Understanding Information Seeking Behavior through Network Traffic Analysis

by

Cyrus-Charles Weaver

Submitted to the Department of Electrical Engineering and Computer Science

in Partial Fulfillment of the Requirements for the Degree of

Master of Engineering in Electrical Engineering and Computer Science

at the Massachusetts Institute of Technology

August 22, 2008

Author_____
Department of Electrical Engineering and Computer Science
August 22, 2008

Certified by_____
Erik Brynjolfsson, Thesis Supervisor

Certified by_____
Sinan Aral, Thesis Co-Supervisor

Accepted by_____
Arthur C. Smith
Professor of Electrical Engineering
Chairman, Department Committee on Graduate Theses

Understanding Information Seeking Behavior through Network Traffic Analysis

by

Cyrus-Charles Weaver

Submitted to the Department of Electrical Engineering and Computer Science

August 22, 2008

In Partial Fulfillment of the Requirements for the Degree of Master of Engineering in Electrical Engineering and Computer Science

ABSTRACT

Many of today's information workers use the Internet as a valuable first-choice source for new knowledge. As such, Internet based information seeking is a key part of how information workers find information. This study develops techniques to quantify the information seeking patterns of information workers by looking at Web Site diversity, page rank, and general statistics of Web Site viewership. Future research by our group will build on these measurement techniques and explore the relationship between information worker productivity and Internet information seeking behavior.

Erik Brynjolfsson
George and Sandi Schussel Professor of Management
Director, Center for Digital Business, MIT Sloan School of Management
Thesis supervisor

Sinan Aral
Assistant Professor, Stern School of Business, New York University
Visiting Professor, MIT Sloan School of Management
Thesis co-supervisor

Table of Contents

1 Introduction

In the U.S. today, Information workers make up almost 70% of the entire labor force and contribute to over 60% of the economy's output (Apte & Nath, 2004, Aral et. Al. 2006,). These workers are primarily responsible for analyzing various sources of information and creating other value-added forms of information. Many of today's information workers use the Internet as a valuable first-choice source for new knowledge. As such, Internet based information seeking is a key part of how information workers find information.

Our research group has previously focused on the relationship between information worker productivity and several IT tools that information workers commonly use. This research has examined email data (Aral et. al. 2006, 2007, Aral & Van Alstyne 2007), email content (Aral et. al. 2007, Aral & Van Alstyne 2007, Farrokhzadi, 2007 and Monoharn, 2006), outlook calendar data (Qian, 2008), phone logs (Qian, 2008), and face to face networks using sociometirc badges (Wu et. al., 2008).

The next branch of research by our group will examine how the use of information sources on the Internet affects the productivity of information workers. Our group's previous research has shown that "diverse [social] networks" can increase the performance of information workers (Aral & Van Alstyne, 2008). Building off of this, we have reason to believe that a user's overall productivity is affected by the diversity of Internet sources they use and the meaningfulness of the Web Sites they visit. In this thesis, we developed techniques to examine information seeking behaviors on the Internet. Future research by our group will build on these

measurement techniques and explore the relationship between worker productivity and Internet seeking behavior.

We have utilized a network traffic monitoring technique to gather data from a live corporate network. From that data, we have profiled the Internet usage habits of both single users and groups of users. From our research, we hope to answer such questions as: Do information workers who view more diverse content in one area also view more diverse content in another area? Is there a relationship between the page rank of content viewed and the diversity of content viewed by an information worker? Does an increased amount of search correspond to an increase in the diversity of data viewed by an information worker? Can an information worker's diversity among certain types of Web Sites predict their overall diversity among all the types of Web Sites they visit?

2 Theory and Literature

Web Usage Mining is a process by which Web access data is "mined" and analyzed to determine patterns of user behavior (Spiliopoulou & Pohle, 2001). This process can provide important insights into key ways that information workers seek out information on the Internet. By quantifying and analyzing the patterns that occur in a user's Internet behavior, future research will be able to classify users into different types and show how their productivity relates to their Internet information seeking patterns.

To date, several techniques have been developed to track user behavior on the Internet.

2.1 Remote Tracking Tools

Using third-party software installed on user's systems, remote tracking techniques can monitor the exact keystrokes, mouse clicks and browser trajectory of a user's Internet surfing time. One research implementation of this technique collects these user activities and events and sends them to a remote tracking system (Ho, 2005). This forms an exact record of a user's Internet behavior and can even provide insights into the thinking patterns and personality of a user based on the timing of his actions.

This technique has also been used to create behavioral models of information seeking on the Web (Choo & Turnbull, 1998). In their study, Choo and Turnbull combined previous behavioral modeling into a comprehensive framework consisting of four complimentary "modes" of organizational scanning (Table 1) and six specific "information seeking activities" (Table 2) that take place during these four modes (Table 3). This research modeled the Internet use habits of an individual as a series of information seeking modes and information handling steps.

Mode	Explanation	Example
Undirected Viewing	Looking around with no formally defined information seeking goal. Coming across new information may "generate new information needs"	Simply clicking around the Internet and browsing whatever may pop up
Conditioned Viewing	Looking around with a defined topic area in mind. Provides adequate background information to allow for informal and formal search	Viewing a car Web Site in order to quickly see different cars
Informal Search	Seeking to "learn more about a specific issue or development" after having gained a basic understanding of the parameters of the topic through conditioned viewing	Viewing information about the basics of buying a car and various add-ons available
Formal Search	"[Investing] substantial time and effort in order to gather information that will enable action to be taken." Knowledge from the	Searching for a specific type of car with well-defined features, comparing the

	informal search mode allows the viewer to understand the key parameters and concepts of the subject to more thoroughly find information. Often results in a decision	various offerings for quality and comfort, and then deciding which car to purchase

Table 1 - Modes of Organizational Scanning

Mode	Explanation	Example
Starting	"Identifying sources of interest that could serve as starting points of [a] search"	Doing a quick Google search on travel to identify a few top Web Sites to look at for information
Chaining	Following up on "new leads from an initial source." Happens after starting	Clicking on links on travel information Web Sites that lead to other pages for travel agents or destination information
Browsing	Scanning through the information from various sources. Occurs after a user finds source of interest	Quickly scrolling down a Web page to see the topics covered on the page
Differentiating	"[Filtering] and [selecting] from among the sources scanned by noticing differences between the nature and quality of the information offered"	Deciding to trust and use one travel Web Site over another because of professional appearance and informative content
Monitoring	"Keeping abreast of developments in an area by regularly following particular sources"	Periodically revisiting travel Web Sites to look for new deals and information
Extracting	"Working through a particular source or sources in order to identify materials of interest"	Planning a vacation by picking out all the sites and tourist locations to visit and the best hotel

Table 2 - Information Seeking Activities

	Starting	Chaining	Browsing	Differentiating	Monitoring	Extracting
Undirected Viewing	✓	✓				
Conditioned Viewing			✓	✓		
Informal Search				✓	✓	✓
Formal Search						✓

Table 3 - Behavior Model of Information Seeking on the Web

This technique would suit our research purposes extremely well, based on the incredible breadth of data that it can generate. Theoretically, if we could install remote tracking software on every user system across a company's computer network, we would be able to capture a full picture of all the Web Sites that every user viewed, how long they viewed each Web Site, and the manner in which they navigated across the Internet.

However, the main downside of this technique lies in its feasibility. To profile many users on a corporate network, the third-party tool must be installed on every single computer system that users use for Internet browsing. For medium-sized and large-sized companies with many computer systems, this would be difficult to setup and maintain. Further difficulties would abound in cases where workers used their personal laptops at work or in which computers had different software versions and operating systems.

Because of these difficulties in deployment and maintenance, this technique was not feasible for the purposes of our research. A more passive technique was better suited for our data gathering needs.

2.2 Web Server Access Logs

A second technique for Web Usage Mining involves the use of a Web Site's access logs (Spiliopoulou, 2005). These log files typically record useful information like the IP address of a user, the exact Web Site visited, and the time and date of a user interaction with the Web Site. By processing these log files and grouping together actions by the same user, this technique can reveal a user's exact navigation path through a Web Site (Facca & Lanzi, 2003). Then, pattern discovery techniques can be used to find particular patterns in a user's behavior on a Web Site,

leading to possible insights into the way the user sees and perceives information. Statistical analysis, association rules, clustering, and sequential patterns are several such techniques that are commonly used for pattern discovery in the field of Web Usage Mining (Srivastava et. al. 2000).

The main advantage of this technique is that it can provide a complete viewing history for one or more users to a single Web Site. This history contains the addresses and timestamps of every individual page on the Web Site that every user visited. All this information resides in one location, the Web server, making it easy to gather and compile.

The main disadvantage of this technique, as it relates to our research, is that the resulting information is strictly limited to a single Web Site. This technique, while having full information of the pages a user viewed on one Web Site, gathers no information about any other Web Sites that a user may have viewed. Information workers may view dozens of different Web Sites every day in the course of their work. Thus, the scope of this technique is far too small for our use. Gaining access to Web Site logs would also be difficult. Furthermore, with the wide-spread use of Internet gateways and NAT boxes, isolating a single user in a Web Site log file would be difficult on multi-user networks.

2.3 Network Traffic

A third technique for Web Usage Mining deals with the use and analysis of network traffic data. Every time a user interacts with a Web Site, the user's computer and the Web Site communicate in the form of data packets. These packets travel through switches that connect the user to the Internet. By monitoring these packets and collecting information from them as

they traverse the network, we can track the viewing habits of many users to any and all Web Sites they visit.

One proposed application of this technique relates to enhancements of Internet search results (Weinman, 2007). Current search engines, such as google.com, return relevant results by ranking Web Sites by the degree to which those Web Sites are referenced by others. However, by using network traffic, one could determine relevance simply by seeing how popular a Web Site is to actual users who frequent the Web Site. Additional information that network traffic could reveal include the number of unique visitors to a Web Site, and the exact timing with which users come and go across a Web Site's many pages.

Theoretically, this approach can generate similarly rich data to the other two approaches, minus the capture of mouse clicks within a browser and a snapshot of the exact data being viewed at any time. Compared to the other techniques, the main benefit of this approach is that monitoring software does not need to be installed on the computers and access to Web Site log files is not required. Gathering network traffic is minimally invasive to the host network and only requires access to a few network switches and the setup of a system to store the resulting data.

Our research used sampled network traffic data to better understand the information seeking habits of information workers. By using sampling, our technique keeps the data set size low while still maintaining the statistical significance of the original network traffic.

In its raw form, the sampled network traffic data merely indicates when and how often a user viewed different Web Sites. By augmenting this data with additional categorical

information, we were able to craft detailed metrics to quantify the patterns of different Internet users.

Web Sites fall into different categories based on their content and purpose. Different users frequent various categories of Web Sites to differing degrees. One user may love cars and spend a lot of time on auto Web Sites while another user may enjoy film and spend a lot of time reading movie reviews. Even within a single category, different users can have widely different patterns. Perhaps one user always reads his news on cnn.com while another user reads his news stories from multiple sources. At the category level, these two users may have the same amount of news network traffic, but their information gathering patterns are drastically different. Ultimately, by analyzing network traffic at the category level and within the category level, we can quantitatively describe how a user seeks out information on the Internet.

The overall focus of our research group is to study how IT affects information worker productivity. By creating these metrics for Internet information seeking, we will be able to quantitatively profile the information seeking habits of individual information workers. In this thesis, users will not be individual identified but rather general trends, relationships, and correlations will be examined. Future research will examine the effect of Internet information seeking behavior on user productivity by looking for relationships between these metrics and productivity.

3 Background and Data

To better understand user information seeking habits, we partnered with a global media company located in California that specializes in multi-lingual media services. The company

allowed us to monitor their internal network traffic during the duration of this research. The company has approximately 250 employees working on site and the IT environment consists of approximately 250 computers and 50 content servers.

After gaining access to their network, we collected summary data derived from millions of actual communication packets. The summary information allowed us to piece together a picture of the Web usage habits of all the users in their network. Specific information that we collected is listed in section 3.1.3.

Due to the multi-lingual nature of the firm's work, we anticipated the presence of international-related network traffic at higher levels than a non-global firm. Similarly, since the firm is a media service company, we expected a higher level of network traffic related to media and entertainment.

3.1 Data Acquisition

3.1.1 SFlow Technology

SFlow is a packet sampling technology that allows for the monitoring of high bandwidth networks (Phaal & Lavine, 2004). In these networks, routing devices shuttle packets from source to destination. sFlow agents operate directly on the input and output interfaces of these devices, allowing for access to all traffic flowing through a device. For a given sample rate of N, sFlow agents collect one out of every N packets traversing through the routing device. The agent then summarizes relevant properties of the packet and sends the results to a collection point in the network. By using a sFlow agent on each major routing switch in a network, a complete view of all the traffic in the network is achieved.

Our partner company has millions of packets crossing its network every day. Trying to process every single packet would not have been feasible due to the sheer amount of data that would need to be collected to a central place. Because sFlow only samples the packets at an appropriately low rate and summarizes the information, we benefited by having a reasonably-sized data set that was statistically representative of all the packets flowing across their network.

3.1.2 Collecting sFlow Data

Our partner company has eight Foundry Fast Iron edge switches in their network. Each of these switches collected live sFlow data from all of the ports to which user systems were connected (Figure 1). As a result of the broad coverage of the switches, any and all user related network traffic was sampled and picked up by our research.

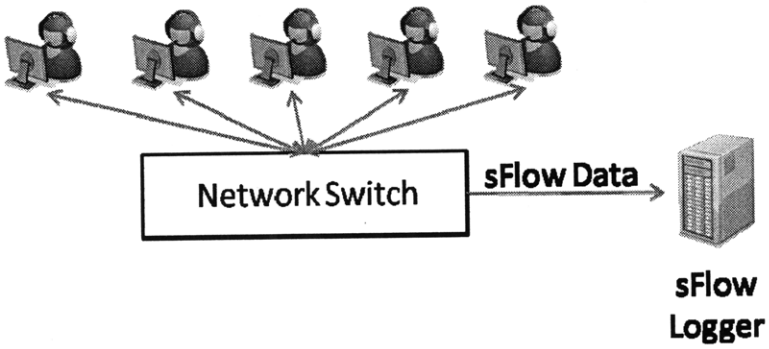

Figure 1 - sFlow Data Collection

The packets passing through the edge switches were sampled at a rate of 1 packet per 1024 packets. The sampling done here captured every 1024th packet that passes through the network switch. Over time, this resulted in an unbiased representation of the overall network traffic for both users with high network traffic and users with low network traffic. This rate allowed us to obtain a meaningful picture of network traffic without overwhelming us with too much data. To

put this into perspective, a visit to cnn.com consists of about 200 packets passing between a computer system and cnn.com's servers. Thus, when a user goes to cnn.com and reads a few articles in one visit, our technique is statistically likely to indicate that the user visited that Web Site. When aggregated over many days, even Web Sites that are infrequently visited appeared in our data set.

One important note about this technique was that the network traffic samples we collected do not represent the actual amount of time that a user spent viewing content. A high-resolution photo will takes thousands of packets to transmit but only a few seconds to look at while a 10 page article will take much fewer packets to transmit but take several minutes to read through. Future work in this area can try to compensate for this effect by estimating how much actual user time different Web Sites are likely to require. Furthermore, video streaming and audio streaming are more bandwidth intensive than viewing of regular text Web Sites. Future work in this area can try to compensate for this by considering how packet intense certain user behaviors are and normalizing the data accordingly to approximate actual numbers of visits to a Web Site. Lastly, some Web Sites, like email Web Sites and online music Web Sites, are often left open and periodically transmit packets all day without user intervention. Future work in this area can try to compensate for this effect by studying the browsing patterns of users and making adjustments to separate intentional and unintentional page views.

3.1.3 Storing sFlow data

At our partner company, all the sFlow data from the eight edge switches was continually entered into a MySQL server on their premises. At MIT, our research group's MySQL server

replicated the first server for data backup and accessibility. At the beginning of our data processing steps, we copied the network data to a third system for data analysis and processing.

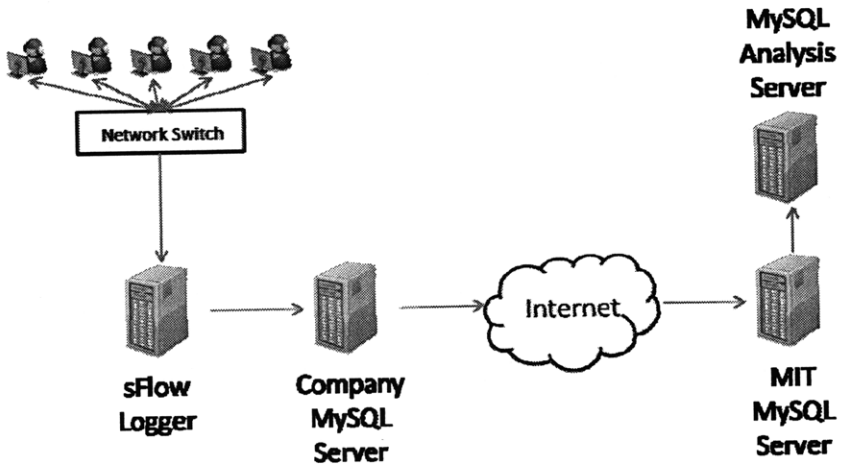

Figure 2 - MySQL Data path

We placed the sFlow data into a table called "network" in the MySQL database (Table 4). This table stored all the relevant properties of the sFlow data required for our analysis.

Field Name	Description
id	An auto-incrementing id value for the entry
date	The date and time of the sampled packet
src_mac	The MAC address of the source of the packet
dst_mac	The MAC address of the destination of the packet
src_host	The hostname of the source of the packet
dest_host	The hostname of the destination of the packet
src_port	The port of the source system
dst_port	The port of the destination system

Table 4 – "Network" MySQL Table Field Descriptions

3.2 Data Characteristics

3.2.1 Data Statistics

There were 11.64 million rows of data in the network table, taking up 2GB of space. Due to a 41-day break in capture of the network data during this research[1], the network data naturally fell into two ranges. Range 1 consisted of 6.56 million rows and covered a 105 day period from November 8, 2007 to February 22, 2008. Range 2 consisted of 5.08 million rows and covered a 72 day period from April 4, 2008 to June 16, 2008.

Network traffic occurs in two forms. In the internally communicated form, traffic both originates and ends inside the network. In the externally communicated form, traffic either comes from an external Web server or goes to an external Web server. The sFlow sampling method we used samples internal and external packets all together. However, our research focused on the externally communicated form of network traffic, specifically traffic traveling from inside the network to an external destination.

Most of the network traffic from our partner company corresponded to internally communicated traffic. We stored all the traffic samples, but did not use the internally communicated traffic for this thesis. Only a small portion of all the sampled data corresponded to externally destined data. For Range 1, there were a total of 341K rows (5.2% of the total number of rows for Range 1) and for Range 2, there were a total of 265K rows (5.22% of the total number of rows for Range 2).

[1] The 41-day break in network traffic was caused by an error in the MySQL database setup that limited the number of rows allowed in the data table.

3.3 Data Usability

In its original form, the raw data we collected could already show us the day and time that a user visited specific Web Sites. However, modeling the *types* of Web Sites a user visited over time was the central objective of this research. For example, the raw data, with slight modification, could already indicate that a user visited cnn.com, abc.com, amazon.com, and nytimes.com. Important additional information was needed to realize that the user visited two news Web Sites, one entertainment Web Site, and one commerce Web Site. The pre-processing steps described in the Methods section enhanced the data set by adding in contextual information about each Web Site. This enhancement process incorporated information from the external Web Sites themselves and other external sources. The pre-processing steps also separated out entries for which contextual information was unavailable or inadequate.

4 Methods

In order to effectively analyze the network traffic to examine the Internet patterns of users, we needed to pre-process the network data. This multi-step process transformed the data into a more readily usable form by adding in additional content used to categorize the visited Web Sites. After the pre-processing was done, we worked to formulate the resulting data into different forms for the different analysis approaches described in the Results section.

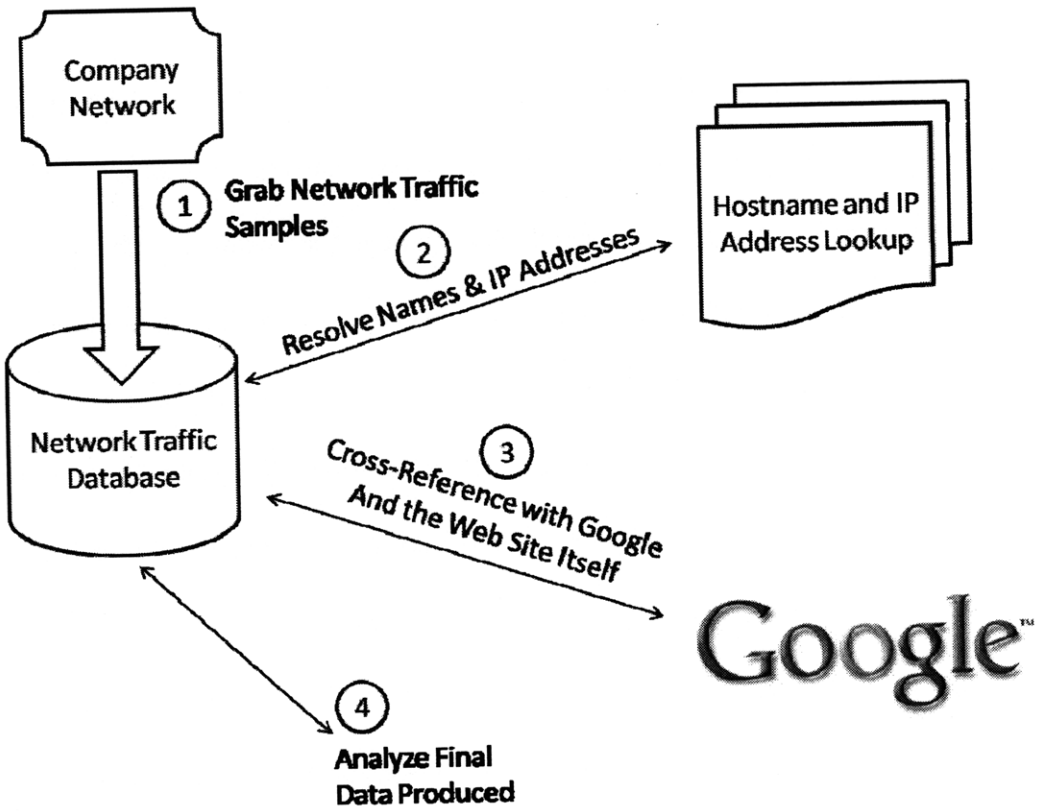

Figure 3 - Overall Processing Steps

4.1 Data Pre-Processing Steps

4.1.1 Data Marking Steps

As previously mentioned, our data fell into two time periods. The first step we took was to augment the network table by adding a range field to indicate whether an entry was Range 1 or Range 2. This would allow for much easier querying since we could simply reference that field instead of having to plug in the start and stop dates for each range. We ran two quick MySQL queries on the data to accomplish this goal:

- UPDATE network SET range=1 WHERE id < 6562136
- UPDATE network SET range=2 WHERE id >= 6562136

Id number 6562136 belonged to the first entry in the time-ordered network table corresponding to the start of Range 2.

Next, we filtered out the internally communicated network traffic from the externally-destined network traffic. The hostnames and IP addresses of our partner company's internal systems had a consistent structure. Internal hostnames contained the name of the company and internal IP addresses were of the standard form for internal network traffic (10.*.*.* and 192.*.*.*). We also had to separate out VRRP (Virtual Router Redundancy Protocol) traffic which looked like external traffic (destined for vrrp.mcast.net), but actually represented internal routing related information. We augmented the network table by adding fields called is_local, is_external, and is_vrrp. We used these fields to mark whether an entry was externally-destined traffic, internally communicated traffic, or VRRP traffic. We ran three MySQL queries on the data to accomplish this goal:

- UPDATE network SET is_local = 1 WHERE dst_host LIKE '192.%' OR LIKE '10.%'
- UPDATE network SET is_local = 1 WHERE dst_host LIKE '%company_name%'
- UPDATE network SET is_vrrp = 1 WHERE dst_host LIKE '%vrrp%'

4.1.2 External Web Site Processing

After separating out the different types of network traffic, we processed all the entries marked as externally-destined. Specifically, we looked at the destination host field of each entry to link-up all externally-destined traffic destined for the same Web domain.

The destination host fields of the entries were structured in one of two possible ways: an IP address or a hostname. For example, a Web server belonging to google.com may appear in the form of an IP address, "209.85.141.176", or in the form of a hostname, "wa-in-f19.google.com."

Both of these representations are for google.com. For added scalability and reliability, most high traffic Web Sites today utilize multiple Web servers to deliver the same content for a single Web address. Google has hundreds of servers that all operate under google.com. Thus, our pre-processing needed to form a link between all the possible hostname representations for a single Web Site. This allowed us to capture all the visits to a Web Site despite the differing forms that a hostname could take.

For destination host fields in the form of IP addresses, we utilized PHP's gethostbyaddr function to resolve the IP addresses to hostnames. After the conversion, we processed the converted hostnames in the same way as the hostname form of the destination host field. Sometimes, an IP address could not be resolved to a hostname. These entries were discarded from the data set since we would be unable to link them with a regular Web Site.

For destination host fields in the form of hostnames, we used a PHP script to reduce them to their second-level domain and top-level domain. For United States related Web Sites, this process was straight forward and involved retaining the part of the hostname right after the second-to-last period, or the whole hostname if no second-to-last period existed in the hostname. For example, "s01.lastfm.com" would become "lastfm.com." The top-level domain, "com," and the second-level domain, "lastfm," would be retained but the third-level domain, "s01," would be discarded. International Web Sites had to be handled more cautiously. These Web Sites typically have country-code top-level domains which are two letter suffixes at the end of an address, such as ".cn" for Chinese Web Sites or ".uk" for United Kingdom Web Sites. We had to retain the country-code top-level domain and the traditional second-level domain as well. For example, "thu2.planetlab.edu.cn" would become "planetlab.edu.cn", retaining "cn" as

the country-code, "edu" as the second-level domain, and "planetlab" as the third level domain. The fourth-level domain, "thu2" would be discarded in this example.

In following with the rules for database normalization, we created a new table, external_web_site, to avoid redundant information. The key fields of this table are listed in Table 5 below.

Field Name	Description
id	An auto-incrementing id value for the entry
web_site	The resulting Web Site address after pre-processing the destination host fields from the network table
instances_range1	The number of times this site was visited in the network table for Range 1
instances_range2	The number of times this site was visited in the network table for Range 2
has_min_instances	Indicates if this Web Site has enough total instances to pass the consideration threshold (10)
is_failed	Indicates if this Web Site failed to be resolved
failed_reason	Indicates the reason why the Web Site failed to be resolved
google_desc	The description of the Web Site from a Google search of the hostname
meta_desc	The contents on the "description" meta tag of the Web Site
meta_keywords	The contents of the "keywords" meta tag of the Web Site
manual_desc	A manually entered description of the Web Site, if needed
is_whq	Indicates if this Web Site falls under the categories of Web hosting, content delivery, proxies, online advertising or questionable content. Used during analysis to exclude Web Sites from the results
page_rank	The Google "page rank" of the Web Site

Table 5 – "External_web_site" MySQL Table Description

We added a field, ext_web_site_id, to the network table to link each row to a single Web Site entry in the external_web_site table. A PHP script processed all the destination host fields in the network table and created a row in the external_web_site table for each Web Site encountered in the data set. After these steps, all network table entries for the same Web Site referred to a single Web Site row in the external_web_site table.

Once all the entries in the network table were linked to their actual Web Sites, we ran two MySQL queries to count the number of times that a Web Site appeared in Range 1 and Range 2.

- UPDATE external_web_site SET instances_range1 = (SELECT count(*) FROM network WHERE range=1 AND network.ext_web_site_id = ext_web_site.id)
- UPDATE external_web_site SET instances_range2 = (SELECT count(*) FROM network WHERE range=2 AND network.ext_web_site_id = ext_web_site.id)

Pre-processing the network table resulted in 13K unique Web Sites with instances numbering from the single digits to the several thousands. Categorizing all 13K Web Sites would have been impractical. Furthermore, since the top 10% of Web Sites visited made up 90% of the total externally-destined traffic, categorizing all 13K would have been unnecessary. Thus, we focused our attention on those Web Sites appearing at least ten times in the data set. Setting the threshold at this level allowed us to consider just 1.4K Web Sites, a much more manageable number for our categorization process.

4.1.3 External Web Site Data Augmentation

As mentioned in the Background section, modeling the *types* of content a user visited over time is the central point of this thesis. A Web Site address does not intrinsically indicate the type of content found on a Web Site. To properly categorize a Web Site, we needed to acquire

additional information about the Web Site's content and purpose. By using three separate types of Web Site content information, we attempted to gather content information for all 1.4K remaining rows in the external_web_site table. For each Web Site, we attempted to gather Google's description of the Web Site, the "description" meta tag from the Web Site itself, and the "keywords" meta tag from the Web Site itself. These three methods were about 50% accurate and did not always produce results in some cases. These methods tended to succeed or fail together since Google descriptions were often snippets of the "description" meta tag and since Web Sites with a "description" meta tag tended to also have a "keywords" meta tag. However, by using these three different methods simultaneously, it maximized our chances of success and decreased the amount of processing needed for future steps.

Next, we manually walked though the 1.4K entries, verifying them for accuracy and fixing or adding information as necessary. Ultimately, we wanted to end up with a list of Web Sites with accompanying descriptive information so that we could properly categorize them. There were several types of problems we fixed during this walk-through process.

Trying to pull down Google's description of the Web Site was not always accurate. Our PHP script ran a Google search for the exact Web Site name and retrieved the description of the first result. However, the first result was not always the correct description. Sometimes the second or third result would be the proper description. Another error we encountered was that some Web Sites lacked a Google description. This was often due to the Web Site having non-textual Flash components (Flash Web Sites are often not properly described by Google because the textual content is embedded in Flash instead of HTML). A third error we encountered was that some Web Sites had incorrect descriptive information. Often, the description of these Web

Sites would simply be the first words that appeared on the Web Site, even though those words may not truly be descriptive of the Web Site's content. We corrected for all of these types of errors by doing a manual Google search for Web Sites that were noticeable mismatched or contained incomplete information.

In terms of the meta tags, some Web Sites had non HTML-compliant "description" and "keyword" meta tags. In other cases, proper tags existed but our PHP script could not reach the Web Site. We visited these Web Sites, retrieved the proper information, and manually inserted the information into the external_web_site table.

Additionally, many Web Sites did not have meta tags or available Google descriptions. To form descriptive information, we visited the Web Sites and constructed a description from available information. This often took the form of a snippet from the "About Us" section. In other cases, we formed a description from the clearly visible content on the page. For example, a Web Site for soccer could easily be seen as being sports related.

Lastly, since our partner company specializes in international media, some Web Sites were from other countries and in different languages. We used free online translation tools, such as freetranslation.com, to translate descriptive information from these Web Sites into English.

This manual walk-through process also turned up instances where differently spelled Web Sites were all directed to the same destination. For example, blackberry.net is an alias of blackberry.com. We moved all the different aliases of a destination to a single Web Site so that our instance counts for each range would be completely accurate.

4.1.4 External Web Site Categorization

With all the necessary descriptive information at hand, we moved to categorize all of the Web Sites. No reliable database of Web Site categorization existed that contained all the information we needed. Thus, we had to do the categorization ourselves. We initially looked into using categorization algorithms such as Vector Space Modeling (Salton et. al. 1975). Past research by our group has used Vector Space Modeling to categorize Wikipedia entries (Manoharn, 2006). In that research, the Wikipedia articles were rich with distinctive worlds that accurately allowed Web Sites dealing with the same content to be grouped together. However, the descriptive information we gathered from Google, the Web Page itself, or manually constructed tended to be short phrases or sentences. The small feature space of our descriptive information made highly-accurate automatic characterization infeasible without a lot of manually intervention. Web Sites belonging to different categories could ultimately have overlapping key words that would prevent Vector Space Modeling from producing accurate results. For example, Web Sites providing information on video cameras would fall under the Information category while Web Sites selling video cameras would fall under to commerce category and Web Sites providing videos would fall under the Entertainment category. These types of small distinction would have been lost if we tried to categorize by Vector Space Modeling on the short Web Site descriptions. Additionally, Web Sites that should be categorized together often had different sets of words referring to the same concepts. For example, one Web hosting Web Site may use the phrase "Web hosting" while another Web Site may prefer to market its "dedicated servers." Both of these terms refer to Web hosting, but use completely different words.

Due to the infeasibility of existing automatic categorization methods, we manually categorized the Web Sites. We created a categorization model with 18 possible main categories that the Web Sites in our data set could fall under. While these categories are not mutually exclusive, 88.5% of the Web Sites fell into only a single category, 11% of the Web Sites fell into exactly two categories, and .05% of the Web Sites fell into exactly three categories. Table 6, below, describes each of the 18 main categories and gives examples for each.

Category	Description	Examples
Access	Companies that offer phone, Internet, or television services	comcast.net rr.com
Blogging	Online blogs and blogging services	xanga.com boingboing.com
Commerce	Online stores that are related to commerce, the selling of a variety of goods, shipping goods, or paying for goods on commerce Web Sites	amazon.com ebay.com ups.com
Education	Educational institutions or educational reference sources	ucla.edu wikimedia.org mit.edu
Entertainment	Arts, film, dance, theater, games, media, television, music, radio, sports, live events, ticketing, or the promotional thereof	youtube.com live365.com ticketmaster.com
Financial	Financial institutions, financial services, investment services, loan companies, payment services or insurance companies	bankofamerica.com fidelity.com wellsfargo.com
Government	Federal, state, or local government entities or services	ca.gov lapdonline.org lacity.org
Informational	Provides information on goods, ideas, locations, or products but do not primarily handle commerce	cnet.com opentable.com slashdot.com
Internet Services & Software	Tools and services available online and for use over the Internet, and software	mozilla.org microsoft.com
News	Sources of news, either general or covering specific topics like entertainment, business, or local news	cnn.com forbes.com msnbc.com
Products	Promotes specific commercial and non-commercial products. These Web Sites may sell	apple.com crateandbarrel.com

	their products also, but unlike "commerce" Web Sites, these Web Sites typically only sell *their own products*	dell.com
Professional Services	Real-world professional services offered by various types of companies such as job searching, media processing, translation services, or design services	monster.com proz.com bydeluxe.com
Search	Used for finding information on the Internet	google.com yahoo.com answers.com
Social Networking	Online communities of people connected for professional or personal purposes	facebook.com myspace.com linkedin.com
Travel & Transportation	Modes of transportation, such as buses or planes, or information related to traveling, booking, or vacation destination	southwest.com travelocity.com priceline.com
Web Hosting & Content Delivery	Companies that provide Web hosting services, Web advertising services, proxies, content distribution services, and Web Site caches	secureserver.net dreamhost.com akamai.com doubleclick.com
Web Portal & Email	Web portals to various types of information on the Internet and Email services	aol.com hotmail.com
Questionable	May potentially be related to Web hosting and content delivery. As explained in later sections, these Web Sites are excluded from our analysis	real.com turn.com

Table 6 - Web Site Main Categories

After manually categorizing the Web Sites, we realized that some of the network traffic needed to be flagged and removed from consideration in our analysis. Many Web Sites across the Internet are hosted by hosting companies such as godaddy.com, or 1and1.com. Often, a single Web server at the hosting company will host many different unrelated Web Sites through the same IP address. In these cases, information about which Web Site a user specifically visited was not available in the sFlow packet summaries. In our analysis, we excluded such Web hosting traffic since we were unable to properly categorize the true destination. We set the

"is_whq" field to "1" for these Web Sites in the external_web_site table, allowing us to exclude those Web Sites in our MySQL queries when needed.

Similarly, there was a high amount of network traffic related to content delivery and web advertising. We reasoned that this traffic, although coming from these specific Web Sites, was actually displayed on other Web Sites. For example, the advertising site doubleclick.net provides paid advertising to thousands of other Web Sites. Resulting network traffic did not represent a visit to the Web Site doubleclick.net, but rather represented ads appearing on other Web Sites. Thus, we could not use these Web Sites in our data analysis. We set the "is_whq" field to "1" for these Web Sites in the external_web_site table, allowing us to exclude those Web Sites in our MySQL queries when needed.

We further divided the Web Sites in each main category into sub-categories. This resulted in 140 sub-categories with some main categories having as many as eighteen sub-categories and other main categories having no sub-categories (See Appendix C: Sub-Category Breakdowns).

To store all the category related information, we created three new tables, main_category, sub_category, and sub_category_assignment. Every main category had an entry in the main_category table with the name of the category and a unique id for the category. Each sub-category had an entry in the sub_category table with the name of the sub-category, a unique id for the sub-category, and the id number of the sub-category's parent category. Lastly, each pairing of a Web Site with a sub-category (and by extension, a parent-category) was inserted into the sub_category_assignment table. This arrangement of information allowed for easier querying of the data. Table 7, Table 8, and Table 9 respectively describe the MySQL layouts of each of the three tables.

Field Name	Description
id	An auto-incrementing id value for the entry
name	The name of this main category

Table 7 – "main_category" MySQL Table Description

Field Name	Description
id	An auto-incrementing id value for the entry
name	The name of this sub-category
parent_id	The id of the main category that this subcategory falls under

Table 8 - "sub_category" MySQL Table Description

Field Name	Description
stc_id	The id of the sub-category that the corresponding external_web_site falls under
e_id	The id of the external Web Site that falls under this category

Table 9 - "sub_category_assignment" MySQL Table Description

4.1.4.1 Yahoo.com and Google.com Exceptions

During our pre-processing phase, we noticed a large amount of traffic from google.com and yahoo.com related servers. While both of these Web Sites are well known for their search capabilities, they also provide Email accounts and other smaller services. We needed to separate out the search portion of the network traffic from the other types of network traffic from these two Web Sites. We did not have enough information to arrive at a precise solution, so we estimated as best as we could in order to separate out email traffic.

Many of yahoo.com's sub-domain servers are structured as follows: "d1.ycs.vip.s1s.yahoo.com."As such, we are not sure what exact service these servers actually carry out. Some other servers are in a form such as "mg2a.*mail*.vip.re1.yahoo.com." We separated out the Yahoo related entries in the network table that contained "mail," in the

destination host and assumed that these handled all the Email traffic. This amounted to 13% of the yahoo related traffic. These Email related instances were categorized under the "Web Portal & Email" category.

Google's sub-domain servers, on the other hand, are all structured in the same non-descriptive form, "wa-in-f19.google.com." We were unable to determine which network traffic packets traveling through these servers corresponded to search and which corresponded to Email and other services. To estimate these values, we used the 13% figure from Yahoo's Email traffic and randomly selected 13% of the google.com entries in the network table to represent email traffic.

4.1.5 Pre-processing Outcome

Starting with the 341K and 265K external network traffic rows in the data set for Range 1 and Range 2, respectively, our pre-processing steps reduced this data to 303K and 238K usable network traffic rows by filtering out rarely viewed Web Sites and invalid Web Sites that could not be reached. This resulted in a roughly 10% reduction in size for both ranges. Additionally, starting from 13K unique Web Sites in the data set, our pre-processing steps reduced this number by almost 90% to 1.4K unique Web Sites while still retaining 90% of the total network traffic.

4.1.6 Processing Code Statistics

Our pre-processing code consisted of about 4.5K lines of code across about 35 script files. The PHP script that processed the network table and pulled out Web Site names, the PHP script that pulled down meta data tags and Google descriptions, and the PHP script which retrieved the page rank of a site collectively took over a full week to completely run.

4.2 Data Analysis Processing Steps

Having categorized and grouped all the externally-destined network traffic in the pre-processing stages, we moved to the analysis phase. We used several MySQL scripts to retrieve the appropriate data from the necessary tables and used PHP scripts to reorganize the data into a form ready for Excel manipulation. For example, in order to create a daily view of each user's Web Site category browsing, we first ran a MySQL query to join together information from the network table, the external_web_site table, and the three category related tables. We then used PHP scripts to transform the MySQL output from tens of thousands of rows, each representing the activity of one user for a single day in a single category, to an Excel spreadsheet in which a single row contained all the daily activity of a user, organized by each day and category combination in the columns.

5 Results

5.1 Introduction

In our analysis, we often selected which data to use by limiting the calculations to only those users who had exceeded a certain threshold of daily average network traffic. The threshold values we commonly used were 5 instances per day, 10 instances per day, 20 instances per day, and 50 instances per day. Making these considerations allowed our metrics to be more meaningful in certain situations by excluding the users in the network who seldom browsed the Internet for whatever reason.

We also excluded network traffic that was related to Web hosting, content delivery, proxies, and advertising placement. As explained in the Methods section, we performed this exclusion because we were unable to pinpoint the category of content originating from these Web Sites.

Lastly, as mentioned in the Methods section, we only considered Web Sites that had at least 10 instances of network traffic in the data set. This allowed us to focus on the top 10% of Web Sites that carried 90% of the network traffic in our data set. Importantly, this prevented rarely visited Web Sites from diluting the results. For example, in calculating diversity, the number of possible Web Sites that could be visited affected how much impact a single instance of network traffic to a single Web Site had on the diversity score. By excluding rarely visited Web Sites from consideration in the diversity calculations, the resulting diversity score was more reflective of the bulk of the network traffic.

5.2 Main Category Breakdowns

As explained in the Methods section, our pre-processing steps divided the network traffic in the data set into different categories, based on the type of Web Site visited. The resulting division of network traffic reveals many interesting facts about the information seeking habits of Internet users.

5.2.1 Overall Traffic

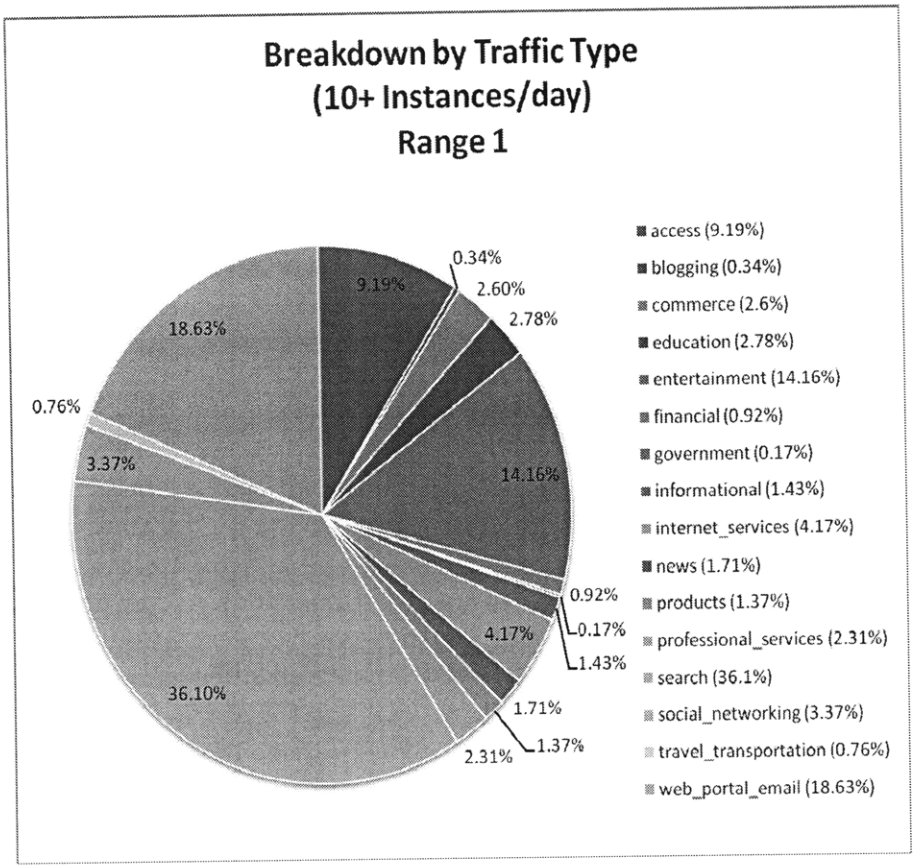

Figure 4 - Breakdown by Traffic Type. 10+ Instances/day. Range 1

Figure 4 shows the overall category proportions of the network traffic for users with an average of 10 instances per day. As first realized in our pre-processing steps, the Search category made up a large portion of our data set with 36.10% of the total network traffic. This is mainly from the Web Sites google.com and yahoo.com. Search engine use is very common for today's Internet users. Many times, users will "Google" to quickly arrive at a Web Site they frequently visit instead of using bookmarks or typing a long url. Additionally, users mostly use search engines as the starting point for finding various forms of information.

The Access and Web Portal & Email categories also make up a large portion of the data set at 9.19% and 18.63% respectively. The relatively high amount of network traffic in the Access

category surprised us. We reasoned that a lot of this traffic involves users at our partner company checking their home email accounts at work (the accounts they have through their Internet providers at home). However, without being able to clearly identify this as email related traffic, we left the data in the "Access" category. The relatively high proportion of Web Portal & Email traffic is likely due to users leaving their email accounts (such as Gmail and Yahoo Mail) open at work during the work day.

To gain a better understanding of the other 13 categories, we looked at the same results, but excluded the Access, Search, and Web Portal & Email categories. As shown in Figure 5, we can now see how often the other Web Site categories appeared in the data set. Entertainment was the biggest contributor to network traffic with 39.24%. Given the prevalence of video-sharing Web Sites like youtube.com and music streaming Web Sites like immem.com, this was not surprising. Streaming content from these Web Sites generates more network traffic than simply viewing a static Web Site. So we expected that this type of traffic would be more present in the data set.

Internet Services & Software, Social Networking, Education, and Commerce Web Sites made up the second groupings of prevalent Web Sites in our network traffic at 11.56%, 9.35%, 7.7%, and 7.21%, respectively. Internet Services & Software appeared at a high proportion due to the corporate nature of our network traffic. With many computers on their network, contact is often made to microsoft.com for software updates or to online collaboration tools like groove.net (such Web Sites account for 37% of the total Internet Services & Software traffic). The other three categories can be understood in the context of typical user Internet activity of

today. The traffic in the Social Networking account is largely due to the popularity of the Web Sites in that category. Many people have myspace.com or facebook.com accounts and frequently check them during the day. Educational Web Sites, largely in the form of wikipedia.org entries, are common among today's Internet users. Commerce Web Sites, like amazon.com, are an often visited destination for many Internet users as more and more people shop online.

The last remaining categories made up smaller pieces of the data set. Web Sites falling under the Blogging and Government categories were comparatively less common than Web Sites falling under the categories of Financial, Informational, News, Products, Professional Services and Travel & Transportation.

Extra figures for Range 1 and figures for Range 2 can be found in Appendix B: Main Category Breakdowns

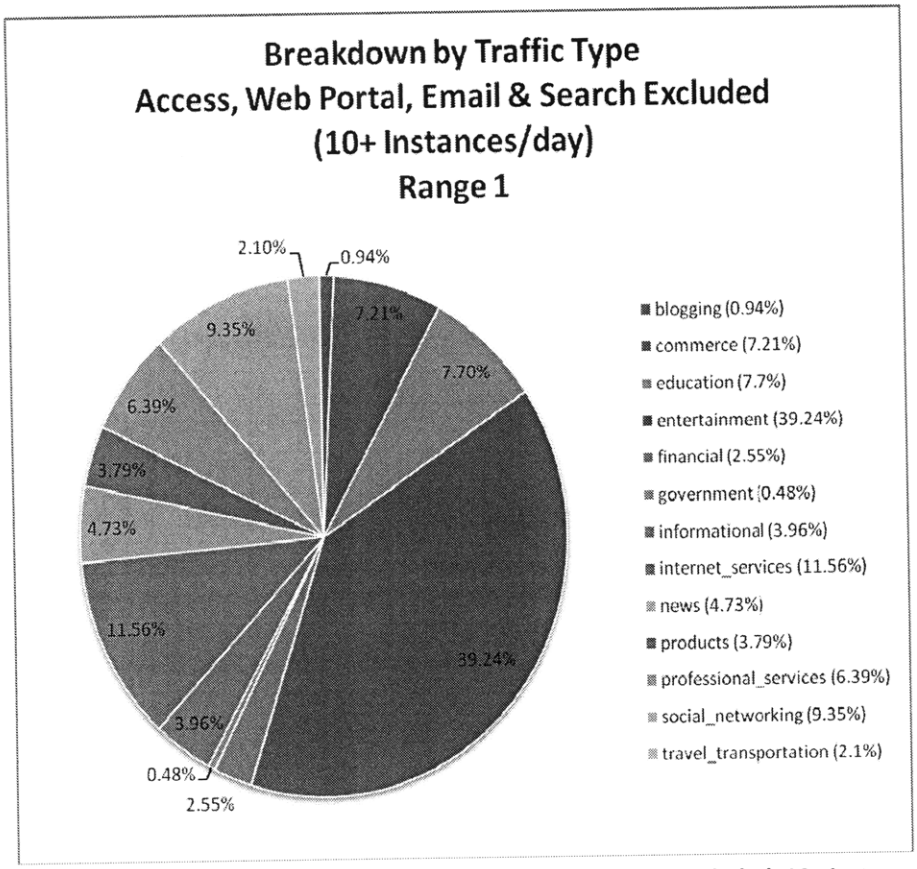

Figure 5 - Breakdown by Traffic Type. Access, Web Portal & Email, & Search Excluded. 10+ Instances/day. Range 1

5.2.2 Top 20 Users

In terms of individual user behavior, one possibility is that all users view categories of Web Sites in the same proportions as all the other users. Alternatively, each individual user may have completely different proportions of Web Site instances compared to other users. To gain a sample view of how various users browsed between the 16 categories, we looked at the main category divisions for the top 20 users by total number of network traffic instances in our data set (Figure 6). The main feature to notice is that these top 20 users had various divisions of Web Site traffic between the 16 different categories. While search tended to dominate for most

users, some users had higher proportions of traffic from one or two unique categories, such as Entertainment or Social Networking.

We also noticed some strange activity that required us to dig back into the raw network data. Looking at the graph, we noticed that user #8 had a markedly high proportion of Professional Services traffic. After looking through the network table, we saw that this network traffic came from a media services company in a related field to our partner company. Perhaps this individual is responsible for communications with the other company. Whatever the case, it was clear that this individual had a special role involving the outside company that other users did not have.

Overall, this result showed that every user in the network was unique in the way they split their Web Site visits between the different categories of Web Sites.

