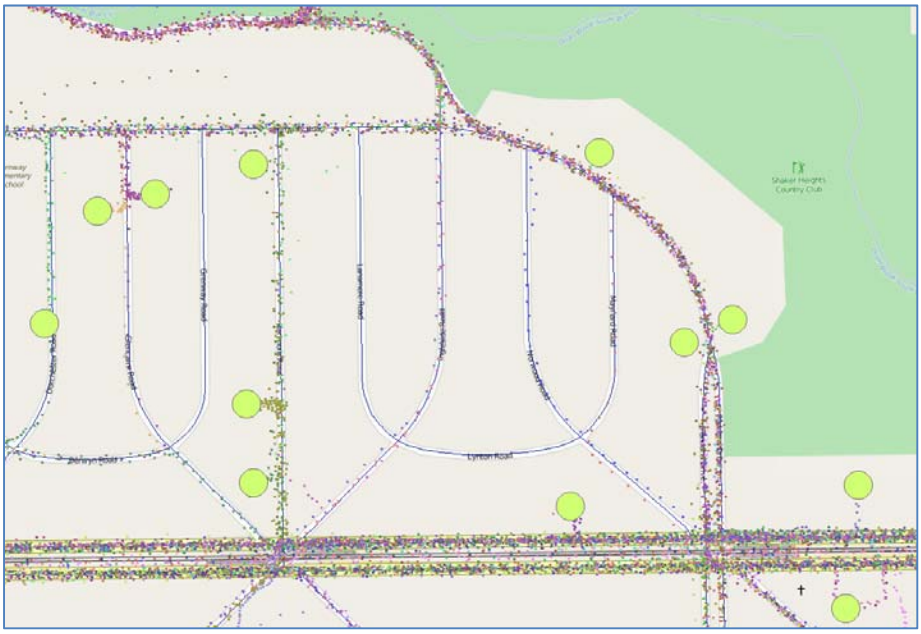

Figure 6 - Example of GPSData sample file with highlighted trip ends

Running through the Time to Confusion approach results in a dataset that has unique traces of GPS data removed. Additionally, all other unique identifiers are removed such that the resulting database is a simple point database of speeds and times. Figure 7 shows the results for the same sample area shown in Figure 6. This restriction eliminates the possibility of recreating long trip traces but retains the point speed, heading, and timestamp estimates. Trips cannot be reconstructed based on speed and heading because the method ensures that records can be uniquely connected after some user-defined distance or time is removed. This approach has some value for certain applications, but limits the uses to point-level speeds or short-space mean speeds.

A full time-to-confusion algorithm can be found in Appendix B.

Figure 7 - Example of GPSData sample file after techniques are applied

Virtual Trip Lines (VTL)

With the advance of science and technology, data collected through mobile sensors such as GPS, include detailed location traces that enable tracking and identifying vehicles and individuals. Mobile data with more detailed spatial information can provide insights on developing new modeling techniques for transportation research. However, it is critical to address the possible privacy breaches for releasing mobile data. The VTL system embeds sophisticated privacy protection techniques into the data collection process. The system ensures that the risk of disclosing privacy is controlled at an acceptable level while maintaining the usefulness of the released data.

As illustrated in Sun et al. (2013), the major component of a VTL system is a trust-worthy location proxy server, which collects data transmitted from mobile sensors and applies the developed privacy protection techniques. Later, the processed data are released through an application server. Data users have access only to the application server, but not to the location proxy server where the privacy sensitive information is stored.

The VTL system only collects location traces within VTL zones, which are areas between two geographic marks, usually placed near intersections. The traces between neighboring VTL zones are removed. The basic privacy protection technique is to remove the trace identifiers and assign random IDs. With the anonymized data there are still chances that the intruders can successfully link the short traces of the same vehicle over multiple intersections. Filtering approaches can be applied on top of anonymization to provide further protection. Instead of releasing all vehicle traces captured by the

location proxy server, only a subset of traces are made available to the application server. Sun et al. (2013) proposed three filtering approaches.

