Re-Evaluating the Wisdom of Crowds in Assessing Web Security

Pern Hui Chia and Svein Johan Knapskog
Q2S* NTNU, Trondheim, Norway

Abstract. We examine the outcomes of the Web of Trust (WOT), a user-based system for assessing web security and find that it is more comprehensive than three automated services in identifying 'bad' domains. Similarly to PhishTank, the participation patterns in WOT are skewed; however, WOT has implemented a number of measures to mitigate the risks of exploitation. In addition, a large percentage of its current user inputs are found to be based on objective and verifiable evaluation factors. We also confirm that users are concerned not only about malware and phishing. Online risks such as scams, illegal pharmacies and misuse of personal information are regularly brought up by the users. Such risks are not evaluated by the automated services, highlighting the potential benefits of user inputs. We also find a lack of sharing among the vendors of the automated services. We analyze the strengths and potential weaknesses of WOT and put forward suggestions for improvement.

Keywords: Web Security, Crowd Sourcing, Web of Trust

1 Introduction

Security on the web remains a challenging issue today. Losses due to online banking fraud in the UK alone stood at 59.7 million pound in 2009, with more than 51,000 phishing incidents recorded (up 16% from 2008) [22]. Provos et al. [8] found that over 3 million malicious websites initiate drive-by downloads and about 1.3% of all queries to the Google search engine get at least one URL labeled as malicious in the results page. Meanwhile, Zhuge et al. [13] found that 1.49% of Chinese websites, sampled using popular keywords on Baidu and Google search engines, are malicious.

There is also a lack of efficient services to identify sites that are not outright malicious, but are 'bad' in the sense that they try to *trick* or *offend* users in many aspects, such as scams, deceptive information gathering and misuse of user data. Several fraudulent online activities such as money-mule recruitment and illegal online pharmacies seem to have fallen out of the specific responsibilities or interests of the authorities and security vendors. While banking-phishing sites have a short life-time between 4 to 96 hours, the average life-time was found to be 2 weeks for mule-recruitment and 2 months for online pharmacy websites [6]. Problems with the potentially inappropriate content may also be serious. The Internet adult industry is assumed to be worth more than 97 billion USD and adult sites regularly rank among the

* Centre of Quantifiable Quality of Service in Communication Systems (Q2S), Centre of Excellence appointed by the Research Council of Norway, is funded by the Research Council, Norwegian University of Science and Technology (NTNU) and UNINETT. http://www.q2s.ntnu.no

top 50 visited websites worldwide [11]. While it is a personal judgment whether adult content in general is inappropriate to the viewers, Wondracek et al. [11] confirmed that adult sites are plagued with issues such as malware and script-based attacks and they frequently use aggressive or inappropriate marketing methods.

Online certification issuers, such as BBBOnline.org and TRUSTe.com strive to distinguish 'good' sites from the 'bad' ones. This is, however, not a straightforward task. Most websites are not entirely good or bad; there is sometimes a conflict of interest. Problems, such as adverse-selection [4] have been observed when certification issuers adopt lax requirements to certify sites belonging to this 'gray' category.

1.1 The wisdom of crowds for security

Despite the fact that user feedback is highly sought after by various online services, sourcing for user inputs for security purposes is still an early concept. A typical argument against the idea is on the limited ability of ordinary users in providing reliable assessment for security. There is a general uneasiness involved in relying on the ordinary users for this seemingly serious task. Indeed, different from the general quality assessment, an incorrect security evaluation can cause harm to the users. Yet, this should not preclude the feasibility of collating user inputs for security purposes. Surowiecki gives multiple real life examples where inputs by non-experts collectively performed better than experts' advices when handling complex and serious tasks [10].

PhishTank [20] and Web of Trust (WOT) [23] are two of the few existing systems that employ the wisdom of crowds to improve web security. PhishTank solicits for user reporting and voting against sites suspected to be phishes, while WOT collects public opinions on the trustworthiness, reliability, privacy and child-safety aspects of websites. Both services operate on the principle that a collective decision by ordinary users, when harnessed wisely, can yield good outcomes as errors made by individuals tend to cancel out each other. There is also the advantage of scale to cater for a large volume of items needing an evaluation. Whether a user-based system can be helpful in the security 'arms race', however, depends on a multitude of factors.

In this work, we measure the reliability of WOT against three automated services by well-known vendors, namely, McAfee's SiteAdvisor [17], Norton's Safe Web [19] and Google's Safe Browsing Diagnostic Page [15]. We also investigate the characteristics of user participation in WOT. Our findings can be summarized as follows:

- Only a few sites are commonly classified as 'bad' by the prominent security vendors, indicating a lack of data sharing.
- The coverage of WOT for general sites is low compared to the automated services.
- WOT's coverage increases when considering only domains registered in regions where active user participation is currently observed.
- WOT is more comprehensive in identifying the 'bad' domains.
- False negatives (in identifying 'bad' domains) are mostly labeled as 'unknown' by WOT, while they are often wrongly classified as 'good' by the other services.
- The contribution ratios in WOT are skewed with the comment contribution following a power law distribution.
- WOT has put in place a number of mitigation measures against manipulation.
- A majority of the current user inputs in WOT is based on objective evaluation criteria and hence verifiable.
- User concerns on web security are not limited to malware and phishing.

2 Related work

The Wisdom of Crowds should be a familiar notion by now. Surowiecki [10] outlines 4 conditions for a wise crowd to outperform a few experts. Firstly, the crowd members should be diverse (not homogenous). They should also have independent thought processes to avoid mere information cascade. Thirdly, they should be organized in a decentralized structure to tap into local knowledge and specialization. Lastly, a good aggregation strategy is needed to collate the inputs from the individuals.

