In Silico Human Mobility Data Science: Leveraging Massive Simulated Mobility Data (Vision Paper)

ANDREAS ZÜFLE, Emory University, USA
DIETER PFOSER, George Mason University, USA
CAROLA WENK, Tulane University, USA
ANDREW CROOKS, University at Buffalo, USA
HAMDI KAVAK, George Mason University, USA
TAYLOR ANDERSON, George Mason University, USA
JOON-SEOK KIM, Oak Ridge National Laboratory, USA
NATHAN HOLT, L3Harris, USA
ANDREW DIANTONIO, L3Harris, USA

Human mobility data science using trajectories or check-ins of individuals has many applications. Recently, we have seen a plethora of research efforts that tackle these applications. However, research progress in this field is limited by a lack of large and representative datasets. The largest and most commonly used dataset of individual human trajectories captures fewer than 200 individuals while data sets of individual human check-ins capture fewer than 100 check-ins per city per day. Thus, it is not clear if findings from the human mobility data science community would generalize to large populations. Since obtaining massive, representative, and individual-level human mobility data is hard to come by due to privacy considerations, the vision of this paper is to embrace the use of data generated by large-scale socially realistic microsimulations. Informed by both real data and leveraging social and behavioral theories, massive spatially explicit microsimulations may allow us to simulate entire megacities at the person level. The simulated worlds, which do not capture any identifiable personal information, allow us to perform "in silico" experiments using the simulated world as a sandbox in which we have perfect information and perfect control without jeopardizing the privacy of any actual individual. In silico experiments have become commonplace in other scientific domains such as chemistry and biology, permitting experiments that foster the understanding of concepts without any harm to individuals. This work describes challenges and opportunities for leveraging massive and realistic simulated alternate worlds for in silico human mobility data science.

Additional Key Words and Phrases: Spatial Simulation, Mobility Data Science, Trajectory Data, Location Based Social Network Data, In Silico

ACM Reference Format:
Andreas Züfle, Dieter Pfoser, Carola Wenk, Andrew Crooks, Hamdi Kavak, Taylor Anderson, Joon-Seok Kim, Nathan Holt, and Andrew DiAntonio. 2021. In Silico Human Mobility Data Science: Leveraging Massive Simulated Mobility Data (Vision Paper). In *Woodstock '18: ACM Symposium on Neural Gaze Detection, June 03–05, 2018, Woodstock, NY*. ACM, New York, NY, USA, 25 pages. https://doi.org/10.1145/1122445.1122456

ACM acknowledges that this contribution was authored or co-authored by an employee, contractor, or affiliate of the United States government. As such, the United States government retains a nonexclusive, royalty-free right to publish or reproduce this article, or to allow others to do so, for government purposes only.

© 2021 Association for Computing Machinery.
Manuscript submitted to ACM

1 INTRODUCTION

Data sets capturing individual-level movements have been used for (1) modeling and analyzing **human mobility patterns** (e.g., [15, 16, 20, 69, 70, 87, 88, 102, 110, 111]), (2) recommending locations to users based on previously visited locations (for example, [11, 12, 24–27, 35, 48–50, 52, 54–56, 58–61, 68, 86, 100, 101, 112, 115, 116, 119–121, 124, 125, 127–131, 133, 134], surveyed in [6]), (3) predicting the next location to be visited by an individual (e.g., [5, 13, 33, 53, 57, 84, 96, 114]), and (4) suggests new friends to individuals based on similar interests observed by visiting similar locations [14, 34, 67, 80, 89, 99, 104, 108, 113]),

Publicly available real-world data sets have been the driving force for human mobility data science in recent years. These data sets mainly comprise *trajectory data* and *location-based social network (LBSN) data*. Trajectory data captures the location of individuals at a relatively high frequency, such as at 1Hz (one location per second per individual). LBSN data captures both 1) so-called check-ins, that is, the location of individuals only when a point of interest (POI) such as a restaurant is visited as well as 2) a social network between individuals. However, publicly available real-world trajectory and LBSN data sets exhibit certain weaknesses:

- **Data sparsity:** In all existing data sets that capture check-ins of individuals, the vast majority of users have less than ten check-ins [60]. This results in the density of the data used in experimental studies on LBSNs being only usually around 0.1% [60] (of a complete dataset that would capture every individual). For trajectory data, the most commonly used data set is the GeoLife dataset [136, 137] which captures the locations of 178 individuals in Beijing. Given the population of 21.58 million (2018) in Beijing, this means that the proportion of captured individuals compared to the real population is less than 0.001%. It becomes very challenging to infer patterns of human behavior from such a small sample. In addition, it is not clear whether the individuals included in real-world datasets are a representative sample of the population, as the demographics of anonymous users are unknown. Furthermore, the study in [51] has shown that the lower bound of predictability of human spatio-temporal behavior (defined in [51]) is as low as 27%. They conclude that "Researchers working with LBSN data sets are often confronted with doubts regarding the quality or potential of their data sets." and that "it is reasonable to be skeptical"[51].
  But imagine the research possibilities in a world where we had complete trajectory data of 100% of individuals of a large population such as Bejing, China.
- **Privacy Concerns:** Individual human mobility data is considered Personal Identifiable Information (PII) as it allows one to trace an individual's identity. Acquiring, storing, and publishing of individual human mobility data requires the consent of individuals. Even if such consent is given, users may later revoke this consent, for instance, by deleting their LBSN account. This limits, for good reasons, our ability to acquire additional individual-level human mobility data.
  But imagine the research possibilities stemming from collecting individual-level trajectory data that does not jeopardize the privacy of any real human individual.
- **No Ground-Truth Behavior:** There is no way to assess, in existing human mobility data, whether location updates or check-ins are accurate and complete or if updates are missing. It is also difficult to derive the underlying behaviors that lead to an individual's decision to visit a particular POI. For example, did an individual visit a restaurant to eat by themselves? To meet a friend? To have a business meal? Or to work in the restaurant? What preferences lead an individual to choose one grocery store over another? In real-world individual mobility data, the link between users in the data and the corresponding individual in the real world is lost (intentionally

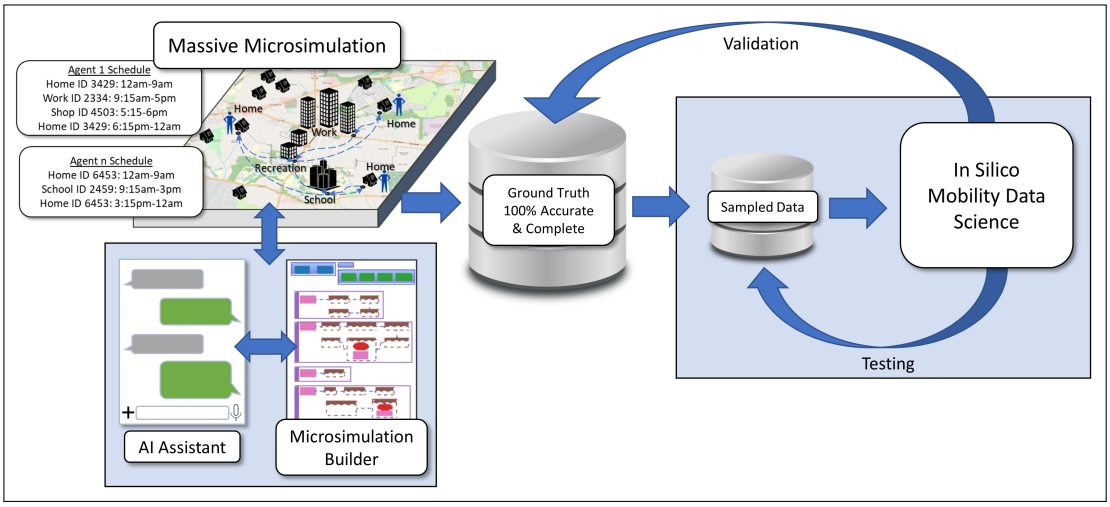


Fig. 1. The envisioned in silico mobility data science process - (left:) A massive microsimulation is created to simulate realistic human behavior specified by a user through an AI-supported builder tool. (middle:) The microsimulation generates massive datasets, including high-fidelity trajectories of all individuals over years of simulation time. This data, which is 100% accurate and complete (in the simulated world) is then sampled to generate realistic datasets. (right:) These datasets are then used to perform mobility data science tasks in the simulated in silico world as if it was the real world. The results of these tasks can then be compared to the ground truth data (of the simulated in silico world) for validation.

for privacy). Without knowing the underlying human behavior that led to the observed mobility, it is difficult to infer patterns of human behavior and to predict future mobility.

But imagine a world where we can go back in time to ask people about the purpose of their mobility to understand why *an individual visited a place of interest.*

1.1 In Silico Human Mobility Data Science

Our vision is to create such a world as illustrated in Figure 1. A digital world or a "sandbox" that 1) provides massive, complete, and high-fidelity trajectory data of 100% of the (simulated) population that 2) does not jeopardize the privacy of any real human, and 3) allows insights into the underlying behavior that led individuals to make the trips they made. We envision creating such a world "in silico" through massive microsimulation of socially plausible human behavior and mobility. To clarify this notion of "in silico" science, let us first define the underlying notion of in vitro science, which, instead of performing experiments on living organisms (in vivo) performs experiments inside a test tube:

Definition 1.1 (In Vitro; Latin: "In Glass"). In vitro refers to a process or studies that are performed with microorganisms, cells, or biological molecules outside their normal biological context [...] performed or taking place in a test tube, culture dish, or elsewhere outside a living organism [19, 64].

In vitro scientific experiments are commonplace outside of computer science. In biology, in vitro experiments are commonly performed to understand the behavior of parts of organisms (such as cells/tissue of humans). For example, the effects of a certain chemical or biotoxin can be evaluated in vitro on a small tissue sample rather than in vivo on a living human person to avoid unethical health risks for the human. Similarly, in human mobility data science, working directly with the trajectories of individual humans without their explicit consent is also unethical, as human

trajectories are considered Personal Identifiable Information (PII) and using such data may put individuals at serious risk. For example, such trajectory data could enable an attacker (robber, stalker) to identify where a vulnerable person (for example, a child) may be found alone and without protection.

For human mobility data science, it is difficult to separate data from an individual in a safe way to enable in vitro science, which is the goal of efforts in the field of location privacy (e.g., [29, 47, 105, 107]. This task is difficult because trajectories are very unique to individual humans. It has been shown to be sufficient to use four spatial points to uniquely identify most individuals, even among a large population of people [17, 90]. Thus, despite efforts to anonymize the data, privacy attacks remain a risk.

In addition to in vitro science, the concept of in silico science has been used recently as a paradigm of scientific discovery.

Definition 1.2 (In Silicio; Latin: "In silicon"). In silico refers to a process or studies that are performed entirely via computer simulation [93].

In biology, in silico science simulates parts of the human body (such as individual cells) and allows the study of the interaction of cells with chemicals without any risk to any (real, not simulated) living being [77]. In the case of human mobility data science, the organism that we want to study is an entire population such as that of a city or another universe of discourse (the world that we want to study). We can take advantage of massive microsimulations of individual humans to create a digital twin of this world. In this in silico world, we have complete access to all of the atoms (the individual agents). Furthermore, we can use the simulation as a sandbox that allows us to make changes to the simulation to see how policy interventions may affect the population over time. Just like in biology and other domains, results obtained from in silico simulation may not fully or accurately predict the effects in the real world. However, in cases where an in vivo (directly in the living body) analysis is too intrusive or even impossible (in the case of individual human mobility data due to privacy), and an in vitro analysis may be difficult on a large scall (in the case of individual human mobility data due to the risk of deindentification even using only a small part of human mobility traces) an in silico analysis allows us to evaluate hypotheses and gain an understanding that may likely be reflected in the real world.

