

A REUSABLE FRAMEWORK FOR SECURITY DATASET ANALYSIS

Approved by:

---

Dr. Tyler Moore

---

Dr. Suku Nair

---

Dr. Frank Coyle

---

Dr. Michael Hahsler

---

Dr. Scott Dynes

A REUSABLE FRAMEWORK FOR SECURITY DATASET ANALYSIS

A Dissertation Presented to the Graduate Faculty of the  
Bobby B. Lyle School of Engineering: Department of Computer Science  
Southern Methodist University

in

Partial Fulfillment of the Requirements

for the degree of

Doctor of Engineering

with a

Major in Software Engineering

by

Lewis A. Sykalski

(B.S. E.E., Computer Science, University of Wisconsin, 2002)  
(M.S. Software Engineering, Southern Methodist University, 2009)

December 19, 2015

## ACKNOWLEDGMENTS

I would like to thank my doctoral advisor, Dr. Tyler Moore, for his continual advice and support in bringing this to fruition. Whether it be steering my course, lending advice regarding the design of the framework, or patiently proofreading my presentation material his involvement has been instrumental. Furthermore, I'd like to thank him for suggesting this research, as it has proved to be a very engaging topic. While engrossing much of my free time, the knowledge gained will be key to later endeavors, and I am confident the work detailed herein will live on in Dr. Moore's later endeavors.

I would also like to thank members of the HACNET Security group for their data which fed into my research – namely the CMS dataset. In particular, Marie Vasek and John Wadleigh for providing me the time-based data which stressed the nonfunctional performance limits of the framework, thus uncovering problems that were necessary to resolve.

I would also like to thank my examiners Dr. Suku Nair, Dr. Scott Dynes, Dr. Michael Hahsler, and Dr. Frank Coyle for their participation in my committee. Hopefully, the end result reflects your inputs. Special thanks goes to Dr. Dynes, who was willing to participate in the committee at the last moment.

I am also grateful for the data provided by the non-profit group Privacy Rights Clearinghouse, their stated mission being to educate and empower individuals to protect their privacy. Additionally, the keepers of Data Driven Security (Bob Rudis & Jay Jacobs) who's honeypot dataset provided further validation of the reusability and extensibility characteristics of the framework.

Finally, I would like to thank my family. My dad for keeping me on the straight

and narrow, and guiding me in the right direction. My beautiful and understanding girlfriend Emily for encouraging me in all my endeavors and helping me by proof-reading my work. And to my loving children, Leah, Lexie, Lauren, and Logan for their unconditional love, support and admiration.

Sykalski , Lewis A. B.S. E.E., Computer Science, University of Wisconsin, 2002  
M.S. Software Engineering, Southern Methodist University, 2009

A Reusable Framework for Security Dataset Analysis

Advisor: Professor Tyler Moore

Doctor of Engineering degree conferred December 19, 2015

Dissertation completed November 20, 2015

## SUMMARY

Visualizations provide researchers a window into understanding interactions within datasets which are difficult to recognize by simply looking at the data. While the value of visualizations is well-known, creating custom visualizations can be costly in both time and resources. Additionally, their one-time static nature limits their utility as empirical analysis is often an iterative process. As research becomes data-driven, there is an increasing need for tools to visualize these datasets.

This praxis describes the creation of a reusable and extensible cybersecurity analysis framework. In addition to a host of visualizations, we provide user interaction through filtering controls, which enable the user to dynamically control the data flow into the visualization. Additionally, side-by-side portaling and odds ratio plots allow the user to perform case-control comparison in order to determine what factors are important between a given treatment and control dataset.

The utility of the framework is then demonstrated by customizing the framework for two separate case studies of data previously collected for research projects undertaken at SMU. In the first case study, a dataset of privacy breaches including attributes of publicly traded firms is used. The second case study examines a dataset on webserver compromises. In both case studies, the data is arranged in case-control study format, where attributes such as content management system type and version

for uncompromised webservers are compared to compromised ones. The framework thus enables examination of distinct sources (e.g., phishing, malware) as well as time-based analysis to present how risk factors evolve. Finally, the framework's reusable nature is further demonstrated by applying to a third dataset, collected external to the institution. The discussion concludes by contrasting with typical analysis methods to ascertain the value added from the framework.

## TABLE OF CONTENTS

LIST OF FIGURES .....	xv
LIST OF TABLES .....	xx
CHAPTER	
1. INTRODUCTION .....	1
1.1. Problem Statement .....	1
1.2. Prior Work .....	3
1.2.1. Origins of Modern-Visualization .....	3
1.2.2. Visualization Framework Research .....	4
1.2.2.1. Operator-Based Models .....	4
1.2.2.2. Reusable Frameworks .....	4
1.2.2.3. Distributed Frameworks .....	5
1.2.3. Application-Based Frameworks .....	6
1.2.3.1. Financial Visualization .....	6
1.2.3.2. Security Visualization .....	7
1.3. Contribution and Structure .....	7
2. REQUIREMENTS & TRADE STUDY .....	10
2.1. Requirements .....	10
2.1.1. Functional Requirements .....	10
2.1.2. Non-Functional Requirements .....	11
2.1.3. Wants .....	12
2.2. Background .....	12
2.3. Comparison Criteria .....	13

2.4.	Commercial Tools .....	15
2.5.	Open Source Tools .....	16
2.6.	Down Select.....	16
2.7.	Tool Selection .....	21
3.	FRAMEWORK DESIGN AND ARCHITECTURE.....	22
3.1.	System-level Design .....	22
3.2.	Component-level Discussion .....	23
3.2.1.	RapidMiner Discussion .....	23
3.2.1.1.	Process Design .....	24
3.2.1.2.	Plot View.....	25
3.2.1.3.	R-Extension .....	26
3.2.1.4.	Reporting-Extension .....	27
3.2.2.	RapidAnalytics Discussion .....	27
3.2.3.	Apache Web Server Discussion .....	30
3.2.4.	Security Database Discussion.....	30
3.2.5.	Repository Discussion .....	31
3.2.6.	Web Client .....	31
3.3.	Visualization Design.....	32
3.3.1.	Tabular View .....	34
3.3.1.1.	Visualization.....	34
3.3.1.2.	Process Design .....	35
3.3.2.	Aggregate Pie Plot .....	36
3.3.2.1.	Visualization.....	36
3.3.2.2.	Process Design .....	36



3.3.3.	Time Aggregate Bar Plot .....	41
3.3.3.1.	Visualization.....	41
3.3.3.2.	Process Design.....	41
3.3.4.	Time Line Plot .....	44
3.3.4.1.	Visualization.....	44
3.3.4.2.	Process Design.....	44
3.3.5.	Mosaic Plot .....	48
3.3.5.1.	Visualization.....	48
3.3.5.2.	Process Design.....	48
3.3.6.	Box Plot .....	52
3.3.6.1.	Visualization.....	52
3.3.7.	Process Design .....	52
3.3.8.	Odds Ratio Plot .....	55
3.3.8.1.	Visualization.....	55
3.3.8.2.	Process Design.....	56
3.3.9.	Time-Based Odds Ratio Plot .....	58
3.3.9.1.	Visualization.....	58
3.3.9.2.	Process Design.....	59
3.3.10.	Geospatial View .....	61
3.3.10.1.	Visualization.....	61
3.3.10.2.	Process Design.....	61
3.3.11.	Web Service Parameters.....	65
3.4.	User-Interface Design Features.....	68
3.4.1.	Portaling .....	68

3.4.2.	Attribute Filtering .....	68
3.4.3.	Time-Based Filtering.....	69
3.4.4.	Download Dataset .....	69
3.4.5.	Download Chart Data.....	70
3.4.6.	Publishing .....	71
3.5.	Installation & Deployment.....	71
3.6.	Extending the Framework .....	71
3.7.	Source-Code .....	71
4.	BREACH CASE STUDY DESCRIPTION .....	72
4.1.	Background .....	72
4.2.	Dataset Description .....	73
4.3.	Data Aggregation .....	75
4.4.	Analysis Goals .....	78
4.5.	Methodology .....	80
4.6.	Data Limitations .....	81
5.	ANALYSIS OF BREACH DATASET .....	83
5.1.	Case-Control Study Comparative Analysis .....	83
5.1.1.	Proportional Analysis .....	84
5.1.2.	Odds Ratio Analysis .....	87
5.2.	Breach-Only Analysis .....	92
5.2.1.	Two-Way Categorical Analysis .....	92
5.3.	Geographic Analysis.....	97
5.3.1.	Median Analysis .....	100
5.4.	Time-Based Analysis .....	102

5.5.	Predictor Summary .....	104
5.6.	Follow-On Research .....	105
6.	CMS CASE STUDY DESCRIPTION .....	107
6.1.	Background .....	107
6.2.	Dataset Description .....	108
6.3.	Data Aggregation .....	109
6.4.	Analysis Goals .....	110
6.5.	Methodology .....	110
6.6.	Data Limitations .....	111
7.	ANALYSIS OF CMS DATASET .....	112
7.1.	Case-Control Study Comparative Analysis .....	112
7.1.1.	Proportional Analysis .....	113
7.1.2.	Odds Ratio Analysis .....	116
7.2.	Compromise-Only Analysis .....	118
7.2.1.	Two-Way Categorical Analysis .....	119
7.3.	Time-Based Analysis .....	122
7.4.	Geospatial Analysis .....	125
7.5.	Predictor Summary .....	127
7.6.	Follow-On Research .....	127
8.	HONEYPOT DATASET .....	129
8.1.	Background .....	129
8.2.	Dataset Description .....	131
8.3.	Data Aggregation .....	131
8.4.	Analysis Goals .....	132

8.5. Methodology .....	132
8.6. Data Limitations .....	133
8.7. Framework Analysis .....	133
8.7.1. Proportional Analysis .....	133
8.7.2. Geospatial Analysis .....	136
8.7.3. Time-Based Analysis .....	139
8.7.4. Two-Way Categorical Analysis .....	142
8.8. Predictor Summary .....	143
8.9. Follow-On Research .....	144
9. CONCLUSION AND FUTURE WORK .....	146
9.1. Reusability Analysis .....	146
9.2. Benefits Over Traditional Analysis .....	147
9.3. Limitations .....	148
9.4. Future Work .....	149
APPENDIX	
A. Deployment .....	151
A.1. Deployment Overview .....	151
A.2. Requirements/Dependencies .....	151
A.3. Download Locations .....	151
A.4. RapidAnalytics Overview .....	152
A.5. RapidMiner Overview .....	153
A.6. Installation .....	153
A.6.1. Java Installation .....	153
A.6.2. R Installation .....	153

A.6.3.	MySQL Installation .....	154
A.6.4.	RapidMiner Installation .....	155
A.6.5.	RapidAnalytics Installation.....	156
A.6.5.1.	RapidAnalytics Method #1 Installation .....	156
A.6.5.2.	RapidAnalytics Method #2 Installation .....	157
A.6.6.	RapidAnalytics Repository Installation .....	157
A.6.7.	RapidAnalytics Startup .....	159
A.6.8.	Website Setup .....	160
A.6.9.	Troubleshooting.....	162
A.6.9.1.	Access Denied .....	162
A.6.9.2.	Extension Failure .....	163
A.6.9.3.	Shared Library Failure .....	164
A.6.9.4.	Stale Images .....	164
B.	Extending the Framework .....	165
B.1.	Overview .....	165
B.2.	New Datasets .....	165
B.2.1.	Importing into MYSQL .....	165
B.2.2.	Configuring the Framework.....	165
B.3.	New Visualizations .....	167
C.	Reference Code.....	173
C.1.	HTML Code .....	173
C.1.1.	bar.html .....	173
C.1.2.	boxplot.html .....	177
C.1.3.	geospatial.html.....	181

C.1.4. help.html .....	184
C.1.5. intro.html .....	186
C.1.6. line.html .....	187
C.1.7. mosaic.html .....	191
C.1.8. odds.html .....	195
C.1.9. oddstime.html .....	199
C.1.10. pie.html .....	204
C.1.11. start.html .....	207
C.1.12. table.html .....	208
C.2. Javascript Code .....	211
C.2.1. common.js .....	211
C.2.2. specific.js .....	218
REFERENCES .....	228

## LIST OF FIGURES

Figure	Page
2.1. RapidMiner .....	19
2.2. Spago BI .....	20
3.1. System-Level Diagram .....	24
3.2. RapidMiner Process Design View .....	25
3.3. RapidMiner Results View .....	26
3.4. RapidMiner R-Perspective.....	27
3.5. RapidAnalytics Repository .....	28
3.6. RapidAnalytics Service.....	29
3.7. RapidAnalytics Chart Definition .....	30
3.8. Table View .....	34
3.9. Table SQL 'Read Database' Script.....	35
3.10. RapidMiner Table Process.....	35
3.11. Aggregate Pie Plot .....	37
3.12. RapidMiner Aggregate Pie Process .....	38
3.13. Aggregate Pie Plot SQL 'Read Database' Script .....	39
3.14. RapidAnalytics Pie Flash Chart Options .....	40
3.15. Time Aggregation(BAR) View .....	42
3.16. RapidMiner Bar Chart Process .....	43
3.17. Time-Aggregate(BAR) SQL 'Read Database' Script.....	43
3.18. Time Line Plot .....	45
3.19. RapidMiner Time Line Process .....	46

3.20. Time-Line SQL ‘Read Database’ Script .....	46
3.21. Time Line Plot Process Design R-Script .....	47
3.22. Mosaic Plot .....	49
3.23. Mosaic Plot SQL ‘Read Database’ Script .....	50
3.24. RapidMiner Mosaic Plot Process .....	51
3.25. Mosaic Plot Process Design R-Script .....	51
3.26. Box Plot.....	53
3.27. Box Plot SQL ‘Read Database’ Script .....	53
3.28. RapidMiner Box Plot Process .....	54
3.29. Box Plot Process Design R-Script .....	54
3.30. Odds Ratio .....	55
3.31. Odds Ratio SQL ‘Read Database’ Script.....	56
3.32. RapidMiner Odds Ratio Process.....	57
3.33. ODDS Ratio Process Design R-Script.....	57
3.34. Time-Based Odds Ratio.....	58
3.35. RapidMiner Time-Based Odds Ratio Process .....	59
3.36. Time-Based ODDS Ratio Process Design R-Script .....	60
3.37. Time-Based Odds Ratio SQL ‘Read Database’ Script .....	61
3.38. Geospatial View .....	62
3.39. Google Fusion Tables .....	63
3.40. Heatmap Options .....	64
3.41. Portaling Feature .....	68
3.42. Attribute Filtering .....	69
3.43. Time-Based Filtering.....	70
4.1. Breach Data Sources & Traceability .....	74
4.2. Breach Type Definitions from PRC .....	77
4.3. Breach Legislation Map .....	82
5.1. Breach vs Clean (By Sector) .....	84



5.2. Breach vs Clean (Financial Sector By Industry) .....	86
5.3. Breach vs Clean (By Fine Cap Size) .....	87
5.4. Odds Ratio By Sector .....	88
5.5. Odds Ratio By Sector .....	90
5.6. Odds Ratio By Industry (Financial) .....	91
5.7. Breach Type Vs CapSize (Coarse) .....	94
5.8. CapSize(Coarse) vs Sector .....	95
5.9. CapSize(Fine) vs Sector .....	96
5.10. Breach Type vs Region .....	97
5.11. Entity Category vs Region .....	98
5.12. Entity Type vs Region .....	99
5.13. City Size vs Breach Type .....	100
5.14. Median Log(Breach) By Entity Type .....	101
5.15. Median Log(Population) By Sector .....	101
5.16. Breached Entities By Year .....	102
5.17. Breach Type By Year .....	103
5.18. Breached Sectors By Year .....	104
7.1. Generator Type - Compromise vs. Control .....	114
7.2. Server Type - Compromise vs. Control .....	115
7.3. Generator Type - Odds Ratio (Day 1) .....	116
7.4. Server Type - Odds Ratio (Day 1) .....	117
7.5. WordPress Version - Odds Ratio (Day 1) .....	118
7.6. Server Type vs. Generator Type(Day 1) .....	119
7.7. TLD vs. Generator Type(Day 1) .....	120
7.8. Country vs. TLD(Day 1) .....	121
7.9. Generator Type Compromises By Month .....	122

7.10. Server Type Compromises By Month .....	123
7.11. Server Type Odds Ratios By Month .....	124
7.12. WordPress Version By Month Aggregate .....	124
7.13. Odds Ratios By Country .....	125
7.14. Geospatial Country Odds Ratios .....	126
8.1. Attacks By Target Machine .....	134
8.2. Attacks By Targeted Service .....	135
8.3. Attacks By Protocol .....	135
8.4. Source of Attacks (Pie Chart) .....	137
8.5. Source of Chinese Attacks (Heatmap).....	138
8.6. Source of U.S. Attacks (By State) .....	138
8.7. Source of U.S. Attacks (By Zip Code) .....	139
8.8. Target Machine By Month .....	140
8.9. Service By Month .....	141
8.10. Country By Month.....	141
8.11. Country vs. Service .....	142
8.12. Country vs. Machine.....	143
A.1. MySQL Privilege Update Script .....	155
A.2. RA Installer Step#1 .....	158
A.3. RA Installer Step#2 .....	159
A.4. RA Installer Step#3 .....	160
A.5. RA Installer Step#4 .....	161
A.6. RA System Settings .....	162
B.1. SQL Table Definition Example .....	166
B.2. SQL Table Import Example .....	166
B.3. Url parameter.....	167
B.4. Entry Function to configure new dataset.....	168

B.5. HTML page references .....	168
B.6. Data-Specific Variables .....	169
B.7. Filter Specific Widget Definition.....	169
B.8. Source-Combo Widget Definition .....	170
B.9. Chart-Specific Widget Definition .....	170
B.10. RA Repos Browser .....	171
B.11. RA Process Scheduler .....	171
B.12. RA Exporting To Service.....	172
B.13. RA Table Service .....	172

## LIST OF TABLES

Table	Page
2.1. Comparison Criteria .....	14
2.2. Commercial Tools .....	15
2.3. Open Source Tools .....	17
2.4. SpagoBI vs. RapidMiner .....	18
3.1. Reporting Extension Plot Types .....	28
3.2. Google Fusion Table Color-Codes .....	62
3.3. Web service Parameters .....	67
4.1. Privacy Rights Breach Record Attributes .....	76
4.2. Privacy Rights Breach Record Additions .....	78
4.3. NASDAQ / NYSE Attributes .....	79
4.4. Census Attributes .....	80
5.1. Breach vs Clean (By Sector) .....	85
5.2. Breach vs Clean (By Financial Industry) .....	85
5.3. Breach vs Clean (By Fine Cap Size) .....	86
5.4. Odds Ratio By Sector .....	88
5.5. Odds Ratio By CapSize .....	89
5.6. Odds Ratio By Industry (Financial) .....	89
5.7. CHISQ Test Results .....	93
5.8. Sample Yahoo API query tags .....	106
6.1. CMS Dataset Attributes .....	109
8.1. Honeypot Dataset Attributes .....	132

*To my four loving children Leah, Lexie, Lauren, and Logan.*

## Chapter 1

### INTRODUCTION

In this first chapter, we present the background to a significant problem affecting security dataset analysis, and lay the groundwork for a solution to the problem. We begin with the problem statement, which discusses the current state of cybersecurity analysis and establishes the need for a data-driven reusable framework. We then examine prior research into visualization framework models. After opening with a synopsis of the origins of modern visualization, we go on to discuss past efforts that made attempts at solving this or a similar problem. Finally, we conclude in Section 1.3 by detailing the contribution of the praxis and giving an overview of forthcoming content.

#### **1.1. Problem Statement**

Cybersecurity is becoming increasingly necessary these days. Every day, it seems one hears of a new incident. The prevalence of common operating systems, fueled by the mass-connectivity provided by the Internet, has afforded hackers, thieves, and other criminals the ability to easily gain access to resources. From passwords being hacked to web-servers being attacked to valuable data being stolen, these incidents have become high profile and frequent.

Determining what factors are indications of risk is imperative, as resources are limited. Once an assessment is made as to what these factors are strategies can then be formulated to deal with an issue. While one is capable of throwing all their resources towards a problem, a more thought-out approach should be employed. In pursuit of

the correct choice, costs and benefits should be examined in order to determine what action or inaction yields the most net benefit. These types of questions have become the foundation of the study of security economics. [3]

So where does one begin to determine what factors affect a security problem? They can rely on expert opinion, news reports, or speculation, which may suffice in situations of micro-scale. But how about the information technology chief of a multinational corporation? What strategy should he/she employ in making security decisions that could potentially cost the company millions of dollars? It is important that these kinds of decisions be well-founded data-driven decisions. Twenty years ago, this kind of data may have been hard to come by, but today we live in a data rich world. An enormous amount of data is generated in almost all fields (including security), which can fortunately aid in decision making.

But what challenges still remain after the data is identified or collected? While the human mind is inherently capable of processing data and drawing suitable conclusions, there remains a limit to its ability to process data without the aid of visualization. To counteract this, solutions should be crafted that aggregate and visualize data from different perspectives, thus enabling decision makers to draw conclusions and subsequently act. However, each dataset is unique; each problem is new. Does a new solution need to be employed every time a new dataset is utilized? Custom solutions are inherently time-consuming. Efficiency suggests reusable software components/tools be created to help facilitate analysis.

Our problem statement is thus summarized by the following points:

- Accurate security decision making is becoming increasingly important
- As research becomes more data-driven, there is an increased need for tools to visualize security datasets.
- Custom visualizations can be costly in both time and resources.

- One-time static nature limits their utility as empirical analysis is often an iterative process.

A reusable framework for security dataset analysis would enable researchers to quickly interpret a dataset by drilling into the data iteratively. If a particular view of the data isn't relevant, they could quickly abandon efforts to focus on the next variable or view, thus allowing them to make meaningful conclusions with short turnarounds. Such a framework would provide efficiencies to both corporate and educational security research. It would also serve as a starting point for improved analysis methods and subsequent collaboration in research endeavors.

## **1.2. Prior Work**

### 1.2.1. Origins of Modern-Visualization

The modern study of information visualization began with computer graphics, which from its inception has been used to study scientific problems. However, during its early years, a lack of computing resources often limited its importance. The recent resurgence of visualization began in the late 1980's with an article in Scientific Computing [1]. This spurred a new renaissance fueled by improved computing resources and the collaborative nature of the internet.

This has in turn, over the past couple of decades given rise to a variety of fields within visualization (Educational, Scientific, Informational, Product, Communication, Visual Analytic, etc.) Two of these areas form the facets for modern day cybersecurity research, namely Scientific Visualization and Visual Analytics.

Scientific Visualization focuses on the exploration, analysis, and understanding of data for scientific purposes. It provides a portal for the mind's eye into complex higher order data using graphical visualization techniques. In essence, it reduces the



complexity of the data into something that is easier to understand while still providing meaningful information. As such Scientific Visualization has become a critical aspect of data visualization.

In contrast, Visual Analytics, focuses on the means of human interaction with the visualization system itself as part of both an iterative and larger process of analysis. This interactivity enables users to determine the next step in an iterative process. This process flow enables researchers to iterate on a dataset visually to reach conclusions with each potentially building on the last.

## 1.2.2. Visualization Framework Research

### 1.2.2.1. *Operator-Based Models*

In 1998, Ed Huai-hsin Chi and John T. Riedl proposed a framework [13] that included both operators and interactions for visualization systems. This was likely one of the first attempts to solidify the tenets of visualization framework that included a focus on user-driven interactivity for analysis purposes. Interactive operators were provided to users to transform data within the data flow. (e.g. import, filter, visualize, etc.) Furthermore, the research focused on an end user's need for viewing intermediate results in determining subsequent analysis steps. In this sense, it empowered users to understand how the system work, and then manipulate the the data flow appropriately to perform the next logical analysis action.

### 1.2.2.2. *Reusable Frameworks*

An early 2003 research framework called EVolve [47] was an early attempt at a reusable visualization framework. EVolve is extensible in that it could be applied to new data sources or visualization types. EVolve came from a need to profile Java applications in various ways. This meant visualizing program behaviors (e.g.

predictability) as well as extending it to custom visualizations that mapped certain program behaviors. In essence, the authors needed a system where new visualization types could be easily created as well as providing new input types to be defined.

A subsequent 2005 research project called Prefuse [25] was a novel attempt to provide building blocks for research analysis and visualization. By providing table, graph, and tree data structures with arbitrary data attributes, Prefuse aimed to provide an abstract interface to data elements. When used in conjunction with its reusable components library for layouts, controls, and operations, it provided an effective means for visualizing data using Java. Prefuse focused on component-based reuse and extensibility, using provided modules (e.g., filters, layouts, renderers, and interactive controls) across visualizations, while providing the developer the opportunity to extend functionality to create customized components. Additionally, Prefuse furnished researchers with novel visualization techniques like Force Directed Graph Drawing among other things. Accumulating tens of thousands of downloads, it enjoyed minor success as an open source Java visualization library, and can still be downloaded today under the sourceforge project.

### *1.2.2.3. Distributed Frameworks*

In 2002, Bender, Klein, Disch, and Ebert proposed the idea of a web-based collaborative distributed framework [8]. This would employ interactive visualizations running on a web-server and visualized within a client browser. Prior to this time, visualization research had focused on employing dedicated application-based software tools running natively on local machines while using resident data-sources (files / databases). Interactive visualizations can now be extended to distributed environments by employing client/server paradigms enabled by client-browser technology. This works as the visualization itself is self-contained as a web-based service or process component. If employed, distributed frameworks can have a sweeping affect of

widening the target audience who uses the associated visualization.

In 2006, Holmberg, Wnsche, and Tempero suggested a web-based framework [27] for visualization. Their work represents an effective analysis and decomposition of requirements for such a framework some of which parallels our own requirements. Additionally, their survey of solutions constitutes a compelling analysis of web-based technologies at the time. However, their research stops short of development of the proposed framework.

### 1.2.3. Application-Based Frameworks

While a plethora of generic frameworks and business intelligence tools exist, limited work has been done to customize visualization frameworks to particular datasets.

#### 1.2.3.1. *Financial Visualization*

In “Analyzing Financial Data Through Visualization,” [50] Yadav creates an applet to visualize Financial Data. This servlet provides the user the ability to draw from financial sources (web sites) and provide visualization of metrics (Net Income, Total Assets, Operating Margin, etc). He then measures the effectiveness of his effort through a control and non-control group, and makes conclusions through testing / surveys of the effectiveness of his tool. Yadav stops short of providing a framework for financial research, however, he does postulate that any interface that might address the needs of the community “would update in real-time so that the users can filter information according to their liking.” He also suggests that the tool “would lay great emphasis on interaction in order for the user to have full control over the information being displayed.” Yadav concludes with the hope that his tool will pave the way for future research in this area.

### *1.2.3.2. Security Visualization*

In 2004, Komlodi, Goodall, and Lutters proposed a framework for Intrusion Visualization [32]. Within their framework, monitoring and analysis activities were developed with regards to user-needs and real life problems. By studying daily activities and understanding of routine work practices, a representative framework was developed that focused on software quality attributes of customizability, reusability, and understandability. Filtering and interaction as well as the ability to choose data sources were central themes within their framework. Their displays sought to refocus Intrusion Detection Visualization tools to focus on analysis and exploration rather than only the monitoring component. Additionally, IDSRadar [51], IMAP [21] and FloVis [42] tools expand on this research.

Further work has been done into visualizing security events (auditing) as well as the visualization of high performance clusters [19] [49]. While network and data security are imperative within the security research community, no tangible work has been done to visualize security economics data in a reusable way.

Additionally, a peer-research group VizSec (Visualization for Cyber Security) was established in 2004 to bring together researchers and practitioners from academia, government, and industry to address the needs of the Cyber Security community [45]. By bringing light to new visualization and analysis techniques within the community, VizSec continues to lead in disseminating research within the community.

## **1.3. Contribution and Structure**

This praxis is organized around three case studies of cyber-security analysis. The first case study, derived from the PrivacyRights website [16], examines a dataset detailing breach incidents reported through government channels or gathered through news outlets since 2005. It highlights relative frequencies and types of security

breaches that occur. The second case study examines a CMS (Content Management System) web-server compromise dataset derived from ongoing cyber-security research channels at SMU (Southern Methodist University). The dataset provides web server attributes (e.g. software, version, country, etc.) and survivability metrics related to web-server attacks. Finally, a third dataset provided by the Data Driven Security Blog [10] details honeypot attacks over a period of 6 months. The collected attributes include variables such as the attacking country, port number, and time. Thus, each case study validates the utility and relative importance of the framework for exploring cyber-security datasets. By applying the same framework across three distinctly different datasets, we hope to demonstrate the reusable nature of the contribution.

We propose this reusable framework within this praxis with the following intent:

- Providing meaningful visualizations to analyze security economics data for both academic research and corporate decision-making.
- Developing reusable visualizations that are extensible to most other datasets.
- Allowing the user to have full control over the data being displayed by providing filtering mechanisms and comparison mechanisms.
- Equipping researchers with a mechanism to collaborate on and share datasets.
- Providing shortened analysis time as time is only an act of importing into the framework.
- Providing deeper relative analysis as more interactions can be looked at without significant costs.
- Supplying an instrument to provide interactive publishing and reporting of their research.

Chapter 2 begins by detailing the software engineering process involved with the creation of the framework. We start by formulating both functional and nonfunctional

requirements from both the prior research and the problem statement. We then carry out a trade study of existing capability, by generating comparison criteria and mapping both open-source and commercial tools to these criteria. In Chapter 3, we define a system-level design for the framework. Following a discussion of each component, we then define the visualizations and discuss their design.

In Chapter 4, we discuss the history of the breach dataset including prior work. Additionally, we discuss aggregation of the data and identify potential risk factors. In Chapter 5, we analyze the breach dataset using the framework. We first outline the process of importing a dataset into the framework and configuring the framework for the dataset. We then demonstrate the framework by taking advantage of the dynamic interactivity and case-control comparison features provided within the framework to perform a top-down analysis of the data. We outline the iterative strategy behind using the framework and highlight the associated features therein.

We then demonstrate the reusable nature of the framework by examining a second dataset. In Chapter 6, we discuss the history of the CMS dataset and how the data was compiled. We also discuss prior work and outline goals for this dataset as it relates to ongoing cybersecurity research. In Chapter 7, we again turn towards performing an analysis using the framework. A similar yet also contrasting analysis from the prior is demonstrated utilizing the framework. In doing so, we justify some of the same conclusions provided by traditional analysis methods, but also highlight new dynamic perspectives into the data utilizing the framework.

In Chapter 8, we examine application of the framework to a third dataset, the honeypot dataset. By presenting an analysis of the data using the framework, we further demonstrate its reusable nature. Finally in Chapter 9, we assess the contribution of this framework and the benefits over traditional analysis methods. We also discuss opportunities for future work with this framework and conclude.

## Chapter 2

### REQUIREMENTS & TRADE STUDY

In this chapter we outline requirements for the framework and survey representative solutions. We begin in Section 2.1 by formulating both functional and non-functional requirements for the architecture. Additionally, stretch goals are defined. We then discuss our approach to the trade study in Section 2.2, and generate both needs and wants that we use as comparison criteria to differentiate between design alternatives (Section 2.3). By surveying popular COTS and open source tools against the comparison criteria, we establish the best solution for our framework. Finally, the rationale behind our decision is presented in Section 2.7.

#### **2.1. Requirements**

By analysis of the problem statement as well as decomposing deficiencies of prior works, a list of capabilities for the visualization tool was created. These capabilities were then subsequently decomposed into requirements. A formal list of functional and non-functional requirements follows:

##### 2.1.1. Functional Requirements

- The system shall be able to allow for analysis of a security dataset using a sortable tabular view depicting the data-itself.
- The system shall be able to allow for analysis of a security dataset using a pie-chart view showing relative proportions for a chosen control variable.

- The system shall allow for analysis of a security dataset using an odds ratio plot.
- The system shall be able to allow for analysis of a security dataset using a geospatial view.
- The system shall allow for analysis of a security dataset using a mosaic Chi-Squared plot.
- The system shall allow for analysis of a security dataset using a time-based plot.
- The system shall be able to allow for analysis of a security dataset using a line view depicting the data-itself.
- The system shall provide for dynamic analysis of a security dataset by providing controls to filter to a subset of the source data.
- The system shall allow for side-by-side comparison of two different datasets of like-formatted data.
- The system shall allow clients to connect and analyze security datasets via the web.
- The system shall allow clients to publish contents of their analysis.
- The system shall allow the user to download the current view context's data.

#### 2.1.2. Non-Functional Requirements

- The system shall be able to operate on large datasets, that being larger than 100K records within a reasonable amount of time per operation (10 seconds).
- The system shall provide for a reusable framework reconfigurable to another dataset with minimal changes.



- The system shall provide for authentication to prevent unauthorized access.
- The system shall be reconfigurable and maintainable remotely via the web.
- The system shall be low to no-cost and provided with permissible licensing.

### 2.1.3. Wants

Additionally, some stretch goals were delineated. These were described as wants, not necessarily requirements, but would be useful if provided by the framework.

- Integration of R within the framework
- Support for non-standard visualization options (Geospatial plot, Bubble-Plot, etc.)
- Cross-platform support for server-side and client-side software within the framework
- Provide the ability to create dashboards, namely preformatted HTML where one can integrate graphs and controls into a unified presentation for the user
- Tool support whether it be online or preferably personalized support.
- Ease of use by providing graphical methods to defining visualizations

## 2.2. Background

Through the last decade the field of visualization has grown exponentially. Visual Analytics tools have now become a recent focus as visualization systems have become of increasing importance to firms. Today, many commercial Business Intelligence (BI) Tools compete in a global market space. Many of these for-profit tools are extraordinarily capable, enabling businesses to make decisions on their data. The usage of

said tools often requires exorbitantly priced licenses and large upfront investments. Indeed, as data models become more and more complicated, advanced knowledge/-expertise has become required leading to widespread usage of support contracts. In contrast, in the open source realm, a smattering of visualization tools and frameworks exist. These tools are also highly capable in some instances and provide meaningful capabilities. Too often, however, these tools can become unstable and limited in nature. Additionally, some tools are offered on a bait-catch model by initially offering limited capability with the intention of vesting you in their product. They then attempt to offer additional feature-sets, reduce limitations, or sell support contracts. Despite these limitations in both the commercial and open-source realm, both subsets of BI tools provide meaningful visualization capability and possibly could serve as components within our reusable framework.

In contrast to a reuse solution, a completely custom solution to the problem statement was determined to be too time consuming to achieve. Any solution would inevitably end up being less polished and less capable. Instead, as part of this trade study, only COTS and Open-source tools were examined for incorporation into the framework.

### **2.3. Comparison Criteria**

Initiating the study, a list of comparison criteria (design primitives) were created from what was thought to be important to the end-solution. The comparison criteria were then ranked as Needs and Wants. A description of the criteria follows below in Table 2.1 and were based off both the functional and non-functional requirements provided in the previous sections. Most important to the comparison criteria is the cost and web-integration aspect. Additionally, the ability to operate on large datasets and have basic visualization options was deemed important. Some stretch goals

Table 2.1. Comparison Criteria

	Comparison Criteria	Description
<b>Need</b>	Easily integrates with web	Can the design alternative support a a web-solution.,For example, does it have web-publishing ability?
	Operate on large data sets	It needs to be able to process large file-based data sets and/or large database sets.
	Low to no cost.	Neither SMU or I have dedicated funds for a COTS solution that does not provide for a free or low-cost academic license. At bare minimum, costs need to be fixed and there can be no user-based costs.
	All common visualization options(line, scatter, pivot)	This solution needs to support all common graphs, chart types: namely line, table, scatter, pivot.
	Permissible license	License needs to be permissive for academic use & provide for distribution of derivative product.
<b>Want</b>	Cross-Platform	The tool is cross-platform compatible.
	Dashboard creation	If it has integrated, gui creation features for the web-page, that is a plus as it would save time in the long-haul and provide for ease of extensibility
	Non-standard visualization options (e.g. geo-plot)	Other not-so-critical visualization options including: Geospatial plot, Bubble-Plot, etc.
	Integration with R	R the statistical and data-mining program language.
	Query/Filter/Transformation Ease of Use	From an ease-of-use standpoint, it would be nice if queries could be created with a graphical tool.,Filtering could easily be achieved and the data can be transformed using some user-friendly methodology.
	Support for tool	Support for the tool., preferably cheap.,This could be through professionals or community-support.
	Authentication	Authentication methods to prevent unauthorized access.

include: cross-platform capability, dashboard creation, R-integration, and advanced visualization options (like geospatial or bubble plots). Additionally, authentication methods, tool support and ease of use would be desired.

## 2.4. Commercial Tools

To begin, COTS solutions were surveyed and a list of candidate options was created. Due to time constraints of not being able to review every tool, tools were down-selected for comparison by market-share and relative popularity in associated BI Forums [30]. The table below (2.2) depicts the tools (design alternatives) that were compared. Most of these commercial tools had features provided with their free licensable versions or purposefully removed key capabilities such as operating on large datasets.

Table 2.2. Commercial Tools

Comparison Criteria	Design Alternatives					
	Oracle (BI)	Tableau	Webalo	Pentaho	Jaspersoft	Ethority
Easily integrates with web	Yes	No(Free) Yes(Paid)	No(Free) Yes(Paid)	No(Free) Yes(Paid)	No(Free) Yes(Paid)	Yes
Operate on large data sets	Yes	No(Free) Yes(Paid)	No(Free) Yes(Paid)	No(Free) Yes(Paid)	No(Free) Yes(Paid)	Yes
Low to no cost.	No(\$10K)	Yes(Free) No(\$1K)	Yes(Free) No(\$100/yr/u)	Yes(Free) No	Yes(Free) No(\$20K)	Not Provided
All common visualization options(line,scatter, pivot)	Yes	Yes	Yes	Yes	Yes	Yes
Permissible license	Yes	No(Free) Yes(Paid)	No(Free) Yes(Paid)	No(Free) Yes(Paid)	No(Free) Yes(Paid)	Yes
Cross-Platform	No	Yes	Yes	Yes	Yes	No
Dashboard creation	Yes	Yes(Free) Yes(Paid)	No(Free) Yes(Paid)	No(Free) Yes(Paid)	No(Free) Yes(Paid)	Yes
Non-standard visualization options(e.g. geo-plot)	Yes	Yes	No	No	No	No
Integration with R	No	No	No	No	No	No
Query/Filter/Transformation Ease of Use	Yes	Yes	Yes	No	Yes	No
Authentication Methods	Yes	Yes	Yes	Yes	Yes	Yes
Support for Tool	Yes (Paid)	Yes (Paid)	Yes (Paid)	Yes (Forum/Paid)	Yes (Forum/Paid)	Yes (Paid)

The majority of tools within the candidate list provide an immense amount of capability. Oracle BI, for example, has a superior dashboard creation capability. Tableau also provided an advanced easy-to-use geospatial view. While many of the tools like Webalo, Pentaho, Tableau and Jaspersoft, provide free licensable version of the software, these versions were severely limited or impaired. For example, the ability for all users to operate and visualize freely or with a low cost was a central theme to the framework. Due to severe limitations with the observed candidates, commercial tools were unfortunately ruled out.

## **2.5. Open Source Tools**

Open-source tools were then selected as candidates for the framework using the same popularity methodology [30] as described in the previous section. Table 3.3 depicts the various options surveyed.

Although there are ample open-source tools out there, most had severe limitations to their capability and could be ruled out without further analysis. Others had severe licensing restrictions, possessed no web interoperability, or lacked the ability to operate on large data sets.

After performing a coarse-grain analysis of the remaining open-source tools, two viable options were down-selected: SpagoBI and RapidMiner. Both met the low-cost and licensing criteria, and had impressive web-integration capability and performance characteristics.

## **2.6. Down Select**

To down select to a final choice, a fine-grain analysis was then performed. A brief description of our evaluation of each product follows as well as a more detailed score-card (Table 2.4).

Table 2.3. Open Source Tools

Comparison Criteria	Design Alternatives							
	Google Charts API	Google Fusion	Spago BI	Rapid Miner/ Rapid Analytics	BIRT	TACTIC	KNIME	R
	Easily integrates with web	No	Yes	Yes	Yes	Yes	Yes	Yes
Operate on large data sets	Yes	No, Max=250MB	Yes	Yes	Yes	Yes	Yes	Yes
Low to no cost.	Free	Free	Free	Free	Free	Free	Free	Free
All common visualization options(line,scatter, pivot)	Yes	Yes	Yes	Yes	Yes	No	No	Yes
Permissible license	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Cross-Platform	Google	Google	Mozilla	AGPL3	Eclipse	Eclipse	GPL	GPL
Dashboard creation	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Authentication	No	No	Yes	Yes	No	Yes	No	No
Non-standard visualization options(e.g. geo-plot)	Yes	Yes	Yes	Yes	No	No	No	Yes
Integration with R	Yes	Yes	Yes	Yes	No	No	No	Yes
Query/Filter/Transformation	No	No	No	Yes	No	No	No	Yes
Ease of Use	No	Yes	Yes	Yes	No	No	No	No
Support for Tool	Limited (Forum)	Limited (Forum)	Limited (Forum) Yes (Paid)	Yes (Forum/ Paid)	No	No	No	Yes (Forum)

SpagoBI is the only BI tool of its caliber that is completely open-source with no attached conditions. Both, its performance on large datasets and interoperability with the web made it a viable option. The creators of SpagoBI have chosen to focus their application on distributed development and collaboration. All operations, design, administration, etc. is performed through using a web client to log in to

Table 2.4. SpagoBI vs. RapidMiner

Comparison Criteria	SpagoBI	RapidMiner/ RapidAnalytics
Easily integrates with web	Yes SpagoBI is a pure web-integration. Everything done in Spago is intrinsically web.	Yes. Web-integration is provided through RapidAnalytics product.
Operate on large data sets	Yes. Performance with queries of >100K records return quickly.	Yes. Performance with queries of >100K records return quickly.
Low to no cost.	Completely Free & Open-Source. No hidden gimmicks. No reduction in capability.	Free. Some reduction in capability as it relates to dashboard creation.
All common visualization options(line, scatter, pivot)	Yes	Yes
Permissible license	Yes Mozilla	Yes, AGPL3
Cross-Platform	Yes, The tool is pure web, and both client and server will run on multiple platforms. Because it is pure web, tool use does not require any install.	Yes Both client and server will run on multiple platforms. Tool use requires an install.
Dashboard Creation	Yes. Dashboard creation is intrinsic to the tool	Yes Dashboard creation is done at additional cost.
Non-standard visualization options	Yes	Yes More non-standard visualization options
Integration with R	No	Yes While not seamless, it still allows for a knowledgeable user to embed said script.
Query/Filter/Transformation Ease of Use	Yes. Operations provided through SQL syntax.	Yes, This is where RapidMiner shines. Extremely easy to use & understand.
Support for tool	Limited forum/community support. Separate support offered by developers for a fee/contract.	A plethora of forum/ community support. Separate support offered by developers for a fee/contract.

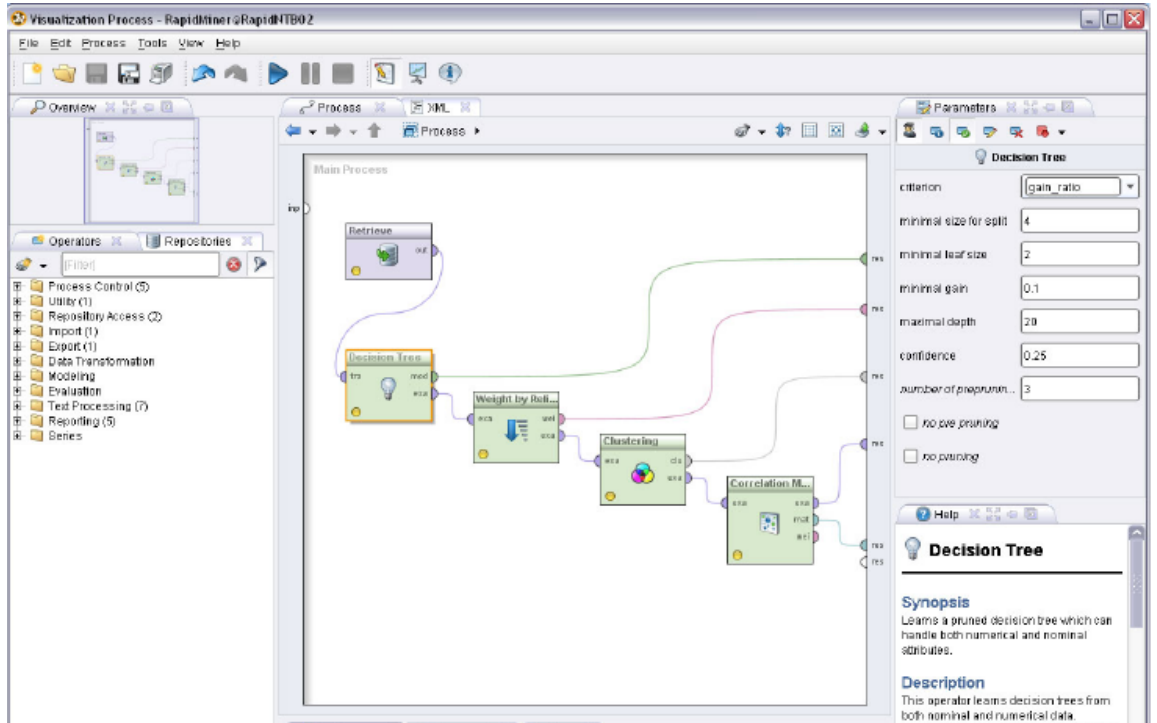


Figure 2.1. RapidMiner

the server through their web portal and operating on associated forms. The user can create custom dashboards containing a variety of different widgets or tables/graphics. The plot operations in SpagoBI is accomplished through drag, drop, link and SQL command definitions. All standard chart types are supported. In addition, some advanced plot types are provided. Once development is complete, the dashboard can be published to a web page.

RapidMiner (RM) is a tool that has captured a sizable market-share of the business-intelligence industry. The RM product specifically deals with local machine analysis of datasets. On the contrary, RapidAnalytics (RA), its sister product, handles the web-integration aspect. RM is flow-based in the sense that it supports operators which together manipulate the dataset to the desired results. These operators can be tweaked and extended in predetermined ways to provide extra versatility in



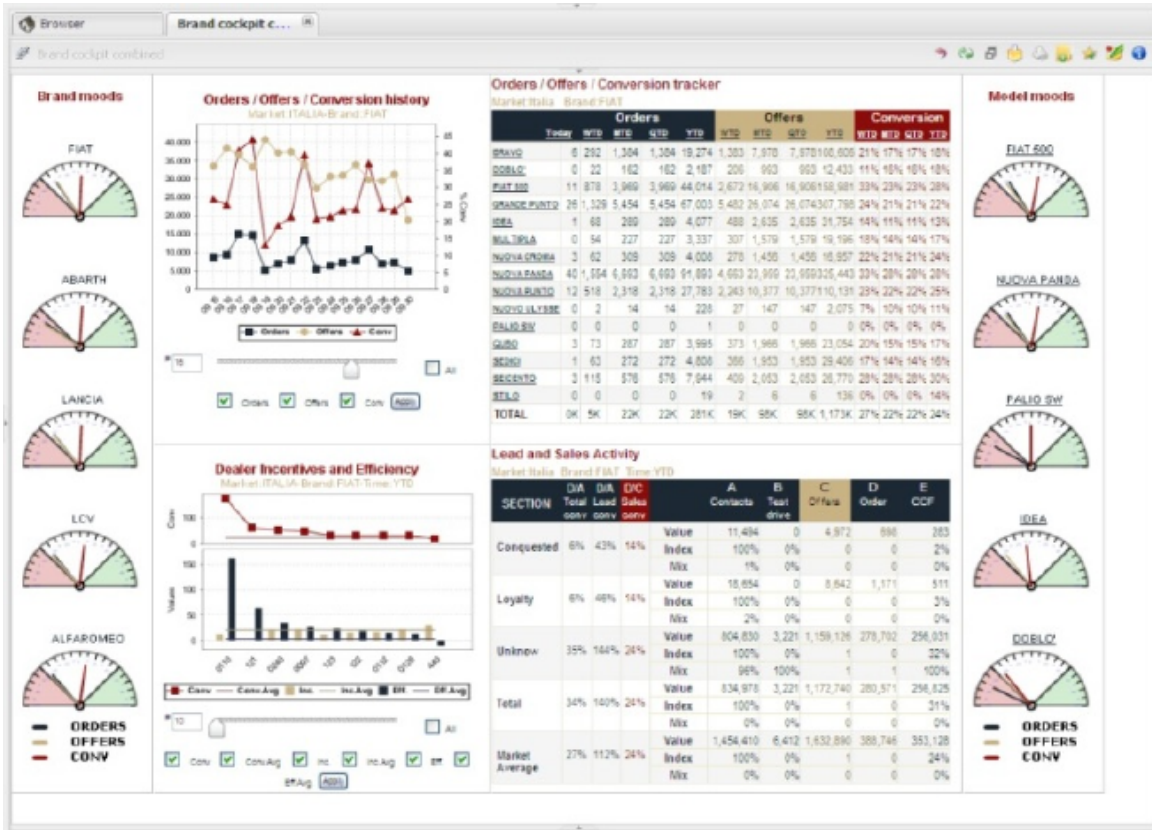


Figure 2.2. Spago BI

process design. In the pursuit of distributed analysis, users have the ability to upload files to the RA web-server, then configure and run these services as web-services. All standard chart types are supported within the product as well as some additional ones. It can also integrate with the R-programming language, something that no other tool could accomplish. This integrates through use of a custom R operator wherein both inputs and outputs for the 'Execute Script' component are mapped, and an associated R script is executed. Because of this feature, the RM/RA solution allows for integration of any chart type that R provides either natively or through an extension package.

## 2.7. Tool Selection

Inevitably, due to ease-of-use, R-integration, permissive licensing, market-share [30], and support within the community [38], the RapidMiner suite was selected. Indeed, the RapidMiner suite of tools (RapidMiner & RapidAnalytics) integrate well with the web, satisfying the distributed analysis requirement. The user-interface is intuitive and the flow-based operators make building processes simple.

While the full version of the RapidMiner product isn't available open-source or free this was determined to be something that we could effectively work around. Indeed, the free-version is capable in and of itself, and contains an incredible amount of features. Additionally, if full-product access ever became necessary, RapidMiner does provide an older version of the software, completely open source, following every new major build release. For example, when RM releases 5.0, they release the last minor version of 4.0, open source. The releasing of a full-featured tool with extensive capability on a open-source business model makes RapidMiner & RapidAnalytics a clear choice. And while the dashboard creation capability was lacking with the community version and ultimately inferior compared to SpagoBI's offering, this was viewed as a limitation that we could work around. Given additional effort and time spent in implementation, this limitation could evolve into an opportunity. By building the user-interface from the ground up and fine-tuning it to meet the needs of the research, we could allow the research needs to drive the associated user-interface design.

## Chapter 3

### FRAMEWORK DESIGN AND ARCHITECTURE

This chapter lays out the design and architecture of the framework. In doing so, we present both high-level and low-level specifics. We also seek to demonstrate the framework's visualizations and features, as well as the reusable strategies involved in its creation. It begins by presenting the high-level design of the system (Section 3.1). In doing so, each component is decomposed and examined in detail (Section- 3.2). We then transition into a discussion of each, wherein we examine both the purpose and associated design for each (Section 3.3) Next, in order to provide context into how the framework can be used for data-driven analysis, the framework's core user-interface features are described (Section 3.4). We then conclude by discussing the logistics for usage of the described framework (Sections 3.5 - 3.7)

#### 3.1. System-level Design

A high-level system diagram (see Figure 3.1) depicts the flow between components of the framework. We enumerate the components below and go on to describe them in further detail in the Component-Level Discussion section.

Components:

- **RapidMiner/R**: The design component to the system allowing the operator to define graphs and filters as well as plot R-enabled visualizations
- **RapidAnalytics Web Server**: The run-time component to the system which web enables RapidMiner / R processes as web-services

- **Apache Web Server:** The component of the system which serves HTML web pages on port 80
- **Repository:** The central location where the RapidMiner process models get stored and managed
- **Security Database:** The data store containing the security data to analyze
- **Web Client:** The browser running on the user’s computer used to access the remote visualization framework
- **HTML/JavaScript:** The client/server web-based code written to access the web process and provide the user-interface

The two components which are chiefly employed by the user to interact with the system

### 3.2. Component-level Discussion

A component-level discussion of each piece of the system follows. The system diagram above should be used for reference (Figure 3.1)).

#### 3.2.1. RapidMiner Discussion

RapidMiner (RM) is a cross-platform Java-based tool that resides on the design station. It can be run in a standalone fashion to work on local files or distributed fashion accessing files remotely. However, within our framework, RapidMiner (RM) serves as the design tool to create web-enabled processes. Because of its cross-platform nature, the application can be targeted for installation on a wide-variety of machines. Using the design perspective within the RM tool one is able to create a visualization by inserting flow-based components or “operators” then subsequently defining the routing between them. The output of our RM processes are images or the data nec-

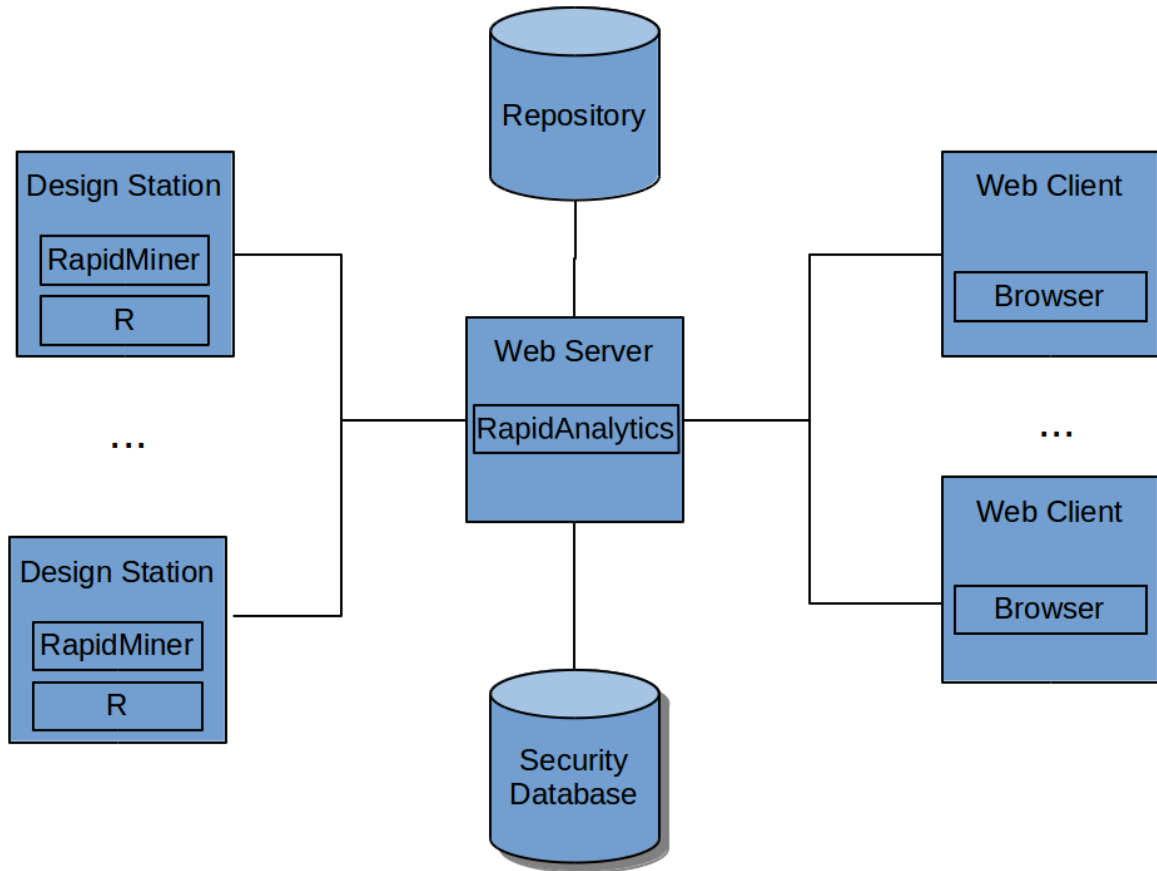


Figure 3.1. System-Level Diagram

essary to drive a web-enabled Flash Chart; however, RM supports building processes with a wide range of utility.

#### 3.2.1.1. Process Design

The internal design of RM is visualized within the process view of RM (see Figure 3.2). In order to create the process, a block schematic is constructed by the user. Data is read from input operators (whether they be file file-based, SQL-based, or network based), directed through a variety of operators, and finally arrive at the processes outputs. In order to provide enhanced flexibility, dynamic macros can be

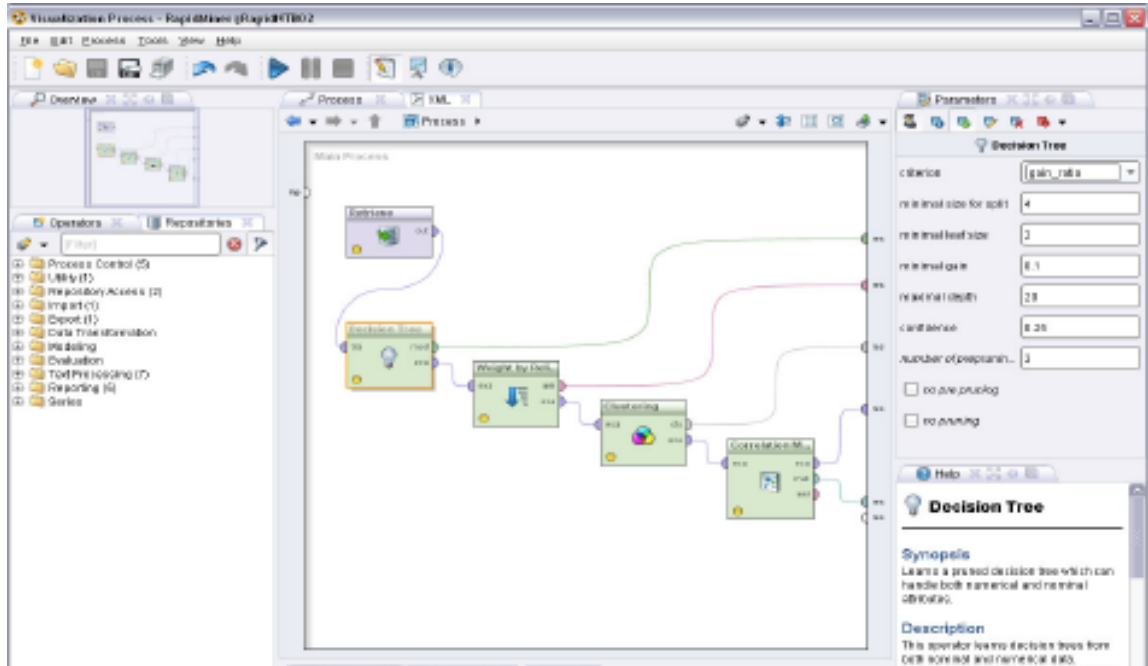


Figure 3.2. RapidMiner Process Design View

supplied, which serve as additional parameters to the process. They can best be described as environment variables globally made available throughout the process. These macros also can be aliased to URL parameter arguments when running in a web-server configuration such that arguments supplied to the webservice are accessible within the process.

