Daniel K. Sääw
St Catharine's College
dks28

Computer Science Tripos Part II Project Proposal

On Spectral Methods for Clustering of Irregular Digraphs

24th October 2019

Project Originators: Dr Luca Zanetti, Dr Thomas Sauerwald, Daniel Sääw

Project Supervisors: Dr Thomas Sauerwald and Dr Luca Zanetti

Director of Studies: Dr Sergei Taraskin

Overseers: Prof. Anuj Dawar and Prof. Andrew Moore

Introduction and Description of the Work

Clustering in general is the process of partitioning some data with the purpose of grouping together data that is ‘similar’ in some way. It has thus become a fundamental, unsupervised, method in Machine Learning that allows an abstraction of some data set into a (comparatively) small number of communities comprising some subset of the original data the members of which are some notion of ‘similar’ to one another. Due to results in spectral graph theory (in particular, the fact that the algebraïc multiplicity of 0 in an undirected graph’s Laplacian’s characteristic polynomial is the number of connected components in this graph, as well as the Davis-Kahan Theorem [2]), when clustering graphs with respect to, for example, conductance, we may apply spectral algorithms.

This means that it is fairly easy to cluster undirected graphs if by ‘communities’ (i.e. vertex subsets that are ‘similar’) we mean well-connected internally compared to the connections to the complement of the community. However, it is less clear what a community might be in a directed setting, since the edge direction makes notions of ‘connectivity’ inherently more complex. For example, some digraph G might exhibit a ‘cyclic flow’ structure in that the connections between individual subgraphs of G might be unidirectional, whilst each such subgraph has equal in- and out-degrees, meaning that the undirected graph constructed from symmetrising G might not exhibit communities in the classical sense. In general, the directional information might crucially change the meaning of the graph. Therefore, a number of techniques for clustering digraphs have emerged based on notions of ‘directional communities’. These techniques are usually supported with some model of directed random graph derived from Stochastic Block-Models (SBMs), which have the perhaps unreasonable property that each node so generated graphs will, in expectation, have equal in-degrees and out-degrees [1, 4].

This project, therefore, will involve attempting to develop a model of random digraphs that will exhibit directional communities in a sense similar to previous directed SBMs, whilst improving over previous such models by also exhibiting properties of real-world graphs not included in previous models. For this purpose, a known model of random graph for an undirected setting will likely be augmented to feature directed edges to enable directional communities. Should this fail, the project would fall back on investigating the performance of the proposed algorithms on real-world graphs themselves.
This model (or a set of real-world graphs with properties not modelled by previous digraph models) will serve as a foundation for a comparison study of several spectral techniques for clustering that will determine if

the proposed techniques perform satisfactory clustering on such ‘irregular’ graphs. The random (or real-world) graphs used for this will have the additional benefit that such graphs lend themselves to regularisation techniques (again, previously developed for the setting of undirected graphs) – that is modifications to the underlying graphs that make certain properties (such as degree) more homogeneous across the vertex set – with the purpose of improving clusters detected by the algorithms, meaning that the final step of the project will be conducting a second comparison study evaluating the improvement of each algorithm under the employment of such regularisation techniques.

Starting Point

The concepts of graph clustering and random (as well as real-world) graphs were briefly introduced in the Paper 3 course *Machine Learning and Real-world Data*. In that course, however, the problem was approached programmatically where my project will employ spectral methods. The linear algebra thus required for the project was touched on in the course *Mathematical Methods I* from the Natural Sciences Tripos, and linear algebra in general is used in a number of courses of Part I of the Computer Science Tripos.

In the specific area of spectral clustering, there have been a few research papers concerning the detection of ‘directional’ communities, introducing models of random digraphs with community structures, which should help extend undirected random graph models for this project. The project will involve implementing and refining algorithms presented in [1, 4]. Furthermore, there is some preëxisting literature on regularising undirected graphs for the purpose of clustering (e.g. [5, 3]), from which I will draw a starting point for the corresponding parts of the project.

Substance and Structure of the Project

The work carried out for this project will be the following:

1. A model of random graph will be developed (and failing that, existing collections of real-world graphs will be sampled to replace the random model) that has the properties of irregularity, in the sense that certain properties of the graphs will be strongly heterogeneous across the graphs’ vertices, and directional communities, in the sense

that certain subsets of vertices will exhibit a notion of similarity internally that would be lost without directional information about the graph. This model will draw inspiration from previous directed SBMs and might be influenced by properties of a certain kind of real-world graph.

2. Implementations of previously presented spectral algorithms for clustering digraphs will be written, and comparatively evaluated in a statistically significant setting to determine whether the proposed algorithms perform satisfactory clustering with respect to the known, underlying communities inherent in the model of random graph.
3. Techniques for regularising graphs that have previously been developed for undirected graphs will be applied (and if necessary, first ported to a directed setting) to graphs drawn from the model developed at first, and the change in performance of the implemented algorithms will be evaluated in a statistically significant setting.

Possible Extensions

The project might also...

... investigate a hypothesis formulated in some previous literature that conjectures that using a graph representation that interpolates between the directed and an abstracted, undirected, representation might improve clustering results by retaining the directionality information whilst making the graph more susceptible to classical spectral clustering techniques.

... apply the algorithms to real-world graphs that should intuitively be similar to those generated by the random model, and evaluate the performance based on node tags and determining the quality of the recovered clusters without a ground truth for underlying clusters.