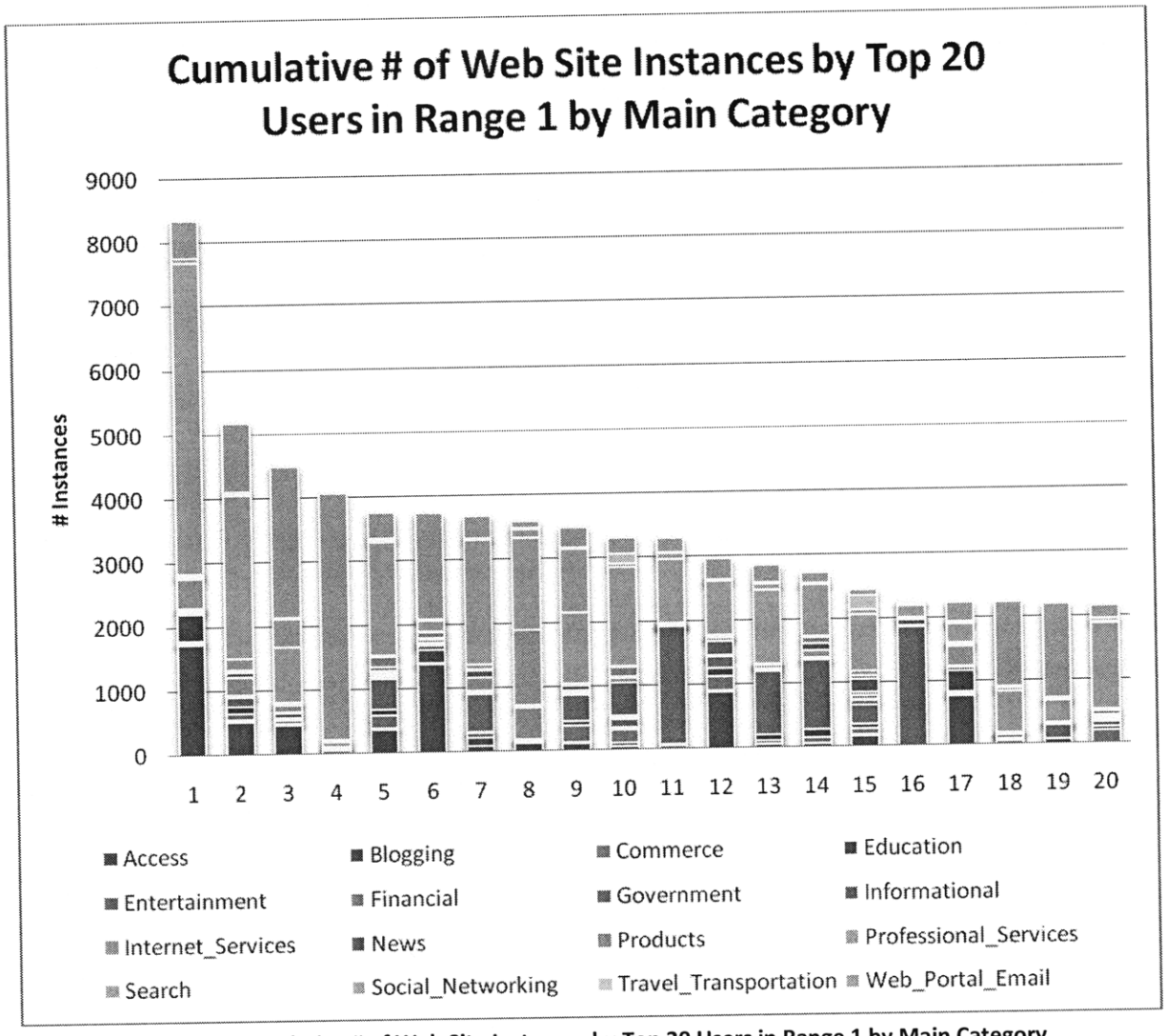

Figure 6 - Cumulative # of Web Site Instances by Top 20 Users in Range 1 by Main Category

Taking the same information found in Figure 6, we looked at the breakdown for the top 20 users across two divisions (Figure 7). The first division contained the Access, Search, and Web Portal & Email categories. The second division contained traffic from all the other 13 categories. For most users, their network traffic fell mostly into the first division. The reasons for this are the same reasons explained in section 5.2.1 describing the pie charts of main category traffic proportions.

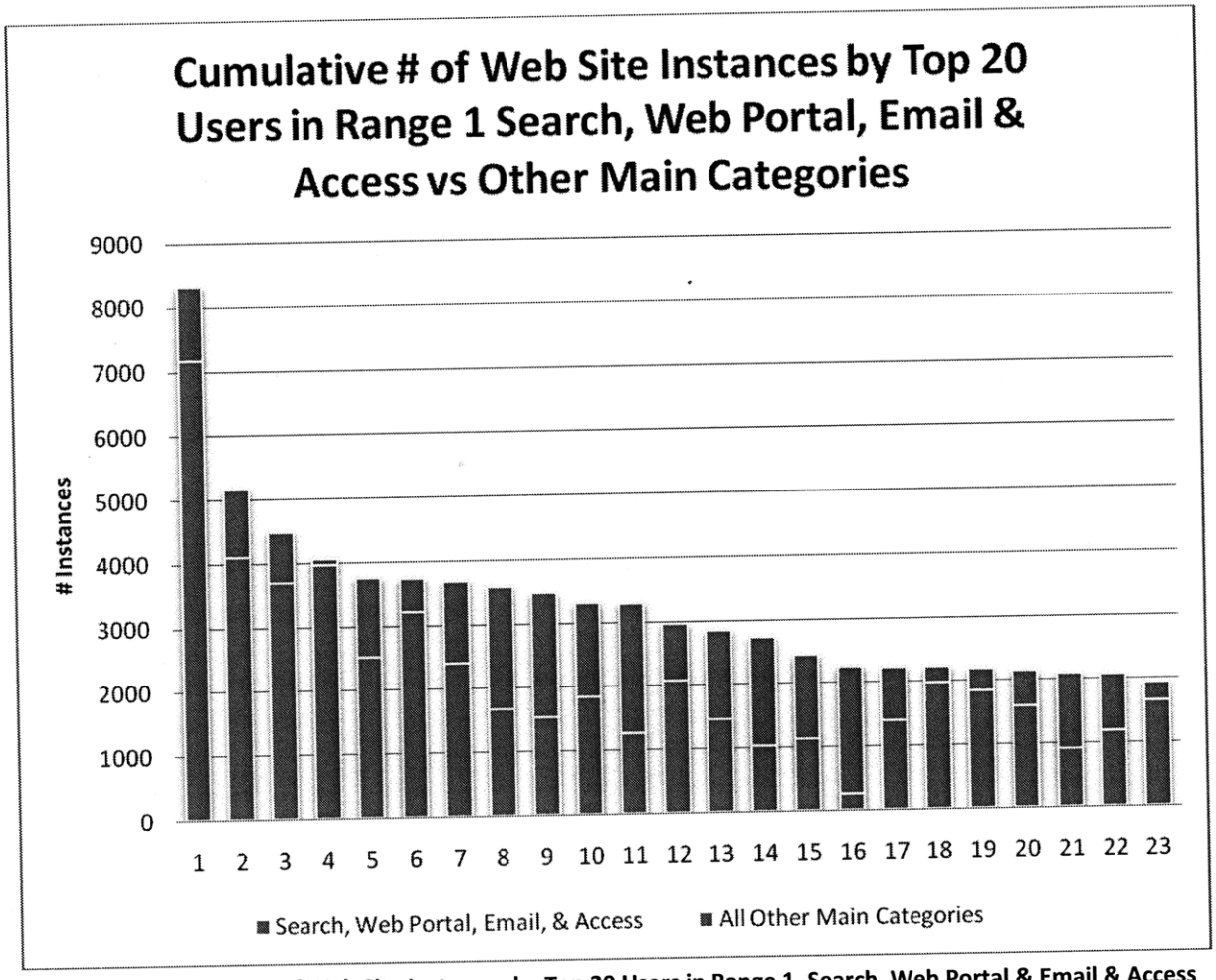

Figure 7 - Cumulative # of Web Site Instances by Top 20 Users in Range 1. Search, Web Portal & Email & Access vs Other Main Categories

Given that much of the network traffic fell into only a few categories, we also looked at the top 20 users in terms of the number of unique Web Sites they visited (Figure 8). These top 20 users had between 204 and 100 unique Web Sites represented in their network traffic. Similar to Figure 6, all these users had differing divisions of Web Site access across the 16 categories.

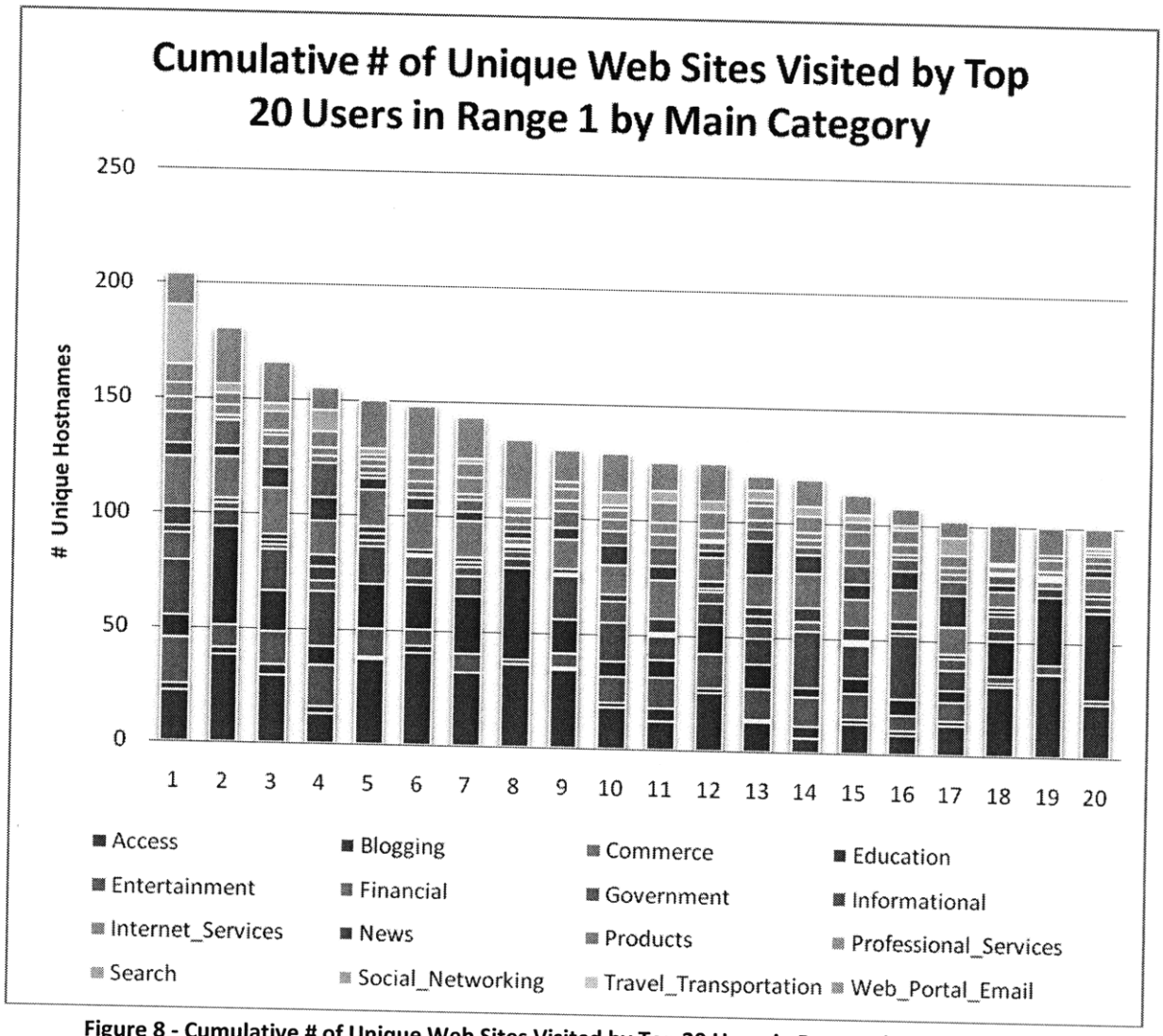

Figure 8 - Cumulative # of Unique Web Sites Visited by Top 20 Users in Range 1 by Main Category

Blogging, Financial, Government, Informational, News, Products, Professional Services, Search, Travel & Transportation and Social Networking all had relatively small numbers of unique Web Sites represented for these top 20 users. This suggested that, in these categories, users viewed a small number of Web Sites to seek out the information they were looking for. This likely occurred for one of two reasons. Some categories have only a few key Web Sites that most users frequent and some categories tend to have users that are loyal to a small number of Web Sites. In the Search category, for example, only a handful of key Web Sites made up for

most of the traffic, namely google.com and yahoo.com. Almost all the users had used one of those two Web Sites many times. But in the News category, users may have tended to look for their news from a consistent set of sources. Thus, while there are many possible news Web Sites available, users may have tended to use the same ones every time they sought out news information.

The Access, Commerce, Education, Entertainment, Internet Services & Software, and Web Portal & Email categories had relatively larger numbers of unique Web Sites represented. This suggested that users find information on these categories from a wider variety of Web Sites. For example, more Web Sites in the Commerce area may reflect users shopping around for goods or going to different Web Sites to purchase different kinds of products. Perhaps a user buys his books from amzon.com, his electronics from bestbuy.com, and his music from iTunes.com.

In Figure 9, we have grouped the categories into two divisions as the grouping for Figure 7. As shown in this graph, most of the users viewed more Web Sites in the other 13 categories than in the three categories of Access, Search, and Web Portal & Email. Although most of the network traffic came from these three categories, these three categories generally accounted for a much smaller portion of the total unique Web Sites viewed.

Extra figures for Range 1 and figures for Range 2 can be found in Appendix B: Main Category Breakdowns.

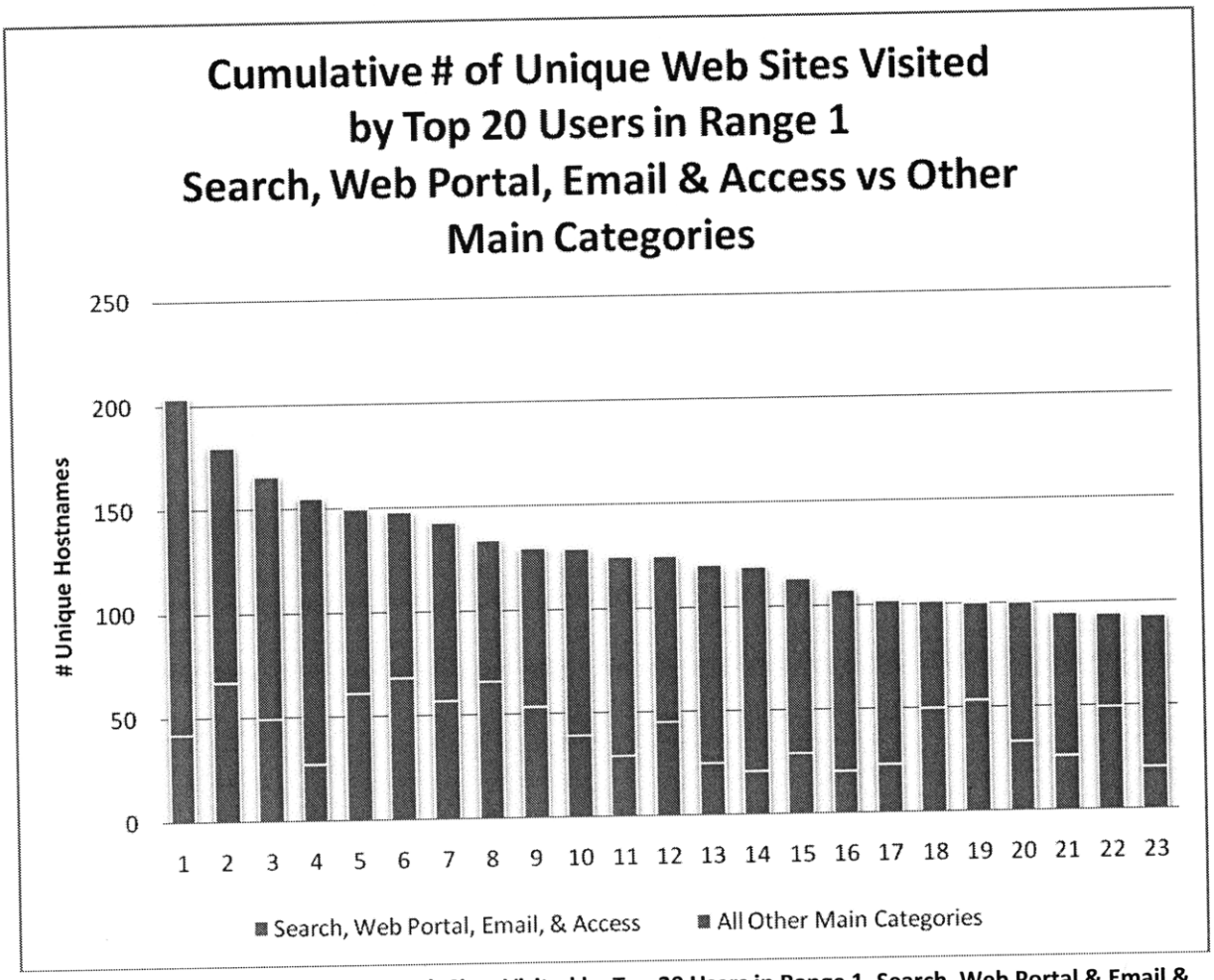

Figure 9 - Cumulative # of Unique Web Sites Visited by Top 20 Users in Range 1. Search, Web Portal & Email & Access vs Other Main Categories

5.2.3 Average Number of Instances by Category over Time

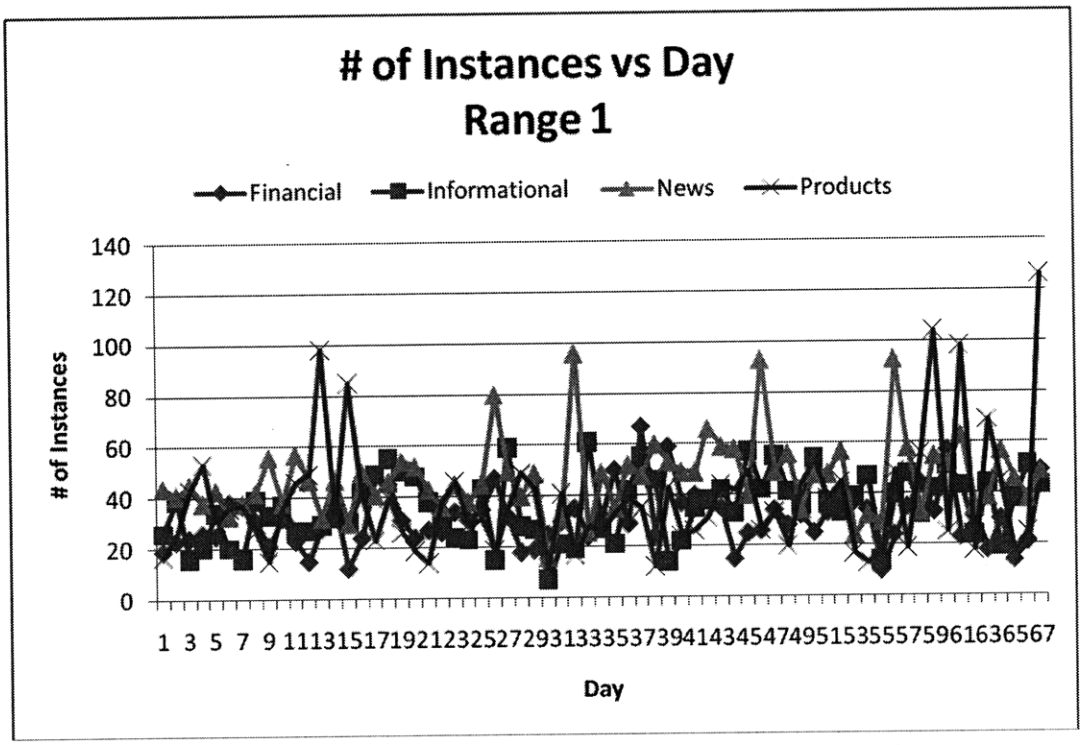

Figure 10 - # of Instances vs Day. Range 1. Financial, Information, News, Products

Figure 10 and Figure 11 show the number of overall instances by day for two sets of categories. Weekends and holidays have been removed from consideration since most of the traffic at our partner company occurred during business hours. The first set consists of the Financial, Informational, News, and Products categories and the second set consists of the Commerce, Education, Internet Services & Software, and Social Networking categories. The groupings in these graphs reflect categories that have similar amounts of daily network traffic during Range 1.

In Figure 10, those 4 categories typically have between 20 and 40 daily instances and in Figure 11, those 4 categories typically have between 50 and 100 daily instances. The spikes in traffic typically happened on a single day and were typically caused by a single user's greater

than average viewing of Web Sites related to that category for that given day. Overall, the pattern of category accesses tended to remain around the same numerical values for most of the data set.

Remaining figures for Range 1 and figures for Range 2 can be found in Appendix B: Main Category Breakdowns.

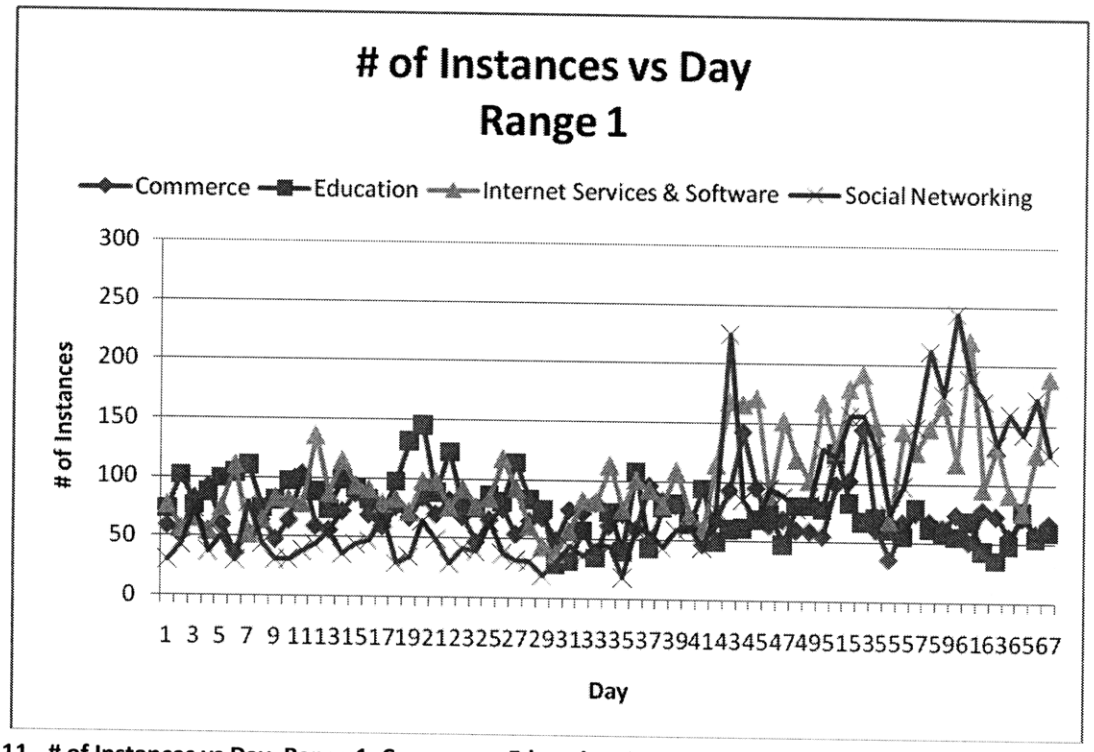

Figure 11 - # of Instances vs Day. Range 1. Commerce, Education. Internet Services & Software, Social Network

5.2.4 Deviation from Average Number of Instances by Category over Time

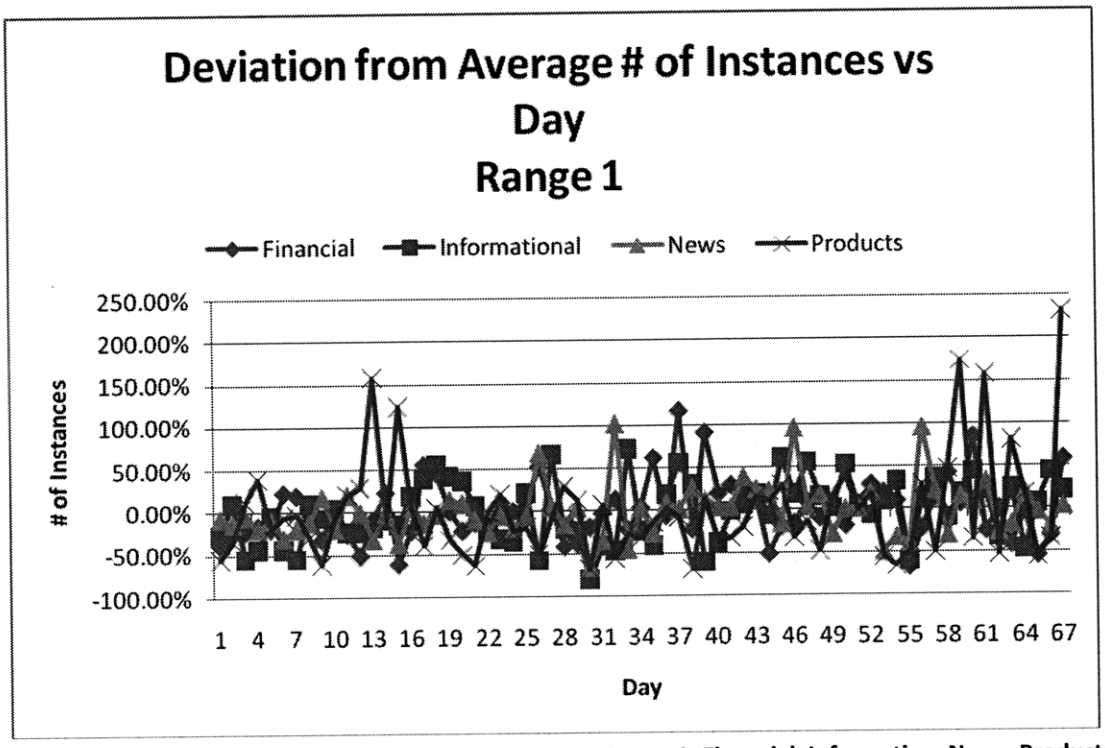

Figure 12 - Deviation from Average # of Instances vs Day. Range 1. Financial, Information, News, Products

Figure 12 and Figure 13 show the deviation from the average number of overall instances by day for two sets of categories. Weekends and holidays have been removed from consideration since most of the traffic at our partner company occurred during business hours. The first set consists of the Financial, Informational, News, and Products categories and the second set consists of the Commerce, Education, Internet Services & Software, and Social Networking categories. The groupings in these graphs reflect categories that have similar amounts of daily network traffic during Range 1.

In Figure 12, Informational, News, and Financial Category Web Site instances tend to remain with 50% of their overall average with occasional spikes 100% from the average. Yet the Products category, while also tending to stay well within 50% of the average, has several spikes

that deviate over 150% above the average. In general, these spikes in instances tended to be caused by the occasional viewing of packet intensive data. For example, a user deciding to watch video reviews of products instead of reading textual reviews of a product would use more bandwidth. In Figure 13, Commerce, Internet Services & Software, and Education tend to stay within 50% of the average while Social Networking, deviating less overall, had several larger spikes in the data set.

Remaining figures for Range 1 and figures for Range 2 can be found in Appendix B: Main Category Breakdowns.

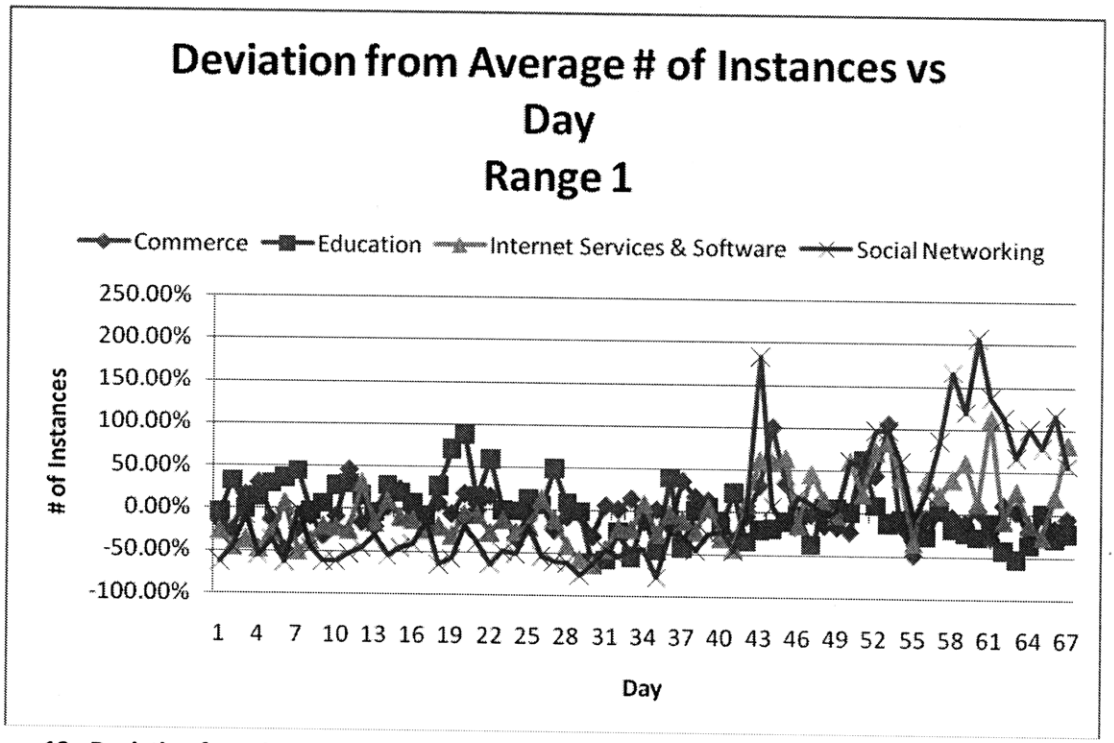

Figure 13 - Deviation from Average # of Instances vs Day. Range 1. Commerce, Education. Internet Services & Software, Social Network

5.2.5 Average Number of Unique Sites by Category over Time

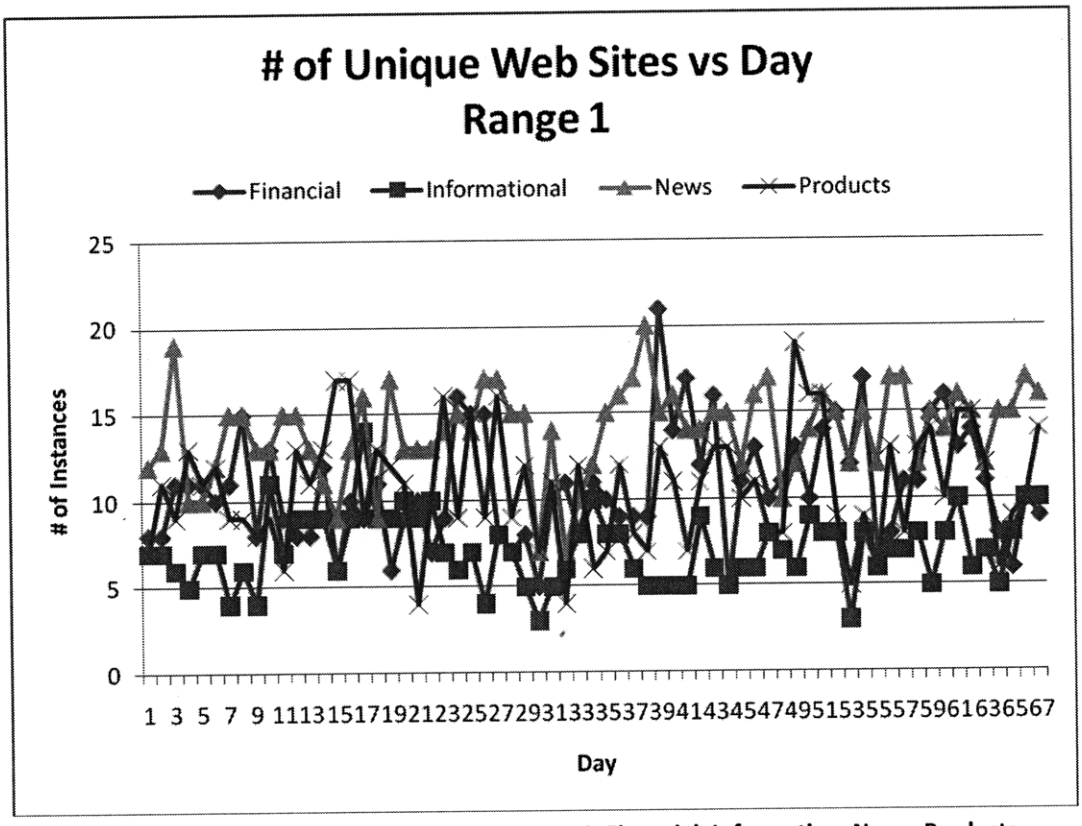

Figure 14 - # of Unique Web Sites vs Day. Range 1. Financial, Information, News, Products

Figure 14 and Figure 15 show the number of unique Web Sites visited by day for the same two sets of categories as in Figure 10 and Figure 11. Once again, weekends and holidays have been removed from consideration.

In Figure 14, there are typically between 5 and 10 unique Web Sites in those four categories and in Figure 15, there are typically between 10 and 25 unique Web Sites. These categories tended to remain around their averages with fluctuations happening from day to day. The spikes in the graphs were typically caused by a single user viewing multiple Web Sites on a specific topic within a short period of time. For example, a single user may decide to research different banks to open a new checking account, thus causing a spike in the number of unique bank Web Sites viewed.

Remaining figures for Range 1 and figures for Range 2 can be found in Appendix B: Main Category Breakdowns.

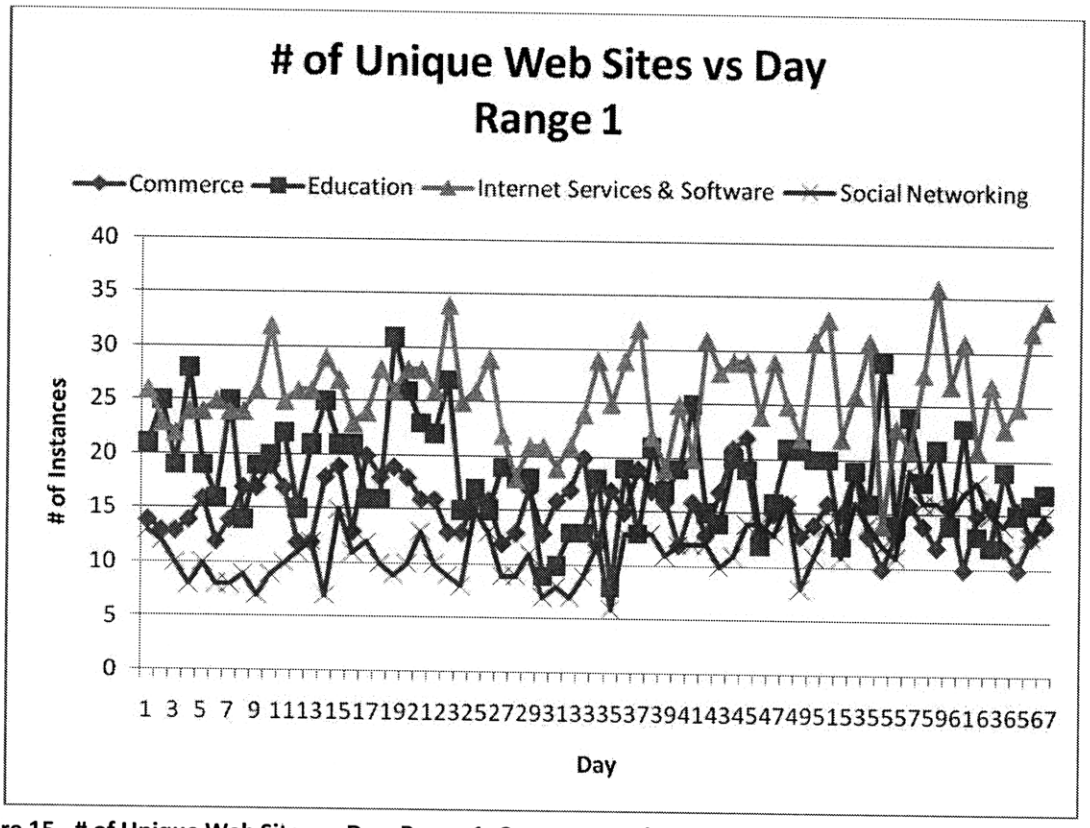

Figure 15 - # of Unique Web Sites vs Day. Range 1. Commerce, Education. Internet Services & Software, Social Network

5.2.6 Deviation from Average Number of Unique Web Sites by Category over Time

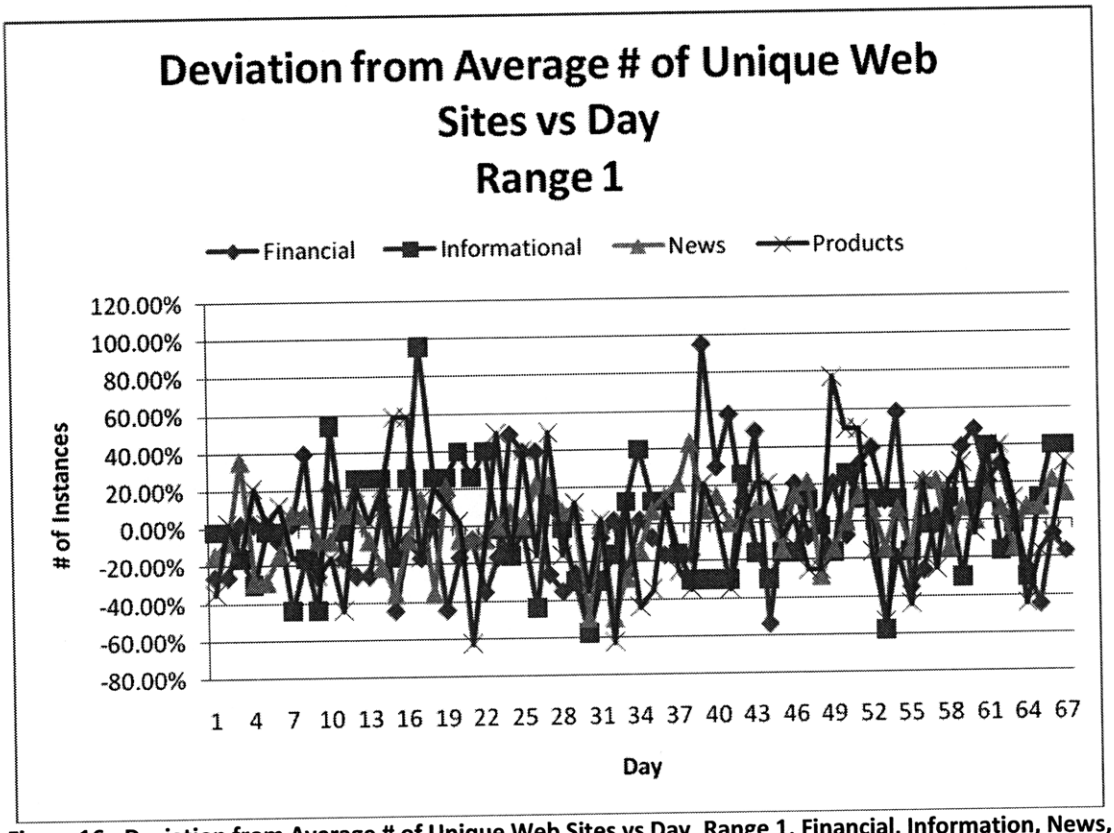

Figure 16 - Deviation from Average # of Unique Web Sites vs Day. Range 1. Financial, Information, News, Products

Figure 16 and Figure 17 show the deviation from the average number of unique Web Sites visited by day for the same two sets of categories as in Figure 14 and Figure 15. Weekends and holidays have been removed from consideration.

In Figure 16, the Internet Services & Software category tends to stay within 20% of its average value while Commerce, Education, and Social Networking fluctuate more wildly and tend to stay within 40% of their average value with occasional spikes.

Remaining figures for Range 1 and figures for Range 2 can be found in Appendix B: Main Category Breakdowns.

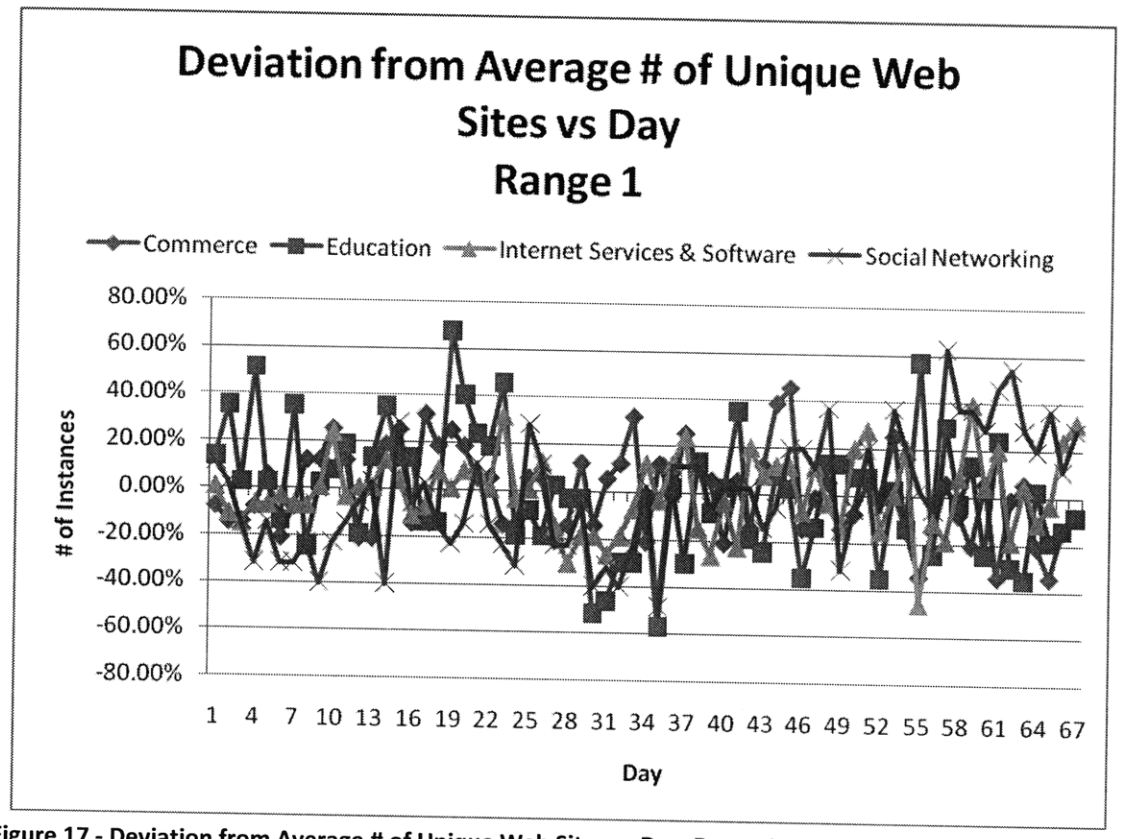

Figure 17 - Deviation from Average # of Unique Web Sites vs Day. Range 1. Commerce, Education. Internet Services & Software, Social Network

5.2.7 Pair-wise Correlation of Instance Counts between Categories

In Table 10 and Table 11, we looked at the pair-wise correlation of instance counts between categories for Range 1 and Range 2. In Range 1, the pairs of categories that appear to be correlated were: Entertainment and Access, Informational and Commerce, Entertainment and Search, and Travel & Transportation and Information. In Range 2, the pairs of categories that appear to be correlated were: Entertainment and Access, Search and Commerce, and Search and Travel & Transportation. Overall, though some categories show some correlation at times, there was no consistent pattern of correlation between Range 1 and Range 2 besides the pair of Entertainment and Access. Thus, we concluded that pairs of categories were uncorrelated overall.

	A	B	C	Ent	Edu	F	G	Inf	Int	N	Pdt	Prf	Sch	Scl	T	Wpe
A	1.00	-0.03	0.02	**0.72**	-0.17	0.01	-0.05	0.00	0.24	-0.05	0.25	-0.05	0.33	-0.02	-0.10	0.20
B	-0.03	1.00	0.34	-0.04	0.10	-0.03	0.20	0.27	0.09	0.33	0.11	0.12	0.35	0.05	0.14	0.03
C	0.02	0.34	1.00	0.09	0.02	0.11	**0.47**	**0.50**	-0.05	0.22	0.33	0.05	0.30	0.35	0.33	-0.01
Ent	**0.72**	-0.04	0.09	1.00	-0.11	0.12	0.08	-0.01	0.34	0.06	0.06	-0.09	**0.45**	0.03	-0.03	0.09
Edu	-0.17	0.10	0.02	-0.11	1.00	-0.14	0.04	0.02	-0.08	0.15	-0.03	-0.04	-0.02	0.04	-0.06	-0.15
F	0.01	-0.03	0.11	0.12	-0.14	1.00	0.19	-0.04	0.05	0.10	0.08	0.04	0.14	0.03	0.24	-0.08
G	-0.05	0.20	**0.47**	0.08	0.04	0.19	1.00	**0.62**	0.04	0.13	**0.47**	-0.07	0.24	-0.02	**0.60**	-0.07
Inf	0.00	0.27	**0.50**	-0.01	0.02	-0.04	**0.62**	1.00	-0.02	0.03	0.26	-0.06	0.16	-0.01	**0.53**	-0.08
Int	0.24	0.09	-0.05	0.34	-0.08	0.05	0.04	-0.02	1.00	0.04	-0.04	0.29	0.28	-0.01	-0.02	-0.02
N	-0.05	0.33	0.22	0.06	0.15	0.10	0.13	0.03	0.04	1.00	0.03	0.03	0.12	0.09	0.23	-0.09
Pdt	0.25	0.11	0.33	0.06	-0.03	0.08	**0.47**	0.26	-0.04	0.03	1.00	0.00	0.06	-0.03	0.30	0.04
Prf	-0.05	0.12	0.05	-0.09	-0.04	0.04	-0.07	-0.06	0.29	0.03	0.00	1.00	0.14	0.10	-0.05	-0.06
Sch	0.33	0.35	0.30	**0.45**	-0.02	0.14	0.24	0.16	0.28	0.12	0.06	0.14	1.00	0.12	0.13	0.02
Scl	-0.02	0.05	0.35	0.03	0.04	0.03	-0.02	-0.01	-0.01	0.09	-0.03	0.10	0.12	1.00	-0.01	0.08
T	-0.10	0.14	0.33	-0.03	-0.06	0.24	**0.60**	**0.53**	-0.02	0.23	0.30	-0.05	0.13	-0.01	1.00	-0.07
Wpe	0.20	0.03	-0.01	0.09	-0.15	-0.08	-0.07	-0.08	-0.02	-0.09	0.04	-0.06	0.02	0.08	-0.07	1.00

Table 10 - Pair-wise Correlation of Instance Counts between Categories. Range 1

	A	B	C	Ent	Edu	F	G	Inf	Int	N	Pdt	Prf	Sch	Scl	T	Wpe
A	1.00	-0.08	-0.03	**0.56**	-0.11	0.09	-0.02	0.03	0.15	-0.04	0.00	0.07	-0.12	-0.10	-0.06	0.14
B	-0.08	1.00	0.08	-0.02	0.16	-0.01	0.01	0.18	0.10	-0.01	-0.05	0.02	0.19	0.05	-0.04	0.04
C	-0.03	0.08	1.00	0.02	0.07	0.26	**0.57**	0.10	0.05	0.04	0.22	-0.02	**0.45**	0.04	0.07	0.03
Ent	**0.56**	-0.02	0.02	1.00	-0.09	-0.03	0.03	0.02	0.24	-0.03	0.02	-0.04	-0.05	-0.05	0.01	0.09
Edu	-0.11	0.16	0.07	-0.09	1.00	-0.06	-0.02	0.00	-0.05	-0.02	-0.04	-0.06	0.06	0.36	-0.09	-0.09
F	0.09	-0.01	0.26	-0.03	-0.06	1.00	0.20	-0.02	0.03	0.02	0.17	0.05	0.35	-0.01	0.35	0.08
G	-0.02	0.01	**0.57**	0.03	-0.02	0.20	1.00	0.26	0.08	0.01	0.22	-0.06	**0.47**	-0.04	0.15	-0.01
Inf	0.03	0.18	0.10	0.02	0.00	-0.02	0.26	1.00	-0.03	-0.05	0.00	-0.03	0.20	-0.04	0.04	-0.04
Int	0.15	0.10	0.05	0.24	-0.05	0.03	0.08	-0.03	1.00	-0.04	0.08	0.19	0.09	-0.01	0.17	0.09
N	-0.04	-0.01	0.04	-0.03	-0.02	0.02	0.01	-0.05	-0.04	1.00	0.19	-0.03	0.00	-0.07	0.10	-0.08
Pdt	0.00	-0.05	0.22	0.02	-0.04	0.17	0.22	0.00	0.08	0.19	1.00	-0.03	0.04	-0.07	0.05	0.00
Prf	0.07	0.02	-0.02	-0.04	-0.06	0.05	-0.06	-0.03	0.19	-0.03	-0.03	1.00	0.07	0.11	-0.05	0.01
Sch	-0.12	0.19	**0.45**	-0.05	0.06	0.35	**0.47**	0.20	0.09	0.00	0.04	0.07	1.00	0.03	**0.45**	0.25
Scl	-0.10	0.05	0.04	-0.05	0.36	-0.01	-0.04	-0.04	-0.01	-0.07	-0.07	0.11	0.03	1.00	0.11	0.11
T	-0.06	-0.04	0.07	0.01	-0.09	0.35	0.15	0.04	0.17	0.10	0.05	-0.05	**0.45**	0.11	1.00	0.03
Wpe	0.14	0.04	0.03	0.09	-0.09	0.08	-0.01	-0.04	0.09	-0.08	0.00	0.01	0.25	0.11	0.03	1.00

Table 11 – Pair-wise Correlation of Instance Counts between Categories. Range 2

5.2.8 Search Category Instances Regression

In Table 12 and Table 13, we list the regression coefficients for Search as a function of the other 15 categories for Range 1 and Range 2. For Range 1, the categories that had a statistically significant impact on Search were Blogging and Education which both had p-values less than .1. In Range 2, the categories that had a statistically significant impact on Search were Commerce, Government, Professional Services, Social Networking, Travel & Transportation and Web Portal & Email. However, because Range 1 and Range 2 did not share the same set of statistically significant categories, we viewed these results as inconclusive for the entire data set. Perhaps the differences between the two ranges were due to noise in the data or other factors which are beyond the scope of this thesis.

