Vision, Modeling, and Visualization (2019)
H.-J. Schulz, M. Teschner, and M. Wimmer (Eds.)

Visual Analysis of Probabilistic Infection Contagion in Hospitals

M. Wunderlich[1,6], I. Block[1], T. von Landesberger[1,2,6], M. Petzold[3,6], M. Marschollek[4,6], and S. Scheithauer[5,6]

[1]TU Darmstadt, Germany [2]Karlsruhe Institute of Technology, Germany [3]University Hospital Heidelberg, Germany
[4]Hannover Medical School, Germany [5]University Medicine Göttingen, Germany
[6]HiGHmed Use Case Infection Control, Germany

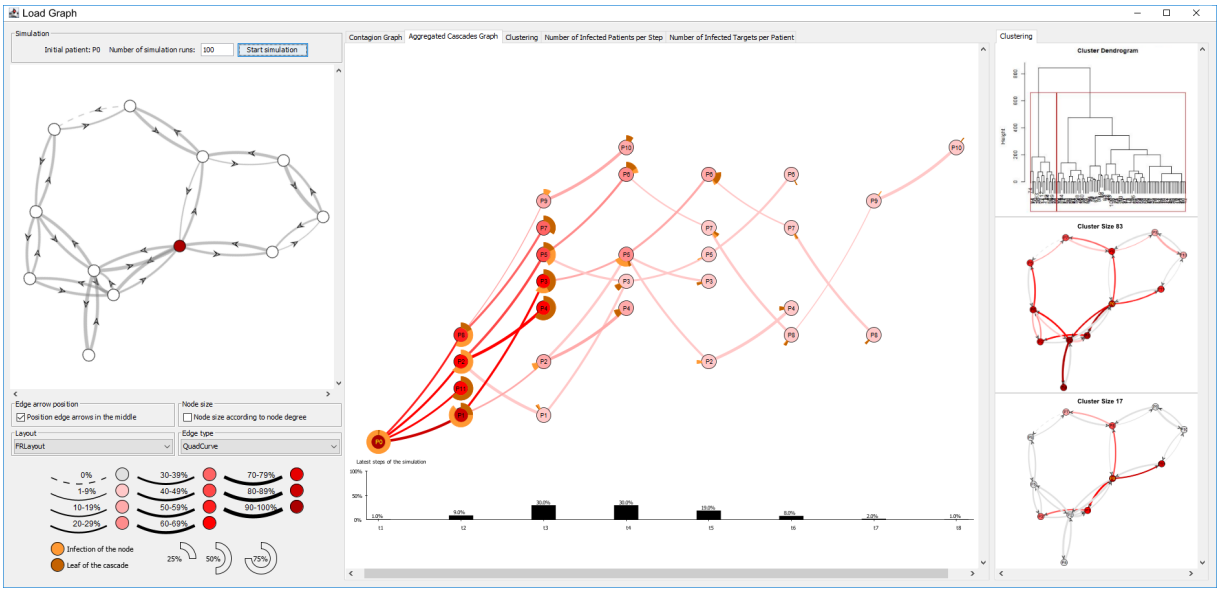

Figure 1: *Visual Interface. The infection contagion in the patient contact graph (left) is simulated. The development of the contagion is shown with* Aggregated Cascades Graph *(center). The contagion results are clustered and displayed with* Contagion Graph*s (right).*

Abstract

Clinicians and hygienists need to know how an infection of one patient could be transmitted among other patients in the hospital (e.g., to prevent outbreaks). They need to analyze how many and which patients will possibly be infected, how fast the infection could spread, and which contacts are likely to transfer the infections within the hospital. Currently, infection contagion is modeled and visualized for populations only on an aggregate level, without identification and exploration of possible infection between individuals. We present a novel visual analytics approach that simulates the contagion in a contact graph of patients in a hospital. We propose a clustering approach to identify probable contagion scenarios in the simulation ensemble. Furthermore, our novel visual design for detailed assessment of transmission shows the temporal development of contagion per patient in one view. We demonstrate the capability of our approach to a real-world use case in a German hospital.

CCS Concepts

• ***Human-centered computing*** → ***Visualization;*** *Graph drawings;* ***Visual analytics;***

1. Introduction

Multi-resistant pathogens and viruses, such as the influenza virus and the noro virus, within hospitals are an acute problem in Germany, USA, Japan, and other countries worldwide [GG11, HSR*18]. The pathogens are transmitted among the patients in the hospital. In a hospital with a few hundred patients, dozens of the patients might get infected, depending on the pathogen type and

DOI: 10.2312/vmv.20191328

hospital ward [HSR*18]. An infection obtained in hospitals is not only a threat for the patients' health, it also raises costs of treatment.

Clinicians and hygienists (or "experts") seek to identify and predict infection contagion to prevent pathogen outbreaks, i.e., a higher number of infected patients than the usual occurrence. Infectious patients are identified with their microbiology results [HSR*18]. However, not all patients are tested regularly for all types of pathogens and infection contagion might be overlooked. Moreover, for a new patient in a hospital, the test results are available only after a couple of days after the test. In the meantime, the infection could already spread. An infected patient may transmit the infection to other patients through contact, e.g., those in the same hospital ward, same operation room, et cetera. The infected patients can then transmit the infection to their contacts and so on [DWSG12]. The prediction of such an infection contagion among the patients would help the experts to understand likely transmission routes to prevent an outbreak by developing appropriate intervening strategies. For instance, patients with a high probability of getting infected or patients with a high probability of transmitting the infection to many others should be isolated early.

Currently, clinicians and hygienists analyze infection contagion mostly manually. This analysis heavily depends on their individual expertise and is error-prone [HSR*18]. In the research literature, disease spreading is simulated and visualized mostly for large populations. The available approaches analyze and visualize the number of infected persons over time and the spatial distribution of infections [MLR*11, AKMR16]. These allow for an aggregated view on the number and location of infected persons. However, the analysis of infection contagion within hospitals requires an exploration of possible infections and routes for individuals. The visual analysis of contagion among individuals exists for other domains such as finance [vLDBF15] or information spreading [AP16]. These approaches assume a deterministic spread model, but infection contagion is probabilistic. The prediction of disease spreading is commonly computed using *Monte-Carlo simulations* of *S-I-models* that use a probability of infecting a patient along a so-called patient contact graph [NS13, SVS*17]. Such simulations result in hundreds of dynamic contagion graphs (or an "ensemble"). Current approaches for the visual comparison of many graphs [FPSG10, vLGS09] do not take these dynamics into account.

