Evaluating the Quality of Clustering Algorithms using Cluster Path Lengths

Faraz Zaidi, Guy Melançon

Evaluating the Quality of Clustering Algorithms using Cluster Path Lengths

Faraz Zaidi, Daniel Archambault, and Guy Melançon

CNRS UMR 5800 LaBRI & INRIA Bordeaux - Sud Ouest
351, cours de la Libération
33405 Talence cedex, FRANCE
{faraz.zaidi, guy.melancon}@labri.fr
daniel.archambault@inria.fr

Abstract. Many real world systems can be modeled as networks or graphs. Clustering algorithms that help us to organize and understand these networks are usually referred to as, graph based clustering algorithms. Many algorithms exist in the literature for clustering network data. Evaluating the quality of these clustering algorithms is an important task addressed by different researchers. An important ingredient of evaluating these clustering techniques is the node-edge density of a cluster. In this paper, we argue that evaluation methods based on density are heavily biased to networks having dense components, such as social networks, but are not well suited for data sets with other network topologies where the nodes are not densely connected. Example of such data sets are the transportation and Internet networks. We justify our hypothesis by presenting examples from real world data sets.
We present a new metric to evaluate the quality of a clustering algorithm to overcome the limitations of existing cluster evaluation techniques. This new metric is based on the path length of the elements of a cluster and avoids judging the quality based on cluster density. We show the effectiveness of the proposed metric by comparing its results with other existing evaluation methods on artificially generated and real world data sets.

Key words: Evaluating Cluster Quality, Cluster Path Length

1 Introduction

Many real world systems can be modeled as networks or graphs where a set of nodes and edges are used to represent these networks. Examples include social networks, metabolic networks, world wide web, food web, transport and Internet networks. *Community detection* or *Clustering* remains an important technique to organize and understand these networks [6] where [22] provides a good survey of graph based clustering algorithms. A cluster can be defined as a group of elements having the following properties as described by [24]:

– Density: Group members have many contacts to each other. In terms of graph theory, it is considered to be the ratio of the number of edges present in a group of nodes to the total number of edges possible in that group.

- Separation: Group members have more contacts inside the group than outside.
- Mutuality: Group members choose neighbors to be included in the group. In a graph-theoretical sense, this means that they are adjacent.
- Compactness: Group members are 'well reachable' from each other, though not necessarily adjacent. Graph-theoretically, elements of the same cluster have short distances.

The Density of a cluster can be measured by the equation $d = e_{actual}/e_{total}$ where e_{actual} represents the actual number of edges present in the cluster and e_{total} represents the total number of possible edges in the cluster. Density values lie between [0,1] where a value of 1 suggests that every node is connected to every other node forming a clique.

The Separation can be calculated by the number of edges incident to a cluster, i.e the number of edges external to the clusters. This is often referred to as the cut size and can be normalized by the total number of edges incident to the cluster. Low values represent that the cluster is well separated from other clusters where high values suggest that the cluster is well connected to other clusters.

Mutuality and Compactness of a cluster can easily be evaluated using a single quantitative measure: the average path length between all the nodes of a cluster. The path length refers to the minimum number of edges connecting node A to node B. The average path length represents how far apart any two nodes lie to each other and is calculated by taking the average for all pairs of nodes. This value can be calculated for a cluster giving us the average path length of a particular cluster. Low values indicate that the nodes of a cluster lie in close proximity and high values indicate that the cluster is sparse and its nodes lie distant to each other.

Cluster Detection has a wide range of applications in various fields. For example, in social networks, community detection could lead us towards a better understanding of how people collaborate with each other. In a transport network, a community might represent cities or countries well connected through transportation means. There are many algorithms addressing the issue of clustering and readers are referred to various surveys on the topic [22, 9, 3] for further information. Evaluating different clustering algorithms remains essential to measure the quality of a given set of clusters. These evaluation metrics can be used for the identification of clusters, choose between alternative clusterings and compare the performance of different clustering algorithms [22].

Most of the evaluation metrics consider *density* as a fundamental ingredient to calculate the quality of a cluster. From the definition of clusters given above, density is an important factor but not the only factor to be considered while evaluating the quality of clustering. Having a densely connected set of nodes might be a good reflection of nodes being adjacent to each other or lying at short distances but the inverse conjecture might not necessarily be true as illustrated in Fig. 1. Consider the set of five nodes in Fig. 1(a,b,c) being identified as clusters by some clustering algorithm. The density of graph in Fig. 1(a) is 1 and that of (b) and (c) is 0.4. Intuitively (b) is more cohesive than (c). Moreover the

average path length of (b) is lower than that of (c) suggesting that the elements of cluster (b) are closer to each other. From this example, we can deduce that, if we consider density as the only criteria, then for such an evaluation metric, (b) and (c) will be assigned a similar value which is not consistent with Mutuality and Compactness.

