ABSTRACT

Title of dissertation	EXPLORING DIFFERENCES IN MULTIVARIATE DATASETS USING HIERARCHIES AN INTERACTIVE INFORMATION VISUALIZATION APPROACH
	John Alexis Guerra Gómez, Doctor of Philosophy, 2013
Directed by	Professor Ben Shneiderman Department of Computer Science

Hierarchies are a useful way of representing data. The parent-child relationships they define facilitate the analysis of a dataset by breaking it down into its component parts. Representing data as hierarchies can also be used to track changes to a dataset over time or between versions. For example, analysts can use hierarchies to uncover changes in the US Federal Budget in the last twenty years, by grouping accounts by Agencies and Bureaus. Similarly, a company manager can analyze changes to their product sales due to the holiday season by breaking them up by markets and product categories. Exploring differences in such trees could help them understand changes in the data. However, comparing hierarchies is a difficult task, even when comparing two trees with a small number of nodes. To address this, information visualization techniques were used to support data comparison tasks using hierarchies. After evaluating my techniques with domain experts on real world problems, I identified and addressed two main research topics:

This dissertation first tackled the problem of comparing two versions of a tree by using two types of change, while most of the significant work on this topic has focused only on *changes in node values* or *changes in topology*. *TreeVersity (*`http://hcil.cs.umd.edu/treeversity`) is a comparison tool that allows users to explore changes between two versions of a tree by tracking *node value differences, and newly created or removed nodes*. Domain experts using TreeVersity were excited to discover differences

in the trees, but expressed a desire to explore the evolution of a dataset over time. To that end, they suggested applying TreeVersity comparison capabilities to datasets that were non inherently hierarchical.

Following users' feedback, the problem of exploring changes over time in datasets that can be categorized as trees was addressed next. *TreeVersity2* (`http://treeversity.cattlab.umd.edu` is a web-based data comparison tool that allows users to explore a tree that changes over time and of datasets that are not inherently hierarchical, by categorizing them by their attributes. TreeVersity2 also helps users navigate the sometimes large amounts of differences between versions of a tree using an interactive textual reporting tool.

My research has resulted in three main contributions: First, the introduction of the *Bullet,* a visualization glyph to represent four characteristics of change (as described in Section 1.2) in tree nodes, and the implementation of the Bullet in *TreeVersity*. Second, the creation of the *StemView,* a tree visualization technique that represents five characteristics of change in all the nodes of a tree (not just the leaves), and the implementation of the StemView in *TreeVersity2*. Furthermore, my research resulted in the development of the *reporting tool,* another feature of TreeVersity2, which helps users navigate outstanding changes in the tree with textual representations and coordinated interactions. Third, the development of 13 case studies with domain experts on real world comparison problems. The case studies have validated the utility and flexibility of my approaches. Finally, my research opens possibilities for future research on comparing hierarchical structures.

EXPLORING DIFFERENCES IN MULTIVARIATE DATASETS USING HIERARCHIES
AN INTERACTIVE INFORMATION VISUALIZATION APPROACH

by

John Alexis Guerra Gómez

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park, in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2013

Advisory Committee:

Professor Ben Shneiderman, Chair
Dr. Catherine Plaisant
Professor Lise Getoor
Professor Ben Bederson
Professor David Lovell

To my life-partner/co-founder/friend/complement/boss/wife
Mafe

Acknowledgments

For many, obtaining a PhD is a big achievement. For others, it is just the next step in their careers. For me, getting this degree has been the dream of my life and, as in any dream, this has been a wonderful journey where I have been blessed to receive the support and help of many loving people.

I would like to first thank my dear wife Mafe. Her support, patience, and trust have been fundamental in achieving this goal. Thanks to Mafe I came here, thanks to her I persisted during the toughest of times, and thanks to her I finished. I am who I am thanks to her. From co-founding and directing our company to encouraging me to apply for the almost "impossible-to-get" Fulbright Science and Technology Scholarship and to putting on hold her dreams so I could achieve mine, Mafe has been my perfect other half, my partner, my life. Moreover, thanks for loving me the way you do everything you do, with your whole heart, and thank you for blessing me with a wife I can respect, admire, and love.

I would like to thank my advisor Ben Shneiderman, whose tireless enthusiasm and passion for his work have been truly inspirational to me. He introduced me to the world of information visualization and guided me through the development of my work. His teachings, which many times came from unexpected situations, have forged me into a researcher and made me a better professional and a better person. Every meeting with him was an opportunity to learn something new, and his teachings extended beyond research principles to life guidelines. Dear Ben, thank you very much for all the time you invested in me, for challenging me to defend my ideas, for sharing your energy with me, and for giving me a wonderful role model of what I want to be as a researcher.

Also, I would like to thank my co-advisor Catherine Plaisant, whose endless joy, energy and enthusiasm deeply influenced me. Catherine was always there to offer me honest and kind advice when I needed it. Thanks to her, our work meetings were always joyful and fruitful. Her vision, experience, and creativity enriched my work immensely. Despite her vast knowledge, she always presented herself as a reachable person who always made an extra effort to understand my ideas and concerns. Catherine and Ben make the perfect research duo, offering different points of view but with the same end goals. Dear Catherine, thank you very much for your infallible support, for your thoughtful advice, for listening and relating to my ideas, and for giving so much of yourself to make me a better researcher.

I express my gratitude also to Audra Buck-Coleman for her thoughtful advice in design. Audra's designs enrichedembellished my research, and I'm sure her influence will permeate my future work. My gratitude also goes to Michel L. Pack for supporting my journey, giving me the freedom to pursue my ideas and a rich working environment in the CATT Lab to develop them. Furthermore, I would like to thank my co-workers at the HCIL for being such a diverse and supportive group of people who were always ready to help. In the same way, I thank the Fulbright Science and Technology Scholarship for helping me achieve my dream by funding the first three years of my PhD. Special mention to Vincent Picket, Sarah Boeving, Anna Rendon, Lori Reynolds, Catalina Ahumada and Ann Mason for all your help solving my day to day problems. On that same note, I thank Colfuturo and Colciencias for funding the last part of my studies. Furthermore, I express my gratitude to my dissertation committee members for their thoughtful advice and constructive comments on my thesis.

As in any important accomplishment in life, this journey was full of challenges and opportunities. During the whole process the support of my friends was fundamental for overcoming the problems and maintaining focus on the big picture of my goal. Because of this I want to thank the many friends that welcomed me into their lives. Dear Caro, thank you very much for offering me a friendship that I wouldn't have thought was even possible, for welcoming me into your life and for sharing a piece of your CAOS with me. Juan, thanks for sharing your philosophy of life with me, and for always being there for me, ready to share your carácter. In the same way, to Andrea & Iván for offering me their selfless friendship, and for being always ready to offer a thoughtful advice. To Fernando & Elena, Camilo & Cindy, Andrew, and Dan, for adopting this lost Colombian and offering me a daily space for fruitful debate, relaxation and lunch. To Dina María for always being there for me despite the passage of time and for offering me your true friendship. To Robin for being mi lanza, and for teaching me with your humbleness and capacity to give. To Felipe, Jose and Walter for being our partners and all your help building DUTO. To Anderson for letting me learn from your courage and perseverance.

To Clara & Oscar for treating me like another son. To Jason & Naomi, and Kevin & Melanie for sharing their country, their traditions and their culture with me. To Jaime and Gilberto, for offering me your thoughtful advice. To Cathe, Maryluz, Gladys and Mayi for always being there for me. To Dica, Jhony, Leandro, Marthica, Nana, Ferney, Jorge, Raúl, Nestor, Pepe, Krist, Shopan, David, Tak, Megan, Suri, Cody, Alex, Chang, Catalin, Marce, Garcés, Aldana, Martha Escobar, Priscilla, Flor, José Guerra, and to the many others that I didn't list, for blessing me with your friendship.

Finally but not less importantly, I want to thank all my family. You planted the seeds that made me who I am now. Dear Mom and Dad, your unconditional support and endless love have always been extremely motivational for me, and I will always feel blessed to have you as my parents. To Edwin, Liza and Magda, I have always been honored to follow your footsteps and to try to fill the immense footprints that you have made with your life achievements. To Santi, Yuli, Johan, Edwar, Jorge, and to my many relatives, both close and far, who always gave me words of support and constantly prayed for me during this journey.

Thank you all, this achievement is also yours, and may your dreams come true as well.

Contents

List of Figures

Chapter 1

Introduction

Trees are one of the most common data structures in Computer Science. The hierarchies they represent help users to organize and categorize data. Many techniques have effectively addressed the problem of exploring, storing and visualizing trees [8, 17, 29, 42, 43, 51, 70, 74]. However the problem of comparing two or more versions of a tree is more complicated even for small trees (with just dozens of nodes) and still leaves space for improvement. Given the adaptability of trees to many different domains, providing solutions for this comparison problem will be applicable to a wide range of domains, from finding changes in reports of adverse effects for a drug, to monitoring lung cancer indexes in the country, or finding changes in traffic bottlenecks.

1.1 Overview of the Dissertation

The main research question of my dissertation is: how can information visualization techniques be effectively designed and implemented to help explore differences between versions of a tree? From my initial explorations, I found that answering this question would help solve real world problems such as identifying what has changed in the US Federal Budget (grouping accounts by Agencies and Bureaus) in the last year or whether there has been a change in the number of passengers flying in the US (broken down by State and then by City). To address these issues, I decided to first approach the problem of comparing two versions of a tree, and then with the feedback I collected, I realized that my approaches were expandable to the problem of determining the change on one tree that evolves over time. Therefore, in this dissertation I address these two main problems:

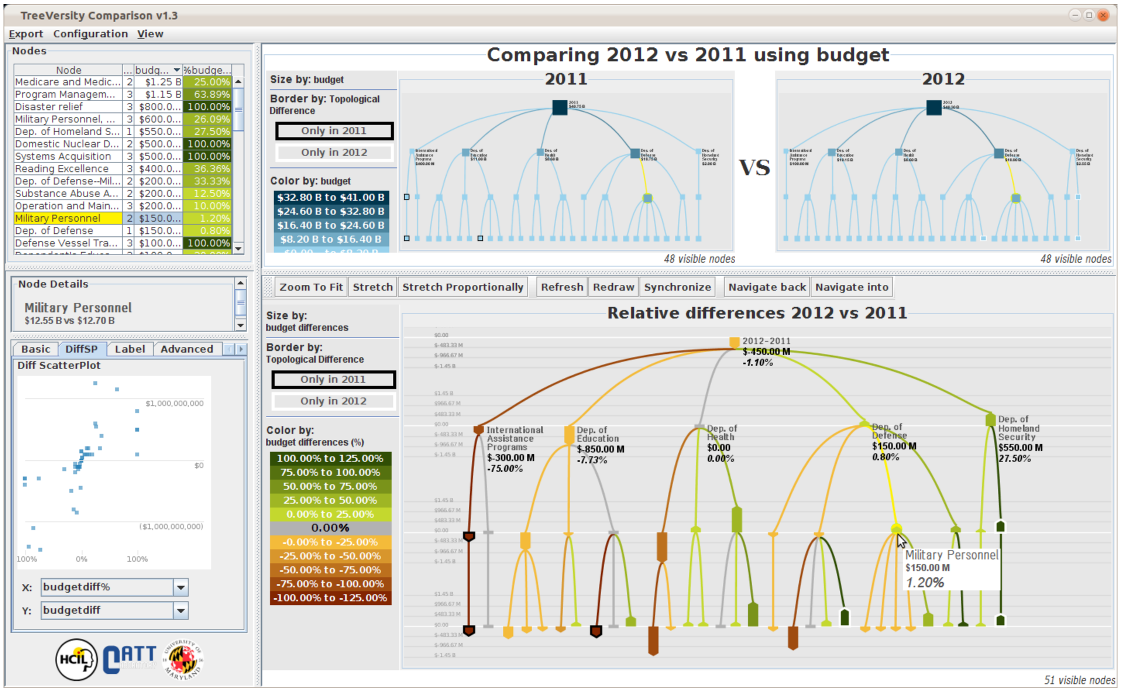
1. **How to help users find differences between two versions of a tree?**

   Comparing two versions of a tree is a significant problem which is why extensive research has previously been conducted on this topic. This research has mainly focused either on finding changes in topology between two trees (without values in the nodes) [13, 20, 28, 62, 66] or in finding changes in the values of a tree with fixed hierarchy [77, 86]. Previous approaches have focused on node value comparison or topological comparison, but none have effectively combined these techniques. While Ying Tu et al. [90] has attempted to combine both approaches, his treemap based solution only showed changes in leaves for aggregated trees. Therefore I aimed to create a solution that addressed both node value changes and some topology differences (created and remove nodes), which also worked for aggregated and non-aggregated trees (those where a node value cannot be calculated as a function of the values of its children).

   In my initial explorations, I found that users wanted to identify four characteristics of change (described in more detail in Section 1.1.2) in tree nodes: *1) direction of change*, *2) actual difference*, *3) relative difference*, and 4) if the *node was created or removed.* As a result, I collaborated with design professor Audra Buck-Coleman from the University of Maryland to design the *Bullet*, a simple but powerful glyph visualization that displayed all of the aforementioned characteristics. Consequently, I implemented the Bullet on *TreeVersity,* which is a tree comparison tool between two versions of a tree that computes and displays the following: 1) differences in node values and 2) nodes that were created or removed. Figure 1.1 shows the TreeVersity main interface using an artificially generated sample budget.

2. **How to help users find changes over time in datasets that can be categorized as trees?**

   With further development of case studies, users reported that they wanted to perform comparisons with more than two time points. Moreover, they wanted to see



TreeVersity Comparison v1.3

Export Configuration View

Nodes

Node	...	budg...	%budge...
Medicare and Medic...	2	$1.25 B	25.00%
Program Managem...	3	$1.15 B	63.89%
Disaster relief	3	$800.0...	100.00%
Military Personnel, ...	3	$600.0...	26.09%
Dep. of Homeland S...	1	$550.0...	27.50%
Domestic Nuclear D...	2	$500.0...	100.00%
Systems Acquisition	3	$500.0...	100.00%
Reading Excellence	3	$400.0...	36.36%
Dep. of Defense--Mil...	2	$200.0...	33.33%
Substance Abuse A...	2	$200.0...	12.50%
Operation and Main...	3	$200.0...	10.00%
Military Personnel	2	$150.0...	1.20%
Dep. of Defense	1	$150.0...	0.80%
Defense Vessel Tra...	3	$100.0...	100.00%

Node Details

Military Personnel
$12.55 B vs $12.70 B

Diff ScatterPlot

X: budgetdiff%

Y: budgetdiff

Comparing 2012 vs 2011 using budget

Size by: budget

Border by: Topological Difference

Color by: budget

48 visible nodes

48 visible nodes

Size by: budget differences

Border by: Topological Difference

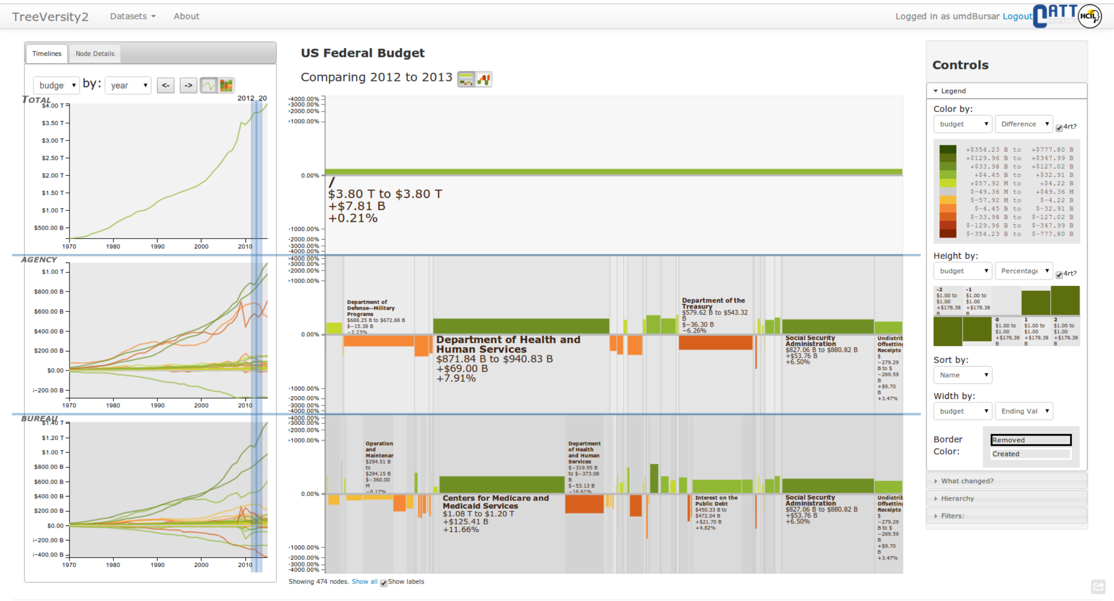
Color by: budget differences (%)

51 visible nodes

Figure 1.1: TreeVersity comparison interface. On the top are the two original trees being compared (budgets for 2011 and 2012). At the bottom the DiffTree shows the amount of change for each node. The glyph called "the bullet" points up to denote increase.

> the starting and ending values in the representations so they could identify the most significant nodes (e.g. Department of Defense, Department of Health and Social Security Administration are three of the main agencies in the US Federal Budget because of the size of their budgets). They also suggested that the comparison techniques would apply datasets that can be categorized as trees, but are not inherently hierarchical (e.g. University of Maryland's Student Information grouped by Ethnicity, Gender, and Type of Student). To support these requests, I designed the *StemView,* which is an area based visualization for tree differences that addresses the four characteristics of change targeted by the Bullet, with the addition of the starting or ending values. I incorporated the StemView in *TreeVersity2,* which is a web based information visualization comparison tool for finding differences in datasets over time using hierarchies. Figure 1.2 illustrates the TreeVersity2 main comparison interface that shows changes in the US Federal Budget.

This dissertation explains how I addressed these two research questions in detail, as well as describes 13 different Multi-dimensional In-depth Longitudinal Case Studies (MILCS) [78] developed with domain experts from different domains that validate the utility and flexibility of my techniques and implementations. Table 1.1 summarizes these case studies.



budge by: year

Comparing 2012 to 2013

Color by: budget Difference ☑ 4rt?

🟩	+$354.23 B to +$777.80 B
🟩	+$129.96 B to +$347.99 B
🟩	+$33.98 B to +$127.02 B
🟩	+$4.45 B to +$32.91 B
🟩	+$57.92 M to +$4.22 B
⬜	$-49.36 M to +$49.36 M
🟨	$-57.92 M to $-4.22 B
🟧	$-4.45 B to $-32.91 B
🟧	$-33.98 B to $-127.02 B
🟥	$-129.96 B to $-347.99 B
🟥	$-354.23 B to $-777.80 B

Height by: budget Percentage ☑ 4rt?

Sort by: Name

Width by: budget Ending Val

Border Color: Removed Created

Figure 1.2: TreeVersity2 comparison interface. On the top are the two original trees being compared (budgets for 2011 and 2012). At the bottom the DiffTree shows the amount of change for each node. The glyph called the Bullet points up to denote increases, and down for decreases. Nodes that have the same value in both trees are shown as small gray rectangles. The created and removed nodes are highlighted with a thick white or black border respectively. In this example, the height of the Bullet is proportional to the absolute change (in Dollars) while the color is mapped to the percentage change making it easy to spot the changes that are significant in both absolute and relative terms, i.e. the dark tall bullets. Novice users can start with a redundant encoding using the same variable for both color and size.

Organization	Case Study	MILCS Stage	Driving Mode	TreeVersity Version	Data Size	Time Points	Example Tree Size	Number Attribs.	Number Vars.	Type of Tree	Tree Comparison Type
DOT	**Airlines Budgets**	Early	Chauffeur	1	216	N/A	187 (2 Levels)	3	1	Dynamic	Type 3: aggregated + different topology
OMB	**US. Federal Budget**	Early	Chauffeur	1 & 2	4,845	56	1,393 (4 Levels)	7	1	Mixed	Type 3: aggregated + different topology
DOT	**TRB Publications**	Early	Chauffeur	1 & 2	52,135	8,012	674 (2 Levels)	20	1	Dynamic	Type 3: aggregated + different topology
DOT	**Nat. Trans. Library Publications**	Early	Chauffeur	1 & 2	38,351	374	294 (3 Levels)	10	1	Dynamic	Type 3: aggregated + different topology
DOT	**Passengers flying in the US**	Early	Chauffeur	1 & 2	65,534	162	4,194 (3 Levels)	4	1	Mixed	Type 3: aggregated + different topology
NCI	**National Cancer Institute**	Early	Chauffeur	2	1,716	13	101 (3 Levels)	3	3	Dynamic	Type 2: non aggregated + same topology
FDA	**FDA Drug Adverse Effects**	Mature	Chauffeur	2	2,964	5	1,614 (4 Levels)	4	4	Fixed	Type 4: non aggregated + different topology
UMD	**UMD Budget**	Early	Chauffeur	2	16,332	5	1,296 (3 levels)	6	1	Mixed	Type 3: aggregated + different topology
UMD Bursar	**UMD Students Information**	Mature	Chauffeur	2	227,158	5	715 (5 Levels)	219	3	Mixed	Type 3: aggregated + different topology
eBay	**eBay Product Sales Data**	Early	User-driven	2	63,098	4	5,443 (4 Levels)	6	2	Fixed	Type 1: aggregated + same topology
CATT Lab	**Transportation Bottleneck Data**	Early	User-driven	2	96,205	24	286 (3 Levels)	7	4	Mixed	Type 3: aggregated + different topology
IDB	**Imports and Exports in the Americas**	Early	User-driven	2	119,741	19	3,766 (4 Levels)	5	1	Dynamic	Type 3: aggregated + different topology
DUTO	**Blind Students in Colombia**	Mature	User-driven	2	33,802	4	1,098 (3 Levels)	21	1	Mixed	Type 3: aggregated + different topology

Table 1.1: Case Studies Summary

1.1.1 Definitions

In this dissertation a tree is treated as the traditional data structure defined in computer science books composed by nodes and links that express the parent-to-child relationship, where each node, regardless of being leaf or inner node, follows three rules: 1) it is uniquely labeled in the tree, 2) contains one or more numeric variables with values over time and 3) contains one or more categorical attributes that may have more than one value.

Much work has been conducted on visualizing [50, 54, 56, 72] and exploring [19, 42, 70] single tree structures; however, the problem of comparing two trees is significantly more difficult. From my explorations, I have identified and classified five types of tree comparison (Figure 1.3):

Type 0: Topological differences between two trees where the nodes only contain a label. For example, finding differences between two phylogetic trees or trees of species where biologists want to identify which species are in the same position on the tree, that have moved, appeared or disappeared.

Type 1: Positive and negative changes in leaf node values with aggregating values in the interior nodes (i.e. trees that can be visualized with a treemap [50]) and no changes in topology. For example, comparing the stock market's closing prices between today and yesterday across a hierarchy of market sectors while assuming no stocks were created or deleted.

Type 2: Positive and negative changes in leaves and interior node values with no changes in topology. For example, comparing the salaries in an organizational chart between two years when no reorganization has occurred.

Type 3: Positive and negative changes in leaf node values with aggregating values in the interior nodes and with changes in topology. For example, finding changes in the U.S. Federal Budget, given that agencies or bureaus have been created or terminated.

Type 4: Positive and negative changes in leaves and interior node values with changes in topology. For example, comparing the number of website visits between two months

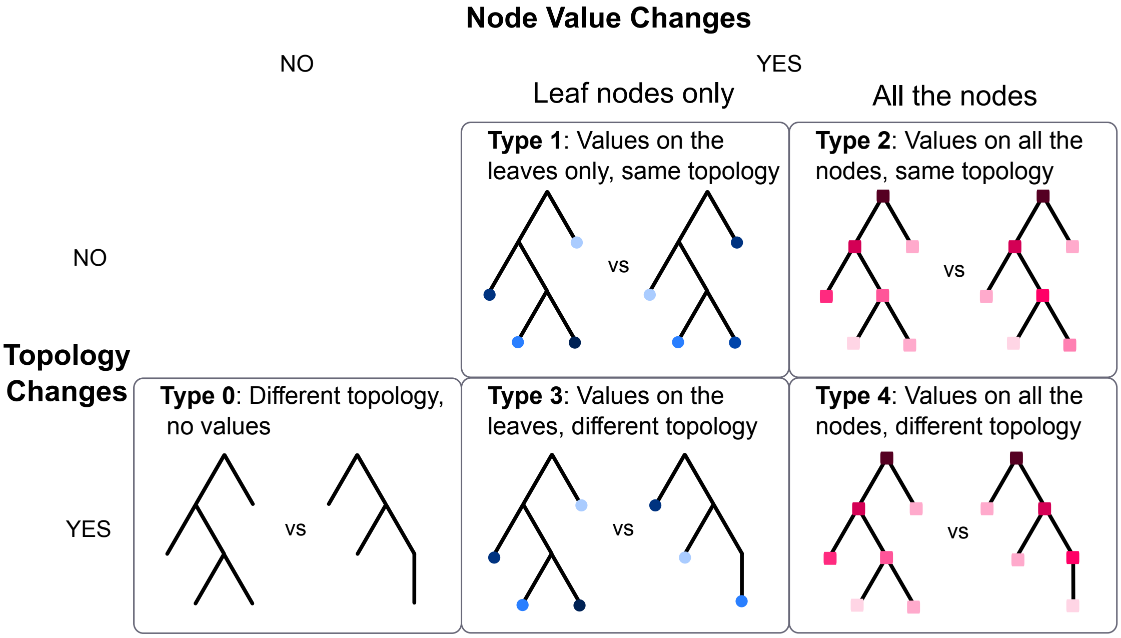

Node Value Changes
NO
YES
Leaf nodes only
All the nodes
Topology Changes
NO
YES
Type 1: Values on the leaves only, same topology
vs
Type 2: Values on all the nodes, same topology
vs
Type 0: Different topology, no values
vs
Type 3: Values on the leaves, different topology
vs
Type 4: Values on all the nodes, different topology
vs

Figure 1.3: Types of tree comparison problems. Current literature has addressed Types 0 and 1, with only one attempt at Type 3 [84]. TreeVersity2 supports all five cases, with emphasis on Types 1-4, the ones that include node value changes.

using the file hierarchy as a natural organization. Some pages may be created or removed and each page in the hierarchy has an independent number of visits.

1.1.2 Characteristics of node changes

According to related work and the feedback from our case studies, analysts that want to perform these types of tree comparisons want to be able to find and understand the following characteristics of a tree node:

Starting and Ending Values: the actual values of a node in the two compared time points. For example, the ending value for the Department of Defense was $672 billion in 2013.

Direction of change: positive, negative or neutral (no change).

Absolute change: the actual amount of change, e.g. the Department of Defense budget will be decreased by $15.99 billion between 2012 and 2013.

Percentage change: the absolute change with respect to the original value, such as the cut in the Department of Defense represents a 2.32% decrease with respect to its budget in 2012.

Created and Removed: nodes that are created, removed, or moved. e.g. The Bureau of Engraving and Printing ($140 million) is scheduled to be removed from the Department of Treasury in 2013 (i.e. does not have budget for 2013).

1.2 Contributions

The contributions of my dissertation are:

- The *Bullet* is a visualization glyph for tree nodes which shows four characteristics of change: direction of change, absolute change, relative change and if the node was created or removed. Moreover, the development of *TreeVersity*, a comparison tool to identify changes between two versions of a tree, which combines an implementation of the Bullet along with coordinated views and interactive filters to explore differences between two versions of a tree.

- The *StemView* is an area based visualization artifact, that shows changes in all the nodes of a tree (including interior nodes) and represents five characteristics of change: direction of change, absolute change, relative change, starting or ending values, and created and removed nodes. The implementation of the StemView in *TreeVersity2,* a web based information visualization tool, allows exploration of changes in datasets over time using hierarchies. Furthermore, designing and implementing the *reporting tool* which helps users navigate outstanding changes in the tree with textual representations and coordinated interactions.

- The development of 13 case studies with domain experts on real world comparison problems validate the utility and flexibility of the TreeVersity tools.

1.3 Dissertation Organization

The rest of this dissertation is organized as follows: Chapter 2 presents a literature review of the state of the art in tree comparison. I then present my approaches for the problem of comparing two versions of a tree on Chapter 3. Then, Chapter 4 details how I expanded my approaches to the problem of comparing one tree changing over time. Later, in Chapter 5 I present the 13 case studies that bring evidence to support the validity of my ideas and finally conclude in Chapter 6, which describes possible future research projects, that although will not be implemented in this dissertation, are envisioned as doors to be opened by this dissertation.

Chapter 2

Related Work on Tree Comparison

This chapter focuses on research that has been conducted on comparing, visualizing and analyzing multiple tree structures. There is substantial work on single tree structures, but since they are not relevant to the main objectives of this dissertation, I will not expand on them here. However I would refer the reader to the comprehensive surveys and compilations on [9, 18, 30, 44, 52, 75].

The related work has been categorized in four areas according to the described project's focus: topological comparison, node value comparisons, algorithmically oriented and other approaches.

2.1 Topological Comparison

Most of the tree comparison work has been done on comparing topological changes between tree structures. This tendency might have been influenced by the well-known problem of comparing different versions of evolutionary or phylogenetic trees. TreeJuxtaposer by Munzer et al. [63] is one of the best examples of topological comparison, presenting an efficient algorithm for comparing hierarchies. TreeJuxtaposer uses a node link representation with side-by-side comparison and a focus+context technique with guaranteed visibility. It scales well with the number of nodes, handling easily trees with two hundred thousand nodes; however, it is commonly limited to comparing two trees at a time. Figure 2.1 shows a tree comparison made with TreeJuxtaposer. Four species (sub-trees) have been selected and highlighted with colors on the tree on the left. The matching nodes are then highlighted with the corresponding colors on the tree on the right. TreeJuxtaposer is limited to topological differences and does not address node values comparisons, in the same way as the rest of the projects described in this section.

Double Tree by Parr et al. [67](Figure 2.2) uses a different approach for comparing two side-by-side phylogenetic trees. It relies on animations and user interactions (inherited from Spacetree [69]) to perform the comparison by collapsing the nodes in the tree, and displaying only the currently selected nodes and their local contexts. This allows for a better use of screen space but sacrifices the overall picture. MultiTrees by Holten & van Wijk [46] (Figure 2.3) also compares two tree structures using side-by-side Icicle-like [55] representations, mirroring one of them and drawing connections between the tree's nodes using Hierarchical Edge Bundling [45] to reduce cluttering. MultiTrees connections can become very busy, but are useful to represent splits and joins between the trees as shown in Figure [46].

Other good examples of side-by-side comparison are Graham and Kennedy's [26] Icicle-like [55] representation (Figure 2.4) and Bremm et al. [14] node-link visualization (Figure 2.5). These two solutions scale to the tens of trees by dividing the screen space into small interconnected views of the compared trees, but are limited by the screen size. In later work [28] (Figure 2.6) Graham & Kennedy addressed this by switching from small multiples to an aggregated representation using directed acyclic graphs (DAGs). Others have used the concept of aggregation of multiple trees in one view, like Furnas et al. [23] who proposed the concept in 1994 and CandidTree [57] (Figure 2.8) which uses a node-link representation combined with color, shapes and dotted lines to convey uncertainty of a node or link among different trees. Amenta and Klingner's TreeSet [6] (Figure 2.7) takes a different approach to comparing a large number of trees by calculating a bi-dimensional metric representing each tree and plotting them in a scatter plot.

TimeTree by Card et al. [21] (Figure 2.9) explored the concept of time changing hierarchies, combining Degree of Interest Trees (DOITrees) [41,64] with time sliders to analyze hierarchies that evolve with time.

The InfoVis2003 contest [40] promoted the development of projects on topological tree comparison. Some of the winning submissions presented innovative solutions for

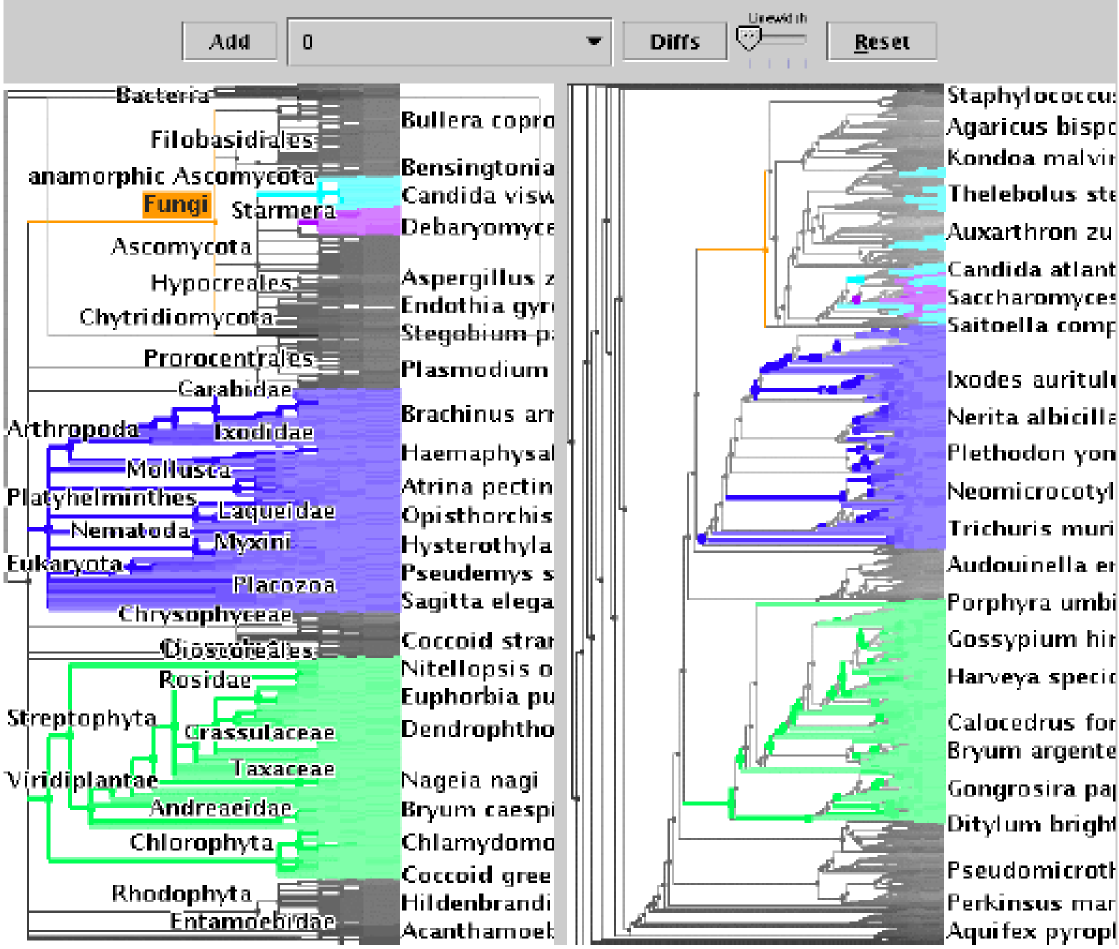
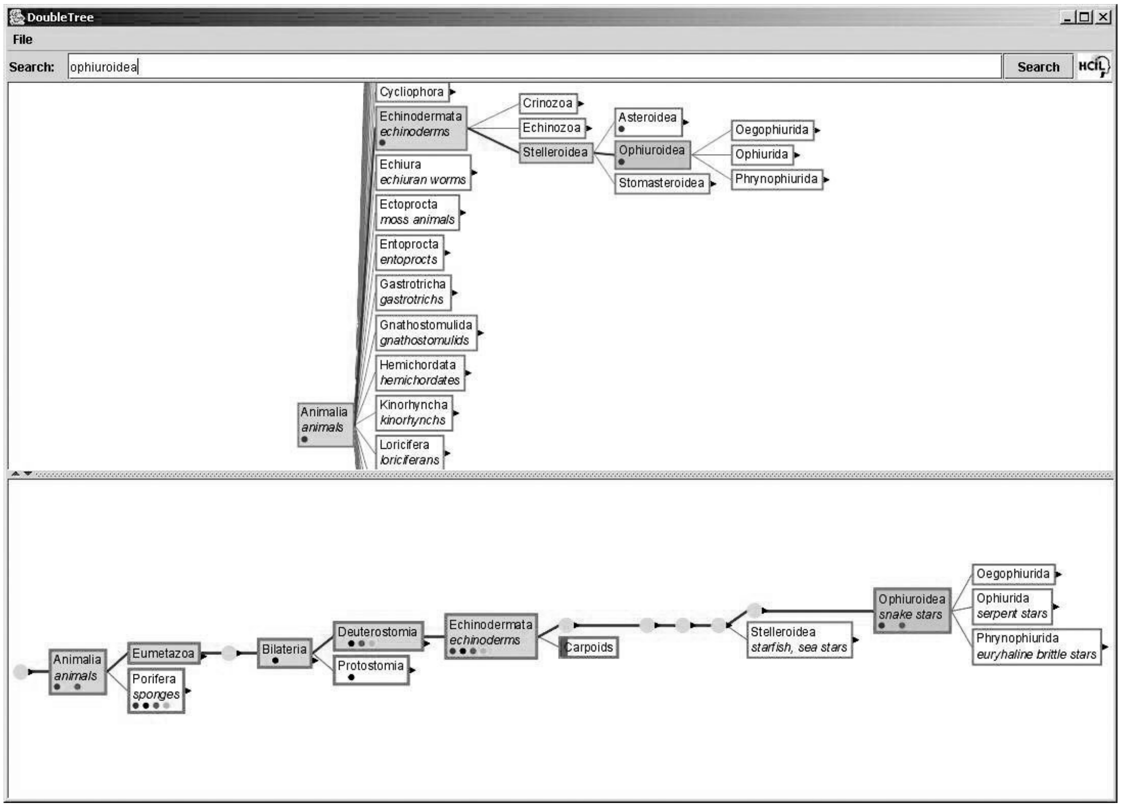


Figure 2.1: TreeJuxtaposer [63] comparing two phylogenetic trees. Four species (sub-trees) have been selected (highlighted colors) on the tree on the left, and the matching nodes are highlighted on the tree on the right. This example shows that the sub-trees on green and purple are staying together on the second tree, while the cyan and fuchsia are mixed together and are in different locations in the tree.



Search: ophiuroidea

Figure 2.2: Double Tree

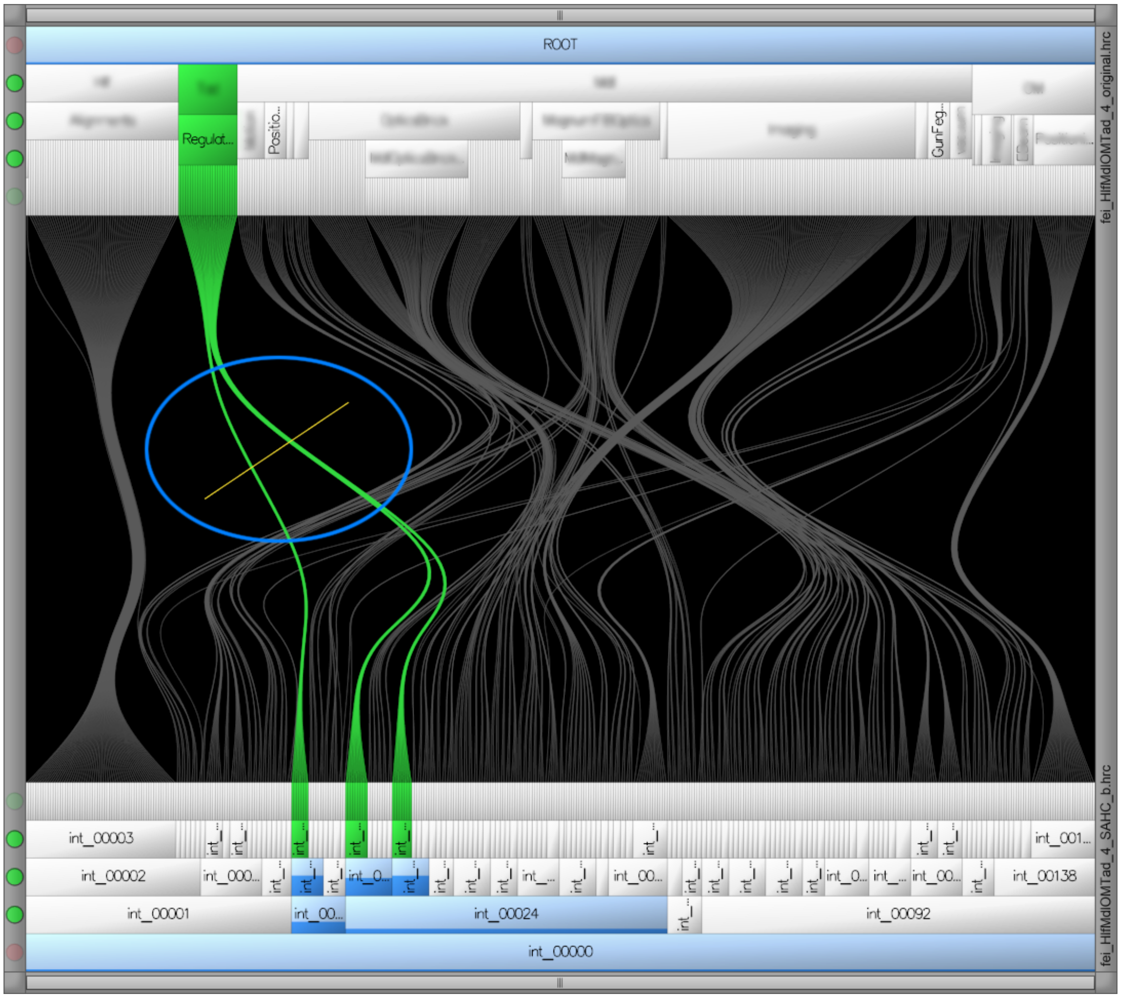


Figure 2.3: MultiTrees by Holten and van Wijk [46]showing the changes between two versions of a software repository. The gray lines in the middle represent connections between the trees. The highlighted nodes on green, show how a package from the top gets split into three different components in the hierarchy on the bottom.

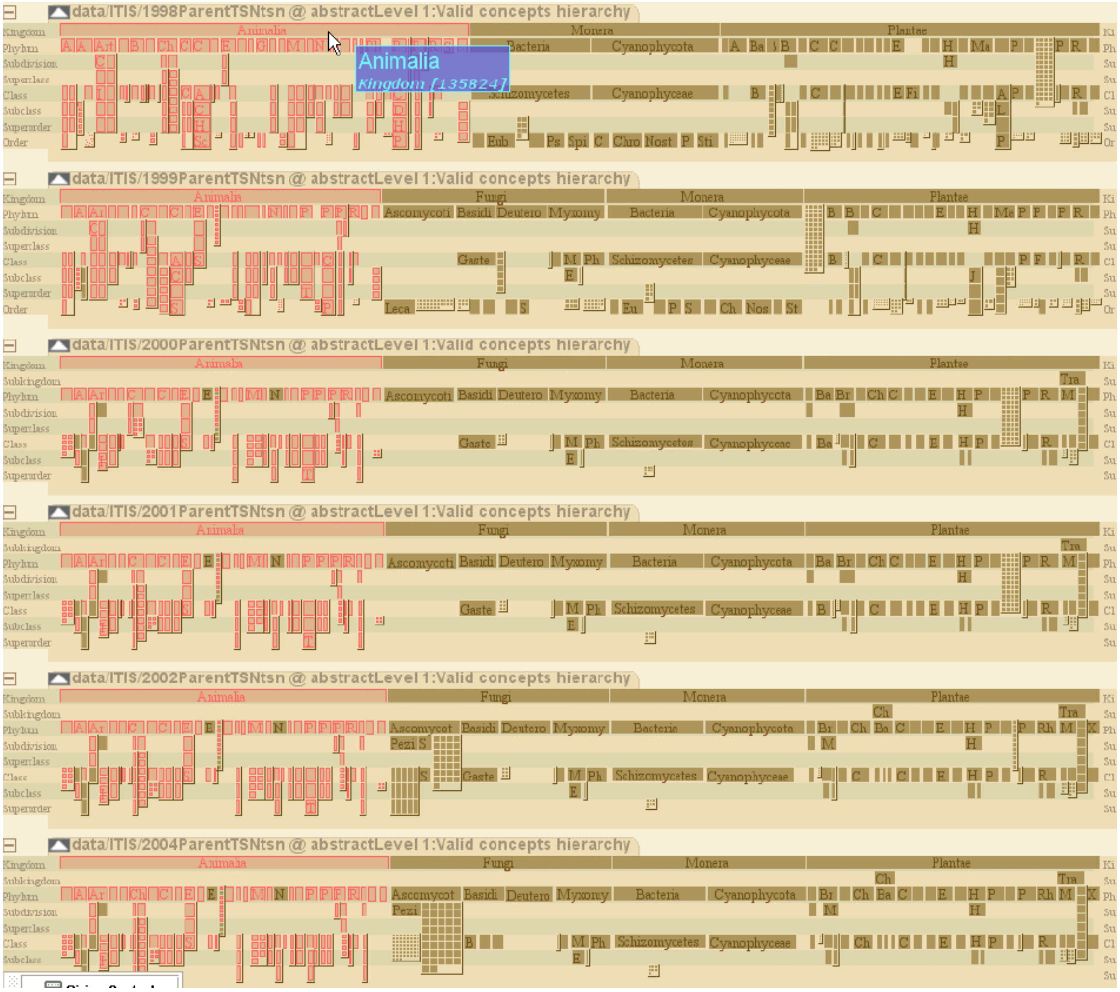

data/ITIS/1998ParentTSNtsn @ abstractLevel 1:Valid concepts hierarchy
Animalia
Kingdom [135824]
Monera
Plantae
Bacteria
Cyanophycota
Schizomycetes
Cyanophyceae
Kingdom
Phylum
Subdivision
Superclass
Class
Subclass
Superorder
Order
data/ITIS/1999ParentTSNtsn @ abstractLevel 1:Valid concepts hierarchy
Fungi
Ascomycota
Basidi
Deutero
Myxomy
Gaste
Leca
data/ITIS/2000ParentTSNtsn @ abstractLevel 1:Valid concepts hierarchy
Subkingdom
data/ITIS/2001ParentTSNtsn @ abstractLevel 1:Valid concepts hierarchy
data/ITIS/2002ParentTSNtsn @ abstractLevel 1:Valid concepts hierarchy
Pezi
data/ITIS/2004ParentTSNtsn @ abstractLevel 1:Valid concepts hierarchy

Figure 2.4: Graham and Kennedy's TaxVis [26, 27, 31, 32]side-by-side comparison of up to ten different taxonomies with hundreds of nodes

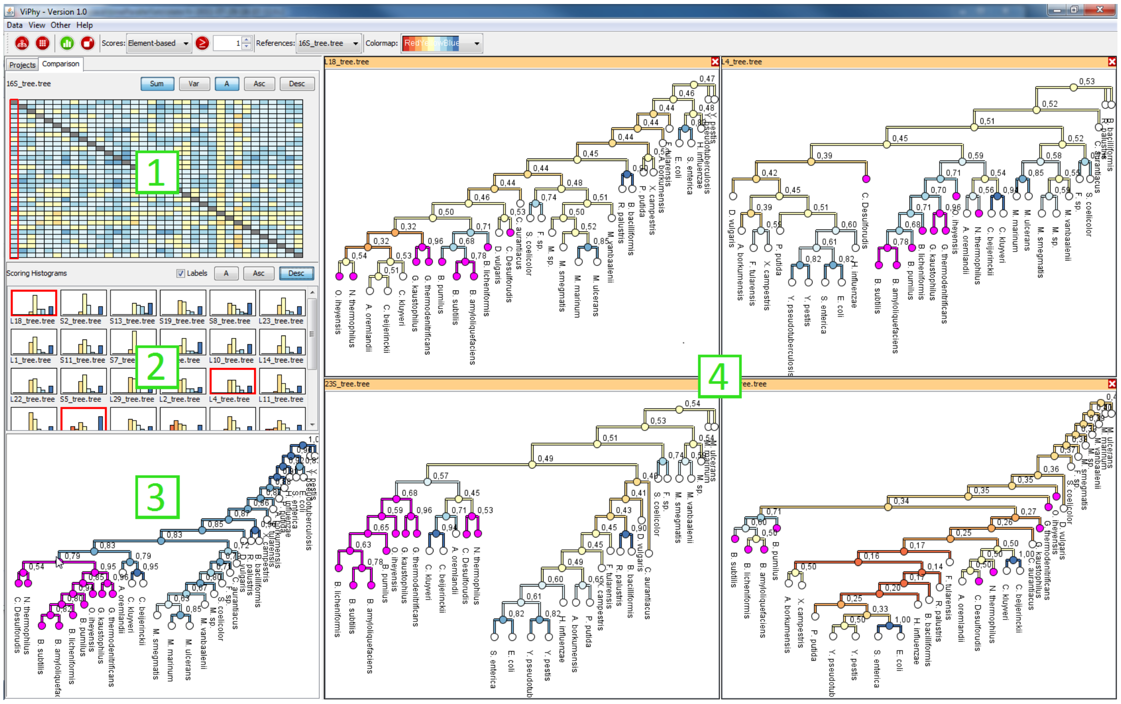


Figure 2.5: ViPhy [14] compares up to dozens of phylogenetic trees with a small number of nodes (around 50) using side-by-side views and color coding based on similarity algorithms.

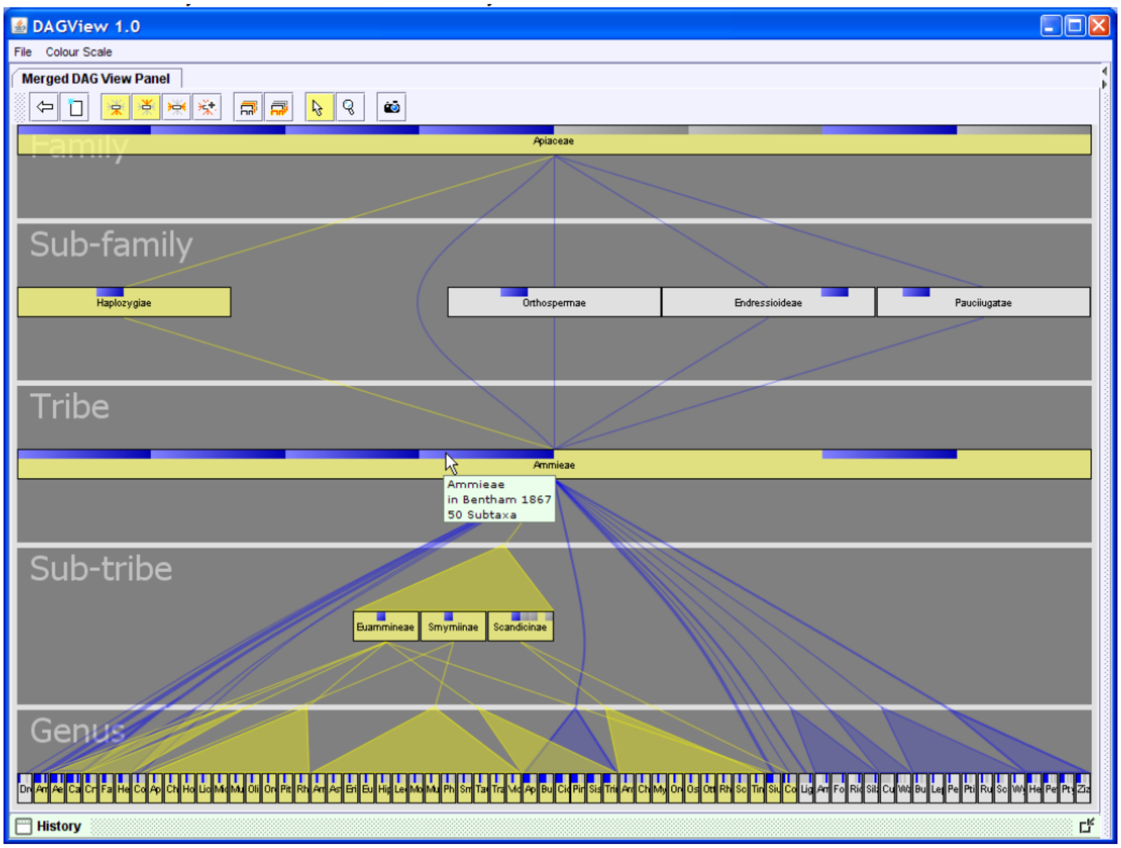
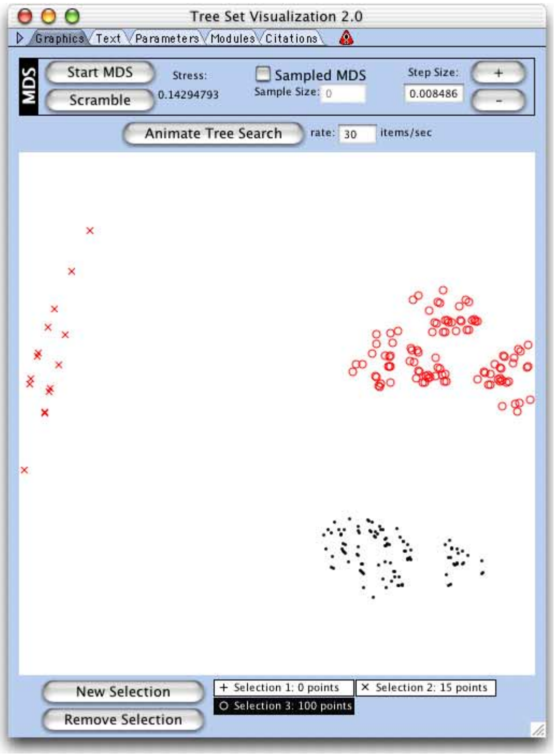


Figure 2.6: TaxVis DAGs comparison of eight different hierarchies. Each node in the graph represent a node in the trees, the ordered blue bars inside each node encode the presence of each node in each of the hierarchies, for example the selected node in the middle (a "Tribe"), exists in the hierarchies 1,2,3,4 and 7.



Stress: 0.14294793

Sample Size: 0

Step Size: 0.008486

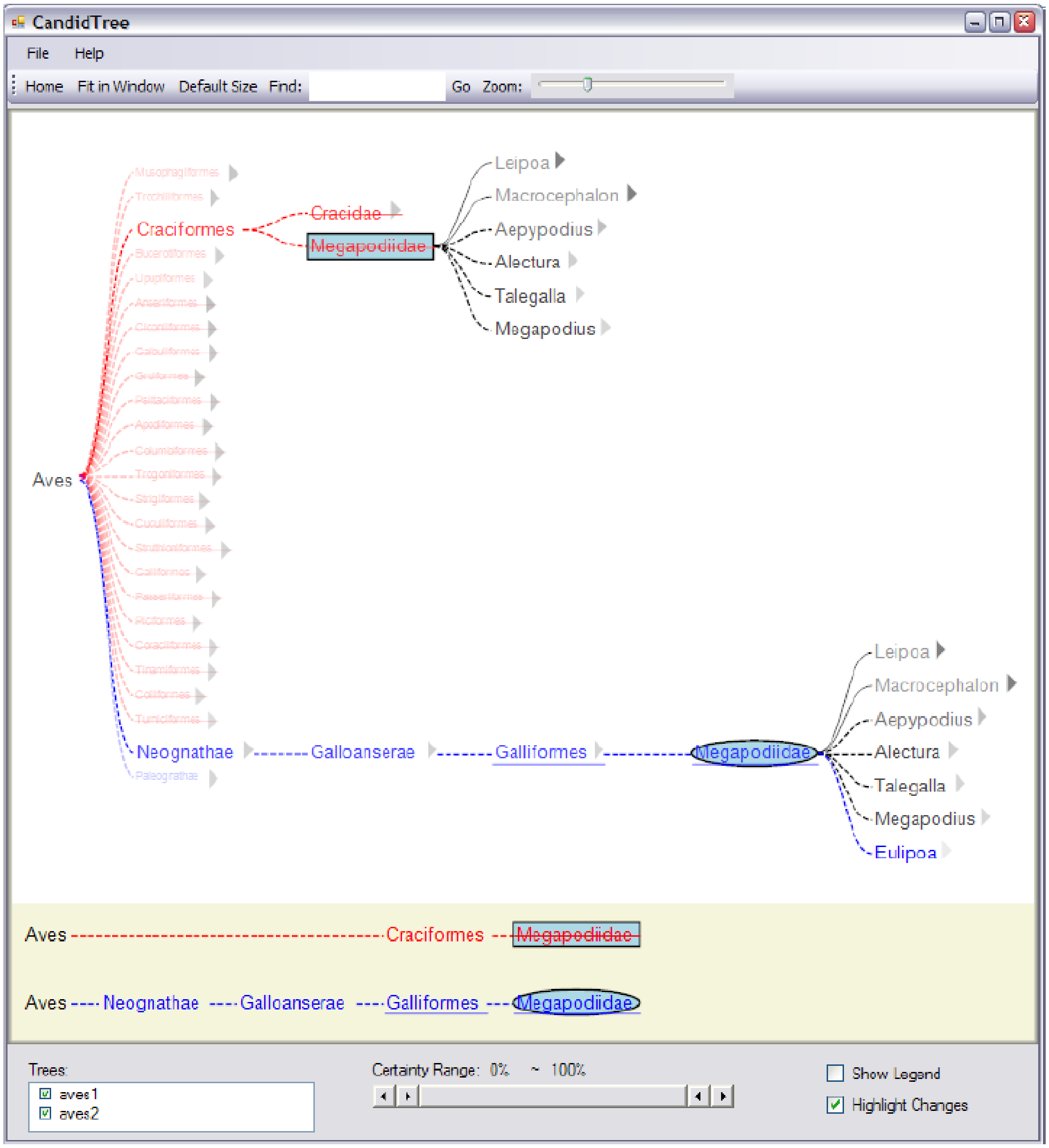
rate: 30 items/sec

+ Selection 1: 0 points

× Selection 2: 15 points

○ Selection 3: 100 points

Figure 2.7: TreeSet [6]point set representation of hundreds of trees. Each point represents a tree, and the distances between trees reflect the distances in a bi-dimensional metric.

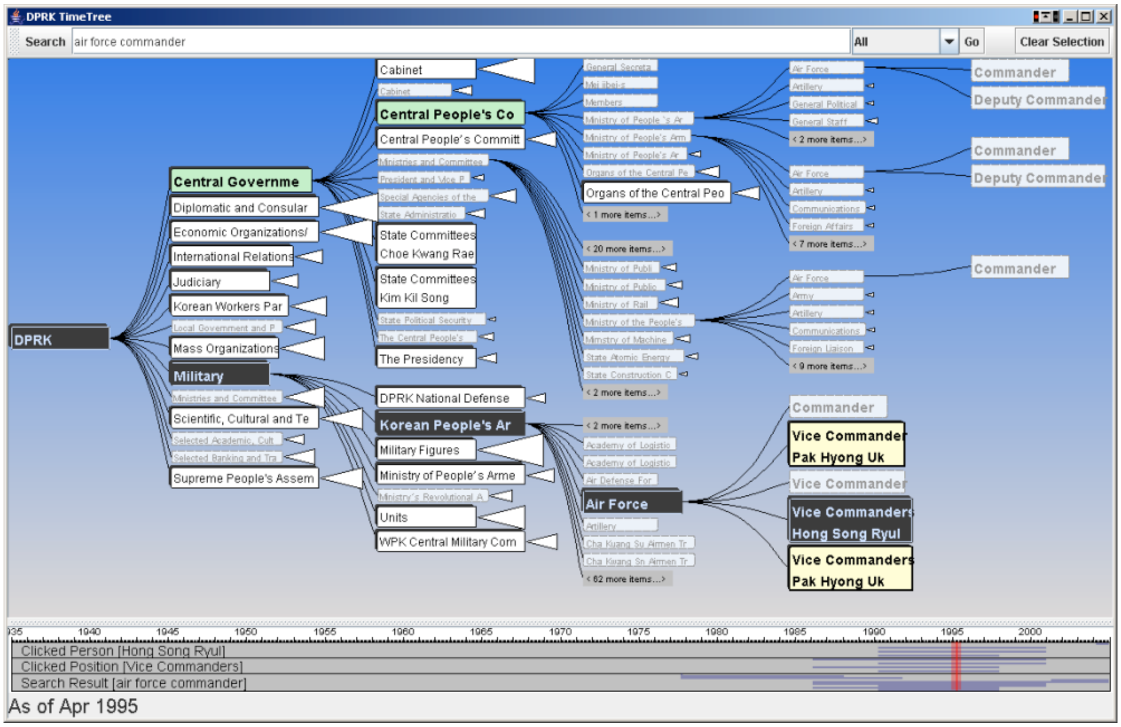


Trees:
☑ aves1
☑ aves2

Certainty Range: 0% ~ 100%

□ Show Legend
☑ Highlight Changes

Figure 2.8: CandidTree [57] visualizes uncertainty in a "merged tree" created from aggregating two trees. The application represents two types of uncertainty: "location of the node relative to its parents and the sub-tree structure of a node." CandidTree uses a SpaceTree-like [69] navigation scheme and dotted lines to represent the uncertainty.



