

AN ABSTRACT OF THE THESIS OF

Prashanth Ayyavu for the degree of Master of Science in Computer Science presented on September 13, 2011

Title: Adding Context and Value to Online Security Ratings by Combining Heuristics with Community Based Ratings.

Abstract approved:

Carlos Jensen

The Internet is growing rapidly in terms of websites, users and uses. People use the Internet for reference, shopping, social networking, communications, business and much more. Though the Internet is useful, there are many risks associated with its use, like malicious websites, identity theft, hateful content and fraudulent practices. Online safe surfing tools help Internet users stay safe from potential threats and have become common. These tools give security ratings for websites or automatically block and filter content. Many users rely on these ratings to identify dangers. This thesis examines different types of online safe surfing tools, analyzes their relative strengths and weaknesses, and ways to improve their quality. This thesis consists of two manuscripts. First, we conducted a study of the security ratings given by these tools. Based on these results, we identified two families of tools; 1) heuristic machine based and 2) community based tools. These differ both in analysis methodology and focus. Heuristics tools advise users about the technical foundations of a website but cannot be used to find whether the website engages in dubious practices, whether the content is for adults only, or whether the content is credible or reliable. Community based tools use user ratings, but determining the reliability of these is a problem. We proposed a method for extracting high quality information from user data that can be added with heuristic techniques. Second, we conducted a controlled lab experiment

and studied whether combining user data with machine-based results added value to users. We also analyzed how users behave when presented with conflicting ratings for a website. Our results show that combining heuristics and community-based information increased the confidence level of users while rating websites.

©Copyright by Prashanth Ayyavu
September 13 2011
All Rights Reserved

Adding Context and Value to Online Security Ratings by Combining Heuristics with
Community Based Ratings.

by
Prashanth Ayyavu

A THESIS

submitted to

Oregon State University

in partial fulfillment of
the requirements for the
degree of

Master of Science

Presented September 13, 2011
Commencement June 2012

Master of Science thesis of Prashanth Ayyavu presented on September 13, 2011

APPROVED:

Major Professor, representing Computer Science

Director of the School of Electrical Engineering and Computer Science

Dean of the Graduate School

I understand that my thesis will become part of the permanent collection of Oregon State University libraries. My signature below authorizes release of my thesis to any reader upon request.

Prashanth Ayyavu, Author

ACKNOWLEDGEMENTS

I would like to thank my Major Professor, Dr. Carlos Jensen for all his guidance and support during this thesis work.

I would also like to thank my parents Mr. K. Ayyavu and Gokila Mani Ayyavu, and my brother Praveen Kumar Ayyavu, for their endless support throughout my study. I feel lucky to have such an amazing and supportive family behind me. Amma, without your love and encouragement I would not have achieved this.

I appreciate and thank Dr. Christopher Scaffidi, who was my academic advisor during Fall 2009 and Winter 2010.

I thank all my friends at OSU and HCI research group members for their support throughout my graduate school life.

DEDICATION

I dedicate this thesis work to my Amma, who never failed to encourage me during the tough times of this thesis project. I wouldn't have been able to finish this project without you.

TABLE OF CONTENTS

	<u>Page</u>
1 Introduction.....	1
2 First manuscript: Integrating User Feedback with Heuristic Security and Privacy Management Systems	3
2.1 Abstract.....	4
2.2 Introduction.....	4
2.3 Related work	7
2.4 Methodology	10
2.5 Results.....	15
2.6 Discussion	23
2.7 Conclusion	26
2.8 Acknowledgements.....	27
2.9 References.....	27
3 Second manuscript: A User Evaluation of Heuristic and Community Based Safe Surfing Ratings: Which Do Users Trust More?	31
3.1 Abstract.....	32
3.2 Introduction.....	32
3.3 Related work	36
3.4 Methodology	39
3.4.1 Setup	39
3.4.2 Prototype system.....	41
3.4.3 Selection of websites	43
3.4.4 Experimental conditions	45
3.4.5 Questionnaire	45
3.4.6 Post-experiment survey	46
3.4.7 Demographics	47
3.5 Results.....	47

TABLE OF CONTENTS (Continued)

	<u>Page</u>
3.5.1 RQ1: Are users more heavily influenced by machine ratings, or user ratings? (Regardless of which is more correct)	47
3.5.2 RQ2: Does combining the two types of ratings add value to users?	49
3.5.3 RQ 3: When presented with conflicting ratings, how do users act?	50
3.5.4 Survey results.....	51
3.6 Discussion	52
3.7 Conclusion	54
3.8 Acknowledgments	54
3.9 References.....	54
4 Conclusion	57
5 Bibliography	58

LIST OF FIGURES

<u>Figure</u>	<u>Page</u>
2.1 At-a-glance security & privacy indicators	11
2.1 McAfee SiteAdvisor site review page	11
2.3 McAfee SiteAdvisor decision-making model.....	12
2.4 Web of Trust site review page	13
2.5 Construction of the ‘Good’ dictionary	20
2.6 Construction of the ‘Bad’ dictionary	20
2.7 % of adult sites in the three samples	21
2.8 Classification of websites in Set B using dictionaries	23
3.1 Various website reputation services and their colored-rating icons.....	34
3.2 Design of study application.	40
3.3 WOT rating icon near the website URL.	41
3.4 a) WOT rating card. b) MSA rating card.	41
3.5 Prototype scorecard.....	42
3.6 Questionnaire used in the experiment.....	46
3.7 Users average confidence in rating the trustworthiness of good and bad sites.....	49
3.8 Difference in confidence levels of participants from different groups	50
3.9 Increase in confidence level of participants while rating trustworthiness of conflicting websites.....	51

LIST OF TABLES

<u>Figure</u>	<u>Page</u>
2.1 System discrimination.....	15
2.2 Degree of mismatch between WOT and MSA	15
2.3 Stability of site ratings	17
2.4 Site reviews	18
3.1 Color of star icon for a website, depends on the type of rating	42
3.2 Classification of websites based on ratings given by MSA and WOT.	44

1. INTRODUCTION

The Internet is growing rapidly and has established itself as a powerful platform for delivering large amounts of information and services. People use the Internet for reference, shopping, social networking, communications, business and much more. Though the Internet is useful, there are many risks associated with its use, like malicious websites, identity theft, hateful content and fraudulent practices. Since the Internet has become a mixture of good and bad websites, and telling these two apart can at times be difficult, users need help to differentiate between them (Dourish et al. 2004). Many users lack the confidence, skill, or time to make careful security decisions on their own (Dourish et al. 2004) (Gross & Rosson 2007), and these tools fill an important niche. Online safe surfing tools have become common and provide different kinds of services aimed at helping Internet users stay safe from potential threats. These tools often give colored-iconic security ratings for websites or automatically block and filter content.

There are two major groups of safe surfing tools: Heuristic-based tools, and community-based or end-user rating tools. Heuristic tools in essence scan websites for previously identified exploits, techniques or technologies that experts have identified as potentially problematic for users privacy or security. Community-based tools are those where the user community in essence scores and reviews websites based on their personal experience, giving other users advice on whether a website is safe or not. Both these families have their own merits and demerits. (Angai et al. 2010) found that different heuristic-based safe browsing services implement and maintain their own malware detection system and algorithm, resulting in major discrepancies between different services. Furthermore, heuristics-based tools cannot be used to check whether a website follows all said privacy policies, whether a site involves in fraudulent practices or not. One of the main problems associated with community-based tools is determining the reliability of user ratings. The problem is that most users are not terribly good at evaluating security or privacy risks. Some users are more

thorough and knowledgeable and can be trusted to do a good job of analyzing sites while others cannot.

This thesis examines different types of online safe surfing tools, analyzes their relative strengths and weaknesses, and ways to improve their quality. To do this we had the following research questions.

- How much conflict is there between heuristic and community-based ratings?
- How stable are community-based ratings when compared to heuristic-based ratings?
- When disagreements occur, are these due to different rating criteria, or to disagreements and inaccuracies in applying these criteria?
- Are users more heavily influenced by machine ratings, or user ratings?
(Regardless of which is more correct)
- Does combining the two types of ratings add value to users?
- When presented with conflicting ratings, how do users act?

This thesis manuscript consists of two papers. In the first paper, we conducted a study of the security ratings given by these tools for about 20,000 websites. Heuristic and community tools differ both in analysis methodology and focus. There was a high level of conflict between these tools in rating a website as “bad”, ”risky”. This makes this work significant, because protecting users from dangerous websites is always a paramount necessity. We proposed a method for constructing dictionaries by extracting high quality information from user data. These dictionaries can add nuance to machine based scan results. In our second manuscript, we conducted a controlled lab experiment and studied whether combining user data with machine based results added value to users. We also analyzed how users behave when presented with conflicting ratings for a website. Our results show that users confidence level in dealing with the trustworthiness of websites increases when presented with combined data.

2 First manuscript: Integrating User Feedback with Heuristic Security and Privacy Management Systems

Prashanth Ayyavu, Carlos Jensen

School of EECS
Oregon State University
Corvallis, Oregon, 97331, USA
{ayyavu, cjensen}@eecs.oregonstate.edu

Proceedings of the 2011 annual conference on Human factors in computing systems
Vancouver, BC, CANADA
SESSION: Security

2.1 ABSTRACT

Tools aimed at helping users safely navigate the web and safeguard themselves against potential online predators have become reasonably common. Currently there are two families of tools; heuristics analysis tools that test websites directly using automated scripts and programs, and community based tools where users rate websites and write reviews for the benefit of others. In this paper we examine the relative strengths and weaknesses of each technique, whether these techniques are compatible, and how community feedback can be combined with heuristic-based evaluations. In order to do this we conduct a large-scale comparison of the ratings of heuristic and community based tools, and explore novel methods for abstracting key information from user comments, which could be used to add context and nuance to heuristic based ratings. We find that heuristic and community based ratings are highly complementary, and can be combined to potentially guide users to make more informed decisions.

2.2 INTRODUCTION

Safe web-surfing tools, designed to help users identify potentially dangerous and malicious websites, are becoming commonplace. Today, most web-browsers and search engines incorporate some safe surfing features, but standalone or plug-in systems are still the norm. Some well-known safe surfing tools are Web of Trust aka. WOT (<http://www.mywot.com>), Alexa (<http://www.alexa.com>), McAfee's SiteAdvisor (<http://www.siteadvisor.com>) and Norton's SafeWeb (<http://us.norton.com/index.jsp>). Perhaps for this reason, tools such as these are gaining market share, as people look for help in making decisions with respect to online privacy and security management (Dourish et al. 2004). Many users lack the confidence, skill, or time to make careful security decisions on their own due to either a lack of interest and/or understanding of the technical details of web security (Dourish et al. 2004) (Gross & Rosson 2007).

There are two major types of safe surfing tools: Heuristic analysis based tools, and community based or end-user rating based tools. Heuristic analysis tools in essence scan websites for previously identified exploits, techniques or technologies that

experts have identified as potentially problematic for users privacy or security. Most antivirus or anti-spam software follows this strategy, though anti-spam tools often include some element of machine learning to allow them to adapt to changing trends. Community based tools are those where the community in essence scores and reviews websites, giving other users advice on whether a website is safe or not.

One shortcoming of heuristic analysis tools is that they can only advise users about the technical foundations and behavior of a website. For instance: Is the web server patched for the latest known vulnerabilities? Is the website vulnerable to known types of attacks? Are downloads infected with viruses or other malware? Does the website engage in drive-by downloads? Does the website associate with those who do? Etc. These analysis techniques, though very sophisticated and useful, do have some inherent shortcomings. Heuristics cannot be used to determine whether a website's plain text privacy policy is desirable or even lawful (one could evaluate a P3P policy though), whether the company follows said policies or engages in other dubious practices not directly linked to their website (such as unwanted telemarketing or credit card fraud). It is also difficult to accurately classify the content of websites, such as whether the content is for adults only, or whether the content is credible or reliable.

Using word of mouth, or user ratings and feedback to determine whether to visit a website is not a novel concept, and a great number of rating sites and systems exist today. These services depend on building a community of reviewers with both a broad set of interests (to ensure broad coverage) of sites, as well as good knowledge of security and privacy issues (to ensure that reviews are accurate and informative). In addition, the rating site must be engineered in such a way as to promote and encourage insightful reviews rather than allowing the site reviews to revert toward the mean.

One of the chief problems associated with this approach is determining the reliability of user ratings. The problem is that most users are not terribly good at evaluating security or privacy risks, and are quite likely to be biased by factors as a sites' visual design or name recognition. Some users are more thorough and knowledgeable and

can be trusted to do a good job of analyzing sites while others cannot. The challenge is identifying which is which, and ensuring that the good reviewers carry more weight than those of the uninformed or biased. Previous research has looked into some of the problems associated with these systems, and how to increase the reliability of the user ratings for such systems (Goecks, Edwards & Mynatt 2009).

A problem associated with both approaches is generating ratings or recommendations that are simple for end-user to understand, and which are at the same time convincing. Sites that rely on numeric ratings are easier to interpret (a low score is less desirable than a high-scoring site for instance), but how much worse is a 4 from a 5 on a 10-point scale? If categories are used, how much risk is associated with visiting a “risky” website, and what am I at risk from? These are issues that heuristic systems often have difficulty addressing. On the other hand, community based systems often suffer from too much data (e.g. hundreds of paragraph long user reviews), or mismatched scales (e.g. is everyone rating on the same scale?) This is a well-known problem for sites like Amazon, and can be corrected for if reviewers score enough items. From our experience, we know that many end-users do not rate a large number of websites.

In this paper we seek to determine the reliability and compatibility of these two techniques: heuristic evaluation and community based reviews. In other words, we seek to determine:

RQ1: How much conflict is there between heuristic and community-based ratings?

RQ2: How stable are community-based ratings when compared to heuristic-based ratings?

RQ3: When disagreements occur, are these due to different rating criteria, or to disagreements and inaccuracies in applying these criteria?

To try and answer these questions, we chose to focus on the ratings and the techniques employed by McAfee’s SiteAdvisor, one of the leading heuristics-based tools, and

Web of Trust, a leading community-based rating system. We chose these because they are quite popular, and because they are at the cutting edge of their respective approaches.

In this paper, we also explore an approach for extracting high-value community data based around identifying key words from reviews, and organizing these into dictionaries. These dictionaries may then be used to extract meaning and context for community-based ratings.

