Towards Generating Realistic Geosocial Networks

Abed Al Rhman Sarsour
Institute of Computer Science
Johannes Gutenberg University Mainz
Germany
asarsour@students.uni-mainz.de

Panagiotis Bouros
Institute of Computer Science
Johannes Gutenberg University Mainz
Germany
bouros@uni-mainz.de

Theodoros Chondrogiannis
Department of Computer and Information Science
University of Konstanz
Germany
theodoros.chondrogiannis@uni.kn

ABSTRACT

The proliferation of location-based services and social networks have given rise to geosocial networks, which model not only the social interactions between users but also their spatial activities. Examples include traditional social networks extended with geo-annotated posts such as Twitter and Facebook, and networks such as Foursquare and Yelp that directly offer geosocial services. Despite the ubiquity of such networks in everyday life and the strong interest by the research community, a limited number of datasets are in fact publicly available. In view of this, we investigate the generation of realistic geosocial networks which find application in benchmarking and testing of analysis tasks, "what-if" scenarios and simulations. The contributions of our work are twofold. We first identify three types of synthetic geosocial networks which mimic the characteristics of real ones and second, we develop a prototype which combines graph and spatial generators, to construct such networks.

CCS CONCEPTS

• **Information systems** → **Social networks**; *Geographic information systems.*

KEYWORDS

Geosocial network, graph, spatial data, generator

ACM Reference Format:
Abed Al Rhman Sarsour, Panagiotis Bouros, and Theodoros Chondrogiannis. 2023. Towards Generating Realistic Geosocial Networks. 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, 5 pages. https://doi.org/10.1145/3615896.3628340

1 INTRODUCTION

The ubiquity of mobile location-aware devices (smart phones and watches, tablets etc.) and the proliferation of social networks have given rise to *geosocial networks*, where users not only form social connections to each other but also perform geo-referenced actions, e.g., posts and check-ins. Examples of such networks include traditional social networks extended with geospatial information such as X (formerly known as Twitter) and Facebook, and networks directly offering geosocial services such as Yelp and Foursquare. Research in geosocial networks has focused on query processing [4, 10, 11, 23, 25–27] and indexing [28, 29], on the collision with Recommender systems [19, 24] and on analysis tasks such as influence maximization [7, 20] and community search [8, 12, 13].

Despite the volume of research in geosocial networks, a limited number of datasets are publicly available for downloading. For example, Yelp[1] offers an official dump for academic purposes, a Brightkite and a Gowalla dump can be found in SNAP's page[2], while a Foursquare[3] dataset is available from [19, 24]. Another option for acquiring such datasets is to use official APIs offered by some geosocial networks, e.g., X[4] and Foursquare[5]. However, these APIs typically restrict the number of queries per day, and therefore, the amount of data to retrieve; for unlimited downloads, fees are charged. In an attempt to fill this availability gap, generated geosocial networks can be used instead, e.g., for benchmarking the efficiency and the robustness of geosocial queries, for hypothesis testing, "what-if" scenarios, and simulations. For example, an extensive performance comparison of the algorithms can be conducted by studying the impact of parameters/factors such as the network size, its topology and the distribution in space.

Existing generators. Network generation has received significant attention in the graph literature. The goal of all proposed models is to generate synthetic networks whose properties match the ones observed in real networks, as well as possible. Real-world social networks in specific, are typically characterized by a vertex-degree distribution that follows a power law; i.e., the number of vertices c_k with degree k is given by the formula $c_k \propto k^{-\gamma}$ where $\gamma > 0$ is called the power-law exponent. In addition, social networks exhibit a small diameter, a.k.a. the "small-world" phenomenon, or "six degrees of separation" [22]. Specifically, a diameter d indicates that every pair of vertices can be connected by a path that contains at most d edges. In this context, the majority of proposed models [1, 2, 16, 17, 32] involve some form of preferential attachment to progressively construct a synthetic network. Under this, the new vertices added are preferentially connected to existing vertices with high degree, adopting a "rich get richer" approach. A different family of models such as the small-world model [31] strives for small diameter, while the Kronecker graph model [18] employs a recursive construction to create self-similar networks, starting from a small initial one as the basis.

