

Project Topics

Below is a list of possible project topics. Some of these are open-ended, meaning that you are required to come up with a new algorithm or model, and formulate it yourselves. Such projects may require more effort, but they will be also graded based on the effort, as well as the final result. Others are more straight-forward, you would need to obtain a complex dataset and apply algorithms on this dataset. There are also more theoretical projects, and more practical ones, so you can pick depending on your preference.

You will also have to present in class one paper related with your project. The list below includes the paper for each project.

Papers also vary in difficulty and scope. For experimental papers, that just report results of experimental studies, we expect that you just present and explain the main findings. Since such papers require less effort, you will be asked to present 2 such papers.

You also need to create a web page for the project (including the final report and dataset used).

Projects should be done **in teams of at most two** students.

Deliverables and Timeline:

- *Week before Christmas vacations*
 - A two-page project proposal outlining what you plan to do. This should include the topic (and papers) of your presentation
 - A 15' presentation of the project proposal (on Tuesday 21/12)
 - First version of the web page (include the link in the project proposal)
- *Week after Christmas vacations*
 - A 20' presentation of 1-2 research papers related to your project (presentations tentatively scheduled for 12/1)
- *End of January*
 - The source code of your project
 - A final report describing your project (the report needs to follow a specific format)
 - Final version of the web page

Topic 1

Structural diversity based on network embeddings

Project:

The main goal is to explore embeddings for defining network diversity. The idea is to define the diversity of a node based on the embeddings of its neighbors, for example, you may define the diversity of a node to be equal to the average distance between the embeddings of its neighbors. Then, study the correlation of the embedding-based diversity of a node with other metrics, such as the degree, clustering co-efficient, or PageRank of the node. Since this is an experimentally-oriented paper, use at least 3 different network embedding methods (e.g., node2vec, DeepWalk, GraphSage), 5 different real graph datasets and potentially a number of synthetic datasets (e.g., forest fire, preferential attachment, random graphs). Optionally, try to relate structural diversity as defined in the papers to be presented with the embedding-based definition.

Papers:

Johan Ugander, Lars Backstrom, Cameron Marlow, Jon M. Kleinberg: *Structural diversity in social contagion*. Proc. Natl. Acad. Sci. U.S.A. 109(16): 5962-5966 (2012)

Xin Huang, Hong Cheng, Rong-Hua Li, Lu Qin, Jeffrey Xu Yu: *Top-K structural diversity search in large networks*. VLDB J. 24(3): 319-343 (2015) (you are asked to study/present only Sections 1, 2, and 3.)

Topic 2

Content homophily in a real social network

Project:

The main goal of this project is to measure content homophily in a real social network. One way to formulate this problem is to test whether friends in a social network post similar content. Choose a social network, e.g., Twitter, and a set of users. Then, construct the ego network of these users and collect the posts of the users belonging to these ego networks. Use word2vec to define similarity between posts. Check the similarity between pairs of users that are friends and pairs of users that are not friends. You can further check whether friends endorse (for example, retweet) similar content. Choose the set of users, using a specific criterion, for example, choose newspapers (or, politicians) with different political orientations, football teams, or players, etc., so that you may be able to draw more general conclusions.

Paper:

Marina Drosou, H.V. Jagadish, Evaggelia Pitoura, and Julia Stoyanovich. *Diversity in Big Data: A Review*. Big Data. Jun 2017

Topic 3

Fairness in a real social network

Project:

The main goal is to measure fairness in a real social network, in particular, Github. Construct a number of friends-networks from Github and use existing software to determine the gender of people in these networks. Consider different ways to rank nodes in the constructed networks, for example, based on degree, PageRank, Personalized PageRank, centrality, etc. Test parity fairness: is the percentage of women in the top positions close to the percentage of women in the whole population? Then, assume a diffusion process in these networks using the IC (Independent Cascade) model with a small number of seeds. Test parity fairness in the affected nodes. You are free to select seeds randomly, using some heuristic, e.g., highest degree, or, any other algorithm.

Paper:

Fariba Karimi, Mathieu Génois, Claudia Wagner, Philipp Singer and Markus Strohmaier. *Homophily influences ranking of minorities in social networks*. *Scientific Reports* 8 (2018)

Topic 4

Link Recommendation for reducing polarization.

Project

In the the paper by Matakos et al., a metric is defined for measuring the polarization in a social network with opinions. Propose algorithms for the problem of suggesting links to reduce the polarization metric. The recommendations should take into account the probability of a recommendation to be accepted.

