Visualization of Dataflows: a Casestudy of COVID-19 Rumors

Mikhail Ulizko [1,2], Evheniy Tretyakov [1,2], Rufina Tukumbetova [1,2], Alexey Artamonov [1,2] and Mikhail Esaulov [2]

[1] *Plekhanov Russian University of Economics, Stremyannyy Pereulok, 36, Moscow, 115093, Russia*
[2] *National Research Nuclear University MEPhI (Moscow Engineering Physics Institute), Kashira Hwy, 31, Moscow, 115409, Russia*

Abstract
One of the most significant and rapidly developing works in the field of data analysis is information flow management. Within the analysis targeted and stochastic dissemination patterns are studied. The solving of such problems is relevant due to the global growth in the amount of information and its availability for a wide range of users. The paper presents a study of dissemination of information messages in open networks on the example of COVID-19. The study was conducted with the use of visual analytics. Informational messages from the largest world and Russian information services, social networks and instant messengers were used as sources of information. Due to the large amount of information on the topic, the authors proposed a pattern of the wave-like dissemination of information on the example of topic clusters on the connection of COVID-19, hydroxychloroquine and 5G. The developed methods can be scaled up to analyze information events of various topics.

Keywords
graph analysis, geospatial analysis, Web-technology, COVID 19, data scrapping, misinformation

1. Introduction

In the digital world, Internet traffic is growing every year [1]. According to various projections, by the end of 2021 Internet traffic will exceed 3 zettabytes (ZB) [2, 3]. At the same time, communication in society also passes into the virtual world [4, 5]. In particular, to obtain information about world events, the user is increasingly using Internet media, social networks and instant messengers [6]. These sources of information promptly react to world events. The exchange of information via the Internet has significantly reduced the time for the delivery of information from the moment of the event to the moment it is received by the consumer / user. However, due to the large amount of data and its heterogeneity, such an environment has become a fertile ground for the dissemination of false information, and in some cases it is generated in larger volumes than true information. False information can form a false point of view, which can lead to the destabilization of society [7].

The paper examines the following sources of information: online media, Twitter and Telegram channels with the following characteristics:

- regularity. It means that a flow of information messages is sent regularly.
- impartiality (excluding Twitter), i.e. without value judgments and distortion of reality.
- citation (including self-citation). It means that sources can rely not only on their own materials, but also refer to other sources.

According to the authors, more than 14000 publications about the coronavirus indexed in WOS and Scopus in 2020 providing sharp rate of publication growth (1600% compared to the previous 5 years) [8, 9]. Authors performed the bibliographic analysis of scientific papers from 2000 to 2020 by existing

visualization tools (for instance, VOSviewer [10, 11]) and noted the most significant terms in this field and emphasized international collaboration as a key for dealing with the COVID-19. Results shows three main clusters for contributing papers: The USA, China and European countries. Meanwhile Belli et al [8] highlights the importance of open science: "Open science is the best method because it is an approach based on collaborative work, openness, and transparency in all stages including not only publication, but also data collection, peer review, and assessment."

Despite the existence of wide variety of software for monitoring coronavirus, a number of authors have proposed new visualization tools [12, 13, 14, 15]. Martínez Beltrán et al. [12] describes developed "web-based, user-friendly dashboard for interactive plotting" which is able to analyze both coronavirus and related data, such as maximum temperature, grocery and farmacy, etc., mostly in Spain. As Marcílio-Jr et al. [13] writes: "Visualization-based strategies for monitoring the dissemination of diseases account for the fact that graphical representations can enhance the ability to identify data patterns and tendencies". So, authors proposed tool for drawing graphics, pie-charts, and especially maps to monitor the evolution of dissemination of coronavirus in Sao Paulo state. Another geospatial techniques are proposed by Mast et al. in their study [14]. Their analysis for the USA allows to reveal vaccinating distribution in two directions: over the time and over the place (cities, states). And finally, Chintala et al. develop software for workflow-based analysis over the world [15]. Their Python's application provides dynamic maps on cumulative daily confirmed COVID-19 cases for different countries.

