Security Informatics
a SpringerOpen Journal

RESEARCH Open Access

Fluency of visualizations: linking spatiotemporal visualizations to improve cybersecurity visual analytics

Zhenyu Cheryl Qian[1*†] and Yingjie Victor Chen[2†]

Abstract

This paper adopts the metaphor of representational fluency and proposes an auto linking approach to help analysts investigate details of suspicious sections across different cybersecurity visualizations. Analysis of spatiotemporal network security data takes place both conditionally and in sequence. Many visual analytics systems use time series curves to visualize the data from the temporal perspective and maps to show the spatial information. To identify anomalies, the analysts frequently shift across different visualizations and the original data view. We consider them as various representations of the same data and aim to enhance the fluency of navigation across these representations. With the auto linking mechanism, after the analyst selects a segment of a curve, the system can automatically highlight the related area on the map for further investigation, and the selections on the map or the data views can also trigger the related time series curves. This approach adopts the slicing operation of the Online Analytical Process (OLAP) to find the basic granularities that contribute to the overall value change. We implemented this approach in an award-winning visual analytics system, SemanticPrism, and demonstrate the functions through two use cases.

Keywords: Representational fluency; Cybersecurity analysis; Spatiotemporal visualization; Interaction design

Introduction

One of the biggest challenges the information security society faces is analyzing large-scale spatiotemporal datasets. In most organizations and companies, their computer networks are routinely capturing huge volumes of historical data describing the network events. Most of these events are recorded as spatiotemporal data because every event takes place at a certain time and in a certain location. The location could be either a physical location (e.g., an office) or a virtual space (e.g., an Internet IP address) [1]. Different kinds of events have more detailed information, such as operations, products, targets, and human involvement, which can add more dimensions to the spatiotemporal database. As a result, such a dataset is usually both high-dimensional and very large.

Peuquet [2] identified three components in spatiotemporal data: space (where), time (when), and objects (what). Foresti et al. [3] also labeled when, where, and what (W3) as the three attributes of cybersecurity alerts and events because of their very nature. According to their definitions, when refers to the point in time where the event happened, where to the location of the event that happened, and what to the type of the event. The space of what and where are finite, and the when space is semi-infinite [3]. Finding the relations among these components and answering related questions are essential to analysis [4].

The two most popular methods to visualize and analyze the cybersecurity spatiotemporal data are geospatial visualization and time series curves. (1) The geospatial visualization is usually integrated with a time slider to adjust the time frame. The high-dimensional data are often displayed on the geospatial map with multiple views and layers overlaid with numerous data points, connections, and details. The visualization can easily overwhelm the display space on a single monitor. These types of visualizations challenge human cognition to remember what was seen previously, where it was, and its potential relationship to current information [5]. (2) The time series

*Correspondence: qianz@purdue.edu
†Equal contributors
[1] Interaction Design, Purdue University, 552 W. Wood Street, 47907 West Lafayette IN, USA
Full list of author information is available at the end of the article

curves focus on providing the analyst an overview of how the data change over time. Significant value changes can be clearly reflected on the curve as peaks or valleys, which hint for the analyst to pay attention to these significant situations. This type of visualization is clean and easy to read, but it skips the context of spatial information.

In a survey of cybersecurity visualization techniques, Shiravi et al. [6] argued that user experience should be one of the key issues a successful visual analytics system should consider. The user experience is not only about elegant appearance or powerful functions, but also, and more importantly, about a smooth and fluent analysis process. Heer and Shneiderman also stressed that "visual analytics tools must support the fluent and flexible use of visualizations at rates resonant with the pace of human thought" [7]. However, in most cases the complex data and multiple visualizations lead to poor user experience.

This paper aims to promote the fluency of navigating in the spatiotemporal visualizations and to enhance the user experience of cybersecurity analysis. It demonstrates a solution to link the time series curves, geospatial visualizations, and data views together and to help the user achieve situational awareness through comprehension of the what, where, and when attributes of cybersecurity issues. This paper was originated from our previous work [8] that attempted to link the user from the temporal time series curve to geospatial visualizations. At this paper, we were able to extend the approach and its application to link the user in multiple directions among temporal visualization, geospatial visualization, and data view. We borrow the term "representational fluency" [9] from psychology and pedagogical literature to describe our efforts of enabling the user to fluently switch among different types of spatiotemporal visualizations and to more efficiently solve analysis tasks. The extensions in this paper include:

- A detailed explanation of the mechanism that selects portions from the time series curves and links to spatial visualizations. This mechanism was revised and extended.
- A new mechanism that reversely selects and links spatial visualizations and data views to time series curves.
- New use cases to demonstrate these two mechanisms.
- Redesigned interaction operations that allow the user to access information more smoothly.

To achieve smooth transitions across interactive visualizations, techniques such as brush and linking have been widely used in VA systems. This paper provides a practical technical mechanisms to link multiple visualizations and aims to help users gain better experience and improve performance when analyzing the big network security data visually.

Related work

To enhance the user experience of cybersecurity visual analytics, we suggest adopting representational fluency in designing the structure of spatiotemporal visualizations because "users of this information will need fluency in the tools of digital access, exploration, visualization, analysis, and collaboration [10]". The literature review inspects two main components: representational fluency and spatiotemporal data visualization methods.

Representational fluency of visualizations

The concept of fluency is originally associated with the ability to express oneself in both spoken and written language and to move effortlessly between the two representations. Although fluency is often associated with language, researchers have extended fluency to other fields such as physics, chemistry, engineering, and mathematics. In these fields, fluency is the ability to understand and translate among commonly used modes of representation, such as verbal, mathematical, graphical, and manipulatable. In the context of information systems, fluency is the ability to access, make sense of, and use information to build new understandings [11]. Defined by Irving Sigel [9], representational fluency is the ability to (1) comprehend equivalence in different modes of expression; (2) comprehend information presented in different representations; (3) transform information from one representation to another: and (4) learn in one representation and apply that learning to another.

Representational fluency is an important aspect of deep conceptual understanding. It was mainly discussed in pedagogical literature about promoting the transfer between learning and the development of "expertise". In our context of visual analytics, we borrowed this concept to describe how to let the analyst better comprehending the multiple visualizations of "when, where, and what" for cybersecurity situational awareness. Representational fluency is more skillfulness than skill [12]. Skillfulness connotes continuous adaptation and dynamism along with the ability to perform with facility, adeptness, and expertise. Skillfulness of representational fluency in visual analytics includes several capabilities, such as abstractly visualizing and conceptualizing transformation processes, qualifying quantitative data, working with patterns, and working with continuously changing qualities and trends. To achieve these goals, analysts should be supported with proper tools to interpret visualizations more efficiently.

Visualization methods of W3 attributes

Much previous research has been devoted to exploring different methods to visualize the large-scale high-dimensional datasets. Keim et al. [13] reviewed and summarized recent visualization techniques to deal with large

multivariate datasets. One of their own techniques is a hybrid approach that is scalable with "big-data" visualization [14]. Guo et al. [15] proposed to use multiple-linked views to visualize the multivariate data. Andrienko et al. [4] created a structured inventory of existing exploratory spatiotemporal visualization techniques related to the types of data and tasks they are appropriate for. Based on the W3 attributes, Foresti et al. [3,16] developed a novel visualization paradigm, VizAlert, to visualize network intrusion from all three "when", "where", and "what" perspectives. Con-centric rings were used to represent different time periods, from inside to outside. Because of the limited screen space, the VizAlert system may be unable to display the history for a long period. The user needs to rely on interaction to pan and zoom for shifting between different periods.

