USING NEO4J DATABASE AND GRAPH MODEL FOR ANALYSIS OF METROPOLITAN RAILWAY CONNECTIONS OF SILESIAN VOIVODESHIP IN POLAND

PAWEŁ BUCHWALD* AND PAWEŁ SOBCZAK**

Department of Transport and Computer Science
WSB University, Cieplaka 1c
41-300 Dąbrowa Górnicza, Poland
*pbuchwald@wsb.edu.pl
**psobczak@wsb.edu.pl

(received: 13.09.2021; revised: 29.10.2021;
accepted: 16.11.2021; published online: 30.12.2021)

Abstract: The study concerns the possibility of using the Neo4j database as a graph analysis tool. The analysis was presented on the example of the railway connection network assessment when designing a new version of the railway infrastructure in the Silesian Voivodeship. The authors present the concepts of a laboratory environment built on the basis of the Neo4j database, and an application that allows obtaining data and modifying the railway infrastructure model. Thanks to this, it is possible to simulate many variants of the designed model and its evaluation using the proposed measurement indicators. The study presents methods and tools of analysis that can be successfully used to assess the topology of a graph in many different research areas, such as analysis of management systems, computer networks or biological systems.

Keywords: Neo4j, graph analysis, Cypher, .Net, railway connections, transport, logistics

DOI: https://doi.org/10.34808/tq2021/25.4/d

1. Analysis of metropolitan rail connections in the Silesian Voivodeship

Currently, road transport is the most popular and easiest to implement. Unfortunately, this is mainly individual road transport (the average level of vehicle load in Poland, despite a number of promotional activities e.g. Carsharing, is only about 1.3 people per vehicle [1], the vast majority of vehicles are used only by the driver [2]. Bus transport is the second most popular means of transport in Poland, also in highly urbanized areas. Unfortunately, this causes a number of unfavorable phenomena related to this mode of transport. The most important of them are:

- noise
- serious environmental pollution
- occupying large areas for the necessary infrastructure
- health effects and costs of road accidents (including medical costs or deaths from road accidents and collisions).

One method of mitigating these negative effects is to create more sustainable transport using several modes of transport, including rail transport. The implementation of sustainable transport (including public transport) is one of the most important elements of activity in urban areas. As is stated in many scientific publications, the infrastructure (including transport) has also a significant impact on the development of an area or a region [3-6]. It is especially important and essential in the case of highly urbanized areas. Transport performs a number of functions. The most important of them are services that enable business activities of many enterprises, and services for the public. In connection with the above, steps have been taken to create the so-called Metropolitan Railway. The concept of metropolitan rail connections must take into account a number of criteria (including cost criteria, environmental impact, accessibility to residents). An additional element worth analyzing, being the main goal and element of this article, is the analysis of connections in terms of structure with the use of the graph theory and modern data processing tools oriented on the graph ontology. Such analysis helps in the detection of key connection nodes and connection networks, which are the most important elements in terms of the proper functioning of the movement of people and goods by rail transport.

2. Research methods using graph theories in analysis of connections between railway network nodes

The study used the computer simulation method based on the graph theory. This theory has been successfully applied for many years to the analysis of management systems [7], computer networks [8] and biological systems [9, 10]. Analysis based on graph notation can be effectively used to assess transport networks [11]. A very important element of analysis of a network of connections with the use of the graph theory is the ability to perform both functional and organizational analysis. As part of analysis, a number of factors are calculated that allow assessment of both individual network nodes (which in the case of the rail transport network are stops located on this network), but also allow analysis of the network as a whole. The analysis allows obtaining information about the role of individual nodes in the network (whether it is a local role, or maybe the role of a node constituting the network center) and, importantly, this analysis allows obtaining estimated information on the resistance to possible disturbances (both unintentional in the form of accidents) and intentional (e.g. sabotage or terrorist attack) [12]. As part of the research, a structural analysis of the connection network was carried out, presenting at the same time the possibilities of the non-relational Neo4j database in the presented range of applications. The architecture of railway connections for the Silesian agglomeration was used to carry out analyses for the needs of the research. It is not a real architecture, but one of the variants proposed for implementation in the future. For this reason, the graph shows the actual railway junctions as well as the proposed ones that are likely to be implemented during the construction of a new version of the metropolitan railway. The graph used to verify the methods and research tools presented in the study is shown in Figure 1.