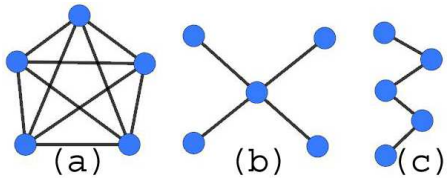

Fig. 1. (a) Represents a *clique* (b) presents a *star-like* structure and (c) is a set of nodes connected to each other in a *chain-like* structure.

Another important class of evaluation metric uses connectivity of clusters to capture the notion of Separation. The simplest way to measure this is the *cut size* which is defined as the minimum number of edges required to be removed so as to isolate a cluster. Consider the graphs in Fig. 2(a,b,c) with enclosed nodes representing clusters. Calculating the cut size for all these clusters will give the same cut size, which is 1 in these examples, as each cluster is connected to the rest of the graph through exactly one edge. The example suggests that cut-size alone is not a good representation of the quality of clustering as all the clusters in Fig. 2 have the same cut-size.

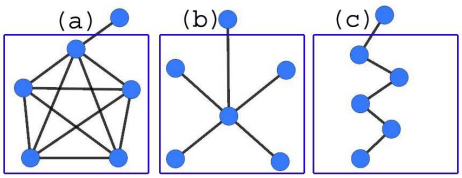

Fig. 2. Represents three graphs with enclosed nodes being the clusters. All the clusters have the same *cut size* which is equal to 1. Based on the *cut size* alone the quality of the clustering cannot be judged.

More sophisticated measures combining *density* and *cut size* have been investigated with the most important example being *relative density* [13]. Even combining these two metrics, the clusters in Fig. 2(b) and (c) will be assigned an equal score, failing to incorporate Mutuality and Compactness of a cluster.

Calculating the density and the cut size of these two clusters will result in the exact same value. We present other cluster evaluation techniques in Sect. 2.

If we consider Density, Mutuality and Compactness together to evaluate the quality of clusters present in Fig. 2, the highest measure should be associated to cluster (a) as it is the cluster with the highest Density, Mutuality and Compactness. Then cluster (b) where it has high Mutuality and Compactness but low density and finally cluster (c) which is the least Dense, Mutual and Compact cluster of the three clusters present in Fig. 2. We show that the existing cluster evaluation metrics do not evaluate the quality of clusters in this order. We discuss the details in Sect. 4.

Until now, we have argued that ignoring Mutuality and Compactness of a cluster to evaluate its quality can give inconsistent results. A simple question can be raised about the importance of these two criterion especially for real world data sets. To answer this question, we turn our focus towards some real world data sets. Consider the example of an **Air Traffic Network** which represents an airport-airport graph where two airports are connected through an edge if a direct flight exists between them [21]. In this particular case, we took Hong Kong as an example by taking some airports directly connected to it as shown in Fig. 3[1]. On one side, we can see some of the world's biggest cities having direct flights to Hong Kong where on the other hand, we have lots of regional airports also directly connected to Hong Kong. If we consider a cluster by putting Hong Kong with the regional airports, the resulting cluster will have very low density and high cut size which are undesirable features for a cluster. In the other case, where we consider Hong Kong as part of the cluster with the biggest cities in the world, the cluster with Hong Kong will have a high cut size. Moreover, the regional airports could not be clustered together as they will no longer remain connected to each other. We will end up with lots of singleton clusters which again will reduce the overall quality of any clustering algorithm.

Another example of these star-like structures comes from **Internet Tomography Networks** which is a collection of routing paths from a test host to other networks on the Internet. The database contains routing and reachability information, and is available to the public from the Opte Project website (`http://opte.org/`). Considering two hubs from this data set and taking all the nodes lying at distance five from these hubs, we obtain a structure as shown in Fig. 4. The two hubs dominate the number of connections in these networks presenting the *star-like* behavior in real world data sets.

