Title: Relative Validity Criteria for Community Mining Algorithms

Name: Reihaneh Rabbany[1], Mansoreh Takaffoli[1], Justin Fagnan[1], Osmar R. Zaïane[1], Ricardo Campello[1,2]
Affil./Addr. 1: Computing Science Department, University of Alberta
Edmonton, Canada T6G-2E5
E-mail: {rabbanyk,takaffol,fagnan,zaiane,rcampell}@ualberta.ca
Affil./Addr. 2: Department of Computer Science, University of São Paulo
São Carlos, SP, Brazil, C.P. 668
E-mail: campello@icmc.usp.br

Relative Validity Criteria for Community Mining Algorithms

Synonyms

Evaluation Approaches, Quality Measures, Clustering Evaluation, Clustering Objective Function, Graph Clustering, Graph Partitioning, Community Mining

Glossary

Social network: a graph of interconnected nodes
Ground-truth: the right answer
A: adjacency matrix
C: clustering
ED: Edge Path
SPD: Shortest Path Distance
ARD: Adjacency Relation Distance
NOD: Neighbour Overlap Distance
PCD: Pearson Correlation Distance
ICD: ICloseness Distance

Definition

Grouping data points is one of the fundamental tasks in data mining, which is commonly known as clustering if data points are described by attributes. When dealing with interrelated data data represented in the form of nodes and their relationships and the connectivity is considered for grouping but not the node attributes, this task is also referred to as community mining. There has been a considerable number of approaches proposed in recent years for mining communities in a given network. However, little work has been done on how to evaluate community mining results. The common practice is to use an agreement measure to compare the mining result against a ground truth, however, the ground truth is not known in most of the real world applications. In this article, we investigate relative clustering quality measures defined for evaluation of clustering data points with attributes and propose proper adaptations to make them applicable in the context of social networks. Not only these relative criteria could be used as metrics for evaluating quality of the groupings but also they could be used as objectives for designing new community mining algorithms.

Introduction

The recent growing trend in the Data Mining field is the analysis of structured/interrelated data, motivated by the natural presence of relationships between data points in a variety of the present-day applications. The structures in these interrelated data are usually represented using net-

works, known as complex networks or information networks; examples are the hyperlink networks of web pages, citation or collaboration networks of scholars, biological networks of genes or proteins, trust and social networks of humans and much more.

All these networks exhibit common statistical properties, such as power law degree distribution, small-world phenomenon, relatively high transitivity, shrinking diameter, and densification power laws [19; 17]. Network clustering, a.k.a. community mining, is one of the principal tasks in the analysis of complex networks. Many community mining algorithms have been proposed in recent years (for a recent survey refer to Fortunato [6]). These algorithms evolved very quickly from simple heuristic approaches to more sophisticated optimization based methods that are explicitly or implicitly trying to maximize the goodness of the discovered communities. The broadly used explicit maximization objective is the modularity, first introduced by Newman and Girvan [21].

Although there have been many methods presented for detecting communities, very little work has been done on how to evaluate the results and validate these methods. The difficulties of evaluation are due to the fact that the interesting communities that have to be discovered are hidden in the structure of the network, thus, the true results are not known for comparison. Furthermore, there are no other means to measure the goodness of the discovered communities in a real network. We also do not have any large enough dataset with known communities, often called ground truth, to use as a benchmark to generally test and validate the algorithms. The common practice is to use synthetic benchmark networks and compare the discovered communities with the built-in ground truth. However, it is shown that the networks generated with the current benchmarks disagree with some of the characteristics of real networks. These facts motivate investigating a proper objective for evaluation of community mining results.

Key Points

Defining an objective function to evaluate community mining is non-trivial. Aside from the subjective nature of the community mining task, there is no formal definition on the term community. Consequently, there is no consensus on how to measure "goodness" of the discovered communities by a mining algorithm. However, the well-studied clustering methods in the Machine Learning field are subject to similar issues and yet there exists an extensive set of validity criteria defined for clustering evaluation, such as Davies-Bouldin index [4], Dunn index [5], and Silhouette [29] (for a recent survey refer to Vendramin et al [30]). In this article, we describe how these criteria could be adapted to the context of community mining in order to compare results of different community mining algorithms. Also, these criteria can be used as alternatives to modularity to design novel community mining algorithms.

In the following, we first briefly introduce well-known community mining algorithms, and common evaluation approaches including available benchmarks. Next, different ways to adapt clustering validity criteria to handle comparison of community mining results is proposed. Then, we extensively compare and discuss the adapted criteria on real and synthetic networks. Finally, we conclude with a brief analysis of these results.

Historical Background

A community is roughly defined as "densely connected" individuals that are "loosely connected" to others outside their group. A large number of community mining algorithms have been developed in the last few years having different interpretations of this definition. Basic heuristic approaches mine communities by assuming that the network of interest divides naturally into some subgroups, determined by the network itself. For instance, the Clique Percolation Method [25] finds groups of nodes that can be reached via chains of k-cliques. The common optimization

approaches mine communities by maximizing the overall "goodness" of the result. The most credible "goodness" objective is known as modularity Q, proposed in [21], which considers the difference between the fraction of edges that are within the communities and the expected such fraction if the edges are randomly distributed. Several community mining algorithms for optimizing the modularity Q have been proposed, such as fast modularity [20]. Although many mining algorithms are based on the concept of modularity, Fortunato and Barthélemy [7] have shown that the modularity cannot accurately evaluate small communities due to its resolution limit. Hence, any algorithm based on modularity is biased against small communities. As an alternative to optimizing modularity Q, we previously proposed TopLeaders community mining approach [27], which implicitly maximizes the overall closeness of followers and leaders, assuming that a community is a set of followers congregating around a potential leader. There are many other alternative methods. One notable family of approaches mines communities by utilizing information theory concepts such as compression (e.g. Infomap [28]), and entropy (e.g. entropy-base [12]). For a survey on different community mining techniques refer to [6].