We present a novel visual analytics approach that simulates and visualizes infection contagion in a contact graph of patients. We predict this infection contagion by computing *Monte-Carlo simulations* of *S-I-model*, which starts from one initial patient (e.g., a new patient in a hospital) and spreads along patient contacts (cf., Subsection 3.1). This method is commonly used [NS13, SVS*17] and fits the experts' model of contagion. Our analysis approach for the resulting ensemble of dynamic contagion graphs is based on two interrelated parts. First, in Subsection 4.3, we present a novel visual design that combines several contagion-relevant data in one view. This data encompasses, i.a., the likelihood of infection for patients, for infecting contacts, and the temporal development of the infection contagion. Second, in Subsection 4.4, we propose a specialized clustering to gain an overview of the possible contagion scenarios, which may differ in, e.g., the length of the contagion, the set of infected patients, and the infecting contacts.

We demonstrate the usefulness of our approach on a real dataset from a German hospital in Section 5. We developed this use case in cooperation with our project partners—clinicians and hygienists.

2. Related Work

Contagion transmission over networks is studied in domains such as health-care and epidemiology for disease spreading [HSR*18, AKMR16, DWSG12], finance for systemic risk analysis [vLDBF15, Sar16], biology for gene mutation analysis [LKB*14], social media for information spreading, and opinion flow [BHBGBM13, ZCW*14, AP16, VWH*13]. Each domain has specific contagion models, that determine the simulation result and, thus, the visualization requirements.

For the simulation of disease spreading, targeted tools such as NEMO [AKMR16] or GEFSim [SVS*17] exist. They focus on the simulation functionality and visualize only population statistics—the number of infected patients in line charts. This is insufficient for a detailed view of stochastic infection contagion over a patient contact network. Such simulations result in an ensemble of dynamic graphs that need to be compared and analyzed in detail for patients individually. This is challenging, with very few existing approaches [LZM19, BBDW17]. Most ensemble visualizations focus on multivariate and temporal data [CZC*15]. For networks, Liu et al. [LZM19] and Bremm et al. [BvLH*11] provide a multi-level approach for visual comparison of static phylogenetic trees. These approaches are restricted to static tree-structured data. Manynets [FPSG10], GraphLandscape [KKNW17], and SOM-based clustering [vLGS09] extract graph properties and compare many static graphs using their properties. However, two differently structured graphs can have the same properties [CSL*18]. Small multipiles [BHRD*15] allow to compare several static graphs in detail but do not scale well with hundreds of nodes and thousands of graphs. The graphs could be reduced to points and their similarity could be shown through dimensionality reduction [vLDBF15, vdEHBvW15]. This would allow to show also the network dynamics but not the graph structure as necessary for infection contagion analysis. The dynamics of the graph structure can be shown by GraphDiaries [BPF13] or by interleaved parallel edge splatting [BHW17]. This is, however, restricted to one dynamic graph.

Visual analytics for epidemiologists focuses mainly on the spatio-temporal evolution of a disease over a population—how many persons will be infected in which geographic area and how fast will the disease spread over these areas [MHR*10, BWMM15, LAS14, BH13, YDH*17]. However, the detailed view on individual infected persons and contacts transmitting the disease, as needed for hospital analysis purpose, is not sufficiently supported. Visual analysis of hospital patients by now concentrates on the treatment history [RWA*13, RFG*17, CCDW17] without a specific view on the infection contagion along patient contacts.

3. Contagion, Simulation, and Analysis Tasks

In the following, we describe the terms regarding infection contagion that we use throughout the paper. We explain the simulation and present the analysis tasks of the hygienists and clinicians.

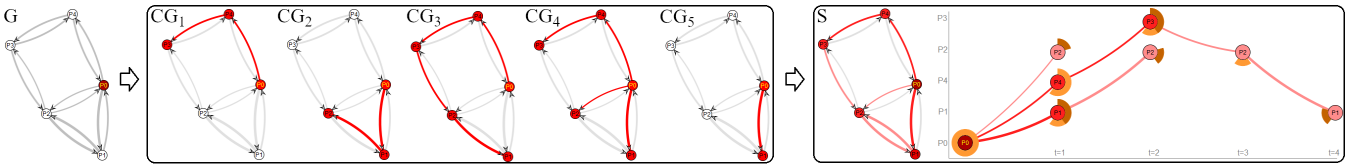

Figure 2: *Five possible contagion graphs CG_i for the contact graph G. Each contagion spreads from the initial patient P0 (red in G). In CG_3 and CG_4 the same nodes are infected but the order and infective edges differ. The similarity of CG_3 and CG_4 depends on the task (cf., Subsubsection 3.3.3). All contagion graphs of the simulation S are shown with the* Contagion Graph View *(left) and the* Aggregated Cascades Graph View *(right).*

3.1. Definitions

A **patient contact graph** (or "contact graph") $G = (N,E)$ is a graph consisting of nodes $N = \{n_1,\ldots,n_{|N|}\}$ and directed edges $E = \{e_1,\ldots,e_{|E|}\} \subseteq N \times N$. The nodes represent patients in a hospital and the edges represent contacts between these patients, e.g., when patients share a room. A contact is bidirectional per-definition, i.e., when patient n_1 was with patient n_2 at a certain location, then n_2 was with n_1 at this location as well: $\forall e_i = (n_k,n_l) \in E : \exists e_j = (n_l,n_k) \in E$. The set of nodes $\{n_l\} : (n_k,n_l) \in E$ are "targets" of n_k.

Various models [All94] exist to model an **infection contagion** (or "contagion/transmission process", "disease spreading") over the patient contact graph G. We use the discrete *S-I-Model*, a simple but powerful model, which fits the hygienists' model of contagion in our use case. Each patient and contact has an infection state $Inf(\cdot,t) \in \{0,1\}$ per step t. 0 means that a patient or contact is susceptible and 1 means that a patient is infected and a contact is infective (i.e., transmitted an infection), respectively. This state can change during the infection contagion. At $t = 0$, all contacts and all patients apart from one initial patient are susceptible.

