

New Insights into Human Mobility with Privacy Preserving Aggregation

Tuesday, November 12, 2019

Posted by Adam Sadilek, Software Engineer and Xerxes Dotiwalla, Product Manager, Google Research

Understanding human mobility is crucial for [predicting epidemics](#), urban and transit [infrastructure planning](#), understanding people's responses to conflict and [natural disasters](#) and other [important domains](#). Formerly, the state-of-the-art in mobility data was [based on cell carrier logs](#) or [location "check-ins"](#), and was therefore available only in limited areas — where the telecom provider is operating. As a result, cross-border movement and long-distance travel were typically not captured, because users tend not to use their SIM card outside the country covered by their subscription plan and datasets are often bound to specific regions. Additionally, such measures involved considerable time lags and were available only within limited time ranges and geographical areas.

In contrast, de-identified *aggregate* flows of populations around the world can now be computed from phones' location sensors at a uniform spatial resolution. This metric has the potential to be extremely useful for urban planning since it can be measured in a direct and timely way. The use of de-identified and aggregated population flow data collected at a global level via smartphones could shed additional light on city organization, for example, while requiring significantly fewer resources than existing methods.

In "[Hierarchical Organization of Urban Mobility and Its Connection with City Livability](#)", we show that these mobility patterns — statistics on how populations move about in aggregate — indicate a higher use of public transportation, improved walkability, lower pollutant emissions per capita, and better health indicators, including easier accessibility to hospitals. This work, which appears in [Nature Communications](#), contributes to a better characterization of city organization and supports a stronger quantitative perspective in the efforts to improve urban livability and sustainability.



Visualization of privacy-first computation of the mobility map. Individual data points are automatically aggregated together with differential privacy noise added. Then, flows of these aggregate and obfuscated populations are studied.

Computing a Global Mobility Map While Preserving User Privacy

In line with our [AI principles](#), we have designed a method for analyzing population mobility with privacy-preserving techniques at its core. To ensure that no individual user's journey can be identified, we create representative models of aggregate data by employing a technique called [differential privacy](#), together with [k-anonymity](#), to aggregate population flows over time. Initially [implemented in 2014](#), this approach to differential privacy intentionally adds random "noise" to the data in a way that maintains both users' privacy and the data's accuracy at an aggregate level. We use this method to aggregate data collected from smartphones of users who have deliberately chosen to opt-in to [Location History, in order to](#) better understand global patterns of population movements.

The model only considers de-identified location readings aggregated to geographical areas of predetermined sizes (e.g., [S2 cells](#)). It "snaps" each reading into a spacetime bucket by discretizing time into longer intervals (e.g., weeks) and latitude/longitude into a unique identifier of the geographical area. Aggregating into these large spacetime buckets goes beyond protecting individual privacy — it can even protect the privacy of communities.

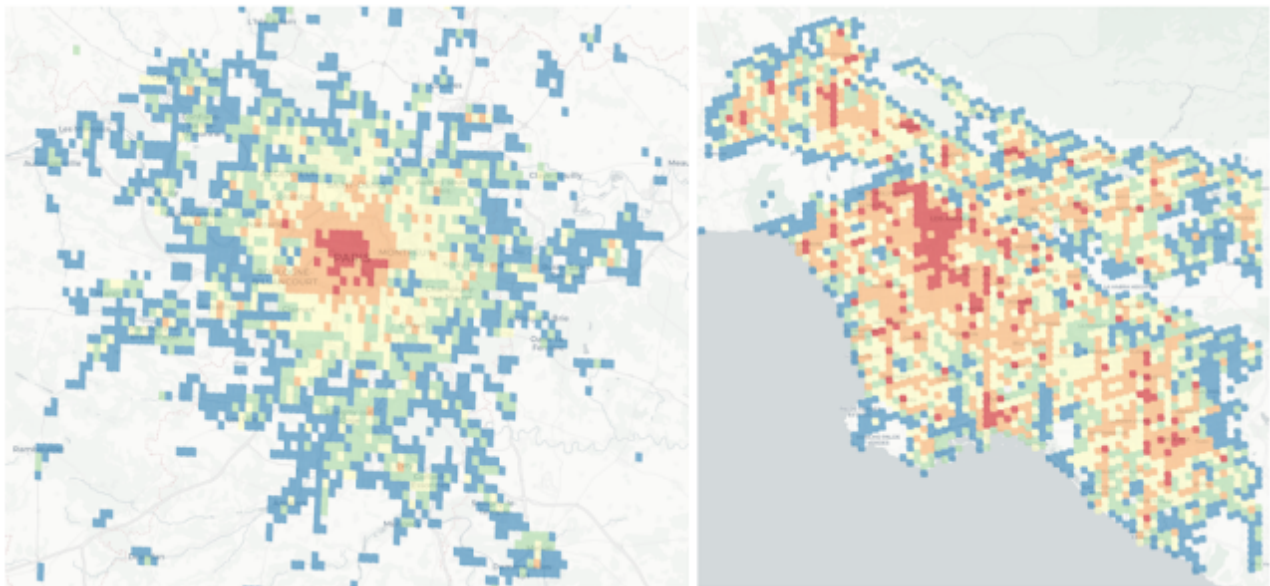
Finally, for each pair of geographical areas, the system computes the relative flow between the areas over a given time interval, applies differential privacy filters, and outputs the global, anonymized, and aggregated mobility map. The dataset is generated only once and only mobility flows involving a sufficiently large number of accounts are processed by the model. This design is limited to heavily aggregated flows of populations, such as that already used as a vital source of information for estimates of [live traffic](#) and [parking availability](#), which protects individual data from being manually identified. The resulting map is indexed for efficient lookup and used to fuel the modeling described below.