The standard procedure for evaluating results of a community mining algorithm is the external evaluation of results, particularly when comparing accuracy of different algorithms; which is assessing the agreement between the results and the ground truth that is known for benchmark datasets. These benchmarks are typically small real world datasets or synthetic networks. On the other hand, there is no well-defined criterion for evaluating the resulting communities for networks without any ground truth, which is the case in most of real world applications. The common practice is to validate the results partly by a human expert.However, the community mining problem is NP-complete; the human expert validation is limited and rather based on narrow intuition than on an exhaustive examination of the relations in the given network. Alternatively, modularity Q is sometimes reported to show the quality of discovered communities. In this article, we investigate other potential measures for comparing different (non-overlapping) community mining results and examine the performance of these measures parallel to the modularity Q. All these new measures are adapted from well-grounded traditional clustering criteria for evaluating data points with attributes. Recently, Vendramin et al. comprehensively compared their performances in [30], based on the idea that the better a criterion the more correlated is its ranking of different partitions to the ranking of an external index.

The external evaluation requires knowing the true communities. For this purpose, several generators have been proposed for synthesizing networks with built-in ground truth. GN benchmark [8] is the first synthetic network generator. This benchmark is a graph with 128 nodes, with expected degree of 16, and is divided into four groups of equal sizes; where the probabilities of the existence of a link between a pair of nodes of the same group and of different groups are z_{in} and $1 - z_{in}$, respectively. However, the same expected degree for all the nodes, and equal-size communities are not accordant to real social network properties. LFR benchmark [16] amends GN benchmark by considering power law distributions for degrees and community sizes. Similar to GN benchmark, each node shares a fraction $1 - \mu$ of its links with the other nodes of its community and a fraction μ with the other nodes of the network. In this article, we generate our synthetic networks using LFR benchmark, due to its more realistic structure.

There are recent studies on the comparison of different community mining algorithms in terms of evaluating their performance on synthetic and real networks. For example, refer to studies by Danon et al. [3]and Lancichinetti and Fortunato [15]. All these studies are based on the agreements of the generated communities with the true one in the ground truth and are using GN and/or LFR benchmarks. Orman et al. [23] further performed a qualitative analysis of the identified communi-

ties by comparing the distribution of resulting communities with the community size distribution of the ground truth. None of these studies, however, considers any different validity criteria other than modularity to evaluate the goodness of the detected communities. In this article, we plan to examine potential validity criteria specifically defined for evaluation of community mining results. In the future, these criteria not only can be used as a means to measure the goodness of discovered communities, but also as an objective function to detect communities.

Community Quality Criteria

In this section, we overview several validity criteria that could be used as relative indexes for comparing and evaluating different partitionings of a given network. All of these criteria are generalized from well-known clustering criteria. The clustering quality criteria are defined with the implicit assumption that data points consist of vectors of attributes. Consequently their definition is mostly integrated or mixed with the definition of the distance measure between data points. The commonly used distance measure is the Euclidean distance, which cannot be defined for graphs. Therefore, we first review different possible distance measures that could be used in graphs. Then, we present generalizations of criteria that could use any notion of distance.

Distance Between Nodes

Let A denote the adjacency matrix of the graph, and let A_{ij} be the weight of the edge between nodes n_i and n_j. The distance $d(i,j)$ denotes the dissimilarity between n_i and n_j, which can be computed by one of the following measures.

Edge Path (ED)

The distance between two nodes is the inverse of their incident edge weight:

$$d_{ED}(i,j) = \frac{1}{A_{ij}}$$

For avoiding division by zero, when A_{ij} is zero, $1/\epsilon$ is returned where ϵ is a very small number; the same is true for all other formula whenever a division by zero may occur.

Shortest Path Distance (SPD)

The distance between two nodes is the length of the shortest path between them, which could be computed using the well-known Dijkstra's Shortest Path algorithm.

Adjacency Relation Distance (ARD)

The distance between two nodes is the structural dissimilarity between them, that is computed by the difference between their immediate neighbourhood.

$$d_{ARD}(i,j) = \sqrt{\sum_{k\neq j,i} (A_{ik} - A_{jk})^2}$$

Neighbour Overlap Distance (NOD)

The distance between two nodes is the ratio of the unshared neighbours between them.

$$d_{NOD}(i,j) = 1 - \frac{|\aleph_i \cap \aleph_j|}{|\aleph_i \cup \aleph_j|}$$

where $\aleph_i$ is the set of nodes directly connected to n_i . Note that there is a close relation between this measure and the previous one, since similarly d_{NOD} could be re-written as:

$$d_{NOD}(i,j) = 1 - \frac{\sum_{k\neq j,i} |A_{ik}+A_{jk}| - \sum_{k\neq j,i} |A_{ik}-A_{jk}|}{\sum_{k\neq j,i} |A_{ik}+A_{jk}| + \sum_{k\neq j,i} |A_{ik}-A_{jk}|}$$

The latter formulation of d_{NOD} in terms of the adjacency matrix can be straightforwardly generalized for weighted graphs.

Pearson Correlation Distance (PCD)

The Pearson correlation coefficient between two nodes is the correlation between their corresponding rows of the adjacency matrix:

$$C(i,j) = \frac{\sum_k (A_{ik} - \mu_i)(A_{jk} - \mu_j)}{N\sigma_i\sigma_j}$$

where N is the number of nodes, the average $\mu_i = (\sum_k A_{ik})/N$ and the variance

$\sigma_i = \sqrt{\sum_k (A_{ik} - \mu_i)^2/N}$. Then, the distance between two nodes is computed as $d_{PCD}(i,j) = 1 - C(i,j)$, which lies between 0 (when the two nodes are most similar) and 2 (when the two nodes are most dissimilar).

