ADAPTATION AND ENHANCEMENT OF EVALUATION MEASURES TO OVERLAPPING GRAPH CLUSTERINGS

Tatiana Gossen, Michael Kotzyba and Andreas Nürnberger
Data and Knowledge Engineering Group, Faculty of Computer Science, Otto-von-Guericke-University Magdeburg
D-39106 Magdeburg, Germany

Keywords: Evaluation measures, Overlapping graph clustering, Clustered graph, Graph generation model.

Abstract: Quality measures are important to evaluate graph clustering algorithms by providing a means to assess the quality of a derived cluster structure. In this paper, we focus on overlapping graph structures, as many real-world networks have a structure of highly overlapping cohesive groups. We propose three methods to adapt existing crisp quality measures such that they can handle graph overlaps correctly, but also ensure that their properties for the evaluation of crisp graph clusterings are preserved when assessing a crisp cluster structure. We demonstrate our methods on such measures as *Density*, *Newman's modularity* and *Conductance*. We also propose an enhancement of an existing modularity measure for networks with overlapping structure. The newly proposed measures are analysed using experiments on artificial graphs that possess overlapping structure. For this evaluation, we apply a graph generation model that creates clustered graphs with overlaps that are similar to real-world networks i.e. their node degree and cluster size distribution follow a power law.

1 INTRODUCTION

Many information spaces from different domains, e.g. life sciences or social sciences, can be modeled in form of graphs or networks. Information concepts or entities represent the nodes of a graph and a pair of nodes has an edge if there is a relationship between corresponding entities (Palla et al., 2005). During the last years many graph based models have been created and analysed, describing e.g. social networks like acquaintance and collaboration networks, technological networks like the Internet, the Worldwide Web and power grid networks, biological networks like neural networks, food webs, and metabolic networks (Girvan and Newman, 2002).

One major task while analysing graphs is to find groups of strongly connected entities that form some kind of cluster. In other words, there exist groups of graph nodes that are more densely connected within the group than to the rest of the graph. Thus, the graph can be seen as a set of such groups also called structural sub-units, communities or clusters (Girvan and Newman, 2002). These clusters correspond to functional units of the underlying systems.

Many data mining algorithms have been proposed to find these units. However, the majority of them provide only separate or "hard" *clusterings* (partition of the graph nodes into clusters) with pairwise disjoint clusters (*crisp* clusters). Unfortunately in practice many structural sub-units are highly *overlapping* cohesive groups (Ahn et al., 2010; Lázár et al., 2010). As an example from the biological domain, in the protein complex network a large fraction of proteins belong to several protein complexes simultaneously (Gavin et al., 2002; Palla et al., 2005).

For each graph a huge amount of partitions into sub-units can be found. However, the task is to find a meaningful one. Depending on the requirements for the clustering e.g. how important dense or separated clusters are, the meaningful partitions may differ in structure. In order to evaluate the specific clustering structure *quality measures* or *indices* are required. To tackle this issue several quality measures for crisp graph clusterings have been introduced e.g. *coverage*, *performance*, *intra- and inter-cluster conductance* (Brandes et al., 2003; Brandes and Erlebach, 2005), *modularity* (Newman and Girvan, 2004), and *density* (Delling et al., 2006). Each index assesses different clustering properties and can be chosen depending on the specific requirements.

As most of the graph algorithms focus on finding a crisp structure, existing measures are optimized to evaluate the quality of such crisp clusterings. To our knowledge, there are too few measures for graph clus-

Gossen T., Kotzyba M. and Nürnberger A. (2012).
ADAPTATION AND ENHANCEMENT OF EVALUATION MEASURES TO OVERLAPPING GRAPH CLUSTERINGS.
In *Proceedings of the 1st International Conference on Pattern Recognition Applications and Methods*, pages 5-14
DOI: 10.5220/0003706400050014

terings with overlaps. So far only one index to measure networks with overlapping communities M^{ov} was proposed (Lázár et al., 2010). Thus, it is important to provide new measures for overlapping graph clusterings to be able to evaluate different cluster structure properties.

In order to do this we can use the existing measures for crisp clusterings and adapt them such that they can handle overlaps correctly. In this paper we therefore propose extensions of crisp quality measures to be able handling the graph overlaps correctly and still remain their original properties. The structure of this paper is as follows. Sect. 2 gives an overview of research that is related to this paper. Sect. 3 introduces the formal concepts we use in the remainder of the paper. We present our main ideas for the adaptation of existing crisp evaluation measures to handle the overlapping graph clustering and enhancement of overlapping measure M^{ov} in Sect. 4. In Sect. 5 and 6 we describe a model to generate clustered graphs and experiments with synthetic clustered graphs using the evaluation measures. We conclude and give directions for future work in Sect. 7.

2 RELATED WORK

We can subdivide related research work into three categories: algorithms for graph clustering, quality measures for graph clustering and generation models for clustered graphs.

Algorithms for Graph Clustering. A good overview of graph clustering algorithms is given by *Schaeffer* (Schaeffer, 2007) and *Fortunato* (Fortunato, 2010). There are many algorithms for crisp graph clusterings. One of the most prominent approaches is to repeatedly decompose the graph structure into sub-units by removing edges with the highest betweenness until the network becomes disconnected (hierarchical top-down algorithm by Girvan and Newman (Girvan and Newman, 2002)).

In order to uncover clusterings with overlaps a set of algorithms for overlapping clusterings has been introduced, e.g. the $LA-IS^2$ two step algorithm (Baumes et al., 2005), the *CONGA* algorithm (Gregory, 2007) that extends Girvan and Newman's algorithm (Girvan and Newman, 2002), the clique percolation algorithm (Palla et al., 2005) implemented in *CFinder* (Adamcsek et al., 2006), and the single-linkage agglomerative hierarchical algorithm which clusters graph links with proposed similarity measure between link groups based on their neighbourhood (Ahn et al., 2010).

