# Synthetic Networked Data: Modeling Social-Graph Theories, and Applications

Lisette Espín-Noboa[1,2][0000−0002−3945−2966], Jan Bachmann[1,2][0000−0002−6153−4714], and Fariba Karimi[1,3][0000−0002−0037−2475]

[1] Complexity Science Hub Vienna, Austria
[2] Central European University, Austria
[3] Graz University of Technology, Austria

**Abstract.** Social networks have been widely studied over the last century from multiple disciplines to understand societal issues such as inequality in employment rates, managerial performance, and epidemic spread. Today, these and more issues can be studied at global scale thanks to the digital footprints that we generate when browsing the Web or using social media platforms. Unfortunately, scientists often struggle to access such data primarily because it is proprietary, and even when it is shared with privacy guarantees, such data is either not representative or too big. Additionally, algorithms using networked data are commonly evaluated on un-realistic synthetic networks or specific real-world network benchmarks. Unfortunately, these approaches neglect the range of the algorithm's performance across diverse network types, and lack interpretability. This tutorial explores recent developments and future directions in *social-graph modeling*, emphasizing the use of *synthetic networked data* to examine real-world issues such as algorithmic bias, ranking inequalities, spreading dynamics, and data privacy. It begins with an overview of various graph models for social networks, including node-attributed, directed, and scale-free networks. The focus then shifts to *auditing algorithms* through synthetic networks and *evaluating interventions* aimed at enhancing minority representation in top-k rankings and access to information. The tutorial concludes by addressing ongoing challenges, future directions, and open questions in this field.

**Keywords:** graph theory · social network modeling · algorithm auditing · impact assessment

## 1 Topic and relevance

While synthetic data generation has proven beneficial across various domains for addressing data imbalances and enhancing algorithmic accuracy, its application has predominantly focused on text, tabular, and image data. However, the complexity of real-world data, especially in social networks, demands a deeper exploration into how networks form and how they impact algorithms. These networks, characterized by intricate node attributes, interconnections, and dependencies, present unique challenges in data structure and algorithms. Our tutorial aims to bridge this gap, offering the ECML PKDD community valuable insights into generating *synthetic networked data*. This knowledge is crucial for developing more effective tools, understanding complex data structures, and implementing algorithms for advanced knowledge extraction and informed decision-making in network-oriented data scenarios. For example, with realistic synthetic social networks, we can scrutinize the robustness and fairness of ML algorithms, particularly when data is biased or unrepresentative of reality [14].

### 1.1 Social theories of edge formation

Social scientists are interested in studying relations between entities within social networks, e.g., how social friendship ties form between actors and explain them based on attributes such as gender, race, political affiliation, or age [38]. The complex networks community suggests a set of generative network models aiming at explaining the formation of edges focusing on the two core principles of *popularity* and *similarity* [34]. Thus, a series of approaches to study edge formation have emerged including statistical tools [25,39] and model-based approaches [40,34,22] specifically established in the physics and complex networks communities. Computer scientists use these tools to detect communities in temporal networks [3], and to study the impact of network structure on GNNs [46].

In terms of similarity, many social networks demonstrate a property known as *homophily*, which is the tendency of individuals to associate with others who are similar to them, e.g., with respect to gender or ethnicity [28]. Alternatively, individuals may also prefer to close triangles by connecting to people whom they already share a friend with [17] which in turn can explain the emergence of communities [6], high connectivity [31], and induced homophily [4]. Furthermore, the class balance or distribution of individual attributes over the network is often uneven, with coexisting groups of different sizes, e.g., one ethnic group may dominate the other in size. Popularity, on the other hand, often refers to how well connected a node is in the network which in turn creates an advantage over poorly connected nodes. This is also known as the rich-get-richer or Matthew effect when new nodes *attach preferentially* to other nodes that are already well connected [5]. Many networks, including the World Wide Web, reflect such property by means of power-law degree distributions [1].

Here, we focus on the main mechanisms of edge formation: homophily, triadic closure, activity, and preferential attachment. Moreover, we pay attention to important properties such as class (im)balance, directed edges, and edge density.

### 1.2 Network models

In this section, we will review a set of well known network generator models found in the netin[4] and networkx[5] Python packages, see Table 1.