DPRK TimeTree

Search air force commander

All

As of Apr 1995

Figure 2.9: TimeTree combines Degree Of Interest Trees with time-sliders to analyze the evolution of a tree structure over time.

the problem, such as TreeJuxtaposer [63], already described. Others include Zoomology [47] which used radial representations combined with zooming interfaces, InfoZoom [80] which used condensed side-by-side tables, EVAT [7] with radial side-by-side comparisons and TaxoNote [61] with a condensed Microsoft Windows Explorer-like representation. However, many of these promising projects did not evolve beyond the competition's two page submission requirement.

Finally, other approaches use zooming interfaces such as MoireTrees [60], which allows navigation of multi hierarchies (different trees that categorize a shared group of leaf nodes) using zooming and radial displays, and the already mentioned DoubleTree [67], that uses two connected, side-by-side SpaceTrees [69] to highlight topological differences between taxonomies.

Despite the substantial work on topological differences between trees, to the best of my knowledge, none of these solutions addresses the problem of comparing changes in node values. TreeVersity takes the task of tree comparison one step further, by looking also at node value changes, and therefore tackling a richer set of problems than those solutions restricted to topological differences only. However, TreeVersity allows the exploration of created and removed nodes, moved nodes are not currently supported; the reason for this decision is that none of the thirteen case studies conducted required the tracking of moved nodes. If it were necessary for different problem domains, and given that the tree nodes used in this dissertation are required to be uniquely labeled, support for moved nodes could be easily added by modifying the comparison algorithm to track changes in node's parents between versions of the tree as done by Graham and Kennedy [[33].

Apart from supporting the comparison of node value changes with identification of created and removed nodes, TreeVersity also improves over the related literature by supporting trees with **fixed**, **dynamic**, and **mixed hierarchies**, as well as **aggregating** and **non aggregating** trees. Most of previous projects, such as those described in this section were designed for a specific problem domain and therefore focused in only one of these tree types.

2.2 Node Values Comparison

The work on comparing node values is more limited than the one on topological differences. Previous projects in this area have usually employed treemaps. The original treemap tool [50] included a menu option to display the changing values on the tree, but was not developed for comparison. Animated Treemaps [24] represented changes in the node's attribute values using animation, focusing on stabilizing the layout. Both projects rely on user's memory to keep track of the amount of change and the location of the nodes which can be taxing and confusing. TreeVersity in contrast combines side-by-side comparison with explicit differences visualizations that allow users to navigate differences in a more explicit way. SmartMoney's Map of the Market [87] (Figure 2.11) represents stock market price changes using colored treemaps[1]. The Map of the Market is easy to understand and is commonly used by stock analysts, however it only presents -relative- node value differences in leaf nodes without topological changes, or what was called problem Type 1 in the introduction.

Contrast Treemap [84] is to the best of my knowledge, the only project that combines topological differences with changes in node values. It modified the traditional treemap technique by splitting each of the nodes' rectangular shapes into two complementary color triangles to represent value changes and structural differences. The shade of color and the area of the triangles represent both the values of the two nodes compared, in the case of Figure 2.10 the points per game of each NBA Player in the 02-03 season (upper left triangle) vs the same statistic in the 03-04 season (bottom right triangle). The hue of the color represents the topological changes: blue to black colors are for players that played in the same team two seasons, lime green for transfered players (moved nodes) and dark yellow for new players (created nodes). Since the size of the rectangles represents the number of total points per game of each player in the second seasons (03-04), removed nodes are not displayed. Compared to the Contrast Treemaps, TreeVersity and TreeVersity2 shows changes in all the nodes of the tree, not only leaf

[1] http://www.smartmoney.com/map-of-the-market/

nodes. Moreover TreeVersity and TreeVersity2 also support non-aggregating trees (tree comparison problems Type 2 and 4), as well as fixed, dynamic and mixed hierarchies.

2.3 Tree metric oriented comparisons (algorithmic comparison)

The final approach for tree comparison makes use of tree metrics, which usually are algorithms that calculate distances between two or more trees. These metrics work on non-labeled trees, and therefore also work with labeled trees (ignoring the labels), and can be classified by the type of comparison they make, and Bille [10] presents an excellent survey of them. According to him, the most important classes of metrics are *Edit Distance*, *Alignment Distance* and *Inclusion (sub-tree finding)*. In this work he describes efficient algorithms for each of this areas that could be used to compare many trees at once.

Another common related strategy for analyzing multiple trees is the consensus tree [6, 58, 81, 83]. This a technique used in phylogenetic analysis for summarizing many trees into one. This dissertation focused on information visualization approaches for comparing for labeled trees, therefore the algorithms described in this section were not used.

2.4 Other Approaches

The Multiple Skylines Graphs by Caemmerer is a visualization artifact designed to show changes in datasets. It uses the concepts of variable width bar charts that are similar to the ideas used on the StemView, however was not designed for tree structures and therefore does not support them. The Skylines, shown in Figure 2.12 (taken from `http://www.slideshare.net/billcaemmerer/telling-the-data-comparison-story-using-a-skyline` were featured in an on-line article in the SAP Design Guild [16] and does not seem to have been academically published anywhere else. On the other hand, Brodbeck et al. work [15]on visualizing survey results use a area filling hierarchical visualization base

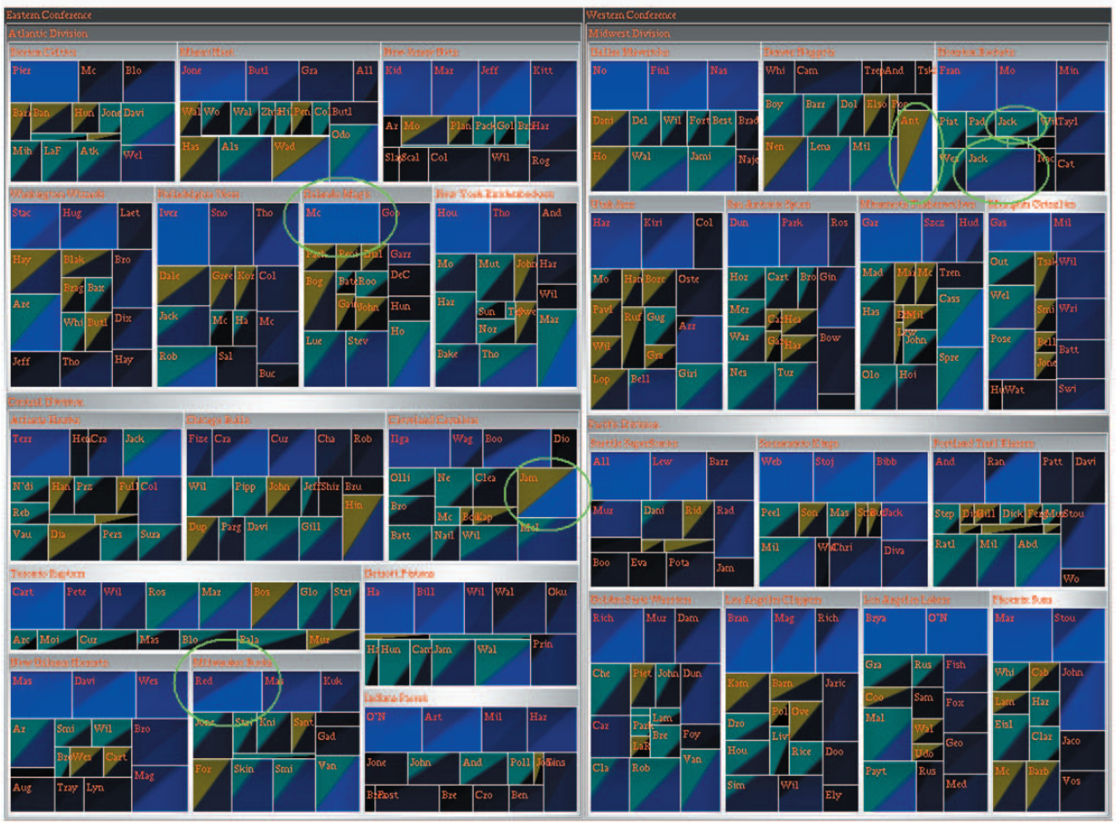
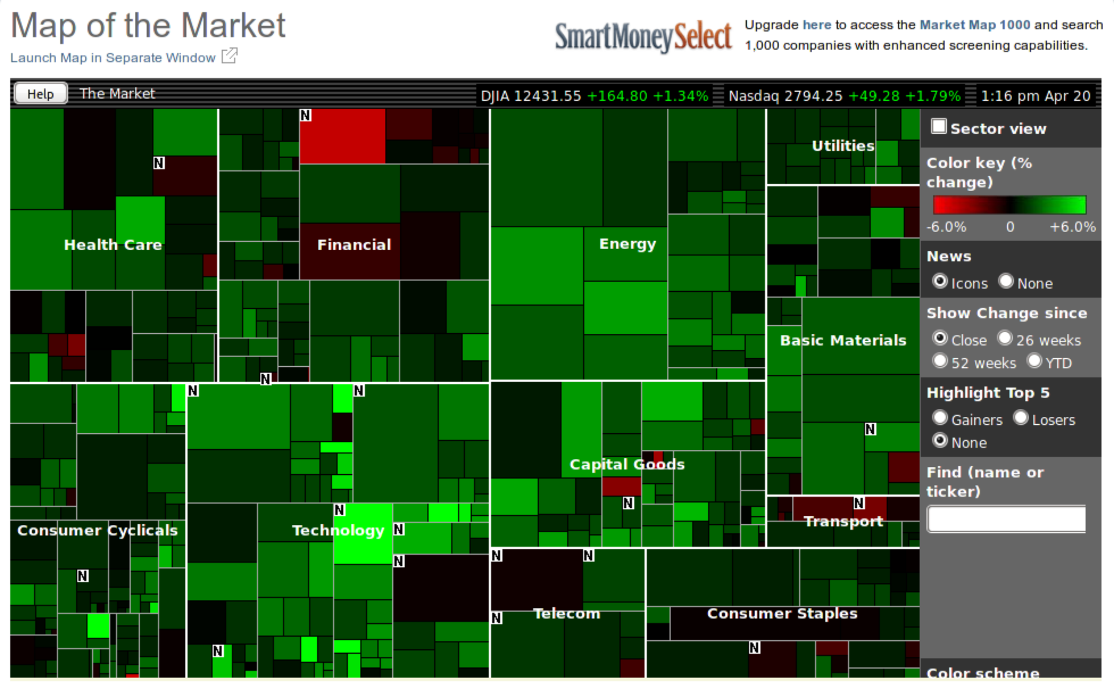


Figure 2.10: Two Corner Contrast Treemaps [84] use color and different shading techniques to encode node value and structural changes. The image shows the differences in NBA players' points per game between two seasons, categorized by teams and conferences. The paper [84] explains: "For an item, if both corners are in the blue to black range, the player was in the same team for both seasons. If the color for the 02-03 season is pine green, it means the player transferred to this team in the second season. If the color for the 02-03 season is dark yellow, the player joined the NBA in the second season."



Map of the Market

Launch Map in Separate Window

SmartMoneySelect Upgrade here to access the Market Map 1000 and search 1,000 companies with enhanced screening capabilities.

Help The Market DJIA 12431.55 +164.80 +1.34% Nasdaq 2794.25 +49.28 +1.79% 1:16 pm Apr 20

□ Sector view

Color key (% change)

-6.0% 0 +6.0%

News

● Icons ○ None

Show Change since

● Close ○ 26 weeks ○ 52 weeks ○ YTD

Highlight Top 5

○ Gainers ○ Losers ● None

Find (name or ticker)

Color scheme

Figure 2.11: SmartMoney's Map of the Market visualizes changes in the stock prices from one day to the next. Shades of red represent decreases and shades of green increases. The stocks are grouped by type and the size of the box represents the company's market capitalization.

with overlying line graphs. This is a similar technique to the one used for the StemView, but was not designed for showing change, and uses a different representation.

LifeFlow [36, 89] a temporal categorical data exploration tool, included an option for using non temporal attributes to compare different trees side by side. LifeFlow was the inspirational work of TreeVersity and did not include techniques for a more in depth comparisons others than visual inspection.

On the area of the reporting tool presented in this dissertation, the related work has focused mainly on annotating charts to offer more information about it. Kong and Agrawala [53] developed a system that analyzes a bar chart, computes and display graphical overlays such as reference structures, highlights, redundant encodings, summary statistics, and annotations; these graphical overlays were designed to enrich the information provided by static graphs, and do not allow for interactions. In contrast, the reporting tool presented in this dissertation was designed to guide users through the most significant changes in a visualization using interactive textual representations.

Contextifier by Hullman et al. [49] is another example of an annotation system for visualizations. Contextifier was designed to generate annotated stock line charts using a news article of a company (owner of the stock). The tool explores the data for points of interest, which could be either visually salient points on the graph, or relevant information in the news article, and using this information produces a customized line chart with textual annotations. Apart from having a different problem domain, Contextifier differs from the reporting tool in that it was not designed to guide users through a list of the most relevant points in the chart, instead it highlights in a static way a selection of points of interest using overlays.

2.5 Summary

This chapter described the related work that inspired this dissertation. The projects outlined were classified in three main groups depending on how they addressed the problem of comparing trees. The first group was composed by the projects that compared trees

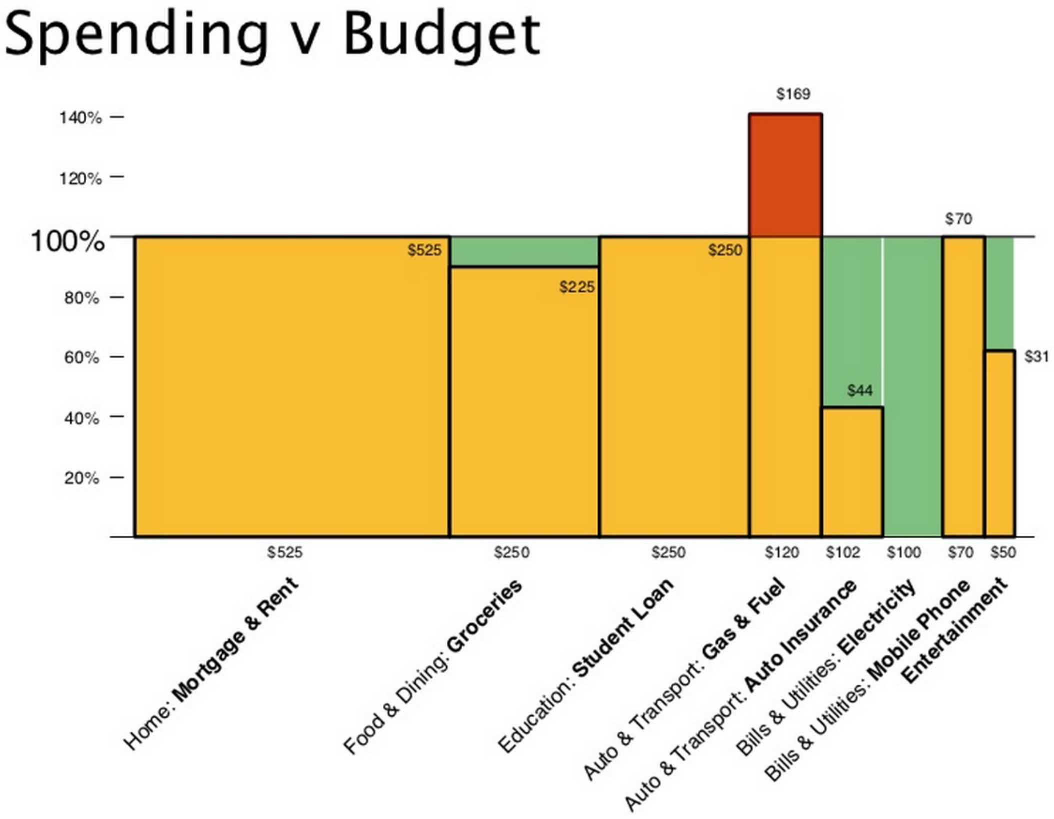


Figure 2.12: Multiple Skylines Graphs by Caemmerer is a technique that shows change using modified barcharts. The figure shows how the Gas & Fuel category increased from $120 to $169, while many others decreased, such as Auto Insurance, which went from $102 to $44. Bars over the horizon represent increases, and under the horizon decreases. The width of the bars represents the starting value of the category.

by looking at topological differences described in Section 2.1. Despite the considerable amount of research that has been conducted in this area, all of the projects found on it were designed for topological comparison only and didn't allow any type of node value comparison.

The second group of projects characterized in Section 2.2, were those performing node value comparisons between trees. A considerable smaller amount of work was found in this area, and most of the projects on it used some modification of treemaps to perform the comparison. From those projects addressing node value comparisons, only Contrast Treemaps [84] offered some type of topological comparison, however it uses an arguably difficult to understand codification and is limited to a subset of the problems that will be addressed in this dissertation.

Section 2.3 presented a general description of the algorithmically oriented projects, and described the main types of algorithms used for compared trees. Finally, I presented other approaches found on the literature related to my work.

The projects described in this chapter helped not only as inspiration for the techniques that this dissertation proposes, but also provide evidence of the relevance and importance of tree comparison as a difficult problem. They also illustrate the novelty of Treeversity and TreeVersity2 as a new approaches for comparing trees both on created and removed nodes, and node value differences.

Chapter 3

TreeVersity and the Bullet: Comparing two trees using node values and created and removed nodes

Hierarchies like those shown in Figure 3.1, help us organize and understand information. Many have researched visualizing, navigating and analyzing tree structures. Techniques such as node link representations [19,20,70], treemaps [77], Radial representations [22] and Icicle trees [54] are now often used in scientific and non-scientific publications. However, visualizing just a single tree representing a snapshot in time has limitations. Significantly less research has been conducted on how to compare tree structures that change over time.

When facing the task of comparing tree structures, users might find themselves asking:

- Where are the significant gains and losses in a complex budget proposal?
- How and where has congestion changed, either nationally and locally?
- How have airline maintenance budgets shifted from year to year?

The answers to these questions could be visualized by identifying changes on each node, a node being the individual elements of the tree, each having a type (e.g. State), a name (e.g. Maryland), and a value (e.g. a budget of US$200 million for the State). These changes can be of two types: topological differences (e.g. what nodes appear, disappear or move), and node attribute value differences (increases and decreases). Most work on tree comparison has focused on one or the other type of change, but not both. Despite the substantial work on topological differences between trees, none of their solutions addresses the problem of comparing changes in node values. Perceiving this limitation,

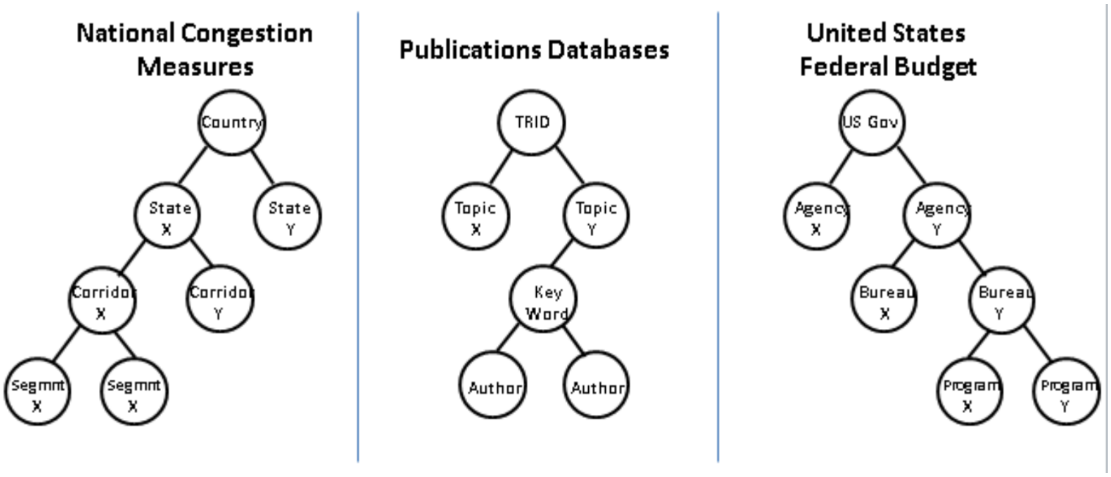


Figure 3.1: Example datasets viewed as hierarchies

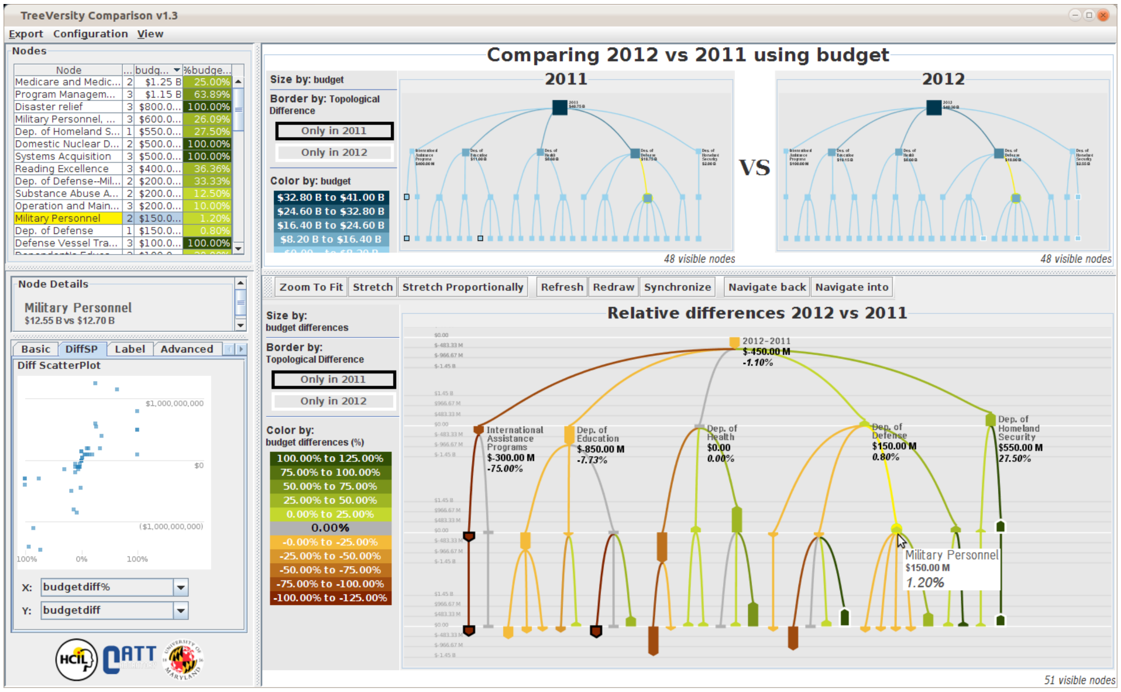
I created TreeVersity, a novel tree comparison tool able to address differences in both node values and changes in topology.

TreeVersity tackles a richer set of problems by combining a novel visualization technique, an interface design with coordinated views, interaction techniques, and comparison algorithms to support all five types of tree comparisons. TreeVersity was designed for power users interested in finding differences in hierarchical datasets (e.g. data analysts, researchers, and policy and planning officials making decisions based on data). However, I also envision TreeVersity as a communication tool useful for a broader audience.

3.1 TreeVersity

TreeVersity combines juxtaposition and aggregation techniques [25] along with interconnected views (Figure 3.2). The top of TreeVersity shows a side-by-side comparison of the two original trees. Below them, a third aggregated view called DiffTree shows the differences between the original trees. The three views are interconnected: selecting one node highlights and centers the two other corresponding nodes in the other views.

TreeVersity also displays the differences between trees in a tabular representation (top



Node	...	budg...	%budge...
Medicare and Medic...	2	$1.25 B	25.00%
Program Managem...	3	$1.15 B	63.89%
Disaster relief	3	$800.0...	100.00%
Military Personnel, ...	3	$600.0...	26.09%
Dep. of Homeland S...	1	$550.0...	27.50%
Domestic Nuclear D...	2	$500.0...	100.00%
Systems Acquisition	3	$500.0...	100.00%
Reading Excellence	3	$400.0...	36.36%
Dep. of Defense--Mil...	2	$200.0...	33.33%
Substance Abuse A...	2	$200.0...	12.50%
Operation and Main...	3	$200.0...	10.00%
Military Personnel	2	$150.0...	1.20%
Dep. of Defense	1	$150.0...	0.80%
Defense Vessel Tra...	3	$100.0...	100.00%

Military Personnel
$12.55 B vs $12.70 B

X: budgetdiff%
Y: budgetdiff

Size by: budget
Border by: Topological Difference
Only in 2011
Only in 2012
Color by: budget
$32.80 B to $41.00 B
$24.60 B to $32.80 B
$16.40 B to $24.60 B
$8.20 B to $16.40 B

Size by: budget differences
Border by: Topological Difference
Only in 2011
Only in 2012
Color by: budget differences (%)
100.00% to 125.00%
75.00% to 100.00%
50.00% to 75.00%
25.00% to 50.00%
0.00% to 25.00%
0.00%
-0.00% to -25.00%
-25.00% to -50.00%
-50.00% to -75.00%
-75.00% to -100.00%
-100.00% to -125.00%

2012-2011
$-450.00 M
-1.10%

International Assistance Programs
$-300.00 M
-75.00%

Dep. of Education
$-850.00 M
-7.73%

Dep. of Health
$0.00
0.00%

Dep. of Defense
$150.00 M
0.80%

Dep. of Homeland Security
$550.00 M
27.50%

Military Personnel
$150.00 M
1.20%

Figure 3.2: TreeVersity comparison interface. On the top are the two original trees being compared (budgets for 2011 and 2012). At the bottom the DiffTree shows the amount of change for each node. The glyph called "the bullet" points up to denote increase.

left of Figure 3.2). The table lists all the nodes currently displayed, also with tightly coupled highlighting. The columns include the name of the node, level in the tree, and absolute and relative differences of each attribute. Sorting columns allows the rapid selection of nodes with extreme values (e.g. largest relative difference).

3.2 The Bullet

The Bullet is a novel visualization glyph that allows the representation of four characteristics of change in tree nodes: direction of change, absolute change, relative change and if the node was created or removed (Figure 3.5). The compared trees seen in the top of Figure 4 use rectangular icons with color and size redundantly representing the attribute values of each node. To build the DiffTree seen in the bottom of Figure 3.3, TreeVersity

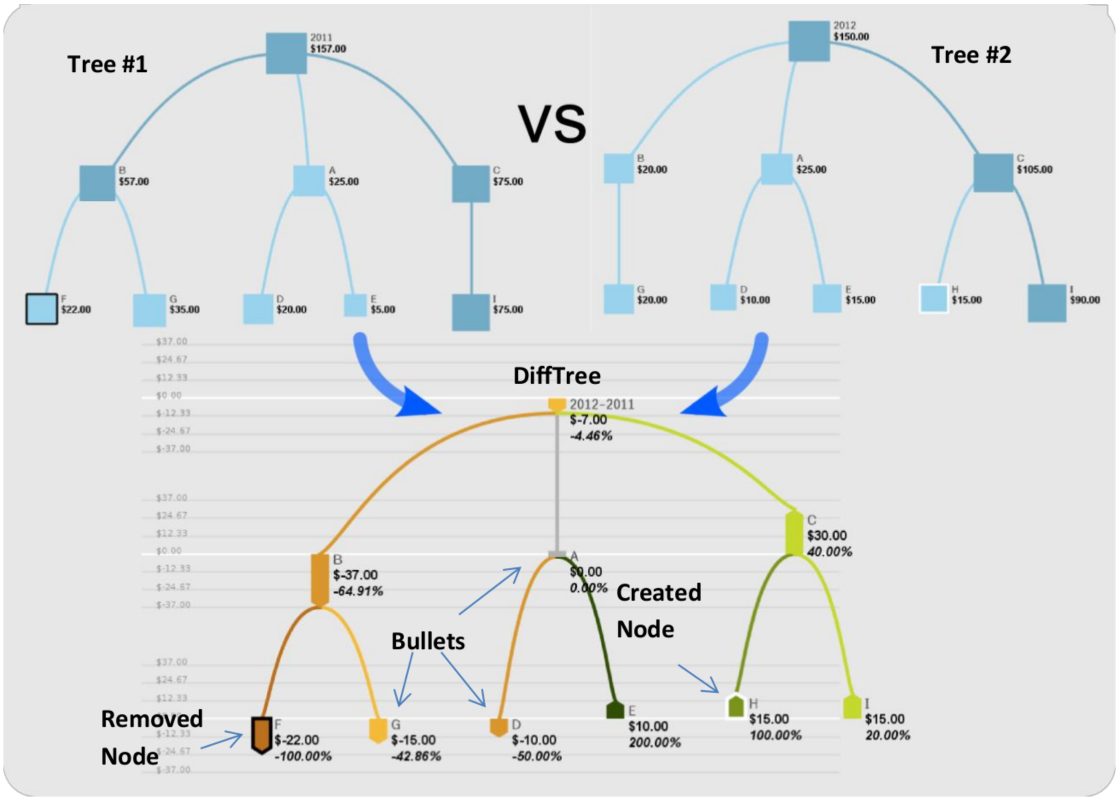


Figure 3.3: DiffTree construction. For each node, a "bullet" is created in the DiffTree whose size represents the absolute change and the color the relative change on it.

computes the difference in all the attributes for each original node. A positive difference indicates that the value on the right is larger than the one on the left. A node present on the left but not on the right is considered a removed node and its value in the DiffTree will appear as negative, assuming the value of absent nodes as zero.

The DiffTree (seen in the lower half of Figure 3.4) uses a novel glyph visualization—the Bullet—to represent differences between the two original trees. The Bullet glyph encodes the direction of the change, the amount of change, and the creating/deletion. The shape's direction represents the cardinality of the change: down for negative and up for positive in the vertical layout and left for negative and right for positive in the horizontal layout. The bullet size represents the amount of change. Color is used to encode both the cardinality and amount of change in the nodes. Users can select from preset color palettes that are binned in five steps to ease differentiation and accommodate visual preferences. The colors in the DiffTree are deliberately different from those in the original trees because they usually use very different value ranges. Gray rectangles represent nodes where the amount of change is zero. Finally, thick black or white borders are added around the bullet to denote removed or newly created nodes (white for added nodes, black for removed). By default both size and color are encoding the absolute amount of change (e.g. the amount in dollars in the case of a budget), but users can switch to relative change (i.e. percent change), or assign color and size to different characteristics of the changes.

Users can filter out specific nodes by differential amounts and/or by topological characteristics (created, removed, or present in both trees). The nodes are sorted according to the amount of change (absolute or relative). Users can also choose a left-to-right (horizontal) or top-to-bottom (vertical) layout. The original trees use rectangular icons with color and size redundantly encoding the attribute values of the nodes; the color and size scale uses the maximum possible values found in either tree, so that the ranges in both original views are the same, facilitating side-by-side comparison.

Different tree visualizations were considered for both the original and DiffTree views, and after a process of selection, the node link representations were chosen. In particular the Treemap was eliminated because - while it shines at showing leaf node values - it cannot show values for internal nodes and does not show the topological structure clearly. The node-link representation seemed to be more versatile to address the four types of tree comparison I wanted to address.

3.3 Interactions

3.3.1 Filtering

Users can filter the nodes by topological change, by range of values, and by maximum depth. Filtering by topological change allows users to see only the nodes that were created, or removed, or that are present on both trees. With the filter by node variables

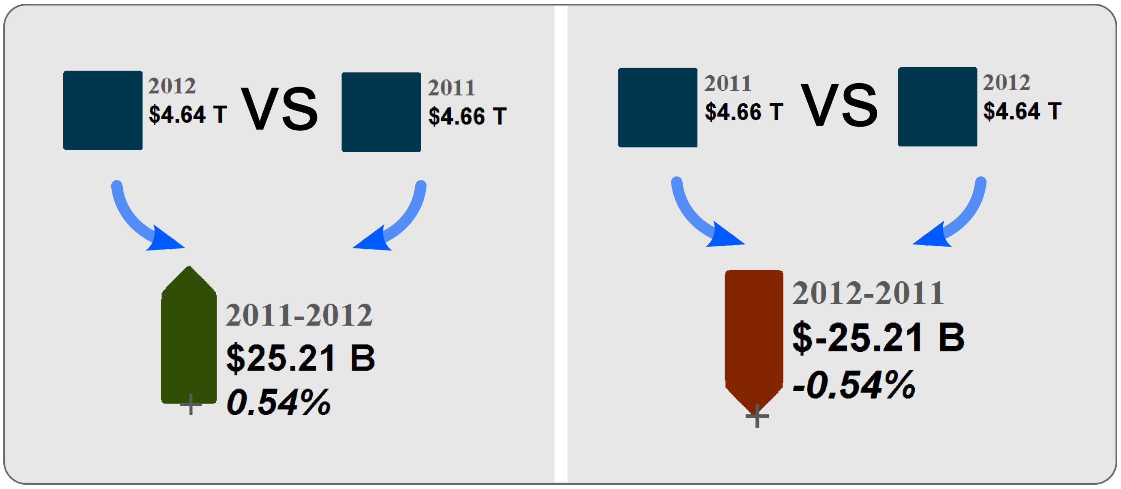


Figure 3.4: The Bullet representation. Shapes pointing up (or right for horizontal layouts) represent nodes with increases in their values (those where the value on the right tree is bigger than the corresponding node on the left tree) while shapes pointing down (or left) represent decreases. The size of the shape represents the amount of absolute (or relative) change compared with the rest of the tree nodes. The biggest shape corresponds to the node with the maximum value overall and the rest are normalized according to it.

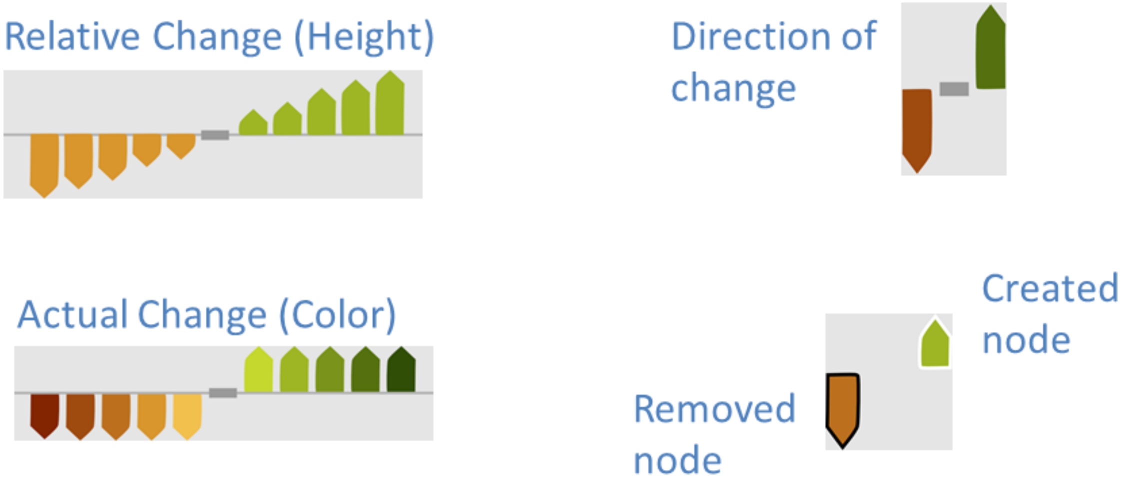


Figure 3.5: The Bullet represents four characteristics of change in tree nodes in one glyph.

range, users can keep visible only nodes whose values fall within a specific range, using absolute or relative amounts of change. Finally, the filter by maximum depth hides all the nodes that are deeper than a specified maximum depth. Another method of filtering uses a visualization called the *DiffScatterPlot*, seen in the call-out of Figure 3.6. The DiffScatterPlot lays out in a scatter-plot the summary of all the changes in the tree, distributing dots according to their absolute amount change on the y- axis, and to their percentage of change in the x-axis. Users drag a bounding rectangle with the mouse to select nodes of interest. This technique is especially useful for selecting outliers.

3.3.2 Overview

All three visualizations offer panning and zooming options for navigation. When analyzing trees with thousands of nodes, a zoomed out (macro) view of the whole tree can produce a cluttered mass of nodes. To reduce clutter, TreeVersity distributes the distance between the layers of nodes to fit the screen. This option is especially useful to understand the structure of the compared trees and of the DiffTree. Figure 3.8 shows the overview of the US Federal Budget grouped by BEA (Budget Enforcement Act) Category, which classifies accounts as Discretionary, Mandatory, or Net Interest.

3.3.3 Navigation

Users can focus on a subtree comparison. This is done by double clicking on the root node of the subtree of interest. After navigating into a subtree, all the views will be updated to display only the nodes on it; this is particularly useful in de-cluttering the screen. A navigation panel records navigated nodes and allows users to return to a previously navigated state.

3.3.4 Labels and Colors

TreeVersity offers multiple options to control the visualizations. Users can display the node's values and other descriptive information as an adjacent label. Users also have

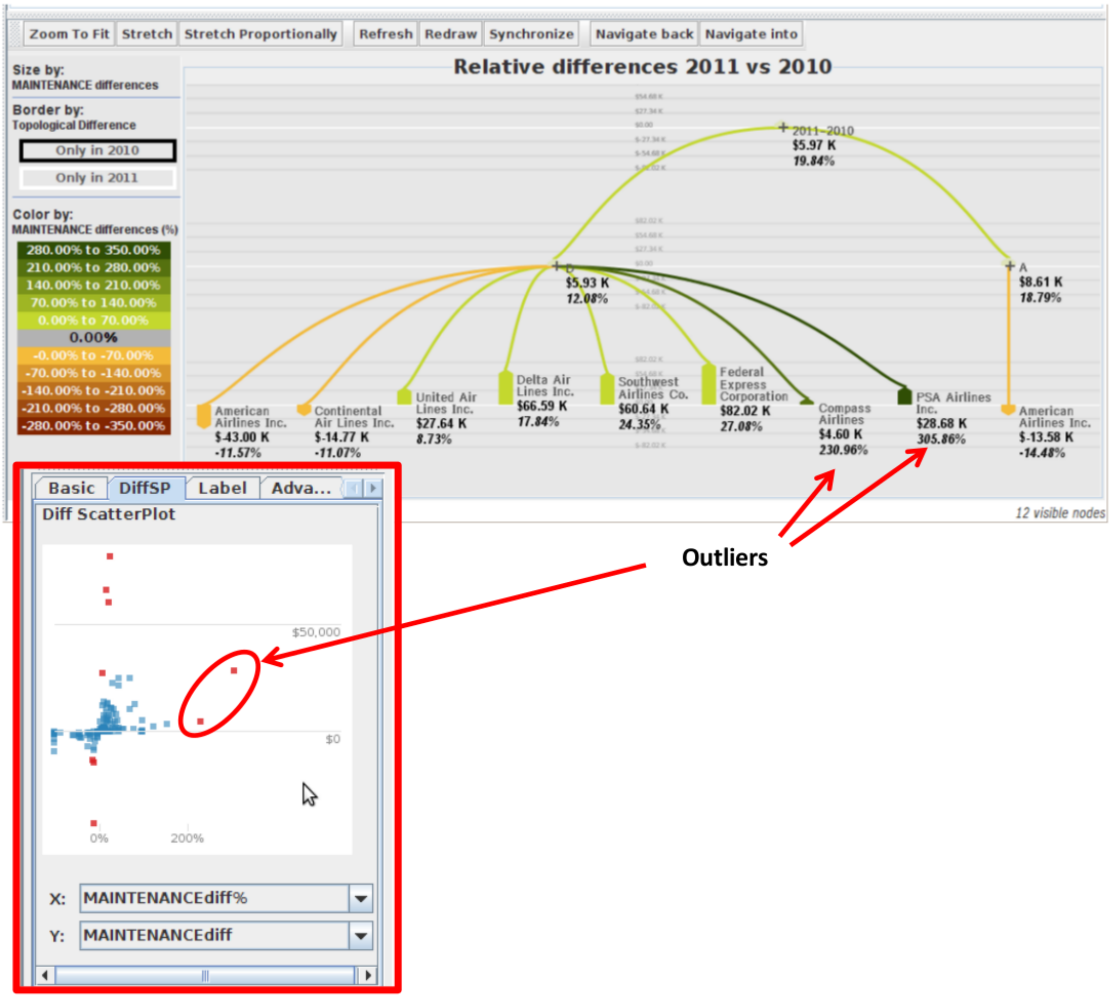


Figure 3.6: Airlines that changed their maintenance budget the most between 2011 and 2010 by region. Airlines were filtered by only including those that increased their budgets by more than $27,000, or by more than 200%, or that reduced their budgets by more than $13,500. The budgets are grouped by region. For the nodes at the first level in the tree, D stands for "Domestic" and A for "Atlantic" (i.e. Airlines operating over the Atlantic Ocean) . The values in the regions represent the average amount of change among all the airlines in that region. The root node shows the average overall.

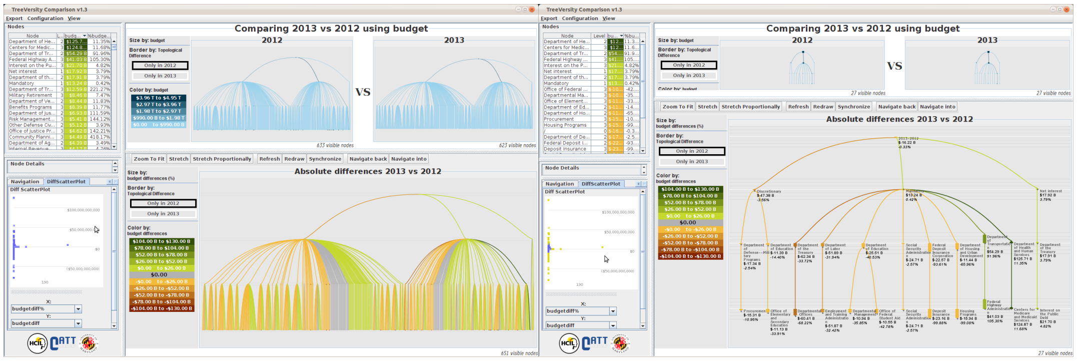
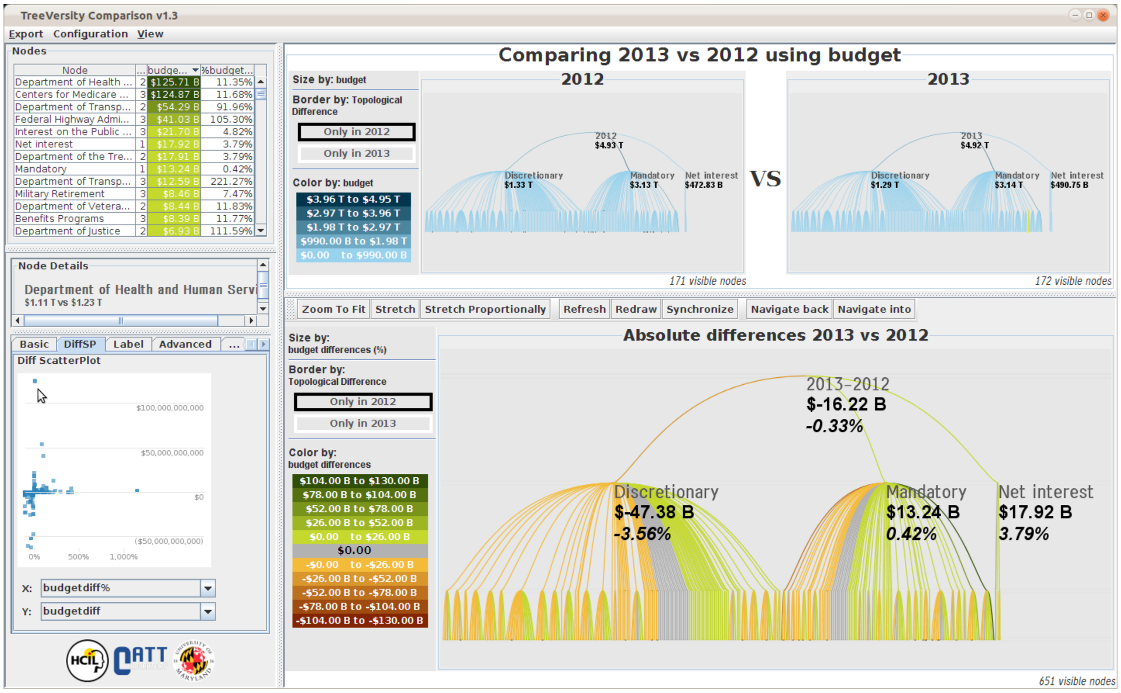

TreeVersity Comparison v1.3
Comparing 2013 vs 2012 using budget
2012
2013
VS
Absolute differences 2013 vs 2012
Discretionary
Mandatory
Net interest

Figure 3.7: US Federal Budget before and after filtering by the biggest changers



Node	...	budge...	%budget...
Department of Health ...	2	$125.71 B	11.35%
Centers for Medicare ...	3	$124.87 B	11.68%
Department of Transp...	2	$54.29 B	91.96%
Federal Highway Admi...	3	$41.03 B	105.30%
Interest on the Public ...	3	$21.70 B	4.82%
Net interest	1	$17.92 B	3.79%
Department of the Tre...	2	$17.91 B	3.79%
Mandatory	1	$13.24 B	0.42%
Department of Transp...	3	$12.59 B	221.27%
Military Retirement	3	$8.46 B	7.47%
Department of Vetera...	2	$8.44 B	11.83%
Benefits Programs	3	$8.39 B	11.77%
Department of Justice	2	$6.93 B	111.59%

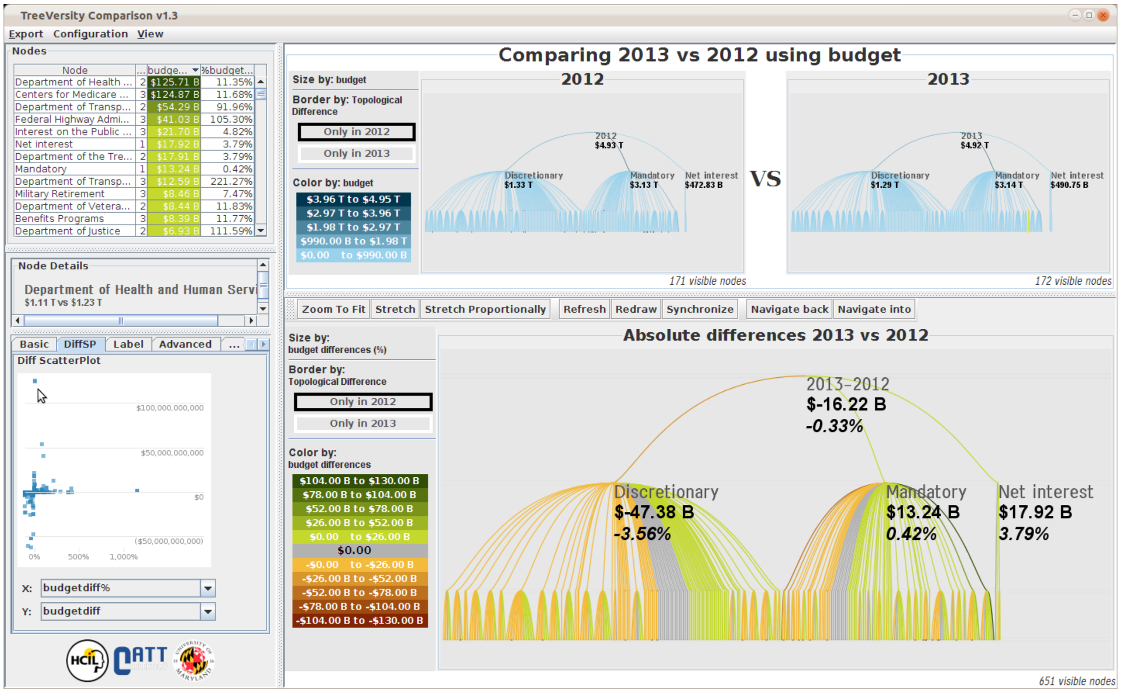
Node Details: Department of Health and Human Servi
$1.11 T vs $1.23 T

X: budgetdiff%
Y: budgetdiff

Size by: budget
Border by: Topological Difference
Only in 2012
Only in 2013
Color by: budget
$3.96 T to $4.95 T
$2.97 T to $3.96 T
$1.98 T to $2.97 T
$990.00 B to $1.98 T
$0.00 to $990.00 B

Size by: budget differences (%)
Border by: Topological Difference
Only in 2012
Only in 2013
Color by: budget differences
$104.00 B to $130.00 B
$78.00 B to $104.00 B
$52.00 B to $78.00 B
$26.00 B to $52.00 B
$0.00 to $26.00 B
$0.00
-$0.00 to -$26.00 B
-$26.00 B to -$52.00 B
-$52.00 B to -$78.00 B
-$78.00 B to -$104.00 B
-$104.00 B to -$130.00 B

Figure 3.8: Overview of the changes in the US Federal Budget between 2013 and 2012 grouped by BEA Category.



Nodes

Node	...	budge...	%budget...
Department of Health ...	2	$125.71 B	11.35%
Centers for Medicare ...	3	$124.87 B	11.68%
Department of Transp...	2	$54.29 B	91.96%
Federal Highway Admi...	3	$41.03 B	105.30%
Interest on the Public ...	3	$21.70 B	4.82%
Net interest	1	$17.92 B	3.79%
Department of the Tre...	2	$17.91 B	3.79%
Mandatory	1	$13.24 B	0.42%
Department of Transp...	3	$12.59 B	221.27%
Military Retirement	3	$8.46 B	7.47%
Department of Vetera...	2	$8.44 B	11.83%
Benefits Programs	3	$8.39 B	11.77%
Department of Justice	2	$6.93 B	111.59%

Node Details

Department of Health and Human Servi
$1.11 T vs $1.23 T

X: budgetdiff%

Y: budgetdiff

Size by: budget

Border by: Topological Difference

Only in 2012

Only in 2013

Color by: budget

$3.96 T to $4.95 T
$2.97 T to $3.96 T
$1.98 T to $2.97 T
$990.00 B to $1.98 T
$0.00 to $990.00 B

Size by:
budget differences (%)

Border by:
Topological Difference

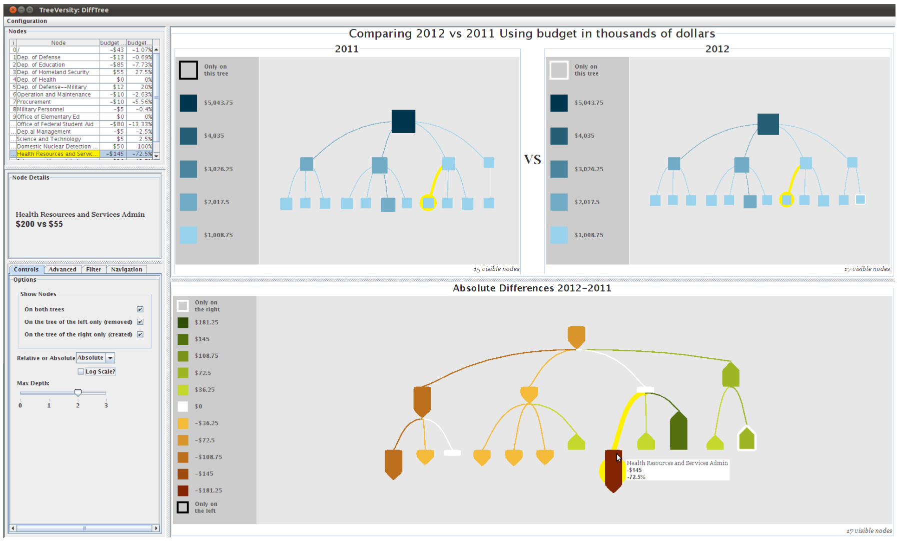
Only in 2012

Only in 2013

Color by:
budget differences

$104.00 B to $130.00 B
$78.00 B to $104.00 B
$52.00 B to $78.00 B
$26.00 B to $52.00 B
$0.00 to $26.00 B
$0.00
-$0.00 to -$26.00 B
-$26.00 B to -$52.00 B
-$52.00 B to -$78.00 B
-$78.00 B to -$104.00 B
-$104.00 B to -$130.00 B

Figure 3.9: Overview of the changes in the US Federal Budget between 2013 and 2012 grouped by BEA Category.



Nodes

i	Node	budget ...	budget...
0	/	-$43	-1.07%
1	Dep. of Defense	-$13	-0.69%
2	Dep. of Education	-$85	-7.73%
3	Dep. of Homeland Security	$55	27.5%
4	Dep. of Health	$0	0%
5	Dep. of Defense--Military	$12	20%
6	Operation and Maintenance	-$10	-2.63%
7	Procurement	-$10	-5.56%
8	Military Personnel	-$5	-0.4%
9	Office of Elementary Ed	$0	0%
...	Office of Federal Student Aid	-$80	-13.33%
...	Dep.al Management	-$5	-2.5%
...	Science and Technology	$5	2.5%
...	Domestic Nuclear Detection ...	$50	100%
...	Health Resources and Servic...	-$145	-72.5%

Node Details

Health Resources and Services Admin
$200 vs $55

Controls | Advanced | Filter | Navigation

Options

Show Nodes

On both trees ☑
On the tree of the left only (removed) ☑
On the tree of the right only (created) ☑

Relative or Absolute: Absolute

☐ Log Scale?

Max Depth: 2 (0 1 2 3)

Health Resources and Services Admin
-$145
-72.5%

Figure 3.10: Screen shot of the TreeVersity version used in the user study. Note the older legend.

control over how much information is provided (i.e. name of the node, its value, relative values, and other descriptions). TreeVersity maximizes spatial considerations when displaying the layout of the nodes and their corresponding labels, and, if necessary, users can select the option to truncate the labels.

The colors and size of the Bullets on the visualization can represent either the absolute or relative values (and differences) of the nodes. By default, the variable used for sorting the nodes is the same as the one used for coloring, so if users change the coloring, the nodes will be rearranged on all views (using animations) to fit the new ordering scheme.

3.4 User Study

A user study with eight participants was conducted to evaluate if users could understand the visual encodings and the basic interface organization of Treeversity without training. The dataset was presented as the budget of a hypothetical country but was in fact a small subset of the U.S. Federal Budget for 2011 and 2012. The node values were modified to include multiple sets of extreme changes that could be easily spotted by the users (if they could interpret the encodings accurately), such as an account receiving a budget increase in a department where all other accounts were getting decreases. The dataset had 46 nodes distributed over four levels with 1 node removed from 2011 and 3 created on 2012. The participants—three females and five males—were new students in their first week of the Master on Human Computer Interaction program. Their background varied from computer science to design and mathematics. They were unfamiliar with TreeVersity. After a one minute introduction explaining the objective of TreeVersity and the nature of the dataset, participants were asked to try the interface and to describe their perception and understanding of the interface using a think aloud protocol. No further training or orientation was provided. At the begin of the study, TreeVerity showed only the first three levels of the tree, a total of 17 nodes. The participants engaged with an earlier, less-evolved version of TreeVersity shown in Figure 3.10, the main differences where in the legend, and the color used to represent no change (white before, gray 20% now).All the desktop interactions and discussions were recorded.

For about five minutes participants explored on their own while I kept track of what had been correctly interpreted and learned (or not), using a checklist of expected concepts to be discovered. After five minutes misunderstandings were discussed and participants questions answered. If a particular feature had not been mentioned in the think aloud exploration, I would request the participant to speculate on the overlooked feature's purpose, and if any misunderstanding remain it was clarified.

Overall, all participants correctly interpreted most of interface components of TreeVersity without training. The first thing participants described was always the side-by-side trees and the DiffTree. They all correctly described the relationship between the three trees, and even that the tight coupling of the highlighting between the views was not described explicitly, all of the users started using it right away. Then participants talked about how they interpreted the glyphs. For the side-by-side trees, everyone immediately associated color with the amount of money at each node, however some people overlooked the size property (both color and size encoded the same information).

While the DiffTree is more complex than the side-by-side visualizations of the original trees, all participants were able to interpret the visualization of changes on their own.

By looking at the matching nodes on each view, and the shape and color of the bullets, participants were able to conclude that each node on the DiffTree was representing the amount of change on that node. The direction of the change was always understood correctly, by commenting on the color and direction of the bullet. Some people focused solely on the color of the bullet and seemed to ignore the meaning of direction and size, while others guided themselves by the shape alone and seemed to ignore the color. Since size and color encoded the same information it was appropriate. Some participants had problems understanding the older color legend used in the study. Whereas they only saw one number on the scale and did not guess that each color represented an interval, other participants thought it was clear and comprehensible. Participants had no problem understanding that nodes being small and white meant that their value had not changed, even in situations where a internal node had not changed but all of its children had changed significantly (with the sum of the changes being zero).

In the user study setup, size and color encoded the same information, but the meaning of the size of the bullets was not mentioned in the legend. Four subjects assumed that the size represented the percentage change while the color represented the absolute change, which was unexpected. To address this a better legend was developed, to make explicit the meaning of each characteristic of the bulltet. This suggests also that encoding both variables at the same time might in fact be a good idea as a default encoding as it fits the expectation of some users.

The created and removed nodes (represented with white and black thick borders respectively), were usually unnoticed initially, but all users eventually recognized them. They didn't seem to immediately understand what they meant, but figured it out either by looking at the legend or by using the coordinated views and noticing the node was missing in one of the views. Some users suggested that labels in the legend could be more meaningful e.g. "only in 2011" instead of "on the left only". The black and white colors I used initially to denote the topological changes were found confusing because white was already associated with nodes without change. I later changed the coding of zero as gray instead of white.

After the initial free exploration the participants were shown the larger tree with 46 nodes and asked to find significant changes. All of the subjects easily found many insights in the data. They followed a fairly consistent process: they started by looking at the created and removed nodes, then pointed out the nodes with the biggest differences, both negative and positive. Then most subjects took a step back to describe the large overall negative or positive changes in the budget between the two years. Finally some of the participants pointed out the more subtle patterns (e.g. a node getting a big increase while all the sibling nodes are being cut). Those who did not spontaneously find those patterns were able to find them when asked to look for them.

To explore further people's comprehension of relative changes, participants were asked to explain how they thought the relative percent change was calculated. While, all of them had been able to read and interpret the absolute and relative differences correctly, more than half of the participants struggled to explain correctly how it was calculated which confirms how complex the task of comparison can be. Finally they were asked to review the operation and labeling of the controls and suggest improvements.

After the test participants were asked three questions about the usefulness of TreeVersity:

q1: How useful do you think TreeVersity is to detect differences in the budget.

q2: How effective do you think the colored Bullets are to codify the changes.

q3: How useful do you find the three interconnected views to understand the changes.

The answers, were recorded on a 7 point Likert scale, where 1 was "Not useful" and 7 "Very useful". The results shown Figure 3.11, suggest that users found TreeVersity useful for the task of comparing the Budgets.

3.5 Implementation and Design

TreeVersity was developed using Java 1.6 and the widely-used Prefuse [48] visualization toolkit using a machine with an Intel Core i7 processor with 4GB of RAM memory using GNU/Linux, but has been successfully tested in computers with smaller specifi-

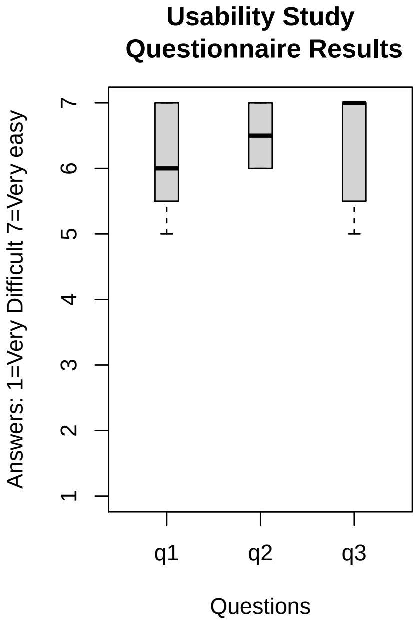


Figure 3.11: Usability Study questionnaire and MySocialTree user survey responses.

cations and with different operative systems (Mac OSX, and Windows 7). In terms of data size, TreeVersity was tested with good response times (no noticeable visual delays) comparing two trees of up to 8,000 nodes each, up to 10 levels deep, with a fan-out of less than 100 children and less than 40% of topological changes. Importers were created for CSV and XML datasets.