Besides crisp and overlapping graph clustering there are also fuzzy clustering methods and measures which search for fuzzy structure in graphs (Nepusz et al., 2008; Nicosia et al., 2009). In this case each vertex of the graph may belong to multiple communities at the same time and its membership is determined by a numerical membership degree. However, the fuzzy approach for graph clustering is not widely used (Schaeffer, 2007).

Quality Measures. After a clustering is obtained one can apply quality measures to evaluate how well the chosen algorithm worked or to compare the results produced by different clustering algorithms. One can distinguish between *unsupervised* and *supervised* quality measures (Tan et al., 2006). Supervised measures (also called *external* indices) require external information about the expected cluster structure and compare it to the structure found by the algorithm to assess the clustering quality. An example of an external measure is the F-measure (Gregory, 2007). Unsupervised measures (also called *internal* indices) evaluate the quality of a clustering structure without considering any external information. They assess how well separated the clusters are (*inter-cluster sparsity*) and how dense the graph nodes are connected within the clusters (*intra-cluster density*). Internal indices for crisp graph clusterings are coverage, performance, intra- and inter-cluster conductance (Brandes et al., 2003), modularity (Newman and Girvan, 2004) and density (Delling et al., 2006). Lázár (Lázár et al., 2010) proposed a modularity measure for networks with overlapping communities M^{ov}. In this paper we concentrate on internal indices. Note that the internal quality indices are used not only for the evaluation of clusterings but also within the clustering algorithm as a fitness function (Schaeffer, 2007).

Generation Models for Clustered Graphs. In order to evaluate clustering algorithms and to analyse the behaviour of different quality indices, clustered graphs with different properties are required. Therefore different models to create clustered graphs have been proposed. These are models to generate different classes of graphs e.g. unweighted and weighted, undirected and directed, uniform random graphs, multigraphs and bipartite graphs with desirable cluster properties e.g. connectivity and density. An overview of generation models for graphs and graphs with clustering structure is given in (Schaeffer, 2007) and (Chakrabarti et al., 2010). A description of generation models for crisp clustered (unweighted and undirected) graphs can be found in (Girvan and Newman, 2002; Newman and Girvan, 2004). Gregory (Gregory, 2007) extends the generation model of crisp clustered graphs to produce clustered graphs with overlaps. The authors in (Lancichinetti and Radicchi, 2008) generalised the method by Girvan and Newman stressing that the distributions of node degrees and of community sizes in real networks are heterogeneous. Their model enables variation of the cluster sizes and non-trivial degree distributions.

3 PRELIMINARIES

In this paper we focus on undirected unweighted graphs. However, the approaches discussed in the next section could be also applied to directed and/or weighted graphs. Let $G = (V, E)$ be such a graph with a non-empty set of nodes V and a set of edges E. $d(v)$ or $|neigh(v)|$ is the number of nodes adjacent to the node v. A *clustering* $\zeta(G) = \{C_1, \ldots, C_k\}$ is a partition of all nodes into k clusters C_i, where $C_i \subseteq V$ is a non-empty subset of nodes, i.e. each node belongs to at least one cluster. $C(v)$ denotes the set of all clusters that contain the node v. A *cut* ξ is a partition of a vertex set V of a graph G into two non-empty subsets (C_1,C_2), i.e. $C_1 = V \setminus C_2$.

The set of all edges between clusters C_i and C_j is $E(C_i, C_j)$, where $i \neq j$. $E(C_i) = E(C_i, C_i)$ is the set of edges within the cluster C_i. They have their origin and destination in C_i. $E(\zeta) := \cup_{i=1}^{k} E(C_i)$ is the set of *intra-cluster edges* and $\overline{E(\zeta)} := E \setminus E(\zeta)$ is the set of *inter-cluster edges*. The set of edges that are incident to any node in a cluster C is denoted by $E_{inc}(C)$ (edges incident to C). We say that an edge is an *overlapping edge* if both its incident nodes are in the same overlap of two or more clusters (see Figure 1). Let m be the number of graph edges and n the number of graph nodes. The maximum possible number of edges is denoted by E_{max}:

$$E_{max}(G) = \frac{n(n-1)}{2} \qquad (1)$$

A quality measure is defined as a function $index(\zeta) \rightarrow \mathbb{R}$ that assigns a real value, usually $index(\zeta) \in [0, 1]$, to a given clustering $\zeta(G)$ (Brandes and Erlebach, 2005).

4 ADAPTATION OF EXISTING MEASURES

In this section we discuss three major ideas how to adapt the existing crisp evaluation measures to handle graph clusterings with overlaps:

1. in a *direct* way, i.e. by incorporating a component that evaluates the quality of overlapping parts.

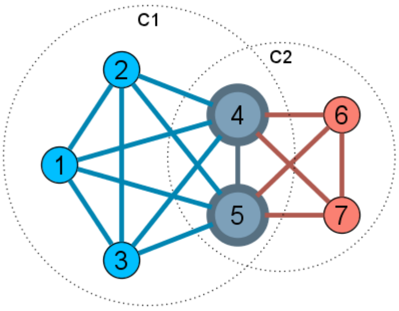

Figure 1: Graph clustering $\zeta = \{C_1, C_2\}$ has an overlapping edge $\{4,5\}$. The set of incident edges for the cluster C_1 is $E_{inc}(C1) = E(C1) \cup \{\{4,6\}, \{4,7\}, \{5,6\}, \{5,7\}\}$

2. by *incorporation* of edge *weights* so that overlapping edges have a lower weight than non-overlapping ones.
3. in an *indirect* way, i.e. depending on the extended measure and its criteria, one has to decide how the overlapping parts are handled.

We also propose an enhancement of the overlapping measure M^{ov}.

