Simulating Urban Patterns of Life: A Geo-Social Data Generation Framework

Joon-Seok Kim
George Mason University
jkim258@gmu.edu

Hamdi Kavak
George Mason University
hkavak@gmu.edu

Umar Manzoor
University of Hull
u.manzoor@hull.ac.uk

Andrew Crooks
George Mason University
acrooks2@gmu.edu

Dieter Pfoser
George Mason University
dpfoser@gmu.edu

Carola Wenk
Tulane University
cwenk@tulane.edu

Andreas Züfle
George Mason University
azufle@gmu.edu

ABSTRACT

Data generators have been heavily used in creating massive trajectory datasets to address common challenges of real-world datasets, including privacy, cost of data collection, and data quality. However, such generators often overlook social and physiological characteristics of individuals and as such their results are often limited to simple movement patterns. To address these shortcomings, we propose an agent-based simulation framework that facilitates the development of behavioral models in which agents correspond to individuals that act based on personal preferences, goals, and needs within a realistic geographical environment. Researchers can use a drag-and-drop interface to design and control their own world including the geospatial and social (i.e. geo-social) properties. The framework is capable of generating and streaming very large data that captures the basic patterns of life in urban areas. Streaming data from the simulation can be accessed in real time through a dedicated API.

CCS CONCEPTS

• **Computing methodologies → Agent / discrete models**; **Simulation tools**.

KEYWORDS

Agent-based simulation, trajectory data, data generator, spatial network, human behavior

ACM Reference Format:
Joon-Seok Kim, Hamdi Kavak, Umar Manzoor, Andrew Crooks, Dieter Pfoser, Carola Wenk, and Andreas Züfle. 2019. Simulating Urban Patterns of Life: A Geo-Social Data Generation Framework. In *SIGSPATIAL'19: ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, November 05–08, 2019, Chicago, IL.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3347146.3359106

SIGSPATIAL'19, November 2019, Chicago, Illinois, USA

ACM ISBN 978-1-4503-9999-9/18/06.
https://doi.org/10.1145/3347146.3359106

1 INTRODUCTION

Trajectory data capturing human mobility is valuable in many applications, including the inference of popular routes [23], traffic prediction [12], and ride-sharing applications [26]. However, having a large set of real-world trajectory data raises serious concerns regarding the spatial privacy of users. For example, Apple's privacy policy allows sharing the spatio-temporal location of their users with "partners and licensees" [1]. Studies on location privacy based on trajectories have already shown that such data allows inferring the identity of individuals based on their locations using taxi trajectories [24], Twitter check-ins [18], and call detail records [7].

In addition to the privacy challenge, the cost of data collection and data quality hinders the broad research community in exploiting real-world trajectory data sets. In particular, publicly available real-world trajectory data sets are highly limited in quality and quantity [10]. For example, existing vehicle trajectories may only report GPS samples in a minute-scale frequency [25] and microblogging data (e.g., Twitter) may have users share their location less than once per day. This data sparsity makes high-fidelity mobility analysis more than challenging [21]. Another type of mobility data used recently by the GIS community is Origin-Destination data, such as provided by the New York City Taxi and Limousine Commission [3]. But, the lack of actual traces, as well as the inability to capture multiple trips by the same user, make this type of data inapplicable for many applications (e.g., Why does a person go from A to B? Why was one route chosen over another etc.). Moreover, non-spatial properties of individuals such as age, occupation, and preference, which are critical to understand causality between mobility and behavior, are omitted for the sake of privacy protection.

While there are numerous limitations with respect to publicly available data sets, imagine the research possibilities with trajectory data that have none of these shortcomings. Having a large data set with high-quality mobility traces, the users' home and work locations, as well as the users travel motivation, will allow one to address many research questions in traffic prediction, urban planing, and location privacy. Along with the need for enriched mobility data, our work has been motivated by the following observations.

