Pern Hui Chia

Information Security on the Web and App Platforms

An Economic and Socio-Behavioral Perspective

Thesis for the degree of Philosophiae Doctor

Trondheim, November 2012

Norwegian University of Science and Technology
Faculty of Information Technology,
Mathematics and Electrical Engineering
Department of Telematics

NTNU – Trondheim
Norwegian University of
Science and Technology

NTNU
Norwegian University of Science and Technology

Thesis for the degree of Philosophiae Doctor

Faculty of Engineering Science and Technology
Department of Engineering Design and Materials

ISBN 978-82-471-3969-1 (printed ver.)
ISBN 978-82-471-3970-7 (electronic ver.)
ISSN 1503-8181

Doctoral theses at NTNU, 2012:324

Printed by NTNU-trykk

Preface

This dissertation is submitted in partial fulfillment of the requirement for the degree of Philosophiae doctor (Ph.d.) at the Norwegian University of Science and Technology (NTNU). The work was performed at the Centre for Quantifiable Quality of Service in Communication Systems (Q2S), and has been supervised by Professor Svein J. Knapskog. Centre for Quantifiable Quality of Service in Communication Systems (Q2S), Centre of Excellence, is established and funded by the Research Council of Norway, NTNU, UNINETT and Telenor.

This work has also benefited from fruitful research visits to well known research groups during the doctoral program. These include a three-week visit to Nokia Research Center (Helsinki) to work with Dr. Andreas P. Heiner and Dr. N. Asokan in August 2009, a six-month stay at the School of Information, University of California (Berkeley) to work with Professor John Chuang in Spring 2011, and a two-week invited visit to Carleton Computer Security Lab at Carleton University, headed by Professor Paul van Oorschot, in May 2012.

Acknowledgements

There are many people to whom I owe my gratitude throughout the doctoral program. First and foremost, I have been very fortunate to have Professor Svein J. Knapskog as my advisor, a sincere and knowledgeable person whom I always respect and admire. I am grateful for his patience and tireless efforts in guiding my research, and the flexibility and trust he has given me to explore my interests, to attend relevant conferences, and to go for international research visits.

I am deeply grateful to Dr. N. Asokan whom I first met when I was a trainee at Nokia Research Center. I am fortunate to have continually received good advice and help from him throughout these years. I am also very grateful to Professor John Chuang who has been very kind to host my research visit to University of California, Berkeley. I appreciate his patience in guiding me. I still remember how our weekly discussion frequently overshot the pre-allocated time, only to be reminded by the bells on Sather tower. I would like to take the opportunity to thank Professor Paul van Oorschot for inviting me for a short visit to Carleton University where I have met and exchanged ideas with a number of good researchers.

I would like to thank all my co-authors: Benedikt Westermann, Yusuke Yamamoto, Georgios Pitsilis, Andreas P. Heiner, Yanling Chen, Gergely Biczók, John Chuang, N. Asokan, Svein J. Knapskog, for the expertise and patience in making our joint papers successful. There are many more researchers and colleagues whom I owe my gratitude to. It is impossible to name all of you here, but I would like to let you know that I always look up to many of you.

I will not forget the valuable friendships I have been blessed with during these years. I thank you all for the encouraging words and for being there for support and listening. Several special people have particularly lightened up my days. I wish all of you happiness and a bright future whichever path you will pursue and wherever you will be.

I have been away from home for more than a decade. I sincerely apologize to my family members for not being there with you in many occasions. Thank you to my parents and siblings for your understanding and unwavering support. Despite the majestic fjords, pure-white snow and fresh salmons in Norway, you can be assured that I think of you always.

Abstract

Various security measures are ineffective having been designed without adequate usability and economic considerations. The primary objective of this thesis is to add an economic and socio-behavioral perspective to the traditional computer science research in information security. The resulting research is interdisciplinary, and the papers combine different approaches, ranging from analytic modeling to empirical measurements and user studies. Contributing to the fields of usable security and security economics, this thesis fulfills three motivations.

First, it provides a realistic game theoretical model for analyzing the dynamics of attack and defense on the Web. Adapted from the classical Colonel Blotto games, our Colonel Blotto Phishing model captures the asymmetric conflict (resource, information, action) between a resource-constrained attacker and a defender. It also factors in the practical scenario where the attacker creates large numbers of phishing websites (endogenous dimensionality), while the defender reactively detects and strives to take them down promptly.

Second, the thesis challenges the conventional view that users are always the weakest link or liability in security. It explores the feasibility of leveraging inputs from expert and ordinary users for improving information security. While several potential challenges are identified, we find that community inputs are more comprehensive and relevant than automated assessments. This does not imply that users should be made liable to protect themselves; it demonstrates the potentials of community efforts in complementing conventional security measures. We further analyze the contribution characteristics of serious and casual security volunteers, and suggest ways for improvement.

Third, following the rise of third party applications (apps), the thesis explores the security and privacy risks and challenges with both centralized and decentralized app control models. Centralized app control can lead to the risk of central judgment and the risk of habituation, while the increasingly widespread decentralized user-consent permission model also suffers from the lack of effective risk signaling. We find the tendency of popular apps requesting more permissions than average. Compound with the absence of alternative risk signals, users will habitually click through the permission request dialogs. In addition, we find the free apps, apps with mature content, and apps with names mimicking the popular ones, request more permissions than typical. These indicate possible attempts to trick the users into compromising their privacy.

Contents

1. Introduction

While much focus has been given to technological advancement, security remains a challenging problem impacting billions of users. Truth is that information security is a multidisciplinary problem. Without a comprehensive view combining the technical, social, behavioral and economic aspects, security measures will fail to serve their purposes in practice.

This thesis contributes to some missing pieces in information security research, particularly from an economic and socio-behavioral perspective. Six papers of analytic modeling, empirical, and experimental natures are included. The thesis will first present the background and related work in Section 1.1, research motivations and methodology in Section 1.2, before describing the included papers and elaborating the contributions in Section 1.3.

1.1. Background

The last decade has seen an exciting development in security research. Researchers are starting to realize the importance of usability for a security measure to be effective. There is also a growing attention on the economic aspect of security problems. This section presents the background and a survey of related works in the two expanding fields of usable security and security economics.

1.1.1. Usable Security

The user is a central aspect of computer security. Strong cryptographic mechanisms and secure protocols must be accompanied by an easy-to-use interface and procedure. Putting unreasonable requirements on the users will risk user mistakes or compelling them to embrace convenient but insecure behaviors. More than ease-of-use, user mental models, behavioral biases and social norms are among the topics of interest of the research community.

Usability

A written requirement of usability for good security can be dated back to Auguste Kerckhoffs's article [75] in 1883. Today, Kerckhoffs is widely known for the principle that a cryptosystem must be secure even if everything about the system, except the key, is public knowledge. Many do not realize that Auguste Kerckhoffs has in fact pointed also the importance of usability [72, 103]. In particular, the sixth principle in Kerckhoffs's article states that a cryptosystem must be easy to use and

must neither require stress of mind nor the knowledge of a long series of rules (as translated by Fabien Petitcolas [98]).

The importance of usability is also highlighted in the influential paper by Saltzer and Schroeder (1975) [106]. The authors identify eight design principles for information protection in computer systems, namely, economy of mechanism, fail-safe defaults, complete mediation, open design, separation of privilege, least privilege, least common mechanism, and psychological acceptability. Psychological acceptability describes the importance of a human interface that is easy to use, and the matching of user's mental image of his protection goals and the mechanisms he must use.

Adam and Sasse [8] observed in 1999 that requiring users to remember several strong passwords and to change them periodically, has led to excessive cognitive strain and the situation where users would simply write the passwords in plaintext beside their computers. Research on the intersection of security and usability has since begun to gain momentum. An area of interests is on improving the usability and security of authentication scheme. Innovations we have seen include federated (single-sign-on) authentication (e.g., OpenID [101], Microsoft Passport), graphical passwords (see a survey of different proposals in [110]), password managers (e.g., on Firefox), and biometrics (e.g., fingerprint, voice).

Yet, despite the numerous innovations, password has remained the most pervasive authentication scheme in practice. Surveying a wide range of web authentication technologies, Bonneau et al. [27] find that no alternative schemes are currently dominant over the traditional passwords, considering the combination of usability, deployability and security perspectives. By usability, the researchers evaluate if an authentication method is memorywise-effortless, scalable-for-users, nothing-to-carry, physically-effortless, easy-to-learn, efficient-to-use, has infrequent-errors, and can be easy-recovered-from-loss. The researchers find that while certain technologies (e.g., federated login schemes) offer a better usability than legacy passwords, they are less easy to deploy. Many other schemes offer a better security than passwords but are more difficult to use or more costly to deploy. The authors note that many academic proposals have failed to gain traction given that researchers rarely take into account a sufficiently wide range of real-world constraints [27].

Improving the usability of a security measure is indeed not a straightforward process, nor it is a standalone problem. Over the years, a number of studies have evaluated the usability of various security technologies, including the use of PGP for email encryption [118], privacy control with peer-to-peer file sharing [62], secure bluetooth pairing [113], and secure identity management [73].