Direct Way. A quality measure usually exploits two functions to evaluate the clustering "goodness": intra-cluster density f and inter-cluster sparsity g. These functions are combined as $index(f(\zeta), g(\zeta)) \rightarrow \mathbb{R}$ e.g.:

$$index(\zeta) = f(\zeta) + g(\zeta) \qquad (2)$$

Given a clustering with overlaps we introduce a third function $h(\zeta)$ for the assessment of the overlaps' "goodness". Therefore, we measure the *complement of the overlap size ratio* (COR) to evaluate the size of overlapping parts and the *membership of the overlapping nodes* (OVM) to measure the overlaps' quality.

One intuitive assumption we make is that a good clustering should not have too many overlapping nodes. This, of course, is dependent on the application and could be changed on demand. Given OV as the set of nodes in overlaps we assess the overlap size as:

$$COR(\zeta) = 1 - \frac{|OV|}{|V|} \qquad (3)$$

The second assumption is that good-quality overlaps should only contain nodes which have a strong membership to all the clusters they belong to. The membership of the overlapping nodes is calculated as follows:

$$OVM(\zeta) = \frac{1}{|OV|} \sum_{v \in OV} \left(\frac{1}{|C(v)|} \sum_{C \in C(v)} LD(C, v) \right) \qquad (4)$$

where $LD(C, v)$ is a link density of a node v in a cluster C and the following holds:

aa

$$LD(C,v) \quad := \quad \begin{cases} \frac{|C \cap \text{neigh(v)}|}{|C|-1}, \text{ if } |C| > 1 \\ 0, \text{ otherwise} \end{cases} \tag{5}$$

For each overlapping node we calculate the number of connections to the nodes in its corresponding clusters. Both functions OVS and OVM return values in the interval $[0,1]$. We calculate the overlaps' "goodness" as $h(\zeta) \in [0,1]$:

$$h(\zeta) = \omega_s \cdot COR(\zeta) + \omega_m \cdot OVM(\zeta), \tag{6}$$

$$\text{where } \omega_s > 0, \omega_m > 0 \text{ and } \omega_s + \omega_m = 1 \tag{7}$$

ω_s and ω_m are used as weighting parameters to influence the importance of overlap size and membership. To demonstrate the idea of the direct way to adapt the existing crisp evaluation measure we employ the density quality index proposed in (Delling et al., 2006):

$$Density(\zeta) := \frac{1}{2} \underbrace{\left(\frac{1}{k} \sum_{C \in \zeta} \frac{|E(C)|}{E_{max}(C)} \right)}_{f(\zeta)} + \frac{1}{2} \underbrace{\left(1 - \frac{|\overline{E(\zeta)}|}{E_{max} - \sum_{C \in \zeta} E_{max}(C)} \right)}_{g(\zeta)} \tag{8}$$

We extend the notion of density to handle the graph clusterings with overlaps by the function $h(\zeta)$:

$$Density^{OV}(\zeta) := \omega_f \cdot f(\zeta) + \omega_g \cdot g(\zeta) + \omega_o \cdot \underbrace{\Big(\omega_s \cdot COR(\zeta) + \omega_m \cdot OVM(\zeta) \Big)}_{h(\zeta)} \tag{9}$$

where ω_f, ω_g and ω_o are positive weighting parameters and their sum is equal to 1. Thus, $Density^{OV} \in [0,1]$. If $|OV| = 0$, then $\omega_o = 0$ and $Density^{OV}(\zeta) = Density(\zeta)$, with $\omega_f = \omega_g = \frac{1}{2}$.

Incorporation of Edge Weights. Our second suggestion for adaptation of the existing crisp evaluation measures follows from the argument that the inaccuracies that occur when we apply crisp measures to clusterings with overlaps occur due to multiple counting of overlapping edges. Some measures e.g. Newman's modularity assess the quality of each cluster separately and sums the values. If applying the modularity to overlapping clustering directly, the overlapping edges would contribute to the index value several times. This results in larger values (e.g. see Figure 2). To solve this problem we redefine the edge weighting, given an edge $e = \{u, v\}$, as follows:

$$\varpi_E(e) := \begin{cases} 1, \text{if } e \in \overline{E(\zeta)}, \\ \frac{1}{|C(u) \cap C(v)|}, \text{otherwise} \end{cases} \tag{10}$$

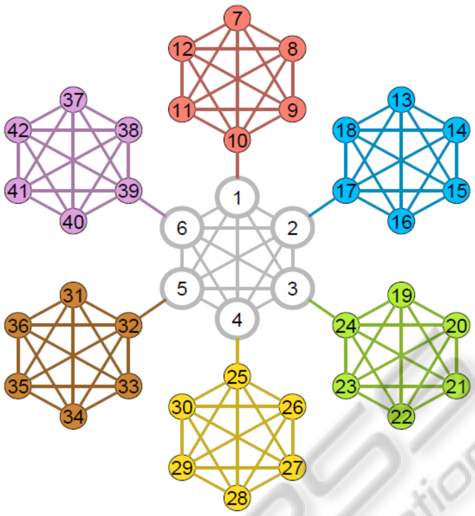

Figure 2: The graph clustering $\zeta = \{C_1 = \{1,\ldots,6,\ 7,\ldots,12\},\ C_2 = \{1,\ldots,6,\ 13,\ldots,18\},\ C_3 = \{1,\ldots,6,\ 19,\ldots,24\},\ C_4 = \{1,\ldots,6,\ 25,\ldots,30\},\ C_5 = \{1,\ldots,6,\ 31,\ldots,36\}, C_6 = \{1,\ldots,6,\ 37,\ldots,42\}\}$ has a modularity value $Q \approx 1.04455$. Newman's modularity, given a crisp clustered, undirected and unweighted graph, has an interval range of $[-\frac{1}{2}, 1]$ (Brandes et al., 2007).