The rest of this paper is organized as follows: First we give a brief overview of the most relevant literature on both community and heuristic based rating techniques and systems. We then give a description of both McAfee SiteAdvisor and Web of Trust, our methodology and data collection practices, followed by our results, and finally our interpretation of these results, conclusions and future work.

2.3 RELATED WORK

Heuristic techniques use filters to find and respond to known risks and potentially risky behaviors. In this case, heuristics refer to simple rules and patterns. This may take the form of simple keyword searches, or much more sophisticated scans of software signatures, resource use, or memory footprint. Many online tools, including anti-virus (McAfee, <http://www.mcafee.com/us>) (Norton, <http://us.norton.com/index.jsp>), anti-spam (SiteAdvisor, <http://www.siteadvisor.com/>) (SpamKiller, <http://us.mcafee.com/root/product.asp?productid=msk>), anti-phishing (Mozilla Firefox anti-phishing, <http://www.mozilla.com/en-US/firefox/phishing-protection/>) (Untangle, <http://www.untangle.com/>), anti-spyware (Avira, <http://www.avira.com/en/pages/index.php>) (SpyBot, <http://www.safer-networking.org/en/index.html>), popup-blockers (Webroot popup washer, http://www.webroot.com/En_US/consumer-products-spysweeper.html) (AOLexplorer, <http://www.aol.com/>), and firewalls (AlpineLinux, <http://www.alpinelinux.org/wiki>) (Ipfire, <http://www.ipfire.org/en/index>) use

heuristics to identify potential computer based risks. Most of these systems however, treat their particular heuristics or approaches as trade secrets.

Two examples of such systems are described in more detail: Brightmail (<http://www.symantec.com/business/products/family.jsp?familyid=brightmail>) is an advanced antispam and email security solution provided by Symantec. Brightmail is based on many heuristic based rules learnt from user reported spam mails. SpamAssassin (<http://spamassassin.apache.org/>) is an Apache project that serves as a mail filter meant to eliminate Spam. It uses a wide variety of local and network tests to identify spam signatures. The testing mechanisms include header and text analysis, Bayesian filtering, DNS blocklists and collaborative filtering.

All heuristic based tools have their own problems. One significant disadvantage with SpamAssassin and most heuristic based tools is that authors of heuristics are forever playing catch-up to the authors and discoverers of exploits. The more public a “fix” is, the quicker spammers etc. are to move to a new system and the easier it is for them to verify whether their new solution evades current filters. Therefore previous research has concluded that complete automation of security decisions through the use of machine based testing is unreasonable and inefficient (Edwards et al. 2008).

The wide adoption of the Internet has given people the power to publish and access a large wealth of information. Many users provide ratings for things like movies, shopping experience, business experience, websites etc. A wide range of research has been done on the use of such rating mechanisms or recommendations, and how they can be used successfully and efficiently (Jin & Si 2004)(Chen & Singh 2001)(Amatriain et al. 2009)(Dellarocas 2000). Ratings for products and services published on the Internet are increasingly important, as they allow users to harvest the wisdom of the community in making decisions (Chen & Singh 2001). Using community data like user feedback is not a new idea - word of mouth and acquaintances have long been valued sources for information, and many studies have

shown the efficiency of peer based community data (Goecks, Edwards & Mynatt 2009). Online reviews are a major information source for consumers (Hu et al. 2006).

Social navigation is the practice of using information from other people, exploiting social practices and behavior, to help users attain their goals. Researchers have applied social navigation systems to many diverse domains (McNee et al. 2006)(Svensson et al. 2001). Social navigation is also potentially useful in online security management (Goecks, Edwards & Mynatt 2009). In such an approach one collects data from many users and aggregates it in order to generate recommendations. It has been theorized that through proper presentation and screening, this data can help users make better decisions related to online privacy and security (Goecks, Edwards & Mynatt 2009).

This approach is not without its problems however. Careful screening is necessary to avoid problems like herding behavior and information cascades. In an area where most users have relatively little technical knowledge (as in online security and privacy), it is important to make sure that knowledgeable voices are promoted and weighted differently. Previous research has addressed these issues and provided ways to mitigate these problems (Goecks, Edwards & Mynatt 2009).

Due to the inherent noise present in most community data, it is important to normalize the ratings of different users to the same scale (Jin & Si 2004). Ratings providing some review rationale are also seen as especially valuable. Explanations provide many benefits, from improving user satisfaction to increasing user confidence to helping users make better decisions by matching their concerns and expectations with different reviewers (Vig et al. 2008). These comments and feedback help users make more reliable and correct decisions (Bilgic & Mooney 2005)(Herlocker et al. 2000).

At the same time, reading a large number of user comments and feedback is not always feasible or desirable. Reviews may be associated with a number of ratings or tags, which can facilitate clustering, analysis, searching, or filtering. Tags have become popular with many websites such as Delicious, Flickr, and Amazon. Tags provide both factual and subjective descriptions (Sen et al. 2006).

2.4 METHODOLOGY

In order to examine our research questions we had to conduct two connected experiments: The first being a large-scale data collection and analysis of website ratings, done over time, and a smaller, more in-depth analysis of user comments and heuristic ratings of a subset of websites.

For the first experiment our goal was to gather a broad statistical base with which we could determine whether there are significant differences in terms of the quality, focus, and stability of the ratings given by MSA and WOT for a website. This required us to gather multiple data sets.

We decided to examine sites that came from the list of top trafficked websites as published by Alexa (<http://www.alexa.com>). In order to cover a wider range of websites we removed all duplicate domains. For example, the inclusion of www.google.com excluded sub-domains like www.mail.google.com and www.google.co.in. Many websites are reviewed by a small number of users, and their ratings are thus inherently unstable.

In order to account for this, WOT provides a popularity rating for each website. While the formula used to derive the popularity rating is proprietary, WOT says this rating reflects their level of confidence in their rating. All sites in our sets have a minimum popularity rating of 3 out of 5. For the second experiment we also had to filter out any website which had comments which were not in English as this complicated our text analysis. Finally, we had to filter out websites not reviewed by both MSA and WOT.

McAfee's SiteAdvisor (MSA) performs extensive and frequent heuristics-based evaluations of websites. These evaluations include tests for phishing, infected downloads, spam, drive-by-downloads, e-commerce vulnerabilities, browser exploits, popups, etc. Each site is then grouped into one of a number of risk categories: Safe, caution, risky, and untested. Each category is associated with a color, used to give users quicker, at-a-glance information (see Figure 2.1).



Figure 2.1 At-a-glance security & privacy indicators

facebook.com

We've tested this site and found it safe to use.
Are you the owner of this site? [Leave a comment](#)

Contact information: Country: United States Popularity: Lots of users

McAfee Secure Search
Search worry free.
We will block risky sites in your search results.

AUTOMATED WEB SAFETY TESTING RESULTS FOR FACEBOOK.COM

E-MAIL TESTS FOR FACEBOOK.COM: [?](#)
< 1 e-mail/ month
After entering our e-mail address on this site, we received less than 1 e-mail per month.
[View detailed analysis](#)

What our inbox looked like after we signed up here:

Subject	Sender	Date
Re: Remove From Database	pri...@support.facebook.com	2010 October
Re: Remove From Database	pri...@support.facebook.com	2010 October

DOWNLOAD TESTS FOR FACEBOOK.COM: [?](#)
67 green downloads
In our tests, we found downloads on this site were free of adware, spyware, and other potentially unwanted programs.
[View detailed analysis](#)
[Submit a download for analysis](#)

Downloads we found on this site:

Download	Analysis
FacebookPhotoUploader3.cab	
dosperl.zip	
perl_t12_archimedes.zip	
hivemind-1.1.zip	
hivemind-2.0-alpha-1-bin.zip	

67 total downloads. [See more.](#)

ONLINE AFFILIATIONS FOR FACEBOOK.COM: [?](#)
Linked to green sites
When we visited this site, we found that most of its links are to sites which are safe or have only minor safety/annoyance issues.

```

graph TD
    facebook[facebook.com] --> apache[apache.org]
    facebook --> apple[apple.com]
    facebook --> flickr[flickr.com]
    facebook --> youtube[youtube.com]
    facebook --> ilike[ilike.com]
    facebook --> facebooknet[facebook.net]
    facebook --> github[github.com]
    facebook --> bitly[bit.ly]
  
```

Figure 2.2 McAfee SiteAdvisor site review page

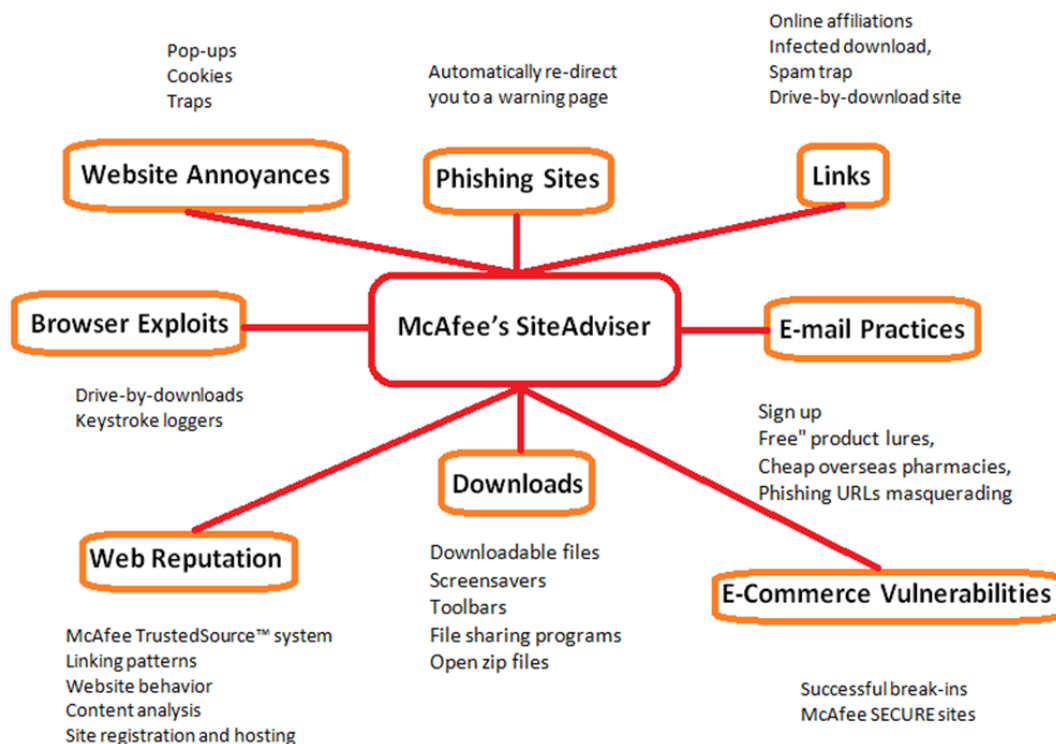


Figure 2.3 McAfee SiteAdvisor decision-making model.

Furthermore, a detailed site report is available describing the type of site, results of email tests, downloads tests, and online affiliations (see Figure 2.2). These tests are updated regularly. As seen in Figure 2.3, McAfee injects some reputation into its model through its McAfee TrustedSource™ system.

Web of Trust (WOT) is a community based safe surfing tool where users rate websites based on four factors: Trustworthiness, vendor reliability, privacy, and child safety. Users can post comments tagged as “good”, “bad”, or “reference or other”. These ratings are processed by WOT and a composite reputation score is given for each site. These ratings are displayed to end-users through either a site warning (see Figure 2.4), or a browser plug-in. Sites are grouped into one of 5 reputation categories: Excellent, good, unsatisfactory, poor, and very poor (see Figure 2.1).

creatives.livejasmin.com

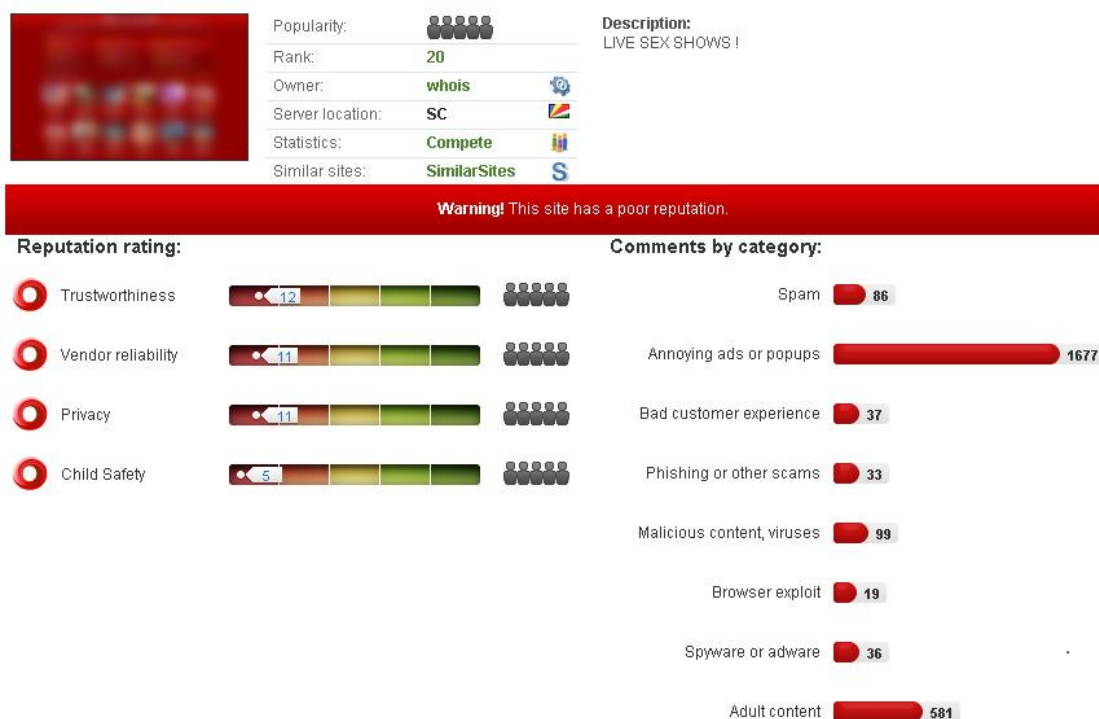


Figure 2.4 Web of Trust site review page

WOT encourage users to give a numerical score for a site as well as a textual feedback in the form of comments. This data can be much more informative than a numeric score, as it exposes the reasoning and criteria used for arriving at said score. At the same time, this unstructured data is difficult to process, both for users as well as the system. Users are often looking for thumbs up or thumb down recommendation when visiting a website, not to read 20-30 reviews. Unfortunately, this unstructured data is often crucial to understanding ratings. Someone who arrives at an adult site while looking for adult content is likely to rate said site differently from someone who is surprised to arrive at the same site through a misleading link or search. Therefore, effort has to be put into extracting as much valuable information from comments, while filtering out the noise as efficiently as possible.

