Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Elsevier hereby grants permission to make all its COVID-19-related research that is available on the COVID-19 resource centre - including this research content - immediately available in PubMed Central and other publicly funded repositories, such as the WHO COVID database with rights for unrestricted research re-use and analyses in any form or by any means with acknowledgement of the original source. These permissions are granted for free by Elsevier for as long as the COVID-19 resource centre remains active.
Special Section on EuroVA 2021

Uncovering chains of infections through spatio-temporal and visual analysis of COVID-19 contact traces

Dario Antweiler a,b,*, David Sessler c, Maxim Rossknecht c, Benjamin Abb c, Sebastian Ginzel a, Jörn Kohlhammer c,d

a Fraunhofer IAIS, Schloss Birlinghoven, Sankt Augustin, 53757, Germany
b Fraunhofer Center for Machine Learning, Schloss Birlinghoven, Sankt Augustin, 53757, Germany
c Fraunhofer IGD, Fraunhoferstraße 5, Darmstadt, 64283, Germany
d TU Darmstadt, Karolinenpl. 5, Darmstadt, 64289, Germany

A R T I C L E   I N F O

Article history:
Received 12 November 2021
Received in revised form 20 May 2022
Accepted 24 May 2022
Available online 26 May 2022

Keywords:
Visual analytics
Healthcare information systems
Public health
SARS-CoV-2 pandemic
Corona virus

A B S T R A C T

A major challenge for departments of public health (DPHs) in dealing with the ongoing COVID-19 pandemic is tracing contacts in exponentially growing SARS-CoV-2 infection clusters. Prevention of further disease spread requires a comprehensive registration of the connections between individuals and clusters. Due to the high number of infections with unknown origin, the healthcare analysts need to identify connected cases and clusters through accumulated epidemiological knowledge and the metadata of the infections in their database. Here we contribute a visual analytics dashboard to identify, assess and visualize clusters in COVID-19 contact tracing networks. Additionally, we demonstrate how graph-based machine learning methods can be used to find missing links between infection clusters and thus support the mission to get a comprehensive view on infection events. This work was developed through close collaboration with DPHs in Germany. We argue how our dashboard supports the identification of clusters by public health experts, discuss ongoing developments and possible extensions.

© 2022 Elsevier Ltd. All rights reserved.

1. Introduction

Public health departments (DPHs) are facing great challenges in obtaining and learning from available data on the spread of SARS-CoV-2, connecting it to other sources of data and analyzing it regarding ever-presenting questions on how to proceed in combating the virus. Since the outbreak of SARS-CoV-2, DPHs have faced a plethora of data that they need to input, compare, consolidate, and recall, yet they lack both the means and the personnel to do so: public health officers (PHOs) in Germany must search for contact information or compare lists in their computer system, which sometimes requires additional effort to find specific individuals and their contacts, and to compare this information. Rapid identification and interruption of infection chains through contact tracing plays a key role in responding to the current pandemic and most other infectious diseases, e.g. the norovirus or tuberculosis. To achieve this, PHOs build and analyze large contact and infection networks, possibly spanning multiple generations of transmissions. By tracing back the source of an infection, individual clusters can be contained and recurring transmission patterns can inform decision making on pandemic responses. However, in practice only around 27% of SARS-CoV-2 cases can be traced back to their source leaving the majority of cases disconnected [1].

This paper reports on technology that was developed as part of an initiative that started during the first wave of the SARS-CoV-2 pandemic (early 2020) and provided several DPHs in Germany with visualization and analytics technologies. The federalism in Germany induced a public health setup with more than 370 regional DPHs and the Robert-Koch-Institute (RKI) as a central, federal health institute. We successfully established collaborations, discussed requirements, and continuously presented interim results with 4 DPHs and the RKI during the course of the project. However, our approach is not dependent on a particular public health organization, but is rather based on data that is routinely collected for contact tracing by DPHs and partly shared with the RKI via their SurvNet@RKI system [2]. In this sense, our approach supports cluster detection that is independent of further technical means (e.g. tracing apps) and without further collecting data beyond the information contained in the DPH databases.

