SafeGraph

Threshold Privacy Analysis

Completed

Smart City PDX

11/05/2021

THRESHOLD PRIVACY ANALYSIS REPORT

City of Portland Privacy Toolkit

WHAT IS THE THRESHOLD PRIVACY ANALYSIS?

The Threshold Privacy Analysis (“TPA”) is a method to quickly evaluate what are the general privacy risks of a technological solution or a specific use, transfer or collection of data to City bureaus or offices. The TPA is a way to identify factors that contribute to privacy risks and lead to proper strategies for risk mitigation or alternatives that may even remove those identified risks.

The Threshold Privacy Analysis may lead to a more comprehensive Privacy Impact Assessment (PIA) and a Surveillance Assessment depending on the level or risks identified and the impacts on civil liberties or potential harm in communities.

In the interests of transparency about data collection and management, the City of Portland has committed to publishing all Privacy Assessments on an outward facing website for public access. TPAs do not include specific uses of technology or data other than those initially evaluated.

WHEN IS AN THRESHOLD PRIVACY ANALYSIS RECOMMENDED?

A TPA is recommended when:

- A project, technology, data sharing agreement, or other review has been flagged as having some privacy risk due to the collection of private or sensitive data.
- A technology has high financial impact and includes the collection, use or transfer of data by city bureaus or third parties working for or on behalf of the city.

HOW TO COMPLETE THIS DOCUMENT?

City staff complete two documents:

- *The Threshold Privacy Analysis form.* This document identifies all important information related to the project description, data collection, use, safekeeping, and management; as well as a verification of existing privacy policies and measures to protect private information.
- *The Privacy Risk Assessment.* This document breaks the privacy risk into six different areas of evaluation: (1) Individual Privacy Harms; (2) Equity, Disparate Community Impact; (3) Political, Reputation & Image; (4) City Business, Quality & Infrastructure; (5) Legal & Regulatory; and, (6) Financial Impact. Then compares risks to the likelihood of happening to create a single risk measure based on the worst case scenario.

Executive summary

Portland Parks and Recreation is planning to use aggregated geolocation data to understand how many visitors are in parks, when and for how long they visit a park, and how far they travel to get to a park.

SafeGraph's Patterns dataset includes visitor and demographic aggregations for points of interest (POIs) in the US. This service uses aggregated raw counts of visits to POIs from a panel of mobile devices, answering how often people visit, how long they stay, where they came from, where else they go, and more. Data is aggregated by month, no IDs, no PII is collected. The main goal is to know how many people visited a park with a minimum threshold of 4.

The assessment risk level is **MEDIUM** due to:

1. Potential equity impacts due to individuals without an electronic device recorded by source data services not being counted. These individuals may include low income people and those experiencing houselessness, as well as children. Data from this service could be biased and additional ground truth is required.

2. Political and reputation risks due to the collection and use of geolocation data with privacy rights issues. These issues include lack of transparency on how data has been collected from people and lack of meaningful consent from users from whom data has been collected. In this application, the City, as a consumer, has only access to anonymized and aggregated data; which minimizes risks for re-identification, but raises the ethical issue of whether to become part of a system that impacts individuals privacy and extracts information without user awareness or consent.

3. Impacts on City business due to using potentially biased data in Portland Parks and Recreation's businesses, services, or decisions around infrastructure. Individuals without tracking devices may not be counted, staff may be overcounted, which may need further studies to understand this impact. The main impacts are likely to happen in BIPOC and communities with a high number of people living in poverty. This situation may include hidden operation costs that may have not been considered.

In order to mitigate these issues, these actions are recommended:

- Inform City staff and users of this data about its bias and potential impacts in reporting and decision making.
- Try to Inform the public about ways in which they can better protect personal privacy when using Portland Parks and Recreation facilities.
- Validate data with ground truth with Portland Park and Recreation's own data.
- Make data available through open data.

The Threshold Privacy Analysis section includes a recollection of information about the application, privacy policy and services offered by the company, and additional

comments connected to the City's Privacy and Information Protection Principles.
The privacy impact risk severity assessment section describes and comments on the issues found in this assessment.

This report is intended to be informative. This report provides an analysis of impacts and risks only for this specific project and context. All recommendations are intended to be used as a reference and inform decision makers and the public. The report has no legal value.