In a closely related work [5], Moore and Clayton evaluated the reliability and contribution patterns in PhishTank. They find that it is particularly susceptible to manipulation since the participation ratio in PhishTank is highly skewed (following a power-law distribution). Compared to a commercial phishing report, they also find that PhishTank is slightly less comprehensive and slower in reaching a decision. Our work is inspired by theirs, combined with the curiosity of why PhishTank has become quite widely adopted despite the criticisms.

While a number of studies look at the efficiency of various blacklists or tools for the specific issue of phishing (e.g., [9,12]), there is little effort in evaluating the tools for web security as a whole. To our knowledge, our study is the first to evaluate the reliability of WOT, comparing it with three automated alternatives.

3 The Web of Trust (WOT)

WOT is a reputation system that aggregates user inputs into global ratings about sites under evaluation. It takes the form of a browser add-on and a centralized database [23]. User inputs and the evaluation outcomes on WOT are structured around four aspects, namely trustworthiness, vendor reliability, privacy and child-safety, with ratings ranging from very poor (0-19), poor (20-39), unsatisfactory (40-59) to good (60-79) and excellent (80-100). WOT describes the four aspects as follows:

- **Trustworthiness** (Tr): whether a site can be trusted, is safe to use, and delivers what it promises. A 'poor' rating may indicate scams or other identified risks e.g., identity theft, credit card fraud, phishing, viruses, adware or spyware. A rating of 'unsatisfactory' indicates that the site may contain annoying advertisements, excessive pop-ups or elements that make browsers crash.
- **Vendor Reliability** (Vr): whether a site is safe for business transactions. A 'poor' rating indicates a possible scam or a bad shopping experience.
- **Privacy** (Pr): whether a site has a privacy policy that protects information regarded as sensitive by the user e.g., whether it has opt-in privacy options or gives users the means to determine what can be made public and what should remain private. A 'poor' rating indicates concern that user data may be sold to third parties, be stored indefinitely or be turned over to law enforcement without a warrant or user consent.
- **Child-Safety** (Cs): whether a site contains material such as adult content, violence, vulgar or hateful language, or content that encourages dangerous or illegal activities

Besides user inputs, WOT also receives inputs from a list of trusted third parties. For example, WOT receives blacklists of phishing, spamming and malware-infesting sites from PhishTank [20], SpamCop [21] and LegitScript [16], respectively.

WOT applies Bayesian inference to weigh user inputs differently based on the reliability of individual contributors, judging from their past rating behaviors. Individual

user ratings are kept private to the contributors. Neither is the actual formula used in the computation publicly available. WOT argues that the hidden formula and individual inputs, plus the Bayesian inference rule, help to mitigate typical threats facing reputation and recommender systems such as a Sybil attack in which dishonest users register multiple identities to attempt influencing the outcomes. The aggregate rating is accompanied by a confidence level (0-100) rather than the count of the individual ratings. The developers argue that the confidence level is more appropriate as it takes into account both the number of inputs and the probable reliability of the contributors. WOT requires a minimal confidence level before publishing the aggregate rating.

Besides numerical ratings, users can also comment about the sites under evaluation. To give a comment, they must first register themselves on WOT's website. Non-registered users can only rate a site via the add-on, which gives a unique pseudonym to every WOT user. Users specify one out of 17 categories (as shown in Figure 5) which best describes their comment. Comments do not count towards the aggregate ratings. Unlike the individual ratings, they are publicly accessible on the website.

The website also supports a number of community features such as a personal page per registered user, a page for each evaluated site, messaging channels between users, a discussion forum, a wiki, as well as mechanisms to make explicit public requests to evaluate certain sites. WOT also presents some rudimentary statistics on the website, such as the total numbers of comments, ratings and dangerous sites.

The browser add-on allows a user to conveniently rate the sites he visits, besides signaling the reputation of different URI links found on web pages and warning the user as he navigates to sites that have been given a credible 'poor' rating (i.e., rating $<$ 40 and confidence level $\geq$8) in either aspect of trustworthiness, vendor reliability or privacy. The child-safety rating is ignored by default but the settings for risk signaling and warning are configurable. The add-on's implementation is open-source.

4 Data Collection

To evaluate the reliability of WOT, we compared its aggregate ratings with the outcomes provided in the querying pages of the three automated services, as identified in Section 1.1. We collected the outcomes on 20,000 sites randomly selected from the top million frequently visited sites, published by Alexa [14]. This gives us a realistic evaluation scenario in which we measure the reliability of WOT for sites that users normally visit. For each site, our program queried the assessment report from each service, parsed and stored the result (referred to as dataset-I). The querying process took place from the end of July to mid of August 2010. We have confirmed with the developers that WOT does not take inputs from any of the three automated services.

In addition to the above, we have requested and obtained two more datasets (hereafter referred to as dataset-II and dataset-III) from the developers. Dataset-II contains the contribution level of 50,000 randomly selected users out of >1.5 million registered users at the time of data collection. It describes the total numbers of ratings and comments which have been contributed by a user, as well as his date of registration. Dataset-III consists of 485,478 comments randomly selected from >8 million at that time. Besides the comment text, it includes the date of writing and a category chosen by the contributor to best describe the comment. Both dataset-II and III contain only information that are publicly accessible for all who have logged in to the WOT's website.

The comments in dataset-III evaluate a total of 412,357 unique sites. To study the users' commenting behavior, we downloaded also the aggregate ratings of all these 412k sites using the public query API of WOT [23].

