Location-Based Social Simulation

Hamdi Kavak
George Mason University
hkavak@gmu.edu

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

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

Location-based social networks (LBSNs) have been studied extensively in recent years. However, utilizing real-world LBSN datasets in such studies has severe weaknesses: sparse and small datasets, privacy concerns, and a lack of authoritative ground-truth. Our vision is to create a large scale geo-simulation framework to simulate human behavior and to create synthetic but realistic LBSN data that captures the location of users over time as well as social interactions of users in a social network. While existing LBSN datasets are trivially small, such a framework would provide the first source of massive LBSN benchmark data which would closely mimic the real world, containing high-fidelity information of location, and social connections of millions of simulated agents over several years of simulated time. Therefore, it would serve the research community by revitalizing and reshaping research on LBSNs by allowing researchers to see the (simulated) world through the lens of an omniscient entity having perfect data. These evaluations will guide future research enabling us to develop solutions to improve LBSN applications such as user-location recommendation, friend recommendation, location prediction, and location privacy.

CCS CONCEPTS

• **Human-centered computing → Social networks**; • **Computing methodologies → Agent / discrete models**; • **Information systems → Location based services**.

KEYWORDS

Agent-based simulation, location-based social network, data generator, spatial network, human behavior

ACM Reference Format:
Hamdi Kavak, Joon-Seok Kim, Andrew Crooks, Dieter Pfoser, Carola Wenk, and Andreas Züfle. 2019. Location-Based Social Simulation. In *16th International Symposium on Spatial and Temporal Databases (SSTD '19), August 19–21, 2019, Vienna, Austria.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3340964.3340995

SSTD '19, August 19–21, 2019, Vienna, Austria

ACM ISBN 978-1-4503-6280-1/19/08…$15.00
https://doi.org/10.1145/3340964.3340995

1 INTRODUCTION

A social network is a social structure consisting of individual users connected by a social relationship such as friendship. Social networking services build on real-world social networks through online platforms, providing ways for users to share ideas, activities, events, and interests. For example, users can: share location-tagged images with their friends (e.g., in Flickr), rate restaurants and bars and recommend them to their friends (e.g., in Foursquare), or log jogging and bicycle trails for sports analysis and experience sharing (e.g., in Bikely). This dimension of location bridges the gap between the physical world and online social networking services. As the location is one of the most important components of user context, extensive knowledge about an individual's interests, behaviors, and relationships with others can be learned from their locations. These kinds of location-embedded and location-driven social structures are known as location-based social networks (LBSNs).

Publicly available real-world datasets have been the driving force for LBSN research in recent years, but such datasets exhibit certain weaknesses:

- **Data sparsity:** LBSN data exhibits an extreme long-tail distribution of user behavior. In all existing datasets, the vast majority of users has less than ten check-ins [12]. Besides, the number of locations visited by a user is usually only a small portion of all locations. The density of the data used in experimental studies on LBSNs is usually around 0.1% [12].
- **Small datasets:** Existing datasets used to train models are small, as detailed in Section 3. They cover only a short period, a small number of users, or a small number of check-ins. Thus, model overfitting becomes a severe concern.
- **Privacy Concerns:** Most LBSN data was published by users and consented for public use. However, some users may revoke this consent, for instance, by deleting their LBSN account. Such changes will not be reflected in existing LBSN datasets, creating severe privacy concerns.
- **No ground-truth:** There is no way to assess whether a user visited a location or if the social network is correct and complete. Thus, it is difficult to assess the accuracy and robustness of existing experimental results using LBSN data.

The vision described in this work is to employ geospatial simulation to create artificial, but socially plausible LBSN datasets. Such large and dense datasets would allow the broad research community to test research hypotheses without any privacy concern. Therefore, the synthetic datasets would enable us to investigate what would be possible if we had such high-fidelity LBSN data of the real world.

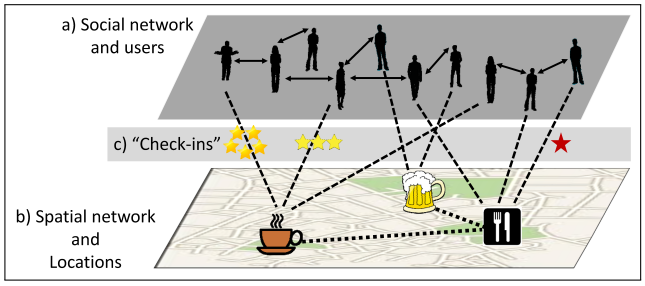

Figure 1: LBSN Overview.

