

# CatTrees: Dynamic Visualization of Categorical Data Using Treemaps

CMSC838b

Erica Kolatch and Beth Weinstein

kolatch@cs.umd.edu

bweinste@wam.umd.edu

May 2001

## **Abstract**

Social scientists collect a significant amount of categorical data. Current visualizations of categorical data are primarily static. They do not accommodate the concept of dynamic queries or implement the essentials of interactive visualizations. We have designed a tool, an enhancement of Treemaps called CatTrees, which takes categorical data, creates a hierarchy from it, and allows direct, interactive manipulation of that hierarchy.

## **Introduction**

Social scientists collect a significant amount of data. Vast portions of this data are answers to qualitative or categorical questions. These questions, framed, for example, in terms of yes/no, male/female, or 1/2/3/4 need to be analyzed and interpreted statistically in order to discover and validate new and interesting relationships. Current visualizations tools are primarily static or provide limited tools for data comparison. They do not fully accommodate the concept of dynamic queries or implement the essentials of interactive visualizations. Our goal is to create a tool that will allow the dynamic, interactive exploration of data sets consisting of primarily categorical data.

Categorical data, sometimes referred to as qualitative data or nominal data (data that can be named), is data that can be separated into different categories distinguished by some non-numeric characteristic. The collection of categorical data involves the counting of occurrences that can be named and enumerated, and it is analyzed using a number of statistical methods, including contingency tables, regression models, conditional inference [Llo99], and correspondence analysis [Hof86, Wat97]. Methods may be both *confirmatory* and *exploratory*. Confirmatory methods have a hypothesis as their basis, such as “income is dependent on race”. Exploratory methods often do not generate strong conclusions, such as “in general blacks earn less than whites”, but frequently social scientists are using exploratory methods to describe the structure of the data rather than model the relationships [BG98]. Visualizations have been used to display the results of both types of methods and as an enhancement of the modeling process. The goal of visualizations has been to discover the structure of the data before modeling begins using interactive visualization tools, provide methods for viewing results often using static visualization tools, and identify further information about the modeled data.

The main difference between static visualizations and interactive visualizations based on dynamic queries are that dynamic visualizations provide immediate feedback and allow

queries that are incremental and reversible [\[Shn94\]](#). Current static visualizations require reprocessing for each query change and are often hard-coded within the statistical package. They are often dependent on the method of statistical analysis and frequently require an in-depth understanding of the domain in order to understand the result.

In our work, we attempt to build on current tools in the interactive information visualization field in order to create a new tool that will display an intuitive visualization of categorical data and that will allow dynamic manipulation of the data. This tool will allow exploration of the data without extensive mathematical calculation or intensive domain knowledge.

The remainder of the paper is organized as follows. We will first look at current related work in both categorical visualization and interactive information visualization. Then, we will describe our implementation, its main features, and its effectiveness for the task. Next, we will outline further enhancements and suggestions for changes to the tool, and finally we will draw some conclusions.

---

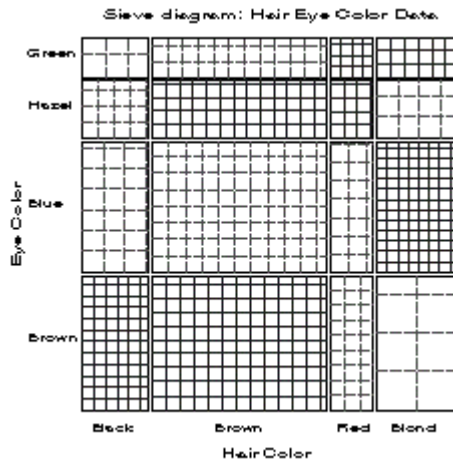
## Related Work

Two parallel paths have developed in the visualization of categorical data. The first follows the statistical visualization path, and the second follows the dynamic information visualization path. The goal of our tool is to join statistical visualization with information visualization.