3. Neo4j database as an example of a data analysis tool in the form of a graph

The NoRel trend which allows data representation using other models has appeared in recent years in addition to the relational model dominating in database systems. Such databases are characterized by a high time efficiency in processing data operations [13]. The graphical representation of the database is based on two concepts: entity (otherwise object, node) and relationships (connections, edges). An entity represents a single node. A relationship is a property that can occur between entities. Entities and relationships can have their own attributes. Speaking in the language of the graph theory, a model is a labeled and directed multigraph with attributes. A labeled graph has edge type labels. The directed graph includes edges with a specific direction from a source node to a destination node. The attribute graph allows assigning a variable list of attributes to each node and each edge. The attribute is a name-related value. A multigraf can contain multiple edges between two nodes. This means that two nodes can be joined multiple times over different edges, even if the edges have the same source node, destination node, and label. Databases that use this way of representing entities and the relationship between them are called graph databases. Due to the use of a data model that differs from the relational data model, graph databases also introduce the query and data modification languages characteristic of the graph model. They are the same as SQL for databases that use relational form. The most popular languages used in graph databases are:

- Gremlin - a graphical programming language. It has built-in mechanisms for graph queries, graph analysis and manipulation. Its service is provided, inter alia. by Neo4j;
- SPARQL - allows querying a variety of data sources where the data is stored in RDF. The result of SPARQL queries is a set of RDF graphs;
- G, G +, GraphLog - allow creating graph queries. The graph query in the case of the G language is a set of tagged directed multigrafs;
- G-Log - contains a declarative language for complex objects. It uses the logical rule satisfaction notation to evaluate the query response.

Neo4j provides native API containing classes and methods that enable programmatic navigation around the graph, i.e. traverse. With this, search methods can be defined, e.g. in depth or across. In addition, a stop condition can be specified.