5 Analysis

We started by studying the characteristics of the automated services:

- **McAfee's SiteAdvisor** [17] evaluates a site based on a variety of proprietary and automated tests on aspects such as downloads, browser exploits, e-mail, phishing, annoyance factors (e.g., pop-ups and cookies) and affiliations with other sites. SiteAdvisor also receives inputs from Trusted Source [18] which evaluates aspects such as website behavior, traffic and linking patterns, as well as site registration and hosting. Among others, it helps SiteAdvisor to identify spamming and phishing sites. SiteAdvisor allows users to comment on a particular site; however, the comments are not considered for the assessment outcomes.
- **Norton's Safe Web** [19] tests if a site imposes threats such as drive-by downloads, phishing attacks, spyware, Trojans, worms, viruses, suspicious browser changes, joke programs and identity theft. It collects also user ratings and comments, but like SiteAdvisor, user inputs do not count towards the overall rating.
- **Google's Safe Browsing Diagnostic Page** [15] warns about sites that have been the hosts or intermediaries which download and install (malicious) software on a user's device without consent. It should be noted that warnings about phishing activities are not included in the diagnostic page. Phishing reports may only be accessible via the Safe Browsing API. We note that this should not affect our results as we do not expect that the frequently visited sites (used in our evaluation) to be phishes.

Table 1. Aligning the different classification classes.

	WOT	SiteAdvisor	Safe Browsing DP	Safe Web
Good	$Tr \geq 60$, and without a credible warning in Vr or Pr	Green: Very low or no risk issues found	The site is not currently listed as suspicious and Google has visited it in the past 90 days.	Safe
Caution	$60 > Tr \geq 40$, and without a credible warning in Vr or Pr	Yellow: Minor risk issues found	The site is not currently listed as suspicious but part of this site was listed for suspicious activity in the past 90 days.	Caution
Bad	$Tr < 40$, or there is a credible warning in Vr or Pr	Red: Serious risk issues found	Site is listed as suspicious.	Warning
Unknown	Tr has no rating, and without a credible warning in Vr or Pr	Gray: Not yet rated	This site is not currently listed as suspicious but Google has not visited it the past 90 days.	Untested

To enable a fair comparison, we mapped the evaluation outcomes of the respective services into 4 classes: good, caution, bad and unknown, as shown in Table 1. We classified WOT's ratings based on the default strategy used by its browser add-on for risk signaling, which regards Trustworthiness (Tr) as the most important evaluation

aspect as it often covers the scopes of Vendor Reliability (Vr) and Privacy (Pr)[1]. A site is considered 'good' if its Tr rating is ≥ 60 without any credible warning in Vr or Pr (i.e., rating < 40 and confidence level ≥ 8). We did not consider child-safety in the classification as it is ignored by the browser add-on in the default settings. Neither is content-appropriateness evaluated by the automated services.

5.1 The reliability of WOT

Table 2. Coverage and the percentage of different evaluation outcomes.

	Coverage (%)	Evaluation outcomes (%)		
		Bad	Caution	Good
WOT	51.23[2]	3.16	2.15	45.93
SiteAdvisor	87.84	1.48	0.47	85.90
Safe Browsing DP	55.65[3]	0.13	1.63	53.90
Safe Web	68.09	0.51	0.38	67.21

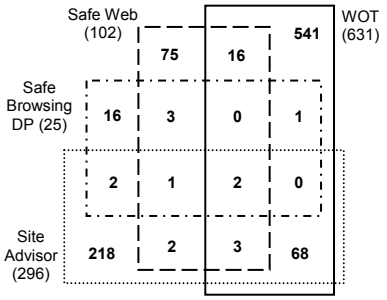

Figure 1. Venn diagram shows the divergence in the classification of bad sites. Out of 948 that have been marked as 'bad' by any service provider, only 2 receive the same verdict from all services, while only 98 sites are classified as 'bad' by more than one services.

We first evaluated the coverage of individual services (see Table 2). Coverage is defined as the ratio of evaluated sites (i.e., not being classified as 'unknown') over the number of total sites. SiteAdvisor has the highest coverage while WOT scores the lowest among the 4 services. This can be attributed to the fact that decisions in WOT depend on manual user contribution. It may be also due to that the popularity of WOT is still limited to in Europe and North America, as shown by the breakdown of user activity by region on its statistics page [23]. Considering only sites registered in the North America, the EU plus Norway and Switzerland, the coverage of WOT increases from 51.23% to 67.46%, while the coverage of SiteAdvisor increases to 94.98%.

The breakdown of the evaluated outcomes is included in Table 2. SiteAdvisor classifies 1.48% sites as 'bad'. This is interestingly close to the result in [13] which found 1.49% of Chinese sites, sampled using popular keywords on Baidu and Google (different from our sampling method), are malicious. WOT classifies 3.16% sites to be 'bad'. This larger value is likely due to the broader evaluation scope of WOT, which is not limited to the malicious sites only. In comparison, results by Safe Web and Safe Browsing Diagnostic Page may be too optimistic.

---

[1] We also tried a different scheme which treats all aspects as equally important (such that a site is classified as 'bad' if either Tr, Vr or Pr is <40), but found no significant changes in results.

[2] Based on the default risk signaling strategy of WOT and not including the child-safety rating.

[3] We regard a site that is not currently blacklisted and that has not been visited (and thus not evaluated) by Google's web bot in the past 90 days as 'not tested'.

The Venn diagram in Figure 1 shows that out of 296 and 102 'bad' sites that SiteAdvisor and Safe Web find respectively, only 8 are on their common blacklist. The small percentage of the common findings about 'bad' sites indicates the different testing methodologies employed and a lack of sharing between the two vendors. The lack of data sharing is also notable in the anti-phishing industry [7]. Previously this was a problem also in the anti-virus industry, but security vendors were found to have learned the lesson and are now sharing virus samples [7]. On the other hand, WOT finds 21 'bad' sites in common with Safe Web and 73 with SiteAdvisor. This hints on the better ability of WOT in identifying 'bad' sites that have been found by the others.

$$R = \frac{\Sigma T_p}{\Sigma T_p + \Sigma F_N} \qquad P = \frac{\Sigma T_p}{\Sigma T_p + \Sigma F_p} \qquad FS = \frac{2RP}{R+P} \qquad \textbf{(E1)}$$

Table 3. The true positives, false positives and false negatives scenarios in the case of Optimistic (left) or Conservative (right, in brackets) consensus. $F_{N,C}$, $F_{N,G}$ and $F_{N,U}$ denote the scenarios of false-negatives being classified as 'caution', 'good' and 'unknown' respectively.