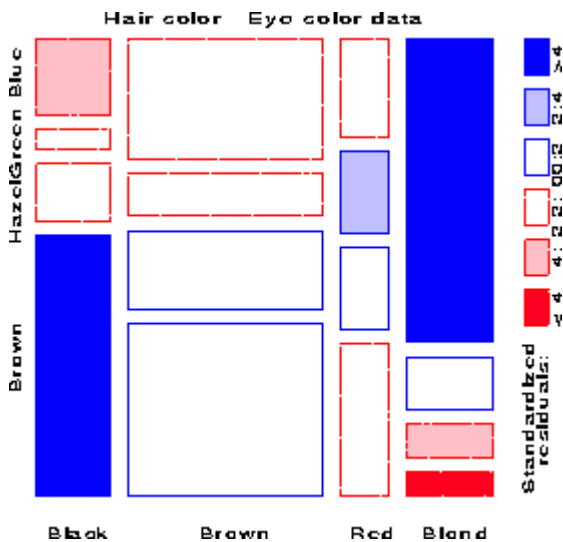
Statistical visualization started with basic graphs and charts. An evolution of simple graphs includes the graphing of correspondence analysis which uses a distance metric to represent relationship between data points [\[Gre93\]](#). Recently, researchers have been working on developing statistical visualizations that focus on categorical data using different forms of mosaics or collections of rectangular tiles.

Information visualization started with simple graphs and evolved to include many tools which are designed to effectively visualize information. With respect to data visualization, these tools are successful at two types of tasks. The first is allowing the user to drill down to discover concrete pieces of information, for example using interactive zooming [\[Bed96\]](#) or focus+context [\[RC94\]](#) techniques. The second is finding interesting patterns in the data using maps, graphs, or plots.

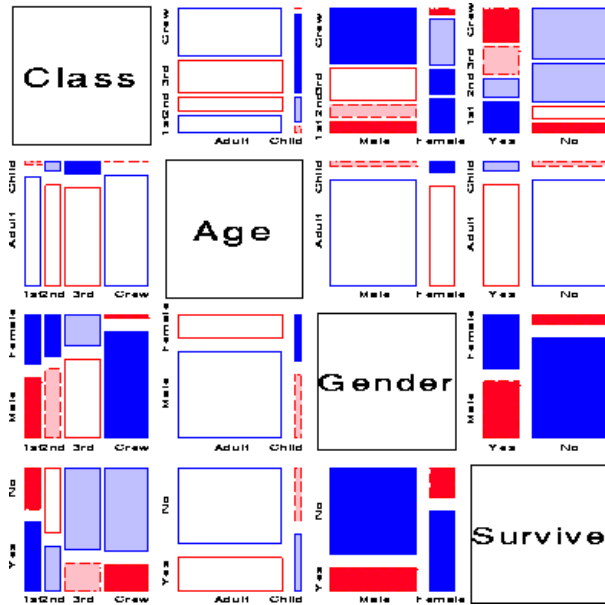
The literature on the statistical analysis of categorical data recognizes the lack of successful visualizations for qualitative data. Several methods have been suggested to remedy the situation.



Sieve or parquet diagrams are used to plot two-way contingency tables. They are based on the premise that the area of each rectangle is proportional to the expected frequency of an element in the table, and observed frequency is reflected in the number of grid squares within each larger rectangle [RS94, Fri00].



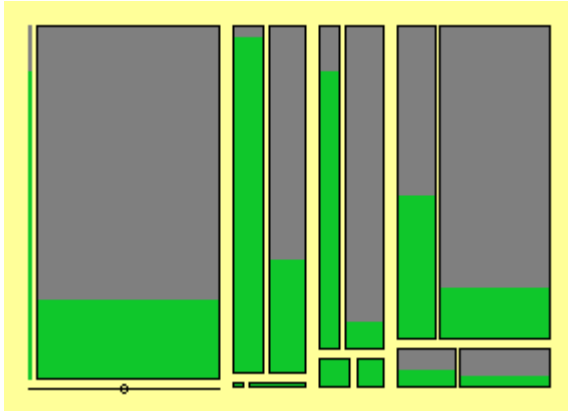
Mosaic displays [Fri92a, Fri92b, Fri00] are a graphical method for visualizing n-way contingency tables. They are similar in appearance to parquet diagrams but have one major difference. The areas of the rectangular tiles are based on the observed frequency of elements in the contingency table rather than the expected frequency. Areas may also be colored and shaded depending on the statistical model used. In this example, the coloration is based on the standardized residual from independence, with positive values in blue with solid borders, and negative values in red with broken borders. Mosaic displays may also be used to determine if a statistical model fits the data, and can be used to suggest an alternative model [Fri00].