[4] `https://pypi.org/project/netin/`

[5] `https://pypi.org/project/networkx/`

**Table 1.** Network models to be covered in this tutorial

| Attributed? | Directed? | Network models |
|---|---|---|
| No | No | Erdős-Rényi, Newman-Watts-Strogatz, Barabási-Albert, Configuration model, Holme-Kim power-law cluster graph, Exponential random graphs. |
| Yes | No | Stochastic block model, Attributed Barabási–Albert (PA), PA with homophily (PAH), PAH with triadic closure (PATCH). |
| Yes | Yes | Directed PA (DPA), DPA with homophily (DPAH). |

### 1.3 Model selection and validation

Identifying the model that best explains a given network remains an open challenge. First, we will show how to infer the hyper-parameters of each network model (e.g., homophily) given a real-world network via maximum likelihood estimation (MLE). Then, we will learn how to use and interpret different approaches including AIC [44], BIC [2], Bayes factors [18], and likelihood ratios [45].

### 1.4 Applications

We will demonstrate how to exploit *synthetic networks* to understand how certain algorithms are influenced by network structure and edge formation. The idea is to evaluate the outcomes of these algorithms, and their impact, while systematically changing the structure of the input network.

**Algorithm Auditing.** An important aspect of *explaining discrimination* [29] via *network structure* is that we gain a better understanding of the direction of bias, e.g., why and when inference discriminates against certain (groups of) people [8]. Similarly, research shows that minority groups' under-representation in top-k rankings is due to poor connectivity rather than their size [14], and this can be shown by systematically varying the structure of synthetic networks.

**Impact Assessment.** Our goal is to present innovative examples of using social network modeling in policy-making to evaluate interventions pre-deployment. This includes studies on the role of quotas and behavioral changes in networks for boosting minority visibility in top-k rankings [30] and improving information access equality and efficiency [43], along with tools for analyzing intervention scenarios to prevent health disparities [36].

## 2 Outline

Table 2 shows a summary of the main topics that we plan to cover in this tutorial. In the first part, we will explore social science theories of edge formation, translating them into practical network models and examining their key structures and properties. The second part will focus on applying these models to evaluate machine learning algorithms, and examining the effects of structural interventions across various scenarios. The tutorial will end with a discussion on current challenges and future directions.

**Table 2.** Proposed outline

| Topic | Tutor | Duration (min.) |
|---|---|---|
| *Social Theories* | | 40 |
| Social theories of edge formation | F.K. | |
| Network properties and structure | F.K. | |
| *Network models (includes hands-on)* | | 60 |
| Undirected networks | J.B. | |
| Directed networks | L.E.N. | |
| *Break* | | 30 |
| *Applications (includes hands-on)* | | 100 |
| Algorithm auditing | L.E.N. | |
| Impact assessment | J.B. | |
| *Conclusion* | | 10 |
| Challenges, future directions, and open questions | F.K. | |

## 3 Tutors

*Lisette Espín-Noboa*[6] is a postdoc at the Complexity Science Hub Vienna (CSH), and at Central European University (CEU). Her research interests lie at the intersection between computational social science, network science, and AI for social good. She is particularly focused on understanding how edges form in social networks [10], and how these mechanisms of edge formation may affect machine learning algorithms [13,8,14], decision-making [7], and human behavior [9,11,42].

*Jan Bachmann*[7] is a senior PhD. student at CEU and affiliated to the CSH. His research interest lies in studying the impact of edge formation mechanisms on the representation of minority or disadvantaged groups through network modeling [30] or collaboration network analysis.

*Fariba Karimi*[8] is a Full Professor at the Graz University of Technology (TU Graz) and a group leader at the CSH. Her research mainly focuses on computational and network approaches to address societal challenges such as gender disparities in collaboration and citation networks [19,24], visibility of minorities in social and technical systems [21,14,33], algorithmic biases [32,8,15], and

[6] `https://www.lisetteespin.info`

[7] `https://mannbach.de`

[8] `https://networkinequality.com/people/fariba-karimi`

sampling hard-to-reach groups [41,13]. Her research also touches upon the emergence of culture in Wikipedia [20,37], spreading of information and norms [23], and perception biases [26] by using mathematical models, digital traces and online experiments.