Risk Communication

Two schools of thought in managing security risks are security by designation and security by admonition [123]. Security by designation builds on the belief that user actions simultaneously express command and extension of authority. Authority can thus be inferred and granted to a system through users' conscious actions, while execution of insecure actions being prohibited altogether. Security by admonition,

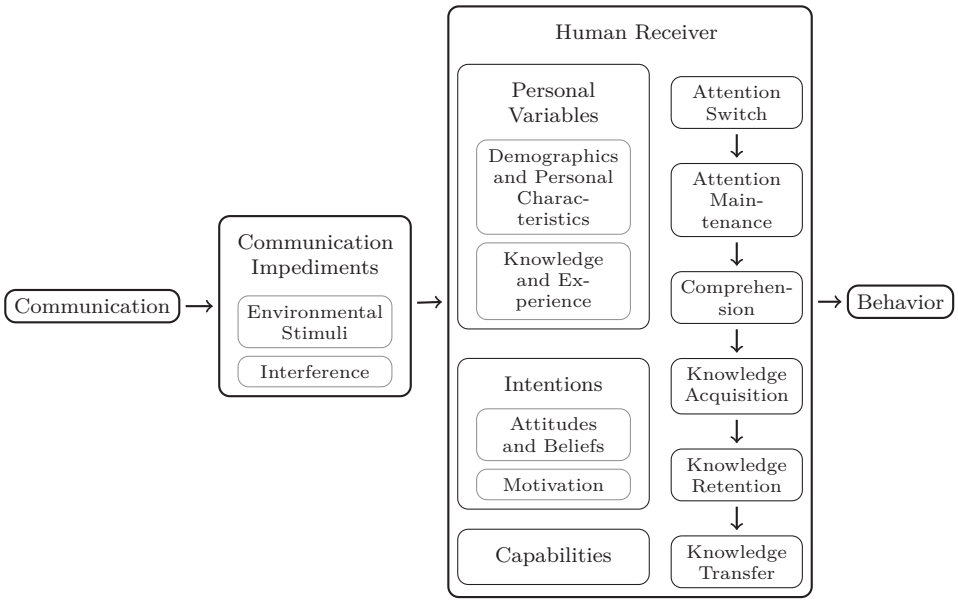


Figure 1.1.: The human-in-the-loop security framework provided by Cranor [44] for reasoning about the cause of security failures attributable to human errors.

on the other hand, disrupts user attention to a secondary source of information such as warning and confirmation dialogs to request for an extension of authority.

While better in usability, security by designation is not always applicable, for example, when inter-operating with another untrustworthy system, or when there are too many fine-grained user actions to consider [123]. Security by admonition can help prompt users about the imminent danger or risk in these situations. However, the secondary source of information such as warning and confirmation dialogs, is often inadequate and not context-aware. The frequent use of admonition dialogs and the relative rare occurrence of insecure events have also led to a high level of false positives. This may in turn cultivate user habituation to ignore and click-through the warning and confirmation dialogs. Given the pervasive reliance on security by admonition today, risk communication is thus an important research area in usable security.

There are plentiful examples of unsuccessful risk signaling in the literature. Wu et al. [122] observe that users fail to notice or act upon risk signals from security toolbars. Schechter et al. [107] find that users also ignore HTTPS indicators and site-authentication images – cues designed to mitigate phishing threats. At the same time, research [61, 24] has found that users click-through the End User License Agreement (EULA) and similar user-consent dialogs.

Cranor [44] presents the human-in-the-loop framework for investigating and reasoning the root cause of security failures that have been attributed to human er-

rors. As depicted in Figure 1.1, the framework examines different phases of risk communication beginning from (i) the type of communication (warnings, notices, status indicators, training or policies), (ii) potential communication impediments (environment stimuli or interference), (iii) the human receiver, which depicts how capabilities, intentions and personal variables interact with information processing by the human receiver, and finally (iv) the induction of a desired behavior. The information processing steps in this framework are adapted from the well-known Communication-Human Information Processing (C-HIP) model by Wogalter [120] in the warning science literature. The adapted information processing model consists of six component steps: attention switch, attention maintenance, comprehension, knowledge acquisition, knowledge retention and knowledge transfer.

The human-in-the-loop framework reiterates the importance of salient risk signals for attention switch and maintenance. Yet, the lack of human attention is only a part of the extensive set of potential failures. Lack of security knowledge among users has also been identified as a factor contributing to security problems (e.g., in phishing [47]). There have been initiatives to help users learn about security. Sesame [109] helps users make security related decisions using interactive system visualization. Meanwhile, Anti-Phishing Phil [108] teaches the users not to fall for online phishing through an interactive and entertaining game starring Phil – a small fish taking advices from his father.

1.1.2. Security Economics

To construct a framework for comparing various authentication schemes, Bonneau et al. [27] has correctly considered the usability, deployability and security as three central evaluation factors. Deployability says much about the economics of implementing a particular scheme. Indeed, an economic perspective on information security is essential because security measures come with a price. With a thriving underground economy, modern perpetrators are incentivized by financial gains; they are no longer mere hobbyists hackers. Information security problems also often arise due to misaligned incentives, externality and information asymmetry, three problems widely studied in economics [15].

The field of security economics was kick-started with the inauguration of the Workshop of Economics of Information Security (WEIS) in 2002. The annual event has since provided a common platform for computer scientists, economists, sociologists, industrial representatives and policymakers to come together and discuss various security problems from different perspectives. Research in security economics has thus far encompassed security incentives and interdependence analysis, investigation of the underground economy and modi operandi, as well as analytic modeling of optimal security investment and analysis on the feasibility of security insurance. A survey of notable works can be found in [16, 17]. The following describes several areas which have received much attention within this fast expanding field.

Misaligned Incentives, Network Externality, Asymmetric Information

Anderson argues in the seminal paper [15] that economics underlies the many security problems we have today:

> According to one common view, information security comes down to technical measures. Given better access control policy models, formal proofs of cryptographic protocols, approved firewalls, better ways of detecting intrusions and malicious code, and better tools for system evaluation and assurance, the problems can be solved. In this note, I put forward a contrary view: information insecurity is at least as much due to perverse incentives. Many of the problems can be explained more clearly and convincingly using the language of microeconomics: network externalities, asymmetric information, moral hazard, adverse selection, liability dumping and the tragedy of the commons. – Anderson [15]

Anderson gives multiple examples in the paper. First is the problem of misaligned incentives and liability dumping by banks in Britain, Norway and the Netherlands in 1990s. At that time, consumers in these countries would need to present proofs to dispute a fraudulent ATM transaction. It was different to the situation in the US where the burden of proof was on the banks. Given the lack of financial responsibility, banks in Britain, Norway and the Netherlands implemented less secure systems and suffered more frauds compared to the US counterparts [15].

Network externality presents another incentive problem. While users are probably willing to spend $100 for purchasing an anti-virus software, they are unwilling to spend $1 for a software that will prevent their computers from causing harm to others (e.g., becoming a bot and used to perform distributed-denial-of-service attack on some other systems) [15]. The absence of incentives to prevent damages external to the users matches the problem of 'the tragedy of the commons' [67] that is long known to economists. Regulatory actions are needed to remedy the problem of network externality [15]. Indeed, Lichtman and Posner, two law professors point out that the best way to mitigate the problem of botnets will be to hold the Internet Service Providers (ISPs) accountable [83]. They note that the ISPs are in the best position to fix the problem due to several reasons. First, direct liability on bad users, whose machines are bot-infested, is unsuitable as some of them would be out of reach of law (e.g., cross-border, or incapable to pay fines) [83]. Furthermore, it can be hard (costly) to expect the users to have the ability to ensure that their machines are clean [83]. On the other hand, the ISPs are the best liability intermediaries given that they can detect bot-infested machines and regulate user access to the Internet in addition to having the contact details of users [83]. Yet, holding the ISPs accountable may not be a straightforward task, especially at places where the risks of surveillance and excess centralized control are feared. Dealing with to what extent the ISPs should be held responsible and be given the power of control will certainly require efforts from the regulators.

Perverse incentives in security can also be attributed to asymmetric information. With a lack of user ability to distinguish between secure and insecure products (e.g.,

software, websites), there will be no incentives for companies to actually invest in security. This can lead to 'the market of lemons' – the scenario sketched by the well known economist George Akerlof in [9] – where bad (insecure) products in the market drive out the good ones eventually. Economic literature suggests to rely on certification intermediaries to approach both cases when the private information is unknown (i) ex-ante, and (ii) ex-post, a user action. Albano and Lizzeri show that if quality is endogenous, the existence of a certification intermediary will improve product quality [10]. If quality is exogenous, an intermediary will also improve welfare by not certifying unsafe products [10]. However, a monopolistic certifier will be keen to disclose only minimal information to induce trade [85].

In practice, we have seen numerous criticisms on security and trust certifications. Anderson [15] points out the faulty incentives with Common Criteria for IT Security Evaluation (CC) [6]. The CC framework is problematic as product evaluation is paid by the vendors rather than the potential users (e.g., the governments). This motivates the vendors to shop for the easiest path, either in terms of cost, strictness, or time, for certification [15]. Although the Commercial Licensed Evaluation Facility (CLEF) can have their licenses withdrawn, Anderson note that there is a lack of sanctions for misbehavior [15]. In addition, it will be wrong to equate a CC-certified product as secure. CC certification only says that a product has been evaluated to meet a set of security requirements and specifications, as documented in the Protection Profile (PP) and Security Target (ST), up to one out of seven different assurance levels. Lax requirements for certification can indeed lead to more harm than good. Edelman [48] reports the situation of 'adverse selection' with online trust certifications. He find that sites certified by a large vendor are in fact twice as likely to be untrustworthy as the uncertified sites [48]. Analogous to certification, sponsored advertisements on leading search engines are also found to be more than twice as likely to be untrustworthy as to the corresponding organic search results [48].

Given the challenges with third party certifiers, should we opt for mandatory regulations, for example, to have the government intervenes and enacts strict security and privacy protection standards? On online privacy protection, however, researchers have shown that when the expected loss due to privacy violation is moderate, mandatory regulation is not socially optimal [112]. Are there alternatives? Part of this thesis will investigate the feasibility of leveraging user inputs against the security and privacy risks on the Web and application platforms.

Are We Investing Enough?

As security risks grow, an important question we may ask is whether we have invested enough in security. How should companies approach an optimal investment in security measures? An answer to this is given by the well known Gordon-Loeb security investment model [63]. Assuming that an increasing security investment decreases the probability of security breach, but at a decreasing rate, Gordon and Loeb show that for two broad classes of security breach probability functions, the optimal security investment does not exceed 37% (=1/e) of the expected loss due

to a breach [63]. This calls for a thorough check on expensive security investment. Their model also shows when the vulnerability is high, it may not be optimal to continue to invest in protection. Security managers should in this case focus on reducing the expected loss. It is necessary to note that, however crisp and simple, the Gordon-Loeb security investment model does come with several limitations. The model assumes a zero fixed cost in security investment. In addition, it is not easy to determine the levels of threats and vulnerabilities as well as the value of the assets to be protected, so to work out the value of expected loss and optimal investment. Assuming that the expected loss is finite, their model is also not applicable to the protection of critical assets or infrastructures where a security breach will be catastrophic.

