Analyzing Hospitalization Records of a Pandemic
VAST 2010 Mini Challenge #2

Yuzuru Tanahashi Yingcai Wu Kwan-Liu Ma
Visualization and Interface Design Innovation (VIDI) Group
University of California, Davis

Abstract
Analyzing data collected from hospitals and health care centers to make prompt responses is crucial in preventing epidemics from getting out of control. Health organizations are constantly on the search of unusual trends in such data, and a single day of delay may cause catastrophic damage. Although rich datasets provide a solid foundation for a deep analysis, it is often difficult to deal with a large quantity of data to make a quick response. In this paper, we describe an intuitive visualization system for users to visually analyze the hospitalization records and gain a quick insight into the data.

1 Overview of the Tool
The system we had developed for this task involves two components, interactive line charts and stack graphs.

The line charts facilitate the users to analyze data of multiple regions simultaneously, such as daily mortality rates shown in Figure 3. Due to the volatility of the data, the line charts are enhanced with an interactive smoothing function, so that the users can quickly switch their focus from a detailed concise view to a simplified abstract view.

The stack graphs facilitate the users to analyze data of multiple symptoms simultaneously. Considering the inefficiency of having a legend for all symptoms, we added an interactive highlighting function that dynamically generates a filtered legend on the stack graphs. Therefore, the users can easily identify the symptoms with significant features.

2 Data
In order to facilitate an effective analysis, the data provided from the VAST 2010 Challenge #2 needed to be filtered into a homogeneous format and compiled into a structure easy to manipulate.

At first, the original data contained almost 1,500 kinds of medical abbreviations of symptoms. By applying a hash table for diverse medical abbreviations, identical symptoms with various descriptions were unified, and the total number of symptoms were reduced to 720.

Next, due to the quantity of the data, we aggregated each regions’ daily data and compiled it into a percentile format. Considering that the objective is to analyze individual symptoms, we assigned each symptom two percentile properties, later referred as HS (Hospitalized Symptom) and DP (Death Probability).

\[ \text{HS-X}^1 = \text{Percentage of Patients with Symptom X}. \]
\[ \text{DP-X}^2 = \text{Mortality Rate of Patients with Symptom X}. \]

\(^*\)ytanahashi@ucdavis.edu, ycwu@ucdavis.edu, ma@cs.ucdavis.com

Table 1: A Prototype of the Data Format

<table>
<thead>
<tr>
<th>Country</th>
<th>Date</th>
<th>Patients</th>
<th>MR (^3)</th>
<th>HS-A</th>
<th>DP-A</th>
<th>HS-B</th>
</tr>
</thead>
<tbody>
<tr>
<td>Colombia</td>
<td>7/22</td>
<td>1043</td>
<td>0.015</td>
<td>0.01</td>
<td>0.008</td>
<td>0.23</td>
</tr>
<tr>
<td>Turkey</td>
<td>7/22</td>
<td>1000</td>
<td>0.010</td>
<td>0.02</td>
<td>0.010</td>
<td>0.41</td>
</tr>
</tbody>
</table>

Table 1 shows how we stored the data. The data of each day in each region contains the number of patients, the mortality rate of the patients hospitalized on that date, and individual HS and DP value per symptom.

Figure 1: Stack graphs of HS data. X-axis: Date. Y-axis(Width): HS value. Drastic changes in HS values are highlighted with high opacity. As a result, "VOMITING", "FEVER", "DIARRHEA", "BACK PAIN", and "ABDOMINAL PAIN" are highlighted in most of the plots.

3 Analysis
3.1 Analysis of the Disease
In Figure 1, symptoms such as "VOMITING", "FEVER", "DIARRHEA", "BACK PAIN", and "ABDOMINAL PAIN" were strongly highlighted for most regions. This indicates that these symptoms had a considerable rise in the percentage of the patients. Therefore, these symptoms can be easily identified as the primary symptoms of the disease.

Figure 2 shows the stack graphs of the DP values of the four symptoms "VOMITING", "FEVER", "DIARRHEA", and "AB-

\(^3\)Mortality rate of the day.
DOMINAL PAIN” for each region. For simplicity, we omitted "BACK PAIN" due to its moderate rise in HS value compared to the other four symptoms. In these stack graphs, all regions except Thailand and Turkey show significant rise in DP values for all four symptoms, each stack stabilizing around the value of 12%. We can also observe the DP values in these stack graphs taking about three to five days to reach their stable values. From these common temporal features, we can derive that the disease had about 12% of mortality rate and three to five days of incubation period.

Figure 2: Stack graphs of DP data of "VOMITING", "FEVER", "DIARRHEA", and "ABDOMINAL PAIN". The red boxes indicate the incubation period of the disease. From May 30th (blue line), the DP value declines significantly in all areas. Each stack stabilizes around 12%, which can be considered as the mortality rate of this disease.

3.2 The Temporal Pattern of the Out-break

From Figure 3 (a) and (c), we can see that the mortality rates start to rise around 23rd of April (circled in red), and reaches its peak around May 15th (lined in red), when smoothing is applied. Considering the epidemic had a direct affect on the mortality rate based on the former analysis, we can conclude these dates as the onset and the peak of the epidemic.

Also, by comparing the onset timings in Figure 3 (especially (c)), we can hypothesize that the epidemic started in Nairobi, then traveled into Middle East (Aleppo, Karachi, Saudi Arabia, and so on) and South America (Colombia and Venezuela). This hypothesis can also be confirmed by closely examining Figure 1 and 2.

However, while all the other regions show significant rise in certain HS values and mortality rates in Figure 1 and 3, Thailand and Turkey show little movement. This indicates that these two regions had no significant influence by any specific disease. Therefore we can conclude that the epidemic had not reached these two regions.

3.3 HIV Patients

Figure 4 shows the stack graphs of DP values for all symptoms. From these stack graphs, we can observe that the patients with not only the key symptoms but also other miscellaneous symptoms had higher mortality rates during the period of the epidemic. Especially in these graphs, many symptoms containing the keyword "VAGINAL" are highlighted. By investigating such symptoms as "VAGINAL BLEEDING", we were convinced that these symptoms have a good chance of indicating HIV patients. Based on this hypothesis, Figure 4 also succeeds in conveying the vulnerability and the distribution across the world of HIV patients.

Figure 4: Stack graphs of DP values for all symptoms. X-axis: Date. Y-axis(Width): DP value. Stacks are highlighted according to their width. As a result, symptoms containing the keyword "VAGINAL" (colored in red) are highlighted.

4 CONCLUSION

This paper describes a visualization system for analysts to visually analyze the hospitalization records of a pandemic to make a quick response. We built the system using two well-established techniques (line charts and stack graphs) to ensure intuitive analysis. This tool enables analysts to visually compare the data among different regions or different symptoms simultaneously. With the tool, we have successfully discovered some interesting patterns such as the incubation period of the disease and the vulnerability and distribution of possible HIV patients.