As opposed to these *star-like* structures, the other most common structure present in most real world data sets is the presence of *cliques*. Social networks are good examples of networks having cliques. As an example data set, consider the collaboration network of researchers usually called the **Co-Authorship Network** [18]. Two authors are connected by an edge if they appear as authors in an article. Scientists co-authoring an article will end up having edges with every

[1] All the images in this paper are generated using TULIP software which is an open source software for the analysis and visualization of large size networks and graphs available at: http://www.tulip.labri.fr/.

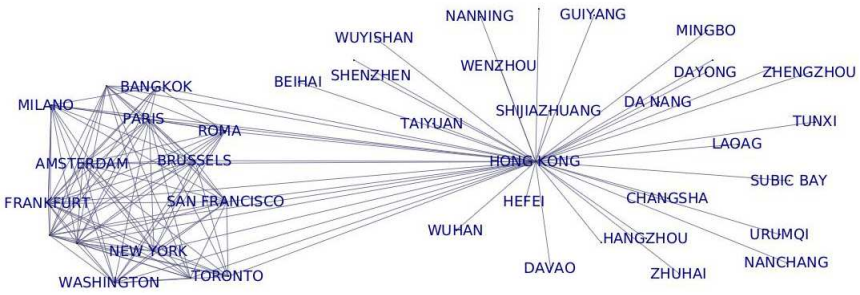

Fig. 3. Air Traffic network drawn using Hong Kong at the center and some airports directly connected to Hong Kong. We can see the worlds most important cities having a direct flight to Hong Kong whereas there are lots of regional airports connected to Hong Kong representing a *star-like* structure as discussed previously in Fig. 1(b) and 2(b).

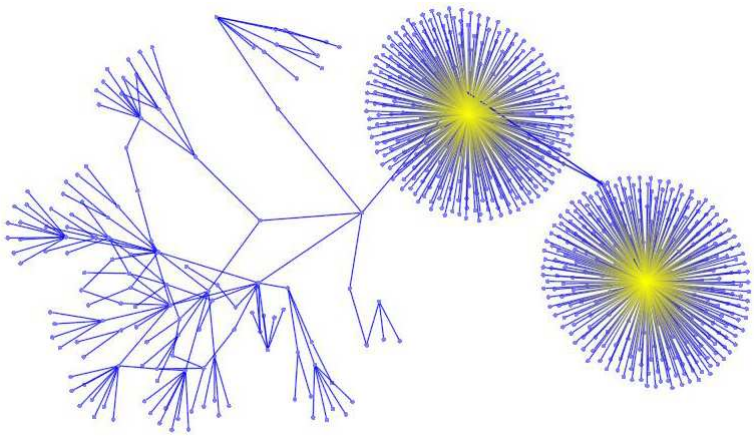

Fig. 4. Internet Tomography Network representing routing paths from a test host to other networks. Two nodes clearly dominate the number of connections as they play the role of hubs to connect several clients. Another example of *star-like* structures in the real world.

other co-author thus forming a clique. Another example of such a network is the Movie network where two actors are connected to each other if they have acted in a movie together [1]. Just as in the case of co-authorship network, actors appearing in a movie together will form a clique and thus represent dense communities.

Metrics based on density and cut size prove to be adequate for networks having densely connected nodes or cliques. Results have shown that different clustering algorithms perform well for these networks [6, 16, 1]. On the other hand, in case where lots of star-like structures exist (see Fig. 3 and 4), an evaluation based on density and cut size fails to perform well as shown in the examples discussed previously. To resolve this problem, we propose a new cluster evaluation metric which takes into account the underlying network structure by considering the average path lengths to evaluate the cluster quality.

Apart from these cliques and star-like structures, other interesting topologies exist in different data sets but are highly dependent on the application domain. Examples include motifs in Chemical Compounds [4] or Metabolic Networks [11] where the goal is to search motifs in graphs and not to cluster them based on some similarity. We focus our attention to generic data sets and evaluating clustering algorithms for specific data sets remains out of the scope of this paper.

The design principle for the proposed metric is very simple and intuitive. Instead of considering *density* as the fundamental component to evaluate the quality of a clustering algorithm, we use the average path length to determine the closeness of the elements of a cluster. It is obvious that in case of a clique, the path length between the nodes is 1 which is the minimum possible value for two connected nodes. But the important aspect here is that a star-like structure will have a higher average path length as compared to a chain like structure thus providing a way to evaluate how close the nodes are of a cluster, irrespective of the density of edges. We discuss the details of the proposed metric further in Sect. 3.