2 LOCATION-BASED SOCIAL NETWORKS: AN OVERVIEW

Users [26] and locations [27] are the two major subjects that interact with each other in an LBSN. We can observe three types of networks that constitute an LBSN, (i) a user-user social network, (ii) a location-location spatial network, and (iii) a user-location bipartite network. Figure 1 gives a schematic overview of these networks and their interaction. Like in a traditional social network, users are connected via relationships such as friend, family, or co-worker. A spatial network defines proximity in terms of path distance, but may also introduce connections between locations that have similar semantic properties, e.g., of the same location type. Finally, the core feature of a location-based social network is the user-location network, which bridges users and locations (Figure 1). This bipartite network between users and locations captures events of users visiting a physical location. Such so-called "check-ins" may be enriched with qualitative information, such as user recommendations, which may be explicit, e.g., on a scale from one to five stars, or implicit, e.g., by observing that a user frequently checks-in at the same location.

LBSN data can be used in a plethora of beneficial applications. Initial LBSN work focused on modeling and describing **Human Mobility Patterns** [23], and explaining why users choose locations and how social ties affect this choice. Then **Location Recommendation** tries to predict edges of the user-location check-in network [24]. Closely related, **Check-in Prediction** [16] tries to predict future check-ins, which is useful in marketing applications. Another research field is LBSN-based **Friend Recommendation** or **Social Link Discovery** [21], which suggest new friends to users that have a similar interest at similar locations, while also having similar social connections. Other research topics include finding communities [25] and efficient query processing [10] in LBSNs.
To sum up, there has been a plethora of diverse LBSN research. However, the impact of LBSN research relies on the quality of data, and the next section will show that there is a considerable shortage of rich datasets. The datasets used in experiments are small, lack sufficient sample size for individual users, and cannot provide authoritative ground-truth knowledge for a meaningful evaluation of all these methods.

3 EXISTING LBSN DATASETS

Meaningful real-world LBSN datasets are a scarce resource considering the privacy implications of making such data public. Also, service providers consider such datasets invaluable when it comes to providing a competitive product and are thus somewhat unwilling to provide researchers even with sizable datasets.

Dataset	#Users	#Locations	#CheckIns	#Links
Gowalla	319K	2.8M	36M	4.4M
BrightKite	58K	971K	4.49M	214k
Foursquare	2.7M	11.1M	90M	0
Yelp	1.00M	144K	4.10M	0

Table 1: Publicly Available Real-World LBSN Datasets

Table 1 summarizes publicly available datasets that are intensively used by the LBSN research community.
Gowalla: Collected by the authors of [11] and retrieved from the LBSN Gowalla, which launched in 2007 and closed in 2012. This dataset has the largest social network of any public LBSN dataset while the majority of 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 [11]. A similar dataset is *Brightkite*, which is available at SNAP [1]. It is smaller than the Gowalla data in every aspect.
Foursquare: In terms of the number of users and check-ins, the largest publicly available LBSN dataset was collected from Foursquare [22]. However, no social network data is available. The Foursquare dataset suffers from the same user-inactivity problem. A recent study [9] has shown that the lower bound of predictability of the human spatiotemporal behavior (defined in [9]) is as low as 27%. They conclude that "Researchers working with LBSN datasets are often confronted by themselves or others with doubts regarding the quality or the potential of their datasets." and that "it is reasonable to be skeptical, indeed".
Yelp: A large dataset is published by Yelp as part of the Yelp Dataset Challenge [2]. This dataset provides additional information, such as user-location ratings, user comments, user information, and location information. However, no social network is known and the social connections can only be estimated, e.g., by similar check-ins at the same time. There is no authoritative ground-truth to validate these connections.
Synthetic Data: The problem of using sparse and noisy real-world LBSN data has already been identified in previous work [4, 5, 8]. However, none of these works proposed a way to obtain plausible check-in data. The authors of [8] generate user-location check-ins uniformly random without considering the semantics of the movement. Armenatzoglou et al. [4, 5] create check-ins randomly following a simple distance-based power-law distribution.