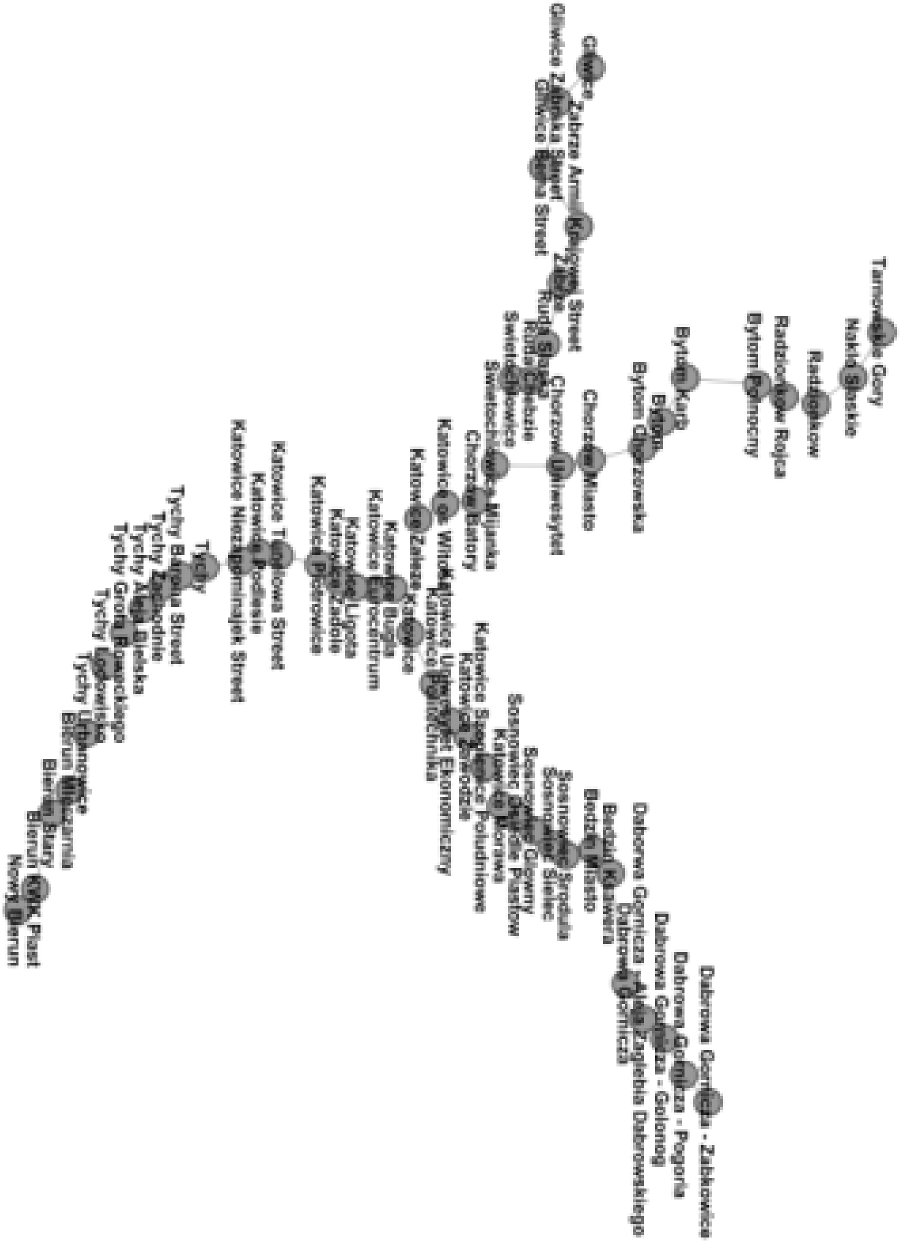


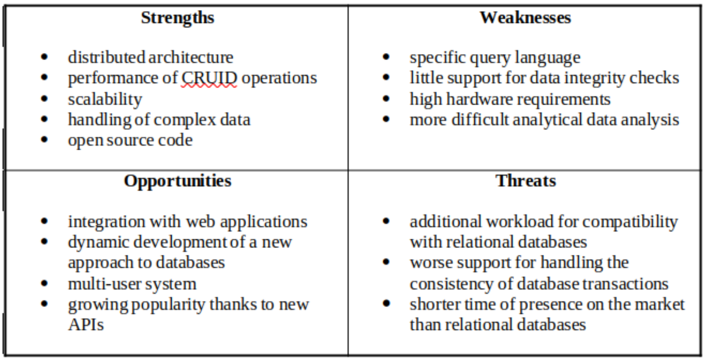
Figure 1. Example of a Railway connections graph

Communication with the database can take place using the REST model. It provides a certain pattern of URLs and their arguments that allow navigating through the space of nodes and dependencies. CRUD operations consist in sending an HTTP request to the appropriate address. Database creators recommend using this API with additional security mechanisms against unauthorized access, as the queries trigger a dynamically generated Groovy code. Access via HTTP makes Neo4j a very universal solution.

The Neo4j database also supports scalability mechanisms[14]. In the case of large graphs, it is not possible to store data on one machine due to the size of the graph [16]. This problem can be eliminated by:

- Grouping queries between machines so that they are executed only at the end of each first width step;;
- Storing multiple levels of relationship associations;;
- Limitation of the search depth for nodes between machines.;

Table 1. SWOT analysis of the Neo4j database application



Strengths	**Weaknesses**
• distributed architecture • performance of CRUID operations • scalability • handling of complex data • open source code	• specific query language • little support for data integrity checks • high hardware requirements • more difficult analytical data analysis
Opportunities	**Threats**
• integration with web applications • dynamic development of a new approach to databases • multi-user system • growing popularity thanks to new APIs	• additional workload for compatibility with relational databases • worse support for handling the consistency of database transactions • shorter time of presence on the market than relational databases

NoSQL databases from the Neo4j family have their advantages and disadvantages. The SWOT analysis of the application of these solutions is presented in Table 1. Additionally, the advantage of using the Neo4j database is the possibility of calculating the coefficients of graph evaluation measures, which, through the prism of the problem area, can be indicators of the analyzed solution evaluation for the selection of railway connections.

4. Graph assessment measures used in analysis of the rail connection graph

Descriptors used in the detection of both nodes and edges of the graph, important for maintaining the efficiency of connections were selected for analysis of the rail connection graph. The selection of appropriate indicators was also dictated as it was possible to determine them on the basis of the implemented libraries for the Neo4j database [15]. The analysis of the rail connection graph was carried out on the basis of the following factors:

- Betweenness centrality

  In the graph theory, the betweenness centrality is a measure of the centrality of a graph based on shortest paths. For example, in a telecommunications network a node with a higher intermediary centrality would have more control of the network as more information would pass through that node.

