Community-based Web Security: Complementary Roles of the Serious and Casual Contributors

Pern Hui Chia
Q2S NTNU*
Trondheim Norway
chia@q2s.ntnu.no

John Chuang
UC Berkeley
Berkeley CA
chuang@ischool.berkeley.edu

ABSTRACT
Does crowdsourcing work for web security? While the herculean task of evaluating hundreds of millions of websites can certainly benefit from the wisdom of crowds, skeptics question the coverage and reliability of inputs from ordinary users for assessing web security. We analyze the contribution patterns of serious and casual users in Web of Trust (WOT), a community-based system for website reputation and security. We find that the serious contributors are responsible for reporting and attending to a large percentage of bad sites, while a large fraction of attention on the goodness of sites come from the casual contributors. This complementarity enables WOT to provide warnings about malicious sites while differentiating the good sites from the unknowns. This in turn helps steer users away from the numerous bad sites created daily. We also find that serious contributors are more reliable in evaluating bad sites, but no better than casual contributors in evaluating good sites. We discuss design implications for WOT and for community-based systems more generally.

Author Keywords
Web Security, Wisdom of Crowds, Web of Trust (WOT)

ACM Classification Keywords
K.4.3 Computers and Society: Organizational Impacts

General Terms
Measurement; Security.

INTRODUCTION
Despite the efforts of a multi-billion dollar computer security industry, web security remains a largely unsolved problem. Large numbers of malicious sites continue to serve as platforms for phishing, malware, and other security exploits. Provos et al. [20] found over 3 million URLs (hosted on more than 180,000 sites) that initiated drive-by downloads – automatic installation and execution of malware on the machines of unsuspecting visitors. Zhuge et al. [26] found that 1.5% of Chinese websites, sampled using popular keywords on Baidu and Google search engines, were malicious. Meanwhile, the Anti-Phishing Working Group recorded more than 67,000 phishing attacks worldwide in the second half of 2010 [2].

Detection, blacklisting, and takedown of malicious sites have been traditionally handled by security vendors such as antivirus and brand protection companies. Detection and blacklisting of suspicious sites are typically done with automated sensing and classification using heuristics and machine learning. Given that the takedown of a malicious site can be cumbersome and protracted in time, tools have been created to warn the users about suspicious sites on the web.

However, many of the automated risk signaling tools, including McAfee's SiteAdvisor [15] and Norton's Safe Web [17], fall short in identifying 'bad' sites that try to trick or harm users in a variety of subtle ways. Problems of increasing concern that are not adequately handled by security vendors include the misuse of personal information, scams and fraudulent sites such as illegal online pharmacies. The automated tools also do not evaluate content appropriateness. While it is a personal judgment whether adult content is appropriate, the fact that adult sites regularly rank among the top 50 most visited sites and are often associated with malware, script-based attacks and aggressive marketing strategies [25], do indicate a serious problem. Furthermore, verifying the goodness of sites is not a straightforward task. Online certification issuers, such as BBBOnline and TRUSTe, strive to distinguish the 'good' sites from the 'bad' ones, but conflicts of interest can sometimes arise. When certifiers adopt lax requirements in certifying sites in the 'gray' category, the problem of adverse selection may result in the certified sites having lower trustworthiness than those that forego certification [9].

The limitations of automated tools and the potential risks of centralized judgment have prompted the alternative approach of leveraging community input for web security. Encouraged by the success of peer-production systems such as Wikipedia, yelp, and reCAPTCHA, the crowdsourcing of website security evaluation holds the promise of scalability. Yet, there remain concerns on the ability of community members in providing timely and reliable evaluation of a large number of websites. In addition to the typical problem of malicious or misinformed contributors present in a peer-production system, there are additional challenges in the context of web security. First, one would expect a certain level of security expertise, and therefore a higher contribution barrier, to evaluate the security of a website. Determining if a website en-

*Centre for 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.

CSCW'12, February 11–15, 2012, Seattle, Washington, USA.

gages in a variety of security exploits is different from bookmarking a page using reddit or reviewing a book on Amazon. Second, attackers play a game of cat-and-mouse by creating large numbers of malicious websites every day, typically with short lifespans. The challenges of coverage and timeliness are therefore different from the case of Wikipedia, where the number and content of articles are relatively static.

Despite these concerns, several community-based systems for web security such as Web of Trust (WOT) [22] and PhishTank [19] have achieved significant impact. PhishTank collects user reporting and voting on suspected phishing sites. Its assessments are used by popular vendors including Yahoo!, McAfee, Mozilla and Opera. On the other hand, WOT collates individual user ratings into aggregate ratings on four aspects of web security, namely trustworthiness, vendor reliability, privacy, and child safety. Facebook has recently incorporated the assessments by WOT to protect its users from potentially harmful URLs [10]. A recent study by Chia and Knapskog [5] found that WOT was more comprehensive than three automated counterparts namely SiteAdvisor, Safe Web, and Safe Browsing Diagnostic Page in identifying bad domains among the frequently visited sites.

There is certainly value in leveraging on community inputs for web security. Beyond comparing the overall reliability of community-based systems against the automated counterparts, in this paper, we set out to study how different types of contributors (casual and serious) play their parts in advancing WOT in the challenging domain of web security. Understanding the roles of different contributors can lead us to a clearer picture of the underlying success factors and potential pitfalls. Specifically, in this work: (I) We ask how do the casual contributors add value to WOT given the steep contribution barrier of assessing web security? (II) We study how different types of contributors may choose to focus on different types of websites (popular/unknown, malicious/benign) or different trustworthiness aspects of websites (e.g., phishing, spam, inappropriate content). (III) We also seek to characterize the contribution patterns of casual and serious contributors, and to examine if there is a room to better coordinate the limited human resources. (IV) We also determine if different types of contributors realize different levels of reliability in their assessments. We hope that these questions can yield insights applicable to other contexts beyond web security.

In the following, we first describe the related works before detailing on how WOT works in practice and our methodology. We then present our analysis results focusing on the coverage, coordination, and reliability of inputs by the serious and casual contributors. Finally, we discuss the design implications to WOT and community-based systems more generally.

