Synthetic Geosocial Network Generation

Ketevan Gallagher
Thomas Jefferson High School for Science and Technology
Alexandria, Virginia, USA
ketevangallagher@gmail.com

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

Andrew Crooks
University at Buffalo
Buffalo, New York, USA
atcrooks@buffalo.edu

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

ABSTRACT

Generating synthetic social networks is an important task for many problems that study humans, their behavior, and their interactions. Geosocial networks enrich social networks with location information. Commonly used models to generate synthetic social networks include the classical Erdős-Rényi, Barabási-Albert, and Watts-Strogatz models. However, these classic social network models do not consider the location of individuals. Real-world geosocial networks do exhibit a strong spatial autocorrelation, thus having a higher likelihood of a social connection between agents that are spatially close. As such, recent variants of the three classical models have been proposed to consider location information. Yet, these existing solutions assume that individuals are located on a uniform lattice and exhibit certain limitations when applied to real-world data that exhibits clusters. In this work, we discuss these limitations and propose new approaches to extend the three classic social network generation models to geosocial networks. Our experiments show that our generated synthetic geosocial networks address the shortcomings of the state-of-the-art models and generate realistic geosocial networks that exhibit high similarity to real-world geosocial networks.

CCS CONCEPTS

• **Information systems → Geographic information systems**.

KEYWORDS

Geosocial Networks, Network Generation, Synthetic Social Networks, Erdos-Renyi, Watts-Strogatz, Barabasi-Albert

ACM Reference Format:
Ketevan Gallagher, Taylor Anderson, Andrew Crooks, and Andreas Züfle. 2023. Synthetic Geosocial Network Generation. In *7th ACM SIGSPATIAL Workshop on Location-based Recommendations, Geosocial Networks and Geoadvertising (LocalRec '23), November 13, 2023, Hamburg, Germany*. ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3615896.3628345

LocalRec '23, November 13, 2023, Hamburg, Germany

ACM ISBN 979-8-4007-0358-4/23/11…$15.00
https://doi.org/10.1145/3615896.3628345

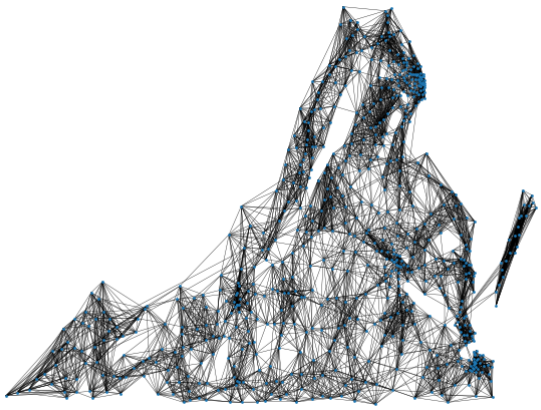


Figure 1: Real- World Geosocial Network using Facebook Social Connectedness Data between Zone Improvement Plan (ZIP) Region Centroids for the State of Virginia, USA.

1 INTRODUCTION

Geosocial networks [5] (often also called location-based social networks [26]) capture both 1) social relationships such as friendship between individuals or populations, and 2) the location of these individuals or populations and their interactions. An example of a real geosocial network is shown in Figure 1 showing the location of ZIP code population centroids that are linked based on the strength of their social connectedness. This network is generated using the Facebook Social Connectedness Index [4]. Such networks are often used to improve the realism of models where social interactions are of importance, such as agent-based models. For example, geosocial networks have been used in disease modeling [11, 15], urban planning [10, 16], and marketing [23]. However, research on geosocial networks 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 [21]. The authors of [20] even conclude that "researchers working with LBSN [Location-Based Social Network] datasets 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 in models, and the more realistic these synthetic social networks are the more realistic these models will be. To generate

synthetic geosocial networks, existing work leverages geosimulation to create a digital twin of a real urban environment and uses co-location between simulated agents [12, 18, 19]. However, such simulation approaches requires a large computational overhead to simulate entire cities.

In terms of synthetic (non-geo) social networks, there are three classical models for generating them: the Erdős-Rényi [9], Barabási-Albert [7], and Watts-Strogatz [24] models. From these classical models, different network structures with properties that can sometimes reflect real observed networks emerge. However, real networks and the dynamics that drive their formation and evolution are embedded within geographic space. Thus, without using location information, we would argue that these classical models can not generate realistic geosocial networks. Methods to improve these classical models have been introduced, such as using geographic coordinates to inform the way nodes in the network connect (link) to one another [1]. However, these methods place nodes randomly or uniformly in geographic space and have not been tested with real-world geospatial data to inform the position of nodes. One could consider this as a drawback, as real-world data capturing the location of populations exhibits spatial heterogeneity where some locations have dense clusters of nodes and others that are sparse [3]. Such heterogeneity is one of the aspects making spatial data special [2].

In this paper, we demonstrate the shortcomings of these existing models and propose algorithms to generate new spatial versions of these three classical models that can be applied to create social networks between vertices having real-world locations. We compare each of these three networks to their non-spatial counterpart, which does not take distance into account when generating the connections. We also compare the generated networks to real-world networks generated from mobility data [17] and Facebook user data to show that our synthetically generated social networks are, to some degree, similar to real-world geosocial networks.

