Encrypted Dataset Collaboration

Intelligent Privacy for Smart Cities

Isaac Potoczny-Jones
Tozny, LLC
Portland, Oregon
ijones@tozny.com

Erin Kenneally
U.S. Dept. of Homeland Security
Science & Technology
erin.kenneally@hq.dhs.gov

John Ruffing
Esri
Portland, Oregon
jruffing@esri.com

ABSTRACT

The past year has seen increasing scrutiny of Smart Cities efforts with regard to privacy. Privacy advocates have criticized Smart City data collection on the whole and critiqued specific city efforts that they feel have crossed a line.

Cities are struggling with a number of privacy issues, including how to address third parties' collection of Smart City data, how cities consume personally identifying information from third-parties, and how public records laws intersect with privacy concerns.

The majority of data that cities collect are subject to disclosure under public record laws, with an attendant obligation to anonymize sensitive private information. However, as the amount and availability of data increases, the ability to cross-reference, correlate, and de-anonymize or re-sensitize datasets also increases. This leads to re-identification attacks that infringe the privacy of individuals in those datasets, and fosters mistrust in city governments and technology vendors. A fundamental challenge is that open data and privacy interact in complex and unpredictable ways. Some cities may choose to allow third parties to collect and manage that data in an effort to encourage innovation in the delivery of city services, while simultaneously wrestling with the legal and policy implications, such as privacy and public records law compliance. Unfortunately, this also may have undesirable privacy outcomes depending on a third-party's use of that data and the city's role in encouraging its collection.

In this paper, we will discuss concrete approaches to smart cities data privacy governance including collection and management, and specifically, an innovative pilot project supported by the U.S. Department of Homeland Security, Science & Technology Directorate aimed at demonstrating how privacy technology can help harmonize data sensitivity risks with intended benefits.

CCS CONCEPTS

• Security and privacy → Cryptography; Systems security; Database and storage security; *Security services;*

KEYWORDS

cryptography, data privacy, smart cities, data trust, data fiduciary

ACM Reference format:

Isaac Potoczny-Jones, Erin Kenneally, John Ruffing. 2019. Encrypted Dataset Collaboration: Intelligent Privacy for Smart Cities. In The 2nd ACM/EIGSCC Symposium On Smart Cities and Communities (SCC '19), September, 2019, Portland, OR, USA. ACM, New York, NY, USA 7 pages. https://doi.org/10.1145/3357492.3358630

1 Introduction

Cities and local governments are not new to managing data in accordance with privacy and public records mandates. But on the whole, that data has either been non-sensitive on its face, clearly designated under open records disclosure exemption, or relatively straightforward to de-sensitize or redact prior to satisfying disclosure demands. This playbook is becoming outdated with the emergence of Smart City efforts generally characterized by increased IoT sensor and actuator devices, along with the associated digitization of behavior and information, and the resultant troves of "Big Data". These capabilities are driving new privacy risk for cities that raise questions around rights and obligations in the overall stewardship and management of data between and among citizens, governments, civil society, and private companies. The strategies that cities have used to classify and then manage public and open datasets will not work for datasets that have emergent sensitivities. We present an instantiation of the "Data Trust" solution, focused primarily on one prominent type of data encountered by Smart Cities- location data.

1.1 Challenges to Privacy in Smart Cities

There are a number of challenges cities face when managing private information in the context of emerging "smart" technology (city-level IoT). We anchor off of the wellspring of foundational aspects of data privacy which is beyond the scope of this paper, and set forth a handful of problems that are specific to cities as they engage smart technology-derived data for new and improved delivery of services.

It is no secret that sensitive information- be it confidential or privacy-relevant- is a target of interest by parties who may have no legitimate right to access or modify data for illegitimate or illegal purposes. While cybersecurity and privacy are not the same,

privacy cannot be maintained without security, so cities need to build new cybersecurity capabilities to control for the confidentiality and integrity of data.