RELATED WORK

Collective Wisdom in General

A large number of prior works on collective wisdom have focused on the participation patterns in Wikipedia (e.g., [3, 12, 18, 23]), its potential pitfalls and risks (e.g., [8]) as well as its success factors and how to improve it (e.g., [6, 7, 11, 13]). Our work is related to that of Kittur et al. [12] and Welser et al. [23] in the way that we are interested in the roles played by different types of contributors for collective intelligence. While Kittur et al. [12] has observed a shift of workload from the elite class contributors to the less active ones over time, Ortega et al. [18] concluded that the contribution pattern in Wikipedia has remained highly skewed even in the stable phase. We note that however contribution should measure more than just the count of article edits and submissions. Indeed, even though the role by the less active contributors might appear overshadowed by the few serious contributors, prior research (e.g., [3]) has pointed out that even the readers (lurkers) can and do contribute to a collaborative system like Wikipedia. In this work, rather than focusing only on the count of comments and ratings, we measure the roles of different types of contributors judging from the coverage, coordination and reliability of their assessments.

Wilkinson [24] describes two macroscopic characteristics in peer production systems and shows how the two regularities arise from simple dynamic rules. First, he demonstrates that the probability a person stops contributing is inversely proportional to the number of contributions he has made, which in turn leads to a power law contribution distribution in all four systems (Wikipedia, Bugzilla, Digg and Essembly) he investigated. He found also a lognormal distribution of per topic activity – a small number of very popular topics accumulate the vast majority of contributions due to a multiplicative popularity reinforcement mechanism. We do not evaluate if the contribution patterns in WOT follows a specific distribution in this paper, but we observe that both the distributions of per person contributions and per site inputs do have a heavy tail. The skewed attention distribution among sites evaluated by the casual contributors is interesting as it suggests the possibility of better coordinating the security crowds for a higher level of efficiency.

Mamykina et al. [14] argued that the success of Stack Overflow attributes not only to the careful design considerations, but also to the high visibility and interactive involvement of the design team in the community. The authors further highlighted that this model of continued community leadership presents challenges to port the success of Stack Overflow easily over to other domain specific systems. This argument has only made it more appealing to better understand the roles played by different contributors as we aim for in this paper.

Collective Wisdom for Web Security

Denning et al. [8] highlighted six areas of potential risks in Wikipedia, namely accuracy, motives of contributor, uncertain expertise, volatility of content, sources of information and coverage. The first five areas relate to the correctness of information, suggesting a heavier focus on content reliability. All six areas are valid concerns facing the community-based web security. We note that however coverage is just as important given that it is the strategy of attackers to create numerous new bad sites to thin the resources of the defenders.

Moore and Clayton [16] evaluated the effectiveness of PhishTank – a community-based system for reporting and voting against suspected phishing sites. They found that the participation ratio in PhishTank was highly skewed (following

a power-law distribution), making it particularly susceptible to manipulation. Compared to a commercial phishing report, they found that PhishTank was slightly less comprehensive and slower in reaching a decision. 3% of the sites reported as suspicious (out of a total of 176,654) were found to be invalid phishes. A large percentage of the incorrect submissions came from the less active contributors. However, considering the eventual assessment outcomes (when the initial reporting is validated or corrected by the subsequent voting mechanism), they have found only 39 false positives and 3 false negatives in total.

The study by Chia and Knapskog [5] was the among the first that evaluated WOT comparing it with the assessment outcomes of SiteAdvisor, Safe Web and Safe Browsing Diagnostic Page. They found that the participation ratio in WOT was also highly skewed. However, WOT was in fact more comprehensive than the three automated systems in identifying bad domains (amongst the top million most visited sites as published by Alexa [1]). The study also concluded that user concerns on web security are not limited to malware and phishing. Scams, illegal pharmacies and misuse of personal information are regular issues raised by WOT's community while they are not evaluated by the automated services. In a similar study, Ayyavu and Jensen [4] rejected the unfair generalization on the low reliability of community-based rating systems as they found that, among frequently visited sites that have been co-evaluated by WOT and SiteAdvisor, the disagreement in assessments was actually less than 10%.

Building on the above studies, our work here looks beyond the overall reliability of crowdsourcing for web security. Acknowledging the potentials of such systems, we examine the roles being played from different types of contributors to better understand the underlying success factors and potential pitfalls. We center our analysis around the coverage, coordination and reliability of user assessments in line with the typical concerns of collaborative systems and web security.

WEB OF TRUST (WOT)

WOT is a reputation system that collates community inputs into aggregate ratings for different websites. It takes the form of an open source browser add-on and a website (mywot.com) with a number of community features including a personal page per registered user, discussion forums, a wiki as well as messaging and polling tools. The add-on has been downloaded for more than 23 million times by August 2011.

User Ratings and Comments

Individual user ratings and the aggregate ratings for different sites in WOT are structured around four aspects: trustworthiness, vendor reliability, privacy and child safety. The ratings range from very poor (0-19%), poor (20-39%), unsatisfactory (40-59%) to good (60-79%) and excellent (80-100%).

WOT weighs the input ratings differently based on the reliability of individual contributors. The reliability of a contributor, decoupled from his activity level or contribution count, is computed with Bayesian inference based on his past contributions. Individual user ratings are kept private to the contributors. The rating aggregation formula is also not publicly available. WOT argues that the hidden formula and individual inputs, plus the Bayesian inference rule, help to mitigate gaming behaviors. We learned from the developers that they have built in automated mechanisms to monitor for suspicious contribution behaviors. They have also factored in the freshness of user ratings by setting the weight of individual ratings to decay over time (until the rater re-visits the site).

Positive category	Negative category	Other
Entertaining	Useless	Other
Useful, informative	Annoying ads or popups	
Child friendly	Ethical issues	
Good customer experience	Hateful, violent/illegal content	
Good site	Bad customer experience	
	Browser exploit	
	Spyware or adware	
	Adult content	
	Phishing or other scams	
	Malicious content, viruses	
	Spam	

Table 1. Comment categories of positive or negative nature in WOT.

