On Community Detection in Real-World Networks and the Importance of Degree Assortativity

Marek Ciglan
Institute of Informatics
Slovak Academy of Sciences
Bratislava, Slovakia
marek.ciglan@savba.sk

Michal Laclavík
Institute of Informatics
Slovak Academy of Sciences
Bratislava, Slovakia
michal.laclavik@savba.sk

Kjetil Nørvåg
Dept. of Computer and
Information Science
Norwegian University of
Science and Technology
Trondheim, Norway
kjetil.norvag@idi.ntnu.no

ABSTRACT

Graph clustering, often addressed as community detection, is a prominent task in the domain of graph data mining with dozens of algorithms proposed in recent years. Community detection algorithms are commonly evaluated against artificially generated networks with planted communities. In this paper, we focus on several popular community detection algorithms with low computational complexity and with decent performance on the artificial benchmarks, and we study their behaviour on the real-world networks. The motivation is that there is a class of networks for which the community detection methods fail do deliver good community structure. For example, when these community detection methods were used to find clusters in the information network of DBPedia, we observed that in the resulting community structure for all the used algorithms, the majority of the nodes belonged to a small number of very large clusters. This result is contradictory to expectations, and a systematic study of the given algorithms behaviour were conducted. In this paper, we first study the relationship between different network properties and the type of community structure unveiled by the given algorithms. Results indicate a statistically significant correlation between the network assortativity coefficient and the size of top k largest clusters. We examine the assortativity of ground-truth communities and show that assortativity of a community structure can be very different from the assortativity of the original network. We then examine the possibility of weighting edges of a network with the aim to improve the community detection outputs for networks with assortative community structure. The evaluation shows that the proposed weighting can significantly improve the results of community detection methods on networks with assortative community structure.

Keywords

community detection, network assortativity, edge weighting

1. INTRODUCTION

The goal of community detection in networks is to identify sets of nodes, *communities,* that are densely connected among themselves and have weaker connections to the other communities in the network. It is a task that should help to analyse large graphs and identify significant structures within. A classical example is to analyse social networks in order to find social groups of users. Community detection faces numerous challenges, the principal one is the lack of a consensus on the formal definition of a network community structure. The unclarity of the task definition results in a significant number of community detection algorithms having been proposed, using different quality definitions of a community structure or leaving the problem formulation in an ambiguous informal description only. In this paper, we do not propose yet another community detection algorithm, nor do we attempt to provide new formalization of the task. Rather, we study community detection methods on real-world networks and the possibilities to improve their precision by pre-processing the network topology.

Motivation. The motivation is that for a class of networks, the community detection techniques fail to deliver a good partition. For example, when using community detection techniques to analyse the semantic network DBPedia[1], where nodes corresponds to the DBPedia concepts and edges denote a relation defined between two concepts, our expectation was that the analysis would reveal small clusters with semantically related concepts and entities. The clusters were expected to be, for example, similar to Wikipedia categories (containing groups of Wikipedia articles handpicked by human contributors and assigned to be a member of the category, class). However, the detected structure contains a few very large communities comprising the majority of the nodes.

Due to the size of the data sets, our choice of community detection methods for the network analysis was limited to a small family of fast community detection algorithms that are pseudolinear in the time complexity. We have analysed the link graph using Label Propagation [16] algorithm, a greedy modularity optimization algorithm [2] and a community detection method with parametrizable community size constraint (SCCD) [3]. The community structure produced by the label propagation algorithm had the largest community, with over 2.96 million of nodes. The SCCD method with default setting yielded a structure with 78% of the nodes in the 20 largest communities, and the greedy modularity optimization method produced partition with 88% of the nodes in the 20 largest clusters. The obtained results clearly did not match our expectations of the community structure of the DBPedia network.

Overview of the study. We first analyse several large social and information networks with known ground-truth communities and compare the detected structure obtained by the three community

[1]Knowledge base derived from Wikipedia, http://dbpedia.org

KDD '13

detection algorithms to the ground-truth clusters. On four of the analysed networks, the detected clusters are a decent approximation of the ground-truth communities. For the rest of the networks, the similarity scores of the yielded partitions with the ground-truth communities are low and at the same time, a few of the largest detected clusters contain the majority of the network nodes.

To find out whether this is a generic behaviour of the algorithms and what are the causes of such behaviour, we have applied the algorithms on a wider range of different networks, including collaboration social networks, citation networks and web graphs. The goal was to identify the correlation between the type of the detected community structure and diverse properties of the analysed networks. The results reveal statistically significant correlations of the size of top largest clusters with two properties — network degree assortativity and number of nodes in the network. We then study the assortativity of the ground-truth communities. Based on the results of the latter, we examine the possibility to modify the network by means of edge weighting. The underlying idea is to examine whether we can increase the precision of the community detection algorithms by the weighting functions for the networks with assortative community structure. We describe the weighting heuristics and show empirically that it is a suitable approach in practice. On our test data, the similarity of the detected community structure to the ground-truth is significantly increased after applying the weightings on network with assortative community structure.

The main contributions of the paper are:

- We show correlation between degree assortativity coefficient of a network and the type of community structure yielded by community detection algorithms with linear time complexity.
- We show that assortativity of the community structure can differ from overall assortativity of the network.
- We propose edge weighting functions designed to decrease the influence of edges connecting disassortative nodes. We show that such edge weighting on networks with assortative community structure can increase significantly the similarity of the communities identified by community detection methods to the ground-truth communities, an increase 2 to 10 times compared to the baseline solution is reported.

The organization of the rest of the paper is as follows. Section 2 describes related work and the community detection methods with near-linear time complexity used in this study. Section 3 evaluates the precision of the algorithms on real-world networks with known ground-truth communities. The finding is, that although on the most of the studied networks the algorithms performs decently, there are networks for which the overlap of the detected clusters and the ground-truth communities is very low. The correlation of the detected community structures and network properties is studied in Section 4. The assortativity of the ground-truth community structures are examined in Section 5. In Section 6, the edge weighting designed to decrease importance of edges connecting disassortative nodes is studied and important increase in precision of community detection methods is observed on networks with assortative community structure. We conclude the paper in Section 7.