### 3.2.1.2. Plot View

The outcome of the process is presented to the user within the results view of RM. In addition, graphing options are available which allow the user to visualize the tabular results graphically. If the user desires a visual representation of the results, controls are provided to choose the chart type as well as the attributes driving the graph (Figure 3.3). Common chart visualizations including scatter, box, and line are available within the plot view. In addition, some more advanced charts like Bubble,

SOM, Distribution, and Density are provided and enable RM to be a very powerful tool for analysis.

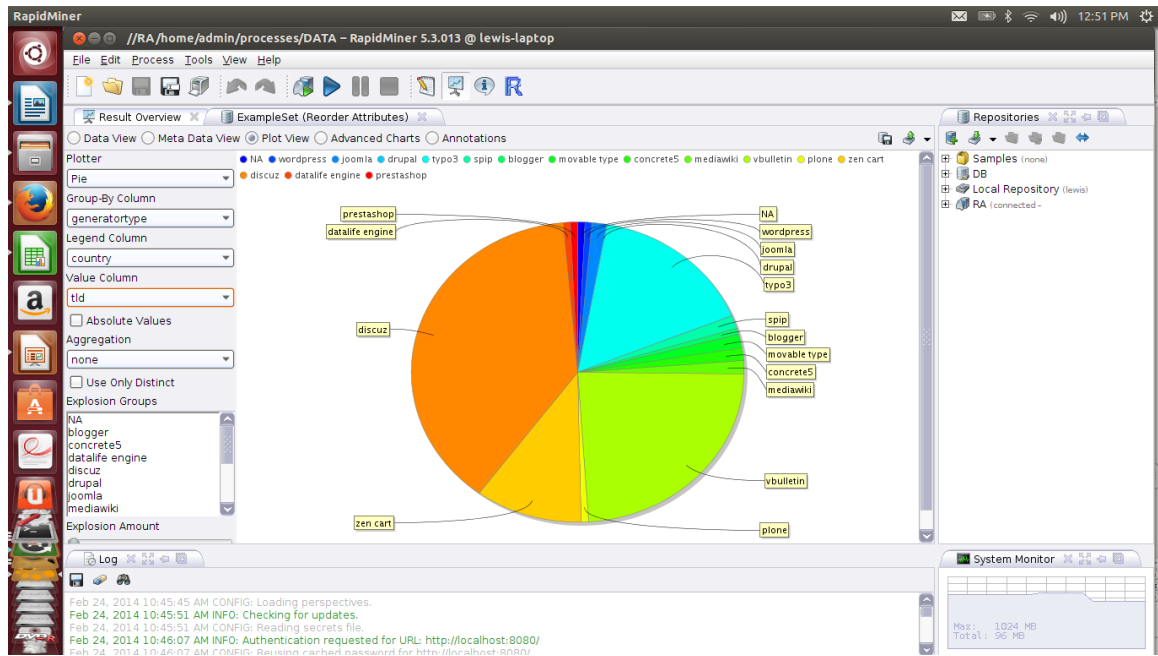


Figure 3.3. RapidMiner Results View

### 3.2.1.3. R-Extension

The R statistical programming language plugin (extension) for RapidMiner provides native R-support. An R-perspective is made available with the intent of fully immersing the user within the R-environment (Figure 3.4). One can use this perspective independently of the process to manipulate and view data just as you would with command-line R. However, most central to the extension is the ability to execute R scripts within RapidMiner processes. In order to accomplish this, the ‘Execute R-Script’ operator is provided. When the extension is loaded, the operator is made available for selection within the process view. The operator, in turn, allows the user to define an associated R-script for a given component block. Multiple inputs

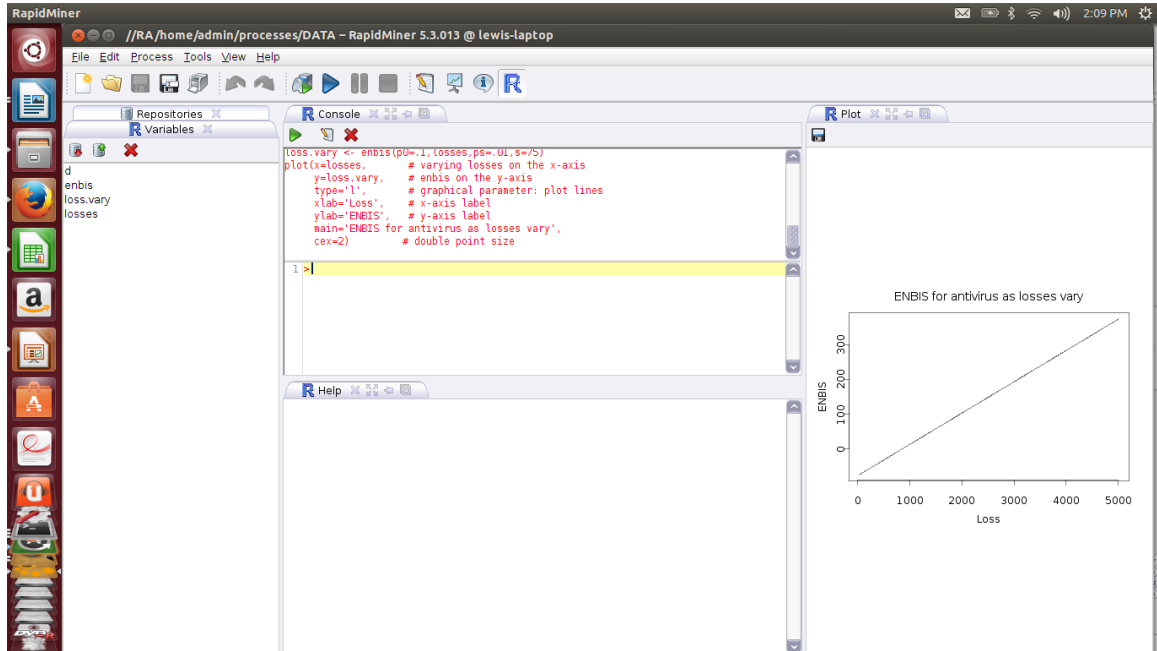


Figure 3.4. RapidMiner R-Perspective

and outputs are allowed per component block providing for maximum versatility in defining processes.

#### 3.2.1.4. Reporting-Extension

The RapidMiner Reporting plugin (extension) provides for automated generation of reports (e.g. HTML or PDF documents) from within RapidMiner processes. Often times this capability is collectively referred to as “web publishing”. These reports are capable of utilizing RapidMiner Plotters (visualizations) of varying types. Of particular use to this framework is the Report operator. Using this operator can produce a wide variety of plot types as defined in Table 3.1. As the core RapidAnalytics flash charts only support six distinct chart types, this drastically increases the quantity of charts available.



Table 3.1. Reporting Extension Plot Types

Scatter Plot (Various)	Series (Various)	Density	Pareto
Bubble	Survey	Pie	Andrews Curves
Quartile (Various)	Sticks (Various)	Box (Various)	Quartile (Various)
Parallel	SOM	Ring	Distribution
Deviation	Block	Bars (Various)	Histogram (Various)

### 3.2.2. RapidAnalytics Discussion

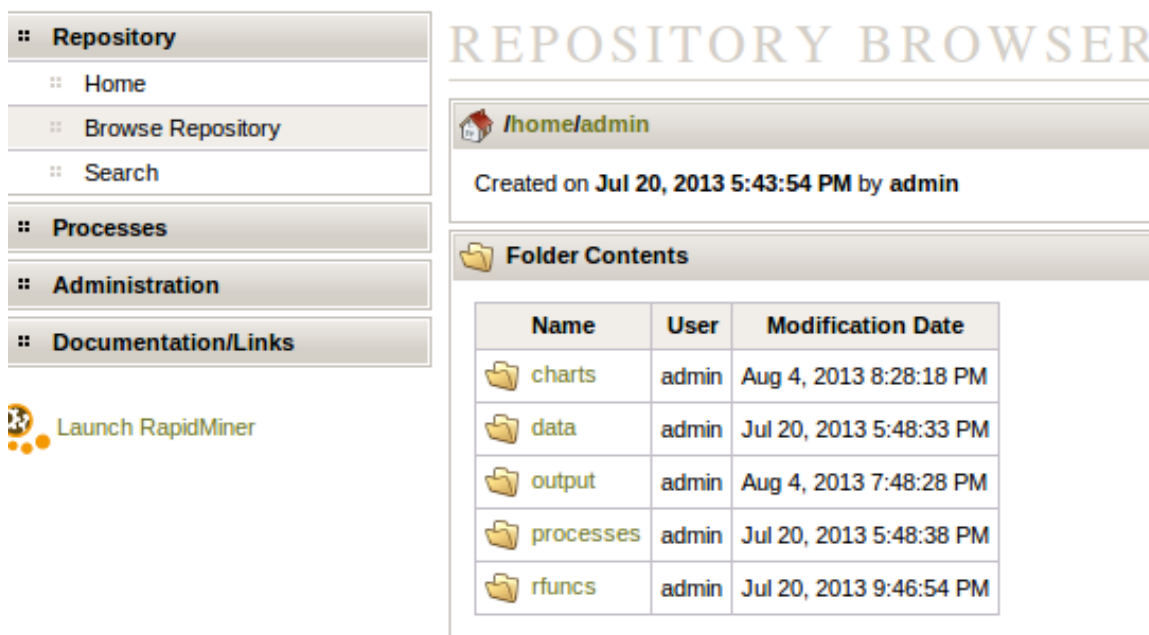


Figure 3.5. RapidAnalytics Repository

RapidAnalytics (RA) serves as the core component of within this framework. It can best be described as a web-server application whose primary goal is to run exported RapidMiner processes as web-services. To serve this means, a RA repository (discussed in Section 3.2.5) can be created where RapidMiner processes are developed, collaborated on, executed, and subsequently exported to services. These operations can be performed by accessing the RA web portal as seen in Figure 3.5 and referenced



## EDIT SERVICE: DATA\_EMILY

**Service Settings**

---

Service ID:

---

Data source:    
 Run process and remember meta data (This will simplify the editing of parameters.)

---

Output format:

---

MIME Type:

---

Cache input:

---

Parameter binding
















URL query parameter	Target (macro/operator parameter)	Mandatory
 <input type="text" value="tid"/>	 <input type="text" value="tid"/> 	<input type="checkbox"/>
 <input type="text" value="country"/>	 <input type="text" value="country"/> 	<input type="checkbox"/>
 <input type="text" value="tidop"/>	 <input type="text" value="tidop"/> 	<input type="checkbox"/>
 <input type="text" value="countryop"/>	 <input type="text" value="countryop"/> 	<input type="checkbox"/>
 <input type="text" value="servertype"/>	 <input type="text" value="servertype"/> 	<input type="checkbox"/>

Figure 3.6. RapidAnalytics Service

from <http://<RA-hostname>:8080/> in a browser URL.

The services depicted in Figure 3.6 and executed by RA can be configured to output as Tables, Flash Charts, XML, JSON, or binary file downloads. Additional attributes are provided in output contexts such as the Flash Charts in order to give more control over the output artifact (see Figure 3.7). This piece of the system is modular, in that it can live on the Apache HTML web server itself or can serve as an external node within the system.

---

Diagram type   Hide style parameters

**General display settings**

Title

**X-axis**

X-axis attribute

Step size

Label step size

**Y-axis**

Tick label

Explicit maximum

Explicit minimum

Explicit step size

Tick length

X-axis label

Tick height

Label rotation angle

Y-axis label

Maximum

Minimum

Step size

Label rotation angle

Show right axis

**Formatting settings**

Date format

Missing data message

Time format

**Diagram specific settings (bar)**

Tooltip

---

Figure 3.7. RapidAnalytics Chart Definition

### 3.2.3. Apache Web Server Discussion

The Apache web server is responsible for serving HTML requests on port 80. It retrieves the HTML and executes server-side JavaScript code discussed in section 3.2.6. The Apache Web Server is modular in that it does not necessarily have to coexist on the same machine as the RA web server discussed previously.

### 3.2.4. Security Database Discussion

A Database Management System (DBMS) is utilized to store and read the data in lieu of reading directly from file. Due to performance limitations with regular file I/O, this was considered the best option as it allows the implementation to have reduced retrieval costs for potentially millions of records contained therein. The

information also becomes portable as it is decentralized from the web-server itself. MySQL was used as the implementation since it comes freely available, is cross-platform, and already preinstalled on Linux. If performance became of increasing concern a nonrelational database might be employed.

### 3.2.5. Repository Discussion

The RM repository is a relational database that is responsible for storing all the RM processes and web services. The RA administrative web portal allows one to manage the content within the repository. There is no hard definition on the setup; the folder structure can be organized by whatever manner makes sense. It can also be permissioned such that different users or groups have the ability to read/write/execute files or folders respectively. All settings are managed within the same RapidAnalytics administration web-portal as discussed previously in Section 3.2.2.

### 3.2.6. Web Client




The front end for this framework is a web page written in HTML with JavaScript events. Common event handling code is relegated to common JavaScript files for improved reusability between datasets. Additionally, a CSS stylesheet is utilized to provide global reformatting ability across all pages. The imagery contained within the prototype framework are provided by either Icon Archive [4] with free licensing terms or by the dataset owner. A listing of files is provided below.







- **index.html**: The front-page of the framework
- **intro.html**: The page where one can select a unique pre-configured dataset for viewing
- **style.css**: Stylesheet defining user-interface feel for webpage
- **common.js**: Common javascript code

- **specific.js**: JavaScript code specific to the extended dataset
- **page.html**: Unique page for each visualization
- **base64.js**: Code for performing screenshots as provided by the html2canvas external API
- **html2canvas.js**: Code for performing screenshots as provided by the html2canvas external API
- **html2canvasproxy.js**: Code for performing screenshots as provided by the html2canvas external API
- **proxy.js**: Code for performing screenshots as provided by the html2canvas external API
- **icons/\***: Icons used within the webpage
- **csvs/\***: Downloadable csvs for data in-question

### 3.3. Visualization Design

Visualizations provided by the framework include:

-  **Tabular View**: A visualization depicting data results in a tabular view.
-  **Aggregate Pie Plot**: A visualization depicting relative percentages of a specified attribute within a dataset in pie chart format.
-  **Time Aggregate Bar Plot**: A visualization depicting relative percentages of a specified attribute over a time period. The results are depicted in stacked bar format.

- 
**Time Line Plot:** A time-based depiction of frequency of an outcome during a given time interval. The frequency is depicted as as a line plot of raw counts.
- 
**Mosaic Plot:** A visualization of two categorical variables depicting chi-squared statistical variance about the attributes.
- 
**Box Plot:** A visualization of the distribution of data depicting the first quartile, median, and third quartile ranges for a given attribute. The 95% confidence interval and outliers are additionally given.
- 
**Odds Ratio Plot:** A visualization depicting the odds of an outcome occurring given a particular exposure, in relation to the odds of the outcome occurring in absence of that exposure. The odds ratio is presented as a bar plot. Confidence intervals (whiskers) are provided to convey whether the results are statistically valid.
- 
**Time-Based Odds Ratio Plot:** A visualization depicting odds of an outcome occurring given a particular exposure, in relation to to the odds of the outcome occurring in the absence of that exposure. The odds ratio is presented as line plot over time. Confidence intervals (whiskers) are provided to convey whether the results are statistically valid.
- 
**Geospatial View:** A world-based map depicting data in a political boundary view. Colors denote variance of the data.

### 3.3.1. Tabular View

#### 3.3.1.1. Visualization

The intended purpose of the “Tabular View” is primarily to view the data at a low level. Viewing the raw data can often give a greater understanding and confidence in the data, and additionally serves as a means for troubleshooting visualizations.

The user-interface for this visualization is shown below in Figure 3.8 The table can be reconfigured to include different attributes (columns) within the dataset. Controls are provided to both filter on specific attribute values as well as filter out attributes that are not of interest as discussed in 3.4.2. The visualization is limited to the first 5,000 results, a constraint imposed for performance reasons. This preset is non-configurable.

year	entityCategory	entityType	stockXchg	symbol	breachType	state	region	citySize	name	market_cap	sector	
2005	EDU	PRE	NA	NA	CARD	Virginia	SA	MED	NA	0	NA	NA
2005	EDU	PUE	NA	NA	CARD	California	PAC	VLRG	NA	0	NA	NA
2005	EDU	PUE	NA	NA	HACK	Colorado	MTN	MED	NA	0	NA	NA
2005	BSO	PUB	NYSE	SAIC	DISC	California	PAC	VLRG	SCIENCE APPLICATIONS INTERNATIONAL CORPORATION	1896267276	TE	EC
2005	BSO	PUB	NYSE	ENL	HACK	Georgia	SA	MED	Reed Elsevier NV	14291980851.400	CS	Pu
2005	MED	NON	NA	NA	HACK	Illinois	ENC	VLRG	NA	0	NA	NA
2005	BSF	PUB	NYSE	BAC	HACK	North Carolina	SA	LRG	Bank of America Corporation	159138718429.560	FI	Ma
2005	BSF	PRI	NA	NA	DISC	Florida	SA	LRG	NA	0	NA	NA
2005	BSR	PUB	NYSE	DSW	INSD	Ohio	ENC	LRG	DSW Inc.	3803691696.700	CS	Clc Str
2005	BSO	PUB	NYSE	ENL	HACK	Ohio	ENC	LRG	Reed Elsevier NV	14291980851.400	CS	Pu
2005	EDU	PUE	NA	NA	HACK	California	PAC	LRG	NA	0	NA	NA
2005	EDU	PRE	NA	NA	DISC	Massachusetts	NE	LRG	NA	0	NA	NA
2005	MED	PRI	NA	NA	PORT	California	PAC	LRG	NA	0	NA	NA
2005	GOV	GOV	NA	NA	DISC	Nevada	MTN	LRG	NA	0	NA	NA
2005	EDU	PUE	NA	NA	DISC	California	PAC	MED	NA	0	NA	NA
2005	EDU	PRE	NA	NA	STAT	Illinois	ENC	MED	NA	0	NA	NA
2005	EDU	PUE	NA	NA	PORT	Nevada	MTN	LRG	NA	0	NA	NA
2005	EDU	PUE	NA	NA	INSD	Indiana	ENC	MED	NA	0	NA	NA
2005	BSO	PUB	NYSE	MCI	STAT	Colorado	MTN	LRG	Babson Capital Corporate	286093937.510	NA	NA

Figure 3.8. Table View

### 3.3.1.2. Process Design

The table process, depicted in Figure 3.10, is a basic 3-step process. Flow starts with the ‘Read Database’ operator which performs a SQL query with the appropriate user-specified filters as depicted in Figure 3.9. RapidMiner does provide an operator for filtering. However, for performance reasons, all filtering is performed within the ‘Read Database’ operator. Additionally, the returned result is limited to the first 5,000 records for similar reasons. From here, results flow to the ‘Select Attributes’ operator which acts as an additional filter to remove unwanted user-specified columns. The data is then written to text file via the ‘Write CSV’ operator. This makes the data available for download by the user if desired.

---

```
SELECT * FROM %{tablename}
WHERE %{filter1attr}%{filter1op}'%{filter1val}'
AND %{filter2attr}%{filter2op}'%{filter2val}'
AND %{filter3attr}%{filter3op}'%{filter3val}'
AND %{filter4attr}%{filter4op}'%{filter4val}'
AND (%{timeattribute} IS NULL OR
(%{timeattribute} >= '%{timebegin}'
AND %{timeattribute} <= '%{timeend}'))
limit 5000
```

---

Figure 3.9. Table SQL ‘Read Database’ Script



Figure 3.10. RapidMiner Table Process



### 3.3.2. Aggregate Pie Plot

#### *3.3.2.1. Visualization*

The intended purpose of the “Aggregate Pie Plot” is primarily to view relative percentages of a specific attribute within a dataset. Aggregation can occur on the raw data itself or after filters have been provided. Viewing these relative percentages as a piece of a larger whole can provide one insight into the makeup of a dataset. Additionally, side-by-side comparison portals discussed later in Section 3.4 are particularly useful when using this visualization type. By comparing side-by-side, one can quickly perform case-study comparisons between multiple datasets (typically a control and treatment dataset). While this visualization does not provide any statistical relevancy, it does help as a first pass discriminator in determining attributes for follow-on analysis with statistically relevant plots. (e.g. odds ratio)

The user-interface for the visualization is depicted below in Figure 3.11. Controls are used to specify an attribute for aggregation. In addition, controls are provided to both filter on specific attribute values as well as filter out attributes that are not of interest as discussed in 3.4.2.

#### *3.3.2.2. Process Design*

The RapidMiner Process for the “Aggregate Pie Plot” takes advantage of flash charting capability within the RapidAnalytics Suite. However, in order to arrive at the end visualization, a RM process is required to appropriately transform the data into percentage or count-based groups. As depicted in Figure 3.12, the first step in the RM process is reading from the database via the ‘Read Database’ operator. We can see in the SQL statement in Figure 3.13 that the operator both applies the user-specified filters and selects a count for the chosen aggregate attribute. This query

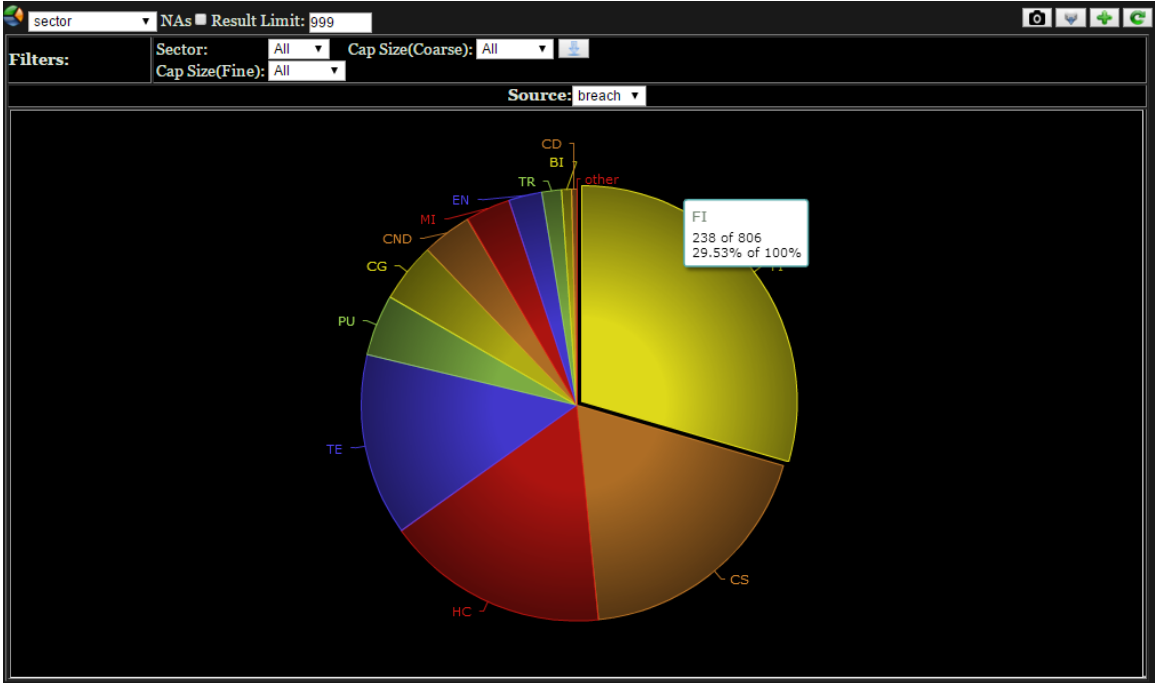


Figure 3.11. Aggregate Pie Plot

is applied after choosing from the top N results (specified from the user-interface). The returned results are then grouped by the aggregate attribute and returned from the 'Read Database' operator. The data then flows to the 'Branch' operator where 'NA's are optionally filtered out by using the filter example-set operator. Data then transitions into the rename operator where the aggregate and count attributes get renamed to something more user friendly. Finally, the data is made available for download via the 'Write CSV' operator before arriving at the output node. After the process was run on the RapidAnalytics Server and exported as a web service, the RapidAnalytics Pie Flash Chart options were defined as seen in Figure 3.14.

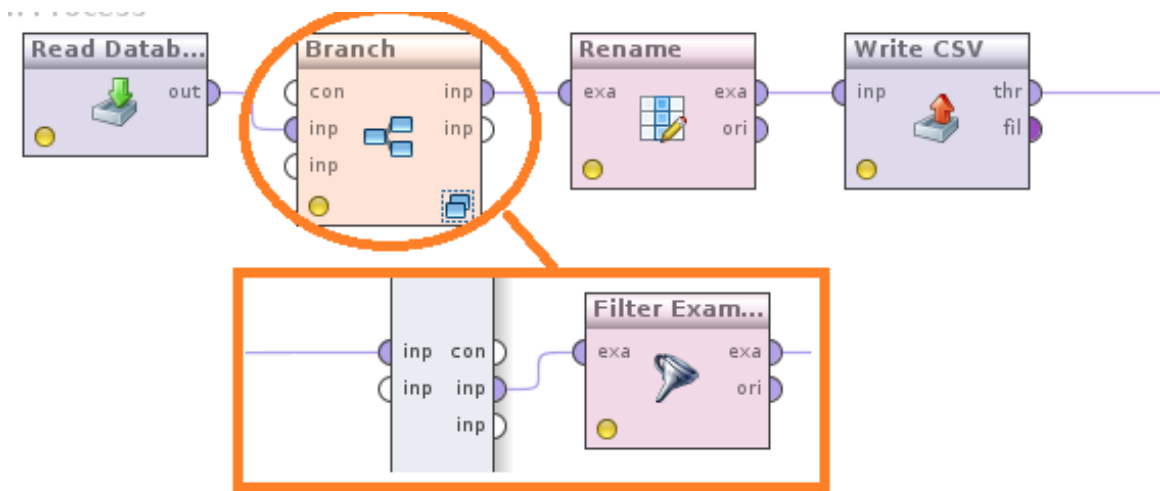


Figure 3.12. RapidMiner Aggregate Pie Process

---

```

(
SELECT {%aggregateattribute}, count(*) as count from ‘{%tablename}’
where {%filter1attr}{%filter1op}’{%filter1val}’
and {%filter2attr}{%filter2op}’{%filter2val}’
and {%filter3attr}{%filter3op}’{%filter3val}’
and {%filter4attr}{%filter4op}’{%filter4val}’
AND ( {%timeattribute} IS NULL OR
( {%timeattribute} >= ‘{%timebegin}’
AND {%timeattribute} <= ‘{%timeend}’))
group by {%aggregateattribute}
ORDER BY count(*) DESC
LIMIT 0, {%resultlimit}
)
UNION
(
SELECT "other" as {%aggregateattribute}, SUM(count) as count FROM(
SELECT {%aggregateattribute}, count(*) as count from ‘{%tablename}’
where {%filter1attr}{%filter1op}’{%filter1val}’
and {%filter2attr}{%filter2op}’{%filter2val}’
and {%filter3attr}{%filter3op}’{%filter3val}’
and {%filter4attr}{%filter4op}’{%filter4val}’
AND ( {%timeattribute} IS NULL OR
( {%timeattribute} >= ‘{%timebegin}’
AND {%timeattribute} <= ‘{%timeend}’))
group by {%aggregateattribute}
ORDER BY count(*) DESC
LIMIT {%resultlimit}, 4565656565
) as count
)

```

---

Figure 3.13. Aggregate Pie Plot SQL ‘Read Database’ Script

<i>Format parameters</i>	Diagram type	pie	<input checked="" type="checkbox"/> Hide style parameters	
	<b>General display settings</b>			
	Title	Aggregate		
	<b>X-axis</b>			
	X-axis attribute	aggregate	X-axis label	percent
	Step size	2.0	Tick height	5
	Label step size	1	Label rotation angle	0
	<b>Y-axis</b>			
	Tick label	percentage_count	Y-axis label	aggregate
	Explicit maximum	<input type="checkbox"/>	Maximum	10.0
	Explicit minimum	<input type="checkbox"/>	Minimum	0.0
	Explicit step size	<input type="checkbox"/>	Step size	10
	Tick length	5	Label rotation angle	horizontal
			Show right axis	<input type="checkbox"/>
	<b>Formatting settings</b>			
Date format	medium	Time format	medium	
Missing data message	No data!			
<b>Diagram specific settings (pie)</b>				
Start angle	-90	Gradient fill	<input checked="" type="checkbox"/>	
Fade	<input checked="" type="checkbox"/>	Distance	6	
Bounce effect	bounce			
Fixed radius				
Tooltip	#label# #val# of #total#<			

Figure 3.14. RapidAnalytics Pie Flash Chart Options

### 3.3.3. Time Aggregate Bar Plot

#### 3.3.3.1. Visualization

The intended purpose of the “Time Aggregate Bar Plot” is to visualize relative percentages of a specific attribute within a dataset over time. Viewing these relative percentages as a piece of a larger whole over time can provide one insight into the how the distribution of a dataset changes. While this visualization does not provide any statistical relevancy, it does help as a first pass discriminator in determining attributes for follow-on time-based analysis with statistically relevant plots. (e.g. time-based odds ratio)

The user-interface for the visualization is depicted below in Figure 3.15. Controls are used to specify an attribute for aggregation and time interval (either month or year) for the x-axis. A stacked bar chart is then presented of the relative percentages of the total for each time interval. In addition, controls are provided to both filter on specific attribute values as well as filter out attributes that are not of interest as discussed in 3.4.2.

#### 3.3.3.2. Process Design

The RapidMiner Process for the “Time Aggregate Bar Plot” takes advantage of charting capability native to the RapidMiner tool. This isn’t the flash charting as detailed with the “Aggregate Pie Plot” discussed above, but a more-scaled back static image. These are the same plot images generated within RapidMiner itself when using the “Plot View”. In order to generate these images by way of RapidAnalytics (RA), we make use of operators within the Reporting Extension. In Figure 3.16 we specify the file URI for the image with the ‘Generate Report’ operator. A subprocess is then invoked to draw the plot and perform filtering. As a first step in this subprocess, the

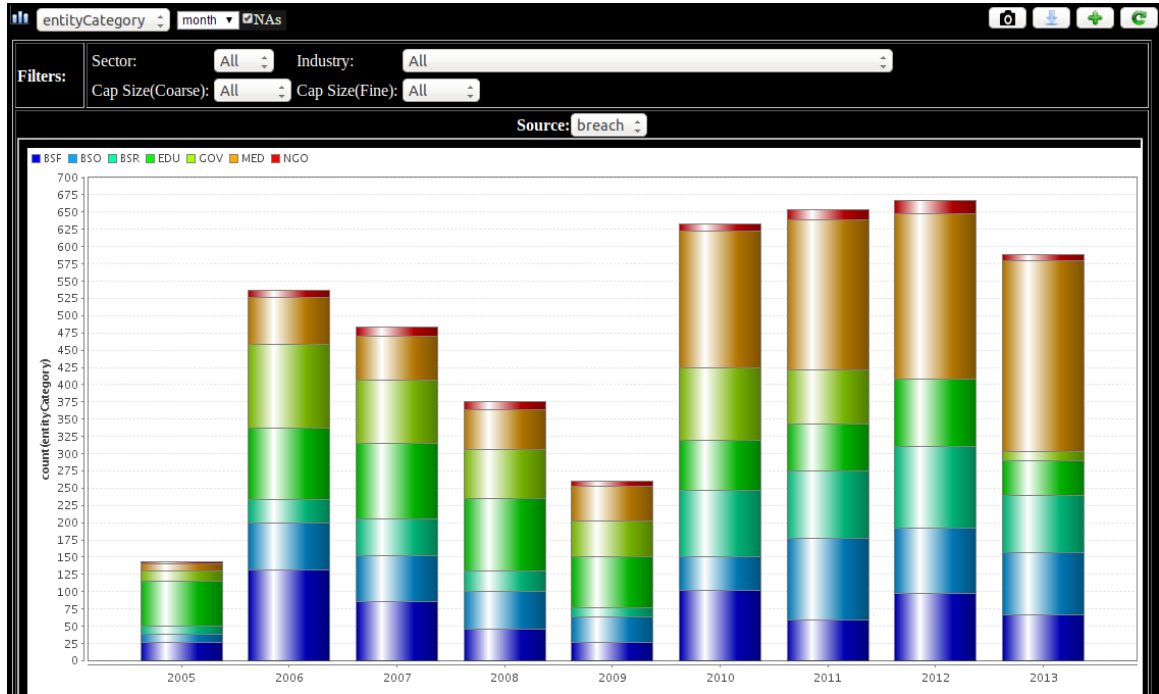


Figure 3.15. Time Aggregation(BAR) View

database is read with the ‘Read Database’ operator with the SQL statement depicted in Figure 3.17 . This goes into an optional filter to filter out ‘NAs’. We then sort by date using the sort operator and finally generate the bar plot by piping into the ‘Generate Report’ operator and defining the associated options. The web client code then periodically refreshes for new content once a request is made.

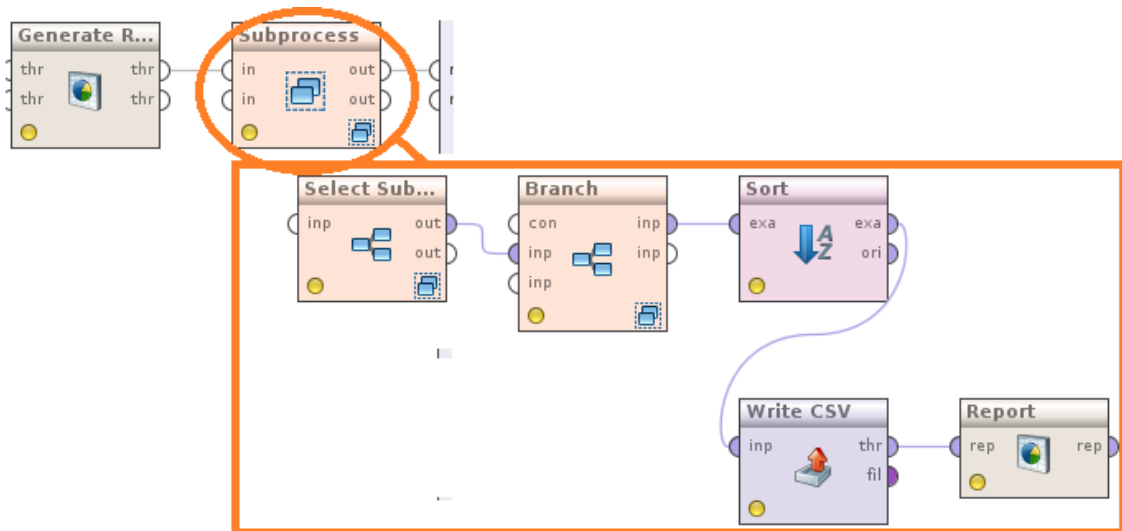


Figure 3.16. RapidMiner Bar Chart Process

---

```

SELECT {%aggregateattribute}, DATE_FORMAT(%{timeattribute}, '%Y/%m'), COUNT(*) FROM {%tablename}
where {%filter1attr}{%filter1op}'{%filter1val}'
and {%filter2attr}{%filter2op}'{%filter2val}'
and {%filter3attr}{%filter3op}'{%filter3val}'
and {%filter4attr}{%filter4op}'{%filter4val}'
AND {%timeattribute} >= '%{timebegin}'
AND {%timeattribute} <= '%{timeend}'
GROUP BY {%aggregateattribute}, DATE_FORMAT(%{timeattribute}, '%Y/%m');

```

---

Figure 3.17. Time-Aggregate(BAR) SQL 'Read Database' Script



### 3.3.4. Time Line Plot

#### 3.3.4.1. Visualization

The intended purpose of the “Time Line Plot” is to visualize frequencies of a specific attribute within a dataset over time. Viewing these frequencies is a quick way to provide insight into the how the distribution of a dataset changes over time. While similar to the “Time Aggregate Bar Plot”, the “Time Line Plot” does not represent data as a piece of a larger whole, but rather instead provides the user raw counts. While this visualization does not provide any statistical relevancy, it does help as a first pass discriminator in determining attributes for follow-on time-based analysis with statistically relevant plots. (e.g. time-based odds ratio)

The user-interface for the visualization is depicted below in Figure 3.15. Controls are used to specify an attribute to track as well as time interval (month or year) for the x-axis. A line chart is then presented of the counts at each time interval. In addition, controls are provided to both filter on specific attribute values, as well as, filter out attributes that are not of interest as discussed in 3.4.2.

#### 3.3.4.2. Process Design

The RapidMiner Process for the “Time Line Plot” takes advantage of charting capability within the R statistical programming language. In order to generate these images by way of RA, we make use of operators within the R Extension. In Figure 3.19 we begin the subprocess by using the ‘Read Database’ operator with the SQL statement depicted in Figure 3.17. The ‘Time Type’ (month or year) varies the context of the SQL statement minimally. After employing the ‘Rename’ operator to rename some attributes, the data flows into optional branching logic to filter out ‘NAs’. We then sort by date using the sort operator and conclude by generating

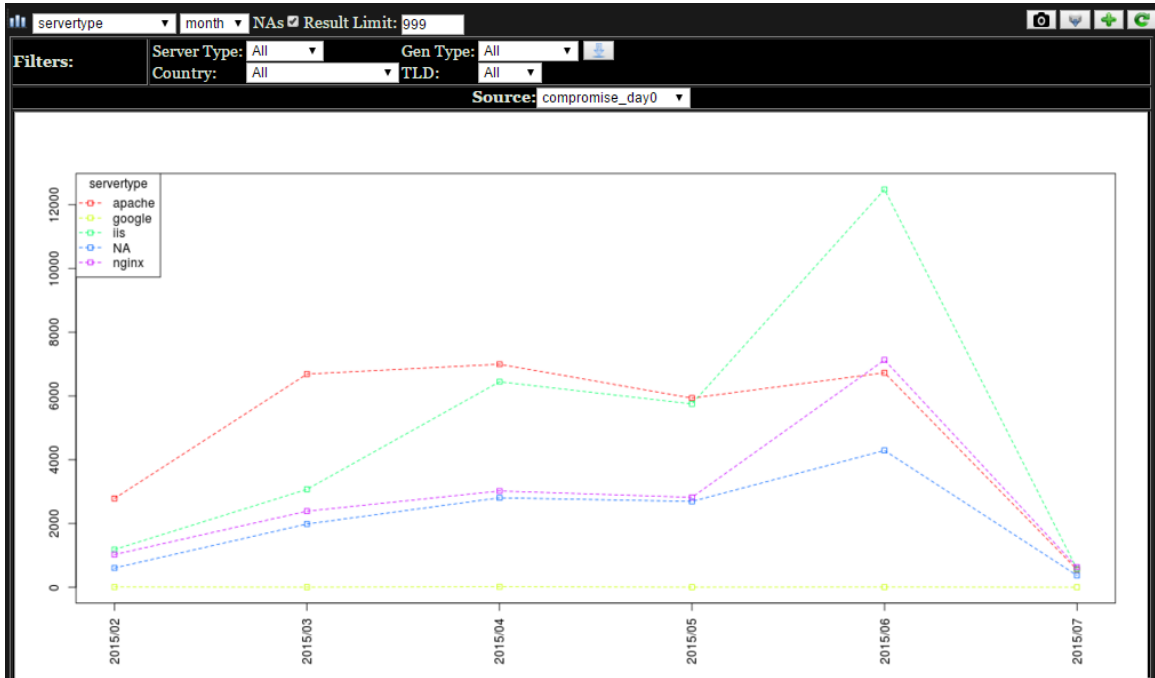


Figure 3.18. Time Line Plot

the line plot with the ‘Execute R Script’ operator. The reference script in Figure 3.21 works by creating a png file stream to paint to. The data is then converted to a two-way occurrence table by invoking the xtabs command and specifying the two attributes of interest. The plot is then opened, and a line is created for each value. The file stream is subsequently closed. The web client code then periodically refreshes for new content once a request is made.

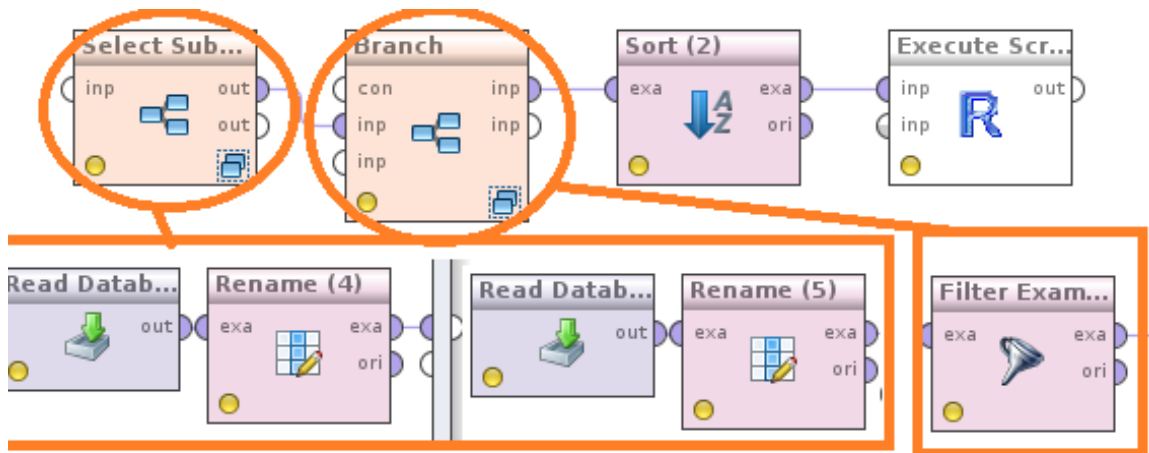


Figure 3.19. RapidMiner Time Line Process

---

```

select i1.#{aggregateattribute} as #{aggregateattribute},
       DATE_FORMAT(#{timeattribute},'%Y/%m'),
       count(*)
from '#{tablename}' join
(
  SELECT #{aggregateattribute}, COUNT(*) as cnt
  FROM '#{tablename}'
  where #{filter1attr}#{filter1op}'#{filter1val}'
  and #{filter2attr}#{filter2op}'#{filter2val}'
  and #{filter3attr}#{filter3op}'#{filter3val}'
  and #{filter4attr}#{filter4op}'#{filter4val}'
  AND #{timeattribute} >= '#{timebegin}'
  AND #{timeattribute} <= '#{timeend}'
  GROUP BY #{aggregateattribute}
  order by cnt desc
  limit #{resultlimit}
) i1
ON #{tablename}.#{aggregateattribute} = i1.#{aggregateattribute}
group by i1.#{aggregateattribute},DATE_FORMAT(#{timeattribute},'%Y/%m');

```

---

Figure 3.20. Time-Line SQL 'Read Database' Script

---

```

#Print out the input table
write.table(MyInput, "%{outdir}/timeline%{chartnum}/download.txt", sep=" ")

ti <- sort(unique(unlist(MyInput$time)))
allaggregate <- unique(unlist(MyInput$aggregate))

png("%{outdir}/timeline%{chartnum}/%{requestid}.png", %{width}, %{height})

y <- xtabs(count ~ aggregate+time, data=MyInput)
y <- y[, order(as.integer(colnames(y)))]

ylim <- c(0, max(MyInput$count))
plot(order(as.integer(ti)), NULL,
      type="n",           # Plot nothing
      main="Aggregate vs. Time", # Main title for the plot
      xlab="Time",        # Label for the x-axis
      ylab="Count",       # Label for the y-axis
      ylim=ylim,         # Range for the y-axis; "xlim" does same for x-axis
      xaxt='n', ann=FALSE)

# initialize random colors
colors <- rainbow(length(allaggregate))

#loop on all oddsvr variables
for(row in 1:length(allaggregate)) {
lines(x=ti, y=y[row,],           # x and y
      type="o",                 # Plot lines and points
      lty=2,                    # Line type: 0=blank, 1=solid, 2=dashed, 3=dotted, 4=dotdash, 5=longdash, 6=twodash
      lwd=1,                    # Line width
      pch=22,                   # Point type
      colors[row])              # Color of the plotted data
}

pchs <- rep(22,length(allaggregate))
ltys <- rep(2,length(allaggregate))
lwds <- rep(1,length(allaggregate))

# Add a legend to the plot
legend("topleft",              # x-y coordinates for legend
      legend=allaggregate,     # Legend labels
      col=colors,              # Color of points or lines
      pch=pchs,                # Point type
      lty=ltys,                # Line type
      lwd=lwds,                # Line width
      title="%{aggregateattribute}") # Legend title

axis(1, at=c(1:length(ti)), labels=unique(unlist(MyInput$time)), col.axis="black", las=2)

dev.off()

```

---

Figure 3.21. Time Line Plot Process Design R-Script

### 3.3.5. Mosaic Plot

#### *3.3.5.1. Visualization*

The “Mosaic Plot” Visualization in Figure 3.22 is useful for visualizing data for two categorical variables. If configured correctly, it can be leveraged to show statistically relevant overlap between multiple attribute values, where said attribute value pairing is over-represented or under-represented and whether that value is statistically relevant or not. From a statistical standpoint, the mosaic plot is really a depiction of the chi-squared residuals. For extra flexibility, the residual method is selectable, which alters the plots slightly.

After choosing the x-axis and y-axis attributes, performing filtering, and optionally limiting the result set, the user is presented with red and blue squares of varying hues. Blues indicate over-represented cross-pairings, while reds indicate under-represented cross-pairings. The darker the hue, the more statistically relevant the cross-pairing. White or hollow squares indicate a cross-pairing that is not statistically relevant. The test statistic, p-value, and degrees of freedom are supplemented in order to assess statistical relevance further.

Using this plot, one can quickly determine statistical relevance of cross-pairings. The onus is then on the user to decide whether the result is meaningful within the overall context of the analysis.

#### *3.3.5.2. Process Design*

The process for the Mosaic Plot is depicted below in Figure 3.24. The process again begins with the ‘Read Database’ operator with SQL statement depicted in Figure 3.23. After applying filtering to remove ‘NAs’, it leverages the ‘Execute R Script’ operator to calculate and display the plot. The reference script in Figure 3.25



Figure 3.22. Mosaic Plot

works by creating a png file stream to paint to. The data is then converted to a 2-way occurrence table by invoking the table command and specifying the two attributes of interest. A mosaic plot is then plotted and written to the file stream. The file stream is subsequently closed. The web client code then periodically refreshes for new content once a request is made.

---

```

select coalesce(i1.#{input1}, 'other') as #{input1},
       coalesce(i2.#{input2}, 'other') as #{input2},
       count(*)
from '#{tablename}' left join
  (select #{input1}, count(*) as cnt
   from '#{tablename}'
   where #{filter1attr}#{filter1op}'#{filter1val}'
   and  #{filter2attr}#{filter2op}'#{filter2val}'
   and  #{filter3attr}#{filter3op}'#{filter3val}'
   and  #{filter4attr}#{filter4op}'#{filter4val}'
   AND (#{timeattribute} IS NULL OR
        (#{timeattribute} >= '#{timebegin}'
         AND #{timeattribute} <= '#{timeend}'))
   group by #{input1}
   order by cnt desc
   limit #{resultlimit1}
  ) i1
on #{tablename}.#{input1} = i1.#{input1} left join
  (select #{input2}, count(*) as cnt
   from '#{tablename}'
   where #{filter1attr}#{filter1op}'#{filter1val}'
   and  #{filter2attr}#{filter2op}'#{filter2val}'
   and  #{filter3attr}#{filter3op}'#{filter3val}'
   and  #{filter4attr}#{filter4op}'#{filter4val}'
   AND (#{timeattribute} IS NULL OR
        (#{timeattribute} >= '#{timebegin}'
         AND #{timeattribute} <= '#{timeend}'))
   group by #{input2}
   order by cnt desc
   limit #{resultlimit2}
  ) i2
on #{tablename}.#{input2} = i2.#{input2}
group by i1.#{input1}, i2.#{input2};

```

---

Figure 3.23. Mosaic Plot SQL ‘Read Database’ Script

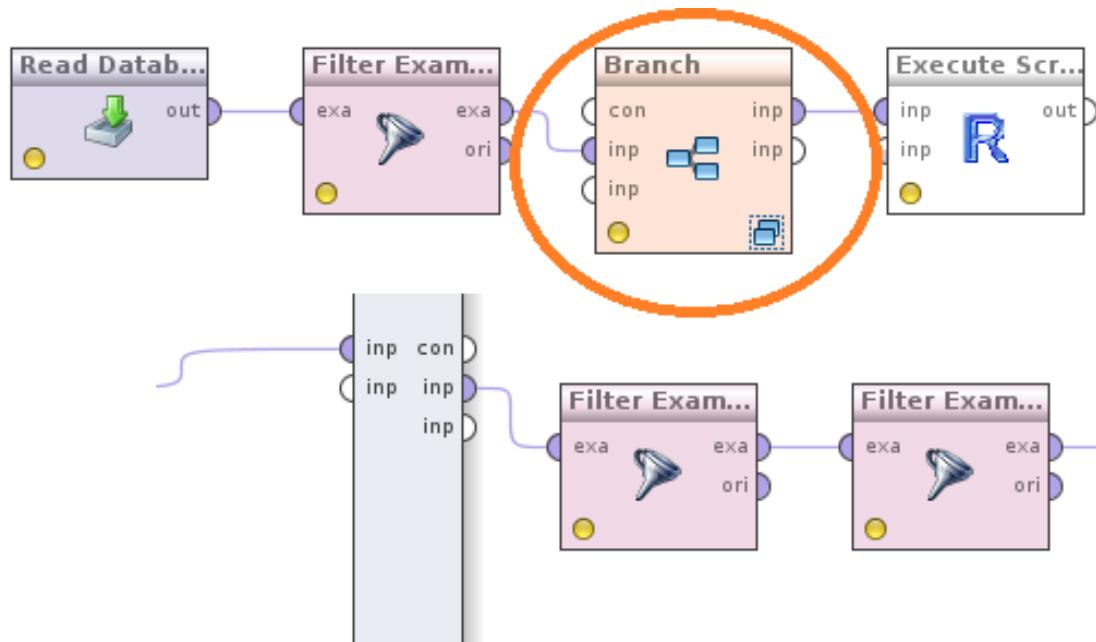


Figure 3.24. RapidMiner Mosaic Plot Process

---

```

png("%{outputdir}/mosaic%{chartnum}/%{requestid}.png", %{width}, %{height})

write.table(MyInput, "%{outputdir}/mosaic%{chartnum}/MyInput.txt", sep=" ")
MyTable <- xtabs(count ~ %{input1}+%{input2}, data=MyInput)
MyTable <- MyTable[rowSums(MyTable[, -1])>0, ]
MyTable <- MyTable[,colSums(MyTable[-1,])>0]
write.table(MyTable, "%{outputdir}/mosaic%{chartnum}/TOH.srt", sep=" ")
TOH.cs<-chisq.test(MyTable)
write.table(MyTable, "%{outputdir}/mosaic%{chartnum}/download.txt", sep=" ")

x <-paste("X-squared=",TOH.cs$statistic,sep=" ")
df <-paste("df=",TOH.cs$parameter,sep=" ")
p <-paste("p-value<",TOH.cs$p.value,sep=" ")
all <-paste(x,"",df,"",p)

mosaicplot(MyTable,main = "%{input1} vs %{input2}", sub=all, xlab="%{input1}", ylab="%{input2}", shade=T, cex.axis = 1.1,
            type="%{residualtype}")
dev.off ()

```

---

Figure 3.25. Mosaic Plot Process Design R-Script



### 3.3.6. Box Plot

#### *3.3.6.1. Visualization*

The “Box Plot” visualization in Figure 3.26 is practical for graphically depicting groups of categorical data through their quartiles. A numerical value must be chosen for the y-axis, however any categorical variable can be chosen for the x-axis. In addition, controls are provided to both filter on specific attribute values as well as filter out attributes that are not of interest as discussed in 3.4.2. The plot then depicts the relative distributions of each value of the categorical x-axis variable. The median is depicted as a black solid line with the 25% to 75% marks forming the “box”. The whiskers representing the 95% confidence intervals protrude out vertically from the box additionally indicate variability outside the upper and lower quartiles. Any remaining dots depict the outliers which are not representative of the bulk of the data.

The “Box Plot” is a statistically relevant plot. If the median of one box does not intersect the entirety of the box of another, there is likely a statistically significant difference in means. On the other hand, if a box (not including whiskers) fails to overlap another box (one can determine this by drawing a horizontal line between them), there is unequivocally a statistically significant difference in means. Keeping this in mind, one can quickly scan for statistical significance for numerical distributions within a dataset.

### 3.3.7. Process Design

The process for the Box Plot Visualization is depicted in Figure 3.28). It begins with the ‘Read Database’ operator which executes the SQL statement depicted in Figure 3.27. From there, the process optionally removes ‘NAs’ with the ‘Filter Exam-

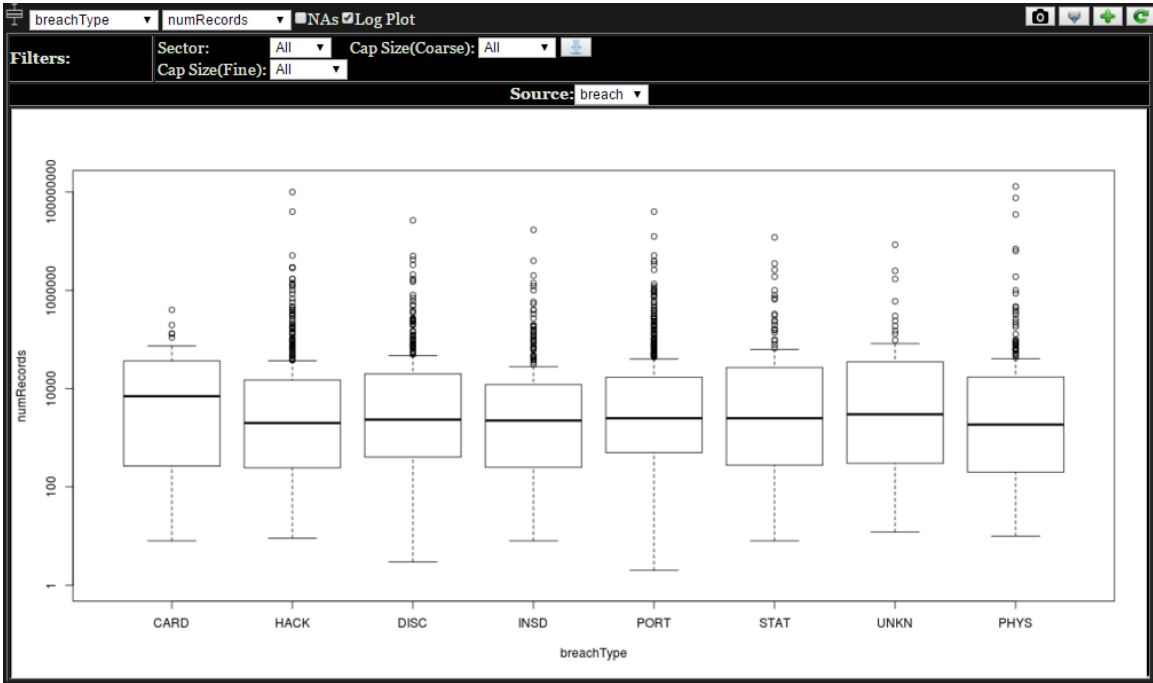


Figure 3.26. Box Plot

pletset' operator. Finally, the 'Execute R script' component is invoked to perform the box plot calculation (shown in Figure 3.29). The web client code then periodically refreshes for new content once a request is made.