All the papers below deal with coronavirus data rather than myths and rumors. Actually, there is much fewer publications which consider dissemination true of false information on COVID-19 around the world via the Internet or mass-media. For instance, one study on the topic shows the statistical analyses to separate myths and facts [16]. Authors illustrates several graphs based on 13-item questionnaire of 125 participants. Pang et al. describes it how governmental social media was used during COVID-19 pandemic [17]. Authors carry out word frequency and contents analyses for Macao governmental social media in Facebook and reveal that it can useful to control rumors dissemination. Also Sond et al. (2021) show different types of rumors and how they can be corrected [18]. They perform their analysis of the data on Sina Weibo, the most popular microblogging site in China. Last but not least papers solve the issue of analyzing misinformation to some extent [19, 20, 21], emphasizing methods of text analysis rather than visualization tools.

It ought to be noted that researchers rarely use visual tools for analysis in their work, except when they are looking at data within their own country. This paper tries to combine approaches from mentioned articles and apply them to main rumors over the world.

This paper presents methods of data processing and data visualization, which is conducted with the use of web technologies for building graphs and plotting data on the globe. The methods are examined and applied on the example of rumors about COVID-19. The described methods can be applied to analyze any data with a similar structure.

2. Methodology

With advances in information technology, researchers in various fields began to pay attention to the analysis of streaming data, which refers to data that is generated continuously from various sources. Also, as in the case of static data, the process of their processing can be represented as follows (Figure 1). In this paper, visualization is used for performing an analysis and showing the results.
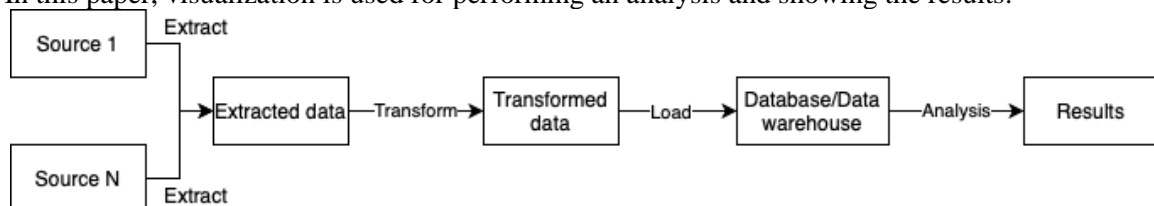

Figure 1: The process of collecting, processing and analyzing data

The paper concerns the following data sources: the world's leading media (CNN, The New York Times, The Verge, etc.), instant messengers (Telegram) and social networks (Twitter). The object of

the study is rumors about COVID-19, which are fake information messages about COVID-19. The data analysis process includes two parts:
- dissemination of rumors and their contradictions between information sources;
- consideration of facts of various degrees of credibility between countries.

2.1. Dissemination of rumors and their contradictions between information sources

Collecting data about rumors related to coronavirus disease comes down to the aggregation of information messages from the mentioned media, instant messengers and social networks using agent technologies [22]. The information message model is based on the collected data. The model consists of the following fields:
- Id;
- Source URL;
- Title of the information message;
- Text of the information message;
- Links to other information sources;
- Date (and time) of posting.

The main feature of such representation is that the objects of this model are connected to each other through the "Links" field. To display such relationships, it is proper to use graph representations of data [23].

Thus, data is represented in the form of a dynamically weighted directed graph, the nodes of which are information sources, edges are information messages, in which one source refers to another. Since information messages differ from each other by the time of posting, the graph is dynamic, that is, it can change its state over time, i.e. new nodes and edges appear over time (eg. the graph is rebuilt at those moments when the information source posts a rumor). The weighting of the graph is applied to both edges and nodes: the weight of a node corresponds to the number of links to it, and the weight of an edge corresponds to the number of links between its initial and final nodes (all information messages are considered to be equal).

The graph consists of two types of edges (color legend of edges) since in the final samples an information message can be either a fake message or a contradiction to a fake message. To determine the type of edge, the text of an information message is analyzed. It is compared with the thesaurus of words, which have a meaning of contradiction, to calculate the value of the criterion. If the value of the criterion exceeds a certain threshold, then the message is recognized as a contradiction.

2.2. Dissemination of facts of varying degrees of credibility between countries

The CoronaVirusFacts/DatosCoronaVirus Alliance Database from the Poynter information resource was used as a data source [24]. The database includes verified facts in over 70 countries and articles published in at least 40 languages. Since the database records are assessments of information messages stored on different web pages, the entire database was collected using scraping. Only records with one of the following credibility levels (50 degrees in total) were selected:
- False;
- Misleading;
- Mostly false;
- No evidence;
- Partially false.