1.2 Challenges for In Silico Human Mobility Data Science

For in silico human mobility data science to be effective, we require a simulated in silico world that mimics the real world. This is challenging, as it requires the microsimulation of realistic individual behavior. As Nobel Prize laureate Murray Gell-Mann famously said, "Think how hard physics would be if particles could think" [76] and as Richard Feynman added, "Imagine how much harder physics would be if electrons had feelings". In our in silico world, the particles are the simulated individual humans, which may act irrationally in the real world. Creating a simulation that captures realistic human behavior (and thus, resulting individual human mobility) requires close collaborations with experts in the social sciences. Towards such a socially realistic massive microsimulation of individual human mobility, Section 2 surveys existing data sets of individual human mobility and their limitations in terms of size and representativeness. Then, Section 3 surveys existing simulation frameworks to generate individual human mobility data and the limitations of state-of-the-art to generate individual human mobility data that captures all of the following: 1) realistic human behavior, defined as having the locations visited by individuals grounded in a realistic purpose that leads to visiting locations such as going to work and visiting friends (rather than individuals following random walks), 2) realistic human movement, defined as a realistic motion that takes an individual from one location to another, following

Table 1. Publicly Available Real-World Trajectory Data Sets.

Data set	#Users	# Trajectories	# Points	Distance	Duration
GeoLife	178	18K	25M	1.3M km	54mo
T-Drive	10K	971K	15M	9.0M km	1 week

a transportation network, waiting at traffic lights, and getting stuck in traffic (rather than simply teleporting between locations), and 3) scalability, allowing simulations to scale to large cities having millions of simulated individual humans. Section 4 then describes our vision of a scalable in silico world that captures realistic human patterns of life and allows us to generate massive datasets as sandboxes of human mobility data science. This vision also describes 1) how this in silico sandbox should be easily extensible by non-experts to include new types of human behavior to enable new research directions, and 2) how such in silico datasets should be informed by real-world trajectory and mobility data. Then, Section 5 describes only a small sample of applications and research directions that would be enabled by massive individual human mobility datasets if our vision came true, and Section 6 concludes this vision paper.

2 LIMITATIONS OF STATE-OF-THE-ART DATASETS

2.1 Existing Trajectory Datasets

Real-world trajectory data sets are a scarce resource due to the privacy implications of making such data public. Also, service providers consider such data sets invaluable when it comes to providing a competitive product and are thus somewhat unwilling to give researchers even sizable data sets. Given these considerations, the mobility data science community is grateful for two datasets that have been made available publicly and that have been used widely within the community and that are summarized in Table 1. The first dataset is the Geolife GPS trajectory dataset [135, 136], which was collected and shared by Microsoft Research Asia. This dataset captures detailed trajectories of 178 users in Beijing over a period of over four years (from April 2007 to October 2011). This dataset recorded a broad range of users' movements, including not only life routines like going home and going to work but also some entertainment and sports activities, such as shopping, sightseeing, dining, hiking, and cycling. Although this data set is excellent in terms of quality and fidelity, it is unfortunately very small. It is difficult to infer broad mobility patterns from a set of only 178 users, especially in a large city such as China. The small sample size makes tasks such as event detection, friend recommendation, or contact tracing difficult, as more than 99.99% of the population of Beijing is missing from the data. It is also not clear if this sample is representative, which allows us to infer patterns learned from the sample to the entire population of Beijing.

The second dataset is the T-Drive trajectory data sample [122, 123] which was (also) collected and shared by Microsoft Research Asia and captures one-week trajectories of 10,357 taxis in Beijing. While the number of individuals captured in this dataset is much larger than in GeoLife, T-Drive captures the trajectories of taxis, not individuals. Thus, consecutive trajectories of the same taxi may not correspond to the same passenger. While useful for applications such as traffic prediction, this dataset is very limited in terms of providing insights into individual human mobility and behavior as it is impossible to understand the sequence of places that individuals have visited.

2.2 Existing LBSN data sets

Table 2 summarizes the main characteristics of publicly available data sets that are used extensively by the LBSN research community.

Table 2. Publicly Available Real-World LBSN Data Sets.

Data set	#Users	#Locations	#CheckIns	#Links	Period
Gowalla	319K	2.8M	36M	4.4M	20mo
BrightKite	58K	971K	4.49M	214k	30mo
Foursquare	2.7M	11.1M	90M	0	5mo
Yelp	1.00M	144K	4.10M	0	36mo

Gowalla: Collected and retrieved from the LBSN Gowalla[58], which was launched in 2007 and closed in 2012. This data set has the largest social network of any public LBSN data set, while most of the users are inactive. After removing users with less than 15 check-ins and removing locations with less than ten visitors, more than half of the visitors are eliminated [58]. A similar data set is that of *Brightkite*, which is available at SNAP [95]. As is seen in Table 2, Brightkite is smaller than the Gowalla data in all aspects.

Foursquare: In terms of the number of users and check-ins, the largest publicly available LBSN data set was collected from Foursquare [109]. However, this data set provides no social network information.

Yelp: A large dataset is published by Yelp as part of the Yelp data set Challenge [117]. This data set provides additional information, such as user location ratings, user comments, user information, and location information. Again, this data set does not provide social network information.

Synthetic Check-In Data: The problem of using sparse and noisy real-world LBSN data has already been identified in previous work (e.g., [3, 4, 48]). However, none of these works have proposed a way to obtain plausible check-in data. For example, [3, 4, 48] generated user location check-ins at random using parametric distributions without considering the semantics of the movement. While [85] created additional check-ins by replicating Gowalla and Brightkite data, thus creating more data for run-time evaluation purposes but without creating more information.

2.3 Mobility Flow Data

There exist many other sources of mobility data that are not at individual level but aggregated to regions for which the pairwise flow between regions is reported. Such data includes the Longitudinal Employer-Household Dynamics (LEHD) Origin-Destination Employment Statistics (LODES) dataset for the United States [31], which contains home-work communiting flows aggregated to census tract level. Safegraph Inc [82] made available free access to high-resolution foot traffic data [83] as part of the SafeGraph COVID-19 Data Consortium. This dataset includes check-ins of 35 million anonymized mobile devices in the US, more than 10% of the US population. However, in this dataset, individuals are aggregated to census block groups (CBGs), which are statistical divisions of the U.S. population containing between 600 and 3,000 individuals. As there are no unique identifiers for individuals, it is not possible to combine flows (from one census block group to another) into trajectories, as there is no way to link multiple flows onto the same individual. Thus, SafeGraph does not provide individual check-in data, but only aggregated data that cannot be used to infer individual check-in sequences of individuals. In addition, there are also cell phone trace datasets, which capture the locations of individuals but aggregated to their nearest cell tower and have been used to help understand aggregated human mobility [97]. However, the problem with cell phone traces is that locations are aggregated to nearest cell towers, which is an aggregation even more coarse than using census block groups covering multiple square kilometers even in cities. Therefore, it is impossible to assess individual visited locations based on cell phone trace data.

Summarizing, all data types mentioned in this chapter do provide information on individual human mobility flow, but due to aggregation to large spatial unites or due to a loss of association of individual users to visited locations, these datasets allow inferring specific locations visited by individuals as possible using trajectory data (see Section 2.1) and location-based social network check-in data (see Section 2.2).

Table 3. Existing Simulators for Individual Trajectory Data. By realistic movement, we refer to realistic of *how* individuals move from one location to another, which may include waiting at traffic lights or delays due to traffic; by realistic behavior, we refer to the underlying causality of *why* individuals move between locations, including behaviors such as commuting, meeting friends, and walking their dog.

Generator	Scalable to Millions?	Realistic Movement?	Realistic Behavior?	Adaptive Geography?
Brinkhoff [9]	Yes	No	No	No
ViewNet [45]	No	Some	No	No
BerlinMod [22]	No	No	Some	No
Hermoupolis [78]	No	No	Some	No
MNTG [66]	No	No	Some	Yes
SUMO [44]	Yes	Yes	No	No
Patterns of Life [39, 139]	No	No	Yes	No
Our Vision	Yes	Yes	Yes	Yes

3 LIMITATIONS OF STATE-OF-THE-ART SIMULATORS

Due to the lack of large and representative individual-level trajectory and check-in data, researchers have challenged the representativeness and reliability of existing work on individual human mobility data science [51]. This vision paper aims to close this gap in the future by proposing the means of generating large-scale and socially plausible synthetic data sets through simulation, which we turn to next. Synthetic individual-level data would allow insight into what is possible concerning new and improved geoinformation systems, but also in terms of privacy and anonymization research without raising any privacy concerns.

3.1 Traditional Individual Human Trajectory Simulators

Individual-level microsimulators have been used for decades to represent and analyze individual human mobility. Early work on massive trajectory generation [79] considers the semantics of movement as well as infrastructure information (buildings and obstacles) in the generation process. Brinkhoff's Moving Objects Generator [9] uses a road network file as input and simulates the shortest paths between random vertices in the network. Moving objects then move at a constant velocity along their shortest path. Once objects reach their destination, a new destination vertex is selected at random. However, the simplified movement at constant speeds along the shortest paths and the selection of random locations as destinations is not a realistic model reflective of real human mobility. A similar simulator is ViewNet [45, 46] which imposes a simple traffic model in which people cannot overtake each other on roads. This model allows for the simulation of traffic jams. However, the interactions between individuals lowers the scalability of this simulator to at most a few thousand simulated individuals. The BerlinMod [22] simulator is the first to simulate some realistic behavior, including the commuting of individuals between home and work locations and visiting random places within their neighborhood outside of work. In contrast, the Hermoupolis [78] simulator simulates individual movement to a random destination but simulates coordinated behavior among individuals, such as groups of individuals moving in flocks thus forming clusters of trajectories. Note that in all of the aforementioned simulators, the goal was not to create a realistic world. Their goal was to create benchmark datasets that could be used to evaluate the efficiency of index structures or to cluster flocks of trajectories.

An important extension of the simulators was proposed by the MNTG [66] simulator. While not a simulator itself, MNTG is a wrapper that enriches any of the aforementioned simulators with a web-interface that allows for the selection of a study region, generates a road network and POIs from OpenStreetMap, feeds these datasets to one of the existing simulators and sends the generated datasets for the selected study region via email. The advantage of MNTG is

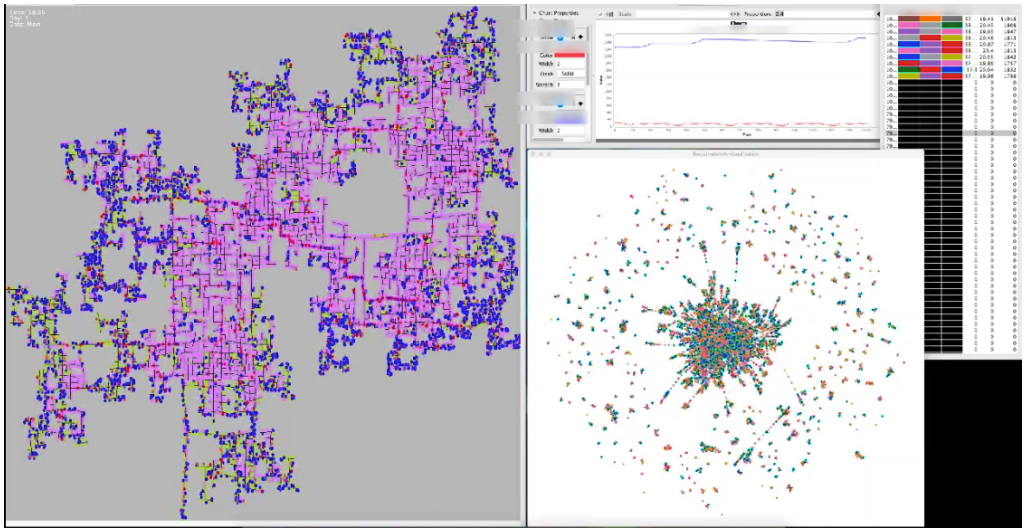


Fig. 2. The Patterns of Life Simulation. A video of the simulation can be found at: https://youtu.be/rP1PDyQAQ5M

that it allows for the extension of existing simulators to any study region of a user's choice, rather than having to use the network dataset(s) provided by the simulators (such as the Berlin network for BerlinMod).

