A Novel Attributed Community Detection by Integration of Feature Weighting and Node Centrality

Mehrdad Rostami (✉ Mehrdad.Rostami@oulu.fi)

University of Kurdistan <https://orcid.org/0000-0001-5710-217X>

Mourad Oussalah

Oulun Yliopisto

Research

Keywords: Social Network Analysis, Community Detection, Attributed Social Network, Attributed Graph Clustering, Feature Weighting, Node Centrality

Posted Date: July 27th, 2021

DOI: <https://doi.org/10.21203/rs.3.rs-707376/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

A Novel Attributed Community Detection by Integration of Feature Weighting and Node Centrality

Mehrdad Rostami^{1,*}, Mourad Oussalah¹

¹ Centre of Machine Vision and Signal Processing, Faculty of Information Technology,
University of Oulu, Oulu, Finland¹

* Corresponding author: Mehrdad Rostami, E-mail: Mehrdad.Rostami@oulu.fi

Abstract

Community detection is one of the basic problems in social network analysis. Community detection on an attributed social networks aims to discover communities that have not only adhesive structure but also homogeneous node properties. Although community detection has been extensively studied, attributed community detection of large social networks with a large number of attributes remains a vital challenge. To address this challenge, a novel attributed community detection method through an integration of feature weighting with node centrality techniques is developed in this paper. The developed method includes two main phases: (1) Weight Matrix Calculation, (2) Label Propagation Algorithm-based Attributed Community Detection. The aim of the first phase is to calculate the weight between two linked nodes using structural and attribute similarities, while, in the second phase, an improved label propagation algorithm-based community detection method in attributed social network is proposed. The purpose of the second phase is to detect different communities by employing the calculated weight matrix and node popularity. After implementing the proposed method, its performance is compared with several other state of the art methods using some benchmarked real-world datasets. The results indicate that the developed method outperforms several other state of the art methods and ascertain the effectiveness of the developed method for attributed community detection.

Keywords: Social Network Analysis, Community Detection, Attributed Social Network, Attributed Graph Clustering, Feature Weighting, Node Centrality

1. Introduction

Community detection is an essential social network analysis method developed for the aim of detecting communities of nodes that are similar to each other with respect to some similarity criteria [1-3]. Especially, community detection is a vital network analysis task that has been applied to many real applications such as social networks analysis from Facebook, Twitter or blogosphere, information networks in web or IT infrastructures, molecular complex network from biomedical data as in protein-to-protein interaction networks [4-6]. Typical community detection algorithms attempt to identify clusters of nodes that have high intra community similarity and low inter community similarity i.e., the nodes inside the community are similar and nodes in outer communities are dissimilar [7-10].

Several types of community detection methods can be distinguished. The traditional stream of methods focuses solely on the structural dimension of the social networks, i.e., the relationship between network nodes and ignore nodes' attributes [11-13]. A such approach uses nodes' structural similarity as a basis for social network analysis and community detection [14-16]. However, most of real-world social networks present more information about social network nodes than just the existing relationships between them. For instance, in several social network graphs, it is rather common that certain nodes' attributes such as age, gender, interests, are available as part of a-priori knowledge [17-19]. For instance, one may find in Facebook network graph, both information concerning the communication between the different users and the profile of these users, which may include gender, age and location. In such cases, the network is called *attributed social network* in which attributes are assigned to each node. In an attributed social network, the network structure determines the first dimension, while attributes ascertain the second dimension of the social network [20-22]. Another stream of traditional community detection algorithms (compared to structural methods) relies only on the node attributes to discover network communities and entirely relinquish the relationship between the nodes of social networks. It is clear that methods that consider only network structure or only node attributes to detect a community do not employ all of the existing knowledge in a attributed social network. As a result, if a method can simultaneously use both structure and vector of attributes in detecting communities, this weakness can be defeated and the accuracy of community detection will be improved [23-26].

It is expected that an ideal community should not only be well linked in terms of its node structure but should also present similar distribution of its node attribute values. This motivates the development of a new challenging stream of methods referred to “attributed community detection methods”. This research area is receiving more and more attention due to the fact that the common use of network structure and node attributes increases the information about social network nodes and makes the combination of network and attribute information very challenging. In this case, similarities between nodes can be calculated using two criteria, including structural and attribute similarity. The structural and attribute similarities can be computed using network topology and nodes' vector attributes, respectively. The network topology structure demonstrates the relationship between nodes, while the node attribute demonstrates the common properties between these nodes. These similarities play vital roles in the organization of the network communities as in [12, 27, 28]. Fig. 1 presents an example of a small co-authorship network, where each scholar is associated with two attributes consisting of “research area” and location “country”. When a traditional (structural) community detection method is applied to this attributed social network, the generated communities would only have a cohesive structure, but their attributes may have random values. In case of an attributed community detection scheme, one shall consider both the topological structure of the network as well as the attributes of its nodes. The objective is to partition the social network into structurally cohesive and attribute homogeneous communities. Fig. 1 illustrates two such communities. In the left community, the attribute "research area" shows a common value of "DM" (which stands for data mining), but the attribute "country" takes random values. In contrast, for the right community, the author labels correspond mainly to "US" but their research interests cover a wide range of topics. Intuitively, the left community is likely to include a closely working international team in data mining, which explains the spread in location attribute, while the right community would suggest an internal collaboration within a large lab in the US. Therefore, attributed community detection algorithm would shed light on the different communities that can be inferred while respecting both the node structure and attributes such detected communities have both high structural cohesiveness and value homogeneity in many attributes.

Fig.1. Example of attributed social network [29]

Although a lot of attributed community detection approaches have been proposed in the recent years, they still suffer from several drawbacks that are summarized below:

- Most of these community detection algorithms rather relinquish node attributes and only focused on the network structural knowledge despite the claim to handle both aspects simultaneously.
- Many of the previous attributed community detection methods utilize all the attributes to calculate the similarity between two nodes, while nodes in some attributed social networks are described through a large-number of features, some of which might be redundant or irrelevant. The presence of a large number of these irrelevant and redundant features in the attributed social network has a negative effect on the performance of the network analysis algorithm and also increases the computational complexity.
- In most of the previous attributed community detection methods, the weight of all attributes is considered equal, while some attributes can be considered more important than others.
- One of the most widely used methods for community detection is Label Propagation Algorithm (LPA). The LPA algorithm, initially devotes a unique label for each node

and then chooses the node with the highest frequency in different updated iteration. As a result, this algorithm faces instability and low performance due to the progress of giant communities resulting from the equal importance of nodes and random behavior in the updating iteration.