The main contribution of this work is the introduction of three geosocial variants of the three classic synthetic social network generation models with the goal of realistically including location information in the social link generation process. Specifically, these three models are:

- **A Geosocial Erdős-Rényi model** In the classical Erdős-Rényi network [9], each node in the network has a constant probability p of connecting to every other node in the network. In the proposed spatial version of this network, the probability that two nodes will connect is not constant, but determined by the distance between the two nodes, using a power law equation following prior work: $(sd)^{-\alpha}$ where s is a scaling factor, d is the distance between nodes, and α is a distance-decay exponent set by the user. The scaling factor was introduced to accommodate for real-world data that may measure distance in different scales and different spaces such as (latitude, longitude) coordinates or absolute coordinates in meters.
- **A Geosocial Barabási-Albert model** The classical Barabási-Albert model [7] starts with a clique (a fully connected graph) of m nodes, which are connected to each other. Then, additional nodes are added by randomly connecting them to the nodes already in the network. Thus, nodes in the starting clique and nodes inserted early end up with a higher degree. The likelihood of making a connection is multiplied by the degree of the selection node. This leads to nodes with higher degrees having more connections, a mechanism known as preferential attachment [22]. The proposed spatial version of this network also starts with a clique of m nodes and nodes are then added one by one. When nodes are added, the likelihood of making a connection is based on the power law equation that is used in the geosocial Erdős-Rényi network. In addition, instead of processing nodes in an arbitrary order, our model adds nodes in a spatial order by iteratively adding nodes by order of increasing distance to the centroid of already added nodes.
- **A Geosocial Watts-Strogatz model** The classical Watts-Strogatz model [24] starts with each node connected to its k nearest neighbors on a random lattice. Each node has a constant probability of "rewiring" to another randomly chosen node. We propose two different geosocial variants using location to different degrees. In both models, each node starts connected to its k nearest neighbors in the location space (rather than a random lattice), but the rewiring is done using two different variants: 1) having a rewiring probability determined by the distance between the two nodes, and 2) having a rewiring probability using a constant parameter. In both models, the new node selected for rewiring is chosen based on the power law used in the geosocial Erdős-Rényi model.

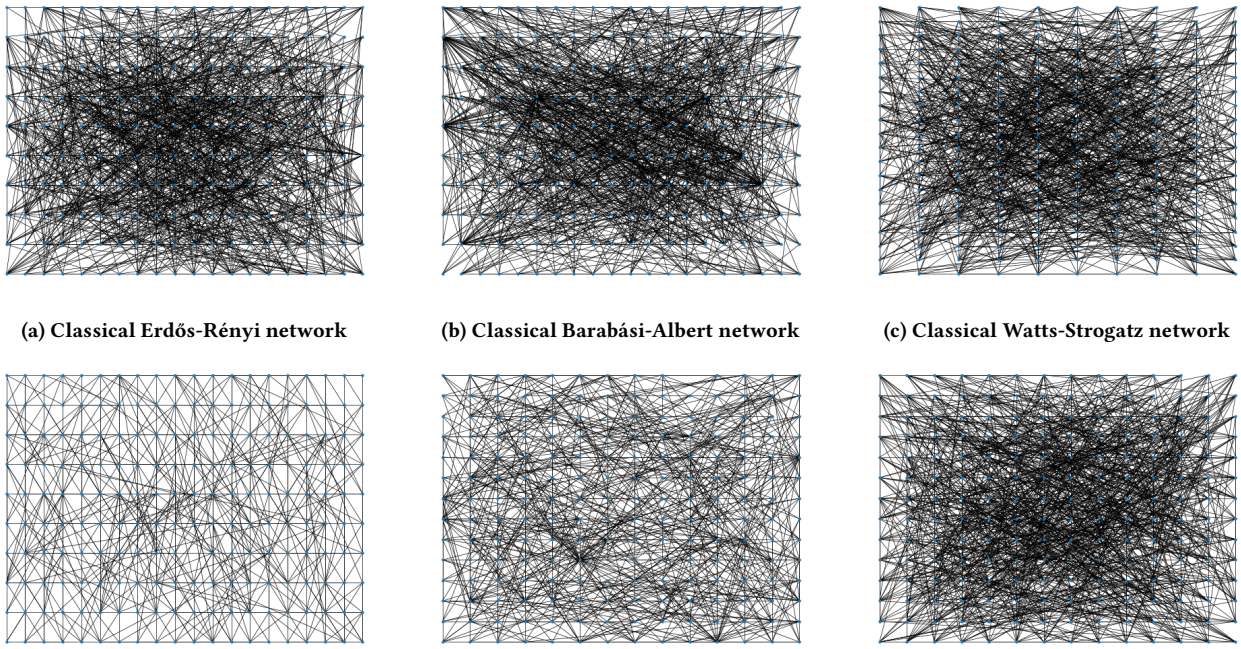
The remainder of this work is organized as follows. Section 2 provides details of the three classic models for synthetic social network generation, their spatial extensions, and the shortcomings of these models. Section 3 then presents our proposed geosocial network generation of the classic models that control for the spatial heterogeneity of the location of nodes. Our experimental evaluation is found in Section 4 showing both qualitative and quantitative experiments to support our claims that our proposed synthetic geosocial network generation algorithms yield realistic (similar to real-world) social networks. Finally, Section 5 provides a brief conclusion to the paper and a discussion of areas of further work.

2 RELATED WORK

Following our brief introduction in Section 1, this section provides an overview of the three classical synthetic social network models and spatial versions of these network models that have been proposed.

2.1 The Erdős-Rényi Model

The Erdős-Rényi model [9] 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 Erdős-Rényi 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. Figure 2a shows an example Erdős-Rényi graph with 200 nodes and an average degree of ten. In all of the graphs, each node is randomly assigned an x coordinate



(a) Classical Erdős-Rényi network (b) Classical Barabási-Albert network (c) Classical Watts-Strogatz network

(d) Spatial Erdős-Rényi network (e) Spatial Barabási-Albert network (f) Spatial Watts-Strogatz network

Figure 2: Graphs generated from the classical models and the spatial models proposed in [1]

from 0 to 9, and a y coordinate from 0 to 19. In order to compare the graphs visually, all of the graphs have the same layout, where the nodes, in blue, are arranged in a rectangular layout. Additionally, all of the graphs have a similar number of edges. For the Erdős-Rényi graph, no identifiable clusters or communities are visible. This is to be expected, as in this model, two nodes have the same likelihood of being connected. While the area in the center of the network appears darker, this is not due to denser connections but due to fact that many edges are crossing the entire network.

