НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ
май–июнь 2023 Том 23 № 3 http://ntv.ifmo.ru/
SCIENTIFIC AND TECHNICAL JOURNAL OF INFORMATION TECHNOLOGIES, MECHANICS AND OPTICS
May–June 2023 Vol. 23 No 3 http://ntv.ifmo.ru/en/
ISSN 2226-1494 (print) ISSN 2500-0373 (online)

ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ И КОГНИТИВНЫЕ ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ

ARTIFICIAL INTELLIGENCE AND COGNITIVE INFORMATION TECHNOLOGIES

doi: 10.17586/2226-1494-2023-23-3-553-563

Role discovery in node-attributed public transportation networks: the study of Saint Petersburg city open data

Yuri V. Lytkin[1], Petr V. Chunaev[2]✉, Timofey A. Gradov[3], Anton A. Boytsov[4], Irek A. Saitov [5]

[1,2,3,4,5] ITMO University, 197101, Saint Petersburg, Russian Federation

[1] jurasicus@gmail.com, https://orcid.org/0000-0001-8140-010X
[2] chunaev@itmo.ru✉, https://orcid.org/0000-0001-8169-8436
[3] timagradov@yahoo.com, https://orcid.org/0000-0003-2537-4087
[4] aboytsov@itmo.ru, https://orcid.org/0000-0001-8343-2519
[5] xanilegendx@gmail.com, https://orcid.org/0000-0002-2805-1323

Abstract

The work presents results of modeling Public Transportation Networks (PTNs) of Saint Petersburg (Russia) and highlights the roles of stations (stops) in this network. PTNs are modeled using a new approach, previously proposed by the authors, based on weighted networks with node attributes. The nodes correspond to stations (stops) of public transport, grouped according to their geospatial location, while the node attributes contain information about social infrastructure around the stations. Weighted links integrate information about the distance and number of transfers in the routes between the stations. The role discovery is carried out by clustering the stations according to their topological and semantic attributes. The paper proposes a software framework for solving the problem of discovering roles in a PTNs. The results of its application are demonstrated on a new set of data about the PTNs of Saint Petersburg (Russia). The significant roles of the nodes of the specified PTNs were discovered in terms of both topological and infrastructural features. The overall effectiveness of the PTNs was assessed. The revealed transportation and infrastructural shortcomings of the PTNs of Saint Petersburg can be considered by the city administration to improve the functioning of these networks in the future.

Keywords

node-attributed network, public transportation network, role discovery, network node classification, network topology, social infrastructure

Acknowledgements

This study is financially supported by the Russian Science Foundation, Agreement 17-71-30029, with co-financing of the “Bank Saint Petersburg”, Russia.

For citation: Lytkin Yu.V., Chunaev P.V., Gradov T.A., Boytsov A.A., Saitov I.A. Role discovery in node-attributed public transportation networks: the study of Saint Petersburg city open data. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2023, vol. 23, no. 3, pp. 553–563. doi: 10.17586/2226-1494-2023-23-3-553-563

УДК 004.94

Выделение ролей в сетях общественного транспорта с атрибутами узлов: исследование открытых данных Санкт-Петербурга

Юрий Всеволодович Лыткин[1], Петр Владимирович Чунаев[2]✉, Тимофей Алексеевич Градов[3], Антон Алексеевич Бойцов[4], Ирек Аликович Саитов[5]

[1,2,3,4,5] Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация

[1] jurasicus@gmail.com, https://orcid.org/0000-0001-8140-010X
[2] chunaev@itmo.ru✉, https://orcid.org/0000-0001-8169-8436
[3] timagradov@yahoo.com, https://orcid.org/0000-0003-2537-4087
[4] aboytsov@itmo.ru, https://orcid.org/0000-0001-8343-2519
[5] xanilegendx@gmail.com, https://orcid.org/0000-0002-2805-1323

Аннотация

Введение. Представлены результаты моделирования и выделения ролей станций (остановок) сетей общественного транспорта на примере Санкт-Петербурга (Россия). **Метод.** Моделирование выполнено с применением нового подхода, предложенного авторами, на основе взвешенных сетей с атрибутами узлов. Узлы представляют собой станции (остановки) общественного транспорта, сгруппированные по их геопространственному положению. Атрибуты узлов содержат информацию о социальной инфраструктуре вокруг станций. Взвешенные связи интегрируют информацию о расстоянии и количестве пересадок в маршрутах движения между станциями. Выделение ролей осуществлено с помощью кластеризации станций по их топологическим и семантическим атрибутам. **Основные результаты.** Предложен программный фреймворк для решения задачи выделения ролей в сети общественного транспорта. Результаты его применения показаны на примере набора данных транспорта Санкт-Петербурга. Выделены значимые роли узлов сети с точки зрения топологических и инфраструктурных особенностей. Выполнена оценка общей эффективности сети общественного транспорта. **Обсуждение.** Результаты работы подхода могут быть использованы городской администрацией для выявления транспортных и инфраструктурных недостатков сети общественного транспорта Санкт-Петербурга для улучшения функционирования сети в будущем.

Ключевые слова

сеть с атрибутами узлов, сеть общественного транспорта, выделение ролей, классификация узлов сети, топология сети, социальная инфраструктура

Благодарности

Исследование выполнено при финансовой поддержке Российского научного фонда, соглашение 17-71-30029, при софинансировании Банком Санкт-Петербург (Россия).

Ссылка для цитирования: Лыткин Ю.В., Чунаев П.В., Градов Т.А., Бойцов А.А., Саитов И.А. Выделение ролей в сетях общественного транспорта с атрибутами узлов: исследование открытых данных Санкт-Петербурга // Научно-технический вестник информационных технологий, механики и оптики. 2023. Т. 23, № 3. С. 553–563 (на англ. яз.). doi: 10.17586/2226-1494-2023-23-3-553-563

Intoduction

This paper is a companion one for the paper [1] where, based on the recent advances in the network theory for Public Transportation Network (PTN) [2–5] and role discovery in social networks [6–8], a novel weighted node-attributed PTN model (using information about a city's social infrastructure to construct the node attributes) was developed and applied by the authors to propose an approach to discover roles of public transport stops and stations.

In this paper, we continue the study in [1], use the terms introduced there, develop a programming framework for solving the problem of role discovery in a PTN and apply it to the newly collected open PTN data of Saint Petersburg, Russia. Our framework aims at discovering meaningful node roles in terms of both network topology (i.e., transition hubs, outskirts, etc.) and node infrastructural attributes — semantics (i.e., tourist, residential, industrial areas, etc.) and is capable of extracting useful insights about the overall PTN's efficiency. As a result of our experimental study, we point out some transportation and infrastructural weaknesses that should be taken into consideration by the city administration of Saint Petersburg to improve the PTN in the future.

The work is organized as follows. Brief information about the previously published approach (model) [1] is given. The results of experiments and their analysis are presented: a review of data, a description of the process of a supernodes network building, a discussion of the nodes functions and their construction, simulation results. The prospects for possible future research are discussed.

