Human Mobility-Based Synthetic Social Network Generation

Ketevan Gallagher*
Thomas Jefferson High School
Alexandria, Virginia, USA
ketevangallagher@gmail.com

Srihan Kotnana*
Westfield High School
Chantilly, Virginia, USA
srihank8@gmail.com

Sachin Satishkumar*
Thomas Jefferson High School
Alexandria, Virginia, USA
skumarr502@gmail.com

Kheya Siripurapu*
Chantilly High School
Chantilly, Virginia, USA
kheya6@gmail.com

Justin Elarde
George Mason University
Fairfax, Virginia, USA
jelarde@gmu.edu

Taylor Anderson
George Mason University
Fairfax, Virginia, USA
tander6@gmu.edu

Andreas Züfle
George Mason University
Fairfax, Virginia, USA
azufle@gmu.edu

Hamdi Kavak
George Mason University
Fairfax, Virginia, USA
hkavak@gmu.edu

ABSTRACT

Location-Based Social Networks (LBSNs) combine location information with social networks and have been studied vividly in the last decade. The main research gap is the lack of available and authoritative social network datasets. Publicly available social network datasets are small and sparse, as only a small fraction of the population is captured in the dataset. For this reason, network generators are often employed to generate social networks to study LBSNs synthetically. In this work, we propose an evolving social network implemented in an agent-based simulation to generate realistic social networks. In the simulation, as agents move to different places of interest have the chance to make social connections with other agents as they visit the same place. A large-scale real-world mobility dataset informs the choice of places that agents visit in our simulation. We show qualitatively that our simulated social networks are more realistic than traditional social network generators, including the Erdos-Renyi, Watts-Strogatz, and Barabasi-Albert.

ACM Reference Format:
Ketevan Gallagher, Srihan Kotnana, Sachin Satishkumar, Kheya Siripurapu, Justin Elarde, Taylor Anderson, Andreas Züfle, and Hamdi Kavak. 2022. Human Mobility-Based Synthetic Social Network Generation. In *The 2nd ACM SIGSPATIAL International Workshop on Animal Movement Ecology and Human Mobility (HANIMOB '22) (HANIMOB '22), November 1, 2022, Seattle, WA, USA.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3557921.3565540

1 INTRODUCTION

Social networks offer a way to represent and examine relationships between individuals. Through these models of relationships, we gain insight into the real world. With the advent of social media and the Internet, the world is more connected and data is more accessible. Especially with mobile devices, data is tracked to a much higher and increasing degree. To gain a more nuanced model, a spatial component can be merged into social networks, creating Location-Based Social Networks (LBSN), bridging the gap between the physical world and online social networking services [20]. LBSNs have been used in disease modeling [8, 13], urban planning [16], and marketing [3]. However, research on LBSNs is limited due to the absence of comprehensive and accurate social network datasets. Publicly available real-world datasets only capture a small fraction of the population [10]. The authors of [9] even conclude that "Researchers working with LBSN data sets are often confronted by themselves or others with doubts regarding the quality or the potential of their data sets.". Therefore, synthetic social networks are often used to study LSBNs [6].

In this study, we aim to understand if social networks created through agent-based simulation informed by real-world human mobility data can create more realistic social network datasets than traditional social network data generators. Specifically, we implement a growing social network using an agent-based simulation based on human mobility data for Fairfax County, USA. Based on interactions between agents within the model, the network continuously evolves based on these interactions, losing and gaining connections [18]. Through fine-scale, real-world foot-traffic data from SafeGraph[1], the agent-based simulation generates realistic mobility data on which the evolving social network is based on. By comparing our generated social networks to commonly used random social network generators, our qualitative evaluation shows that this approach promises to yield more realistic social networks, which include large and small social sub-communities of closely connected agents. To close the gap of a lack of publicly available large real-world social network datasets, we make our simulated social networks available for up to 1.1 million agents having an average number of friends of six, and we provide the source code of our simulation for researchers to create their social networks for other study regions.

*These authors contributed equally to the paper and are named in alphabetical order.

