Replica Data Disclosure Documentation

This information is considered confidential and proprietary. The purpose of this document is to provide necessary documentation to Replica customers and partners related to data use in Replica. In the event that a customer would like to share this information with the general public and / or outside parties, Replica will work with the customer to determine an appropriate summary and medium to share

Overview

Replica is a high-fidelity, synthetic representation of travel-related outcomes that can improve the monitoring and planning of transportation and land use systems.

Technically, Replica is a ***fully calibrated, regional-scale, travel demand model*** offered via software-as-a-service (SaaS). Replica has pioneered a technical approach to using data in a privacy-sensitive way.

Data Disclosure

Replica is built using a combination of the following data sources:

Mobile Location Data: Rather than rely on people's self-reported travel behaviors, Replica uses mobile location data to understand individual travel preferences and behaviors through the choices they've actually made. Replica secures this data from several vendors, including but not limited to; telecommunications companies, third-party app data aggregators, freight and logistics companies, and payment-processing companies. Replica then creates a composite data set ensuring the largest sample and the most accurate baseline. This baseline data set includes all location data from all device types across all mobile operating systems. Replica does not source raw location data directly from Google or any other Alphabet companies. Equally important, at no point in the sourcing or building process does Replica ever handle, process, or store personally identifiable information.

Replica uses modern machine learning techniques to create travel personas from mobile location data. Personas are an extraction of behavioral patterns from individual devices that live in, work in, travel to, travel from, or pass through a particular region. The purpose of creating personas versus directly using or simple factoring of the raw mobile location data is three-fold; (1) personas preserve the privacy of the source data / user, (2) personas provide an explanation for mobility behaviors and choices actually made by households and individual travelers, and (3) personas enable model sensitivity to physical or policy intervention that occurs in the real-world.

Sidewalk Labs independently verifies and audits our providers to ensure stringent privacy protections are in place for data collection. This process also verifies that end-users have provided explicit consent and have the ability to opt-out of collection at any time. This

verification process helps to ensure that Replica can confidently build a high-fidelity travel model while respecting individual privacy.

Population / Household Data: Replica uses a combination of census data and consumer marketing data to generate a synthetic population. This synthetic population is statistically equivalent to the real population of any given region and contains all household characteristics typically associated with the American Community Survey (i.e. income, age, children, car ownership, etc..). The primary reason for creating a synthetic population is two-fold; (1) census data is limited to aggregate geographies, which in its raw form, limits the ability to assign attributes to individuals or single households, and (2) using a synthetic population is another method of protecting privacy without compromising the spatial fidelity of the end model.

The synthetic population is created using Bayesian Networks and allocation based on convex optimization. These data science techniques allow for: (1) modeling the dependencies in socio-demographic parameters and structure of the households, and (2) synthesis of the population at the level of individual households so that it matches aggregate census information at the required level of aggregation such as block groups or tracts.

Replica does not handle, process, or story any personally identifiable information at any point in the creation of the synthetic population.

Ground-Truth Data: Replica uses ground-truth data, provided by local public agencies as a means of calibration. This data includes traffic, transit, pedestrian and bicycle counts, land use and parcel information, and any other relevant data sets made available. This data is used to create a fully-calibrated travel-demand output.

While unlikely, if a public agency attempts to provide Replica with data containing personally identifiable information, Replica requests the data be scrubbed of that information prior to obtaining. In the event that an agency cannot scrub the personal information, Replica takes the necessary steps to; 1) remove all personally identifiable information, and 2) immediately deprecate the original data once complete.

Summary of Vendor Audit Questionnaire

Sidewalk Labs requires all data vendors to complete an audit prior to agreeing to work with the vendor. At any point the vendor fails to meet the privacy requirements, Sidewalk Labs terminates the relationship and no longer uses data provided. The audit is completed across several dimensions of the vendor operations:

- Ensure data supplier has a current, clear, and conspicuous privacy notice that describes what data is collected, how it is used and shared, and user choices for opt-out
- Ensure data supplier contractually requires it's sources (if different) to maintain a current, clear, and conspicuous privacy notice that describes what data is collected, how it is used and shared, and user choices for opt-out

- Verification that data supplier takes steps to confirm and periodically review data samples and publishers to ensure appropriate notice
- Ensure data supplier obtains opt-in consent for collection of any location data
- Confirm data supplier has provided opt-out mechanisms and that user preferences for opt-out are propagated throughout the supply-chain
- Contractual obligation to protect data at the same or higher level than required by applicable law and makes no attempt to merge data or identify individuals within the data sets
- Ensure supplier implements technical measures to prevent re-identification, like encryption of potential identifiers
- Confirm that suppliers make the data collected available to users
- Ensure that suppliers implement technical and operational measures to protect the security of the data
- Confirm supplier has well defined limits on data retention and data is kept only as long as needed to fulfill a legitimate business need

Summary of Findings from Third Party Review

"Studies on the privacy of mobility data often centre on the property of *k-anonymity* [1], which requires the set of characteristics of each individual in a database to be shared with at least k-1 others in the database. The value k is not specified, but the larger it is the greater the level of protection (with k=2 a re-identification claim has a 50% chance of being right). K-anonymity can be attained by coarsening the data granularity – perhaps only for individuals at risk – and/or suppressing attributes or entire records for individuals at risk.

Based on the concept of k-anonymity, studies of de-identified mobile databases have shown that it is possible to identify unique individuals in a mobile database based on their spatio-temporal data. These studies have shown that coarsening the location or time data, or producing only partial data, does not necessarily resolve the problem. Some of the reasons for the identifiability of mobile user data are that individual travels are fairly unique and fairly consistent. Most individuals spend the most time at few top locations, e.g., home, work, shopping, and two or three of these points may suffice to locate unique individuals. One study has shown that the median size of an individual's home-work anonymity set (number of persons with the same home-work locations) in the U.S. working population is 1 and 21 for locations known at the granularity of a census block and census tract, respectively.

In contrast, Replica data have undergone much manipulation including suppression of activity sequences, replication from multiple persona matches, perturbation of location – more than just location coarsening. Time is also perturbed; Replica data for Tuesday may include a collection of data from Tuesdays over a three-month interval, and event start/end times were simulated. The property of k-anonymity is not as necessary when data have undergone replication and perturbation.

Sidewalk Labs goes to great lengths to protect the confidentiality of their data. As noted, there is no risk of disclosure from the synthetic persons created, and the many steps taken to protect the identity and characteristics of personas seriously hinder attempts at identifying them."