Each edge $e_i = (n_k,n_l)$ has an attribute $\overline{p(e_i)} \in [0,1]$ that denotes the probability of **infection transmission** from patient n_k to patient n_l. This probability is given externally. It depends on the general probability of infection transmission and individual patient characteristics (e.g., elderly persons are more likely to get infected).

During infection contagion the infection of the initial patient n_0 $Inf(n_0,0) = 1$ at $t = 0$ is transmitted along the contacts in the graph G. In each step $t > 0$ the patients n_k infected at step $t-1$ infect their susceptible targets $\{n_l\} : e_i = (n_k,n_l) \in E \wedge Inf(n_l,t-1) = 0$ with probability $\overline{p(e_i)}$. This possibly causes an infection state change of any target patient and the respective contact: $\Delta Inf(\cdot,t+1) = Inf(\cdot,t) \vee Inf(\cdot,t-1)$. According to the hygienists' and clinicians' knowledge of infectious diseases in the analytical focus, e.g., MRSA or herpes, infected patients remain infectious even though they may not have any illness symptoms. The contagion continues as long as new patients are infected: $\forall t \in (0,T] \exists n_l \in N : \Delta Inf(n_l,t) = 1 \wedge \forall t > T \forall n_l \in N : \Delta Inf(n_l,t) = 0$. T is the length of the contagion. Each step t may be seen as a time interval of infection transmission, such as one day for bacterial diseases. Note that every edge has exactly one chance to transmit an infection.

Contagion creates a dynamic graph $CG = (G_0,\ldots,G_T)$. This **contagion graph** is composed of the same nodes and edges as the contact graph: $N_t = N$, $E_t = E$. Only the infection states of the patients and contacts $Inf(\cdot,t)$ change between G_t and G_{t+1} according to the description above. Note that a contact of two patients in G is bidirectional whereas the infection transmission between two patients is unidirectional. For instance, in Figure 2, the patients $P3$ and $P4$ had contact. In the simulation run for CG_1, $P4$ infects $P3$, e.g., via cough/air. A **cascade** C is a directed, acyclic subgraph of G_T that contains only the infected nodes and edges: $C = (N_C,E_C) : \forall n \in N_C \subseteq N_T Inf(n) = 1 \wedge \forall e \in E_C \subseteq E_T Inf(e) = 1$.

3.2. Simulation

The infection contagion is a stochastic process (cf., Subsection 3.1). Its realization depends on the realization of the probabilistic infection transmission $\overline{p}$ in each step t for each contact with an infected patient $n, Inf(n) = 1$. Hence, the analysis of infection contagion is performed using *Monte-Carlo simulations* [SVS*17,JRS09,NS13].

For one patient contact graph G, the *Monte-Carlo simulation* calculates R contagion graphs CG_r: $S = \{CG_1,\ldots,CG_R\}$ (also called "ensemble"). The number of **simulation runs** R is usually high, such as $R = 100$ or $R = 1000$. Determining the most appropriate value of R is a subject of research [LM05]. In the illustrative example of Figure 2, R equals 5. Each run r starts with the same initial patient n_0, which simulates a disease starting to spread from an infected patient. The infection of n_0 may, e.g., stem from outside of the hospital ("non-nosocomial") or its source can be an operation. The stochastic infection contagion creates possibly different contagion graphs in each simulation run $CG_r \neq CG_q, r \neq q$: A different set of nodes or edges may be infected and/or they are infected in different steps (cf., CG_1–CG_5 in Figure 2).

3.3. Tasks

The discussions with hygienists and clinicians in our project team [HSR*18,LWB*19,SvLB*19] as well as the study of medical literature [SVS*17,JRS09,NS13] have revealed the following tasks for infection contagion among hospital patients.

3.3.1. Tasks on the Infection Contagion for Individual Patients

Task-*n* What is the probability of infection of a patient n?

Task-*n*-*t* What is the probability of infection of a patient n until step t and in step t?

Task-*e* What is the probability of a contact e to transmit an infection?

Task-*e*-*t* What is the probability of a contact e to transmit an infection in step t?

Task-n_l What is the probability for a patient n infecting no other patients n_l?
Task-n_l-t What is the probability of a patient n infecting no other patients n_l in step t?
Task-$|\{n_l\}|$ What is the probability of a patient n infecting a specific number of other patients?

Note, the probability of contact $e = (n_i, n_k)$ to transmit an infection $p(Inf(e,\cdot))$ equals the probability n_i to infect the patient n_k.

3.3.2. Tasks on the Length of the Infection Contagion

Task-t What are the likely lengths of the contagion T and how probable are they?
Task-$\Delta|N|$-t How many patients may be infected in step t and how probable are these numbers?

Note, we focus on the new infections per step as the hygienists and clinicians in our project are especially interested in outbreak detection. An **outbreak** is defined as the number of new infections $|\{n \in N : \Delta Inf(n,\cdot) = 1\}|$ exceeding a user-defined threshold. Task-$\Delta|N|$-t analyzes when outbreaks probably occur during infection contagion and together with the other tasks, which patients are likely to be part of the outbreak. Related work analyzed the alternative task of the total number of infected patients, i.e., the sum of new infections (cf., Section 2).

3.3.3. Tasks on the Contagion Result and Temporal Progress

Task-N Which same patients are infected in the infection contagion result G_T across simulation runs and what is its probability?
Task-N-t Which same patients are infected in the infection contagion steps $G_0, \ldots, G_T$ across simulation runs. How likely is it?
Task-E Which same contacts are infected in the infection contagion G_T across simulation runs and how likely is it?
Task-E-t Which same contacts are infected in the infection contagion steps $G_0, \ldots, G_T$ across simulation runs. How likely is it?

4. Visual Analysis of Infection Contagion

Our user interface (cf., Figure 1) for the simulation and exploration of infection contagion has two parts. On the left, the user loads the contact graph G and sets the contagion simulation settings. On the right, the user visually explores the results in several complementary views, which answer the user tasks, cf., Subsection 3.3.