	Findings of other services				
Finding of this service	Bad w/o any good	Mixed of good and bad	Caution only	Good w/o any bad	All unknown
Bad	$T_{P,B}$ [$T_{P,B}$]	$F_{P,M}$ [$T_{P,M}$]	$F_{P,C}$ [$F_{P,C}$]	$F_{P,G}$ [$F_{P,G}$]	$F_{P,U}$ [$F_{P,U}$]
Caution	$F_{N,C}$ [$F_{N,C}$]	- [$F_{N,C}$]	-	-	-
Good	$F_{N,G}$ [$F_{N,G}$]	- [$F_{N,G}$]	-	-	-
Unknown	$F_{N,U}$ [$F_{N,U}$]	- [$F_{N,U}$]	-	-	-

Table 4. Recall (R), Precision (P) and F-Score (FS) of individual services. R, $F_{N,C}$, $F_{N,G}$ and $F_{N,U}$ add up to 100%. The 5th row considers 'bad' sites to include only those with a credible warning (such that the add-on will prompt an explicit warning dialog to the user). The last 3 rows consider only the assessment of other automated services in the consensus outcomes.

	Optimistic consensus (%)						Conservative consensus (%)					
	R	P	FS	$F_{N,G}$	$F_{N,U}$	$F_{N,C}$	R	P	FS	$F_{N,G}$	$F_{N,U}$	$F_{N,C}$
WOT	15.3	1.7	3.1	11.1	72.2	1.4	22.1	14.3	17.3	22.6	49.4	5.9
SiteAdvisor	8.3	3.4	4.8	57.5	27.5	6.7	10.7	26.4	15.2	69.7	15.6	4.0
Safe Web	4.1	8.8	5.6	59.0	34.2	2.7	3.1	26.5	5.5	71.6	23.6	1.7
Safe Browsing DP	2.5	16.0	4.3	40.0	55.6	1.9	1.0	36.0	1.9	47.6	46.2	5.2
WOT [credible warning]	13.9	2.5	4.3	-	-	-	17.2	17.8	17.5	-	-	-
SiteAdvisor [auto]	10.0	2.0	3.4	68.3	18.3	3.3	8.3	3.4	4.8	68.6	20.7	2.5
Safe Web [auto]	2.9	4.9	3.6	65.7	28.0	3.4	3.5	10.8	5.3	66.1	26.9	3.5
Safe Browsing DP [auto]	3.3	16.0	5.4	42.3	51.2	3.3	2.1	32.0	3.9	43.6	44.6	9.7

To quantify the reliability in identifying 'bad' sites, we measured Recall (R), Precision (P) and F-Score (FS) i.e., three popular metrics used in the field of information retrieval. A challenge here is to determine the values of true positives (T_P), false positives (F_P) and false negatives (F_N) given that we do not know the 'correct' assessment outcomes which otherwise could be used as reference. We approached this by comparing the outcomes of a particular service with the consensus result of the three others. Thus, in this context, Recall (R) describes the success rate of a service in recognizing all consensus 'bad' sites, while Precision (P) measures the fraction of 'bad' sites identified by a service matching the consensus of the others (see E1). We define two types of consensus: *optimistic* and *conservative*. In the optimistic case, the consensus 'bad' sites are the ones classified as 'bad' by other services without any contradictory classification of 'good'. In the conservative case, the consensus 'bad' sites

include those that have mixed 'bad' and 'good' verdicts by individual services. We note that the conservative case may depict a more realistic scenario given the divergence in the classification of 'bad' sites. Table 3 shows the definitions of T_P, F_P and F_N. Table 4 shows the R, P and FS values.

Having the highest R in both optimistic and conservative cases, we find that WOT renders a more comprehensive protection against 'bad' sites in comparison to the three automated services. On a closer look, we also find that in the event that WOT fails to warn against sites having a 'bad' consensus rating, a higher percentage of these false-negatives are classified by WOT as 'unknown' or 'caution' rather than 'good' as indicated by the $F_{N,U}$, $F_{N,C}$, $F_{N,G}$ values in Table 4. Conversely, most of the false-negatives by SiteAdvisor and Safe Web are classified as 'good' rather than 'unknown' or 'caution'. This adds on to the reliability of WOT. Meanwhile, users should remain cautious even when a site has a 'good' rating from SiteAdvisor or Safe Web.

However, WOT has a low Precision (P) value in comparison to the others. As we learned from the developers that the browser add-on will only prompt the user an explicit warning when a 'poor' or 'very poor' rating (in either aspect of Tr, Vr or Pr) has a confidence level ≥ 8 (i.e., credible), we measured the precision of WOT considering 'bad' sites to be only those that will be explicitly warned against. As shown in the 5th row of Table 4, the Precision of WOT increases, but only slightly. The low P value may reflect that that WOT covers a broader evaluation scope than the others. Yet, a very low P value may result in a situation where users habitually regard all warnings as false positives as they do not observe similar warnings from the other services. It is thus important for WOT to inform the users about the differences.

If we weigh false-positives and false-negatives equally, the tradeoff between Recall and Precision can be measured by FS – the harmonic mean of R and P. In the optimistic case, all FS values are small (3.1% to 5.6%) with Safe Web having the highest FS, despite a low R. In the conservative case, the difference in FS values becomes more evident. WOT has the highest FS value of 17.3%. SiteAdvisor has a FS value of 15.2% and interestingly, the FS value of Safe Web remains at 5.5%.