Category	*Coefficient*	*Standard Error*	*t Stat*	*P-value*
Intercept	221.81	122.174	1.816	0.074
Access	0.31	0.320	0.967	0.337
Blogging	12.40	4.115	3.014	0.004
Commerce	1.84	1.663	1.109	0.271
Education	3.14	1.481	2.119	0.038
Entertainment	0.04	0.177	0.236	0.814
Financial	1.16	2.104	0.553	0.582
Government	18.75	15.270	1.228	0.224
Informational	-1.07	1.474	-0.723	0.472
Internet Services & Software	0.34	0.521	0.652	0.516
News	-0.99	1.202	-0.827	0.411
Products	-2.31	1.785	-1.293	0.200
Professional Services	0.51	0.471	1.089	0.280
Social Networking	0.20	0.511	0.389	0.699
Travel & Transportation	1.38	2.855	0.485	0.629
Web Portal & Email	-0.04	0.115	-0.311	0.757

Table 12 – Regression Coefficients for # of Search Instances as a Function of the # of Instances of the Other Categories. Range 1

Category	*Coefficient*	*Standard Error*	*t Stat*	*P-value*
Intercept	203.304	44.455	4.573	1.86E-05
Access	-0.126	0.104	-1.214	0.228

Blogging	3.827	2.325	1.646	0.104
Commerce	1.113	0.388	2.868	0.005
Education	0.018	0.256	0.069	0.945
Entertainment	0.144	0.095	1.514	0.134
Financial	0.973	0.811	1.200	0.234
Government	17.393	8.668	2.007	0.048
Informational	0.686	0.634	1.082	0.283
Internet Services & Software	-0.159	0.421	-0.379	0.706
News	-0.051	0.253	-0.202	0.841
Products	-0.436	0.362	-1.206	0.231
Professional Services	0.163	0.096	1.692	0.095
Social Networking	-0.156	0.110	-1.410	0.163
Travel & Transportation	4.532	0.981	4.619	1.57E-05
Web Portal & Email	0.434	0.124	3.508	0.001

Table 13 - Regression Coefficients for # of Search Instances as a Function of the # of Instances of the Other Categories. Range 2

5.3 Sub-Category Breakdowns

Within each main category, we also divided the network traffic into sub-categories. The following are results and analysis related to that process.

5.3.1 Overall Traffic

This section focuses on the five main categories of Commerce, Entertainment, Financial, News, and Products. Remaining figures for Range 1 and figures for Range 2 can be found in Appendix C: Sub-Category Breakdowns.

For the Commerce category in Figure 18, the majority of Web Site instances were from commerce Web Sites that deal with a variety of goods, such as amazon.com. This is not surprising as users tend to go to these types of Web Sites to buy any and all types of goods they seek. The remaining sub-categories roughly split the rest of the network traffic with some categories having more traffic than others.

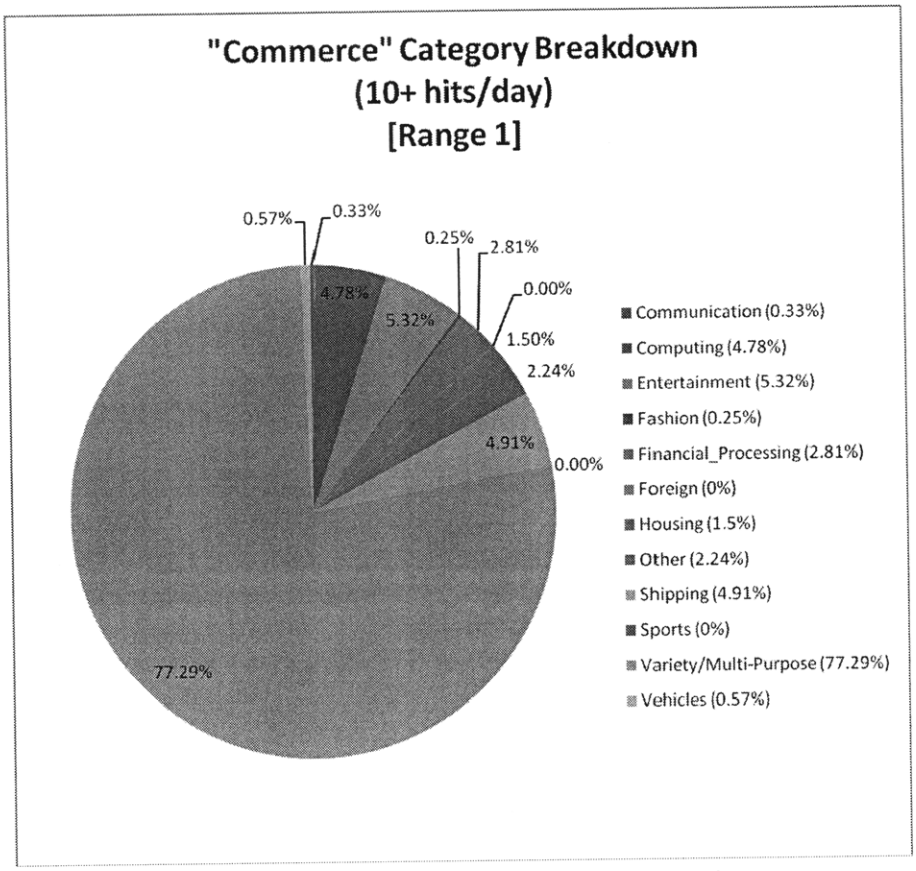

Figure 18 - "Commerce" Category Breakdown. 10+ Instances/day. Range 1

For the Entertainment category in Figure 19, video-related Web Sites were the largest sub-category with 43.07%. Next were Web Sites with a broad range of entertainment content, at 18.44%, followed by Web Sites with music related content, at 13.65%. Given the prevalence of media related Web Sites, like youtube.com and imeem.com, the high proportion of video and music related instances in the Entertainment main category is reasonable. Since our partner company deals with media and entertainment, we would expect to see this wide variety of different Web Sites.

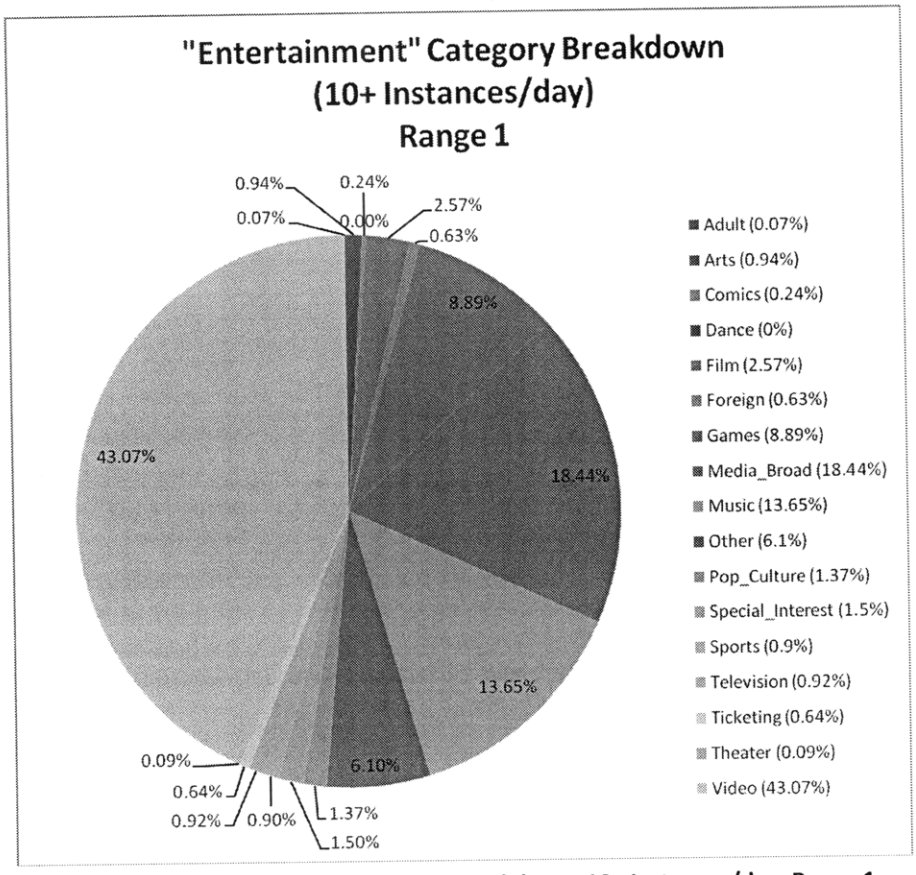

Figure 19 - "Entertainment" Category Breakdown. 10+ Instances/day. Range 1

For the Financial category In Figure 20, banking related Web Sites made up the majority of this traffic with 54.39% followed by investing Web Sites, at 18.03%, and credit related Web Sites, at 9.95%. Matching with our real life experience, the high amount of banking related traffic was likely related to users checking their bank statements or making payments online. In general, smaller numbers of people deal with investments in the real world, so there was less network traffic in that area as compared to banking. Lastly, since many banks today handle checking and credit card accounts together, there was a smaller amount of exclusively credit card oriented Web Site traffic. Payment services made up the fourth highest category at 8.00% and mostly consisted of PayPal use. The remaining sub-categories roughly split the remaining Financial category traffic.

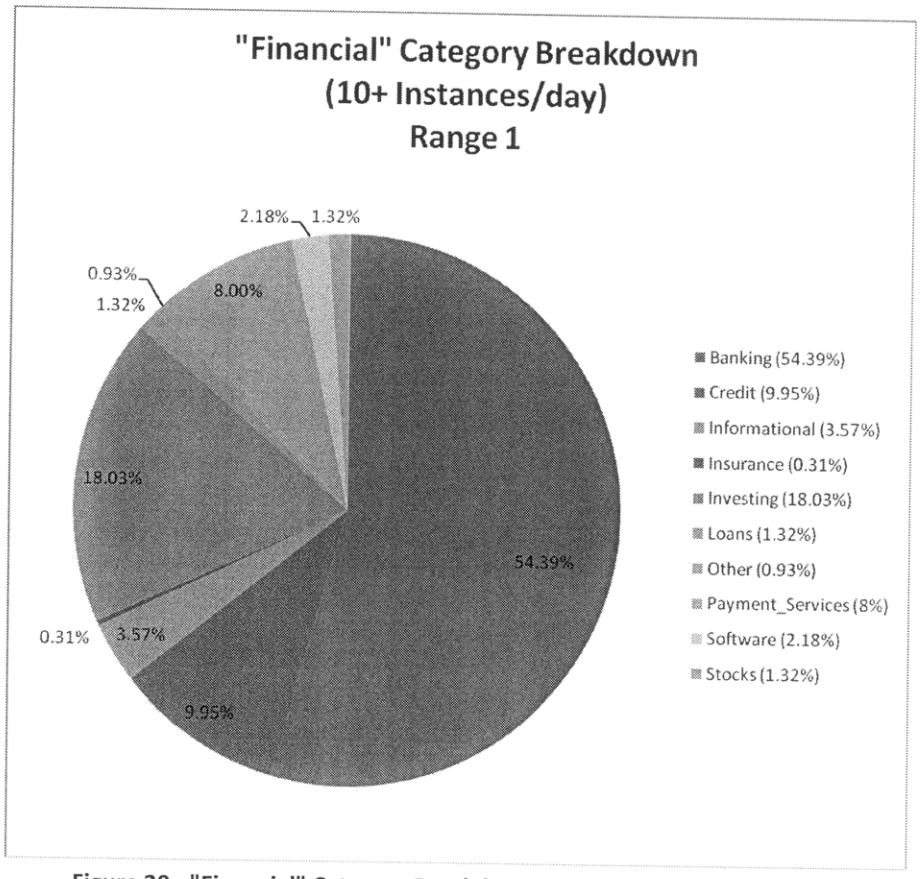

Figure 20 - "Financial" Category Breakdown. 10+ Instances/day. Range 1

For the News category in Figure 21, national, special interest, and foreign news made up the top three sub-categories with 36.08%, 21.98%, and 13.98% respectively. National news as the top category was not surprising since the most common news outlets, such as cnn.com, ntimes.com, latimes.com are also some of the most widely viewed Web Sites in the data set. Looking back at the raw data, the special interest sub-category was largely driven by a small number of users who frequently visited Web Sites related to particular ethnic or social groups. Lastly, given that our partner company deals with international media, it is not surprising that there was a relatively high proportion of foreign related network traffic as users viewed Web Sites in their native languages or related to their countries of origin.

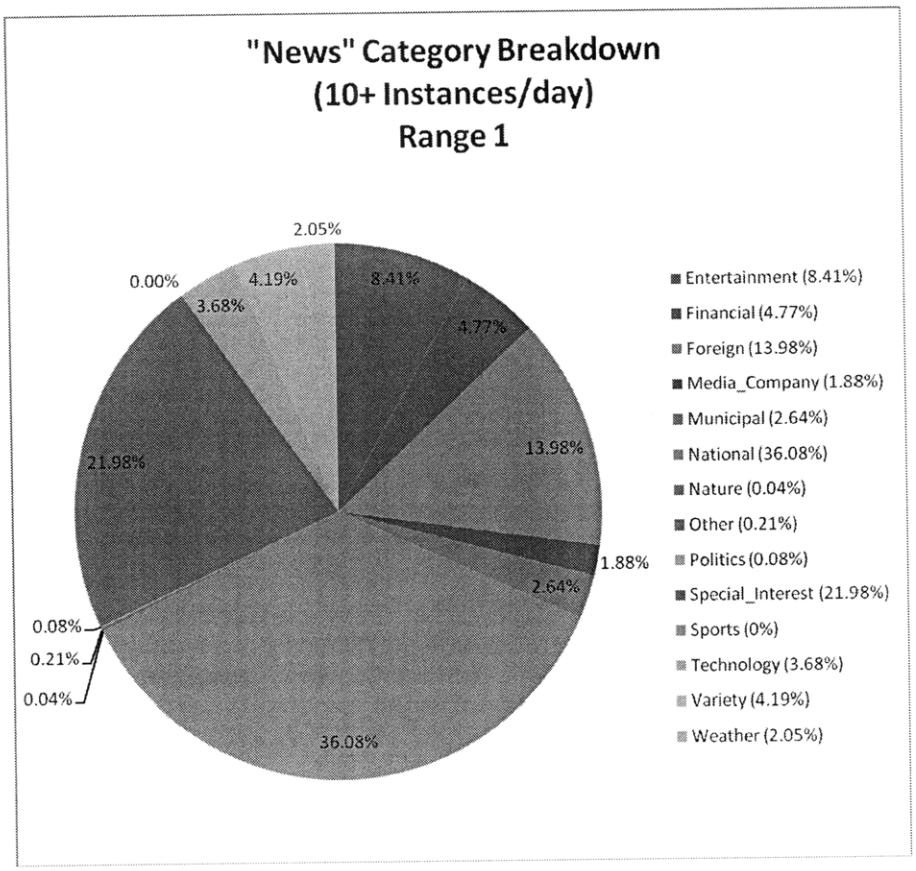

Figure 21 - "News" Category Breakdown. 10+ Instances/day. Range 1

For the Products category in Figure 22, network traffic is roughly split between a set of 4 larger sub-categories, and a set of 7 smaller sub-categories. The overall breakdown between these many categories showed that, in the aggregate, many users seek out information about a wide variety of different product types. Looking back over the raw data, this typically took the form of a handful of users who were very interested in a single area rather than all users being interested in all areas.

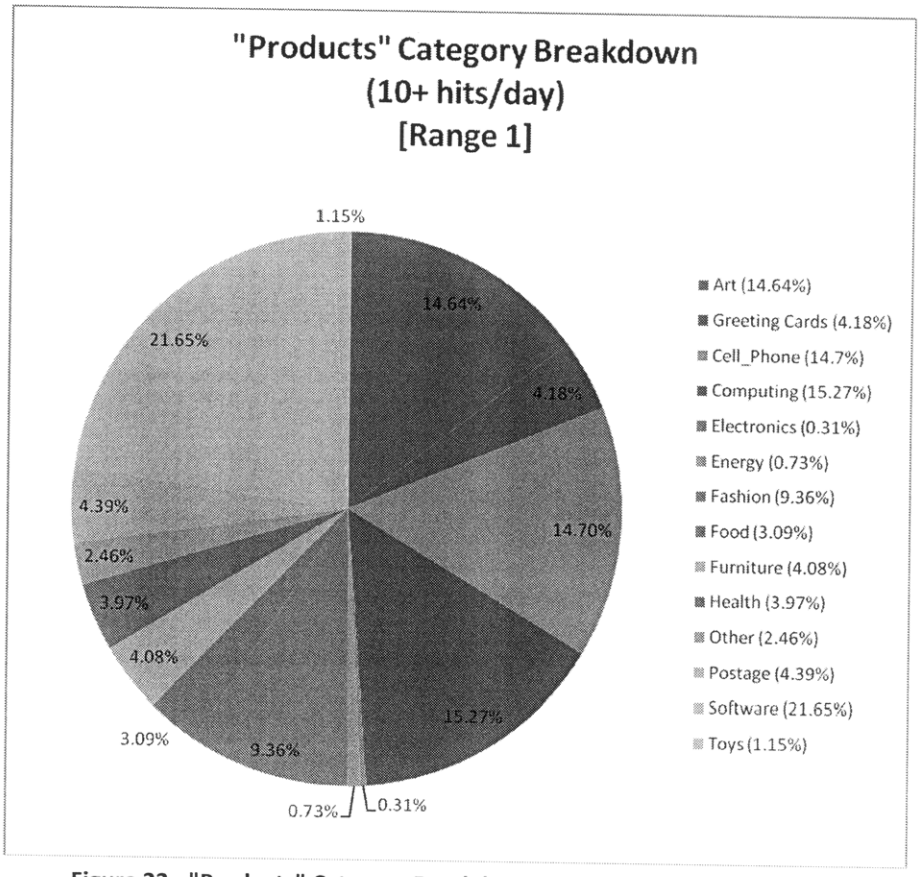

Figure 22 - "Products" Category Breakdown. 10+ Instances/day. Range 1

5.4 Top Web Sites

5.4.1 Overall

Table 14 shows the top 40 Web Sites in Range 1 by their total number of instances. As expected from our main category breakdown results, google.com and yahoo.com were on top with a collective 48.59% of the network traffic. The remaining Web Sites spanned across multiple main categories and consisted of many prominently known and popular Web Sites, such as youtube.com, wikimedia.org (wikipedia.org), amazon.com, ebay.com, and facebook.com. Given that there were over 1000 more Web Sites present in Range 1, this table shows how most of the network traffic, about 72%, was concentrated in 4% of the total number

of Web Sites present. The top 40 Web Sites for Range 2 can be found in Appendix D: Top Web Sites.

Top 40 Web Sites (Range 1)			
	Web Site	**# Instances**	**Percentage**
1	google.com	35475	28.45%
2	yahoo.com	25110	20.14%
3	youtube.com	12169	9.76%
4	webaccess.umail.ucsb.edu	8560	6.86%
5	google.com (email)	5195	4.17%
6	yahoo.com (email)	4016	3.22%
7	aol.com	3947	3.17%
8	sony.com	2766	2.22%
9	live.com	2697	2.16%
10	mailcenter.comcast.net	2354	1.89%
11	warnerbros.com	2039	1.64%
12	wikimedia.org	1897	1.52%
13	intdm.com	1726	1.38%
14	amazon.com	1688	1.35%
15	meebo.com	1530	1.23%
16	microsoft.com	1523	1.22%
17	live365.com	1137	0.91%
18	cnet.com	1098	0.88%
19	orb.com	984	0.79%
20	msn.com	924	0.74%
21	sbcglobal.net	889	0.71%
22	ign.com	775	0.62%
23	cnn.com	742	0.60%
24	ip.fastwebnet.it	730	0.59%
25	imeem.com	730	0.59%
26	mac.com	707	0.57%
27	ebay.com	705	0.57%
28	nonfatmedia.com	672	0.54%
29	groove.net	668	0.54%
30	myspace.com	636	0.51%
31	ebuddy.com	616	0.49%
32	aol.com (email)	604	0.48%
33	facebook.com	590	0.47%
34	pacbell.net	553	0.44%
35	craigslist.org	542	0.43%

36	userplane.com	532	0.43%
37	apartmenttherapy.com	524	0.42%
38	mediaserver.kataweb.it	495	0.40%
39	bankofamerica.com	494	0.40%
40	last.fm	436	0.35%
	Remaining	35205	28.23%

Table 14 - Top 40 Web Sites. Range 1

5.4.2 By Main Category

This section focuses on the five main categories of Commerce, Entertainment, Financial, News, and Social Networking. Remaining tables for Range 1 and for Range 2 can be found in Appendix D: Top Web Sites.

For the top 10 Web Sites in the Commerce category in Table 15, the Web Sites we would expect to be prominent made up most of the traffic, including amazon.com, ebay.com and craigslist.org. The remaining top Web Sites were the prominent shipping Web Sites and other specialized merchandise Web Sites.

'Commerce' Category Top Web Sites by Overall Instances							
	Range 1				Range 2		
	Web Site	**# Instances**	**Percentage**		**Web Site**	**# Instances**	**Percentage**
1	amazon.com	1688	34.48%	**1**	amazon.com	2186	43.26%
2	ebay.com	705	14.40%	**2**	ebay.com	858	16.98%
3	craigslist.org	542	11.07%	**3**	craigslist.org	506	10.01%
4	paypal.com	170	3.47%	**4**	paypal.com	139	2.75%
5	slickdeals.net	169	3.45%	**5**	slickdeals.net	99	1.96%
6	costco.com	124	2.53%	**6**	fedex.com	98	1.94%
7	ups.com	104	2.12%	**7**	revolveclothing.com	81	1.60%
8	startrek.com	103	2.10%	**8**	dhl-usa.com	79	1.56%
9	dhl-usa.com	88	1.80%	**9**	costco.com	71	1.41%
10	thinkgeek.com	74	1.51%	**10**	ioffer.com	70	1.39%
	remaining(58)	1128	23.04%		remaining(58)	866	17.14%

Table 15 - 'Commerce' Category Top Web Sites by Overall Instances

For the top 10 Web Sites in the Entertainment category in Table 16, the Web Site we would expect to be first, youtube.com, was indeed first. The remaining top Web Sites included several popular streaming music Web Sites. The other media related Web Sites in the listing likely appeared because of the nature of our partner company's work.

'Entertainment' Category Top Web Sites by Overall Instances							
	Range 1				Range 2		
	Web Site	# Instances	Percentage		Web Site	# Instances	Percentage
1	youtube.com	12169	45.25%	1	youtube.com	8509	44.32%
2	sony.com	2766	10.29%	2	imeem.com	2188	11.40%
3	warnerbros.com	2039	7.58%	3	warnerbros.com	2117	11.03%
4	live365.com	1137	4.23%	4	sony.com	1523	7.93%
5	orb.com	984	3.66%	5	live365.com	586	3.05%
6	ign.com	775	2.88%	6	scopeseven.com	258	1.34%
7	imeem.com	730	2.71%	7	dailymotion.com	246	1.28%
8	last.fm	436	1.62%	8	netflix.com	215	1.12%
9	dailymotion.com	338	1.26%	9	mediatakeout.com	210	1.09%
10	warfish.net	309	1.15%	10	volia.net	178	0.93%
	remaining(205)	5209	19.37%		remaining(205)	3170	16.51%

Table 16 - 'Entertainment' Category Top Web Sites by Overall Instances

For the top 10 Web Sites in the Financial category in Table 17, Bankofamerica.com, the Web Site for one of the nation's most popular banks, came in first with another banking Web Site, wellsfargo.com, in second place. The remaining Web Sites were split between banking, investing, and personal finance Web Sites.

'Financial' Category Top Web Sites by Overall Instances							
	Range 1				Range 2		
	Web Site	# Instances	Percentage		Web Site	# Instances	Percentage
1	bankofamerica.com	494	23.31%	1	bankofamerica.com	442	21.49%
2	wellsfargo.com	366	17.27%	2	wellsfargo.com	244	11.86%
3	paypal.com	170	8.02%	3	fidelity.com	214	10.40%
4	wamu.com	142	6.70%	4	wamu.com	158	7.68%
5	ta-retirement.com	124	5.85%	5	paypal.com	139	6.76%
6	hsbccreditcard.com	86	4.06%	6	ta-retirement.com	130	6.32%
7	fidelity.com	75	3.54%	7	yodlee.com	70	3.40%

8	washingtonmutualfinance.org	53	2.50%	8	transaccessonline.com	51	2.48%
9	firstent.org	45	2.12%	9	washingtonmutualfinance.org	48	2.33%
10	bigcharts.com	44	2.08%	10	mybills.com	45	2.19%
	remaining(33)	520	24.54%		remaining(33)	516	25.09%

Table 17 - 'Financial' Category Top Web Sites by Overall Instances

For the top 10 Web Sites in the News category in Table 18, the very prominent cnn.com Web Site was the top Web Site. The remaining Web Sites were mostly for other national or international news Web Sites, or for specific purposes like entertainment, finance, or special interest.

'News' Category Top Web Sites by Overall Instances							
	Range 1				Range 2		
	Web Site	# Instances	Percentage		Web Site	# Instances	Percentage
1	cnn.com	742	21.04%	1	cnn.com	1205	30.31%
2	datalounge.com	413	11.71%	2	bbc.co.uk	314	7.90%
3	msnbc.com	371	10.52%	3	nationalgeographic.com	277	6.97%
4	bbc.co.uk	269	7.63%	4	msnbc.com	226	5.68%
5	eonline.com	188	5.33%	5	mediatakeout.com	210	5.28%
6	mediatakeout.com	187	5.30%	6	datalounge.com	182	4.58%
7	washingtonpost.com	146	4.14%	7	sandrarose.com	163	4.10%
8	guardian.co.uk	106	3.01%	8	bossip.com	139	3.50%
9	forbes.com	104	2.95%	9	guardian.co.uk	109	2.74%
10	nytimes.com	81	2.30%	10	anapixelbsl.elmundo.es	90	2.26%
	remaining(44)	919	26.06%		remaining(44)	1061	26.69%

Table 18 - 'News' Category Top Web Sites by Overall Instances

For the top 10 Web Sites in the Social Networking category in Table 19, meebo.com, an online communications platform that supports instant messaging services like AIM and Yahoo, took first place. Closely following was immem.com, a social networking and music streaming Web Site, and the more typical social networking Web Sites, myspace.com, facebook.com, ebuddy.com, and linkedin.com.

'Social Networking' Category Top Web Sites by Overall Instances			
	Range 1		Range 2

	Web Site	# Instances	Percentage		Web Site	# Instances	Percentage
1	meebo.com	1530	27.76%	1	meebo.com	3555	39.35%
2	imeem.com	730	13.24%	2	imeem.com	2188	24.22%
3	myspace.com	636	11.54%	3	facebook.com	1879	20.80%
4	ebuddy.com	616	11.18%	4	myspace.com	379	4.19%
5	facebook.com	590	10.70%	5	migente.com	187	2.07%
6	linkedin.com	183	3.32%	6	linkedin.com	126	1.39%
7	migente.com	144	2.61%	7	digg.com	116	1.28%
8	digg.com	121	2.20%	8	ebuddy.com	87	0.96%
9	plaxo.com	114	2.07%	9	buzznet.com	82	0.91%
10	blackpeoplemeet.com	78	1.42%	10	xanga.com	48	0.53%
	remaining(31)	770	13.97%		remaining(31)	388	4.29%

Table 19 - 'Social Networking' Category Top Web Sites by Overall Instances

5.4.3 Unique Source Macs for Top Web Sites

In Table 20, we list the top 40 Web Sites for Range 1, ordered by the total number of instances that the Web Site was visited, alongside the number of unique users who visited that Web Site. As expected, many users have visited popular Web Sites like google.com, yahoo.com, youtube.com, amazon.com, and wikimedia.org (wikipedia.org). Smaller cluster of users have visited slightly less popular Web Sites like live.com, cnet.com, ebay.com and Microsoft.com. Lastly, for some Web Sites, only a handful of users were driving all the network traffic to that Web Site. For example, meebo.com was only used by 8 users. However, since meebo.com is a chat application, it made sense that it would be in the top 40 for network traffic since users probably left this chat application open at work for long periods of time. Similarly, orb.com is a Web Site that allows users to reach media on their home systems from other locations. Thus, the single user using orb.com generated high amounts of network traffic over the data set. A listing of top Web Sites for Range 2 can be found in Appendix D: Top Web Sites.

Unique User Statistics for Top 40 Web Sites for Range 1			
Web Site	Main Category	# Instances	# of Unique Users

google.com	Search	35475	229
yahoo.com	Search	25110	213
youtube.com	Entertainment	12169	179
webaccess.umail.ucsb.edu	Web Portal & Email	8560	5
google.com (email)	Web Portal & Email	5195	191
comcast.net	Access	4381	71
yahoo.com (email)	Web Portal & Email	4016	107
aol.com	Web Portal & Email	3947	151
sony.com	Entertainment	2766	34
live.com	Search	2697	83
mailcenter.comcast.net	Web Portal & Email	2354	2
rr.com	Access	2244	91
warnerbros.com	Entertainment	2039	12
wikimedia.org	Education	1897	152
charter.com	Access	1819	61
intdm.com	Professional Services	1726	3
amazon.com	Commerce	1688	144
meebo.com	Social Networking	1530	8
microsoft.com	Internet Services & Software	1523	91
pccwglobal.net	Access	1343	131
live365.com	Entertainment	1137	11
cnet.com	Informational	1098	116
orb.com	Entertainment	984	1
msn.com	Web Portal & Email	924	121
sbcglobal.net	Web Portal & Email	889	43
cox.net	Access	846	62
mpowercom.net	Access	815	23
ign.com	Entertainment	775	42
cnn.com	News	742	68
imeem.com	Entertainment	730	45
ip.fastwebnet.it	Web Portal & Email	730	6
metrored.net.mx	Access	725	7
mac.com	Web Portal & Email	707	47
ebay.com	Commerce	705	92
nonfatmedia.com	Professional Services	672	5
groove.net	Internet Services & Software	668	33
myspace.com	Social Networking	636	62
ebuddy.com	Social Networking	616	3
aol.com (email)	Web Portal & Email	604	23

Table 20 - Unique User Statistics for Top 40 Web Sites for Range 1

5.4.4 Unique Users vs Instances

The following 4 figures are graphs of the number of unique users for a Web Site vs the number of instances that a Web Site appeared in the network traffic data set.

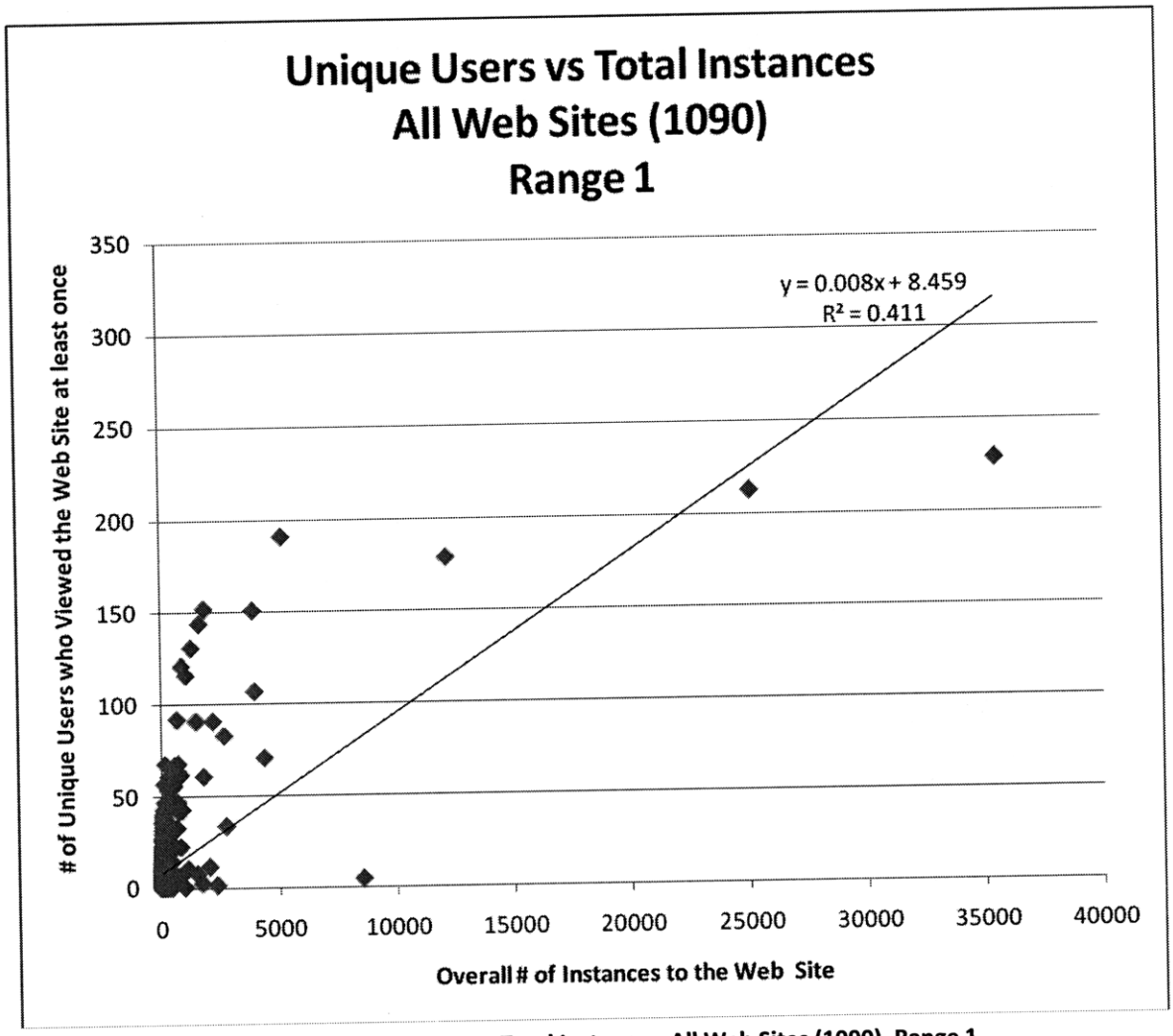

Figure 23 - Unique Users vs Total Instances. All Web Sites (1090). Range 1

Figure 23 displays all Web Sites in the data set. The two prominent outliers to the right are google.com and yahoo.com. Youtube.com is third from the right and webaccess.umail.ucsb.edu is fourth from the right. This forth Web Site seemed out of place. Looking at the raw data, we noticed that a small handful of users had massive amounts of network traffic to the UCSB webmail servers over a small period of time. We reasoned that they must have been

transferring files online via email. The large majority of Web Sites clustered around the bottom left corner and indicated Web Sites that a small number of users infrequently visit.

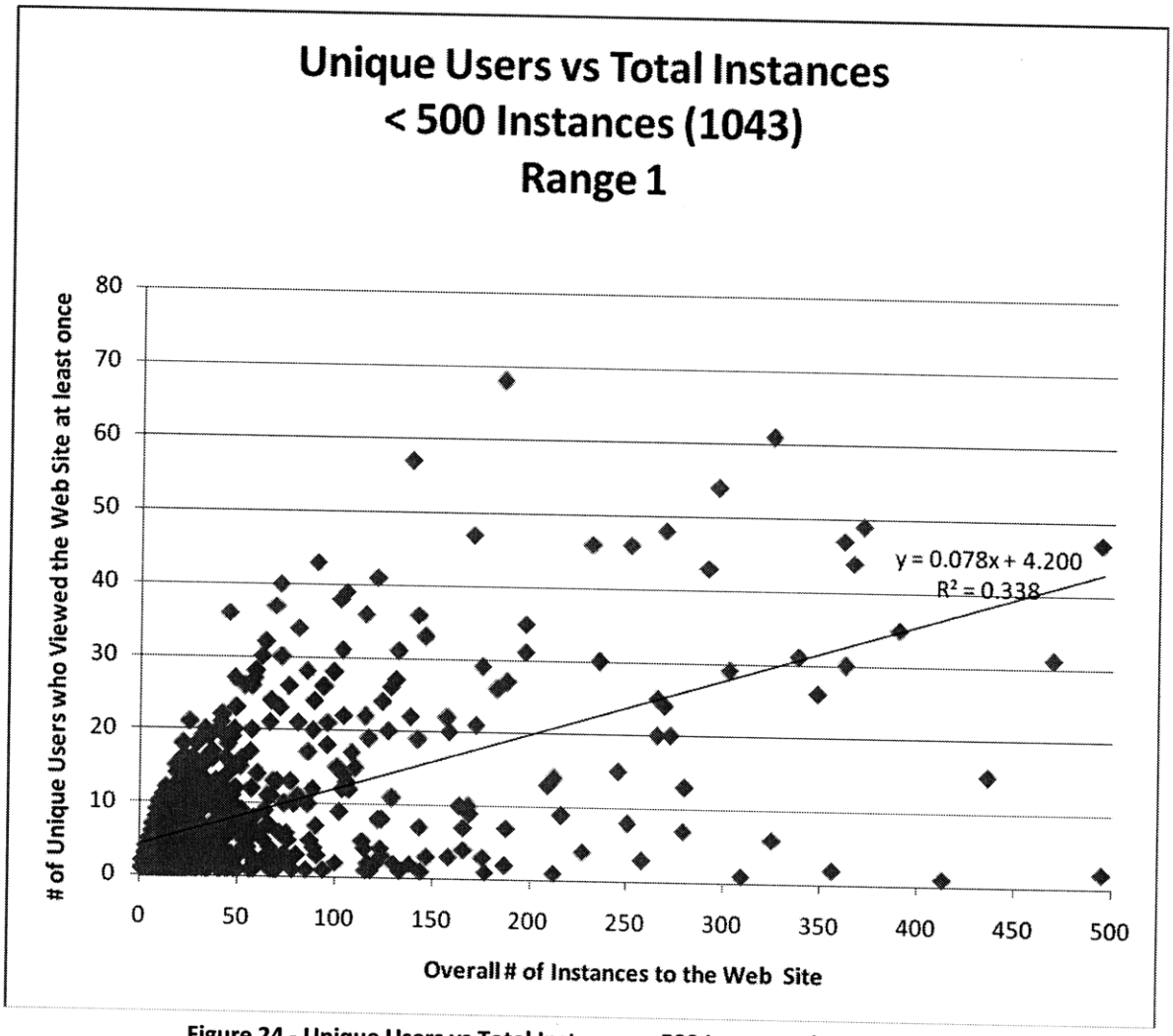

Figure 24 - Unique Users vs Total Instances. <500 Instances (1043). Range 1

To gain a better view of the bottom left quadrant, Figure 24 looks at Web Sites with less than 500 total instances in the data set. Once again, most of the 1043 Web Sites crowded around the bottom left corner. However, slightly more than 100 Web Sites were outside of that area, representing Web Sites that were popular across many users.

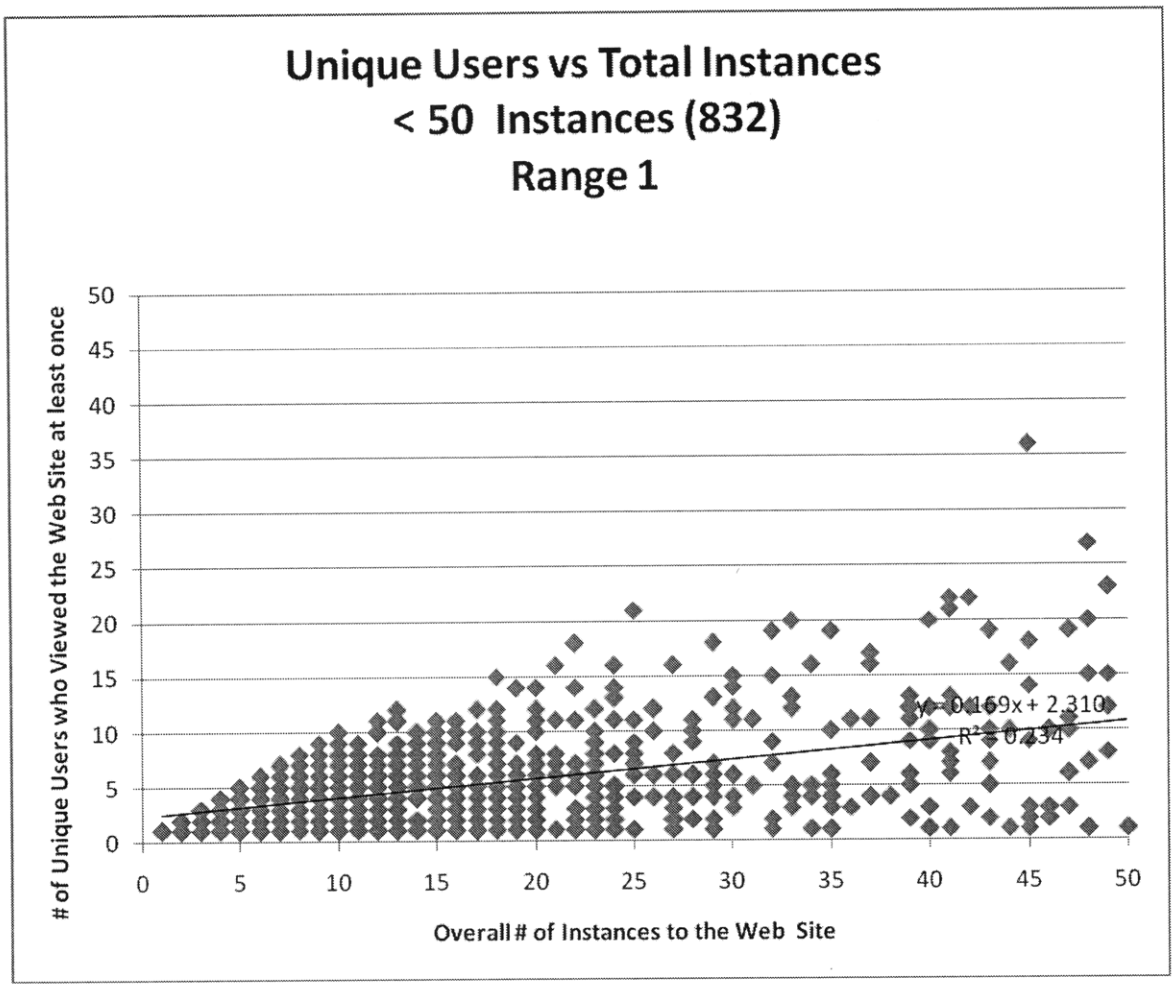

Figure 25 - Unique Users vs Total Instances. <50 Instances (832). Range 1

Figure 25 looks at Web Sites with less than 50 instances in the data set. This figure cannot portray a third dimension, namely the number of Web Sites that contained the same combination of instances and unique users, but the dots in the bottom left quadrant represent many Web Sites each. Above 13 instances on the Web Site, most Web Sites had more instances than viewers, representing the concentrating of network traffic to a small subset of users who viewed a particular Web Site. Given that a popular Web Site would not only be viewed by more users but also be viewed more times in general, it was not surprising that the Web Sites in this figure rarely exceeded 15 unique visitors.

Ultimately, the trend expressed in all of these figures was that a small number of users typically dominated the Web Site activity of most Web Sites. Exceptions to this pattern include the widely popular Web Sites which are viewed by more users and viewed more often overall per user. Remaining figures for Range 1 and Range 2 can be found in Appendix D: Top Web Sites.

5.5 Diversity

5.5.1 Background

The Herfindahl index is an economic metric that represents the amount of competition between firms in a particular industry. A low Herfindahl index represents an industry in which many firms equally share the market. A high Herfindahl index represents an industry in which a single firm has a monopoly on the market. By viewing Web Site categories as unique firms and the number of instances of a category as the amount of market share a category has, we used the Herfindahl index to quantitatively measure the degree to which a user's Web Site visitation habits were diverse (spread equally between multiple categories).

The formula for diversity, based on the Herfindahl Index, is: $1 - \sum_{i=1}^{N} p_i^2$. In this formula, N is the total number of categories of Web Sites and p_i is the fractional proportion of Web Site i. The resulting value ranges from 0 to 1 – 1/n. A value of 0 indicates no diversity. An example of this would be a user who only visited news Web Sites. Higher values indicate increasing levels of diversity. For the purpose of our calculations, we normalized this value so that it always ranges from 0 to 1 with 0 being no diversity and 1 being complete diversity. An example of

complete diversity is a user who equally splits his Web Site views between all Web Site categories.

By calculating the diversity of a user's Web Site habits, we were able to quantitatively measure the degree to which users sought out information from a variety of different sources. Future researchers can use the methods presented here to profile information workers and relate their productivity to their Web Site diversity.

5.5.2 Overall Diversity

Level	Range 1	Range 2
50+ Instances/day	79.39%	92.09%
20+ Instances/day	86.31%	89.48%
10+ Instances/day	85.68%	88.30%
5+ Instances/day	85.81%	88.41%
All	86.19%	88.43%

Table 21 - Overall Diversity of Network Traffic

Table 21 lists the overall diversity of categories of the network traffic data set. We calculated the diversity score based on different levels of user inclusion. Each row represents a different cutoff on the average number of daily instances over the data set. Cutoffs of 20, 10, 5, and 0 produced strongly similar overall diversity values while a cutoff of 50 produces a noticeable different diversity value. The following section explains the reason for this difference by looking at cumulative daily diversity.

5.5.3 Cumulative Daily Diversity

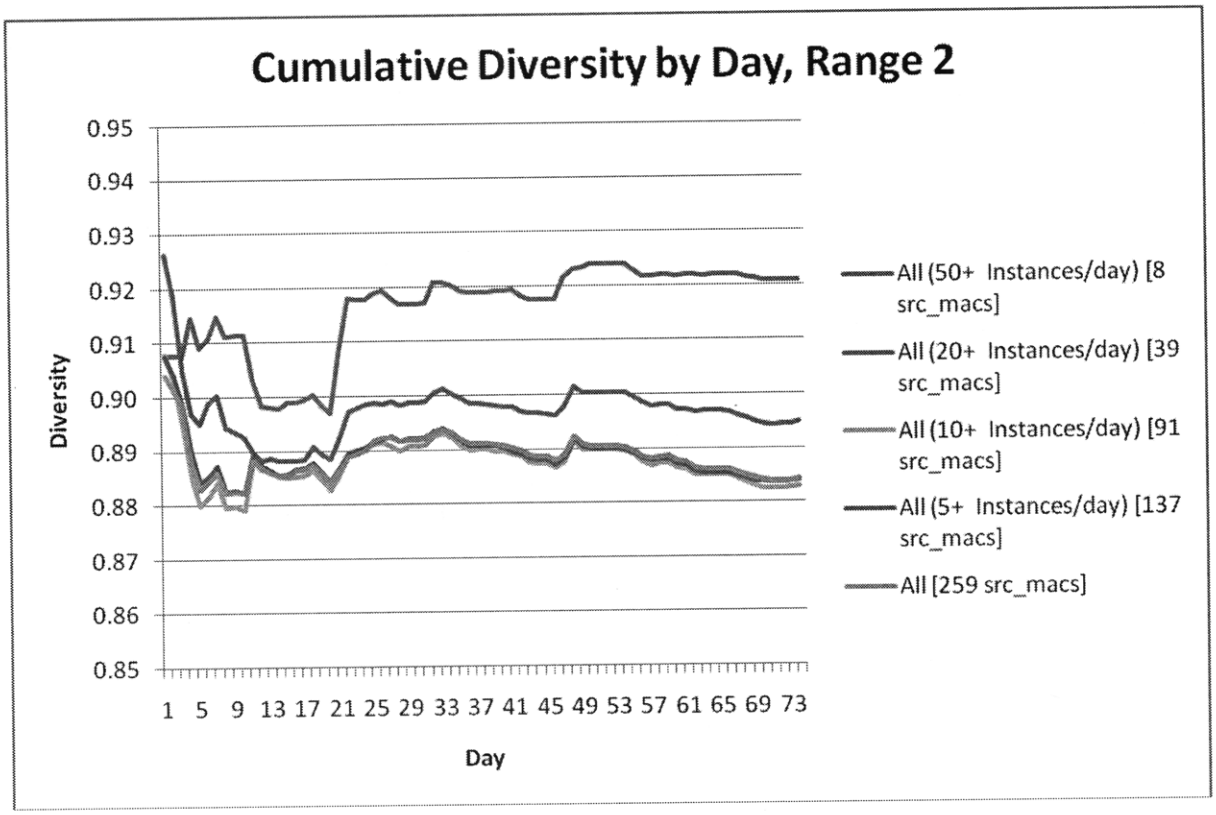

Figure 26 - Cumulative Diversity by Day, Range 2

Figure 29 shows the cumulative diversity of Range 2 on a day-by-day basis. For a given day, we calculated the cumulative diversity by considering all the traffic on that day and all the days before it. As expected from the use of the sFlow sampling technique to capture network traffic, it took several days for this metric to stabilize over the data set. Over the first few weeks, the diversity fluctuateed, driven by instances of a few users having network traffic heavily focused in one main category or on one Web Site. As time continued, the diversity values settled and approached their final values.

Another feature to note is that the cumulative diversity metrics for cutoff values of 0, 5, and 10 all remain very close throughout the data range. But the cumulative diversity metric with a cutoff

value of 50. The problem with the cutoff value of 50 was that too few users were considered in the calculation. Even though they made up a large portion of the network traffic, the resulting diversity values were noticeable off from the overall diversity values for all more users. For the same reasons, the cutoff value of 20 also starts to deviate. Overall, to arrive at diversity values that properly represent the entire user population, we had to include users with at least 10 instances of network traffic per day. The lower thresholds added successively less instances to the data set and thus could be approximated accurately by a cutoff value of 10.

