**Calhoun: The NPS Institutional Archive**

**DSpace Repository**

Theses and Dissertations | 1. Thesis and Dissertation Collection, all items

2009-06

# Visualization of client-side web browsing and email activity

Roussas, Gregory.

Monterey, California. Naval Postgraduate School

https://hdl.handle.net/10945/4738

*Downloaded from NPS Archive: Calhoun*

http://www.nps.edu/library

Calhoun is the Naval Postgraduate School's public access digital repository for research materials and institutional publications created by the NPS community. Calhoun is named for Professor of Mathematics Guy K. Calhoun, NPS's first appointed -- and published -- scholarly author.

Dudley Knox Library / Naval Postgraduate School
411 Dyer Road / 1 University Circle
Monterey, California USA 93943

# NAVAL POSTGRADUATE SCHOOL

## MONTEREY, CALIFORNIA

# THESIS

**VISUALIZATION OF CLIENT-SIDE WEB BROWSING AND EMAIL ACTIVITY**

by

Gregory Roussas

June 2009

| | |
|---|---|
| Thesis Advisor: | Cynthia E. Irvine |
| Second Reader: | Chris S. Eagle |

**Approved for public release; distribution is unlimited**

THIS PAGE INTENTIONALLY LEFT BLANK

# REPORT DOCUMENTATION PAGE

*Form Approved OMB No. 0704–0188*

The public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden to Department of Defense, Washington Headquarters Services, Directorate for Information Operations and Reports (0704–0188), 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202–4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to any penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number. **PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.**

| 1. REPORT DATE *(DD–MM–YYYY)* | 2. REPORT TYPE | 3. DATES COVERED *(From — To)* |
|---|---|---|
| 22–6–2009 | Master's Thesis | 2007-09-21—2009-06-19 |

**4. TITLE AND SUBTITLE**

Visualization of Client-Side Web Browsing and Email Activity

**5a. CONTRACT NUMBER**

**5b. GRANT NUMBER**

DUE 0414102

**5c. PROGRAM ELEMENT NUMBER**

**6. AUTHOR(S)**

Gregory Roussas

**5d. PROJECT NUMBER**

**5e. TASK NUMBER**

**5f. WORK UNIT NUMBER**

**7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)**

Naval Postgraduate School
Monterey, CA 93943

**8. PERFORMING ORGANIZATION REPORT NUMBER**

**9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)**

N/A

**10. SPONSOR/MONITOR'S ACRONYM(S)**

**11. SPONSOR/MONITOR'S REPORT NUMBER(S)**

**12. DISTRIBUTION / AVAILABILITY STATEMENT**

Approved for public release; distribution is unlimited

**13. SUPPLEMENTARY NOTES**

**14. ABSTRACT**

Both web browsers and email clients provide records of user activity, the former as part of the history mechanism for revisitation purposes, and the latter as part of each message. Both are highly valuable from a forensic perspective, with elements such as visited site, mail contact, and event timestamp revealing a wealth of information about the user's browsing and communication behavior. The ability of the forensic analyst to quickly and efficiently explore and understand this volume of information and reconstruct the user's online activity is important, and can contribute to the progress of the investigation. The objective of this thesis is the design and construction of a set of tools to transform this textual history into a visual format, thus facilitating the analysis, interpretation and identification of trends and relationships that may exist. The result of transforming textual histories into visual images and presenting them in a single summary report is the effective distillation of large amounts of information into minimal space, thereby enabling the analyst to form a high-level profile of the user who generated the data. This allows him to better understand the user's online activity in the context of the specific investigation, and effectively prioritize his limited time and attention.

**15. SUBJECT TERMS**

Forensic, Browsing, Email, Automation, Visualization

| 16. SECURITY CLASSIFICATION OF: | | | 17. LIMITATION OF ABSTRACT | 18. NUMBER OF PAGES | 19a. NAME OF RESPONSIBLE PERSON |
|---|---|---|---|---|---|
| a. REPORT | b. ABSTRACT | c. THIS PAGE | | | 19b. TELEPHONE NUMBER *(include area code)* |
| Unclassified | Unclassified | Unclassified | UU | 152 | |



**Standard Form 298 (Rev. 8–98)**
**Prescribed by ANSI Std. Z39.18**

THIS PAGE INTENTIONALLY LEFT BLANK

**Approved for public release; distribution is unlimited**

**VISUALIZATION OF CLIENT-SIDE WEB BROWSING AND EMAIL ACTIVITY**

Gregory Roussas
Civilian, Naval Postgraduate School
B.S., University of California Davis, 1995

Submitted in partial fulfillment of the
requirements for the degree of

**MASTER OF SCIENCE IN COMPUTER SCIENCE**

from the

**NAVAL POSTGRADUATE SCHOOL**
**June 2009**

Author: Gregory Roussas

Approved by: Cynthia E. Irvine
Thesis Advisor

Chris S. Eagle
Second Reader

Peter J. Denning
Chair, Department of Computer Science

THIS PAGE INTENTIONALLY LEFT BLANK

# ABSTRACT

Both web browsers and email clients provide records of user activity, the former as part of the history mechanism for revisitation purposes, and the latter as part of each message. Both are highly valuable from a forensic perspective, with elements such as visited site, mail contact, and event timestamp revealing a wealth of information about the user's browsing and communication behavior. The ability of the forensic analyst to quickly and efficiently explore and understand this volume of information and reconstruct the user's online activity is important, and can contribute to the progress of the investigation. The objective of this thesis is the design and construction of a set of tools to transform this textual history into a visual format, thus facilitating the analysis, interpretation and identification of trends and relationships that may exist. The result of transforming textual histories into visual images and presenting them in a single summary report is the effective distillation of large amounts of information into minimal space, thereby enabling the analyst to form a high-level profile of the user who generated the data. This allows him to better understand the user's online activity in the context of the specific investigation, and effectively prioritize his limited time and attention.

THIS PAGE INTENTIONALLY LEFT BLANK

# Contents


**1 Introduction** **1**
1.1 Motivation . . . . . 1
1.2 Purpose of Study . . . . . 4
1.3 Organization of Thesis . . . . . 5

**2 Background and Existing Tools** **7**
2.1 Browser History File Formats . . . . . 7
2.2 Mailbox File Formats . . . . . 8
2.3 Information Visualization . . . . . 10
2.4 Current Tools . . . . . 12
2.5 Summary . . . . . 15

**3 Features, Requirements and Design Overview** **23**
3.1 Key Stakeholder and High-Level Goals . . . . . 23
3.2 System Features Summary . . . . . 26
3.3 Choice of Extracted Data . . . . . 28
3.4 Selected Use Cases . . . . . 31
UC1: Create New Case File . . . . . 34
UC2: Ingest Browser History File and Populate System Data Structure . . . . . 34
UC3: Ingest Mailbox File and Populate System Data Structure . . . . . 35
UC4: Calculate Summaries – Visit Counts by Time Period . . . . . 36
UC5: Derive Metadata – Site Country . . . . . 37
UC6: Create Visualization – Counts Over Time . . . . . 38
3.5 Supplementary Specification . . . . . 40
3.6 Domain Model . . . . . 43
3.7 Data Structures . . . . . 43

**4 Visualization of Historical Web and Mail Activity** **49**
4.1 The Visualization Process . . . . . 49
4.2 Discussion of Visualizations . . . . . 55

**5 Sample Analysis Session** **83**
5.1 Ingest, Category Resolution and Report Generation . . . . . 83
5.2 Summary Report Analysis . . . . . 85
5.3 Details . . . . . 89

5.4 Summary . . . . . 92

**6 Conclusion and Future Work** **97**
6.1 Conclusion . . . . . 97
6.2 Future Work . . . . . 98

**Appendix A Glossary** **111**

**Appendix B Utility Console Output** **113**
B.1 Ingest . . . . . 113
B.2 Category Resolution . . . . . 115
B.3 Report Generation . . . . . 117

**Appendix C Detail Output** **121**
Visited Embedded URLs . . . . . 121
Downloads . . . . . 122
Typed Sites . . . . . 124
Search Queries . . . . . 126

**Referenced Authors** **131**

**Initial Distribution List** **133**

# List of Figures


2.1 Mozilla Firefox `moz_places` and `moz_historyvisits` history tables . . . 8

2.2 Tabled output format of two current forensic utilities . . . 16

2.3 EnCase timeline view From [52] . . . 17

2.4 WebViz visit categories web visualization From [53] . . . 17

2.5 Two different network graph layouts used for web visualization . . . 18

2.6 Two visualizations based on the network graph applied to web traffic . . . 19

2.7 Visualizations based on the network graph applied to mail From [62] . . . 20

2.8 Visualizations based on the network graph applied to mail From [63, 64] . . . 21

2.9 Novel visualizations applied to mail . . . 22

3.1 Concept of Operations . . . 32

3.2 Domain Model . . . 44

4.1 Transforming Raw Data to Data Tables . . . 50

4.2 Transforming Data Tables to Visual Structures . . . 54

4.3 Card Reference Model for Visualization . . . 55

4.4 Understanding the Visual Structure . . . 56

4.5 Understanding and context in the conversion of data to wisdom . . . 57

4.6 Web browsing and mail summary tables . . . 58

4.7 Tooltips for displaying additional detail . . . 70

4.8 Navigation history details from section of web history . . . 71

4.9 Top visited base and full URLs from web summary . . . 72

4.10 Top senders and recipients from mail summary . . . 73

4.11 Top senders and recipients from mail summary ignoring most frequent . . . 74

4.12 Top domains from web summary ignoring most frequent . . . . . 75

4.13 Top visited base and full URLs by frecency from web summary . . . . . 76

4.14 Daily visit and message counts . . . . . 77

4.15 Top categories and search queries from web summary . . . . . 78

4.16 Totals by visit type and access scheme
Top occurring countries and TLDs . . . . . 79

4.17 Daily visit and message counts time-series . . . . . 80

4.18 Timeline of web history . . . . . 81

4.19 Timeline of mail history (single day detail) . . . . . 82

5.1 Summary table of unique counts . . . . . 86

5.2 Top visited base and full URLs . . . . . 87

5.3 Top visited base and full URLs by count and frecency . . . . . 87

5.4 Totals by visit type and access scheme . . . . . 93

5.5 Top visited countries and TLDs . . . . . 94

5.6 Top categories and search queries . . . . . 95

5.7 Daily site visit counts . . . . . 96

# List of Tables


2.1 Common browser history file locations . . . . . 7

2.2 Forensic tools for web browser history display and analysis . . . . . 12

3.1 Stakeholder goals . . . . . 24

3.2 Key high-hevel goals . . . . . 26

3.3 GeoIP resolution for diamond.nps.edu . . . . . 30

3.4 Web browser history record data structure . . . . . 46

3.5 Mail message record data structure . . . . . 47

4.1 Structure of a Data Table . . . . . 50

4.2 Data Table containing navigation records from a browser history . . . . . 51

4.3 Data Table illustrating relationship between users exchanging mail messages . 51

4.4 Web browser history elements data types . . . . . 53

5.1 Sample history record after ingest phase . . . . . 84

5.2 Unsuccessful category lookup . . . . . 85

THIS PAGE INTENTIONALLY LEFT BLANK

# Acknowledgements

I would like to express my sincerest thanks to my advisor, Dr. Cynthia Irvine, for supporting and standing behind this work from its inception, while providing valued time, effort and insight over the course of its development. To Chris Eagle, your capacity to simultaneously mentor and allow independent development in your students has greatly expanded and shaped my outlook during my time here, and contributed significantly to this work. A thanks to all others who made suggestions and provided guidance, with a very special mention for my fiance, Laura, without whom I would never have had the will and energy to dedicate to this effort. Finally, I would like to thank Dr. Irvine and Valerie Linhoff for everything they have done to make the Scholarship For Service program at NPS what it is.

This material is based on work supported by the National Science Foundation, under Grant DUE 0414102. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author and do not reflect the views of the National Science Foundation.

THIS PAGE INTENTIONALLY LEFT BLANK

# CHAPTER 1:
# Introduction

## 1.1 Motivation

From its origin as an information-presentation medium to its increasing use as an infrastructure hosting dynamic service-oriented applications, the World Wide Web (WWW) has evolved into a platform for entertainment, commerce, communication and collaboration, and has seen the number of users increase along with its functionality [1]. In the largest study of its kind, the World Internet Project Report by the University of Southern California Annenberg School for Communication, two-thirds of the surveyed users ranked the Internet as a very important source of information. Even larger percentages ranked it as a more important source of information than television, newspapers and radio [1]. In addition to this, traffic once carried over closed, proprietary networks is being moved to the Internet for reasons of convenience and cost-saving, resulting in the Internet becoming part of the worldwide crucial infrastructure.

Throughout this evolution, the web browser has remained the primary means of interfacing and interacting with Internet resources, and is increasingly becoming the interface for services that traditionally had specialized clients such as email, chat, file storage, document editing and remote desktop access. Web browsers have effectively become the universal front-end to interactive network-hosted applications, and it is precisely this phenomenon that makes a user's navigational history, i.e, a complete log of their browsing activity, of forensic interest. The navigational history has the potential to become a central repository for understanding the user's online interests and behavior, and as more services migrate to the WWW and become accessible via web browsers, an increasing amount of user activity will be reflected in this history.

Three scenarios would bring a user's navigational or email history to a forensic analyst's attention: the user is a victim, insider threat, or malicious perpetrator. In the first scenario, the user is a victim of the exploitation of a vulnerability in the web browser, the associated plugin architecture or the email client, and the forensic analyst is tasked with determining when and how this exploitation occurred and what repercussions there may have been. In the second scenario, the user was an insider whose actions resulted in damage to their organization, and the forensic analyst is tasked with reconstructing the when and how as well as the extent and degree of the damage. Finally, in the third scenario, the user was the perpetrator of an illegal

act and has used the WWW through the browser or email client in some or all of that act. The role of the web browser or email client in all three scenarios and their respective histories will be discussed in more detail below.

### 1.1.1 User as Victim

The WWW has become the primary conduit for attack activity and the attack focus is shifting from entire networks to targeting specific clients [2, 3, 4, 5]. Exploitation of web browser vulnerabilities can be triggered by a simple visit to a malicious web page with no user interaction and can lead to memory corruption, spoofing and remote execution of arbitrary code or scripts. In certain situations these vulnerabilities can be further leveraged to exploit more serious vulnerabilities in core operating system components as has been demonstrated by exploits against the Microsoft Windows HTML Help facility and the Graphics Rendering Engine [6]. The delivery mechanisms for many of these exploits — malware-infected web pages — are appearing at a rate of one every four-and-a-half seconds, with the highest percentage (37%) hosted in the U.S. [4]. Furthermore, many of these sites are widely-used and trusted by the general public [4]. Based on the details outlined in the Symantec Internet Threat Report 2008 [2, 3], the number of web-oriented vulnerabilities has climbed significantly from past years and is outpacing all other vulnerability categories. Of all documented vulnerabilities, 58% affected web applications, and the numbers of Cross Site Scripting (XSS) vulnerabilities are outpacing traditional vulnerabilities (11,253 versus 2,134) [2, 3]. Web browser vulnerabilities are also increasing - 88 in Mozilla, 22 in Safari, 18 in Internet Explorer and 12 in Opera [2, 3]. There were 239 vulnerabilities found in third-party browser plugins of which 79% were in ActiveX, 8% in QuickTime, 5% in Flash and 5% in Java [2, 3].

Web sites infected with malicious programs which exploit vulnerabilities in the visitor's web browser may be explicitly owned and operated by the attacker, or could be victims themselves. Any vulnerability in the software infrastructure used to host and render a website can be exploited by the attacker in order to insert malware somewhere within the site structure, which will, in turn, infect all the visitors of that site with browsers vulnerable to the particular exploit leveraged by the malware. In a recent example of a well timed and targeted attack, the website of former Beatle Paul McCartney was infected with a particular crimeware toolkit right before a fundraiser concert in New York city which was widely publicized as the first reunion of Paul

McCartney and Ringo Starr in seven years [7]. The malware was inserted in a web page but hidden from view in an iFrame and targeted vulnerable browsers by loading a rootkit on the visitor's machines.

In many cases, a user with a vulnerable browser or plugin has to do little more than visit an infected web page to have his or her machine compromised, as was illustrated in the example above, and the browser history will contain a navigation activity record before, during, and after the compromise. This history can be used to identify, or at least narrow down, the origin of the attack.

### 1.1.2 User as Insider Threat

Users who are allowed by their employers to browse the WWW have become a major security risk for their organizations, both from being victimized by existing web browser vulnerabilities as outlined above, but also by using the web browser to perform malicious or illegal acts. The results of the 2008 report assembled by US-CERT, Department of Homeland Security and the Secret Service, titled "Illicit Cyber Activity in the Government Sector," found 149 cases across 12 of 13 critical infrastructure sectors involving insiders [8]. The majority of those insiders had very limited technical skills and were historically policy-abiding employees: 58% were in administrative or support positions, 84% had no recorded incidents of violating policies and 85% had authorized access to company systems and networks.

With 250 million confidential records reported lost or stolen in 2008 alone [9], data spillage or leakage is becoming increasingly common and may be the result of intentional or unintentional actions on the part of those responsible. Common acts by insiders that result in data spillage are cross-domain information postings, such as using a personal blog or email to communicate sensitive or classified information, or posting sensitive documents in public locations. In a report on the insider threat released by Cisco Systems [10], 63% of employees admit to using a work computer for personal use every day and 78% accessed personal email from the same computer. The results of such user behavior were highlighted by a 2006 incident in which the personal and financial details of 28,000 U.S. Navy personnel were revealed after they were posted to a public website [11]. The details surrounding the incident were not made public, but it is reasonable to speculate that a web browser was used in the process of posting the documents. Shuttling sensitive or confidential information between work and home via personal email or web storage is another common cause of data spillage and 46% of employees admitted to having

done so when working from home, with another 13% admitting they use their personal email service to send business email because they can't connect to corporate networks [10]. Once again, the web browser history file or mailbox will contain a record of every transaction, and can be used to help identify the cause and extent of the data spillage.

### 1.1.3 User as Malicious Perpetrator

Data spillage may be unintentional and the result of otherwise well-meaning employees bypassing their organization's security policy for the sake of convenience, but a malicious user wishing to exfiltrate sensitive or classified information with ill intentions may do so with very little effort. A web browser could be used to post photos of documents or screen captures to public photo-sharing sites, or the materials could be forwarded to an account owned by the malicious user with a mail client. In a recently publicized example of such an incident [12] a Countrywide Financial employee exfiltrated and sold 2 million customer records over a period of two years. Using computers at Kinko's business center stores he distributed them to his buyers by mail, leaving a record of each transaction on each machine. Another common reason for the forensic analyst to scrutinize web browser and mail histories is the downloading or sharing of illegal content such as pornography or copyrighted material using the WWW or mail.

## 1.2 Purpose of Study

The constantly increasing size of storage media and the ever growing amount of stored content makes forensic analysts' jobs time-intensive and difficult, forcing them to prioritize their analyses. The importance of the web browser and mail client in peoples' personal and professional lives results in large volumes of historical information which may be of great value to the forensic analyst. Manualy analyzing these histories, working through thousands of navigation records or mail messages individually, may not be feasible unless the analyst has a specific piece of information he is looking for. Even with unlimited time to perform this analysis, a manual analysis of such a large dataset is unlikely to result in the distillation of any useful knowledge or high-level understanding of the user's habits and behaviors.

This thesis proposes the design and development of a set of tools to ingest, parse and generate a brief summary report of the contents of web browser and mail histories. Visualizations will be used in the report to condense and present the extracted information into a form easily understood by analysts, allowing them to quickly form a high-level profile of the user by discovering trends, patterns and relationships in his browsing and communication activity.

Armed with such profiles, the analysts can return to a more in-depth analysis and investigation of the histories or other sections of the media with a set of priorities and potential targets.

## 1.3 Organization of Thesis

The remainder of this thesis is organized as follows. In Chapter 2 we provide some background on web browser history file formats and information visualization, along with an overview of existing forensic tools used in the analysis of web browser and mail histories. In Chapter 3 we describe the goals and desired features of the tool built to generate a visual summary of Mozilla Firefox v3 history files. Chapter 4 details a model for generating visualizations from data and discusses the visualizations presented in the summary report within that context. In Chapter 5 the tool is applied to a web browser history and the amount of knowledge gained from the resulting output is compared to a manual analysis of the original history file. Finally, in Chapter 6, we present our conclusions and describe some possibilities for future work.

THIS PAGE INTENTIONALLY LEFT BLANK

# CHAPTER 2:
# Background and Existing Tools

## 2.1 Browser History File Formats

Web browser developers have used different storage formats for the navigational history file over the years. Microsoft Internet Explorer uses a closed proprietary binary format which has been successfully reverse engineered and is well understood [13]. The Netscape family of browsers used the Berkeley DB format from versions 2.0 to 4.0, but Mozilla migrated to a newly designed format named *Mork*, which uses markup in a plain text format to describe uninterpreted binary content [14, 15]. The *Mork* format persisted through Mozilla Version 2 and with Version 3 the *mozStorage* SQLite API was introduced which employs the SQLite 3 embedded single-file relational database for all historical information storage [16, 17]. All historical browsing information is stored on disk in the `places.sqlite` database. Common filesystem locations of the mentioned history files on various operating systems can be seen in Table 2.1 [18].

The `places.sqlite` database contains 11 tables, of which only `moz_places` and `moz_historyvisits` are needed in order to extract the information that will be used for forensic analysis by this utility. Figure 2.1 shows the two tables and the relationship between them.

| Browser | OS | Filesystem Location |
|---|---|---|
| Internet Explorer | Windows | `%systemdir%\Documents and Settings\%username%\ Local Settings\History\history.ie5` |
| Mozilla Firefox 3 | Windows | `%systemdir%\Documents and Settings\%username%\ Application Data\Mozilla\Profiles\%profile%*.slt` |
| | Linux | `/home/%username%/.mozilla/firefox/%profile%/` |
| | OS-X | `/Macintosh HD/Users/%username%/Library/Mozilla/ Firefox/Profiles/%profile%/` |

Table 2.1: Common browser history file locations

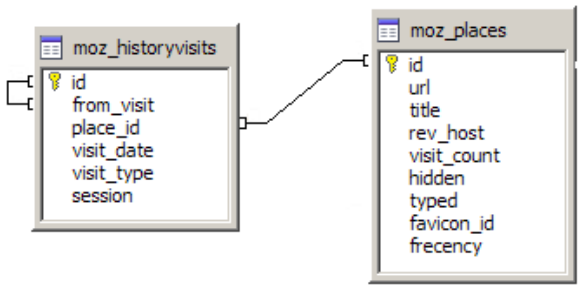



Figure 2.1: Mozilla Firefox `moz_places` and `moz_historyvisits` history tables

## 2.2 Mailbox File Formats

The storage of electronic mail, referred to as mail for the remainder of this document, for use by client utilities is file-based and may be maintained in a number of different formats. An overview of the three most popular formats follows.

### 2.2.1 mbox

*mbox* is a file format for the storage of mail in which messages are stored in a single text file in sequence. Each message begins with a *From_* line followed by a message in RFC2822 format [19], and ending with a blank line which must be nothing other than a newline character. The *From_* line is any line consisting of the characters {F,r,o,m} in that order, followed by a space [20]. An example message in mbox format:

```
From ggroussa@nps.edu   Mon Jun 15 18:13:12 2009
Return-Path: <ggroussa@nps.edu>
Subject: This is the subject
From: Greg Roussas <ggroussa@nps.edu>
To: DJB <djb@cr.yp.to>
Content-Type: text/plain
Date: Mon, 14 Jun 2009 18:13:12 -0800

This is the first line of the body.
...
This is the last line of the body, followed by a blank line.
```

The specific *mbox* variant described above is the *mboxd* format made popular by Rahul Dhesi in June 1995, but there are many different variations of the *mbox* format including *mboxo*, *mboxrd*, *mboxcl* and *mboxcl2* [20]. Support for reading and writing mailboxes in *mbox* format is widespread among mail clients and there are APIs for many popular programming and scripting languages for parsing mbox-formatted mailboxes [21, 22].

### 2.2.2 Maildir

Originally implemented in the Qmail SMTP server to address reliability and performance issues inherent with the single-file *mbox* format, mailboxes in the *maildir* format store each message in a separate file with a unique name [23]. The name is generated by the Mail Transfer Agent (MTA) [App A] or Mail User Agent (MUA) [App A] responsible for delivering the mail, and consists of three period-delimited sections: the output of the `time()` or the second counter of the `getimeofday()` system call, a delivery identifier and the result of the `gethostname()` system call [24]. As with the *mbox* format, there exist APIs for many popular programming and scripting languages for parsing maildir-formatted mailboxes [21, 25].

### 2.2.3 Personal Folder File (PFF)

This format was developed by Microsoft for the storage of mail, appointments, tasks, contacts and notes by the Outlook personal information manager. The *PFF* is used in the following three file formats:

- Personal Address Book (PAB): used to represent an address book containing contacts. Files have *.pab* extension.
- Offline Storage Table (OST): used to contain mail, appointment, tasks, etc. in an offline format by Microsoft Exchange. Files have *.ost* extension.
- Personal Storage Table (PST): used to contain mail, appointment, tasks etc. in active and archived mailboxes. Files have *.pst* extension.

The *PFF* format has been studied in detail and successfully reverse-engineered with libraries available for the parsing, listing and conversion of PST-format mailboxes to *mbox* and other formats [26, 27, 28, 29].

## 2.3 Information Visualization

### 2.3.1 Background

Visualization is used extensively in many scientific disciplines such as chemistry and the biological and the earth sciences as a tool to better understand the data and discover relationships that may exist in large datasets. *Information visualization* is a more recent application of visualization to abstract datasets and may be formally defined as the "use of computer-supported, interactive visual representations of abstract data to amplify cognition" with the main goals of discovery, decision making and explanation [30]. Abstract data, for example financial data, information flows or the structure of a computer program, lacks physical structure and, as a result, does not have any straight-forward spatial mapping which may be used to render it in visual form. It is therefore of utmost importance to design an appropriate model to map the information to its visual form in order to allow the visualization to capture the essence of the information and provide added value.

### 2.3.2 Cognition Amplification

Cognition is "the acquisition or use of knowledge" [30] and it can be made more efficient or "amplified" by the visualization of information in six major ways: [30]

1. by increasing the memory and processing resources available to the users
2. by reducing the search for information
3. by using visual representations to enhance the detection of patterns
4. by enabling perceptual inference operations
5. by using perceptual attention mechanisms for monitoring
6. by encoding information in a manipulatable medium

Human memory and processing resources are maximized by reducing the memory load and encoding the information in a visual format accessible immediately by the viewer. Instead of searching for related information throughout the entire data set, similar information may be grouped in the visualization facilitating its simultaneous analysis. This allows large sets of

identical information to be aggregated and represented in a more compact form. Patterns that would be difficult or impossible to discern in the raw data may be discovered in a visualization which presents a higher-level overview of the information. The process of visualizing abstract datasets and its advantages are discussed in more detail in Chapter 4.

### 2.3.3 A Forensic Application of Visualization

The traditional and current approach to maintaining web history and mail stores is file-based and relies on a textual presentation to the user. The manual analysis of datasets generated from these files is largely a sequential or random-access process of scanning, searching and filtering, in an effort to understand the contents or locate a specific piece of information. This approach is tedious and inherently not scalable by its very nature, and may require the resources of a highly experienced or well-trained analyst who is short on time and long on workload. The appropriate use of visualization to represent and explore such a dataset leverages the cognition-enhancing features discussed above and takes advantage of the parallel and preattentive nature of the visual-spatial cognitive modality [31]. Preattentive processing is a method derived from human cognitive psychology by which humans organize what is in their visual field based on cognitive operations believed to be rapid, automatic and spatially parallel [32]. Hue, intensity, orientation, size and motion are all examples of visual features that can be detected in this manner [32].

The forensic analyst attempting to gain a high-level understanding of the browsing or communication behavior contained in a dataset generated from a web browser history or a mail store will benefit from the advantages conferred by visualizing the data. Aggregation and grouping of select elements of forensic interest will allow the analyst to quickly develop an overview of the information and the underlying behavior of the user who generated the data. The patterns that may be displayed in a visualization will allow the analyst to make inferences about the user and suggest further areas on which to focus the investigation. Due to its high-level nature, the use of visualizations in the analysis may not be the final step, but when used in conjunction with the original it will improve the accuracy of the analysis and help in formulating a final determination. The application of visualization to web browsing and mail histories with the goal of discovering forensically interesting information is discussed in more detail in Chapters 4 and 5.

## 2.4 Current Tools

### 2.4.1 Web Browser History Forensics Tools

Due to the many different formats in use by web browser developers for the representation and storage of historical navigation data, there is a division in the existing forensic tools between those offering support across multiple formats and those targeted to a specific format. The tools may be organized into those that support either the Internet Explorer or the Mozilla Firefox history file formats, and those that support the majority of the existing major browser formats including Internet Explorer, Mozilla Firefox, Safari, Opera and Chrome. The tools supporting specific formats tend to be freely available and in some cases open source, whereas those supporting multiple formats are commercial and closed-source.

| | **Web Browsers** | | | | | | |
|---|---|---|---|---|---|---|---|
| **Utility [Ref]** | **IE** | **Moz v2** | **Moz v3** | **Opera** | **Safari** | **Licensed** | **Latest** |
| FTK 2.x [33] | ✔ | ✔ | ? | | | ✔ | 2008 |
| EnCase [34] | ✔ | ✔ | ? | ✔ | ✔ | ✔ | 2008 |
| CacheBack 2.0 [35] | ✔ | ✔ | | | | ✔ | 2008 |
| NirSoft IEHistoryView [36] | ✔ | | | | | | 2008 |
| FoxAnalysis [37] | | ✔ | ✔ | | | | 2008 |
| Firefox 3 Extractor [38] | | | ✔ | | | | 2008 |
| Passcape Password Recovery [39] | | ✔ | | | | | 2008 |
| NirSoft MozillaHistoryView [40] | | ✔ | | | | | 2007 |
| NetAnalysis [41] | ✔ | ✔ | ✔ | ✔ | ✔ | ✔ | 2007 |
| Index.dat Viewer and Zapper [42] | ✔ | | | | | ✔ | 2007 |
| X-Ways Trace [43] | ✔ | ✔ | | ✔ | | ✔ | 2007 |
| Firefox Forensics [44] | | ✔ | ✔ | | | ✔ | 2007 |
| Cleanersoft IEHistory [45] | ✔ | | | | | | 2006 |
| Index.dat Analyzer [46] | ✔ | | | | | ✔ | 2006 |
| Enhanced History Manager Add-on [47] | | ✔ | | | | | 2006 |
| Browser History Viewer (BHV) [48] | ✔ | ✔ | | ✔ | ✔ | | 2006 |
| Web Historian [49] | ✔ | ✔ | | ✔ | ✔ | | 2005 |
| Pasco [50] | ✔ | | | | | | 2004 |
| mork.pl [51] | | ✔ | | | | | 2004 |

Table 2.2: Forensic tools for web browser history display and analysis

The tools shown in Table 2.2 support opening and parsing the history file, but additional analytic functionality varies greatly from one tool to the next. At one extreme of the feature-set is the commercial tool CacheBack 2.0 which rebuilds web pages in history from cache, builds thumbnails, performs link analysis, allows filtering based on file type, extension and hostname, and exports to a delimited text or html report. At the other end is a tool like NirSoft IEHistoryView

[36] or MozillaHistoryView [40] which displays the extracted information in a tabled format, allows sorting by column and supports export to a text-delimited, HTML or XML file, but lacks any other analytic or reporting functionality. Figure 2.2 shows the typical interface of two utilities — Passcape History viewer and Nirsoft MozillaHistoryView — with their tabled output format [39, 40].

EnCase, the commercial forensic suite by Guidance Software, supports the ingestion and parsing of various web history file formats shown in the second column of Table 2.2. It offers a general timeline visualization which can be applied to most types of data, ranging from filesystem contents to web history. Shown in Figure 2.3 is a view of EnCase's timeline visualization where each event is represented by a green square. It is easy to see that with the many thousands of navigation events in web browser history files, this visualization would quickly become occluded to the point that it offered very little value.

There have been various approaches to visualizing different portions of web browsing in academia, but none that do so comprehensively and with a forensic perspective. The focus in the academic literature is more on the novelty of the visualization, not on the aggregation and summary presentation of specific elements of the history using common visualizations. As an example, the *WebViz* utility aims to provide a view of the types of subjects people are browsing for [53], and using the URL and visit count, the visualization shown in Figure 2.4 is generated. The category of each URL is looked up in the Dmoz Open Directory Project (ODP), discussed in detail in Section 3.3, and visit counts to each category are calculated. Each circle in the visualization represents a time interval, each category a distinct color, and saturation and brightness of each circle indicate the visit frequency.

Network link graphs have been used heavily both in the academic literature and the website architecture and optimization sectors [54]. Examples of 2D and 3D implementations shown in Figure 2.5 include the *StarTree* visualization by SAP Software [55] which represents site structure using 2D hyperbolic geometry, and the H3 utility from the graphics group at Stanford [56], which lays out each site as a large directed graph in 3D hyperbolic space. Visualizations have been created that are based on the network graph but incorporate additional information, such as *VISVIP* [57], part of the WebMetrics suite developed by NIST, which represents the path using a smooth curve showing the trajectory and direction of movement and time spent on each page. A sample output from *VISVIP* is shown in Figure 2.6.

Even more abstract visual metaphors have been developed based on the network graph. One such example is the *Anemone* visualization developed at the MIT Media Lab by Ben Fry as part of his thesis on *organic information design* [58]. The ideas behind *Anemone* can be used to visualize any information flow, and when applied to navigation paths the visualization shown in 2.6b is produced. Sites or pages with high visit counts produce thick nodes or lines compared to less visited pages which are rendered with thin lines. The many paths through the network are juxtaposed with lighter lines. This class of visualizations is very visually appealing and attracts heavy attention in the design sector, but has not yet been integrated into commercial information visualization tools.

