Bachelor projects 2021

Data Intensive systems group

In addition to the projects described in this document, we can also provide several projects in **machine learning algorithms, data mining, and data management**. Feel free to contact any of us (Ira Assent, Panos Karras, Davide Mottin, Cigdem Aslay) for more information.

1. Content-Based Influence by the Linear Threshold Model

Advisor: *Panos Karras (panos@cs.au.dk)*

How can a message best spread in a network after we share it with friends and followers? What if the desired effect requires the simultaneous spread of several messages? How can we improve our performance on such tasks in an online manner, learning from experience? Significant research activity has been dedicated to the problem of strategically selecting a seed set of initial adopters so as to maximize a message's spread in a network. Yet this line of work assumes that the success of such a campaign depends solely on the choice of a tunable seed set of adopters, regardless of how users perceive the propagated meme, which is fixed. Yet in many real-world settings, the opposite holds:
a meme's propagation depends on users' perceptions of its tunable characteristics, while the set of initiators is fixed.
In this project, you will build on previous work on the problem that arises in such circumstances: find the optimal content characteristics for a creative promotion campaign, to maximize its expected spread over a social network, pursuing two breakthroughs:

1. Addressing the problem by a popular model for information diffusion in networks, the linear threshold model.
2. Handling the case in which the propagated message consists of multiple parts, which need to be allocated to initial propagators.

In all cases, you will be building on top of an existing code base; the conducted work will be publishable.

Related work:

1. Amit Goyal, Wei Lu, Laks V. S. Lakshmanan: SIMPATH: An Efficient Algorithm for Influence Maximization under the Linear Threshold Model. ICDM 2011
2. Sergei Ivanov, Konstantinos Theocharidis, Manolis Terrovitis, Panagiotis Karras: Content Recommendation for Viral Social Influence. SIGIR 2017

2. Authenticating Geometric Queries

Advisor: *Panos Karras (panos@cs.au.dk)*

The wide use of mobile connected devices allows for sophisticated spatial queries to be performed by roaming users in location-based services, involving merchandizing and logistics. At the same time, there is growing need on behalf of users to verify that query results are genuine, especially in an outsourced data model, where data owners publish their data to a third-party service provider who delivers results to the users. As such a service provider may manipulate results in favor of certain sponsors, it becomes compulsory to provide to users not only those results, but also a proof of their correctness; in other words, to authenticate results.

In a *query authentication* mechanism, the data owner publishes not only data, but also signed endorsements of the data. Given a query, the service provider returns both the query results and a proof, called verification object (VO). The client uses this VO, together with the query results, to reconstruct the endorsements and thus verify the correctness of the results.

Significant progress has been made towards developing authentication mechanisms for several ordinary spatial queries, such as range queries [1] and top-k queries [2]. However, the spatial queries that are relevant to users usually involve more complex geometric operations, whose results also need to be authenticated. An example of such an operation is the *Centerpoint Query*, which returns a point such that a line through it divides a given point-set into two roughly equal subsets.

In this thesis project, you will study previous works in the area and develop a novel solution for authenticating of geometric queries such as the centerpoint query. The conducted work will be suitable for publication.

Related work:

1. Dimitrios Papadopoulos, Stavros Papadopoulos, Nikos Triandopoulos: Taking Authenticated Range Queries to Arbitrary Dimensions. ACM Conference on Computer and Communications Security 2014: 819-830
2. Qian Chen, Haibo Hu, Jianliang Xu: Authenticating Top-k Queries in Location-based Services with Confidentiality. PVLDB 7(1): 49-60 (2013).
3. Shreesh Jadhav, Asish Mukhopadhyay: Computing a Centerpoint of a Finite Planar Set of Points in Linear Time. Symposium on Computational Geometry 1993: 83-90.

3. Similarity Search with Dynamic Time Warping

Advisor: *Panos Karras (panos@cs.au.dk) and Ira Assent (ira@cs.au.dk)*

Similarity search in time-series (or data-series) databases finds application in finance, health care, energy installation monitoring, and bioinformatics. In its offline formulation, the problem is to extract records from a repository, which are the most similar to a given query record. The way similarity is defined plays a cardinal role. The Euclidean distance metric, which is most popular, cannot capture similarity in the presence of shifts, skews, and discontinuities. For that purpose, one can use other distance measures, such as Dynamic Time Warping (DTW), which allows the flexibility to capture such features. However, it remains a challenge to devise index that enables efficient DTW- based similarity search over large time series databases. The curse of dimensionality, which arises in such problems, becomes even more debilitating in the case of DTW measures. Past approaches to indexing DTW lack the versatility to offer multiple levels of resolution that can speed up search and retrieve results more efficiently.

