TreeCovery: Synchronized Dual Treemap Visualization for Exploring the Recovery Act Report

Miguel Rios  
Department of Computer Science  
University of Maryland  
migue.rum@gmail.com

Puneet Sharma  
Department of Computer Science  
University of Maryland  
puneet@cs.umd.edu

Rachel Schwartz  
Department of Computer Science  
University of Maryland  
rachel13@cs.umd.edu

Tak Yeon Lee  
Department of Computer Science  
University of Maryland  
tylee@umd.edu

ABSTRACT
The American Recovery and Reinvestment Act dedicated $787 billion to stimulate the US economy and mandated the release of the data describing the exact distribution of that money. The dataset is a large and complex one; one of its distinguishing features is its bi-hierarchical structure, arising from the distribution of money through agencies to specific projects and the natural aggregation of awards based on location. To offer a comprehensive overview of the data, a visualization must incorporate both these hierarchies. We present TreeCovery, a tool that accomplishes this through the use of two synchronized treemaps. The tool includes a number of innovative features, including coordinated zooming and filtering and a proportional highlighting technique across the two trees. TreeCovery facilitates the exploration of the data in an intuitive way and promises to be helpful in insight generation.

Keywords
Information visualization, treemap, polyarchy, multiple hierarchy, Recovery Act

1. INTRODUCTION
In February 2009, President Obama signed an economic stimulus package into law, dedicating $787 billion to create jobs and give a boost to the economy, with the provision that the distribution of the money would be completely transparent. In fulfillment of this requirement, the agencies in charge of distributing the money and all recipients issued periodic reports detailing how the money they controlled was spent. These publicly available reports comprise a large amount of data, containing information about the effectiveness of the stimulus package, the general trends of distribution, and potentially interesting outliers.[7]

Some effort has already been expended toward producing visualizations of this data that could assist in revealing such details. The government commissioned a website, Recovery.gov, dedicated to this purpose, and several independent journalism outlets have produced their own applications, all offering a particular take on the data. Most of the existing visualizations consist primarily of either tabular or geographical displays. In addition, while many provide some interaction, most do not allow users full control over the data visualized.

While the data lends itself well to geographical layout, given that states and counties are convenient schemas for chunking the data, exclusive use of maps cannot adequately portray alternate views of the monetary distribution. Specifically, money was distributed through 28 agencies, who assigned it to projects at their discretion; funding was placed in the charge of the prime recipient, who in turn funded sub-recipients and/or vendors as necessary for the project. Agencies naturally funded projects nationwide, and recipients for each project were not necessarily all located in the same area. This view of the data - an agency > project > recipient hierarchy - cannot be adequately conveyed by a geographical substrate.

Our tool, TreeCovery, offers a way to explore data both geographically and according to the monetary outlays. TreeCovery accomplishes this goal through the use of two synchronized treemaps, one drawn with a geographic hierarchy and the other one with levels corresponding to the agency > project > recipient money flow. While the views presented by the two treemaps differ, the underlying data remains identical at all times. Filtering is synchronized across the views and a proportional highlighting technique is used for coordination.

In addition to the synchronized treemap design, we incorporated a few other features to improve exploration techniques. From news articles about Recovery Act we found many of them using demographical statistics such as population or unemployment rates. We thus included census data for each
Figure 1: Recovery.gov offers geographical maps showing each award as a dot.

Figure 2: The Stimulus Tracker from msnbc provides combinations of bar charts, line graphs and geographical maps.

county, and also made it possible to filter by demographic attributes. We also added the ability to save snapshots of the current state of the treemap for later comparison. Finally, we included support for emphasizing invalid data values.

The rest of the paper is organized as follows. In Section 2, we discuss related work. Section 3 provides an explanation of our analysis process, including a detailed illustration of Spotfire’s ability to support exploration of Recovery Act data. Section 4 explains TreeCovery in detail, while Section 5 offers some sample insights found by the tool. Section 6 contains our conclusion and future work.