The paper is organized as follows. In the following section, we provide a brief overview of some widely used metrics to evaluate cluster quality. In Sect. 3 we present the proposed metric and we discuss our findings by performing a comparative study of the different evaluation metrics in Sect. 4. Finally in Sect. 5 we present our conclusions and future research directions in light of the newly proposed metric.

2 Related Work

The different approaches to evaluate cluster quality can be classified as *external*, *relative* or *internal*. The term *external* validity criteria is used when the results of the clustering algorithm can be compared with some pre-specified clustering structures [7] or in the presence of ground truth [20]. *Relative* validity criteria measure the quality of clustering results by comparing them with the results of other clustering algorithms [12]. *Internal* validity criteria involve the development of functions that compute the cohesiveness of a clustering by using density,

cut size, distances of entities within each cluster, or the distance between the clusters themselves etc [14, 19, 8].

For most real world data sets, an external validity criteria is simply not available. In the case of relative validity criteria, as Jain[9] argues, there is no clustering technique that is universally applicable in uncovering the variety of structures present in multidimensional data sets. Thus we do not have an algorithm that can generate a bench mark clustering for data sets with varying properties. For these reasons we focus our attention on internal quality metrics only. Further more, we deal with quality metrics for partitional or flat clustering algorithms that are non-overlapping.

Modularity(Q) [16] (Q metric) is a metric that measures the fraction of the edges in the network that connect within-community edges minus the expected value of the same quantity in a network with the same community divisions but random connections between the vertices. If the number of within-community edges is no better than random, we will get Q = 0. Values approaching Q = 1, which is the maximum, indicate strong community structure.

Another metric used by Auber *et al.* [1] to effectively evaluate the quality of clustering for small world graphs is the MQ metric initially proposed by Mitchell *et al.* [15] as a partition cost function in the field of software reverse engineering. It comprises of two factors where the first term contributes to the positive weight represented by the mean value of edge density inside each cluster. The second term contributes as a negative weight and represents the mean value of edge density between the clusters.

The Relative Density [13] of a cluster calculates the ratio of the edge density inside a cluster to the sum of the edge densities inside and outside that cluster. The final Relative Density is the averaged sum of the these individual relative densities for all clusters.

For our experimentation and comparison, we use the three metrics presented above. Other notable metrics used to evaluated the quality of clustering include coverage [2], conductance [10], performance [2] but since they are based on more or less the same principles to evaluate the quality of clusterings, we do not include them in this study.

3 Proposed Metric For Cluster Evaluation: Cluster Path Lengths

As we discussed earlier, the design principle which makes our metric novel, is the fact that we consider the path length of elements of a cluster. The metric is composed of two components, the positive component($M^+(G)$) which assigns a positive score to a cluster and a negative component($M^-(G)$) which attributes a negative score to edges between clusters. The positive score is assigned on the basis of the density, compactness and mutuality of the cluster whereas the negative score is assigned on the basis of the separation of the cluster from other clusters. The final quality of a cluster is simply the sum of the two components given by the equation:

$$M(G) = M^{+}(G) - M^{-}(G) \tag{1}$$

In the above equation, the two components are weighted equally. An option can be to assign different weights to the two components, for example a higher weight to the positive component, for the sake of simplicity, we have not experimented with different weights. We discuss the details of how the positive and the negative components are calculated below.

3.1 Positive Component:

The goal is to assign a quantitative value to a cluster based on its density, compactness and mutuality. Looking at the different clusters in Fig. 2, if we calculate the average path length of the nodes within the cluster, the least value would be assigned to cluster (a), then cluster (b) and finally (c). This is quite intuitive as we reduce the average distance between nodes of a cluster, the density tends to increase. Lets call the average path length of each cluster Cluster Path Length. The best possible average path length for any cluster can be 1 in the case when every node is connected to every other node forming a clique. The normalized cluster path length can be given by the following equation:

$$CPL_i = \frac{1}{AvgPathLen_i} \tag{2}$$

Where CPL_i represents the normalized cluster path length of cluster i and $AvgPathLen_i$ represents the average path length of the nodes in cluster i. Higher this value is for a cluster, better is the quality of the cluster where the values lie in the range of [0,1]. The overall cluster path length is then averaged for all clusters where k is the total number of clusters, giving us the value for the positive component to evaluate the quality of the clustering:

$$M^{+}(G) = CPL_{1\cdots k} = \frac{1}{k}\sum_{i=1}^{k} CPL_i \tag{3}$$

3.2 Negative Component:

The next step is to assign a negative score to penalize the inter-cluster edges. The value of M^{-} evaluates the separation of the two clusters. This score is calculated for each pair of clusters and is based on the number of edges that link two clusters i and j compared to the total number of edges possible between these two clusters. Let n_i and n_j be the number of nodes contained in clusters i and j respectively. Therefore, the edge penalty for the edges present between these two cluster would be given by the equation:

