Noname manuscript No.
(will be inserted by the editor)

Communities Validity:

Methodical Evaluation of Community Mining Algorithms

Reihaneh Rabbany · Mansoureh Takaffoli · Justin Fagnan · Osmar R. Zaïane · Ricardo J. G. B. Campello

the date of receipt and acceptance should be inserted later

Abstract 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, that is represented in the form a graph wherein a link between two nodes indicates a relationship between them, 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 the community mining algorithms. The common practice is to evaluate the algorithms based on their performance on standard benchmarks for which we know the ground-truth. This technique is similar to external evaluation of attribute-based clustering methods. The other two well-studied clustering evaluation approaches are less explored in the community mining context; internal evaluation to statistically validate the clustering result, and relative evaluation to compare alternative clustering results. These two approaches enable us to validate communities discovered in a real world application, where the true community structure is hidden in the data. In this article, we investigate different clustering quality criteria applied for relative and internal evaluation of clustering data points with attributes, and also different clustering agreement measures used for external evaluation; and incorporate proper adaptations to make them applicable in the context of interrelated data. We further compare the performance of the proposed adapted criteria in evaluating community mining results in different settings through extensive set of experiments.

Keywords Evaluation Approaches · Quality Measures · Clustering Evaluation · Clustering Objective Function · Community Mining

Reihaneh Rabbany · Mansoureh Takaffoli · Justin Fagnan · Osmar R. Zaïane · Ricardo J. G. B. Campello
Department of Computing Science, University of Alberta, Edmonton, Canada
Tel.: +1-780-492-{3978,2860}, Fax: +1-780-492-1071
E-mail: {rabbanyk, takaffol, fagnan, zaiane, rcampell}@ualberta.ca

1 Introduction

Data Mining is the analysis of large scale data to discover meaningful patterns such as groups of data records (cluster analysis), unusual records (anomaly detection) or dependencies (association rule mining) which are crucial in a very broad range of applications. It is a multidisciplinary field that involves methods at the intersection of artificial intelligence, machine learning, statistics and database systems. 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 typically modeled by a graph of interconnected nodes, known as complex networks or information networks. Examples of such networks are hyperlink networks of web pages, citation or collaboration networks of scholars, biological networks of genes or proteins, trust and social networks of humans among others.

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 (Leskovec et al. 2005, Newman 2010). 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 surveys refer to Fortunato (2010), Porter et al. (2009). 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 introduced by Newman and Girvan (2004).

Although there have been many methods proposed for community mining, very little research has been done to explore evaluation and validation methodologies. Similar to the well-studied clustering validity methods in the Machine Learning field, we have three classes of approaches to evaluate community mining algorithms; external, internal and relative evaluation. The first two are statistical tests that measure the degree to which a clustering confirms a-priori specified scheme. The third approach compares and ranks clusterings of a same dataset discovered by different parameter settings (Halkidi et al. 2001).

In this article, we investigate the evaluation approaches of the community mining algorithms considering the same classification framework. We classify the common evaluation practices into external, internal and relative approaches, and further extend these by introducing a new set of adapted criteria and measures that are adequate for community mining evaluation. More specifically, the evaluation approaches are defined based on different clustering validity criteria and clustering similarity measures. We propose proper adaptions that these measures require to handle comparison of community mining results. Most of these validity criteria, that are introduced and adapted here, are for the first time applied to the context of interrelated data i.e. used for the community mining evaluation. These criteria not only can be used as means to

measure the goodness of discovered communities, but also as objective functions to detect communities. Furthermore, we propose the adaptation of the clustering similarity measures for the context of interrelated data, which has been overlooked in the previous literature. Apart from the evaluation, these clustering similarity measures can also be used to determine the number of clusters in a data set, or to combine different clustering results and obtain a consensus clustering (Vinh et al. 2010).

The remainder of this paper is organized as follows. In the next section, we first present some background, where we briefly introduce the well-known community mining algorithms, and the related work regarding evaluation of these algorithms. We continue the background with an elaboration on the three classes of evaluation approaches incorporating the common evaluation practices. In the subsequent section, we overview the clustering validity criteria and clustering similarity measures, and introduce our proposed adaptions of these measures for the context of interrelated data. Then, we extensively compare and discuss the performance of these adapted validity criteria and the properties of the adapted similarity measures, through a set of carefully designed experiments on real and synthetic networks. Finally, we conclude with a brief analysis of these results.

2 Background and Related Works

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 (Palla et al. 2005) 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 (Newman and Girvan 2004), 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 (Newman 2006), and Max-Min modularity (Chen et al. 2009).

Although many mining algorithms are based on the concept of modularity, Fortunato and Barthélemy (2007) 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 (Rabbany et al. 2010), 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 mine communities by utilizing information theory concepts such as compression e.g. Infomap (Rosvall and Bergstrom 2008), and entropy e.g. entropy-base (Kenley and Cho 2011). For a survey on different community mining techniques refer to (Fortunato 2010).

Fortunato (2010) shows that the different community mining algorithms discover communities from different perspective and may outperform others in specific classes of networks and have different computational complexities. Therefore, an important research direction is to evaluate and compare the results of different community mining algorithms, and select the one providing more meaningful clustering for each class of networks. An intuitive practice is to validate the results partly by a human expert (Luo et al. 2008). However, the community mining problem is NP-complete; the human expert validation is limited, and is based on narrow intuition rather than on an exhaustive examination of the relations in the given network, specially for large real networks. To validate the result of a community mining algorithm, three approaches are available: *external evaluation*, *internal evaluation*, and *relative evaluation*; which are described in the following.

2.1 Evaluation Approaches

2.1.1 External Evaluation

External evaluation involves comparing the discovered clustering with a prespecified structure, often called ground-truth, using a clustering agreement measure such as Jaccard, Adjusted Rand Index, or Normalized Mutual Information. In the case of attribute-based data, clustering similarity measures are not only used for evaluation, but also applied to determine the number of clusters in a data set, or to combine different clustering results and obtain a consensus clustering i.e. ensemble clustering (Vinh et al. 2010). In the interrelated data context, these measures are used commonly for external evaluation of community mining algorithms, where the performance of the algorithms are examined on standard benchmarks for which the true communities are known(Chen et al. 2009, Danon et al. 2005, Lancichinetti and Fortunato 2009, Orman et al. 2011). There are few and typically small real world benchmarks with known communities available for external evaluation of community mining algorithms. While the current generators used for synthesizing benchmarks with built-in ground-truth, overlook some characteristics of the real networks (Orman and Labatut 2010). Moreover, in a real-world application the interesting communities that need to be discovered are hidden in the structure of the network, thus, the discovered communities can not be validated based on the external evaluation. These facts motivate investigating the other two alternatives approaches – internal and relative evaluation. Before describing these evaluation approaches, we first elaborate more on the synthetic benchmark generators and the studies that used the external evaluation approach.

To synthesize networks with built-in ground truth, several generators have been proposed. GN benchmark (Girvan and Newman 2002) 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 (Lancichinetti et al. 2008) 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. However, having the same mixing parameter μ for all nodes, and not satisfying the densification power laws and heavy-tailed distribution are the main drawback of this benchmark.

Apart form many papers that used the external evaluation to assess the performance of their proposed algorithms, there are recent studies specifically on comparison of different community mining algorithms using the external evaluation approach. Gustafsson et al. (2006) compare hierarchical and k-means community mining on real networks and also synthetic networks generated by the GN benchmark. Lancichinetti and Fortunato (2009) compare a total of a dozen community mining algorithms. Where the performance of the algorithms is compared against the network generated by both GN and LFR benchmark. Orman et al. (2011) compare a total of five community mining algorithms on the synthetic networks generated by LFR benchmark. They first assess the quality of the different algorithms by their difference with the ground truth. Then, they perform a qualitative analysis of the identified communities by comparing their size distribution with the community size distribution of the ground truth. All these mentioned works borrow clustering agreement measures from traditional clustering literature. In this article we overview different agreement measures, and also provide an alternative measure which is adapted specifically for clustering of interrelated data.

2.1.2 Internal and Relative Evaluation

Internal evaluation techniques verify whether the clustering structure produced by a clustering algorithm matches the underlying structure of the data, using only information inherent in the data. These techniques are based on an internal criterion that measures the correlation between the discovered clustering structure and the structure of the data, represented as a proximity matrix –a square matrix in which the entry in cell (j, k) is some measure of the similarity (or distance) between the items i, and j. The significance of this correlation is examined statistically based on the distribution of the defined criteria, which is usually not known and is estimated using Monte Carlo sampling method (Theodoridis and Koutroumbas 2009). An internal criterion can also be considered as a quality index to compare different clusterings which overlaps with relative evaluation techniques. The well-known modularity of

Newman (2006) can be considered as such, which is used both to validate a single community mining result and also to compare different community mining results (Clauset 2005, Rosvall and Bergstrom 2007). Modularity is defined as the fraction of edges within communities, i.e. the correlation of adjacency matrix and the clustering structure, minus the expected value of this fraction that is computed based on the configuration model (Newman 2006). Another work that could be considered in this class is the evaluation of different community mining algorithms studied in (Leskovec et al. 2010). Where they propose network community profile (NCP) that characterizes the quality of communities as a function of their size. The quality of the community at each size is characterized by the notion of conductance which is the ratio between the number of edges inside the community and the number of edges leaving the community. Then, they compared the shape of the NCP for different algorithms over random and real networks.

Relative evaluation compares alternative clustering structures based on an objective function or quality index. This evaluation approach is the least explored in the community mining context. 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. Nevertheless, 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 (Davies and Bouldin 1979), Dunn index (Dunn 1974), and Silhouette (Rousseeuw 1987); for a recent survey refer to Vendramin et al. (2010). In the next section, 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.

3 Evaluation of Community Mining Results

In this section, we elaborate on how to evaluate results of a community mining algorithm based on external and relative evaluation. *External evaluation* of community mining results involves comparing the discovered communities with a prespecified community structure, often called ground truth, using a clustering agreement measure, while the *relative evaluation* ranks different alternative community structures based on an objective function – quality index (Theodoridis and Koutroumbas 2009). To be consistent with the terms used in attribute-based data, we use *clustering* to refer to the result of any community mining algorithm, and *partitioning* to refer to the case where the communities are mutually exclusive. Note that, in this study we only focus on non-overlapping community mining algorithms that always produce disjoint communities. Thus, in the definition of the following quality criteria and

agreement measures, *partitioning* is used instead of *clustering* which implies that these are only applicable in the case of mutually exclusive communities. In the rest, we first overview relative community quality criteria, then describe different clustering agreement measures.

3.1 Community Quality Criteria

Here, 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 originally 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 proximity measures that could be used in graphs. Then, we present generalizations of criteria that could use any notion of proximity.

3.1.1 Proximity 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 proximity between n_i and n_j, $p_{ij} = p(i, j)$, can be computed by one of the following distance or similarity measures. The latter is more typical in the context of interrelated data, therefore, we tried to plug-in similarities in the relative criteria definitions. When it is not straightforward, we used the inverse of the similarity index to obtain the according dissimilarity/distance. For avoiding division by zero, when P_{ij} is zero, if it is a similarity ϵ and if it is distance $1/\epsilon$ is returned, where ϵ is a very small number, i.e. 10E-9.