Some significant approaches were to analyze spatiotemporal patterns by making separate use of multiple maps and statistical graphs. Alan M. MacEachren's GeoVISTA Center [9] uses highlighting, brushing, and linking, and filtered and linked selections to help users analyze georeferenced time-varying multivariate data. IEEE VAST 2012 Mini-Challenge 1 (MC1) asked researchers to analyze a high-dimensional spatiotemporal dataset [17]. Most of the challenge entries used maps and statistical graphs. For example, Chen et al. [18] and Choudury et al. [19] used one 2-D map to visualize the overall computer statuses in a given time and a slider to adjust the time. Dudas et al. [20] used time series curves to show the aggregate trend of certain qualities.

Analysis process for spatiotemporal cybersecurity data sets

The analysis process on a spatiotemporal dataset often happens conditionally and in sequence [21]. At first the temporal aspect is analyzed, and then the spatial aspect, or vice versa. It is difficult to have a joint integral modeling approach. We have observed such sequential analysis processes in our own practice [22] while solving the VAST 2012 challenge MC2 [17], and in other winning entries [23,24] when they tried to solve the VAST 2013 challenge MC3 [25]. Many times when looking for issues, the user first examined the temporal aspect by looking at the time series curves to find out the anomalies (e.g., huge peak in the curve), then checked out other detailed visualizations to allocate the affecting hosts (IP addresses). Sometimes the analysis starts with a detailed visualization, e.g., an IP address showing abnormal behavior. To understand the overall picture of the affected computers, the user will then need to examine the time series curves. Sometimes this process happens iteratively. The user starts from one visualization, then goes to others, returns to the first visualization with a different parameter (e.g. time or place), and goes on to gain comprehensive cybersecurity awareness. To investigate the detail, the analyst usually need to narrow down and even to read the raw data such as the log file.

Context - data and system

Our implementation of representational fluency was developed on a visual analytics system SemanticPrism [18]. It won the award of "outstanding integrated analysis and visualization" in the VAST 2012 MC1. From 2011 to 2013, the IEEE VAST challenges committee created three cyber-network visual analytics tasks [25] to simulate the complex nature of cyber security. VAST 2011 MC2 data contain 3-day logs of a small computer network. VAST 2012 MC1 data record 2-day logs of a huge global network. VAST 2013 MC3 data include 2-week logs of a 1200-computer network. All the datasets provided are spatiotemporal.

The high-dimensional spatiotemporal dataset we used in this paper was from the VAST 2012 MC1. It simulates a large enterprise network named the BankWorld, which contains approximately a million computers in about 4000 offices. Offices have latitude and longitude information that can be marked on the map. Computers are divided into three classes, server, workstation, and ATM (Automated teller machine). By their functions, Servers are further divided as web, email, file server, compute, or multiple, and Workstations are further divided into teller, loan, or office. Every 15 minutes each computer generates a status log. Within the 48 hour period, the network accumulated approximately 160 million logs. Each log contains a time stamp, IP address, activity flag, policy status, and number of connections (NOC). Policy status has a value range from 1 to 5 to represent healthy status from normal to severe condition. Value 1 means the machine is healthy. 2 means the machine is suffering from mild policy deviation. 3 means the machine has non-critical patches failing and is suffering from serious policy deviations. 4 means critical policy deviations and many patches are failing. 5 means the machine may be infected by virus or unknown files are found. Activity flags also have 5 possible values range from 1 to 5. Value 1 means normal activities have been detected on the machine. 2 means the machine is going down for maintenance. It may appear offline for the next couple time slots. 3 means there were more than 5 invalid login attempts. 4 means the machine's CPU is running at 100% capacity. 5 means a device (e.g. an USB drive or a DVD) has been added to the machine.

The spatial part of this VAST 2012 MC1 dataset contains two layers, the physical geographic location and virtual IP addresses. Its IP space ranges from 172.1.1.2 to 172.56.39.254. The information of both has a hierarchical structure enabling the top-level larger range to be divided into several lower-level smaller ranges. In a real computer network, the geographic locations may range from a continent, a country, a state, or a province to a specific office

in a building. For IP addresses, the network can be divided into multiple levels of subnetworks that are connected through gateways. Each subnetwork occupies a partial IP address space.

With the system SemanticPrism, the analyst is able to see and compare data of different dimensions at multiple granularities. We chose visualization methods and designed interactions based on the nature of the data and the problems faced. SemanticPrism uses a multilinked-view approach to explore the data from different perspectives. Using different transformation methods, data are visualized by using the geospatial map, time series curves, and pixel-oriented visualizations. The technique of semantic zooming [26] was used as the basic interaction technique to navigate through these visualizations. Each visualization has multiple zoom levels to present different levels of details. The analyst can scan to quickly understand the overall situation of the enterprise network and navigate further to read more details of regions, offices, and even the level of individual computers.

Geospatial visualization with a time slider

The default view in SemanticPrism is a geospatial visualization with a time slider that helps to aware the network status at a given time (Figure 1). Offices of the BankWorld are marked as square dots on the analogous world map. Their different color shades indicate the maximum policy violation statuses of the computers within the offices at that time. The analyst can slide the pointer on the time slider to update the geospatial visualization to a different time frame. Different dimensions of the information (e.g. policy status and activity flags) were stacked on the map as different layers. To let the analyst see the global status, SemanticPrism provides a time-zone layer to indicate the local times of different regions in this global organization.

Besides zooming in on and out of the map, the analyst can focus and investigate the data at different levels of details through semantic zooming. Depending on the size of available space, an office can be dynamically visualized at four levels: (1) an individual dot when using the default full-map view or when the space is still quite dense after zooming; (2) a horizontal color bar to show the percentage of computers with different policy statuses in the office; (3) a series of growth curves of all policies in the office where the X axis presents the temporal direction and the Y axis the number of computers; (4) history diagrams of each computer within the office.