3.2 Realistic Traffic Movement Microsimulation

In addition, there are also microscopic traffic simulation toolkits such as SUMO [44] and MATSim [98] that are used widely to simulate the flow of traffic. Such platforms allow for the simulated movement along road networks at very high fidelity, including realistic traffic (where individual cars may not pass through each other), multiple lanes, traffic lights, and different transportation modes. However, such simulations do not simulate realistic human behavior. That is, the origin/destination pairs that are used to simulate traffic are not simulated but have to be provided through an external file, which can be randomly generated or informed by commuting patterns dataset such as the Longitudinal Employer-Household Dynamics (LEHD) Origin-Destination Employment Statistics (LODES) dataset for the United States.

3.3 Human Mobility Microsimulation

To simulate realistic individual human behavior, a simulation was recently published called "*Patterns of Life*" simulation [39, 139] using Maslow's Hierarchy of Needs [62] as a driver of simulated individual human behavior. In this simulation, driven by physiological, safety, love, and esteem needs, agents perform activities to satisfy these needs by moving across a spatial road network to find places to eat, work, take shelter, engage in recreational activities, meet friends, and return home to be with family and sleep. By meeting others, agents make new friends and strengthen their friendships. When meeting at places, agents form social links which, in turn, affect the places agents visit. The purpose of this model is to act as a digital sandbox environment for social scientists to assess different methods and tools for analyzing complex social phenomena in an agent-based simulated world. The simulation was used in DARPA's Ground Truth Program as a "plausible alternate world" in which we have perfect knowledge of the ground truth of how the world works, how individuals make their decisions, and how social ties are formed.

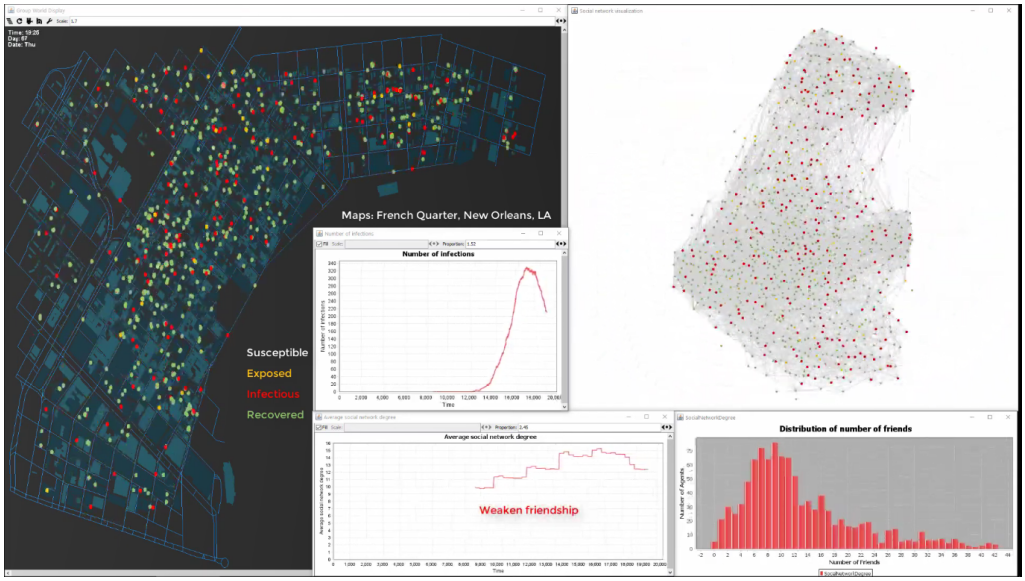


Fig. 3. Patterns of Life Simulation of the French Quarter, New Orleans, LA. A video of this simulation can be found at: https://youtu.be/zXs_1MWItX4

Figure 2 shows a screenshot of this simulation and provides a link to a video of a simulation run (the spatial network populated with agents on the left and the social network between agents on the right). As agents go to work, restaurants, and recreational sites, they meet with new friends and create a social fabric. A detailed description of this simulation can be found in [139]. The Patterns of Life model used synthetic environments and synthetic population [40]. This choice was deliberate to prevent other performers[1], who were tasked to explain, to predict, and to prescribe changes to this world

Figure 3 shows a screenshot and provides a link to a video of a simulation of the French Quarter, New Orleans, LA, using real-world road and building data from OpenStreetMap [75] and U.S. Census information to initialize the agent population. Again, the location of agents in the geographical space (located at places of interest and across the road network) is shown on the left and their location in the social network is shown on the right. Using this model, we simulated the outbreak of an infectious disease. The color of agents (both in geographical and social spaces) corresponds to the disease status of an agent using the SEIR (Susceptible, Exposed, Infectious, Recovered) model. This simulation allows tracking the spread of diseases simultaneously across the geographical map and the social network. Other graphs in Figure 3 show the time series of infected agents (center), the time series of the average number of friends per agent (bottom, center), and the current distribution of the number of friends per agent (bottom, right). Note that this is **not** a COVID-19 simulation. This simulation was created and published [42] in 2019 - before COVID-19 was discovered. The contagion used in this simulation was a generic flu-like disease.

Although the goal of this simulation was to represent socially plausible human behavior, agents in this simulation move along the shortest paths (at constant speed) to reach their intended destinations without any realistic movement or

[1]The simulation was designed for DARPA's Ground Truth Program. In Ground Truth, Technical Area 1 performers developed simulators for artificial but socially-plausible worlds. Technical Area 2 performers that did not know how simulations worked had to explain how the world works, predict what would happen in the future, and prescribe some remedy to achieve certain goals (e.g., minimizing the number of infected agents). Here, 'other performers' indicate TA2 teams.

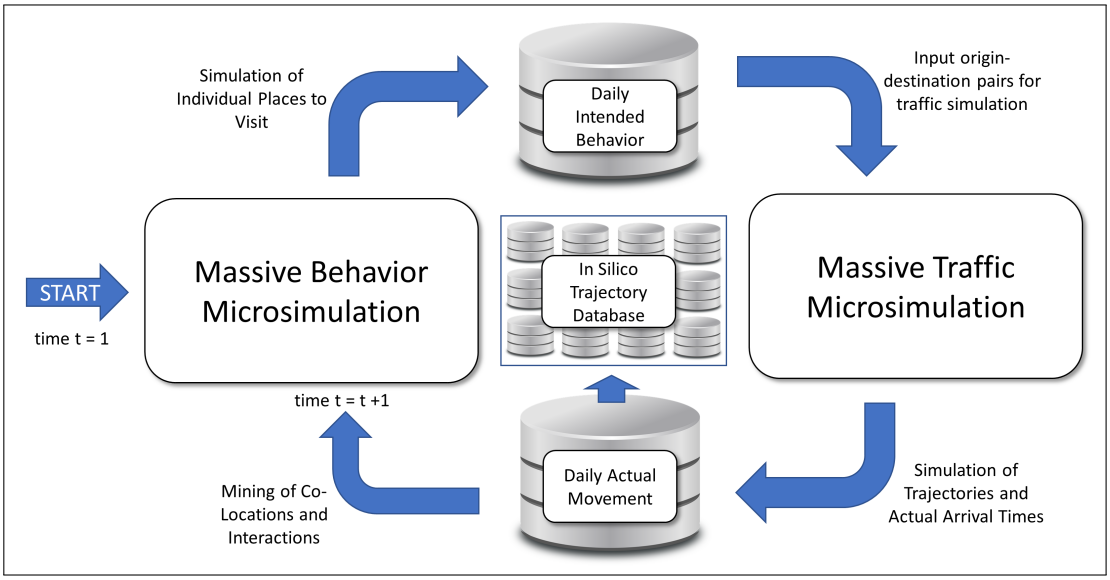


Fig. 4. Envisioned framework for a simulation that exhibits both realistic behavior and realistic movement.

traffic model. Also, due to the simulated complex human behavior based on agent needs on agents' social networks, this simulation can only scale to a few thousand individual agents. In addition, while realistically modeling an individual's needs to visit places, this simulated world may not reflect the real world, as "all models are wrong, but some are useful" [8]. For example, it may not reflect that some individuals may have a favorite restaurant or a hairdresser they have been going to for years. But informed by real-world data about the environment (road network and buildings of a city), we hope that such a simulation, while not being a true representation of the real world, can be representative of some aspects of a real population to generate realistc trajectories.

4 A VISION OF SCALABLE SIMULATION OF REALISTIC MOVEMENT AND BEHAVIOR

As we observe from Table 3, there exist individual-level trajectory simulations that can be scaled to millions of individuals. However, these simulations lack realistic behavior (as in what, when, and why to visit certain places) and lack realistic movement (as in how to move between locations). We also see that there are traffic microsimulations that capture realistic mobility, including traffic jams and traffic lights. But in these simulations, the simulated origin and destination pairs of individuals are chosen at random or extrapolated from aggregated flow datasets as surveyed in Section 2.3. There are also simulations that have agents exhibit realistic human behavior, but these simulations do not simulate realistic movement and are not yet scalable to millions of agents.

Our vision is a simulation that satisfies all three requirements, as follows: 1) scalability to millions of agents, 2) realistic human behavior, and 3) realistic human movement. From this simulation, individual-level trajectory mobility datasets can be generated that could be leveraged for human mobility data science research without risk to privacy for real human individuals. We envision that this could be achieved through a framework depicted in Figure 4 that combines current capabilities for realistic individual behavior simulation (such as in [39, 139]) with current capabilities for realistic individual movement simulation (such as in [44]). For the purpose of scalability, we propose iterating between a closely coupled realistic behavior simulation and a realistic movement simulation. At the beginning of each

simulation day, the individual human behavior simulation creates a daily plan for an agent prioritized by the needs of an agent. For example, such a daily plan of an agent may consist of the following: 1) Wake up at 6:00am, 2) Go to work at 8:00am, 3) Go to a restaurant for lunch at 12:00pm, 4) Go to a bar to meet friends at 6:00pm, 5) Go home at 10pm, with a list of priorities such as: I) work for eight hours, II) eat food, III) meet friends. This agent plan and set of priorities are then passed on to the traffic simulation generate movement data. During the execution of the traffic simulation, delays may occur. For example, the individual may arrive at work one hour late. The prioritization of the behavior simulation will then be used to decide that, in order to work for the full eight hours (which is the agent's first priority obtained from the behavior simulation, for example, due to a pressing financial safety need), the agent has to skip lunch or skip meeting friends. Prioritization will cause this agent to go to the restaurant, but skip going to the bar as a consequence of arriving late at work. The movement simulation will yield trajectories and check-ins of the current day. These are then passed back to behavior simulation to inform decision-making for the next day. The idea here is to inform the behavioral simulation with information about what has actually happened. For example, in the example above, the simulation will be informed that the individual was able to satisfy their financial safety and food needs, but did not meet any of their friends, such that the social links that would be reinforced through the planned meeting are not actually reinforced, and the individual will also not meet any new friends they would have met if they had gone to the bar instead of working longer. Once the behavior simulation is updated giving the actual movement of the previous day, the loop in Figure 4 restarts by planning the next day for each agent.