## 4 Previous editions

This is the second time we have conceptualized and planned this tutorial. The first edition[9] was taught at The ACM Web Conference 2023 in Austin, Texas [16,12]. Additionally, we have experience organizing and teaching network science topics to broad audiences. *Fariba Karimi* has given lectures and seminars on network and data science, theory, and dynamics to a broad audience including computer scientists and social scientists at the University of Koblenz-Landau and GESIS — The Leibniz Institute for the Social Sciences. *Lisette Espín-Noboa* has given network and data science lectures at TU Wien and TU-Graz[10], and co-organized and co-lectured a 4-day virtual hands-on seminar for social scientists on how to do network analysis in Python [27]. *Jan Bachmann* has contributed to teaching network science topics in courses at CEU. Last but not least, we have organized multiple workshops on "Network Structure" at Networks 2021 and NetSci 2023[11], and on "Network Inequality" at NetSci 2022 and 2023[12], where we invited a diverse group of distinguished scientists to talk about the relationship between network structure and social inequalities.

## 5 Target Audience

This tutorial is targeted to researchers who want to learn more about (i) random network generator models, (ii) deployment of synthetic networks with and without attributes and specific edge formation mechanisms, (iii) model selection given an empirical network, and (iv) how to exploit synthetic networks to tackle real-world issues including algorithmic bias, ranking inequalities, spreading dynamics, and data privacy. Participants must have a basic knowledge of coding, preferable in Python.

## 6 Requirements

### 6.1 Style and Duration

This 4-hour hybrid tutorial includes hands-on sessions. Participants can join either online or in-person with their own computers and should have all prerequisites installed.

[9] `https://bit.ly/snma2023`

[10] `https://github.com/lisette-espin/TeachingMaterials`

[11] `https://bit.ly/NetStructure`

[12] `https://sites.google.com/view/netin-satellite-2023/home`

### 6.2 Pre-requisites

All materials will be available beforehand on our website and on a GitHub repository. These include: pre-prepared Jupyter notebooks, Python scripts, libraries, settings, and slides. We will also use publicly available real-world networks [35].

### 6.3 Equipment

For on-site participants at the conference venue, a projector and pointer are needed. For online attendees, we require a stable internet connection and host permissions on Zoom[13] (or the conference's preferred platform) for screen sharing, breakout rooms, and remote access if needed.

### 6.4 Contingency plan

All the tutors are planning to be physically present at the conference. However, in case of unexpected events (e.g., restricted mobility, sickness, or bad internet connection) we will provide pre-recorded lectures of the entire tutorial. Moreover, all exercises will be given in advance as python scripts and Jupyter notebooks.

## References

1. L. A. Adamic and B. A. Huberman. Power-law distribution of the world wide web. *science*, 287(5461):2115–2115, 2000.
2. E. M. Airoldi, D. Blei, S. Fienberg, and E. Xing. Mixed membership stochastic blockmodels. *Advances in neural information processing systems*, 21, 2008.
3. A. P. Appel, R. L. Cunha, C. C. Aggarwal, and M. M. Terakado. Temporally evolving community detection and prediction in content-centric networks. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2018, Dublin, Ireland, September 10–14, 2018, Proceedings, Part II 18*, pages 3–18. Springer, 2019.
4. A. Asikainen, G. Iñiguez, J. Ureña-Carrión, K. Kaski, and M. Kivelä. Cumulative effects of triadic closure and homophily in social networks. *Science Advances*, 6(19):eaax7310, 2020.
5. A.-L. Barabási and R. Albert. Emergence of scaling in random networks. *science*, 286(5439):509–512, 1999.
6. G. Bianconi, R. K. Darst, J. Iacovacci, and S. Fortunato. Triadic closure as a basic generating mechanism of communities in complex networks. *Physical Review E*, 90(4):042806, 2014.
7. L. Espín-Noboa et al. Network fairness, its ethical issues and conflicts. (work-in-progress).
8. L. Espín-Noboa, F. Karimi, B. Ribeiro, K. Lerman, and C. Wagner. Explaining classification performance and bias via network structure and sampling technique. *Applied Network Science*, 6(1):1–25, 2021.

[13] We have Pro Zoom accounts through our institutions.

