# QGraph: A quality assessment index for graph clustering

Maria Halkidi[1] and Iordanis Koutsopoulos[2]

[1] University of Piraeus, mhalk@unipi.gr
[2] Athens University of Economics and Business, jordan@aueb.gr

**Abstract.** In this work, we aim to study the cluster validity problem for graph data. We present a new validity index that evaluates structural characteristics of graphs in order to select the clusters that best represent the communities in a graph. Since the work of defining what constitutes cluster in a graph is rather difficult, we exploit concepts of graph theory in order to evaluate the cohesiveness and separation of nodes. More specifically, we use the concept of *degeneracy*, and *graph density* to evaluate the connectivity of nodes *in* and *between* clusters. The effectiveness of our approach is experimentally evaluated using real-world data collections.

**Keywords:** Cluster validity; Graph clustering; Data analysis

## 1 Introduction

In recent years, there are many application domains (web applications, biomedicine, social networks) where the available data are represented as graphs and thus the requirement for graph data analysis techniques is stronger than ever. *Graph clustering* is one of the main tasks in graph data analysis and has attracted the interest of data mining research community. A graph clustering can be defined as a set of subgraphs, further referred to as graph clusters or communities, characterized by dense connections between vertices in clusters and low density between vertices of different clusters. The last few decades, a number of methods for graph clustering (community detection) have been proposed [3], [4].

Clustering algorithms with different cost functions give different results, and there is no single optimal choice of the algorithm and the cost function for all available data sets. Even the same clustering algorithm under different assumptions and input parameter values could result in different partitionings of a data set. Then a challenging issue is how to evaluate the quality of different clustering results and select the best possible clustering for a data set. This is the well known *cluster validity problem*.

The problem of *cluster validity* has been widely studied and there is a number of indices for evaluating clustering results [6], [8]. They measure the compactness and separability of clusters using variance or density analysis methods. The majority of cluster validity indices are applied to Euclidean space while there are only few works on graph data.

Since the need for new data analysis techniques that deal with the graph structure increases, the requirement of evaluating the quality of analysis results also arizes. Figure

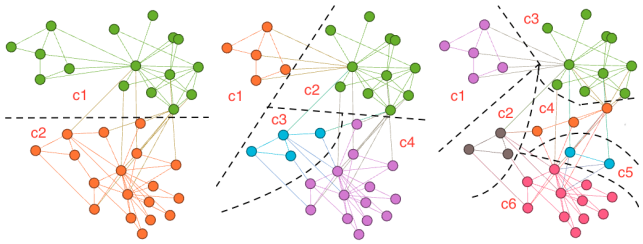



**Fig. 1.** Zachary karate club data set: Partitioning into 2, 4 and 6 clusters.

1 shows the partitioning of graph under different input parameters using a modularity-based clustering algorithm [1]. In most of the cases we are not able to have a visualization of our data and thus it is difficult to identify which is the partitioning that best fits them. Moreover the characteristics and properties of graphs are different from other data types, such as numerical, categorical data, and thus new metrics have to been studied in order to evaluate the structure of graph clusters.

In this work we aim to study the characteristics of graph clusters and develop a new cluster validity approach for evaluating their quality. We exploit concepts of graph theory such as *graph degeneracy* and *density* to evaluate both locally and globally the connectivity in and between graph clusters.

## 2 Related work

Clustering algorithms extract clusters from data, which are not known a priori, and thus the final partition of a data set requires some sort of evaluation in most applications [5], [6]. The evaluation of clustering results is well known in the research community and a number of research works have been made especially in the area of machine learning, pattern recognition [8]. Since the application of graphs is high, the interest of researchers to develop graph clustering algorithms increases. Fortunato [3] provides an extended overview on community detection methods in graphs while he also discusses issues regarding the significance of clustering and how methods should be tested and compared against each other. However, there is little work on cluster validity approaches for graph data.

Boutin et al. [2], present an overview of validity indices for graph clustering while they also propose some normalized version of the available indices. There are indices that extend widely used cluster validity indices, such as David Bouldin, Dunn's index, to deal with graph structures. Also there are indices that use number of links and vertices in a graph to evaluate the connectivity within and between graph clusters. A metric that uses the concepts of cohesion and separation to assess the quality of clusters is the Silhouette index. Its definition is based on the distance between vertices within clusters and between clusters. One of the limitation of this index is the calculation cost. Also it presents a tendency of giving better scores for clusterings with many singletons.