1. **Random sampling.** A portion of the traces, say, 50 percent, are randomly selected and released at each VTL zone. This approach, to some extent, reduces the likelihood of an intruder tracking traces of the same vehicles across zones, but does not take the tracking probability into account.
2. **Individual tracking probability-based filtering.** The decision of releasing or suppressing a trace is based on the individual tracking probability. For example, only release the traces with tracking probability smaller than 0.2. This guarantees that no more than 20 percent of the traces can be successfully tracked. The individual tracking probability depends on path likelihood and travel time distribution. For a given pair of upstream and downstream zones, the path likelihood is calculated as the proportion of the traces that pass both zones. Sun et al. (2013) assumes the travel time of a vehicle between two VTL zones follows a log-normal distribution.
3. **Entropy-based filtering.** This approach is mathematically related to individual tracking-probability filtering but can sometimes have different privacy implications. The entropy-based filtering technique attempts to measure the overall risk in a dataset. The dataset can be released if the risk is lower than an acceptable level. On the other hand, the objective of the individual tracking probability-based filtering is to suppress individual traces that are subject to high privacy risk.

All three filtering approaches make it more difficult to conduct successful links to the anonymized short traces, and therefore the approaches keep the privacy risk under control. In an evaluation study, Sun et al. (2013) showed that the individual probability- and entropy-based approaches provide better privacy protection than the random filtering approach, while retaining similar level of data utility.

Example

Using the GPSData from the sample datasets, a selected set of four major intersections was selected. At each intersection, a 100-meter buffer was defined and GPS data within that buffer was extracted. Figure 8 shows the raw GPS data and the four buffers. Once extracted, identifiers such as PersonID are removed and replaced with a unique ID for each trace through each buffer area (no connection between GPS records in different intersection buffers). Figure 9 shows the extracted GPS data. This resulting database is useful for evaluating the intersection vehicle dynamics while eliminating any potential PII issues. It should be noted that this approach is not limited to intersections, but any spatial extraction defined by the user. While not explicitly mentioned in VTL, users of this approach should ensure that extracted sections do not include trip end information or very small result sets.