After collecting the data, preliminary data processing was carried out: unification of the countries to which the record belongs, highlighting the date, collecting the link to the primary source of information. To visualize the dissemination of rumors between countries, a model is built, which consists of the following fields (Figure 2):

- category/title of rumor;
- a description of information message;
- a country in which the rumor was spread;
- a source of a message (media, Facebook, etc.);
- a degree of credibility.

Fact-checked by: Newschecker

2021/06/28 | India

MISLEADING: **The third wave of Corona has started wreaking havoc in Delhi and Madhya Pradesh. Four people have died from the 'Delta Plus variant'. Cases of the Delta Plus variant have been reported in Maharashtra, Delhi, Kerala, and Madhya Pradesh. It is also being said in the viral post that those who died of this variant had received both doses of the vaccine.**

Explanation: We found that the third wave of the corona has not arrived in the country yet. During the investigation, we found that cases of 'Delta plus variant' have been found in some states of the country, but on the basis of some cases, it cannot be said that the third wave of the corona epidemic has arrived in the country.

READ THE FULL ARTICLE (NEWSCHECKER)

Figure 2: An example of the record

3. Visualization and data analysis

3.1. Dissemination of rumors and their contradictions between information sources

It is customary to build a graph with the help of such tools as Gephi, igraph, LargeViz, etc., or using additional libraries for programming languages. For dynamic and interactive work with graphs 3D Force-Directed Graph [25], a component for visualizing graphs was chosen, which uses ThreeJS/WebGL for rendering and force-directed graph drawing algorithms to build a graph.

To analyze information sources, the data is classified into separate rumors. The paper considers information messages on the topics '5G' and 'hydroxychloroquine'. The graph with '5G' rumors is represented below (Figure 3).

The graph was built according to the following rules:
- nodes of the graph are information sources;
- if two nodes have an edge, which connects them, it means that the two sources are interconnected by an information message, in which one information source refers to another (henceforth such nodes will be called adjacent or connected);
- the graph is oriented. The direction of an edge is shown by the motion of the ball along the edge from the source of the information message to the link;
- a size of a node is proportional to the number of posts associated with this information source (both as a source and as a link);
- the graph is displayed using the force-directed graph drawing algorithm;
- an edge can belong to one of the two types: without contradiction (green edges) and with refutation (red edges);
- the nodes are color-coded: without selecting an active node, nodes have colors according to the color bar (Figure 4); when an active node is selected, it and all nodes adjacent to it become orange.
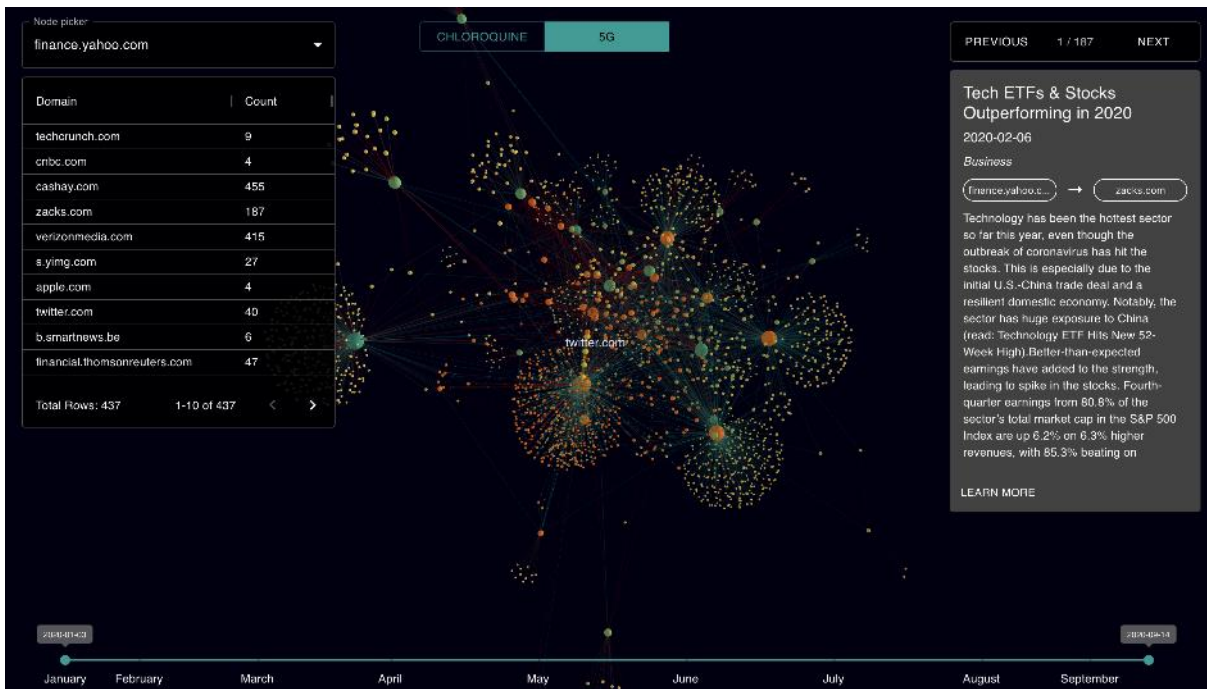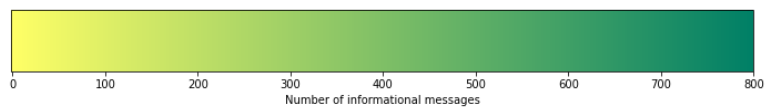
Figure 3: Graph web interface

