Generating and Understanding Large Datasets of Social Contact Patterns from Foot-Traffic Data

Hossein Amiri
hossein.amiri@emory.edu
Emory University
Atlanta, USA

James Song
james.song2@emory.edu
Emory University
Atlanta, USA

Li Xiong
lxiong@emory.edu
Emory University
Atlanta, USA

Andreas Züfle
azufle@emory.edu
Emory University
Atlanta, USA

Abstract

Recent advances in large-scale human mobility datasets have opened new opportunities to improve public health through data-driven strategies, advanced computational methods, and interdisciplinary approaches. A key focus in epidemiological research is the estimation and analysis of social contact patterns—representing the frequency and nature of interactions among different demographic groups. These patterns are vital for modeling disease transmission, evaluating public health interventions, and guiding resource allocation. However, obtaining accurate and representative contact data remains a major challenge. In this paper, we propose a novel, scalable framework for generating and analyzing large-scale social contact datasets derived from foot-traffic data. Our approach integrates statistical modeling to estimate demographic distributions—such as age groups—at millions of points of interest (POIs) across the United States and globally, including restaurants, stores, hospitals, and schools. This framework enables actionable insights to inform public health strategies and improve population health outcomes. Moreover, the resulting datasets have broad cross-sector utility, supporting applications in strategic business planning, resource distribution, and personalized marketing and advertising.

CCS Concepts

• **Information systems → Geographic information systems**; **Location based services**; • **Applied computing → Health care information systems**.

Keywords

Human Mobility, Contact Tracing, Social Contact Patterns

ACM Reference Format:
Hossein Amiri, James Song, Li Xiong, and Andreas Züfle. 2025. Generating and Understanding Large Datasets of Social Contact Patterns from Foot-Traffic Data. In *The 1st ACM SIGSPATIAL International Workshop on Geospatial Computing for Public Health (GeoHealth '25), November 3–6, 2025, Minneapolis, MN, USA.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3764918.3770162

GeoHealth '25, Minneapolis, MN, USA

ACM ISBN 979-8-4007-2181-6/2025/11
https://doi.org/10.1145/3764918.3770162

Table 1: A simple example of a contact pattern showing the number of contacts between three age groups (young, middle-aged, and elderly) in a population.

Elderly	20	30	40
Middle	50	80	30
Young	100	50	20
From / To	Young	Middle	Elderly

1 Introduction

The emergence of large-scale datasets has revolutionized public health research by enabling the discovery of previously inaccessible insights and patterns [1]. These datasets provide a critical foundation for understanding complex health issues, identifying trends, and informing evidence-based interventions. However, data availability and quality often vary significantly across regions and populations [2, 3]. Researchers frequently face obstacles in accessing comprehensive datasets that accurately capture the demographics and behaviors of the populations under study. Large-scale datasets enriched with demographic information are vital for generating meaningful insights and guiding effective public health responses [4]. To effectively guide public health responses, it is crucial to understand social contact patterns; defined as the frequency and nature of interactions between different demographic groups within a population. These patterns are essential for modeling disease transmission, evaluating public health interventions, and informing resource allocation strategies. A simple example of a contact pattern is depicted in Table 1, which shows the number of contacts between three age groups (young, middle-aged, and elderly) in an imaginary population. For instance, the last row indicates that young people have 100 contacts with other young people, 50 contacts with middle-aged individuals, and 20 contacts with elderly individuals.

Contact patterns are traditionally collected through surveys or diaries, which are often time-consuming and expensive, and have been used intensively by the public health community [5–11] to estimate contact patterns in specific populations and settings. However, these methods typically involve small populations, such as manually collected contact diaries at specific healthcare facilities [5], rural and urban areas in Mozambique [6–8], or long-term care facilities [9]. These studies are usually limited to specific populations and settings, making it challenging to generalize findings to larger populations. One of the largest studies, which surveyed 4654 participants across the United States, can be found in [10]. However, even with national sampling, the density of data for individual states, cities, regions, and specific places of interest remains insufficient for comprehensive analysis.

In this paper, we propose an approach to generate social contact patterns for every place of interest in the United States and across the world using publicly available foot-traffic data capturing the mobility of tens of millions of individuals. Foot-traffic data refers to information about the movement of individuals to and from specific locations, often aggregated and anonymized to protect privacy. Advan Research [12] provides foot-traffic data for 7 million places of interest (POI) in the United States, capturing approximately 50 million cell phone users. For each POI, the home locations of the visiting individuals are aggregated at the census block group (CBG) levels, offering insight into the number of visitors, their stay duration, and their home CBGs. Such datasets are invaluable for understanding human mobility patterns and their implications for public health, urban planning, and commercial applications. Through extensive exploration of various aspects of spatial data [13–19], we have identified demographic information as a critical factor in understanding social behaviors and public health outcomes. Unfortunately, such demographic data are frequently incomplete, unavailable, or deliberately excluded due to privacy concerns and participant reluctance.