- The *Contagion Graph View* visualizes the result of the contagion: who is infected how likely and by whom.
- The *Aggregated Cascades Graph View* shows details of the contagion development for individual patients, contacts, and steps.
- A specialized clustering of contagion cascades identifies likely contagion scenarios. In this view, the user selects the distance function corresponding to the task at hand. The *Contagion Graph* and *Aggregated Cascades Graph Views* show the output.
- Supportive views provide a broad overview of the simulation: the length of the contagion, the number of infected patients, and the number of their targets.

All these views and analysis options can be used in combination within an exploratory process. The views are linked to, e.g., highlight a user-selected patient in all views.

4.1. Simulation Input View

The loaded contact graph is shown as a node-link diagram with *Fruchterman and Reingold* layout [FR91] of *JUNG* [OFS*05] (cf., Figure 1). The contact graph displays the patients as white nodes and their contacts as black edges. The width of the edges corresponds to the probability of infection transmission $\overline{p}$. The user chooses the initial patient n_0, which is then highlighted in red, in the graph (cf., Figure 2). He then sets the number of simulation runs R and starts the simulation (cf., Section 3.2).

4.2. Contagion Graph View

The *Contagion Graph View (CGV)* shows the probabilities of patient infection and a contact transmitting the infection within the node-link diagram of the patient contact graph (cf., Figure 2). We use color saturation [GHL15, BHR17] to show the infection probability: light red indicates low probability, whereas dark red indicates high probability, and light gray shows a probability value of 0 (cf., Figure 1, bottom left). The probability of the patient being infected (Task-n) determines the fill color of the nodes. The edges' color shows the probability of infection transmission (Task-e). The probabilities are calculated as the number of simulation runs in which a patient was infected respectively an edge transmitted the infection divided by the total number of simulation runs R. The calculated probability of being infective may not equal the origin contagion transmission probability $\overline{p}$. The original probability is indicated by the edges' width.

4.3. Aggregated Cascades Graph View

The *Aggregated Cascades Graph View (ACGV)* focuses on the development of the infection contagion on the level of patients, their contacts, and contagion steps. Inspired by [LKB*14, vLDBF15, AP16], we propose to transform the cascades of all dynamic contagion graphs S into one static visualization (cf., Figure 2).

We combine the cascades across all runs into one supergraph—a directed, acyclic graph with the initial patient n_0 as the root node. The other nodes of this supergraph are patient infections in individual steps. Edges are infection transmissions in individual steps. In different cascades, one patient n_i may get infected at different steps t, thus, forms several nodes of the supergraph. The same applies to infectious edges. Node and edge attributes denote the probabilities of infection of a patient or contagion transmission per step.

The *ACGV* shows the infected patients $n_i : Inf(n_i,t) = 1$ per contagion step t as nodes along the vertical and horizontal axis. Each patient has a unique vertical position and time steps are on the horizontal axis. The initial patient n_0 is at the bottom left ($t = 0, y = 0$). The subsequent patients n_i, infected in step τ, may differ in each cascade. They are shown at position $t = \tau$ and $y = i$. In different cascades, one patient n_i may get infected at different steps, thus, may occur at multiple positions t (with fixed $y = i$). The graph contains all infective edges. According to the layout, these connect patients at subsequent steps $t = \tau$ and $t = \tau + 1$. Hence, their direction follows from the t-position of the nodes.

The coloring of nodes and edges is consistent with the *CGV*. The darker a node's red fill is, the higher is the patient's infection

probability (part of Task-*n*-*t*). As the initial patient is infected in every cascade per definition, its node is filled with the darkest red. The darker an edge's red line color is, the higher is its transmission probability (Task-*e*-*t*). In contrast to the *CGV*, these probabilities are presented for each step t separately. Hence, the node fill saturation for patient n_i may differ for the steps $t = \tau_1, \tau_2; \tau_1 \neq \tau_2$. The edge width, again, represents the origin transmission probability $\overline{p}$, which is independent of time, thus, does not vary along the steps.

A yellowish circular arc around each node shows the patient's infection probability per step t (Task-*n*-*t*). A full circle represents a certain infection (probability of 100%), a half-circle 50% infection probability, a quarter circle 25% et cetera. At the first step $t = \tau$ a node n_i appears, its circular arc starts at the top center position above the node. The circular arc for the next appearance of n_i at $t = \tau + \delta, \delta > 0$ starts where the circular arc for $t = \tau$ ends. Hence, at $t = \tau + \delta$ the end of the circular arc represents the probability of infection for n_i in the first $\tau + \delta$ steps ($0 < t \leq \tau + \delta$; Task-*n*-*t*). Consequently, the end of the circular arc at the latest appearance of the node represents the same probability as the node's fill color saturation in the *CGV* (Task-*n*). A fraction of the yellow circular arc may be filled with darker yellow. This part denotes the number of cascades in which the corresponding patient was infected but did not infect other patients (i.e., the node is a leaf of the cascade; Task-n_l-*t*). For instance, in Figure 2, $P1$ gets infected at $t = 1$ with a probability of 63%. He or she infects other patients (here, $P2$) at $t = 2$ with a probability of 15%. The total probability of infection is 80% (i.e., the end of the circular arc at $t = 4$).

The user can highlight a patient by clicking on one of the corresponding nodes (cf., Figure 5). This patient is then emphasized with the edges to its infected targets at all steps t the patient gets infected. Clicking on a contact emphasizes the contact at all steps it was infective to identify transmission patterns. Patients or contacts with an infection probability below a threshold can be hidden.

4.4. Clustering of Cascades

The number of possible cascades may be high, even for smaller patient contact graphs. The user can cluster similar cascades of different simulation runs and view each cluster's representative. This reduces the number of cascades to display, thereby improves the overview of the infection contagion scenarios. Furthermore, the size of the cluster is equivalent to the probability of the corresponding contagion scenario.