2. RELATED WORK

Because the stimulus information is both newsworthy and publically available, many visualizations of the data are already available. First and foremost, recovery.gov offers geographical maps displaying award locality (Figure 1). The maps can be zoomed in to state and zip code level and show dots each representing a project colored by its award type - contract, grant and loan. The site also offers some pie and bar chart summaries, as well as tabular data. While the basic information is thus available, interaction with the visualization and customization are limited.

Many other websites offer similar tools to those of recovery.gov. The Federal Procurement Data System\(^1\), which has the raw data available for download, offers a few selected slightly interactive visualizations as well; again, interactivity is limited and the data available through the visualizations is limited as well. The website ProPublica hosts a feature with recovery act information, Eye on the Stimulus\(^2\). Most of the site is devoted to text articles, but tabular and geographical visualizations of spending progress area offered as well. Other sites following the geographical/tabular trend include the Wall Street Journal \([8]\) and msnbc (Figure 2)\([9]\).

The existing visualizations of stimulus data, while informative, do not support exploratory analysis of the data. Because the data contains dual hierarchies, the geographical one and the monetary distribution, it can be most effectively portrayed using visualizations tailored toward this structure.

\(^1\)https://www.fpds.gov

\(^2\)http://www.propublica.org/ion/stimulus

Several methods of achieving this have been discussed in the literature. Polyarchy Visualization allows representation of intersecting hierarchies \([14]\). While useful, this approach does not provide well for visualization of multidimensional data, as is necessary for recovery.gov data. Multitrees describe the general structure of the data: nodes that are shared by multiple ancestor trees \([13]\). However, they are presented as a graph theoretic rather than visual concept and, as such, are not particularly helpful in building our tool.

Treemaps have proved very useful for displaying hierarchical data\([12]\). Burch and Diehl introduce the “Trees in a Treemap” technique to represent trees with an associated taxonomy \([10]\). The taxonomy is represented as a treemap, and the related tree is drawn on top. In this representation, the two (or more) hierarchies are not represented in a symmetric fashion; attribute information about the hierarchy can be easily integrated, but the nodes of the trees are not well described besides for their structural position. Therefore, this approach does not suit our data. Jern, et al., suggest using a treemap visualization in combination with a cartographic one \([11]\). They demonstrate the value of multiple displays of the same data, but their treemap hierarchy is regional. The stimulus visualization requires a method that represents data in multiple hierarchies.

One final treemap-based hierarchical representation is that of Wood et al. \([15, 16, 17]\). A specialized ordering is used to facilitate special and temporal locality so that the layout of nodes is more intuitive. Wood’s approach combines two hierarchies (temporal and spatial) into a single treemap. We felt it was important for users to be able to explore each hierarchy in isolation, as well as in conjunction with the other. TreeCovery therefore uses two treemaps to represent the data. In this way, both hierarchies are transparently represented, and each can display its own attribute values. Cross-hierarchical exploration is supported through synchronization of the treemaps.

3. ANALYSIS

To design our tool, we first needed to determine the chief goals of stimulus data visualization. As recipient reports of Recovery Act had just been released, it was not easy to find end-users who had already done extensive work on the
data. Thus, instead of using direct interviews or a survey, we decided to analyze headlines of relevant news articles in order to understand the process of journalists analyzing it. Further, we analyzed the data with Spotfire, one of the most versatile visual analytic tools. These exploratory tasks helped us to understand the data better and to develop the concept of our tool.

3.1 News Headlines

Journalists are the most prominent group investigating Recovery Act data. From their work we can infer what people want to know and how the data can be analyzed. We collected headlines from the first two weeks of November 2009, just after the recipient reports had been publicized. We searched in the Google News search engine by the keyword ‘Recovery Act’ and focused on articles referring to the recipient reports. We found some interesting patterns of analytic tasks done on the Recovery Act report.