2. PRELIMINARIES AND RELATED WORK

This section summarizes related work and preliminaries. We first discuss the community detection task and then provide more detailed overview of the three algorithms with near-linear time complexity that we use in our study.

2.1 Community Detection

The task of community detection is to, given an input network, find the community structure of the network. A community structure is considered to be a collection of clusters of densely connected vertices that are less densely connected to other parts or communities in the network. The problem has been a very popular research topic and has been extensively studied in recent years; good evidence of the topic's popularity is the overview paper by Fortunato [6] with more than 450 references. Since its publication in 2010 (until Feb 2013), it has attracted more than 1200 citations according to the Google Scholar service. Despite the significant research effort on this problem, there is no consensus on the formalization of the task and authors often use different definitions of a community or even leave the notion of the community structure in an informal description. The most widely used approach is to focus on maximizing the modularity measure (introduced by Newman and Girvan[14]) that compares how community-like is the partition of the input network to a random network with the same degrees of vertices. Modularity is a quality function for estimation of how good the partition of a network is. The basic formulation of the community detection task expects as an output a partition of a network; that is, each node is a member of exactly one community. Numerous variants of the problem have been studied, including detection of overlapping communities (a vertex can belong to multiple communities) (e.g., works by Gregory [7] and Zhang et al. [19] , clustering of bipartite graphs (e.g., Papadimitriou et al. [15]) or detection of clusters exploiting additional information than network structure (e.g., attributes on nodes/edges) (e.g., Yang et al. [18]).

Community-detection algorithms are usually evaluated against artificial benchmark graphs, where a community structure has been injected (e.g., [9]). The advantage is that the evaluator can tune the parameters of the generated network, the disadvantage is the artificiality itself. The real-world networks with known community structure studied in the literature are usually small ones (e.g., Zachary's karate club (36 nodes) or Dolphin social network (62 nodes)), with few exceptions; e.g., in a recent study by Yang and Leskovec [17], the authors identify the ground-truth communities for several large networks and study their properties. In this work, we reuse their data sets.

Leskovec et al. in [12] provide an empirical comparison of community detection algorithms, studying community quality scores of clusters detected by various algorithms, with the focus on the conductance measure. One of the main finding of their study was that although community detection methods often optimize the clusters quality scores nicely, there are classes of networks where the community detection techniques perform sub-optimally. Our work confirms that observation for the pseudo-linear community detection approaches and we show that, to a certain extent, we can overcome that behaviour. We demonstrate that a weighting of the edges, based on the assortativity of connected nodes, can increase the precision of the studied algorithms for a class of networks. The concept of edge weighting as a preprocessing for community detection has been explored before, e.g., in [1][8] where authors use different approach to the weighting; they re-weight the edges based on edge betweenness centrality and common neighbour ratio.

2.2 Algorithms with Near-Linear Complexity

A large number of the community detection methods proposed in the literature are heuristics with polynomial time complexity. In practice, the time complexity often limits the usability of a large part of community detection algorithms to small networks (e.g., see the comparative study in [10]). In this work, we intend to study the behaviour of community detection methods on large real-world

networks, which limit our options to a small family of community detection techniques with near-linear time complexity. This section discusses the fast greedy algorithms we have used. The discussed algorithms are the Label Propagation approach by Raghavan et al. [16], a heuristic for modularity optimization by Bondel et al. [2] and a heuristic for Size-constrained Community Detection (SCCD) [3]. The base method is the Label Propagation (LP), the two other use the same underlying principle as LP with several modifications.

Label Propagation algorithm. The Label Propagation approach is based on the simple idea that a node should be assigned to the community to which most of its neighbours belong to. When starting from scratch, i.e., the community structure is unknown, the label propagation algorithm assigns a unique label to all the vertices. In randomized order, the algorithm iterates over vertices and recomputes the the label for each vertex in the following way: we compute number of neighbours with distinct labels and set the vertex label to the label with the most members. It is repeated in iterations in which the community membership of all the nodes are updated in random order. If, during an iteration, none of the vertices changes its label, the algorithm stops. However, this stop condition is not guaranteed as there might be nodes changing their membership in each iteration (e.g., nodes with equally strong ties to several stable clusters). It has been argued by the authors that in practice a good community structure is found after a few iterations and we can set the maximal number of the iterations to be performed. This results in linear time complexity (a constant number of iterations over N vertices). The problem of the method is that one (or a few) label(s) often becomes dominant and the majority of the network collapses into one single community.

Louvain method. The Lovain method proposed by Bondel et al. in [2] has two phases. The first one is based on a process very similar to the label propagation. The main difference is the label assignment, the Louvain method does not use the number of neighbours with the same labels as principal factor. Rather, it computes a possible gain in the modularity in case a vertex changes its label to other given label. In the second phase, a new community network is constructed by contracting community members into single nodes. The first phase is then used on the derived network and a hierarchy of communities is constructed.

SCCD. The Size-Constrained Community Detection method [3] is another method based on the Label Propagation principle. The main difference is in the scoring function that governs the label assignment of a node, where SCCD method discriminates the scores according to the sizes of target communities. The underlying idea of the approach is to increase the ability to detect small-sized, compact clusters independently of the network size, as opposes to the modularity optimization methods that are known to have a resolution limit and tend to increase the size of generated communities with the increase of the network size.

All the three algorithms have near-linear time complexity, and the two latter have a confirmed decent precision on the artificial benchmark networks [3, 10], competitive to other well-performing methods.

3. DETECTED COMMUNITIES VS. GROUND-TRUTH CLUSTERS

In this section, we evaluate the three studied algorithms on large, real-world data sets with ground-truth clusters. The main motivation is to observe, whether the partitions produced by community detection algorithms approximate well the ground-truth communities. We consider this section particularly interesting as this exercise puts algorithms to a tough test because of the large sizes of the networks that are studied. We first describe the data sets, focusing mainly on how the ground-truth communities were identified, and introduce a ground-truth communities data set for the DBPedia knowledge graph. We then describe the evaluation functions used to assess the quality of the detected partition compared to ground-truth communities. We summarize the results of the analysis and discuss questions raised by the experiment; namely, the suitability of the ground-truth communities data for the community detection task and the failure of the used community detection methods to approximate well ground-truth communities for a class of networks.

