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

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

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.

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 $\mathbf{f}_i$. d attributes are associated with each of the n nodes, so $F = (\mathbf{f}_1, \ldots, f_n)$, where $\mathbf{f}_i \in \mathbb{R}^d$. 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* $D = \mathrm{diag}(A \cdot \mathbf{1})$ is the diagonal matrix that contains degrees of all the nodes on the main diagonal; $\mathbf{1} = (1, \ldots, 1)^T$. Given A and D, $L = D - A$ is the *Laplacian matrix*, and $P = D^{-1} A$ is the *Markov matrix*.

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

$$a_{ij}^s = \beta a_{ij} + (1-\beta)s_{ij}, \tag{1}$$

where $\beta \in [0, 1]$ and $s_{ij} = s(\mathbf{f}_i, \mathbf{f}_j)$ 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 β 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 $\mathbf{f}_i = (f_i^1, \ldots, f_i^d)$ and $\mathbf{f}_j = (f_j^1, \ldots, f_j^d)$ be the attribute vectors of nodes i and j, respectively. In this paper, we consider five attribute similarity measures:

- Matching coefficient (MC) [31]:

$$s^{\mathrm{MC}}\left(\mathbf{f}_i, \mathbf{f}_j\right) = \frac{\sum_{k=1}^{d} \mathbb{1}\left(f_i^k = f_j^k\right)}{d}.$$

$\mathbb{1}(x)$ 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]:

$$s^{\mathrm{CS}}\left(\mathbf{f}_i, \mathbf{f}_j\right) = \frac{\mathbf{f}_i \cdot \mathbf{f}_j}{\|\mathbf{f}_i\|_2 \|\mathbf{f}_j\|_2};$$

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

$$s^{\mathrm{JS}}\left(\mathbf{f}_i, \mathbf{f}_j\right) = \frac{\mathbf{f}_i \cdot \mathbf{f}_j}{\|\mathbf{f}_i\|_2^2 + \|\mathbf{f}_j\|_2^2 - \mathbf{f}_i \cdot \mathbf{f}_j};$$

- Manhattan similarity (MS) [33]:

$$s^{\mathrm{MS}}\left(\mathbf{f}_i, \mathbf{f}_j\right) = \frac{1}{1 + \|\mathbf{f}_i - \mathbf{f}_j\|_1};$$

- Euclidean similarity (ES) [33]:

$$s^{\mathrm{ES}}\left(\mathbf{f}_i, \mathbf{f}_j\right) = \frac{1}{1 + \|\mathbf{f}_i - \mathbf{f}_j\|_2}.$$

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.

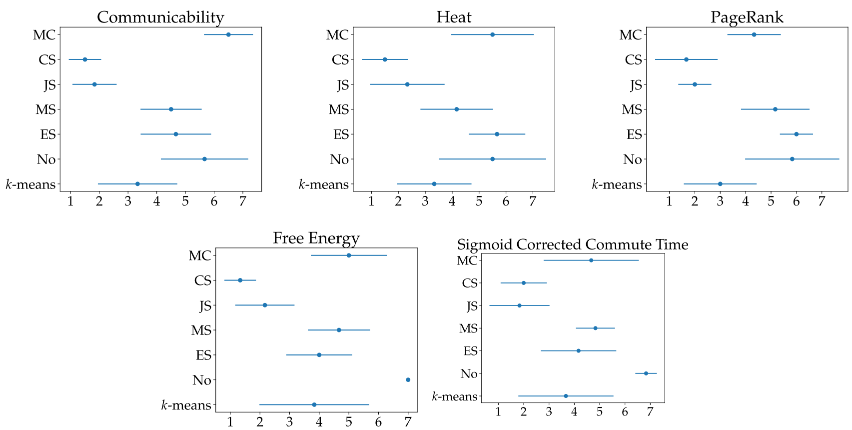

Figure 1. Average rank and standard deviation for attribute similarity measures and k-means.

Nmber	Proximity Measure	Similarity Measure	Average Rank
1	FE	CS	2.833
2	FE	JS	6.333
3	communicability	CS	6.667
4	SCCT	JS	7.333
5	SCCT	CS	7.667
6	communicability	JS	8.333
7	PR	CS	8.333
8	heat	CS	8.667

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

7. Varying Attribute-Relation Ratio

The coefficient β 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 β 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 β: 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 β 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 β, while the value of ARI is plotted on the y axis.