Another challenge facing cities is in managing the requirements of public records laws as they intersect with privacy. In the US, the federal government and most cities have public records laws that require that data maintained by public entities be made available to the public… and often in electronic form. Most such laws have an exemption for private data, or a requirement that the data is redacted ("de-identified") prior to release of the containing record, but it is difficult to predict what data may be identifying. The released data can be aggregated, cross-referenced, and linked with any number of commercial and openly-available sources to re-identify the disclosed data or create new privacy sensitivities, such as with the compilation of digital dossiers by private companies. Governments engage an inexact analysis of balancing public interest in disclosure versus public interest in withholding. And in general courts interpret and allow the public interest exemption very narrowly. The significance for Smart Cities is that the reigning general presumption of disclosure (even for private information) will prevail, resulting in a negative privacy impact.

While the definition of "personally identifying information" (PII) and its kin are relatively well agreed upon, it is very difficult to predict how PII can be derived from datasets (especially large ones). This is a function of the fact that the disclosing entity does not know or control the growing spectrum of other data that may be available to cross-reference and re-identify seemingly benign data, and furthermore, advanced computational analytics is enabling new or exposing latent data sensitization. For instance, a dataset that includes two GPS locations, a start and an end point of a trip, might or might not indicate any private data. If those points start at a densely populated downtown street and end at another densely populated downtown street, it may be very hard to determine who the traveler is. On the other hand, if the starting point is an individual's home and the ending point is a cancer treatment center, that single trip can uniquely identify a person and perhaps a medical condition, when correlated with public information about those addresses. The relative density of identifiable objects (persons, places, things) is a significant factor affecting the likelihood of uncovering private information.

Another challenge is that technology vendors that are involved in Smart City efforts have an interest in monetizing data, but cities get criticism for facilitating access to those datasets. The cities might do so as a condition to use the services, as part of a development investment, or in order to avoid holding the data themselves. For instance, the City of Toronto has come under fire from privacy advocates (including one of their advisory panel's privacy experts) for partnering with Sidewalk Labs, a sister company to Google, due to their approach to data minimization or redaction and fears over how Google might use the data *[1]*.

1.2 Current Approaches to Protecting Privacy

There are numerous technologies, policy and hybrid approaches to addressing privacy in Smart Cities.

In some cases, cities may choose to not collect PII or reduce the amount of PII collected. This is a sound approach and is typically known as "data minimization" which roughly means that if you do not need to collect the data, then do not collect it. This mitigates a wide variety of security and privacy concerns since the data does not exist in the first place.

However, cities are increasingly interested in collecting data, including PII, because of its utility for informing policy with a goal of benefiting their constituents. This creates a tension since minimizing the collection of certain data can reduce the utility of data and the potential for personalization or localization of the policy (viz the attendant technology as is the case with Smart Cities) that would be impacted by that data. Less data could mean ill-informed policy for the constituents whose privacy is being protected. Similar to the data sensitivity prediction challenge, utility values are often hard to know a priori. As the geolocation example illustrates, sensitive data may not be explicit in the dataset, but rather, carry a latent privacy risk. In the end, minimization is an insufficient solution for Smart Cities' objectives.

For most companies and other organizations that collect PII, the most common and effective way to protect data is simply not to release it to the public. This prevents unauthorized parties from using it in ways that would impact data subjects' privacy. Due to public records laws, this is not an option available to most cities.

Anonymization, redaction, and de-identification are approaches that attempt to remove PII from datasets. Unfortunately, it has been shown repeatedly that for large datasets, especially datasets with many fields, re-identification is possible, and not necessarily even difficult *[2]*.

Privacy risk management approaches like NIST's PRAM *[3]* advocate for a trade-off analysis; identifying the risks of collecting and using the data and implementing mitigations to those risks. These mitigations can include data minimization, consent, encryption, or other protections.