Contact patterns can accurately model the spread of infectious diseases within a population. They help researchers and policymakers predict and prevent pandemics, analyze social behavior for public safety or commercial purposes, and make data-driven decisions about resource allocation. By examining interactions between demographic groups, high-risk populations can be identified, enabling targeted interventions to reduce transmission, such as in the case of monkeypox, where age significantly influences the risk factors [20]. Incorporating age-specific transmission rates allows for more precise estimates of disease propagation across population segments. This information is essential for guiding public health strategies, including prioritizing vaccinations for vulnerable groups and implementing targeted social distancing measures. Additionally, understanding contact patterns can reveal potential superspreader events or high-risk locations, supporting more effective containment. Businesses can also apply these insights to optimize service delivery and enhance customer engagement by analyzing social interaction dynamics at key points of interest.

Our goal in this paper is to address these challenges by developing methods to generate detailed and accurate social contact patterns for any POI in the world. One input to the algorithm is the number of visitors to each location grouped by their home CBG available from Advan data. The second input to the algorithm is the demographic distribution of the local population, available from the U.S. Census Bureau. By combining these two datasets, we can estimate the demographic composition of visitors to each POI using the proposed statistical techniques, as detailed in Section 4. In general, we will estimate the demographic composition of visitors, such as age groups, gender distribution, and socioeconomic status, in a range of POIs, including retail settings, healthcare facilities, educational institutions, and public spaces. Our approach systematically incorporates demographic data into spatial analyses, enhancing predictive modeling capabilities and enabling more targeted public health interventions. With these insights, the research objectives of this paper are as follows:

- Estimate the demographic distribution of visitors to various points of interest (POIs) using direct probability assignment
- Provide the steps to generate comprehensive datasets tailored to key demographic variables, including age, gender, socioeconomic status, and other relevant attributes.
- Generate social contact patterns that capture interactions between demographic groups within each POI.

In the remainder of this paper, we present related work in Section 2, define the problem in Section 3, describe our proposed methodology in detail in Section 4, present the contact matrices for selected POIs in Section 5, and conclude with a discussion of the future works in Section 6.

2 Related Works

The social contact patterns is important and has been studied in different research contexts. Studies using contact diaries have shown stable patterns of interaction among healthcare workers and their role in transmission dynamics [5]. Data from rural and urban Mozambique highlighted age-related and setting-specific contact patterns, revealing the limitations of synthetic models and the need for localized empirical data [6, 7]. Changes in U.S. employee contact rates during COVID-19 indicated rising transmission potential in community settings [8]. Long-term care facility employees exhibited distinct contact behaviors across household and workplace settings, showing how staff may bridge community and institutional transmission [9]. Nationwide surveys revealed disparities in contact rates by socioeconomic status, linking structural inequalities to differential COVID-19 outcomes [10]. Outbreak modeling at a scout jamboree emphasized the role of superspreaders and network clustering in transmission [11]. The challenge of gathering comprehensive social contact data lies in the infeasibility of direct collection across all populations and settings. However, due to the widespread availability of location-sharing services, datasets such as Advan foot traffic data [12], and tools that report place-specific visitation patterns, it is increasingly feasible to estimate social contact patterns. These sources provide information on the number and timing of visitors to points of interest (POIs), and when combined with statistical demographic data from sources such as the U.S. Census, can yield meaningful estimates even without individual-level demographics.

3 Problem Definition

Figure 1 illustrates a simple example involving three age groups visiting the Kaldi's Coffee at Emory University from four different census block groups (CBGs). Each CBG has a different age distribution, and the goal is to calculate the aggregate age distribution of all visitors to Kaldi's Coffee.

- CBG1: 15 visitors, with 60% old, 30% middle, and 10% young
- CBG2: 30 visitors, with 30% old, 30% middle, and 40% young
- CBG3: 35 visitors, with 20% old, 40% middle, and 40% young
- CBG4: 55 visitors, with 30% old, 10% middle, and 60% young

The question is: What is the overall percentage of each age group among all visitors to Kaldi's Coffee?