To defeat the aforementioned disadvantages of the previously attributed community detection approaches, in this study, an improved LPA algorithm is devised. The proposed method attempts to detect efficient communities at an acceptable time. In this proposed method, a combination of structural and attributed similarity is put forward to detect the final communities. Moreover, in this developed algorithm, utilizing the node centrality measure, the importance of the nodes in label updating process is employed. The proposed community detection method has several novelties compared to the standard LPA and other previous attributed community detection methods, summarized into the following:

- An improved LPA algorithm that combines structural and attributed similarity through a new similarity measure in a way to reduce iteration time and keeps the original time efficiency is proposed.
- A new method for attribute weighting and attribute feature selection is proposed. The proposal is shown to, on one hand, to increase the quality of the similarity measure in terms of accommodating of both structural and attribute information, and, on the other hand, to reduce the burden of complexity by eliminating irrelevant and redundant attributes.
- An improved criterion of node importance in social networks is developed, that can be employed as a communities heads. Utilizing the node centrality criterion to calculate the importance of attributes and applying this criterion in propagating label is shown to improve the community detection performance.
- A comprehensive testing using some publicly available dataset with a comparison of some state-of-the-art community detection methods have been conducted and whose results demonstrated the technical soundness and practical usefulness of the proposal.

In the remainder of this study in Section 2 previous work in the area of community detection is reviewed. Section 3 details the proposed attributed community detection method. Section 4 reports the experimental results on different real-world attributed social network datasets. Finally, Section 5 presents the conclusion and some perspective work.

2. Related Works

Community detection has been a challenging research field in the social network analysis, machine learning and graph mining where several extensive reviews on the community detection methods have been produced [11, 30-33]. In this section, some of the popular community detection methods is reviewed.

In terms of the type of social network, previous community detection approaches can be grouped into two categories: Non-Attributed and Attributed community detection approaches [34]. In the Non-Attributed community detection, only the linkage network structures are considered, while the node attributes are fully ignored. This category of community detection can be divided into four main classes consisting of Hierarchical community detection, Modularity algorithms, Random Walk models, and Label propagation models [35-37]. On the other hand, Attributed graph methods focus on both structural and attribute knowledge. This category also be partitioned into four popular groups consisting of Edge Weighting, Augmented network, Fitness Function Optimization, and Unified Distance [38].

Hierarchical community detection is a useful and an old algorithm for discovering groups in social network analysis [2, 39, 40], which starts by defining a similarity criterion. Hierarchical community detection-based model can be grouped into two categories: agglomerative-based model and divisive-based model. In random walk-based models, each vertex includes an initial walker's state. Next, walkers select the current vertex neighborhood at random for localization. Modularity-based models use modularity criterion to identify clusters. These community detection algorithms assume a high modularity quantity for the detected community. The aim of modularity models is to identify clusters of graphs to optimize modularity performance [41, 42].

The before-mentioned community detection methods consider only one factor of the network structure but ignore node attributes information.

Attributed community detection methods employ information from both network structural and attributes to detect communities in social network [43], where the detected communities are clusters of densely linked vertices and also very similar to each other in their attributes. Many community detection methods is developed to employ attribute knowledge in addition to the social network structure. These previous attributed community detection methods are

grouped into four classes: (1) Converting an attributed social network to a weighted network, (2) distance models, (3) model-based methods, and (4) subspace models.

The first group try to convert the original attributed network to a weighted graph. In these methods, The weighted graph of the network is constructed based on the attribute information of nodes. The distance-based group consists of distance models and the network structural knowledge is remained in a vertices similarity measure, that integrated with the attribute similarity measure. The model-based group are utilized probabilistic model try to avoid the design of an unrealistic distance criterion. The subspace-based models detect the communities only on the subspace-based of their related properties as a subset of all nodes' attributes, particularly for large social network. In Fig. 2, a categorization of the previously mentioned community detection algorithms is highlighted where one distinguishes the Attributed community detection algorithms and the Non-attributed community detection algorithms we well their various ramifications.

Fig.2. Category of different community detection methods

In overall, we noticed that the detecting communities with previously proposed attributed community detection methods is too slow to be effective in big data scenarios, in aspect of computational complexity and aspect of execution time. Indeed, They utilize data structures that must be entirely reconstructed if the input data modifies; and, sometimes, the number of social network clusters must be determined by the user. Furthermore, most previous attributed community detection methods require categorical attributes, these methods force the user to discretize non-categorical attribute ranges, resulting in a loss of information in similarity criteria. As a result, there is a need for novel efficient community detection algorithma that

consider both structural and attributed information with the capability to detect communities in a reasonable time.

3. The proposed Attributed Community Detection Method

In this section, we detail our innovative attributed community detection method by integrating the concept of Label Propagation Algorithm with Feature Selection (LPAFS). The LPAFS method is cast into the class of attributed community detection methods and considers importance of nodes attributes in its community detection process. LPAFS includes two main phases: (1) Weight Matrix Calculation, (2) Label Propagation Algorithm-based Attributed Community Detection. The aim of the first phase is to calculate the weight between two nodes that have edges with each other using structural and attribute similarities. For this purpose, two techniques of attribute selection and attribute weighting were used as will be detailed later. In the next phase, we utilize the previously calculated weight matrix and node popularity to detect the different communities. The flowchart of the developed attributed community detection method is shown in Fig.3. Also, the details of these two phases are explained in the reminder of this section.

Fig.3. Flowchart of the proposed attributed community detection method

3.1 Weight Matrix Calculation

Although several community detection approaches has been developed, the community detection of real social network with a large number attributes, remains challenging [13]. In this regard, similarity of nodes can be calculated using two criteria: structural similarity and attributed similarity. The structural similarity is calculated considering network structure, while the attribute similarity is evaluated by utilizing the internal properties of nodes that are completely autonomous of the network structure.

More formally, let $G = (V, E, A)$ be the attributed social network, where V is the set of vertices, E the set of links denoting the existing relationshipb between the vertices, and A shows the set of attribute vectors of each node. $n=|V|$ is the number of nodes, $m=|E|$ is the number of edges, and $A(attr_1, attr_2, attr_3 \dots attr_k)$ indicates to each node in V , its corresponding k characteristic (attributes), where k stands for the dimension of node attributes.

Nodes in attributed social networks are usually described by a large number of attributes. Many of these attributes may be irrelevant to the given network analytic application, which may impact the efficiency and computational complexity. Therefore, reducing the dimensions of network attributes is a fundamental task in the applications of social network analysis.

One distinguishes two main ways to reduce the size of dimension: feature extraction and feature selection [44, 45]. In feature extraction, the initial space of features is mapped to a smaller space [46-49] where a small number of features are generated by combining the initial features in a way to preserve the information in the original inputs [50-53]. On the other hand, in feature selection, a subset of the primary attributes is selected according to some predefined criterion in a way to improve prediction and performance [54-57].

In the first step of this phase, the weight of an individual attribute is calculated using Laplacian Score (LS) feature selection method [58]. Laplacian Score is filter-base based feature weighting method that. The purpose of LS is to measure the importance of features based on their ability to preserve locality. The Laplacian Score for Attribute i is calculated using Equation (1):

$$LS(i) = \frac{\sum_{j,k} (A_j(i) - A_k(i)) S_{jk}}{\sum_j (A_j(i) - \bar{A}(i)) D_{jj}} \quad (1)$$

where, $A_j(i)$ denotes the value of the attribute i in the j -th node, $\bar{A}(i)$ determines the average of the attribute i in all nodes, D is a diagonal matrix such that $D_{jj} = \sum_k S_{jk}$, and S_{jk} shows the neighborhood relationship between different nodes, calculated as Equation (2):

$$S_{ij} = \begin{cases} e^{-\frac{JD(j,k)}{t}}, & \text{if } V_j \text{ and } V_k \text{ are neighbors} \\ 0, & \text{Otherwise} \end{cases} \quad (2)$$

where, t is a suitable constant, V_j represents j -th node. V_j and V_k are neighbors if there is an edge between them and $JD(j, k)$ denotes the Jaccard Distance between two nodes of j and k calculated as follows:

$$JD(j, k) = \frac{|\Gamma(j) \cup \Gamma(k)|}{|\Gamma(j) \cap \Gamma(k)|} \quad (3)$$

where $\Gamma(j)$ corresponds to the first-order neighborhood of node j .