For data collection of open datasets, it has been common for cities to provide open Abstract Programming Interfaces (APIs) or download points for datasets. Such datasets are typically open on the Internet and available for anyone. This approach is likely to be part of any solution for open data, and for private data, a similar approach with access control will likely become common, as it has for all kinds of data access across the Internet.

Differential privacy, a statistically sound approach to managing queries against private datasets, is in its early stages. Few practitioners and tools exist, and awareness and usability, let along understanding of the techniques is not widespread.

Another emerging approach is that of a data trust or data fiduciary. This legal and policy framework approach contemplates a first or third party entity (apart from the data subject), governed by an approved set of legally-enforceable obligations, to manage the data. It is akin to a collective power of attorney for data or public land trust. While data trust definitions and implementations vary, they all are predicated on contractual relationships steeped in data stewardship policy that include privacy and security requirements *[4]*.

2 Identifiability of Location Data

Location data and travel data is difficult to anonymize, and determining whether the data is effectively anonymized is also a significant challenge. Previous research has shown that an individual's travel path, even with reduced precision, is often unique *[5] [6]*. For instance, while thousands of people may travel on the same roads, a single individual can be isolated in a data set via the specific set of roads and the specific order of travel.

For travel location, start and end coordinates can be identifying to a specific individual. For instance, a trip that starts at a person's home and ends at a person's place of work can be uniquely attributed to a specific person by looking up the address at the origin.

Human location data can be reduced in accuracy, making it hard to correlate data sets, but it is difficult to determine how much accuracy identifies individuals. For instance, if the GPS is configured to only report a specific block, this may preserve individual privacy in a crowded city, but in a sparsely populated area, such an individual may be the only person in the GPS radius.

In 2019, a cycling and running app location tracker released a dataset of anonymized user locations. Some of the users in that dataset were US Department of Defense personnel who were in a secret facility, which was revealed in maps of the dataset. In this case, population density or user density was so low that those facilities could be identified for lack of other data points nearby *[7]*.

Each of these issues is exacerbated, again in a very unpredictable way, by the availability of other datasets that can re-identify individuals. In one study, privacy researchers were investigating the sensitivity of a dataset with taxi trips that included pick-up location, drop-off location and time, but no names. The authors of the study found photos on the internet of celebrities getting into taxi cabs and used metadata from those photos to match up with starting times and locations in the taxi dataset. Armed with the knowledge of the celebrities based on their appearance, the metadata from the photos, and the taxi dataset, researchers found the drop-off location of those individuals as well as their home addresses; information they had gone to some lengths to protect *[8]*.

The lesson from the taxi dataset, among many other academic studies of differential privacy, is that any large dataset can expose PII when combined with other datasets. There are many examples of such re-identification, and it is well understood among privacy researchers, and some policymakers, that no practical and simple solutions exist.

These correlation and re-identification challenges are not limited to publicly released datasets. Recently, Google acquired a dataset of "anonymized" medical records from the University of Chicago Medical Center. A lawsuit alleges that Google has the ability to correlate these datasets with other information at its disposal, such as location tracking data from mobile devices, thereby re-identifying individuals and circumventing health-care related privacy regulations like HIPAA *[9]*.

2.1 Case Study: Location Data Opportunities

Human location and other geospatial analysis is an important method for cities to make policy decisions. This can include zoning analysis, transportation planning, and behavioral data involving citizen-consumer movement- all relevant to Smart Cities' remit to improve delivery of services and be responsible stewards of tax payers' funding.

This human location data can be produced by city busses with integrated GPS units, ridesharing services via the drivers' and passengers' phones (e.g., Uber Lyft), bike share and scooter services with integrated GPS units (e.g., Lime, BikeTown), and even apps collecting location data on mobile phones (e.g., Google Maps, Apple Maps, Strava), which can be shared with cities.

In some cases, location data indicates the start and end points of trips, waypoints along travel routes, current locations of persons, or the locale of stationary objects like scooters parked on sidewalks. Technology defaults for location measurements can be very precise or intentionally reduced in precision in order to increase privacy protections.