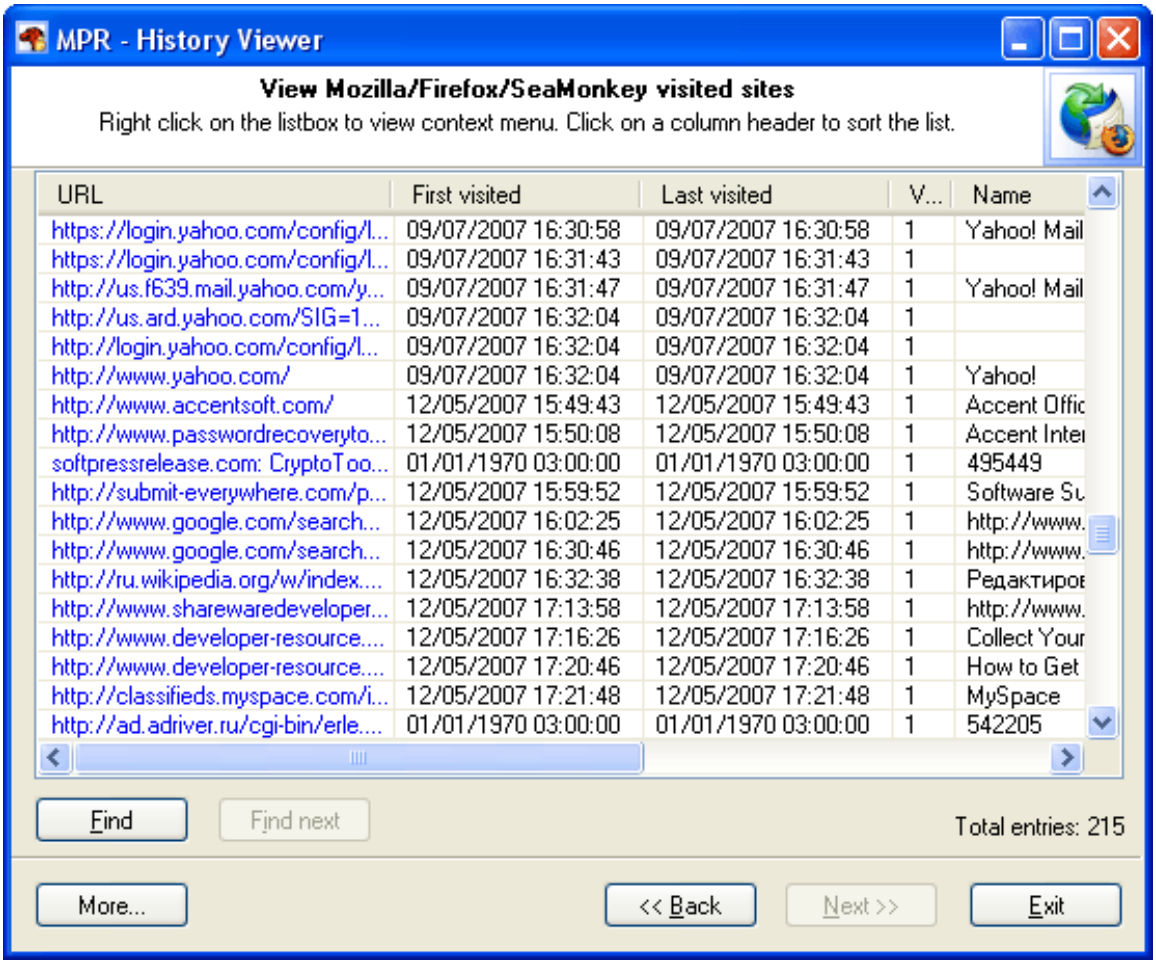
### 2.4.2 Mail Forensic Tools

The two large commercial forensic analysis software suites, EnCase and FTK, both support the extraction and presentation of mail. All the common mailbox formats are supported including *mbox*, *Maildir* and the Microsoft *PFF* format, as well as web-based mail such as Yahoo! and Hotmail [59, 33, 34]. The mail content can be presented in a report format, and most of the components of the ingested mailboxes can be searched and filtered based on various fields. Network forensic utilities such as PyFlag are able to reconstruct the full contents of web navigation and webmail sessions, but require full network packet dumps in order to do so [60]. None of these utilities offer any application-specific visualization functionality for mail, and present the information in a report or tabled format similar to that shown above for web browsing history. In the academic literature, analysis of mail communication falls under the study of social or organizational network analysis which is different in nature than simple navigation records due to the connections between multiple communicating parties [61]. The use of visualization has a long history in this field of study, and the types of visualizations are very similar in appearance to those presented above for site architecture. Many of the visualizations presented in the academic literature are integrated in some fashion into commercial software applications, as shown in Figure 2.7. Figure 2.7b shows the detailed view of the communication graph generated by the InFlow social network analysis tool [62]. This particular visualization, representative of this class of communication graph visualizations, portrays each communication participant as a square node whose color is keyed to the participant's department and the gray edge indicates communication volume above a certain threshold. Examples of similar representations can be seen in Figure 2.8 which shows the *Email Mining Toolkit* "clique" visualization [63] and the Enron Corpus Viewer [64] showing a particular view of a portion of the Enron mail archive.

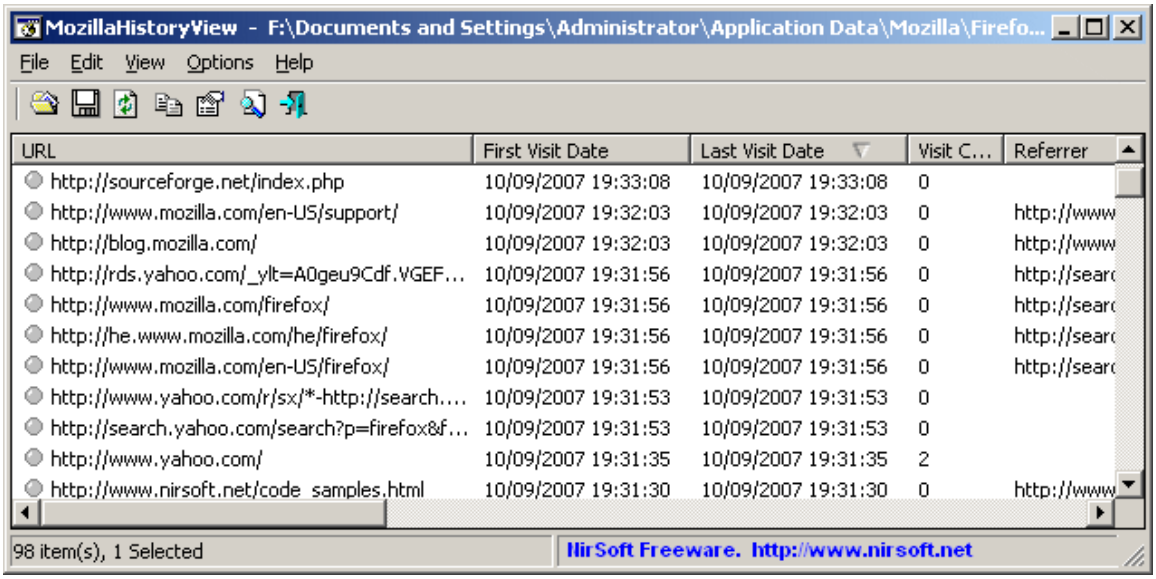
There are also a large number of visualizations that don't fit into any particular class due to their novelty. The *Thread Arcs* visualization by IBM [65] shown in 2.9a displays the branching tree structure of individual communication threads, and was designed to highlight connections between messages and people. The "Author Lines" visualization [66] in Figure 2.9b shows patterns of communication activity over time in terms of the initiation of new conversations and the replies to those conversations. The initial conversations are represented as vertically stacked bubbles above a center line and replies appear below the center dividing line. The conversational history visualization generated by *Themail* utility seen in Figure 2.9c shows the exchange between two parties over time with frequently occuring words stacked vertically based on count.

## 2.5 Summary

This chapter presented background information on the different browser history file and mailbox formats used by various browsers and mail clients. The field of Information Visualization was introduced and the different ways in which visualization can be used to amplify cognition were outlined. Visualization was discussed in the context of forensic analysis. Finally, current forensic tools supporting the analysis of web browser histories and mailboxes were discussed, and the visualization of these histories in the academic literature was covered. In the next chapter I will describe the work that will be done prior to the development of the utilities central to this work, including stakeholder descriptions, high-level goals and proposed system features.




(a) Passcape History Viewer




(b) Nirsoft MozillaHistoryView

Figure 2.2: Tabled output format of two current forensic utilities

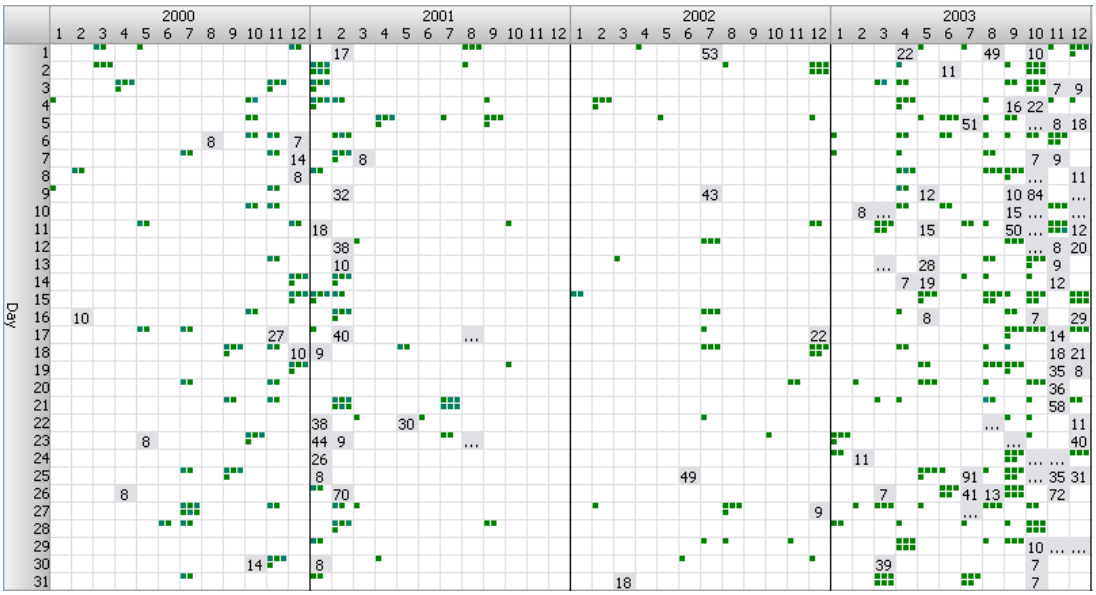



Figure 2.3: EnCase timeline view From [52]

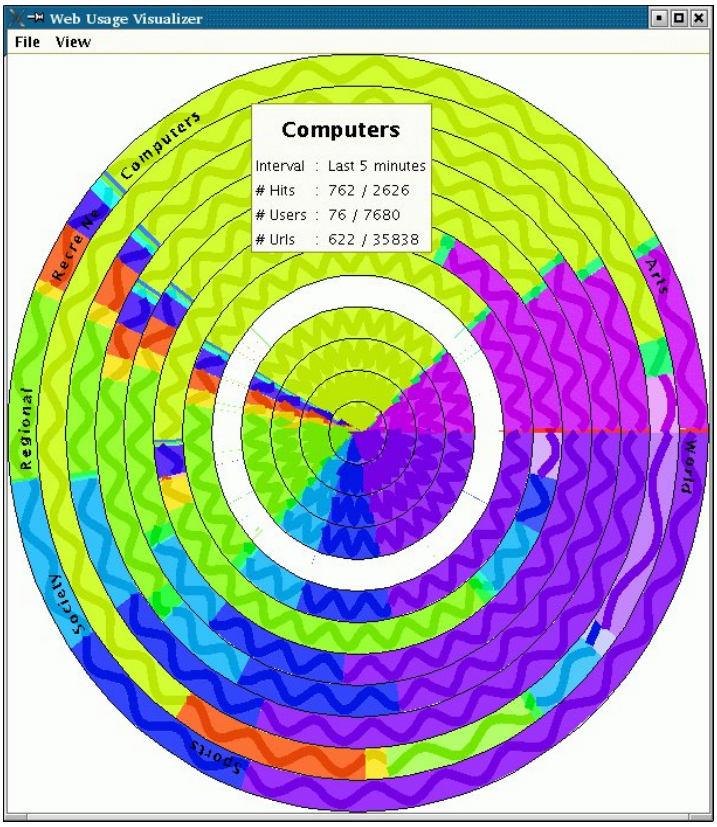
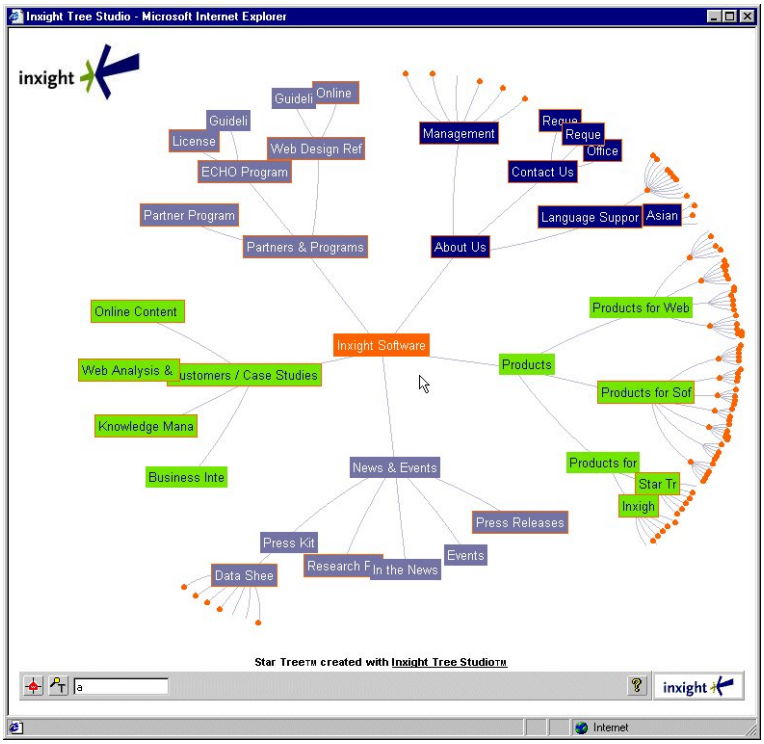
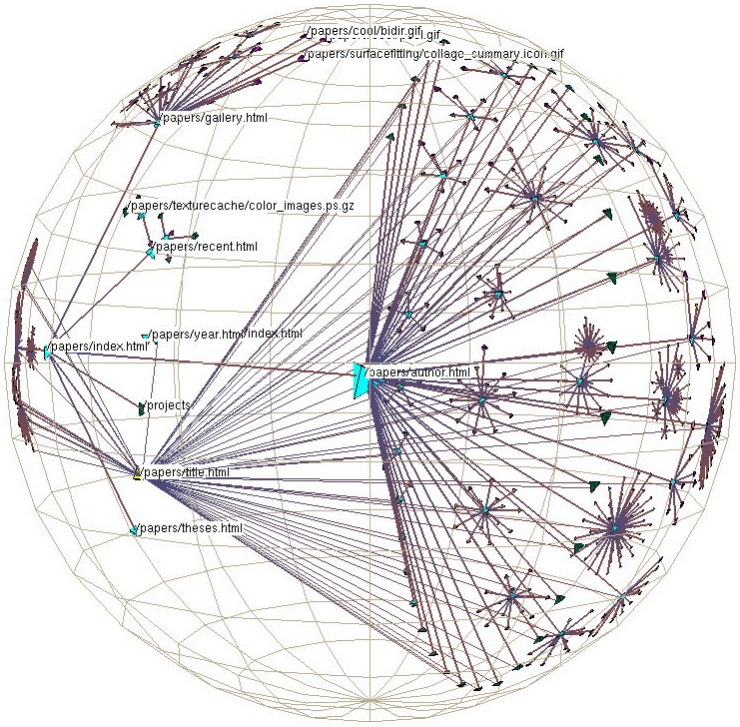



Figure 2.4: WebViz visit categories web visualization From [53]

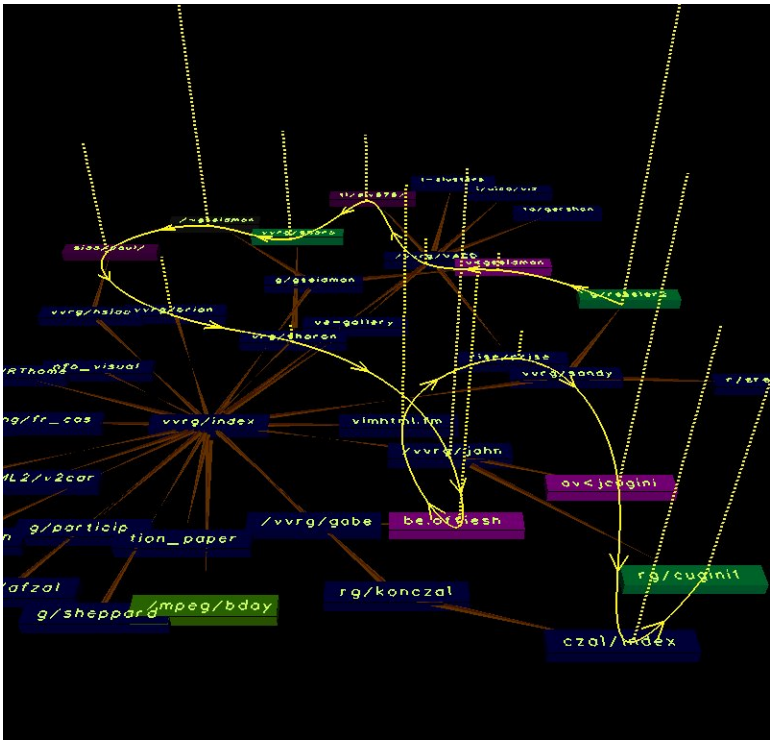



(a) SAP StarTree 2D Hyperbolic Tree




(b) H3 3D Hyperbolic Tree

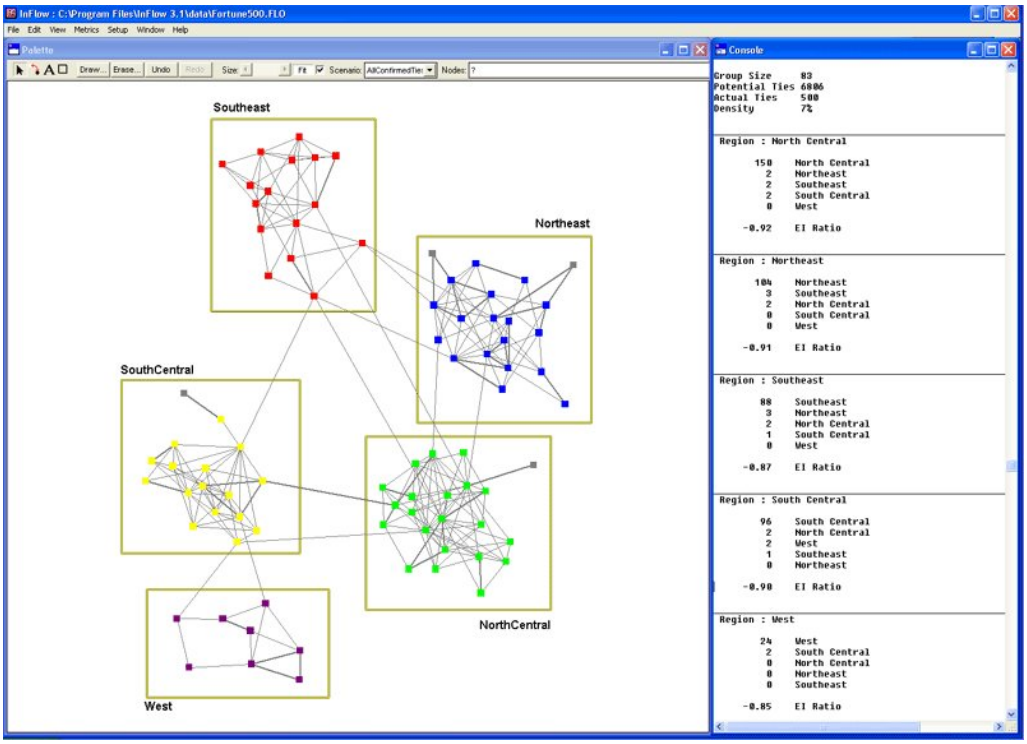
Figure 2.5: Two different network graph layouts used for web visualization



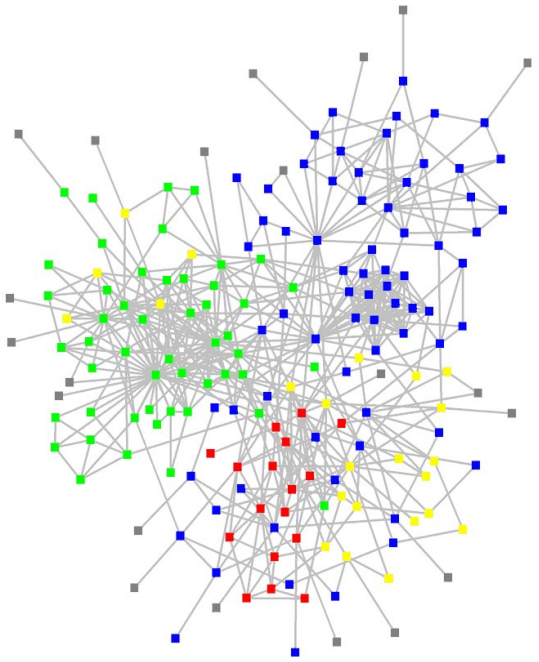

(a) VISVIP navigation path

(b) Anemone path visualization

Figure 2.6: Two visualizations based on the network graph applied to web traffic

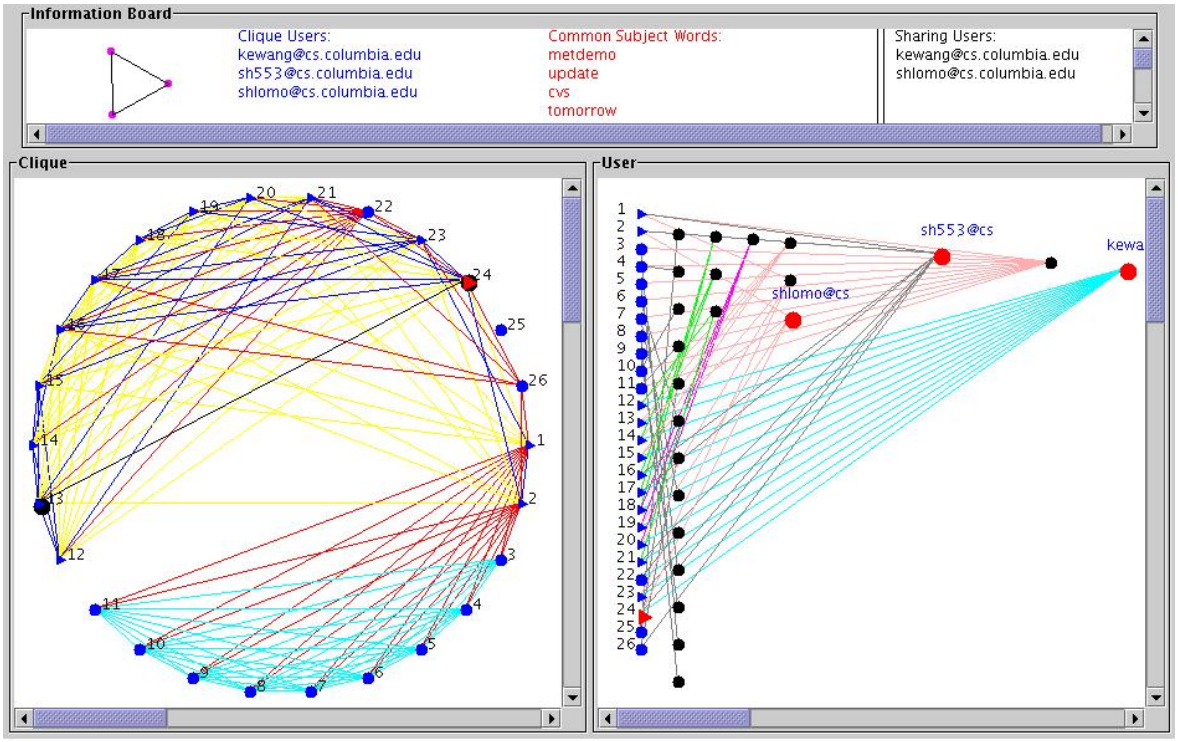



(a) InFlow network visualization interface

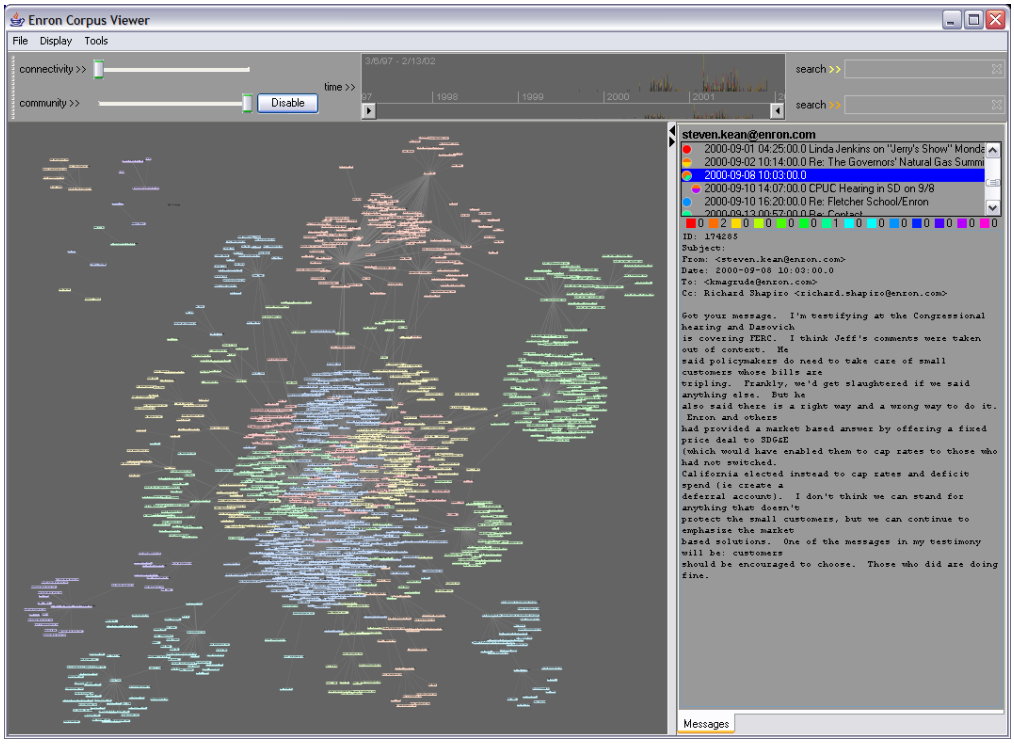


(b) Inflow communication graph detail

Figure 2.7: Visualizations based on the network graph applied to mail From [62]



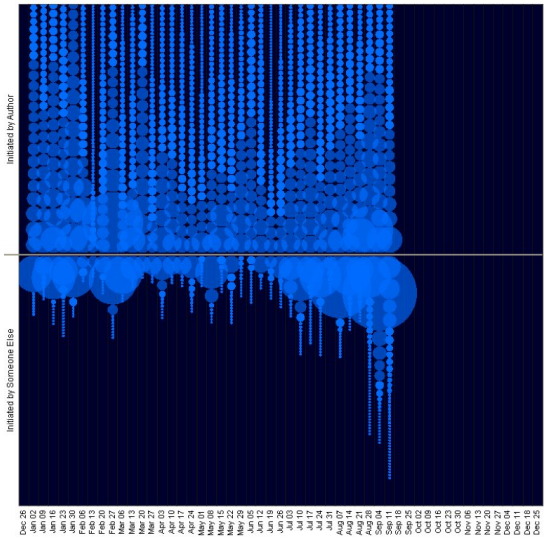

(a) Email Mining Toolkit clique analysis



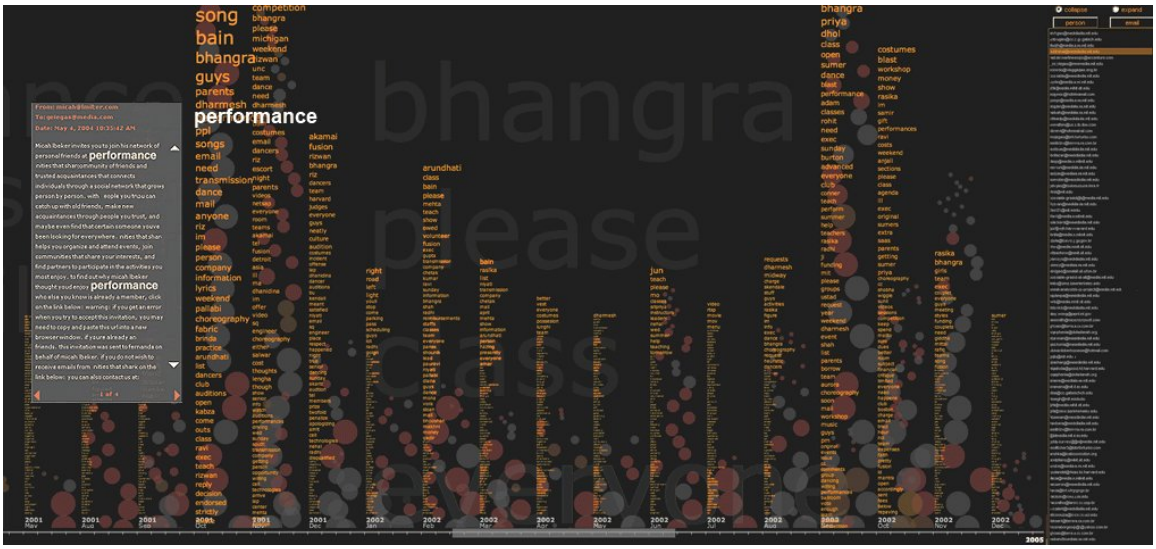

(b) Enron mail archive community view

Figure 2.8: Visualizations based on the network graph applied to mail From [63, 64]

(a) IBM Thread Arcs[67]




(b) “Author Lines” visualization



(c) Themail conversational history

Figure 2.9: Novel visualizations applied to mail

# CHAPTER 3:
# Features, Requirements and Design Overview

## 3.1 Key Stakeholder and High-Level Goals

### 3.1.1 Stakeholder Descriptions

The high-level goals of this toolset are dependent on the needs of the stakeholders — those who care about the system and have a stake in its function — and the context in which it is used. In order to define these goals, the stakeholders and their goals were identified, and then those goals were applied to the problem domain to define the high-level goals of the application itself.

As a case moves through different phases of discovery and analysis during the course of an investigation, there will be users of varying technical knowledge and ability at each phase who need access to the information at different levels of depth and understanding. Because of the diverse levels of technical sophistication of the potential users and the requirement that the underlying information be made accessible to each of them, the report that is produced must strike a balance between depth and density of information and simplicity of presentation. Good user interfaces (UI) are easy to use and understand, meet the needs of the intended users and support the users in the tasks they wish to undertake [68]. In the case of this toolset, the UI is composed of two different elements: the interface to this utility used to extract, manipulate and generate summary reports, and the interface used to view those reports. Each of these interfaces has different requirements due to the specific needs of each target user. The utility to extract, manipulate and generate summary reports will be used by the forensic analyst who is presumed to have had a reasonable amount of exposure to commonly used forensic utilities and therefore be of a higher level of technical sophistication than the case manager or division head who simply wishes to view and understand the final report.

The overarching goal of this utility, and in fact any software application that aims to be accepted by its users, is usability, defined in ISO 9241-11 as "the extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use." [69] Effectiveness refers to the accuracy and completeness with which specified goals in specified environments can be achieved. Efficiency is a measure of the resources expended with relation to the accuracy and completeness of the achieved goals,

and satisfaction is the comfort and acceptability of the system to all its stakeholders [69]. It is possible for an application to be usable in one context but unusable in another, so in order to attempt to identify all the contexts in which the system would be used and ensure it satisfies the usability requirement in those contexts, the potential stakeholders were identified. The identification of stakeholders for this utility was the result of converstations with a forensic analyst actively working in the field who has experience in both developing utilities for others and making the output of those utilities accessible to all stakeholders. Each of these stakeholders will make use of this utility and the report which it produces in a different context, so each stakeholder and their respective qualitative goals are outlined in Table 3.1.

| **Stakeholder** | **Goals** |
|---|---|
| All | accurate results<br>ensure integrity of source data<br>consistent structures, language and typographic conventions<br>high-quality print-ready report<br>easily-comprehended visualizations<br>ability to save visualizations for use in custom reports or presentations |
| Analyst | ease-of-use, minimal interaction<br>automate information extraction and report generation<br>readily accessible profile of user's online activities<br>flexibility to modify both report generation process and resulting report |
| Case Manager | quick turnaround of newly acquired data<br>efficient processing of case material<br>easily understood reports |
| Government Agency | potentially court-admissible evidence |

Table 3.1: Stakeholder goals

To satisfy the goals in Table 3.1 across all contexts in which the application is to be used and fulfill the usability requirements of effectiveness, efficiency and satisfaction for all stakeholders, a number of design decisions were made. The report generation utility that will be used by the forensic analyst will be a UNIX command-line oriented tool allowing it to be used in environments in which GUI-based operating systems are absent. Application functionality will be accessed through the use of options or command-line switches requiring no interaction once executed, making the use of the application in different contexts amenable to automation. The application will be written in Python, an object-oriented, interpreted scripting language with a clean and very readable syntax [70] allowing full transparency into the methods used and easy modification or customization.

The report will be generated using a template-based system and presented as a web page structured in plain Hypertext Markup Language (HTML) styled with Cascading Style Sheets (CSS) and viewable in any web browser. The web browser may be chosen by the viewer of the report and will present a familiar interface allowing him to concentrate on the content. The use of HTML to structure the report satisfies two requirements: the transparent and easily modified nature of HTML markup will allow the analyst to make modifications or customizations to the report and the template-based approach to report generation allows any changes to be made in a single place in order to be applied to all generated reports. The appearance and all stylistic elements (e.g. fonts, colors, spacing) of the report are achieved through the use of a separate CSS style file which may also be easily modified or customized to suit the desires of the analyst or department. The template-based nature of the report structure and the separation of its structural and stylistic elements also satisfy the requirement to present consistent structures, language and typographic conventions, by maintaining a single place where changes may be made that will have a global effect.

The visualizations that are generated and included in the report will be Scalable Vector Graphics (SVG), an XML-based language for describing two-dimensional graphics developed by the World Wide Web Consortium (W3C) [71]. Since the graphics are represented using XML, they can be easily modified with any text editor, and have the ability to scale up or down in size with no loss of resolution or change in file size. This enables the viewing of the generated visualizations with an SVG viewer, allowing a more detailed and focused analysis using features of the viewer such as filtering, selection and magnification of specific areas of interest. The reasons behind the choice of specific visualizations is further detailed in Section 4.1.2, but with regard to the requirements, the visualizations must be easy to comprehend by everyone involved in the analysis process, from the forensic analyst to the management-level stakeholder to whom the analyst may have to present and explain the results. For this reason, commonly used visualizations were chosen that could easily be understood by the average person with no specific training in mathematics or information visualization. These included tables, bar charts, pie charts, histograms, and time-series plots.

Having addressed the goals from the perspective of the user, the application of those goals to the problem space will result in the high-level goals of the application itself. Table 3.2 describes each of the high-level goals of this toolset along with problems and concerns that may affect the successful implementation of each goal.

| High-Level Goal | Priority | Problems & Concerns |
|---|---|---|
| Efficient, automated processing of history file | high | Longer processing times as history filesize increases<br>Inability to process corrupted files |
| Easily-comprehended visualizations | high | Depends on sophistication level of analyst<br>Graphing packages suited for sophisticated modeling |
| Generation of high-level comprehensive report | high | Report format lacks interactivity<br>May be large depending on size of input file |
| Accurate calculation of time summaries | high | Corrupted or unreadable timestamp entries<br>Depends on accurate timestamp at time of visit or message delivery<br>Trust in system clock of target computer |
| Accurate geolocation of domains | medium | Depends on existence in GeoIP database<br>Depends on accuracy of GeoIP database |
| Accurate derivation of website categories | low | Depends on existence in category database<br>Depends on accuracy of category database |

Table 3.2: Key high-hevel goals

Each of the high-level goals outlined in Table 3.2 will be further refined by mapping them to features that will be supported by the application. The next section details the process by which this was performed and the features that were identified.

## 3.2 System Features Summary

The features to be provided by the application were derived based on the nature of the data being analyzed, high-level goals common to all forensic software and features offered by current forensic tools in the space. Through the use of various visualizations, we hope to supplement the extracted information and convey the maximum amount of information in the least amount of space. Various techniques used in the *Unified Software Development Process* were applied such as use cases, supplementary specifications and the development of a high-level domain model. The selected use cases that were written during requirements analysis provide an early examination of how the application would be used and the manner in which the users would interact with it.

The *Unified Software Development Process* or *Unified Process* (UP) is a model or framework for developing software in an iterative and incremental method that was first proposed by Jacobson, Booch and Rumbaugh in 1999 [72]. The UP makes heavy use of the Universal Modeling Language (UML) and bases the software development lifecycle from inception, requirements and design to implementation and testing on *use cases* and an iterative and incremental development process. In contrast to earlier models that enjoyed popularity until recently, such as the sequential or Waterfall model, the UP emphasizes early development of high-risk functions

and constant testing, resulting in working portions of the final product at each step. Agile development methods are emphasized, including timeboxed iterative and evolutionary development, adaptive planning and incremental delivery, encouraging rapid and flexible response to change [73]. This approach allows for the mitigation of problems that may arise early in the development process as opposed to at the time of product delivery when it becomes much more difficult to make changes.

During the features discovery phase, the basic functionality to be provided by the application was motivated by high-level goals of computer forensic software as well as features that are currently implemented across various forensic tools and expanded to include visualizations. The primary function of computer forensics software is the automation of trivial processing activities in order to assist the investigator in his analysis and reduce investigation time and complexity [74]. Freed from the burden of having to manually process large amounts of information, the analyst can use any insight gleaned from the output of the particular tool to guide or focus the investigation. Additionally, the high-level overview offered by an extraction and summarization utility such as this, will assist the analyst in forming a profile of the user behind the history files. The high-level features to be provided by the application are outlined below.