3.1 Data Sets

We reuse five networks with ground-truth communities from the Stanford Large Network Dataset Collection (SLNDC) and we introduce ground-truth communities for the network of DBPedia.

SLNDC data sets. We have used five graphs from the SLNDC collection containing the ground-truth communities, introduced in [17]. LifeJournal social network containing friendship network where user-defined groups are considered as the ground-truth communities. Similarly, user defined groups are considered as ground-truth communities in the Orkut and YouTube datasets. The DBLP dataset provides co-authorship network. Here, the ground-truth communities are created by grouping authors publishing on same venue or journal. The Amazon data network contains products as nodes and the edge indicates co-purchasing relation. The ground-truth clusters are equivalent to the product category in Amazon. In all the data sets, the identified groups were further split to connected components and each connected component is regarded as a distinct ground-truth community. In addition, all communities with less than 3 nodes were removed.

Ground-truth communities for DBPedia. We describe the data set containing ground-truth communities for the DBPedia separately, as it has not been used so far in the context of the community detection. The DBPedia is a knowledge base derived from Wikipedia, mostly by parsing the infoboxes of Wikipedia articles. It can be viewed as a graph of interconnected entities, where the entities can have properties (e.g., a concept related to a person can have attributes such as birth date, height, occupation) and are linked by labelled relations or edges to other entities. In our previous work on ad-hoc retrieval from semantic data we were faced with the challenge to extract semantically related sets of entities from the DBPedia knowledge base [4]. As the first approximation we took the members of Wikipedia categories as such semantic sets. In DBPedia, the category membership translates into relation labelled 'subject' connecting members to the entity representing the category. This approach has two disadvantages: a) even though large number of Wikipedia categories group semantically related entities, there is a lot of trivia categories (e.g., Category:1970_births containing people born in 1970 - those entities can hardly be considered similar in other respects than the date of the birth); b) data set size - the categories contain only subset of such the semantically related sets (e.g., for some music bands, there could be a category grouping 'members_of ..', while there is no such category for a number of other entities of the same type).

We have approached b) by exploiting the semantic relations (labelled edges) in DBPedia to identifying additional potentially semantically related sets. We have selected all the sets of vertices that fit to the following two patterns: a set of vertices connected by an outgoing edge of the same label to a common vertex v, or a set of vertices that have an incoming edge of the same label from a single vertex. We have used all the labels, except the '*wikilink*' label that denotes an existence of a hyperlink between the two Wikipedia articles, but the true semantics of the relation is not given.

The remaining problem was to distinguish good semantic groups and the trivia groups. We have used several similarity scores to measure the relatedness of the entities in the sets. We have used text-based similarity (as the DBPedia nodes contain also abstracts of related articles, we have quite a rich textual component) as well as structural similarities (using the topology of the DBPedia graph). The details are provided in [4], and to summarize, most of the used similarity measures had a high correlation, even the text based cosine similarity with the structural similarities. We then chose the sets with similarity scores above certain threshold.

We exploit the data set of the semantically related entities and use them to identify the ground-truth communities. From the candidate set we select groups with high internal density measure. The internal density expresses how clique-like is the subgraph generated by the given set of vertices; more formally, let $a_{i,j}$ be an element of the adjacency matrix for G, the internal density is $\psi(S) = \sum_{i,j \in S; i \neq j} a_{ij}/|S| \times |S-1|$.

We have selected all sets having more than 3 members and having score of internal density higher than or equal to 0.1. A group of nodes with high internal density intuitively corresponds to a good community. We have also selected 5000 communities based on their internal density scores. The data sets are provided on the support web-page[2].

3.2 Comparing Detected and Ground Truth Clusters

For comparison of the detected and ground-truth communities, we use two measures. The first one is based on set similarity and is introduced in the following text; the second one is Normalized Mutual Information which is popular in the community detection literature. We report values for both in the presented results.

Comsim - Set similarity based measure. Algorithms studied in this work output a partition of a network, where each node is assigned to exactly one community. The ground-truth communities in our collection contain groups of nodes that can have non-empty intersection, i.e. a vertex can be a member of multiple communities. In addition, not all the vertices of the network have to belong to a ground-truth community. To compare such structures, we need a function that measures their similarity. In addition, the ability to assess approximation of the ground-truth communities one by one would be of advantage, allowing us to analyse the produce results on a cluster level, rather than on the partition level. To achieve that, we compute a similarity score of each of the ground-truth community with the most overlapping group from the network partition identified by community detection algorithm. The overall score would be an average of the similarity scores of distinct ground-truth communities. As we need to compare two subsets, the straightforward approach is to use Jaccard's similarity coefficient: $J(A,B) = |A \cap B|/|A \cup B|$. The remaining point to solve is how to find the best fitting cluster from a network partition for a given set. We select the cluster with the largest intersection with the given ground-truth set. In case there are multiple sets with the same, maximal size of intersection, we select the smallest one.

More formally, let $G = (V, E)$ be the graph, let $D = \{P_1, P_2, \ldots, P_k | \forall P_i, P_j : P_i \subset V, P_j \subset V, P_i \cap P_j = \emptyset \wedge \bigcup_{i=1..k} P_i = V\}$ be the partition produced by the community detection algorithm, let $E = \{T_1, T_2, \ldots, T_l | \forall_i T_i \subset V\}$ be the set of ground-truth communities. $o(T, D)$ is a set of clusters from D having maximal intersection with T :

$$o(T, D) = \{P : P \in D \wedge \forall_{P_i \in D} |P_i \cap T| \leq |P \cap T|\}$$

[2]http://ups.savba.sk/~marek/communities.html

Table 1: Similarity scores - comsim and NMI (in parenthesis) for detected network partitions and the ground-truth data sets.