9. L. Espín Noboa, F. Lemmerich, P. Singer, and M. Strohmaier. Discovering and characterizing mobility patterns in urban spaces: A study of manhattan taxi data. In *Proceedings of the 25th International Conference Companion on World Wide Web*, pages 537–542, 2016.
10. L. Espín-Noboa, F. Lemmerich, M. Strohmaier, and P. Singer. Janus: A hypothesis-driven bayesian approach for understanding edge formation in attributed multigraphs. *Applied Network Science*, 2(1):1–20, 2017.
11. L. Espín-Noboa, F. Lemmerich, S. Walk, M. Strohmaier, and M. Musen. Hoprank: How semantic structure influences teleportation in pagerank (a case study on bioportal). In *The World Wide Web Conference*, pages 2708–2714, 2019.
12. L. Espín-Noboa, T. Peixoto, and F. Karimi. Social network modeling and applications, a tutorial. *arXiv preprint arXiv:2306.11004*, 2023.
13. L. Espín-Noboa, C. Wagner, F. Karimi, and K. Lerman. Towards quantifying sampling bias in network inference. In *Companion Proceedings of the The Web Conference 2018*, pages 1277–1285, 2018.
14. L. Espín-Noboa, C. Wagner, M. Strohmaier, and F. Karimi. Inequality and inequity in network-based ranking and recommendation algorithms. *Scientific reports*, 12(1):1–14, 2022.
15. A. Ferrara, L. Espín-Noboa, F. Karimi, and C. Wagner. Link recommendations: Their impact on network structure and minorities. *arXiv preprint arXiv:2205.06048*, 2022.
16. V. Fionda, O. Hartig, R. Abdolazimi, S. Amer-Yahia, H. Chen, X. Chen, P. Cui, J. Dalton, X. L. Dong, L. Espin-Noboa, W. Fan, M. Fritz, Q. Gan, J. Gao, X. Guo, T. Hahmann, J. Han, S. Han, E. Hruschka, L. Hu, J. Huang, U. Jaimini, O. Jeunen, Y. Jiang, F. Karimi, G. Karypis, K. Kenthapadi, H. Lakkaraju, H. W. Lauw, T. Le, T.-H. Le, D. Lee, G. Lee, L. Levontin, C.-T. Li, H. Li, Y. Li, J. C. Liao, Q. Liu, U. Lokala, B. London, S. Long, H. K. Mcginty, Y. Meng, S. Moon, U. Naseem, P. Natarajan, B. Omidvar-Tehrani, Z. Pan, D. Parekh, J. Pei, T. Peixoto, S. Pemberton, J. Poon, F. Radlinski, F. Rossetto, K. Roy, A. Salah, M. Sameki, A. Sheth, C. Shimizu, K. Shin, D. Song, J. Stoyanovich, D. Tao, J. Trippas, Q. Truong, Y.-C. Tsai, A. Uchendu, B. Van Den Akker, L. Wang, M. Wang, S. Wang, X. Wang, I. Weber, H. Weld, L. Wu, D. Xu, E. Y. Xu, S. Xu, B. Yang, K. Yang, E. Yom-Tov, J. Yoo, Z. Yu, R. Zafarani, H. Zamani, M. Zehlike, Q. Zhang, X. Zhang, Y. Zhang, Y. Zhang, Z. Zhang, L. Zhao, X. Zhao, and W. Zhu. Tutorials at the web conference 2023. In *Companion Proceedings of the ACM Web Conference 2023*, WWW '23 Companion, page 648–658, New York, NY, USA, 2023. Association for Computing Machinery.
17. M. S. Granovetter. The strength of weak ties. *American journal of sociology*, 78(6):1360–1380, 1973.
18. J. M. Hofman and C. H. Wiggins. Bayesian approach to network modularity. *Physical review letters*, 100(25):258701, 2008.
19. M. Jadidi, F. Karimi, H. Lietz, and C. Wagner. Gender disparities in science? dropout, productivity, collaborations and success of male and female computer scientists. *Advances in Complex Systems*, 21(03n04):1750011, 2018.
20. F. Karimi, L. Bohlin, A. Samoilenko, M. Rosvall, and A. Lancichinetti. Mapping bilateral information interests using the activity of wikipedia editors. *Palgrave Communications*, 1(1):1–7, 2015.
21. F. Karimi, M. Génois, C. Wagner, P. Singer, and M. Strohmaier. Homophily influences ranking of minorities in social networks. *Scientific reports*, 8(1):1–12, 2018.