The model and the role discovery approach

The model of a node-attributed PTN. The terms used here and below are described in detail in the companion paper [1]. Let us just recall that formally the model of a PTN as a weighted node-attributed network can be defined as a tuple

$$G = (V, E, A),$$

where V is the set of nodes, $E \subseteq V \times V \times \mathbb{R}$ is the set of undirected weighted links, and $A: V \rightarrow \mathbb{R}^n$ is a mapping that defines the set of node-attributed vectors. Here the nodes are stations (stops) grouped with respect to their geospatial position, node attributes store information about social infrastructure around the stations (stops), and weighted links integrate information about the travelling distance and the number of hops in the transportation routes between the stations (stops). Recall that each component of G is described in detail in the companion paper [1].

The role discovery task for the node-attributed PTN. Recall that the task of role discovery originated in the field of social network analysis, but found its way into a variety of different domains of science [1]. The basic approach to this task is to extract some features of the network nodes and then use machine learning algorithms (i.e., K-Means [9]) to extract clusters based on these features. Even though originally only topological features were used in this approach, the basic framework can naturally be extended to include also node-attributed vectors. For example, one can obtain two separate clustering (with respect to topological features and node attributes) and then analyze their relationship, for instance, using a contingency table. In this study, as proposed in [1], we adopt this approach.

Recall that the reason for this choice is also provided in the companion paper [1].

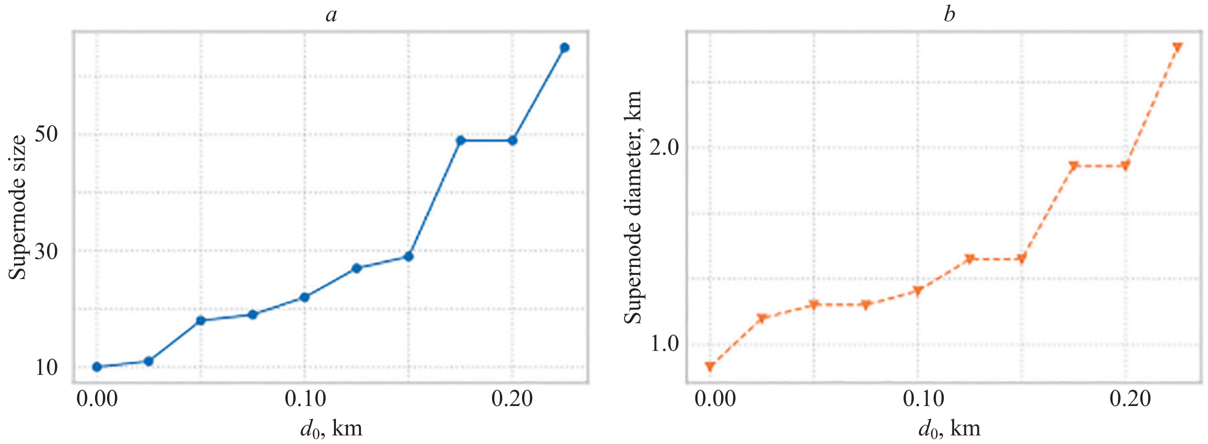

Fig. 1. Maximal supernode size (*a*) and diameter (*b*) for different values of d_0. Even for relatively small values ($d_0 > 0.15$) these characteristics grow quite rapidly resulting in some supernodes having diameter as large as 2 km and more

Role discovery in the PTN of Saint Petersburg. We apply the model for the task of role discovery on the PTN of Saint Petersburg, Russia. The task is to analyze and separate in an *unsupervised* manner various public transportation stops into homogeneous clusters, i.e., groups of nodes with similar characteristics including infrastructural attributes and topological features. This task is useful in the analysis of public transportation systems since these clusters can then be further analyzed in terms of their role within the public transportation system and reachability between them.

Dataset description. First, we describe the data[1] needed to build the PTN model proposed above. As we mentioned before, we only use the general public transportation and infrastructural data that is available for most of the cities in the world, thus making it possible to use the proposed techniques for analyzing public transportation systems of virtually any city.

The two sources of original data are:

— Saint Petersburg city Open Data[2], containing information about different public transportation stops and routes operating in Saint Petersburg, Russia;

— OpenStreetMap (OSM)[3], containing information about various infrastructural objects around Saint Petersburg.

The public transportation data is presented in the form of a table with each row containing information about a stop as part of some transportation route. Each such row consists of information about the current route (its ID and mode of transportation), the current stop (its ID and coordinates), and the next stop corresponding to the route. The three modes of transportation presented here are: bus, tramway, and trolleybus. Note that there is no subway data available here, thus it will be extracted from the second source of data.

The data from OSM is presented in form of a JSON list containing information on various objects in and around Saint Petersburg. These objects can be either *nodes* or *ways*. A *node* usually denotes a single point on the map, and they are used in cases where the size of an object does not matter too much (for instance, bus stops, historical monuments, etc.). By contrast, *ways* (sequences of nodes) are usually used to represent larger objects (big buildings, industrial areas, etc.). Each object is represented as a dictionary that contains information on this object (namely, its ID, coordinates and some attributes corresponding to its type). These attributes are usually quite precise and can be used to extract more topic-specific information about each object. This data is used in two ways. Firstly, we extract information on subway routes and stations to add another mode of transportation to those presented in the first data source. The basic statistics on the completed public transportation data can be seen in Table 1.

Secondly, we group various infrastructural objects into 20 groups related to different types of social infrastructure (i.e., housing, shopping, restaurants, medicine, etc.). All of this is done using the OSM attributes of these objects, and the specific correspondences between these attributes and the resulting infrastructure types can be found in the Github repository. The number of infrastructural objects of each type can be seen in Fig. 2.

Supernode network. This data is then used to build the model described in previous section. First of all, supernodes are produced by combining the closely situated stops and stations. We use the distance threshold $d_0 = 0.1$, i.e., all stops that are closer than d_0 are combined into a single supernode (Fig. 3 and the definitions in [1]). In practice, this can be achieved using the following algorithm:

1. Calculate the distances between all pairs of stops.
2. Build a graph in which a link between two stops s_1 and s_2 exists if $d(s_1, s_2) \leq d_0$.
3. Each connected component of this graph is a separate supernode.

Secondly, we need to construct links for our network. As we mentioned in [1], we adopt the *P*-space model for constructing links, i.e., a link between nodes s_i and s_j means that there exists a route connecting these nodes (not necessarily consecutively). We also use link weights

[1] The data along with all preprocessing and analysis procedures is available in the Github repository at https://github.com/AlgoMathITMO/public-transport-network (accessed: 01.10.2022).

[2] Available at: https://classif.gov.spb.ru/irsi/7830001067-marshruty-dvizheniya-gorodskogo-transporta/structure_version/186/ (accessed: 01.10.2022).

[3] Available at: https://www.openstreetmap.org/ (accessed: 01.10.2022).