---

```

SELECT %{input1},{input2} FROM %{tablename}
WHERE %{filter1attr}%{filter1op}'%{filter1val}'
AND %{filter2attr}%{filter2op}'%{filter2val}'
AND %{filter3attr}%{filter3op}'%{filter3val}'
AND %{filter4attr}%{filter4op}'%{filter4val}'
AND (%{timeattribute} IS NULL OR
(%{timeattribute} >= '%{timebegin}'
AND %{timeattribute} <= '%{timeend}'));

```

---

Figure 3.27. Box Plot SQL 'Read Database' Script

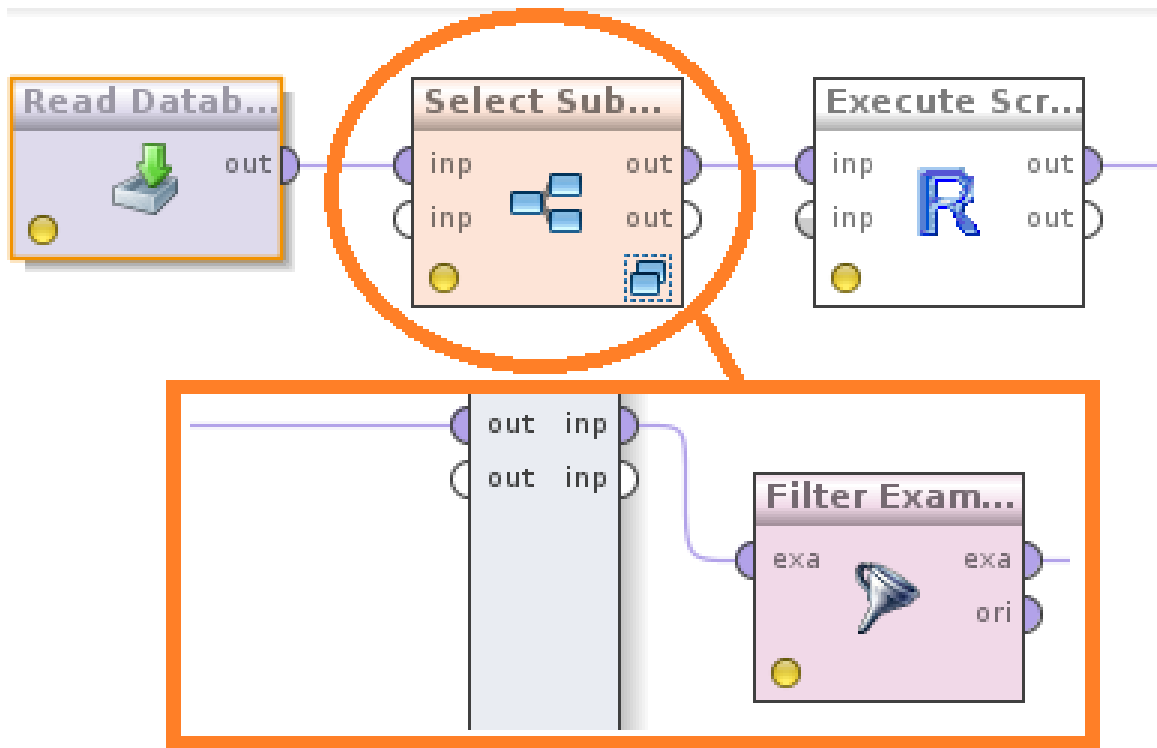


Figure 3.28. RapidMiner Box Plot Process

---

```

png("%{outputdir}/box%{chartnum}/%{requestid}.png", %{width}, %{height})

MyFrame <- as.data.frame(MyInput)
MyFrame <- MyFrame[MyFrame$%{input2}!=0,]

write.table(MyFrame, "%{outputdir}/box%{chartnum}/download.txt", sep=" ")

logopt = ""
if (%{loggraph}) {
logopt = "y"
}
boxplot(%{input2}~%{input1}, data=MyFrame, log=logopt, ylim=c(1, max(MyFrame$%{input2})), xlab="%{input1}",
ylab="%{input2}")
dev.off()

```

---

Figure 3.29. Box Plot Process Design R-Script

### 3.3.8. Odds Ratio Plot

#### 3.3.8.1. Visualization

The “Odds Ratio Plot” (Figure 3.30) is a necessary tool for case-study analysis. The visualization depicts the odds of an outcome occurring given a particular exposure, in relation to the odds of the outcome occurring in absence of that exposure.

In the user-interface depicted, after the user selects an odds ratio table combination to analyze, the user chooses an odds variable. In addition, controls are provided to both filter on specific attribute values as well as filter out attributes that are not of interest as discussed in 3.4.2. After refreshing, the odds ratios are then visualized as bar plots with the left-most element chosen as the reference normal (1.0). Additionally, 95% confidence intervals (whiskers) are provided to convey whether the results

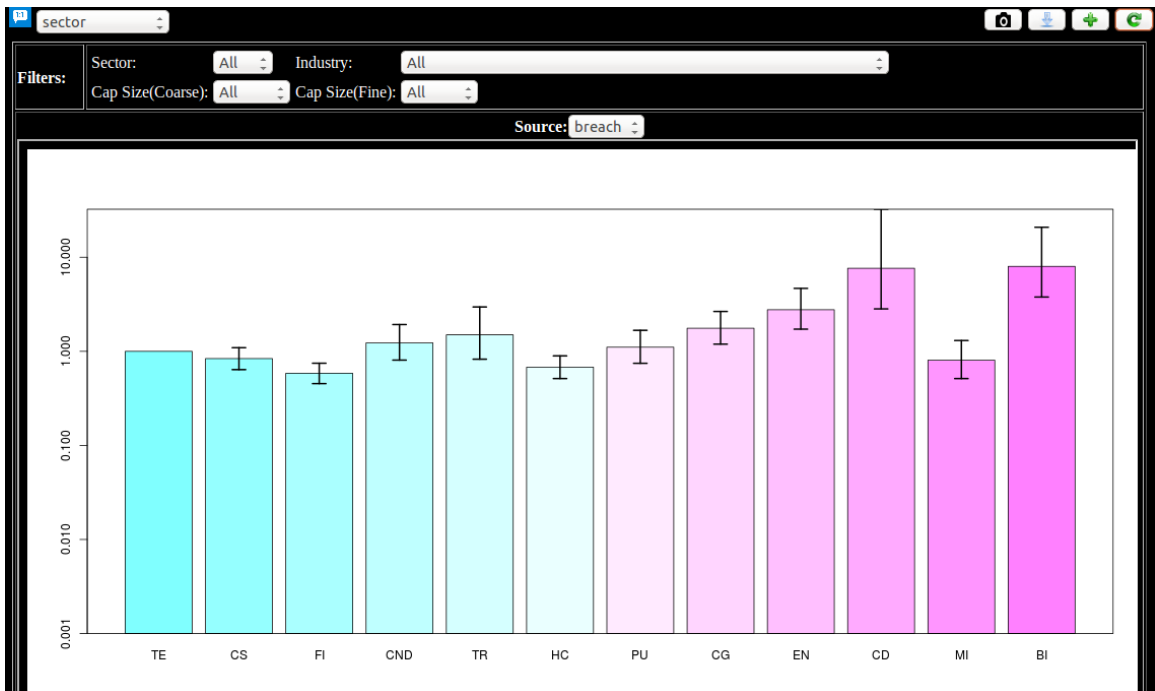


Figure 3.30. Odds Ratio

---

```
SELECT %oddsvar, COUNT(*) FROM %ctl_tablename}
where %filter1attr}%filter1op}'%filter1val}'
and %filter2attr}%filter2op}'%filter2val}'
and %filter3attr}%filter3op}'%filter3val}'
and %filter4attr}%filter4op}'%filter4val}'
AND (%timeattribute} IS NULL OR
(%timeattribute} >= '%timebegin}'
AND %timeattribute} <= '%timeend}'))
GROUP BY %oddsvar};
```

---

Figure 3.31. Odds Ratio SQL ‘Read Database’ Script

are statistically valid. Non-overlapping odds ratio ranges are considered statistically significant.

### 3.3.8.2. Process Design

The process (Figure 3.32) for odds ratio visualization starts off with a dual read from both the control table and the treatment table via the ‘Read Database’ operator (Figure 3.31). A new attribute called “non-control” set to “True” or “False” is then concatenated to each example-set respectively designating which table it originated from. Both example-sets are then joined (i.e. appended) to one another creating a super-set of both example-sets. From there, a filter is provided to hone in on specific attribute values of interest. An ‘Execute R script’ component utilizing the epitools library (Figure 3.33) is then invoked to perform the odds ratio calculation and generate the odds ratio bar plot with confidence interval bars. Finally, the web client code then periodically refreshes for new content once a request is made.

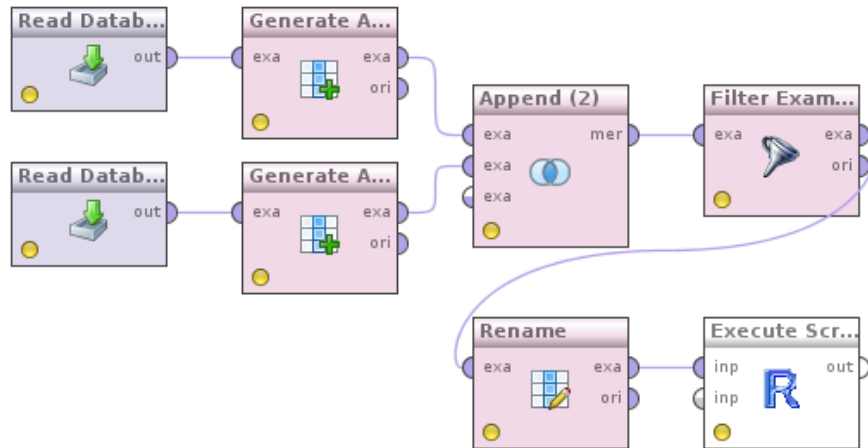


Figure 3.32. RapidMiner Odds Ratio Process

---

```

png("%{outputdir}/odds%{chartnum}/%{requestid}.png", %{width}, %{height})
write.table(MyInput, "%{outputdir}/odds%{chartnum}/MyInput.txt", sep=" ")

MyTable <- xtabs(count ~ %{oddsvar}+noncontrol, data=MyInput)
MyTable <- MyTable[!rowSums(MyTable == 0), ]
MyTable <- MyTable[,c(2,1)]

write.table(MyTable, "%{outputdir}/odds%{chartnum}/MyTable.txt", sep=" ")
library(epitools)
odds <- oddsratio(MyTable, rev="both", conf.level=%{confevel})
write.table(odds$measure, "%{outputdir}/odds%{chartnum}/download.txt", sep=" ")
transpodds <- t(odds$measure)
options(scipen=5)
ylim <- c(0.001, max(transpodds, na.rm=TRUE))

logopt = ""
if (%{loggraph}) {
logopt = "y"
}
mp <- barplot(transpodds[1,], log=logopt, ylim=ylim, col=cm.colors(length(transpodds[1,])))

box()
abline(h=1.0)
segments(mp, transpodds[2,], mp, transpodds[3,], lwd=2)
segments(mp - 0.1, transpodds[2,], mp + 0.1, transpodds[2,], lwd=2)
segments(mp - 0.1, transpodds[3,], mp + 0.1, transpodds[3,], lwd=2)
dev.off()

```

---

Figure 3.33. ODDS Ratio Process Design R-Script

### 3.3.9. Time-Based Odds Ratio Plot

#### 3.3.9.1. Visualization

The “Time-Based Odds Ratio Plot” (see Figure 3.34) helps us visualize how the odds ratios change over time. Confidence intervals markers (whiskers are provided which allow one to determine the statistical ranges of each data point in relation to the normal.

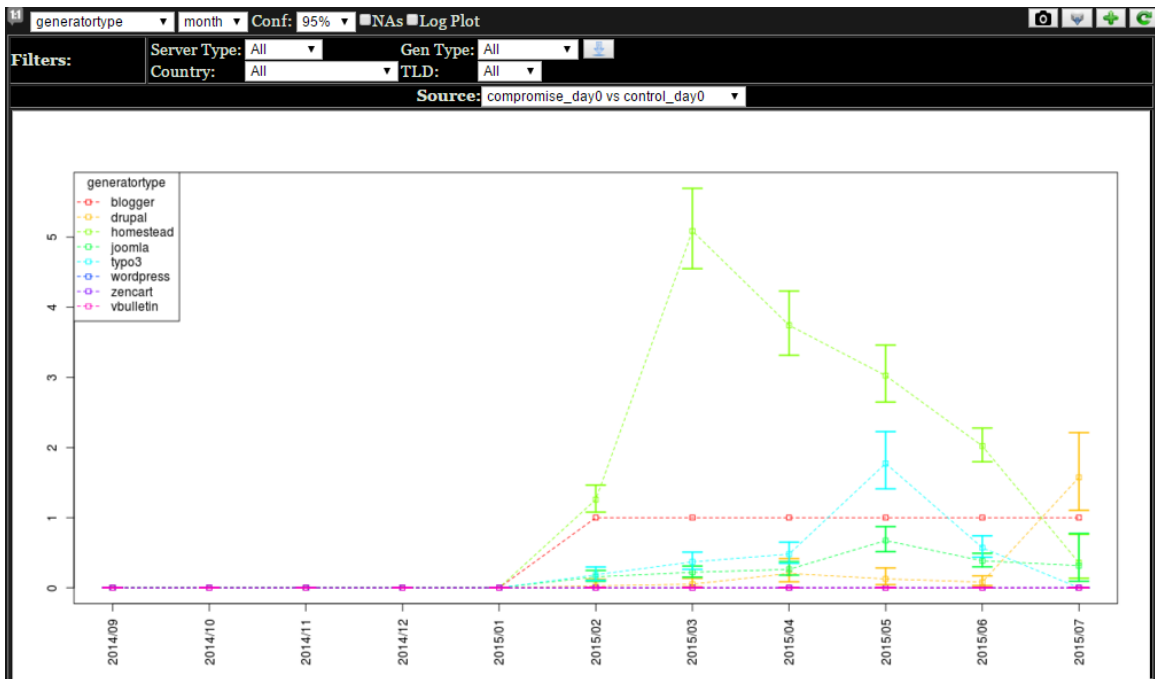


Figure 3.34. Time-Based Odds Ratio

In the user-interface depicted, after the user selects an odds ratio table combination to analyze, the user chooses an odds variable. In addition, controls are provided to both filter on specific attribute values as well as filter out attributes that are not of interest as discussed in 3.4.2. After refreshing, the odds ratios are then visualized as a line plot with the one element chosen as the reference normal (1.0). Addition-

ally, 95% confidence intervals (whiskers) are provided to convey whether the results are statistically valid. Non-overlapping odds ratio ranges are considered statistically significant.

### 3.3.9.2. Process Design

The process (Figure 3.35) for the time-based odds ratio plot mirrors the odds ratio process above (Section 3.3.8.2). The ‘Read Database’ Operator executes the SQL statement depicted in Figure 3.37. The ‘Execute R script’ operator plots these odds ratios over the associated time intervals by running the R-script delineated in Figure 3.36. The web client code then periodically refreshes for new content once a request is made.

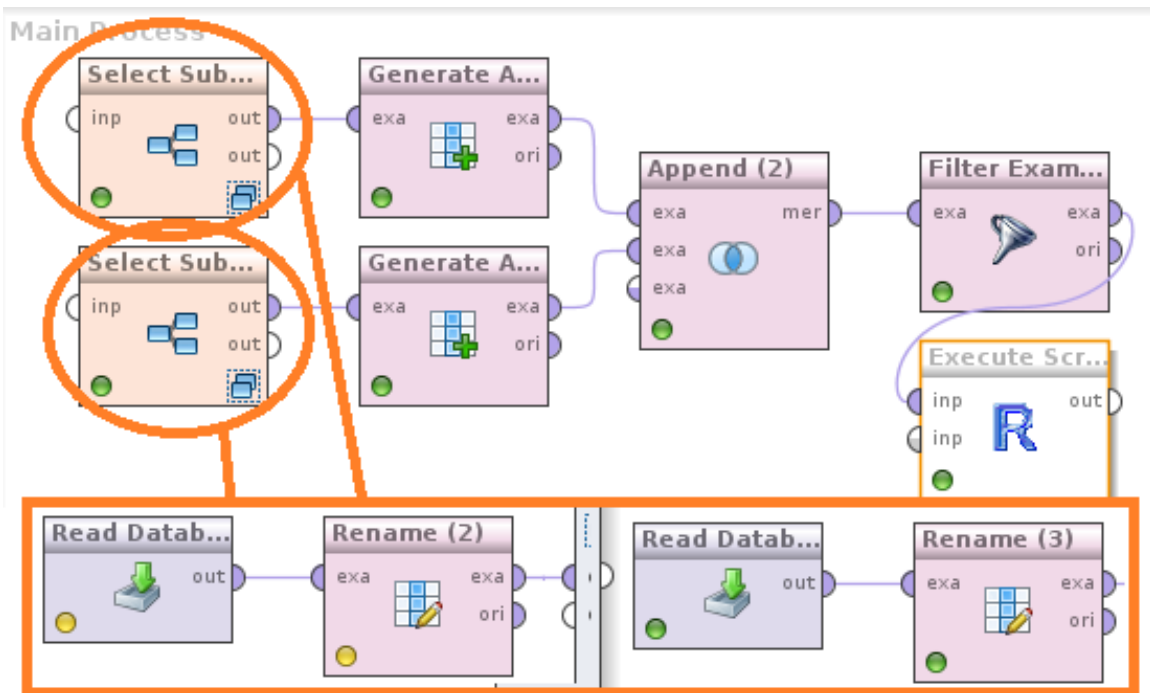


Figure 3.35. RapidMiner Time-Based Odds Ratio Process



```

ti <- unique(unlist(MyInput$time)); alloddsvar <- unique(unlist(MyInput$%{oddsvar}))

library(epitools)
y <- matrix(, nrow = length(ti), ncol = length(alloddsvar))
yconfmin <- matrix(, nrow = length(ti), ncol = length(alloddsvar)); yconfmax <- matrix(, nrow = length(ti), ncol =
length(alloddsvar))

tilength = length(ti); oddslength = length(alloddsvar)
for(row in 1:length(ti)) { for(col in 1:length(alloddsvar)) {
  yconfmin[row,col] <- 0.0; y[row,col] <- 0.0; yconfmax[row,col] <- 0.0
} }

ylim <- c(0.001, 1.0)
for(row in 1:length(ti)) {
  MyInputSubset <- MyInput[MyInput$time==ti[row],]
  inputname <- paste("%{outputdir}/oddstime%{chartnum}/MyInput",row, sep="");

  MyTable <- xtabs(count ~ %{oddsvar}+noncontrol, data=MyInputSubset, )
  MyTable <- MyTable[!rowSums(MyTable == 0), ]; MyTable <- MyTable[,c(2,1)]

  if(NROW(MyTable)>=2) {
    odds<-oddsratio(MyTable,rev="both", conf.level=%{conlevel})
    oddsname <- paste("%{outputdir}/oddstime%{chartnum}/odds",row, sep="");
    transpodds <- t(odds$measure); options(scipen=5)

    maxy <- max(transpodds, na.rm=TRUE); ylim <- c(0.001, max(ylim[2], maxy))

    for(col in 1:length(transpodds[1,])) {
      y[row,col] <- transpodds[1,col]; yconfmin[row,col] <- transpodds[2,col]; yconfmax[row,col] <- transpodds[3,col]
    } }

write.table(y, "%{outputdir}/oddstime%{chartnum}/download.txt", sep=" ")
png("%{outputdir}/oddstime%{chartnum}/%{requestid}.png", %{width}, %{height})
plot(ti, NULL, type="n", main="Odds Ratio vs. Time", xlab="Time", ylab="Odds Ratio", ylim=ylim, xaxt='n',
ann=FALSE)

colors <- rainbow(length(alloddsvar))
for(col in 1:length(alloddsvar)) {
  lines(ti, y[,col], type="o", lty=2, lwd=1, pch=22, colors[col])
}

pchs <- rep(22,length(alloddsvar)); ltys <- rep(2,length(alloddsvar)); lwds <- rep(1,length(alloddsvar))
legend("topleft", legend=alloddsvar, col=colors, pch=pchs, lty=ltys, lwd=lwds, title="%{oddsvar}")
for(col in 1:length(alloddsvar)) {
  segments(c(1:tilength), yconfmin[,col], c(1:tilength), yconfmax[,col], lwd=2, col=colors[col])
  segments(c(1:tilength) - 0.1, yconfmin[,col], c(1:tilength) + 0.1, yconfmin[,col], lwd=2, col=colors[col])
  segments(c(1:tilength) - 0.1, yconfmax[,col], c(1:tilength) + 0.1, yconfmax[,col], lwd=2, col=colors[col])
}

axis(1, at=c(1:tilength), labels=ti, col.axis="black", las=2); dev.off()

```

Figure 3.36. Time-Based ODDS Ratio Process Design R-Script

---

```

SELECT %oddsvar, DATE_FORMAT(%timeattribute, '%Y/%m'), COUNT(*) FROM %ctltablename}
where %filter1attr}%filter1op}'%filter1val}'
and %filter2attr}%filter2op}'%filter2val}'
and %filter3attr}%filter3op}'%filter3val}'
and %filter4attr}%filter4op}'%filter4val}'
AND %timeattribute} >= '%timebegin}'
AND %timeattribute} <= '%timeend}'
GROUP BY %oddsvar, DATE_FORMAT(%timeattribute, '%Y/%m');

```

---

Figure 3.37. Time-Based Odds Ratio SQL ‘Read Database’ Script

### 3.3.10. Geospatial View

#### 3.3.10.1. Visualization

The “Geospatial View” is a way of visualizing data geographically. Proof of concept was done with countries but could be easily expanded to states or other geographic regions (time zones, MGRS, UTM, etc.). A control to select the source data is provided. The visualization can be utilized for an odds ratio representation or an aggregate. In fact, any set of geospatial metric can be presented by mapping the metric to a range of colors. For example, in the prototype visualization an odds ratio is presented by mapping the value of the risk-factor to a color-intensity pairing, with deeper greens indicating more negative risk and deeper reds indicating more positive risk. Neutral grays indicate statistically insignificant risk factor. The user can then pan, scale, and manipulate the map in other ways as it behaves just like a regular Google Map.

#### 3.3.10.2. Process Design

The Geospatial View utilizes Google Fusion Tables to display data as an “iframe” within the portal window. The chart is produced by uploading data with geographic

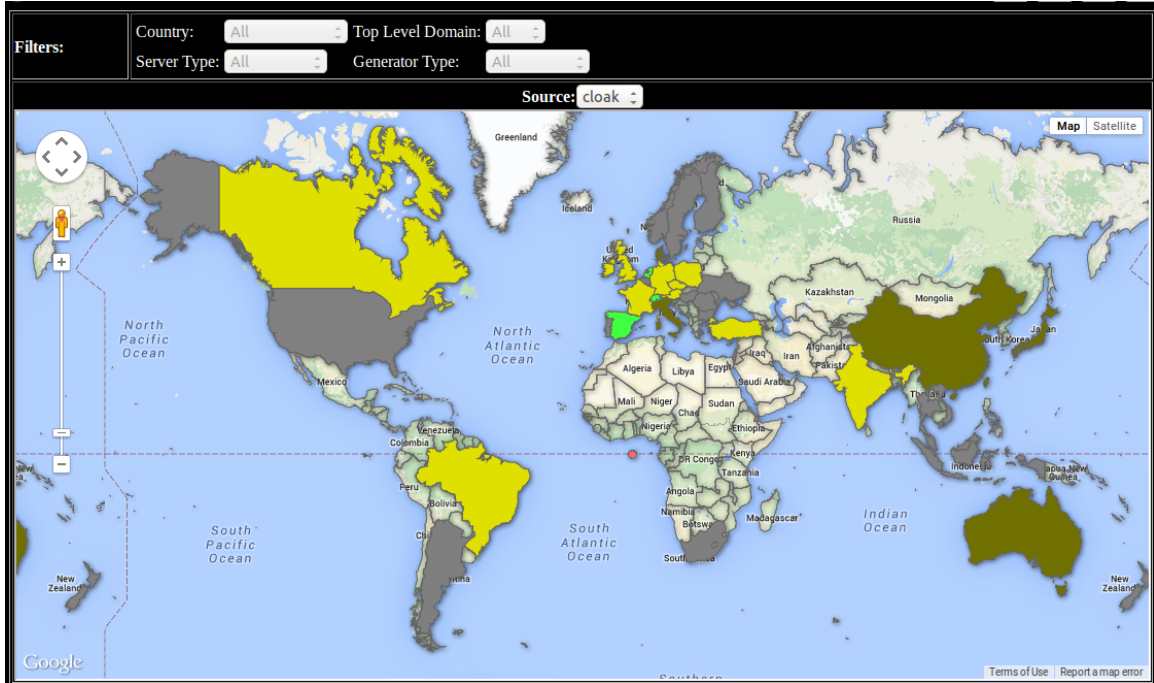


Figure 3.38. Geospatial View

Table 3.2. Google Fusion Table Color-Codes

odds range	color code
0.0 - 0.33	#707000FF
0.33 - 0.66	#DFDF00FF
0.66 - 1.51	#808080FF
1.51 - 3.03	#EE9F9FFF
3.03+	#FF0000FF

information (latitude longitude) to Google Fusion Table site. One can then display the information as markers or as a “heat map”.

In order to produce a “heat map”, we simply associate the latitude with the longitude by selecting the column Fusion Tables and selecting “Change”. We then associate with the longitude column as “Two Column Location”. After this association is performed, we can define the heat map options by going to “Map of Latitude”

## CloakWorld

Imported at Fri Sep 27 09:16:41 PDT 2013 from CloakWorld1.xls.

[Add Attribution](#) - Edited on September 27, 2013

File Edit Tools Help Rows 1 Cards 1 Map of geometry +

Filter No filters applied

1-100 of 236

Name	geometry	ISO	estimate	riskfactor	color
Afghan...	KML...	AF			#00000000
Albania	KML...	AL			#00000000
Algeria	KML...	DZ			#00000000
American Samoa	KML...	AS			#00000000
Andorra	KML...	AD			#00000000
Angola	KML...	AO			#00000000
Anguilla	KML...	AI			#00000000
Antarct...	KML...	AQ			#00000000
Antigua & Barbuda	KML...	AG			#00000000
Argentina	KML...	AR	0.7262730951	none	#808080FF
Armenia	KML...	AM			#00000000
Aruba	KML...	AW			#00000000
Australia	KML...	AU	0.2407572471	neg.	#707000FF
Austria	KML...	AT	0.4108708231	neg.	#DFDF0...
Azera...	KML...	AZ			#00000000

Figure 3.39. Google Fusion Tables

and selecting “Change Map”. We can then configure the radius and opacity and give additional weighting to a specific attribute if desired (Figure 3.40). Visualization is limited to the first 1,000 rows.

Additionally, by merging data containing geospatial information with KML data for geospatial boundaries (see Figure 3.39) one can produce geographic zones for display. In order to do this, we take the download data from the country odds ratios and merge it with with KML data for every political boundary in the world to arrive at a merged dataset for use with Google Fusion Tables. These KMLs contain polygon information in geodetic coordinates for every country in the world. The framework,

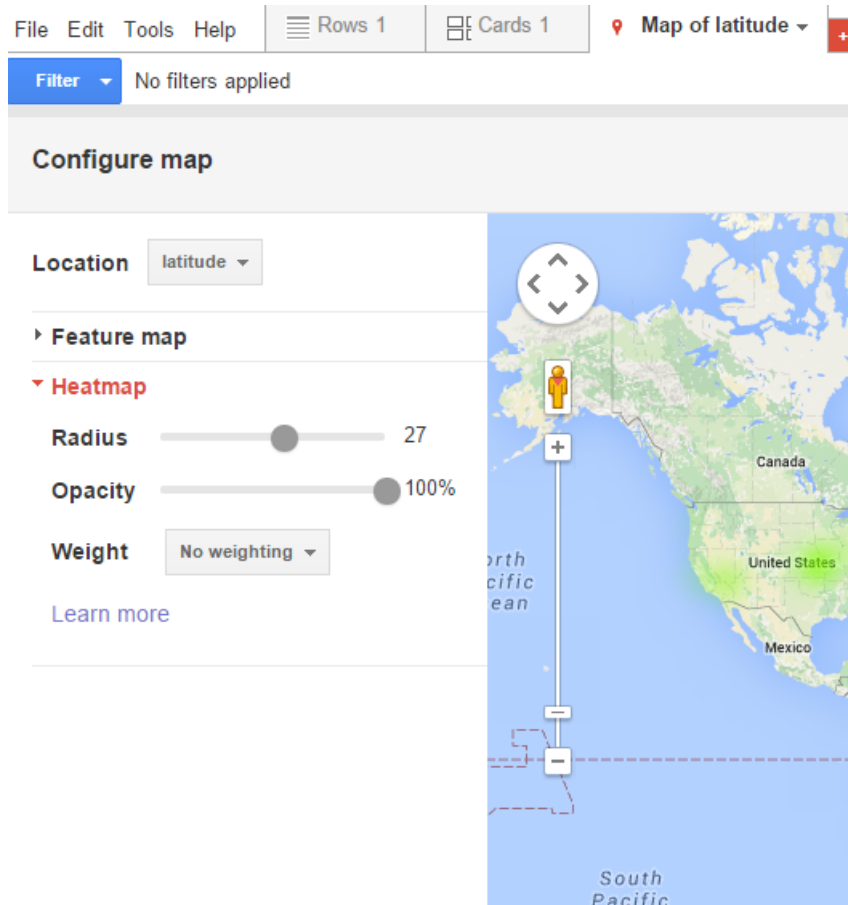


Figure 3.40. Heatmap Options

however, does not limit to this use case. A developer is free to merge any type of data with any set of geospatial KMLs. This merge process can happen within the tool itself or external to the tool (R). After this is completed, we then map colors to each odds ratio value with a conditional formula for a new column, the color column. In our framework, we choose to associate more saturation within the hue with larger positive (red) or negative (green) risk factors. A sample utilized mapping is provided below in Table 3.2, however, one can really apply any coloring one desires by mapping the odds ratio estimate to a hexadecimal color value.

After this data is uploaded to Google, the researcher can go to the “Map of Geometry” tab to observe the results in a “Google Maps” style frame. Additionally, a data-enabled URL is provided for publishing purposes, which we are able to link to within the framework. As there is no connectivity between the filtering controls and the geospatial map, the data can be considered static (i.e. no additional filtering can be performed to alter the data). While limiting the dynamic interaction with the data, this static Google Maps View is a powerful tool in representing the data geographically.

### 3.3.11. Web Service Parameters

The RM processes are both reusable and dynamic through usage of web service parameters. These parameters can be mapped as macros that are made available within a RM process by delimiting in the following way: `%{PARAMETER}`. These can be used as options within operators, provided as SQL options, or specified in R-scripts. Each parameter provides extra flexibility in transforming the data within a RM process in a unique way. When exporting a RM process to a web service (RA), these macros can then be passed into the URL for the Rapid Analytics service. Example usage is depicted below.

```
http://<RA server hostname>:8080?serviceID=?&param1=A&param2=B
```

The specific web service parameters used in the framework are described below in Table 3.3. Some parameters are general and apply to all chart types, while others only exist within the context of a one or more charts. Parameters will take on the default value saved within the RM process if not present in the URL, so it is important to properly define the state.

A survey of the parameters shows the facets of configurability necessary for dynamic interaction. The parameters fall into the following category types: attribute selection (timeattr, aggregateattr, oddsattr input1, and input2), filtering (includeNAs, filter(n)attr, filter(n)op, and filter(n)val), infrastructure (db\_url, chartnum, height, width), data selection (tablename, nonctl\_tablename, ctl\_tablename), and usability options(loggraph, resultlimit, keepattr, timetype).

Table 3.3. Web service Parameters

parameter	chart type	description
aggregateattr	all aggregate types	the attribute to use for the aggregate
chartnum	all	the chart number: 1 (left) or 2 (right) to generate
db_url	all	jdbc url to use for chart to get data e.g. jdbc:mysql://127.0.0.1:3306/breach
filter(n)attr	all	attribute (n=1:4) to use for filtering result set
filter(n)op	all	SQL operation (n=1:4) to use for filtering result set <!= , =, ... >
filter(n)val	all	value (n=1:4) to use for filtering result set <null — value >
height	all	height of generated chart in pixels
includeNAs	all	whether or not to include NAs in results (TRUE or FALSE)
input1—input2	mosaic—boxplot	attributes to use for x-axis (1) and y-axis (2)
keepattr	table	attributes to display with or delimiter e.g. attr1 — attr2
loggraph	boxplot—timeline—oddstime	Whether or not to use a log-based y-axis (TRUE or FALSE)
(non)ctl.tablename	odds—oddstime	tables to compare for odds ratio calculation
oddsattr	odds—oddstime	attribute to use in calculating odds ratios
residual	mosaic	Residual method to use <pearsons—freeman—likelihood>
requestid	all R-based	A unique image id for filename
resultlimit	mosaic—timeline	Only display the top N occurring results
serviceID	all	service ID for visualization e.g. <PIE — ODDS_RATIO — BOXPLOT — etc.>
tablename	all besides odds—oddstime	SQL tablename to use for visualization
timeattr	all	attribute to use for time
timebegin	all	time filter start (date)
timeend	all	time filter stop (date)
timetype	all time-based	time interval for time-based plots (1=year or 2=month)
width	all	width of generated chart in pixels



### 3.4. User-Interface Design Features

#### 3.4.1. Portaling

The ability to do side-by-side monitoring of charts is one of the more powerful features of the framework. Up to two side-by-side portals (see Figure 3.41) are allowed. A second portal can be instantiated with the “+” symbol icon. The “x” button is then subsequently utilized to remove the second portal and restore the user-interface back to the 1 portal configuration. While in the two-portal configuration, the operator can reference different source tables (e.g. control table and treatment table for case-study research) as well as define different filters. The two portals operate on the same source metric (aggregate, odds ratio variable, etc.) and are refreshed in tangent.

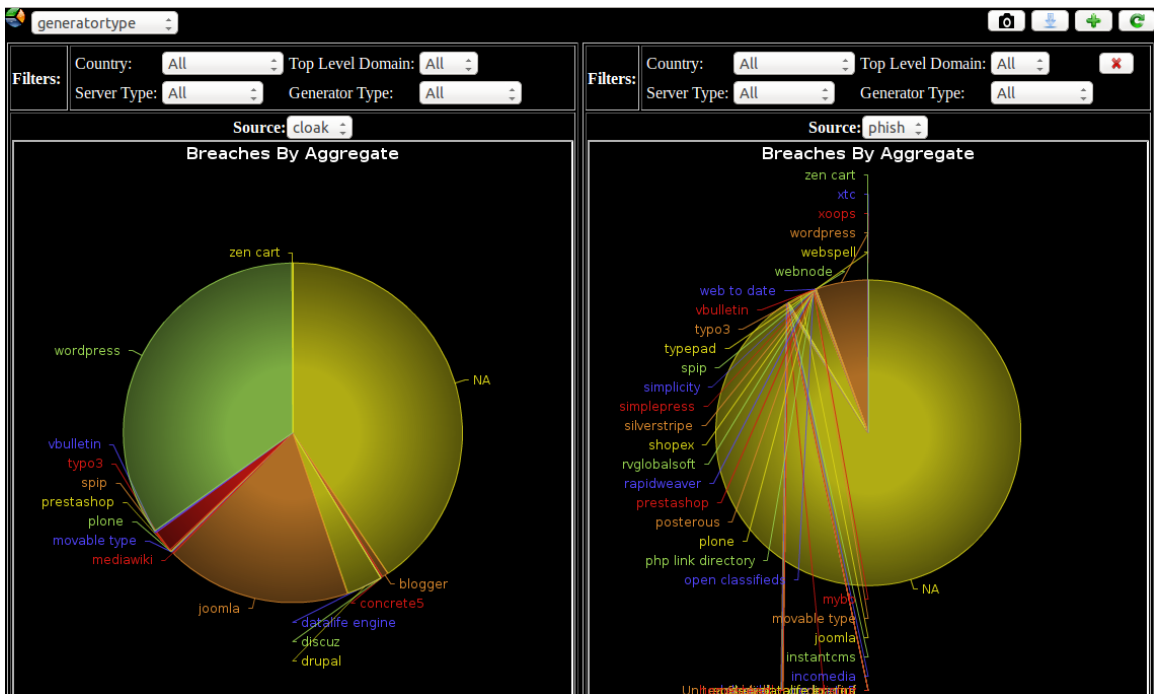


Figure 3.41. Portaling Feature

### 3.4.2. Attribute Filtering

Attribute filtering controls as depicted in Figure 3.42 are provided to allow the user the ability to manipulate the data flow into the visualization. A user can select “All” for each filter control effectively ignoring filtering or select a specific value. The filtering controls options can be tailored to any desired dataset by defining the options within the “specific.js” JavaScript file. This capability thus enhances the utility of visualizations by providing dynamic filtering.

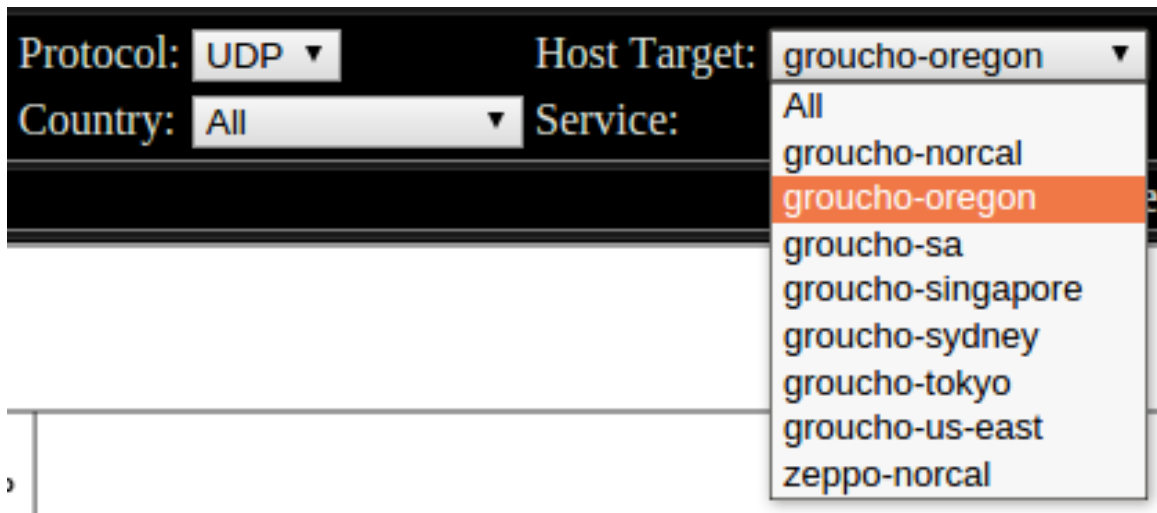


Figure 3.42. Attribute Filtering

### 3.4.3. Time-Based Filtering

Time-based filtering controls as depicted in Figure 3.43 are provided to allow the user the ability to manipulate the time-scope for the visualization. This applies to all chart types regardless of whether or not it is time-based in nature. If there is no time attribute for a dataset, a dummy attribute must be created (e.g. alter table tablename add time datetime)

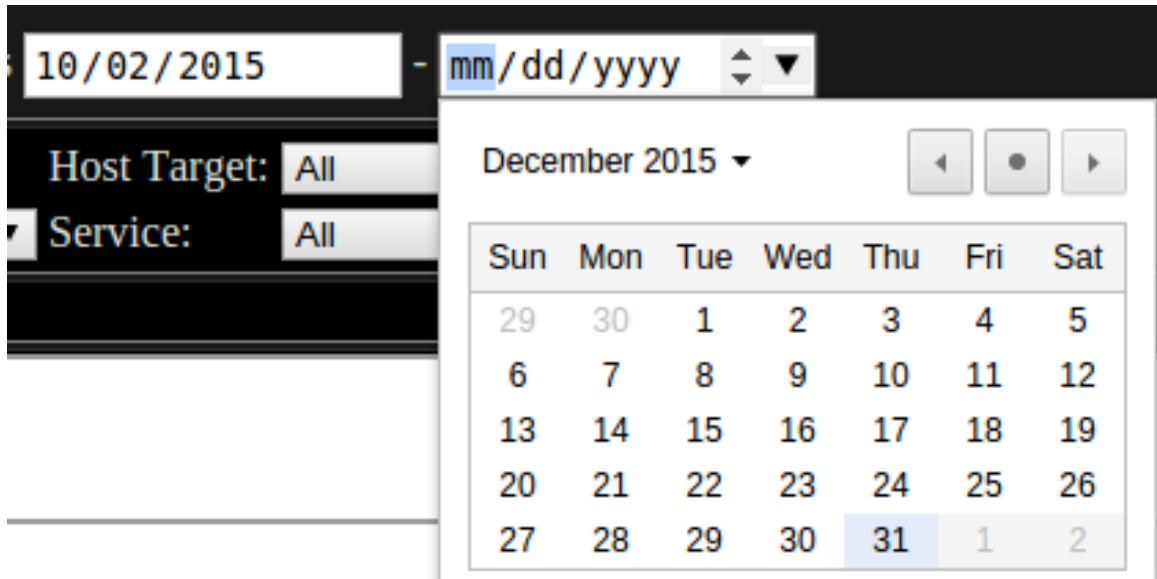


Figure 3.43. Time-Based Filtering

#### 3.4.4. Download Dataset

The ability to download a dataset is provided for data integrity and off-line analysis. Accessing the download button will return the selected chart's tabular data for a given dataset as either a csv file or zip package.

#### 3.4.5. Download Chart Data

The ability to download the data representing any given chart type is provided. The data is provided in txt format back to the user in a logical tabular space-delimited format. For example, if the PIE chart is selected, the relative percentages or counts of each pie slice are provided. This could be useful if one wants to see the fine-grain visualization data itself for distribution among stakeholders.

#### 3.4.6. Publishing

The ability to publish a particular web-portal is prototyped but not entirely implemented. This is provided through the `html2canvas` package [46] by Niklas von Hertzen. After the cross-site scripting domain problems are solved, this would allow the user to take precision screenshots of preformatted charts/tables along with controls. The intended artifact would be an image (png) provided for download back to the user which the user could then disseminate to others how he/she saw fit. Currently, the provided prototype mirrors the image back to the user at the bottom of the web page. As its in a nonfunctional state, we have descoped it from the framework. It can be added back into any given page by uncommenting the HTML input “`screenshotBtn`”.

### 3.5. Installation & Deployment

Appendix A describes installation & deployment instructions necessary to install this framework on a web-server. It also describes steps necessary to set up the client design machine.

### 3.6. Extending the Framework

Appendix B details instructions necessary to extend this framework by adding new datasets or visualizations. This section also goes into finer grain detail on how to use the RapidAnalytics (RA) tool. For example, it details the conversion process necessary to migrate a RapidMiner (RM) process to a RA web service.

### 3.7. Source-Code

Appendix C captures the framework HTML / JavaScript code discussed by this paper.

## Chapter 4

### BREACH CASE STUDY DESCRIPTION

In this chapter we introduce the breach dataset. We begin by presenting a background to the breach dataset (Section 4.1), including the origins and the intentions behind its compilation. The collection process as well as a description of the surrounding attributes is then detailed in Section 4.2. We then discuss aggregation of the data with other source data (Section 4.3) to expand the dataset analysis. Next, potential risk factors and analysis goals are presented in relation to the data (Section 4.4). Finally a brief overview of the analysis methodology is presented along with perceived limitations inherent to the data (Section 4.5 & 4.6).

#### 4.1. Background

For many years people had no recourse for knowing when their most private data had been disclosed. With no accountability, companies and non-profits alike would leak data with no legislation or requirements to disclose said incidents. However, beginning in 2002 with the passage of SR 1386 (California), breach disclosure laws have slowly come into being, eventually matriculating to most states. This has helped the public remain aware of entities mishandling their sensitive data providing much needed accountability. Breach disclosure laws have also had a positive impact on cybersecurity research by providing valuable data on security incidents.

Privacy Rights Clearinghouse (PRC) [16], a non-profit group dedicated to consumer information and advocacy, has led this effort. Their stated mission – to engage, educate and empower individuals to protect their privacy [14] – has empowered

consumers in becoming aware of these disclosures. Privacy Rights Clearinghouse provides consumer resources while raising awareness on the effects of technology on personal privacy. Services include: responding to specific privacy-related complaints from consumers, interceding on their behalf, and, when appropriate, referring them to the other organizations for further assistance. In addition to these resources, PRC engages on the political front by relaying consumer complaints to policy makers, industry representatives, and the media, including but not precluding participation in political commissions, legislative testimony, and task forces.

## 4.2. Dataset Description

PRC has collected privacy breach related data spanning all 50 states dating back to the year 2005. This data is compiled within a public breach database provided at [privacyrights.org](http://privacyrights.org). [15] The database consists of data that has been collected from multiple sources including: other breach databases, state disclosures that require mandatory data-breach reporting, as well as media. Incidents are included if they meet a threshold of nine or more affected individuals or if there is a compelling reason to include it within the dataset from a consumer protection standpoint. The dataset is by no means a complete listing of breaches, however, it should be indicative of frequency, size, and source.

The vast majority of the data is derived from the Open Security Foundation list-serve [35], who has pulled the data from government web sites, blog posts, and verifiable news stories. As of January 2010, the sources were expanded to include three additional sources: Databreaches.net [18], Personal Health Information (PHI) Privacy [36], and the National Association for Information Destruction (NAID) [34]. In particular Databreaches.net, one of the source inputs, has compiled a large volume of reports dating back to January 2009 when it was incepted as [pogowasright.org](http://pogowasright.org).

PHI Privacy is a site that compiles only medical data breaches. The majority of their records are sourced from the US Department of Health and Human Services medical data breach list [26]. Due to it being protected medical information, only a minimally sufficient amount of information is recorded. Lastly, the National Association for Information Destruction (NAID) the last large-scale organization providing input reports events of improper document destruction – a key source for disclosures. A figure depicting both the data sources as well as their associated traceability is included below in Figure 4.1.

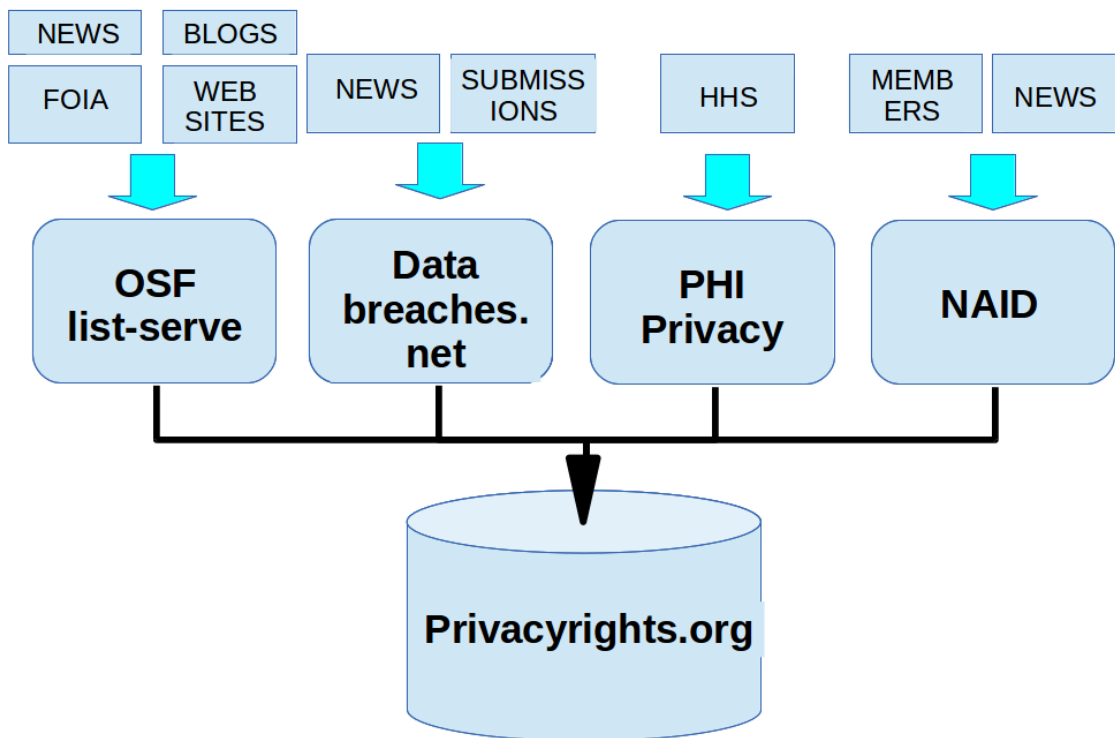


Figure 4.1. Breach Data Sources & Traceability

The Breach dataset comes as a downloadable csv file. Each row constitutes a breach incident. In turn, each incident can have one or more responsible or affected

parties. The attributes provided for each record are detailed below in Table 4.1. In addition to generic information such as the date, names, location and source of the breach disclosure, PRC provides the entity category, the number of records, and the breach type. The entity category refers to what type of entity incurred the breach – government, business, educational institute, etc. Additionally, the number of records of sensitive information breached is provided (if compiled from by the source). Lastly, the breach type is compiled, referring to the method that the breach occurred. A description of each brief type as provided by PRC [17] is referenced below in Figure 4.2. This attribute can range from traditional hacking breach (HACK), to a lost portable device (PORT), to human in the loop unintended disclosures (DISC).

### **4.3. Data Aggregation**

We further refined the dataset by adding more granularity as defined in Table 4.2. This involved examining each row where there was an incident and if multiple parties were involved further breaking the row into multiple records. In doing so we assigned a “MultiRow” field of “true”. We also created an “IsAddendum” field if it was an addendum to an initial record provided by a multi-row record. Our intent was to gain traceability into all parties affected without losing traceability to the original distinct record. Any statistical variance had by linking multiple entities to a breach occurrence, we postulate should be evenly distributed across the dataset. Additionally, we sought more granularity into the type of entity affected. In support of this interest, we created an “entity type” attribute by expanding whether the entity was a private or public business/educational institute. We also allowed for cooperatives for business entity types (if applicable). Additionally, if the company was a publicly traded company, we assigned a stock symbol for the company. The new attributes created by this effort are summarized in Table 4.2.



Table 4.1. Privacy Rights Breach Record Attributes

Attribute	Description
Date	Date made public
Name	Responsible/Affected Parties
Entity Category	Category of Entity for which the breach occurred. <b>BSF</b> : Bus - Financial & Ins Services <b>BSR</b> : Businesses Retail/Merchant <b>EDU</b> : Educational Institutions <b>GOV</b> : Government and Military <b>MED</b> : Healthcare - Medical Providers <b>NGO</b> : Nonprofit Organizations
Entity(s)	Entity(s) who-for the breach occurred
City	City that the breach occurred in
State	State that the breach occurred in
Breach Type	Generic category of breach type: <b>CARD</b> : Payment Card Fraud <b>DISC</b> : Unintended Disclosure <b>HACK</b> : Hacking or Malware <b>INSD</b> : Insider <b>PHYS</b> : Physical Loss <b>PORT</b> : Portable Device <b>STAT</b> : Stationary Device
Breach Description	Detailed Description of Breach
Breach Records	Quantity of Records Breached
Location	Location of breach
Source	Source that reported breach

After creation of these new fields, we merged the data with two datasets. The first dataset, a list of all publicly traded companies on the NASDAQ and NYSE, was merged with the original dataset by utilizing the newly created stock ticker symbol as a key. This provided additional elements, specifically for the publicly-traded companies, including: sector, industry, and market cap size. Market cap size was furthermore decomposed categorically into a fine-grain and coarse-grain representation

### Breach Types:

- **Unintended disclosure (DISC)** - Sensitive information posted publicly on a website, mishandled or sent to the wrong party via email, fax or mail.
- **Hacking or malware (HACK)** - Electronic entry by an outside party, malware and spyware.
- **Payment Card Fraud (CARD)** - Fraud involving debit and credit cards that is not accomplished via hacking. For example, skimming devices at point-of-service terminals.
- **Insider ( INSD)** - Someone with legitimate access intentionally breaches information - such as an employee or contractor.
- **Physical loss (PHYS)** - Lost, discarded or stolen non-electronic records, such as paper documents
- **Portable device (PORT)** - Lost, discarded or stolen laptop, PDA, smartphone, portable memory device, CD, hard drive, data tape, etc
- **Stationary device (STAT)** - Lost, discarded or stolen stationary electronic device such as a computer or server not designed for mobility.
- **Unknown or other (UNKN)**

Figure 4.2. Breach Type Definitions from PRC

as detailed in Table 4.3.

In addition, we merged the original dataset with population census data from the U.S. Census Bureau containing population data for all Incorporated Places [23]. In doing this we utilized the city/state provided in the original dataset as the merge key. After population was aggregated into the final dataset, we created a categorical repre-

sentation of city-size based off of logarithmic-based ranges. Census regions were then appended by cross-correlating the city/state attribute with the region assignments provided by the U.S. Census Bureau [24]. The attributes appended to the dataset as part of the merge process with the Census data are summarized below in Table 4.4.

The scripts for the import / merge process discussed above (using AWK and SQL) are made available at download location #9 in Section A.3.

#### 4.4. Analysis Goals

By examining the breach dataset, we hope to determine which attributes of a company have correlation to either breach vulnerability or breach exposure. This might include market share of the company, industry/sector, or region of the United States. For example, are there certain industries or sectors that are more susceptible

Table 4.2. Privacy Rights Breach Record Additions

Attribute	Description
Entity Type	Finer-grain type of Entity for which the breach occurred. <b>COOP</b> : Mutual or Cooperative <b>GOV</b> : Government entity <b>NON</b> : Org. w/ non-profit status <b>PUB</b> : Public Company <b>PRI</b> : Private Company <b>PUE</b> : Public Educational Institution <b>PRE</b> : Private Educational Institution
MultiRow	Whether this record represents one where-in multiple entities were involved
IsAddendum	Whether this record represents an addendum to a multi-entity record
Stock Exchange	For Public companies, the stock exchange, that the stock is traded on
Stock Ticker	For Public companies, the stock associated stock ticker symbol for the company.

Table 4.3. NASDAQ / NYSE Attributes

Attribute	Description
MarketCap	Market-capitalization - the total marketvalue of a company's outstanding shares.
CapSize Coarse	Coarse-grain Market-Cap Size <b>SMALL:</b> Below \$1 billion <b>MID:</b> \$1 billion - \$5 billion <b>LARGE:</b> Over \$5 billion
CapSize Fine	Fine-grain Market-Cap Size <b>MEGA:</b> Over \$200 billion <b>LARGE:</b> Over \$10 billion <b>MID:</b> \$2 billion\$10 billion <b>SMALL:</b> \$250 million\$2 billion <b>MICRO:</b> Below \$250 million <b>NANO:</b> Below \$50 million
Sector	Generic Sector category <b>BI:</b> Basic Industries <b>CD:</b> Consumer Durables <b>CND:</b> Consumer Non-Durables <b>CS:</b> Consumer Services <b>EN:</b> Energy <b>FI:</b> Finance <b>HC:</b> Health Care <b>MI:</b> Miscellaneous <b>PU:</b> Public Utility <b>TE:</b> Technology <b>TR:</b> Transportation
Industry	Fine-grain industry of sector

Table 4.4. Census Attributes

Attribute	Description
Population 2012	Estimated population for 2012 based on 2010 census data & historical trends
CitySize	<b>VSML:</b> Population under 1000 <b>SML:</b> 1000-9999 people <b>MED:</b> 10000 99999 people <b>LRG:</b> 100000 999999 people <b>VLRG:</b> Over 1000000 people
Region	The region where-in the breach occurred divided by census region. <b>ENC:</b> East North Central <b>ESC:</b> East South Central <b>MA:</b> Mid Atlantic <b>MTN:</b> Mountain <b>NE:</b> New England <b>PAC:</b> Pacific <b>SA:</b> South Atlantic <b>WNC:</b> West North Central <b>WSC:</b> West South Central

to certain breach types? Are there certain regions/states more susceptible to breach? Does city population have any bearing on breach exposure? Our intent in performing this analysis is not to make any substantive claims, but to demonstrate dynamic iterative analysis using the framework.

#### 4.5. Methodology

For this analysis we predominantly focus on the publicly traded companies. However, other entity types are examined. We arrange this analysis in case-control study format. Breach data is compared in relation to controlled data. The control dataset is the full list of NASDAQ and NYSE publicly traded companies. It serves as the control group and a logical comparison alternative for percentages and counts. In addition to aggregation, odds ratio plots can be used to determine statistical measures

of merit.

We also perform a breach-only analysis, wherein the breach is analyzed on its own. A box plot can be used for median analysis of population and number of records breached. Two-way categorical plots can provide insight into whether a combination of two attributes yields higher risk. Additionally, a time-based attribute means we can set up time plots to visualize the data in yet another dimension.

#### **4.6. Data Limitations**

This breach dataset represents a subset of all breaches that have occurred in the time-frame since 2005. Having said this, its quite likely that some companies are more likely to report breaches than others. This can be confirmed by looking at the data on an individual company basis. Despite this discrepancy occurring, it is likely applied uniformly throughout the dataset. We postulate the tendency of some companies to report more than others should have minimal impact on the results if a large enough sample-size is chosen. Considering there are 3,792 records in the sample-set, the effect should be minimized. However, any hopes to analyze the data from an individual company perspective (are there frequent offenders?) are likely futile due to this inconsistency.

Noting another limitation to the data, breach disclosures are reported in only 46 states, with legislation in some states being more strict than others. While media disclosures are still represented in the data for all 50 states, some states and subsequently some regions are likely over-represented in the data due to the presence of stricter breach legislation. We attempted to gauge the relative effect of this on the dataset by plotting breach legislation across the United States and assessing if there were any regions more heavily represented than others. The map below (Figure 4.3) provided by Janco Associates [5] attempts to visualize the distribution of

breach disclosure laws across the United States. It can be seen in this map that the absence of state disclosure laws is evenly distributed across four regions with one occurring in both East North Central, East South Central, Mountain, and South Atlantic respectively. Additionally, stricter legislation appears evenly dispersed. Thus the cumulative effects of uneven geographical breach legislation are very likely quite marginal. If further solvency of the data is required, a fine-grain description of each state's associated breach legislation can be garnered through the IT Governance website [22] who provides state disclosure laws on a state-by-state basis.