Cities are interested in using location data in a variety of ways:

- To shape demand for traffic and parking, and ensure safety, by encouraging modes of travel that are more effective for those goals.
- To manage curbs and sidewalks for instance for parking cars, shared scooters, and bicycles in appropriate spaces.
- To ensure that companies make products and services available equitably, as well as providing city services to lower income areas and individuals with disabilities.
- To govern micro-mobility no go areas for regulating where they are allowed and their speed for particular areas.
- To understand where to spend money on infrastructure.

Traffic and parking of single occupancy vehicles is extremely costly and difficult to scale as city population sizes increase. Cities see the use of human location and transportation mode data as a potential boon for discovering ways to reduce costs by shifting

demand to more efficient transportation modes and cultivating community, sustainable growth, and improve urban living.

2.2 Examples

Next we offer a few representative examples of human location data generation and use for city planning purposes across Smart Cities.

In the Portland, Oregon metro area, TriMet manages buses, light rail trains, and other services. TriMet collects GPS data from their buses including passenger pick-up and drop-off stops and other types of transportation activity and measurements such as rider count data. This data is used for a wide variety of services, primarily to manage demand for TriMet services *[10]*.

Ride sharing companies like Uber and Lyft are required to share data with the city of Portland, including trip origin and destination GPS latitude and longitude. The city code requires that data should be anonymized prior to providing it to the city *[11]*. (The authors note that the city code requires both GPS data and anonymization, but we do not know whether this apparent contradiction is addressed in practice and if so how.) The city uses this data to manage these companies' compliance with various regulations. A similar effort in the city of Los Angeles, CA engendered strong push-back from privacy advocates like the Electronic Frontier Foundation, which argued that the data should be handled like PII and not collected without more privacy protections *[12]*. In the city of Austin, TX similar data is collected, but the location accuracy is reduced in order to reduce privacy exposure *[13]*.

3 Proposed Approach

In this section we outline a model solution that addresses this Smart City privacy and utility data challenge, based on a pilot project funded by the Department of Homeland Security (DHS), Science & Technology Directorate, which at the time of article submission is in the design and analysis stage. A significant amount of technology and social capital has already been developed to implement this model, with an anticipated roll-out of technology on an initial set of data and users in the Fall of 2019.

3.1 Policy Approach: Data Trust

Our pilot approach is an innovative, real world implementation of a "data trust" solution for responsibly managing the rights, needs, and obligations between citizens, governments and private companies related to data in smart cities *[4]*. Metaphorically it is a container that can hold data assets, define terms and conditions, and manage oversight of data by technological stewardship of data access and use rights and needs between smart city data stakeholders. This model provides a repeatable, scalable framework to govern and manage shared resources… one that ensures that data needed for public services is stewarded in accordance with privacy and accountable innovation.

For instance, a city government could establish a data trust to receive data collected by vendors and the city itself according to an established set of rules for collection, minimization, and appropriate use. In order to anticipate future advancements in reidentification and privacy technologies, the city could establish principles about the balance between public access and privacy, and legally bind the data trust to those principles. The data trust would steward the technological enforcement of privacy rights and obligations in the management of data between citizens, governments, and private companies. This protects privacy along three common dimensions but in different ways:

- **Organization**: The organization that convenes/stewards the data is legally accountable for its proper access and use. This removes a potential for conflict of interest where technology providers or other parties use data outside of its intended scope and purpose.
- **Policy**: For cities and other governments, data trusts can address the balance between governments' public interest obligation to provide access to public records and the right to privacy.
- **Technology**: The pilot platform uses technology to enforce privacy policies and minimize sensitive data exposure ex ante. Technology can be used to ensure that the legal framework and the implementation of policy are in synch. As outlined below, our approach uses access control and cryptographically-enforced privacy rules, which can be aligned with the legal policy framework to enforce policy, detect violations of policy, and provide transparency to the public that no such violations have occurred.