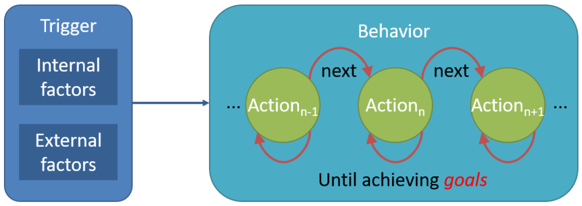


Figure 1: Causality in human behavior

- *Trajectory data generators* have been proposed in the past [5, 17, 19, 20] as an alternative method to create trajectories conveniently and inexpensively. These data generators, however, often overlook non-spatial properties of individuals. They tend to focus on mimicking the mobility phenomena without deep understanding of the intrinsic motivations of why people move. For example, existing data generators specify origin and destination regions, and simulate moving objects (or agents) between these regions on shortest paths. This oversimplification of human behavior leads to generating data that is inexplicable and implausible.
- *Agent-based simulation* is a flexible paradigm commonly applied in computational social science research to study complex social and spatial phenomena, such as the outbreak of riots, the disease spread, and or slum emergence in urban environments (see [6] for a review). Such simulations consider the motivations that drive individual agents based on socially plausible rules and theory. In addition, analysis can be carried out at an aggregate level, such that there is no need for large-scale data generation [6]. While, urban environments and places significantly influence citizens [14], such environments including buildings and spatial networks are often simplified or abstracted in agent-based simulations [6].

These observations drive us to identify the requirements for a geo-social data generator: flexibility, scalability, and a spatially explicit realistic representation. First and foremost, the generator should be equipped with a set of atomic building blocks with the *flexibility* to develop various behaviors of agents. Thus, the generator needs to easily facilitate changes to the model logic. Second, the simulator should be *scalable* to generate and stream very large geo-social datasets to support data-driven research. Instead of downloading data, users should be able to simply download the generator and run their model. Finally, the simulation should be *spatially aware*, allowing simulated agents to choose and travel to different locations depending on their characteristics and motivation.

In the remainder of this paper, we present our flexible, scalable, and spatially aware agent-based simulation framework to generate and stream large geo-social data. Specifically, in Section 2 we present the design of our framework which allows researchers to develop simulations based on user-specified human behavior concepts. Section 3 describes features and implementation of the demonstrator to be showcased at ACM SIGSPATIAL 2019, and Section 4 concludes the paper and discusses future research directions.

2 CONCEPTS AND ARCHITECTURE

Most human actions are triggered by a cause and have consequences/effects [6]. Ranging from simple repetitive behaviors like eating to a more complex behavior such as learning how to use a tool, this simple fact allows us to capture the patterns of daily human life (e.g., commuting to and from work, eating, sleeping, etc.) in a reflexive way. This forms the basis of our human behavior model along with human motivation theories [4, 8, 15, 16]. In simple terms, we define four concepts to capture the necessary factors and effectors of human behavior in our framework. Our framework allows us to define and combine each of these four concepts to create unique simulation and data generators.

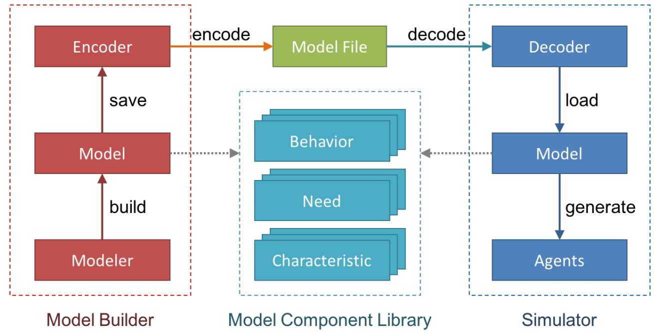


Figure 2: Architecture of framework

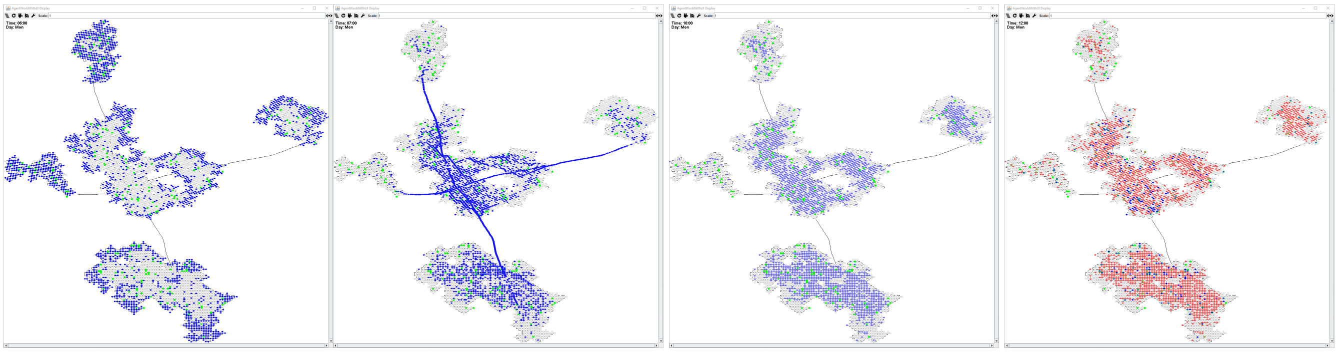
- *Trigger* is a mechanism that is initiated by a change or event in internal or external factors. Internal factors include one's needs, beliefs, and characteristics, while external factors include the environment or other individuals. For example, a trigger called hunger can be defined as an agent having a food level less than a specified threshold.
- *Behavior* is a construct that is directly initiated by a trigger. For instance, we assume that eating is a behavior because it is directly initiated by the trigger of hunger. Inside a behavior, one can define multiple actions that make up the lifetime of the behavior.
- *Action* is the process of performing certain step(s) that directly produce an output once it reaches the defined goal. Sticking with the eating behavior example, each process leading towards becoming full is considered an action. For instance, deciding whether eating at home or outside is a one-step action that generates a decision. Similarly, the process of grocery shopping is an action with multiple steps.
- *Goal* is simply the end condition for an action, such as an agent reaching a food level of 100%.