For our first experiment we collected reports for 20,000 websites on May 2010. These reports were collected from WOT and MSA, thus sites had to be reviewed by both

services. Of these 20,000 websites, 18,650 met this criterion. These 18,650 sites were selected based on traffic, geographic representation, and topical distribution. Next, we gathered data on the top 10,000 trafficked sites, as specified by Alexa on July 2010, from WOT and MSA.

For our second experiment, our goal was to collect data to determine how closely aligned the review criteria of heuristic and community-based review sites were, and what value one technique could add to the other. This required a careful analysis of user comments, thus forcing us to look at a smaller data set. Given that we were primarily interested in understanding why disagreements emerge, we carefully constructed three data sets to examine this question.

Set A: Set of websites rated as ‘good’ or ‘safe’ by both SiteAdvisor and Web of Trust.

Set B: Set of sites with conflicting rating between SiteAdvisor and Web of Trust.

Set C: Set of sites rated as ‘poor’ or ‘risky’ by both SiteAdvisor and Web of Trust.

We sampled from the list of top 10,000 websites, as listed by Alexa in June 2010. We filtered this list according to a set of selection criteria (listed below). Because of this filtering, the samples ended up being of different sizes, but roughly comparable.

Set A consisted of websites which were rated as ‘Excellent’ or ‘Good’ by WOT and ‘Safe’ by MSA. All sites in this set have a minimum of 100 user comments, and 90 % of these tagged as ‘Good’ in WOT. We had 82 websites in Set A.

Set B consisted of 97 websites where MSA and WOT were in disagreement. All sites in this set had a minimum of 5 user comments in WOT. We further divided this set into two according to the nature of the disagreement. B1 consisted of the 47 websites that were rated ‘Good’ by WOT and ‘Risky’ by MSA. B2 consisted of the 50 websites rated ‘Poor’ by WOT and ‘Safe’ by MSA.

Set C consists of websites that were rated as ‘Poor’ or ‘Very poor’ by WOT and ‘Risky’ by MSA. There were a totally of 54 sites in Set C. All these sites had a minimum of 5 user comments in WOT.

After the sets were identified, we collected all WOT user comments for these sites. The number of comments differed from site to site. Overall, the number of comments for Set A was higher than for the sites in Set B and Set C.

2.5 RESULTS

Starting with **RQ1**: How often are there conflicts between heuristic and community-based ratings? We used the first data set, containing 18,650 websites to examine this issue. We first filtered out the websites which did not have a ‘good’ or ‘bad’ rating with both WOT and MSA. In other words, we removed the sites for which MSA or WOT were unsure about. Table 2.1 gives an overview of the different sets.

Table 2.1 System discrimination

System	Set	Sample Size	% of sample
WOT	Good	16,968	90.79%
	Bad	1,722	9.21%
MSA	Good	18,062	96.64%
	Bad	628	3.36%

As we see from Table 2.1, the vast majority of sites in our sample were seen as safe, or ‘good’. This is not surprising as we sampled from top websites. If a significant number of these sites were engaging in dubious behavior, they likely would be punished by users and quickly find their way out of the rankings.

More interesting, we find that WOT users seem more selective in evaluating websites, finding more faults (9.21% of sites found to be ‘bad’ by WOT users compared to 3.36% by MSA). We performed Pearson’s chi-square test to check whether the difference is statistically significant or not. We found that these differences were highly significant ($\chi^2=543.466$, $p<0.0001$, $DF=1$). This is important when we consider

the risks associated with Type I and Type II errors, as we will in the discussion section of this paper.

The sum difference does not tell the whole story however. Two review systems might arrive at the same number of ‘bad’ sites, yet disagree on which sites are ‘bad’. Therefore the next step is to determine the amount of disagreement between these sources. Table 2.2 gives an overview of the degree of disagreement in our sample.

Table 2.2 Degree of mismatch between WOT and MSA

Input set	Sample Size	Mismatch MSA/WOT	# Mismatch
Rated “Good” on WOT	16,968	2.06%	350
Rated ”Bad” by WOT	1,722	81.54%	1,404
Total	18,690	9.38%	1,754

A mismatch occurs when either WOT says the site is ‘good’ while MSA says it is ‘bad’, or when MSA says a site is ‘good’ and WOT says it is ‘bad’. What we found was quite surprising: MSA very rarely disagreed with WOT users when these claimed that sites were safe (only 2.06% of cases), whereas MSA claimed that 81.54% of the sites WOT users considered to be unsafe were safe. This is a surprising degree of disagreement, the roots of which we will explore when discussing RQ3.

Turning to **RQ2**: How stable are community-based ratings when compared to heuristic-based ratings?

To answer this question we turn to the second data set, based on the top 10,000 sites in July 2010. We compared this list to the 18,690 sites we collected in May of the same year. From this list we found an overlap of 5,743 sites, the remaining 4,257 sites were either not part of our original set or were being reexamined by MSA. What we were looking for was how many websites had gone from one rating to another (i.e. from ‘good’ to ‘bad’ or vice versa). This means that we were not interested in sites that got additional reviews if these were consistent with prior reviews.

As we can see in Table 2.3, we found surprisingly few changes over the two-month period. Doing a Pearson's chi-square test, we do find that there was a statistically significant difference in terms of the rate of change between WOT ratings and MSA ($\chi^2 = 31.208$, $p < 0.0001$, $DF = 1$), however, at least for top websites, this rate of change is very small.

Table 2.3: Stability of site ratings

Data	# of sites	# of sites changed	% change
WOT	5,743	34	0.59%
MSA	5,743	1	0.02%

Finally, **RQ3**: When disagreements occur, are these due to the use of different rating criteria, or are they due to disagreements and inaccuracies evaluating the same criteria?

In order to tackle this question, we turn to our second experiment, where we analyzed a smaller set of sites in greater detail and extracted rating rationales from the many user comments. In total we collected 12,677 user comments from the 233 websites evaluated in this smaller experiment. Of these, 7,158 comments were related to sites in Set A, 4,688 comments were related to sites in Set B, and 831 comments were related to sites in Set C.

We can confirm after this analysis that the quality, focus, and accuracy of user comments varies wildly. User comments differ in quality and reliability, as well as what users choose to focus on in their review. Reviews of a website often include contradictory statements. Therefore, these comments need filtering and normalization before comparisons can be made.

WOT allows its users to categorize their comments when they post them. Reviewers can categorize a comment as 'Good', 'Bad', or 'Other'. Of the 12,677 user comments, 7,453 comments were tagged as 'Good', 4,691 comments were tagged as 'Bad', and

533 comments were tagged as ‘other’. Each of these categories were tagged with terms like ‘adult’, ‘spam’, ‘viruses’, etc. (see Figure 2.4).

Table 2.4 Site reviews

Data	# of reviews	Good reviews	Bad reviews
Set A	7,158	86.70%	9.80%
Set B	4,688	24.18%	70.43%
Set C	831	14.07%	82.67%

Based on the chi-square test result we obtained (Pearson’s chi-square test, $\chi^2=5501.475$, $p<0.0001$, $DF=2$), the difference in ratio of good to bad reviews between Sets A, B, and C was statistically significant. This means that as we’d expect disputed sites get more bad reviews than ‘good’ sites, and ‘bad’ sites get the most bad reviews. Interestingly, ‘good’ sites are more likely to get reviews than the ‘bad’ sites. Part of the reason might be due to sites in Set A being the most popular in terms of visitor numbers while Set C sites were the least popular.

We used our own filter and tag cloud algorithm to organize and visualize the most common words in user comments. Our algorithm uses an English dictionary to filter and process the comments. The English dictionary was used to help identify words and variations of words for clustering. This algorithm was used to identify a set of high frequency words, based both on the total number of times a word appeared, as well as how many comments it appeared in overall. This was done to ensure that repeated use of a word in a single post would not skew results. To ensure that no important words were missed, we compared our results to those of 3 online tag clouding websites Wordle (<http://www.wordle.net/create>), Artviper (<http://www.artviper.net/texttagcloud/>), TagCrowd (<http://tagcrowd.com>). We used these online services to build clouds, which we used to identify key “non-words” that we needed to grandfather into our English dictionary, like ‘NIX4KIDS’. ‘NIX4KIDS’

is not an English word, but many reviewers used the term in their comments related to adult sites.

We did a separate analysis for the tags in the good reviews and the bad reviews. This was done to identify words that were uniquely associated with positive and negative reviews, as opposed to generally popular words. We of course looked for modifier words, like ‘not’ before doing this filtering. We also tagged technical words, but left them in their respective good and bad dictionaries.

A word was included in our “Good” dictionary if it met the following 3 conditions:

1. At least 0.5% of the total comments contained the word
2. At least 0.75% of the good comments contained the word
3. The rate of use of the word in good comments should be at least 0.5 percentage points higher than that of the bad comments.

The same rules were followed for the “bad” dictionary:

1. At least 0.5% of the total comments contained the word
2. At least 0.75 % of the bad comments contained the word.
3. The rate of use of the word in bad comments should be at least 0.5 percentage points higher than that of the good comments.

The statistical limits were set after some experimentation, but this would likely benefit from more rigorous study, especially as the sets grow.

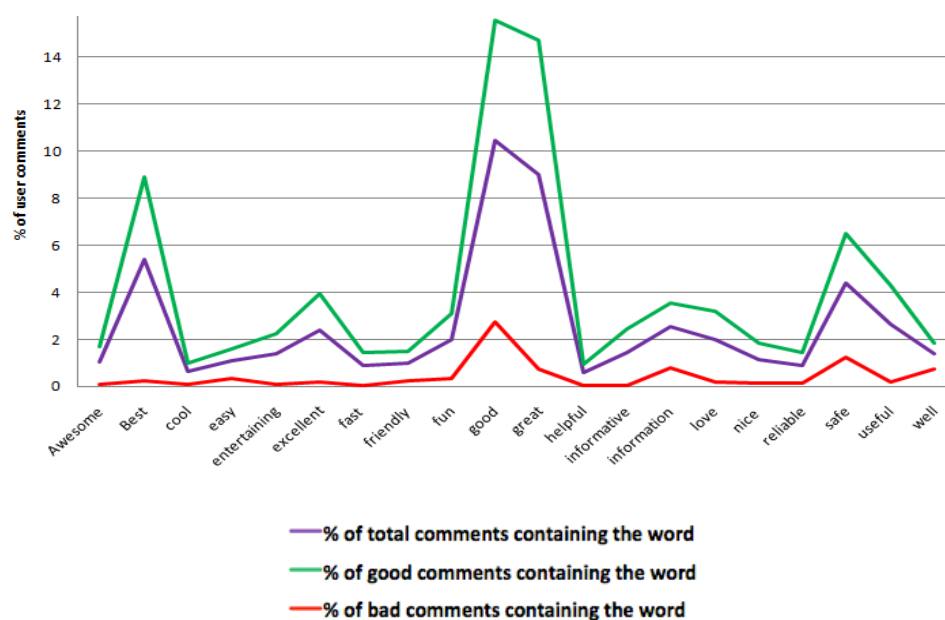


Figure 2.5 Construction of the 'Good' dictionary

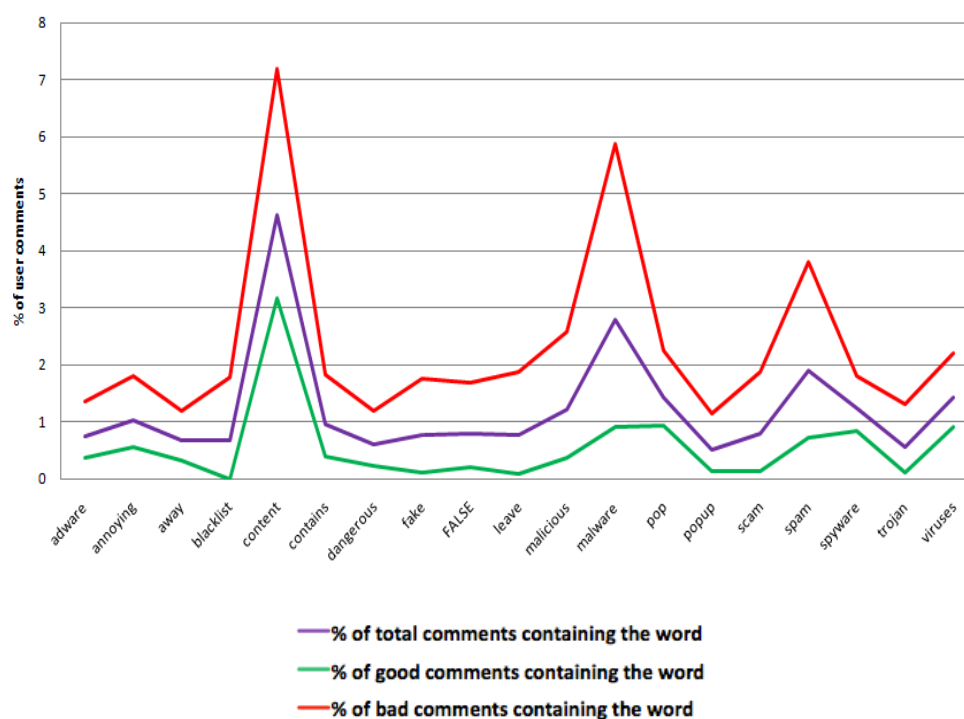


Figure 2.6 Construction of the 'Bad' dictionary

There were many websites with adult content in our three data sets. Adult content can sometimes be controversial. For someone looking for adult content, finding it is a good thing. Finding adult content when one is not looking for it is often undesirable. Adult words can therefore be both positive and negative, and can sometimes be used as both in the reviews of the same site. We therefore decided to build a special dictionary of adult words.

We calculated the number of user comments classifying a site as an ‘Adult site’. There were naturally more adult websites in Set B and in Set C compared with Set A (see Figure 2.7). We defined a site as an adult site if at least 50% of comments stated that the site contained adult contents.