A collection of related mosaics, or a mosaic matrix, can be used to show all pair-wise relationships of a set of elements in a multi-way contingency table of categorical variables. The concept of matrices can be extended to display additional relationships among the data including marginal or conditional relationships [\[Fri99\]](#). In theory, a mosaic matrix could accommodate a large number of variables, but it becomes difficult to visualize when greater than four or five are selected [\[Fri99\]](#).

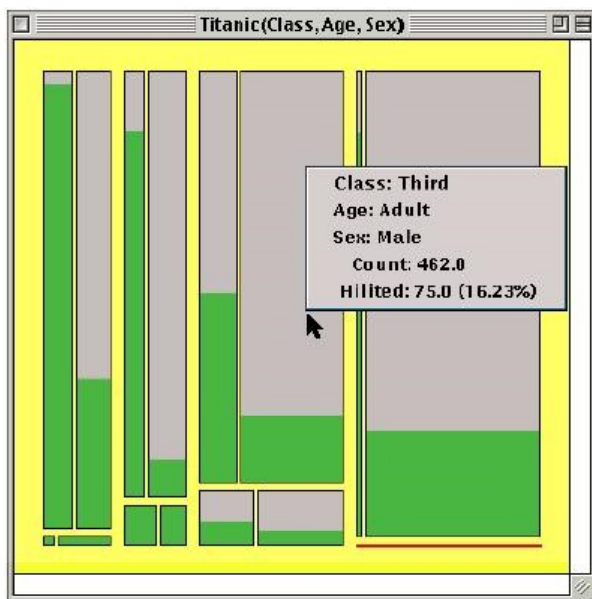
(The example shown displays information on the survival rates for passengers and crew on the Titanic.)

Mosaic displays, as portrayed above, are static visualizations, created with data that has already been manipulated and analyzed. They are frequently products of large statistical packages (eg SAS) and require hard coding of the data as part of the construction process. Some research has been done on more interactive versions of mosaic plots. Interactive mosaic plots would allow the user to switch between different low-dimensional views of multivariate data.



MANET [[UHHS96](#), [THSU97](#)] provides facilities for interactive statistical techniques including mosaic plots in a package designed for the Macintosh OS. MANET can handle up to 256 combinations (number of categories \* number of variables), and can show bar charts, histograms, trellis displays, and scatterplots, as well as, mosaic plots from both raw data and the values for certain loglinear models. Aside from the limits to input size, MANET's output is less effective because rectangles are not labeled.

(This mosaic plot represents the same data as the mosaic matrix above, but requires explanations before it can be interpreted.)



MONDRIAN [[The98](#)], an enhanced version of MANET, is a package designed to provide interactive statistical techniques written in JAVA. MONDRIAN can show bar charts, histograms, maps, mosaic plots, and parallel coordinates. As in MANET, the mosaic plots can display raw data or modeled data and can be manipulated dynamically. In addition, MONDRIAN has facilities for selecting sub-groups of data and zooming is available in plots with data coordinates. Although some query information is available in the title bar, and the

mouse can be used to interpret details, labeling is still a significant issue in this implementation.

(Again, this plot shows an interpretation of the Titanic dataset.)

MANET and MONDRIAN can show the relationships between variables in a dataset resulting from a particular query such as "What was the survival rate of passengers on the Titanic with relation to passage class, age, and sex?" In effect, these queries are creating a hierarchy among the variables, performing a count based on the ordering in the hierarchy, and answering the question, "How many of each combination of responses is present in the data set?" The ability to dynamically examine and manipulate hierarchical data is an essential feature of two other visualization tools, Table Lens [RC94, PR96] and Treemaps [JS91, Shn00].

