Project Title: Combining modularity-based and Markov-based community detection algorithms to accurately and reliably cluster networked data

This project involves evaluating and combining algorithms for solving an unsupervised learning task involving network/graph data.
Community detection is the process of inductively identifying groups within a networked system. It is a core problem in data science with a wide range of applications in life sciences (e.g., biological systems, gene regulatory networks), social sciences (offline/online social networks), physical sciences (VLSI design, complex systems), mathematical sciences (agent-based models, computer vision), and even health sciences (drug discovery).
Among numerous approaches, the most common method is maximizing an objective function, *modularity*, across different partitions of the graph nodes. Despite their design philosophy and wide adoption, heuristic modularity-maximization algorithms rarely return an optimal partition or anything similar which is shown to limit their predictive capabilities. The Bayan algorithm has been developed in a recent U of T project. Within its computational limits, it outperforms existing methods in accurate retrieval of communities on standard benchmarks. A Python implementation of Bayan (bayanpy) is publicly accessible through the package installer for Python (pip). Alternative methods which also show remarkable capabilities are algorithms which rely on the concept of Markov stability and random walks.
In this project, the student will work alongside a senior researcher from Huawei and the project supervisor to complete a series of weekly tasks. After reviewing the literature, and receiving training on the details of the Bayan algorithm, weekly research tasks will be assigned which may involve data analyses, computational experimentation, and implementing and testing new algorithm features and speed-ups in a GitHub environment.
This project leverages the state-of-the-art methods in computing and optimization to push the limits for solving an unsupervised learning problem accurately and efficiently. The output of this research project contributes to the development of a reliable, open-source, and reproducible algorithm for a robust and theoretically grounded detection of clusters in any network dataset, thereby improving upon a widely used computational tool for data science and machine learning.

Unfortunately, there is no funding for the student who works on this project.

Required skills:

1. Python
2. Python libraries for data analysis, machine learning, and optimization (pandas, numpy, Gurobi, scikit-learn)
3. Familiarity with graph theory, network science, and operations research

Other desired skills (to have or acquire during the project):

4. Python libraries for large-scale/network data analysis (networkX, CDlib, igraph, seaborn, joblib, bayanpy)
5. Python libraries and methods for graph representation learning (graph2vec, node2vec, deepwalk)
6. machine learning explainability (Shap values, partial dependence plots)
7. familiarity with data science methods and discrete optimization
8. academic writing and research experience
9. Julia programming language (basic usage)

Interested MEng students should send one email to Prof. Samin Aref (aref@mie.utoronto.ca) with the following documents combined into one pdf file: CV, unofficial transcript(s), one paragraph describing their interest in the project, one paragraph describing how they match the skills listed in the description of the project.