Professor Buck-Coleman's Designs

TreeVersity was developed on a close collaboration with Audra Buck-Coleman, a Professor of Design at the University of Maryland. Prof. Buck-Coleman entered the project on it's early stages and helped me explore different visualization glyphs that could be used for representing change in the diffTree. After some exploration we agreed at the concept of the Bullet described in this chapter. Figure 3.12 shows one of Prof. Buck-Coleman's original sketches for the Bullet, which evolved later to a more mature design as the one shown in 3.13. Prof. Buck-Coleman's help was also crucial on the selection of color palettes that were appropriate to the task of showing change, while

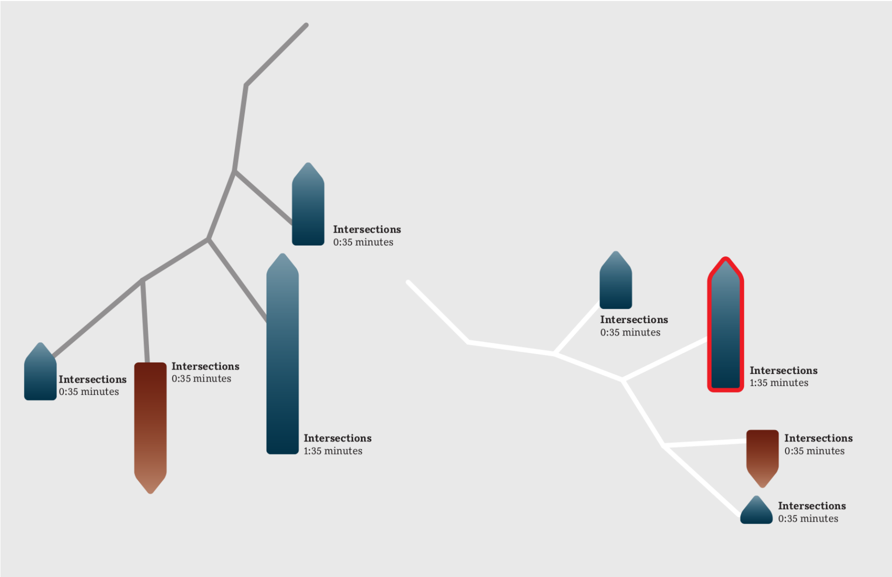


Figure 3.12: Original Bullet Sketches

still be intuitive. Figure 3.14 shows the palettes used in TreeVersity (and later migrated to TreeVersity2). These palettes were manually selected to guarantee that each color was distinct enough from the others.

Not all of Prof. Buck-Coleman's designs could be implemented in TreeVersity. Figures 3.15 and 3.16 show two of those ideas.

3.6 Summary

This chapter described my approaches to the research question: *How to help users find differences between two versions of a tree?* I presented the Bullet (Section 3.2) as a visualization glyph that effectively represents four characteristics of change in tree nodes: *1) direction of change*, *2) actual difference*, *3) relative difference*, and 4) if the *node was created or removed.* Furthermore, I described my implementation of the Bullet on *TreeVersity* (Section 3.1) a tree comparison tool between two versions of a tree that computes and displays the following: 1) *differences in node values* and 2) *nodes that were created or removed*. In addition, I described the user interactions (Section 3.3) that en-

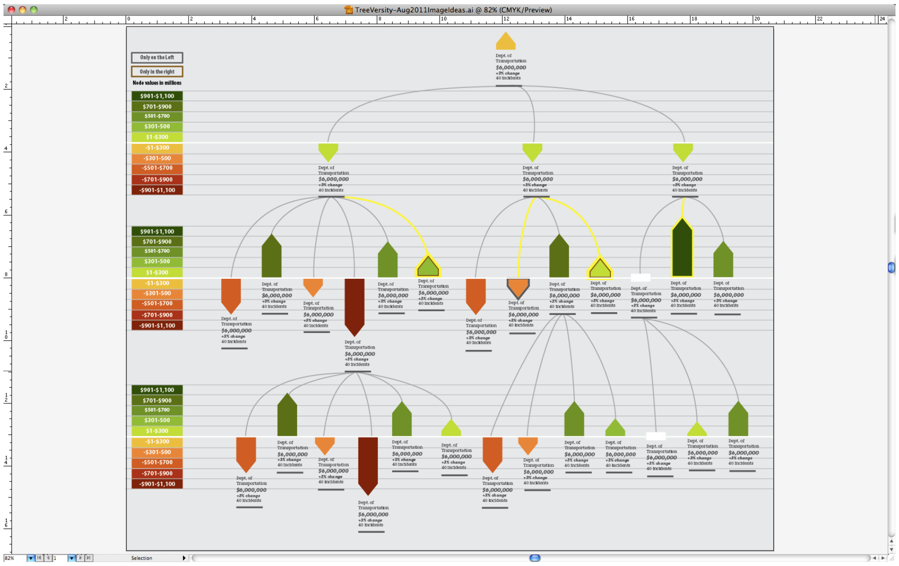


Figure 3.13: More mature bullet sketch

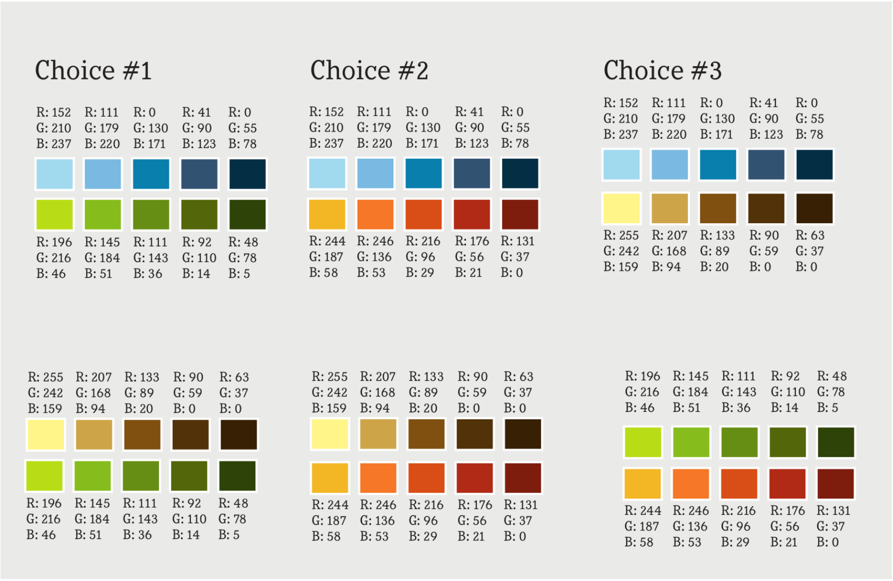


Figure 3.14: TreeVersity Color Palettes

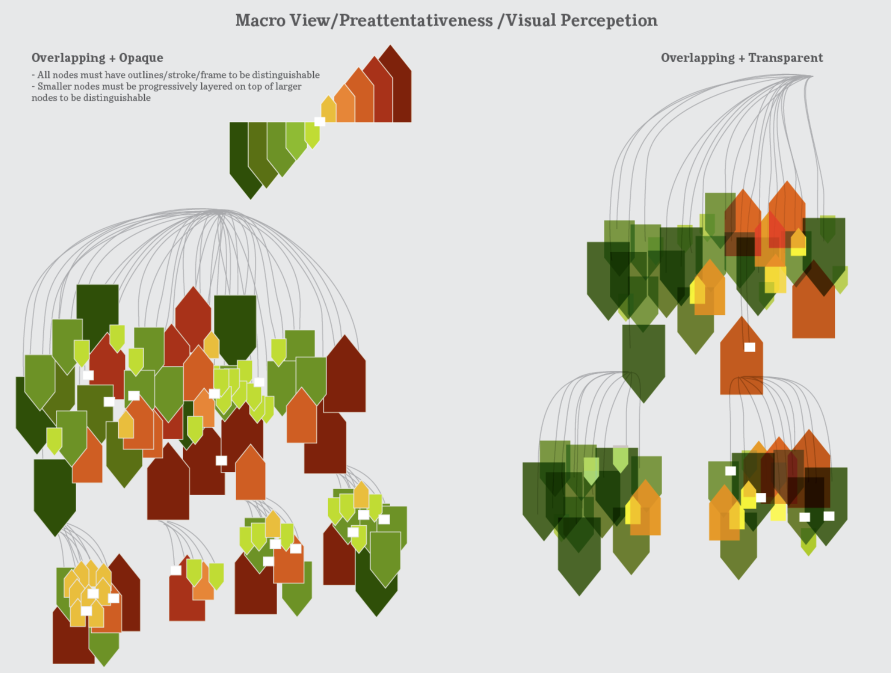

Macro View/Preattentativeness /Visual Percepetion
Overlapping + Opaque
- All nodes must have outlines/stroke/frame to be distinguishable
- Smaller nodes must be progressively layered on top of larger nodes to be distinguishable
Overlapping + Transparent

Figure 3.15: Macro View for the bullets, one of Prof. Buck-Coleman's designs that could not be implemented on TreeVersity. This view was design to help users grasp a better idea of the overview of the changes in the tree, while using less space than on the implemented overview mode.

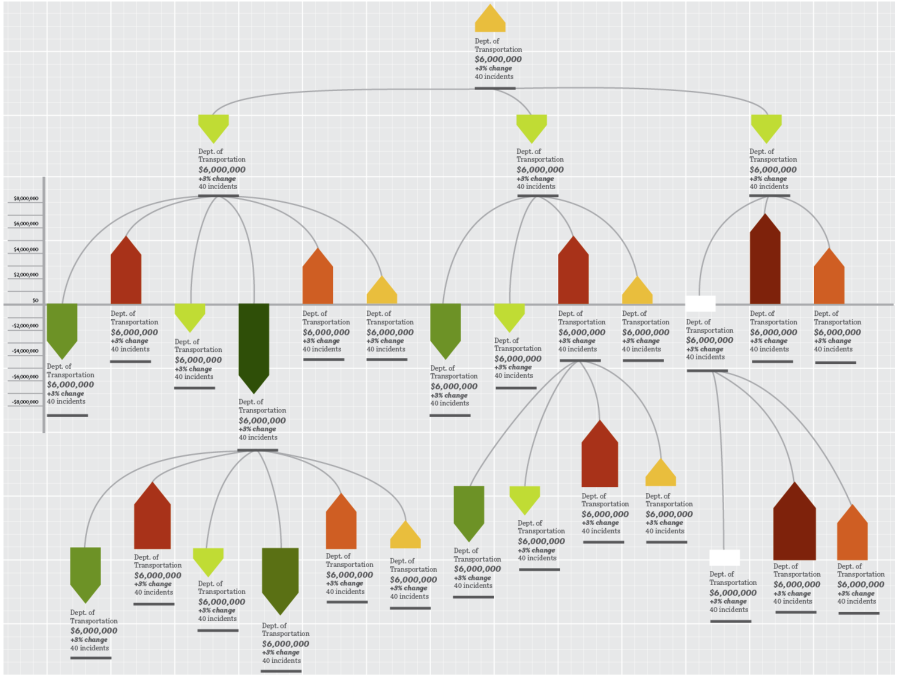

Dept. of Transportation $6,000,000 +3% change 40 incidents
Dept. of Transportation $6,000,000 +3% change 40 incidents
Dept. of Transportation $6,000,000 +3% change 40 incidents
Dept. of Transportation $6,000,000 +3% change 40 incidents
$8,000,000
$6,000,000
$4,000,000
$2,000,000
$0
-$2,000,000
-$4,000,000
-$6,000,000
-$8,000,000
Dept. of Transportation $6,000,000 +3% change 40 incidents
Dept. of Transportation $6,000,000 +3% change 40 incidents
Dept. of Transportation $6,000,000 +3% change 40 incidents
Dept. of Transportation $6,000,000 +3% change 40 incidents
Dept. of Transportation $6,000,000 +3% change 40 incidents
Dept. of Transportation $6,000,000 +3% change 40 incidents
Dept. of Transportation $6,000,000 +3% change 40 incidents
Dept. of Transportation $6,000,000 +3% change 40 incidents
Dept. of Transportation $6,000,000 +3% change 40 incidents
Dept. of Transportation $6,000,000 +3% change 40 incidents
Dept. of Transportation $6,000,000 +3% change 40 incidents
Dept. of Transportation $6,000,000 +3% change 40 incidents
Dept. of Transportation $6,000,000 +3% change 40 incidents
Dept. of Transportation $6,000,000 +3% change 40 incidents
Dept. of Transportation $6,000,000 +3% change 40 incidents
Dept. of Transportation $6,000,000 +3% change 40 incidents
Dept. of Transportation $6,000,000 +3% change 40 incidents
Dept. of Transportation $6,000,000 +3% change 40 incidents
Dept. of Transportation $6,000,000 +3% change 40 incidents
Dept. of Transportation $6,000,000 +3% change 40 incidents
Dept. of Transportation $6,000,000 +3% change 40 incidents
Dept. of Transportation $6,000,000 +3% change 40 incidents
Dept. of Transportation $6,000,000 +3% change 40 incidents
Dept. of Transportation $6,000,000 +3% change 40 incidents
Dept. of Transportation $6,000,000 +3% change 40 incidents
Dept. of Transportation $6,000,000 +3% change 40 incidents

Figure 3.16: Variable width bullets, other of Prof. Buck-Coleman's designs that could not be implemented in TreeVersity. This mode was aimed to represent an extra variable on the bullets by using the widths to show the number of instances. It was discarded, as many other ideas, because of time limitations.

able data exploration and insight discovery on TreeVersity. The results of an exploratory user study demonstrated (Section 3.4) that users were able to understand the Bullet on TreeVersity even without previous training. The chapter concludes with some notes on the implementation and the design guidelines by Professor Audra Buck-Coleman (Section 3.5).

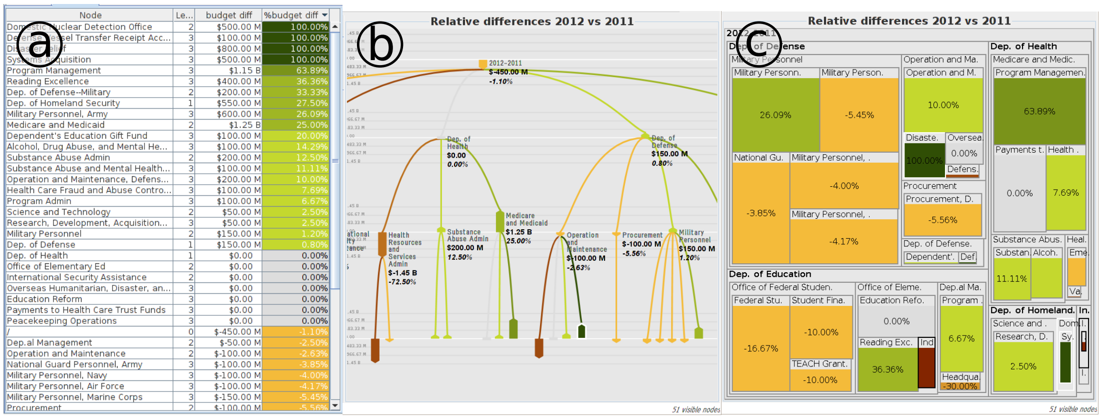
Chapter 4

TreeVersity2 and the StemView: Comparing one tree over multiple points in time with node values and created and removed nodes

Analyzing the changes of a dataset over time is one of the most common and useful techniques of data exploration. More specialized analyses can be made if the datasets have the characteristics of a tree. For example, if users want to explore changes in the U.S. Federal Budget for the past 20 years, they can look at the Budget as a tree by grouping the different budget accounts by their Agencies and Bureaus. Each node can be labeled by their organizational name (e.g. the Agency Department of Treasury) and have information on the amount of dollars spent during a fiscal year. In addition, each node in the tree can be categorized by their Discretionary/Mandatory/Net Interest nature.

From collaborations with domain experts in our case studies, analysts asked the following question: which accounts increase or decrease the most compared to their previous budgets (both in relative and absolute values)? These questions suggest that a visual analytics tool to explore these changes should represent the **direction of change** (to highlight increases and decreases), the **actual amount** of change (dollars in the budget example) and the **relative** change (the percentage of change compared to the previous year). One simple solution for this problem would be to use a table that shows all of the actual and percentages of change for each account in the budget (Figure 4.1(a)).

However, the table would be insufficient if users wanted to perform tasks that maintain the **context of the hierarchy** and involve **inner node's values**, such as finding Bureaus that change significantly inside Agencies that do not change as much or at all. Moreover



Node	Le...	budget diff	%budget diff
Domestic Nuclear Detection Office	2	$500.00 M	100.00%
Defense Vessel Transfer Receipt Acc...	3	$100.00 M	100.00%
Disaster Relief	3	$800.00 M	100.00%
Systems Acquisition	3	$500.00 M	100.00%
Program Management	3	$1.15 B	63.89%
Reading Excellence	3	$400.00 M	36.36%
Dep. of Defense--Military	2	$200.00 M	33.33%
Dep. of Homeland Security	1	$550.00 M	27.50%
Military Personnel, Army	3	$600.00 M	26.09%
Medicare and Medicaid	2	$1.25 B	25.00%
Dependent's Education Gift Fund	3	$100.00 M	20.00%
Alcohol, Drug Abuse, and Mental He...	3	$100.00 M	14.29%
Substance Abuse Admin	2	$200.00 M	12.50%
Substance Abuse and Mental Health...	3	$100.00 M	11.11%
Operation and Maintenance, Defens...	3	$200.00 M	10.00%
Health Care Fraud and Abuse Contro...	3	$100.00 M	7.69%
Program Admin	3	$100.00 M	6.67%
Science and Technology	2	$50.00 M	2.50%
Research, Development, Acquisition...	3	$50.00 M	2.50%
Military Personnel	2	$150.00 M	1.20%
Dep. of Defense	1	$150.00 M	0.80%
Dep. of Health	1	$0.00	0.00%
Office of Elementary Ed	2	$0.00	0.00%
International Security Assistance	2	$0.00	0.00%
Overseas Humanitarian, Disaster, an...	3	$0.00	0.00%
Education Reform	3	$0.00	0.00%
Payments to Health Care Trust Funds	3	$0.00	0.00%
Peacekeeping Operations	3	$0.00	0.00%
/	0	$-450.00 M	-1.10%
Dep.al Management	2	$-50.00 M	-2.50%
Operation and Maintenance	2	$-100.00 M	-2.63%
National Guard Personnel, Army	3	$-100.00 M	-3.85%
Military Personnel, Navy	3	$-100.00 M	-4.00%
Military Personnel, Air Force	3	$-100.00 M	-4.17%
Military Personnel, Marine Corps	3	$-150.00 M	-5.45%
Procurement	2	$-100.00 M	-5.56%

Figure 4.1: Different ways of showing changes between trees: (a) table representation, (b) bullet visualization, (c) treemap representation.

users might want to find nodes that are **created** or **removed** in the tree, like finding all the Bureaus that were created in 2012. Using a node-link based tree visualization with special glyphs for the nodes to illustrate change, such as the Bullet visualization [34, 35] shown in Figure 4.1(b), will allow the exploration of tasks that require the context of the hierarchy while still providing insight about absolute and relative changes.

Despite being easy to understand on small trees, node link representations may become too crowded with even trees of hundreds of nodes which may hide the **starting** and **ending** values of the nodes (e.g. if comparing the 2013 and 2012 budgets, the starting values are the actual budgets for 2012 and the ending values those for 2013) that are required to answer questions such as what is the biggest decreasing Agency in the budget.

A treemap where the color of the nodes represents the change and area of the boxes (the actual values as the one shown in Figure 4.1(c)) can be used for this task (i.e. showing changes on a aggregating tree while also showing actual values) [77, 86, 90]; however treemaps would only beable to display one variable at a time (actual or relative change), would not show negative values, and would hide the values of the inner nodes.

As a result, I present TreeVersity2 which is an interactive data visualization tool that allows the exploration of changes in trees addressing direction of change, actual and relative change, starting and ending values, created and removed nodes, and inner node values while keeping the hierarchy context. TreeVersity2 allows the exploration of change over time in trees using novel interactive data visualizations for exploring changes in the tree between two time points (e.g. two years) coordinated with time based visualizations to explore the time context. Moreover TreeVersity2 includes a reporting tool that guide users through the most significant differences in the tree based on outlier detection algorithms.

I evaluated TreeVersity2 using 12 case studies developed with partners from organizations as diverse as the National Cancer Institute, Food and Drug Administration, Department of Transportation, Office of the Bursar of the University of Maryland, and even eBay. The diversity of the characteristics of the datasets of these case studies showcased the flexibility of TreeVersity2 and suggests that it is a useful tool to support the exploration of changes over time in datasets. Details on the case studies may be found in Chapter 5

4.1 TreeVersity2

TreeVersity2 is a interactive web data visualization tool that allows the exploration of time changing datasets using hierarchies. Users can navigate the time range using controls that allow them to analyze the changes between two time points, while still being aware of the context using time based visualizations. As a example (that is explained in more detail in Section 5.1), TreeVersity2 allowed data analysts from the Food and Drug Administration to compare changes in the number of adverse effects reports generated for a drug between any two years between 2008 and 2012, while keeping the overall context of the tendencies for the whole period.

The time-based visualizations are displayed on TreeVersity2's main interface on the left side where users have the option to switch between traditional timelines to compare actual values or TimeBlocks for comparing differential values. The TimeBlocks depicted in Figure 4.2 use color boxes to represent differential change between sequential years. For example, (shown in Figure 4.2), decreases in a National Cancer Institute's (NCI) lung cancer death index are represented by green boxes (green because they are good) while increases are shown in red. Each horizontal line in the TimeBlocks represents a corresponding attribute (or node in the tree), so in the example for the timelines in the second row top to bottom, there are three TimeBlocks with one for each race.

TreeVersity2 allows users to explore the detailed changes between any two time points. The range of the time points can be setup by the user according to the datasets and can range from seconds (e.g. comparing number of tweets in periods of five seconds) to tens of years (e.g. compare the number of publications in a research field in the last twenty years by decades).

The StemView is a novel tree visualization artifact that enable users to examine the detailed differences in the tree between two time points by highlighting actual and relative changes, positive and negative changes, created and removed nodes, and starting or ending values of the nodes, all while keeping the context of the tree and showing inner nodes changes. The StemView is shown in the center of TreeVersity2's main interface and is explained in more detail in Section 4.2.

Finally, to enable customization and allow exploration, a Control Panel is presented on the right side of the interface. Controls enable users to change the different visual attributes of the visualizations to adjust to their exploration tasks. Users can assign the available variables to the color, height, width, and sorting order of the boxes. Since the visualizations on the StemView represent change, users can select one of five modifiers for each variable:1) actual difference, 2) relative difference, 3) starting value, 4) ending value, or 5) maximum of the starting and ending values. Different combinations of these parameters allow richer explorations, for example in Figure 4.2 NCI analysts were able to explore the actual and relative changes in their lung cancer death rate (represented with the color and height of the StemView boxes respectively) while still analyze the sizes of the populations compared (depicted by the width of the StemView boxes).

The control panel also includes a novel textual reporting tool that helps users navigate significant differences by exploring a textual list of outliers calculated for each pair of

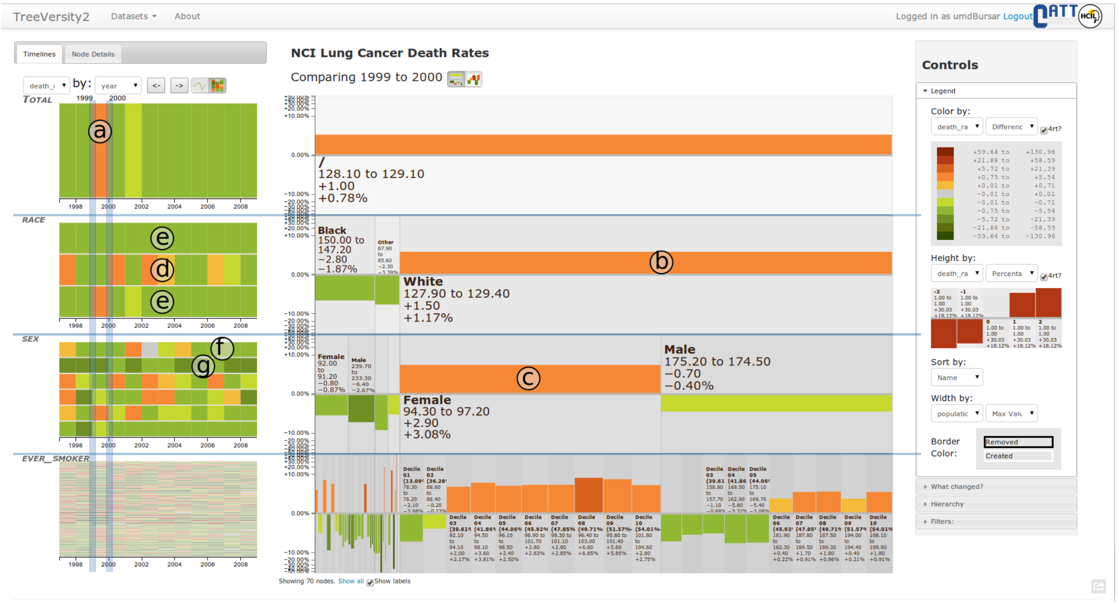
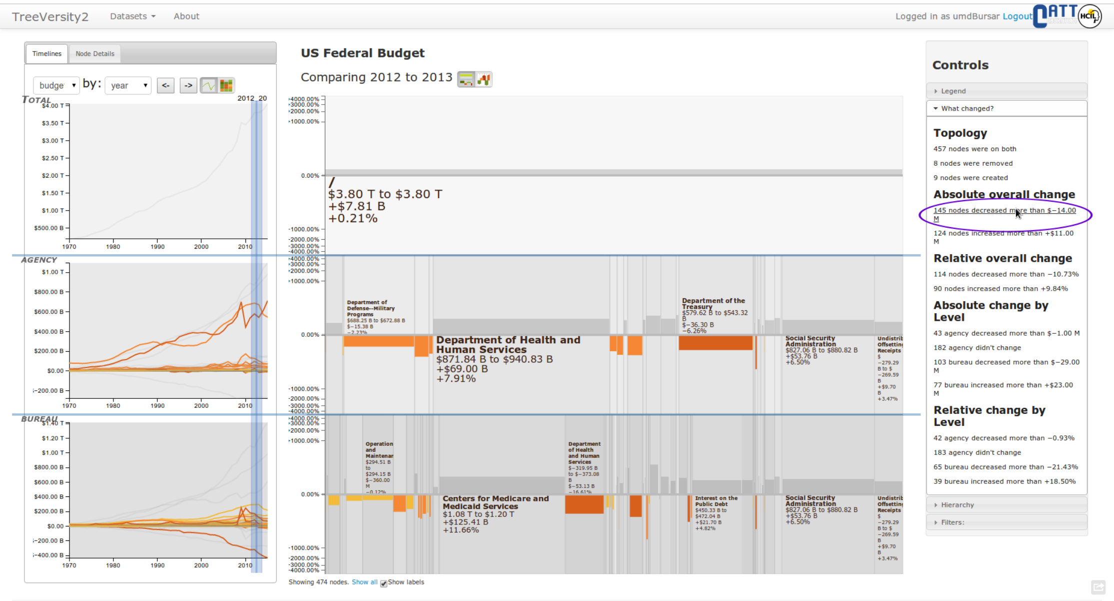


Figure 4.2: National Cancer Institute Lung Cancer related death-rate change between 1999 and 2000 in the US. Color shows relative change in the rate, and height represents relative change in the rate, the width encodes the population counts for each group. The TimeBlocks show that the (a) overall rate increases only in 2000, however (b) the only race increasing is "White", that also happens to be more than 80% of the population. Among whites though, (c) women seem to be the ones contributing the most to the increase.



Controls

Legend

What changed?

Topology

457 nodes were on both

8 nodes were removed

9 nodes were created

Absolute overall change

145 nodes decreased more than $−14.00 M

124 nodes increased more than +$11.00 M

Relative overall change

114 nodes decreased more than −10.73%

90 nodes increased more than +9.84%

Absolute change by Level

43 agency decreased more than $−1.00 M

182 agency didn't change

103 bureau decreased more than $−29.00 M

77 bureau increased more than +$23.00 M

Relative change by Level

42 agency decreased more than −0.93%

183 agency didn't change

65 bureau decreased more than −21.43%

39 bureau increased more than +18.50%

Hierarchy

Filters:

Figure 4.3: The Reporting tool highlighting all the agencies and bureaus in the US Federal Budget that decrease more than $14 million dollars. Users can filter down to only those accounts by clicking on the corresponding line of text in the reporting tool.

compared time points. For instance, Figure 4.3 shows how the analysts at the Office of the Management of the Budget (OMB) can identify all of the accounts decreasing more than $14 million dollars in the US Federal Budget between 2012 and 2013, all while keeping the context. This reporting tool is described in more detail in Section 4.4. Lastly, users can apply specific range filters on each one of the characteristics of change. For example, users can explore all of the changes in all accounts in the US Federal Budget that have a budget higher than $10 million dollars or all accounts increasing or decreasing more than $1 million dollars. Smooth animations and transitions allow users to keep track of the changes in the tree. When filtering, the nodes not matching the criteria are removed and the filtered nodes are animated to occupy all of the available space.

TreeVersity2 allows the exploration of change over time in datasets using hierarchies. These hierarchies can be either **fixed** when there is an inherent parent-to-child relationship (e.g. the Agency->Bureau classification in the U.S. Federal Budget where Bureau-

>Agency does not make sense), **dynamic** when the hierarchy is constructed by grouping rows by their attributes as defined in the original treemap paper [77] (e.g. Census population group by gender, race and age range), or **mixed** where some levels of the hierarchy are fixed and some dynamic (e.g. grouping the U.S. Federal Budget by Discretionary/-Mandatory Accounts that are dynamic and then by Agency/Bureau that are fixed). On the other hand, each one of these hierarchies can be **aggregated** if the values for the parent node are calculated as a function of the values of the children (e.g. adding up the values) or **non aggregating** if the values of the parent nodes are independent from the values of the children (e.g. The FDA's hierarchy of adverse effects of a drug presented in Figure 5.2 where the values of the parent nodes are not a function of the values of the children).

4.2 The StemView

The StemView is a novel visualization that represents five characteristics of change in tree nodes: direction of change, absolute change, relative change, starting or ending values, and created and removed nodes (Figure 4.4). It uses an area filling representation based on icicle trees [55] where the levels of the hierarchy are distributed vertically in equally sized rows. Figure 4.5 shows an example StemView constructed for the US Federal Budget between 2008 and 2009, aggregating the budget accounts by their pertinence to the budget (On or Off budget) and by their Budget Enforcement Act Category (BEA, that determines if they are Discretionary, Mandatory or Net Interest). The vertical space available is distributed equally among the levels, and inside each level, the horizontal space is distributed among the nodes represented as boxes according to their respective ending budget. Figure 4.5(a) shows this first step, which is basically an icicle-tree showing the budgets of each node for 2009. The StemView builds on top of the icicle to show the actual and relative changes of each node. For this purpose, it splits each level vertically using a horizontal line that will represent zero change, as shown in Figure 4.5(b). From that zero line, a sub-box is drawn with the same width of the node's containing

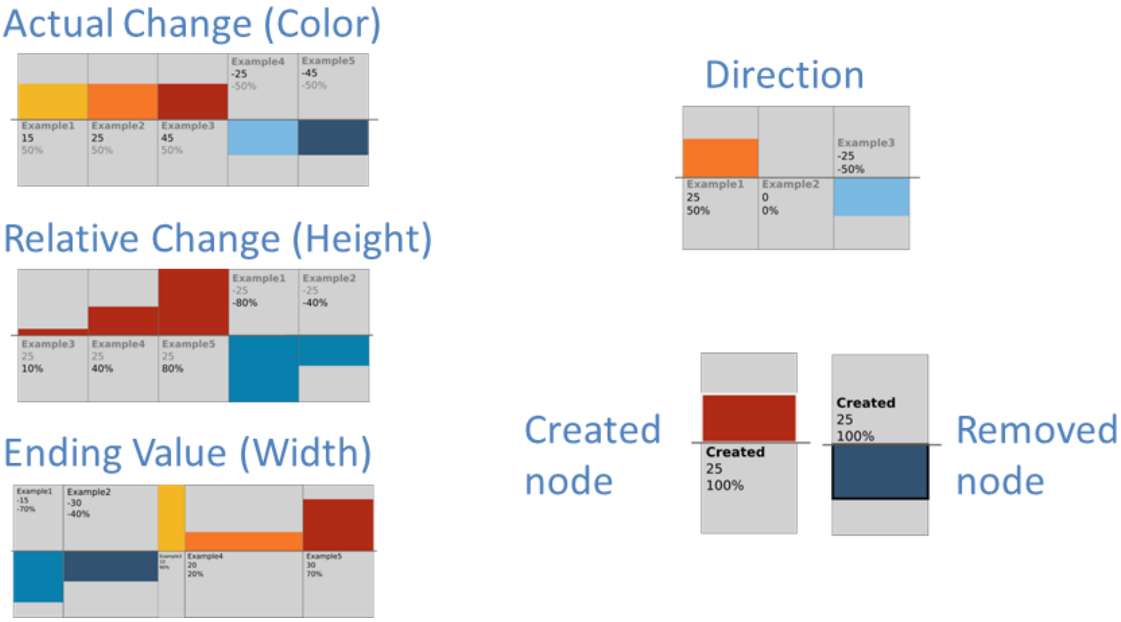


Figure 4.4: The StemView represents five characteristics of change in tree nodes in one visualization.

box but with a height relative to the relative change of the node (e.g. +17.94% for the overall budget). The sub-boxes go upward from the horizontal line for increasing nodes and downward for decreasing nodes (Figure 4.5(c)). Figure 4.5(d) shows the final step where the sub-boxes are colored using the actual amount of change of each node (e.g. +$535.12 billion dollars) using two color scales. These scales are usually green for increasing values and yellows-to-reds for decreasing values; however, this parameter may be customized to show other color schemes for special purposes as shown in the Section 5.1. Finally, the StemView uses white borders around the sub-box to represent created nodes and black borders for the deleted ones. Each one of the characteristics of the StemView, height, width, color of the boxes, and the order in which they are distributed among their parents, can be assigned to different variables of the dataset and their modifiers (starting value, ending value, actual difference or relative difference). Figures 4.6, 4.7, 4.8 and 4.9 illustrate the default parameters of the StemView.

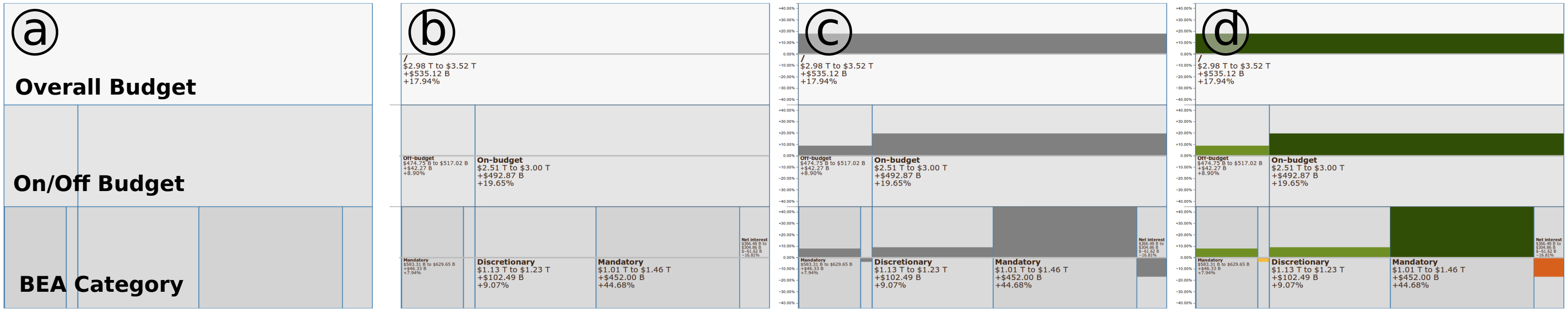


Figure 4.5: Steps for the StemView construction: (a) First an icicle tree for the ending values is used as the base of the visualization, (b) then inside each level a horizon line is drawn representing no change. (c) Sub-boxes with height corresponding to the relative change are drawn inside each node. (d) Finally the nodes are colored using the actual amount of change.

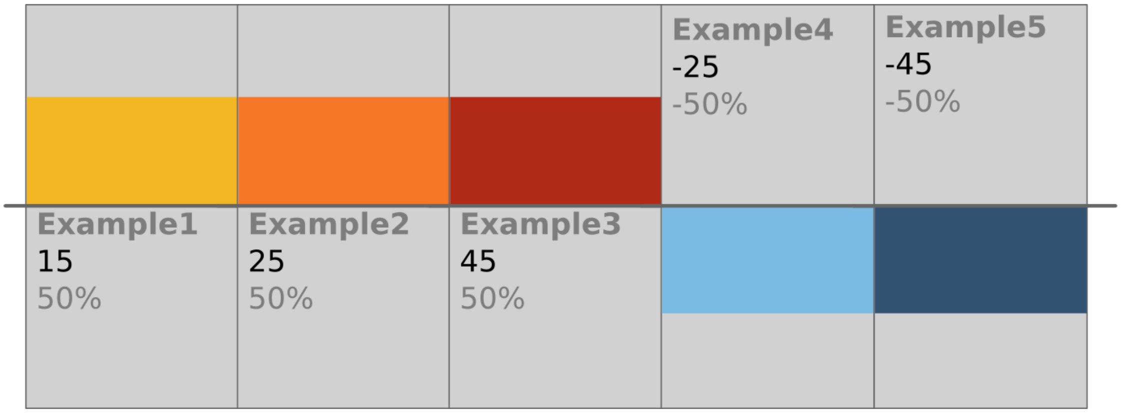


Figure 4.6: The StemView: color represent actual differences, as in dollars for the U.S. Budget

The following is a list of three examples of how the StemView compares to the Bullet, drawn from the informal comments given by the domain experts that participated in the thirteen case studies conducted in this dissertation (described in Chapter 5):

- The Bullet is better at showing the hierarchy in a more intuitive way, especially for small trees. Figure 4.10 shows one small tree (less than 10 nodes) represented with the Bullet and the StemView. The tree was created by grouping the U.S. Federal Budget Accounts into its type of spending (BEA Category: Discretionary, Mandatory or Net Interest) and if it is on/off budget. The bullet does not show the actual values of the accounts, and because of this it fails to reveal information, such as that the *Mandatory* accounts are about three times bigger than the *Net Interest* accounts, which can be easily grasped with the StemView.

- The StemView scales better than the Bullet on bigger trees (more than 20 nodes). Figure 4.11 shows all the U.S. Federal Budget Agencies grouped again by its type of spending. Despite this tree being relatively small (198 nodes), the Bullet is so crowded that labels needed to be removed to avoid overdrawing. On the right side of the Figure, the StemView scales better to show the change of the biggest agencies such as *Social Security Administration*, *Department of Defense* and *Department of Health.* Labels are placed whenever there is enough space

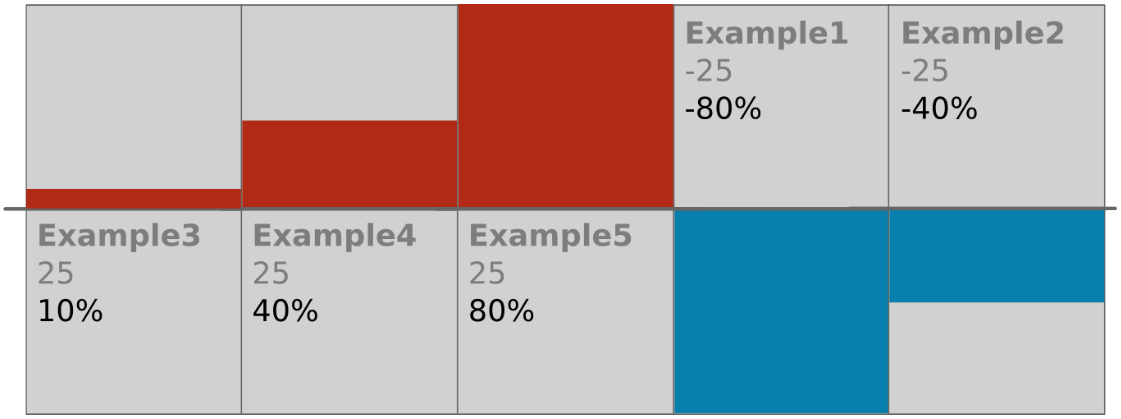


Figure 4.7: The StemView: height represent percentage of change

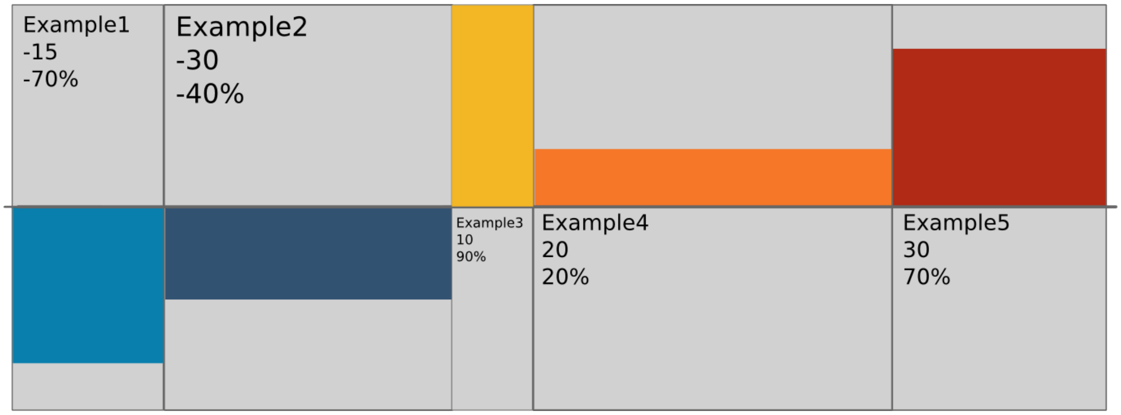


Figure 4.8: The StemView: width represents the ending value

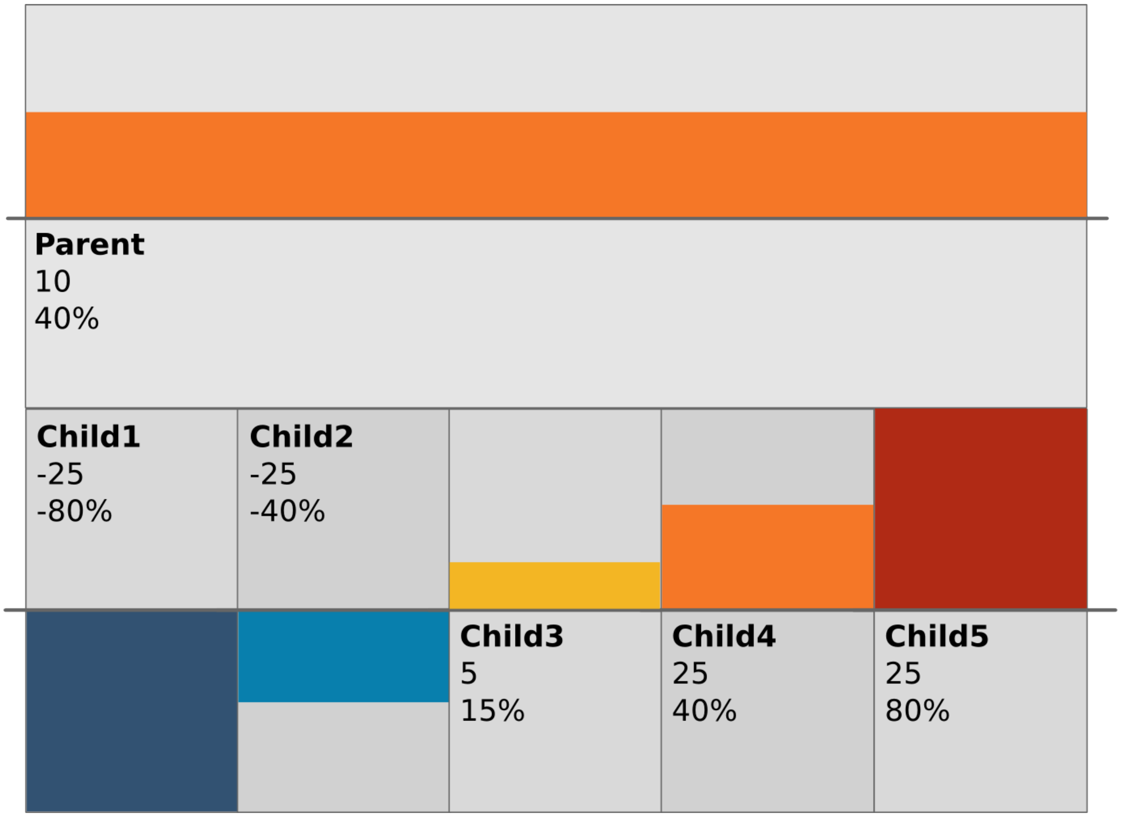


Figure 4.9: The StemView: the hierarchy is represented in the same way Icicle trees work, children are layered out one level below their parents.

available, so users can still read some details.

- The StemView shows one more characteristic of change than the Bullet, that is assigned by default to the ending value of the node. This allows analysts to recognize the most significant elements in a dataset, such as the biggest agencies in the U.S. Federal Budget. Figure 4.12 show the change in the US Federal Agencies between 2010 and 2011 grouped by the type of spending. The Figure demonstrate how the StemView shows all the characteristics of change of the Bullet and also provides information on the sizes of the Agencies.

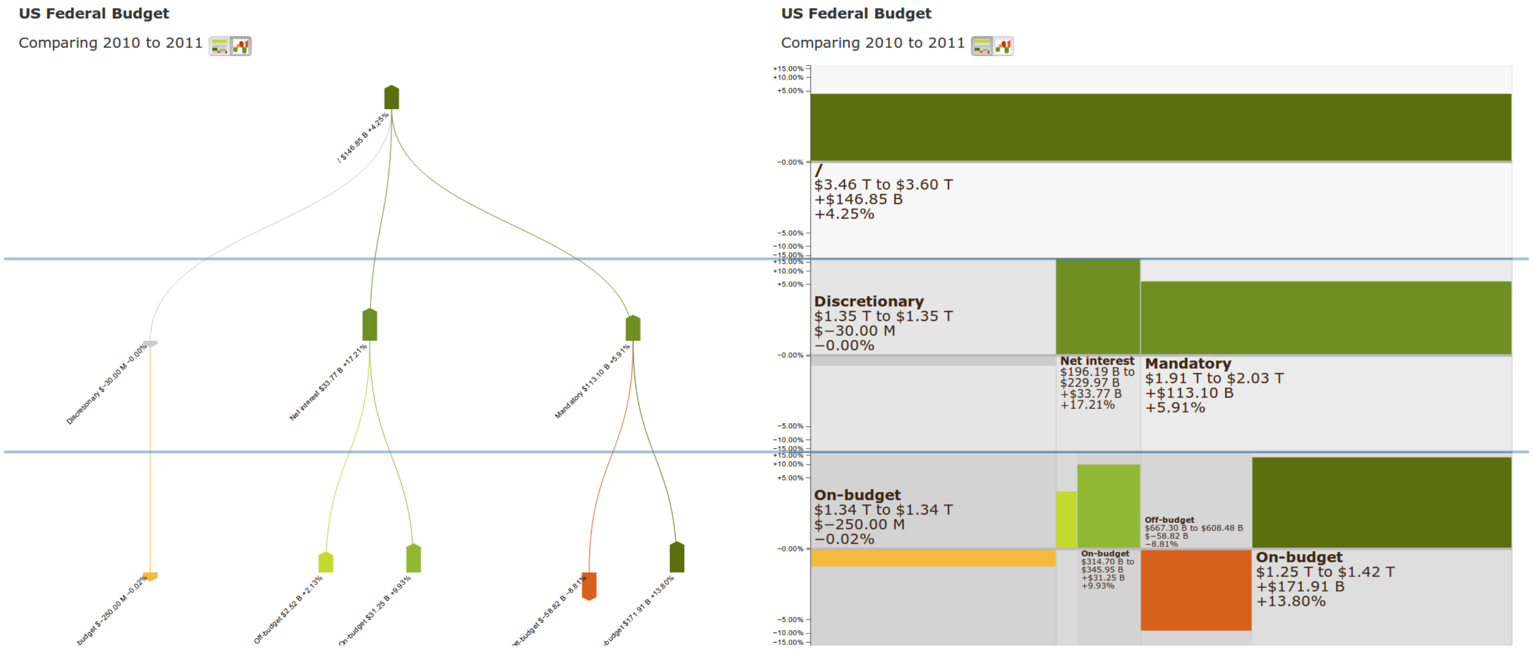


Figure 4.10: Bullet vs StemView: US Federal Budget grouped by Mandatory/Discretionary/Net Interest (BEA Category) and then by on/off-budget. On small trees like these the bullet tends to be easier to understand (for the untrained eye) than the StemView. However, because the bullet does not show actual values, it fails to represent that the Mandatory part of the Budget is about three times bigger than the Net Interest.

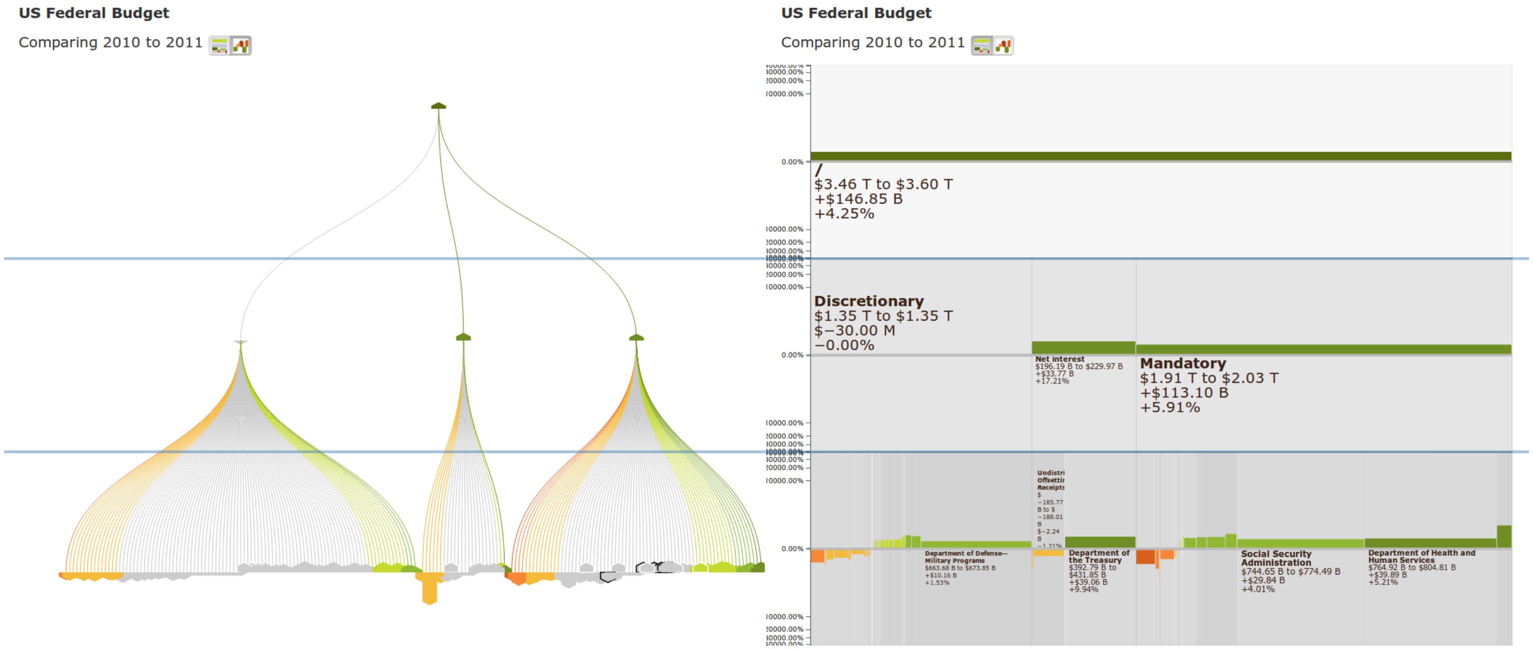


Figure 4.11: Bullet vs StemView: U.S. Federal Budget grouped by BEA Category and then by Agency. On bigger trees the bullet helps finding outliers, but it is very difficult to label and does not emphasize the big agencies in the budget, such as the Social Security Administration, the Deparment of Defense and the Department of Health.

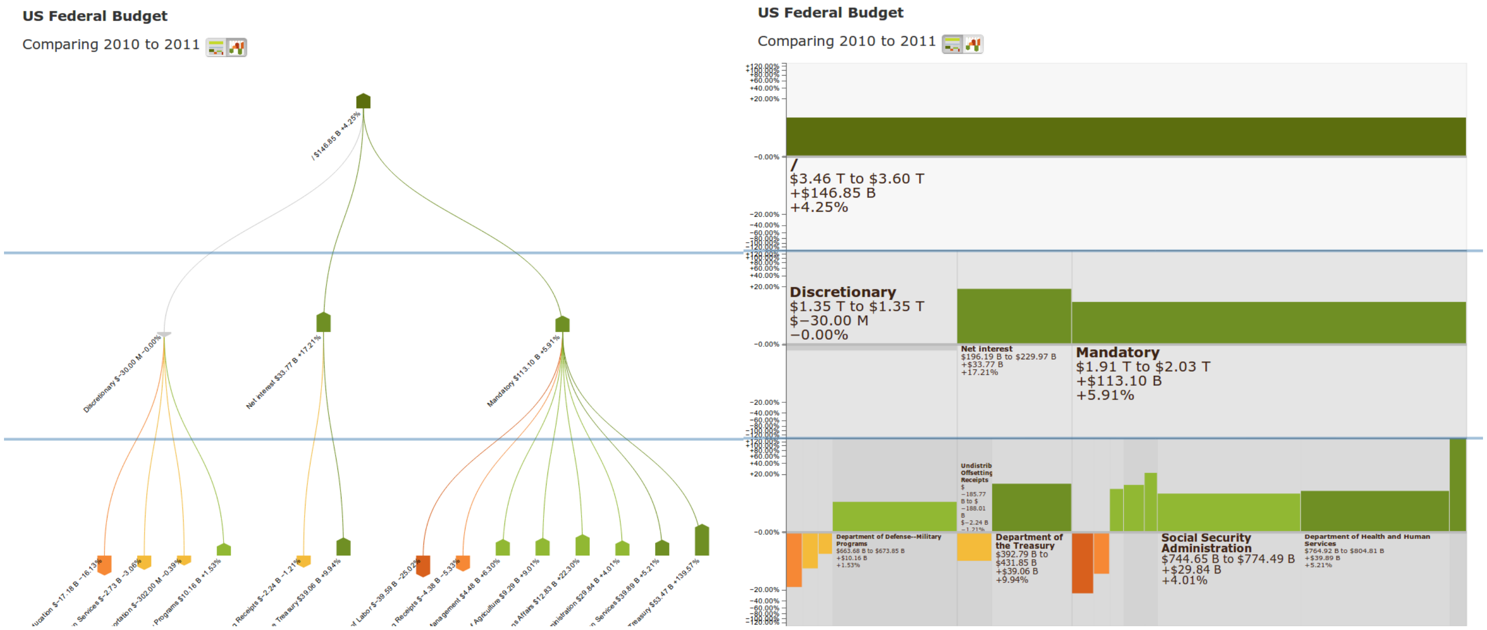


Figure 4.12: Bullet vs StemView: U.S. Federal Budget grouped by BEA Category and then by Agency, filtering by the biggest agencies. The bullet behaves better on filtered results that reduce the number of bullets on the screen and allow the placement of labels. However, the bullet fails to show the actual size of the agencies

4.3 Implementation

TreeVersity2 was designed to allow fluent interactions with datasets in the order of hundred of thousands of records which generate trees with thousands of nodes. This is achieved due to the way it distributes the workload between the client and server applications. When the browser sends a petition to the TreeVersity2 server (Figure 4.13(1) and (2)), it processes it using an application written with Python using the Django application server and then accesses (Figure 4.13(3)) a PostgreSQL database that hosts the full extent of the datasets. The returned value of the SQL query (Figure 4.13(4)) is a preprocessed data structure that is significantly smaller than the full database and contains the information necessary to build the tree according to the parameters sent by the user (encoded in the URL). This data structure is then sent back to the browser ((Figure 4.13(5)) where a JavaScript application processes it using a Crossfilter library ((Figure 4.13(6)) then draws all the visualizations using the D^3 [11] visualization library, and updates the reporting tool ((Figure 4.13(7)).

This section describes some details of TreeVersity2 implementation.

4.3.1 Back-end

On the server side, TreeVersity2 has two main components: 1) an application built on Python using Django and 2) a PostgreSQL database. The Django application has the main objective of answering requests from the client, parsing URL parameters, and generating dynamic SQL scripts to be issued to the Database. This application generates two types of scripts, one for *aggregated trees* and the other for *non-aggregated trees*.

The SQL generated for aggregating trees uses a GROUP BY clauses to aggregate values on the interior nodes of the tree using an aggregating function, such as average, or summation. Figure 4.14 shows a SQL query generated for the US Federal Budget tree grouping by Agency and Bureau, which is an aggregated tree (the budget of the agencies is the sum of the budgets of the bureaus).