Table 1. Statistics on public transportation of Saint Petersburg

	Bus	Subway	Tram	Trolley
Number of stops	5511	71	887	1192
Number of routes	1070	10	83	90
Average route length, km	14.15	24.02	10.38	11.32

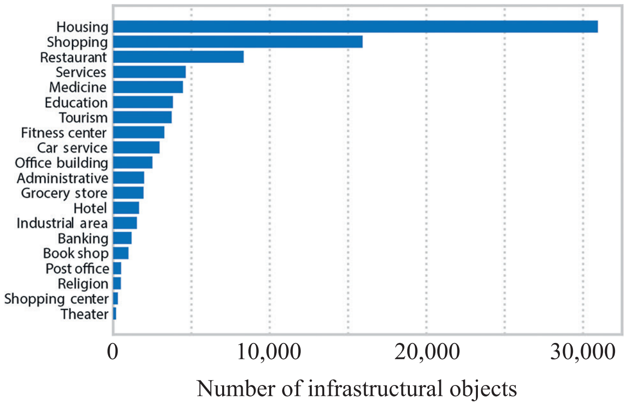

Fig. 2. Number of infrastructural objects of various types in Saint Petersburg

defined in [1] with α = 0.2. In practice, this can be done using the following algorithm:

1. Iterate through all routes. For each route r do the following.
2. Iterate through all pairs of nodes in r. For each pair of nodes s_i, s_j calculate and store the route distance $rd_r(s_i, s_j)$ [1].
3. When this is done, for each pair of nodes s_i, s_j take the minimal route distance $rd(s_i, s_j)$ and use it to calculate the link weight $w(s_i, s_j)$ [1].

Note that the built graph is not always connected. In Fig. 4 we can see two small portions of nodes disconnected from the main body of the graph. This can happen if each of the route stops is too far away from the rest of the nodes in the graph, thus making it impossible to make a short walk to reach it.

Finally, we need to construct the supernode attributes based on the social infrastructure around them. As it was described in [1], we use a distance window $d_1 = 0.2$ here. This value controls the additional distance (with respect to the minimal distance to any supernode) allowed in order to assign a given infrastructure object to a supernode (Fig. 5). In general, assigning infrastructure objects to supernodes can be done using the following algorithm:

1. Iterate through all infrastructure objects. For each object i do the following.
2. Calculate distances from i to each supernode.
3. Take the minimal distance $d_{\min}$. Assign object i to each stop s such that $d(i, s) \leq d_{\min} + d_1$.

Note that in general not all nodes in the graph will be assigned to infrastructure objects, and there can be some nodes with no infrastructure around them. Fig. 6 shows all the supernodes of the Saint Petersburg PTN with colors indicating the number of infrastructure objects assigned to them.

Supernode features. Recall that in this study we perform separate clustering with respect to topological features (derived from the network structure) and infrastructure features (using the supernode attributes, see previous section) and then compare the two.

Firstly, we consider the supernode attributes which are in [1] and were built above in this paper. To construct *infrastructure features* from these attributes, we additionally

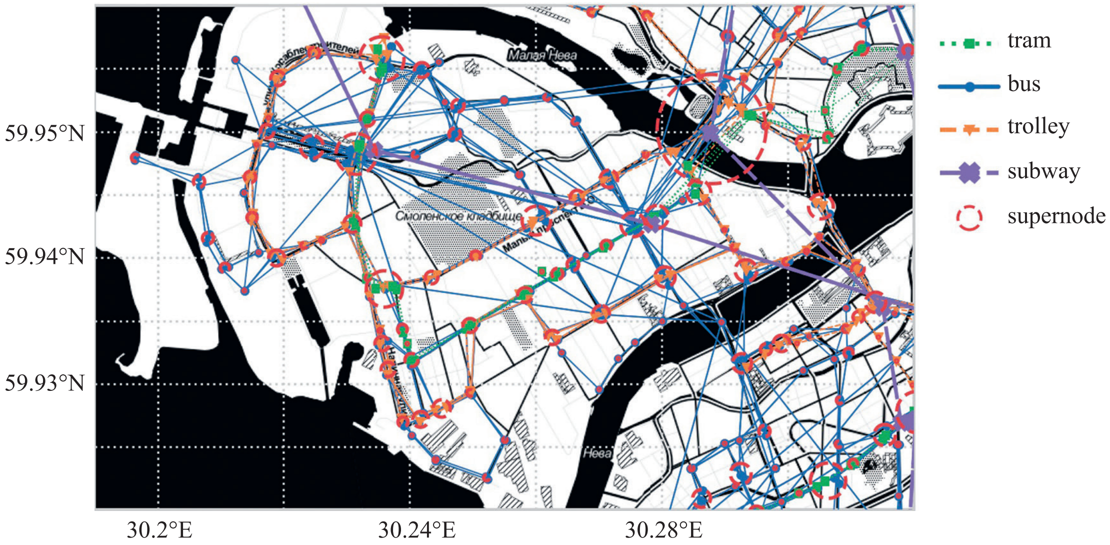

Fig. 3. The map of Vasileostrovsky District in Saint Petersburg indicating stops and routes for different modes of transportation as well as the supernodes (groups of nearby stops).

The picture is taken from [1]

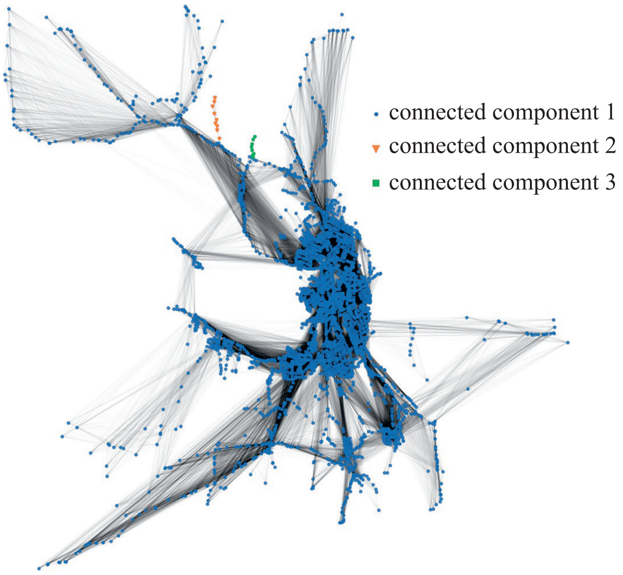

Fig. 4. The *P*-space supernode graph built using the Saint Petersburg data. The node positions correspond to the geographical locations of the corresponding nodes. There is a small portion of nodes that are disconnected from the main body of the graph

divide each vector $\mathbf{v} = A(s)$ by its sum $\sum \mathbf{v}$ (obviously, excluding the cases where $\sum \mathbf{v} = 0$). Therefore, each such infrastructure feature vector shows the orientation of a given supernode towards one or multiple infrastructure types (for example, some nodes can be mainly housing-oriented having a large value corresponding to the housing infrastructure type and smaller values on other positions), regardless of the total number of infrastructure objects around the node.

The second set of supernode *topological features* is constructed based on the topology, i.e., the connectivity structure of the network. Here we use various well-known centrality measures, namely, *degree centrality*, *betweenness centrality* [10] (considering weights w when calculating shortest paths), and *closeness centrality* [11] (using both weights w and the number of hops) as well as other topological features like the *local clustering coefficient* [12] and *PageRank* [13].

As mentioned in [1], the choice of a network model (*L*-space or *P*-space) is of paramount importance when using and interpreting such topological features. For instance, we argue against using centrality measures based on shortest paths with the *L*-space model since such shortest paths are not indicative of the optimal travelling routes of passengers (Fig. 4), thus making any analysis of these centralities (as in [5]) rather questionable.