Although our paper focuses on the COVID-19 pandemic as the use case on hand, the dashboard can be applied for other
infectious diseases as well. The two main differences regard the availability of collected data by the DPH on the one hand and the urgency in contact tracing to prevent an overburdening of the healthcare system on the other hand.

We first describe a general scenario that crystallized during the collaboration with experts from Germany’s largest DPH in Cologne. This will motivate our contributions presented in this paper. We then describe related work in the areas of epidemiological visual analytics and link prediction on complex networks. The proposed visual analytics dashboard is then presented and evaluated with an exemplary inspection.

As an extension to our original publication [3], our added contributions are as follows: We complemented the temporal and graph-based dashboard with a geospatial visualization component, displaying user-selected chains of infections on an interactive map.

2. Background

We worked with five PHOs at two DPHs, conducting 6 requirement analyses and feedback sessions across half a year in 2020. The PHOs regularly trace infections in a broader view and time frame, and need an overview of the spread of infections to inform local authorities and help advise the implementation or adaptation of restrictions and quarantine measures. All DPHs observe and trace single infected citizens, but also look at the broader developments and, most challenging, clusters of infections. Due to the low number of infections that can be traced back to the host of infection, the PHOs try to identify clusters through accumulated epidemiological knowledge and the metadata of the infections in their database. This includes the time of infection, the address of the citizen, and the provided contacts that the citizen can remember during the telephone interview. Learning more about the clusters of infection and understanding the infection routes both promise an important insight into the epidemiological dynamic. This in turn drives intervention policies and the possibility to strategically prioritize contact tracing and impose (or lift) restrictions. In many DPH regions the number of infections in care homes were particularly high. Detecting whether infections in such facilities were driven by internal infection chains or were continuously induced from outside is difficult without known infection origins. Therefore, our motivation was to visually provide all the available information in an interactive epidemiological network visualization and a temporal visualization of all contacts within a cluster. In addition, we employed a link prediction algorithm to suggest connections between formerly separated clusters if such a connection can be justified by the data, i.e. the temporal and geographical relation of infected contacts. The combination of the need to visualize and interact with complex data as well as the benefits of identification of possible missing links motivates our use of visual analytics (VA) and machine learning (ML) approaches.

3. Related work

Our work is closely related to the research areas of visual analytics for public health and infectious diseases [4] and link prediction in complex networks.

3.1. Visual analytics for public health

There have been approaches in the past that especially detect and visualize infection clusters. Si et al. examined the data of the SARS outbreaks from 2003 onward and developed a prototype system that tracks the patients’ histories in detail to find contacts and visiting activities [5]. This is a good example for approaches that heavily depend on a fine-grained record about the movement and activities of each infected person and their contacts, which our approach could not rely on. There have been a number of SARS-CoV-2 related publications in the visualization community. For example, Leite et al. use temporal visualization to understand the prediction of different COVID-19 scenarios, though not on the contact level [6]. Similarly, Afzal et al. focus on geographical visualization to help public health officials prepare and exercise response plans in pandemic outbreak scenarios [7]. In comparison, our spatial visualization module not only displays information on a single geographical unit, but also regarding inter-unit connections. Various visualization approaches included a spatial visualization of infections [8,9], mostly through choropleth maps [10]. None of the visualizations we analyzed combined graph view, temporal view, and geographic view in one analytical dashboard.

Other approaches take a more general epidemiological approach and address emergent pandemics [11], the role of social networks in epidemiology [12,13], or rather epidemic management [14,15]. Epidemic management and immediate outbreak control are also the focus of SORMAS [16] (originally developed for Ebola outbreaks), which is used in some German DPHs in addition to SurvNet®/RKI, but not for the interactive detection of infection clusters with link predictions.