***Shortest Path (SP)*:** 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 (A)*:** similarity between the two nodes n_i and n_j is considered their incident edge weight, $p^A_{ij} = A_{ij}$; accordingly, the distance between these nodes is derived as:

$$d^A_{ij} = M - p^A_{ij} \tag{1}$$

where M is the maximum edge weight in the graph; $M = A_{max} = \max_{ij} A_{ij}$.

***Adjacency Relation (AR)*:**distance between two nodes is their structural dissimilarity, that is computed by the difference between their immediate neighbourhoods (Wasserman and Faust 1994):

$$d^{AR}_{ij} = \sqrt{\sum_{k \neq j,i} (A_{ik} - A_{jk})^2} \tag{2a}$$

This definition is not affected by the (non)existence of an edge between the two nodes. To remedy this, Augmented AR ($\hat{AR}$) can be defined as;

$$d_{ij}^{\hat{AR}} = \sqrt{\sum_k (\hat{A}_{ik} - \hat{A}_{jk})^2} \tag{2b}$$

where $\hat{A}_{ij}$ is equal to A_{ij} if $i \neq j$ and A_{max} otherwise.

***Neighbour Overlap (NO)*:** similarity between two nodes is the ratio of their shared neighbours (Fortunato 2010):

$$p_{ij}^{NO} = |\aleph_i \cap \aleph_j| / |\aleph_i \cup \aleph_j| \tag{3a}$$

where $\aleph_i$ is the set of nodes directly connected to n_i, $\aleph_i = \{n_k | A_{ik} \neq 0\}$. The corresponding distance is derived as $d_{ij}^{NO} = 1 - p_{ij}^{NO}$.

There is a close relation between this measure and the previous one, since d^{AR} can also be computed as: $d_{ij}^{AR} = \sqrt{|\aleph_i \cup \aleph_j| - |\aleph_i \cap \aleph_j|}$. while $d_{ij}^{\hat{AR}}$ is also derived from the same formula, if neighbourhoods are considered closed, i.e. $\hat{\aleph}_i = \{n_k | \hat{A}_{ik} \neq 0\}$. We also consider the closed neighbour overlap similarity, $p^{\hat{NO}}$, with the same analogy that two nodes are more similar if directly connected. The closed overlap similarity, $p^{\hat{NO}}$, could be rewritten in terms of the adjacency matrix which can be straightforwardly generalized for weighted graphs.

$$p_{ij}^{\hat{NO}} = \frac{\sum_k \hat{A}_{ik}\hat{A}_{jk}}{\sum_k [\hat{A}_{ik}^2 + \hat{A}_{jk}^2 - \hat{A}_{ik}\hat{A}_{jk}]} \tag{3b}$$

$$p_{ij}^{\hat{NOV}} = \frac{\sum\limits_k (\hat{A}_{ik} + \hat{A}_{jk})(\hat{A}_{ik} + \hat{A}_{jk}) - \sum\limits_k (\hat{A}_{ik} - \hat{A}_{jk})(\hat{A}_{ik} - \hat{A}_{jk})}{\sum\limits_k (\hat{A}_{ik} + \hat{A}_{jk})(\hat{A}_{ik} + \hat{A}_{jk}) + \sum\limits_k (\hat{A}_{ik} - \hat{A}_{jk})(\hat{A}_{ik} - \hat{A}_{jk})} \tag{3c}$$

***Topological Overlap (TP)*:** similarity measures the normalized overlap size of the neighbourhoods (Ravasz et al. 2002), which we generalize as:

$$p_{ij}^{TP} = \frac{\sum_{k \neq j,i} (A_{ik}A_{jk}) + A_{ij}^2}{min(\sum_k A_{ik}^2, \sum_k A_{jk}^2)} \tag{4}$$

and the corresponding distance is derived as $d_{ij}^{TO} = 1 - p_{ij}^{TO}$.

***Pearson Correlation (PC)*:** coefficient between two nodes is the correlation between their corresponding rows of the adjacency matrix:

$$p_{ij}^{PC} = \frac{\sum_k (A_{ik} - \mu_i)(A_{jk} - \mu_j)}{N\sigma_i\sigma_j} \tag{5a}$$

where N is the number of nodes, the average $\mu_i = (\sum_k A_{ik})/N$ and the variance $\sigma_i = \sqrt{\sum_k (A_{ik} - \mu_i)^2/N}$. This correlation coefficient lies between -1 (when the two nodes are most similar) and 1 (when the two nodes are most

dissimilar). Most relative clustering criteria are defined assuming distance is positive, therefore we also consider the normalized version of this correlation, i.e. $p^{NPC} = (p_{ij}^{PC} + 1)/2$. Then, the distance between two nodes is computed as $d_{ij}^{(N)PC} = 1 - p_{ij}^{(N)PC}$.

In all the above proximity measures, the iteration over all other nodes can be limited to iteration over the nodes in the union of neighbourhoods. More specifically, in the formulas, one can use $\sum_{k \in \hat{\aleph}_i \cup \hat{\aleph}_j}$ instead of $\sum_{k=1}^{N}$. This will make the computation local and more efficient, especially in case of large networks. This trick will not work for the current definition of the pearson correlation, however, it can be applied if we reformulate it as follows:

$$p_{ij}^{PC} = \frac{\sum_k A_{ik}A_{jk} - (\sum_k A_{ik})(\sum_k A_{jk})/N}{\sqrt{((\sum_k A_{ik}^2) - (\sum_k A_{ik})^2/N)((\sum_k A_{jk}^2) - (\sum_k A_{jk})^2/N)}} \tag{5b}$$

We also consider this correlation based on $\hat{A}$, $p^{\hat{PC}}$, so that the existence of an edge between the two nodes, increases their correlation. Note that since we are assuming a self edge for each node, $\hat{N} = N + 1$ should be used.

The above formula can be further rearranged as follows:

$$p_{ij}^{PC} = \frac{\sum\limits_k \left[A_{ik}A_{jk} - (\sum\limits_{k'} A_{ik'})(\sum\limits_{k'} A_{jk'})/N^2\right]}{\sqrt{(\sum\limits_k \left[A_{ik}^2 - (\sum\limits_{k'} A_{ik'})^2/N^2\right])(\sum\limits_k \left[A_{jk}^2 - (\sum\limits_{k'} A_{jk'})^2/N^2\right])}} \tag{5c}$$

Where if the k iterates over all nodes, it is equal to the original pearson correlation; however, this is not true if it only iterates over the union of neighbourhoods, $\sum_{k \in \hat{\aleph}_i \cup \hat{\aleph}_j}$, which we call pearson overlap (NPO).

***Number of Paths (NP)*:** between two nodes is the sum of all the paths between them, which is a notion of similarity. For the sake of time complexity, we consider paths of up to a certain number of hops i.e. 2 and 3. The number of paths of length l between nodes n_i and n_j can be computed as $np_{ij}^l = (A^l)_{ij}$. More specifically we have:

$$np_{ij}^1 = A_{ij}, \quad np_{ij}^2 = \sum_k A_{ik}A_{jk}, \quad np_{ij}^3 = \sum_{kl} A_{ik}A_{kl}A_{jl} \tag{6a}$$

where p^{NP} is defined as a combination these; $p^{NP^2} = np^1 + np^2$ and $p^{NP^3} = np^1 + np^2 + np^3$. We also considered two alternatives for this combination;

$$p^{NP_L^3} = np^1 + \frac{np^2}{2} + \frac{np^3}{3}, \quad and \quad p^{NP_E^3} = np^1 + \sqrt[2]{np^2} + \sqrt[3]{np^3} \tag{6b}$$

***Modularity (M)*:** similarity is defined inspired by the Modularity of Newman (2006) as:

$$p_{ij}^M = A_{ij} - \frac{(\sum_k A_{ik})(\sum_k A_{jk})}{\sum_{kl} A_{kl}} \tag{7a}$$

$$p_{ij}^{MD} = \frac{A_{ij}}{\frac{(\sum_k A_{ik})(\sum_k A_{jk})}{\sum_{kl} A_{kl}}} \tag{7b}$$

The distance is derived as $1 - p^{M(D)}$.

***ICloseness (IC)*:** similarity between two nodes is computed as the inverse of the connectivity between their scored neighbourhoods:

$$p_{ij}^{IC} = \frac{\sum\limits_{k \in \aleph_i \cap \aleph_j} ns(k,i) ns(k,j)}{\sum\limits_{k \in \aleph_i} ns(k,i)^2 + \sum\limits_{k \in \aleph_j} ns(k,j)^2 - \sum\limits_{k \in \aleph_i \cap \aleph_j} ns(k,i) ns(k,j)} \tag{8a}$$

$$p_{ij}^{ICV} = \frac{a-b}{a+b}\,, \quad where \tag{8b}$$

$$a = \sum_{k \in \hat{\aleph}_i \cup \hat{\aleph}_j} (ns(k,i) + ns(k,j))(ns(k,i) + ns(k,j))$$

$$b = \sum_{k \in \hat{\aleph}_i \cup \hat{\aleph}_j} (ns(k,i) - ns(k,j))(ns(k,i) - ns(k,j))$$

where $ns(k,i)$ denotes the neighbouring score between nodes k and i that is computed iteratively; for complete formulation refer to (Rabbany and Zaïane 2011). In Icloseness, the neighbourhood is defined with a depth; here we consider 3 variations: direct neighbourhood (IC1), neighbourhood of depth 2 i.e. neighbours up to one hop apart (IC2) and up to two hops apart (IC3). The distance is then derived as $d^{IC(V)} = 1 - p^{IC(V)}$.

3.1.2 Community Centroid

In addition to the notion of proximity 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 the community C can be obtained as the medoid:

$$\overline{C} = \arg\min_{m \in C} \sum_{i \in C} d(i,m) \tag{9}$$

3.1.3 Relative Validity Criteria

Here, we present our generalizations of well-known clustering validity criteria defined as quality measures for internal or relative evaluation of clustering results. All these criteria are originally defined based on distances between data points, which in all cases is the Euclidean or other inner product norms of difference between their vectors of attributes; refer to (Vendramin et al. 2010) 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 proximity 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 partitioning $C = \{C_1 \cup C_2 \cup ... \cup C_k\}$ of N data points, where $\overline{C}$ denotes the (generalized) centroid of data points belonging to C and $d(i,j)$ denotes the (generalized) distance between point i and point j. The quality of C can be measured using one of the following criteria.

***Variance Ratio Criterion (VRC)*:** measures the ratio of the between-cluster/community distances to within-cluster/community distances which could be generalized as follows:

$$VRC = \frac{\sum_{l=1}^{k} |C_l| d(\overline{C}_l, \overline{C})}{\sum_{l=1}^{k} \sum_{i \in C_l} d(i, \overline{C}_l)} \times \frac{N-k}{k-1} \tag{10}$$

where $\overline{C}_l$ is the centroid of the cluster/community C_l, and $\overline{C}$ is the centroid of the entire data/network. Consequently $d(\overline{C}_l, \overline{C})$ is measuring the distance between centroid of cluster C_l and the centroid of the entire data, while $d(i, \overline{C}_l)$ is measuring the distance between data point i and its cluster centroid.

The original clustering formula proposed by Calinski and Harabasz (1974) for attributes vectors is obtained if the centroid is fixed to averaging of vectors of attributes and distance to (square of) Euclidean distance. Here we use this formula with one of the proximity measures mentioned in the preious section; if it is a similarity measure, we either transform the similarity to its distance form and apply the above formula, or we use it directly as a similarity and inverse the ratio to within/between while keeping the normalization, the latter approach is distinguished in the experiments as VRC'.