Contrarily to this, the mentioned centrality measures offer a natural interpretation when using them with the *P*-space model. For instance, degree centrality emphasizes the so-called accessibility hubs, i.e., nodes from which a lot of other nodes are accessible without need to make a connection. Betweenness centrality emphasizes the transportation hubs, i.e., the nodes at which a lot of connections happen. Closeness centrality emphasizes the nodes that on average require the least travelling distance (in sense of either weighted distance w, or the number of connections) to reach. PageRank is similar to degree centrality, but it also promotes the nodes that are connected to many important nodes of the graph. All these centrality measures therefore highlight different aspects of centrality that can occur in a PTN.

The least intuitive feature here is the local clustering coefficient which is the fraction of closed triangles that exist in the neighborhood of a given node. Since in the *P*-space model all the node pairs inside each route are connected with a link (and therefore all possible triangles exist around these nodes in these cases), local clustering coefficient does the contrary to the measures described above and actually emphasizes the nodes that are a part of the least number of different routes.

Before turning to the clustering task, we examine these features a bit more. Fig. 7 shows the heatmap of Spearman correlations between the features. Note that some

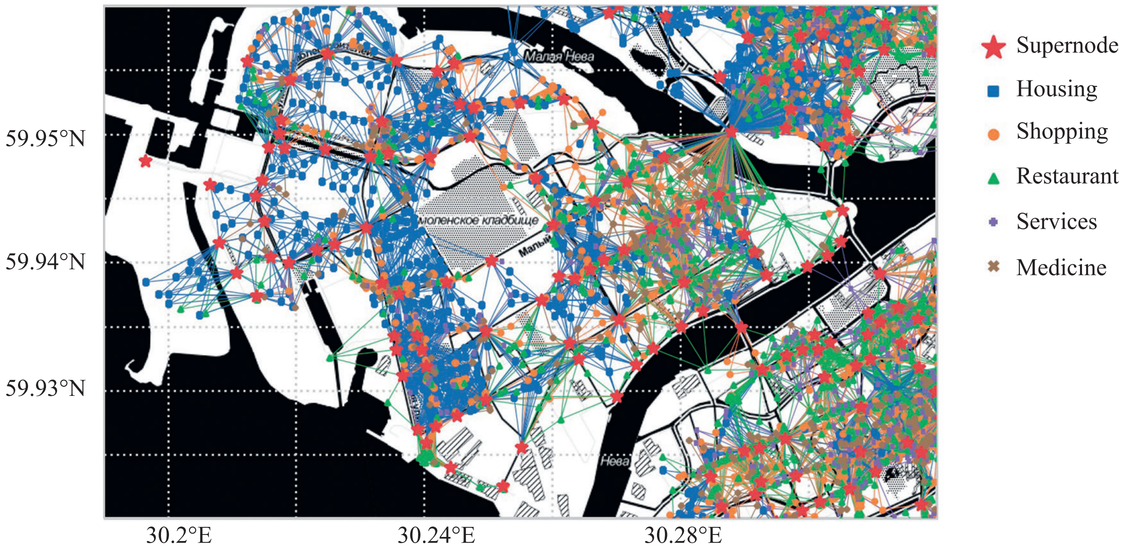

Fig. 5. The map of Vasileostrovsky District in Saint Petersburg indicating supernodes and various infrastructural objects attached to them. The picture is taken from [1]

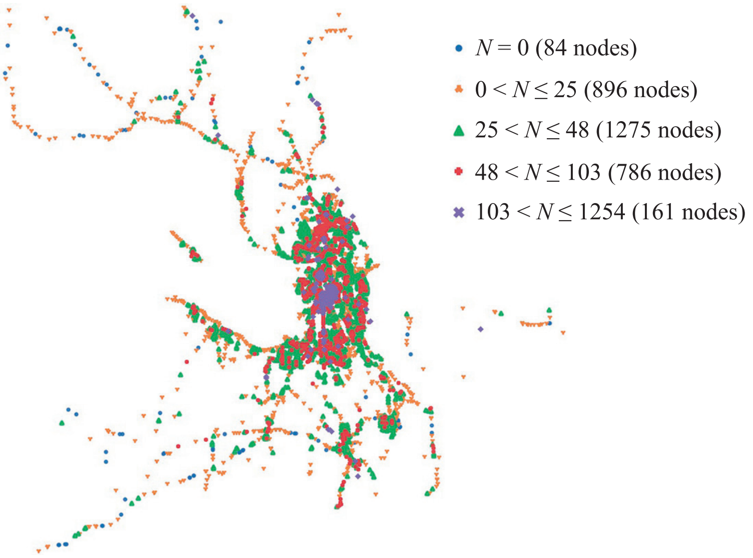
● $N = 0$ (84 nodes)
▼ $0 < N \leq 25$ (896 nodes)
▲ $25 < N \leq 48$ (1275 nodes)
● $48 < N \leq 103$ (786 nodes)
✖ $103 < N \leq 1254$ (161 nodes)

Fig. 6. The nodes of the Saint Petersburg graph with colors indicating the number of infrastructure objects assigned to them (N)

of the higher correlations (in absolute terms) are actually expected, such as the strong positive correlation between the centrality measures and the strong negative correlation between the latter and the local clustering coefficient. The other correlations are more interesting though. For instance, we can see a significant positive correlation between the centrality measures and some infrastructure features, such as Restaurant, Service, Office building and Banking. These (expectedly) indicate that there are on average more of such infrastructure objects towards the city center. Such correlation is less noticeable for Shopping centers and Post offices, indicating that these infrastructure objects can generally be found everywhere throughout the city, not just in the center. On the other hand, we can note a slight negative correlation between Housing and many of the other features (notably, except Grocery stores and Education), which indicates a tendency towards higher isolation of the residential districts in Saint Petersburg.

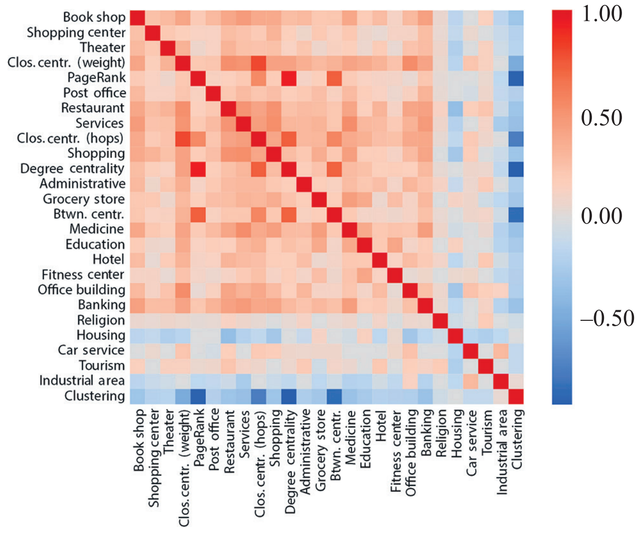

Fig. 7. The heatmap of Spearman correlations between various supernode graph features.

(For interpretation of the references to color in this heatmap, the reader is referred to the web version of this article)

Supernode clustering. Finally, in this section we perform the cluster analysis of the supernodes with respect to both infrastructure and graph features (see previous section). The framework of this analysis is as follows. For a given set of features, we first obtain their clusters using the K-Means algorithm [9]. The number of clusters is determined based on the inertia plot and interpretability of the clusters. The clusters are then plotted in different views and interpreted based on their features.