Mobility Map Applications

Aggregate mobility of people in cities around the globe defines the city and, in turn, its impact on the people who live there. We define a metric, the flow hierarchy (Φ), derived entirely from the mobility map, that quantifies the hierarchical organization of cities. While hierarchies across cities have been extensively studied since [Christaller's work](#) in the 1930s, for individual cities, the focus has been primarily on the differences between core and peripheral structures, as well as on whether cities are more "central" or "peripheral." Our results instead show that the reality is more

as whether cities are mono- or poly-centric. Our results instead show that the reality is much more rich than previously thought. The mobility map enables a quantitative demonstration that cities lie across a spectrum of hierarchical organization that strongly correlates with a series of important quality of life indicators, including health and transportation.

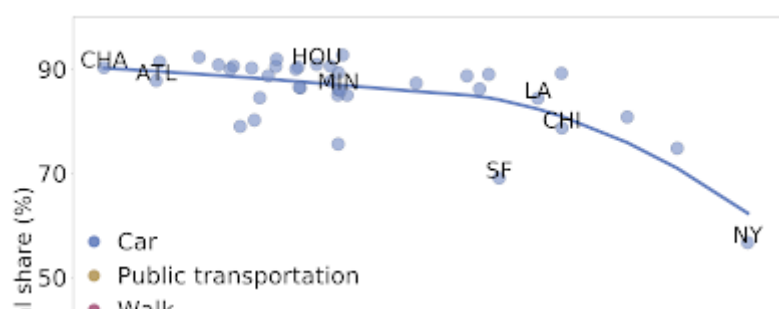
Below we see an example of two cities — Paris and Los Angeles. Though they have almost the same population size, those two populations *move* in very different ways. Paris is mono-centric, with an "onion" structure that has a distinct high-mobility city center (red), which progressively decreases as we move away from the center (in order: orange, yellow, green, blue). On the other hand, Los Angeles is truly poly-centric, with a large number of high-mobility areas scattered throughout the region.

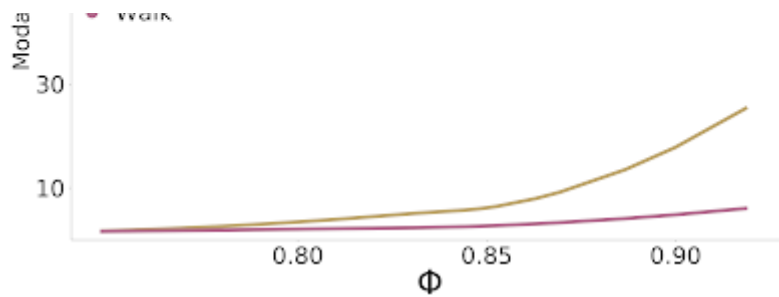


Mobility maps of Paris (**left**) and Los Angeles (**right**). Both cities have similar population sizes, but very different mobility patterns. Paris has an "onion" structure exhibiting a distinct center with a high degree of mobility (red) that progressively decreases as we move away from the center (in order: orange, yellow, green, blue). In contrast, Los Angeles has a large number of high-mobility areas scattered throughout the region.