The remaining breakdowns for the other main categories can be found in Appendix F: Diversity.

5.5.4 Overall Category Diversity

Main Category	Range 1	Range 2
Access	91.47%	86.85%
Blogging	88.19%	75.12%
Commerce	86.07%	78.49%
Education	88.20%	89.10%
Entertainment	77.72%	77.43%
Financial	91.73%	93.16%
Government	91.88%	84.74%
Informational	76.62%	68.03%
Internet Services & Software	94.12%	96.15%
News	92.75%	89.38%
Products	95.81%	73.18%
Professional Services	83.18%	72.76%
Search	56.83%	58.67%
Social Networking	88.64%	76.10%
Travel & Transportation	94.40%	94.69%
Web Portal & Email	88.41%	89.59%

Table 22 – Overall Diversity of Network Traffic by Diversity

Table 23 lists the overall diversity of each category of the network traffic data set. This calculation was of the normalized Herfindahl index with consideration of all the Web Sites that

fell within a certain category. Specifically, we formed this calculation by considering each possible Web Site in a category as one entity and calculating the diversity of the category as a collection of various numbers (instances) of those entities (Web Sites).

All of the main categories had a differing number of total unique Web Sites. As a side-effect of this changing number, it was more difficult to achieve a high diversity in a category with lots of possible Web Sites than it was in a category with much fewer Web Sites. For example, given the Blogging category with less than 15 Web Sites and the Entertainment with over 100 Web Sites, the situation with exactly 5 Web Sites evenly splitting 100 instances of network traffic will result in a higher diversity score for the Blogging category than for the Entertainment category. In the Blogging category case, these 5 Web Sites make up 33% of the total possible Web Sites, so they have a greater impact on the diversity score than 5 Web Sites do in the Entertainment category that has over 100 Web Sites. Overall, the Search category has the lowest diversity score. This occurred because, while there are many search engines on the Internet, google.com and yahoo.com tend to monopolize the Search category.

5.6 Page Rank

5.6.1 Background

Google's page rank is a numerical value that Google assigns to Web Sites to help create search results. This score represents how popular a Web Site is based on the degree to which other popular Web Sites link to it (Rogers, 2002). A score of 10, the highest possible score, represents a very well referenced and important Web Site and a score of 0, the lowest possible score, represents a very poorly referenced and unimportant Web Site.

By gathering the scores for all the Web Sites in the data set, we were able to quantitatively measure the relevance of a user's Internet browsing habits as well as the overall relevance of different categories of Web Sites. Future researchers can use the methods presented here to profile information workers and relate their productivity to the relevance of the Web Sites they frequent.

5.6.2 Overall for Entire Data Set

Range 1	Range 2	Range 1+2
7.84	7.83	7.834

Table 23 - Overall Weighted Average Page Rank of Visited Web Sites

Table 23 lists the overall Page Rank score of the Web Sites in our data set. To put this into perspective, google.com, real.com, and w3.org were the only 10s in our data set. There were 63 9s in our data set including such popular Web Sites as yahoo.com, youtube.com, facebook.com, cnn.com, wikimedia.org (wikipedia.org) nytimes.com, apple.com, slashdot.com, ibm.com, ucsc.edu, yale.edu, and amazon.com. There were 140 8s in the data set with prominent Web Sites including ebay.com, cnet.com, craigslist.com, myspace.com, hostmail.com, usc.edu, linkedin.com, fedex.com, delta.com and ticketmaster.com. There were 219 7s in the data set with prominent Web Sites including meebo.com, netflix.com, warnerbros.com, southwest.com, walmart.com, blackberry.com, and bestbuy.com. There were 253 6s in order data set with prominent Web Sites including imeem.com, verizon.net, virginamerica.com, and virtualearth.net. Lastly, there were 179 5s, 84 4s, 47 3s, 12 2s, 7 1s, and 94 0s in the data set. Web Sites with page ranks lower than 6 tended to be much more obscure and less popular than Web Site with higher page ranks.

5.6.3 Main Category Breakdowns

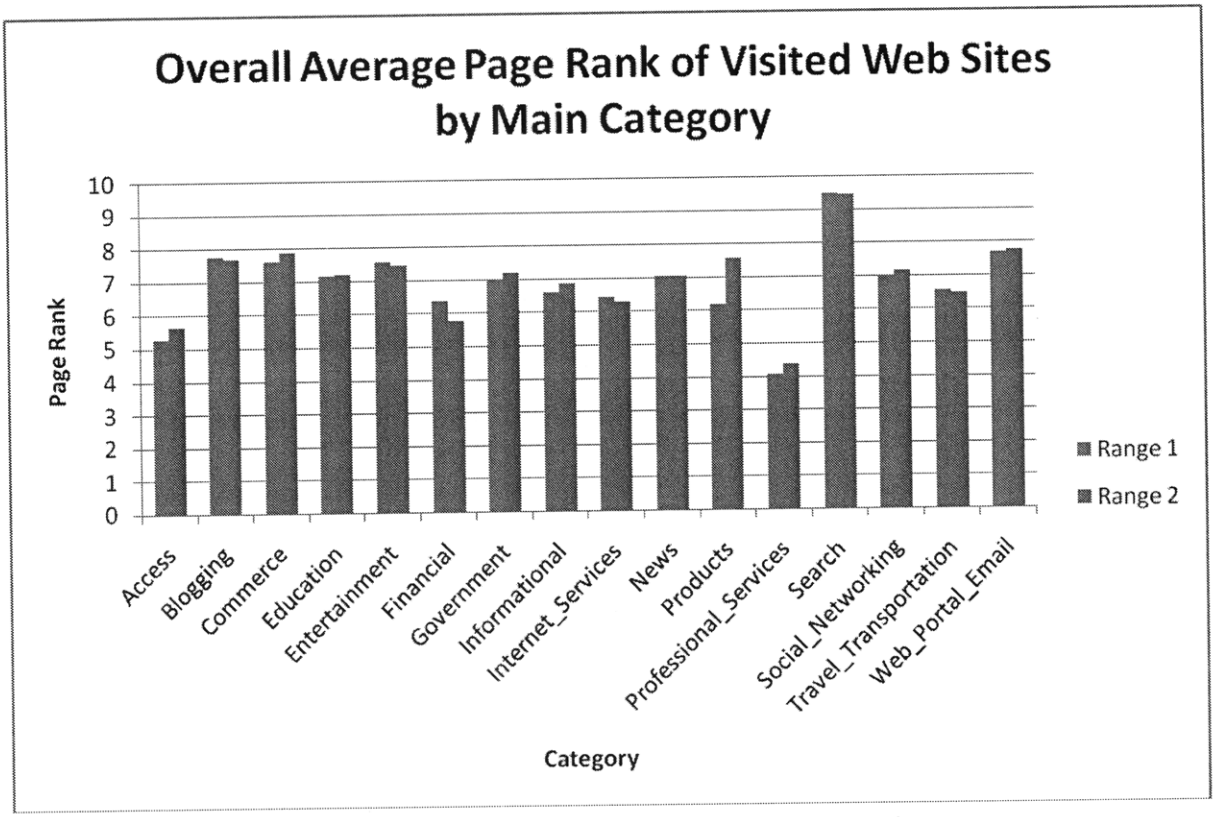

Figure 27 - Overall Average Page Rank of Visited Web Sites by Main Category

In Figure 27, the overall page ranks by category are shown for Range 1 and Range 2. Given that google.com and yahoo.com made up a large proportion of this traffic and have page ranks of 10 and 9 respectively, it was not surprising that the Search category had the highest page rank. Additionally, since a large portion of the data set consisted of Search, it is easy to see how the overall page ranks from Table 23 were above 7.8 when only 3 total categories had averages above 7.8.

Eight of the other categories had an average page rank of at least 7. These categories were Blogging, Commerce, Education, Entertainment, Government, News, Social Networking, and Web Portal & Email. Seven of the categories had an average page rank less than 7. These categories were Access, Financial, Informational, Internet Services & Software, Products,

Professional Services, and Travel & Transportation. What this roughly indicated was that the latter categories were often more specialized and had Web Sites that were less well-known than Web Sites that fell into the former categories. Professional Services was the lowest ranked category. We reasoned that this occurred because of the niche nature and smaller audience size of the professional services Web Sites present in the data set. For example, Web Sites related to specialized firms in specific industries, like design or fashion, are going to be less popular among the general public than more common search and news related Web Sites that everyone often uses.

5.6.4 Sub-Category Breakdowns

Table 24 lists the page rank of the sub-categories for the Commerce, Entertainment, Financial, Informational, News, Products, and Social Networking main categories. One important feature to note is that, even within the same main category, different sub-categories had drastically different average page ranks. For example, in the News category, national news had an overall page rank of 8.90 while special interest news had an overall page rank of 5.00. This seems reasonable since special interest and niche Web Sites are inherently less popular by their nature. Similarly, housing related commerce Web Sites had an overall lower page rank than often used variety related commerce Web Sites like amazon.com and ebay.com.

Given that many main categories have sub-categories with low page ranks, it is clear that most of the instances within a main category fell into a few high page rank sub-categories and supported the overall main category page rank averages in Figure 27.

The remaining breakdowns for the other main categories can be found in Appendix E: Page Rank.

Average Page Rank of Visited Web Sites by Sub-Category				
Main Category	**Sub-Category**	**Range 1**	**Range 2**	**Range 1+2**
Commerce	Communication	7.00	7.00	7.00
Commerce	Computing	5.90	5.81	5.87
Commerce	Entertainment	6.29	6.15	6.26
Commerce	Fashion	5.10	4.75	4.83
Commerce	Financial Processing	7.99	7.90	7.95
Commerce	Foreign	5.00	n/a	5.00
Commerce	Housing	3.58	4.29	3.77
Commerce	Other	3.93	6.05	4.69
Commerce	Shipping	7.72	7.69	7.70
Commerce	Sports	n/a	5.00	5.00
Commerce	Variety	8.06	8.19	8.13
Commerce	Vehicles	6.76	5.03	5.48
Entertainment	Adult	0.00	n/a	0.00
Entertainment	Art	6.00	6.00	6.00
Entertainment	Arts	5.88	6.34	6.01
Entertainment	Comics	6.81	6.21	6.57
Entertainment	Dance	3.00	3.00	3.00
Entertainment	Film	6.23	6.26	6.24
Entertainment	Foreign	3.19	4.10	3.54
Entertainment	Games	5.84	5.52	5.77
Entertainment	Media Broad	7.56	7.40	7.49
Entertainment	Music	6.00	5.93	5.97
Entertainment	Other	5.45	4.25	5.13
Entertainment	Pop Culture	5.26	5.82	5.54
Entertainment	Special Interest	5.19	5.37	5.30
Entertainment	Sports	6.16	5.92	6.11
Entertainment	Television	5.49	4.80	5.06
Entertainment	Theater	6.37	6.60	6.47
Entertainment	Ticketing	5.36	5.66	5.49
Entertainment	Video	8.90	8.85	8.88
Financial	Banking	7.07	7.14	7.10
Financial	Credit	5.19	5.93	5.50
Financial	Informational	3.93	3.31	3.60
Financial	Insurance	5.00	5.00	5.00
Financial	Investing	3.10	1.91	2.36
Financial	Loans	6.00	6.00	6.00
Financial	Other	8.00	8.00	8.00
Financial	Payment Services	7.90	8.00	7.94
Financial	Software	7.23	6.48	6.64

Financial	Stocks	6.60	6.17	6.38
Informational	Arts	4.00	4.00	4.00
Informational	Foreign	7.00	n/a	7.00
Informational	General	7.75	7.87	7.82
Informational	Other	6.32	6.00	6.23
Informational	Societies	1.89	4.32	2.99
Informational	Special Interest	5.59	5.43	5.55
Informational	Technology	7.74	7.91	7.81
News	Entertainment	6.80	6.25	6.56
News	Financial	7.11	7.10	7.11
News	Foreign	6.06	5.11	5.48
News	Media Company	6.96	7.43	7.16
News	Municipal	5.98	5.50	5.73
News	National	8.87	8.92	8.90
News	Nature	8.00	8.00	8.00
News	Other	5.00	5.00	5.00
News	Politics	7.00	7.00	7.00
News	Special Interest	4.80	5.24	5.00
News	Technology	7.12	8.48	7.40
News	Variety	4.33	4.41	4.37
News	Weather	7.88	7.88	7.88
Products	Art	6.91	6.95	6.93
Products	Cards	6.00	6.00	6.00
Products	Cell Phone	7.00	7.00	7.00
Products	Computing	7.26	8.74	8.34
Products	Electronics	4.71	5.16	4.89
Products	Energy	6.69	6.20	6.50
Products	Fashion	4.97	4.97	4.97
Products	Food	5.25	5.15	5.23
Products	Furniture	4.12	5.83	5.08
Products	Health	6.37	6.41	6.38
Products	Other	6.86	6.77	6.84
Products	Postage	1.72	0.00	1.33
Products	Software	6.06	6.31	6.16
Products	Toys	5.82	5.00	5.57
Social Networking	Blogging	7.77	7.38	7.64
Social Networking	Communication	6.99	6.98	6.98
Social Networking	General	8.01	8.72	8.44
Social Networking	Links	7.83	7.58	7.71
Social Networking	Music	6.00	6.00	6.00
Social Networking	Other	3.78	4.74	4.29

Social Networking	Special Interest	5.42	5.93	5.63

Table 24 - Average Page Rank of Visited Web Sites by Sub-Category

5.6.5 Top Users

Figure 28 shows the top 40 users by instances in the data set verses their average page rank for Range 1. As shown in the figure, most users had a page rank between 6 and 9. As the number of instances in the data set decreased, the spread tended to widen and users became just as likely to have a low or high page rank. Overall, there was not a strong relationship between number of instances in the data set and page rank that would allow us to predict one's average page rank from the statistics of these top 40 users.

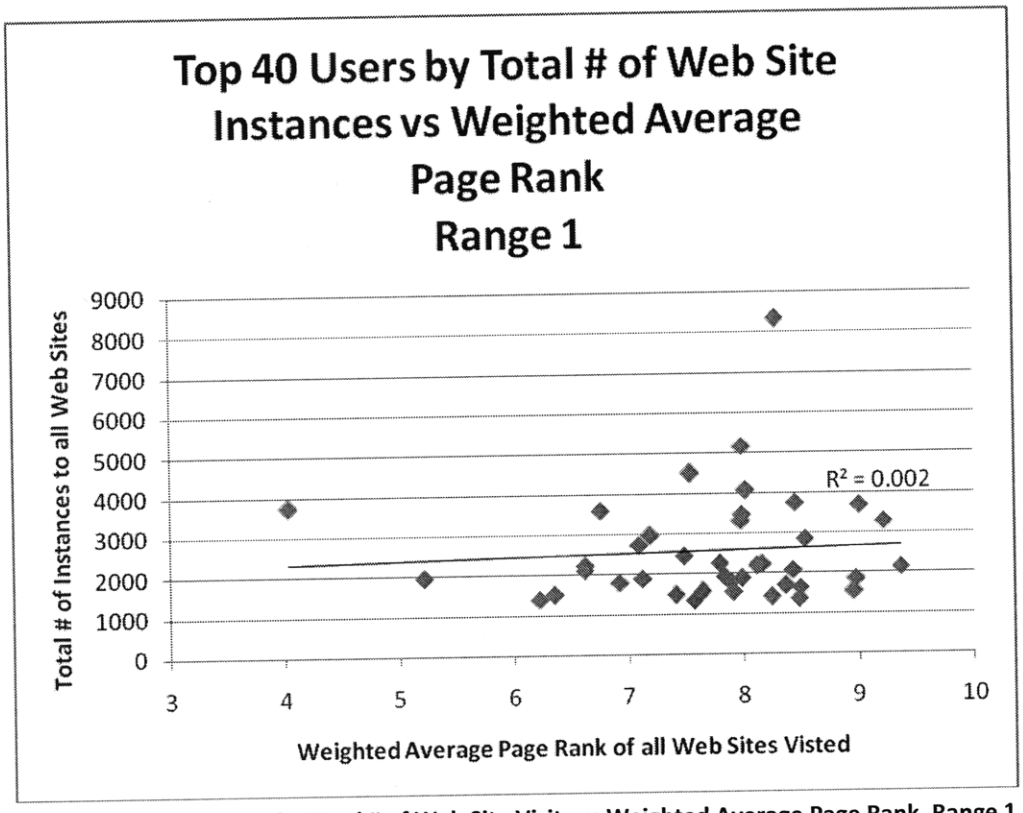

Figure 28 - Top 40 Users by Total # of Web Site Visits vs Weighted Average Page Rank. Range 1

Figure 29 shows the top 40 users by number of unique Web Sites visited in the data set verses their average page rank. Similar to Figure 28, most users had a page rank between 6 and

8.5. As the number of instances in the data set decreased, the spread tended to widen and users became just as likely to have a low or high page rank. Overall, there was not a strong relationship between number of unique Web Sites visited and page rank that would allow us to predict one's average page rank by the number of unique Web Sites accessed by the top 40 users.

The remaining breakdowns for the other main categories can be found in Appendix E: Page Rank.

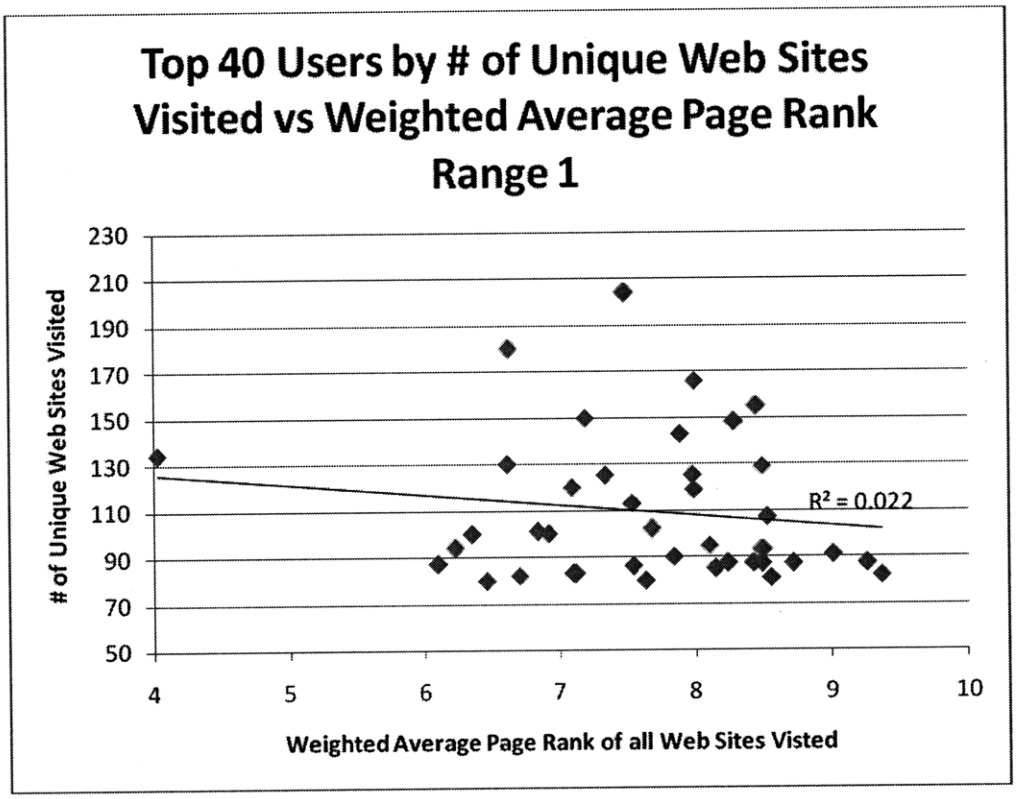

Figure 29 - Top 40 Users by # of Unique Web Sites Visited vs Weighted Average Page Rank. Range 1

5.7 Predictive Statistics and Correlations

5.7.1 Overall Page Rank vs Overall Diversity

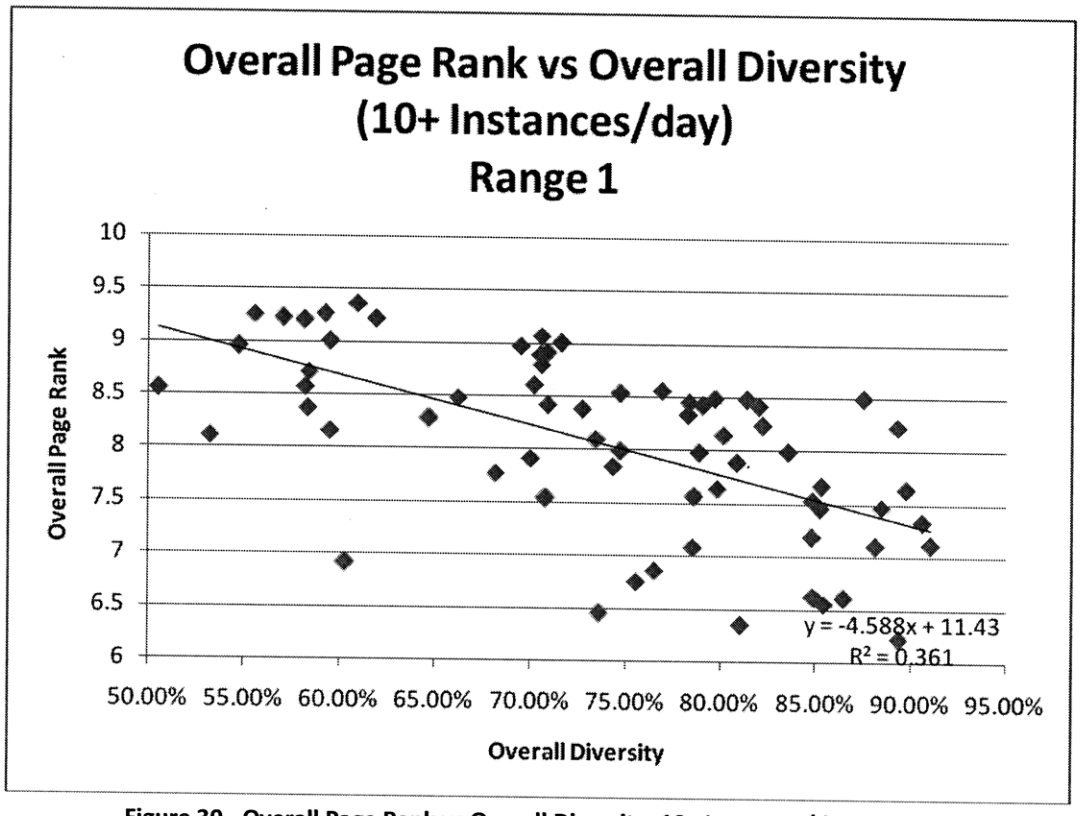

Figure 30 - Overall Page Rank vs Overall Diversity. 10+ Instances/day. Range 1

In Figure 30 and Figure 31, we have plotted a linear regression of overall page rank vs overall diversity for users in Range 1 and Range 2, respectively, with at least 10 instances of network traffic in the data set. We removed 14 outliers from Range 1 with low page ranks below 6 and low diversity values below 50%, leaving 66 sample points. We removed 18 outliers with low page ranks or low diversity values from Range 2, leaving 73 data points. Overall, there was an inverse relationship between overall page rank and overall diversity. As a user's diversity increases, his page rank tended to decrease. Theoretically, this made sense as users with more diverse content had a more equal share of network traffic to all the Web Site categories,

including the categories that were more obscure and tended to have lower page ranks. Being able to seek out information on the limitless Internet has allowed for what is known as the "Long Tail" phenomenon and has likely contributed to this relationship between diversity and page rank (Anderson, 2004). In a user's initial search, he likely first visits the most popular Web Sites on a subject. However, the more obscure Web Sites are only a few lines away in the search results, leading a user's search to the more obscure Web Sites which also have lower page ranks. Thus, users who look around more often, having higher diversity, tend to achieve lower average page rank among the Web Sites they visit.

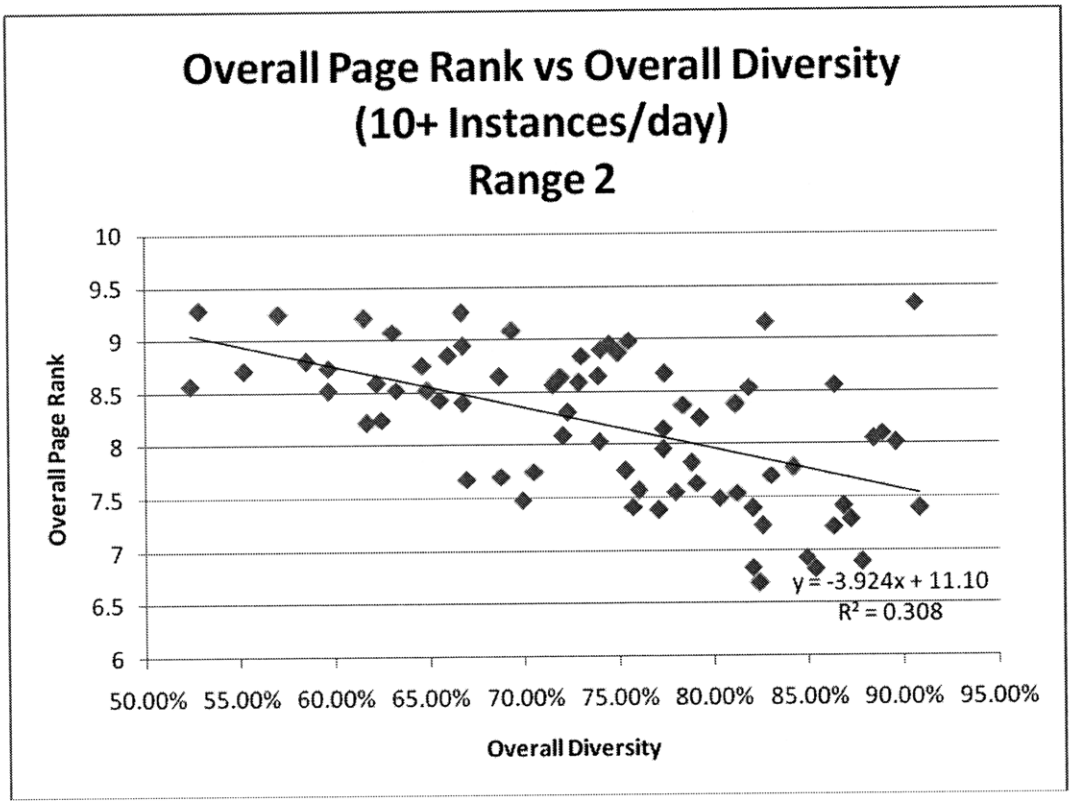

Figure 31 - Overall Page Rank vs Overall Diversity. 10+ Instances/day. Range 2

5.7.2 Overall Diversity vs Amount of Search

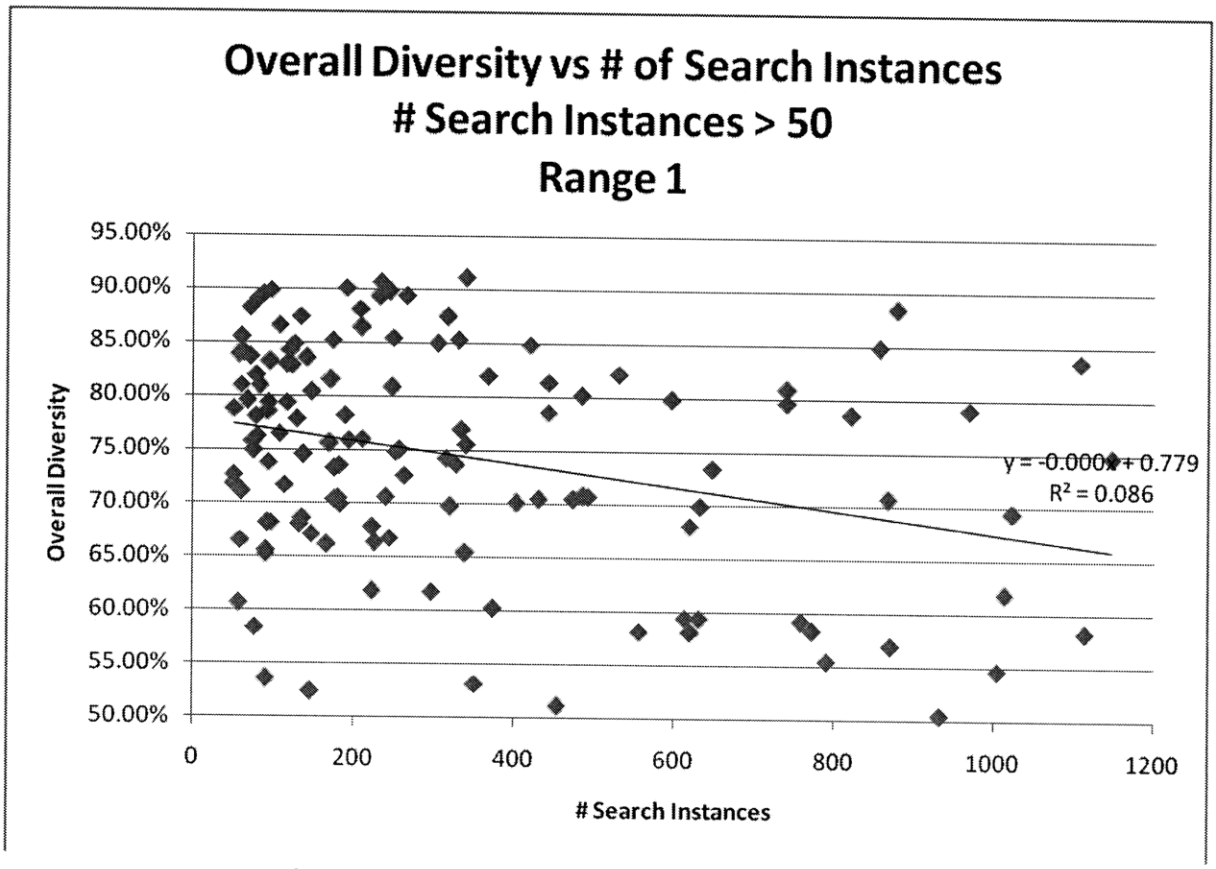

Figure 32 - Overall Diversity vs # of Search Instances. Range 1

In Figure 32 and Figure 33 we have plotted a linear regression of overall diversity vs number of search instances in the data set for users having at least 50 search instances. We removed 19 outliers from Range 1, leaving 129 data points and we removed 15 outliers from Range 2, leaving 137 data points. The outliers that were removed had unusually high search instances or markedly low diversity.

The surprising inverse relationship between overall diversity and number of search instances is faint. We suspect that our inability to precisely distinguish all search traffic on google.com and yahoo.com from other traffic to those servers may be responsible for this weak

relationship. We would expect that higher search leads to more diversity since a wider variety of Web Sites would possibly be encountered through increased search.

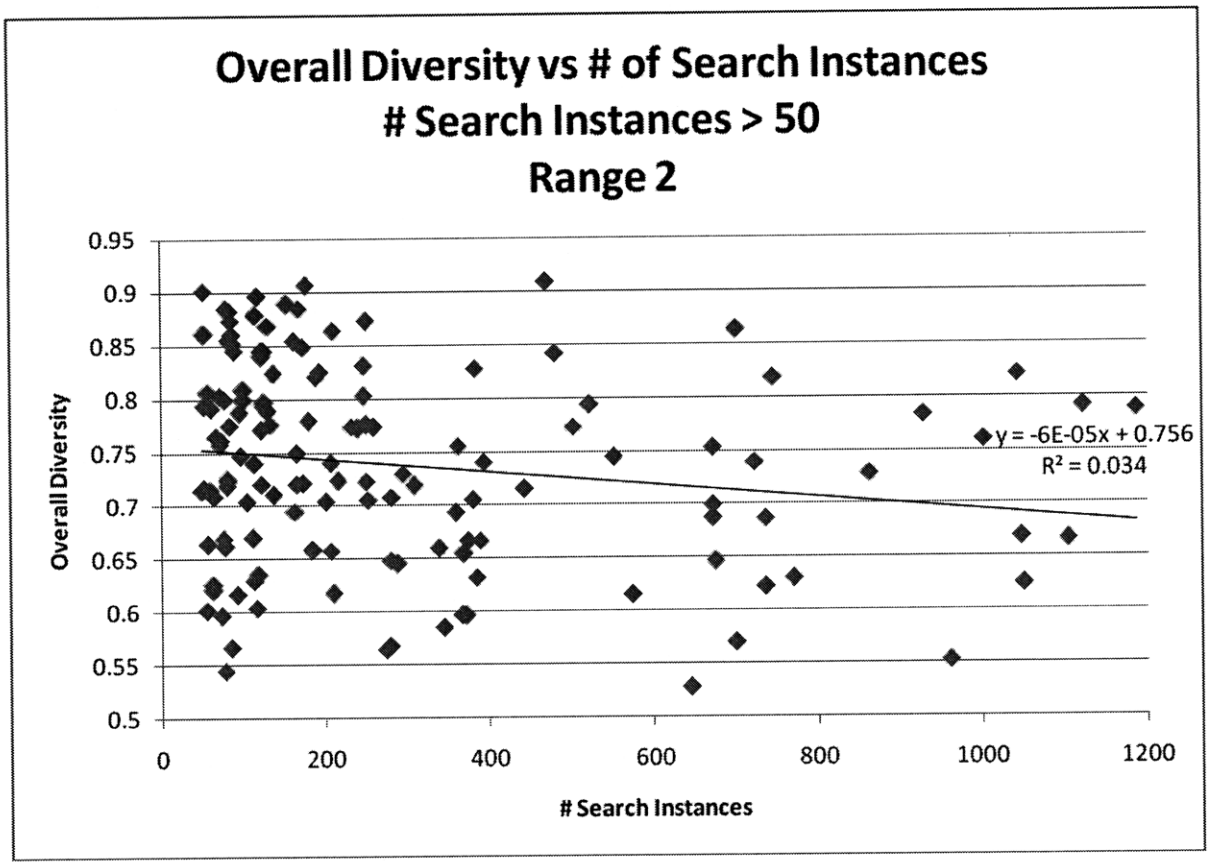

Figure 33 - Overall Diversity vs # of Search Instances. Range 2

5.7.3 Overall Page Rank Deviation vs Overall Diversity Deviation

In Figure 34 and Figure 35 we have plotted a linear regression of the percent deviation from the overall weighted average page rank vs the percent deviation from the overall data set weighted diversity for users having an average of least 10 instance per day. We removed 16 outliers from Range 1, leaving 66 data points and we removed 11 outliers from Range 2, leaving 80 data points. The outliers that were removed had unusually high deviations from the average or were the top users with unusually high amounts of network traffic.

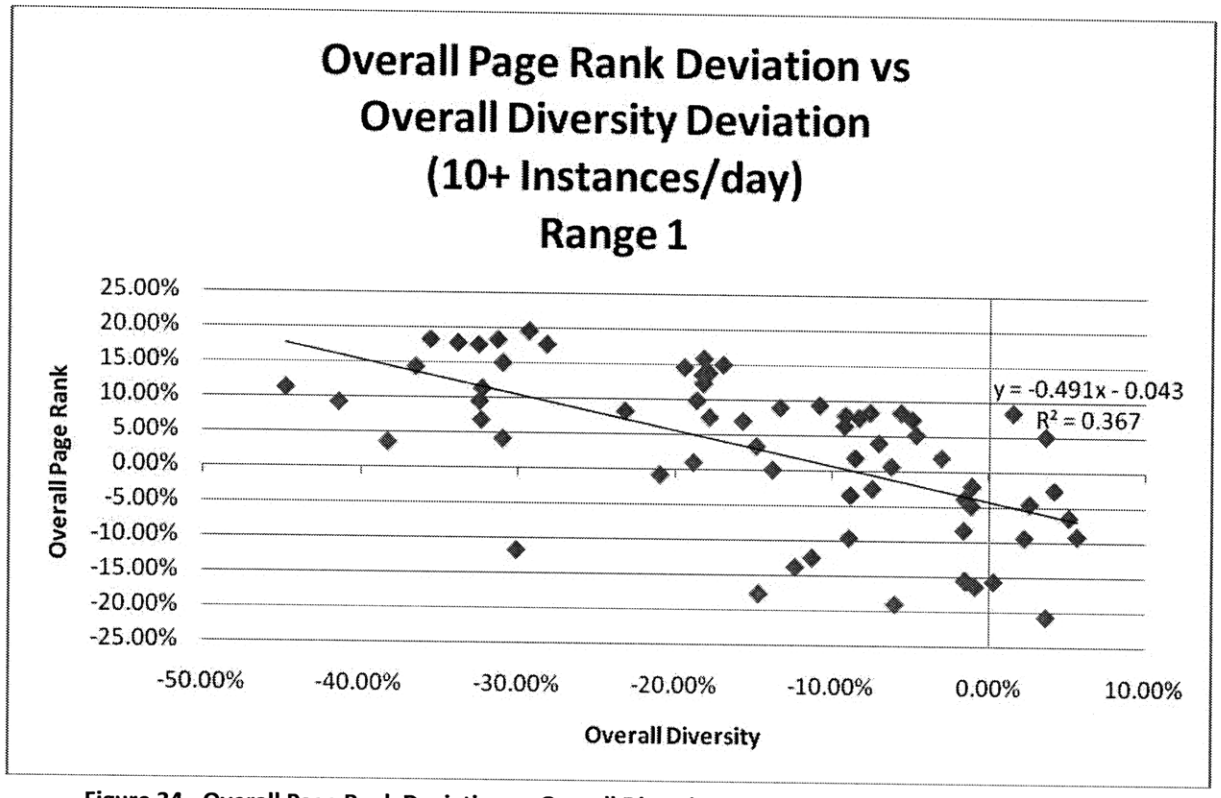

Figure 34 - Overall Page Rank Deviation vs Overall Diversity Deviation. 10+ Instances/day. Range 1

The inverse relationship between these two variables was much more pronounced for Range 1 than for Range 2. The six data points at the bottom left area for Range 2 are the likely cause of this. Their removal increased the R-Squared value to .344, as illustrated in Figure 36.

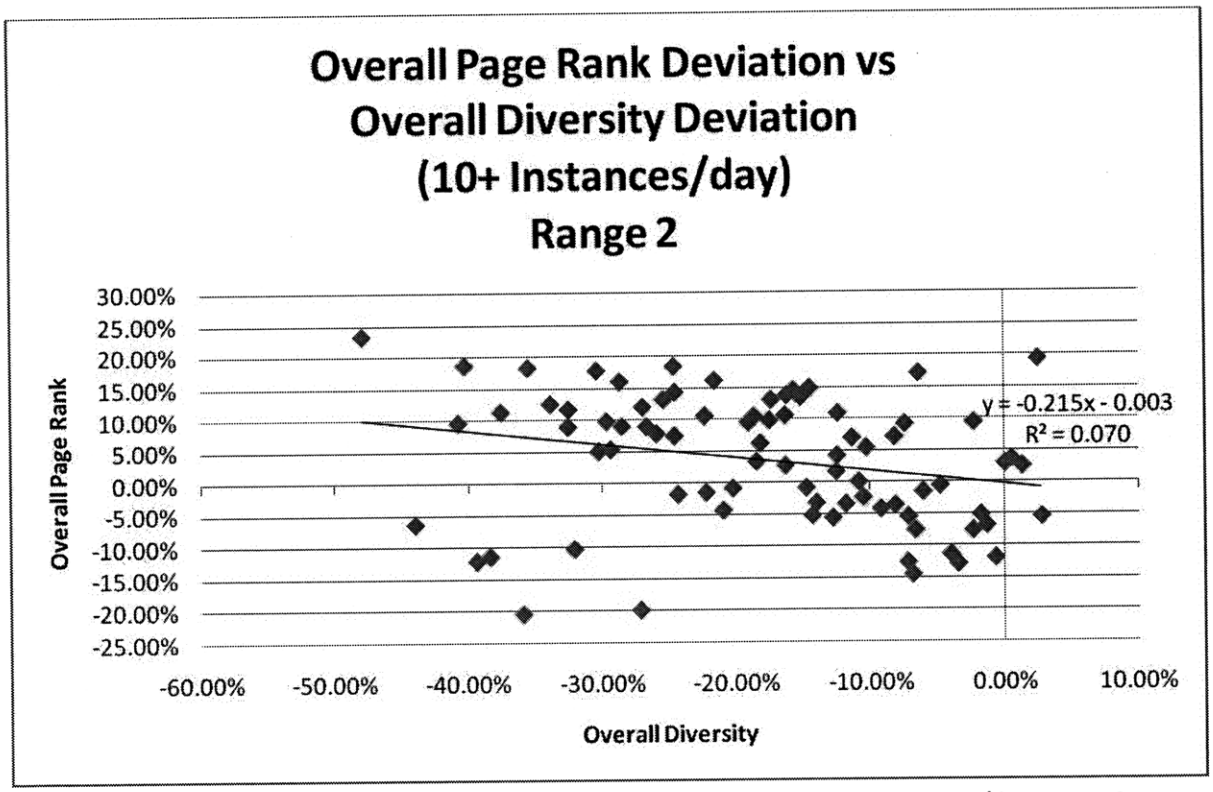

Figure 35 - Overall Page Rank Deviation vs Overall Diversity Deviation. 10+ Instances/day. Range 2

Overall, this relationship showed that as diversity approaches the overall diversity of the data set, the page rank tended to deviate further down below the average page rank. This reinforces the conclusion found in section 5.7.1. that diversity and page rank are inversely proportional.

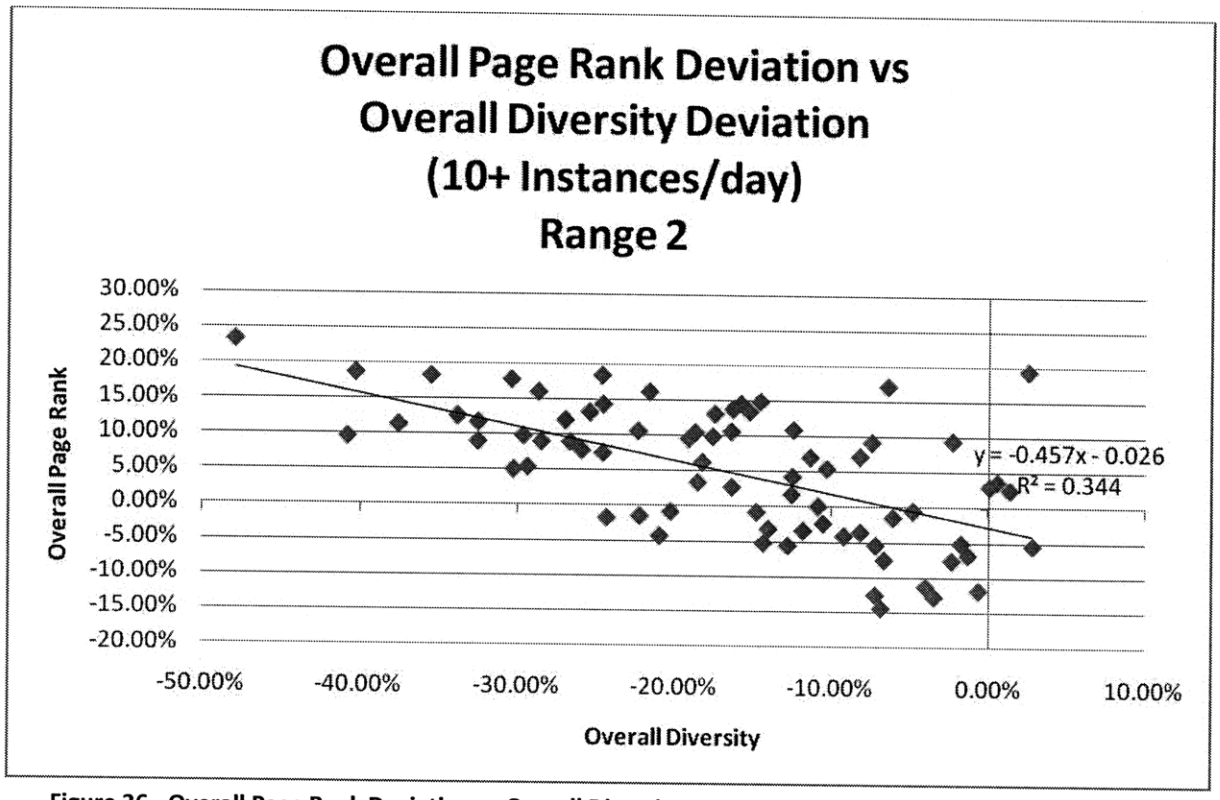

Figure 36 - Overall Page Rank Deviation vs Overall Diversity Deviation. 10+ Instances/day. Range 2. Extra Outliers Removed

5.7.4 Category Diversity vs Overall Page Rank

In Figure 37 and Figure 38, we have plotted linear regressions of category diversity vs overall page rank for the categories of Commerce, Entertainment, News, and Social Networking for Range 1 and Range 2 respectively. We removed 3 outliers from Range 1 and 4 outliers from Range 2 with low page ranks. We chose these categories as a representative sample of the 16 categories in the data set. The graphs also include the R-squared values for each linear regression.

As illustrated by the low R-squared values and the wide spread data of points from each category, there was no correlation between a category's diversity and the user's overall page rank. Commerce, with an R-squared value of .004 and .064 for Range 1 and Range 2,

respectively, changed from almost flat to a negative slope from Range 1 to Range 2. Entertainment, with an R-squared value of 0.00 and .005 for Range 1 and Range 2, respectively, changed from almost completely flat in Range 1 to a slightly negative slope in Range 2. News, with an R-squared value of .006 and 004 for Range 1 and Range 2, respectively, retained its slightly positive slope between Range 1 and Range 2. Lastly, Social Networking, with an R-Squared value of .023 and .005 in Range 1 and Range 2, respectively, also retained its slightly positive slope between Range 1 and Range 2.

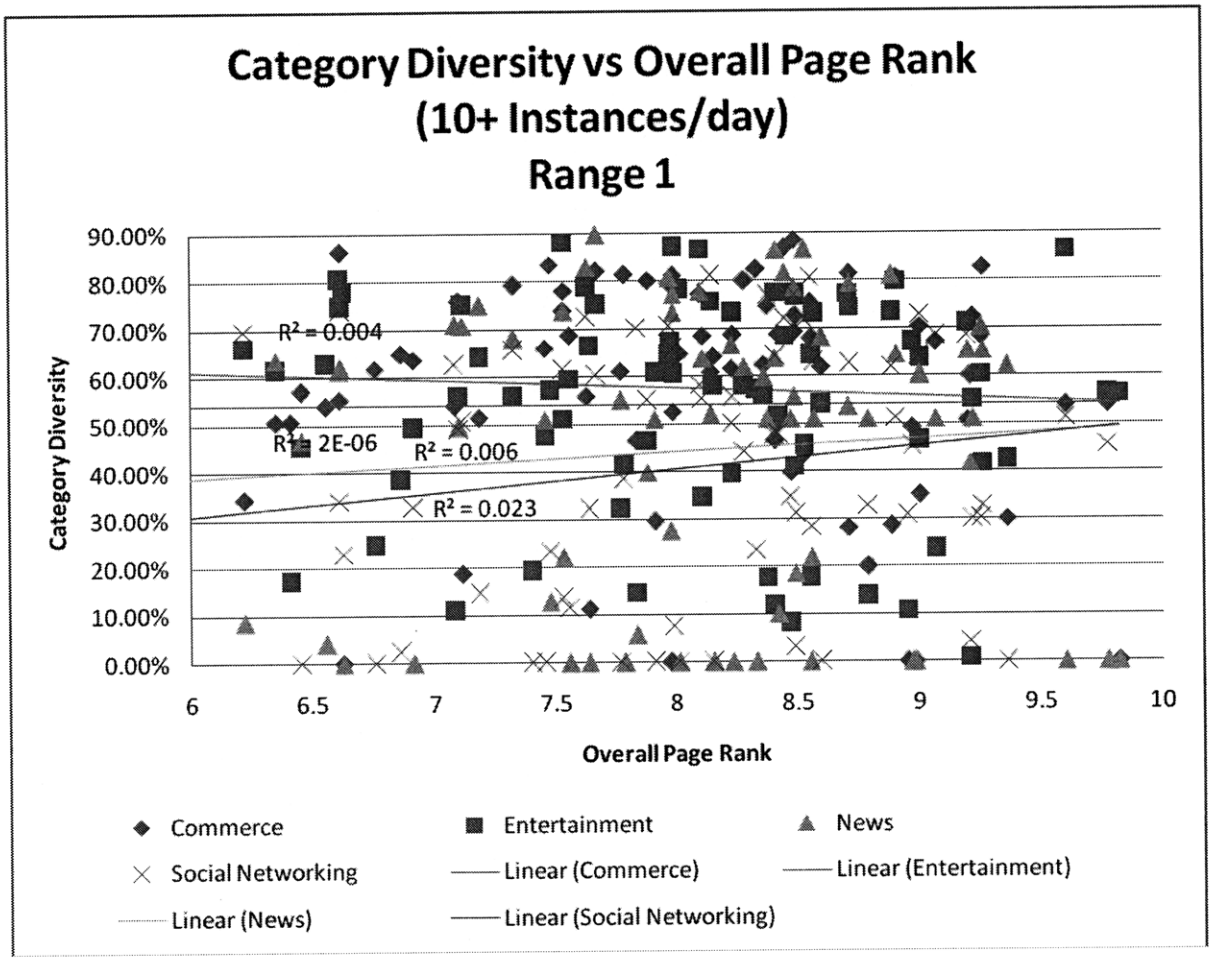

Figure 37 - Category Diversity vs Overall Page Rank. 10+ Instances/day. Range 1

We would expect to see the same results for all the categories, namely that category diversity and overall page rank are not correlated in the network traffic data set. Thus, when viewed in the light of page rank vs overall diversity graphs, overall diversity is needed to determine page rank rather than single category diversity.