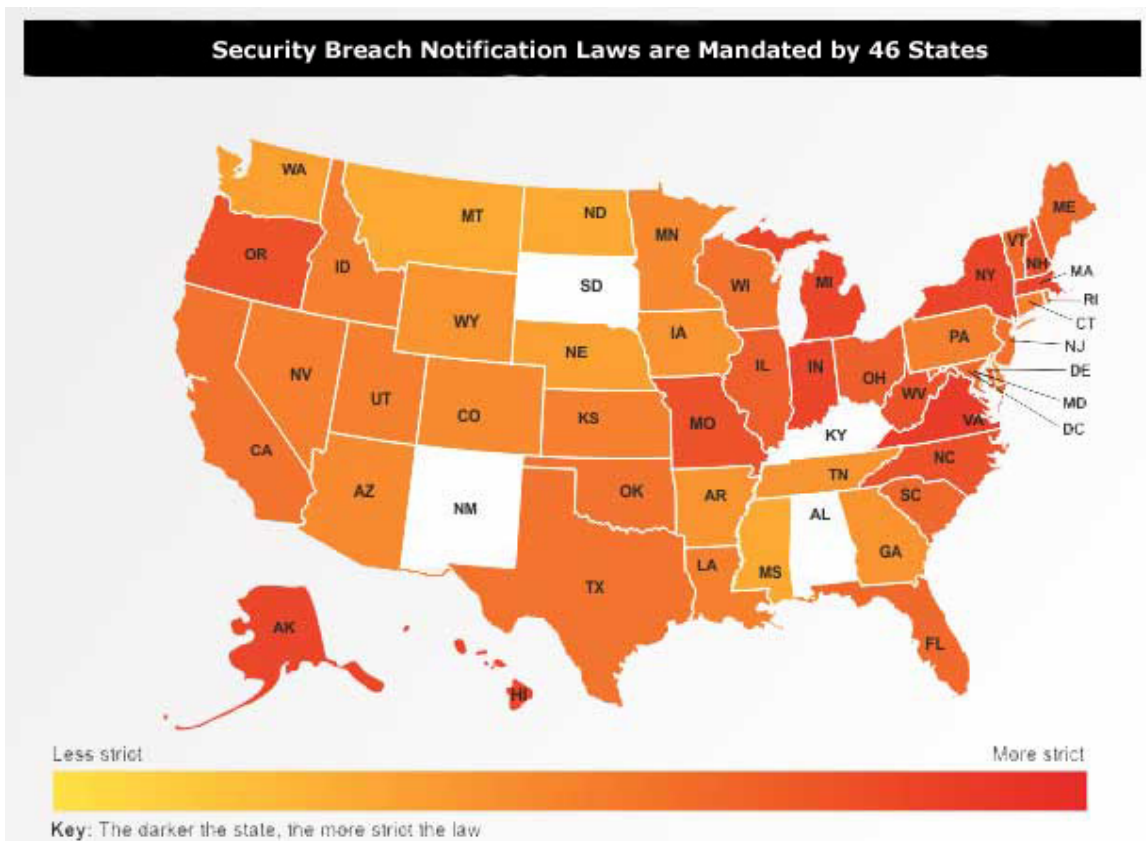


Figure 4.3. Breach Legislation Map

## Chapter 5

### ANALYSIS OF BREACH DATASET

After the data was compiled and migrated to the database, we performed an analysis of the breach dataset. In doing so, we demonstrate both the iterative process and utility provided by the framework. Indeed, analysis could be performed any number of ways. However, by highlighting the representative steps involved, we seek to differentiate the framework by validating both its dynamic reusable nature and streamlined iterative process. The aforementioned analysis is decomposed into two distinct strategies: case study comparative analysis (Section 5.1) as well as looking at the breach data by itself (Section 5.2). We then supplement with both a geospatial analysis (Section 5.3) and a time-based analysis (Section 5.4). Finally we conclude by summarizing predictors for webserver compromise, and ponder future possibilities regarding this dataset. (Section 5.5 & 5.6)

#### **5.1. Case-Control Study Comparative Analysis**

A case-control study comparative analysis seeks to compare two sets, typically a control group and a experiment or “treatment” group to explain differences between the two sample sets. As discussed in Section 4.4, we examined the breach data (treatment) from PRC in relation to the set of all NASDAQ and NYSE publicly traded companies (control a.k.a. clean).



### 5.1.1. Proportional Analysis

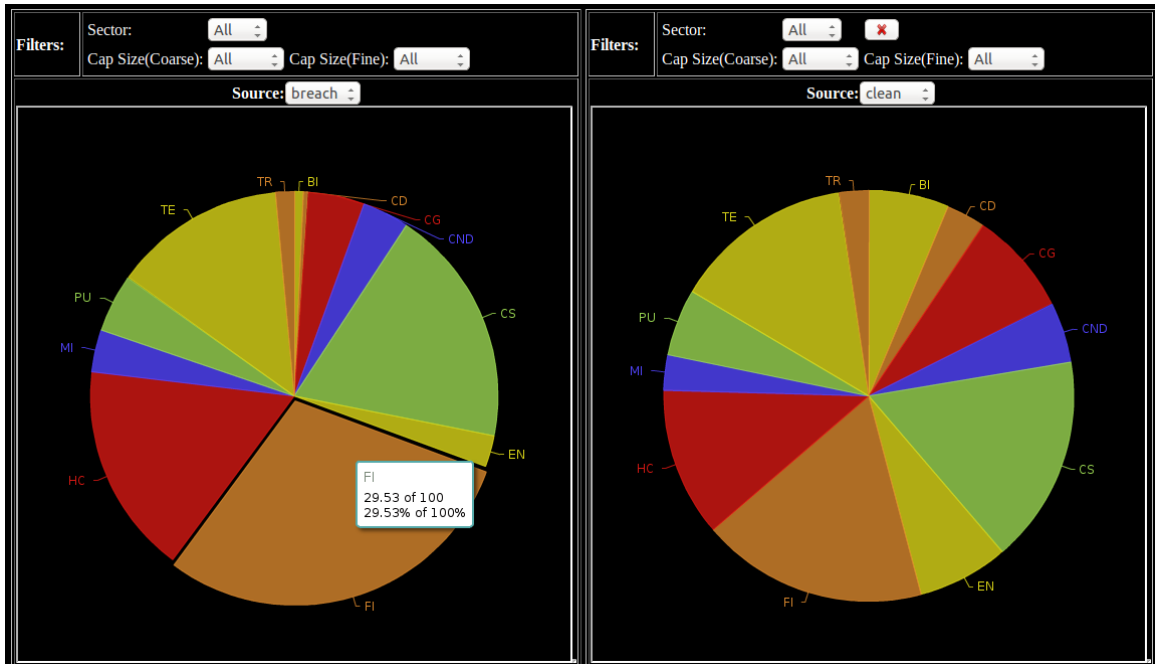


Figure 5.1. Breach vs Clean (By Sector)

To perform a comparative analysis and begin the iterative process of looking for relevant results, we first create a split portal view with the breach dataset on the left and the clean dataset on the right. By utilizing an “Aggregate Pie Plot”, we can quickly scan the datasets for relative differences in percentages. For example, we find that when analyzing by sector in Figure 5.1, that the Financial sector and Health Care sectors are over-represented in the breach set (30% v.s. 18% and 17% v.s. 12% respectively). The table summarizes the results (Table 5.1). These results leave more questions than answers. Are criminals targeting Financial Institutions for personal gain? Does the Health Care industry have stricter disclosure laws which cause it to be over-represented?

Table 5.1. Breach vs Clean (By Sector)

<b>Sector</b>	<b>Clean</b>	<b>Breach</b>
<b>Finance (FI)</b>	<b>17.84%</b>	<b>29.53%</b>
<b>Health Care (HC)</b>	<b>11.76%</b>	<b>16.75%</b>
Basic Industries (BI)	6.33%	0.74%
Energy (EN)	7.16%	2.48%
Consumer Services (CS)	16.34%	18.86%
Consumer Non-Durables (CND)	4.76%	3.72%
Consumer Goods (CG)	8.17%	4.47%
Consumer Durables (CD)	3.10%	0.37%
Transportation (TR)	2.33%	1.49%
Technology (TE)	14.14%	13.65%
Public Utilities (PU)	5.27%	4.59%
Miscellaneous (MI)	2.81%	3.35%

Table 5.2. Breach vs Clean (By Financial Industry)

<b>Industry</b>	<b>Clean</b>	<b>Breach</b>
<b>Major Banks</b>	<b>38.2%</b>	<b>51.7%</b>
<b>Life Insurance</b>	<b>3.5%</b>	<b>9.2%</b>
Investment Managers	3.9%	2.9%
Investment Bankers	6.2%	6.8%
Finance: Consumer Services	4.8%	6.7%
Diversified Financial Services	0.4%	0.4%
Diversified Com Services	0.5%	0.8%
Commercial Banks	4.2%	4.6%
Business Services	2.9%	0.4%
Banks	3.8%	0.4%
Accident & Health Insurance	1.3%	1.3%
Specialty Insurance	2.1%	5.9%
<b>Saving Institutions</b>	<b>13.0%</b>	<b>2.5%</b>
<b>Real Estate</b>	<b>4.7%</b>	<b>0.4%</b>
Property-Casualty Insurance	9.4%	5.9%

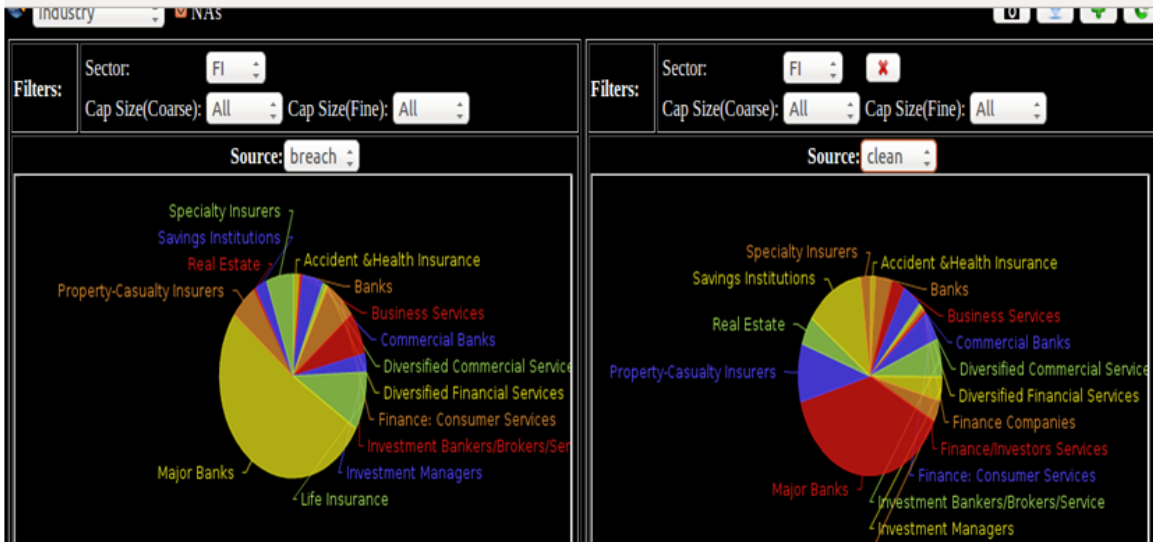


Figure 5.2. Breach vs Clean (Financial Sector By Industry)

Table 5.3. Breach vs Clean (By Fine Cap Size)

	Clean	Breach
<b>Nano</b>	8.89%	1.11%
<b>Micro</b>	25.66%	3.32%
<b>Small</b>	34.37%	9.72%
<b>Mid</b>	19.91%	21.65%
<b>Large</b>	10.84%	55.97%
<b>Mega</b>	0.33%	8.24%

Based on the previous result, we then set the filter for both portals to the “Financial Sector (FI)” and delve into the data further by looking at the relative breakdown by industry (Figure 5.2). Here we find that “Major Banks and Life Insurance” companies make up a larger share of breached companies than what is to be expected from the control group (Table 5.2). Both of these are common targets for criminals seeking monetary gain.

Additionally, we can quickly assess whether market capitalization has any affect on the breach susceptibility. Figure 5.3 reveals that large companies experience more



Figure 5.3. Breach vs Clean (By Fine Cap Size)

breaches (large  $\geq 64\%$ ) than smaller companies in relation to the control group. (large  $\leq 11\%$ ). Table 5.3 summarizes these findings. The discovery that larger companies have a higher breach percentage begs some questions. Are larger companies more likely to comply with breach disclosure laws? Or are they the target for more criminal activity?

### 5.1.2. Odds Ratio Analysis

To further investigate the results of the previous section, we use the “Odds Ratio Plot”. This is an efficient way to determine the odds of an outcome occurring (breach) given a particular exposure in relation to the odds of the outcome occurring in absence of that exposure (clean). The results are considered statistically significant at the associated confidence if the the confidence interval (whiskers) fail to intersect other bars’ relative ranges.

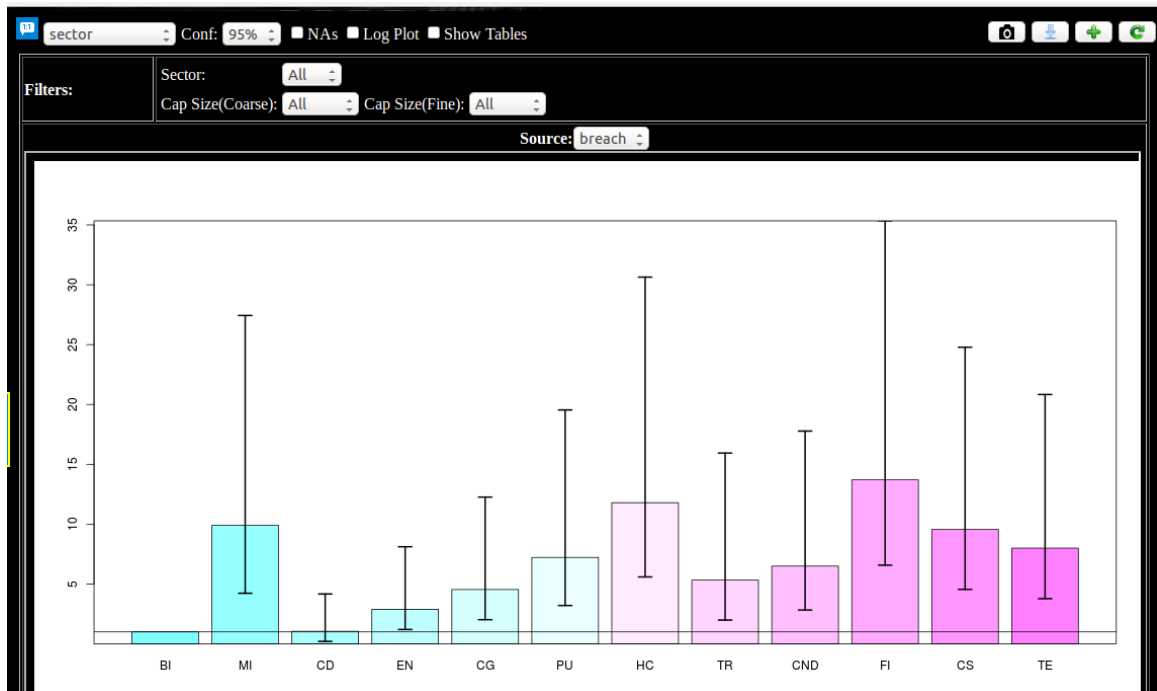


Figure 5.4. Odds Ratio By Sector

Table 5.4. Odds Ratio By Sector

sector	estimate	lower	upper
<b>Basic Industries (BI)</b>	<b>1</b>	<b>NA</b>	<b>NA</b>
<b>Miscellaneous (MI)</b>	<b>9.90</b>	<b>4.23</b>	<b>27.44</b>
Consumer Durables (CD)	1.05	0.21	4.17
Energy (EN)	2.89	1.20	8.11
Consumer Goods (CG)	4.54	2.02	12.26
Public Utilities (PU)	7.22	3.21	19.54
<b>Health Care (HC)</b>	<b>11.79</b>	<b>5.59</b>	<b>30.64</b>
Transportation (TR)	5.33	1.99	15.94
Consumer Non-Durables (CND)	6.50	2.83	17.78
<b>Finance (FI)</b>	<b>13.70</b>	<b>6.57</b>	<b>35.34</b>
Consumer Services (CS)	9.56	4.55	24.78
Technology (TE)	8.00	3.78	20.83

Table 5.5. Odds Ratio By CapSize

capsize	estimate	lower	upper
NANO	1	NA	NA
MEGA	191.60	86.01	481.00
MICRO	1.03	0.50	2.35
MID	8.58	4.61	18.31
LARGE	40.67	22.06	86.18
SMALL	2.23	1.17	4.84

Table 5.6. Odds Ratio By Industry (Financial)

Industry	estimate	lower	upper
Diversified Financial,Services	1	NA	NA
Real Estate	0.09	0.00	4.19
Banks	0.11	0.00	5.16
Savings Institutions	0.16	0.02	5.34
Diversified Commercial Services	1.39	0.07	58.21
Commercial Banks	0.92	0.09	28.60
Business Services	0.14	0.00	6.73
Finance: Consumer,Services	1.16	0.12	35.22
Accident &Health,Insurance	0.86	0.06	31.22
Property-Casualty,Insurers	0.52	0.05	15.75
Specialty Insurers	2.23	0.23	69.38
Life Insurance	2.14	0.23	64.69
Investment Managers	0.63	0.06	20.26
Investment Bankers/Brokers/Service	0.90	0.09	27.32
Major,Banks	1.11	0.12	32.20

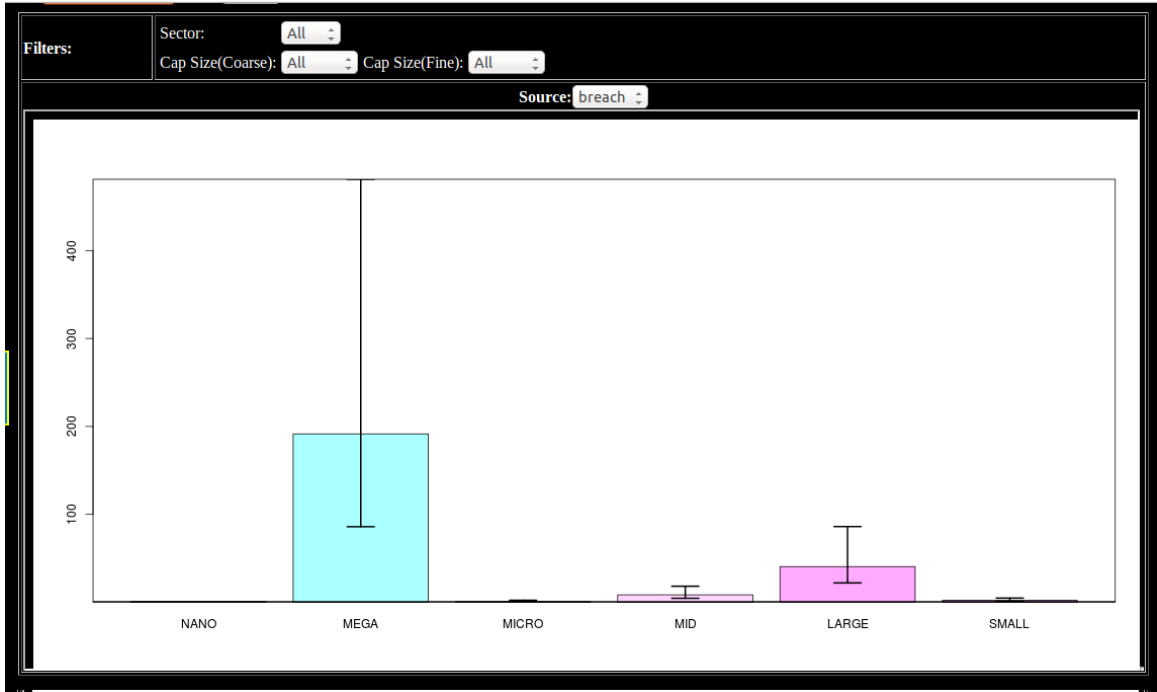


Figure 5.5. Odds Ratio By Sector

With this in mind, we can quickly assess whether the findings in the previous section are statistically significant. In the first plot (Figure 5.4), we assess sector odds ratios by choosing the applicable aggregate as well as the desired confidence level (95%). We then look for overlap of the confidence intervals (whiskers). Again, if the upper or lower confidence bars intersect the plot, the value is statistically relevant in relation to the other value. The plot selects ‘Basic Industries’ as the baseline value (1.0). We then observe ‘Finance’ (FI), ‘Health Care’ (HC), and ‘Miscellaneous’ (MC) sectors as having the highest odds ratios and statistical significance difference from the chosen baseline value ‘Basic Industries’ (BI). All three are also significant in relation to ‘Consumer Durables’ (CD). Table 5.4 summarizes the ratio values as well as their confidence intervals.

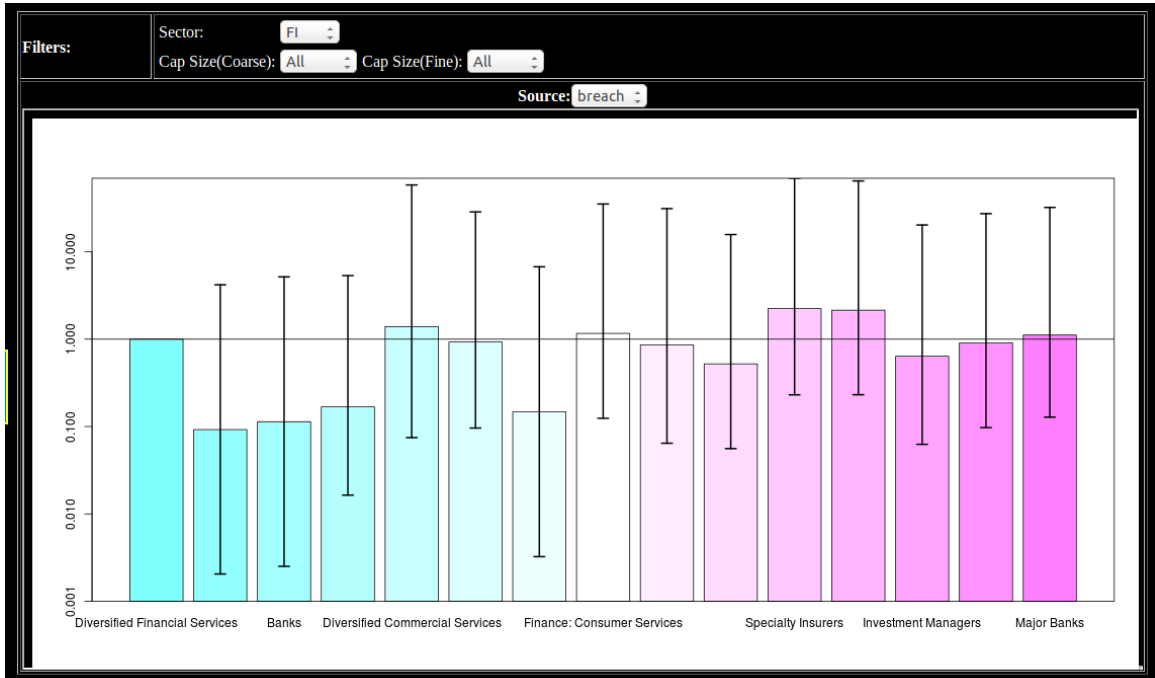


Figure 5.6. Odds Ratio By Industry (Financial)

We next plot the Cap Size using the odds ratio plot. (see Figure 5.5 Large & Mega size companies are over-represented (40.67 and 191.60 respectively). There are no intersection points of these ranges with any of the other ranges including the baseline value NANO. Table 5.5 summarizes these ratios and their associated confidence intervals.

We then attempt to hone in on data within the Financial Industry by setting the filter options to only display the Financial (FI) sector. We immediately notice extreme variation in the confidence levels. Switching to the log plot within the framework smooths out the variation (See Figure 5.6). Even then, we find too much variation within the confidence intervals (bars overlap significantly). The odds ratio ranges are provided in Table 5.6 for the Financial Sector odds ratio calculation. Switching to the Health Care sector, we find similar results (similar overlap). It appears as if there



are insufficient samples to arrive at any statistical significance.

## 5.2. Breach-Only Analysis

We then demonstrate analysis using only the breach dataset. The dataset has more attributes than the required intersection of attributes necessary for a comparative analysis, and can provide additional factors for examination. In doing so we will utilize visualizations that provide meaningful statistical relevance in relation to a single table or dataset (box-plot and mosaic plot).

### 5.2.1. Two-Way Categorical Analysis

Chi-squared statistical relevance in relation to two categorical variables can be analyzed through the “Mosaic Plot”. To be fully inclusive, we plotted all combinations of variables determined to be significant. We observe statistically relevant values within the data by observing the p-value depicted in Table 5.7 in conjunction with the test statistic ( $\chi^2$ ) (not depicted).

As a sanity-check we first plot the attributes ‘Breach Type’ vs ‘Cap-Size’ with the “Mosaic Plot” (Figure 5.7). This plot depicts sample size and statistical variance between breach types and company cap-sizes. The result correlates well with chi-squared test, depicting no statistical significance (absence of colored squares). We can conclude that regardless of ‘Cap-Size’, the types of breach that seem to occur appears to be uniformly distributed.

Subsequently, we plot ‘Cap-Size’ (Coarse) vs ‘Region’ as well as ‘Breach Type’ vs ‘Entity Type’, & ‘City Size’ vs ‘Cap-Size’ and see little statistical correlation. We must accept the null hypothesis that neither of these variables taken in conjunction with one other, have much statistical significance.

Table 5.7. CHISQ Test Results

Comparison Attributes	P-value
Breach Type vs Cap Size Coarse	0.2716
Breach Type vs Cap Size Fine	0.7626
Breach Type vs Entity Type	0.005742
Breach Type vs Region	2.12E-008
Cap Size Coarse vs Region	0.03723
Cap Size Coarse vs Sector	8.83E-010
Cap Size Fine vs Region	0.00583
Cap Size Fine vs Sector	<2.2e-16
Entity Category vs Region	6.08E-014
Entity Type vs Region	<2.2e-16
Entity Type vs Region	<2.2e-16
City Size vs Region	<2.2e-16
City Size vs Breach Type	8.57E-006
City Size vs Entity Category	<2.2e-16
City Size vs Entity Type	<2.2e-16

Note that we used the framework to select the residual method to create the plot (e.g. “Pearson’s Chi-Squared”). Different residual methods would create slightly different graphs depending on the threshold for their inclusion into the plot. The blue indicates statistically significant over-representation, while the red indicates under-representation. Deeper colors imply stronger significance. However, altering the residual method has little affect on these plots.

Many of the other plots do have significance. In comparing ‘Cap-Size’ to ‘Sector’ (Figure 5.8), we find that there are certain combinations that have higher breach susceptibility. Again this raises more questions than it answers. For example, are small technology companies a target for breaches? Is the technology / increased competition in the sector fueling breaches? Do these types of companies handle more of this type of information (CC#S, SSN, etc?). However companies that belong within certain swatches of sector/cap-size might consider investing more interest into

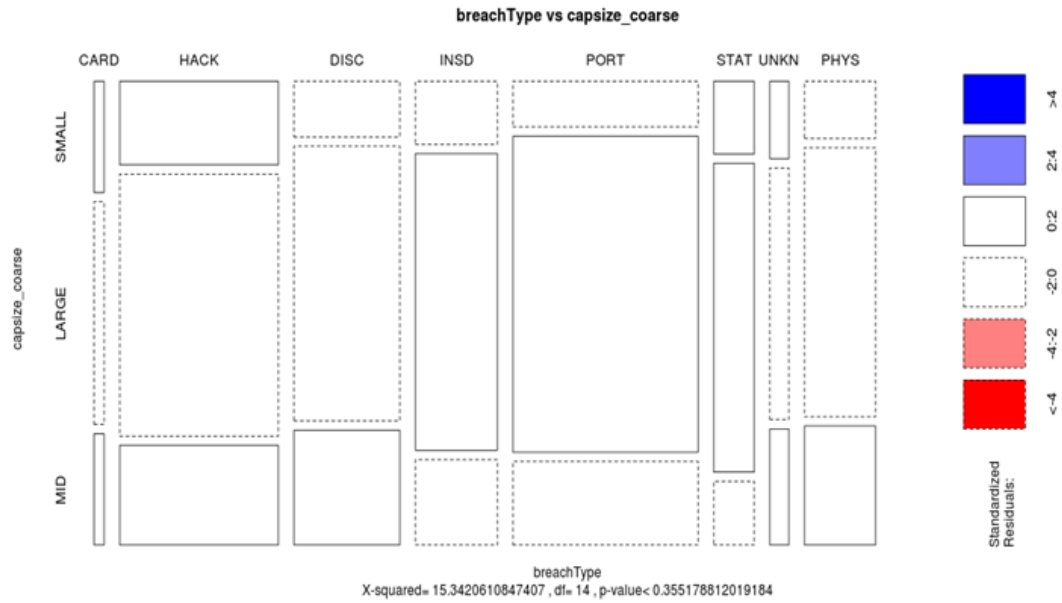
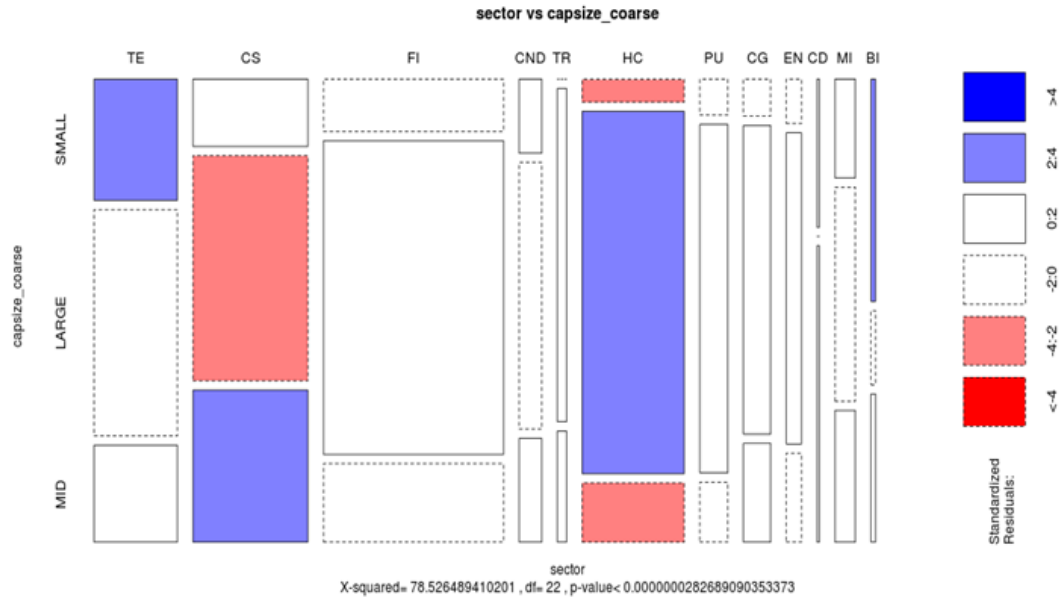


Figure 5.7. Breach Type Vs CapSize (Coarse)

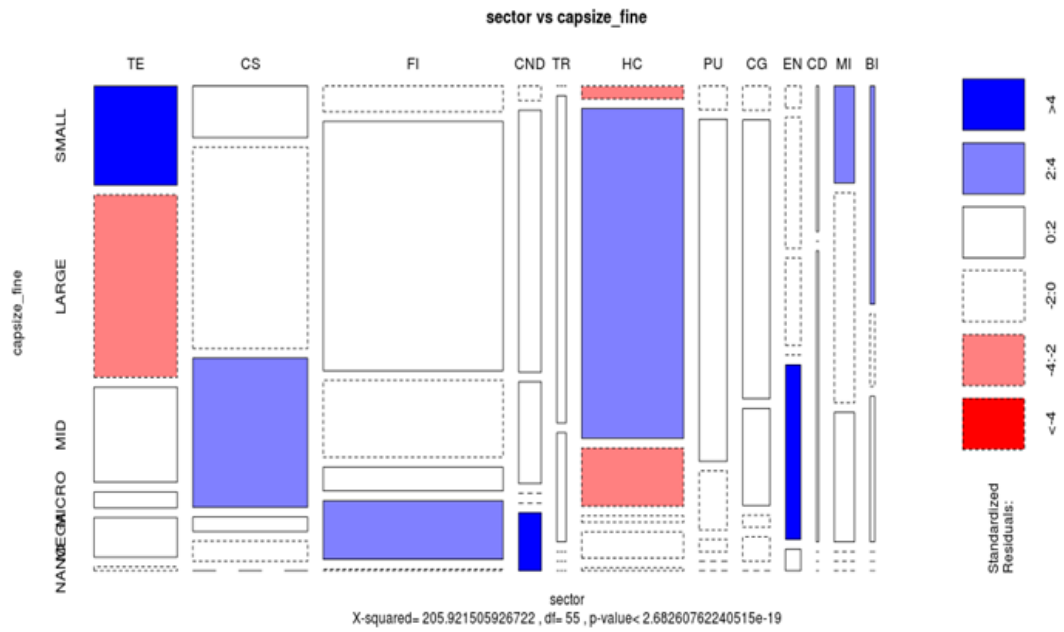
determining if they are at risk.

Further refining this analysis, we can break up the cap-size more discretely. This would work well if the sample-size is sufficiently large enough to justify it. After reconfiguring this plot, we can still see small technology companies as over-representation, but now mega energy and nano-sized consumer durables firms are over-represented. This seems to indicate that perhaps our critical energy infrastructure is being targeted. However, again, there is no conclusive proof as to why this is the case, just that over-representation exists.



Over-represented	Under-represented
<i>SML - Tech Companies</i>	LRG - Consumer Services
LRG - Health Care	MID - Health Care
MID - Consumer Services	Small - Health Care
SML - Basic Industries	

Figure 5.8. CapSize(Coarse) vs Sector



Over-represented	Under-represented
<i>SML - Tech Companies</i>	LRG - Technology
<i>MEGA - Energy</i>	LRG Consumer Services
<i>NANO - Consumer Durables</i>	MID - Health Care
LRG - Health Care	SML - Health Care
MEGA - Finance	
SML - Basic Industries	

Figure 5.9. CapSize(Fine) vs Sector

### 5.3. Geographic Analysis

We postulate there are certain areas that are targeted more based off of our chi-squared testing. This could be due to breach type or other categorical variables. So, we analyze the geographic aspect of the data-set. Does where the entity reside influence probability of breach? Our framework provides a geographic plot, however, it is configured to work with numeric odds ratios (of which we don't have regarding regions). Instead, we will use mosaic plots as a stand-in in this section and utilize the geographic plot for the second dataset.

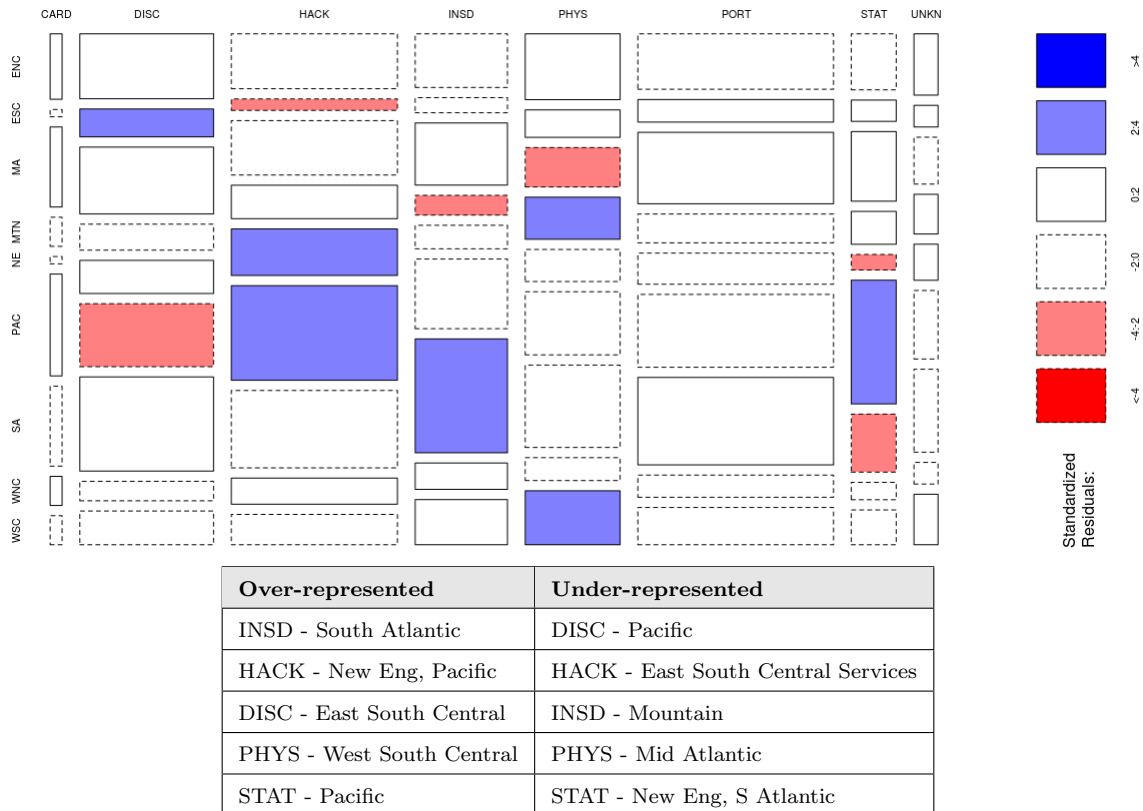
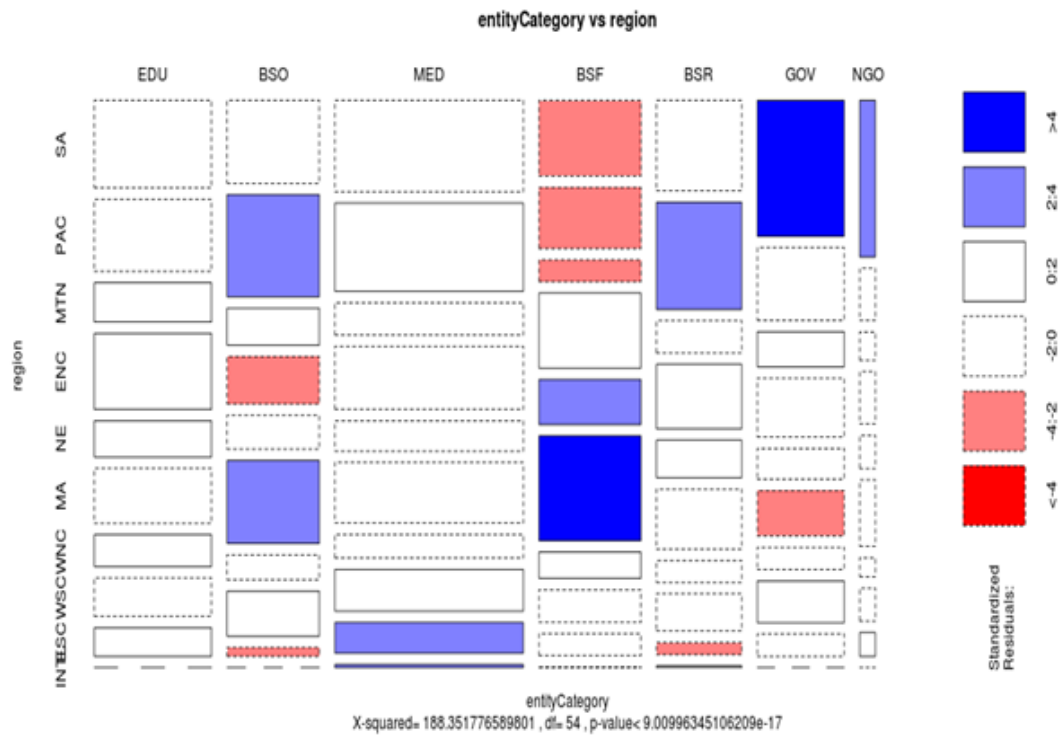


Figure 5.10. Breach Type vs Region

In the first plot (Figure 5.10) we see how regions are affected by breach-types. Of interest here is that the technology corridors (New England/Pacific) are more



Over-represented	Under-represented
<b>GOV-South Atlantic</b>	GOV-Mid Atlantic
<b>BSF-Mid Atlantic</b>	BSF-South Atlantic
BSF-North East	BSF-Pacific
NGO-South Atlantic	BSF-Mountain
BSO-Pacific	BSO-East North Central
BSO-Mid Atlantic	BSO-East South Central
BSR-Pacific	
MED-East South Central	

Figure 5.11. Entity Category vs Region

susceptible to HACK breach types. The remaining combinations are depicted within the plot, however, there doesn't seem to be any particular obvious rationale associated with it.

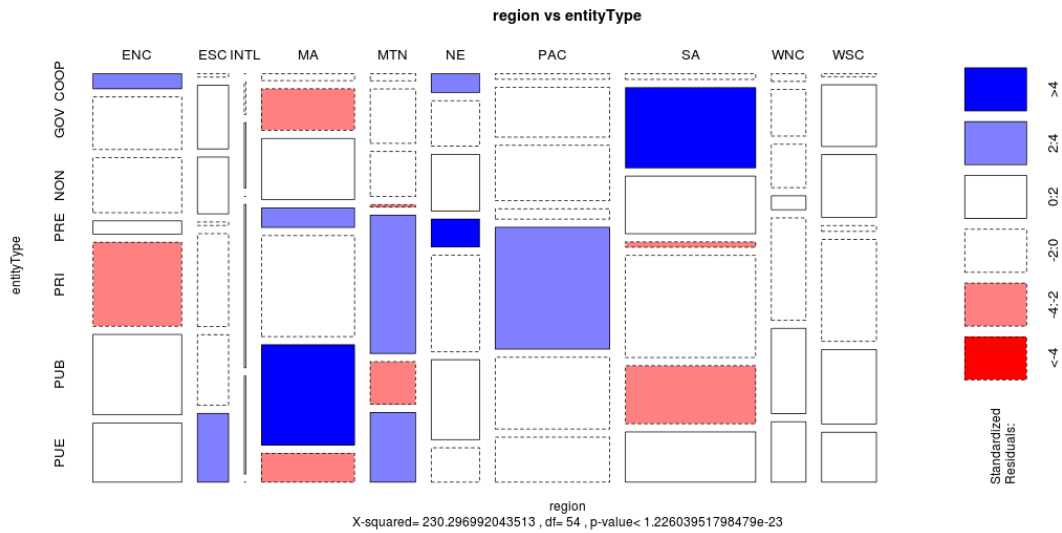


Figure 5.12. Entity Type vs Region

We then turn towards what types of organizations are affected in census regions. By comparing the PrivacyRights provided ‘Entity Category’ (Figure 5.11) and subsequently the more refined ‘Entity Type’ (Figure 5.12) attribute, we find there is statistical correlation between the type of entity and the census region. Government entities in the South Atlantic Region, as well as public companies in Mid Atlantic are over-represented in the sample-set.

‘City Size’ seems to also have a correlation with ‘Breach Type’. In Figure 5.13, it appears to correlate with different breach types. In particular, disclosures seem to happen more in larger cities, and Insider/Portable Device Breaches seem to happen more frequently with very large cities. This assumes an even distribution of breaches of companies across cities. Because its likely that companies tend to congregate in larger cities, we can reject this graph as irrelevant.



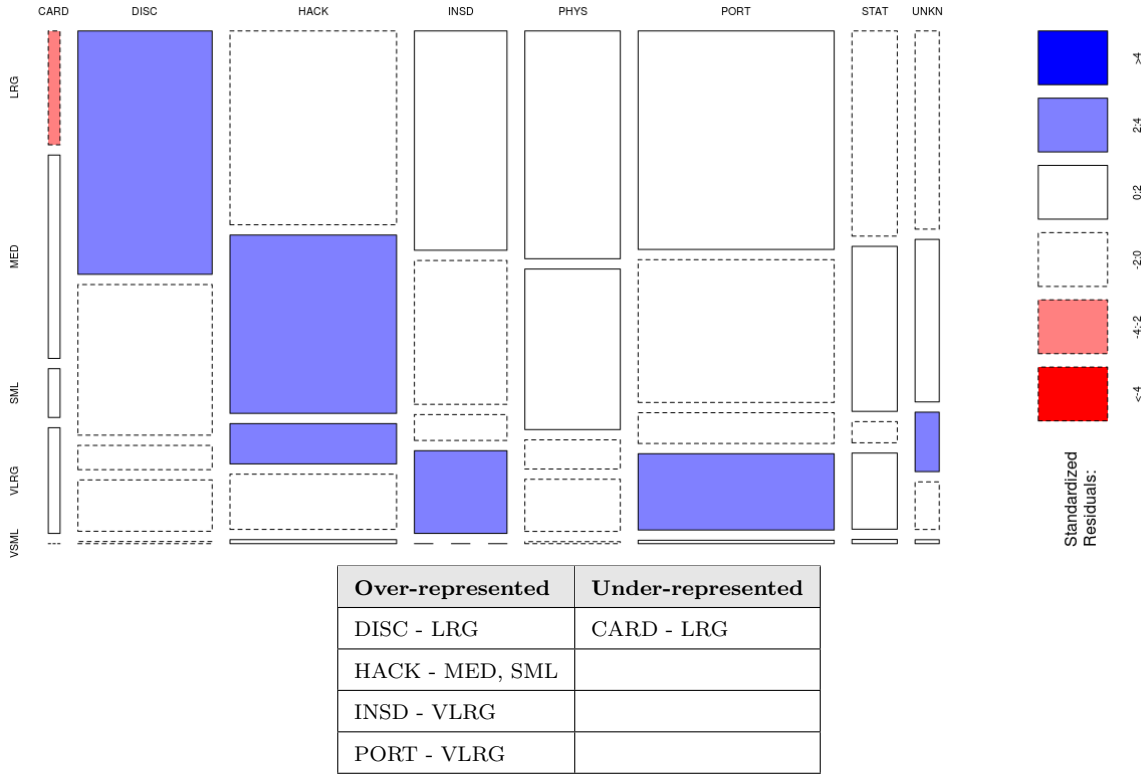


Figure 5.13. City Size vs Breach Type

### 5.3.1. Median Analysis

As discussed in Section 3.3.6, the “Box Plot” visualization is useful for graphically depicting groups of categorical data through their quartiles. The plot depicts the relative median, 25%, and 75% regions. The whiskers extending vertically from the box indicate variability outside the quartiles. With the breach data we have two numerical attributes available: ‘Breach Record Count’ and ‘Population Size 2012’.

Examining ‘Breach Record Count’ and opting for a log plot representation, we find that the categorical variable with the strongest correlation is ‘Entity Type’ as depicted in Figure 5.14. However, there isn’t much quantifiable difference between them.

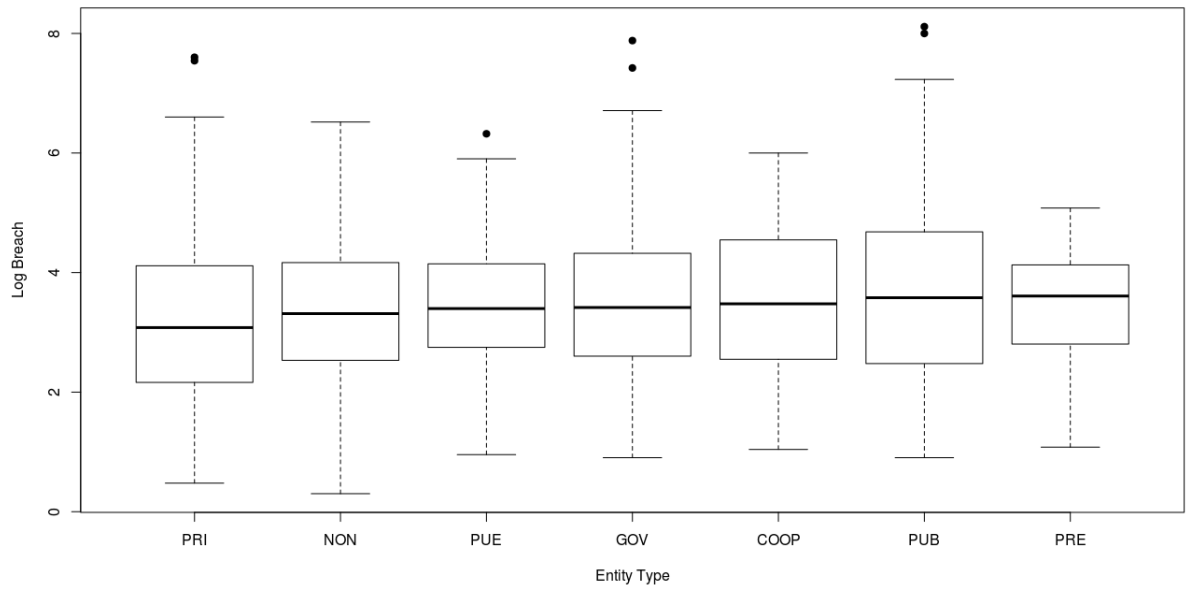


Figure 5.14. Median Log(Breach) By Entity Type

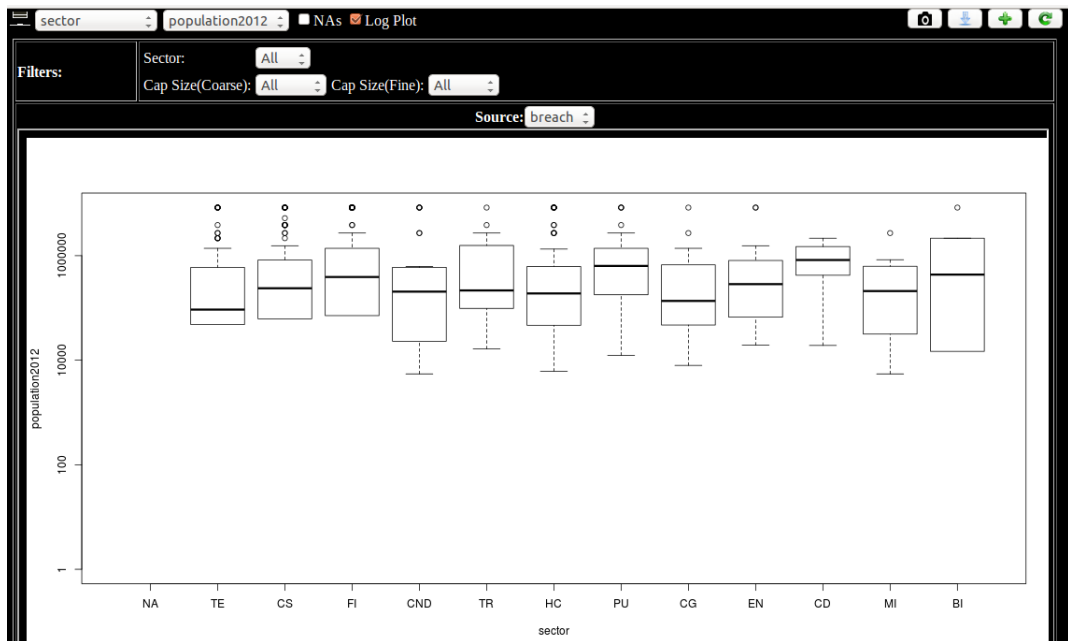


Figure 5.15. Median Log(Population) By Sector

Additionally, we analyze population city-size with respect to each categorical attribute. By cycling through each x-axis possibility, we find that ‘Sector’ has the most correlation. 5.15. However, this may not be meaningful, as its possible that firms of a given sector may congregate around larger or smaller cities respectively.

#### 5.4. Time-Based Analysis

Using the “Time Aggregate Bar Plot”, we turn towards a new dimension of the data, time. Using this plot we are able to assess how an aggregate attribute changes over time. Note that this plot does not necessarily indicate statistical relevance of its results. It merely is a facet for detecting time trends within the data with respect to a variable. To assess statistical relevance, the time-based odds ratio can be used. (not utilized as we do not have times for the control group)

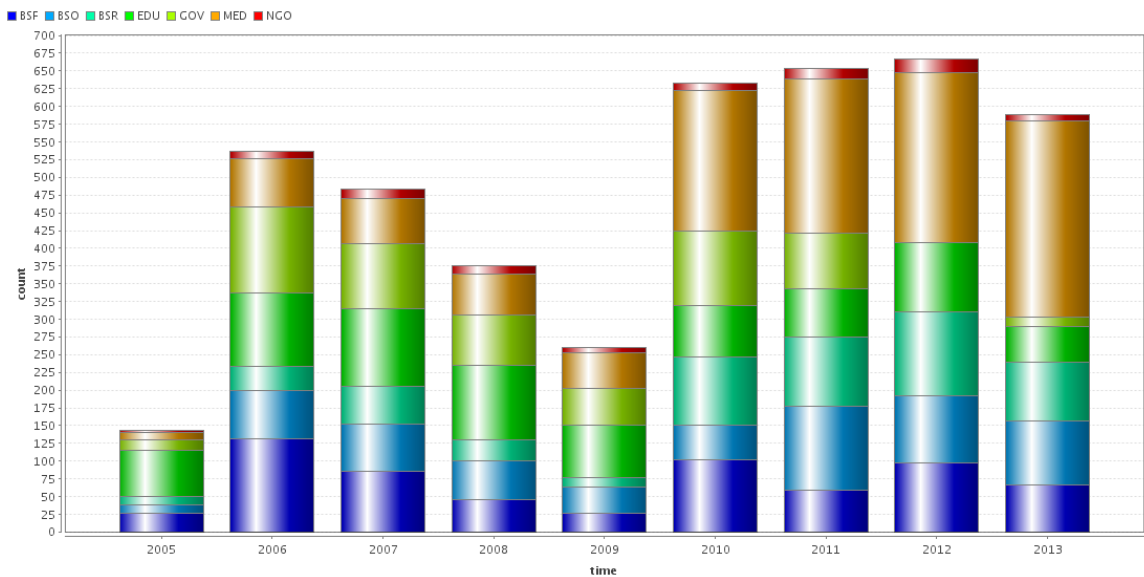


Figure 5.16. Breached Entities By Year

In the first plot (Figure 5.16), we can see a trend toward targeting medical entities post 2009. Was this due to a change in the nature of their reporting mechanisms, or

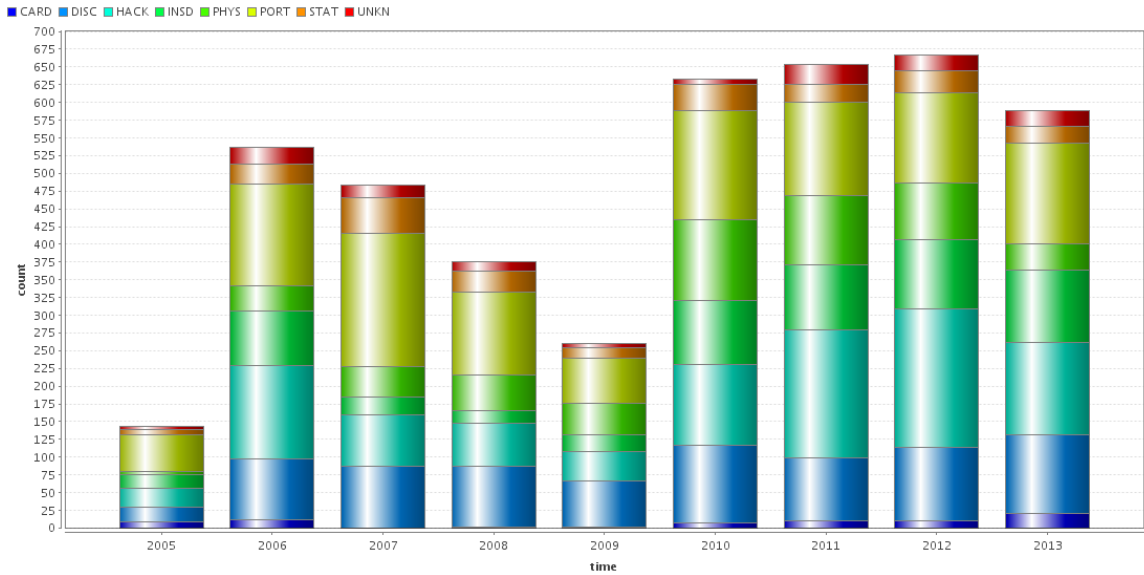


Figure 5.17. Breach Type By Year

is the medical industry a more lucrative industry for breaches?

In the second plot (Figure 5.17) we observe the breach type per year (though varying in overall record count) is stable percentage-wise. There does not appear to be a trend toward targeting different breach methods (at least on a year interval).

In the third plot (Figure 5.18), we see the financial industry as the largest breached sector, dipping in relative percentage post-recession (2009) as well as the Technology sector encompassing a larger relative percentage of breaches in later years. Was it less lucrative post-recession to target these types of firms or did the quantity of these firms decrease?

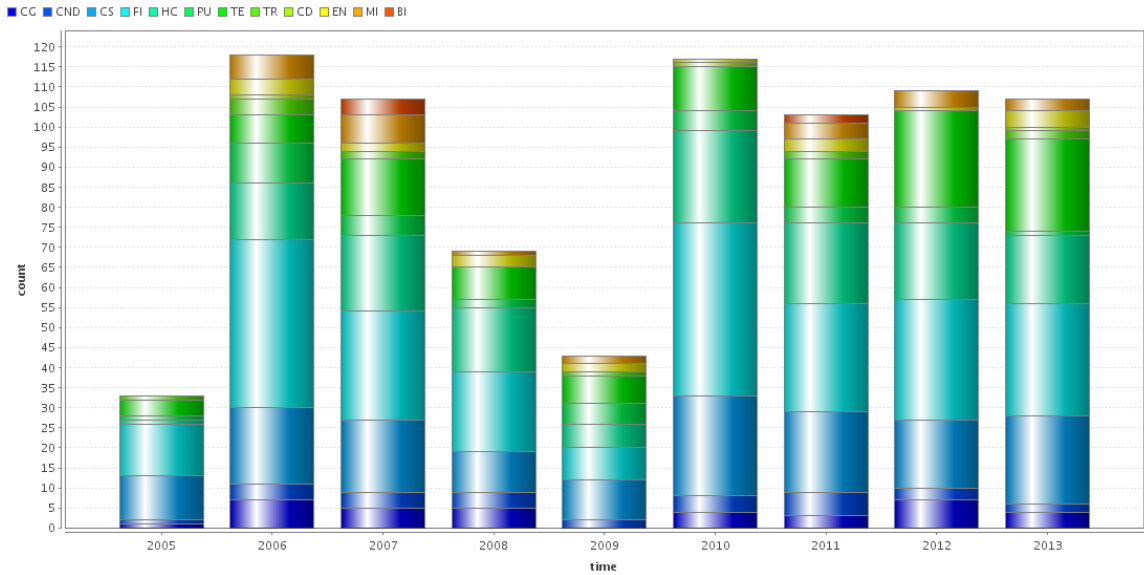


Figure 5.18. Breached Sectors By Year

### 5.5. Predictor Summary

Our intention in performing this analysis is to demonstrate the dynamic iterative nature of exploratory analysis using the framework not to draw any definite conclusions.

Having said that, our case study comparative analysis for this dataset showed that ‘Market Cap Size’ and ‘Sector’ are likely factors for breach susceptibility (assuming minimal noise or external factors within the data). This can be first detected by looking at the ‘Aggregate Pie Plot’ in a side-by-side configuration (breach set and control set), and further verified by using the odds ratio plots to determine statistical significance. This may be because they handle more types of sensitive data, are targeted more due to their revenue pots, or due to unknown reasons. Industry, on the other hand, is inconclusive due to lack of samples.