There is a wider range of possibilities for tracking patients and infections as well as finding clusters of infections in hospitals and similar confined environments. The project HiGHmed focused one of its use cases on infection control in hospitals to elicit transmission paths of multi-resistant bacteria within and between clinics [17,18]. Their literature analysis confirms that there is only little related work on the visual analysis of disease spread in dynamic networks. While the HiGHmed pathway interface also combines network and temporal visualizations, it does not enable the detection of new infection clusters. Instead, our network visualization shows all infection sub-networks, including predicted links, and a temporal visualization that is focused on the characteristics of the COVID-19 viral transmission and quarantine measures.

3.2. Link prediction

Link prediction is a common task when dealing with graph-structured data. Any typical dataset may exhibit missing or unobserved edges. Given two nodes in a network, that are not connected through an edge, we want to estimate the probability of this edge being absent based on the currently observed links between nodes and their properties. Regular applications include predicting relationships in social networks [19], classifying unseen triplets in knowledge graphs [20] or identifying user-product interactions in recommendation systems [21]. An additional major area of application is the prediction of links that may or may not appear in dynamic and temporal networks changing over time [22]. Instead of predicting links between nodes, some authors explore prediction of risk for individuals within an epidemic, regardless their specific pattern of contacts [23]. Simple heuristics to estimate the probability of a missing link use common neighbors and exploit the topological structure of the graph to assess whether the two nodes belong to the same community and should therefore be connected. Other methods include random walks, factorization of the adjacency matrix of the graph and probabilistic graphical models. In recent years, deep learning based methods such as DeepWalk [24], node2vec [25] and Graph Convolutional Networks (GCNs) [26] introduced the concept of node embeddings which, together with various similarity metrics, can be used to estimate the likelihood of an edge.

Most of the discussed approaches are ill-suited for our problem at hand. Firstly, they assume the graph is composed of large
connected components and predict edges within these components with far greater accuracy than between them. Contact tracing data in epidemiology takes the form of many smaller components and users are especially interested in identifying connections between disconnected clusters [27]. Secondly, the considered models do not account for interpretability of results and are therefore unsuitable for applications in a risk-averse healthcare setting. Overall, to the best of our knowledge, there are no references that combine link prediction in infection networks with time-oriented visualization of epidemiological contacts. Therefore, in the following section we propose a simple, interpretable, and accurate approach to identify possible missing edges in a given contact tracing network.

4. Data

The dataset used for our analysis consists of 44,634 records, collected via structured telephone interviews carried out by employees of Cologne's DPH during the pandemic in 2020. The data is recorded in a database system with the purpose of tracking and tracing the pandemic and to support decision-makers in imposing or lifting restrictions. Contacted individuals include those tested positive for COVID-19 as well as individuals identified as direct contacts of infected individuals. Each one is asked for information on their demography including age, sex, country of origin, language, type of employment and housing block as well as epidemiological information such as dates of testing, start of the symptoms, hospital stays, return from risk area, and start/end of mandated quarantine. Additionally, if applicable, a list of contacts is collected and a single other individual is identified as the most probable source of infection [28]. Among 11,652 'index cases' that tested positive for COVID-19, there are 7,846 (67.3%) for whom the infection source is either unknown or not part of the dataset.

All individuals and their potential source of infection form a graph \( D = (V, A) \) where \( V \) is the set of all persons included in the dataset and the set \( A \subseteq V \times V \) contains a directed edge \( e = (u, v) \) if and only if \( u \) is the specified source of infection for \( v \). This makes \( D \) a directed forest (i.e. without cycles) consisting of individual infection trees or clusters.

It is important to note limitations and uncertainty of the available data. The quality of the collected information is dependent on multiple factors, including the ability of the individuals being interviewed to recall contacts and symptoms as well as detailedness of the interview. The incidence rate of COVID-19 in Cologne was highly volatile during the observed timeframe and impacted the disposable time per telephone interview at times. Still, the large dataset enabled a rare detailed view on the pandemic and a valuable starting point for data analytics.

An anonymized snapshot of the collected data was provided to us within the research project to perform the described analysis and develop the dashboard.

