Bachelor projects 2023

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. Evolutionary games and spatial upstream reciprocity

Advisor: *P. Karras (piekarras@gmail.com), A. Pavlogiannis (pavlogiannis@cs.au.dk)*

Game theory studies how selfish individuals shape their behavior according to the behavior of their neighbors. One apparent paradox to selfishness is *cooperation: why do we spend time/effort/resources in cooperating with others?* Evolutionary game theory takes a Darwinian viewpoint to behavior, and looks for simple mechanisms that make cooperation evolutionary advantageous. One such mechanism is *upstream reciprocity: when you receive help, you feel good and might help others*. Upstream reciprocity alone does not promote cooperation, but it does so when the population interacts on a simple 1-dimensional structure.

The project aims to extend this study of upstream reciprocity to more realistic, 2-dimensional population structures, such as 2-dimensional grids. It consists of 2 stages:

1. Implement the model of evolutionary games with upstream reciprocity on graphs: The model takes as input a 2-dimensional graph (e.g., a grid, or more generally a planar graph), and some parameters of the game, and simulates the evolution of the game on the graph.
2. Use the model to phrase and answer questions in upstream reciprocity, such as “does upstream reciprocity favor cooperation on 2-dimensional structures”, and “for what parameters of the evolutionary game it does so?”

The project requires interest in discrete math/graph theory, probability, programming and simulation.

Related work:

https://royalsocietypublishing.org/doi/10.1098/rspb.2006.0125

2. Structural social balance under controversy

Advisor: *P. Karras (piekarras@gmail.com), A. Pavlogiannis (pavlogiannis@cs.au.dk)*

"The enemy of my enemy is my friend". This social axiom has been shaping interpersonal relations since the dawn of society. Structural balance theory models relationships as a signed graph: nodes correspond to entities (people, nations etc), and a signed edge between two nodes captures their relationship ("+" means "friends", "-" means "enemies"). A triangle is imbalanced when the product of its signs is “-". Imbalance creates social unrest that eventually gets resolved by one sign (randomly) changing (e.g., two enemies make amends under the uniting influence of a common friend). What is the state of the final population?

The project will study a similar setting, in the presence of a controversial subject X. Signs between nodes still capture interpersonal relations (friends/enemies). Every individual also has an opinion on X: “+” denotes support, “-“ denotes resentment. This time, triangles are formed between two nodes and X. If the nodes disagree on X, they will either become enemies or one will convince the other to change stance on X.

The project aims to develop and study this stochastic model of social balance. It consists of 2 stages:

1. Implement this model on graphs. Given an input graph and other parameters of the process, the task is to simulate the evolution of the graph until balance is reached.
2. Use the model to study basic questions of social balance, such as the number of friendships made and broken in this process.

The project requires interest in discrete math/graph theory, probability, programming and simulation.

Related work:
https://en.wikipedia.org/wiki/Signed_graph
http://physics.bu.edu/~redner/pubs/pdf/dresden.pdf

3. Similarity Search with Dynamic Time Warping

Advisor: *P. Karras (piekarras@gmail.com), 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:

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)
Shrikant Kashyap, Panagiotis Karras: Scalable kNN search on vertically stored time series. KDD 2011: 1334-1342

4. BESo: Beyond Entity Summarization

Advisor: *Davide Mottin* (*davide@cs.au.dk*) and *Ira Assent* (ira@cs.au.dk)

Entity summarization [1] refers to the problem of finding a small and descriptive version of an entity. For instance, if we want to provide a short description of Denmark we could say that Denmark is a country, located in Europe, part of Scandinavia, with capital Copenhagen. Entity summarization studies the problem in knowledge graphs, where each node is an entity (Copenhaghen, Denmark) and each connection is a relationship (capital of).

A number of methods have been proposed to solve the problem [1], all of them relying on a benchmark called ESBM. Recently, we have found two problems:

1. The benchmark is not complete nor clean.
2. The methods focus on the problem of finding the most relevant immediate connections to a node. However, sometimes we would like to provide a hierarchical summary.

Two potential projects arise:

1. Cleaning the ESBM dataset, reimplement some state-of-the-art algorithms, and reproduce the results of some of these algorithms.
2. Extend the Entity summarization problem to Hierarchical summarization.