To achieve realistic human behavior, an in-silico world also needs to consider a realistic environment including daily weather changes and seasonal patterns. We know that weather conditions significantly affect traffic [122]. General human behavior such as going for a walk, going to build a snowman, or having ice cream also depends on weather conditions. To the best of our knowledge, there is no existing large-scale agent-based simulation of human behavior that simulates weather conditions and seasonal weather and temperature patterns. But weather patterns can be learned and simulated for any city (using history weather patterns) and we hypothesize that including weather will be paramount in obtaining realistic human mobility data.

4.1 Simpifying Model Building

A necessary step in white-box model-based approaches is to build models and configure their parameters. However, this often requires expertise in modeling and can be very time-consuming for even expert modelers. To accelerate this process, we envision an advanced user interface beyond traditional programming languages or scripts. Human vision is the most efficient sensory system to comprehend a large amount of information in a short period of time. As such, well-defined diagrams are an effective means to conceive model concepts. Taking this into account, a model builder can provide a graphical user interface (GUI) that visually encodes entities, relationships, and constraints [41].

Although a GUI is intuitive, editing models can still be overwhelming, especially when the models are complex and require many template constructs. The modeling and configuration process requires multiple iterations. To advance the iterative process, we envision a multimodal large language model (MLLM) [21] to create an initial graphical model and update the model via chat and voice interfaces similar to ChatGPT [71, 72]. Large Language Models (LLM) (e.g., ChatGPT, Llama2 [63]) pre-trained from a vast corpus of text data (e.g., 500B tokens of code for Llama2) can capture complex context from user queries and generate comprehensive embeddings (e.g., 70B parameters for Llama2).

LLMs are also at the heart of AI-assisted programming tools such as Github Copilot [30] and Amazon Codewhisperer [1], which, when provided with a programming problem in natural language, are capable of generating solution code.

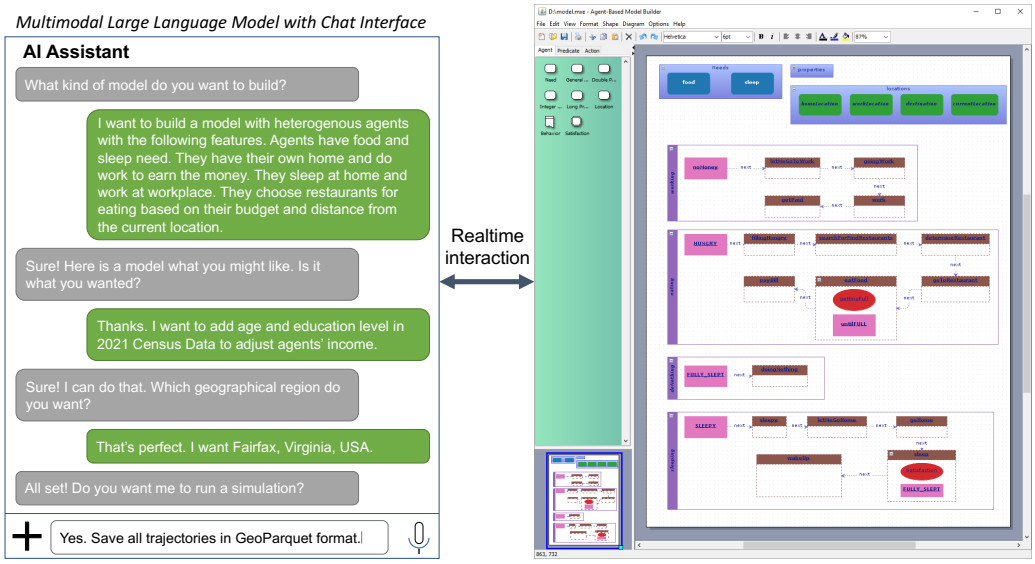


Fig. 5. Interactive model builder powered by multimodal large language model with chat interface .

Amazon Codewhisperer even introduced Amazon Q, an interactive generative AI-powered assistant that provides guidance such as code explanations, transformations, and suggestions through a simple *conversational interface*.

In envisioning such a conversational approach, Figure 5 illustrates how a modeler can interact with an AI assistant to build a model and configure simulations through the chat interface. Compared to earlier attempts to integrate natural language processing (NLP) [91] or natural language understanding (NLU) [92], our vision is to take advantage of a prompt incorporated into the GUI model builder. The AI assistant can be considered a mediator between the GUI model builder and the user. State-of-the-art MLLMs [106] that interact with modelers will dramatically improve processes and productivity. The more context we can provide in this process, the better and more specific the final model will be. Recent ChatGPT improvements, such as a code interpreter or Advanced Data Analysis [73] enable downloading and visualizing data within a prompt, as well as browsing available datasets. Key capabilities of the AI assistant for model building can be summarized as follows:

- Manipulate (load/save/add/update/remove) model components.
- Recommend/suggest datasets or components.
- Download/catalog datasets.
- Configure parameters and settings.

To the best of our knowledge, the idea of AI-assisted world building has never been used in the model and simulation community and has the potential to disrupt the field of simulation and modeling. Specifically, such an approach promises the bridge the gap between domain experts (who know what to model) and system builders (who know how to implement it).

4.2 Simulation vs. (Deep) Generation

This work is motivated by the fact that since large trajectory datasets are sorely missing, realistic synthetic trajectory datasets are of great importance. The authors of [132] propose a new framework called "End-to-End Trajectory

Generation with spatiotemporal validity constraints" (EETG). The EETG framework employs factorized latent sequential deep generative models to disentangle global and local semantics while *learning trajectory representations in an end-to-end fashion*. Global semantics capture trip reasons, e.g., commuting, a stroll downtown, airport pickup. Local semantics are related to a specific location and time, and capture spatiotemporal autoregressive patterns, effectively modeling the dependency of subsequent trajectory samples. To ensure that this generative process also produces trajectories that resemble real-world trajectories, i.e., obey motion physics, the work in [132] also introduces a novel constrained optimization solution that reduces the probability of generating invalid and irrational vehicle trajectories, such as speed limits in combination with turn angles.

Figure 6 provides a heatmap representation (stacked semi-opaque points representing trajectory position samples) of input/original trajectory datasets (Figures 6a and 6c) and the corresponding generated datasets (Figures 6b and 6d) for Beijing (T-drive data [123]) and Porto, Portugal [38].

The results of Figure 6 and other examples presented in [132] show that end-to-end trajectory generation without any knowledge of the underlying road network can produce sets of trajectories that resemble network-constrained movement. As such, end-to-end generation is a viable alternative to simulation-based generation.

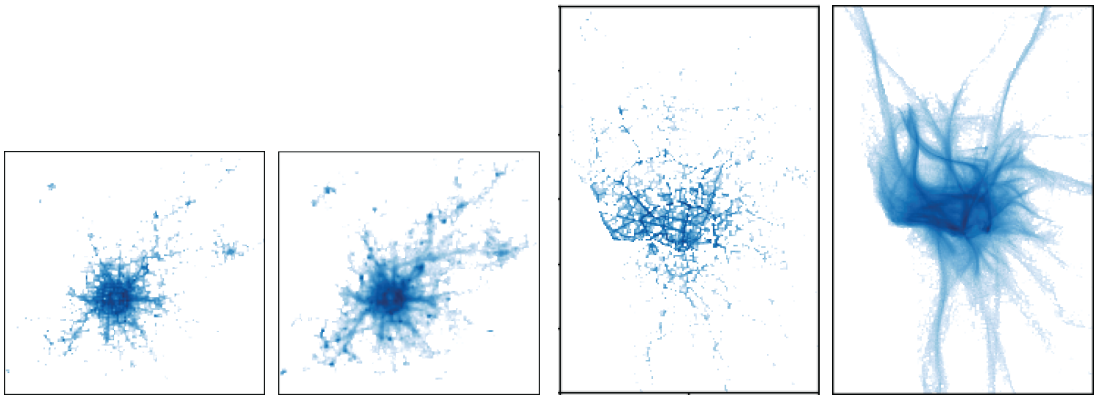
But one of the challenges in deep learning generative model is hallucination, that is, generating responses that are either factually incorrect, nonsensical, or disconnected from the input prompt. To ensure that this generative process also produces trajectories that resemble real-world trajectories, i.e., obey motion physics, the work in [132] also introduces a novel constrained optimization solution that reduces the probability of generating invalid and irrational vehicle trajectories, such as speed limits in combination with turn angles. As such, end-to-end generation is a viable alternative to simulation-based generation in scenarios where it is not critical to explain underlying motivation and decision-making process of human behaviors. The more context we can provide in this process, the better and more specific the final model will be. The AI assistant can be beneficial for non-domain experts as well as domain experts. Model components developed by other users can be integrated using the model builder, and the AI assistant can suggest model components regardless of one's expertise. Even if users adopt model components outside of their expertise, their model is explainable because such simulation models consist of white box components unlike black box models.

4.3 Bias and Ethical Considerations of AI-Generated In-Silco Worlds

Our main idea of using an AI assistant to support the building of an agent-based model is to abstract the implementation aspect and allow non-technical users, such as social scientists and epidemiologists to implement aspects of human behavior without having to write code. But one can take this vision one step ahead, and use an AI to generate the rules of human behavior for the AI itself to implement. Such an approach would take the human out of the look and let an AI generate an in-silico world that reflects the AI's understanding of human behavior.

Aside from the computational challenges of a model of human mobility that not only reflects behavior relevant to a research area (such as the spread of an infectious disease) but all human behavior that an AI can "think" of, we also see potential challenges concerning data bias and privacy.

Data Bias: Abstractly, an AI captures human knowledge that was used to train it. This human knowledge is often taken from public textual data sources such as Wikipedia or news sources. But such sources are often biased to high entropy (or interesting) events. For example, news articles may report criminal activity in one city, but won't report the lack of a such activity in another city. Using such data, that is biased to high entropy events may cause the AI to overestimate the frequency of such events. Data generated from such an AI-modelled (and implemented) simulation may thus overestimate the frequency and magnitude of rare events. Using such a dataset for research may confirm



(a) Beijing - original (b) Beijing - generated (c) Porto - original (d) Porto - generated

Fig. 6. Trajectory datasets: point count heatmap comparison for original and generated trajectory datasets [132].

Application	Data Available?	Representative Data?	Unbiased Data?	Spatial Information?	Temporal Information?	Demographic Information?
Social Link Prediction	✓	✗	✗	✓	✓	✓
Disaster Relief	(✓)	✗	✗	(✓)	(✓)	(✓)
Epidemiology	(✓)	✗	✗	(✓)	(✓)	(✓)
Mobility Modeling	✓	✗	✗	✓	✓	✗
Benchmark Data	✓	✗	✗	✓	✓	✗
Location Recommend.	✓	✗	✗	✗	✓	✓
Community Detection	✓	✗	✗	✗	✓	✓

Table 4. Properties of real-world datasets available in different applications described in Section 5.

spurious hypothesis relating to such rare events. For example, a study modeling and simulating the number of crime events may yield an "interesting" result only because the simulation it was trained on overrepresents interesting data.

Data Privacy: Another problem with having the AI not only guide the implementation but also the modeling is the problem of data privacy. A big advantage of our envisioned in-silico simulation is that simulated individual agents to not related to any real-world humans, as all data is based on aggregate information such as census data. But by using an AI (and thus, the knowledge stored in the AI using a large language model) to generate worlds, we may indeed capture information about real-world individuals that may have been captured in the data used for training the large language model. Thus, any claims that research using such an in-silico world does not include any personally identifiable information may not longer hold.

Due to these considerations, we would like to advice caution in using a large language model-based AI in the modeling of an in-silico world for human mobility data science.