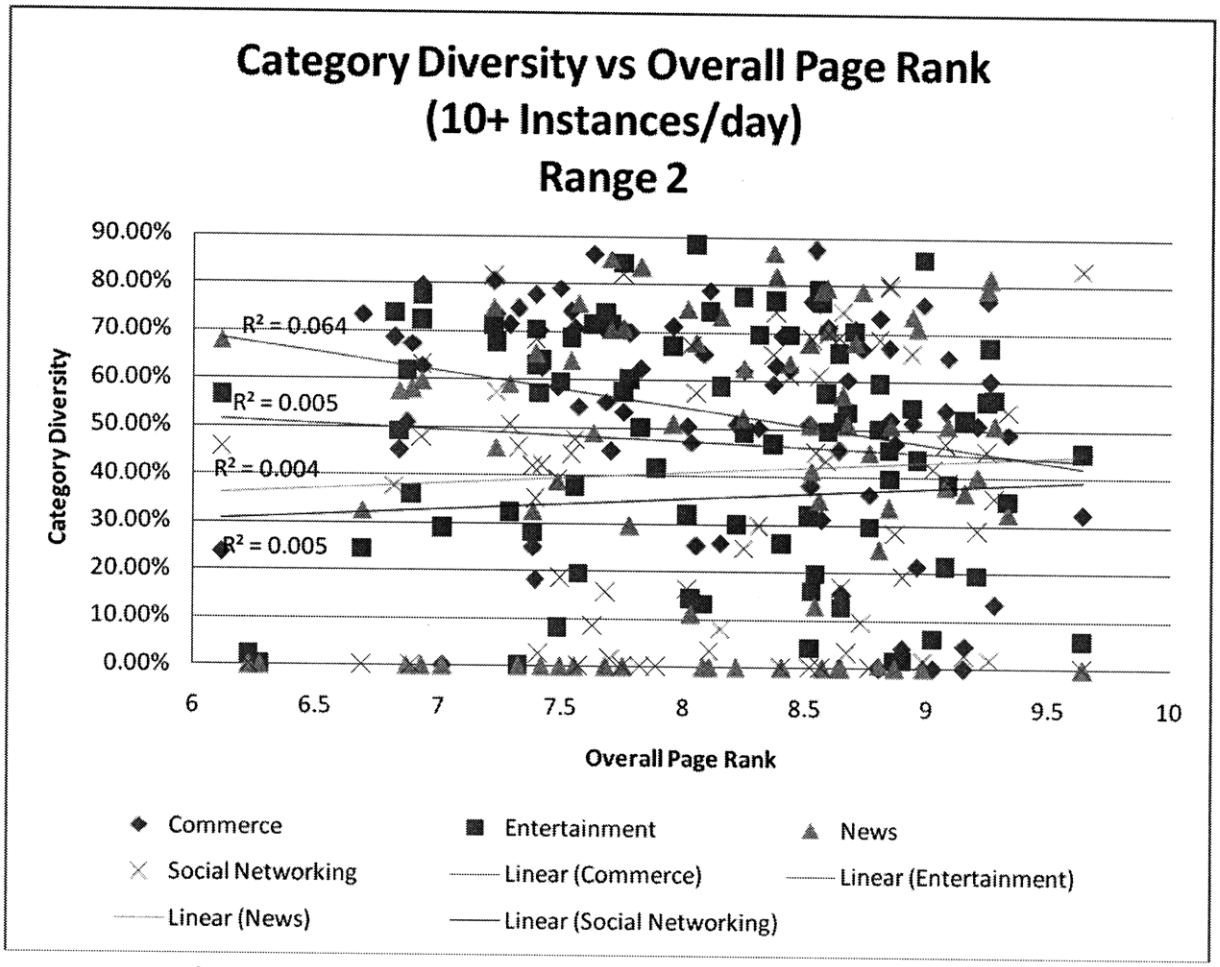

Figure 38 - Category Diversity vs Overall Page Rank. 10+ Instances/day. Range 2

5.7.5 Category Page Rank vs Overall Diversity

In Figure 39 and Figure 40, we have plotted linear regressions of category page rank vs overall diversity for the categories of Commerce, Entertainment, News, and Social Networking for Range 1 and Range 2 respectively. We removed outliers with low diversity scores or page

ranks. We removed 10 and 6 outliers from the Commerce category in Range 1 and Range 2, respectively. We removed 8 and 10 outliers from the Entertainment category in Range 1 and Range 2 respectively. We removed 16 and 15 outliers from the News category in Range 1 and Range 2 respectively. We removed 7 and 3 outliers from the Social Networking category in Range 1 and Range 2, respectively.

We chose these categories as a representative sample of the 16 categories in the data set. The graphs also include the R-squared values for each linear regression.

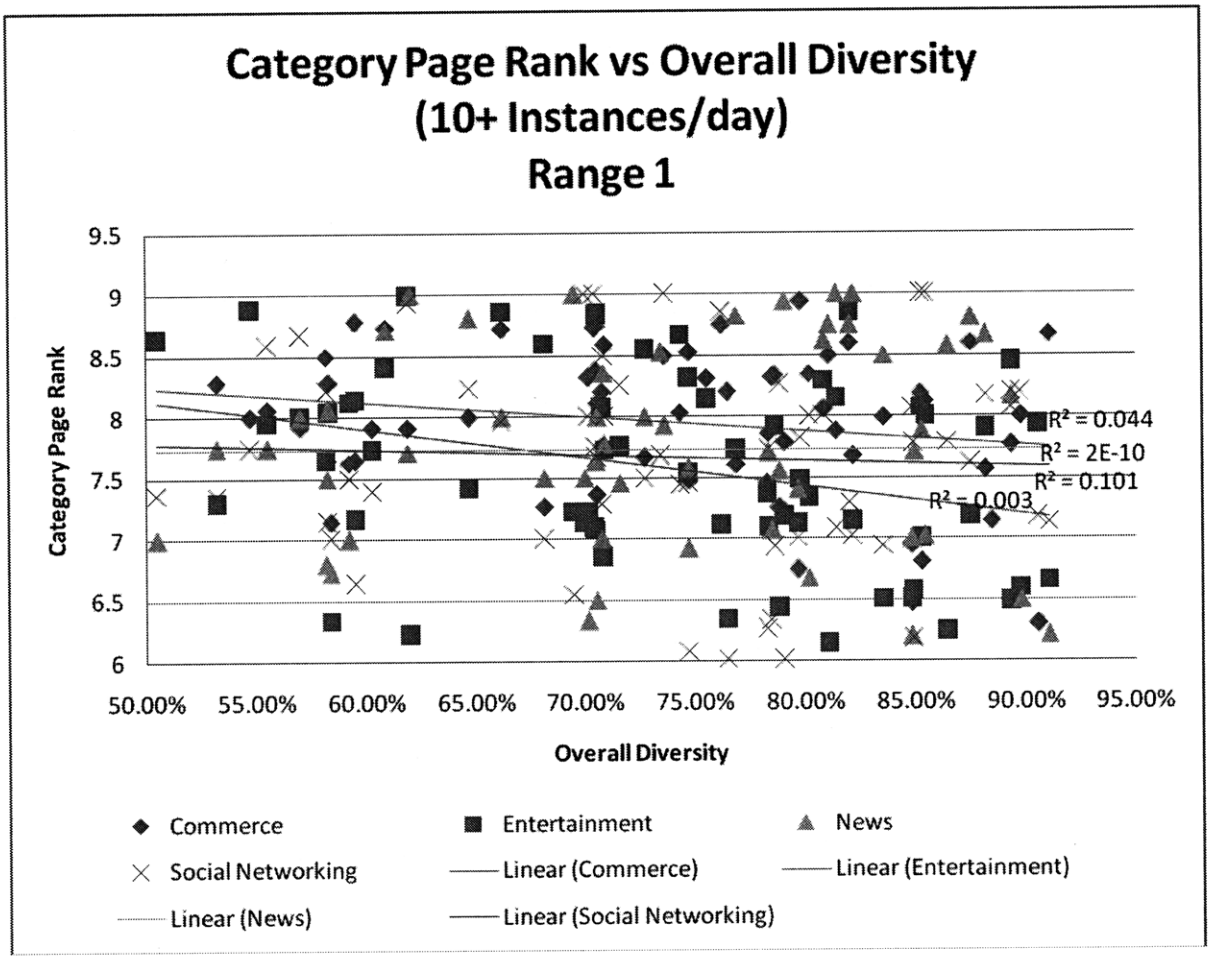

Figure 39 - Category Page Rank vs Overall Diversity. 10+ Instances/day. Range 1

As illustrated by the low R-squared values and the changing slope of the regression lines between the Range 1 and Range 2 graphs, there was no relationship between a category's page

rank and the user's overall diversity. Commerce, with an R-squared value of .044 and .026 for Range 1 and Range 2, respectively, had a slight negative slope in each graph. Entertainment, with an R-squared value of .101 and .002 for Range 1 and Range 2, respectively, changed from a negative slope in Range 1 to almost completely flat in Range 2. News, with an R-squared value of 0.00 and .007 for Range 1 and Range 2, respectively, changed from mostly flat in Range 1 to a positive slope in Range 2. Lastly, Social Networking, with an R-Squared value of .003 and .024 in Range 1 and Range 2, respectively, also changed from mostly flat to a slight negative slope.

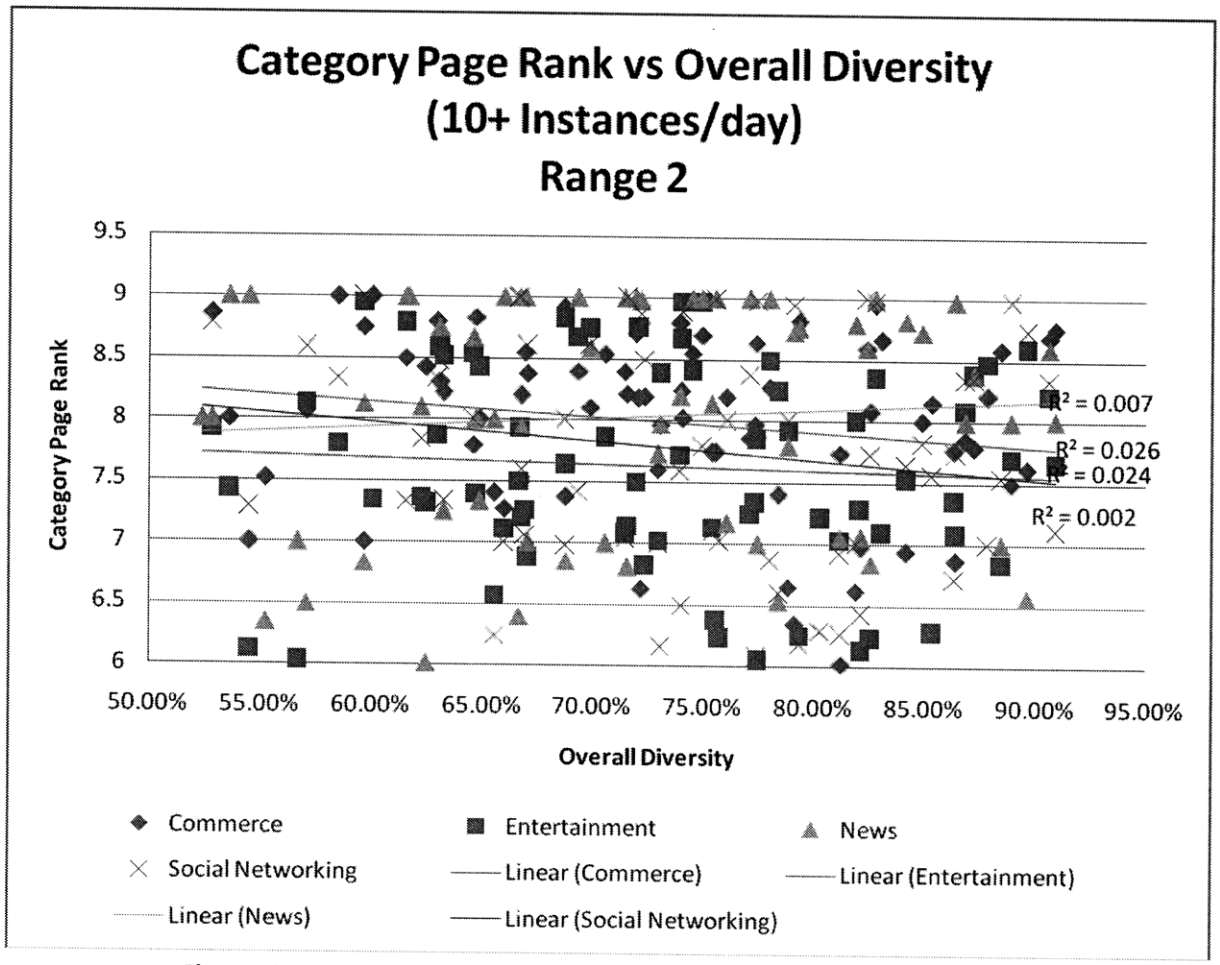

Figure 40 - Category Page Rank vs Overall Diversity. 10+ Instances/day. Range 2

We would expect to see the same results for all the categories, namely that category page rank and overall diversity are not correlated in the network traffic data set. Thus, when viewed

in the light of page rank vs overall diversity graphs, overall page rank is needed to determine overall diversity rather than single-category page rank.

5.7.6 Category Page Rank vs Category Diversity

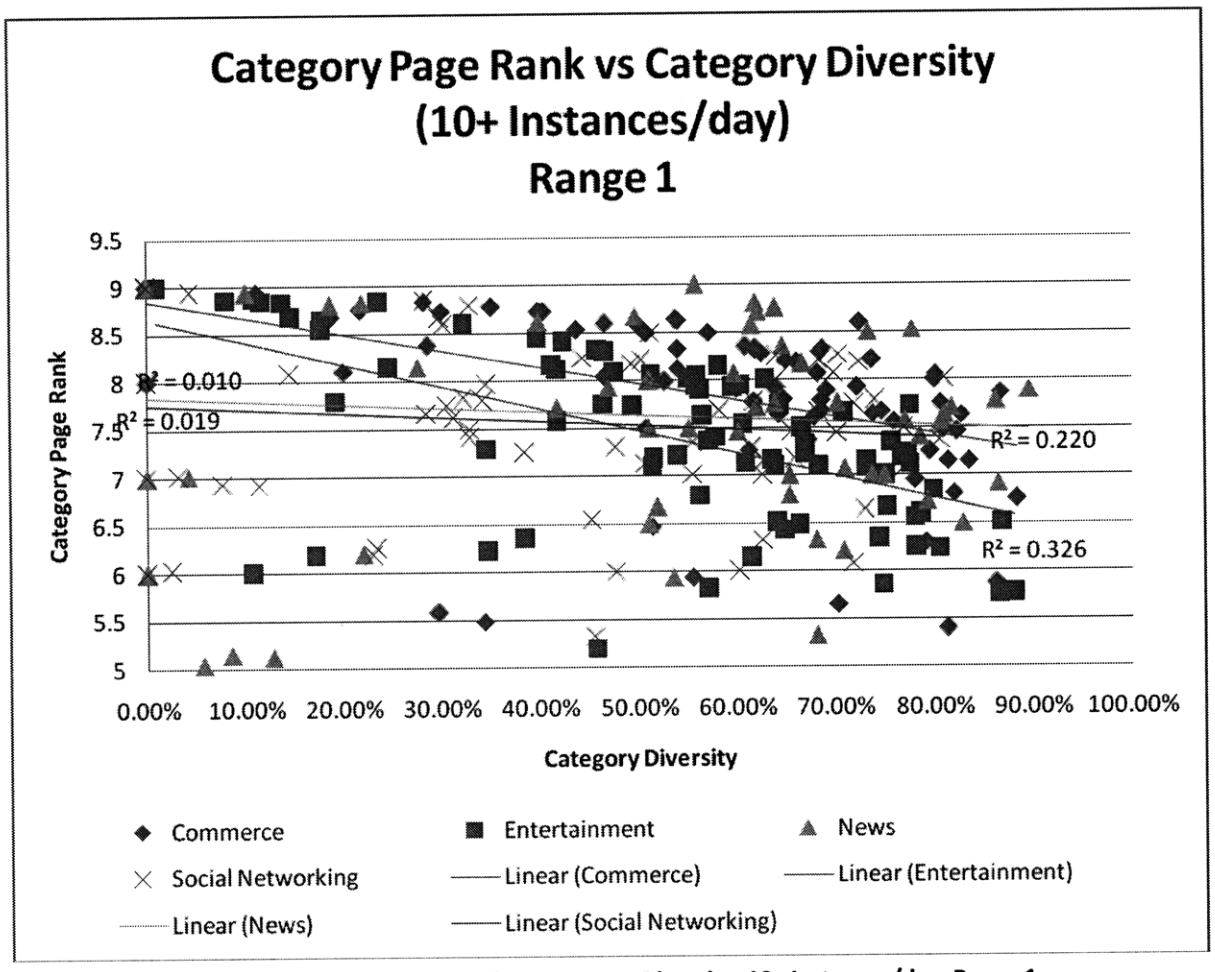

Figure 41 - Category Page Rank vs Category Diversity. 10+ Instances/day. Range 1

In Figure 41 and Figure 42, we have plotted linear regressions of category page rank vs category diversity for the categories of Commerce, Entertainment, News, and Social Networking for Range 1 and Range 2 respectively. We removed outliers with low page ranks. We removed 4 outliers from the Commerce category in both Range 1 and Range 2. We removed 2 and 5 outliers from the Entertainment category in Range 1 and Range 2 respectively.

We removed 10 and 8 outliers from the News category in Range 1 and Range 2 respectively. We removed 3 outliers from the Social Networking category in Range 2. We chose these categories as a representative sample of the 16 categories in the data set. The graphs also include the R-squared values for each linear regression.

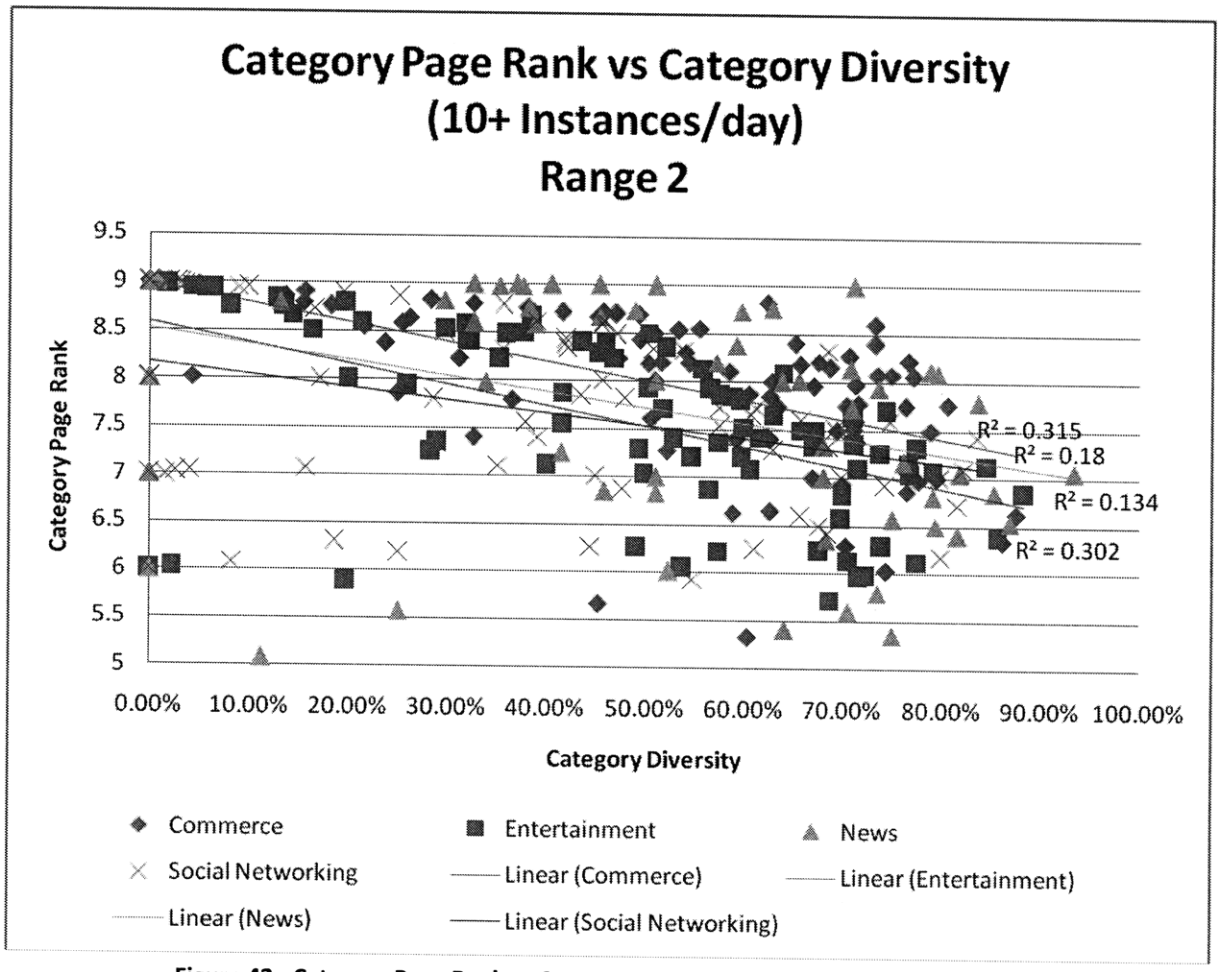

Figure 42 - Category Page Rank vs Category Diversity. 10+ Instances/day. Range 2

In this graph, we retained users with low category diversity because, unlike overall diversity, it is much more likely that a user may only frequent a small number of Web Sites within a given category. For example, there were dozens of news Web Sites in the data set, but many people may solely get their news from cnn.com

Overall, our results are mixed. Commerce had R-Squared values of .220 and .315 for Range 1 and Range 2, respectively. Entertainment also had relatively high R-Squared values of .326 and .302 for Range 1 and Range 2, respectively. This suggested that diversity in both of these categories was inversely related to the resulting page rank of that category. Intuitively, we would expect this to be the case. Users with more category diversity visited the more obscure Web Sites more often in addition to the more popular and higher ranked Web Sites that most users view.

For the other two categories, the findings were less pronounced. News had R-Squared values of .010 and .18 for Range 1 and Range 2, respectively. Social Networking had R-squared values of .019 and .134 for Range 1 and Range 2, respectively. While both of these categories showed an inverse relationship between category diversity and category page rank, the findings were fainter in Range 1 and more pronounced in Range 2, although still not as pronounced as the Commerce and Entertainment category for either Range. We reasoned that the R-squared values were lower for these categories because of the degree to which users were highly loyal to the few Web Sites they use. This loyalty produced low diversity and consistently contributed to a high or low page rank, depending on the Web Site chosen, making it more difficult to correlate between the two variables.

We would expect to see similar results in the other categories, namely that category page rank is inversely proportional to category diversity with some categories exhibiting this correlation more strongly than others.

6 Discussion and Conclusions

In this study, we have used network traffic data from a medium-sized company in California to develop meaningful metrics for quantifying the information seeking habits of information workers. We first collected network traffic data using sFlow technology to sample the company's network. Than we pre-processed this data to pull out valid externally destined network traffic and categorize the Web Sites visited into a set of main categories and sub-categories. Afterwards, we studied the data to better understand user behavior and trends. Finally, we looked at specific metrics that would allow us to quantify the information seeking habits of information works.

Overall, we discovered that by dividing Web Site traffic into meaningful categories, we could use diversity measurements and page rank calculations to gain a strong sense of the manner in which users sought information from either a wide variety or small amount of sources. These metrics are a major first step in a toolbox of metrics for studying the information seeking behavior of information workers. Future research by our team will examine the relationship between information worker productivity and the metrics presented in this thesis.

7 Limitations and Future Work

An important limiting factor of our research was that the network data we gathered only showed the external hostname of outbound Internet connections and not the URL address as seen by the actual users. Having the full URL would greatly increase the accuracy of our categorization steps for Web Sites that have multiple purposes. For example, google.com

services like email, chat, and search, could be fully distinguished from each other based on the URL of the visited Web Site.

Additionally, a different arrangement of main categories and sub-categories may increase the responsiveness of our metrics to the individual nuances of each user. Future research in this area could explore using different main and sub-categories in order to determine an optimal categorization structure.

The manual categorization processes, although resulting in high accuracy, was extremely time consuming. Future work in this area could better explore automatic categorization techniques that may sample a Web Site's content to estimate its category. This could decrease the amount of user intervention needed to categorize visited Web Sites.

Our research was conducted offline in a piece-by-piece fashion by running various scripts. A nice advancement would be a start-to-finish implementation that could actively incorporate all the sFlow data coming from a network and automatically run it through the necessary pre-processing and analysis steps. This approach would reduce the time and energy needed to manually process the data and would allow for a real-time view instead of an after-the-fact view of user information seeking patterns and behavior.

Lastly, our research only looked at externally destined network traffic, but internal network traffic could also shed light on the information seeking habits of computer users. Future work in this area could develop metrics that quantify how a user spends his time working with internal company systems like email servers, file servers, knowledge resources, and databases.

Appendix A: MySQL Table Definitions

"Network" Table Definition				
Field	Type	Null	Default	Extra
id	int(11)	NO	NULL	auto_increment
date	datetime	YES	NULL	
src_mac	char(17)	YES	NULL	
dst_mac	char(17)	YES	NULL	
src_host	text	YES	NULL	
dst_host	text	YES	NULL	
src_port	int(11)	YES	NULL	
dst_port	int(11)	YES	NULL	
ext_web_site_id	int(11)	NO	0	
is_external	tinyint(1)	YES	0	
is_vrrp	tinyint(1)	YES	0	
is_local	tinyint(1)	YES	0	
range	tinyint(2)	YES	0	

Table 25 - "Network" Table Definition

"External_web_site" Table Definition				
Field	Type	Null	Default	Extra
id	int(11)	NO	NULL	auto_increment
web_site	varchar(255)	YES	NULL	
instances_range1	mediumint(9)	NO	0	
instances_range2	mediumint(9)	NO	0	
has_min_instances	tinyint(4)	YES	0	
is_failed	tinyint(1)	YES	0	
failed_reason	varchar(50)	YES	NULL	
google_desc	varchar(1000)	YES	NULL	
meta_desc	varchar(1000)	YES	NULL	
meta_keywords	varchar(1000)	YES	NULL	
manual_desc	varchar(1000)	YES	NULL	
is_whq	tinyint(1)	YES	0	
page_rank	tinyint(4)	YES	0	

Table 26 - "External_web_site" Table Definition

"Main_category" Table Definition				
Field	Type	Null	Default	Extra
id	tinyint(4)	NO	NULL	auto_increment
name	varchar(50)	YES	NULL	

description	varchar(100)	YES	NULL	

Table 27 - "Main_category" Table Definition

"Sub_category" Table Definition				
Field	Type	Null	Default	Extra
id	mediumint(9)	NO	NULL	auto_increment
parent_id	tinyint(4)	NO	NULL	
name	varchar(50)	YES	NULL	

Table 28 - "Sub_category" Table Definition

"Sub_category_assignment" Table Definition				
Field	Type	Null	Default	Extra
stc_id	int(11)	NO	NULL	
e_id	int(11)	NO	NULL	

Table 29 - "Sub_category_assignment" Table Definition

Appendix B: Main Category Breakdowns

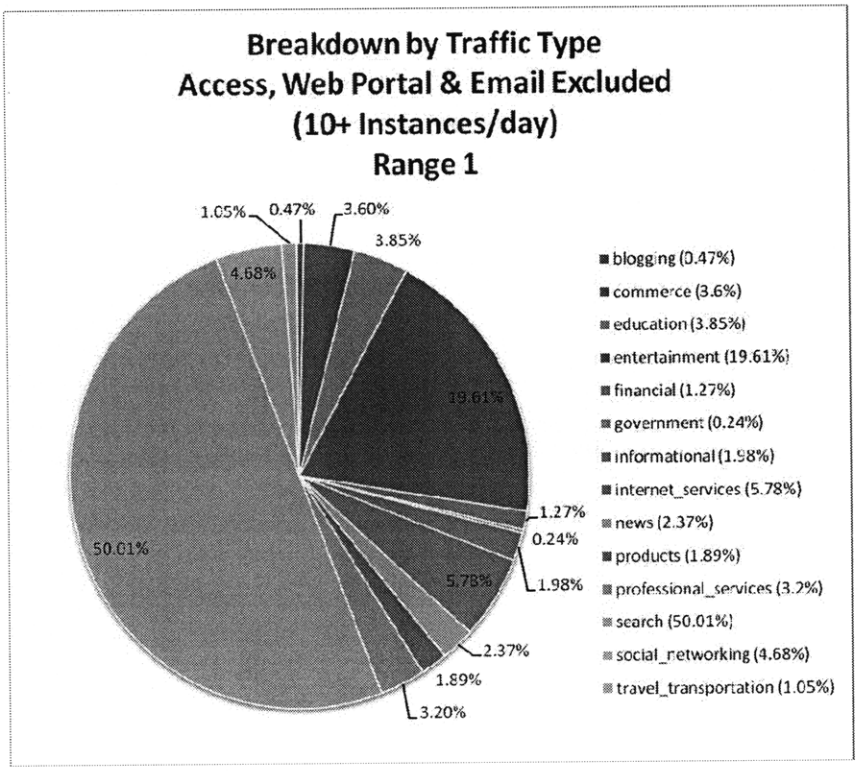

Figure 43 - Breakdown by Traffic Type. Access, Web Portal & Email Excluded. 10+ Instances/day. Range 1

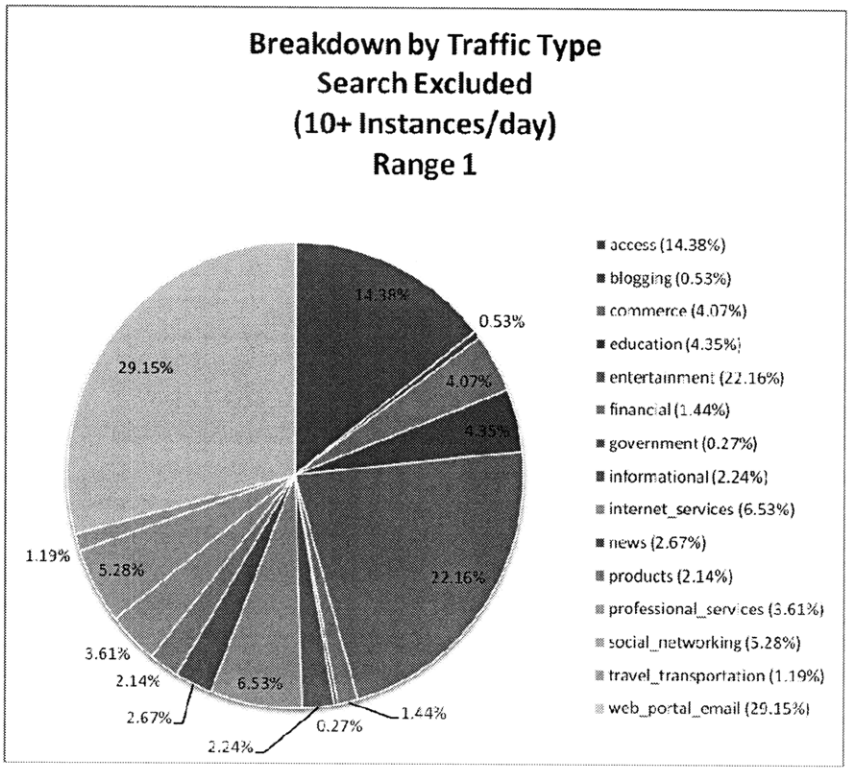

Figure 44 - Breakdown by Traffic Type. Search Excluded. 10+ Instances/day. Range 1

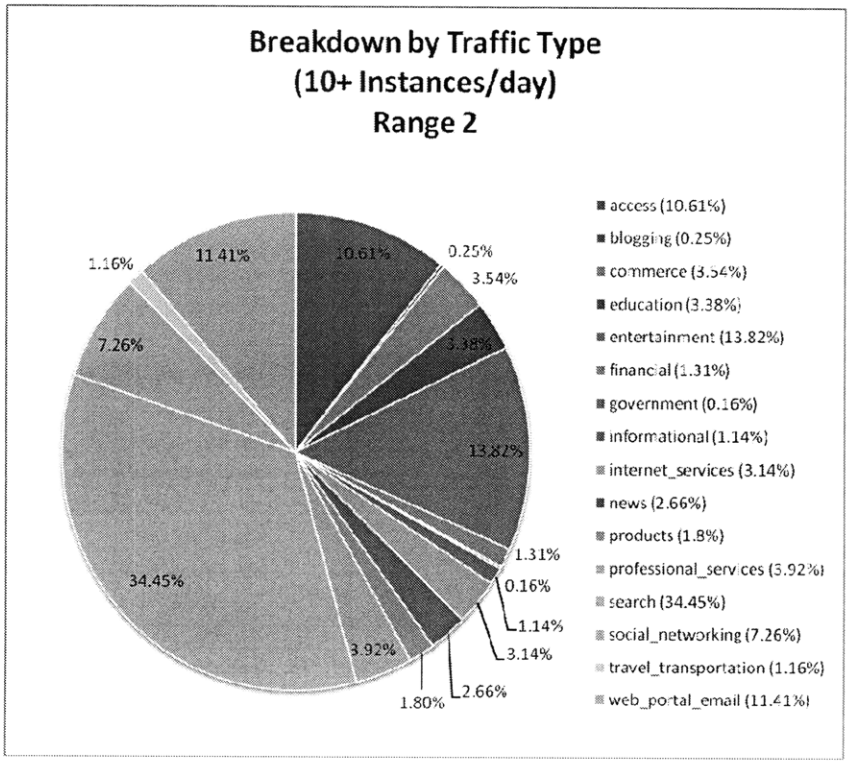

Figure 45 - Breakdown by Traffic Type. 10+ Instances/day. Range 2

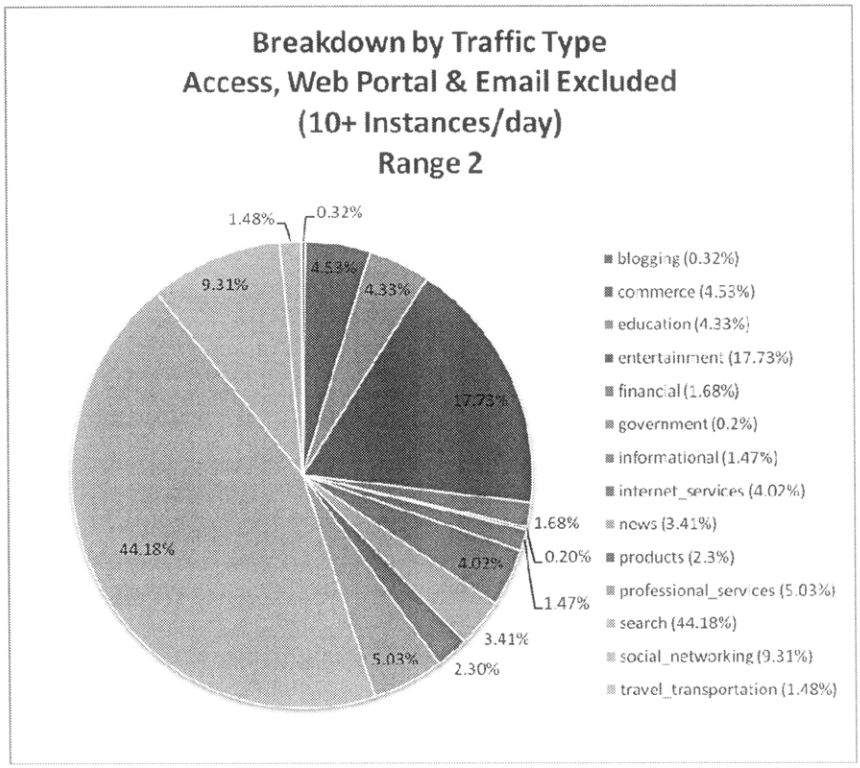

Figure 46 - Breakdown by Traffic Type. Access, Web Portal & Email Excluded. 10+ Instances/day. Range 2

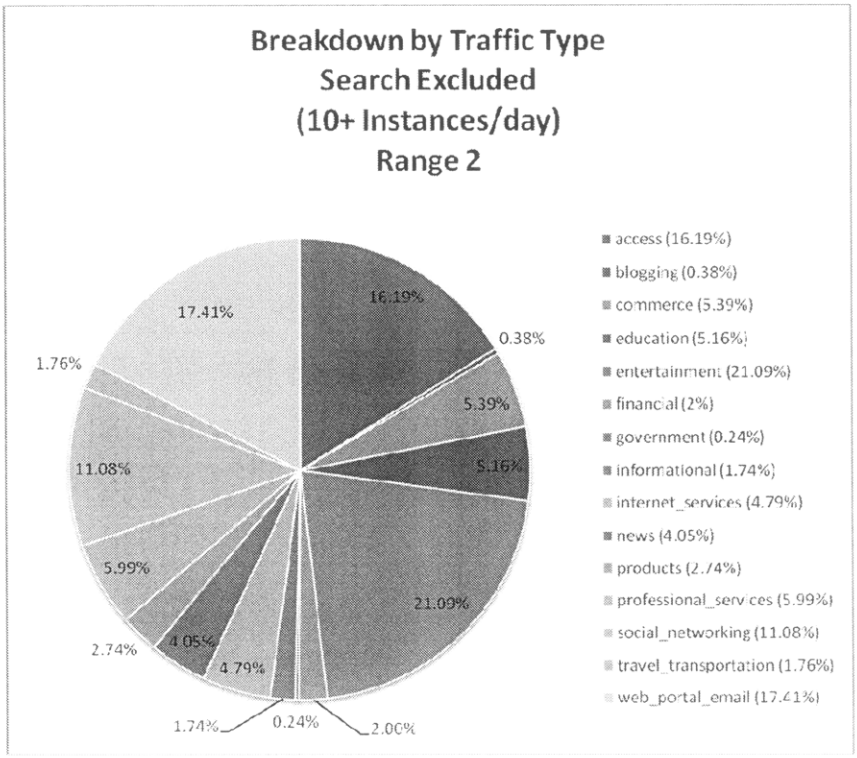

Figure 47 - Breakdown by Traffic Type. Search Excluded. 10+ Instances/day. Range 2

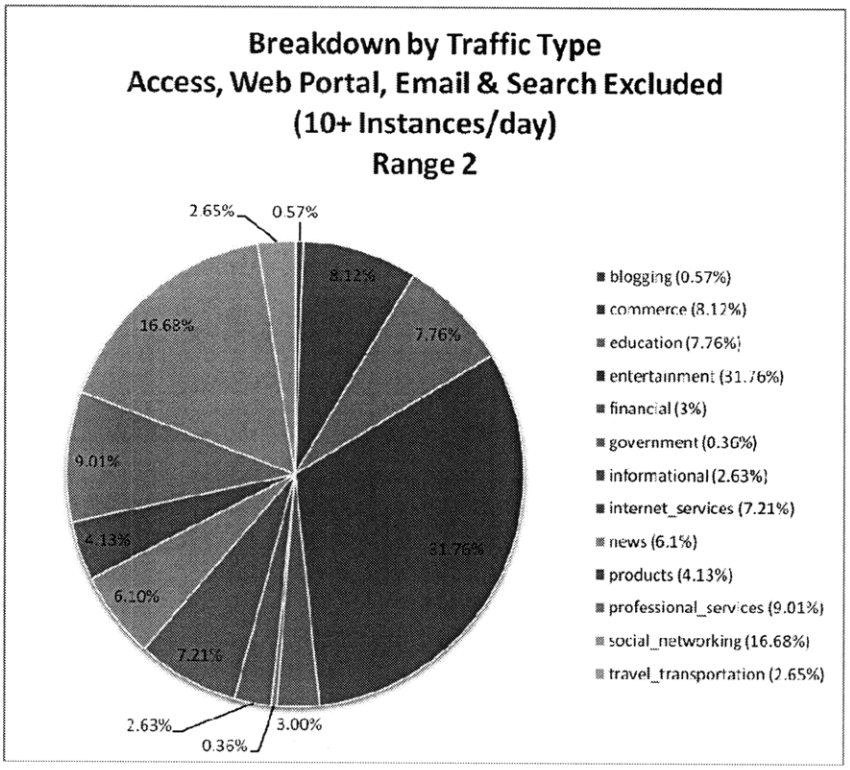

Figure 48 - Breakdown by Traffic Type. Access, Web Portal & Email, & Search Excluded. 10+ Instances/day. Range 2

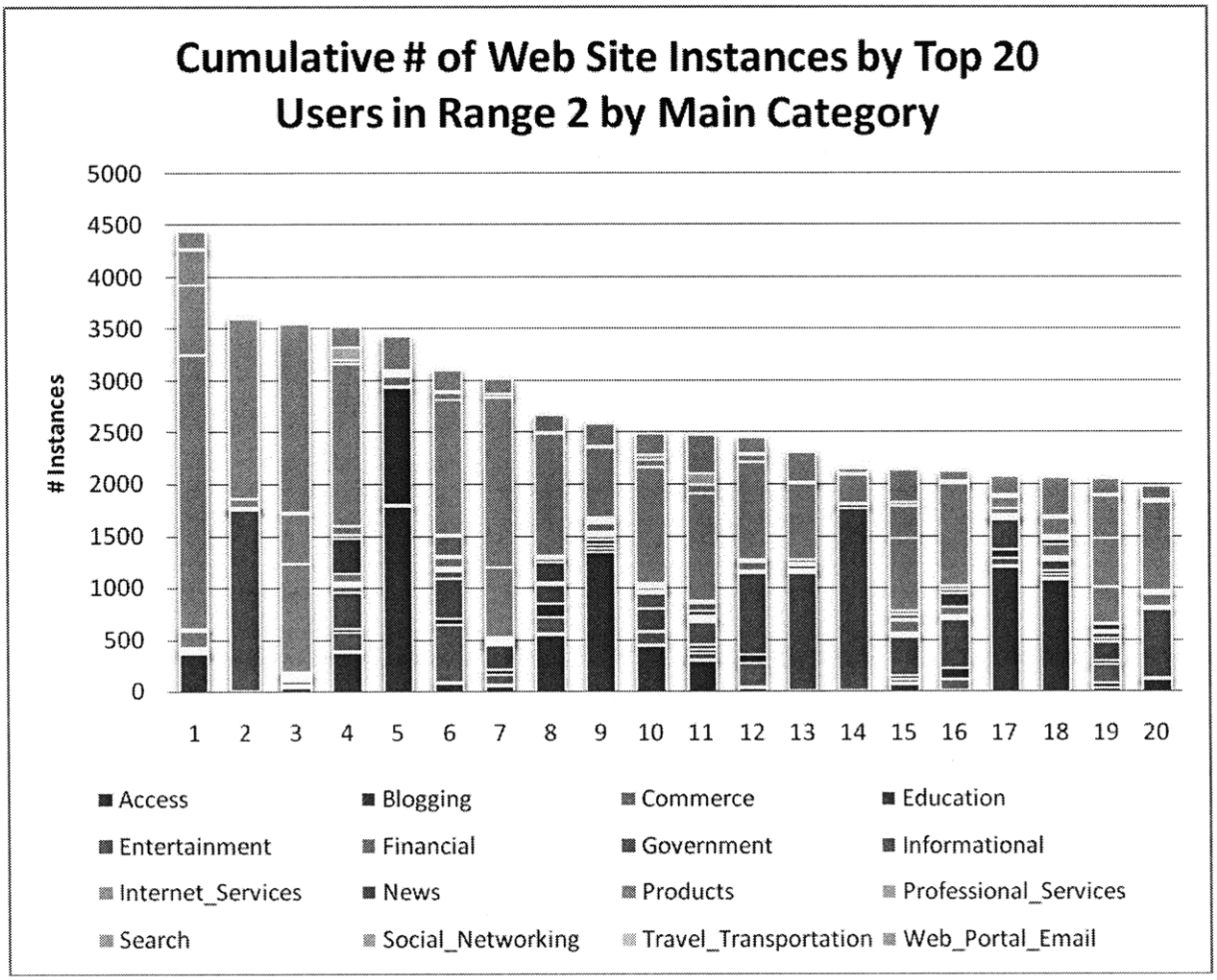

Figure 49 - Cumulative # of Web Site Instances by Top 20 Users in Range 2 by Main Category

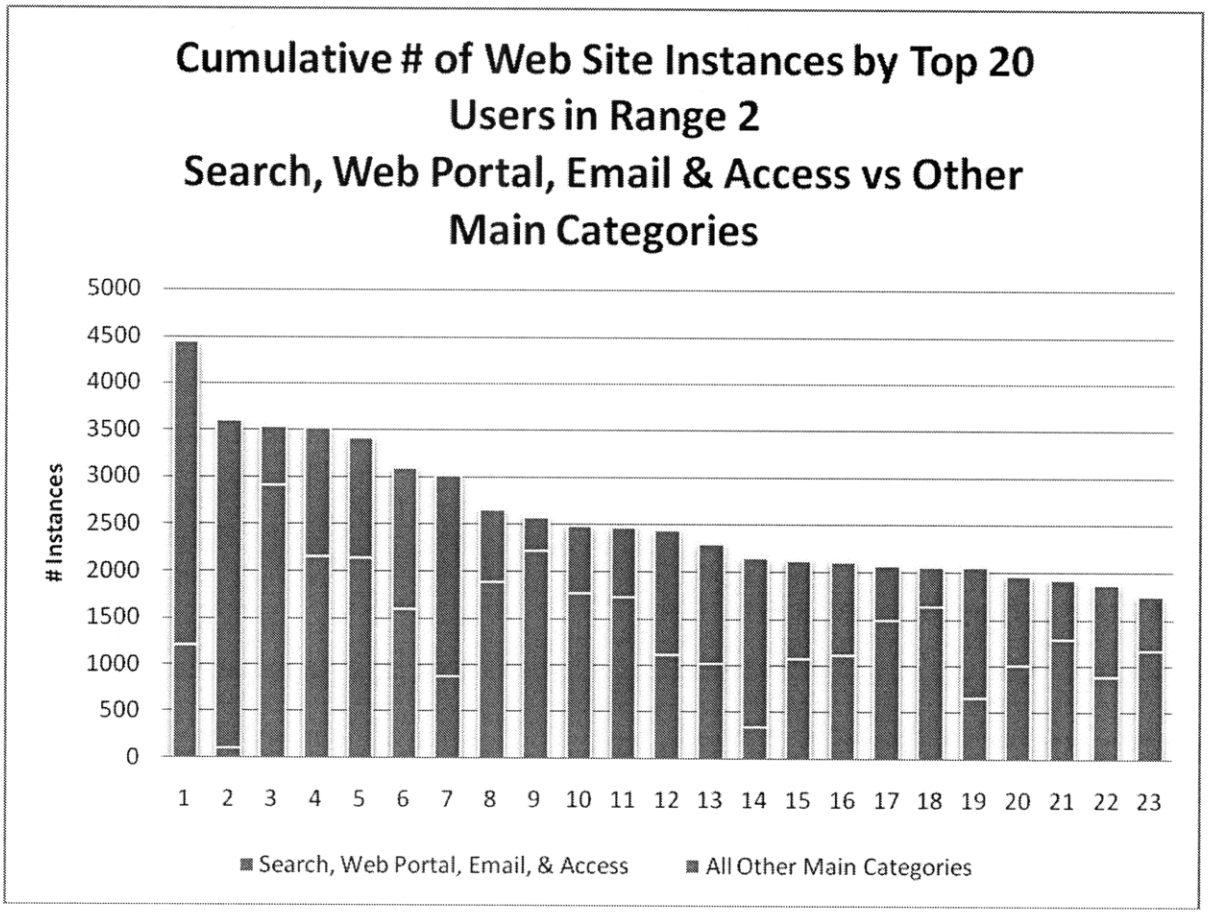

Figure 50 - Cumulative # of Web Site Instances by Top 20 Users in Range 2. Search, Web Portal & Email & Access vs Other Main Categories

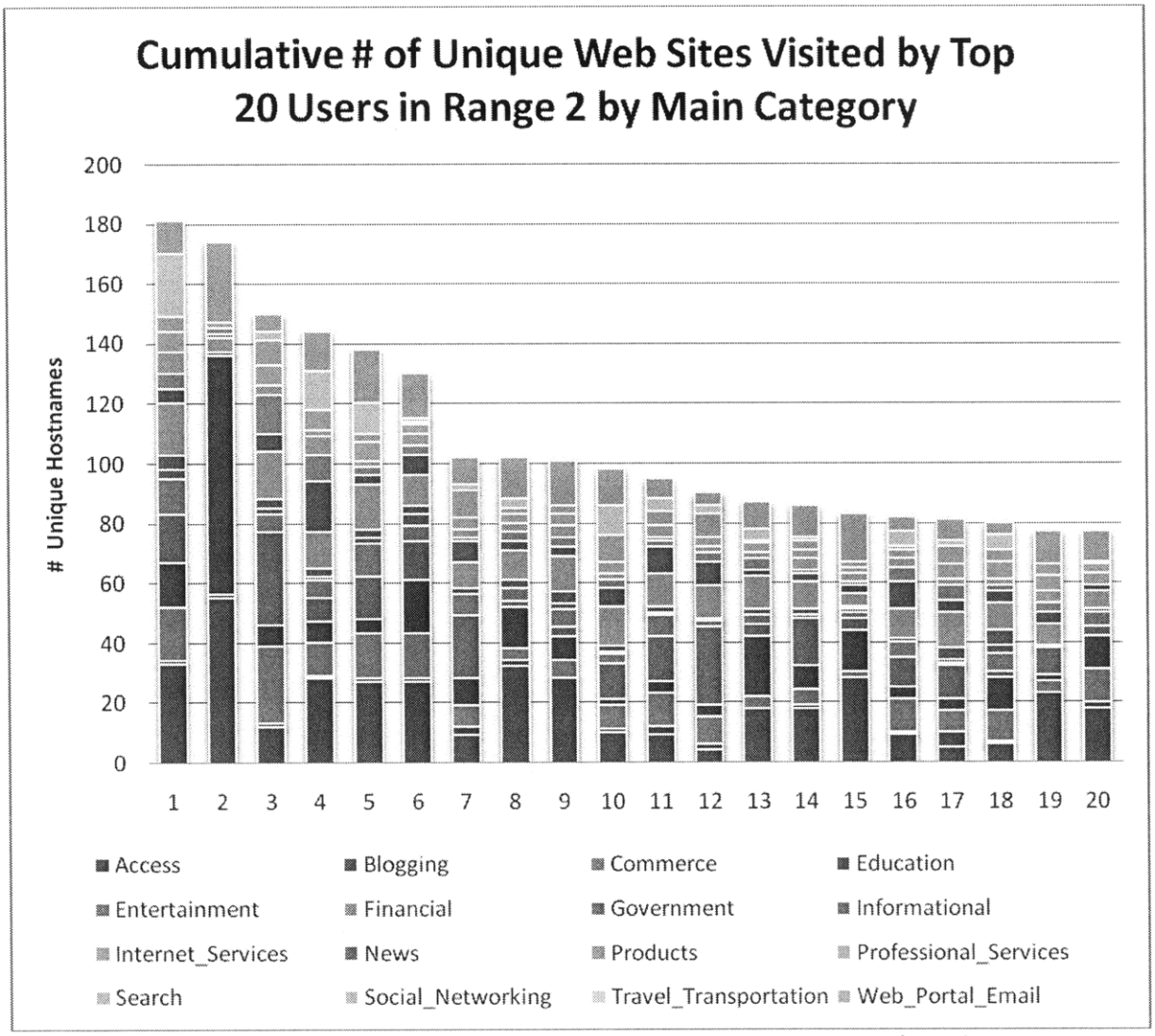

Figure 51 - Cumulative # of Unique Web Sites Visited by Top 20 Users in Range 2 by Main Category