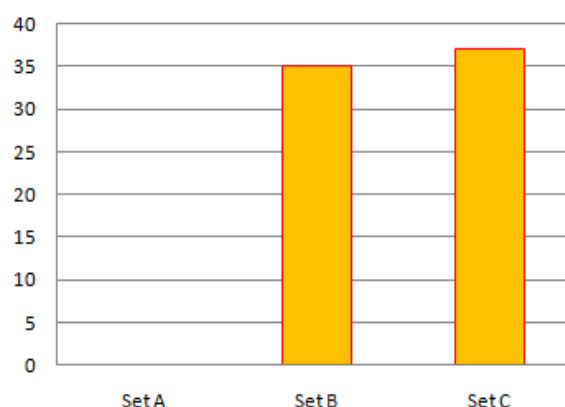


Figure 2.7 % of adult sites in the three samples

The dictionaries we ended up with contained the following words:

Good dictionary: Awesome, Best, Cool, Easy, Entertaining, Excellent, Fast, Friendly, Fun, Good, Great, Helpful, Informative, Love, Nice, Reliable, Safe

Bad dictionary: Adware, Annoying, Away, Blacklist, Content, Dangerous, Fake, False, Leave, Malicious, Malware, Popup, Scam, Spam, Spyware, Trojan, Virus,

Adult dictionary: Adult, Pornography, Children, Explicit, NIX4KIDS, Sex

The exact content of these dictionaries is affected by the small sample sizes in this part of the study. With a larger sample size, both the content and the size of these dictionaries would likely change. Idiosyncratic words and classifications (such as the word ‘Content’ being in our ‘Bad’ dictionary) would be less likely.

Looking at these dictionaries, we notice that a large number of the words in the bad dictionary are technical terms, and represent technologies and techniques that would have been sought out by MSA. Examples include “Adware”, “Malware”, “Spam”, “Spyware”, “Trojan”, and “Virus”. Thus looking back to our research question, we must conclude that there is some overlap in the rating criteria, but looking at Figure 2.6, we see that except for the terms “Malware” and “Spam”, these terms were not outstanding features of the bad dictionary.

When we look at the content of the comments in Set B, we see that the most important ‘bad’ word is ‘content’ (as in contains), and most of the disagreements with the MSA ratings (where MSA has declared the site safe), are due to either the presence of adult content, or undesirable practices not related to technology. Looking at Figure 2.8, we can see how the sites in Set B were truly a mixed bag, and adult content was very prevalent. We therefore conclude that while there is some overlap between the rating criteria of WOT and MSA, it is not the only source of disagreement, and that user comments add a significant amount of data in the form of tags of unsavory content and business practices.

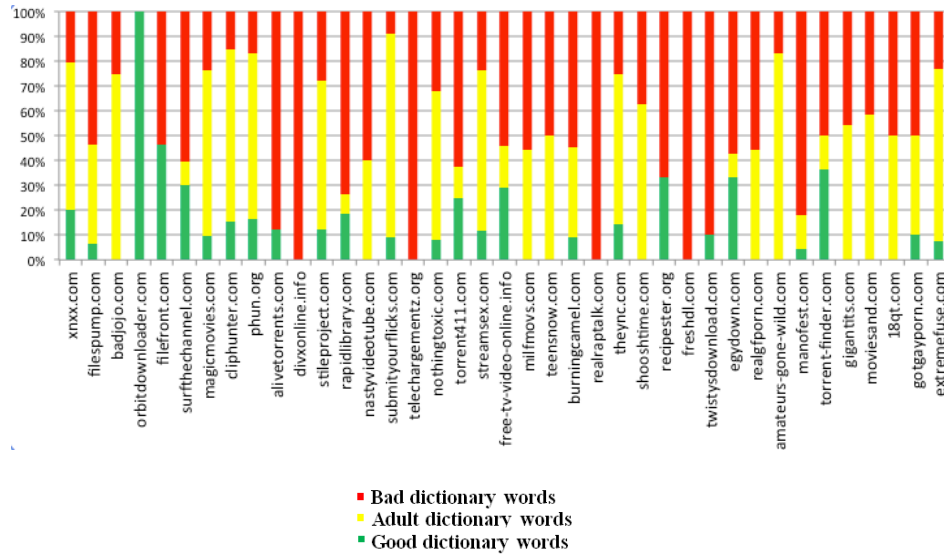


Figure 2.8 Classification of websites in Set B using dictionaries

2.6 DISCUSSION

The sampling decisions in our experiments were based on the goals of the experiments. Experiment 1 was done to study the focus and review criteria of MSA and WOT and the causes of disagreements between the two, thus we wanted as large a sample as possible. However, the second part of that experiment was aimed at analyzing how stable MSA and WOT ratings were over time. In order to determine whether the ratings were stable, some time had to pass. However, if too much time was allowed to pass, chances are that the site's practices would change, thus triggering a justified change in rating which might have confounded our study. We decided to study sites over a period of 2 months, deeming this to be an acceptable compromise. This decision means that for this part of the experiment we lost a significant number of sites from our sample set at this point. We realize however that not all will agree with this definition of stability, or with our concerns that site-evolution might taint our stability findings. We do recognize that the loss of sites at this step in the process is problematic, and if we were to redo this study, we would have looked for ways to avoid such a loss of data.

For our second experiment we were interested in determining what differentiated ‘good’ sites from ‘bad’, which required us to review user comments. Unfortunately, while it was relatively easy to find examples of popular well-reviewed ‘good’ sites, it was considerably harder to find examples of ‘bad’ sites. There exists no directory of such sites, so we were forced to resort to trial and error as well as a certain amount of random chance to find enough sites for sample sets B and C. This meant that we had a much harder time collecting data for ‘bad’ sites, both in terms of the sample sizes as well as the number of reviews available for each sample. It is therefore important that the reader not look to our dictionaries as the authoritative list of what defines a ‘good’ or ‘bad’ site, but rather as a proof of concept for how to extract such knowledge from user comments. Further study should in any case be done to determine what acceptable and useful thresholds should be for inclusion or exclusion into a dictionary, as well as determining what the most useful dictionaries for this application might be. While we think our three dictionaries (‘good’, ‘bad’, and ‘adult’) are useful, there might be room for more (a ‘technical’ dictionary perhaps?)

One of the most important findings from this study is the low rate of change of ratings (less than 0.6% over 2 months for WOT), both for MSA and WOT. At the same time, the total amount of disagreement between MSA and WOT was less than 10% over more than 16,000 websites. One of the main criticisms of community based rating systems has been that there is no guarantee that they will result in reliable ratings, or that these ratings will be stable. As we have seen from this study, those criticisms are not fair for all community-based rating systems.

Another interesting finding was the fact that WOT users were more critical of websites than MSA was, marking more websites as ‘bad’. While this study did not try to objectively determine which rating was most correct, if such a thing could be determined objectively, when there is risk for fraud or privacy invasions, it might be better to play it safe. In this sense, encouraging WOT users to mark up sites which are borderline or dubious is probably a good idea.

People are generally poor at evaluating risk and technology, but good at evaluating the social, legal, and intangibles. Therefore, it might be wise to construct a separate dictionary of technical terms, and use this to carefully evaluate the reliability of technical claims users make about sites.

Our work to date has been limited to a relatively small sample of sites. More importantly, the bulk of these sites have been selected from the most popular sites on the Internet. To validate the scalability and generalizability of this work, and our tag extraction approach, we should extend our research to cover a larger set of sites, spanning a greater set of topic areas, and engaging in a wider set of business practices. This means of course, identifying a set of relatively obscure sites as well, as they are more likely to be engaging in unsavory or dubious behavior.

Rather than relying on lists and ratings to do this, we instead propose to develop a browser plug-in that will record more realistic browsing patterns. This plug-in can of course also be used to provide feedback to users, which would allow us to determine whether the combined data is more successful at spurring user to action, or helping them make the right decisions.

In the work we have done to date, we have not done any research to determine the error rates of either MSA or WOT. We know that at least some of the disagreements between the two must be due to faulty information with one of the two review systems. While it is easy to assume that MSA will be more accurate when looking for security and technology-specific privacy threats, this is a hypothesis that should be tested. Understanding how big a disconnect there is, or how big the false positive and false negative rates are would be an important contribution to our understanding of the value of user ratings.

Another future step is to examine how dictionaries should be weighted (how bad is the inclusion of an ‘adult’ word in a ‘good’ site), and if and how words should be weighted within dictionaries. This is an important next step when one considers how to practically combine heuristic data, which tends to generate a numerical or

categorical score, and community feedback, which is inherently verbose and qualitative. Each word in a dictionary may be assigned a weight or numerical score depending on its relationship with user reviews and comments about that specific website. In fact each dictionary itself may have a numerical score for a specific type of website.

In this paper we have demonstrated how this approach can be used to potentially deal with morally ambiguous content such as adult sites. This is just one example of many areas of the World Wide Web where a certain amount of context needs to be applied in order to give users more nuanced and useful warnings and recommendations. This same approach might be useful when dealing with other troublesome dimensions like security, privacy, business practices, etc. where different users are going to have different concerns with regards to different websites and contexts.

User comments and feedback have their own problems, like information cascades, herding behavior, incorrect comments, misusing of community data etc. These problems are not addressed in this paper, but have been described eloquently in (Goecks, Edwards & Mynatt 2009), which also shows some successful mitigating measures for them. In the future, it would be interesting to examine whether mitigating behaviors such as those described in (Goecks, Edwards & Mynatt 2009) are naturally occurring in some of the community sites such as WOT, whether administrators promote them, or whether these communities have developed their own norms and techniques to avoid these pitfalls.

2.7 CONCLUSION

Users were found to be more discriminating than heuristic-based systems, at least for the sample used in our study. This may be an indicator of a higher false positive rate with community-based systems, or it may be evidence that users have access to more meaningful information (again, at least for this sample), or that users are more risk-averse than heuristic-based systems give them credit for.

User ratings and heuristic-based systems are both relatively stable, and do not change significantly over short periods of time (though user ratings change more rapidly than heuristic-based reviews). Given the low rate of change, we cannot determine how much of the change is due to genuine change in the site policies, and how much is due to shifting opinion.

Our results show that information abstracted from user comments may add value in terms of additional data to the evaluation and communication of risk to users by automatically analyzing large numbers of user comments. This may help people in at-a-glance digesting large volumes of user feedback easily. Previous research has shown that users are influenced by user feedback; this abstracted form of that information in form of dictionaries may add sufficient nuance to the heuristic machine based testing systems like McAfee SiteAdvisor and Norton SafeWeb.

2.8 ACKNOWLEDGEMENTS

We wish to thank our colleagues in the HCI group at the School of EECS, Oregon State University for their support and help in preparing the paper. We'd also like to thank Web of Trust and it's users for giving us access to their data for this study. Finally, we'd like to thank our reviewers for thoughtful feedback and suggestions on how to improve this paper.

2.9 REFERENCES

- Amatriain, X. et al., 2009. Rate it again. In *Proceedings of the third ACM conference on Recommender systems - RecSys '09*. New York, New York, USA: ACM Press, p. 173. Available at: <http://portal.acm.org/citation.cfm?id=1639744> .
- Bilgic, M. & Mooney, R.J., 2005. Explaining recommendations: Satisfaction vs. promotion. *Proceedings of beyond personalization*. Available at: <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.107.488&rep=rep1&type=pdf#page=13>.
- Chen, M. & Singh, J.P., 2001. Computing and using reputations for internet ratings. In *Proceedings of the 3rd ACM conference on Electronic Commerce - EC '01*. New York, New York, USA: ACM Press, pp. 154-162. Available at: <http://portal.acm.org/citation.cfm?doid=501158.501175>.

- Dellarocas, C., 2000. Immunizing online reputation reporting systems against unfair ratings and discriminatory behavior. In *Proceedings of the 2nd ACM conference on Electronic commerce*. New York, New York, USA: ACM, p. 157. Available at: <http://portal.acm.org/citation.cfm?doid=352871.352889>.
- Dourish, P. et al., 2004. Security in the wild: user strategies for managing security as an everyday, practical problem. *Personal and Ubiquitous Computing*, 8(6), pp.391-401. Available at: <http://www.springerlink.com/index/10.1007/s00779-004-0308-5>.
- Edwards, W.K., Poole, E.S. & Stoll, J., 2008. Security automation considered harmful? *Proceedings of the 2007 Workshop on New Security Paradigms - NSPW '07*, p.33. Available at: <http://portal.acm.org/citation.cfm?doid=1600176.1600182>.
- Goecks, J., Edwards, W.K. & Mynatt, E.D., 2009a. Challenges in supporting end-user privacy and security management with social navigation. *Proceedings of the 5th Symposium on Usable Privacy and Security - SOUPS '09*, p.1. Available at: <http://portal.acm.org/citation.cfm?doid=1572532.1572539>.
- Gross, J.B. & Rosson, M.B., 2007. Looking for trouble: understanding end-user security management. In *Proceedings of the 2007 symposium on Computer human interaction for the management of information technology, March*. pp. 30–31. Available at: http://www.cc.gatech.edu/classes/AY2008/cs4235b_fall/Group1/UnderstandingEndUserSecurityMgt.pdf.
- Herlocker, J.L., Konstan, J. a & Riedl, J., 2000. Explaining collaborative filtering recommendations. *Proceedings of the 2000 ACM conference on Computer supported cooperative work - CSCW '00*, pp.241-250. Available at: <http://portal.acm.org/citation.cfm?doid=358916.358995>.
- Hu, N., Pavlou, P.A. & Zhang, J., 2006. Can online reviews reveal a product's true quality?: empirical findings and analytical modeling of Online word-of-mouth communication. In *Proceedings of the 7th ACM conference on Electronic commerce*. New York, New York, USA: ACM, pp. 324–330. Available at: <http://portal.acm.org/citation.cfm?id=1134707.1134743>.
- Jin, R. & Si, L., 2004. A study of methods for normalizing user ratings in collaborative filtering. In *Proceedings of the 27th annual international conference on Research and development in information retrieval - SIGIR '04*. New York, New York, USA: ACM Press, p. 568. Available at: <http://portal.acm.org/citation.cfm?doid=1008992.1009124>.

McNee, S.M., Kapoor, N. & Konstan, J.A., 2006. Don't look stupid. In *Proceedings of the 2006 20th anniversary conference on Computer supported cooperative work - CSCW '06*. New York, New York, USA: ACM Press, p. 171. Available at: <http://portal.acm.org/citation.cfm?doid=1180875.1180903> .

Sen, S. et al., 2006. Tagging, Communities, Vocabulary, Evolution. In *Proceedings of the 2006 20th anniversary conference on Computer supported cooperative work - CSCW '06*. New York, New York, USA: ACM Press, p. 181. Available at: <http://portal.acm.org/citation.cfm?doid=1180875.1180904>.

