Hands-on tutorial on analyzing social network data in Jupyter Python: The essentials, signed networks, and network optimization

Rezvaneh Rezapour[1], Samin Aref[2], Ly Dinh[1], and Jana Diesner[1]
[1]School of Information Sciences, University of Illinois at Urbana-Champaign, USA
[2]Lab of Digital and Computational Demography, Max Planck Institute for Demographic Research, Germany

Abstract: This tutorial teaches new methods for analyzing the structure of networks with positive and negative edges (*signed networks*). Depending on the context, signed networks can be more flexible than their unsigned counterparts as they allow modeling relations of opposite nature [1]. In social contexts, examples of such relations include friendship vs. enmity, trust vs. distrust, or alliance vs. antagonism. Signed networks also have a range of other modeling applications, from biology [2] (gene regulatory networks) and finance [3] (financial portfolios) to political science [4] (collaboration and avoidance between legislators), physics, and chemistry [3].

In this tutorial, after covering general preliminaries and essentials, we focus on different methods for analyzing the structure of signed directed networks [5] based on balance theory [6], which explains the forces behind the structure of social signed networks. A signed network is *balanced* if its set of vertices can be partitioned into two subsets such that each negative edge joins vertices belonging to different subsets and each positive edge joins vertices belonging to the same subset [6]. This definition is often expressed in terms of network triads; which considers triads with an even number of negative edges to be balanced [6]. We expand this modeling approach to incorporate edge directionality [7], and consider three levels of analysis: micro- (triads), meso- (subgroups), and macro-level (whole network) [1]. For assessing micro-level balance, we use semicycles of length 3 that satisfy the condition of transitivity and sign consistency. For meso-level balance, we derive measures of cohesiveness (internal solidarity) and divisiveness (external antagonism) to capture balance in subgroups using the most fitting partition of nodes into two groups [8, 9]. For macro-level balance, we use the normalized line index [10, 11], which relies on the proportion of edges whose position suits balance.

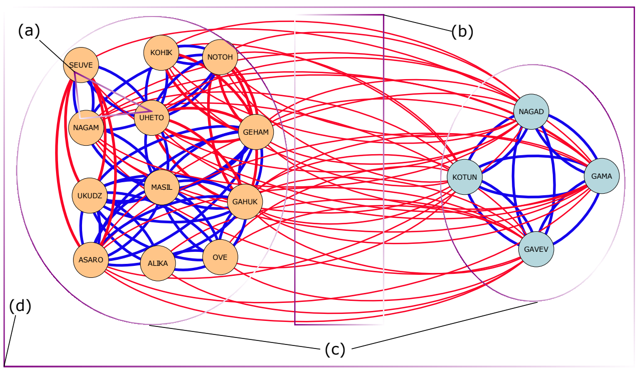

Fig. 1. Signed network of alliance (blue) and antagonism (red) between 16 tribes and our proposed approach to evaluating balance at the micro level (a), meso level (b,c), and macro-level (d).

Figure 1 shows the signed network of 16 tribes in Papua New Guinea [12]. We measure micro-level balance (a) by evaluating all transitive semicycles of length 3 in a network. We quantify divisiveness (b) by looking at the signs of the external edges between two subgroups. We quantify cohesiveness (c) by looking at the signs of internal edges in subgroups. We measure macro-level balance (d) by partitioning the network into two subgroups (node colors) with minimum intra-group negative and inter-group positive edges.

This tutorial is based on recent methodological advancements [1] at the intersection of social network analysis and graph optimization. The underlying learning objective of this workshop is gaining proficiency in evaluating the structure of signed networks at different levels, for which we propose a comprehensive yet parsimonious approach.

Organizers:

Rezvaneh (Shadi) Rezapour is a Ph.D. candidate at the School of Information Sciences (the iSchool) at the University of Illinois at Urbana-Champaign (UIUC) **where she has served as a teaching assistant for social network analysis courses**. Shadi's research is focused on computational social science. In particular, she develops models to extract meaningful information from (online) social discourse. More broadly, she is interested in combining methods from natural language processing, machine learning, and network analysis with social science theories to better understand real world behaviors, attitudes and cultures. For more information see `https://sites.google.com/view/rezapour/home`.

