Extending Proximity Measures to Attributed Networks for Community Detection

Rinat Aynulin
Moscow Institute of Physics and Technology
9 Institutskii per.
Dolgoprudny, 141700, Russia
and
Kotel’nikov Institute of Radio-engineering and Electronics of the
Russian Academy of Sciences
11 Mokhovaya str.
Moscow, 125009, Russia

Pavel Chebotarev
Trapeznikov Institute of Control Sciences of the
Russian Academy of Sciences
65 Profsoyuznaya str.
Moscow, 117997, Russia

Proximity measures on graphs are extensively used for solving various problems in network analysis, including community detection. Previous studies have considered proximity measures mainly for networks without attributes. However, attribute information, node attributes in particular, allows a more in-depth exploration of the network structure. This paper extends the definition of a number of proximity measures to the case of attributed networks. To take node attributes into account, attribute similarity is embedded into the adjacency matrix. Obtained attribute-aware proximity measures are numerically studied in the context of community detection in real-world networks.

Keywords: attributed networks; community detection; proximity measure; kernel on graph

1. Introduction

Networks can model a wide variety of real-world systems [1] from the fields of computer science, biology, economy, social science and others. Graphs are a natural way to represent networks: nodes represent objects, and edges represent connections between them. In this paper, the terms “graph” and “network” are used interchangeably.

Many real-world networks exhibit community structure, where nodes are divided into groups called communities or clusters. Nodes
in a group are densely connected, while few edges connect nodes from different groups. The problem of extracting such groups is called community detection or clustering. Community detection is an active field, and many methods have been proposed to solve this problem. Some community detection methods use the notion of proximity or distance measure on the set of graph nodes.

A proximity (or distance) measure on a graph is a function that shows proximity (or distance) between pairs of its nodes. One of the simplest measures of this kind is the shortest path distance. However, there are plenty of other measures on the set of graph nodes [2–4]. Many proximity measures are defined with the help of kernels on graphs (i.e., positive semidefinite matrices) [5]. In this paper, the term “proximity measure” is used in a broader sense than in [5], namely, we do not require them to satisfy the triangle inequality for proximities.

In previous studies, kernels on graphs have been considered mainly for networks without attributes. However, nodes in real-world networks are often associated with attributes describing their personal features. Node attributes play an important role in the emergence of communities, so taking them into account can provide a better understanding of the network community structure.

A number of algorithms for community detection in attributed networks have been proposed in recent years. Surveys [6, 7] provide a comprehensive review of existing methods, some of which are discussed in Section 2. However, kernel-based community detection has not yet been considered in the context of attributed networks.

This paper extends the definition of several well-studied proximity measures to the case of networks with node attributes. Node attributes information is embedded into proximity measures using a number of attribute similarity measures. The obtained attribute-aware proximity measures are then applied for community detection in several real-world networks.

The results show that proximity measures based on attribute and structural information generally outperform both plain proximity measures and pure attribute similarity measures in solving community detection tasks. On the other hand, the variation of the ratio of attribute and structural information shows that for some networks, attribute information is enough to detect the community structure.

This paper is an extended version of the paper [8] presented at Complex Networks 2020. Compared to the conference paper, we search for the optimal ratio of the attribute and structural information and provide new numerical results.
2. Related Work

This section consists of two parts. First, we discuss proximity and distance measures on the set of graph nodes and related studies. In the second part, we provide a brief overview of previously proposed methods for community detection in attributed networks.

2.1 Measures on the Set of Graph Nodes

Until the 1960s, the shortest path distance was the only widely used measure on the set of graph nodes [9]. Deza and Deza [2, Chapter 15] review an assortment of measures on the set of graph nodes suggested in recent decades, including the physics-inspired electric metric [10] (also known as resistance distance) and many others.

Avrachenkov et al. [5] analytically and numerically study the properties of several proximity measures, including PageRank, Communicability, Walk, Heat and a number of logarithmic measures. In [11], the authors review nine kernels on graphs and numerically compare them in the context of link prediction and semi-supervised classification.

Sommer et al. [12] compare the Randomized Shortest Path, Free Energy, Sigmoid Commute Time, Corrected Commute Time and Logarithmic Forest measures in application to community detection in 15 datasets. According to the results of experiments, the best-performing measures are Randomized Shortest Path and Free Energy.

In [13, 14], the authors propose to transform proximity measures using simple mathematical functions like the logarithm. Experimental results show that such transformations improve the effectiveness of proximity measures in application to community detection.

