Statement of Work

GHOST: Global Hepatitis Outbreak and Surveillance Technology

Team 23:

Raghav Kaul

Jeongsoo Kim

Ernest Lai

Sarthak Mohapatra

Lovissa Winyoto

Georgia Institute of Technology

In partnership with:

David S. Campo, Ph.D.

Centers for Disease Control (CDC)

Introduction

Hepatitis C is a contagious liver disease that is hard to detect and spreads rapidly. According to the Centers for Disease Control and Prevention (CDC), almost 3 million individuals in the US may be afflicted with chronic Hepatitis C. In response, the Division of Viral Hepatitis at the CDC has developed Global Hepatitis Outbreak and Surveillance Technology (GHOST), a system for the detection of Hepatitis outbreaks in the United States. The system is being used internally at the CDC as well as by affiliated State Departments of Health (SDH), who may submit “jobs” involving genomic analysis for the CDC’s high-performance computing (HPC) cluster to process. After the HPC cluster sequences the genomes, it returns a dataset to the user containing information on the relationship between potential patients, based on the similarity of those genomes.

GHOST accomplishes two goals: first, it is the front end for a database of previously sequenced patient genomes. Similarities between the genomes of two patients, A and B, imply that patient A may have infected patient B, vice versa, or that the two patients may be linked through a third patient, C. Second, once a researcher navigates to a specific genomic analysis, GHOST allows visualization of the data in the form of a node-link diagram of patient-to-patient relationships, where nodes are patients and links represent the strength of their genomic relationship.

Problem Statement

GHOST enables public health researchers, who generally are not specialized bioinformaticians, to glean actionable insights from scientific data. However, feature and interface improvements are needed in order to allow GHOST to better achieve this goal. For GHOST, a less intuitive, less easy-to-use system implies fewer tangible returns from health research. Users are currently unable to fully utilize the substantial computational power from the CDC’s system either for lack of knowledge or because the options for computational job submission and visualization are not easily accessible.

The system is also inaccessible to the users without the technical background to understand the role of genomics in Hepatitis transmission, which creates a problem when users are overwhelmed with data they are unable to understand or make inferences from.

The specific problems with GHOST's user interface that our project will address are divided into the following two major categories:

User interface issues

1. The landing page does not allow the users to visualize the structure of the system/environment.
2. The dashboard does not allow all of the information to be viewed in a single area and rather includes extraneous and redundant links.

Data visualization issues

1. The data visualization, in the form of a node-link diagram, does not encode enough of the patient data (geographic, demographic, or temporal) and their history causing poor functionality.
2. The data visualization does not allow users to select/highlight patients and get details on demand of the selection, slowing users down by forcing them to read large volumes of plain text.
3. The data visualization does not have an option to reduce data through filtering or aggregation, necessitating the redundant and time-consuming process of submitting new jobs.
4. The nodes in the data visualization do not convey useful information based on their shape, color, size, or orientation.

Objectives

User interface redesign is important for GHOST because a more user-friendly "sink" for the data pipeline - from molecular sequencing processing and analytics, all the way to actionable insights, can move the needle for patient outcomes by enabling researchers in high-risk areas where Hepatitis C is prevalent.

Functional Requirements

1. The landing page must provide an opening dialog to first-time users, introducing them to the system.
2. The nodes in the node-link diagram must provide more detailed information (e.g. geographic distribution of candidate genomes, demographic information on the candidate, and temporal data on the date of detection) upon selection.
3. The visualization interface must provide a feature to filter out nodes by characteristics or user selection.
4. The nodes in the node-link diagram must be reconfigured to provide additional information based on their shape, color, size, or orientation.
5. The user dashboard link must redirect users to a single page with all of the relevant information consolidated into a single visualization.