We first perform cluster analysis of the infrastructure features. Fig. 8 shows the t-SNE projection [14] and geographical positions of the supernodes, colored based on the obtained clusters. To analyze the difference between these clusters and interpret them, we also plot the aggregated features over each cluster (Fig. 9). In the upper part of the figure the mean feature values are plotted for each cluster as well as the global mean. Also, to emphasize the difference between the feature values in each cluster, in the lower part we plot the values of the 2-sample Welch's t-test statistic [15] for comparing the mean of each feature over the given cluster compared to the mean of this feature over the rest of the clusters (this is the so-called one-vs-rest strategy).

The obtained clusters can be summarized as follows:

1. Tourism area. Nodes with mostly tourist attractions around them, not much else.
2. Residential area — unimproved. Nodes with mostly housing around them and no other common urban amenities, such as grocery stores, hospitals, etc.
3. Center. Nodes located around shops, restaurants, office buildings, banks, etc.
4. Residential area — improved. Nodes with housing as well as various amenities, like schools, hospitals, grocery stores, fitness centers, etc.
5. Industrial area. Nodes surrounded by industrial areas and not much else.
6. No infrastructure. Nodes that have no social infrastructure around.

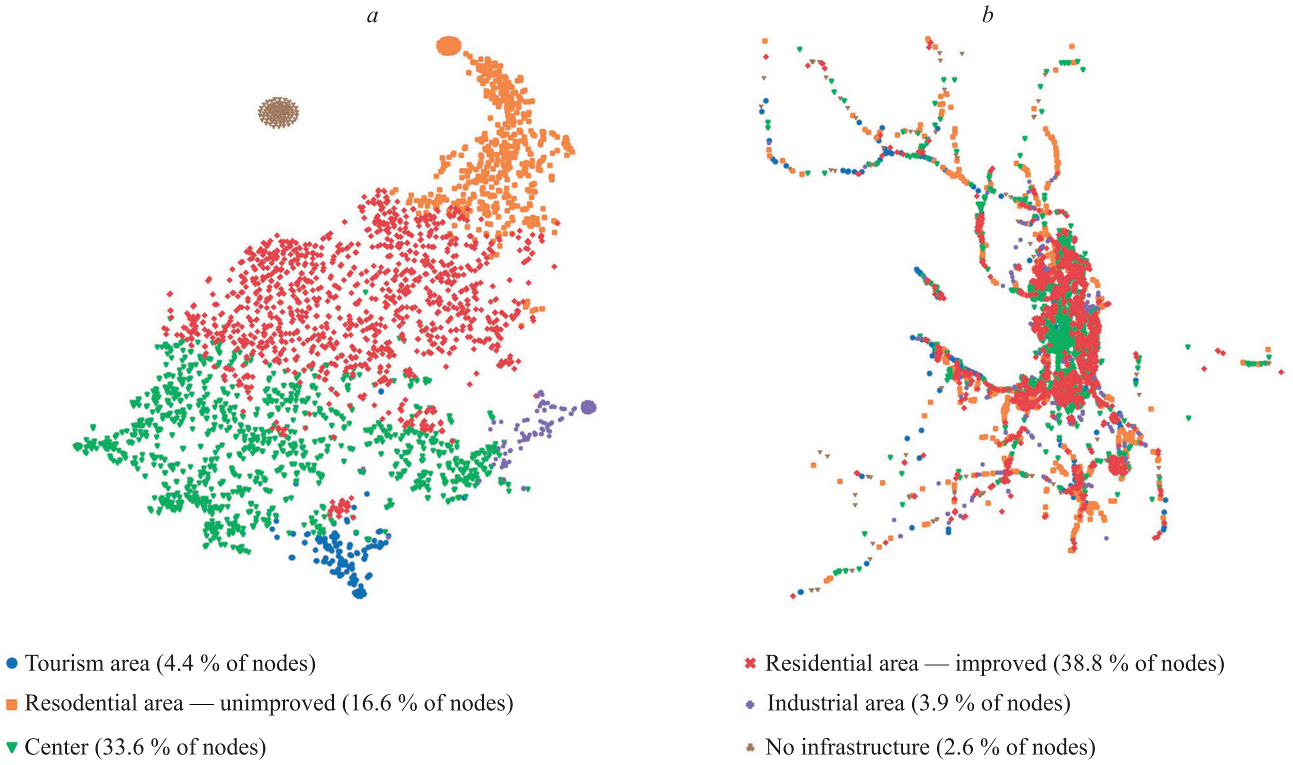

Fig. 8. t-SNE projection (*a*) and geographical positions (*b*) of supernodes, colored based on the clusters by their infrastructure features. In (*a*), the circle-shape clusters are groups of supernodes with the same attribute vectors; e.g., the brown cluster is the group of supernodes with zero-vector attributes (i.e., the groups of stops having no infrastructure around)