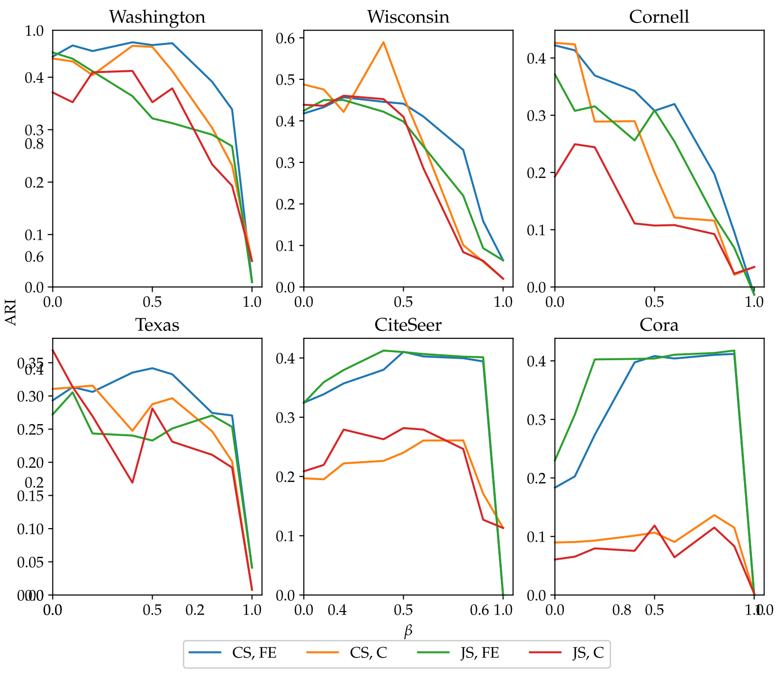

Figure 2. Searching for the optimal β.

As can be seen, the optimal value of β depends on the dataset as well as the proximity and similarity measures. However, a common pattern can be identified. Even a small amount of attribute information (changing β 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 β.

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 β 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

[1] L. da F. Costa, O. N. Oliveira Jr., G. Travieso, F. A. Rodrigues, P. R. V. Boas, L. Antiqueira, M. P. Viana and L. E. C. Rocha, "Analyzing and Modeling Real-World Phenomena with Complex Networks: A Survey of Applications," *Advances in Physics*, **60**(3), 2011 pp. 329–412. doi:10.1080/00018732.2011.572452.

[2] M. M. Deza and E. Deza, *Encyclopedia of Distances*, 4th ed., Heidelberg: Springer, 2016.

[3] P. Yu. Chebotarev and E. V. Shamis, "On Proximity Measures for Graph Vertices," *Automation and Remote Control*, **59**(10), 1998 pp. 1443–1459.

[4] P. Chebotarev, "Studying New Classes of Graph Metrics," in *Proceedings of the SEE Conference Geometric Science of Information (GSI-2013)*, Paris, France (F. Nielsen and F. Barbaresco, eds.), Berlin, Heidelberg: Springer, 2013 pp. 207–214. doi:10.1007/978-3-642-40020-9_21.

[5] K. Avrachenkov, P. Chebotarev and D. Rubanov, "Kernels on Graphs as Proximity Measures," in *Proceedings of the 14th International Workshop on Algorithms and Models for the Web Graph (WAW 2017)*, Toronto, Canada (A. Bonato, F. C. Graham and P. Prałat, eds.), Cham, Switzerland: Springer, 2017 pp. 27–41. doi:10.1007/978-3-319-67810-8_3.

[6] C. Bothorel, J. D. Cruz, M. Magnani and B. Micenkova, "Clustering Attributed Graphs: Models, Measures and Methods," *Network Science*, **3**(3), 2015 pp. 408–444. doi:10.1017/nws.2015.9.

[7] P. Chunaev, "Community Detection in Node-Attributed Social Networks: A Survey," *Computer Science Review*, **37**, 2020 100286. doi:10.1016/j.cosrev.2020.100286.

[8] R. Aynulin and P. Chebotarev, "Measuring Proximity in Attributed Networks for Community Detection," in *Proceedings of the Ninth International Conference on Complex Networks and Their Applications (Complex Networks 2020)*, Madrid, Spain (R. M. Benito, C. Cherifi, H. Cherifi, E. Moro, L. M. Rocha and M. Sales-Pardo, eds.), Cham, Switzerland: Springer, 2021 pp. 27–37. doi:10.1007/978-3-030-65347-7_3.