***Davies-Bouldin index (DB)*:** calculates the worst-case within-cluster to between-cluster distances ratio averaged over all clusters/communities (Davies

and Bouldin 1979):

$$DB = \frac{1}{k}\sum_{l=1}^{k}\max_{m\neq l}((\overline{d}_l + \overline{d}_m)/d(\overline{C}_l, \overline{C}_m)) \ , \tag{11}$$
$$where \quad \overline{d}_l = \frac{1}{|C_l|}\sum_{i\in C_l} d(i, \overline{C}_l)$$

If used directly with a similarity measure, we change the max in the formula to min and the final criterion becomes a maximizer instead of minimizer, which is denoted by DB'.

***Dunn index*:** 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) (Dunn 1974):

$$Dunn = \min_{l\neq m}\{\frac{\delta(C_l, C_m)}{max_p \Delta(C_p)}\} \tag{12}$$

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)*:** measures the average silhouette scores, which is computed individually for each data point. The silhouette score of a point shows the goodness of the assignment of this point to the community it belongs to, by calculating the normalized difference between the distance to its nearest neighbouring community and the distance to its own community (Rousseeuw 1987). Taking the average one has:

$$SWC = \frac{1}{N}\sum_{l=1}^{k}\sum_{i\in C_l}\frac{\min_{m\neq l} d(i, C_m) - d(i, C_l)}{\max\{\min_{m\neq l} d(i, C_m), d(i, C_l)\}} \tag{13}$$

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 SWC0) (i.e. $1/|C_l|\sum_{j\in C_l} d(i,j)$) or could be the distance to its centroid (called SWC1) (i.e. $d(i, \overline{C_l})$). An alternative

formula for Silhouette is proposed in (Vendramin et al. 2010) :

$$ASWC = \frac{1}{N}\sum_{l=1}^{k}\sum_{i\in C_l}\frac{\min_{m\neq l} d(i,C_m)}{d(i,C_l)} \tag{14}$$

Similar to DB, if used directly with a similarity proximity measure, we change the min to max and the final criterion becomes a minimizer instead of maximizer, which is denoted by $(A)SWC'$.

***PBM*:** criterion is based on the within-community distances and the maximum distance between centroids of communities(Pakhira and Dutta 2011):

$$PBM = \frac{1}{k}\times\frac{\max_{l,m} d(\overline{C}_l,\overline{C}_m)}{\sum_{l=1}^{k}\sum_{i\in C_l} d(i,\overline{C}_l)} \tag{15}$$

Again similar to DB, here also if used directly with a similarity measure, we change the max to min and consider the final criterion as a minimizer instead of maximizer, which is denoted by PBM'.

***C-Index*:** criterion compares the sum of the within-community distances to the worst and best case scenarios (Dalrymple-Alford 1970). 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.

$$CIndex = \frac{\theta - \min\theta}{\max\theta - \min\theta}\ ,\ where\quad \theta = \frac{1}{2}\sum_{l=1}^{k}\sum_{i,j\in C_l} d(i,j) \tag{16}$$

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

C-Index can be directly used with a similarity measure as a maximization criterion, whereas with a distance measure it is a minimizer. This is also true for the two following criteria.

***Z-Statistics*:** criterion is defined similar to C-Index (Hubert and Levin 1976):

$$ZIndex = \frac{\theta - E(\theta)}{\sqrt{var(\theta)}}\ ,\quad where\quad \bar{d} = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N} d(i,j)\ , \tag{17}$$

$$E(\theta) = \Theta\times\bar{d}\ ,\qquad Var(\theta) = \frac{1}{4}\sum_{l=1}^{k}\sum_{i,j\in C_l}(d(i,j)-\bar{d})^2$$

***Point-Biserial (PB)*:** This criterion computes the correlation of the distances between nodes and their cluster co-membership which is dichotomous variable (Milligan and Cooper 1985). Intuitively, nodes that are in the same community should be separated by shorter distances than those which are not:

$$PB = \frac{M_1 - M_0}{S}\sqrt{\frac{m_1 m_0}{m^2}} \tag{18}$$

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*:** Modularity is the well-known criterion proposed by Newman et al. (Newman and Girvan 2004) 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:

$$Q = \frac{1}{2E}\sum_{l=1}^{k}\sum_{i,j \in C_l}[A_{ij} - \frac{\sum_k A_{ik}\sum_k A_{kj}}{2E}] \tag{19}$$

3.1.4 Computational Complexity

The computational complexity of different clustering validity criteria is provided in the previous work by Vendramin et al. (2010). For the adapted criteria, the time complexity of the indexes is affected by the cost of the chosen proximity measure. All the proximity measures we introduced here can be computed in linear time, $\mathcal{O}(n)$, except for the A (adjacency) which is $\mathcal{O}(1)$, the NP (number of paths) which is $\mathcal{O}(n^2)$ and the IC (Icloseness) which is $\mathcal{O}(E)$. However, for the case of sparse graphs and using a proper graph data structure such as incidence list, this complexity can be reduced to $O(\hat{d})$, where $\hat{d}$ is the average degree in the network, i.e. the average neighbors of a node in the network. For example lets revisit the formula for AR (adjacency relation): $d_{ij}^{AR} = \sqrt{\sum_{k \neq j,i}(A_{ik} - A_{jk})^2}$. In this formula we can change $\sum_k$ to $\sum_{k \in \aleph_i \cup \aleph_j}$ since the expression $(A_{ik} - A_{jk})^2$ is zero for other values of k, i.e. for nodes that are not neighbor to either i or j and therefore have $A_{ik} = A_{jk} = 0$. The same trick could be applied to other proximity measures.

The other cost that should be considered is the cost of computing the medoid of m data points, which is $\mathcal{O}(pm^2)$, where p is the cost of the proximity measure. Therefore the VRC criterion that require computing the overall centroid, is in order of $\mathcal{O}(pn^2)$. This is while the VRC for traditional clustering is linear with respect to the size of the dataset, since it uses averaging for computing the centroid which is $\mathcal{O}(n)$. Similarly, any other measure that requires computing all the pairwise distances will have the $\Omega(pn^2)$. This holds for the adapted $Dunn$ index which is in order of $\mathcal{O}(pn^2)$, becuase for finding the minimum distances between any two clusters, it requires to compute the distances between all pair of nodes. Similarly, the $ZIndex$ computes all the pairwise distances, and is in order of $\mathcal{O}(pn^2)$. The same also holds for the PB. The $CIndex$ is even more expensive since it not only computes all the pairwise distances but also sorts them, and hence is in order of $\mathcal{O}(n^2(p+logn))$. These orders (except for VRC) are along the computational complexities previously reported in Vendramin et al. (2010), where the cost of the p is the size of the feature vectors there.

The adapted DB and PBM, on the other hand, do not require computing the medoid of the whole dataset nor all pairwise distances. Instead they only compute the medoid of each cluster, which makes them in $\Omega(pk\hat{m}^2)$, where k is the number of clusters and the $\hat{m}$ is the average size of the clusters. Consequently, this term will be added to the complexity of these criteria, giving them the order of $\mathcal{O}(p(n+k^2+k\hat{m}^2))$. Finally for the silhouette criterion, the $(A)SWC0$ that uses the average distance, has the order of $\mathcal{O}(pn^2)$, however the order for $(A)SWC1$ is simplified to $\mathcal{O}(kp(n+\hat{m}^2))$ since it uses the distance to centroid instead of averaging. The latter is similar to the order for modularity Q which is $\mathcal{O}(k(n+\hat{m}^2))$. To sum up, none of the adapted criteria is significantly superior or inferior in terms of its order, therefore one should focus on which criterion is more appropriate according to its performance which is demonstrated in the experiments.

3.2 Clustering Agreement Measures

Here, we formally review different well-studied partitioning agreement measures used in the *external evaluation* of clustering results. Consider two different partitionings U and V of data points in D. There are several measures to examine the agreement between U and V, originally introduced in the Machine Learning field. These measures assume that the partitionings are disjoint and cover the dataset. More formally, consider D consist of n data items, $D = \{d_1, d_2, d_3 \dots d_n\}$ and let $U = \{U_1, U_2 \dots U_k\}$ denotes the k clusters in U then $D = \cup_{i=1}^{k} U_i$ and $U_i \cap U_j = \emptyset \quad \forall i \neq j$.

3.2.1 Pair Counting Based Measures

Clustering agreement measures are originally introduced based on counting the pairs of data items that are in the same/different partitions in U and

V. Generally, each pair (d_i, d_j) of data items is classified into one of these four groups based on their co-membership in U and V; which results in the following four pair counts:

$V \backslash U$	Same	Different
Same	M_{11}	M_{10}
Different	M_{01}	M_{00}

These pair counts can be translated considering the contingency table (Hubert and Arabie 1985). The contingency table, consists of all the possible overlaps between each pair of clusters in U and V, where $n_{ij} = |U_i \cap V_j|$ and $n_{i.} = \sum_j n_{ij}$. Considering the contingency table, we could compute the pair counts using following formulae.

	V_1	V_2	...	V_r	sums
U_1	n_{11}	n_{12}	...	n_{1r}	$n_{1.}$
U_2	n_{21}	n_{22}	...	n_{2r}	$n_{2.}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
U_k	n_{k1}	n_{k2}	...	n_{kr}	$n_{k.}$
sums	$n_{.1}$	$n_{.2}$	...	$n_{.r}$	n

$$M_{10} = \sum_{i=1}^{k} \binom{n_{i.}}{2} - \sum_{i=1}^{k}\sum_{j=1}^{r} \binom{n_{ij}}{2}, \quad M_{01} = \sum_{j=1}^{r} \binom{n_{.j}}{2} - \sum_{i=1}^{k}\sum_{j=1}^{r} \binom{n_{ij}}{2}$$

$$M_{11} = \sum_{i=1}^{k}\sum_{j=1}^{r} \binom{n_{ij}}{2}, \quad M_{00} = \binom{n}{2} + \sum_{i=1}^{k}\sum_{j=1}^{r} \binom{n_{ij}}{2} - \sum_{i=1}^{k} \binom{n_{i.}}{2} - \sum_{j=1}^{r} \binom{n_{.j}}{2}$$

These *pair counts* have been used to define a variety of different clustering agreement measures. In the following, we briefly explain the most common pair counting measures; the reader can refer to Albatineh et al. (2006) for a recent survey.

***Jaccard*:** similarity coefficient measures similarity of two sets as the fraction of their intersection to their union. If we consider co-membership of data points in the same or different clusters as a binary variable, Jaccard agreement between co-memberships in clustering U and V is defined as follows (Manning et al. 2008):

$$J = \frac{M_{11}}{M_{01} + M_{10} + M_{11}} \tag{20}$$

***Rand Index*:** is defined similar to Jaccard, but it also prizes the pairs that belong to different clusters in both partitioning (Manning et al. 2008):

$$RI = \frac{M_{11} + M_{00}}{M_{11} + M_{01} + M_{10} + M_{00}} = 1 + \frac{1}{n^2 - n}(2\sum_{i=1}^{k}\sum_{j=1}^{r} n_{ij}^2 - (\sum_{i=1}^{k} n_{i.}^2 + \sum_{j=1}^{r} n_{.j}^2)) \quad (21)$$

***F-measure*:** is a weighted mean of the precision (P) and recall (R) (Manning et al. 2008) defined as:

$$F_\beta = \frac{(\beta^2 + 1)PR}{\beta^2 P + R} \text{ , } where \quad P = \frac{M_{11}}{M_{11} + M_{10}} \text{ , } \quad R = \frac{M_{11}}{M_{11} + M_{01}} \quad (22)$$

The parameter β indicates how much recall is more important than precision. The two common values for β are 2 and .5; the former weights recall higher than precision while the latter prizes the precision more.