Other than numerical ratings, users can also evaluate a site by textual comments. To give a comment, they must first register themselves on mywot.com. There are more than 2 millions registered users to date. Unregistered users (i.e., anyone who has downloaded the add-on) can only rate a site through the add-on, which assigns a unique pseudonym to the user. When submitting a comment, the user selects one out of 17 comment categories that best describes their concern. As shown in Table 1, excluding the category 'Other', 5 of the comment categories are positive in nature, while the remaining 11 are negative. Comments do not count towards the aggregate ratings, but they provide a way of reasoning as to how a user has rated a particular site. Unlike the individual ratings, comments are publicly accessible on the *scorecard* of each evaluated site. The scorecard of a particular site refers to a uniquely reserved page on mywot.com that shows the aggregate ratings and user comments given to the site along with other details such as its traffic ranking, server location, description and links for further information.

Mass Rating Tool

WOT ranks the community members starting from rookie, bronze, silver, gold to the platinum level. The ranking is done based on the activity score which is computed from the total ratings and comments contributed, different from the reliability score that is kept private and designed to incentivize the users to contribute responsibly. Platinum members are given the privilege to use the *mass rating tool* which allows them to evaluate (at maximum) 100 sites at the same time with the same rating and comment. It is a handy tool for those who have access to some blacklists (e.g., on spamming, phishing and malicious sites) to submit the bulk evaluations conveniently.

Trusted sources

Besides user ratings and comments, WOT does factor in inputs given by trusted third parties. For example, it receives blacklists of phishes, spamming sites and illegal online pharmacies from PhishTank, SpamCop.net and LegitScript.com

respectively. Inputs from the trusted third parties play an important role in improving the coverage and timeliness of WOT in responding to new bad sites created by the attackers daily. We do not have access to the inputs from these trusted sources (nor the ratings from individual contributors). We will focus only analyzing the user comments in this paper.

Risk Signaling and Warning

WOT signals the reputation of different URLs through the browser add-on using colored rings (red for 'bad', yellow for 'caution', green for 'good', gray for 'unknown'). By default, the reputation of a site is computed based on the trustworthiness (tr) rating which covers whether a site can be trusted and is safe to use (without malicious content). A site is considered bad if tr$<$40, caution if 40$\leq$tr$<$60, good if tr$\geq$60, and unknown if tr is not available or if a minimal confidence level has not been obtained. A special case is when WOT finds a *credible warning* in either aspect of vendor reliability or privacy and thus treating the site as bad. By credible warning, we refer to the case when a particular aspect is given an aggregate rating below 40% with a confidence level above 8%. The confidence level is computed based on both the number of ratings and the reliability scores of the contributors. In the presence of a credible warning, besides displaying a red ring next to the URL, WOT prompts a large warning dialog to the user if he clicks on the link. The child safety rating is ignored by default. The settings for risk signaling and warning can however be configured to suit the needs of different users.

Evaluation Statistics

According to its statistics page, WOT has evaluated more than 32 million sites by August 2011. The community may however have quite some catch-up to do considering that there are more than 205 million domain names now (as estimated in [2] and [21]), giving WOT an overall coverage of 15.6%. As found in [5], the coverage of WOT among Alexa's top million most visited sites was 51.2%, but still lower than SiteAdvisor (87.9%) and Safe Web (68.1%). Among the 32 million sites evaluated by WOT, 3.4 millions (10.6%) are regarded as bad with a low trustworthiness rating. While no one can be sure about the total bad sites on the web (given that many of them are undetected), researchers found that 1.5% of the frequently visited Chinese sites were malicious [26], and 1.3% of Google search queries received more than one malicious URL in the result page [20]. Putting the above figures together, WOT does appear to have a better coverage for bad sites than the good ones in its current state. Indeed, WOT was found to be more comprehensive than SiteAdvisor, Safe Web and Safe Browsing Diagnostics Tool in identifying bad domains among the frequently visited sites [5].

METHODOLOGY AND DATA COLLECTION

For this study, we have obtained two valuable datasets from the WOT developers (hereafter referred to as DS-Comment and DS-Activity). DS-Comment consists of 600,000 comments randomly selected from more than 12 millions in total in WOT in early 2011. The comments evaluate a total of 504,874 sites and were submitted by 20,657 unique contributors. Each comment in the dataset is accompanied by details including the user ID, date of writing, evaluated domain as well as a comment category as specified by the contributor. We made use of the positive or negative nature of a comment category (as classified in Table 1) to determine the positive or negative sentiment of the contributor's assessment. We thus refer to a negative (positive) comment as one that has been given a negative (positive) comment category in this article. On the other hand, DS-Activity describes the total ratings and comments that each of the 20,657 contributors has given considering the entire database of WOT. The dataset thus indicates the activity level of the contributors in WOT in entirety; we made use of it to distinguish between different types of contributors (casual or serious). Put together, these two datasets allow us to evaluate the coverage (attention) provided by different types of contributors, various characteristics of sites they attend to, and the potential coordination among themselves.

To evaluate the reliability of inputs given by different contributor types, we then randomly selected 5,000 domains from the 504,874 evaluated in DS-Comment and queried for their aggregate ratings from WOT. Note that the aggregate ratings of WOT have factored in the reliability scores of different contributors and additional inputs (if any) from trusted third parties. For each of the 5,000 sites, we queried also the assessments by SiteAdvisor (SA) and Safe Web (SW) – two services provided by McAfee and Norton respectively. SA evaluates a site based on proprietary and automated tests on aspects such as downloads, browser exploits, email, phishing and annoyance factors (e.g., pop-ups). It also receives inputs from TrustedSource.org (also owned by McAfee) which evaluates aspects including site behavior, traffic and linking patterns, and site registration and hosting. Similarly, SW run automated tests to determine if a site imposes threats such as drive-by download, phishing, spyware, Trojan, worm, virus, suspicious browser change, joke program and identity theft. Both SA and SW do collect user comments (and ratings in the case of SW) but these inputs do not count towards the eventual assessment. We parsed the reports and obtained the assessment outcomes which constitute our third dataset, DS-Reliability. The querying process took place in April 2011. We repeated the queries in mid May and found no significant changes in the assessment outcomes by all three services.

Limitations

We list here several limitations to our study. First, given that we do not have access to the ratings given by individual contributors (which are kept private in WOT), we will be projecting the attention and concern of different contributor types judging from their comment contribution. This should not be problematic as we find a strong correlation (r=0.89 with $p<.001$) between the total ratings and total comments one has contributed from DS-Activity.