We discussed predictors for probability of breach using the framework visualizations. Using these predictors, one can attempt to evaluate risk and relative probability

of breach. The strongest predictors we found in our research are delineated below, however, there are likely others to be found.

- **Cap-Size:** Larger.
- **Region:** Pacific
- **States:** Arizona, California, Florida, Nevada, Rhode Island, D.C.
- **Population:** Inconclusive
- **Industry:** Inconclusive
- **Sector:** Finance / Health Care
- **City Sizes:** Inconclusive
- **Cap-Size vs. Sector:** SML Technology, MEGA-Energy
- **Entity Type vs. Region:** Private Education in New England, Government in South Atlantic, Public Companies in Mid Atlantic

## 5.6. Follow-On Research

As part of a follow-on study, more data could be pulled from external sites, namely the Yahoo API for financial data. This API allows for CSV download of a plethora of attributes for stock-ticker symbols, all provided in the URL argument list. Some representative sample tags are provided below. Merging this data with the breach dataset could provide more variables for breach-only analysis as well as historical trend analysis.

Table 5.8. Sample Yahoo API query tags

<b>Name</b>	<b>Tag</b>
AfterHoursChangeRealtime	c8
AnnualizedGain	g3
Ask	a0
AskRealtime	b2
AskSize	a5
AverageDailyVolume	a2
...	...
Volume	v1

## Chapter 6

### CMS CASE STUDY DESCRIPTION

In this chapter we introduce the CMS dataset. We begin by presenting a background to the CMS dataset (Section 6.1), including prior work and the intentions behind its compilation. The collection process as well as a description of the surrounding attributes is then detailed in Section 6.2. We then discuss transforming the data to provide for additional attributes for dataset analysis (Section 6.3). Next, potential risk factors and analysis goals are presented in relation to the data (Section 6.4). Finally a brief overview of the analysis methodology is presented along with perceived limitations inherent to the data (Section 6.5, 6.6).

#### **6.1. Background**

The case study is loosely based on the analysis of the paper: “Hacking is not random: a case-control study of webserver-compromise risk” [33], wherein the authors identify risk factors that are associated with higher rates of webserver compromise. By utilizing reported data of compromised web servers from the nonprofit Stop Badware [41], an organization whose mission includes “making the Web safer by fighting badware”, the authors analyze attributes of these hosts. Looking at the relative distributions, they find that websites derived from WordPress and Joomla content generators are more likely to be hacked than those not running any Content Management System (CMS). Also, servers using Apache and Nginx for their webserver infrastructure are more likely to be hacked than those running Microsoft Internet Information Services (IIS). The paper[33] in turn builds upon analysis from a previous



paper: Identifying Risk Factors for Web-Server Compromise [44]. The authors find that country of origin, generator type, and server type are important risk factors in webserver compromise. Again, WordPress and Joomla as well as Apache and Nginx are contributing factors.

After publication, the authors of [44] and [33] continued to collect data on web-servers. We utilize this data within the framework in order to carry out longitudinal analysis.

## 6.2. Dataset Description

Just as in the previous case study, the data was collected from two data sources following the case-control study comparative analysis paradigm. The control set originates from a random sample of domains listed in the .com zone file. While limited to a single top level domain, the .com domain encompasses close to half of all registered domains across a dispersed set of countries. More so than any other domain, .com contains diverse content, so it should represent a wide distribution of generators, web-servers, and countries. Thus it should be broad enough to be representative of most web content. Although this strategy effectively eliminates top level domain (TLD) as an analysis parameter, the increased diversity leads to a more representative sample of all web content.

The compromise set derives from a merged feed of “phishing” urls and search-redirect reports initially. The time-based collection consists of hacked websites reported to StopBadware’s datasharing program. The servers included in the list are added if they have observed to redirect to a third-party website and engage in cloaking. Reports originate from URL reports from concerned parties as well as companies with a vested interest in removing nefarious operations from the web (e.g. Google, Mozilla).

At chosen days after blacklisting, the websites were sampled (day 0, 1, 2, 5, 7, 15). The URLs exist on these lists until they are unblacklisted or reach day 15. The authors collected this additional data to help understand what happened to web servers after compromise and whether that action was effective. While the framework may not be capable of fully demonstrating what happens on a URL by URL basis, as it would require correlating ids between tables, this additional data is still made available for trend analysis. Segregating the data in this fashion (day 0, 1, 2, 5, 7, and 15 as distinct SQL tables) allows the data to be compared without excessively adding to the size of any one table. This additionally has the benefit of increased performance on the ‘Read Database’ operator.

Both datasets have been collected with the the same attributes, which makes case study comparison methods possible, as detailed in Table 6.1. Additionally, because WordPress is such a significant portion of the market, we transformed the generator string to a separate attribute ‘wordpressVersion’ through a series of AWK and SQL scripts. This attribute is a derivative from the collected ‘generator’ attribute and represents an applicable major version of the software.

Table 6.1. CMS Dataset Attributes

Attribute	Description
id	unique id corresponding to IP / url
day	days since blacklisting
server	server string from header
servertype	webserver type
country	server country of origin
generator	content generator string from header
generatortype	content generator type from header
tld	top level domain of the website
collectiondate	timestamp that the data was collected
wordpressVersion	wordpress major version, numeric (e.g. 3.0,3.1, etc)

### **6.3. Data Aggregation**

The attribute “wordpressVersion” as discussed in Section 6.2 was the only data element derived from the original dataset. However, in order to manipulate and create separate tables, create SQL-based timestamps, and perform general cleanup of the raw data, we had to run a series of 8 scripts. The scripts used in this transformation / import process utilized AWK and SQL and are provided at download location #9 in Section A.3.

### **6.4. Analysis Goals**

By examining the CMS dataset, we hope to determine which attributes of a website correlate with compromise. This might include generator type, server type, or country of origin. Are certain content generators more susceptible to risk of webserver compromise? What webserver software also results in increased risk? How do these factors change over time?

We seek to both validate results found in the previous two papers and examine the data from different perspectives using the framework. The intent is not to make any substantive claims, but to demonstrate dynamic iterative analysis using the framework.

### **6.5. Methodology**

For this analysis we predominantly focus on a case-control study format. Compromised data is compared in relation to controlled data. The control dataset, as previously stated, is a random sample of .com websites from the .com zone file. It serves as the control group and a logical comparison alternative. After looking at relative percentages and counts, odds ratio plots can be used to determine statistical measures of the findings.

We also perform a compromise-only analysis, wherein the compromised data is analyzed on its own accord. In looking at just the compromise data on its own, two-way categorical plots can provide insight into whether a combination of two attributes yields higher risk. Additionally, a time-based attribute means we can set up time plots to visualize the data in yet another dimension, so we can attempt to understand how this data changes over time.

## **6.6. Data Limitations**

There are some limitations inherent in this dataset. The five month time period that the compromise dataset corresponds to (February 2015 to July 2015) is only half the time duration as that of the control dataset (September 2014 to July 2015). This time interval is brief and may not prove useful for comparative case-study analysis. Additionally, the generator and server versions strings represent raw server strings. While we were able to construct a major WordPress version through a complex series of data transformations. There is not enough consistency in the raw string to extrapolate versions for other generator types (e.g. Joomla) or for webserver software (e.g. Apache, IIS, etc.).

## Chapter 7

### ANALYSIS OF CMS DATASET

After the webserver compromise data was compiled and imported to the database, we performed an analysis of the CMS data. In performing this analysis, we again demonstrate both the iterative process and utility provided by the framework. We also demonstrate the reusable nature of the framework by applying it across a distinctly different dataset. The analysis is decomposed into two distinct strategies: case study comparative analysis (Section 7.1) as well as looking at the compromise data by itself (Section 7.2). We then supplement with both a geospatial analysis (Section 7.4) and a time-based analysis (Section 7.3). Finally we conclude by summarizing predictors for webserver compromise, and describe the ongoing effort regarding this dataset. (Section 7.5 & 7.6)

#### **7.1. Case-Control Study Comparative Analysis**

By comparing a control group and an experiment or "treatment" group, we can uncover factors in the two datasets that lead to increased risk of webserver compromise. As discussed in Section 6.4, we examined the compromised webserver set (treatment) in relation to the set of servers originating from the .com zone file (control).

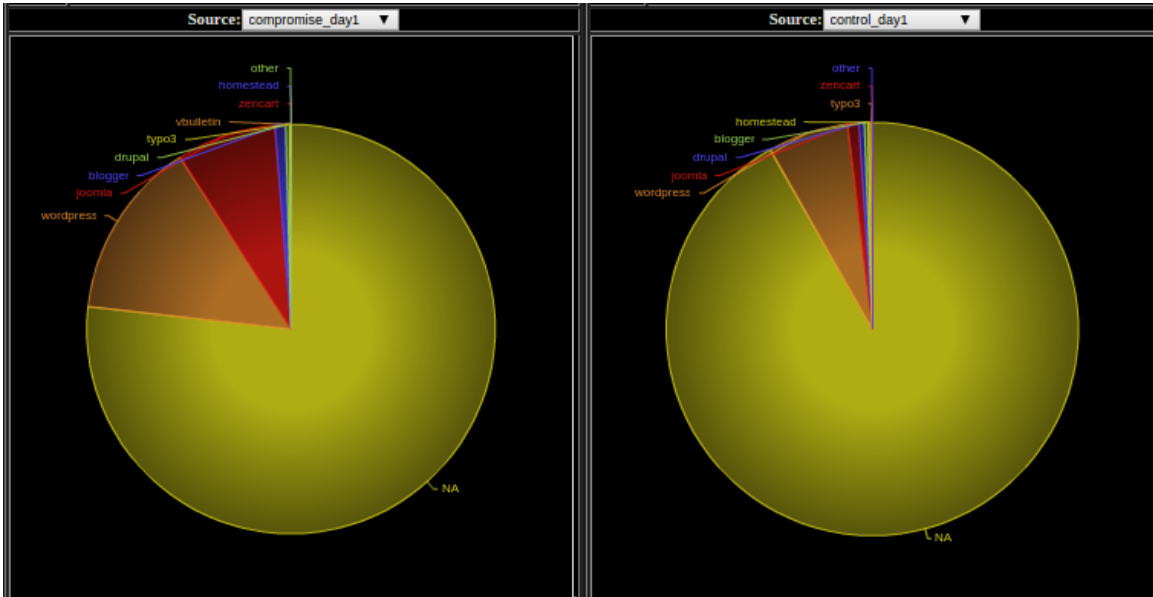
Throughout this comparative analysis, we opt to examine risk factors using day one data. For the reason that day 1 represents the first day the URL was known to be compromised, it should be beneficial in determining what attributes contribute to increased risk for day 1. Alternatively if we were to use a different day (e.g. day 5), the analysis effort would be focused on determining attributes that lead to increased

risk for that later day.

### 7.1.1. Proportional Analysis

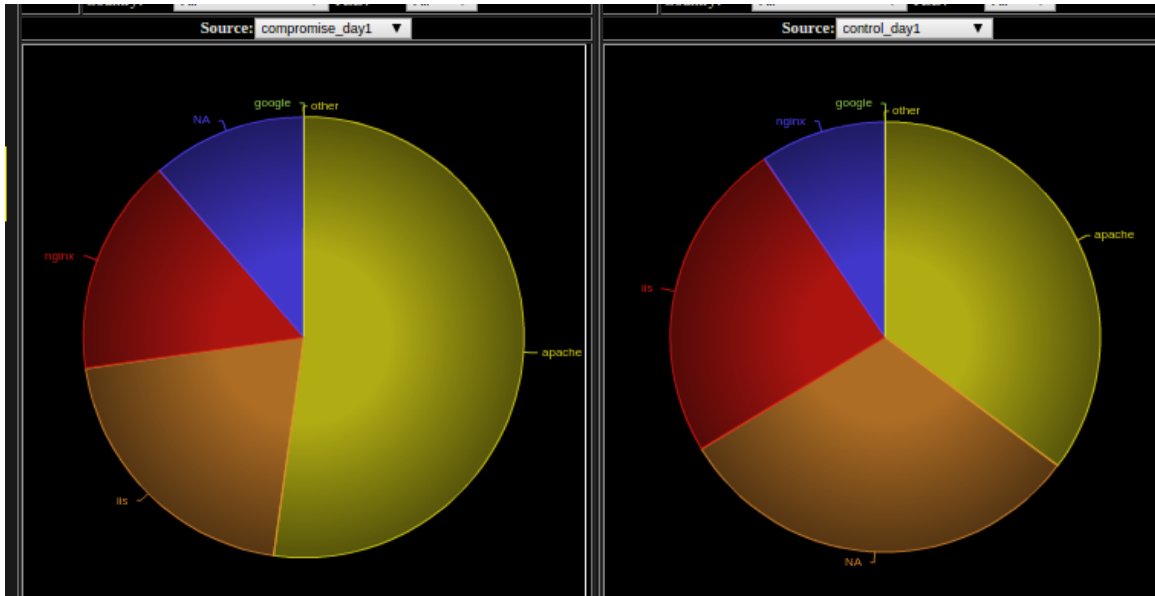
To perform a proportional analysis and begin the iterative process of looking for relevant results, we first create a split portal view with the compromise day 1 dataset on the left and the control day 1 dataset on the right. By utilizing an "Aggregate Pie Plot", we can quickly scan the datasets for relative differences in percentages. For example, we find that when analyzing by 'generatortype' in Figure 7.1, that WordPress and Joomla are over-represented in the compromise set (14% and 8% vs. 6% and 1% respectively). In contrast, absence of a generator ("NA"), is over-represented in the control dataset (92% vice 77%). So presence of a content generator could be a risk factor for compromise.

Next we examine 'servertype' utilizing a similar strategy. Apache and NGINX are over-represented in the compromise set (52% and 16% vice 35% and 9%), while the absence of webserver information in the server header ("NA") is over-represented in the control set (31% vice 11%) (Figure 7.2). So in this case, hiding webserver information appears to be a negative risk factor for webserver compromise.



Compromise Day 0		Control Day 0	
NA	76.8%	NA	91.8%
wordpress	14.1%	wordpress	6.3%
joomla	7.9%	joomla	0.9%
blogger	0.8%	drupal	0.4%
drupal	0.4%	blogger	0.3%
typo3	<0.1%	homestead	0.2%
vbulletin	<0.1%	typo3	0.1%
		zencart	<0.1%

Figure 7.1. Generator Type - Compromise vs. Control



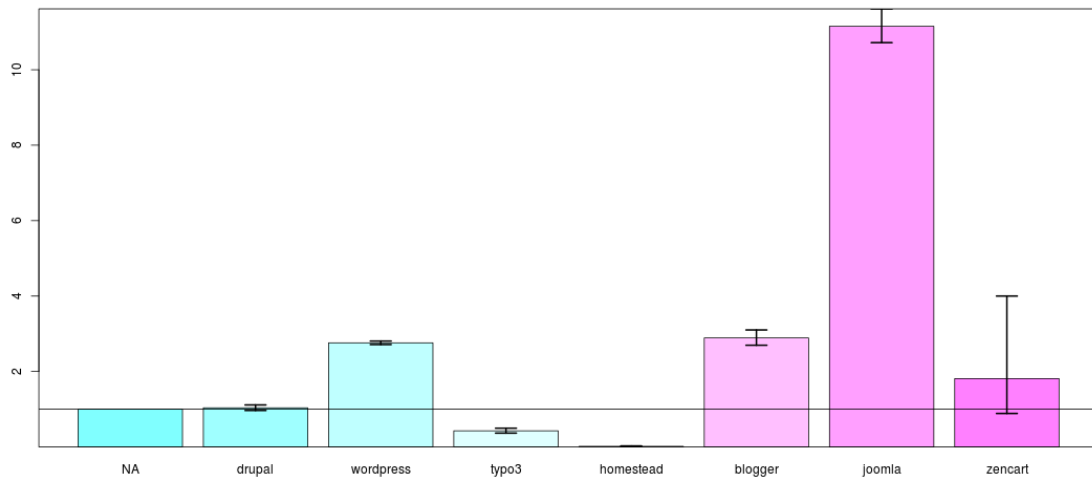
Compromise Day 1		Control Day 1	
apache	52.2%	apache	35.2%
iis	20.5%	NA	31.1%
nginx	15.9%	iis	24.2%
NA	11.3%	nginx	9.4%
google	<0.1%	google	<0.1%

Figure 7.2. Server Type - Compromise vs. Control



### 7.1.2. Odds Ratio Analysis

To further validate the results of the previous section, we use the "Odds Ratio Plot". In this situation the odds ratio represents the odds of an outcome occurring (webserver compromise) given a particular exposure in relation to the odds of the outcome occurring in absence of that exposure (control set). The results are statistically significant at the associated confidence if the the confidence interval (whiskers) fail to intersect other bars' relative ranges.



generator type	estimate	lower	upper
NA	1.0	NA	NA
wordpress	2.76	2.71	2.81
typo3	0.43	0.37	0.50
joomla	11.15	10.72	11.61
drupal	1.04	0.97	1.11
blogger	2.89	2.70	3.10

Figure 7.3. Generator Type - Odds Ratio (Day 1)

We then verify our results from our proportional analysis for 'generator type'. The treatment table was selected as "compromise\_day1" and the control as "control\_day1".

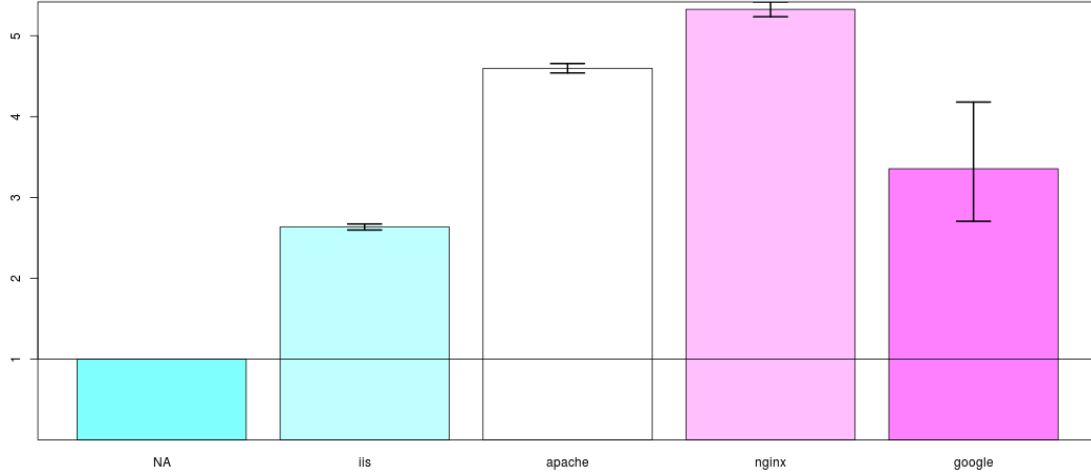
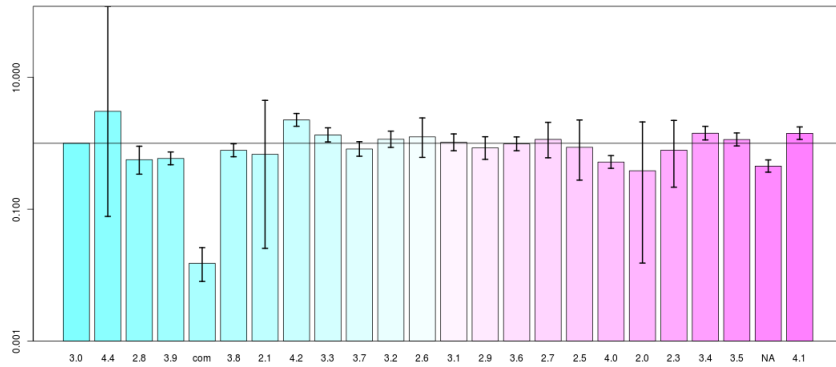


Figure 7.4. Server Type - Odds Ratio (Day 1)

The odds ratios, depicted below in Figure 7.3, show that Joomla (2.88) and WordPress (1.0) are positive risk factors for webserver compromise in relation to not having a content generator. The absence of a content generator is, on the other hand a negative risk factor. The results are statistically significant for all options as none of the confidence intervals intersect, and correlate well with the paper’s findings [33].

Plotting odds ratios for ‘servertype’ (Figure 7.4) we find that Apache and NGINX are positive risk factors for compromise (OR=4.60 and 5.33) in relation to not providing server software information. Thus, the best strategy to avoid web server compromise might be to omit server information on a http response.

Finally, we look at whether a specific version of WordPress is at higher risk than others. We do this by filtering on WordPress, then selecting ‘wordpressVersion’ as the odds variable. The associated odds for different versions of WordPress in relation to version ‘3.0’ are shown in Figure 7.5. Although some versions are statistically significant in relation to one another, there doesn’t appear to be any trend in the data. Thus the odds ratio plot appears to be inconclusive.



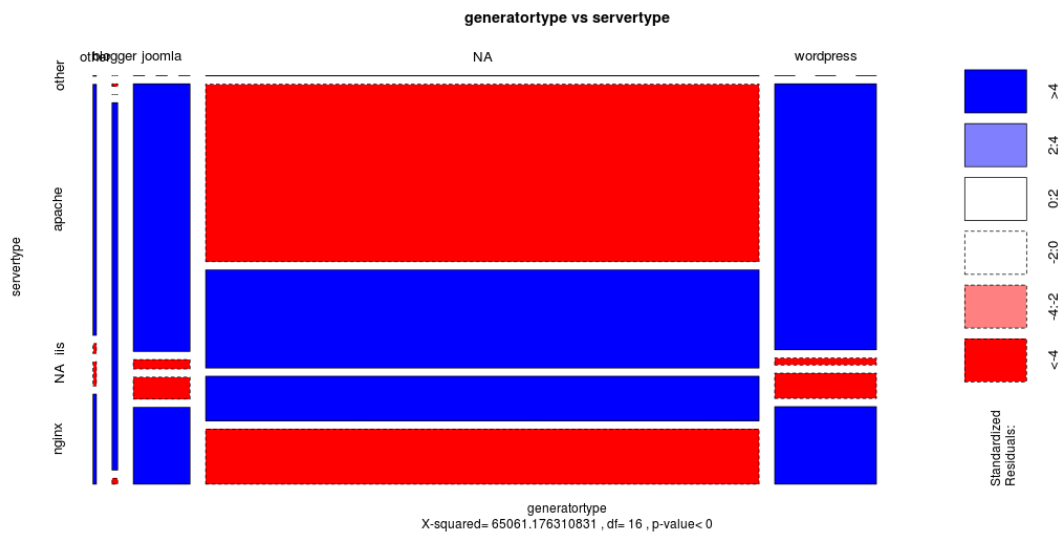
wordpress version	estimate	lower	upper	wordpress version	estimate	lower	upper
3.0	1.0	NA	NA	3.2	1.15	0.87	1.53
4.4	3.05	0.08	119.6	2.6	1.25	0.61	2.43
2.8	0.56	0.34	0.90	3.1	1.03	0.77	1.39
3.9	0.59	0.47	0.74	2.9	0.85	0.57	1.26
com	0.02	0.01	0.03	3.6	0.98	0.77	1.25
3.8	0.78	0.63	0.98	2.7	1.15	0.60	2.08
2.1	0.68	0.03	4.48	2.5	0.87	0.28	2.26
4.2	2.26	1.81	2.84	4.0	0.52	0.42	0.65
3.3	1.34	1.05	1.72	2.0	0.38	0.02	2.11
3.7	0.82	0.64	1.06	2.3	0.79	0.22	2.22
3.4	1.42	1.12	1.81	3.5	1.14	0.91	1.44
NA	0.45	0.37	0.56	4.1	1.42	1.14	1.77

Figure 7.5. WordPress Version - Odds Ratio (Day 1)

## 7.2. Compromise-Only Analysis

We then turn towards a demonstration of analysis using only the compromise data. In doing so we utilize visualizations that provide meaningful statistical relevance in relation to a single table or dataset (e.g. mosaic plot).

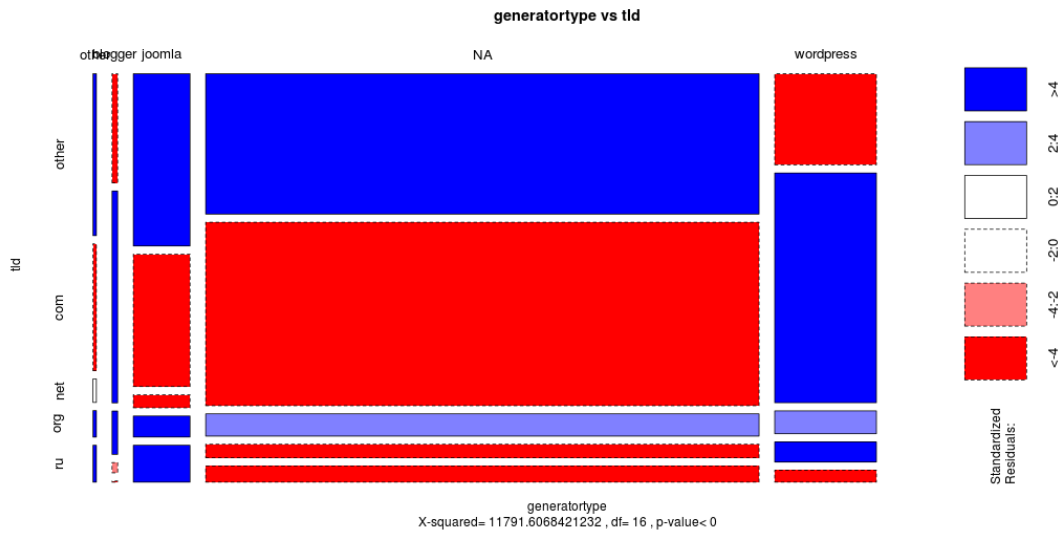
### 7.2.1. Two-Way Categorical Analysis



Over-represented	Under-represented
wordpress-apache	NA-apache
NA-iis	joomla-iis
NA-NA	wordpress-NA
joomla-nginx	NA-nginx
wordpress-nginx	wordpress-iis

Figure 7.6. Server Type vs. Generator Type(Day 1)

Chi-squared statistical relevance in relation to two categorical variables was analyzed through the "Mosaic Plot". Most every combination of inputs proved to be statistically relevant, with small p-values. Since there were many options allowed, we had to limit the output by only using the top four results. We provide a couple of



Over-represented	Under-represented
ru-joomla	com-Joomla
org-joomla	com-NA
com-wordpress	ru-wordpress
org-wordpress	org-NA

Figure 7.7. TLD vs. Generator Type(Day 1)

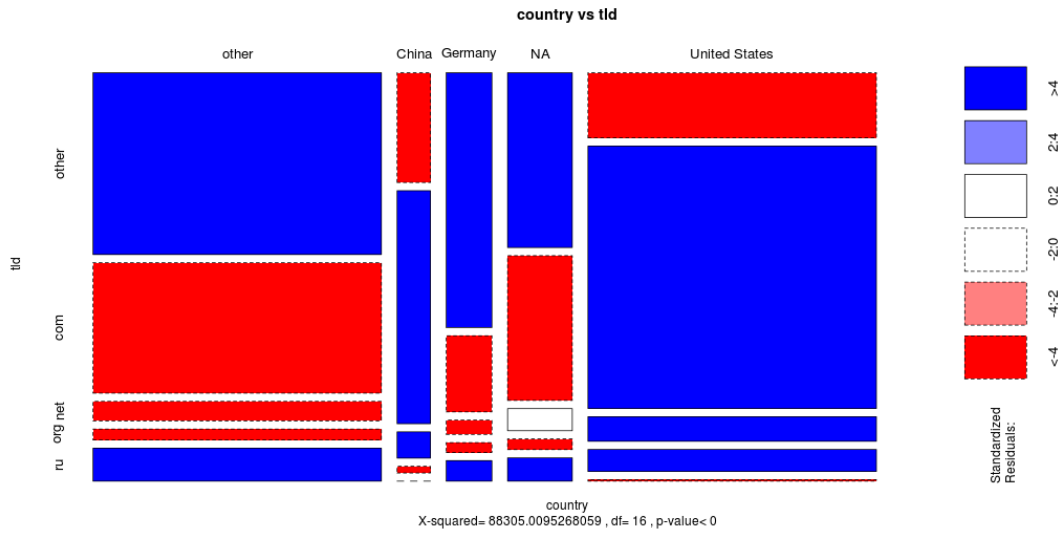
the more interesting plots.

Are there certain content generator pairings that interact in a negative way together or are more preferred by cybercriminals? For ‘servertype’ vs. ‘generatortype’ (Figure 7.6), we observe several pairings that are both over-represented and under-represented within the compromise dataset. Thus attributes in this table have a higher observed frequency than would be present if there were linear independence of these variables. These overrepresentations include wordpress/apache, joomla/nginx, No CMS/iis, No CMS/No Webserver, and wordpress/nginx.

Are there certain domains where cybercriminals operate to defeat specific content management systems? For ‘tld’ vs. ‘generatortype’ (Figure 7.7), we observe higher residuals swaths. Over-represented pairings include com-wordpress, org-joomla, org-

joomla, and ru-joomla.

Are there certain country / domain pairings that cybercriminals prefer? We depict this visualization in Figure 7.8. United States and China are over-represented within the .com and .net domains. Perhaps these pairings are more likely to be reported or perhaps there are more nefarious reasons at play.



Over-represented	Under-represented
United States-com	Germany-com
China-com	China-org
United States-net	Germany-net
United States-org	Germany-org
China-net	

Figure 7.8. Country vs. TLD(Day 1)

### 7.3. Time-Based Analysis

To analyze how risk factors change over time, we have a few methods available to us. We start with the "Time Line Plot" which allows us to visualize raw counts within a time period: in this case February to July of 2015. Both WordPress (Figure 7.9) and Apache (Figure 7.10) seem to spike heavily relative to others options in April of 2015. Looking back to news during that time period we find that there was a WordPress vulnerability that was being exploited (Double Zero Day) [20]. This effected all versions of the software.

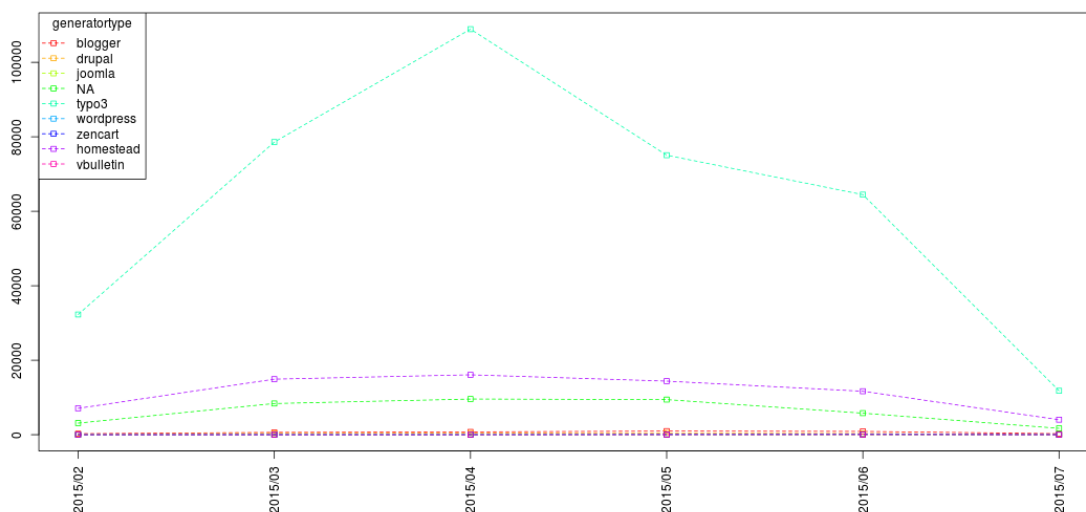


Figure 7.9. Generator Type Compromises By Month

We then look at the "Time-Based Odds Ratio" plot, which visualizes odds ratios within time intervals. Visualizing 'serverType' odds ratios by month (depicted in Figure 7.11), we find that Apache (which also serves as our normal) becomes clearly defined as a positive risk factor in later months (May onward). It is unsure if this is a manifestation of collection processes or if there is any clear explanation for this.

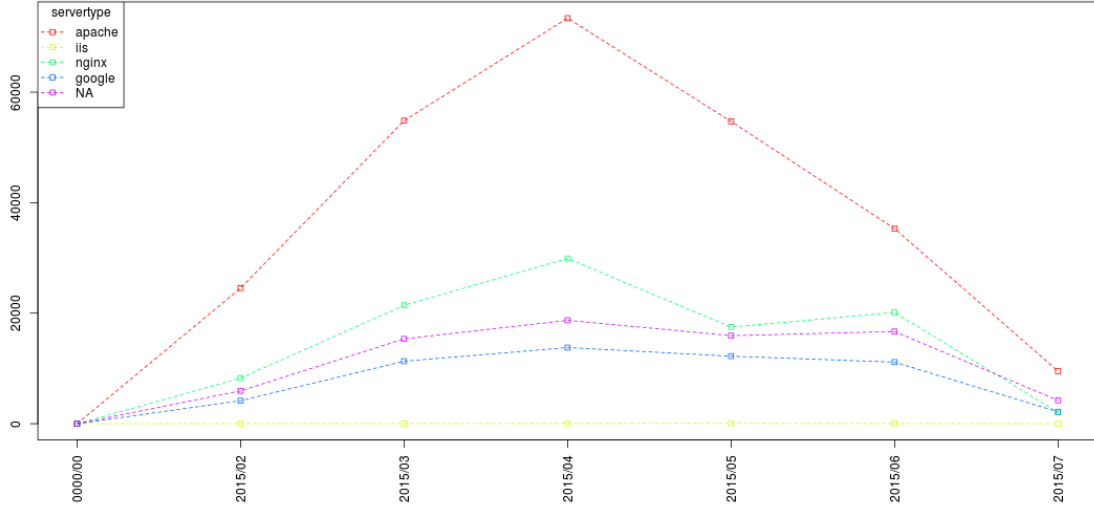


Figure 7.10. Server Type Compromises By Month

We then visualize percentages with the "Time Aggregate Bar Plot" to try to understand the decomposition of the WordPress spike discussed earlier in this section. Visualizing WordPress versions in Figure 7.12 we find that version 4.1 remained the highest percentage from March to May but was eventually overtaken by version 4.2 in June. There doesn't appear to be anything suspicious here, as this trend can alternatively be attributed to the natural process of upgrading to a newer release.



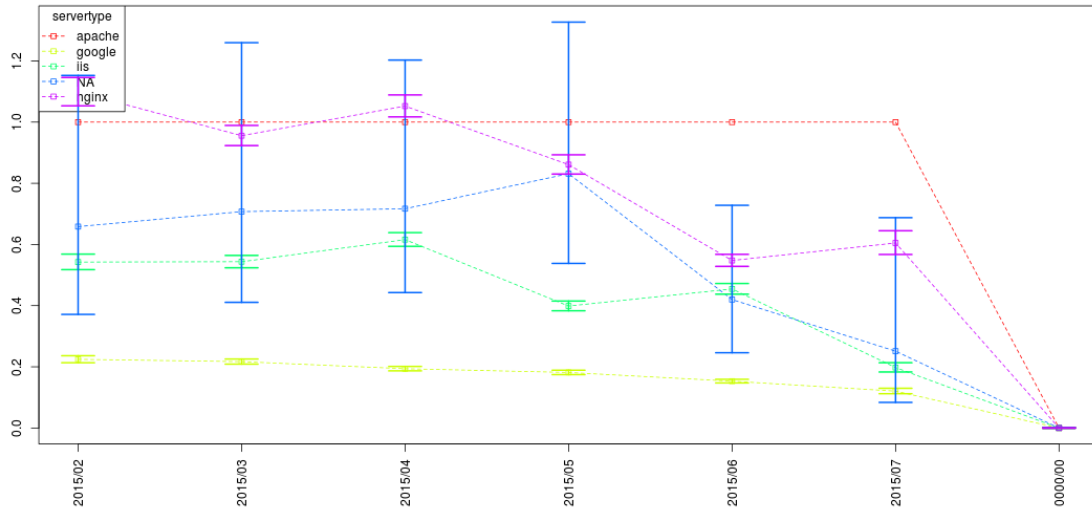


Figure 7.11. Server Type Odds Ratios By Month

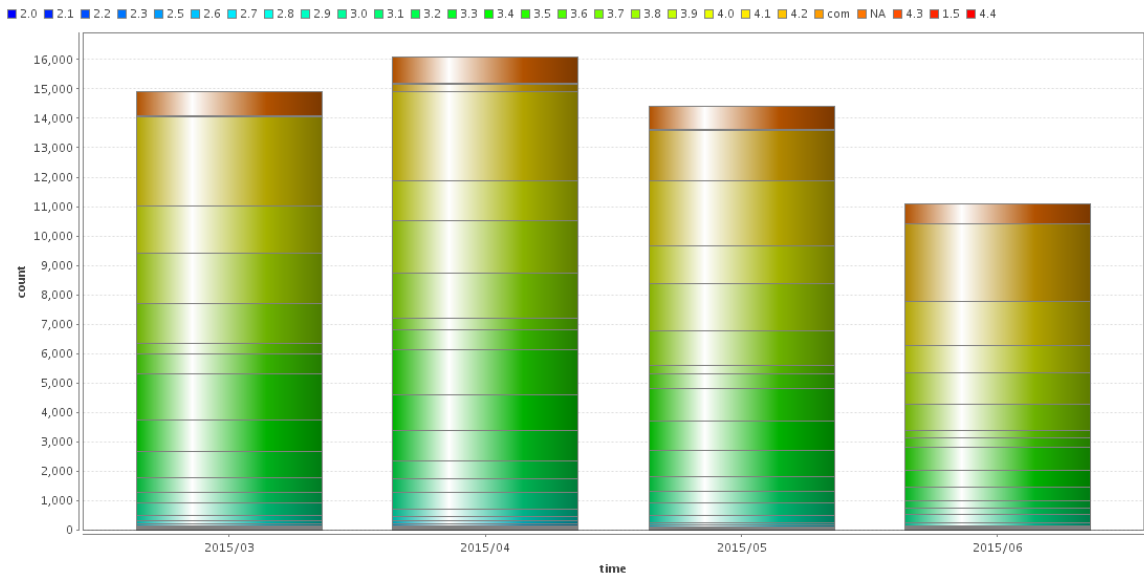
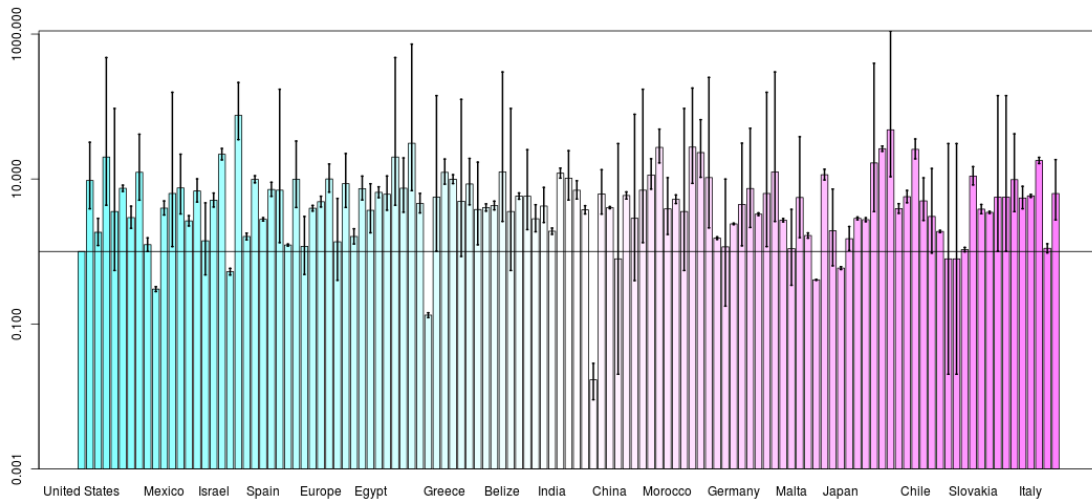


Figure 7.12. WordPress Version By Month Aggregate

## 7.4. Geospatial Analysis

We then examined the relative risk of compromise from a geographic perspective. We first used the "Odds Ratio Plot" to determine where web servers had the highest risk of being compromised. We computed these odds in relation to "NA", where country of origin was unknown or not reported. The resulting graphic is shown in Figure 7.13. Countries like Russia and China have high risk (46.0 and 7.0), while the United States appears to be relatively neutral (1.75). Other countries like Switzerland (0.53) and the British Virgin Islands (0.23) appear to have negative risk.



country	estimate	lower	upper
NA	1.0	NA	NA
United States	1.75	1.73	1.77
China	7.07	6.83	7.31
Russia	45.93	42.27	49.95
Germany	4.21	4.11	4.32
Switzerland	0.53	0.49	0.57
Virgin Islands(U.K)	0.23	0.21	0.25

Figure 7.13. Odds Ratios By Country

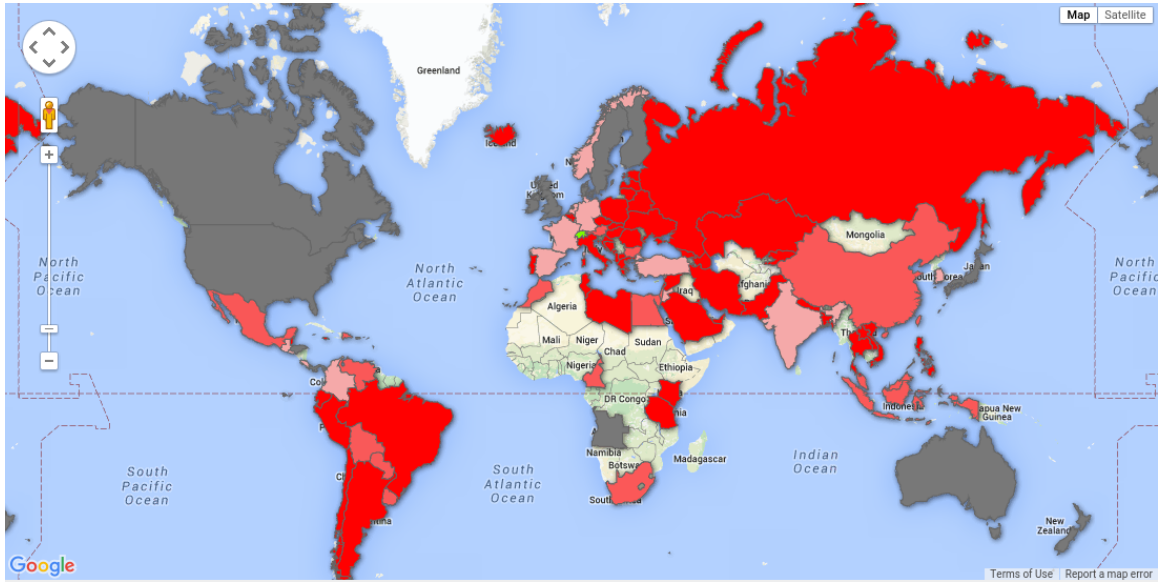


Figure 7.14. Geospatial Country Odds Ratios

To expand our perspective, we can then map this same odds ratio output to a Google Fusion Table geospatial representation. We started by creating a color column and assigning colors to each of the odds ratio estimates as discussed in Section 3.3.10.2. Negative risk is indicated by green, while positive risk is indicated by darker red. We then uploaded the resultant csv file to Google Fusion Tables and merged it with our World Country Boundaries (KML). After configuring the geospatial map to use the ‘geometry’ key and use the colors provided by our new color column, a publishable map is provided (Figure 7.14). Here we find that Western Asia/Eastern Europe as well as South America tend to represent a disproportionate amount of servers that were compromised. Since a lot of criminal activity also emanates from these regions, it begs the question whether this is due to geographic proximity or some other corollary.

## 7.5. Predictor Summary

As previously discussed, our intention in performing this analysis is to demonstrate the dynamic iterative nature of exploratory analysis using the framework not to draw any definite conclusions. In pursuit of such, we discussed predictors for the CMS dataset using the framework visualizations. With a firm grasp of these predictors and trust in the dataset, one can attempt to evaluate risk and implement remedies to avoid webserver compromise. With regards to webserver compromise, the strongest risk factors from this dataset are delineated below, however, there are likely others to be found.

- **Server Type:** Apache
- **Generator Type:** Joomla and WordPress
- **Country:** China, Russia
- **WordPress Version:** Inconclusive
- **Server Type By Month:** Apache spike April 2015
- **Generator Type By Month:** WordPress spike April 2015
- **WordPress Version By Month:** Inconclusive
- **Generator Type vs Server Type:** wordpress/apache, joomla/nginx, No CMS/iis, No CMS/No Webserver, and wordpress/nginx
- **Country vs TLD:** United States and China / .com and .net
- **Generator Type vs TLD:** com-wordpress, org-joomla, org-joomla, and ru-joomla

## 7.6. Follow-On Research

The Security Economics Lab at the University of Tulsa has plans to use the framework to track and analyze how compromises of webservers develop and subsequently get resolved. As more data is collected, tighter confidence intervals will emerge. The framework can also be expanded to provide additional visualizations as they see fit. If the data grows to critical mass, a nonrelational database should be examined to house the data.

## Chapter 8

### HONEYPOT DATASET

In this chapter we operate on a third dataset, the Honeypot dataset, collected externally. In performing the analysis, the reusable nature of the framework is validated further. We begin by presenting a background to the Honeypot dataset (Section 8.1), including an overview of both the authors and the collection process, as well as a synopsis of analysis that was performed thus far. A description of the surrounding attributes is then presented in Section 8.2. We then discuss aggregating additional data to the dataset for enhanced analysis (Section 8.3). Next, potential risk factors and analysis goals are presented in relation to the data (Section 8.4). We then present a brief overview of the analysis methodology along with perceived limitations inherent to the data (Section 8.5, 8.6) before diving into analysis of the data.

The analysis of the data is performed much like the other two studies. However, since we have no control data, we only perform an analysis of the honeypot data on its own. This analysis is further decomposed into the following parts: a proportional analysis (Section 8.7.1), a geospatial analysis (Section 8.7.2), a time-based analysis (Section 8.7.3), and two-way categorical analysis (Section 8.7.4). Finally we conclude by summarizing predictors for honeypot susceptibility and ponder future efforts regarding this dataset. (Section 8.8 & 8.9)

#### **8.1. Background**

Data Driven Security (DDS) [10], a collaboration effort between Jay Jacobs and Bob Rudis, is a security visualization research hub. Using their website, blog, book,

and podcasts, Jacobs & Rudis aim to “help security domain practitioners embrace and engage all elements of security data science to help defend their organizations.” [11] They focus on teaching people how to better understand security data through analytics and visualization demonstrations, effectively bringing meaning to security related datasets. In doing so, they employ programming tools such as Python and R, presenting the data in significant ways.

The “Honeypot” dataset, collected and examined by Jacobs on his DDS blog [12], is a third dataset intended to further validate the framework. The data was originally collected by Daniel Blander, a contract information security and risk management specialist [40]. In an effort to collect security intrusion information on hackers, Blander setup Amazon Web Service (AWS)[2] virtual machines around the world to act as honeypots. These honeypots were setup to attract cyber criminals who attempt to penetrate other people’s systems (hackers, crackers, script kiddies, etc.). Details about the exact installation process, default load-out of software, or collection methodologies are unknown. We do know that the data was collected from the packets from the iptables from nine hosts during the period of March to September of 2013. The data was then subsequently provided to Jay Jacobs at DDS who performed an R-based analysis of the data in a two part blog post.

In part one of his blog post[28], Jacobs views the data from a high-level summary viewpoint. He finds that the TCP packets vastly outnumber the UDP. He then performs an aggregate calculation to see the relative percentages of target host machines, and extrapolates this to the time domain by viewing this breakdown on a daily basis. Lastly, he filters the data to remove duplicate IP addresses, and presents the metric as the number of unique IP addresses per day accessing a particular host. He sees that Oregon, Tokyo, and Singapore are seeing about twice as many unique hosts as the others, but has no answers for why.

In part two of his blog post[29], Jacobs uses a box plot to depict the average number of unique IP accesses per day for the top 10 ports (services) being attacked. Additionally, through a series of data transformations, he plots this metric over time. Despite, visualizing this data in effective ways, Jacobs concludes with more unanswered questions than answered.

## 8.2. Dataset Description

The data [9] was well structured and required little cleanup. However, we felt it necessary to correlate the destination (attack) port to a tangible service name by joining it with a list of service names. This was performed by merging with a csv file provided by the Internet Assigned Numbers Authority[6]. The scripts for this import / merge process (using awk and SQL) are provided at download location #9 in Section A.3.

We detail the respective fields within the dataset in Table 8.1. Each row presumably corresponds to an audit log artifact with an iptable existing within one of the 9 target hosts. Within the remaining portion of this chapter we refer to each record as an “attack”. However, we have no way of knowing the intent of the IP access.

## 8.3. Data Aggregation

As previously discussed in Section 8.2, the data was merged with the IANA data to determine the service name associated with the “attack” port using AWK. The merged data was then imported from csv into mysql using the scripts provided at at Download Location #9 in Section A.3 and methodology described in Section B.2.1



Table 8.1. Honeybot Dataset Attributes

Fields	Description
datetime	The date/time that the “attack” occurred
host_target	The AWS machine that was targetted in the “attack” <norcal, sa, sydney, eu, norcal, us-east, singapore, tokyo, oregon>
ip_src/ipaddr	The source IP (if known) in long int and octet format
protocol	The protocol used in the “attack” (TCP or UDP)
attack_port	The port used for the “attack”
country/code	The country and abbreviated country code that the “attack” emanated from
locale/abbr	The locale of the attacker (typically city) as well as an abbreviated version (number)
postal_code	Zip Code (if known) of the attacker
lat/lon	Latitude / Longitude of the attacker
service	The service that was targeted by the “attack”

#### 8.4. Analysis Goals

By examining the Honeybot dataset, we hope to discover patterns of cyber criminals as well as determine which characteristics of an “attack” lead to higher risk. This might include installed services, virtual machine location, or geographic region. For example, “do attacks target a certain set of services?” and “where are these attacks originating from?” We also hope to use the framework to discover time-based trends in the data from month to month. For example, “do criminals change their habits from month-to-month?”

#### 8.5. Methodology

For this analysis we have no control group, so a comparative analysis is not possible. We instead focus on analysis from an “attack” only perspective. This means choosing charts that make sense in relation to a single dataset.

This dataset lends itself well to heavy proportional and two-way categorical analysis. We can not use the box plot for lack of numerical attributes. However, since the dataset provides heavy access to geospatial data, we should be able to use the geospatial plot type. Additionally, a time-based attribute means we can set up time plots to visualize in another dimension.

## **8.6. Data Limitations**

There are some limitations inherent in this dataset. The seven month time period may not provide significant results. We also have no relevant information as to the makeup of the target host machines. For example, what services are installed? Are they different in any way? There also exists no information on the collection strategies. Also, how does one distinguish a single instance of an attack? Are some services over-represented if their communication protocol required more packet handshakes? Also, is this an evenly dispersed Honeypot-Net? Or are some geographic regions over-represented (i.e. United States)?

## **8.7. Framework Analysis**

In this section, we apply the framework to the dataset using a variety of analysis strategies. A summary of the findings is made in Section 8.8. We conclude by discussing analysis opportunities for this dataset.

### **8.7.1. Proportional Analysis**

Using our framework we first perform some basic analysis using the “Aggregate Pie Plot.” Just as Jacobs found in his analysis, we immediately find that Oregon, Singapore, and Tokyo have the most spurious traffic (Figure 8.1). Their activity accounts for close to 60% of the dataset.

We then turn towards examining what services (ports) are of interest to “attackers” for gaining entry into a system. We limit the results to the Top 20 returns and find in Figure 8.2 that “ms-sql-s” (1433), “http” (80), and “epmap” (135) are the most commonly attacked services. This correlates well with a list of commonly hacked ports found on the web [7].

If we look at the protocol involved we find that TCP accounts for 87% of all activity (Figure 8.3). Our postulation for this discrepancy is that services predominantly operate off UDP protocol. So “attackers” opt to target the service (not the protocol) for exploit in order to gain access into the system.

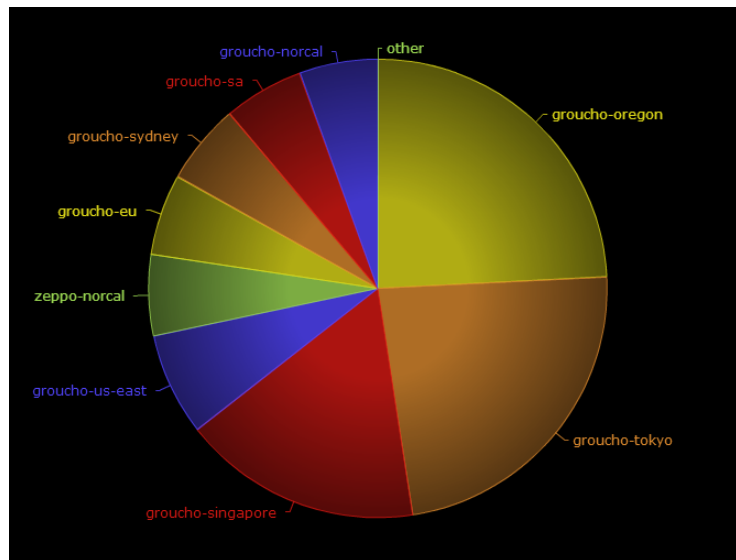


Figure 8.1. Attacks By Target Machine

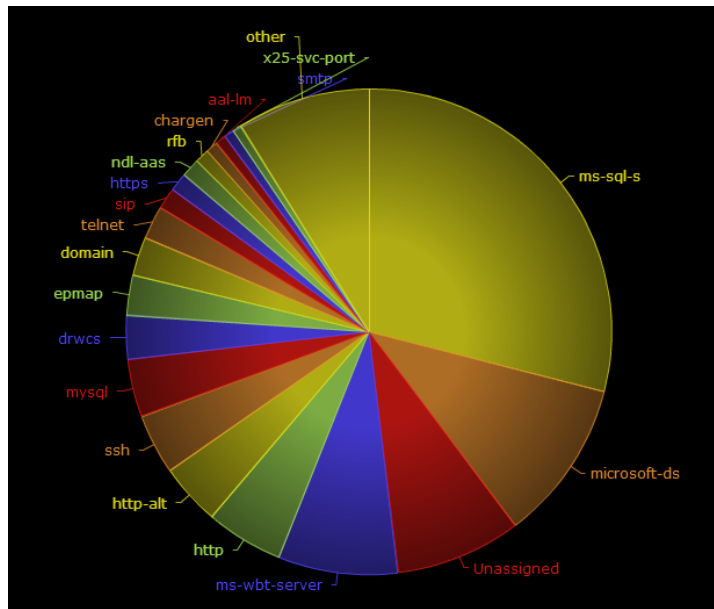


Figure 8.2. Attacks By Targeted Service

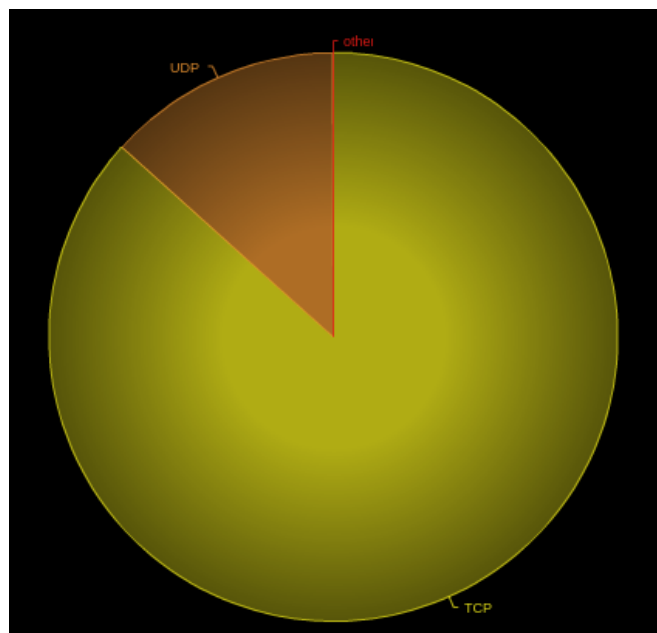


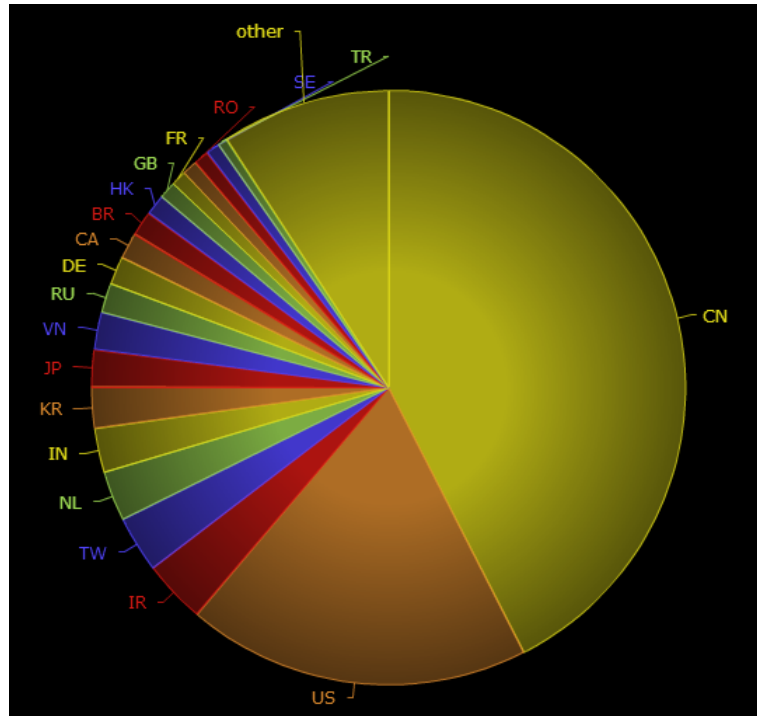
Figure 8.3. Attacks By Protocol

### 8.7.2. Geospatial Analysis

We examine the geographic makeup of the dataset by looking at the top 20 countries from which the attacks originate. The “Aggregate Pie Plot” is used to make this assessment by selecting a result limit of “20.” The chart (Figure 8.4) shows that China and the United States are responsible for the majority of attacks against the deployed honeypot virtual machines (42.5% and 18.7% respectively).