5. Users and tasks

As introduced above, the users for which we developed this approach are public health officers (PHOs) working at a DPH in Germany. PHOs are medical specialists, i.e. approved doctors with an additional qualification in public health. In our design study, we especially worked with the PHOs who oversaw the response to COVID-19 in their jurisdiction. PHOs have a profound knowledge of their geographical area of responsibility, including the type of housing, transport capabilities, and demographic details of the subregions. DPHs are typically subordinate to the regional government or the mayor's office in cities. They thus work closely with the executive bodies of a jurisdiction. While our contacts at the DPH in Cologne have a breadth of duties (also beyond SARS-CoV-2), we looked at specific domain tasks to support, which led to the requirements of our approach.

As introduced in the background section, we observed a number of tasks that our PHOs have to regularly perform and that we aimed to support with visualization and analytics methods:

- **T1**: Observe and trace single infected citizens in their area of responsibility.
- **T2**: Trace infections across different time frames and subregions, observing broader developments.
- **T3**: Identify clusters of infections to learn about the characteristics of the epidemiologic dynamic, especially typical infection routes.
- **T4**: Consult on prioritizing contact tracing and imposing/lifting restrictions in subregions.

It should be noted again that the currently available systems installed in DPHs are not supporting these tasks in an adequate way. They are rather meant to support the acquisition of data than to analyze the acquired data. For our visual analytics dashboard, we thus identified the following requirements:

- **R1**: Different visualizations and methods should be accessible through one web-based interface. Not being able to provide different viewpoints on the gathered data is a major downside of the current system environment at DPHs.
- **R2**: Integrate models of the DPH with the visualizations to enhance the situation awareness of the PHOs. There is profound domain knowledge of PHOs that cannot be easily applied to the gathered data and analyzed on this basis.
- **R3**: Use widely familiar visualizations that are accessible both by PHOs and the political stakeholders that the DPHs collaborate with.
- **R4**: Contact data must not leave the premises of the DPH. Privacy and data security is of utmost important for the sensitive data that the PHOs handle.

In the remainder of the paper, we will refer to the tasks and the requirements to clearly show how our approach supports the domain users and their tasks through visualization and analytics.

6. Proposed visual analytics dashboard

Based on the above tasks and requirements, we designed and developed a visual analytics dashboard in close collaboration with PHOs. It consists of a machine learning module that predicts missing links in the contact data and a web application that visualizes this data enriched with the prediction results plus additional visualizations in one environment. We use the anonymized data from Cologne’s health department as the basis for the machine learning and the visualization dashboard. In the following, we present the main components of our visual analytics dashboard in detail and conclude with a practical use case.

6.1. Dashboard overview

Our visual interface consists of three main interactive views (see Fig. 1): A graph-based view to identify infection clusters, a temporal view to analyze the infection dynamic, and a geographical view to perceive the spatial distribution of cases for specific DPHs (R3). In the following subsections, we describe the different views in detail and provide further technical details. All visualizations are linked and allow the selection of groups of persons for a more detailed analysis (R1).
Fig. 1. Dashboard overview: Visual interface for contact tracing and missing link detection: Our web application consists of three interactive views: In the top-center the Contact Network shows persons with their documented contacts (blue edges) as well as predicted transmissions (red edges) ①. In the bottom-center the Epi-Gantt diagram shows events associated with the virus for each person of a selected cluster ②. On the right a geospatial view to perceive the spatial distribution of selected cases for specific DPHs ③. At the top and the left the user can apply various filters to select specific groups of persons and control the parameters of the Contact Network.

6.1.1. Node-link visualization

In order to facilitate a fast identification of infection clusters we implemented a graph-based view of the contacts in the dataset (T₄). We chose a node link diagram for this task in accordance to the preferences we acquired with the PHOs (R₂). At the top-center of Fig. 1 the graph-based visualization shows the contact network ①. Each node represents a person and describes their state. The color-coding indicates whether a person was tested positive or negative. If no information about the test outcome was recorded, we display if the person had symptoms prior or shortly after a contact with other people. The directed edges show contacts between people and thus indicate potential infection paths of the virus. The blue edges represent the registered contacts in the dataset and denote documented infections from the DPH dataset. Additional red edges represent the predicted links according to their calculated probability, ranging from 100% (saturated red) to 50% (slightly red). Predicted links below 50% probability are omitted due to their unreliability. The main benefit of the node link diagram is that it allows to identify individuals that were in contact with each other at a glance. Therefore, it speeds up the identification of infection chains which is crucial for timely countermeasures. Further, this view is a good place to start an analysis because it provides an overview over the data. The node-link visualization supports zooming and panning for navigation and allows the users to interact with the graphs to select interrelated contacts of interest for further investigation in the other two views. Additionally, the predicted links make it possible to identify additional potential contacts, which can be verified by the DPH staff to fill the missing links in their data. Thus, a fast and more profound assessment can be made about the extent of infection clusters, which support informed decisions to contain the spread of the virus for future cases (T₄).