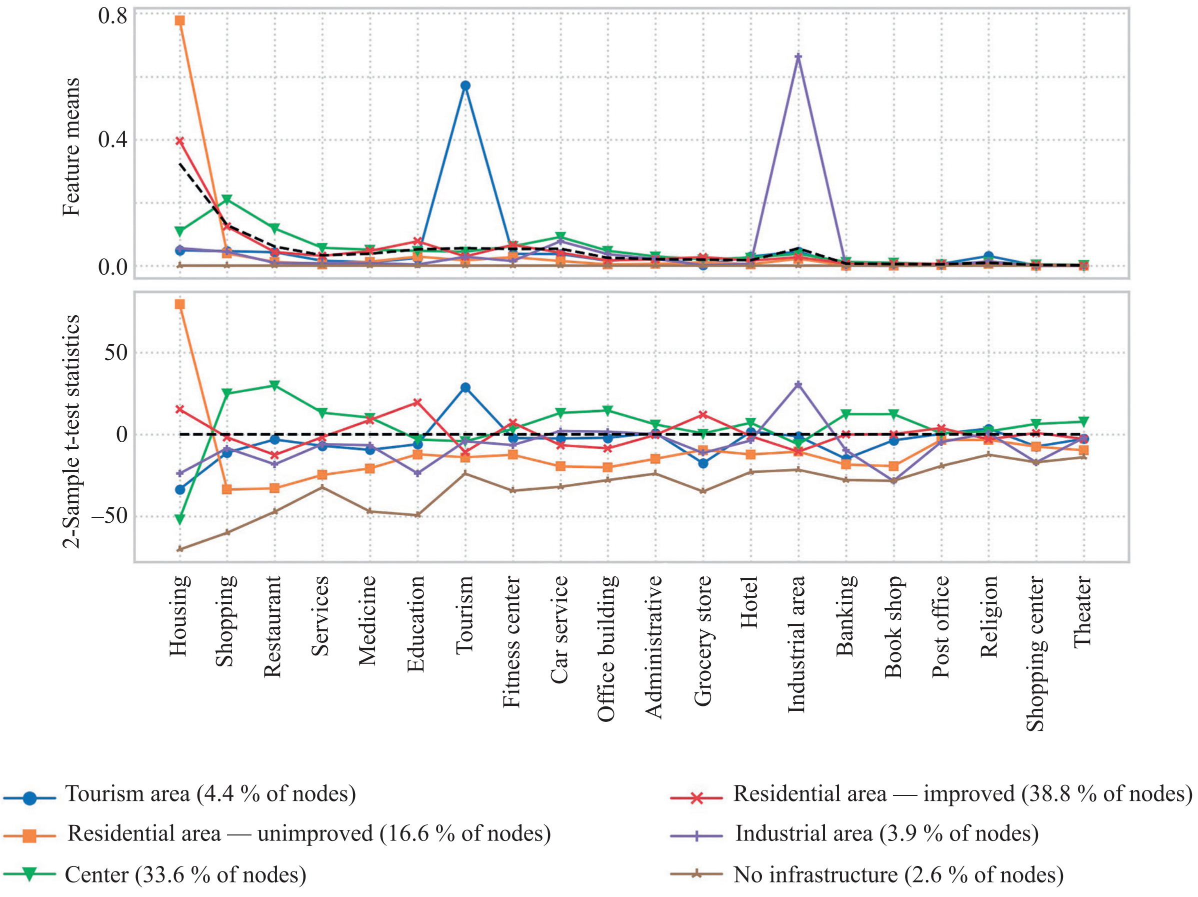

Fig. 9. Aggregated features of supernodes from different infrastructure clusters. The upper plot shows a mean value of each feature across each cluster as well as the global mean. The lower plot shows the values of the 2-sample Welch's t-test statistic [15] for comparing the mean of each feature over the given cluster, compared to the mean of this feature over the rest of the clusters

It should be noted that the proposed interpretation is not exactly strict since, as it can be seen in the t-SNE projection in Fig. 8, there are no clear boundaries between these clusters (except for the cluster with no social infrastructure). Thus by means of social infrastructure, one can see smoothly changing and highly various set of supernode types in the PTN under consideration.

The cluster analysis of graph features is done in a similar fashion: the clustering is performed using the K-Means algorithm, the clusters are plotted using their t-SNE projection as well as the geographical positions in Fig. 10, and their aggregated features are shown in Fig. 11. Based on the presented data, these clusters can be summarized as follows:

1. Hubs. Nodes that serve as points of transition between different routes when travelling through a city. These nodes have higher degree and betweenness centrality, compared to the other nodes.
2. Center. Nodes that represent the well-accessible part of the city center. These nodes have high values of centralities.
3. Inaccessible center. Nodes that represent the less accessible part of the city center. These nodes have high closeness centrality based on the link weights which indicate their close proximity to the center, but at the same time these nodes have low closeness centrality based on hops (which indicates that these nodes on average require much more connections to reach) as well as low betweenness and degree centrality.
4. Towns. Nodes located outside the main city in separate towns (low betweenness, closeness and degree centrality).
5. Suburbs. Nodes that are located moderately far away from the city center (lower closeness centrality), but they are still well-connected to the transportation network (moderate betweenness and degree centrality).
6. Disconnected nodes. A few nodes that are not connected to the transportation network and form separate connected components. (Note that this cluster is not presented in Fig. 11 since its 2-test statistic values are extremely low and render all the other plots impossible to read.)

As with the infrastructure-based clustering, it should be noted that there are no clear boundaries between the topology-based clusters. Nevertheless, this clustering shows the different high-level roles of the network nodes and provides insight into the relations between these clusters.

Finally, in order to assess the relations between the infrastructure and topology feature clusters, we build a contingency table by counting the number of nodes in different intersections of these clusters. These values are presented in Table 2. The rows of the table represent the infrastructure clusters and the graph feature clusters are represented by the columns.

From this table some interesting interconnections between the two clusterings arise. We can see that most of the infrastructure clusters are well-represented in all of the graph-feature clusters (and vice versa), which means that these two clusterings both carry important and unique information about the roles of each node. For instance, we can see that the nodes corresponding to improved residential areas (with better-developed urban amenities) have more members in graph-based clusters ‘Center’, ‘Hub’, and ‘Inaccessible center’, as well as ‘Suburbs’, at the same time there are more undeveloped residential areas in the ‘Towns’ cluster.

Another important cluster to consider from the urban development point of view is the graph-feature cluster

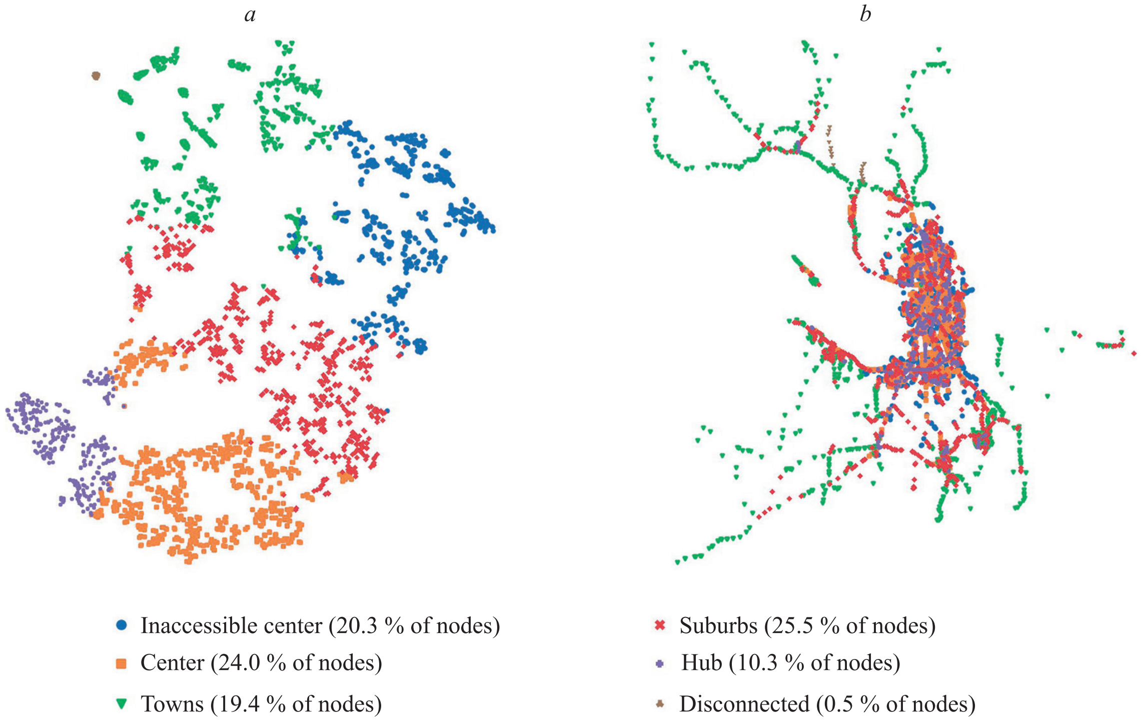

Fig. 10. t-SNE projection (*a*) and geographical positions (*b*) of supernodes, colored based on the clusters obtained using their graph features