After calculating the importance value for all attributes, the weight of the edges between different nodes in the social network should be determined based on node attributes and network structure. In the proposed method, a convex combination of structural similarity and attributed similarity is utilized to calculate the weight between nodes. The weight of the edge between the nodes i and j is calculated based on the following equation:

$$W_{ij} = \alpha Ssim(i, j) + (1 - \alpha) Asim(i, j) \quad (4)$$

where, $Ssim_{ij}$ and $Asim_{ij}$ indicate structural similarity and attributed similarity between node i and j nodes, respectively. α is a parameter that controls the balance between structural and attribute components.

In this paper, the structural similarity between the two nodes is calculated using the concept of Network Motifs [59]. For this purpose, network neighbors and triple censuses were used to calculate the Network Motifs-based similarity method. The proposed similarity measurement was calculated using the node conflict distribution in network patterns. Each pair of network nodes and each of their common neighbors create a triple network motif. Therefore, each pair of network nodes can be a member of many three network motifs. The effect of these three network patterns is the main difference between this method of calculating similarity and other methods of neighborhood similarity. Network Motifs-like similarity is defined by the following equation:

$$Ssim(i, j) = \frac{\sum_{k \in \Gamma(i) \cap \Gamma(j)} \phi(i, j, k) \times \frac{1}{13}}{|\Gamma(i) \cap \Gamma(j)|} \quad (5)$$

where, $\Gamma(i)$ represents the neighbors of nodes i and $\phi(i, j, k)$ represents the number of Network Motifs including nodes i, j and k .

On the other hand, the attributed similarity criterion is calculated using the proposed attribute similarity method. One of the drawbacks in the previous methods is that the weight of all the attributes is constant in calculating the similarity. While in many real networks the value of attributes will not be equal. To solve this limitation, this paper presents an improved similarity criterion called weighted attributed similarity. This criterion is calculated as follows:

$$Asim(i, j) = \frac{Num_N}{Num} \times Asim_N(i, j) + \frac{Num_C}{Num} \times Asim_C(i, j) \quad (6)$$

where Num , Num_N and Num_C represent the total number of attributes, the total number of numerical attributes and the total number of non-numerical attributes, respectively. Moreover $Asim_N(i, j)$, $Asim_C(i, j)$ indicate the similarity of nodes i and j based on numerical attributes and non-numerical attributes, respectively, modelled using (7) and (8):

$$Asim_N(i, j) = \sum_{l=1}^{Num_N} w_l \times \left(\exp\left(-\frac{V_i^N(l) - V_j^N(l)}{2}\right) \right) \quad (7)$$

$$Asim_C(i, j) = \sum_{l=1}^{Num_C} w_l \times \left(Sim(V_i^C(l), V_j^C(l)) \right) \quad (8)$$

where, w_l indicates the weight of attribute l calculated using Laplacian Score. Moreover $Sim(V_i^C(l), V_j^C(l))$ is calculated using Equation (9).

$$Sim(V_i^C(l), V_j^C(l)) = \begin{cases} 1, & \text{if } V_i^C(l) = V_j^C(l) \\ 0, & \text{if } V_i^C(l) \neq V_j^C(l) \end{cases} \quad (9)$$

where $V_i^N(l)$ and $V_i^C(l)$ denote the l -th numerical attributes of node V_i and l -th non-numerical attributes of node V_i

3.2 Label Propagation Algorithm-based Attributed Community Detection

By considering an attributed social network $G = (V, E, A)$ and the number of clusters K , the proposed community detection method in this paper is to group the initial social network vertex-set into K disjoint communities $V_1, V_2, \dots, V_n$, where $V = \bigcup_{i=1}^n v_i$ and $V_i \cap V_j = \emptyset$ for any $i \neq j$, so that: (1) communities contain intensively interconnected nodes while different communities contain sparsely connected nodes; and (2) the nodes of the same community have low levels of diversity in attribute values, while nodes belonging to separate communities may exhibit disparate attribute values.

The standard Label Propagation community detection, which utilizes only social network structure to detect communities, first, defines a unique label to each node and next chooses the most frequent node in different steps of update process. All nodes under the dense sub community which attained the similar label can be grouped as community if the algorithm reaches an iteration in which there is a maximum number of adjacent tags [60].

Due to the assumption of similar influence of nodes, random process in the updating step and tiebreaker mode, the standard LPA faces instability and low performance. In each community, there are more distinguished nodes that make a major importance (the center of community), therefore, social network nodes located near the center of community compared to border nodes, more influential nodes. Typically, the nodes in the community have different influence scores and the nodes with higher importance play the role of the dominant nodes, and the nodes with lower importance is in charge of non-domination. All the nodes in the community are not of similar importance, and a given node either affects other nodes (dominant) or is affected by another node (non-dominance).

As mentioned, in the first phase, the graph $G = (V, E, A)$ was converted into the graph $G = (V, E, W)$ with structure-attributes similarity. In this phase, the popularity of each node in social network is calculated using Laplacian Node Centrality. The nodes with higher popularity in the weighted social network will be penetrable on their neighbor in terms of structure and attributes. Those nodes that occupy a central position in the community and have a significant number of links with other nodes of the community make a remarkable portion in leading, guiding, and constructing ability and consistency in the group, while the members of the community that are on the border of clusters may lead and guide a mediator duty between different clusters. Calculating the measure of centrality or identifying more "central"

and Influential nodes has been a major impact in the analysis of social networks [61, 62]. In this work, the Laplacian Centrality (LC) [62] is employed to calculate the centrality of nodes. Formally, for the generated weighted matrix of graph G , X is defined as below:

$$X(G) = \begin{pmatrix} X_1 & 0 & \dots & 0 \\ 0 & X_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & X_n \end{pmatrix} \quad (10)$$

Where $X_i = \sum_{j=1}^n w_{i,j} = \sum_{u \in N(v_i)} w_{v_i,u}$ and X_i is titled the sum weight of nodes v_i and $N(v_i)$ is the set of v_i 's neighbors. Also, the Laplacian Energy of G is calculated as below:

$$E_L(G) = \sum_{i=1}^n X_i + 2 \sum_{i < j} w_{i,j}^2 \quad (11)$$

Finally, the Laplacian Centrality $C_L(v_i, Graph)$ of node v_i is calculated using the following formula:

$$NC_L(v_i, Graph) = \frac{(\Delta E)_i}{E_L(Graph)} = \frac{E_L(Graph) - E_L(Graph_i)}{E_L(Graph)} \quad (12)$$

Where $Graph_i$ is the graph acquired using deleting v_i from graph $Graph$.