Threshold Privacy Analysis

Risk and impact level	**Medium**

	Portland threshold privacy analysis for a technology, project, data sharing agreement or app solution
Version 0.3	**This information is considered restricted and for internal use only until the client clears it for public release. This notice must be remove when authorized for publication**
Information	**Request information**
Bureau	Portland Parks & Recreation Asset Management
Prepared by (name/email)	Hector Dominguez - hector.dominguez@portlandoregon.gov Privacy work group coordinator
Reviewed by	City Attorney's office Office of Equity and Human Rights
Date of Assessment	October 26, 2021
Document status	**Delivered**
Date of Acceptance	**10/19/2021**
Name of the assessment	**Parks Usage Statistics from SafeGraph Aggregated Data**
General description	Portland Parks & Recreation is planning to use aggregated geolocation data to understand how many visitors are in parks, when and for how long they visit a park, and how far they travel to get to a park.
Evaluation topic	**Assessment**
Purpose of the technology, project, data sharing or application	SafeGraph's Patterns dataset includes visitor and demographic aggregations for points of interest (POIs) in the US. This contains aggregated raw counts of visits to POIs from a panel of mobile devices, answering how often people visit, how long they stay, where they came from, where else they go, and more. Data aggregated by month, no IDs, no PII is collected. The main data question to answer is, how many people visited a park? With a minimum threshold of 4.

Name of the entity owner of the application and website	SafeGraph builds and maintains data sets to help accelerate machine learning and AI. Its high-precision places data covers business listings, building footprints, and aggregated foot traffic data for millions of points of interest (POI) and thousands of brands in the US, UK, and Canada. With detailed brand affiliation, spatial hierarchy, and other key business details, SafeGraph data informs market analytics, investment research, site selection, and more. http://www.safegraph.com/ SafeGraph's places dataset includes a breadth of information about physical places in the US, UK and Canada. This includes core location data, spatial hierarchy metadata, place traffic data, and more. The schema is defined here https://docs.safegraph.com/docs/core-places SafeGraph's Patterns dataset includes visitor and demographic aggregations for points of interest (POIs) in the US. This contains aggregated raw counts of visits to POIs from a panel of mobile devices, answering how often people visit, how long they stay, where they came from, where else they go, and more. https://docs.safegraph.com/docs/monthly-patterns
Type of Organization	Private entity
Scope of personal data collected. List all sources of data and information.	SafeGraph collects the following Information: Mobile ad identifiers, primarily Apple iOS IDFAs or Google Android IDs; The precise geographic location of a device at a certain time, usually expressed in latitude/longitude coordinates along with a timestamp; The horizontal accuracy of the latitude/longitude coordinates. SafeGraph may sometimes also collect: Phone carrier and connection type (e.g., cellular, wifi); The direction in which the device is traveling as a degree coordinate and speed of device; Information about a device such as device type and model and OS type and version; Device language; Public facing (external) IP address of device; Available or connected wifi SSID and/or BSSID names; Altitude and vertical accuracy; Whether device is charging; Bluetooth connected/available devices; Beacon proximity or any beacon metadata; Name of the mobile app the consumer was using during data collection;
How personal data is collected	SafeGraph obtains a variety of information from trusted third-party data partners such as mobile application developers and companies that aggregate information from those developers' mobile apps.

	SafeGraph works with data companies to offer more than 1,000 additional data attributes on places. SafeGraph may collect certain demographic data, but that is not collected by the City. Demographic data may increase risks and impacts and it is important to disclose what specific data is used.
Who can access the data	SafeGraph customers – a variety of companies and organizations. SafeGraph staff and developers have access to data in order to provide the service and develop products, for operation and internal purposes. Authorities and third parties for legal, auditing and accounting purposes.
Purposes the data is used for	SafeGraph customers – a variety of companies and organizations – use the Information collected by SafeGraph for a variety of commercial and research purposes, including but not limited to, ad targeting (for instance, building models of inferred audience preferences), foot and vehicle traffic analysis (for instance, tracking which parts of a city or neighborhood are most busy, at what times), retail site selection (for instance, determining where to open a new restaurant) and market research (for instance, tracking consumer shopping trends based on foot traffic concentration). SafeGraph may also use the Information to develop derivative products, such as determining whether a certain device visited a certain retail store at a given time. SafeGraph may also use the Information for our internal and operational purposes, such as to consider or make internal service improvements or quality checking, or for the company's sales and marketing purposes, and more generally to operate, maintain and improve the services we offer. SafeGraph may also use and share the Information for legal, auditing and accounting purposes, such as (a) in good faith compliance with a request from law enforcement or a governmental agency, including law enforcement, (b) protect or enforce our rights or those of others, (c) to evaluate or enhance the security or quality of our Information, or (d) to investigate potential wrongdoing. Likewise, in the event of any potential merger or acquisition, any Information we hold (including information collected on our website) will likely be transferred to the successor entity, and shared with others in preparation or anticipation of such an event (e.g., during due diligence).
Where the data is stored	**data is stored in US-based servers.** SafeGraph protects Information in our possession against unauthorized access, disclosure, alteration, or destruction. We regularly review our physical security, storage, and processing to ensure compliance with industry best practices.