[9] E. W. Dijkstra, "A Note on Two Problems in Connexion with Graphs," *Numerische Mathematik*, **1**(1), 1959 pp. 269–271. doi:10.1007/BF01386390.

[10] G. Sharpe, "Solution of the (m+1)-Terminal Resistive Network Problem by Means of Metric Geometry," in *Proceedings of the First Asilomar Conference on Circuits and Systems*, Pacific Grove, CA, 1967 pp. 319–328.

[11] F. Fouss, K. Francoisse, L. Yen, A. Pirotte and M. Saerens, "An Experimental Investigation of Kernels on Graphs for Collaborative Recommendation and Semisupervised Classification," *Neural Networks*, **31**, 2012 pp. 53–72. doi:10.1016/j.neunet.2012.03.001.

[12] F. Sommer, F. Fouss and M. Saerens, "Comparison of Graph Node Distances on Clustering Tasks," in *Proceedings of the 25th International Conference on Artificial Neural Networks (ICANN 2016)*, Barcelona, Spain (A. Villa, P. Masulli and A. Pons Rivero, eds.), Cham, Switzerland: Springer, 2016 pp. 192–201. doi:10.1007/978-3-319-44778-0_23.

[13] V. Ivashkin and P. Chebotarev, "Do Logarithmic Proximity Measures Outperform Plain Ones in Graph Clustering?," in *Proceedings of the International Conference on Network Analysis (NET 2016)*, Nizhny Novgorod, Russia (V. Kalyagin, A. Nikolaev, P. Pardalos and O. Prokopyev, eds.), Cham, Switzerland: Springer, 2016 pp. 87–105. doi:10.1007/978-3-319-56829-4_8.

[14] R. Aynulin, "Efficiency of Transformations of Proximity Measures for Graph Clustering," in *Proceedings of the 16th International Workshop on Algorithms and Models for the Web-Graph (WAW 2019)*, Brisbane, Australia (K. Avrachenkov, P. Prałat and N. Ye, eds.), Cham, Switzerland: Springer, 2019 pp. 16–29. doi:10.1007/978-3-030-25070-6_2.

[15] M. Girvan and M. E. J. Newman, "Community Structure in Social and Biological Networks," *Proceedings of the National Academy of Sciences*, **99**(12), 2002 pp. 7821–7826. doi:10.1073/pnas.122653799.

[16] Y. Zhou, H. Cheng and J. X. Yu, "Graph Clustering Based on Structural/Attribute Similarities," *Proceedings of the VLDB Endowment*, **2**(1), 2009 pp. 718–729. doi:10.14778/1687627.1687709.

[17] Y. Ruan, D. Fuhry and S. Parthasarathy, "Efficient Community Detection in Large Networks Using Content and Links," in *Proceedings of the 22nd International Conference on World Wide Web (WWW '13)*, Rio de Janeiro, Brazil, New York: ACM, 2013 pp. 1089–1098. doi:10.1145/2488388.2488483.

[18] J. Neville, M. Adler and D. Jensen, "Clustering Relational Data Using Attribute and Link Information," in *Proceedings of the Text Mining and Link Analysis Workshop, 18th International Joint Conference on Artificial Intelligence*, Acapulco, Mexico, San Francisco, CA: Morgan Kaufmann Publishers Inc., 2003 pp. 9–15.

[19] J. Yang, J. McAuley and J. Leskovec, "Community Detection in Networks with Node Attributes," in *Proceedings of the 13th International Conference on Data Mining (ICDM 2013)*, Dallas, TX, Piscataway, NJ: IEEE, 2013 pp. 1151–1156. doi:10.1109/ICDM.2013.167.

[20] J. MacQueen, "Some Methods for Classification and Analysis of Multivariate Observations," in *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, Oakland, CA, Piscataway, NJ: IEEE, 1967 pp. 281–297.

[21] J. Shi and J. Malik, "Normalized Cuts and Image Segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **22**(8), 2000 pp. 888–905. doi:10.1109/34.868688.

[22] U. von Luxburg, "A Tutorial on Spectral Clustering," *Statistics and Computing*, **17**(4), 2007 pp. 395–416. doi:10.1007/s11222-007-9033-z.