	Louvain	Label Prop.	SCCD
DBLP-top5K	0.53 (0.24)	0.49 (0.25)	0.51 (0.26)
Amazon-top5K	0.76 (0.37)	0.85 (0.46)	0.86 (0.46)
YouTube-top5K	0.23 (0.09)	0.08 (0.02)	0.19 (0.05)
Orkut-top5K	0.04 (0.03)	0.03 (0.02)	0.24 (0.06)
LJ-top5K	0.52 (0.28)	0.57 (0.32)	0.60 (0.32)
DBPedia-top5K	0.004 (0.0008)	0.003 (0.003)	0.13 (0.06)
DBLP-all	0.34 (0.12)	0.32 (0.14)	0.32 (0.14)
Amazon-all	0.42 (0.28)	0.26 (0.24)	0.28 (0.25)
YouTube-all	0.11 (0.03)	0.03 (0.01)	0.10 (0.02)
Orkut-all	0.001 (0.05)	0.0002 (0.04)	0.0091 (0.03)
LJ-all	0.03 (0.02)	0.01 (0.03)	0.04 (0.02)
DBPedia-all	0.004 (0.005)	0.003 (0.004)	0.06 (0.02)

best fitting clusters for a ground-truth community T is:

$$b(T, D) = \{P : P \in o(T, D) \wedge \forall_{P_i \in o(T,D)} |P_i| \geq |P|\}$$

Fitting score for a ground-truth community T is then $sim(T, D) = J(T, B) : B \in b(T, D)$. Overall score for D and E is

$$comsim(E, D) = \frac{\sum_{T_i \in E} sim(T_i, D)}{|E|}$$

.

Normalized mutual information. The second measure we use is Normalized Mutual Information (NMI) [5]. It is a standard measure used in the community detection literature to compare two partitions of a network. It has been designed to evaluate non-overlapping partitions. Lancichinetti et al. in [11] have proposed generalized NMI able to compare overlapping communities as well, and we also report the latter in our experiments.

Evaluation. The number and quality of the ground-truth communities can vary for different data sets. We follow the approach of [17] where they selected the top 5000 communities for each of the ground-communities data set, based on average rank of the community for six different community scores. Extracting and using top k communities enables experimentation with high quality clusters. We have selected the top 5000 communities for the DBPedia data set as well. In our case, we have used the rank of the community according to conductance and internal density measures. For the evaluation purposes, we have for all the networks used both the complete and top-5000 ground-truth communities data sets. We have used the three studied community detection algorithms to cluster the six networks and we have compared received partitions with ground-truth communities with $comsim$ score and generalized *NMI* measure.

The results are summarized in Table 1, where the top half presents scores for the top 5000 community sets, the bottom half for full ground-truth communities sets. Scores for the top 5000 communities are, not surprisingly, significantly higher than the scores for the data sets with all ground-truth. Another important observation is that for some networks, the similarity score is quite high (DBLP, Amazon, LJ, Youtube), for the Orkut and DBPedia data sets, the scores are low. There are two possible explanations that we are going to discuss in more detail. First possible explanation could be that the ground-truth data sets for the three networks with low scores are flawed, and do not capture the real communities within the networks and are unsuitable for evaluation of the community detection task in general. The second explanation could be that the ground-truth communities are suitable for the evaluation, but the

studied algorithms have been unable to approximate the network community structure precisely.

Suitability of the used ground-truth communities for the community detection task. The ground-truth communities were derived from network data where nodes where explicitly assigned to sets (e.g., user defined groups). As the results in Table 1 indicate, the community detection algorithms were not very successful in approximating the ground-truth communities. Legitimate concern is whether the used data sets are suitable for testing community detection in general. The question is whether the assignment to ground-truth communities corresponds to the community structure of a network. There is no consensus on the formal definition of a community structure, the general notion is that a good community has nodes that are densely linked among themselves and are less densely linked to other communities in the network. The latter requirement becomes questionable under the model of overlapping communities. We thus focus on the notion of a community as a densely linked group of nodes. Internal density function is a suitable measure to capture how clique-like a community is (the measure has been described in Section 3.1). We have computed the internal density of the ground-truth communities and have compared it to internal density of a same sized group of nodes in a random graph with identical number of network elements (nodes and edges). The results are presented in Table 2 and show that the average internal density of ground-truth communities is high, orders of magnitude higher than in the case of random graphs. Conclusion is that the ground-truth communities used in the experiments are related to the informal notion of a community as a densely connected subgraph. A possibility that cannot be dismissed from the results in Table 2 is that some of the ground-truth set are possibly subsets of the 'real' communities. Even in that case, an important overlap with the correctly identified communities should exist.

Failure of community detection algorithms. Based on the observations above, we conclude that the low scores for community detection methods on three networks indicate their inability to detect the communities precisely. Observation that there are classes of networks where the community detection techniques perform suboptimally has been also reported in [12]. Based on our preliminary observation from the introduction about the sizes of top largest detected communities on DBPedia, we have looked at the community sizes for the partitions detected on the six networks with ground-truth communities. Table 3 presents the sizes of top 50 largest communities in the partitions obtained by the Louvian method. As the results show, for the ORKUT and DBPEDIA data sets, the majority of the nodes have been members of a few largest clusters. For those networks, also the similarity scores with the ground-truth communities have been very low. In the following section, we study systematically the relation of the sizes of the largest communities and network properties on a larger set of real-world networks.

Table 2: Average internal density for the ground-truth communities and equally sized groups of nodes in random graphs. Columns prefixed by '$top5k$' report numbers related to top 5000 best communities, columns prefixed by 'all 'concerns all the ground-truth communities in the data set.

	$top5k$	$tok5k - RND$	all	$all - RND$
DBLP	0.71	9.17×10^{-4}	0.52	2.21×10^{-3}
AMAZON	0.62	4.89×10^{-4}	0.49	8.99×10^{-4}
YOUTUBE	0.35	2.31×10^{-4}	0.37	1.64×10^{-4}
LiveJournal	0.78	2.54×10^{-4}	0.43	2.77×10^{-4}
ORKUT	0.31	1.00×10^{-2}	0.45	1.00×10^{-3}
DBPEDIA	0.98	8.34×10^{-5}	0.32	1.58×10^{-4}

Table 3: Fraction of nodes belonging to the top 50 communities in the partition identified by Louvain method.

net	fract. of N	net	fract. of N
DBLP	0.11	LJ	0.67
Amazon	0.03	Orkut	0.97
Youtube	0.50	DBP	0.98

4. NETWORK PROPERTIES AND DETECTED COMMUNITIES

In this section, we study the structures identified by community detection algorithms with pseudo-linear complexity on real-world networks and the relation of the detected structures to the properties of the analysed networks. The motivation is the finding from the previous section that community detection methods, when applied on the real world networks, can fail to approximate well their community structure and, at the same time, the majority of the nodes are assigned to a few large communities containing the majority of the nodes.