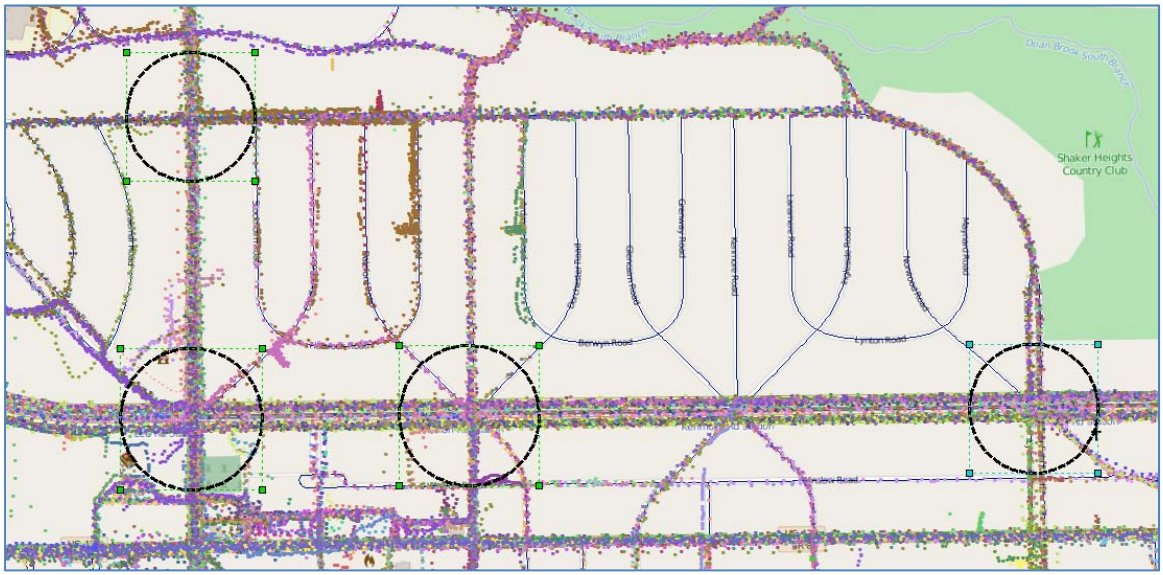

Figure 8 - GPSData sample with four selected intersections and 100-meter buffers

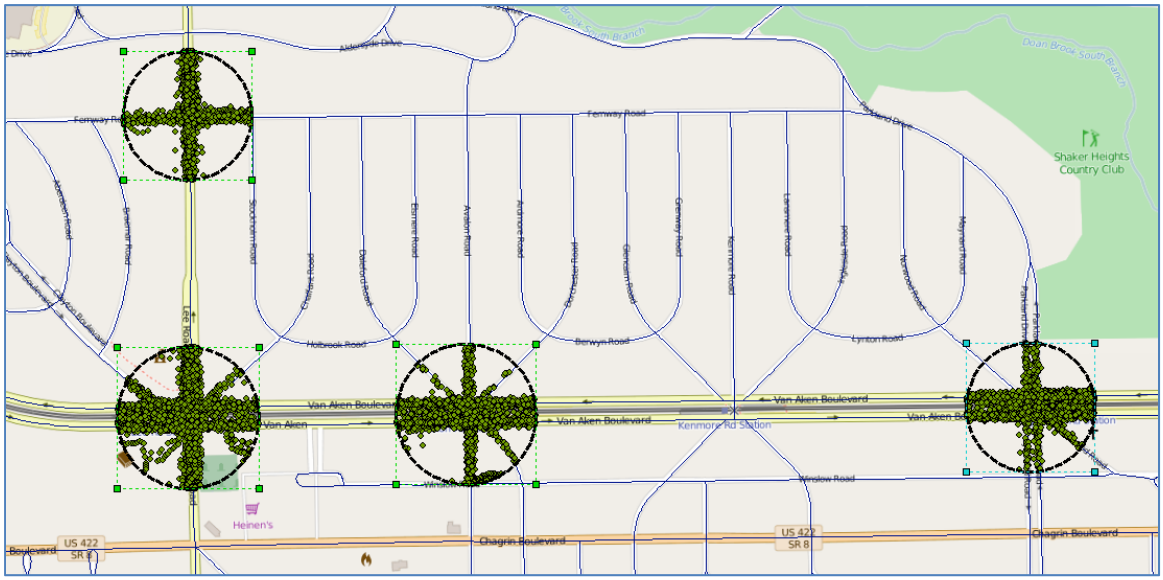

Figure 9 - VTL extracted GPSData usable for analysis

Risk and data quality impacts

Measures of the privacy impact depend on adversary models. Assume an intruder attempts to link traces between two VTL zones by accounting for travel time and speed of travel. The travel time can be estimated by the length between two zones divided by the average travel speed. The average speed is calculated based on the speeds reported in each pair of VTL zones. Knowing the departure time at an upstream zone, the intruder can estimate the arrival time at a downstream zone. The intruder may make an inference

that among all the traces entering the downstream zone, the trace with the closest arrival time belongs to the same vehicle. In the evaluation, Sun et al. (2013) computed the risk measures including the percentage of tracked traces among all released traces and the percentage of correct inferences among all inferences an intruder can make using the model.

The quality of released data can be evaluated with respect to whether the data are sufficient for the travel models in which the data users are interested. Sun et al. (2013) used the real-time queue-length estimation model (Ban et al., 2012). In general, the more data are released, the more successfully the travel models work. The metric (such as filtering rate or threshold) used in the filtering approaches should be chosen carefully to balance the privacy protection and data utility. In other words, the privacy risk needs to be kept at an acceptably low level while achieving the targeted success rate of travel models.

Minimum conditions

Prior to implementing the VTL-zone system and filtering approaches, experiments should be conducted to determine proper privacy metric using historical data. The determination of privacy level should account for not only the adversary models, but also different travel networks, geographic locations, data users, and the availability of historic or other external data. A limitation of the individual probability- and entropy-based approaches is that the filtering process targets the traces of high privacy risk. Without doing appropriate adjustment, the travel models that are based on the released data may be biased. The random filtering approach does not have this concern.

Process for applications

The filtering approaches in the VTL-zone system could be implemented on GPS and sensor data in the following steps.

1. Define the VTL zones.
2. Remove the traces between the VTL zones.
3. Remove the identifiers of the traces within the VTL zones and assign a random ID.
4. For a vehicle passing a given upstream VTL zone, model its probability of passing the neighboring downstream VTL zone, using travel time from one zone to the other.
5. Implement the three filtering approaches, respectively. The individual probability- and entropy-based approaches use the estimated probability in Step 2 to determine the privacy metric for filtering.
6. Evaluate the filtering approaches using the measures of privacy and data quality impact. Adjust the privacy metric, if necessary, to satisfy the desired level of privacy. Step 3 and Step 4 may need to be repeated a few times.

A full VTL algorithm can be found in Appendix C.

Further work on VTL

Further work may be done to evaluate the impact of individual probability- and entropy-based filtering approaches on the success rates of specific travel models, with and without adjustment to account for targeted filtering. Also, the privacy protection techniques can be tested and further improved using more sophisticated adversary models and data collected from different scenarios in the real world.