$$EdgePenalty_{(i,j)} = \frac{e_{ij}}{n_i * n_j} \tag{4}$$

Where e_{ij} is the number of edges present between clusters i and j. The overall Edge Penalty ($M^-(G)$) is the average calculated for all pair of clusters given by the equation:

$$M^-(G) = \frac{2}{k*(k-1)} \sum_{i=1,j=1}^{k} EdgePenalty_{(i,j)} \quad where(i \neq j) \tag{5}$$

The negative score sums all edge penalties over all pairs of clusters and then normalizes the value by $k(k-1)/2$ to produce an overall penalty in the range [0,1]. This value is linearly proportional to the number of edges present between clusters where low values correspond to few broken edges and a better clustering quality.

To summarize the proposed metric, we use the cluster path lengths to assign a positive score to evaluate the quality of clustering subtracted by a negative score which is based on the inter-cluster density. The values lie in the range of [0,1] where low values indicate poor clustering and high values indicate better clustering. We refer to the metric as *CPL* for Cluster Path Lengths (although we subtract the Edge penalties from the CPLs calculated).

4 Experimentation

For evaluating different cluster quality metrics, we use two different experiments. The first, where we generate artificial data sets and the second where we use real world data sets.

4.1 Artificial and Clustered Data Set

For the artificial data set, we directly generate clusters to avoid biasing the experiment using any particular clustering algorithm. We generate three clustered graphs of size n. We generate a random number k between 1 and Max to determine the size of a cluster. For the first graph, we add k nodes such that each node is connected to the other forming a *clique*. For the second graph, k nodes are added such that a *star-like* is formed and finally k nodes are added to the third graph forming a chain like structure. The process is repeated until the maximum number of nodes in the graphs reach n. The clusters in each of these graphs are connected by randomly adding $RandE$ edges. This number decides the number of inter-cluster edges that will be produced for each graph. The choice of selecting the variables n, Max and $RandE$ are independent of the experiment and do not change the final evaluation. For our experiment, we used $n = 200$, $Max = 20$ and $RandE = 40$.

Two important inferences can be drawn from the experiment described above. The first, where we compare how the different evaluation metrics perform for

Cluster Quality Metric	**Cliques**	**Star-like**	**Chain-like**
Cluster Path Length	0.998	0.611	0.374
Q metric	0.975	0.281	0.281
MQ metric	0.998	0.844	0.844
Relative Density	0.862	0.711	0.711

Table 1. Evaluating the quality of clustering using three topologically different and artificially generated clustered data sets.

evaluating the quality of clusters where each cluster is a clique with some inter-cluster edges. Looking at the high values for the all the evaluation metrics, we can justify that all the metrics are consistent in evaluating the quality of clusters including the newly proposed metric. As discussed previously, density based metrics perform well when the clusters are densely connected, and so does the proposed metric.

The other important result can be derived by comparing the values assigned to the *star-like* clusters and *chain-like* clusters by different evaluation metrics. Clearly the other metrics fail to differentiate between how the edges are distributed among the clusters ignoring the Mutuality and Compactness of a cluster whereas CPL does well by assigning higher values to *star-like* clusters as compared to *chain-like* clusters. This justifies the use of cluster path length as a metric to evaluate the quality of clusters specially where dense clusters are not expected.

4.2 Real World Data Sets and Clustering Algorithms

The second experiment uses real world data sets. We use four different data sets, two of them were briefly introduced earlier in Sect. 1. We give the source and description of each data set below.

The Co-authorship network is network of scientists working on network theory and experiments, as compiled by M. Newman in May, 2006 [18]. The network was compiled from the bibliographies of two review articles on networks, M. E. J. Newman, SIAM Review and S. Boccaletti et al., Physics Reports, with a few additional references added by hand. The biggest connected component is considered for experimentation which contains 379 nodes and 914 edges.

The Air Transport Network is an undirected graph where nodes represent airports and edges represent a direct flight from one airport to the other. The network contains 1540 nodes and 16523 edges. The node-edge density of the graph indicates that the average degree of node is around 10, but actually the graph follows a scale free degree distribution where some nodes have very high degree and many nodes have low degree (see [21] for more details). This is quite understandable because the worlds busiest airports like Paris, New York, Hong Kong, London etc have flights to many other destinations and small cities or regional airports have very restricted traffic as shown in Fig. 3.