Paper

Kiran Garimella, Gianmarco De Francisci Morales, Aristides Gionis, and Michael Mathioudakis. 2017. *Reducing Controversy by Connecting Opposing Views*. In Proceedings of the Tenth ACM International Conference on Web Search and Data Mining (WSDM '17). ACM, New York, NY, USA, 81-90.

Original paper:

A. Matakos, E. Terzi, P. Tsaparas. "Measuring and Moderating Polarization in Social Networks", ECMP/PKDD 2017

Also useful:

Xi Chen, Jefrey Lijffijt, and Tijl De Bie. 2018. *Quantifying and Minimizing Risk of Conflict in Social Networks*. In Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD '18). ACM, New York, NY, USA, 1197-1205.

Topic 5

Empirical study of approximate fairness for PageRank algorithm

Project

In [1] fair variants of the PageRank algorithm were proposed. A fair PageRank algorithm performs a modified version of the PageRank random walk that guarantees that the protected group in the network will receive exactly ϕ fraction of the Pagerank value. In the modified random walk, for every node, the probability of transitioning to the protected group is exactly ϕ . The goal of this project is to study approximations of this algorithm that modify partially the PageRank random walk, and understand which modifications play an important role in achieving fairness. As a starting point, you will assume that only a percent of the nodes performs the fair walk. You could also consider approximate versions of the Fairness-sensitive PageRank algorithm defined in the paper.

(There exists an updated version of the paper that is not yet publicly available. We will give the paper to anyone who decides to take this project).

Paper

[1] Sotiris Tsioutsoulis, Evaggelia Pitoura, Panayiotis Tsaparas, Ilias Kleftakis, Nikos Mamoulis: Fairness-Aware Link Analysis. CoRR abs/2005.14431 (2020)

Topic 6

Fairness of the SALSA algorithm

Project

In [1] a fair variant of the PageRank algorithm was proposed. A fair PageRank algorithm performs a modified version of the PageRank random walk that guarantees that the protected group in the network will receive exactly ϕ fraction of the PageRank value. The goal of this project is to define and experiment with fairness for the SALSA algorithm. In SALSA there are two random walks, you may modify one, or, both of them. You will need to evaluate your algorithm for a recommendation problem.

(There exists an updated version of the paper that is not yet publicly available. We will give the paper to anyone who decides to take this project).

Paper

[1] Sotiris Tsioutsoulis, Evaggelia Pitoura, Panayiotis Tsaparas, Ilias Kleftakis, Nikos Mamoulis: Fairness-Aware Link Analysis. CoRR abs/2005.14431 (2020)

Topic 7

Using fair random walks in graph embeddings

Project

When nodes in a network belong to different groups (e.g., female/male), we would like all groups to be fairly represented in the embeddings. Previous work focused on the node2vec embedding and proposed a modified fair walk to achieve equal representation of groups in the produced embeddings [1]. In this project, you will evaluate this approach. Specifically, you will implement [1] (or use an existing implementation of [1]) and evaluate the produced embeddings for the link recommendation problem. You may also use and evaluate other variants of the fair random walk.

Paper

[1] Tahleen A. Rahman, Bartłomiej Surma, Michael Backes, Yang Zhang: Fairwalk: Towards Fair Graph Embedding. IJCAI 2019: 3289-3295

Topic 8

Finding lasting dense connected subgraphs

Project

An evolving graph is a graph that changes over time. It can be represented as a set of graph snapshots, $G_0, G_1, \dots, G_n$, where each snapshot G_i corresponds to the state of the graph at time instance i .

Previous work studied the following problem: Given a set of graph snapshots, identify the set of nodes that are the most densely connected in all snapshots [1]. The algorithms proposed in [1] are based on a popular greedy algorithm for static graphs that works in rounds and at each round, it removes from the graph the node having the smallest degree. The goal of this project is to extend the proposed algorithms with the requirement that the most densely connected subgraph is also connected in all, or, a subset of the snapshots.

Paper

[1] Konstantinos Semertzidis, Evaggelia Pitoura, Evimaria Terzi, Panayiotis Tsaparas: Finding lasting dense subgraphs. Data Min. Knowl. Discov. 33(5): 1417-1445 (2019)

Topic 9

Another option is to suggest a project of your own, based on what you have seen in the class so far, questions you may have thought of, and things that are related to your research area. In this case you should create a project proposal (initially just a paragraph or an idea) and contact us to discuss it.