Nonfunctional Requirements

1. The code for the web interface should follow the open-closed principle, where it needs to be closed for modifications but open for extensions.
2. The web interface should take no more than 3 clicks from the main screen to navigate to any other screen.
3. The web interface should be usable by multiple users by performing tasks concurrently with no decrease in job turnaround time from the original GHOST system.
4. The code for the interface should be documented for all methods with top-level visibility.

Applicable Standards

Since the CDC is a federal agency under the Department of Health and Human Services (HHS), regulations for federal public websites are the only applicable standards which affect our project. We will therefore be complying to Section 508 and Accessibility standards that ensure that all government owned websites or documents have to be accessible to people with disabilities.

Solution

The solution that we have proposed for this project is to redesign the interface and visualizations for the CDC's Project GHOST platform by modifying and rewriting the existing code in D3.js, a Javascript framework for data visualization and user interface design. This redesign will be broken up into three major parts: the dashboard, the landing page, and the node-link visualization. The dashboard revision will involve a consolidation of related but external functions and an expansion of the core visualization to improve the usability of the section. The landing page revision will improve the modularity of the design to allow individual user to become more quickly acquainted with the layout and function of the system. Finally, the redesigned node-link visualization will allow users to gain further insights into the data being displayed by encoding additional information and adding new controls to the visualization.

Functional Requirements

1. The landing page will provide an opening dialog to first-time users, introducing them to the system with a modal introduction with an attached walkthrough of the system that will be provided in video and text format.
2. The nodes in the node-link diagram will provide more detailed information (e.g. geographic distribution of candidate genomes, demographic information on the candidate, and temporal data on the date of detection) upon selection.
    - Demographic information will be toggled on and off in a toolbox appended to the corresponding node
    - Selecting a node or group of nodes will bring in a side pane for additional geographic, demographic, and temporal data on the node or group of nodes.
3. The visualization interface will provide a feature to filter out nodes by characteristics or by user selection.
    - The visualization will allow users to click-and-drag select one/multiple nodes to put into focus.
    - Buttons for filtering nodes by shared, strain, and genotype traits, the button for exporting the canvas image, and sliders for threshold and view will be replaced by widgets in the diagram canvas and click listeners for nodes that follow D3's enter-update-exit data model.

4. The nodes in the node-link diagram will be reconfigured to provide additional information based on their shape, color, size, or orientation. The nodes will have high/low saturation corresponding to high/low v-frequency, an attribute that reflects of the experimental significance of a certain node.
5. The user dashboard links will redirect users to pages with all of the relevant information, consolidated, on a selected topic. Links, Alerts, and News will be contained within a single Notifications section and will be shifted to the top of the dashboard to be displayed within the visualization space to allow for a single point of focus.

Nonfunctional Requirements

1. The code for the web interface should follow the open-closed principle, whereby it needs to be open for extension but closed for modifications. The Javascript overlay will be implemented in an object-oriented and modular fashion to allow for future extension.
2. The web interface should take no more than three clicks from the main screen to navigate to any other screen. Since there is only one interaction that takes more than three clicks, this will be achieved by removing the additional data analysis pop-up box interface and appending it to the existing landing page.
3. The web interface should be usable by multiple users by performing tasks concurrently with no decrease in job turnaround time from the original GHOST system. The separation of the front-end and the back end will allow this, since the front-end will be updated in a way that will allow it to maintain its independence from the GHOST back-end.
4. The code for the interface will be documented as it is written and will undergo documentation revision/updates as the solution is being constructed. Before submitting to the client, a formal code review process will be undertaken in order to ensure validity and consistency in both code and documentation.