The Internet network is a network mapping data which consists of paths from a test host towards other networks on the Internet containing routing and reach-

ability information. The complete data set is available from the Opte Project website (`www.opte.org`). The entire data set contains 35836 nodes and 42387 edges. Since the Divisive Clustering algorithm has a high time complexity, we only consider a subset of the actual data set constructed by considering a hub and the nodes connected at distance 5 from it. The subset consists of 1049 nodes and 1319 edges.

The fourth data set is a Protein-Protein interactions network. The data represents a set of S. cerevisiae interactions identified by TAP purification of protein complexes followed by mass-spectrometric identification of individual components used by [5]. The data is available from `http://dip.doe-mbi.ucla.edu/dip` and contains 1246 nodes and 3142 edges. Around 80 nodes were disconnected from the biggest connected component and were removed for this experimentation.

The choice of Air Traffic, the Internet Tomography and the Protein network is purely based on the fact that these networks do not have densely connected components. Rather there are components that have chain-like structures and star-like structures. On the other hand we use the co-authorship network to show the efficiency of the clustering algorithms used as they perform well in detecting communities present in the network.

To cluster these data sets, we use two known clustering algorithms, the Bisecting K-Means algorithm [23] and the Divisive Clustering algorithm based on Edge Centrality [6]. The choice of these algorithms is based on the criteria that these algorithms do not try to optimize or influence the clustering algorithm based on the density or some other cluster quality metric as compared to other algorithms present in the literature such as [17]. We also use the Strength Clustering algorithm proposed by [1] which was initially introduced to cluster social networks. The algorithm has been shown to perform well for the identification of densely connected components as clusters.

The Bisecting K-Means algorithm and the Divisive Clustering algorithm based on Edge Centrality are both divisive algorithms, i.e. they start by considering the entire graph as a single cluster and repeatedly divide the cluster into two clusters. Both these algorithms can be used to create a hierarchy where the divisive process stops when each cluster has exactly one node left. Instead of generating the entire hierarchy, we stop the process as soon as the minimum number of nodes in the cluster reaches around 20 nodes. Moreover since we do not propose a method to evaluate the quality of a hierarchical clustering algorithm, we consider the leaves as a single partitional clustering. Note that the clustering algorithm might create singletons but while evaluating the quality of clusters we do not consider clusters having a single element. The results for evaluating the clusters obtained for the two data sets are given in Table 2.

The Strength clustering algorithm uses the strength metric for clustering. This metric quantifies the neighborhoods cohesion of a given edge and thus identifies if an edge is an intra-community or an inter-community edge. Based on these strength values, nodes are judged to be part of the same cluster (see [1] for more details). The reason for using this clustering algorithm is to demon-

strate that irrespective of the clustering algorithm, the CPL metric evaluates the quality of a clustering. Since the other two algorithms do not force the detection of strongly connected components, we use Strength clustering as a representative of clustering algorithms that try to detect densely connected nodes.

		Cluster Quality Metric			
Data Set	Clustering Algorithm	CPL	MQ	Q	Relative Density
Co-Authorship	Divisive Clustering	0.672	0.531	0.772	0.630
	Bisecting K-Means	0.589	0.425	0.775	0.636
	Strength Clustering	0.846	0.832	0.264	0.232
Air Traffic	Divisive Clustering	0.614	0.399	0.093	0.105
	Bisecting K-Means	0.499	0.238	0.012	0.122
	Strength Clustering	0.676	0.528	0.024	0.078
Internet	Divisive Clustering	0.498	0.324	0.790	0.697
	Bisecting K-Means	0.581	0.415	0.592	0.582
	Strength Clustering	0.666	0.503	0.356	0.554
Protein	Divisive Clustering	0.527	0.315	0.638	0.498
	Bisecting K-Means	0.595	0.410	0.336	0.316
	Strength Clustering	0.683	0.529	0.165	0.291

Table 2. Evaluating the quality of clustering real world data sets using the existing and the proposed cluster evaluation technique.