After calculating the Laplacian Centrality of each node, each node's Label Influence (LI) is computed as follows:

$$LI(i, l) = NC_L(v_i, G) \quad (13)$$

where, $LI(i, l)$ denotes the influence of the label l on the node i , that $NC_L(v_i, G)$ is a laplacian centrality of a node i . Unlike standard Label Propagation Algorithm where all neighbors have the equal probability to propagate a label, in the proposed attributed and improved Label Propagation Algorithm method, Label influence of neighbors is employed to choose the optimum label for propagation. Label Acceptance (LA) is calculated as below:

$$LA(u) = \left[\arg \max_{v \in \Gamma(u)} LI(v, l) \right] \quad (14)$$

The pseudocode of the developed attributed community detection is shown in Algorithm 1.

Algorithm 1: The developed Attributed Community Detection Method**Input:** Social Network $G = (V, E)$ **Output:** Detected Communities $C = \{C_1, C_2, \dots, C_k\}$

1. Calculating Weighing Matrix using Eq. (4)
2. Allocate a unique label to initial social network node
3. Compute the label influence using Eq. (13)
4. **While** the label of nodes change or $i < \text{Threshold}$ **do**
 5. Vector X is constructed by ordering nodes based on their strength.
 6. Set $i = 1$.
 7. **For** each node in X update its labels acceptance using Eq. (14)
 8. In case of a tie-break, calculate the sun label of a neighbor node and choose the sun label with the highest importance.
 9. $i = i + 1$.
10. **End while**
11. Generate communities according to equal labels.
12. **Return** Detected Communities.

3.3 Computational Complexity

The complexity of the developed algorithm is computed in this subsection. Suppose that a graph $G = (V, E)$ with $n = |V|$ nodes and $m = |E|$ links and let k be the average node degree. Our method described in Algorithm 1 contains several repetitions and each iteration independently runs with itself compitional complexity. First phase, the complexity of the computation to initialize all nodes is indicated by unique labels with $O(n)$ complicity. Then, in the next phase, label influence is calculated. For this purpose, Weight Matrix should be computed. The complexity of this phase is $O(nk)$, which can be approximated by $O(m)$ because in a such large-scale social network, it holds $k \ll n$, and node centrality is calculated as $O(m)$. In the next phase, we perform the sorting the nodes using their influences. This yields a complexity of $O(n)$. The next pahse performs the label propagation procedure. The conoutiional complexity of this procedure is calculated with $O(nk)$ complexity that equivalent to $O(m)$. In the final phase, each nodes are assigned to their communities with complexity of $O(n)$. Generally, the final complexity of developed method is $O(nk + m + 2n) \approx O(m)$. As a result, since real social networks tend to have sparse edges, the nodes will be approximately equal to edges, therefore, $O(m) \approx O(n)$. In other words, the time complexity

of the proposed method is reduced to linear complexity of the number nodes of the network, which corresponds to a significant improvement over alternative implementation schemes.

4. Experimental Result

To measure the efficiency of the developed community detection approach, various experiments are designed. The performance of the developed approach is compared with four new methods of attributed community detection: Adapt-SA-Soft [63], Adapt-SA-PCA [63], and Subspace Stochastic Block Model (SSBM) [64]. The detail of these methods is listed in Table 1.

In these experiments the developed approach and other compared community detection methods were implemented using Python language programming on an Intel Core-i7 CPU with 8GB of RAM. The results are acquired over ten separate and autonomous runs to attain more precise and acceptable assessments.

Table 1. Characteristics of comparative methods

Method	Description	Year
Adapt-SA-Soft	Adapt-SA-soft is a weighted K-means method with local learning and an additional step to the fuzzy K-means for attributed community detection.	2017
Adapt-SA-PCA	Adapt-SA-PCA is a weighted K-means method with local learning and an PCA dimensionality reduction phase for attributed community detection.	2017
SSBM	Subspace stochastic block model (SSBM) to detect the community structures in attributed social network. The main purpose of this method is to consider both topological structure and attribute information as the latent factors to analysis the organization of communities.	2020

In the reminder of this section, the details of the used datasets, evaluation criteria, proposed method parameter and experimental results, are described.

4.1. Used Dataset

In this study, several datasets with different properties are used in the experiments to show the effectiveness and robustness of the developed attributed community detection approach.

These attributed social network real-world datasets consist of Citeseer, Cora, Cornell, Texas, Washington, Wisconsin and Political Blogs.

Cora [65] contains a set of vertex demonstrating scientific publications where an edge between two vertex is a citation from a publication to another. The attributes of this dataset is demonstrated using a set of unique words.

Citeseer [65] is another citations network where each vertex belongs to one groups of agents, artificial intelligence, databases, human–computer interaction, information retrieval, and machine learning.

WebKB [65] citations network includes scientific publications, which belong to Web page networks of four universities: 1) **Cornell**; 2) **Texas**; 3) **Washington**, and 4) **Wisconsin**. Each page network is classified into five classes: course, faculty, student, project, and staff.

The main specifications of these attributed social network datasets are detailed in Table 2. These social network datasets have been selected according to various properties such as number of attributes, number of nodes, number of edges, and number of communities. Because datasets with large attributes are one of the most challenging aspects of attributed community detection, more attention has been paid to the number of attributes to choose the dataset.

Table 2. Characteristics of the used social network datasets.

Dataset	Nodes	Edges	Attributes	Community
Cora	2708	5429	1433	5
Citeseer	1787	3285	3703	6
Cornell	195	304	1703	5
Texas	187	328	1703	5
Washington	230	446	1703	5
Wisconsin	265	530	1703	5

4.2. Evaluation Criteria

Two major categories of criteria; namely, external and internal criteria, are utilized to measure the efficiency of our method as well as alternative different methods. In the reminder of this subsection, these measures will be detailed.

The normalized Mutual Information $NMI(A, B)$ of two community partitions A and B of a network is calculated as follows. Assume C a confusion matrix where an element C_{ij} is the number of nodes of community i of the partition A that are also in the community j of the partition B .

$$NMI(A, B) = \frac{-2 \sum_{i=1}^{C_A} \sum_{j=1}^{C_B} \log(nC_{ij}/C_i C_j)}{\sum_{i=1}^{C_A} C_i \cdot \log(C_i/n) + \sum_{j=1}^{C_B} C_j \cdot \log(C_j/n)} \quad (15)$$

where C_A (resp. C_B) is the number of groups in the partition A (resp. B), C_i (resp. C_j) is the sum of the elements of C in row i (resp. column j), and n is the total number of nodes in the network. Especially, it holds that if $A = B$, then $NMI(A, B) = 1$. Similarly, if A and B are completely dissimilar, $NMI(A, B) = 0$.

Accuracy is another popular statistical criterion that refers to the correlation of the evaluations to a determined value. In the community detection, a high accuracy value indicates a high correlation between the detected communities and the real communities in the social network. This criterion shows the ratio of the number of nodes whose community has been correctly detected to the total number of nodes.