Thus, intra-cluster edges that belong to only one cluster and all inter-cluster edges have a weight 1. Intra-cluster edges that belong to multiple clusters are weighted accordingly lower.

In the following we illustrate the idea of edge weight incorporation using modularity and density as examples. Newman's modularity is defined in the following way (Newman and Girvan, 2004):

$$Q(\zeta) := \sum_i (\psi_{i,i} - a_i^2) \tag{11}$$

$$\psi_{i,i} = \frac{|E(C_i)|}{m} \tag{12}$$

$$a_i = \frac{|E_{inc}(C_i)|}{m} \tag{13}$$

Using our approach we introduce the modified modularity:

$$Q^{ov}(\zeta) := \sum_i (\psi_{i,i}^{ov} - (a_i^{ov})^2) = \sum_{C \in \zeta} \left(\frac{1}{m} \sum_{e \in E(C)} \varpi_E(e) - \left(\frac{1}{m} \sum_{e \in E_{inc}(C)} \varpi_E(e) \right)^2 \right) \tag{14}$$

Given a clustering with overlaps, edges incident to a cluster C could be also intra-cluster edges of other clusters. An overlapping clustering with a relatively high intra-cluster density has a larger value of $(a^{ov})^2$ (than a crisp clustering) which results in a decrease of

Q^{ov}. Therefore, our modularity is modified to assess clusterings with overlaps but still prefers fully separated clusters.

Q^{ov} achieves its theoretical maximum if all the clusters are disconnected and have their maximum density. Q^{ov} achieves its theoretical minimum if, given a graph in form of a single edge, k clusters are distributed over each node: $k(\frac{0}{m} - (\frac{1}{m})^2) = -k$. Thus, we have $-\infty < Q^{ov} < 1$.

We can also further adjust the $Density^{OV}$ by incorporation of edge weights. The intra-cluster density $f(\zeta)$ remains unchanged as the ratio of intra-cluster edges of a cluster to the maximum possible number of edges in the cluster is independent of whether the cluster nodes are overlapping or not. We modify the inter-cluster sparsity $g(\zeta)$:

$$\widetilde{g}(\zeta) = 1 - \frac{\overline{E(\zeta)}}{E_{max} - \sum_{C\in\zeta}\left(\sum_{e\in E(K_{|C|}(C))} \varpi_E(e)\right)} \tag{15}$$

where $K_{|C|}(C)) = (C,(C \times C))$ is a complete graph that consists of the nodes in the cluster C and all possible connections between them.

$$\widetilde{Density}^{ov}(\zeta) := \omega_f \cdot f(\zeta) + \omega_g \cdot \widetilde{g}(\zeta) + \omega_o \cdot h(\zeta) \tag{16}$$

Indirect Way. There are quality measures which assess the quality of each cluster separately and then use the obtained values to calculate the "goodness" of the whole clustering. A good quality cluster should not only be dense inside, but also have a low degree of connectivity to other clusters. Thus, the quality of the *cut* between each cluster and the rest of the graph is important.

In the case of graph clustering with overlaps a question arises about how to produce a cut. There are actually three possible ways:

1. Include the overlapping nodes in the observed cluster and exclude them from the rest of the graph.
2. Include the overlapping nodes in the observed cluster and consider them also as belonging to the rest of the graph.
3. Exclude the overlapping nodes from the observed cluster and include them in the rest of the graph.

Thus, e.g. using Figure 2, observing cluster C_1 and considering the first way, the cut $\xi(C_1, V \setminus C_1)$ is $(C_1 = \{1,\ldots,6, 7,\ldots,12\}, V \setminus C_1 = \{13,\cdots,42\})$. We consider this first way to be the most intuitive one.

To demonstrate the indirect way we use the quality index *(inter-cluster) conductance* (Brandes et al., 2003; Brandes and Erlebach, 2005). The conductance of a cut compares the size of the cut and the number of edges in either of the two induced subgraphs. However, the definitions for the conductance in (Brandes et al., 2003; Brandes and Erlebach, 2005) slightly differ. In this paper, the size of the cut corresponds to the number of edges between the two components of the cut and the edges of the two induced subgraphs correspond to all edges incident to a node in the subgraphs.

The conductance of a graph clustering $\sigma(\zeta)$ is the maximum conductance value over all induced cuts $(C_i,V \setminus C_i)$. The conductance value of a cut $\xi(C,C')$ is defined as:

$$\phi(C) := \begin{cases} 1, \text{if } C \in \{\emptyset, V\} \\ 0, \text{if } C \notin \{\emptyset, V\} \text{ and } |E(C,C')| = 0 \\ \frac{|E(C,C')|}{min(|E_{inc}(C)|,|E_{inc}(C')|)} \end{cases} \tag{17}$$

where $C' = V \setminus C$. A cut can be considered as a bottleneck if its size is small relative to the density of either side of the cut.

The conductance of a graph clustering is:

$$\sigma(\zeta) = 1 - max_{C\in\zeta}\phi(C) \tag{18}$$

If applying the first intuitive way of cut definition, the formula of conductance given an overlapping clustering remains the same:

$$\sigma^{ov}_{ExFromRest}(\zeta) = \sigma(\zeta) \tag{19}$$

If applying the second way of cut definition, both the observed cluster and the rest of the graph contain the overlapping nodes. We should redefine the cut $\xi(C,C')$ as:

$$C' = V \setminus \{v \in C : |C(v)| = 1\} \tag{20}$$

We also have to modify the formula for cut conductance as in this case we actually have two bottlenecks which should be considered. To assess the bottlenecks between the two subgraphs C and C' we use the following formula:

$$\phi^{ov}_{Inc}(C) := max\Big(\frac{|E_{inc}(C) \setminus E(C)|}{|E_{inc}(C)|}, \frac{|E_{inc}(C') \setminus E(C')|}{|E_{inc}(C')|}\Big) \tag{21}$$