2.2 Community Detection in Attributed Networks

Classically, community detection methods used only structural information and ignored node attributes (see, e.g., [15]). Recently, a number of methods that use both types of information have been proposed.

The SA-Cluster algorithm is presented in [16]. This algorithm creates an attribute node for all the values of all the attributes. If a node has an attribute value specified in an attribute node, an edge is created between this node and this attribute node. After the creation of new edges, the algorithm estimates the distance between nodes using a random walk model. The $k$-medoids method is then used to detect communities. Ruan et al. [17] propose the CODICIL method. The idea of this method is as follows: an additional edge (a content edge) is created between nodes with similar attributes. The resulting graph with additional edges is then clustered using the Multi-level Regularized Markov Clustering and Metis algorithms.

https://doi.org/10.25088/ComplexSystems.30.4.441
In [18], the authors propose a weight modification–based method for community detection in attributed networks. For each edge in the network, its weight is changed to the matching coefficient (i.e., the number of common attribute values) between the nodes connected by this edge. After the weight modification, the network is clustered using the Spectral, Karger’s Min-Cut, and MajorClust algorithms. Yang et al. [19] present the CESNA method, which assumes that the attributed network is generated by a probabilistic model. Maximum-likelihood estimation on this model allows detection of communities in the network.

For a thorough survey of various techniques for community detection in attributed networks, we refer to [6, 7]. According to the classification suggested in [7], our method belongs to the class of weight-based methods.

## 3. Preliminaries

### 3.1 Definitions

An undirected weighted attributed graph $G = (V, E, F)$ is considered, where $V (|V| = n)$ is the set of nodes, $E (|E| = m)$ is the set of edges, and $F$ is the tuple of attribute (or feature) vectors $f_i$. $d$ attributes are associated with each of the $n$ nodes, so $F = (f_1, \ldots, f_n)$, where $f_i \in \mathbb{R}^d$. We consider networks with binary attributes.

The adjacency matrix $A = (a_{ij})$ of a graph is a square matrix with $a_{ij}$ equal to the weight of the edge $(i, j)$ if nodes $i$ and $j$ are connected, and equal to zero otherwise. Sometimes the cost matrix $C = (c_{ij})$ can also be defined for the graph $G$ with elements $c_{ij}$ equal to the cost of following this edge. If the cost is not defined independently, it can be taken as the inverse to the weight: $c_{ij} = 1 / a_{ij}$.

The node degree is the sum of the weights of the edges connected to this node. The degree matrix $D = \text{diag}(A \cdot 1)$ is the diagonal matrix that contains degrees of all the nodes on the main diagonal; $1 = (1, \ldots, 1)^T$. Given $A$ and $D$, $L = D - A$ is the Laplacian matrix, and $P = D^{-1}A$ is the Markov matrix.

Any function $\kappa : V(G) \times V(G) \rightarrow \mathbb{R}$ that shows proximity, similarity or distance between the nodes of $G$ will be referred to as a measure on the node set of $G$. A kernel on $G$ is a proximity measure that can be represented by a positive semidefinite matrix $K$ (or Gram matrix), with $k_{ij}$ showing closeness between nodes $i$ and $j$. For more details on measures and kernels on graphs, see [5].
### 3.2 Community Detection Algorithms

Community detection algorithms allow the community structure of a network to be detected. In this paper, we consider two algorithms: \( k \)-means and Spectral.

The \( k \)-means algorithm [20] is used to detect communities based on the node attributes. For community detection based on the network structure and node attributes, we use the version of the Spectral algorithm presented in [21]. This method applies the \( k \)-means algorithm to the eigenvectors of the Laplacian matrix. For a thorough review of the Spectral algorithm, see [22].

### 3.3 Proximity Measures

In this paper, we consider five proximity measures that have proven to be efficient in the context of community detection in previous studies (e.g., [5, 12]):