5 APPLICATIONS AND CHALLENGES FOR HUMAN MOBILITY DATA SCIENCE USING SCALABLE AND REALISTIC IN SILICO WORLDS

Consider an in silico world that represents the realistic behavior and movement of a large number of individuals and generates large sets of individual-level mobility data from this world. This section describes some possible applications

Table 5. Data Sets Resulting from Location-Based Social Network Simulation

Settings	Period (mo.)	# of CheckIns	# of Links	Runtime (min)	Size (MB)
GMU-1K	15	2,082,788	9,114,337	53	398
GMU-1K	121	16,210,909	75,747,439	372	3,235
GMU-3K	15	6,229,293	27,650,685	342	1,247
GMU-5K	15	11,189,377	54,250,961	1,099	2,389
NOLA-1K	15	2,099,867	9,160,459	52	400
NOLA-1K	221	29,597,885	141,425,945	774	5,502
NOLA-3K	15	6,886,573	27,284,999	362	1,282
NOLA-5K	15	12,007,415	48,710,881	1,233	2,284
TownS-1K	15	2,101,620	7,643,374	44	359
TownS-3K	15	6,454,785	26,364,057	319	1,227
TownS-5K	15	10,760,008	45,118,825	867	2,093
TownL-1K	15	2,030,688	6,418,473	49	320
TownL-3K	15	6,340,360	22,655,915	400	1,109
TownL-5K	15	10,548,956	40,431,579	942	1,937

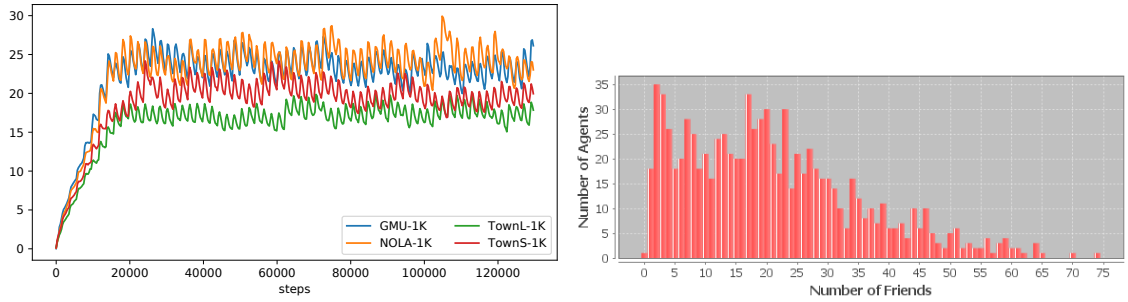
using such data for broad impacts in science and everyday life. To illustrate how in silico data collection may support research in each of the following application domains, Table 4 shows the limitations of publicly available real-world datasets for each application which can be addressed by using data generated in silico. For example, in all of the following applications, real-world datasets pertain to a small subset of the population selected by participation in a corresponding data collection application (such as a location-based social network or a contract tracing application). We know that such data will overrepresent the technically savvy and underrepresent old and vulnerable populations. By using data simulated in-silico, we can evaluate how such lack of representative data collection may affect applications such as community detection and contact tracing. For each application, details are described in the following.

5.1 In Silico Human Mobility Data Science for Social Link Prediction

This use case envisions a social link prediction data science challenge using existing simulation capabilities detailed in [39]. The idea is to generate a very large set of individual trajectories and check-ins (many orders of magnitude larger than existing real-world datasets) along with the dynamic social network of friendship between the simulated individuals. Using this dataset as ground truth, we can publish a sample of "observed" trajectories, check-ins, and social connections to challenge the community to solve the problem of estimating missing social connections that are not captured in this sample.

Example ground-truth data sets for this use case can be found at OSF (https://osf.io/e24th/). Due to the excessive size of some of these LBSN data sets introduced in the following, we recommend that researchers interested in using the data re-run the simulation locally instead of downloading the data directly. Parameterized executables are available for download (https://github.com/gmuggs/pol), and our simulation is fully serialized and deterministic, such that the data generated locally is guaranteed to be identical to the downloadable data. Details on the underlying simulation to model realistic human behavior can be found in [139] with specific settings (including simulation initialization) to generate realistic datasets detailed in [39].

The goal of the generated data sets discussed in the following is to act as benchmark data sets for the LBSN community, where publicly available datasets (as surveyed in Table 2) are very small and biased to a small subset of the population



(a) Average Social Network Degree over Time (1K).

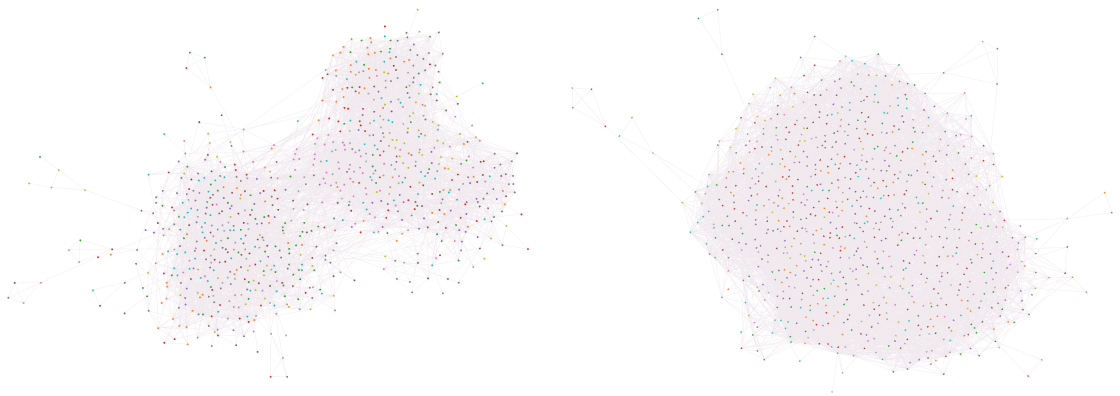
(b) Social Network Degree Distribution - TownS-1K

Fig. 7. Characteristics of the simulated social network for the TwonS-1K simulation.

using the respective data collection phone applications. We generated a mix of real and synthetic urban settings. Real road network and point-of-interest data were obtained from OpenStreetMap (OSM) [74], where we downloaded data for the greater New Orleans, LA (NOLA) metropolitan area, and the George Mason University (GMU) office for Geo-Information Systems (GIS) Facilities Archives [28] provided us with data for the Fairfax, VA campus of GMU. We also generated two synthetic urban datasets of different size and layout denoted as *Small* (TownS) and *Large* (TownL). These synthetic urban components were created using a spatial network and a place generator based on a generative grammar similar to L-systems described in [40].

Within NOLA, the area we concentrated on was the historic French Quarter (FQ). The GMU area captures the main campus in Fairfax, VA. Both areas were prepared using QGIS desktop GIS software [81] and JOSM [37]. Data preparation involved editing the data sets to produce three separate shapefiles [23]: (i) building footprints, (ii) transportation networks (road and sidewalk layers), and (iii) building purpose (i.e., residential, commercial, etc.).

Table 5 provides an overview of the generated output data from the location-based social network simulation. It shows the number of agent check-ins and the number of social links attributed to each of the scenarios. We observe that the number of check-ins increases for all study areas linearly with the number of agents. This is plausible, as the number of hours per day that agents can spend to satisfy their needs and visit sites is independent of other agents. However, we do see that the number of social links increases super-linearly with the number of agents. This can be explained by more agents leading to larger co-locations of agents, creating chances for each *pair* of agents in the same co-location to become friends. We note that the generated temporal social network may have more edges than we have agent pairs. This is due to the temporal nature of the network. It reports changes over time and as such a single pair of agents can have multiple friends and unfriend events. The reported number corresponds to the number of new edges added to the temporal social network, regardless of the duration of these events. The super-linear growth of the social network also explains the super-linear run-time to create each data set, ranging from less to one hour for the 1000 agent instances to 10.5h for 5000 agents. Besides (i) the number of check-ins and (ii) social links, we also report (iii) the run-time of each simulation and (iv) the resulting data size in Table 5. It is interesting to see that even small simulations can create sizable results with a longer duration. For example, the smaller synthetic urban component that had the longest (221mo) simulation period produced a 5.5GB result data set. However, the actual run-time of this simulation is shorter than a simulation with more agents that had a shorter simulation period, e.g., NOLA-5K - 15mo produced only a 2.3GB data set and took 1233min to run vs. NOLA-1k - 221mo produced a 5.5GB data set with a run-time of 774min. Increasing the number of agents results in more complex data structures, e.g., social networks, which in turn increases



(a) TownL-1K (b) NOLA-1K

Fig. 8. Simulated social networks for two different in silico worlds.

the run-time of the algorithms to process them at each step of the simulation. Figure 7a shows the average number of friendships per simulation time measured in 5-minute steps for all the 1K networks. In all cases there is a three-month (one month is equal to $12 * 24 * 30 = 8640$ steps) settling time during which agents establish friendships (starting with an empty social network). After this phase, the friendship degrees fluctuate around the mean values for each simulation. We observe that the two real networks exhibit a denser social network, due to a more uneven distribution of agents, leading to large groups of agents to co-locate at sites to become friends.

For a more detailed view of the resulting social network, Figure 8 shows two visualizations of the social networks of 1K agents exemplary for the large synthetic network and NOLA at the end of the 15mo simulation. These visuals show different types of network structures, such as two to three large social communities for the synthetic TownL, and one large community for NOLA. Since it is hard to describe the evolution of a social network over time, we have created a video for each of the four spatial areas showing the social network evolution over the 129, 600 steps within the 15mo simulation time. These videos can be found at https://mdm2020.joonseok.org and show how social networks evolve from small isolated cliques into a large and complex network showing different sub-structures. The video can also help explain the patterns observed in Figure 7a in which agents make friends during the day, while losing some at midnight at which time we periodically lower the weights of the social network.

These videos also show, at each step of the simulation, the distribution of the number of friends per agent, an exemplary one is shown in Figure 7b for the small synthetic network (TownS-1K). We observe that all case studies exhibit realistic long-tail distributions of the number of friends: While most agents have 5-25 friends, there are outliers having 50+ friends, but also agents having only three or fewer friends. This observation agrees well with a limit on the number of people with whom one can maintain stable social relationships (cf. the Dunbar number [138]). In our simulated world, we observe an emerging limit of about 35 friends resulting from agents striving to maximize their number of friends (to satisfy their love need) while having to prioritize lower needs in Maslow's hierarchy (sleep, food, money). More details on how agents become friends through collocation and and how friendship ties decay over time if not maintained can be found in [139].

5.2 Emergency Response

Emergency response is driven by planning in relation to different crisis scenarios and informed by considerable amounts of information and data. No unifying model is available that integrates the available information and goes beyond a static view of a community and its infrastructure. We can use in silico mobility science in emergency preparedness exercises to better gauge situations and to simulate potential response scenarios, e.g., experimentally evaluating evacuation scenarios due to extreme weather or due to an anthropogenic disaster. By capturing the complex dynamics of human behavior during emergencies, simulations can help identify potential bottlenecks, evacuation challenges, and areas that require additional resources. This information can be used to refine hazard mitigation plans and allocate resources more effectively, ultimately reducing the potential impact of disasters.

Simulating emergency events in our in silico world allows us to study the behavior of simulated individuals given different emergency response strategies. Many people may immediately leave the city using a shortest-path approach, but others may first want to ensure the safety of their dependents, such as children and the elderly. Other individuals may simply stay at home and forego evacuation due to their beliefs, or due to a lack of communication of the situation (e.g., due to language or technological barriers which can be simulated). Such a result would provide us with a very large trajectory dataset of all individuals capturing their behavior and movements. Data may show bottlenecks (for example, in traffic or communication) that may be difficult to predict without a simulation, or using existing evacuation simulations that assume perfect (yet unrealistic) behavior in which all individuals immediately seek the shortest path out of the evacuation area [7, 94].