In this thesis project, you will study previous works in the area, using existing code base, and develop a novel solution for DTW-based data-series similarity search using adaptations of state-of-the-art methodologies. The conducted work will have components of research, data preprocessing, implementation, and experimentation.

Related work:

4. Ira Assent, Marc Wichterich, Ralph Krieger, Hardy Kremer, Thomas Seidl: Anticipatory DTW for Efficient Similarity Search in Time Series Databases. PVLDB 2(1): 826-837 (2009)
5. Shrikant Kashyap, Panagiotis Karras: Scalable kNN search on vertically stored time series. KDD 2011: 1334-1342

4. Interactive Explainable AI

Advisor: *Davide Mottin* (*davide@cs.au.dk*)

Explainable AI allows for interpretable explanations for complex machine learning prediction models. Recently, there has been quite an interest in this area that has led to the development of a number of techniques to explain complex models [1]. Unfortunately, most of these explainers fail to capture when the machine learning model is "right for the wrong reasons", which means that the model is correct but the explanation why it is correct is wrong. To circumvent this problem, recent works have proposed approaches that involve the user in the loop and ask for the right explanation [2]. While useful, these approaches do not allow to improve the machine learning model, because they provide explanations only after the training. We have invented a method that learns the right prediction and explanation at the same time and involves the user in the process. Currently, the method does not scale to large datasets with a lot of features and needs to find the best explanation among *all* possible explanations, which is infeasible.

This project aims at studying fast GPU algorithms and sampling methods to overcome the aforementioned limitations and will implement a testing framework for experiments with multiple parameters.

Tasks:

- Reading and understanding how the method works
- Construct an experimental framework for testing multiple parameters (using e.g., sacred library)
- Find and include more datasets in the evaluation
- Implement advanced sampling or pruning techniques to reduce the search space
- Improve performance and scalability by moving to GPU computation (e.g., using gpytorch library)

Related work:

[1] Das, A. and Rad, P., 2020. Opportunities and challenges in explainable artificial intelligence (xai): A survey. arXiv preprint arXiv:2006.11371.
[2] Teso, S.; and Kersting, K. 2019. Explanatory Interactive Machine Learning. In Proceedings of AIES'19

5. Graph alignment: the good, the bad, and the ugly

Advisor: *Davide Mottin* (*davide@cs.au.dk*)

Graph alignment is the problem of finding correspondences among nodes in two (or more) graphs. The problem has been studied extensively and multiple solutions have been proposed to tackle the problem (for instance [1,2,3]). However, the current research seems to focus on limited experimental evaluations in which an existing graph is compared to its copy with random edges deleted. This kind of perturbation seems unrealistic as multiple models proved that graph evolution does not happen randomly [4].

In this project, we will test and explore different graph alignment algorithms and real datasets and evaluate their performance on real datasets and under different evolution models. If successful, the project could open the way to new methods that finally take into account real graphs for boosting performance of graph alignment.

Tasks:

- Reading, understanding the relevant literature in graph alignment
- Implementation of algorithms into an experimental framework
- Collection of datasets with time evolving features and real datasets with ground truth alignment
- Generation of realistic synthetic datasets that test different hypotheses for graph alignment
- Experimental evaluation of the methods on the dataset

Related work:

[1] Heimann, M., Shen, H., Safavi, T. and Koutra, D., 2018, October. Regal: Representation learning-based graph alignment. In Proceedings of the 27th ACM International Conference on Information and Knowledge Management (pp. 117-126). ACM.
[2] Nassar, H., Veldt, N., Mohammadi, S., Grama, A. and Gleich, D.F., 2018, April. Low rank spectral network alignment. In Proceedings of the 2018 World Wide Web Conference (pp. 619-628). International World Wide Web Conferences Steering Committee.
[3] Bayati, M., Gleich, D.F., Saberi, A. and Wang, Y., 2013. Message-passing algorithms for sparse network alignment. ACM Transactions on Knowledge Discovery from Data (TKDD), 7(1), p.3.
[4] Leskovec, J., Kleinberg, J. and Faloutsos, C., 2007. Graph evolution: Densification and shrinking diameters. ACM Transactions on Knowledge Discovery from Data (TKDD), 1(1), p.2.