The golden rule of an 1/e upper limit for optimal security investment has been challenged in several subsequent publications. In particular, considering four classes of security breach functions with different characteristics of marginal security improvement, namely (i) decreasing, (ii) first increasing but later decreasing (logistic function), (iii) increasing, and (iv) constant, Hausken [68] shows that optimal security investment is not universally capped at 1/e. Depending on the security breach function, it may also be optimal to invest heavily to protect the extremely vulnerable information or system, opposed to the recommendation from Gordon-Loeb's model. Indeed, it remains an empirical question as to which (if any) of the security breach functions best captures the real world phenomenon.

In another extension work, Matsuura [89] introduces the concept of 'productivity space' of information security to model the fact that security investment can reduce both vulnerability and threat, making it harder or more costly for the attackers. This extends Gordon-Loeb's model which considers an exogenous threat level and that security investment reduces only the vulnerability level. However, as it is with deciding the best fitting vulnerability-driven security breach probability function, it is not straightforward how we should model the security threats and how they would be reduced with an increasing investment. Indeed, uncertainty can make a big difference in defender's optimal strategy. Böhme and Moore [25] show that under a high uncertainty about the security threats (e.g., costs of attacks), assuming the attacker will always go for the easiest or cheapest attack, it could be optimal for the defender to protect nothing (in a static game) or to have a wait-and-see reactive strategy (in a repeated game setting). This highlights the importance of information and to have a better understanding of the attack modi operandi. It also leads the controversial implication that security under-investment can in fact be a rational strategy, calling for the need to rethink the wide condemnation on seemingly lax security practices by the defenders [25]. While incentive misalignment often leads to security under-investment, it is not a necessary condition [25].

Can We Insure Security Risks?

Apart from deciding the optimal security investment, an idea that has captured the interests of many researchers in the field of security economics is on the viability of cyber insurance in improving information security. An early account on the

advantages of cyber insurance is given in Varian (2000) [114]. Varian envisions a two-step market approach in managing security risks. First, liability should be assigned to parties that have the best access to relevant manpower and technical resources for managing risks. For example, banks should be given the most of liability in ATM frauds although a small share of liability can also be assigned to users so that they will be careful. Secondly, as liability is straightened out, Varian argues that liable parties will no doubt want to buy insurance. This may seem counter-intuitive at first, but factoring in that insurers will only insure good clients, liable parties will be incentivized to comply to good security practices [114].

The conjecture that cyber insurance can improve information security has been echoed widely but there remains little uptake of the idea in practice. Several analytic works have highlighted the challenges. In [23] for example, Böhme shows that one particular challenge with cyber insurance, different from other insurance businesses, lies with the dominance of certain IT systems. This leads to the threat of tremendous correlated losses. Indeed, a virus infecting a client's system will hit many others at the same time, causing the business of cyber insurance to be particularly risky. Thus, Böhme [23] suggests that policies in support of cyber insurance should simultaneously consider supporting the diversification of IT systems. Apart from correlated cyber risks due to monocultures of IT systems, there is the problem of interdependent risks [80]. The security risks one faces depend on his and others' actions. The reward of protection and insurance thus depends on the security of other interconnected systems.

Yet, Lelarge and Bolot [80] show that in the presence of interdependent risks, insurance remains a viable scheme to incentivize users to adopt good security practices. This optimistic view is perhaps not shared by the majority of other modeling works, as surveyed by Böhme and Schwartz [26]. The authors find a discrepancy between the conjecture favoring cyber insurance as a tool for aligning incentives for good security practices, and the majority of analytical results challenging the viability of a market for cyber insurance. They conclude by calling for future works that will address the discrepancy so to advance the research of cyber insurance.

More than the Weakest Link

Security is often regarded as the problem of the weakest link – attackers will exploit the most vulnerable part of a security system. Yet, following the analysis by Hirshleifer on public provisioning [71], security researchers have started to realize the importance of an interdependency analysis in information security [115, 65]. Consider the case of a walled village, defending the village from the attackers is more than the weakest link problem. Depending on the underlying interdependency, the probability of successful defense can be modeled as a function of multiple forms:

- Weakest link – if successful defense depends on the lowest part of the wall
- Weakest target – if only the villager who has the lowest part of wall suffers
- Total effort – if the villagers build the wall together; the strength of the wall and thus successful defense depends on the combined effort of the villagers

- Best shot – if the villagers build multiple layers of walls; successful defense thus depends on the strongest layer

Game theory can be used to analyze the incentives of the villagers – whether they will contribute to the defense of the village. Knowing the equilibrium outcomes, a social planner (the village leader) can react to design a strategy that will incentivize the villagers to achieve the social optimum. Note that the above list is by no means exhaustive. Practical security scenarios can be a hybrid combination of the four security games or other relevant models. In addition, there may be occasions where we do not know the underlying interdependency structure; reverse-engineering from empirical data to reveal the structure is a potential direction for research [42].

We can already obtain some useful insights into various practical security scenarios with the above four security games. The weakest link game models the perimeter defense in network security; censorship resistance where one standing server defeats the attacker is an example of a best shot game; the strength of anonymity networks such as Tor which depends on the number of users can be modeled as a total effort game [42]. These security games can also model the case of secure software development. Given that the mistakes by any careless programmers can introduce vulnerabilities (weakest link) to the system, one should consider hiring fewer but better programmers [17]. At the same time, the best security architect available should be hired for designing the system, while more testers should be employed given the total effort nature of software testing in removing bugs and vulnerabilities [17]. Another application of the total effort game is given by Florêncio and Herley [55]. The authors argue that the password based authentication is a total effort game from the perspective of an attacker. While there remain many who will use an unsafe password such as the name of their pet, these users are spared from the attackers who must guess a large number of 'easy' passwords correctly in order to become profitable. In practical terms, the diversity of user passwords can thus be more important the strength of individual passwords [55].

On the other hand, the weakest target game, introduced by Grossklags et al. [65], can be used to model various types of Internet-scale attacks, such as phishing and drive-by downloads, in which the perpetrators set out to victimize not all, but the subset of the easiest targets or 'low hanging fruits'. Grossklags et al. [65] consider also the scenario where users are able to either protect themselves (through actions such as installing firewalls and regular software patching), or insure themselves to control the extent of losses (through actions such as regular backup and purchase of cyber insurance). In this setting, Grossklags et al. [65] find an important difference between the weakest link and the weakest target models. As the number of users increases, players will tend to protect themselves in the weakest target game, while the players will shift from protection actions to insure themselves in the weakest link game [65].

Underground Economy and Modi Operandi

We often hear astronomical figures of cyber crime profitability and security losses. In 2009, the chief security officer of AT&T testified to the US congress citing the

global cyber crime revenues to be more than $1 trillion per annum [14]. To put in perspective, $1 trillion is two times the Gross Domestic Product (GDP) of Norway. Meanwhile, Detica, an information intelligence company part of BAE Systems, and the UK Cabinet Office provided a joint report in 2011 which estimates the cost of cyber crimes in the UK to be £27 billion per annum [46]. A large portion of the cost (£21 billion) goes to the corporate sector, which includes losses due to theft of intellectual properties and industrial espionages [46]. Cyber crimes thus seem extremely lucrative. One should however take the figures with a pinch of salt.

An example of a gross mismatch in loss estimations can be seen with phishing. In 2007, Gartner estimated a loss of $3.2 billion due to phishing in the United States with 3.6 million victims and a $886 average per person loss [58]. With a conservative set of parameters, however, Herley and Florêncio [69] estimate the loss to be much smaller. Leverage their earlier study that 0.4% users do enter their passwords at phishing sites [54], and a phishing victim rate of 0.34% estimated by Moore and Clayton [91], Herley and Florêncio estimate that 0.185% (half of the average victim rate) users will really lose money to phishing activities. Considering the online population to be 165 millions in the US, and a median loss of $200 per person, their estimate for phishing losses in the US is $61 million per annum [69]. Although it remains a non-negligible figure, there is a stark difference compared to the estimate given by Gartner.

Indeed there is a lack of reliable estimates of security losses. Many 'guesstimates' are extrapolated from self-reported surveys. Moreover, there are incentives for security vendors to report over-estimated figures. Researchers have been critical with the estimates of the underground economy. Herley and Florêncio [70] challenge the reliability of the estimates of underground economy obtained by monitoring the Internet Relay Chat (IRC) channels. The duo argue that cheating is a way of life in the IRC channels. Yet, while we should question the astronomical loss estimates, what we currently know about the underground economy could well be just the tip of the iceberg. We should also be aware of the tendency of under-reporting from corporate victims to exercise reputation damage control. Many of them may even not realize that an attack has taken place.

Hence, there is great scientific importance to dissect the underground economy of cyber crimes in the academic settings, however challenging it may be. An applaudable work is by Levchenko et al. [82] who have conducted an end-to-end measurement on how spams are being delivered through botnets, how spam-advertised items are merchandised, and how the payments flow. The researchers find that it will be more effective to seek cooperation from a few banks to disrupt the financing of spammers, instead of improving on the detection, blacklisting and takedown of spamming servers and domain names – areas where computer scientists conventionally focus on. Such a measurement study is thus valuable and needed. Not only can an in-depth measurement study provide good insights into the structure and the state of a given problem, it can also allow the defenders to strategically allocate their resources to the most effective security measures.

A closely related work is by Kanich et al. [74]. The authors present two methods to estimate the rate of orders received by enterprises whose revenue drives spams,

and to characterize the spam-advertised products and customers. They find well over 100,000 orders of spam-advertised products per month [74]. In addition, they find that the online illegal pharmacy market is huge with a projected annual revenue in tens of millions, largely supported by a Western consumer base [74]. However, the figure is much less than guesstimates given by others, and is also much less than the annual expenditures on anti-spam solutions [74]. Besides providing a reality check to the anti-spam industry, it certainly cautions us to rethink our security strategies. Have we invested too much? How well are the resources allocated? Are there alternative resources for information security?