Secondly, for our analysis on coverage and attention, we will assume that the contributors have specified a category that fits their comment correctly. The same assumption is used as we will leverage on the nature of a comment category (positive or negative) to evaluate the reliability of different contributor groups in assessing bad and good sites. We note that this as-

sumption is reasonable given that there is no motivation for a user to cheat or game the system by choosing a false category as comments do not affect the aggregate outcome.

Another limitation relates to the fact that we will measure the coverage and reliability of different contributor groups based on the sites evaluated in DS-Comment. The ratio of comments given a category of negative nature is much higher than the positive ones in DS-Comment, in line with the statistics on mywot.com. While this gives us an accurate representation of the state of contribution distribution in WOT, it may be misleading to take, for example, the loss of coverage in the absence of the casual contributors to be minimal (2.16% as we will show in the next section). The impact will be larger if we consider sites relevant to the daily browsing patterns of ordinary users. For example, among the top million most visited sites, WOT rated 45.9% of them as good [5] – a stark difference to the small proportion of positive comments (5%) in DS-Comment. This suggests a more important role played by the casual contributors than it may appear.

We note that also DS-Comment contains comments that may have been later removed by the contributors (e.g., when a negative comment is disputed by other users as a false positive). We pay attention to this when evaluating the reliability of different contributor groups given that the test sample is smaller.

ANALYSIS / RESULTS

We will first describe the macroscopic contribution patterns in WOT and how we categorize the contributors based on their activity level. Then, we delve into the characteristics of the two extreme types of contributors (serious or casual) and measure their roles in covering sites of different natures as well as evaluating them reliably. We study also how the different contributors may have (mis-)coordinated themselves.

Characterizing Different Types of Contributors

Figure 1 plots the contribution distributions of ratings and comments using DS-Activity. Both of them do not fit a power-law distribution (different from in [5] where the comment contribution was found to be following a power-law distribution). We did not test if they fit some other types of heavy tailed distributions (e.g., log-normal, Weibull) but it is visually intuitive that the distributions are skewed. This is not entirely unexpected; a skewed contribution distribution of a community-based system can be characterized by the 'participation momentum' [24] – the more contributions one has made, the lower it is the likelihood of him quit contributing. An interesting observation (not shown in figure) is that not all the highly active contributors actually arrived from the beginning. WOT has managed to attract new highly active members as the community evolves.

While more ratings have been given than comments (per person) on average, the difference is not statistically significant. There is a strong correlation (r=0.89 with p<.001) between the number of ratings and number of comments contributed per person. This indicates the feasibility to study the different characteristics of the contributors based on the comments given instead of ratings that are not publicly available.

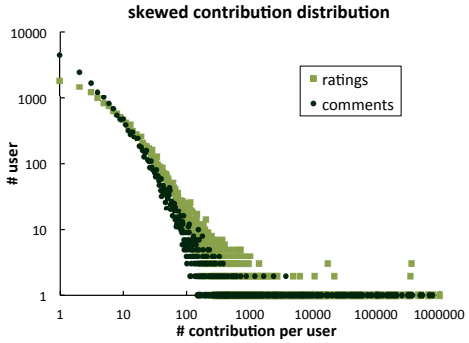

Figure 1. Distribution of comment and rating contribution.

		Statistics from DS-Comment		
Contr. group	Total comments (from DS-Activity)	# comments contributed	# sites evaluated	# unique users
u0	1 – 9	15,493	12,932	13,924
u1	10 – 99	18,727	17,306	5,850
u2	100 – 999	8,569	8,197	703
u3	1000 – 9999	16,956	16,641	106
u4	10000 – 99999	80,607	73,965	44
u5	100000 or more	459,648	407,778	30
	All groups	600,000	504,874	20,657

Table 2. Grouping based on total comments one has given in WOT.

We categorize the contributors according to the number of comments one has given, with *u0* denoting the group of *casual* contributors who have provided less than 10 comments, and *u5* denoting the group of *serious* contributors who have given at least 100,000 comments. In other words, each contributor group corresponds to a different contribution level measured in terms of the base 10 magnitude order of the total comments contributed. Table 2 details on the categorization rules, total comments, total unique sites covered and the size of each contributor group. 67.41% of the contributors belong to the casual type while more than 76.61% of the comments comes from the few serious contributors. As demonstrated in [14, 23], there may be ways to refine the contributor categorization using various structural attributes (e.g., the temporal patterns of comment submission, the nature of sites evaluated). However, we note that an activity-based categorization scheme does serve the research questions we pursue in this paper. In addition to comparing the characteristics of the casual (u0) and serious (u5) contributors, we include also the results of comparing the combinations of u0+u1 (less active members) and u4+u5 (highly active members) whenever suitable. We expect the combination of u4 and u5 to represent those who have the privilege of using the mass rating tool.

Coverage: Complementary Attention and Concern

We first analyze the attention and concern by different contributor groups judging from the comments they have given.

Attention Divide on Goodness and Badness of Sites

Figure 2 (right) depicts the percentage breakdown of comments given by different contributor groups in each comment category. Notice that the first 5 categories are positive in nature, followed by 11 others that are negative, as classified in Table 1. We made use of the positive or negative nature of a comment category to determine the contributor's attention