Svensson, M. et al., 2001. Social navigation of food recipes. In *Proceedings of the SIGCHI conference on Human factors in computing systems - CHI '01*. New York, New York, USA: ACM Press, pp. 341-348. Available at: <http://portal.acm.org/citation.cfm?doid=365024.365130>.

Vig, J., Sen, S. & Riedl, J., 2008. Tagsplanations. In *Proceedings of the 13th international conference on Intelligent user interfaces - IUI '09*. New York, New York, USA: ACM Press, p. 47. Available at: <http://portal.acm.org/citation.cfm?doid=1502650.1502661>.

Alexa., <http://alexa.com>

AlpineLinux., <http://www.alpinelinux.org/wiki>

AOL explorer., <http://www.aol.com/>

Artviper., <http://www.artviper.net/texttagcloud/>

Avira., <http://www.avira.com/en/pages/index.php>

Brightmail.,
<http://www.symantec.com/business/products/family.jsp?familyid=brightmail>

Ipfire., <http://www.ipfire.org/en/index>

McAfee., <http://www.mcafee.com/us/>

McAfee SiteAdvisor., <http://www.siteadvisor.com/>

McAfee SpamKiller., <http://us.mcafee.com/root/product.asp?productid=msk>

Mozilla Firefox anti-phishing., <http://www.mozilla.com/en-US/firefox/phishing-protection/>

Norton., <http://us.norton.com/index.jsp>

Norton SafeWeb., <http://safeweb.norton.com/>

Spamassassin., <http://spamassassin.apache.org/>

SpyBot., <http://www.safer-networking.org/en/index.html>

Tagcrowd., <http://tagcrowd.com>

Untangle., <http://www.untangle.com/>

Web of Trust., <http://www.mywot.com/>

Webroot popup washer., http://www.webroot.com/En_US/consumer-products-spysweeper.html

Wordle., <http://www.wordle.net/create>

3 Second manuscript: A User Evaluation of Heuristic and Community Based Safe Surfing Ratings: Which Do Users Trust More?

Prashanth Ayyavu, Carlos Jensen

School of EECS
Oregon State University
Corvallis, Oregon, 97331, USA
{ayyavu, cjensen}@eecs.oregonstate.edu

Proceedings of the 2012 The ACM SIGCHI Conference on Human Factors in Computing Systems
Austin, TX
SESSION: Security

(Submission in process)

3.1 ABSTRACT

Online safe surfing tools advise users about the safety of websites. They have become a standard tool for many users, and are commonly integrated into browsers and anti-virus suites. Some of these tools derive recommendations based on heuristics and others are based on community experience and recommendations. Due to this divergent focus, tools sometimes produce conflicting recommendations for websites. These conflicts are not necessarily an indication of failure, but rather of the non-overlapping nature of the reviewed. In this paper we examine how users react to security ratings, especially conflicting or mixed reviews, what sources of information users find more compelling and trustworthy, and whether combining the two approaches leads to improved confidence. To answer these questions we conducted a controlled user study. We found that confidence levels increase when heuristics are combined with community reports in a measured fashion, and that users do not trust sites with conflicting ratings.

3.2 INTRODUCTION

The Internet has become a ubiquitous part of everyday life in the industrialized world. Naturally, the number of websites users visit and rely on, and the speed with which these grow, change, and adopt new technologies, presents a huge challenge in terms of managing ones exposure and security online. Though security has been a major focus for many years, the Internet is still fraught with malicious software, and websites featuring browser exploits or engaging in identity theft and fraudulent practices (NCSA Norton Online Safety Study., <http://staysafeonline.mediaroom.com/index.php?s=67&item=57>). While our technical understanding of exploits and users awareness and acceptance of the need for security measures have increased, it is still very much an arms race between security professionals and users on one hand, and hackers on the other.

Attacks are not limited to the purely technical in nature however. Equally damaging to users is the proliferation of websites engaging in fraudulent or just poor business practices. This can range from simply poor customer support and turn-around times to

sites engaging in identity theft or fraudulent charges. While security firms and software developers keep racing to identify the latest technical exploits, these “low-tech” or “non-tech” threats are much harder to detect. It is therefore obvious that security threats cannot be detected or addressed solely through machine based rules and scanning software. Rather, it is a process in which user’s feedback and experience reports need to be factored in. This form of feedback is something that many users have grown to both expect and trust. However, this type of feedback is not without its problems, which include dealing with irate users with an axe to grind, less than knowledgeable reviewers, disparate expectations, and often long-winded and unstructured reviews.

One long-standing criticism of community or reviewer-based security systems has been that normal users often do not understand the technical aspects of online privacy and security exploits and risks, nor do they have the time or inclination to carefully inspect the technical elements of the websites they visit (Dourish et al. 2004). In reviewing such matters, users are often confronted with technologies and practices which they do not really understand, or whose consequences they do not fully understand (Dourish et al. 2004). Prior research has shown that users often prefer to delegate privacy and security management to others, as they often feel uncomfortable making highly technical security decisions (Dourish et al. 2004). It is for this reason that many users choose to rely on website reputation services, or safe surfing tools, like McAfee SiteAdvisor (SiteAdvisor., <http://www.siteadvisor.com/>), Web of Trust (Web of Trust., www.mywot.com), Alexa (Alexa., www.alexa.com) and Norton SafeWeb (SafeWeb., www.safeweb.com). These tools help users use the Internet more safely with a minimum amount of effort by identifying malicious websites and blocking these or warning users about the risks before they land on the site.

These tools try to present users with a very concise, at-a-glance indicator of the safety of a site. Often, a traffic light metaphor is used; ratings are translated to a colored coded and iconic representation (see Figure 3.1).



Figure 3.1 Various website reputation services and their colored-rating icons

Safe surfing tools can be divided into two categories based on how sites are evaluated: Heuristic analysis/machine-based tools, and community/end user rating based tools. Heuristic analysis tools in essence use scripts or heuristics to automatically scan websites for known exploits, or techniques and technologies that experts have identified as potentially problematic for users privacy or security. Most antivirus or anti-spam software build on this approach, though anti-spam tools often include some element of machine learning to allow them to better adapt and respond to changing trends and contexts. Community based tools are those that primarily rely on a user community to provide scores and reviews for websites. These ratings are based on users personal experience with these sites, and often provide a more holistic and less technical review of the site.

Though relatively little work has looked at the differences between these two types of tools, previous work (Ayyavu & Jensen 2011) has shown that these two families of tools focus on different aspects of the sites they review; heuristic tools focus on the technical foundations, and community-based tools focus on business practices. The same study concluded that, for the two tools studied, the two approaches led to very similar ratings, and that community-based reviews were more likely to err on the side of caution (Ayyavu & Jensen 2011). This study looked at the reviews for a large number of websites, but did not examine how end-users reacted to such ratings,

whether they felt compelled to take action, or whether they found these ratings trustworthy. What would happen if a user got conflicting recommendations, what would they do? A rating or recommendation system is only as good as the end-user allows it to be; if recommendations don't compel users to take action, or users don't have confidence in the recommendations made, they have limited value.

Understanding user's perception of ratings and warnings, how they react to these, and how they weigh the different warnings given based on available evidence is crucial to the design of better security and safe surfing tools. To add to our understanding of these decision-making processes, as well as answer outstanding questions regarding the relative merits of, or the possible value of combining elements of these two families of safe-surfing tools, as suggested in (Ayyavu & Jensen 2011), we conducted a controlled user study. Our goal for this study was to answer the following research questions:

- **RQ1:** Are users more heavily influenced by heuristic-based ratings, or community ratings? (Regardless of which is more correct)
- **RQ2:** Does combining the two types of ratings add value to users?
- **RQ3:** When presented with conflicting ratings, how do users act?

To try and answer these questions, and to align ourselves as closely as possible with the work of (Ayyavu & Jensen 2011), we chose to use SiteAdvisor (SiteAdvisor., <http://www.siteadvisor.com/>) as our sample heuristic-based tool, and Web of trust (Web of Trust., www.mywot.com) as our sample community based tool. These tools are popular in their respective methodologies, and have a large and active user base. We refer these tools as MSA and WOT in the rest of the paper. For each of these tools we will use the data they provide for real websites, as well as the UI mechanisms they use to communicate their recommendations to users in order to keep the study as grounded as possible. This means that some of the idiosyncrasies of these tools may influence our results, but on the other hand, we are basing our study on real production systems.

The rest of this paper is organized as follows: First we present a brief overview of the most relevant literature on safe surfing tools and people's perception and decision-making process as it relates to online security. Then we give a brief description of our methodology and experiment setup, followed by our results and discussions. We wrap up with conclusions and future work.

3.3 RELATED WORK

Heuristic-based safe surfing tools like McAfee SiteAdvisor (SiteAdvisor., www.siteadvisor.com/) , Norton SafeWeb (Norton SafeWeb., www.safeweb.com), Alexa (Alexa., www.alexa.com) , AVG LinkScanner (LinkScanner., www.linkscanner.avg.com/) and Kaspersky Safe Surf (Kaspersky., www.usa.kaspersky.com) and many others, use heuristics and pre-programmed patterns to automatically visit, test and scan large numbers of websites for known vulnerabilities and exploits. Heuristic-based tools can inform users about the risks associated with the technical foundations or infrastructure of websites, information which most naïve users either do not have access to, or lack the skills to successfully interpret. Such tools are not without problems. (Angai et al. 2010) found that different heuristic-based safe browsing services implement and maintain their own malware detection systems and algorithms, resulting in major discrepancies between providers. Each provider flags websites as malicious based on different criteria, and criteria that are not public. Algorithms cannot be used to check human-readable or experiential factors such as whether a website follows its privacy policies, or whether it engages in fraudulent practices. Another problem is the more public a “fix”, the quicker those with malicious intent move on to new exploits, and the easier it is for them to verify whether their new solution evades current filters. Furthermore attackers have been able to manipulate innocent websites to make them appear malicious to algorithms (krebsonsecurity., <http://krebsonsecurity.com/2010/04/hiding-from-anti-malware-search-bots/>), thus eroding user confidence in such ratings.

The growth of the Internet has given people the power to publish and access a wealth of information. Users routinely provide ratings for things like movies, products,

experiences, businesses, websites etc. Community-based safe surfing systems like Web of Trust (Web of Trust., www.mywot.com) collect user ratings and comments, process them and provide a reputation score for each website. Ratings for products and services published on the Internet are increasingly important decision-making tools for people, as these allow people to harvest the wisdom and experience of the community (Chen & Singh 2001). Some problems associated with this approach are 1) ensuring the reliability of user ratings, 2) herding behavior and 3) information cascades. Previous research has looked into these problems, and how to increase the reliability of the user ratings for such systems (Goecks et al. 2009).

Online Security is not a purely technological problem anymore, but an emerging societal problem (Colwill 2009). Previous research (Proctor et al. 2009) has shown that human factors will be critical in resolving issues surrounding user privacy and online security. No matter how well designed, security methods and tools rely on individuals to implement and use them. These tools may not accomplish their intended objectives if they are not used properly (Proctor et al. 2009). User feedback has become an important factor in online security management. According to the Google security blog (Google online security blog., [http:// googleonlinesecurity.blogspot.com/2007/ 11/help-us-fill-in-gaps.html](http://googleonlinesecurity.blogspot.com/2007/11/help-us-fill-in-gaps.html)), the search giant already knows about hundreds of thousands of "bad" websites, and is hoping that its users will add to the list by completing an online form to report new malicious sites.

Very little research has been published on website security indicators. (Maurer et al. 2011) found that passive indicators like rating icons, warning symbols were overlooked by users and even many blocking site mechanisms are routinely bypassed when users feel these get in the way of accomplishing their task. Security is not a goal in and of itself to most users; a security breach is an undesirable outcome to the process of accomplishing other goals. Thus, if risks are perceived as low, security measures are given low priority. Most non-blocking indicators are routinely overlooked by users busy with their primary task, or are quickly dismissed by users who get habituated to them.

According to one study of privacy policies, though users are concerned about their security and online privacy, they do not take time to thoroughly examine things like privacy policies (Tsai et al. 2009). (Wu et al. 2006) did a comparative study of different passive toolbar security indicators and found them ineffective in preventing most phishing attacks. They also identified some drawbacks with such tools: 1) A toolbar is a small display in the peripheral area of the browser, and thus does not always get noticed. 2) Security is rarely the user's primary goal in web browsing 3) Users may not care about the toolbar's display even if they do notice it. This last point is probably especially true for websites with which a user already has a relationship.

(Angai et al. 2010) analyzed the effectiveness of different safe browsing services, but focused only on heuristic-based services like McAfee, Google and Norton. This study compiled a list of potentially malicious websites and tested it against the above mentioned 3 services. They found that relying on a single service to protect a user from malicious attacks is insufficient. This study had the added drawback of focusing exclusively on sites in blacklists.

(Ayyavu & Jensen 2011) found that heuristics-based and community-based tools differ in quality, focus and methodology. They found that these two families of tools are good at addressing different areas of concern for online security and user privacy. WOT users seem more selective in evaluating websites, finding more faults compared to MSA. Disagreements between WOT and MSA were interesting because, MSA very rarely disagreed with WOT users when these claimed that sites were safe, whereas MSA claimed that more than 80% of the sites WOT users considered to be unsafe were safe (Ayyavu & Jensen 2011). They also explored some methods for abstracting key information from user comments, which could be used to add context and nuance to heuristic based ratings. What did paper (Ayyavu & Jensen 2011) did not explore was whether users found the two families of tools equally reliable, or whether they had a preference for one over the other.

3.4 METHODOLOGY

This paper builds on, and seeks to extend the work presented by Ayyavu and Jensen. (Ayyavu & Jensen 2011) in that we seek to examine what systems inspire greater confidence in users, which systems are more likely to compel users to act, how users react to websites with conflicting ratings from the two families of systems. (Angai et al. 2010) argued that data aggregation from different services increases user confidence, though this study only looked at combining ratings from services using the same general approach. No information from community-based tools was used. This means that in (Angai et al. 2010), some of these rating services had to be wrong, even if by omission. In the case of someone combining heuristic-based recommendations and community-based recommendations, conflicts are more likely caused by non-overlapping domains of investigation.

In order to answer our research questions, we conducted a limited controlled experiment using two existing security rating systems, MSA and WOT, as well as a prototype combining the ratings of both.