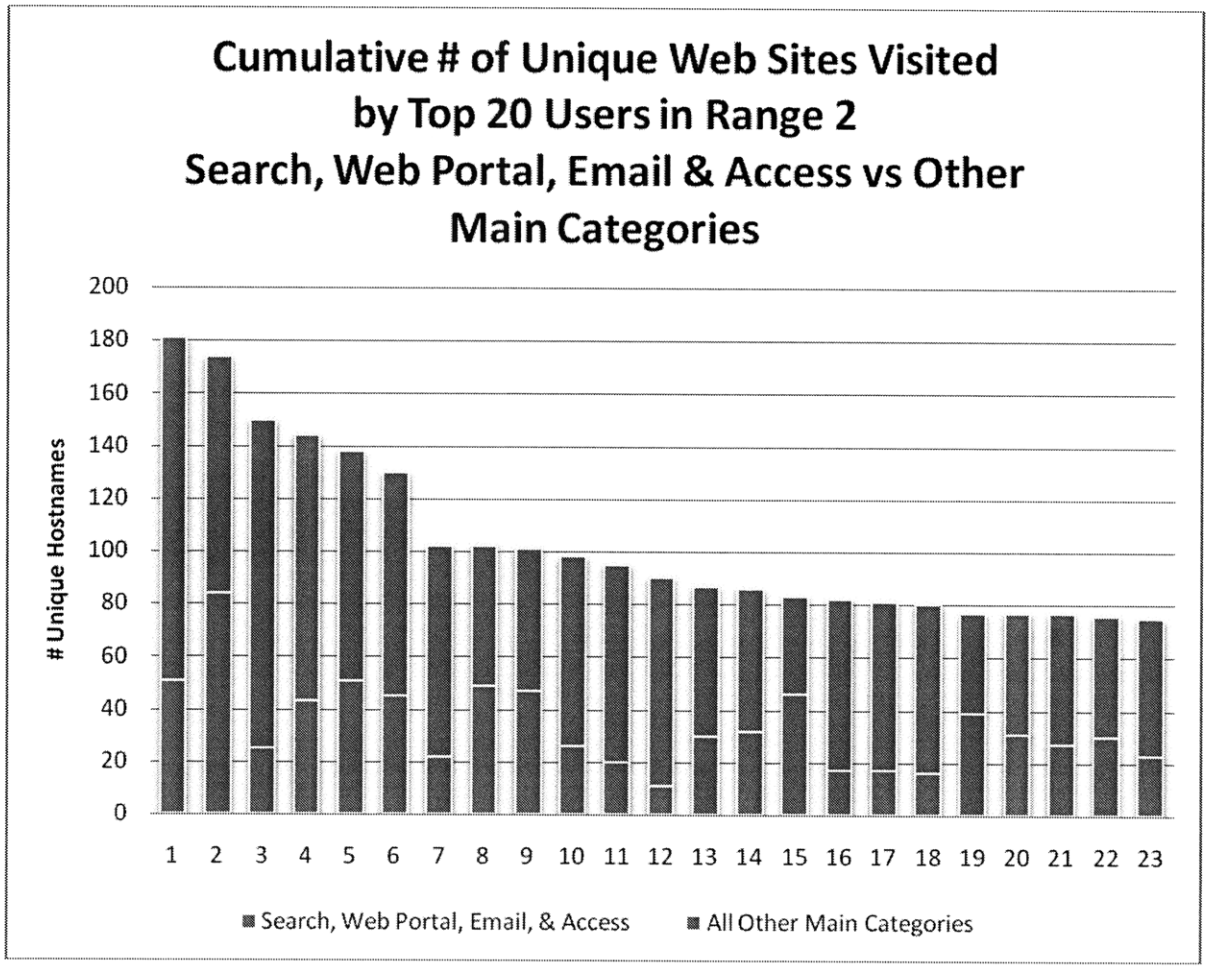

Figure 52 - Cumulative # of Unique Web Sites Visited by Top 20 Users in Range 2. Search, Web Portal & Email & Access vs Other Main Categories

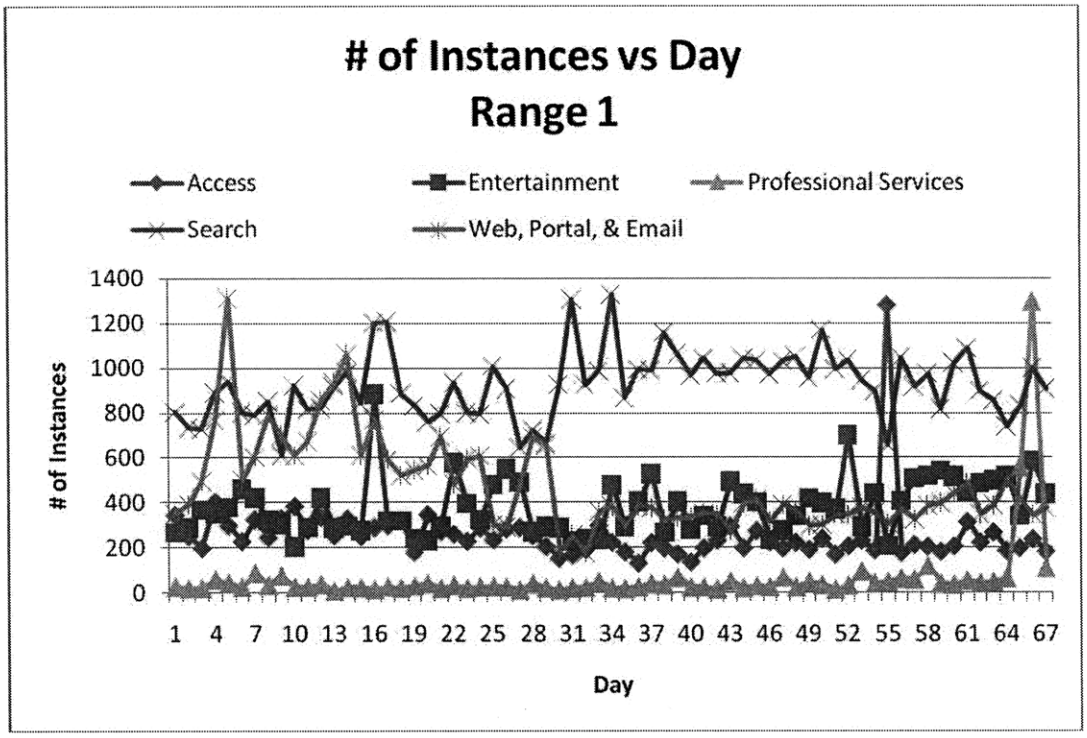

Figure 53 - # of Instances vs Day. Range 1. Access, Entertainment, Professional Services, Search, Web Portal & Email

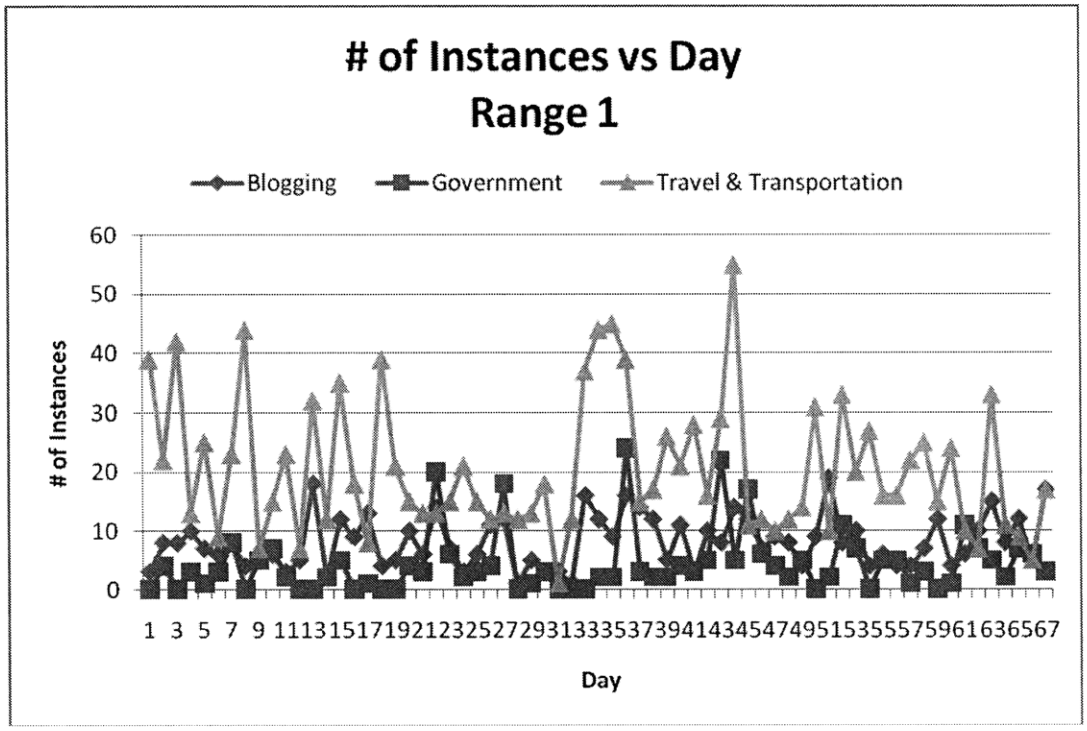

Figure 54 - # of Instances vs Day. Range 1. Blogging, Government, Travel & Transportation

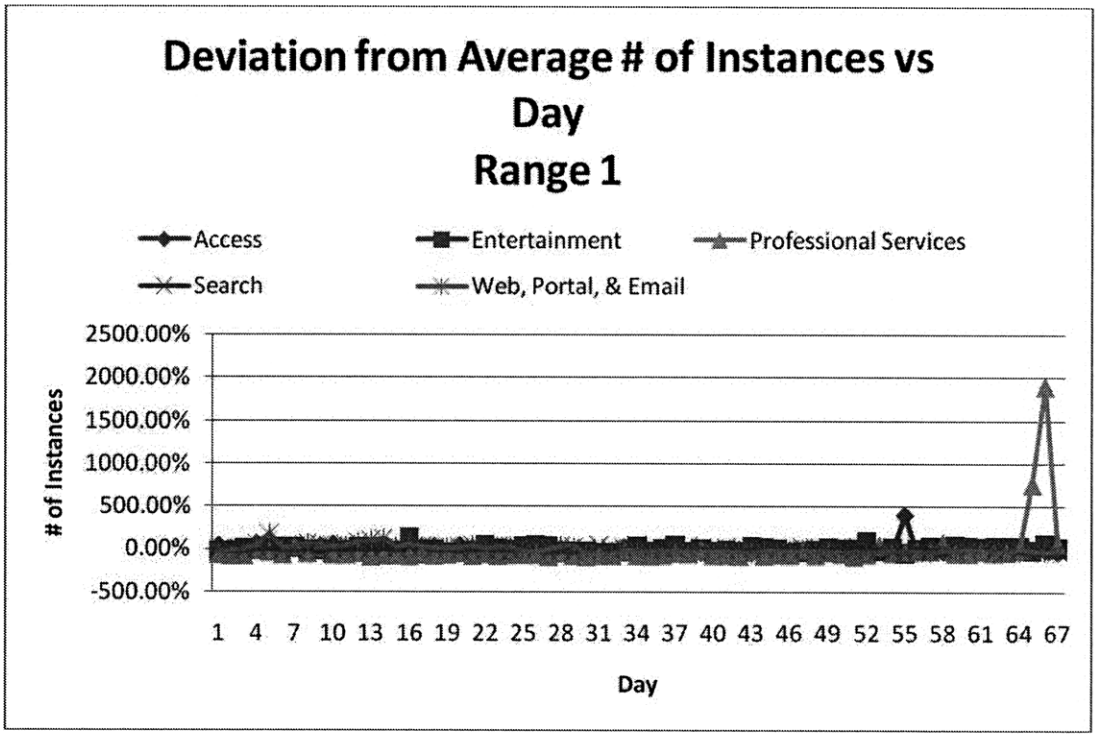

Figure 55 - Deviation from Average # of Instances vs Day. Range 1. Access, Entertainment, Professional Services, Search, Web Portal & Email

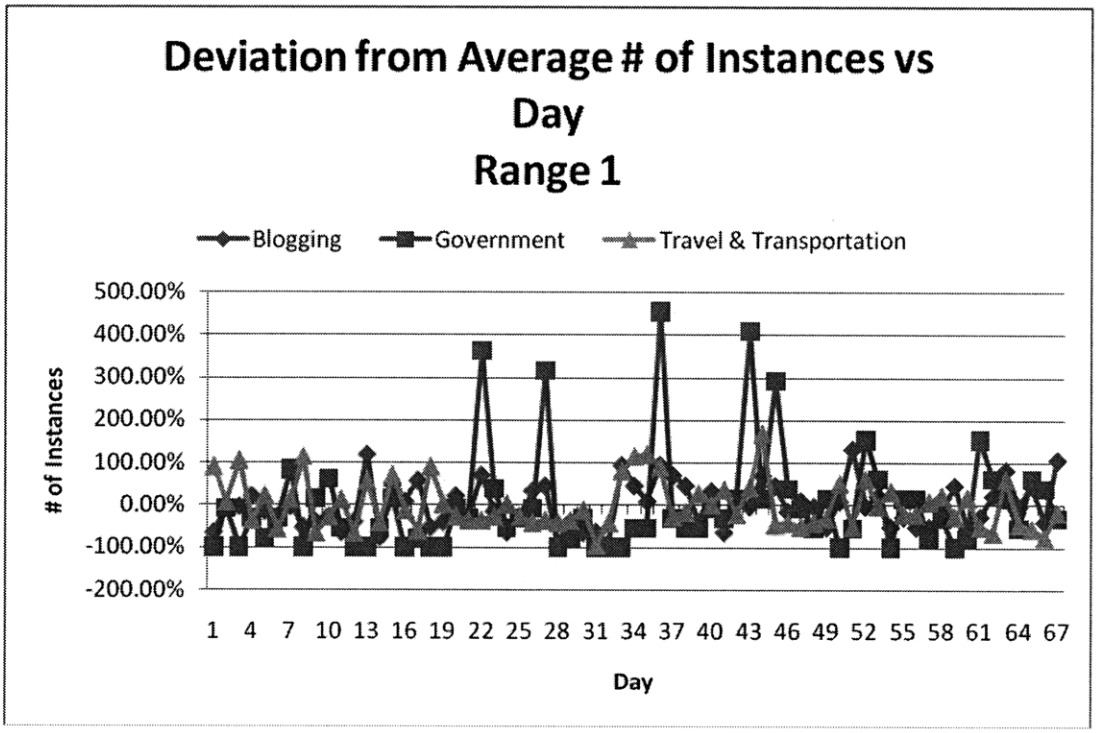

Figure 56 - Deviation from Average # of Instances vs Day. Range 1. Blogging, Government, Travel & Transportation

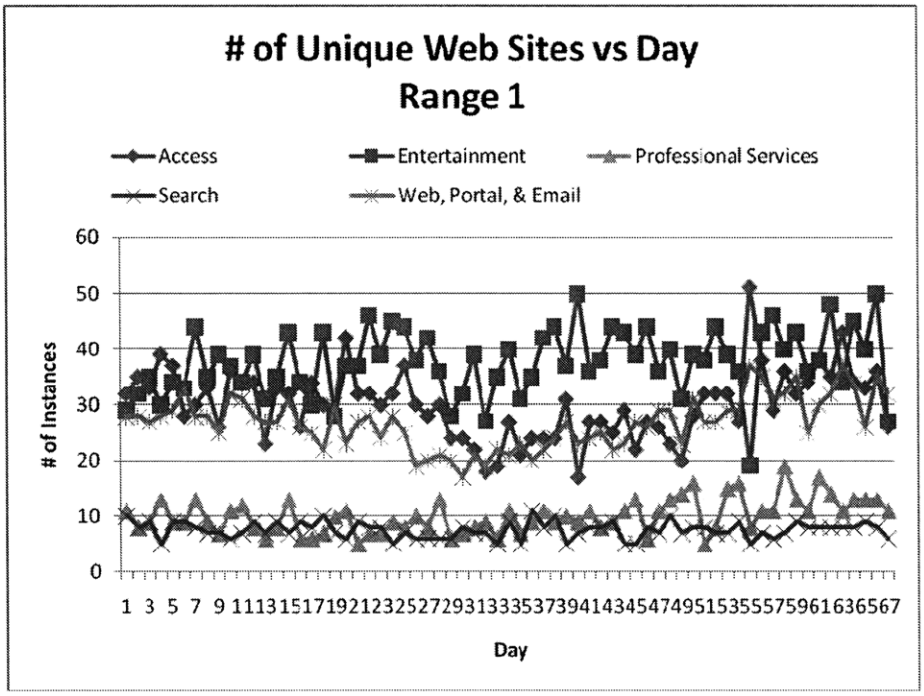

Figure 57 - # of Unique Web Sites vs Day. Range 1. Access, Entertainment, Professional Services, Search, Web Portal & Email

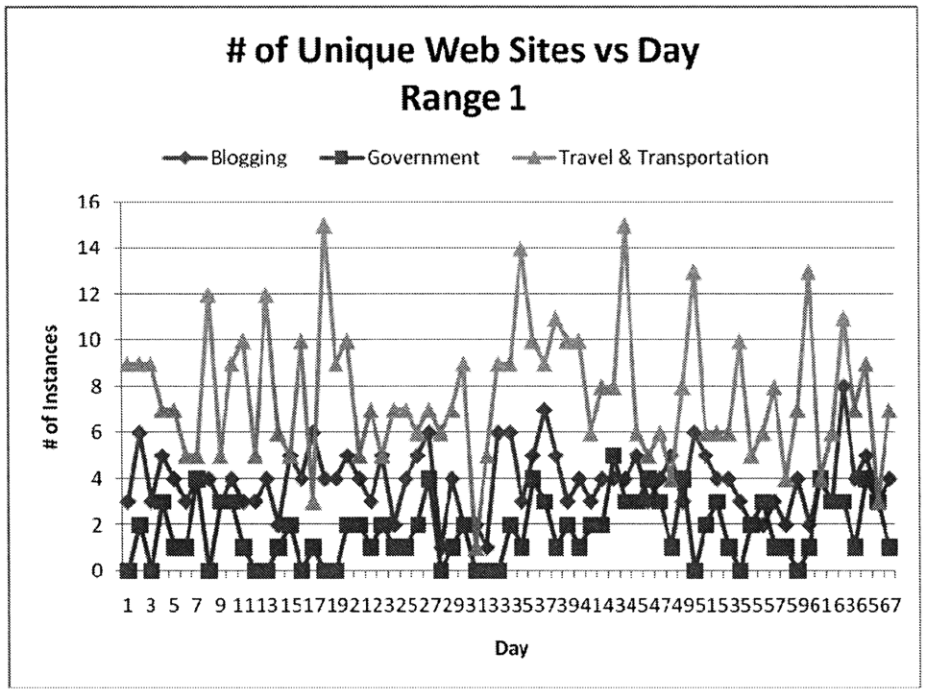

Figure 58 - # of Unique Web Sites vs Day. Range 1. Blogging, Government, Travel & Transportation

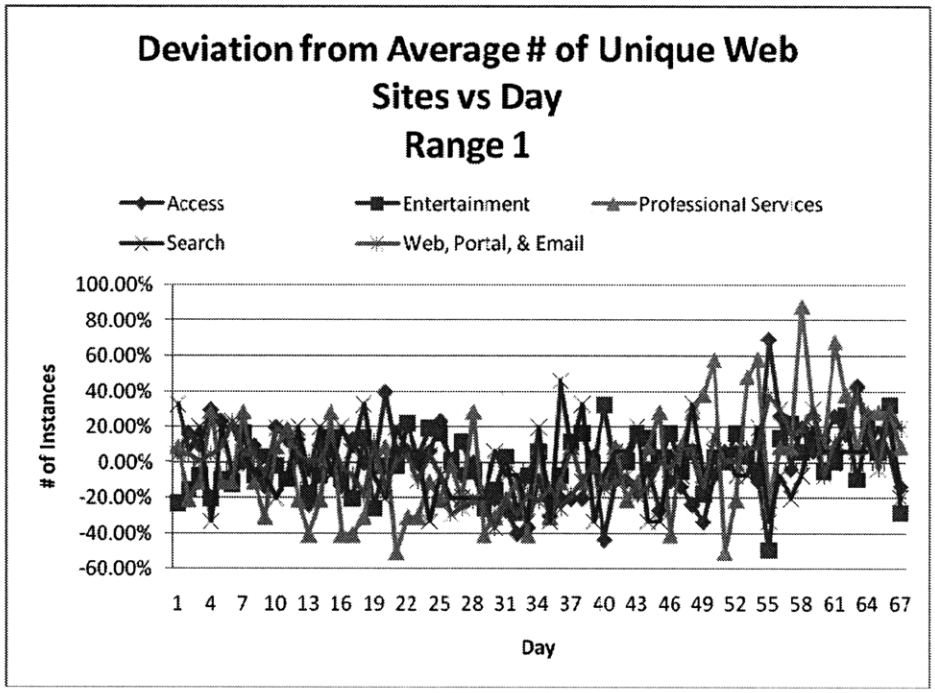

Figure 59 - Deviation from Average # of Unique Web Sites vs Day. Range 1. Access, Entertainment, Professional Services, Search, Web Portal & Email

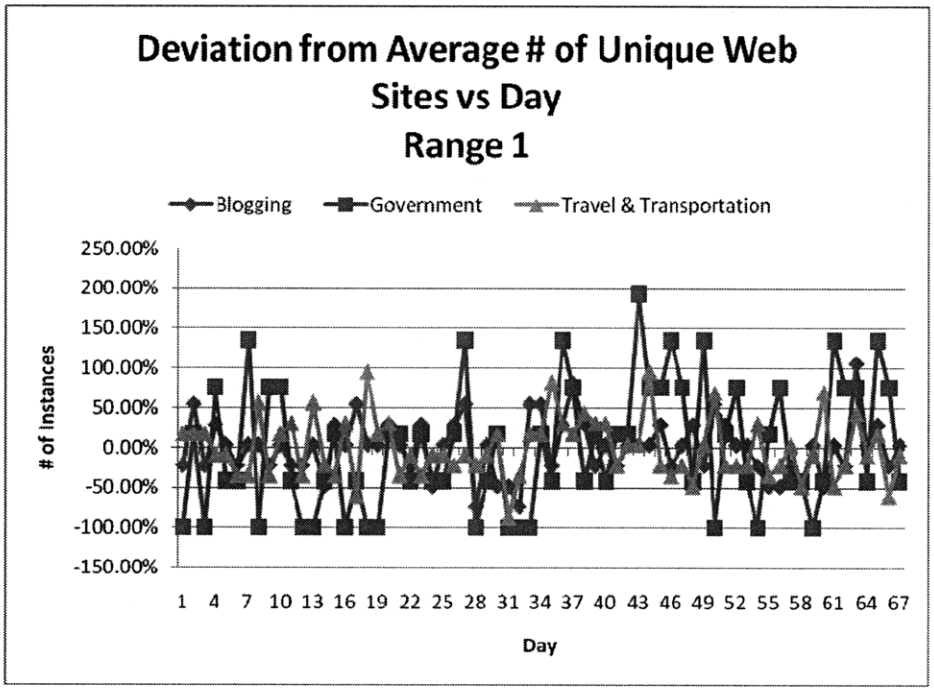

Figure 60 - Deviation from Average # of Unique Web Sites vs Day. Range 1. Blogging, Government, Travel & Transportation

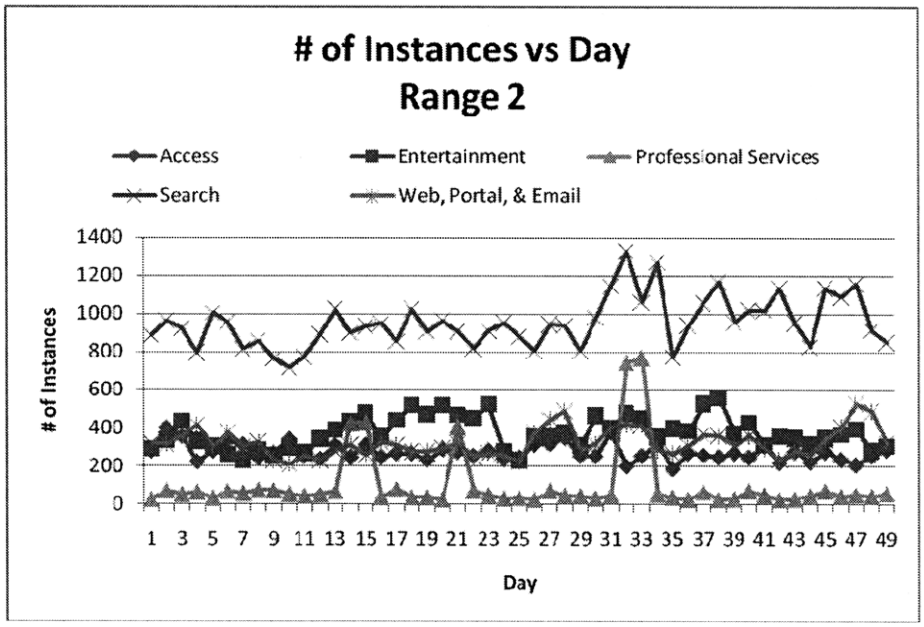

Figure 61 - # of Instances vs Day. Range 2. Access, Entertainment, Professional Services, Search, Web Portal & Email

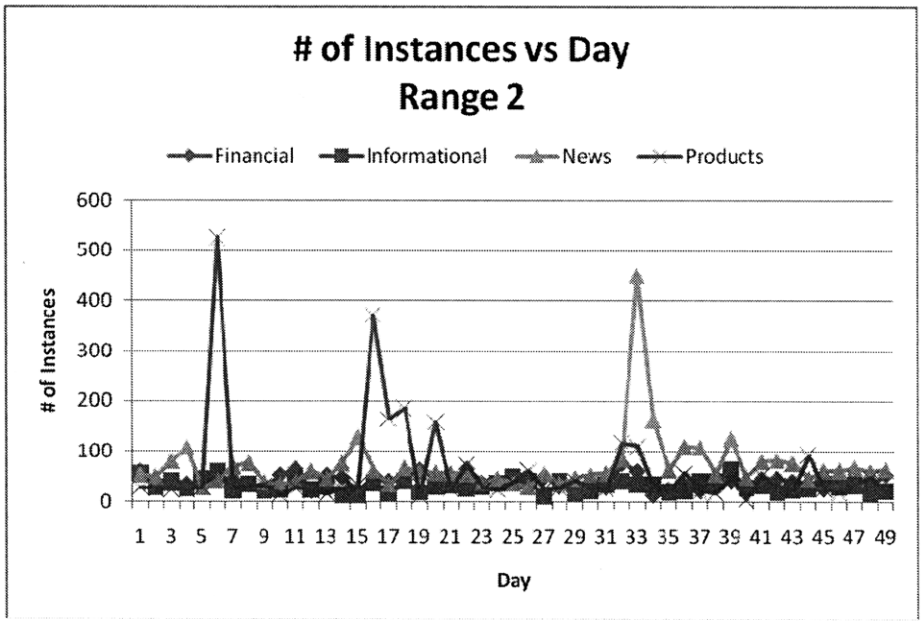

Figure 62 - # of Instances vs Day. Range 2. Financial, Information, News, Products

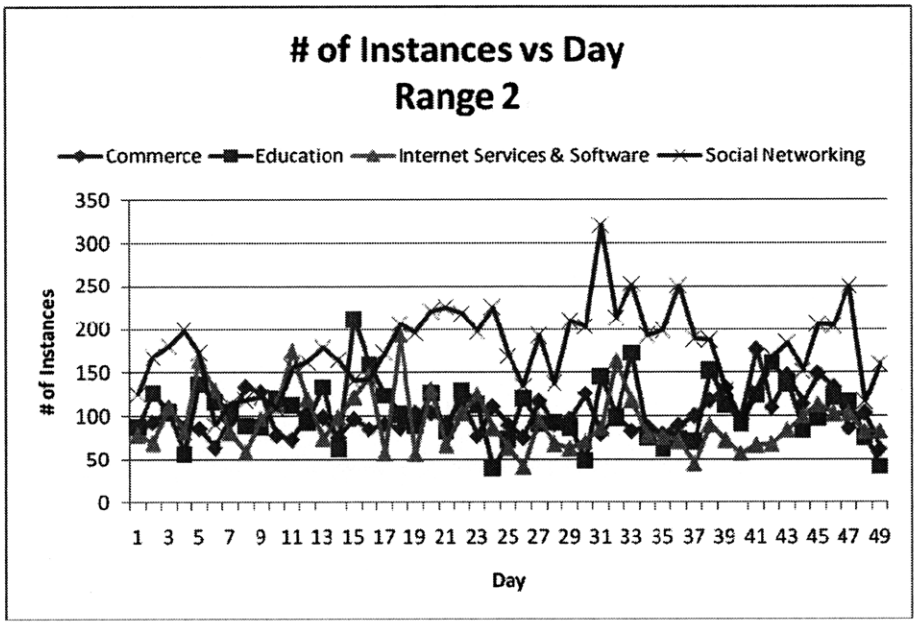

Figure 63 - # of Instances vs Day. Range 2. Commerce, Education. Internet Services & Software, Social Network

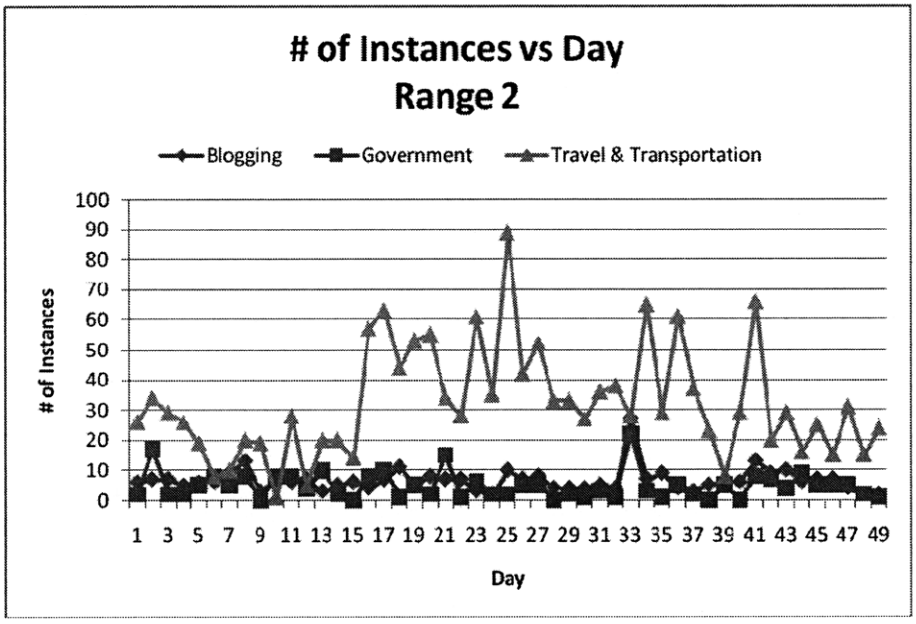

Figure 64 - # of Instances vs Day. Range 2. Blogging, Government, Travel & Transportation

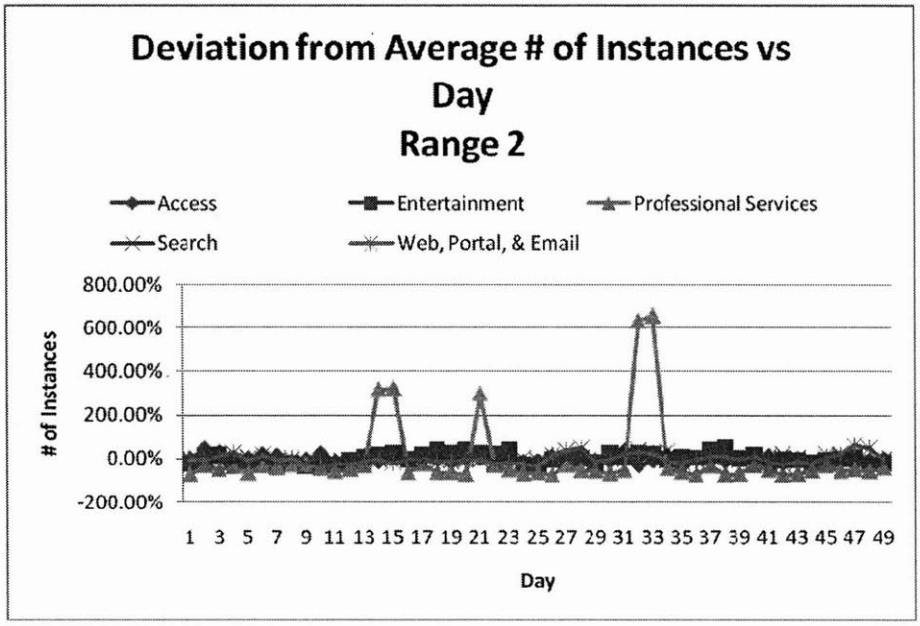

Figure 65 - Deviation from Average # of Instances vs Day. Range 2. Access, Entertainment, Professional Services, Search, Web Portal & Email

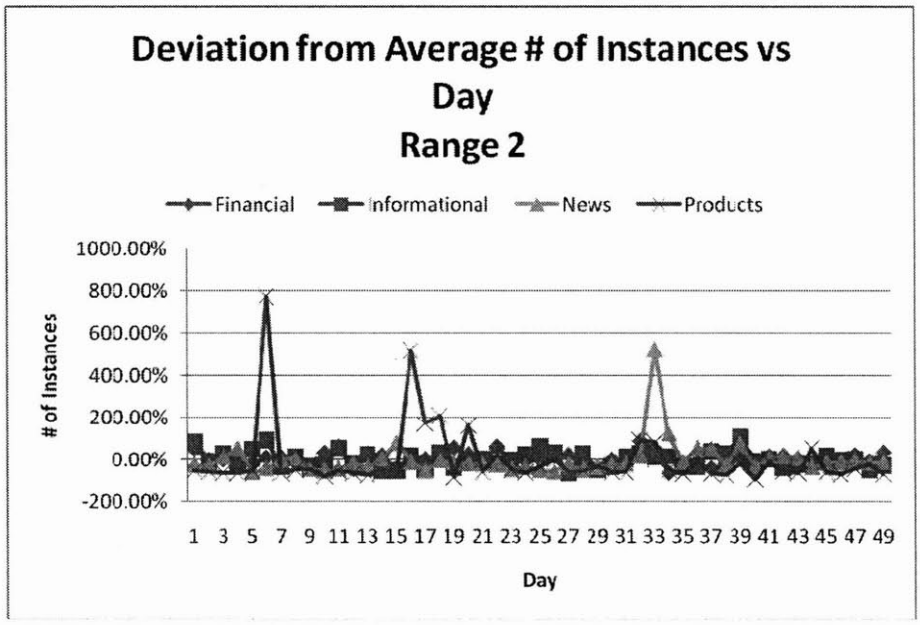

Figure 66 - Deviation from Average # of Instances vs Day. Range 2. Financial, Information, News, Products

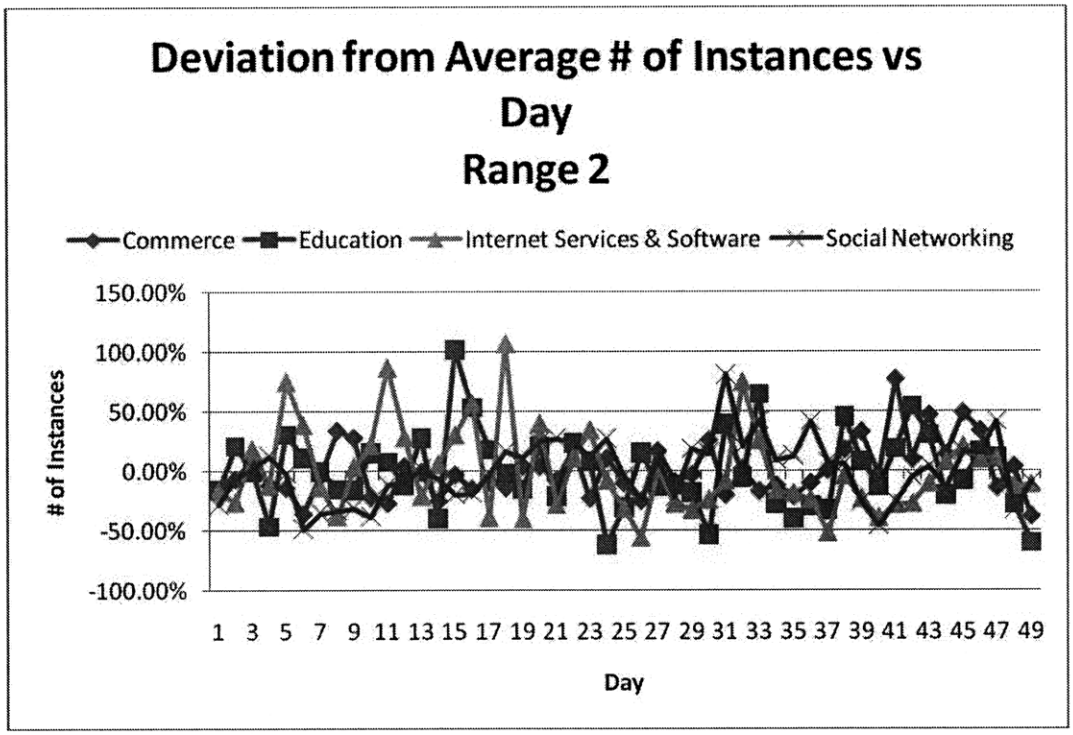

Figure 67 - Deviation from Average # of Instances vs Day. Range 2. Commerce, Education. Internet Services & Software, Social Network

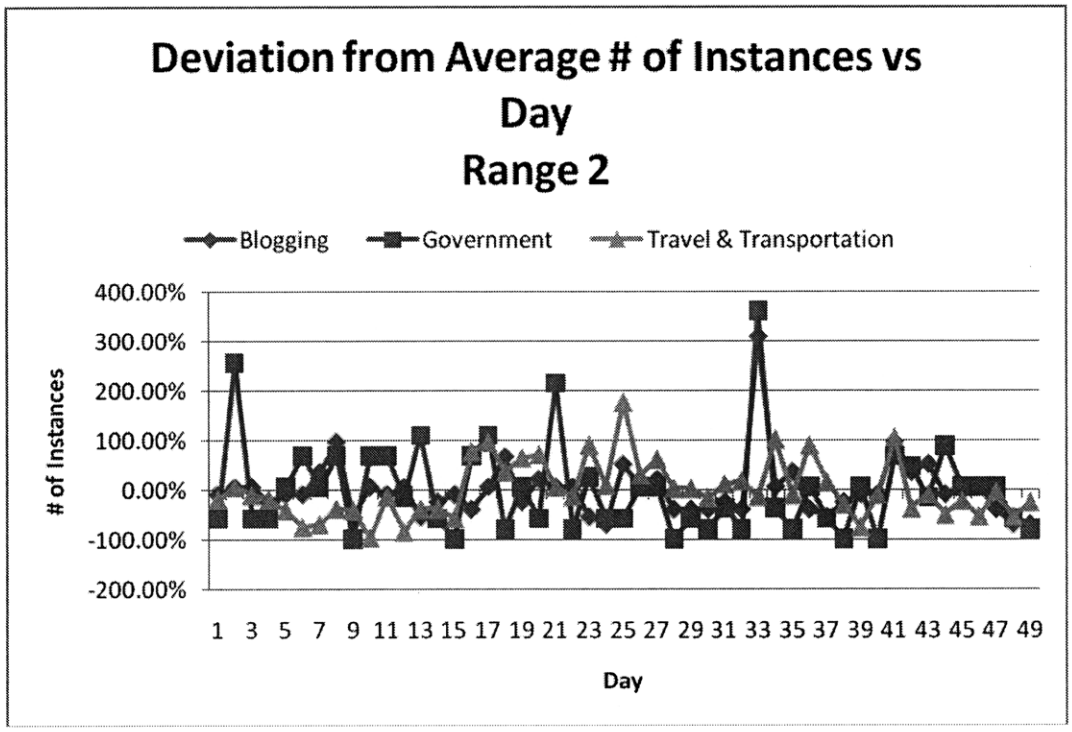

Figure 68 - Deviation from Average # of Instances vs Day. Range 2. Blogging, Government, Travel & Transportation

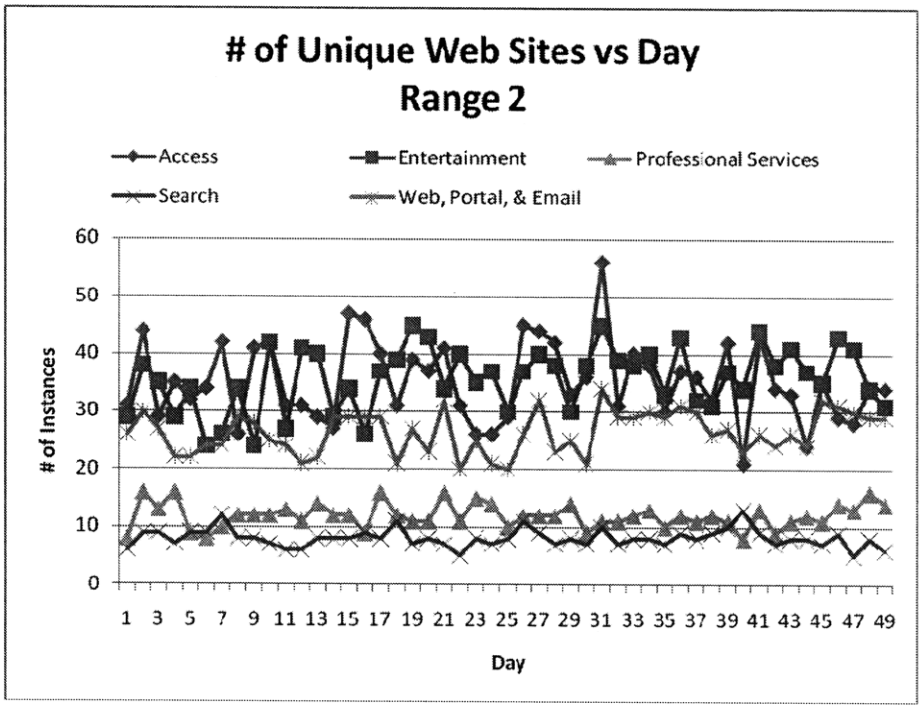

Figure 69 - # of Unique Web Sites vs Day. Range 2. Access, Entertainment, Professional Services, Search, Web Portal & Email

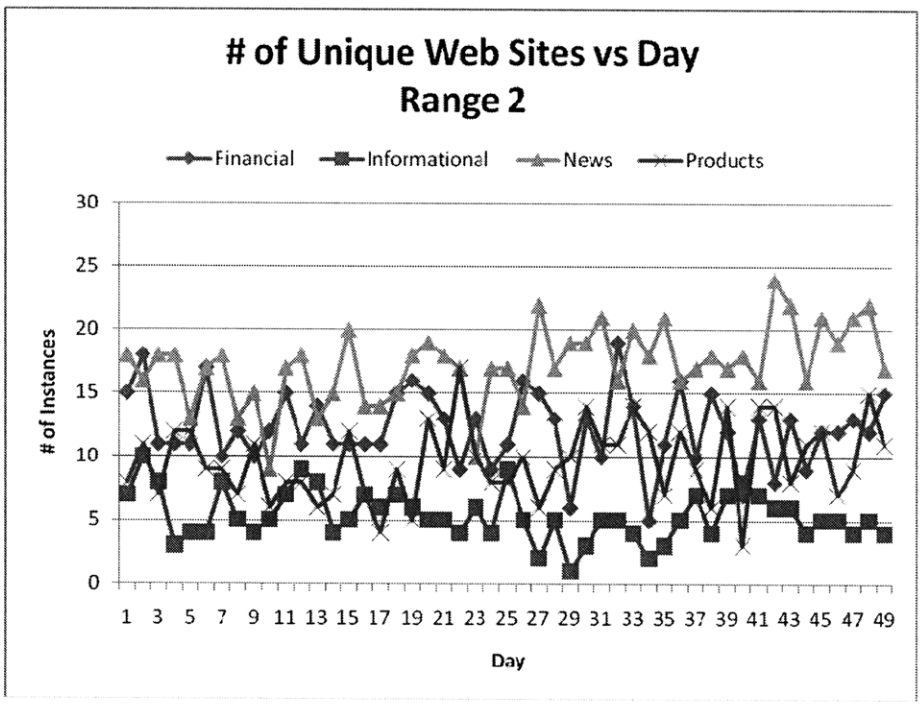

Figure 70 - # of Unique Web Sites vs Day. Range 2. Financial, Information, News, Products

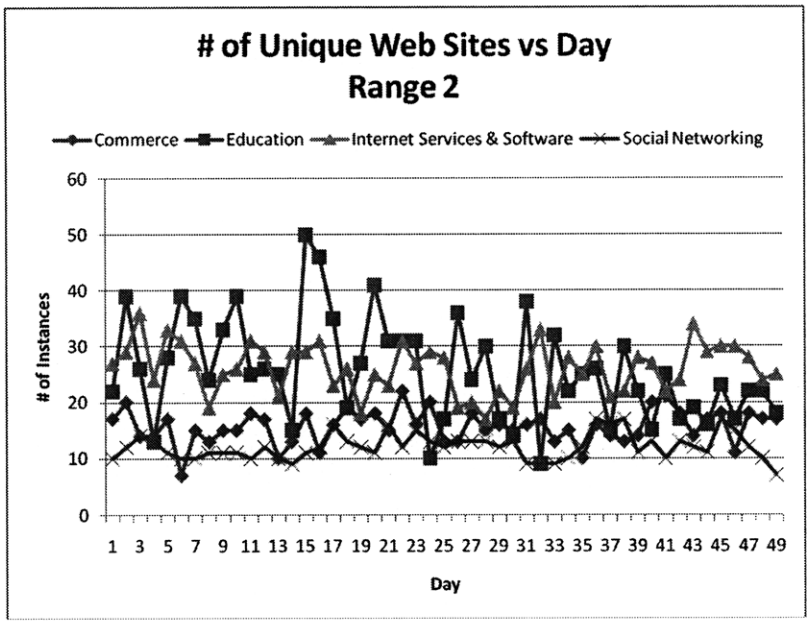

Figure 71 - # of Unique Web Sites vs Day. Range 2. Commerce, Education. Internet Services & Software, Social Network

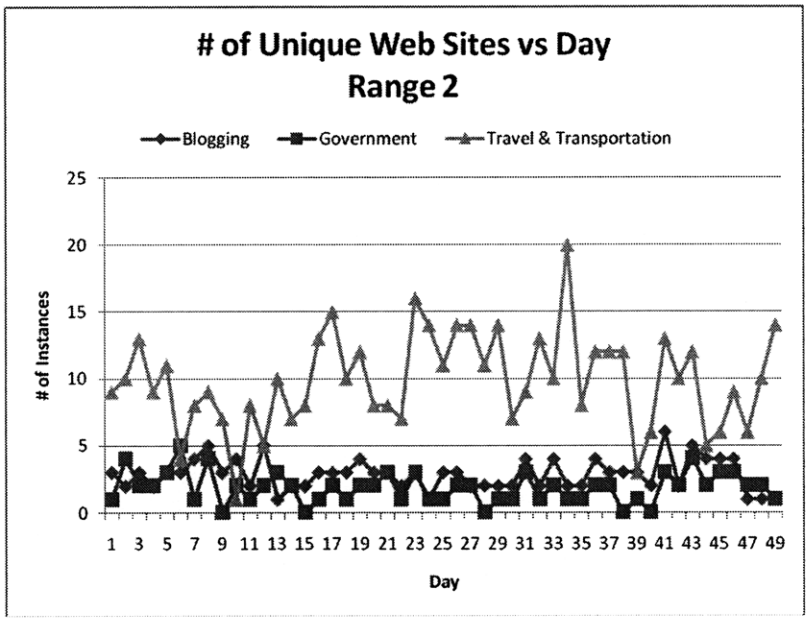

Figure 72 - # of Unique Web Sites vs Day. Range 2. Blogging, Government, Travel & Transportation

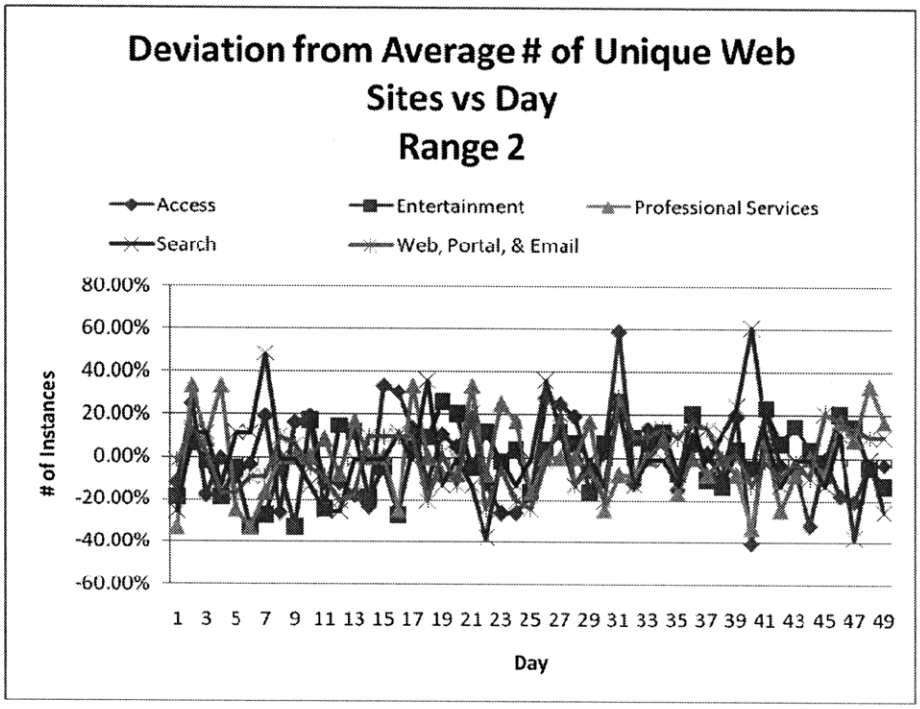

Figure 73 - Deviation from Average # of Unique Web Sites vs Day. Range 2. Access, Entertainment, Professional Services, Search, Web Portal & Email