3.2.2 Information Theoretic Based Measures

There is also a family of information theoretic based measures defined based on *Mutual Information* between the two clusterings. These measures consider the cluster overlap sizes of U and V, n_{ij}, as a joint distribution of two random variables – the cluster memberships in U and V. Then, entropy of cluster U ($H(U)$), joint entropy of U and V ($H(U,V)$), and their mutual information($I(U,V)$) are defined as follows; based on which several clustering agreements have been derived.

$$H(U) = -\sum_{i=1}^{k} \frac{n_{i.}}{n} \log(\frac{n_{i.}}{n}) \text{ , } \quad H(V) = -\sum_{j=1}^{r} \frac{n_{.j}}{n} \log(\frac{n_{.j}}{n}) \quad (23a)$$

$$H(U,V) = -\sum_{i=1}^{k}\sum_{j=1}^{r} \frac{n_{ij}}{n} \log(\frac{n_{ij}}{n}) \quad (23b)$$

$$I(U,V) = \sum_{i=1}^{k}\sum_{j=1}^{r} \frac{n_{ij}}{n} \log(\frac{n_{ij}/n}{n_{i.}n_{.j}/n^2}) \quad (23c)$$

***Variation of Information (VI)*:** is specifically proposed for comparing two different clusterings as (Meil 2007):

$$VI(U,V) = \sum_{i=1}^{k}\sum_{j=1}^{r} \frac{n_{ij}}{n} \log(\frac{n_{i.}n_{.j}/n^2}{n_{ij}^2/n^2}) \quad (24)$$

All the pair counting measures defined previously have a fixed range of $[0, 1]$, i.e. are *normalized*. The above information theoretic definitions however are

not normalized; the mutual information for example, ranges between $(0, logk]$, while the range for variation of information is $[0, 2\log\max(k, r)]$ (Wu et al. 2009). Therefore, to be applicable for comparing different clusterings, the mutual information has been normalized in several different ways (Vinh et al. 2010):

***Normalized Mutual Information (NMI)*:** is defined in several ways (Vinh et al. 2010), while the followings are are the most commonly used forms:

$$NMI_{sum} = \frac{2I(U,V)}{H(U)+H(V)}, \quad NMI_{sqrt} = \frac{I(U,V)}{\sqrt{H(U)H(V)}} \tag{25}$$

Vinh et al. (2010) discussed another important property that a proper clustering agreement measure should comply with: *correction for chance*, which is adjusting the agreement index in a way that the expected value for agreements no better than random becomes a constant, e.g. 0. As an example, consider that the agreement between a clustering and the ground-truth is measured as .7 using an unadjusted index, i.e. a measure without a constant baseline where its baseline may be .6 in one settings or .2 in another; therefore this .7 value can not be interpreted directly as strong or weak agreement without knowing the baseline.

None of the measures we reviewed to this point are adjusted to have a constant baseline value. The adjustment/correction for chance is usually performed using the following formula which is defined based on the expected value of the index, $E(index)$, and its upper bound, $Max(index)$ (Hubert and Arabie 1985):

$$adjusted\_index = \frac{index - E(index)}{Max(index) - E(index)} \tag{26}$$

***Adjusted Rand Index*:** is the adjusted version of Rand Index (ARI) which is proposed by Hubert and Arabie (1985), which returns 0 for agreements no better than random and ranges between $[-1, 1]$.

$$ARI = \frac{\sum_{i=1}^{k}\sum_{j=1}^{r}\binom{n_{ij}}{2} - \sum_{i=1}^{k}\binom{n_{i.}}{2}\sum_{j=1}^{r}\binom{n_{.j}}{2}/\binom{n}{2}}{1/2[\sum_{i=1}^{k}\binom{n_{i.}}{2} + \sum_{j=1}^{r}\binom{n_{.j}}{2}] - \sum_{i=1}^{k}\binom{n_{i.}}{2}\sum_{j=1}^{r}\binom{n_{.j}}{2}/\binom{n}{2}} \tag{27}$$

The necessity of correction for chance for the information theoretic based measures has been discussed quite recently by Vinh et al. (2009; 2010). They have shown that the unadjusted indexes such as the widely-used NMI, do not have a constant baseline and in fact are biased in favor of large number of clusters. We will illustrate this bias of the unadjusted indexes further in the experiments.

***Adjusted Mutual Information (AMI)*:** is proposed by Vinh et al. (2010) using the similar adjustment approach as the ARI, please refer to the

main source, or the supplementary materials for the exact formula. They have shown that after correction for chance, the adjusted variation of information, AVI, is equivalent to AMI when the $1/2(H(U)+H(V))$ upper bound is used, i.e.:

$$AVI = AMI = \frac{I(U,V) - E(I(U,V))}{1/2(H(U) + H(V)) - E(I(U,V))} \tag{28}$$

3.2.3 Graph Agreement Measures

The result of a community mining algorithm is a set of sub-graphs. To also consider the structure of these sub-graphs in the agreement measure, we first define a weighted version of these measures; where nodes with more importance affect the agreement measure more. Second, we alter the measures to directly assess the structural similarity of these sub-graphs by focusing on the edges instead of nodes.

More specifically, instead of $n_{ij} = |U_i \cap V_j|$, we first use:

$$\eta_{ij} = \sum_{l \in U_i \cap V_j} w_l \tag{29}$$

where w_l is the weight of item l. If we assume all items are weighted equally as 1, then $\eta_{ij} = n_{ij}$. Instead, we can consider weight of a node equal to its degree in the graph. Using this degree weighted index can be more informative for comparing agreements between community mining results, since nodes with different degrees have different importance in the network, and therefore should be weighted differently in the agreement index. Another possibility is to use the clustering coefficient of a node as its weight, so that nodes that contribute to more triangles – have more connected neighbours – weight more.

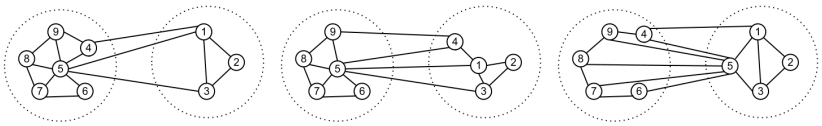
Second, we consider the structure in a more direct way by counting the edges that are common between U_i and V_j. More formally, we define;

$$\xi_{ij} = \sum_{k,l \in U_i \cap V_j} A_{kl} \tag{30}$$

which sums all the edges in the overlap of cluster U_i and V_j. Applying ξ_{ij} instead of n_{ij}, in the agreement measures defined above, is more appropriate when dealing with inter-related data, since it takes into account the structural information of data i.e. the relationship between data points, whereas the original agreement measures that completely overlook the existence of these relationships, i.e edges. For more clarification see Figure 1.

4 Comparison Methodology and Results

In this section, we first describe our experimental settings. Then, we examine behaviour of different external indexes in comparing different community mining results. Next, we report the performances of the proposed community quality criteria in relative evaluation of communities.



(a) V (b) U_1 (c) U_2

Fig. 1 Example for the benefits of the altered agreement indexes for graphs. Partitioning U_1 and $U2$ of the same graph with true partitioning V. Both partitionings have the exact same contingency table with V, $\{\{5,0\}\{1,3\}\}$, and therefore the same agreement value regardless of the agreement method used, however, U_1 looks more similar to the true partitioning V, which is reflected in the adapted measure: in the degree weighted index, we have $\eta(U_1,V) = \{\{18,0\}\{3,9\}\}$ and $\eta(U_2,V) = \{\{14,0\}\{7,9\}\}$. And in the edge based measure we have $\xi(U_1,V) = \{\{6,0\}\{0,3\}\}$ and $\xi(U_2,V) = \{\{4,0\}\{0,3\}\}$.

4.1 Experiment Settings

We have used three set of benchmarks as our datasets: Real, GN and LFR. The Real dataset consists of five well-known real-world benchmarks: Karate Club (weighted) by Zachary (Zachary 1977), Sawmill Strike data-set (Nooy et al. 2004), NCAA Football Bowl Subdivision (Girvan and Newman 2002), and Politician Books from Amazon (Krebs 2004). The GN and LFR datasets, each include 10 realizations of the GN and LFR synthetic benchmarks (Lancichinetti et al. 2008), which are the benchmarks widely in use for community mining evaluation.

For each graph in our datasets, we generate different partitionings to sample the space of all possible partitionings. For doing so, given the perfect partitioning, we generate different randomized versions of the true partitioning by randomly merging and splitting communities and swapping nodes between them. The sampling procedure is described in more details in the supplementary materials. The set of the samples obtained covers the partitioning space in a way that it includes very poor to perfect samples.

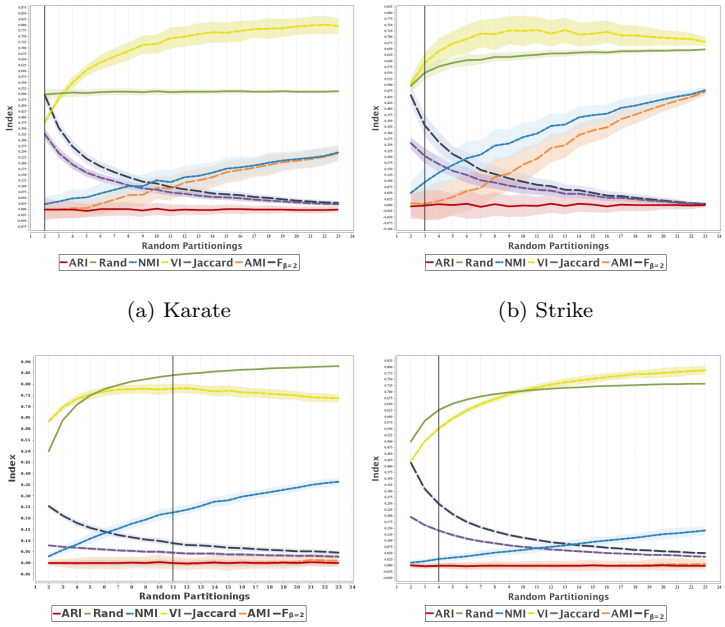
4.2 Agreement Indexes Experiments

Here we first examine two desired properties for general clustering agreement indexes, and then we illustrate these properties in our adapted indexes for graphs.

4.2.1 Bias of Unadjusted Indexes

In Figure 2, we show *the bias of the unadjusted indexes*, where the average agreement of random partitionings to a true partitioning is plotted as a function of number of clusters, similar to the experiment performed in (Vinh et al. 2010). We can see that the average agreement increases for the unadjusted indexes when the number of clusters increases, while the adjusted rand index,

ARI, is unaffected. Interestingly, we do not observe the same behaviour from AMI in all the datasets, while it is unaffected in football and GN datasets (where $k \ll N$), it increases with the number of clusters in the strike and karate dataset (where $k \ll N$ is not true).