To estimate the age group distributions of visitors to a POI, we can use the known age distributions of the CBGs from which they

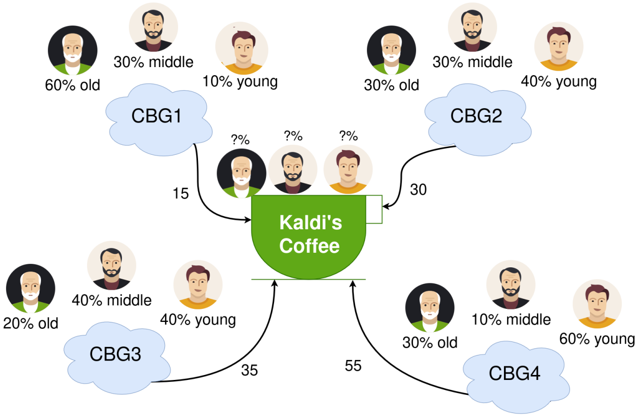

Figure 1: Simplified problem of estimating visitor ages based on their CBG origins. Kaldi's Coffee is a café at Emory University.

originate. A naive solution can be achieved by multiplying the total number of visitors by the proportion of each age group of the CBGs. This can give an estimate of the number of visitors in each age group. Applying this approach to the problem in Figure 1, we estimate the number of visitors from each age group for the POI in Table 2. Notably, Figure 1 only shows a toy example having only four CBGs and three age groups. In practice, a real-world dataset would involve more age groups and hundreds of CBGs, resulting in a highly overdetermined linear system with many more equations than unknowns.

4 Methodology

To conduct this research, we collected two primary datasets: Demographic data and visitors data. The demographic data were sourced from the Census Bureau, providing detailed demographic distributions for various Census Block Groups (CBGs). Visitor data was obtained from the Foot-traffic data provided by a location-based services, Advan [12], which tracks anonymized movement patterns to Points of Interest (POIs). The demographic data includes population proportions for different demographic groups for each CBG, while the visitor data provide the total number of visitors to specific POIs, aggregated by CBG. By combining these datasets, we can estimate the demographic distribution of visitors to POIs using the proposed methodology.

4.1 Data

We construct contact matrices based on two primary dataset: the Advan Weekly Patterns dataset [12] and the American Community Survey (ACS) 5-year Estimates [21]. The Advan dataset provides information at the point of interest (POI) level, including POI IDs, POI names, weekly aggregated CBG distributions of visitors, median dwell times, etc. The ACS data supplies the demographic composition, specifically age distribution of each CBG in the study area.

For this analysis, we focus on greater Atlanta, Georgia, in 2019. The area covers 5 counties in total: Fulton, Gwinnett, Cobb, DeKalb, and Clayton. From the Advan dataset, we extract the visitor data and other relevant information of all available POIs located within those counties, including POI IDs, the home CBGs of visitors each week, hourly visit counts, and median dwell time. ACS data from the same five counties in 2019 is used to determine the age group composition of each CBG. It is aggregated into the following groups: 0-9 years, 10-19 years, 20-29 years, 30-39 years, 40-49 years, 50-59 years and 60+ years old. These groups are chosen to balance demographic granularity with privacy and statistical robustness.

Table 2: Estimated visitor count in each age group for Figure 1

	Old	Middle	Young
CBG1	15*0.6	15*0.3	15*0.1
CBG2	30*0.3	30*0.3	30*0.4
CBG3	35*0.2	35*0.4	35*0.4
CBG4	55*0.3	55*0.1	55*0.6
Estimated	**42**	**33**	**61**

4.2 Implementation

The construction of contact matrices involves estimating the expected number of co-presences between visitors of different age groups at each POI across time. The process proceeds in three main steps:

Step 1 – Sampling Age Distribution of POI Visitors: For each POI in each week, we retrieve the distribution of home CBGs of visitors from the Advan dataset and use the ACS age compositions of those CBGs to calculate an estimated age distribution of visitors. The visitor age groups of each CBG are randomly sampled based on the age distribution of that CBG, and then aggregated together to represent the visitor age composition of the selected POI in that selected week.

Step 2 – Hourly Sampling and Presence Modeling: Based on the composition of the visitor's age and the hourly visitor counts, the hourly visitor age composition is sampled. Using the median dwell time of visitors in the selected POI as the dwell time of all visitors, we can then infer the visitor count of each age group in each hour.