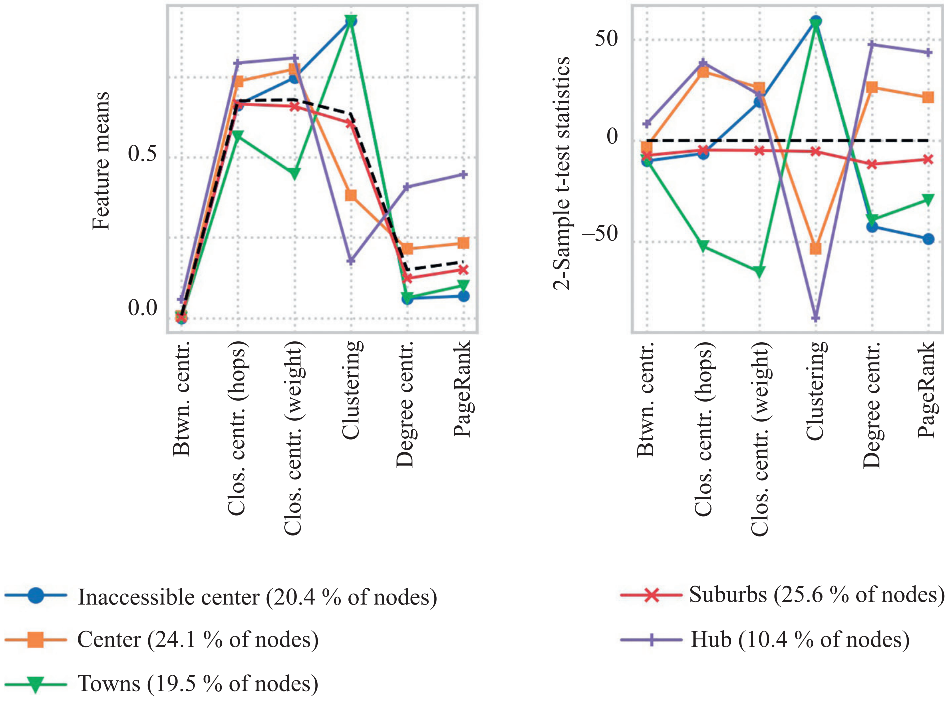

Fig. 11. Aggregated features of supernodes from different graph feature clusters. The plot (*a*) shows a mean value of each feature across each cluster as well as the global mean. The plot (*b*) shows the values of the 2-sample Welch's t-test statistic [15] for comparing the mean of each feature over the given cluster, compared to the mean of this feature over the rest of the clusters

Table 2. Contingency table showing the relations between infrastructure clusters and graph feature clusters. The rows of the table represent the infrastructure clusters and the graph feature clusters are represented by the columns

Area	Center	Disconnected	Hub	Inaccessible center	Suburbs	Towns
Center	300	3	169	265	225	115
Industrial area	3	0	0	33	43	47
No infrastructure	0	3	0	2	11	68
Residential area (improved)	371	2	144	249	344	131
Residential area (unimproved)	74	9	15	92	139	203
Tourism area	20	0	2	10	54	56

'Inaccessible center', which contains nodes that are fairly central in terms of closeness centrality (i.e., average distance from the rest of the nodes), but they are low on betweenness and degree centrality, which means that public transportation is under-developed in these areas. We can see that this cluster contains many members of the infrastructure clusters 'Center' and 'Residential area — improved', which means that these areas are well-developed in terms of urban amenities, but are quite separated from the rest of the PTN, which means less convenience of daily commuting, for instance.

Conclusions and future work

In this work, we applied a novel weighted node-attributed PTN model (using information about a city's social infrastructure to construct the node attributes) and the approach to discover roles of public transport stops and stations to the Saint Petersburg open PTN data. For this purpose, a novel role discovery programming framework was introduced which uses both structural (i.e., network topology) and semantic (i.e., social infrastructure around the nodes) aspects of a node-attributed PTN. This framework is shown to be capable of extracting useful information about the properties and overall efficiency of a city's public transportation system from both the structural and infrastructure standpoints. For instance, in case of Saint Petersburg, it is able to point out some under-developed areas of the city, e.g., less accessible parts of the city center or residential areas that are low on urban amenities. These weaknesses can lead to better development of the city in the future if taken into consideration by the city administration. The performed analysis uses only the generally available data, which means that similar analysis can be performed on any large city's public transportation system. In general, the proposed approach to role discovery in node-attributed networks can be applied beyond the scope of PTNs and to any other kind of network (e.g. social, biological, technical, etc.), given the appropriate set of node attributes.

Among other things, a potential direction of future research is found: to develop more interpretable graph-based node metrics that would highlight even more peculiarities in the different roles of the nodes in a PTN. For instance, as it is already mentioned, the metric of betweenness centrality over a *P*-space model graph highlights the nodes at which a lot of transfers happen. At the same time, the actual stops through which these shortestroutes and a -space graph pass are not highlighted by any of the existing metrics (and are not actually even considered in a *P*-space model). A metric like this could bring up very important information about the actual workload of different PTN nodes without the need for any dynamic data like transportation of passenger flows.

It is also reasonable to further enhance the results of this study by performing the node classification task on the data, which could be also a validation of clustering approach for role detection.

References

1. Lytkin Yu.V., Chunaev P.V., Gradov T.A., Boytsov A.A., Saitov I.A. Role discovery in node-attributed public transportation networks: the model description. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2023, vol. 23, no. 2, pp. 340–351. https://doi.org/10.17586/2226-1494-2023-23-2-340-351
2. Haznagy A., Fi I., London A., Nemeth T. Complex network analysis of public transportation networks: A comprehensive study. *Proc. of the 2015 International Conference on Models and Technologies for Intelligent Transportation Systems (MT-ITS)*, 2015, pp. 371–378. https://doi.org/10.1109/mtits.2015.7223282
3. Yang X.-H., Chen G., Chen S.-Y., Wang W.-L., Wang L. Study on some bus transport networks in china with considering spatial characteristics. *Transportation Research Part A: Policy and Practice*, 2014, vol. 69, pp. 1–10. https://doi.org/10.1016/j.tra.2014.08.004
4. Wang L.-N., Wang K., Shen J.-L. Weighted complex networks in urban public transportation: Modeling and testing. *Physica A: Statistical Mechanics and its Applications*, 2020, vol. 545, pp. 123498. https://doi.org/10.1016/j.physa.2019.123498
5. Shanmukhappa T., Ho I.W.-H., Tse C.K. Spatial analysis of bus transport networks using network theory. *Physica A: Statistical Mechanics and its Applications*, 2018, vol. 502, pp. 295–314. https://doi.org/10.1016/j.physa.2018.02.111
6. Rossi R.A., Ahmed N.K. Role discovery in networks. *IEEE Transactions on Knowledge and Data Engineering*, 2015, vol. 27, no. 4, pp. 1112–1131. https://doi.org/10.1109/tkde.2014.2349913
7. Gupte P.V., Ravindran B., Parthasarathy S. Role discovery in graphs using global features: Algorithms, applications and a novel evaluation strategy. *Proc. of the 2017 IEEE 33rd International Conference on Data Engineering (ICDE)*, 2017, pp. 771–782. https://doi.org/10.1109/icde.2017.128
8. Revelle M., Domeniconi C., Johri A. Persistent roles in online social networks. *Lecture Notes in Computer Science*, 2016, vol. 9852, pp. 47–62. https://doi.org/10.1007/978-3-319-46227-1_4
9. MacQueen J. Some methods for classification and analysis of multivariate observations. *Proc. of the Fifth Berkeley Symposium on Mathematical Statistics and Probability. V. 1: Statistics*, 1967, pp. 281–297.
10. Freeman L.C. A set of measures of centrality based on betweenness. *Sociometry*, 1977, vol. 40, no. 1, pp. 35–41. https://doi.org/10.2307/3033543
11. Freeman L.C. Centrality in social networks conceptual clarification. *Social Networks*, 1978, vol. 1, no. 3, pp. 215–239. https://doi.org/10.1016/0378-8733(78)90021-7
12. Onnela J.-P., Saramäki J., Kertész J., Kaski K. Intensity and coherence of motifs in weighted complex networks. *Physical Review E*, 2005, vol. 71, no. 6, pp. 065103. https://doi.org/10.1103/physreve.71.065103
13. Page L., Brin S., Motwani R., Winograd T. *The pagerank citation ranking: Bringing order to the web*: Technical Report 1999-66, Stanford InfoLab, November 1999.
14. Van der Maaten L., Hinton G. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 2008, vol. 9, no. 86, pp. 2579–2605.
15. Welch B.L. The generalization of `student's' problem when several different population variances are involved. *Biometrika*, 1947, vol. 34, no. 1-2, pp. 28–35. https://doi.org/10.2307/2332510