In the analysis of railway connections, a node with a higher value of the coefficient will in many cases be an intermediate node between the beginning and the end of a route. The method of determining the value of this coefficient in the Neo4j database is shown in Listing 1.

```
CALL gds.betweenness.stream('myGraph')
YIELD nodeId, score
RETURN gds.util.asNode(nodeId).name AS name, score
ORDER BY name ASC
```

Listing 1. Determining the betweenness centrality ratio in Neo4j

- PageRank

It was initially used by Google to position websites depending on the flow of links. Web links are treated as a Markov chain in which a ”random user” is simulated. Each incoming edge is seen as a ”vote” for the website. Websites getting more votes are interpreted as ”Hubs” on the network, while those with more outgoing votes are considered ”Authorities”. The center and authority results for each node are updated recursively until either the maximum iteration limit is reached or the results converge. Determining the PageRank is possible on the basis of the following equation:

$$PR(A)=(1-d)+d\left(\frac{PR(T_1)}{C(T_1)}+\ldots+\frac{PR(T_n)}{C(T_n)}\right),$$

where it is assumed that side A is linked to parties T_1 to T_n which point to it, while d is a damping factor which can be set from 0 (inclusive) to 1 (exclusive). Usually it is set to 0.85. $C(A)$ is the number of links coming from page A.

```
CALL gds.pageRank.stream('myGraph')
YIELD nodeId, score
RETURN gds.util.asNode(nodeId).name AS name, score
ORDER BY score DESC, name ASC
```

Listing 2. Calculation of betweenness centrality

As in the case of Internet sources, many references to a given source indicate its high popularity, and in the case of railway nodes, many connections leading to it may indicate the importance of this node [18].

- Authority and hubs

An indicator used in various research fields, such as finance and analysis of currency flows between angles, and in the analysis of stock transactions. Similarly, in the case of important railway nodes, the value of this coefficient can be considered as an indicator of the attractiveness of individual stations in terms of the number of incoming and outgoing connections. The code showing how to calculate this coefficient using the Neo4j database is presented in Listing 3.

```
CALL gds.alpha.hits.stream('myGraph', {hitsIterations: 20})
YIELD nodeId, values
RETURN gds.util.asNode(nodeId).name AS Name, values.auth AS auth, values.hub as hub
ORDER BY Name ASC
```

Listing 3. Determining the authority and hubs factor.

5. Illustrative diagram of building data acquisition application for analysis

The publicly available data on passenger train timetables was used to create a graph of railway connections. The data was obtained from websites with passenger train timetables.

The analyzed network of connections was modified manually. Therefore, the examined graph does not fully reflect the real infrastructure, but was built on the basis of data on real connections. Additionally, processing of websites was required to create a coherent database on the basis of which it would be possible to create a graph. Unfortunately, none of the rail transport companies provides an API interface or data in the form of an easy-to-analyze data set in formalized notation, such as XML or a relational database. For this reason, the data was obtained with the help of the created tool allowing the website interpretation and the processing of the obtained data into a graph database. The data ingestion application was implemented in c using the Visual Studio environment. This language was used due to the availability of high-level libraries, which on the one hand, enable communication with the Neo4j database, and on the other hand, simplify access to data made available in the form of websites.

The modular architecture of the created programming environment for analyzing railway connections is presented in Figure 2.

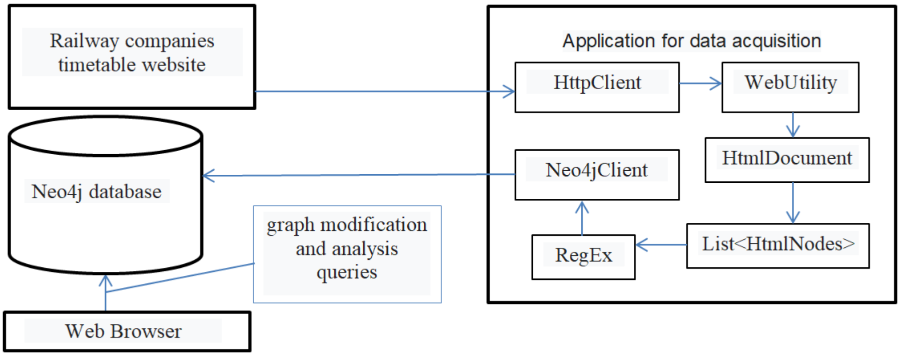

Railway companies timetable website
Application for data acquisition
HttpClient
WebUtility
HtmlDocument
Neo4jClient
Neo4j database
RegEx
List<HtmlNodes>
graph modification and analysis queries
Web Browser