Olympic Diving Medal Results

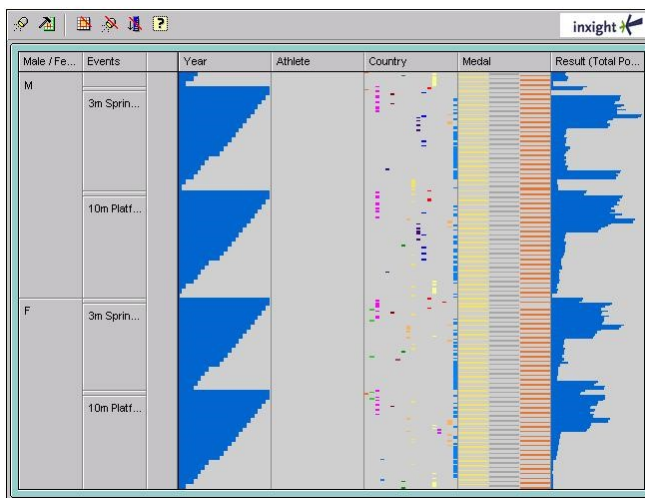
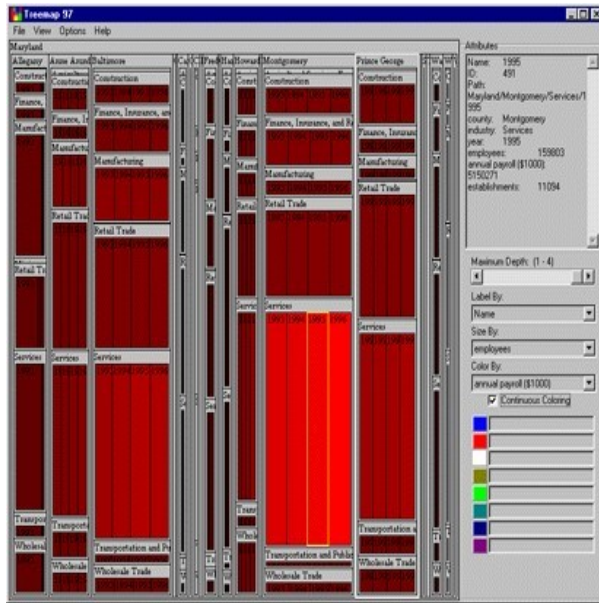


Table Lens, developed originally by Xerox Park, Palo Alto, CA, uses a focus+context technique to display tabular information. Table Lens supports interaction with very large tables. It provides an overview of the sorted data, and also allows the user to isolate individual records or group records using row focusing techniques. Although not a hierarchical tool, it allows the user to sort and filter the data based on values of individual columns. A user may also isolate a single variable or group of variables and sort successively on those variables thus creating a virtual hierarchy. In this example the data was first sorted by gender and then by diving event. In effect, gender is at the top of the hierarchy, or tree, event is at the next level, and all the other variables, including year, athlete, country, medal and result are in the leaf nodes. A commercial version of Table Lens is available from Inxight software. Called Eureka, demos of the product can be viewed at [http://www.inxight.com/products\\_eu/eureka/index.html](http://www.inxight.com/products_eu/eureka/index.html).



Treemaps was developed by Ben Shneiderman at the Human-Computer Interaction Laboratory (HCIL) of the University of Maryland during the 1990s. Treemaps is a tool for visualizing hierarchical structures in a minimum amount of space. It allows rapid comparison of the size of nodes or the shape of sub-trees and creates a display of leaf node values based on size and color [JS91]. Dynamic queries, added in recent implementations of Treemaps, allow rapid and reversible selection of attribute values which create shrinking sub-tree structures and encourage data exploration. For current work on Treemaps under the auspices of HCIL go to: <http://www.cs.umd.edu/hcil/treemaps/treemap2001/>.

## Implementation

### Motivation

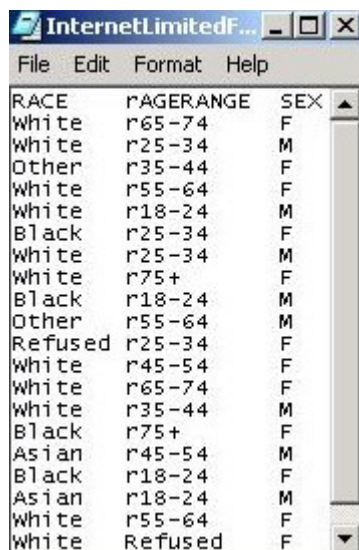
Although Treemaps allows the manipulation of variable ranges and hierarchical zooming, it does not permit the manipulation of the hierarchy itself. An interest in manipulating hierarchies to explore underlying relationships in data is a basic motivation for our work. Our speculations centered on the concept that queries on categorical data focus on relationships between groups and frequently are posed in terms of size, for example “which is larger” or “which has more”. These questions can be answered in a simple data set by counting items. However, in a larger data set answering these questions becomes time-consuming. In multi-dimensional data sets, data with many variables for each data point, questions can become extremely complex, for example, “Do white males in the North East use the Internet more than white males in the South?” This question requires examining data points to find all subjects who answered “yes” to white, male, use the Internet, North East, and South, and then counting each group. If the data were arranged in a tree, or

hierarchy, then each possible pattern of answers would have a leaf node with a count in it, and we could simply follow the correct paths to two leaf nodes and compare the counts. Basic Treemaps could be used to arrange the data in a hierarchy. The question that arises is how to determine the hierarchical order of the data. If the hierarchy selected for the data did not provide the answer to our query then we would have to reformat the data and start over again. An ability to manipulate the hierarchy would allow us to answer specific queries, test hypotheses about the data, experiment with selecting variables and arrangements, and promote a better understanding of the entire data set.