Analyzing the results presented in Table 2, first we look at the Co-authorship network. The high values of the Divisive algorithm for all the evaluation metric suggest that the algorithm does well to find the good clusters. Bisecting K-Means seem to perform quite well also for this data set although values for the CPL and MQ metric are comparatively lower than the divisive algorithm. Looking at the results of Strength Clustering using CPL and MQ, the values are quite high indicating that the algorithm found high quality clusters but the low Q metric and Relative Density values create some doubt about the performance of the algorithm. This variation is due to the large number of clusters generated by Strength clustering (122) as compared to Divisive (23) and Bisecting K-Means (38) algorithm. While evaluating the quality using Q metric and Relative Density, this high number of clusters reduces its quality as it results in high number of inter-cluster edges.

In case of the Air Traffic network, the clusterings generated by the Bisecting K-Means and Divisive algorithms are relatively poorly judged as compared to the CPL and MQ metric. This is a clear indication that when considering the star-like structures as clusters which are present in abundance in the Air-Traffic network, the evaluation metrics judge the performance of the clustering algorithms to be poor. This is because there are not many densely connected airports in the network. High values of CPL indicate that even though, the clusters are not densely connected, they lie in close proximity and thus are judged to be good clusters. The overall node-edge density plays an important role as well since

the entire network has a high node-edge density, Q metric and Relative Density expect highly dense clusters to be found and their absence results in low values for these metrics. As mentioned in the introduction, there are a few nodes that have a very high number of connections, airports such as Paris, London and New York, which increases the overall density of the network, but most of the airports have a very low number of connections. Thus many clusters found are representatives of regional or with-in country airports connecting all its cities, as shown in Fig 3. These results are a good justification of why the CPL is a good cluster evaluation metric as it does not rate the quality of such clusters poorly as compared to the other metrics.

Next, we look at the Internet Network. Almost all the evaluation metrics rate the quality of clustering highly for the three clustering algorithms except for the Strength clustering-Q metric value. Again, we refer to the overall node-edge density of this graph which is quite low. Due to this, Q metric and Relative Density do not expect highly dense clusters and thus even though there are lots of star-like clusters found in this network, their quality is rated as good.

Finally the analysis of the Protein network is quite close to that of the Airport network. The overall density is not that high, but still the node-edge ratio is 1:3. The network is a good mix of some highly dense clusters and some star-like and/or chain-like clusters. The strength algorithm again generates a very high number of clusters (169) as compared to Divisive (91) and Bisecting K-Means (117). The divisive algorithm has the lowest number of clusters and thus has relatively high Q metric and Relative Density values.

For all the different data sets and algorithms, the CPL metric assigns high values consistently. This is an indication that by definition and from previous experimental results on a wide variety of data sets, these algorithms perform well in grouping similar items together. The Q metric and the Relative density are heavily dependent on the overall node-edge density for the evaluation of a clustering. In case of high node-edge density, these metrics expect highly dense clusters and in case of low node-edge density, less dense clusters can be rated as high quality irrespective of the underlying cluster topology, where we have argued that Mutuality and Compactness should be taken into consideration. The CPL metric is consistent with algorithms and dense data sets where tightly connected clusters are expected as is the case with the co-authorship network and to some extent, the protein network.

We would like to mention that the experimentation and the results described in this paper compare different cluster evaluation techniques and should not be generalized to compare the different clustering algorithms. This is because the number of clusters and their sizes vary from one clustering algorithm to the other. Specially, Bisecting K-Means and Divisive Clustering based on Edge Centrality can not be compared with the Strength clustering algorithm in terms of performance and quality of clusters generated as strength clustering generates many small size clusters as compared to the other clustering algorithms.

5 Conclusion and Future Research Directions

In this paper we introduced a new metric called the CPL metric to evaluate the quality of clusters produced by clustering algorithms. We argued that Density and Cut Size based metrics play an important role in the evaluation of dense graphs but Mutuality and Compactness are also important for the evaluation of clusters in graphs that are not densely connected. The proposed metric takes into account the underlying network structure and considers the average path length as an important factor in evaluating the quality of a cluster. We evaluated the performance of some existing cluster evaluation techniques showing that the new metric actually performs better than the metrics used largely by the research community.

As part of future work, we intend to extend the metric to evaluate the quality of hierarchical clustering algorithms based on the principles introduced in this paper. A more extended study is needed to compare different clustering algorithms for data sets having varying network topologies to comprehend the behavior of different clustering algorithms which in turn can lead us towards a better understanding of how to judge these algorithms.