Thus we set up experiments where we apply the chosen algorithms to a variety of real-world networks with the goal to observe the cardinality of the largest communities. The rationale is that a presence of very large groups in the detected community structure probably indicates the inability of the community detection techniques to identify densely connected clusters in the network and assigns the members of its dense core to a small number of large groups. During the first inspection of the preliminary results on different network types, we have observed that clusters detected from collaboration networks fit our expectations of the community structure, having at most several hundreds of nodes in the largest communities. Other network types of comparable sizes had identified community structures comprising several thousands or tens of thousands of nodes. Similar partition of networks was observed in the study by Newman et al. [13], where the authors have shown the tendency of the collaboration and social networks to be assortative (considering the degrees of vertices), while other studied network types (technological and biological) were found to be dissasortative. We have thus set up an experiment to study whether there is a relation between the properties of the analysed networks (including the assortativity) and the sizes of the largest clusters detected by the fast heuristic community detection methods. In the rest of the section, we describe the examined network properties, setup of the experiment, and its results.

4.1 Examined Network Properties

The analysis of real-world networks has shown that a lot of those network types have several common properties, such as power-law degree distribution, small-world property (any two random nodes can be connected by a path with short distance), and high clustering coefficient. In this section, we provide a short description of the properties used in this work to identify correlations with the detected community structures.

Degree assoratitivity coefficient. Degree assoratitivity coefficient (AC) denotes a tendency of nodes to be connected with other nodes of similar degree. It is defined as the Pearson correlation coefficient of degrees of pairs of nodes connected by an edge in the network ([13]). Let M be the number of edges, j_i and k_i be the degrees of the i-th edge, the assortativity coefficient can be computed

as follows:

$$r = \frac{M^{-1}\sum_i j_i k_i - [M^{-1}\sum_i (j_i + k_i)/2]^2}{M^{-1}\sum_i (j_i^2 + k_i^2) - [M^{-1}\sum_i (j_i + k_i)/2]^2}$$

According to [13], social networks tend to have positive assortativity coefficient so the networks are assortative, while networks such as internet, biological networks or other information network tend to have negative assortative coefficient and we refer to them as being disassortative. For directed networks, we can distinguish in-degree (number of edges oriented towards the nodes) and out-degree (number of edges oriented from the node). In our study we consider assortativity of in-degree, out-degree and total degree (in+out degree).

Average clustering coefficient. The local clustering coefficient is a measure expressing how much of the node's neighbours are also neighbours. It is done by measuring number of existing edges between the neighbours of a vertex and all possible edges between them. The clustering coefficient is equal to 1 when the nodes neighbours form a clique. Clustering coefficient of a vertex i with out-degree k_i is: $C(i) = |e_{j,k} : e_{i,j}, e_{i,k}, e_{j,k} \in E|/k_i(k_i - 1)$

Fraction of closed triangles. Fraction of closed triangles is the number of triangles in the network divided by number of paths of length two.

Network diameter. Diameter is the longest of the shortest paths in the network. Due to computational complexity, it is computed from a sample of 1000 randomly chosen nodes.

4.2 Correlation of Network Properties and the Discovered Community Structure

In this subsection, we first discuss the data set used in the experiments; we then describe the experimental setup and present results and their interpretation.

The data set. We have used a subset of the Stanford large network dataset collection[3], a collection of graph data comprising real-world networks of various types, including social, citation, collaboration networks or web graphs. The collection contains data sets of various sizes, ranging from hundred thousands of graph elements (nodes and edges) to hundred millions of graph elements.

Methodology. For the experiments, we have split the data sets according to their sizes into three groups; networks with number of elements in A) hundred of thousands, B) millions, C) tens of millions. In the following text, we will refer to those sets as A or medium sized networks, B or large networks and C or very large networks. The split and subsequent experiments allow us to observe the behaviour on classes of networks with different sizes. We have run the community detection algorithms on the networks and obtained the partitions of the networks into disjoint clusters. We have then computed the Pearson product-moment correlation coefficient between the network properties described in 4.1 and the cardinality of top k largest communities. The Pearson correlation coefficient measures the linear dependency between variables, giving a value between -1 and 1. We have performed separate evaluation for the network sets A, $\{ A \cup B \}$ and $\{ A \cup B \cup C \}$, to see whether results differs with inclusion of larger data sets.

Results. Reported results are correlations of network properties and sum of top 50 communities for a network partition by a given community detection algorithm. We have done experiments with cardinality of top k, with k ranging from 20 to 200, the differences where minor and we consider the results for $k = 50$ representative for the performed experiments. We summarise results in tables, where we highlight statistically significant correlations using bold and underlined text. Bold text is used to denote statistical significance (at the 5% level) for a non-directional hypotheses test, underlined text denotes significance for a directional hypotheses test. In our case, a directional hypothesis would postulate a positive or negative correlation, a non-directional hypothesis just assumes a correlation. As in our study, we were testing for a correlation without the directional assumption, the values in bold text are the ones to observe. The abbreviations in the tables are as follows: *#nodes* is number of nodes, *#edges* is the number of edges, *In AC, Out ACAC* denote the assortativity coefficient of nodes in-degrees / out-degrees and the total degree (in plus out-degree), *Ccoef* denotes the average local clustering coefficient, *tri* denotes percentile of triangles and *diam* denotes the diameter. We have used directed as well as undirected networks in the experiments, for the undirected networks the values of *In AC, Out ACAC* are identical.