ICloseness Distance (ICD)

The distance between two nodes is computed as the inverse of the connectivity between their common neighbourhood:

$$d_{ICD}(i,j) = \frac{1}{\sum\limits_{k \in \aleph_i \cap \aleph_j} ns(k,i) ns(k,j)}$$

where $ns(k,i)$ denotes the neighbouring score between nodes k and i that is computed iteratively (for complete formulation refer to [26]).

Community Centroid

In addition to the notion of distance measure, most of the cluster validity criteria use averaging between the numerical data points to determine the centroid of a cluster. The averaging is not defined for nodes in a graph, therefore we modify the criteria definitions to use a generalized centroid notion, in a way that, if the centroid is set as averaging, we would obtain the original criteria definitions, but we could also use other alternative notions for centroid of a group of data points.

Averaging data points results in a point with the least average distance to the other points. When averaging is not possible, using medoid is the natural option, which is perfectly compatible with graphs. More formally, the centroid of a community can be obtained as:

$$\overline{C} = \arg\min_{m \in C} \sum_{i \in C} d(i,m)$$

Relative Validity Criteria

Here we present our generalizations of well-known clustering validity criteria defined as quality measures for internal evaluation of clustering results. All these criteria are originally defined based on distances between data points, which is in all cases the Euclidean or other inner product norms of difference between their vectors of attributes; refer to [30] for comparative analysis of these criteria in the clustering context. We alter the formulae to use a generalized distance, so that we can plug in our graph distance measures. The other alteration is generalizing the mean over data points to a general centroid notion, which can be set as averaging in the presence of attributes and the *medoid* in our case of dealing with graphs and in the absence of attributes.

In a nutshell, in every criterion, the average of points in a cluster is replaced with a generalized notion of centroid, and distances between data points are generalized from Euclidean/norm to a generic distance. Consider a clustering $C = \{C_1 \cup C_2 \cup ... \cup C_k\}$ of N data points, where $\overline{C}$ denotes the centroid of data points belonging to C. The quality of C can be measured using one of the following criteria.

Variance Ratio Criterion (VRC)

This criterion measures the ratio of the between-cluster/community distances to within-cluster/community distances which could be generalized as follows:

$$VRC = \frac{\sum\limits_{l=1}^{k} |C_l| d(\overline{C}_l, \overline{C})}{\sum\limits_{l=1}^{k} \sum_{i \in C_l} d(i, \overline{C}_l)} \times \frac{N-k}{k-1}$$

where $\overline{C}_l$ is the centroid of the cluster/community C_l, and $\overline{C}$ is the centroid of the entire data/network. The original clustering formula proposed in [1] for attributes vectors is obtained if the centroid is fixed to averaging of vectors of attributes and distance to (square of) Euclidean distance.

Davies-Bouldin index (DB)

This minimization criterion calculates the worst-case within-cluster/community to between-cluster/community distances ratio averaged over all clusters/communities [4]:

$$DB = \frac{1}{k} \sum_{l=1}^{k} \max_{m \neq l} ((\overline{d}_l + \overline{d}_m)/d(\overline{C}_l, \overline{C}_m))$$

$$\overline{d}_l = \frac{1}{|C_l|} \sum_{i \in C_l} d(i, \overline{C}_l)$$

Dunn index

This criterion considers both the minimum distance between any two clusters/communities and the length of the largest cluster/community diameter (i.e. the maximum or the average distance between all the pairs in the cluster/community) [5]:

$$Dunn = \min_{l \neq m} \{ \frac{\delta(C_l, C_m)}{max_p \Delta(C_p)} \}$$

where δ denotes distance between two communities and Δ is the diameter of a community. Different variations of calculating δ and Δ are available; δ could be single, complete or average linkage, or only the difference between the two centroids. Moreover, Δ could be maximum or average distance between all pairs of nodes, or the average distance of all nodes to the centroid. For example, the single linkage for δ and maximum distance for Δ are $\delta(C_l, C_m) = \min_{i \in C_l, j \in C_m} d(i,j)$ and $\Delta(C_p) = \max_{i,j \in C_p} d(i,j)$. Therefore, we have different variations of Dunn index in our experiments, each indicated by two indexes for different methods to calculate δ (i.e. single(0), complete(1), average(2), and centroid(3)) and different methods to calculate Δ (i.e. maximum(0), average(1), average to centroid(3)).

Silhouette Width Criterion (SWC)

This criterion measures the average of silhouette score for each data point. The silhouette score of a point shows the goodness of the community it belongs to by calculating the normalized difference between the distance to its nearest neighbouring community and its own community [29]. Taking the average one has:

$$SWC = \frac{1}{N} \sum_{l=1}^{k} \sum_{i \in C_l} \frac{\min_{m \neq l} d(i, C_m) - d(i, C_l)}{\max \{ \min_{m \neq l} d(i, C_m), d(i, C_l) \}}$$

where $d(i, C_l)$ is the distance of point i to community C_l, which is originally set to be the average distance (called SWC2) (i.e. $1/|C_l| \sum_{j \in C_l} d(i,j)$) or could be the distance to its centroid (called SWC4) (i.e. $d(i, \overline{C_l})$). An alternative formula for Silhouette is proposed in [30] :

$$ASWC = \frac{1}{N} \sum_{l=1}^{k} \sum_{i \in C_l} \frac{\min_{m \neq l} d(i, C_m)}{d(i, C_l)}$$

PBM

This criterion is based on the within-community distances and the maximum distance between centroids of communities[24]:

$$PBM = \frac{1}{k} \times \frac{\max_{l,m} d(\overline{C}_l, \overline{C}_m)}{\sum_{l=1}^{k} \sum_{i \in C_l} d(i, \overline{C}_l)}$$

C-Index

This criterion compares the sum of the within-community distances to the worst and best case scenarios [2]. The best case scenario is where the within-community distances are the shortest distances in the graph, and the worst case scenario is where the within-community distances are the longest distances in the graph.

$$\theta = \frac{1}{2} \sum_{l=1}^{k} \sum_{i,j \in C_l} d(i,j)$$

$$CIndex = \frac{\theta - \min \theta}{\max \theta - \min \theta}$$

The $\min \theta / \max \theta$ is computed by summing the m_1 smallest/largest distances between every two points, where $m_1 = \sum_{l=1}^{k} \frac{|C_l|(|C_l|-1)}{2}$.