Figure 4: Color bar

The interface of the constructed web application makes it possible to analyze rumors on a selected topic by selecting individual nodes, choosing a time interval and considering information messages between two nodes connected by an edge.

This approach allows us to identify the primary sources of the dissemination of false information and the main distribution nodes. For instance, the rumor about chloroquine shows that "Global Banking & Finance Review" released has released only rumors with refutations (Figure 5). On the other hand, the figure (Figure 6) represents the possibilities of the graph. According to the graph (Figure 5a), "MarketBeat" was the first resource which published information about connection between 5G and coronavirus, but with time its contribution to the total share of messages decreased. Also, analysis of the data using the graph revealed that most often information sources refer to Twitter (as evidenced by the maximum total input flow and the location of the node in the center of the graph), followed by "Verizon Media". The graph also shows that the most active source of informational messages is "Yahoo! Finance", "The Conversation", etc. For instance, in this time interval "The Conversation" released most of the messages in May and then reduced the intensity.
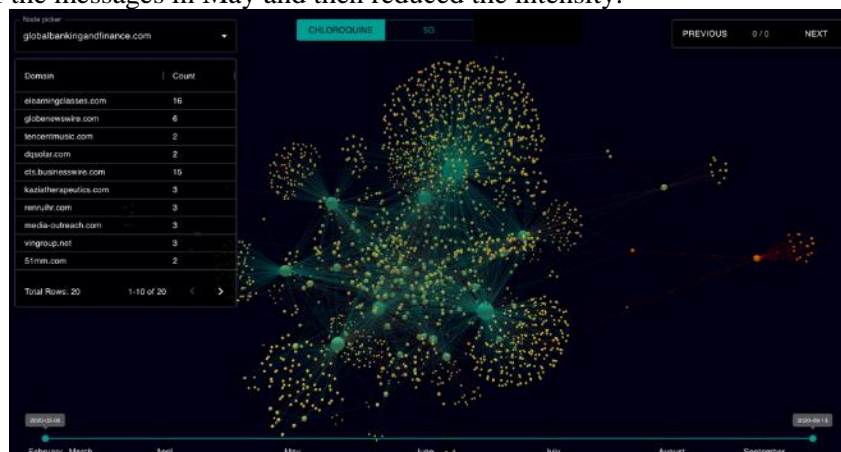

Figure 5: Rumor "chloroquine"

Figure 6: Graph analysis

3.2. Dissemination of facts of varying degrees of credibility between countries

To analyze the dissemination of facts of various degrees of credibility between countries, a proper method is to plot them on the world globe. The authors used the Globe.GL web component [26], which enables applying data visualization layers to a three-dimensional globe. The final form of the web application is demonstrated in Figure 7.

The globe has an intuitive interface: the fewer information messages related to the country, the darker the country looks and the closer it is to the globe; and vice versa: the more information messages related to the country, the more reddish the country and the higher it is raised above the globe. The web interface enables analyzing data by creating a query. It may include:

- a word or phrase;
- a type of rumors (7 main rumors and all the rest);
- a country to which a rumor belongs;
- a source of an information message;
- a degree of credibility of a rumor;
- time range.

The search query tool together with the user-friendly interface enables carrying out sufficient research. The analysis revealed that most of the rumors about the coronavirus came from India and the United States, while information is spreading in neighboring countries.