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

Table 4: Pearson correlation of network properties and cardinality of largest 50 communities for collection containing medium networks (set A) only. Number of networks used: 29, bold/underlined correlations are statistically significant for non-directional/directional test.

prop. \ CD algo	Louvain	Label Prop.	SCCD
#nodes	**0.538**	**0.453**	**0.571**
#edges	0.362	0.306	**0.416**
In AC	-0.358	-0.318	-0.333
Out AC	-0.111	-0.1	-0.044
AC	**-0.432**	**-0.39**	**-0.398**
Ccoef	-0.112	-0.159	-0.041
tri	-0.316	-0.313	-0.263
diam	-0.238	-0.317	-0.251

Table 4 summarises the correlations on medium networks (set A containing networks with elements in order of hundred thousands). The data set consists of 29 networks of different types (e.g., social, citation, collaboration, P2P). We can observe a correlation with the number of nodes meaning the larger the analysed network was the more nodes were members of the largest communities. There is also a negative correlation with the assortativity coefficient, which we can interpret in the sense that assortative networks (network AC > 0) had fewer nodes in the largest communities compared to disassortative networks (network AC < 0). A possible explanation of the correlation with degree assortativity could be that the community detection algorithms produce results correlated with the network size and in our data sample, there is a strong correlation between the number of nodes and the assortativity of a network and therefore the correlation of with assortativity coefficient could be due to the bias of the data sample. To verify this hypothesis, we have computed the correlations of the nodes and edges of the networks with degree assortativity; the correlation of the number of nodes and AC is -0.08 and the correlation of the number of edges and AC is -0.05. As the correlation of network size and the assortativity coefficient is insignificant for the data set, we can dismiss the latter hypothesis and we can assume that assortative networks have smaller largest communities than the dissassortative ones.

As we had only few networks in data sets B and C to produce credible statistics we have decided that instead of studying correlations on B and C alone, we will do the experiments on $\{A \cup B\}$ and $\{A \cup B \cup C\}$. Table 5 summarises the results for data set $\{A \cup B\}$. We can observe that for the Louvain method, the correlation with assortativity coefficient have disappeared while it remains significant for Label Propagation and SCCD algorithms. We can also observe correlation with the number of network elements. The cor-

Table 5: Pearson correlation of network properties and cardinality of largest 50 communities for collection containing medium and large networks (sets $A \cup B$). Number of networks: 39, bold/underlined correlations are statistically significant for non-directional/directional test.

	Louvain	Label Prop.	SCCD
#nodes	**0.686**	0.077	0.255
#edges	0.31	0.29	**0.502**
In AC	-0.255	**-0.316**	**-0.359**
Out AC	-0.058	0.171	0.17
AC	-0.19	**-0.376**	**-0.385**
Ccoef	-0.042	0.224	**0.325**
tri	-0.243	-0.253	-0.207
diam	-0.029	-0.165	-0.188

Table 6: Pearson correlation of network properties and cardinality of largest 50 communities for collection containing very large networks (set $A \cup B \cup C$). Number of networks: 45, bold/underlined correlations are statistically significant for non-directional/directional test.

	Louvain	Label Prop.	SCCD
#nodes	**0.812**	**0.872**	**0.746**
#edges	**0.915**	**0.924**	**0.776**
In AC	-0.063	-0.104	-0.231
Out AC	-0.036	-0.066	0.032
AC	-0.088	-0.097	-0.257
Ccoef	-0.12	-0.168	0.025
tri	0.008	-0.052	-0.112
diam	-0.102	-0.097	-0.181

relations using also very large networks are reported in Table 6. The only remaining significant correlation is with the number of network elements. For the SCCD algorithm, correlation with degree assortativity remains on the level of directional significance.

To summarise, we have observed a strong correlation of the sizes of largest clusters and the number of the network elements (nodes/ edges), which indicates the bias of the studied algorithms to increase the size of largest components with increasing size of the network. in addition, a (negative) correlation with degree assortativity coefficient has been observed for data samples containing medium and large networks. The interpretation is that the community detection algorithms were able to detect smaller, compact communities for the assortative networks.

5. ASSORTATIVITY OF THE COMMUNITY STRUCTURE

As shown in previous section, the network assortativity coefficient is correlated with the sizes of detected cluster. We thus examine the assortativity of the networks with ground-truth communities more closely. Namely, we compute the assortativity of the whole networks and assortativity of the community structure; i.e. for the latter, we use only edges belonging to ground-truth communities, while keeping the original node degree for the computation of the assortativity coefficient. The results are reported in Table 7, showing assortativity coefficient of the whole network, assortativity computed using only top 5000 clusters and assortativity computed using all of ground-truth clusters. The results show that the assortativity of the community structure can be very different compared to the assortativity of the original network, e.g., Youtube and DBPedia are disassortative, while having assortative community structure. Five of the analysed networks have positive assortativity coefficient of the ground-truth communities, four of which are strongly assortative. The interpretation is that the networks with assortative community structure have important parts of communities composed of edges connecting nodes with similar degrees.

Table 7: Assortativity of the networks with ground-truth communities and the assortativity of their community structures.

	net AC	Top5k Comm. AC	All Comm. AC
DBLP	0.267	0.436	0.446
AMAZON	-0.059	-0.077	-0.026
YOUTUBE	-0.037	0.067	0.068
LJ	0.045	0.464	0.365
ORKUT	0.016	0.233	0.326
DBPEDIA	-0.018	0.958	0.973

6. EDGE WEIGHTING

In the Section 3, we have studied the similarity of the partitions yielded by the community detection algorithms and the ground-truth communities. The results reveal that for DBpedia and Orkut networks the similarity scores are very low. In the previous section (Section 5), we have shown that those two networks have an assortative community structure, which means that the communities they comprise are composed of an important portion of edges connecting nodes with similar degrees. Based on these observations, we examine the possibility of modifying the network structure by weighting its edges, in a way to lower the weight of the edges connecting disassrtative nodes. The hypothesis is that, for networks with assortative community structure, such a modification could affect positively the similarity of the detected clusters with ground-truth clusters. We first propose two edge weighting functions designed to lower the influence of the edges connecting low and high degree nodes. We then study the effects of such weighting on the network, namely we expect the overall network assortativity coefficient to change. Finally, we study the clusters detected on the weighted versions of the networks by the community detection algorithms; we compute their similarity to the ground-truth communities and compare it to the results obtained from original, unweighted networks.