We use hierarchical clustering because it is fast and requires only a similarity function and no further user-defined parameters. We propose four specialized similarity functions (sim) for analyzing cascades according to task-group 3 (cf., Subsection 3.3):

sim-*N* Infection of the same patients $n_i : Inf(n_i, t \leq T) = 1$
sim-*E* Infection transmission via the same contacts $e_i : Inf(e_i, t \leq T) = 1$
sim-*N*-*t* Infection of the same patients in the same steps $n_i, t : Inf(n_i, t) = 1$
sim-*E*-*t* The same infection contagion $e_i, t : Inf(e_i, t) = 1$

The similarity functions calculate the similarity of contagions as a similarity of boolean feature vectors extracted from the contagion data. The feature vector depends on the task (cf., Figure 2):

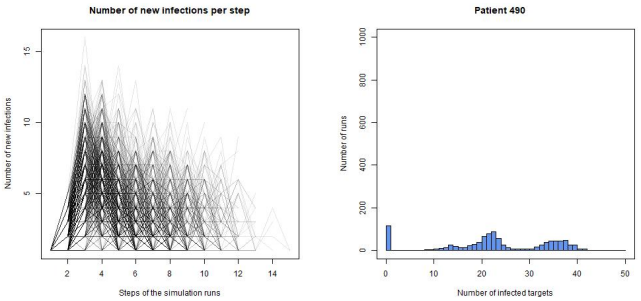

Figure 3: *Supportive Visualizations.* Number of Infected Patients per Step *(left) and* Number of Infected Targets per Patient *(right).*

1. sim-*N*: Feature vector of all patients. With 1 at position i if patient n_i was infected and 0 if not.
2. sim-*E*: Feature vector of all contacts. With 1 at position i if contact e_i was infectious and 0 if not.
3. sim-*N*-*t*: Feature vectors of all patients and all steps. With 1 at position i if patient n_i was infected in this step and 0 if not.
4. sim-*E*-*t*: Feature vectors of all contacts and all steps. With 1 at position i if contact e_i was infectious in step t and 0 if not.

We utilize the *Tanimoto* similarity [But99] to calculate the similarity of these feature vectors, as it is recommended for such fingerprints [BRH15]. The first two similarity functions consider only the final result of the contagion but not its development. The latter two also consider the contagion development. Therefore, the similarity of feature vectors in each step is calculated and then summed up.

The users may choose the similarity function that fits their task (Task-*N*, Task-*N*-*t*, Task-*E*, Task-*E*-*t*). In the GUI, the user can choose whether the temporal order should be considered for similarity calculation (sim-*N* or sim-*E* vs. sim-*N*-*t* or sim-*E*-*t*). The trade-off between the usage of nodes and edges for similarity calculation is set as $\alpha \in [0,1]$ whereas $sim = \alpha * \text{sim-}N + (1-\alpha) * \text{sim-}E$. Hence, $\alpha = 0$ calculates sim-*N* while $\alpha = 1$ calculates sim-*E*. sim-*N* and sim-*E* may be replaced with sim-*N*-*t* and sim-*E*-*t*, respectively.

We display the calculated similarity values of the iterative hierarchical clustering as a dendrogram (cf., Figure 4, right). Furthermore, the user can view the clusters with one representative cascade as well as the center per cluster in multiple views (juxtaposed). Both can be viewed with the *CGV* or the *ACGV*.

4.5. Supportive Visualizations

The following visualizations support the analysis. They can be viewed independently or in conjunction with the main visualizations (cf., Figure 1).

Length of the Contagion A histogram shows the lengths of the infection contagion T and their probabilities (Task-*t*). In Figure 1, 30% of the cascades ended at $t = 4$.

Number of Infected Patients per Step A line graph shows the number of patients infected per step t (cf., Figure 3; Task-$\Delta|N|$-*t*). The graph contains one semi-transparent line per simulation run

(CG_r), thus the superposition of all lines presents an overview for all simulation runs (S). The end of a line ends at the step at which the infection contagion stopped (Task-t).

Number of Infected Targets per Patient A histogram for each patient shows the possible number of its infected targets and their corresponding probability. In Figure 3, patient $P490$ infected zero other patients in 148 out of 1,000 runs, i.e., this patient infects no others with a chance of approx. 15% (Task-n_I). The infection of P490 likely spreads to more than 20 other patients (Task-$|\{n_I\}|$).

The upper bounds for the values shown in these supportive views are also retrievable from the *Aggregated Cascades Graph View*.

5. Use Case and Expert-Feedback

In this use case, we demonstrate the usage of our visual analytics approach by a hygienist from a German hospital from our project team. We use an anonymized dataset (.csv) from a German hospital. The data on patient locations over time (stored in hospital-intern accounting systems) was used to build the contact graph. The experts expect patients to have contact when they are at the same location at one moment. The initial probabilities of infection transmission $\overline{p}$ were given externally. The graph is built for three hospital wards of interest and all patients that are connected to the subgraph that includes $P498$. $P498$ can not infect not-connected patients. This results in a contact graph with 50 nodes and 327 edges.

The hygienist wishes to examine possible infection contagion among the patients in a hospital within three days. Even though new patients are screened routinely after hospitalization, it takes one to three days until the results are available. Within this time, a newly infected patient could have contact with several other patients. The hygienist sets $P498$ as the initial patient because this patient was hospitalized recently and might have been infected with a multi-resistant pathogen (via non-nosocomial infection).

The hygienist starts a simulation with the default recommendation of 1,000 runs. He first inspects the number of new infections per step (Task-$\Delta|N|$-t; cf., Figure 3, left) to look for outbreaks. As many lines have a high slope at early steps, he expects that the possibly infected $P498$ could have infected many patients until the screening results are available. This could cause an outbreak.

Then, the hygienist looks at the *Aggregated Cascades Graph View* (cf., Figure 5). He notices that the initial patient $P498$ might transmit the infection to six patients. For each of them, he inspects the histograms of the number of infected targets. For $P490$, he sees a trimodal distribution of the number of infected patients (cf., Figure 3, right). If $P490$ is really infected, he might not infect others at all but also highly likely about 20, or even about 40 patients.