One may reason that the low R and high P values of the automated services could be an artifact of comparing them with WOT which has a broader evaluation scope. As a robustness check, we measured the reliability of the automated services using only the outputs of the other two automated services to determine the consensus outcomes. As shown in the last three rows of Table 4, the P values drop without an evident improvement in R. All FS values are low (3.4% to 5.4%) even in the conservative case.

The above findings show that WOT is reliable in comparison to the three investigated automated services, especially when users should be cautious about web security, as captured in the case of conservative consensus. Overall, WOT has shown a better ability in recognizing 'bad' sites among the popular online destinations. Some of its warnings may concern online risks that are not currently evaluated by the others.

5.2 The few dominating contributors

According to Moore and Clayton [5], a highly skewed participation ratio increases the risks of manipulation in PhishTank. They argue that the corruption of a few highly active users can completely undermine the validity and availability of PhishTank. It is also not difficult for a highly active user to disrupt the system under cover of a large body of innocuous behavior [5]. We investigated if similar problems exist in WOT.

We analyzed dataset-II which describes the contribution level of 50,000 randomly selected users. Of these users, the total rating and comments posted are 214,872 and 20,420 respectively. Many of the registered users have not made any ratings or comments. Only 38.34% of them have rated and 7.16% have commented about a site at least once. On a closer look, we find a pattern similar to the one in [5]. Few users have contributed greatly while many have only made a modest contribution.

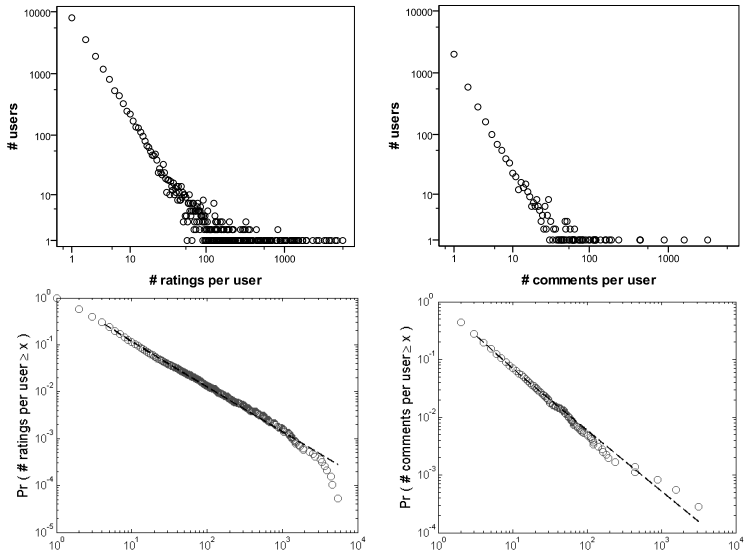

Figure 2. (Top) Number of users against the number of ratings and comments contributed. (Bottom) The complementary CDF of ratings and comments. Dashed lines depict the best fitted power law distribution with α=1.95, x_{min}=4 (rating, left) and α=2.05, x_{min}=3 (comment, right).

The seemingly straight lines in the log-log graphs (in Figure 2) suggest that the contribution of ratings and comments could be following the power law distribution. We computed the best-fit of power-law scaling exponent α and the lower cutoff x_{min} using maximum-likelihood estimation, based on the approach in [3]. We obtained the best fitted scaling exponent α=1.95 and lower cut-off x_{min}=4 for rating contribution, and α=2.05 and x_{min}=3 for comment contribution. The goodness-of-fit of these values were evaluated using the Kolmogorov-Smirnov (KS) test. We obtained a high *p*-value (0.76) for the parameters of comment contribution, indicating that it is likely to follow a power law distribution. This is, however, not the case for rating contribution where we rejected the null hypothesis that it is power-law at the 5% significance level.

We did not proceed to test if the rating contribution follows other types of heavy-tailed distributions (e.g., log-normal, Weibull) given that it is visually intuitive that a large percentage of the contribution comes from a small group of users. We observed that the complementary cumulative distribution function (CDF) of rating contribution begins to curve-in among the top contributors[4] (Figure 2, bottom left). Adapting from the 80:20 rule of the Pareto principle, we measured the skewness *S* such that *S* is the

[4] Excluding users who have contributed >3000 ratings, the KS test for a power-law fit gives a *p*-value of 0.36, indicating that it may be a power law distribution only with an upper cut-off.

largest *k*% of the total inputs coming from (100-*k*)% of the contributors. We found that *S* is 89 for rating and 95 for comment contribution. Put in words, 89% of the total ratings are provided by 11% of the rating contributors while the top 5% comment contributors gave 95% of the total comments. Both contribution ratios are skewed.

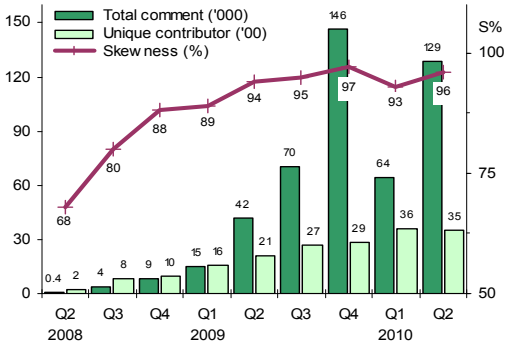

Figure 3. Total comment (in thousands), Unique contributor (in hundreds) and Skewness (%).

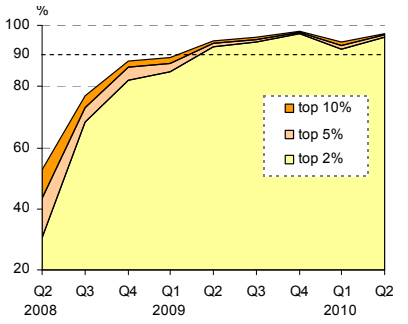

Figure 4. Percentage of comment provision by the top 2, 5 and 10% contributors.