System Architecture

System Architecture Components

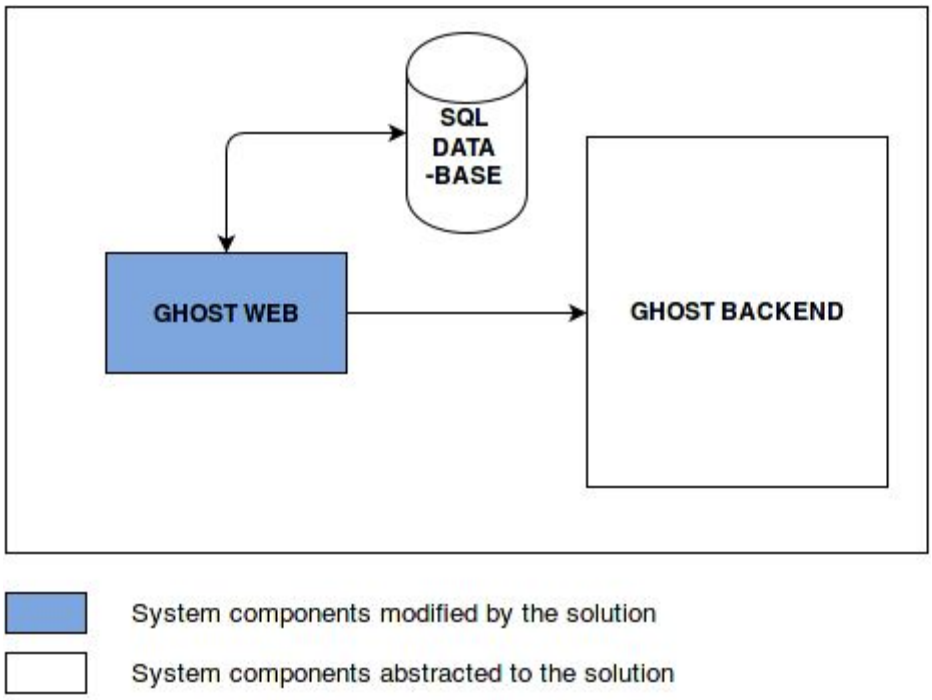

Figure 1. System Architecture Diagram.

System Architecture Components

The components shown in Figure 1 are:

1. GHOST Web - The CDC-hosted website for the GHOST project that contains the visualization.
2. SQL Database - The database that contains the information of the nodes.
3. GHOST Backend - This refers to the entire existing CDC system, including a filesystem for analyzed sequences, an intermediate “worker” system, and a high-performance computing cluster that performs the analysis.

Traceability Analysis

FR - Functional Requirement NFR - Nonfunctional Requirement	**GHOST Web**	**SQL Database**	**GHOST Backend**
FR1 - Landing page navigation	X		X
FR2 - Node detail upon selection	X	X	X
FR3 - Node filter ability	X	X	X
FR4 - Node representation configuration	X		X
FR5 - User dashboard link	X		X
NFR1 - Open closed principle	X	X	X
NFR2 - Three of clicks per screen	X		
NFR3 - Task performance	X		X
NFR4 - Documentation	X		