## Methodology

Starting with the basic structure of Treemaps, we adapted it to allow dynamic manipulation of the hierarchical structure of the tree, creating CatTrees. Since our data was not hierarchical in nature, we assumed that the current order of the data reflected the initial hierarchy. Once the initial hierarchy is created, CatTrees allows the user to dynamically change the order of the hierarchy using drag-and-drop, thus changing the relationships in the hierarchy, and control which columns are included in the tree, thus controlling the depth of the hierarchy.

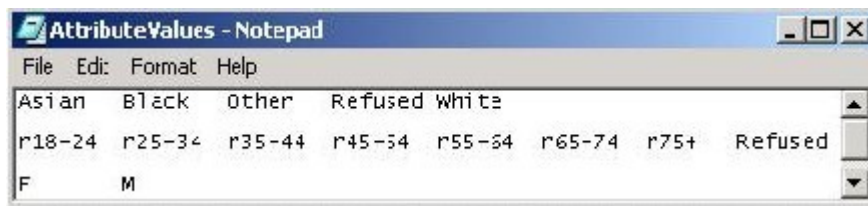
## Design Decisions



RACE	RAGERANGE	SEX
White	r65-74	F
White	r25-34	M
Other	r35-44	F
White	r55-64	F
White	r18-24	M
Black	r25-34	F
White	r25-34	M
White	r75+	F
Black	r18-24	M
Other	r55-64	M
Refused	r25-34	F
White	r45-54	F
White	r65-74	F
White	r35-44	M
Black	r75+	F
Asian	r45-54	M
Black	r18-24	F
Asian	r18-24	M
White	r55-64	F
White	Refused	F

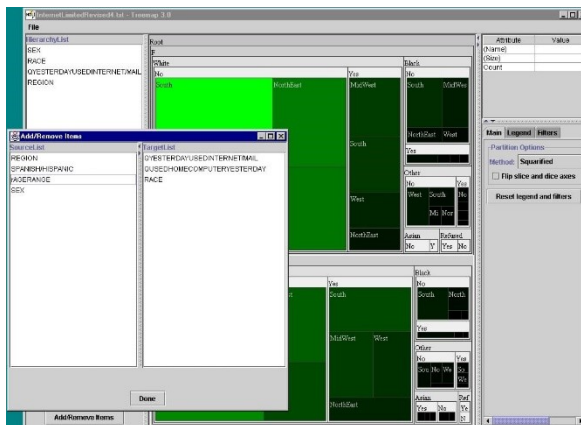
Our first concern was how to present the data to CatTrees. We altered Treemaps so that it now recognizes tab-delimited files which can be created from a spreadsheet or database. Tab-delimited files similar to the abbreviated table shown at left have one row for each survey participant. However, CatTrees also requires an additional file identifying the initial tree structure. Each row in the file represents the possible values for a particular variable. An example of this type of file is shown below.





AttributeValues - Notepad							
File Edit Format Help							
Asian	Black	Other	Refused	White			
r18-24	r25-34	r35-44	r45-54	r55-64	r65-74	r75+	Refused
F	M						

From the tab delimited presentation, data is transported to a sorted two-dimensional array with one more column than is present in the original data. This column is used to hold master counts for the data. These counts, calculated as the data are read in, are associated with each row of data. Even though the order of data is changed, the count for the row will still be the same. If only some of the columns are selected for the hierarchy, the counts for rows with similar patterns are added to create a cumulative total for the pattern. In this way, counting is done only once for each data importation. In addition, data is never moved in the array once it is inserted, changed hierarchies use a mapping from the array structure to the desired column order or column selection. Although counting is never repeated, the CatTree must be redrawn each time the hierarchy changes.