For non-aggregated trees, the back-end stores the pre-calculated values of the inte-

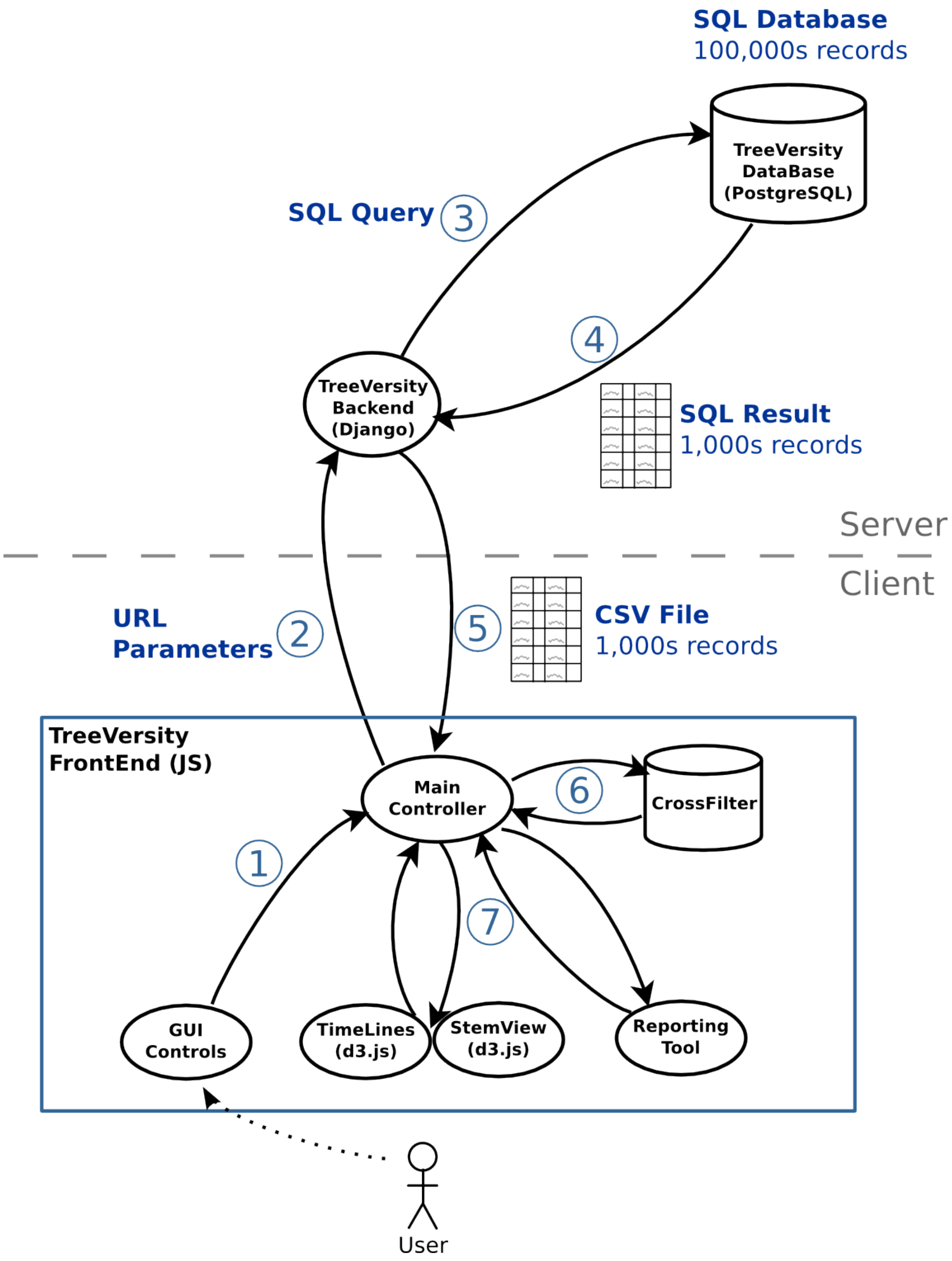


Figure 4.13: TreeVersity2 architecture. The tool divides the processing between the server and client. When the user starts an exploration, (1) TreeVersity2 queries the back-end (2) with the parameters of the exploration encoded in the URL. The back-end then generates a dynamic SQL query (3) and sends it to the PostgreSQL database. The database returns the result in SQL format (4) and the back-end creates a CSV file that is passed back to the front-end (5). The front-end loads the CSV file in a CrossFilter (6) and then passes it (7) to the visualizations and the reporting tool. Once the data is loaded into the CrossFilter all of the following interactions which do not change the hierarchy, are processed on the front-end without having to query again the back-end.

```
SELECT *,   nonNullDiff(sum0_0,sum0_1) AS dt0_1_0,
                nonNullPct(sum0_0,sum0_1) AS dRelt0_1_0,
                topol(sum0_0,sum0_1) AS topol0_1_0
 FROM
(SELECT agency,bureau,sum(values.budget) AS sum0_0
 FROM attribs NATURAL JOIN values
 WHERE seq_val>='2012-01-01T00:00:00'::timestamp
  AND seq_val<'2013-01-01T00:00:00'::timestamp
 GROUP BY 1,2) as t0 NATURAL FULL JOIN
(SELECT agency,bureau,sum(values.budget) AS sum0_1
 FROM attribs NATURAL JOIN values
 WHERE seq_val>='2013-01-01T00:00:00'::timestamp
  AND seq_val<'2014-01-01T00:00:00'::timestamp
 GROUP BY 1,2) as t1
```

Figure 4.14: Example SQL query generated for the US Federal Budget dataset grouping by Agencies and Bureaus, which is an aggregated tree (the budget of the Agencies can be calculated as the sum of the budgets of the Bureaus)

rior nodes on the database and queries them directly (i.e. not calculating them on the fly). Figure 4.14 shows an example SQL query generated for the FDA's drug adverse effects tree, which is a non-aggregating tree since the values in the inner nodes cannot be calculated as a function of the leafs.

Importing Scripts

I created a general purpose importing script that facilitates the addition of new datasets to TreeVersity2. Figure 4.16 shows the use of the importing script to load one of the FDA datasets. The script supports aggregated and non-aggregated hierarchies, attributes with one or multiple values, multiple numeric variables, and different date formats.

4.3.2 Front-end

TreeVersity2 front-end was built using well known web development libraries, namely:

- D^3, used as the main visualization library.
- JQuery, used for DOM manipulation.

```
SELECT *,   nonNullDiff(sum0_0,sum0_1) AS dt0_1_0,
                nonNullPct(sum0_0,sum0_1) AS dRelt0_1_0,
                topol(sum0_0,sum0_1) AS topol0_1_0 ,
                nonNullDiff(sum1_0,sum1_1) AS dt1_1_0,
                nonNullPct(sum1_0,sum1_1) AS dRelt1_1_0,
                topol(sum1_0,sum1_1) AS topol1_1_0 ,
                nonNullDiff(sum2_0,sum2_1) AS dt2_1_0,
                nonNullPct(sum2_0,sum2_1) AS dRelt2_1_0,
                topol(sum2_0,sum2_1) AS topol2_1_0 ,
                nonNullDiff(sum3_0,sum3_1) AS dt3_1_0,
                nonNullPct(sum3_0,sum3_1) AS dRelt3_1_0,
                topol(sum3_0,sum3_1) AS topol3_1_0 ,
                nonNullDiff(sum4_0,sum4_1) AS dt4_1_0,
                nonNullPct(sum4_0,sum4_1) AS dRelt4_1_0,
                topol(sum4_0,sum4_1) AS topol4_1_0 ,
                nonNullDiff(sum5_0,sum5_1) AS dt5_1_0,
                nonNullPct(sum5_0,sum5_1) AS dRelt5_1_0,
                topol(sum5_0,sum5_1) AS topol5_1_0 ,
                nonNullDiff(sum6_0,sum6_1) AS dt6_1_0,
                nonNullPct(sum6_0,sum6_1) AS dRelt6_1_0,
                topol(sum6_0,sum6_1) AS topol6_1_0
 FROM
(SELECT soc,hlgt,hlt,pt,values.count AS sum0_0,
        values.eb05 AS sum1_0,values.ebgm AS sum2_0,
        values.eb95 AS sum3_0,values.n AS sum4_0,
        values.dim AS sum5_0,values.e AS sum6_0
 FROM attribs NATURAL JOIN values
 WHERE seq_val>='2011-01-01T00:00:00'::timestamp AND
        seq_val<'2012-01-01T00:00:00'::timestamp
 AND soc IS NOT NULL
 AND hlgt IS NOT NULL
 AND hlt IS NOT NULL
 AND pt IS NOT NULL) as t0 NATURAL FULL JOIN
(SELECT soc,hlgt,hlt,pt,values.count AS sum0_1,
        values.eb05 AS sum1_1,values.ebgm AS sum2_1,
        values.eb95 AS sum3_1,values.n AS sum4_1,
        values.dim AS sum5_1,values.e AS sum6_1
 FROM attribs NATURAL JOIN values
 WHERE seq_val>='2012-01-01T00:00:00'::timestamp AND
        seq_val<'2013-01-01T00:00:00'::timestamp
 AND soc IS NOT NULL
 AND hlgt IS NOT NULL
 AND hlt IS NOT NULL
 AND pt IS NOT NULL) as t1
```

Figure 4.15: Example SQL query generated for the non-aggregating FDA dataset.

```
import Importer
infile_src='./data/fda/fda_6_consolidated.csv'
attribs="ITEM1,ITEM2,SUBSET,P1,P2,PT,HLT,HLGT,SOC,SOC_ABBREV,PRIMARY_SOC_FG".split("
vals=["EB05","EBGM", "EB95", "N", "DIM", "E"]
dbname="tv2_sectorMaps6"
seqAttrib="SUBSET"
seqFormat="%Y"

im=Importer.Importer(dbname, infile_src, attribs, vals, seqAttrib,seqFormat)
im.run()
```

Figure 4.16: TreeVersity2 importing script.

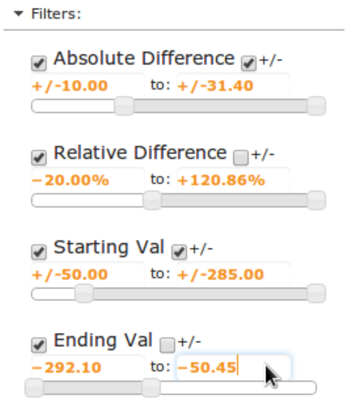
- Bootstrap and JQuery UI, for the UI layout and widgets.
- RequireJS for modularizing the pieces of code.
- LESS, for the use of variables inside CSS.
- CrossFilter, for data manipulation.

The use of this libraries made TreeVersity2 a more standard compliant, and therefore facilitated it's development and deployment. Some other parts of the front-end required special effort or provided special features and are therefore described in the rest of this section.

Local data management CrossFilter

Crossfilter is a wonderful data filtering tool for JavaScript design by Mike Bostock with many contributions from Jason Davies. Crossfilter allows fast data filtering on the browser without querying back the server, which allows TreeVersity2 to offer fast interactions that were not possible if the browser had to be queried on every user move. TreeVersity2 was designed to balance the processing load between the server and client where the server side performs actions on the full dataset using PostgreSQL while the client side uses Crossfilter (on a pre-processed dataset that is one order of magnitude smaller than the original data) to perform the data handling.

TreeVersity2 loads the data obtained from the server on a Crossfilter and then queries it to obtain the values to be displayed according to the user interface. With these values,



▾ Filters:

☑ Absolute Difference ☑ +/-
+/-10.00 to: +/-31.40

☑ Relative Difference □ +/-
−20.00% to: +120.86%

☑ Starting Val ☑ +/-
+/-50.00 to: +/-285.00

☑ Ending Val □ +/-
−292.10 to: −50.45

Figure 4.17: Modified JQuery UI Slider for four modifiers of a variable "Absolute Difference" (range $[-31.40\,,\,-10]\cup[10\,,\,31.40]$), "Relative Difference" (range $[20\%\,,\,120.86\%]$), "Starting Val" (range $[-285\,,\,-50]\cup[50\,,\,285]$), and "Absolute Difference" (range $[-292\,,\,-50.45]$).

TreeVersity2 builds a tree and timelines needed by $D^3$to create the visualizations.

Labeling

Adding meaningful labels to the StemView was as challenging as adding labels to any area based visualization. Different techniques were used to add labels and a dynamic algorithm that draws labels on the StemView boxes whenever the label can fit. The algorithm also checks to see if larger fonts can be used and still fit the StemView box. The resulting experience shows animated fonts that readjust, appear or dissapear to fit the changing sizes of the StemView boxes.

MySlider

A special class was created to extend the JQueryUI slider class to include extra functionalities. First, the slider includes labels with the current values of the range which can be used as text entries by the user to enter the exact number. Second, the slider includes a $+/-$ check-box that switches, enables the bidirectional mode, and allows bidirectional ranges such as $[-20,-10]\cup[10,20]$. Figure 4.17 shows four sliders with different configurations.

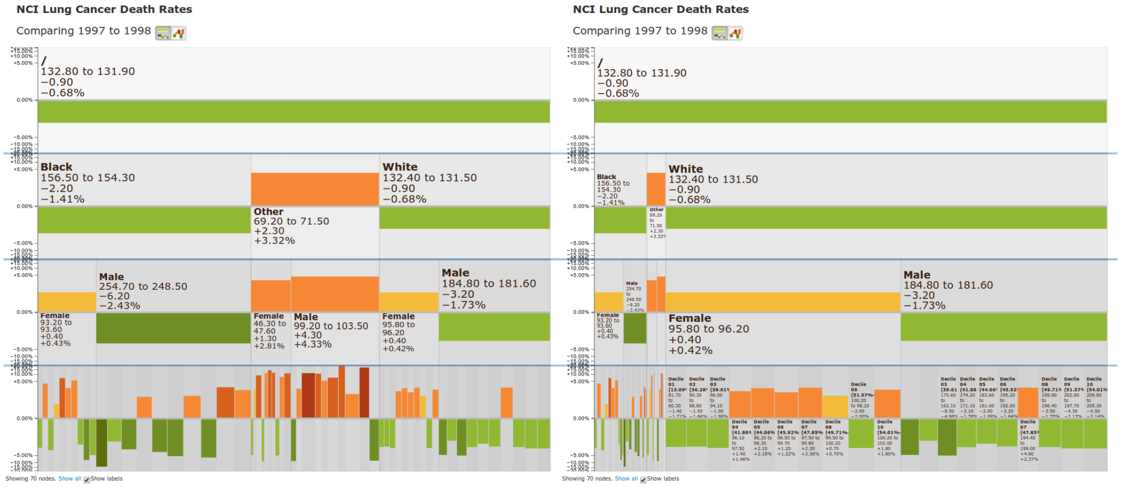


Figure 4.18: TreeVersity Dynamic labels adjust to show text onlywhen there is enough space available, and adjust the font size to make values more readable. The image shows the same StemView but changing the width of the boxes, notice how the leave nodes on the right get labels once there is enough space available. Also notice the changes in the font sizes to adjust the text to fit the boxes.

4.4 Finding and reporting significant differences: The Reporting Tool

We know from the HCIL expertise on development of advanced information visualization tools that users may find them complicated to use. During the development of TreeVersity and TreeVersity2, I collected similar feedback from users. Analysts were comfortable exploring their datasets when I was controlling the tool but some of them felt challenged by the many controls and configurations that TreeVersity offer, and therefore were scared of conducting the explorations by themselves. To address this issue, I developed an interactive, text-based, reporting tool for TreeVersity2 which aids users finding significant changes in their datasets. The reporting tool was praised by users as an easy and fast way of navigating through the differences in TreeVersity2.

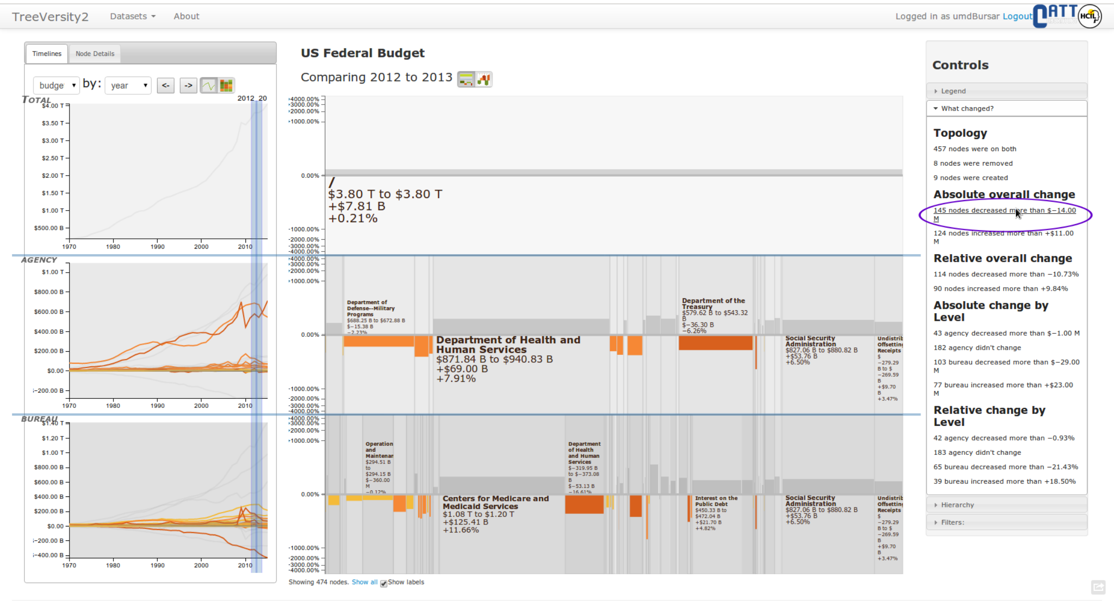
Every time users change the compared time points in an exploration, the reporting tool generates a new textual list with what has changed in the tree which is then grouped by a *reporting metric*. A reporting metric can be any algorithm that returns a list of interesting nodes in the tree. The current version of TreeVersity2 includes five reporting metrics:

- *Topological changes*, which finds the number of nodes that were created, the number removed, and the number that appear in both compared time points.
- *Absolute changes overall*, which finds outliers on the actual changes (i.e. changes in dollars for the US Federal Budget dataset) on the tree node values compared with all the other nodes in the tree.
- *Relative changes overall*, same as the previous one but with the relative change (i.e. percentages).
- *Absolute change by level*, similar to the *absolute changes overall* but compared against the nodes in a certain level only. For example, this metric will find a Bureau that increased more than normal compared to the other Bureaus in the Budget, but that would have been hidden by the even bigger changes on the agencies.
- *Relative changes overall,* same as before but for the relative change.

The reporting tool was designed to be extensible, so adding new metrics requires a simple modification in the source code. The current version is a proof of concept, but I envision a more developed implementation has many more metrics and includes controls to allow users to select which metrics to use and even add their own.

User Interactions

Every time users change the compared time points in the dataset, the reporting tool is recalculated. The results are presented in a textual list of reporting items that provide a human readable description of each group of interesting nodes, such as *145 nodes decreased more than $-14 M*. When users hover over one reporting item, the corresponding nodes are highlighted on the StemView and timelines as shown on Figure 4.19. This allows users to identify nodes while keeping the context of the tree. If the user wants to



Controls

Legend

What changed?

Topology

457 nodes were on both

8 nodes were removed

9 nodes were created

Absolute overall change

145 nodes decreased more than $−14.00 M

124 nodes increased more than +$11.00 M

Relative overall change

114 nodes decreased more than −10.73%

90 nodes increased more than +9.84%

Absolute change by Level

43 agency decreased more than $−1.00 M

182 agency didn't change

103 bureau decreased more than $−29.00 M

77 bureau increased more than +$23.00 M

Relative change by Level

42 agency decreased more than −0.93%

183 agency didn't change

65 bureau decreased more than −21.43%

39 bureau increased more than +18.50%

Hierarchy

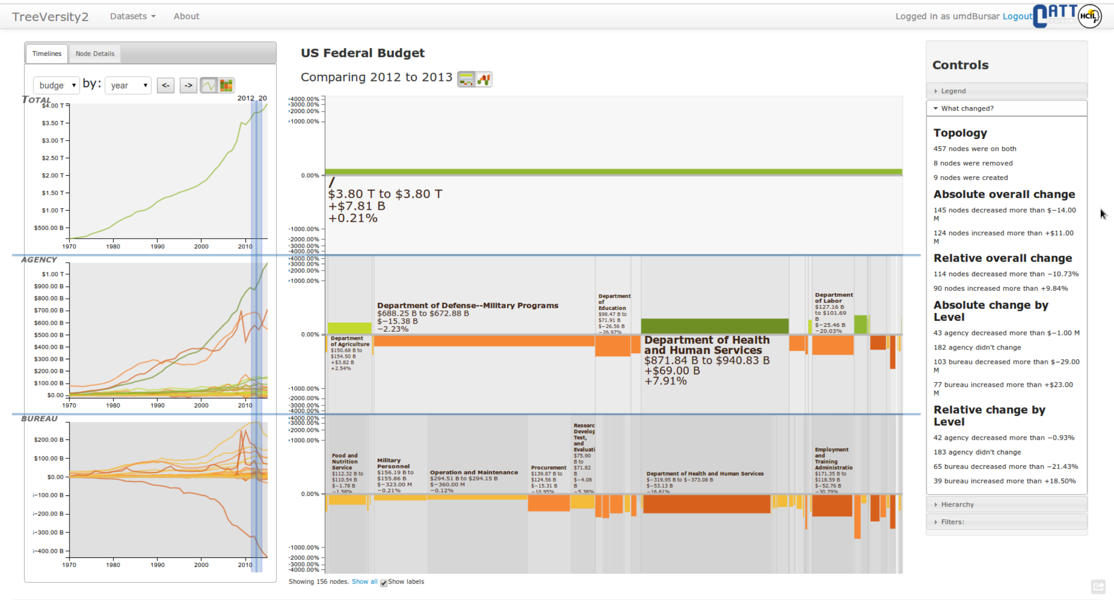
Filters:

Figure 4.19: The Reporting tool highlighting all the agencies and bureaus in the US Federal Budget that decrease more than $14 million dollars. Users can filter down to only those accounts by clicking on the corresponding line of text in the reporting tool.

explore further, they can click on the reporting item, which will filter with smooth animation categories of change (e.g. change by topology or change by level). Each item in the list describes a group of nodes with a node count and why they are interesting (e.g. *145 nodes decreased more than $-14 M*). Users can hover over an item in the report to highlight the corresponding nodes in the StemView and time visualizations as shown in Figure 4.19. Users who want to explore further can click over the report item to filter out all the non matching nodes, leaving only the nodes referenced by the report item as demonstrated in Figure 4.20.

4.5 Summary

This chapter presented my solutions to the research question: *How to help users find changes over time in datasets that can be categorized as trees?* I described the *StemView* (Section 4.2) a tree visualization technique that effectively represents five characteristics



Controls

- Legend
- What changed?

Topology

457 nodes were on both

8 nodes were removed

9 nodes were created

Absolute overall change

145 nodes decreased more than $−14.00 M

124 nodes increased more than +$11.00 M

Relative overall change

114 nodes decreased more than −10.73%

90 nodes increased more than +9.84%

Absolute change by Level

43 agency decreased more than $−1.00 M

182 agency didn't change

103 bureau decreased more than $−29.00 M

77 bureau increased more than +$23.00 M

Relative change by Level

42 agency decreased more than −0.93%

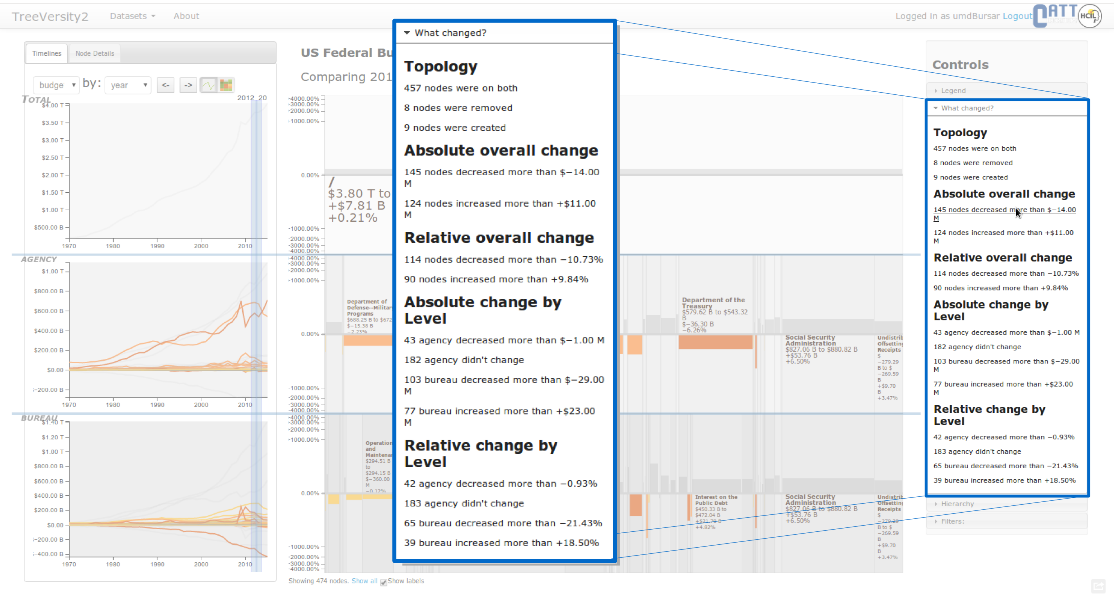
183 agency didn't change

65 bureau decreased more than −21.43%

39 bureau increased more than +18.50%

- Hierarchy
- Filters:

Figure 4.20: Same as in Figure 4.19 after the user clicked on the reporting item. The reporting tool then filters the views to show only the "145 nodes decreased more than $14 million dollars". The remaining nodes on the StemView that are increasing (i.e. Department of Health and Human Services) are shown to maintain the hierarchy.



What changed?

Topology

457 nodes were on both

8 nodes were removed

9 nodes were created

Absolute overall change

145 nodes decreased more than $−14.00 M

124 nodes increased more than +$11.00 M

Relative overall change

114 nodes decreased more than −10.73%

90 nodes increased more than +9.84%

Absolute change by Level

43 agency decreased more than $−1.00 M

182 agency didn't change

103 bureau decreased more than $−29.00 M

77 bureau increased more than +$23.00 M

Relative change by Level

42 agency decreased more than −0.93%

183 agency didn't change

65 bureau decreased more than −21.43%

39 bureau increased more than +18.50%

Figure 4.21: Zoomed in view of the reporting tool

of change on trees: *1) direction of change*, *2) actual difference*, *3) relative difference*, 4) if the *node was created or removed* and 5) *starting* or *ending values.* I implemented the StemView on *TreeVersity2* (Section 4.1) a web based information visualization tool that allows exploration of changes in datasets over time using hierarchies. The architecture of TreeVersity2 was described (Section 4.3) with details on how it balances the processing load between the server and the client components to offer responsive interactions to the users. Finally I outlined the *Reporting Tool* (Section 4.4), a feature of TreeVersity2 that helps users navigate significant data changes using human readable descriptions and coordinated interactions.

Chapter 5

Case Studies

Many interactive information visualization systems have been evaluated using methods that are restricted to short in-laboratory studies. Those methods are usually targeted to measure the speed and accuracy of subjects performing artificial exploration tasks on the tools or explore the usability characteristics of a certain component of the tool. Data exploration tools such as TreeVersity and Treeversity2 are composed of many coordinated components that need to be used as a whole to obtain its real benefit. Moreover, the exploration tasks that they allow require deep understanding of the analyzed datasets and real interest on extracting insights from them which is very difficult, if not impossible, to simulate in one to two hour experiments in a laboratory. Because of this, there is a movement by visualization researchers towards alternative evaluation methods [48, 59, 65, 68, 73, 76, 85]. Among these, Shneiderman and Plaisant have advocated a Multi-dimensional In-depth Long-term Case Study (MILCS) method to study how visualization systems are used in real systems [78]. This methodology encourages the use of the system outside laboratory settings, working with users as partners rather than as subjects, and making the success of the user as a measure of the success of the system. During this lengthy process (a few weeks to several months), bugs are fixed, adaptations are made for specific user needs, and detailed logs are taken. Users are encouraged to use the systems on their own or explore their datasets with the aid of a researcher interacting with the system (i.e. chauffeur mode).

During this research, a total of 13 case studies where developed that demonstrate the flexibility and usefulness of the TreeVersity system (in both versions). As shown in Table 5.1, the case studies represent a broad spectrum of the tree comparison problem domain. In each study there are one or more partners who have a deep understanding of their datasets and are able to confirm expected changes, identify unexpected behaviors, and find anomalies in their data using one of the two versions of TreeVersity. This chapter presents a detailed description of five of these case studies, including the motivation behind it, how it was developed and what was learned from it (both by the user and by myself). Out of the five studies presented, four were conducted using TreeVersity2 and one using a combination of both tools (TreeVersity and Treeversity2). Summaries of the remaining case studies can be found in Annex B.

The chapter concludes with a discussion of the results of the exit questionnaire conducted with the domain experts that assesses their experiences using the TreeVersity tools. The complete questionnaire can submitted to the experts and can be found in Annex C.

Organization	Case Study	MILCS Stage	Driving Mode	TreeVersity Version	Data Size	Time Points	Example Tree Size	Number Attribs.	Number Vars.	Type of Tree	Tree Comparison Type
DOT	**Airlines Budgets**	Early	Chauffeur	1	216	N/A	187 (2 Levels)	3	1	Dynamic	Type 3: aggregated + different topology
OMB	**US. Federal Budget**	Early	Chauffeur	1 & 2	4,845	56	1,393 (4 Levels)	7	1	Mixed	Type 3: aggregated + different topology
DOT	**TRB Publications**	Early	Chauffeur	1 & 2	52,135	8,012	674 (2 Levels)	20	1	Dynamic	Type 3: aggregated + different topology
DOT	**Nat. Trans. Library Publications**	Early	Chauffeur	1 & 2	38,351	374	294 (3 Levels)	10	1	Dynamic	Type 3: aggregated + different topology
DOT	**Passengers flying in the US**	Early	Chauffeur	1 & 2	65,534	162	4,194 (3 Levels)	4	1	Mixed	Type 3: aggregated + different topology
NCI	**National Cancer Institute**	Early	Chauffeur	2	1,716	13	101 (3 Levels)	3	3	Dynamic	Type 2: non aggregated + same topology
FDA	**FDA Drug Adverse Effects**	Mature	Chauffeur	2	2,964	5	1,614 (4 Levels)	4	4	Fixed	Type 4: non aggregated + different topology
UMD	**UMD Budget**	Early	Chauffeur	2	16,332	5	1,296 (3 levels)	6	1	Mixed	Type 3: aggregated + different topology
UMD Bursar	**UMD Students Information**	Mature	Chauffeur	2	227,158	5	715 (5 Levels)	219	3	Mixed	Type 3: aggregated + different topology
eBay	**eBay Product Sales Data**	Early	User-driven	2	63,098	4	5,443 (4 Levels)	6	2	Fixed	Type 1: aggregated + same topology
CATT Lab	**Transportation Bottleneck Data**	Early	User-driven	2	96,205	24	286 (3 Levels)	7	4	Mixed	Type 3: aggregated + different topology
IDB	**Imports and Exports in the Americas**	Early	User-driven	2	119,741	19	3,766 (4 Levels)	5	1	Dynamic	Type 3: aggregated + different topology
DUTO	**Blind Students in Colombia**	Mature	User-driven	2	33,802	4	1,098 (3 Levels)	21	1	Mixed	Type 3: aggregated + different topology

Table 5.1: Case Studies Summary

5.1 FDA Adverse Drug Effects

5.1.1 Case Study Sheet

Partner: Anna Szarfman, Medical Officer at FDA

Organization: Food and Drug Administration (FDA)

MILCS level: Mature, Chauffeur

Duration: July 2012 - April 2013 (10 Months, 5 meetings)

System used: TreeVersity2

Data: EBGM values, an index representing the number of adverse effects reported for a drug compared to the expected number of reports

Number of rows: 2,964

Number of time points: 5 (years)

Example tree size: 1,614 (4 Levels)

Number of numeric variables: 6 (EB05, EB95, EBGM, n, dim, e)

Number of Attributes: 4 (SOC, HLGT, HLT, PT)

Type of Hierarchy: Fixed

Type of comparison: Type 2: non aggregating + same topology

New Features: support for multiple variables in the visualizations at the same time, support for fixed non aggregating hierarchies, support for different color palettes, support for localized navigation

Limitations: Problems authorizing Chrome at the FDA, need to show confidence intervals.

URL: `https://treeversity.cattlab.umd.edu/cs/fda`

5.1.2 Organization and personnel

In this case study I worked with Dr. Ana Szarfman who is a Medical Officer at the Food and Drug Administration (FDA). One of the roles of the FDA is to oversee the number of adverse reports received for different drugs available in the US. Dr. Szarfman has extensive experience on data analytics and has created different techniques to analyze big volumes of data at the FDA. For this purpose, she has used data mining techniques as well as data visualization.

5.1.3 Datasets

Dr. Szarfman has been interested in analyzing the adverse effects[1] reported for different drugs. There are thousands of possible adverse effects for each drug and the FDA has been collecting reports for several years, thus historical information is available. Finding which adverse effects are significant for a certain drug is a complicated task; therefore, Dr. Szarfman and other analysts at the FDA created the Empiric Bayes Geometric Mean (EBGM). The EBGM provides an index of reported adverse effects compared to the expected [82] which serves as an indicator of how significant an adverse effect is for a certain drug. An EBGM value of 1.0 says that the expected number of reports was received for a certain adverse effect. Values greater than 1.0 represent adverse effects that have received more reports than expected (which is bad). The EBGM values are organized in a *fixed***,** *non-aggregated* hierarchy defined by the Medical Dictionary for Regulatory Activities (MedDRA[2]). The MedDRA hierarchy is organized by System Organ Class (SOC), High-Level Group Terms (HLGT), High-Level Terms (HLT) and Preferred Terms (PT). The EBGM index is a non-aggregated value because as an index, the value of a node cannot be calculated using the EBGM values of its children. Furthermore, an average MedDRA tree contains 1614 nodes.

I received seven datasets from Dr. Szarfman, with the first four containing artificially generated values and the last three including real adverse effects *EBGM*s for an undisclosed drug. Each dataset contained similar information but with small modifications. The final dataset contained 2,964 rows, 5 time points, 6 numeric variables ($EB05$, $EB95$, $EBGM$, n, dim, e), and 4 attributes (SOC, $HLGT$, HLT, PT the elements of the hierarchy). From the numeric variables, the $EBGM$ was used as the main comparison

[1] An adverse event is any undesirable experience associated with the use of a medical product in a patient (From FDA's website)

[2] `http://www.meddramsso.com/`

variable. The $EB05$ and $EB95$ represented the confidence intervals of the $EBGM$ value. n contained the number of reports and was used for the width of the StemView boxes to emphasize the relevance of each effect. *dim* and e were not used in the study.

5.1.4 The task

Dr. Szarfman wanted to identify significant changes in the adverse effects reported for a drug over time. For example, she wanted to know if there was a significant increase on the reports of *Haematomas* (a *Vascular* adverse effect) for drug X over the past three years. Moreover, Dr Szarfman wanted to identify adverse effects with significant changes that were not previously considered. To identify those changes, Dr. Szarfman looked for changes in the $EBGM$ index that provided an index of reported adverse effects compared to the expected. As an index, the $EBGM$ value has a confidence interval defined by $[EB05, EB95]$. A significant change is defined by a major shift in the $EBGM$ for an adverse effect with a large number of reports, with no overlapping between the confidence intervals .

Before this case study, analysts at the FDA could explore changes over time in the $EBGM$ and its confidence intervals *for a specific adverse effect (*using time lines)*;* however, they cannot explore all the effects at once. In addition, they have been using a treemap based visualization called the Sector Maps [71] that shows the EBGM values for the adverse effects reported for a drug in a certain year. They also wanted to find changes in the EBGM values between years, and the only way of doing it was to switch back a forth between the Sector Maps or use side by side comparisons (Figure 5.1). A new treemap visualization could have been used where the color represented the change in the EBGM value, but doing so would hide the changes in the inner nodes of the hierarchy. This was undesired since analysts wanted to explore changes in the EBGM values in all levels of the MedDRA hierarchy while still keeping the number of reports per adverse effect. Moreover, they wanted to highlight the adverse effects with non-overlapping confidence intervals, but their current solutions were insufficient for addressing all these requirements.

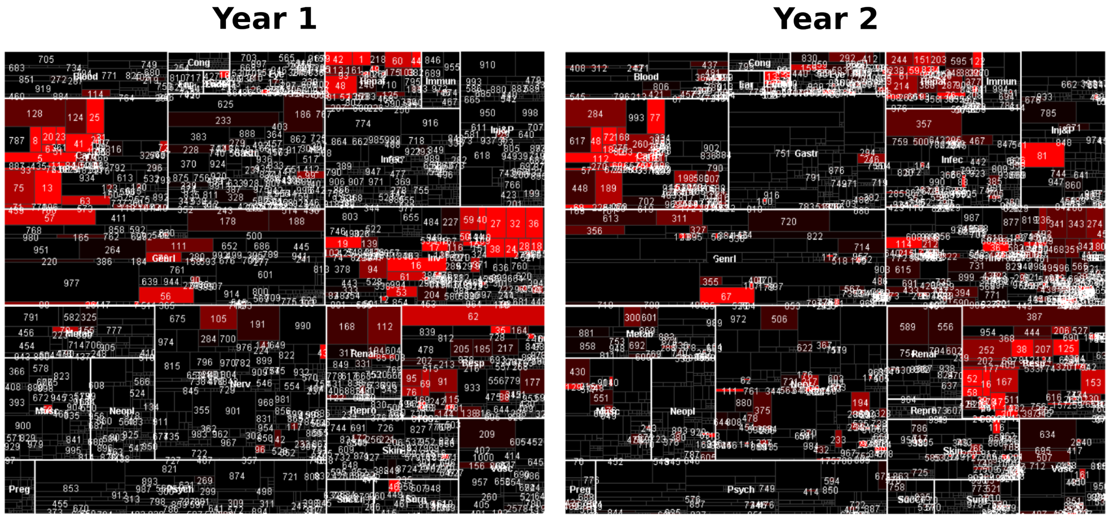


Figure 5.1: Sector Maps for the EBGM values of a drug for two years. Each box represents an adverse effect, with red values encoding high EBGM, which has a bad connotation. Values of the inner nodes were occluded, and could be exposed only by redrawing the Sector Map at a different level. FDA analysts relied on side by side comparisons like this to identify changes before using TreeVersity2.

5.1.5 Work procedures

This case study was developed during a period of ten months where a total of five meetings were held mainly in the FDA offices. The initial meetings were used to characterize the problem and understand the tasks, then an iterative process was followed where new features were implemented as they were identified during the different meetings. During the meetings, the chauffeur mode was used with Dr. Szarfman guiding the analysis while I controlled the interface. I kept a log of each dataset received and made notes of each meeting. Finally, the IRB's exit questionnaire[3] was administered to collect the final impressions of Dr. Szarfman.

5.1.6 Outcomes

Outcomes for TreeVersity refinement

These are the features of TreeVersity2 mainly used in this case study:

- Filters where extensively used to find adverse effects that started in small *EBGM* values and then increased significantly.

- The StemView feature of displaying change in the interior nodes was especially useful for this case study. This was the first time that Dr. Szarfman was able to see changes in the interior nodes of the MedDra hierarchy.

- Localized navigation to allow analysts to zoom into an internal node of the hierarchy, focus on its subtree, and still be able to see changes over time and apply filters without zooming out. This feature was designed for this case study.

- Support for multiple variables at the same time which enabled Dr. Szarfman to explore changes in the *EBGM* values (using the height of the StemView boxes) while also displaying the total number of reports. This feature was especially developed for this study.

[3] `http://goo.gl/zoSbT`

- Support for different color palettes that were closer to the Sector Maps.
- Support of confidence intervals. Since TreeVersity2 was not designed to display changes in confidence intervals, I added a special condition in the code that used color to highlight adverse effects with non overlapping confidence intervals. For this feature I computed the following function: $Max(ending\_EB05 - starting\_EB95, ending\_EB95 - starting\_EB05)$, which returns a positive number only when the confidence intervals don't overlap.

Outcomes for the users

- Figure 5.2 shows the changes between EBGM values for an undisclosed drug between 2010 and 2011 using TreeVersity2 as shown during the final meeting. Each box in the StemView represents an adverse effect, and yellow-to-red colored sub-boxes denote adverse effects with non-overlapping confidence levels. Height corresponded with the relative change of the EBGM index, so sub-boxes going up represented adverse effects that have received more reports (with a fourth root scale). Finally the width of the boxes shows the total number of reports by effect, therefore adverse effects with more reports have wider boxes. With this configuration, Dr. Szarfman was able to find that in 2011 the *Pulmonary Embolism* went from having no reports in 2010 to having a EBGM score of 25.20 which is undesirable. She reported that "*it was incredible that we can see that important effect this way*" and "*it was significant given the drug in question*". Dr. Szarfman also praised TreeVersity2's visualizations for encoding many of the variables required for the comparison in one single view, as well as the possibility of exploring changes over time, "*It looks awesome!*" she said.

5.1.7 Discussion

In this case study I helped Dr. Ana Szarfman explore changes in the index of adverse effects reports (EBGM) for a specific drug over time. For this purpose we used a *fixed,*

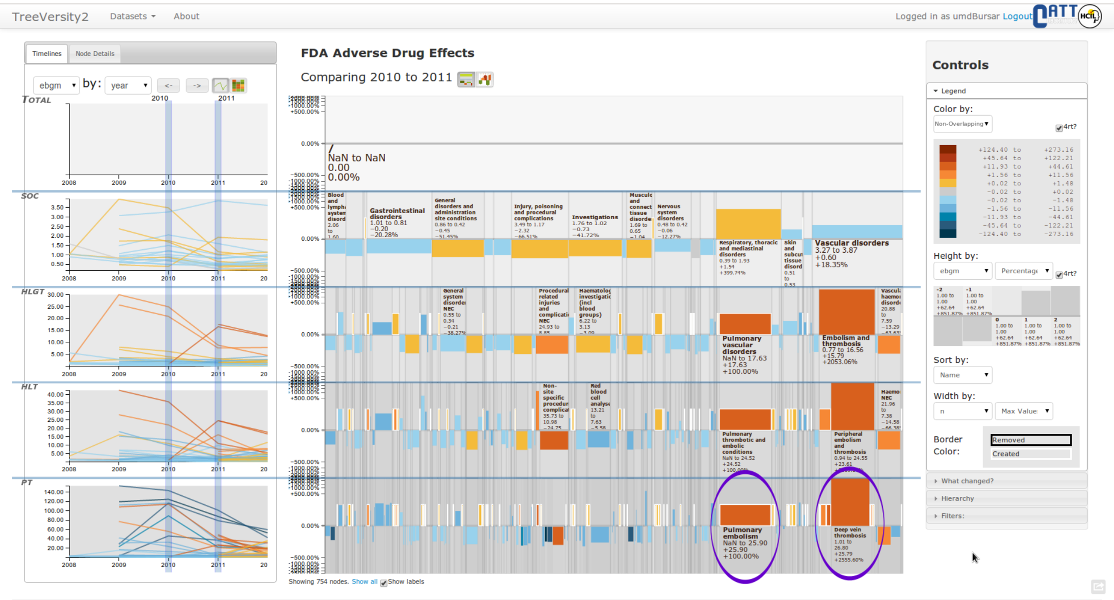


Figure 5.2: Changes in the FDA's EBGM index of adverse effects (e.g. Pulmonary Embolism) for a non-disclosed drug between 2011 and 2010 (Sixth dataset). Using the StemView, analysts were able to identify two relevant adverse effects that received more reports than expected in the 2011 *Pulmonary Embolism* and *Deep Vein Thrombosis* that were not previously reported in 2010 (i.e. created node denoted with white border). The EBGM index is distributed in a fixed, non-aggregated tree and it is a measure of the number of reported adverse effects compared to the expected. A value of 1.0 indicates that the expected number of reports for a certain adverse effect were received, and decreasing values are desirable. The change of each the index is shown using the height of the boxes, so boxes going up indicate adverse effects getting more reports and boxes going down the opposite. The width of the boxes in the StemView represent the total number of reports, so wide boxes are more important. The color was especially crafted to meet a special requirement from the FDA, and to highlight adverse effects with non-overlapping confidence intervals (shown on yellow and red). Therefore, analysts searched for wide, red/yellow boxes going up.

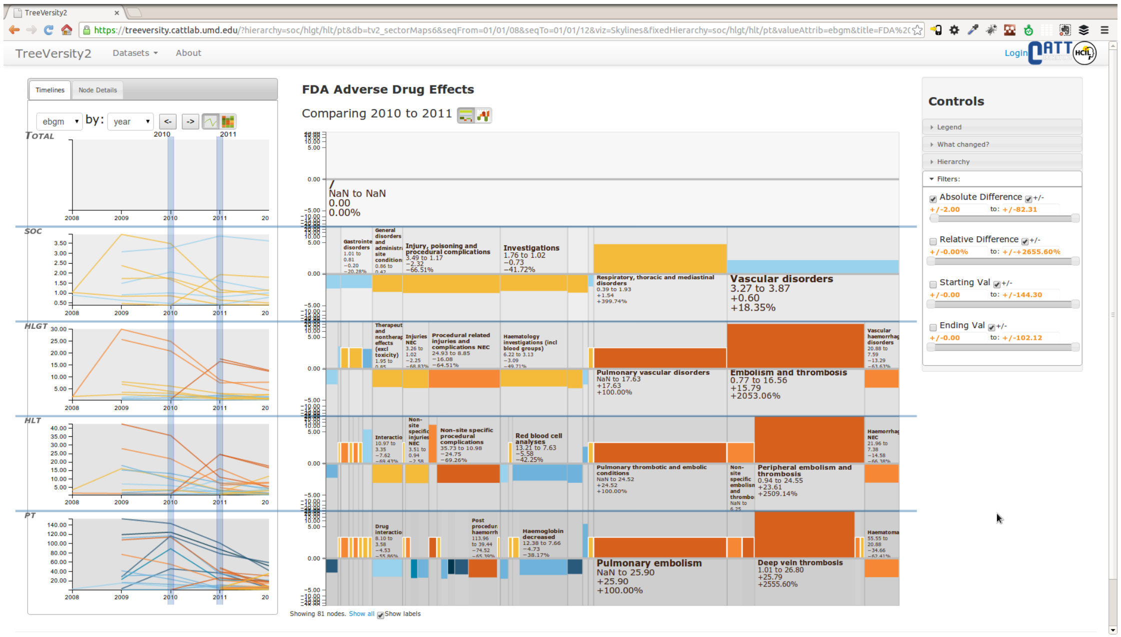


Figure 5.3: FDA's sixth dataset with the same configuration as in Figure 5.2 after filtering for adverse effects changing more than 2.0 in the EBGM value.

non-aggregated hierarchy with *changes in topology* (Type 4). Dr. Szarfman was thrilled that she could use TreeVersity2 to find significant changes in the adverse effects for an undisclosed drug over five years of data, such as the 2011 "Pulmonary Embolism" increase that went from no reports to an EBGM value of 25.20. Several features were added to TreeVersity2 to allow the type of comparisons that Dr. Szarfman required. The case study revealed some limitations on the tool, especially on issues for installing TreeVersity2 on FDA servers and disparities with color encoding when compared to the Sector Maps. I received seven versions of the data and a total of five visits were made to the FDA to show new features and collect feedback. Given the success of the case study and the interest of continued use of TreeVersity2 for her daily work, Dr. Szarfman is seeking to implement TreeVersity2 at the FDA.

Dr. Szarfman was extremely excited to see the final results, she said "*Awesome findings*" and added "*It looks awesome!*". However she expressed some issues with the color codification representing the non overlapping adverse effects, which is significantly different to what they are used to with the Sector Maps. She said that it might take some training to adjust to that change. She also mentioned that it will be useful to have the StemView box widths match the areas used for the Sector Maps, and she agreed to send that information to me. Despite these issues, Dr. Szarfman was very interested in using TreeVersity2 in they day to day work, so a proposal is in the works to get a third party consultant to implement a especially designed version for FDA's needs.

5.2 National Cancer Institute

5.2.1 Case Study Sheet

Partner: Dr. Carol Kosary, Program Manager for the Surveillance, Epidemiology, and End Results (SEER) program and Dr. Bradford Hesse, Chief of the Health Communication and Informatics Research Branch (HCIRB)
Organization: National Cancer Institute
MILCS level: Early, Chauffeur Mode
Duration: November 2 2013 - April 2013 (6 Months, 4 meetings)
System used: TreeVersity2
Data: Lung cancer related death index in the U.S. between 1997 and 2009
Number of rows: 1,716
Number of time points: 13 (years)
Example tree size: 101 (3 Levels)
Number of numeric variables: 3 (death rate, death count, population)
Number of Attributes: 3 (Gender, Race, Counties by Deciles for Ever Smoked question)
Type of Hierarchy: Dynamic
Type of comparison: Type 2: non aggregating + same topology
New Features: Support for multiple attributes, Timeblocks
Limitations: The StemView might be too complicated for the general public
URL: `https://treeversity.cattlab.umd.edu/cs/nci`

5.2.2 Organization and personnel

For this case study I worked with Dr. Carol Kosary the Program Manager for the Surveillance, Epidemiology, and End Results (SEER) of the National Cancer Institute (NCI). The initial approaches with the NCI were made with Dr. Bradford Hesse that helped me identify projects in the NCI that could benefit from TreeVersity2 and directed me to Dr. Kosary. As manager of the SEER, Dr. Kosary has access to extensive Cancer related datasets and "*oversees the development, enhancement, maintenance, and deployment of large IT systems that support central cancer registry operations such as SEER*DMS and the development of electronic tools for data capture, including but not limited to electronic pathology (ePath)*"[4].

5.2.3 Datasets

I worked with Dr. Kosary to explore changes in the lung cancer related death-rates in the US between 1997 and 2009. They calculated a normalized lung death-rate across the counties in the US, splitting them in ten comparable groups (i.e. by deciles) according to what percentage of the population have ever smoked. The dataset was also subsequently divided by ethnicity and gender, moreover the population and death counts were also included with the data. The dataset contained 1716 rows, distributed in 13 years, with 3 variables: death rate, death count and population. An example tree built from the dataset contains around 100 nodes with all the 3 levels. The index was normalized to allow the comparison between county deciles, because of this the generated tree is non-aggregating (the values of the interior nodes cannot be calculated using the values of the leafs).

5.2.4 The task

Dr. Kosary wanted to analyze the change on the Lung Cancer related death index over time and compare it with the percentages of how many people have ever smoked in the

[4] Taken from `http://surveillance.cancer.gov/about/bios/kosaryc.html`

different counties in the US (the counties were grouped in deciles). First she wanted to confirm her intuition that less smoking counties would have smaller death rates, but with the use of TreeVersity2 she realized that the tool could be used also to find anomalies in the data.

5.2.5 Work procedures

The case study was developed over a period of four months, with five meetings held in the National Cancer Institute building in Bethesda, Maryland. After an introduction via email to Dr. Hesse, I did an introductory presentation of TreeVersity2 to a group of NCI's analysts, which included Dr. Kosary. After this we met to discuss possible applications of TreeVersity2 on their datasets, and followed an iterative process to agree on data format. To conclude the study, we held a meeting with more than ten NCI's data analysts to present my findings. The meetings lasted for about one hour each one.

5.2.6 Outcomes

Outcomes for TreeVersity refinement

- The Lung Cancer death rate is a commonly decreasing index, however in one of the 13 years, the index increased its value. In order to find this type of changes more easily I implemented the TimeBlocks.

- NCI's analysts wanted to compare the death rates for the different races in the population, while keeping into account the population numbers. For this I used the multiple variables feature of TreeVersity2.

Outcomes for the users

- Dr. Kosary was able to confirm with TreeVersity2 that the Lung Cancer death rates for the counties are correlated with the number of smoking people on them. This was not a new discovery, but Dr. Kosary was excited to see it clearly displayed on TreeVersity2 visualizations.

- Although it was not planned initially, we were able to find anomalies in the dataset, especially with the "other" race, which oscillated between years in patterns different to the ones of the other races. When discussing this with Dr. Kosary and her data management staff, they explained that this might be due to inconsistencies in the data collection methods used by the different counties, which might differ in their definition the "other" race.

5.2.7 Discussion

For a first exploration of the dataset a dynamic, non aggregating (the inner nodes were normalized) hierarchy was used, that grouped it by ethnicity, then by gender and finally by the counties deciles, as shown in Figure 5.4. In the image, color represents the relative change of the death_rate (decreasing values on green), and the height of the sub-boxes encode the actual change in the death-rate. In order to highlight the groups sizes, the value of the population counts (the max between the values of 1997 and 1998) of each node of the tree was selected for the width. Finally the TimeBlocks were used to compare the change of each group across time.

As shown in Figure 5.4, analysts were interested when seeing that the death-rate increased only in 2000 (a). They also found interesting to be able to see how this increase was due mainly to whites (b) in general, and to white females in specific (c). Other relevant findings show how the "other" race fluctuated between increases and decreases between years (d), when the remaining races decreased more consistently (e). They explained that it might have been due to inconsistencies in the definition of the race "other" between years for the population count purposes. The initial exploration also suggested that African American men death rates (f) decreased more significantly than those of African American Females (g). With this information and given that the hierarchy is dynamic, the grouping order was changed to Gender->Ethnicity->Counties-Deciles, which confirmed the tendency (not shown in the Figure). Analysts explained that this might have been to smoking reduction campaigns being targeted mainly to men. Finally the hierarchy was changed again to put the grouping of the counties at the top, which revealed

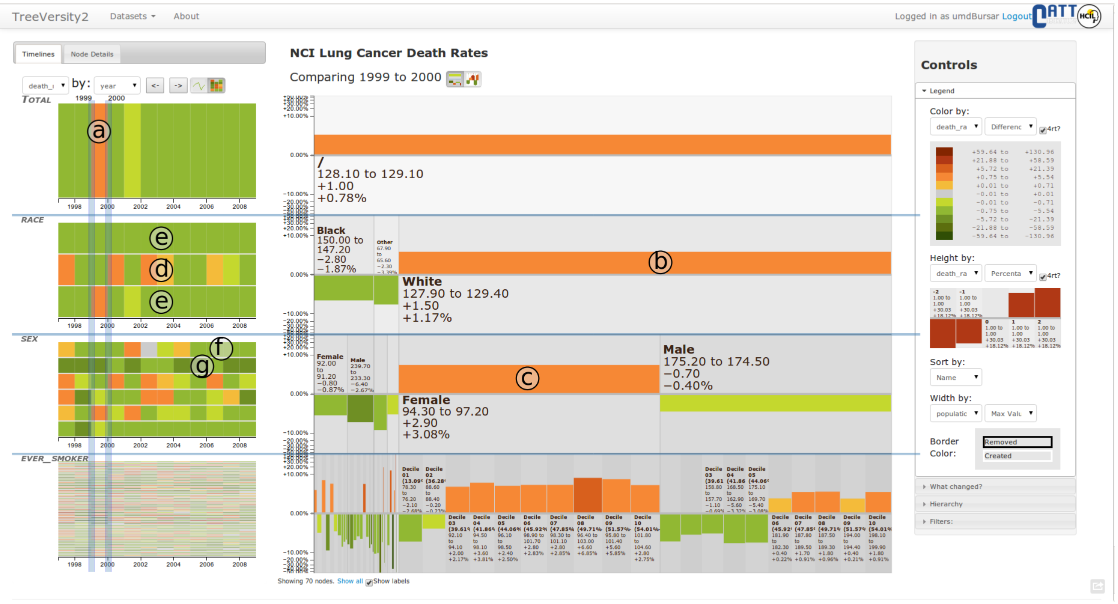


Figure 5.4: National Cancer Institute Lung Cancer related death-rate change between 1999 and 2000 in the US. Color shows relative change in the rate, and height represents relative change in the rate. the width encodes the population counts for each group. The TimeBlocks show that the (a) overall rate increases only in 2000, however (b) the only race increasing is "White", that also happens to be more than 80% of the population. Among whites though, (c) women seem to be the ones contributing the most to the increase.

the expected correlation between the smoking and lung cancer death.

Analysts at the NCI were excited to see the changes in their datasets in a visual way, and liked the flexibility of TreeVersity2 to switch parameters. They express that tools like TreeVersity2 could be used by them to communicate in a more effective way their findings to the general public, however they they were concerned with the learning curve required to understand the StemView.

5.3 eBay Products Category Tree

5.3.1 Case Study Sheet

Partner: Andy Edmonds, Distinguished Product Manager, Experimentation & Learning
Organization: eBay
MILCS level: Early, self driven
Duration: August 2012 to February 2013 (7 months, 2 meetings)
System used: TreeVersity2
Data: eBay's product sales and inventories for a period of four weeks during holidays 2012
Number of rows: 63,098
Number of time points: 4 (weeks)
Example tree size: 5,443 (4 Levels)
Number of numeric variables: 2 (Product sales, product counts)
Number of Attributes: 6 **(**Ebay's six levels product hierarchy)
Type of Hierarchy: Fixed
Type of comparison: Type 1: aggregating + same topology
New Features:
Limitations: The full eBay product tree is too big for the current implementation of TreeVersity2
URL: restricted

5.3.2 Organization and personnel

In this case study Mr. Andy Edmonds Distinguised Product Manager at eBay, used TreeVersity2 to analyze the changes in the sales and inventories of products of the sales site during four weeks of the holidays season of 2012.

5.3.3 The task

Mr Edmonds was interested in visualizing changes in eBay's category in order to communicate to other members of the company the relevance of the different branches of the tree. eBay's category tree groups all the products available on the site by its characteristics.

5.3.4 Datasets

eBay's Category tree contains more than 10,000 nodes distributed in 6 levels. The dataset provided contained 63,098 rows that contained information for the four weeks between 11 November 2012 and 12 December 2012. The data included the number of product sold on each category for that week, as well as the total product count.

5.3.5 Work procedures

I met with Mr. Edmonds on two occasions, one at the CHI conference in Austin Texas where I first demonstrated TreeVersity, and the second at eBay's headquarters in San Jose, California. Mr. Edmonds used the tool on his own and also with the "inventory/structured data/classification, search science, and research lab teams".

5.3.6 Outcomes

Outcomes for TreeVersity refinement

- This case study pushed TreeVersity2 limits on terms of the number of nodes that can be represented at the same time on the screen. The categories tree had to be analyzed using two

Outcomes for the users

- Mr Edmonds reported that "*the classification team came up with a large number of use cases most of which required flexible date aggregation and multi-year* *datasets. Additional opportunities for inventory sourcing were also imagined. For search science, a compelling case was made for seasonal demand changes that further motivated using past year data in feedback loops (e.g. what kind of items should we show for the query fossil? dinosaur bones or purses?). The research team appreciated the visualization accomplishment.*"

- When asked to summarize his discoveries Mr. Edmonds answered: "*Numerous examples of obvious, but heretofore unrevealed patterns were discovered. The ebay US category tree, at > 9k nodes, has a huge number of category branches that many of our staff are not aware of. Browsing these fairly obvious seasonal changes provided a better understanding of the depth and utility of the tree.*

- *More directly impactful, the use case for seasonal variability and using last year's data of the specific time frame was strengthened, rather than just a trailing temporal set of data from recent logs.*"

- Mr. Edmonds also had to say this when comparing TreeVersity2 to previous approaches used by them: "*I've spent many hours in color coded Excel files expressing changes at query and category levels. The visualization and what changed views made this process much more efficient and easier to share.*"

5.3.7 Discussion

Figure 5.5 shows the changes in the total number of products for the first four levels of the tree, between 18 November 2012 and 25 November 2012. The full tree was not used for the analysis because it was over the sweet spot that TreeVersity2 supported (beyond 7,000 nodes the interactions become slow). The figure highlights how the Collectibles category is by far the biggest category on the site, presenting an increase of 15 new items listed during that week. Other interesting patterns were found but they are not included here because of the sensitivity of the data.

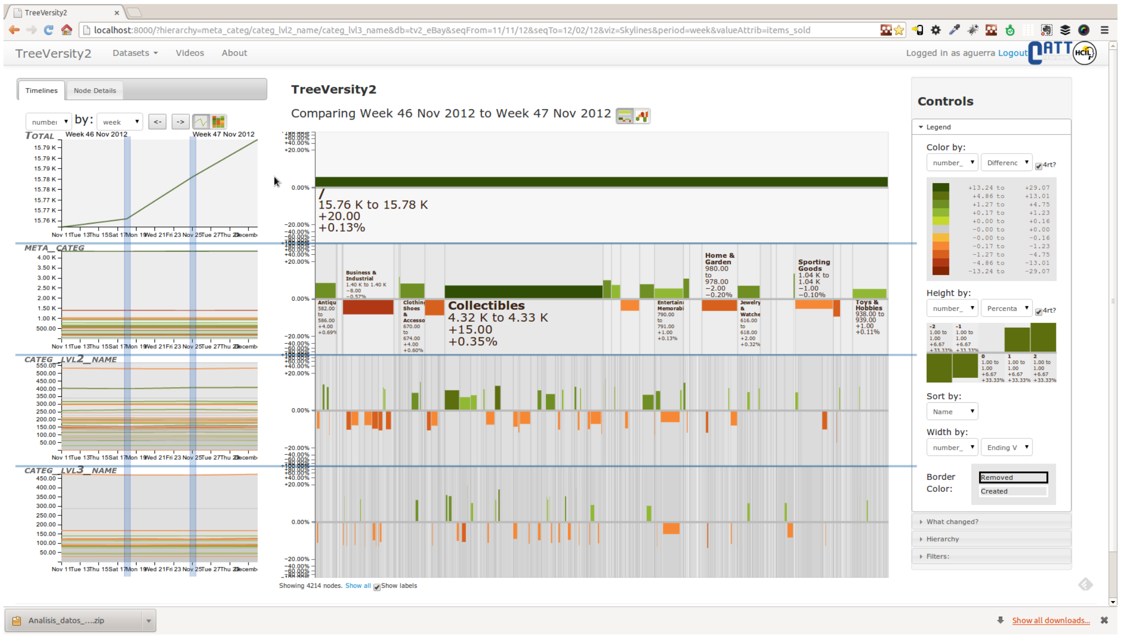


Figure 5.5: Changes on the number of items offered on eBay's during 2012 holidays

5.4 University of Maryland students

5.4.1 Case Study Sheet

Partner: Stephanie Lee David Coordinator, Data Integrity and Pamela M. Phillips Associate Director

Organization: Office of Institutional Research, Planning & Assessment (IRPA), University of Maryland (UMD)

MILCS level: Mature, Chauffeur Mode

Duration: November 2 2013 - April 2013 (6 Months, 4 meetings)

System used: TreeVersity2

Data: UMD students demographic information between 2008 and 2013

Number of rows: 227,158

Number of time points: 5 (years)

Example tree size: 715 (5 Levels)

Number of numeric variables: 6 (number of students, last accumulative GPA, age)

Number of Attributes: 219 (Gender, Race, Origin Country, College, Major, etc)

Type of Hierarchy: Mixed

Type of comparison: Type 3: aggregating + different topology

New Features: Support difference aggregation functions when using multiple variables (e.g. aggregate number of students by summing up, and GPA by averaging)

Limitations: TreeVersity requires training to get its full power

URL: Restricted

5.4.2 Organization and personnel

In this case study I worked with a group of domain experts from the Office of Institutional Research, Planning & Assessment (IRPA) of the University of Maryland at College Park, led by Dr. Mona Levine the Associate Vicepresident of IRPA. After a series of meetings were TreeVersity2 was demonstrated, I started cooperating with Stephanie Lee David Coordinator, Data Integrity and Pamela M. Phillips Associate Director of IRPA.