[1] https://www.safegraph.com/guides/foot-traffic-data

HANIMOB '22, November 1, 2022, Seattle, WA, USA

ACM ISBN 978-1-4503-9534-2/22/11...$15.00
https://doi.org/10.1145/3557921.3565540

2 AGENT-BASED SIMULATION

We use an agent-based simulation of Fairfax County, VA, with a population of over 1.1 million individuals. To inform the human mobility of the simulation, we use empirical data obtained from SafeGraph foot-traffic data. This dataset provides data about visits of individuals grouped into home census block groups (CBGs) to points of interest (POI). A CBG is a geographical unit used by the United States Census Bureau, having a population of 600 to 3,000 people. Points of interest data are taken from SafeGraph's POI databases and are places where money can be spent[2], and includes places such as restaurants, schools, and gas stations.

We use this data by estimating, for each CGB *cbg* and each POI *poi*, the probability $\hat{p}(cbg,poi)$ that an individual from *cbg* visits *poi*. Using foot-traffic data from 2019-2021, this probability is estimated empirically, by dividing the number $\text{count}(cbg, poi)$ of observed visits of individuals from *cbg* to *poi*, divided by the total number of visits $\text{count}(cbg)$ from *cbg* to any place of interest, that is:

$$\hat{p}(cbg, poi) = \frac{\text{count}(cbg, poi)}{\text{count}(cbg)}$$

Based on this mobility model, each individual (having a home CBG c) creates a new schedule of POIs to visit each day. POIs to visit are selected from probability distribution defined by

$$P_c(poi) = \hat{p}(c, poi),$$

for each POI in the simulation. For example, if we have observed $4,000$ POI visits in a CBG c, out of which 20 visits were made to POI poi, then we would have a probability of $\frac{20}{4000} = 0.5\%$ for an agent from CBG c to visit poi. Based on the average duration of a stay at a POI and arrival times (also taken from SafeGraph data), agents repeat drawing new POIs until they fail to add an additional POI due to a lack of time in their schedule. Due to arrival times learned empirically from the data, most agents visit locations between 8 am-11 pm but can occasionally make trips outside of that.

Once the daily schedule of all agents has been computed, agents execute their schedule by visiting their planned POIs. Whenever an agent visits a POI, they choose another agent located at the same POI (if any) to create a social connection with. An edge in the social network is established between this pair of agents with a probability of p_{focal}, which is a parameter of our simulation. These connections denoted as focal closure, are based on a small probability that two strangers may become friends when meeting [7, 17]. In addition, when an agent enters a POI and finds that two existing friends are already located at the same POI, then there is a chance of $p_{triadic}$ that the agent closes the triangle by introducing these two friends to each other, a concept denoted as triadic closure [15].

Our simulation runs until the average network degree reaches a specified parameter *degree*. This parameter is used to make our resulting social network comparable to other synthetic social network generators, which also allows us to specify the desired average network degree.

3 QUALITATIVE EVALUATION

Our goal is to compare our simulated social network with the three common classical models, Erdos-Renyi, Watts-Strogatz, and Barabasi-Albert. To scale these network generators to large networks, we describe our implementation details for each of the networks and illustrate the resulting network using a small sample of agents for visualization.

[2]https://www.safegraph.com/products/places

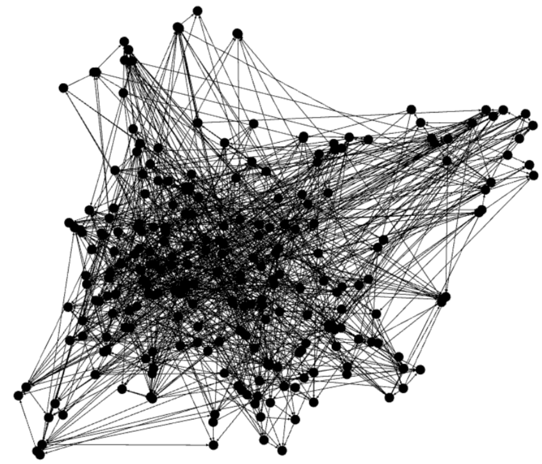

Figure 1: Erdos-Reyni model visualization with 250 nodes

3.1 Erdos-Renyi

The Erdos-Renyi model [5] is a randomly generated model. Each edge is only included in the graph based on a chosen probability. Edges are chosen independently of other edges and nodes. Typically, the Erdos-Renyi model is generated by looping through every edge pair between every node, choosing a random probability p_{Erdos}, and adding the edge only if the random probability is below a certain threshold established beforehand. However, this naive implementation is inapplicable for large social networks due to its quadratic run-time. We implemented a faster algorithm to generate Erdos-Renyi graphs by using, for each of the n nodes, a binomial distribution $B(n-1, p_{Erdos})$ to determine the number of friends that an agent will have. This random variable can be approximated for large values of n in constant time using a Normal approximation with continuity correction. Once the number of friends an agent will have is determined, these are drawn randomly from the set of all agents. This approach allows us to determine the friends of an agent in $O(degree)$ rather than in $O(n)$, where the degree is the average degree of the network, which is (for large networks) generally much larger than the number of nodes n.