The F-measure integrates the precision and recall concepts from an information retrieval perspective. The recall and precision evaluations are calculated as follows:

$$Precision = \frac{|S \cap T|}{|S|} \quad (16)$$

$$Recall = \frac{|S \cap T|}{|T|} \quad (17)$$

where S is the set of node pairs that are assigned to the same community and T is the set of node pairs that have the same label.

F1-Score integrates the precision and recall concepts into a single evaluation as follows:

$$F1_Score = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (18)$$

4.3. Proposed Method Parameters

Unlike many community detection methods that have many parameters, the proposed method involves only one user-defined parameters. The α parameter in Equation (4) determines the importance of structural similarity with respect to node attribute information, where a higher value indicates a willingness to favor network structure over node attributes.

This parameter is important for the proposed community detection method because it directly controls the behavior of the learning model and has a considerable impact on the efficiency of the final community detection. To optimally choose this parameter, it is necessary to iteratively set the α value and detect the final communities and then evaluate the final performance to set the best value. Based on the Trial and Error strategy, the α parameter is set to 0.6 that the results of this experiments are shown on 4.5 (Sensitivity analysis) subsection.

4.4. Experimental Results

In this subsection, to evaluate the efficiency of the proposed method, the results of our community detection method are compared with previous attributed community detection methods that employ both network structural knowledge and node attributes. The community detection results are demonstrated in Table 3-5. To evaluate the community detection methods, NMI, Accuracy, and F1-score criteria are employed.

Table 3 records the average NMI value of ten autonomous runs. The best average values of accuracy are marked in boldface. The reported results of Table 3 show that in all cases the proposed method outperforms other community detection methods. It also reveals that the average accuracy of the developed approach in all social network datasets was 0.375, which is 0.026 higher than the average NMI for the second-ranked method (i.e. SSBM).

Table 3. Average NMI and standard deviation (shown in parenthesis) of different attributed community detection methods.

Dataset	Adapt-SA-Soft	Adapt-SA-PCA	SSBM	Proposed Method
Cora	0.441 (0.012)	0.206 (0.002)	0.462 (0.012)	0.479 (0.013)
Citeseer	0.293 (0.003)	0.231 (0.012)	0.381 (0.021)	0.398 (0.017)
Cornell	0.134 (0.031)	0.122 (0.012)	0.212 (0.002)	0.249 (0.006)
Texas	0.221 (0.061)	0.136 (0.058)	0.318 (0.007)	0.376 (0.012)
Washington	0.262 (0.046)	0.216 (0.013)	0.361 (0.004)	0.381 (0.008)
Wisconsin	0.324 (0.021)	0.109 (0.015)	0.359 (0.009)	0.364 (0.002)

Moreover, Table 4 shows that in all social network datasets except Wisconsin, the developed approach had the highest accuracy and ranked first among all compared attributed community methods. Also, in the Wisconsin dataset, the accuracy of the developed approach ranked second with a difference of 0.001 compared to the SSBM community detection method.

Table 4. Average Accuracy and standard deviation (shown in parenthesis) of different attributed community detection methods.

Dataset	Adapt-SA-Soft	Adapt-SA-PCA	SSBM	Proposed Method
Cora	0.583 (0.008)	0.385 (0.005)	0.632 (0.001)	0.634 (0.003)
Citeseer	0.542 (0.003)	0.489 (0.011)	0.611 (0.003)	0.620 (0.001)
Cornell	0.451 (0.015)	0.443 (0.002)	0.559 (0.007)	0.562 (0.002)
Texas	0.558 (0.038)	0.551 (0.012)	0.598 (0.002)	0.601 (0.007)
Washington	0.573 (0.019)	0.502 (0.008)	0.658 (0.006)	0.682 (0.005)
Wisconsin	0.552 (0.034)	0.508 (0.028)	0.513 (0.001)	0.512 (0.003)

Tables 5-7 show the average Precision, Recall and F1-Score for the different community detection methods, respectively. Based on the exhibited numerical results, it can be concluded that in all social network datasets, the proposed method outperformed other attributed

community detection methods. For example, for the Texas dataset, the proposed method obtained 0.631, F1-Score while for Adapt-SA-Soft, Adapt-SA-PCA, and SSBM, this metric scores 0.491, 0.539, and 0.620, respectively.

Table 5. Average Precision and standard deviation (shown in parenthesis) of different attributed community detection methods.

Dataset	Adapt-SA-Soft	Adapt-SA-PCA	SSBM	Proposed Method
Cora	0.427 (0.012)	0.282 (0.008)	0.492 (0.002)	0.503 (0.003)
Citeseer	0.398 (0.015)	0.341 (0.019)	0.494 (0.012)	0.512 (0.001)
Cornell	0.372 (0.010)	0.369 (0.012)	0.442 (0.021)	0.452 (0.002)
Texas	0.499 (0.013)	0.531 (0.010)	0.631 (0.002)	0.681 (0.007)
Washington	0.512 (0.023)	0.539 (0.011)	0.543 (0.002)	0.569 (0.001)
Wisconsin	0.481 (0.018)	0.451 (0.019)	0.501 (0.012)	0.512 (0.004)

Table 6. Average Recall and standard deviation (shown in parenthesis) of different attributed community detection methods.

Dataset	Adapt-SA-Soft	Adapt-SA-PCA	SSBM	Proposed Method
Cora	0.440 (0.015)	0.291 (0.001)	0.498 (0.002)	0.513 (0.012)
Citeseer	0.397 (0.012)	0.349 (0.010)	0.492 (0.001)	0.502 (0.006)
Cornell	0.378 (0.016)	0.362 (0.012)	0.441 (0.006)	0.450 (0.010)
Texas	0.491 (0.012)	0.539 (0.012)	0.620 (0.003)	0.631 (0.012)
Washington	0.519 (0.021)	0.536 (0.013)	0.552 (0.008)	0.562 (0.002)
Wisconsin	0.486 (0.019)	0.457 (0.018)	0.504 (0.001)	0.519 (0.007)

Table 7. Average F1-Score and standard deviation (shown in parenthesis) of different attributed community detection methods.

Dataset	Adapt-SA-Soft	Adapt-SA-PCA	SSBM	Proposed Method
Cora	0.433 (0.008)	0.286 (0.003)	0.495 (0.008)	0.508 (0.009)
Citeseer	0.397 (0.001)	0.345 (0.008)	0.493 (0.018)	0.507 (0.007)
Cornell	0.375 (0.012)	0.365 (0.019)	0.441 (0.026)	0.451 (0.011)
Texas	0.495 (0.018)	0.535 (0.006)	0.625 (0.007)	0.655 (0.004)
Washington	0.515 (0.019)	0.537 (0.004)	0.547 (0.013)	0.565 (0.021)
Wisconsin	0.483 (0.027)	0.454 (0.025)	0.502 (0.024)	0.515 (0.013)

Moreover, in Fig. 4-8, we show the average value over all the social network datasets for the NMI, Accuracy, Precision, Recall and F1-Score, respectively. As it can be seen from these figures, for these five measures, the proposed method had the highest performance. The results of Fig. 4 show that the developed community detection method achieves 0.375 average NMI and ranks first with a margin of 0.026 compared to the SSBM algorithm, which obtained the second-best average NMI measure. Furthermore, from results in Fig. 5, it can be observed that the differences between the obtained Accuracy of the developed attributed community detection technique and the second-best ones (SSBM) and third-best ones (Adapt-SA-Soft) are evaluated to 0.007 (i.e., $0.602 - 0.595$) and 0.059 (i.e., $0.602 - 0.543$), respectively. Also, considering reported result of Fig. 8 and for the F1-Score criterion, the proposed community detection approach achieved the first rank with an average F1-Score of 0.534, and the SSBM and Adapt-SA-Soft community detection approaches were ranked second and third with an average F1-Score of 0.517 and 0.450, corespondingly.