Phishing has also received much academic attention in the recent years. Moore and Clayton [91] investigate the modi operandi of phishers, the effectiveness of take-downs, and the victimization rate based on the lifetime of phishing sites. Another of their work [92] finds how non-cooperation between the defenders contributes to the long lifetime of phishing sites, and calls for information sharing in the anti-phishing industry. The same authors investigate how vulnerable servers are being exploited through the use of search engines for recompromise in another work [95]. Interestingly, they find that phishing websites and thus the susceptible servers, placed onto a public blacklist are recompromised no more frequently than the list of susceptible servers only known within closed communities [95]. This adds to the value of a public blacklist for giving better information to the defenders, although the authors do caution for the need of continued monitoring so that the public blacklist does not adversely favor the attackers [95].

Apart from spam and phishing, there have been also a number of research investigations on the ecosystems of fraudulent online activities, including online bullying and threatening in Japan [41], illegal online pharmacies [81], and typo-squatting domains [96]. These studies are particularly interesting. While fraudulent online activities may not be explicitly harmful, there is no reason to assume that they are separated from the economies of malicious activities. Furthermore, there is often no clear assignment of responsibilities – which authorities should act upon the gray areas of the Web – in practice.

1.2. Rethinking Information Security

Given the relative short history, there remain plentiful research problems and potentials in the fields of usable security and security economics. While both fields are cross disciplinary in nature, they tend to be treated separately in the research community. This thesis looks at both the economic and socio-behavioral aspects to provide new insights and to challenge conventional beliefs.

Figure 1.2 presents a framework to relate three different perspectives – technical, economic, and socio-behavioral – of information security. Traditionally, technical research activities have encompassed areas including cryptography, cryptanalysis, protocol design, trusted hardware, authentication, access control, anomaly detection, and privacy enhancing technologies. Technical security and privacy research

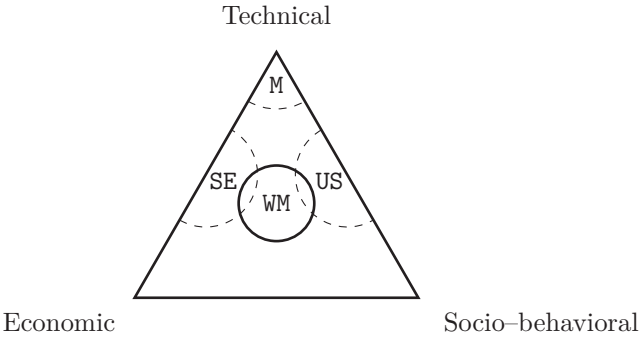


Figure 1.2.: Technical, economic and socio-behavioral perspectives on information security. SE and US denote the focuses of the fields of security economics and usable security, respectively. M indicates the need for high security assurance in military security, while WM denotes the need for a balanced trade-off of different perspectives on the web and app platforms.

ensures confidentiality, integrity, availability, authenticity, non-repudiation, in addition to anonymity and unlinkability.

Meanwhile, research activities on the economics of information security can involve a macro- or microscopic analysis. A macroscopic economic perspective on security problems includes research activities on the optimal security investment, risk management, feasibility of cyber insurance, as well as empirical investigations of underground economies. On the other hand, a microscopic view on the economics of information security typically concerns the analysis of incentives, liabilities and strategies of interdependent actors using tools such as game and contract theories.

Thirdly, security measures can be ineffective without adequate consideration to the users and society. A socio-behavioral perspective looks at the alignment of security measures with social expectations and user behaviors. This encompasses multiple areas investigated by the community of usable security, including the ease-of-use of security features, risk communication, user habituation and cognitive biases, as well as the attitudes, knowledge and awareness of the public.

A point in the triangle indicates the relative weights or focuses of a research activity on different perspectives. The area labelled as SE, for example, represents a wide range of research works in security economics which look at the intersection of technical security and economics. On the other hand, the area US indicates the field of usable security which looks at the socio-behavioral aspects of security measures. Assuming finite resources and excluding those (e.g., nation state actors) who may have access to enormous resources for a comprehensive program, focusing on a particular perspective naturally comes at the expense of the other two. One may thus want to attend to different perspectives of information security strategically depending on the contexts and requirements.

This does not suggest that a security measure must always be cross disciplinary. To illustrate, security measures for military purposes, as indicated by the area `M` in Figure 1.2, may want to focus on the highest level of security assurance albeit they may be more expensive and less user friendly. On the other hand, when designing security measures for the public, one will inevitably need to trade off his focus on security assurance with attentions to economic viability and usability. Yet, the distinction of different security contexts and requirements is often neglected in the research community. Policymakers can play a role to guide and correct the attentions by different research groups accordingly.

Security on the Web, which remains a challenging issue today, will benefit from a balanced treatment of the three perspectives. A security measure to improve web security would not be feasible without considerations to economic deployability and usability, including the ability of users to comprehend the risk signals and react to them expectedly. At the same time, while we should not underestimate the web perpetrators, it will not be helpful to over-assume their abilities. The level of security assurance on the Web differs from the assurance level needed for the military purposes, or for the protection of critical infrastructures. Shouldn't we model the web attack and defense accordingly?

There is also a lack of efficient services and clear cut assignment of responsibilities against various fraudulent online activities. Although not outright malicious, fraudulent websites trick or harm the users through scams, illegal product sales, deceptive information gathering, misuse of user data, and so on. While malicious phishing sites are taken down between 4 to 96 hours, fraudulent websites for mule-recruitment and illegal online pharmacies have an average life-time of two weeks and two months respectively [94]. Other fraudulent activities on the Web include the sales of counterfeit luxury goods or software [82, 74], adult sites (typically plagued with malware and aggresive marketing [121]), typo-squatting domains mimicking the URLs of popular brands [96], as well as online bullying and threatening [41]. Security vendors avoid flagging fraudulent websites fearing the complication of litigations, especially on subjective and potentially contentious matters. On the other hand, online certification issuers and search engines may have conflicts of interests in certifying or accepting advertisement orders from websites in the gray category [48]. The gap of responsibility leads to the question of whether we can leverage inputs from volunteers (expert and ordinary users) in improving web security.

Besides the Web, another domain needing an economic and socio-behavioral perspective is the third party application (app). As mobile device platforms compete for third party applications to be more attractive to the users, more and more device functionalities and personal information are made available to third party developers. The openness and richness in functionalities and information improve user experience, but increase also the incentives for malicious and fraudulent activities. While the motivations of malicious or fraudulent third party apps may be similar to that of bad websites, installing an app involves a different mental process, and can impose a higher level of risks to the users. This constitutes the third motivation of this thesis to examine the risks and challenges following the popularity of third party apps.

1.2.1. Motivation and Related Work

This section details the motivations and related works of the thesis. As briefly sketched out earlier, there are three motivations (M1, M2, M3) in this thesis. Specially, the thesis will investigate the security and privacy risks facing the users on the web and app platforms (M1 and M3), and the potentials of leveraging volunteering efforts, from expert and ordinary users, in mitigating the risks (M2).

M1: Realistic Economic Modeling of Web Attack and Defense

To learn about the attackers we are defending against is crucial for designing an effective defense measure. While we should never underestimate the perpetrators, over-assuming their capabilities, resources and profitability will do a disservice to our community. Researchers find that, for example, the actual losses due to phishing activities can be of a few magnitude orders lower than the figures reported by industrial players [69]. Not only can this lead to an over-spending for security, a 'rosy' picture painted for the profitability of online crimes will only serve to attract more perpetrators, stressing the defense mechanisms even though many of the perpetrators will not be profitable [69]. Yet, it is not trivial to learn more about the attackers through measurement experiments. A few papers measuring the modi operandi and the economics of perpetrators have emerged over the last few years (e.g., [91, 100, 121, 41, 81, 82]); the fact that one of the most complete studies [82] involves 15 co-authors says a lot about the complexity behind the setup of the end-to-end practical measurement.

Without an easy access to good empirical data, it is important to inform our community on the strategies of rational perpetrators and how to mitigate their attempts effectively through analytic modeling. Yet, models capturing the incentives and interdependence of different actors are only useful when constructed to reflect the practical scenarios. How should we realistically model the threats facing users on the Web?

Use of game theoretical analysis in security has gained its popularity in the past few years. An early work is by Liu and Zang (2003) [84] which advocates the use of game theory in reasoning about the attacker behaviors. The authors propose a conceptual framework that formalizes the modeling of attacker intent, objectives and strategies in game theoretical settings. Further, there have been attempts to integrate the modeling of system security and dependability, factoring in both cases whether the underlying failure causes are intentional or not [105]. A comprehensive survey of game theoretical literature for security and privacy problems can be found in [87]. While there are numerous studies that look at the dynamics between an attacker and a defender, they usually model the attack and protection of a set of network systems (e.g., intrusions [11, 12, 20, 25]) or resources (e.g., jamming attacks, denials of service [30, 13, 104]).

The interaction between the defender (e.g., takedown specialists, security vendors) and the perpetrators on the Web is different from the dynamics in network security. First, web perpetrators should be distinguished from state sponsored at-

tackers with potentially unlimited resources so not to focus an overly secure solution at the expense of cost and usability. To be realistic, the actions of the defender and attacker should be constrained by finite resources. Secondly, there is the difference that web perpetrators create new malicious or fraudulent websites on the Web compared to the context of network security where the defender protects a fixed set of systems or resources. The newly created bad websites are unknown to the defender. Furthermore, the defender is limited to use reactive strategies, acting to detect and take down the bad websites created by the perpetrators. To summarize, web security is hence a finite resource allocation problem between the defender and attacker with information asymmetry (unknown bad websites) and action asymmetry (reactive detect-and-takedown defense). How should we model this analytically, and what can we learn from it?

M2: Exploring the Potentials of Community Inputs for Security

Have we invested too little for security, or have we not been able to better coordinate our resources? Can we leverage voluntary efforts in online communities as alternative resources to improve information security?

The notion of 'wisdom of crowds' has gained much popularity ever since the book by Surowiecki in 2005 [111]. Articles on the value of collective judgements can in fact be traced back to more than a century ago. Sir Francis Galton observed in 1907 that the aggregate values (median and mean) of the entries to an ox weight-judging competition were more accurate than individual guesses, indicating the trustworthiness of a democratic judgement [57, 56]. Collective judgements are however not always better. Surowiecki outlines four conditions for a wise crowd to outperform a few experts [111]. He notes that the crowd members should be diverse (not homogenous), have independent thought processes to avoid mere information cascade, be decentralized (to tap into local knowledge and specialization) in addition to the need of a good aggregation strategy to collate the inputs from the individuals.