3.2 Example Use Case

In this section, we describe our approach in the context of a specific but generalizable scenario and a reference technical implementation.

3.2.1 Scenario. By way of example, several mobility companies may produce datasets with GPS coordinates reflecting trips that users take when utilizing those companies' services. As outlined above, the location data itself will almost certainly trigger privacy risk since some subset of the trips can be correlated with other data inputs like individuals' home addresses. This linking can be done by any party with access to the data, whether it's publicly released or not.

For the mobility companies this location data is linked to the user's account, which typically includes their name and some contact information. Some mobility companies collect photographs of the user's driver's license in order to verify their identity and ability to operate a motor vehicle. Account information therefore contains PII that can include license number phone number, address, email, a photograph of the user, height and weight, and birthdate, among others.

By way of example, a city may require that the mobility company provide trip data as a condition to operate their mobility service within the city. This trip data can be used by the city to help with last-mile planning, to ensure equitable distribution of mobility options to under-served parts of the city, or for other legitimate planning purposes.

Mobility companies have an interest having this data protected: If the city leaks the data by way of a cyber attack or due to a public records request, or by sharing it with law enforcement, the privacy of the company's users could be negatively impacted, and their reputation and/or ability to engage in future business damaged. One would expect that the city, not the company would be held accountable, but it is notoriously difficult to determine the root cause of cyber attacks. If and by the time the forensics could attribute the leak, the fact that the leak involved the mobility company's data is enough to incite negative blowback for the company. This damage would occur even if the company could not have prevented it in the case where it was the city's systems that were compromised.

Furthermore, the mobility companies have confidentiality interests in the intellectual property within the data. If details of their services, customer base, how many vehicles were on the road, etc., were revealed to competitors it could harm their competitive advantage. If some of their competitors are subject to different regulations (e.g. taxi company datasets may be handled differently from scooter datasets) the information disclosures would be asymmetric. One party would have knowledge of the other party's business, but not vice-versa, resulting in unfair competition.

The need is a solution that applies privacy protecting technology that is informed by the policy needs of Smart Cities and similar jurisdictions.

3.2.2 Approach to Scenario. Our pilot model addresses user privacy rights, technology vendor intellectual property and reputation interests, and cities' data planning analysis and services obligations. At a high-level, the policy-informed technology approach to data and information flow is as follows:

1. The encrypted location *dataset* is provided by the mobility company to the data trust in a particular format that enforces the policy (more details on this below).
2. The city provides their analytic *computation* to the data trust in a particular format that allows for automation. This can include geospatial analysis, statistical analysis, or any type of computation that would benefit from the data.
3. The data trust applies policy rules to the dataset. This could include rules provided by the user themselves (e.g. consent or lack thereof), provided by the city, or provided by the vendor. This policy check happens before executing the computation in a secure environment. By way of example, the policy could exclude:
   - Specifically identifying information such as email addresses.
   - Users who have not consented to having their data analyzed for this purpose.
   - Computations that extract and exfiltrate PII.
4. The results of the computation are returned to the city. The following diagram illustrates this approach:

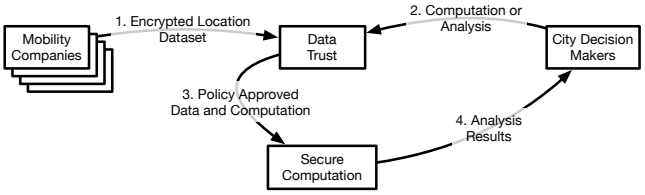

Figure 1: Encrypted data collection and secure computation to protect privacy and IP while allowing city planning.

Since the data is decrypted during computation, potentially exposing it to cyber attack, practical mitigations to protect the data are required. These can include multi-party control of encryption keys, hardware-enforced secure enclaves (e.g. with systems like TrustZone or SGX), virtualization such as what is implemented by modern cloud providers, and software containerization.