Most prominently, many findings focused on a State/County-wise comparison. Although geographical region is not the main hierarchy of the Recovery Act plan, the most frequently asked question was something like: “How much money is given to our state/county?” For example, an article compares the amounts of awards given to two states - “Idaho Gets Four times More Stimulus Money in contracts Than Louisiana” [4]. It is noteworthy that state/county-wise comparisons require aggregation of projects in each state or county.

The second pattern we perceived was usage of census data. In order to find states/counties in similar context or to validate fairness of funding from a specific agency, census data is quite useful. For example, an article referred to high-school graduation rates, infant mortality rates, unemployment rates, and juvenile justice incarceration to pick the 5 worst cities for youth and compared numbers of jobs created by the Recovery Act in those places[1]. Both of those patterns indicate that the agency and spatial hierarchies are equally important when analyzing the Recovery Act data.

The third pattern we noticed was validity checking to reveal unlikely numbers and non-existing categorical values. As each recipient report was submitted through an online form by the recipient, it is natural for the reports to have some errors due to simple mistakes. Non-existing congressional district codes are a typical case of simple mistakes criticized by CNN[3]; however, other cases more complicated than simple mistakes can be identified. For instance, ‘Number of jobs created’, which is an important gauge of success, can be interpreted in different ways. Several news headlines pointed out projects that created too many or too few jobs[2][6]. Usually, invalid values are either resolved or ignored in information visualization, although they have significant importance especially for a federal government website.

Summing up the findings above, we came up with the idea of bi-hierarchical data exploration in Figure 3. While the Recovery Act funding is distributed to recipients along the agency tree, the information of recipients is also aggregated by county and state in combination with census data.

3.2 Spotfire

Before designing our own tool, we tried to analyze the recipient data report of contracts with an existing application. Spotfire was chosen because of its wide set of features for dealing with multivariate data. This analysis had three purposes. First of all, analyzing the actual data gave us a deeper understanding of the data. Second, we determined the capabilities and limitations of current visualization techniques on this specific data. Lastly, the exploration was helpful in suggesting opportunities for improvements on existing visualization techniques.

In addition to the recipient report, we also incorporated the State and County QuickFacts dataset from the US Census Bureau. The census data consists of a wide variety of demographic profiles for each county, including Population, Infant deaths, Housing unit, Household income and Unemployment rate. (For the full list, see Appendix A)

Paragraphs below are examples of headlines we found by using Spotfire.

The most effective job creators are suspect. Knowing that there are projects with an improbably high number of jobs created, we drew a treemap visualization (Figure 4) showing which departments or states were related to those projects. Colors of elements represented money per job, which means how much money was spent for creating each job in a project. We had to filter projects by money per job in order that small important outliers be visible in the treemap. We found that treemap visualization was suitable for displaying job creation anomalies.

Florida, the most senior state in the US, gets the most money from military sources. We tried to incorporate census data with recipient reports in this example. In Fig-

---

1. http://spotfire.tibco.com
Figure 4: Treemaps can show the distribution of Recovery funds effectively with other attributes as well. TORO COMPANY is an exceptional recipient created a number of jobs with very small money.

Figure 5: Brushing and highlighting technique is useful for connecting multiple visualizations. In Figure 5, the scatter plot (bottom right) shows the distribution of counties with percentage of population over 65 along the x axis. By selecting counties with a high percentage of senior citizens, all the related counties and projects were highlighted, which is called the brushing and highlighting technique [13]. Although very useful, the highlighting technique is often misleading because it always fills up the entire area of corresponding elements. Thus even when one state contains only a tiny single project related to the brushed selection, it would look as if the state has a lot of related projects.

4. TREECOVERY

We designed TreeCovery to be useful for investigative journalists and citizen watchdogs [5] who have some domain knowledge and experience in data analysis. It streamlines the exploration process available through existing visualization techniques and adds more features for data analysis. This section elaborates on the development platform, data and UI components of TreeCovery.

