Final Project Guidelines and Network Datasets

ELEC 573: Network Science and Analytics

1 General description

The goal of the final project is for you to investigate and implement state-of-the-art network analysis tools and algorithms in an application of your preference. You should select a specific topic related to network science and perform a relatively in-depth survey of the topic. This requires finding good literature sources, possibly performing some analysis and/or numerical simulations to experiment on interesting network datasets, and providing a detailed summary of the main ideas. I am flexible with the type of study you may choose to carry out. However, keep in mind that your project is an in-depth study of a specific topic with potential original contributions; it is not just a summary of a few research papers. Please see Section 4 on what constitutes (and what does not constitute) an admissible project.

You are encouraged to team up with other students in groups of up to three people for the project. Still, if you wish so, you can work alone.

2 Potential topics

Your first task is to pick a topic that you like for the project. You are encouraged to pick an application area that is related to your current or future research, thus, I suggest that you first speak with your advisor (if you have one) about a possible topic that relates to networks. In addition, you are welcome to talk with me during office hours and I will be happy to make suggestions and brainstorm with you, and point you towards useful datasets, code, papers, and other resources.

A few exciting areas from where you could explore project ideas are:

- Signal processing on graphs: An emerging field with the goal of extending high-dimensional data analysis to networks and other irregular domains. A few good tutorial references to get started are [1–3]. There are multiple sub-areas in this field, and I'd be happy to point you to more specialized literature if needed.

- Deep learning for network data: An accessible tutorial paper on early work in this timely field [4]. See also `http://geometricdeeplearning.com/` for comprehensive resources about geometric deep learning.

- Graph anomaly detection: A comprehensive tutorial paper on the state of the art [5].

- Topological data analysis: Using algebraic topology tools (such as persistence homology) to perform data analysis of network data [6].
- Network visualization: Plotting a network in low dimensions while preserving important features is a research direction on its own [7–9].
- Diffusion and cascading processes over networks: Opinion formation and influence maximization. A seminal paper on influence maximization [10].
- Community detection and clustering: A challenging problem that remains an active area of research. A comprehensive tutorial paper [11]. Axiomatic theoretical foundations can be found for traditional [12] and hierarchical [13, 14] clustering.
- Graphons: A modern graph-theoretical concept that serves both as the limit for graphs of increasing size as well as a non-parametric network model [15, 16].
- Games on graphs: Game theoretical problems where the agents' communication or observations are described via a graph [17].
- Social learning: Accumulation of social knowledge in network-based systems [18, 19].
- Compressed sensing, sparsity and low-rank structures for network analytics: A few related pointers include a tutorial paper on dynamic communication network health monitoring [20], applications to smart grid load curve cleansing [21] and estimation of diffusion network structure [22].

In deciding what to work on, I suggest that you explore the datasets stated in Section 3.

3 Network datasets

Researchers and institutions have compiled valuable network datasets. Some of these resources follow:

- Stanford Large Network Dataset Collection compiled by Jure Leskovec (`http://snap.stanford.edu/data/index.html`).
- Network Repository is a scientific network data repository with interactive visual analytics tools (`http://networkrepository.com/index.php`).
- The unified New York City taxi and Uber data with information of more than a billion trips (`https://github.com/toddwschneider/nyc-taxi-data`).
- Graph and Social Data compiled by the Yahoo Webscope Program (`https://webscope.sandbox.yahoo.com/catalog.php?datatype=g&guccounter=1`).

- The Social Computing Data Repository hosts datasets from a collection of many different social media sites (`http://socialcomputing.asu.edu/pages/datasets`).
- A list of network datasets used in Eric Kolaczyk's book on Statistical Analysis of Network Data (`http://math.bu.edu/people/kolaczyk/datasets.html`).
- The Yelp Dataset Challenge offers a social review dataset including a large social graph (`https://www.yelp.com/dataset/challenge`).
- Network datasets compiled over the years by Mark Newman (`http://www-personal.umich.edu/~mejn/netdata/`).
- The UCI Network Data Repository is an effort to facilitate the scientific study of networks (`https://networkdata.ics.uci.edu/`); see also their own list of additional network dataset resources (`https://networkdata.ics.uci.edu/resources.php`).
- KONECT (the Koblenz Network Collection) is a project to collect large network datasets of all types in order to perform research in network science and related fields, collected by the Institute of Web Science and Technologies at the University of Koblenz-Landau (`http://konect.uni-koblenz.de/`).
- A collection of complex networks compiled by Uri Alon (`http://www.weizmann.ac.il/mcb/UriAlon/download/collection-complex-networks`).

4 On what constitutes an admissible project

We may roughly divide the type of potential projects into theory-based and application-based projects (although most projects will have elements from both subtypes):

- **Theory-based project:** An analytical project that considers a model, an algorithm or a network property and derives a rigorous theoretical result about it.
- **Application-based project:** An experimental evaluation of algorithms and models on interesting network data, implementing your own code and/or investigating existing software for network analysis.