- Support for extracting history data from native browser format (Mozilla Firefox v3)
- Support for extracting history data from native mailbox format (mbox)
- Support for separating the extraction phase from the processing or report generation phase
- Extraction, calculation and summarization of various browsing attributes:
  - full and base URL, referrer, countries, categories, search queries
  - navigation events by time period (hourly/daily/monthly/yearly)
  - navigational events aggregated by protocol, top level domain
  - navigational trigger - bookmark, history, typed, embedded
- Extraction, calculation and summarization of various mail attributes:
  - message sender and recipient addresses, times
  - top senders and recipients, domains, countries
  - messages sent/received by time period (hourly/daily/monthly/yearly)

- Derivation and presentation of forensically relevant information:
  - extract search queries from URL
  - derive geographic location of website and domain of message sender/recipient
  - derive category of website visited
- Summarization of results in intuitive and easily-comprehended visualizations
- Presentation of report in compact and portable format
- Storage of results in a transparent format accessible using easily accessible open-source tools on multiple operating systems

These features will be implemented in the tool, which will have the ability to generate a final summary report striving to meet all the goals outlined above, thereby allowing each stakeholder to comprehend and extract the necessary relevant information. With this set of features, it is hoped that the analyst will be spared the task of manually processing the information and have results that support an informed determination of the areas on which to focus the next phase of his investigation.

## 3.3 Choice of Extracted Data

The desired user and high-level goals and features outlined in Sections 3.1 and 3.2 are primarily concerned with accurately and summarily portraying a high-level view of the user's activity and providing the analyst insight into the data without overwhelming him. This requirement dictates the selection of relevant data to extract from the available sets of fields in the web browsing and mail history files. Data in both histories that is necessary to depict the user's activity may be divided into two general categories: navigational or communicational, and contextual. Contextual information may be further decomposed into temporal and informational.

In order to accurately and completely visualize the user's activity, the URLs of visited sites and the mail addresses of those with whom the user communicates must be extracted. This navigational or communicational category includes the *Session*, *Visited URL* and *Referring URL* fields from the web browsing history and the *Source Address* and *Recipient Address(es)* from

the mail history. The extraction of these fields will allow a reconstruction and visualization of the users' navigational and communication history providing the analyst with a view of their online activity.

To provide temporal context, the *Visit Timestamp* and *Message Timestamp* must be extracted, which serve in placing the navigation event or communication somewhere in the linear temporal continuum. This temporal context provides the analyst with the ability to view a high-level summary of the volume of navigation or communication the user was engaged in over time. This will facilitate the identification of time periods exhibiting out-of-the-ordinary usage patterns, potentially highlighting an anomaly that needs to be investigated more closely.

The last category includes attributes that provide informational context to the navigation or communication record and includes the *Site Title*, *Site Visit Count*, *Site Visit Type*, whether the URL was *Typed* or *Embedded*, and *Frecency* [App A] — a metric of visit frequency and recency — from the web history. The *Sender Name*, *Recipient Name(s)* and *Message Subject* from the mail history provide similar information. Each of these fields is secondary to the information defining the navigation path or message exchange, but provides supplementary information that may be used to place that particular record or message in context, or allow the analyst to make a previously unseen connection. Each *navigation action* — an event that initiates a site visit comprising all user actions that lead to a new entry in the history [75] — is captured in the browser history and may be very useful information for the analyst, as it indicates something about the user's behavior in connection with his visit to that site. A high visit count or a visit type of *Typed* or *Bookmark* indicates revisitation frequency and familiarity with the site respectively. When combined, many of these supplementary attributes provide even deeper insight into the user's behavior. For example, a particular navigation record may gain additional significance to the analyst when the supporting informational context attributes of a high *Frecency* score and a `Bookmark` visit type are observed, highlighting a frequently and recently visited site that originated from the user's bookmarks and indicating a high degree of familiarity and interest on the part of that user.

Additional data will be derived using the extracted fields outlined above, which will provide additional context to the navigational or communication record. This data includes *Site Country*, *Site Category* and a thumbail of each site in the web browsing history, and *Sender* and *Recipient(s)* country in the mail history. This data does not exist in the original history and is derived using an external reference such as a *GeoIP database* or a *Website Directory* service

and therefore should not be given as much weight by the analyst, however it may be useful in providing additional context when supplementing the existing history.

The MaxMind GeoIP database, which is widely used for fraud detection, targeted advertisement services, traffic analytics and content customization, provides a mapping of IP address to geographic location with detail down to the latitude and longitude of each location [76]. Table 3.3 shows the level of detail resulting from a lookup of the NPS gateway `diamond.nps.edu` in the GeoIP database, although this utility will only make use of information at the country level. MaxMind provides free and paid versions of the product, both including worldwide coverage and claiming 99.5% and 99.8% accuracy respectively at the country level.

| | |
|---|---|
| **IP Address** | 205.155.65.226 |
| **Ctry Code (2)** | US |
| **Ctry Code (3)** | USA |
| **Ctry Name** | United States |
| **City Name** | Monterey |
| **Region Code** | CA |
| **Region Name** | California |
| **Postal Code** | None |
| **Latitude** | 36.3698997498 |
| **Longitude** | -121.84059906 |
| **Area Code** | 831 |
| **Organization** | California State University Network |
| **Time Zone** | America/Los Angeles |
| **Metro Code** | 828 |

Table 3.3: GeoIP resolution for diamond.nps.edu

The Dmoz Open Directory Project (ODP) will be used to associate each visited site with a category. The ODP is the "largest, most comprehensive human-edited directory of the Web" and provides the data to other large-scale web directories such as Google Directory and AOL [77]. Other URL-based classification schemes include MeURLin [78], which itself offers three different schemes, and Google Directory which uses a modified form of the ODP data, but places daily limits on the number of queries that can be made. The ODP organizes and classifies websites in a hierarchical set of categories, for example, Google appears under "Computers/Internet/Searching/Search Engines/Google," a clear human-readable category hierarchy that can provide context for anyone unfamiliar with the nature of a given site. One potential problem with using any external directory service is the accuracy of the results and the degree of coverage provided, both problems which may be encountered when working with large amounts of data. In an attempt to ameliorate this problem and provide flexibility and customizability for

the analyst, a local database will be created during the category lookups including only those mappings between URL and category for data in the browsing history. This database may be modified in any manner the analyst sees fit in order to better reflect the goals of the investigation or department.

The ODP provides a dump of the contents of its entire directory in two separate files, one describing the structure of the directory and containing the entire category hierarchy, and the other containing the content which has been categorized and assigned to a category. The data is provided in the Resource Description Format (RDF), a language with an XML syntax developed by the World Wide Web Consortium (W3C) for representing information about resources on the World Wide Web [79]. The ODP defined its description format before the official specification for RDF was finalized, so lookups against the original files are both slow due to their size — the structure file is nearly 700MB and the content file is nearly 2GB — and extremely inefficient due to divergence from the official specification. For this reason, the data will first be checked for consistency and then ingested into an SQL database. This will both increase efficiency and provide consistency with the external data store interface of this toolset.

The Thumbshots service [80] will be used to generate a thumbnail of each visited site which will be displayed alongside the site details, providing a graphical representation of the visited site as it would be rendered by a web browser. The base URL of each visited site will be sent to the remote Thumbshots server which will render and display the thumbnail. This feature will require WWW access and will need to be explicitly enabled since the analyst may have concerns about associating his source IP with each URL request.

## 3.4 Selected Use Cases

Following the feature discovery phase, requirements analysis is performed with *use cases* – narratives or scenarios describing how people will use the application [72]. By describing the use of various core functions of the application in a narrative fashion, the requirements begin to take form, and may be further defined or expanded during the elaboration phase. Prior to outlining the use cases, a high-level concept of operations is in order showing the sequence of operations that takes place upon program execution and the flow of information between components.

The operation of the application is separated into three high-level phases: the ingestion of a web or mail history file and its subsequent storage in an internal database, the calculation of

summaries and derivation of various supplemental attributes, and the rendering of the tables and visualizations and generation of the summary report. Each phase may be completed in sequence, or they can be separated by an indeterminate amount of time, allowing the analyst the flexibility to batch operations in a manner that is most beneficial and convenient to his schedule and available system resources. Figure 3.1 outlines this workflow.

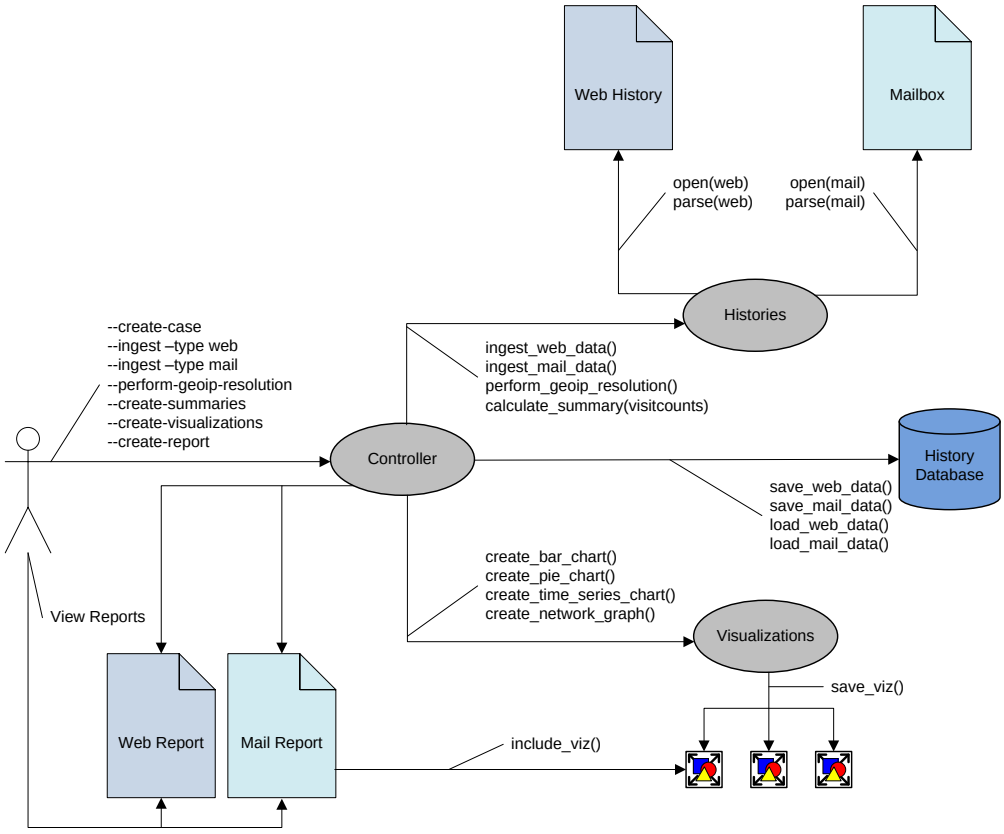



Figure 3.1: Concept of Operations

The following use cases for the application have been outlined in summary format for the most visible and high-risk operations:

- UC1: Create New Case File
- UC2: Ingest Browser History File and Populate System Data Structure
- UC3: Ingest Mailbox File and Populate System Data Structure
- UC4: Calculate Summaries – Visit Counts by Time Period
- UC5: Derive Metadata – Site Country
- UC6: Create Visualization – Counts Over Time

## UC1: Create New Case File

**Preconditions**

Forensic analyst has access to existing location where he is able to create directories and files.

**Postconditions**

A new data storage directory in the structure to be used to store files imported or generated during use of the utility is created. The analyst's output directory is set to point to newly created data directory.

**Stakeholders**

Analyst wants the ability to easily create new case file with specified name in location of his choosing. Case manager or investigator may dictate case naming convention based on agency policy.

**Related**

**Basic Flow**

1. Execute utility with `--create-new-case <case name>` argument
2. System checks specified case name/number for existence of illegal characters
   *Repeat until a case name/number is entered containing ony legal characters*
3. System prompts for location to create new case directory offering current directory as default
4. Analyst enters fully-qualified directory path to create new case directory or `Enter` to accept proposed default
5. System checks that user has appropriate permissions to write to current location
   *Repeat until a location is entered which is writable by current user*
6. System creates new data storage directory structure in specified location with specified case name
7. If the user specified a location for creation of data storage directory, system outputs message instructing user to change to that directory

**UC2: Ingest Browser History File and Populate System Data Structure**

**Preconditions**

Analyst has created a new case and the web browser history file is in a location he has read access to. Analyst is executing the utility from within the appropriate case directory.

**Postconditions**

Contents of web browser history file are read completely into in-memory system data structure and the extracted contents are stored in internal database.

**Stakeholders**

Analyst wants ability to easily and completely read information from existing browser history file into in-memory data structure and store all extracted information and derived metadata in internal database for further analysis.

**Related**

UC1: CREATE NEW CASE FILE

**Basic Flow**

1. Execute utility with `--ingest --type web --file <history file>` argument
2. System checks for web browser history file in specified location and opens for read
3. System reads information from web browser history file into in-memory data structure one record at a time until EOF
4. System closes connection to file and indicates to user that the ingestion task has been completed
5. System checks for internal database in current data storage directory for persistent storage of extracted information and opens for write
6. System writes each in-memory record to database
7. System closes connection to database and indicates to user that the data preservation task has been completed with the number of records processed

**UC3: Ingest Mailbox File and Populate System Data Structure**

**Preconditions**

Analyst has created a new case and completed the ingest step. Analyst is executing the utility from within the appropriate case directory.

**Postconditions**

Contents of mailbox file are read completely into in-memory system data structure and the extracted contents are stored in internal database.

**Stakeholders**

Analyst wants ability to easily and completely read information from existing mailbox into in-memory data structure and store all extracted information and derived metadata in internal database for further analysis.

**Related**

UC1: CREATE NEW CASE FILE

**Basic Flow**

1. Execute utility with `--ingest --type mail --file <mailbox file>` argument
2. System checks for mailbox file in specified location and opens for read
3. System opens each message in mailbox and reads relevant fields from header portion until EOF: `sender name`, `sender address`, `recipient addresses`, `timestamp`, `subject`
4. System closes connection to file and indicates to user that the ingestion task has been completed
5. System checks for internal database in current data storage directory for persistent storage of extracted information and opens for write
6. System modifies list of recipients for each message to format suitable for storage in database
7. System writes all information to database
8. System closes connection to database and indicates to user that the data preservation task has been completed with the number of records processed

### UC4: Calculate Summaries – Visit Counts by Time Period

**Preconditions**

Analyst has created a new case and completed the ingest step. Analyst is executing the utility from within the appropriate case directory.

**Postconditions**

Counts of unique visits during each specified time interval within the date range specified by argument to `--date-begin` and `--date-end`.

**Stakeholders**

Analyst wants to know visit counts per specified time interval within a specific date range.

**Related**

UC1: CREATE NEW CASE FILE, UC2: INGEST BROWSER HISTORY FILE AND POPULATE SYSTEM DATA STRUCTURE, UC3: INGEST MAILBOX FILE AND POPULATE SYSTEM DATA STRUCTURE

**Basic Flow**

1. Execute utility with `--calculate-summaries --date-begin <beginning date> --date-end <ending date> --interval <interval number>` argument
2. System checks that the beginning and ending dates fall within date range for which history records exist, interval is one of `daily/hourly/monthly/yearly`, and requested interval will not result in records that fall outside of valid date range
3. System checks for internal database in current data storage directory, opens for read
4. System reads records grouped by specified interval, and for each record:
   (a) retrieves the full URL for that record, adds it to list of unique URLs visited during that interval, updates count

   *Repeat until all counts have been calculated for each time interval*
5. System closes connection to database and indicates to user that the calculations have been completed with a visit count for each time interval

**UC5: Derive Metadata – Site Country**

**Preconditions**

Analyst has created a new case and completed the ingest step. Analyst is executing the utility from within the appropriate case directory.

**Postconditions**

The country where each visited site is hosted is looked up, the navigation record for that site is updated and stored in internal database.

**Stakeholders**

Analyst wants to know the country where each visited site is hosted.

**Related**

UC1: CREATE NEW CASE FILE, UC2: INGEST BROWSER HISTORY FILE AND POPULATE SYSTEM DATA STRUCTURE, UC3: INGEST MAILBOX FILE AND POPULATE SYSTEM DATA STRUCTURE

**Basic Flow**

1. Execute utility with `--perform-geoip-resolution` argument
2. System checks for internal database in current data storage directory, opens for read
3. System checks for geolocation database and opens for read
4. System reads one record at a time and executes a lookup for the base URL in the geolocation database
5. The country in which the given website is hosted is recorded in the `country` field
   *Repeat until all records have been read and countries resolved*
6. System updates internal database with new information
7. System closes connection to both databases and indicates to user that the operation has been completed with a count of successful and unsuccessful country lookups

**UC6: Create Visualization – Counts Over Time**

**Preconditions**

Analyst has created a new case and completed the ingest and summary calculations. Analyst is executing the utility from within the appropriate case directory.

**Postconditions**

Visit counts per time interval within specified date range are plotted on a time-series chart.

**Stakeholders**

Analyst wants to be able to see a user's historic browsing volume by specified time interval within a specific date range.

**Related**

UC1: CREATE NEW CASE FILE, UC2: INGEST BROWSER HISTORY FILE AND POPULATE SYSTEM DATA STRUCTURE, UC3: INGEST MAILBOX FILE AND POPULATE SYSTEM DATA STRUCTURE, UC4: CALCULATE SUMMARIES – VISIT COUNTS BY TIME PERIOD

**Basic Flow**

1. Execute utility with `--create-visualizations --date-begin <beginning date> --date-end <ending date> --interval <interval number>` argument
2. System checks that the beginning and ending dates fall within date range for which history records exist, interval is one of `daily/hourly/monthly/yearly`, and requested interval will not result in records that fall outside of valid date range
3. System checks the number of resulting time intervals within specified date range and adjusts the type of plot to be created
4. System plots each interval in a time-series chart sorted chronologically on the x-axis and by increasing count on the y-axis
5. System creates image file containing rendered chart to appropriate location in data storage directory
6. System closes connection to database and indicates to user calculations have been completed with a count for each time interval

## 3.5 Supplementary Specification

### 3.5.1 Introduction

This section outlines system requirements not captured explicitly in the requirements summary or by a use case. General categories of requirements covered are reports, documentation and packaging.

### 3.5.2 Logging and Error Handling

Depending on the level of verbosity selected by the user, all system operations (optional) and errors that occur during operation will be logged to the console and may easily be captured in a file by the user.

### 3.5.3 Configuration

Since the application is largely a command-line, batch-oriented utility, it is not worthwhile to introduce additional complexity in order to provide the ability to read and write an externally maintained configuration file.

### 3.5.4 Human Factors

The human factors elements that are of highest importance depend on the interface with which the user is currently working, however the elements that are common to both the command-line application and the generated report include:

- font size - should be no smaller than 10pt
- font color - should be of a color and contrast level that is not fatiguing to the eyes
- font typeface - either Monospace or a Sans-Serif typeface
- font encoding - all fonts will be Unicode
- error and other interactive notifications must be presented in the foreground or inline to any existing textual output and must be easily distinguished from other output

### 3.5.5 Recoverability

All initial input data is read from a file opened in read-only mode, therefore any failure of the system on which this utility is installed will only affect data that may be reconstructed by re-executing the interrupted function.

All processed data is saved to an internal database in order to allow the user to complete the analysis over a number of different sessions, and is opened and updated during normal processing. Any failure of the system on which this utility is installed will affect all data not already captured in this store, and therefore will require those operations whose output was not recorded to be repeated. It is beyond the scope of this system to provide a transactional capability to safeguard against this minimal risk.

### 3.5.6 Performance

There are two primary types of system resource-intensive operations performed by this utility – disk I/O and in-memory processing. The disk I/O is limited to the reading of the initial web browser history file from disk, and all subsequent reads and writes of the processed data to the internally maintained database. The amount of data to be read and written is not foreseen to be of large enough volume to negatively affect the I/O performance of the system on which this utility is installed. In-memory processing is foreseen to be the highest resource consumer and potential tax on the performance of the host system. For this reason, it is recommended that this utility be given priority during all processing operations and that multitasking be kept to a minimum.

### 3.5.7 Adaptability

The utility can not dictate a hardware baseline, but should be capable of providing acceptable performance on any relatively modern (post-2004) hardware.

### 3.5.8 Implementation Constraints

The implementation will be object-oriented in nature, using the Python programming language. Any third-party packages or plugins that are used to extend the base functionality of this utility must conform to this requirement.

### 3.5.9 Purchased Components

There are currently no plans to purchase any components for this utility.

### 3.5.10 Free Open Source Components

The utility will be built on a foundation of best-of-breed open source components including but not limited to the following:

- Python programming language and native package library
- Matplotlib graphing and modeling library
- GraphViz graphing and modeling library
- MaxMind GeoIP IP geolocation database
- Dmoz Open Directory Project categogory and content RDF dump
- Various 3rd party packages with a Python API

### 3.5.11 Hardware Interfaces

- Directly connected or network-accessible printer for preserving hardcopy version of rendered reports
- CD/DVD writer or USB drive in order to backup or save rendered reports
- Network interface with Internet connectivity may be necessary for full functionality of certain operations

### 3.5.12 Software Interfaces

Any architecture- or operating system-specific functions will be avoided in the implementation of this utility in order to maximize its potential for installation across multiple platforms.

### 3.5.13 Application-Specific Domain Rules

All policies and business rules are dictated by the specific agency making use of this utility.

### 3.5.14 Legal Issues

The analyst has prior knowledge of the identity of the user whose history he is examining, and there is no additional data stored by this utility than already exists in the original history. The information that is processed by this utility is potentially of sensitive or classified nature, therefore the proper safeguards and privacy procedures must have been arranged in advance by the user. The data that is processed by this utility will need to be given the same protection, attention and classification as the source data, and no effort is made by this utility to safeguard its confidentiality. The utility does not transmit any information to the Internet.

## 3.6 Domain Model

The high-level conceptual model of the toolset architecture is captured by the domain model in Figure 3.2 which attempts to illustrate the class hierarchies and associations. The problem space is decomposed into individual comprehensible units in order to effectively capture the essential abstractions and convey all the information that is required to understand the domain.

## 3.7 Data Structures

The data structures that were composed for storing a single record of historical information are illustrated and annotated in order to give the user a deeper understanding of the internals of the toolset. Each record will be stored in an list during in-memory processing and as a single record in a relational database on disk for long-term storage.

### 3.7.1 Browser History Record

Each web navigational record that is read from the browser history file on disk must be stored in core in an efficiently indexed and easily iterated form. The Python *dictionary* data type, known in other languages as *associative array*, *hash* or *HashMap*, was chosen for this purpose since it

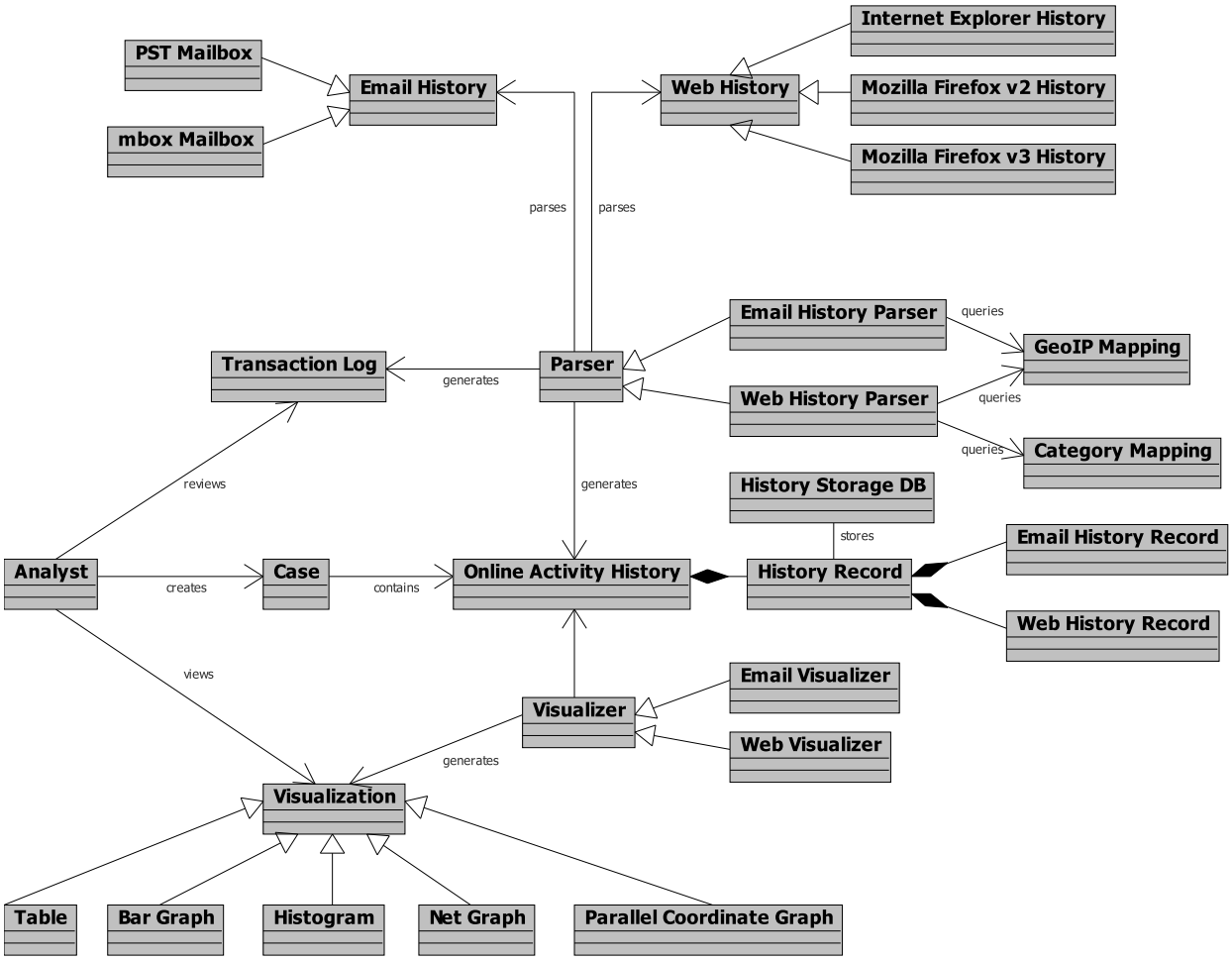



Figure 3.2: Domain Model

allows fast indexing of its contents by an arbitrary developer-assigned immutable key and allows for representing sparse data structures. The potential for sparsity in the resulting extracted data exists due to the choice to keep the identification indexes that appear in the source data, and the possibility that certain records may not be parsed due to corruption or missing fields. Each element of the history record that is read from the history file on disk is stored in its original form in order to maintain data integrity, but in certain cases the element may need to be manipulated in order to present it in a more user-friendly format. In this case the modified version is stored in an additional field which is named based on the modification that took place. As an example, the `url` field contains the full URL of the visited site as read from the history file on disk which may be long and unwieldy when attempting to display it in a visualization or a table, so the base URL — the portion of the URL beginning with the protocol identifier (*e.g.,* www.) and ending

with the top level domain (*e.g.,* .com) — is extracted and stored in a field named `url_base`. The storage of each derived element offers the advantage of only incurring the processing cost a single time, thereby reducing processing time at the expense of a small amount of additional storage space. Table 3.4 shows the format of the data structure used to contain a web history record with an explanation of each field.

### 3.7.2 Mail History Record

Similar to the web browsing record, the attributes of each mail message header that is read from the mailbox file are stored in a dictionary in their original form. All derived formats of the original contents are stored in an additional field. Table 3.5 outlines the format of the data structure with an explanation of each field.

### 3.7.3 Summary

This chapter outlined all the considerations taken into account prior to the design of this utility. This included identifying the stakeholders and their goals, the high-level goals to meet the stakeholder needs and the resulting features that this utility will provide. Reasoning was provided for the choice of data to be extracted from the histories and six selected high-risk use cases were outlined in summary format as a means of sketching the operation of this utility and identifying specific requirements prior to implementation. A concept of operations was provided to show the individual operation of each component of this utility, as well as the communication between all the different components and the analyst. Finally, a supplementary specification was provided to capture any requirements not expressly specified in the previous sections. The high-level conceptual model of the toolset architecture was illustrated in the domain model diagram and the data structures used for storing a single history record were outlined and annotated to provide further insight into the operation of this utility.

The next chapter will formalize and describe in detail the visualization process from raw data to the resulting visual structure. It will also address the types of data encountered in web browser and mail histories, along with the choice of visualizations with which to represent that data graphically.

```
mozhistrec = [
   # ID for this record taken from history file
   ('id','URL ID','INTEGER PRIMARY KEY'),
   # Session number taken from history file
   ('session','Session','INTEGER'),
   # Full URL of visited site
   ('url','URL','LONGVARCHAR'),
   # Base URL: access scheme prefix, domain and TLD suffix
   ('url_base','Base URL','LONGVARCHAR'),
   # URL access scheme (protocol) specifier
   ('scheme','Protocol','VARCHAR'),
   # TLD of domain
   ('tld','Top Level Domain','VARCHAR'),
   # ID of referrer
   ('ref_url_id','Referring URL ID','INTEGER'),
   # Previous site in navigation path
   ('ref_url','Referring URL','LONGVARCHAR'),
   # Base URL of referrer
   ('ref_url_base','Referring URL Base URL','LONGVARCHAR'),
   # Title of web page from <title></title>
   ('title','Page Title','LONGVARCHAR'),
   # Date of site visit in seconds since the epoch
   ('visit_date','Visit Date','INTEGER'),
   # Date of site visit formatted for viewing
   ('visit_date_fmt','Visit Date Formatted','VARCHAR'),
   # Duration of site visit in seconds
   ('visit_duration','Visit Duration','INTEGER NOT NULL DEFAULT 0'),
   # Total number of times site visited
   ('visit_count','Visit Count','INTEGER DEFAULT 0'),
   # Index into site visit type table
   ('visit_type','Visit Type','INTEGER NOT NULL DEFAULT 0'),
   # URL included in object embedded in current page (0/1)
   ('embedded','Embedded','INTEGER NOT NULL DEFAULT 0'),
   # URL was typed into navigation bar (0/1)
   ('typed','Typed','INTEGER NOT NULL DEFAULT 0'),
   # Mozilla frequency/recency metric
   ('frecency','Frecency','INTEGER NOT NULL DEFAULT 0'),
   # URL for site favicon
   ('favicon_url','Favicon URL','LONGVARCHAR'),
   # Country code
   ('country_code','Country Code','VARCHAR'),
   # Country name
   ('country_name','Country Name','VARCHAR'),
   # Category of base URL
   ('category','Category','VARCHAR NOT NULL DEFAULT ""'),
   # If this was a search engine query
   ('search_query','Search Query','VARCHAR')
]
```

Table 3.4: Web browser history record data structure

```
mboxhistrec = [
   # Unique ID for this message
   ('id','ID', 'INTEGER PRIMARY KEY'),
   # Name of sender
   ('source_name','Source Name', 'LONGVARCHAR'),
   # Address of sender
   ('source_addr','Source Address', 'LONGVARCHAR'),
   # Address(es) of recipient(s) (to: cc: bcc:)
   # List of (name, address) tuples
   ('recipient_addrs','Recipient Addresses','BLOB'),
   # Date and time message was sent
   ('datetime','Date & Time','REAL'),
   # Date and time message was sent (formatted)
   ('datetime_fmt','Date & Time','VARCHAR'),
   # Subject of message
   ('subject','Subject','LONGVARCHAR')
]
```

Table 3.5: Mail message record data structure

THIS PAGE INTENTIONALLY LEFT BLANK

# CHAPTER 4:
# Visualization of Historical Web and Mail Activity

## 4.1 The Visualization Process

At a fundamental level, visualization is the mapping of some dataset into a *Visual Structure* that somehow represents those original contents. A formal treatment of the process by which *Raw Data* is transformed into a visual representation and the quality of the resulting *Visual Structure* is given using the reference model for visualization proposed by Card et al.[30]. In this chapter, each step in the process of transforming the *Raw Data* into a *Visual Structure* is described and illustrated with examples from web browser and mail histories.

### 4.1.1 From Raw Data to Data Tables

*Raw Data* is data maintained or represented in some "idiosyncratic format," [30] and the first step in preparing data for use in visualizations is its extraction from the underlying dataset. This idiosyncratic format is defined by the application which is used to collect, manipulate and store the data, in this case a web browser and MUA, and may be structured in any of the formats described in Sections 2.1 and 2.2. For the purposes of this thesis, the web browsing history is stored in the format dictated by the *mozStorage* API [16, 17] and mail is stored in the *mbox* mailbox format[20]. These formats provide a particular structure to the data, and, with an understanding of that structure, access and extraction of the desired elements as determined in Section 3.3 becomes possible.

The extraction of the desired information from Firefox v3 history files described in Section 3.7.1 is performed according to the process outlined in use case UC2: Ingest Browser History File and Populate System Data Structure in Section 3.4 with use of the Python `sqlite3` SQL interface for SQLite databases[81]. This interface, which is compliant with the DB-API 2.0 specification as defined in the Python Database API Specification v2.0 [82], has been included in the Python Standard Library since version 2.5 [81] and allows connection to, and manipulation of, SQLite databases using native SQL commands.

The extraction of desired information from mailboxes in the *mbox* format is performed according to the process outlined in use case UC3: Ingest Mailbox File and Populate System Data Structure in Section 3.4 using the Python `mailbox` module[21]. The `mailbox` module is

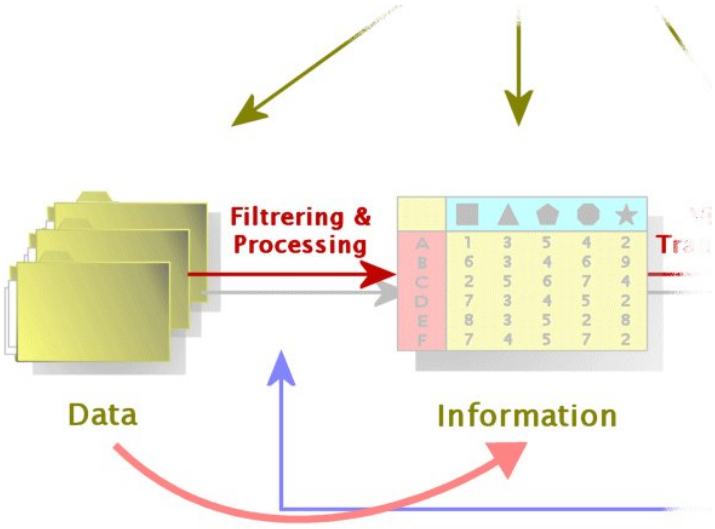



Figure 4.1: Transforming Raw Data to Data Tables (partial view) [83]

also part of the Python Standard Library and allows for "accessing and manipulating on-disk mailboxes and the messages they contain."[21] The desired information is extracted from the header portion of each message and is described in detail in Section 3.7.2.

