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 GitHub 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:

Friday, 15/4/2022	A one-page project proposal outlining what you plan to do. This should include the topic (and papers) of your presentation
Wednesday, 4/5/2022	A 15' presentation of the project proposal
Wednesday, 25/5/2022	Paper presentation
Wednesday, 15/6/2022	Submit GitHub page: ○ Source code of the project ○ Datasets used ○ Project report

Topic 1

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 a distributed representation of the text (e.g., BERT, word2vec, etc) 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. You may 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 can draw more general conclusions.

Paper:

Luca Maria Aiello, Alain Barrat, Rossano Schifanella, Ciro Cattuto, Benjamin Markines, Filippo Menczer: *Friendship prediction and homophily in social media.* ACM Trans. Web 6(2): 9:1-9:33 (2012)

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

Topic 2

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, or Reddit 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. In addition to the real dataset that you collect, you will also repeat your experiments in 4 real datasets that we will provide.

Paper:

Lisette Espín-Noboa, Claudia Wagner, Markus Strohmaier, Fariba Karimi: *Inequality and Inequity in Network-based Ranking and Recommendation Algorithms*. CoRR abs/2110.00072 (2021)

Topic 3

Link Recommendation for reducing polarization.

Project

In 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:

Antonis Matakos, Evimaria Terzi, Panayiotis Tsaparas: *Measuring and moderating opinion polarization in social networks.* Data Min. Knowl. Discov. 31(5): 1480-1505 (2017)

Other:

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.

Topic 4

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 replace random walk in node2vec by the residual fair random walk proposed by Tsioutsiouliklis et al. You will evaluate the produced embeddings for the link recommendation problem.

Paper

Sotiris Tsioutsiouliklis, Evaggelia Pitoura, Panayiotis Tsaparas, Ilias Kleftakis, Nikos Mamoulis: *Fairness-Aware PageRank*. WWW 2021: 3815-3826

Other:

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

Topic 5

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, $G_0, G_1, .., G_n$, 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 6

Diffusion of toxicity

Project

The goal of the project is to study how toxicity is propagated in discussions. You will collect discussions from Reddit and build discussion trees with root the initial post and nodes the comments. Test whether the toxicity of the original post correlates with the toxicity of the comments, how fast toxicity propagates, the size of the trees with respect to the toxicity of the post and comments, etc. To measure toxicity use Perspective (https://www.perspectiveapi.com/). Use your findings to discuss whether the propagation of toxicity follows any of the known opinion, or diffusion propagation models.

Paper

Marinos Poiitis, Athena Vakali, Nicolas Kourtellis: *On the Aggression Diffusion Modeling and Minimization in Twitter.* ACM Trans. Web 16(1): 5:1-5:24 (2022)

Other

Chrysoula Terizi, Despoina Chatzakou, Evaggelia Pitoura, Panayiotis Tsaparas, Nicolas Kourtellis: *Modeling aggression propagation on social media*. Online Soc. Networks Media 24: 100137 (2021)

Topic 7

Fair community detection

Project

In this project, you will study community detection in terms of fairness. For the definition of fairness, we will use the definition from [2], where given groups of nodes defined based on some protected attribute (e.g., gender, race), we ask that each group is equally represented in each community. You will study the fairness of at least 3 of the community algorithms that were presented in class. In addition, you will evaluate the ML algorithm proposed by Cavallari, et al. For the evaluation, you need to use at least 5 networks. In addition, derive and evaluate a post-processing method that tries to achieve fairness with minimum changes from the initial clustering.

Paper

Sandro Cavallari, Vincent W. Zheng, Hongyun Cai, Kevin Chen-Chuan Chang, Erik Cambria: *Learning Community Embedding with Community Detection and Node Embedding on Graphs*. CIKM 2017: 377-386

Other

Flavio Chierichetti, Ravi Kumar, Silvio Lattanzi, Sergei Vassilvitskii: *Fair Clustering Through Fairlets.* NIPS 2017: 5029-5037

Topic 8

Fair node classification

Project

In this project, you will study node classification in terms of fairness. For the definition of fairness, you will use a statistical group-based definition that looks at the classification errors for each group (see, Verma and Rubin, below). You will study the fairness of at least 4 of the embedding methods for the classification problem using at least 5 datasets. In addition, derive and evaluate a post-processing method that tries to achieve fairness with minimum changes from the initial classification.

Paper

Verma, S., Rubin, J.: *Fairness definitions explained*. In: Proceedings of the International Workshop on Software Fairness, FairWare@ICSE 2018, pp. 1–7. ACM (2018)

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.