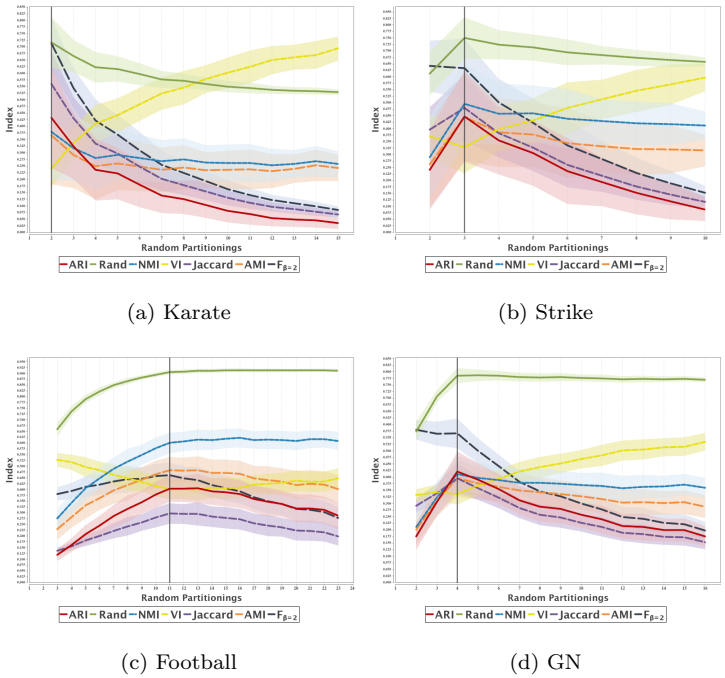
(a) Karate (b) Strike

(c) Football (d) GN

Fig. 2 Necessity of adjustment of external indexes for agreement at chance. Here we generated 100 sample partitionings for each k, then for each sample, we computed its agreement with the true partitioning for that dataset. The average and variance of these agreements are plotted as a function of the number of clusters. We can see that the unadjusted measures of $Rand$, VI, $Jaccard$, $Fmeasure$ and NMI tend to increase/decrease as the the number of clusters in the random partitionings increases. While the Adjusted Rand Index (ARI) is unaffected and always returns zero for agreements at random.

4.2.2 Knee Shape

Figure 3, illustrates the behaviour of these criteria on different fragmentations of the ground-truth as a function of the number of clusters. The ideal behaviour is that the index should return relatively low scores for partitionings/fragmentations in which the number of clusters is much lower or higher than what we have in the ground-truth. In this figure, we can see that ARI exhibits this *knee shape* while NMI does not show this clearly. Table 1, reports the average correlation of these external indexes over these four datasets. Here we used the similar sampling procedure described before but we generate



(a) Karate (b) Strike

(c) Football (d) GN

Fig. 3 Behaviour of different external indexes around the true number of clusters. We can see that the *ARI* exhibits a clear knee behaviour, i.e., its values are relatively lower for partitionings with too many or too few clusters. While others such as *NMI* and *Rand* comply less with this knee shape.

merge and split versions separately, so that the obtained samples are fragmentations of the ground-truth obtained from repeated merging or splitting. Refer to the supplementary materials for the detailed sampling procedure.

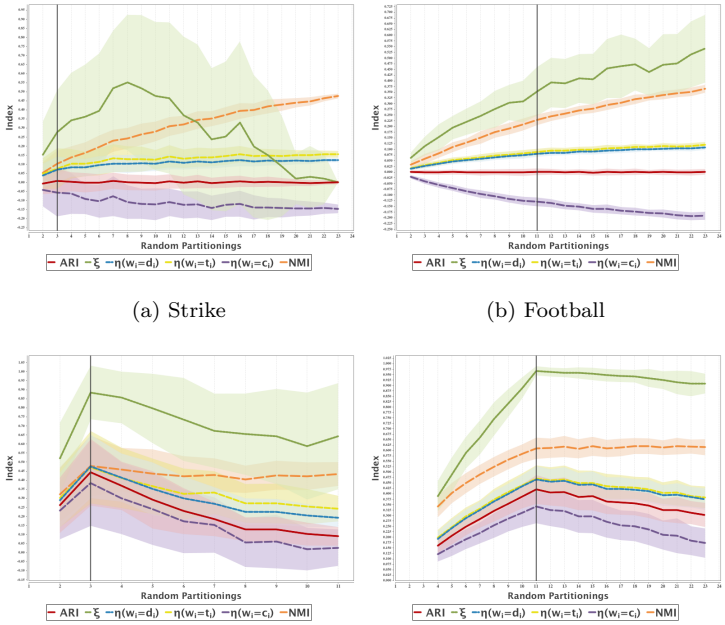
There are different ways to compute the correlation between two vectors. The classic options are Pearson Product Moment coefficient or the Spearman's Rank correlation coefficient. The reported results in our experiments are based on the Spearman's Correlation, since we are interested in the correlation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

4.2.3 Graph Partitioning Agreement Indexes

Finally, we examine the adapted versions of agreement measures described in Section 3.2.3. Figure 4 shows the constant baseline of these adapted criteria for agreements at random, and also the knee shape of the adapted measures around the true number of clusters, same as what we have for the original

Index	ARI	Rand	NMI	VI	Jaccard	AMI	$F_{\beta=2}$
ARI	1	0.73±0.18	0.67±0.07	-0.80±0.17	0.85±0.08	0.76±0.15	0.64±0.16
Rand	0.73±0.18	1	0.83±0.12	-0.46±0.42	0.41±0.32	0.71±0.11	0.13±0.46
NMI	0.67±0.07	0.83±0.12	1	-0.43±0.27	0.31±0.17	0.93±0.07	0.04±0.10
VI	-0.80±0.17	-0.46±0.42	-0.43±0.27	1	-0.93±0.02	-0.54±0.27	-0.82±0.21
Jaccard	0.85±0.08	0.41±0.32	0.31±0.17	-0.93±0.02	1	0.46±0.28	0.90±0.13
AMI	0.76±0.15	0.71±0.11	0.93±0.07	-0.54±0.27	0.46±0.28	1	0.25±0.13
$F_{\beta=2}$	0.64±0.16	0.13±0.46	0.04±0.10	-0.82±0.21	0.90±0.13	0.25±0.13	1

Table 1 Correlation between external indexes averaged for datasets of Figure 3, computed based on Spearman's Correlation. Here we can see for example that ARI behaves more similar to, has a higher correlation with, AMI compared to NMI respectively.



(a) Strike (b) Football

(c) Strike (d) Football

Fig. 4 Adapted agreement measures for graphs. On top we see that the adapted measures, specially the weighted indexes by degree (d_i) and the number of triangles (t_i), are adjusted by chance, which can not be seen for the structural edge based version (ξ). The bottom figures illustrate the perseverance of the knee behaviour in the adapted measures.

ARI. Therefore, one can safely apply one of these measures depending on the application at hand. Table 2 summarizes the correlation between each pair of the external measures.

In the following we compare the performance of different quality indexes, defined in Section 3.1, in relative evaluation of clustering results.

4.3 Quality Indexes Experiments

The performance of a criterion could be examined by how well it could rank different partitionings of a given dataset. More formally, consider for the dataset

Index	ARI	ξ	$\eta_{w_i=d_i}$	$\eta_{w_i=t_i}$	$\eta_{w_i=c_i}$	NMI
ARI	1±0	0.571±0.142	0.956±0.031	0.819±0.135	0.838±0.087	0.736±0.096
ξ	0.571±0.142	1±0	0.623±0.133	0.572±0.169	0.45±0.109	0.497±0.2
$\eta_{w_i=d_i}$	0.956±0.031	0.623±0.133	1±0	0.876±0.097	0.777±0.106	0.787±0.094
$\eta_{w_i=t_i}$	0.819±0.135	0.572±0.169	0.876±0.097	1±0	0.848±0.056	0.759±0.107
$\eta_{w_i=c_i}$	0.838±0.087	0.45±0.109	0.777±0.106	0.848±0.056	1±0	0.6±0.064
NMI	0.736±0.096	0.497±0.2	0.787±0.094	0.759±0.107	0.6±0.064	1±0

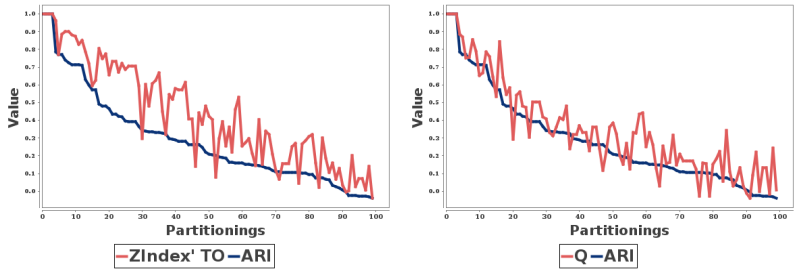
Table 2 Correlation between adapted external indexes on karate and strike datasets, computed based on Spearman's Correlation. Here, $\eta_{w_i=d_i}$, $\eta_{w_i=t_i}$, and $\eta_{w_i=c_i}$ denote the weighted *ARI* where each node is weighted respectively by, its degree, the number of triangles it belongs to, or its clustering coefficient. The ξ, on the other hand, stands for the structural agreement based on number of edges (see Section 3.2.3 for more details).

Table 3 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 100 different partitionings with average 3.82±1.51 clusters ranging from 2 to 7 and the "goodness" of the samples is on average 0.29±0.26 in terms of their *ARI* agreement.

Dataset	K^*	#	$\overline{K}$	$\overline{ARI}$
strike	3	100	3.2±1.08∈[2,7]	0.45±0.27∈[0.01,1]
polboks	3	100	4.36±1.73∈[2,9]	0.43±0.2∈[0.03,1]
karate	2	100	3.82±1.51∈[2,7]	0.29±0.26∈[-0.04,1]
football	11	100	12.04±4.8∈[4,25]	0.55±0.22∈[0.16,1]

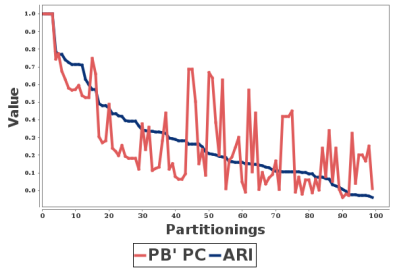
d, we have a set of m different possible partitionings: $P(d) = \{p_1, p_2, \ldots, p_m\}$. Then, the performance of criterion c on dataset d could be determined by how much its values, $I_c(d) = \{c(p_1), c(p_2), \ldots, c(p_m)\}$, correlate with the "goodness" of these partitionings. Assuming that the true partitioning (i.e. ground truth) p^* is known for dataset d, the "goodness" of partitioning p_i could be determined using partitioning agreement measure a. Hence, for dataset d with set of possible partitionings $P(d)$, the external evaluation provides $E(d) = \{a(p_1, p^*), a(p_,p^*), \ldots, a(p_m, p^*)\}$, where (p_1, p^*) denotes the "goodness" of partitioning p_1 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. The following procedure summarizes our comparison approach.