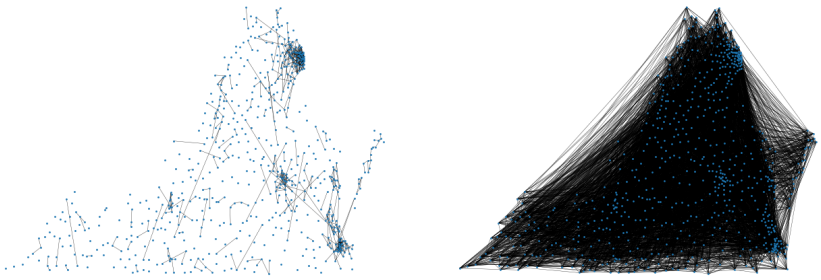
2.2 The Spatial Erdős-Rényi Model

The spatial version of the Erdős-Rényi model that has been proposed [1] does not use a predetermined consent probability to determine whether two nodes connect, but instead calculates a probability based on the distance between those two nodes. The model initializes a number of nodes, and assigns them Cartesian coordinates. For each possible connection between two nodes, the probability of connection is calculated using a power law equation: $p(d) = Cd^{-\alpha}$. Here, d is the distance between two nodes, and α is a distance-decay exponent set beforehand. C is a normalizing coefficient calculated by dividing $1 - \alpha$ by the difference of the minimum distance raised to $1-\alpha$ and the maximum distance raised to $1-\alpha$. C is also included in the power law equations for the Watts-Strogatz and Barabási-Albert networks included in [1], but in the code for that paper it is not included, so we do not include C when discussing or generating the spatial Watts-Strogatz and Barabási-Albert networks. If the probability of connection is greater than a randomly generated number, then an edge is drawn between those two nodes. Figure 2d shows a spatial Erdős-Rényi graph. Despite having the same number of edges, this graph appears less "dark" than the graph resulting from the classical model in Figure 2a. This is to be expected as the nodes are more likely to connect to nodes that are close, leading to shorter lines and more white space in the visualization. Still, this model does not exhibit any spatial clustering as all nodes are located in regions of equal density of nodes (except for border effects). When applying the spatial Erdős-Rényi model proposed in [1] to Virginia ZIP code data, the resulting graph as shown in Figure 3a having a large number of edges in dense regions of highly-populated areas where the distances between ZIP-code centroids are small. However, in sparsely populated regions where the distances between ZIP code centroids are large, almost no connections are made. Since the model originally assumes that all locations are located on a uniform grid, the model does not specify a parameter to account for differences in the density of locations.

A different variant of a spatially-aware version of the Erdős-Rényi model has been proposed in [25] using a constant threshold H (called neighborhood radius). Instead of giving each pair of vertices the same probability p of being connected, this model uses a probability p_1 for pairs of vertices having a spatial distance less or equal to H and a probability $p_2 < p_1$ for pairs of vertices having a distance larger than H. In practice, however, it is difficult to define a single such threshold. Thus, a continuous function that maps distance (between two vertices) to a probability of being connected as proposed in [1] and also adopted in this work is preferable.

2.3 Barabási-Albert

The Barabási-Albert model [7] is a randomly generated model that follows a power law in the distribution of the number of edges



(a) Spatial Erdős-Rényi Network with Virginia ZIP code data **(b) Spatial Watts-Strogatz Network with Virginia ZIP code data**

Figure 3: Graphs generated using the code from [1] implemented with Virginia ZIP code data

per node. The Barabási-Albert model helps explain networks such as the World Wide Web and social networks [6]. 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 higher number of edges. This model aims to create a network where there are few nodes with many connections and a large number of nodes that are sparsely connected. Figure 2b depicts a generated Barabási-Albert network. The typical Barabási-Albert structure can be observed, with some nodes having a high number of connections, but most having only a few connections.

2.4 The Spatial Barabási-Albert Model

The proposed spatial version of the Barabási-Albert model [1] starts with a clique with m number of nodes, and adds nodes iteratively. The first node that is added is connected to each of the starting m nodes. Each node that is added afterwards selects an m number of nodes from the existing network to connect to. Nodes are chosen randomly, and the probability of connection is calculated using a variation of the power law equation used for the Erdős-Rényi and Watts-Strogatz models, $p(d) = kd^{-\alpha}$. d and α are once again the distance between nodes and the distance-decay exponent respectively. k is the degree of node that is being considered for connection. If a randomly generated probability is less than the calculated probability and the two nodes are not already connected, an edge is made between those two nodes. Once the original node creates m number of edges, another node is added to the network and the process is repeated until the desired population number is reached. Figure 2e shows the spatial version of the Barabási-Albert network. Highly connected nodes can be seen, and there is more white space in this graph than in many of the others. This is due to the short length of edges, which is a result of nodes prioritizing connections to other nodes that are spatially close. When incorporating Virginia ZIP code data into the spatial Barabási-Albert network, the model does not terminate. This is caused by nodes located in sparse regions having a very low spatial density of nodes. For any such node, the likelihood of creating an edge will be very low, even for the nearest node, due to the exponential decay of probability in distance. For such nodes, the algorithm keeps attempting to create edges but keeps failing due to having a probability of succeeding of almost zero. This problem of non-termination was not evident in the experiments performed in [1] as this work assumes nodes to be located on a random lattice. Thus, the distances between nodes does not vary in these experiments. But using real-world data, we have both dense and sparse regions. Our proposed approach will account for these differences in the density of nodes and the exponential difference in resulting probabilities.

2.5 The Watts-Strogatz Model

The Watts-Strogatz model [24] is a randomly generated model that exhibits small world properties. The network is initialized having each node connected to a specified number of nearest neighbors. In the original version of the model, because the network is not spatially informed, its neighbors are not chosen spatially. Instead, each node connects to a given number of the nodes that have a similar identifier. For example, node #100 may be connected to nodes #96 to #99 and nodes #101 to #104.