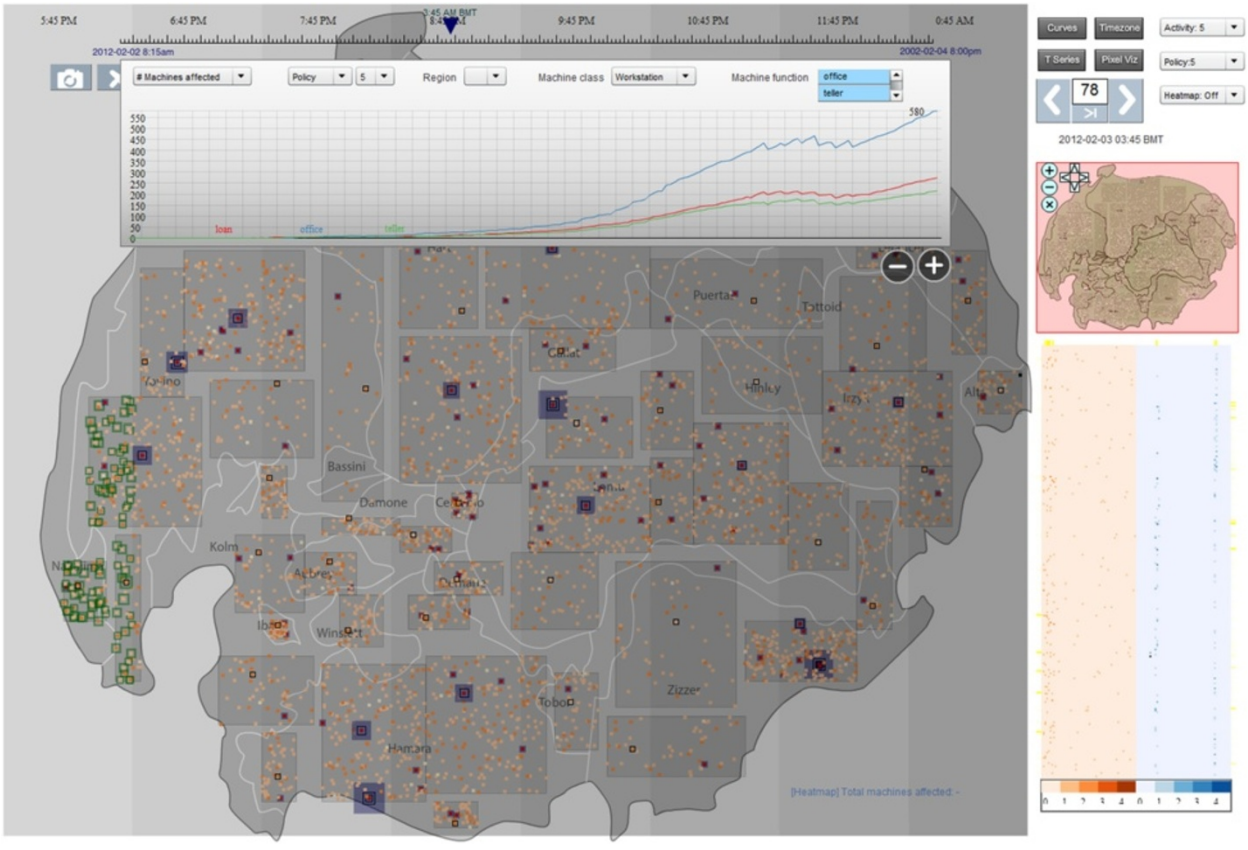

Figure 1 **Default view of the visual analytics system SemanticPrism.**

Time series curves

We adopted Ben Shneiderman's visual information-seeking mantra to guide the design of the SemanticPrism's information query process, "overview first, then zoom and filter, and lastly details on demand" [27]. The time series curves (the curve graph in Figure 1) can be configured to provide an overview of the growth trends of policy statuses, activities, server populations, and NOC (number of connections) over the given period. Figures 2 and 3 curves show the total number of workstations in different policies (2 5) and activities (2 5) over the 2-day period. With the support of time series curves, the analyst can easily identify the overall trend of policy violation growth and patterns of activities. By relying solely on the curves, however, the analyst cannot see the cause, details, and effects of an event. Usually he/she must manually switch to other views to investigate, such as what causes the curve to change, and where the change takes place. This significant user-experience problem motivates our new development of extending user interactions from semantic zooming to marking interesting segments on the curve.

Pixel-based visualizations

IP addresses indicate the virtual locations of network computers. For cybersecurity issues, they provide a different perspective of spatial information than physical locations. The classification of IP addresses also partially reflect the organization's network structure. SemanticPrism incorporated a pixel-based visualization to show many IP blocks. In the default zoom level, five rectangular panels show the number of computers within an IP block that are affected by each activity and policy. In the panels, each pixel represents a group of computers in a particular class-C block. The X axis consists of the IP's class-B block, and the Y axis consists of the values of class-C blocks. The colors of the pixels encode the number of computers that carry the selected policy status or activity flags in the C-level blocks. Through semantic zooming, the analyst is able to overview time series curves of all C-level blocks within one B-level block and all individual computers within a C-level bock.

Mechanism and implementation

SemanticPrism's comprehensive visualizations and interactions show multiple visualizations of where, when, and what data components (Figure 4). With it, we were able to discover all anomalies hidden within the large dataset in the competition. In this paper, we implement the representational fluency concept by extending the interaction design in this system. We consider three important representations for the user to be truly aware of the situation – raw data, spatial visualization, and temporal visualization. We seek to allow the user to shift fluently back and forth among these three representations of the cybersecurity information without losing the analysis context.

Dimension Hierarchy in SemanticPrism

To enhance the efficiency while analyzing a large multidimensional dataset, we adopt the OLAP (online analytical processing) [21] approach to execute analytical queries. OLAP's slicing operation enables the user to take out one specific part of data. SemanticPrism [23] pre-computed the aggregation values along necessary dimensions and storing them into several database tables. The dimension hierarchy is essential for these computations. Pre-computing all possible aggregations on all different granularities, however, will use too many resources. We selected several dimensions to compute in certain granularities.

SemanticPrism maintains a set of dimension hierarchies so that the analyst can have multiple navigation paths to narrow down and examine computers with a certain status (e.g., policy or activity) at a certain time slot and in a certain region. Spatially on a map, a computer is located at the following hierarchy:

Company $\Rightarrow$ *Region* $\Rightarrow$ *Office* $\Rightarrow$ *Computer class*

As virtual IP space, an IP address is located at

Whole IP space $\Rightarrow$ *B-level IP blocks* $\Rightarrow$ *C-level IP Blocks*

In this dataset, computers within one C-level block belong to the same office and are in the same class of server or workstation. But one office may contain many C-level blocks. Therefore the basic aggregation of computers we choose is the number of computers in one C-level IP block with a given policy/activity status at a given time. From such basic units, we can compute the number of computers of on-policy status at one office at one time, then the policy status at the region level, then to the whole company level. Thus we can have different levels of time series curves of different activity/policy statuses, from the basic level of computer classes, to regions, and lastly the entire company. The number of connections (NOC) are more related to IP-related attacks (e.g., port scan); thus we can simply use the IP address hierarchy to divide it.

With the spatial hierarchy and policy/activity status, the system has multiple paths to aggregate the basic units based on the user's analysis needs.

Link data query to visualizations

For visual analysis systems, the raw data are the resource of everything. The more details the dataset contains, the more insights and discoveries can be found. For solving cybersecurity issues, the datasets are usually very large and comprehensive. It is impossible for human beings to read through, compare, and identify issues in the large-scale datasets. Visualization becomes the only feasible way to allow the analyst to make sense of the large amount of data. However, visualizations cannot show all