Fig. 4. Average NMI on different datasets for all community detection approaches

Fig. 5. Average Accuracy on different datasets for all community detection approaches

Fig. 6. Average Precision on different datasets for all community detection approaches

Fig. 7. Average Recall on different datasets for all community detection approaches

Fig. 8. Average F1-Score on different datasets for all community detection approaches

4.5. Sensitivity analysis

In this subsection, we investigate the sensitivity analysis with respect to parameter α in the proposed method. Fig. 9 shows the NMI performance metric when varying the α parameter from 0.2 to 0.8 in the different dataset. The result shows that setting the α parameter to 0.6 yields optimal NMI value. The choice of the parameter values α between 0.2 and 0.8 is motivated by the desire to avoid extreme cases where the tradeoff between attributed similarity and structural similarity scores is not exhibited.

Fig. 9. Average NMI with different α values

4.6. Statistical analysis

In this subsection, the nonparametric Friedman test [66] is applied to assess the statistical significance of our findings. In other words, we use Friedman test to compare the efficiency of the different community detection approaches on different social network data. For this purpose, each community detection approach is ranked on each social network dataset the SPSS statistical package was used. Table 8 reports the results of the Friedman test for these evaluated community detection methods. The results of this Table indicated a p-value of 0.000939, 0.001817, 0.001072, 0.000939 and 0.000939 in NMI, Accuracy, Precision, Recall, F1-Score criteria, respectively. Since these values are lower than 0.05, it can be asserted that the performance of the developed attributed community detection approach is considerably better than those of the other compared methods.

Table 8. The results of the Friedman statistics test.

	Criterion				
	NMI	Accuracy	Precision	Recall	F1-Score
Chi-Square	16.400	15.000	16.119	16.400	16.400
df	3	3	3	3	3
Asymp.Sig (p-value)	0.000939	0.001817	0.001072	0.000939	0.000939

4.7. Discussion

The developed approach has at least three major innovations that have made it perform better than other methods:

1- The majority of previous community detection approaches mostly eliminate node attributes information and only look at the network structure despite the claim to handle both aspects of the network at the same time. In this study, through a novel similarity measure, an improved attributed community detection method that integrates structural and attributed similarity is developed and successfully tested.

2- Irrelevant attributes, as well as redundant attributes, strongly affect the efficiency of the learning model and the result of attributed community detection method. Therefore, attributed community detection method should recognize and weigh original attributes of nodes and

ignore the irrelevant and redundant attributes as far as possible. Most previous attributed community detection algorithms, due to the neglectance of attribute weighting phase fail to ignore the irrelevant and redundant attributes accurately. To accommodate this objective, in this study, an efficient attributed community detection method that can efficiently and effectively ignore irrelevant attributes is developed.

3- In popular community detection algorithms such as LBA, the node centrality is ignored and all nodes are assumed to be equally important, while it is acknowledged that such hypothesis does not hold in many real dataset as those considered in our study. Therefore, our approach was devised to account for such discrepancy and applied to the label updating process.

5. Conclusion and future works

Community detection is one of the most important research topics in graph mining and social network analysis. Community detection includes clustering the nodes of an input social network into several communities satisfying some specified measures. In the last decade, it has been widely studied how to detect communities in attributed social network. Although a large number of community detection methods have been developed, most of these methods are not appropriate for social networks with large number of attributes due to the low efficiency, high computational complexity, and not being parameter-free. To defeat these weaknesses, in this paper, a novel attributed community detection method is developed by combination of feature weighting and node centrality techniques. After implementation of the developed method, its performance is compared with those from different community detection methods in attributed social networks using popular six real-world datasets. The results denote that the proposed method outperforms other previous methods, ascertain the effectiveness of the proposed method for attributed community detection problem and lay down foundations for the emergence of new social network based analytics that supports large-scale community discovery taking into multiple attributes information.

Due to the fact that this article developed a new measure for calculating similarity in attributed social networks, this measure can be utilized in other fields of social network analysis, such as link prediction, Information diffusion, and node reputation. For example, in future research, this new similarity calculation measure can be employed to predict links in protein-protein interaction networks.

Declaration

• **Ethics approval and consent to participate:** Not applicable

• **Consent for publication:** Not applicable

• **Availability of data and materials:** The dataset used in this study can be obtained from the corresponding author on reasonable request.

• **Competing interests:** The authors declare that they have no competing interests.

• **Funding:** This work is supported by the Academy of Finland Profi5 Project No. 3261291 on DigiHealth.

Authors' contributions

The specific contributions made by each author is as follows:

Mehrdad Rostami: Conceptualization, Methodology, Validation, Implementation, Writing-Original Draft, Writing - Review & Editing

Mourad Oussalah: Conceptualization, Methodology, Writing-Original Draft, Writing - Review & Editing

Abbreviations:

LPA: Label Propagation Algorithm

LPAFS: Label Propagation Algorithm with Feature Selection

LS: Laplacian Score

LC: Laplacian Centrality

LI: Label Influence

LA: Label Acceptance

SSBM: Subspace Stochastic Block Model

NMI: Normalized Mutual Information

Acknowledgements:

This work is supported by the Academy of Finland Profi5 Project No. 3261291 on DigiHealth, which gratefully acknowledged