After these connections are initialized, each connection is iterated through. For each node in the network, and for each connection between that node and another, a random probability is generated, and if this random probability is lower than a threshold parameter p (the likelihood of rewiring), the original node is rewired to another, randomly chosen node. Figure 2c shows a generated Watts-Strogatz network having $p = 0.2$. This was chosen so that on average one of a node's connections will be rewired, as this network has an average degree of ten. Because the coordinates are random and the nodes initially connect by identifier, it looks random and similar to the Erdős-Rényi network. Much like the Erdős-Rényi network, there is not much white space due to the long length of edges.

2.6 The Spatial Watts-Strogatz Model

Two spatial variations of the Watts-Strogatz network have been proposed [1]. In both variations, nodes are initialized like in the non-spatial Watts-Strogatz model using their nearest neighbors on the (arbitrarily chosen and non-spatial) unique identifiers of nodes. In the first version, each node and each connection to that node is iterated through and the connection's re-wiring probability is calculated using the same power law equation used in the spatial Erdős-Rényi network. Thus, distant nodes connected in the initial network are more likely to be rewired than close nodes.

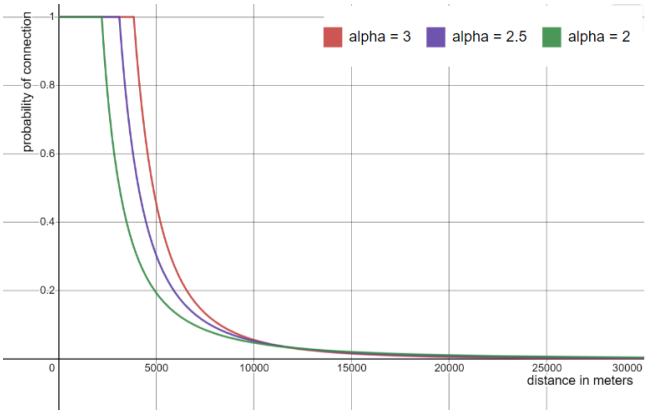


Figure 4: A graph showing how distances is mapped to probability of connection for different alpha values.

The second variation also starts by iterating through each node and their connection, but the probability that the node will rewire is determined by a constant probability p. The second node is chosen based on the power law equation used in the first version of the spatial Watts-Strogatz network. Thus, nodes are selected randomly for rewiring, but the rewiring is based on distance. Figure 2f depicts the first version of the spatial Watts-Strogatz network. The network still appears to be random due to the random coordinates and the initial connections that are made without considering spatial distance similar to the nonspatial Watts-Strogatz graph and the Erdős-Rényi graph. For both of the spatial Watts-Strogatz graphs, when implementing Virginia Zip code data, because the nodes initially connect based on unique identifiers, which in this case is assigned ID numbers, and not location, the graph is very random and is very dark due to the long edges that stretch across the graph, which is shown in Figure 3b. No clustering can be identified in the graph.

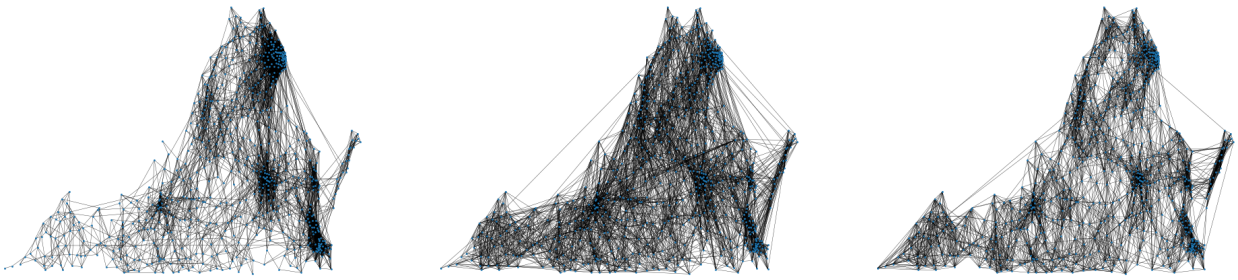
3 GEOSOCIAL SYNTHETIC SOCIAL NETWORK GENERATION

As discussed in Section 2, existing solutions for spatial synthetic network generation may lead to graphs that are very sparse, overly dense, or may not produce any graphs due to non-termination. The reason for these shortcoming is that the spatial models were designed for cases where the locations of nodes are placed in geographic space, but abstractly on a uniform grid. But using real-world locations such as locations of uses, buildings, ZIP codes or census tracts, we may get more realistic geosocial networks.

In this section we propose three new approaches to generating realistic synthetic geosocial networks by allowing to integrate real-world location data to represent the location of populations. Section 3.1 describes our geosocial variant of the classic Erdős-Rényi; Section 3.2 presents our geosocial variant of the classic Barabási-Albert Model; and Section 3.3 introduces our geosocial variant of the classic Watts-Strogatz model. In addition to the descriptions found in this section, the interested reader may find implementations and algorithms published on GitHub at https://github.com/KetevanGallagher/Synthetic-Geosocial-Networks. We do this not only to allow readers to replicate what we present here but to extend what we have presented here if they see fit to do so.