The University of Maryland has more than 46,000 students distributed between the many colleges and programs that it offers. The Office of Institutional Research, Planning, and Assessment of the University is in charge on analyzing their information "for the purposes of decision-making, policy analysis, strategic planning, mandated reporting, and academic program review" (quote taken from the IRPA website).

5.4.3 Datasets

The data used for this case study is the biggest sample used with TreeVersity2 to date. It contains information about the University of Maryland's students between 2008 to 2013. The dataset contained 227,158 rows and 219 attributes, including information about the students' majors, demography, place of origin and grades among others. Some attributes of the dataset are fixed hierarchies, such as School->Department, were most of the others can be used to create dynamic hierarchies. This dataset is a good example of a tree comparison Type 3, aggregating tree with changes in topology.

5.4.4 The task

In this case study, I helped IRPA's Stephanie David and Pam Phillips analyze the changes in number of students, ages and GPAs for the students at the University of Maryland for the five years. They were especially interested in finding differences between the current student database and the previous years (that they call frozen).

5.4.5 Work procedures

The case study was developed during a period of six months, where four meetings were conducted, normally on the office of IRPA at the University of Maryland. The initial meeting organized by Dr. Levine, which gathered twelve members of the IRPA and other administrative organs of the University, was used for brainstorming about where to better use the tool. The outcome of the meeting was the selection of the process of comparison of the students information between their frozen databases (previous years) and the current version. For this specific task I started working with Ms. Phillips and Ms. David, who in cooperation with Mr. Kyle Langford (a manager at the Enterprise Database Services of the University) provided me with a dataset with the information of the students registered in the University.

5.4.6 Outcomes

Outcomes for TreeVersity refinement

- Access control was implemented for this case study to restrict access to this dataset.
- Given the high number of attributes in the dataset, TreeVersity2's dynamic hierarchy control needed to be modified to allow navigation of the attributes. This feature was extensively used because through the organization of attributes into trees in different orders allowed analysts to explore different questions.

- The numeric variables of this dataset required different aggregation functions, i.e. while the number of students can be aggregated summing up the values, the GPA required averages for aggregation. To support this feature, I implemented a especial parameter in TreeVersity2 that allowed the selection of different aggregation functions for the variables in the dataset.

Outcomes for the users

- Thanks to this case study, Ms. Phillips and Ms. David were able to start asking new types of questions (more oriented to comparisons) on their data. They have expressed their intention on continuing using TreeVersity2, and for this they are preparing new datasets and questions.

5.4.7 Discussion

Ms. David and Ms. Phillips were very excited to see the flexibility of the tool to explore a broad range of questions on their data. They organized a staff meeting to present it to twelve analysts and staff members of the IRPA and the Bursar office to find how they can use TreeVersity2 in their day to day work. During this meeting analysts reported that TreeVersity2 "gave me a picture of the data really quickly", that it "could help us look for what they don't know", and that it was "fun to look at" which could ease the cognitive load of performing tedious exploration tasks. They also found especially useful the reporting tool, and reported that it could be useful in their work to have the tool compute their common queries so they can explore them more easily.

5.5 Change in passengers flying in the US between 1990-2003

5.5.1 Case Study Sheet

Partner: Martin Akerman, Analyst; Pat Hu, Associate Administrator and Director, Bureau of Transportation Statistics
Organization: Department of Transportation

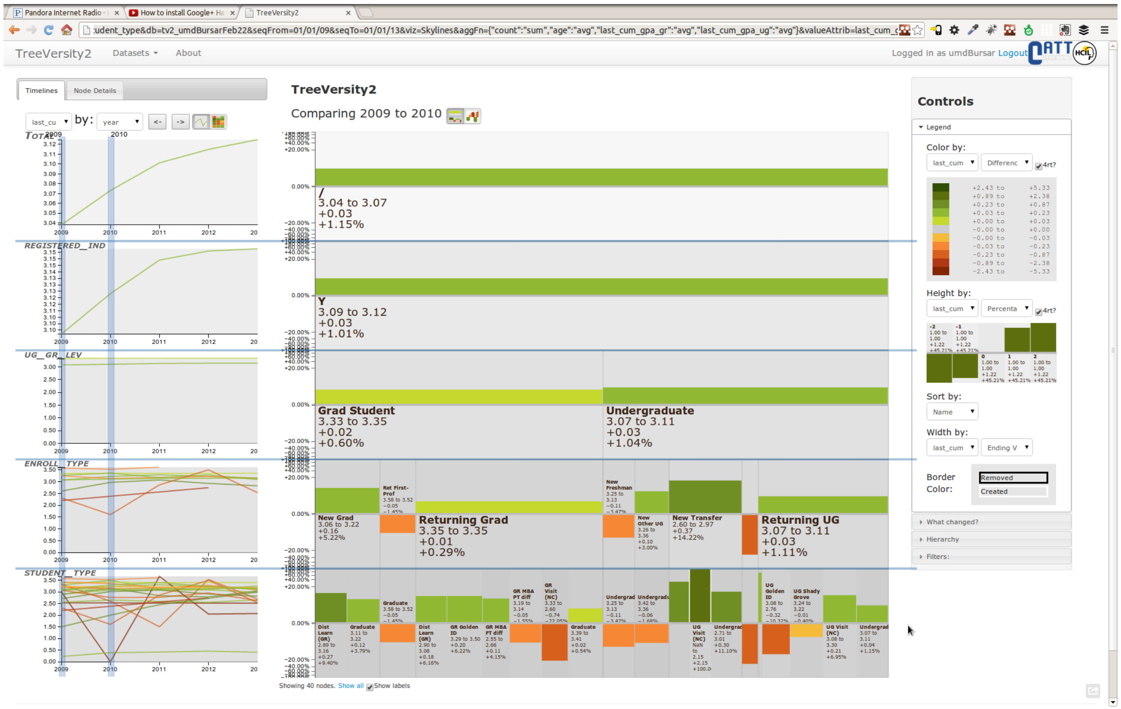


Figure 5.6: UMD Returning students grouping by graduate level, enroll type and student type

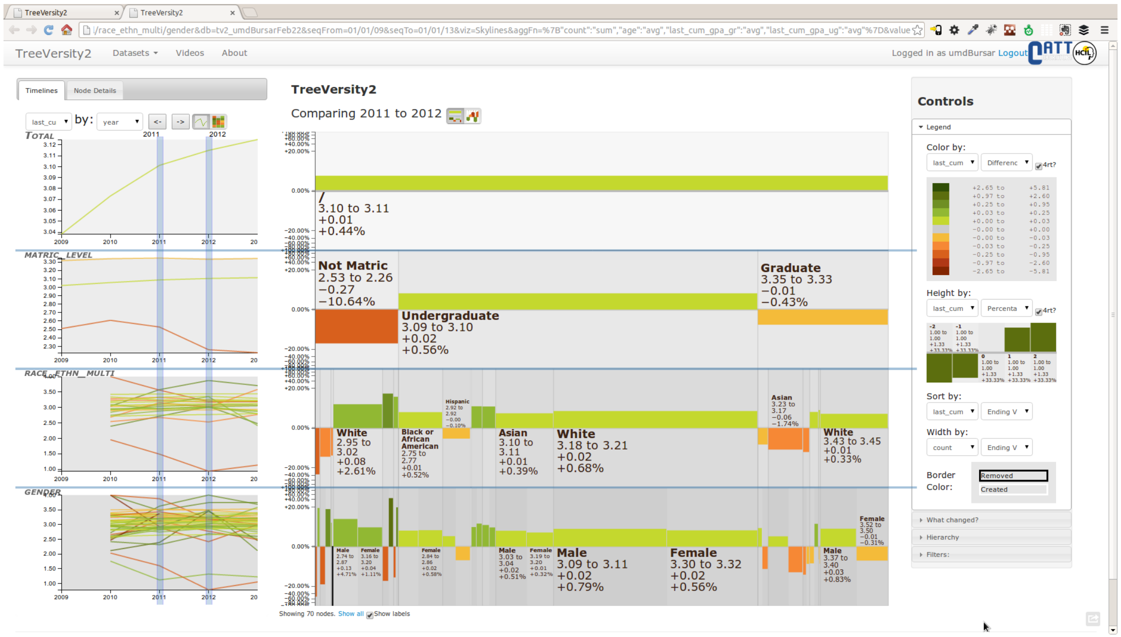


Figure 5.7: UMD students average GPA by level races and gender

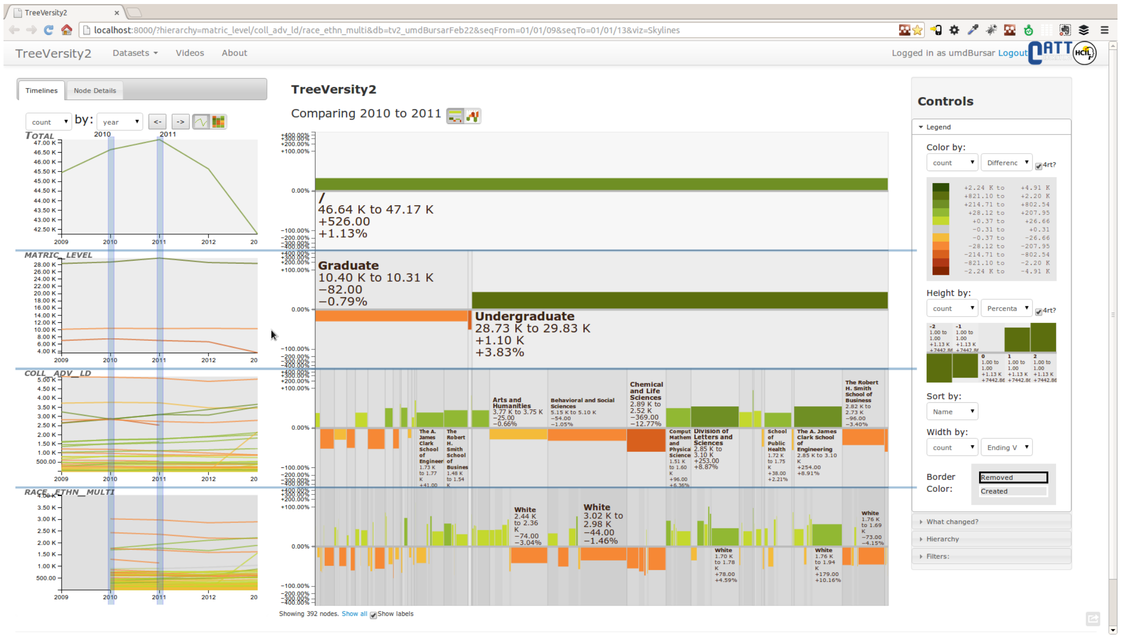


Figure 5.8: UMD students by level college and race

MILCS level: Early, Chauffeur Mode

Duration: November 2011 to July 2012 (9 months, 3 meetings)

System used: TreeVersity and TreeVersity2

Data: Number of passengers that took off from US airports between 1990 to 2003

Number of rows: 65,534

Number of time points: 162 (months)

Example tree size: 4,194 (3 Levels)

Number of numeric variables: 3 (death rate, death count, population)

Number of Attributes: 3 (Gender, Race, Counties by Deciles for Ever Smoked question)

Type of Hierarchy: Dynamic

Type of comparison: Type 2: non aggregating + same topology

New Features: This was one of the inspirational datasets to build TreeVersity2

URL: `https://treeversity.cattlab.umd.edu/cs/pax`

5.5.2 Organization and personnel

In this case study I worked with Mr. Martin Akerman a data analyst that used to work for the Department of Transportation (DOT) under the supervision of Ms. Patricia S. Hu the Associate Administrator and Director of the Bureau of Transportation Statistics (BTS). The BTS was "created to administer data collection, analysis, and reporting and to ensure the most cost-effective use of transportation-monitoring resources" [5].

5.5.3 Datasets

The dataset used in this case study contained 65,535 records representing the number of passengers reported to have taken off an U.S. airport, month by month, between 1990 and 2003. The data also included the Airport's State and City, which where used to create the hierarchy.

5.5.4 The task

Mr. Akerman wanted to explore the dataset to find interesting patterns. We knew before hand that an interesting pattern would emerge from the attacks of 9/11, but he wanted to explore that and other changes over time in the data.

5.5.5 Work procedures

The case study was developed during a period of nine months between November 2011 and July 2012. As in the previous case studies, an initial meeting was held to discuss possible datasets that could be analyzed with my approaches. In this case study we started the analysis using TreeVersity, but after showing the results to Mr. Akerman and other members of Ms. Hu team, we realized that the analysis to this dataset could benefit even more from looking at the whole extension of the time line rather than just looking at two years at a time.

[5] Taken from BTS website: `http://www.rita.dot.gov/bts/about`

5.5.6 Outcomes

Outcomes for TreeVersity refinement

- This case study was conducted initially with TreeVersity and the analysts at the DOT were excited to see changes in the number of passengers between two years, however they wanted to expand the exploration to multiple time points. This was one of the inspirations to design and build TreeVersity2.
- TreeVersity's Zooming and filtering functions were extensible used for finding airports with an increasing number of passengers after 9/11.
- TreeVersity's navigate into subtree feature was helpful to explore the changes within a month.

Outcomes for the users

- We discovered something unexpected: not all the airports reported decreases in September 2001. MOT in North Dakota is one of those airports, that stands out because it reported an increase of more than 442% compared to the previous year, going from 673 to 3,654 enplanements. Ms. Hu and Mr. Akerman were excited to find this on their data.

5.5.7 Discussion

This case study was done on the total number of enplanements on the different U.S. Airports, by state and by month. As an initial approach the dataset was explored using Spotfire (an off-the-shelf visualization tool) using line charts and heat-mats over time. As shown on Figure 5.9, the visualizations suggested that the most relevant changes in the data occurred between 2001 and 2000, were a big drop on the number of enplanements was reported, probably because of the attacks of 9/11. To further analyze this change, TreeVersity was used to compare the years 2001 and 2000. The hierarchies for each year were built grouping the data by Airport, then by State and finally by Month. So the hierarchy used was Total -> Month -> State -> Airport.

Figure 5.10 shows an overview of the data by Month and then by State. The nodes in the tree are sorted by the absolute amount of change, and the color was used for the change in percentage with respect to 2000. The four dark subtrees on the left show the big decrease on the number of enplanments that occurred from September to December, while months previous to the attack like January and August were the ones increasing the most. Figure 5.11 shows the same overview but adding labels to all the nodes that change in more than one million passengers, it shows how states like California, Texas and New York suffered big decreases in 2001.

Performing a more in-depth analysis, Figure 5.12 shows the changes on the enplanements by State and then by Airport for the month of September only. The prominent yellow to red colors shows that decreases were reported by all States, and almost all the Airports, however some green nodes revealed that even after 9/11, some Airports actually reported increases in their number of passengers boarding. Figure 5.13 shows all the airports that reported increases over 300 passengers, with Minot (MOT) in North Dakota reporting on September 2001 an increase of more than 442% on the number of enplanements compared to the same month on 2000.

Replicating this analysis with TreeVersity2 I was able to find even more interesting patterns. Figure 5.14 shows the overall comparison of the changes by Month and State, which highlights the big decrease of September 2001. Moreover, exploring the changes between 2001 and 2000 by State and then by city, revealed that most States decreased their number of passengers in 2001, however some states, such as Maryland, presented increases, as shown in Figure 5.15. Figures 5.16 expands this exploration by filtering the cities to show only those presenting increases, which showed that important cities such as Oakland, Pittsburgh, Fort Lauderdale and Baltimore increased their number of passengers overall. Figure 5.17 shows the zoomed in view of the changes in Baltimore by month.

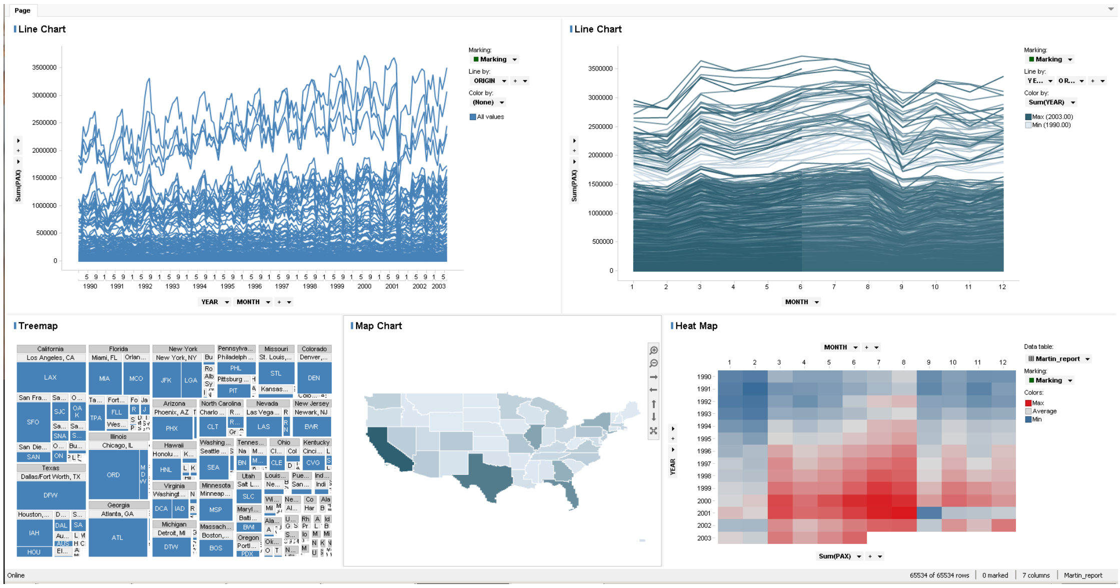
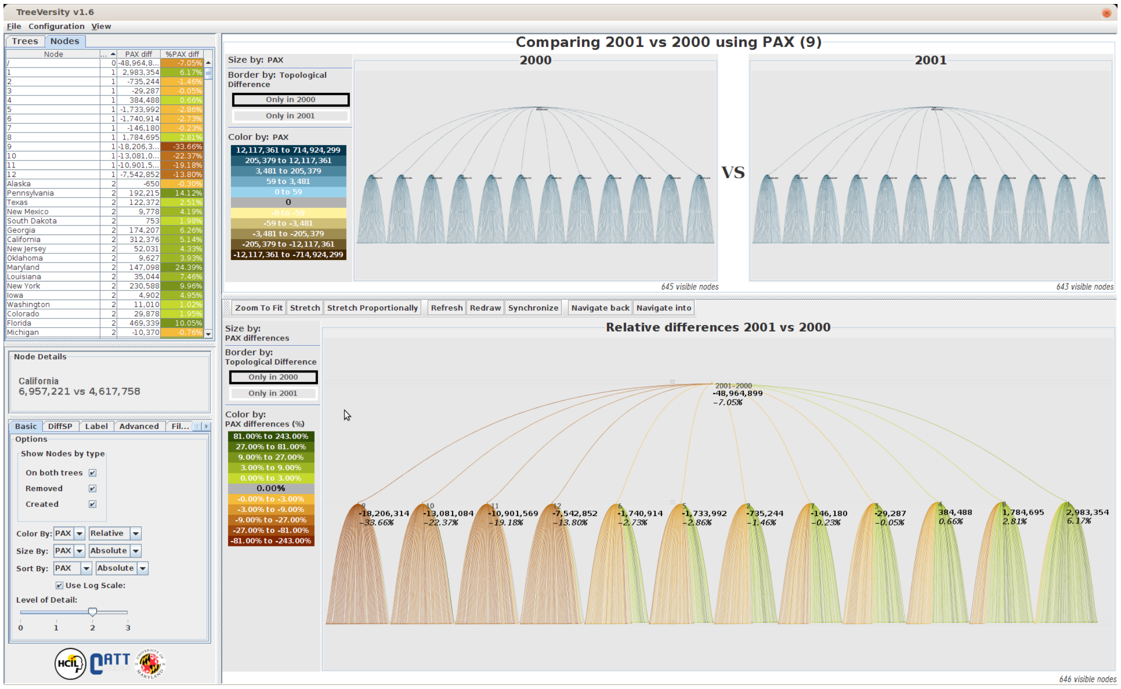


Figure 5.9: Initial Exploration in Spotfire shows a big decrease on the number of passengers reported country wide on September 2001


TreeVersity v1.6

File Configuration View

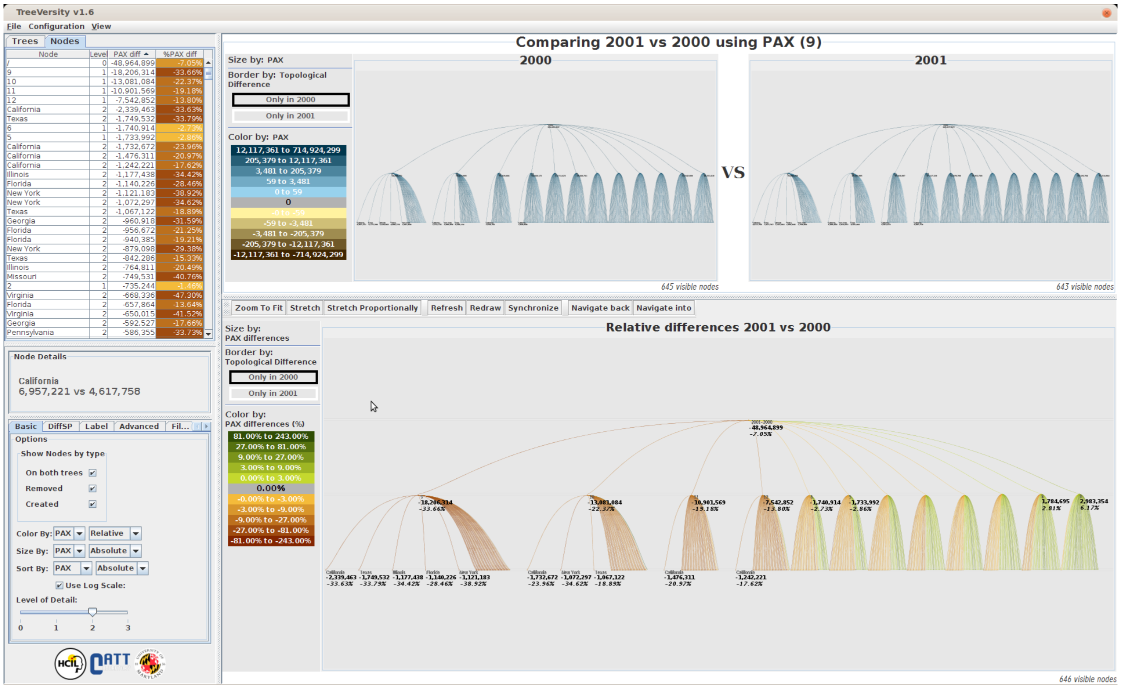
Trees Nodes

Node	...	PAX diff	%PAX diff
/	0	-48,964,8...	-7.05%
1	1	2,983,354	6.17%
2	1	-735,244	-1.46%
3	1	-29,287	-0.05%
4	1	384,488	0.66%
5	1	-1,733,992	-2.86%
6	1	-1,740,914	-2.73%
7	1	-146,180	-0.23%
8	1	1,784,695	2.81%
9	1	-18,206,3...	-33.66%
10	1	-13,081,0...	-22.37%
11	1	-10,901,5...	-19.18%
12	1	-7,542,852	-13.80%
Alaska	2	-650	-0.30%
Pennsylvania	2	192,215	14.12%
Texas	2	122,372	2.51%
New Mexico	2	9,778	4.19%
South Dakota	2	753	1.98%
Georgia	2	174,207	6.26%
California	2	312,376	5.14%
New Jersey	2	52,031	4.33%
Oklahoma	2	9,627	3.93%
Maryland	2	147,098	24.39%
Louisiana	2	35,044	7.46%
New York	2	230,588	9.96%
Iowa	2	4,902	4.95%
Washington	2	11,010	1.02%
Colorado	2	29,878	1.95%
Florida	2	469,339	10.05%
Michigan	2	-10,370	-0.76%

Node Details

California
6,957,221 vs 4,617,758

Figure 5.10: Change in number of passengers by month and State between 2001 and 2000. Sorted by Absolute Change



Node	Level	PAX diff	%PAX diff
/	0	-48,964,899	-7.05%
9	1	-18,206,314	-33.66%
10	1	-13,081,084	-22.37%
11	1	-10,901,569	-19.18%
12	1	-7,542,852	-13.80%
California	2	-2,339,463	-33.63%
Texas	2	-1,749,532	-33.79%
6	1	-1,740,914	-2.73%
5	1	-1,733,992	-2.86%
California	2	-1,732,672	-23.96%
California	2	-1,476,311	-20.97%
California	2	-1,242,221	-17.62%
Illinois	2	-1,177,438	-34.42%
Florida	2	-1,140,226	-28.46%
New York	2	-1,121,183	-38.92%
New York	2	-1,072,297	-34.62%
Texas	2	-1,067,122	-18.89%
Georgia	2	-960,918	-31.59%
Florida	2	-956,672	-21.25%
Florida	2	-940,385	-19.21%
New York	2	-879,098	-29.38%
Texas	2	-842,286	-15.33%
Illinois	2	-764,811	-20.49%
Missouri	2	-749,531	-40.76%
2	1	-735,244	-1.46%
Virginia	2	-668,336	-47.30%
Florida	2	-657,864	-13.64%
Virginia	2	-650,015	-41.52%
Georgia	2	-592,527	-17.66%
Pennsylvania	2	-586,355	-33.73%

Node Details

California
6,957,221 vs 4,617,758

Options

Show Nodes by type

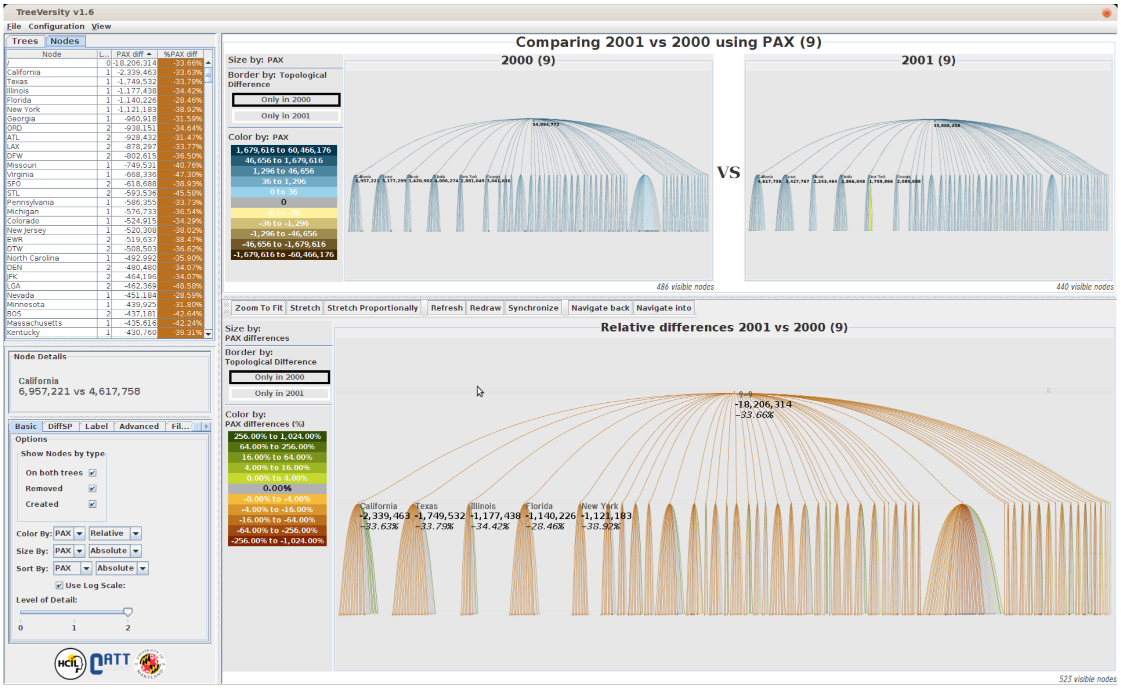
On both trees ☑
Removed ☑
Created ☑

Color By: PAX Relative
Size By: PAX Absolute
Sort By: PAX Absolute
☑ Use Log Scale:
Level of Detail: 2 (0 1 2 3)

Size by: PAX
Border by: Topological Difference
Color by: PAX

Size by: PAX differences
Border by: Topological Difference
Color by: PAX differences (%)

Figure 5.11: 2001vs2000 Months and States decreasing more than 1 million. There is a huge decrease in September, October, November and December



TreeVersity v1.6

Node	L...	PAX diff	%PAX diff
/	0	-18,206,314	-33.66%
California	1	-2,339,463	-33.63%
Texas	1	-1,749,532	-33.79%
Illinois	1	-1,177,438	-34.42%
Florida	1	-1,140,226	-28.46%
New York	1	-1,121,183	-38.92%
Georgia	1	-960,918	-31.59%
ORD	2	-938,151	-34.64%
ATL	2	-928,432	-31.47%
LAX	2	-878,297	-33.77%
DFW	2	-802,615	-36.50%
Missouri	1	-749,531	-40.76%
Virginia	1	-668,336	-47.30%
SFO	2	-618,688	-38.93%
STL	2	-593,536	-45.58%
Pennsylvania	1	-586,355	-33.73%
Michigan	1	-576,733	-36.54%
Colorado	1	-524,915	-34.29%
New Jersey	1	-520,308	-38.02%
EWR	2	-519,637	-38.47%
DTW	2	-508,503	-36.62%
North Carolina	1	-492,992	-35.90%
DEN	2	-480,480	-34.07%
JFK	2	-464,196	-34.07%
LGA	2	-462,369	-48.58%
Nevada	1	-451,184	-28.59%
Minnesota	1	-439,925	-31.80%
BOS	2	-437,181	-42.64%
Massachusetts	1	-435,616	-42.24%
Kentucky	1	-430,760	-39.31%

Node Details

California
6,957,221 vs 4,617,758

Options

Show Nodes by type

On both trees ☑

Removed ☑

Created ☑

Color By: PAX Relative

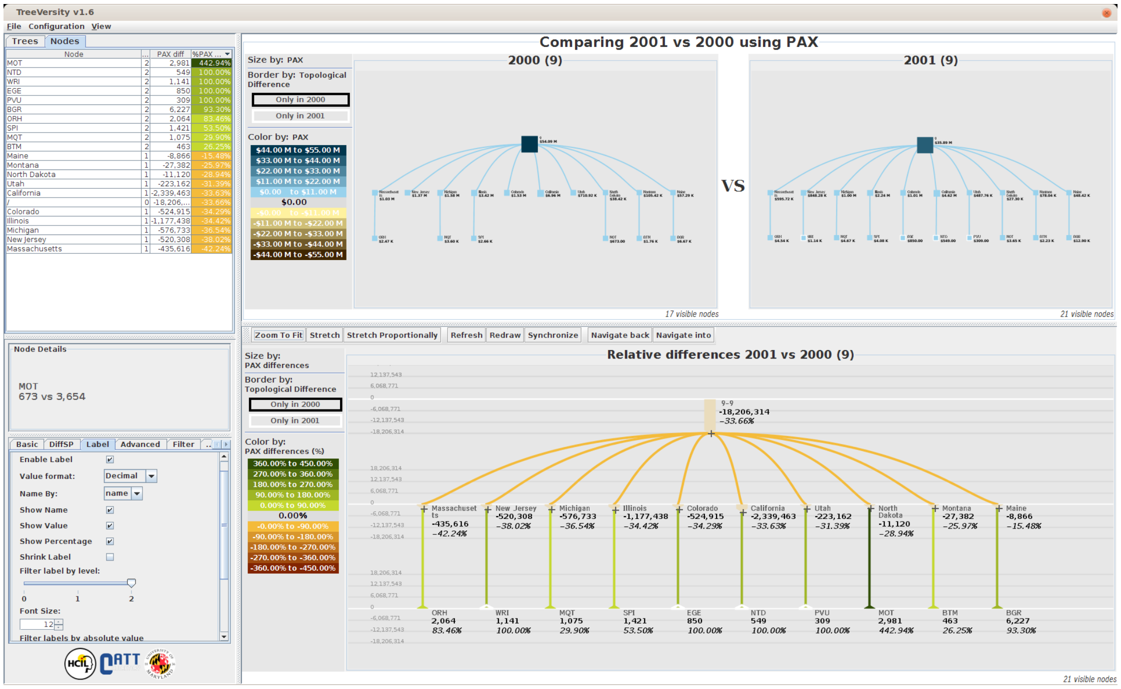
Size By: PAX Absolute

Sort By: PAX Absolute

☑ Use Log Scale:

Level of Detail: 2

Figure 5.12: Change in number of passengers in States and Airports in September 2001 compared to September 2000. California, Texas, Illinois, Florida and New York saw the biggest absolute decreases. However even thought most of the airports reported decreases in the number of passengers, there are some showing increases (green edges).



Node	...	PAX diff	%PAX ...
MOT	2	2,981	442.94%
NTD	2	549	100.00%
WRI	2	1,141	100.00%
EGE	2	850	100.00%
PVU	2	309	100.00%
BGR	2	6,227	93.30%
ORH	2	2,064	83.46%
SPI	2	1,421	53.50%
MQT	2	1,075	29.90%
BTM	2	463	26.25%
Maine	1	-8,866	-15.48%
Montana	1	-27,382	-25.97%
North Dakota	1	-11,120	-28.94%
Utah	1	-223,162	-31.39%
California	1	-2,339,463	-33.63%
/	0	-18,206,...	-33.66%
Colorado	1	-524,915	-34.29%
Illinois	1	-1,177,438	-34.42%
Michigan	1	-576,733	-36.54%
New Jersey	1	-520,308	-38.02%
Massachusetts	1	-435,616	-42.24%

Node Details

MOT
673 vs 3,654

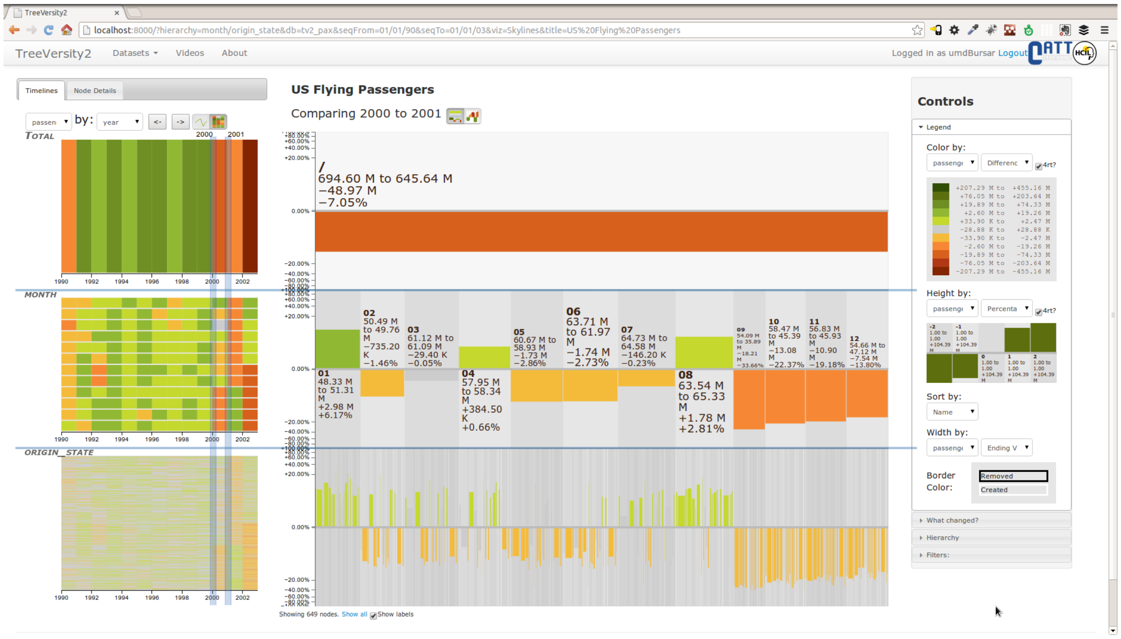
Enable Label ☑
Value format: Decimal
Name By: name
Show Name ☑
Show Value ☑
Show Percentage ☑
Shrink Label □
Filter label by level: 2
Font Size: 12

Comparing 2001 vs 2000 using PAX

Size by: PAX
Border by: Topological Difference
Color by: PAX

Size by: PAX differences
Border by: Topological Difference
Color by: PAX differences (%)

Figure 5.13: Airports increasing in more than 300 passengers in September 2001 compared to September 2000. Airports like MOT in North Dakota, didn't seem affected by 09/11, they reported an increase of 442% on their number of passengers, from 673 to 3,654.



TreeVersity2

localhost:8000/?hierarchy=month/origin_state&db=tv2_pax&seqFrom=01/01/90&seqTo=01/01/03&viz=Skylines&title=US%20Flying%20Passengers

Logged in as umdBursar Logout

US Flying Passengers

Comparing 2000 to 2001

Showing 649 nodes. Show all ☑Show labels

Controls

Legend

Color by: passenge Differenc ☑4rt?

🟩	+207.29 M to +455.16 M
🟩	+76.05 M to +203.64 M
🟩	+19.89 M to +74.33 M
🟩	+2.60 M to +19.26 M
🟩	+33.90 K to +2.47 M
⬜	-28.88 K to +28.88 K
🟨	-33.90 K to -2.47 M
🟧	-2.60 M to -19.26 M
🟧	-19.89 M to -74.33 M
🟥	-76.05 M to -203.64 M
🟥	-207.29 M to -455.16 M

Height by: passenge Percenta ☑4rt?

Sort by: Name

Width by: passenge Ending V

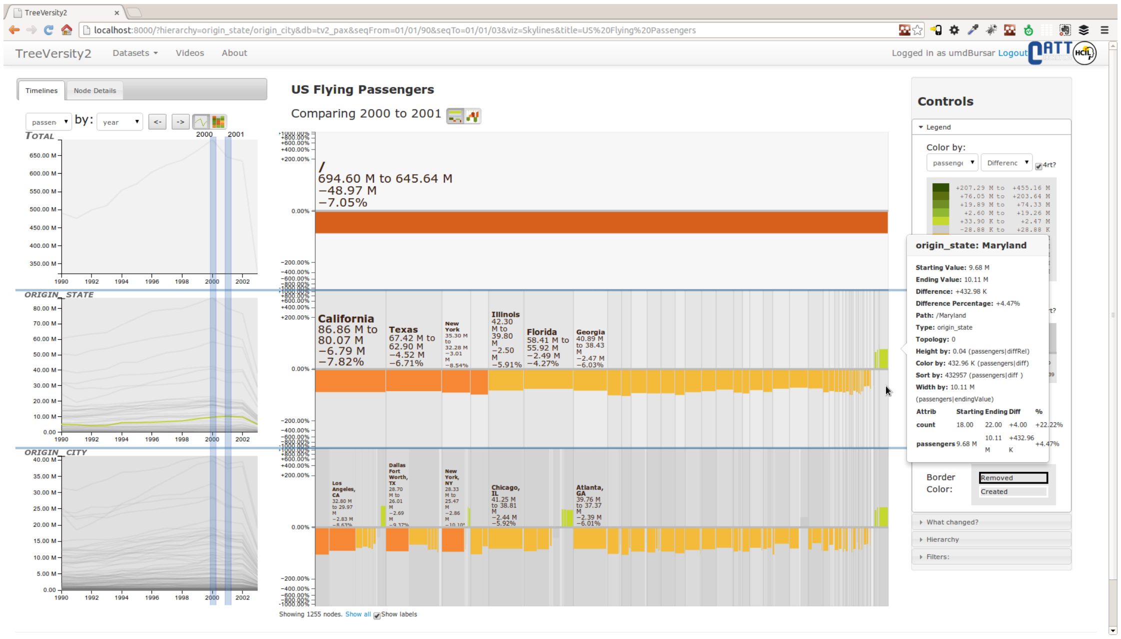
Border Color: Removed Created

What changed?

Hierarchy

Filters:

Figure 5.14: Change in the number of passengers among US States between 2001 and 2000 divided by month and by state. The time blocks show that 2001 presents the first decrease in the total number of passengers since 1990, and that in September 2001 there was a significant decrease on the number of passengers, a pattern that continued until September 2002



US Flying Passengers

Comparing 2000 to 2001

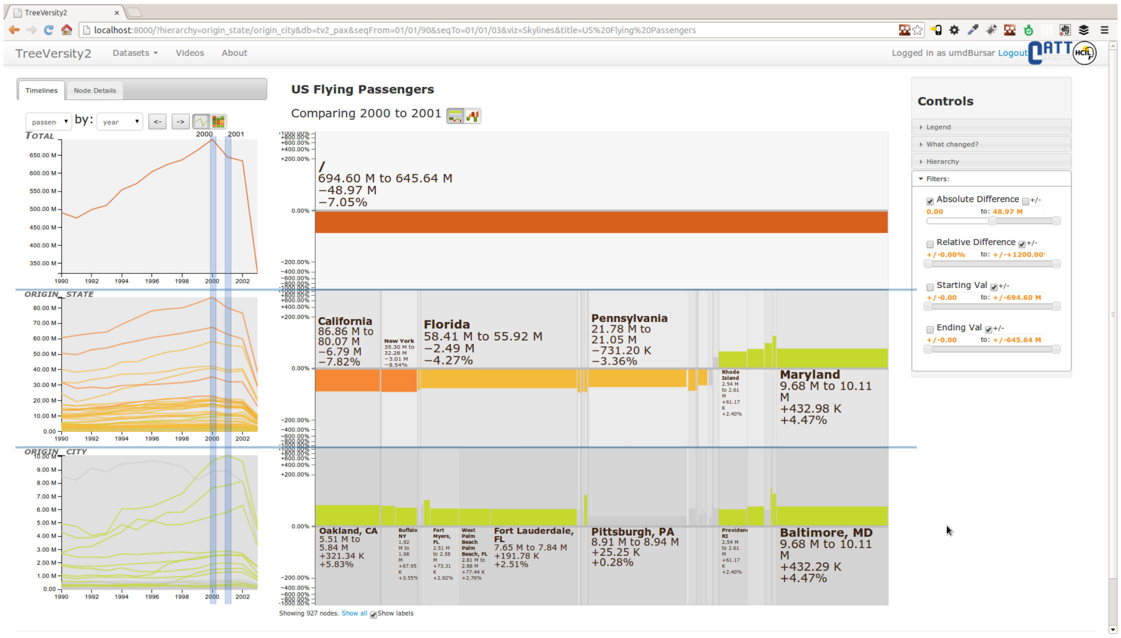
origin_state: Maryland

Starting Value: 9.68 M
Ending Value: 10.11 M
Difference: +432.98 K
Difference Percentage: +4.47%
Path: /Maryland
Type: origin_state
Topology: 0
Height by: 0.04 (passengers|diffRel)
Color by: 432.96 K (passengers|diff)
Sort by: 432957 (passengers|diff)
Width by: 10.11 M (passengers|endingValue)

Attrib	Starting	Ending	Diff	%
count	18.00	22.00	+4.00	+22.22%
passengers	9.68 M	10.11 M	+432.96 K	+4.47%

Showing 1255 nodes. Show all ☑ Show labels

Figure 5.15: Change in the number of passengers among US Airports between 2001 and 2000 divided by states. Nodes are sorted by their actual difference. The StemView shows how despite in 2001 there was a significant decrease on the number of passengers, some states such as Maryland actually reported increases.



localhost:8000/?hierarchy=origin_state/origin_city&db=tv2_pax&seqFrom=01/01/90&seqTo=01/01/03&viz=Skylines&title=US%20Flying%20Passengers

US Flying Passengers

Comparing 2000 to 2001

Controls

Filters:

☑ Absolute Difference □ +/- 0.00 to: 48.97 M

□ Relative Difference ☑ +/- +/-0.00% to: +/-+1200.00

□ Starting Val ☑ +/- +/-0.00 to: +/-694.60 M

□ Ending Val ☑ +/- +/-0.00 to: +/-645.64 M

Showing 927 nodes. Show all ☑ Show labels

Figure 5.16: Change in the number of passengers among US Cities between 2001 and 2000 divided by states, filtered by increasing cities only. This view shows that apart from Baltimore, other cities like Fort Lauderdale, Oakland and Pittsburg also reported increases in 2001

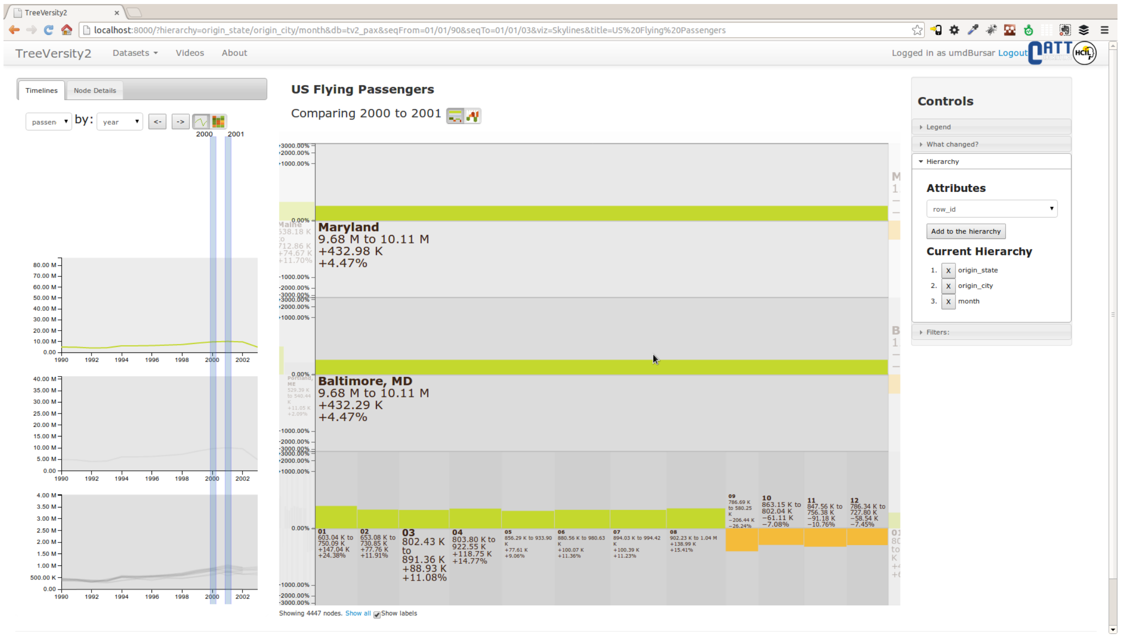


Figure 5.17: Change in the number of passengers for Maryland between 2001 and 2000 divided by cities and months. The StemView shows that Baltimore decreased their number of passengers after September 2001, but the overall sum gives a positive increase for the year.

5.6 Exit questionnaire

As part of the MILCS methodology, an exit questionnaire was distributed to the domain experts to gather their impressions on TreeVersity and TreeVersity2. The questionnaire [6] included open ended questions about the way they used TreeVersity or TreeVersity2 and the findings they made with it. It also included requests for permission to disclose the information collected during the case study and eight close-ended questions:

- q1: For this particular case study TreeVersity was:
- q2: In general the tool is likely to be:
- q3: Did the reporting tool help direct your exploration?
- q4: Did you find the StemView comprehensible?
- q5: Did you find the Bullet comprehensible?
- q6: Would you like to continue working with Treeversity?
- q7: Would you be willing to install and use Treeversity on your own?
- q8: How does this compare to your original expectations before starting with the tool.

The answers to these close ended questions where collected in a 7-point Likert scale, where the answers ranged from 1 being "Not useful at all" or "not comprehensible at all", to 7 being "Extremely useful" or "Extremely easy to comprehend". Users where allowed to skip questions, but only 4 questions were left unanswered. Figure 5.18 shows the boxplots of the answers for the 10 answers collected (some case studies were conducted with the same expert, so only one answer was collected for those). The answers showed mostly positive answers for most of the questions, with some outliers for the question

[6] The questionnaire can be found in this link: `https://docs.google.com/forms/d/1cfzl054M2kuB-nfJ6vI1iIuk0EoSA-9CxHTrFkGLfwk/viewform`

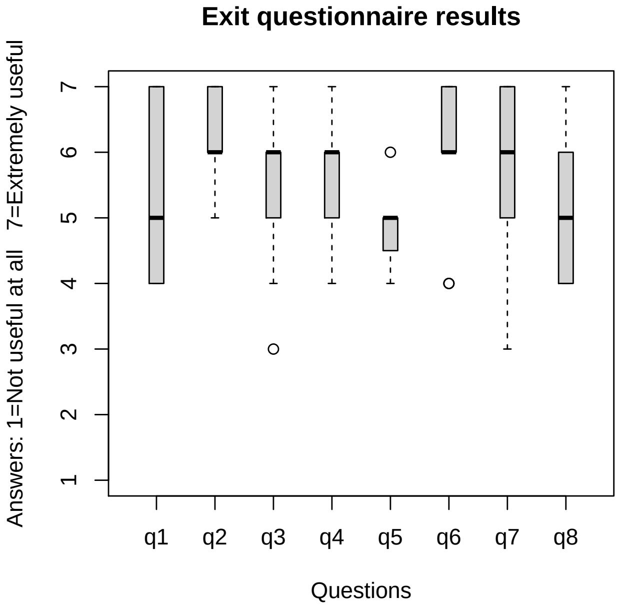


Figure 5.18: Exit questionnaire results for ten domain experts that completed it. Two experts collaborated with me in more than one case study and two did not complete the questionnaire.

about the reporting tool (q3) that may be due to the fact that not all the experts where exposed to it. A negative answer was also found on the question on the willingness to install TreeVersity on their own (q7), which is understandable for domain experts that are not necessarily computer experts. Answers also showed that experts found TreeVersity to be very useful (q2) and are willing to continue using it (q6). Moreover, experts found the StemView and Bullet comprehensible (q4, q5).

The open ended questions provided valuable feedback and quotes that are included on each of the case studies described on this chapter.

5.7 Summary

This chapter presented a sample of the 13 case studies conducted with TreeVersity and TreeVersity2, which validated the effectiveness of my techniques to help users find changes in datasets using hierarchies. For the sake of brevity, five case studies were thoroughly described (Sections 5.1, 5.2, 5.3, 5.4 and 5.5), information about the remaining studies can be found in Appendix B. In addition, the chapter described the exit questionnaire conducted with the domain experts (Section 5.6), which provided evidence that users found my approaches useful and understandable.

Chapter 6

Conclusions and Future Work

6.1 Conclusions

With the number of datasets being collected over time growing, finding what has changed in the data becomes an important problem. This dissertation presents three main contributions to this problem:

- The *Bullet* a visualization glyph for tree nodes that shows four characteristics of change: direction of change, absolute change, relative change and if the node was created or removed. Moreover, The development of *TreeVersity* a comparison tool to identify changes between two versions of a tree, that combines an implementation of the Bullet along with coordinated views and interactive filters to explore differences between two versions of a tree.

- The *StemView* an area based visualization artifact to show changes in all the nodes of a tree (including interior nodes) that represents five characteristics of change: direction of change, absolute change, relative change, starting or ending values, and created and removed nodes. The implementation of the StemView in *TreeVersity2,* a web based information visualization tool, that allows exploration of changes in datasets over time using hierarchies. Also, the implementation of a *reporting tool* that help users navigate outstanding changes in the tree with textual representations and coordinated interactions.

- The development of 13 case studies with domain experts on real world comparison problems validate the utility and flexibility of my approaches.

6.2 Future Work

This dissertation opened the door for follow up research ideas that could not being developed in this project, but that are worth pursuing. This section describes some of them.

6.2.1 Usability

TreeVersity and TreeVersity2 were developed for power users. However, some of the domain experts that worked with TreeVersity and TreeVersity2 to analyze their datasets expressed concerns on the many features that the tools offer, which seemed too challenging for the untrained user. Among other improvements, the reporting tool was developed to address this concerns, and users praised it as very useful to help navigating changes in their datasets. Further improvements on terms of usability and new features for the reporting tool could make TreeVersity and TreeVersity2 even more accessible to wider audiences.

6.2.2 Scalability

The current implementation of TreeVersity2 supports datasets of about 250,000 rows, with around 200 attributes, which generate trees with up to 7000 nodes and that can have thousands of time points. After these limits the interface starts being un-responsive and difficult to use. Despite all but one (the eBay product tree was bigger than 7,000 nodes when displayed with it's six levels) of the case studies described in this dissertation were under this limits, TreeVersity2 could be improved to support bigger datasets by optimizing the amounts of data transfered, using caches and redistributing the processing tasks between the server and the client.

6.2.3 Database caching

TreeVersity2 generates new SQL queries to the database every time the parameters of the tree are changed. These queries tend to be repetitive when users are performing reiterative analysis of the same dataset. Despite that, the database backend reprocesses the same queries over and over again. This could be improved by adding a caching layer on the database that sacrifice storage for efficiency, and will check if a query has already been calculated and return it from a cache instead of re computing it again. Being this a common problem for other domains, I searched for existing solutions that have already implemented this simple idea, but I couldn't find any out of the box solution for PostgreSQL.

6.2.4 Better support for outliers

During the development of the case studies it was common to find data ranges with extreme outliers (e.g. a data range where 95% of the data is in the $[-10, 10]$ range, but there are some outliers with values over 1000). TreeVersity2 allows the use of a quadratic root scale to emphasize small values in such data ranges, but a more customizable solution that allow users to specify the range they want to use for their data would be more beneficial.

6.2.5 Node selection

In some of our case studies, domain experts wanted to perform selection operation on the tree, such as analyzing only the Exports of US in the Interamerican Development Bank dataset. A full set of operations could be created to address this requests. These operations should support selection by filter and by manual pointing. Once the selections are in place, operations with them would help users too, such as union of selections, hiding the selected nodes, hiding the remaining nodes, etc.

6.2.6 Better labels

Appropriate labeling is always a challenge in dynamic visualizations. I implemented an adaptive labeling technique (described in Section 4.3.2) that draws labels on the StemView boxes whenever there is space available and that resizes the fonts to use as much space as possible. This technique could be improved even further by adding smarter label placing and user customizations.

6.2.7 Split numeric variables in groups

TreeVersity2 allows the creation of dynamic trees grouping data rows by its attributes. In the current implementation this is only useful if the attribute has a small number (around and less than 50) categoric values, for attributes with discrete ranges or bigger number of values it would be useful to implement grouping by deciles or quartiles.

6.2.8 Algorithm to find the best possible hierarchy

Users can use TreeVersity2 to find changes in their data using different hierarchies created by grouping by attributes in specific orders. By changing the order users can explore the differences in their data in different ways. It would be interesting to develop an algorithm to calculate the entropy of each hierarchy combination by calculating how many interesting outliers can be found on that ordering of attributes. The result of this algorithm could be incorporated on the reporting tool helping users find even more interesting changes on the data.

6.2.9 Better support for confidence intervals

Analysts at the FDA required that confidence intervals were considered when comparing their EBGM (Empiric Bayes Geometric Mean) index. Since this was the only case study to request that, and since TreeVersity2 wasn't designed to support confidence intervals, I implemented a hack to support their request. A new project could be started to address the problem of comparing indexes with confidence intervals using hierarchies building on top of the contributions of this work.

6.2.10 Reporting metric design tool

The current implementation of the Reporting tool includes only metrics handcrafted on the code. The architecture used to build the tool is flexible enough to create new metrics and add them to the tool, however to add new metrics users would have to modify TreeVersity2 source code. A GUI design application could be created to allow users define their own metrics by graphically selecting parameters to filter, or writing simple scripting code.

6.2.11 Direct connection to real datasets

TreeVersity2 current implementation uses a database that stores locally the data analyzed. The code base includes importing scripts that allow the incorporation of new datasets, however there is no support for direct connection to the original databases. Some of the domain experts in our case studies reported that they would like to have TreeVersity2 connecting directly to their databases and therefore avoid the importing process. Developing such connection in an efficient way would be an interesting research project.

Appendix A

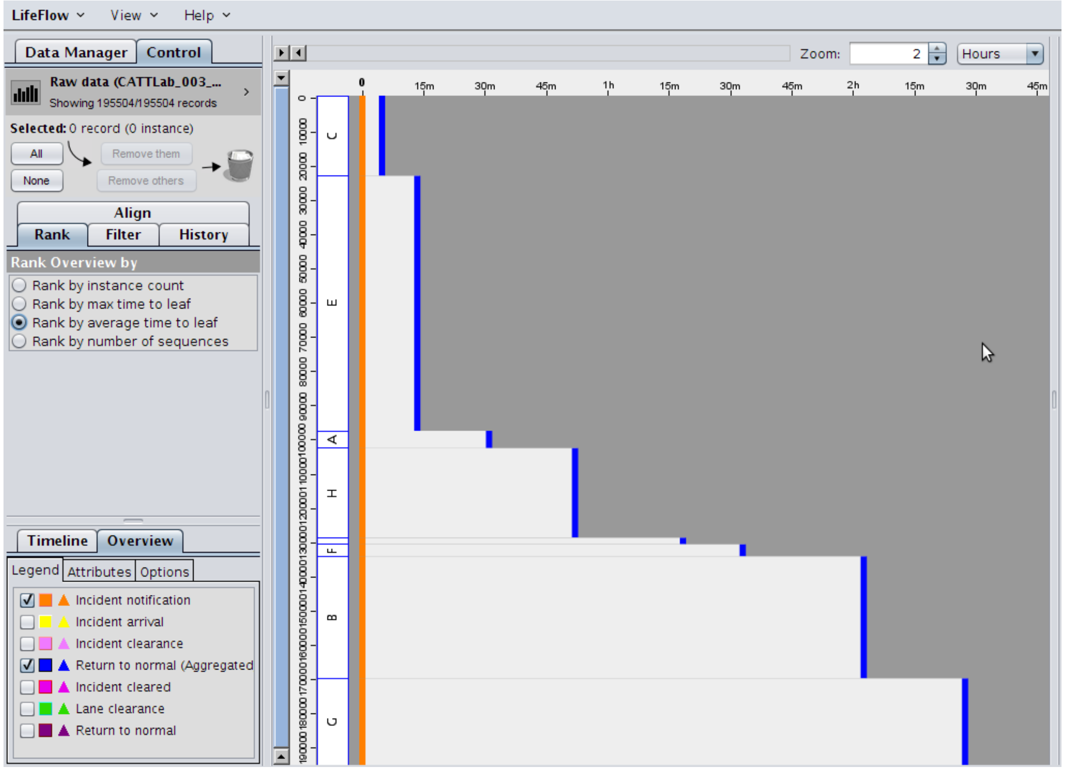
Early approaches

This chapter describes my previous work that led me to propose this dissertation. I start by describing my early work on LifeFlow a tool for visualizing summaries of temporal categorical data [36, 88], where I designed and implemented new features to include non-temporal data into the LifeFlow visualization. I used these improvements to illustrate the complexity of finding differences in hierarchical data by comparing the performance of eight traffic management agencies in the US. Then, I describe my initial approaches to the tree comparison problem with TreeVersity [38] where I identified the types of problems I will address in this dissertation, enumerated the characteristics of the differences that I wanted to illustrate, and explored alternatives to represent those characteristics. Next, I will recount the improvements I made on TreeVersity, thanks to the thoughtful advice in design of Professor Audra Buck-Coleman, which led to the design and implementation of TreeVersity v1.0 [39]. In this work, I performed a user study with 8 participants, that showed evidence that users could identify topological and node value differences between two small trees (50 nodes) using TreeVersity, without initial training and after minutes of usage. Finally, I outline my work on MySocialTree [37], a web implementation of TreeVersity used to navigate a user's Facebook News Feed that I built to explore the concept of comparing one evolving tree over time (or any other sequential variable).