6.1.2. Epi-gantt

The graph view is combined with an event-based visualization that helps to analyze the temporal characteristics of the infection dynamic. The PHOs suggested a Gantt chart for this, which we incorporated into the visualization ②. The Epi-Gantt diagram at the bottom-center of Fig. 1 shows a selection of event records of connected persons. Each row in this view represents a person and lists important time-dependent events, like contacts with other people, start of symptoms, and the outcome of a SARS-CoV-2 test. Bars represent continuous events, and circles represent events that can be considered as a point in time. Detailed information about these events can be displayed via a tooltip by hovering the mouse over the corresponding bar or circle. Given a selected timeframe at the top of the dashboard, the Epi-Gantt shows all persons, who contain at least a single data point (test result, symptoms or quarantine) within the timeframe. Data of the respective persons is displayed in full, this implies that the timeframe of the Epi-Gantt can extend over the global timeframe selected by the user. This visualization is suited to analyze the temporal dependencies between contacts in a cluster of limited size and thus should be used after a drill-down. Otherwise, the limited vertical display space might hinder the comparison of events because of scrolling. This visualization can help to get insights on the sequence of events that led to the propagation of the infection (T₂). These insights support the identification of potential weaknesses in the containment strategy and improved countermeasures in the future, as further elaborated in Section 6.3.

6.1.3. Geospatial visualization

As an extension of our initial dashboard, we now provide a geospatial visualization that shows contacts from the DPH
database in two modes: A flow map that shows the general transmission dynamic in an area ($T_2$); a case-specific contact network that is provided for a selected case ($T_1$). Initially three layers are shown as in Fig. 2. A base map is overlaid with official city district data [29], displaying the seven-day incidence at district level. The seven-day incidence is calculated for the last day of the selected date range using population data provided by the city of Cologne [30]. The color palette utilized follows the RKI’s official range [31]. On top of this base layer, a flow map is drawn connecting the districts via clustered documented contacts. The number of contacts to each district is mapped to the thickness of the arrows. Here, the focus can be set on a single node by hovering over it. Edges not connected to this node are then grayed out with reduced opacity. Clicking on an individual person in the contact network removes the top layer and replaces it with the contact network of this case. This is shown in the map view of Fig. 1 (9). Here the documented contacts (see Fig. 1 blue arrows), as well as the predicted transmissions (see Fig. 1 red arrows) to and from this person are mapped to the corresponding districts. The number of contacts and predicted transmissions is mapped to the thickness of the arrows. This is giving a more general overview of the infection spread, while retaining control over individual cases.

The benefit of this new geospatial view on the contact traces stems from the mental map of the DPH experts who divide their area of responsibility into various districts ($R_2$). Each district has its own characteristic, most prominently the population density and the percentage of the commuting workforce. The district-level layer shows the infection contacts between these districts and complements the mental map with a data-driven view of the inter-district dynamic. One particular contact who is selected in one of the views is now mapped to the home district of this contact with links to the other contacts in his or her contact network. While the contact graph was already provided in the former version of the dashboard, the geographical mapping gives a further impression of the locality of the contact events. The predominance of local vs. cross-district contacts would have a clear influence on the DPH’s policy decisions ($T_4$).