A rule of thumb to have in mind is that the more original work there is in your project, the less (grading) importance will be given to the in-depth analysis of existing work. As an illustration, imagine that you choose the topic of graph signal processing and within that topic you are interested in a theory-based project on optimal graph filter design. This is an area with dozens of published papers at this point. Hence, a few examples of admissible and non-admissible projects in this topic follow:

i) *Non-admissible*: Select two journal papers, read them, and write a report comparing them.
ii) *Admissible*: Go over 10-15 papers in the area, identify main directions of thought and how they evolved, based on this determine the most probable future directions for the field and the associated potential applications.
iii) *Admissible*: Go over 3-4 seminal papers in the area, come up with an original idea based on those papers (and make sure that this idea was not proposed elsewhere). Develop the idea and illustrate its application.

Notice that for example *ii)*, where the original contribution is smaller, more importance is given to the in-depth survey, whereas the opposite is true for example *iii)*. Assume now that, by contrast, you are interested in developing an application-based project, where you apply network tools to an interesting dataset. Here, focus will be given to the originality of the project and the creativity of the network techniques. A few examples follow:
i) *Non-admissible*: Choose a known network dataset, compute a few network features, and report them (like problem 1.3 in your homework).
ii) *Admissible*: Select a topic of your interest (e.g. music), and build or download an interesting dataset (e.g. the graph of related artists in Spotify). Analyze how structural properties of this graph relate to objective measures of popularity (weeks in world-wide top rankings); correlate outputs of community detection with music genre and unveil a notion of distance between genres; consider data associated with the nodes of this graph (monthly listeners) and evaluate the success of different semi-supervised learning techniques.

If you are considering a project plan and you are unsure if it constitutes an admissible project or not, please come talk to me during office hours. Do not wait until the project proposal deadline (see Section 5 for more info).

5 Deliverables and grading

As stated in the syllabus, the project will have three associated due dates with specific deliverables and grade weights (all submissions will be done through Canvas):
i) Project proposal (10% of course grade): Brief summary of what you plan to do for your project. Due 10/18/18.
ii) Progress report (15% of course grade): First (incomplete) draft of your final report, though naturally shorter and most likely without your major results. Should serve as a skeleton for the final report and as a checkpoint to assess feasibility. Due 11/15/18.
iii) Final report and presentation (35% of course grade): Should provide a clear and detailed description of what you did, what results you obtained, and what you have learned and concluded from your work. Apart from the written report, you will prepare a short presentation for the rest of the class. Final presentations will be on the last week of class. The final report submission deadline is 12/12/18.

Projects will be evaluated based on:

- Technical quality: Is the project technically sound? Are the modeling assumptions made and the algorithms tried reasonable? Do the conclusions suggest in-depth critical thinking about the chosen topic, possibly conveying novel insights about the problem and/or chosen algorithms?
- Significance: Is this an interesting and timely problem to work on? Is this work useful and the underlying research area likely to have impact?
- Clarity of presentation: How effectively are the research findings conveyed orally (during the oral presentation) and in writing (in the final report)?

To build your reports (proposal, progress report, and final report), please use the template files based on the NIPS conference paper kit (`https://nips.cc/Conferences/2018/PaperInformation/StyleFiles`).

5.1 Project proposal

The project proposal should summarize what you plan to do for your project. The writeup should not exceed 4 pages, and you should try to include: i) A clear description of the problem that you will be addressing; ii) Preliminary ideas on how you plan to address it (models/algorithms/techniques); iii) Basic literature references you will be consulting; iv) If applicable what software tools you will need for your work (or if you plan to write your own code what language you will use); v) Network datasets that you will be working with; vi) What you expect to produce as a result of your work and how you will judge success of the project; and vii) Any additional information that you think that will be useful in evaluating your plans.

5.2 Progress report

The progress report should look like a first (incomplete) draft of your final report, but naturally shorter and most likely without your major results. The write-up should not exceed 8 pages, and you should try to include:

i) An introduction, literature review of relevant prior work (with corresponding list of references), and clear problem statement in finalized form; ii) If you collected your own data to construct a network, describe that process; iii) For all those applicable, provide mathematical derivations, detailed model descriptions, and algorithms you have used, adapted or developed; iv) Summary of preliminary results obtained so far and datasets you have analyzed; v) Anything worth commenting on unforeseen complications that arose, and workarounds; and vi) Outline of the work-to-do, including any portions that you see infeasible and why.

5.3 Final report

The final report should naturally build on your progress report, providing a clear and detailed description of what you did, what results you obtained, and what you have learned and concluded from your work. The write-up should not exceed 12 pages, and you should try to include:

i) A motivating introduction, literature review of relevant prior work, and clear problem statement in finalized form; ii) If you collected your own data to construct a network, describe that process; iii) For all those applicable, provide mathematical derivations, detailed model descriptions, and algorithms you have used, adapted or developed; iv) Description of your experiments, showcasing the obtained results and a relevant discussion based on your observations; v) Conclusions indicating the accomplished goals and what you learned, as well as possible extensions or future directions; and vi) A list of relevant references.