This technical data trust approach is a significant improvement over the current and common practice of data sharing (e.g. in smart cities, scientific research, and collaboration between companies). Typically datasets are furnished directly to data consumers who perform analysis in their own environments, and thus in the case of smart cities, on the policy and the security of the city's environment. In current practice, when there are multiple consumers, the data is moved to multiple environments, and the risk to privacy and security of that data is the risk in the least protected environment. In contrast, the goal of the data trust would be to establish a shared and trustworthy "high bar" for data management.

This model cannot address every potential privacy or security issue, and in the following sections, we summarize the strengths and weaknesses of this approach.

3.2.3 Benefits of this approach. This model addresses a number of issues that are well beyond standard security and privacy practices. The following classes of problems, associated with publicly-known data privacy breaches, are addressed:

- **Exclusion of PII from Data**: In many cases, the technology platform can enforce data exclusions in a manner that no party can bypass, including the data trust. For instance, the user's email address can be encrypted separately and no key provided to the data trust or computation.
- **Use of PII Beyond Consent**: Users benefit from the data encryption to enforce that their data cannot be used if they do not consent to it.
- **Unintentional or Accidental Data Leaks**: Data cannot be leaked from the city's environment since it does not reside

there. While in the data trust environment, the data remains encrypted at all times except during computation.

- **Breach of IP confidentiality:** The datasets are not released outside the scope of the computation required, which prevents other organizations from obtaining access to the data.
- **Compromised Data trust:** Even if the data trust is malicious or compromised, they cannot decrypt or access the data.
- **Utility needs:** This approach still maximizes utility of the data by allowing authorized parties to compute and analyze the data, and receive the results of that computation, without risking the data itself.

3.2.4 Concerns with this approach. In limited instances, it is still possible to circumvent the privacy protections. These are neither new attacks nor are these attacks made worse by this approach, but they cannot be fully mitigated:

- **Trojan Computations**: The city could construct computations that expose PII. For instance, if the computation intentionally identifies trips starting at a particular individual's address and analyzes the locations they went to, the analysis could infringe individual privacy expectations.
- **Compromised Secure Computation:** The data is decrypted during computation, and a successful cyberattack against the data at this point could expose sensitive data upon which those analytics are running. Note that when compared to common practice, this approach does still reduce the privacy attack surface. The data is encrypted at all points except during computation, so there is only a single potential point of compromise.

3.2.5 Potential Risk Mitigations. In light of the residual privacy vulnerabilities above, the following approaches will be considered to mitigate/prevent these risks:

- **Inspection to Prevent Trojan Computations:** The computation itself could be proactively or retroactively audited to prevent leak of data or identify it after the fact. For instance, the mobility companies could require that they approve the code.
- **Differential Privacy to Mitigate Trojan Computations:** Differential privacy is a statistical approach to ensure privacy of analytic query outputs that prevents identification of specific individuals in a dataset. This approach is not appropriate for the analytic context of the pilot.
- **Homomorphic Encryption to Mitigate Compromise**: Homomorphic encryption performs computation on encrypted data, reducing the attack surface further. This is an advanced technique with few pragmatic applications due to high computational overhead costs (i.e., it is slow for large-scale datasets). This approach is not a good fit for the data needs and privacy risk involved in the pilot.

3.3 Details of Technical Approach

The reference implementation of this model engages a privacy platform developed by Tozny, based on advanced privacy technology development support from DHS, NIST and DARPA.

The privacy platform employs *end-to-end encryption*. This is a security approach in the field of cryptography where data is encrypted at the point of its creation and only decrypted (if at all) at its point of use. Contrast this with more typical approaches to encryption like *data in transit* (DIT) or *data at rest* (DAR) encryption which secure data at a point in time, but not for its entire lifecycle. For instance, data transmitted over HTTPS is typically decrypted immediately when it's received, so the receiver or an attacker targeting the receiver can see the plain text data. Such an attack is not possible with end-to-end encryption.