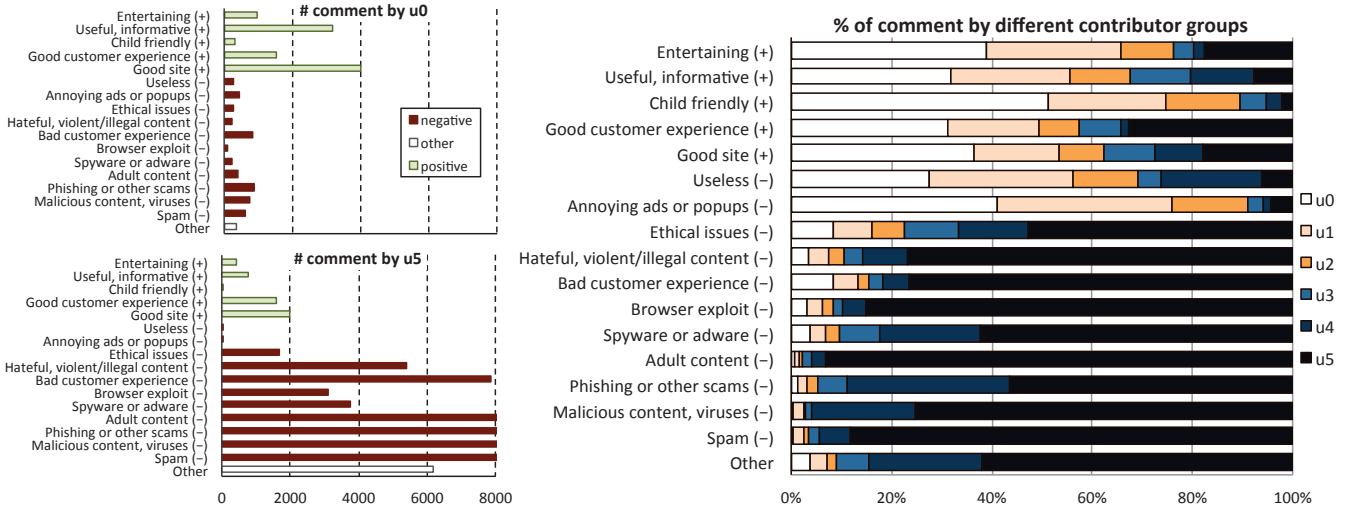

Figure 2. (Left) Total comments given by u0 and u5 in each comment category with the horizontal bars truncated at value=8000. (Right) Percentage breakdown of comments given by different contributor groups. + and – denote the positive and negative nature of a comment category respectively.

on the goodness or badness of a site. An interesting finding is that a large percentage of attention or concern on the goodness of sites (i.e., whether they are entertaining, useful, child friendly, offers good customer experience, or good) actually comes from the less active contributors (especially u0). Conversely, other than the 'Useless' and 'Annoying ads or pop-ups' categories, a large percentage of comments among the negative categories actually come from the highly active contributors (especially u5). These include the attention on the technical security of sites (e.g., whether a site contains malware, spyware, browser exploits, or whether a site is related to spamming, phishing or scams) as well as the attention on adult and other potentially inappropriate content. While this may be largely due to the fact that the serious contributors have been rating and commenting a large number of sites based on some blacklists they maintain or reference to, the distinctive divide on the attention for the goodness and badness of sites does highlight the role of the casual contributors.

The finding is consistent when we look at the ratio of positive versus negative comments that have been given by the casual and serious contributors respectively. As shown in Figure 2 (left), the casual contributors (u0) are indeed more inclined to comment about the good aspects of a site, different from the serious contributors (u5) who have produced much more negative comments versus the positive ones. Next, we quantify the roles of serious and casual contributors in covering for sites of specific nature in Table 3. Specifically, we measure the loss of coverage should a particular contributor is absent in the community. Most notable is that, without the inputs from the highly active contributors (u4 and u5), 92.77% of the 473,273 potentially bad sites would have gone undetected. Meanwhile, the loss of coverage for potentially bad sites should u0 and u1 are absent is 2.55%.

While it may appear that the casual contributors provide little value to web security, we argue the reverse is true as they enable a system like WOT to signal against sites that are good from those that have not been evaluated. Indeed, without the less active contributors (u0 and u1), 50.1% of potentially good domains (i.e., those that have received at least a positive comment in the our dataset) would have been given an 'unknown' status. This is important, as attackers tend to leverage on a large volume of bad sites to thin the defenders' resources. Given an adequate coverage of good sites, users who are conservative on web security can regard sites with an unknown status as potentially questionable.

Attention Divide for Popular Sites and The Long Tail

Among sites that have been attended to by more than one contributors, we find that 4.91% of these sites were in fact first discovered (first commented) by a member of either u0 or u1. This is close to the 4.69% loss of coverage on all kinds of sites (as shown in Table 3) should we ignore the inputs from u0 and u1 contributors. The corresponding figures considering the u0 contributors only are 1.98% and 2.16%.

Note that however the above figures are computed based on the total number of sites covered in DS-Comment, which contains a disproportionately large share of bad sites (93%) typically found in the long tail of the web popularity. The value of the casual contributors will be larger should we consider only sites that are more relevant for daily browsing. Not surprisingly, we find that only 3.4% of the sites evaluated by the u5's comments appear on Alexa's list of top million most visited sites. On the other hand, 51.9% of the sites attended to by u0 are among the top million most visited sites. The corresponding figure is 29.6% for u1 and 4.3% for u4 respectively. Considering only those that appear on Alexa's list, the mean traffic ranking of sites attended to by u0 is lower than that of attended to by u5 ($p<.01$). The mean traffic ranking comparing u0+u1 to u4+u5 is also significantly lower ($p<.01$).

While serious contributors identify most of the bad sites in the long tail of web popularity, we note that the coverage for the more popular sites by the casual contributors is equally important. Following our earlier argument, coverage for known sites would allow the users to be correctly cautious about un-

	Comment category	# comments	# unique sites	Loss of coverage (%) without u0	without u0 & u1	without u4 & u5	without u5
	All	600,000	504,874	2.16	4.69	89.60	75.53
	Positive only	29,018	24,697	31.04	50.10	27.01	17.79
	Negative only	561,006	473,273	0.80	2.55	92.77	78.61
positive	Entertaining	2,496	2,219	36.64	62.33	21.54	19.51
positive	Useful, informative	9,985	8,841	29.51	51.85	21.66	7.86
positive	Child friendly	607	577	49.39	73.31	5.37	2.25
positive	Good customer experience	4,948	4,595	29.36	46.49	36.28	34.86
positive	Good site	10,982	9,999	33.69	50.06	28.58	18.43
negative	Useless	950	929	26.37	55.33	26.91	6.46
negative	Annoying ads or popups	1,040	905	37.46	72.71	5.75	4.64
negative	Ethical issues	3,149	3,038	7.83	14.91	67.08	53.32
negative	Hateful, violent or illegal content	7,044	6,927	2.86	6.63	86.73	77.83
negative	Bad customer experience	10,300	10,117	7.44	12.32	82.67	77.38
negative	Browser exploit	3,660	3,609	3.02	5.99	89.86	85.23
negative	Spyware or adware	6,052	5,902	3.49	6.25	82.68	62.15
negative	Adult content	53,879	52,856	0.60	1.19	95.99	93.74
negative	Phishing or other scams	65,259	58,011	1.29	3.13	88.01	54.67
negative	Malicious content, viruses	171,175	139,711	0.47	1.84	95.19	72.82
negative	Spam	238,498	210,529	0.26	2.13	93.89	87.12
	Other	9,976	9,861	3.31	6.45	85.16	62.29

Table 3. Total comments and unique sites evaluated per comment category, and the loss of coverage in the absence of casual or serious contributors.