3.1 A Geosocial Erdős-Rényi Model

In the power law equation used in the spatial versions of the Erdős-Rényi Model (Section 2.2), each distance is raised to the power of negative alpha, which gives the weight for each possible connection. The spatial versions used random integer Cartesian coordinates as the locations for each node, so there was not a wide range of distances. However, when using longitude and latitude coordinates as the locations for each node, many of the distances between nodes are relatively small. This leads to very large probabilities of connections for those distances that are small, and very low probabilities of connection for all other distances, which made the power law equation ineffective. To combat this, a scaling factor was introduced. The scaling factor is chosen by raising the smallest desired distance to the power of -1. The distance between nodes is multiplied by the scaling factor, which leads to a power law equation as such: $p(d) = (sd)^{-\alpha}$. Additionally, distances that are below the smallest distance chosen are set to that distance, which solves an issue when using the code and power law equation from [1], which is if there is a wide range of distances, the weights for the larger distances could become very small when compared to the smaller distances, leading to a very sparse graph. In a more extreme example, if a distance is zero, all of the weights for the other distances will approach infinity and a network will not be generated. Figure 4 shows a graph of the new power law equation for three different alpha values, where the x-axis is the distance between nodes in meters and the y-axis is the probability of connection generated from the power law equation. Three different scaling factors were used for the three different alpha values, and were chosen in order to make networks with an average degree of 20 when used with Virginia ZIP code data. Because distances that are less than the distance used for the scaling factor are set to that distance, the power law equation results in a probability of one for all of those distances. Distances greater than the one chosen for the scaling factor will follow the power law equation and the probability will decrease with respect to the α value that is set. Figure 5a shows the geosocial Erdős-Rényi graph using Virginia ZIP code data, where each node is the centroid for a different ZIP code in Virginia. Upon visual inspection, we observe that this graph exhibits a certain similarity to the real-world data graph derived from Facebook Social Conectedness data depicted in Figure 1. A main difference is that sparse regions (such as in the west, on the left side of the network) have fewer connections and gaps in the real-world geosocial network that are caused by mountains are disregarded in this synthetic network. We note that the former shortcoming may be addressed by calibration of the distance-decay parameter α, and the later may be addressed by using a more sophisticated distance function that considers network distance (instead of Euclidean distance) between two nodes to account for barriers such as mountains and bodies of water. We also observe that this model yields a geosocial network that is much more similar to the real-world than the Spatial Erdős-Rényi model proposed in [1] which is depicted in Figure 3a.



(a) Geosocial Erdős-Rényi Network with Virginia ZIP code data

(b) Geosocial Barabási-Albert Network with Virginia ZIP code data

(c) Geosocial Watts-Strogatz Network with Virginia ZIP code data

Figure 5: Geosocial graphs using Virginia ZIP code data

3.2 A Geosocial Spatial Barabási-Albert Model

When utilizing real-world data to inform the Barabási-Albert network, the modified power law equation was also used, although with the addition of a variable k corresponding to the degree of the node to which a connection is made: $p(d) = k(sd)^{-\alpha}$. This gives nodes with a higher degree a higher probability of connecting to new nodes that are added to the network. In the Barabási-Albert network, nodes are added iteratively, so a node order was implemented to increase the likelihood that nodes will connect to those that are near them. This is the same approach used by the spatial Barabási-Albert network proposed in [1]. The difference proposed by our version of a spatial Barabási-Albert is the order in which nodes are processed. The model proposed in [1] adds nodes to the network in arbitrary order (without considering location of the nodes). This approach creates problems in the initial iterations, where only very few nodes have been added to the network, and thus, distances between these nodes may be very large. These large distances yield exponentially low (in these distances) probabilities of making connections, leading to exponentially large run-times due to repeating attempts to connect until one of the corresponding Bernoulli trials of making a connection succeeds. This methodology also creates unrealistically large edges (i.e., connecting nodes having a large distance) by forcing to connect randomly selected nodes.

Our approach initially starts with a seed node chosen randomly from the network (our experiments use the westernmost node). Then, nodes are added iteratively using the node having the shortest distance to the centroid of the nodes that have already been added. This process repeats until all nodes have been added to the list. Implementing this spatial node order substantially increases the run-time, as nodes are more likely to connect to those that are close to them and the number of nodes to connect to will be chosen faster if the nodes available are close in distance. Another change that was implemented to speed up the network generation was normalizing the weights given from the power law equation. By normalizing the probabilities of creating edges from a new node to all existing nodes, we avoid a large number of low-probability Bernoulli trials. Although the we cannot compare the run time of the spatial Barabási-Albert network to the geosocial Barabási-Albert as the spatial Barabási-Albert network does not terminate, we are able to compare the run time of the geosocial Barabási-Albert with and without the implementation of node order and normalizing distances. Over twenty trials (repetitions) it took an average of 103.1283 seconds to generate the geosocial Barabási-Albert graph without using the node order and normalizing distances. For the geosocial Barabási-Albert graph that used node distances and normalized distances, it took an average of 49.5471 seconds to generate the graphs. Figure 5b shows the spatial Barabási-Albert graph with Virginia ZIP code data. We again observe a certain similarity to the real-world graph in Figure 1. We observe more links in the west compared to the Erdős-Rényi model which is due to the (arbitrary) choice of using the westernmost node as the seed node, thus having nodes in the west added first and thus benefiting from the preferential attachment (by having a chance to connect to all other nodes).

3.3 A Geosocial Spatial Watts-Strogatz Model

In the spatial Watts-Strogatz network proposed in [1], nodes are initially connected to their nearest neighbors based on arbitrary order (without considering spatial information). This leads to much spatial randomness in the network, and in fact both spatial and non-spatial Watts-Strogatz networks look very similar to the (non-spatial) Erdős-Rényi network because spatial information is not used in the initialization of the nearest-neighbor lattice. The approach in [1] uses spatial information only for the rewiring.

Thus, in our version of the spatial Watts-Strogatz model, the nearest neighbors that a node initially connects to are defined as the nodes that have the shortest distances to that node. Figure 5c shows this updated version of the spatial Watts-Strogatz network, which also implements the same power law as the geosocial Erdős-Rényi network. Compared to the network generated by the spatial Watts-Strogatz model proposed in [1] and depicted in Figure 3b we observe much more structure in this graph. There is also visible clustering, which is not seen in the spatial Watts-Strogatz graph. The edges are much shorter and an outline of the shape of Virginia can be seen. The resulting network also visually resembles the real-world geosocial network shown in Figure 1.