Once the desired information has been extracted from the history files into a raw form, the next step in the visualization reference model is the transformation of this *Raw Data* into a relation or set of relations that are more structured and therefore more easily mapped to a visual structure. This is done by performing transformations on the *Raw Data* to obtain *Data Tables*, "relational descriptions of data extended to include metadata" [30] as shown in Figure 4.1. *Metadata* is used to describe the relations between different data elements, and in Table 4.1, which shows the structure of a data table, it takes the form of labels on the rows and columns. In the visualization reference model, the rows represent *variables* "sets that represent the range of values" and the columns represent *cases*, "sets of values for each of the variables."[30].

| **Case** | $Case_i$ | $Case_j$ | $Case_k$ | ... |
|---|---|---|---|---|
| $Variable_x$ | $Value_{ix}$ | $Value_{jx}$ | $Value_{kx}$ | ... |
| $Variable_y$ | $Value_{iy}$ | $Value_{jy}$ | $Value_{ky}$ | ... |
| ••• | ... | ... | ... | ... |

Table 4.1: Structure of a Data Table [30]

*Data Tables* clearly show the attributes of the collected data, and by the application of variables allow selection of those attributes that will be used to create the mapping between the data and

the resulting visual structure. Table 4.2 shows the data that is extracted from a browser history file in the format of a data table with a single case titled *ID* and multiple variables describing each value. The single case *ID* is composed of unique values which are necessary in order to disambiguate each record and provide an index for mapping each record to the visual space.

| **ID** | 300 | 450 | 535 | ... |
|---|---|---|---|---|
| **Session** | 3 | 4 | 5 | ... |
| **URL** | www.nbc.com/shows | www.nps.edu | www.cnn.com | ... |
| **Referrer** | www.google.com | intranet.nps.edu | www.ask.com | ... |
| **Title** | NBC::Shows | NPS | CNN::Home | ... |
| **Visit Date** | Apr 1 15:33 2009 | Apr 1 15:42 2009 | Apr 1 15:54 2009 | ... |
| **Visit Count** | 4 | 1 | 9 | ... |
| **Visit Type** | 2 | 1 | 3 | ... |
| **Embedded** | 0 | 0 | 0 | ... |
| **Typed** | 1 | 1 | 0 | ... |
| **Frecency** | 5000 | 10 | 1000 | ... |

Table 4.2: Data Table containing navigation records from a browser history

*Data Tables* may also be used to represent network data with the creation of a variable describing the edge or link metric that captures the relationship between connected cases, which represent nodes in the graph. In the example in Table 4.3 the *MsgExch* variable establishes the relationship between the *Users* cases, and the *Count* variable provides the metric that quantifies that relationship.

| **Users** | bob@edu | alice@com | carol@org | ... | mallory@ru |
|---|---|---|---|---|---|
| **Count** | 14 | 37 | 22 | ... | 1 |
| **MsgExch** | {alice@com} | {bob@edu,carol@org} | {alice@com} | ... | {carol@org} |
| **...** | ... | ... | ... | ... | ... |

Table 4.3: Data Table illustrating relationship between users exchanging mail messages

This section described the transformation of *Raw Data* into *Data Tables* which are given structure and meaning using *metadata*. The last step in the reference model — the transformation from *Data Table* to *Visual Structure* — will be described next, following an explanation of the different data types and the classification of the information extracted from web browsing and mail histories.

### 4.1.2 Categorical, Ordinal, Interval, Hierarchical and Graph Data Types

There are five different data types encountered in web browsing and email histories that must be described by variables: categorical, ordinal, interval, hierarchical and graph. Following are definitions for each variable type and an example from the web or mail history. Table 4.4 summarizes the types of each variable used to describe values extracted from the web browser and mail histories.

*Categorical*, also known as discrete or nominal data, are qualitatively assigned, lack any intrinsic order and are based on two or more names or categories[84, 85]. An example of a categorical variable used to describe a field from the browsing history is the visit type which may take one of seven possible values and indicates the nature of the navigational action taken by the user that resulted in that navigational event. The country in which the site or domain of the mail address is hosted also constitutes a categorical dimension and lacks any innate order.

*Ordinal* data are also categorical and qualitatively assigned, and describe values that possess an inherent order but lack any measurable interval between them[86]. An ordinal variable is used to describe the field containing the Mozilla *Frecency* metric [App A] — the combined recency and frequency of a particular site visit, outlined in more detail in Section 4.2.3. The count of unique URLs and messages is also an ordinal value with may be sorted in descending order from highest to lowest count in order to identify frequently occurring entries.

*Interval* and *ratio* data are continuous and may be plotted on a continuum or scale. Interval values are quantitative in nature — they refer to the quantity of what is measured [84] — and there is a measurable interval between them[86]. Interval data has no breaks or gaps and lacks a true zero point[85]. Ratio data may be expressed as real numbers and have a true zero point[85]. The index used as a unique ID to represent each history record in the database is an interval value beginning at one and increasing sequentially. The date and time of a site visit or message receipt is an example of an interval value which is measured internally as a floating point number representing seconds since the epoch, midnight (00:00:00) 1 January 1970 Coordinated Universal Time (UTC).[App A] A time after the epoch has a positive value, and a time before the epoch has a negative value, however in the context of this work, negative values for timestamps will be viewed as erroneous or spurious since the Internet was not in existence in 1970.

*Hierarchical* data represents a ranking of items such that each item is subordinate to only one other item[86]. There are examples of both individual variables describing hierarchical values, as well as hierarchical arrangements composed of tuples of variables. An example of a hierarchical variable is the site URL which forms a tree rooted at the access scheme, followed by any number of optional hostnames, a mandatory domain name and a mandatory TLD followed by any number of optional path components. An example of a hierarchical arrangement across the data results when the *url* and *visit_date* fields of each record are combined to form the *(url,visit_date)* tuple, each of which is unique and subordinate to only one other such tuple.

*Graph* data are composed of nodes (vertices) connected by any number of links (edges), which may be directed with the link specifying the direction of the connection, or undirected, in which case each link is bidirectional. An example of graph data in mail histories is the directed link that may be drawn between the message sender and the recipient representing the communication that occurred between them in a single message. When more than one message is taken into account for the same sender and recipient, that link may become undirected if there are one or more replies sent to the original message.

| **Categorical** | **Ordinal** | **Interval** | **Ratio** | **Hierarchical** | **Graph** |
|---|---|---|---|---|---|
| URL Access Scheme | Frecency | Site/Message ID | Site Visit Timestamp | Site URL | Sender/Recipient Pair |
| Visit Type | Site Visit Count | Session Number | Message Timestamp | Email Address | |
| Site Country | Message Count | | | | |
| Site Category | | | | | |
| Domain Country | | | | | |

Table 4.4: Web browser history elements data types

### 4.1.3 From Data Tables to Visual Structures

Visualization in a security or forensic context is the process of generating a picture based on the underlying data and defines how that data is mapped into a visual representation[85]. Section 4.1.1 described the mapping of *Raw Data* to *Data Table*, and this section will address the final transformation from *Data Table* to *Visual Structure* as shown in Figure 4.2.

In the final step from *Raw Data* to visualization, the *Data Tables* are mapped to *Visual Structures* which "augment a spatial substrate with marks and graphical properties to encode information."[30] A good *Visual Structure* is created from the *Data Table* in such a manner that the data is preserved, and is *expressive* if all and only the existing data is also represented in the *Visual Structure*[87]. Due to limits in the user's perceptual system, representational limits of the graphical medium and a limited number of components used to construct the *Visual Structure*,

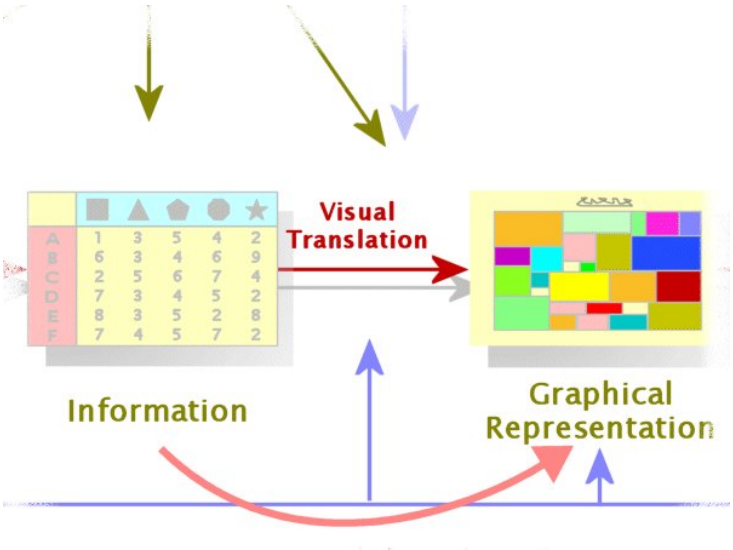



Figure 4.2: Transforming Data Tables to Visual Structures (partial view) [83]

there are a corresponding limited number of mappings of *Data Tables* to *Visual Structures*. The basic set of components used to construct the *Visual Structure* are the *spatial substrate*, *marks*, and the graphical properties of the *marks*, and although there are additional properties, these are the core set used in most visualizations[88, 87].

Space is the dominantly perceived property of any visualization, and its use often determines the effectiveness of the visualization. The choice of which and how many of the variables from the *Data Table* will be mapped onto the *spatial substrate* is a series of choices of which variables will be represented at the expense of others, since many variables will overlap and result in a highly occluded visualization which fails in representing the data. The *spatial substrate* is bounded by *axes* of which there are four elementary types: *Unstructured*, *Nominal*, *Ordinal* and *Quantitative* [30], each of which maps to the type of data that is being visualized as outlined in Section 4.1.2.

*Marks* are the visible attributes that are placed on the *spatial substrate* and there are four primary types: *Points*, *Lines*, *Areas* and *Volumes*[30]. The placement of *marks* on the *spatial substrate*, and the relationships between them, is another area where the designer of the visualization must make a decision on which variables to represent and which to suppress. Relationships between *points* may be highlighted by connecting them with *lines*, and space may be filled with color in order to convey *areas* or *volumes*. *Marks* may also be used to express hierarchical or graph relations when the data is mapped to the *Visual Structure* in such a way that the relative spatial

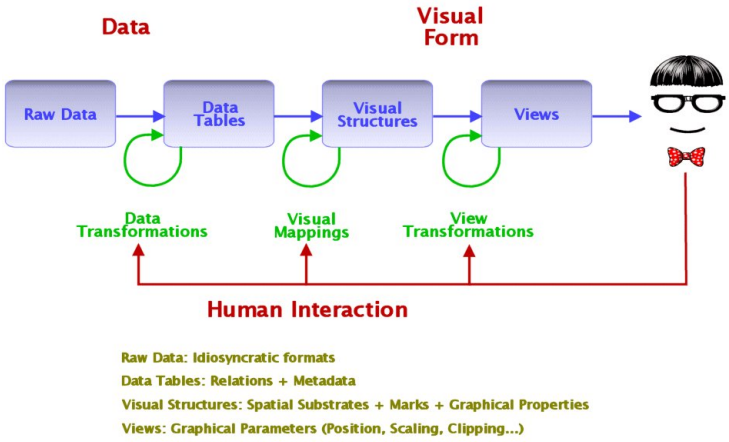



Figure 4.3: Card Reference Model for Visualization [89]

positioning of individual elements serves to convey information. For example, the size of a *point* may serve to encode the visit frequency to a particular site or the number of recipients to a particular message. Alternatively, properties such as proximity and clustering may be used when placing and connecting *marks* in space to encode information, for example encoding the amount of time elapsed between navigational events in the length of the line used to connect them.

The visualization process — the series of steps in which *Raw Data* undergoes various transformations to finally be rendered as a *Visual Structure* — has been outlined in detail in Sections 4.1.1 and 4.1.3 and may be seen in its entirety in Figure 4.3.

## 4.2 Discussion of Visualizations

The final step in the visualization process is the generation of the *Visual Structure* that will be viewed and manipulated with the desired result of the conversion of that *Visual Structure* into understanding and knowledge by the consumer of the visualization through revealing underlying patterns in the data. This transformation of the perceptual representation into understanding by the user as seen in Figure 4.4 is dependent on many factors, many of which are out of the control of the designer of the visualization. The amount of knowledge gained from the visualization is also dependent on the number and diversity of the users that will be viewing them and the context in which they are viewed as seen in Figure 4.5.

In his landmark book “The Visual Display of Quantitative Information,” Edward Tufte states that “good graphic design reveals the greatest number of ideas in the shortest time with the least

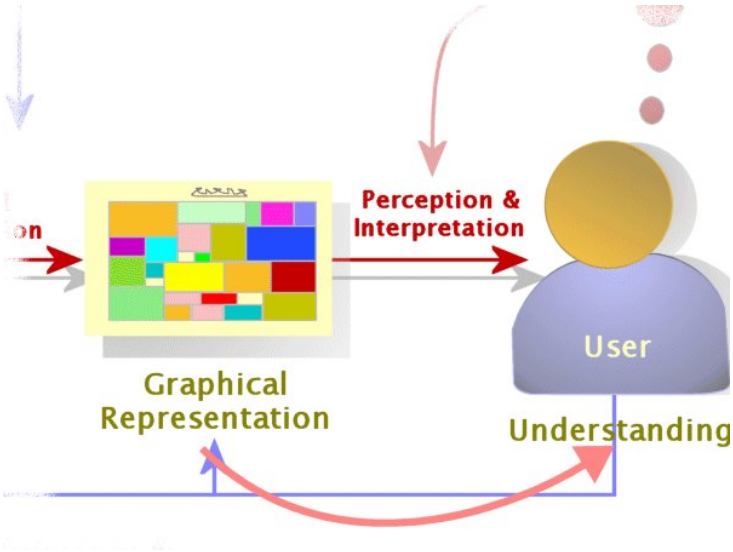



Figure 4.4: Understanding the Visual Structure (partial view) [83]

ink in the smallest space."[90] A very concise and easy to understand concept, it is one of the most-cited rules of visual design and was applied during the design of the visualizations used to represent the web browsing and mail histories. The primary challenge of visualization is in choosing how to convert the *Raw Data* into a graphical format that provides insight into that data. In order to satisfy all the requirements previously discussed and simultaneously appeal to the entire set of stakeholders identified in Section 3.1.1, the choice of visualizations for the summary report was limited to those familiar to most people due to their common use in mainstream, financial and scientific publications. The data table, bar chart, pie chart and time-series plot were chosen as the primary means of presenting a summary view of the histories and are discussed in detail in the following section along with the more detailed timeline and network graphs which do not appear in the summary report. The timeline and network graph visualizations are more specialized in nature and therefore potentially less comprehensible to the entire range of stakeholders, but will prove valuable to the analyst when attempting to gain knowledge of the details of a particular series of events or period.

### 4.2.1 Data Tables

*Data Tables*, introduced and defined in Section 4.1.1 as "relational descriptions of data extended to include metadata," [30] are utilized in the report to highlight and provide compact summaries of various elements of the browsing and mail history. The web browsing history summary report presents summary tables of total unique counts, top visited base and full URLs, visit

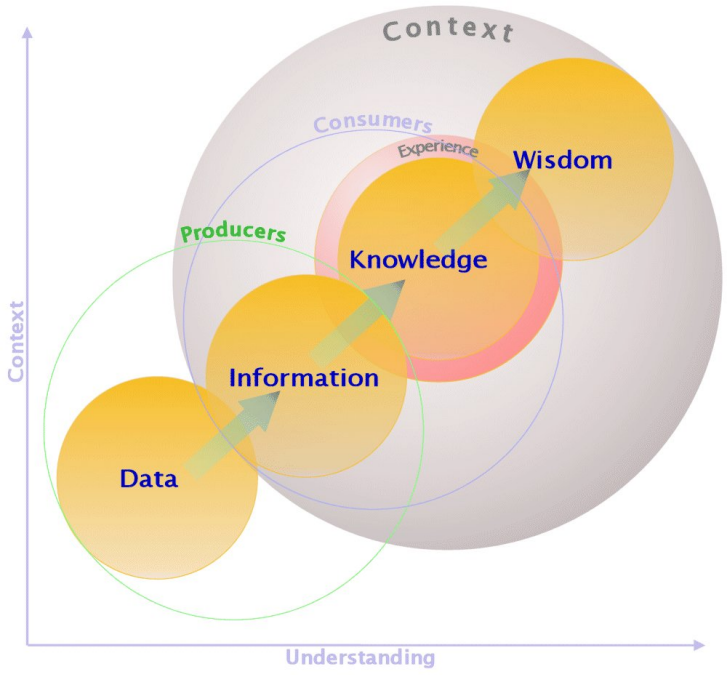



Figure 4.5: Understanding and context in the conversion of data to wisdom [89]

types, access protocol, top countries and TLDs, top search queries and top visited categories. The mail summary report presents tables for total unique counts, top senders recipients and domains, and top TLDs and countries. Each table is displayed alongside the visualization most appropriate to represent it, combining the detail provided by the table with all the previously mentioned advantages of visualization. The data that is presented in the tables is discussed in more detail in the remainder of this section followed by a discussion of the visualizations.

The tables shown in Figure 4.6 appear at the top of the report and provide an aggregate summary of the data that will be presented in more detail below, allowing the analyst to see both the volume and type of information that will be elaborated on. The tabulated counts are unique counts with all duplicates removed, so for example, if the user visits a particular website daily during the period reflected in the report, the URL is only counted once for the purposes of this table. Similarly, mail addresses are only counted once, regardless of whether they occur in the role of sender or recipient in any given message. This presentation choice reduces the weight of frequently visited sites and oft-used addresses, allowing the analyst to form a balanced overview of the user's browsing and mail behavior.

| Total Unique Counts | | | |
|---|---|---|---|
| Type | Total | Visited | Not Visited |
| Full URLs | 15,642 | 15,555 | 87 |
| Base URLs | 2,563 | 2,534 | 29 |
| Top Level Domains | 58 | 53 | 5 |
| Countries | 54 | 45 | 9 |
| Embedded URLs | 127 | 17 | 110 |
| Downloads | 2,875 | - | - |
| Typed URLs | 103 | - | - |
| Categories | 410 | - | - |
| Search Queries | 505 | - | - |

(a) Web browsing summary table

| Total Unique Counts | |
|---|---|
| Type | Total |
| Messages | 3,167 |
| Senders | 227 |
| Recipients | 1,633 |
| Aliases | 1,685 |
| Domains | 99 |
| Top Level Domains | 8 |

(b) Mail summary table

Figure 4.6: Web browsing and mail summary tables

The leading table in the web browsing summary shows total unique counts of full and base URLs with counts of visited and not visited for each category. Full URLs include the entire address of the visited website, while base URLs are the full URL with the path component removed. Using the base URL is a method of further condensing the user's browsing habits through the aggregation of all intra-site traffic, since at the overview level, the analyst may be more interested in the user's overall browsing patterns than the specific pages visited within each site. This intra-site aggregation can be seen when comparing rows one and two in Figure 4.6a: there are 15,642 unique visits when counting full URLs, but of those visits only 2,563 or

approximately 16 percent were unique sites when counting the base URL. This leaves approximately 84 percent of the browsing during this period as intra-site in nature.

Displaying both the sites that were visited and not will provide additional insight into the nature of the user's browsing behavior due to the way site visits are classified by the browser. The Firefox browser may classify a URL as not visited for a number of reasons. When the user requests a page and the browser renders it, every link within that page which refers to an embedded or external object is also loaded and recorded in the browsing history. Embedded or external objects may be images, videos, music, compressed archives, executable binaries or any other object that can be linked to in a web page. Unless the user explicitly follows the link to one of these embedded or external references, it will be classified in the history as not visited, so all embedded references in a page are recorded in the history regardless of whether or not they were visited. The total count of embedded links is displayed in row five in Figure 4.6a with separate counts for visited and not visited.

The Top Level Domain (TLD) of each visited site and mail address is extracted, and the count of unique occurrences is shown in row three of the web table and row six of the mail table. This helps supplement the profile of the user by highlighting their high-level browsing and mail patterns and potentially their language preference. To further supplement this information, the country where each visited site is physically hosted is resolved through a lookup in the MaxMind GeoIP IP geolocation database as outlined in Section 3.3. Resolving the actual geographic location of the site using a geolocation database attempts to associate each site with its geographic location independent of its TLD. The top-occurring TLDs and countries are shown in Figure 4.16 and are further discussed in Section 4.2.4.

Additional attributes that help shape the user's online behavior profile include the number of file downloads they complete, the method by which they initiate each navigation event, the types of sites they visit and the search queries they execute. Row six in Figure 4.6a shows the total count of downloaded objects for the period in question, and when viewed relative to the proportion of traditional informational-oriented browsing, may provide additional insight into the nature of the user's online activity. Row seven shows the count for navigation events initiated by the user explicitly typing the URL into the address bar of the browser, potentially indicating a higher degree of familiarity with that site than one which was simply followed from a link appearing in a page. The proportion of sites of type `Download` and `Typed` relative to all other visit types can be seen in Figure 4.16 which is discussed in Section 4.2.4.

The count of unique categories for visited sites is shown in row eight for those sites having entries in the Dmoz ODP web directory [App A], outlined in Section 3.3. Those sites in the user's browsing history which are successfully looked up in the directory and paired to a category will further assist the analyst in composing a complete profile of that user by providing context through a more abstract view of their browsing than can be attained from the visited sites alone. The top categories for those sites visited by the user can be seen in Figure 4.15a, and are further discussed in Section 4.2.3.

The number of unique search queries successfully extracted from the browsing history for this period is listed in row nine. Search engines using the `GET` method of form submission encode the form content directly in the URL upon submission, which is then recorded in the browsing history. This utility provides the ability to extract the search queries submitted to a number of different search engines including the largest — Google, Yahoo!, MSN, AOL, Ask — as well as some of their commonly used services such as Google Images and Yahoo! Answers. A user's search queries can provide a wealth of information about their interests and is provided as an additional point of information for the analyst to use in forming their profile. The display of the extracted search queries can be seen in Figure 4.15b and is discussed in Section 4.2.3.

### 4.2.2 Displaying Details

In order to provide more detail than can be displayed in the summary report, a number of features were implemented to overlay additional information on the existing presentation or link to another page which provides the information in a different format. Mouseovers — tooltips that appear when the mouse is hovered over the data — were used to overlay details on demand, and links to the relevant section of detail pages were provided for individual entries.

Mouseovers were employed to increase the amount of information presented in the summary report while consuming the least amount of screen space. For all fields containing URLs, mouseovers display additional details relevant to, or contained in, the URL that would not fit in the table row. In the case of the table showing top visited full URLs seen in Figure 4.7a, the mouseover displays a tooltip containing the page title, the date and time of first visit, and a breakdown of the URL into base, path, parameter and query components. This technique was also used for the tables displaying base URL, search queries and categories. Each visualization that includes a label referencing a URL is limited in space, and in order to avoid the occlusion that results when the full URL is used, each URL was reduced to its base component for

presentation as a label, and the full URL was encoded as a hyperlink within the image, allowing the contents of the full URL to be displayed in the status area of the browser by hovering the mouse over the label. Mail addresses where expanded with tooltips to contain the full name associated with each address as shown in Figure 4.7b.

In addition to the mouseovers that overlay information on demand, each field contained in the table links to the relative location of that entry in the navigation history details page. For example, each URL shown in the base and full URL visit count tables links to its entry in the detail page, which lists all extracted components of each navigation record as shown in Figure 4.8. The detailed navigation history includes each site in the navigation chain with all the information that has been extracted from the web browser history and all derived information including visit duration, country, category and search query. If enabled by the analyst, a thumbnail view of each site is generated dynamically using the Thumbshots service [80] and displayed alongside the site details, providing a graphical representation of the visited site as it would be rendered by a web browser.

### 4.2.3 Bar Charts

Bar charts are used primarily with categorical data to visualize the frequency of occurrence for each value of a particular dimension. They are an intuitive and widely-used visualization that facilitate a comparative view of the frequency of each categorical value, and were used in the summary report to compare various top counts. Counts of the top visited full and base URLs, top sites by *frecency* [App A], and top search queries and categories were displayed from the web browser history, and top senders, receivers and domains from the mail history.

Arranging the visit count of the top unique full and base URLs on a bar chart as shown in Figure 4.9, allows the analyst to see at a glance those sites which the user devoted the majority of his attention and browsing volume to. Presenting the unique base URL counts aggregates all the user's intra-site traffic into a single variable — the base URL — and gives an indication of the comparative browsing volume between frequently accessed sites. The base URL is plotted on the y-axis and the percentage of each base URL as a proportion of the total is plotted on the x-axis. The bar chart is oriented horizontally in order to display the URL on the chart itself without consuming additional space with a ledger. There are a number of scenarios in which a view of the frequency counts per site or address may be beneficial to the analyst's understanding of the user's behavior that will be detailed below.

Figure 4.9a shows the bar chart for the user's top twenty visited base URLs, with the relative proportions arranged from highest to lowest, top to bottom. The analyst can quickly see that this user's number one, top five and top ten most frequently visited sites accounted for approximately eight, 27 and 40 percent of their total browsing history respectively. This list of sites will give the analyst a possible starting point on which to focus his attention, or alternately to selectively ignore them if they are of no interest based on previous knowledge about the user. The comparative proportions will highlight cases in which browsing is very heavily concentrated in a few sites, possibly indicating a very specific focus of interest on the part of the user, or an unknown cause such as some form of malware making connections without the user's knowledge. Furthermore, when paired with previous knowledge the analyst may have about the specific case, the information obtained from the top-visited sites can be very instructive. For example, if the browsing history in question was obtained from a work machine residing on a corporate or government network which has a policy of mandatory proxy usage for all employees, any URLs on the top-visited list that fall outside of permitted corporate intranet sites or the proxy itself, indicate a user who has been engaging in behavior that is in violation of the corporate security policy and may need to be investigated further.

Charting the top senders and recipients in the mail history offers the analyst insight into who the user is communicating with in a comparative fashion. It is to be expected that the address or addresses of the user himself will appear with the highest frequency, so the ability to supply a list of addresses to ignore was provided in the utility. The charts shown in Figure 4.10 include the address of the user, and as expected, it constitutes a far higher proportion than any of the other addresses. When generating the chart and ignoring this most frequently occurring address, the result is markedly different as shown in Figure 4.11. This set of charts offers a much clearer view of the proportion of the addresses with which the user is communicating.

The display of the top occurring domains from the mail history employs an aggregation technique similar to that used in the web history with the base URLs. In the same manner that aggregating on base URL serves to abstract intra-site navigation actions, extracting the domain from the address and charting it as shown in Figure 4.12a, highlights the type of organizations the user is communicating with. From this chart it is obvious that the domain of the user's address, and the one he communicates with most often, is `roussas.com` from its disproportionately large occurrence. This results in a highly unbalanced chart, so the same technique of specifying a domain to ignore is used once again, and results are shown in the chart in Figure 4.12b. This chart provides a clearer view of the organizations the user is communicating with and their proportions.

The display of the top-ranked set of visits to full and base URLs ranked by *frecency* is shown in Figure 4.13, and highlights those sites visited both frequently and recently based on the Mozilla *frecency* metric. According to the application of this algorithm, each site will be assigned a whole number weighting: -1 for any valid site visited once, 0 for invalid entries, or a number whose value increases based on revisitation count, visit type, how recent the visit was, and whether the site was bookmarked or tagged[91]. From a forensic perspective, an aggregate metric identifying sites with a high *frecency* weighting are of particular interest to the analyst since they indicate a visit that was repeated frequently, recently and may have been initiated from bookmarks or typed into the location bar, indicating that it was extremely familiar to the user. Since it is common for the 15 most recently visited pages to account for 88 percent of all revisits [75], with a single glance at this chart, the analyst is able to simultaneously see those sites visited most often, most recently and most likely to be the highest proportion of the user's total revisitation traffic.

Figure 4.14a shows the visit count over time using one-day intervals and may reveal patterns in the user's browsing behavior that are of interest to the analyst. The knowledge gained from this visualization depends on the usage context of the machine from which the browsing history was taken and background knowledge that may be available on the user. Returning to the example of the browser history from a machine installed on a corporate or government network, if the graph shows a pattern of browsing activity on Saturday or Sunday when it is uncommon or prohibited for employees to have access to work resources on weekends, this may be an indication of suspicious activity on the part of the user, or an indication that someone else has access to the user's machine. In order to investigate the specific browsing activity that was performed during that time interval in more detail, the analyst can turn to the timeline visualization which is discussed in Section 4.2.6.

The count of sent and received messages over one month using one-day intervals is shown in Figure 4.14b, and will highlight patterns in the user's communication behavior. Similar to the daily site visit counts, the usage context of the machine from which the mailbox was taken and any background knowledge that may be available on the user, will both serve to provide context. In this case, the drop in message volume that can be seen over most weekends and the spike that occurs on Mon 01/07 is not surprising with the knowledge that this mailbox belonged to a student whose winter break ended on that day. To investigate the specific communication activity in more detail, the timeline visualization pictured in Figure 4.19 can be consulted.

As described in Section 4.2.1, each visited site is mapped to a category based on a lookup in the ODP directory. The base URL for each visited site is looked up in the category database that was generated from the ODP RDF dump as described in Section 3.3, and the matching category is stored in both the navigation record and a local database containing domain-to-category mappings of all sites looked up by this utility. This serves to speed up the lookup process for subsequent sites since the local database is much smaller in size and contains only a single mapping for each site without any of the additional information that is provided by the ODP data.

The categories for the top-visited sites are shown in Figure 4.15a arranged in a horizontal bar chart from highest to lowest visit count. The choice of number of categories to plot is based on a configurable percentage threshold representing the proportion of each category to the whole. For example, the threshold that was set on the generation of Figure 4.15a was one percent, so only those categories for sites accounting for one percent or greater of the total browsing history were displayed. The distribution of categories is very instructive and provides a direct and understandable depiction of the user's interests, and in this example shows that over half of the user's browsing was accounted for and could be described by thirteen categories. Even categories which are different and plotted on separate bars can easily be distinguished by the human analyst and aggregated even more, further condensing the user's interests. For example, the third and sixth bar may be combined as "Search Engines," bars one, three and seven as "File Sharing," and bars eight and nine as "Online Auctions." The combination or aggregation of large numbers of site visits into a hierarchically arranged set of categories with the visual simplicity of a single bar chart allows the analyst to easily profile the browsing behavior of the user.

In order to provide more specific insight into the user's interests, the search queries that were extracted as described in Section 4.2.1 were plotted on a horizontal bar chart as seen in Figure 4.15b. There is not as much aggregation possible with search queries when compared to all the other attributes such as sites and categories due to the fact that most search queries are unique, and once the results are returned there is not a great deal of motivation to repeat the search. Similar to all the other bar charts, searches are ordered with highest occurring at the top and the least frequent at the bottom. The count is plotted on the x-axis. The same percentage threshold mechanism used with the categories was applied in this case to limit the number of searches plotted in the chart.

Search queries are useful in that they can provide context and motivation for the ensuing navigation events. For example, a navigation tree beginning with a search for ``osama bin laden'' will produce numerous results. The user may visit multiple results leading to sites with a wide, and not necessarily related, set of topics such as the Wikipedia page on Al Qaeda, the FBI ten most wanted list, a PBS documentary about the war on terror, a site dealing with conspiracy theories on the 9/11 attacks, and a blog espousing militant jihad. The user may have initiated the search with the intention of reading the latest news on the search for Osama Bin Laden and have no interest in the ideas espoused by the militant blog, but since he clicked on the link and the blog site appears in the browsing history, the analyst may want to understand the reason behind that. Since the site was navigated to from the results of a search, and was not explicitly initiated by the user, the analyst can take this into account to place the visit in context. He could verify this further by consulting the details for this visit and checking the visit type which will be `Link`, the referrer which will be `www.google.com/search` and the visit count which should be `1`. Without the knowledge that this visit originated from following a set of search results, the analyst would have no way of placing it in context and may form an incorrect impression of the user and waste valuable time focusing in the wrong direction.

### 4.2.4 Pie Charts

The pie chart is a ubiquitous visualization that has been widely used to compare categorical data as proportions or percentages of a whole. Anything that can be represented using a pie chart may also be represented as a bar chart, however the pie chart can be very useful and may be better suited to situations where it is desired to view each category as a proportion of the total number of categories, with the requirement that the number of different categories is small. The attributes of the browsing history that were chosen for visualization using pie charts are visit types, access schemes, visited countries and visited TLDs. The pie chart was chosen to visualize these attributes since each has a relatively small number of possible categories, allowing the presentation of all the categories together in a compact form with minimal occlusion.

The visit type field of each navigation event record indicates the navigation action or trigger that caused that event and may be one of `Link`, `Typed`, `Bookmark`, `Embedded`, `Permanent Redirect`, `Temporary Redirect` or `Download`, each of which is represented by an integer from one to seven. Summing the visit types for the entire history, or a selected period, and displaying them as relative proportions in a pie chart as shown in Figure 4.16a, allows the analyst to make a judgment as to the type of browsing behavior the user was engaged in based on

comparing the relative proportions. For example, a user with a high proportion of visits of type `Download` relative to others, is using the browser for more than informational or recreational browsing, and instead is using it as a data transfer utility. With only 17 percent of U.S. Internet users reporting going online at least weekly to download or watch videos and 30 percent to listen or download music [1], a high proportion of `Download` visits may indicate an atypical browsing behavior pattern, and therefore a user whose download history should be analyzed in more detail. Having observed a high proportion of `Download` site visits in the pie chart, the analyst could follow up by consulting the list of downloads for names of files which may be of interest and proceed to locate them in the user's files for further inspection. A large proportion of navigational events initiated from bookmarks also alerts the analyst to the high value of the user's bookmarks file in the future of the investigation since they show a pattern of heavy revisitation. The presence of `Embedded` and `Typed` visit types and their implications was discussed in Sections 4.2.1 and 4.2.3.

The display of total visit counts for access scheme as relative proportions in a pie chart, as seen in Figure 4.16b, helps the analyst understand the user's browsing behavior by providing an additional perspective on the manner in which the user uses the web browser. The access scheme indicates the protocol used to access each URL and may be any type supported by the Firefox browser, but most commonly is one of `http`, `https`, `ftp` or `file`. Similar to the example above regarding the `Download` visit type, a high proportion of File Transfer Protocol (FTP) accesses, indicated by `ftp` in the access scheme field, points to a user who is engaged in a high number of file transfers relative to web browsing. The Firefox browser supports the FTP protocol natively, and with many available plugins that make interacting with an FTP server from within the browser both easy and intuitive, it can readily become a replacement for any external FTP client. The analyst may once again consult the list of downloaded files or execute the utility with the `--get-records --type web --access-scheme ftp` command-line argument to obtain a list of all navigation records using FTP as an access scheme.

Presenting the relative proportions of countries and TLDs as shown in Figures 4.16c, 4.16d, 4.16e and 4.16f provides insight into the geographical distribution of the user's interests and communications as was discussed in Section 4.2.1. Viewing the two pie charts next to each other allows the information in each to reinforce the other, and makes it very easy to see how the TLD may correspond to a country. For example, the `ru` TLD accounted for 4.8 percent of the user's browsing and Russia accounted for 3.8 percent, so two proportions which were derived independently reinforce each other. A high proportion of visits or messages sent to

specific countries and TLDs might be normal for a user who speaks the native language of that country or has some cultural, entertainment or business interest there, but would be unexpected for a user with none of those. As another example, it is possible for a site to be registered under one TLD and hosted in a geographic area with no relation to that TLD. For example, it is not uncommon for many file-sharing and hosting sites to distribute their machines geographically based on such variables as cost and network proximity to their customers, so assuming the user is contacting a site hosted in a particular country based solely on the TLD may be erroneous. Geolocating each site offers added context to the TLD and would correctly classify a site such as the popular file-storage service by `depositfiles.com` which has a `.com` TLD, is registered in Cyprus and hosts its servers in a number of different countries including Russia and the Netherlands.

### 4.2.5 Time-Series Graphs

The display of daily visit counts using a bar chart discussed in Section 4.2.3 as shown in Figure 4.14a, is most suited for relatively brief periods, which can be selected at the time of report generation by the analyst. For longer periods, the time series graph as shown in Figure 4.17a was chosen for its ability to accommodate more information in the same amount of space and display trends over time. Figure 4.17a shows an example of a time-series graph for a three-month period with a daily interval, but the interval may be scaled up or down depending on the length of the period being displayed. The daily interval is instructive for longer periods measured in months, whereas a weekly or monthly interval would be more appropriate for periods measured in years. When the period is measured on the order of days, an hourly interval may be more useful in identifying intraday browsing trends, if the analyst wishes to view that level of granularity. A trend of the user's mail communication over a six month period is shown in Figure 4.17b.

As mentioned in the discussion of the bar graph showing counts over time, this visualization is used in order to provide a high-level view of the user's browsing or communication trends over time, but in order to view a detailed depiction of the specific activity during that time, the analyst can turn to the timeline visualization discussed in the next section.

