Graph Clustering Performance

Pierre Miasnikof[1], **Alexander Y. Shestopaloff**[2], **Yuri Lawryshyn**[1], **Anthony J. Bonner**[3]

[1]University of Toronto, Dept. of Chemical Engineering and Applied Chemistry

[2]The Alan Turing Institute, London, UK

[3]University of Toronto, Dept. of Computer Science

Contact: p.miasnikof@mail.utoronto.ca

Background

- Graph clustering and network community detection is a topic that has gained much attention recently
- However, performance evaluation of clustering algorithms remains an open problem
- Reliance on "ground truth" data sets does indeed provide objective reproducible performance measurements
- Unfortunately, it does not provide guarantees the algorithm will perform similarly well on an unlabeled data set
- Estimating the number of clusters (communities) in a graph is another open problem
- The current practice is to begin with an "educated guess" and iteratively re-apply the clustering algorithm with different inputs, until reasonable results are obtained
- This iterative process is very time-consuming and may be infeasible when dealing with very large data sets
- A suitably approximated starting point estimate may be very useful in streamlining this process
- Also, for algorithms that do not require it as input parameter, an estimate of the number of clusters provides an additional benchmark
- The eigengap heuristic has been suggested as a possible estimate for the number of clusters, but it is costly to compute
- Finding suitable approximations could improve the efficiency of analyses

Objectives

1. Define statistical measures and tests of clustering quality
2. Approximate the eigengap heuristic, to obtain an estimate of the number of communities
3. Use the approximate eigengap heuristic as performance benchmark

Clustering Quality

Arguably, a cluster (community) exhibits a high-level of interconnections between vertices within itself and a low-level of connections to vertices in the rest of the graph. Our statistical benchmarks are an attempt to formally measure the strength of the clusters, across the graph. Our goal is to formally test the null hypothesis that the algorithm's clustering is the result of a random assignment.

Variables

- The set of all clusters: $C = \{C_1, \ldots, C_\kappa\}$, with $|C| = \kappa$
- Total number of vertices in the graph: N
- Total number of vertices in cluster i: $|C_i| = n_i$
- Note that communities may overlap, so

$$\sum_{i=1}^{\kappa} n_i \geq N$$

- The set of all edges on the graph: $E = \{e_1, \ldots, e_m\}$, where $|E| = m$
- The sets of edges connecting two vertices in cluster i:
  $E_{i,i} \in \{E_{1,1}, E_{2,2}, \ldots, E_{\kappa,\kappa}\}$, where $|E_{i,i}| = m_{i,i}$
- The sets of edges connecting two vertices in two different clusters C_i and C_j:
  $E_{i,j} \in \{E_{1,2}, \ldots, E_{1,\kappa}, \ldots, E_{\kappa-1,\kappa}\}$, where $i \neq j$, and $|E_{i,j}| = m_{i,j}$
  (Note the symmetry, $E_{i,j} = E_{j,i}$)

Clustering Statistics

- Graph's connections ratio:

$$\bar{K} = \frac{|E|}{0.5 \times N(N-1)}$$

- Mean intra-cluster connections ratio:

$$\bar{K}_{\text{intra}} = \frac{1}{\kappa} \sum_{i=1}^{\kappa} \frac{|E_{i,i}|}{0.5 \times n_i(n_i - 1)}$$

- Mean inter-cluster connections ratio:

$$\bar{K}_{\text{inter}} = \frac{1}{0.5 \times \kappa(\kappa-1)} \sum_{i=1}^{\kappa} \sum_{j=i+1}^{\kappa} \frac{|E_{i,j}|}{0.5 \times \big((n_i + n_j)(n_i + n_j - 1) - n_i(n_i - 1) - n_j(n_j - 1)\big)}$$

Null Hypotheses

- $H_o^{(a)}$: The clustering is the result of a random assignment, therefore the mean intra-cluster connections ratio and the graph's connections ratio are statistically indistinguishable, or $\bar{K}_{\text{intra}} \approx \bar{K}$
  $H_a^{(a)}$: $\bar{K}_{\text{intra}} > \bar{K}$
- $H_o^{(b)}$: The clustering is the result of a random assignment, therefore the mean inter-cluster connections ratio and the graph's connections ratio are statistically indistinguishable, or $\bar{K}_{\text{inter}} \approx \bar{K}$
  $H_a^{(b)}$: $\bar{K}_{\text{inter}} < \bar{K}$