Existing System

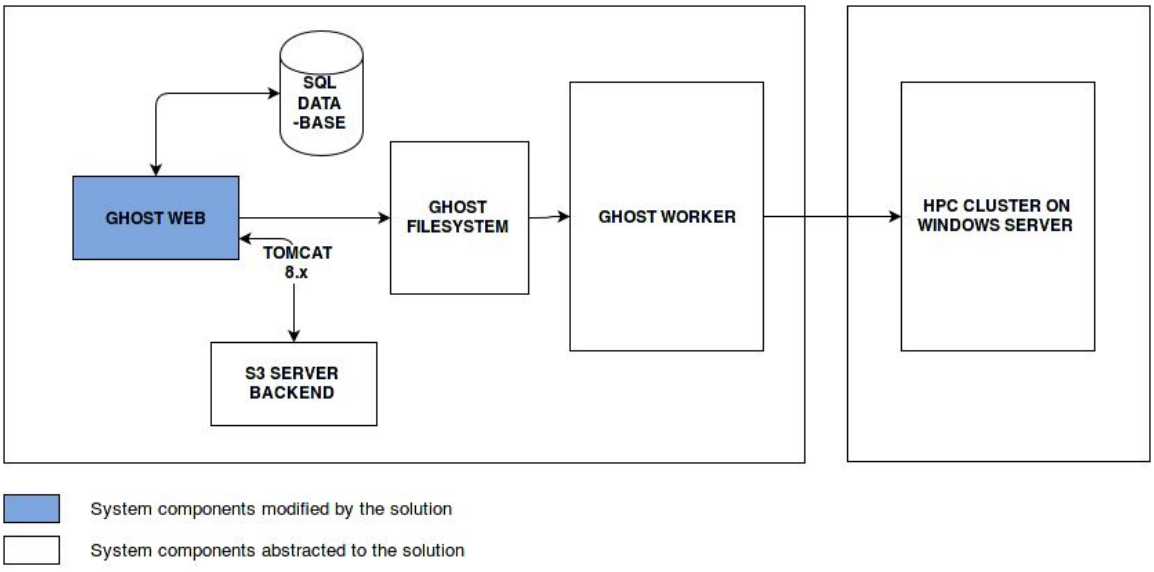

Figure 2. Existing System Architecture Diagram.

User Demographics

The users of the system will be members of the CDC and state general health practitioners involved in the reporting and tracking of Hepatitis C outbreaks. It is assumed that there is no difference in usability scenarios between these two types of users. It is also assumed that the users will have minimal technical knowledge of the system and should be able to use it as easily as possible to complete their task.

Usability Scenarios

Dashboard Visualizations - The user would like to view a visualization of the different tasks that have been submitted on a state-by-state basis. The user begins at the landing page after logging in. The user selects the dashboard link. The system displays the dashboard visualization. The user scrolls to and selects the appropriate visualization from the left side of the page. The system highlights this option and opens the visualization in the center of the page as displayed in Figure 3.