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

ACM ISBN 979-8-4007-0358-4/23/11…$15.00
https://doi.org/10.1145/3615896.3628340

[1] https://www.yelp.com/dataset/
[2] http://snap.stanford.edu/data/index.html#locnet
[3] https://archive.org/details/201309_foursquare_dataset_umn
[4] https://developer.twitter.com/en/products/twitter-api
[5] https://location.foursquare.com/developer/reference/places-api-overview

Generating spatial data has been also studied in the past. Beckmann and Seeger presented a generator for benchmarking multi-dimensional indices in [5], which can be used to produce spatial points or rectangles following different distributions in space, e.g., uniform, clustered, diagonal etc. Recently, this generator was further extended in [30] and the Spider Web-based interface[6] was presented in [15].

Contributions. Despite the efforts on generating social networks and spatial data, to the best of our knowledge, the process of generating realistic *geosocial* networks has not be investigated. To fill this gap, our paper first discusses three types of synthetic geosocial networks that mimic the characteristics of real networks. Then, we describe the process of generating such networks and discuss our open-source prototype in Python which combines graph generators offered by the NetworkX[7] graph library with the spatial generator in [15, 30].

2 SYNTHETIC GEOSOCIAL NETWORKS

We first introduce necessary notation and then describe three types of realistic synthetic geosocial networks.

2.1 Notation

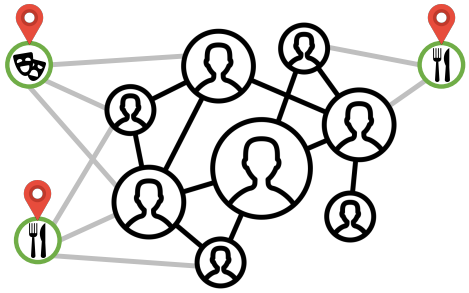
We model a *social* network as a graph $G = (V, E)$ where every vertex $v \in V$ represents an entity of the network and every edge $(u, v) \in E \subseteq V \times V$ indicates a relationship between the entities modelled by vertices u and v. The edges in E can be either *directed* or *undirected* depending on the application and the nature of the modelled relationships between the graph vertices; without loss of generality, we draw only undirected edges in the rest of the text. A *geosocial* network is a social network where a vertex v can be associated with the geometry of an object in the two or three-dimensional space, denoted by $v.geom$, e.g., a point, a polygon etc. For simplicity, we call such v, a *spatial vertex*.

2.2 Types

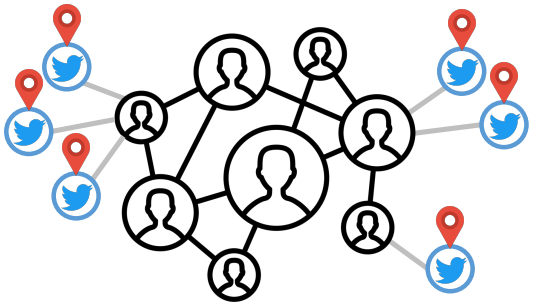
In the first type of synthetic geosocial networks, denoted by G_s, all vertices in set V represent the same type of entities, e.g., users of the network, and the edges in set E model relationships between such entities, e.g., *FRIEND_OF*, *FOLLOWS* etc. Geospatial information assigned to (spatial) vertices stores specific location such as a person's workplace or residence. As an example, consider the scenario of an academic geosocial network created based on the co-authorship relationship. Every vertex on the network represents a researcher and an edge is defined for each pair of researchers who have co-authored a publication. The $v.geom$ of a spatial vertex v stores the coordinates for the current affiliation of the corresponding researcher.[8] Figure 1(a) exemplifies a type G_s geosocial network; spatial vertices are marked with a red pin.