In addition to understanding "what will happen?" in a disaster scenario, our in silico world also allows us to run experiments directly in the simulated world by using prescriptive analytics focusing on answering "how to make X happen?" questions. For example, we can evaluate the effectiveness of different traffic management policies and different strategies to effectively communicate evacuation notifications to individuals and optimize optimal strategies.

In this specific context, an in silico simulation can become the computational foundation of a digital twin, which can serve as a powerful tool to raise public awareness of hazards and mitigation strategies. By visualizing the potential impacts of disasters and the benefits of mitigation measures, the digital twin can engage and educate citizens, encouraging their participation in community preparedness efforts.

Improved emergency preparedness and response strategies can lead to reduced response times, more efficient resource allocation, and minimized damage. Raising public awareness encourages citizens to actively participate in related initiatives, building more resilient communities. A better prepared community will experience fewer casualties and reduced property damage during emergencies, leading to significant savings, especially considering the long-term impacts of human losses and the high costs associated with property damage and recovery efforts. Furthermore, the intangible benefits of an in silico "sandbox" would include enhanced public safety, improved inter-agency collaboration, and increased public awareness.

5.3 Infectious Disease Spread Prediction and Digital Contract Tracing

During the outbreak of an infectious disease, contact tracing allows health officials to warn individuals who may have been exposed to an infectious disease. Digital contact tracing applications allow to systematically detect such contacts, even contacts with strangers that an individual may not be able to remember. Unfortunately, due to the COVID-19 pandemic, it has been shown that contact tracing applications were ineffective due to an insufficient adoption of the population [10]. In an in-silico world, we can learn how underrepresented groups in contact tracing applications will affect the spread of an infectious disease by giving insights on what we don't know in the real world.

We can simulate the spread of an emerging infectious disease and evaluate optimal policies to mitigate the spread. By having a ground truth of all infection states of all agents (in our in silico world) we can evaluate how much sampled data we need to be able to successfully trace and mitigate a disease. For example, we know that most contact tracing apps have an uptake rate of less than 10%. That means that for any disease transmission event, we only have a chance of $0.1^2 = 1\%$ of capturing both the infector and the infected in the contact tracing app [65]. To answer the question of what degree of participation of a contact tracing app is needed to be effective has been tackled in recent work [126] using up-sampling of real-world data (which only captures a small portion of the ground truth population). Such up-sampling however, assumes that the data is a representative example of the population, whereas we know that real-world human mobility datasets (see Section 2) and real-world infectious disease datasets are a biased sample of the population [32]. Having a ground truth of all individual human mobility data as well as a ground truth of all infections would allow us to gain a deeper understanding of the capabilities of different infectious disease spread prediction and contact tracing models.

5.4 Human Mobility Modeling

High fidelity, fine-grained, modeling of human mobility is of critical interest to the Intelligence Community, in particular for establishing models of "normal" movement capable of encoding the diversity of human movement present across times, locations, and people [18]. Having realistic models of normal models may allow for the identification of anomalous movements such as loitering or circling around an area or event of interest, unusual movement between locations, unusual volume of movement to a point of interest, or unusual congregations of individuals. But real-world human mobility datasets are very sparsely sampled and biased to users for specific apps.

Current research for understanding human mobility only provides a high-level understanding. The key limitation in achieving this goal is the lack of ground-truthed mobility datasets. But without full knowledge of the underlying activities, it is impossible to characterize what movement is practically detectable as anomalous. In silico individual mobility data science will allow us to understand what types of activities can be modeled and provide datasets that are grounded in these activities, thus allowing the generation of mobility datasets where each trajectory is labeled with the ground truth activity that caused the individual to perform the observed mobility. Having such a ground truth will allow for testing our capabilities of modeling such activities and of identifying outliers. For example, we may simulate the trajectories of a pickpocket circling near a landmark, and test our abilities of identifying such behavior directly from trajectory data.

5.5 Trajectory Benchmark Data.

For a fair comparison of different research methods, it is paramount to compare solutions on the same data sets. As discussed in Section 2, publicly available data sets lack volume, temporal information, and ground truth to reliably generalize knowledge that can be mined from them. As an alternative to using these existing data sets, the mobility data science community has proposed solutions to efficiently crawl data from LBSN data providers [36]. However, this data is the intellectual property of the respective LBSN providers, and publishing their data for research benchmarks will violate their license agreements. As it is not possible to crawl data from the past, researchers will ultimately find themselves comparing their solutions on similar, but not identical data sets crawled at different times. Generated data from our in silicio simulations can fill this gap. They allow different research groups, at different times, to evaluate their algorithms for different problems on the same data sets. Furthermore, simulated benchmark data is extensible. If researchers choose to use a simulation to generate a new data set for their particular application, then the corresponding parameter file can be added to a repository. However, for very large trajectory and LBSN data sets, which may exceed 10TB of

filesize, instead of downloading the data directly, researchers can share their data by only providing the self-executable simulation for local re-generation of data due to bandwidth constraints. Towards this vision of large-scale simulated trajectory benchmark data, a first Data and Resource paper has recently been published [2] providing terabytes of trajectory data (more than a million times larger than existing trajectory data sets), simulating four different regions of interest for tens of years of simulation time. The simulation can easily be adapted to new study regions as described in [43].

5.6 Location Recommendation.

While there exists a large body of work on the problem of location recommendation (e.g., [6, 13, 48, 100, 115]) these works use the publicly available datasets surveyed in Section 2. Due to the small sampling size of these datasets, it becomes difficult to understand how the proposed solutions would generalize to a full population. Using an in-silico world, we are able to bridge this gap.

To recommend locations, a simulation may allow agents to *rate* sites on a one to five-star scale. This rating could be determined by a deterministic function of the agents' preferences and the locations' attributes. To leverage this simulation for location recommendation, our simulation can be extended to obfuscate ratings by random noise (of parameterizable degree). This obfuscation can be deliberately biased, such as giving low ratings a higher chance to appear, with medium ratings more likely to be omitted. Such data would allow researchers to experimentally compare existing methods and evaluate the effect of bias between observed and ground truth ratings. Such comparison enables us to answer the question of recommendation systems' generalizability to the whole population, or if they overfit their models towards a sub-population of individuals that use the recommendation service.

5.7 Community Detection.

For the task of Community Detection (often referred to as social network clustering), existing work (e.g., [103, 118]) also uses the relatively small datasets surveyed in Section 2. This makes it difficult to understand if the communities found on these small datasets are representative of the entire population. To leverage in-silico human mobility simulation for community detection and social network clustering, we can extend the simulation to impose circles of friends (i.e., strongly connected groups) in our social network. Then, by observing co-locations from the data, we can see which existing solutions can best approximate the imposed ground truth social networks. This data generation provides the ground truth for communities which can be used to evaluate the accuracy of community detection algorithms.

6 CONCLUSION

Human mobility data science using individual-level data has the potential to improve our understanding of human behavior and has many applications with broad impacts on society. However, existing datasets of human mobility are too limited to trust results and generalize to broad populations and existing simulation frameworks do not capture realistic human behavior. The vision presented herein is to bridge this gap and reignite research on mobility data science by breaking the chains of small data sets and embracing in silico human mobility data science. In silico human mobility data science can leverage socially realistic and scalable simulation to generate massive datasets of individual-level trajectories. We envision a world builder which allows lay users to rapidly create an in silico world that captures human behavior of interest to be studied, and which allows learning from and adapting to samples of real-world trajectory data. Based on such in silico worlds for which we have limitless human mobility data, we describe applications that transcend the boundaries of current human mobility data science, such as broadly identifying friendship from (co-)locations, recommending locations to users, and identifying trajectories that correspond to malicious behavior.

ACKNOWLEDGMENTS

Notice: This manuscript has been authored by UT-Battelle, LLC, under contract DE-AC05-00OR22725 with the US Department of Energy (DOE). The US government retains and the publisher, by accepting the article for publication, acknowledges that the US government retains a nonexclusive, paid-up, irrevocable, worldwide license to publish or reproduce the published form of this manuscript, or allow others to do so, for US government purposes. DOE will provide public access to these results of federally sponsored research in accordance with the DOE Public Access Plan (http://energy.gov/downloads/doe-public-accessplan).

REFERENCES

[1] Amazon AWS. Amazon CodeWhisperer. (https://aws.amazon.com/codewhisperer/), 2023.
[2] H. Amiri, S. Ruan, J.-S. Kim, H. Jin, H. Kavak, A. Crooks, D. Pfoser, C. Wenk, and A. Zufle. Massive trajectory data based on patterns of life (best data & resource track paper award). In *Proceedings of the 31st ACM International Conference on Advances in Geographic Information Systems*, pages 1–4, 2023.
[3] N. Armenatzoglou, R. Ahuja, and D. Papadias. Geo-social ranking: functions and query processing. *VLDB Journal*, 24(6):783–799, 2015.
[4] N. Armenatzoglou, S. Papadopoulos, and D. Papadias. A general framework for geo-social query processing. *Proc. of the VLDB Endowment*, 6(10):913–924, 2013.
[5] R. Assam and T. Seidl. Check-in location prediction using wavelets and conditional random fields. In *ICDM*, pages 713–718. IEEE, 2014.
[6] J. Bao, Y. Zheng, D. Wilkie, and M. Mokbel. Recommendations in location-based social networks: a survey. *GeoInformatica*, 19(3):525–565, 2015.
[7] B. Barnes, S. Dunn, C. Pearson, and S. Wilkinson. Improving human behaviour in macroscale city evacuation agent-based simulation. *International Journal of Disaster Risk Reduction*, 60:102289, 2021.
[8] G. Box. All models are wrong, but some are useful. *Robustness in Statistics*, 202(1979):549, 1979.
[9] T. Brinkhoff. A Framework for Generating Network-Based Moving Objects. *GeoInformatica*, 6(2):153–180, 2002.
[10] K. E. Cevasco and A. A. Roess. Adaptation and utilization of a postmarket evaluation model for digital contact tracing mobile health tools in the united states: Observational cross-sectional study. *JMIR Public Health and Surveillance*, 9:e38633, 2023.
[11] X. Chen, Y. Zeng, G. Cong, S. Qin, Y. Xiang, and Y. Dai. On information coverage for location category based point-of-interest recommendation. In *AAAI*, pages 37–43, 2015.
[12] C. Cheng, H. Yang, I. King, and M. R. Lyu. Fused matrix factorization with geographical and social influence in location-based social networks. In *Aaai*, volume 12, pages 17–23, 2012.
[13] C. Cheng, H. Yang, M. R. Lyu, and I. King. Where you like to go next: Successive point-of-interest recommendation. In *IJCAI*, volume 13, pages 2605–2611, 2013.
[14] R. Cheng, J. Pang, and Y. Zhang. Inferring friendship from check-in data of location-based social networks. In *ACM International Conference on Advances in Social Networks Analysis and Mining 2015*, pages 1284–1291. ACM, 2015.
[15] Z. Cheng, J. Caverlee, K. Lee, and D. Z. Sui. Exploring millions of footprints in location sharing services. *ICWSM*, 2011:81–88, 2011.
[16] E. Cho, S. A. Myers, and J. Leskovec. Friendship and mobility: user movement in location-based social networks. In *KDD*, pages 1082–1090. ACM, 2011.
[17] 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*, 3:1376, 2013.
[18] Department of Interior In partnership with The Intelligence Advanced Research Projects Activity (IARPA). Broad Agency Announcement for: Hidden ActivitY Signal and Trajectory Anomaly Characterization (HAYSTAC) (https://sam.gov/opp/07a559beb9db44d2b7fb23b22ce5ead4/view), 2022.
[19] O. E. Dictionary. Oxford english dictionary. *Simpson, Ja & Weiner, Esc*, 3, 1989.
[20] T.-N. Doan and E.-P. Lim. Modeling location-based social network data with area attraction and neighborhood competition. *Data Mining and Knowledge Discovery*, 33(1):58–95, 2019.
[21] D. Driess, F. Xia, M. S. Sajjadi, C. Lynch, A. Chowdhery, B. Ichter, A. Wahid, J. Tompson, Q. Vuong, T. Yu, et al. Palm-e: An embodied multimodal language model. *arXiv preprint arXiv:2303.03378*, 2023.
[22] C. Düntgen, T. Behr, and R. H. Güting. Berlinmod: a benchmark for moving object databases. *The VLDB Journal*, 18:1335–1368, 2009.
[23] I. Environmental Systems Research Institute. ESRI Shapefile Technical Description. Technical report, Environmental Systems Research Institute, Inc, 1998.
[24] S. Feng, X. Li, Y. Zeng, G. Cong, Y. M. Chee, and Q. Yuan. Personalized ranking metric embedding for next new poi recommendation. In *IJCAI*, pages 2069–2075, 2015.
[25] G. Ference, M. Ye, and W.-C. Lee. Location recommendation for out-of-town users in location-based social networks. In *CIKM*, pages 721–726. ACM, 2013.