### 4.2.6 Timeline

The timeline visualization is a snapshot of the user's browsing or communication history over time and was designed to depict all events for the selected period with a higher level of detail

than that used in the summary report. Figure 4.18 shows three sections of the timeline for a one-month period in the user's browsing history, with thirty-minute major intervals and five-minute minor intervals. Once again, as with the time-series graph, the intervals can be chosen based on the length of the period and the desired level of detail. Each individual navigation event is represented by a horizontal bar, displayed in order with time progressing left to right, and the length keyed to the visit duration for that particular site. Each bar is labeled on the left with the base URL for the visited site, and in the SVG version of the image, each bar is hyperlinked to the full URL. The bars oscillate from top to bottom within the channel as time progresses, and are alternately staggered on even and odd y-axis coordinates in order to avoid the occlusion that presents when a series of navigation events are made in rapid succession and therefore in close proximity on the graph.

A portion of the timeline for the user's mail history is shown in Figure 4.19, and there are some minor differences from the web browsing timeline in the representation. The width of each horizontal bar is keyed to the number of recipients in that particular message and the label shows the sender and recipient in the format `sender > recipient`. If there are multiple recipients, the labels are shown in the format `sender > first_recipient...last_recipient` in order to avoid the occlusion that results with large numbers of recipients.

This visualization combines multiple elements of the browsing and mail history into a single integrated display maximizing the amount of information conveyed. Each navigation event includes the site visited, the duration of the visit, the date and time of the visit and the previous and next visits. Each communication event includes the sender, recipient(s), and the date and time the message was sent or received. At this level of detail, the analyst can focus on the exact series of navigation events that were made by the user and reconstruct his exact behavior over any period and therefore validate the profile or clear up any misconceptions he may have formed when reviewing the summary report.

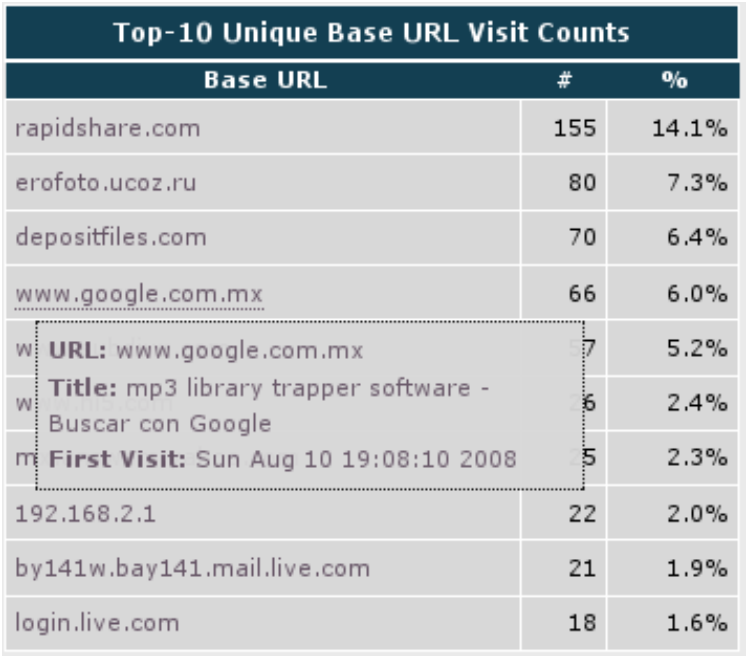
### 4.2.7 From Summary to Details

This collection of visualizations is presented in the summary report in a manner that allows for a review of the user's web browsing history from the general to the specific, and illustrates the synergistic nature of taking different views of the same data. The bar charts of top visited base and full URLs and top visited URLs by *frecency* provide the "what" and "how much," and are bolstered further by the charts of top searches and categories. The pie charts of visit types and access protocols summarize the "how" and the top visited countries and TLDs summarize

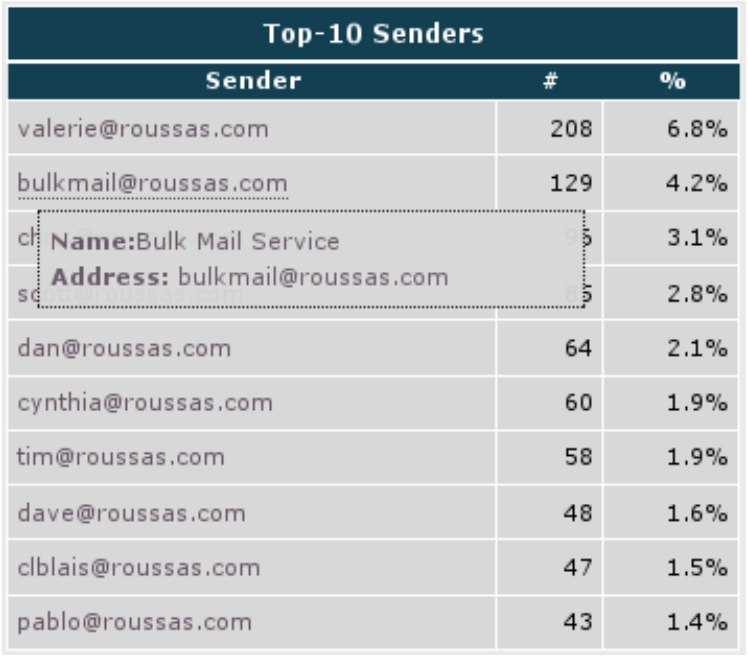
the “where” while providing a very clear depiction of the relative proportions. The bar chart showing visit counts over time provides the “when” and “how much,” and when combined with the “what” uncovers the user’s interests and browsing patterns based on the exact sites and categories of sites he visits, the searches he executes, and the dates he makes those visits.

Having reviewed the summary report, the analyst has already formed a high-level profile of the user’s browsing activity which can be further verified or investigated in detail through examination of the timeline visualization. No one image can completely capture the entirety of a large dataset in a limited amount of space, however, the combination of visualizations that was presented give the analyst a huge advantage over having to manually sort through the raw data and attempt to understand the contents in any meaningful way. With a high-level understanding of the history and the detailed presentation of the timeline, the analyst can turn to the original dataset with something specific to search for.

A demonstration of the utility on a web browsing history will be presented in the next chapter, followed by a narrative analysis of the summary report. This will be followed by various usage examples to extract history details.




(a) Tooltip from mouseover on full URL in web summary




(b) Tooltip from mouseover on address in mail summary

Figure 4.7: Tooltips for displaying additional detail

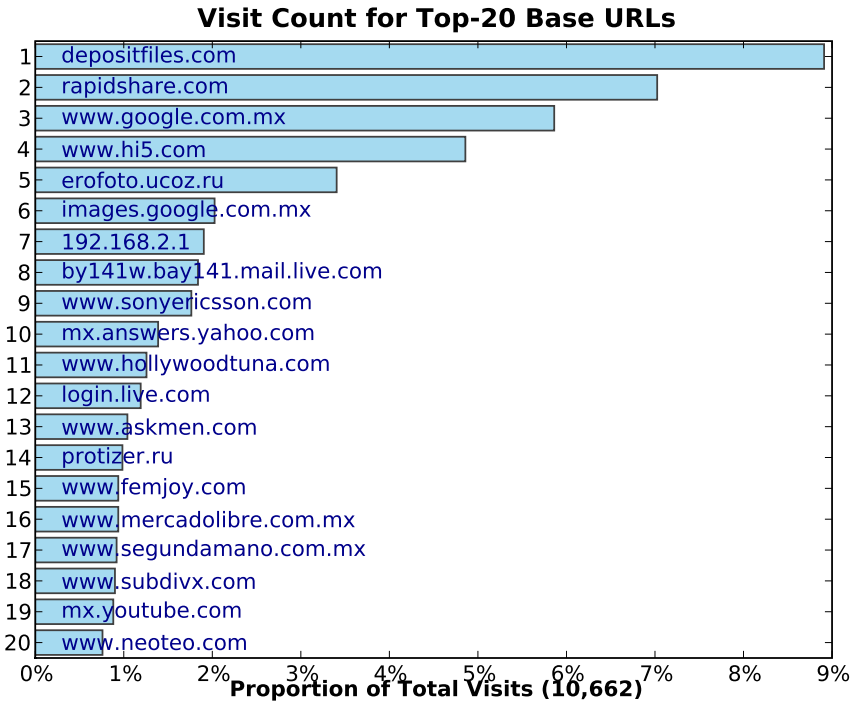
**Title:** Google
**URL:** www.google.com
**Referrer:** -
**Visit Date:** Sun May 31 14:47:58 2009
**Visit Count:** 1
**Visit Type:** Typed
**Visit Duration:** 8 seconds
**Country Name:** United States
**Country Code:** US
**Category:** Computers:Internet:Searching:Search Engines:Google
**Search Query:** -

**Title:** xss vulnerability - Google Search
**URL:** www.google.com/search
**Referrer:** www.google.com/
**Visit Date:** Sun May 31 14:48:06 2009
**Visit Count:** 1
**Visit Type:** Link
**Visit Duration:** 32 seconds
**Country Name:** United States
**Country Code:** US
**Category:** Computers:Internet:Searching:Search Engines:Google
**Search Query:** xss vulnerability

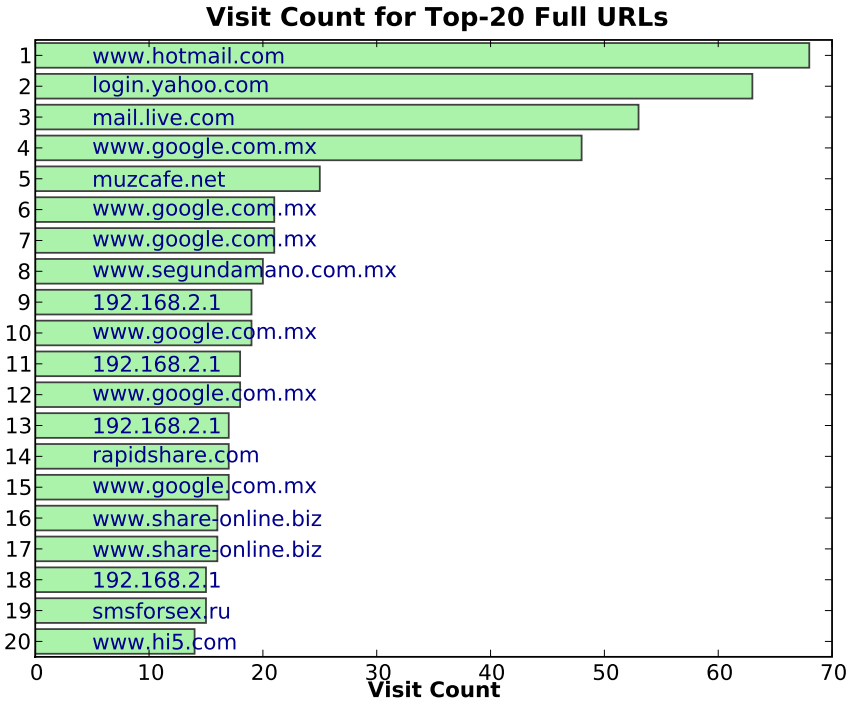
**Title:** PayPal XSS Vulnerability Undermines EV SSL Security - Netcraft
**URL:** news.netcraft.com/archives/2008/05/
**Referrer:** www.google.com/search
**Visit Date:** Sun May 31 14:48:38 2009
**Visit Count:** 1
**Visit Type:** Link
**Visit Duration:** 2 minutes 9 seconds
**Country Name:** United Kingdom
**Country Code:** UK
**Category:** Computers:Internet:Statistics and Demographics
**Search Query:** -

**Title:** Netcraft What's That Site Running Results
**URL:** uptime.netcraft.com/up/graph/
**Referrer:** news.netcraft.com/
**Visit Date:** Sun May 31 14:50:47 2009
**Visit Count:** 1
**Visit Type:** Link
**Visit Duration:** 37 seconds
**Country Name:** United Kingdom
**Country Code:** UK
**Category:** Computers:Internet:Statistics and Demographics
**Search Query:** -

Figure 4.8: Navigation history details from section of web history

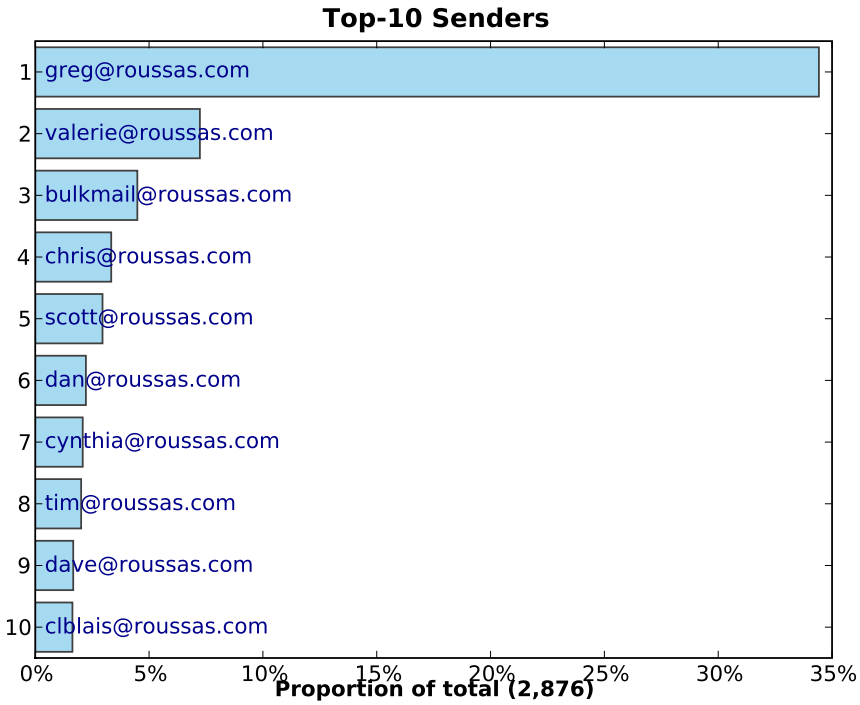



(a) Top-10 visited base URLs

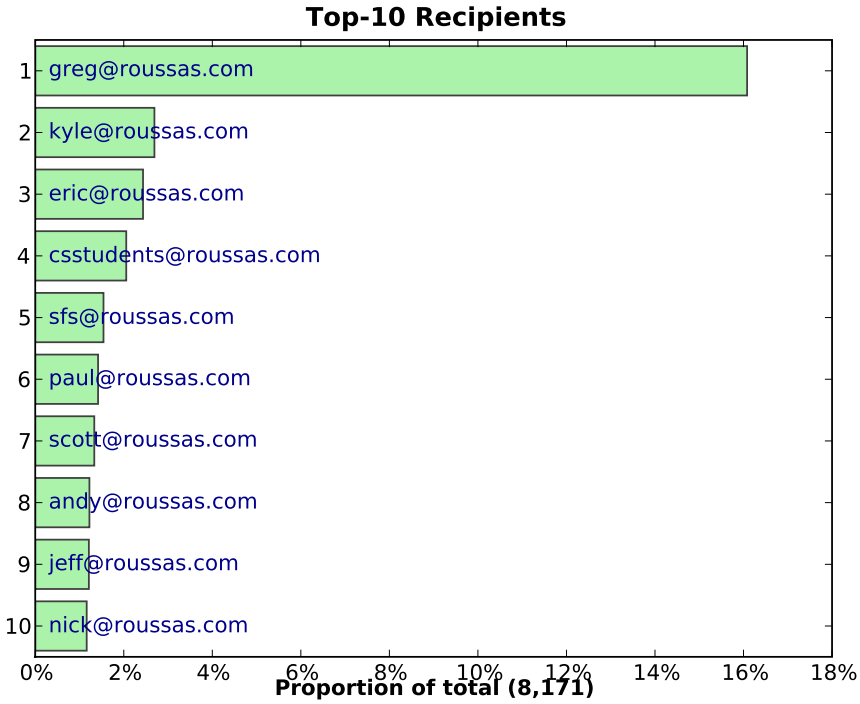



(b) Top-10 visited full URLs

Figure 4.9: Top visited base and full URLs from web summary

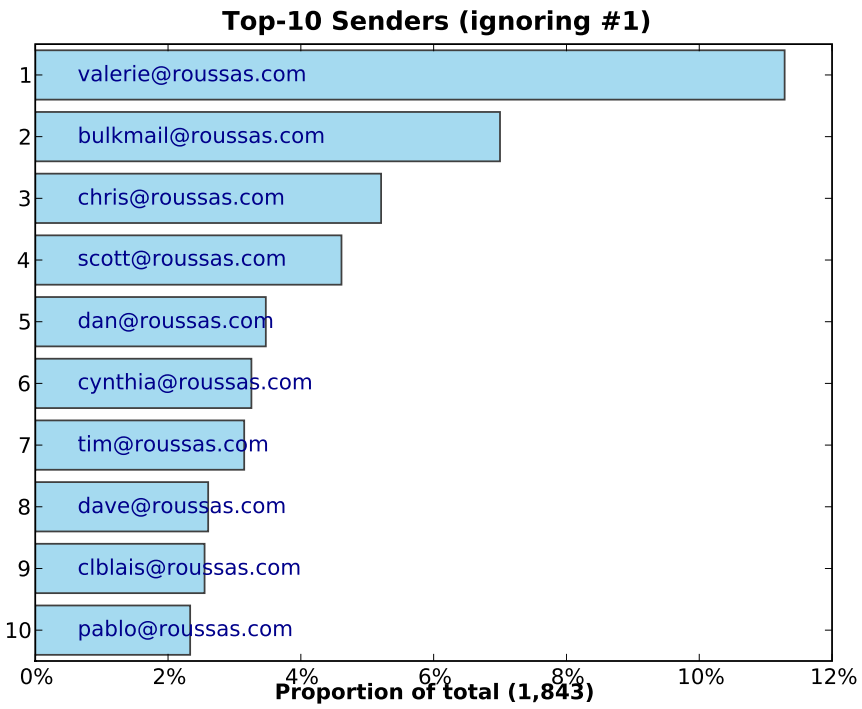



(a) Top-10 senders

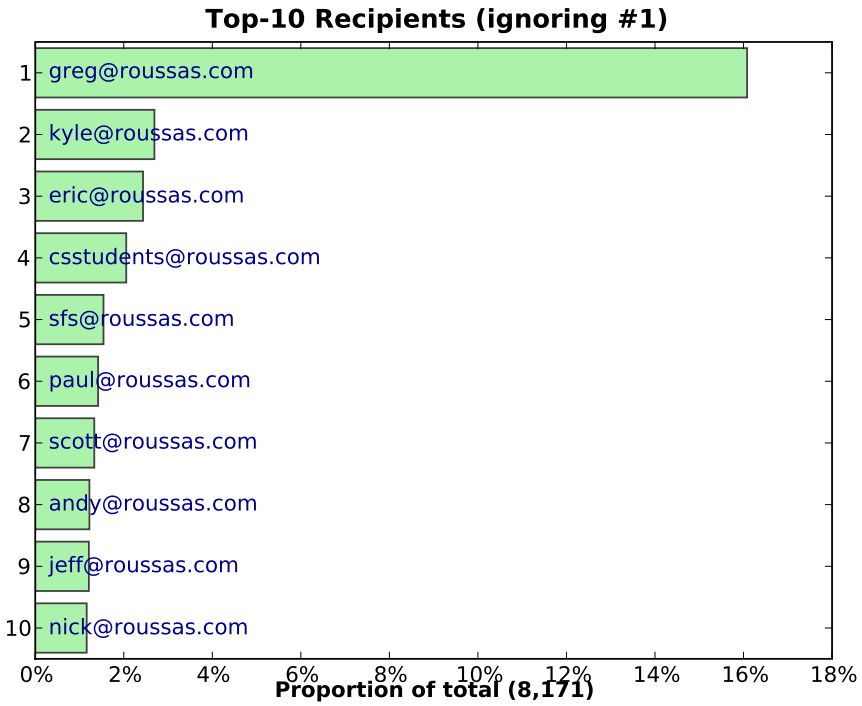



(b) Top-10 recipients

Figure 4.10: Top senders and recipients from mail summary

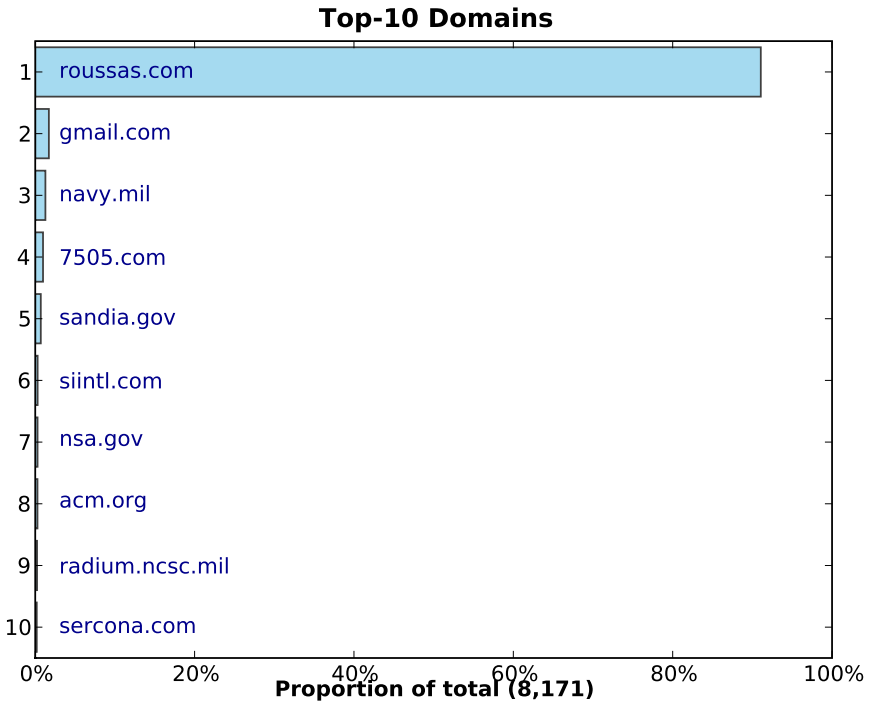



(a) Top-10 senders ignoring most frequent

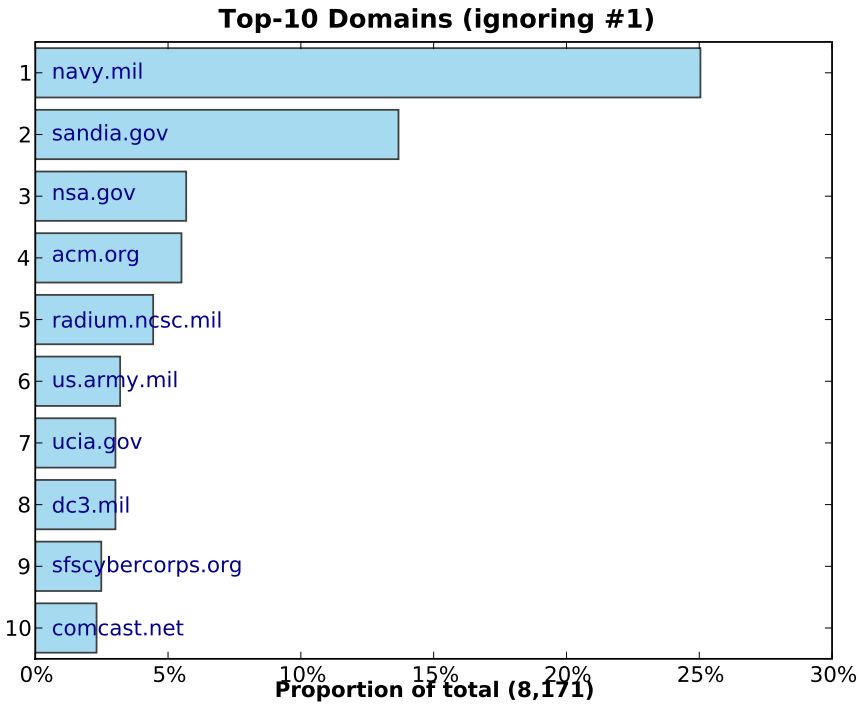



(b) Top-10 recipients ignoring most frequent

Figure 4.11: Top senders and recipients from mail summary ignoring most frequent

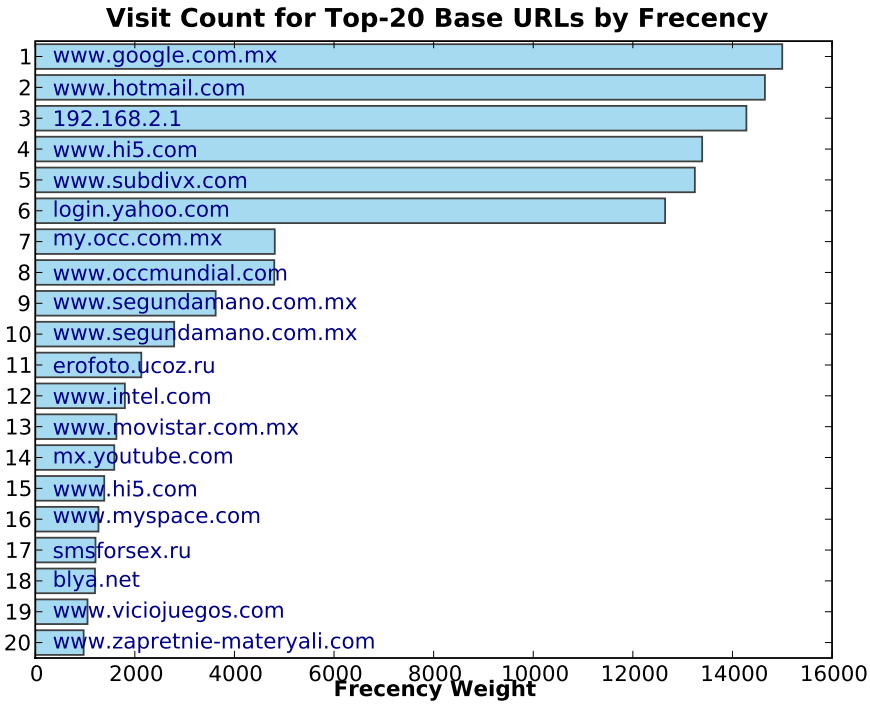



(a) Top-10 domains

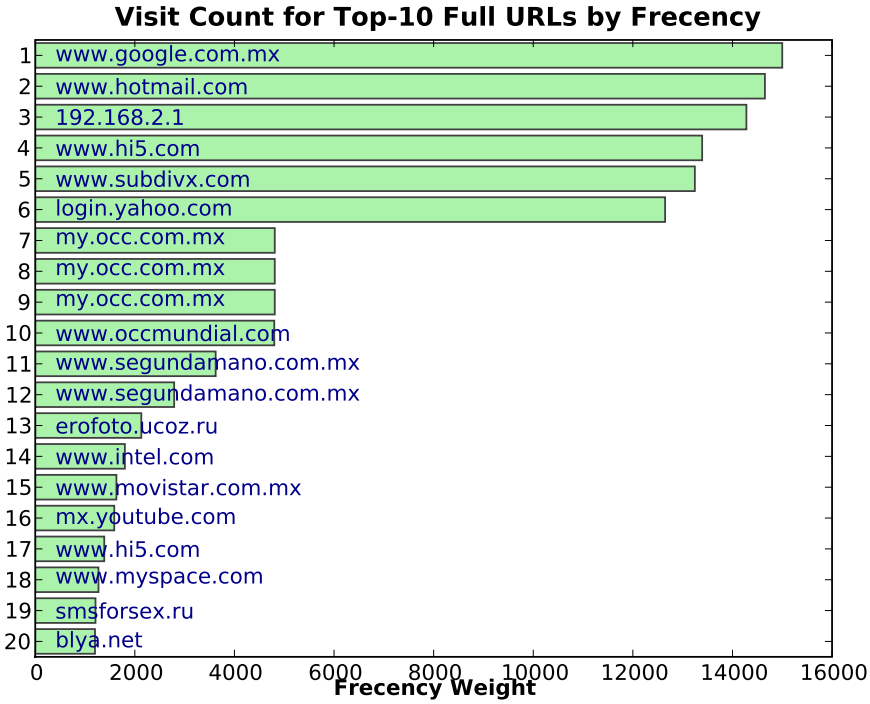



(b) Top-10 domains ignoring most frequent

Figure 4.12: Top domains from web summary ignoring most frequent

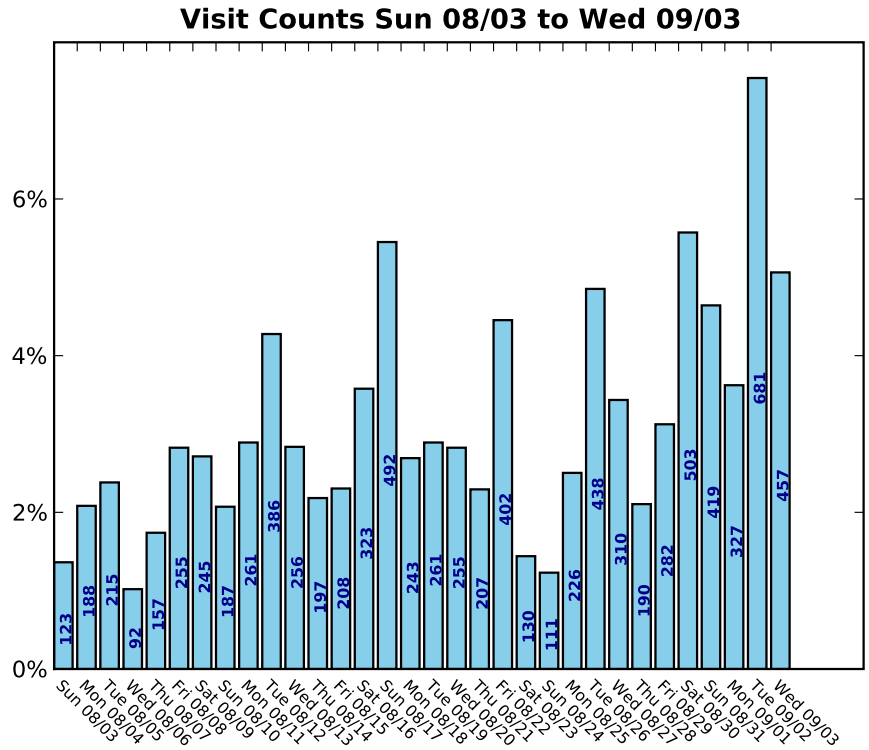



(a) Top-10 Base URLs by Frecency

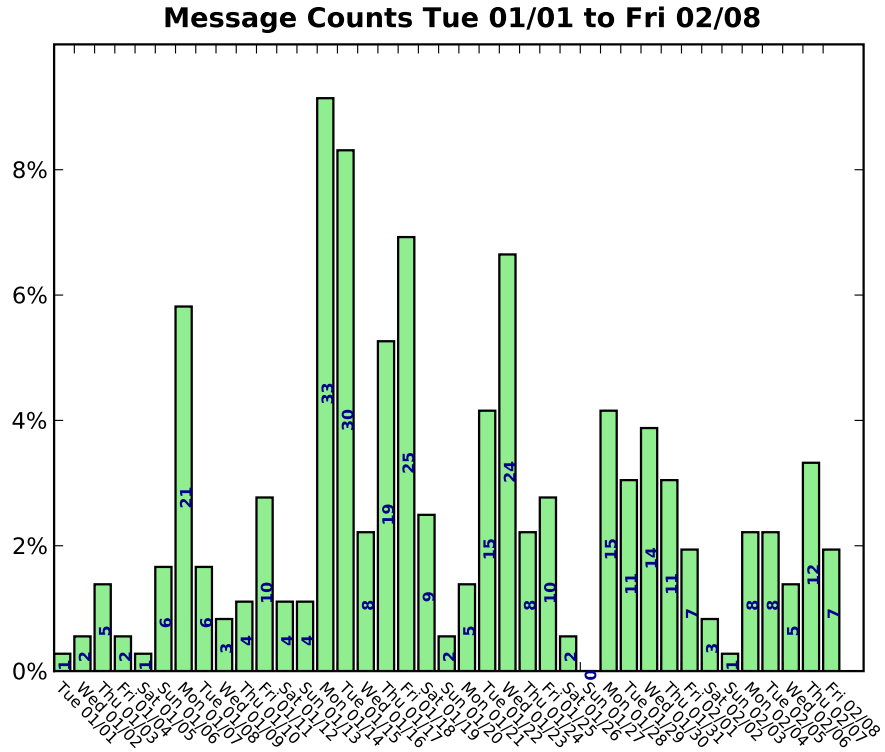



(b) Top-10 Full URLs by Frecency

Figure 4.13: Top visited base and full URLs by frecency from web summary

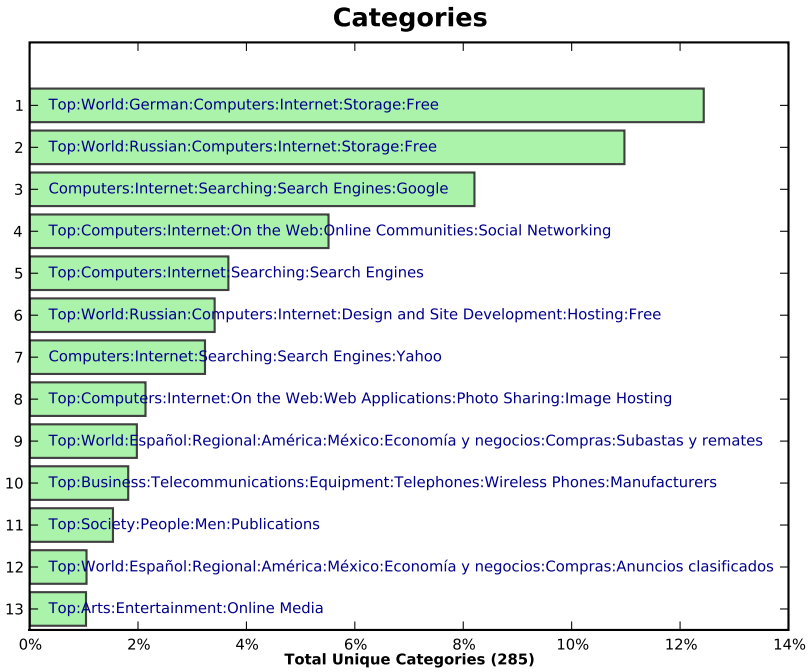



(a) Daily site visits from web history

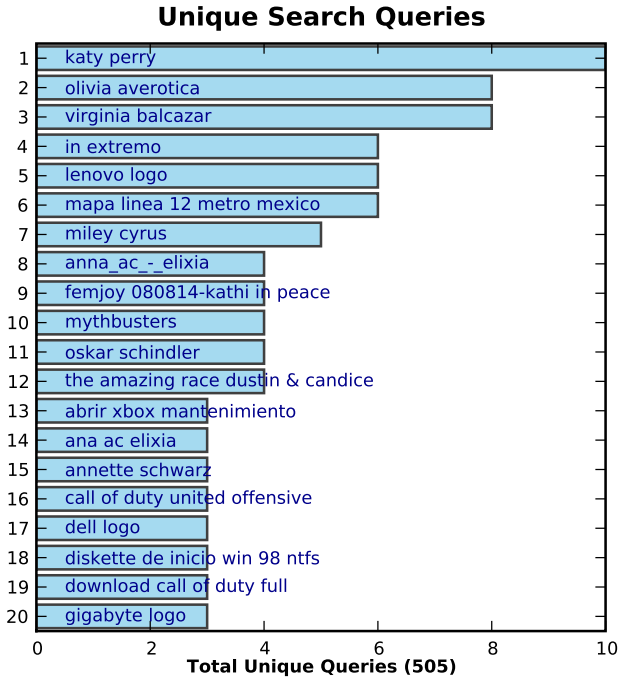



(b) Daily message counts from mail summary

Figure 4.14: Daily visit and message counts

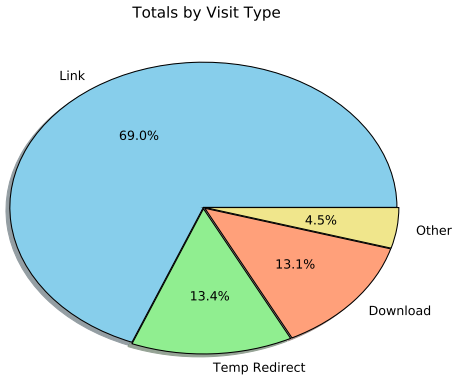



(a) Top categories for visited sites

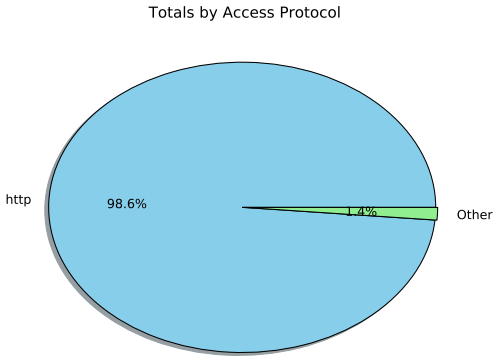



(b) Top search queries

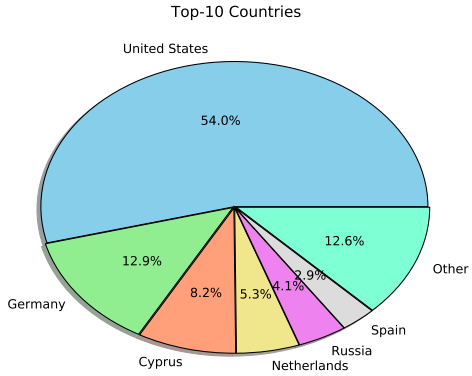
Figure 4.15: Top categories and search queries from web summary