Step 3 – Contact Matrix Construction: Assuming that all the visitors at the same POI and at the same hour made contacts with each other, we can now compute contact matrices using the hourly visitor age group compositions. Then, all 168 hourly contact matrices are aggregated to get the contact matrix of the selected POI in the selected week.

By repeating the above process, we can now compute weekly contact matrices of different age groups for all available POIs in the US. Figure 2, 3, and 4 show examples of the contact matrices generated by the described method for a primary school, a restaurant, and a gas station near Atlanta, respectively. The code for this approach is available on https://github.com/onspatial/sigspatial2025-social-contact-patterns.

5 Experimental Results

Using our methodology, we generate weekly contact matrices for selected POIs—each reflecting distinct interaction patterns based on the POI's function and demographics, with weekly resolution aligned to the ADVAN data format. As shown in Figure 2, Ocee Elementary School has a high concentration of visitors in the 0–9 and 50–59 age groups, resulting in the strongest contact intensity between these two cohorts. This likely captures interactions between young students and adult staff or parents. We note that this elementary school has approximately 700 students. If each student comes into contact with every other student, this would result in roughly 0.5 million contacts per day. In contrast, the matrix in Figure 3, representing Nan Thai Fine Dining, displays concentrated interactions among adults aged 30 and above, which is consistent with the expected clientele of upscale restaurants. Meanwhile, the

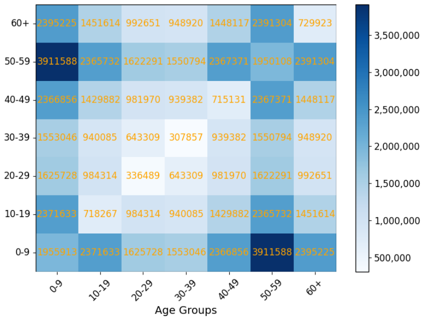

Figure 2: POI “Ocee Elementary School” in Johns Creek, GA

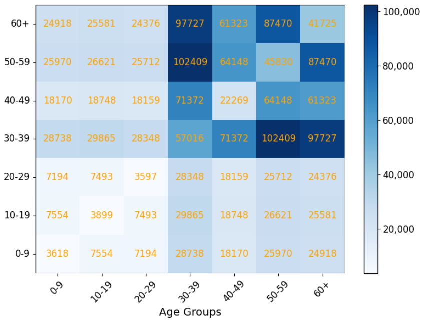

Figure 3: POI “Nan Thai Fine Dining” in Atlanta, GA

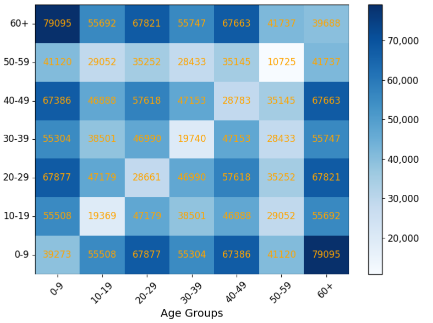

Figure 4: POI “Exxon” in Atlanta, GA

Exxon gas station in Figure 4 presents a more diffused contact matrix, with moderate contact across all age groups and no strong clustering. This pattern reflects the unstructured nature of gas station visits, where brief and non-repetitive visits lead to broad but low-intensity contacts across demographics. Collectively, these results demonstrate the method’s ability to extract age-based contact patterns at POI level from mobility traces, capturing venue-specific demographic differences. We note that the resulting dataset cannot be shared due to Advan’s data policies.

6 Conclusions and Future Work

This paper presents a novel, scalable framework for generating social contact matrices from foot-traffic data, enabling high-resolution, demographic-aware insights into human interactions at millions of POIs. By inferring age distributions from home census block group (CBG) demographics and applying probabilistic sampling, our method replaces costly surveys with a data-driven, repeatable, and adaptable solution. The resulting contact matrices offer valuable inputs for epidemiological modeling, guiding public health interventions, and informing strategies for disease prevention and resource allocation. Additionally, the framework holds significant potential beyond health, providing actionable intelligence for retail site selection, marketing optimization, and urban planning.

Despite its strengths, the proposed approach has inherent limitations. Demographic inference is based on CBG-level aggregation, which can introduce bias in areas with high heterogeneity or unusual visitor patterns. Additionally, assuming uniform contact probabilities among co-present individuals overlooks important factors such as spatial layout, dwell time, and behavioral variation that influence real-world interactions. The stochastic nature of the sampling process also introduces variability, particularly affecting precision in low-traffic or demographically skewed POIs. To mitigate these limitations, incorporating alternative data sources with finer granularity could improve accuracy. Furthermore, optimization techniques can be applied to shift the focus toward POI-level statistics rather than relying solely on CBG-level information. In addition, the validation of the method is an ongoing process that requires collecting ground truth data and comparing the results of the proposed approach with real-world data.