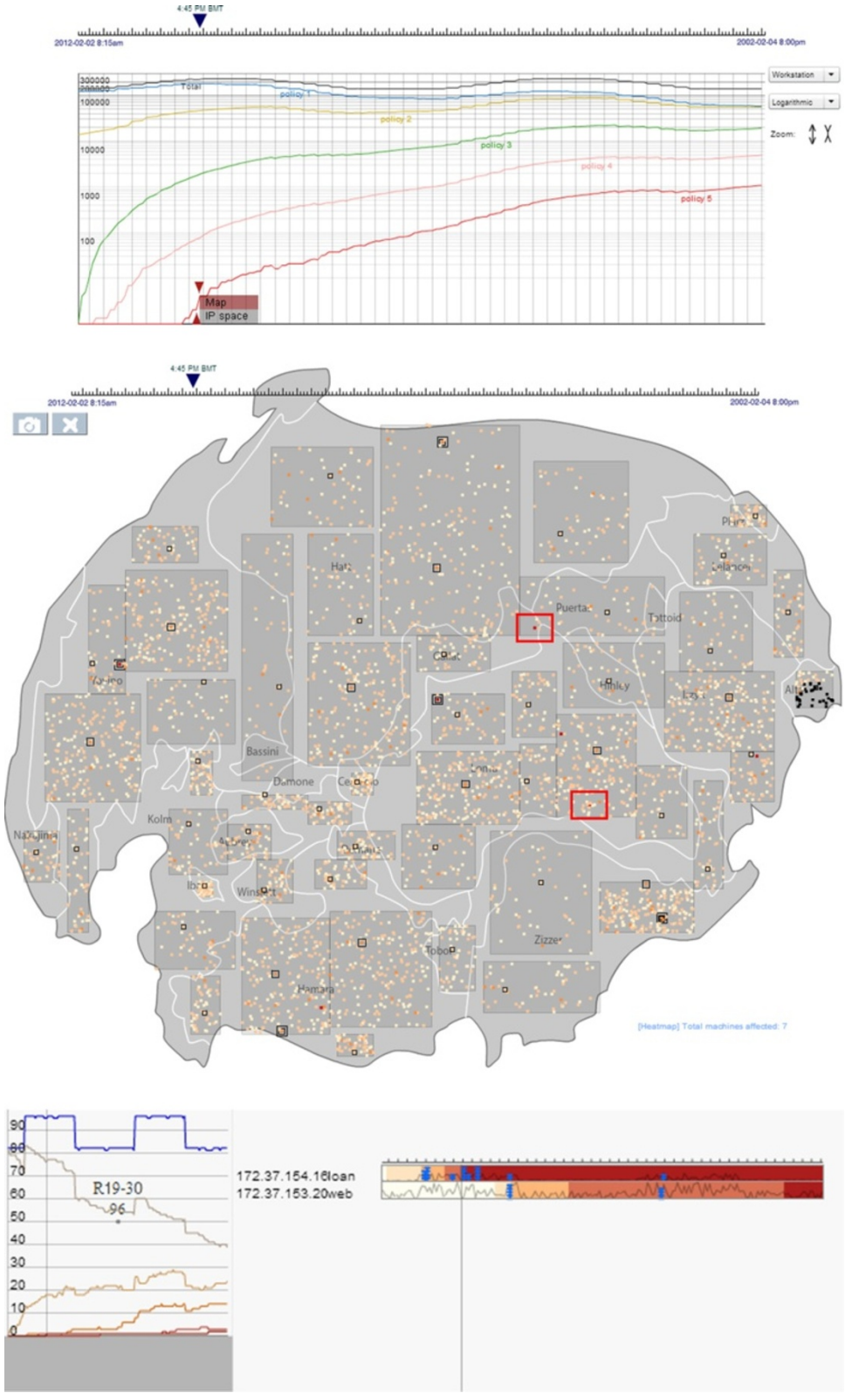

Figure 2 Investigate the number of workstations that violate policy 5. Top: We want to check out which workstations are new to violation of policy 5 at 2012-02002 4:45 p.m. by clicking on the segment in the curve. Middle: The map marks by red squares the two new offices with policy 5 violations. Bottom: Clicking on the top marked square to see the details.

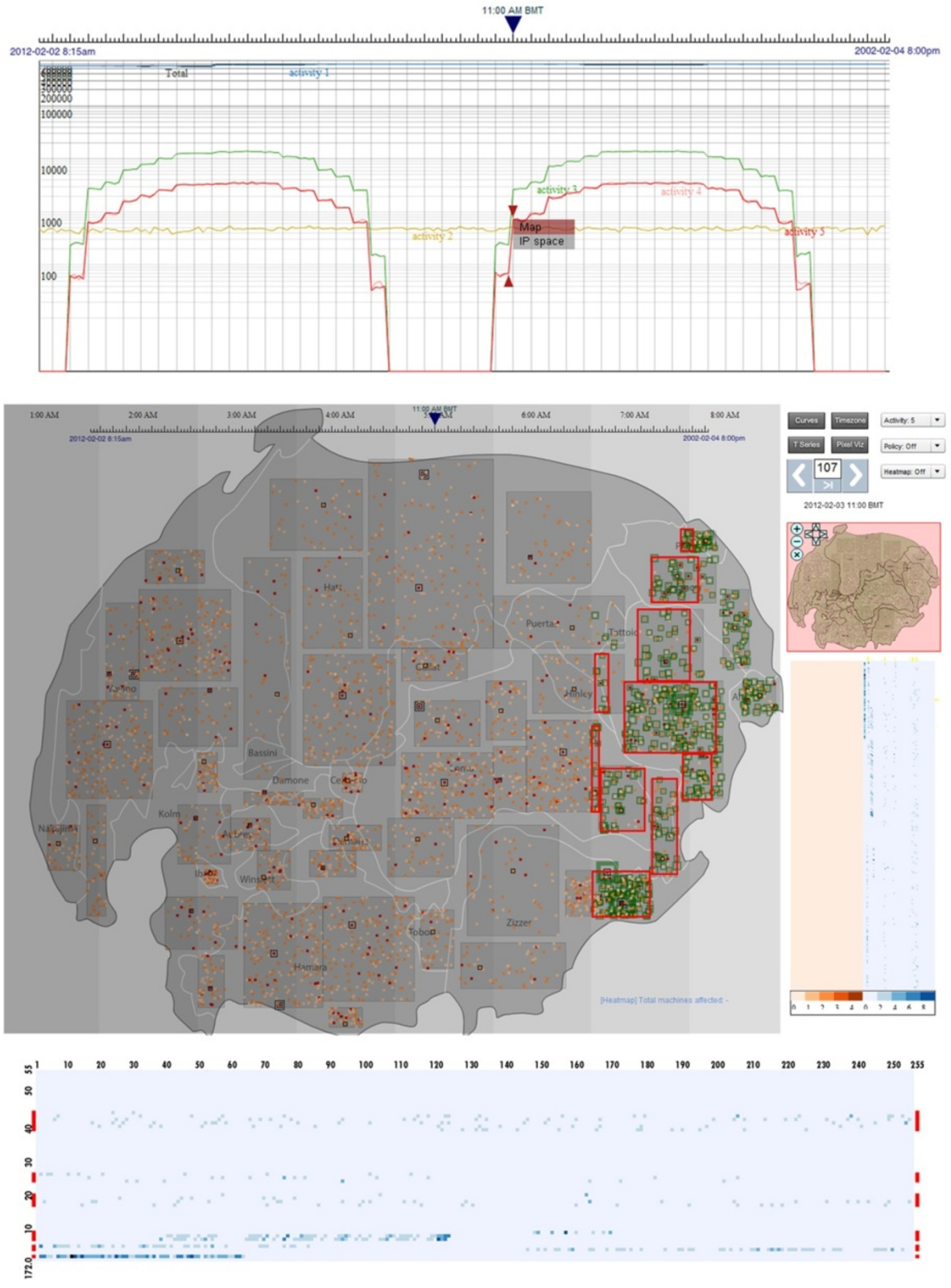

Figure 3 Investigate the number of workstations that have activity flag 5. Top: we want to check out what happened on activity 5 at that time. Middle: The map marks offices by regions. Bottom: These computers are also marked by their IP addresses.