The second type of synthetic geosocial networks, denoted by G_c, models two types of entities as vertices. Intuitively, the graph of the network contains a social core of non-spatial vertices representing for instance network users; the edges connecting these vertices

(a) Type G_s



(b) Type G_c



(c) Type G_p

Figure 1: Types of synthetic geosocial networks: non-spatial vertices and edges in between them are drawn in black, spatial vertices are marked with a red pin.

model again typical social relationships. In addition, G_c contains a set of vertices associated with geospatial information. These spatial vertices do not represent users and they are never connected to each other; instead, they are always connected to one or more vertices of the social core. As an example of the G_c type, consider Foursquare. The social core of its network graph represents registered users who are connected to each other via the *FRIEND_OF* relationship. The graph also contains a set of spatial vertices modelling the location of venues, businesses, attractions and other points of interest. Every such spatial vertex is connected to one or more non-spatial vertices via a *CHECK_IN* relationship, to indicate the users having visited the corresponding location. Figure 1(b) shows an example of a G_c network, inspired by Foursquare. Notice the social core of the network drawn in black color, which includes the non-spatial vertices, representing users, and the edges between them, capturing friendship relationships. On the other hand, the spatial vertices, drawn in green, model attractions; specifically, two restaurants and a theatre house. Last, the gray edges model the *CHECK_IN* activity

[6] https://spider.cs.ucr.edu
[7] https://networkx.org
[8] It is possible that only a subset of vertices V are spatial as information about the affiliation of some researchers may not be available.

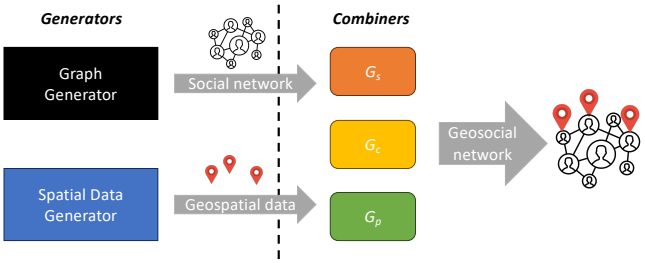


Figure 2: Generation process

of the users to these attractions; as expected, a location can be visited by multiple users.

The third type of synthetic geosocial networks, denoted by G_p, also models two types of entities. Similar to G_c, the network contains a social core of non-spatial vertices with typical social relationships. However, every spatial vertex in set V is exclusively connected to a single non-spatial vertex. As an example, consider X or Facebook. Non-spatial vertices represent registered users, connected via a *FOLLOW* or a *FRIEND_OF* relationship, respectively. Spatial vertices on the other hand, represent geo-annotated posts (namely, tweets in case of X); this geospatial information can be either explicitly provided by users via a mobile application on their smartphone or even automatically extracted from the text through a process known as geo-tagging [3, 9, 21]. Figure 1(c) exemplifies type G_p geosocial networks. We observe similar to the G_c case, the social core of the network, drawn in black. The spatial vertices drawn in blue, represent geo-annotated tweets. However, this time, every spatial vertex is always connected to exactly one user vertex, i.e., the author of the corresponding post. Naturally, as users can post multiple geo-annotated posts, a non-spatial vertex can be connected to multiple spatial ones.

3 GENERATION PROCESS

We next elaborate on the process of generating realistic synthetic geosocial networks and discuss our prototype generator.[9]

3.1 Overview

Figure 2 illustrates the generation process. Essentially, the process comprises two phases. First, a graph and a spatial generator are independently employed to create a synthetic social network and a collection of geospatial objects, respectively. In the second phase, these intermediate datasets are then combined to construct a geosocial network. A combiner is defined for each type of synthetic geosocial network, i.e., G_s, G_c and G_p.

We developed a prototype in Python that implements the above generation process. The system is designed with modularity and extensibility in mind. The output datasets of the generators in Figure 2 are stored inside a .gr file (for the graph of the social network) and a .co file (for the geospatial data), which are then scanned by one of the combiners to produce the final geosocial network (also outputted inside a .gr and a .co file). This design enables us to consider different models, implementations and libraries for each component without the need to reimplement the entire workflow, as long as the output files of the generators comply with the .gr and .co format. In addition, it is also possible for the user to provide an existing (potentially real) social network graph and/or a collection of spatial objects as input(s) instead of using the generators, and then directly employ one of the combiners to construct the final geosocial network.