Figure 1 shows an example Erdos-Renyi graph with 250 nodes (for visualization) and having an expected degree of 4, that is, each edge has a probability of $p_{Erdos} = \frac{4}{250}$ of existing. We observe that the resulting graph has no structure; Using OpenOrd, an open-source toolbox to visualize large graphs [11], we cannot observe any meaningful clusters. This is expected, as the network is entirely random. We note that all visualization in this work uses the OpenOrd graph visualization algorithm implemented in Gephi, which is an open-source software for exploring and manipulating networks [2].

3.2 Watts-Strogatz

The Watts-Strogatz model [19] is a randomly generated model that aims to have social networks with a high degree of clustering. It is a small-world model with short average path lengths. The We use

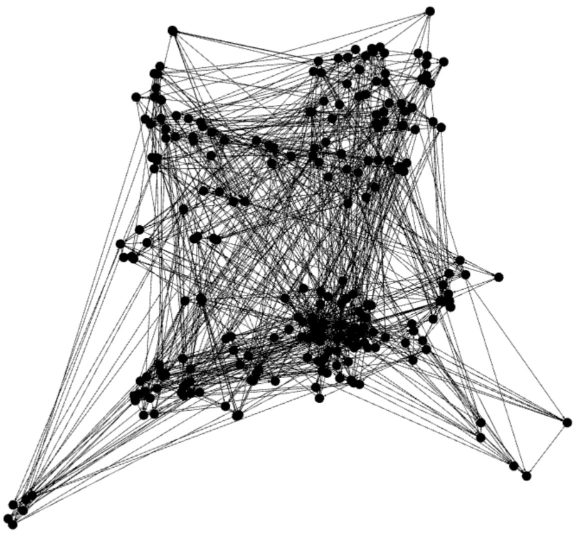

Figure 2: Watts-Strogatz model visualization with 250 nodes

the traditional algorithm that starts by constructing a ring lattice in which each node is circularly connected to its d next nodes in arbitrary order, where d is the desired degree of the network. With a probability of p_{Watts} each connection is rewired to a randomly selected node. Figure 2 shows a generated Watts-Strogatz network having $p = 0.1$. We observe that even after using OpenOrd [11], it is difficult to discriminate any interesting clusters of communities. We still observe the single ring-shaped cluster that stems from the initial ring lattice. Beyond the single-ring cluster, there is little evidence of interesting sub-clusters or communities that one would expect in a real-world social network.

3.3 Barabasi-Albert

The Barabasi-Albert model [1] is a randomly generated model that aims to create a scale-free network, one that follows a power law in the distribution of the number of edges per node. To do this, the model utilizes preferential attachment. Nodes are added to the network iteratively and are more likely to connect to nodes with a high number of edges. This model aims to simulate the emergence of very strongly connected influencers. Figure 3 depicts a generated Barabasi-Albert network. We once again observe no realistic structure in the network. While some nodes have a very large number of friends, all friends are still selected randomly, so there are no clusters or communities.

3.4 Our Proposed ABM-Based Social Network Generator

We ran our simulation of Fairfax County with 250 agents for comparison. For such a small simulation, we reduced the study to only ten CBGs located in the southeast of Fairfax County. Figure 4 depicts our simulation for 250 agents to allow a fair comparison to Figures 1-3. In this case, we can now observe reasonable clusters. We find one cluster that is largely isolated from the rest and one large cluster which has three smaller clusters within it.