Table 3 lists the anonymization methods identified during the literature review and highlights those methods deemed the most valuable for further exploration based on four end-use scenarios. Those methods listed as "high" have significant relevance as an anonymization method. Likewise, "low" or blank methods are unlikely to be helpful in the designated use vase. The first two methods (Safe harbor protection and k-anonymity/l-diversity/t-closeness) are generic concepts that can be universally applied. They are valuable to all anonymization efforts. The remaining methods are targeted techniques for special data and end-use circumstances.

Table 3 - Relevance of protection methods for scenarios

Protection Method	Fixed-sensor travel time study	Fixed-sensor OD study	Mobile device traffic study	Mobile device travel behavior study
Safe Harbor identifier protection	High	High	High	High
K-anonymity/L-diversity/T-closeness	High	High	High	High
Geographic masking	Low	Medium	Medium	Medium
Spatial and temporal cloaking			Medium	Medium
"CliqueCloak"		Low	Low	Low
Path confusion			Medium	Medium
Location suppression	Medium	High	High	High
"Never Walk Alone"			Medium	Medium
"Time to Confusion"			Medium	Medium
Mix-Zone algorithm			Medium	Medium
Virtual trip lines (VTL)	High	High	High	High

Future Research Needs

This review of anonymization methods illustrated two important issues that transportation agencies should consider when collecting, analyzing, and sharing data. First, the guidance on the legal requirements for PII protection in spatial data is lacking. Many of the policies and methods conducted by agencies may be locally defined in the

absence of specific industry guidance. PII issues that are prominent in other disciplines (i.e., health research) are sometimes not strongly considered in transportation data.

The second issue is that maximizing data utility while simultaneously protecting PII requires the knowledge and implementation of specific technical methods. An approach to generate data in support of an OD analysis may be very different from an approach that targets segment-level speeds. Transportation agencies are encouraged to ensure that their approach targets the objectives mentioned in this research: protect Safe Harbor identifiers, use k-anonymity principles to protect pseudo-identifiers, and protect the sequences of behavioral data that can be connected to *k*-1 individuals. A generally accepted value *k* is 3, however, larger values offer more protection.

Mathematical solutions will continue to evolve in transportation as the issue of PII protection gains importance. Advances in connected vehicles and consumer-based passive data products suggest that microdata availability will continue to increase. This development is beneficial to the transportation community as it provides a wealth of observational information to evaluate performance and support forecasts. Future research in PII protection should expand on reviews of existing methods (as described in this document) and target mathematical solutions for specific datasets and end uses. This approach will allow the development of better alternatives as opposed to trying to force an existing method / solution that may have been defined for a different purpose.

Bibliography

Abedi N., A. Bhasker, and E. Chung, "Bluetooth and WI-FI MAC Address Based Crowd Data Collection and Monitoring: Benefits, Challenges, and Enhancement," Australasian Transport Research Forum 2013 Proceedings, October 2013, Brisbane, Australia.

Abowd J., F. Andersson, M. Graham, L. Vilhuber, and J. Wu, "Formal Privacy Guarantees and Analytical Validity of OnTheMap Public-Use Data," NSF-Census-IRS workshop on Synthetic Data and Confidentiality Protection, Suitland, MD, July 31, 2009. http://www2.vrdc.cornell.edu/news/wp-content/uploads/2009/08/1-5-Andersson.pdf.

Armstrong M., G. Rushton, D. Zimmerman, "Geographically Masking Health Data to Preserve Confidentiality," *Statistics in Medicine*, April 1999.

Baik H., M. Gruteser, H. Xiong, and A. Alrabady, "Achieving Guaranteed Anonymity in GPS Traces via Uncertainty-Aware Path Cloaking," *IEEE Transactions on Mobile Computing*, Vol. 9, No. 8 (2010).

Bamba B., L. Liu, P. Pesti, T. Wang, "Supporting Anonymous Location Queries in Mobile Environments with PrivacyGrid," *WWW*, pp. 237-246. ACM 2008.

Ban X. J. and M. Gruteser, "Towards Fine-Grained Urban Traffic Knowledge Extraction Using Mobile Sensing," Proceedings of the ACM SIGKDD International Workshop on Urban Computing (2012) pp. 111-117.

Beresford A. R. and F. Stajano, "Mix Zones: User Privacy in Location-Aware Services," Proc. IEEE Int'l Workshop Pervasive Computing and Comm. Security, *PerSec* 2004.