Tasks:

- Reading and understanding the main concepts of Entity summarization (start from [1])
- Familiarize with ESBM and our current extension
- Implement some baseline or extend the problem to hierarchical summarization
- Evaluate the methods.
- Report the results

Related work:

[1] https://sites.google.com/view/entity-summarization-tutorials/www2020

[2] https://github.com/nju-websoft/ESBM

5. SaSSo: Scalable (Spectral) Subgraph Localization

Advisor: *Davide Mottin* (*davide@cs.au.dk*) *and P. Karras* (piekarras@gmail.com)

A number of problems can be formulated as finding a graph (i.e., a network of nodes connected through edges) inside another graph. For instance, detecting patterns of fraud transactions inside a network of transactions, or detecting potential proteins into a virus. This problem we call subgraph localization and is computationally hard to solve (NP-hard).

We have devised a method, SSL [1], that detects subgraphs by solving the simplified problem of thinking of the graph as a geometrical shape. Although more scalable than previous methods, our algorithm takes a long time to find subgraphs in graphs with more than 100 nodes.

In this project, we intend to improve its scalability to 1000 nodes or more with no loss in quality. We will provide all the necessary knowledge to understand the details of the algorithm and carry on the project without problems. You will delve into sophisticated theories, such as, the approximation of graph spectrum [2] and perturbation theory [3], experiment with state-of-the-art methods, and potentially influence the research community.

An **alternative project**, but similar in spirit and techniques, is the study of scalability for clustering with Graph Neural Networks, a special Neural Networks for graphs. Contact davide@cs.au.dk if interested.

Disclaimer: This project is for students interested in algorithmic aspects and math theory.

Tasks:

- Reading, understanding the relevant literature in graph localization
- Replicate some of the experiments in SSL
- Devise and implement algorithms to speed up the computation of SSL.
- Test the results on large datasets of increasing size.

Related work:

[1] https://openreview.net/pdf?id=TS_VsCpuWr
[2] Cohen-Steiner, D., Kong, W., Sohler, C. and Valiant, G., 2018, July. Approximating the spectrum of a graph. KDD.
[3] https://en.wikipedia.org/wiki/Perturbation_theory

6. 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

7. Automatically labeling data for classifying rhetorical appeals

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

This project is within natural language processing and the aim is to detect rhetorical appeals in texts. The long-term aim of detecting rhetorical appeals is to understand and identify misinformation.

We want to construct a text dataset containing labels to train machine learning models for detecting rhetorical appeals on. We approach this by writing label functions through the Snorkel framework and by using state-of-the art Transformer models for zero-shot classification. The idea is to write multiple functions to (roughly) identify and predict the labels e.g. using different models. The advantage of this is the fact that with multiple predictions we can adjust the noise from each labelling function.

We also need to create some ground-truth labels by humans for testing and evaluation. The data can be in Danish or English.

Tasks:

- Preprocess the data source
- Discuss and define guidelines for annotations
- Annotate a part of the data to use in the evaluation
- Get familiar with the Snorkel package (python) and run experiments
- Devise and write different label functions to build a dataset
- Evaluate, discuss and reflect on the results

References:

https://cs.brown.edu/people/sbach/files/ratner-vldb17.pdf
https://www.snorkel.org/use-cases/01-spam-tutorial
https://en.wikipedia.org/wiki/Modes_of_persuasion

8. Using 3D-Shape Descriptors for Clustering of Molecular Dynamics Data

Advisor: *Ira Assent (ira@cs.au.dk), Anna Beer (beer@cs.au.dk)*

This project is in clustering, i.e. a data mining approach for automatically grouping data based on mutual similarity. Clustering can help gain an overview over data or understand complex processes, e.g. protein movements in chemistry or biology. When modelling a protein's movements over time, clustering can reveal different states or conformations in a time series describing those movements [1]. Where many approaches apply simple dimensionality reduction before clustering, we want to investigate the usage of (3D-) shape descriptors [2,3] for clustering a protein's conformations. Additionally to well-known descriptors, we also want to apply some new shape descriptors we developed in previous work (python).
Applying shape descriptors on the accessible surface area of the proteins could yield significantly better clusterings than common methods while offering a high explainability.