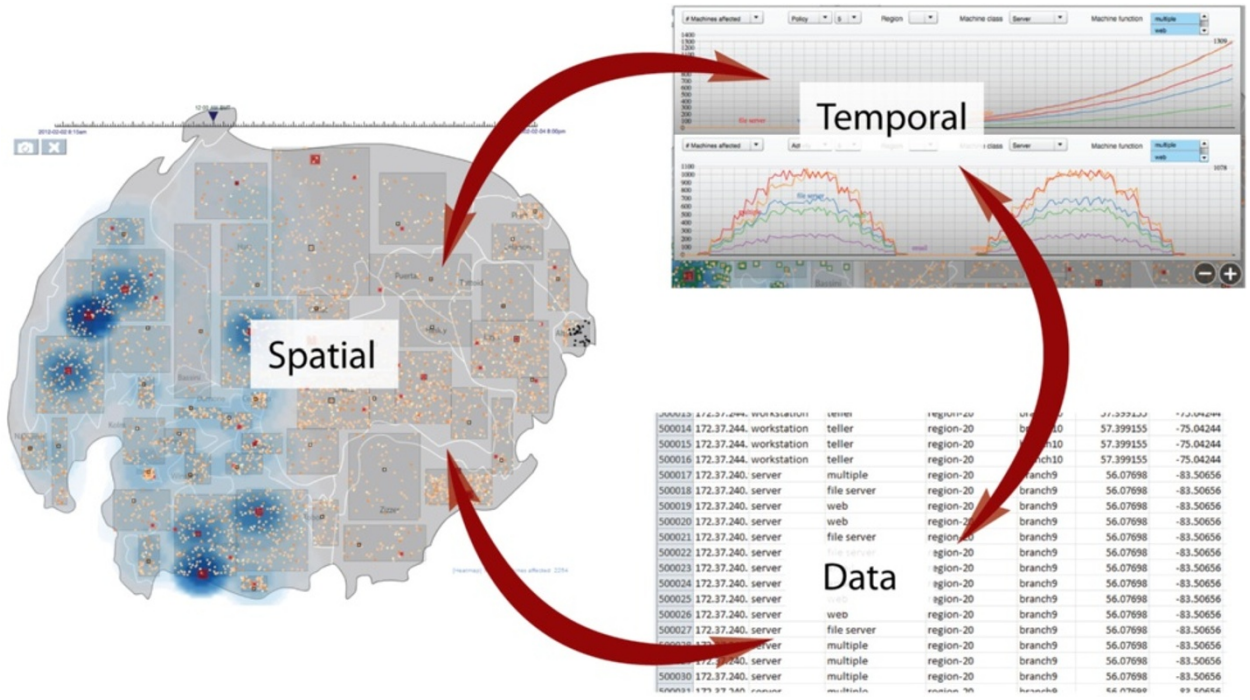

Figure 4 Three components in spatiotemporal data. SemanticPrism's visualizations show: where (spatial data in geospatial visualization), when (temporal data in time curves), and what (data objects in data tables).

the information in the dataset. Through categorization, aggregation, and visualization, only part of the information has been presented in graphs. An analyst still needs to frequently examine the original raw data (e.g., a recorded log or an event report) to determine the exact issue. Thus we should provide a direct-query interface for the user to search the raw data. Based on the searched criteria, the user can pop up visualizations, for example, to display the locations of the computers under investigation in the geographic visualization. Also from the visualizations, the analyst should be able to open, allocate, and read the piece of the raw data of an interesting point.

Link from time series curves to spatial visualization

The time series curve is the visualization to show the plot of the data narration. The data are measured at successive points in the temporal direction at uniform time intervals. In computer networks, it is a common strategy to aggregate (or count) certain network incidents at a given time interval (e.g., 15 minutes in the VAST 2012 data). Thus a series of data points along the time will be generated and can be visualized as time series curves. In our implementation, the curves are plotted on a 2-D Cartesian system with line segments connecting a series of points. X axis is the time direction and Y axis is the value of data. Thus such data have a natural temporal ordering. The user should be able to see the overall trend of the network status through the temporal curve. For a running system, its temporal curve can present certain kinds of patterns (e.g., fixed frequency and amplitude, or various grow rates). For a complex system, such patterns are sometimes hard to define by mathematic equations, and therefore hard to be detected solely by machines. Temporal visualizations rely on a human's visual perception and pattern recognition to help the analyst to detect such potential attacks through recognizing abnormal patterns in time series curves.

Figure 5 lists six popular abnormal situations, including a sudden jump, dive, peak, valley, slop gradient change, and frequency or amplitude change in oscillating curves. In the figure, blue squares mark the data points, and red line segments label the abnormal sections. Such abnormal segments on a curve imply that there are some computers behaved abnormally during that time period. This paper only focuses on the abnormal segments as the four scenarios on the top image of Figure 5. After detecting an abnormal segment on the curve, the analyst needs to investigate what caused the change. He/she should switch from the overview curve to more detailed curves or other visualizations to investigate its when, where, and what details. Although sometimes the abnormal behavior may happen globally, in our observation such behaviors most of the time happen on computers within a small region. Again, such a region could be a physical location or in the virtual space of IP addresses. It is essential for the analyst to find out which region(s) causes these problems in detail.

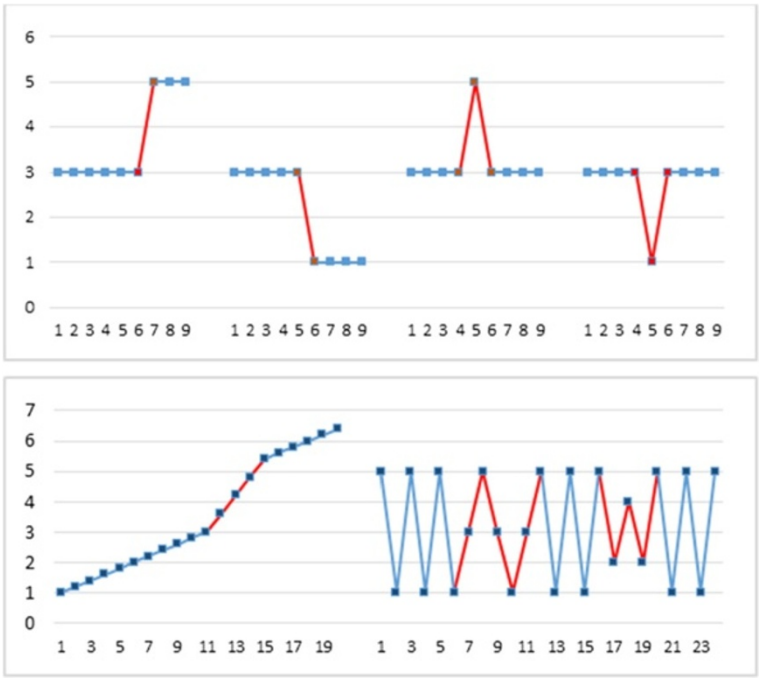

Figure 5 **Examples of abnormal sections (marked in red) on curves.**

While working on a large-scale complex dataset, an analyst will find it tedious and exhausting to examine each individual curve section to learn the related spatial information. This VAST 2012 Challenge dataset includes 4,000 offices and 13,000 C-level blocks. Manually examining each office or C-level block is simply impossible.

A time series curve of higher level granularity (e.g., a region) can be divided into several curves of its subgranularities (e.g., all sub-regions). Spatial data have hierarchy and can be divided into many levels of sub-regions. The aggregated value of the upper-level region is the total of all its sub-regions. For example, the total number of computers in one company must be equal to adding up all computers in its regional offices. Thus an anomaly (e.g., a jump) on a higher level curve must appear on some of its sub-curves. According to our observation, usually only a few sub-curves contribute most of the change in the higher-level curve. Thus it is essential for us to find these sub-curves and allocate the spatial information from them. In reality, curves will fluctuate slightly even in normal conditions. While finding the cause for the anomalies in the curve, we must filter out these small fluctuations.

We store the time series data according to the dimension hierarchies we discussed in the previous section. Using the OLAP slicing operation, we are able to divide an aggregated value into different granularity levels. The overall process of detecting anomaly from aggregated time series curves to geospatial details can be described as follows:

- The system maintains the hierarchies and relations of different levels of subdimensions in different directions. This information enables the system to iteratively check all subdimensions until it reaches the grounded basic granularity.
- The analyst anchors a suspicious segment on the curve. In this operation, the user defines the following parameters: start-and-end times, start-and-end data values, and value difference at this dimension.
- The detailed spatiotemporal data of the suspicious segment can be shown in two ways: locations on the geospatial map at the current time, or many curves "sliced" from the original curve. Based on the nature of the sliced curve, the system may automatically select one direction to show the details, or prompt to ask the user to select a direction to show the captured segment in detail.
- The system checks all of its subdimensions to learn which contribute the most to the overall value difference. This is done by sorting them by their percentages of value changes in the given period. If the percentages are within the same range, the system will rank the subdimensions by their absolute values. The higher the percentage of value change, the more contribution we must consider it will be giving to the overall value change. In some scenarios, some subdimensions may have much smaller values than others. The absolute value change in one smaller region might be too small to contribute to the overall,

although it is still significant enough locally. Thus we consider the relative change of numbers instead of the absolute change.