To further investigate which contagion scenarios might occur and how likely they are, the hygienist clusters the simulated cascades. He decides to cluster based on infected nodes (Task-N) as he is primarily interested in the infected patients. This would allow him to identify patients for screening or isolation. After viewing the cluster dendrogram, the hygienist looks at the four possible scenarios (cf., Figure 4, left). The *Contagion Graph View* for the clusters shows that almost all patients will get infected eventually in the first

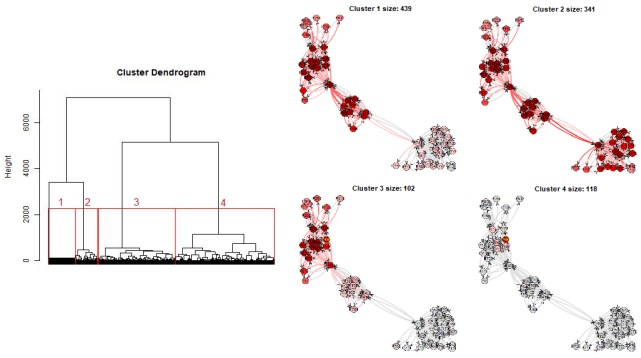

Figure 4: *Dendrogram and cluster representatives as small multiples for the clustering of possible infections starting from* $P498$.

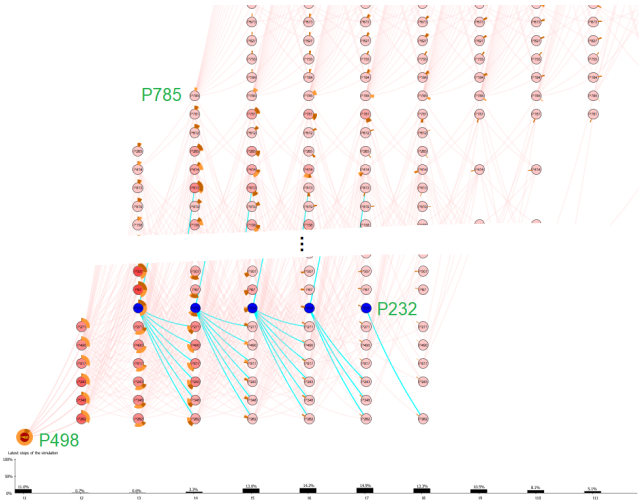

Figure 5: *Excerpt of the* Aggregated Cascades Graph View *for possible infections starting from* $P498$ *with the* Length of the Contagion *displayed below. The user highlighted* $P232$.

and second scenario. According to the cluster sizes, the contagion affects two wards with a chance of 10% (cluster 3), or the contagion stops after only a few infections with a likelihood of 12% (cluster 4). The notable probability of an outbreak that affects more than one ward causes the hygienist to report it to his managers.

Getting infected is particularly risky for patients with immunodeficiency. One of them is $P232$. Hence, this patient is especially interesting for the hygienist (e.g., to relocate the patient, if needed). $P232$ might be infected in steps 3 to 7 (cf., Figure 5). According to the circular arc, $P232$ gets infected with a chance of approx. 45% at $t = 3$ (Task-n-t) and with a chance of approx. 85% in total (Task-n). He clicks on $P232$ in the *ACGV* to highlight the patient with its outgoing edges in blue. Combining the yellow and brown circular arcs as well as the number of outgoing edges, the hygienist understands that both the infection probability and the maximum number of patients infected by $P232$ decreases with increasing steps (Task-n_I-t, Task-e-t). If possible, this patient should be relocated early.

The hygienist also notices very few edges between patients at the upper part of the *ACGV*. This is caused by the patients being at different wards. *P*785 has contact with patients of both wards, hence, is critical for driving the disease spreading. We should prevent his infection, e.g., by isolation, to stop spreading to another ward.

Expert-Feedback After using our system, the hygienist described it as a valuable support in infection analysis. Currently, their combined analysis of microbiological test results and the movement of patients in the hospital is a cumbersome work. The hygienist has to use multiple software systems to access the data and views it with common spreadsheet software. Hence, he finds it very helpful to have a contact graph displayed. The graph layout that indicates which patients are locally close together is intuitive. Currently, he can analyze the infection contagion only retrospectively. This system could provide new perspectives as it shows the possible upcoming contagion. As the hygienist mentions, the simulation results can be combined with actually proven transmission identified via genome sequencing. This will be part of our future work. He was surprised by the possible maximum number of infected patients calculated by the *S-I-model*. Hence, he also wanted to use the visualizations for other simulation models that are published in medical research for various pathogens. The hygienist praised the clustering of cascades to identify likely probable contagion scenarios. Clustering according to infected patients without time (i.e., sim-*N*) corresponds to the view on the data in retrospective analysis. However, he liked the temporal view on the contagion as offered by the *Aggregated Cascades Graph View* together with the ability to highlight transmissions reoccurring in different steps.

6. Limitations & Future Work

We use the simple but powerful *S-I-Model* with two assumptions. Based on the expert's feedback, we could extend our work to other disease spreading models such as *S-I-R* by adapting the similarity function to several states. We assume a static patient contact graph with externally given transmission probabilities $\overline{p}$. A dynamic patient contact graph would allow for, e.g., movements of patients among wards, leaving the hospital, and hospitalization of new patients. Furthermore, the attributes (e.g., age) of patients, with which the infection probabilities are calculated externally, might be included in the analysis process. We will extend our work in this direction in the future.

Another direction for further research is the scalability of our solution, especially the scalability of the *Aggregated Cascades Graph View*. We designed our analysis approach for the analysis of nosocomial infections, i.e., for a patient contact graph with only up to a few hundreds of patients. Usually, the hygienist analyses hospital wards separately (instead of the whole hospital), resulting in contact graphs with approx. 20–60 patients. Nevertheless, as indicated in our use case, the height of the *ACGV* increases rapidly and edge overplotting makes tracking transmission routes more difficult. Clustering of the cascades reduces the number of cascades to display in one view and the linked views help to gain overview in the larger *ACGV*. Further interaction and filtering could improve the analysis as well. Research on the visual analysis of higher-order networks [JJCC17] could also guide further work on the scalability.

7. Conclusion

We presented a visual analytics approach for infection transmissions from one initial patient in contact graphs of hospital patients. Our approach includes a *Monte-Carlo simulation* of contagion according to the *S-I-model*. We designed novel visualizations and a specialized clustering of possible infection contagion, all for the tasks of the hygienists and clinicians. The visualizations show the likely contagion results as well as the development of the infection contagion among individual patients (in addition to overviews). For the clustering, we propose a similarity function that considers patient infections, infecting contacts, and the steps of contagion as weighted by the user. We showed the usage of our approach in a use case of a hygienist in a German hospital.