We then studied the evolution of user participation using dataset-III, which contains 485,478 comments contributed by 16,030 unique users. Figure 3 shows an increasing number of comments and unique contributors per quarter. Unfortunately, the contribution ratio has become more skewed as WOT evolves, as shown by the *S* values in Figure 3. Since 2009 Q2, more than 90% of the total comments are actually provided by the top 2% active users (see Figure 4). The increasing trend of skewness is likely to be caused by the mass rating tool which allows one to rate and comment 100 sites at once. The privilege to use the tool was initially given to both the Gold and Platinum users since Sep 2008 (according to WOT's wiki [23]). As cases of misuse were detected, WOT began to award new privileges only to the Platinum users, from 28 Dec 2009 [23]. Revoking the privilege from those who have misused the tool might be the reason that has caused the dip in *S* and total comment during 2010 Q1.

We cannot inspect the evolution of rating contribution as individual ratings are kept private in WOT. Our guess is that rating contribution evolves similarly but not as skewed given that it does not fit well with a Power Law distribution and that it has a smaller *S* value than that of comment. In addition, WOT has made the rating process more convenient that commenting. Using the browser add-on, users neither need to first register themselves, nor visit the WOT's website in order to give a rating.

Skewed participation patterns are not entirely unexpected; some users are naturally more inclined to contribute than the others. WOT also has put in place a number of features to mitigate the risks of exploitation. First, in its current form, security decisions in WOT are not easily guessable due to the hidden nature of the aggregation formula and individual ratings. WOT also states that it does not weigh the user inputs based on the activity level of individual contributors; the weights are computed from the reliability of their past rating behavior. These measures make the repeated cheating by a single highly active user difficult. One may be able to cast biased ratings unnoticed amidst a large number of innocuous inputs, but this is only valid if it is cost-efficient for the attacker to build up a reputation in order to rate up or down a few targeted sites. An attack may be more easily done with the help of several accomplices,

or through a 'pseudo reliability' built by providing automatic ratings with reference to some public blacklists. The developers state that there are automatic mechanisms in WOT which monitor and investigate suspicious user behavior. Yet, to the root of the challenges, WOT should work towards diversifying the user contribution so that it does not becoming a centralized/homogenous system overwhelmed with the inputs of a few. The mass rating privilege should be handled with care.

5.3 Exploitability, disagreement and subjectivity

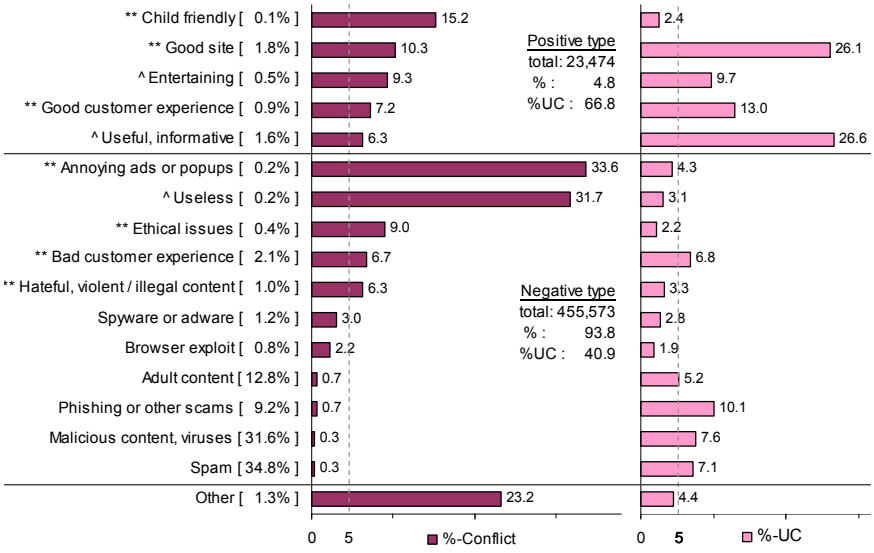

Figure 5. %-Conflict, % of Unique Contributor (%-UC) and %-count (in brackets) of different comment categories. [^ denotes not in the scope of rating, ** denotes having a %-conflict value > 5]

Grouping the comments according to their respective category, we observed that there are many more comments of negative type than positive (see Figure 5). We measured the percentages of conflict (%-conflict) and unique contributors (%-UC) of each comment category. A 'conflict' is defined to arise when a comment of positive type is given to a site that has a poor rating (<40 in either Tr, Vr, Pr or Cs aspect), or when a comment of negative type is given to a site that has a good rating (≥60 for all aspects). A conflict can happen due to several reasons. Firstly, it can be due to the difference in scope between the comment and rating. Specifically, whether a site is useful or not, and whether it is entertaining are factors not evaluated by the four rating aspects. Secondly, assuming that the ratings reflect the true state of a site, a conflict can be due to user attempts to cheat (e.g., to defame or lie about a site of interests) or simply divergent views. We could not easily differentiate between exploitation and disagreement, but underlying the two are common factors of subjectivity and non-verifiability.

Excluding categories that are not in the scope of rating (i.e., Entertaining, Useful and informative, and Useless), we found that categories that concern user experience and content (except for 'adult content') have a %-conflict value of >5. In comparison, there is little conflict resulting from comments which warn about browser exploits, phishing sites or adult content. We attribute this to the different levels of objectivity.

For example, feedback on whether a site has annoying ads and whether a site provides good customer experience are subjective. Meanwhile, one cannot believably allege a site for phishing, browser exploit or adult content without verifiable evidence.

In addition, we found no association between a low %-conflict value and a small group of contributors. Comments with categories such as 'Adult content', 'Malicious content, viruses' and 'Spam' are provided by more than 5% of total contributors but have a low level of conflict. Conversely, comments about 'Child friendliness' and 'Ethical issues' are given by fewer users but result a higher level of conflict.