Hypothesis Tests

To test $H_o^{(a)}$, we use a t distribution with $\kappa - 1$ degrees of freedom. The test statistic in this case is

$$t_a = \frac{\bar{K}_{\text{intra}} - \bar{K}}{se(\text{Mean intra-cluster connections ratio})}$$

To test $H_o^{(b)}$, we use a t distribution with $0.5\kappa(\kappa-1) - 1$ degrees of freedom. The test statistic in this case is

$$t_b = \frac{\bar{K}_{\text{inter}} - \bar{K}}{se(\text{Mean inter-cluster connections ratio})}$$

Approximating the Eigengap Heuristic

- The eigengap heuristic offers a reasonable estimate of the number of clusters on a graph
- According to this heuristic, the number of communities is approximately the index at which a *spike* or *jump* occurs in the sorted eigenvalues (in ascending order)

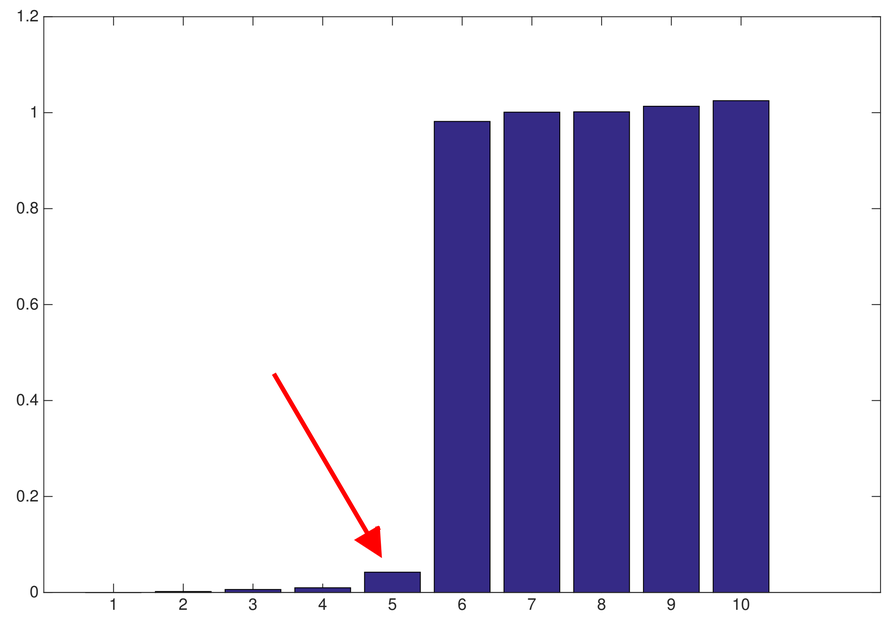

Figure 1: An educated guess is the graph contains FIVE Clusters

Eigengap Estimation via Random Sampling

- The eigengap heuristic requires the spectral decomposition of the Laplacian matrix (costly)
- We suggest approximating the eigengap via random sampling the adjacency matrix
- Using the bounds of the Gershgorin discs (line segments in this case) as an approximation has been suggested in the literature
- We compare our approximation technique, based on random sampling, to the one based on Gershgorin discs

Approximate Eigengap Heuristic as a Performance Benchmark

- An approximation of the eigengap heuristic can be used to obtain input parameters for clustering algorithms that require it
- It can also be used as a benchmark for the number of communities returned by an algorithm that does not require the number of communities as input

Conclusions

- We developed novel statistical benchmarks and tests for the quality of a graph's partitioning
- We proposed an approximation technique based on random sampling of the adjacency matrix
- Our technique appears to yield better results than a technique presented in the literature which is based on Gershgorin discs

Forthcoming Research

- This project only recently began, we are still in the process of formulating our research path & methodology
- Test on large data sets
- Explore dimensionality reduction of the Laplacian

Acknowledgements

PM offers thanks to Prof Derek Corneil of the University of Toronto Dept of Computer Science and Amit Bermanis of the University of Toronto Dept of Mathematics.

PM is supported by a MITACS-CIBC Accelerate grant.