3.4.1 SETUP

Our setup enabled participants to browse websites and rate them at the same time. A web application consisted of two frames, as shown in Figure 3.2 was used. The Left (main) side of the web page displayed a randomly assigned website for users to study, complete with the browser chrome (the borders and headers), while the right hand frame displayed a simple questionnaire the users would use to rate the site and their confidence in their rating. When they were done rating the site, they would be taken to a new site to review, and possibly presented with a different rating/warning system in the browser chrome. To avoid confusion, the site was run in full-screen mode so the real browser chrome was hidden from view.

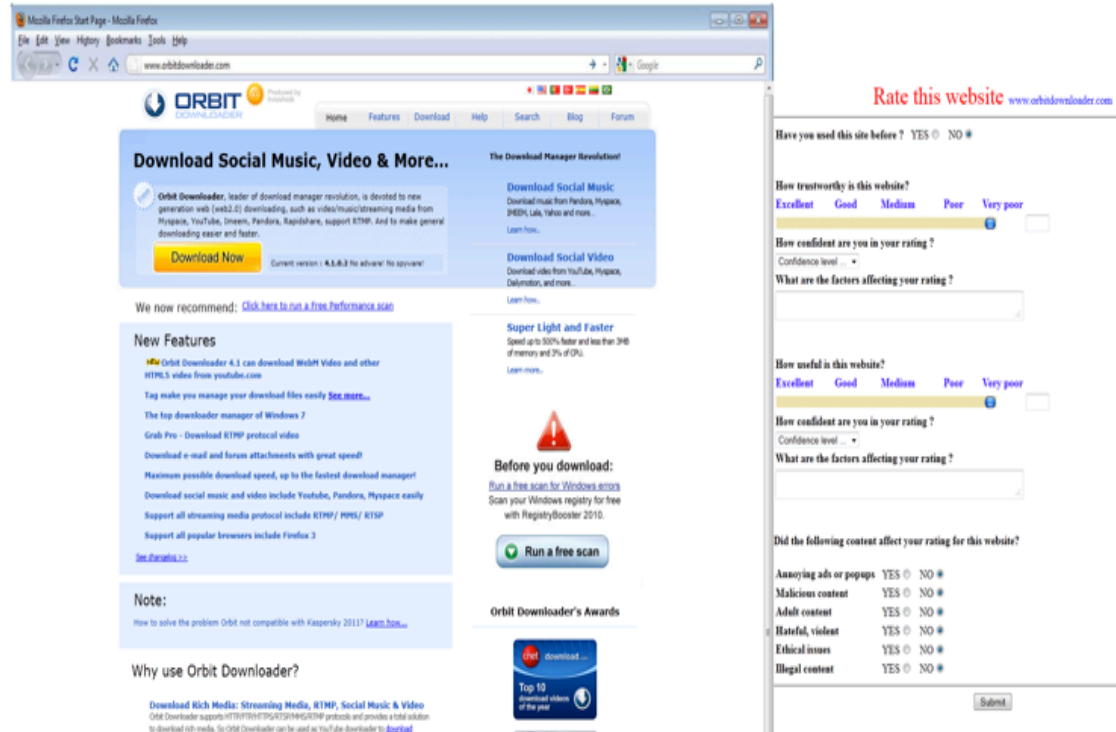


Figure 3.2 Design of study application. Left shows a website and right frame displays the questionnaire.

When a rating was displayed, all three rating systems would display an icon in the simulated browser chrome, next to the URL. The placement was identical for all 3 systems being evaluated (see Figure 3.3 for example). These icons were color and shape coded to convey a recommendation (see Figure 3.1). Clicking on one of these icons would bring up a more in-depth explanation of the rating (see figures 3.4 and 3.5). Participants were free to look up more details as they saw fit, and could spend as much time as they wanted examining the website before giving a rating.

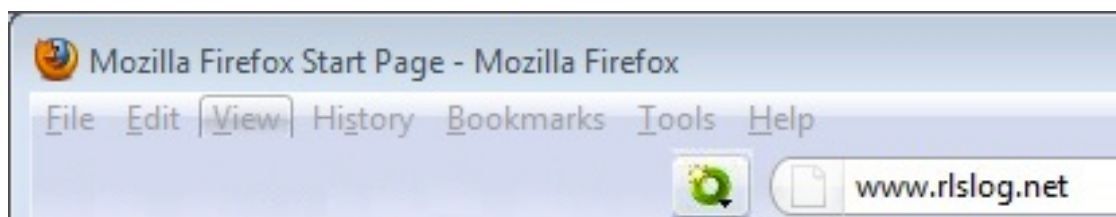


Figure 3.3 WOT rating icon near the website URL.



Figure 3.4 a) WOT rating card. b) MSA rating card.

3.4.2 PROTOTYPE SYSTEM

Our prototype combined information from WOT and MSA. High value keywords were extracted from user reviews by processing the hundreds of WOT user comments using the techniques discussed in (Ayyavu & Jensen 2011). In essence, the frequency of non-trivial words were measured, and categorized based on whether they were affiliated with positive or negative reviews. Keywords were grouped into dictionaries and associated with specific colors depending on the dictionaries in which they were present. Red for words frequently used in negative reviews, green for words commonly associated with positive reviews. Technical words were grouped in a category of their own and marked yellow, as these were often more thoroughly treated by heuristic-based tools. Our prototype featured a star shaped icon near the website URL similar to WOT and MSA. The color of the star icon depends on the type of ratings a particular site get from WOT and MSA (see Table 3.1).




Color	WOT rating	MSA rating
 Good	Excellent, Good	Safe
 Bad	Poor, Very poor	Risky
 Conflicting	Excellent, Good	Risky
	Poor, Very poor	Safe

Table 3.1 Color of star icon for a website, depends on the type of rating it gets from WOT and MSA

Users can click on this icon to see more details on a combination of ratings from MSA and WOT as shown in Figure 3.5.

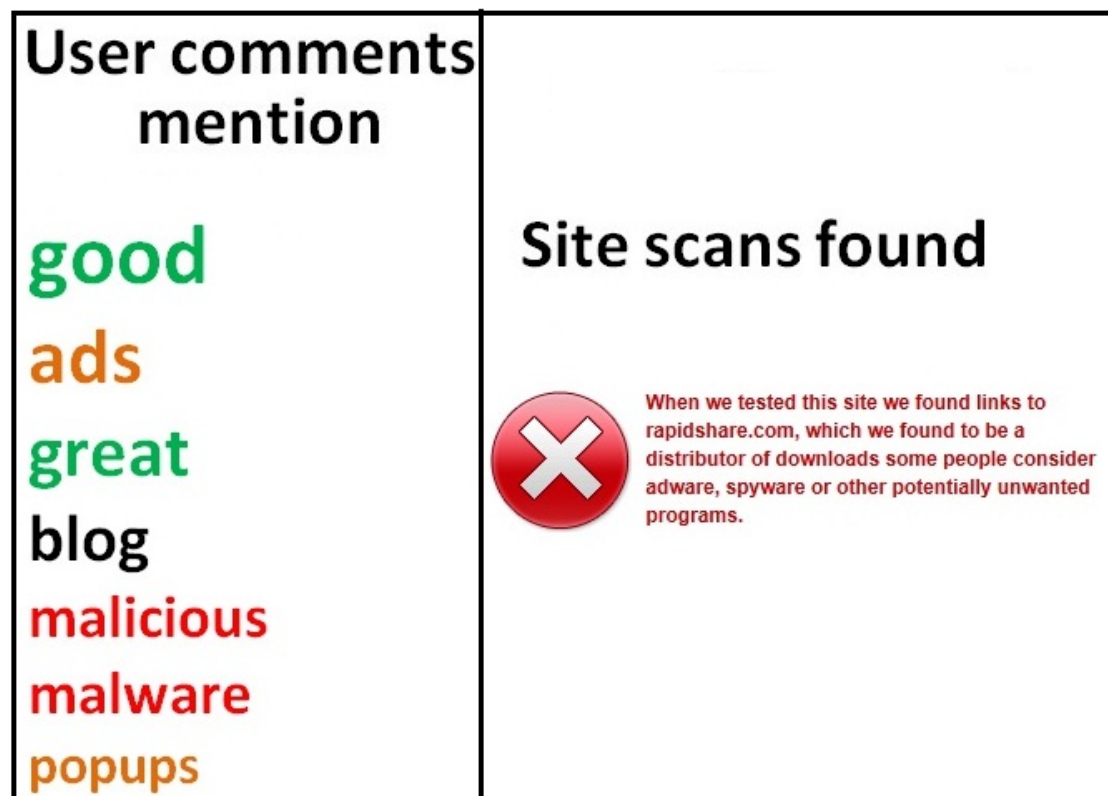


Figure 3.5 Prototype scorecard.

3.4.3 SELECTION OF WEBSITES

We selected 49 websites after analyzing more than of 20,000 websites. All 49 websites were unique domains and had been rated by both MSA and WOT close to the time of our study (spring/summer 2011). To avoid sudden changes in the ratings during the course of the study, we collected their ratings and used a static local copy, current as of August 1, 2011. The websites were selected based on the following criteria:

3.4.3.1 USER TRAFFIC

All websites used in this study had high user traffic. We used a number of third party tools and services like Alexa (Alexa., www.alexa.com), Statbrain (Statbrain., www.statbrain.com), Google Trends (Google Trends, www.google.com/trends), Compete (Compete., www.compete.com), and Quantcast (Quantcast., www.quantcast.com) to rank the websites in terms of users and daily hits. We felt that this was important because we wanted to ensure that the sample was as representative and relevant as we could make it. At the same time, we avoided the very most popular websites in order to maximize the chance that users would not have encountered the site previously and thus already formed an opinion about the safety of the site.

3.4.3.2 SITE POPULARITY

Related to, but different from traffic numbers, Web of Trust (Web of Trust., www.mywot.com) similarly gives a 5-point popularity rating for each website. This popularity rating is a metric to denote, among other things, how often a site has been reviewed. WOT does not disclose what goes into calculating the popularity metric, but this is what they refer to as their “reliability” metric for reviews. We felt this was an important metric because WOT relies on users for ratings. We wanted to use the websites which were not only generally popular, but also reasonably popular with the WOT user population in order to ensure we had enough data. All websites in our study had at least a 3 out 5 popularity rating, as given by WOT.

Each website used in this study had a minimum of 20 user comments in WOT. We wanted to make sure that the websites in our study had a good number user reviews, and that we had enough data to use. All user comments used in this study were in English.

3.4.3.3 CLASSIFICATION BASED ON RATINGS

We wanted to make sure we could cover a wide range of situations in this study, and thus sought to include sites with some dubious ratings. Ideally we would have an equal number of sites in each category of ratings, however, finding popular sites with poor reviews is a challenging problem. No service that we are aware of lets you rank or search for websites based on poor reputation. Ratings given by MSA and WOT were used in the selection process, and this divided our sample as shown in Table 3.2. 61% of the sites we selected had conflicting ratings between WOT and MSA. Among these 30 websites, 20 websites were rated “Excellent” or ”good” by WOT but rated as “Risky” by MSA. The remaining 10 had the inverse relationship. This helped us to focus more on how participants reacted to websites given with opposite ratings by community-based system and machine-based system.

# of sites	Rating	WOT	MSA
11	Good, Safe	Agree	Agree
8	Poor, Risky	Agree	Agree
10	Safe, Good	Agree	Disagree
20	Poor, Risky	Agree	Disagree

Table 3.2 Classification of websites under study based on ratings given by MSA and WOT.

3.4.3.4 PSEUDO-RANDOM ASSIGNMENT

Each participant was assigned to review 20 websites from our sample. These websites were selected in a pseudo-random manner in order to ensure that each participant saw some good and bad sites, as well as sites with mixed reviews (only relevant for the condition where they were using our prototype). Each participant rated 5 websites with either an “Excellent” or “Good” reputation from WOT and a “Safe” rating from MSA, 10 websites with conflicting ratings, and 5 websites with a “Poor” reputation score from WOT and a “Risky” rating from MSA. Within these categories the selection was random, as was the order in which these sites were presented.

3.4.4 EXPERIMENTAL CONDITIONS

We had three experimental conditions in this study, one for each of the three tools under study; MSA, WOT, and our hybrid prototype. Subjects were randomly assigned to conditions. Given 23 total subjects, (one of which we had to disqualify due to a failure to follow directions) we had 7 subjects in each of the condition MSA and WOT, and 8 for the hybrid prototype. Each subject had to rate half of their assigned sites without the help of a tool, and half of the sites with the help of a tool, as assigned to them at the start of the study. The selection of which sites they got to rate with or without a tool was done randomly. This was done to check that the presence of a tool influenced subjects’ actions.

Subjects were given a brief tutorial in the tool they were assigned, and were allowed to see the rating for a demo site and familiarize themselves with how the tool worked, and what the ratings meant before stating their test.

3.4.5 QUESTIONNAIRE

Subjects were asked to rate each site using a simple questionnaire. The questionnaire had 2 parts, the first focusing on rating the trustworthiness and usefulness of the website. Participants were asked to answer these questions using a 5-point Likert scale ranging from “Excellent” to “Very poor”. They were asked to enter their confidence level with respect to their ratings using a Likert scale ranging from “High confidence” to “Low confidence”. Figure 3.6 shows a screen-shot of this part of the

questionnaire. The second half of the questionnaire asked subjects to justify their ratings with a number of binary questions about the suspected content, business practices, and/or suspected technical vulnerabilities.

How trustworthy is this website?

Excellent Good Medium Poor Very poor



How confident are you in your rating ?

What are the factors affecting your rating ?

How useful is this website?

Excellent Good Medium Poor Very poor



How confident are you in your rating ?

What are the factors affecting your rating ?

Figure 3.6. Questionnaire used in the experiment.

3.4.6 POST-EXPERIMENT SURVEY

After reviewing the 20 websites, participants were asked to fill out another short survey about their experience. This survey focused on questions like

1. How easy was the experiment?
2. How they decide what sites were trustworthy?

3. How much confidence they have in the recommendations of the tool used?
4. What affected their confidence rating?
5. Recommendations to improve these tools?

3.4.7 DEMOGRAPHICS

Subjects were recruited from the student population at Oregon State University. Of the 23 subjects recruited for the experiment, one had to be excluded because he refused to follow directions, or even complete the tutorial properly. Recruitment was done through flyers and emails, targeting student groups, dorms, and social gathering places on campus. Subjects were offered a \$10 USD compensation for participating.

Of the 22 subjects for that we got data, only 4 were women. Most subjects had ample computer experience (average of 10 years), and ranged in age from 18 to 33 years, with the bulk of subjects being in the 22-28-age range. We were successful in attracting subjects with a wide range of backgrounds, with a total of 13 majors represented in our sample.