7 Acknowledgment

This publication was made possible by the Insight Net cooperative agreement CDC-RFA-FT-23-0069 from the CDC’s Center for Forecasting and Outbreak Analytics. Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the Centers for Disease Control and Prevention. This research was also supported by the NSF #2109647 and #2125530.

References

[1] Kornelia Batko and Andrzej Slezak. The use of big data analytics in healthcare. *Journal of big Data*, 9(1):3, 2022.
[2] Alexis Pengfei Zhao, Shuangqi Li, et al. Ai for science: predicting infectious diseases. *Journal of safety science and resilience*, 2024.
[3] Hong Chen, David Hailey, Ning Wang, and Ping Yu. A review of data quality assessment methods for public health information systems. *International journal of environmental research and public health*, 11(5):5170–5207, 2014.
[4] Jana Sedlakova, Paola Daniore, , et al. Challenges and best practices for digital unstructured data enrichment in health research: A systematic narrative review. *PLOS Digital Health*, 2(10):e0000347, 2023.
[5] Lauren Pischel, Obianuju G Aguolu, et al. P-271. contact patterns of united states health care workers at quaternary health center: Stability of contacts during the post-pandemic era. In *Open Forum Infectious Diseases*, volume 12, pages ofae631–475, 2025.
[6] Moses C Kiti, Charfudin Sacoor, et al. Social contact patterns in rural and urban settings, mozambique, 2021–2022. *Emerging Infectious Diseases*, 31(1):94, 2025.
[7] Moses C Kiti, Charfudin Sacoor, et al. Characterizing social contact patterns in rural and urban mozambique: the globalmix study, 2021-2022. *medRxiv*, pages 2024–06, 2024.
[8] Moses C Kiti, Obianuju G Aguolu, Alana Zelaya, et al. Changing social contact patterns among us workers during the covid-19 pandemic: April 2020 to december 2021. *Epidemics*, 45:100727, 2023.
[9] Seth Zissette, Moses C Kiti, Brady W Bennett, et al. Social contact patterns among employees in us long-term care facilities during the covid-19 pandemic, december 2020 to june 2021. *BMC Research Notes*, 16(1):294, 2023.
[10] Kristin N Nelson, Aaron J Siegler, Patrick S Sullivan, et al. Nationally representative social contact patterns among us adults, august 2020-april 2021. *Epidemics*, 40:100605, 2022.
[11] Jon Zelner, Carly Adams, Joshua Havumaki, and Ben Lopman. Understanding the importance of contact heterogeneity and variable infectiousness in the dynamics of a large norovirus outbreak. *Clinical Infectious Diseases*, 70(3):493–500, 2020.
[12] AdvanResearch. https://www.advanresearch.com, 2025. Accessed: [2025-04-11].
[13] Hossein Amiri, Shiyang Ruan, et al. Massive trajectory data based on patterns of life. In *SIGSPATIAL’23*, pages 1–4. ACM, 2023.
[14] Will Kohn, Hossein Amiri, and Andreas Züfle. Epipol: An epidemiological patterns of life simulation (demonstration paper). In *SIGSPATIAL SpatialEpi’23 Workshop*, pages 13–16. ACM, 2023.
[15] Zheng Zhang, Hossein Amiri, et al. Large language models for spatial trajectory patterns mining, 2023.
[16] Zheng Zhang, Hossein Amiri, et al. Transferable unsupervised outlier detection framework for human semantic trajectories, 2024.
[17] Hossein Amiri, Will Kohn, et al. The patterns of life human mobility simulation, 2024.
[18] Hossein Amiri, Ruochen Kong, and Andreas Zufle. Urban anomalies: A simulated human mobility dataset with injected anomalies, 2024.
[19] Hossein Amiri, Richard Yang, and Andreas Züfle. Geolife+: Large-scale simulated trajectory datasets calibrated to the geolife dataset. In *SIGSPATIAL GeoSpatial’24 Workshop*, pages 25–28, 2024.
[20] Chigozie Louisa Jane Ugwu, Nicola Luigi Bragazzi, et al. Risk factors associated with human mpox infection: a systematic review and meta-analysis. *BMJ Global Health*, 10(2), 2025.
[21] Census Data. https://data.census.gov/, 2025. Accessed: [2025-04-11].