In summary, the experimental results of existing work on LBSNs may be considered inconclusive, both in terms of scalability and effectiveness due to a lack of available datasets. This vision paper aims at closing this gap by proposing the means to generate large scale and ground-truth based synthetic datasets through simulation. Synthetic data would allow insights into what is possible concerning new and improved geoinformation systems, but also in terms of privacy and anonymization research without having to raise any privacy concerns.

4 LOCATION-BASED SOCIAL SIMULATION

Our vision is to create a geo-simulation framework to generate high-quality LBSN data. The framework will model users living and traveling in an urban environment, going home at night, working during work days, and visiting recreational locations. Individual

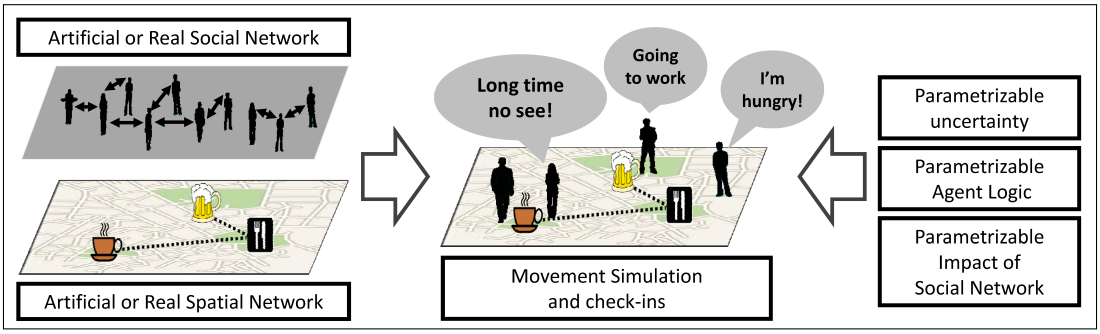

Figure 2: LBSN Check-in Data Simulation Framework: Overview.

user preferences will guide the choice of locations. Simulated individuals co-locating in space and time may become friends, depending on the type of place where they meet. Following this underlying social network, friends will travel and visit locations together. Based on psychology and social science theories [3, 14], individuals will exhibit a socially plausible behavior. Simulation parameters will be calibrated to create scenarios similar to the real world, thus allowing us to create massive sets of simulated LBSN data, capturing all individuals of our simulated world with a valid ground-truth (having no uncertainty) and without impacting the privacy of any human subject in the real world. As a deliverable, this research will yield synthetic LBSN data sets of hundreds of thousands of users, scaling to years of observed user data, and thus creating gigabytes of meaningful check-in and social interaction data.

4.1 Challenges

The first challenge in this vision is to enrich the simulation with *plausible human behavior* by integrating psychological/social theories such as Maslow's hierarchy of needs [14] and the theory of planned behavior [3]. As such, individuals' actions are driven by their needs, e.g., physiological needs such as food, financial needs, and social needs to meet friends and family. More importantly, such actions should align statistically with real-world measurements.

The second challenge is the creation of a *scalable and efficient geo-simulation design* to accommodate millions of individuals to be simulated simultaneously. At each decision point, individuals need to efficiently decide when and where to move based on limited information about their environment. Using pruning techniques, we plan to avoid evaluating predicates that did not change between decision points. At the same time, navigation on a spatial network needs to be efficient. Pre-computing and preferentially caching shortest paths will speed up the spatial network-related operations. Parallelization of the simulation is challenging, since social and spatial networks are not independent and social networks may change dynamically.

4.2 Agent-Based Modeling Approach

One possible approach to implement the envisioned framework is to employ agent-based modeling. The main idea is to create a massively scalable implementation of the agent-based model based on algorithmic innovation. There are many agent-based modeling platforms; one example is the MASON (Multi-Agent Simulation of Neighborhoods) open-source simulation toolkit [13] and its GIS extension, GeoMASON [20]. MASON has been used in the past to develop agent-based models to describe complex social interaction that is based on the agent's location in space and time, including models for riot prediction [19], simulating the spread of disease [6], and the emergence of slums in urban environments [18]. These simulations consider a limited number of agents, over only a few days of simulation time. One of the main challenges to be addressed in such an agent-based simulation is to scale the system, including spatial properties, by using efficient resource distribution algorithms and index structures to avoid computational bottlenecks.