6.1 Weighting Heuristic

So far, we have analysed the networks without weights on edges, which is equivalent to the situation where weights are equal on all the edges. In the following, we propose heuristic weighting functions, designed to decrease the importance of edges connecting nodes with very different degrees, which could lead to an increase of the network's assortativity. Instead of the vertex degree, we will compute the assortativity of the sum of weights of edges adjacent to a vertex; we will refer to it as weighted degree assortativity. It is a necessary change to capture the effect of the weight on edges. For the most of the networks with ground-truth communities, we have no information other than the network structure (that is, we do not have any attributes on nodes or edges that could be used to derive nodes similarity). We thus have to base our weighting functions purely on the network structure.

The assortativity measures the tendency to link similar nodes and the similarity in our case is the sum of edge weights. We thus try to use weighting functions that would penalize the links between highly disassotative vertices, i.e., vertices with very different degrees. The underlying idea is to decrease the importance of edges connecting disassortative nodes. We propose two weighting

functions that differ in the degree of penalization of the edges linking disassortative nodes. Our first experimental weighting function would be 10^{1-x}, where x is the number of digits in a decimal notation of the division of nodes degrees.

More formally, let $d(v)$ denote the degree of a vertex v. Let

$$f_1(z, y) = floor(log_{10}(max(z, y)/min(z, y))$$

The weighting function $w(e_{i,j})$ is:

$$w1(e_{i,j}) = 10^{-f_1(d(i),d(j))}$$

In practice, this weighting would assign weight of 1 to an edge connecting nodes with the degrees of a same magnitude, weight of 10^{-1} to an edge connecting a node with degree in one order of magnitude lower that the other's node degree, and so on.

The second weighting function assigns a weight of 10^{1-x}, where x is the number of digits in a decimal notation of the difference of nodes degrees. Let

$$f_2(z, y) = floor(log_{10}(|z - y|)$$

The weighting function $w(e_{i,j})$ is:

$$w2(e_{i,j}) = 10^{-f_2(d(i),d(j))}$$

This weighting assigns weights as follows: edge linking nodes with $|d(i) - d(j)| < 10$ would be assigned with weight 1, edge $e_{i,j}$ with $|d(i) - d(j)| < 100$ would be assigned with weight of 0.1, and so on. The difference between $w1$ and $w2$ can be significant, e.g., let us compute weights for an edge connecting nodes with the degrees of 50 and 600; the $w1$ would give the weight of 0.1 while the weighting $w2$ would produce the number 0.01.

6.2 Assortativity of the Weighted Networks

We have applied the weighting functions and built a weighted version for all the networks we have used in the experiments. First, we look at the changes in the networks assortativity coefficient. Note that instead of degree assortativity, we use the weighted degree assoratativity. The assortativity coefficient has increased for most of the networks (for $w1$: 38 nets, for $w2$: 33 nets). 24 unweighted networks of the data set have been disassortative, after reweighting of edges 15 (for $w1$; 16 for $w2$) of them have change the polarity of the assortativity, switching from disassortative (AC<0) to assortative (AC>0).

Next, we have re-run the community detection algorithms on the re-weighted networks (all of the used algorithms are able to work with weighted networks). As the weighting has changed the assortativity, our goal was to verify whether the correlation between assortativity and largest clusters cardinality still holds. We have used the original unweighted as well as weighted networks. We have observed that after the addition of the weighted networks to the data sets, the results stay consistent with those reported in Section 4.

6.3 Detected Community Structure of Weighted Networks

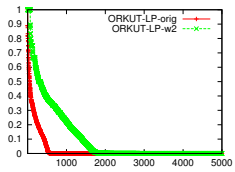
The next step is to analyse effects of the weighting on the results of community detection method compared to the ground-truth communities. We compare the similarity score $comsim$ (we omit *NMI* due to the space constraints) achieved by the very same algorithms on the original unweighted networks with the scores achieved on the weighted versions of the same networks.

The results are summarized in Table 8. For each network, we list the similarity scores of the ground-truth communities with the partitions detected by community detection methods on: original unweighted network (column *Orig*), on network re-weighted with function $w1$ and on network re-weighted with function $w2$. The columns *Imp1* and *Imp2* denotes improvement of the score on $w1$ (or $w2$) compared to the original solution $Orig$ (i.e., $Imp1 = w1/Orig$). We provide results of the comparison for the sets of top 5000 ground-truth communities. We highlight by bold font the best similarity scores for all the networks and we underline the important improvements in similarity score achieved by weighting. The first observation is that there are networks (DBLP, Amazon and LiveJournal and Youtube data sets) for which the community detection methods achieve high similarity scores with the ground-truth communities even without weighting. For those networks, the weighting caused small increases in similarity scores (in several case the score has been marginally decreased). The best performing algorithm for this class has been the Louvain method. The second class, containing the Orkut and DBPedia data sets, was quite challenging for the community detection techniques and the similarity score were very low on unweighted networks. After the re-weighting, the detected communities were significantly better approximations of the ground-truth communities. For the SCCD and Louvain methods on top 5000 datasets, similarity score increased 2 to 14 times. For the Label Propagation, the increase is even more dramatic (e.g., factor of 116 for DBPedia data set). The reason for such a dramatic increase in similarity is that the algorithm often collapses the majority of the node's network into a single community. The weighing of the edges was a successful strategy to prevent this behaviour. We can conclude that for the networks on which the the algorithms were achieving marginal similarity scores, the weighting had a very positive effect on the similarity of the detected communities with ground-truth. Figure 1 depicts similarity scores for distinct communities in the ground-truth data sets with the best fitting clusters in the detected network partitions for the Orkut and DBPedia data sets. The x-axis is the rank of a ground-truth communities, y-axis depicts the Jaccard similarity of the compared sets. The figure visualizes how the weighting improved similarity of the detected clusters with the ground-truth.

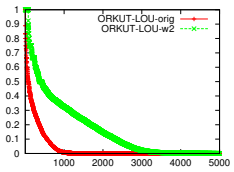
The goal of the work has been to verify whether weighting of the edges can help in better approximating the clusters in networks

Table 8: $Comsim$ similarity scores of ground-truth communities with the detected communities on: original unweighted network (*Orig*), on network re-weighted with *w1* , on network re-weighted with *w2*. *Imp1 = w1 / Orig, Imp2 = w2 / Orig*.