$D \leftarrow \{d_1, d_2, \ldots, d_n\}$
for all dataset $d \in D$ **do**
  $P(d) \leftarrow \{p_1, p_2, \ldots, p_m\}$ # generate m possible partitionings
  $E(d) \leftarrow \{a(p_1, p^*), a(p_2, p^*), \ldots, a(p_m, p^*)\}$ # compute the external scores
  for all $c \in Criteria$ **do**
    $I_c(d) \leftarrow \{c(p_1), c(p_2), \ldots, c(p_m)\}$ # compute the internal scores
    $score_c(d) \leftarrow correlation(E, I)$ # compute the correlation
  end for
end for
$score_c \leftarrow \frac{1}{n} \sum_{d=1}^{n} score_c(d)$ # rank criteria based on their average scores

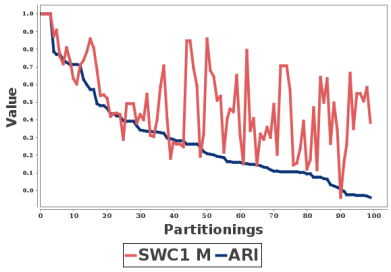


(a) ZIndex with Topological Overlap

(b) Q modularity



(c) Point-Biserial with Pearson Correlation



(d) Silhouette with Modularity Proximity

Fig. 5 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 ARI. For the sake of illustration therefore, each criterion's values are scaled to be in the same range as of the external index.

4.3.1 Results on Real World Datasets

Table 3 shows general statistics of our real world datasets and their generated samples. We can see that the randomized samples cover the space of partitionings according to their external index range.

Figure 5 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 3. Similar analysis is done for all 4 datasets $\times$ 645 criteria (combination of relative indexes and distances variations) $\times$ 5 external indexes, which produced over 12900 such correlations. The top ranked criteria based on their average performance over these datasets are summarized in Table 4. Based on these results, *ZIndex* when used with almost all of the proximity measures, such as Topological Overlap (TO), Pearson Correlation Similarity (PC) or Intersection Closeness (IC); has a higher correlation with the external index comparing to the modularity Q.

And this is true regardless of the choice of *ARI* as the external index, since it is ranked above Q by other external indexes, e.g., *NMI* and *NMI*. Other criteria, on the other hand, are all ranked after the modularity Q, except the CIndex *SP*. One may conclude based on this experiment that *ZIndex* is a more accurate evaluation criterion comparing to Q. We can also examine the ranking of different proximity measures in this table. For example, we can see that the Number of Paths of length 2, $NP2$, performs better than length 3, $NP3$; and that the exponential combination of NPE performs better than linear, NPL, and uniform, NP, alternatives.

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 5. For example, *SWC1* (Silhouette with Criterion where distance of a node to a community is computed by its distance to the centroid of that community) with the Modularity M proximity 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. We can also see the similar pattern in the Point-Biserial with PC proximity. 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 5, 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 near or medium far from the true partitioning, *PB' PC* is between top criteria, while its performance drops significantly for samples very far from the ground truth.

4.3.2 Synthetic Benchmarks Datasets

Similar to the last experiment, Table 7 reports the ranking of the top criteria according to their average performance on synthesized datasets of Table 6. Based on which, *ZIndex* overall outperforms other criteria including the modularity Q, this is more significant in ranking finner partitionings, near optimal; while it is less significant in ranking poor partitionings.

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 becomes the overall superior criterion for synthetic benchmarks with higher level of mixed communities ($.3 \leq \mu \leq .5$). Table 8 reports the overall ranking of the criteria for a difficult set of datasets that have high mixing parameter. We can see that although Q is the overall superior criterion, *ZIndex* still significantly outperforms Q in ranking finer partitionings. Results for other settings are available in the supplementary materials.

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

Table 4 Overall ranking of criteria on the real world datasets, based on the average Spearman's correlation of criteria with the ARI external index, ARI_{corr}. Ranking based on correlation with other external indexes is also reported. The full ranking of the 654 criteria, which is not reported here due to space limit, can be accessed in the supplementary materials.

Rank	Criterion	ARI_{corr}	Rand	Jaccard	NMI	AMI
1	ZIndex' TO	0.925±0.018	9	148	9	7
2	ZIndex' $\hat{PC}$	0.923±0.012	2	197	2	2
3	ZIndex' $N\hat{P}C$	0.923±0.012	3	198	1	1
4	ZIndex' IC2	0.922±0.024	8	182	5	3
5	ZIndex' $\hat{TO}$	0.922±0.016	10	153	8	8
6	ZIndex' $N\hat{P}O$	0.921±0.014	6	204	3	4
7	ZIndex' ICV2	0.919±0.04	18	163	12	10
8	ZIndex' PC	0.918±0.018	4	207	10	11
9	ZIndex' IC3	0.918±0.039	19	165	15	12
10	ZIndex' $N\hat{O}V$	0.915±0.014	11	213	6	9
11	ZIndex' IC1	0.912±0.02	5	235	13	20
12	ZIndex' NPE2.0	0.911±0.03	26	168	21	15
13	ZIndex' NOV	0.91±0.023	12	225	18	21
14	ZIndex' ICV1	0.91±0.023	13	226	19	22
15	ZIndex' $NP\hat{E}$2.0	0.91±0.025	23	184	22	19
16	ZIndex' NPL2.0	0.909±0.02	24	202	14	13
17	ZIndex' M	0.908±0.028	25	149	26	23
18	ZIndex' ICV3	0.908±0.057	29	176	28	25
19	ZIndex' NP2.0	0.907±0.021	20	212	16	14
20	ZIndex' $NP\hat{L}$2.0	0.906±0.022	21	216	17	17
21	ZIndex' $N\hat{P}$2.0	0.906±0.022	22	217	20	18
22	ZIndex' $\hat{NO}$	0.905±0.022	16	253	11	16
23	ZIndex' NO	0.904±0.034	7	250	23	31
24	ZIndex' $\hat{MM}$	0.903±0.037	17	233	24	30
25	CIndex SP	0.9±0.02	1	251	31	42
26	ZIndex' $NP\hat{L}$3.0	0.899±0.032	30	200	27	24
27	ZIndex' $N\hat{P}$3.0	0.899±0.033	33	196	29	27
28	ZIndex' $NP\hat{E}$3.0	0.899±0.048	31	205	35	33
29	ZIndex $\hat{AR}$	0.898±0.035	14	264	30	36
30	ZIndex' NPE3.0	0.897±0.052	35	187	39	34
31	ZIndex' NPL3.0	0.897±0.038	36	170	32	28
32	ZIndex SP	0.895±0.036	28	215	40	41
33	ZIndex' NP3.0	0.895±0.039	37	166	34	29
34	ZIndex AR	0.895±0.039	15	255	36	38
35	ZIndex' A	0.894±0.045	32	158	38	35
36	ZIndex' MD	0.894±0.048	34	179	33	32
37	ZIndex' $\hat{A}$	0.891±0.05	27	241	37	37
38	Q	0.878±0.034	45	110	45	44
39	CIndex' NPE3.0	0.876±0.054	43	9	4	6
40	CIndex' ICV3	0.869±0.069	44	4	7	5
41	CIndex AR	0.864±0.031	40	268	42	40
42	CIndex $\hat{AR}$	0.861±0.032	42	266	41	39
43	CIndex' $NP\hat{E}$3.0	0.858±0.07	47	8	25	26
44	ZIndex' $\hat{MD}$	0.856±0.101	38	323	43	45
45	SWC0 IC1	0.847±0.09	41	108	46	47
46	SWC0 IC2	0.838±0.092	49	11	50	49
47	SWC0 NO	0.837±0.106	39	146	48	50
48	SWC0 IC3	0.819±0.104	57	7	58	52
49	SWC0 NOV	0.814±0.094	52	26	54	56
50	SWC0 ICV1	0.814±0.094	53	27	55	57

Table 5 Difficulty analysis of the results: considering ranking for partitionings near optimal ground truth, medium far and very far. Reported result are based on ARI and the Spearman's correlation.

Near Optimal Samples						
Rank	Criterion	ARI_{corr}	Rand	Jaccard	NMI	AMI
1	ZIndex' $N\hat{P}C$	0.851±0.081	1	3	4	5
2	ZIndex' $\hat{PC}$	0.851±0.081	2	4	3	3
3	ZIndex SP	0.847±0.084	18	2	8	8
4	ZIndex' $N\hat{P}O$	0.845±0.088	3	9	6	6
5	DB ICV2	0.845±0.065	30	1	31	30
6	ZIndex' $NP\hat{E}$3.0	0.842±0.082	10	5	2	2
7	ZIndex' ICV3	0.839±0.084	4	20	20	21
8	ZIndex' $N\hat{O}V$	0.835±0.093	11	14	15	15
9	ZIndex' $T\hat{O}$	0.835±0.09	9	10	7	7
10	ZIndex' $NP\hat{E}$2.0	0.834±0.089	13	8	1	1
11	ZIndex' TO	0.834±0.089	7	16	11	11
12	ZIndex' IC2	0.834±0.095	5	23	18	18
		⋮				
36	ZIndex' M	0.763±0.139	33	29	30	31
37	Q	0.762±0.166	39	21	41	41
38	DB ICV3	0.757±0.126	37	35	38	36
39	DB IC3	0.753±0.176	35	36	39	39
40	PB' PC	0.753±0.289	45	26	71	71
Medium Far Samples						
Rank	Criterion	ARI_{corr}	Rand	Jaccard	NMI	AMI
1	ZIndex' TO	0.775±0.087	5	361	22	20
2	ZIndex' $T\hat{O}$	0.771±0.091	6	386	19	17
3	ZIndex' IC3	0.768±0.134	2	372	16	13
4	ZIndex' ICV2	0.766±0.124	3	370	2	2
5	ZIndex' NPL3.0	0.762±0.079	12	349	28	27
6	ZIndex' ICV3	0.757±0.12	4	376	21	19
7	ZIndex' NP3.0	0.756±0.085	15	354	29	28
8	ZIndex' $\hat{PC}$	0.755±0.122	9	417	4	4
9	ZIndex' $N\hat{P}C$	0.755±0.122	11	418	3	3
10	ZIndex' NPE2.0	0.753±0.107	10	373	14	14
11	ZIndex' NPE3.0	0.746±0.093	8	369	24	24
12	ZIndex' $N\hat{P}O$	0.744±0.123	14	437	5	5
		⋮				
29	ZIndex' $M\hat{M}$	0.694±0.168	31	458	40	32
30	Q	0.69±0.151	58	70	79	72
		⋮				
31	ZIndex' A	0.69±0.144	34	366	58	54
46	PB' PC	0.623±0.06	112	28	200	157
Far Far Samples						
Rank	Criterion	ARI_{corr}	Rand	Jaccard	NMI	AMI
1	ZIndex' ICV2	0.724±0.066	36	520	4	9
2	ZIndex' IC3	0.72±0.062	40	523	11	19
3	ZIndex' ICV3	0.717±0.059	47	511	23	25
4	ZIndex' IC2	0.715±0.072	35	540	3	6
5	ZIndex' TO	0.706±0.064	49	519	16	14
6	ZIndex' $N\hat{P}O$	0.704±0.076	44	547	1	3
7	ZIndex' $T\hat{O}$	0.704±0.062	51	522	13	5
8	ZIndex' NPE2.0	0.701±0.057	55	505	15	7
9	ZIndex' $N\hat{P}C$	0.698±0.083	45	552	6	10
10	ZIndex' $\hat{PC}$	0.697±0.083	46	553	9	11
11	ZIndex' $NP\hat{E}$2.0	0.688±0.047	57	521	24	23
12	ZIndex' NPL2.0	0.688±0.072	58	529	12	4
		⋮				
30	ZIndex' IC1	0.655±0.132	43	566	34	40
31	ZIndex' $\hat{NO}$	0.651±0.106	52	567	22	26
32	Q	0.643±0.033	86	444	50	45
33	ZIndex' NO	0.638±0.158	38	572	38	47
34	ZIndex' MD	0.63±0.099	78	513	43	41
		⋮				
117	PB' PC	0.372±0.126	197	170	159	129