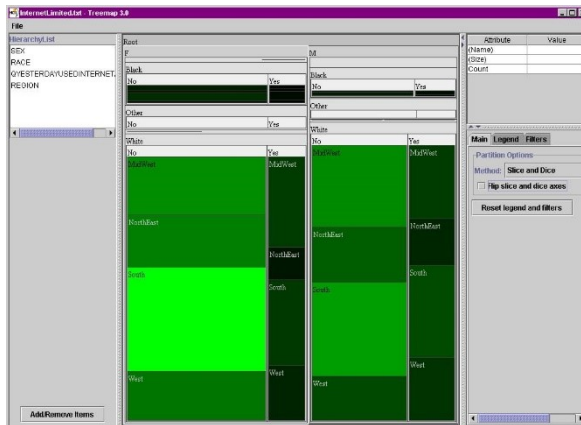


We also had to determine how to modify the Treemaps interface to accommodate hierarchy manipulation and selection. We decided on two separate elements. The first, present at all times in the window, is a box labeled "Hierarchy Order", showing the current order of variables. Within this box, users can manipulate the variable names using drag-and-drop in order to manipulate the Treemaps hierarchy. The second element, activated by a button labeled "Add/Remove Items" in the bottom left hand corner, is a pop-up box. This pop-up box has two sides. When first opened, all variable names are on the left, the *source* side. Some or all of the variable names can be dragged to the *target* side thus allowing the user to change the number of variables included in the visualization without changing the original dataset.

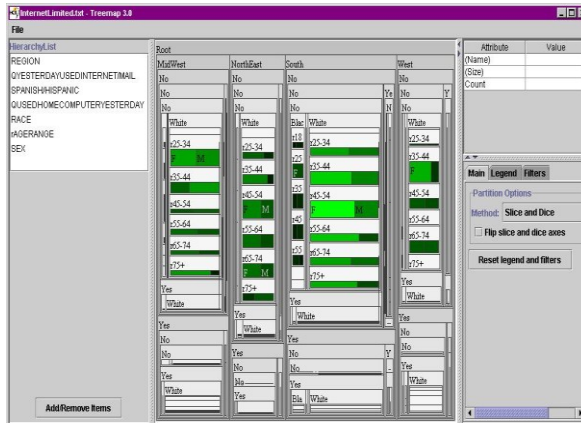
(Click image to enlarge.)

## Results

CatTrees successfully allows the user to manipulate categorical data arranged in a hierarchy. It provides the capability of examining as many or as few variables at a time as desired, and varying the depth those variables take in the hierarchy. In addition, CatTrees is enhanced by the basic elements of Treemaps which allows the user to zoom in or out of the hierarchy and determine the actual count of a variable in the total count of the unit with a simple mouse click.



The [demonstration data](#) provided at the end of the paper operates on a subset of the University of Maryland Internet Usage Survey Data, 1999. The full dataset, including the codebook is available for download from the [WebUse](#) site at the University of Maryland. Users can quickly compare data groupings and the implementation allows users to visually explore both explicit queries as well as discovering unanticipated results in the data which may open new avenues for consideration. The question previously stated asking for a comparison of usage by white males in the North East and white males in the South can easily be answered by the visualization shown at left. (*Click on the image to enlarge it.*) In addition, the visualization can be used to explore the dataset. Sometimes the results can quickly confirm previous suspicions, other times, results may be more unexpected. For example, from this hierarchy we can confirm that a digital divide still exists between men and women. Men use the Internet more than women. However, we can also see that Internet usage among women is proportionally higher in the MidWest than in the South even though more women from the South responded to the survey, but Internet usage among men in the South and MidWest appears to be proportional to the number of respondents.



There were some problems with the implemented version. We identified some memory issues related to Treemaps 3.0. Treemaps worked best with categorical data variables containing a limited number of values. When presented with multiple data points containing large numbers of values, for example a list of State names (52 values), or a group of ranges containing 18 values, Treemaps 3.0 could not build the initial tree. the size of the tree grows exponentially. A hierarchy with six variables (six levels) and two values for each variable will have 64 leaf nodes. A hierarchy with six variables and five values for each variable will have 15,625 leaf nodes. Even larger numbers of values will create even larger numbers of leaf nodes, and this can and did cause program failure in some test cases. In addition, there is a visual problem when the initial CatTree is created. Visually, a large tree in CatTrees may cause confusion because there is no apparent order to the boxes. *(Click on side image to enlarge it)*. This question of order dissipates quickly as the hierarchy is manipulated.