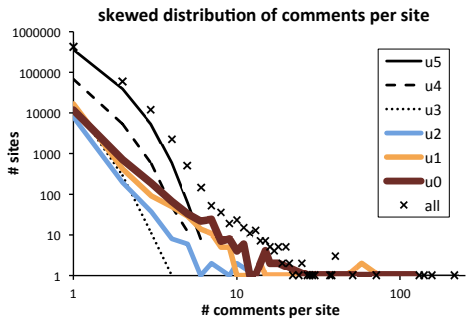

Figure 3. Distribution of comments per site considering all contributors ($\times$ markers) and each contributor group separately (black/color lines).

rated or unknown websites. Evaluations for the more frequently visited sites are also of a higher relevance and can thus be of a higher value. There is however a potential pitfall. Sites from well-known vendors should not be unnecessarily occupying the attention of the contributors. We look into the issue of efficiency and redundancy in the next section.

Coordination: Redundancy versus Efficiency

Figure 3 plots the distribution of comments per site considering all contributors together (depicted by the $\times$ markers), and considering the individual contributor groups separately (depicted by the black and color lines). An interesting observation is the heavy tail of the overall distribution of comments. This is largely due to the highly redundant coverage given by the less active contributors (u0 and u1). Notice that both the u0 and u1 lines exhibit also a long tail. In particular, a number of sites have received tens of comments from u0 alone. The redundancy is much larger in practice given that our dataset represents only a 5% sample of all available comments.

While a certain degree of redundancy is needed to ensure a reliable assessment outcome (the law of large number), excessive redundancy indicates inefficiency. It may be reasonable to expect controversial sites to receive more attention than the others. We examine the cases where u0 contributors have given more than one comments to a site, and measure the controversy of a site as $1 - |n_i - p_i|/(n_i + p_i)$ with n_i and p_i denoting the number of negative and positive comments given to site i respectively. However, we find only a low correlation (r=0.08, p<.01) between the number of comments of a site and its level of controversy. Indeed, the top most commented sites by u0 (and their controversy level) include WOT's own website, mywot.com (0.17) and other well-known sites such as google.com (0.38), facebook.com (0.95), youtube.com (0.66) and mail.google.com (0.10). This suggests the potential to coordinate the casual contributors for a higher efficiency.

The distributions of comments given by u4 and u5 seem to follow a different trend. A large number of sites evaluated by the serious contributors (u5) have actually received only one or two comments (in DS-Comment). While it may appear that there is an implicit coordination, we find that the three most common issues (spam, phishing and malicious sites) are actually attended to by 27 out of the 30 serious contributors. 15 of them have used the blacklists on malwaredomains.com for malicious sites, while 14 have referred to joewein.net for spamming activities. While further investigation is necessary, there may be also some room to better coordinate the volunteering efforts by the serious contributors.

Reliability and Verifiability

Thus far, we have studied various characteristics of the comments given by different contributor groups but we have yet to consider the reliability of their inputs. We would expect some of the comments (and ratings) to be invalid due to errors or potentially gaming behaviors. To evaluate the validity of the individual inputs, we first work out the *true risk status* of different sites using the dataset DS-Reliability, which contains the assessment outcomes by WOT, SiteAdvisor (SA) and Safe Web (SW) on 5,000 sites randomly selected from DS-Comment. This is however not a straight forward task; the assessment outcomes of different services are known to

Contr. group	Among sites identified as bad by WOT		Among sites identified as bad by WOT & SA		Among sites identified as good by WOT		Among sites identified as good by WOT & SA	
	# c (–)	# c (+)	# c (–)	# c (+)	# c (–)	# c (+)	# c (–)	# c (+)
u0	30	34	1	1	15	50	15	48
u1	99	7	34	-	11	28	11	28
u2	32	4	7	-	8	27	7	27
u3	107	3	20	-	3	18	3	15
u4	698	1	326	1	1	20	1	16
u5	3,826	4	1,232	-	9	41	7	38

Table 4. Error rate of different contributor groups in assessing bad and good sites comparing the number of positive (+) and negative (–) comments to (i) the sole assessments by WOT, and (ii) the common assessment outcomes (bad/good) of WOT and SiteAdvisor (SA). Texts in red denote the counts of false-negative or false-positive cases accordingly.

Contr. group	% comments with web link	mean comment length (# char)
u0	3.10	91
u1	8.07	118
u2	10.07	76
u3	26.56	99
u4	64.29	108
u5	49.38	138

Table 5. Percentage of comments containing a web link, and average comment length (in terms of the number of characters) excluding comments containing non-Latin characters.

be disagreeing with each other [5]. We map the assessment outcomes of SA (Green, Yellow, Red, Gray) and SW (Safe or VerisignTrusted, Caution, Warning, Untested) to the default risk signals of WOT (Green: good, Yellow: caution, Red: bad, Gray: unknown). We find that, among the 302 good sites identified by WOT, a majority of them receive the same verdict from SA and SW respectively. However, out of the 4544 sites identified as bad by WOT, 1230 are co-identified as bad by SA while only 47 sites are warned by SW. The large discrepancy between SW and WOT can be attributed to the extremely low coverage of SW (29%) on the 5000 sites in our test sample. This leads us to ignore the assessments from SW in the subsequent analysis. On the other hand, SA (with a coverage of 78%) has come short in evaluating sites with an IP address and those hosted on shared domain or free hosting services. Another factor contributing to the discrepancy between SA and WOT is the larger evaluation scope of WOT. For example, SA does not evaluate the vendor reliability aspect as WOT does. For these reasons, to study the reliability of different contributor groups, we approximate the true risk status of sites based on (i) the aggregate assessment outcomes by WOT alone, (ii) the common outcomes of WOT and SA.

Reliability in Evaluating Good and Bad Sites

Table 4 shows the number of positive and negative comments given by different contributor groups that match the assessments by WOT alone, and that match the common verdicts by WOT and SA. Note that we have excluded comments with the 'Adult content', 'Child friendly', 'Hateful, violent or illegal content', 'Ethical issues' and 'Entertaining' categories given that both SA and the default risk signaling strategy of WOT do not evaluate content appropriateness or fun level.