- The value changes might be caused by one or several subdimensions, or by most (or all) subdimensions. In the evenly distributed scenario, the relative percentage value of each subdimension should be very close.
- The system sums up the absolute number of changes from the highest-ranked subdimension to the lower-ranked subdimensions and tracks all subdimensions until the summary value reaches a certain user-defined threshold (e.g., 95%) of the original value. If we consider threshold as 100%, all small fluctuations will be counted and may blur the focus of the problem.
- If these subdimensions can be divided further to the next level of subdimensions, the system will iterate back to step 4 until the subdimensions are the most-grounded basic granularity.
- Group these basic granularities into a cluster if necessary. Based on the nature of these granularities, some spatial-clustering algorithms, such as DBSCAN (Density-Based Spatial Clustering of Application with Noise) [28], can be used.
- Mark these clusters on the map if these basic granularities are geographic based, or display them as a collection of time series curves if they are still time based.

Link from spatial visualization to time series curves

From a spatial visualization back to a time series curve is relatively straightforward. The spatial visualization normally presents the geographic distribution of different types of data. In the SemanticPrism (and many other spatial visualization systems), which type of data to be visualized can be controlled by menus that turn the data layers on and off. Also with zooming technology, the analyst can zoom in a smaller area on the visualization to open the view of a region or a subnetwork. Therefore from one spatial visualization we can capture a list of parameters, including the type of data items being visualized, current time, and current display area/region/subnetwork. Based on these parameters, popping up the related time series curve is simple.

Interaction design to support the fluency

An analyst may start the investigation by analyzing the curves. The abnormal segments in a curve, like sudden jumps, dives, peaks, or valleys, reflect the value change and therefore present us with a hint that something worthwhile is waiting to be investigated further. As discussed earlier, the time series curve can be seen as a series of vertices with connecting line segments. Thus the analyst can interact with two types of objects on the curve, the vertices and line segments, and can mark the suspicious segments in two ways. The first way is to mark a suspicious one-time-unit segment by simply clicking on the segment. To mark a segment across several time-unit periods, the user clicks on the starting point and end point on the curve and leaves two red triangle marks (top screen shot in Figure 2). After selecting one segment or two vertices, the system will present a pop-up menu for the analyst to select from if there are several possible subdimensions. In the example shown by Figure 3, further details can be shown in either the map or the IP pixel-based visualization. If there is only one subdimension, the system will automatically jump to the detailed view and display the marked area. Because the data are discrete with time intervals, selecting a partial segment is unnecessary. The minimal selectable range should be one segment (or the two neighboring data vertices).

The area of interest on the map is indicated by a red rectangle. Although the offices are spatially spread across the map, they are hierarchically grouped by regions. Therefore we did not use particular spatial clustering algorithms, but rather cluster offices by regions. If two or more offices are in one region, they will be marked together within one block. The boundary of the rectangle is defined by the spatial elements (offices in the middle screen shot, Figures 2 and 3). Sometimes the affected area will be tiny, for example, containing only one office. Marking the tiny area may not be visually significant enough to be noticed. Thus we define the minimum size of a marking as a rectangle measuring 45×35 pixels (Figure 2). The analyst can click on the rectangle to zoom in. The semantic zoom mechanism will automatically display details of the affected offices (middle and bottom screen shots).

Use cases

We use some examples below to show how we implement visual fluency in SemanticPrism, which tries to provide the user a smooth and efficient method to link information from different visualizations.

From time series curve to spatial visualization

In Figure 2, from the time series curve, the analyst saw the increasing number of computers are falling into high policy statuses. To accommodate multiple curves in one graph, we used thin lines in SemanticPrism to draw the curves. To identify how the policy violence spread spatially, the analyst needs to examine the locations of the computers. The user can inspect each segment on the policy-5 curve to check new computers that violate the policy. After clicking on the segment between 4:30 to 4:45 p.m., the user chooses the map from the pop-up menu to see which offices have new computers are new in policy

status 5 starting at 4:45 p.m. The spatial view highlights the two new offices as the middle image in Figure 2. It is possible for the user to highlight all computers by selecting one time point. The user can simply click on the vertex in the curve to highlight all computers having the problem at that time.

In our current implementation, we simply use regions to cluster offices. Thus the two offices are marked separately. Clicking on the red marked boundary will lead the spatial view to zoom to that area. But because only one office is in that region, the system automatically zooms to the maximum level, which shows the detailed information of the office, including time series curves about policies in this office, and shows all computers with that policy 5 violation. We can see the IP 172.37.154.15 just started in policy 5 status at the given time (marked by the gray vertical line to indicate the current time).

Figure 3 demonstrated how the same curve jumps can be marked on either maps or IP addresses. The top image has 6 curves about the number of computers at different activities status (including total number of online computers) along the two days. At each hour there is a step (up or down) on the curves of activity 3 and 4. To find out what causes these steps, the analyst selects and examines one of the jumping segments (top image). By checking out the affected area on the map, he/she can see that they are actually caused by time zones – Offices open at 7 a.m. and close at 5 p.m. As time passes, offices open to turn on computers and close to shut off computers, which causes the sudden steps on the curve. The red squares mark the offices with computers that are newly emerging in activity 5. However, in here the marks are not 100% accurately aligned with the time zone because of the threshold we used (defined in step 6 the previous mechanism section). Small fluctuations happen all the time everywhere, especially for these computer activities, such as log-in errors. We assume that within a large area (e.g., a region), these small fluctuations that happened in small sub regions (e.g., in an office) will be counteracted with each other and make the regional number relatively stable in a normal situation. Therefore smaller areas might sometimes be neglected, or mismarked, as shown in the middle image of Figure 3. But the areas that contributed much to the change will be clearly marked out.

The bottom image of Figure 3 shows the distribution of new computers in the IP space. Each small square in the image represents a C-level IP block. Rows from bottom to top are the 2nd byte of the IP address (from 172.0.xx.xx to 172.55.xx.xx). Columns from left to right are the 3rd byte of the IP address (0 to 255). Besides marking each C-level block with blue squares, we also mark the B-blocks on both the left and right sides with red indicators (Figure 3 bottom).

From spatial visualization to time series curves

SemanticPrism provides a semantic zooming mechanism to change the details of display while the user is zooming in [22]. Offices on the map can change into 4 levels of details, depending on the available on-screen space. When zoomed in enough, the user is able to see the time series curves for individual offices (Figure 6).

Besides using semantic zooming to check out time series curves of different offices, the user can also click on a region or an office to see the temporal summary. Region 25 (the right-most region Alta at the top image of Figure 6) has many blacked-out offices, which means that these offices are disconnected from the Internet, possibly because of a power outage in the area. We can see that the distribution of blacked-out offices changes as time passes. To get to the affected computers over time, we can click on the region to bring out a regional time series curve (top image of Figure 7). The black curve shows that the overall computers sent out status reports during the period. A big valley on the curve shows that more and more computers lost connections in the middle of the first day. The worst time was at 11 p.m. BMT (Bankworld Mean Time). The situation recovered in the next 4 hours. The analyst can also choose to turn on the layers to highlight one activity status or one policy status. The green squares surrounding offices on the top image show offices with computers at activity 2 (going down for maintenance). Since the activity 2 layer is currently turned on, the time series curve of activity 2 is also included in the curves. However, the number of computers with activity 2 is so small, at the pixel level it is at the baseline and hides behind the policy 5 curve. The middle image in Figure 7 uses a logarithmic scale to boost these curves with extremely small values on the screen. Zooming in on this curve will break down the time series data of the region into individual offices. These curves for all offices are displayed in a grid as the bottom image of Figure 7.