Table 6 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^*	#	$\overline{K}$	$\overline{ARI}$
network1	4	100	5.26±2.45∈[2,12]	0.45±0.18∈[0.13,1]
network2	3	100	4±1.7∈[2,8]	0.47±0.23∈[0.06,1]
network3	2	100	4±1.33∈[2,6]	0.36±0.22∈[0.07,1]
network4	7	100	10.68±3.3∈[4,19]	0.69±0.21∈[0.25,1]
network5	2	100	4.68±1.91∈[2,9]	0.32±0.22∈[-0.01,1]
network6	5	100	5.98±2.63∈[2,14]	0.52±0.21∈[0.12,1]
network7	4	100	6.62±2.72∈[2,12]	0.52±0.22∈[0.11,1]
network8	5	100	5.8±2.45∈[2,12]	0.55±0.22∈[0.15,1]
network9	5	100	6.54±2.08∈[3,11]	0.64±0.2∈[0.25,1]
network10	6	100	8.88±2.74∈[4,15]	0.59±0.19∈[0.21,1]

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 algorithm is producing high quality results close to the optimal, modularity Q might not distinguish the good and bad partitionings very well. While if we are choosing between mixed and not well separated clusterings, it is the superior criterion. Please note that these results and criteria are different from our earlier work (Rabbany et al. 2012), particularly, ZIndex is defined differently in this paper.

5 Summary and Future Perspectives

In this section, we summarize our paper and elaborate on the findings of our results and suggest some line of works that could be followed.

In this article, we examined different approaches for evaluating community mining results. Particularly we examined different external and relative measures for clustering validity and adapted these for community mining evaluation. Our main contribution is the generalization of 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 interrelated data, where the only commonly used criterion to evaluate the goodness of detected communities is currently the modularity Q. Providing a more extensive set of 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. Unlike most of the original clustering validity criteria that are defined specifically based on the Euclidean distance, our generalized formulation is independent of any particular distance measure, .

In our experiments, several of these adapted criteria exhibit high performances on ranking different partitionings of a given dataset, which makes them

Table 7 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	ARI_{corr}	Rand	Jaccard	NMI	AMI
1	ZIndex' ICV2	0.96±0.029	5	32	3	3
2	ZIndex' IC3	0.958±0.028	4	42	2	2
3	ZIndex' IC2	0.958±0.033	1	58	1	1
4	ZIndex' $\hat{PC}$	0.953±0.04	3	78	6	6
5	ZIndex' $N\hat{P}C$	0.953±0.04	2	79	7	7
6	ZIndex' ICV3	0.953±0.027	8	44	4	5
7	ZIndex' $N\hat{P}O$	0.951±0.041	6	83	9	9
8	ZIndex' $\hat{TO}$	0.949±0.045	13	60	17	17
9	ZIndex' $N\hat{O}V$	0.949±0.042	7	90	8	8
10	ZIndex' TO	0.948±0.046	16	50	21	21
11	ZIndex' PC	0.947±0.043	10	77	16	15
12	ZIndex' $NP\hat{E}$2.0	0.947±0.042	11	68	13	13
13	ZIndex' NPE2.0	0.946±0.043	17	51	20	20
14	ZIndex' NOV	0.941±0.047	14	95	18	18
15	ZIndex' ICV1	0.941±0.047	15	96	19	19
		⋮				
29	ZIndex' NPL3.0	0.895±0.072	31	121	38	37
30	Q	0.893±0.046	33	33	26	22
31	ZIndex' NP3.0	0.89±0.076	32	130	39	39
Near Optimal Results						
Rank	Criterion	ARI_{corr}	Rand	Jaccard	NMI	AMI
1	ZIndex' IC2	0.826±0.227	2	10	4	6
2	CIndex' ICV2	0.822±0.132	7	1	11	7
3	ZIndex' IC3	0.821±0.232	1	16	5	9
4	CIndex' ICV3	0.818±0.237	4	9	3	5
5	ZIndex' ICV2	0.816±0.232	3	18	7	10
6	ZIndex' $\hat{A}$	0.813±0.225	5	19	2	2
7	CIndex' IC3	0.8±0.2	31	2	13	8
8	ZIndex' A	0.795±0.177	30	20	6	4
9	ZIndex' $\hat{MM}$	0.794±0.221	9	33	1	1
		⋮				
206	SWC1' $\hat{NO}$	0.591±0.179	225	194	244	233
207	Q	0.589±0.161	222	198	138	110
Medium Far Results						
Rank	Criterion	ARI_{corr}	Rand	Jaccard	NMI	AMI
1	ZIndex' ICV2	0.741±0.177	4	231	22	22
2	ZIndex' IC2	0.738±0.181	1	247	16	20
3	ZIndex' IC3	0.728±0.188	5	252	18	21
4	ZIndex' ICV3	0.721±0.177	8	258	21	23
5	ZIndex' $\hat{PC}$	0.719±0.204	3	285	30	35
6	ZIndex' $N\hat{P}C$	0.719±0.204	2	286	31	36
7	CIndex' ICV3	0.713±0.151	28	21	33	27
8	ZIndex' $N\hat{P}O$	0.709±0.205	7	278	32	38
9	ZIndex' $\hat{TO}$	0.703±0.216	12	240	42	48
10	ZIndex' TO	0.702±0.217	14	239	45	53
		⋮				
37	Q	0.62±0.139	42	167	56	47
Far Far Results						
Rank	Criterion	ARI_{corr}	Rand	Jaccard	NMI	AMI
1	ZIndex' ICV2	0.834±0.062	9	464	5	3
2	ZIndex' IC3	0.832±0.06	7	469	4	2
3	ZIndex' TO	0.825±0.098	22	423	29	27
4	ZIndex' ICV3	0.823±0.063	12	458	6	6
5	ZIndex' $\hat{TO}$	0.823±0.096	18	446	27	25
		⋮				
30	ZIndex' M	0.638±0.151	31	537	9	4
31	Q	0.581±0.155	95	368	69	32
32	ZIndex SP	0.58±0.158	72	539	25	29

Table 8 Overall ranking of criteria based on AMI & Spearman's Correlation on the synthetic benchmarks with the same parameters as in Table 6 but much higher mixing parameter, .4. We can see that in these settings, modularity Q overall outperforms the ZIndex while the latter is significantly better in differentiating finer results near optimal.

Overall Results						
Rank	Criterion	ARI_{corr}	Rand	Jaccard	NMI	AMI
1	Q	0.854±0.039	11	1	4	2
2	ZIndex' M	0.839±0.067	2	5	1	1
3	ZIndex' A	0.813±0.071	4	11	3	3
4	ZIndex' $\hat{MM}$	0.785±0.115	1	63	2	4
5	ZIndex' $\hat{A}$	0.767±0.101	3	86	5	5
6	ZIndex' $\hat{PC}$	0.748±0.19	5	108	7	7
7	ZIndex' $N\hat{P}C$	0.748±0.19	6	109	8	8
8	ZIndex' $N\hat{P}O$	0.745±0.191	7	110	9	9
9	ZIndex' $\hat{TO}$	0.738±0.197	13	88	16	15
10	ZIndex' $N\hat{O}V$	0.738±0.197	8	134	10	10
Near Optimal Results						
Rank	Criterion	ARI_{corr}	Rand	Jaccard	NMI	AMI
1	ZIndex' M	0.825±0.105	1	1	1	1
2	ZIndex' A	0.8±0.184	2	2	2	2
3	ZIndex' $\hat{MM}$	0.768±0.166	3	4	3	3
4	ZIndex' $\hat{A}$	0.76±0.192	4	6	4	4
5	Q	0.72±0.209	34	3	34	34
6	ASWC0 $NP\hat{L}2.0$	0.719±0.248	22	8	5	5
7	SWC0 $NP\hat{L}2.0$	0.718±0.247	23	9	6	6
8	ZIndex' $NP\hat{E}2.0$	0.714±0.259	5	21	7	8
9	ASWC0 SP	0.71±0.286	28	5	29	26
10	ZIndex' $NP\hat{L}2.0$	0.702±0.261	6	29	13	18
Medium Far Results						
Rank	Criterion	ARI_{corr}	Rand	Jaccard	NMI	AMI
1	Q	0.578±0.124	106	22	3	1
2	CIndex' $N\hat{P}C$	0.522±0.146	154	12	78	69
3	CIndex' $\hat{PC}$	0.521±0.146	155	13	79	70
4	CIndex' $N\hat{P}O$	0.519±0.142	176	5	120	100
5	CIndex' $N\hat{O}V$	0.501±0.14	209	4	142	135
6	ZIndex' M	0.498±0.199	4	364	2	2
7	CIndex' IC2	0.492±0.146	227	9	176	173
8	CIndex' ICV2	0.483±0.193	149	79	119	115
9	CIndex' IC3	0.478±0.191	187	43	148	146
10	CIndex' TO	0.478±0.175	179	31	204	203
Far Far Results						
Rank	Criterion	ARI_{corr}	Rand	Jaccard	NMI	AMI
1	ZIndex' $\hat{PC}$	0.527±0.169	61	501	5	4
2	ZIndex' $N\hat{P}C$	0.527±0.169	62	502	6	5
3	Q	0.523±0.192	128	73	93	25
4	ZIndex' M	0.522±0.121	77	465	8	2
5	ZIndex' $N\hat{P}O$	0.518±0.168	63	504	10	6
6	ZIndex' $N\hat{O}V$	0.515±0.166	60	518	11	7
7	ZIndex' $\hat{TO}$	0.489±0.171	78	485	15	9
8	ZIndex' $NP\hat{E}2.0$	0.481±0.168	79	491	24	14
9	ZIndex' $\hat{MM}$	0.48±0.15	30	553	2	3
10	ZIndex' $\hat{NO}$	0.48±0.17	43	552	7	8

useful alternatives for the Q modularity. Particularly the $ZIndex$ criterion exhibits good performance almost regardless of the choice of the proximity measure. This makes $ZIndex$ also an attractive objective for finding communities. We intend to further investigate this direction in the future work.

Our results suggests that the performances of different criteria and their rankings changes in different settings. Here we examined the effects of how well-separated are the communities in the ground truth and also the general distance of a clustering from the ground truth. We further observed that the quality of different criteria is also affected by the choice of benchmarks: Synthetic v.s. Real benchmarks. This difference motivates further investigation in order to produce more realistic synthetic generators (Aldecoa and Marin (2012)). Another direction is to *classify the criteria* according to their performance based on different network characteristics; Onnela et al. (2010), Sallaberry et al. (2013) provide examples of network characterisation.