... develop formally a theoretical performance bound for the algorithms when applied to the graphs sampled from the random model.

Success Criteria

The project should be deemed a success if the following criteria have been met:

1. Code implementing a (to be developed) model of random directed graph with highly irregular degrees and directional communities has

been written, or a suitable set of real-world graphs with irregular features has been collected[1]

2. Spectral clustering algorithms proposed in previous research papers have been implemented, both base versions and variants conducting initial graph regularisation
3. A statistically significant verdict has been reached about whether the proposed algorithms are suitable for the clustering of highly irregular graphs
4. A statistically significant verdict has been reached about whether the clusters detected by the proposed algorithms are improved by the application of graph regularisation

Timetable and Milestones

I will divide my time into 2 week slots, to allow for sufficient granularity while simultaneously not committing myself to an excessively tight schedule.

Slot 1 - *12th October – 25th October*

- finalise and submit project proposal
- research backgrounds of spectral clustering more thoroughly

Deadlines

- Phase 1 Proposal Deadline – 14th October, 3 PM.
- Draft Proposal deadline – 18th October, 12 noon.
- Proposal Deadline – 25th October, 12 noon.

Slot 2 - *26th October – 8th November*

- research models of random graphs with highly irregular degrees
- determine a general type of real-world graph that is of interest to decide details of random graph model

Slot 3 - *9th November – 22nd November*

- formalise random-graph model

[1] Data sets corresponding to, for example, citation networks in academia are available, and are prone to having the irregularity requirements (in the case of citation networks, with respect to in-degree) as desired.

- begin planning components of experimental set-up

Slot 4 - *23rd November – 6th December*
Note: Due to my choice of Units of Assessment, this period will likely be very stressful. I therefore foresee less time to work on the project during this slot.

- implement code to generate graphs drawn from the random model

Slot 5 - *7th December – 20th December*

- implement previously proposed spectral algorithms for digraph clustering
- implement code for graph normalisation

Slot 6 - *21st December – 3rd January*

- develop a measure for extent of success of clustering procedure
- begin drafting introduction, preparation and implementation chapters of the dissertation

Slot 7 - *4th January – 17th January*

- write code to generate graphs for test bench
- complete preparation for comparison study of algorithms, including generating a set of graphs to compose the 'test bench' to ensure internal consistency

Slot 8 - *18th January – 31st January*

- conduct experiments to evaluate clustering algorithms with regard to performance measure developed in slot 6
- write and submit progress report

Deadlines

- Progress Report Deadline – 31st January, 12 noon.

Slot 9 - *1st February – 14th February*

- begin evaluating results of experiments on unnormalised algorithms
- conduct experiments to evaluate performance change due to normalisation
- prepare progress report presentation

Deadlines

- Progress Report Presentations – 6th, 7th, 10th, 11th February, 2 PM.

Slot 10 - *15th February – 28th February*

- complete performance evaluation
- reach verdict about statistical significance of results

Slot 11 - *29st February – 13th March*

- allow time to ensure that success criteria have been met
- consult with project supervisors on best continuing work (e.g. which of the potential extensions would now be best approached)

Slot 12 - *14th March – 27th March*

- carry out further work according to supervisor consultation
- complete first chapters drafted in slot 6

Slot 13 - *28th March – 10th April*

- write evaluation and conclusion chapters
- continue and complete any work that is yet to be completed

Slot 14 - *11th April – 24th April*

- complete draft dissertation
- revise details of work to improve overall quality of dissertation and ensure consistency in dissertation

Slot 15 - *25th April – 8th May*

- allow for time to complete any remaining work in case this timeline has not been accurate
- submit dissertation as early as possible to allow transition to revision

Deadlines

- Dissertation Deadline – 8th May, 12 noon.
- Source Code Deadline – 8th May, 5 PM.

Resources Declaration

For the programming work to be done, I intend to use my own personal laptop. This has an Intel i7-7700HQ 4.0 GHz processor as well 16 gigabytes of random-access memory. To guard against the case that my personal machine fails I shall back up all my work, including any progress on the dissertation, using version control via GitHub, and regular commits to the working repository.

Should my computer thus fail, I should be able to transition fairly smoothly to the machines provided by the managed cluster service (MCS).

Experiments will be constructed, and any code required, using MatLab, a license for which the University provides me with; any graphs that need to be visualised will be visualised using either MatLab or GePhi.

Should the development of a model for random graphs fail, publically available real-world data sets (yielding directed graphs) will be sampled to generate test benches for the project instead of random graphs generated from an original random model.

References

[1] Mihai Cucuringu et al. *Hermitian matrices for clustering directed graphs: insights and applications.* 2019. arXiv: 1908.02096.

[2] Chandler Davis and W. M. Kahan. “The Rotation of Eigenvectors by a Perturbation. III”. In: *SIAM Journal on Numerical Analysis* 7.1 (Mar. 1970), pp. 1–46. DOI: 10.1137/0707001. URL: https://doi.org/10.1137/0707001.

[3] Antony Joseph and Bin Yu. *Impact of regularization on Spectral Clustering.* 2013. arXiv: 1312.1733.

[4] Karl Rohe, Tai Qin and Bin Yu. *Co-clustering for directed graphs: the Stochastic co-Blockmodel and spectral algorithm Di-Sim.* 2012. arXiv: 1204.2296.

[5] Yilin Zhang and Karl Rohe. *Understanding Regularized Spectral Clustering via Graph Conductance.* 2018. arXiv: 1806.01468.