Discussions

Visual analytics is the process for an analyst to learn the facts from the large volume of raw data through different forms of visualization. Representational fluency is the ability to comprehend equivalence in different modes of expression [9]. We borrow this term from psychology and pedagogical literature to describe our efforts to enable the analyst to fluently switch among different types of visualizations and data views to build up the understanding of facts. Cybersecurity issues can be visualized in temporal, in geospatial, in structural, or in raw data as logs. Visual analytics fluency allows the ability (1) to transform information from one representation to another; (2) to comprehend the equivalence in different modes of representations, including data and visualizations; and

Figure 6 Zoom in to see different levels of office details.

(3) to comprehend information presented in different representations.

In this paper, we propose an auto linking mechanism that can smoothly transfer the analyst from one view to the other and thus effectively improve the speed of visual data analysis. Cognitively, a person can pay attention to only 3 or 4 things at one time. Our fluency metaphor may also reduce the cognitive load, helping the analyst to focus

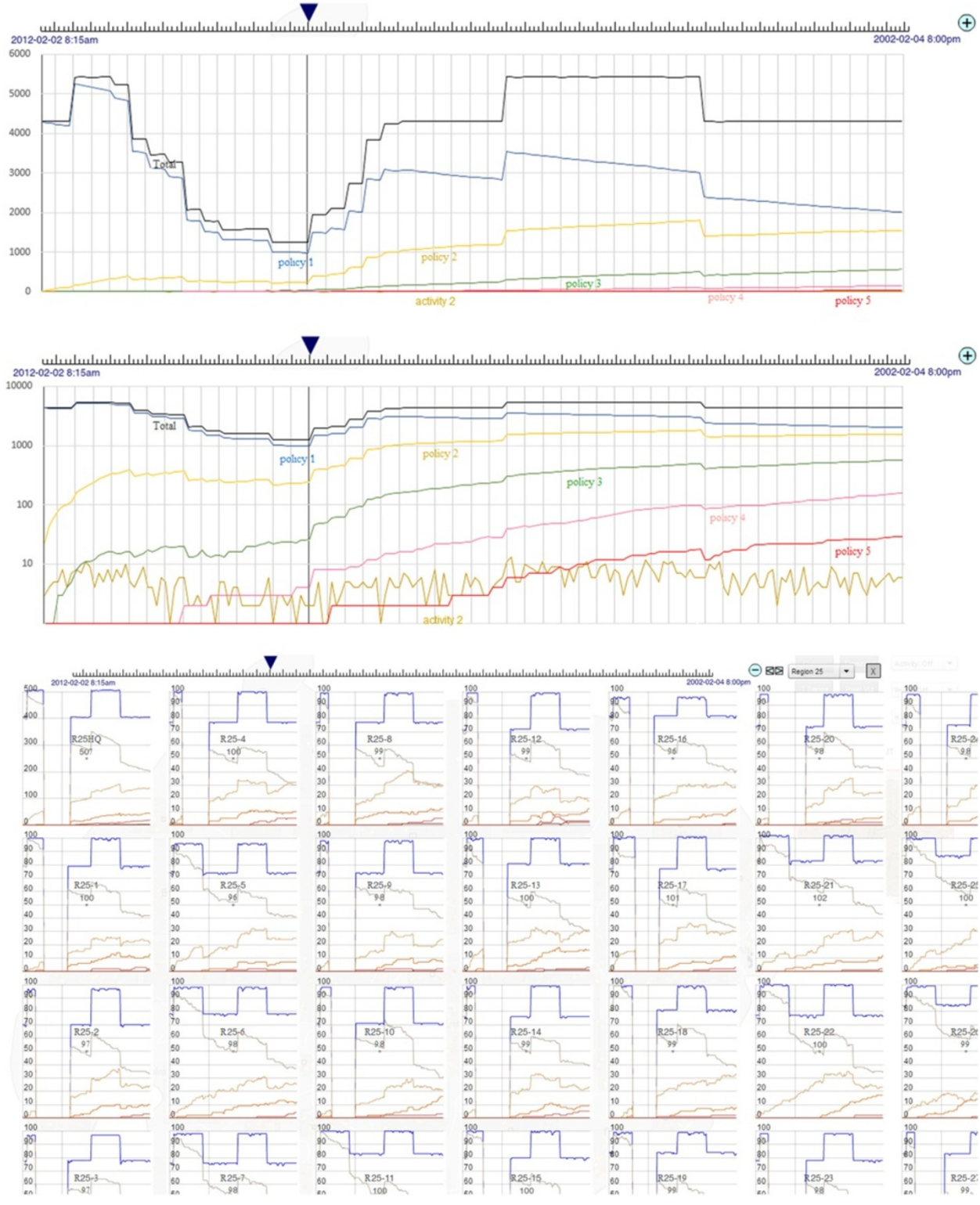

Figure 7 Time series curves corresponding to the region 25 in Figure 6. Top: Overall number of computers in different policy flags in region 25. Middle: The logarithmic scale version of the top visualization. Bottom: Zoom in on the curve to break down the regional time series curves into offices.

on some important incidents. At the stage of submitting SemanticPrism to the VAST 2012 challenge (July 2012), the four team members needed several days to identify all the anomalies by manually going over suspicious areas on all the curves and jumping across different views to examine and filter information. Most of the energy and time was exhausted during the back-and-forth navigation. With this newly developed linking mechanism, on one hand an analyst can mark suspicious segments on the time series curves and go directly to its related spatial visualization and data view. On the other hand, the analyst can simply right-click on the map, opening the menu to show one or several related time series curves.

We plan to improve this mechanism and its direct interaction design in the following directions.

First, we should extend our approach to other types of data and visualizations. The VAST 2012 MC1 dataset contains no data about computer network connectivity. In some cybersecurity analysis scenarios, visualizing such connections as the network intrusions from external IPs to internal hosts is crucial. Most often, connection data of these kinds can be visualized as a tree, or a network graph, with different layout variations (e.g., layout nodes in radial fashion). How to anchor parts of such spatial visualizations and link them to their related time series curves, geographic visualizations, or data views comprise the new domain we want to explore.

Second, we should find a method to automatically detect anomalies on the curves. A curve must be displayed at a certain resolution to allow the analyst to identify problematic areas. However, because the curves are mostly based on aggregation, the user sometimes cannot visually detect the problem when the number is too small to cause a significant visual change on the curve. Some literature on data mining and statistics [29,30] shows that allowing the system to detect anomalies on the curves by itself is possible. We will consider integrating this effective approach.

This approach can also be easily extended to handle streaming data such as real time analysis. In such case, the time series curve will become dynamic by updating itself in regular time intervals. Visually the curve will grow, extend, and slide from right to left (if the new data starts from the right end) just like the electrocardiography. Old part of curve will disappear on the left end. The user still be able to notice the anomaly happened during the recent past time intervals. For the just past time interval, the aggregations should be computed across the hierarchy of the spatial structure from top to bottom. The computing resource needed for pre-compute the aggregation depends on the length of the time interval and the complexity of the spatial structure. For this VAST 2012 MC1 data, since the time interval is pretty long as 15 minutes and there are only several thousands of spatial units, computing aggregations for one time interval is very fast. For existing computed aggregations of each time interval, there is no need to re-compute them. The only aggregations need to be updated are the aggregations about recent past history (e.g. recent two days). But normally there is no urgent need to get the aggregation for the past history in real-time.