Samin Aref holds a Ph.D. in computer science from University of Auckland (New Zealand) with a dissertation on structural analysis of signed networks. He works as a research scientist at the Laboratory of Digital and Computational Demography, Max Planck Institute for Demographic Research (Germany) **where he has taught hand-on networks courses**. His research interests include computational social science, analyzing big data, network science, bibliometrics, and machine learning. For more information, see `https://saref.github.io/`.

Ly Dinh is a Ph.D. student at the iSchool of UIUC, **where she teaches a graduate-level course on social network analysis**. Her research topics focus on how research methods, such as network analysis, social simulation models, and text analysis, can be used to advance our understanding of various social and organizational systems. Her current projects place network science at the core to understand and explain a number of social and organizational phenomena ranging from egocentric networks to interagency emergency response networks. For more information, see `https://publish.illinois.edu/lydinh-uiuc/`.

Jana Diesner is an Associate Professor at the iSchool of UIUC, where she leads the Social Computing Lab. Her research in social computing and human-centered data science combines methods from natural language processing, social network analysis and machine learning with theories from the social sciences to advance knowledge and discovery about interaction-based and information-based systems. Jana got her Ph.D. (2012) in Societal Computing from the School of Computer Science at Carnegie Mellon University. For more information, see `http://jdiesnerlab.ischool.illinois.edu`.

Topic and Relevance: We are surrounded by different types of social networks in our daily lives. Social network analysis is a longstanding methods toolbox used to examine the structures of relations between social entities, which can represent individuals, groups, or organizations, among other entity types. Existing network metrics and models are flexible in that they can detect structural dynamics that exist at three fundamental levels of analysis, namely the micro, meso, and macro levels of networks. While several open-source tools for social network analysis are available, there is a need for a pipeline that guides scholars through a multi-level analysis of networks.

This tutorial aims to help participants from various disciplinary backgrounds to learn the state-of-the-art methods for processing and analyzing both synthetic (e.g. random, small-world, scale-free) and real-world social networks (e.g. social networking sites, weblogs, online encyclopedias). Participants will be working with sample data provided by the Tutorial Track co-chairs. Sample data is publicly available for research purposes.

Topics of Interest: 1) Social network analysis, 2) graph optimization, 3) using public social network datasets, 4) open source tools and models for analyzing social networks, and 5) visualizing social networks using NetworkX.

Duration: The duration of this workshop is 6 hours. We use a range of teaching methods and environments to ensure that all learning objectives are met despite the possibility of attendees coming from different fields with different backgrounds on network analysis.

Interaction Style: Hands-on-tutorial. Combination of both group- and individual-oriented learning activities.

Intended Audience and Level: The intended audience are researchers who use networks or plan to start using networks in their work. We do not assume any prior knowledge other than basic level of mathematics and basic familiarity with Jupyter Python (being able to run "Hello World!" in Jupyter).

Learning objectives:

- Loading synthetic and real-world networks using NetworkX.
- Describing networks via statistical measures at multiple levels of analysis.
- Visualizing networks using NetworkX.
- Analyzing degree distribution of networks.
- Detecting communities in networks.
- Working with signed directed networks.
- Evaluating micro-level balance.
- Solving graph optimization problems using Gurobi.
- Evaluating meso-level balance.
- Evaluating macro-level balance.

Tutorial materials All our material will be available for free on our tutorial website, and participants are permitted to reuse the material.

Online format (plans for ensuring "success")

One of the main objectives of this tutorial is to make sure that scholar with different backgrounds and levels of familiarity with social network analysis can follow and learn the methods that we present in each module. For this purpose we will follow the steps below:

- We will send clear and detailed instructions and preparatory material in advance to participants so that everyone can start on equal footing.
- Before the session, we will ask participants to fill out a short survey on (1) their main objectives for participating in the tutorial and (2) their background. We then modify the modules accordingly.
- We provide a continuous feedback and interaction with participants in each training module using synchronous Mentimeter polls and questions, and active monitoring of Zoom Q/A.
- Each module has one main instructor and one assistant. The instructor manages the main Zoom session while there is a breakout room for one-to-one discussion or troubleshooting with the instructor assistant. Participants who have any technical questions can move from the main session to a breakout room in Zoom where the instructor assistant helps them resolve the issue so that they come back to the main session and continue following along with the rest of the class.
- Our training modules are interactive, with frequent usage of whiteboards to visually explain concepts. - We will create a Piazza forum for a wiki-style documentation of the topics covered and the discussions at the workshop so that participants can access a knowledge base later on.
- We plan to run a post-workshop evaluation survey and share the results with Tutorial Track co-chairs.