An example good use of the wisdom of crowds in modern IT systems is the Wikipedia. Denning et al. [45] highlight six potential risks with Wikipedia, namely accuracy, motives of contributors, uncertain expertise, volatility of content, sources of information, and coverage. Despite critiques and skepticisms, Wikipedia has evolved to be one of the most important information sources on the Web. Studies on Wikipedia are plenty. Many of them contribute to analyzing its reliability (e.g., [60]), the contribution patterns (e.g., [77, 97, 119]), as well as its success factors and suggestions for improvement (e.g., [43, 78, 59]). Researchers have also examined the success factors of other collaborative systems, such as the Stack Overflow [4], one of the fastest growing Question and Answer (Q&A) systems [86].

Can we leverage the wisdom of crowds for security purposes? PhishTank [3] is among the first out of the few practical systems that leverage crowd wisdom to improve web security. PhishTank collates user reporting and voting against suspect phishing sites. Another example is the Web of Trust (WOT) [7] which aggregates both human and automated inputs from trusted blacklists to evaluate four aspects of websites, namely trustworthiness, vendor reliability, privacy and child-safety.

Moore and Clayton [93] evaluate the reliability of PhishTank. They find that the participation ratio in PhishTank is highly skewed, following a power-law distribution. They argue that this makes PhishTank particularly susceptible to manipulation. Compared to a commercial phishing blacklist, they find that PhishTank is less comprehensive and slower. In addition, they find that inexperienced users make many errors. However, most of the mistakes are corrected in the voting process. The eventual assessment outcomes contain only few incorrect decisions, all of which are later reversed.

Indeed, two challenges of collective efforts for security purposes are the reliability of user inputs, and the incentives of the contributors (e.g., whether there will be adequate and sustainable volunteering efforts in the long term). Compared to an encyclopedia or a question & answer system, security may impose an even higher bar of contribution barrier given the complexity of security evaluation. Further, there are questions on why and how users, with limited resources, would keep up with the large numbers of malicious and fraudulent websites created daily.

Yet, can we generalize the pessimisms on PhishTank to the use of crowd wisdom for general security evaluation?[1] Is the skewed contribution ratio, commonly found in peer-production systems [119], a real threat? Will the less active users evolve to play a more important role, as observed in Wikipedia [77]? Can ordinary users ever contribute to information security? Are the mistakes by inexperienced users outweighing the potentials of volunteering efforts in complementing the existing measures, and in evaluating aspects that are potentially contentious or subjective, and not covered by security vendors and service providers? How about leveraging inputs from sources which individual users trust?

M3: Risks and Challenges transitioning from the Web to Apps

The mobile industry has been through an exciting revolution over the past few years. An exciting change to the mobile industry is arguably the opening up of the access to various functionalities of the mobile devices and user information to third party developers, as well as the setup of an application store (app store) that channels the third party applications (apps) conveniently to the users.[2] This creates a win-win-win situation – users can now add advanced functionalities to their devices, while third party developers profit from selling apps, and platform owners make a cut from the app sales besides gaining competitive advantages over others as apps increase the attractiveness of a platform. To date, there are more than half a million third party apps available for download on the App Store and Google Play for iOS and Android mobile device platforms, respectively.

The rise of applications is not just a phenomenon on the mobile platforms. One can develop third party apps on web platforms such as on Facebook and Google

[1] Mamykina et al. [86] note that the success of a collaborative system may depend not only on tangible design decisions, but also an active community leadership by the developers. This makes it hard to port the lessons learned from one to other community-based systems.

[2] Symbian and Java platforms has long allowed third party apps to gain access to some device capabilities before the advent to the iOS. Yet, Apple was the first to setup the App Store as a centralized venue for distributing third party apps made for its iOS platform.

Chrome. The HTML5 web standards has built in capabilities for developers to build web apps that can run across different browsers on different device platforms. The availability of comprehensive APIs including offline caching makes it possible for HTML5 web apps to offer functionalities similar to native device applications.

Thus we are witnessing a transition from websites to mobile and web apps. Users are, for example, installing an app to read online news, another app to check flight schedule, and yet another to access Internet banking. The growing popularity of rich and integrated services by third party apps increases the incentives for activities with security and privacy implications. Much research attention has been given to the Android mobile device platform given that its 'laissez-faire' design which allows anyone to develop and distribute an Android application without much scrutiny from Google. Research on the Android platform has focused on platform security architecture [53, 40, 50] and on identifying malicious applications automatically [49, 124, 126]. Others have looked at the problem of a non-global application identification (appID) system and the emergence of alternative application marketplaces [21]. In addition, there have been a number of surveys on malicious applications on Android and mobile device platforms in general [51, 125].

Access to device functionalities and user information by third party apps is governed by the operating system and runtime platform security schemes to apply the principle of least authority – one of the eight design principles for computer security outlined by Saltzer and Schroeder [106]. The most common is the permission-based platform security that has been adopted by modern mobile device application platforms [79] as well as web application platforms on Facebook and Google Chrome. Some of these platforms such as Apple's iOS rely on a central authority to decide what permissions can be granted to a given application while others (Android, Facebook, Google Chrome) rely on the user making the authorization decisions. We refer to the former category 'centralized permission systems' and the latter 'user-consent permission systems'.[3]

What are the different security and privacy challenges with the centralized and user-consent models? Intuitively, centralized permission systems take the burden of judgment away from users. However, there is the question of whether centralized judgment will always be suitable. Apple has received numerous objections for disallowing or removing certain apps from the App Store, prompting some users to 'jail-break' the phone to be able to install the apps from alternative sources [2]. The appropriateness of an application, for example, whether it is privacy-invasive or has offensive content, is a subjective matter, and may be problematic when judged by a central authority.

On the other hand, there are also numerous challenges in the user-consent permission systems. Do users understand the permission systems and pay attention to them? A few studies have looked at the effectiveness of user-consent permission models. King et al. [76] survey the privacy knowledge, behaviors and concerns of Facebook app users. More than a quarter of the survey participants report that

[3]Several HTML5 APIs, such as the geolocation API, currently support a user-consent permission system. The decentralized nature of the Web implies that the user-consent permission systems will become more widespread, if HTML5 web apps become dominant [88].

they have never read the permission request dialog. While half of a quarter of the participants are knowledgeable about Facebook apps, a quarter of them do not even realize that apps are both created by Facebook and third party developers [76]. This highlights the challenge of risk communication, especially when third party apps are tightly integrated onto the platforms, and distributed through official channels provided by well known platform owners (e.g., Facebook, Google, Apple). Meanwhile, Felt et al. [52] analyze the permissions requested by the most popular Android apps and Google Chrome extensions. They conclude that as dangerous permissions are being requested frequently by the popular apps, the user-consent permission model may not be an effective tool for preventing the installation of malware or alerting the users.

Will the above findings generalize to different applications, popular or new, and across different platforms? How are users reacting to apps that request for more permissions than average? Are there reliable risk signals at all that are assisting the users to distinguish the potentially suspicious apps from the good ones? How could we potentially cater for a subjective evaluation? How can we signal risks to users effectively? What are the trends of security and privacy risks facing the users? This thesis contributes by providing answers to the above questions.

1.2.2. Research Methodology

Inline with the multidisciplinary research, the methodology of this thesis is manifold, combining analytic modeling, empirical measurements, and user studies.

Game theoretical analysis is chosen as the tool to model the incentives and dynamics between the web perpetrator and defender (takedown specialist). In particular, we have surveyed for games that incorporate resource constraints and that can be extended to model the practical information and action asymmetries. We find the Colonel Blotto model to be particularly suitable. This class of games has a long history, first introduced by Borel in 1921 [28] and studied by a few others in [29, 64], before being neglected due to its complexity until a reemerged interest in 2006 following the work by Roberson [102]. To have a realistic economic model for web attack and defense, this thesis has taken on the phishing problem as a case study, and surveyed the modi operandi and economics of phishing. In particular, we have constructed our model with reference to practical measurement findings, including those provided by academics (e.g., [94, 92, 90]) and the Anti-Phishing Working Group (APWG) – a consortium of industrial, academic and governmental partners (e.g., [18, 19]).

To evaluate the feasibility of leveraging user inputs for security purposes, the thesis investigates the Web of Trust (WOT) and compares its reliability to non-human based automated systems provided by three popular vendors, namely McAfee, Google and Symantec Norton. We have also managed to obtain multiple data sets from the developers of WOT, based on which we have investigated the contribution patterns in WOT, and evaluated its strengths and potential weaknesses.

On the other hand, the thesis has investigated the limitations of the current user-consent based permission models, and some trends of exploitations on Android,

Facebook and Google Chrome platforms through a large scale data collection and analysis. Through an online survey, we have also studied the self-reported user behaviors during the application installation process, and the attitudes on the security and privacy risks of third party applications. In addition, we have conducted laboratory user studies to evaluate the effectiveness of habituation mitigation mechanisms and integrated risk signals from personalized sources, leveraging a prototype the thesis author developed during his master thesis project [31]. Results from the survey and laboratory experiments are used to construct guidelines for designing a trustworthy application installation process.

Working in relatively new research fields, the thesis has benefited tremendously from international contacts and collaborations. Research visits to well known security groups at Nokia Research Center, University of California (Berkeley) and Carleton University have helped to form interesting research ideas besides laying the foundation for joint papers. The thesis has also managed to tap into the talent of master students at the home institution. In particular, an extended understanding for Android and Facebook applications is gained through supervising a master project focusing on implementing friends based risk signaling on Facebook, and two master theses focusing on data analytics and machine learning methods for identifying suspicious Android applications.