4.1 Prefuse Toolkit

TreeCovery is implemented as a desktop application written in the Java programming language using the Prefuse visualization toolkit [11]. Prefuse supplies a basic treemap class, as well as support for in-memory querying. TreeCovery uses a modified version of the Prefuse Treemap class. All treemap filtering is done using Prefuse Predicate objects, which filter the underlying data table.

4.2 Data Pre-processing

TreeCovery relies on a modified version of the recipient report data from recovery.gov [7]. Information is available for each recipient of any stimulus dollars, including those attributes that are necessary to build the hierarchies (department, project, state, and county) as well as attributes that are used for filtering, size-by, and color-by (number of jobs, award amount, and others), and other attributes that can be viewed in the details on demand panel (congressional district, project description, and many others). Our dataset also includes some fields added during preprocessing because of their potentially interesting values. For instance, we included money per job for each recipient, as that proportion may be more germane to the questions of stimulus effectiveness than either job creation or award amount alone. Besides the standard recipient data, TreeCovery also integrates census data fields such as population, education level, and unemployment rate. We found that many of the headlines discussing recovery data also used census information so we felt it would be very useful to integrate that information directly into the tool.

In addition to this standard processing of the data, we analyzed the data to find invalid values. We looked for three specific types of invalid values: invalid zip codes, invalid congressional districts and projects that had no primary recipient, only sub-recipients. Invalid zip codes were simply those that do not exist in the standard zip codes table; we found 253 recipients who reported non-existent zip codes. Invalid congressional districts were determined on a state-by-state basis, so that if a district did not exist in that recipient’s state, it was flagged as invalid. We found 38 errors of this kind.

7 http://www.recovery.gov/FAQ/Pages/DownloadCenter.aspx
Finally, we found 561 projects with no prime recipient reported at all. All these errors were flagged and integrated into the TreeCovery display.

### 4.3 Visualization

The chief component of the tool is, of course, the dual treemap display (Figure 6). The left treemap displays the agency (department>project) hierarchy, while the right one is spatial (state>county). Although the two share identical underlying data (the recipients) as their leaves, this level is never visible on the treemap. The shared leaf level data does, however, make it possible to synchronize the displays. Essentially, filtering occurs simultaneously (Figure 7); zooming in on one treemap, which amounts to filtering on the recipient leaves of the zoomed node, causes the other treemap to be filtered on those same recipients. Although Spotfire allows users to both zoom and filter, the two actions are independent, so that side by side treemaps will not stay synchronized automatically. In TreeCovery, users can zoom in and out on both treemaps in any arbitrary order, and the recipient leaves included in the layout will remain coordinated.

In addition to the filtering capabilities provided through zooming, TreeCovery also allows customization of the treemap displays using a set of controls. Double-sided sliders allow filtering on various attributes of the recipients and the census information for their areas. One innovative filter we added was the ability to filter on only invalid or valid date values, as found in the preprocessing stage. Each treemap can be sized and colored on a chosen attribute as well. These capabilities mirror those offered by Spotfire.
TreeCcovery’s main innovation lies in its **proportional highlighting capability**. This feature completes the synchronization of the two treemaps. Although Spotfire highlights child nodes in a treemap when the parent is selected in another one, our highlighting technique is much more finely tuned. When a node is clicked in one treemap, a highlighting square is placed inside all nodes in the other treemap that share any child recipients with the selected node. The square’s size is proportional to the ratio of the shared children’s areas to the total area of the node. For example, if the selected node has a child with an area of 10, and its parent in the other tree has an area of 100, then the highlighted square will take up 1/10 of the parent in the other tree’s area. In this way, the exact distribution of the size by attribute for the selected attribute across the other hierarchy becomes easily apparent. Using the default proportionality constant of 1, the total highlighted areas will, in fact, equal the area of the selected node. Users can change the proportionality constant for sizing the squares, in case they want to emphasize the highlighted areas. They can also control the opacity of the squares, so that the underlying labels will be visible. This highly flexible highlighting capability displays extensive information about attribute distribution across hierarchies in a powerful and intuitive manner.