References

[1] D. I. Shuman, S. K. Narang, P. Frossard, A. Ortega, and P. Vandergheynst, "The emerging field of signal processing on graphs: Extending high-dimensional data analysis to networks and other irregular domains," *IEEE Signal Process. Mag.*, vol. 30, no. 3, pp. 83–98, May 2013.

[2] A. Sandryhaila and J. M. F. Moura, "Discrete signal processing on graphs," *IEEE Trans. Signal Process.*, vol. 61, no. 7, pp. 1644–1656, April 2013.

[3] A. Ortega, P. Frossard, J. Kova?evi?, J. M. F. Moura, and P. Vandergheynst, "Graph signal processing: Overview, challenges, and applications," *Proc. IEEE*, vol. 106, no. 5, pp. 808–828, May 2018.

[4] M. M. Bronstein, J. Bruna, Y. LeCun, A. Szlam, and P. Vandergheynst, "Geometric deep learning: Going beyond euclidean data," *IEEE Signal Process. Mag.*, vol. 34, no. 4, pp. 18–42, July 2017.

[5] L. Akoglu, H. Tong, and D. Koutra, "Graph based anomaly detection and description: a survey," *Data Mining and Knowledge Discovery*, vol. 29, no. 3, pp. 626–688, May 2015.

[6] G. Carlsson, "Topology and data," *Bulletin of the American Mathematical Society*, vol. 46, no. 2, pp. 255–308, 2009.

[7] M. J. McGuffin, "Simple algorithms for network visualization: A tutorial," *Tsinghua Science and Technology*, vol. 17, no. 4, pp. 383–398, Aug 2012.

[8] K. W. Boyack, R. Klavans, and K. Börner, "Mapping the backbone of science," *Scientometrics*, vol. 64, no. 3, pp. 351–374, Aug 2005.

[9] J. I. Alvarez-Hamelin, L. Dall'Asta, A. Barrat, and A. Vespignani, "Large scale networks fingerprinting and visualization using the k-core decomposition," in *Advances in neural information processing systems*, 2006, pp. 41–50.

[10] D. Kempe, J. Kleinberg, and É. Tardos, "Maximizing the spread of influence through a social network," in *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2003, pp. 137–146.

[11] S. Fortunato, "Community detection in graphs," *Physics reports*, vol. 486, no. 3-5, pp. 75–174, 2010.

[12] J. M. Kleinberg, "An impossibility theorem for clustering," in *Advances in neural information processing systems*, 2003, pp. 463–470.

[13] G. Carlsson and F. Mémoli, "Characterization, stability and convergence of hierarchical clustering methods," *Journal of machine learning research*, vol. 11, no. Apr, pp. 1425–1470, 2010.

[14] G. Carlsson, F. Mémoli, A. Ribeiro, and S. Segarra, "Hierarchical clustering of asymmetric networks," *Advances in Data Analysis and Classification*, vol. 12, no. 1, pp. 65–105, 2018.

[15] L. Lovász and B. Szegedy, "Limits of dense graph sequences," *Journal of Combinatorial Theory, Series B*, vol. 96, no. 6, pp. 933–957, 2006.

[16] M. Avella-Medina, F. Parise, M. T. Schaub, and S. Segarra, "Centrality measures for graphons: Accounting for uncertainty in networks," *arXiv preprint arXiv:1707.09350*, 2017.

[17] A. Galeotti, S. Goyal, M. O. Jackson, F. Vega-Redondo, and L. Yariv, "Network games," *The Review of Economic Studies*, vol. 77, no. 1, pp. 218–244, 2010.

[18] B. Golub and M. O. Jackson, "Naive learning in social networks and the wisdom of crowds," *American Economic Journal: Microeconomics*, vol. 2, no. 1, pp. 112–49, February 2010.

[19] ——, "How homophily affects the speed of learning and best-response dynamics," *The Quarterly Journal of Economics*, vol. 127, no. 3, pp. 1287–1338, 2012.

[20] G. Mateos and K. Rajawat, "Dynamic network cartography: Advances in network health monitoring," *IEEE Signal Process. Mag.*, vol. 30, no. 3, pp. 129–143, May 2013.

[21] G. Mateos and G. B. Giannakis, "Load curve data cleansing and imputation via sparsity and low rank," *IEEE Trans. Smart Grid*, vol. 4, no. 4, pp. 2347–2355, Dec 2013.

[22] H. Daneshmand, M. Gomez-Rodriguez, L. Song, and B. Schölkopf, "Estimating diffusion network structures: Recovery conditions, sample complexity & soft-thresholding algorithm," in *International Conference on Machine Learning*, 2014, pp. 793–801.