1.3. Thesis Contribution

A total of eleven research papers are co-authored during the doctoral program. As depicted in Figure 1.3, the papers can be broadly categorized into two abstract themes: (i) exploring security and privacy risks facing the users, or Security For Users (`SFU`), and (ii) exploring the potentials of leveraging volunteering efforts from expert and ordinary users for security purposes, or Users For Security (`UFS`). On the other hand, the thesis has focused on two problem domains: issues on the `web`, or on the `app` platforms. Figure 1.3 also indicates how a paper relates to the three thesis motivations `M1`, `M2` and `M3`, and whether a paper takes mainly an economic, or a socio-behavioral perspective of information security, using the circled `e` or `sb`.

Six papers (A–F [33, 37, 34, 32, 36, 39]) on security and privacy problems facing the web and app users, and the potentials of users in contributing to mitigate such problems are included in this thesis, with minor editorial changes. They are depicted in Figure 1.3 as solid nodes. Non-included papers are shown in nodes with dashed line. Paper G [35] is in submission to a conference, while H [22] is a working paper. Meanwhile, paper I [117] analyzes the potential ramifications of an incentive scheme to anonymity management in Tor. Paper J [99] and paper K [38] take a different focus to study the use of trust information in recommender systems. The list of papers and their publication venue, name of publisher, abstract and acceptance rate, if available, are shown in the following. Section 1.3.2 further summarizes the contributions of the included papers.

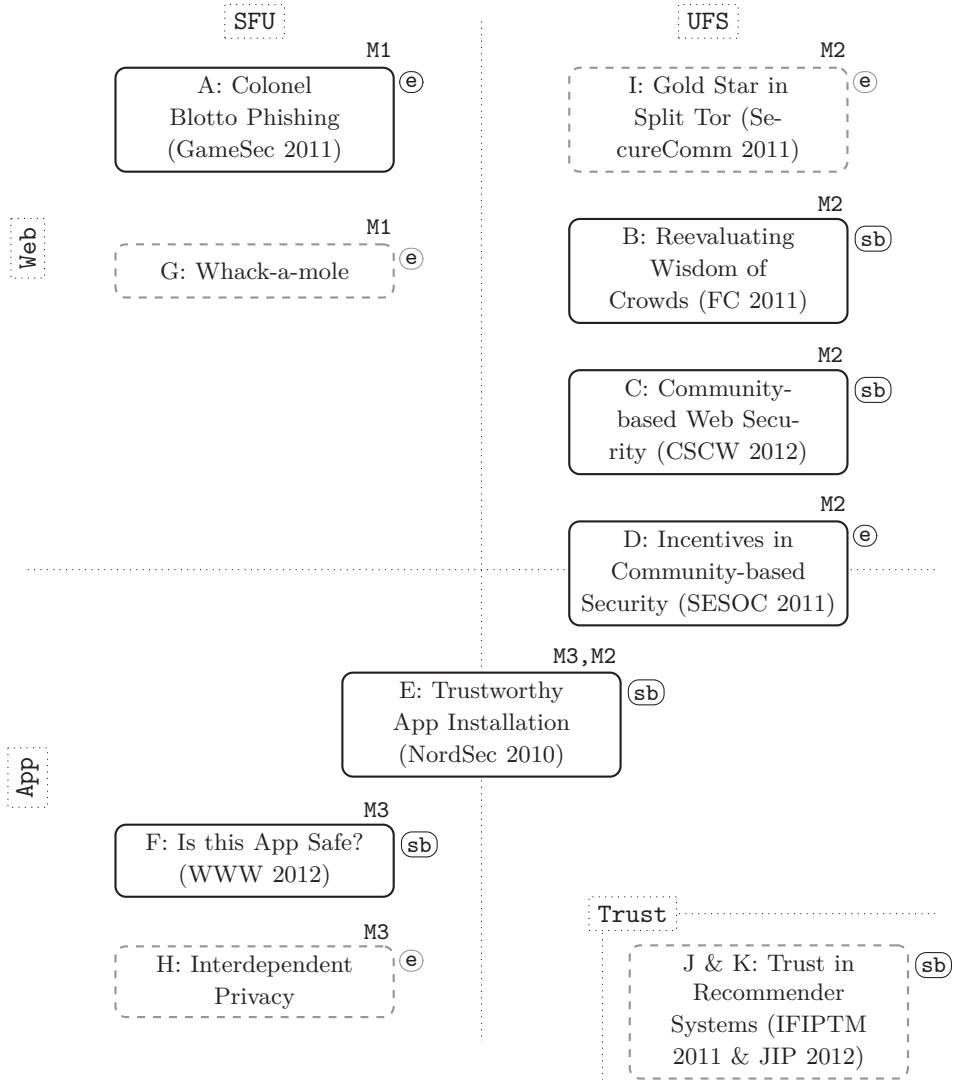


Figure 1.3.: List of papers co-authored on the theme of security for users (SFU), or users for security (UFS), concerning web or app issues. Papers included in the thesis are shown in solid nodes; non-included papers are depicted in dashed nodes. M1, M2 and M3 indicate how a paper relates with three thesis motivations. Circled labels e and sb indicate if a paper provides mainly an economic or socio-behavioral perspective of information security. Paper J and paper K have a different focus; they study the use of trust information in recommender systems.

1.3.1. List of Papers

A. P. H. Chia, J. Chuang. **Colonel Blotto in the Phishing War**, In *Proceedings of 2011 Conference on Decision and Game Theory for Security (GameSec)*, LNCS vol. 7037, pp. 201–218, Springer (2011).

Abstract. Phishing exhibits characteristics of asymmetric conflict and guerrilla warfare. Phishing sites, upon detection, are subject to removal by takedown specialists. In response, phishers create large numbers of new phishing attacks to evade detection and stretch the resources of the defenders. We propose the Colonel Blotto Phishing (CBP) game, a two-stage Colonel Blotto game with endogenous dimensionality and detection probability. We find that the optimal number of new phishes to create, from the attacker's perspective, is influenced by the degree of resource asymmetry, the cost of new phishes, and the probability of detection. Counter-intuitively, we find that it is the less resourceful attacker who would create more phishing attacks in equilibrium. And depending on the detection probability, an attacker will vary his strategies to either create even more phishes, or to focus on raising his resources to increase the chance he will extend the lifetime of his phishes. We discuss the implications to anti-phishing strategies and point out that the game is also applicable to web security problems more generally.

Contribution statement. Pern Hui Chia was the main contributor of this work. Analysis and programming in Mathematica were due to Pern Hui Chia. John Chuang contributed to active discussion and parts of the writing, as well as the original ideation to adapt Colonel Blotto games for phishing dynamics.

B. P. H. Chia, S. J. Knapskog. **Re-Evaluating the Wisdom of Crowds in Assessing Web Security**, In *Proceedings of 15th International Conference on Financial Cryptography & Data Security (FC)*, LNCS vol. 7035, pp. 299–314, Springer (2011). *(Acceptance rate: 16/56=29%)*

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.

Contribution statement. Pern Hui Chia was the main contributor of this work. Data collection and analysis were due to Pern Hui Chia. Svein J. Knapskog provided valuable feedbacks and contributed to parts of the writing. Two of the data sets used in this work were provided by the developers of WOT.

C. P. H. Chia, J. Chuang. **Community-based Web Security: Complementary Roles of the Serious and Casual Contributors**, In *Proceedings of 15th Conference on Computer Supported Cooperative Work (CSCW)*, pp. 1023–1032, ACM (2012). *(Acceptance rate: 40%, note the new two-phase conference-journal hybrid reviewing system of CSCW starting from 2012:* `http://cscw.acm.org/CSCW-review-process-statement.pdf`*)*

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.

Contribution statement. Pern Hui Chia was the main contributor of this work. Data collection and analysis were due to Pern Hui Chia. John Chuang contributed to active discussion and parts of the writing, particularly the abstract and introduction. Two of the data sets used in this work were provided by the developers of WOT.

D. P. H. Chia. **Analyzing the Incentives in Community-based Security Systems**, In *9th International Conference on Pervasive Computing and Communications (PerCom)*, Workshop Proceedings – *4th International Workshop on Security and Social Networking (SESOC)*, pp. 270–275, IEEE (2011). *(Acceptance rate: 7/23=30%)*

Abstract. Apart from mechanisms to make crowd-sourcing secure, the reliability of a collaborative system is dependent on the economic incentives of its potential contributors. We study several factors related to the incentives in a community-based security system, including the expectation on the social influence and the contagion effect of generosity. We also investigate the ef-

fects of organizing community members differently in a complete, random and scale-free structure. Our simulation results show that, without considering any specific incentive schemes, it is not easy to encourage user contribution in a complete-graph community structure (global systems). On the other hand, a moderate level of cooperative behavior can be cultivated when the community members are organized in the random or scale-free structure (social networks).

Contribution statement. Pern Hui Chia was the sole author of this work.

E. P. H. Chia, A. P. Heiner, N. Asokan. **Use of Ratings from Personalized Communities for Trustworthy Application Installation**, In *Proceedings of 15th Nordic Conference in Secure IT Systems (NordSec)*, LNCS vol. 7127, pp. 71–88, Springer (2010).

Abstract. The problem of identifying inappropriate software is a daunting one for ordinary users. The two currently prevalent methods are intrinsically centralized: certification of "good" software by platform vendors and flagging of "bad" software by antivirus vendors or other global entities. However, because appropriateness has cultural and social dimensions, centralized means of signaling appropriateness is ineffective and can lead to habituation (user clicking-through warnings) or disputes (users discovering that certified software is inappropriate). In this work, we look at the possibility of relying on inputs from personalized communities (consisting of friends and experts whom individual users trust) to avoid installing inappropriate software. Drawing from theories, we developed a set of design guidelines for a trustworthy application installation process. We had an initial validation of the guidelines through an online survey; we verified the high relevance of information from a personalized community and found strong user motivation to protect friends and family members when know of digital risks. We designed and implemented a prototype system on the Nokia N810 tablet. In addition to showing risk signals from personalized community prominently, our prototype installer deters unsafe actions by slowing the user down with habituation-breaking mechanisms. We conducted also a hands-on evaluation and verified the strength of opinion communicated through friends over opinion by online community members.