6.1.4. Technical details
To facilitate efficient contact tracing and analysis of the data we implemented various filtering capabilities ($R_4$). At the top of the dashboard (see Fig. 1) is a time filter to narrow down the displayed data to the time frame of interest. On the left are additional filters to specify a person (via a pseudo ID) or to set the minimum sizes of displayed contact networks. Further, it is possible to display only those persons that had a positive SARS-CoV-2 test result or had symptoms associated with the virus. It is also possible to adjust the amount of displayed predicted links by adjusting the minimum probability threshold via a slider control.

We use a JavaScript web frontend implemented with the framework react.js [32] and material-ui [33]. For the graph visualization we use the library react-force-graph [34] that implements a force directed layout. The event based visualization is implemented with d3.js [35]. For the map view, we use the visualization framework deck.gl [36]. The extension flowmap.gl [37] provides a flow map layer that is used for visualizing infection contacts. The geospatial data is stored and served with GeoRocket [38]. The entire system is able to run on premise without the need of external services ($R_4$).

6.2. Predicting missing links
We predict missing edges with high probability inside our contact tracing network based on the known connections and their attributes. Feature selection was based on DPH experts’ feedback to find feasible real-world features that are most relevant for our task ($R_2$). Given an ordered pair of nodes ($u,v$) inside our graph, we calculate the days between infection reports, the difference in years of age and a binary feature that encodes if $u$ and $v$ live in the same housing block, an anonymized proxy for the home address. The age difference augments the classification, as the difference between source and target infection typically is around zero due to people with similar age that are learning and working together or separated by a generation gap ($\approx 30$ years) due to families living together, at least for the dataset used for our experiments. The generated feature vectors are the input to multiple different classification models including DecisionTrees, RandomForest, NaiveBayes and Support Vector Classification. This selection includes interpretable models as well as widely used classification architectures. We omit the comparison to highly complex model architectures, such as Graph Neural Networks (GNNs) [39], as they are hard to interpret and therefore limit the understanding of the results. Hyperparameters are selected by performing a grid search and cross-validation over a predefined parameter space. We evaluate our approach on the anonymized data from Cologne’s containing 44,634 persons with 11,652 ‘index cases’ who were tested positive for COVID-19 among them. To create a dataset for training our models, we select all observed edges and add 40,000 negative samples of random pairs of persons, who do not share a connection inside our dataset. We perform a random 85%/15% train-test split of the resulting dataset and train our model to predict the probability of an infection. For model performance comparison, we compute precision, recall and the receiver operating characteristic (ROC) curve for all classifiers plotting true positive rate (TPR) versus false positive rate (FPR) on the test set. Results are reported in Table 1. Area under the
ROC curve (AUROC) is best for the RandomForest model with 0.98, but all other models do achieve similar results. For the presented use case, a high precision is particularly desirable as resources to track additional infection transmissions at the DPH might be scarce and false positive predictions could lead to negative consequences (e.g. quarantine) for the citizens involved (T3). To retain high interpretability of results, we confine the model to the attributes selected above and integrate the best performing DecisionTree model into our visualization dashboard. The model can be easily scaled up to a large number of nodes and generates classification results for our 44k patients dataset in a matter of milliseconds.

6.3. Use case

In the following, we revisit the care home example in Section 2. A PHO wants to find out whether the recent increase of infections in a care home was induced from the outside (T3). To focus on the events of last week the user specifies a time filter, thus reducing the number of observed cases from thousands to a few hundred. Further, the PHO may query for a specific person from the care home that was tested positive in this time span (T1). The infected person and their direct contacts are automatically highlighted in the node link diagram. Therefore, the user can identify the relevant contact graph at a glance and zoom in to investigate its structure. The contact network (see (1) in Fig. 1) includes not only blue edges, which represent known contacts to the other residents of the care home, but also two outgoing edges and one incoming red edge. These predicted links show additional potential contacts, which may have led to the spread of the infection into the care home or from the care home to the outside. Further, the user may inspect the temporal dependencies between the recorded events in the Epi-Gantt diagram (see (2) in Fig. 1). This way the PHO can get insights about the sequence of events that led to the infection cluster (T2). These insights might help to pinpoint weaknesses in the containment strategy and thus may lead to better countermeasures in the future (T4). Additionally, the PHO can use the Epi-Gantt diagram to check if the predicted transmissions between the care home cluster and outside groups are plausible regarding their temporal succession of events. In this case, these potential contacts can be verified by the DPH and added to their data (R2). Moreover, the Geospatial visualization provides deeper insight into the spread of infections throughout the city (see (3) in Fig. 1). This allows the PHO to determine which city districts are associated with the infection cluster to and from the care home (T2) and supports the implementation of appropriate countermeasures (T4). The PHO notices a medium infection situation in the district the care home resides in and that most of the infection come from the near vicinity. The interlinked views provide different viewpoints on the gathered data regarding the use case and allow for a more comprehensive analysis than separate visualizations (R1).