- Communicability (C) [23, 24]: \( K^C = \sum_{n=0}^{\infty} a^n A^n / n! = \exp(aA), a > 0; \)
- Heat (H) [25]: \( K^H = \sum_{n=0}^{\infty} a^n (-L)^n / n! = \exp(-aL), a > 0; \)
- PageRank (PR) [11, 26]: \( K^{PR} = (I - aP)^{-1}, 0 < a < 1; \)
- Free energy (FE) [27]. Given a Markov matrix \( P \), a cost matrix \( C \) and a parameter \( a \), the matrix \( W \) is defined as \( W = \exp(-aC) \circ P \) (where the “\( \circ \)” symbol is the elementwise multiplication). Then, we can define \( Z = (I - W)^{-1} \) and \( S = (Z(C \circ W)) / Z \) (where the “\( \div \)” symbol is the elementwise division). The distance matrix is then equal to \( \Delta^FE = (\Phi + \Phi^T) / 2 \), where \( \Phi = \log(Z) / a \). Finally, \( K^{FE} \) is calculated from \( \Delta^FE \) using the transformation \( K = -1 / 2(\Delta^H); \)
- Sigmoid corrected commute time (SCCT) [12, 28]. First, the corrected commute time (CCT) kernel is defined as \( K^{CCT} = HD^{-1/2} M(I - M)^{-1} MD^{-1/2} H \), where \( H = I - (1 \cdot 1^T) / n \), \( M = D^{-1/2}(A - (d \cdot d^T) / vol(G))D^{-1/2} \), \( d \) contains the elements on the main diagonal of the degree matrix \( D \), \( vol(G) = \sum_{ij=1}^n a_{ij} \). Finally, the elements of the sigmoid corrected commute-time kernel \( K^{SCCT} \) are defined as \( K_{ij}^{SCCT} = 1 / (1 + \exp(-aK_{ij}^{CCT} / \sigma)), \) where \( \sigma \) stands for the standard deviation of the \( K^{CCT} \) elements, and \( a \) is a positive parameter.

### 3.4 Quality Evaluation of Community Detection

For the community detection quality evaluation, the adjusted Rand index (ARI) is used. ARI was introduced in [29], and [30] presented its advantages over some other quality indices. ARI takes the values ranging from zero to one, and the larger the value, the better.

ARI measures the agreement between two partitions of \( n \) elements. In the case of community detection, one partition is the ground-truth.
community assignment, and the other partition is the community assignment according to a clustering algorithm.

ARI is an enhanced version of the Rand index. Let \( a \) be the number of pairs of elements that are in the same communities in both partitions, and \( b \) the number of pairs of elements that are in different communities in both partitions. Then, the Rand index is calculated as follows:

\[
\frac{a + b}{\binom{n}{2}}.
\]

ARI transforms the Rand index in such a manner that it will have a value close to zero for random partitions:

\[
\text{ARI} = \frac{\text{index} - \text{expected index}}{\text{max index} - \text{expected index}}.
\]

## 4. Proximity-Based Community Detection in Attributed Networks

A proximity measure matrix \( K \) is used as an input to community detection algorithms. In order to take node attributes into account for community detection, we should consider attributes during proximity calculation.

As we saw in Section 3.3, proximity measures are functions of the adjacency matrix. Thus, attributes should be embedded into the adjacency matrix. This can be achieved by modifying edge weights based on the attributes in the following way:

\[
a^s_{ij} = \beta a_{ij} + (1 - \beta)s_{ij},
\]

where \( \beta \in [0, 1] \) and \( s_{ij} = s(f_i, f_j) \) is the similarity between nodes \( i \) and \( j \) calculated using some attribute similarity measure. An attribute similarity measure, as the name suggests, quantifies the level of similarity between the attribute vectors of two nodes.

The coefficient \( \beta \) allows a tradeoff between the impact of the network structure and attributes on the resulting weight. In the case when \( \beta = 0 \), the attributed adjacency matrix \( A^s \) shows only the similarity according to attributes, while with \( \beta = 1 \) it represents only the network structure and coincides with the adjacency matrix. Given \( A^s \), the attributed versions of the preceding proximity measures can be derived and applied for community detection.

Let \( f_i = (f^1_i, \ldots, f^d_i) \) and \( f_j = (f^1_j, \ldots, f^d_j) \) be the attribute vectors of nodes \( i \) and \( j \), respectively. In this paper, we consider five attribute similarity measures:
Matching coefficient (MC) [31]:

$$s_{MC}(f_i, f_j) = \frac{\sum_{k=1}^{d} \mathbb{1}(f_i^k = f_j^k)}{d}.$$  

$\mathbb{1}(x)$ is the indicator function that equals one if the condition $x$ is satisfied, and zero otherwise. The MC is generally used for discrete attributes (especially binary) since equality is rare for continuous attributes:

- Cosine similarity (CS) [32, Chapter 2]:

$$s_{CS}(f_i, f_j) = \frac{f_i \cdot f_j}{\|f_i\|_2\|f_j\|_2};$$

- Extended Jaccard similarity (JS) [32, Chapter 2]:

$$s_{JS}(f_i, f_j) = \frac{f_i \cdot f_j}{\|f_i\|_2^2 + \|f_j\|_2^2 - f_i \cdot f_j};$$