Carianha A. M., L. P. Barreto, G. Lima, "Improving Location Privacy in Mix-Zones for VANETs" In: Proceedings of the 30th International Performance Computing and Communications Conference (2011).

Cavoukian A., "Privacy by Design, Take the Challenge," Information and Privacy Commissioner of Ontario, Toronto (2009).

Clifton K. J. and S. R. Gehrke, "Wider Dissemination of Household Travel Survey Data Using Geographic Perturbation Methods," *OTREC* (January 2014).

Cox L. H., "Suppression Methodology and Statistical Disclosure Control," *Journal of the American Statistical Association,* 75, (1980) pp. 377-385.

Dahl M., P. Delaune, G. Steel, "Formal Analysis of Privacy for Vehicular Mix-Zones," In: Proceedings of the 15th European Conference on Research in Computer Security (2010).

Dandekar R., "Maximum Utility-Minimum Information Loss Table Server Design for Statistical Disclosure Control of Tabular Data," in Lecture Notes in Computer Science: CASC Project Final Conference, PSD 2004, Barcelona, Spain, June 9-11, 2004

Denning D. E. R., *Cryptography and Data Security*, Purdue University, Addison-Wesley Publishing Company (1982).

Dewri R., "Location Privacy and Attacker Knowledge: Who Are We Fighting Against?" Security and Privacy in Communication Networks: 7th International ICST (2008).

Douma F., S. Frooman, and J. Deckenbach, "What You Need to Know – And Not Know: Current and Emerging Privacy Law for ITS," In: Proceedings of the 87th Annual Meeting of Transportation Research Board (2008).

DPPA (The Drivers Privacy Protection Act), Public Law No. 103-322 (1994).

Duncan G. T. and R. Pearson, "Enhancing Access to Microdata While Protecting Confidentiality: Prospects for the Future (with discussion)." *Statistical Science* 6 (1991) pp. 219-239.

Elliot, M. "Disclosure Risk Assessment in Confidentiality, Disclosure, and Data Access: Theory and Practical Applications for Statistical Agencies," (P. Doyle, J. Lane, J. Theeuwes, and L. Zayatz, eds.), *Elsevier* (2001) pp. 75-95.

FCSM, *Report on Statistical Disclosure Methodology*, Statistical Policy Working Paper 22 of the Federal Committee on Statistical Methodology, 2nd version. Revised by Confidentiality and Data Access Committee 2005, Statistical and Science Policy, Office of Information and Regulatory Affairs, Office of Management and Budget. Available at: http://www.hhs.gov/sites/default/files/spwp22.pdf. (2005).

FHWA, *FHWA Strategic Plan*, (January 2014 revision), FHWA-PL-08-027.

Filgueiras J. "Sensing Bluetooth Mobility Data: Potentials and Applications," J. Filgueiras, R. J. F. Rossetti, Z. Kokkinogenis, M. Ferreira, and C. Olaverri-Monreal, *Computer-Based Modelling and Optimization in Transportation*, (2013) pp. 419-431.

Gedik B. and L. Liu, "Location Privacy in Mobile Systems: A Personalized Anonymization Model," In: Proceedings of the 25th IEEE International Conference on Distributed, Computing Systems, (2005) pp. 620-629.

Google Street View https://epic.org/privacy/streetview/ (2011).

Gruteser M. and D. Grunwald, "Anonymous Usage of Location-Based Services Through Spatial and Temporal Cloaking," In Proc. of USENIX MobiSys (2003).

Henschen D., "Mining WIFI Data: Retail Privacy Pitfalls," Information Week, (http://www.informationweek.com/mobile/mobile-applications/mining-wifi-data-retail-privacy-pitfalls/a/d-id/1315679) September 15, 2014.

HIPAA, "The Health Insurance Portability and Accountability Act of 1996 (HIPAA) Privacy, Security and Breach Notification Rules," HIPAA, Pub.L. (enacted August 21, 1996) pp. 104-191, 110 Stat. 1936.

Hoh B. and M. Gruteser, "Protecting Location Privacy Through Path Confusion," In: Proceedings of IEEE/Create-Net SecureComm, Athens, Greece (2005).

Hoh B., M. Gruteser, H. Xiong, and A. Alrabady, "Achieving Guaranteed Anonymity in GPS Traces via Uncertainty-Aware Path Cloaking," *IEEE Trans. Mob. Comput.,* 9 (8) (2010) pp. 1089-1107.