	Ground Truth	Classic ER	Classic BA	Classic WS	Spatial ER	Spatial WS	Geosocial ER	Geosocial BA	Geosocial WS	KNN
Avg. Degree	19.59	19.78	19.84	19.77	2.94	20.00	20.05	19.73	20.00	20.00
Std. Dev. Degree	3.40	3.83	18.49	2.70	4.29	1.37	14.70	5.06	4.36	4.50
Max. Degree	32.00	32.60	163.25	29.40	22.55	25.00	69.40	37.50	33.90	34.00
Radius of Gyration	55292	513022	489469	509785	12539	510990	87529	133198	68856	37152
Std. Dev Radius of Gyration	51072	132646	134790	132887	23579	130288	67073	107183	64809	15330
Avg. Length of Edge	20934	206714	208432	201354	7739	188125	18440	29880	18792	16087
Std. Dev. Length of Edge	21175	166824	161757	168058	14522	167691	25926	45530	22785	12822
# Triangles	26357	1251	6742	4717	1258	22709	38421	13283	23522	27596
Jaccard Index	1	0.0078	0.0118	0.0136	0.1167	0.0266	0.3615	0.1660	0.2367	0.2560

Table 1: Statistics for graphs using Virginia ZIP code data

	Ground Truth	Classic ER	Classic BA	Classic WS	Spatial ER	Spatial WS	Geosocial ER	Geosocial BA	Geosocial WS	KNN
Avg. Degree	13.58	13.62	13.77	13.69	12.48	14.00	13.99	13.63	14.00	14.00
Std. Dev. Degree	8.69	3.12	11.52	2.24	4.68	1.35	4.96	4.07	3.11	3.16
Max. Degree	72.00	22.95	82.75	20.65	24.45	18.35	26.95	26.30	22.70	22.00
Radius of Gyration	11836	29287	29047	26435	9049	28719	9675	12637	7434	4401
Std. Dev Radius of Gyration	6935	6843	8267	7847	5597	7770	5505	8532	4814	970
Avg. Length of Edge	4109	13852	14036	9506	2731	12338	2842	3594	2462	2068
Std. Dev. Length of Edge	3968	9930	10414	9008	3018	9951	3049	4587	2345	1333
# Triangles	3074	401	1606	900	2491	3251	3176	2075	3687	4597
Jaccard Index	1	0.0163	0.0317	0.0643	0.2814	0.0295	0.2861	0.1179	0.1880	0.2038

Table 2: Statistcs for graphs using Fairfax Census Tract data

4 EXPERIMENTAL EVALUATION

This section provides qualitative and quantitative experimental evaluations showing that our three proposed synthetic geosocial network generators are able to produce networks that are more similar to real-world networks than existing solutions. In Section 4.1 we first provide additional details on two real-world datasets that were used for this study: Human mobility data and social connectedness data. Then, Section 4.2 provides a more detailed qualitative evaluation of visualizations of the generated geosocial networks and Section 4.3 provides a quantitative evaluation by comparing network measures and statistics between the generated geosocial networks and the real-world geosocial networks.

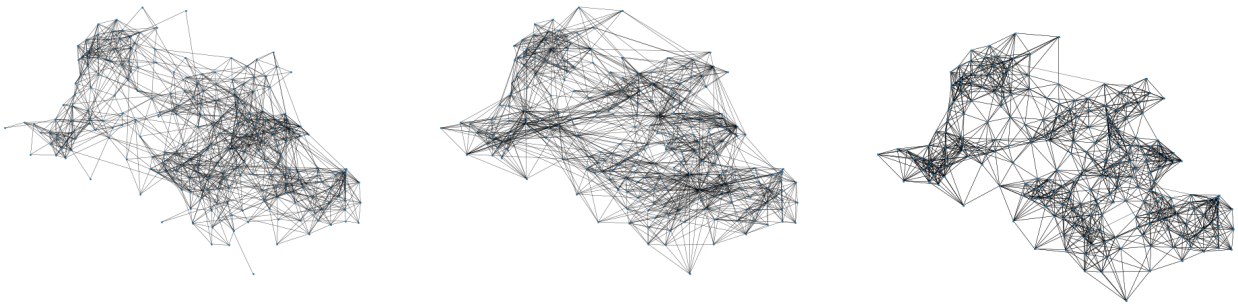
4.1 Datasets

To evaluate the similarity between generated geosocial networks to real-world data, and to show that our proposed geosocial networks can be used as a proxy in lieu of real-world data, we used two different types of real-world geosocial network data. Both datasets delineate geographic space into administrative units that contain a population. The first dataset delineates space into census tracts for the study region of Fairfax County, VA, USA. Each census tract can be abstracted to a node in the network where the node's spatial properties, latitude and longitude, correspond to the census tract's center of population. To create the ground truth graph depicted in 6b, we used human mobility data from SafeGraph [17], and connected each census tract to the seven other census tract for which it had the highest estimated population flows between for the date 1/4/2020 to estimate geosocial connections between census tracts in Fairfax county. This resulted in an average degree of 14. Thus, the geospatial network derived from this dataset is composed of nodes corresponding to the population centroid of census tracts and edges that correspond to a high estimated population flow between the nodes.

The second dataset delineates geographic space based on ZIP codes for Virginia USA. We leverage data provided by Facebook Data for Good, referred to as the Social Connectedness Index [4]. The dataset provides a measure of social connectness between all pairs of ZIP codes. We connect each ZIP code to the ten other ZIP codes with which it had the highest Social Connectedness Index. The resulting network is depicted in Figure 1, and has an average degree of 20.