Figure 2. Illustrative diagram of the application architecture for data analysis

An object of the HttpClient class was used for the automatic connection with the website providing the railway timetable. It supports communication with a remote server using http and https protocols. Additionally, it has the ability to handle user authorization and authentication. This gives the ability to access websites that require a username and a password, or authorization methods with a certificate. Data interpretation at the level of a website created in HTML was realized by the WebUtility class. It allows the interpretation of the incoming data stream together with the received response to the http protocol query as a website in the form of an HTML code. The very processing of website data saved in the form of HTML tags is quite laborious, therefore the HtmlDocument class was used, which facilitates this task. It allows accessing the markup of the analyzed page as one-way list objects. Due to the repeatability of the graphic interface of the timetable visualization, the RegEx class was used to read individual connections. It allows handling regular expressions in c. Based on the data obtained in this way, classes were created for individual railway nodes and the edges of connections between them. Such data was saved on the Neo4j database side using the Neo4jClent class. It is a class that allows cooperation with a local database and a database shared in a distributed architecture (e.g. in a computing cloud). It uses data access interfaces based on the REST software model. After creating a graph in the Neo4j database, it is possible to use the access to the Neo4j database via a web browser, it is possible to explore the connection graph using query languages native to graph databases.

Saving data in the form of a graph in the Neo4j database allows a simple modification of the existing connection infrastructure. It is possible to run alternative connections and delete the existing ones. This allows reflecting incidents related to the shutdown of a railway junction, which may occur for random or planned reasons.

6. Analysis results

As a result of the conducted analyzes, information was obtained on the properties and topography of the Metropolitan Railway network. The conducted analyses showed that the obtained network of connections was a free-scale network. Networks of this type are characterized by a relatively high global resistance to random disruptions (e.g. failures), as there is a high probability that the failure will occur in a smaller network node and will not affect the functioning of other nodes. Unfortunately, this type of network is very sensitive to intentional disruptions (e.g. terrorist attacks) if it occurs especially at the main network nodes. The conducted analyses have shown that there are 2 such main nodes for the analyzed network: the Katowice and Świętochłowice Mijanka interchanges. The Katowice junction is located in the very center of the agglomeration. Importantly, after the initial stage of using the existing railway line for the implementation of metropolitan railway connections, an additional railway line dedicated to the Metropolitan Railway is to be built in this area. This should largely make the network immune to any disturbances in this area (the multitude of tracks on this section will allow an emergency temporary redirection of traffic), therefore, it is highly advisable to create additional lines in this area. Obviously, this does not protect the network from a complete paralysis of the node. The data reflecting the significance of railway junctions on the basis of the determined values of the coefficients used to evaluate the graph are presented in Table 2.

Table 2. Calculated parameters of the network nodes

Node (Train Station) Name	Betweenness centrality	Authority	Pageranks
Katowice	0.672932	0.387368	0.023253
Swietochlowice Mijanka	0.489975	0.386459	0.023312
Katowice Zaleze	0.473684	0.291266	0.016094
Katowice os Witosa	0.463659	0.291077	0.016267
Chorzow Batory	0.452381	0.290875	0.016108
Katowice Bugla	0.43985	0.241726	0.016279
Katowice Politechnika	0.394737	0.241725	0.016279
Katowice Eurocentrum	0.426065	0.175682	0.0167
Katowice Uniwesytet Ekonomiczny	0.377193	0.175681	0.016701
Katowice Ligota	0.411028	0.109629	0.016945
Chorzow Uniwesytet	0.270677	0.241106	0.016313
Katowice Zadole	0.394737	0.079677	0.017072
Katowice Zawodzie	0.358396	0.109628	0.016947
Swietochlowice	0.219298	0.240919	0.016351
Katowice Piotrowice	0.377193	0.04972	0.017152
Chorzow Miasto	0.245614	0.175152	0.016756
Katowice Szopienice Poludniowe	0.338346	0.079674	0.017077
Katowice Tunelowa Street	0.358396	0.036136	0.017189
Katowice Morawa	0.317043	0.049717	0.017157
Ruda Chebzie	0.191729	0.174715	0.016847
Katowice Podlesie	0.338346	0.022549	0.017219
Katowice Niezapominajek Street	0.317043	0.016388	0.017231
Sosnowiec Osiedle Piastow	0.294486	0.03613	0.017203
Bytom Chorzowska	0.219298	0.10919	0.017026
Tychy	0.294486	0.010226	0.017249
Sosnowiec Glowny	0.270677	0.022542	0.017237
Tychy Barona Street	0.270677	0.00743	0.017259
Ruda Slaska	0.162907	0.108504	0.0172
Bytom	0.191729	0.079185	0.017225
Sosnowiec Sielec	0.245614	0.016376	0.017272
Tychy Zachodnie	0.245614	0.004635	0.017287