References

- [1] W. Zhang, R. Zhang, R. Shang, J. Li, and L. Jiao, "Application of natural computation inspired method in community detection," *Physica A: Statistical Mechanics and its Applications*, vol. 515, pp. 130-150, 2019/02/01/ 2019.
- [2] C. Li, J. Bai, Z. Wenjun, and Y. Xihao, "Community detection using hierarchical clustering based on edge-weighted similarity in cloud environment," *Information Processing & Management*, vol. 56, no. 1, pp. 91-109, 2019/01/01/ 2019.
- [3] Z.-H. Deng, H.-H. Qiao, M.-Y. Gao, Q. Song, and L. Gao, "Complex network community detection method by improved density peaks model," *Physica A: Statistical Mechanics and its Applications*, vol. 526, p. 121070, 2019/07/15/ 2019.
- [4] X. Li, S. Zhou, J. Liu, G. Lian, G. Chen, and C.-W. Lin, "Communities detection in social network based on local edge centrality," *Physica A: Statistical Mechanics and its Applications*, p. 121552, 2019/05/24/ 2019.
- [5] X. Li, S. Zhou, J. Liu, G. Lian, G. Chen, and C.-W. Lin, "Communities detection in social network based on local edge centrality," *Physica A: Statistical Mechanics and its Applications*, vol. 531, p. 121552, 2019/10/01/ 2019.
- [6] X. Zhao, J. Liang, and J. Wang, "A community detection algorithm based on graph compression for large-scale social networks," *Information Sciences*, vol. 551, pp. 358-372, 2021/04/01/ 2021.
- [7] H. Roghani, A. Bouyer, and E. Nourani, "PLDLS: A novel parallel label diffusion and label Selection-based community detection algorithm based on Spark in social networks," *Expert Systems with Applications*, vol. 183, p. 115377, 2021/11/30/ 2021.
- [8] V. Moscato and G. Sperli, "A survey about community detection over On-line Social and Heterogeneous Information Networks," *Knowledge-Based Systems*, vol. 224, p. 107112, 2021/07/19/ 2021.
- [9] Y. Chen and J. W. Baker, "Community detection in spatial correlation graphs: Application to non-stationary ground motion modeling," *Computers & Geosciences*, vol. 154, p. 104779, 2021/09/01/ 2021.
- [10] H. Ma, Z. Liu, X. Zhang, L. Zhang, and H. Jiang, "Balancing topology structure and node attribute in evolutionary multi-objective community detection for attributed networks," *Knowledge-Based Systems*, vol. 227, p. 107169, 2021/09/05/ 2021.
- [11] C. Bothorel, J. D. Cruz, M. Magnani, and B. MicenkovÁ, "Clustering attributed graphs: Models, measures and methods," *Network Science*, vol. 3, no. 3, pp. 408-444, 2015.
- [12] X. Huang, H. Cheng, and J. X. Yu, "Dense community detection in multi-valued attributed networks," *Information Sciences*, vol. 314, pp. 77-99, 9/1/ 2015.
- [13] K. Berahmand, S. Haghani, M. Rostami, and Y. Li, "A new Attributed Graph Clustering by using Label Propagation in Complex Networks," *Journal of King Saud University - Computer and Information Sciences*, 2020/09/01/ 2020.
- [14] Z. Liu and M. Barahona, "Graph-based data clustering via multiscale community detection," *Applied Network Science*, vol. 5, no. 1, p. 3, 2020/01/08 2020.
- [15] K. M. Yazdi *et al.*, "Improving Recommender Systems Accuracy in Social Networks Using Popularity," in *2019 20th International Conference on Parallel and Distributed Computing, Applications and Technologies (PDCAT)*, 2019, pp. 301-307.
- [16] K. Majbouri Yazdi *et al.*, "Prediction optimization of diffusion paths in social networks using integration of ant colony and densest subgraph algorithms," *Journal of High Speed Networks*, vol. 26, pp. 141-153, 2020.
- [17] P. Wu and L. Pan, "Mining application-aware community organization with expanded feature subspaces from concerned attributes in social networks," *Knowledge-Based Systems*, vol. 139, pp. 1-12, 2018/01/01/ 2018.

- [18] A. Khan, L. Golab, M. Kargar, J. Szlichta, and M. Zihayat, "Compact group discovery in attributed graphs and social networks," *Information Processing & Management*, vol. 57, no. 2, p. 102054, 2020/03/01/ 2020.
- [19] P. Chunaev, T. Gradov, and K. Bochenina, "Community detection in node-attributed social networks: How structure-attributes correlation affects clustering quality," *Procedia Computer Science*, vol. 178, pp. 355-364, 2020/01/01/ 2020.
- [20] A. Reihanian, M.-R. Feizi-Derakhshi, and H. S. Aghdasi, "Community detection in social networks with node attributes based on multi-objective biogeography based optimization," *Engineering Applications of Artificial Intelligence*, vol. 62, pp. 51-67, 2017/06/01/ 2017.
- [21] Y. Ding, S. Yan, Y. Zhang, W. Dai, and L. Dong, "Predicting the attributes of social network users using a graph-based machine learning method," *Computer Communications*, vol. 73, pp. 3-11, 2016/01/01/ 2016.
- [22] R. Fei, J. Sha, Q. Xu, B. Hu, K. Wang, and S. Li, "A new deep sparse autoencoder for community detection in complex networks," *EURASIP Journal on Wireless Communications and Networking*, vol. 2020, no. 1, p. 91, 2020/05/11 2020.
- [23] Y. Gao, H. Zhang, and Y. Zhang, "Overlapping community detection based on conductance optimization in large-scale networks," *Physica A: Statistical Mechanics and its Applications*, vol. 522, pp. 69-79, 2019/05/15/ 2019.
- [24] N. Kasoro, S. Kasereka, E. Mayogha, H. T. Vinh, and J. Kinganga, "PercoMCV: A hybrid approach of community detection in social networks," *Procedia Computer Science*, vol. 151, pp. 45-52, 2019/01/01/ 2019.
- [25] J. Zhang, X. Ding, and J. Yang, "Revealing the role of node similarity and community merging in community detection," *Knowledge-Based Systems*, vol. 165, pp. 407-419, 2019/02/01/ 2019.
- [26] E. A. Corrêa and D. R. Amancio, "Word sense induction using word embeddings and community detection in complex networks," *Physica A: Statistical Mechanics and its Applications*, vol. 523, pp. 180-190, 2019/06/01/ 2019.
- [27] W. Dhifli, N. E. I. Karabadi, and M. Elati, "Evolutionary mining of skyline clusters of attributed graph data," *Information Sciences*, vol. 509, pp. 501-514, 2020/01/01/ 2020.
- [28] A. Baroni, A. Conte, M. Patrignani, and S. Ruggieri, "Efficiently clustering very large attributed graphs," in *2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, 2017, pp. 369-376: IEEE.
- [29] Y. Yang, H. Liu, Z. Guan, X. He, and G. Liu, "CoHomo: A cluster-attribute correlation aware graph clustering framework," *Neurocomputing*, vol. 412, pp. 327-338, 2020/10/28/ 2020.
- [30] P. Chunaev, "Community detection in node-attributed social networks: A survey," *Computer Science Review*, vol. 37, p. 100286, 2020/08/01/ 2020.
- [31] F. D. Malliaros and M. Vazirgiannis, "Clustering and community detection in directed networks: A survey," *Physics Reports*, vol. 533, no. 4, pp. 95-142, 2013/12/30/ 2013.
- [32] Y. Zhao, "A survey on theoretical advances of community detection in networks," *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 9, no. 5, p. e1403, 2017.
- [33] S. Harenberg *et al.*, "Community detection in large-scale networks: a survey and empirical evaluation," *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 6, no. 6, pp. 426-439, 2014.
- [34] E. Akbas and P. Zhao, "Attributed graph clustering: An attribute-aware graph embedding approach," in *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017*, 2017, pp. 305-308.
- [35] Y. Gandica, A. Decuyper, C. Cloquet, I. Thomas, and J.-C. Delvenne, "Measuring the effect of node aggregation on community detection," *EPJ Data Science*, vol. 9, no. 1, p. 6, 2020/03/11 2020.
- [36] V. d. F. Vieira, C. R. Xavier, and A. G. Evsukoff, "A comparative study of overlapping community detection methods from the perspective of the structural properties," *Applied Network Science*, vol. 5, no. 1, p. 51, 2020/08/20 2020.