Our approach can help the experts in both understanding infection contagion within real patient contact graphs (e.g., to reason routine screening of pathogens on hospitalization) and guiding the prevention of outbreaks (e.g., by isolating highly infectious patients). The hygienists and clinicians in our project team agreed on this. Future research can improve such simulation-based forecasts by including larger, dynamic, or multi-attribute contact graphs.

Acknowledgments

This work was supported in part by the German Federal Ministry of Education and Research (BMBF) within the framework of the research and funding concepts of the Medical Informatics Initiative (01ZZ1802B/HiGHmed) as well as by the Deutsche Forschungsgemeinschaft e.V. (DFG, LA 3001/2-1) and the Austrian Science Fund (FWF, I 2703-N31).

References

[AKMR16] ABDELHAMID S. E., KUHLMAN C. J., MARATHE M. V., RAVI S. S.: Interactive exploration and understanding of contagion dynamics in networked populations. In *Int. Conf. Behavioral, Economic and Socio-cultural Computing* (2016), pp. 1–6. 2

[All94] ALLEN L. J.: Some discrete-time si, sir, and sis epidemic models. *Mathematical Biosciences 124*, 1 (1994), 83–105. 3

[AP16] ARCHAMBAULT D., PURCHASE H. C.: On the effective visualisation of dynamic attribute cascades. *Information Visualization 15*, 1 (2016), 51–63. 2, 4

[BBDW17] BECK F., BURCH M., DIEHL S., WEISKOPF D.: A taxonomy and survey of dynamic graph visualization. *Computer Graphics Forum 36*, 1 (2017), 133–159. 2

[BH13] BROCKMANN D., HELBING D.: The hidden geometry of complex, network-driven contagion phenomena. *Science 342*, 6164 (2013), 1337–1342. 2

[BHBGBM13] BORGE-HOLTHOEFER J., BAÑOS R. A., GONZÁLEZ-BAILÓN S., MORENO Y.: Cascading behaviour in complex socio-technical networks. *J. Complex Networks 1*, 1 (2013), 3–24. 2

[BHR17] BAE J., HELLDIN T., RIVEIRO M.: Understanding indirect causal relationships in node-link graphs. *Computer Graphics Forum 36*, 3 (2017), 411–421. 4

[BHRD*15] BACH B., HENRY-RICHE N., DWYER T., MADHYASTHA T., FEKETE J.-D., GRABOWSKI T.: Small multipiles: Piling time to explore temporal patterns in dynamic networks. *Computer Graphics Forum 34*, 3 (2015), 31–40. 2

[BHW17] BURCH M., HLAWATSCH M., WEISKOPF D.: Visualizing a sequence of a thousand graphs (or even more). *Computer Graphics Forum 36*, 3 (2017), 261–271. 2

[BPF13] BACH B., PIETRIGA E., FEKETE J.-D.: Graphdiaries: Animated transitions andtemporal navigation for dynamic networks. *IEEE TVCG 20*, 5 (2013), 740–754. 2

[BRH15] BAJUSZ D., RÁCZ A., HÉBERGER K.: Why is tanimoto index an appropriate choice for fingerprint-based similarity calculations? *J. Cheminformatics 7*, 1 (2015), 20. 5

[But99] BUTINA D.: Unsupervised data base clustering based on daylight's fingerprint and tanimoto similarity: A fast and automated way to cluster small and large data sets. *J. Chem. Inf. and Comp. Sci. 39*, 4 (1999), 747–750. 5

[BvLH*11] BREMM S., VON LANDESBERGER T., HESS M., SCHRECK T., WEIL P., HAMACHERK K.: Interactive visual comparison of multiple trees. In *IEEE VAST* (2011), IEEE, pp. 31–40. 2

[BWMM15] BRYAN C., WU X., MNISZEWSKI S., MA K.-L.: Integrating predictive analytics into a spatiotemporal epidemic simulation. In *IEEE VAST* (2015), IEEE, pp. 17–24. 2

[CCDW17] CABALLERO H. S. G., CORVO A., DIXIT P. M., WESTENBERG M. A.: Visual analytics for evaluating clinical pathways. In *IEEE VAHC* (2017), IEEE, pp. 39–46. 2

[CSL*18] CHEN H., SONI U., LU Y., MACIEJEWSKI R., KOBOUROV S.: Same stats, different graphs. In *Int. Symp. Graph Drawing and Network Visualization* (2018), Springer, pp. 463–477. 2

[CZC*15] CHEN H., ZHANG S., CHEN W., MEI H., ZHANG J., MERCER A., LIANG R., QU H.: Uncertainty-aware multidimensional ensemble data visualization and exploration. *IEEE TVCG 21*, 9 (2015), 1072–1086. 2

[DWSG12] DONKER T., WALLINGA J., SLACK R., GRUNDMANN H.: Hospital networks and the dispersal of hospital-acquired pathogens by patient transfer. *PLoS One 7*, 4 (2012), e35002. 2

[FPSG10] FREIRE M., PLAISANT C., SHNEIDERMAN B., GOLBECK J.: Manynets: an interface for multiple network analysis and visualization. In *SIGCHI Conf. on Human Factors in Computing Systems* (2010), ACM, pp. 213–222. 2

[FR91] FRUCHTERMAN T. M., REINGOLD E. M.: Graph drawing by force-directed placement. *Software: Practice and experience 21*, 11 (1991), 1129–1164. 4

[GG11] GEFFERS C., GASTMEIER P.: Nosocomial Infections and Multidrug-resistant Organisms in Germany. *Deutsches Ärzteblatt International 108*, 6 (2011), 87–93. 1

[GHL15] GUO H., HUANG J., LAIDLAW D. H.: Representing uncertainty in graph edges: An evaluation of paired visual variables. *IEEE TVCG 21*, 10 (2015), 1173–1186. 4