4.2 Qualitative Evaluation

In this section we qualitatively/visually compare our proposed synthetic geosocial network generation models to both the classical (non-spatial) social network models and the existing spatial network models. To qualitatively evaluate the graphs we generate all networks in a way that they have a similar average degree. For graphs using the Virginia ZIP code data the average degree was set to 20. An average degree of 20 was chosen because it created graphs that were dense enough that clustering can be observed. For graphs generated using the Fairfax Census Tract data, the average degree was chosen to be 14. Similarly, the average degree of 14 was chosen



(a) Geosocial Erdős-Rényi graph with Fairfax Census Tract data **(b) Graph generated from mobility data from [17]** **(c) KNN graph with Fairfax Census Tract data**

Figure 6: Graphs using Fairfax Census Tract data

so that clustering could be seen in the figures that are included. For the spatial and geosocial graphs, the alpha value was set to $\alpha = 3$ so all the graphs would have a similar degree of clustering. This value was chosen so that graphs would have more clustering, as lower alpha values lead to graphs that are more random. Much like the geosocial Erdős-Rényi graph in Figure 5a, the graph generated from Facebook data in Figure 1 has regions that appear dark, where clusters of nodes are densely connected. However, while both graphs are visually dark in the northeast region of the graph, we visually observe that the southwest region is more connected in the graph generated from the Facebook data than the geosocial Erdős-Rényi graph in figure 5a is. Additionally, there is a section in the northern region of the Facebook graph that has very few connections across this area and is very sparse. This is slightly replicated in the geosocial Watts-Strogatz graph in figure 5c, but the same area in the geosocial Watts-Strogatz graph is not as long or as sparsely connected as it is in the Facebook graph. None of the other graphs replicate this area. We believe that it may be caused by the Appalachian mountains and Shenandoah National Park, which are in this area. Because the area is mountainous and heavily wooded due to it being a national park, it is sparsely populated and difficult to travel through. Thus, people who live on either side of this region would not interact much, and there are very few connections across this region. All of the graphs generated from the classical social network models and the spatial Watts-Strogatz graph are very dark and have long edges that reach across the graph. This is because the distance between nodes are not taken into account when the graph is being generated, so they look very random. The spatial Erdős-Rényi graph has some clusters, but the clusters in the geosocial Erdős-Rényi graph are more defined. The graph generated from the mobility data in [17] has dark clusters where nodes are highly connected, but also has some long edges that stretch across the graph. This is similar to the geosocial Barabassi-Albert graph, where there are some clusters but also long edges that are connected to nodes with a high degree. The geosocial Watts-Strogatz graph looks very similar to the KNN graph which connects each node to exactly their k (spatially) nearest neighbors, which makes sense as the Watts-Strogatz graph starts as a KNN graph but becomes more random as nodes are rewired. Clustering can also be seen in the geosocial Erdős-Rényi graph, in Figure 6a, when it is used with Fairfax Census Tract data. When compared to the graph generated from mobility data from [17] in Figure 6b, clustering can be seen in many of the same places. The KNN graph generated from Fairfax Census Tract data in Figure 6c also has similar clustering when compared to these graphs, but the geosocial Erdős-Rényi graph has some longer random edges that make it more similar to the graph generated from the mobility data.

4.3 Quantitative Evaluation

For our quantitative evaluation, we used the same setting having $\alpha = 3.0$ and set the average degree so that all of the graphs using the same location data have a similar number of edges and could be compared to each other. The exception to this is the Erdős-Rényi model, which does not have a parameter to change the density of the graph, and thus has a lower degree than the other graphs using the same datasets. The spatial Barabási-Albert graph is not included in Tables 1 and 2 because it did not terminate. For the Watts-Strogatz graphs, probability that nodes would rewire was set to the inverse of half of the desired average degree, so that each node had the probability of having one of its edges rewired. Each statistic in the table is the average of twenty trials. The average radius of gyration (the maximum distance of a node to its neighbors) [14], standard deviation of radius of gyration, average distance between connected nodes (labeled average length of edge in 1 and 2), and standard deviation of the distance between connected edges (labeled standard deviation of length of edge in 1 and 2) are all measured in meters. The radius of gyration was calculated by taking the average of the maximum length in meters between two connected nodes for each node in the graph. For the graphs generated with Virginia ZIP code data, all of the average degrees are around 20, except for that of the spatial Erdős-Rényi graph. The standard deviation of degree of the classical Barabási-Albert graph is very high relative to the other graphs, which is to be expected as nodes that have many connections are likely to gain more connections. We believe the reason the standard deviation of the degree in the geosocial Barabási-Albert graph is not as high is because the node order begins in the southwest of the graph, where nodes are more sparse, and thus the two components that decide if a node becomes connected to another work against each other. Because the node order begins in the southwest, those nodes are likely to gain connections

because they already have many connections, but nodes that are not well connected but that are closer to another nodes are also likely to gain connections. This leads to a lower standard deviation of degree. One might expect that the standard deviation of degree for the KNN graph should be zero, as each node has ten outgoing edges. However, the number of incoming edges (which is the number of reverse nearest neighbors) is not constant and may vary depending on the local density of a node. Thus, some nodes may have more or less than 20 neighbors. We also measured the maximum degree which, to be realistic, should not exceed a reasonable number [8, 13]. The geosocial Erdős-Rényi graph had a very high maximum degree, most likely due to dense clusters where many nodes are very close to each other. The radius of gyration for the classical graphs is very high, as these graphs have long edges that reach across the graph. Thus the average distance between connected nodes is also very high for these graphs. The graph that had a radius of gyration most similar to that of the graph generated with Facebook data was the geosocial Erdős-Rényi graph. When using Virginia ZIP code data, the geosocial Erdős-Rényi graph overestimated the radius of gyration, but when Fairfax Census Tract data was used, the geosocial Erdős-Rényi graph underestimated the radius of gyration. All of the classic graphs and the spatial Watts-Strogatz graphs had very long distances between connected nodes as all of these graphs are quite random and have long edges that stretch across the graph. The geosocial Watts-Strogatz graph had the average edge length that was closest to the ground truth, but the geosocial Erdős-Rényi graph was also close to the ground truth. The KNN graph had a very low radius of gyration and average edge length, which makes sense because by definition a KNN graph has the smallest total length of edges possible for a set degree. The geosocial Erdős-Rényi graph and the geosocial Barabási-Albert graph had a higher and lower number of triangles than the ground truth, while both spatial Watts-Strogatz and the KNN network had similar numbers to the ground truth, indicating a similar level of clustering. Because all of these graphs have the same nodes, we were able to measure the Jaccard index of the edges for these graphs. The graph with the highest Jaccard index was the geosocial Erdős-Rényi graph. All of the geosocial graphs and the KNN network had higher Jaccard indexes than the other graphs. The spatial Erdős-Rényi graph did not have an especially low Jaccard index, but the average degree of this graph is so low that it cannot be compared to the other graphs. For the graphs generated with Fairfax Census Tract data, all of the graphs had an average degree of around 14, except for the Spatial Erdős-Rényi graph, which is due to the fact the average degree of the graph cannot be changed without changing the alpha value. The mobility graph had a very high maximum degree, which surpassed all the other graphs except for the classic Barabási-Albert graph. The geosocial Barabási-Albert and Erdős-Rényi graphs had the radius of gyration closest to that of the graph generated from the mobility data. These graphs also had the closest average length of connected edges to the graph generated from the mobility data. The geosocial Erdős-Rényi graph had the closest number of triangles to the ground truth. The geosocial Erdős-Rényi graph had the highest Jaccard index. The spatial Erdős-Rényi graph had the second highest Jaccard index, but that may be due to the fact that the average degree in that graph is lower than the others.