Table 2 – continued.

Sosnowiec Srodula	0.219298	0.010209	0.017311
Tychy Aleja Bielska	0.219298	0.003365	0.017317
Bytom Karb	0.162907	0.049176	0.017409
Zabrze	0.132832	0.078023	0.017548
Bedzin Miasto	0.191729	0.007404	0.017387
Tychy Grota Roweckiego	0.191729	0.002096	0.017392
Bytom Polnocny	0.132832	0.035362	0.017667
Bedzin Ksawera	0.162907	0.004598	0.017496
Tychy Lodowisko	0.162907	0.001516	0.017498
Zabrze Armii Krajowej Street	0.101504	0.04754	0.017988
Dabrowa Gornicza	0.132832	0.003307	0.017718
Tychy Urbanowice	0.132832	0.000936	0.01772
Radzionkow Rojca	0.101504	0.021546	0.018054
Daborwa Gornicza - Aleja Zaglebia Dabrowskiego	0.101504	0.002015	0.018081
Gliwice Bema Street	0.068922	0.032712	0.018712
Bierun Mleczarnia	0.101504	0.000664	0.018081
Radzionkow	0.068922	0.014825	0.018753
Dabrowa Gornicza - Golonog	0.068922	0.001386	0.018771
Bierun Stary	0.068922	0.000392	0.018772
Gliwice Zabrska Street	0.035088	0.017882	0.019932
Naklo Slaskie	0.035088	0.008104	0.019957
Dabrowa Gornicza - Pogoria	0.035088	0.000758	0.019967
Bierun KWK Piast	0.035088	0.00025	0.019967
Gliwice	0	0.008941	0.011063
Tarnowskie Gory	0	0.004052	0.011075
Dabrowa Gornicza - Zabkowice	0	0.000379	0.01108
Nowy Bierun	0	0.000107	0.01108

7. Summary

The conducted research and analyzes allowed obtaining information on the network of railway connections which is to play an important role in ensuring efficient and effective transport in one of the most densely populated, but at the same time polluted and exposed to the effects of traffic pollution, region of Poland. It is proposed to perform similar analyzes using the graph theory for other new investments of this type in Poland and Europe - similar activities are carried out to analyze the current and planned railway infrastructure in Poland. The presented article shows the possibilities of using graph analysis with the Neo4j tool on the basis of analysis of railway connections in the Silesian metropolis.

Such analysis allows calculating the coefficients of assessing the importance of individual network nodes, but also for simulating changes in the topology and assessing the impact of these changes on the global and local network parameters. There are many tools on the market that allow analyzing graphs, however, due to the properties of the Neo4j database, this solution was selected as flexible and with high scalability [17]. For the analyzed example, the version of the database installed on one computer workstation was used, however, if it is necessary to analyze a network with a larger number of nodes, it is possible to migrate the solution to the cloud computing architecture or add multiple computer nodes to the database. The flexible nature of the Neo4j database in terms of communication interfaces, a wide selection of libraries that allow the database to work with the application layer, and a high time efficiency in graph processing and exploration make this database a perfect tool for analyzing all kinds of issues that can be described in the form of a graph.