- Manhattan similarity (MS) [33]:

$$s_{MS}(f_i, f_j) = \frac{1}{1 + \|f_i - f_j\|_1};$$

- Euclidean similarity (ES) [33]:

$$s_{ES}(f_i, f_j) = \frac{1}{1 + \|f_i - f_j\|_2}.$$

5. Experimental Methodology

To test the effectiveness of the attribute-aware proximity measures, we perform a number of experiments on real-world datasets. For each of the datasets, communities are detected using multiple methods. First, the $k$-means algorithm is applied, which deals only with node attributes and ignores the network structure. Then, communities are detected using the Spectral algorithm in combination with five proximity measures of Section 3. Finally, each dataset is clustered using the Spectral algorithm in combination with attribute-aware proximity measures. The quality of produced communities is evaluated each time using the ARI.

Since there is a parameter in the definition of each of the proximity measures, we also search for the optimal parameter. In the results, we use the optimal parameters while estimating the community detection quality.
Experiments are performed on the following datasets:

- WebKB [34, 35]: a dataset of computer science department webpages at several universities. Each webpage has a binary feature vector \((d = 1703)\) that shows the presence or absence of words from a special vocabulary. There are five classes that describe types of the webpages. The dataset contains four unweighted graphs that correspond to University of Washington \((n = 230, m = 446)\), University of Wisconsin \((n = 265, m = 530)\), Cornell University \((n = 195, m = 304)\) and University of Texas at Austin \((n = 187, m = 328)\).

- CiteSeer [36]: a citation graph of machine learning papers with 3312 nodes and 4732 edges. Nodes are divided into one of six classes (the topic of the paper). A binary attribute vector \((d = 3703)\), which indicates the presence or absence of words from a dictionary, is associated with each of the nodes.

- Cora [36]: a citation graph of machine learning papers with 2708 nodes and 5429 edges. As in the CiteSeer dataset, nodes are divided into classes according to the topic of the paper (the number of classes is seven) and a binary attribute vector \((d = 1433)\) is associated with each of the nodes.

### 6. Results

In this section, results of the experiments with a balanced version of the attribute-aware proximity measures \((\beta = 1/2\) in equation \((1)\)) are presented.

Figure 1 shows the average rank and standard deviation for the attribute similarity measures combined with the network information in the framework of the Spectral algorithm and \(k\)-means applied to the attribute information only. The rank is averaged over the six graphs discussed earlier. There are five subfigures, one for each of the proximity measures. The horizontal axis represents the average rank and standard deviation, while the vertical axis shows the used attribute similarity measure (or just \(k\)-means if the network structure was ignored). “No” on the horizontal axis stands for the network proximity measure that does not use any attribute information.

As can be seen in the figure, node attributes information improves the quality of community detection for all the proximity measures: the plain proximity measure, “No”, has one of the lowest ranks on each of the subfigures. Not all of the attribute similarity measures have shown a good community detection quality. For example, the Euclidean similarity (ES) and the matching coefficient (MC) have worsened the effectiveness of some of the proximity measures. However, the cosine similarity (CS) and the Jaccard similarity (JS) always improve the results in comparison with the plain proximity measure.
When the cosine similarity or the Jaccard similarity is used, the attribute-aware proximity measures also outperform the \( k \)-means method. Therefore, with a good attribute similarity measure (according to our experiments, CS or JS), the attribute-aware proximity measures perform better than both the \( k \)-means method, which deals only with attributes, and plain proximity measures, which use only the network structure.

Table 1 shows the best-performing pairs of a proximity measure and an attribute similarity measure. The indisputable leader is free energy paired with the cosine similarity measure.

![Figure 1](https://doi.org/10.25088/ComplexSystems.30.4.441)

### Table 1. The best-performing pairs of proximity measure and similarity measure.

<table>
<thead>
<tr>
<th>Number</th>
<th>Proximity Measure</th>
<th>Similarity Measure</th>
<th>Average Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>FE</td>
<td>CS</td>
<td>2.833</td>
</tr>
<tr>
<td>2</td>
<td>FE</td>
<td>JS</td>
<td>6.333</td>
</tr>
<tr>
<td>3</td>
<td>communicability</td>
<td>CS</td>
<td>6.667</td>
</tr>
<tr>
<td>4</td>
<td>SCCT</td>
<td>JS</td>
<td>7.333</td>
</tr>
<tr>
<td>5</td>
<td>SCCT</td>
<td>CS</td>
<td>7.667</td>
</tr>
<tr>
<td>6</td>
<td>communicability</td>
<td>JS</td>
<td>8.333</td>
</tr>
<tr>
<td>7</td>
<td>PR</td>
<td>CS</td>
<td>8.333</td>
</tr>
<tr>
<td>8</td>
<td>heat</td>
<td>CS</td>
<td>8.667</td>
</tr>
</tbody>
</table>