Additional info for hands-on tutorials:
Tutorial duration: 6 hours.
Operating system and required installed tools on attendees' devices: Any OS, Anaconda (Jupyter notebook), Gurobi installed on Jupyter (see setup instructions below)
List of software licenses required for the tools: (free academic) license for Gurobi (see setup instructions below).
Setup instructions for attendees: step-by-step setup instructions available at github.com/saref/multilevel-balance.

Video snippet: Link for related talk at CCS'2020

Tentative Outline:

- Module 1 (60 minutes):
  - Loading synthetic and real-world networks using NetworkX
  - Describing the statistics of a network at multiple levels (ego, dyad, triad, subgroup, whole-network)
  - Break (10 minutes)
- Module 2 (60 minutes):
  - Analyzing the degree distribution of a network
  - Mixing patterns (assortativity)
  - Detecting communities in networks
  - Break (10 minutes)
- Module 3 (60 minutes):
  - Visualizing a network using NetworkX
  - Motifs and components
  - Operations on networks
  - Break (10 minutes)
- Module 4 (60 minutes):
  - Operations on directed and signed networks
  - Evaluating micro-level balance
  - Break (10 minutes)
- Module 5 (60 minutes):

- Modeling and solving graph optimization problems using Gurobi in Jupyter
- Break (10 minutes)

- Module 6 (50 minutes):
  - Evaluating meso-level balance
  - Evaluating macro-level balance
- Final Q&A (10 minutes)

---

[1] S. Aref, L. Dinh, R. Rezapour, and J. Diesner, "Multilevel structural evaluation of signed directed social networks based on balance theory," *Scientific Reports*, vol. 10, no. 1, pp. 1–12, 2020.

[2] G. Iacono, F. Ramezani, N. Soranzo, and C. Altafini, "Determining the distance to monotonicity of a biological network: a graph-theoretical approach," *IET Systems Biology*, vol. 4, no. 3, pp. 223–235, 2010.

[3] S. Aref and M. C. Wilson, "Balance and frustration in signed networks," *Journal of Complex Networks*, vol. 7, pp. 163–189, 08 2019.

[4] S. Aref and Z. Neal, "Detecting coalitions by optimally partitioning signed networks of political collaboration," *Scientific Reports*, vol. 10, no. 1, pp. 1–10, 2020.

[5] S. Aref, L. Dinh, and R. Rezapour, "Dataset of directed signed networks from social domain." *figshare* `http://dx.doi.org/10.6084/m9.figshare.12152628`, 2020.

[6] D. Cartwright and F. Harary, "Structural balance: a generalization of Heider's theory," *Psychological Review*, vol. 63, no. 5, pp. 277–293, 1956.

[7] L. Dinh, R. Rezapour, L. Jiang, and J. Diesner, "Structural balance in signed digraphs: Considering transitivity to measure balance in graphs constructed by using different link signing methods," *arXiv preprint arXiv:2006.02565*, 2020.

[8] S. Aref, A. J. Mason, and M. C. Wilson, "Computing the line index of balance using integer programming optimisation," in *Optimization Problems in Graph Theory* (B. Goldengorin, ed.), pp. 65–84, Springer, 2018.

[9] S. Aref, A. J. Mason, and M. C. Wilson, "A modeling and computational study of the frustration index in signed networks," *Networks*, vol. 75, no. 1, pp. 95–110, 2020.

[10] F. Harary, "On the measurement of structural balance," *Behavioral Science*, vol. 4, no. 4, pp. 316–323, 1959.

[11] S. Aref and M. C. Wilson, "Measuring partial balance in signed networks," *Journal of Complex Networks*, vol. 6, no. 4, pp. 566–595, 2018.

[12] K. E. Read, "Cultures of the central highlands, New Guinea," *Southwestern Journal of Anthropology*, vol. 10, no. 1, pp. 1–43, 1954.