How data is shared	Data is shared through the services set for customers which include: **Bulk delivery data**. SafeGraph offers a number of ways to access our data in bulk on various cadences. Learn how to access SafeGraph data through tools like S3, Databricks, AWS Data Exchange, and more. https://docs.safegraph.com/docs/bulk-data-delivery **Application Programming Interface (API).** The SafeGraph Place API is a GraphQL API that is useful for looking up attributes of a point of interest (POI) by Placekey or location name. https://docs.safegraph.com/docs/places-api **SafeGraph Shop.** This service provides control over purchased datasets. This service allows data browsing to purchase by choosing the places you want data for, and then filtering to find the attributes you need. You can also integrate with an API, or use the free Match Service to upload a CSV of your own data and enrich it with more attributes. https://docs.safegraph.com/docs/safegraph-shop
How long is the data stored?	Information we collect is retained **indefinitely** provided that SafeGraph will comply with the opt-out procedures described in the privacy policy.
Effectiveness	The SafeGraph service collects information needed for the offered service. However; some inaccuracies may show due to the nature of errors in geolocation generation. Also, visitors without electronic devices reporting geolocation or tagging services (like bluetooth sensor) will be invisible to this service and some ground truth may be required to validate their inaccuracies. **Data on visits reported nuances** **Worker & non-worker visits** - visitor data is aggregated from all the geolocated electronic devices within a geofenced location. However, this bucket of data does not separate visitors from workers. These visitors will be disproportionately represented in the highest bucket. **GPS data** - The visits are determined using GPS data, SafeGraph does not include any GPS data with a horizontal accuracy >160 meters. **Very long visits** - Sometimes a visit lasts a very long time (e.g., > 24 hours). This is likely due to picking up an employee device or picking up someone in a place above a POI (such as residential or retail over office). **Relationship with places of interest (POI) opening and closing dates** - Where the dates of POI opening and closing are known (see opened_on and closed_on columns in Core Places), no visits will be attributed to the POI before the opened_on date or after the closed_on date. If these dates are not known, it is possible for the algorithm to mis-attribute visits for those POIs (e.g., after a POI has closed in real life but where this knowledge has not yet been captured in our data). Other consequences in data accuracy may be connected to times of interest as the analysis is done monthly, yearly, or by weekends, etc.

Proportionality, fundamental rights, frequency of the collection, and data protection and privacy issues line unintended data collection or processing.	Data provided to users is already anonymized; however, access to some derivative products may create unintended risks to specific individuals. Some of these and other privacy rights issues are described below: **Consent**. SafeGraph offers a service that relies on aggregation of data coming from third parties. Managing consent (or informed consent) in this ecosystem does not seem feasible. However, the measures developed by the company seem to follow privacy protection best practices and anonymization of data is done in a reasonable way. People whose information is collected do not have a viable way to ask for any privacy right to SafeGraph given that the company is also a customer of data aggregators. However, the company provides reasonable privacy protection services as described in their public documentation and privacy policy. **Re-identification risk.** There are certain cases where data fields offer through the patterns dataset (https://docs.safegraph.com/docs/monthly-patterns) may increase the risk for re-identification of data, this includes: Number of visitors to the points of interest (POI) from each census block group or census tract based on the visitor's home location Number of visitors to the POI from each country based on the visitor's home country (or state) code. Median distance from home travelled by visitors (of visitors whose home we have identified) Number of visitors to the POI that are using Android vs. iOS. Number of visitors to the POI based on the wireless carrier of the device. SafeGraph does not report data unless at least 2 visitors are observed from that group. If there are between 2 and 4 visitors this is reported as 4. This measure reduces the likelihood for re-identification. **Tracking and surveillance**. Fields like the Median distance from home travelled by visitors seem to suggest that some tracking is done, not available to customers, and SafeGraph has access to more sensitive data. “Home location” is an abbreviation for “common nighttime location”. **Sampling and geographic bias**. Sampling error is the difference between a sample and the population. Small geographic bias exists in our panel based on our understanding of the home locations of the devices in the panel. SafeGraph tested for geographic bias by comparing its determination of the state-by-state numbers of home location of