Without loss of generality, our prototype currently uses the NetworkX graph library (also in Python) to generate the graph of a social network. Specifically, we use:

- the barabasi_albert_graph function which constructs random graphs according to the preferential attachment model in [1],
- the scale_free_graph function which constructs scale-free graphs according to the model proposed in [6], and
- the powerlaw_cluster_graph function which constructs graphs according to the model in [14], with power law degree distribution and approximate average clustering

For generating geospatial data, we use the generator from [15, 30], which is also implemented in Python.[10]

3.2 Implementation

Last, we discuss the implementation of our prototype. For simplicity, we focus on the case when both the social graph and the geospatial data are generated during the first phase, and not uploaded by the user. The system receives as input the number of vertices of the social network graph and the number of geometries, which also equals the number of spatial vertices to be created in the output geosocial graph. The remaining parameters for the first phase depend on the model to used by the graph generator and on the distribution and the geometry type for the spatial generator; both are specified by the user.[11] Depending on the graph generation model, the user can request a directed or an undirected social network graph.

We next elaborate on the combiners. All three receive as inputs a .gr file which contains the generated social graph and a .co file, which contains the geometries for the spatial vertices. Type G_c and G_p require extra inputs which we will discuss in the following. We start with the G_s combiner, which is the simplest of the three. Essentially, the output geosocial network contains exactly the same vertices and edges as the generated social graph in the previous phase. Hence, the combiner randomly selects a subset of the social vertices to become spatial, by assigning them a geometry from the .co file. These vertex-to-geometry assignments are stored in the output .co file. Note that the output .gr file is identical to the input.

Unlike G_s, the G_c and G_p combiners will extend the set of vertices and edges of the network graph. Both will create and add new vertices, one for each geometry found in the input .co file, and new edges to connect these spatial vertices to the social ones found in the input .gr file. The key difference between the two combiners lies on how these new edges are created. The G_c combiner will connect every created spatial vertex to one or more social. We assume that the number of edges created for each spatial vertex follows a normal distribution; the user can specify the mean value

[9] Code for our generator can be found in https://github.com/pbour/geosocialgenerator.

[10] Code available in https://github.com/aseldawy/spider and https://github.com/tinvukhac/spatialdatagenerators

[11] For more information about the input parameters in each case, please refer to NetworkX documentation https://networkx.org/documentation/stable/ and the https://github.com/aseldawy/spider, https://github.com/tinvukhac/spatialdatagenerators repositories.

and standard deviation for the distribution. In contrast, the G_p combiner will connect every created spatial vertex to exactly one social. In this case, we consider a normal distribution for the number of edges towards spatial vertices, every social vertex can have; for example, the number of posts a user made in the network. The parameters for the distribution can also be provided by the user.

4 CONCLUSIONS

In this paper, we studied the generation of realistic geosocial networks. We described three types of synthetic networks and presented a prototype for the generation process. Our prototype capitalizes on the NetworkX graph library to generate the network graph and on a spatial data generator.

In the future, we plan to extend our work towards multiple directions. First, we will investigate additional types of synthetic geosocial networks, able to cover more application scenarios. We will also study how realistic are the generated networks. For this purpose, similarity measures for comparing generated networks to a repository of existing real geosocial networks can be employed. Further, we plan to include more models for generating the social network graph and to offer more features for the graph edges, e.g., supporting multiple types of relationships via labeling, and weights to capture the strength of a relationship Finally, we also intend to develop an interactive Web-based user interface, able to both generate and visualize geosocial networks.

ACKNOWLEDGMENTS

This work is partially supported by Grant No. CH 2464/1-1 of the Deutsche Forschungsgemeinschaft (DFG). This work is based on the BSc thesis of Abed Al Rhman Sarsour at Johannes Gutenberg University Mainz, Germany.