### 7. Varying Attribute-Relation Ratio

The coefficient \( \beta \) in equation (1) denotes the ratio between attribute and relation (graph structure) information in the resulting attributed
adjacency matrix $A^s$. In this section, we analyze the effectiveness of community detection when various values of $\beta$ are used.

For the coefficient search, we consider the most promising proximity and similarity measures according to the results of the previous section. The cosine similarity and extended Jaccard similarity are used as similarity measures, and the free energy and communicability measures are used as proximity measures.

Similar to the previous section, we detect communities in WebKB, CiteSeer and Cora datasets using the Spectral algorithm and attribute-aware proximity measures. We employ the following values of $\beta$: 0, 0.1, 0.2, 0.4, 0.5, 0.6, 0.8, 0.9, 1. With $\beta = 0$, elements of the attributed adjacency matrix $A^s$ involve only similarity by attributes.

Figure 2 presents the results of searching for the optimal coefficient $\beta$ for six attributed networks. Each subfigure contains four plots for various combinations of attribute similarity and proximity measures: cosine similarity and free energy (CS, FE), cosine similarity and communicability (CS, C), extended Jaccard similarity and free energy (JS, FE), and extended Jaccard similarity and communicability (JS, C).

The $x$ axis shows the values of the varying coefficient $\beta$, while the value of ARI is plotted on the $y$ axis.

![Figure 2](image-url)
As can be seen, the optimal value of $\beta$ depends on the dataset as well as the similarity and similarity measures. However, a common pattern can be identified. Even a small amount of attribute information (changing $\beta$ from 1 to 0.9) significantly improves the ARI score. This further underscores the importance of attributes for community detection.

Moreover, for networks in the WebKB dataset, the top or close to top ARI score can be achieved without using graph structure information at all ($\beta = 0$). This means that for these graphs, using hyperlinks between the webpages in the present way does not improve the quality of community detection. Paying more attention to the graph structure starts to sharply reduce quality after some low value of $\beta$.

On the contrary, for CiteSeer and Cora datasets, the graph structure is significant for detecting communities. Only when attributes are not used at all ($\beta = 1$) does the ARI value drastically decrease.

Thus, the optimal $\beta$ strongly depends on the dataset. For some datasets, the graph structure is more important for determining the community structure than attributes. For others, including WebKB, using the structural information in the present way does not assist in community detection.

The latter result is easy to interpret. The types of webpages to be recognized on WebKB networks are: faculty (member), staff, student, research project and course. In contrast to CiteSeer and Cora, where an edge/arc between nodes (papers) reflects semantic similarity, which is typical of papers in the same category, the situation with WebKB is radically different. For instance, a faculty member’s webpage usually links to courses and projects, as well as Ph.D. and graduate students, rather than to other faculties. From a student webpage, typical links are to courses and advisors rather than to other students or staff members, and so on.

A general conclusion might be that in community detection tasks, the semantics of network links is of paramount importance. In some networks, the density of intraclass connections is much higher than that of interclass ones; other networks resemble multipartite graphs whose edges mainly connect different classes; some networks like WebKB have specific link patterns. Taking such patterns into account may improve the accuracy of community detection, while ignoring them can degrade the results. Further research should suggest alternative approaches to combining attribute and topological information to reveal the structure of semantically complex networks, such as those in the WebKB collection.
8. Conclusion

In this study, we extended the definitions of proximity measures such as communicability, heat, PageRank, free energy and sigmoid corrected commute time to attributed networks. Node attributes were embedded into the proximity measures using five attribute similarity measures: the matching coefficient, the cosine similarity, the extended Jaccard similarity, the Manhattan similarity and the Euclidean similarity. The obtained attribute-aware proximity measures were applied for community detection in attributed networks.

The experiments on the real-world datasets showed that node attributes information provides a better understanding of the community structure and improves the quality of community detection. Moreover, for some networks, attribute information is enough for community detection. The best-performing attribute similarity measures in the experiments were the cosine similarity and extended Jaccard similarity.

References


Extending Proximity Measures to Attributed Networks for Community Detection


https://doi.org/10.25088/ComplexSystems.30.4.441