Besides the main treemap functionality, TreeCcovery provides additional features for data exploration. Firstly, details of the recipients that comprise each node are available when selected. A simple table displays the data for each recipient, including values for all attribute fields in our data set. Each row of the table can be double-clicked to bring up the recovery.gov page on that project. TreeCcovery also allows users to save the current view of a treemap for later viewing. The image of a single treemap is saved, along with the current filter, zoom, size-by, and color-by settings. All saved images are shown as thumbnails in the shoebox area. Users can select and view any number of the saved images side by side in a separate window.

5. **EVALUATION**

5.1 **Insights**

To demonstrate the utility of TreeCcovery, we had a short exploratory session. Below, we present three insights that illustrate the use of TreeCcovery in data analysis.

**Washington State is one of the major recipients of money distributed by the Department of Energy (Figure 8).** TreeCcovery was run with the contract data and both treemaps were sized by the award amount. The Department of Energy received the highest amount of contract money, which was easily visible in the agency treemap. The Department of Energy node was selected, quickly highlighting the related counties in the spatial treemap. The highlighted area of each county showed the fraction of money given by the Department of Energy. Washington State received the highest portion of contract award. The highlighting feature enabled us to quickly examine the overview and motivated further exploration.

**DC, the federal hub, received the highest amount of contract money from the General Service Administration (Figure 9).** To illustrate TreeCcovery’s zooming feature, both treemaps were sized by the award amount and the General Service Administration (GSA) node was selected in the agency treemap. Various counties were highlighted in the spatial Treemap, and it was clear that DC had received the highest amount of contract money. We zoomed in on the GSA node (Figure 10) and found that one of the biggest projects of GSA was allocated to DC and all of its recipients (Prime and Sub-Prime) were in DC.

**Georgia is creating more jobs while getting less contract award money (Figure 11).** To see the utility of the shoebox in analysis, the spatial treemap was sized by award amount and a snapshot was taken. Another snapshot was taken after sizing the spatial treemap by jobs created. Both images were selected in the shoebox and opened in the comparison window. In the figure, the left treemap shows the money distribution and the right treemap shows jobs creation. Usually, states getting more money create more jobs but Georgia (GA) stood out as an outlier, with more jobs created and a comparatively small contract award amount.

![Figure 11: Comparison window showing two spatial treemaps, one sized by the award amount (left) and another sized by the jobs created (right)](image)

5.2 **Expert Evaluation**

We demonstrated our tool to the recovery.gov development team at Synteractiv. Their feedback was very positive. They felt the tool was “one of the coolest demos” they’ve seen and that TreeCcovery was the best analytic tool developed as of yet for the stimulus data. Most significantly, they were interested in incorporating a web-based version of the tool into the recovery.gov website.

6. **CONCLUSION AND FUTURE WORK**
Figure 8: Washington State (Highlighted) received the biggest portion of contract awards from the Department of Energy

Figure 9: The District of Columbia (Highlighted) received the highest contract amount from the General Service Administration Department
Figure 10: The District of Columbia has one of the biggest projects of the GSA department.

The American Recovery and Reinvestment Act provided for a substantial sum of money, $787 billion, to be distributed with the goal of economic stimulus. Tracking that distribution involves a large, multi-attribute set that can be organized as a dual hierarchy of money flow and geographical allocation. Many visualizations of the stimulus data have already been developed, but none of them adequately portray this dual hierarchy or offer flexible exploration capabilities. Our tool, TreeCover, uses synchronized treemaps to accomplish exactly that task. We use coordinated zooming and filtering and finely tuned highlighting to streamline exploration across the two hierarchies. The tool incorporates a number of other features to aid in customization and flexibility of the display. Insights that would be difficult or impossible to see with previously available tools become readily apparent when TreeCover is used to visualize the data.