The above observations have several implications. First, signaled by the low %-conflict, identifying a phishing site is an objective process. Given that an objective evaluation is verifiable, there is a reduced chance for successful manipulation going unnoticed, even by the highly active users. This may have served to mitigate the risks and incentives of exploitation in PhishTank. Indeed, despite the early criticisms on its highly skewed (power law) participation ratio [5], PhishTank is now adopted by multiple vendors including Yahoo!, Mozilla, Kaspersky Lab, Opera and McAfee [20].

Risks of exploitation can, however, be a real issue for WOT since several of its evaluation aspects, such as trustworthiness and vendor reliability are subjective in nature. Fortunately, in its current state, we found that a large majority of the user comments actually come under categories that have a low level of conflict e.g., 'adult content', 'malicious content, viruses', 'spam' and 'phishing or other scams'. Although we cannot know for sure, the pattern exhibited here does imply that the existing user ratings are largely based on objective criteria. While evaluation based on objective criteria do not equate honest assessment, for example one can accuse an innocent site to be malicious, such manipulation can be discovered and punished with an appropriate level of monitoring. This reduces the incentives of such an attack.

Yet, it is not unreasonable to expect an increase of subjective user inputs in the long run. 7 of the 13 comment categories in the scope of rating actually have a %-conflict value of more than 5. Comments that come under these categories were also in fact contributed by more than half of the unique contributors. Subjective opinions, if provided honestly, are valuable to a user-based system as they mark the diversity of the participants. The challenge lies in that we cannot assume the honesty of users. Subjective and non-verifiable evaluation criteria can be exploited easily.

5.4 User concerns on web security

We also looked at popular words used in user comments and how the trend may have changed over time. As we discovered that a large number of comments are made with exactly the same description likely to be caused by the mass rating tool, we used only unique comments in our analysis. We parsed for nouns and transformed them into the singular form. Table 5 shows the most frequently used words ranked in popularity. We observe that 'spam' and 'scam' are among the most common issues discussed in user comments. The word 'information' is also frequently used in conjunction with 'personal' and 'sensitive' describing privacy concerns. Another popular word is 'pharmacy' which is found in warnings against fake or illegal online pharmacy sites. The use of the word 'phishing' becomes dominant since late 2008. Meanwhile, concern about malware on the web, virus and Trojan included, is increasing.

Overall, this analysis indicates that user concerns on web security are not limited to only phishing and malware. This brings up the limitation of the automatic services in

catering for user concerns on online risks such as scams, illegal pharmacies, information protection and inappropriate content in general (as highlighted in Table 5).

Table 5. Popular words used in user comments per year quarter.

08'Q2	08'Q3	08'Q4	09'Q1	09'Q2	09'Q3	09'Q4	10'Q1	10'Q2
site	site	site	site	site	site	site	site	site
info	spam	spam	spam	spam	spam	spam	malware	spam
spam	criminal	info	scam	scam	malware	scam	Trojan	scam
email	email	phishing	phishing	phishing	Trojan	phishing	virus	phishing
link	info	software	software	malware	scam	info	spam	malware
people	trade	scam	pharmacy	software	phishing	malware	threat	info
pharmacy	scam	criminal	virus	info	info	pharmacy	exploit	pharmacy
software	gang	security	info	pharmacy	software	software	scam	credit card
service	porn	service	download	virus	exploit	email	phishing	abuse
child	pharmacy	warning	porn	registrar	virus	virus	info	software
privacy	brand	content	link	exploit	content	link	pharmacy	risk
product	child	email	malware	Trojan	download	Trojan	software	virus

* words with the same meaning e.g., domain, website, page (~site), scammer (~scam), program (~software) were omitted.

6 Discussion

The strengths of WOT lies in a number of its characteristics. First, it caters for different user concerns about web security and does so reliably. Its overall ratings are not easily guessable and hence there is little chance of manipulation. The browser add-on has also made the process of rating a site very easy. Sub-domains are designed to inherit the reputation of the parent domain unless there are sufficient ratings for the sub-domain itself, avoiding redundant user effort. WOT also encourages users to contribute responsibly by weighing the inputs according to the reliability of individual contributors through statistical analysis. In a private communication with the developers, we were told that WOT has also factored in the dynamics of aggregate ratings as the weight of individual ratings is set to decay (until the respective contributors re-visit the sites). The system is also capable of ignoring spammers and suspicious ratings as WOT monitors for unusual rating behavior automatically. Finally, the community features such as discussion forum, messaging channels between users and the ability to request for public reviews have all contributed to a reliable reviewing process.

Yet, WOT is not without several weaknesses. We discuss some of them below:

- **Skewed contribution patterns.** The contribution patterns of rating and comment are skewed, most likely due to the mass rating tool. A highly skewed contribution pattern can cause WOT to be overwhelmed by the inputs of a few, violating the diversity and decentralization conditions of the wisdom of crowds. While the risks of exploitation due to a skewed participation is expected to be limited given the measures taken in WOT and the observation that a majority of the current user inputs are based on objective evaluation factors (in Section 5.3), we suggest to handle the mass rating tool with a greater care. It may be wise to restrict the highly active users to use the tool only for evaluation aspects that are objective and verifiable. At the time of writing, it is also not mandatory for them to provide the evidence of their mass ratings, although they are required to submit a comment in which it is recommended to include the relevant references and that they must be contactable by anyone who disagrees with the rating. Attention must also be given to potential gaming behavior such as building up a pseudo reputation by simply referencing the

publicly available blacklists. Essentially, WOT should work on diversifying the sources of bulk contribution.