The inspiration and implementation of this fluency mechanism were based on the visual analytics system SemanticPrism and the VAST 2012 challenge dataset. To understand its generalizability and limits, we will use other datasets to test the possibility of linking the W3 structure visualizations. Furthermore, we aim to study the possibility of representational fluency being a suitable and valid design goal in the context of visual analytics and how to promote it to different platforms and systems.

Competing interests
The authors declare that they have no competing interests.

Author details
[1]Interaction Design, Purdue University, 552 W. Wood Street, 47907 West Lafayette IN, USA. [2]Computer Graphics Technology, Purdue University, 402 S. Grant Street, 47907 West Lafayette IN, USA.

Received: 24 January 2014 Accepted: 4 June 2014
Published online: 15 July 2014

References

1. G Jiang, G Cybenko, Temporal and spatial distributed event correlation for network security, in *American Control Conference, 2004, Proceedings of the 2004*, vol. 2 (IEEE Boston, MA, USA, 2004), pp. 996–1001
2. DJ Peuquet, It's about time: A conceptual framework for the representation of temporal dynamics in geographic information systems. Ann. Assoc. Am. Geographers. **84**(3), 441–461 (1994)
3. S Foresti, J Agutter, Y Livnat, Moon S, R Erbacher, Visual correlation of network alerts. IEEE Comput. Graphics Appl. **26**(2), 48–59 (2006)
4. N Andrienko, G Andrienko, P Gatalsky, Exploratory spatio-temporal visualization: an analytical review. J. Visual Languages & Comput. **14**(6), 503–541 (2003)
5. J Booker, T Buennemeyer, A Sabri, C North, High-resolution displays enhancing geo-temporal data visualizations, in *Proceedings of the 45th Annual Southeast Regional Conference* (ACM New York, NY, USA, 2007), pp. 443–448
6. H Shiravi, A Shiravi, AA Ghorbani, A survey of visualization systems for network security. IEEE Trans. Visualization Comput. Graphics. **18**(8), 1313–1329 (2012)
7. J Heer, B Shneiderman, Interactive dynamics for visual analysis. Mag. Queue - Microprocessors. **55**(4), 45–54 (2012)
8. YV Chen, ZC Qian, From when and what to where: Linking spatio-temporal visualizations in visual analytics, in *IEEE International Conference on Intelligence and Security Informatics* (IEEE Seattle, WA, USA, 2013), pp. 39–45
9. IE Sigel, Approaches to representation as a psychological construct: a treatise in diversity, in *Development of Mental Representation: Theories and Applications* (Psychology Press East Sussex, UK, 1999), pp. 3–12
10. M Stone, Challenge for the humanities, in *Working Together or Apart: Promoting the Next Generation of Digital Scholarship* (Washington, DC, USA, 2009). The Council on Library and Information Resources and The National Endowment for the Humanities
11. B Stripling, Assessing information fluency: gathering evidence of student learning. School Library Media Activities Monthly. **23**(8), 25–29 (2007)
12. (R Lesh, H Doerr, eds.), *Beyond constructivism: a models and modeling perspective on mathematics teaching, learning, and problem solving*. (Lawrence Erlbaum Associates, Hillsdale, NJ). ISBN 0-8058-3822-8

13. D Keim, C Panse, M Sips, ed. by Dykes J, Maceachren A, and Kraak M, Information visualization: Scope, techniques and opportunities for geovisualization, in *Exploring Geovisualization* (Elsevier Ltd Oxford, UK, 2005), pp. 23–52
14. DA Keim, C Panse, M Sips, SC North, Pixelmaps: a new visual data mining approach for analyzing large spatial data sets, in *Proceedings of the Third IEEE International Conference on Data Mining* (IEEE Los Alamitos, CA, USA, 2003), pp. 565–568
15. D Guo, J Chen, Eachren MacAM, K Liao, A visualization system for space-time and multivariate patterns (vis-stamp). IEEE Trans. Visualization Comput. Graphics. **12**(6), 1461–1474 (2006)
16. Y Livnat, J Agutter, S Moon, S Foresti, Visual correlation for situational awareness, in *Information Visualization, 2005. INFOVIS 2005. IEEE Symposium on* (IEEE Minneapolis, MN, USA, 2005), pp. 95–102
17. KA Cook, G Grinstein, M Whiting, M Cooper, M Havig, K Liggett, B Nebesh, CL Paul, VAST challenge 2012, visual analytics for big data, in *Visual Analytics Science and Technology, 2012 IEEE Conference on* (IEEE Seattle, WA, USA, 2012), pp. 151–155
18. VY Chen, AM Razip, S Ko, CZ Qian, DS Ebert, SemanticPrism: a multi-aspect view of large high-dimensional data: VAST 2012 mini challenge 1 award: Outstanding integrated analysis and visualization, in *Visual Analytics Science and Technology, 2013 IEEE Conference on* (IEEE Seattle, WA, USA, 2012), pp. 259–260
19. S Choudury, N Kodagoda, P Nguyen, C Rooney, S Attfield, K Xu, Y Zheng, BLW Wong, R Chen, G Mapp, L Slabbert, M Aiash, A Lasebae, M-sieve: a visualisation tool for supporting network security analysts, in *VisWeek 2012*, 165–166 (2012)
20. L Dudas, Z Fekete, J Gobolos-Szabo, A Radnai, A Salanki, A Szabo, G Szucs, OWLAP - using OLAP approach in anomaly detection, in *Visual Analytics Science and Technology, 2012 IEEE Conference on* (IEEE Seattle, WA, USA, 2012), pp. 167–168
21. O Schabenberger, CA Gotway, *Statistical Methods for Spatial Data Analysis*. (CRC Press, Boca Raton, FL, USA, 2004)
22. VY Chen, AM Razip, S Ko, ZC Qian, DS Ebert, Multi-aspect visual analytics on large-scale high-dimensional cyber security data, in *Information Visualization 2013* (Sage Publications Thousand Oaks, CA, 2013)
23. Y Zhao, X Liang, Y Wang, M Yang, F Zhou, X Fan, MVSec: a novel multi-view visualization system for network security, in *VisWeek* (2013)
24. S Chen, F Merkle, H Schaefer, C Guo, H Ai, X Yuan, T Ertl, Annette - collaboration oriented visualization of network data, in *VisWeek* (2013)
25. M Whiting, KA Cook, CL Paul, K Whitley, G Grinstein, B Nebesh, K Liggett, M Cooper, J Fallon, VAST challeng 2013: Situation awareness and prospective analysis, in *Visual Analytics Science and Technology, 2013 IEEE Conference on* (IEEE Atlanta, GA, USA, 2013)
26. K Perlin, D Fox, Pad: an alternative approach to the computer interface, in *Proceedings of the 20th Annual Conference on Computer Graphics and Interactive Techniques* (ACM New York, NY, USA, 1993), pp. 57–64
27. B Shneiderman, The eyes have it: A task by data type taxonomy for information visualizations, in *Proceedings of 1996 IEEE Symposium on Visual Languages* (IEEE Boulder, CO, USA, 1996), pp. 336–343
28. M Ester, H-P Kriegel, J Sander, X Xu, A density-based algorithm for discovering clusters in large spatial databases with noise, in *KDD, vol. 96 1996* (AAAI Portland, OR, USA), pp. 226–231
29. RS Tsay, Outliers, level shifts, and variance changes in time series. J. Forecasting. **7**(1), 1–20 (1988)
30. JD Hamilton, *Time series analysis*, vol. 2. (Cambridge University Press, Cambridge, UK, 1994)

doi:10.1186/s13388-014-0006-4
Cite this article as: Qian and Chen: Fluency of visualizations: linking spatiotemporal visualizations to improve cybersecurity visual analytics. *Security Informatics* 2014 **3**:6.