Then the conductance is calculated as:

$$\sigma^{ov}_{Inc}(\zeta) = 1 - max_{C\in\zeta}\phi^{ov}_{Inc}(C) \tag{22}$$

If applying the third way of cut definition, where the nodes in overlaps are excluded from the observed cluster, we should redefine the cut $\xi(C,C')$ as:

$$\xi(C_{ExFromCl}, V \setminus C_{ExFromCl}) \tag{23}$$

where

$$C_{ExFromCl}(C) = C \setminus \{v \in C : |C(v)| > 1\} \tag{24}$$

The conductance calculation given an overlapping clustering is then defined as following:

$$\sigma^{ov}_{ExFromCl}(\zeta) = 1 - max_{C\in\zeta}\phi(C_{ExFromCl}(C)) \qquad (25)$$

Enhancement of M^{ov}. The authors in (Lázár et al., 2010) define the modularity measure of networks with overlapping communities M^{ov} as follows:

$$M(C) \ := \underbrace{\frac{1}{|C|}\sum_{v\in C}\frac{in_C(v)-out_C(v)}{d(v)|C(v)|}}_{\text{node justifiability}} \cdot \underbrace{\frac{|E(C)|}{E_{max}(C)}}_{\text{cluster density}} \qquad (26)$$

where $in_C(v)$ is the number of *inward edges* of v (edges that are incident to v and are intra-cluster edges of C) and $out_C(v)$ is the number of *outward edges* (edges that are incident to v and have their destination not in C).

$$M^{ov}(\zeta) = \frac{1}{k}\sum_{C\in\zeta} M(C) \qquad (27)$$

The M^{ov} measure assesses each cluster separately and calculates the average of the ratings. The "goodness" $M(C)$ of a cluster depends on two criteria: how "justifiable" the cluster nodes are assigned to the cluster and how dense the cluster is. The first criteria means that a given node should primarily go inward towards its cluster(s) and should not go outward.

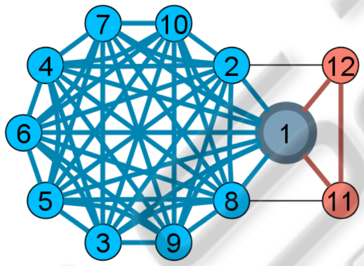

Figure 3: Graph clustering ζ: $\{C_1 = \{1,\ldots,10\}$, $C_2 = \{1,11,12\}$, has value $M^{ov}(\zeta) \approx 0.50$, while $\widetilde{M^{ov}_{mod}} \approx 0.73$

We discovered one drawback of M^{ov} measure which appears in case the given clustering contains large clusters that are well separated and dense and some small clusterings that have a small $M(C)$ value. The ratings for these "bad quality" small clusters decrease the value of M^{ov} (see Figure 3). However, it is more rational if the contribution of the cluster rating is proportional to the cluster size:

$$M^{ov}_{mod}(\zeta) := \sum_{C\in\zeta}\left(\frac{|C|}{n}M(C)\right) \qquad (28)$$

The larger a cluster is, the larger is also its influence and therefore also the influence of its quality on the whole clustering. If we do not take the cluster sizes into account, the evaluation of the entire clustering may become unbalanced or biased.

Given a clustering with overlaps we should pay attention that a node may contribute to the measure calculation several times. This leads to the increase of the measure value, although the clustering becomes harder to interpret, as the cluster borders get fuzzier. To get rid of this effect we can use a weighting for a single node (similar to the edge weighting above):

$$\varpi_V(v) = \frac{1}{|C(v)|} \qquad (29)$$

The modification of M^{ov} using the node weightings is:

$$\widetilde{M^{ov}_{mod}}(\zeta) := \sum_{C\in\zeta}\left(\left(\frac{1}{n}\sum_{v\in C}\varpi_V(v)\right)M(C)\right) \qquad (30)$$

Note, that the weighting component $\varpi_V(v)$ is already used in a node justifiability part of $M(C)$ (Formula 26). While we use it more for an appropriate calculation of a cluster's size, Lázár et al. (2010) use it to weight the contribution of each node.

The quality of the graph in Figure 3 using formula 30 is 0.73 in comparison to the original value of 0.50. The range value for $\widetilde{M^{ov}_{mod}}$ remains between -1 and 1 as for the original measure M^{ov}.

5 GENERATION MODEL FOR CLUSTERED GRAPHS

The previous section illustrated the main ideas to extend a quality measure for crisp graph clusterings to a quality measure for overlapping graph clusterings. Although the extensions take care to preserve the original measure criteria, one has to keep in mind that an extension always creates a new index. For this reason, it is important to analyse the properties and the behavior of the new indices on different overlapping graph clusterings. As there is a lack of realistic benchmark graphs with known overlapping structure, we are forced to use computer generated clustered graphs. In this way, we can analyse the behaviour of indices on clustered graphs with different properties, which is a major advantage.

To generate an artificial overlapping graph clustering, we use the idea from (Gregory, 2007) and modify some parts to create more realistic graphs and clusterings. At first Gregory generates n nodes and divides them into k clusters. He uses a parameter r to specify the fraction of overlaps, so that each cluster contains nr/k nodes. Afterwards the edges are randomly placed between pairs of nodes with the probability p_{in} if the nodes belong to the same cluster and p_{out} otherwise. In our generation model the clusters are not equally-sized and the nodes do not possess the same degree in average to make them more realistic. Many real-world graphs have a power-law degree distribution (Lancichinetti and Radicchi, 2008), e.g. the Internet graph (Chakrabarti et al., 2010). They also have a broad distribution of community sizes, i.e. many small communities coexis with some much larger ones. The tail of the community size distribution can be often quite well described by a power law (Lancichinetti and Radicchi, 2008; Lancichinetti et al., 2010). Our generation model requires six parameters to generate an overlapping graph clustering in three steps: n, k, α, r, p_{in} and p_{out}. In the following the three steps and the related parameters are explained in more detail.