3.3.1 Data Production and Encryption. The data producer is the party that creates or originates the data, such as the end user or the mobility company. Data is encrypted by way of a Software Development Kit (SDK) or library embedded into the end user application or the server of the mobility company. Tozny's SDK is available for a wide variety of programming languages.

Applying encryption as early as possible in the data lifecycle ensures that the data enjoys the protection of encryption from its point of creation to its point of use. It is therefore secure across the various networks, servers, mobile devices, and backup systems that it may reside on.

3.3.2 Encryption Standards and Access Control. While the use of encryption for privacy and consent is innovative in its application (many models are either theoretical or suffer from usability challenges), the encryption itself utilizes standard algorithms and industry best practices. In particular, the data itself is encrypted with fast symmetric algorithms, and the key for that data is encrypted with an asymmetric cipher (also known as public / private keys). Each party who is authorized to access data only needs to maintain his or her own private key or passphrase, and the software development kits make this easy by leveraging local biometrics and key management capabilities that are built into modern consumer devices and cloud service platforms. All of the intermediate keys can be decrypted with that private key, allowing for simple key management and efficient encryption.

3.3.3 Cloud API Storage. Once the data is encrypted, it is transmitted to a cloud API for storage. This system allows for collection of large amounts of data, redundancy, security, backups, and access control. The following diagram depicts the major components of the cloud-based encryption platform:

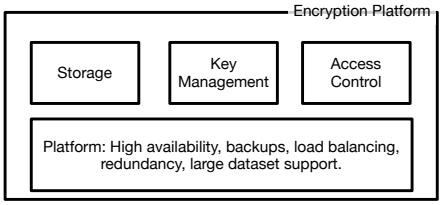

Figure 2: Major components in encryption platform

3.3.4 Data Compatibility and Interoperability. Data can be generated or constructed in structured and unstructured data formats, such as JavaScript Object Notation (JSON), which is a standard approach for modern web interfaces. This allows for compatibility and interoperability across a wide variety of systems that can easily read and write JSON. Unstructured data like images, documents, or other types of files can also be stored and retrieved from the system. Both structured and unstructured data can be tagged with metadata to allow for SQL-type queries.

3.3.5 Data Flow Overview. The following diagram depicts the data flow throughout the system:

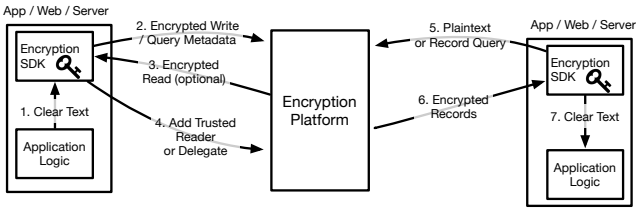

Figure 3: Encryption dataflow high-level overview

1. The data producer function can be a user mobile app, a web server, or any software system that originates the data. This application includes the encryption SDK. The application logic sends structured JSON data in clear text to the SDK, which encrypts the data.
2. The encryption SDK sends the encrypted data to the encryption platform with optional metadata that can be used in queries.
3. The data producer can also optionally read that data.
4. When policy permits, the data producer can add a *trusted reader* to the data. The addition of a trusted reader is accomplished via performing a cryptographic key exchange. Alternately, the data producer can *delegate* this authority to another human or software system to make policy decisions on their behalf.
5. The data consumer can be a mobile app, server, or other software system. In our example, the data consumer can represent the secure computation platform. The data reader can query the encryption platform for subsets of the data that they are interested in (and that policy allows them to access).
6. After enforcing policy, the encryption platform sends the encrypted records to the data consumer.
7. The application logic itself (e.g. the computation) can now decrypt and access the data in clear text.

This approach reduces the need to trust various parties to follow privacy policy since the data itself is not available in plain text.