While TreeCover provides some innovative features and encompasses many exploration aids, it can, of course, be greatly improved. The shoebox feature could potentially allow more extensive comparison among saved treemaps if the saved views were more interactive. The future version of TreeCover will allow entire treemaps to be saved and loaded for viewing, rather than just a screenshot. Future versions will also include support for data manipulation, including for user-defined columns, in the manner of Spotfire. This will allow greater flexibility in the way users build treemaps. We would also like to incorporate an advanced color scheme, where the color gradient follows the distribution of the data, rather than staying linear. This will allow close but not identical values to be easily differentiable in color. These extra features will enhance the TreeCover exploration experience, and hopefully lead to more insight generation.

7. ACKNOWLEDGMENTS

We would like to thank the people of Synteractive for their guidance and helpful suggestions, and Dr. Shneiderman for constant support and suggestions.

8. REFERENCES

<table>
<thead>
<tr>
<th>Population</th>
<th><a href="http://www.recovery.gov/About/Pages/The_Act.aspx">http://www.recovery.gov/About/Pages/The_Act.aspx</a></th>
</tr>
</thead>
<tbody>
<tr>
<td>Female Percentage</td>
<td></td>
</tr>
<tr>
<td>White people percentage</td>
<td></td>
</tr>
<tr>
<td>Black people percentage</td>
<td></td>
</tr>
<tr>
<td>Hispanic percentage</td>
<td></td>
</tr>
<tr>
<td>Infant deaths</td>
<td></td>
</tr>
<tr>
<td>High school graduate percentage</td>
<td></td>
</tr>
<tr>
<td>Bachelor degree percentage</td>
<td></td>
</tr>
<tr>
<td>Housing unit</td>
<td></td>
</tr>
<tr>
<td>Housing unit percent change</td>
<td></td>
</tr>
<tr>
<td>Median household income</td>
<td></td>
</tr>
<tr>
<td>People in poverty</td>
<td></td>
</tr>
<tr>
<td>Labor force</td>
<td></td>
</tr>
<tr>
<td>Unemployment rate</td>
<td></td>
</tr>
<tr>
<td>Number of firms</td>
<td></td>
</tr>
<tr>
<td>Women-owned firms percentage</td>
<td></td>
</tr>
</tbody>
</table>

**APPENDIX**

A. CENSUS DATA
downloaded from http://quickfacts.census.gov/qfd/download_data.html

<table>
<thead>
<tr>
<th>Population</th>
<th><a href="http://www.recovery.gov/About/Pages/The_Act.aspx">http://www.recovery.gov/About/Pages/The_Act.aspx</a></th>
</tr>
</thead>
<tbody>
<tr>
<td>Female Percentage</td>
<td></td>
</tr>
<tr>
<td>White people percentage</td>
<td></td>
</tr>
<tr>
<td>Black people percentage</td>
<td></td>
</tr>
<tr>
<td>Hispanic percentage</td>
<td></td>
</tr>
<tr>
<td>Infant deaths</td>
<td></td>
</tr>
<tr>
<td>High school graduate percentage</td>
<td></td>
</tr>
<tr>
<td>Bachelor degree percentage</td>
<td></td>
</tr>
<tr>
<td>Housing unit</td>
<td></td>
</tr>
<tr>
<td>Housing unit percent change</td>
<td></td>
</tr>
<tr>
<td>Median household income</td>
<td></td>
</tr>
<tr>
<td>People in poverty</td>
<td></td>
</tr>
<tr>
<td>Labor force</td>
<td></td>
</tr>
<tr>
<td>Unemployment rate</td>
<td></td>
</tr>
<tr>
<td>Number of firms</td>
<td></td>
</tr>
<tr>
<td>Women-owned firms percentage</td>
<td></td>
</tr>
</tbody>
</table>