All procedures were documented and approved by IRB. All subjects volunteered to participate and signed informed consent forms.

3.5 RESULTS

3.5.1 RQ1: Are users more heavily influenced by machine ratings, or user ratings? (Regardless of which is more correct)

Because subjects typically only had access to one rating system, and because they lacked the technical knowledge to determine whether a rating was correct, we focused on adherence in this study. That means that we do not care if ratings were correct, but rather whether subjects chose to believe in the recommendation.

To answer this first question, we only looked at data from the WOT and MSA conditions. The confidence level of users while rating whether a website is trustworthy or not varied significantly between the WOT group and the MSA. We broke down our analysis for overall ratings as well as by “good” and “bad” websites. While subjects

followed recommendations, we found that there was a statistically significant difference in participant's confidence level (see Figure 3.7).

Confidence levels are important indicators because they tell us how likely users are to follow recommendations or override the system under less controlled conditions. While we saw almost perfect adherence to recommendations in our experiment, we would not expect the same in a real-world setting; subjects knew the purpose of the experiment, and they were therefore more likely to obey. For the overall case (both good and bad sites, we found a statistically significant difference in the confidence levels, with WOT users being more confident of their ratings (*ANOVA* $F=7.24$, $df=1$, $p = 0.0087$). In the case of good sites, we found no significant difference. This was not unexpected, as there is clearly a ceiling effect taking place.

There was a significant difference in the confidence levels of MSA users and WOT users when rating bad websites (*ANOVA* $F=7.56$, $df=1$, $p=0.0090$). MSA users clearly heeded the warnings issued by MSA, but were not terribly confident in doing so for "bad" websites. This seems to indicate that subjects were either unsure or uncomfortable with the basis for flagging a "bad" website, or the way the MSA system communicates that information. Though subjects almost unanimously followed recommendations in our experiment, subjects were a lot more confident about recommendations when they came from a community-based service.

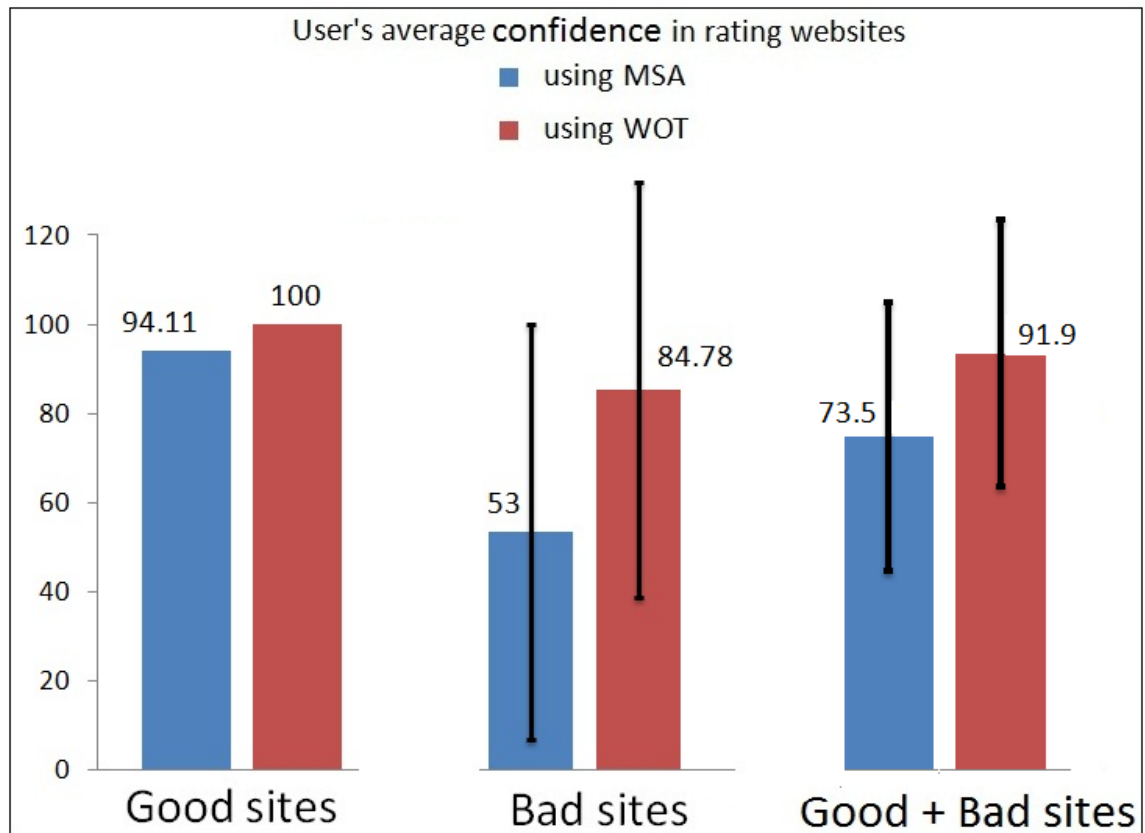


Figure 3.7 Users average confidence in rating the trustworthiness of good and bad sites. Confidence intervals shown for statistically significant cases.

3.5.2 RQ2: Does combining the two types of ratings add value to users?

Subjects in the hybrid prototype group were more confident in their ratings than the participants in either of the two other groups (see Figure 3.8) for the average case. The differences are statistically significant ($ANOVA F=7.48, df=2, p=0.000893$). Though the differences between the WOT and hybrid conditions is not significant, users confidence levels differ significantly between the WOT and MSA conditions ($ANOVA F=7.24, df=1$ and $p=0.0087$) and between the MSA and the Hybrid condition ($ANOVA F=9.33, df=1$ and $p \text{ value} = 0.0031$). It is therefore clear that users react positively to additional sources of information, at least when that information is internally consistent.

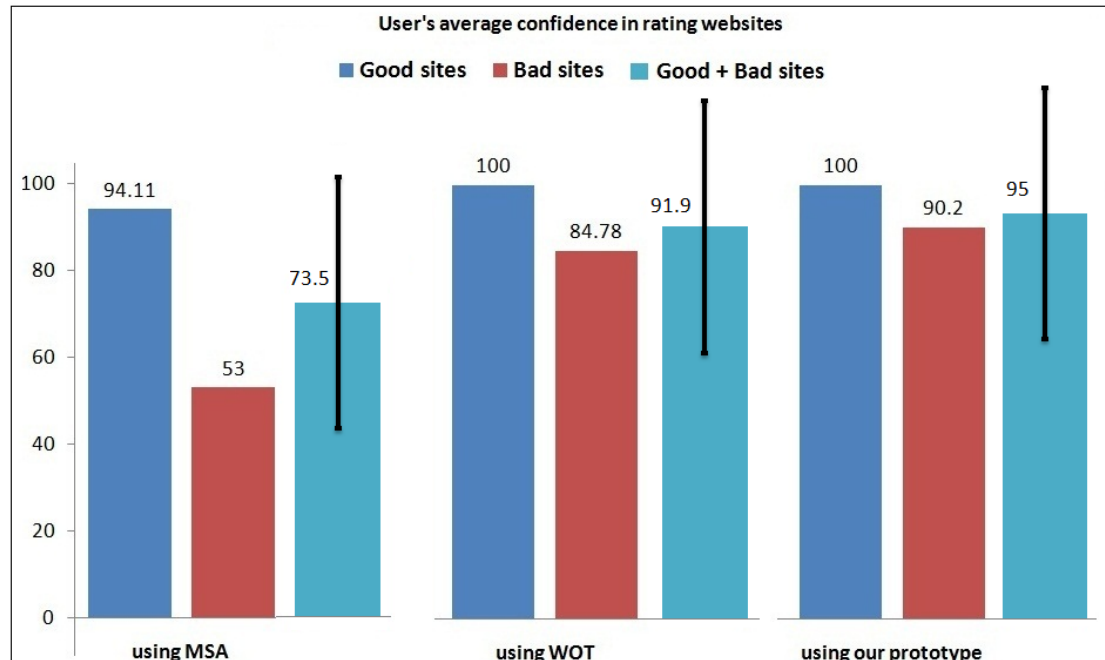


Figure 3.8 Difference in confidence levels of participants from different groups.

3.5.3 RQ3: When presented with conflicting ratings, how do users act?

Again, turning to the participants in the hybrid condition (the only ones aware of a conflict), participants' almost uniformly miss-trusted websites with conflicting ratings (90% of the participants rated these sites as "untrustworthy"). This means, that at least under the artificial conditions created in a controlled lab experiment, users are more comfortable taking a risk-averse strategy. What is even more interesting is looking at how these conflicting websites stack up against uniformly "bad" websites in the different conditions (see Figure 3.9). There is a statistical difference between users confidence in rating bad sites using MSA and rating conflicting sites using our prototype (ANOVA $F=22.65$, $df=1$, $p \text{ value} < 0.001$).

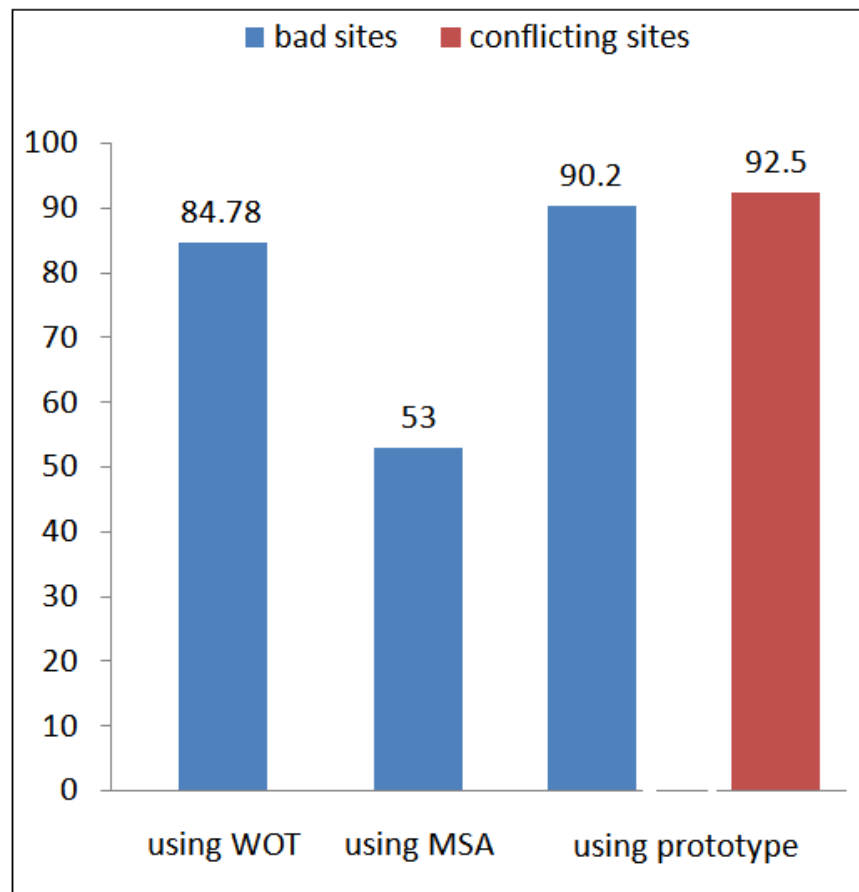


Figure 3.9 Increase in confidence level of participants while rating trustworthiness of conflicting websites

3.5.4 SURVEY RESULTS

In the post-experiment questionnaire, we explored how people make decisions about the trustworthiness of sites and the recommendations given by different safe-surfing tools. More than 75% of the participants stated that their confidence level was routinely affected by at least one of the following:

1. Previous knowledge or familiarity with a website
2. Brand-name recognition of safe-surfing tools
3. Popularity of the website.

While subjects in our study did not have access to statistics about the popularity of websites, we did not attempt to in any way disguise the names of the tools used in our experiment. McAfee is a household name for computer security, much more so than Web of Trust, and it is therefore interesting to see that subjects were less confident when relying on McAfee SiteAdvisor to make decisions.

While deciding on the trustworthiness of a website, participants said they also relied on:

1. Visual Appearance of the website
2. What information the site collects (explicitly) from them

These are relatively naïve approaches to judging the trustworthiness of websites, but handily beat out more technical explanations, which few end-users understand or have access to. This shows how important it is to promote the adoption of safe-surfing tools that inspire high confidence in users, thus decreasing the likelihood that they would override or ignore the recommendations given.

3.6 DISCUSSION

When we compare the WOT and MSA conditions, users confidence in deciding whether a website is trustworthy was higher for users of Web of Trust than for those using McAfee SiteAdvisor, especially in the case of “bad” websites. Though they followed the recommendations given by MSA, and the McAfee brand as an anti-virus company, most subjects did not have high confidence in their recommendation to avoid sites. This is troubling because we believe this is a strong indicator that subjects will at least in some cases override or ignore these recommendations in a real-world setting.

As Ayyavu and Jensen showed in (Ayyavu & Jensen 2011), MSA tends to be more conservative when it comes to giving a “bad” rating to a website, which means that only the most egregiously bad sites get such a rating. When quizzed about their reluctance to trust the ratings, some participants reported that their confidence were

affected by a lack of understanding of the technical foundations and the in-depth security reports given by MSA. This is a tough challenge for MSA-like tools to overcome. Users seem to indicate that unless they can understand the basis for a negative review, they don't have high confidence in it, yet the basis for such recommendations are highly technical and inaccessible to most users.

One of the most important findings in our study was the significant increase in confidence of subjects rating websites with the help of the hybrid tool. Previous research (Angai et. Al. 2010) had found that users confidence level increased when data was aggregated across different heuristic-based services, but no research had been done on the combination of community and heuristic-based data. Part of the reason for this increase may lie in that many Internet users do not understand technical security discussions and that when looking for justifications for why a site should be avoided, human-generated and experiential justifications carry a lot more weight.

In our prototype, the basic result contained simple description from MSA like "The site is safe" or "The site is risky", and high quality words extracted from WOT user comments. This was a very condensed view of what the community-based tools were reporting about the sites, but it seemed to be sufficient for users to make high-confidence decisions. Participants reported that this representation was easy to digest and many were satisfied and not compelled to drill down the UI to see full reports. This does not mean that this UI could not be further improved for more positive results.

The inclusion of all kinds of websites with varying ratings, like "good", "bad", "may be risky", as well as sites with conflicting ratings in our sample set helped us do a more nuanced study of user reactions and attitudes. More of the "good, safe" websites in our study had been seen previously by our subjects, so this may have affected confidence ratings for those sites.