Contribution statement. Pern Hui Chia was the main contributor of this work. Andreas P. Heiner contributed to active discussion and parts of the writing, particularly the section on user cognition during application installation. N. Asokan also contributed to active discussion and parts of the writing, particularly the abstract and introduction. The online survey and hands-on evaluation were jointly designed by all three authors. The hands-on evaluation was conducted by Pern Hui Chia and Andreas P. Heiner. The two also jointly analyzed the findings of the survey and hands-on evaluation. The web implementation of the survey was due to Pern Hui Chia. Meanwhile, the prototype system was adapted from the master thesis project of Pern Hui Chia, which was conducted at Nokia Research Center and supervised by Andreas P. Heiner.

F. P. H. Chia, Y. Yamamoto, N. Asokan. **Is this App Safe? A Large Scale Study on Application Permissions and Risk Signals**, In *Proceedings of 21st International World Wide Web Conference (WWW)*, pp. 311–320, ACM (2012). *(Acceptance rate: 108/885=12%)*

Abstract. Third-party applications (apps) drive the attractiveness of web and mobile application platforms. Many of these platforms adopt a decentralized control strategy, relying on explicit user consent for granting permissions that the apps request. Users have to rely primarily on community ratings as the signals to identify the potentially harmful and inappropriate apps even though community ratings typically reflect opinions about perceived functionality or performance rather than about risks. With the arrival of HTML5 web apps, such user-consent permission systems will become more widespread. We study the effectiveness of user-consent permission systems through a large scale data collection of Facebook apps, Chrome extensions and Android apps. Our analysis confirms that the current forms of community ratings used in app markets today are not reliable indicators of privacy risks of an app. We find some evidence indicating attempts to mislead or entice users into granting permissions: free applications and applications with mature content request more permissions than is typical; "look-alike" applications which have names similar to popular applications also request more permissions than is typical. We also find that across all three platforms popular applications request more permissions than average.

Contribution statement. Pern Hui Chia was the main contributor of this work. Yusuke Yamamoto contributed to active discussion and parts of the writing, particularly the section on look-alike third party applications. N. Asokan also contributed to active discussion and parts of the writing, particularly the abstract, introduction and results sections. Data collection and a large part of the analysis were due to Pern Hui Chia. The analysis of look-alike application names was done by Yusuke Yamamoto.

Other Papers (not included in thesis)

G. P. H. Chia, J. Chuang, Y. Chen. **Whack-a-mole: Asymmetric conflict and guerrilla warfare in web security**. (*in submission*)

H. G. Biczók, P. H. Chia. **Interdependent privacy: Your actions affect my privacy**. (*working paper*)

I. B. Westermann, P. H. Chia, D. Kesdogan. **Analyzing the gold star scheme in a split Tor network**, In *Proceedings of 8th International Conference on Security and Privacy in Communication Networks (SecureComm)*, Springer (2011). *(Acceptance rate: 23/95=24%)*

J. G. Pitsilis, P. H. Chia. **Does trust matter for user preferences? A study on Epinions ratings**, In *Proceedings of 4th IFIP International Conference on Trust Management (IFIPTM)*, pp. 232–247, Springer (2010). *(Acceptance rate: 18/61=30%)*

K. P. H. Chia, G. Pitsilis. **Exploring the use of explicit trust links for filtering recommenders: a study on Epinions.com**, *Journal of Information Processing (JIP)*, 19:332–344, Information Processing Society of Japan (IPSJ), (2011).

1.3.2. Summary of Contribution

The contribution of this thesis according to the three identified motivations are summarized in the following:

M1. Realistic Economic Modeling of Web Attack and Defense (Paper A)

- We observe that realistic web attack and defense exhibits characteristics of asymmetric conflict and guerrilla warfare, different from protecting a known set of assets in the context of network security. The defenders are constrained by information and action asymmetries, not knowing which new malicious sites have been created, and reactively taking down the detected ones. While the defenders may be more resourceful given the help from service providers, policymakers and law enforcers, web perpetrators create large numbers of malicious sites to stretch the resources of the defenders.

- We propose a two-step resource allocation game with endogenous dimensionality and detection probability to model the asymmetric conflict on the Web. Our model is adapted from the Colonel Blotto two-player game, first introduced by Borel in 1921 [28], studied by Borel and Ville in 1938 [29], and Gross and Wagner in 1950 [64]. Colonel Blotto games have been largely neglected arguably due to a lack of pure-strategy equilibriums and the complexity of the solution for the case of asymmetric resources, until the work by Roberson (2006) [102], which successfully characterizes the unique equilibrium payoffs under all configurations of resource asymmetry.

- When applied to the scenario of phishing, our Colonel Blotto Phishing (CBP) model gives several interesting insights. Somewhat counter-intuitively, we find that a less resourceful attacker will create more phishing attacks than a resourceful counterpart in equilibrium. Further, we find that it is optimal for an attacker to vary his strategies, to reduce cost so as to create large numbers of new phishes given a low detection probability, or to focus on raising his resources when the detection probability is high.

- Our findings provide some implications to the anti-phishing industry. Increasing the degree of resource asymmetry by either raising the defender's resources or disrupting the attacker's infrastructures, increasing the cost of

new phishes, and increasing the probability of detection, can all reduce the attacker's utility. However, raising the cost of new phishes may be hard in practice given tricks such as the use of stolen credit card numbers for registering new domains, and the use of compromised servers as phishing hosts. An increased detection probability can be achieved through data sharing among the defenders, or volunteering efforts such as user reporting. Lastly, we note the importance of good estimates of the state of the problem to help the defender to prioritize different measures accordingly. For example, it will be optimal to focus on disrupting the attacker's infrastructures when the detection probability is high, and to focus on making it harder to create large numbers of new phishes by patching vulnerable servers or cooperating with domain registries, otherwise.

M2. Exploring the Potentials of Community Inputs for Security (Paper B,C,D,E)

- One of the few community-based systems for web security is the Web of Trust (WOT). WOT comes in the form of a browser add-on which has been downloaded more than 30 million times, and a central community portal which has more than 3 million registered contributors by early 2012. We study the performance and characteristics of WOT inline with our motivation to explore the potentials of volunteering efforts in contributing to web security.

- We evaluate the reliability of WOT comparing it to three well known services based on automated assessments, namely, McAfee's SiteAdvisor, Norton's Safe Web and Google's Safe Browsing Diagnostic Page. We find that WOT's general coverage is low, but it is actually more comprehensive than the three automated services in identifying bad domains. While this could be due to the larger evaluation scope of WOT and the tendency of service providers to be more conservative in order to avoid potential litigations, it is to the credit of the WOT community in providing comprehensive warnings against potentially suspicious websites.

- We 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. There is also a lack of data sharing between service providers; only a few sites are commonly classified as bad by the automated services.

- We find that the contribution ratios in WOT are skewed, similar to those in PhishTank as found by Moore and Clayton [93]. We further study the characteristics of different contributors, and observe the complementary roles by serious and casual members. Serious contributors are responsible for the bulk of inputs in WOT, and they report the majority of bad websites, often based on some blacklists. Negative aspects of sites evaluated in WOT are mostly objective; this mitigates the risk of a skewed contribution ratio. On

the other hand, a large percentage of attention to the goodness of sites comes from the casual contributors. While they do not evaluate bad sites extensively and reliably, assessments for the good sites are equally valuable in the context of web security. Whitelisting can play an equally important role 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.

- We present an infinitely repeated total effort security game, adapted from the static versions presented in [65, 66], to model the incentives in a community-based security system. In the reverse manner of the tragedy of the commons [67], contributing to common protection is a problem of public goods provisioning. Our model reflects that, without any incentive schemes, the level of user contribution can be expected to be low. Cultivating a sense of social responsibility can increase contribution level; however, this is challenging in practice. Further, we find that in the presence of a few generous users, who contribute unconditionally, the overall contribution level increases but with only a limited contagion effect. Over-reliance on a small group of generous users, on the other hand, is responsible for a skewed contribution ratio, reflecting the case in many practical systems including WOT. We find, however, that it is possible to encourage a moderate level of contribution when organizing the users in a random or scale-free structure. This points to a potential research direction in exploiting community structures in social networks for increasing user contribution. Indeed, we find in an online survey that users are motivated to protect their friends and family members when they know of digital risks. We look into the use of inputs from personalized communities for trustworthy application installation in Paper E (see M3).

- We find that WOT is not without several potential weaknesses. A potential pitfall we identify is the current lack of distinction between subjective and objective evaluation factors. Evaluating if a site has malicious content, browser exploits, or is a phish is an objective process and can be verified. However, deciding if a site is good or has ethical issues is subjective and can be contentious. The risks of manipulation given a skewed contribution ratio and a black-box rating aggregation process, can be high when the underlying evaluation factors are subjective. We suggest that WOT should restrict the use of the mass-rating-tool, which enables the serious contributors to evaluate multiple sites conveniently, to only objective evaluation factors. We also suggest to improve the verifiability of user inputs through a better referencing system when it comes to objective evaluations, and a structured evaluation process when it comes to subjective factors. In addition, we note the importance of evaluating the reliability of individual contributors per context basis; WOT should not mix the reliability of contributors across subjective and objective evaluation factors.

- We find multiple strengths of WOT. First, the browser add-on is easy to use and saliently signals against potentially suspicious sites. It caters for multiple user concerns on the Web not evaluated by automated services. It integrates inputs from trusted blacklists and has multiple anti-gaming measures built in, including abilities to automatically monitor for suspicious rating behaviors, and a rating aggregation algorithm which factors in the reliability of individual contributors. The rating process is easy. We also observe the developers to be actively engaged with the community of volunteers, in addition to sharing the aggregate assessments back to the public. WOT exemplifies the feasibility of leveraging the wisdom of crowds to improve web security.

M3. Risks and Challenges transitioning from the Web to Apps (Paper E, F)

- Installing an application (app) on the mobile device or in the browser imposes a higher level of security and privacy risks than browsing on the Web. It also involves a different mental process. We define the terminology of 'inappropriate software' to refer to the set of applications that may cause a bad user experience. It includes applications that fundamentally disregard user choice, have malicious intents, or adopt bad practices such as installing additional unexpected software and using incomprehensible End User License Agreement (EULA) that hinder an informed consent. More than the notion of badware by StopBadware.org [5], we regard inappropriate software to include those that may be culturally or socially offensive.