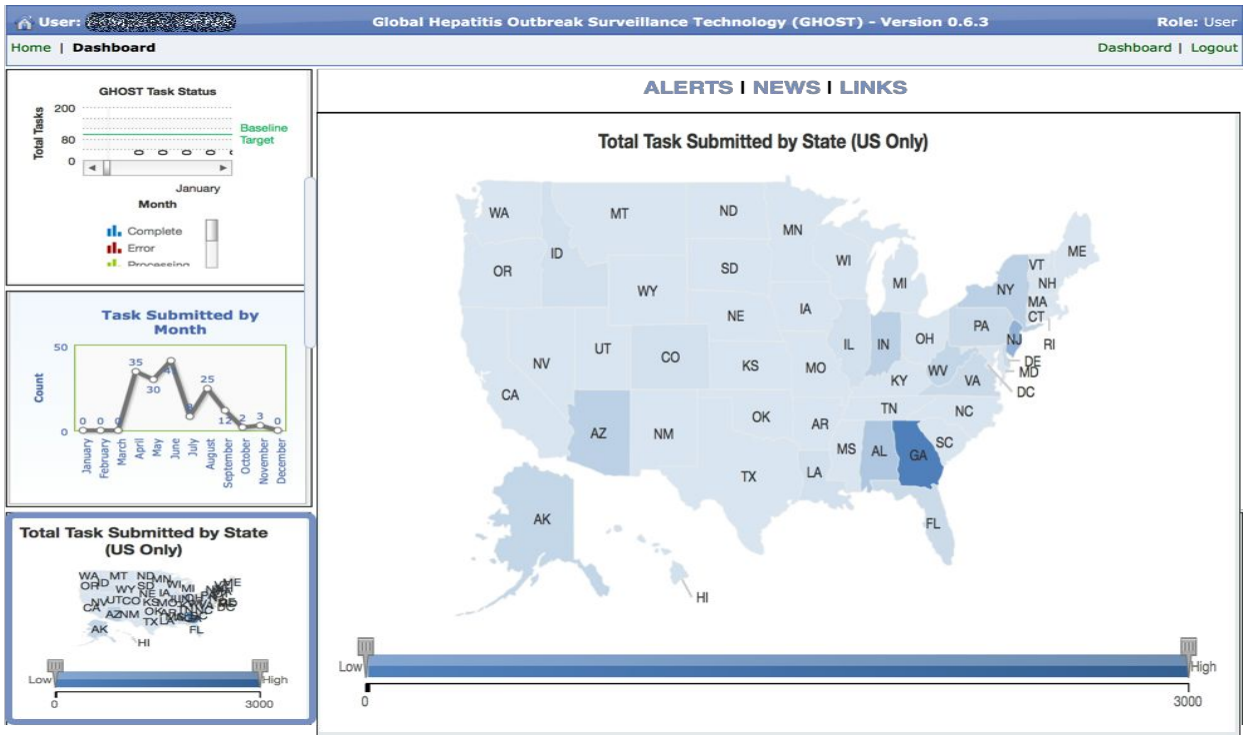

Figure 3. GHOST task submission screen..

Dashboard Notifications - The user would like to view the various alerts that have been added to the section. The user begins at the landing page after logging in. The user selects the dashboard link. The system displays the dashboard visualization. The user selects the appropriate link in the notifications section above the central visualization. The system highlights this option and opens the information in the center of the page as displayed in Figure 4 below.