Z-Statistics

This criterion is similar to C-Index, however with different formulation [10]:

$$ZIndex = \frac{\theta - E(\theta)}{\sqrt{var(\theta)}}$$

$$E(\theta) = \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{N} d(i,j)$$

$$Var(\theta) = \frac{\left(\sum_{i=1}^{N} \sum_{j=1}^{N} d(i,j) \right)^2 - 2 \sum_{i=1}^{N} \left(\sum_{j=1}^{N} d(i,j) \right)^2}{N(N-1)}$$

$$-\frac{\left(\sum_{i=1}^{N}\sum_{j=1}^{N} d(i,j)\right)^2}{N^2} + \frac{\sum_{i=1}^{N}\sum_{j=1}^{N} d(i,j)^2}{N}$$

Point-Biserial (PB)

This criterion computes the correlation of the distances between nodes and their cluster co-membership which is dichotomous variable [18]. Intuitively, nodes that are in the same community should be separated by shorter distances than those which are not:

$$PB = \frac{M_1 - M_0}{S}\sqrt{\frac{m_1 m_0}{m^2}}$$

where m is the total number of distances i.e. $N(N-1)/2$ and S is the standard deviation of all pairwise distances i.e. $\sqrt{\frac{1}{m}\sum_{i,j}(d(i,j) - \frac{1}{m}\sum_{i,j} d(i,j))^2}$, while M_1, M_0 are respectively the average of within and between-community distances, and m_1 and m_0 represent the number of within and between community distances. More formally:

$$m_1 = \sum_{l=1}^{k} \frac{N_l(N_l-1)}{2} \quad m_0 = \sum_{l=1}^{k} \frac{N_l(N-N_l)}{2}$$

$$M_1 = 1/2\sum_{l=1}^{k}\sum_{i,j\in C_l} d(i,j) \quad M_0 = 1/2\sum_{l=1}^{k}\sum_{\substack{i\in C_l \\ j\notin C_l}} d(i,j)$$

Modularity

is the well-known criterion proposed by Newman et al. [21] specifically for the context of community mining. This criterion considers the difference between the fraction of edges that are within the community and the expected such fraction if the edges were randomly distributed. Let E denote the number of edges in the network i.e. $E = \frac{1}{2}\sum_{ij} A_{ij}$, then Q-modularity is defined as:

$$Q = \frac{1}{2E}\sum_{l=1}^{k}\sum_{i,j\in C_l}[A_{ij} - \frac{\sum_k A_{ik}\sum_k A_{kj}}{2E}]$$

The computational complexity of different validity criteria is provided in the previous work by Vendramin et al. [30].

Comparison Methodology and Results

In this section, we compare the proposed relative community criteria. First, we describe the approach we have used for the comparison. Then, we report the criteria performances in different settings. The following procedure summarizes our comparison approach.

$D \leftarrow \{d_1, d_2 \dots d_n\}$
for all dataset $d \in D$ **do**
  {generate m possible partitionings}
  $P(d) \leftarrow \{p_{d_1}, p_{d_2} \dots p_{d_m}\}$
  {compute external scores}
  $E(d) \leftarrow \{a(p_{d_1}, p_d^*), a(p_{d_2}, p_d^*) \dots a(p_{d_m}, p_d^*)\}$
  for all $c \in Criteria$ **do**
    {compute internal scores}
    $I_c(d) \leftarrow \{c(p_{d_1}), c(p_{d_2}) \dots c(p_{d_m})\}$
    {compute the correlation}
    $score_c(d) \leftarrow correlation(E, I)$
  end for
end for
{rank criteria based on their average scores}
$score_c \leftarrow \frac{1}{n}\sum_{d=1}^{n} score_c(d)$

The performance of a criterion could be examined by how well it could rank different partitionings of a given dataset. More formally, consider we have a dataset d and a set of m different possible partitionings, i.e. $P(d) = \{p_{d1}, p_{d2}, \dots, p_{dm}\}$. Then, the performance of criterion c on dataset d could be determined by how much its values, $I_c(d) = \{c(p_{d1}), c(p_{d2}), \dots, c(p_{dm})\}$, correlate with the "goodness" of these partitionings. Assuming that the true partitioning (i.e. ground truth) p_d^* is known for dataset d, the "goodness" of partitioning p_{di} could be determined using partitioning agreement measure a, a.k.a. external evaluation. Hence, for dataset d with set of possible partitionings $P(d)$, the external evaluation provides $E(d) = \{a(p_{d1}, p_d^*), a(p_{d2}, p_d^*), \dots, a(p_{dm}, p_d^*)\}$, where (p_{d1}, p_d^*) denotes the "goodness" of partitioning p_{d1} comparing to the ground truth. Then, the performance score of criterion c on dataset d could be examined by the correlation of its

values $I_c(d)$ and the values obtained from the external evaluation $E(d)$ on different possible partitionings. Finally, the criteria are ranked based on their average performance score over a set of datasets.

External evaluation is done with an agreement measure, which computes the agreement between two given partitionings or between a partitioning and the ground truth. There are several choices for the partitioning agreement measure. The commonly used ones are pair counting based, such as Adjusted Rank Index (ARI) [9] and Jaccard Coefficient [11], and the information theoretic-based, such as Normalized Mutual Information (NMI) [14; 3] and the Adjusted Mutual Information (AMI) [31].