This approach allows us to identify the hidden relationships between countries. For example, when analyzing rumors connected with the United States, it can be seen that there are similar rumors in Spain (35 common rumors), France (20 common rumors), Canada (12 common rumors) and Ukraine (9 common rumors). This suggests that there are close connections between these countries.
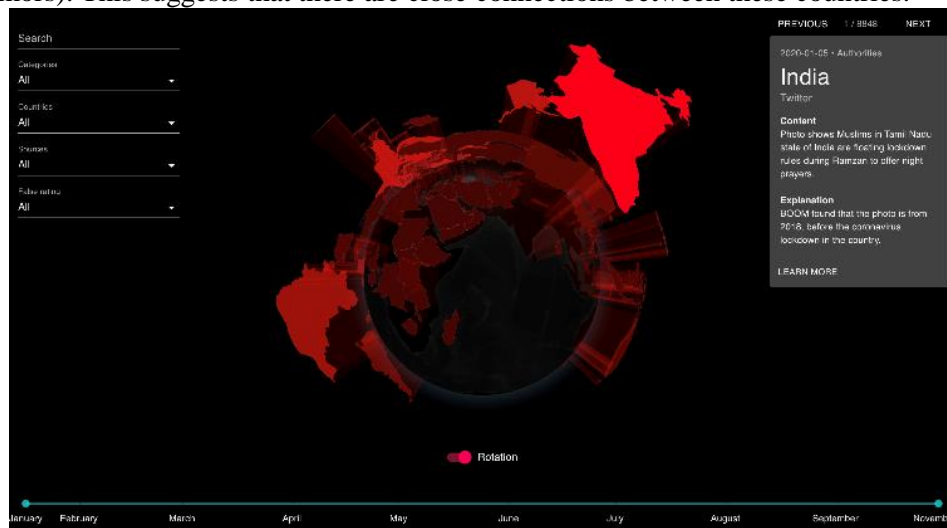

Figure 7: Globus web interface

4. Conclusion

Even though various objects from physical installations to social networks and Internet media can be a source of streaming data, the analysis of such sources can be carried out using similar methods and tools. The paper considers the task of visualizing the dissemination of rumors about coronavirus disease in online media and between countries. The key feature of this study is the statistical analysis of the dynamic system, so with the help of visualization, the spread of information and assess the intensity of information dissemination both locally in the country and around the world is explicitly shown.

A dynamic graph was built to analyze the dissemination of rumors on 5G and hydroxychloroquine in the world media. The graph enables identifying the most "important" nodes (information sources that produce a lot of information messages), considering the process of dissemination rumors over time and visually determine a cluster structure of objects.

To analyze the global situation about coronavirus disease according to rumors, a globe with a search engine was built, which demonstrates the spread of rumors between countries. With the help of this application, the countries in which the rumors appeared most often and their influence on the rest of the world have been identified.

According to the authors, the obtained data model together with the described tools may serve as a good basis for analyzing streaming data of various nature. Further research will be devoted to improving

the accuracy of the model due to natural language processing and expanding the possibilities for statistical analysis.

5. References

[1] Cisco.Com, Cisco Annual Internet Report (2018–2023) White Paper, 2018. URL: https://www.cisco.com/c/en/us/solutions/collateral/executive-perspectives/annual-internet-report/white-paper-c11-741490.html.

[2] Cisco.Com, VNI Complete Forecast Highlights, 2017. URL: https://www.cisco.com/c/dam/m/en_us/solutions/service-provider/vni-forecast-highlights/pdf/Global_2022_Forecast_Highlights.pdf.

[3] The World Bank, World Development Report 2021, 2021. URL: https://wdr2021.worldbank.org/stories/crossing-borders.

[4] W.-L. Shiau, Y.K. Dwivedi, H.S. Yang, Co-citation and cluster analyses of extant literature on social networks, International Journal of Information Management 37(5) (2017) 390-399. doi: 10.1016/j.ijinfomgt.2017.04.007.

[5] W.-L. Shiau, Y.K. Dwivedi, H.-H. Lai, Examining the core knowledge on facebook, International Journal of Information Management 43 (2018) 52-63. doi: 10.1016/j.ijinfomgt.2018.06.006

[6] S. Athey, M. Mobius, J. Pal, (2017) "The Impact of Aggregators on Internet News Consumption." Stanford University Graduate School of Business Research Paper No. 17-8.