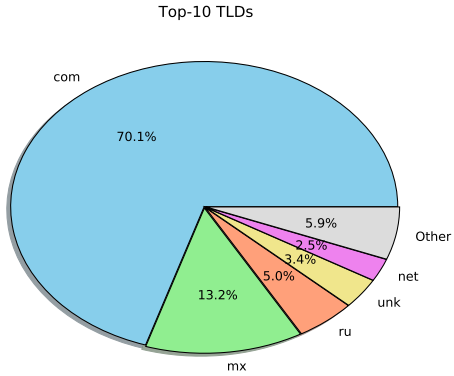

(a) Totals by Visit Type from web summary



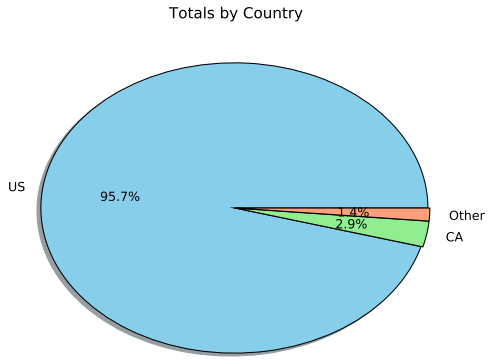

(b) Totals by Access Scheme from web summary



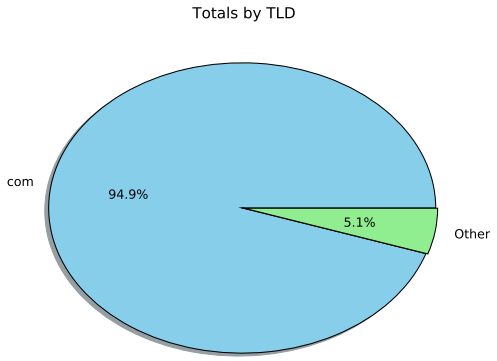

(c) Top visited countries from web summary




(d) Top visited TLDs from web summary

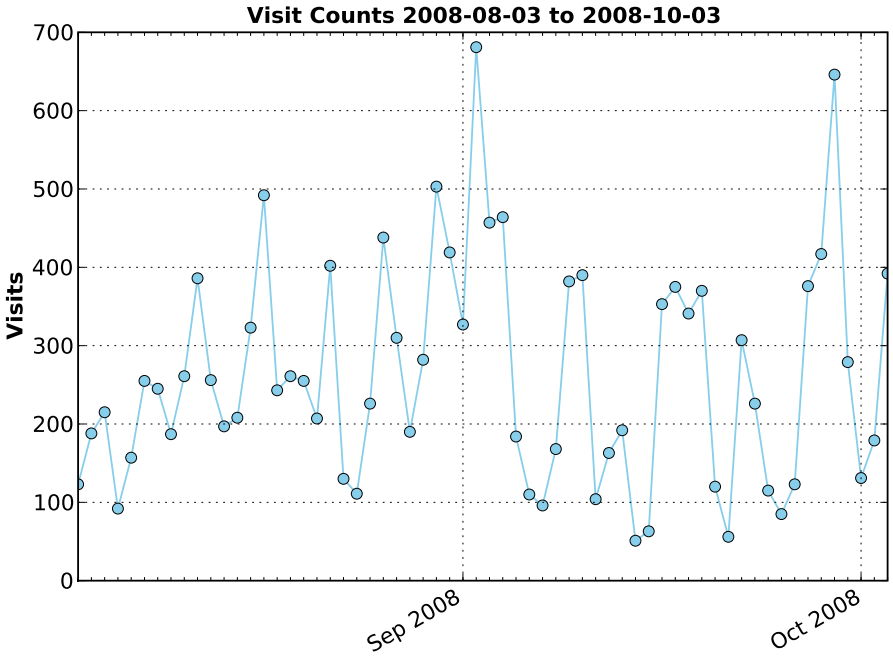



(e) Top mailed countries from mail summary

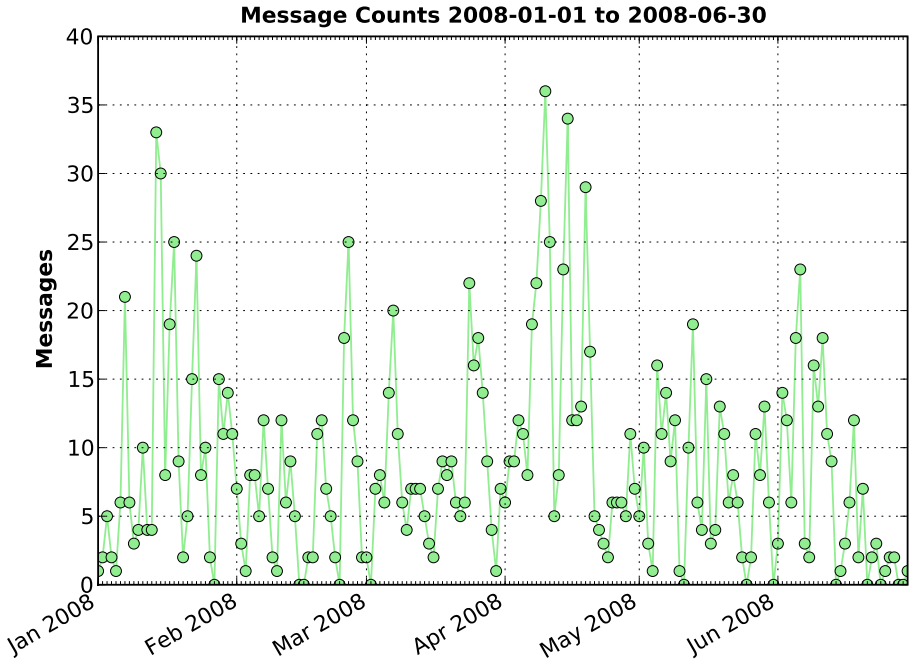



(f) Top mailed TLDs from mail summary

Figure 4.16: Totals by visit type and access scheme
Top occurring countries and TLDs




(a) Daily site visit counts from web summary




(b) Daily message counts from mail summary

Figure 4.17: Daily visit and message counts time-series

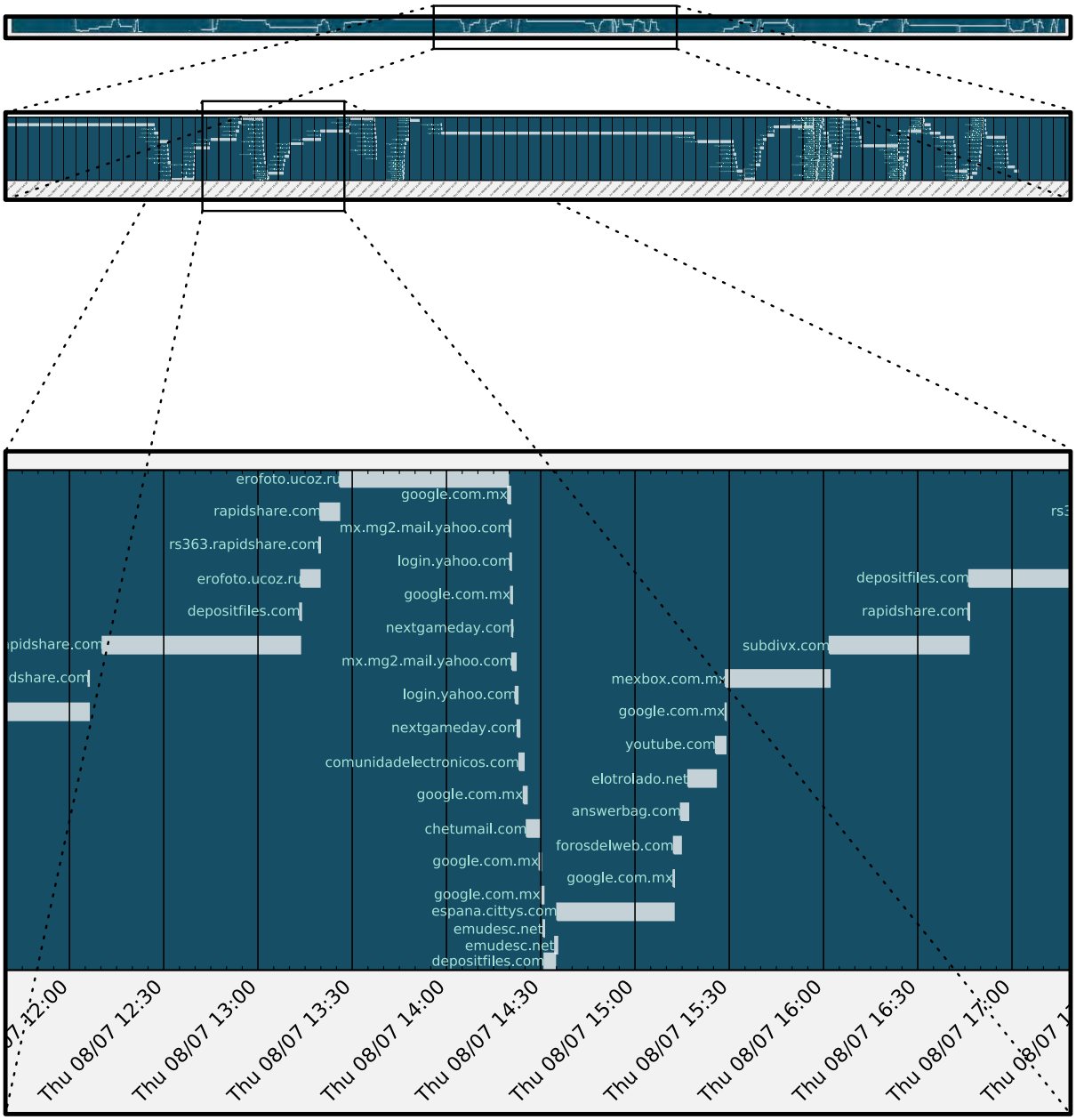



Figure 4.18: Timeline of web history

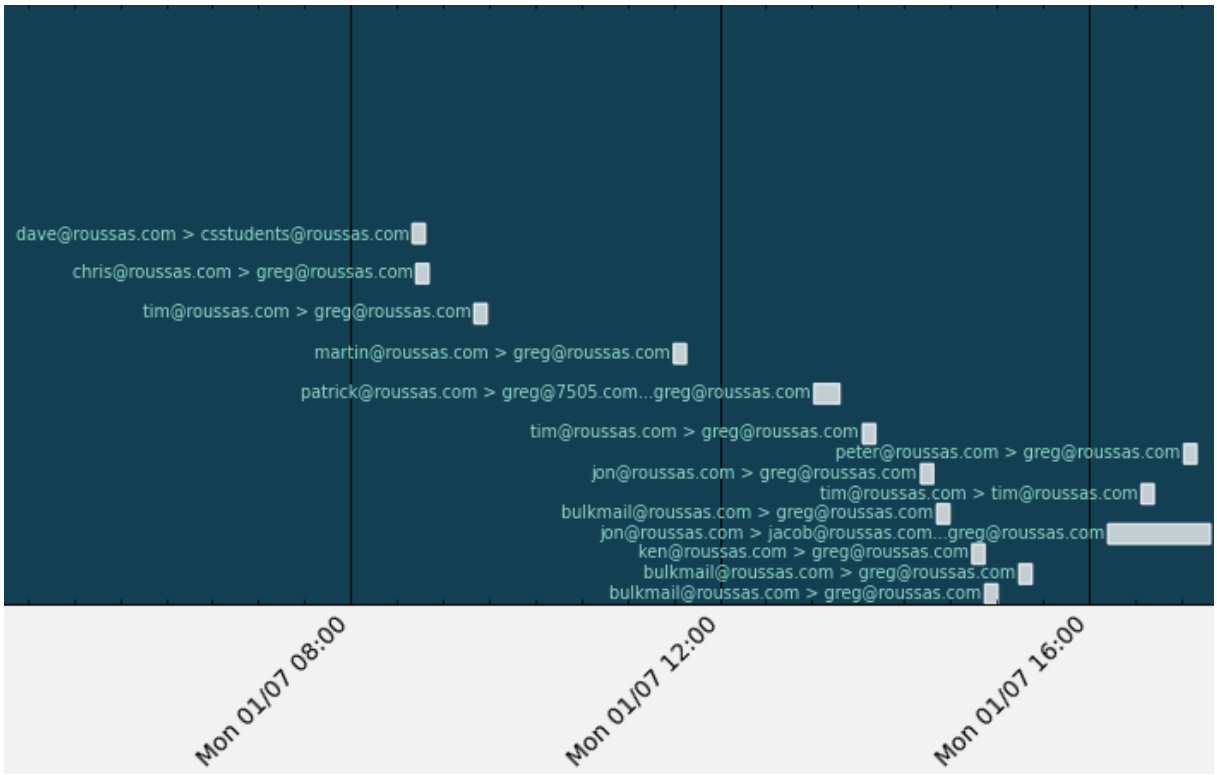



Figure 4.19: Timeline of mail history (single day detail)

# CHAPTER 5:
# Sample Analysis Session

This chapter will demonstrate the use of this utility to generate the visualizations and summary report of a web browser history file. The functionality of the utility will be illustrated and the generated summary report will be presented in narrative fashion, similar to an examination made by a forensic analyst. After forming a high-level profile of the user from the summary report, additional functionality of this utility will be demonstrated to search for details in the history database.

There are an abundance of web history files freely available on the web, possibly as a result of social bookmarking services offering centralized storage for all a user's bookmarks making them easy to share publicly. It also appears to be common practice for people to maintain web-accessible copies of their browsing history, possibly to facilitate sharing among different machines. The history file used in the analysis below was one of a number that were downloaded based on the results of a search for *places.sqlite*. It was chosen from among the others for the diverse nature of the browsing activity which spans many different countries, TLDs and categories. All work is performed within a virtual machine instance of Ubuntu 8.10 running on a 2.0 GHz single-core processor with 512 MB of allocated memory. The databases queried during the ingest and category resolution phase are located on a different machine, and all lookups are performed over the network. All reported execution times should be viewed with this in mind, and all processing times would be greatly improved by running everything on a standalone workstation.

## 5.1 Ingest, Category Resolution and Report Generation

A typical session will be separated into a number of phases which can be performed in sequence or separately, depending on the availability of machine resources and the analyst's schedule. The phases are: ingest, category resolution, summary report generation, report examination and detail extraction. Beginning with the ingest phase, the contents of the history file are parsed and inserted in the internal database that will be used in all subsequent phases. The utility is executed with the following command line arguments: `--ingest --type web --file places.sqlite`. The debug-level console output of this, and all subsequent commands, can be seen in Appendix B. The time to complete each phase is dependent on the number of records in

the history file and the utilization level of the machine at the time of execution. The ingest phase for this history file, which is nearly 15 MB in size, took approximately 33 minutes to complete. At the end of this phase each history record will be fully populated as seen in Table 5.1 with the exception of the category field.

```
            id = 525
       session = 156
           url = http://www.google.com.mx/search?num=20&hl=es&safe=off&
                 q=firefox+3.0.1+windows+live+hotmail&btnG=Buscar&meta=l
                 r\%3Dlang_es
      url_base = www.google.com.mx
        scheme = http
           tld = mx
    ref_url_id = 462
       ref_url = http://www.google.com.mx/search?hl=es&q=firefox+3.0.1+
                 windows+live+hotmail&meta=&btnG=Buscar+con+Google
  ref_url_base = www.google.com.mx
         title = firefox 3.0.1 windows live hotmail - Buscar con Google
    visit_date = 1217916275
visit_date_fmt = Mon Aug  4 23:04:35 2008
visit_duration = 12
   visit_count = 1
    visit_type = 1
      embedded = 0
         typed = 0
      frecency = 10
   favicon_url = http://www.google.com.mx/favicon.ico
  country_code = US
  country_name = United States
      category = Computers/Internet/Searching/Search_Engines/Google
  search_query = firefox 3.0.1 windows live hotmail
```

Table 5.1: Sample history record after ingest phase

If we want to see the category bar chart in the summary report, the categories will need to be resolved for each history record by executing the utility with the `--resolve-categories` argument. This function was separated from the initial ingest phase because it can be lengthy depending on the types of URLs in the history file and whether they have been encountered before. If a URL has been resolved in any previous history file ingest, it will have a URL-to-category mapping in the local category database, but if not, a lookup in the ODP database will be necessary. The completion time of this lookup depends on how quickly the URL can be matched to a category, for example, the debug output in Table 5.2 shows an instance of a lookup that was performed against a number of different URL patterns in an attempt to match the URL with a category, but was unsuccessful since there was no entry in the ODP for this URL.

```
www.nctc.gov (1 visit)
Domain name: "nctc.gov"
Unknown URL; searching on "http://www.nctc.gov/"
Searching again with "www" removed: "http://nctc.gov/"
Searching again without trailing "/": "http://www.nctc.gov"
Searching again with wildcard as "http://www.nctc.gov/\%"
Searching again with wildcard: "\%nctc.gov\%"
Category not found for www.nctc.gov
```

Table 5.2: Unsuccessful category lookup

After the category resolution is complete, each record with an entry in the ODP database will have its category field populated as seen in Table 5.1. The next phase will be the summary generation phase, whose completion time once again depends on the number of records in the history file. The generation of the summary report for this particular history file took nearly two minutes, and once complete, the summary report is ready for analysis.

## 5.2 Summary Report Analysis

The summary report was designed to be read from the top down with the displayed information progressing from general to detailed. Glancing at the summary table shown in Figure 5.1, we can see a number of unique counts that were defined in Section 4.2.1 of Chapter 4. The aggregation of inter-site browsing shows that of the 15,555 visited unique URLs only 2,534 or approximately 16 percent had unique base URLs leaving approximately 84 percent of the browsing during this period as inter-site in nature. If we are not interested in the user's inter-site browsing, this can greatly reduce the number of sites we need to inspect in more detail.

The number of visited TLDs and countries is fairly high indicating a diverse geographic distribution in the user's interests. Of the 127 embedded URLs 17 were visited, so this might indicate something to follow up on during the detail extraction phase. There are 2,875 downloads accounting for approximately 15 percent of all visits, which may be another area to focus on later. The user typed 103 URLs directly into the address bar, which we will also focus on later. There were 410 unique categories resolved and over 500 search queries extracted, which should tell us even more about the user's interests. Next, we turn to the tables and charts showing the top-visited base and full URLs to see what type of sites the user frequents.

A glance at the top-10 base URLs shown in Figures 5.2 and 5.3 shows they account for approximately 40 percent of the total visits, giving us a good indication of the types of sites this user is

| Total Unique Counts | | | |
|---|---|---|---|
| Type | Total | Visited | Not Visited |
| Full URLs | 15,642 | 15,555 | 87 |
| Base URLs | 2,563 | 2,534 | 29 |
| Top Level Domains | 58 | 53 | 5 |
| Countries | 54 | 45 | 9 |
| Embedded URLs | 127 | 17 | 110 |
| Downloads | 2,875 | - | - |
| Typed URLs | 103 | - | - |
| Categories | 410 | - | - |
| Search Queries | 505 | - | - |

Figure 5.1: Summary table of unique counts

interested in. The top two sites, accounting for 13 percent of the visits, are file-sharing services, confirming the large number of downloads we saw in the summary table above. Google search and Google images account for nearly eight percent of visits, very useful from our perspective since this utility extracts all these search queries. The fourth, sixth and tenth site are not familiar, so hovering over the URL in the table to display the tooltip, we see that all of the page titles contain the name "Katy Perry," which a quick search confirms is a recording artist. Nearly five percent of this user's total visits were directly to the artist's fan site, and another five percent were to sections of entertainment news sites possibly associated with the same name. The fifth most-visited site is a social networking site, so examination of the corresponding full URLs could tell us a lot about other people this user associates with. The seventh most visited site is a Russian photo- and video-sharing site, and the ninth is the Microsoft `live.com` webmail service.

Viewing the visualizations for the top-visited base and full URLs ordered by *frecency*, which assigns more weight to sites that were more recently visited, typed directly in to the address bar or initiated from bookmarks, we can see that there is a re-ordering of the sites. Depending on the details of the case, the most recently visited sites may be important, in which case we would concentrate on these two charts.

| Top-10 Unique Base URL Visit Counts | | |
|---|---|---|
| **Base URL** | **#** | **%** |
| depositfiles.com | 1,424 | 7.7% |
| rapidshare.com | 983 | 5.3% |
| www.google.com.mx | 935 | 5.1% |
| www.katyperry.com.mx | 889 | 4.8% |
| www.hi5.com | 838 | 4.5% |
| www.buzznet.com | 598 | 3.2% |
| erofoto.ucoz.ru | 573 | 3.1% |
| images.google.com.mx | 472 | 2.6% |
| by141w.bay141.mail.live.com | 365 | 2.0% |
| cdn.buzznet.com | 342 | 1.8% |

| Top-10 Unique Full URL Visit Counts | | |
|---|---|---|
| **Full URL** | **#** | **%** |
| www.google.com.mx/ | 150 | 0.8% |
| www.hotmail.com/ | 111 | 0.6% |
| login.yahoo.com/config/mail.intl=mx&.lg= | 102 | 0.6% |
| mail.live.com/ | 53 | 0.3% |
| www.segundamano.com.mx/login.aspx | 45 | 0.2% |
| www.segundamano.com.mx/Particulares/Gest | 31 | 0.2% |
| muzcafe.net/id=612 | 25 | 0.1% |
| www.hi5.com/ | 24 | 0.1% |
| smsforsex.ru/ | 22 | 0.1% |
| www.hi5.com/friend/profile/displaySamePr | 22 | 0.1% |

Figure 5.2: Top visited base and full URLs

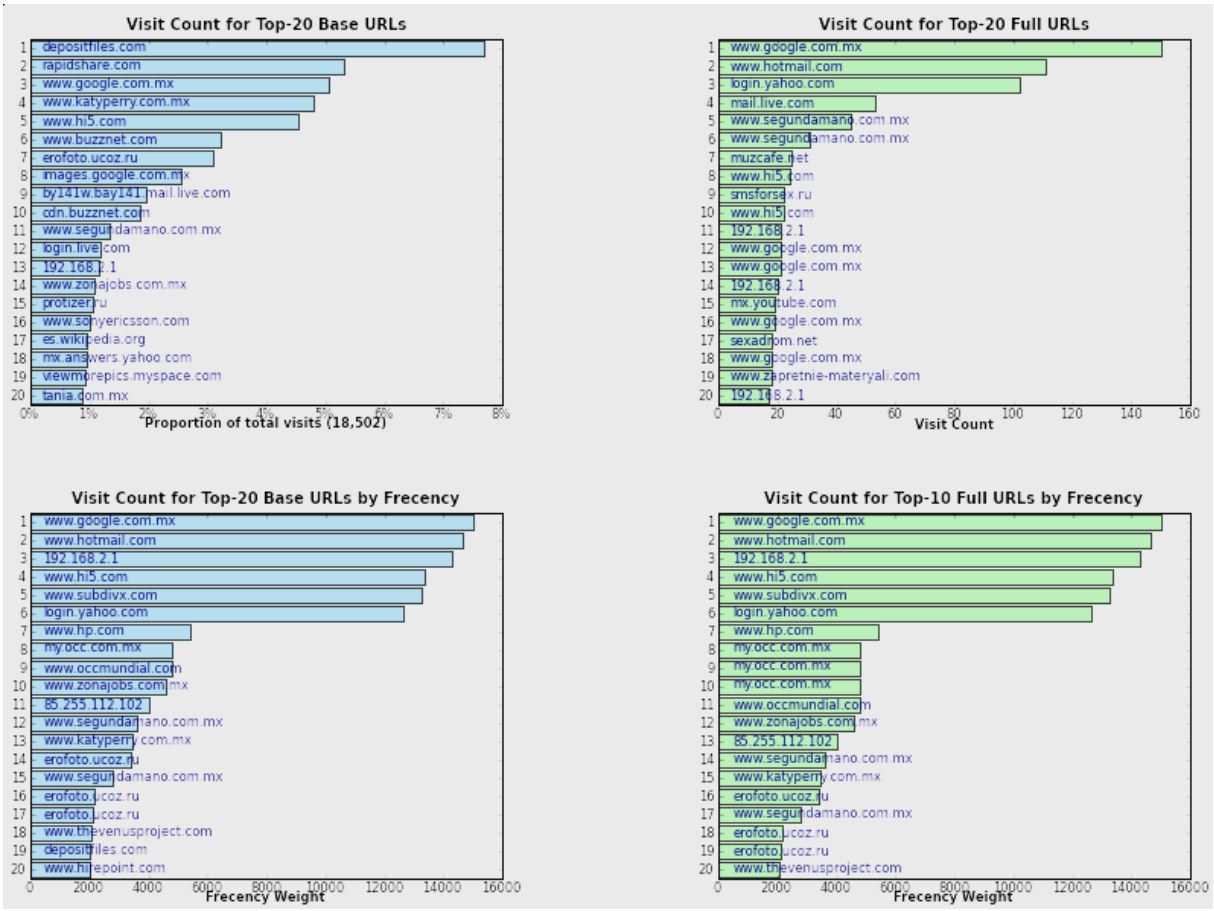



Figure 5.3: Top visited base and full URLs by count and frecency

Continuing with the report, we see the breakdown in type of browsing by looking at the totals by visit type and access protocol tables and pie charts as shown in Figure 5.4. We confirm what we had already determined by viewing the summary table above: 15 percent of this user's total browsing was downloads. We can see that nearly 70 percent of his browsing is link-following and only two percent of his visits are initiated from bookmarks. The user typed in 125 URLs of which 103 were unique as we saw in the summary table. The temporary and permanent redirects are probably not interesting, so we ignore them. The "totals by access protocol" table and pie chart shown in Figure 5.4 show that the majority of the user's browsing is standard `http`, but there were 11 visits using the `file` protocol indicating local file on the machine or network, and seven using `ftp` which were file transfers. We will look at these in detail in the next section.

The top visited countries and TLDs shown in Figure 5.5 will give us insight into where the user browses. The majority of his browsing is to the `.com` TLD and sites located in the US, followed by nearly 27 percent of browsing concentrated in Germany, Cyprus, Netherlands and Russia. This would seem to confirm what we noted above about the top two visited sites, which were file sharing sites, accounting for 13 percent of all visits. A quick check confirms that the most visited site (`depositfiles.com`) is hosted in Cyprus and the second most visited site (`rapidshare.com`) in Germany. There is one additional piece of information to note: based on the top TLDs table, nearly 20 percent of the user's browsing was to the `.mx` TLD, which increases the probability that his native language is Spanish.

Moving on to the top search queries table and bar chart shown in Figure 5.6 we can see that there were over 500 queries extracted during the ingest phase. The top query which was repeated 10 times is for the same recording artist mentioned in the top base URLs. The majority of the top-10 search queries all of which were repeated multiple times are for actors or musicians which may have a connection to the high number of downloads. We will look at the downloads in more detail in the next phase.

The top categories bar chart shown in Figure 5.6 confirms the profile we have already formed from the rest of the report by giving us a ranked list of the user's interests. As was discussed in Section 4.2.3, multiple bars which represent different categories can easily be aggregated into one. For example bars one, two and seven represent "file sharing" and bars three, six and eight "search engines." Using this generalization we can deduce that this user is someone whose primary interests are sharing files using a variety of German and Russian file-swapping sites,

using Google and Yahoo! for search, Wikipedia for answers to specific questions, buying or selling and communicating locally with specific regional auction and social networking sites, sharing or downloading photos, and keeping up on entertainment news.

Turning to the last summary visualization shown in Figure 5.7, we are looking for any patterns or anomalies in the user's browsing over time. We can see that there seems to be a regular increase in browsing activity near the end or the beginning of each month and a distinct peak in the middle of the last month. Once again, whether this is relevant to the case depends on the information the analyst has on the user and the motivation behind the investigation. Now that we understand the user's browsing behavior and have formed a high-level profile, we will look at some of the information we identified above in more detail.

## 5.3 Details

In the previous section we identified a number of attributes that we wanted to see more detail on. All the functionality to extract the details from the utility database shown below has not been fully implemented, but since all the information is in an SQL database we will simulate the output by running the necessary queries manually. In many cases the output below was modified for presentation, so a more complete view can be seen in Appendix C.

First we will look at the embedded links the user followed. When this functionality is completely integrated, the `--get-records --type embedded --visited true --field url` argument will be used to execute the search and the URL for each of these records will be returned as shown below.

```
http://www.google.com/ig/modules/translatemypage.xml&up_source_language=ru
  &w=160&h=60&source=http://erofoto.ucoz.ru/news/10-0-1
http://www.google.com/ig/modules/translatemypage.xml&up_source_language=ru
  &w=160&h=60&source=http://erofoto.ucoz.ru/news/2-0-1
  ... output truncated ...
http://www.google.com/ig/modules/translatemypage.xml&up_source_language=ru
  &w=160&h=60&source=http://erofoto.ucoz.ru/news/2008-09-20-1295
http://www.google.com/ig/modules/translatemypage.xml&up_source_language=ru
  &w=160&h=60&source=http://erofoto.ucoz.ru/news/2008-10-31-1515
```

Examining each URL we can see that the user has clicked on an embedded Google gadget module which translates between languages. This module can be inserted into any web page and

the contents of that page will be translated in the language you select. In this case, the user was browsing a number of different pages within the `/news` section of a Russian content-sharing site. This not only confirms his interest in file-swapping, but also that his native language is not Russian, however, the target language for the translation is not encoded in the URL, so we can not say anything about the native language of the user. Examination of the page referenced by each URL will reveal more detail about the nature of the content the user was translating.

Now we will extract the downloads the user has completed to see the nature of the files he is interested in. The utility will be executed with the `--get-records --type download --field url_base --field filename` and the base URL and filename of each download will be extracted as shown here:

```
files.hamachi.cc           HamachiSetup-1.0.3.0-es.exe
rs526tg.rapidshare.com     MC_-_2008-03-12_-Tropical_Lounge.part2.rar
www.katyperry.com.mx       0729.jpg
www.katyperry.com.mx       0631.jpg
download.divx.com          DivXUserGuide52.exe
download.skype.com         SkypeSetup.exe
ulises2k.googlepages.com   WGA_crack_by_Ulises2k.zip
ulises2k.googlepages.com   LegitCheckControl_Universal_by_Ulises2k.zip
files2.subdivx.com         113244.rar
files2.subdivx.com         86519.rar
rs164l34.rapidshare.com    Mila_B_-_Presenting_Mila.part1.rar
upload.wikimedia.org       Oktoberfest.JPG/800px-Oktoberfest.JPG
download.nai.com           sdat5419.exe
  ... output truncated ...
```

Even a small sample of the user's downloads can reveal a wealth of information about his interests, and a quick examination of the full list confirms our profile of the user formed from the summary report. For example, approximately 24 percent of the downloads had the name "katy perry" in either the URL or filename, 54 percent of the downloaded files were `jpg` images, and 34 percent were `rar` and `zip` archives which can contain any filetype, but in this case, were primarily movies and executable programs based on the filename.

The URLs directly typed in by the user will be extracted by executing the utility with `--get-records --type typed --field visit_count --field url` resulting in this output:

```
24        http://www.hi5.com/
21        http://192.168.2.1/
15        http://www.subdivx.com/
 9        http://www.zonajobs.com.mx/
 6        http://depositfiles.com/es/files/7859768
 4        http://www.myspace.com/
 4        http://www.occmundial.com/
 4        http://www.zeitgeistmovie.com/
 3        http://erofoto.ucoz.ru/news/2008-09-28-1400
 3        http://www.femjoy.com/models.php
 3        http://www.hp.com/
 3        http://www.intel.com/
 3        http://www.katyperry.com.mx/
 3        http://www.movistar.com.mx/
 ... output truncated ...
```

Once again, examination of this list can be used to confirm our existing profile of the user, since the typed URLs follow the theme we have already outlined. To round out our detail extraction phase, we retrieve the search queries with the `--get-records --type search_queries --field url_base --field search_query` argument, resulting in the following output:

```
www.google.com.mx       olivia averotica
images.google.com.mx    site:www.filmgecko.com becki newton
images.google.com.mx    "oskar schindler""
images.google.com.mx    "virginia balcazar"
images.google.com.mx    annette schwartz
images.google.com.mx    diskette de inicio win 98 ntfs
images.google.com.mx    gugabyte ga-945gzm-s2
images.google.com.mx    gigabyte ga-945gzm-s2
images.google.com.mx    large hadron collider
images.google.com.mx    ntfs desde el diskette de inicio win 98
www.google.com.mx       abrir eliminador de laptop
www.google.com.mx       abrir x box mantenimiento
www.google.com.mx       alcohol isopropilico
www.google.com.mx       descargas rapidshare corta final firefox
www.google.com.mx       desactivar actualizacion de windows media player 11
www.google.com.mx       metallica death magnetic lyrics
www.google.com.mx       metallicas new single
 ... output truncated ...
```

This is a small selection of the 505 search queries, and when the full list is examined with the search queries ordered by visit date, it reads like a narrative of the user's browsing session. You can see the natural progression of the browsing sessions with searches building on one another, each one getting closer to the user's final objective.

Finally, if we were interested in viewing the navigation details for browsing that occurred during the spikes at the month boundaries, we could generate a summary report for that period with `--create-summaries --type web --date-begin 2008-09-29 --date-end 2008-10-02` and extract the individual records with `--get-records --type web --date-begin 2008-09-29 --date-end 2008-10-02`. The timeline visualization could be used by itself or in conjunction with the extracted records to see the exact details of the user's browsing during this period.

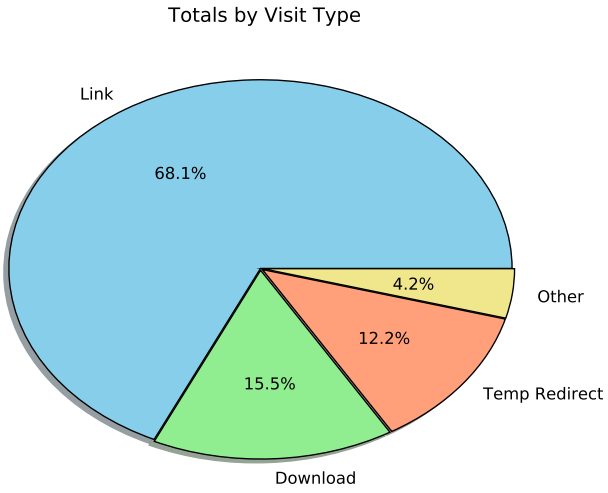
## 5.4 Summary

The breakdown and ranking of various attributes of the user's browsing in the summary report allows us to form a high-level profile of the user and reveal specific attributes to focus on in more detail. The details that we extract in the following phase can then be used to confirm this profile and support it with more specific information. There will undoubtedly be certain attributes that are more interesting than others depending on the background knowledge we have on the user and the particular details of the investigation. The summary report will allow us to quantify just how much we need to focus on and how much we can safely ignore. For example, if the user's interest in "Katy Perry" is not relevant to the investigation we know that we can ignore nearly 10 percent of the entire history file, as was established above by examination of the top-visited base URLs bar chart.