We further show our simulated social network for 10,000 nodes in Figure 5. While, at first glance, the resulting network appears to be a single random furball, a closer look shows texture created

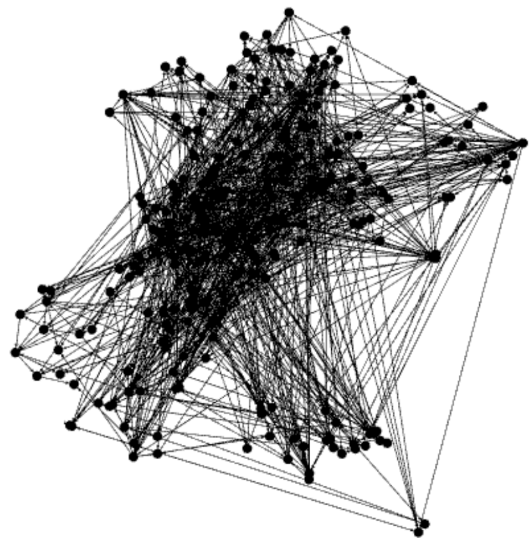

Figure 3: Barabasi-Albert model visualization with 250 nodes

by smaller and larger clusters and communities throughout the network. As these clusters are difficult to discern on paper, we provide a high-resolution version of this network on our project website at https://github.com/STIP-Summer-2021/Growing-Social-Networks. We believe that our generated social network can closely mimic real friends' networks.

4 QUANTITATIVE EVALUATION

We ran all four models using 100,000 nodes and set the average degree to 10. For the Watts-Strogatz network, we used a rewiring probability of 50 percent, and we initialized the Barabasi-Albert network with a set of 10 nodes. Table 1 shows network statistics for all four considered networks.

Network Diameter: Out of the three classical networks, the Watt-Strogatz model had the highest network diameter of 62, which is expected due to the initial ring-like structure, which may create long dead-ends after rewiring. Barabasi-Albert model had a diameter of 25, which is still very large, explained by potentially long chains of agents having very few friends. The Erdos-Reyni model had a diameter of 36. Interestingly, our model had the lowest diameter, showing that all agents can reach each other on the network with at most 12 hops, which seems realistic for a 100K agent network.

Number of Triangles: The Erdos-Reyni model had the least number of triangles: 155, as a triangle has a meager chance of p_{Erdos}^3. The Barabasi-Albert had a total of 35,153 triangles, mainly stemming from neighbors on the initial ring lattice, and the Watts-Strogatz model had the greatest number of triangles: 565,730, which is due to some agents having a very large number of friends, thus turning any pairs of friends among them into triangles. Our simulation has 211,305 triangles, which is a high number given the low variance in the number of friends, showing a high degree of clusters of the simulated network.

Standard Deviation: In the Watts-Strogatz model, all agents have the same number of friends, so there is no standard deviation. As expected, the Barabasi-Albert network has an extreme standard deviation of 30.5, as some agents have hundreds of friends. Our

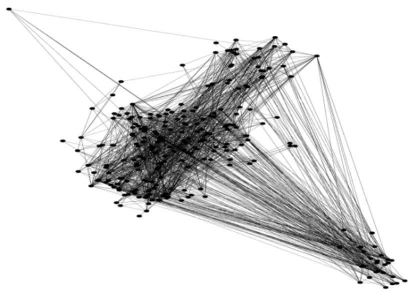

Figure 4: Simulated social network with 250 nodes

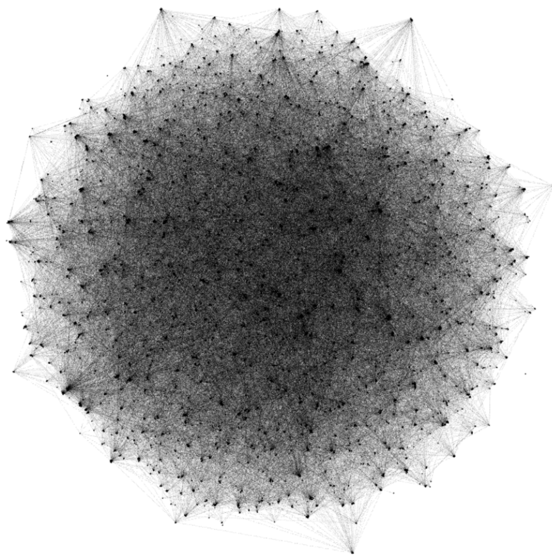

Figure 5: Simulated social network with 10,000 nodes

	Erdos	Watts	Barabasi	Simulated
Diameter	36	62	25	12
# Triangles	155	35,153	565,730	211,305
Degree std. dev.	2.23	0.00	30.5	8.54

Table 1: Network Statistics for the three baseline network generators and our proposed simulated network.

simulated network has a moderate standard deviation of 8.54, with 75% of agents having between 3 and 20 friends while also having 12K agents with less than three friends and 13K agents having more than 20 friends. We show a detailed distribution of the number of friends, visualizations of the network with different parameters, simulation source code, and datasets on our Github page at https://github.com/STIP-Summer-2021/Growing-Social-Networks .