There are also different ways to compute the correlation between two vectors. The classic options are Pearson Product Moment coefficient or the Spearman's Rank correlation coefficient. The reported results in our experiments are based on the Spearman's Correlation, since we are interested in the correlation of rankings that a criterion provides for different partitionings and not the actual values of that criterion. However, the reported results mostly agree with the results obtained by using Pearson correlation, which are reported in the supplementary materialsavailable from: `http://cs.ualberta.ca/~rabbanyk/criteriaComparison`.

Sampling the Partitioning Space

In our comparison, we generate different partitionings for each dataset d to sample the space of all possible partitionings. For doing so, given the perfect partitioning, p^*_d, we randomized different versions of p^*_d by randomly merging and splitting communities and swapping nodes between them. The sampling procedure is described in more details in the supplementary materials.

Table 1. Statistics for sample partitionings of each real world dataset. For example, for the Karate Club dataset which has 2 communities in its ground truth, we have generated 60 different partitionings with average 3.57±1.23 clusters ranging from 2 to 6 and the "goodness" of the samples is on average 0.46±0.27 in terms of their AMI agreement.

Results on Real World Datasets

We first compare performance of different criteria on five well-known real-world benchmarks: Karate Club (weighted) by Zachary [32], Sawmill Strike data-set [22], NCAA Football Bowl Subdivision [8], and Politician Books from Amazon [13]. Table 1 shows general statistics about the datasets and their generated samples. We can see that the randomized samples cover the space of partitionings according to their external index range.

Fig. 1. Visualization of correlation between an external agreement measure and some relative quality criteria for Karate dataset. The x axis indicates different random partitionings, and the y axis indicates the value of the index. While, the blue/darker line represents the value of the external index for the given partitioning and the red/lighter line represents the value that the criterion gives for the partitioning. Please note that the value of criteria are not generally normalized and in the same range as the external indexes, in this figure AMI. For the sake of illustration therefore, each criterion's values are scaled to be in the same range as of the external index.

Figure 1 exemplifies how different criteria exhibit different correlations with the external index. It visualizes the correlation between few selected relative indexes and an external index for one of our datasets listed in Table 1.

Similar analysis is done for all 4 datasets × 19 criteria × 7 distances × 4 external indexes, which produced over 2000 such correlations. The top ranked criteria based on their average performance over these datasets are summarized in Table 2. Based on these results, *CIndex* when used with *PCD* distance has a higher correlation with the external index comparing to the modularity Q. And this is true regardless of the choice of AMI as the external index, since it is ranked above Q also by ARI and NMI.

Table 2. Overall ranking of criteria on the real world datasets, based on the average Spearman's correlation of criteria with the AMI external index, AMI_{corr}. Ranking based on correlation with other external indexes is also reported.

The correlation between a criterion and an external index depends on how close the randomized partitionings are from the true partitioning of the ground truth. This can be seen in Figure 1. For example, *Dunn01* (single linkage network diameter and average linkage within community scores) with the *ICD* distance agrees strongly with the external index in samples with higher external index value, i.e. closer to the ground truth, but not on further samples. On the other hand, Q is very well matched for the samples too far or too close to the ground truth, but is not doing as well as others in the middle. With this in mind, we have divided the generated clustering samples into three sets of easy, medium and hard samples and re-ranked the criteria in each of these settings. Since the external index determines how far a sample is from the optimal result, the samples are divided into three equal length intervals according to the range of the external index. Table 3, reports the rankings of the top criteria in each of these three settings. We can see that these average results support our earlier hypothesis, i.e., when considering partitionings medium far from the true partitioning, *CIndex PCD* performs significantly better than modularity Q, while their performances are not very different in the near optimal samples or the samples very far from the ground truth. One may conclude based on this experiment that *CIndex PCD* is a more accurate evaluation criterion comparing to Q, especially when the results might not be very accurate or very poor.

Table 3. Difficulty analysis of the results: considering ranking for partitionings near optimal ground truth, medium far and very far. Reported result are based on AMI and the Spearman's correlation.

Synthetic Benchmarks Datasets

Lastly, we compare the criteria on a larger set of synthetic benchmarks. We have generated our dataset using the LFR benchmarks [16] which are the generators widely in use for community mining evaluation. Similar to the last experiment, Table 5 reports the ranking of the top criteria according to their average performance on synthesized datasets of Table 4. Based on which, modularity Q overall outperforms other criteria especially in ranking poor partitionings; while *CIndex PCD* performs better in ranking finner results.

Table 4. Statistics for sample partitionings of each synthetic dataset. The benchmark generation parameters: 100 nodes with average degree 5 and maximum degree 50, where size of each community is between 5 and 50 and mixing parameter is *0.1*.

Table 5. Overall ranking and difficulty analysis of the synthetic results. Here communities are well-separated with mixing parameter of .1. Similar to the last experiment, reported result are based on AMI and the Spearman's correlation.

The LFR generator can generate networks with different levels of difficulty for the partitioning task, by changing how well separated the communities are in the ground truth. To examine the effect of this difficulty parameter, we have ranked the criteria for different values of this parameter. We observed that modularity Q is the superior criterion for these synthetic benchmarks up to some level of how mixed are the communities, but this changes in more difficult settings. Results for other settings are available in the supplementary materials.

Table 6 reports the overall ranking of the criteria for a difficult set of datasets that have high mixing parameter. We can see that in this setting *PB* index used with *PCD*, *NOD*, *SPD* or *ARD* distances, is significantly more reliable than modularity Q, particularly considering the much higher variance of the latter.

Table 6. Overall ranking of criteria based on AMI & Spearman's Correlation on the synthetic benchmarks with the same parameters as in Table 4 but much higher mixing parameter, .7. We can see that in these settings, PB indexes outperform modularity Q.