[26] H. Gao, J. Tang, X. Hu, and H. Liu. Exploring temporal effects for location recommendation on location-based social networks. In *Proc. 7th ACM Conf. Recommender systems*, pages 93–100. ACM, 2013.

[27] H. Gao, J. Tang, X. Hu, and H. Liu. Content-aware point of interest recommendation on location-based social networks. In *AAAI*, pages 1721–1727, 2015.

[28] George Mason University campus data . http://librarygeodata.gmu.edu.mutex.gmu.edu/campus.html, 2009.

[29] G. Ghinita, P. Kalnis, A. Khoshgozaran, C. Shahabi, and K.-L. Tan. Private queries in location based services: anonymizers are not necessary. In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, pages 121–132, 2008.

[30] Github. Github Copilot - The world's most widely adopted AI developer tool. (https://github.com/features/copilot), 2023.

[31] M. Graham, M. Kutzbach, and B. McKenzie. Design comparison of lodes and acs commuting data products. Technical report, US Census Bureau, Center for Economic Studies, 2014.

[32] G. J. Griffith, T. T. Morris, M. J. Tudball, A. Herbert, G. Mancano, L. Pike, G. C. Sharp, J. Sterne, T. M. Palmer, G. Davey Smith, et al. Collider bias undermines our understanding of covid-19 disease risk and severity. *Nature communications*, 11(1):5749, 2020.

[33] J. He, X. Li, L. Liao, D. Song, and W. K. Cheung. Inferring a personalized next point-of-interest recommendation model with latent behavior patterns. In *AAAI*, pages 137–143, 2016.

[34] H.-P. Hsieh, R. Yan, and C.-T. Li. Where you go reveals who you know: analyzing social ties from millions of footprints. In *CIKM*, pages 1839–1842. ACM, 2015.

[35] B. Hu and M. Ester. Spatial topic modeling in online social media for location recommendation. In *Proc. 7th ACM conference on Recommender systems*, pages 25–32. ACM, 2013.

[36] S. Isaj and T. B. Pedersen. Seed-driven geo-social data extraction. In *Proceedings of the 16th International Symposium on Spatial and Temporal Databases*, pages 11–20, 2019.

[37] JOSM: An extensible editor for OpenStreetMap . https://josm.openstreetmap.de/wiki/Download, 2020.

[38] Kaggle. Ecml/pkdd 15: Taxi trajectory prediction challenge. https://www.kaggle.com/c/pkdd-15-predict-taxi-service-trajectory-i/data, Accessed Mar 20, 2023.

[39] J. S. Kim, H. Jin, H. Kavak, O. C. Rouly, A. Crooks, D. Pfoser, C. Wenk, and A. Züfle. Location-based social network data generation based on patterns of life. In *2020 21st IEEE International Conference on Mobile Data Management (MDM)*, pages 158–167, 2020.

[40] J.-S. Kim, H. Kavak, and A. Crooks. Procedural city generation beyond game development. *SIGSPATIAL Special*, 10(2):34–41, 2018.

[41] J.-S. Kim, H. Kavak, U. Manzoor, A. Crooks, D. Pfoser, C. Wenk, and A. Züfle. Simulating urban patterns of life: A geo-social data generation framework. In *Proceedings of the 27th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pages 576–579, 2019.

[42] J.-S. Kim, H. Kavak, C. O. Rouly, H. Jin, A. Crooks, D. Pfoser, C. Wenk, and A. Züfle. Location-based social simulation for prescriptive analytics of disease spread. *SIGSPATIAL Special*, 12(1):53–61, 2020.

[43] W. Kohn, H. Amiri, and A. Züfle. Epipol: An epidemiological patterns of life simulation (demonstration paper). In *Proceedings of the 4th ACM SIGSPATIAL International Workshop on Spatial Computing for Epidemiology*, pages 13–16, 2023.

[44] D. Krajzewicz. Traffic simulation with sumo–simulation of urban mobility. *Fundamentals of traffic simulation*, pages 269–293, 2010.

[45] H.-P. Kriegel, P. Kunath, M. Pfeifle, and M. Renz. Viewnet: visual exploration of region-wide traffic networks. In *22nd International Conference on Data Engineering (ICDE'06)*, pages 166–166. IEEE, 2006.

[46] H.-P. Kriegel, M. Renz, M. Schubert, and A. Zuefle. Statistical density prediction in traffic networks. In *Proceedings of the 2008 SIAM International Conference on Data Mining*, pages 692–703. SIAM, 2008.

[47] J. Krumm. A survey of computational location privacy. *Personal and Ubiquitous Computing*, 13:391–399, 2009.

[48] J. J. Levandoski, M. Sarwat, A. Eldawy, and M. F. Mokbel. Lars: A location-aware recommender system. In *ICDE*, pages 450–461. IEEE, 2012.

[49] H. Li, Y. Ge, R. Hong, and H. Zhu. Point-of-interest recommendations: Learning potential check-ins from friends. In *KDD*, pages 975–984, 2016.

[50] H. Li, R. Hong, S. Zhu, and Y. Ge. Point-of-interest recommender systems: A separate-space perspective. In *ICDM*, pages 231–240. IEEE, 2015.

[51] M. Li, R. Westerholt, H. Fan, and A. Zipf. Assessing spatiotemporal predictability of lbsn: a case study of three foursquare datasets. *GeoInformatica*, pages 1–21, 2016.

[52] X. Li, G. Cong, X.-L. Li, T.-A. N. Pham, and S. Krishnaswamy. Rank-geofm: A ranking based geographical factorization method for point of interest recommendation. In *SIGIR*, pages 433–442. ACM, 2015.

[53] D. Lian, V. W. Zheng, and X. Xie. Collaborative filtering meets next check-in location prediction. In *WWW*, pages 231–232. ACM, 2013.

[54] G. Liao, S. Jiang, Z. Zhou, C. Wan, and X. Liu. Poi recommendation of location-based social networks using tensor factorization. In *MDM*, pages 116–124. IEEE, 2018.

[55] B. Liu, Y. Fu, Z. Yao, and H. Xiong. Learning geographical preferences for point-of-interest recommendation. In *ACM SIGKDD*, pages 1043–1051, 2013.

[56] B. Liu, H. Xiong, S. Papadimitriou, Y. Fu, and Z. Yao. A general geographical probabilistic factor model for point of interest recommendation. *IEEE Transactions on Knowledge and Data Engineering*, 27(5):1167–1179, 2015.

[57] Q. Liu, S. Wu, L. Wang, and T. Tan. Predicting the next location: A recurrent model with spatial and temporal contexts. In *AAAI*, pages 194–200, 2016.

[58] X. Liu, Y. Liu, K. Aberer, and C. Miao. Personalized point-of-interest recommendation by mining users' preference transition. In *CIKM*, pages 733–738. ACM, 2013.

[59] Y. Liu, C. Liu, B. Liu, M. Qu, and H. Xiong. Unified point-of-interest recommendation with temporal interval assessment. In *KDD*, pages 1015–1024, 2016.

[60] Y. Liu, T.-A. N. Pham, G. Cong, and Q. Yuan. An experimental evaluation of point-of-interest recommendation in location-based social networks. *Proc. VLDB Endowment*, 10(10):1010–1021, 2017.

[61] Y. Liu, W. Wei, A. Sun, and C. Miao. Exploiting geographical neighborhood characteristics for location recommendation. In *CIKM*, pages 739–748. ACM, 2014.

[62] A. H. Maslow. A theory of human motivation. *Psychological Review*, 50:370–396, 1943.

[63] Meta AI. Llama 2 (https://ai.meta.com/llama/), 2023.

[64] F. C. Mish. *Merriam-Webster's collegiate dictionary*, volume 1. Merriam-Webster, 2004.

[65] M. Mokbel, S. Abbar, and R. Stanojevic. Contact tracing: Beyond the apps. *SIGSPATIAL Special*, 12(2):15–24, 2020.

[66] M. F. Mokbel, L. Alarabi, J. Bao, A. Eldawy, A. Magdy, M. Sarwat, E. Waytas, and S. Yackel. Mntg: An extensible web-based traffic generator. In *Advances in Spatial and Temporal Databases: 13th International Symposium, SSTD 2013, Munich, Germany, August 21-23, 2013. Proceedings 13*, pages 38–55. Springer, 2013.

[67] G. S. Njoo, M.-C. Kao, K.-W. Hsu, and W.-C. Peng. Exploring check-in data to infer social ties in location based social networks. In *PAKDD*, pages 460–471. Springer, 2017.

[68] A. Noulas, S. Scellato, N. Lathia, and C. Mascolo. A random walk around the city: New venue recommendation in location-based social networks. In *PASSAT (socialcom)*, pages 144–153. Ieee, 2012.

[69] A. Noulas, S. Scellato, C. Mascolo, and M. Pontil. An empirical study of geographic user activity patterns in Foursquare. *ICWSM*, 11:70–573, 2011.

[70] A. Noulas, S. Scellato, C. Mascolo, and M. Pontil. Exploiting semantic annotations for clustering geographic areas and users in location-based social networks. *The social mobile web*, 11(2), 2011.

[71] OpenAI. ChatGPT can now see, hear, and speak (https://openai.com/blog/chatgpt-can-now-see-hear-and-speak), 2023.

[72] OpenAI. ChatGPT for AI (https://chatgpt4.ai), 2023.

[73] OpenAI. Code Interpreter - ChatGPT (https://chat.openai.com/?model=gpt-4-code-interpreter), 2023.

[74] OpenStreetMap. https://www.openstreetmap.org/, 2019.

[75] OpenStreetMap Foundation. OpenStreetMap API. http://wiki.openstreetmap.org/wiki/API.

[76] S. E. Page. Computational models from a to z. *Complexity*, 5(1):35–41, 1999.

[77] B. Palsson. The challenges of in silico biology. *Nature biotechnology*, 18(11):1147–1150, 2000.

[78] N. Pelekis, C. Ntrigkogias, P. Tampakis, S. Sideridis, and Y. Theodoridis. Hermoupolis: a trajectory generator for simulating generalized mobility patterns. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2013, Prague, Czech Republic, September 23-27, 2013, Proceedings, Part III 13*, pages 659–662. Springer, 2013.

[79] D. Pfoser and C. S. Jensen. Indexing of network constrained moving objects. In *GIS '03: Proceedings of the 11th ACM international symposium on Advances in geographic information systems*, pages 25–32. ACM, 2003.

[80] H. Pham, C. Shahabi, and Y. Liu. Ebm: an entropy-based model to infer social strength from spatiotemporal data. In *ACM SIGMOD*, pages 265–276. ACM, 2013.