We then assess the data from a different perspective. Namely, are there certain geographic locations within China that are responsible for more attacks than others? In order to answer this, we use the “Geospatial Plot” in the heat map configuration, and assess whether or not attacks originate from a specific geographic location. Visualization is limited to the first 1,000 rows, so we randomly sample the data. Figure 8.5 exhibits a heavy concentration of attacks originating from Beijing, along with other heavy concentrations in Jiangsu Sheng, GungZhao, Tian Shui (approximate), and Shanghai. This seems to agree with the locality information provided within the dataset. Why are there heavy concentrations in Beijing? Could this be government funded activity? Examining the highest population cities (#1: GungZhao, #2: Shanghai, and #3: Beijing) on Wikipedia [48], we can dismiss this assertion, as the heat map seems to generally correlate with centers of population.

In order to assess attacks emanating from within the United States, we set the country filter to United States. After limiting the results, we find that California is the largest offender for this type of spurious activity with a relative percentage of 41% (see Figure 8.6). Additionally, we aggregate by zip code and find that within the United States: Walnut, CA (91789), Los Angeles, CA (90017), and Sunnyvale, CA (94085) are collectively responsible for 27% of this these attempts (see Figure 8.7).



Country	Percentage	Country	Percentage
China	42.5%	Germany	1.5%
United States	18.7%	Canada	1.4%
Iran	3.4%	Brazil	1.3%
Taiwan	3.1%	Hong Kong	1.1%
Netherlands	2.7%	Great Britain	0.9%
India	2.4%	France	0.8%
South Korea	2.2%	Unknown	0.8%
Japan	2.0%	Romania	0.8%
Vietnam	2.0%	Sweden	0.8%
Russia	1.6%	Turkey	0.5%
		Other	9.2%

Figure 8.4. Source of Attacks (Pie Chart)

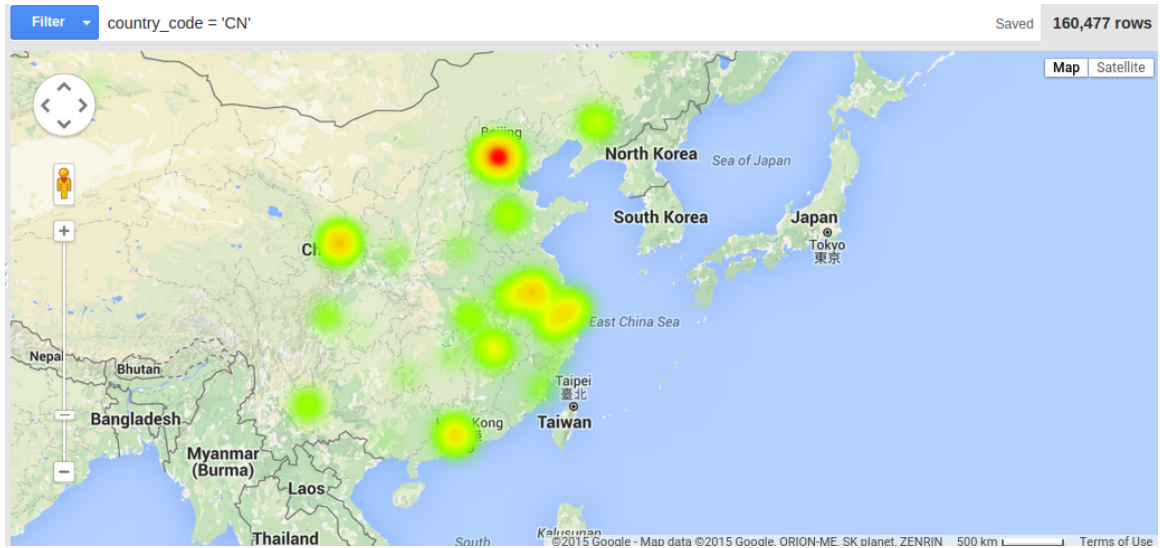


Figure 8.5. Source of Chinese Attacks (Heatmap)

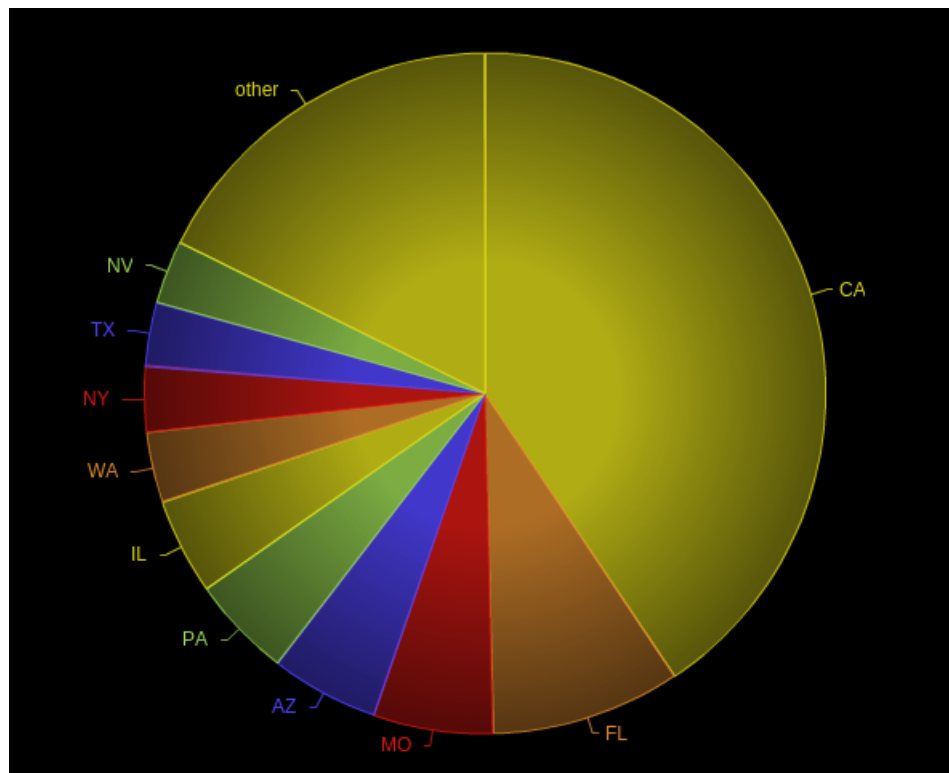


Figure 8.6. Source of U.S. Attacks (By State)

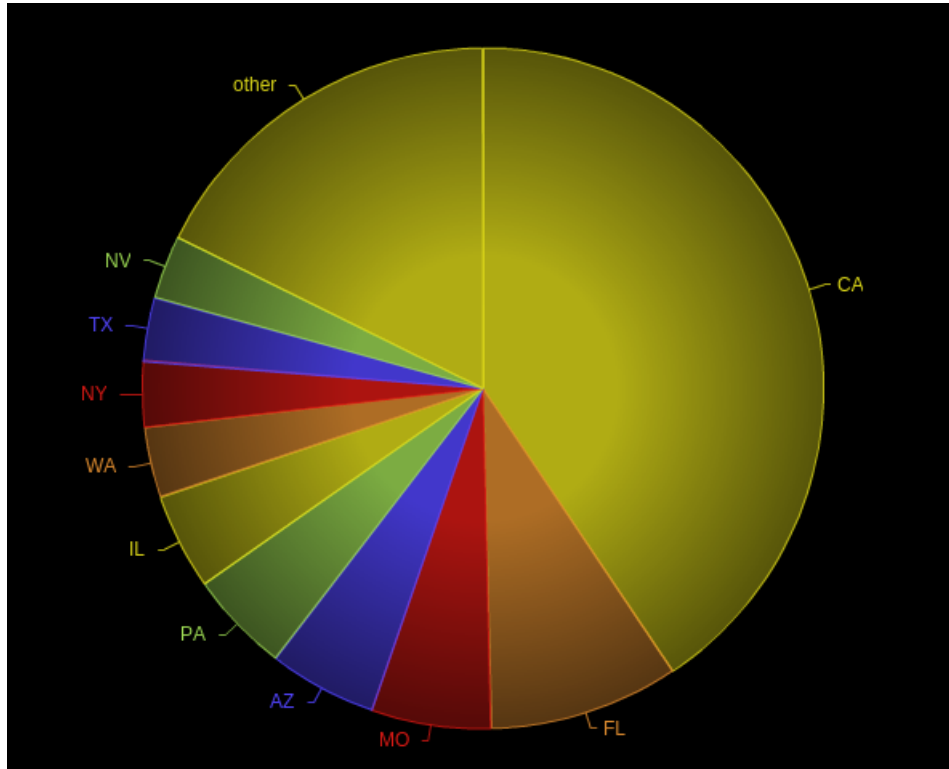


Figure 8.7. Source of U.S. Attacks (By Zip Code)

### 8.7.3. Time-Based Analysis

The target machines were tracked by month to mirror Jacob’s analysis. The results (see Figure 8.8) agree with Jacob’s findings 8.1 (Oregon, Tokyo, and Singapore see about twice as much traffic). Why are these locations more attractive to “attackers?” Perhaps these IP addresses had a previous history under a former life?

We then attempt to detect trends in cyber criminals attack strategies, by tracking the service by month (Figure 8.9). While we only have a span of about seven months, this graph shows the utility in one might analyze this metric over a multi-year effort to determine how cyber-criminals operate. While Microsoft SQL Server is the overall favorite service of “attackers” from month-to-month, we do see other services (microsoft-ds, http, ms-wbt-server) trading positions monthly.



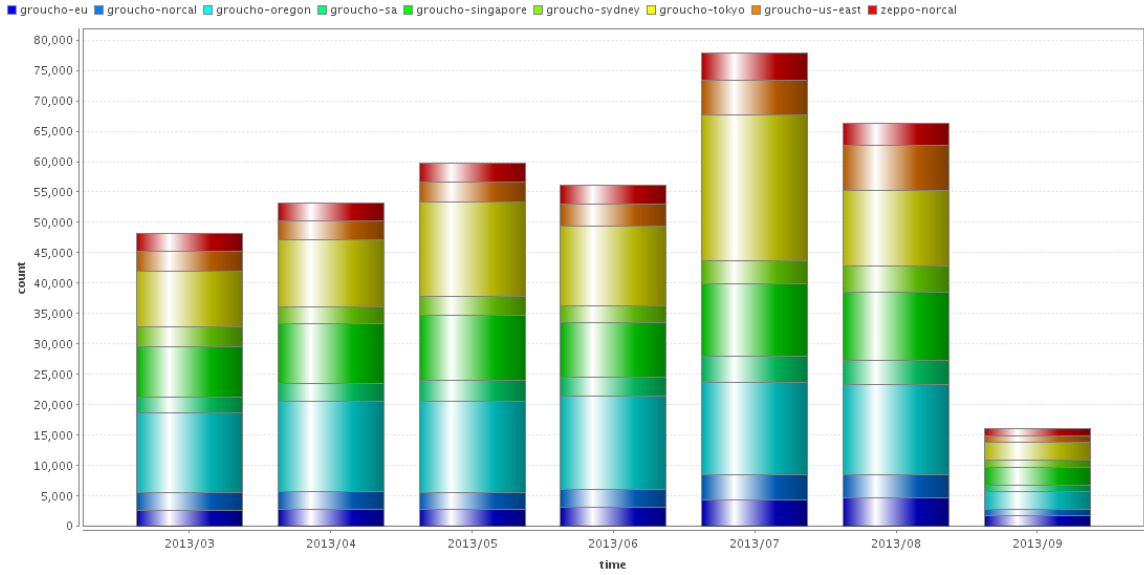


Figure 8.8. Target Machine By Month

We then examine the top countries per month. All hold their positions during the chosen time interval with China being the clear-leader. However, Iran seems to spike with 8x as much traffic in July of 2013 (Figure 8.10). “Googling” for news around that time we find that the United States expanded their sanctions July 1st of that year [31]. Is this just an anomaly, or could this news story have some correlation?

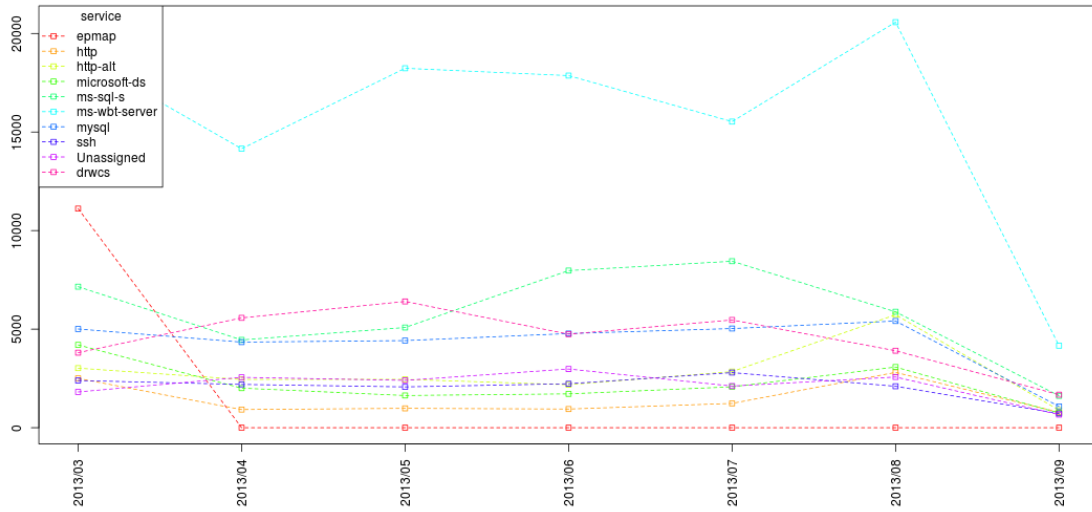


Figure 8.9. Service By Month

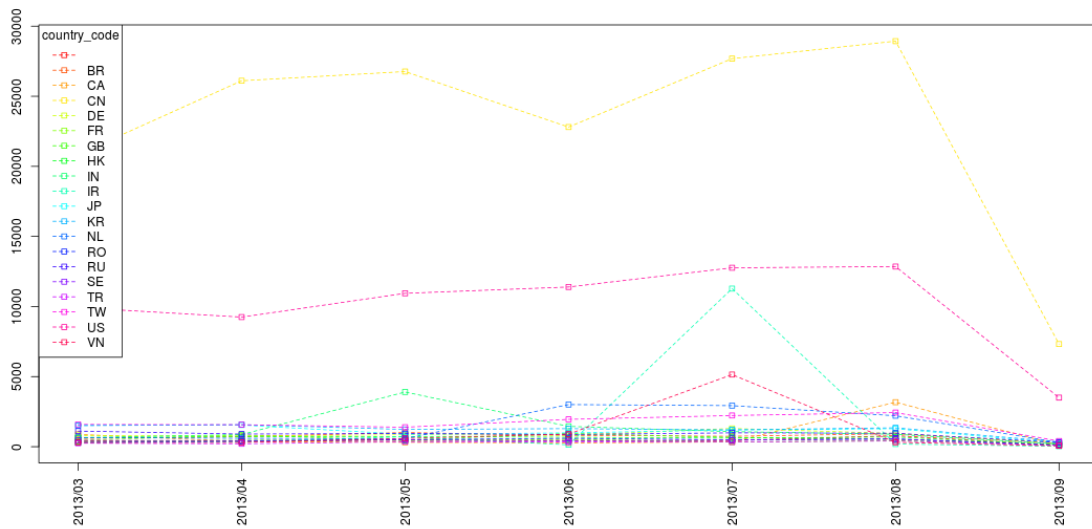
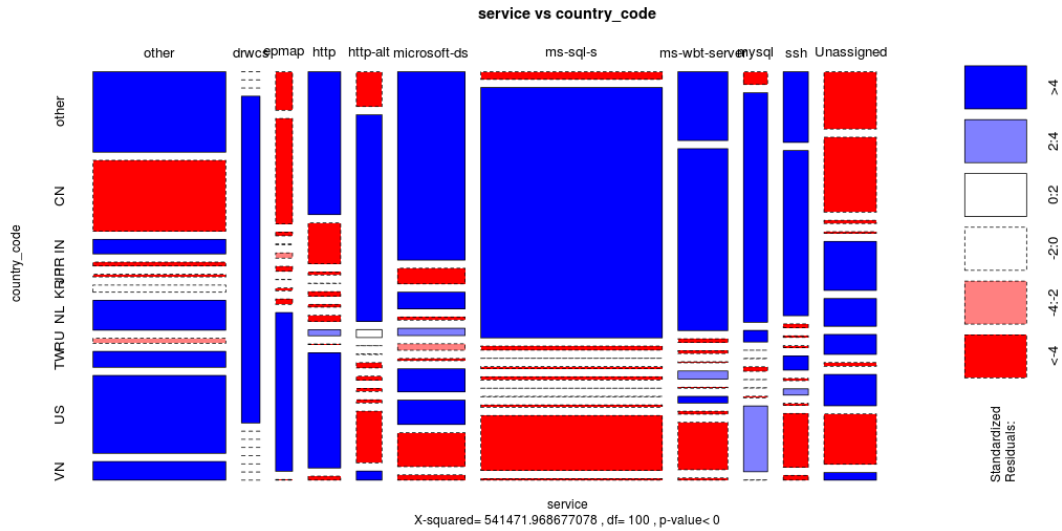


Figure 8.10. Country By Month

### 8.7.4. Two-Way Categorical Analysis

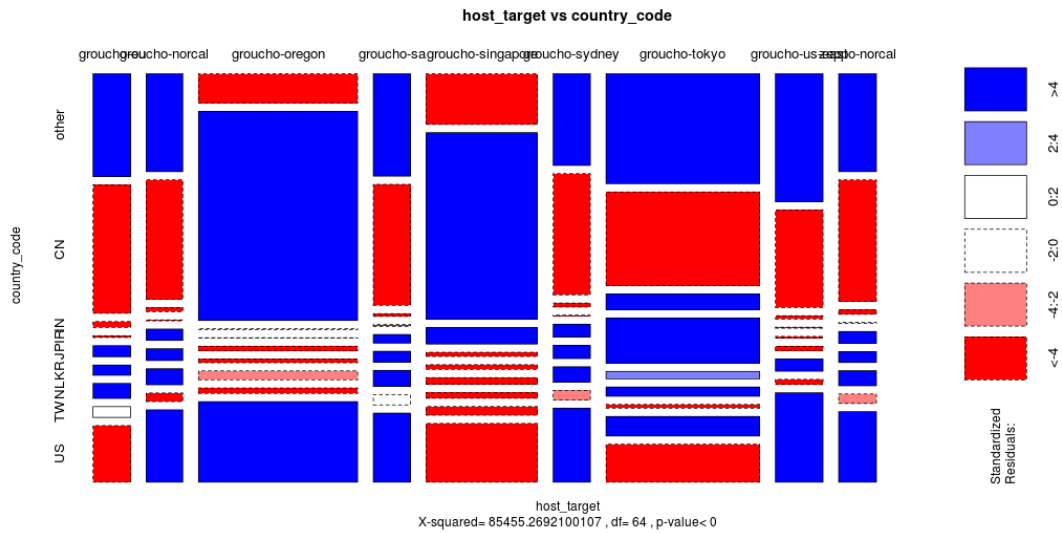
As detailed in the previous two datasets, Chi-squared statistical relevance in relation to two categorical variables can be extrapolated from the “Mosaic Plot.”



Over-represented	Under-represented
<i>China:</i> http-alt (8080), ms-sql-s ms-wbt-server, ssh	<i>China:</i> http, microsoft-ds
<i>India:</i> microsoft-ds	<i>United States:</i> microsoft-ds, ms-sql-s, ms-wbt-server, ssh
<i>Taiwan:</i> microsoft-ds	<i>India:</i> http
<i>South Korea:</i> ssh	<i>Iran:</i> http
<i>United States:</i> http	<i>South Korea:</i> http

Figure 8.11. Country vs. Service

By configuring for a two-way categorical analysis of ‘Country Code’ vs. ‘Service’, we are able to explore which “attack” strategies specific countries prefer (Figure 8.11). China seems to strongly prefer http-alt, ms-sql-s, ms-wbt-server, and ssh as a means of “attack.” The United States, on the other hand, had a strong preference for http as well as “other” means.



Over-represented	Under-represented
<i>China</i> : singapore	<i>China</i> : norcal, eu, sa, sydney, tokyo,us
<i>U.S.</i> : norcal, oregon, sydney,us	<i>U.S.</i> : eu, singapore, tokyo

Figure 8.12. Country vs. Machine

Another interesting plot for analysis is “attacker” vs “attacked.” Configuring for ‘Country Code’ vs. ‘Host Target’ (Figure 8.12) we find a number of results with statistical relevance. For the most part countries stay within their geographic location (e.g. Chinese criminals attacking Singapore, and United States criminals attacking U.S.-based servers). Outside of this general rule, China does possess an over-representation of “attacks” against “groucho-oregon.”

### 8.8. Predictor Summary

As previously discussed, our intention in performing this analysis is to demonstrate dynamic iterative analysis using the framework, not to make any substantive claims. In pursuit of such, we discussed predictors for the honeypot dataset using the framework visualizations. With a firm grasp of these predictors and trust in the

dataset, a researcher can attempt to evaluate risk and relative probability of “attack.” In order to get to a level of trust in the dataset, one might need to create a larger sample-set or tailor the collection strategy towards the business process to be targeted. With regards to this dataset, the strongest predictors we found in our research are delineated below, however, there are likely others to be found.

- **Country:** China, United States, Iran
- **China Localities:** Beijing, Jiangsu Sheng, GungZhao, Tian Shui (appr.), and Shanghai
- **U.S. Localities:** California: Walnut, CA (91789), Los Angeles, CA (90017), and Sunnyvale, CA (94085)
- **Target Machine:** oregon, singapore, and tokyo
- **Service:** ms-sql-s (1433), http (80), and epmap (135)
- **Protocol:** TCP
- **Country vs. Service:** *China:* http-alt , ms-sql-s, ms-wbt-server, ssh *India* : microsoft-ds. *United States:* http *Taiwan:* microsoft-ds *South Korea:* ssh
- **Service By Month:** Inconclusive
- **Country By Month:** July, abnormally high spike in activity from Iran

## 8.9. Follow-On Research

Mirroring the collection strategy for this dataset, follow-up research can be performed using an Amazon Web Services (AWS) HoneyNet. Because AWS [39] generously allows for honeypots in their free-tiered usage terms, setting up a controlled

experiment with well-defined parameters would provide a significant research opportunity. Control should be had to evenly disperse the honeypots geographically as well as limit the environment of the machines. This could be expanded beyond the initial iptable attributes collected by Blander. By monitoring the data on an ongoing basis, researchers could gain a greater understanding of how cyber criminals operate.

## Chapter 9

### CONCLUSION AND FUTURE WORK

Within this praxis we demonstrated increased efficiencies for analysis of a diverse set of security data using a reusable framework. After pointing the framework at the data and defining attributes of interest, one can dynamically and iteratively drill into the data. Such a framework could provide increased efficiencies to both corporate and educational research, enabling decision makers to come to isolate risk factors quicker without expending significant resources. A distributed research framework could also serve as a starting point for collaboration in research endeavors.

In our concluding remarks we evaluate the reusability aspect of the framework (Section 9.1) as well as the benefits over traditional analysis (Section 9.2). We also scope out the current perceived limitations of the framework (Section 9.3). Finally, we conclude by discussing planned usage of the framework, as well as ponder desirable improvements to the framework. (Section 9.4)

#### **9.1. Reusability Analysis**

Using our prototype framework and three distinctly different datasets, we drew several conclusions. Data from each set flowed through the exact same web process and HTML presentation layer without requiring any changes. The only change required was some upfront configuration to point it to the SQL data and populate the HTML controls.

By applying the framework to three distinctly different datasets, we not only demonstrate the reusable nature from a software perspective, but from a utility

standpoint. We analyzed datasets that range the spectrum of security relevant issues. From consumer-related privacy data (Breach) to webmaster-related software decisions (CMS) as well as intrusion detection modeling (Honeybot), the framework can be applied within most constructs. The range of visualizations and utility contained therein undoubtedly make the framework applicable to other sciences outside security research ranging from genetics to epidemiology.

## 9.2. Benefits Over Traditional Analysis

While custom analysis still has its uses, there are clearly benefits that the framework provides over traditional methodologies. We discuss these benefits below.

- **Shortened Analysis Time:** Because custom visualizations don't need to be created, the analyst spends less time developing these front-ends. Analysis time is only an act of importing the data into the framework, and configuring dataset specific controls.
- **Iterative Dynamic Analysis:** An analyst can explore the data through filtering and chart-specific controls that alter visualizations according to the user's wants. This allows the researcher to explore attributes before fully vesting himself in a projected conclusion.
- **Comparative Analysis:** The analyst can utilize side-by-side portaling features and dynamic odds ratio plots to perform case-control study comparative analysis.
- **Collaborative Environment:** Because data and processes are stored, saved, and operated on a remote server, researchers now have a means to collaborate on and share datasets. As the research network expands, users and groups can be configured with different privileges and permissions. Additionally, any work



done on visualizations by external entities can be pulled back into the base framework.

- **Deeper Analysis:** Because more interactions can be looked at without significant costs, this framework allows researchers to perform a deeper more profound analysis given limited resources (time and cost).
- **Publication:** Because of its web nature, the framework intrinsically allows researchers the ability to publish their research if they choose to do so.
- **Security:** The RapidAnalytics webserver requires authentication before any webprocesses can be executed. If important, user passwords and privileges can be configured, by the administrator to provide a security layer.
- **Easily Leverage New Core Development:** Because the core framework is linked to external products (RapidAnalytics, RapidMiner, R) with significant active development, any fixes or new features that are made, can be leveraged simply by upgrading to a new release of the product.
- **Increased Dependability:** As these visualizations get utilized over and over again and problems with them get resolved, their behavior becomes dependable in nature.

### 9.3. Limitations

One limitation that becomes obviously apparent to any user is the inability to save settings. When one transitions between chart types, this becomes immediately obvious. In order to solve this, both the filtered data and selected table could be saved as cookies. The data could then be reconstructed on the new form thus improving the user experience.

Also further limiting its utility from a usability standpoint, feedback is limited during the chart-generation process. Indeed, a true percent complete might be difficult to obtain. However, it is important to at least inform the user that the process is still running. Additionally, if the webprocess fails, error information should be presented to the user which right now is limited.

In addition, we have demonstrated usage on over one million records for the CMS dataset “Odds Ratio Plot.” Larger datasets with millions of records might necessitate, however, transitioning to a nonrelational database. This would widen its application for big data.

Finally, an involved static code analysis and penetration effort should be performed on the framework before disseminating it further.

#### **9.4. Future Work**

The Security Economics Lab at University of Tulsa plans to apply and evolve the framework to multiple ongoing research efforts. As previously discussed in Section 7.6, it will be used to track CMS data over time. It may also be applied to ongoing bitcoin and banking data collection.

We don’t, however, intend to limit its usage to efforts at the University of Tulsa. In support of this, we’ve made it freely available by hosting it on the SMU web servers with detailed deployment instructions contained herein within this praxis (Section A). Our hopes are by making it freely available through a permissible Apache license that other organizations (both educational and corporate alike) will use it for multiple research applications.

Beyond research efforts, long-term, we envision that the user-interface could further be improved. Beyond the nonfunctional usability improvements discussed in the previous section (status and performance), more statically relevant visualizations

(e.g. Pareto, Normal, Half-normal plots) could be added to expand its utility. Also, the back end could be improved to automate both the creation of the user-interface options as well as the importing of data into the database. If the configuration of the framework for a given dataset was migrated to web-based forms, research opportunities would have limited upfront costs and a true distributed research environment could be formed.

## Appendix A

### Deployment

#### **A.1. Deployment Overview**

This section describes the instructions and necessary requirements for setting up the reusable security economics framework. Installation instructions are targeted for an Ubuntu Linux / MySQL environment, however, could be tailored to a different operating system if desired.

#### **A.2. Requirements/Dependencies**

1. MYSQL v5.5 or other SQL server installation
2. Windows/Linux Machine
3. JRE 1.7 or higher
4. RapidAnalytics 1.3.008
5. RapidMiner 5.3
6. R v2.15.2 or higher

#### **A.3. Download Locations**

1. **Rapidminer v5.3:**

<http://lyle.smu.edu/~lsykalski/downloads/rapidminer.tar.gz>

**2. Rapidminer v5 Repository:**

<http://lyle.smu.edu/~lsykalski/downloads/rapidminer5repos.tar.gz>

**3. RapidAnalytics Installer v1.3.008:**

<http://lyle.smu.edu/~lsykalski/downloads/RapidAnalyticsInstaller.zip>

**4. RapidAnalytics Preinstalled v1.3.008:**

<http://lyle.smu.edu/~lsykalski/downloads/RapidAnalytics.tar.gz>

**5. Breach & RA Database:**

<http://lyle.smu.edu/~lsykalski/downloads/breach.sql.gz>

**6. CMS Database:**

<http://lyle.smu.edu/~lsykalski/downloads/cms.sql.gz>

**7. Honeypot Database:**

<http://lyle.smu.edu/~lsykalski/downloads/honeypot.sql.gz>

**8. Website:**

<http://lyle.smu.edu/~lsykalski/downloads/sececon.tar.gz>

**9. Import Scripts:**

<http://lyle.smu.edu/~lsykalski/downloads/import.tar.gz>

#### **A.4. RapidAnalytics Overview**

RapidAnalytics (RA) is the web-server which serves the RapidMiner processes into web services (charts/tables/etc.) that are viewable by a user through a URL. RA serves as the core component of this framework. It has attachments to an installed RapidAnalytics repository, where RapidMiner processes can be collaborated on. In the deployed version of this framework, this should reside on the a web-server or some place accessible from an external node.

## A.5. RapidMiner Overview

RapidMiner (RM) is the web-service design tool that resides on the design station. RapidMiner is a multi-platform Java-based application whose role is as the design component of the RapidMiner suite. Using the design view, one can transform inputs in a meaningful way to create new outputs using flow-based components and routing. The outputs of RM can be file, tabular or chart-based and many visualization options are available for presenting the data in a meaningful way.

## A.6. Installation

### A.6.1. Java Installation

From an Ubuntu command-line terminal type the following:

```
$sudo apt-get update $sudo apt-get install default-jre
```

### A.6.2. R Installation

The framework requires installation of R on the server as well as development-machines. A version of R 2.15.2 or higher is required. Additionally, a version of the package epitools is required. Instructions are provided below for Ubuntu to be executed from an Ubuntu terminal

```
$sudo apt-get update
$sudo apt-get install r-base r-base-core r-base-dev
sudo $R
>install.packages("rJava")           >install.packages("JavaGD")
>install.packages("epitools")
```

### A.6.3. MySQL Installation

This brief installation excerpt for MySQL is taken from UBUNTU.com [43]

Now, to install MySQL, run the following command from a terminal prompt (the first command is optional if you are unsure if you have mysql 5.5 or want to reinstall the package):

```
$ sudo apt-get purge mysql-client-core-5.5
$ sudo apt-get install mysql-server
$ sudo apt-get install mysql-client
```

During the installation process you will be prompted to enter a password for the MySQL root user. Once the installation is complete, the MySQL server should be started automatically. You can run the following command from a terminal prompt to check whether the MySQL server is running:

```
$sudo netstat -tap |grep mysql—
```

When you run this command, observe the following line or something similar:

```
tcp 0 0 localhost.localdomain:mysql *.* LISTEN -
```

If the server is not running correctly, you should restart with the following

```
$ sudo /etc/init.d/mysql restart
```

You may then edit the `/etc/mysql/my.cnf` file to configure the basic settings: (log file, port number, etc.) if desired.

---

```

$>mysql -u root
FLUSH PRIVILEGES;
/* Now reset/update your password */
>SET PASSWORD FOR root@'localhost' = PASSWORD('password');
/*If you have a mysql root account that can connect from everywhere*/
/*you should also do: */
>UPDATE mysql.user SET Password=PASSWORD('newpwd') WHERE User='root';
/*Alternate Method:*/
>USE mysql;
>UPDATE user SET Password = PASSWORD('newpwd')
  WHERE Host = 'localhost' AND User = 'root';
/*And if you have a root account that can access from everywhere:*/
>USE mysql;
>UPDATE user SET Password = PASSWORD('newpwd')
  WHERE Host = '%' AND User = 'root';

```

---

Figure A.1. MySQL Privilege Update Script

In order to verify the privileges are set correctly, run through the following sequence detailed below in Figure A.1.

#### A.6.4. RapidMiner Installation

A prepackaged copy of RapidMiner with the necessary extensions (Reporting & R) come preloaded at download location #1 and #2 above. Installation instructions follow. Alternatively, one can download a new version of RapidMiner (RapidMiner studio) with necessary plugins from <https://rapidminer.com/>. However, the model processes would need to be converted, and no support for the framework is claimed.



If downloading a newer version, we would recommend downloading the Community Version as it is an open-source free implementation.

If downloading directly, follow the following directions to install and verify installation by launching RapidMiner:

```
$gunzip rapidminer5repos.tar.gz
$mv rapidminer5repos.tar /
$star -xvf rapidminer5repos.tar
$gunzip rapidminer.tar.gz
$mv rapidminer.tar /intended/RM/Install/Location
$star -xvf rapidminer.tar
$cd /intended/RM/Install/Location/rapidminer
$./scripts/RapidMinerGUI.
```

#### A.6.5. RapidAnalytics Installation

A prepackaged copy of the RapidAnalytics installer is provided at download location #3 above. Alternatively, one can download the installer (also provided at download location #2) and run the jar file to install the application (java -jar RapidAnalytics-CE-Installer-1.3.008.jar). This installation method is not preferred but provides insight into typical installation methodologies for the product. Alternatively, one can download a new version of RapidAnalytics (now called RapidMiner Server) from the RapidMiner product website [37]. However, files may need to be converted, and no support is claimed for newer versions. If downloading a newer version, we would recommend downloading the Community Version.

##### *A.6.5.1. RapidAnalytics Method #1 Installation*

If following recommended procedure, download the tar-ball directly from download location #3, and follow the below directions:

```
$gunzip rapidanalytics.tar.gz
$mv rapidanalytics.tar /intended/RA/Install/Location
$gunzip rapidanalytics.tar.gz
$tar -xvf rapidanalytics.tar
$cd /intended/RA/Install/Location/rapidanalytics/bin
$./run.sh
```

#### *A.6.5.2. RapidAnalytics Method #2 Installation*

If you decide to use the installer, use the command: (java -jar RapidAnalytics-CE-Installer-1.3.008.jar) and follow the steps below outlined in Figures A.2, A.3, A.4, and A.5 respectively. Next, ensure that the R shared object library files (.so) are copied into the bin directory of RapidAnalytics. Please note that this could be in many locations, but most commonly for Ubuntu, R is installed to /usr/share/local/lib/R. You'll also need to copy the plugins from rapidminer/lib/plugins to the RA bin directory. Also ensure the variables are set correctly by configuring the system settings variables to what is shown in Figure A.6.

#### A.6.6. RapidAnalytics Repository Installation

The repositories containing the processes, services, and data for both the breach, cms, & honeypot dataset can be downloaded at download locations #5 and #6, and #7 respectively. Additionally, utility import scripts are provided at download location #9. Please note that the breach database also additionally contains the RapidAnalytics Tables containing the processes and other associated data necessary

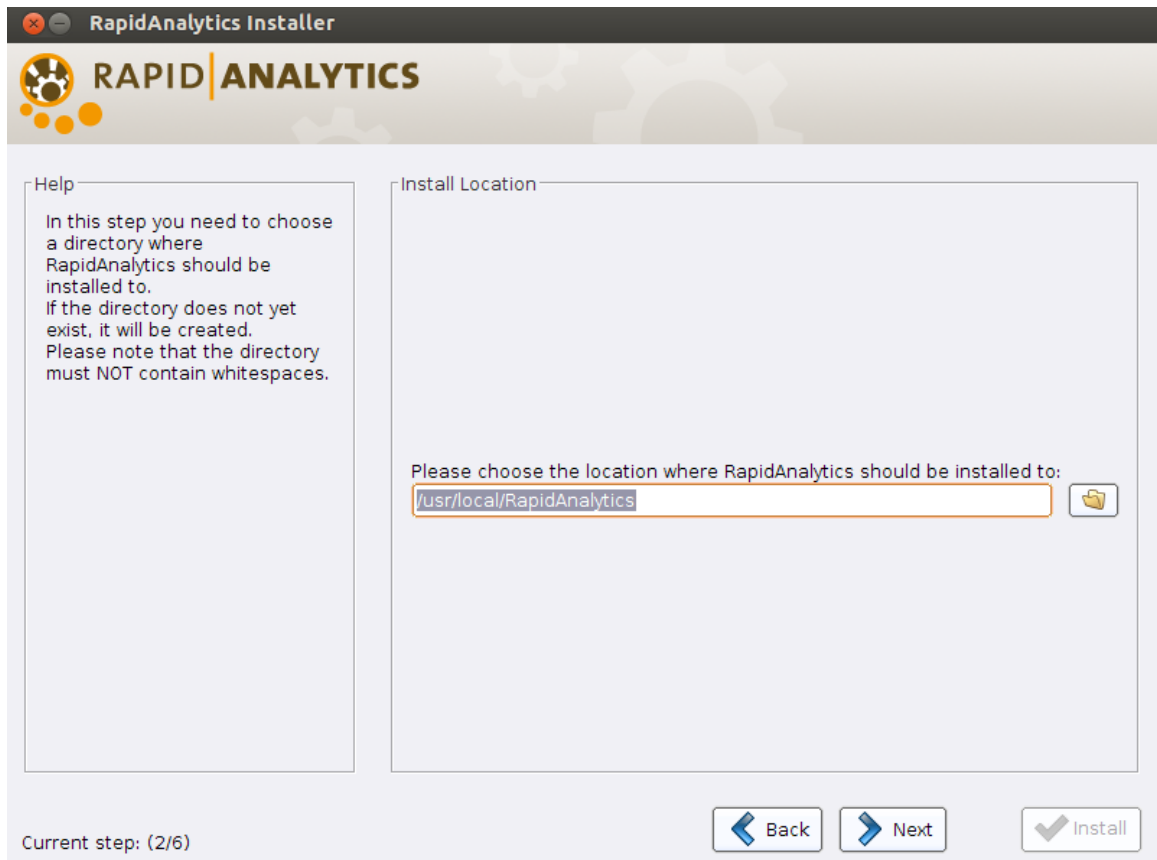


Figure A.2. RA Installer Step#1

for the repository tables. After downloading, please follow the below instructions to install the databases.

```
$mysql -u root -p
mysql>CREATE DATABASE breach;
mysql>CREATE DATABASE cms;
mysql>quit
$gunzip breach.sql.gz
$mysql -u root -p breach <breach.sql
```

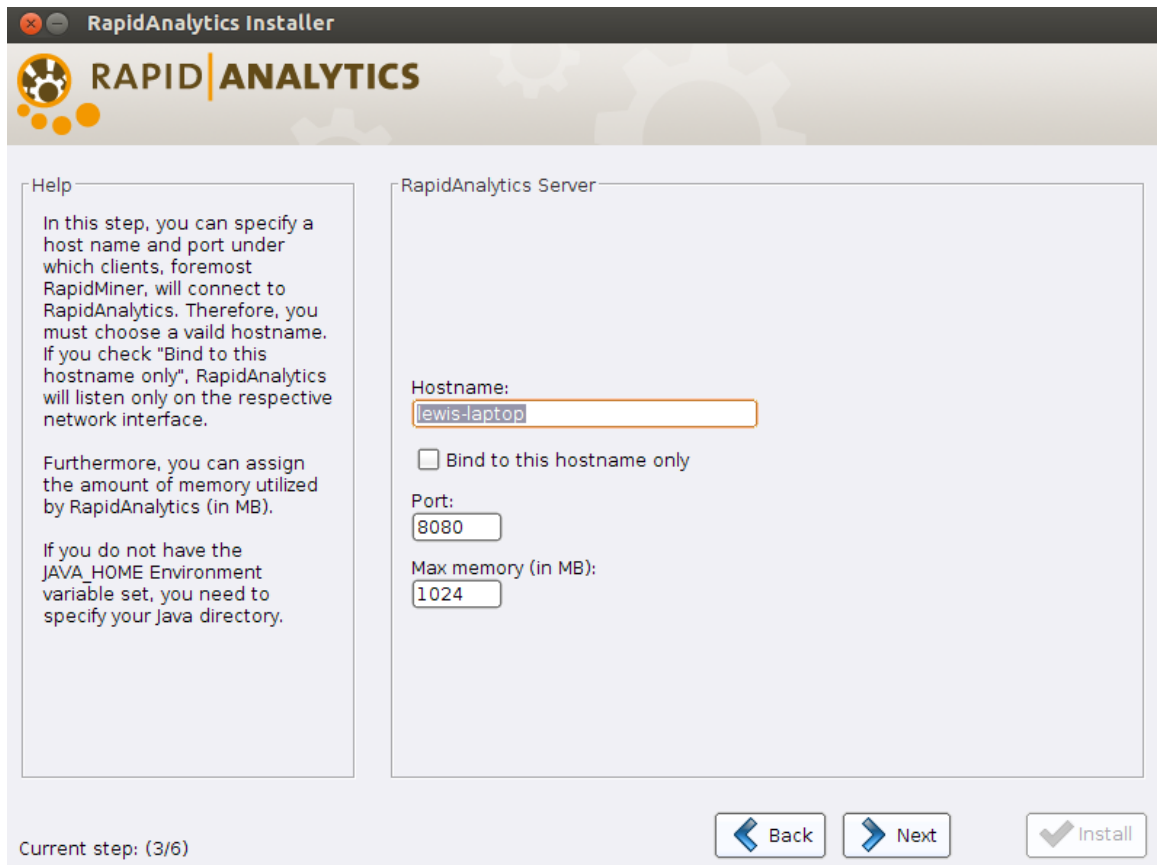


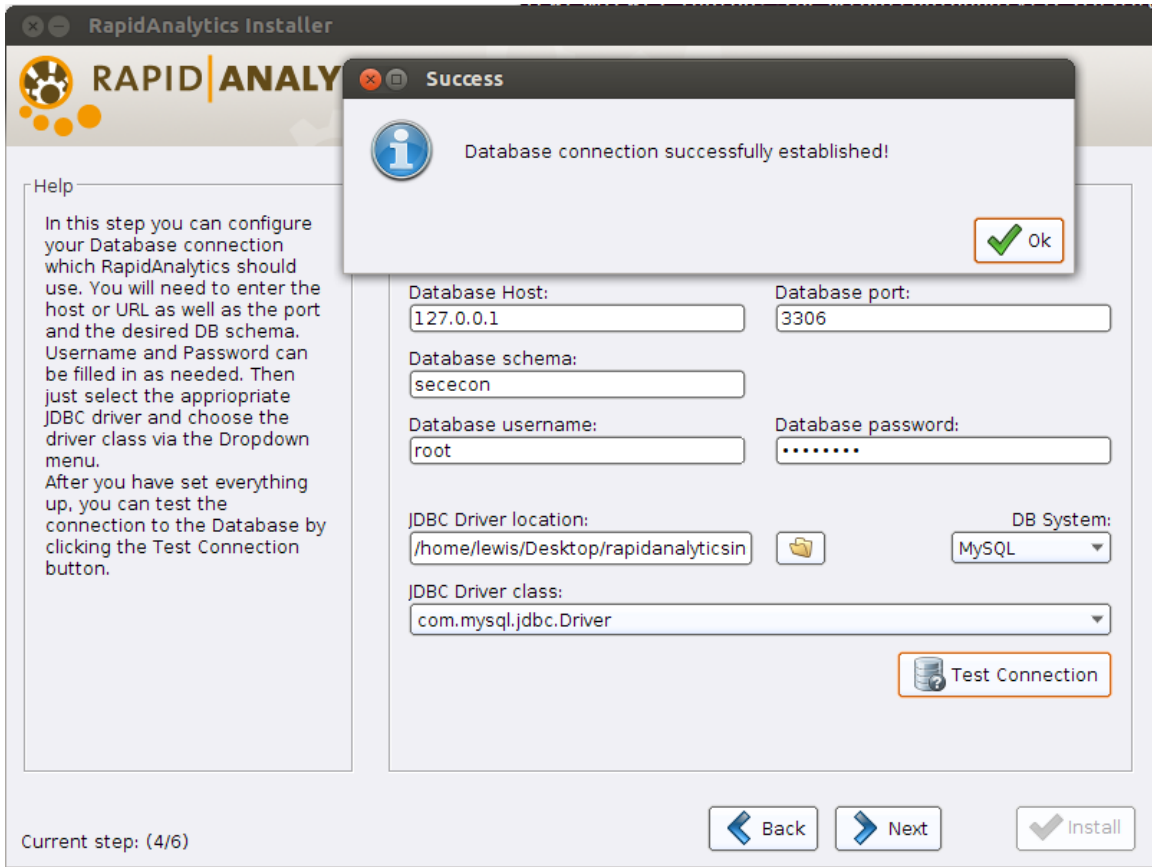
Figure A.3. RA Installer Step#2

```
$gunzip cms.sql.gz  
$mysql -u root -p cms <cms.sql
```

#### A.6.7. RapidAnalytics Startup

The RapidAnalytics startup script will then need to put into the start-up configuration so it can be started automatically on machine reboot.

Figure A.4. RA Installer Step#3



```
$ln -s /etc/rc.d/init.d/myscript  
/intended/RA/Install/Location/rapidanalytics/bin/run.sh
```

Finally, to test the installation, reboot the computer, and enter the URL

```
http://<ipaddress>:8080/RA
```

from another machine. The user should then be presented with the RapidAnalytics login screen. The default password for the rapidanalytics installation is "abc123".

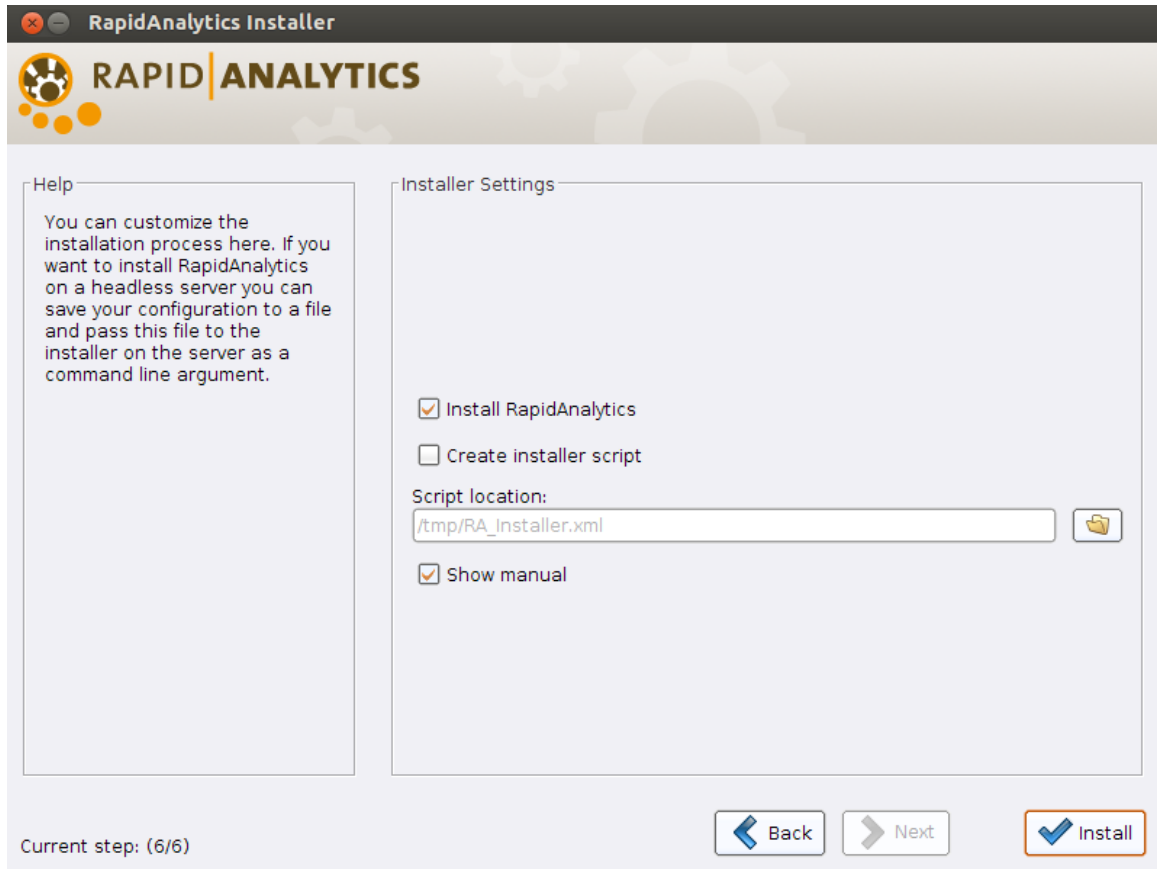


Figure A.5. RA Installer Step#4

#### A.6.8. Website Setup

The deployable web framework HTML and JavaScript code can be downloaded at download location #8. Once downloaded and unzipped to the apache location (e.g. `/var/lib/www/html`), ensure that proper permissions exist for all users (e.g. `chmod -R go+r sececon`). Also ensure all lower-level directories have proper execute permissions (e.g. `chmod go+x lower-directory`). Once complete ensure that the variables referenced below are set correctly to indicate the proper IP-addresses of both the RA-webserver and the mysql server respectively. Finally, test your installation by accessing `intro.html` first locally and then remotely. A RapidAnalytics web-portal

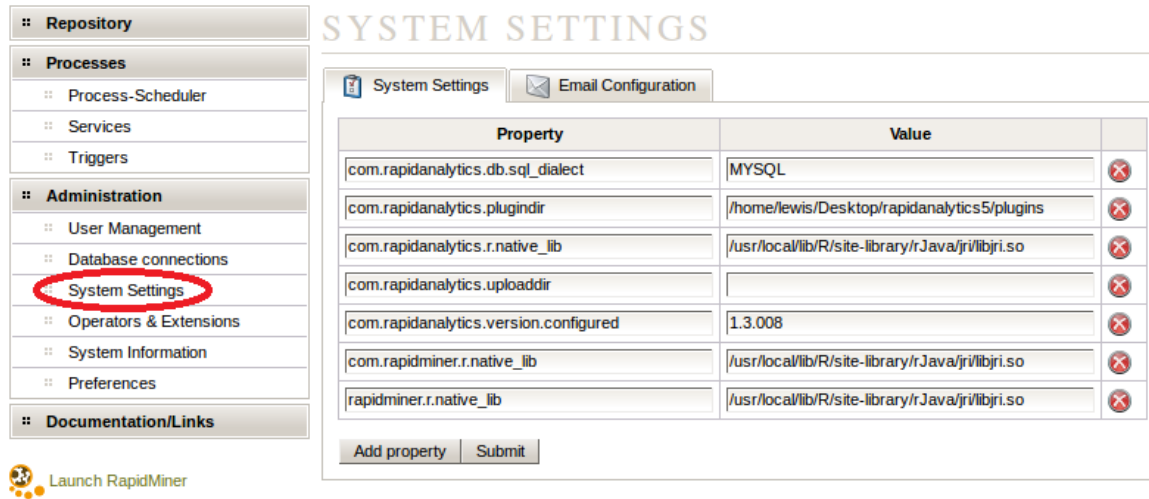


Figure A.6. RA System Settings

should be presented when you access the breach-data (pie.html). Enter the rapidanalytics password as given in the previous section and ensure the pie-chart populates. Also, ensure that the R-enabled RapidMiner processes work (odds, mosaic, box) by going to the web-portal.

```

\\In common.js
var RA_WEBSERVER = "lewis-laptop:8080";

\\In specific.js (2 places)
DB_URL="jdbc:mysql://127.0.0.1:3306/cms";
DB_URL="jdbc:mysql://127.0.0.1:3306/breach";

```

## A.6.9. Troubleshooting

### A.6.9.1. Access Denied

**Problem:** RapidAnalytics yields Access denied for user 'user'@'localhost' error when attempting to view visualizations

**Solution:** Connections.xml must be defined in the root home directory: ~/.RapidMiner5/connections.xml. This should contain a cached encoded version of the password for each connection (cms and breach). Example provided below. This is required to be present .

```
<?xml version="1.0" encoding="UTF-8" standalone="no"?>
<jdbc-entries \ key="4z07g9DlPmca3M0xkdXQeupUaMSMlDEN">
<field-entry>
<name>cms</name>
<system>MySQL</system>
<host>localhost</host>
<port>3306</port>
<database>cms</database>
<user>root</user>
<password>WvUuua0ndcc=</password>
<properties/>
</field-entry>
</jdbc-entries>
```

#### A.6.9.2. Extension Failure

**Problem:** Extensions will not work (reporting, R, etc.)



**Solution:** Verify extensions are loaded in the Operator and Extensions pane within the RapidAnalytics web-server. If not, ensure they (rapidminer-Parallel Processing-5.3.000.jar, rapidminer-R Extension-5.3.000.jar, and rapidminer-Reporting-5.3.000.jar) are present in /opt/rapidanalytics/plugins

#### *A.6.9.3. Shared Library Failure*

**Problem:** Complaints of libjmx.so or libjni.so in RapidAnalytics console or attempted R package installation.

**Solution:** One of the packages was not installed correctly or the path is not correct in the system settings. Verify RapidAnyanalytics/Java pathing of R in run.sh as well as the path the the libjni.so in the system settings within the RapidAnalytics web-portal. Additionally, verify R-installation. If necessary, reinstall R.

#### *A.6.9.4. Stale Images*

**Problem:** Web-Page Images are being cached.

**Solution:** Clear web-caching setting in browser. (browser specific)

## Appendix B

### Extending the Framework

#### **B.1. Overview**

This appendix covers extending the framework to utilize new data for analysis or for creating new visualizations to supplement existing data. Detailed Instructions follow.

#### **B.2. New Datasets**

Relevant data should be imported into a MySQL database. Data attributes should be chosen appropriately to have relevance to the architecture (e.g. a time-based attribute or geospatial attribute). Importing into MYSQL can be accomplished by using the scripts provided in the FRAMEWORK DESIGN AND ARCHITECTURE:Importing into MYSQL section.

##### B.2.1. Importing into MYSQL

Data can be imported into MYSQL or other relational SQL database using a preformatted set of scripts. Given a csv file, one can alter the provided scripts (see Figure B.1 to create table definitions as well as import the data into the tables (see Figure B.2 using the source command within MYSQL. It would have been nice to have a graphical user-interface to import data remotely through the web, however, this was scrubbed due to time considerations.

```
DROP TABLE breach;
CREATE TABLE breach(
symbol VARCHAR(12) NOT NULL
, name VARCHAR(302)
, market_cap DOUBLE
, sector VARCHAR(6)
, industry VARCHAR(141)
, capsize_coarse VARCHAR(8)
, capsize_fine VARCHAR(8)
);
```

```
DROP TABLE clean;
CREATE TABLE clean(
symbol VARCHAR(12) NOT NULL
, name VARCHAR(302)
, market_cap DOUBLE
, sector VARCHAR(6)
, industry VARCHAR(141)
, capsize_coarse VARCHAR(8)
, capsize_fine VARCHAR(8)
);
```

Figure B.1. SQL Table Definition Example

```
mkfifo my_pipe chmod 666 my_pipe cat *.csv | my_pipe
mysql| use sececon; mysql| source import.sql;
import.sql: load data local infile 'my_pipe' into table breach fields
terminated by ',' enclosed by '"' lines terminated by '\n'
(symbol, name, market_cap, sector, industry, capsize_coarse, capsize_fine);
```

Figure B.2. SQL Table Import Example

```
http://myhost.edu/website/pie.html?data=cms
```

Figure B.3. Url parameter

### B.2.2. Configuring the Framework

We can then configure the framework for a certain dataset by creating the filtering combo-boxes, source table combo-boxes, url definitions, and specific chart widgets. Functions are provided to do this easily and all code is relegated to a single file (specific.js)

While there is some rewriting of client-side code, this effort is minimal and localized to a single-file. A url-parameter, that being the dataset itself is used to initialize the portal-specific widgets and variables appropriately. Figure B.3 depicts the how the URL is formed. Based on this parameter the code can then be overloaded to reference a separate initialization function as depicted in Figure B.4.

Within the initialization function the page layout, data-specific variables, and widgets and can be defined as depicted in Figure B.5, B.6, B.7, B.8, B.9 respectively.

## B.3. New Visualizations

In order to create new visualizations, new processes need to be created and saved to the RapidAnalytics repository. (/home/admin/processes/\*) They can be based off the existing processes in the repository. Care needs to be made to save & run processes on the RapidAnalytics web-server itself instead of locally. Options are provided in the RapidMiner Toolbar to run on the RapidAnalytics web-server ("Run on RapidAnalytics Now")

```

function onSpecificInit()
{
    var param = location.search.split('data=')[1] ?
        location.search.split('data=')[1] : 'breach';
    if(param == "breach") {
        setupBreach();
    }
    else if(param == "cms") {
        setupCms();
    }
    --> else if(param == "newdata") {
    -->     setupNewData();
    --> }
}

```

Figure B.4. Entry Function to configure new dataset

```

//Add pages (Maximum 8)
document.getElementById('page1').innerHTML =
    '<a href="table.html?data=cms" title="Table" >
        </a>';
document.getElementById('page2').innerHTML =
    '<a href="mosaic.html?data=cms" title="Mosaic Plot">
        </a>';
...