Step 1, Initial Cluster Allocation. In the first step, the nodes of the graph are created and partitioned into clusters. To assign the nodes to the clusters, the parameters n and k are used. The parameter n specifies the number of nodes in the graph and k the number of clusters, to which the n nodes are allocated. We use a power law distribution to assign the nodes to the clusters, in particular the inverse cumulative distribution function of the power law distribution:

$$x = \Phi(y) = t(1-y)^{-\frac{1}{\alpha}} \tag{31}$$

The parameter t indicates the minimum for the value range and will be always equal to 1 in this paper for simplicity. The parameter α can be used to manipulate the degree of the slope and thus changes the differences in the cluster sizes. With a low α the variability between the cluster sizes is higher. To derive the k clusters, we use the inversion method. At first, we generate k random uniform distributed values $p_1, \ldots, p_k$ with $p_i \in [0,1]$. The k function values of the $\Phi(p_i)$ represent the probabilities to assign a node to a cluster C_i. Afterwards the k function values will be normalized with

$$norm_i = \frac{\Phi(p_i)}{\sum_{j=1}^{k} \Phi(p_i)} \tag{32}$$

and mapped to the unit interval. Finally for each node a new random uniform distributed value assigns the node to the cluster C_i which is represented though the $norm_i$ on the unit interval. At the end of step 1, every node is assigned to one cluster and the cluster sizes are power law distributed.

Step 2, Overlap Generation. The overlapping parameter r indicates the number of overlapping nodes. If $r > 1$, then $nr - n$ random nodes are assigned to an additional cluster. With $r = 1$ the graph clustering possesses no overlaps. The overlapping nodes are randomly selected with replacement. Thus, a node can belong to more than two clusters.

Step 3, Edge Generation. In the last step the edges between the nodes are created using the probabilities p_{in} and p_{out}. To avoid giving all nodes the same degree, we use the following scheme: All nodes are sequently added to the graph. While adding a new node $u \in V$ to the graph, all possible node pairs (u,v), consisting of u and an already present node v, are considered. If the two nodes belong to different clusters, a new edge between u and v is created with the probability p_{out}. If the nodes belong to the same cluster, the probability p_{in} is used and will be increased in dependence to degree $d(v)$ of the already present node v to $p_{in}^{(u,v)}$. That is, we do not use p_{in} directly, but use the probability $p_{in}^{(u,v)}$ to create an edge between u and v. Using the degree of the nodes to calculate the probability for a new edge, is a common method for generating artificial graphs, to make them more realistic, and is called *preferential attachment* (Aggarwal and Wang, 2010). In this way, the nodes, rich on edges, get richer as the graph grows, leading to power law effects.

To calculate $p_{in}^{(u,v)}$ we use:

$$p_{in}^{(u,v)} = \frac{\left(\frac{d(v)}{n} + p_{in}\right)}{\left(\frac{d(v)}{n} + 1\right)} \tag{33}$$

The parameter n specifies the number of nodes in the graph we want to generate and remains constant.

With this calculation, nodes with a high degree get an additional edge with higher probability than nodes with a low degree. An analog calculation for p_{out} is not necessary and has drawbacks: Clusterings with no inter-cluster edges can not be created. Even with $p_{out} = 0.0$ the node degrees would increase the values for $p_{out}^{(u,v)}$ and inter-cluster edges can be generated with a low probability. One has to consider that our generation model in theory can create empty clusters. If this is the case, the whole clustering is generated again.

6 EXPERIMENTS

In the following we analyse the properties and the behavior of the new indices in comparison to their original indices. Unfortunately, there is a lack of realistic benchmark graphs with known overlapping structure and with different properties, i.e. number and size of

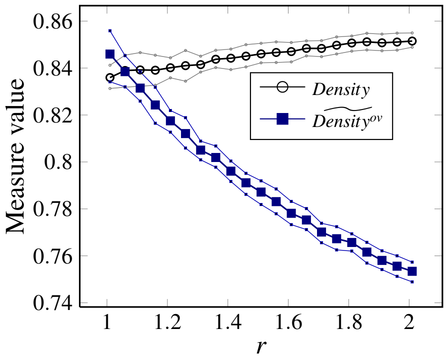

Figure 4: Experiment E01; Parameters: $n = 128$, $k = 8$, $1.01 \leq r \leq 2.01$, $p_{in} = 0.7$, $p_{out} = 0.05$, $\alpha = 3.0$. Each main line shows the mean values, the thin lines correspond to the first and third quintiles.

clusters, size of overlapping parts, degree distribution etc. In order to overcome this problem we use computer generated clustered graphs.

For this analysis, different graph clusterings are generated and evaluated through the indices. To generate the graph clusterings, we use our model from the previous section. Four experiments are presented. In each of them we generate graph clusterings with 128 nodes and 8 clusters. To get an appropriate clustering structure we choose $p_{in} = 0.7$ and $p_{out} = 0.05$ for the edge probabilities. The parameters r or α are varied. In the following, values we report were calculated by taking the mean, first and third quintile of the respective measure on 100 different graph clusterings generated with the same parameters.

In the first experiment $E01$ we compare the original Density (Formula 8) and its weighted extension (Formula 16). We vary the overlap parameter r from 1.01 to 2.01, for $\alpha = 3.0$. In the original density, the weighting for the two criteria intra-cluster density and inter-cluster sparsity are equal to $\frac{1}{2}$, to calculate the average. For this experiment we adapt the weights for the extended density to $\frac{1}{3}$ for ω_f, ω_g and ω_o and $\frac{1}{2}$ for ω_s and ω_m. In Figure 4, the results of experiment $E01$ are illustrated. Both indices show a diverging behaviour. For increasing r the value of density increases slightly. However, the clusterings become harder and harder to interpret as r increases, as the cluster borders get more fuzzy. With $r = 2$ almost each node belongs to two clusters. Thus the original density is inappropriate to evaluate this overlapping graph clustering. In contrast, the value of the weighted extension of density decreases with the increasing degree of overlapping and the consequently decreasing interpretability. Therefore the extension produces an improvement.