Another well-known metric is the conductance of a cut [12]. It compares the number of edges cut (i.e. between clusters) and the number of edges in either of the two

clusters induced by the cut. Also the coverage [13] of clustering is a metric that used for evaluating clustering results. It is defined as the fraction of intra-cluster edges with respect of the edges of the whole graph. In [14], another cluster validity index, called *performance* metric, is presented. It counts the number of edges withing clusters along with the edges that do not exist between the cluster vertices and the other vertices in the graph.

## 3 Problem statement

We assume a graph $G = (V,E)$, where $V$ is the set of nodes (vertices) and $E$ is the set of edges. Let $SC = \{C_1, \ldots, C_N\}$ be a set of different partitionings (clusterings) of $G$. The number of clusters (also known as communities) could be different in each partitioning $C_i$.

We desire to define an index that assigns to each $C_i \in SC$ a value. This value should be indicative of the quality of clustering $C_i$, i.e. it shows how well $C_i$ captures the structure of clusters in the graph $G$. The definition of such an index should be compatible with the main idea of clustering that is the extraction of compact and well separated clusters. Among the available partitionings in $SC$, we expect that $C_i$ that best fits the clusters in graph $G$, would correspond to the optimal value of the validity index.

In summary, given an index $Q$, a graph $G$ and a set of its partitionings $SC$, we aim to find the partitioning $C_i \in SC$ such that

$$max/min_{C_i \in SC} Q(C_i)$$

The selection of *max* or *min* depends on the index definition.

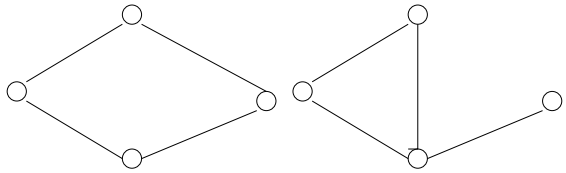
## 4 Definition of a cluster validity index for graphs

We consider an undirected graph $G = (V,E)$ comprising a set $V$ of vertices together with a set $E$ of edges and let $SC = \{C_1, \ldots, C_N\}$ be the set of different partitionings (clusterings) of $G$.

We denote the degree of a vertex $v \in V$ as $d_G(v) = |\{u|(u,v) \in E\}|$. The $k$-core of a graph is the maximal subgraph of $G$, $G' = (V', E')$ where $\forall v \in V'$, $d_{G'}(v) \geq k$. The core number of a vertex $v$, $core(v)$, is the order of the highest-order core that $v$ belongs to. A vertex has core number $k$ if $v$ belongs to the $k$-core but not to the $(k+1)$-core.

The *degeneracy* of a graph $G$, denoted by $deg(G)$, is the largest value $k$ such that it has a $k$-core. A $k-$degenerate graph is an undirected graph in which every sub-graph has a vertex of degree at most $k$. Then the degeneracy of a graph can be defined as the maximum core number of vertices in $V$: $deg(G) = k_{max}\text{-}core = max_{v \in V} core(v)$. The $k_{max}$-core is also called *degeneracy*-core.

**Evaluating the compactness of clusters**. The degeneracy of a graph $G$ has been extensively used for evaluating and detecting strongly cohesive communities in real-word graphs [7]. It indicates the existence of sub-graphs in $G$ where each vertex has at least $deg(G)$ neighbors. Then the *degeneracy* can be considered as measure of graph's sparsity.



(a) G1 Graph (b) G2 graph

**Fig. 2. Degeneracy vs Degeneracy coverage**: The graphs G1 and G2 have the same degeneracy ($deg(G1) = deg(G2) = 2$) while their degeneracy coverage is different $deg\_coverage(G1) = 1$, $deg\_coverage(G2) = \frac{3}{4}$

.

A question that arises at this point is what percentage of graph vertices participate to the *degeneracy*-core (i.e. they have degree at least $deg(G)$ ). In Fig 2 the graphs $G1$, and $G2$ have the same degeneracy but the *degeneracy*-core of each graph covers different part of the graph. In $G1$ all the vertices are part of the *degeneracy*-core while in $G2$ three of the four vertices participate in *degeneracy*-core. Thus in order to evaluate the connectivity of the clusters (communities) in a graph we have to take into account both the degeneracy of the graph cluster and the coverage of its *degeneracy*-core.