There are several interesting findings here. First, notice that among sites that have been co-identified as bad by WOT and SA, there are only two positive comments wrongly made for these sites (see Table 4, 5th column). A similar trend can be observed for sites that have been identified as bad by WOT (a superset of the previous case); the ratio of positive comments (error rate) is small except for the case of u0 (see Table 4, column 2-3). Here, the casual contributors (u0) could be misinformed about the badness of the sites or attempting to game the aggregate outcomes. Either way, the large error rate suggests the limitation of the casual contributors as a whole in assessing bad sites reliably. On the other hand, the reliability of serious contributors in assessing bad sites is applaudable. In fact, u5 has found many more bad sites than SA.

Next, we look at the reliability of different contributor groups in assessing the good sites. Notice that there is a higher error rate (the ratio of negative to positive comments) in general. Indeed, labeling a site as good involves a higher level of subjectivity. Different from the objective assessment on whether a site is malicious, is a phishing site and so on, there is also a lack of well-defined terminologies in general to measure the good properties of a site. Interestingly, the error rate does not differ much across different contributor groups. To be exact, the difference in the ratios of positive to negative comments given by u0 and u5 is not statistically significant considering sites evaluated as good by WOT (Table 4, column 6-7) (Fisher's exact test, p=0.64), as well as sites co-evaluated as good by both WOT and SA (column 8-9) (p=0.34). The casual contributors are thus not inferior to the serious contributors when it comes to evaluating a good site correctly. We look into the error cases by u5 and find that 4 out of 9 false positive comments have actually been removed from the scorecards of the related sites.

Verifiability: Reference and Comment Length

Table 5 shows the percentage of comments that come with at least a URL link. While it is not always the case, URLs in the user comments often lead to some specific resources (e.g., further discussion) or references (e.g., to some online blacklists) where the contributors have become aware of the evaluated sites. We use the presence of a URL as an estimator of the verifiability of a comment. Notice that only 3% of the comments given by the casual contributors (u0) contain a URL. At the same time, only 49% of the comments given by u5 potentially contain a reference URL, typically pointing to a blacklist provided by, for example, joewein.net, cert.at, uribl.com, atma.es, malwareurl.com, spamtrackers.eu and malwaredomains.com. Also given in Table 5 is the mean length of comments provided by different contributor groups, excluding comments containing some non-Latin characters. The mean comment length increases going from casual to serious contributors; however, the increment is not statistically significant. These findings do signal the need of actions from WOT to improve the verifiability of user inputs. We outline some potential pitfalls and suggestions in the following.

DISCUSSION

Complementary Roles in Web Security

An important lesson learnt from our study is the complementary roles of casual and serious contributors for community-based web security. Contrary to the skepticisms that security

is out of reach for ordinary users given that it is a highly specialized domain requiring expert knowledge, our work shows that the casual contributors can be helpful in differentiating the good and known sites from those that have yet to be evaluated. Availability of such a 'whitelist' is valuable considering the large number of bad and gray sites created daily. In addition, while serious contributors may be sharp in evaluating the badness of a site (given the access to some reliable blacklists and expert knowledge on malicious activities on the web), their judgment on good sites (subjective) is not significantly better than the less active contributors.

Applicability to Other Contexts

The complementary roles we find in this paper are probably unique to web security where conventional approaches are being overloaded with a large number of new sites, and where there is a need for subjective judgment (where personal experience matters) and objective evaluation (where expert knowledge is required) on different aspects of sites. While the exact complementarily may not apply to other collaborative systems directly, the finding that different members play different roles and exhibit different potentials should be capitalized by community-based systems across different domains. Leveraging on the different roles and natures of tasks, we outline several design implications relevant to WOT and community-based systems more generally in the following.

Design Implications

Context based Reliability

The ability to gauge and make use of the reliability of a contributor is an important building block to many community-based systems. WOT currently weighs the user inputs differently based on the reliability of individual contributors when computing the aggregate outcomes. This provides an incentive for the community members to contribute responsibly. Nevertheless, the actual formula used is hidden (arguably to mitigate potential gaming behaviors). Our findings that different contributor types attend to sites of different natures and realize different levels of reliability in evaluating bad and good sites, raise several important issues in designing the reliability weighting mechanism. First, should the weighting be computed at per contributor or per contributor-and-site-category level? We argue that the latter would be more appropriate. Specific to WOT, a serious contributor who has been consistently giving reliable evaluations on potentially malicious sites should not be automatically given a heavy weight when, for example, evaluating the goodness or content appropriateness of a site. Another issue lies with the subjective evaluation aspects that WOT and many other reputation systems actually deal with. Does the reliability weighting punish those who may have a different expectation and opinion than the majority on a subjective aspect? This is a tricky matter which further highlights the need to acknowledge the differences across multiple evaluation aspects: subjective or objective, requiring expert knowledge or not, and so on.

Verifiability of Objective and Subjective Evaluation

A way to increase the reliability of a community-based system is probably by improving the verifiability of user contributions. WOT details on the use of inputs from third party sources (if any) on the scorecard of each evaluated site, but the community inputs seem to be lacking in verifiability currently. WOT requires the users of the mass rating tool (the highly active members) to include a comment describing the reasons of their ratings and to be always contactable on mywot.com. However, our analysis shows that only 49% of the comments provided by the serious contributors do potentially contain a reference URL. The percentage is much lower among the comments given by the casual contributors. These suggest a lack of verifiability that could affect user confidence in the long run. We suggest putting in place a referencing system for objective evaluation (e.g., on whether a site is malicious) and a more structured process when eliciting subjective evaluation. For example, requiring the mass rating tool users to always include the supporting reference will help especially in the cases of false positives. This will also restrict the tool from being wrongly used on aspects that are subjective and objectionable in nature. On the other hand, casual contributors who we find to be more likely to attend to subjective aspects such as the goodness of a site, should be guided to detail on their personal experience in a more structured manner. This can include indicating if they are affiliated with the site, how frequently they visit it, how does it matches a list of keywords, and so forth. The use of a referencing system and a structured way of eliciting subjective inputs are not new. However, different from the Wikipedia and many other systems that deal with either objective or subjective contributions solely, WOT exemplifies the case where both methods will be needed at the same time to cater for a mix of objective and subjective evaluations.