4 Conclusion

Our pilot implementation of the increasingly-popular data trust model allows multiple stakeholders with different interests in Smart City data to achieve responsible innovation, be they technology providers and platforms, governments and citizen-users. Although this approach is technology-enforced and policy-driven, it allows for agility in defining, negotiating and enabling a balancing of interests and priorities so that mutual benefits can be realized from the collection and use of sensitive data in Smart Cities. This technical framework enforces personally identifiable data minimization, access control, and flexible and granular disclosure/redaction control, all with provable accountability and in conjunction with legally-enforceable obligations and responsibilities for data use. This pilot is demonstrating how Smart City data can be engaged productively without undermining individuals' privacy, public interest in government data, and industry's drive toward innovation.

ACKNOWLEDGMENTS

This material is based on research sponsored by DHS, Science & Technology under agreement number FA8750-19-C-0069. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of DHS S&T or the U.S. Government.

REFERENCES

[1] L. Cecco, "'Surveillance capitalism': critic urges Toronto to abandon smart city project," 6 June 2019. [Online]. Available: https://www.theguardian.com/cities/2019/jun/06/toronto-smart-city-google-project-privacy-concerns.

[2] L. Rocher, J. M. Hendrickx and Y.-A. de Montjoye, "Estimating the success of re-identifications in incomplete datasets using generative models," *Nature Communicationsvolume ,* vol. 10, no. Article number: 3069, 2019.

[3] S. Brooks, M. Garcia, N. Lefkovitz, S. Lightman and E. Nadeau, "NISTIR 8062: An Introduction to Privacy Engineering and Risk Management in Federal Systems," January 2017. [Online]. Available: https://nvlpubs.nist.gov/nistpubs/ir/2017/NIST.IR.8062.pdf. [Accessed 2019].

[4] B. Wylie and S. McDonald, "What Is a Data Trust?," 9 October 2018. [Online]. Available: https://www.cigionline.org/articles/what-data-trust.

[5] Y.-A. de Montjoye, C. A. Hidalgo, M. Verleysen and V. D. Blondel, "Unique in the Crowd: The privacy bounds of human mobility," *Scientific Reports,* vol. 3, no. Article number: 1376, 2013.

[6] D. Kondor, B. Hashemian, Y. de Montjoye and C. Ratti, "Towards matching user mobility traces in large-scale datasets," in *IEEE Transactions on Big Data doi: 10.1109/TBDATA.2018.2871693.*, 2018.

[7] A. Hern, "Fitness tracking app Strava gives away location of secret US army bases," 28 January 2018. [Online]. Available: https://www.theguardian.com/world/2018/jan/28/fitness-tracking-app-gives-away-location-of-secret-us-army-bases.

[8] A. Tockar, "Riding with the Stars: Passenger Privacy in the NYC Taxicab Dataset," 15 September 2014. [Online]. Available: https://research.neustar.biz/2014/09/15/riding-with-the-stars-passenger-privacy-in-the-nyc-taxicab-dataset/.

[9] L. Schencker, "How much is too much to tell Google? Privacy lawsuit alleges U. of C. Medical Center went too far when sharing patient data.," 27 June 2019. [Online]. Available: https://www.chicagotribune.com/business/ct-biz-lawsuit-university-of-chicago-google-patient-records-20190627-4vnmvfdnv5gcdl5fakgp5zwtna-story.html.

[10] TriMet, "TriMet Developer Resources," [Online]. Available: https://developer.trimet.org/.

[11] Portland City Code, *16.40.140 - Taxi Company Operating Responsibilities and Prohibitions.,* 2018.

[12] N. Sheard, "The Los Angeles Department of Transportation's Ride Tracking Pilot is Out of Control," 9 April 2019. [Online]. Available: https://www.eff.org/deeplinks/2019/04/los-angeles-department-transportations-ride-tracking-pilot-out-control.

[13] M. Clarke and J. Clary, "Explore Dockless Data with Austin Transportation," 14 January 2019. [Online]. Available: https://medium.com/civiqueso/explore-dockless-data-with-austin-transportation-4a308aa5c18.