```

Figure B.5. HTML page references

```

controlTable="control_day0";
treatmentTable="compromise_day0";
controlFile="csvs/cms/cms_control_csvs.tar.gz";
treatmentFile="csvs/cms/cms_compromise_csvs.tar.gz";

DB_URL="jdbc:mysql://127.0.0.1:3306/cms";
document.getElementById('logo').src = "cmslogo.png";

TIME_ATTR = "time";

```

Figure B.6. Data-Specific Variables

```

FILTER2_ATTR = "generatortype";
FILTER2_LABEL = "Gen Type:";
FILTER2_HTML="<select id='filter2ComboIDNUM'>
  <option value='All' selected='selected'>All</option>
  <option value='blogger'>blogger</option>
  <option value='drupal'>drupal</option>
  <option value='homestead'>homestead</option>
  <option value='joomla'>joomla</option>
  <option value='typo3'>typo3</option>
  <option value='vbulletin'>vbulletin</option>
  <option value='wordpress'>wordpress</option>
  <option value='zencart'>zencart</option>
</select>";

```

Figure B.7. Filter Specific Widget Definition

```
SOURCECOMBO_HTML = "<option value='breach'>breach</option>  
<option value='clean'>clean</option>";
```

Figure B.8. Source-Combo Widget Definition

```
var oddsCombo = document.getElementById('oddsCombo');  
if(oddsCombo) {  
    var opts = ["country", "generator", "generatortype",  
"server", "servertype", "tld"];  
    addComboOptions(oddsCombo, opts, "servertype");  
}
```

Figure B.9. Chart-Specific Widget Definition

All processes and their associated web-services can be viewed in the Repository Browser (see Figure B.10 within the RapidAnalytics web-portal). Processes can be triggered to run through the RapidAnalytics web-server or through RapidMiner. Once triggered to run on RapidAnalytics, their status can be viewed in the Process-Scheduler (see Figure B.11 within the RapidAnalytics web-portal.)

If successfully run, a process can then be Exported to a web-service (see Figure B.12). Macros are mapped to web-parameters. RapidAnalytics provides a couple of chart types (table, pie, line, bar charts). When using these native chart types, more display options are provided. An example is provided below in Figure B.13. Alternatively one can choose to run the web-service under JSON by selecting the JSON option. When doing such, the user is relying on your RM process to generate the imagery. This can be achieved by either using the RapidMiner Reporting Extension or

**REPOSITORY BROWSER**

Created on Jul 13, 2015 9:45:22 PM by admin

**Folder Contents**

Name	User	Modification Date
AGGREGATE_BREACH	admin	Jul 13, 2015 10:21:58 PM
BOX_PLOT_BREACH	admin	Jul 14, 2015 3:08:58 PM
DATA_BREACH	admin	Jul 13, 2015 10:30:15 PM
DEFINITION_BREACH	admin	Jul 14, 2015 1:01:44 AM
MOSAIC_PLOT_BREACH	admin	Jul 14, 2015 11:55:33 AM
ODDS_RATIO_BREACH	admin	Jul 14, 2015 2:22:58 PM
TIME_BREACH	admin	Jul 13, 2015 10:30:00 PM

Figure B.10. RA Repos Browser

**PROCESS-SCHEDULER**

Currently Defined Triggers  
There are currently no triggers defined.

Running and completed processes

Show only running processes

Process	Queue	User	Enqueue time	Execution time	Log
/home/admin/processes/MOSAIC_PLOT_BREACH	DEFAULT	admin	Jul 14, 2015 11:15:04 AM	Jul 14, 2015 11:15:04 AM - 11:15:05 AM	0 s
/home/admin/processes/MOSAIC_PLOT_BREACH	DEFAULT	admin	Jul 14, 2015 2:01:46 AM	Jul 14, 2015 2:01:46 AM - 2:01:46 AM	0 s
/home/admin/processes/MOSAIC_PLOT_BREACH	DEFAULT	admin	Jul 14, 2015 1:51:01 AM	Jul 14, 2015 1:51:01 AM - 1:51:02 AM	0 s
/home/admin/processes/MOSAIC_PLOT_BREACH	DEFAULT	admin	Jul 14, 2015 1:20:40 AM	Jul 14, 2015 1:20:40 AM - 1:20:40 AM	0 s
/home/admin/processes/MOSAIC_PLOT_BREACH	DEFAULT	admin	Jul 14, 2015 1:18:09 AM	Jul 14, 2015 1:18:09 AM - 1:18:10 AM	0 s

Figure B.11. RA Process Scheduler

by executing an R-script with the R extension. Please reference processes delineated in Chapter 3 to pick a starting-point process and associated web-service to emulate.



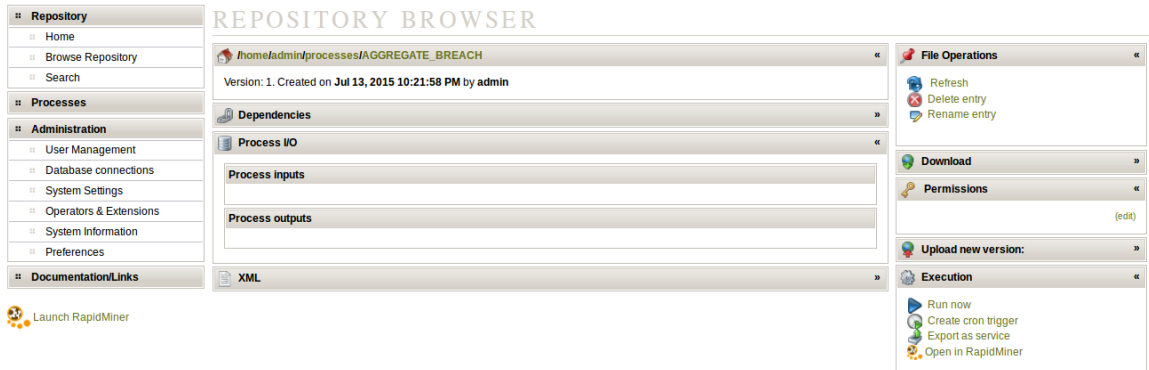


Figure B.12. RA Exporting To Service

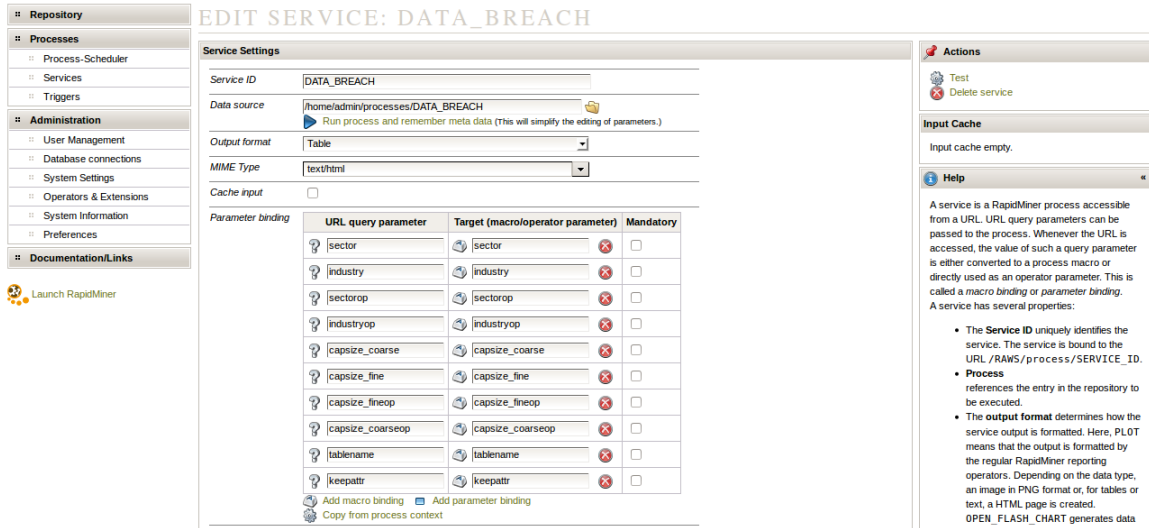


Figure B.13. RA Table Service

## Appendix C

### Reference Code

The reference HTML and Javascript code is provided below. Additional code (SQL & R snippets) for each RapidMiner process are provided within the framework design section (Section 3.)

#### C.1. HTML Code

##### C.1.1. bar.html

---

```
<!--
-- Copyright [2013] [Lewis Sykalski]
--
-- Licensed under the Apache License, Version 2.0 (the "License");
-- you may not use this file except in compliance with the License.
-- You may obtain a copy of the License at
--
-- http://www.apache.org/licenses/LICENSE-2.0
--
-- Unless required by applicable law or agreed to in writing, software
-- distributed under the License is distributed on an "AS IS" BASIS,
-- WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied.
-- See the License for the specific language governing permissions and
-- limitations under the License.
--
-- source: bar.html
-->
<html>
<link rel="stylesheet" href="style.css">
<head>
  <meta http-equiv='cache-control' content='no-cache'>
  <meta http-equiv='expires' content='0'>
  <meta http-equiv='pragma' content='no-cache'>
  <script type="text/javascript" src="canvas2image.js"> </script>
  <script type="text/javascript" src="//ajax.googleapis.com/ajax/libs/jquery/1.7.1/jquery.min.js"></script>
  <script type="text/javascript" src="html2canvas.js"></script>
  <script type="text/javascript" src="jquery.plugin.html2canvas.js"></script>
  <script type="text/javascript" src="specific.js"> </script>
```

```

<script type="text/javascript" src="common.js"> </script>
<script language="javascript" type="text/javascript">

function doDownloadChartData(name)
{
    window.location.href = 'dynamicCharts/timebar' + name + '/download.txt';
}

function setUrl(name, width, filter1Op, filter1Value, filter2Op, filter2Value, filter3Op, filter3Value, filter4Op, filter4Value,
    tableValue) {

    var aggregateIndex = document.getElementById('aggregateCombo').selectedIndex;
    var aggregateValue = document.getElementById('aggregateCombo')[aggregateIndex].value;

    var includeNAs = document.getElementById('includeNAs').checked;

    var timeTypeIndex = document.getElementById('timeTypeCombo').selectedIndex;
    var timeTypeValue = document.getElementById('timeTypeCombo')[timeTypeIndex].value;

    var TIME_TYPE = "1"; // 1 = year / 2 = month
    if (timeTypeValue == "month")
        TIME_TYPE = "2";

    var timebegin = document.getElementById('beginDate').value;
    var timeend = document.getElementById('endDate').value;

    var requestId = generateRequestID();

    var url = "http://" + RA_WEBSERVER + "/RA/faces/restricted//embed.xhtml?serviceId=" + BAR_AGGR_SERVICE + "&width=" + width +
        "&height=550&db_url=" + DB_URL + "&db_username=" + DB_USERNAME + "&outputdir=" + OUTPUT_DIR + "&filter1attr=" +
        FILTER1_ATTR + "&filter1op=" + filter1Op + "&filter1val=" + filter1Value + "&filter2attr=" + FILTER2_ATTR +
        "&filter2op=" + filter2Op + "&filter2val=" + filter2Value + "&filter3attr=" + FILTER3_ATTR + "&filter3op=" + filter3Op +
        "&filter3val=" + filter3Value + "&filter4attr=" + FILTER4_ATTR + "&filter4op=" + filter4Op + "&filter4val=" +
        filter4Value + "&aggregateattribute=" + aggregateValue + "&tablename=" + tableValue + "&chartnum=" + name +
        "&includeNAs=" + includeNAs + "&timeattribute=" + TIME_ATTR + "&timetype=" +
        TIME_TYPE+"&timebegin="+timebegin+"&timeend="+timeend+ "&requestid="+requestId;

    document.getElementById("serviceIframe" + name).src = url;
    document.getElementById("raIframe" + name).src = 'dynamicCharts/timebar' + name + '/image1.png';

    if(name == "1") {
        if(INTERVAL_TIMER1 == 0)
        {
            INTERVAL_TIMER1 = setInterval(function () {
                var time = +new Date;
                document.getElementById("raIframe" + name).src = 'dynamicCharts/timebar' + name + '/image1.png?' + time;
            }, 2500);
        }
    }
    if(name == "2") {
        if(INTERVAL_TIMER2 == 0)
        {
            INTERVAL_TIMER2 = setInterval(function () {

```

```

        var time = +new Date;
        document.getElementById("raiframe" + name).src = 'dynamicCharts/timebar' + name + '/image1.png?' + time;
    }, 2500);
    }
}
}

function configurePortalHTML(portalHTML) {
    return portalHTML;
}

function postConfigurePortal(name) {
    //do nothing
}

</script>
</head>
<body onload="onPortalInit();" >
    <div class="wrapper" align="center">
        <table cellspacing="0" cellpadding="0" border="0" id="shell">
            <tr>
                <td colspan="2">
                    
                    
                </td>
            </tr>
            <tr>
                <td width="70" valign="top">
                    <table id="navigation" title="Navigation" border="0">
                        <tr>
                            <td id="page1" />
                        </tr>
                        <tr>
                            <td id="page2" />
                        </tr>
                        <tr>
                            <td id="page3" />
                        </tr>
                        <tr>
                            <td id="page4" />
                        </tr>
                        <tr>
                            <td id="page5" />
                        </tr>
                        <tr>
                            <td id="page6" />
                        </tr>
                        <tr>
                            <td id="page7" />
                        </tr>
                        <tr>
                            <td id="page8" />
                        </tr>
                    </table>
                </td>
            </tr>
        </table>
    </div>

```

```

        <td id="page9" />
    </tr>
</table>
</td>
<td>
    <table id="content">
        <tr>
            <td>
                
                <select id="aggregateCombo" onchange="refresh(this);">
                </select>
                <select id="timeTypeCombo" onchange="refresh(this);">
                </select>
                <input id="includeNAs" type="checkbox" onchange="refresh(this);" checked />NAs
                <input id="beginDate" type="date"> -
                <input id="endDate" type="date"><br />
            </td>
            <td align="right">
                <!-- <button id="screenshotBtn" type="submit" onclick="doScreenshot();">
                    </button> -->
                <button id="downloadBtn" type="submit" onclick="doDownload();">
                    </button>
                <button id="addBtn" type="submit" onclick="addIFrame();">
                    </button>
                <button id="refreshBtn" type="submit" onclick="refreshAll();">
                    </button>
            </td>
        </tr>
        <tr>
            <td colspan="2">
                <table>
                    <tr>
                        <td id="tableCell1" align="center" nowrap>
                        </td>
                        <td id="tableCell2" align="center" nowrap>
                        </td>
                    </tr>
                </table>
            </td>
        </tr>
    </table>
</td>
</tr>
<tr>
    <td />
    <td align="center">
        <table id="banner" border="0">
            <tr>
                <td>
                    <h2>
                        Selected Dataset:</h2>
                </td>
                <td>
                    <img id="icon" height="60">
                </td>
            </tr>
        </table>
    </td>
</tr>

```

```

                </td>
                <td>
                    <img id="logo" height="60">
                </td>
            </tr>
        </table>
    </td>
</tr>
</table>
</div>
<div class="footer" align="center">
    <p>
        Copyright (c) 2013 -- contact: <a href="http://lyle.smu.edu/~lsykalski/">Lewis Sykalski</a>
        <a href="http://tylermoore.ens.utulsa.edu/">Tyler Moore</a></p>
    </div>
</body>
</html>

```

---

## C.1.2. boxplot.html

---

```

<!--
-- Copyright [2013] [Lewis Sykalski]
--
-- Licensed under the Apache License, Version 2.0 (the "License");
-- you may not use this file except in compliance with the License.
-- You may obtain a copy of the License at
--
--    http://www.apache.org/licenses/LICENSE-2.0
--
-- Unless required by applicable law or agreed to in writing, software
-- distributed under the License is distributed on an "AS IS" BASIS,
-- WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied.
-- See the License for the specific language governing permissions and
-- limitations under the License.
--
-- source: boxplot.html
-->
<html>
<link rel="stylesheet" href="style.css">
<head>
    <meta http-equiv='cache-control' content='no-cache'>
    <meta http-equiv='expires' content='0'>
    <meta http-equiv='pragma' content='no-cache'>
    <script type="text/javascript" src="canvas2image.js"> </script>
    <script type="text/javascript" src="//ajax.googleapis.com/ajax/libs/jquery/1.7.1/jquery.min.js"></script>
    <script type="text/javascript" src="html2canvas.js"></script>
    <script type="text/javascript" src="jquery.plugin.html2canvas.js"></script>
    <script type="text/javascript" src="specific.js"> </script>
    <script type="text/javascript" src="common.js"> </script>
    <script language="javascript" type="text/javascript">

```

```

function doDownloadChartData(name)
{
    window.location.href = 'dynamicCharts/box' + name + '/download.txt';
}

function setUrl(name, width, filter1Op, filter1Value, filter2Op, filter2Value, filter3Op, filter3Value, filter4Op, filter4Value,
    tableValue) {

    var inputIndex1 = document.getElementById('boxComboX').selectedIndex;
    var input1 = document.getElementById('boxComboX')[inputIndex1].value;
    var inputIndex2 = document.getElementById('boxComboY').selectedIndex;
    var input2 = document.getElementById('boxComboY')[inputIndex2].value;

    var logGraph = "FALSE";
    var logPlot = document.getElementById('logPlot').checked;
    if (logPlot) {
        logGraph = "TRUE";
    }

    var includeNAs = document.getElementById('includeNAs').checked;

    var includeNAsIndex = 2;
    if (includeNAs) {
        includeNAsIndex = "1";
    }

    var timebegin = document.getElementById('beginDate').value;
    var timeend = document.getElementById('endDate').value;

    var requestId = generateRequestID();

    var url = "http://" + RA_WEBSERVER + "/RA/faces/restricted//embed.xhtml?serviceId=" + BOXPLOT_SERVICE + "&width=" + width +
        "&height=550&db_url=" + DB_URL + "&db_username=" + DB_USERNAME + "&outputdir=" + OUTPUT_DIR + "&filter1attr=" +
        FILTER1_ATTR + "&filter1op=" + filter1Op + "&filter1val=" + filter1Value + "&filter2attr=" + FILTER2_ATTR +
        "&filter2op=" + filter2Op + "&filter2val=" + filter2Value + "&filter3attr=" + FILTER3_ATTR + "&filter3op=" + filter3Op +
        "&filter3val=" + filter3Value + "&filter4attr=" + FILTER4_ATTR + "&filter4op=" + filter4Op + "&filter4val=" +
        filter4Value + "&ctl_tablename=clean" + "&nonctl_tablename=breach" + "&chartnum=" + name + "&loggraph=" + logGraph +
        "&includeNAs=" + includeNAsIndex + "&input1=" + input1 + "&input2=" +
        input2 + "&timebegin=" + timebegin + "&timeend=" + timeend + "&timeattribute=" + TIME_ATTR + "&requestid=" + requestId;

    document.getElementById("serviceIframe" + name).src = url;

    document.getElementById("raiframe" + name).src = 'dynamicCharts/box' + name + '/' + requestId + '.png';

    if(name == "1") {
        if (INTERVAL_TIMER1 != 0) {
            clearInterval(INTERVAL_TIMER1);
        }
        INTERVAL_TIMER1 = setInterval(function () {
            var time = +new Date;
            document.getElementById("raiframe" + name).src = 'dynamicCharts/box' + name + '/' + requestId + '.png';
        }, 2500);
    }
}

```

```

if(name == "2") {
    if(INTERVAL_TIMER2 != 0) {
        clearInterval(INTERVAL_TIMER2);
    }
    INTERVAL_TIMER2 = setInterval(function () {
        var time = +new Date;
        document.getElementById("raiframe" + name).src = 'dynamicCharts/box' + name + '/' + requestId + '.png';
    }, 2500);
}
}

function configurePortalHTML(portalHTML) {
    return portalHTML;
}

function postConfigurePortal(name) {
    //do nothing
}
</script>
</head>
<body onload="onPortalInit();">
    <div class="wrapper" align="center">
        <table cellspacing="0" cellpadding="0" border="0" id="shell">
            <tr>
                <td colspan="2">
                    
                    
                </td>
            </tr>
            <tr>
                <td width="70" valign="top">
                    <table id="navigation" title="Navigation" border="0">
                        <tr>
                            <td id="page1" />
                        </tr>
                        <tr>
                            <td id="page2" />
                        </tr>
                        <tr>
                            <td id="page3" />
                        </tr>
                        <tr>
                            <td id="page4" />
                        </tr>
                        <tr>
                            <td id="page5" />
                        </tr>
                        <tr>
                            <td id="page6" />
                        </tr>
                        <tr>
                            <td id="page7" />
                        </tr>
                    </table>
                </td>
            </tr>
        </table>
    </div>

```



```

        <td id="page8" />
    </tr>
    <tr>
        <td id="page9" />
    </tr>
</table>
</td>
<td>
    <table id="content">
        <tr>
            <td>
                
                <select id="boxComboX" onchange="refresh(this);">
                </select>
                <select id="boxComboY" onchange="refresh(this);">
                </select>
                <input id="includeNAs" type="checkbox" onchange="refresh(this);" /> NAs
                <input id="logPlot" type="checkbox" onchange="refresh(this);" /> Log Plot
                <input id="beginDate" type="date"> -
                <input id="endDate" type="date"><br />
            </td>
            <td align="right">
                <!-- <button id="screenshotBtn" type="submit" onclick="doScreenshot();">
                </button> -->
                <button id="downloadBtn" type="submit" onclick="doDownload();">
                </button>
                <button id="addBtn" type="submit" onclick="addIFrame();">
                </button>
                <button id="refreshBtn" type="submit" onclick="refreshAll();">
                </button>
            </td>
        </tr>
        <tr>
            <td colspan="2">
                <table>
                    <tr>
                        <td id="tableCell1" align="center" nowrap>
                        </td>
                        <td id="tableCell2" align="center" nowrap>
                        </td>
                    </tr>
                </table>
            </td>
        </tr>
    </table>
</td>
</tr>
<tr>
    <td />
    <td align="center">
        <table id="banner" border="0">
            <tr>
                <td>
                    <h2>

```

```

                Selected Dataset:</h2>
            </td>
            <td>
                <img id="icon" height="60">
            </td>
            <td>
                <img id="logo" height="60">
            </td>
        </tr>
    </table>
</td>
</tr>
</table>
</div>
<div class="footer" align="center">
    <p>
        Copyright (c) 2013 -- contact: <a href="http://lyle.smu.edu/~lsykalski/">Lewis Sykalski</a>
        <a href="http://tylermoore.ens.utulsa.edu/">Tyler Moore</a></p>
    </div>
</body>
</html>

```

---

### C.1.3. geospatial.html

---

```

<!--
-- Copyright [2013] [Lewis Sykalski]
--
-- Licensed under the Apache License, Version 2.0 (the "License");
-- you may not use this file except in compliance with the License.
-- You may obtain a copy of the License at
--
-- http://www.apache.org/licenses/LICENSE-2.0
--
-- Unless required by applicable law or agreed to in writing, software
-- distributed under the License is distributed on an "AS IS" BASIS,
-- WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied.
-- See the License for the specific language governing permissions and
-- limitations under the License.
--
-- source:geospatial.html
-->
<html>
<link rel="stylesheet" href="style.css">
<head>
    <meta http-equiv='cache-control' content='no-cache'>
    <meta http-equiv='expires' content='0'>
    <meta http-equiv='pragma' content='no-cache'>
    <script type="text/javascript" src="canvas2image.js"> </script>
    <script type="text/javascript" src="//ajax.googleapis.com/ajax/libs/jquery/1.7.1/jquery.min.js"></script>
    <script type="text/javascript" src="html2canvas.js"></script>
    <script type="text/javascript" src="jquery.plugin.html2canvas.js"></script>

```

```

<script type="text/javascript" src="specific.js"> </script>
<script type="text/javascript" src="common.js"> </script>
<script language="javascript" type="text/javascript">

    function setUrl(name, width, filter1Op, filter1Value, filter2Op, filter2Value, filter3Op, filter3Value, filter4Op, filter4Value,
        tableValue) {
        var url = "http://lyle.smu.edu";
        if (tableValue in GEOSPATIAL_URL) {
            url = GEOSPATIAL_URL[tableValue];
        }
        document.getElementById("raIframe" + name).src = url;
    }

    function configurePortalHTML(portalHTML) {
        return portalHTML;
    }

    function postConfigurePortal(name) {
        document.getElementById("countryCombo" + name).disabled = true;
        document.getElementById("tldCombo" + name).disabled = true;
        document.getElementById("serverTypeCombo" + name).disabled = true;
        document.getElementById("generatorTypeCombo" + name).disabled = true;
    }

</script>
</head>
<body onload="onPortalInit();">
    <div class="wrapper" align="center">
        <table cellspacing="0" cellpadding="0" border="0" id="shell">
            <tr>
                <td colspan="2">
                    
                    
                </td>
            </tr>
            <tr>
                <td width="70" valign="top">
                    <table id="navigation" title="Navigation" border="0">
                        <tr>
                            <td id="page1" />
                        </tr>
                        <tr>
                            <td id="page2" />
                        </tr>
                        <tr>
                            <td id="page3" />
                        </tr>
                        <tr>
                            <td id="page4" />
                        </tr>
                        <tr>
                            <td id="page5" />
                        </tr>
                    </table>
                </td>
            </tr>
        </table>
    </div>

```

```

        <td id="page6" />
    </tr>
    <tr>
        <td id="page7" />
    </tr>
    <tr>
        <td id="page8" />
    </tr>
    <tr>
        <td id="page9" />
    </tr>
</table>
</td>
<td>
    <table id="content">
        <tr>
            <td>
                
            </td>
            <td align="right">
                <button id="screenshotBtn" type="submit" onclick="doScreenshot();">
                    </button>
                <button id="downloadBtn" type="submit" onclick="doDownload();">
                    </button>
                <button id="addBtn" type="submit" onclick="addIFrame();">
                    </button>
                <button id="refreshBtn" type="submit" onclick="refreshAll();">
                    </button>
            </td>
        </tr>
        <tr>
            <td colspan="2">
                <table>
                    <tr>
                        <td id="tableCell1" align="center" nowrap>
                        </td>
                        <td id="tableCell2" align="center" nowrap>
                        </td>
                    </tr>
                </table>
            </td>
        </tr>
    </table>
</td>
</tr>
<tr>
    <td />
    <td align="center">
        <table id="banner" border="0">
            <tr>
                <td>
                    <h2>
                        Selected Dataset:</h2>
                </td>
            </tr>
        </table>
    </td>
</tr>

```

```

        <td>
            <img id="icon" height="60">
        </td>
        <td>
            <img id="logo" height="60">
        </td>
    </tr>
</table>
</td>
</tr>
</table>
</div>
<div class="footer" align="center">
    <p>
        Copyright (c) 2013 -- contact: <a href="http://lyle.smu.edu/~lsykalski/">Lewis Sykalski</a>
        <a href="http://tylermoore.ens.utulsa.edu/">Tyler Moore</a></p>
    </div>
</body>
</html>

```

---

#### C.1.4. help.html

---

```

<!--
-- Copyright [2013] [Lewis Sykalski]
--
-- Licensed under the Apache License, Version 2.0 (the "License");
-- you may not use this file except in compliance with the License.
-- You may obtain a copy of the License at
--
--   http://www.apache.org/licenses/LICENSE-2.0
--
-- Unless required by applicable law or agreed to in writing, software
-- distributed under the License is distributed on an "AS IS" BASIS,
-- WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied.
-- See the License for the specific language governing permissions and
-- limitations under the License.
--
-- source: help.html
-->

<html>

<link rel="stylesheet" href="style.css">
<head>

<script type="text/javascript" src="canvas2image.js"> </script>
<script type="text/javascript" src="//ajax.googleapis.com/ajax/libs/jquery/1.7.1/jquery.min.js"></script>
<script type="text/javascript" src="html2canvas.js"></script>
<script type="text/javascript" src="jquery.plugin.html2canvas.js"></script>
<script type="text/javascript" src="specific.js"> </script>
<script type="text/javascript" src="common.js"> </script>

```

```

<script language="javascript" type="text/javascript">

function setUrl(name, width, filter10p, filter1Value, filter20p, filter2Value, filter30p, filter3Value, filter40p, filter4Value,
    tableValue)
{
    var url =
        "http://"+RA_WEBSERVER+"/RA/faces/restricted//embed.xhtml?serviceId="+DEFINITION_SERVICE+"&width="+width+"&height=550&tablename="+tableValue+"&db
    document.getElementById("raIframe"+name).src=url;
}

function configurePortalHTML(portalHTML)
{
    return portalHTML.replace(/BGCOLOR/g, "white");
}

function postConfigurePortal(name)
{
    //do nothing
}

</script>

</head>
<body onload="onPortalInit();">

<table cellspacing="0" cellpadding="0" border="0"
bgcolor="black" id="shell">
    <tr>
        <td colspan="2" bgcolor="black">
            <table id="banner" border="0">
                <tr>
                    <td>
                        <img id="logo" width=600 height=50>
                    </td>
                </tr>
            </table>
        </td>
    </tr>
    <tr>
        <td bgcolor="black" width="70" valign="top">
            <table id="navigation" title="Navigation" border="0">
                <tr><td id="page1" ></td></tr>
                <tr><td id="page2" ></td></tr>
                <tr><td id="page3" ></td></tr>
                <tr><td id="page4" ></td></tr>
                <tr><td id="page5" ></td></tr>
                <tr><td id="page6" ></td></tr>
                <tr><td id="page7" ></td></tr>
                <tr><td id="page8" ></td></tr>
            </table>
        </td>
        <td bgcolor="black">

```

```

<table id="content">
  <tr bgcolor="black">
    <td>
    </td>
    <td align="right">
      <button id="screenshotBtn" type="submit" onclick="doScreenshot();"></button>
      <button id="downloadBtn" type="submit" onclick="doDownload();"></button>
      <button id="addBtn" type="submit" onclick="addIFrame();"></button>
      <button id="refreshBtn" type="submit" onclick="refreshAll();"></button>
    </td>
  </tr>

  <tr>
    <td colspan="2">
      <table>
        <tr>
          <td id="tableCell1" align="center" nowrap></td>
          <td id="tableCell2" align="center" nowrap></td>
        </tr>
      </table>
    </td>
  </tr>
</table>
</td>
</tr>
</table>
</body>

</html>

```

---

## C.1.5. intro.html

---

```

//
// source: intro.html
//

<html>

<link rel="stylesheet" href="style.css">
<head>

</head>
<body style="text-align:center;">


<h1></h1>
<h3>contact: <a href="mailto:lsykalski@smu.edu">lsykalski@smu.edu</a> <a href="mailto:tmoore@smu.edu">tmoore@smu.edu</a></h3>

```

```
<h1>NEXT <a href="start.html" title="Next page"></a></h1>
</html>
```

---

## C.1.6. line.html

---

```
<!--
-- Copyright [2013] [Lewis Sykalski]
--
-- Licensed under the Apache License, Version 2.0 (the "License");
-- you may not use this file except in compliance with the License.
-- You may obtain a copy of the License at
--
--   http://www.apache.org/licenses/LICENSE-2.0
--
-- Unless required by applicable law or agreed to in writing, software
-- distributed under the License is distributed on an "AS IS" BASIS,
-- WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied.
-- See the License for the specific language governing permissions and
-- limitations under the License.
--
-- source: line.html
-->
<html>
<link rel="stylesheet" href="style.css">
<head>
  <meta http-equiv='cache-control' content='no-cache'>
  <meta http-equiv='expires' content='0'>
  <meta http-equiv='pragma' content='no-cache'>
  <script type="text/javascript" src="canvas2image.js"> </script>
  <script type="text/javascript" src="//ajax.googleapis.com/ajax/libs/jquery/1.7.1/jquery.min.js"></script>
  <script type="text/javascript" src="html2canvas.js"></script>
  <script type="text/javascript" src="jquery.plugin.html2canvas.js"></script>
  <script type="text/javascript" src="specific.js"> </script>
  <script type="text/javascript" src="common.js"> </script>
  <script language="javascript" type="text/javascript">

    function doDownloadChartData(name)
    {
      window.location.href = 'dynamicCharts/timeline' + name + '/download.txt';
    }

    function setUrl(name, width, filter1Op, filter1Value, filter2Op, filter2Value, filter3Op, filter3Value, filter4Op, filter4Value,
      tableValue) {

      var aggregateIndex = document.getElementById('aggregateCombo').selectedIndex;
      var aggregateValue = document.getElementById('aggregateCombo')[aggregateIndex].value;

      var includeNAs = document.getElementById('includeNAs').checked;

      var timeTypeIndex = document.getElementById('timeTypeCombo').selectedIndex;
```



```

var timeTypeValue = document.getElementById('timeTypeCombo')[timeTypeIndex].value;

var TIME_TYPE = "1"; // 1 = year / 2 = month
if (timeTypeValue == "month")
    TIME_TYPE = "2";

var resultlimit = document.getElementById('resultlimit').value;

var url = "http://" + RA_WEBSERVER + "/RA/faces/restricted//embed.xhtml?serviceId=" + LINE_SERVICE + "&width=" + width +
    "&height=550&db_url=" + DB_URL + "&db_username=" + DB_USERNAME + "&outputdir=" + OUTPUT_DIR + "&filter1attr=" +
    FILTER1_ATTR + "&filter1op=" + filter1Op + "&filter1val=" + filter1Value + "&filter2attr=" + FILTER2_ATTR +
    "&filter2op=" + filter2Op + "&filter2val=" + filter2Value + "&filter3attr=" + FILTER3_ATTR + "&filter3op=" + filter3Op
    + "&filter3val=" + filter3Value + "&filter4attr=" + FILTER4_ATTR + "&filter4op=" + filter4Op + "&filter4val=" +
    filter4Value + "&aggregateattribute=" + aggregateValue + "&tablename=" + tableValue + "&chartnum=" + name +
    "&timeattribute=" + TIME_ATTR + "&timetype=" + TIME_TYPE + "&includeNAs=" + includeNAs + "&resultlimit=" + resultlimit;

document.getElementById("serviceIframe" + name).src = url;
document.getElementById("raIframe" + name).src = 'dynamicCharts/timeline' + name + '/test.png';

if(name == "1") {
    if(INTERVAL_TIMER1 == 0)
    {
        INTERVAL_TIMER1 = setInterval(function () {
            var time = +new Date;
            document.getElementById("raIframe" + name).src = 'dynamicCharts/timeline' + name + '/test.png?' + time;
        }, 2500);
    }
}
if(name == "2") {
    if(INTERVAL_TIMER2 == 0)
    {
        INTERVAL_TIMER2 = setInterval(function () {
            var time = +new Date;
            document.getElementById("raIframe" + name).src = 'dynamicCharts/timeline' + name + '/test.png?' + time;
        }, 2500);
    }
}
}

function configurePortalHTML(portalHTML) {
    return portalHTML;
}

function postConfigurePortal(name) {
    //do nothing
}

</script>
</head>
<body onload="onPortalInit();">
    <div class="wrapper" align="center">
        <table cellspacing="0" cellpadding="0" border="0" id="shell">
            <tr>
                <td colspan="2">

```

```

        
        
    </td>
</tr>
<tr>
    <td width="70" valign="top">
        <table id="navigation" title="Navigation" border="0">
            <tr>
                <td id="page1">
                </td>
            </tr>
            <tr>
                <td id="page2">
                </td>
            </tr>
            <tr>
                <td id="page3">
                </td>
            </tr>
            <tr>
                <td id="page4">
                </td>
            </tr>
            <tr>
                <td id="page5">
                </td>
            </tr>
            <tr>
                <td id="page6">
                </td>
            </tr>
            <tr>
                <td id="page7">
                </td>
            </tr>
            <tr>
                <td id="page8">
                </td>
            </tr>
            <tr>
                <td id="page9" />
            </tr>
        </table>
    </td>
    <td>
        <table id="content">
            <tr>
                <td>
                    
                    <select id="aggregateCombo" onchange="refresh(this);">
                    </select>
                    <select id="timeTypeCombo" onchange="refresh(this);">
                    </select>
                    </select> NAs
                </td>
            </tr>
        </table>
    </td>
</tr>

```

```



```

```
</div>
</body>
</html>
```

---

## C.1.7. mosaic.html

---

```
<!--
-- Copyright [2013] [Lewis Sykalski]
--
-- Licensed under the Apache License, Version 2.0 (the "License");
-- you may not use this file except in compliance with the License.
-- You may obtain a copy of the License at
--
--   http://www.apache.org/licenses/LICENSE-2.0
--
-- Unless required by applicable law or agreed to in writing, software
-- distributed under the License is distributed on an "AS IS" BASIS,
-- WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied.
-- See the License for the specific language governing permissions and
-- limitations under the License.
--
-- source: mosaic.html
-->
<html>
<link rel="stylesheet" href="style.css">
<head>
  <meta http-equiv='cache-control' content='no-cache'>
  <meta http-equiv='expires' content='0'>
  <meta http-equiv='pragma' content='no-cache'>
  <script type="text/javascript" src="canvas2image.js"> </script>
  <script type="text/javascript" src="//ajax.googleapis.com/ajax/libs/jquery/1.7.1/jquery.min.js"></script>
  <script type="text/javascript" src="html2canvas.js"></script>
  <script type="text/javascript" src="jquery.plugin.html2canvas.js"></script>
  <script type="text/javascript" src="specific.js"> </script>
  <script type="text/javascript" src="common.js"> </script>
  <script language="javascript" type="text/javascript">

    function doDownloadChartData(name)
    {
      window.location.href = 'dynamicCharts/mosaic' + name + '/download.txt';
    }

    function setUrl(name, width, filter1Op, filter1Value, filter2Op, filter2Value, filter3Op, filter3Value, filter4Op, filter4Value,
      tableValue) {
      var mosaicIndex1 = document.getElementById('mosaicCombo1').selectedIndex;
      var mosaicValue1 = document.getElementById('mosaicCombo1')[mosaicIndex1].value;
      var mosaicIndex2 = document.getElementById('mosaicCombo2').selectedIndex;
      var mosaicValue2 = document.getElementById('mosaicCombo2')[mosaicIndex2].value;

      var residualIndex = document.getElementById('residualCombo').selectedIndex;
```

```

var residualType = document.getElementById('residualCombo')[residualIndex].value;

var includeNAs = document.getElementById('includeNAs').checked;

var resultlimit1 = document.getElementById('resultlimit1').value;
var resultlimit2 = document.getElementById('resultlimit2').value;

var timebegin = document.getElementById('beginDate').value;
var timeend = document.getElementById('endDate').value;

var requestId = generateRequestID();

var url = "http://" + RA_WEBSERVER + "/RA/faces/restricted//embed.xhtml?serviceId=" + MOSAIC_SERVICE + "&width=" + width +
"&height=550&db_url=" + DB_URL + "&db_username=" + DB_USERNAME + "&outputdir=" + OUTPUT_DIR + "&filter1attr=" +
FILTER1_ATTR + "&filter1op=" + filter1Op + "&filter1val=" + filter1Value + "&filter2attr=" + FILTER2_ATTR +
"&filter2op=" + filter2Op + "&filter2val=" + filter2Value + "&filter3attr=" + FILTER3_ATTR + "&filter3op=" + filter3Op +
"&filter3val=" + filter3Value + "&filter4attr=" + FILTER4_ATTR + "&filter4op=" + filter4Op + "&filter4val=" +
filter4Value + "&tablename=" + tableValue + "&chartnum=" + name + "&input1=" + mosaicValue1 + "&input2=" +
mosaicValue2 + "&includeNAs=" + includeNAs + "&residualtype=" + residualType + "&resultlimit1=" + resultlimit1 +
"&resultlimit2=" + resultlimit2 + "&timebegin=" + timebegin + "&timeend=" + timeend + "&timeattribute=" + TIME_ATTR +
"&requestid=" + requestId;

document.getElementById("serviceIframe" + name).src = url;
document.getElementById("raIframe" + name).src = 'dynamicCharts/mosaic' + name + '/' + requestId + '.png';

if(name == "1") {
    if(INTERVAL_TIMER1 != 0) {
        clearInterval(INTERVAL_TIMER1);
    }
    INTERVAL_TIMER1 = setInterval(function () {
        var time = +new Date;
        document.getElementById("raIframe" + name).src = 'dynamicCharts/mosaic' + name + '/' + requestId + '.png';
    }, 2500);
}
if(name == "2") {
    if(INTERVAL_TIMER2 != 0) {
        clearInterval(INTERVAL_TIMER2);
    }
    INTERVAL_TIMER2 = setInterval(function () {
        var time = +new Date;
        document.getElementById("raIframe" + name).src = 'dynamicCharts/mosaic' + name + '/' + requestId + '.png';
    }, 2500);
}
}

function configurePortalHTML(portalHTML) {
    return portalHTML;
}

function postConfigurePortal(name) {
    //do nothing
}
}
</script>

```

```

</head>
<body onload="onPortalInit();" >
  <div class="wrapper" align="center">
    <table cellspacing="0" cellpadding="0" border="0" id="shell">
      <tr>
        <td colspan="2">
          
          
        </td>
      </tr>
      <tr>
        <td width="70" valign="top">
          <table id="navigation" title="Navigation" border="0">
            <tr>
              <td id="page1" />
            </tr>
            <tr>
              <td id="page2" />
            </tr>
            <tr>
              <td id="page3" />
            </tr>
            <tr>
              <td id="page4" />
            </tr>
            <tr>
              <td id="page5" />
            </tr>
            <tr>
              <td id="page6" />
            </tr>
            <tr>
              <td id="page7" />
            </tr>
            <tr>
              <td id="page8" />
            </tr>
            <tr>
              <td id="page9" />
            </tr>
          </table>
        </td>
        <td>
          <table id="content">
            <tr>
              <td>
                
                <select id="mosaicCombo1" onchange="refresh(this);">
                </select>
                (<input id="resultlimit1" type="number" min="0" max="20" value="999" required>) vs
                <select id="mosaicCombo2" onchange="refresh(this);">
                </select>(<input id="resultlimit2" type="number" min="0" max="20" value="999" required>)
                <input id="includeNAs" type="checkbox" onchange="refresh(this);" checked />NAs
                <input id="beginDate" type="date"> -
              </td>
            </tr>
          </table>
        </td>
      </tr>
    </table>
  </div>

```

```

        <input id="endDate" type="date">
        <br />
    </td>
    <td align="right">
        Residual:
        <select id="residualCombo" onchange="refresh(this);">
            <option value="pearson">Pearson's ChiSquared</option>
            <option value="deviance">Likelihood Ratio</option>
            <option value="FT">Freeman-Turkey</option>
        </select>
        <!-- <button id="screenshotBtn" type="submit" onclick="doScreenshot();">
            </button> -->
        <button id="downloadBtn" type="submit" onclick="doDownload();">
            </button>
        <button id="addBtn" type="submit" onclick="addIFrame();">
            </button>
        <button id="refreshBtn" type="submit" onclick="refreshAll();">
            </button>
    </td>
</tr>
<tr>
    <td colspan="2">
        <table>
            <tr>
                <td id="tableCell1" align="center" nowrap>
                </td>
                <td id="tableCell2" align="center" nowrap>
                </td>
            </tr>
        </table>
    </td>
</tr>
<table>
    <tr>
        <td align="center">
            <table id="banner" border="0">
                <tr>
                    <td>
                        <h2>
                            Selected Dataset:</h2>
                    </td>
                    <td>
                        <img id="icon" height="60">
                    </td>
                    <td>
                        <img id="logo" height="60">
                    </td>
                </tr>
            </table>
        </td>
    </tr>
</table>
</tr>

```

```

    </table>
</div>
<div class="footer" align="center">
  <p>
    Copyright (c) 2013 -- contact: <a href="http://lyle.smu.edu/~lsykalski/">Lewis Sykalski</a>
    <a href="http://tylermoore.ens.utulsa.edu/">Tyler Moore</a></p>
  </div>
</body>
</html>

```

---

## C.1.8. odds.html

---

```

<!--
-- Copyright [2013] [Lewis Sykalski]
--
-- Licensed under the Apache License, Version 2.0 (the "License");
-- you may not use this file except in compliance with the License.
-- You may obtain a copy of the License at
--
--   http://www.apache.org/licenses/LICENSE-2.0
--
-- Unless required by applicable law or agreed to in writing, software
-- distributed under the License is distributed on an "AS IS" BASIS,
-- WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied.
-- See the License for the specific language governing permissions and
-- limitations under the License.
--
-- source: odds.html
-->
<html>
<link rel="stylesheet" href="style.css">
<head>
  <meta http-equiv='cache-control' content='no-cache'>
  <meta http-equiv='expires' content='0'>
  <meta http-equiv='pragma' content='no-cache'>
  <script type="text/javascript" src="canvas2image.js"> </script>
  <script type="text/javascript" src="//ajax.googleapis.com/ajax/libs/jquery/1.7.1/jquery.min.js"></script>
  <script type="text/javascript" src="html2canvas.js"></script>
  <script type="text/javascript" src="jquery.plugin.html2canvas.js"></script>
  <script type="text/javascript" src="specific.js"> </script>
  <script type="text/javascript" src="common.js"> </script>
  <script language="javascript" type="text/javascript">

    function doDownloadChartData(name)
    {
      window.location.href = 'dynamicCharts/odds' + name + '/download.txt';
    }

    function setUrl(name, width, filter10p, filter1Value, filter20p, filter2Value, filter30p, filter3Value, filter40p, filter4Value,
      tableValue) {

```



```

var oddsIndex1 = document.getElementById('oddsCombo').selectedIndex;
var oddsValue1 = document.getElementById('oddsCombo')[oddsIndex1].value;

var confIndex = document.getElementById('confCombo').selectedIndex;
var confValue = document.getElementById('confCombo')[confIndex].value;

var logGraph = "FALSE";
var logPlot = document.getElementById('logPlot').checked;
if (logPlot) {
    logGraph = "TRUE";
}

var includeNAs = document.getElementById('includeNAs').checked;

var includeNAsIndex = 2;
if (includeNAs) {
    includeNAsIndex = "1";
}

var tables = tableValue.split(" vs ");

var timebegin = document.getElementById('beginDate').value;
var timeend = document.getElementById('endDate').value;

var requestId = generateRequestID();

var url = "http://" + RA_WEBSERVER + "/RA/faces/restricted//embed.xhtml?serviceId=" + ODDS_RATIO_SERVICE + "&width=" + width
+ "&height=550&db_url=" + DB_URL + "&db_username=" + DB_USERNAME + "&outputdir=" + OUTPUT_DIR + "&filter1attr=" +
FILTER1_ATTR + "&filter1op=" + filter1Op + "&filter1val=" + filter1Value + "&filter2attr=" + FILTER2_ATTR +
"&filter2op=" + filter2Op + "&filter2val=" + filter2Value + "&filter3attr=" + FILTER3_ATTR + "&filter3op=" + filter3Op
+ "&filter3val=" + filter3Value + "&filter4attr=" + FILTER4_ATTR + "&filter4op=" + filter4Op + "&filter4val=" +
filter4Value + "&ctl_tablename=" + tables[1] + "&nonctl_tablename=" + tables[0] + "&chartnum=" + name + "&loggraph=" +
logGraph + "&includeNAs=" + includeNAsIndex + "&oddsvar=" + oddsValue1 + "&confllevel=" +
confValue+"&timebegin="+timebegin+"&timeend="+timeend+"&timeattribute="+TIME_ATTR+ "&requestid="+requestId;

document.getElementById("serviceIframe" + name).src = url;

document.getElementById("raIframe" + name).src = 'dynamicCharts/odds' + name + '/' + requestId + '.png';
document.getElementById("serviceIframe" + name).src = url;

if(name == "1") {
    if (INTERVAL_TIMER1 != 0) {
        clearInterval(INTERVAL_TIMER1);
    }
    INTERVAL_TIMER1 = setInterval(function () {
        var time = +new Date;
        document.getElementById("raIframe" + name).src = 'dynamicCharts/odds' + name + '/' + requestId + '.png';
    }, 2500);
}
if(name == "2") {
    if (INTERVAL_TIMER2 != 0) {
        clearInterval(INTERVAL_TIMER2);
    }
}

```

```

    INTERVAL_TIMER2 = setInterval(function () {
        var time = +new Date;
        document.getElementById("raiframe" + name).src = 'dynamicCharts/odds' + name + '/' + requestId + '.png';
    }, 2500);
}

function configurePortalHTML(portalHTML) {
    return portalHTML;
}

function postConfigurePortal(name) {
}

</script>
</head>
<body onload="onPortalInit();" >
    <div class="wrapper" align="center">
        <table cellspacing="0" cellpadding="0" border="0" id="shell">
            <tr>
                <td colspan="2">
                    
                    
                </td>
            </tr>
            <tr>
                <td width="70" valign="top">
                    <table id="navigation" title="Navigation" border="0">
                        <tr>
                            <td id="page1" />
                        </tr>
                        <tr>
                            <td id="page2" />
                        </tr>
                        <tr>
                            <td id="page3" />
                        </tr>
                        <tr>
                            <td id="page4" />
                        </tr>
                        <tr>
                            <td id="page5" />
                        </tr>
                        <tr>
                            <td id="page6" />
                        </tr>
                        <tr>
                            <td id="page7" />
                        </tr>
                        <tr>
                            <td id="page8" />
                        </tr>
                        <tr>
                            <td id="page9" />
                    </table>
                </td>
            </tr>
        </table>
    </div>

```

```

        </tr>
    </table>
</td>
<td>
    <table id="content">
        <tr>
            <td>
                
                <select id="oddsCombo" onchange="refresh(this);">
                </select>
                Conf:
                <select id="confCombo" onchange="refresh(this);">
                    <option value="0.85">85%</option>
                    <option value="0.86">86%</option>
                    <option value="0.87">87%</option>
                    <option value="0.88">88%</option>
                    <option value="0.89">89%</option>
                    <option value="0.90">90%</option>
                    <option value="0.91">91%</option>
                    <option value="0.92">92%</option>
                    <option value="0.93">93%</option>
                    <option value="0.94">94%</option>
                    <option value="0.95" selected="true">95%</option>
                    <option value="0.96">96%</option>
                    <option value="0.97">97%</option>
                    <option value="0.98">98%</option>
                    <option value="0.99">99%</option>
                </select>
                <input id="includeNAs" type="checkbox" onchange="refresh(this);" /> NAs
                <input id="logPlot" type="checkbox" onchange="refresh(this);" /> Log Plot
                <input id="beginDate" type="date"> -
                <input id="endDate" type="date"><br />
            </td>
            <td align="right">
                <!-- <button id="screenshotBtn" type="submit" onclick="doScreenshot();" -->
                </button> -->
                <button id="downloadBtn" type="submit" onclick="doDownload();" -->
                </button>
                <button id="addBtn" type="submit" onclick="addIFrame();" -->
                </button>
                <button id="refreshBtn" type="submit" onclick="refreshAll();" -->
                </button>
            </td>
        </tr>
    </table>
    <tr>
        <td colspan="2">
            <table>
                <tr>
                    <td id="tableCell1" align="center" nowrap>
                    </td>
                    <td id="tableCell2" align="center" nowrap>
                    </td>
                </tr>
            </table>
        </td>
    </tr>
</table>

```

```

        </td>
      </tr>
    </table>
  </td>
</tr>
<tr>
  <td />
  <td align="center">
    <table id="banner" border="0">
      <tr>
        <td>
          <h2>
            Selected Dataset:</h2>
        </td>
        <td>
          <img id="icon" height="60">
        </td>
        <td>
          <img id="logo" height="60">
        </td>
      </tr>
    </table>
  </td>
</tr>
</table>
</div>
<div class="footer" align="center">
  <p>
    Copyright (c) 2013 -- contact: <a href="http://lyle.smu.edu/~lsykalski/">Lewis Sykalski</a>
    <a href="http://tylermoore.ens.utulsa.edu/">Tyler Moore</a></p>
</div>
</body>
</html>

```

---

## C.1.9. oddstime.html

---

```

<!--
-- Copyright [2013] [Lewis Sykalski]
--
-- Licensed under the Apache License, Version 2.0 (the "License");
-- you may not use this file except in compliance with the License.
-- You may obtain a copy of the License at
--
--   http://www.apache.org/licenses/LICENSE-2.0
--
-- Unless required by applicable law or agreed to in writing, software
-- distributed under the License is distributed on an "AS IS" BASIS,
-- WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied.
-- See the License for the specific language governing permissions and
-- limitations under the License.
--

```

```

-- source: odds.html
-->
<html>
<link rel="stylesheet" href="style.css">
<head>
  <meta http-equiv='cache-control' content='no-cache'>
  <meta http-equiv='expires' content='0'>
  <meta http-equiv='pragma' content='no-cache'>
  <script type="text/javascript" src="canvas2image.js"> </script>
  <script type="text/javascript" src="//ajax.googleapis.com/ajax/libs/jquery/1.7.1/jquery.min.js"></script>
  <script type="text/javascript" src="html2canvas.js"></script>
  <script type="text/javascript" src="jquery.plugin.html2canvas.js"></script>
  <script type="text/javascript" src="specific.js"> </script>
  <script type="text/javascript" src="common.js"> </script>
  <script language="javascript" type="text/javascript">

    function doDownloadChartData(name)
    {
      window.location.href = 'dynamicCharts/oddstime' + name + '/download.txt';
    }

    function setUrl(name, width, filter1Op, filter1Value, filter2Op, filter2Value, filter3Op, filter3Value, filter4Op, filter4Value,
      tableValue) {
      var oddsIndex1 = document.getElementById('oddsCombo').selectedIndex;
      var oddsValue1 = document.getElementById('oddsCombo')[oddsIndex1].value;

      var confIndex = document.getElementById('confCombo').selectedIndex;
      var confValue = document.getElementById('confCombo')[confIndex].value;

      var logGraph = "FALSE";
      var logPlot = document.getElementById('logPlot').checked;
      if (logPlot) {
        logGraph = "TRUE";
      }

      var includeNAs = document.getElementById('includeNAs').checked;

      var includeNAsIndex = 2;
      if (includeNAs) {
        includeNAsIndex = "1";
      }

      var timeTypeIndex = document.getElementById('timeTypeCombo').selectedIndex;
      var timeTypeValue = document.getElementById('timeTypeCombo')[timeTypeIndex].value;

      var TIME_TYPE = "1"; // 1 = year / 2 = month
      if (timeTypeValue == "month")
        TIME_TYPE = "2";

      var timebegin = document.getElementById('beginDate').value;
      var timeend = document.getElementById('endDate').value;

```

```

var tables = tableValue.split(" vs ");

var requestId = generateRequestID();

var url = "http://" + RA_WEBSERVER + "/RA/faces/restricted//embed.xhtml?serviceId=" + ODDS_RATIO_TIME_SERVICE + "&width=" +
width + "&height=550&db_url=" + DB_URL + "&db_username=" + DB_USERNAME + "&outputdir=" + OUTPUT_DIR + "&filter1attr="
+ FILTER1_ATTR + "&filter1op=" + filter1Op + "&filter1val=" + filter1Value + "&filter2attr=" + FILTER2_ATTR +
"&filter2op=" + filter2Op + "&filter2val=" + filter2Value + "&filter3attr=" + FILTER3_ATTR + "&filter3op=" + filter3Op
+ "&filter3val=" + filter3Value + "&filter4attr=" + FILTER4_ATTR + "&filter4op=" + filter4Op + "&filter4val=" +
filter4Value + "&ctl_tablename=" + tables[1] + "&nonctl_tablename=" + tables[0] + "&chartnum=" + name + "&loggraph=" +
logGraph + "&includeNAs=" + includeNAsIndex + "&oddsvar=" + oddsValue1 + "&conflevel=" + confValue + "&timetype=" +
TIME_TYPE+"&timebegin="+timebegin+"&timeend="+timeend+"&timeattribute="+TIME_ATTR+ "&requestid="+requestId;

document.getElementById("serviceIframe" + name).src = url;

alert("done here");

document.getElementById("raiframe" + name).src = 'dynamicCharts/oddstime' + name + '/' + requestId + '.png';
document.getElementById("serviceIframe" + name).src = url;

if(name == "1") {
    if(INTERVAL_TIMER1 != 0) {
        clearInterval(INTERVAL_TIMER1);
    }
    INTERVAL_TIMER1 = setInterval(function () {
        var time = +new Date;
        document.getElementById("raiframe" + name).src = 'dynamicCharts/oddstime' + name + '/' + requestId + '.png';
    }, 2500);
}
if(name == "2") {
    if(INTERVAL_TIMER2 != 0) {
        clearInterval(INTERVAL_TIMER2);
    }
    INTERVAL_TIMER2 = setInterval(function () {
        var time = +new Date;
        document.getElementById("raiframe" + name).src = 'dynamicCharts/oddstime' + name + '/' + requestId + '.png';
    }, 2500);
}

}

function configurePortalHTML(portalHTML) {
    return portalHTML;
}

function postConfigurePortal(name) {
}

</script>
</head>
<body onload="onPortalInit();" >
    <div class="wrapper" align="center">
        <table cellspacing="0" cellpadding="0" border="0" id="shell">
            <tr>

```

```

<td colspan="2">
    
    
</td>
</tr>
<tr>
<td width="70" valign="top">
    <table id="navigation" title="Navigation" border="0">
        <tr>
            <td id="page1" />
        </tr>
        <tr>
            <td id="page2" />
        </tr>
        <tr>
            <td id="page3" />
        </tr>
        <tr>
            <td id="page4" />
        </tr>
        <tr>
            <td id="page5" />
        </tr>
        <tr>
            <td id="page6" />
        </tr>
        <tr>
            <td id="page7" />
        </tr>
        <tr>
            <td id="page8" />
        </tr>
        <tr>
            <td id="page9" />
        </tr>
    </table>
</td>
<td>
    <table id="content">
        <tr>
            <td>
                
                <select id="oddsCombo" onchange="refresh(this);">
                </select>
                <select id="timeTypeCombo" onchange="refresh(this);">
                </select>
                Conf:
                <select id="confCombo" onchange="refresh(this);">
                    <option value="0.85">85%</option>
                    <option value="0.86">86%</option>
                    <option value="0.87">87%</option>
                    <option value="0.88">88%</option>
                    <option value="0.89">89%</option>
                    <option value="0.90">90%</option>
                </select>
            </td>
        </tr>
    </table>
</td>
</tr>
</table>

```

```

        <option value="0.91">91%</option>
        <option value="0.92">92%</option>
        <option value="0.93">93%</option>
        <option value="0.94">94%</option>
        <option value="0.95" selected="true">95%</option>
        <option value="0.96">96%</option>
        <option value="0.97">97%</option>
        <option value="0.98">98%</option>
        <option value="0.99">99%</option>
    </select>
    <input id="includeNAs" type="checkbox" onchange="refresh(this);" />NAs
    <input id="logPlot" type="checkbox" onchange="refresh(this);" />Log Plot
    <input id="beginDate" type="date"> -
    <input id="endDate" type="date"><br />
</td>
<td align="right">
    <!-- <button id="screenshotBtn" type="submit" onclick="doScreenshot();">
        </button> -->
    <button id="downloadBtn" type="submit" onclick="doDownload();">
        </button>
    <button id="addBtn" type="submit" onclick="addIframe();">
        </button>
    <button id="refreshBtn" type="submit" onclick="refreshAll();">
        </button>
</td>
</tr>
<tr>
<td colspan="2">
    <table>
        <tr>
            <td id="tableCell1" align="center" nowrap>
            </td>
            <td id="tableCell2" align="center" nowrap>
            </td>
        </tr>
    </table>
</td>
</tr>
</table>
</td>
</tr>
<tr>
<td />
<td align="center">
    <table id="banner" border="0">
        <tr>
            <td>
                <h2>
                    Selected Dataset:</h2>
            </td>
            <td>
                <img id="icon" height="60">
            </td>
        </tr>
    </table>
</td>

```



```

        <img id="logo" height="60">
      </td>
    </tr>
  </table>
</td>
</tr>
</table>
</div>
<div class="footer" align="center">
  <p>
    Copyright (c) 2013 -- contact: <a href="http://lyle.smu.edu/~lsykalski/">Lewis Sykalski</a>
    <a href="http://tylermoore.ens.utulsa.edu/">Tyler Moore</a></p>
  </div>
</body>
</html>

```

---

### C.1.10. pie.html

---

```

<!--
-- Copyright [2013] [Lewis Sykalski]
--
-- Licensed under the Apache License, Version 2.0 (the "License");
-- you may not use this file except in compliance with the License.
-- You may obtain a copy of the License at
--
-- http://www.apache.org/licenses/LICENSE-2.0
--
-- Unless required by applicable law or agreed to in writing, software
-- distributed under the License is distributed on an "AS IS" BASIS,
-- WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied.
-- See the License for the specific language governing permissions and
-- limitations under the License.
--
-- source: pie.html
-->
<html>
<link rel="stylesheet" href="style.css">
<head>
  <meta http-equiv='cache-control' content='no-cache'>
  <meta http-equiv='expires' content='0'>
  <meta http-equiv='pragma' content='no-cache'>
  <script type="text/javascript" src="canvas2image.js"> </script>
  <script type="text/javascript" src="//ajax.googleapis.com/ajax/libs/jquery/1.7.1/jquery.min.js"></script>
  <script type="text/javascript" src="html2canvas.js"></script>
  <script type="text/javascript" src="jquery.plugin.html2canvas.js"></script>
  <script type="text/javascript" src="specific.js"> </script>
  <script type="text/javascript" src="common.js"> </script>
  <script language="javascript" type="text/javascript">

    function doDownloadChartData(name)
    {

```

```

window.location.href = 'dynamicCharts/pie' + name + '/download.txt';
}

function setUrl(name, width, filter1Op, filter1Value, filter2Op, filter2Value, filter3Op, filter3Value, filter4Op, filter4Value,
    tableValue) {

    var aggregateIndex = document.getElementById('aggregateCombo').selectedIndex;
    var aggregateValue = document.getElementById('aggregateCombo')[aggregateIndex].value;

    var includeNAs = document.getElementById('includeNAs').checked;
    var resultlimit = document.getElementById('resultlimit').value;

    var timebegin = document.getElementById('beginDate').value;
    var timeend = document.getElementById('endDate').value;

    var url = "http://" + RA_WEBSERVER + "/RA/faces/restricted//embed.xhtml?serviceId=" + PIE_SERVICE + "&width=" + width +
        "&height=550&db_url=" + DB_URL + "&db_username=" + DB_USERNAME + "&outputdir=" + OUTPUT_DIR + "&filter1attr=" +
        FILTER1_ATTR + "&filter1op=" + filter1Op + "&filter1val=" + filter1Value + "&filter2attr=" + FILTER2_ATTR +
        "&filter2op=" + filter2Op + "&filter2val=" + filter2Value + "&filter3attr=" + FILTER3_ATTR + "&filter3op=" + filter3Op +
        "&filter3val=" + filter3Value + "&filter4attr=" + FILTER4_ATTR + "&filter4op=" + filter4Op + "&filter4val=" +
        filter4Value + "&aggregateattribute=" + aggregateValue + "&tablename=" + tableValue + "&chartnum=" + name +
        "&includeNAs=" + includeNAs
        + "&resultlimit=" + resultlimit + "&timebegin=" + timebegin + "&timeend=" + timeend + "&timeattribute=" + TIME_ATTR;
    document.getElementById("raIframe" + name).src = url;
}

function configurePortalHTML(portalHTML) {
    return portalHTML;
}

function postConfigurePortal(name) {
    //do nothing
}

</script>
</head>
<body onload="onPortalInit();" >
    <div class="wrapper" align="center">
        <table cellspacing="0" cellpadding="0" border="0" id="shell">
            <tr>
                <td colspan="2">
                    
                    
                </td>
            </tr>
            <tr>
                <td width="70" valign="top">
                    <table id="navigation" title="Navigation" border="0">
                        <tr>
                            <td id="page1" />
                        </tr>
                        <tr>
                            <td id="page2" />
                        </tr>
                    </table>
                </td>
            </tr>
        </table>
    </div>

```

```

<tr>
  <td id="page3" />
</tr>
<tr>
  <td id="page4" />
</tr>
<tr>
  <td id="page5" />
</tr>
<tr>
  <td id="page6" />
</tr>
<tr>
  <td id="page7" />
</tr>
<tr>
  <td id="page8" />
</tr>
<tr>
  <td id="page9" />
</tr>
</table>
</td>
<td>
  <table id="content">
    <tr>
      <td>
        
        <select id="aggregateCombo" onchange="refresh(this);">
        </select><input id="resultlimit" type="number" min="0" max="20" value="999" required>
        <input id="includeNAs" type="checkbox" onchange="refresh(this);" checked />NAs
        <input id="beginDate" type="date"> -
        <input id="endDate" type="date"><br />
      </td>
      <td align="right">
        <!-- <button id="screenshotBtn" type="submit" onclick="doScreenshot();">
          </button> -->
        <button id="downloadBtn" type="submit" onclick="doDownload();">
          </button>
        <button id="addBtn" type="submit" onclick="addIFrame();">
          </button>
        <button id="refreshBtn" type="submit" onclick="refreshAll();">
          </button>
      </td>
    </tr>
  </table>
  <tr>
    <td colspan="2">
      <table>
        <tr>
          <td id="tableCell1" align="center" nowrap>
          </td>
          <td id="tableCell2" align="center" nowrap>
          </td>
        </tr>
      </table>
    </td>
  </tr>

```

```

        </table>
      </td>
    </tr>
  </table>
</td>
</tr>
<tr>
  <td />
  <td align="center">
    <table id="banner" border="0">
      <tr>
        <td>
          <h2>
            Selected Dataset:</h2>
          </td>
          <td>
            <img id="icon" height="60">
          </td>
          <td>
            <img id="logo" height="60">
          </td>
        </tr>
      </table>
    </td>
  </tr>
</table>
</div>
<div class="footer" align="center">
  <p>
    Copyright (c) 2013 -- contact: <a href="http://lyle.smu.edu/~lsykalski/">Lewis Sykalski</a>
    <a href="http://tylermoore.ens.utulsa.edu/">Tyler Moore</a></p>
</div>
</body>
</html>

```

---

### C.1.11. start.html

---

```

<!--
-- Copyright [2013] [Lewis Sykalski]
--
-- Licensed under the Apache License, Version 2.0 (the "License");
-- you may not use this file except in compliance with the License.
-- You may obtain a copy of the License at
--
-- http://www.apache.org/licenses/LICENSE-2.0
--
-- Unless required by applicable law or agreed to in writing, software
-- distributed under the License is distributed on an "AS IS" BASIS,
-- WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied.
-- See the License for the specific language governing permissions and
-- limitations under the License.