5 CONCLUSIONS, AND FUTURE WORK

As we noted in at the start of the paper, generating synthetic social networks is an important task for many problems that study humans, their behavior, and their interactions. However, until now many models used to generate such networks do not consider location when making new links and those that do assume a uniform distribution of the population which we would argue is not often the case in reality. To overcome this issue, we have proposed geosocial extensions to three widely used network models (i.e., Erdős-Rényi, Watts-Strogatz, and Barabási-Albert) which specifically includes location information in the social link generation process. Not only do we describe implementations of these networks (Section 3), but our results look similar to what one finds in reality. For example, our updated geosocial Erdős-Rényi graph looks more similar to the ground truth graphs generated from data from Facebook and human mobility data. The features, such as radius of gyration and average distance between connected nodes, were more similar to those of the ground truth when compared against other graphs which have the same nodes and a similar number of edges. The geosocial Erdős-Rényi graph also had the highest Jaccard index when compared with both ground truth graphs.

However, while the graph generated from Facebook data has an area with very few connections across it in the middle of the northern region, none of the generated geosocial graphs have this sparse region. We hypothesize that the reason for this sparse area is the Appalachian mountains, which are forested, difficult to cross and have a low population count. Future work needs to explore how instead of using the flat Euclidean distances between nodes, using the actual topographical distances may lead to increased accuracy. Another area of further work is to apply our geosocial methods to other areas in order to see what other differences emerge and explore why this might be the case. With this being said, this paper paves the way for creating a new way to generate synthetic geosocial networks that consider individuals' social relationships and their locations.

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] M. Alizadeh, C. Cioffi-Revilla, and A. Crooks. Generating and analyzing spatial social networks. *Computational and Mathematical Organization Theory*, 23:362–390, 2017.
[2] L. Anselin. What is special about spatial data? alternative perspectives on spatial data analysis (89-4). 1989.
[3] L. Anselin. Thirty years of spatial econometrics. *Papers in regional science*, 89(1):3–25, 2010.
[4] M. Bailey, R. Cao, T. Kuchler, J. Stroebel, and A. Wong. Social connectedness: Measurement, determinants, and effects. *Journal of Economic Perspectives*, 32(3):259–280, 2018.
[5] J. Bao, Y. Zheng, and M. F. Mokbel. Location-based and preference-aware recommendation using sparse geo-social networking data. In *Proceedings of the 20th international conference on advances in geographic information systems*, pages 199–208, 2012.
[6] A.-L. Barabási. Scale-free networks: a decade and beyond. *science*, 325(5939):412–413, 2009.

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

[8] M. Barthélemy. Crossover from scale-free to spatial networks. *Europhysics letters*, 63(6):915, 2003.

[9] P. Erdős, A. Rényi, et al. On the evolution of random graphs. *Publ. math. inst. hung. acad. sci*, 5(1):17–60, 1960.

[10] M. Fire, D. Kagan, R. Puzis, L. Rokach, and Y. Elovici. Data mining opportunities in geosocial networks for improving road safety. In *2012 IEEE 27th Convention of Electrical and Electronics Engineers in Israel*, pages 1–4. IEEE, 2012.

[11] S. R. Friedman and S. Aral. Social networks, risk-potential networks, health, and disease. *Journal of Urban Health*, 78:411–418, 2001.

[12] K. Gallagher, S. Kotnana, S. Satishkumar, K. Siripurapu, J. Elarde, T. Anderson, A. Züfle, and H. Kavak. Human mobility-based synthetic social network generation. In *Proceedings of the 2nd ACM SIGSPATIAL International Workshop on Animal Movement Ecology and Human Mobility*, pages 23–26, 2022.

[13] L. Hamill and G. Gilbert. Social circles: A simple structure for agent-based social network models. *Journal of Artificial Societies and Social Simulation*, 12(2), 2009.

[14] A. Hernando, D. Mateo, J. Bayer, and I. Barrios. Radius of gyration as predictor of covid-19 deaths trend with three-weeks offset. *medRxiv*, pages 2021–01, 2021.

[15] A. L. Hill, D. G. Rand, M. A. Nowak, and N. A. Christakis. Infectious disease modeling of social contagion in networks. *PLOS computational biology*, 6(11):e1000968, 2010.

[16] L. Humphreys. Mobile social networks and urban public space. *New Media & Society*, 12(5):763–778, 2010.

[17] Y. Kang, S. Gao, Y. Liang, M. Li, J. Rao, and J. Kruse. Multiscale dynamic human mobility flow dataset in the us during the covid-19 epidemic. *Scientific data*, 7(1):390, 2020.

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] M. E. Newman. Clustering and preferential attachment in growing networks. *Physical review E*, 64(2):025102, 2001.

[23] I. P. Tussyadiah. A concept of location-based social network marketing. *Journal of Travel & Tourism Marketing*, 29(3):205–220, 2012.

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

[25] L. H. Wong, P. Pattison, and G. Robins. A spatial model for social networks. *Physica A: Statistical Mechanics and its Applications*, 360(1):99–120, 2006.

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