[HSR*18] HAARBRANDT B., SCHREIWEIS B., REY S., SAX U., SCHEITHAUER S., RIENHOFF O., KNAUP-GREGORI P., BAVENDIEK U., DIETERICH C., BRORS B., ET AL.: Highmed–an open platform approach to enhance care and research across institutional boundaries. *Methods of information in medicine 57*, S 01 (2018), e66–e81. 1, 2, 3

[JJCC17] JUN TAO, JIAN XU, CHAOLI WANG, CHAWLA N. V.: Honvis: Visualizing and exploring higher-order networks. In *IEEE PacificVis* (2017), pp. 1–10. 7

[JRS09] JUHER D., RIPOLL J., SALDAÑA J.: Analysis and monte carlo simulations of a model for the spread of infectious diseases in heterogeneous metapopulations. *Phy. Rev. E 80*, 4 (2009), 041920. 3

[KKNW17] KENNEDY A., KLEIN K., NGUYEN A., WANG F. Y.: The graph landscape: using visual analytics for graph set analysis. *J. Visualization 20*, 3 (2017), 417–432. 2

[LAS14] LAN R., ADELFIO M. D., SAMET H.: Spatio-temporal disease tracking using news articles. In *SIGSPATIAL Int. Workshop on Use of GIS in Public Health* (2014), ACM, pp. 31–38. 2

[LKB*14] LENZ O., KEUL F., BREMM S., HAMACHER K., VON LANDESBERGER T.: Visual analysis of patterns in multiple amino acid mutation graphs. In *IEEE VAST* (2014), IEEE, pp. 93–102. 2, 4

[LM05] LERCHE I., MUDFORD B. S.: How many monte carlo simulations does one need to do? *Energy exploration & exploitation 23*, 6 (2005), 405–427. 3

[LWB*19] LANDESBERGER T. V., WUNDERLICH M., BAUMGARTL T., HÖHN M., MARSCHOLLEK M., SCHEITHAUER S.: Visual-Interactive Exploration of Pathogen Outbreaks in Hospitals. In *EuroVis-Posters* (2019), The Eurographics Association. 3

[LZM19] LIU Z., ZHAN S. H., MUNZNER T.: Aggregated dendrograms for visual comparison between many phylogenetic trees. *IEEE TVCG early access*, published online (2019), 1–1. 2

[MHR*10] MACIEJEWSKI R., HAFEN R., RUDOLPH S., LAREW S. G., MITCHELL M. A., CLEVELAND W. S., EBERT D. S.: Forecasting hotspots–a predictive analytics approach. *IEEE TVCG 17*, 4 (2010), 440–453. 2

[MLR*11] MACIEJEWSKI R., LIVENGOOD P., RUDOLPH S., COLLINS T. F., EBERT D. S., BRIGANTIC R. T., CORLEY C. D., MULLER G. A., SANDERS S. W.: A pandemic influenza modeling and visualization tool. *J. Vis. Lang. and Computing 22*, 4 (2011), 268–278. 2

[NS13] NAGY N., SIMON P.: Monte carlo simulation and analytic approximation of epidemic processes on large networks. *Open Mathematics 11*, 4 (2013), 800–815. 2, 3

[OFS*05] O'MADADHAIN J., FISHER D., SMYTH P., WHITE S., BOEY Y.-B.: Analysis and visualization of network data using jung. *J. Statistical Software 10*, 2 (2005), 1–35. 4

[RFG*17] RIND A., FEDERICO P., GSCHWANDTNER T., AIGNER W., DOPPLER J., WAGNER M.: Visual analytics of electronic health records with a focus on time. In *New Perspectives in Medical Records*. Springer, 2017, pp. 65–77. 2

[RWA*13] RIND A., WANG T. D., AIGNER W., MIKSCH S., WONGSUPHASAWAT K., PLAISANT C., SHNEIDERMAN B.: Interactive information visualization to explore and query electronic health records. *Foundations and Trends in HCI 5*, 3 (2013), 207–298. 2

[Sar16] SARLIN P.: Macroprudential oversight, risk communication and visualization. *J. Financial Stability 27* (2016), 160–179. 2

[SvLB*19] SARGEANT A., VON LANDESBERGER T., BAIER C., BANGE F., DALPKE A., ECKMANNS T., GLÖCKNER S., KAASE M., KRAUSE G., MARSCHOLLEK M., MALONE B., NIEPERT M., REY S., WULFF A., SCHEITHAUER S.: Early detection of infection chains & outbreaks: Use case infection control. In *ICT for Health Science Research: EFMI* (2019), vol. 258, IOS Press, p. 245. 3

[SVS*17] SAHNEH F. D., VAJDI A., SHAKERI H., FAN F., SCOGLIO C.: Gemfsim: a stochastic simulator for the generalized epidemic modeling framework. *J. Computational Science 22*, September (2017), 36–44. 2, 3

[vdEHBvW15] VAN DEN ELZEN S., HOLTEN D., BLAAS J., VAN WIJK J. J.: Reducing snapshots to points: A visual analytics approach to dynamic network exploration. *IEEE TVCG 22*, 1 (2015), 1–10. 2

[vLDBF15] VON LANDESBERGER T., DIEL S., BREMM S., FELLNER D. W.: Visual analysis of contagion in networks. *Information Visualization 14*, 2 (2015), 93–110. 2, 4

[vLGS09] VON LANDESBERGER T., GORNER M., SCHRECK T.: Visual analysis of graphs with multiple connected components. In *IEEE VAST* (2009), IEEE, pp. 155–162. 2

[VWH*13] VIÉGAS F., WATTENBERG M., HEBERT J., BORGGAARD G., CICHOWLAS A., FEINBERG J., ORWANT J., WREN C.: Google+ ripples: A native visualization of information flow. In *Int. Conf. World Wide Web* (2013), ACM, pp. 1389–1398. 2

[YDH*17] YANEZ A., DUGGAN J., HAYES C., JILANI M., CONNOLLY M.: Pandemcap: Decision support tool for epidemic management. In *IEEE VAHC* (2017), IEEE, pp. 24–30. 2

[ZCW*14] ZHAO J., CAO N., WEN Z., SONG Y., LIN Y.-R., COLLINS C.: # fluxflow: Visual analysis of anomalous information spreading on social media. *IEEE TVCG 20*, 12 (2014), 1773–1782. 2