REFERENCES

[1] Réka Albert and Albert-László Barabási. 1999. Emergence of Scaling in Random Networks. *Science* 286 (1999), 509–512.
[2] Réka Albert and Albert-László Barabási. 2002. Statistical mechanics of complex networks. *Rev. Mod. Phys.* 74 (2002), 47–97. Issue 1.
[3] Einat Amitay, Nadav Har'El, Ron Sivan, and Aya Soffer. 2004. Web-a-where: geotagging web content. In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Sheffield, UK, July 25-29, 2004*. ACM, 273–280.
[4] Nikos Armenatzoglou, Stavros Papadopoulos, and Dimitris Papadias. 2013. A General Framework for Geo-Social Query Processing. *Proc. VLDB Endow.* 6, 10 (2013), 913–924.
[5] Norbert Beckmann and Bernhard Seeger. 2008. *A Benchmark for Multidimensional Index Structures.* Technical Report. Philipps-Universität Marburg. https://www.mathematik.uni-marburg.de/ rstar/benchmark/distributions.pdf.
[6] Béla Bollobás, Christian Borgs, Jennifer T. Chayes, and Oliver Riordan. 2003. Directed scale-free graphs. In *Proceedings of the Fourteenth Annual ACM-SIAM Symposium on Discrete Algorithms, January 12-14, 2003, Baltimore, Maryland, USA*. 132–139.
[7] Panagiotis Bouros, Dimitris Sacharidis, and Nikos Bikakis. 2014. Regionally influential users in location-aware social networks. In *Proceedings of the 22nd ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, Dallas/Fort Worth, TX, USA, November 4-7, 2014*. 501–504.
[8] Lu Chen, Chengfei Liu, Rui Zhou, Jianxin Li, Xiaochun Yang, and Bin Wang. 2018. Maximum Co-located Community Search in Large Scale Social Networks. *Proc. VLDB Endow.* 11, 10 (2018), 1233–1246.
[9] Junyan Ding, Luis Gravano, and Narayanan Shivakumar. 2000. Computing Geographical Scopes of Web Resources. In *Proceedings of 26th International Conference on Very Large Data Bases, VLDB 2000, September 10-14, 2000, Cairo, Egypt*. 545–556.
[10] Yerach Doytsher, Ben Galon, and Yaron Kanza. 2010. Querying geo-social data by bridging spatial networks and social networks. In *Proceedings of the 2010 International Workshop on Location Based Social Networks, LBSN 2010, November 2, 2010, San Jose, CA, USA*. 39–46.
[11] Yerach Doytsher, Ben Galon, and Yaron Kanza. 2012. Querying socio-spatial networks on the world-wide web. In *Proceedings of the 21st World Wide Web Conference, WWW 2012, Lyon, France, April 16-20, 2012 (Companion Volume)*. 329–332.
[12] Yixiang Fang, Reynold Cheng, Xiaodong Li, Siqiang Luo, and Jiafeng Hu. 2017. Effective Community Search over Large Spatial Graphs. *Proc. VLDB Endow.* 10, 6 (2017), 709–720.
[13] Yixiang Fang, Zheng Wang, Reynold Cheng, Xiaodong Li, Siqiang Luo, Jiafeng Hu, and Xiaojun Chen. 2019. On Spatial-Aware Community Search. *IEEE Trans. Knowl. Data Eng.* 31, 4 (2019), 783–798.
[14] Petter Holme and Beom Jun Kim. 2002. Growing scale-free networks with tunable clustering. *Phys. Rev. E* 65 (Jan 2002), 026107. Issue 2.
[15] Puloma Katiyar, Tin Vu, Ahmed Eldawy, Sara Migliorini, and Alberto Belussi. 2020. SpiderWeb: A Spatial Data Generator on the Web. In *Proceedings of the 28th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, Seattle, WA, USA, November 3-6, 2020*. 465–468.
[16] Jon M. Kleinberg, Ravi Kumar, Prabhakar Raghavan, Sridhar Rajagopalan, and Andrew Tomkins. 1999. The Web as a Graph: Measurements, Models, and Methods. In *Computing and Combinatorics, Proceedings 5th Annual International Conference, COCOON 1999, Tokyo, Japan, July 26-28, 1999*. 1–17.
[17] Ravi Kumar, Prabhakar Raghavan, Sridhar Rajagopalan, and Andrew Tomkins. 1999. Extracting Large-Scale Knowledge Bases from the Web. In *VLDB'99, Proceedings of 25th International Conference on Very Large Data Bases, September 7-10, 1999, Edinburgh, Scotland, UK*. 639–650.