Литература

1. Lytkin Yu.V., Chunaev P.V., Gradov T.A., Boytsov A.A., Saitov I.A. Role discovery in node-attributed public transportation networks: the model description // Научно-технический вестник информационных технологий, механики и оптики. 2023. Т. 23. № 2. С. 340–351. https://doi.org/10.17586/2226-1494-2023-23-2-340-351
2. Haznagy A., Fi I., London A., Nemeth T. Complex network analysis of public transportation networks: A comprehensive study // Proc. of the 2015 International Conference on Models and Technologies for Intelligent Transportation Systems (MT-ITS). 2015. P. 371–378. https://doi.org/10.1109/mtits.2015.7223282
3. Yang X.-H., Chen G., Chen S.-Y., Wang W.-L., Wang L. Study on some bus transport networks in china with considering spatial characteristics // Transportation Research Part A: Policy and Practice. 2014. V. 69. P. 1–10. https://doi.org/10.1016/j.tra.2014.08.004
4. Wang L.-N., Wang K., Shen J.-L. Weighted complex networks in urban public transportation: Modeling and testing // Physica A: Statistical Mechanics and its Applications. 2020. V. 545. P. 123498. https://doi.org/10.1016/j.physa.2019.123498
5. Shanmukhappa T., Ho I.W.-H., Tse C.K. Spatial analysis of bus transport networks using network theory // Physica A: Statistical Mechanics and its Applications. 2018. V. 502. P. 295–314. https://doi.org/10.1016/j.physa.2018.02.111
6. Rossi R.A., Ahmed N.K. Role discovery in networks // IEEE Transactions on Knowledge and Data Engineering. 2015. V. 27. N 4. P. 1112–1131. https://doi.org/10.1109/tkde.2014.2349913
7. Gupte P.V., Ravindran B., Parthasarathy S. Role discovery in graphs using global features: Algorithms, applications and a novel evaluation strategy // Proc. of the 2017 IEEE 33rd International Conference on Data Engineering (ICDE). 2017. P. 771–782. https://doi.org/10.1109/icde.2017.128
8. Revelle M., Domeniconi C., Johri A. Persistent roles in online social networks // Lecture Notes in Computer Science. 2016. V. 9852. P. 47–62. https://doi.org/10.1007/978-3-319-46227-1_4
9. MacQueen J. Some methods for classification and analysis of multivariate observations // Proc. of the Fifth Berkeley Symposium on Mathematical Statistics and Probability. V. 1: Statistics. 1967. P. 281–297.
10. Freeman L.C. A set of measures of centrality based on betweenness // Sociometry. 1977. V. 40. N 1. P. 35–41. https://doi.org/10.2307/3033543
11. Freeman L.C. Centrality in social networks conceptual clarification // Social Networks. 1978. V. 1. N 3. P. 215–239. https://doi.org/10.1016/0378-8733(78)90021-7
12. Onnela J.-P., Saramäki J., Kertész J., Kaski K. Intensity and coherence of motifs in weighted complex networks // Physical Review E. 2005. V. 71. N 6. P. 065103. https://doi.org/10.1103/physreve.71.065103
13. Page L., Brin S., Motwani R., Winograd T. The pagerank citation ranking: Bringing order to the web: Technical Report 1999-66, Stanford InfoLab, November 1999.
14. Van der Maaten L., Hinton G. Visualizing data using t-SNE // Journal of Machine Learning Research. 2008. V. 9. N 86. P. 2579–2605.
15. Welch B.L. The generalization of `student's' problem when several different population variances are involved // Biometrika. 1947. V. 34. N 1-2. P. 28–35. https://doi.org/10.2307/2332510

Authors

Yuri V. Lytkin — PhD (Physics & Mathematics), Senior Researcher, ITMO University, Saint Petersburg, 197101, Russian Federation, sc 57155292900, https://orcid.org/0000-0001-8140-010X, jurasicus@gmail.com

Petr V. Chunaev — PhD (Physics & Mathematics), Senior Researcher, ITMO University, Saint Petersburg, 197101, Russian Federation, sc 36522457300, https://orcid.org/0000-0001-8169-8436, chunaev@itmo.ru

Timofey A. Gradov — Engineer, ITMO University, Saint Petersburg, 197101, Russian Federation, sc 57221121540, https://orcid.org/0000-0003-2537-4087, timagradov@yahoo.com

Anton A. Boytsov — Engineer, ITMO University, Saint Petersburg, 197101, Russian Federation, https://orcid.org/0000-0001-8343-2519, aboytsov@itmo.ru

Irek A. Saitov — Engineer, ITMO University, Saint Petersburg, 197101, Russian Federation, sc 57215429754, https://orcid.org/0000-0002-2805-1323, xanilegendx@gmail.com

Авторы

Лыткин Юрий Всеволодович — кандидат физико-математических наук, старший научный сотрудник, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, sc 57155292900, https://orcid.org/0000-0001-8140-010X, jurasicus@gmail.com

Чунаев Петр Владимирович — кандидат физико-математических наук, старший научный сотрудник, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, sc 36522457300, https://orcid.org/0000-0001-8169-8436, chunaev@itmo.ru

Градов Тимофей Алексеевич — инженер, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, sc 57221121540, https://orcid.org/0000-0003-2537-4087, timagradov@yahoo.com

Бойцов Антон Алексеевич — инженер, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, https://orcid.org/0000-0001-8343-2519, aboytsov@itmo.ru

Саитов Ирек Аликович — инженер, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, sc 57215429754, https://orcid.org/0000-0002-2805-1323, xanilegendx@gmail.com

Received 09.12.2022
Approved after reviewing 06.03.2023
Accepted 16.05.2023

Статья поступила в редакцию 09.12.2022
Одобрена после рецензирования 06.03.2023
Принята к печати 16.05.2023