	the devices in the panel to the true proportions reported by the 2016 US Census. Based on that analysis, SafeGraph panel density closely mirrors true population density. The overall average percentage point difference is < 1% with a maximum of +/-3% per state. https://colab.research.google.com/drive/1u15afRytJMsizySFqA2EPlXSh3KTmNTQ#offline=true&sandboxMode=true **Forecasting and predictive algorithms.** SafeGraph claims to use AI in data processing, curation, and analysis. SafeGraph also develops derivative products using AI algorithms. For instance, in predicting financial indicators, SafeGraph data can be used to estimate foot traffic and predict financial indicators of companies (eg. number of visitors, revenue, etc.).

Privacy safeguards	Data is anonymized and aggregated by time and geolocation. Industry standards for privacy and information protection are used. Best practices and clear and accessible documentation for delivering data to users are part of good privacy practices. To preserve privacy, SafeGraph applies differential privacy techniques to the following columns: visitor_home_cbgs, visitor_home_aggregation, visitor_daytime_cbgs, visitor_country_of_origin, device_type, carrier_name. SafeGraph has added Laplacian noise to the values in these columns. After adding noise, only attributes (e.g., a census block group) with at least two devices are included in the data. If there are between 2 and 4 visitors this is reported as 4. SafeGraph offers an acceptable and above the standard technique for privacy protection. SafeGraph provide contact information for privacy related issues at privacy@safegraph.com or by writing at: Attention: Data Privacy Officer 1624 Market Street, Suite 226 #53755 Denver, CO 80202
Open source	No
AI/ML claims	Yes
Privacy Policy (link)	https://www.safegraph.com/privacy-policy
Privacy risk	Medium
Surveillance Tech?	The data service is based on surveillance technologies
Portland Privacy Principles (P3)	
Data Utility	We recommend defining performance metrics that represent the value of using this data and assess it in a predetermined time period.
Full lifecycle stewardship	There are still big black boxes regarding initial geolocated data and the ability of users to control tracking from their service providers. This may include electronic devices attached or used to minors that would be hard to filter out. However, SafeGraph implements reasonable anonymized data provided to users. The risk transferred to consumers of data is already minimized. Some derivative services may increase that risk in a marginal way.

Transparency and accountability	SafeGraph offers a variety and acceptable channels to learn about their service and privacy measures. However, it is not clear what process will be followed when a privacy breach happens, considering that some data fields used in their analysis can lead to re-identification. This service is subjected to all applicable consumer data regulation and laws.
Ethical and non-discriminatory use of data	This service does not include or mentions equity impacts. Those individuals without an electronic device recorded by source data services won't be counted. These individuals may include low income people and those experiencing houselessness, as well as children.
Data openness	Data offered by this service is mostly proprietary. Some free and open data is also included in their services and available to users with free accounts. Different privacy risks may be added when an individual account is created in their system.
Equitable data management	This data is insufficient to measure outcomes or when designing or implementing programs, services, and policies, connected to people that do not have electronic devices.
Automated Decision Systems	This data service may include biases when added to automatic decision systems. Information should be verified with more reliable and equitable sources. SafeGraph offers good documentation to understand bias and limitation of data for different use cases: https://colab.research.google.com/drive/1tD12yigZ6rCzYdNIavBeFuLHduDyMSji?usp=sharing
Optional	
Consent	Full data lifecycle consent is not possible. In practice, the use of data delivered by SafeGraph is anonymized and aggregated and offers low risk of re-identification when using basic queries and services. Advanced queries that included specific socioeconomic, geographic, or other tracking information may increase that risk, and impacted individuals do not have a way to remove any access to their personal data.