In short, the relative performances of different criteria depends on the difficulty of the network itself, as well as how far we are sampling from the ground truth. Altogether, choosing the right criterion for evaluating different community mining

results depends both on the application, i.e., how well-separated communities might be in the given network, and also on the algorithm that produces these results, i.e., how fine the results might be. For example, if the problem is hard and communities are heavily mixed, modularity Q might not distinguish the good and bad partitionings very well. While if we are choosing between fine and well separated clusterings, it indeed is the superior criterion.

Conclusion

In this article, we generalized well-known clustering validity criteria originally used as quantitative measures for evaluating quality of clusters of data points represented by attributes. The first reason of this generalization is to adapt these criteria in the context of community mining of interrelated data. The only commonly used criterion to evaluate the goodness of detected communities in a network is the modularity Q. Providing more validity criteria can help researchers to better evaluate and compare community mining results in different settings. Also, these adapted validity criteria can be further used as objectives to design new community mining algorithms. Our generalized formulation is independent of any particular distance measure unlike most of the original clustering validity criteria that are defined based on the Euclidean distance. The adopted versions therefore could be used as community criteria when plugged in with different graph distances. In our experiments, several of these adopted criteria exhibit high performances on ranking different partitionings of a given dataset, which makes them possible alternatives for the Q modularity. However, a more careful examination is needed as the rankings depends significantly on the experimental settings and the criteria should be chosen based on the application.

Cross-References

1. Community Detection, Current and Future Research Trends (00027, 33/33)
2. Competition Within and Between Communities Within and Across Social Networks(00216, 36/36)
3. Combining Link and Content for community detection (00214, 28/28)
4. Communities Discovery and Analysis in Online and Offline Social Networks (00006, 30/30)
5. Extracting and Inferring Communities via Link Analysis (00218, 83/83)
6. Game Theoretic Framework for Community Detection (00350, 94/94)
7. Inferring Social Ties and Communities in Social Networks (00177, 116/116)

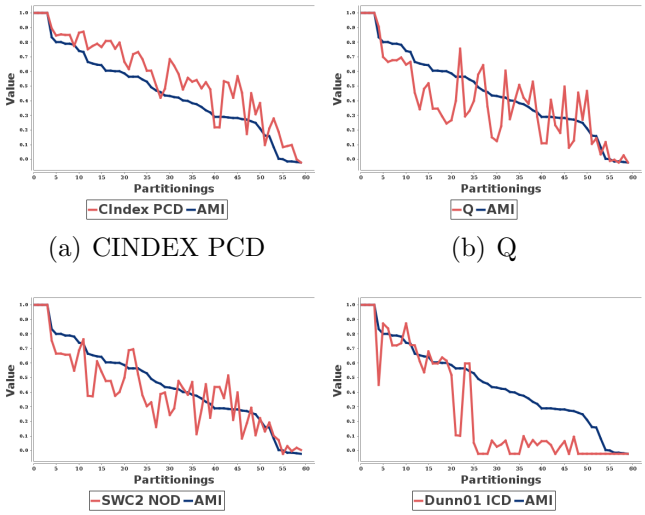
Acknowledgement

The authors are grateful for the support from Alberta Innovates Centre for Machine Learning and NSERC. Ricardo Campello also acknowledges the financial support of Fapesp and CNPq.

References

1. Calinski T, Harabasz J (1974) A dendrite method for cluster analysis. Communications in Statistics-theory and Methods 3:1–27
2. Dalrymple-Alford EC (1970) Measurement of clustering in free recall. Psychological Bulletin 74:32–34
3. Danon L, Guilera AD, Duch J, Arenas A (2005) Comparing community structure identification. Journal of Statistical Mechanics: Theory and Experiment (09):09,008
4. Davies DL, Bouldin DW (1979) A cluster separation measure. Pattern Analysis and Machine Intelligence, IEEE Transactions on PAMI-1(2):224 –227
5. Dunn JC (1974) Well-separated clusters and optimal fuzzy partitions. Journal of Cybernetics 4(1):95–104
6. Fortunato S (2010) Community detection in graphs. Physics Reports 486(35):75–174
7. Fortunato S, Barthélemy M (2007) Resolution limit in community detection. Proceedings of the National Academy of Sciences 104(1):36–41
8. Girvan M, Newman MEJ (2002) Community structure in social and biological networks. Proceedings of the National Academy of Sciences 99(12):7821–7826
9. Hubert L, Arabie P (1985) Comparing partitions. Journal of Classification 2:193–218