Tasks:

- Get familiar with working with molecular dynamics (MD) data (libraries: e.g., mdtraj or pyemma)
- Apply existing well-known 3D shape descriptors on MD data
- Apply new shape descriptors from python code-base
- Implement shape descriptor from [3]
- Compare clustering on descriptors with common clustering techniques

Related Work:

[1] Glielmo, A., Husic, B. E., Rodriguez, A., Clementi, C., Noé, F., & Laio, A. (2021). Unsupervised learning methods for molecular simulation data. *Chemical Reviews*, *121*(16), 9722-9758.

[2] Lara Lopez, G., Peña Pérez Negrón, A., De Antonio Jimenez, A., Ramirez Rodriguez, J., & Imbert Paredes, R. (2017). Comparative analysis of shape descriptors for 3D objects. Multimedia Tools and Applications, 76(5), 6993-7040.

[3] Ankerst, M., Kastenmüller, G., Kriegel, H. P., & Seidl, T. (1999, July). 3D shape histograms for similarity search and classification in spatial databases. In International symposium on spatial databases (pp. 207-226). Springer, Berlin, Heidelberg.

9. Densely Connected Subgraphs in Dual Graphs (DCS-Dual)

Advisor: *Petros Petsinis* (petsinis@cs.au.dk) *and P. Karras* (piekarras@gmail.com)

There have been proposed multiple measurements that quantify how densely and well-connected a graph is; density, k-edge connectivity, k-core, algebraic connectivity etc. Discovering dense components in a graph has been extensively studied [1,3] since it is the core-part of many problems in Social Network Analysis, Biology and Recommendation Systems.

In this project we seek to study a recently introduced research problem on dual graphs [2,4] called DCS-Dual problem:
Given a pair of graphs G, H on the same set of nodes V , how do we find a subset of nodes S ⊆ V that induces a well-connected subgraph in G and a dense subgraph in H?

By [4]: *"Numerous real-world applications, ranging from computational biology and computational neuroscience to computational social science, take as input a **dual** graph, namely a **pair** of graphs on the **same set of nodes**."*

The project has three main components:

A) Study related works for: i) **dense (connected) Subgraph Discovery** and ii) **DCS-Dual** problems.

B) Inspired by Ai) combine a density measure with a connectivity measure to formulate a new **DCS-Dual** problem.

C) Inspired by Aii) implement an efficient and effective algorithm for the formulated problem of B).

The project requires interest in math and graph theory.

[1] Greedy Approximation Algorithms for Finding Dense Components in a Graph (APPROX 2000)

[2] Finding Dense and Connected Subgraphs in Dual Networks (ICDE 2015)

[3] Finding densest k-connected subgraphs (Discrete Applied Mathematics 2021)

[4] Dense and well-connected subgraph detection in dual networks (SIAM 2022)

10. Elastic Graph Indexing

Advisor: *Davide Mottin* (*davide@cs.au.dk*) *and P. Karras* (piekarras@gmail.com)

The indexing of large graph databases poses a challenge that conventional data indexing does not. Graphs are complex structures that are hard to put in order or even partition with respect to their outstanding attributes. Therefore, graph indexing is usually done by ad hoc methods and relying on computationally hard similarity measures [1]. An alternative approach to indexing large graph databases relies on *spectral* properties of graphs [2]. This approach is promising, yet has only been attempted by extracting local vertex signatures, without trying to rely on representations of complete graphs and subgraphs

In this project, we aim to develop a method for graph indexing relying on spectral signatures of the complete graphs as well as their subgraphs.We will build a hierarchical index structures that allow for efficient and effective similarity search in graph databases based on the properties and interrelationships of such signatures. In addition, we will explore the potential for building such indexes adaptively, as in [3], by constructing additional parts of the spectral index hierararhc on demand, in response to queries that express user interest.

The project will involve collaboration with PhD student K. Skitsas.
The project requires a strong interest in math and graph theory.

[1] https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=6228085
[2] https://dl.acm.org/doi/10.1145/1353343.1353369
[3] http://vldb.org/pvldb/vol5/p502_felixhalim_vldb2012.pdf