In the next experiment $E02$ we compare Newmans's modularity (Formulas 11–13) and its weighted extension (Formula 14). The values for r and α are the same as in the previous experiment $E01$. In Figure 5 the diagram for $E02$ is illustrated. Both indices show a similar behaviour. Not or only marginally overlapping graph clusterings get positive values. The more r increases the stronger the measure values decrease towards negative values (-0.5). This behaviour results from the strong connectivity between the clusters and confirms that the modified modularity handles clusterings with overlaps but still prefers fully separated clusters. One can discover a slight difference for $r \geq 1.5$. The original modularity evaluates these clusterings adequately and therefore the extension is not necessary in this example. Nevertheless there are clusterings, where the value for the original Newman's modularity exceeds the upper interval boundary (see Figure 2). Therefore our weighted extension is a reliable modification to adapt Newmans's modularity for an overlapping graph clusterings.

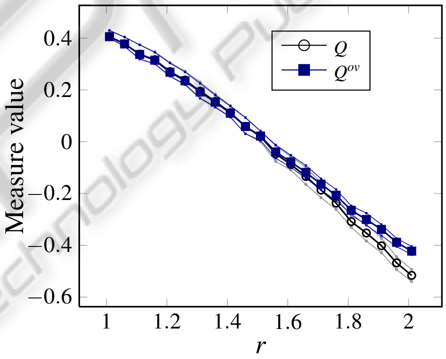

Figure 5: Experiment E02; Parameters: $n = 128$, $k = 8$, $1.01 \leq r \leq 2.01$, $p_{in} = 0.7$, $p_{out} = 0.05$, $\alpha = 3.0$. Each main line shows the mean values, the thin lines correspond to the first and third quintiles.

$E03$ is the third experiment and its results are illustrated in Figure 6. Here the three possible extensions for conductance are compared. In contrast to the other experiments, the parameter $r = 1.1$ is fixed and α is varied between 1.0 and 4.0. Note that for real world networks the scaling exponent α is usually between 2.0 and 3.5 (Lancichinetti et al., 2010; *Appendix S1*). With a low α the variability between the cluster sizes is higher. Thus, the probability to generate a clustering with at least one small cluster that possesses a strong connection outwards is high even with a low α. The value of the conductance is dominated by the lowest partial value for a single cluster. This is why all three extensions return a low value given a low α. The values of the extensions $\sigma^{ov}_{ExFromRest}$ (Formula 19) and σ^{ov}_{Inc} (Formula 22) are almost identical. This results from the similar calculation. The two extensions only differ if the ratio for $\frac{E(C')}{E_{inc}(C')}$ is greater

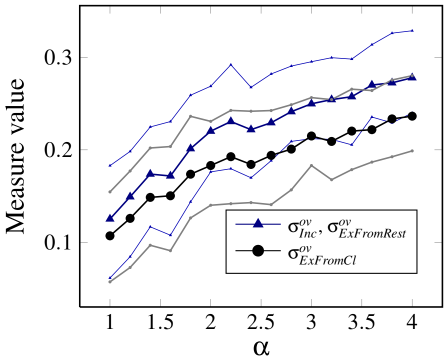

Figure 6: Experiment E03; Parameters: $n = 128$, $k = 8$, $r = 1.1$, $p_{in} = 0.7$, $p_{out} = 0.05$, $1.0 \leq \alpha \leq 4.0$. Each main line shows the mean values, the thin lines correspond to the first and third quintiles.

than the ratio $\frac{E(C)}{E_{inc}(C)}$, which is very seldom. The extension $\sigma^{ov}_{ExFromCl}$ (Formula 25) continuously returns a lower value. Overlapping nodes do not belong to the evaluated cluster in this extension. As there is a high probability that overlapping nodes possess a strong connection to all of their clusters (also due the relative high value p_{in} which leads to the appearance of nodes with high degree), excluding these nodes from the observed cluster causes a strong connection outwards, to the rest of the graph. Therefore the conductance value of the cut induced by the cluster has a low value. This example illustrates how important the manner in which the evaluation for overlaps is integrated into an index is.

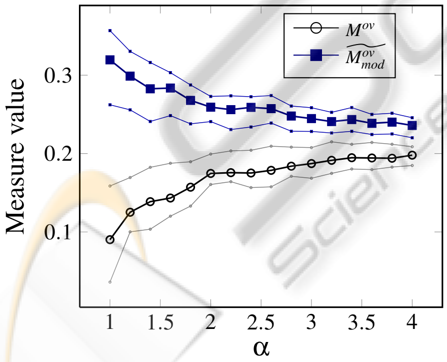

Figure 7: Experiment E04; Parameters: $n = 128$, $k = 8$, $r = 1.1$, $p_{in} = 0.7$, $p_{out} = 0.05$, $1.0 \leq \alpha \leq 4.0$. Each main line shows the mean values, the thin lines correspond to the first and third quintiles.

In the last experiment $E04$ the index M^{OV} and its modification are compared. The parameters for the cluster generation model are identical to experiment $E03$. That means α is the variable parameter again. The experiment $E04$ is illustrated in Figure 7. One can see that the values are almost mirrored horizontally. Low values for α cause relatively high M^{OV} and respectively low $\widetilde{M^{ov}_{mod}}$ values. The higher α is, the more equal the indices values are. In the original M^{OV} every cluster is weighted equally. With a low α there are some small clusters which possess a strong connection outwards and decrease the value. Even well separated large clusters cannot avoid this, because of the equal weighting of the clusters. The modified M^{OV} considers the different cluster sizes, therefore well separated large clusters increase the value and small clusters are neglected.

