Reaction Paper # 2: Spectral Clustering

Mohammad Elkholy  
Computer Science and Engineering  
 The American University in Cairo  
 New Cairo, Cairo, Egypt  
moelkholy@aucegypt.edu

SUMMARY

Clustering is an important technique to use in network exploration and analysis. It is important to be able to find tightly knit components in networks as they give us a lot of insight on how information flows between the different nodes and the topology of the network. It is thus important to have techniques that are able to identify these clusters as accurately as possible as efficiently as possible. It turns out that algorithms that are able to exploit structural information perform way better and are more efficient and as such it is important to understand the characteristics of different networks. Since networks in the real world come in varyingly different sizes, finding algorithms that are fast and reliable by exploiting their structures are particularly useful for very large networks. The first paper *Self-Similarity in the web* discusses an important structural feature of the world wide web. The second paper *On Clusterings: Good, Bad and Spectral* provides a bicriteria measure that can assess the quality of clusters produced by clustering algorithms as well as provide worst-case guarantees for a specific variant of spectral clustering.

**Self-Similarity in the web**

The authors in this paper examined several parameters of graphs over several subgraphs of the web. The parameters studied were the distributions of in and out-degrees, connected component sizes, bipartite cores and URL compressibility. These parameters were studied for several subgraphs: slices based on contest, slices based on relative location on the web, slices based on geographic location, random set of websites and a general subgraph of the web. These subgraphs were chosen to be cohesive clusters and not merely random. The authors found that many of these subgraphs generally behave in the same way for most of these parameters regardless of location on the web and that cohesive clusters provide the entire web with a backbone that can be used to navigate to any part of it, making it resistant to node deletion. This leads to the idea that the web is “fractal” in nature, i.e., that smaller subgraphs induced from the web graph behave in the same way as the web graph. This is observed for most of the parameters covered across the different subgraphs examined.

**On Clusterings: Good, Bad and Spectral**

Chart, scatter chart

Description automatically generatedThe authors start the paper by discussing a new measure to assess the quality of clusters produced by clustering algorithms. This new bicriteria measure tries to combat the problems that were faced by previously suggested measures like minimum diameter, k-center, k-median and minimum sum. A small example (Figure 1) was given for the minimum diameter measure to motivate the need for better measures. For a human it is obvious that the two clusters are separated by the dashed line (A), but by minimizing the maximum diameter of the points, we get the clustering given by the dashed line (B). This is also observed for many other measures with similar outcomes. They then introduce a parameter that measures the minimum quality of a cluster based on its conductance, and then a parameter that represents the ratio of the weights of edges not in the cluster compared to the overall edge weight. This parameter helps prevent clusters of low quality that maximizes .

The applicability of this measure was not discussed in the paper. They then provide two approximation algorithms that can be used to obtain an approximation function for the monotonic function governing the two parameters and , thus one does not need to know values for either of them before using them. The first approximation algorithm tries to find a cut that approximates the minimum conductance cut in the graph, then recurses on the pieces induced by the cut. As finding a cut of minimum conductance is hard, they resort to approximate cut algorithms with worst case guarantees. The one they used had a running time of with log n suppressed. This is rather slow for many real-world applications, which led them to the next approximation algorithm based on spectral clustering. They suggested a recursive version of the spectral clustering algorithm that approximates the best ratio cut based on the 2nd right eigenvector of the normalized matrix A, then recursing on the pieces induced by the cut. They then proved guarantees for the clusters produced.

**COMMENTS**

**Self-Similarity in the web**

This paper was rather simple in its goal, the authors wanted to find a structural property that can be exploited.   
For the most part they managed to show some practical evidence for some sense of self-similarity in the web. There are several problems I have with this paper, however. The first reason is that this paper is very old. The world wide web has evolved to a much greater extent since this paper was published. The whole worldwide web manifests itself in a completely different way now, with websites like YouTube, Wikipedia, Google, Facebook, and many other social media websites being the backbone that connects it together. However, the communities within these websites need not mimic the web, and thus this “fractal” nature of the web needs to be examined again. It would be particularly interesting to see if the fractal nature still exists or if there is some other structure in place, and if there is a way to model the change from the fractal nature first found to the new structure, if a new one exists. The next problem I find with the paper is the way some of the subgraphs were induced. Firstly, the subgraph induced by content was made by picking webpages that shared a set of keywords. However, keywords used like “math”, “mp3” and “restaurant” does not mean that these subgraphs are cohesive. It is not very hard to realize that such words are very vague and are by no means keywords that would connect a community together but would connect so many different webpages that simply mention them but have nothing to do with each other. A better keyword would be something domain specific, an example would be the word “particles” which would connect a community of physicists and chemists. This would be considered a cohesive community and might behave differently. The subgraphs made based on common keywords were basically redundant as they are very hardly any different than a random set of webpages, which they induced in another group. The next problem I have is the use of the bipartite core parameter. It was randomly introduced, without any justifications for the values of 5, 7 that they picked for this parameter, and it was not discussed at all in the analysis of their results. They only stated that higher its value, the less well defined a community is, but even then, they did not specify a value that would be used as a benchmark. In the tables provided, the values for this parameter ranged very widely between 22 to 410, with no explanations for that. Another problem is in the crawling of pages from websites, limiting the number of pages crawled from any specific site was not explained, this is particularly problematic as sites may have many pages with very little degrees, and many of the important pages on the side (specifically ones with higher in and out degrees) might be excluded. No justification was also given for why they only specifically only took 12K pages. The paper leaves one with more questions than answers.

**On Clusterings: Good, Bad and Spectral**

This paper was very hard to read to say the least, as it was highly theoretical and as such, I do not really have much to say about the methods they used to derive the worst case guarantees for their algorithm. I do however appreciate the idea behind the bicriteria measure for assessing the quality of clusters and the need for better measurements that do not simply work on a specific dataset. Although they managed to provide a proof for the clustering that can be achieved by the approximate clustering algorithms a bound on the number of misclassifications, the criteria for the number of misclassifications are very specific and might not be very applicable for many networks. It is worth noting that since this is a theoretical paper, there is no data about how the algorithms and bicriteria measure fair next to other measures and algorithms. Another problem is that the number of clusters was not really addressed. Typically, spectral clustering requires the number of clusters to be known beforehand, which is a big disadvantage especially in very complex networks where clusters might not be as obvious.

**CONCLUSIONS**

It is important to understand the underlying structure of networks, particularly ones with very large numbers of nodes as the algorithms are typically very computationally intensive and finding ways to exploit their structures can go a long way in decreasing the computational complexity and even improving the quality and reliability of clustering algorithms used. It is also very important to find measures that can better assess the quality of clusters.

REFERENCES

[1] Stephen Dill, Ravi Kumar, Kevin S. Mccurley, Sridhar Rajagopalan, D. Sivakumar, and Andrew Tomkins. 2002. Self-similarity in the web. ACM Trans. Internet Technol. 2, 3 (August 2002), 205–223. https://doi.org/10.1145/572326.572328

[2] Ravi Kannan, Santosh Vempala, and Adrian Vetta. 2004. On clusterings: Good, bad and spectral. J. ACM 51, 3 (May 2004), 497–515. https://doi.org/10.1145/990308.990313

Conference Name:ACM Woodstock conference

Conference Short Name:WOODSTOCK’18

Conference Location:El Paso, Texas USA

ISBN:978-1-4503-0000-0/18/06

Year:2018

Date:June

Copyright Year:2018

Copyright Statement:rightsretained

DOI:10.1145/1234567890

RRH: F. Surname et al.

Price:$15.00