```

```

--
-- source: start.html
-->

<html>

<link rel="stylesheet" href="style.css">
<head>

</head>
<body style="text-align:center;">

<a href="pie.html?data=cms" title="Next page">
<a href="pie.html?data=breach" title="Next page">
<h1>
<h3>contact: <a href="mailto:lsykalski@smu.edu">lsykalski@smu.edu</a> <a href="mailto:tmoore@smu.edu">tmoore@smu.edu</a></h3>

</html>

```

---

## C.1.12. table.html

```

<!--
-- Copyright [2013] [Lewis Sykalski]
--
-- Licensed under the Apache License, Version 2.0 (the "License");
-- you may not use this file except in compliance with the License.
-- You may obtain a copy of the License at
--
--   http://www.apache.org/licenses/LICENSE-2.0
--
-- Unless required by applicable law or agreed to in writing, software
-- distributed under the License is distributed on an "AS IS" BASIS,
-- WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied.
-- See the License for the specific language governing permissions and
-- limitations under the License.
--
-- source: table.html
-->

<html>
<link rel="stylesheet" href="style.css">
<head>
  <meta http-equiv='cache-control' content='no-cache'>
  <meta http-equiv='expires' content='0'>
  <meta http-equiv='pragma' content='no-cache'>
  <script type="text/javascript" src="canvas2image.js"> </script>
  <script type="text/javascript" src="//ajax.googleapis.com/ajax/libs/jquery/1.7.1/jquery.min.js"></script>
  <script type="text/javascript" src="html2canvas.js"></script>
  <script type="text/javascript" src="jquery.plugin.html2canvas.js"></script>
  <script type="text/javascript" src="specific.js"> </script>
  <script type="text/javascript" src="common.js"> </script>
  <script language="javascript" type="text/javascript">

```

```

function doDownloadChartData(name)
{
    window.location.href = 'dynamicCharts/table' + name + '/download.txt';
}

function setUrl(name, width, filter1Op, filter1Value, filter2Op, filter2Value, filter3Op, filter3Value, filter4Op, filter4Value,
    tableValue) {
    var keepattr = getKeepAttr();

    var timebegin = document.getElementById('beginDate').value;
    var timeend = document.getElementById('endDate').value;

    var url = "http://" + RA_WEBSERVER + "/RA/faces/restricted//embed.xhtml?serviceId=" + TABLE_SERVICE + "&width=" + width +
        "&height=550&db_url=" + DB_URL + "&db_username=" + DB_USERNAME + "&outputdir=" + OUTPUT_DIR + "&filter1attr=" +
        FILTER1_ATTR + "&filter1op=" + filter1Op + "&filter1val=" + filter1Value + "&filter2attr=" + FILTER2_ATTR +
        "&filter2op=" + filter2Op + "&filter2val=" + filter2Value + "&filter3attr=" + FILTER3_ATTR + "&filter3op=" + filter3Op +
        "&filter3val=" + filter3Value + "&filter4attr=" + FILTER4_ATTR + "&filter4op=" + filter4Op + "&filter4val=" +
        filter4Value + "&chartnum=" + name + "&tablename=" + tableValue + "&keepattr=" +
        keepattr + "&timebegin=" + timebegin + "&timeend=" + timeend + "&timeattribute=" + TIME_ATTR;

    document.getElementById("raIframe" + name).src = url;
}

function configurePortalHTML(portalHTML) {
    return portalHTML.replace(/BGCOLOR/g, "white");
}

function postConfigurePortal(name) {
    //do nothing
}

</script>
</head>
<body onload="onPortalInit();">
    <div class="wrapper" align="center">
        <table cellspacing="0" cellpadding="0" border="0" id="shell">
            <tr>
                <td colspan="2">
                    
                    
                </td>
            </tr>
            <tr>
                <td width="70" valign="top">
                    <table id="navigation" title="Navigation" border="0">
                        <tr>
                            <td id="page1" />
                        </tr>
                        <tr>
                            <td id="page2" />
                        </tr>
                        <tr>
                            <td id="page3" />
                        </tr>
                    </table>
                </td>
            </tr>
        </table>
    </div>

```

```

</tr>
<tr>
  <td id="page4" />
</tr>
<tr>
  <td id="page5" />
</tr>
<tr>
  <td id="page6" />
</tr>
<tr>
  <td id="page7" />
</tr>
<tr>
  <td id="page8" />
</tr>
<tr>
  <td id="page9" />
</tr>
</table>
</td>
<td>
  <table id="content">
    <tr>
      <td>
        
        Columns:
      </td>
      <td>
        <div id="tableCols" class="container" style="height: 60px;">
        </div>
      </td>
      <td align="right">
        <input id="beginDate" type="date"> -
        <input id="endDate" type="date">
        <!-- <button id="screenshotBtn" type="submit" onclick="doScreenshot();">
          </button> -->
        <button id="downloadBtn" type="submit" onclick="doDownload();">
          </button>
        <button id="addBtn" type="submit" onclick="addIFrame();">
          </button>
        <button id="refreshBtn" type="submit" onclick="refreshAll();">
          </button>
      </td>
    </tr>
    <tr>
      <td colspan="3">
        <table>
          <tr>
            <td id="tableCell1" align="center" nowrap>
            </td>
            <td id="tableCell2" align="center" nowrap>
            </td>
          </tr>
        </table>
      </td>
    </tr>
  </table>

```

```

        </table>
      </td>
    </tr>
  </table>
</td>
</tr>
<tr>
  <td />
  <td align="center">
    <table id="banner" border="0">
      <tr>
        <td>
          <h2>
            Selected Dataset:</h2>
          </td>
          <td>
            <img id="icon" height="60">
          </td>
          <td>
            <img id="logo" height="60">
          </td>
        </tr>
      </table>
    </td>
  </tr>
</table>
</div>
<div class="footer" align="center">
  <p>
    Copyright (c) 2013 -- contact: <a href="http://lyle.smu.edu/~lsykalski/">Lewis Sykalski</a>
    <a href="http://tylermoore.ens.utulsa.edu/">Tyler Moore</a></p>
</div>
</body>
</html>

```

---

## C.2. Javascript Code

### C.2.1. common.js

```

//
// Copyright [2013] [Lewis Sykalski]
//
// Licensed under the Apache License, Version 2.0 (the "License");
// you may not use this file except in compliance with the License.
// You may obtain a copy of the License at
//
// http://www.apache.org/licenses/LICENSE-2.0
//
// Unless required by applicable law or agreed to in writing, software

```



```

// distributed under the License is distributed on an "AS IS" BASIS,
// WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied.
// See the License for the specific language governing permissions and
// limitations under the License.
//
// source: common.js
//

var widthSetting = "1100";

var PORTAL_HTML = "<td id='tableCellIDNUM' align='center'><table border='1'><tr
    bgcolor='black'><td><b>Filters:</b></td><td><table><tr><td id='filter1Label_IDNUM'></td><td id='filter1_IDNUM'></td><td
    id='filter2Label_IDNUM'></td><td id='filter2_IDNUM'></td><td><button id='removeBtnIDNUM' type='submit'
    onclick='removeIFrame(this);'><img src='icons/remove.png' alt='remove' width='12' height='15'></button></td><td><button
    id='downloadData_IDNUM' align='right' type='submit' onclick='doDownloadChartDataBase(this);'><img src='icons/download.png'
    alt='download' width='15' height='15' align='right'></button></td></tr><tr><td id='filter3Label_IDNUM'></td><td
    id='filter3_IDNUM'></td><td id='filter4Label_IDNUM'></td><td id='filter4_IDNUM'></td></tr></table></td></tr><tr><td colspan='2'
    bgcolor='BGCOLOR' align='center'><b>Source:</b><select id='sourceComboIDNUM' align='center'
    onchange='refresh(this);'></select><br></td></tr><tr><td colspan='2' align='center' bgcolor='BGCOLOR'><iframe id='raIframeIDNUM'
    scrolling='no' width='WIDTH' height='550' /></iframe><iframe id='serviceIframeIDNUM' width='0' height='0'></iframe></td></tr>";
//
//var RA_WEBSERVER = "wilkes.ens.utulsa.edu:8080";
var RA_WEBSERVER = "127.0.0.1:8080";
//var OUTPUT_DIR = "/var/www/html/sececon/dynamicCharts";
var OUTPUT_DIR = "/home/lewis/Desktop/thesis/sececonHTML/dynamicCharts";

var DB_USERNAME = "user";

var BAR_AGGR_SERVICE = "TIME_BAR";
var BOXPLOT_SERVICE = "BOX_PLOT";
var LINE_SERVICE = "TIME_LINE";
var MOSAIC_SERVICE = "MOSAIC_PLOT";
var ODDS_RATIO_SERVICE = "ODDS_RATIO";
var ODDS_RATIO_TIME_SERVICE = "ODDS_RATIO_TIME";
var PIE_SERVICE = "PIE";
var TABLE_SERVICE = "TABLE";
var DEFINITION_SERVICE = "DEFINITION";
var INTERVAL_TIMER1 = 0;
var INTERVAL_TIMER2 = 0;

function setIFrameSrc(name, width) {

    var filter1 = document.getElementById('filter1Combo' + name);
    var filter2 = document.getElementById('filter2Combo' + name);
    var filter3 = document.getElementById('filter3Combo' + name);
    var filter4 = document.getElementById('filter4Combo' + name);

    var filter1Value = "All";
    var filter2Value = "All";
    var filter3Value = "All";
    var filter4Value = "All";

    if (filter1) {
        var filter1Index = filter1.selectedIndex;

```

```

        filter1Value = filter1[filter1Index].value;
    }

    if (filter2) {
        var filter2Index = filter2.selectedIndex;
        filter2Value = filter2[filter2Index].value;
    }

    if (filter3) {
        var filter3Index = filter3.selectedIndex;
        filter3Value = filter3[filter3Index].value;
    }

    if (filter4) {
        var filter4Index = filter4.selectedIndex;
        filter4Value = filter4[filter4Index].value;
    }

    var tableIndex = document.getElementById('sourceCombo' + name).selectedIndex;
    var tableValue = document.getElementById('sourceCombo' + name)[tableIndex].value;

    var filter1Op = "=";
    var filter2Op = "=";
    var filter3Op = "=";
    var filter4Op = "=";
    if (filter1Value == "All")
        filter1Op = "!=";
    if (filter2Value == "All")
        filter2Op = "!=";
    if (filter3Value == "All")
        filter3Op = "!=";
    if (filter4Value == "All")
        filter4Op = "!=";

    setUrl(name, width, filter1Op, filter1Value, filter2Op, filter2Value, filter3Op, filter3Value, filter4Op, filter4Value, tableValue);

    if (name == "1")
        document.getElementById("removeBtn1").style.display = "none";

    postConfigurePortal(name);
}

function addIFrame() {
    var sourceCombo1 = document.getElementById('sourceCombo1');
    var filterCombo1 = document.getElementById('filter1Combo1');
    var filterCombo2 = document.getElementById('filter2Combo1');
    var filterCombo3 = document.getElementById('filter3Combo1');
    var filterCombo4 = document.getElementById('filter4Combo1');

    widthSetting = "550";
    var portalHTML1 = PORTAL_HTML.replace(/IDNUM/g, "1").replace(/WIDTH/g, widthSetting);
    var portalHTML2 = PORTAL_HTML.replace(/IDNUM/g, "2").replace(/WIDTH/g, widthSetting);

    //configure any specifics for portal
    portalHTML1 = configurePortalHTML(portalHTML1);

```

```

portalHTML2 = configurePortalHTML(portalHTML2);

//defaults if not set...
portalHTML1 = portalHTML1.replace(/BGCOLOR/g, "black");
portalHTML2 = portalHTML2.replace(/BGCOLOR/g, "black");

document.getElementById('tableCell1').innerHTML = portalHTML1;
setFilterControls("1");

document.getElementById('tableCell2').innerHTML = portalHTML2;
setFilterControls("2");

if (sourceCombo1) {
    document.getElementById('sourceCombo2').selectedIndex = sourceCombo1.selectedIndex;
}
if (filterCombo1) {
    document.getElementById('filter1Combo2').selectedIndex = filterCombo1.selectedIndex;
}
if (filterCombo2) {
    document.getElementById('filter2Combo2').selectedIndex = filterCombo2.selectedIndex;
}
if (filterCombo3) {
    document.getElementById('filter3Combo1').selectedIndex = filterCombo3.selectedIndex;
}
if (filterCombo4) {
    document.getElementById('filter4Combo1').selectedIndex = filterCombo4.selectedIndex;
}

document.getElementById('addBtn').disabled = true;

refreshAll();
}

function setSrcIFrame(width) {
    setIFrameSrc("1", width);
    setIFrameSrc("2", width);
}

function doDownload() {
    var tableIndex = document.getElementById('sourceCombo1').selectedIndex;
    var tableValue = document.getElementById('sourceCombo1')[tableIndex].value;

    if (tableValue == controlTable)
        window.location.href = controlFile;
    if (tableValue == treatmentTable)
        window.location.href = treatmentFile;
}

function doDownloadChartDataBase(button)
{
    var id = "unknown";
    if (button.id.indexOf("1") >-1)
        id = "1";
    else if (button.id.indexOf("2") >-1)

```

```

        id = "2";

    if(id != "unknown") {
        doDownloadChartData(id);
    }
}

function resizeIframe(id) {
    var newheight;
    var newwidth;

    if (document.getElementById) {
        newheight = document.getElementById(id).contentWindow.document.body.scrollHeight;
        newwidth = document.getElementById(id).contentWindow.document.body.scrollWidth;
    }

    document.getElementById(id).height = (newheight) + "px";
    document.getElementById(id).width = (newwidth) + "px";
}

function removeIframe(sourceBtn) {
    var id = "unknown";
    if (sourceBtn.id.indexOf("1") >-1)
        id = "1";
    else if (sourceBtn.id.indexOf("2") >-1)
        id = "2";
    else if (sourceBtn.id.indexOf("3") >-1)
        id = "3";

    if (id != "unknown") {
        document.getElementById('tableCell' + id).innerHTML = "";
        widthSetting = "1100";
        document.getElementById('raIframe1').width = widthSetting;
        document.getElementById('addBtn').disabled = false;
        refreshAll();
    }
}

function refreshAll() {
    setSrcIframe(widthSetting);
}

function refresh(document) {
    var id = "unknown";
    if (document.id.contains("1"))
        id = "1";
    else if (document.id.contains("2"))
        id = "2";

    if (id != "unknown")
        setIframeSrc(id, widthSetting);
    else
        refreshAll();
}

```

```

function setFilterControls(idNUM) {
    //set the filter controls
    document.getElementById('filter1Label_' + idNUM).innerHTML = FILTER1_LABEL;
    document.getElementById('filter1_' + idNUM).innerHTML = FILTER1_HTML.replace(/IDNUM/g, idNUM);
    document.getElementById('filter2Label_' + idNUM).innerHTML = FILTER2_LABEL;
    document.getElementById('filter2_' + idNUM).innerHTML = FILTER2_HTML.replace(/IDNUM/g, idNUM);
    document.getElementById('filter3Label_' + idNUM).innerHTML = FILTER3_LABEL;
    document.getElementById('filter3_' + idNUM).innerHTML = FILTER3_HTML.replace(/IDNUM/g, idNUM);
    document.getElementById('filter4Label_' + idNUM).innerHTML = FILTER4_LABEL;
    document.getElementById('filter4_' + idNUM).innerHTML = FILTER4_HTML.replace(/IDNUM/g, idNUM);

    var pageTitle = window.location.pathname.replace(/.*\.[^/]*$/, "$1");
    if(pageTitle=="odds.html") {
        document.getElementById('sourceCombo' + idNUM).innerHTML = ODDS_SOURCECOMBO_HTML;
    }
    else if(pageTitle=="oddstime.html") {
        document.getElementById('sourceCombo' + idNUM).innerHTML = ODDS_SOURCECOMBO_HTML;
    }
    else {
        document.getElementById('sourceCombo' + idNUM).innerHTML = SOURCECOMBO_HTML;
    }
}

function addComboOptions(combo, options, selectedOpt) {
    var temp = "";

    for (var i = 0; i < options.length; i++) {
        if (selectedOpt == options[i]) {
            temp = temp + "<option value='" + options[i] + "' selected='true'" + options[i] + "</option>";
        }
        else {
            temp = temp + "<option value='" + options[i] + "'" + options[i] + "</option>";
        }
    }
    // end for

    combo.innerHTML = temp;
}

function addDivOptions(div, options, selected) {
    var temp = "";
    for (var i = 0; i < options.length; i++) {
        var sel = false;
        for (var j = 0; j < selected.length; j++) {
            if (options[i] == selected[j]) {
                sel = true;
            }
        }
        if (sel) {
            temp = temp + "<input id='tableCheck" + options[i] + "' type='checkbox' checked/> " + options[i] + " <br />";
        }
        else {

```

```

        temp = temp + "<input id='tableCheck" + options[i] + "' type='checkbox' /> " + options[i] + " <br />";
    }
} // end for

div.innerHTML = temp;
}

function getKeepAttr() {

    var keepattr = "";
    var eleChild = document.getElementById("tableCols").childNodes;
    for (i = 0; i < eleChild.length; ++i) {
        var widget = eleChild[i];
        var id = widget.id;
        if (id != null) {
            var index = id.indexOf("tableCheck");
            if (index >= 0) {
                if (widget.checked) {
                    if (keepattr != "") {
                        keepattr += "%7C";
                    }
                    keepattr += id.replace(/tableCheck/g, "");
                }
            }
        }
    }

    return keepattr;
}

function generateRequestID()
{
    var date = new Date();
    var components = [
        date.getYear(),
        date.getMonth(),
        date.getDate(),
        date.getHours(),
        date.getMinutes(),
        date.getSeconds(),
        date.getMilliseconds()
    ];

    return components.join("");
}

function onPortalInit() {

    onSpecificInit();

    //set the selected page

    var filename = location.pathname.substring(location.pathname.lastIndexOf("/") + 1);
    document.getElementById(filename).border = "2";
}

```

```

var portalHTML = PORTAL_HTML.replace(/IDNUM/g, "1").replace(/WIDTH/g, widthSetting);
portalHTML = configurePortalHTML(portalHTML);

//default if not set...
portalHTML = portalHTML.replace(/BGCOLOR/g, "black");
document.getElementById('tableCell1').innerHTML = portalHTML;

setFilterControls("1");

refreshAll();
}

function saveCanvas(oCanvas, strType) {
    var success = false;

    if (strType == "PNG")
        success = Canvas2Image.saveAsPNG(oCanvas);
    if (strType == "BMP")
        success = Canvas2Image.saveAsBMP(oCanvas);
    if (strType == "JPEG")
        success = Canvas2Image.saveAsJPEG(oCanvas);

    if (!success) {
        alert("Sorry, this browser is not capable of saving " + strType + " files!");
        return false;
    }
    return true;
}

function doScreenshot() {
    var iframe = document.getElementById('raIframe1');
    var iframeDoc = iframe.contentDocument || iframe.contentWindow.document;
    //var iframeBody = iframeDoc.body;
    var content = document.getElementById('content');
    html2canvas(content, {
        onrendered: function (canvas) {

            // saveCanvas(canvas, "PNG");
            document.body.appendChild(canvas);
        }
    });
}

```

---

## C.2.2. specific.js

---

```

//
// Copyright [2013] [Lewis Sykalski]
//

```

```

// Licensed under the Apache License, Version 2.0 (the "License");
// you may not use this file except in compliance with the License.
// You may obtain a copy of the License at
//
//      http://www.apache.org/licenses/LICENSE-2.0
//
// Unless required by applicable law or agreed to in writing, software
// distributed under the License is distributed on an "AS IS" BASIS,
// WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied.
// See the License for the specific language governing permissions and
// limitations under the License.
//
//
// source: specific.js
//

var controlFile;
var treatmentFile;

var FILTER1_LABEL;
var FILTER1_HTML;

var FILTER2_LABEL;
var FILTER2_HTML;

var FILTER3_LABEL;
var FILTER3_HTML;

var FILTER4_LABEL;
var FILTER4_HTML;

var SOURCECOMBO_HTML;
var ODDS_SOURCECOMBO_HTML;

var FILTER1_ATTR;
var FILTER2_ATTR;
var FILTER3_ATTR;
var FILTER4_ATTR;

var TIME_ATTR;
var GEOSPATIAL_URL;
var TIME_BEGIN;
var TIME_END;

function setupBreach() {

    controlFile = "csvs/breach/Clean.csv";
    treatmentFile = "csvs/breach/Breached.csv";

    FILTER1_ATTR = "sector";
    FILTER1_LABEL = "Sector:";
    FILTER1_HTML = "<select id='filter1ComboIDNUM'><option value='All' selected='selected'>All</option><option
        value='BI'>BI</option><option value='CD'>CD</option><option value='CG'>CG</option><option value='CND'>CND</option><option
        value='CS'>CS</option><option value='EN'>EN</option><option value='FI'>FI</option><option value='HC'>HC</option><option

```



```

        value='MI'>MI</option><option value='NA'>NA</option><option value='PU'>PU</option><option value='TE'>TE</option><option
        value='TR'>TR</option></select>";

FILTER2_ATTR = "capsize_coarse";
FILTER2_LABEL = "Cap Size(Coarse).:";
FILTER2_HTML = "<select id='filter2ComboIDNUM'><option value='All' selected='selected'>All</option><option
        value='SMALL'>SMALL</option><option value='MID'>MID</option><option value='LARGE'>LARGE</option></select>";

FILTER3_ATTR = "capsize_fine";
FILTER3_LABEL = "Cap Size(Fine).:";
FILTER3_HTML = "<select id='filter3ComboIDNUM'><option value='All' selected='selected'>All</option><option
        value='MICRO'>MICRO</option><option value='SMALL'>SMALL</option><option value='MID'>MID</option><option
        value='LARGE'>LARGE</option><option value='MEGA'>MEGA</option></select>";

FILTER4_ATTR = "industry";
FILTER4_LABEL = "";
FILTER4_HTML = "";
// var FILTER4_LABEL ="TLD:"
// var FILTER4_HTML="<select id='filter4ComboIDNUM'><option value='All' selected='selected'>All</option><option
        value='com'>com</option><option value='edu'>edu</option><option value='info'>info</option><option
        value='org'>org</option></select>";

SOURCECOMBO_HTML = "<option value='breach'>breach</option><option value='clean'>clean</option>";
ODDS_SOURCECOMBO_HTML = "<option value='breach vs clean'>breach vs clean</option>";

DB_URL = "jdbc:mysql://127.0.0.1:3306/breach";

//Create logo for page
document.getElementById('icon').src = "breach-icon.png";
document.getElementById('logo').src = "breach-logo.png";

//Add pages (Maximum 8)
document.getElementById('page1').innerHTML =
    '<a href="table.html?data=breach" title="Table"></a>';
document.getElementById('page2').innerHTML =
    '<a href="mosaic.html?data=breach" title="Mosaic Plot"></a>';
document.getElementById('page3').innerHTML =
    '<a href="bar.html?data=breach" title="Bar chart"></a>';
document.getElementById('page4').innerHTML =
    '<a href="pie.html?data=breach" title="Aggregation"></a>';
document.getElementById('page5').innerHTML =
    '<a href="boxplot.html?data=breach" title="box plot"></a>';
document.getElementById('page6').innerHTML =
    '<a href="odds.html?data=breach" title="odds ratio"></a>';
document.getElementById('page7').innerHTML =
    '<a href="line.html?data=breach" title="Line"></a>';
document.getElementById('page8').innerHTML =
    '<a href="help.html?data=breach" title="Table Definitions"></a>';

var aggCombo = document.getElementById('aggregateCombo');
if (aggCombo) {
    var opts = ["breachType", "capsize_coarse", "capsize_fine", "citySize", "entityCategory",

```

```

        "entityType", "industry", "sector", "stockXchg", "region", "state"];
    addComboOptions(aggCombo, opts, "sector");
}

var oddsCombo = document.getElementById('oddsCombo');
if (oddsCombo) {
    var opts = ["capsize_coarse", "capsize_fine", "industry", "sector"];
    addComboOptions(oddsCombo, opts, "sector");
}

var mosCombo1 = document.getElementById('mosaicCombo1');
if (mosCombo1) {
    var opts = ["breachType", "capsize_coarse", "capsize_fine", "citySize", "entityCategory",
        "entityType", "industry", "sector", "stockXchg", "region", "state", "year"];
    addComboOptions(mosCombo1, opts, "entityCategory");
    addComboOptions(document.getElementById('mosaicCombo2'), opts, "region");
}

var boxComboX = document.getElementById('boxComboX');
if (boxComboX) {
    var opts = ["breachType", "capsize_coarse", "capsize_fine", "citySize", "entityCategory",
        "entityType", "industry", "sector", "stockXchg", "region", "state", "year"];
    addComboOptions(boxComboX, opts, "sector");
}

var boxComboY = document.getElementById('boxComboY');
if (boxComboY) {
    var opts = ["population2012", "market_cap", "numRecords"];
    addComboOptions(boxComboY, opts, "population2012");
}

var tableCols = document.getElementById('tableCols');
if (tableCols) {
    var opts = ["breachType", "capsize_coarse", "capsize_fine", "citySize", "entityCategory",
        "entityType", "industry", "market_cap", "name", "sector", "stockXchg", "symbol", "region", "state", "year"];
    var sel = ["breachType", "capsize_coarse", "capsize_fine", "citySize", "entityCategory",
        "entityType", "industry", "market_cap", "name", "sector", "stockXchg", "symbol", "region", "state", "year"];
    addDivOptions(tableCols, opts, sel);
}

var timeTypeCombo = document.getElementById('timeTypeCombo');
if (timeTypeCombo) {
    var opts = ["year", "month"];
    addComboOptions(timeTypeCombo, opts, "month");
}

TIME_ATTR = "time";
TIME_BEGIN = "1979-01-01";
TIME_END = "2015-12-31";
var timeBegin = document.getElementById('beginDate');
if (timeBegin) {
    timeBegin.min = TIME_BEGIN; timeBegin.max = TIME_END; //timeBegin.value = TIME_BEGIN;
}

```

```

}
var timeEnd = document.getElementById('endDate');
if (timeEnd) {
    timeEnd.min = TIME_BEGIN; timeEnd.max = TIME_END; //timeEnd.value = TIME_END;
}
}

function setupCms() {

    controlFile = "csvs/cms/cms_control_csvs.tar.gz";
    treatmentFile = "csvs/cms/cms_compromise_csvs.tar.gz";

    /* google, apache, nginx, iis */

    FILTER1_ATTR = "servertype";
    FILTER1_LABEL = "Server Type:";
    FILTER1_HTML = "<select id='filter1ComboIDNUM'><option value='All' selected='selected'>All</option><option
        value='apache'>apache</option><option value='google'>google</option><option value='iis'>iis</option><option
        value='nginx'>nginx</option></select>";

    /* blogger, drupal, homestead, joomla, typo3, vbulletin, wordpress, zencart */

    FILTER2_ATTR = "generatortype";
    FILTER2_LABEL = "Gen Type:";
    FILTER2_HTML = "<select id='filter2ComboIDNUM'><option value='All' selected='selected'>All</option><option
        value='blogger'>blogger</option><option value='drupal'>drupal</option><option value='homestead'>homestead</option><option
        value='joomla'>joomla</option><option value='typo3'>typo3</option><option value='vbulletin'>vbulletin</option><option
        value='wordpress'>wordpress</option><option value='zencart'>zencart</option></select>";

    //FILTER3_ATTR = "day";
    // FILTER3_LABEL = "Day:";
    // FILTER3_HTML="<select id='filter3ComboIDNUM'><option value='All' selected='selected'>All</option><option
        value='0'>0</option><option value='1'>1</option><option value='2'>2</option><option value='5'>5</option><option
        value='9'>9</option><option value='15'>15</option></select>";
    FILTER3_ATTR = "country";
    FILTER3_LABEL = "Country:";
    FILTER3_HTML = "<select id='filter3ComboIDNUM'><option value='All' selected='selected'>All</option><option
        value='Canada'>Canada</option><option value='China'>China</option><option value='France'>France</option><option
        value='Germany'>Germany</option><option value='Hong Kong'>Hong Kong</option><option value='Japan'>Japan</option><option
        value='Portugal'>Portugal</option><option value='Russian Federation'>Russian Federation</option><option value='United
        Kingdom'>United Kingdom</option><option value='United States'>United States</option></select>";

    FILTER4_ATTR = "tld";
    FILTER4_LABEL = "TLD:";
    FILTER4_HTML = "<select id='filter4ComboIDNUM'><option value='All' selected='selected'>All</option><option
        value='com'>com</option><option value='cn'>cn</option><option value='co.uk'>co.uk</option><option
        value='de'>de</option><option value='edu'>edu</option><option value='net'>net</option><option value='org'>org</option><option
        value='ru'>ru</option></select>";

    SOURCECOMBO_HTML = "<option value='compromise_day0'>compromise_day0</option><option
        value='compromise_day1'>compromise_day1</option><option value='compromise_day2'>compromise_day2</option><option
        value='compromise_day5'>compromise_day5</option><option value='compromise_day9'>compromise_day9</option><option
        value='compromise_day15'>compromise_day15</option><option value='control_day0'>control_day0</option><option
        value='control_day1'>control_day1</option><option value='control_day2'>control_day2</option><option

```

```

value='control_day5'>control_day5</option><option value='control_day9'>control_day9</option><option
value='control_day15'>control_day15</option>;

ODDS_SOURCECOMBO_HTML = "<option value='compromise_day0 vs control_day0'>compromise_day0 vs control_day0</option><option
value='compromise_day1 vs control_day1'>compromise_day1 vs control_day1</option><option value='compromise_day2 vs
control_day2'>compromise_day2 vs control_day2</option><option value='compromise_day5'>compromise_day5 vs
control_day5</option><option value='compromise_day9 vs control_day9'>compromise_day9 vs control_day9</option><option
value='compromise_day15 vs control_day15'>compromise_day15 vs control_day15</option>";

DB_URL = "jdbc:mysql://127.0.0.1:3306/cms";

//Create logo for page
document.getElementById('icon').src = "cms-icon.png";
document.getElementById('logo').src = "cms-logo.png";

//Add pages (Maximum 8)
document.getElementById('page1').innerHTML =
'<a href="table.html?data=cms" title="Table"></a>';
document.getElementById('page2').innerHTML =
'<a href="mosaic.html?data=cms" title="Mosaic Plot"></a>';
document.getElementById('page3').innerHTML =
'<a href="bar.html?data=cms" title="Bar chart"></a>';
document.getElementById('page4').innerHTML =
'<a href="pie.html?data=cms" title="Aggregation"></a>';
document.getElementById('page5').innerHTML =
'<a href="oddstime.html?data=cms" title="odds time">';
document.getElementById('page6').innerHTML =
'<a href="odds.html?data=cms" title="odds ratio"></a>';
document.getElementById('page7').innerHTML =
'<a href="line.html?data=cms" title="Line"></a>';
document.getElementById('page8').innerHTML =
'<a href="help.html?data=cms" title="Table Definitions"></a>';
document.getElementById('page9').innerHTML =
'<a href="geospatial.html?data=cms" title="Geospatial"></a>';

var aggCombo = document.getElementById('aggregateCombo');
if (aggCombo) {
    var opts = ["country", "generator", "wordpressVersion", "genMajorVersion", "generatortype", "server", "servertype", "tld"];
    addComboOptions(aggCombo, opts, "servertype");
}

var oddsCombo = document.getElementById('oddsCombo');
if (oddsCombo) {
    var opts = ["country", "wordpressVersion", "generatortype", "server", "servertype"];
    addComboOptions(oddsCombo, opts, "servertype");
}

var mosCombo1 = document.getElementById('mosaicCombo1');
if (mosCombo1) {
    var opts = ["country", "wordpressVersion", "generatortype", "server", "servertype", "tld"];
    addComboOptions(mosCombo1, opts, "generatortype");
}

```

```

        addComboOptions(document.getElementById('mosaicCombo2'), opts, "servertype");
    }

    var boxComboX = document.getElementById('boxComboX');
    if (boxComboX) {
        var opts = ["country", "wordpressVersion", "generatortype", "server", "servertype", "tld"];
        addComboOptions(boxComboX, opts, "servertype");
    }

    var boxComboY = document.getElementById('boxComboY');
    if (boxComboY) {
        var opts = ["day"];
        addComboOptions(boxComboY, opts, "day");
    }

    var tableCols = document.getElementById('tableCols');
    if (tableCols) {
        var opts = ["collectiondate", "country", "day", "generator", "wordpressVersion", "generatortype", "id", "server", "servertype",
            "tld"];
        var sel = ["collectiondate", "country", "day", "generator", "wordpressVersion", "generatortype", "id", "server", "servertype",
            "tld"];
        addDivOptions(tableCols, opts, sel);
    }

    var timeTypeCombo = document.getElementById('timeTypeCombo');
    if (timeTypeCombo) {
        var opts = ["year", "month"];
        addComboOptions(timeTypeCombo, opts, "month");
    }

    TIME_ATTR = "collectiondate";
    TIME_BEGIN = "1979-01-01";
    TIME_END = "2015-12-31";
    var timeBegin = document.getElementById('beginDate');
    if (timeBegin) {
        timeBegin.min = TIME_BEGIN; timeBegin.max = TIME_END; //timeBegin.value = TIME_BEGIN;
    }
    var timeEnd = document.getElementById('endDate');
    if (timeEnd) {
        timeEnd.min = TIME_BEGIN; timeEnd.max = TIME_END; //timeEnd.value = TIME_END;
    }

    GEOSPATIAL_URL = new Object();
    GEOSPATIAL_URL["compromise_day0"] =
        "https://www.google.com/fusiontables/embedviz?q=select+col1+from+i3buur4LN9UJbb-dXna-gSIUd2M8At7Lb10Wx1zA&viz=MAP&h=false&lat=19.75167645062207&lng=-";
    GEOSPATIAL_URL["compromise_day1"] = GEOSPATIAL_URL["compromise_day0"];
    GEOSPATIAL_URL["compromise_day2"] = GEOSPATIAL_URL["compromise_day0"];
    GEOSPATIAL_URL["compromise_day5"] = GEOSPATIAL_URL["compromise_day0"];
    GEOSPATIAL_URL["compromise_day9"] = GEOSPATIAL_URL["compromise_day0"];
    GEOSPATIAL_URL["compromise_day15"] = GEOSPATIAL_URL["compromise_day0"];
    GEOSPATIAL_URL["control_day0"] =
        "https://www.google.com/fusiontables/embedviz?q=select+col1+from+iRVg2wbxs7s6JkPIAbRhr7FRKMAe5_FB-j8kC8Ks&viz=MAP&h=false&lat=19.75167645062207&lng=-";

```

```

    GEOSPATIAL_URL["control_day1"] = GEOSPATIAL_URL["control_day0"];
    GEOSPATIAL_URL["control_day2"] = GEOSPATIAL_URL["control_day0"];
    GEOSPATIAL_URL["control_day5"] = GEOSPATIAL_URL["control_day0"];
    GEOSPATIAL_URL["control_day9"] = GEOSPATIAL_URL["control_day0"];
    GEOSPATIAL_URL["control_day15"] = GEOSPATIAL_URL["control_day0"];
}

function setupHoneypot() {

    controlFile = "csvs/honeypot/honeypot.csv";
    treatmentFile = "";

    /* google, apache, nginx, iis */

    FILTER1_ATTR = "protocol";
    FILTER1_LABEL = "Protocol:";
    FILTER1_HTML = "<select id='filter1ComboIDNUM'><option value='All' selected='selected'>All</option><option
        value='TCP'>TCP</option><option value='UDP'>UDP</option></select>";

    /* blogger, drupal, homestead, joomla, typo3, vbulletin, wordpress, zencart */

    FILTER2_ATTR = "host_target";
    FILTER2_LABEL = "Host Target:";
    FILTER2_HTML = "<select id='filter2ComboIDNUM'><option value='All' selected='selected'>All</option><option
        value='groucho-norcal'>groucho-norcal</option><option value='groucho-oregon'>groucho-oregon</option><option
        value='groucho-sa'>groucho-sa</option><option value='groucho-singapore'>groucho-singapore</option><option
        value='groucho-sydney'>groucho-sydney</option><option value='groucho-tokyo'>groucho-tokyo</option><option
        value='groucho-us-east'>groucho-us-east</option><option value='zeppo-norcal'>zeppo-norcal</option></select>";

    FILTER3_ATTR = "country";
    FILTER3_LABEL = "Country:";
    FILTER3_HTML = "<select id='filter3ComboIDNUM'><option value='All' selected='selected'>All</option><option
        value='Austria'>Austria</option><option value='Brazil'>Brazil</option><option value='Canada'>Canada</option><option
        value='China'>China</option><option value='France'>France</option><option value='Germany'>Germany</option><option value='Hong
        Kong'>Hong Kong</option><option value='India'>India</option><option value='Iran'>Iran</option><option
        value='Japan'>Japan</option><option value='Netherlands'>Netherlands</option><option value='South Korea'>South
        Korea</option><option value='Taiwan'>Taiwan</option><option value='United Kingdom'>United Kingdom</option><option
        value='United States'>United States</option><option value='Vietnam'>Vietnam</option></select>";

    FILTER4_ATTR = "service";
    FILTER4_LABEL = "Service:";
    FILTER4_HTML = "<select id='filter4ComboIDNUM'><option value='All' selected='selected'>All</option><option
        value='chargen'>chargen</option><option value='domain'>domain</option><option value='drwcs'>drwcs</option><option
        value='epmap'>epmap</option><option value='http'>http</option><option value='http-alt'>http-alt</option><option
        value='https'>https</option><option value='microsoft-ds'>microsoft-ds</option><option
        value='ms-sql-s'>ms-sql-s</option><option value='ms-wbt-server'>ms-wbt-server</option><option
        value='mysql'>mysql</option><option value='ndl-aas'>ndl-aas</option><option value='sip'>sip</option><option
        value='ssh'>ssh</option><option value='telnet'>telnet</option><option value='Unassigned'>Unassigned</option></select>";

    SOURCECOMBO_HTML = "<option value='honeypot'>honeypot</option>";

    DB_URL = "jdbc:mysql://127.0.0.1:3306/honeypot";
}

```

```

//Create logo for page
document.getElementById('icon').src = "honeypot-icon.png";
document.getElementById('logo').src = "honeypot-logo.png";

//Add pages (Maximum 8)
document.getElementById('page1').innerHTML =
  '<a href="table.html?data=honeybot" title="Table"></a>';
document.getElementById('page2').innerHTML =
  '<a href="mosaic.html?data=honeybot" title="Mosaic Plot"></a>';
document.getElementById('page3').innerHTML =
  '<a href="pie.html?data=honeybot" title="Aggregation"></a>';
document.getElementById('page4').innerHTML =
  '<a href="line.html?data=honeybot" title="Line"></a>';
document.getElementById('page5').innerHTML =
  '<a href="bar.html?data=honeybot" title="Bar chart"></a>';
document.getElementById('page6').innerHTML =
  '<a href="geospatial.html?data=honeybot" title="Geospatial"></a>';
document.getElementById('page7').innerHTML =
  '<a href="help.html?data=honeybot" title="Table Definitions"></a>';

var aggCombo = document.getElementById('aggregateCombo');
if (aggCombo) {
  var opts = ["country", "country_code", "host_target", "locale_abbr", "postal_code", "protocol", "service"];
  addComboOptions(aggCombo, opts, "protocol");
}

var mosCombo1 = document.getElementById('mosaicCombo1');
if (mosCombo1) {
  var opts = [ "host_target", "protocol", "country_code", "locale_abbr", "postal_code", "service"];
  addComboOptions(mosCombo1, opts, "host_target");
  addComboOptions(document.getElementById('mosaicCombo2'), opts, "service");
}

var tableCols = document.getElementById('tableCols');
if (tableCols) {
  var opts = ["datetime", "host_target", "ip_src", "protocol", "attack_port", "ipaddr", "country_code", "country", "locale",
    "locale_abbr", "postal_code", "latitude", "longitude", "service"];
  var sel = ["datetime", "host_target", "ip_src", "protocol", "attack_port", "ipaddr", "country_code", "country", "locale",
    "locale_abbr", "postal_code", "latitude", "longitude", "service"];
  addDivOptions(tableCols, opts, sel);
}

var timeTypeCombo = document.getElementById('timeTypeCombo');
if (timeTypeCombo) {
  var opts = ["year", "month"];
  addComboOptions(timeTypeCombo, opts, "month");
}

TIME_ATTR = "datetime";
TIME_BEGIN = "1979-01-01";
TIME_END = "2015-12-31";

```

```
var timeBegin = document.getElementById('beginDate');
if (timeBegin) {
    timeBegin.min = TIME_BEGIN; timeBegin.max = TIME_END; //timeBegin.value = TIME_BEGIN;
}
var timeEnd = document.getElementById('endDate');
if (timeEnd) {
    timeEnd.min = TIME_BEGIN; timeEnd.max = TIME_END; //timeEnd.value = TIME_END;
}

GEO_SPATIAL_URL = new Object();
GEO_SPATIAL_URL["honeypot"] =
    "https://www.google.com/fusiontables/embedviz?q=select+col11%2C+col12+from+1IC4dAFEQ4PsPzC0t468HOY8GM0Viiwsc8mu4M1fm+where+col6+%3D+'CN'+limit+1000&v";
}

function onSpecificInit() {
    var param = location.search.split('data=')[1] ? location.search.split('data=')[1] : 'breach';
    if (param == "breach") {
        setupBreach();
    }
    else if (param == "cms") {
        setupCms();
    }
    else if (param == "honeypot") {
        setupHoneypot();
    }
}
}
```

---



## REFERENCES

- [1] *SIGGRAPH Comput. Graph.* 21, 6 (1987).
- [2] AMAZON. Amazon web services. <https://aws.amazon.com/>, 2015.
- [3] ANDERSON, R., AND MOORE, T. Information security economics - and beyond. In *CRYPTO* (2007), pp. 68–91.
- [4] ARCHIVE, I. Icon archive - search 523,525 icons. <http://www.iconarchive.com/>, 2014.
- [5] ASSOCIATES, J. State security breach notification laws. <http://www.e-janco.com/Articles/2012/201209-Security-Breach-State-Laws.html>, 2014.
- [6] AUTHORITY, I. A. N. Service name and transport protocol port number registry. <http://www.iana.org/assignments/service-names-port-numbers/service-names-port-numbers.csv>, 2015.
- [7] BEAVER, K. Commonly hacked ports. <http://www.dummies.com/how-to/content/commonly-hacked-ports.html>, 2015. Online; accessed 19-Sept-2015.
- [8] BENDER, M., KLEIN, R., DISCH, A., AND EBERT, A. A functional framework for web-based information visualization systems. *IEEE Transactions on Visualization and Computer Graphics* 6, 1 (Jan. 2000), 8–23.
- [9] BLANDER, D. Honeypot dataset marx csv. <http://datadrivensecurity.info/blog/data/2014/01/marx-geo.tar.gz>, 2015.
- [10] BOB RUDIS, J. J. Data driven security. <http://datadrivensecurity.info/blog/>, 2015.
- [11] BOB RUDIS, J. J. Data driven security. <http://datadrivensecurity.info/blog/pages/about-dds.html>, 2015.
- [12] BOB RUDIS, J. J. Inspecting internet traffic. <http://datadrivensecurity.info/blog/posts/2014/Jan/blander-part1/>, 2015.
- [13] CHI, E. H.-H., AND RIEDL, J. An operator interaction framework for visualization systems. In *Proceedings of the 1998 IEEE Symposium on Information Visualization* (Washington, DC, USA, 1998), INFOVIS '98, IEEE Computer Society, pp. 63–70.

- [14] CLEARINGHOUSE, P. R. About privacy rights clearinghouse. <https://www.privacyrights.org/node/1398>, 2014. [Online; accessed 08-December-2014].
- [15] CLEARINGHOUSE, P. R. Privacy rights clearinghouse breach dataset. <http://www.privacyrights.org/data-breach>, 2014. [Online; accessed 08-December-2014].
- [16] CLEARINGHOUSE, P. R. Privacyrightsclearinghouse - empowering consumers protecting privacy. <http://privacyrights.org>, 2014. [Online; accessed 08-December-2014].
- [17] CLEARINGHOUSE, P. R. Privacyrightsclearinghouse - empowering consumers protecting privacy. <http://www.privacyrights.org/data-breach>, 2014. [Online; accessed 08-December-2014].
- [18] DATABREACHES. Data breaches. <http://www.databreaches.org>, 2014.
- [19] DOERR, K.-U., AND KUESTER, F. Cglx: A scalable, high-performance visualization framework for networked display environments. *IEEE Trans. Vis. Comput. Graph.* 17, 3 (2011), 320–332.
- [20] FORBES. Wordpress under attack as double zero-day trouble lands. <http://www.forbes.com/sites/thomasbrewster/2015/04/27/wordpress-zero-day-exploits/>, 2015. [Online; accessed 08-September-2015].
- [21] FOWLER, J. J., JOHNSON, T., SIMONETTO, P., SCHNEIDER, M., ACEDO, C., KOBOUROV, S., AND LAZOS, L. Imap: Visualizing network activity over internet maps. In *Proceedings of the Eleventh Workshop on Visualization for Cyber Security* (New York, NY, USA, 2014), VizSec '14, ACM, pp. 80–87.
- [22] GOVERNANCE, I. Us data breach notifications by state. [http://www.itgovernanceusa.com/data-breach-notification-laws.aspx#.VVjagEPoe\\_o](http://www.itgovernanceusa.com/data-breach-notification-laws.aspx#.VVjagEPoe_o), 2014.
- [23] GOVERNMENT, U. S. 2010 census population data. all incorporated places. <http://www.census.gov/popest/data/cities/totals/2011/SUB-EST2011-3.html>, 2014.
- [24] GOVERNMENT, U. S. Census regions. [http://www.census.gov/geo/www/us\\_regdiv.pdf](http://www.census.gov/geo/www/us_regdiv.pdf), 2014.
- [25] HEER, J., CARD, S. K., AND LANDAY, J. A. Prefuse: A toolkit for interactive information visualization. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (New York, NY, USA, 2005), CHI '05, ACM, pp. 421–430.

- [26] HHS. Health and human services. <http://www.hhs.gov/ocr/privacy/hipaa/administrative/breachnotificationrule/postedbreaches.html>, 2014.
- [27] HOLMBERG, N., WÜNSCHE, B., AND TEMPERO, E. A framework for interactive web-based visualization. In *Proceedings of the 7th Australasian User Interface Conference - Volume 50* (Darlinghurst, Australia, Australia, 2006), AUIC '06, Australian Computer Society, Inc., pp. 137–144.
- [28] JACOBS, J. Honeypot dataset marx csv. <http://datadrivensecurity.info/blog/posts/2014/Jan/blander-part1/>, 2015.
- [29] JACOBS, J. Honeypot dataset marx csv. <http://datadrivensecurity.info/blog/posts/2014/Jan/blander-part2/>, 2015.
- [30] KDNUGGETS. Kdnuggets 15th annual analytics, data mining, data science software poll: Rapidminer continues to lead. <http://www.kdnuggets.com/polls/2014/analytics-data-mining-data-science-software-used.html>, 2014.
- [31] KNIGHT, H. . Expanded u.s. sanctions on iran effective july, 1st 2013. <http://www.hklaw.com/publications/Expanded-US-Sanctions-on-Iran-Effective-July-1-2013-06-06-2013/>, 2015. Online; accessed 19-Sept-2015.
- [32] KOMLODI, A., GOODALL, J. R., AND LUTTERS, W. G. An information visualization framework for intrusion detection. In *CHI '04 Extended Abstracts on Human Factors in Computing Systems* (New York, NY, USA, 2004), CHI EA '04, ACM, pp. 1743–.
- [33] MARIE VASEK, JOHN WADLEIGH, T. M. Hacking is not random: a case-control study of webserver-compromise risk. *IEEE Transactions on Dependable and Secure Computing* 8437 (2015).
- [34] NAID. National association for information destruction. <http://www.naid.org>, 2014.
- [35] OSF-LISTSERVE. Open software foundation listserve. <http://datalossdb.org/>, 2014.
- [36] PRIVACY, P. Protected health information - privacy. <http://www.phip.org>, 2014.
- [37] RAPIDMINER. Rapidminer. <https://rapidminer.com/>, 2013.
- [38] RAPIDMINER. Rapidminer forums. <http://forum.rapid-i.com/>, 2014.
- [39] SERVICES, A. W. Aws developer forums: Honeypot. <https://forums.aws.amazon.com/thread.jspa?threadID=105126>, 2015.

- [40] SHAKACON. Shakacon security conference. <http://www.shakacon.org/2012/speakers.htm>, 2015.
- [41] STOPBADWARE. Stopbadware: A nonprofit organization that makes the web safer. <https://www.stopbadware.org/>, 2015. Online; accessed 19-Sept-2015.
- [42] TAYLOR, T., AND (CANADA)., D. U. *FloVis: A Network Security Visualization Framework*. Canadian theses. Dalhousie University (Canada), 2009.
- [43] UBUNTU.COM. Mysql. <https://help.ubuntu.com/lts/serverguide/mysql.html>, 2014. [Online; accessed 08-December-2014].
- [44] VASEK, M., AND MOORE, T. Identifying risk factors for webserver compromise. *IEEE Transactions on Visualization and Computer Graphics* 8437 (2014), 326–345.
- [45] VIZSEC. Vizsec. <http://www.vizsec.org/>, 2015.
- [46] VON HERTZEN, N. html2canvas - screenshots with javascript. <http://html2canvas.hertzen.com/>, 2014.
- [47] WANG, Q., WANG, W., BROWN, R., DRIESEN, K., DUFOUR, B., HENDREN, L., AND VERBRUGGE, C. Evolve: An open extensible software visualization framework. In *Proceedings of the 2003 ACM Symposium on Software Visualization* (New York, NY, USA, 2003), SoftVis '03, ACM, pp. 37–ff.
- [48] WIKIPEDIA. List of cities in china by population. [https://en.wikipedia.org/wiki/List\\_of\\_cities\\_in\\_China\\_by\\_population](https://en.wikipedia.org/wiki/List_of_cities_in_China_by_population), 2015. Online; accessed 19-Sept-2015.
- [49] WILLIAM YURCIK, XIN MENG, N. K. Nvisioncc: a visualization framework for high performance cluster security.
- [50] YADAV, F. Analyzing financial data through visualization.
- [51] ZHAO, Y., ZHOU, F., FAN, X., LIANG, X., AND LIU, Y. Idsradar: a real-time visualization framework for ids alerts. *Science China Information Sciences* 56, 8 (2013), 1–12.