7 CONCLUSIONS AND FUTURE WORK

In this paper, we discussed the problem of finding appropriate measures to evaluate overlapping graph clusterings. In particular, we proposed three methods to adapt existing crisp evaluation measures to handle overlapping graph clusterings in an appropriate manner. We proposed to modify the quality indices in a direct way, by incorporation of edge weights and in an indirect way. When taking a direct way, the quality measure evaluates not only the intra-cluster density and the inter-cluster sparsity but also measures the quality of the overlapping parts e.g. considering the overlap size and the membership of the overlapping nodes. We demonstrated the first extension method on the density measure.

If a crisp evaluation measure is applied directly to a graph clustering with overlaps, the calculation contains inaccuracies because overlapping nodes and edges can be considered multiple times. Our second extension method uses edge weights, so that overall each element is considered exactly one time. Incorporation of the edge weights was demonstrated on Newman's modularity. We applied the incorporation of the node weights to the M^{OV} index.

In the third method, the evaluation of the overlapping parts is integrated indirectly. That is, depending on the extended measure and its criteria, one has to decide how the overlapping parts are handled. There are indices, which assess the quality of each cluster separately, and then use the obtained values to calculate the overall "goodness" of the clustering. For these measures, the overlapping parts can be handled in different ways depending on the decision where to make the cut between the observed cluster and the rest of the graph. We showed three possible extensions and gave an example using the conductance measure.

M^{OV} is one of the few already existing indices for overlapping graph clusterings. In this paper, we modified it to make M^{OV} more sensitive for different cluster sizes. The idea is that the quality of large clusters should have more influence on the index than the quality of small clusters. This can be done using the weighting of clusters qualities depending on the cluster size.

To analyse the new measures, in particular the influence of the modification on the original measure, we used a generation model for overlapping graph clusterings. The model is a modification of a common method. We enhanced it using a power law distribution of cluster sizes and node degrees to produce more realistic clusterings. The experiments with this generation model confirmed that all our extensions for crisp evaluations measures provide an appropriate and reliable adaption to handle overlapping graph clusterings.

In the future we are also going to test the new measures on data from real-world networks. Another potential research topic for future work is the adaption of our extension methods to overlapping clustering on directed or/and weighted graphs. One more interesting research question is, whether the new measures can also be successfully used to generate overlapping graph clusterings. In future work we will study if using the proposed measures as fitness functions within overlapping clustering algorithms will improve the clustering performance for overlapping clusterings.

REFERENCES

Adamcsek, B., Palla, G., Farkas, I., Derényi, I., and Vicsek, T. (2006). Cfinder: locating cliques and overlapping modules in biological networks. *Bioinformatics*, 22(8):1021.

Aggarwal, C. and Wang, H. (2010). *Managing and Mining Graph Data*, volume 40. Springer-Verlag New York Inc.

Ahn, Y., Bagrow, J., and Lehmann, S. (2010). Link communities reveal multiscale complexity in networks. *Nature*, 466(7307):761–764.

Baumes, J., Goldberg, M., and Magdon-Ismail, M. (2005). Efficient identification of overlapping communities. *Intelligence and Security Informatics*, pages 27–36.

Brandes, U., Delling, D., Gaertler, M., et al. (2007). On modularity clustering. *IEEE Transactions on Knowledge and Data Engineering*, pages 172–188.

Brandes, U. and Erlebach, T. (2005). *Network analysis: methodological foundations*, volume 3418. Springer Verlag.

Brandes, U., Gaertler, M., and Wagner, D. (2003). Experiments on graph clustering algorithms. *Algorithms-ESA 2003*, pages 568–579.

Chakrabarti, D., Faloutsos, C., and McGlohon, M. (2010). Graph mining: Laws and generators. *Managing and Mining Graph Data*, pages 69–123.

Delling, D., Gaertler, M., Görke, R., Nikoloski, Z., and Wagner, D. (2006). *How to evaluate clustering techniques*. Univ., Fak. für Informatik, Bibliothek.

Fortunato, S. (2010). Community detection in graphs. *Physics Reports*, 486(3-5):75–174.

Gavin, A., Bösche, M., Krause, R., Grandi, P., Marzioch, M., Bauer, A., Schultz, J., Rick, J., Michon, A., Cruciat, C., et al. (2002). Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature*, 415(6868):141–147.

Girvan, M. and Newman, M. (2002). Community structure in social and biological networks. *Proceedings of the National Academy of Sciences of the United States of America*, 99(12):7821.

Gregory, S. (2007). An algorithm to find overlapping community structure in networks. *Knowledge Discovery in Databases: PKDD 2007*, pages 91–102.

Lancichinetti, A., Kivelä, M., and Saramäki, J. (2010). Characterizing the community structure of complex networks. *PloS one*, 5(8):e11976.

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

Lázár, A., Ábel, D., and Vicsek, T. (2010). Modularity measure of networks with overlapping communities. *EPL (Europhysics Letters)*, 90:18001.

Nepusz, T., Petróczi, A., Négyessy, L., and Bazsó, F. (2008). Fuzzy communities and the concept of bridgeness in complex networks. *Physical Review E*, 77(1):016107.

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

Nicosia, V., Mangioni, G., Carchiolo, V., and Malgeri, M. (2009). Extending the definition of modularity to directed graphs with overlapping communities. *Journal of Statistical Mechanics: Theory and Experiment*, 2009:P03024.

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

Schaeffer, S. (2007). Graph clustering. *Computer Science Review*, 1(1):27–64.

Tan, P., Steinbach, M., Kumar, V., et al. (2006). *Introduction to data mining*. Pearson Addison Wesley Boston.