More hierarchical cities — in terms of flows being primarily between hotspots of similar activity levels — have values of flow hierarchy Φ closer to the upper limit of 1 and tend to have greater levels of uniformity in their spatial distribution of movements, wider use of public transportation, higher levels of walkability, lower pollution emissions, and better indicators of various measures of health. Returning to our example, the flow hierarchy of Paris is $\Phi=0.93$ (in the top quartile across all 174 cities sampled), while that of Los Angeles is 0.86 (bottom quartile).

We find that existing measures of urban structure, such as [population density](#) and [sprawl](#) composite indices, correlate with flow hierarchy, but in addition the flow hierarchy conveys comparatively more information that includes behavioral and socioeconomic factors.





Connecting flow hierarchy Φ with urban indicators in a sample of US cities. Proportion of trips as a function of Φ , broken down by model share: private car, public transportation, and walking. Sample city names that appear in the plot: ATL (Atlanta), CHA (Charlotte), CHI (Chicago), HOU (Houston), LA (Los Angeles), MIN (Minneapolis), NY (New York City), and SF (San Francisco). We see that cities with higher flow hierarchy exhibit significantly higher rates of public transportation use, less car use, and more walkability.

Measures of urban sprawl require composite indices built up from much more detailed information on land use, population, density of jobs, and street geography among others (sometimes up to [20 different variables](#)). In addition to the extensive data requirements, such metrics are also costly to obtain. For example, censuses and surveys require a massive deployment of resources in terms of interviews, and are only standardized at a country level, hindering the correct quantification of sprawl indices at a global scale. On the other hand, the flow hierarchy, being constructed from mobility information alone, is significantly less expensive to compile (involving only computer processing cycles), and is available in real-time.

Given the ongoing [debate](#) on the [optimal](#) structure of cities, the flow hierarchy, introduces a different conceptual perspective compared to existing measures, and can shed new light on the organization of cities. From a public-policy point of view, we see that cities with greater degree of mobility hierarchy tend to have more desirable urban indicators. Given that this hierarchy is a measure of proximity and direct connectivity between socioeconomic hubs, a possible direction could be to shape opportunity and demand in a way that facilitates a greater degree of hub-to-hub movement than a hub-to-spoke architecture. The proximity of hubs can be generated through appropriate land use, that can be shaped by data-driven zoning laws in terms of business, residence or service areas. The presence of efficient public transportation and lower use of cars is another important factor. Perhaps a combination of policies, such as congestion-pricing, used to disincentivize private transportation to socioeconomic hubs, along with building public transportation in a targeted fashion to directly connect the hubs, may well prove useful.

Next Steps

This work is part of our larger [AI for Social Good](#) efforts, a program that focuses Google's expertise on addressing humanitarian and environmental challenges. These mobility maps are only the first step toward making an impact in epidemiology, infrastructure planning, and disaster response, while ensuring high privacy standards.

The work discussed here goes to great lengths to ensure privacy is maintained. We are also working on newer techniques, such as [on-device federated learning](#), to go a step further and enable computing aggregate flows without personal data leaving the device at all. By using distributed [secure aggregation protocols](#) or [randomized responses](#), global flows can be computed without even the aggregator having knowledge of individual data points being aggregated. This technique has also been applied to help [secure Chrome from malicious attacks](#).

Acknowledgements

This work resulted from a collaboration of Aleix Bassolas and José J. Ramasco from the [Institute](#)

for [Cross-Disciplinary Physics and Complex Systems](#) (IFISC, CSIC-UIB), Brian Dickinson, Hugo Barbosa-Filho, Gourab Ghoshal, Surendra A. Hazarie, and Henry Kautz from the [Computer Science Department](#) and [Ghoshal Lab](#) at the University of Rochester, Riccardo Gallotti from the [Bruno Kessler Foundation](#), and Xerxes Dotiwalla, Paul Eastham, Bryant Gipson, Onur Kucuktunc, Allison Lieber, Adam Sadilek at Google.

The [differential privacy library](#) used in this work is open source and available on our [GitHub repo](#).



Google

Google · Privacy · Terms