A.1 LifeFlow comparison

My initial work on comparison and original inspiration for this research topic comes from my work on LifeFlow [36, 88]. LifeFlow is a visual analytics tools for temporal categorical data. It allows the understanding of large datasets by the creation of a summary of all the possible temporal sequences present on the data, and represent them using a modified Icicle tree, that can display both the temporal component and the number of records on each sequence. As an example of the type of analysis that can be performed on LifeFlow, it has been used to analyze all the different patterns that patients follow in a emergency unit in a hospital. This information includes hundreds of patients and several temporal events per patient describing their movement through the different areas in the hospital, like the Intensive Care Unit (ICU) and the Floor (normal care unit). One of the questions the doctors were able to answer using LifeFlow, was how many patients "bounce back" to the ICU, this is a sequence of ICU->Floor-> ICU.

My contribution to the project was the incorporation of non-temporal attributes to the visualization, that allows the categorization of the data by different parameters. After this, I developed new algorithms for sorting the different trees, according to the average or maximum time between events in them. Using these two improvements I was able to compare 8 different traffic agencies in their efficiency of clearing traffic accidents. I used a dataset from the National Cooperative Highway Research Program (NCHRP) that includes 203,214 traffic incidents from 8 agencies. I was able to rank the agencies by the data that I had, and I found some possible causes for the best and worst performers. Figure A.1 shows a screen capture of the LifeFlow analysis of the dataset, the orange bar represents when the agency was notified and the blue one when it cleared the scene. The x axis represents the time and the y axis the number of incidents attended by each agency. From the image is easy to see that agency C is the fastest agency, clearing the accidents in less than 10 minutes, while agency G is taking more than 2 hours in average to resolve their incidents. Further analysis of the type of incidents cleared by both agencies revealed that agency C had such a good average because most of their incidents were cleared at the exact same time they were reported (which might indicate a data entry problem), while for agency G they reported only two types of accidents, and while they were performing very well in one of them, they were taking more than 5 hours to clear the second type, as can be seen on Figure A.2.



Raw data (CATTLab_003_...
Showing 195504/195504 records

Selected: 0 record (0 instance)

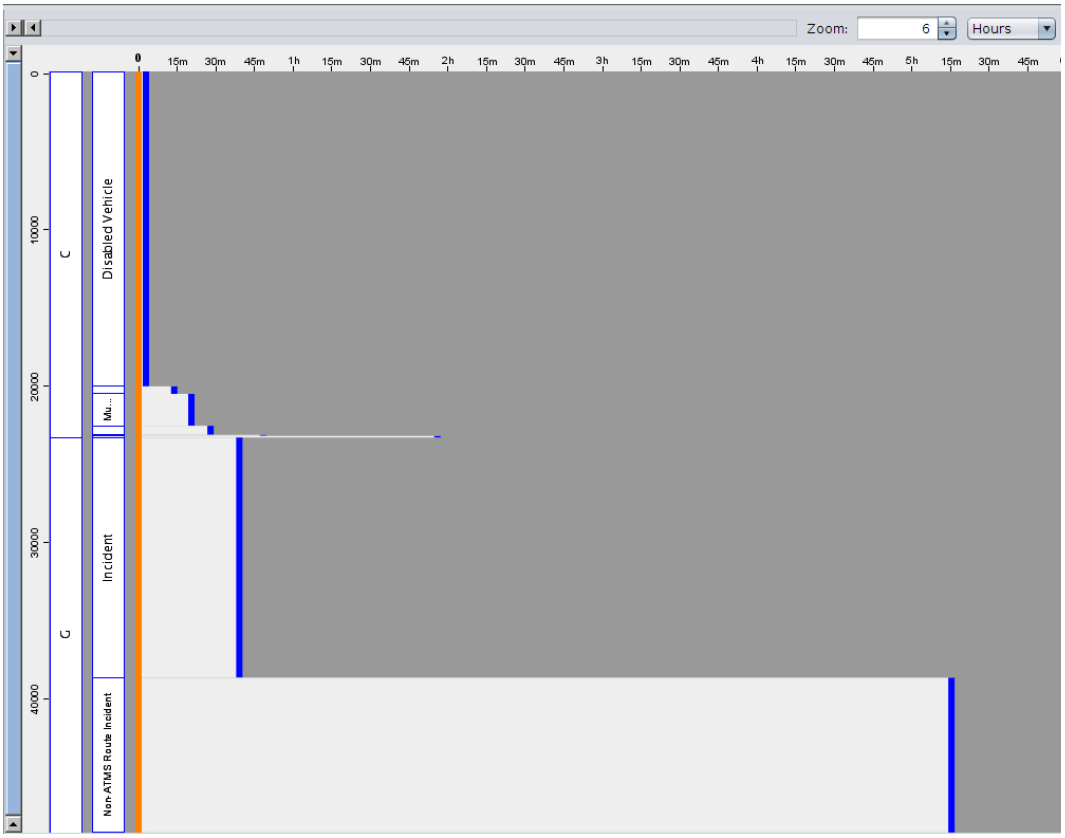
Zoom: 2 Hours

Rank Overview by

○ Rank by instance count
○ Rank by max time to leaf
● Rank by average time to leaf
○ Rank by number of sequences

☑ 🟧 Incident notification
□ 🟨 Incident arrival
□ Incident clearance
☑ 🟦 Return to normal (Aggregated
□ Incident cleared
□ 🟩 Lane clearance
□ 🟪 Return to normal

Figure A.1: Comparison of 8 different traffic agencies on their performance on clearing 203,214 traffic incidents across the USA.



Zoom: 6 Hours

Figure A.2: Detailed comparison of agencies C and G. The LifeFlow shows that most of the accidents attended by C are reported cleared right after they were notified which can explain the good overall result obtained by the agency, while in agency G the type "Non-ATMS Route Incident" is taking more than 5 hours to be cleared, which is affecting the agency average.

A.2 TreeVersity Initial approaches

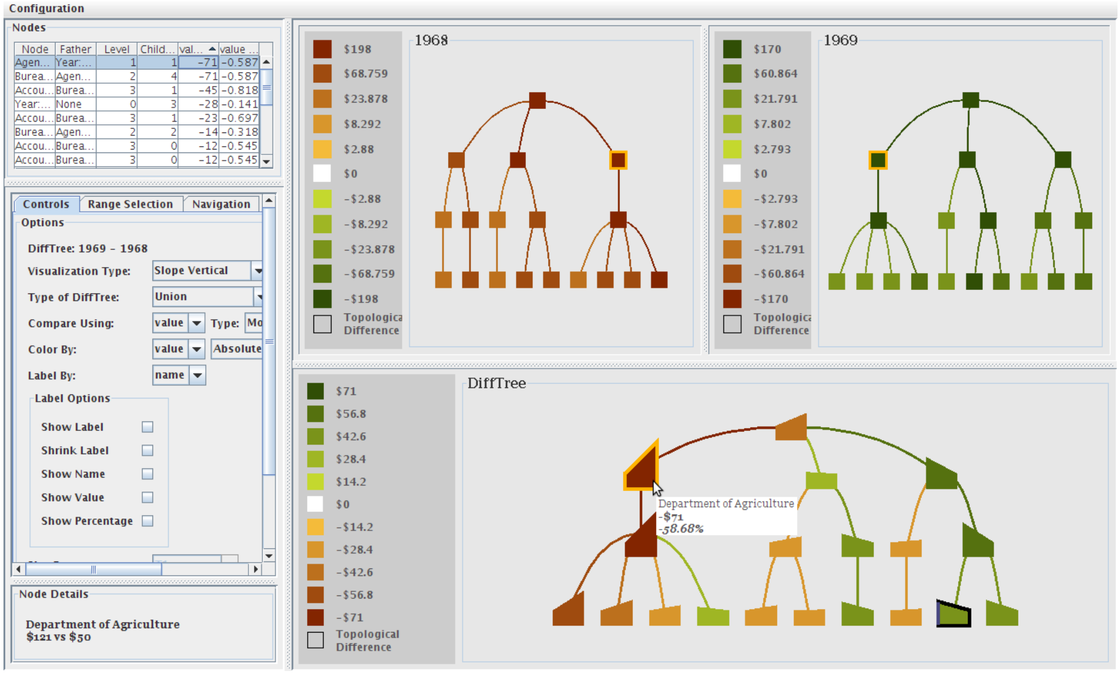
The type of comparison capabilities of LifeFlow was useful for the task at hand, but at the same time it reveals some limitations of the technique. For example, comparing more than two events at a time is very difficult in LifeFlow because of screen overcrowding. Because of this, I decided to build TreeVersity [38], a tool that specializes in comparing tree structures, both on topological changes and in node attribute values.

My main goal was to create an interactive visualization that allows the comparison of two trees by looking at: 1) Created and removed nodes. 2) Absolute and relative differences of the node attributes values. 3) Cardinality of the differences. 4) Differences in attributes of leaf nodes only or differences in attributes of interior nodes also. 5) Amount of change compared with the other nodes on the tree (or compared with the siblings).

For this, I have implemented a prototype of TreeVersity that uses a mixed approach for comparison. First, it presents a connected side by side comparison of the trees, that allows synchronized navigation and identification of unique versus created/removed nodes. Second, It displays an aggregation of both trees that represents the differences between the node attribute values of the trees, I call this the *"diffTree"*. For this I have been experimenting with two different visualizations, I have called them informally the "*slope*" and the "*gas tank*" approaches. In the next sections I describe them in more detail.

A.2.1 The Slope

The slope is a tree visualization based on a node link representation. It uses shape, color and size to represent the (absolute or relative) amount and direction of change. The shape uses a slope that can be decreasing (if the compared tree on the left is bigger) or decreasing (if the bigger is the tree on the right). The amount of slope is relative to the amount of change of each node compared with all the other nodes on the tree. The color represents the same amount of change using a gradient of tree user selectable colors.



Configuration

Nodes

Node	Father	Level	Child...	val... ▲	value ...
Agen...	Year:...	1	1	-71	-0.587
Burea...	Agen...	2	4	-71	-0.587
Accou...	Burea...	3	1	-45	-0.818
Year:...	None	0	3	-28	-0.141
Accou...	Burea...	3	1	-23	-0.697
Burea...	Agen...	2	2	-14	-0.318
Accou...	Burea...	3	0	-12	-0.545
Accou...	Burea...	3	0	-12	-0.545

Controls | Range Selection | Navigation

Options

DiffTree: 1969 – 1968

Visualization Type: Slope Vertical

Type of DiffTree: Union

Compare Using: value Type: Mo

Color By: value Absolute

Label By: name

Label Options

Show Label □

Shrink Label □

Show Name □

Show Value □

Show Percentage □

Node Details

Department of Agriculture
$121 vs $50

Figure A.3: Example of the slope visualization representing a subset of a made up Federal Budget. The image shows the comparison of years 1968 vs 1969, where a cut of 58% was made on the "Department of Agriculture". Red nodes represent cuts, while green ones represent increases. The node with the black border represents a created node (topological difference).

The topological differences are also represented in this visualization. I used a special marker (a different color and a line) on the nodes to differentiate created and removed nodes from the others. An example of this visualization is shown on Figure A.3.

I believe the slope visualization is good to recognize the changes in all the levels of the tree, to identify the topological changes and to understand the structure of the aggregating tree. However, it works better with smaller trees, and can only represent one magnitude of change at a time (either absolute or relative). Because of this, I developed a complementary visualization that is presented in the next section.

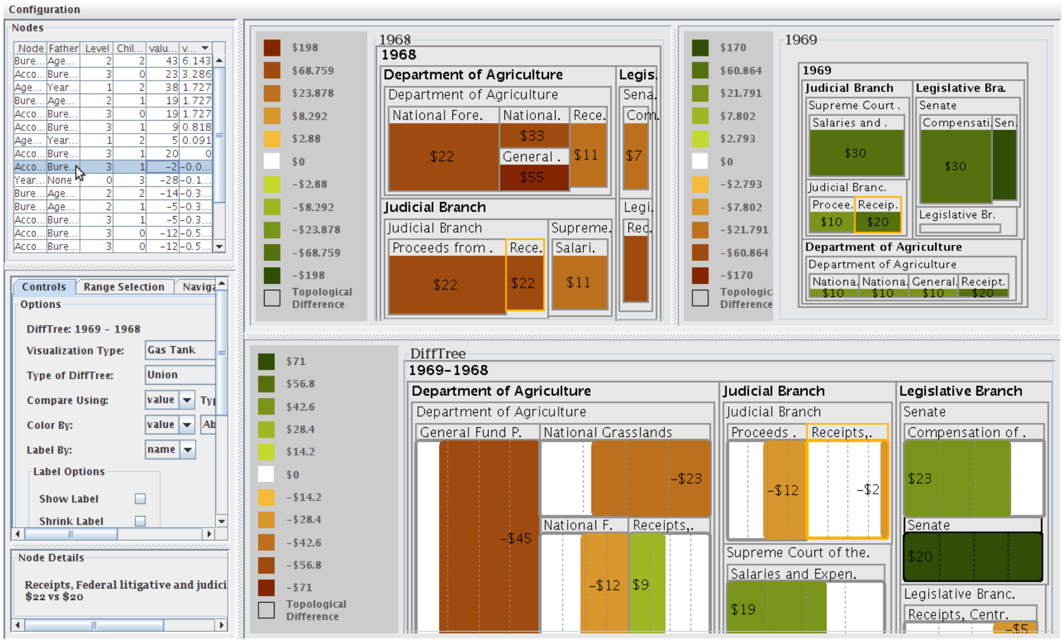
A.2.2 The Gas Tank

The "gas tank" representation uses a space filling approach based on Treemaps [50] to represent changes in the leaf nodes of the diffTree, displaying at the same time the absolute and relative amounts of change. It is especially good to highlight the "biggest players" on the diffTree (nodes with the biggest values overall). To create the gas tank representation I first take the two individual Treemap nodes representation, I combine them and obtain the difference, and finally I draw the difference as a filling portion of the area of the biggest of the original nodes. This way I avoid nodes of size zero. The process is illustrated on Figure A.4 on the left. I then use these node representations of the leaves and aggregate them in the same way Treemaps does to represent the hierarchy. Although the gas tank visualization only represent the changes on the leaf nodes, TreeVersity has a control that allows the selection of a level of the tree to represent, that way the gas tank diffTree is redrawn to represent only nodes in that level (or leaves on previous levels). An example of the gas tank representation with the same artificial budget data is shown in Figure A.4 on the right.

Like the slope visualization, the gas tank also has its strong and weak points. It is especially useful to compare the absolute and relative changes in the biggest nodes, but it isn't so helpful to represent the hierarchy of the tree, and it hides information like the amount of change in the interior nodes.

A.3 MySocialTree: browsing the Facebook feed using hierarchies [37]

Looking for new application domains for TreeVersity and wanting to start exploring the concept of comparing an evolving tree structure over time, I decided to build a Facebook feed navigation tool using hierarchies for my CMSC838C Social Computing course called MySocialTree. The goal of this project was to build a web application that displayed the updates in the user's Facebook feed, by exploring them using a user



Configuration

Nodes

Node	Father	Level	Chil...	valu...	v...
Bure...	Age...	2	2	43	6.143
Acco...	Bure...	3	0	23	3.286
Age...	Year...	1	2	38	1.727
Bure...	Age...	2	1	19	1.727
Acco...	Bure...	3	0	19	1.727
Acco...	Bure...	3	1	9	0.818
Age...	Year...	1	2	5	0.091
Acco...	Bure...	3	1	20	0
Acco...	Bure...	3	1	-2	-0.0...
Year...	None	0	3	-28	-0.1...
Bure...	Age...	2	2	-14	-0.3...
Bure...	Age...	2	1	-5	-0.3...
Acco...	Bure...	3	1	-5	-0.3...
Acco...	Bure...	3	0	-12	-0.5...
Acco...	Bure...	3	0	-12	-0.5...

Controls | Range Selection | Naviga

Options

DiffTree: 1969 - 1968

Visualization Type: Gas Tank

Type of DiffTree: Union

Compare Using: value

Color By: value

Label By: name

Label Options

Show Label □

Shrink Label □

Node Details

Receipts, Federal litigative and judici
$22 vs $20

Figure A.4: Gas Tank Visualization representing the data from Figure A.3. It shows the absolute amount of change using color and a label, and the relative difference using the amount of space filled in each node. It also shows the topological differences using again a black border. This view also allows the selection of a node along all the compared trees and the diffTree to see its details, like it was done with the "Receipts, Federal litigative and judiciary" that decreased only on $2 from 1968 to 1969.

defined hierarchical criteria. A common example of this criteria (that was used as default setting in the application) would be to classify the feed by poster's membership to the user's lists (Close Friends, Family, School, etc) and the type of post (Photo, Status Update, Video, Link, Activity, etc). By grouping the posts in a tree, users can find the posts they care the most by clicking on the nodes of the tree, that answers queries like "show me photo updates from my close friends". Moreover, since the tree will show the positive and relative amount of change in each one of its nodes, users should be able to identify trends like "there are a lot of status updates from my university friends". This section describes the design and implementation of MySocialTree, as well as presents a small user survey that seems to suggest that it is an effective and easy to understand tool for finding relevant posts in a user's feed. A running version of the tool can be found at: `http://mysocialtree.appspot.com`.

Given the time constraints of the class project, the initial implementation of MySocialTree doesn't include the tree comparison component, however this is described and designed in the paper, and is part of the future work of the project.

MySocialTree has two main components, first a categorization technique that organizes the posts as nodes on a tree, and second a visualization tool to display the nodes highlighting the number of posts or the change against the average rate of posts per unit of time. In this section I describe in more detail those components:

A.3.1 Encoding the feed as a tree

MySocialTree uses trees to organize a user's social feed. These trees are created by categorizing the items in the feed (Posts or Tweets) using classifiers according to certain user criteria, like friend's lists and type of Facebook's posts. A classifier is a function $C(p)$ that maps each post to a number of 1 or more classes. The user selects the desired criteria by choosing and ordering a list of classifiers, like *Type of Post*, *List membership of the poster* and *Location*. Being p a post. The user criteria is therefore just an ordered list of classifiers $\{C_1, C_{2,}..., C_n\}$. The initial implementation of MySocialTree uses the fixed criteria $\{ListMembershipOfPoster, TypeOfPost\}$, future implementations should

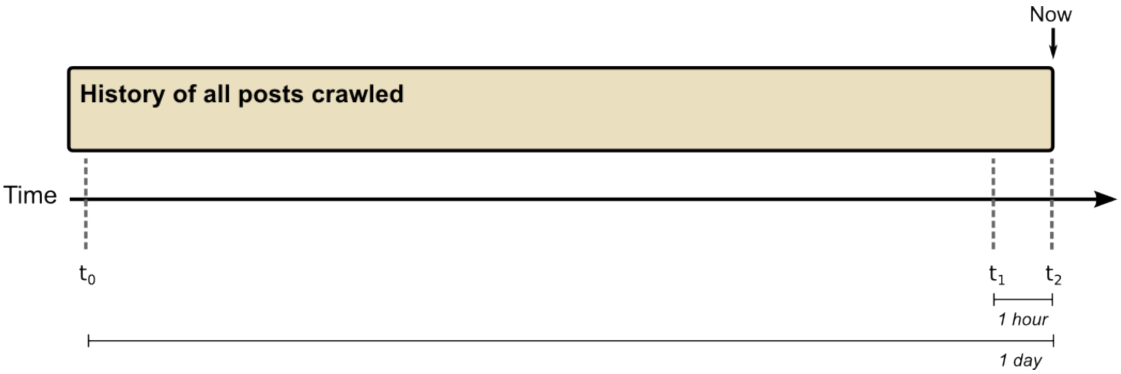


Figure A.5: Timeline used for the example of how to calculate the diffTree in MySocialTree

include an user interface for selecting and ordering the classifiers.

Currently MySocialTree represents only the tree corresponding to all the posts in the user feed for a certain period of time T_{count}, however the final objective of this project is to display the change of number posts compared to the average rate $T_{diff} = T_{count} - T_{avg}$. Let's explain this in more detail with an example: let's say than a user want to see her T_{diff} for the last hour$\Delta t = 1hr$, then, the application calculate $T_{avg}(\Delta t, t_0, t_2)$, the tree of the average number of posts in each of the categories of the hierarchy for the time window Δt in the period of time between $t_o$and t_2, where $t_o$and t_2 are usually the time frame of the total history of posts stored in the database for the user, or a constant frame bigger than Δt (e.g. if $\Delta t = 1hour$ then $t_2 - t_0$ can be $1\,day$). Then the application calculates $T_{count}(t_1, t_2)$ (The tree of the counts of the number of posts in each of the categories of the hierarchy for the period of time $t_1$to t_2, where $t_2 - t_1 = \Delta t$). The diffTree in MySocialTree is the difference between $T_{diff}(\Delta t, t_0, t_1, t_2) = T_{count}(t_1, t_2) - T_{avg}(\Delta t, t_0, t_2)$, where the difference is calculated node to node. Figure A.5 shows the timeline used for the example.

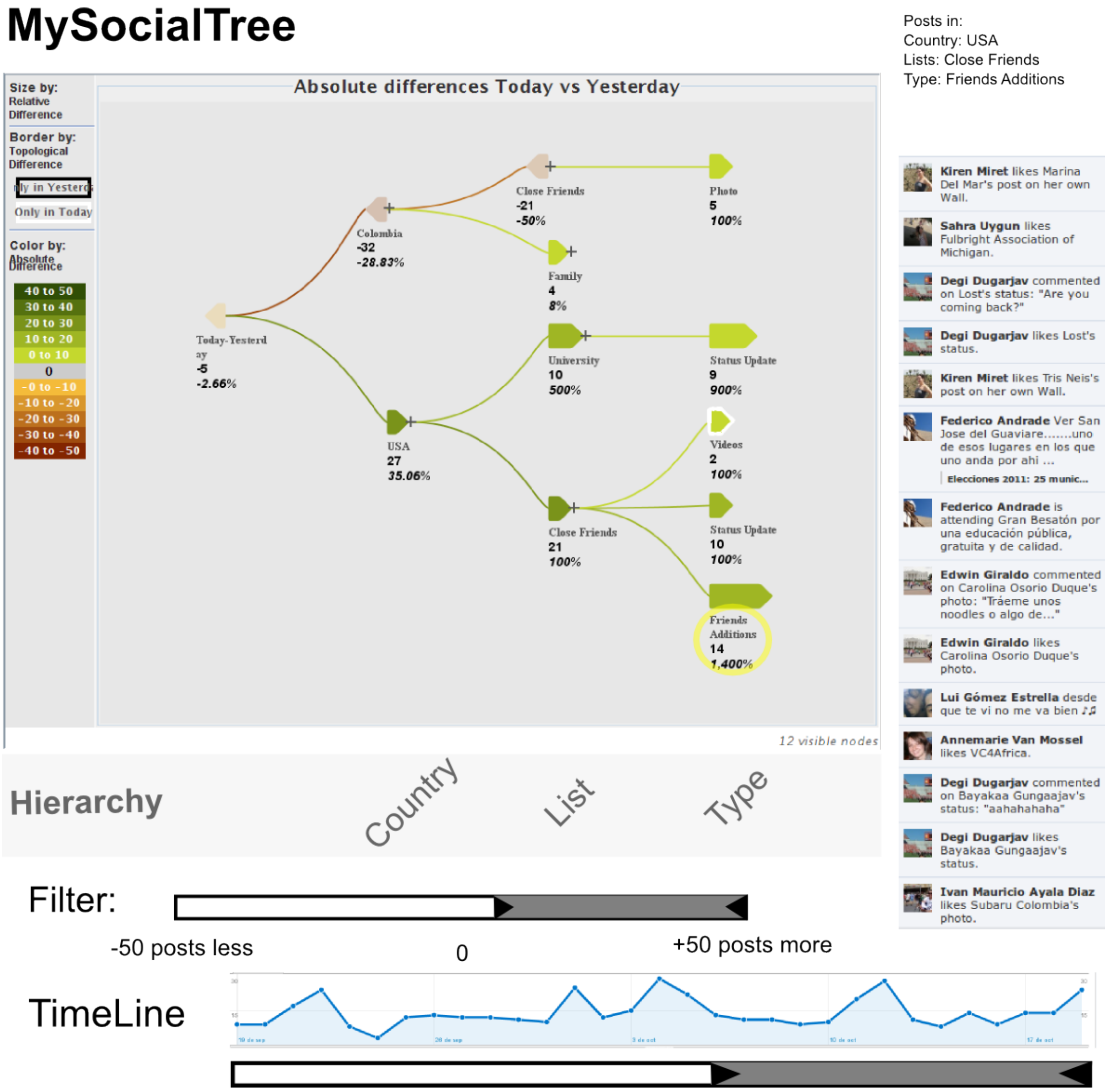
For an example using numbers, suppose we want to calculate MySocialTree for $\Delta t = 1\,hour$. Then we need to count the number of posts in the last hour ($T_{count}(t_1, t_2) : t_2 = now, \; t_1 = 1\,hour\,ago$) for each category in the hierarchy, say is 15 *Friends Additions* for my *Close Friends* in *USA*. Now, to compare to the history we calculate the average number of posts in that same category, in periods of time of 1 hour for the last day

($T_{avg}(\Delta t, t_0, t_2)$ $t_0 = 1\, day\, ago$), say it is 2. Then, in difference MySocialTree there is going to be a node with path */USA/Close Friends/Friends Additions* with value $15 - 1 = 14$ that shows an increment of 1,400% as highlighted in the Figure A.6.

A.3.2 Visualizing the tree

MySocialTree visualize the trees using TreeVersity [38, 39], an interactive visualization technique for tree comparison. TreeVersity can be used to visualize topological changes and node value differences between two versions of a tree. It uses a modified node-link representation enhanced with an special glyph called The Bullet that encodes the cardinality of the change, and the amount of change. The shape's direction represents the cardinality of the change: left for negative and right for positive in the horizontal layout, and down for negative and up for positive in the vertical layout. The bullet size represent the amount of change. Color is used to encode both the cardinality and amount of change in the nodes. To accommodated for color-blind users can select from preset color palettes that are binned in five steps to ease differentiation. Gray rectangles represent nodes where the amount of change is zero. The bullet can represent also topological differences by means of different border colors on the nodes, but this feature was not used in MySocialTree. By default, both size and color are redundantly encoding the absolute amount of change (e.g. the amount in dollars in the case of a budget), but users can switch to relative change (i.e. percent change), or assign color and size to different characterization of the changes.

Each node in the tree displays its current name, that is the combination of the classes assigned by the classifiers in the hierarchy criteria (e.g. the node */Close Friends/Photos* contains all of the photos posted by users in the *Close Friends* list). The label also includes the number of posts that fell into that node. Figure A.8 shows a screen shot of the initial implementation of MySocialTree that uses T_{count} only (therefore, no negative values are displayed), the mock-up on Figure A.6 shows the concept using T_{diff}.



Posts in:
Country: USA
Lists: Close Friends
Type: Friends Additions

Kiren Miret likes Marina Del Mar's post on her own Wall.

Sahra Uygun likes Fulbright Association of Michigan.

Degi Dugarjav commented on Lost's status: "Are you coming back?"

Degi Dugarjav likes Lost's status.

Kiren Miret likes Tris Neis's post on her own Wall.

Federico Andrade Ver San Jose del Guaviare.......uno de esos lugares en los que uno anda por ahi ...

Elecciones 2011: 25 munic...

Federico Andrade is attending Gran Besatón por una educación pública, gratuita y de calidad.

Edwin Giraldo commented on Carolina Osorio Duque's photo: "Tráeme unos noodles o algo de..."

Edwin Giraldo likes Carolina Osorio Duque's photo.

Lui Gómez Estrella desde que te vi no me va bien ♪♫

Annemarie Van Mossel likes VC4Africa.

Degi Dugarjav commented on Bayakaa Gungaajav's status: "aahahahaha"

Degi Dugarjav likes Bayakaa Gungaajav's status.

Ivan Mauricio Ayala Diaz likes Subaru Colombia's photo.

Figure A.6: MySocialTree idea mock-up

A.3.3 Architecture

MySocialTree was developed using a combination of tools that include: the Google App Engine [5], the Facebook Developers Platform [3] and the Python SDK for it [4] with some modifications[1], the Data Driven Document framework D^3 [2, 12], and Bootstrap from Twitter [1]. Putting all of this pieces to work together was a big technical challenge, but each one of them gave the application different qualities. The whole application was built for the Google App Engine, which gave it the scalability and robustness of Google's technologies. The direct connection with the Facebook Platform, allows users to access their feeds directly without the need of external importers. The use of D^3 allowed the creation of the tree visualization and most of the JavaScript interactions, while using open web standards that offers portability. And finally Bootstrap from Twitter gave it a nice look and feel from the beginning.

A running version of the application should be available on `http://mysocialtree.appspot.com`. Figure A.7 presents an overview of the architecture of the application. The user is faced with a main view with one main use case show the *MySocialTree,* depending of the state of the system this can trigger three actions *showTree, showPosts* and *crawlPosts.* The *crawlPosts* actions triggers the connection to Facebook to obtain the latest posts in the feed, that are later stored in a table. This table is used to display the feed to the user and to calculate the trees using the current hierarchy (initially Facebook Lists then Posts Type). The trees are then stored in another table that is crawled later for the tree comparison and the final display of the trees.

A.3.4 Evaluation

MySocialTree included a small survey requesting feedback from users, that with the following questions:

[1] The original version of the Python SDK contained some bugs that were corrected. Also, in the middle of this project Facebook enforced OAuth login only for the apps, and since this feature wasn't available in the original SDK, the functionality was added

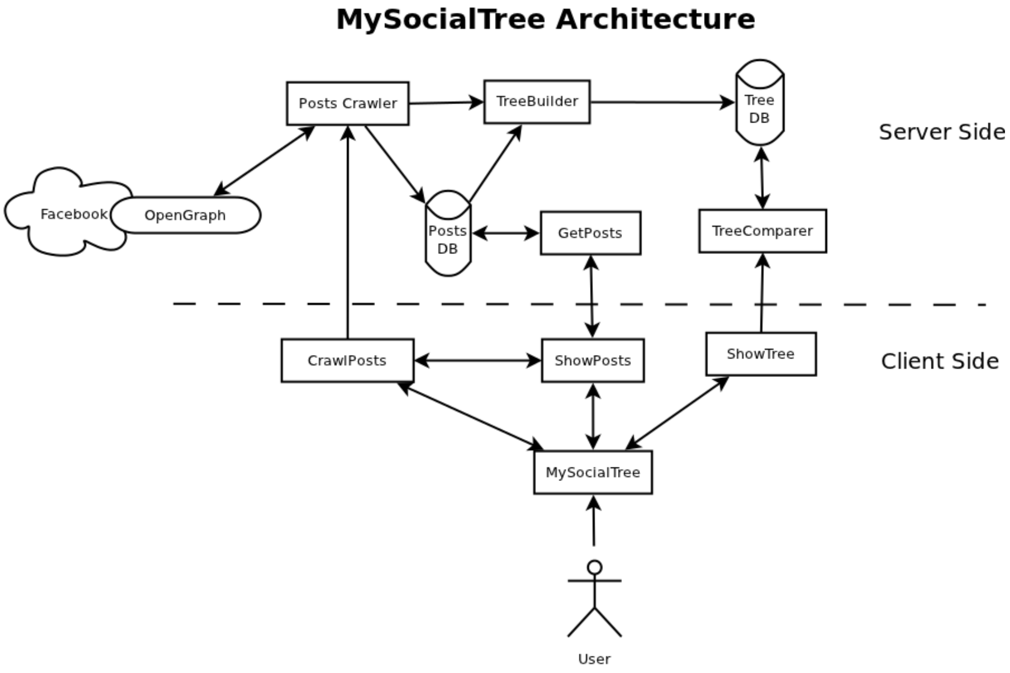
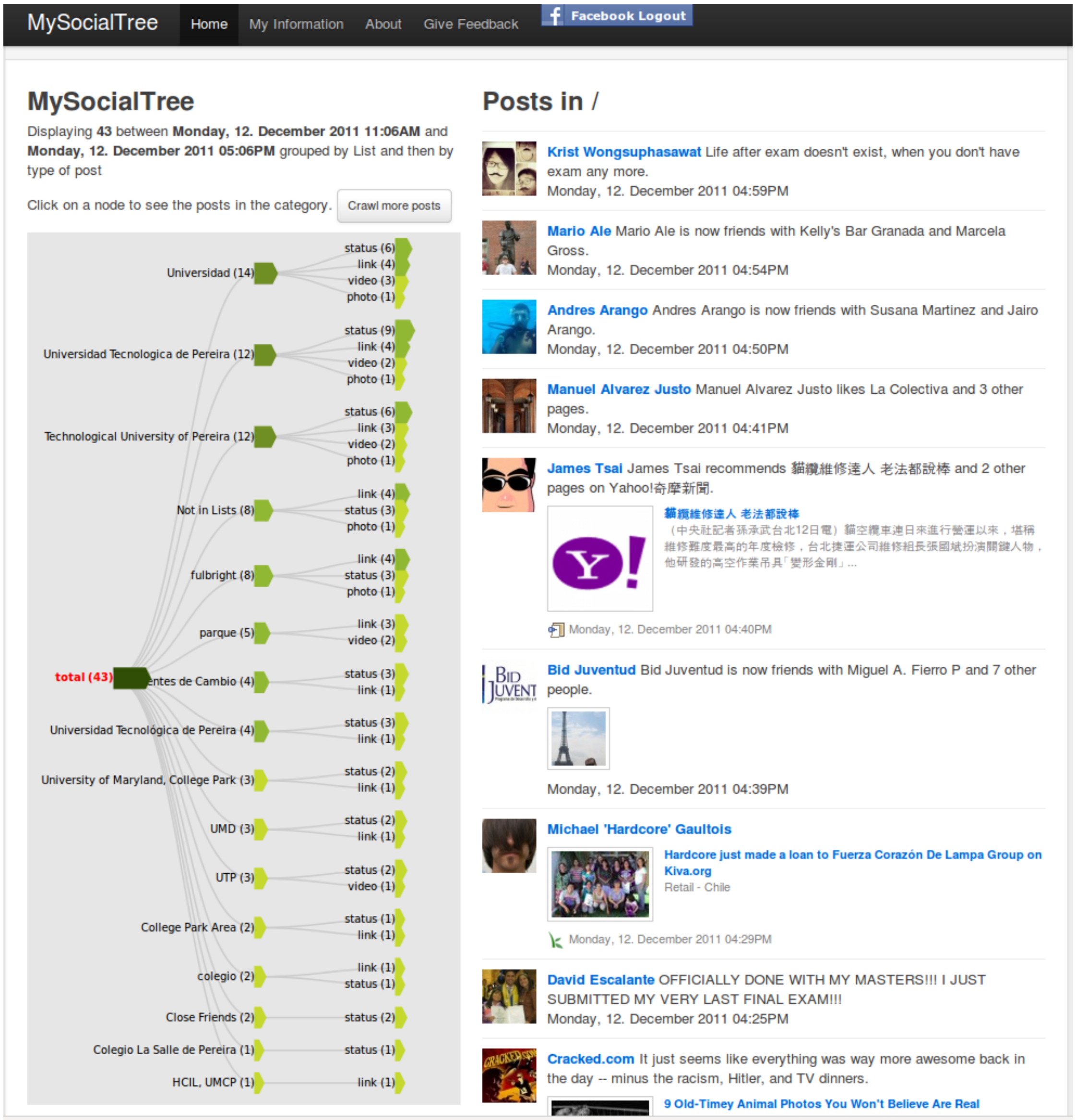


Figure A.7: MySocialTree Architecture

- q1: How easy is for you to find the posts that matter the most to you on the traditional Facebook Feed?
- q2: How useful do you think MySocialTree can be for organizing and browsing your Facebook posts?
- q3: How easy was it for you to understand and navigate the tree?
- q4: How likely would you be of using this MySocialTree in the future to navigate your Facebook Feed?
- q5: Any other comments, suggestions, bugs?

The answers were rated in a 7 point Likert-scale, with the answers ranging from 1: *Very difficult/Not useful/Not likely* to 7: *Very easy/Very useful/Very Likely.* The survey was completed by 15 of the 47 users of the application between December 9th and 13th 2011, their responses are summarized in the Figure A.9. According to the responses, q1 seems



MySocialTree

Displaying **43** between **Monday, 12. December 2011 11:06AM** and **Monday, 12. December 2011 05:06PM** grouped by List and then by type of post

Click on a node to see the posts in the category.

Posts in /

Krist Wongsuphasawat Life after exam doesn't exist, when you don't have exam any more.
Monday, 12. December 2011 04:59PM

Mario Ale Mario Ale is now friends with Kelly's Bar Granada and Marcela Gross.
Monday, 12. December 2011 04:54PM

Andres Arango Andres Arango is now friends with Susana Martinez and Jairo Arango.
Monday, 12. December 2011 04:50PM

Manuel Alvarez Justo Manuel Alvarez Justo likes La Colectiva and 3 other pages.
Monday, 12. December 2011 04:41PM

James Tsai James Tsai recommends 貓纜維修達人 老法都說棒 and 2 other pages on Yahoo!奇摩新聞.

貓纜維修達人 老法都說棒
（中央社記者孫承武台北12日電）貓空纜車連日來進行營運以來，堪稱維修難度最高的年度檢修，台北捷運公司維修組長張國斌扮演關鍵人物，他研發的高空作業吊具「變形金剛」...

Monday, 12. December 2011 04:40PM

Bid Juventud Bid Juventud is now friends with Miguel A. Fierro P and 7 other people.

Monday, 12. December 2011 04:39PM

Michael 'Hardcore' Gaultois

Hardcore just made a loan to Fuerza Corazón De Lampa Group on Kiva.org
Retail - Chile

Monday, 12. December 2011 04:29PM

David Escalante OFFICIALLY DONE WITH MY MASTERS!!! I JUST SUBMITTED MY VERY LAST FINAL EXAM!!!
Monday, 12. December 2011 04:25PM

Cracked.com It just seems like everything was way more awesome back in the day -- minus the racism, Hitler, and TV dinners.

9 Old-Timey Animal Photos You Won't Believe Are Real

Figure A.8: MySocialTree screen shot

to indicate that users find it difficult to identify the important posts in the traditional Facebook news feed (75% of the responses where less than 5, between somehow easy and neutral). According to q2 and q3 most users (more than 75%) found MySocialTree useful (only one person responded less than 4) and easy to understand, while on q4 more than 50% of the users expressed that they would be likely to use the application in the future.

No personal information was collected from the participants, however since the announcement of availability of the application and the call for participation in the survey was made through the author's personal Facebook page, and different university email lists, most of the participants can be assumed to directly know the author and therefore might have some type of bias in favor of this work.

The survey also included an open ended question, requesting more feedback and comments. This question prove to very informative, because participants use it to express experience with MySocialTree in a more open way. The responses for this question were overall positive and constructive. Participants submitted all sorts of comments, ranging from the very excited and supportive of MySocialTree like: *"Great idea!"*, *"...it was *really* nice to have that news feed grouped in sensible way! It was fantastic to be able to tease out the posts..."* and *"As a visualization, I think MySocialTree is neat and pretty. It is much more organized than a simple list of posts...",* to the very frustrated like: *"...the category nodes were not useful for me.", "...I am not sure that it is offering me anything different than what Facebook already provides through its list views."* and *"...I'm not sure exactly what this does :-P".* The wide range of responses received helped in a way to validate the results, and dismisses some of fear of that the participants might be biases towards the applications.

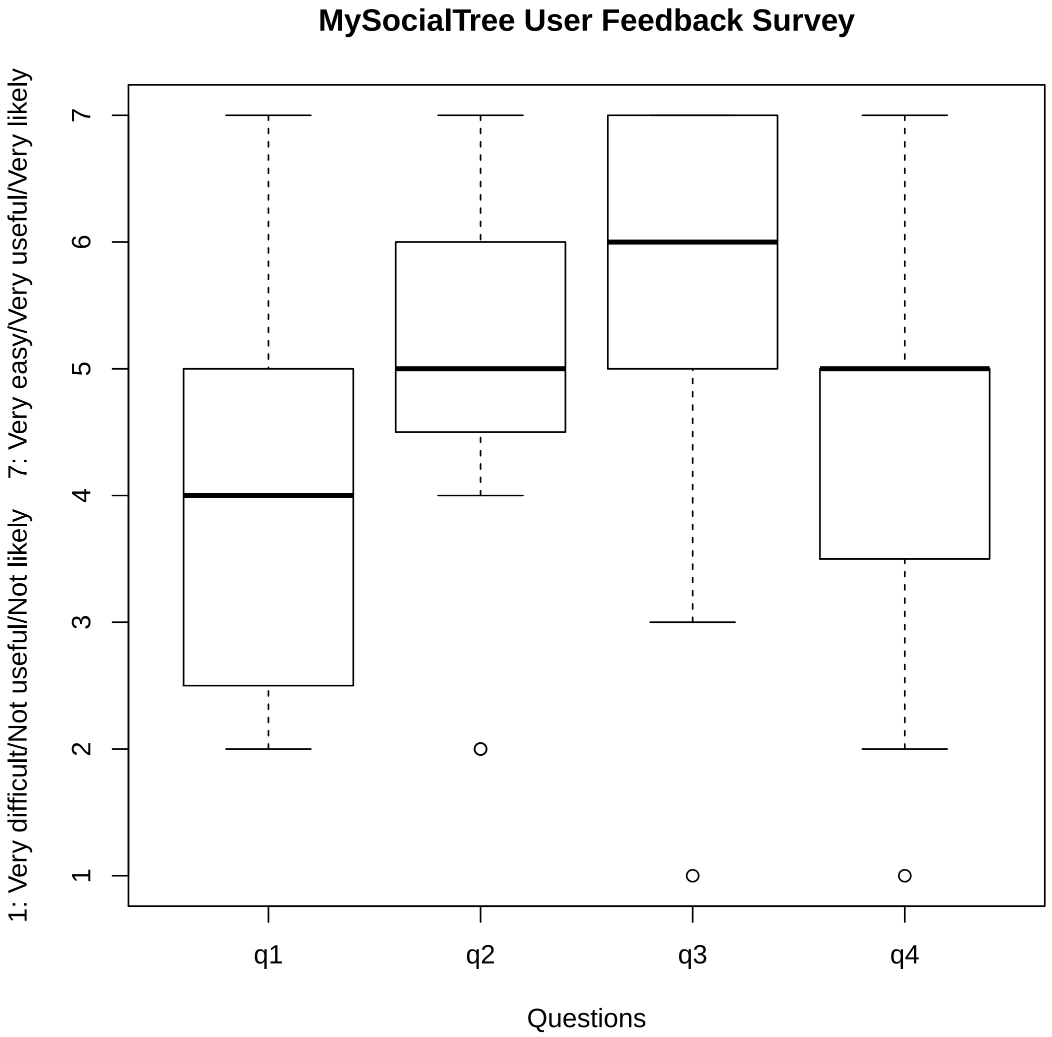


Figure A.9: MySocialTree user survey responses. Includes the feedback of 15 participants that responded four questions about their perception of MySocialTree.

Appendix B

Other Case Studies

B.1 Transportation Bottleneck data

B.1.1 Case Study Sheet

Partner: Michael L. Pack, Lab Director

Organization: Center for Advanced Transportation Technologies (CATT Lab)

MILCS level: Early, self driven

Duration: March 2013 to April 2013 (2 months, virtual meetings)

System used: TreeVersity2

Data: Traffic log data for six states between 2010 and 2011

Number of rows: 96,205

Number of time points: 24 (months)

Example tree size: 286 (3 Levels)

Number of numeric variables: 4 (average duration, average length, impact factor, occurrences)

Number of Attributes: 7 (Direction, Longitude, Latitude, Location, State, County, Road Class)

Type of Hierarchy: Mixed

Type of comparison: Type 3: aggregating + different topology

URL: `https://treeversity.cattlab.umd.edu/cs/catt_bottlenecks`

B.1.2 Discussion

***Disclosure**: The CATT Lab partially founded my research during the last years of my PhD, and therefore Mr. Pack was my direct manager.*

In this case study, Mr. Michael Pack the director of the Center for Advanced Transportation Technologies (CATT Lab) used TreeVersity2 to analyze changes in traffic congestion levels over time for various states, counties, roads, etc. The CATT Lab has developed visualization tools that allow users to explore congestion levels in single geographies with aggregates for data ranges, but they don not have tools to analyze change over time. Mr. Pack wanted to use TreeVersity2 to explore trends and perform change analysis.

The data contained 96,205 rows representing aggregated traffic congestion bottlenecks for six U.S. States between 2010 and 2011. The data also included geographical information, as well as details of the roads where the bottlenecks occurred. The numeric variables included the average duration and length of the bottlenecks, as well as the number of occurrences and a measure of its impact.

Mr. Pack demonstrated TreeVersity2 in a meeting at the North Carolina's Department of Transportation to several analysts and engineers. Not being an expert TreeVersity2 user himself, Mr. Pack was amazed to find that with his guidance, meeting participant were able to identify within minutes a "*major long-term congestion event*" that significantly affected the congestion statistics for North Carolina. The meeting participants then started asking follow up questions to dig deeper into the data, and were happy with what they found. The only concerns the analysts found with TreeVersity2 on the meeting was that since the tool was not specifically designed to represented traffic data, the labels didn't included proper units of metric.

The CATT Lab will continue the development of TreeVersity2, and will support the tool to some of the case studies developed in this thesis.

B.2 US Federal Budget

B.2.1 Case Study Sheet

Partner: David Rowe, Education Branch Chief, drowe at omb.eop.gov

Organization: at Office of Management and Budget (OMB)

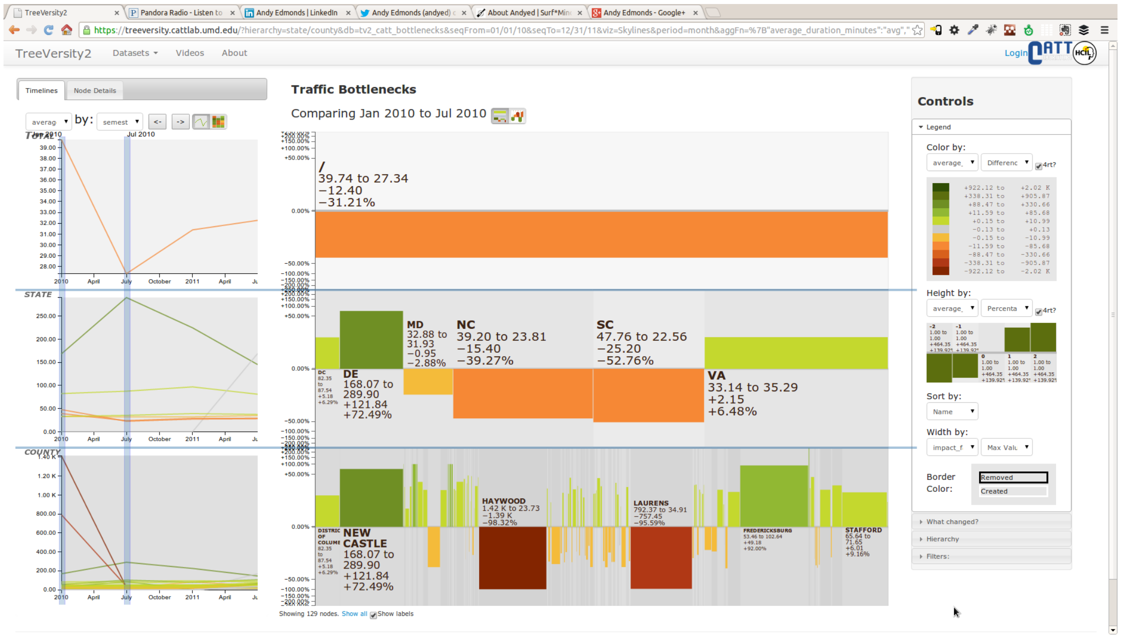


Figure B.1: CATT Lab's average traffic bottlenecks in six states of the U.S.

MILCS level: Early, Chauffeur Mode

Duration: March 14 2012 - June 4 2012 (10 Months)

System used: TreeVersity and TreeVersity2

Data: US Federal Budget as published by the White House[1]

Number of rows: 4,845

Number of time points: 56 (years)

Example tree size: 1,393 (4 Levels)

Number of numeric variables: 1 (budget outlays)

Number of Attributes: 7 (Agency name, Bureau name, Account name, Subfunction Title, BEA category, Grant/non-grant, On/off budget)

Type of Hierarchy: Mixed

Type of comparison: Type 3: aggregating + different topology

URL: `https://treeversity.cattlab.umd.edu/cs/budget`

[1] `http://www.whitehouse.gov/sites/default/files/omb/budget/fy2013/assets/outlays.csv`

B.2.2 Discussion

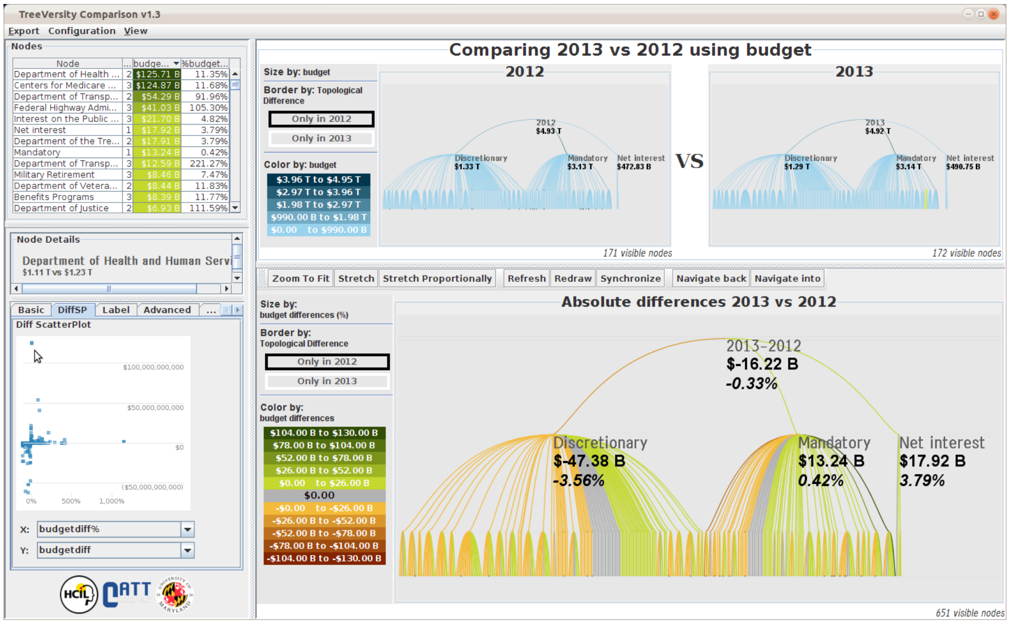
In this case study I used the 2012 and 2013 U.S. Federal Budget outlays [2] as published on the White House website [3] on March 2012. The Federal Budget has an explicit hierarchy composed by the Agencies, and their Bureaus, but deeper hierarchies can be built grouping by these attributes Account Name, Sub-Function Name, Budget Enforcement Act (BEA) category (either Discretionary, Mandatory or Net Interest), Grant/No-Grant and On-Budget/Off-Budget; grouped by these attributes the 2012 budget generates a tree with 7,511 nodes while the one for 2013 has 7,085 nodes. The original budget included negative values for some accounts, although TreeVersity supports those values, they weren't considered for the examples that follow to avoid confusion for the readers.

Overview of the changes

The first task performed on the Budget was to try to understand the overall changes between 2012 and 2013. For this, the compared trees were reorganized to be grouped first by BEA category, and then by Agency and Bureaus. The color differences represent absolute changes. The comparison, shown in Figure B.2, allows many immediate conclusions. First there is a total expected decrease of $16.22 Billions for 2013 compared to 2012 on the overall Budget. Also the tree types of BEA Categories are clear (first level nodes), and is clear that from those Net Interest is the one with fewer accounts. The other two categories separate the budget between the accounts that can be changed by the government (Discretionary) and the ones that cannot (Mandatory). The overall budget of Mandatory accounts is planned to increase by $13.24 Billions (+0.42%) while the Discretionary are scheduled to be reduced on $47.38 (-3.56%). Many Agencies (level 2 nodes) are staying with the same budget as can be seen by the nodes with gray edges, however none of those contains more than one Bureau (level 3 nodes). Finally some dark green nodes can be spotted at the far right of the Mandatory subtree, those are the *Department of Health and Human Services* and its Bureau *Centers for Medicare and Medicaid Services* which are scheduled to be increased $125.71 Billions (+11.35%) and

[2] Amount of money that is expected to be spent

[3] http://www.whitehouse.gov/omb/budget/Supplemental/



TreeVersity Comparison v1.3

Nodes

Node	...	budge...	%budget...
Department of Health ...	2	$125.71 B	11.35%
Centers for Medicare ...	3	$124.87 B	11.68%
Department of Transp...	2	$54.29 B	91.96%
Federal Highway Admi...	3	$41.03 B	105.30%
Interest on the Public ...	3	$21.70 B	4.82%
Net interest	1	$17.92 B	3.79%
Department of the Tre...	2	$17.91 B	3.79%
Mandatory	1	$13.24 B	0.42%
Department of Transp...	3	$12.59 B	221.27%
Military Retirement	3	$8.46 B	7.47%
Department of Vetera...	2	$8.44 B	11.83%
Benefits Programs	3	$8.39 B	11.77%
Department of Justice	2	$6.93 B	111.59%

Node Details

Department of Health and Human Servi
$1.11 T vs $1.23 T

Diff ScatterPlot

X: budgetdiff%

Y: budgetdiff

Comparing 2013 vs 2012 using budget

Size by: budget

Border by: Topological Difference

Color by: budget

Absolute differences 2013 vs 2012

Size by: budget differences (%)

Border by: Topological Difference

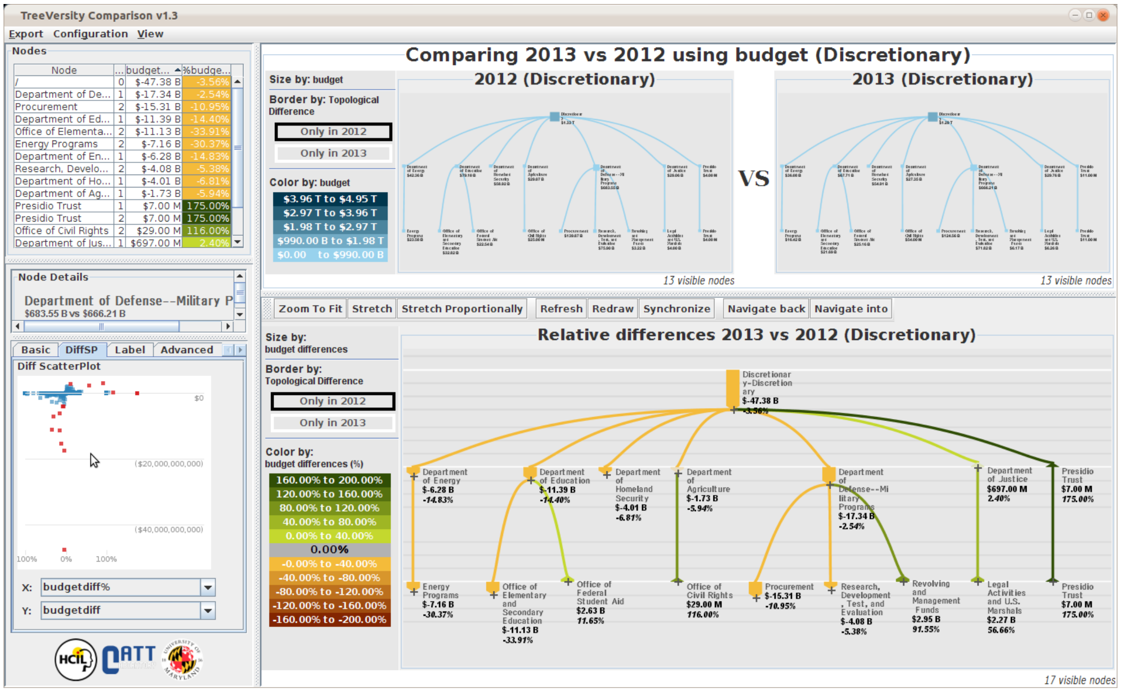
Color by: budget differences

Figure B.2: Overview of changes in the Federal Budget between 2013 and 2012 grouped by the Budget Enforcement Act category (Mandatory, Discretionary or Net Interest), and then by Agency (leaf nodes). The color here shows the absolute node changes. TreeVersity shows how there is a budgeted cut of $47.38 (-3.56%) Billion on all the Discretionary accounts, the only ones that the Government can actually modify.

$124.87 Billions (+11.68%) respectively, big outliers compared to the rest of the Budget changes.

Agencies and Bureaus changing the most

After analyzing the overview, the next task performed was to analyze the Agencies and Bureaus that changed the most. As shown in Figure B.3 the DiffScatterPlot (on the left of the figure) shows interesting outliers, e.g. one node increasing by more than $100 Billion. All the nodes changing by more than $10 Billion were selected (yellow dots on the DiffScatterPlot). The resulting DiffTree uses color for absolute change and bullet height for the percentage change. Some relevant nodes in the Mandatory subtree are the *Federal Deposit Insurance Corporation* that has a significant percentage decrease (-$62.34 Billion, -33.72%) and the *Department of Transportation* that is changing in the opposite direction (+$54.29 Billion, +91.96%). With respect to the absolute differences,



Node	...	budget...	%budge...
/	0	$-47.38 B	-3.56%
Department of De...	1	$-17.34 B	-2.54%
Procurement	2	$-15.31 B	-10.95%
Department of Ed...	1	$-11.39 B	-14.40%
Office of Elementa...	2	$-11.13 B	-33.91%
Energy Programs	2	$-7.16 B	-30.37%
Department of En...	1	$-6.28 B	-14.83%
Research, Develo...	2	$-4.08 B	-5.38%
Department of Ho...	1	$-4.01 B	-6.81%
Department of Ag...	1	$-1.73 B	-5.94%
Presidio Trust	1	$7.00 M	175.00%
Presidio Trust	2	$7.00 M	175.00%
Office of Civil Rights	2	$29.00 M	116.00%
Department of Jus...	1	$697.00 M	2.40%

Node Details

Department of Defense--Military P
$683.55 B vs $666.21 B

X: budgetdiff%

Y: budgetdiff

Size by: budget

Border by: Topological Difference

Color by: budget

Size by: budget differences

Border by: Topological Difference

Color by: budget differences (%)

TreeVersity Comparison v1.3
Export
Configuration
View
Nodes
Basic
DiffSP
Label
Advanced
Diff ScatterPlot
$0
($20,000,000,000)
($40,000,000,000)
100%
0%
100%
Comparing 2013 vs 2012 using budget (Discretionary)
2012 (Discretionary)
2013 (Discretionary)
VS
Only in 2012
Only in 2013
$3.96 T to $4.95 T
$2.97 T to $3.96 T
$1.98 T to $2.97 T
$990.00 B to $1.98 T
$0.00 to $990.00 B
13 visible nodes
13 visible nodes
Zoom To Fit
Stretch
Stretch Proportionally
Refresh
Redraw
Synchronize
Navigate back
Navigate into
Relative differences 2013 vs 2012 (Discretionary)
Only in 2012
Only in 2013
160.00% to 200.00%
120.00% to 160.00%
80.00% to 120.00%
40.00% to 80.00%
0.00% to 40.00%
0.00%
-0.00% to -40.00%
-40.00% to -80.00%
-80.00% to -120.00%
-120.00% to -160.00%
-160.00% to -200.00%
Discretionary-Discretionary $-47.38 B -3.56%
Department of Energy $-6.28 B -14.83%
Department of Education $-11.39 B -14.40%
Department of Homeland Security $-4.01 B -6.81%
Department of Agriculture $-1.73 B -5.94%
Department of Defense--Military Programs $-17.34 B -2.54%
Department of Justice $697.00 M 2.40%
Presidio Trust $7.00 M 175.00%
Energy Programs $-7.16 B -30.37%
Office of Elementary and Secondary Education $-11.13 B -33.91%
Office of Federal Student Aid $2.63 B 11.65%
Office of Civil Rights $29.00 M 116.00%
Procurement $-15.31 B -10.95%
Research, Development, Test, and Evaluation $-4.08 B -5.38%
Revolving and Management Funds $2.95 B 91.55%
Legal Activities and U.S. Marshals $2.27 B 56.66%
Presidio Trust $7.00 M 175.00%
17 visible nodes

Figure B.3: Biggest changers in the among the Discretionary category of the Federal Budget between 2013 and 2012. Color represents percentage changes while size represents absolute changes. Nodes were filtered to show only Agencies and Bureaus and the DiffScatterPlot, on the left, was used to select the biggest changers. The size of the bullets show how the *Departments of Defense* and *Education* are getting the biggest cuts in absolute value (-$17.34 Billion and -$11.39 Billion respectively), while *Presidio Trust* with a dark green on the right is the most significant percentage difference with 175% increase.

the Mandatory side of the *Department of Health and Human Services* is again the most notable Agency (+$125.71 Billion, +11.35%). It is also interesting to see that only two Agencies changed by more than $10 Billions under the Discretionary side (the one that the Government can actually change) the *Departments of Defense* (-$17.34 Billion, -2.54%) and of *Education* (-$11.39 Billion, -14.40%).