[23] E. Estrada, "The Communicability Distance in Graphs," *Linear Algebra and Its Applications*, **436**(11), 2012 pp. 4317–4328. doi:10.1016/j.laa.2012.01.017.

[24] F. Fouss, L. Yen, A. Pirotte and M. Saerens, "An Experimental Investigation of Graph Kernels on a Collaborative Recommendation Task," in *Proceedings of the Sixth International Conference on Data Mining (ICDM 2006)*, Hong Kong, China (C. W. Clifton, N. Zhong, J. Liu, B. W. Wah and X. Wu, eds.), Los Alamitos, CA: IEEE, 2006 pp. 863–868. doi:10.1109/ICDM.2006.18.

[25] R. I. Kondor and J. Lafferty, "Diffusion Kernels on Graphs and Other Discrete Input Spaces," in *Proceedings of the 19th International Conference on Machine Learning (ICML '02)*, Sydney, Australia (C. Sammut and A. G. Hoffmann, eds.), San Francisco, CA: Morgan Kaufmann Publishers Inc., 2002 pp. 315–322.

[26] L. Page, S. Brin, R. Motwani and T. Winograd, *The PageRank Citation Ranking: Bringing Order to the Web*, Technical Report 422, Stanford InfoLab, 1999. ilpubs.stanford.edu:8090/422.

[27] I. Kivimäki, M. Shimbo and M. Saerens, "Developments in the Theory of Randomized Shortest Paths with a Comparison of Graph Node Distances," *Physica A: Statistical Mechanics and Its Applications*, **393**, 2014 pp. 600–616. doi:10.1016/j.physa.2013.09.016.

[28] U. von Luxburg, A. Radl and M. Hein, "Getting Lost in Space: Large Sample Analysis of the Commute Distance," in *Proceedings of the 23rd International Conference on Neural Information Processing Systems (NIPS 2010)*, Vancouver, Canada (J. D. Lafferty and C. K. I. Williams, eds.), Red Hook, NY: Curran Associates Inc., 2010 pp. 2622–2630.

[29] L. Hubert and P. Arabie, "Comparing Partitions," *Journal of Classification*, **2**(1), 1985 pp. 193–218. doi:10.1007/BF01908075.

[30] G. W. Milligan and M. C. Cooper, "A Study of the Comparability of External Criteria for Hierarchical Cluster Analysis," *Multivariate Behavioral Research*, **21**(4), 1986 pp. 441–458. doi:10.1207/s15327906mbr2104_5.

[31] Z. Sulc and H. Řezanková, "Evaluation of Recent Similarity Measures for Categorical Data," in *Proceedings of the 17th International Conference Applications of Mathematics and Statistics in Economics (AMSE 2014)*, Wroclaw, Poland (Z. Rusnak and B. Zmyslona), Wrocław Poland: Wrocław University of Economics, 2014 pp. 249–258. doi:10.15611/amse.2014.17.27.

[32] P.-N. Tan, M. Steinbach, A. Karpatne and V. Kumar, *Introduction to Data Mining*, 2nd ed., Pearson Education India, 2016.

[33] T. Dang and E. Viennet, "Community Detection Based on Structural and Attribute Similarities," in *Proceedings of the 6th International Conference on Digital Society (ICDS 2012)*, Valencia, Spain (J. L. Mauri, G. Martinez, L. Berntzen and A. Smedberg, eds.), IARIA, 2012 pp. 7–12.

[34] M. Craven, D. DiPasquo, D. Freitag, A. McCallum, T. Mitchell, K. Nigam and S. Slattery, *Learning to Extract Symbolic Knowledge from the World Wide Web*, Working paper CMU-CS-98-122, School of Computer Science, Carnegie Mellon University, 1998. reports-archive.adm.cs.cmu.edu/anon/usr/ftp/1998/CMU-CS-98-122.pdf.

[35] Q. Lu and L. Getoor, "Link-Based Classification," in *Proceedings of the 20th International Conference on Machine Learning (ICML 2003)*, Washington, DC (T. Fawcett and N. Mishra, eds.), Menlo Park, CA: The AAAI Press, 2003 pp. 496–503. www.aaai.org/Papers/ICML/2003/ICML03-066.pdf.

[36] P. Sen, G. Namata, M. Bilgic, L. Getoor, B. Galligher and T. Eliassi-Rad, "Collective Classification in Network Data," *AI Magazine*, **29**(3), 2008 p. 93. doi:10.1609/aimag.v29i3.2157.