22. B. Karrer and M. E. Newman. Stochastic blockmodels and community structure in networks. *Physical Review E*, 83(1):016107, 2011.
23. J. Kohne, N. Gallagher, Z. M. Kirgil, R. Paolillo, L. Padmos, and F. Karimi. The role of network structure and initial group norm distributions in norm conflict. In *Computational Conflict Research*, pages 113–140. Springer, Cham, 2020.
24. H. Kong, S. Martin-Gutierrez, and F. Karimi. Influence of the first-mover advantage on the gender disparities in physics citations. *Communications Physics*, 5(1):1–11, 2022.
25. D. Krackhardt. Predicting with networks: Nonparametric multiple regression analysis of dyadic data. *Social networks*, 10(4):359–381, 1988.
26. E. Lee, F. Karimi, C. Wagner, H.-H. Jo, M. Strohmaier, and M. Galesic. Homophily and minority-group size explain perception biases in social networks. *Nature human behaviour*, 3(10):1078–1087, 2019.
27. H. Lietz, O. Zagovora, and L. Espin-Noboa. Introduction to social network science with python, 2020. Accessed: 2022-11-10.
28. M. McPherson, L. Smith-Lovin, and J. M. Cook. Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, 27(1):415–444, 2001.
29. N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan. A survey on bias and fairness in machine learning. *arXiv preprint arXiv:1908.09635*, 2019.
30. L. Neuhäuser, F. Karimi, J. Bachmann, M. Strohmaier, and M. T. Schaub. Improving the visibility of minorities through network growth interventions. *Communications Physics*, 6(1):108, 2023.
31. M. E. Newman. Clustering and preferential attachment in growing networks. *Physical review E*, 64(2):025102, 2001.
32. E. Ntoutsi, P. Fafalios, U. Gadiraju, V. Iosifidis, W. Nejdl, M.-E. Vidal, S. Ruggieri, F. Turini, S. Papadopoulos, E. Krasanakis, et al. Bias in data-driven artificial intelligence systems—an introductory survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 10(3):e1356, 2020.
33. M. Oliveira, F. Karimi, M. Zens, J. Schaible, M. Génois, and M. Strohmaier. Group mixing drives inequality in face-to-face gatherings. *Communications Physics*, 5(1):1–9, 2022.
34. F. Papadopoulos, M. Kitsak, M. Á. Serrano, M. Boguná, and D. Krioukov. Popularity versus similarity in growing networks. *Nature*, 489(7417):537–540, 2012.
35. T. P. Peixoto. The netzschleuder network catalogue and repository, 2020.
36. S. Sajjadi, P. T. Simin, M. Shadmangohar, B. Taraktas, U. Bayram, M. V. Ruiz-Blondet, and F. Karimi. Structural inequalities exacerbate infection disparities: A computational approach, 2022.
37. A. Samoilenko, F. Karimi, D. Edler, J. Kunegis, and M. Strohmaier. Linguistic neighbourhoods: explaining cultural borders on wikipedia through multilingual coediting activity. *EPJ data science*, 5:1–20, 2016.
38. S. F. Sampson. *A novitiate in a period of change: An experimental and case study of social relationships*. Cornell University, 1968.
39. T. Snijders, M. Spreen, and R. Zwaagstra. The use of multilevel modeling for analysing personal networks: Networks of cocaine users in an urban area. *Journal of quantitative anthropology*, 5(2):85–105, 1995.
40. T. A. Snijders. Statistical models for social networks. *Review of Sociology*, 37:131–153, 2011.
41. C. Wagner, P. Singer, F. Karimi, J. Pfeffer, and M. Strohmaier. Sampling from social networks with attributes. In *Proceedings of the 26th international conference on world wide web*, pages 1181–1190, 2017.

42. S. Walk, L. Esín-Noboa, D. Helic, M. Strohmaier, and M. A. Musen. How users explore ontologies on the web: A study of ncbo's bioportal usage logs. In *Proceedings of the 26th International Conference on World Wide Web*, pages 775–784, 2017.
43. X. Wang, O. Varol, and T. Eliassi-Rad. Information access equality on generative models of complex networks. *Applied Network Science*, 7(1):1–20, 2022.
44. R. J. Williams and D. W. Purves. The probabilistic niche model reveals substantial variation in the niche structure of empirical food webs. *Ecology*, 92(9):1849–1857, 2011.
45. X. Yan, C. Shalizi, J. E. Jensen, F. Krzakala, C. Moore, L. Zdeborová, P. Zhang, and Y. Zhu. Model selection for degree-corrected block models. *Journal of Statistical Mechanics: Theory and Experiment*, 2014(5):P05007, 2014.
46. J. Zhu, Y. Yan, L. Zhao, M. Heimann, L. Akoglu, and D. Koutra. Beyond homophily in graph neural networks: Current limitations and effective designs. *Advances in neural information processing systems*, 33:7793–7804, 2020.