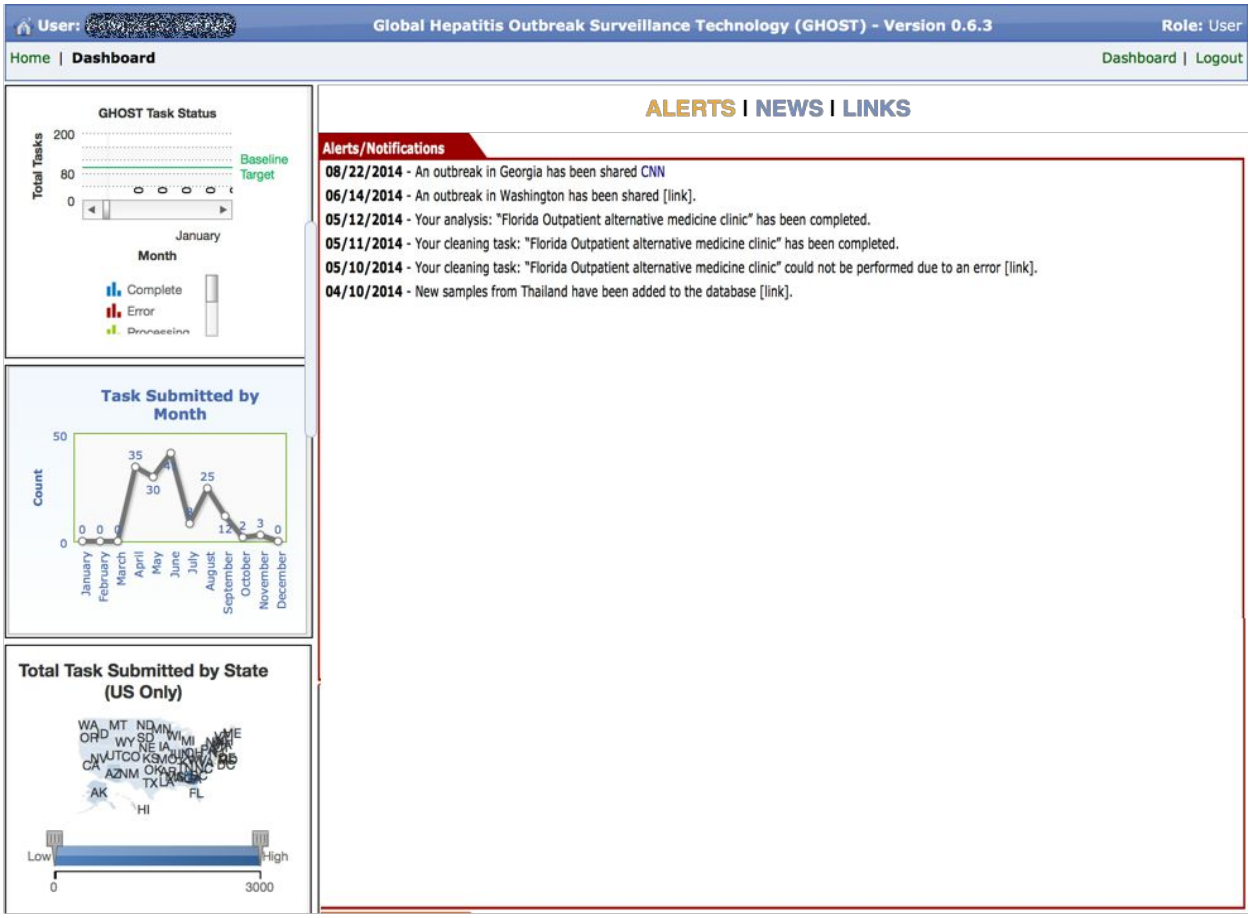

Figure 4. GHOST notification screen.

Landing Page - The user would like to log in to the system in order to upload and analyze data. When the user accesses the GHOST website, the system shows the user a warning message about authorization policies to access GHOST website. If the user clicks “I Accept the Terms”, the system displays the main page of the GHOST website to the user. Alternatively, if the user clicks “I Decline the Terms”, the system directs the user to CDC homepage. Then the system continues to the GHOST landing page.

1. If the user is a first time user of the GHOST website, then the system displays the main page of GHOST website with a pop-up dialog including the option of a text or video tutorial to help the user become familiar with the layout and function of the website. When the user selects “continue”, the system returns to the landing page. This has been displayed in Figure 5 below.
2. If the user has previously used GHOST website, the system displays the main landing page of GHOST system to the user directly.