Hoh B., M. Gruteser, H. Xiong, and A. Alrabady, "Enhancing Security and Privacy in Traffic-Monitoring Systems," IEEE Pervasive Computing, vol. 5, no. 4, (Oct. 2006) pp. 38-46.

Hoh B., T.Iwuchukwu, Q. Jacobson, M. Gruteser, A. Bayen, J.C. Herrera, R. Herring, D. Work, M. Annavarum, and J. Ban, "Enhancing privacy and accuracy in probe vehicle based traffic monitoring via virtual trip lines", *IEEE Transactions on Mobile Computing*, (2012)11, 849-864

Hundepool A., J. Domingo-Ferrer, L. Franconi, S. Giessing, E. S. Nordholt, K. Spicer, and P. P. de Wolf, *Statistical Disclosure Control,* Chichester, UK: John Wiley & Sons (2012).

Katz vs. U.S., 389 U.S. 347 (1967).

Krenzke T., J. Li, and L. Zayatz, "Balancing Use of Weights, Predictions, and Locality Effects in a Model-Assisted Constrained Hot Deck Approach for Random Perturbation," in JSM Proceedings, Survey Research Methods Section. Alexandria, VA: American Statistical Association, (2013) pp. 1598-1612.

Krenzke T., J. Li, M. Freedman, D. Judkins, D. Hubble, R. Roisman, and M. Larsen, "Producing Transportation Data Products from the American Community Survey That Comply With Disclosure Rules," Washington, DC: National Cooperative Highway Research Program, Transportation Research Board, National Academy of Sciences (2011).

Krumm J, "Inference Attacks on Location Tracks," Proc. Fifth Int'l Conf. *Pervasive Computing* (May 2007).

Li N., L. Tiancheng, and S. Venkatasubramanian, "t-Closeness: Privacy Beyond k-Anonymity and l-Diversity" *ICDE* (2007) pp. 106-115. IEEE.

Liebig T. and A. U. Kemloh Wagoum, "Modelling Microscopic Pedestrian Mobility Using Bluetooth," in *ICAART* (2012) pp. 270-275.

Machanavajjhala A., D. Kifer, J. Abowd, J. Gehrke, and L. Vilhuber, "Privacy: Theory Meets Practice on the Map," in ICDE: Proceedings of the 2008 IEEE 24th International Conference on Data Engineering, Washington, DC: *IEEE Computer Society* (2008) pp. 227-286.

Machanavajjhala A., D. Kifer, J. Gierke, and M. Venkitasubramanian, "*l*-diversity: Privacy Beyond *k*-anonymity," *ACM Transactions on Knowledge Discovery from Data*, Vol. 1, No. 1, Article 3 (2007).

Massell P., M. Freiman, and L. McHenna, "Data Masking for Disclosure Limitation," Wiley StatsRef: Statistics Reference Online, John Wiley & Sons, Ltd (2015).

Montjoye Y.-A., C. A. Hidalgo, M. Verleysen, and V. D. Blondel, "Unique in the Crowd: The Privacy Bounds of Human Mobility," *Scientific Reports* 3 (2013).

National Research Council (NRC), "Putting People on the Map," Panel on Confidentiality Issues Arising from the Integration of Remotely Sensed and Self-identifying Data, *National Academy Press,* Washington, DC, (2007).

National Workrights Institute, "On Your Tracks: GPS Tracking in the workplace", Available at https://epic.org/privacy/workplace/gps-traking.pdf

NCHRP 775, "Applying GPS to Understand Travel Behavior," *National Cooperative Highway Research Program,* TRB, Washington, DC, (2014).

NCHRP W174, "Performance Measurement and Evaluation of Tolling and Congestion Pricing Projects," NCHRP web-only document W174, TRB, Washington DC, (2012).

NCHS, "NCHS Staff Manual on Confidentiality," *National Center for Health Statistics* (2004), Available at http://www.cdc.gov/nchs/data/misc/staffmanual2004.pdf.

Nergiz M. E., M. Atzori, Y. Saygin, and B. Guc, "Towards Trajectory Anonymization: A Generalization-Based Approach," *Transactions on Data Privacy* (2009) 2 (1), pp. 47-75.

NIST, "Guide to Protecting the Confidentiality of Personally Identifiable Information (PII)," Recommendations of the National Institute of Standards and Technology, U.S. Department of Commerce (2010).