10. Hubert LJ, Levin JR (1976) A general statistical framework for assessing categorical clustering in free recall. Psychological Bulletin 83:1072–1080
11. Jaccard P (1901) Étude comparative de la distribution florale dans une portion des alpes et des jura. Bulletin del la Société Vaudoise des Sciences Naturelles 37:547–579
12. Kenley EC, Cho YR (2011) Entropy-based graph clustering: Application to biological and social networks. In: IEEE International Conference on Data Mining
13. Krebs V (2004) Books about us politics, http://www.orgnet.com/
14. Kvalseth TO (1987) Entropy and correlation: Some comments. Systems, Man and Cybernetics, IEEE Transactions on 17(3):517 –519, DOI 10.1109/TSMC.1987.4309069
15. Lancichinetti A, Fortunato S (2009) Community detection algorithms: A comparative analysis. Physical Review E 80(5):056,117
16. Lancichinetti A, Fortunato S, Radicchi F (2008) Benchmark graphs for testing community detection algorithms. Physical Review E 78(4):046,110
17. Leskovec J, Kleinberg J, Faloutsos C (2005) Graphs over time: densification laws, shrinking diameters and possible explanations. In: ACM SIGKDD international conference on Knowledge discovery in data mining, pp 177–187
18. Milligan G, Cooper M (1985) An examination of procedures for determining the number of clusters in a data set. Psychometrika 50(2):159–179
19. Newman M (2010) Networks: An Introduction. OUP Oxford, URL `http://books.google.ca/books?id=q7HVtpYVfC0C`
20. Newman MEJ (2006) Modularity and community structure in networks. Proceedings of the National Academy of Sciences 103(23):8577–8582
21. Newman MEJ, Girvan M (2004) Finding and evaluating community structure in networks. Physical Review E 69(2):026,113
22. Nooy Wd, Mrvar A, Batagelj V (2004) Exploratory Social Network Analysis with Pajek. Cambridge University Press
23. Orman GK, Labatut V, Cherifi H (2011) Qualitative comparison of community detection algorithms. In: International Conference on Digital Information and Communication Technology and Its Applications, vol 167, pp 265–279
24. Pakhira M, Dutta A (2011) Computing approximate value of the pbm index for counting number of clusters using genetic algorithm. In: International Conference on Recent Trends in Information Systems, pp 241 –245
25. Palla G, Derenyi I, Farkas I, Vicsek T (2005) Uncovering the overlapping community structure of complex networks in nature and society. Nature 435(7043):814–818
26. Rabbany R, Zaïane OR (2011) A diffusion of innovation-based closeness measure for network associations. In: IEEE International Conference on Data Mining Workshops, pp 381 –388
27. Rabbany R, Chen J, Zaïane OR (2010) Top leaders community detection approach in information networks. In: SNA-KDD Workshop on Social Network Mining and Analysis
28. Rosvall M, Bergstrom CT (2008) Maps of random walks on complex networks reveal community structure. Proceedings of the National Academy of Sciences 105(4):1118–1123
29. Rousseeuw P (1987) Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics 20(1):53–65
30. Vendramin L, Campello RJGB, Hruschka ER (2010) Relative clustering validity criteria: A comparative overview. Statistical Analysis and Data Mining 3(4):209–235
31. Vinh NX, Epps J, Bailey J (2010) Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. Journal of Machine Learning Research 11:2837–2854
32. Zachary WW (1977) An information flow model for conflict and fission in small groups. Journal of Anthropological Research 33:452–473



(a) CINDEX PCD (b) Q

(c) SWC2 NOD (d) Dunn01 ICD

Fig. 1. Visualization of correlation between an external agreement measure and some relative quality criteria for Karate dataset. The x axis indicates different random partitionings, and the y axis indicates the value of the index. While, the blue/darker line represents the value of the external index for the given partitioning and the red/lighter line represents the value that the criterion gives for the partitioning. Please note that the value of criteria are not generally normalized and in the same range as the external indexes, in this figure AMI. For the sake of illustration therefore, each criterion's values are scaled to be in the same range as of the external index.

Table 1. Statistics for sample partitionings of each real world dataset. For example, for the Karate Club dataset which has 2 communities in its ground truth, we have generated 60 different partitionings with average 3.57±1.23 clusters ranging from 2 to 6 and the "goodness" of the samples is on average 0.46±0.27 in terms of their AMI agreement.

Dataset	K^*	#	$\overline{K}$	$\overline{AMI}$
Strike	3	60	3.17±1∈[2,5]	0.59±0.27∈[-0.04,1]
Polboks	3	60	3.17±1.13∈[2,6]	0.44±0.25∈[0.04,1]
Karate	2	60	3.57±1.23∈[2,6]	0.46±0.27∈[-0.02,1]
Football	11	60	10.17±4.55∈[4,19]	0.68±0.16∈[0.4,1]

Table 2. Overall ranking of criteria on the real world datasets, based on the average Spearman's correlation of criteria with the AMI external index, AMI_{corr}. Ranking based on correlation with other external indexes is also reported.

Rank	Criterion	AMI_{corr}	ARI	Jaccard	NMI
1	CIndex PCD	0.907±0.058	1	1	1
2	SWC2 NOD	0.857±0.031	4	4	2
3	Q	0.85±0.083	2	2	3
4	CIndex ARD	0.826±0.162	6	15	5
5	CIndex SPD	0.811±0.126	3	10	4
6	ASWC2 NOD	0.809±0.043	5	11	6
7	CIndex NOD	0.794±0.096	12	3	9
8	SWC2 PCD	0.789±0.103	7	7	8
9	SWC4 NOD	0.778±0.075	9	5	7
10	ASWC2 PCD	0.772±0.088	10	9	10
11	SWC2 SPD	0.751±0.121	8	6	11
12	Dunn01 ICD	0.742±0.111	18	24	12
13	ASWC2 SPD	0.733±0.116	11	8	13
14	Dunn00 PCD	0.721±0.1	21	30	14
15	DB ICD	0.712±0.063	24	22	16
16	Dunn00 ICD	0.707±0.133	28	28	15
17	Dunn03 ICD	0.703±0.055	25	23	17
18	SWC4 PCD	0.7±0.072	14	12	21

Table 3. Difficulty analysis of the results: considering ranking for partitionings near optimal ground truth, medium far and very far. Reported result are based on AMI and the Spearman's correlation.

Near Optimal Samples					
Rank	Criterion	AMI_{corr}	ARI	Jaccard	NMI
1	Q	0.736±0.266	5	5	2
2	CIndex PCD	0.72±0.326	1	1	3
3	SWC2 SPD	0.718±0.389	3	3	4
4	CIndex SPD	0.716±0.14	4	4	1
5	SWC2 ICD	0.713±0.396	2	2	5
6	ASWC2 ICD	0.687±0.334	11	10	7
Medium Far Samples					
Rank	Criterion	AMI_{corr}	ARI	Jaccard	NMI
1	CIndex PCD	0.608±0.202	8	18	1
2	CIndex NOD	0.58±0.053	39	13	2
3	CIndex ARD	0.513±0.313	26	62	5
4	Dunn01 ICD	0.457±0.173	58	83	8
5	SWC2 NOD	0.447±0.19	5	9	3
6	ASWC2 PCD	0.446±0.191	7	3	9
7	SWC2 PCD	0.446±0.19	6	2	10
8	Dunn03 ICD	0.439±0.109	43	37	11
9	Dunn31 SPD	0.437±0.177	56	47	15
10	Dunn01 SPD	0.434±0.205	29	67	7
11	Q	0.409±0.353	4	7	16
12	DB ICD	0.405±0.072	40	38	18
Far Far Samples					
Rank	Criterion	AMI_{corr}	ARI	Jaccard	NMI
1	SWC2 NOD	0.634±0.217	3	13	1
2	ASWC2 NOD	0.583±0.191	5	21	2
3	Q	0.498±0.179	4	38	5
4	CIndex PCD	0.493±0.282	2	4	13
5	CIndex SPD	0.437±0.291	1	11	4
6	SWC3 NOD	0.436±0.344	8	2	25