The implementation provided a number of satisfactory results. First, it has the ability to create a hierarchy in non-hierarchical data, and second, it provides the user with the ability to manipulate that hierarchy in several different ways. A brief pilot study was completed by four subjects comparing CatTrees to Microsoft Excel™. Users were asked to discover the answers to three different questions.

1. "Which region had the most participants?"
2. "Who used the Internet or Email more yesterday, white males or white females?" -> **I think a better question would have been "Did more white males or more white females use the Internet/Email yesterday?" (Their framing makes it sound like you're supposed to work it out based on overall time used rather than discrete number of people. Also, the 'Internet or Email' adds ambiguity) But this is interesting to think about - if internet or email were different variables, we wouldn't want to count the same person twice.**
3. "How many Asians responded to the survey?"

After a short introduction to CatTrees and Treemaps users were able to complete tasks quickly and successfully. User satisfaction was also much higher with Treemaps than with Excel. Users found it both easy and intuitive to manipulate the hierarchy. CatTrees presents

several different methods for arriving at the answer to each question. Users could either eyeball the size of each node or left-click the node to get the count. They could also drill down in the hierarchy or add and remove items from the tree to bring nodes into view. Different users preferred different methods, and perhaps with more experience a different method might be chosen in the end, but all methods provided quick and accurate responses. Tasks were complicated in Excel because the tool required more mouse clicks, addition and subtraction, and actual counting. The speed of success with CatTrees increased user satisfaction, and Excel caused frustration, because it required significant manipulation, and finding the correct method to manipulate the data was not intuitive.

---

## **Future Work**

A number of suggestions can be made for future improvements for CatTrees.

- In order to combat the size limitations of the tree mentioned in the Results section, the implementation should allow the user to pick specific columns from the dataset before the tree is built, eliminating the need to change the original data file at any time.
- Additional dynamic query elements should be added to the interface. Slider bars or choice boxes should be added for each attribute to provide a method of limiting the variables in the hierarchy. These could be used to filter out values, for example limiting the regions of the country, or provide a range of values, for example only look at users between the ages of 35 and 64. Once the filtering is complete additional attributes could be dynamically included in the hierarchy.
- A user should be able to click through a series that includes all permutations of a set of chosen variables in the Hierarchy List. This would allow the user to quickly compare a set of hierarchies.
- CatTrees should be enhanced to include facilities for calculating additional statistics aside from a basic count.
- More formal usability studies are needed to document CatTrees' potential as a visualization tool, and its value to the statistical and social science communities.

---

## **Conclusion**

Building on existing work in categorical visualization, including static visualizations in the form of mosaic plots and matrixes, and dynamic visualizations, in the form of MANET and MONDRIAN, and work with hierarchical dynamic visualizations, in particular Treemaps, we created a tool, CatTrees that facilitates the dynamic visualization of categorical data. CatTrees takes flat data files, creates a hierarchy of the variables in the datafile, and facilitates the exploration of that data through dynamic hierarchical manipulation. Initial

studies showed that CatTrees could be used to successfully explore categorical datasets and discover interesting patterns and relations in the data. In fact, CatTrees permits flexible manipulation of all forms of hierarchical data, and provides facilities for both exploration and dynamic queries. Future enhancements to CatTrees will enhance the facilities for dynamic queries, and increase the potential usability for both users familiar with a particular dataset as well as users who have never seen the particular dataset before.

---

## References

- [Bed96] Bederson, Benjamin, et al. (1996). "Pad++: A Zoomable Graphical Sketchpad for Exploring Alternate Interface Physics." *Journal of Visual Languages and Computing*. 7, 3-31.
- [BG98] Blasius, Jorg and Michael Greenacre, editors. (1998). *Visualization of categorical data*. Academic Press.
- [Fri92a] Friendly, Michael. (1992). "Graphical methods for Categorical Data." *SAS Users Group International 17th Annual Conference*.
- [Fri92b] Friendly, Michael. (1999). "Visualizing Categorical Data." In Sirken, Monroe G. et al. (eds.) *Cognition and Survey Research*. New York: John Wiley & Sons.
- [Fri99] Friendly, Michael. (1999). "Extending Mosaic Displays: Marginal, Partial, and Conditional Views of Categorical Data." *Journal of Computational Statistics and Graphics*, 8, 373-395.
- [Fri00] Friendly, Michael. (2000). "Visualizing Categorical Data: Data, Stories, and Pictures." *SAS Users Group International, 25th Annual Conference*.  
<http://www.math.yorku.ca/SCS/vcd/vcdstory.pdf>
- [Gre93] Greenacre, M. J. (1993). *Correspondence Analysis in Practice*. London: Academic Press.
- [Hof86] Hoffman, Donna L. (1986). "Correspondence Analysis: The Graphical Representation of Categorical Data in Marketing Research." *Journal of Marketing Research*, 213-227.
- [Hof00] Hoffman, Heike. (2000). "Exploring Categorical Data: Interactive Mosaic Plots." *Metrika*, 51(1), 11-26.