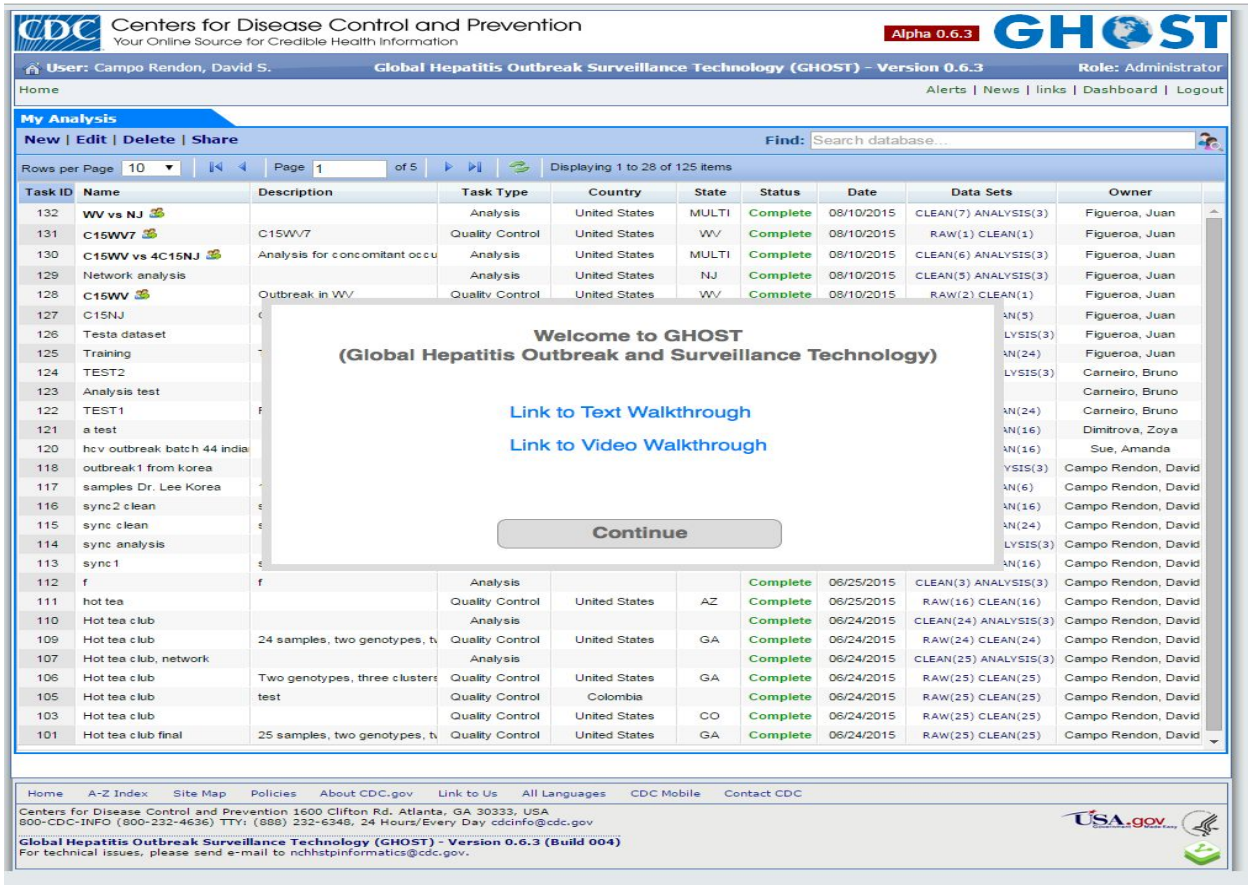

Figure 5. GHOST main page screen.

Data Visualization - The user has uploaded and analyzed files and has pulled up the node-link visualization. The system has displayed the diagram for the currently selected data sets. The datasets in their raw form (JSON and plain text) may be downloaded by clicking the filenames in the page header. The user may select, for further analysis, one or multiple nodes by

clicking/holding down the Ctrl key and clicking. The visualization then outlines the nodes in a contrasting color to identify which ones are in the current selections as displayed in Figure 6.

Further options for filtering include showing patient similarities by the viral strain ID or genotype. The user may adjust the threshold for the link strength between nodes, with a higher threshold eliminating links between nodes with weaker significance, casting those nodes away from their cluster. User updates are made dynamically on the client side, with no interaction with the backend or database except the initial data load. With the information from the visualization, the user may navigate back to the main page to submit an additional job, or export the visualization using the “download” icon in the bottom right hand side of the canvas.