The next chapter will present a summary of findings, issues and features that have not been addressed and possible avenues for future work.

| Totals by Visit Type | | |
|---|---|---|
| **Visit Type** | **#** | **%** |
| Link | 12,591 | 68.1% |
| Download | 2,875 | 15.5% |
| Temp Redirect | 2,255 | 12.2% |
| Bookmark | 396 | 2.1% |
| Perm Redirect | 260 | 1.4% |
| Typed | 125 | 0.7% |

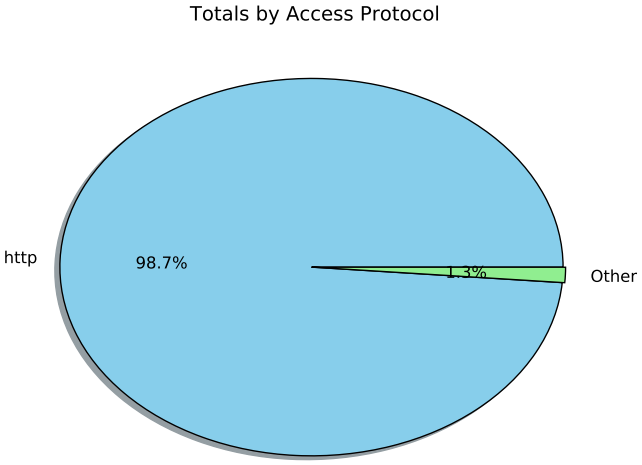
(a) Totals by visit type table




(b) Totals by visit type pie chart

| Totals by Access Protocol | | |
|---|---|---|
| **Protocol** | **#** | **%** |
| http | 18,260 | 98.7% |
| https | 224 | 1.2% |
| file | 11 | 0.1% |
| ftp | 7 | 0.0% |

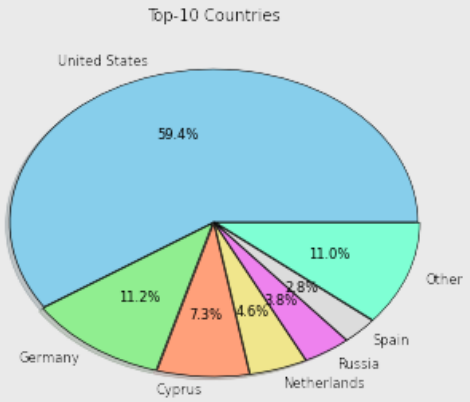
(c) Totals by access protocol table




(d) Totals by access protocol pie chart

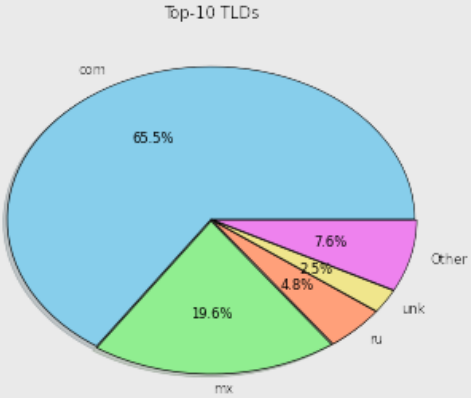
Figure 5.4: Totals by visit type and access scheme

| Top-10 Visited Countries | | |
|---|---|---|
| **Country** | **#** | **%** |
| United States | 10,983 | 59.4% |
| Germany | 2,066 | 11.2% |
| Cyprus | 1,347 | 7.3% |
| Netherlands | 847 | 4.6% |
| Russia | 698 | 3.8% |
| Spain | 525 | 2.8% |
| Mexico | 420 | 2.3% |
| Argentina | 277 | 1.5% |
| Unknown | 275 | 1.5% |
| Israel | 247 | 1.3% |




(a) Top visited countries

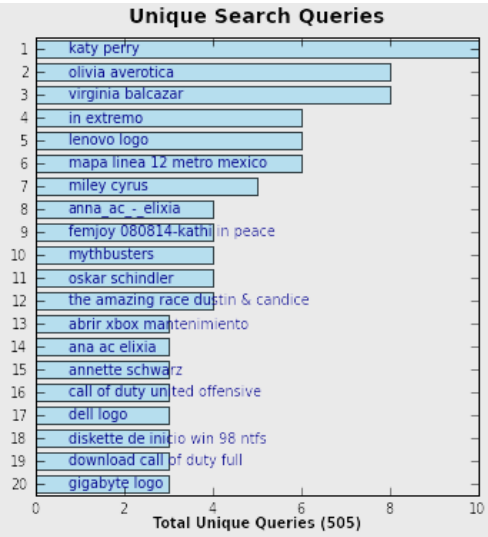
| Top-10 Visited TLDs | | |
|---|---|---|
| **TLD** | **#** | **%** |
| com | 12,123 | 65.5% |
| mx | 3,618 | 19.6% |
| ru | 886 | 4.8% |
| unk | 468 | 2.5% |
| org | 400 | 2.2% |
| net | 379 | 2.0% |
| ua | 65 | 0.4% |
| biz | 61 | 0.3% |
| es | 56 | 0.3% |
| in | 48 | 0.3% |



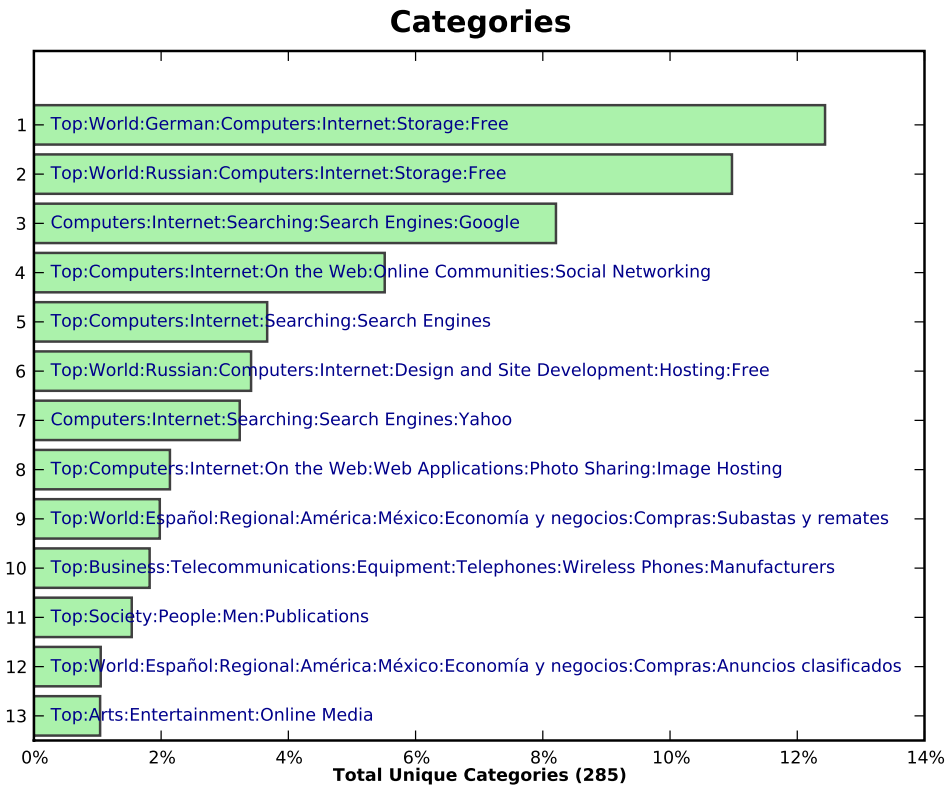

(b) Top visited TLDs

Figure 5.5: Top visited countries and TLDs

| Top-10 Unique Search Queries | | |
|---|---|---|
| Search Query | # | % |
| katy perry | 10 | 1.5% |
| olivia averotica | 8 | 1.2% |
| virginia balcazar | 8 | 1.2% |
| in extremo | 6 | 0.9% |
| lenovo logo | 6 | 0.9% |
| mapa linea 12 metro mexico | 6 | 0.9% |
| miley cyrus | 5 | 0.8% |
| anna_ac_-_elixia | 4 | 0.6% |
| femjoy 080814-kathi in peace | 4 | 0.6% |
| mythbusters | 4 | 0.6% |




(a) Top search queries




(b) Top categories

Figure 5.6: Top categories and search queries

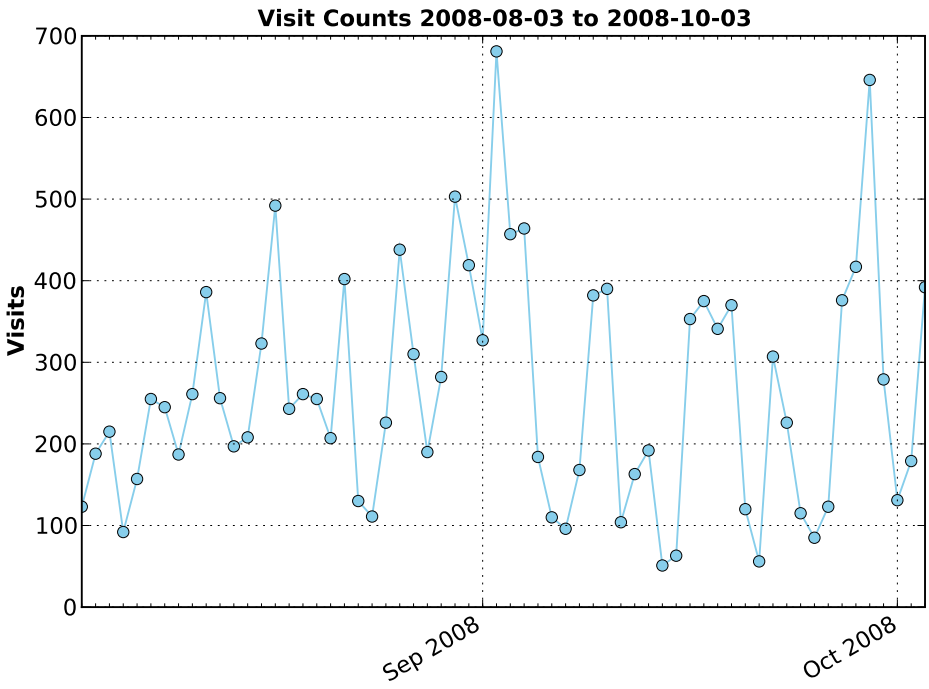



Figure 5.7: Daily site visit counts

# CHAPTER 6:
# Conclusion and Future Work

## 6.1 Conclusion

The tools currently available for forensic analysis of web browsing and mail histories vary greatly in their features and sophistication, but are similar in their almost exclusive use of tables and lists to present their findings. The large size of history files coupled with the current growth in both storage capacities and data retention, renders the exhaustive display of this information using tables obsolete. Use of visualization for presenting these textual histories is a very efficient method of distilling and summarizing the content in a format that is both easily and quickly understood at a high level. Collecting multiple visualizations with different perspectives on the data into a single summary report allows the analyst to generate a profile of the user and prioritize the information so that he can direct his focus to the highest import areas first.

This thesis was focused on the design and development of a set of tools to transform textual browser and mail histories into visual representations with the goal of facilitating the discovery of trends, patterns and relationships that may exist in the data. Background information was given on the existing file storage formats for browser and mail histories, followed by an overview of the currently available tools used in their analysis. A survey of different methods of visualization developed in the academic space to represent web traffic and mail communication was presented. Prior to the development of this toolset, attention was given to the high-level goals and requirements of the target stakeholders in order to define the features to be implemented. The operation of the tools was then sketched out with use cases demonstrating a number of different core features identified in the previous step. The tools were written to ingest and parse the histories and supplement the extracted information with metadata before storage in a database. Another set of utilities were developed to generate various visualizations and present them together in a summary report for use by the forensic analyst. A number of features were implemented to enable the analyst to extract specific details from the application database for the purpose of supplementing the knowledge he gained from the summary report.

Many visualizations that have been used to represent web browsing and mail communication are very sophisticated and effective designs, however, their novelty and complexity does not make them highly useful in an operational environment. Current best-of-breed commercial forensic

analysis suites like EnCase and FTK offer many features useful for detailed analysis of web and mail histories, but have not integrated visualization into their interfaces. EnCase has the ability to find, parse, analyze and display various elements including all visited URLs and mail messages with dates and times, the contents of the browser cache and all mail attachments, as well as reconstructing visited pages from cache including a number of webmail services. All web and mail history that is parsed can be searched by keyword and findings can be exported in a number of document formats. FTK has a similar set of features, slightly differing on the supported file formats.

There is a paucity of tools for interfacing with visualization APIs at a high level without making a considerable investment in learning. It is this absence of easy-to-use visualization support that motivated the work presented here, and drove the choice of familiar and easily-comprehended visualizations including summary tables, bar charts, pie charts and time-series plots. The automated ingest and report generation of textual browsing and mail histories, and the presentation of the data from a forensic perspective using accessible visualizations, will free the analyst from the tedious job of performing this manually. The knowledge gained from these visualizations can be used to guide further in-depth examination of specific sections of the history using tools designed for detailed analysis such as EnCase or FTK. Alternately, this knowledge may also allow the analyst to disregard any sections he deemed unimportant in the context of the investigation.

The analysis of a user's online history is only one portion of any investigation, and with people conducting an increasing amount of their lives online with the main tools being a web browser and mail client, it is sure to become a central element. Automating the processing of potentially useful forensic data, and presenting the analyst with a summary of this information in a concise graphical format, will almost certainly contribute to his ability to more effectively focus his attention and apply his expertise where it is needed.

## 6.2 Future Work

There is much work still to be done in this area. Support for browser history and mailbox file formats was limited to Mozilla Firefox v3 and `mbox` in this set of tools. This toolset was designed to be modular and extensible, so the ability to integrate support for additional file formats by inserting modules into the toolset framework should be relatively straightforward. The current iteration of this utility supports the features necessary to ingest, parse, supplement, generate a report and extract details of certain attributes, but it could be greatly expanded.

Extraction of additional details based on any field in the application database, and on-demand generation of individual visualizations are natural progressions. Finally, this toolset would benefit from a GUI framework, presenting the analyst with a single centralized interface from which to both manipulate the information, and view the resulting visualizations in a dynamic manner, while still maintaining the ability to conduct batch operations on the commandline.

There is plenty of room to add to the visualizations presented in the summary report. Commonly-used summary-style visualizations like bar charts and pie charts were used due to the static nature of the report and their ability to convey significant information in a limited amount of space. More elaborate browsing and communication graphs could be used to represent the information if a GUI interface existed, and the occlusion that results when graphing thousands of entries could be reduced through dynamic filtering. The timeline visualization is very effective in graphically representing navigation and communication over time, but can quickly grow to unmanageable sizes when the periods are long. A GUI would once again allow filtering and panning so that the entire timeline didn't have to be generated and viewed at once.

The metadata that is derived to supplement the existing histories could also be expanded. Currently, the category resolution based on the ODP data is sufficient, and the ability to control this information by providing a built-in URL-to-category mapping provides the analyst with flexibility and customizability. The interface to this built-in database could be integrated with the rest of the toolset so that manipulation of the categories could be performed dynamically while viewing the report. Additionally, the ODP data and GeoIP data will each need to be updated independently to stay current, offering another set of functions that could be integrated. Finally, the search query extraction feature is limited to the large search engines and some of their specialized services. Support for additional search engines and query extraction could be added by expanding the module that performs this function. Finally, additional data could be extracted from the browser history. In addition to the navigation history, Mozilla Firefox maintains separate tables for the information input into forms, downloaded files and cookies. The extraction and integration of this information into the existing navigation history record and its visualization could offer an even richer view into the user's behavior.

The visualizations presented for mail communication were of a similar summary nature as those presented for web browsing, but there is potential to create visualizations that are explicitly designed to capture the multi-directional nature of mail exchanges. The top communicating pairs of addresses could be calculated and represented in a chart similar to the top senders and

recipients bar charts. This statistic, which could be expanded to any relatively small group of addresses, would convey the communication between parties, and offer the same insight as the top senders and recipients with the added dimension of creating connections between addresses. Subjects are currently parsed and stored in the application database, but they are not included in any visualization. Each subject could be tokenized and the most frequently occurring tokens displayed in a bar chart. The message counts over time bar chart and time-series plot could be enhanced to show the proportion of sent and received messages for each day, highlighting anomalous days where the proportion was drastically different than others. This same technique could be applied to the top senders, recipients and domains bar charts by rendering them as stacked bars with each section representing the proportion sent or received. The analyst could then more easily see the communication habits of each address without requiring any additional space in the visualization.

The direction of communication between addresses and the corresponding message volume could be visualized together, further condensing the amount of information represented. This could be done using a network graph with two edges between each node engaging in bi-directional communication. The thickness or length of each edge could be keyed to the amount of message traffic between the two parties, allowing the analyst to quickly see not only who was communicating back-and-forth, but also who was responsible for the bulk of that communication. This idea could be expanded even further by identifying the address in a pair or group of communicating addresses that is most frequently the initiating party in message exchanges, and keying the node color to that attribute. Multiple attributes could be aggregated into a single metric, for example high-volume senders who frequently initiate threads but rarely receive mail could be represented with a different node shape or size than those that rarely initiate threads and mainly receive mail. These features would allow the analyst to quickly pinpoint these addresses in a large network graph, and may allow him to more easily deduce the nature of the relationship between individuals involved in an exchange.

The amount of information extracted from each message could be expanded to include content that is not found in the header. Currently only message headers are parsed, but there is at least one case in which there is information in the body of the message containing addressing information. When a message is forwarded, the header will contain the source address of the initiating party and a number of recipient addresses, but a minimal version of the header of the forwarded message will be inserted in the body of the message. The manner in which this header is displayed is dependent on the mail client, and therefore requires parsing the mail body and knowledge of common client behavior with regards to this function.

There are numerous additions and enhancements that can be made to this work, some of which were presented above. The overarching goal of this set of tools, and any future work, should be in increasing the analyst's knowledge of the underlying data while simultaneously minimizing his effort.

THIS PAGE INTENTIONALLY LEFT BLANK

# List of References


[1] USC Annenberg School for Communication, "World Internet Project Report," Center for the Digital Future, Tech. Rep., April 2009, last accessed 03/2009. [Online]. Available: http://www.digitalcenter.org/

[2] J. Blackbird, S. Entwisle, M. K. Low, D. McKinney, and C. Wueest, "Symantec Inc. Internet Security Threat Report Volume XIII: April, 2008," Symantec Inc., Tech. Rep., 2008, last accessed 02/2009. [Online]. Available: http://www.symantec.com/business/theme.jsp?themeid=threatreport

[3] J. Blackbird, S. Entwisle, M. K. Low, D. McKinney, and C. Wueest, "Symantec Inc. Government Internet Security Threat Report Volume XIII: April, 2008," Symantec Inc., Tech. Rep., 2008, last accessed 02/2009. [Online]. Available: http://www.symantec.com/business/theme.jsp?themeid=threatreport

[4] Sophos, Inc., "Sophos Security Threat Report 2009," Sophos, Inc., Tech. Rep., 2008, last accessed 05/2009. [Online]. Available: http://www.sophos.com/securityreport2009

[5] Computer Security Institute, "CSI Computer Crime and Security Survey 2008," Computer Security Institute (CSI), Tech. Rep., 2008, last accessed 02/2009. [Online]. Available: http://www.gocsi.com/forms/csi_survey.jhtml

[6] SANS, "SANS Top-20 2007 Security Risks (2007 Annual Update)," SANS, Tech. Rep., Nov. 2007, last accessed 02/2009. [Online]. Available: http://www.sans.org/top20/

[7] C. Beaumont, "Paul McCartney website hacked by cybercriminals," Apr. 2009, last accessed 05/2009. [Online]. Available: http://www.telegraph.co.uk/scienceandtechnology/technology/technologynews/5131987/Paul-McCartney-website-hacked-by-cybercriminals.html

[8] US-CERT, DHS, and SS, "The Insider Threat Study: Illicit Cyber Activity in the Government," US-CERT, DHS, Secret Service, Tech. Rep., 2008, last accessed 02/2009. [Online]. Available: http://www.cert.org/archive/pdf/insiderthreat_gov2008.pdf

[9] Privacy Rights Clearinghouse, "A Chronology of Data Breaches," Dec. 2008, last accessed 03/2009. [Online]. Available: http://www.privacyrights.org/ar/ChronDataBreaches.htm

[10] Cisco Systems, Inc., "Data Leakage Worldwide: Common Risks and Mistakes Employees Make," Cisco Systems, Inc., Tech. Rep., 2008, last accessed 03/2009. [Online]. Available: http://www.cisco.com/en/US/solutions/collateral/ns170/ns896/ns895/white_paper_c11-499060.html

[11] J. White, "Navy Finds Data on Thousands of Sailors on Web Site," *Washington Post*, June 2006, last accessed 04/2009. [Online]. Available: http://www.washingtonpost.com/wp-dyn/content/article/2006/06/23/AR2006062301493.html

[12] N. Gohring, "Security Oversight May Have Enabled Countrywide Breach," *Washington Post*, Aug. 2008, last accessed 03/2009. [Online]. Available: http://www.washingtonpost.com/wp-dyn/content/article/2008/08/04/AR2008080401886.html

[13] K. J. Jones, "Forensic Analysis of Internet Explorer Activity Files," Mar. 2003, last accessed 02/2009. [Online]. Available: www.foundstone.com/us/pdf/wp_index_dat.pdf

[14] Mozilla Developer Center, "Mork Why," Sep. 2007, last accessed 02/2009. [Online]. Available: https://developer.mozilla.org/en/Mork_Why

[15] Mozilla Developer Center, "Mork What Is It," Sep. 2007, last accessed 02/2009. [Online]. Available: https://developer.mozilla.org/en/Mork_What_Is_It

[16] Mozilla Developer Center, "Places," Aug. 2008, last accessed 02/2009. [Online]. Available: https://developer.mozilla.org/en/Places

[17] Mozilla Developer Center, "Storage," Oct. 2008, last accessed 02/2009. [Online]. Available: https://developer.mozilla.org/en/Storage

[18] mozillaZine, "Profile folder - Firefox," March 2009, last accessed 06/2009. [Online]. Available: http://kb.mozillazine.org/Profile_folder_-_Firefox

[19] Network Working Group, "Internet Message Format," IETF, Tech. Rep., Apr. 2001, last accessed 03/2009. [Online]. Available: http://tools.ietf.org/html/rfc2822

[20] D. J. Bernstein, *Qmail Manual*, 1st ed., qmail.org, June 1998, last accessed 03/2009. [Online]. Available: http://www.qmail.org/qmail-manual-html/man5/mbox.html

[21] P. S. Foundation, "Python mailbox module," 2008, last accessed 03/2009. [Online]. Available: http://docs.python.org/library/mailbox.html

[22] D. Coppit, *Perl module Mail::Mbox::MessageParser::Perl*, last accessed 03/2009. [Online]. Available: http://search.cpan.org/~dcoppit/Mail-Mbox-MessageParser-1.5000/lib/Mail/Mbox/MessageParser/Perl.pm

[23] D. J. Bernstein, *maildir manual*, last accessed 03/2009. [Online]. Available: http://www.qmail.org/man/man5/maildir.html

[24] D. J. Bernstein. Using maildir format. Last accessed 03/2009. [Online]. Available: http://cr.yp.to/proto/maildir.html

[25] M. Overmeer, *Perl module Mail::Box::Maildir*, last accessed 03/2009. [Online]. Available: http://search.cpan.org/~markov/Mail-Box-2.088/lib/Mail/Box/Maildir.pod

[26] J. Metz, "libpff," Mar. 2009, last accessed 05/2009. [Online]. Available: http://sourceforge.net/projects/libpff/

[27] C. Byington, "libpst," Mar. 2009, last accessed 05/2009. [Online]. Available: http://hg.five-ten-sg.com/libpst/

[28] J. Metz, "Personal Folder File (PFF) file format specification," Hoffmann Investigations, Tech. Rep., Mar. 2009, last accessed 05/2009. [Online]. Available: http://downloads.sourceforge.net/libpff/Personal_Folder_File_format.pdf

[29] J. Metz, "Message API (MAPI) definitions," Hoffmann Investigations, Tech. Rep., 2009, last accessed 05/2009. [Online]. Available: http://downloads.sourceforge.net/libpff/MAPI_definitions.pdf

[30] S. K. Card, J. Mackinlay, and B. Shneiderman, *Readings in Information Visualization: Using Vision to Think*. Morgan Kaufmann, January 1999. [Online]. Available: http://www.amazon.com/dp/1558605339/

[31] C. Wickens, D. Sandry, and M. Vidulich, "Compatibility and resource competition between modalities of input, central processing, and output," *Human Factors*, vol. 25, no. 2, pp. 227–248, 1983.

[32] C. G. Healey, K. S. Booth, and J. T. Enns, "High-speed visual estimation using preattentive processing," *ACM Trans. Comput.-Hum. Interact.*, vol. 3, no. 2, pp. 107–135, 1996.

[33] AccessData Corp., *Forensic Toolkit 2.1.0 Release Notes*, 2nd ed., Access Data, 2008, last accessed 02/2009. [Online]. Available: http://ftk21.accessdata.com/ftk2_readme.pdf

[34] Guidance Software, Inc., *EnCase Forensic Detailed Product Description*, 2nd ed., Access Data, 2008, last accessed 02/2009. [Online]. Available: http://www.guidancesoftware.com/downloads/DetailedProductDescription.pdf

[35] Digital Investigations Group, "CacheBack 2.0 Internet Cache and History Analysis," Oct. 2008, windows. [Online]. Available: http://www.cacheback.ca/

[36] N. Sofer, "IEHistoryView," 2008, last accessed 02/2009. [Online]. Available: http://www.nirsoft.net/utils/iehv.html

[37] forensic software.co.uk, "FoxAnalysis," Apr. 2008, last accessed 02/2009. [Online]. Available: http://forensic-software.co.uk/foxanalysis.aspx

[38] firefoxforensics.com, "Firefox 3 Extractor," June 2008, last accessed 02/2009. [Online]. Available: http://www.firefoxforensics.com/

[39] Passcape Software, "Mozilla/Thunderbird/Firefox Password Recovery URL History Viewer," Oct. 2008, last accessed 02/2009. [Online]. Available: http://www.passcape.com/firefox_url_history.htm

[40] N. Sofer, "MozillaHistoryView," 2007, last accessed 02/2009. [Online]. Available: http://www.nirsoft.net/utils/mozilla_history_view.html

[41] Digital Detective, Inc., "NetAnalysis," 2007, last accessed 02/2009. [Online]. Available: http://www.digital-detective.co.uk/netanalysis.asp

[42] A. Stuart, "Index.Dat Viewer and Zapper," June 2007, last accessed 02/2009. [Online]. Available: http://www.scanraid.com/indexdat.htm

[43] X-Ways Software Technology AG, "X-Ways Trace," 2007, last accessed 02/2009. [Online]. Available: http://www.x-ways.net/trace/

[44] M. Machor, "FireFox Forensics," 2007, last accessed 02/2009. [Online]. Available: http://www.machor-software.com/firefox_forensics

[45] Cleanersoft, "IEHistory," 2006, last accessed 02/2009. [Online]. Available: http://www.cleanersoft.com/iehistory/iehistory.htm

[46] Systenance Software Solutions, "Index.Dat Analyzer," 2006, last accessed 02/2009. [Online]. Available: http://www.systenance.com/indexdat.php

[47] Mozilla, "Enhanced History Manager," Nov. 2006, last accessed 02/2009. [Online]. Available: https://addons.mozilla.org/en-US/firefox/addon/420

[48] lbtechservices.com, "Browser History Viewer," June 2006, last accessed 02/2009. [Online]. Available: http://www.lbtechservices.com/software/oss/bhv/

[49] R. Cliff, "Web Historian," July 2005, last accessed 02/2009. [Online]. Available: http://mandiant.invisionzone.com/index.php?s=e677a00e28c7f80d77e4b443378484a9\&showtopic=8

[50] K. J. Jones, "Pasco," May 2004, last accessed 02/2009. [Online]. Available: http://www.foundstone.com/us/resources/proddesc/pasco.htm

[51] J. Zawinski, Anonymous, and J. Post, "mork.pl - a Mozilla v2 URL History File Parser," Mar. 2004, last accessed 02/2009. [Online]. Available: http://www.jwz.org/hacks/mork.pl

[52] G. Black, "Timeline Analysis," Computer and Enterprise Investigations Conference, Tech. Rep., 2007.

[53] S. P. Reiss and G. Eddon, "Visualizing what people are doing on the Web," pp. 305–307, 2005, iD: 1.

[54] S. G. Eick, "Visualizing online activity," *Commun.ACM*, vol. 44, no. 8, pp. 45–50, 2001. [Online]. Available: http://doi.acm.org/10.1145/381641.381710

[55] SAP Business Software. (2009) Business Objects Tools for Advanced Visualization. Last accessed 05/2009. [Online]. Available: http://www.sap.com/solutions/sapbusinessobjects/large/intelligenceplatform/bi/dashboard-visualization/advanced-visualization/index.epx

[56] T. Munzner, "H3: Laying out large directed graphs in 3D hyperbolic space," in *IEEE Symposium on Information Visualization, 1997. Proceedings.*, 1997, pp. 2–10, last accessed 05/2009. [Online]. Available: http://www-graphics.stanford.edu/papers/h3/

[57] J. Cugini and J. Scholtz, "VISVIP: 3D visualization of paths through web sites," in *Database and Expert Systems Applications, 1999. Proceedings. Tenth International Workshop on*, 1999, pp. 259–263, iD: 1.

[58] B. Fry and J. Maeda, "Organic information design," 2000, last accessed 05/2009. [Online]. Available: http://benfry.com/organic/

[59] AccessData Corp., *Forensic Toolkit 2.0.2 Readme*, 2nd ed., Access Data, 2008, last accessed 02/2009. [Online]. Available: http://www.accessdata.com/downloads/media/ftk202_readme.pdf

[60] M. Cohen, "PyFlag–An advanced network forensic framework," *Digital Investigation*, vol. 5, pp. 112–120, 2008.

[61] X. Fu, S. Hong, N. Nikolov, X. Shen, Y. Wu, K. Xu, and N. Australia, "Visualization and analysis of email networks," in *Visualization, 2007. APVIS'07. 2007 6th International Asia-Pacific Symposium on*, 2007, pp. 1–8.

[62] V. Krebs. (2009) Orgnet InFlow Social Network Analysis Software. Last accessed 06/2009. [Online]. Available: http://www.orgnet.com/inflow3.html

[63] W. Li, S. Hershkop, and S. Stolfo, "Email archive analysis through graphical visualization," in *Proceedings of the 2004 ACM workshop on Visualization and data mining for computer security*. ACM New York, NY, USA, 2004, pp. 128–132.

[64] J. Heer, "Exploring Enron: Visualizing ANLP results," in *Applied Natural Language Processing*, 2004.

[65] B. Kerr, "Thread arcs: An email thread visualization," in *Proceedings of the 2003 IEEE Symposium on Information Visualization*, 2003, p. 27.

[66] A. Perer and M. Smith, "Contrasting Portraits of Email Practices: Visual approaches to reflection and analysis," in *Proceedings of the working conference on Advanced visual interfaces*. ACM New York, NY, USA, 2006, pp. 389–395.

[67] IBM. IBM ReMail Visualizations. Last accessed 06/2009. [Online]. Available: http://www.research.ibm.com/remail/visualizations.html

[68] D. Stone, C. Jarrett, M. Woodroffe, and S. Minocha, *User Interface Design and Evaluation*. Morgan Kaufmann Publishers Inc. San Francisco, CA, USA, 2005.

[69] International Organization for Standardization, "ISO 9241-11," International Organization for Standardization, Tech. Rep., Sep. 2008.

[70] P. S. Foundation, "Python Documentation," Dec. 2008, last accessed 05/2009. [Online]. Available: http://docs.python.org/

[71] World Wide Web Consortium (W3C), "Scalable Vector Graphics (SVG) — XML Graphics for the Web," Apr. 2009, last accessed 05/2009. [Online]. Available: http://www.w3.org/Graphics/SVG/

[72] J. Ivar, G. Booch, and J. Rumbaugh, "The Unified Software Development Process," *Reading: Addison-Wesley*, 1999.

[73] C. Larman, *Applying UML and Patterns*. Prentice Hall PTR Upper Saddle River, NJ, USA, August 2007.

[74] E. Casey, *Digital Evidence and Computer Crime: Forensic Science, Computers, and the Internet*. Academic Press, Inc. Orlando, FL, USA, 2000.

[75] H. Weinreich, H. Obendorf, E. Herder, and M. Mayer, "Not quite the average: An empirical study of Web use," *ACM Trans.Web*, vol. 2, no. 1, pp. 1–31, 2008, last accessed 02/2009. [Online]. Available: http://doi.acm.org/10.1145/1326561.1326566

[76] MaxMind, Inc., "MaxMind GeoIP," Dec. 2008, last accessed 06/2009. [Online]. Available: http://www.maxmind.com/app/geolitecountry

[77] Open Directory Project, "Open Directory Project," Dec. 2008, last accessed 06/2009. [Online]. Available: http://www.dmoz.org/

[78] M.-Y. Kan. (2008) MeURLin: URL-based classification of web pages. Last accessed 05/2009. [Online]. Available: http://wing.comp.nus.edu.sg/meurlin/

[79] W3C, "Resource Description Format," Feb. 2004, last accessed 05/2009. [Online]. Available: http://www.w3.org/TR/rdf-primer/

[80] Thumbshots.com, "Thumbshots," May 2009, last accessed 06/2009. [Online]. Available: http://www.thumbshots.org/

[81] P. S. Foundation, "Python Documentation – sqlite3," 2009, last accessed 05/2009. [Online]. Available: http://docs.python.org/library/sqlite3.html

[82] P. S. Foundation, "Python Database API Specification v2.0," 2008, last accessed 05/2009. [Online]. Available: http://www.python.org/dev/peps/pep-0249/

[83] J. C. Dürsteler, "InfoVis Diagram," Feb. 2007, last accessed 05/2009. [Online]. Available: http://www.infovis.net/printMag.php?num=187\&lang=2

[84] D. Bock, P. Velleman, and R. De Veaux, *Intro Stats*. Boston, MA: Pearson Education, 2009.

[85] R. Marty, *Applied Security Visualization*, 1st ed. Addison-Wesley Professional, August 2008.

[86] G. Conti, *Security Data Visualization*. No Starch Press, 2007.

[87] J. Mackinlay, "Automating the design of graphical presentations of relational information," *ACM Trans. Graph.*, vol. 5, no. 2, pp. 110–141, 1986.

[88] S. Card and J. Mackinlay, "The structure of the information visualization design space," in *IEEE Symposium on Information Visualization, 1997. Proceedings.*, Oct. 1997, pp. 92–99.

[89] J. C. Dürsteler, "Diagrams for Visualization," Jan. 2007, last accessed 05/2009. [Online]. Available: http://www.infovis.net/printMag.php?num=186\&lang=2

[90] E. Tufte, *The Visual Display of Quantitative Information*. Graphics Press, 1983.

[91] Mozilla Developer Center, "The Places frecency Algorithm," June 2008, last accessed 06/2009. [Online]. Available: https://developer.mozilla.org/en/The_Places_frecency_algorithm

[92] W. Feibel, *The Encyclopedia of Networking*. Sybex INc, 1996.

[93] Mozilla Developer Center, "PRTime," June 2008, last accessed 03/2009. [Online]. Available: https://developer.mozilla.org/en/PRTime

# APPENDIX A:
# Glossary

**Domain Name** Abstraction allowing the mapping of a human-readable name to an IP address.

**Frecency** Mozilla frequency/recency metric. A score given to each unique URI in Places, encompassing bookmarks, history and tags. This score is determined by the amount of revisitation, the type of those visits, how recent they were, and whether the URI was bookmarked or tagged.[91]

**Geolocation Database** Database containing mapping of IP address to geographic location.

**GDI** Graphics Device Interface; the portion of the Windows operating system that provides two-dimensional vector graphics, imaging, and typography.

**MTA** Mail Transfer Agent; the component of a message handling system responsible for storing or forwarding messages to another MTA, to a Mail User Agent (MUA), or to another authorized recipient.[92]

**MUA** Mail User Agent; the application process that provides access for a human user to a mailbox.[92]

**ODP** Open Directory Project; the largest most comprehensive human-edited directory of the Web, serving as the backend data provider to directories such as Google Directory and AOL[77].