- [JS91] Johnson, Brian, and Ben Shneiderman (1991). "Treemaps: A Space-Filling Approach to the Visualization of Hierarchical Information Structures." *Proceedings of IEEE Information Visualization '91*, 275-282.
- [Lau97] Lauer, Stephen. (1997). "Interactive Modelling of Categorical Data." <http://www1.math.uni-augsburg.de/~lauer/IMCD.html>
- [Llo99] Lloyd, Chris. (1999). *Statistical Analysis of Categorical Data*. New York: John Wiley & Sons.
- [RC94] Rao, Ramana, and Stuart Card. (1994). "The Table Lens: Merging Graphical and Symbolic Representations in an Interactive Focus + Context Visualization for Tabular Information." *Proceedings CHI'94*, 318-322.
- [PR96] Pirolli, P and Ramana Rao. (1996). "Table Lens as a Tool for Making Sense of Data." In Catarcum T., et al. eds. *Workshop on Advanced Visual Interfaces: AVI-96*, 67-80.
- [RS94] Riedwyl, Hans and H. Schupback. (1994). "Parquet Diagram to Plot Contingency Tables" *IEEE Softstat '93: Advances in Statistical Software*. 293-299.
- [Shn94] Shneiderman, Ben. (1994). "Dynamic Queries for Visual Information Seeking." *IEEE Software*. 11(6), 70-77.
- [Shn00] Shneiderman, Ben. (1998, 2000). "Treemaps for Space-Constrained Visualization of Hierarchies." <http://www.cs.umd.edu/hcil/treemaps/>
- [THSU97] Theus, Martin, Heike Hofmann, Bernd Siegl, and Antony Unwin. (1997). "MANET: Extensions to Interactive Statistical Graphics for Missing Values." In *New Techniques and Technologies for Statistics II*. Amsterdam: IOS-Press, 247-259.
- [The98] Theus, Martin. (1998). "MONDRIAN -- Interactive Graphical Data Analysis." *Graphics Workshop, Drew University*. Madison, NJ. <http://www.planet-interkom.de/martin.theus/Mondrian.pdf>
- [UHHS96] Unwin, Antony, George Hawkins, Heike Hofmann, and Bernd Siegl. (1996). "Interactive Graphics for Data Sets with Missing Values – MANET." *Journal of Computational and Graphical Statistics*. 5 (2), 113 – 122.  
<http://www1.math.uni-augsburg.de/Manet/>
- [Wat97] Watts, Dorraine Day. (1997). "Correspondence analysis: a graphical technique for examining categorical data." *Nursing Research*. 46 (4), 235-9.
-

## Demo

Please download these data files first....

[Data](#) -- right-click and save as "InternetLimited.txt"

[Values for Attributes](#) -- right-click and save as "AttributeValues.txt"

Internet Explorer is required to open the demo. [Demo](#)

1. If given a choice, please select the option "Open this file from its current location."
  2. When TreeMaps opens, click "File" and then click "Open Categorical."
  3. A pop-up window, "Open a Data File" will appear.
  4. Find the two files you downloaded previously.
  5. Select the Data file (saved as "InternetLimited.txt") first, and click "Open".
  6. Select the Values for Attributes file (saved as "AttributeValues.txt") second, and click "Open".
  7. The full version of CatTrees should open in a few seconds.
- 

## Credits

We would like to thank Dr. Ben Shneiderman and Dr. Catherine Plaisant for their help with this project.

We would also like to thank Dr. Alan Neustadt, Sociology Department, University of Maryland, College Park for his assistance.