The envisioned simulation framework will create simulated worlds in which virtual individuals (agents) move and interact with the environment and with each other. A sketch of this simulation framework is illustrated in Figure 2. Each simulation will be based on (real or synthetic) spatial networks with locations and social networks of users. Agents will have home and work locations, which may change over time. They will follow daily patterns of life such as going to work and visiting recreational places. Agents may aim to maximize attributes such as "happiness" and "cash" over years of simulation time. Different agents may have different goals, thus maximizing different attributes. Each agent will have choices, such as preferring a certain type of restaurant, cafe, or bar. The simulation will store spatial and social information of agents into log-files and use Google BigTable for distributed and compressed storage once the files become too large.

4.3 Research Applications

The envisioned simulation would directly benefit many research endeavors using location-based social simulation, including social link discovery, location recommendation, community detection, and others, as described in the following.

Social Link Prediction. Traditionally, the quality of existing link prediction methods is evaluated by removing a fraction of links from the social network, and testing how well existing solutions can predict these links using the remaining links for training. A major problem with this approach is that it is unclear how accurately the LBSN social network reflects the real world. Are there missing links? Are there false links? How much correct signal per noise do these datasets yield? Are existing solutions overfitting to this noise? Unfortunately, these questions are challenging to answer, as there is no way to validate whether two friends that are reported in any of the real-world LBSN datasets (see Section 3) are actual friends. We can fill this gap by exploiting the fact that our agents are real friends, who have deliberately chosen to visit a location together, and which agents are sharing the same location coincidentally.

Location Recommendation. To recommend locations, we have simulated individuals rate locations (on a five "star" scale - as illustrated in Figure 1). This rating will be determined by a deterministic function of the agents' preferences and the locations' attributes. The result will be obfuscated by random noise (of parameterizable degree). Then, we can evaluate how existing methods, such as regression models and matrix factorization models can exploit the (latent) agent preferences and location attributes for prediction. Unlike in real-world recommendation systems where the recommendation matrix is sparse, we can simulate restaurant visits and ratings. Such ground-truth data would enable 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.

Community Detection. For the tasks of community detection and social network clustering, we can impose circles of friends (strongly connected groups) in our social network. Then, by observing co-locations from the data, we can see which existing solutions are able to best approximate the imposed ground truth social networks. Thus, we can obtain a ground-truth of communities to evaluate the accuracy of algorithms against.

Other Applications. For all LBSN problems mentioned, an additional application that simulated data will allow evaluating whether existing algorithms able to scale to large and dense datasets. Such results may reveal computational bottlenecks that were invisible on small and sparse real-world datasets. Another application is the simulation and analysis of traffic solutions. We can examine traffic solutions with centralized control, which is a realistic scenario of the autonomous vehicle future, as envisioned by Bryan Mistele of INRIX [15]. This will allow to evaluate different fleet-based routing strategies and compare them to traditional self-optimizing driving strategies. Finally, synthetic LBSN datasets will benefit research towards location privacy. Existing work [7, 17] points out the sensitive location privacy aspect of LBSN datasets. Recent work [17] shows that an in-depth study requires much larger datasets. Given the general lack of such data, synthetic data will include high-fidelity trajectories of individual users and help show how location privacy is even more at a risk if more user data was available.

5 CONCLUSIONS

Our vision is to employ geo-simulation to generate large-scale and high-fidelity location-based social network datasets. We see this as an open problem, as existing real-world LBSN datasets are insufficient in terms of size and data reliability, which inhibits the broader impacts of data-driven research using LBSN data. Towards the vision of location-based social simulation, we identify two main challenges: (i) plausibility, in terms of generating data that exhibits realistic social behavior, in order to make inductions from results on the simulated data onto the real world, and (ii) scalability, to simulate millions of agents over years of simulation time, thus potentially generating LBSN data of whole generations. Once such a location-based social simulation framework has been developed, research on LBSN, including link prediction, next-location prediction, location recommendation, and community detection, will be revitalized given the availability of massive datasets and by having an authoritative ground-truth pertaining to the correctness of data.