Privacy Impact Risk Severity Assessment form

WORST CASE SCENARIO	**Medium**

1. Individual Privacy Harms

Impact: MODERATE
Likelihood: UNLIKELY
Total Risk level: **LOW**

Justification:
There is a moderate risk of re-identification when using advanced queries that include socioeconomic, geographic, or other tracking information may increase that risk. Basic anonymized data have a minimum individual privacy harm risk.

Comments:
The amount of information and contextualization needed to re-identify an individual after all the privacy safeguards makes this event unlikely to happen.

2. Equity, Disparate Community Impact

Impact: HIGH
Likelihood: POSSIBLE
Total Risk level: **MEDIUM**

Justification:
Individuals without an electronic device recorded by source data services won't be counted. These individuals may include low income people and those experiencing houselessness, as well as children. A large and important group of people and service area might be invisible in this service.

Comments:
If this data is not contextualized and used as is in impactful community decisions, it can create harm. The recommendation is to contextualize data bias and include multiple sources of data that complement the invisible data from this service.

3. Political, Reputation & Image

Impact: MODERATE

Likelihood: POSSIBLE
Total Risk level: **MEDIUM**

Justification:

The collection and use of geolocation data has been highlighted with some privacy rights issues. These issues include lack of transparency on how data has been collected from people and lack of meaningful consent from users from whom data has been collected. In this application, the City, as a consumer, has only access to anonymized and aggregated data; which minimizes risks for re-identification, but raises the ethical issue of whether to become part of a system that impacts individuals privacy indirectly.

Comments:

In order to mitigate the risk coming from the initial collection of mobility data from electronic devices, it is needed to inform the public properly that the City only collects aggregated and anonymized data. In addition, the City can also inform on how to manage geolocation and third-party tracking in electronic devices for those who may want to opt-out.

Inform City data analysts about the biases and limitations of this source data. Encourage analysts to look for alternative and inclusive data sources.

Inform the community about how this data will be used and create governance structures that make data analysts aware of these impacts.

This technology relies on unnamed sources for data collection. The company states that they can provide demographic data which implies that they use sources that collect PII. Use of the technology by the City can be seen to be encouraging PII collection.

The SafeGraph service has been already exposed as using invasive data collection techniques in their services. The app has been banned by Google due to privacy concerns. Using this service may be challenged by privacy advocates.

4. City Business, Quality & Infrastructure

Impact: MODERATE
Likelihood: POSSIBLE
Total Risk level: **MEDIUM**

Justification:

Depending on whether this data will be used in Portland Parks and Recreation's businesses, services, or decisions around infrastructure, this risk could go from low to high. The main issue is that individuals may not be counted and further studies are needed in order to understand this impact.

From a City business perspective, the issue of using biased data in business processes or for decision making, may create derivative problems in services and outcomes provided to users. Bias data emerges from missing important groups of individuals and having to collect missing information from other sources. The main impacts are likely to happen in BIPOC and communities with a high number of people living in poverty. This situation may include hidden operation costs that may have not been considered.

Comments:

The risk of missing groups that do not use or are opting out these services, to whom the City may consider priority service stakeholders, there are some recommendations for specific cases:

1. Contextualize data based on identifiable biases.
2. Aggregate data sources that complement missing information from the data provided by the SafeGraph service.
3. Understand the existing documentation provided by SafeGraph on how to manage bias.
4. Use this data source only as a reference, but follow equitable data collection practices for high impact decisions.

5. Legal & Regulatory

Impact: LOW
Likelihood: UNLIKELY
Total Risk level: **LOW**

Justification:

This data is under commercial information data laws and regulations. SafeGraph offers services complying with these laws. No personal identifiable information (PII) is collected by the City.

Some Public Records may be requested and the amount of data may become large with time. Existing retention time for this data is about three years, after that period data must be destroyed. Failure of doing that may come with extra costs.

Comments:

The City uses best practices and complies with public records laws. Identifying the right record type may allow the agency to keep historic records beyond the three year period when its value is demonstrated.

6. Financial Impact

Impact: LOW
Likelihood: UNLIKELY
Total Risk level: **LOW**

Justification:
No PII is collected. There are general data storage, processing, and management costs. Costs are expected to be reasonable for the described uses of data.

Comments:
Data privacy breach risks and incident management may add some additional costs. However, it is an unlikely event given that no PII is collected.