Role based Coordination and Socialization

Reliability issues aside, under-provisioning is perhaps the most challenging problem in community-based systems. Our study yields interesting insights on how we can better coordinate and socialize the contributors depending on the roles they play in the community. Currently, WOT does allow the site owners to reach out to the community members and call for their assessments. We note that it could be interesting to automatically distribute these requests to selected highly active members with the necessary skill sets and experience, similar to SuggestBot presented in [7]. WOT can also consider introducing a more explicit community structure (e.g., establishing sub-communities that would specialize on certain objective aspects e.g., privacy and malicious contents) such that a core team of contributors could help to set directions and guide the new contributors (as proposed for Wikipedia in [13]). On the other hand, the heavy tailed attention distribution among the less active contributors (as shown in Figure 3) hints on the possibility of coordinating the ordinary members to increase productivity. While there may be a privacy issue, WOT could innovate on an *opt-in* feature that would automatically suggest to the casual contributors to rate the unevaluated sites that they have been visiting. Generally, these should be done with care as certain socialization tactics may adversely turn away the contributors [6]. All in all, coordination and socialization efforts should be done with a good understanding on the different roles and potentials of different contributors.

CONCLUSIONS

We have found the interesting complementary roles of serious and casual contributors in Web of Trust (WOT). Serious contributors play an important role in reporting most of the malicious sites while casual contributors provide a large percentage of attention to the goodness of sites. Although the casual contributors do not evaluate malicious sites extensively and reliably, their evaluations on the good sites are valuable as they enable WOT to differentiate the good and known sites from those that have yet to be evaluated accordingly. This helps to steer users away from the numerous bad sites created daily. In addition, while serious contributors give reliable evaluations on bad sites, their evaluations on good websites are not significantly more reliable than the casual contributors. While the complementarity we find in this paper may be specific to web security, the finding that different community members contribute in different roles and exhibit different potentials in different tasks should be better capitalized by community-based systems across different domains.

ACKNOWLEDGMENTS

This research was supported in part by the National Science Foundation under award CCF-0424422 (TRUST). We are grateful to the anonymous reviewers for their constructive comments. We thank also the WOT developers for providing us two valuable datasets.

REFERENCES

1. Alexa Top Million Sites. http://www.alexa.com/topsites.
2. Anti-Phishing Working Group. Global Phishing Survey: Trends and Domain Name Use in 2H2010. http://www.antiphishing.org/reports/APWG_GlobalPhishingSurvey_2H2010.pdf.
3. Antin, J., and Cheshire, C. Readers are not free-riders: reading as a form of participation on Wikipedia. In *Proc. CSCW*, ACM (2010), 127–130.
4. Ayyavu, P., and Jensen, C. Integrating user feedback with heuristic security and privacy management systems. In *Proc. CHI*, ACM (2011), 2305–2314.
5. Chia, P. H., and Knapskog, S. J. Re-evaluating the wisdom of crowds in assessing web security. In *Proc. Financial Cryptography and Data Security (FC)*, Springer (2011).
6. Choi, B., Alexander, K., Kraut, R. E., and Levine, J. M. Socialization tactics in Wikipedia and their effects. In *Proc. CSCW*, ACM (2010), 107–116.
7. Cosley, D., Frankowski, D., Terveen, L., and Riedl, J. Suggestbot: using intelligent task routing to help people find work in wikipedia. In *Proc. IUI*, ACM (2007), 32–41.
8. Denning, P., Horning, J., Parnas, D., and Weinstein, L. Wikipedia risks. *Communications of the ACM 48*, 12 (2005), 152.
9. Edelman, B. Adverse selection in online "trust" certifications and search results. *Electronic Commerce Research and Applications* (2010).
10. Facebook partners with WOT to protect its users. http://www.arcticstartup.com/2011/05/12/facebook-partners-with-wot-to-protect-its-700-million-users.
11. Geiger, R. S., and Ribes, D. The work of sustaining order in Wikipedia: the banning of a vandal. In *Proc. CSCW*, ACM (2010), 117–126.
12. Kittur, A., Chi, E., Pendleton, B. A., Suh, B., and Mytkowicz, T. Power of the few vs. wisdom of the crowd: Wikipedia and the rise of the bourgeoisie. *Algorithmica 1*, 2 (2007), 1–9.
13. Kittur, A., and Kraut, R. E. Harnessing the wisdom of crowds in Wikipedia: quality through coordination. In *Proc. CSCW*, ACM (2008), 37–46.
14. Mamykina, L., Manoim, B., Mittal, M., Hripcsak, G., and Hartmann, B. Design lessons from the fastest Q&A site in the west. In *Proc. CHI*, ACM (2011), 2857–2866.
15. McAfee SiteAdvisor. http://www.siteadvisor.com.
16. Moore, T., and Clayton, R. Evaluating the wisdom of crowds in assessing phishing websites. In *Proc. Financial Cryptography and Data Security (FC)*, Springer (2008).
17. Norton Safe Web. http://safeweb.norton.com.
18. Ortega, F., Gonzalez-Barahona, J. M., and Robles, G. On the Inequality of Contributions to Wikipedia. In *Proc. HICSS*, IEEE Computer Society (2008).
19. PhishTank. http://www.phishtank.com.
20. Provos, N., Mavrommatis, P., Rajab, M. A., and Monrose, F. All your iframes point to us. In *Proc. Security*, USENIX (2008), 1–15.
21. Verisign Domain Name Report. http://www.verisigninc.com/en_US/why-verisign/research-trends/domain-name-industry-brief/index.xhtml.
22. Web of Trust. http://www.mywot.com.
23. Welser, H. T., Cosley, D., Kossinets, G., Lin, A., Dokshin, F., Gay, G., and Smith, M. Finding social roles in wikipedia. In *Proc. of the 2011 iConference*, ACM (2011), 122–129.
24. Wilkinson, D. M. Strong regularities in online peer production. In *Proc. of the 9th ACM Conference on Electronic commerce (EC)*, ACM (2008), 302–309.
25. 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. of the 9th Workshop on the Economics of Information Security (WEIS)* (2010).
26. 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. of the 7th Workshop on the Economics of Information Security (WEIS)* (2008).