[18] Jure Leskovec, Deepayan Chakrabarti, Jon M. Kleinberg, and Christos Faloutsos. 2005. Realistic, Mathematically Tractable Graph Generation and Evolution, Using Kronecker Multiplication. In *Proceedings of the 9th European Conference on Principles and Practice of Knowledge Discovery in Databases, PKDD 2005, Porto, Portugal, October 3-7, 2005*. 133–145.
[19] Justin J. Levandoski, Mohamed Sarwat, Ahmed Eldawy, and Mohamed F. Mokbel. 2012. LARS: A Location-Aware Recommender System. In *Proceedings of the 28th IEEE International Conference on Data Engineering (ICDE 2012), Washington, DC, USA (Arlington, Virginia), 1-5 April, 2012*, Anastasios Kementsietsidis and Marcos Antonio Vaz Salles (Eds.). 450–461.
[20] Guoliang Li, Shuo Chen, Jianhua Feng, Kian-Lee Tan, and Wen-Syan Li. 2014. Efficient location-aware influence maximization. In *Proceedings of the International Conference on Management of Data, ACM SIGMOD 2014, Snowbird, UT, USA, June 22-27, 2014*. 87–98.
[21] Michael D. Lieberman, Hanan Samet, and Jagan Sankaranarayanan. 2010. Geotagging with local lexicons to build indexes for textually-specified spatial data. In *Proceedings of the 26th IEEE International Conference on Data Engineering, ICDE 2010, March 1-6, 2010, Long Beach, California, USA*. 201–212.
[22] Stanley Milgram. 1967. The Small World Problem. *Psychology Today* 2 (1967), 60–67.
[23] Kyriakos Mouratidis, Jing Li, Yu Tang, and Nikos Mamoulis. 2015. Joint Search by Social and Spatial Proximity. *IEEE Trans. Knowl. Data Eng.* 27, 3 (2015), 781–793.
[24] Mohamed Sarwat, Justin J. Levandoski, Ahmed Eldawy, and Mohamed F. Mokbel. 2014. LARS*: An Efficient and Scalable Location-Aware Recommender System. *IEEE Trans. Knowl. Data Eng.* 26, 6 (2014), 1384–1399.
[25] Jieming Shi, Nikos Mamoulis, Dingming Wu, and David W. Cheung. 2014. Density-based place clustering in geo-social networks. In *Proceedings of the International Conference on Management of Data, ACM SIGMOD 2014, Snowbird, UT, USA, June 22-27, 2014*. 99–110.
[26] Ammar Sohail, Arif Hidayat, Muhammad Aamir Cheema, and David Taniar. 2018. Location-Aware Group Preference Queries in Social-Networks. In *Databases Theory and Applications - 29th Australasian Database Conference, ADC 2018, Gold Coast, QLD, Australia, May 24-27, 2018, Proceedings*. 53–67.
[27] Yuhan Sun, Nitin Pasumarthy, and Mohamed Sarwat. 2017. On Evaluating Social Proximity-Aware Spatial Range Queries. In *Proceedings of the 18th IEEE International Conference on Mobile Data Management, MDM 2017, Daejeon, South Korea, May 29 - June 1, 2017*. 72–81.
[28] Yuhan Sun and Mohamed Sarwat. 2018. A generic database indexing framework for large-scale geographic knowledge graphs. In *Proceedings of the 26th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, SIGSPATIAL 2018, Seattle, WA, USA, November 06-09, 2018*. 289–298.
[29] Yuhan Sun and Mohamed Sarwat. 2021. Riso-Tree: An Efficient and Scalable Index for Spatial Entities in Graph Database Management Systems. *ACM Trans. Spatial Algorithms Syst.* 7, 3 (2021), 12:1–12:39.
[30] Tin Vu, Sara Migliorini, Ahmed Eldawy, and Alberto Bulussi. 2019. Spatial data generators. In *1st ACM SIGSPATIAL International Workshop on Spatial Gems (SpatialGems 2019)* (Chicago, Illinois USA). ACM.
[31] Duncan J. Watts and Steven H. Strogatz. 1998. Collective dynamics of 'small-world' networks. *Nature* 393, 6684 (1998), 440–442.
[32] Jared Winick and Sugih Jamin. 2022. *Inet-3.0: Internet Topology Generator.* Technical Report. University of Michigan, Ann Arbor.