References

[1] Website of the social campaign 2021 *"Drive the right lane"*, *siskom.waw.pl/kp-buspas1.html*

[2] Zawisza T and Dębowska-Mróz M 2017 *Assessment of passenger car filling in terms of improving the use of transport space in cities*, *Journal Buses - Technika, Eksploatacja, Systemy Transportowe*

[3] Miller P 2014 *Sustainability and Public Transportation: Theory and Analysis. PhD Thesis.*, University of Calgary, Calgary, Canada

[4] Van Uytven A 2016 *Sustainable Urban Transport Planning (SUTP)*, Eltis the Urban Mobility Observatory

[5] Adell E and Ljungberg C 2014 *The Poly-SUMP Methodology. How to Develop a Sustainable Urban Mobility Plan for A Polycentric Region*, *In European Platform on Sustainable Urban Mobility Plans*, European Commission: Brussels, Belgium

[6] West Midlands Combined Authority West Midlands Metropolitan County 2013 *West Midlands Metropolitan Freight Strategy 2030; Supporting our Economy; Tackling Carbon*, West Midlands Metropolitan County; Birmingham, United Kingdom of Great Britain and Northern Ireland

[7] Buchwald P and Lis M 2019 *Cloud Computing and the Internet of Things in Advanced Planning and Scheduling Systems*, *Proceedings of the 34th International Business Information Management Association Conference (IBIMA), Madrid, Spain*, Education Excellence and Management of Innovations through Sustainable Economic Competitive Advanta. Vision 2025

[8] Buchwald P, Mączka K, Pikiewicz P and Rostański M 2016 *Proceedings of the 11th Scientific Conference Internet in the Information Society 2016*, Scientific Publishing. University of Dąbrowa Górnicza

[9] Sporns O 2002 *Network analysis, complexity, and brain function*, *Complexity* **8** 56

[10] Stam C J and Reijneveld J C 2007 *Graph theoretical analysis of complex networks in the brain*, *Nonlinear Biomed. Phys.* **1**

[11] Sobczak P, Stawiarska E, Oláh J, Popp J and Kliestik T 2018 *Logistics management of the rail connections using graph theory: the case of a public transportation company on the example of Koleje Dolnośląskie S.A*, *Engineering Management in Production and Services* **10** (8)

[12] Bukowski L and Sobczak P 2019 *Resilience assessment of heterogeneous complex transport networks - a general framework and a case study, Proceedings of the 29th European Safety and Reliability Conference (ESREL)*

[13] Buchwald P, Rostański M and Arkadiusz J 2013 *Relative and non-relative databases performance with an Android platform application, Theoretical and Applied Informatics* **25** (3) 223

[14] Baton J and Rik Van Bruggen R 2017 *Learning Neo4j 3.x - Second Edition: Effective data modeling, performance tuning and data visualization techniques in Neo4j*, Packt Publishing

[15] Hodler A E and Needham M 2019 *Graph Algorithms: Neo4j version*, O'Reilly Media

[16] Huang H and Dong Z 2013 *Research on architecture and query performance based on distributed graph database Neo4j, 2013 3rd International Conference on Consumer Electronics, Communications and Networks*

[17] Mueller W Sr, Idziaszek P, Gierz Ł, Przybył K, Wojcieszak D, Frankowski J, Koszela K, Boniecki P and Kujawa S 2019 *Mapping and visualization of complex relational structures in the graph form using the Neo4j graph database, Proc. SPIE 11179, Eleventh International Conference on Digital Image Processing (ICDIP 2019)* 1117924

[18] Page L, Brin S, Motwani R and Winograd T 1999 *The PageRank citation ranking. Bringing order to the web*, Stanford InfoLab

Paweł Buchwald PhD in computer science, database specialties. Scientifically interested in data processing systems, mobile applications and Internet of Things solutions. A research and teaching worker at the WSB Academy in Dąbrowa Górnicza and the Silesian University of Technology in Gliwice. Also professionally involved in software architecture and designing IT systems for industry and production management. Implementer of scientific projects in the areas of ICT security, artificial intelligence, virtual reality and augmented reality.

Paweł Sobczak PhD, WSB Academy. Department of Transport and Informatics. His interests include analysis and optimization of transport networks with particular emphasis on public transport as well as the use of modern information technologies in logistics, with particular emphasis on warehouse processes. Passionate about modern technologies in logistics and transport. Certified Flexsim software consultant: quot;FlexSim - The Business Process Simulation Consultantquot; and quot;The Certification for FlexSim Trainer - University Levelquot;. Certified 1st degree tutor of Collegium Wratislaviense.