ACKNOWLEDGMENT

This project is sponsored by the Defense Advanced Research Projects Agency (DARPA) under cooperative agreement No.HR00111820005. 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] Stanford network analysis project. https://snap.stanford.edu/index.html. Accessed: 2019-07-01.
[2] Yelp dataset challenge. round 9. https://www.yelp.com/dataset_challenge. Accessed: 2017-07-30.
[3] I. Ajzen. The theory of planned behavior. *Organizational behavior and human decision processes*, 50(2):179–211, 1991.
[4] N. Armenatzoglou, R. Ahuja, and D. Papadias. Geo-social ranking: functions and query processing. *The VLDB Journal*, 24(6):783–799, 2015.
[5] 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.
[6] A. T. Crooks and A. B. Hailegiorgis. An agent-based modeling approach applied to the spread of cholera. *Environmental Modelling & Software*, 62:164–177, 2014.
[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] J. J. Levandoski, M. Sarwat, A. Eldawy, and M. F. Mokbel. Lars: A location-aware recommender system. In *Data Engineering (ICDE), 2012 IEEE 28th International Conference on*, pages 450–461. IEEE, 2012.
[9] 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.
[10] Y. Li, R. Chen, J. Xu, Q. Huang, H. Hu, and B. Choi. Geo-social k-cover group queries for collaborative spatial computing. *IEEE Transactions on Knowledge and Data Engineering*, 27(10):2729–2742, 2015.
[11] X. Liu, Y. Liu, K. Aberer, and C. Miao. Personalized point-of-interest recommendation by mining users' preference transition. In *Proceedings of the 22nd ACM International Conference on Information & Knowledge Management*, CIKM '13, pages 733–738, New York, NY, USA, 2013. ACM.
[12] Y. Liu, T.-A. N. Pham, G. Cong, and Q. Yuan. An experimental evaluation of point-of-interest recommendation in location-based social networks. *Proceedings of the VLDB Endowment*, 10(10):1010–1021, 2017.
[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] A. H. Maslow. A theory of human motivation. *Psychological review*, 50(4):370, 1943.
[15] B. Mistele. Building smarter cars & cities from spatial data. SIGSPATIAL 2018 Conference Keynote (http://sigspatial2017.sigspatial.org/keynotes/#bryan).
[16] A. Noulas, S. Scellato, N. Lathia, and C. Mascolo. Mining user mobility features for next place prediction in location-based services. In *Data mining (ICDM), IEEE 12th international conference on*, pages 1038–1043. IEEE, 2012.
[17] J. D. Park, E. Seglem, E. Lin, and A. Züfle. Protecting user privacy: Obfuscating discriminative spatio-temporal footprints. In *ACM SIGSPATIAL LocalRec Workshop*, page 2. ACM, 2017.
[18] A. Patel, A. Crooks, and N. Koizumi. Slumulation: an agent-based modeling approach to slum formations. *Artificial Societies and Soc. Simulation*, 15(4), 2012.
[19] B. Pires and A. T. Crooks. Modeling the emergence of riots: a geosimulation approach. *Computers, Environment and Urban Systems*, 61:66–80, 2017.
[20] K. Sullivan, M. Coletti, and S. Luke. Geomason: Geospatial support for mason. Technical report, George Mason University, 2010.
[21] Y.-T. Wen, Y. Y. Fan, and W.-C. Peng. Mining of location-based social networks for spatio-temporal social influence. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 799–810. Springer, 2017.
[22] D. Yang, B. Qu, J. Yang, and P. Cudre-Mauroux. Revisiting user mobility and social relationships in lbsns: A hypergraph embedding approach. In *Proceedings of The Web Conference*, WWW'19, 2019.
[23] 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.
[24] M. Ye, P. Yin, and W.-C. Lee. Location recommendation for location-based social networks. In *Proceedings of the 18th SIGSPATIAL international conference on advances in geographic information systems*, pages 458–461. ACM, 2010.
[25] Y.-L. Zhao, Q. Chen, S. Yan, D. Zhang, and T.-S. Chua. Community understanding in location-based social networks. In *Human-Centered Social Media Analytics*, pages 43–74. Springer, 2014.
[26] Y. Zheng. Location-based social networks: Users. In *Computing with spatial trajectories*, pages 243–276. Springer, 2011.
[27] Y. Zheng and X. Xie. Location-based social networks: Locations. In *Computing with spatial trajectories*, pages 277–308. Springer, 2011.