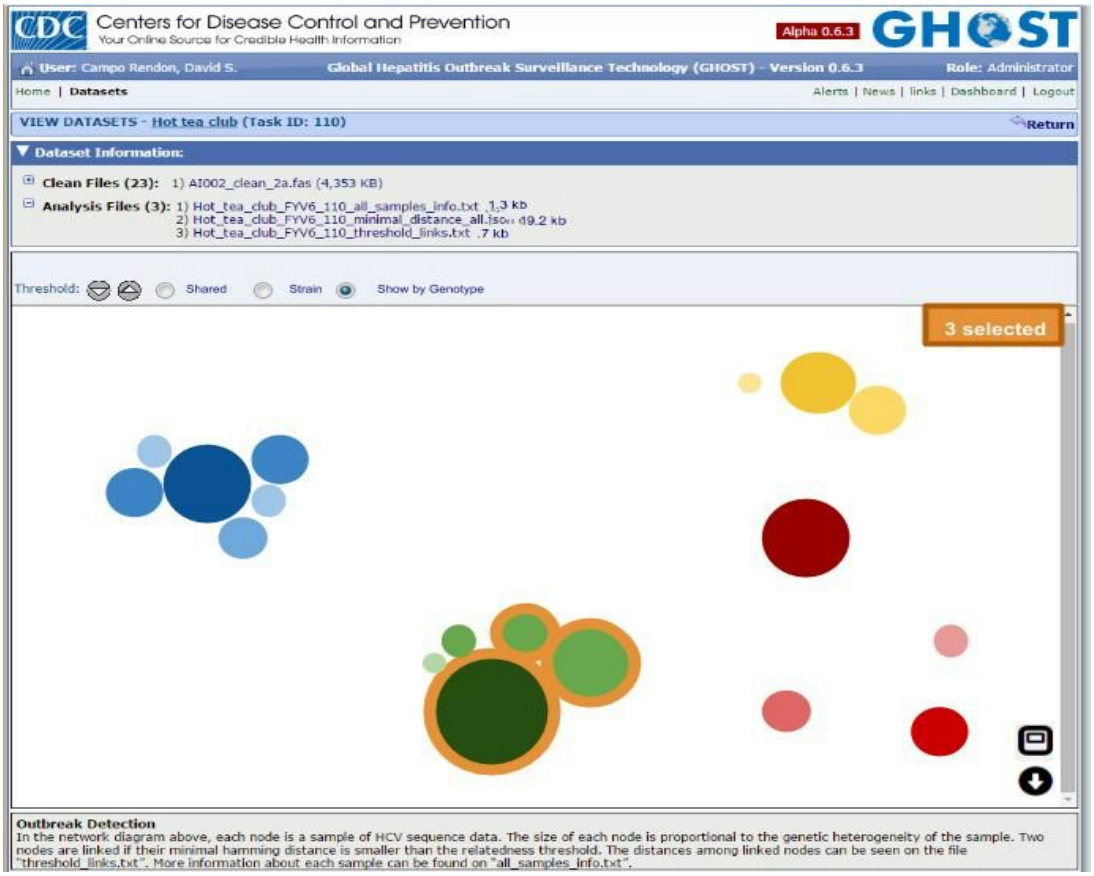

Figure 6. GHOST node-diagram visualization.

Resources

Provided By Team

Resource name	Purpose	Source
Development machines	Laptops or desktops are needed to research and develop the application.	Team members (Cost: $0)
Visualization libraries	Visualization libraries are needed to implement a force-directed layout for the graph visualization.	Michael Bostock, BSD license (Cost: $0)
Development environment	Development environments are needed to write code, test, and deploy the application.	Eclipse foundation, Eclipse Public License (Cost: $0)

Provided By Client

Resource	Purpose	Source
Sample bulk datasets	Sample datasets allow us to test the design of the visualization on JSON objects	CDC DVH Team (Cost: $0)
Web source code	Web source code is needed to test integration and compatibility with the existing web interface.	CDC DVH Team (Cost: $0)

Schedule

Release 1: Redesign node-link diagram visualization

Feature	Requirements Met FR - Functional Requirement NFR - Nonfunctional Requirement
Node zoom and filter	FR3, FR4, NFR1, NFR4
Node-link details on demand	FR2, NFR1, NFR4

Release 2: Redesign landing page

Feature	Requirements Met FR - Functional Requirement NFR - Nonfunctional Requirement
Landing page user profiles	FR1, NFR1, NFR4
Landing page job processing options modal	FR1, NFR1, NFR4

Release 3: Redesign user dashboard

Feature	Requirements Met FR - Functional Requirement NFR - Nonfunctional Requirement
User dashboard consolidated links	FR5, NFR1, NFR2, NFR3, NFR4

Release 4: Testing and Documentation

Feature	Requirements Met FR - Functional Requirement NFR - Nonfunctional Requirement
Documentation and code review	FR2, FR3, FR4, NFR4

References

[1] http://www.cdc.gov/amd/pdf/factsheets/amd-projects-ghost.pdf

[2] http://www.cdc.gov/hepatitis/hcv/cfaq.htm#cFAQ21