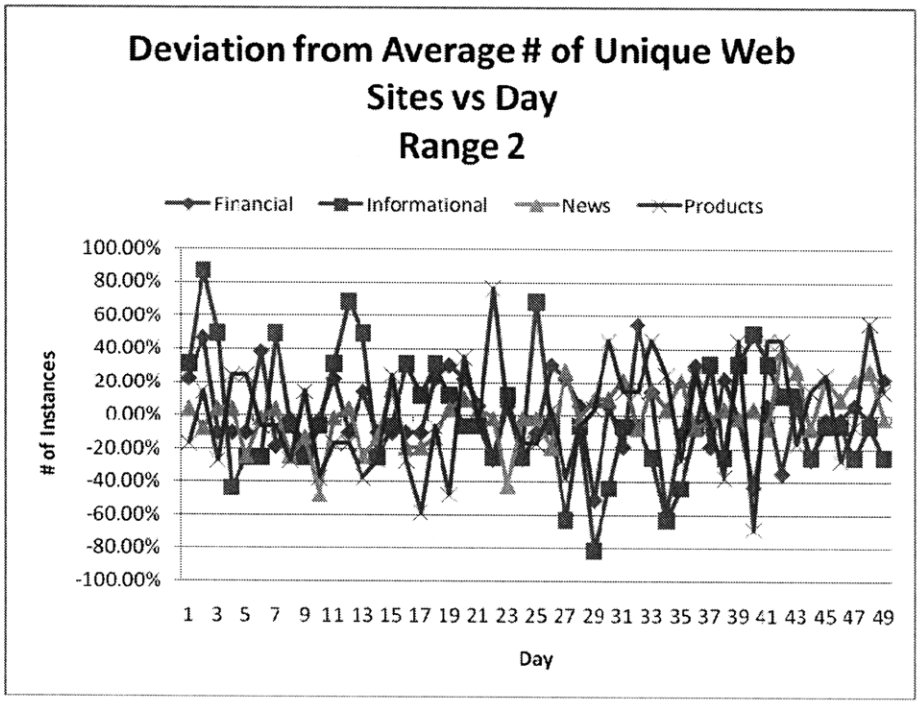

Figure 74 - Deviation from Average # of Unique Web Sites vs Day. Range 2. Financial, Information, News, Products

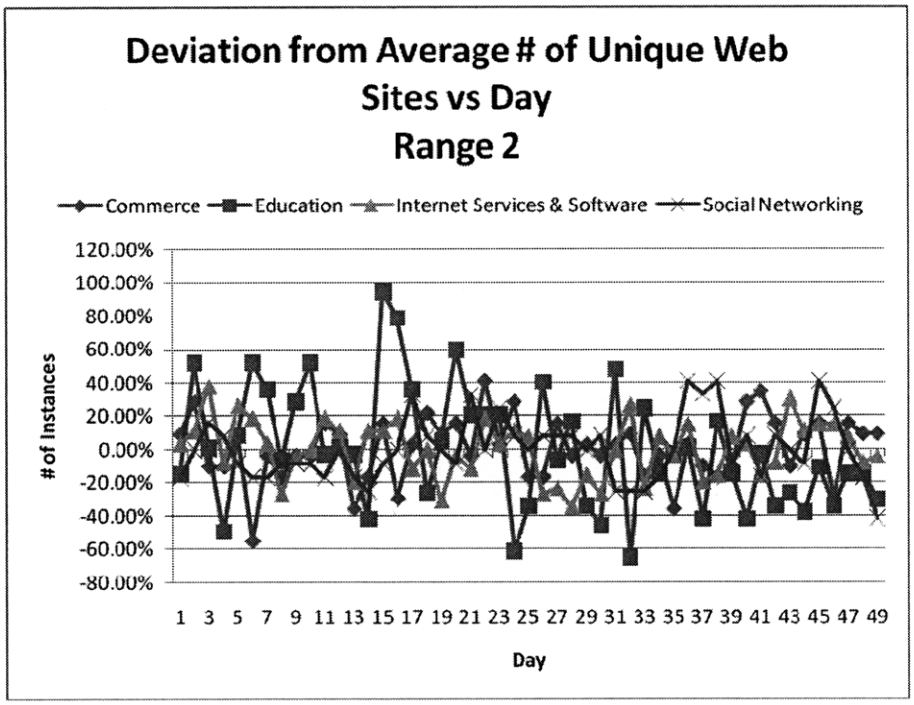

Figure 75 - Deviation from Average # of Unique Web Sites vs Day. Range 2. Commerce, Education. Internet Services & Software, Social Network

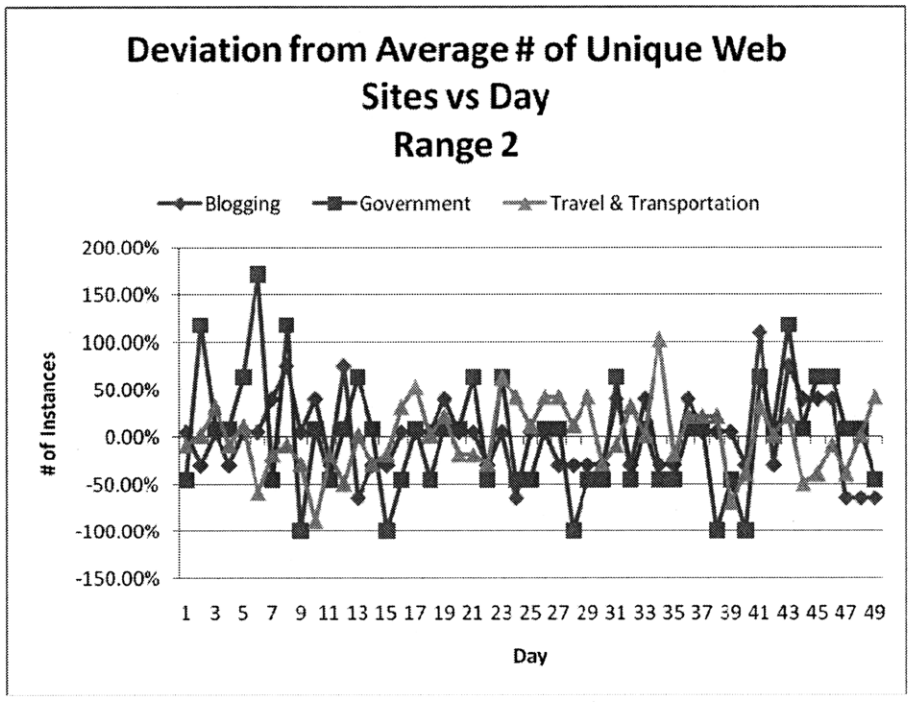

Figure 76 - Deviation from Average # of Unique Web Sites vs Day. Range 2. Blogging, Government, Travel & Transportation

Appendix C: Sub-Category Breakdowns

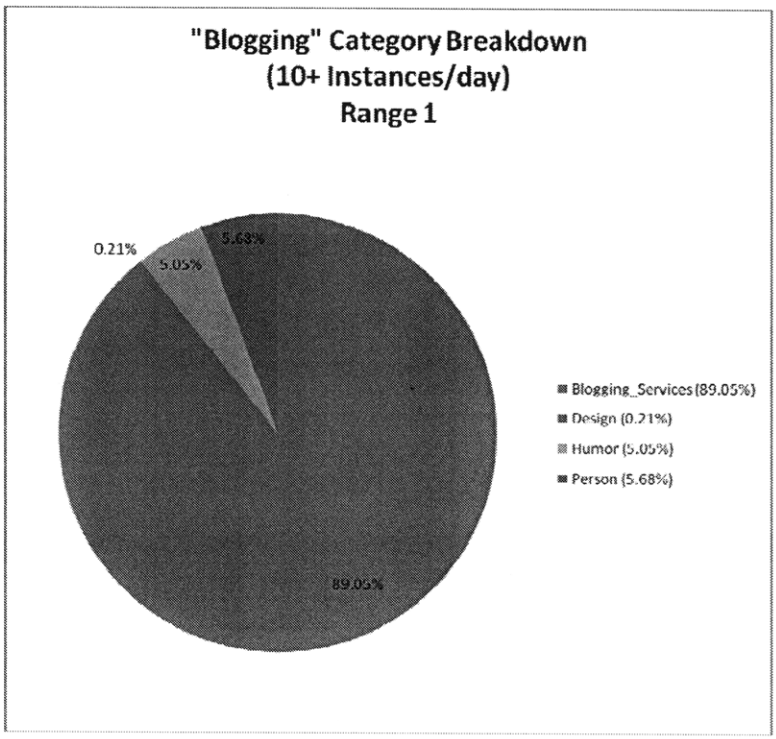

Figure 77 - "Blogging" Category Breakdown. 10+ Instances/day. Range 1

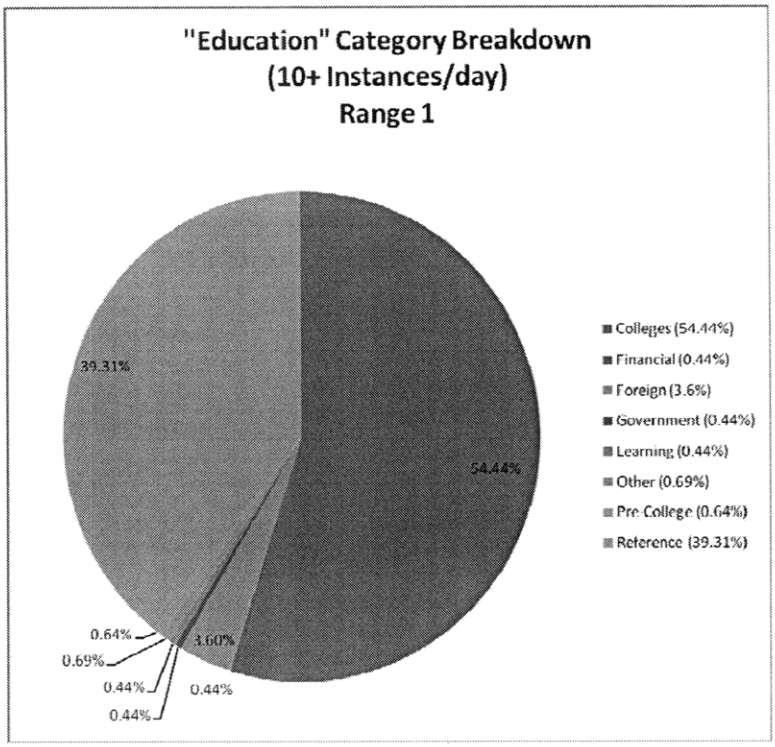

Figure 78 - "Education" Category Breakdown. 10+ Instances/day. Range 1

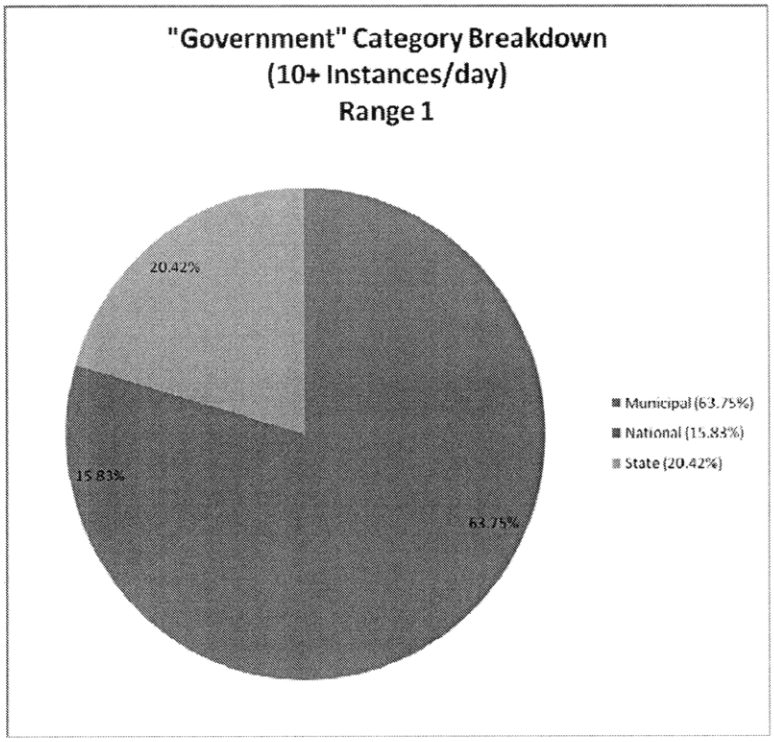

Figure 79 - "Government" Category Breakdown. 10+ Instances/day. Range 1

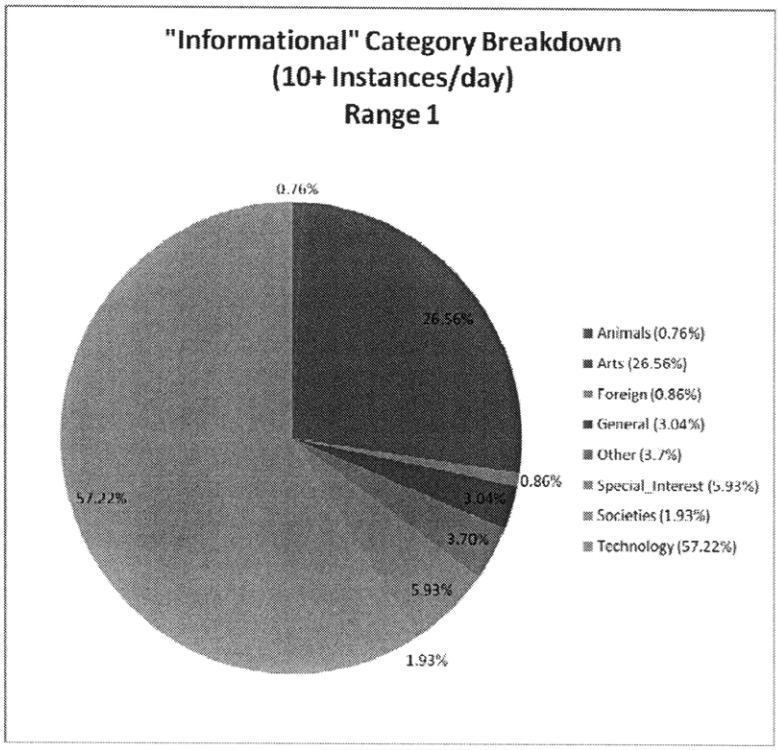

Figure 80 - "Informational" Category Breakdown. 10+ Instances/day. Range 1

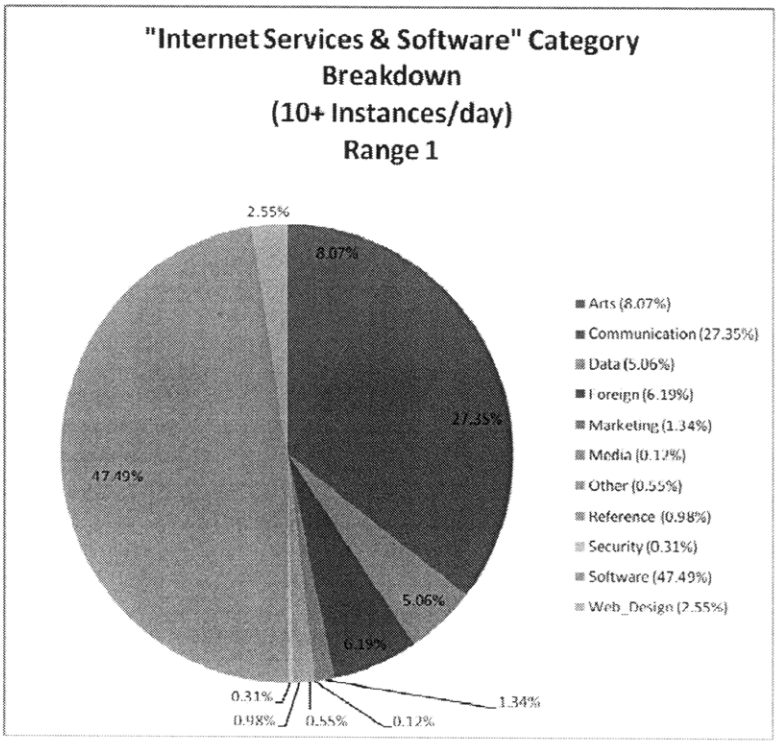

Figure 81 - "Internet Services & Software" Category Breakdown. 10+ Instances/day. Range 1

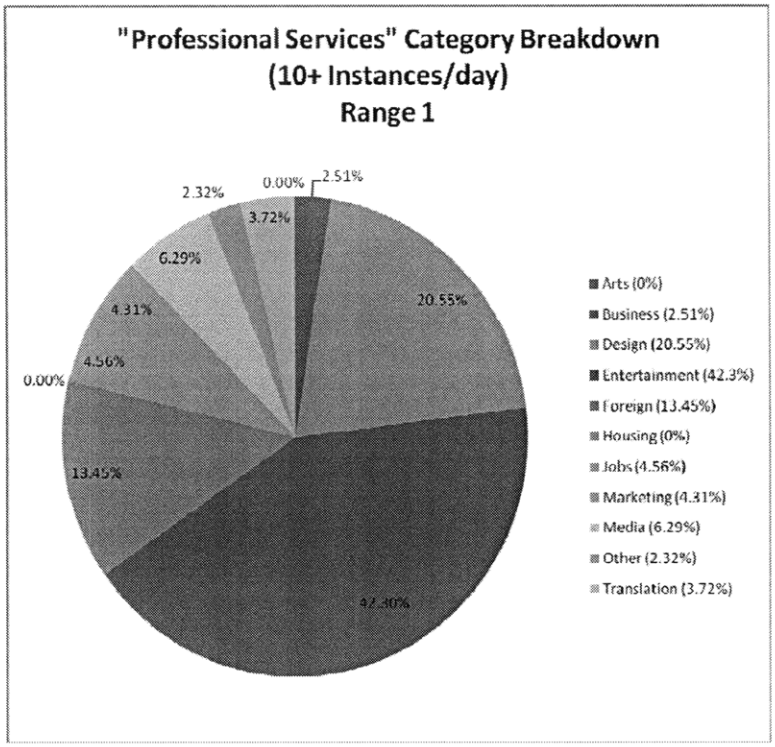

Figure 82 - "Professional Services" Category Breakdown. 10+ Instances/day. Range 1

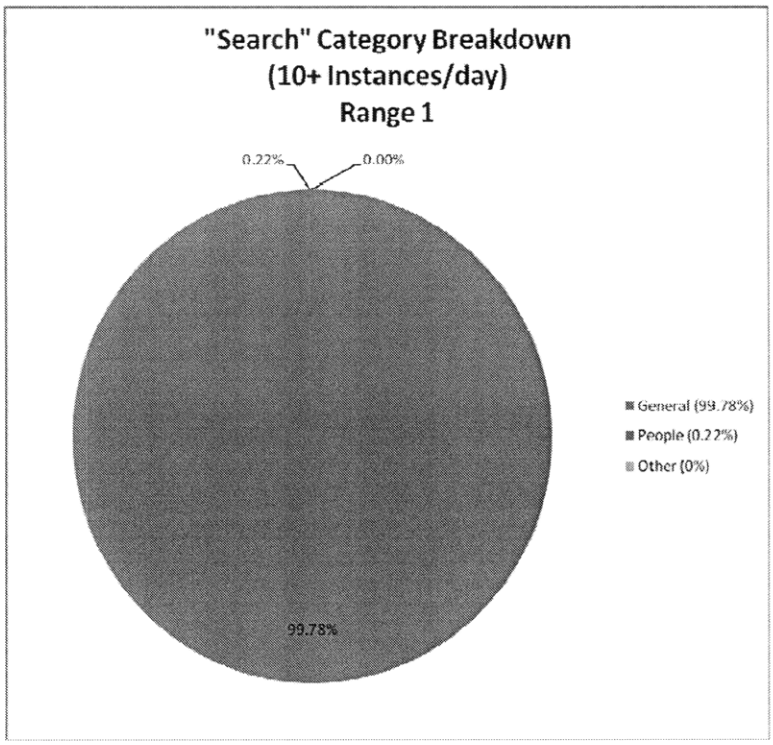

Figure 83 - "Search" Category Breakdown. 10+ Instances/day. Range 1

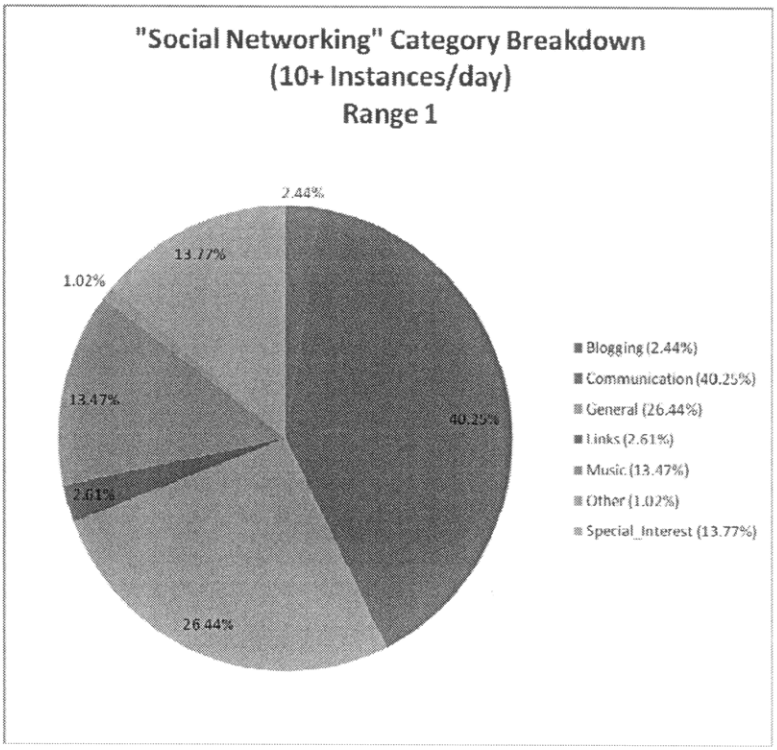

Figure 84 - "Social Networking" Category Breakdown. 10+ Instances/day. Range 1

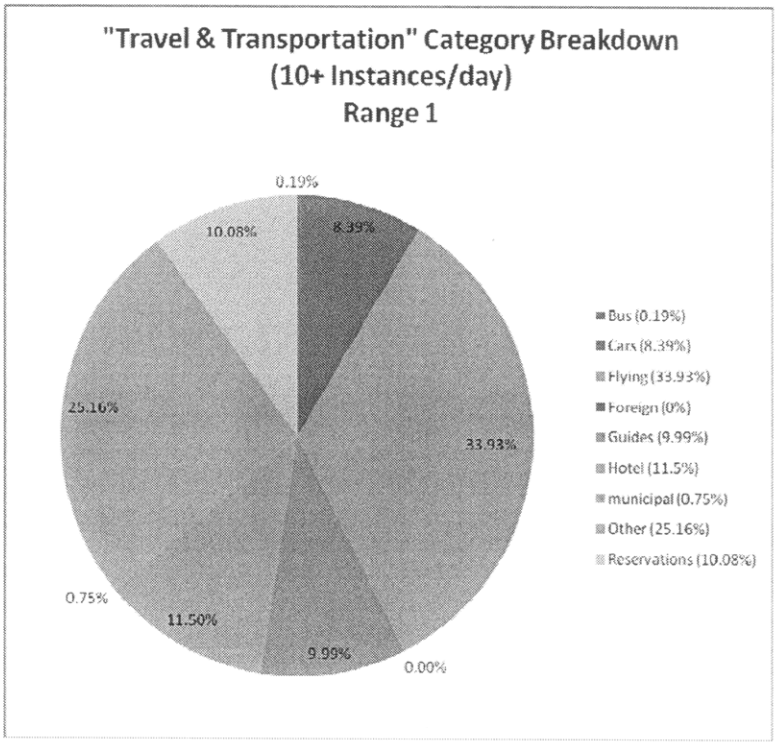

Figure 85 - "Travel & Transportation" Category Breakdown. 10+ Instances/day. Range 1

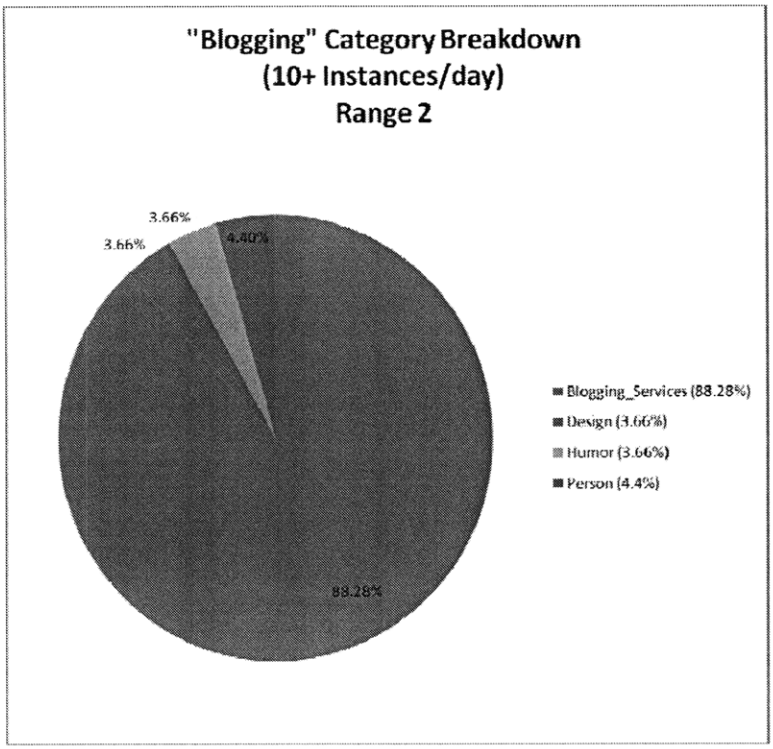

Figure 86 - "Blogging" Category Breakdown. 10+ Instances/day. Range 2

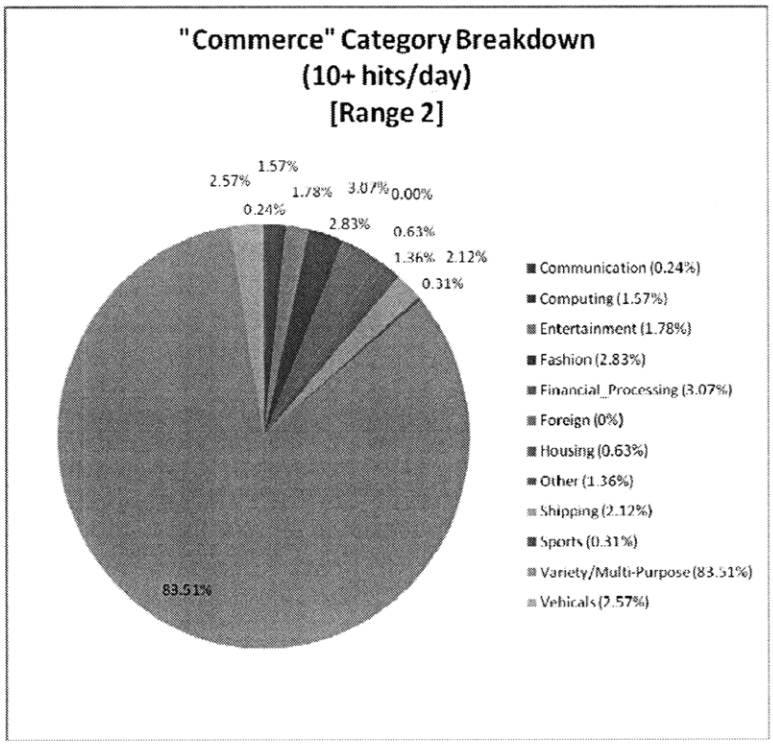

Figure 87 - "Commerce" Category Breakdown. 10+ Instances/day. Range 2

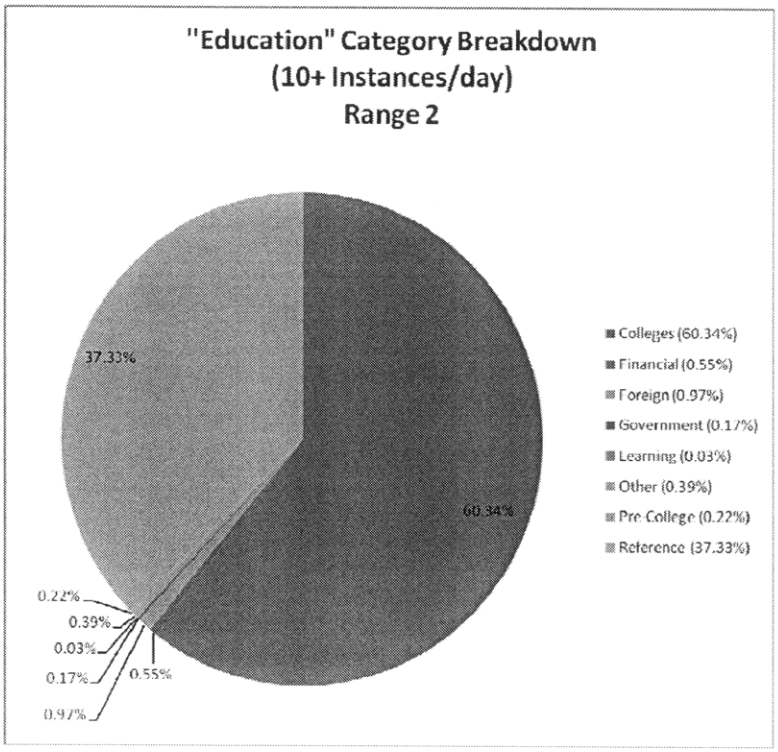

Figure 88 - "Education" Category Breakdown. 10+ Instances/day. Range 2

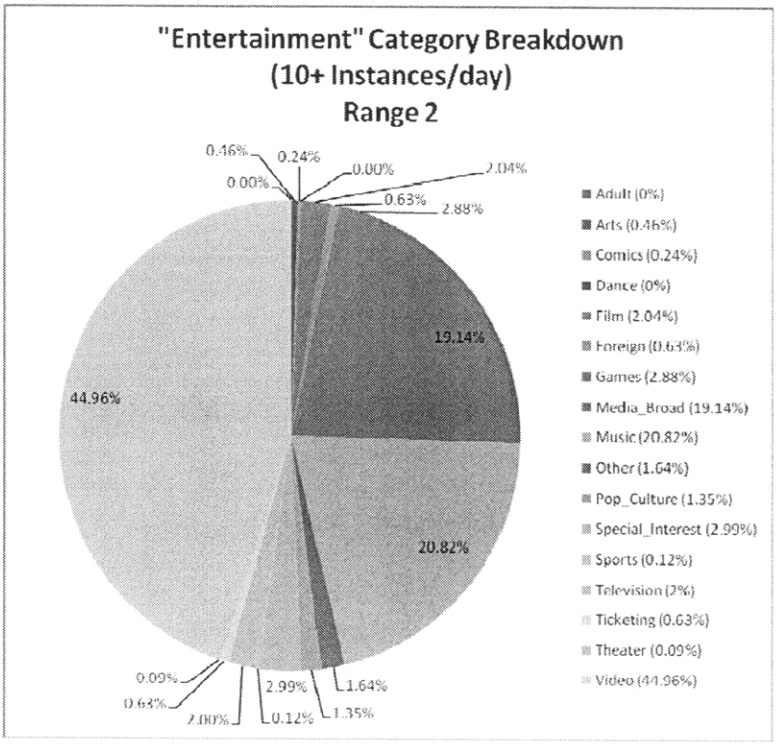

Figure 89 - "Entertainment" Category Breakdown. 10+ Instances/day. Range 2

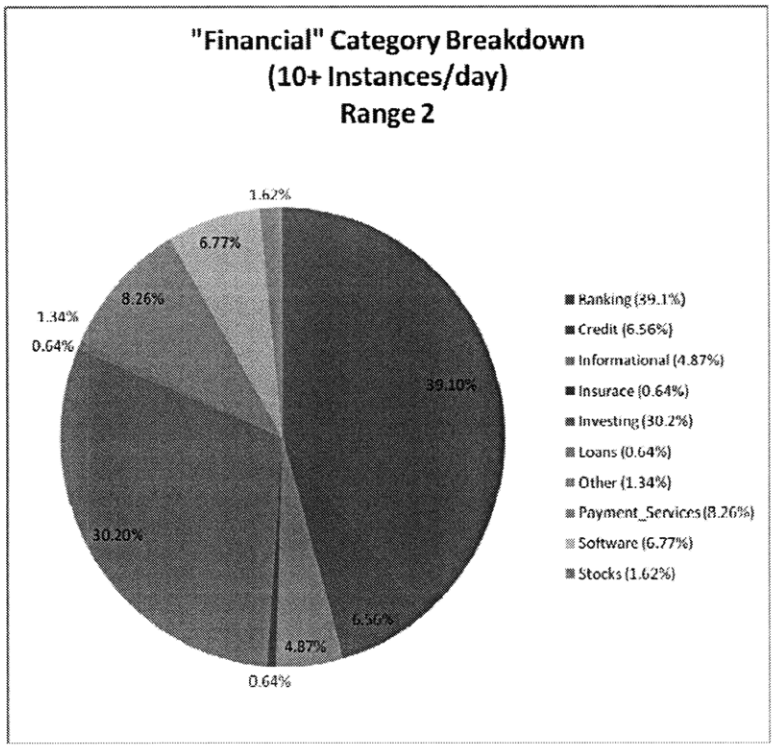

Figure 90 - "Financial" Category Breakdown. 10+ Instances/day. Range 2

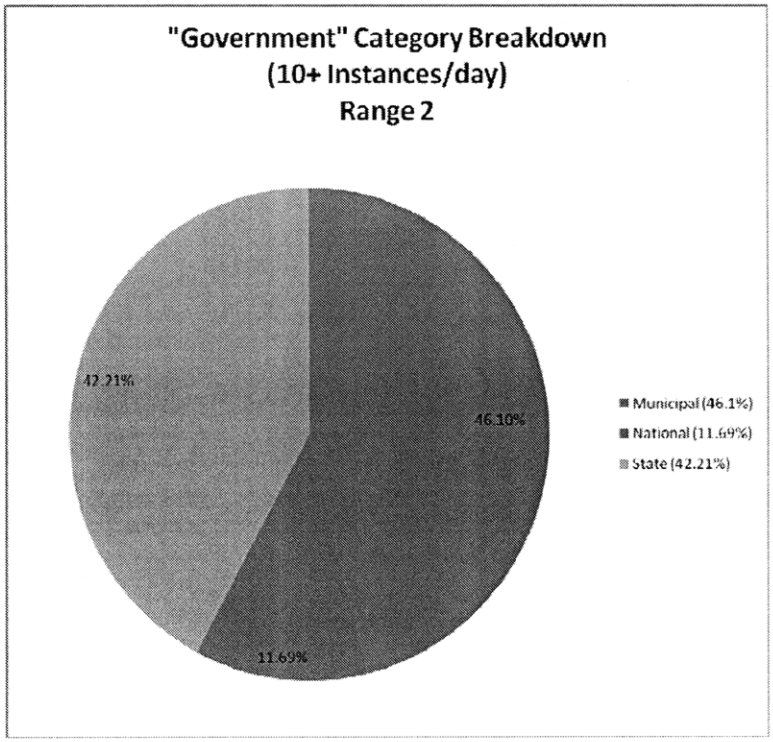

Figure 91 - "Government" Category Breakdown. 10+ Instances/day. Range 2

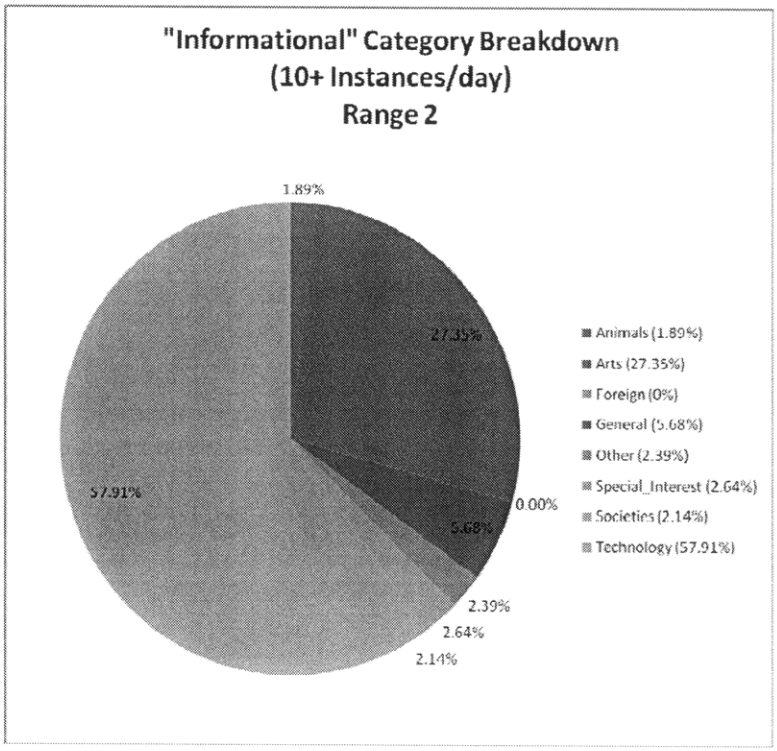

Figure 92 - "Informational" Category Breakdown. 10+ Instances/day. Range 2

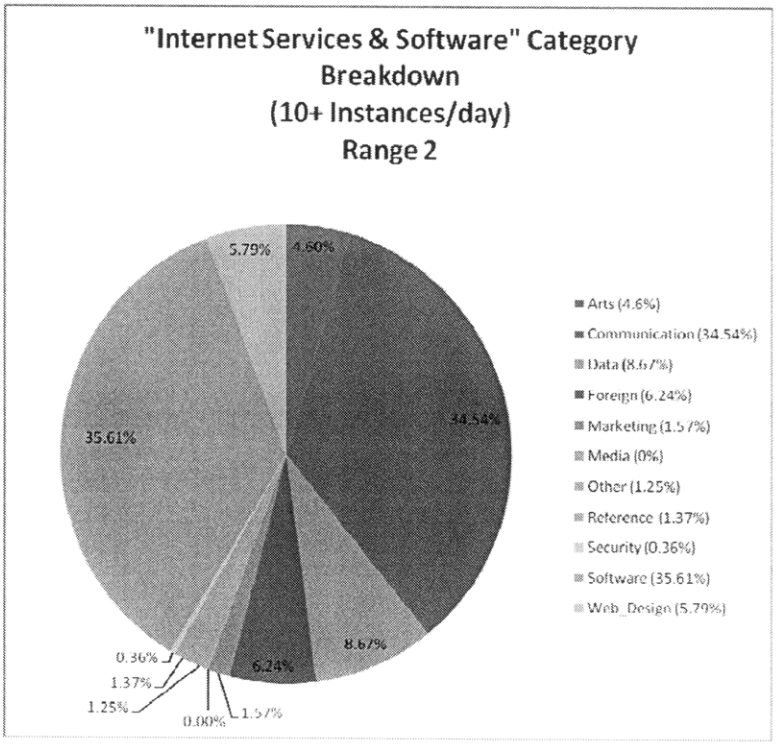

Figure 93 - "Internet Services & Software" Category Breakdown. 10+ Instances/day. Range 2

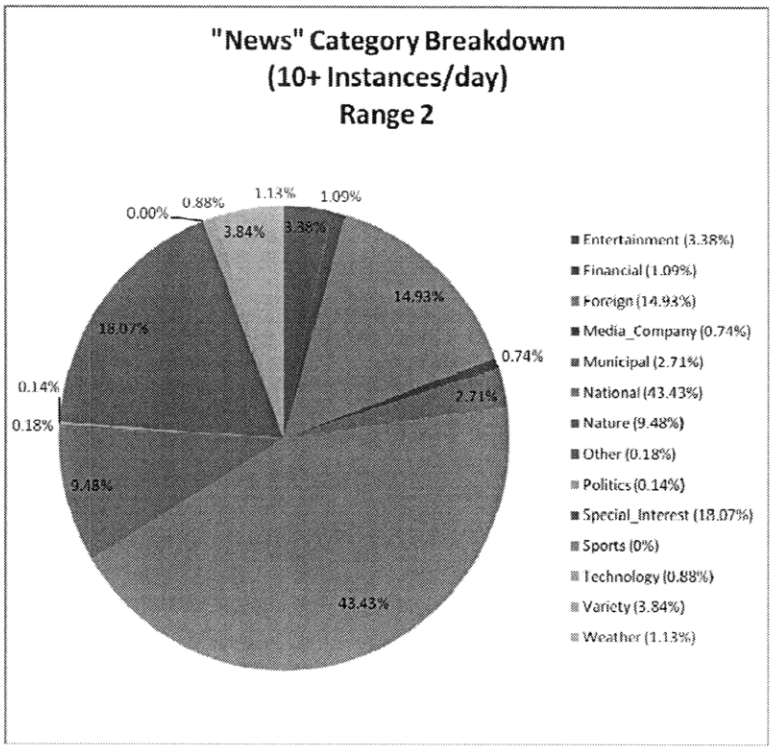

Figure 94 - "News" Category Breakdown. 10+ Instances/day. Range 2

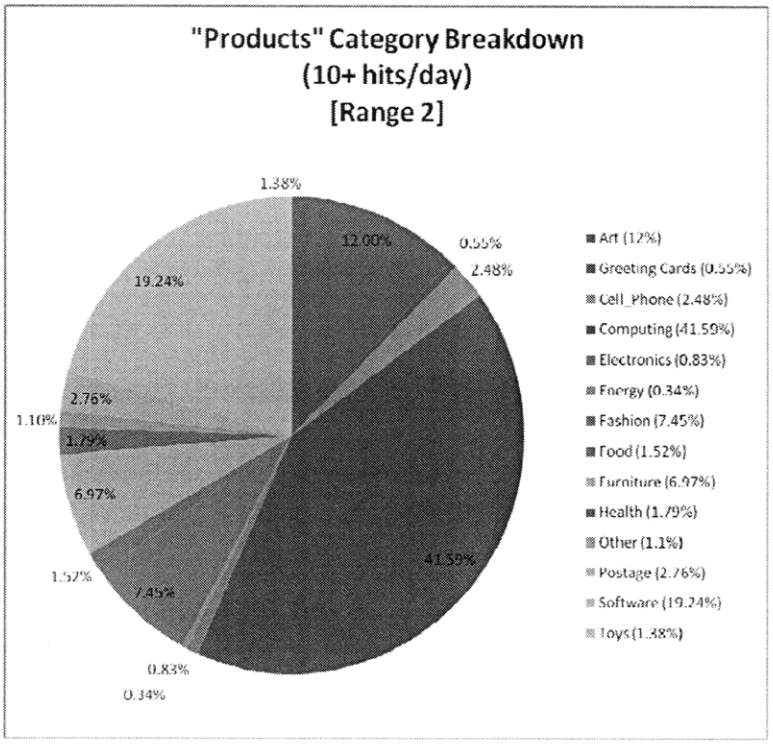

Figure 95 - "Products" Category Breakdown. 10+ Instances/day. Range 2

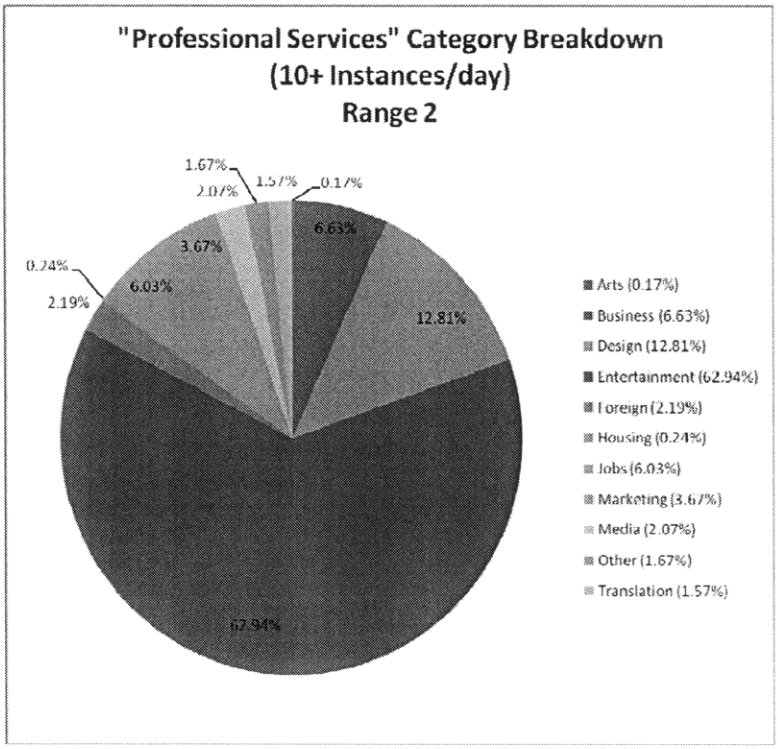

Figure 96 - "Professional Services" Category Breakdown. 10+ Instances/day. Range 2

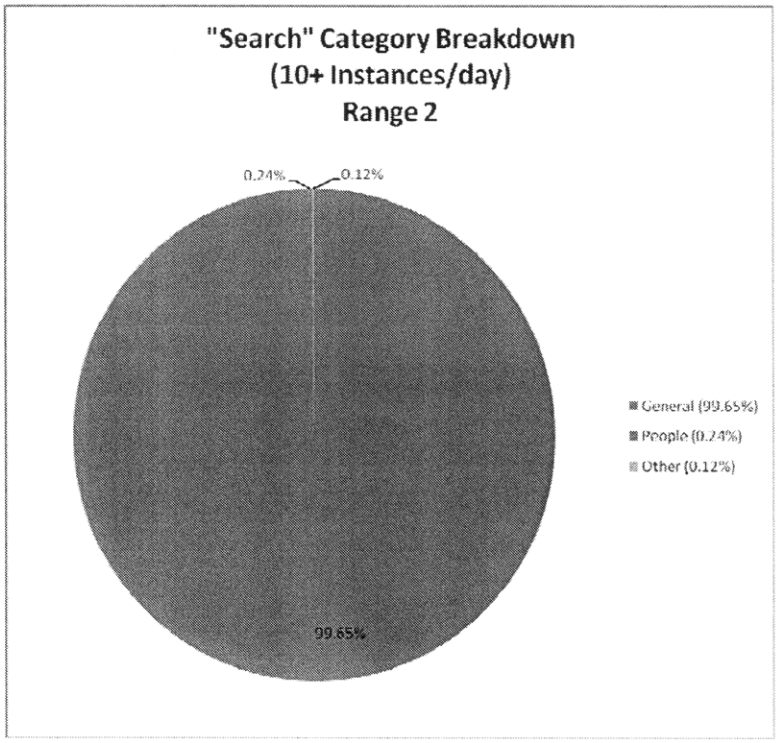

Figure 97 - "Search" Category Breakdown. 10+ Instances/day. Range 2

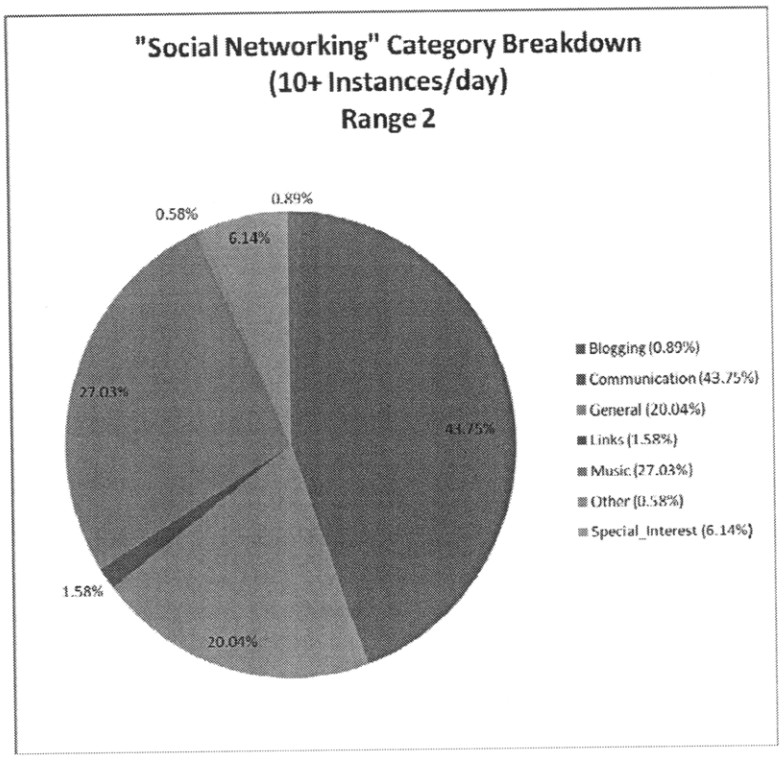

Figure 98 - "Social Networking" Category Breakdown. 10+ Instances/day. Range 2

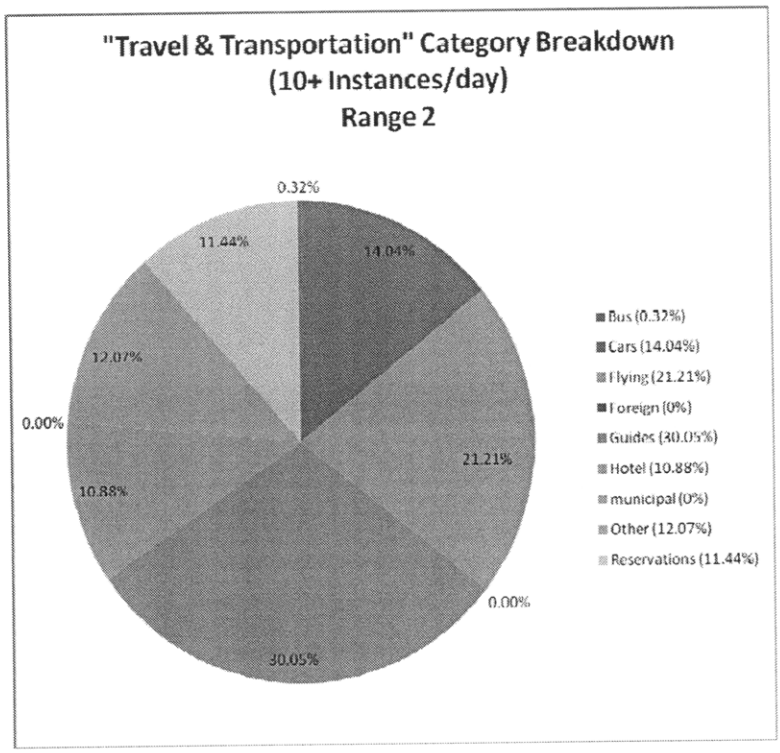

Figure 99 - "Travel & Transportation" Category Breakdown. 10+ Instances/day. Range 2