- [37] C. Durán, A. Muscoloni, and C. V. Cannistraci, "Geometrical inspired pre-weighting enhances Markov clustering community detection in complex networks," *Applied Network Science*, vol. 6, no. 1, p. 29, 2021/04/09 2021.
- [38] M. P. Boobalan, D. Lopez, and X. Z. Gao, "Graph clustering using k-Neighbourhood Attribute Structural similarity," *Applied soft computing*, vol. 47, pp. 216-223, 2016.
- [39] L. Zhou, K. Lü, P. Yang, L. Wang, and B. Kong, "An approach for overlapping and hierarchical community detection in social networks based on coalition formation game theory," *Expert Systems with Applications*, vol. 42, no. 24, pp. 9634-9646, 2015/12/30/ 2015.
- [40] B. Yang, J. Di, J. Liu, and D. Liu, "Hierarchical community detection with applications to real-world network analysis," *Data & Knowledge Engineering*, vol. 83, pp. 20-38, 2013/01/01/ 2013.
- [41] I. Mitrai and P. Daoutidis, "Decomposition of integrated scheduling and dynamic optimization problems using community detection," *Journal of Process Control*, vol. 90, pp. 63-74, 2020/06/01/ 2020.
- [42] M. Zhang and Z. Zhou, "Structural Deep Nonnegative Matrix Factorization for community detection," *Applied Soft Computing*, vol. 97, p. 106846, 2020/12/01/ 2020.
- [43] G. P. Guedes, E. Ogasawara, E. Bezerra, and G. Xexeo, "Discovering top-k non-redundant clusterings in attributed graphs," *Neurocomputing*, vol. 210, pp. 45-54, 2016.
- [44] M. Rostami, K. Berahmand, and S. Forouzandeh, "A novel community detection based genetic algorithm for feature selection," *Journal of Big Data*, vol. 8, no. 1, p. 2, 2021/01/04 2021.
- [45] M. Rostami, K. Berahmand, E. Nasiri, and S. Forouzandeh, "Review of swarm intelligence-based feature selection methods," *Engineering Applications of Artificial Intelligence*, vol. 100, p. 104210, 2021/04/01/ 2021.
- [46] S. R. Bandela and T. K. Kumar, "Unsupervised feature selection and NMF de-noising for robust Speech Emotion Recognition," *Applied Acoustics*, vol. 172, p. 107645, 2021/01/15/ 2021.
- [47] N. Maleki, Y. Zeinali, and S. T. A. Niaki, "A k-NN method for lung cancer prognosis with the use of a genetic algorithm for feature selection," *Expert Systems with Applications*, vol. 164, p. 113981, 2021/02/01/ 2021.
- [48] Y. Zhou, W. Zhang, J. Kang, X. Zhang, and X. Wang, "A problem-specific non-dominated sorting genetic algorithm for supervised feature selection," *Information Sciences*, vol. 547, pp. 841-859, 2021/02/08/ 2021.
- [49] F. Amini and G. Hu, "A two-layer feature selection method using Genetic Algorithm and Elastic Net," *Expert Systems with Applications*, vol. 166, p. 114072, 2021/03/15/ 2021.
- [50] M. Rostami, S. Forouzandeh, K. Berahmand, and M. Soltani, "Integration of multi-objective PSO based feature selection and node centrality for medical datasets," *Genomics*, vol. 112, no. 6, pp. 4370-4384, 2020/11/01/ 2020.
- [51] P. Nimbalkar and D. Kshirsagar, "Feature selection for intrusion detection system in Internet-of-Things (IoT)," *ICT Express*, 2021/05/07/ 2021.
- [52] Y. Zhang, D.-w. Gong, X.-z. Gao, T. Tian, and X.-y. Sun, "Binary differential evolution with self-learning for multi-objective feature selection," *Information Sciences*, vol. 507, pp. 67-85, 2020/01/01/ 2020.
- [53] E. Hancer, "A new multi-objective differential evolution approach for simultaneous clustering and feature selection," *Engineering Applications of Artificial Intelligence*, vol. 87, p. 103307, 2020/01/01/ 2020.
- [54] M. P. Rostami, M. "A clustering based genetic algorithm for feature selection," *Information and Knowledge Technology (IKT)*, pp. pp. 112-116 2014.
- [55] P. Moradi and M. Rostami, "A graph theoretic approach for unsupervised feature selection," *Engineering Applications of Artificial Intelligence*, vol. 44, pp. 33-45, 9// 2015.
- [56] P. Moradi and M. Rostami, "Integration of graph clustering with ant colony optimization for feature selection," *Knowledge-Based Systems*, vol. 84, pp. 144-161, 2015/08/01/ 2015.

- [57] M. Rostami, K. Berahmand, and S. Forouzandeh, "A novel method of constrained feature selection by the measurement of pairwise constraints uncertainty," *Journal of Big Data*, vol. 7, no. 1, p. 83, 2020/10/02 2020.
- [58] Xiaofei He, Deng Cai, and P. Niyogi, "Laplacian Score for Feature Selection," *Adv. Neural Inf. Process. Syst.*, vol. 18, pp. 507-514, 2005.
- [59] F. Aghabozorgi and M. R. Khayyambashi, "A new similarity measure for link prediction based on local structures in social networks," *Physica A: Statistical Mechanics and its Applications*, vol. 501, pp. 12-23, 2018/07/01/ 2018.
- [60] U. N. Raghavan, R. Albert, and S. Kumara, "Near linear time algorithm to detect community structures in large-scale networks," *Physical review E*, vol. 76, no. 3, p. 036106, 2007.
- [61] Xiangbin Yan, Li Zhaia, and W. Fan., "C-index: A weighted network node centrality measure for collaboration competence," *Journal of Informetrics*, vol. 7, no. 1, pp. Pages 223–239, 2013.
- [62] Xingqin Qi, Eddie Fuller, Qin Wu, Yezhou Wu, and C.-Q. Zhang., "Laplacian centrality: A new centrality measure for weighted networks," *Information Sciences*, vol. 194, pp. Pages 240–253, 2012.
- [63] Y. Li, C. Jia, X. Kong, L. Yang, and J. Yu, "Locally weighted fusion of structural and attribute information in graph clustering," *IEEE transactions on cybernetics*, vol. 49, no. 1, pp. 247-260, 2017.
- [64] H. Chen, Z. Yu, Q. Yang, and J. Shao, "Attributed graph clustering with subspace stochastic block model," *Information Sciences*, vol. 535, pp. 130-141, 2020/10/01/ 2020.
- [65] P. Sen, G. Namata, M. Bilgic, L. Getoor, B. Galligher, and T. Eliassi-Rad, "Collective classification in network data," *AI magazine*, vol. 29, no. 3, pp. 93-93, 2008.
- [66] M. Friedman, "A comparison of alternative tests of significance for the problem of m rankings," *Annals of Math. Statistics*, vol. 11, no. 1, pp. 86-92, 1940.