- **A hidden approach.** While the hidden aggregation formula and user ratings may have played a part in making the assessment outcomes in WOT less easily guessable and less vulnerable to manipulation, a hidden approach may in general result in a lack of user confidence. The situation can be more tricky when users compare the assessment outcomes of WOT with other services given that warnings by WOT may not be frequently supported by the automated services (as characterized by the low precision value). Users who are not aware of the broader evaluation scope of WOT, may doubt the reliability of the black-box computation and regard its warnings as mere false-positives. Neither will a hidden approach benefit from the scrutiny and suggestions for improvement from the community. It may be worth the effort for WOT to educate the users concerning its differences from the automated services. A more transparent approach capable of withstanding manipulation, especially by the highly active users, would be the preferred option in the long run.
- **Subjective evaluation criteria.** Subjective evaluation factors can result in contentious outcomes besides increasing the risk of manipulation. In the current state, WOT does not seem to differentiate between objective and subjection evaluation criteria. Improvement can be made in this respect. For example, the rating aggregation strategy may factor in the subjectivity level of the underlying evaluation factor. WOT may also consider tapping into the potentials of personalized communities as proposed in [1,2] to deal with subjective factors. Inputs from personalized communities have the advantages of being more trustworthy, relevant and thus more impactful than those provided by unknown community members [1,2].

There are several limitations to our study in the work as presented here. First, as our evaluation sample consists of sites randomly chosen from the one million most-frequently visited sites, the work has not evaluated the reliability of WOT when dealing with bad sites that are more frequently found in the long-tail of web popularity. Further, the timeliness of WOT's assessment is not tested. It may appear that an assessment by WOT can take a longer time than the automated systems as it depends on user inputs and can miss out on malicious sites which are often online for a short period of time only. While these concerns are valid, we note that they are being covered in WOT by the inclusion of blacklists from trusted third party sources. Future investigation on these concerns would be interesting.

7 Conclusions

We have found that the Web of Trust (WOT) is more comprehensive than three popular automated services in identifying 'bad' domains among the frequently visited sites. The contribution patterns in WOT are found to be skewed with the comment contribution following a power-law distribution. However, WOT has implemented a number of measures to mitigate the risks of exploitation. In addition, a large majority of its current user inputs is found to be based on objective evaluation factors and hence verifiable. This may have also helped to reduce the risks and incentives of exploitation in PhishTank. We find that user concerns on web security are not limited to malware and phishing. Scams, illegal pharmacies and lack of information protection are regular issues raised but are not evaluated by the automated services. There is also

an evident lack of sharing among the vendors of the automated services. We include a discussion on the strengths and weaknesses of WOT which may be helpful for designing user-based security systems in general. In short, WOT clearly exemplifies that the wisdom of crowds for assessing web security can work, given a careful design.

Acknowledgement

We thank N. Asokan, B. Westermann and the anonymous reviewers for their comments, and the WOT developers for dataset-II, III and details about the WOT system.

References

1. Camp, J.L. Reliable, Usable Signally to defeat Masquerade Attacks. In *Proc. WEIS* 2006.
2. Chia, P.H., Heiner, A.P., and Asokan, N. Use of Ratings from Personalized Communities for Trustworthy Application Installation. In *Proc. NordSec* 2010.
3. Clauset, A., Shalizi, C. R., and Newman, M. E. J. Power-law distributions in empirical data, *SIAM Review* 51(4), 661-703 (2009).
4. Edelman, B. Adverse selection in online 'trust' certifications. In *Proc. WEIS* 2006.
5. Moore, T., and Clayton, R. Evaluating the Wisdom of Crowds in Assessing Phishing Websites. In *Proc. FC* 2008.
6. Moore, T., and Clayton, R. The Impact of Incentives on Notice and Takedown. In *Managing Information Risk and the Economics of Security,* ed. Johnson, M.E., 2008.
7. Moore, T., Clayton, R., and Anderson, R. The economics of online crime. Journal of Economic Perspectives, 23(3):3-20, 2009.
8. Provos, N., Mavrommatis, P., Rajab, M.A., and Monrose, F. All your iFRAMEs point to Us, In *Proc. USENIX Security* 2008.
9. Sheng, S., Wardman, B., Warner, G., Cranor, L.F., Hong, J., and Zhang, C. An Empirical Analysis of Phishing Blacklists. In *Proc. CEAS* 2009.
10. Surowiecki, J. The wisdom of crowds. Anchor Books, 2005.
11. Wondracek, G., Holz, T., Platzer, C., Kirda, E., and Kruegel, C. Is the Internet for Porn? An Insight into the Online Adult Industry, In *Proc. WEIS* 2010.
12. Y. Zhang, S. Egelman, L. Cranor, and J. Hong. Phinding Phish: An Evaluation of Anti-Phishing Toolbars. In *Proc. NDSS* 2007.
13. Zhuge, J., Holz, T., Song, C., Guo, J., Han, X., and Zou, W. Studying Malicious Websites and the Underground Economy on the Chinese Web. In *Proc. WEIS* 2008.
14. Alexa: Top million sites. http://www.alexa.com/topsites
15. Google Safe Browsing diagnostic page.
    Query page: http://www.google.com/safebrowsing/diagnostic?site=<site>
16. LegitScript. http://www.legitscript.com/
17. McAfee SiteAdvisor. http://www.siteadvisor.com
    Query page: http://www.siteadvisor.com/sites/<site>
18. McAfee TrustedSource. http://www.trustedsource.org
19. Norton Safe Web. http://safeweb.norton.com
    Query page: http://safeweb.norton.com/report/show?url=<site>
20. PhishTank. http://www.phishtank.com
    Vendors that have adopted PhishTank: http://www.phishtank.com/friends.php
21. SpamCop. http://www.spamcop.net
22. The UK Card Association. New Card and Banking Fraud Figures, 10 March 2010.
    http://www.theukcardsassociation.org.uk/media_centre/press_releases_new/-/page/922/
23. Web of Trust (WOT). http://www.mywot.com
    Query API: http://api.mywot.com/0.4/public_query2?url=<site>