We also compared common clustering similarity/agreement measures used in *external* evaluation of community mining results. Our results confirms that the commonly used agreement measure NMI is biased in favour of large number of communities and falls short of detecting the true number of communities compared to other measures. In contrast ARI possess both of these desirable properties. We further proposed the need for modified measures specific to communities agreement, pointing out that the current clustering agreement measures completely ignore the edges. We have proposed few straightforward extensions for the agreement measures to adapt them for the context of interrelated data, including a degree weighted variation of ARI. The resulting agreement measures, are more appropriate for external evaluation of community mining results while exhibiting the desirable qualities of ARI (i.e. adjustment for chance and detecting true number of clusters). Our results also motivate further investigation into the properties of these extensions and also examining alternative extensions (for example incorporating the notion of assortativity); this is mainly because despite being unbiased, these extensions are not as stable as ARI in evaluating random clusterings. Another line of work following the agreement similarity measures is investigating their application in consensus or ensemble clustering, for example see (Lancichinetti and Fortunato 2012, Strehl and Ghosh 2003).

As a part of future work we intend to provide extensions of the criteria and measures defined here for more general cases of community mining: *overlapping communities*, *dynamic communities* and also *local communities*. For example in the literature on cluster analysis, there are clustering algorithms and validation indexes specially designed to deal with data involving overlapping categories. In particular, fuzzy clustering algorithms produce clustering results in which data objects may belong to multiple clusters at different degrees (Bezdek 1981, Dumitrescu and Jain. 2000, Hppner et al. 1999). In order to evaluate the results of such algorithms, a number of relative, internal, and external fuzzy clustering validation indexes have been proposed (Campello 2010, Campello and Hruschka 2006, Collins and Dent 1988, Dumitrescu and Jain. 2000, Halkidi et al. 2001, Hppner et al. 1999). Furthermore some recent

works study methods of finding and evaluating overlapping communities in the context of interrelated data (Gregory 2011, Lancichinetti et al. 2009, Rees and Gallagher 2012, Yoshida 2013).

Acknowledgment

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

Albatineh, A. N., Niewiadomska-Bugaj, M., and Mihalko, D. (2006). On similarity indices and correction for chance agreement. *Journal of Classification*, 23:301–313. 10.1007/s00357-006-0017-z.

Aldecoa, R. and Marin, I. (2012). Closed benchmarks for network community structure characterization. *Phys. Rev. E*, 85:026109.

Bezdek, J. C. (1981). *Pattern Recognition with Fuzzy Objective Function Algorithms*. Kluwer Academic Publishers, Norwell, MA, USA.

Calinski, T. and Harabasz, J. (1974). A dendrite method for cluster analysis. *Communications in Statistics-theory and Methods*, 3:1–27.

Campello, R. (2010). Generalized external indexes for comparing data partitions with overlapping categories. *Pattern Recognition Letters*, 31(9):966–975.

Campello, R. and Hruschka, E. R. (2006). A fuzzy extension of the silhouette width criterion for cluster analysis. *Fuzzy Sets and Systems*, 157(21):2858–2875.

Chen, J., Zaïane, O. R., and Goebel, R. (2009). Detecting communities in social networks using max-min modularity. In *SIAM International Conference on Data Mining*, pages 978–989.

Clauset, A. (2005). Finding local community structure in networks. *Physical Review E (Statistical, Nonlinear, and Soft Matter Physics)*, 72(2):026132.

Collins, L. M. and Dent, C. W. (1988). Omega: A general formulation of the rand index of cluster recovery suitable for non-disjoint solutions. *Multivariate Behavioral Research*, 23(2):231–242.

Dalrymple-Alford, E. C. (1970). Measurement of clustering in free recall. *Psychological Bulletin*, 74:32–34.

Danon, L., Guilera, A. D., Duch, J., and Arenas, A. (2005). Comparing community structure identification. *Journal of Statistical Mechanics: Theory and Experiment*, (09):09008.

Davies, D. L. and Bouldin, D. W. (1979). A cluster separation measure. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, PAMI-1(2):224 –227.

Dumitrescu, D., B. L. and Jain., L. C. (2000). Fuzzy sets and their application to clustering and training. *CRC Press, Boca Raton*.

Dunn, J. C. (1974). Well-separated clusters and optimal fuzzy partitions. *Journal of Cybernetics*, 4(1):95–104.

Fortunato, S. (2010). Community detection in graphs. *Physics Reports*, 486(35):75–174.

Fortunato, S. and Barthélemy, M. (2007). Resolution limit in community detection. *Proceedings of the National Academy of Sciences*, 104(1):36–41.

Girvan, M. and Newman, M. E. J. (2002). Community structure in social and biological networks. *Proceedings of the National Academy of Sciences*, 99(12):7821–7826.

Gregory, S. (2011). Fuzzy overlapping communities in networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2:17.

Gustafsson, M., Hörnquist, M., and Lombardi, A. (2006). Comparison and validation of community structures in complex networks. *Physica A Statistical Mechanics and its Applications*, 367:559–576.

Halkidi, M., Batistakis, Y., and Vazirgiannis, M. (2001). On clustering validation techniques. *Journal of Intelligent Information Systems*, 17:107–145.

Hppner, F., Klawonn, F., Kruse, R., and Runkler, T. (1999). *Fuzzy cluster analysis: methods for classification, data analysis and image recognition*. J. Wiley.

Hubert, L. and Arabie, P. (1985). Comparing partitions. *Journal of Classification*, 2:193–218.

Hubert, L. J. and Levin, J. R. (1976). A general statistical framework for assessing categorical clustering in free recall. *Psychological Bulletin*, 83:1072–1080.

Kenley, E. C. and Cho, Y.-R. (2011). Entropy-based graph clustering: Application to biological and social networks. In *IEEE International Conference on Data Mining*.

Krebs, V. (2004). Books about us politics. http://www.orgnet.com/.

Lancichinetti, A. and Fortunato, S. (2009). Community detection algorithms: A comparative analysis. *Physical Review E*, 80(5):056117.

Lancichinetti, A. and Fortunato, S. (2012). Consensus clustering in complex networks. *Nature Scientific Reports*, 2.

Lancichinetti, A., Fortunato, S., and Kertsz, J. (2009). Detecting the overlapping and hierarchical community structure in complex networks. *New Journal of Physics*, 11(3):033015.

Lancichinetti, A., Fortunato, S., and Radicchi, F. (2008). Benchmark graphs for testing community detection algorithms. *Physical Review E*, 78(4):046110.

Leskovec, J., Kleinberg, J., and Faloutsos, C. (2005). Graphs over time: densification laws, shrinking diameters and possible explanations. In *ACM SIGKDD international conference on Knowledge discovery in data mining*, pages 177–187.

Leskovec, J., Lang, K. J., and Mahoney, M. (2010). Empirical comparison of algorithms for network community detection. In *International conference on World wide web*, pages 631–640.

Luo, F., Wang, J. Z., and Promislow, E. (2008). Exploring local community structures in large networks. *Web Intelligence and Agent Systems*

Manning, C. D., Raghavan, P., and Schtze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press, New York, NY, USA.

Meil, M. (2007). Comparing clusteringsan information based distance. *Journal of Multivariate Analysis*, 98(5):873 – 895.

Milligan, G. and Cooper, M. (1985). An examination of procedures for determining the number of clusters in a data set. *Psychometrika*, 50(2):159–179.

Newman, M. (2010). *Networks: An Introduction*. Oxford University Press, Inc., New York, NY, USA.

Newman, M. E. J. (2006). Modularity and community structure in networks. *Proceedings of the National Academy of Sciences*, 103(23):8577–8582.

Newman, M. E. J. and Girvan, M. (2004). Finding and evaluating community structure in networks. *Physical Review E*, 69(2):026113.

Nooy, W. d., Mrvar, A., and Batagelj, V. (2004). *Exploratory Social Network Analysis with Pajek*. Cambridge University Press.

Onnela, J.-P., Fenn, D. J., Reid, S., Porter, M. A., Mucha, P. J., Fricker, M. D., and Jones, N. S. (2010). Taxonomies of Networks. *ArXiv e-prints*.

Orman, G. K. and Labatut, V. (2010). The effect of network realism on community detection algorithms. In *Proceedings of the 2010 International Conference on Advances in Social Networks Analysis and Mining*, ASONAM '10, pages 301–305.

Orman, G. K., Labatut, V., and Cherifi, H. (2011). Qualitative comparison of community detection algorithms. In *International Conference on Digital Information and Communication Technology and Its Applications*, volume 167, pages 265–279.

Pakhira, M. and 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*

Palla, G., Derenyi, I., Farkas, I., and Vicsek, T. (2005). Uncovering the overlapping community structure of complex networks in nature and society. *Nature*, 435(7043):814–818.

Porter, M. A., Onnela, J.-P., and Mucha, P. J. (2009). Communities in Networks. *ArXiv e-prints*.

Rabbany, R., Chen, J., and Zaïane, O. R. (2010). Top leaders community detection approach in information networks. In *SNA-KDD Workshop on Social Network Mining and Analysis*.

Rabbany, R., Takaffoli, M., Fagnan, J., Zaiane, O., and Campello, R. (2012). Relative validity criteria for community mining algorithms. In *Advances in Social Networks Analysis and Mining (ASONAM), 2012 International Conference on*.

Rabbany, R. and Zaïane, O. R. (2011). A diffusion of innovation-based closeness measure for network associations. In *IEEE International Conference on Data Mining Workshops*, pages 381 –388.

Ravasz, E., Somera, A. L., Mongru, D. A., Oltvai, Z. N., and Barabsi, A.-L. (2002). Hierarchical organization of modularity in metabolic networks.

Science, 297(5586):1551–1555.

Rees, B. S. and Gallagher, K. B. (2012). Overlapping community detection using a community optimized graph swarm. *Social Network Analysis and Mining*, 2(4):405–417.

Rosvall, M. and Bergstrom, C. T. (2007). An information-theoretic framework for resolving community structure in complex networks. *Proceedings of the National Academy of Sciences*, 104(18):7327–7331.

Rosvall, M. and Bergstrom, C. T. (2008). Maps of random walks on complex networks reveal community structure. *Proceedings of the National Academy of Sciences*, 105(4):1118–1123.

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.

Sallaberry, A., Zaidi, F., and Melançon, G. (2013). Model for generating artificial social networks having community structures with small-world and scale-free properties. *Social Network Analysis and Mining*, pages 1–13.

Strehl, A. and Ghosh, J. (2003). Cluster ensembles — a knowledge reuse framework for combining multiple partitions. *J. Mach. Learn. Res.*, 3:583–617.

Theodoridis, S. and Koutroumbas, K. (2009). Cluster validity. In *Pattern Recognition*, chapter 16. Elsevier Science, 4 edition.

Vendramin, L., Campello, R. J. G. B., and Hruschka, E. R. (2010). Relative clustering validity criteria: A comparative overview. *Statistical Analysis and Data Mining*, 3(4):209–235.

Vinh, N. X., Epps, J., and Bailey, J. (2009). Information theoretic measures for clusterings comparison: is a correction for chance necessary? In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, pages 1073–1080, New York, NY, USA. ACM.

Vinh, N. X., Epps, J., and Bailey, J. (2010). Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. *Journal of Machine Learning Research*, 11:2837–2854.

Wasserman, S. and Faust, K. (1994). *Social Network Analysis: Methods and Applications*. Cambridge University Press.

Wu, J., Xiong, H., and Chen, J. (2009). Adapting the right measures for k-means clustering. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '09, pages 877–886, New York, NY, USA. ACM.

Yoshida, T. (2013). Weighted line graphs for overlapping community discovery. *Social Network Analysis and Mining*, pages 1–13.

Zachary, W. W. (1977). An information flow model for conflict and fission in small groups. *Journal of Anthropological Research*, 33:452–473.