7. Studying the fairness of ML models

Advisor: *Ira Assent (ira@cs.au.dk)*

In this project, we study a novel approach for evaluating the fairness of Machine Learning models. As Machine Learning is increasingly used to support decision-making that affects people's lives, e.g. when granting access to bank loans or in deciding recidivism cases, issues of fairness are of increasing concern. Specifically, the question is whether ML reproduce or even introduce discriminative patterns into decision-making, typically indirectly via biased data sources.
Building on an existing code base in our research group (python), publicly available data sets with sensitive attributes, as well as a draft report, the project should investigate a recent proposal by two students to check the fairness of models using a sampling-based approach. The goal is to devise a proper experimental evaluation that validates the claims in the report, and that further clarifies the methodology and its underlying reasoning, and possibly even expands it further.

Related work:
Surveys with overview over fairness in machine learning:
[1] https://arxiv.org/pdf/2010.04053.pdf
[2] https://arxiv.org/pdf/1908.09635.pdf

Tasks

- Create an overview over existing fairness in machine learning models
- Review claims and approach in draft report
- Familiarize with existing code base, re-run experiments
- Identify and pre-process suitable data sources (setting and discussing requirements)
- Devise new experiments (using tools, data), evaluate, and derive conclusions
- Describe and analyze methodology in accordance with experimental findings

8. Algorithms for Approximating Betweenness Centrality on Large Graphs

Advisor: *Cigdem Aslay* (*cigdem@cs.au.dk*)

Centrality measures are fundamental concepts in graph analysis that help to quantify the importance of nodes in a variety of applications including social, biological and communication networks. Betweenness centrality is a widely used centrality measure that defines the importance of a node proportionally to the fraction of all-pairs shortest paths passing through that node. Computation of betweenness centrality is very costly for large graphs, therefore, many approximate methods based on sampling have been proposed.

The goal of the project is to provide a theoretical and experimental comparison of the existing algorithms approximating betweenness centrality on static and dynamic graphs. The final report should include

- literature review on exact and approximation algorithms,
- a theory section covering the basic analysis of different approximate methods,
- a summary of the implemented algorithms,
- an experimental evaluation of the implemented algorithms comparing their accuracy and efficiency.

Related work:
[1] Brandes, Ulrik. "A faster algorithm for betweenness centrality." Journal of mathematical sociology 25.2 (2001): 163-177.
[2] Yoshida, Yuichi. "Almost linear-time algorithms for adaptive betweenness centrality using hypergraph sketches." Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining. 2014.
[3] Riondato, Matteo, and Eli Upfal. "ABRA: Approximating betweenness centrality in static and dynamic graphs with rademacher averages." ACM Transactions on Knowledge Discovery from Data (TKDD) 12.5 (2018): 1-38.

9. Creating and personalizing knowledge graphs

Advisor: *Ira Assent (ira@cs.au.dk)*

This is a rather open-ended project, in which the goal is to analyze and evaluate empirically techniques for creating, maintaining, expanding knowledge graphs. Knowledge graphs are increasingly common, and have found use in a number of applications [1]. Given the dynamic development of methods and tools, we would like to establish an overview, in particular for the medical domain [2,3].
The project can be adapted to the interests of the students within this general setup. In particular, recent research articles on knowledge graph analysis provide a theoretical foundation [4]; the empirical evaluation can either focus on an in-depth analysis of a single technique against requirements or on a comparative evaluation of a number of methods. Projects include, among others, machine learning methods to complete knowledge graphs and predict new connections, explanations of machine learning methods, graph summarization, modern queries on knowledge graphs.

Related work:

[1] Industry-scale knowledge graphs https://queue.acm.org/detail.cfm?id=3332266
[2] Tutorial with overview over recent research: https://kgtutorial.github.io/
[3] Blog article discussion use of tool Grakn.AI which can serve as inspiration: https://blog.grakn.ai/text-mined-knowledge-graphs-beyond-text-mining-1ff207a7d850
[4] Song, Q., Wu, Y., Lin, P., Dong, L. X., & Sun, H. (2018). Mining summaries for knowledge graph search. *IEEE Transactions on Knowledge and Data Engineering*, *30*(10), 1887-1900.

Tasks

- Select methods and tools that are adequate for managing and analyzing medical records
- Identify and pre-process suitable data sources (setting and discussing requirements, basic coding)
- Create knowledge graphs (using tools, data)
- Evaluate the performance (coding test scripts, running experiments)
- (optional) Discuss and relate to recent research articles