Osman A., B. Francesco, and N. Mirco, "Never Walk Alone: Uncertainty for Anonymity in Moving Objects Databases," *ICDE*, pp. 376-385. IEEE 2008.

Pensa R. G., A. Monreale, F. Pinelli, and D. Pedreschi, "Pattern-Preserving k-Anonymization of Sequences and its Application to Mobility Data Mining," Proceedings of the International Workshop on Privacy in Location Based Applications (2008).

Privacy Act of 1974, 5 U.S.C. § 552a, http://www.justice.gov/opcl/privacy-act-1974.

Rainie L., "TRB Webinar: Tracking Public Perceptions Related to Data Privacy," 12/16/2014.

Riley vs. California, 573 U.S. ___ (2014).

Sun Z., R. Zan, X. Ban, and M. Gruteser, "Privacy Protection Method for Fine-Grained Urban Traffic Modeling Using Mobile Sensors," *Transportation Research* Part B, 56 (2013) pp. 50-69, Elsevier.

Sweeney L., "K-anonymity: A Model for Protection Privacy," *International Journal on Uncertainty, Fuzziness and Knowledge-Based Systems* (2002) 10 (5) pp. 557-570.

Sweeney L., "Simple Demographics Often Identify People Uniquely," Carnegie Mellon University, Data Privacy Working Paper 3. Pittsburgh (2000).

Tang K. P., P. Keyani, J. Fogarty, J. I. Hong, "Putting People in Their Place: An Anonymous and Privacy-Sensitive Approach to Collecting Sensed Data in Location-

Based Applications," In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (2006) pp. 93-102.

Terrovitis M. and N. Mamoulis, "Privacy Preservation in the Publication of Trajectories," In MDM'08: Proceedings of the Ninth International Conference on Mobile Data Management, pp. 65-72, Washington, DC, USA, 2008. *IEEE Computer Society*.

Title 45 CFR Part 46, http://www.hhs.gov/ohrp/humansubjects/guidance/45cfr46.html.

TomTom (2011), http://www.loc.gov/law/help/online-privacy-law/netherlands.php.

U.S. vs. Jones, 132 S. Ct. 945, 565 U.S. ___ (2012).

U.S. vs. Knotts, 460 U.S. 276 (1983).

Wasson J. S., J. R. Sturdevant, and D. M. Bullock "Real-time travel time estimates using media access control address matching" *ITE Journal* (2008) 78 (6), 20-23.

Zang H. and J. C. Bolot, "Anonymization of Location Data Does Not Work: A Large-Scale Measurement Study," In Proceedings of the 17th Annual International Conference on Mobile Computing and Networking (Las Vegas, Sept. 19–23). ACM Press, New York (2011) pp. 145-156.

Appendix A - Terrovitis Algorithm

The full published Terrovitis algorithm (Terrovitis, 2008) is shown in Figure 10 and includes a routine for estimating privacy breaches.

Algorithm 1 Anonymization algorithm

Anonymization(T, P_{br})
1: Construct projection DB T_i for each attacker $i \in \mathcal{V}$
2: *Breach_identification*(T, P_{br}) ▷ call Alg. 2
3: T':=T ▷ Initialize output database
4: **while** there are still privacy breaches **do**
5: $U(t_x^i, t_y^i)$:=lowest cost unification ($t_y^i \subset t_x^i$, t_x^i and/or t_y^i are/is problematic)
6: **for all** $t' \in S(t_x^i, T')$ **do** ▷ traj. in T' supporting t_x^i
7: Suppress all $p \in t'$, $p \in t_x^i$, $p \notin t_y^i$
8: remove t_x^i from T_i ▷ t_x^i not supported anymore in T'
9: *Breach_identification*(T', P_{br})

Algorithm 2 Identification of privacy breaches

Breach_identification(T, P_{br})
1: **for all** i in $\mathcal{V}$ **do** ▷ Initialization
2: **for all** $p_j \notin \mathcal{P}_i$ **do**
3: **for all** $t^i \in T_i$ **do**
4: $sup(p_j, t^i, T)$:=0;
5: **for all** t in T **do** ▷ DB scan to update counters
6: **for all** i in $\mathcal{V}$ **do**
7: t^i:=projection of t in T_i
8: **if** $t^i \neq \emptyset$ **then**
9: **for all** $p_j, p_j \in t, p_j \notin \mathcal{P}_i$ **do**
10: $sup(p_j, t^i, T)$++;
11: **for all** i in $\mathcal{V}$ **do** ▷ Initialization
12: **for all** $p_j \notin \mathcal{P}_i$ **do**
13: **for all** $t^i \in T_i$ **do**
14: **if** $\frac{sup(p_j, t^i, T)}{S(t^i, T)} > P_{br}$ **then**
15: Mark (p_j, t^i) as problematic