6.4. Discussion

We discussed our proposed visual analytics dashboard and its features with PHOs from multiple DPHs to evaluate the benefits for the established use cases. The feedback was strongly positive. A combined view of the temporal epidemiological data via the time scale together with the network-based view of the disease spreading between infected people was perceived as supportive to comprehend the spatio-temporal dynamic of the pandemic. The geospatial visualization allows for extended analysis of infection spread across the city. Additionally, the presentation of possible missing edges was perceived as helpful in the detection of plausible connections between known clusters. Identifying infection sources and tracing the spread of a virus through communities is especially important with the new mutations of SARS-CoV-2. However, the current dashboard is limited to the contact data from individual DPHs, while in practice infection clusters can span across jurisdictions that cover different populations, cities and countries. This limitation could be addressed with the availability of standardized contact tracing data and would require sophisticated data and a visualization architecture that allows the rendering of possibly millions of nodes. Additionally, the current link prediction model is driven by routinely collected observations. While this allows a flexible use of our dashboard, it is important to note that modeling the effect of changes in epidemiological parameters as well as public restrictions (e.g. in care homes) over time may provide further, more detailed insights. Right now, we do not visualize infections that were caused by a person outside our focus region, or infected contacts residing outside Cologne, since the particular DPH will not have responsibility for these individuals. However, we plan to denote infection links to regions out the DPH responsibility with a special node. This will show the DPH how much of the infection situation is caused by other regions in Germany or how many infections spread over from Cologne to these regions.

7. Conclusion and future work

We presented a novel visual analytics approach to detect SARS-CoV-2 infection clusters. Our dashboard calculates and visualizes possible missing infection routes inside the contact tracing network and supports the analysis of time-dependent events that led to the spread of the virus. The evaluation together with PHOs shows the benefits of our approach. In future work we would like to extend our environment with model explanations to better understand the complex interactions underlying SARS-CoV-2 infection clusters. Based on the presented results, link prediction methods together with larger and more detailed contact tracing data, while still preserving privacy, could achieve even better visual and analytical performance and support the control of pandemics. The addition of a visualization component displaying the spatial extent of the infection chains improves traceability.

CRediT authorship contribution statement

Dario Antweiler: Conceptualization, Methodology, Software, Validation, Formal analysis, Writing – original draft, Writing – review & editing, Visualization, Project administration. David Sessler: Conceptualization, Methodology, Software, Validation, Formal analysis, Writing – original draft, Writing – review & editing, Visualization. Maxim Rossknecht: Conceptualization, Methodology, Software, Validation, Formal analysis, Writing – original draft, Writing – review & editing, Visualization. Benjamin Abb: Conceptualization, Methodology, Software, Validation, Formal analysis, Writing – original draft, Writing – review & editing, Visualization. Sebastian Ginzel:
Appendix A. Supplementary data
Supplementary material related to this article can be found online at https://doi.org/10.1016/j.cag.2022.05.013.