Appendix D: Top Web Sites

Top 40 Visited Web Sites (Range 2)			
	Web Site	**# Instances**	**Percentage**
1	google.com	24819	26.22%
2	yahoo.com	20111	21.25%
3	youtube.com	8509	8.99%
4	google.com (email)	3852	4.07%
5	meebo.com	3555	3.76%
6	yahoo.com (email)	2770	2.93%
7	intdm.com	2639	2.79%
8	live.com	2256	2.38%
9	imeem.com	2188	2.31%
10	amazon.com	2186	2.31%
11	warnerbros.com	2117	2.24%
12	facebook.com	1879	1.99%
13	mailcenter.comcast.net	1805	1.91%
14	wikimedia.org	1695	1.79%
15	mac.com	1562	1.65%
16	apple.com	1548	1.64%
17	sony.com	1523	1.61%
18	aol.com	1282	1.35%
19	cnn.com	1205	1.27%
20	ebay.com	858	0.91%
21	cnet.com	855	0.90%
22	start.com	757	0.80%
23	live365.com	586	0.62%
24	nonfatmedia.com	577	0.61%
25	aol.com (email)	544	0.57%
26	groove.net	539	0.57%
27	microsoft.com	507	0.54%
28	craigslist.org	506	0.53%
29	sbcglobal.net	504	0.53%
30	hotmail.com	482	0.51%
31	bankofamerica.com	442	0.47%
32	arttoday.com	429	0.45%
33	attens.net	427	0.45%
34	rae.es	408	0.43%
35	usc.edu	396	0.42%
36	myspace.com	379	0.40%
37	ip.fastwebnet.it	345	0.36%
38	msn.com	339	0.36%

39	apartmenttherapy.com	336	0.35%
40	bydeluxe.com	334	0.35%
	Remaining	27439	28.99%

Table 30 - Top 40 Visited Web Sites (Range 2)

'Access' Category Top Web Sites by Overall Instances							
	Range 1				Range 2		
	Web Site	# Instances	Percentage		Web Site	# Instances	Percentage
1	comcast.net	4381	23.30%	1	charter.com	4044	26.11%
2	rr.com	2244	11.93%	2	comcast.net	3523	22.75%
3	charter.com	1819	9.67%	3	rr.com	1686	10.89%
4	pccwglobal.net	1343	7.14%	4	vzavenue.net	782	5.05%
5	cox.net	846	4.50%	5	cox.net	541	3.49%
6	mpowercom.net	815	4.33%	6	optonline.net	395	2.55%
7	metrored.net.mx	725	3.86%	7	rogers.com	272	1.76%
8	optonline.net	545	2.90%	8	pccwglobal.net	221	1.43%
9	rogers.com	469	2.49%	9	swbell.net	215	1.39%
10	verizon.net	362	1.92%	10	verizon.net	201	1.30%
	remaining(130)	5257	27.95%		remaining(130)	3608	23.30%

Table 31 - 'Access' Category Top Web Sites by Overall Instances

'Blogging' Category Top Web Sites by Overall Instances							
	Range 1				Range 2		
	Web Site	# Instances	Percentage		Web Site	# Instances	Percentage
1	feedburner.com	197	34.68%	1	feedburner.com	175	52.55%
2	sixapart.com	99	17.43%	2	xanga.com	48	14.41%
3	livejournal.com	71	12.50%	3	livejournal.com	21	6.31%
4	xanga.com	53	9.33%	4	straightdope.com	19	5.71%
5	boingboing.net	37	6.51%	5	boingboing.net	16	4.80%
6	typepad.com	30	5.28%	6	typepad.com	13	3.90%
7	straightdope.com	22	3.87%	7	waxy.org	12	3.60%
8	wordpress.com	20	3.52%	8	notcot.com	10	3.00%
9	vapid.com	12	2.11%	9	sixapart.com	10	3.00%
10	dooce.com	11	1.94%	10	vox.com	3	0.90%
	remaining(3)	16	2.82%		remaining(3)	6	1.80%

Table 32 - 'Blogging' Category Top Web Sites by Overall Instances

'Education' Category Top Web Sites by Overall Instances

	Range 1				Range 2		
	Web Site	# Instances	Percentage		Web Site	# Instances	Percentage
1	wikimedia.org	1897	33.63%	1	wikimedia.org	1695	31.27%
2	upenn.edu	272	4.82%	2	rae.es	408	7.53%
3	ucla.edu	197	3.49%	3	usc.edu	396	7.30%
4	mit.edu	172	3.05%	4	ucla.edu	287	5.29%
5	smc.edu	164	2.91%	5	mit.edu	125	2.31%
6	lihti.org	133	2.36%	6	leo.org	112	2.07%
7	wordreference.com	128	2.27%	7	wordreference.com	99	1.83%
8	usc.edu	117	2.07%	8	smc.edu	88	1.62%
9	yale.edu	107	1.90%	9	mtu.edu	87	1.60%
10	leo.org	75	1.33%	10	rit.edu	72	1.33%
	remaining(139)	2378	42.16%		remaining(139)	2052	37.85%

Table 33 - 'Education' Category Top Web Sites by Overall Instances

'Government' Category Top Web Sites by Overall Instances							
	Range 1				Range 2		
	Web Site	# Instances	Percentage		Web Site	# Instances	Percentage
1	lapdonline.org	71	23.28%	1	ca.gov	85	36.17%
2	ca.gov	69	22.62%	2	metro.net	49	20.85%
3	ladwp.com	44	14.43%	3	lapdonline.org	43	18.30%
4	metro.net	39	12.79%	4	ladwp.com	18	7.66%
5	lacity.org	15	4.92%	5	state.gov	9	3.83%
6	loc.gov	12	3.93%	6	lacounty.gov	7	2.98%
7	laanimalservices.org	11	3.61%	7	lacity.org	6	2.55%
8	ssa.gov	11	3.61%	8	nist.gov	5	2.13%
9	senate.gov	10	3.28%	9	culvercity.org	4	1.70%
10	lacounty.gov	9	2.95%	10	loc.gov	3	1.28%
	remaining(3)	14	4.59%		remaining(3)	6	2.55%

Table 34 - 'Government' Category Top Web Sites by Overall Instances

'Informational' Category Top Web Sites by Overall Instances							
	Range 1				Range 2		
	Web Site	# Instances	Percentage		Web Site	# Instances	Percentage
1	cnet.com	1098	44.58%	1	cnet.com	855	53.84%
2	apartmenttherapy.com	524	21.27%	2	apartmenttherapy.com	336	21.16%
3	tomshardware.com	115	4.67%	3	about.com	87	5.48%
4	opentable.com	86	3.49%	4	opentable.com	48	3.02%

5	hardocp.com	74	3.00%	5	w3.org	31	1.95%
6	about.com	71	2.88%	6	tomshardware.com	29	1.83%
7	agapelive.com	70	2.84%	7	slashdot.org	25	1.57%
8	citysearch.com	64	2.60%	8	citysearch.com	17	1.07%
9	one.org	41	1.66%	9	flashkit.com	16	1.01%
10	slashdot.org	40	1.62%	10	polishforums.com	13	0.82%
	remaining(22)	280	11.37%		remaining(22)	131	8.25%

Table 35 - 'Informational' Category Top Web Sites by Overall Instances

'Internet Services & Software' Category Top Web Sites by Overall Instances							
	Range 1				Range 2		
	Web Site	# Instances	Percentage		Web Site	# Instances	Percentage
1	microsoft.com	1523	20.36%	1	groove.net	539	11.17%
2	groove.net	668	8.93%	2	microsoft.com	507	10.50%
3	userplane.com	532	7.11%	3	attens.net	427	8.85%
4	attens.net	245	3.28%	4	mozilla.com	247	5.12%
5	mozilla.com	231	3.09%	5	webex.com	210	4.35%
6	sheepserver.net	227	3.03%	6	amazonaws.com	206	4.27%
7	salesforce.com	209	2.79%	7	ovh.net	178	3.69%
8	feedburner.com	197	2.63%	8	feedburner.com	175	3.63%
9	expertcity.com	188	2.51%	9	userplane.com	163	3.38%
10	equals.com	177	2.37%	10	photos.com	161	3.34%
	remaining(110)	3283	43.89%		remaining(110)	2014	41.72%

Table 36 - 'Internet Services & Software' Category Top Web Sites by Overall Instances

'Products' Category Top Web Sites by Overall Instances							
	Range 1				Range 2		
	Web Site	# Instances	Percentage		Web Site	# Instances	Percentage
1	apple.com	324	12.11%	1	apple.com	1548	50.49%
2	virtualearth.net	291	10.88%	2	arttoday.com	429	13.99%
3	blackberry.com	280	10.47%	3	virtualearth.net	197	6.43%
4	arttoday.com	258	9.64%	4	crateandbarrel.com	67	2.19%
5	delias.com	118	4.41%	5	delias.com	66	2.15%
6	t-mobile.com	103	3.85%	6	priberam.pt	61	1.99%
7	internetpostage.com	80	2.99%	7	t-mobile.com	56	1.83%
8	multitran.ru	80	2.99%	8	blackberry.com	40	1.30%
9	123greetings.com	76	2.84%	9	internetpostage.com	40	1.30%

10	applestore.com	67	2.50%	10	dell.com	35	1.14%
	remaining(57)	998	37.31%		remaining(57)	527	17.19%

Table 37 - 'Products' Category Top Web Sites by Overall Instances

'Professional Services' Category Top Web Sites by Overall Instances							
	Range 1				Range 2		
	Web Site	# Instances	Percentage		Web Site	# Instances	Percentage
1	intdm.com	1726	38.47%	1	intdm.com	2639	51.37%
2	nonfatmedia.com	672	14.98%	2	nonfatmedia.com	577	11.23%
3	ftp.sonopress.de	325	7.24%	3	bydeluxe.com	334	6.50%
4	crestdigital.com	250	5.57%	4	monster.com	230	4.48%
5	bydeluxe.com	143	3.19%	5	crestdigital.com	223	4.34%
6	monster.com	142	3.16%	6	quiettouch.com	151	2.94%
7	quiettouch.com	105	2.34%	7	adp.com	144	2.80%
8	proz.com	94	2.09%	8	collective-media.net	93	1.81%
9	comchoice.com	89	1.98%	9	fancorps.com	79	1.54%
10	adp.com	85	1.89%	10	ftp.sonopress.de	54	1.05%
	remaining(47)	856	19.08%		remaining(47)	613	11.93%

Table 38 - 'Professional Services' Category Top Web Sites by Overall Instances

'Search' Category Top Web Sites by Overall Instances							
	Range 1				Range 2		
	Web Site	# Instances	Percentage		Web Site	# Instances	Percentage
1	google.com	35475	55.17%	1	google.com	24819	51.32%
2	yahoo.com	25110	39.05%	2	yahoo.com	20111	41.59%
3	live.com	2697	4.19%	3	live.com	2256	4.67%
4	start.com	390	0.61%	4	start.com	757	1.57%
5	answers.com	265	0.41%	5	answers.com	100	0.21%
6	atlanticbb.net	102	0.16%	6	ask.com	84	0.17%
7	ask.com	81	0.13%	7	usa.dict.cn	32	0.07%
8	go.com	42	0.07%	8	findabeautysalon.com	27	0.06%
9	adsonar.com	41	0.06%	9	atlanticbb.net	23	0.05%
10	atomz.com	17	0.03%	10	reachlocal.com	22	0.05%
	remaining(13)	79	0.12%		remaining(13)	126	0.26%

Table 39 - 'Search' Category Top Web Sites by Overall Instances

'Travel & Transportation' Category Top Web Sites by Overall Instances			
	Range 1		Range 2

	Web Site	# Instances	Percentage		Web Site	# Instances	Percentage
1	virtualearth.net	291	20.46%	1	yelp.com	316	19.49%
2	southwest.com	132	9.28%	2	virtualearth.net	197	12.15%
3	aa.com	115	8.09%	3	southwest.com	101	6.23%
4	delta.com	108	7.59%	4	virginamerica.com	94	5.80%
5	virginamerica.com	70	4.92%	5	aa.com	79	4.87%
6	citysearch.com	64	4.50%	6	priceline.com	68	4.19%
7	thumpertalk.com	58	4.08%	7	travelocity.com	61	3.76%
8	travelocity.com	47	3.31%	8	virtuallythere.com	60	3.70%
9	hotwire.com	39	2.74%	9	united.com	50	3.08%
10	priceline.com	39	2.74%	10	starwoodhotels.com	44	2.71%
	remaining(44)	459	32.28%		remaining(44)	551	33.99%

Table 40 - 'Travel & Transportation' Category Top Web Sites by Overall Instances

'Web Portal & Email' Category Top Web Sites by Overall Instances							
	Range 1				Range 2		
	Web Site	# Instances	Percentage		Web Site	# Instances	Percentage
1	webaccess.umail.ucsb.edu	8560	25.50%	1	google.com (email)	3852	23.27%
2	google.com (email)	5195	15.47%	2	yahoo.com (email)	2770	16.73%
3	yahoo.com (email)	4016	11.96%	3	mailcenter.comcast.net	1805	10.90%
4	aol.com	3947	11.76%	4	mac.com	1562	9.44%
5	mailcenter.comcast.net	2354	7.01%	5	aol.com	1282	7.74%
6	msn.com	924	2.75%	6	aol.com (email)	544	3.29%
7	sbcglobal.net	889	2.65%	7	sbcglobal.net	504	3.04%
8	ip.fastwebnet.it	730	2.17%	8	hotmail.com	482	2.91%
9	mac.com	707	2.11%	9	optonline.net	395	2.39%
10	aol.com (email)	604	1.80%	10	ip.fastwebnet.it	345	2.08%
	remaining(87)	5645	16.82%		remaining(87)	3013	18.20%

Table 41 - 'Web Portal & Email' Category Top Web Sites by Overall Instances

Unique User Statistics for Top Web Sites for Range 2			
Web Site	Main Category	# Instances	# of Unique Users
google.com	Search	24819	230
yahoo.com	Search	20111	213

youtube.com	Entertainment	8509	167
charter.com	Access	4044	63
google.com (email)	Web Portal & Email	3852	188
meebo.com	Social Networking	3555	13
comcast.net	Access	3523	67
yahoo.com (email)	Web Portal & Email	2770	98
intdm.com	Professional Services	2639	1
live.com	Search	2256	71
imeem.com	Entertainment	2188	47
imeem.com	Social Networking	2188	47
amazon.com	Commerce	2186	146
warnerbros.com	Entertainment	2117	14
facebook.com	Social Networking	1879	98
mailcenter.comcast.net	Web Portal & Email	1805	4
wikimedia.org	Education	1695	152
rr.com	Access	1686	76
mac.com	Web Portal & Email	1562	31
apple.com	Products	1548	55
sony.com	Entertainment	1523	25
aol.com	Web Portal & Email	1282	115
cnn.com	News	1205	58
ebay.com	Commerce	858	108
cnet.com	Informational	855	94
vzavenue.net	Access	782	31
start.com	Search	757	28
live365.com	Entertainment	586	5
nonfatmedia.com	Professional Services	577	8
aol.com (email)	Web Portal & Email	544	14
cox.net	Access	541	48
groove.net	Internet Services & Software	539	20
microsoft.com	Internet Services & Software	507	78
craigslist.org	Commerce	506	63
sbcglobal.net	Web Portal & Email	504	46
hotmail.com	Web Portal & Email	482	59
bankofamerica.com	Financial	442	43
arttoday.com	Products	429	6
attens.net	Internet Services & Software	427	19
rae.es	Education	408	6

Table 42 - Unique User Statistics for Top Web Sites for Range 2

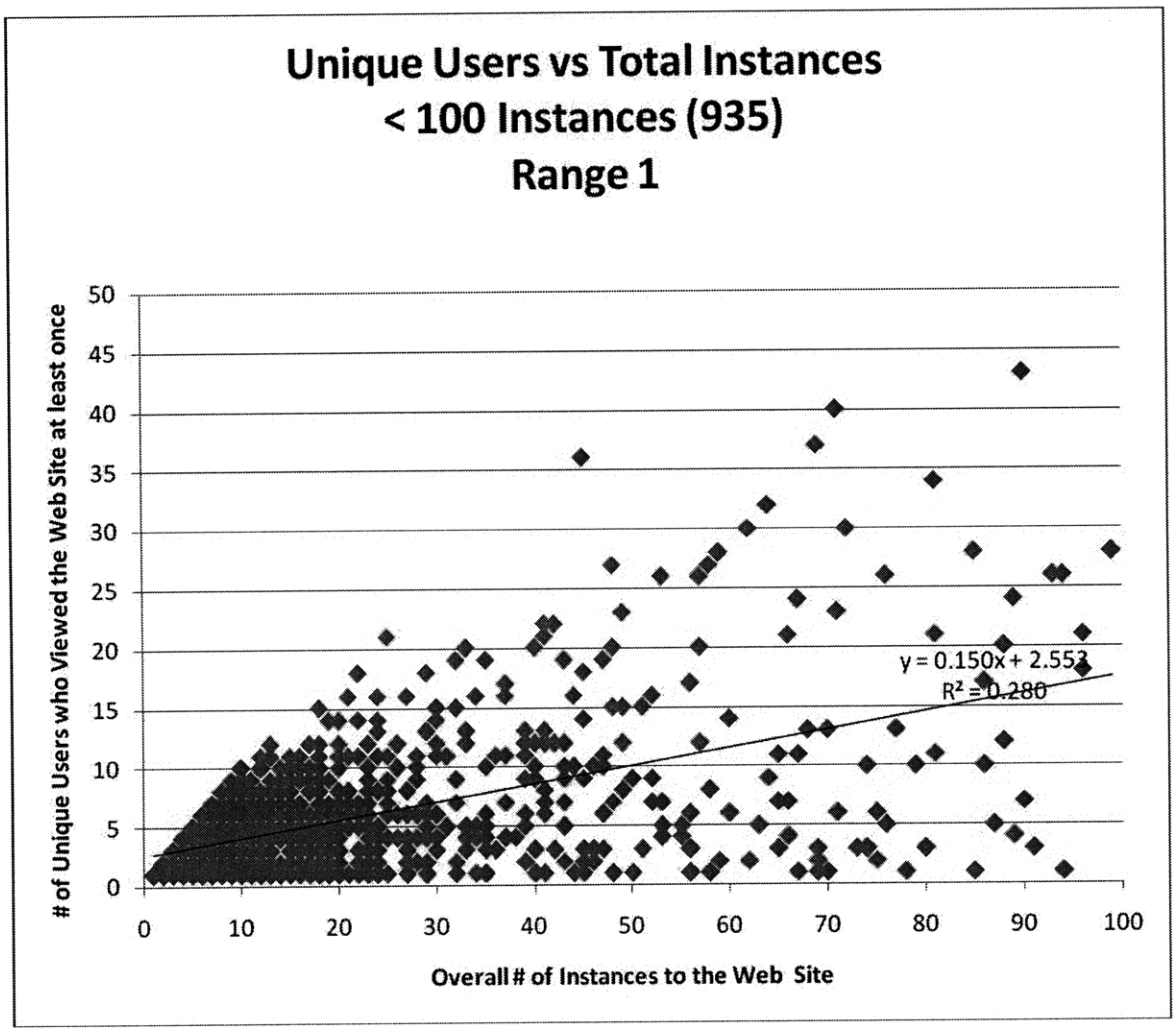

Figure 100 - Unique Users vs Total Instances. <100 Instances (935). Range 1

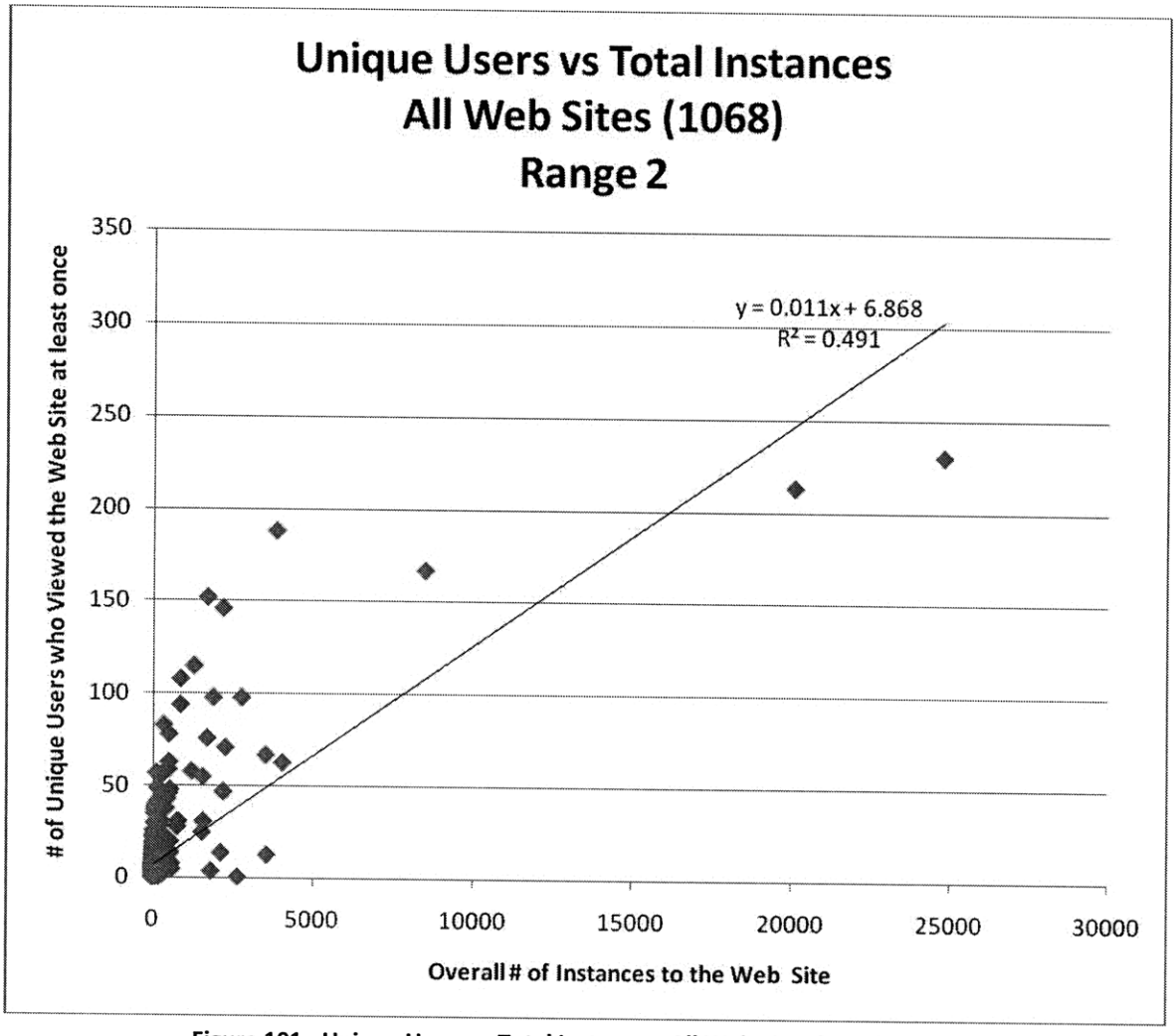

Figure 101 - Unique Users vs Total Instances. All Web Sites (1068). Range 2

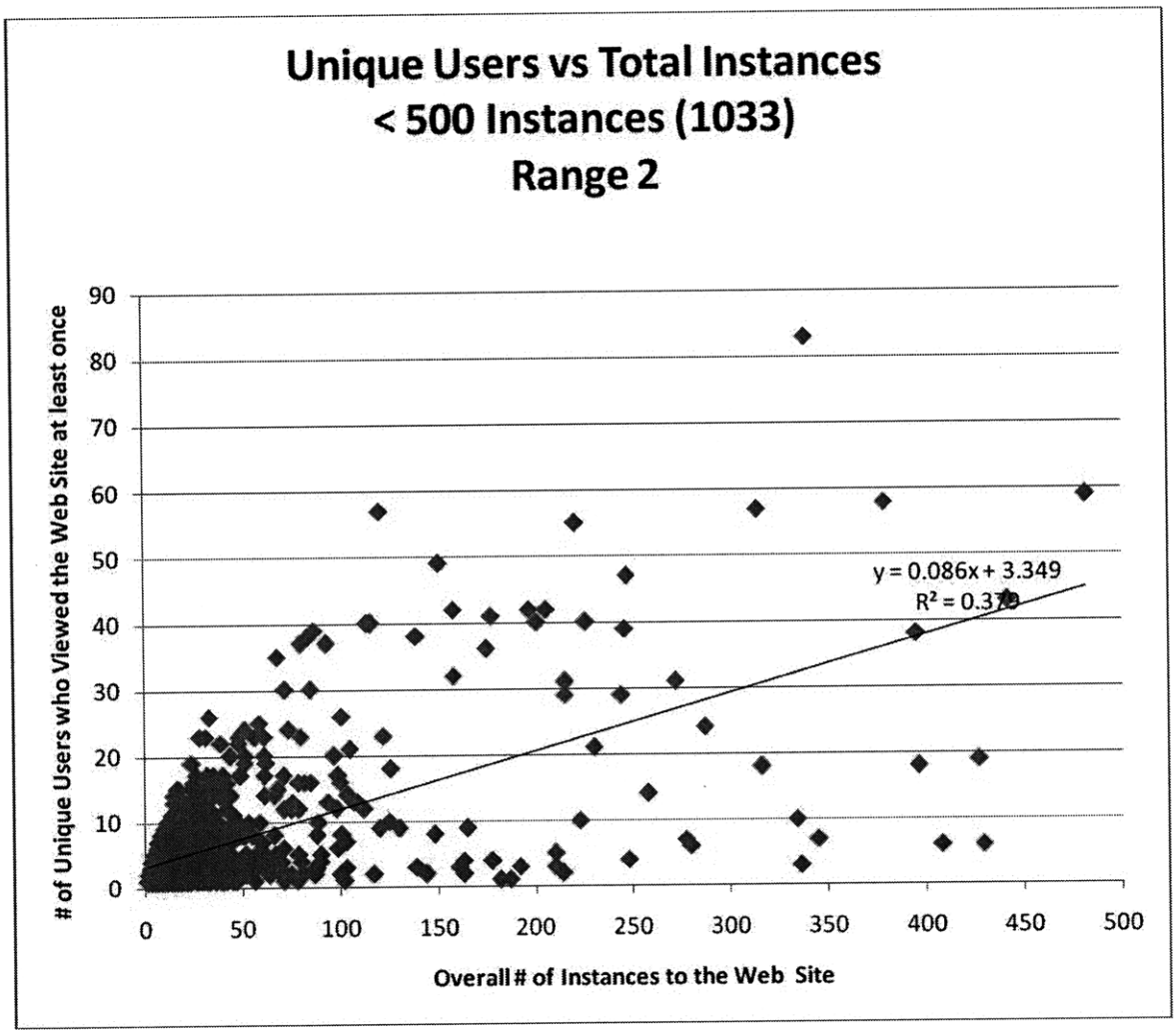

Figure 102 - Unique Users vs Total Instances. <500 Instances (1033). Range 2

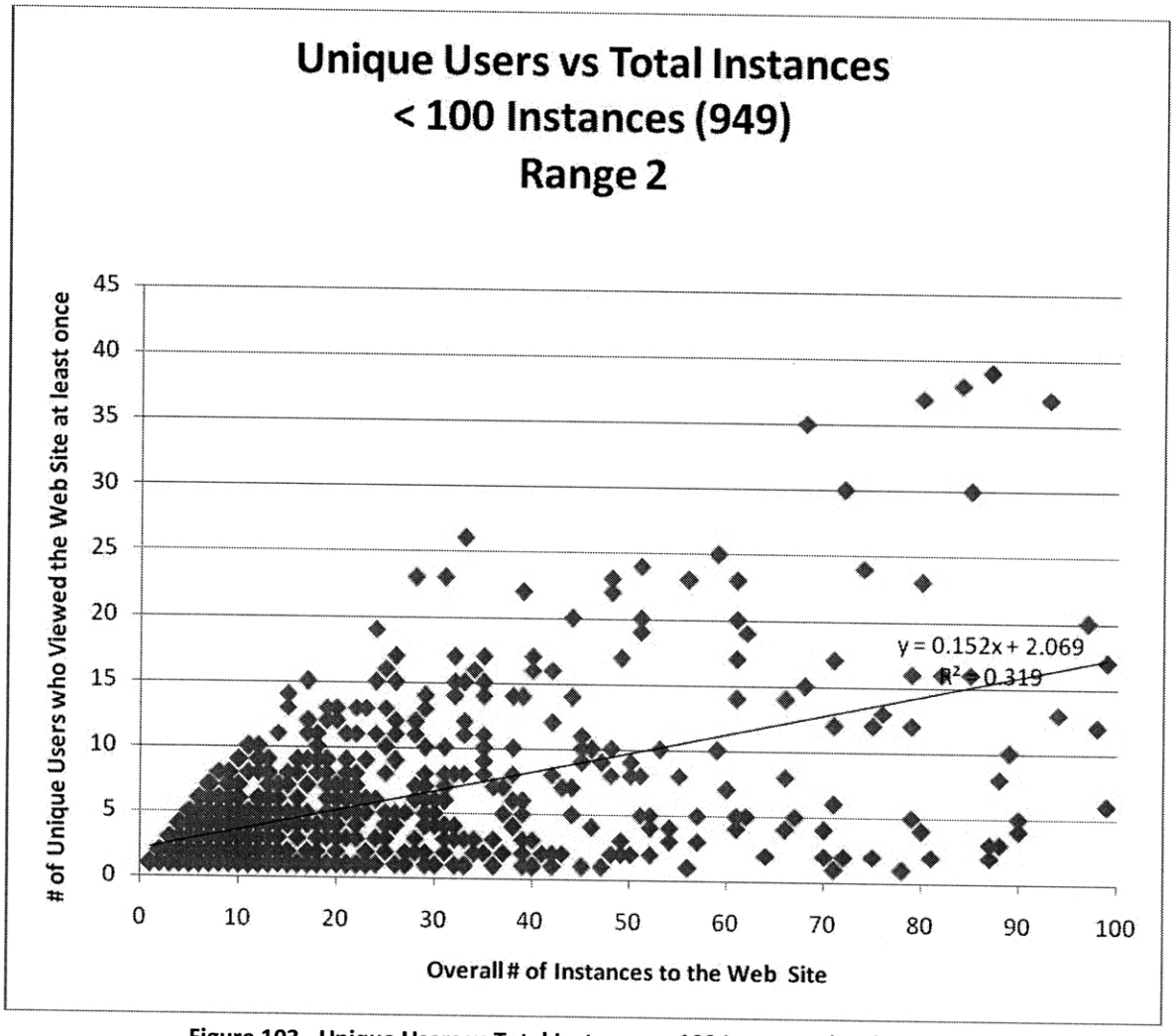

Figure 103 - Unique Users vs Total Instances. <100 Instances (949). Range 2

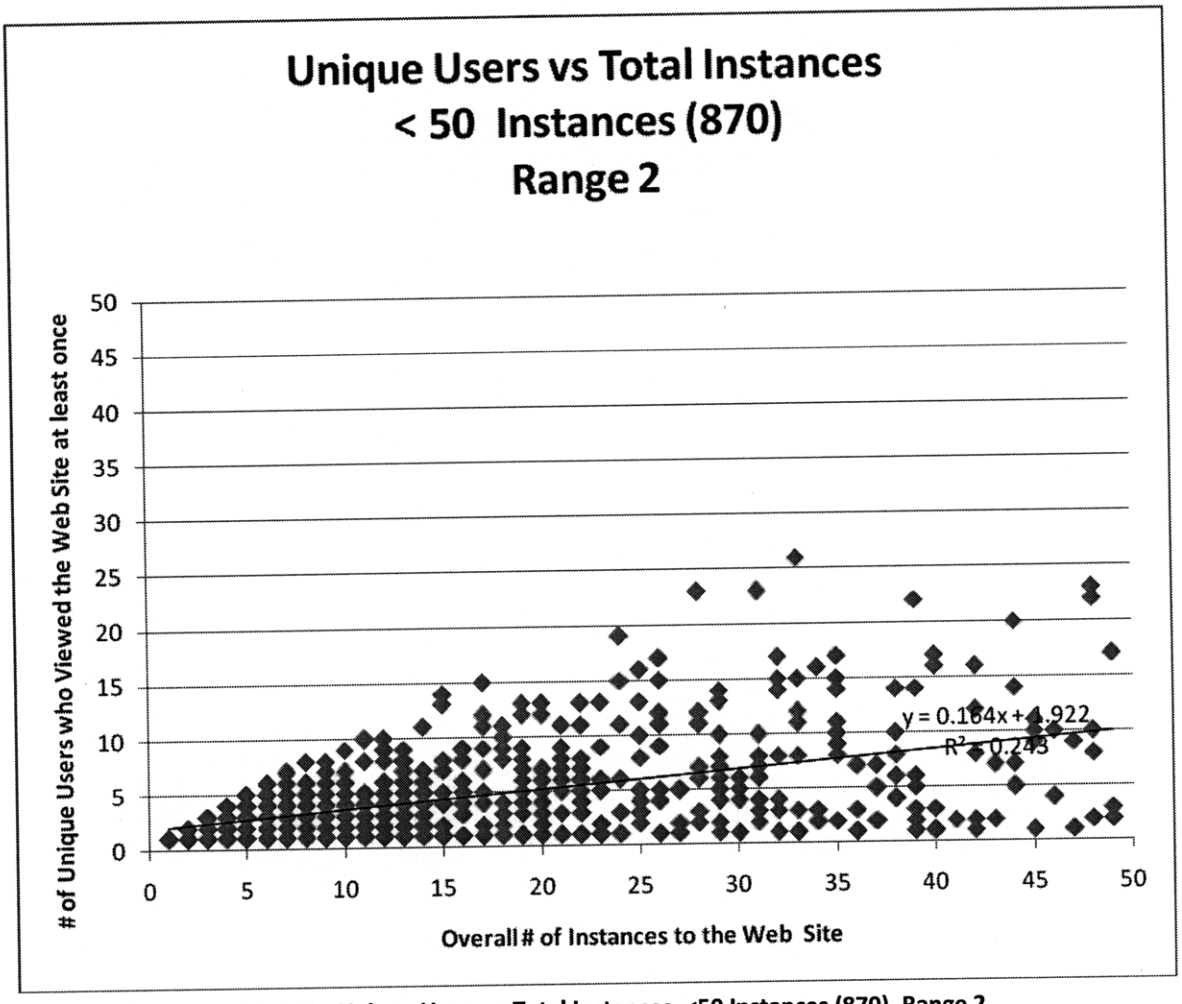

Figure 104 - Unique Users vs Total Instances. <50 Instances (870). Range 2

Appendix E: Page Rank

Average Page Rank of Visited Web Sites by Sub-Category				
Main Category	**Sub-Category**	**Range 1**	**Range 2**	**Range 1+2**
Access	Access	5.27	5.64	5.44
Blogging	Blogging Services	7.93	7.84	7.90
Blogging	Design	6.00	6.00	6.00
Blogging	Humor	7.00	7.00	7.00
Blogging	Person	6.53	7.00	6.72
Education	Colleges	8.21	8.28	8.25
Education	Financial	6.06	6.55	6.32
Education	Foreign	7.17	4.88	6.58
Education	Government	9.00	9.00	9.00
Education	Learning	6.05	5.00	6.00
Education	Other	6.23	6.29	6.24
Education	Pre-College	0.00	0.00	0.00
Education	Reference	6.05	6.13	6.09
Government	Municipal	6.28	6.43	6.34
Government	National	8.68	8.48	8.61
Government	State	8.00	8.00	8.00
Internet Services & Software	Arts	6.08	6.92	6.40
Internet Services & Software	Communication	6.45	6.44	6.44
Internet Services & Software	Data	7.34	7.30	7.32
Internet Services & Software	Foreign	3.16	4.62	3.78
Internet Services & Software	Marketing	5.70	6.48	6.02
Internet Services & Software	Media	1.00	n/a	1.00
Internet Services & Software	Other	6.34	6.80	6.59
Internet Services & Software	Reference	0.05	0.66	0.32
Internet Services & Software	Security	8.11	6.76	7.60
Internet Services & Software	Software	7.18	6.73	7.03
Internet Services & Software	Web Design	4.56	4.02	4.26
Professional Services	Arts	n/a	3.00	3.00
Professional Services	Business	4.79	4.98	4.90
Professional Services	Design	3.97	4.00	3.98
Professional Services	Entertainment	4.00	4.00	4.00
Professional Services	Foreign	0.65	1.42	0.80
Professional Services	Housing	n/a	6.00	6.00
Professional Services	Jobs	7.89	7.86	7.87
Professional Services	Marketing	5.49	4.95	5.21
Professional Services	Media	4.21	4.48	4.35
Professional Services	Other	4.52	4.71	4.60

Professional Services	Translation	5.95	5.23	5.63
Search	General	9.49	9.45	9.47
Search	Other	0.00	1.94	1.84
Search	People	4.85	5.66	5.16
Travel & Transportation	Bus	8.00	8.00	8.00
Travel & Transportation	Cars	5.58	4.50	4.89
Travel & Transportation	Flying	7.17	7.02	7.11
Travel & Transportation	Foreign	0.00	0.00	0.00
Travel & Transportation	Guides	6.58	6.84	6.78
Travel & Transportation	Hotel	6.09	6.22	6.15
Travel & Transportation	Municipal	7.00	7.00	7.00
Travel & Transportation	Other	5.91	6.00	5.95
Travel & Transportation	Reservations	7.48	7.33	7.39
Web Portal & Email	Web Portal & Email	7.67	7.76	7.70

Table 43 - Average Page Rank of Visited Web Sites by Sub-Category

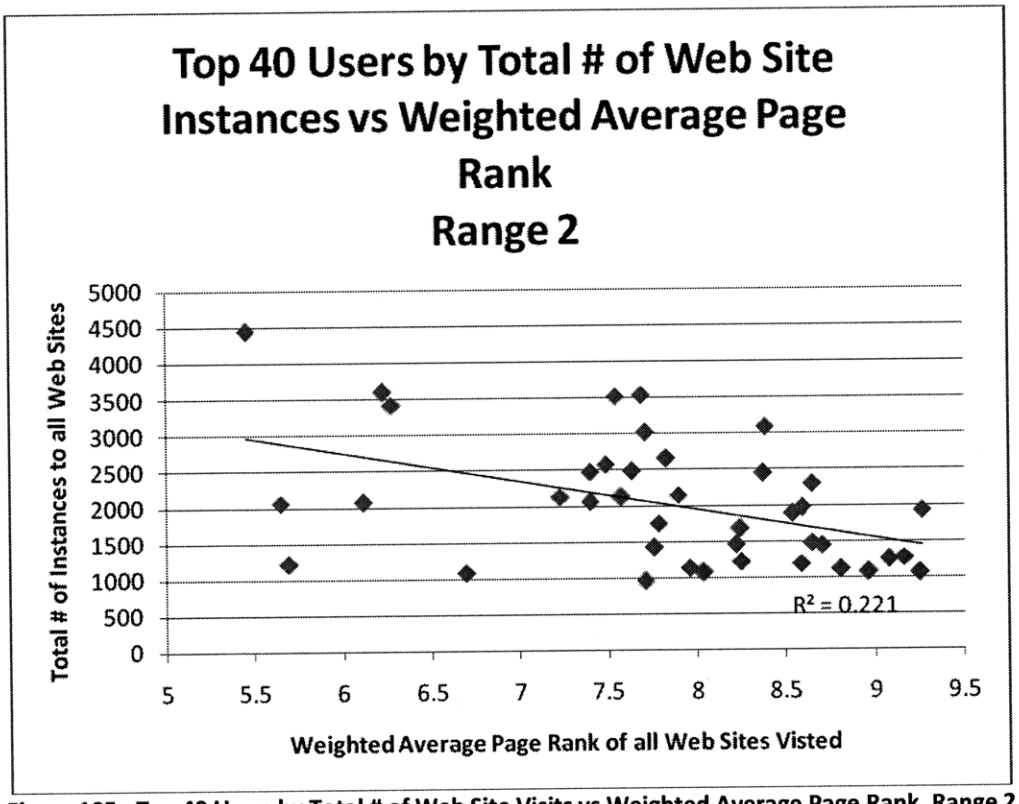

Figure 105 - Top 40 Users by Total # of Web Site Visits vs Weighted Average Page Rank. Range 2

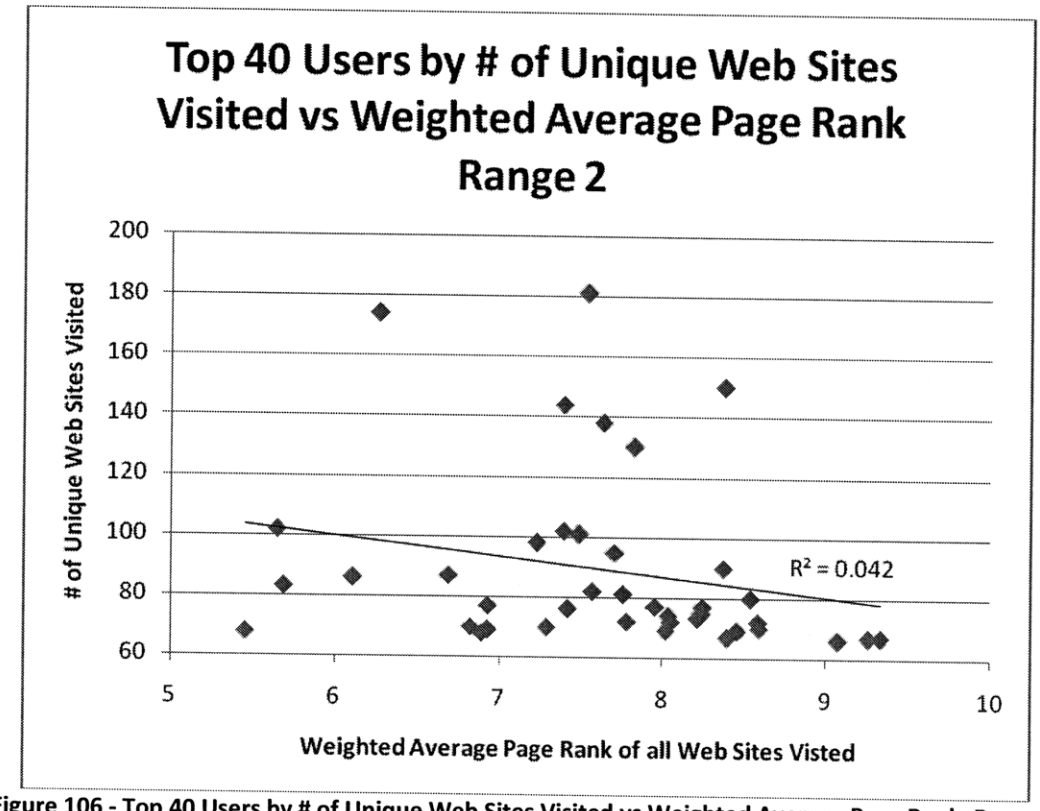

Figure 106 - Top 40 Users by # of Unique Web Sites Visited vs Weighted Average Page Rank. Range 2

Appendix F: Diversity

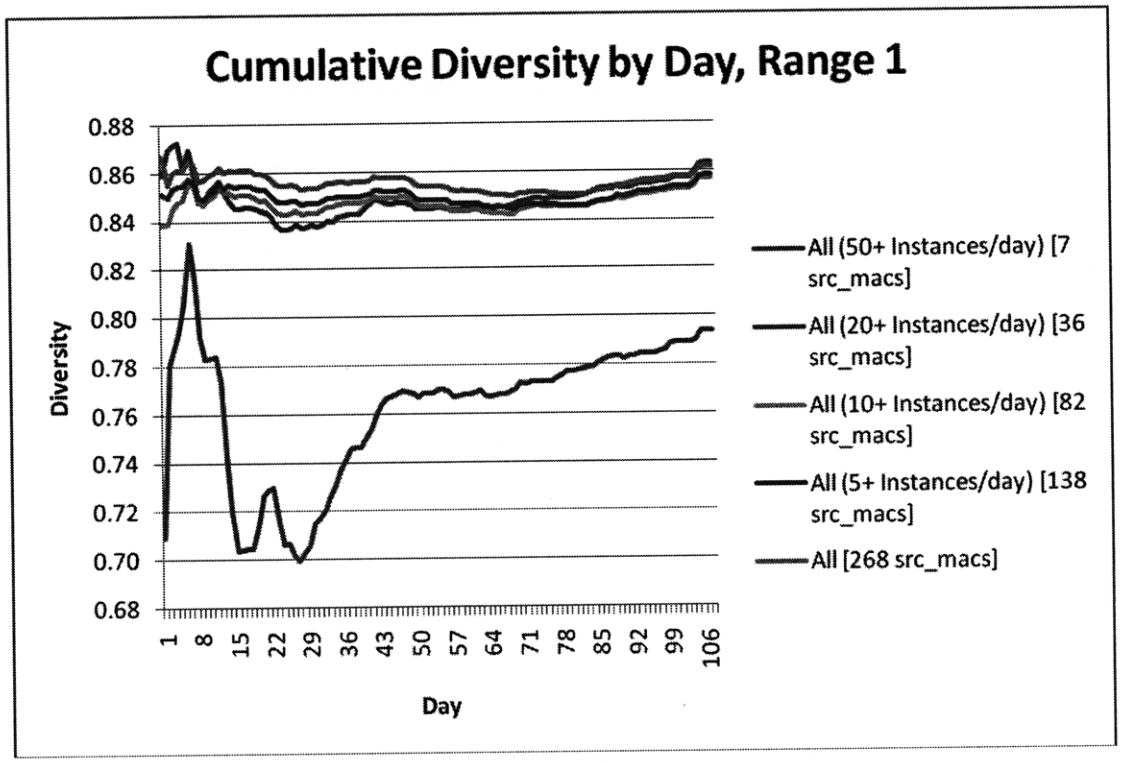

Figure 107 - Cumulative Diversity by Day, Range 1

References

Anderson, C. (2004). The Long Tail. http://www.wired.com/wired/archive/12.10/tail.html

Apte, U.M., & Nath, H.K. (2004). Size, Structure and Growth of the U.S. Informational Economy. In *Managing in the Information Economy (pp.1-28). Springer US.*

Aral, S., Brynjolfsson, E., Van Alstyne, M. (2006). Information, Technology and Information Worker Productivity.

Aral, S., Van Alstyne, M. (2008). Networks, Information & Social Capital.

Choe, T. (2006, August). Identifying Word Categories for Diffusion Studies in an Email Social Network. *Master of Engineering Thesis.* Cambridge, MA: Massachusetts Institute of Technology.

Choo, C.W., Turnbull, D. (1998). A Behavioral Model of Information Seeking on the Web -- Preliminary Results of a Study of How Managers and IT Specialists Use the Web.

Diversity Indices. *Wikipedia*. http://en.wikipedia.org/wiki/Diversity_indices

Facca, F. M., Lanzi, P.L. (2003). Mining Interesting Knowledge From Weblogs: A Survey.

Farrokhzadi, M. (2007, May). Entropy, Information Rate and Mutual Information Measures for the Email Content of Information Workers. *Master of Engineering Thesis.* Cambridge, MA: Massachusetts Institute of Technology.

Ho, S. Y. (2005). An Exploratory Study of Using a User Remote Tracker to Examine Web Users' Personality Traits.

Manoharn, P. (2006, August). Diversity Measurement for the Email Content of Information Workers. *Master of Engineering Thesis*. Cambridge, MA: Massachusetts Institute of Technology.

Phaal, P., Lavine, M. (2004). sFlow Version 5. http://www.sflow.org/sflow_version_5.txt

Qian, J. (2008, May). Structure and Evolution of Communication Networks in Organizations. *Masters of Engineering Thesis.* Cambridge, MA: Massachusetts Institute of Technology.

Rogers, I. (2002). The Google Pagerank Algorithm and How It Works. http://www.ianrogers.net/google-page-rank

Salton, G., Wong, A., Yang, C.S. (1975). A Vector Space Model for Automatic Indexing. *Association for Computing Machinery Inc.*

Spiliopoulou, M. (2000). Web Usage Mining for Web Site Evaluation. *Communications of the ACM* , 127-134.

Spiliopoulou, M., Pohle, C. (2001). Data Mining for Measuring and Improving the Success of Web Sites. Kluwer Academi Publishers.

Srivastava, J., Cooley, R., Deshpande, M., Tan, P. (2000). Web Usage Mining: Discovery and Application of Usage Patterns from Web Data.

Weinman, L. (2007). A New Approach To Search. *Business Communications Review.*

Wu, L., Waber, B., Aral, S., Brynjolfsson, E., Pentland, A. (2008). Mining Face-To-Face Interaction Networks Using Sociometric Badges: Predicting Productivity in an IT Configuration Task.