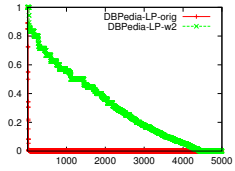
Network	Alg.	Orig	w1	w2	Imp1	Imp2
DBLP	Lou	0.54	0.54	**0.56**	1.00	1.04
	LP	0.49	0.49	0.53	1.00	1.07
	Sccd	0.52	0.51	0.54	0.99	1.05
Amazon	Lou	0.76	0.84	0.83	1.11	1.10
	LP	0.85	0.85	0.83	0.99	0.98
	Sccd	**0.86**	0.86	0.83	1.00	0.97
Youtube	Lou	0.23	0.23	**0.31**	1.12	1.36
	LP	0.08	0.14	0.24	1.78	3.04
	Sccd	0.19	0.21	0.26	1.11	1.42
LJ	Lou	0.52	0.49	0.52	0.94	0.99
	LP	0.57	0.59	0.62	1.04	1.08
	Sccd	0.6	0.61	**0.62**	1.01	1.03
Orkut	Lou	0.04	0.04	0.17	1.14	4.44
	LP	0.03	0.03	0.11	1.18	4.18
	Sccd	0.24	0.20	**0.25**	0.82	1.03
DBPedia	Lou	0.004	0.016	0.053	4.3	14.5
	LP	0.003	0.054	0.31	20.3	116.7
	Sccd	0.13	0.26	**0.34**	1.97	2.60



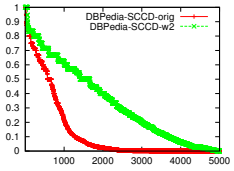
(a) Label Propagation on ORKUT



(b) Lovain method on ORKUT



(c) Label Propagation on DBPEDIA



(d) SCCD on DBPEDIA

Figure 1: Similarity of the community detection algorithms results with Top 5000 ground-truth communities. The x-axis is the rank of the ground-truth community, y-axis depicts the Jaccard similarity coefficient with the best fitting cluster.

with assortiative community structure. An important observation is that both proposed weighting functions caused significant increase in similarity scores for the partitions detected by the studied algorithms for the DBPedia and Orkut networks. For those two networks, the proposed topology preprocessing allowed to increase the similarity of the detected communities from marginal values. Interesting improvements can be observed also for Youtube. DBLP, Amazon and LJ networks received high similarity scores for the partitions detected from original unweighted networks. After the weighting the scores for those networks stayed at approximately the same values. For the Amazon network, which has slightly disassortative community structure, the results after the weighting even marginally decreased. The weighting function $w2$ is more aggressive in penalizing the links between disassortative nodes and it has shown to be causing higher scores than $w1$.

7. CONCLUSION

In this work, we have studied the behaviour of the community detection algorithms with near-linear time complexity on real-world networks. We have observed that for several networks the studied algorithms fail to approximate the ground-truth communities well. By studying network properties and the community structure unveiled by the algorithms, we have shown statistically significant correlation of the network assortativity coefficient and the sizes of the largest detected communities. In addition, we have shown that the assortativity of the community structure is independent of the overall network assortativity. Two networks for which the community detection algorithms failed to deliver good partitioning have assortative community structure. We have proposed weighting functions designed to decrease the disassortativity of the connected nodes and we have studied the effect of such weighting on the networks. The empirical observation is that for the class of networks with assortative community structure, the weighting of the edges can result in significant improvements in the similarity of detected cluster. In several cases, improvements in the similarity score of an order of magnitude have been observed.

8. REFERENCES

[1] J. W. Berry, B. Hendrickson, R. A. LaViolette, and C. A. Phillips. Tolerating the community detection resolution limit with edge weighting. *Phys. Rev. E*, 83:056119, May 2011.

[2] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10).

[3] M. Ciglan and K. Nørvåg. Fast detection of size-constrained communities in large networks. In *Proceedings of WISE'2010*, pages 91–104, 2010.

[4] M. Ciglan, K. Nørvåg, and L. Hluchý. The SemSets model for ad-hoc semantic list search. In *WWW'2012*, pages 131–140, 2012.

[5] L. Danon, J. Duch, A. Diaz-Guilera, and A. Arenas. Comparing community structure identification. *Journal of Statistical Mechanics: Theory and Experiment*, Oct 2005.

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

[7] S. Gregory. A fast algorithm to find overlapping communities in networks. In *Proceedings of PKDD'2008*, pages 408–423. Springer, September 2008.

[8] A. Khadivi, A. Ajdari Rad, and M. Hasler. Network community-detection enhancement by proper weighting. *Phys. Rev. E*, 83:046104, Apr 2011.

[9] A. Lancichinetti and S. Fortunato. Benchmarks for testing comm. detection algorithms on directed and weighted graphs with overlapping communities. *Phys. Rev. E*, 80(1), 2009.

[10] A. Lancichinetti and S. Fortunato. Community detection algorithms: A comparative analysis. *Phys. Rev. E*, 80(5):056117, Nov 2009.

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

[12] J. Leskovec, K. J. Lang, and M. Mahoney. Empirical comparison of algorithms for network community detection. In *Proceedings of WWW'2010*, pages 631–640, 2010.

[13] M. E. J. Newman. Assortative mixing in networks. *Physical Review Letters*, 89(20):208701, 2002.

[14] M. E. J. Newman and M. Girvan. Finding and evaluating community structure in networks. *Phys. Rev. E*, 69(2):026113, Feb 2004.

[15] S. Papadimitriou, J. Sun, C. Faloutsos, and P. S. Yu. Hierarchical, parameter-free community discovery. In *Proceedings of PKDD'2008*, pages 170–187, 2008.

[16] U. N. Raghavan, R. Albert, and S. Kumara. Near linear time algorithm to detect community structures in large-scale networks. *Phys. Rev. E*, 76(3):036106, Sep 2007.

[17] J. Yang and J. Leskovec. Defining and evaluating network communities based on ground-truth. Proceedings of ICDM'2012, pages 745–754, 2012.

[18] T. Yang, R. Jin, Y. Chi, and S. Zhu. Combining link and content for community detection: a discriminative approach. In *Proceedings of KDD'2009*, pages 927–936, 2009.

[19] Y. Zhang, J. Wang, Y. Wang, and L. Zhou. Parallel community detection on large networks with propinquity dynamics. In *Proceedings of KDD'2009*, pages 997–1006, 2009.