[7] A. Gruzd, M. De Domenico, P.L. Sacco, S. Briand, Studying the COVID-19 infodemic at scale, Big Data and Society 8(1). doi: 10.1177/20539517211021115.

[8] S. Belli, R. Mugnaini, J. Baltà, E. Abadal, Coronavirus mapping in scientific publications: When science advances rapidly and collectively, is access to this knowledge open to society? Scientometrics 124(3) (2020) 2661-2685.

[9] J.K. Pal, Visualizing the knowledge outburst in global research on COVID-19. Scientometrics 126 (2021) 4173–4193.

[10] A.Kh. Khakimova, O.V. Zolotarev, M.A. Berberova, Coronavirus infection study: Bibliometric analysis of publications on COVID-19 using PubMed and Dimensions databases, Scientific Visualization 12(5) 112-129. doi: 10.26583/SV.12.5.10.

[11] VOSViewer.Com, VOSViewer, 2021. URL: https://www.vosviewer.com.

[12] E.T. Martínez Beltrán, M. Quiles Pérez, Pastor-Galindo et al. COnVIDa: COVID-19 multidisciplinary data collection and dashboard. Journal of Biomedical Informatics 117 (2021) 103760.

[13] Marcílio-Jr, W.E., Eler, D.M., Garcia, R.E., Correia, R.C.M., Rodrigues, R.M.B. Visual analytics of COVID-19 dissemination in São Paulo state, Brazil. Journal of Biomedical Informatics 117 (2021) 103753.

[14] T.C. Mast, D. Heyman, E. Dasbach, et al. (2021) Planning for monitoring the introduction and effectiveness of new vaccines using real-word data and geospatial visualization: An example using rotavirus vaccines with potential application to SARS-CoV-2. Vaccine: X, 7, 100084.

[15] S. Chintala, R. Dutta, D. Tadmor, COVID-19 spatiotemporal research with workflow-based data analysis, Infection, Genetics and Evolution 88 (2021) 104701.

[16] K. Konar, N. Kabli, A statistical analysis on Covid-2019 to distinguish between myths and facts with data visualization. IOP Conference Series: Materials Science and Engineering 1022(1) (2021) 012043.

[17] P.C.-I. Pang, Q. Cai, W. Jiang, K.S. Chan, Engagement of government social media on facebook during the COVID-19 pandemic in Macao. International Journal of Environmental Research and Public Health 18(7) (2021) 3508.

[18] Song, Y., Kwon, K.H., Lu, Y., Fan, Y., Li, B. The "Parallel Pandemic" in the Context of China: The Spread of Rumors and Rumor-Corrections During COVID-19 in Chinese Social Media. American Behavioral Scientist (2021).

[19] G.K. Shahi, D. Nandini, FakeCovid--A multilingual cross-domain fact check news dataset for COVID-19, arXiv preprint arXiv:2006.11343. doi: 10.36190/2020.14.

[20] P. Patwa, S. Sharma, S. Pykl, et al., Fighting an Infodemic: COVID-19 Fake News Dataset, Communications in Computer and Information Science 1402, 21-29. doi: 10.1007/978-3-030-73696-5_3.

[21] P. Mookdarsanit, L. Mookdarsanit, The covid-19 fake news detection in thai social texts, Bulletin of Electrical Engineering and Informatics 10(2), 988-998. doi: 10.11591/eei.v10i2.2745.

[22] M. Ulizko, L. Pronicheva, A. Artamonov, R. Tukumbetova, E. Tretyakov, Complex Objects Identification and Analysis Mechanisms, Advances in Intelligent Systems and Computing 1310, 517-526. doi: 10.1007/978-3-030-65596-9_63.

[23] M.S. Ulizko, E.V. Antonov, A.A. Artamonov, R.R. Tukumbetova, Visualization of graph-based representations for analyzing related multidimensional objects, Scientific Visualization 12(4) 133-142. doi: 10.26583/sv.12.4.12.

[24] Poynter.Org, The CoronaVirusFacts/DatosCoronaVirus Alliance Database, 2020. URL: https://www.poynter.org/ifcn-covid-19-misinformation.

[25] Github.Com, 3D Force-Directed Graph, 2020. URL: https://github.com/vasturiano/3d-force-graph.

[26] Github.Com, Globe.GL, 2020. URL: https://github.com/vasturiano/globe.gl.