One significant necessary future step is to study the usability of our prototype, and whether the amount of information presented, and the way it is presented could be

improved. We would like to deliver as much information to end-users as is necessary to raise confidence, and thus adherence to our recommendations, but to do so in a way that will not distract users from their primary surfing task. How to strike that balance is one of the perennial challenges for designers of safe-surfing tools; being over-demanding has been shown to be as ineffective a strategy as being too subtle. We would also like to study more specialized types of websites to determine whether attitudes and confidence ratings are more likely to vary (for instance between a “for work” site vs. a leisure site). It would also be interesting to examine how confidence is affected as the amount of available community-based data decreases, a common problem for many less-popular sites.

3.7 CONCLUSION

Discrepancies exist between different online safe surfing services in rating websites. This user study of online safe surfing tools found that 1) Users’ confidence in deciding whether a website is trustworthy or not is higher when using community-based rating services like Web Of Trust than heuristics-based services like SiteAdvisor. 2) When presented with websites with conflicting ratings from MSA and WOT, users defaulted to erring on the side of caution, and did so with high levels of confidence. 3) When users were presented with combined results from WOT and MSA, their confidence level increases when deciding the trustworthiness of websites. Factors like “brand-name” and “previous experience” affected their confidence.

3.8 ACKNOWLEDGMENTS

We wish to thank our colleagues in the HCI group at the School of EECS, Oregon State University for their support and help in preparing the paper. We would like to thank our study participants. We’d also like to thank Web of Trust and it’s users for giving us access to their data for this study.

3.9 REFERENCES

Angai, F. et al., 2010. Analysis on the Effectiveness of Safe Browsing Services. *Computer Engineering*. Available at:

<http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:Analysis+on+the+Effectiveness+of+Safe+Browsing+Services#0> [Accessed August 29, 2011].

Ayyavu, P. & Jensen, C., 2011. Integrating user feedback with heuristic security and privacy management systems. In *Proceedings of the 2011 annual conference on Human factors in computing systems - CHI '11*. New York, New York, USA: ACM Press, p. 2305. Available at: <http://portal.acm.org/citation.cfm?id=1979281> [Accessed August 1, 2011].

Chen, M. & Singh, J.P., 2001. Computing and using reputations for internet ratings. In *Proceedings of the 3rd ACM conference on Electronic Commerce - EC '01*. New York, New York, USA: ACM Press, pp. 154-162. Available at: <http://portal.acm.org/citation.cfm?doid=501158.501175> [Accessed August 27, 2011].

Colwill, C., 2009. Human factors in information security: The insider threat – Who can you trust these days? *Information Security Technical Report*, 14(4), pp.186-196. Available at: <http://linkinghub.elsevier.com/retrieve/pii/S1363412710000051>.

Dourish, P. et al., 2004. Security in the wild: user strategies for managing security as an everyday, practical problem. *Personal and Ubiquitous Computing*, 8(6), pp.391-401. Available at: <http://www.springerlink.com/index/10.1007/s00779-004-0308-5>.

Goecks, J., Edwards, W.K. & Mynatt, E.D., 2009. Challenges in supporting end-user privacy and security management with social navigation. *Proceedings of the 5th Symposium on Usable Privacy and Security - SOUPS '09*, p.1. Available at: <http://portal.acm.org/citation.cfm?doid=1572532.1572539>.

Maurer, M.-emanuel, Luca, A.D. & Kempe, S., 2011. Using Data Type Based Security Alert Dialogs to Raise Online Security Awareness Categories and Subject Descriptors. *Symposium on Usable Privacy and Security (SOUPS) 2011, July 20–22, 2011, Pittsburgh, PA USA*.

Proctor, R., Schultz, E. & Vu, K., 2009. Human factors in information security and privacy. *Research on Information Security*. Available at: <http://www.igi-global.com/viewtitlesample.aspx?id=20669> [Accessed July 25, 2011].

Tsai, J. et al., 2009. The impact of privacy indicators on search engine browsing patterns. *Under review*, p.1. Available at: <http://portal.acm.org/citation.cfm?doid=1572532.1572568> [Accessed August 28, 2011].

Wu, M., Miller, R.C. & Garfinkel, S.L., 2006. Do security toolbars actually prevent phishing attacks? In *Proceedings of the SIGCHI conference on Human Factors in computing systems*. New York, New York, USA: ACM, pp. 601–610. Available at: <http://portal.acm.org/citation.cfm?doid=1124772.1124863> [Accessed August 29, 2011].

Alexa., www.alexa.com

Google online security blog., <http://googleonlinesecurity.blogspot.com/2007/11/help-us-fill-in-gaps.html>

Kaspersky., www.usa.kaspersky.com

Krebsonsecurity., <http://krebsonsecurity.com/2010/04/hiding-from-anti-malware-search-bots/>

LinkScanner., www.linkscanner.avg.com

McAfee SiteAdvisor., www.siteadvisor.com

Norton SafeWeb., www.safeweb.com

NCSA Norton Online Safety Study.,
<http://staysafeonline.mediaroom.com/index.php?s=67&item=57>

Web Of Trust., www.mywot.com

4 CONCLUSION

Discrepancies exist between different online safe surfing services in rating websites. Users were found to be more discriminating than heuristic-based systems, at least for the sample used in first our study. This may be an indicator of a higher false positive rate with community-based systems, or it may be an evidence that users have access to more meaningful information (again, at least for this sample), or that user are more risk-averse than heuristic-based systems give them credit for. User ratings and heuristic-based systems are both relatively stable, and do not change significantly over short periods of time (though user ratings change more rapidly than heuristic-based reviews). Given the low rate of change, we cannot determine how much of the change is due to genuine change in the site policies, and how much is due to shifting opinion.

Our results from first manuscript also show that information abstracted from user comments may add value in terms of additional data to the evaluation and communication of risk to users by automatically analyzing large numbers of user comments. This may help people in at-a-glance digesting large volumes of user feedback easily. Previous research has shown that users are influenced by user feedback; this abstracted form of that information in form of dictionaries may add sufficient nuance to the heuristic machine based testing systems like McAfee SiteAdvisor and Norton SafeWeb.

The user study done in the second manuscript found that

- 1) Users' confidence in deciding whether a website is trustworthy or not is higher when using community-based rating services like Web Of Trust than heuristics-based services like SiteAdvisor.
- 2) When presented with websites with conflicting ratings from MSA and WOT, users defaulted to erring on the side of caution, and did so with high levels of confidence.
- 3) When users were presented with combined results from WOT and MSA, their confidence level increases when deciding the trustworthiness of websites.

5 BIBLIOGRAPHY

- Amatriain, X., Pujol, J. M., Tintarev, N., & Oliver, N. (2009). Rate it again. *Proceedings of the third ACM conference on Recommender systems - RecSys '09* (p. 173). New York, New York, USA: ACM Press. doi:10.1145/1639714.1639744
- Angai, F., Ching, C., Ng, I., & Smith, C. (2010). Analysis on the Effectiveness of Safe Browsing Services. *Computer Engineering*. Retrieved from <http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:Analysis+on+the+Effectiveness+of+Safe+Browsing+Services#0>
- Ayyavu, P., & Jensen, C. (2011). Integrating user feedback with heuristic security and privacy management systems. *Proceedings of the 2011 annual conference on Human factors in computing systems - CHI '11* (p. 2305). New York, New York, USA: ACM Press. doi:10.1145/1978942.1979281
- Bilgic, M., & Mooney, R. J. (2005). Explaining recommendations: Satisfaction vs. Promotion. *IUI05 Beyond Personalization Workshop*, 13-18. Citeseer. Retrieved from <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.107.488&rep=rep1&type=pdf#page=13>
- Chen, M., & Singh, J. P. (2001). Computing and using reputations for internet ratings. *Proceedings of the 3rd ACM conference on Electronic Commerce - EC '01* (pp. 154-162). New York, New York, USA: ACM Press. doi:10.1145/501158.501175
- Colwill, C. (2009). Human factors in information security: The insider threat – Who can you trust these days? *Information Security Technical Report*, 14(4), 186-196. doi:10.1016/j.istr.2010.04.004
- Dellarocas, C. (2000). Immunizing online reputation reporting systems against unfair ratings and discriminatory behavior. *Proceedings of the 2nd ACM conference on Electronic commerce* (p. 157). New York, New York, USA: ACM. doi:10.1145/352871.352889
- Dourish, P., Grinter, R. E., Delgado de la Flor, J., & Joseph, M. (2004). Security in the wild: user strategies for managing security as an everyday, practical problem. *Personal and Ubiquitous Computing*, 8(6), 391-401. doi:10.1007/s00779-004-0308-5
- Edwards, W. K., Poole, E. S., & Stoll, J. (2008). Security automation considered harmful? *Proceedings of the 2007 Workshop on New Security Paradigms - NSPW '07*, 33. New York, New York, USA: ACM Press. doi:10.1145/1600176.1600182

Egelman, S., King, J., Miller, R. C., Ragouzis, N., & Shehan, E. (2007). Security user studies. *CHI '07 extended abstracts on Human factors in computing systems - CHI '07* (p. 2833). New York, New York, USA: ACM Press.

doi:10.1145/1240866.1241089

Goecks, J., Edwards, W. K., & Mynatt, E. D. (2009). Challenges in supporting end-user privacy and security management with social navigation. *Proceedings of the 5th Symposium on Usable Privacy and Security - SOUPS '09*, 1. New York, New York, USA: ACM Press. doi:10.1145/1572532.1572539

Gross, J. B., & Rosson, M. B. (2007). Looking for trouble: understanding end-user security management. *Proceedings of the 2007 symposium on Computer human interaction for the management of information technology, March* (pp. 30–31).

Retrieved from

http://www.cc.gatech.edu/classes/AY2008/cs4235b_fall/Group1/UnderstandingEndUserSecurityMgt.pdf

Herlocker, J. L., Konstan, J. a, & Riedl, J. (2000). Explaining collaborative filtering recommendations. *Proceedings of the 2000 ACM conference on Computer supported cooperative work - CSCW '00*, 241-250. New York, New York, USA: ACM Press. doi:10.1145/358916.358995

Hu, N., Pavlou, P. A., & Zhang, J. (2006). Can online reviews reveal a product's true quality?: empirical findings and analytical modeling of Online word-of-mouth communication. *Proceedings of the 7th ACM conference on Electronic commerce* (pp. 324–330). New York, New York, USA: ACM. doi:10.1145/1134707.1134743

Jin, R., & Si, L. (2004). A study of methods for normalizing user ratings in collaborative filtering. *Proceedings of the 27th annual international conference on Research and development in information retrieval - SIGIR '04* (p. 568). New York, New York, USA: ACM Press. doi:10.1145/1008992.1009124

Maurer, M.-emanuel, Luca, A. D., & Kempe, S. (2011). Using Data Type Based Security Alert Dialogs to Raise Online Security Awareness Categories and Subject Descriptors. *Symposium on Usable Privacy and Security (SOUPS) 2011, July 20–22, 2011, Pittsburgh, PA USA*.

McNee, S. M., Kapoor, N., & Konstan, J. A. (2006). Don't look stupid. *Proceedings of the 2006 20th anniversary conference on Computer supported cooperative work - CSCW '06* (p. 171). New York, New York, USA: ACM Press.

doi:10.1145/1180875.1180903

Proctor, R., Schultz, E., & Vu, K. (2009). Human factors in information security and privacy. *Research on Information Security*. Retrieved from <http://www.igi-global.com/viewtitlesample.aspx?id=20669>

Sen, S., Lam, S. K., Rashid, A. M., Cosley, D., Frankowski, D., Osterhouse, J., Harper, F. M., et al. (2006). Tagging, Communities, Vocabulary, Evolution. *Proceedings of the 2006 20th anniversary conference on Computer supported cooperative work - CSCW '06* (p. 181). New York, New York, USA: ACM Press. doi:10.1145/1180875.1180904

Svensson, M., Höök, K., Laaksolahti, J., & Waern, A. (2001). Social navigation of food recipes. *Proceedings of the SIGCHI conference on Human factors in computing systems - CHI '01* (pp. 341-348). New York, New York, USA: ACM Press. doi:10.1145/365024.365130

Tsai, J., Egelman, S., Cranor, L., & Acquisti, A. (2009). The impact of privacy indicators on search engine browsing patterns. *Under review*, 1. New York, New York, USA: ACM Press. doi:10.1145/1572532.1572568

Vig, J., Sen, S., & Riedl, J. (2008). Tagsplanations. *Proceedings of the 13th international conference on Intelligent user interfaces - IUI '09* (p. 47). New York, New York, USA: ACM Press. doi:10.1145/1502650.1502661

Alexa., www.alexa.com

AlpineLinux., <http://www.alpinelinux.org/wiki>

AOL explorer., <http://www.aol.com/>

Artviper., <http://www.artviper.net/texttagcloud/>

Avira., <http://www.avira.com/en/pages/index.php>

Brightmail.,
<http://www.symantec.com/business/products/family.jsp?familyid=brightmail>

Google online security blog., [http:// googleonlinesecurity. blogspot.com/2007/11/help-us-fill-in-gaps.html](http://googleonlinesecurity.blogspot.com/2007/11/help-us-fill-in-gaps.html)

Ipfire., <http://www.ipfire.org/en/index>

Kaspersky., www.usa.kaspersky.com

Krebsonsecurity., <http://krebsonsecurity.com/2010/04/hiding-from-anti-malware-search-bots/>

LinkScanner., www.linkscanner.avg.com

McAfee., <http://www.mcafee.com/us/>

McAfee SiteAdvisor., www.siteadvisor.com

McAfee SpamKiller., <http://us.mcafee.com/root/product.asp?productid=msk>

Mozilla Firefox anti-phishing., <http://www.mozilla.com/en-US/firefox/phishing-protection/>

NCSA Norton Online Safety Study.,
<http://staysafeonline.mediaroom.com/index.php?s=67&item=57>

Norton., <http://us.norton.com/index.jsp>

Norton SafeWeb., www.safeweb.com

Spamassassin., <http://spamassassin.apache.org/>

SpyBot., <http://www.safer-networking.org/en/index.html>

Tagcrowd., <http://tagcrowd.com>

Untangle., <http://www.untangle.com/>

Web Of Trust., www.mywot.com

Webroot popup washer., http://www.webroot.com/En_US/consumer-products-spysweeper.html

Wordle., <http://www.wordle.net/create>