**Referring site** The URL for the previous website in the navigation path from which the current site was navigated to. Also known simply as *referrer*.

**TLD** Top-Level Domain. The group of letters following the dot after the domain name. Common TLDs include: com, org, net, edu, gov, mil, uk, fr, ca, jp, gr.

**Time (epoch seconds)** A floating-point number representing the number of seconds since the epoch, midnight (00:00:00) 1 January 1970 Coordinated Universal Time (UTC). A time after the epoch has a positive value, and a time before the epoch has a negative value.

**Time (PRTime)** A 64-bit integer representing the number of microseconds since the epoch, midnight (00:00:00) 1 January 1970 Coordinated Universal Time (UTC).[93]

**URI** Uniform Resource Identifier. A compact sequence of characters that identifies an abstract or physical resource. A URL is a more specific instance or a URI.[92]

**URL** Uniform Resource Locator. Provides a means of identifying a document on the Internet. For example `http://www.mozilla.org/`. [92]

# APPENDIX B:
# Utility Console Output

The content below is the console debug output produced by this toolset during the ingest, category resolution and summary report generation phases demonstrated in Chapter 5. Due to the large volume of information that is generated, the output is truncated where appropriate, and annotated with `output truncated`. Detailed discussion of the demonstration is in Section 5.1.

## B.1 Ingest

The information shown below is output by this utility during the ingest phase when executed with `--ingest --type web --file places.sqlite`.

```
$ ./controller.py --ingest --type web --file data/places.sqlite

Parsing web file data/places.sqlite
3 [Sun Aug  3 13:56:15 2008] www.mozilla.com
5 [Sun Aug  3 13:56:20 2008] addons.mozilla.org
1 [Sun Aug  3 13:56:21 2008] es-ar.www.mozilla.com
2 [Sun Aug  3 13:56:21 2008] es-ar.www.mozilla.com
6 [Sun Aug  3 13:56:22 2008] addons.mozilla.org
4 [Sun Aug  3 13:56:27 2008] addons.mozilla.org
7 [Sun Aug  3 13:56:35 2008] addons.mozilla.org
8 [Sun Aug  3 13:56:41 2008] addons.mozilla.org
9 [Sun Aug  3 13:58:55 2008] addons.mozilla.org
10 [Sun Aug  3 13:58:55 2008] addons.mozilla.org
11 [Sun Aug  3 13:59:11 2008] addons.mozilla.org
12 [Sun Aug  3 13:59:32 2008] addons.mozilla.org
13 [Sun Aug  3 14:03:21 2008] erofoto.ucoz.ru
14 [Sun Aug  3 14:03:21 2008] erofoto.ucoz.ru
15 [Sun Aug  3 14:03:21 2008] erofoto.ucoz.ru
16 [Sun Aug  3 14:03:21 2008] depositfiles.com
17 [Sun Aug  3 14:03:21 2008] depositfiles.com
18 [Sun Aug  3 14:03:21 2008] rapidshare.com
19 [Sun Aug  3 14:03:21 2008] erofoto.ucoz.ru
20 [Sun Aug  3 14:03:21 2008] erofoto.ucoz.ru
```

```
21 [Sun Aug  3 14:03:21 2008] erofoto.ucoz.ru
28 [Sun Aug  3 14:03:40 2008] erofoto.ucoz.ru
35 [Sun Aug  3 14:04:53 2008] nguest145.depositfiles.com
36 [Sun Aug  3 14:07:23 2008] nguest110.depositfiles.com
38 [Sun Aug  3 14:27:13 2008] www.google.com.mx
39 [Sun Aug  3 14:27:23 2008] www.google.com.mx
Extracting search query from:
hl=es&q=Sony+Sound+Forge+v9+Keygen&meta=&btnG=Buscar+con+Google
41 [Sun Aug  3 14:27:28 2008] www.google.com.mx

... output truncated ...

32435 [Thu Nov 13 21:47:29 2008] support.microsoft.com
32436 [Thu Nov 13 22:31:55 2008] protizer.ru
32437 [Thu Nov 13 22:31:55 2008] protizer.ru
32438 [Thu Nov 13 22:31:55 2008] protizer.ru
32439 [Thu Nov 13 22:31:55 2008] protizer.ru
32440 [Thu Nov 13 22:31:59 2008] protizer.ru
32441 [Thu Nov 13 22:31:59 2008] protizer.ru
32442 [Thu Nov 13 22:32:12 2008] www.gmodules.com
32443 [Thu Nov 13 22:32:12 2008] www.gmodules.com
32444 [Thu Nov 13 22:32:12 2008] www.gmodules.com
32445 [Thu Nov 13 22:32:12 2008] www.gmodules.com
32446 [Thu Nov 13 22:32:12 2008] www.gmodules.com
32447 [Thu Nov 13 22:32:18 2008] www.gmodules.com
32448 [Thu Nov 13 22:32:48 2008] protizer.ru
32449 [Thu Nov 13 22:32:48 2008] protizer.ru
32450 [Thu Nov 13 22:33:09 2008] www.gmodules.com
32451 [Thu Nov 13 22:33:09 2008] www.gmodules.com
32452 [Thu Nov 13 22:34:47 2008] protizer.ru
32453 [Thu Nov 13 22:34:50 2008] protizer.ru
32454 [Thu Nov 13 22:34:50 2008] intimchik.net
32455 [Thu Nov 13 22:34:57 2008] protizer.ru
32456 [Thu Nov 13 22:35:02 2008] protizer.ru
32459 [Thu Nov 13 22:35:06 2008] sexadrom.net
32457 [Thu Nov 13 22:35:07 2008] protizer.ru
32460 [Thu Nov 13 22:35:08 2008] kino-x.biz
32458 [Thu Nov 13 22:35:16 2008] adulttraffic.ru
32461 [Thu Nov 13 22:35:16 2008] protizer.ru
32465 [Thu Nov 13 22:35:18 2008] x-rolik.net
32462 [Thu Nov 13 22:35:20 2008] protizer.ru
```

```
32464 [Thu Nov 13 22:35:25 2008] sexadrom.net
32463 [Thu Nov 13 22:35:26 2008] adulttraffic.ru
32466 [Thu Nov 13 22:35:26 2008] protizer.ru
32468 [Thu Nov 13 22:35:26 2008] xxx-student.biz
32467 [Thu Nov 13 22:35:33 2008] sexadrom.net
32469 [Thu Nov 13 22:35:57 2008] rapidshare.com
32470 [Thu Nov 13 22:38:39 2008] rs283.rapidshare.com
32471 [Thu Nov 13 22:38:39 2008] rs283gc2.rapidshare.com

Successfuly parsed 18,629 records

Elapsed time:         32 min 45 sec
```

## B.2 Category Resolution

The information shown below is output by this utility during the category resolution when executed with `--resolve-categories` as seen in the first line below.

```
$ ./controller.py --resolve-categories

Loading web records beginning Sun Aug  3 13:56:15 2008

www.mozilla.com (2 visits)
Domain name: "mozilla.com"
=> Top/Computers/Open_Source/Organizations/Mozilla_Foundation/News_and_Media

addons.mozilla.org (1 visits)
Domain name: "mozilla.org"
=> Top/Computers/Open_Source/Organizations/Mozilla_Foundation

erofoto.ucoz.ru (1 visits)
Domain name: "ucoz.ru"
=> Top/World/Russian/Computers/Internet/Design_and_Site_Development/Hosting/Free

depositfiles.com (2 visits)
Domain name: "depositfiles.com"
=> Top/World/Russian/Computers/Internet/Storage/Free

www.google.com.mx (2 visits)
```

```
Domain name: "google.com.mx"
=> Computers/Internet/Searching/Search_Engines/Google

www.taringa.net (1 visits)
Domain name: "taringa.net"
Unknown URL; searching on "http://www.taringa.net/"
Searching again with "www" removed: "http://taringa.net/"
Searching again without trailing "/": "http://www.taringa.net"
Searching again with wildcard as "http://www.taringa.net/"
Searching again with wildcard: "%taringa.net%"
Category not found for www.taringa.net

www.megaupload.com (2 visits)
Domain name: "megaupload.com"
=> Top/Computers/Internet/On_the_Web/Web_Applications/Storage/Free

www.mcafee.com (11 visits)
Domain name: "mcafee.com"
=> Top/Computers/Security/Malicious_Software/Viruses/Products

mx.altavista.com (4 visits)
Domain name: "altavista.com"
=> Computers/Internet/Searching/Search Engines

av.rds.yahoo.com (1 visits)
Domain name: "yahoo.com"
=> Computers/Internet/Searching/Search_Engines/Yahoo

www.hotmail.com (111 visits)
Domain name: "hotmail.com"
=> Top/Computers/Internet/E-mail/Free/Web-Based/H/Hotmail

www.megaupload.com (1 visits)
Domain name: "megaupload.com"
=> Top/Computers/Internet/On_the_Web/Web_Applications/Storage/Free

www60.megaupload.com (1 visits)
Domain name: "megaupload.com"
=> Top/Computers/Internet/On_the_Web/Web_Applications/Storage/Free

www.hi5.com (24 visits)
```

```
Domain name: "hi5.com"
=> Top/Computers/Internet/On_the_Web/Online_Communities/Social_Networking

foro.portalhacker.net (1 visits)
Domain name: "portalhacker.net"
=> Top/World/Espaol/Computadoras/Hacking

mx.youtube.com (4 visits)
Domain name: "youtube.com"
=> Top/Arts/Entertainment/Online_Media

... output trucated ...

Successfuly resolved 410 unique categories

Elapsed time:         143 min 37 sec
```

## B.3 Report Generation

The information shown below is output by this utility during the creation of a summary report for the specified period. The tool was executed with `--create-summaries --type web --date-begin 2008-08-03 --date-end 2008-11-13` as seen in the first line below.

```
$ ./controller.py --create-summaries --type web \
      --date-begin 2008-08-03 --date-end 2008-11-13

Creating web report for Sun Aug  3 00:00:00 2008 - Thu Nov 13 23:59:59 2008

SELECT * FROM web
WHERE visit_date BETWEEN 1217746800.0 AND 1226649599.0
ORDER BY visit_date ASC

Loading web records for Sun Aug  3 00:00:00 2008 - Thu Nov 13 23:59:59 2008

Generating lists from history

SELECT DISTINCT id, url, COUNT() FROM web
WHERE visit_date BETWEEN 1217746800.0 AND 1226649599.0
AND url <> "" AND visit_count <> 0
```

```
GROUP BY url ORDER BY COUNT() DESC LIMIT 20

... output trucated ...

SELECT DISTINCT id, country_code, COUNT() FROM web
WHERE visit_date BETWEEN 1217746800.0 AND 1226649599.0
AND country_code <> "" AND visit_count == 0
GROUP BY country_code ORDER BY COUNT() DESC

Generating daily counts for Sun Aug  3 00:00:00 2008 - Thu Nov 13 23:59:59 2008

Calculating visits per day:
Sun Aug  3 00:00:00 2008
Mon Aug  4 00:00:00 2008
Tue Aug  5 00:00:00 2008

... output trucated ...

Tue Nov 11 23:00:00 2008
Wed Nov 12 23:00:00 2008
Thu Nov 13 23:00:00 2008

Total visits: 18,502

Calculating unique counts for visits
category: 411
url: 15,545
search_query: 506
url_base: 2,533
country_code: 45
tld: 53

Calculating unique counts for non-visits
url: 97
url_base: 30
country_code: 9
tld: 5

Generating formatted base URLs for tables
Generating formatted full URLs for tables
Generating summary tables
```

```
Generating visualizations for web history:
- bar chart for base URLs
- bar chart for base URLs by frecency
- bar chart for full URLs
- bar chart for full URLs by frecency
- pie chart for visit type counts
- pie chart for scheme counts
- pie chart for TLD counts
- pie chart for country counts
- bar chart for search query count
- bar chart for category counts
- time-series plot for counts over time

Elapsed time:         1 min 53 sec
```

THIS PAGE INTENTIONALLY LEFT BLANK

# APPENDIX C:
# Detail Output

The content below is the output produced by this toolset during the various detail extraction phases of the analysis session in Chapter 5. Due to the large volume of information that is presented in some cases, the output is truncated and annotated with `output truncated`. The embedded URLs, completed downloads, typed URLs and search queries are included, and are discussed in detail in Section 5.3.

## Visited Embedded URLs

This is the complete set of embedded URLs visited by the user, and is the result of executing this tool with the `--get-records --type embedded --visited true --field url` commandline argument.

```
http://www.google.com/ig/modules/translatemypage.xml&up_source_language=ru
  &w=160&h=60&source=http://erofoto.ucoz.ru/news/10-0-1
http://www.google.com/ig/modules/translatemypage.xml&up_source_language=ru
  &w=160&h=60&source=http://erofoto.ucoz.ru/news/2-0-1
http://www.google.com/ig/modules/translatemypage.xml&up_source_language=ru
  &w=160&h=60&source=http://erofoto.ucoz.ru/news/2008-09-20-1297
http://www.google.com/ig/modules/translatemypage.xml&up_source_language=ru
  &w=160&h=60&source=http://erofoto.ucoz.ru/news/2008-09-20-1295
http://www.google.com/ig/modules/translatemypage.xml&up_source_language=ru
  &w=160&h=60&source=http://erofoto.ucoz.ru/news/2-0-22-lhoKEL
http://www.google.com/ig/modules/translatemypage.xml&up_source_language=ru
  &w=160&h=60&source=http://erofoto.ucoz.ru/news/2008-04-06-855
http://www.google.com/ig/modules/translatemypage.xml&up_source_language=ru
  &w=160&h=60&source=http://erofoto.ucoz.ru/news/1-0-33
http://www.google.com/ig/modules/translatemypage.xml&up_source_language=ru
  &w=160&h=60&source=http://erofoto.ucoz.ru/news/1-0-33
http://www.google.com/ig/modules/translatemypage.xml&up_source_language=ru
  &w=160&h=60&source=http://erofoto.ucoz.ru/news/2008-10-31-1515
http://www.google.com/ig/modules/translatemypage.xml&up_source_language=ru
  &w=160&h=60&source=http://erofoto.ucoz.ru/news/2-0-1
http://www.google.com/ig/modules/translatemypage.xml&up_source_language=ru
  &w=160&h=60&source=http://erofoto.ucoz.ru/news/3-0-10
```

```
http://www.google.com/ig/modules/translatemypage.xml&up_source_language=ru
  &w=160&h=60&source=http://erofoto.ucoz.ru/news/2-0-22-lhoKEL
http://www.google.com/ig/modules/translatemypage.xml&up_source_language=ru
  &w=160&h=60&source=http://erofoto.ucoz.ru/news/1-0-33
http://www.google.com/ig/modules/translatemypage.xml&up_source_language=ru
  &w=160&h=60&source=http://erofoto.ucoz.ru/news/2008-04-06-855
http://www.google.com/ig/modules/translatemypage.xml&up_source_language=ru
  &w=160&h=60&source=http://erofoto.ucoz.ru/news/2008-04-13-863
http://www.google.com/ig/modules/translatemypage.xml&up_source_language=ru
  &w=160&h=60&source=http://erofoto.ucoz.ru/news/2008-09-20-1295
http://www.google.com/ig/modules/translatemypage.xml&up_source_language=ru
  &w=160&h=60&source=http://erofoto.ucoz.ru/news/2008-10-31-1515
```

# Downloads

This is a partial listing of the 2,875 downloads completed by the user, and is the result of executing this tool with `--get-records --type download --field url_base --field filename`.

```
65.55.152.121             03c6861f-0501-4e91-8444-34576b9b00bf.xls
65.55.152.121             04354d64-650f-4854-b6f6-900fe199435d.pps
65.55.152.121             07a5873c-9d8c-4d4a-a766-246724e572f9.pps
65.55.152.121             0861468c-aad7-4687-b07d-385c6b26af54.wmv
65.55.152.121             09fcc99d-1c6f-463a-9374-df53f5fc3496.pps
65.55.152.121             0f48cd7f-f6df-40be-abba-ee14ba8343a4.mpe
65.55.152.121             0fe41c2c-1d6e-4937-9720-de607d5b3d04.txt
65.55.152.121             10896d65-7354-4d2e-bc5e-d968ab60e01c.zip
65.55.152.121             11495684-f6cc-47b1-8a13-8c109202ca91.pps
65.55.152.121             1294dedf-3651-4fad-97b1-39d37d829076.pps
65.55.152.121             149e14ec-8721-448a-99df-7ca62364afb9.pps
65.55.152.121             173095f2-22a4-4006-848e-5e8d58fe67a5.pps
65.55.152.121             1ca130ec-7de0-47d0-995c-852b1d073b58.wmv
............              output truncated
65.55.152.121             e13616c4-956b-4770-b816-63c0e8e5d424.pps
65.55.152.121             e1bc6a24-83dd-4c55-ab63-85d7bd7e46d3.wmv
65.55.152.121             e1e0adb2-8ae7-412b-a835-388f89978464.pps
65.55.152.121             e4c3fc21-47b3-4560-8e83-1317b3de23c1.pps
65.55.152.121             e5bd959c-a429-4169-9d42-0c9133da481e.pps
65.55.152.121             e9847d5c-b990-41a1-82d4-14f2b1d67c51.MPG
```

```
65.55.152.121             eb5c7346-ad88-479d-aa54-17ac0c42266d.pps
65.55.152.121             f2fa05d9-d0f5-4e2d-8751-3f7a29bed37f.ppS
65.55.152.121             f52132b5-cc18-48b5-b925-fa78e28f6bcd.pps
65.55.152.121             fb83f5ca-6e0b-4ec5-8a47-8e5110594818.wmv
67.198.206.18             CallofDuty2Patchv1_3.zip
67.198.206.18             CallofDuty2Patchv1_3.zip
america.giga-byte.com     motherboard_driver_audio_microsoft_bus.exe
america.giga-byte.com     motherboard_driver_audio_realtek_azalia.exe
america.giga-byte.com     motherboard_driver_chipset_intel.exe
america.giga-byte.com     motherboard_driver_lan_realtek_8111_vista.exe
ardownload.adobe.com      AdbeRdr90_es_ES.exe
definitions.symantec.com  20080804-003-i32.exe
definitions.symantec.com  20080830-036-i32.exe
definitions.symantec.com  20080908-036-i32.exe
definitions.symantec.com  20080922-003-i32.exe
definitions.symantec.com  20081017-003-i32.exe
definitions.symantec.com  20081031-007-i32.exe
dl.google.com             picasaweb-current-setup.exe
dl3.upload.com.ua         Olivia777.rar
dl5.share-online.biz      Simlock.zip
download.betanews.com     RegSeeker.zip
download.divx.com         DivXInstaller.exe
download.divx.com         DivXUserGuide52.exe
download.microsoft.com    IE7-WindowsXP-x86-esn.exe
download.microsoft.com    SyncToySetupPackage.exe
files.hamachi.cc          HamachiSetup-1.0.3.0-es.exe
files1.subdivx.com        1102.zip
files1.subdivx.com        18262.zip
files1.subdivx.com        22653.zip
.................         output truncated
files3.subdivx.com        128960.zip
files3.subdivx.com        129687.rar
files3.subdivx.com        130632.rar
images.starpulse.com      Katy-Perry-c01.jpg
images.starpulse.com      Katy-Perry-c02.jpg
images.starpulse.com      Katy-Perry-c03.jpg
images.starpulse.com      Katy-Perry-c05.jpg
images.starpulse.com      Katy-Perry-c06.jpg
images.starpulse.com      Katy-Perry-c07.jpg
images.starpulse.com      Katy-Perry-c08.jpg
images.starpulse.com      Katy-Perry-c09.jpg
```

```
.....................       output truncated
images.starpulse.com        Katy-Perry-c200.jpg
images.starpulse.com        Katy-Perry-c201.jpg
images.starpulse.com        Katy-Perry-c202.jpg
images.starpulse.com        Katy-Perry-c203.jpg
images.starpulse.com        Katy-Perry-c204.jpg
.....................       output truncated
guest100.depositfiles.com  J_-_2008-04-20_-_Cayenne_-_Cayenne.rar
guest101.depositfiles.com  a_-_2008-04-14_-_Koty_-_Warm_Sand.part1.rar
guest102.depositfiles.com  D_-_2008-09-01_-_Pela_03E.rar
guest102.depositfiles.com  J_-_2008-09-02_-_Maria_-_CoastwiseE.rar
guest104.depositfiles.com  M_-_2008-04-24_-_Natasha_-_Sensual_Simplicity.rar
guest105.depositfiles.com  F_-_2008-03-24_-_Katalin_-_All_About_Me.rar
guest105.depositfiles.com  AA_-_2007-07-21_-_Sveta_-_Alpine_Meadow.rar
guest106.depositfiles.com  M_-_2008-09-29_-_Alida_-_Pastel.rar
guest106.depositfiles.com  BT_-_2008-04-12_-_Carmen_-_Earthy.part2.rar
guest107.depositfiles.com  M_-_2008-09-02_-_Gabrielle_-_GlitterE.rar
guest107.depositfiles.com  J_-_2008-04-06_-_Michaela_-_Casablanca.part2.rar
guest107.depositfiles.com  M_-_2008-09-02_-_Gabrielle_-_GlitterE.rar
```

# Typed Sites

This is a complete listing of all URLs typed directly into the address bar by the user, and is the result of executing this tool with `--get-records --type typed --field visit_count --field url`.

```
24  http://www.hi5.com/
21  http://192.168.2.1/
15  http://www.subdivx.com/
 9  http://www.zonajobs.com.mx/postulante/micuenta.asp
 6  http://depositfiles.com/es/files/7859768
 4  http://www.myspace.com/
 4  http://www.occmundial.com/
 4  http://www.zeitgeistmovie.com/
 3  http://erofoto.ucoz.ru/news/2007-10-11-47
 3  http://erofoto.ucoz.ru/news/2008-09-23-1345
 3  http://erofoto.ucoz.ru/news/2008-09-28-1400
 3  http://www.femjoy.com/models.php
 3  http://www.hp.com/
```

3 http://www.intel.com/
3 http://www.katyperry.com.mx/
3 http://www.movistar.com.mx/
2 http://85.255.112.102/
2 http://blya.net/
2 http://depositfiles.com/en/files/5223237
2 http://depositfiles.com/en/files/5937977
2 http://depositfiles.com/es/files/2839021
2 http://depositfiles.com/es/files/2849760
2 http://depositfiles.com/es/files/2850051
2 http://depositfiles.com/es/files/5236823
2 http://depositfiles.com/es/files/7963833
2 http://depositfiles.com/es/files/8223390
2 http://depositfiles.com/es/files/8264585
2 http://erofoto.ucoz.ru/
2 http://nude.hu/313872
2 http://rapidshare.com/files/102956582/VANESSA_A__HELIUX.rar
2 http://www.mediafire.com/?wjnyhvixx0b
2 http://www.thevenusproject.com/
1 http://42.com.mx/
1 http://borda.ca/
1 http://divx.com/vod
1 http://es.youtube.com/watch?v=E2UaVcZPsYU
1 http://go.microsoft.com/fwlink/?linkid=63939
1 http://maps.google.com/
1 http://mx.youtube.com/user/xxElAnticristo2007xx
1 http://rapidshare.com/files/133598420/Rasatello.part2.rar
1 http://rapidshark.net/You_And_Me_Together.rar
1 http://winamp.com/
1 http://winmap.com/
1 http://www.acer.com/
1 http://www.acerpanam.com/
1 http://www.cablevision.com.mx/
1 http://www.divx.com/
1 http://www.doxpara.com/
1 http://www.ezbsystems.com/ultraiso
1 http://www.femjoy.com/
1 http://www.google.com/
1 http://www.hirepoint.com/
1 http://www.idealite.com.mx/
1 http://www.java.com/

```
 1  http://www.ligasmx.com/
 1  http://www.livesonyericsson.com/
 1  http://www.metallica.com/
 1  http://www.panoramio.com/
 1  http://www.seznam.cz/
 1  http://www.sonyericsson.com/
 1  http://www.standup2cancer.org/
 1  http://www.teendreams.com/tgp/707/verity1/?nats=NzMyOjk6MQ,0,0,0,
 1  http://www.teendreams.com/tgp/707/verity2/?nats=NzMyOjk6MQ,0,0,0,
 1  http://www.teendreams.com/tgp/707/verity3/?nats=NzMyOjk6MQ,0,0,0,
 1  http://www.teendreams.com/tgp/707/verity5/?nats=NzMyOjk6MQ,0,0,0,
 1  http://www.thezeitgeistmovement.com/
 1  http://www.thinviewer.com/es
 1  http://www.toshiba.com/
 1  http://www.ulises2k.com.ar/
 1  http://www.videolan.org/
 1  http://www.winiso.com/
 1  http://www.workspace.office.live.com/
 1  http://www.youtube.com/watch?v=AhvfCFCfdNk
 1  http://www.zonajobs.com/
 1  https://www.bolsadetrabajotelevisa.com.mx/
```

## Search Queries

This is a partial listing of the 505 search queries that were performed by the user and extracted by this toolset. They were queried from the history database with the `--get-records --type search_queries --field url_base --field search_query` argument.

```
images.google.com.mx | site:www.filmgecko.com becki newton
images.google.com.mx | "oskar schindler""
images.google.com.mx | "virginia balcazar"
images.google.com.mx | acer logo
images.google.com.mx | acteck knox
images.google.com.mx | acteck stutgart
images.google.com.mx | aletta alien
images.google.com.mx | alison carroll
images.google.com.mx | alison playboy
images.google.com.mx | alison waite
images.google.com.mx | ana ac elixia
```

images.google.com.mx | anna ac elixia
images.google.com.mx | anna_ac_-_elixia
images.google.com.mx | annette schwartz
images.google.com.mx | annette schwarz
images.google.com.mx | ariadne diaz
images.google.com.mx | ashley tisdale
images.google.com.mx | asus logo
images.google.com.mx | audrina
images.google.com.mx | audrina partridge
images.google.com.mx | aurika - laos
images.google.com.mx | beauty queens
images.google.com.mx | becki newton
images.google.com.mx | colleen shannon
images.google.com.mx | colleen shanon
images.google.com.mx | compaq logo
images.google.com.mx | daria peter janhans
images.google.com.mx | dell logo
images.google.com.mx | diskette de inicio win 98 ntfs
images.google.com.mx | dustin & candice
images.google.com.mx | dustin & candice beauty queens
images.google.com.mx | eva rse
images.google.com.mx | femjoy 080814-kathi in peace
images.google.com.mx | gigabyte ga-945gzm-s2
images.google.com.mx | gigabyte logo
images.google.com.mx | gigabyte m61-sme
images.google.com.mx | gugabyte ga-945gzm-s2
images.google.com.mx | heidi montag
images.google.com.mx | hp logo
images.google.com.mx | ibm logo
images.google.com.mx | in extremo
images.google.com.mx | in extremo saengerkrieg
images.google.com.mx | intel logo
images.google.com.mx | iveta - glowed
images.google.com.mx | jacque fresco
images.google.com.mx | kari byron
images.google.com.mx | kathi femjoy peace pass
images.google.com.mx | kathi in peace
images.google.com.mx | katy perry
images.google.com.mx | kaylah welivetogether
images.google.com.mx | kaylah welivetogheter
images.google.com.mx | large hadron collider

```
images.google.com.mx | lenovo logo
images.google.com.mx | leona - noblessa
images.google.com.mx | m1 garand
images.google.com.mx | mapa de mexico
images.google.com.mx | mapa linea 12 metro
images.google.com.mx | mapa linea 12 metro mexico

... output truncated ...

www.google.com.mx | office 2007 sp1
www.google.com.mx | office lock walkspace
www.google.com.mx | office lock wallspace
www.google.com.mx | office lock workspace
www.google.com.mx | password femjoy 080814-kathi in peace
www.google.com.mx | pc satellite tv
www.google.com.mx | picasa
www.google.com.mx | problema liberar db2020 cid52
www.google.com.mx | problema puerto usb frontal voltaje no reconoce
www.google.com.mx | problema usb frontal
www.google.com.mx | problema usb frontal voltaje no reconoce
www.google.com.mx | procesador e6550
www.google.com.mx | programas liberar sony
www.google.com.mx | puerto usb falla
www.google.com.mx | puerto usb frontal falla
www.google.com.mx | puerto usb frontal no trabaja
www.google.com.mx | puxin 8.11 camdriver
www.google.com.mx | puxin 8.11 camdriver w610/k550
www.google.com.mx | que carpetas puedo borrar en windows xp
www.google.com.mx | que es usb u3
www.google.com.mx | que hace divx ekg
www.google.com.mx | que hace pbsetup
www.google.com.mx | radeon 9200 se
www.google.com.mx | rasuradora sterling 6
www.google.com.mx | razuradora sterling 6
www.google.com.mx | readaheadthresold
www.google.com.mx | recuperar descargas fallidas firefox
www.google.com.mx | recuperar descargas fallidas firefox windows
www.google.com.mx | recuperar descargas fallidas trabadas firefox
www.google.com.mx | redes tac
www.google.com.mx | registro cedula direccion general de profesiones
www.google.com.mx | registro direccion general de profesiones
```

www.google.com.mx | remover alerta de windows pirata
www.google.com.mx | remover amvo desde ms-dos
www.google.com.mx | remover amvo desde msdos
www.google.com.mx | rutas internas del fs
www.google.com.mx | seagate
www.google.com.mx | sebastian lange
www.google.com.mx | sebastian lange in extremo
www.google.com.mx | seleccion de sede olimpica
www.google.com.mx | sepomex
www.google.com.mx | servers crackeados 1.3 cod2
www.google.com.mx | servidores cod2
www.google.com.mx | significado siglas oem
www.google.com.mx | skins originales walkman sony ericsson
www.google.com.mx | skins originaleswalkman sony ericsson
www.google.com.mx | skins sony ericsson
www.google.com.mx | skins walkman sony ericsson
www.google.com.mx | skype
www.google.com.mx | smc
www.google.com.mx | soap opera
www.google.com.mx | tema musical promo smallville warner latinoamerica
www.google.com.mx | tema promo smallville warner latinoamerica
www.google.com.mx | temperatura normal nvidia geforce 7300 gt
www.google.com.mx | temperatura normal zogis nvidia geforce 7300 gt
www.google.com.mx | temperatura nvidia geforce 7300 gt

THIS PAGE INTENTIONALLY LEFT BLANK

# Referenced Authors


AccessData Corp. 12, 14
Anonymous 12
Australia, N.I.C.T. 14

Beaumont, Claudine 3
Bernstein, D. J. 8, 9, 49
Black, Geoff ix, 17
Blackbird, Joseph 2
Bock, DE 52
Booch, Grady 26, 31
Booth, Kellogg S. 11
Byington, Carl 9

Card, S.K. 54
Card, Stuart K. 10, 49, 50, 53, 54, 56
Casey, E. 27
Cisco Systems, Inc. 3, 4
Cleanersoft 12
Cliff, Red 12
Cohen, MI 14
Computer Security Institute 2
Conti, Greg 52
Coppit, David 9
Cugini, J. 13

De Veaux, RD 52
DHS 3
Digital Detective, Inc. 12
Digital Investigations Group 12
Dürsteler, Juan C. 50, 54–57

Eddon, G. ix, 13, 17
Eick, Stephen G. 13
Enns, James T. 11
Entwisle, Stephen 2

Feibel, Werner 111, 112
firefoxforensics.com 12
forensic software.co.uk 12
Foundation, Python Software 9, 24, 49, 50
Fry, B. 14
Fu, X. 14

Gohring, Nancy 4
Guidance Software, Inc. 12, 14

Healey, Christopher G. 11
Heer, J. ix, 14, 21
Herder, Eelco 29, 63
Hershkop, S. ix, 14, 21
Hong, S.H. 14

IBM 22
International Organization for Standardization 23, 24
Ivar, Jacobson 26, 31

Jarrett, C. 23
Jones, Keith J. 7, 12

Kan, Min-Yen 30
Kerr, B. 15
Krebs, Valdis ix, 14, 20

Larman, C. 27
lbtechservices.com 12
Li, W.J. ix, 14, 21
Low, Mo King 2

Machor, Mitchell 12
Mackinlay, J. 54
Mackinlay, Jock 10, 49, 50, 53, 54, 56
Maeda, J. 14
Marty, Raffael 52, 53
MaxMind, Inc. 30
Mayer, Matthias 29, 63
McKinney, David 2
Metz, Joachim 9
Minocha, S. 23
Mozilla 12
Mozilla Developer Center 7, 49, 63, 111
mozillaZine 7
Munzner, T. 13

Network Working Group 8
Nikolov, NS 14

Obendorf, Hartmut 29, 63
Open Directory Project 30, 111
Overmeer, Mark 9

Passcape Software 12, 13
Perer, A. 15
Post, Jacob 12
Privacy Rights Clearinghouse 3

Reiss, S. P. ix, 13, 17
Rumbaugh, James 26, 31

Sandry, D. 11
SANS 2
SAP Business Software 13
Scholtz, J. 13
Shen, X. 14

Shneiderman, Ben 10, 49, 50, 53, 54, 56
Smith, M.A. 15
Sofer, Nir 12, 13
Sophos, Inc. 2
SS 3
Stolfo, S.J. ix, 14, 21
Stone, D. 23
Stuart, Andrew 12
Systenance Software Solutions 12

Thumbshots.com 31, 61
Tufte, E.R. 55

US-CERT 3
USC Annenberg School for Communication 1, 66

Velleman, PF 52
Vidulich, M. 11

W3C 31
Weinreich, Harald 29, 63
White, Josh 3
Wickens, C. 11
Woodroffe, M. 23
World Wide Web Consortium (W3C) 25
Wu, Y. 14
Wueest, Candid 2

X-Ways Software Technology AG 12
Xu, K. 14

Zawinski, Jamie 12

# Initial Distribution List

1. Defense Technical Information Center
   Fort Belvoir, VA

2. Dudley Knox Library
   Naval Postgraduate School
   Monterey, CA

3. Susan Alexander
   OASD/NII DOD/CIO
   Washington, DC

4. George Bieber
   OSD
   Washington, DC

5. Kris Britton
   National Security Agency
   Fort Meade, MD

6. Ed Bryant
   Unified Cross Domain Management Office
   Maryland

7. John Campbell
   National Security Agency
   Fort Meade, MD

8. Deborah Cooper
   DC Associates
   LLC Roslyn, VA

9. Dr. Steven C. Cooper
   National Science Foundation
   Arlington, VA

10. Grace Crowder
    National Security Agency
    Fort Meade, MD

11. Louise Davidson
    National Geospatial Agency
    Bethesda, MD

12. Steve Davis
    National Reconnaissance Office
    Chantilly, VA

13. Vincent J. DiMaria
    National Security Agency
    Fort Meade, MD

14. Rob Dobry
    National Security Agency
    Fort Meade, MD

15. Jennifer Guild
    SPAWAR
    Charleston, SC

16. CDR Scott Heller
    SPAWAR
    Charleston, SC

17. Steve LaFountain
    National Security Agency
    Fort Meade, MD

18. Dr. Greg Larson
    Institute for Defense Analyses
    Alexandria, VA

19. Dr. Karl Levitt
    National Science Foundation
    Arlington, VA

20. Dr. John Monastra
    Aerospace Corporation
    Chantilly, VA

21. John Mildner
    SPAWAR
    Charleston, SC

22. Dr. Victor Piotrowski
    National Science Foundation
    Arlington, VA

23. Jim Roberts
    Central Intelligence Agency
    Reston, VA

24. Ed Schneider
    Institute for Defense Analyses
    Alexandria, VA

25. Mark Schneider
    National Security Agency
    Fort Meade, MD

26. Keith Schwalm
    Good Harbor Consulting, LLC
    Washington, DC

27. Ken Shotting
    National Security Agency
    Fort Meade, MD

28. CDR Wayne Slocum
    SPAWAR
    San Diego, CA

29. Boyd Fletcher
    SPAWAR
    San Diego, CA

30. Dr. Ralph Wachter
    Office of Naval Research
    Arlington, VA

31. Dr. Cynthia E. Irvine
    Naval Postgraduate School
    Monterey, CA

32. Chris Eagle
    Naval Postgraduate School
    Monterey, CA

33. Gregory Roussas
    Civilian, Naval Postgraduate School
    Monterey, CA