Figure 10 - Terrovitis algorithm

Appendix B - Time-to-Confusion Algorithm

An implementation of the time-to-confusion algorithm (Hoh, 2010) is as follows:

```
// Determines which location samples can be release
while maintaining privacy guarantee.
releaseSet = releaseCandidates = {}
for all vehicles v do
        if start of trip then
                v.lastConfusionTime = t
        else
                v.predictedPos = v.lastVisible.position +
                (t-v.lastVisible.time) _ v.LastVisible.speed
        end if

// release all vehicles below timeout
if t - v.lastConfusionTime < confusionTimeout then
        add v to releaseSet
else
        // consider release of others dependent on
        uncertainty
        v.dependencies = k vehicles closest to the
        predictedPos
        if uncertainty(v.predictedPos, v.dependencies) >
        confusionLevel then
                add v to releaseCandidates
        end if
 end if
end for

// prune releaseCandidates
for all v ∈ releaseCandidates do
        if ∃ w ∈ v.dependencies, w ∋ releaseCandidates
        ∪ releaseSet then
                if uncertainty(v.predictedPos, v.dependencies ∩
                (releaseCandidates ∪ releaseSet)) <
                        confusionLevel then
                                delete v from releaseCandidates
                end if
        end if
end for
repeat pruning until no more candidates to remove
releaseSet = releaseSet ∪ releaseCandidates

// release GPS samples and update time of confusion
for all v ∈ releaseSet do
        publish v.currentGPSSample
        v.lastVisible = v.currentGPSSample
        neighbors = k closest vehicles to v.predictedPos
        in releaseSet
        if uncertainty(v.predictedPos, neighbors) >¼ confusionLevel then
                v.lastConfusionTime=t
        end if
end for
```

Appendix C – Virtual Trip Lines (VTL) Algorithm

An implementation of the VTL algorithm (Hoh, 2012) is as follows:

Algorithm 1 Tripline Crossing Detection Algorithm

```
θ = thresholdToSwitchBadToGood
T = subsampling interval
for all GPS sample l do
  if PrevLocationFiltered is null then
    CurrLocationFiltered = LastGoodRefPoint = l;
    LastLocationUpdateTimestamp = l.t; goto TripLineChecking;
  end if
  TimeGap = l.t - LastLocationUpdateTimestamp;
  LastLocationUpdateTimestamp = l.t;
  if TimeGap is too large then
    LastGoodRefPoint = l; LastBadRefPoint = null; n = 0;
    CurrLocationFiltered = l; PrevLocationFiltered = null;
    goto TripLineChecking;
  end if
  Calculate speed against LastGoodRefPoint;
  if a vehicle has not moved far enough then
    LastBadRefPoint = null; n = 0; CurrLocationFiltered = null;
    goto TripLineChecking;
  else if speed glitch is true then
    Re-calculate speed against LastBadRefPoint;
    if speed glitch is false then
      if ++n is greater than θ then
        n = 0; LastBadRefPoint = null; LastGoodRefPoint = l;
        filteredLoc = SmoothingFilter(LastBadRefPoint, l);
        CurrLocationFiltered = checkReportingInterval(filteredLoc, T);
      end if
      goto TripLineChecking;
    end if
    LastBadRefPoint = l; goto TripLineChecking;
  end if
  n = 0; filteredLoc = SmoothingFilter(LastGoodRefPoint, l);
  LastBadRefPoint = null; LastGoodRefPoint = l;
  CurrLocationFiltered = checkReportingInterval(filteredLoc, T);
  // TripLineChecking
  if both CurrLocationFiltered and PrevLocationFiltered not null then
    traj = SetTrajectory(PrevLocationFiltered, CurrLocationFiltered);
    for all tripline j in each tile(i) do
      if tile(i).status is valid then
        triplineCrossed = CheckCrossing(tripline j, traj);
        if triplineCrossed is true then
          Compute speed and heading with traj for triplineMeasurement;
        end if
      end if
    end for
  end if
  if CurrLocationFiltered is not null then
    PrevLocationFiltered = CurrLocationFiltered;
  end if
end for
```

Figure 11 - VTL algorithm by Hoh, 2012