Fig. 1 depicts how these concepts relate to each other. A trigger initiates a behavior, which is composed of a set of actions. Each action could provide an output that feeds into the next action and involves one or multiple steps that are executed until an explicitly defined goal is met.

An agent may have multiple behaviors that are executed in order. One challenge is to determine the priority of competing behaviors. A hypothesis based on Maslow's hierarchy of needs [15] is one approach to this challenge. Our framework allows users to identify the priorities of behaviors so that they can customize their models to their own research needs.

We capture the above concepts in an extensible architecture (cf. Fig. 2) consisting of three parts: the Model Builder, the Model Component Library, and the Simulator.

- The *Model Component Library* is a library that contains essential model components (i.e., function blocks) to define agent-based



(a) Morning (6:00am) **(b) Rush hour (7:00am)** **(c) Working (8:30am)** **(d) Going for lunch (12:00pm)**

Figure 3: Several screenshots from the simulator

models including behaviors, actions, needs, and characteristics. It provides a common library for both the *Model Builder* and the *Simulator* to import components.

- The *Model Builder* enables users to focus on the logic of agent-based models and eases programming efforts for modeling. Basically, all components a model developer builds shall be in the library. The *Encoder* encodes models into files so that they can be exchanged between the *Model Builder* and the *Simulator*.
- The *Simulator* loads all model components and configurations from model files through the *Decoder*. The *Simulator* generates agents and runs simulation with them. Data generated by agents (such as check-in data at locations, transactions such as purchasing food, and meetings with other agents) is written on an output device, which can be a file or a stream.

3 DEMONSTRATION

In our demonstration, we will show three aspects of our data generation framework: (1) the Model Builder, which allows users to customize the logic for a model, (2) the Simulation and Data Generator to visualize the simulation using a graphical user interface, and (3) the Streaming Interface allowing users to connect, subscribe, and obtain a stream of generated data.

Model Builder: Fig. 4 illustrates the model builder with our sample model which will be demonstrated. It is implemented in Java based on JGraphX[2], a Java Swing diagramming library. Users can simply drag-and-drop components from the template panel to the drawing canvas. Our sample model consists of two *triggers* (*food*, *sleep*), eight properties (*money*, *wage*, *workHour*, *buildings*, *homeLocation*, *workLocation*, *destination*, *currentLocation*), and four *behaviors* (*eating*, *sleeping*, *working*, *doNothing*). Fig. 4 demonstrates how to edit the *HUNGRY* trigger in the *eating* behavior. In this example, an agent feels hungry when the strength of need for *food* is greater than 50, and looks for restaurants and determines one of them as a destination. Then, the agent performs an *action* of *traveling* to the restaurant along the shortest path in the road network. Afterwards, the agent eats food by fulfilling the need for *food* until achieving the *goal untilFULL*. After the meal, he/she will pay his/her bill using the *money* property. Modelers can save their working files and export model files that can load in our simulator.