References

1. D. Auber, Y. Chiricota, F. Jourdan, and G. Melancon. Multiscale visualization of small world networks. In *INFOVIS '03: Proceedings of the IEEE Symposium on Information Visualization*, pages 75–81, 2003.
2. U. Brandes and T. Erlebach. *Network Analysis : Methodological Foundations (Lecture Notes in Computer Science)*. Springer, March 2005.
3. U. Brandes, M. Gaertler, and D. Wagner. Engineering graph clustering: Models and experimental evaluation. *ACM Journal of Experimental Algorithmics*, 12, 2007.
4. D. G. Corneil and C. C. Gotlieb. An efficient algorithm for graph isomorphism. *Journal of the ACM (JACM)*, 17:51–64, 1970.
5. A.-C. Gavin, M. Bosche, R. Krause, P. Grandi, M. Marzioch, A. Bauer, J. Schultz, J. M. Rick, A.-M. Michon, C.-M. Cruciat, M. Remor, C. Hofert, M. Schelder, M. Brajenovic, H. Ruffner, A. Merino, K. Klein, M. Hudak, D. Dickson, T. Rudi, V. Gnau, A. Bauch, S. Bastuck, B. Huhse, C. Leutwein, M.-A. Heurtier, R. R. Copley, A. Edelmann, E. Querfurth, V. Rybin, G. Drewes, M. Raida, T. Bouwmeester, P. Bork, B. Seraphin, B. Kuster, G. Neubauer, and G. Superti-Furga. Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature*, 415(6868):141–147, January 2002.
6. M. Girvan and M. E. J. Newman. Community structure in social and biological networks. *Proc. Natl. Acad. Sci. USA*, 99:8271–8276, 2002.
7. M. Halkidi, Y. Batistakis, and M. Vazirgiannis. Cluster validity methods: Part i. *ACM SIGMOD Record*, 31:2002, 2002.
8. M. Halkidi and M. Vazirgiannis. Clustering validity assessment: Finding the optimal partitioning of a data set, 2001.
9. A. K. Jain, M. N. Murty, and P. J. Flynn. Data clustering: a review. *ACM Comput. Surv.*, 31(3):264–323, 1999.
10. R. Kannan, S. Vempala, and A. Vetta. On clusterings good, bad and spectral. *Journal of the ACM*, 51 (3):497–515, 2004.

11. V. Lacroix, C. Fernandes, and M.-F. Sagot. Motif search in graphs: Application to metabolic networks. *Computational Biology and Bioinformatics, IEEE/ACM Transactions on*, 3(4):360–368, Oct.-Dec. 2006.
12. O. Maimon and L. Rokach. *Data Mining and Knowledge Discovery Handbook*. Springer, September 2005.
13. M. Mihail, C. Gkantsidis, A. Saberi, and E. Zegura. On the semantics of internet topologies, tech. rep. gitcc0207. Technical report, College of Computing, Georgia Institute of Technology, Atlanta, GA, USA, 2002.
14. G. W. Milligan. A monte-carlo study of 30 internal criterion measures for cluster-analysis. *Psychometrica*, 46:187–195, 1981.
15. B. MITCHELL, M. S., Y.-F. C., and G. E. Bunch: A clustering tool for the recovery and maintenance of software system structures. In *International Conference on Software Maintenance, ICSM.*, 1999.
16. M. E. Newman and M. Girvan. Finding and evaluating community structure in networks. *Phys Rev E Stat Nonlin Soft Matter Phys*, 69(2 Pt 2), February 2004.
17. M. E. J. Newman. Fast algorithm for detecting community structure in networks. *Physical Review E*, 69:066133, 2004.
18. M. E. J. Newman. Finding community structure in networks using the eigenvectors of matrices. *Physical Review E (Statistical, Nonlinear, and Soft Matter Physics)*, 74(3), 2006.
19. Q. H. Nguyen, Rayward, and V. J. Smith. Internal quality measures for clustering in metric spaces. *Int. J. Bus. Intell. Data Min.*, 3(1):4–29, 2008.
20. W. M. Rand. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, 66(336):846–850, 1971.
21. C. Rozenblat, G. Melançon, and P.-Y. Koenig. Continental integration in multilevel approach of world air transportation (2000-2004). *Networks and Spatial Economics*, 2008.
22. S. E. Schaeffer. Graph clustering. *Computer Science Review*, 1(1):27–64, August 2007.
23. M. Steinbach, G. Karypis, and V. Kumar. A comparison of document clustering techniques. Technical report, Departement of Computer Science and Engineering, University of Minnesota, 2000.
24. S. Wasserman and K. Faust. *Social Network Analysis: Methods and Applications*. Cambridge University Press, Cambridge, 1994.