5 CONCLUSIONS, AND FUTURE WORK

We proposed an agent-based simulation to create realistic social networks informed by real-world human mobility data where agents travel to points of interest to befriend other agents. We have shown first results of our model simulating realistically looking social networks and hope that our published social network dataset may be useful to other researchers as a testbed to evaluate algorithms and theories on large and realistic social networks. As next steps, our goal is to calibrate the distribution of the number of friends of agents to match social science results such as Dunbar's Number [4]. We note that calibration to real-world data is challenging as no ground truth datasets exist. Social media network datasets (such as Facebook, Twitter) do not exactly capture friendship as no human may have hundreds of millions of (true and stable) friendship relationships. We also plan on implementing more realistic agent mobility having agents adapt their mobility to more frequently meet friends. Given our understanding that human mobility affects infectious disease spread [12, 14], we want to investigate how realistic social networks affect infectious disease spread models.

ACKNOWLEDGEMENTS

This work is supported by National Science Foundation Grant #2109647 titled "Data-Driven Modeling to Improve Understanding of Human Behavior, Mobility, and Disease Spread" and by the Aspiring Scientists Summer Internship Program (ASSIP) at George Mason University.

REFERENCES

[1] A.-L. Barabási and R. Albert. Emergence of scaling in random networks. *science*, 286(5439):509–512, 1999.

[2] M. Bastian, S. Heymann, and M. Jacomy. Gephi: an open source software for exploring and manipulating networks. In *Proceedings of the international AAAI conference on web and social media*, volume 3, pages 361–362, 2009.

[3] J. Capdevila, M. Arias, and A. Arratia. Geosrs: A hybrid social recommender system for geolocated data. *Information Systems*, 57:111–128, 2016.

[4] R. I. Dunbar. Neocortex size as a constraint on group size in primates. *Journal of human evolution*, 22(6):469–493, 1992.

[5] P. Erdős and A. Rényi. On random graphs i. *Publicationes mathematicae*, 6(1):290–297, 1959.

[6] 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 *MDM*, pages 158–167. IEEE, 2020.

[7] 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 *ACM SIGSPATIAL*, pages 576–579, 2019.

[8] 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.

[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. 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.

[11] S. Martin, W. M. Brown, R. Klavans, and K. W. Boyack. Openord: an open-source toolbox for large graph layout. In *Visualization and Data Analysis 2011*, volume 7868, pages 45–55. SPIE, 2011.

[12] M. Mokbel, M. Sakr, L. Xiong, A. Züfle, J. Almeida, T. Anderson, W. Aref, G. Andrienko, N. Andrienko, Y. Cao, et al. Mobility data science (dagstuhl seminar 22021). In *Dagstuhl Reports*, volume 12. Schloss Dagstuhl-Leibniz-Zentrum für Informatik, 2022.

[13] M. Morris. Epidemiology and social networks: Modeling structured diffusion. *Sociological methods & research*, 22(1):99–126, 1993.

[14] J. Pesavento, A. Chen, R. Yu, J.-S. Kim, H. Kavak, T. Anderson, and A. Züfle. Data-driven mobility models for covid-19 simulation. In *Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Advances in Resilient and Intelligent Cities*, pages 29–38, 2020.

[15] G. Simmel. *The sociology of georg simmel*, volume 92892. Simon and Schuster, 1950.

[16] R. Smarzaro, T. F. d. M. Lima, and C. A. Davis Jr. Could data from location-based social networks be used to support urban planning? In *Proceedings of the 26th International Conference on World Wide Web Companion*, pages 1463–1468, 2017.

[17] J. Urata and E. Hato. Dynamics of local interactions and evacuation behaviors in a social network. *Transportation research part C*, 125:103056, 2021.

[18] S. Wasserman, K. Faust, et al. Social network analysis: Methods and applications. 1994.

[19] D. J. Watts and S. H. Strogatz. Collective dynamics of 'small-world'networks. *nature*, 393(6684):440–442, 1998.

[20] Y. Zheng. Location-based social networks: Users. In *Computing with spatial trajectories*, pages 243–276. Springer, 2011.