Table 4. Statistics for sample partitionings of each synthetic dataset. The benchmark generation parameters: 100 nodes with average degree 5 and maximum degree 50, where size of each community is between 5 and 50 and mixing parameter is *0.1*.

Dataset	K^*	#	$\overline{K}$	$\overline{AMI}$
network1	4	60	3.4±1.17∈[2,6]	0.46±0.23∈[0,1]
network2	3	60	3.1±1.27∈[2,7]	0.49±0.22∈[0.13,1]
network3	2	60	3.3±1.13∈[2,6]	0.47±0.23∈[0.11,1]
network4	7	60	5.17±2.49∈[2,12]	0.57±0.2∈[0.18,1]
network5	2	60	3.5±1.36∈[2,8]	0.44±0.22∈[0.11,1]
network6	5	60	5.8±2.55∈[2,12]	0.68±0.2∈[0.27,1]
network7	4	60	5.2±2.65∈[2,12]	0.47±0.19∈[0.13,1]
network8	5	60	5.37±2.04∈[2,10]	0.67±0.21∈[0.32,1]
network9	5	60	5.5±2.05∈[2,10]	0.69±0.19∈[0.37,1]
network10	6	60	5.33±2.51∈[2,11]	0.63±0.19∈[0.24,1]

Table 5. Overall ranking and difficulty analysis of the synthetic results. Here communities are well-separated with mixing parameter of .1. Similar to the last experiment, reported result are based on AMI and the Spearman's correlation.

Overall Results					
Rank	Criterion	AMI_{corr}	ARI	Jaccard	NMI
1	Q	0.894±0.018	1	2	1
2	ASWC2 NOD	0.854±0.056	3	4	2
3	SWC2 NOD	0.854±0.051	4	3	3
4	CIndex PCD	0.826±0.07	2	1	4
5	CIndex SPD	0.746±0.137	8	24	5
6	SWC2 PCD	0.743±0.047	5	5	6
7	ASWC2 PCD	0.739±0.048	6	6	7
8	Dunn00 PCD	0.707±0.11	11	26	8
9	SWC4 NOD	0.699±0.131	7	7	9
10	SWC4 ARD	0.689±0.124	9	8	10
11	ASWC2 ARD	0.683±0.108	15	21	11
12	ASWC2 ED	0.665±0.139	10	11	12
13	SWC2 SPD	0.657±0.124	14	16	13
14	ASWC2 SPD	0.651±0.196	16	17	15
15	Dunn03 NOD	0.645±0.156	23	33	14
Near Optimal Results					
Rank	Criterion	AMI_{corr}	ARI	Jaccard	NMI
1	CIndex PCD	0.729±0.17	1	1	1
2	Q	0.722±0.111	6	5	5
3	SWC2 SPD	0.717±0.185	18	18	2
4	SWC4 NOD	0.709±0.201	5	6	4
5	SWC2 ICD	0.704±0.216	15	15	3
6	SWC4 ARD	0.674±0.183	7	7	6
7	ASWC2 NOD	0.66±0.261	20	19	7
8	SWC2 NOD	0.649±0.264	14	14	9
Medium Far Results					
Rank	Criterion	AMI_{corr}	ARI	Jaccard	NMI
1	SWC2 NOD	0.455±0.191	5	11	3
2	CIndex PCD	0.453±0.245	1	2	5
3	Q	0.45±0.236	2	9	2
4	ASWC2 NOD	0.435±0.187	4	14	1
5	Dunn00 ARD	0.386±0.243	119	111	7
6	Dunn00 PCD	0.38±0.195	58	91	6
7	CIndex NOD	0.373±0.213	7	1	14
8	Dunn01 NOD	0.358±0.146	108	95	15
Far Far Results					
Rank	Criterion	AMI_{corr}	ARI	Jaccard	NMI
1	Q	0.63±0.139	1	4	2
2	ASWC2 NOD	0.596±0.164	2	2	3
3	SWC2 NOD	0.57±0.159	3	3	5
4	CIndex SPD	0.565±0.132	4	25	1
5	CIndex PCD	0.446±0.142	5	1	21
6	CIndex ARD	0.433±0.25	10	106	4
7	ASWC4 NOD	0.397±0.119	15	63	11
8	SWC2 PCD	0.356±0.143	6	6	25

Table 6. Overall ranking of criteria based on AMI & Spearman's Correlation on the synthetic benchmarks with the same parameters as in Table 4 but much higher mixing parameter, .7. We can see that in these settings, PB indexes outperform modularity Q.

Rank	Criterion	AMI_{corr}	ARI	Jaccard	NMI
1	PB PCD	0.454±0.15	1	1	1
2	PB NOD	0.448±0.146	2	2	2
3	PB SPD	0.445±0.144	3	3	4
4	PB ARD	0.44±0.149	4	4	5
5	VRC ICD	0.424±0.117	5	5	3
6	Q	0.391±0.381	17	6	12
7	CIndex ARD	0.365±0.173	6	7	6
8	ASWC4 SPD	0.358±0.101	12	12	7
9	DB PCD	0.358±0.108	15	9	10
10	ASWC4 NOD	0.357±0.114	10	10	8