References
[1] Buda S, an der Heiden M, Altmann D, Diercke M, Hamouda O, Reixroth U. Infektionsumfeld von erfassten COVID-19-ausbrüchen in deutschland. RKI Epidemiologisches Bull 2020;38:3–12. http://dx.doi.org/10.25646/7093.
[2] Krause G, Altmann D, Faensen D, Porten K, Benzler J, Pfoch T, et al. SurvNet electronic surveillance system for infectious disease outbreaks, Germany. Emerg Infect Diseases 2007;13(10):1548–55. http://dx.doi.org/10.3201/eid1310.070253, URL: http://wwwnc.cdc.gov/eid/article/13/10/07-0253_article.htm.
[3] Antweiler D, Sessler D, Ginzel S, Kohlhammer J. Towards the detection and visualization of COVID-19 infection clusters. In: Vrotsou K, Bernard J, editors. EuroVis workshop on visual analytics. The Eurographics Association, 2021, http://dx.doi.org/10.2312/eurova.20211097.
[4] Carroll LN, Au AP, Detwiler LT, chieh Fu T, Painter IS, Abernethy NF. Carroll LN, Au AP, Detwiler LT, chieh Fu T, Painter IS, Abernethy NF. Conceptualization, Methodology, Software, Validation, Formal analysis, Writing – original draft, Writing – review & editing, Visualization. Jörn Kohlhammer: Conceptualization, Methodology, Software, Validation, Formal analysis, Writing – original draft, Writing – review & editing, Visualization.
[5] Si Y-W, Leong K-I, Biuk-Aghai RP, Fong S. Hidden cluster detection for infectious disease epidemiology: A systematic review. J Biomed Inform 2014;51:287–98. http://dx.doi.org/10.1016/j.jbi.2014.04.006.
[6] Leite RA, Schetinger V, Ceneda D, Henz B, Milksch S. COViS: Supporting temporal visual analysis of Covid-19 events usable in data-driven journalism. In: 2020 IEEE visualization conference. 2020, p. 56–60. http://dx.doi.org/10.1109/VIS1454.2020.00018.
[7] Assaf S, Ghani S, Jenkins-Smith HC, Ebert DS, Hadwiger M, Hottei L. A visual analytics based decision making environment for COVID-19 modeling and visualization. In: 2020 IEEE visualization conference. 2020, p. 86–90. http://dx.doi.org/10.1109/VIS1454.2020.00024.
[8] Burmeister J, Bernard J, Kohlhammer J, LPeers: Temporal similarity search in Covid-19 data. In: Vrotsou K, Bernard J, editors. EuroVis workshop on visual analytics. The Eurographics Association, 2021, http://dx.doi.org/10.2312/eurova.20211098.
[9] Ponjavić M, Karabegović A, Ferhatbegović E, Tahirović E, Uzunović S, Travar M, et al. Spatio-temporal data visualization for monitoring of control measures in the prevention of the spread of COVID-19 in bosnia and herzegovina. Medicinski Glasnik 2020. URL: https://pubmed.ncbi.nlm.nih.gov/32602300/.
[10] Leach A, Clarke S, Kirk A. COVID UK: coronavirus cases, deaths and vaccinations today. 2021. URL: https://www.theguardian.com/world/2021/ aug/20/covid-uk-coronavirus-cases-deaths-and-vaccinations-today. [Accessed 12 November 2021].
[34] React-force-graph - vasturiano. 2021. URL: https://github.com/vasturiano/react-force-graph.

[35] Bostock M, Ogievetsky V, Heer J. D3 data-driven documents. IEEE Trans Vis Comput Graphics 2011;(12):2301–9, URL: https://d3js.org/.

[36] deckgl. WebGL-powered visualization framework for large-scale datasets. 2021, URL: https://deck.gl/.

[37] flowmapgl - teralytics. 2021. URL: https://github.com/teralytics/flowmap.gl.

[38] Krämer M. GeoRocket: A scalable and cloud-based data store for big geospatial files. SoftwareX 2020;11:100409. http://dx.doi.org/10.1016/j.softx.2020.100409, URL: https://www.sciencedirect.com/science/article/pii/S2352711018302036.

[39] Scarselli F, Gori M, Tsoi AC, Hagenbuchner M, Monfardini G. The graph neural network model. IEEE Trans Neural Netw 2009;20(1):61–80. http://dx.doi.org/10.1109/TNN.2008.2005605.