Created and Removed Agencies and Bureaus

The final analysis task performed on the US Federal Budget was to identify the added and removed Agencies and Bureaus. For this task the topology filtering of TreeVersity was used. The results shown in Figure B.4 displays two removed Agencies and 3 cre-

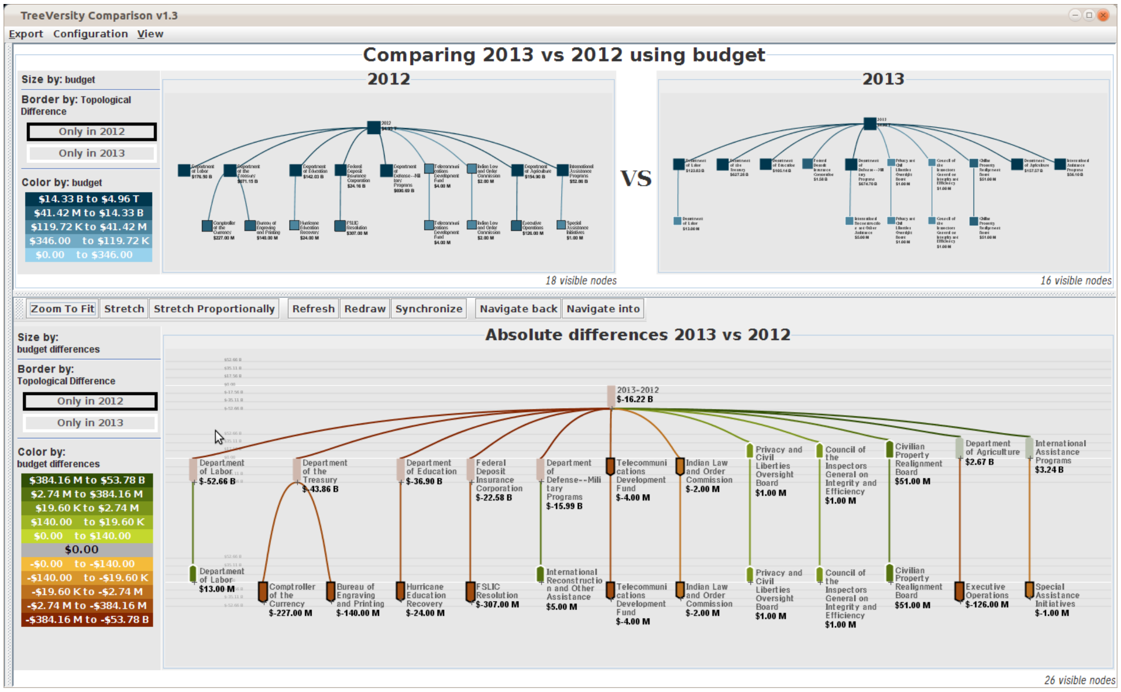
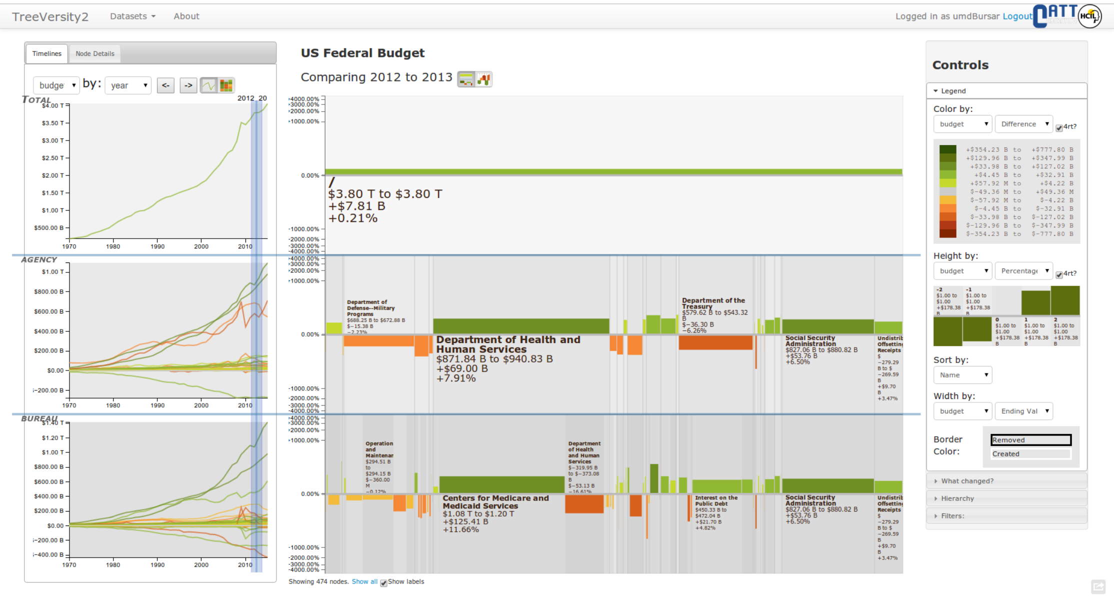

TreeVersity Comparison v1.3
Export Configuration View
Comparing 2013 vs 2012 using budget
Size by: budget
Border by: Topological Difference
Only in 2012
Only in 2013
Color by: budget
$14.33 B to $4.96 T
$41.42 M to $14.33 B
$119.72 K to $41.42 M
$346.00 to $119.72 K
$0.00 to $346.00
2012
2013
VS
Department of Labor $176.50 B
Department of the Treasury $871.15 B
Department of Education $142.03 B
Federal Deposit Insurance Corporation $24.16 B
Department of Defense--Military Programs $690.03 B
Telecommunications Development Fund $4.00 M
Indian Law and Order Commission $2.00 M
Department of Agriculture $154.90 B
International Assistance Programs $52.86 B
Comptroller of the Currency $227.00 M
Bureau of Engraving and Printing $140.00 M
Hurricane Education Recovery $24.00 M
FSLIC Resolution $307.00 M
Telecommunications Development Fund $4.00 M
Indian Law and Order Commission $2.00 M
Executive Operations $126.00 M
Special Assistance Initiatives $1.00 M
18 visible nodes
16 visible nodes
Zoom To Fit
Stretch
Stretch Proportionally
Refresh
Redraw
Synchronize
Navigate back
Navigate into
Absolute differences 2013 vs 2012
Size by: budget differences
Border by: Topological Difference
Only in 2012
Only in 2013
Color by: budget differences
$384.16 M to $53.78 B
$2.74 M to $384.16 M
$19.60 K to $2.74 M
$140.00 to $19.60 K
$0.00 to $140.00
$0.00
-$0.00 to -$140.00
-$140.00 to -$19.60 K
-$19.60 K to -$2.74 M
-$2.74 M to -$384.16 M
-$384.16 M to -$53.78 B
2013-2012 $-16.22 B
Department of Labor $-52.66 B
Department of the Treasury $-43.86 B
Department of Education $-36.90 B
Federal Deposit Insurance Corporation $-22.58 B
Department of Defense--Military Programs $-15.99 B
Telecommunications Development Fund $-4.00 M
Indian Law and Order Commission $-2.00 M
Privacy and Civil Liberties Oversight Board $1.00 M
Council of the Inspectors General on Integrity and Efficiency $1.00 M
Civilian Property Realignment Board $51.00 M
Department of Agriculture $2.67 B
International Assistance Programs $3.24 B
Department of Labor $13.00 M
Comptroller of the Currency $-227.00 M
Bureau of Engraving and Printing $-140.00 M
Hurricane Education Recovery $-24.00 M
FSLIC Resolution $-307.00 M
International Reconstruction and Other Assistance $5.00 M
Telecommunications Development Fund $-4.00 M
Indian Law and Order Commission $-2.00 M
Privacy and Civil Liberties Oversight Board $1.00 M
Council of the Inspectors General on Integrity and Efficiency $1.00 M
Civilian Property Realignment Board $51.00 M
Executive Operations $-126.00 M
Special Assistance Initiatives $-1.00 M
26 visible nodes

Figure B.4: Created and Removed Agencies (nodes at level 1) and Bureaus (nodes at level 2) in the Federal Budget between 2013 and 2011. Created nodes are denoted with thick white borders, while black was used for removed ones. Note that the topology differences are evident on the two compared trees on the top of the image. For this image size represents absolute change and color the percentage differences and a logarithmic scale was used for both.

ated and 8 Bureaus deleted and 5 created. In contrast to the two previous examples, in this task the Budget Agencies were not grouped by their BEA Category. Also, TreeVersity was configured to use color and Bullet height to represent absolute differences, and a logarithmic scale to accentuate the smaller differences. Worth noticing is the *FSLIC Resolution* Bureau of the *Federal Deposit Insurance Corporation* Agency, that is scheduled to be removed in 2013 with its $307 Millions in budget.



budge by: year

Color by: budget Difference ☑ 4rt?

Height by: budget Percentage ☑ 4rt?

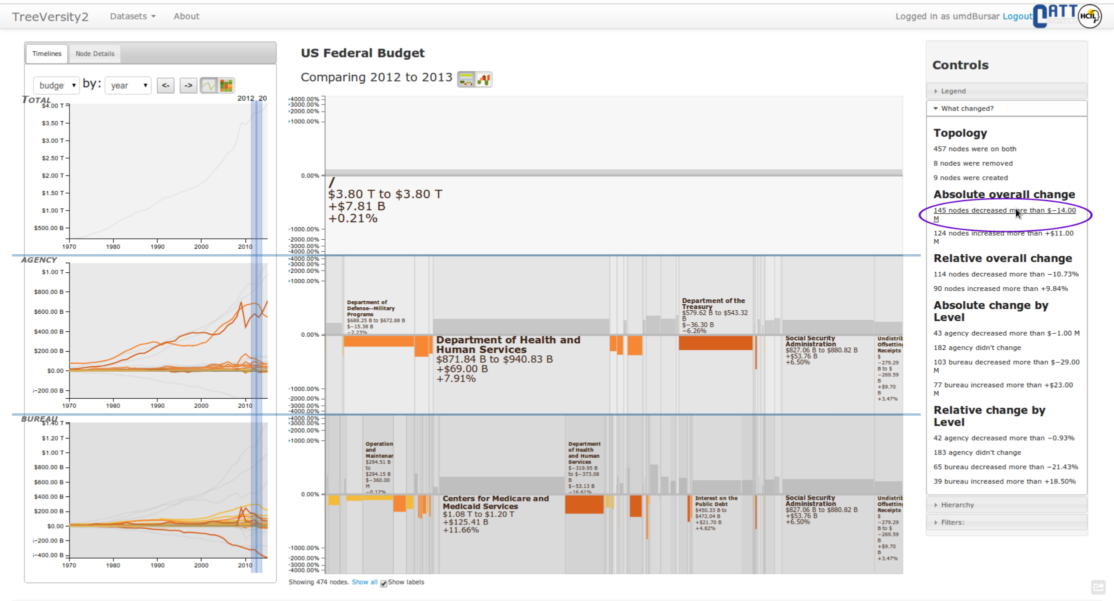
Sort by: Name

Width by: budget Ending Val

Border Color:

Showing 474 nodes. Show all ☑ Show labels

Figure B.5: Changes in the US Federal Budget Between 2013 and 2012. The left side shows the timelines of the actual budgets by element in the tree: overall on the top, by Agency on the middle and by Bureau on the bottom. The StemView (center of the screen) illustrate the changes between 2013 and 2012. Each box in the StemView represents an element in the Budget. The green box on the top tell us that overall the Budget increased in US\$7.81 Billion. The middle row shows the changes in by Agency, where Defense, Health, Treasury and Social Security are the main players, and all are increasing. The colors represent the change in dollars while the height of the boxes show the percentage of change. The width shows the actual budget in 2013.



TreeVersity2
Datasets
About
Logged in as umdBursar Logout
Timelines
Node Details
budge
by:
year
TOTAL
AGENCY
BUREAU
US Federal Budget
Comparing 2012 to 2013
/
$3.80 T to $3.80 T
+$7.81 B
+0.21%
Department of Defense--Military Programs $688.25 B to $672.88 B $−15.38 B −2.23%
Department of Health and Human Services $871.84 B to $940.83 B +$69.00 B +7.91%
Department of the Treasury $579.62 B to $543.32 B $−36.30 B −6.26%
Social Security Administration $827.06 B to $880.82 B +$53.76 B +6.50%
Undistrib Offsettin Receipts $ −279.29 B to $ −269.59 B +$9.70 B +3.47%
Operation and Maintenar $294.51 B to $294.15 B $−360.00 M −0.12%
Department of Health and Human Services $−319.95 B to $−373.08 B $−53.13 B −16.61%
Centers for Medicare and Medicaid Services $1.08 T to $1.20 T +$125.41 B +11.66%
Interest on the Public Debt $450.33 B to $472.04 B +$21.70 B +4.82%
Showing 474 nodes. Show all ☑ Show labels
Controls
Legend
What changed?
Hierarchy
Filters:

Topology

457 nodes were on both

8 nodes were removed

9 nodes were created

Absolute overall change

145 nodes decreased more than $−14.00 M

124 nodes increased more than +$11.00 M

Relative overall change

114 nodes decreased more than −10.73%

90 nodes increased more than +9.84%

Absolute change by Level

43 agency decreased more than $−1.00 M

182 agency didn't change

103 bureau decreased more than $−29.00 M

77 bureau increased more than +$23.00 M

Relative change by Level

42 agency decreased more than −0.93%

183 agency didn't change

65 bureau decreased more than −21.43%

39 bureau increased more than +18.50%

Figure B.6: The Reporting tool highlighting all the agencies and bureaus in the US Federal Budget that decrease more than $14 million dollars. Users can filter down to only those accounts by clicking on the corresponding line of text in the reporting tool.

B.3 UMD Budget

B.3.1 Case Study Sheet

Partner: Theresa Gill Beck, Assistant Director (tbeck at umd.edu)

Organization: Office of Budget & Fiscal Analysis, University of Maryland

MILCS level: Early, Chauffeur Mode

Duration: October 2012 to November 2013 (2 months, 2 meetings)

System used: TreeVersity2

Data: University of Maryland Budget

Number of rows: 16,332

Number of time points: 5 (years)

Example tree size: 1,296 (3 levels)

Number of numeric variables: 1 (budget)

Number of Attributes: 6 (Department, Division, MFS, Support Ind., Program, Major

Classification)

Type of Hierarchy: Mixed

Type of comparison: Type 3: aggregating + different topology

New Features:

Limitations:

URL: Restricted

B.3.2 Discussion

In this case study I worked with Ms Theresa Gill Beck Assistant Director of the Office of Budget and Fiscal Analysis of the University of Maryland. Ms Beck was interested in using TreeVersity2 capabilities to explore the changes in the budget of the University of Maryland. I worked with here in two meetings were we first discussed possible paths of exploration and then demonstrated TreeVersity2 to a group of analysts from the University of Maryland. This case study helped to setup the starting of the Case Study developed with the Office of Institutional Research, Planning & Assessment of the University.

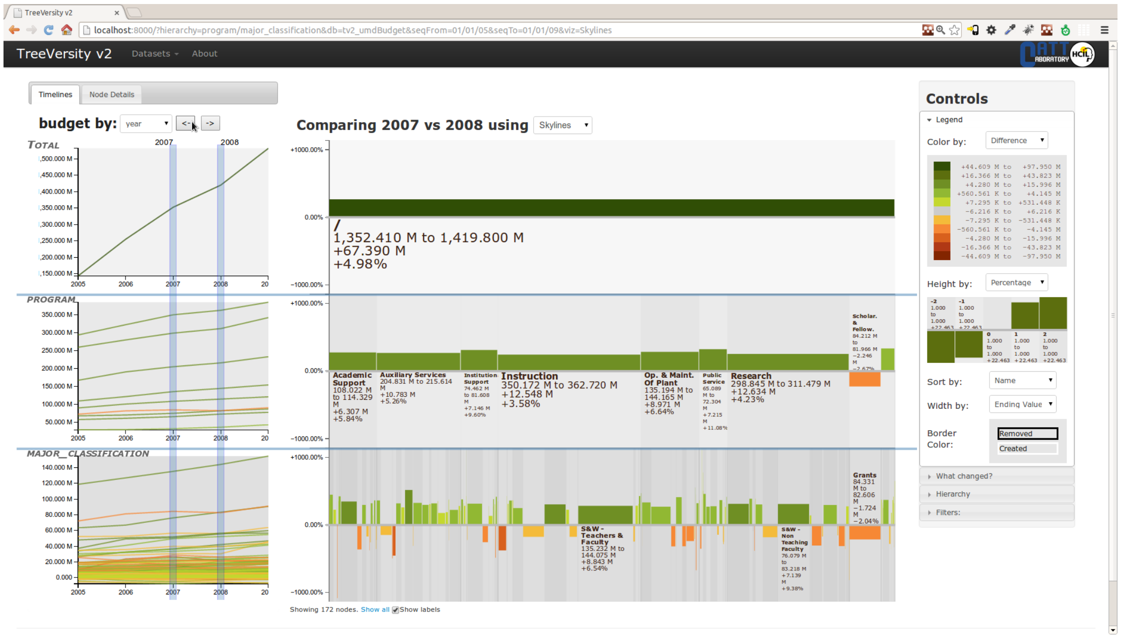
When asked to comment about her experience with TreeVersity2 Ms Beck expressed: "*TreeVersity easily showed us multi-years of the data in a graphical form at once without having to download the data into excel and create graphics. We can do that analysis now, but TreeVersity was much quicker.*". Figure B.7 shows an example analysis of the Budget of the University.

B.4 Department of Transportation Airlines Maintenance Budgets

B.4.1 Case Study Sheet

Partner: Martin Akerman; Pat Hu, Associate Administrator and Director, Bureau of Transportation Statistics

Organization: Department of Transportation


TreeVersity v2

localhost:8000/?hierarchy=program/major_classification&db=tv2_umdBudget&seqFrom=01/01/05&seqTo=01/01/09&viz=Skylines

budget by: year

Comparing 2007 vs 2008 using Skylines

Controls

Color by: Difference

Height by: Percentage

Sort by: Name

Width by: Ending Value

Border Color: Removed, Created

Showing 172 nodes. Show all ☑ Show labels

Figure B.7: University of Maryland Budget grouped by program and major classifications

MILCS level: Early, Chauffeur Mode

Duration: November 2011, July 2012 (9 Months, 3 meetings)

System used: TreeVersity

Data: Amounts of money spent in maintenance as reported by Airlines operating in the US

Number of rows: 216

Example tree size: 187 (2 Levels)

Number of numeric variables: 1 (maintenance budget)

Number of Attributes: 2 (Region, Carrier)

Type of Hierarchy: Dynamic

Type of comparison: Type 3: aggregating + different topology

B.4.2 Discussion

I cooperated with the U.S. Department of Transportation (DOT) to analyze the changes in the maintenance budgets of the different airlines that operate in the US. The dataset contained the reported amount of money spent by the airlines in maintenance by quarters and years, and other attributes such as the regions where the airlines operate and their net incomes. Two hierarchies where built at the request of the DOT's analysts, first grouping the airlines by region of operation (A: Atlantic, L: Latin America, D: Domestic, P: Pacific, I: International) and then by the carriers, and the opposite direction. A total of 67 carriers were compared (only those with operation revenues by $20 millions or more were available) between 2011 and 2010. The dataset is available on the Research and Innovative Technology Administration Bureau of Transportation Statistics website under the *Schedule P-1.2* link under *Air Carrier Financial Reports* [4]

In order to present TreeVersity and then to analyze the airlines dataset, I held two one-hour meetings with Ms Patricia Hu, Associate Administrator and Director of the Bureau of Transportation and Statistics and other members of her staff. The first meeting served as an introduction to the tool, where I described the types of tasks that can be performed with TreeVersity and some example datasets were presented. During this meeting the officials of the DOT brainstormed ideas of comparisons they would like to make using using TreeVersity on their datasets. For the second meeting, after a ten minutes introduction of the tool (for the first time attendees), I presented the different comparisons of the airlines maintenance budgets between 2010 and 2011.

Figure B.8 shows an example of the visualizations discussed during the second meeting. The officials wanted to know which airlines changed their budget the most when grouped by regions, so using the DiffScatterPlot I filtered those changing the most, both in absolute and in percentage differences. After the filter, only 8 out of the 67 carriers remained. Looking at these airlines, Ms Hu immediately noticed that *PSA* and *Compass Airlines* were big outliers in their percentage changes, with +305.86% and +230.96% respectively. She then asked her staff the reasons behind it, and they explain that both companies had been involved recently in merges that would explain the big increases. About this, she expressed "It's great that we could identify these airlines..." and "... if it weren't for this visualization I wouldn't have noticed this". They also complimented

[4] http://www.transtats.bts.gov/

the aesthetic quality of the framework's design. They also found it interesting that *Delta* and *Southwest Airlines* presented significant absolute increases, and that *American Airlines* was the biggest decreasing carrier. Moreover, they mentioned that they would like to breakdown the differences to the level of aircrafts, so they can compare for example, maintenance budgets for the Boeings 747 and 767. Ms Hu also started suggesting other datasets that she would like to compare using TreeVersity. A Multi-dimensional In-depth Long-term Case Study [79] is being planed with Ms Hu and her staff, to measure insight development using TreeVersity in their datasets in a longer period of time.

Note that in Figure B.8 the carriers are grouped by regions, and the node values for the regions are the average amount of change in the carrier's budgets. Because of this the values in the interior nodes are not aggregating, and therefore this is a tree comparison problem Type 4 (it also includes topological differences) as defined in Section 1.1.1.

B.5 Transportation Research Board publications dataset

B.5.1 Case Study Sheet

Partner: Amanda Wilson, Director of the National Transportation Library (NTL); Pat Hu, Associate Administrator and Director, Bureau of Transportation Statistics

Organization: Department of Transportation

MILCS level: Early, Chauffeur Mode

Duration: November 2011, December 2012 (13 Months, 4 meetings)

System used: TreeVersity and TreeVersity2

Data: Number of Publications for the TRB conferences

Number of rows: 52,135

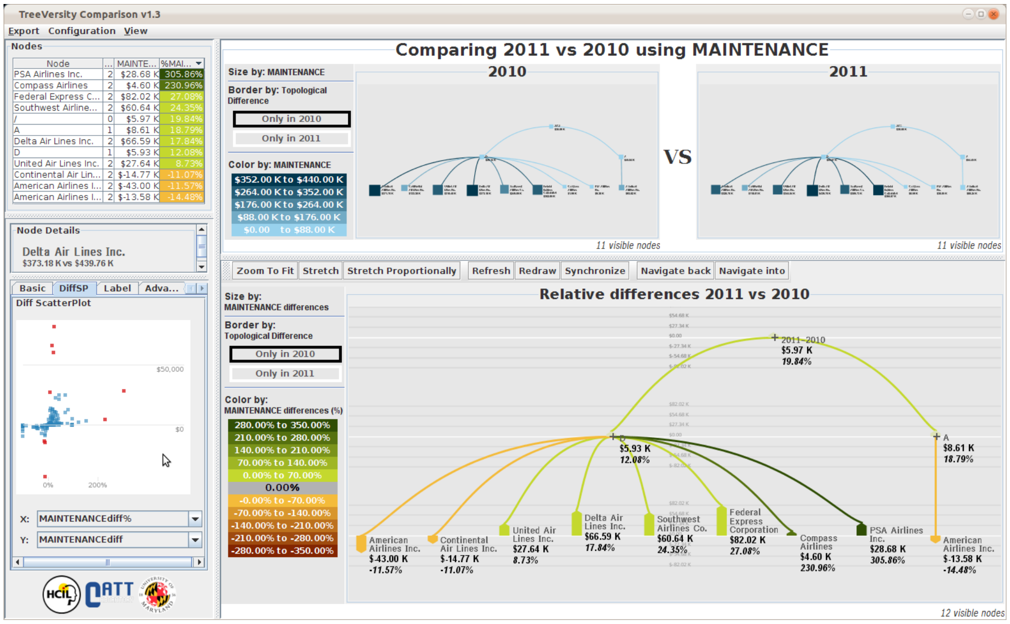
Number of time points: 8,012 (days)

Example tree size: 674 (2 Levels)

Number of numeric variables: 1 (number of papers)

Number of Attributes: 20 (Author, Subject, Conference, TRIS Classification, etc)

Type of Hierarchy: Dynamic



Nodes

Node	...	MAINTE...	%MAI...
PSA Airlines Inc.	2	$28.68 K	305.86%
Compass Airlines	2	$4.60 K	230.96%
Federal Express C...	2	$82.02 K	27.08%
Southwest Airline...	2	$60.64 K	24.35%
/	0	$5.97 K	19.84%
A	1	$8.61 K	18.79%
Delta Air Lines Inc.	2	$66.59 K	17.84%
D	1	$5.93 K	12.08%
United Air Lines Inc.	2	$27.64 K	8.73%
Continental Air Lin...	2	$-14.77 K	-11.07%
American Airlines I...	2	$-43.00 K	-11.57%
American Airlines I...	2	$-13.58 K	-14.48%

Node Details

Delta Air Lines Inc.
$373.18 K vs $439.76 K

X: MAINTENANCEdiff%

Y: MAINTENANCEdiff

Size by: MAINTENANCE

Border by: Topological Difference

Only in 2010

Only in 2011

Color by: MAINTENANCE

- $352.00 K to $440.00 K
- $264.00 K to $352.00 K
- $176.00 K to $264.00 K
- $88.00 K to $176.00 K
- $0.00 to $88.00 K

Size by: MAINTENANCE differences

Border by: Topological Difference

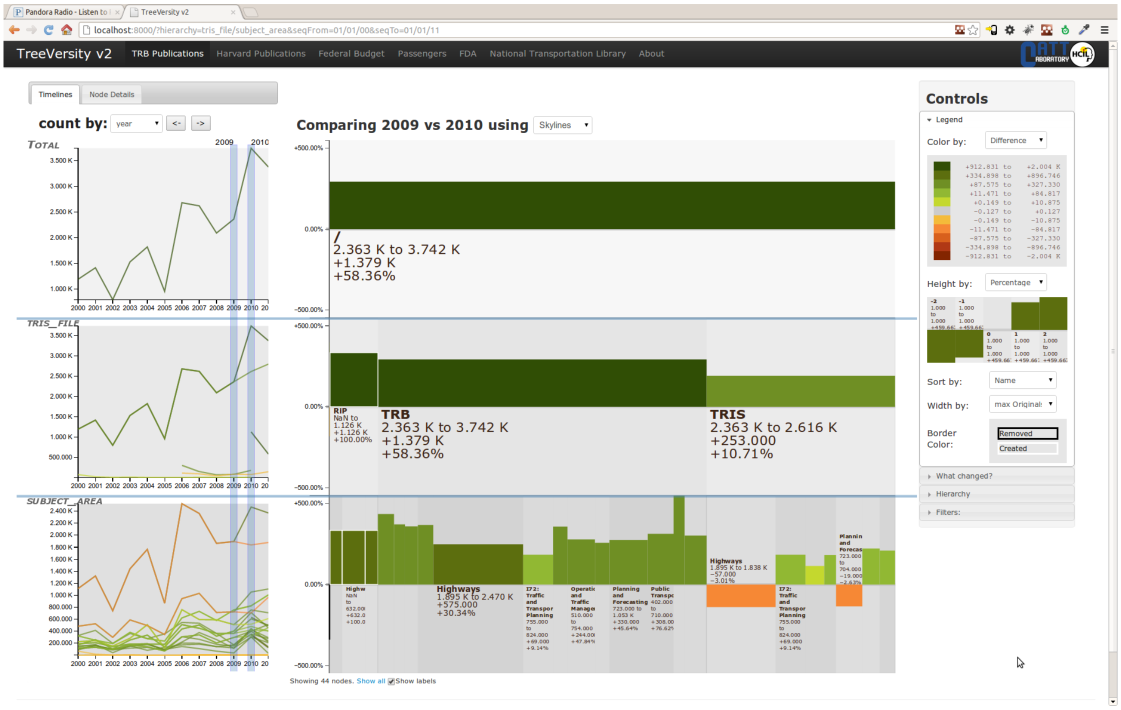
Only in 2010

Only in 2011

Color by: MAINTENANCE differences (%)

- 280.00% to 350.00%
- 210.00% to 280.00%
- 140.00% to 210.00%
- 70.00% to 140.00%
- 0.00% to 70.00%
- 0.00%
- -0.00% to -70.00%
- -70.00% to -140.00%
- -140.00% to -210.00%
- -210.00% to -280.00%
- -280.00% to -350.00%

Figure B.8: Airlines that change their maintenance budget the most between 2011 and 2010 by region. Airlines were filtered by those that incremented their budgets in more than $27,000 or more than 200% or that reduced their budgets in more than $13,500. The budgets are grouped by regions, the nodes at the first level in the tree. D stands for Domestic and A for Atlantic. The values in the regions represent the average amount of change in all the airlines in that region, the root node shows the average overall.



count by: year

Comparing 2009 vs 2010 using Skylines

/
2.363 K to 3.742 K
+1.379 K
+58.36%

RIP
NaN to
1.126 K
+1.126 K
+100.00%

TRB
2.363 K to 3.742 K
+1.379 K
+58.36%

TRIS
2.363 K to 2.616 K
+253.000
+10.71%

Highw
NaN
to
632.00
+632.0
+100.0

Highways
1.895 K to 2.470 K
+575.000
+30.34%

I72: Traffic and Transpor Planning
755.000 to 824.000
+69.000
+9.14%

Operatic and Traffic Manage
510.000 to 754.000
+244.00
+47.84%

Planning and Forecasting
723.000 to 1.053 K
+330.000
+45.64%

Public Transpo
402.000 to 710.000
+308.00
+76.62%

Highways
1.895 K to 1.838 K
−57.000
−3.01%

I72: Traffic and Transpor Planning
755.000 to 824.000
+69.000
+9.14%

Plannin and Forecas
723.000 to 704.000
−19.000
−2.63%

Showing 44 nodes. Show all ☑ Show labels

Color by: Difference

🟩	+912.831 to	+2.004 K
🟩	+334.898 to	+896.746
🟩	+87.575 to	+327.330
🟩	+11.471 to	+84.817
🟩	+0.149 to	+10.875
⬜	-0.127 to	+0.127
🟨	-0.149 to	-10.875
🟧	-11.471 to	-84.817
🟧	-87.575 to	-327.330
🟥	-334.898 to	-896.746
🟥	-912.831 to	-2.004 K

Height by: Percentage

-2 1.000 to 1.000 +459.66
-1 1.000 to 1.000 +459.66
0 1.000 to 1.000 +459.66
1 1.000 to 1.000 +459.66
2 1.000 to 1.000 +459.66

Sort by: Name

Width by: max Original:

Border Color: Removed Created

Figure B.9: Change in the number of publications of the Transportation Research Board between 2009 and 2010, grouped by tris_file (attribute used to represent the source of the publication) and subjects. The figure shows how the TRB and TRIS classifications that mirror each other from 2000 to 2009 behave differently in 2010.

Type of comparison: Type 3: aggregating + different topology

Limitations: URL: restricted

B.5.2 Discussion

The Transportation Research Board Publication dataset contains 52,135 papers published between 1968 and 2012. The analysis of this dataset was complicated by the state of the data, which was not consistent across time. The case study helped DOT Analysts to realize the state of their data and find inconsistencies on it, FigureB.9 shows one of this anomalies.

B.6 National Transportation Library Publications

B.6.1 Case Study Sheet

Partner: Amanda Wilson, Director of the National Transportation Library (NTL) (amanda.wilson dot.gov); Pat Hu, Associate Administrator and Director, Bureau of Transportation Statistics (patricia.hu at dot.gov)
Organization: Department of Transportation
MILCS level: Early, Chauffeur Mode
Duration: November 2011, December 2012 (13 Months, 4 meetings)
System used: TreeVersity and TreeVersity2
Data: Number of Publications for the TRB conferences
Number of rows: 38,351
Number of time points: 374 (days)
Example tree size: 294 (3 Levels)
Number of numeric variables: 1 (number of papers)
Number of Attributes: 10 (Agency, Contributor Name, Contributor Type, Country, Document Type, Group Name Publication Type, Region, Resource Type, Subject)
Type of Hierarchy: Dynamic
Type of comparison: Type 3: aggregating + different topology
Limitations: The data contained many inconsistencies that made it difficult to compare
URL: `https://treeversity.cattlab.umd.edu/cs/ntl`

B.6.2 Discussion

This case study was an extension of the TRB Publications case study developed also with Ms. Amanda Wilson. The dataset Ms. Wilson provided for this study was inconsistent and not appropriate for comparison. This was evident when loading the data into TreeVersity2. As shown on Figure B.10 the number of publications peaked twice on 2003 and 2008, and the values did not persist between years. Ms. Wilson took note of the findings and explained that some of this data was imported from a different source

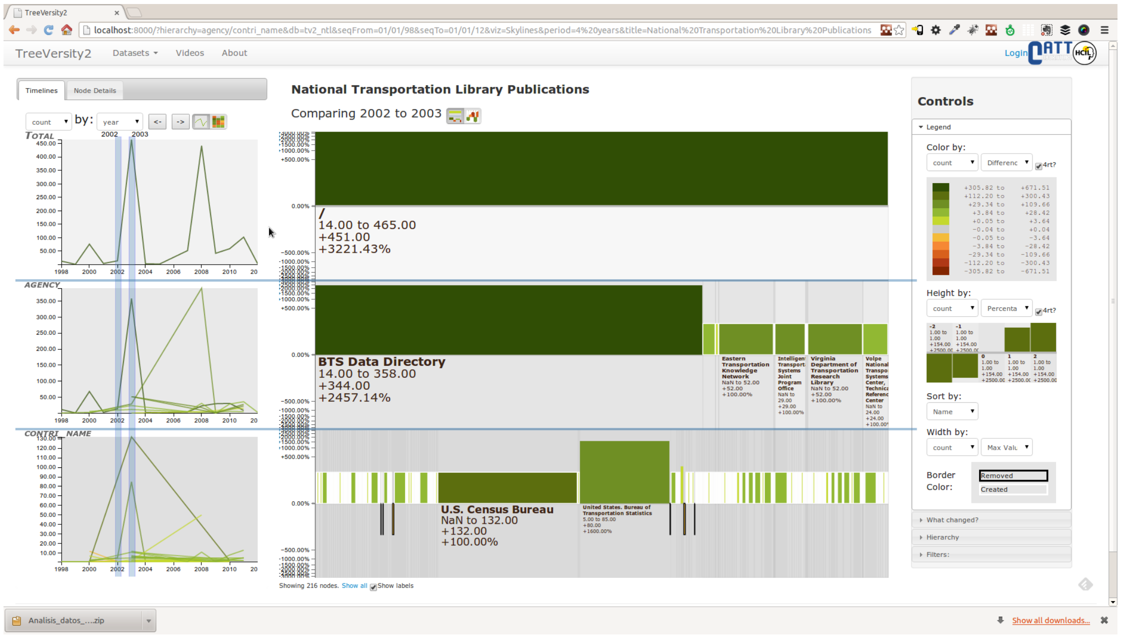


Figure B.10: National Transportation Library publications

during those years which could explain the inconsistency.

B.7 Colombian Blind Students

B.7.1 Case Study Sheet

Partner: María Fernanda Zúñiga Zabala, CEO

Organization: DUTO S.A. (`http://duto.org`)

MILCS level: Mature, user-driven

Duration: 12 February 2013 to 18 February 2013 (6 days, virtual meetings)

System used: TreeVersity2

Data: Number of Blind Student in Colombia as reported by the Government

Number of rows: 33,802

Number of time points: 4 (years)

Example tree size: 1,098 (3 Levels)

Number of numeric variables: 1 (number of students)

Number of Attributes: 21 (Department, City, Type of Disability, School Name, Level, etc)
Type of Hierarchy: Mixed
Type of comparison: Type 3: aggregating + different topology
URL: `https://treeversity.cattlab.umd.edu/cs/inci`

B.7.2 Discussion

Disclosure: *María Fernanda Zúñiga Zabala is my wife, and also co-founder and my boss at DUTO S.A.*

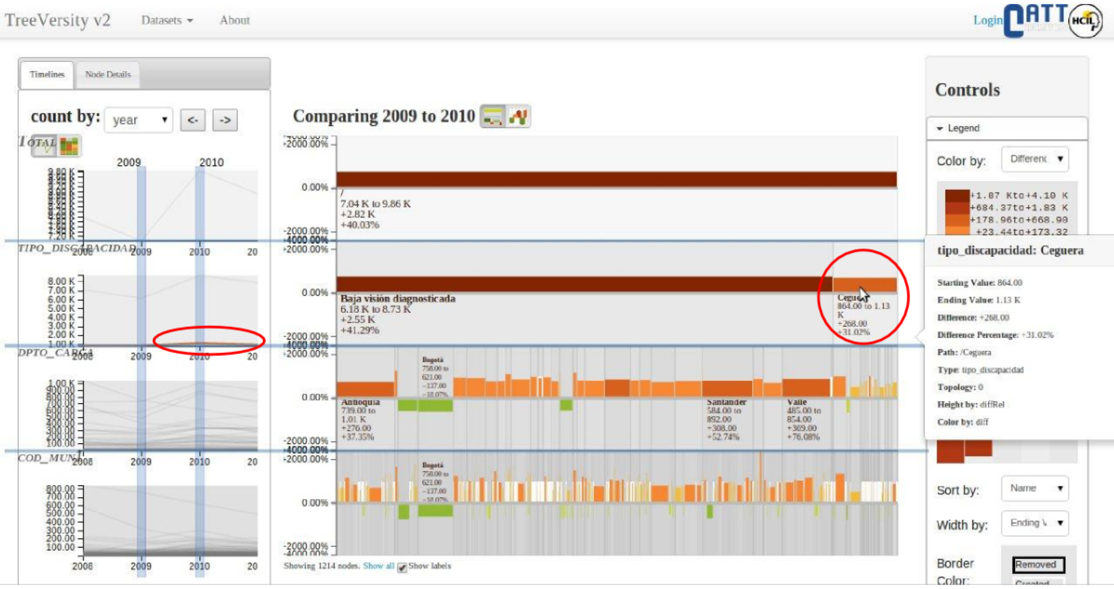
In this case study Ms María Fernanda Zúñiga Zabala used TreeVersity2 to analyze the changes in the number of blind students in Colombia between 2008 and 2011. Ms. Zúñiga leads DUTO a Colombian startup that develops a device for blind student called IRIS. Because of this, Ms. Zúñiga was interested in identifying patterns of increases and/or decreases of the number of blind students in the country.

The dataset was provided to us by the Colombian Education Ministry . The data included 33,802 rows with the information of all the registered blind students in the Country between 2008 and 2011. I loaded the data into TreeVersity2 and let Ms. Zúñiga do the explorations on her own.

Ms. Zúñiga reported that : "TreeVersity allowed me to compare the Colombian blind students information (mainly type of disability and geographical information). Without doubt TreeVersity was way more effective than the data exploration methods that we used before (Spreadsheets and programming scripts) both in terms of the comparisons it allowed us to do and the meaningful insights found"

Ms. Zúñiga reported several insights in the data such as "*The significant increment (40%) of the number of students registered between 2009 and 2010*". She also commented that "*the tool allows the easy detection of anomalies in the analyzed datasets"*. She even mentioned that "*the work I can achieve with TreeVersity is equivalent to years of analysis with our previous methods"*.

Ms Zúñiga also made a list of recommendations for future implementations that in-



tipo_discapacidad: Ceguera

Starting Value: 864.00
Ending Value: 1.13 K
Difference: +268.00
Difference Percentage: +31.02%
Path: /Ceguera
Type: tipo_discapacidad
Topology: 0
Height by: diffRel
Color by: diff

Figure B.11: Changes in the number of blind students in Colombia between 2009 and 2010 grouped by type of disability, by region and by city.

clude "*multiple node selection*", "*generate a printed version of the reporting tool*" or "*add metrics to the reporting tool to highlight the biggest nodes in the tree*". She really liked the TimeBlocks and the Reporting tool and commented that they were very useful in her explorations.

Figures B.11 and B.12 were two of the images generated by Ms. Zuñiga in full report she prepared for the company.

B.8 Imports and Exports in the Americas

B.8.1 Case Study Sheet

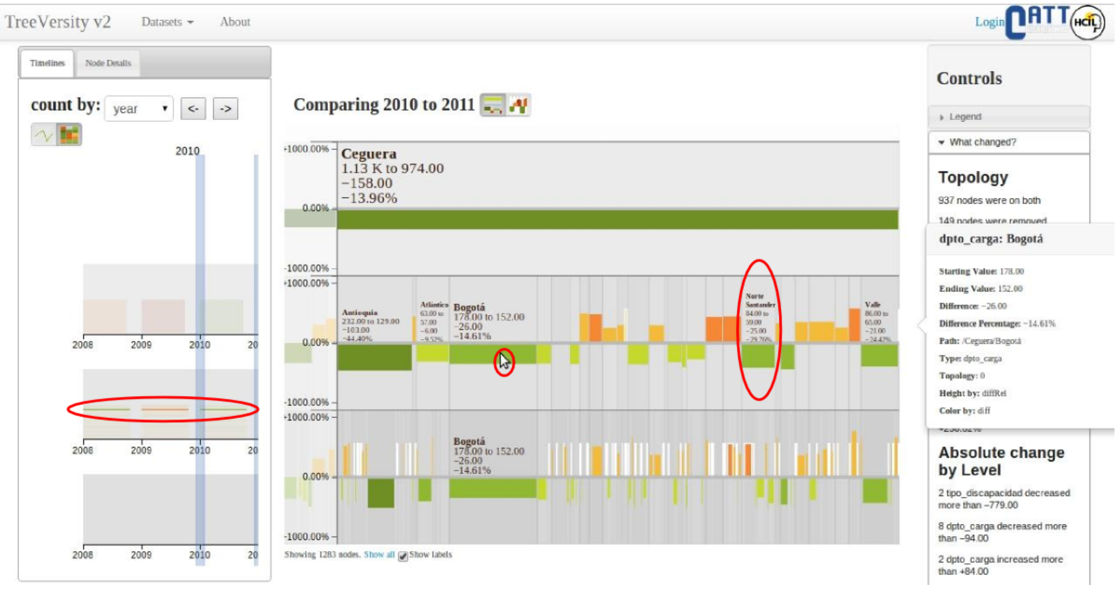
Partner: Dr. Jeremy Harris Economist and Trading Specialist

Organization: Inter American Development Bank

MILCS level: Early**,** self driven

Duration: June 2012 to April 2013 (11 months, virtual)

System used: TreeVersity2



Topology

937 nodes were on both

149 nodes were removed

dpto_carga: Bogotá

Starting Value: 178.00
Ending Value: 152.00
Difference: −26.00
Difference Percentage: −14.61%
Path: /Ceguera/Bogotá
Type: dpto_carga
Topology: 0
Height by: diffRel
Color by: diff

Absolute change by Level

2 tipo_discapacidad decreased more than −779.00

8 dpto_carga decreased more than −94.00

2 dpto_carga increased more than +84.00

Figure B.12: Changes in the number of blind students in Colombia between 2009 and 2010, zoomed in view into the students that are fully blind.

Data: Imports and Exports for the Countries in the Americas between 1992 and 2010

Number of rows: 119,741

Number of time points: 19 (years)

Example tree size: 3,766 (4 Levels)

Number of numeric variables: 1 (Amount of transaction)

Number of Attributes: 5 (Direction of Trade, Origin Country Code, Origin Country Name, Target Country Code, Target Country Name)

Type of Hierarchy: Mixed

Type of comparison: Type 3: aggregating + different topology

URL: `https://treeversity.cattlab.umd.edu/cs/bidTrade`

B.8.2 Discussion

This case study was developed as part of a consulting on Information Visualization that I developed with the Inter American Development Bank. I worked with Dr. Jeremy Harris an economist and trading specialist at the Inter American Bank. Dr. Harris was

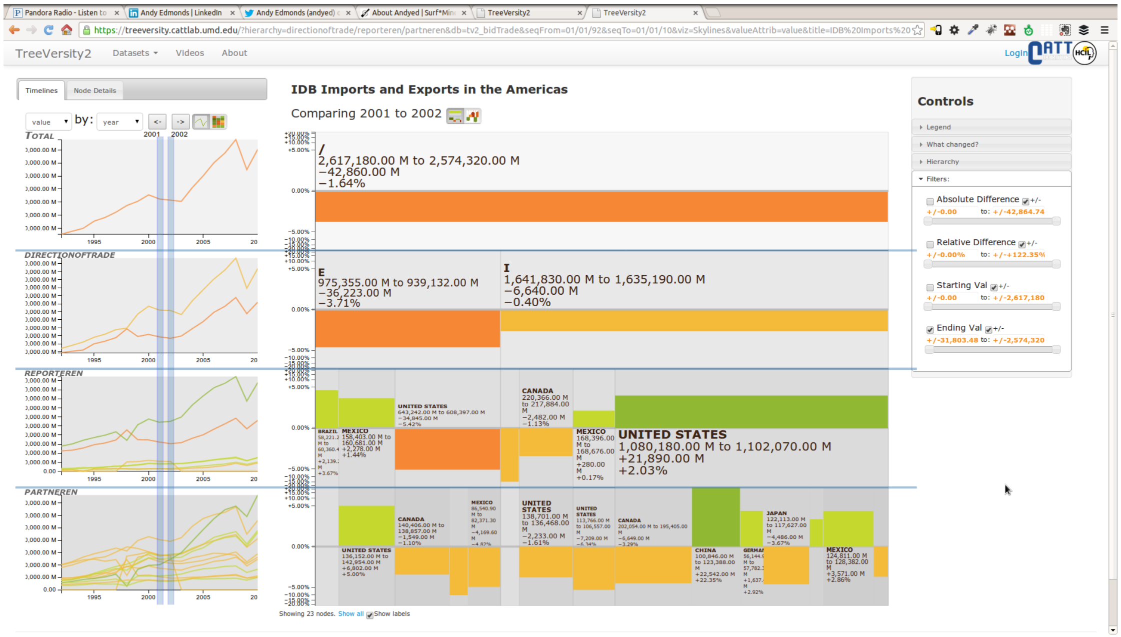


Figure B.13: Change in the imports and exports in the American continent between 2001 and 2002 for the trades of more than US$30 million. The trades are grouped by type of trade (I for imports and E for exports) and then by the exchanging countries.

interested in finding patterns in the trades between countries in the American continent between 1992 and 2010.

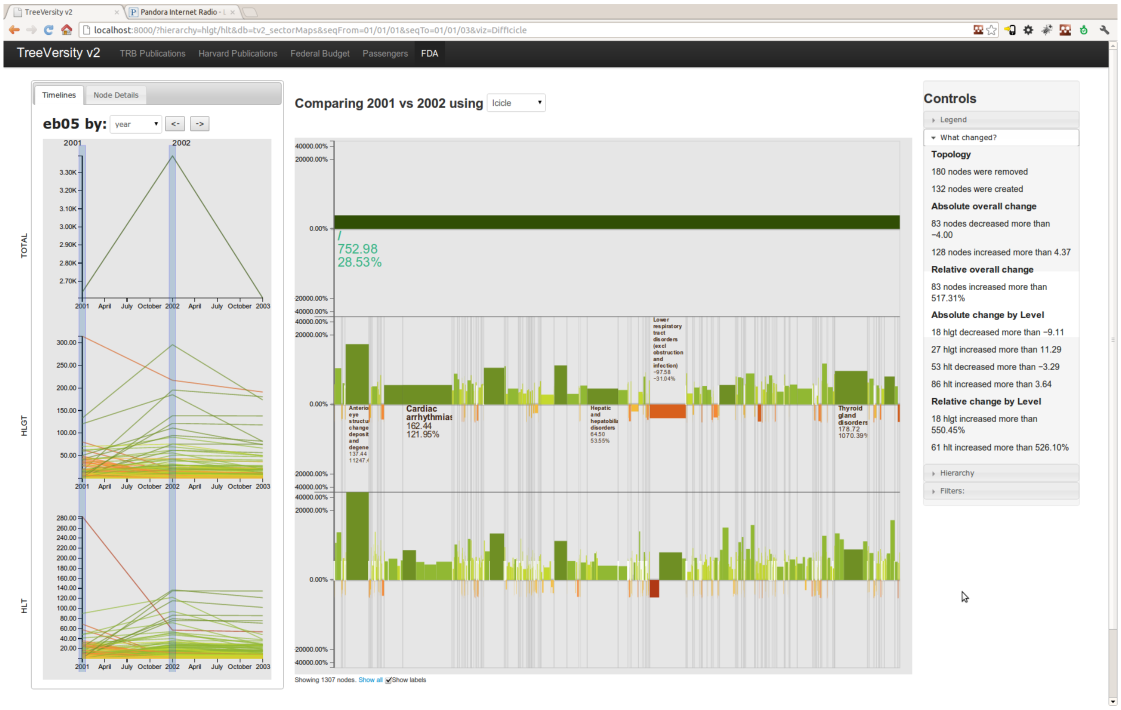
The dataset used for this case study contained 119,741 rows and included the names of the two trading countries, the amount of trade and the direction of the operation (Import or Export).

Dr. Harris was excited to start the case study, but didn't followed up with the process. Figure B.13 shows one of the example analysis that I performed with the data.

B.9 FDA Extended Case Study

B.9.1 Initial Approach

After an initial meeting with Dr. Ana Szarfman on July 18 2012, where Ben Shneiderman, Seth Powsner and I visited the FDA and demonstrated TreeVersity, Dr. Szarfman



TreeVersity v2
Pandora Internet Radio
localhost:8000/?hierarchy=hlgt/hlt&db=tv2_sectorMaps&seqFrom=01/01/01&seqTo=01/01/03&viz=DiffIcicle
TreeVersity v2
TRB Publications
Harvard Publications
Federal Budget
Passengers
FDA
Timelines
Node Details
eb05 by: year
<-
->
TOTAL
HLGT
HLT
2001
2002
2001 April July October 2002 April July October 2003
Comparing 2001 vs 2002 using Icicle
40000.00%
20000.00%
0.00%
/
752.98
28.53%
Cardiac arrhythmias
162.44
121.95%
Thyroid gland disorders
178.72
1070.39%
Lower respiratory tract disorders (excl obstruction and infection)
-97.58
-31.04%
Hepatic and hepatobiliary disorders
Showing 1307 nodes. Show all ☑ Show labels

Controls

- Legend
- What changed?

Topology

180 nodes were removed

132 nodes were created

Absolute overall change

83 nodes decreased more than −4.00

128 nodes increased more than 4.37

Relative overall change

83 nodes increased more than 517.31%

Absolute change by Level

18 hlgt decreased more than −9.11

27 hlgt increased more than 11.29

53 hlt decreased more than −3.29

86 hlt increased more than 3.64

Relative change by Level

18 hlgt increased more than 550.45%

61 hlt increased more than 526.10%

- Hierarchy
- Filters:

Figure B.14: FDA First dataset on an early version of TreeVersity2 (Sept. 10 2012)

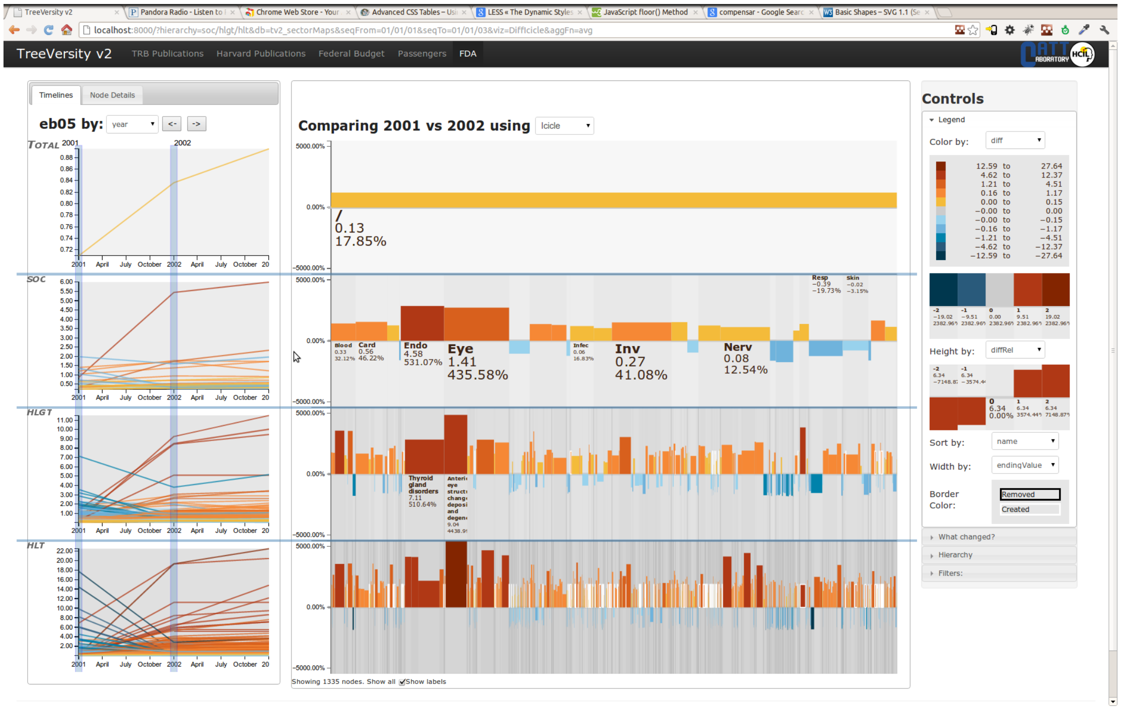
showed us her Sector Maps and we started the case study. Later on August 7th 2012 I received the first dataset from her. It contained the Sector Maps for three different drugs. Since, TreeVersity2 was designed for comparing one tree changing over time, I loaded each tree as a different year into TreeVersity. After the first conversations it was clear that Dr. Szarfman was happy with the Sector Maps for looking at one year of data, but was limited in comparing multiple years at a time. Moreover, since the Sector Maps are treemap based, she was limited to visualizing only EBGM values of the PT nodes (because treemaps only show leaf nodes). Because of this Dr. Szarfman was not expecting to be able to compare changes in all the levels of the MedDRA hierarchy as TreeVersity2 allows, and in the first datasets only sent the EBGM values of the PT nodes (leaves). Figure B.14 shows an early version of TreeVersity2 with the first FDA dataset.

B.9.2 Request for features: allocate for non aggregating fixed hierarchies and different color schema

During August and September 2012 I worked on implementing new features and fix bugs, and by September 14 2012 I managed to load the first dataset into TreeVersity2. This dataset didn't include EBGM values for the nodes for the SOC, HLGT or HLT levels, so I calculated them using the average of the children's values. I presented this on our second meeting on September 24 2012 (as, where Dr. Szarfman explained that the values of the inner levels could not be calculated as a function of the leaf nodes, so she provided the precalculated values. At that moment TreeVersity2 did not supported *non-aggregating hierarchies*, because of the underlying architecture based on SQL aggregative functions (explained in Section 4.3). Therefore the whole underlying architecture had to be changed to allocate for inner node values that were given explicitly rather than calculated. This was done by storing the calculated inner values in the Postgres database, and passing the *fixedHierarchy* parameter on the URL. TreeVersity2's color scheme also needed to leverage the knowledge of the FDA's analysts previous experience on working with the SectorMaps. I created a set of different palettes using Prof. Buck-Coleman's designs as described in Section 3.3.4 and added parameter *theme* to the URL.

B.9.3 Request for features: allow localized navigation, widths for number of reports and confidence intervals.

To allow a time based comparison, Dr. Szarfman gave me a new dataset on September 27 2012 including EBGM values for one drug over time, but she did not include the values for the inner nodes. By October 2 2012 she sent the third dataset including precalculated values for the HLGT and HLT (levels 2 and 3 on the tree), but did not include the values of the SOC. On October 12 2012 Dr. Szarfman sent a new dataset including the SOC values, and we agreed to meet again at the HCIL on October 30 2012 as part of the EventFlow user group meeting, however because of Huricane Sandy the meeting had to



localhost:8000/?hierarchy=soc/hlgt/hlt&db=tv2_sectorMaps&seqFrom=01/01/01&seqTo=01/01/03&viz=DiffIcicle&aggFn=avg

eb05 by: year

Comparing 2001 vs 2002 using Icicle

Controls

Color by: diff

12.59	to	27.64
4.62	to	12.37
1.21	to	4.51
0.16	to	1.17
0.00	to	0.15
−0.00	to	0.00
−0.00	to	−0.15
−0.16	to	−1.17
−1.21	to	−4.51
−4.62	to	−12.37
−12.59	to	−27.64

Height by: diffRel

Sort by: name

Width by: endingValue

Border Color: Removed Created

Showing 1335 nodes. Show all ☑Show labels

Figure B.15: FDA First dataset on an early version of TreeVersity2 (Sept. 18 2012) with the new color coding but without the independent values for the inner nodes.

be postponed to November 14 2012. During this meeting Dr. Szarfman was thrilled to see the new interactions on TreeVersity2 and how easy it was to navigate the differences, "this is amazing!" she said. She was excited at the representations and called some of her coworkers (including Jonathan Levine) to see the system. She also described that she was interested in finding big changes (more than 2.0 in the value) in EBGM values that started in a value of less than 1.5 (e.g. an adverse effect going from 1.1 to 15.2). She explained that when an adverse effect gets to a big value (5.0 or more) it skyrockets and stops being that interesting to them. I showed her that she could explore those changes using the filters and she was excited to find how easy it was to highlight the interesting nodes. The problem with the filters at that time was that every time a filter was changed, or the compared time points were moved, the StemView zoomed out to show the whole tree. Because of this a new localized navigation, that kept the context when changing parameters, filters or time points was implemented in TreeVersity2.