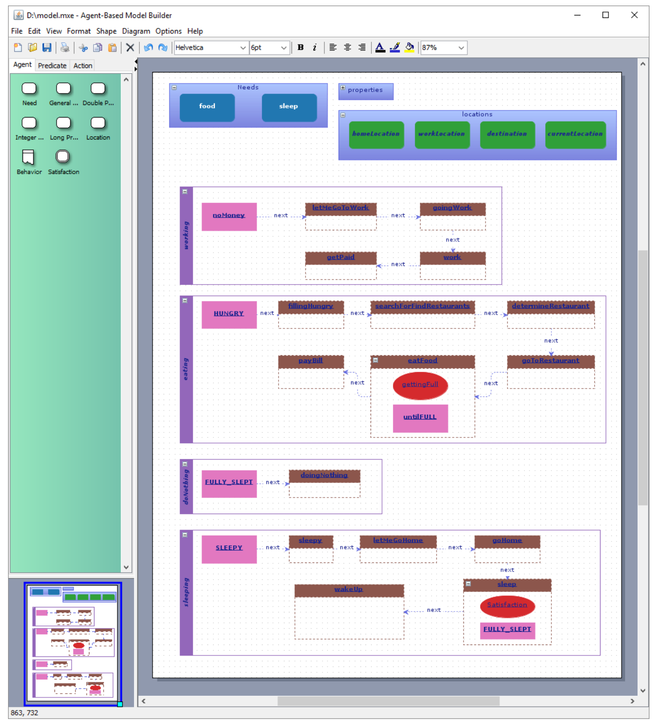


Figure 4: Layout of model builder and sample model

Simulation and Data Generation: Our framework leverages the open source simulation framework MASON [13] and its geospatial extension GeoMASON [22]. Fig. 3 shows screenshots of simulations with 10,000 agents, reflecting highly stylized daily patterns of life. The spatial environment uses synthetic data (14,181 road segments) created by a spatial network and place generator [11]. We captured several screenshots from the simulator representing different times of the day. In the simulation, we visualize hunger levels of agents with color: red means hungry while blue denotes full. In Fig. 3a, around 6 a.m., agents are still at home, distributed in a suburb of the city. Agents go to work in the morning and experience traffic congestion as shown in Fig. 3b. Fig. 3c implicitly depicts the spatial distribution of the workplace. We can notice that restaurants are evenly distributed compared to workplaces (see Fig.

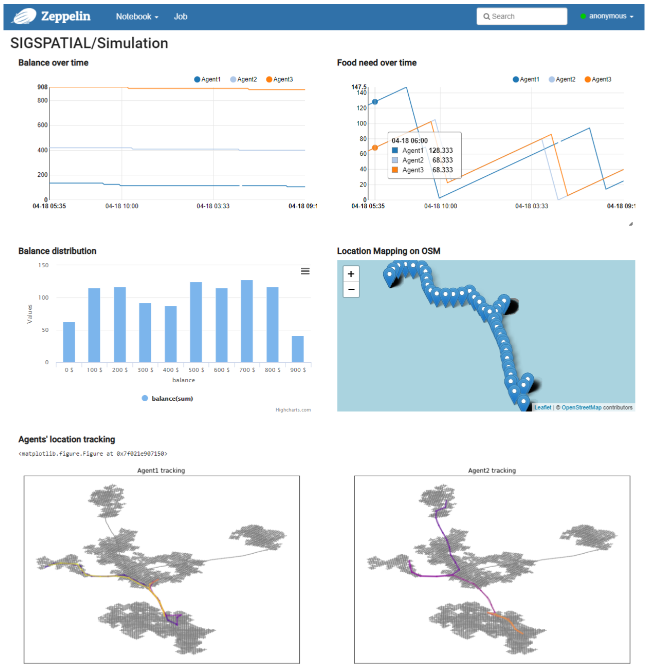


Figure 5: Analysis using Zeppelin + Spark Streaming

3d). While running the simulation, an expected daily routine and the variation of states of all agents over time are observed. However, such results only provide a visual validation. Empirical validation such as trace validation [9] can be conducted as well.

Streaming Interface: Our running simulation streams geo-social data of agents such as agent ID, time stamp (both in simulation time and ticks), current location, and current status of the agent. Our framework streams data using the ZeroMQ client library. We demonstrate a use case of streaming data analysis on our own stream on a Zeppelin instance using Spark Streaming as shown in Fig. 5. Supplementary materials including a demo video and sample data can be found at http://sigspatial19demo.joonseok.org.

4 CONCLUSION

We present an agent-based simulation framework for geo-social data generation. Our framework allows users to easily develop and customize the logic of agent behaviors for different applications, run the simulation to generate geo-social data, and explore the generated datasets. During the demonstration users will be able to test the *Model Builder*, as shown in Figure 4and to see how changes of agent logic are reflected in the simulation visualization. Attendees can also explore their own generated data. We hope that our simulation framework will be used to generate large sets of synthetic data for experimentation in many domains and that simulation settings will be shared to regenerate identical data sets for reproducible experimentation.

Our future work plans are three-fold. (1) We will enrich the atomic elements of the model builder to allow for expressing more complex behaviors, e.g., blocks that are related to agent interactions and social network will be added. (2) We will employ distributed computing systems to further the scalability of our framework. Such an addition is planned to be abstracted from model design so the framework users should not be affected by it. (3) Once the model builder is enriched, we will develop a patterns of life model that produces plausible behaviors of mobility. We will share the generated datasets with the scientific community.