We denote the *degeneracy-core* of a graph $G$ as $dG = (dV, dE)$, $dV \subset V$ and $dE \subset E$. We define the coverage of *degeneracy*-core as: $deg\_coverage(G) = \frac{|dV|}{|V|}$

We evaluate the *linkage of a graph cluster* $c_i$ based on the concepts of *degeneracy* and the *coverage of its degeneracy-core*. More specifically, we define the $intra\_linkage$ of a graph cluster as: $intra\_linkage(c_i) = deg(c_i) \cdot deg\_coverage(c_i)$

Considering a clustering of $G$ into $m$ clusters $C = \{c_1, \ldots, c_m\}$, the linkage within $C$ is defined as average $intra\_linkage$ of all clusters in $C$:

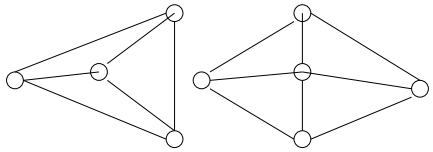
$$intraLink(C) = \frac{1}{m} \sum_{c_i \in C} intra\_linkage(c_i)$$

Another metric that we use to evaluate the cohesion of a graph is the *density* defined as the percentage of the expected edges that exists in the graph. Then the *density of a graph cluster* $c_i = (V_i, E_i)$ is given by: $dens(c_i) = \frac{2 \cdot |E_i|}{|V_i|(|V_i|-1)}$

There are cases that graphs have similar $intra\_linkage$ but the density of graphs is different. Figure 3 shows two graphs $G3$, $G4$ that have the same $intra\_linkage$ but the $G3$ is densest than $G4$. Moreover Figure 2 shows the existence of graphs with the same number of nodes and edges (i.e. same density) but their $intra\_linkage$ is different ($intra\_linkage(G1) = 1$, $intra\_linkage(G2) = \frac{3}{2}$). Then both *density* and *intra-linkage* should be taken into account in order to evaluate the quality of a graph cluster.

The *density of a cluster*, $dens(c_i)$, measures the connectivity among all the vertices in the cluster while $intra\_linkage$ concentrates its evaluation at the densest parts of a cluster.

**Evaluating the separation of clusters.** Assume a pair of graph clusters $c_i$ and $c_j$ with vertices $VC_i$, $VC_j$, respectively. We denote $G(c_i, c_j) = (VC_i \cup VC_j, E(c_j, c_j))$ the graph



(a) G3 Graph (b) G4 graph

**Fig. 3. Intra-Linkage vs Density**: The graphs G3 and G4 have the same intra-linkage($inta\_linkage(G3) = intra\_linkage(G4) = 3$) while their *intra_density* is different $intra\_dens(G3) = 1$, $intra\_dens(G4) = \frac{4}{5}$.

that contains the vertices of the clusters $c_i$ and $c_j$ and the set of edges $E(c_j, c_j)$ which connect the vertices in $VC_i$ with the vertices in $VC_j$.

The $inter\_linkage$ of a pair of clusters $(c_i, c_j)$ measures the linkage of two clusters. It is defined based on the degeneracy of the graph $G(c_i, c_j)$ and the coverage of respective *degeneracy-core*. That is:

$$inter\_linkage(c_i, c_j) = deg(G(c_i, c_j)) * deg\_coverage(G(c_i, c_j))$$

Then the *inter-linkage* of a partitioning of $G$ into $m$ clusters is given by:

$$interLink(C) = \frac{2}{m(m-1)} \sum_i \sum_{j,j<i} inter\_linkage(G(c_i, c_j))$$

**Defining the validity index**. A partitioning that best fits the graph structure of $G$ is expected to contain graph clusters with high *intra_linkage* and low *inter_linkage*. Then a high difference between *intra-linkage* and *inter-linkage* of a clustering $C$ for graph $G$ indicates that $C$ is a good partitioning of $G$.

The *inter-linkage* of clusters is mainly focused on the densest parts between clusters. The density of the graph defined by two graph clusters is also used to evaluate the separation of clusters.

The *inter-density* of a pair of clusters $c_i, c_j$ is defined as the percentage of expected edges across the clusters that exist in the graph: $interDens(c_i, c_j) = \frac{|E(c_i, c_j)|}{|V_i||V_j|}$

The connectivity of two clusters should be evaluated in comparison with the connectivity within the clusters. Given a pair of clusters $(c_i, c_j)$ we define the connectivity of these clusters as follows: $InterCon(c_i, c_j) = \frac{InterDens(c_i, c_j)}{min\{dens(c_i), dens(c_j)\}}$

A partitioning $C$ whose clusters are well separated are expected to have low inter-connectivity, that is $Separation(C) = \frac{2}{m \cdot (m-1)} \sum_i \sum_j \frac{1}{InterCon(c_i, c_j)}$

Then we define the following index as indicator of the quality of graph clustering.

$$QGraph(C) = (intraLink(C) - interLink(C)) + Separation(C)$$

The focus of the first part of the proposed index is on the densest areas between and within clusters while the second part refers to a more global evaluation of clusters connectivity. Based on the above definition we could infer that the higher the value of $QGraph$, the better the quality of clustering.