Apart from this, Dr. Szarfman mentioned that it was very important for them to highlight the adverse effects that have non overlapping confidence intervals (defined by the $EB05$ and $EB95$variables), moreover they wanted to see the number of reports of each adverse effect represented somehow in the visualization. In their previous analysis they were able to look for non overlapping intervals over time but for one adverse effect at a time only. TreeVersity2 compared all the adverse effects at once, so it was more powerful than the previous approach, but it needed to show this extra information. This new feature was difficult to implement as it required to combine different variables ($EBGM$, $EB05$,$EB95$ and n) and their modifiers (actual difference, relative difference, starting value and ending value). For this to work, I needed to rewrite the back-end to include the new variables in the SQL query, and to pass all that information back to the client. Controls were added to allow every feature in TreeVersity2 to be linked to an specific variable with a modifier (e.g. height for $EBGM - difference$ width for $n - endingValue$)

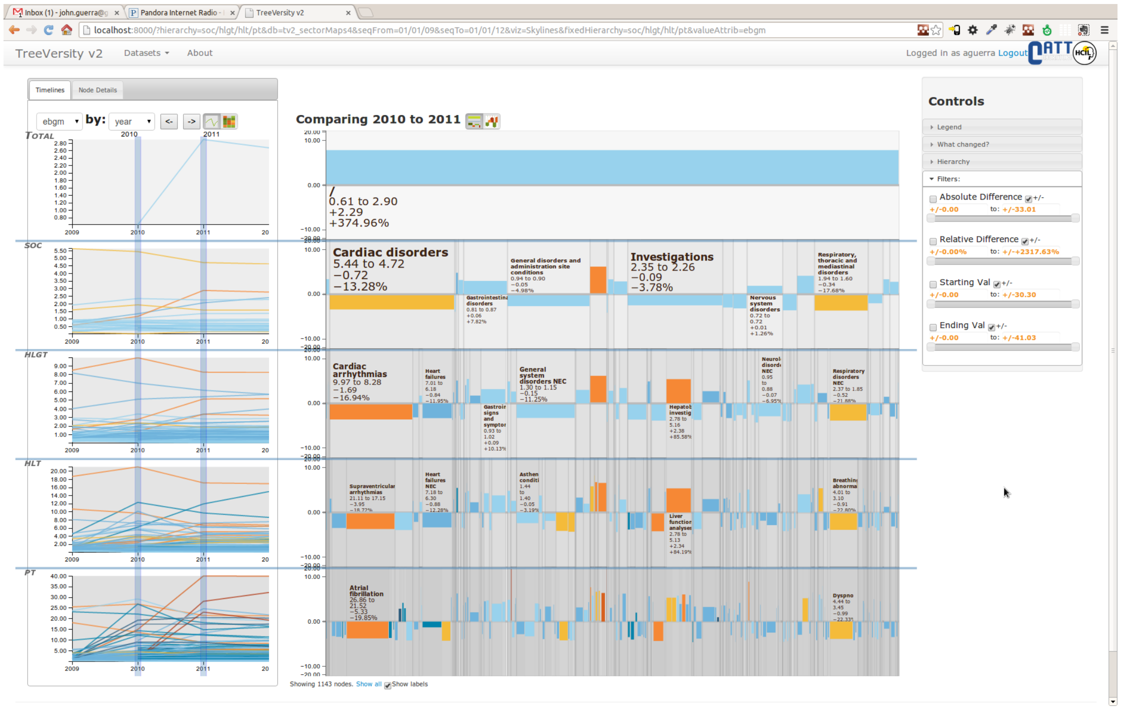
By this time a project was also started to use TreeVersity2 inside the FDA, which revealed some difficulties with the technologies required for running TreeVersity2. On one hand, the FDA had issues with installing the latest versions of Chrome on their computers because of internal protocols. On the other hand, they wanted to run the system locally but it was not possible because of the broad set of tools required to install a TreeVersity2 server (as described in Section 4.3).

Finally to allocate for the confidence intervals, and given that this was a requirement that was outside of the core objective of TreeVersity2, I developed a special hack to show the adverse effects with non overlapping confidence intervals with the colors. This was complicated to implement because to check if the intervals were overlapping the system had to compare values from the starting and ending point, so this could not be made in the pre-proccessing phase. To implement the hack I created a special conditional that check for the FDA case study, and for a special parameter combination (attribute $count - maxValues$) and return the result of $Max(ending\_EB05 - starting\_EB95, ending\_EB95 - starting\_EB05)$, which returns a positive number only when the confidence intervals don't overlap. This was used as the coloring attribute, and as a result Analysts at the FDA could search for yellow to red boxes to identify non overlapping changes, as shown in Figure B.18.

B.9.4 Final meeting

After addressing all of Ana's requests, I requested a final meeting to present the results in March 11 2013. Figures B.16 and B.17 shows two of the screenshots illustrating this results. After seeing this Dr. Szarfman explained that the fourth dataset was artificial and she sent a new one based on real data. The fifth dataset had some issues because it didn't included the full path for each PT, so I asked Dr. Szarfman for a new one that became the sixth dataset. With this dataset we agreed on a final March 25 2013.

Figure B.18 shows the changes between EBGM values for an undisclosed drug between 2010 and 2011 using TreeVersity2 as shown during the final meeting. Each box in the StemView represents an adverse effect, yellow-to-red colored sub-boxes denote adverse effects with non-overlapping confidence levels. Height encoded the relative change of the EBGM index, so sub-boxes going up represent adverse effects getting



localhost:8000/?hierarchy=soc/hlgt/hlt/pt&db=tv2_sectorMaps4&seqFrom=01/01/09&seqTo=01/01/12&viz=Skylines&fixedHierarchy=soc/hlgt/hlt/pt&valueAttrib=ebgm

Logged in as aguerra Logout

ebgm by: year

Comparing 2010 to 2011

Controls

Filters:

□ Absolute Difference ☑+/- +/-0.00 to: +/-33.01

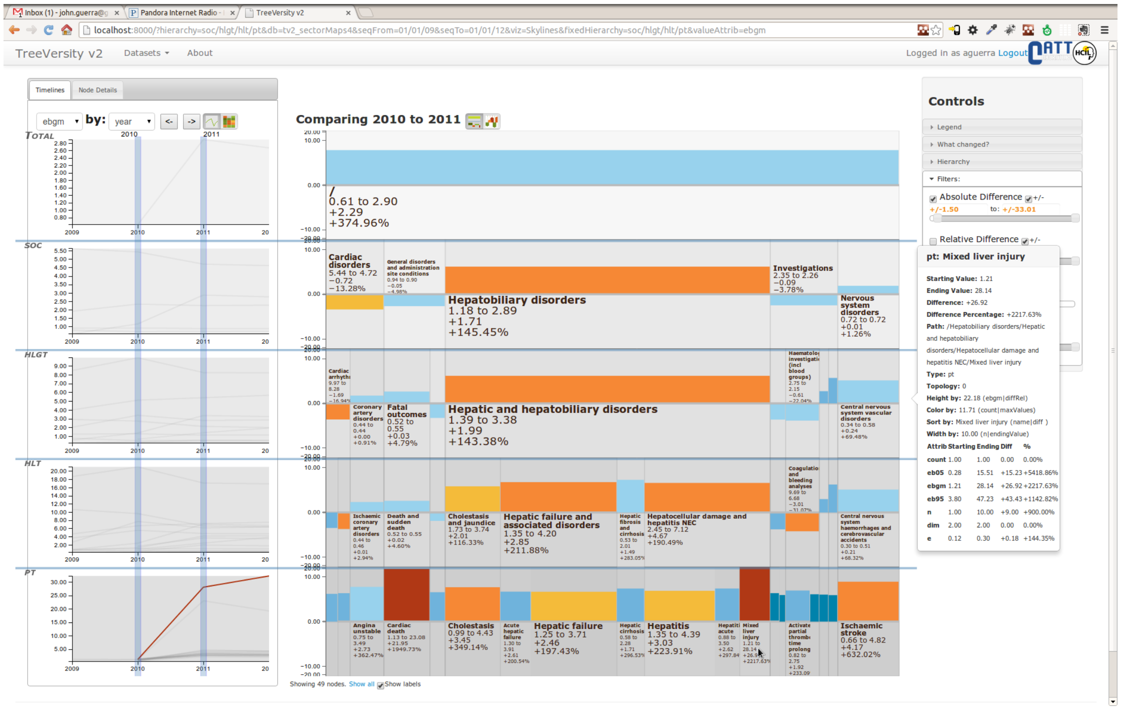
□ Relative Difference ☑+/- +/-0.00% to: +/-+2317.63%

□ Starting Val ☑+/- +/-0.00 to: +/-30.30

□ Ending Val ☑+/- +/-0.00 to: +/-41.03

Showing 1143 nodes. Show all ☑ Show labels

Figure B.16: FDA's fourth dataset one undisclosed drug over four years. Yellow to red boxes represent adverse effects with non-overlapping confidence intervals. The height encodes the actual difference in the *EBGM* value and the width the number of reports *n*. The image shows the change between 2011 and 2010 for all the adverse effects without filters.



☑ Absolute Difference ☑ +/-
+/-1.50 to: +/-33.01

☐ Relative Difference ☑ +/-

pt: Mixed liver injury

Starting Value: 1.21
Ending Value: 28.14
Difference: +26.92
Difference Percentage: +2217.63%
Path: /Hepatobiliary disorders/Hepatic and hepatobiliary disorders/Hepatocellular damage and hepatitis NEC/Mixed liver injury
Type: pt
Topology: 0
Height by: 22.18 (ebgm|diffRel)
Color by: 11.71 (count|maxValues)
Sort by: Mixed liver injury (name|diff)
Width by: 10.00 (n|endingValue)

Attrib	Starting	Ending	Diff	%
count	1.00	1.00	0.00	0.00%
eb05	0.28	15.51	+15.23	+5418.86%
ebgm	1.21	28.14	+26.92	+2217.63%
eb95	3.80	47.23	+43.43	+1142.82%
n	1.00	10.00	+9.00	+900.00%
dim	2.00	2.00	0.00	0.00%
e	0.12	0.30	+0.18	+144.35%

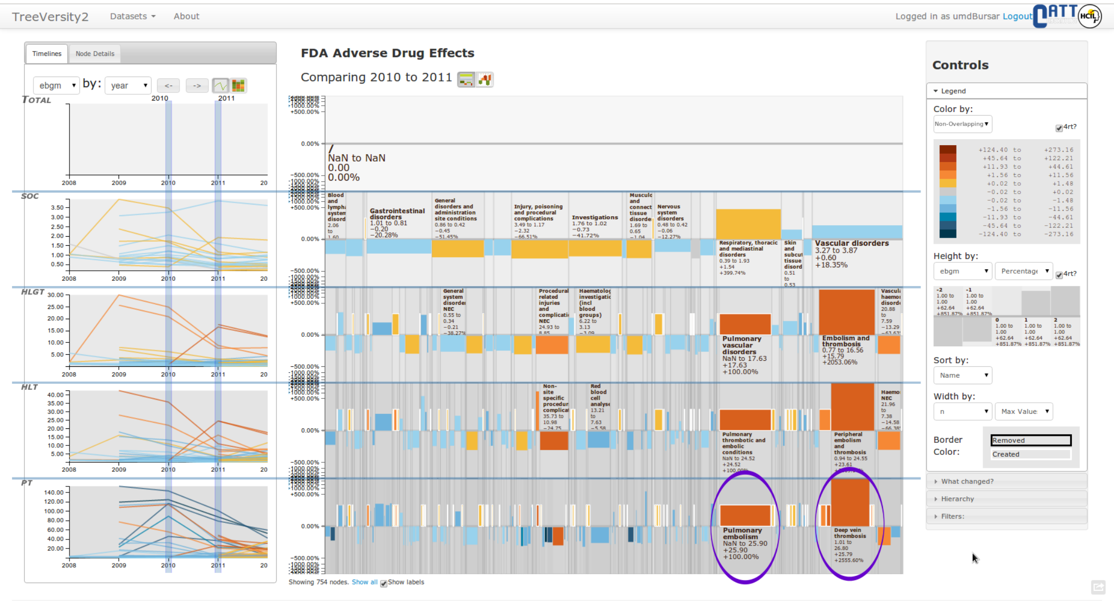
Figure B.17: Same as Figure B.16 after filtering the adverse effects to those having a *EBGM* of less than 1.0 in 2010 and that increased in more than 1.5. The filter shows that "Mixed Liver Injury" and "Cardiac Death" are two adverse effects increasing significantly both because of the height of the boxes and the dark red color.

more reports (with a fourth root scale). Finally the width of the boxes shows the total number of reports by effect, so more significant effects have wider boxes. With these encoding, Dr. Szarfman was able to find that in 2011 the *Pulmonary Embolism* went from not having any reports in 2010 to having a EBGM score of 25.20 which is really bad. She said reported that "*it was incredible that we can see that important effect this way*" and that "*it was significant given the drug in question*". Dr. Szarfman also praised TreeVersity2's visualizations for encoding many of the variables they needed for the comparison in one single view, as well as the possibility of exploring the changes by time, "*It looks awesome!*" she said.

Dr. Szarfman was extremely excited to see the final results, she said "*Awesome findings*" and added "*It looks awesome!*". However she expressed some issues with the color codification representing the non overlapping adverse effects, which is significantly different to what they are used to with the Sector Maps. She said that it might take some training to adjust to that change. She also mentioned that it will be useful to have the StemView boxes width match the areas used for the Sector Maps, and she agreed to sent that information to me. Despite these issues, Dr. Szarfman was very interested in using TreeVersity2 in they day to day work, so a proposal is in the works to get a third party consultant to implement a especially designed version for their needs.

B.9.5 Discussion

TreeVersity2 proved to be a flexible and useful tool to help Dr. Szarfman Szarfman, Medical Officer at the FDA, find significant changes in the adverse effects reported for a drug over years. During this case study five meetings were held, where four different datasets were a analyzed. Dr. Szarfman was extremely excited of the functionality that TreeVersity2 provided, which is significantly better than the comparisons she was having to do with side by side Sector Maps. Many features were added to TreeVersity2 to support this case study and although some limitations were found with the constraints to install the required software at FDA, Dr. Szarfman is looking forward to use TreeVersity2 for her daily work at FDA.



FDA Adverse Drug Effects

Comparing 2010 to 2011

Color by: Non-Overlapping ☑ 4rt?

+124.40	to	+273.16
+45.64	to	+122.21
+11.93	to	+44.61
+1.56	to	+11.56
+0.02	to	+1.48
−0.02	to	+0.02
−0.02	to	−1.48
−1.56	to	−11.56
−11.93	to	−44.61
−45.64	to	−122.21
−124.40	to	−273.16

Height by: ebgm Percentage ☑ 4rt?

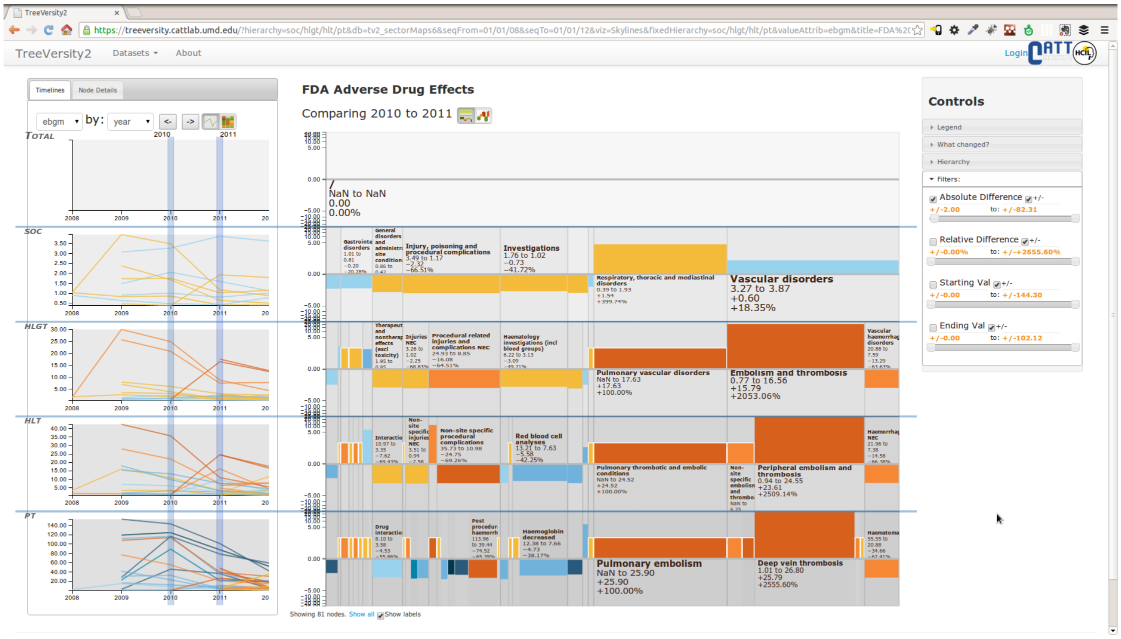
Sort by: Name

Width by: n Max Value

Border Color: Removed Created

Showing 754 nodes. Show all ☑ Show labels

Figure B.18: Changes in the FDA's EBGM index of adverse effects (e.g. Pulmonary Embolism) for a non-disclosed drug between 2011 and 2010 (Sixth dataset). Using the StemView analysts were able to identify two relevant adverse effects that received more reports than expected for 2011, *Pulmonary Embolism* that wasn't reported in 2010 (i.e. created node denoted with white border) and *Deep Vein Thrombosis*. The EBGM index is distributed in a fixed, non-aggregating tree and it is a measure of how many more reports than expected are received for a certain adverse effect. A value of 1.0 indicates that the expected number of reports for a certain adverse effect were received, decreasing values are good. The change of each the index is shown using the height of the boxes, so boxes going up are effect getting worse and boxes going down the opposite. The width of the boxes in the StemView represents the total number of reports, so wide boxes are more important. The color was especially crafted to meet a special requirement from the FDA, to highlight adverse effects with non-overlapping confidence intervals (shown on yellow and red). Therefore, analysts searched for wide, red/yellow boxes going up.



☑ Absolute Difference ☑ +/-
+/-2.00 to: +/-82.31

□ Relative Difference ☑ +/-
+/-0.00% to: +/-+2655.60%

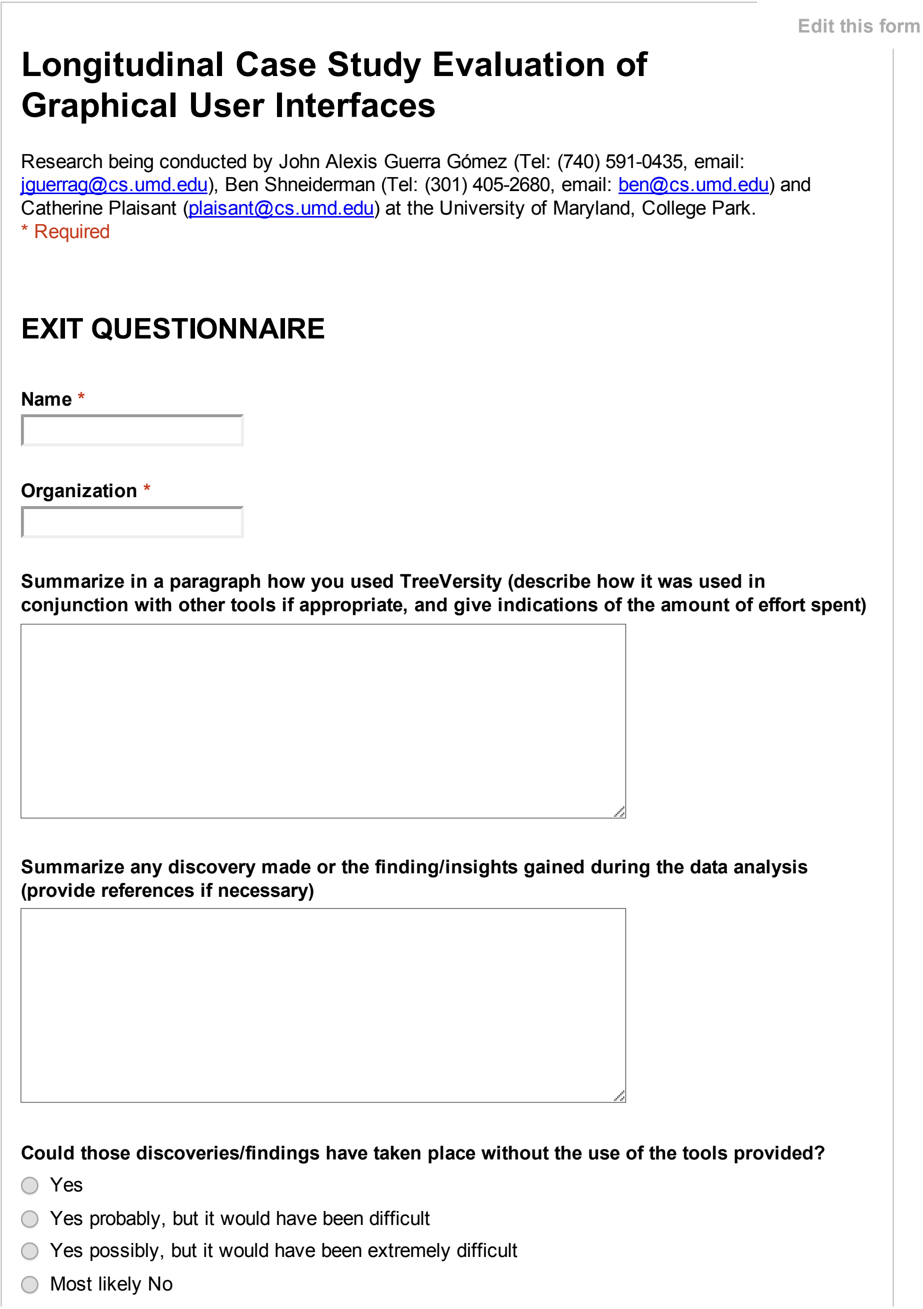
□ Starting Val ☑ +/-
+/-0.00 to: +/-144.30

□ Ending Val ☑ +/-
+/-0.00 to: +/-102.12

Figure B.19: FDA's sixth dataset with the same configuration as in Figure B.18 after filtering for adverse effects changing more than 2.0 in the EBGM value.

Appendix C

TreeVersity Exit Questionnaire



Longitudinal Case Study Evaluation of Graphical User Interfaces

Research being conducted by John Alexis Guerra Gómez (Tel: (740) 591-0435, email: jguerrag@cs.umd.edu), Ben Shneiderman (Tel: (301) 405-2680, email: ben@cs.umd.edu) and Catherine Plaisant (plaisant@cs.umd.edu) at the University of Maryland, College Park.

* Required

EXIT QUESTIONNAIRE

Name *

Organization *

Summarize in a paragraph how you used TreeVersity (describe how it was used in conjunction with other tools if appropriate, and give indications of the amount of effort spent)

Summarize any discovery made or the finding/insights gained during the data analysis (provide references if necessary)

Could those discoveries/findings have taken place without the use of the tools provided?

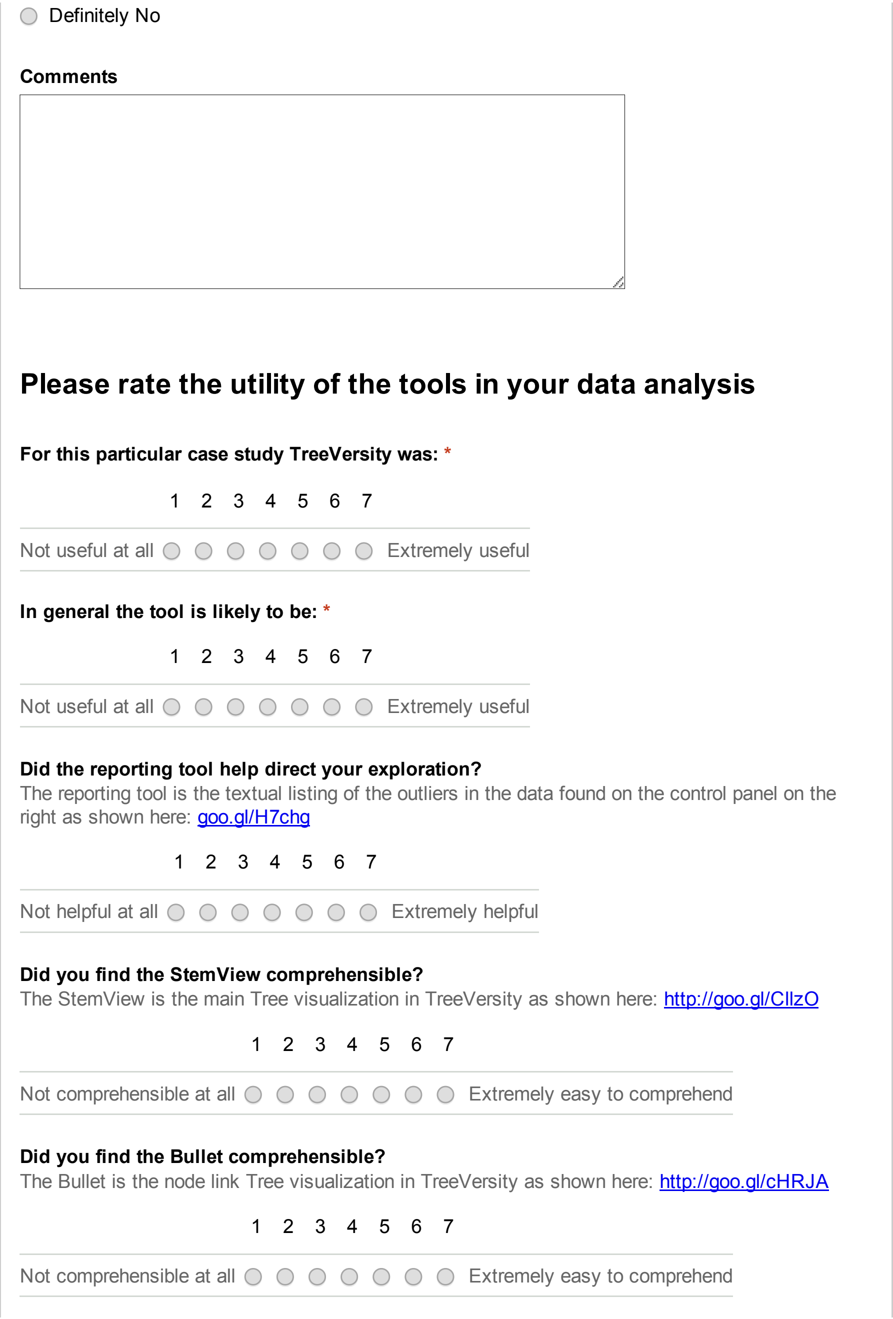
○ Yes

○ Yes probably, but it would have been difficult

○ Yes possibly, but it would have been extremely difficult

○ Most likely No

Figure C.1: Exit questionnaire page 1



○ Definitely No

Comments

Please rate the utility of the tools in your data analysis

For this particular case study TreeVersity was: *

	1	2	3	4	5	6	7	
Not useful at all	○	○	○	○	○	○	○	Extremely useful

In general the tool is likely to be: *

	1	2	3	4	5	6	7	
Not useful at all	○	○	○	○	○	○	○	Extremely useful

Did the reporting tool help direct your exploration?
The reporting tool is the textual listing of the outliers in the data found on the control panel on the right as shown here: goo.gl/H7chg

	1	2	3	4	5	6	7	
Not helpful at all	○	○	○	○	○	○	○	Extremely helpful

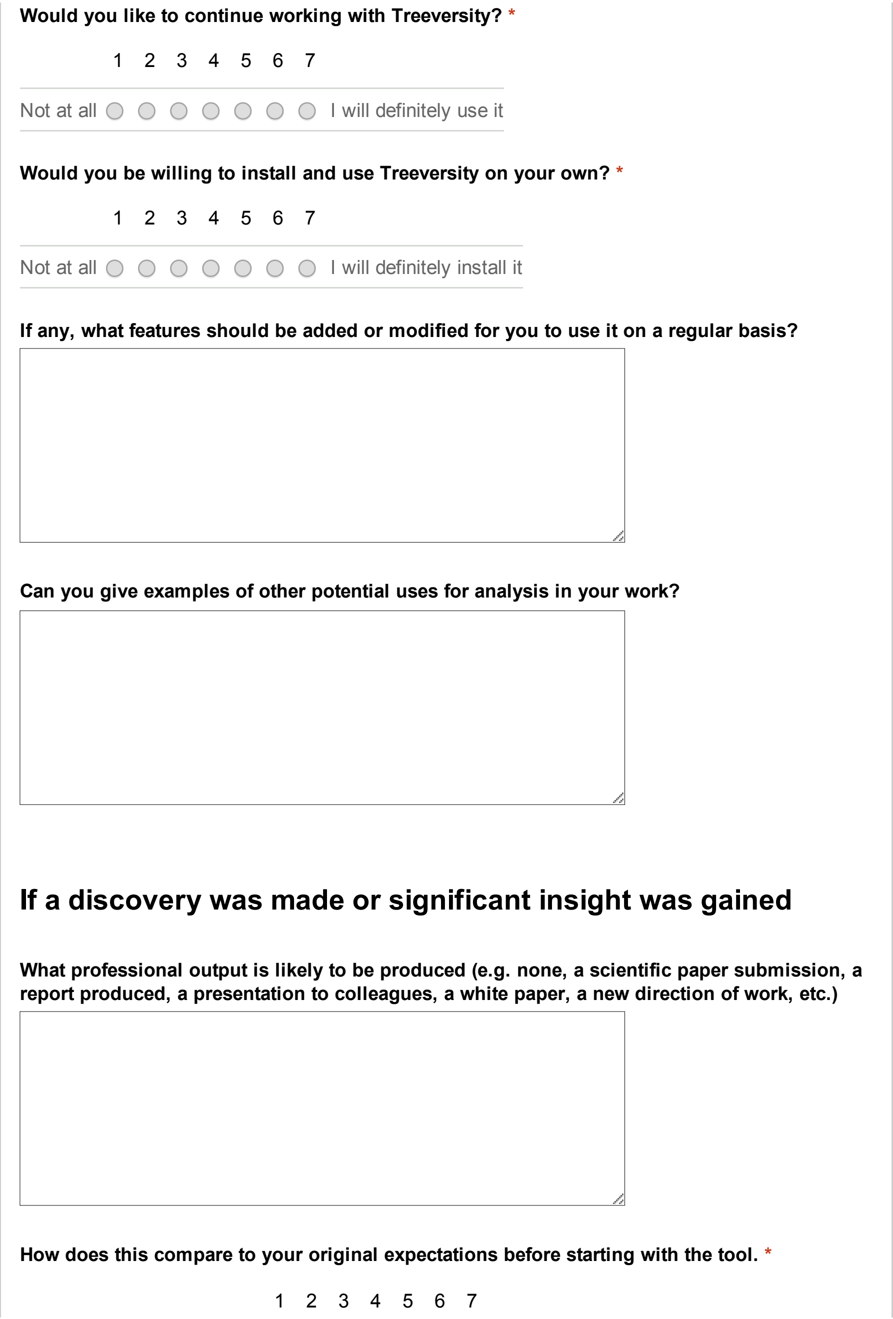
Did you find the StemView comprehensible?
The StemView is the main Tree visualization in TreeVersity as shown here: http://goo.gl/CllzO

	1	2	3	4	5	6	7	
Not comprehensible at all	○	○	○	○	○	○	○	Extremely easy to comprehend

Did you find the Bullet comprehensible?
The Bullet is the node link Tree visualization in TreeVersity as shown here: http://goo.gl/cHRJA

	1	2	3	4	5	6	7	
Not comprehensible at all	○	○	○	○	○	○	○	Extremely easy to comprehend

Figure C.2: Exit questionnaire page 2



Would you like to continue working with Treeversity? *

	1	2	3	4	5	6	7	
Not at all	○	○	○	○	○	○	○	I will definitely use it

Would you be willing to install and use Treeversity on your own? *

	1	2	3	4	5	6	7	
Not at all	○	○	○	○	○	○	○	I will definitely install it

If any, what features should be added or modified for you to use it on a regular basis?

Can you give examples of other potential uses for analysis in your work?

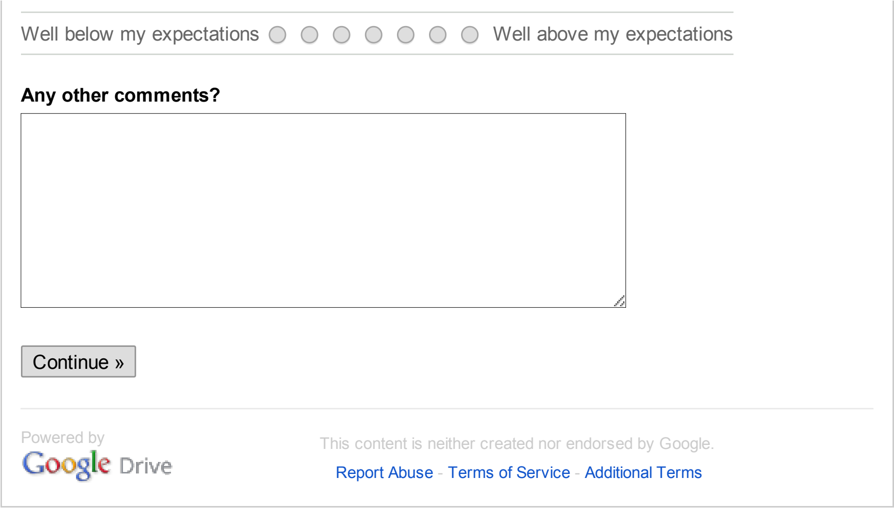
If a discovery was made or significant insight was gained

What professional output is likely to be produced (e.g. none, a scientific paper submission, a report produced, a presentation to colleagues, a white paper, a new direction of work, etc.)

How does this compare to your original expectations before starting with the tool. *

1 2 3 4 5 6 7

Figure C.3: Exit questionnaire page 3



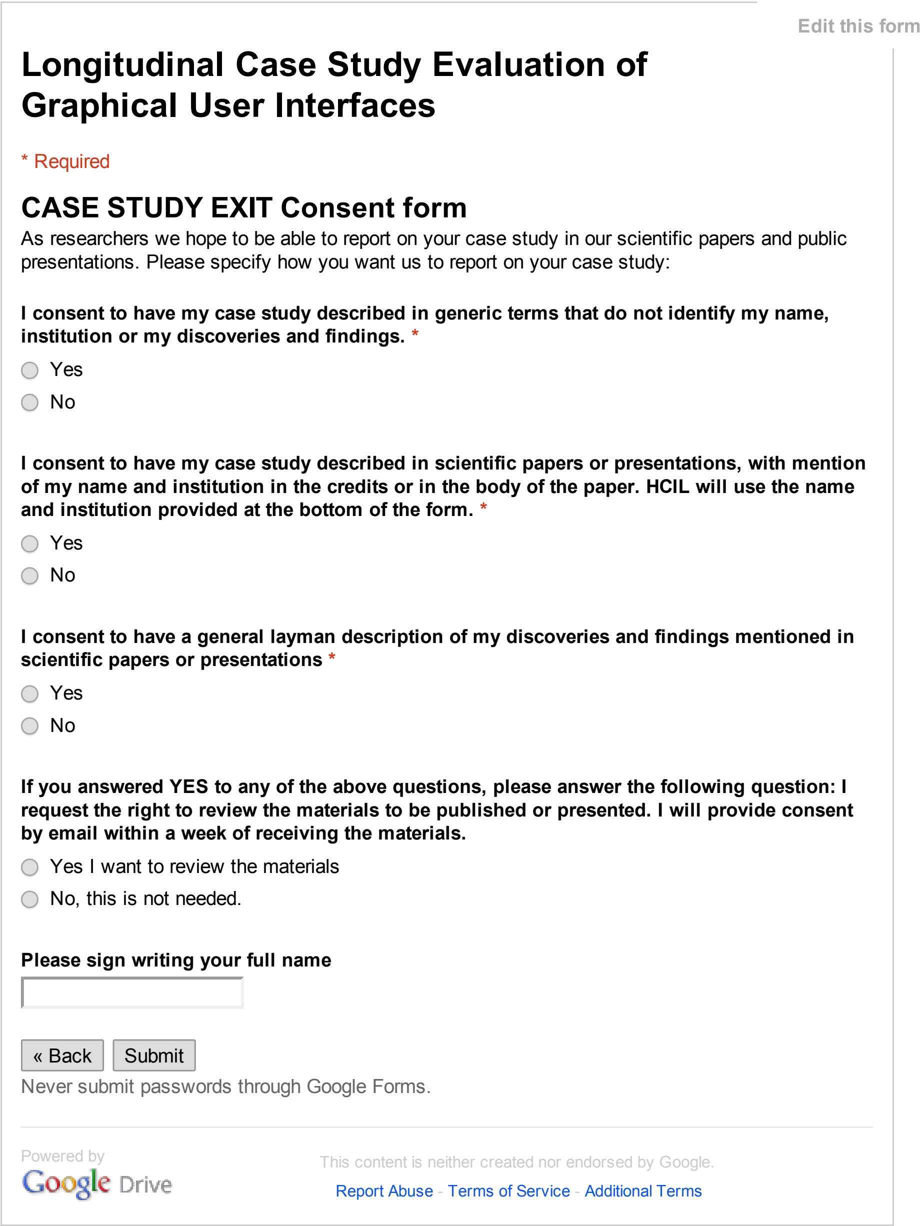
Well below my expectations ○ ○ ○ ○ ○ ○ ○ Well above my expectations

Any other comments?

This content is neither created nor endorsed by Google.

Report Abuse - Terms of Service - Additional Terms

Figure C.4: Exit questionnaire page 5



Longitudinal Case Study Evaluation of Graphical User Interfaces

* Required

CASE STUDY EXIT Consent form

As researchers we hope to be able to report on your case study in our scientific papers and public presentations. Please specify how you want us to report on your case study:

I consent to have my case study described in generic terms that do not identify my name, institution or my discoveries and findings. *

○ Yes

○ No

I consent to have my case study described in scientific papers or presentations, with mention of my name and institution in the credits or in the body of the paper. HCIL will use the name and institution provided at the bottom of the form. *

○ Yes

○ No

I consent to have a general layman description of my discoveries and findings mentioned in scientific papers or presentations *

○ Yes

○ No

If you answered YES to any of the above questions, please answer the following question: I request the right to review the materials to be published or presented. I will provide consent by email within a week of receiving the materials.

○ Yes I want to review the materials

○ No, this is not needed.

Please sign writing your full name

Never submit passwords through Google Forms.

This content is neither created nor endorsed by Google.

Figure C.5: Exit questionnaire page 5

Bibliography

[1] "Bootstrap, from twitter," http://twitter.github.com/bootstrap/. [Online]. Available: http://twitter.github.com/bootstrap/

[2] "d3.js," http://mbostock.github.com/d3/. [Online]. Available: http://mbostock.github.com/d3/

[3] "Facebook developers HomePage," https://developers.facebook.com/. [Online]. Available: https://developers.facebook.com/

[4] "Facebook platform python SDK," https://github.com/facebook/python-sdk. [Online]. Available: https://github.com/facebook/python-sdk

[5] "Google app engine - google code," http://code.google.com/appengine/. [Online]. Available: http://code.google.com/appengine/

[6] N. Amenta and J. Klingner, "Case study: Visualizing sets of evolutionary trees," in *Information Visualization, 2002. INFOVIS 2002. IEEE Symposium on*, 2002, p. 71–74.

[7] D. Auber, M. Delest, J. P. Domenger, P. Ferraro, and R. Strandh, "EVAT: environment for vizualisation and analysis of trees," *IEEE InfoVis Poster Compendium*, p. 124–125, 2003.

[8] G. Battista, "Algorithms for drawing graphs: an annotated bibliography," *Computational Geometry*, vol. 4, no. 5, pp. 235–282, Oct. 1994. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/092577219400014X

[9] ——, "Algorithms for drawing graphs: an annotated bibliography," *Computational Geometry*, vol. 4, no. 5, p. 235–282, Oct. 1994. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/092577219400014X

[10] P. Bille, “A survey on tree edit distance and related problems,” *Theoretical Computer Science*, vol. 337, no. 1-3, pp. 217–239, Jun. 2005. [Online]. Available: http://www.sciencedirect.com/science/article/B6V1G-4FF68TJ-1/2/01da3ee1e6b602d737851fde3040a149

[11] M. Bostock, V. Ogievetsky, and J. Heer, “D&#x0B3; Data-Driven Documents,” *IEEE Transactions on Visualization and Computer Graphics*, vol. 17, no. 12, pp. 2301–2309, Dec. 2011. [Online]. Available: http://ieeexplore.ieee.org/xpl/downloadCitationshttp://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6064996http://vis.stanford.edu/files/2011-D3-InfoVis.pdf

[12] ——, “D&#x0B3; Data-Driven documents,” *IEEE Transactions on Visualization and Computer Graphics*, vol. 17, no. 12, pp. 2301–2309, Dec. 2011. [Online]. Available: http://ieeexplore.ieee.org/xpl/downloadCitations

[13] S. Bremm, T. von Landesberger, M. Hess, T. Schreck, P. Weil, and K. Hamacherk, “Interactive visual comparison of multiple trees.” IEEE, Oct. 2011, pp. 31–40. [Online]. Available: http://ieeexplore.ieee.org/ielx5/6093899/6102423/06102439.pdf?tp=&arnumber=6102439&isnumber=6102423

[14] S. Bremm, T. von Landesberger, and K. Hamacher, “Interactive visual comparison of multiple phylogenetic trees,” http://www.gris.tu-darmstadt.de/research/vissearch/projects/ViPhy/, Oct. 2011. [Online]. Available: http://www.gris.tu-darmstadt.de/research/vissearch/projects/ViPhy/

[15] D. Brodbeck and L. Girardin, “Visualization of large-scale customer satisfaction surveys using a parallel coordinate tree,” in *IEEE Symposium on Information Visualization 2003 (IEEE Cat. No.03TH8714)*. IEEE, pp. 197–201. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=1249026

[16] B. Caemmerer, “Skyline Graphs – New Insights on the Horizon...” 2013. [Online]. Available: http://www.sapdesignguild.org/community/blinks/ui_blinks_gw_03.asp#skyline

[17] S. Card, J. Mackinlay, and B. Shneiderman, *Readings in information visualization: using vision to think*. Morgan Kaufmann, 1999. [Online]. Available: http://books.google.com/books?hl=en&lr=&id=wdh2gqWfQmgC&oi=fnd&pg=

PR13&dq=Readings+in+Information+Visualization:+Using+Vision+to+Think.&ots=olAJ3CmGMz&sig=NEQyA6OjT5yVq8Fp8n1oDktMCLk

[18] ——, *Readings in information visualization: using vision to think*. Morgan Kaufmann, 1999.

[19] S. Card and D. Nation, "Degree-of-interest trees: A component of an attention-reactive user interface," 2002, pp. 231–245. [Online]. Available: http://dl.acm.org/citation.cfm?id=1556300

[20] S. K. Card, B. Suh, B. A. Pendleton, J. Heer, and J. W. Bodnar, "Timetree: exploring time changing hierarchies," vol. 7. IEEE, 2006, pp. 3–10. [Online]. Available: http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:time+tree:+exploring+time+changing+heirarchies#0

[21] ——, "Timetree: exploring time changing hierarchies," in *Proceedings of the IEEE Symposium on Visual Analytics Science and Technology*, vol. 7. IEEE, 2006, p. 3–10. [Online]. Available: http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:time+tree:+exploring+time+changing+heirarchies#0

[22] D. Fisher, R. Dhamija, and M. Hearst, "Animated exploration of dynamic graphs with radial layout," vol. 2001. IEEE, 2001, pp. 43 – 50. [Online]. Available: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=963279

[23] G. W. Furnas and J. Zacks, "Multitrees: enriching and reusing hierarchical structure," in *Proceedings of the SIGCHI conference on Human factors in computing systems: celebrating interdependence*, ser. CHI '94. New York, NY, USA: ACM, 1994, p. 330–336, ACM ID: 191778.

[24] M. Ghoniem and J. D. Fekete, "Animating treemaps," in *Proc. of 18th HCIL Symposium-Workshop on Treemap Implementations and Applications*, 2001.

[25] M. Gleicher, D. Albers, R. Walker, I. Jusufi, C. D. Hansen, and J. C. Roberts, "Visual Comparison for Information Visualization," *Information Visualization*, 2011. [Online]. Available: http://graphics.cs.wisc.edu/Papers/2011/GAWJHR11https://graphics.cs.wisc.edu/Papers/2011/GAWJHR11/paper.pdf

[26] M. Graham and J. Kennedy, “Combining linking and focusing techniques for a multiple hierarchy visualisation,” in *Information Visualisation, 2001. Proceedings. Fifth International Conference on*, 2001, p. 425–432.

[27] ——, “Extending taxonomic visualisation to incorporate synonymy and structural markers,” *Information Visualization*, vol. 4, no. 3, p. 206–223, 2005.

[28] ——, “Exploring multiple trees through DAG representations,” *IEEE Transactions on Visualization and Computer Graphics*, p. 1294–1301, 2007.

[29] ——, “A survey of multiple tree visualisation,” *Information Visualization*, 2009. [Online]. Available: http://www.palgrave-journals.com/ivs/journal/vaop/ncurrent/abs/ivs200929a.html

[30] ——, “A survey of multiple tree visualisation,” *Information Visualization*, 2009.

[31] M. Graham, J. B. Kennedy, and C. Hand, “A comparison of set-based and graph-based visualisations of overlapping classification hierarchies,” in *Proceedings of the working conference on Advanced visual interfaces*, 2000, p. 41–50.

[32] M. Graham, M. F. Watson, and J. B. Kennedy, “Novel visualisation techniques for working with multiple, overlapping classification hierarchies,” *Taxon*, vol. 51, no. 2, p. 351–358, 2002.

[33] M. Graham and J. Kennedy, “Multiform Views of Multiple Trees,” *Proceedings of the 2008 12th International Conference Information Visualisation*, pp. 252–257, 2008. [Online]. Available: http://portal.acm.org/citation.cfm?id=1439280.1440153http://portal.acm.org/citation.cfm?id=1440153

[34] J. A. Guerra-Gomez, A. Buck-Coleman, C. Plaisant, and B. Shneiderman, “TreeVersity: Interactive Visualizations for Comparing Two Trees with Structure and Node Value Changes,” 2012. [Online]. Available: http://hcil2.cs.umd.edu/trs/2012-04/2012-04.pdf

[35] J. A. Guerra Gómez, M. L. Pack, C. Plaisant, and B. Shneiderman, “Visualizing changes over time in datasets using dynamic hierarchies,” 2013. [Online]. Available: http://hcil2.cs.umd.edu/trs/2013-06/2013-06.pdf

[36] J. A. Guerra Gómez, K. Wongsuphasawat, T. D. Wang, M. L. Pack, and C. Plaisant, "Analyzing incident management event sequences with interactive visualization," in *Transportation Research Board 90th Annual Meeting Compendium of Papers*, 2011.

[37] J. Guerra Gómez, "MySocialTree: browising the facebook feed using hierarchies."

[38] J. Guerra Gómez, A. Buck-Coleman, C. Plaisant, and B. Shneiderman, "TreeVersity: comparing tree structures by topology and node's attributes differences."

[39] J. Guerra Gómez, C. Plaisant, B. Shneiderman, and A. Buck-Coleman, "Interactive visualizations for comparing two trees with structure and node value changes."

[40] HCIL, "Infovis benchmark - PairWise comparison of trees," http://www.cs.umd.edu/hcil/InfovisRepository/contest-2003/, Aug. 2011. [Online]. Available: http://www.cs.umd.edu/hcil/InfovisRepository/contest-2003/

[41] J. Heer and S. Card, "DOITrees revisited: scalable, space-constrained visualization of hierarchical data," in *Proceedings of the working conference on Advanced visual interfaces*, 2004, p. 421–424.

[42] J. Heer and S. K. Card, "DOITrees revisited: scalable, space-constrained visualization of hierarchical data," 2004, pp. 421–424. [Online]. Available: http://delivery.acm.org/10.1145/990000/989941/p421-heer.pdf?ip=128.8.127.181&acc=ACTIVESERVICE&CFID=79010598&CFTOKEN=15248593&__acm__=1327094423_fdc44ad7184d10e8848d4742c707a3bbhttp://dl.acm.org/citation.cfm?id=989941

[43] I. Herman, G. Melancon, and M. S. Marshall, "Graph visualization and navigation in information visualization: A survey," *IEEE Transactions on Visualization and Computer Graphics*, vol. 6, no. 1, pp. 24–43, 2000. [Online]. Available: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=841119http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=841119

[44] ——, "Graph visualization and navigation in information visualization: A survey," *IEEE Transactions on Visualization and Computer Graphics*, vol. 6, no. 1, p. 24–43, 2000. [Online]. Available: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=841119

[45] D. Holten, "Hierarchical edge bundles: Visualization of adjacency relations in hierarchical data," *IEEE Transactions on Visualization and Computer Graphics*, vol. 12, no. 5, pp. 741–748, Oct. 2006.

[46] D. Holten and J. J. van Wijk, "Visual comparison of hierarchically organized data," *Computer Graphics Forum*, vol. 27, no. 3, pp. 759–766, May 2008. [Online]. Available: http://onlinelibrary.wiley.com/doi/10.1111/j.1467-8659.2008.01205.x/full

[47] J. Y. Hong, J. D'Andries, M. Richman, and M. Westfall, "Zoomology: comparing two large hierarchical trees," *Poster Compendium of IEEE Information Visualization*, 2003.

[48] W. Huang, P. Eades, and S.-H. Hong, "Beyond time and error," in *Proceedings of the 2008 conference on BEyond time and errors novel evaLuation methods for Information Visualization - BELIV '08*. New York, New York, USA: ACM Press, Apr. 2008, p. 1. [Online]. Available: http://dl.acm.org/citation.cfm?id=1377966.1377970

[49] J. Hullman, N. Diakopoulos, and E. Adar, "Contextifier: Automatic Generation of Annotated Stock Visualizations," in *Proceedings of the 31th ACM Conference on Human Factors in Computing Systems (CHI)*, 2013, pp. 2707–2716. [Online]. Available: http://misc.si.umich.edu/media/papers/vis_messaging_CHI_20130120_submit.pdf

[50] B. Johnson and B. Shneiderman, "Tree-maps: A space-filling approach to the visualization of hierarchical information structures," in *Proceedings of the IEEE Conference on Visualization (Vis)*. IEEE, 1991, pp. 284 – 291. [Online]. Available: http://portal.acm.org/citation.cfm?id=949654&amp;dl=

[51] S. Jurgensmann and H. Schulz, "A Visual Survey of Tree Visualization," *Proceedings of IEEE Information Visualization (Salt Lake City, USA, 2010), IEEE Press*, 2010. [Online]. Available: http://www.informatik.uni-rostock.de/~hs162/treeposter/oldposter/treevis_abstract.pdfhttp://vcg.informatik.uni-rostock.de/~hs162/treeposter/oldposter/poster.html

[52] S. J\ürgensmann and H. Schulz, "Poster: a visual survey of tree visualization," *Proceedings of IEEE Information Visualization (Salt Lake City, USA, 2010), IEEE Press*.

[53] N. Kong and M. Agrawala, "Graphical Overlays: Using Layered Elements to Aid Chart Reading," *Visualization and Computer Graphics, IEEE Transactions on*, vol. 18, no. 12, pp. 2631 – 2638, 2012. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6327269

[54] J. B. Kruskal and J. M. Landwehr, "Icicle Plots: Better Displays for Hierarchical Clustering," *The American Statistician*, vol. 37, no. 2, pp. 162 – 168, 1983. [Online]. Available: http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:Icicle+Plots:+Better+Displays+for+Hierarchical+Clustering#0

[55] ——, "Icicle plots: Better displays for hierarchical clustering," *The American Statistician*, vol. 37, no. 2, pp. 162 – 168, 1983. [Online]. Available: http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:Icicle+Plots:+Better+Displays+for+Hierarchical+Clustering#0

[56] J. Lamping, "The Hyperbolic Browser: A Focus+Context Technique for Visualizing Large Hierarchies," *Journal of Visual Languages & Computing*, vol. 7, no. 1, pp. 33–55, Mar. 1996. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S1045926X96900038

[57] B. Lee, G. G. Robertson, M. Czerwinski, and C. S. Parr, "CandidTree: visualizing structural uncertainty in similar hierarchies," *Information Visualization*, vol. 6, no. 3, p. 233–246, 2007.

[58] T. Margush and F. R. McMorris, "Consensusn-trees," *Bulletin of Mathematical Biology*, vol. 43, no. 2, pp. 239–244, Mar. 1981. [Online]. Available: http://www.springerlink.com/content/3311mu4180803341/

[59] P. McLachlan, T. Munzner, E. Koutsofios, and S. North, "LiveRAC," in *Proceeding of the twenty-sixth annual CHI conference on Human factors in computing systems - CHI '08*. New York, New York, USA: ACM Press, Apr. 2008, p. 1483. [Online]. Available: http://dl.acm.org/citation.cfm?id=1357054.1357286

[60] M. J. Mohammadi-Aragh and T. J. Jankun-Kelly, “MoireTrees: visualization and interaction for multi-hierarchical data,” 2005.

[61] D. R. Morse, N. Ytow, D. M. Roberts, and A. Sato, “Comparison of multiple taxonomic hierarchies using TaxoNote,” in *Compendium of Symposium on Information Visualization*, 2003, p. 126–127.

[62] T. Munzner, F. Guimbretière, S. Tasiran, L. Zhang, and Y. Zhou, “TreeJuxtaposer: scalable tree comparison using Focus+Context with guaranteed visibility,” in *ACM Transactions on Graphics*, vol. 22, no. 3. San Diego, California: ACM, 2003, pp. 453–462. [Online]. Available: http://portal.acm.org/citation.cfm?id=1201775.882291http://portal.acm.org/ft_gateway.cfm?id=882291&type=pdf&coll=GUIDE&dl=GUIDE&CFID=109923626&CFTOKEN=71313789http://portal.acm.org/citation.cfm?doid=882262.882291

[63] T. Munzner, F. Guimbretière, S. Tasiran, L. Zhang, and Y. Zhou, “TreeJuxtaposer: scalable tree comparison using Focus+Context with guaranteed visibility,” *ACM Transactions on Graphics*, vol. 22, no. 3, p. 453, 2003. [Online]. Available: http://portal.acm.org/citation.cfm?doid=882262.882291

[64] D. Nation, D. Roberts, and S. Card, “Browse hierarchical data with the degree of interest tree,” *submitted to CHI*, 2002.

[65] C. North, “Toward measuring visualization insight,” *IEEE Computer Graphics and Applications*, vol. 26, no. 3, pp. 6–9, May 2006. [Online]. Available: http://ieeexplore.ieee.org/xpl/articleDetails.jsp?arnumber=1626178

[66] C. S. Parr, B. Lee, D. Campbell, and B. B. Bederson, “Visualizations for taxonomic and phylogenetic trees,” *Bioinformatics*, vol. 20, no. 17, p. 2997, 2004. [Online]. Available: http://bioinformatics.oxfordjournals.org/content/20/17/2997.full.pdfhttp://bioinformatics.oxfordjournals.org/content/20/17/2997.short

[67] ——, “Visualizations for taxonomic and phylogenetic trees,” *Bioinformatics*, vol. 20, no. 17, p. 2997, 2004.

[68] A. Perer and B. Shneiderman, “Integrating statistics and visualization,” in *Proceeding of the twenty-sixth annual CHI conference on Human factors in*

computing systems - CHI '08. New York, New York, USA: ACM Press, Apr. 2008, p. 265. [Online]. Available: http://dl.acm.org/citation.cfm?id=1357054.1357101

[69] C. Plaisant, J. Grosjean, and B. B. Bederson, "SpaceTree: supporting exploration in large node link tree, design evolution and empirical evaluation," in *Proceedings of the IEEE Symposium on Information Visualization*. IEEE, 1998, p. 57–64. [Online]. Available: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=1173148

[70] ——, "SpaceTree: supporting exploration in large node link tree, design evolution and empirical evaluation." IEEE, 1998, pp. 57–64. [Online]. Available: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=1173148http://hcil.cs.umd.edu/trs/2002-05/2002-05.pdfhttp://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=1173148http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=1173148&tag=1

[71] S. A. Rivkees and A. Szarfman, "Dissimilar hepatotoxicity profiles of propylthiouracil and methimazole in children." *The Journal of clinical endocrinology and metabolism*, vol. 95, no. 7, pp. 3260–7, Jul. 2010. [Online]. Available: http://jcem.endojournals.org/content/95/7/3260.full.pdf+html

[72] G. Robertson, J. Mackinlay, and S. Card, "Cone trees: animated 3D visualizations of hierarchical information," 1991, pp. 189–194. [Online]. Available: http://dl.acm.org/citation.cfm?id=108883

[73] P. Saraiya, C. North, and K. Duca, "An Insight-Based Longitudinal Study of Visual Analytics," *IEEE Transactions on Visualization and Computer Graphics*, vol. 12, no. 6, pp. 1511–1522, Nov. 2006. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=1703371

[74] H.-J. Schulz, S. Hadlak, and H. Schumann, "Point-based tree representation: A new approach for large hierarchies." IEEE, Apr. 2009, pp. 81–88. [Online]. Available: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=4906841

[75] H. Schulz, S. Hadlak, and H. Schumann, "Point-based tree representation: A new approach for large hierarchies," in *Proceedings of the IEEE Pacific*

Visualization Symposium. IEEE, Apr. 2009, p. 81–88. [Online]. Available: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=4906841

[76] J. S. J. Seo and B. Shneiderman, "Knowledge discovery in high-dimensional data: case studies and a user survey for the rank-by-feature framework." *IEEE Transactions on Visualization and Computer Graphics*, vol. 12, no. 3, pp. 311–322, 2006. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/16640245

[77] B. Shneiderman, "Tree visualization with tree-maps: 2-d space-filling approach," *ACM Transactions on graphics (TOG)*, vol. 11, pp. 92–99, 1992. [Online]. Available: http://citeseer.ist.psu.edu/viewdoc/summary?doi=10.1.1.29.1549http://citeseerx.ist.psu.edu/viewdoc/download;jsessionid=C0276DE28F82A834AC81CB3391411583?doi=10.1.1.29.1549&rep=rep1&type=pdfhttp://dl.acm.org/citation.cfm?id=115768

[78] B. Shneiderman and C. Plaisant, "Strategies for evaluating information visualization tools," in *Proceedings of the 2006 AVI workshop on BEyond time and errors novel evaluation methods for information visualization - BELIV '06*. New York, New York, USA: ACM Press, May 2006, p. 1. [Online]. Available: http://dl.acm.org/citation.cfm?id=1168149.1168158

[79] ——, "Strategies for evaluating information visualization tools: multi-dimensional in-depth long-term case studies," in *Proceedings of the 2006 AVI workshop on BEyond time and errors: novel evaluation methods for information visualization*, ser. BELIV '06. New York, NY, USA: ACM, 2006, p. 1–7. [Online]. Available: http://doi.acm.org/10.1145/1168149.1168158

[80] M. Spenke, "Visualization and interactive analysis of blood parameters with InfoZoom," *Artificial Intelligence in Medicine*, vol. 22, no. 2, pp. 159–172, May 2001. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0933365700001056

[81] C. Stockham, L. Wang, and T. Warnow, "Statistically based postprocessing of phylogenetic analysis by clustering," *Bioinformatics*, vol. 18, no. Suppl 1, pp. S285–S293, Jul. 2002. [Online]. Available: http://bioinformatics.oxfordjournals.org/content/18/suppl_1/S285.short

[82] A. Szarfman, J. M. Tonning, J. G. Levine, and P. M. Doraiswamy, “Atypical antipsychotics and pituitary tumors: a pharmacovigilance study.” *Pharmacotherapy*, vol. 26, no. 6, pp. 748–58, Jul. 2006. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/16716128

[83] J. L. Thorley and R. D. Page, “RadCon: phylogenetic tree comparison and consensus,” *Bioinformatics*, vol. 16, no. 5, p. 486, 2000.

[84] Y. Tu and H. Shen, “Visualizing changes of hierarchical data using treemaps,” *Visualization and Computer Graphics, IEEE Transactions on*, vol. 13, no. 6, pp. 1286–1293, 2007.

[85] E. R. A. Valiati, C. M. D. S. Freitas, and M. S. Pimenta, “Using multi-dimensional in-depth long-term case studies for information visualization evaluation,” in *Proceedings of the 2008 conference on BEyond time and errors novel evaLuation methods for Information Visualization - BELIV ’08*. New York, New York, USA: ACM Press, Apr. 2008, p. 1. [Online]. Available: http://dl.acm.org/citation.cfm?id=1377966.1377978

[86] M. Wattenberg, “Visualizing the stock market,” 1999, pp. 188–189. [Online]. Available: http://portal.acm.org/citation.cfm?id=632716.632834

[87] ——, “Visualizing the stock market,” in *CHI’99 extended abstracts on Human factors in computing systems*, 1999, p. 188–189.

[88] K. Wongsuphasawat, J. A. Gomez, C. Plaisant, T. D. Wang, B. Shneiderman, and M. Taieb-Maimon, “LifeFlow: visualizing an overview of event sequences,” in *Proceeding of the twenty-ninth annual SIGCHI conference on Human factors in computing systems. ACM*, 2011.

[89] K. Wongsuphasawat, J. A. Guerra Gómez, C. Plaisant, T. D. Wang, M. Taieb-Maimon, and B. Shneiderman, “LifeFlow: visualizing an overview of event sequences,” in *Proceedings of the 2011 annual conference on Human factors in computing systems - CHI ’11*. New York, New York, USA: ACM Press, 2011, p. 1747. [Online]. Available: http://dl.acm.org/citation.cfm?id=1979196http://dl.acm.org/citation.cfm?doid=1978942.1979196

[90] Ying Tu and Han-Wei Shen, “Visualizing Changes of Hierarchical Data using Treemaps,” *Visualization and Computer Graphics, IEEE Transactions on*, vol. 13, no. 6, pp. 1286–1293, 2007. [Online]. Available: http://ieeexplore.ieee.org/stampPDF/getPDF.jsp?tp=&arnumber=4376152