ACKNOWLEDGMENT

This work was supported by the Defense Advanced Research Projects Agency (DARPA) under cooperative agreement No.HR00111820005 and the National Science Foundation Grants CCF-1637541 and CCF-1637576. The content of the information does not necessarily reflect the position or the policy of the Government, and no official endorsement should be inferred.

REFERENCES

[1] Apple privacy policy. https://www.apple.com/legal/privacy/, Accessed Mar 15, 2019.
[2] JGraphX. https://github.com/jgraph/jgraphx, Accessed Mar 15, 2019.
[3] New york city taxi and limousine commission, trip record data. http://www.nyc.gov/html/tlc/html/about/trip_record_data.shtml, Accessed Mar 15, 2019.
[4] C. P. Alderfer. An empirical test of a new theory of human needs. *Organizational behavior and human performance*, 4(2):142–175, 1969.
[5] T. Brinkhoff. A Framework for Generating Network-Based Moving Objects. *GeoInformatica*, 6(2):153–180, 2002.
[6] A. Crooks, N. Malleson, E. Manley, and A. Heppenstall. *Agent-Based Modelling and Geographical Information Systems: A Practical Primer*. SAGE Publications Limited, 2018.
[7] 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.
[8] E. Deci and R. M. Ryan. *Intrinsic motivation and self-determination in human behavior*. Springer Science & Business Media, 1985.
[9] R. J. Gore, C. J. Lynch, and H. Kavak. Applying statistical debugging for enhanced trace validation of agent-based models. *SIMULATION*, 93(4):273–284, 2017.
[10] H. Kavak, J.-S. Kim, A. Crooks, D. Pfoser, C. Wenk, and A. Züfle. Location-based social simulation. In *SSTD*, pages 218–221. ACM, 2019.
[11] J.-S. Kim, H. Kavak, and A. Crooks. Procedural city generation beyond game development. *SIGSPATIAL Special*, 10(2):34–41, 2018.
[12] W. Liu, Y. Zheng, S. Chawla, J. Yuan, and X. Xing. Discovering spatio-temporal causal interactions in traffic data streams. In *ACM SIGKDD*, pages 1010–1018. ACM, 2011.
[13] S. Luke, C. Cioffi-Revilla, L. Panait, K. Sullivan, and G. Balan. Mason: A multiagent simulation environment. *Simulation*, 81(7):517–527, 2005.
[14] K. Lynch. *The image of the city*, volume 11. MIT press, 1960.
[15] A. H. Maslow. A theory of human motivation. *Psychological review*, 50(4):370, 1943.
[16] D. C. McClelland, J. W. Atkinson, R. A. Clark, and E. L. Lowell. The achievement motive. 1953.
[17] 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 *SSTD*, pages 38–55. Springer, 2013.
[18] J. D. Park, E. Seglem, E. Lin, and A. Züfle. Protecting user privacy: Obfuscating discriminative spatio-temporal footprints. In *ACM SIGSPATIAL LocalRec Workshop*, article 2. ACM, 2017.
[19] D. Pfoser and Y. Theodoridis. Generating semantics-based trajectories of moving objects. *Computers, Environment and Urban Systems*, 27(3):243–263, 2003.
[20] K. Ramamohanarao, H. Xie, L. Kulik, S. Karunasekera, E. Tanin, R. Zhang, and E. B. Khunayn. Smarts: Scalable microscopic adaptive road traffic simulator. *ACM TIST*, 8(2):26, 2017.
[21] A. Sadilek, H. A. Kautz, and V. Silenzio. Predicting disease transmission from geo-tagged micro-blog data. In *AAAI*, pages 136–142, 2012.
[22] K. Sullivan, M. Coletti, and S. Luke. Geomason: Geospatial support for mason. Technical report, George Mason University, 2010.
[23] L.-Y. Wei, Y. Zheng, and W.-C. Peng. Constructing popular routes from uncertain trajectories. In *ACM SIGKDD*, pages 195–203. ACM, 2012.
[24] A. Y. Xue, R. Zhang, Y. Zheng, X. Xie, J. Huang, and Z. Xu. Destination prediction by sub-trajectory synthesis and privacy protection against such prediction. In *ICDE*, pages 254–265. IEEE, 2013.
[25] 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. ACM, 2010.
[26] J. Yuan, Y. Zheng, L. Zhang, X. Xie, and G. Sun. Where to find my next passenger. In *Ubiquitous Computing*, pages 109–118. ACM, 2011.