| (A) Zachary karate club | | | (B) Euro-core email | | |
|---|---|---|---|---|---|
| **number of clusters** | **Modularity** | **QGraph** | **number of clusters** | **Modularity** | **QGraph** |
| **2** | 7.23 | **8.43** | 25 | **72.48** | 2.82 |
| 4 | **9.13** | 5.38 | 27 | 43.0043 | 2.85 |
| 6 | 9.03 | 4.85 | 29 | 30.596 | 3.36 |
| | | | 42 | 11.569 | **20.83** |

**Table 1.** The values of cluster validity indices for the data sets: A) Zachary Karate club, B) Euro-core emails.

## 5 Experimental study

**Data sets.** We have experimented with the real world dataset *EU-core network* [10] for which there is available the "ground-truth" community memberships of nodes [3]. The network was generated using email data from a large European research institution. It contains 1005 nodes and 25571 edges. The netowrk is organized into 42 communities. The average clustering coefficient of the network is 0.3994. Moreover we used the *Zachary karate club* network [11]. This is a social network of friendships between 34 members of a karate club at a US university. In this data set we can find two groups of people into which the karate club was split after an argument between two teachers.

**Discussion on experimental results.** Each dataset is partitioned in different number of clusters using the clustering algorithm presented in [1]. We evaluate the clustering results using the *modularity measure* presented in [2] and the proposed *QGraph* index. Table 1 presents in a comparative fashion the values of the validity indices with respect to the number of clusters. The highest value of the indices indicates the partitioning that the index selects as the best one for the considered dataset.

In case of the EU core (see Table 1 (B)), the *QGraph* index takes its highest value for the partitioning of 42 communities (corresponding to ground truth) while the *modularity index* select 25 clusters as the best partitioning.

Figure 1 depicts different partitionings of the *Zachary karate club* dataset. The results of cluster validity indices for each of the defined partitionings are presented in Table 1(A). We can observe that *QGraph* takes its highest value for the partitioning of two clusters while the *modularity index* selects 4 clusters as the best partitioning.

The above experimental study shows that *QGraph* achieves in all cases to select the partitioning that best fits the underlying graph data.

## 6 Conclusion

In this paper we proposed a new validity index *QGraph* for evaluating graph clustering results. The concepts of graph degeneracy and graph density are properly combined to assess the compactness and separation of extracted clusters. As further work, we plan to evaluate the scalability of the approach and its performance using data sets with various structures and sizes.

[3] http://snap.stanford.edu/data/

**Acknowledgment**
This work has been partly supported by the University of Piraeus Research Center. I. Koutsopoulos acknowledges the support from the AUEB internal project "Original scientific publications".

## References

1. Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, Etienne Lefebvre: "Fast unfolding of communities in large networks". Journal of Statistical Mechanics: Theory and Experiment (2008).
2. F. Boutin, M. Hascoet: "Cluster validity indices for graph partitioning". In: Proceedings of the International Conference of Information Visualisation (2004).
3. S. Fortunato: newblock "Community detection in graphs". Physics Reports (2010).
4. Satu Elisa Schaeffer: "Graph clustering" Computer Science Review (2007).
5. M. Halkidi, Y. Batistakis, M. Vazirgiannis: "On Clustering Validation Techniques" Intelligent Information systems (2001).
6. M. Halkidi, M. Vazirgiannis: "Quality Assessment Approaches In Data Mining". The Data Mining and Knowledge Discovery Handbook: A Complete Guide for Practitioners and Researchers, by Kluwer Academic Publishers (2005).
7. C. Giatsidis: "Graph Mining and Community Evaluation with Degeneracy". Phd Thesis, (2013).
8. S. Theodoridis, K. Koutroubas: Pattern recognition. Academic Press (1999).
9. Andrea Lancichinetti, Santo Fortunato, Filippo Radicchi: "Benchmark graphs for testing community detection algorithms". Physical Review (2008).
10. Hao Yin, Austin R. Benson, Jure Leskovec, and David F. Gleich: "Local Higher-order Graph Clustering." In: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (2017).
11. Ravi Kannan, Santosh Vempala, and Adrian Vetta: "An information flow model for conflict and fission in small groups", In:Journal of Anthropological Research, 452-473 (1977).
12. Ravi Kannan, Santosh Vempala, and Adrian Vetta: "On clusterings: Good, bad and spectral". Journal ACM, 51(3):497 515 (2004).
13. Ulrik Brandes, Marco Gaertler, and Dorothea Wagner: "Engineering graph clustering: Models and experimental evaluation". J. Exp. Algorithmics, 12:126 (2008).
14. S. M. van Dongen: "Graph Clustering by Flow Simulation" PhD thesis, University of Utrecht, The Netherlands (2000).