[81] QGIS. Qgis: A free and open source geographic information system. https://qgis.org/en/site/forusers/download.html, 2019.

[82] SafeGraph Inc. Stopping COVID-19 with New Social Distancing Dataset (https://www.safegraph.com/blog/stopping-covid-19-with-new-social-distancing-dataset).

[83] SafeGraph Inc. Weekly Patterns Dataset (https://docs.safegraph.com/docs/weekly-patterns).

[84] M. A. Saleem, F. S. Da Costa, P. Dolog, P. Karras, T. B. Pedersen, and T. Calders. Predicting visitors using location-based social networks. In *MDM*, pages 245–250. IEEE, 2018.

[85] M. A. Saleem, X. Xie, and T. B. Pedersen. Scalable processing of location-based social networking queries. In *MDM*, volume 1, pages 132–141. IEEE, 2016.

[86] M. Sarwat, J. J. Levandoski, A. Eldawy, and M. F. Mokbel. Lars*: An efficient and scalable location-aware recommender system. *TKDE*, 26(6):1384–1399, 2014.

[87] S. Scellato, C. Mascolo, M. Musolesi, and V. Latora. Distance matters: Geo-social metrics for online social networks. In *WOSN*, 2010.

[88] S. Scellato, A. Noulas, R. Lambiotte, and C. Mascolo. Socio-spatial properties of online location-based social networks. *ICWSM*, 11:329–336, 2011.

[89] S. Scellato, A. Noulas, and C. Mascolo. Exploiting place features in link prediction on location-based social networks. In *ACM SIGKDD*, pages 1046–1054, 2011.

[90] E. Seglem, A. Züfle, J. Stutzki, F. Borutta, E. Faerman, and M. Schubert. On privacy in spatio-temporal data: User identification using microblog data. In *Advances in Spatial and Temporal Databases: 15th International Symposium, SSTD 2017, Arlington, VA, USA, August 21–23, 2017, Proceedings 15*, pages 43–61. Springer, 2017.

[91] D. Shuttleworth and J. Padilla. From narratives to conceptual models via natural language processing. In *2022 Winter Simulation Conference (WSC)*, pages 2222–2233. IEEE, 2022.

[92] D. Shuttleworth and J. J. Padilla. Towards semi-automatic model specification. In *2021 Winter Simulation Conference (WSC)*, pages 1–12. IEEE, 2021.

[93] H. B. Sieburg. Physiological studies in silico. In *1990 Lectures in Complex Systems*, pages 367–390. CRC Press, 1990.

[94] Z. Sinuany-Stern and E. Stern. Simulating the evacuation of a small city: the effects of traffic factors. *Socio-Economic Planning Sciences*, 27(2):97–108, 1993.

[95] Stanford network analysis project. https://snap.stanford.edu/index.html.

[96] Y. Su, X. Li, W. Tang, J. Xiang, and Y. He. Next check-in location prediction via footprints and friendship on location-based social networks. In *MDM*, pages 251–256. IEEE, 2018.

[97] Y. Vigfusson, T. A. Karlsson, D. Onken, C. Song, A. F. Einarsson, N. Kishore, R. M. Mitchell, E. Brooks-Pollock, G. Sigmundsdottir, and L. Danon. Cell-phone traces reveal infection-associated behavioral change. *Proceedings of the National Academy of Sciences*, 118(6):e2005241118, 2021.

[98] K. W Axhausen, A. Horni, and K. Nagel. *The multi-agent transport simulation MATSim*. Ubiquity Press, 2016.

[99] H. Wang, Z. Li, and W.-C. Lee. PGT: Measuring mobility relationship using personal, global and temporal factors. In *ICDM*, pages 570–579. IEEE, 2014.

[100] H. Wang, M. Terrovitis, and N. Mamoulis. Location recommendation in location-based social networks using user check-in data. In *ACM SIGSPATIAL*, pages 374–383, 2013.

[101] W. Wang, H. Yin, L. Chen, Y. Sun, S. Sadiq, and X. Zhou. Geo-sage: A geographical sparse additive generative model for spatial item recommendation. In *ACM SIGKDD*, pages 1255–1264, 2015.

[102] X. Wang, L. Wang, and P. Yang. Prevalent co-visiting patterns mining from location-based social networks. In *MDM*, pages 581–586. IEEE, 2019.

[103] Z. Wang, D. Zhang, X. Zhou, D. Yang, Z. Yu, and Z. Yu. Discovering and profiling overlapping communities in location-based social networks. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 44(4):499–509, 2014.

[104] Y.-T. Wen, Y. Y. Fan, and W.-C. Peng. Mining of location-based social networks for spatio-temporal social influence. In *PAKDD*, pages 799–810. Springer, 2017.

[105] M. Wernke, P. Skvortsov, F. Dürr, and K. Rothermel. A classification of location privacy attacks and approaches. *Personal and ubiquitous computing*, 18:163–175, 2014.

[106] C. Wu, S. Yin, W. Qi, X. Wang, Z. Tang, and N. Duan. Visual chatgpt: Talking, drawing and editing with visual foundation models. *arXiv preprint arXiv:2303.04671*, 2023.

[107] Y. Xiao and L. Xiong. Protecting locations with differential privacy under temporal correlations. In *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*, pages 1298–1309, 2015.

[108] G. Xu-Rui, W. Li, and W. Wei-Li. Using multi-features to recommend friends on location-based social networks. *Peer-to-Peer Networking and Applications*, pages 1–8, 2016.

[109] D. Yang, B. Qu, J. Yang, and P. Cudre-Mauroux. Revisiting user mobility and social relationships in lbsns: A hypergraph embedding approach. In *Proc. Web Conference*, WWW'19, 2019.

[110] D. Yang, D. Zhang, L. Chen, and B. Qu. Nationtelescope: Monitoring and visualizing large-scale collective behavior in LBSNs. *Journal of Network and Computer Applications*, 55:170–180, 2015.

[111] D. Yang, D. Zhang, V. W. Zheng, and Z. Yu. Modeling user activity preference by leveraging user spatial temporal characteristics in LBSNs. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 45(1):129–142, 2015.

[112] G. Yang and A. Züfle. Spatio-temporal site recommendation. In *Data Mining Workshops (ICDMW)*, pages 1173–1178. IEEE, 2016.

[113] G. Yang and A. Züfle. Spatio-temporal prediction of social connections. In *ACM Workshop on Managing and Mining Enriched Geo-Spatial Data (GeoRich)*. ACM, 2017.

[114] J. Ye, Z. Zhu, and H. Cheng. What's your next move: User activity prediction in location-based social networks. In *SIAM Data Mining (SDM)*, pages 171–179. SIAM, 2013.

[115] M. Ye, P. Yin, and W.-C. Lee. Location recommendation for location-based social networks. In *ACM SIGSPATIAL*, pages 458–461, 2010.

[116] M. Ye, P. Yin, W.-C. Lee, and D.-L. Lee. Exploiting geographical influence for collaborative point-of-interest recommendation. In *SIGIR*, pages 325–334. ACM, 2011.

[117] Yelp dataset challenge. round 9. https://www.yelp.com/dataset_challenge. Accessed: 2017-07-30.

[118] H. Yin, Z. Hu, X. Zhou, H. Wang, K. Zheng, Q. V. H. Nguyen, and S. Sadiq. Discovering interpretable geo-social communities for user behavior prediction. In *ICDE*, pages 942–953. IEEE, 2016.

[119] H. Yin, Y. Sun, B. Cui, Z. Hu, and L. Chen. Lcars: a location-content-aware recommender system. In *ACM SIGKDD*, pages 221–229, 2013.

[120] H. Yin, X. Zhou, Y. Shao, H. Wang, and S. Sadiq. Joint modeling of user check-in behaviors for point-of-interest recommendation. In *CIKM*, pages 1631–1640. ACM, 2015.

[121] F. Yu, Z. Li, S. Jiang, and S. Lin. Point-of-interest recommendation for location promotion in location-based social networks. In *MDM*, pages 344–347. IEEE, 2017.

[122] J. Yuan, Y. Zheng, X. Xie, and G. Sun. Driving with knowledge from the physical world. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 316–324, 2011.

[123] J. Yuan, Y. Zheng, C. Zhang, W. Xie, X. Xie, G. Sun, and Y. Huang. T-drive: driving directions based on taxi trajectories. In *ACM SIGSPATIAL*, pages 99–108, 2010.

[124] Q. Yuan, G. Cong, Z. Ma, A. Sun, and N. M. Thalmann. Time-aware point-of-interest recommendation. In *SIGIR*, pages 363–372. ACM, 2013.

[125] Q. Yuan, G. Cong, and A. Sun. Graph-based point-of-interest recommendation with geographical and temporal influences. In *CIKM*, pages 659–668. ACM, 2014.

[126] S. Zeighami, C. Shahabi, and J. Krumm. Estimating spread of contact-based contagions in a population through sub-sampling. *Proceedings of the VLDB Endowment*, 14(9):1557–1569, 2021.

[127] J.-D. Zhang and C.-Y. Chow. igslr: personalized geo-social location recommendation: a kernel density estimation approach. In *ACM SIGSPATIAL*, pages 334–343, 2013.

[128] J.-D. Zhang and C.-Y. Chow. Geosoca: Exploiting geographical, social and categorical correlations for point-of-interest recommendations. In *SIGIR*, pages 443–452. ACM, 2015.

[129] J.-D. Zhang and C.-Y. Chow. Spatiotemporal sequential influence modeling for location recommendations: A gravity-based approach. *TIST*, 7(1):11, 2015.

[130] J.-D. Zhang, C.-Y. Chow, and Y. Li. Lore: Exploiting sequential influence for location recommendations. In *ACM SIGSPATIAL*, pages 103–112, 2014.

[131] J.-D. Zhang, C.-Y. Chow, and Y. Zheng. Orec: An opinion-based point-of-interest recommendation framework. In *CIKM*, pages 1641–1650. ACM, 2015.

[132] L. Zhang, L. Zhao, and D. Pfoser. Factorized deep generative models for end-to-end trajectory generation with spatiotemporal validity constraints. In *Proc. 30th Int'l Conference on Advances in Geographic Information Systems (ACM SIGSPATIAL)*, pages 1–12, 2022.

[133] K. Zhao, G. Cong, Q. Yuan, and K. Q. Zhu. Sar: A sentiment-aspect-region model for user preference analysis in geo-tagged reviews. In *ICDE*, pages 675–686. IEEE, 2015.

[134] S. Zhao, T. Zhao, H. Yang, M. R. Lyu, and I. King. Stellar: Spatial-temporal latent ranking for successive point-of-interest recommendation. In *AAAI*, pages 315–322, 2016.

[135] Y. Zheng, Y. Chen, X. Xie, and W.-Y. Ma. Geolife 2.0: a location-based social networking service. In *mobile data management: systems, services and middleware*, pages 357–358. IEEE, 2009.

[136] Y. Zheng, X. Xie, and W.-Y. Ma. Geolife: A collaborative social networking service among user, location and trajectory. *IEEE Data Eng. Bull.*, 33(2):32–39, 2010.

[137] Y. Zheng, L. Zhang, X. Xie, and W.-Y. Ma. Mining interesting locations and travel sequences from gps trajectories. In *Proceedings of the 18th international conference on World wide web*, pages 791–800, 2009.

[138] W.-X. Zhou, D. Sornette, R. A. Hill, and R. I. Dunbar. Discrete hierarchical organization of social group sizes. *Proceedings of the Royal Society B: Biological Sciences*, 272(1561):439–444, 2005.

[139] A. Züfle, C. Wenk, D. Pfoser, A. Crooks, J.-S. Kim, H. Kavak, U. Manzoor, and H. Jin. Urban life: a model of people and places. *Computational and Mathematical Organization Theory*, pages 1–32, 2021.