- We identify two different risks with a centralized approach to trustworthy application installation – risk of centralized judgment, and risk of habituation. On platforms such as the iOS, a central authority like Apple decides if an application can be distributed in the official app store. Such centralized judgment has not been without any controversies. Apple's decisions have been objected in numerous occasions [2]. Software certification (e.g., Symbian Signed) is another prominent centralized approach to govern if an application can access certain device capabilities. Application installers will normally display warning and disclaimer notices when an application to be installed is not certified. However, such notices are often context insensitive and uninformative. Besides degrading user experience, they rarely indicate a true risk, causing many users to habitually click-through them.

- Drawing from cognitive and information flow theories, we develop a set of design guidelines for a trustworthy application installation process. We identify three guidelines, namely: (i) avoid requiring user actions that can be easily habituated, (ii) employ signals that are visually salient, relevant and of high impact, and (iii) incorporate mechanisms to gather and utilize feedbacks from personalized communities. To mitigate habituation, we note that frequent user actions in the normal context could be made implicit, and replaced with an attention capture mechanism that will signal any deviations from this context. By personalized communities, we refer to friends and expert users whom individual users trust. Experts could be vendors or gurus

who are knowledgeable in the technical evaluation of software, while friends refer to ones whom users have personal contacts with and who could help by sharing their personal experience with apps or relaying information.

- We validate the design guidelines through an online survey. Specifically, most of the survey participants admit that they seldom read the EULA, privacy policy, and disclaimer notices during application installation. The self-report survey also find the high relevance of inputs from personalized communities, and a strong motivation to protect friends and family members when users know of digital risks.

- We conduct a laboratory experiment using a prototype application installer adapted from the doctoral candidate's master thesis project [31]. In addition to showing risk signals from personalized communities prominently, the prototype installer deters unsafe actions by slowing the user down through a habituation-breaking mechanism. Through the user study, we verify that opinions given by friends are of a higher impact than those given by unknown online community members. Warnings by friends overrule the positive feedbacks by online community members, but not vice-versa. The habituation-breaking mechanism receives a mixed response, and can be improved. At the same time, we find that most of the participants are positive with the idea of an integrated app appropriateness rating from personalized communities.

- Opposite to a centralized approach to application scrutiny, more and more platforms including the Android, Facebook, Chrome and HTML5 web apps, are adopting a decentralized 'laissez-faire' control strategy. Application installers on these platforms request for explicit user consent for granting the permissions that the apps request. Users are to left to rely primarily on community ratings to avoid the potentially harmful and inappropriate apps. This is despite that the current rating systems typically reflect opinions about perceived functionality or performance of apps, instead of risks or appropriateness. We analyze the current state of risk signaling on app intrusiveness, and if there is any evidence of attempts to mislead or entice users into compromising their privacy. This is done through a large scale data analytics, covering three user-consent permission based platforms, namely Facebook, Chrome, and Android. The data sets are shared on our project website [1].

- In spite of the different UI designs and permission granularities, we find that the popularity of an app correlates with the number of permissions it requests, even when considering information-sensitive permissions only. This has two implications. First, displaying the permissions an app requests to signal the potential security and privacy risks is likely to be ineffective. As popular apps request for more permissions, users will be trained to habitually accept permission requests. Second, there appear no disincentives for developers who intentionally or mistakenly over-privilege their apps currently.

- Our analysis confirms that the current community rating systems used in app marketplaces are not reliable indicators of app risks. The same goes for signals such as the availability of a developer website, and the number of apps published by the developer. If they are to be reliable to help users detect privacy intrusive apps, they should exhibit a negative correlation with the number of dangerous permissions requested by an app. However such negative correlations are not observed. On the other hand, we find some external services that show potentials in signaling app risks. One is the website reputation scores from WOT, and another is the flagging of spam apps by AppBrain.com. We suggest to prominently display such external signals in app marketplaces to help users recognize those that are potentially intrusive.

- We find some evidence indicating attempts to mislead or entice users into granting permissions; free apps and applications with mature content request more permissions than is typical. The trends hold even when considering only the information-sensitive permissions. Excluding permissions that are commonly required by third party advertisement libraries, we find that free apps still request for more permissions than the paid apps.

- We find the 'look-alike' applications, which have names similar to popular applications, also request more permissions than is typical. While the fraction of look-alike apps is small, there is an underlying problem of 'cheap identity' with app names and IDs currently.

1.4. Conclusions

This thesis contributes to the fields of security economics and usable security in three ways. First, a realistic model of web attack and defense, adapted from the classical Colonel Blotto game, is presented. The model factors in the practical asymmetric conflict and endogenous dimensionality, to reflect the realistic scenario of a resource constrained attacker creating large numbers of malicious sites to evade detection and to stretch the resources of the defender.

Second, this thesis highlights the feasibility of community-based efforts for security purposes. While several potential challenges are identified, leveraging community inputs for information security has the benefits of being more comprehensive and relevant than automated assessments. This does not imply that the responsibility of security should be removed from service providers or security vendors, and assigned to the users. It demonstrates the potentials of volunteering efforts in complementing the conventional security measures.

Third, the thesis studies the risks and challenges of centralized and decentralized controls of third party applications. Centralized app control causes the risk of central judgment and the risk of habituation, leading to user disputes and careless behaviors. The increasingly widespread decentralized user-consent model also suffers from the lack of effective risk signaling. Permissions shown on the user-consent dialogs will be habitually ignored given the tendency of popular apps requesting

more permissions than average, and the absence of alternative risk signals. Free apps, mature apps, and apps with names mimicking the popular ones, also request more permissions than typical, indicating possible trends of exploitations.

1.4.1. Directions for Future Research

There are multiple potential directions for future research. The economic model presented in this thesis has not factored in the practical competition and collaboration between multiple attackers. In addition, it does not consider heterogeneous 'battlefields', and thus does not model targeted attacks on the Web, such as spear-phishing. To improve on realistic economic modeling of web attack and defense, further measurement studies will continue to be helpful to inform about attacker modi operandi and the state of the underground economies.

Apart from the dark side of information security, it will also be interesting to measure the resources of the defenders. Should we invest more for security, or have we not been able to coordinate our resources well enough? Security volunteers represent the alternative resources of the defenders. This thesis is among the few to evaluate the reliability of volunteering efforts for security purposes. Future studies can take a deeper look into the sustainability of security volunteering, as well as the risk of gaming given the increasing threat of malicious crowdsourcing [116].

Against the backdrop of ever increasing malicious and fraudulent activities on the web and app platforms, there is a lack of good metrics to quantify the security and privacy risks facing the users currently. App permissions are problematic as they are being used to apply the principle of least privilege on developers, and to convey risks to users simultaneously. This creates conflicting signals. In addition to risk metrics, further research to improve the effectiveness of risk signaling, taking into account of user behaviors and cognitions, will remain important and interesting.

References

[1] Is this App Safe – Our Project Website. http://aurora.q2s.ntnu.no/app.

[2] Objections towards iTunes Appstore approval process. http://news.cnet.com/8301-13506_3-10317057-17.html, http://www.thelocal.de/society/20091125-23501.html, http://www.eff.org/deeplinks/2009/06/oh-come-apple-reject, http://www.eff.org/deeplinks/2009/05/apple-says-public-do, http://www.eff.org/deeplinks/2009/02/south-park-iphone-app-denied. Last accessed: June 2012.

[3] PhishTank. http://www.phishtank.com.

[4] Stack Overflow. http://stackoverflow.com.

[5] StopBadware. http://www.stopbadware.org.

[6] The Common Criteria Portal. http://www.commoncriteriaportal.org.

[7] Web of Trust. http://www.mywot.com.

[8] A. Adams and M. A. Sasse. Users are not the enemy. *Commun. ACM*, 42(12):40–46, 1999.

[9] G. A. Akerlof. The Market for 'Lemons': Quality Uncertainty and the Market Mechanism. *The Quarterly Journal of Economics*, 84(3):488–500, 1970.

[10] G. L. Albano and A. Lizzeri. Strategic certification and provision of quality. *International Economic Review*, 42(1):267–283, 2001.

[11] T. Alpcan and T. Başar. A game theoretic approach to decision and analysis in network intrusion detection. In *Proceedings of the 42nd IEEE Conference on Decision and Control*, pages 2595–2600, December 2003.

[12] T. Alpcan and T. Başar. A game theoretic analysis of intrusion detection in access control systems. In *Proceedings of the 43rd IEEE Conference on Decision and Control*, pages 1568–1573, December 2004.

[13] E. Altman, K. Avrachenkov, and A. Gamaev. Jamming in wireless networks: the case of several jammers. In *Proceedings of the First ICST international conference on Game Theory for Networks*, GameNets'09, pages 585–592. IEEE Press, 2009.

[14] E. Amoroso. Written testimony to US Senate Committee on Commerce, Science and Transportation: Hearing on Improving Cybersecurity (March 19, 2009). http://commerce.senate.gov/public/?a=Files.Serve&File_id=e8d018c6-bf5f-4ea6-9ecc-a990c4b954c4. Last accessed: June 2012.

[15] R. Anderson. Why information security is hard-an economic perspective. In *ACSAC*, pages 358–365. IEEE Computer Society, 2001.

[16] R. Anderson and T. Moore. The economics of information security. *Science*, 314(5799):610–613, 2006.

[17] R. Anderson and T. Moore. Internet security. In M. Peitz and J. Waldfogel, editors, *The Oxford Handbook of the Digital Economy*, chapter 21. Oxford University Press, 2012. (to appear).

[18] Anti-Phishing Working Group (APWG). Global phishing survey: Trends and domain name use in 2H2009. `http://www.antiphishing.org/reports/APWG_GlobalPhishingSurvey_2H2009.pdf`. Last accessed: June 2012.

[19] Anti-Phishing Working Group (APWG). Global phishing survey: Trends and domain name use in 2H2010. `http://www.antiphishing.org/reports/APWG_GlobalPhishingSurvey_2H2010.pdf`. Last accessed: June 2012.

[20] A. Årnes, K. Sallhammar, K. Haslum, T. Brekne, M. Moe, and S. Knapskog. Real-time risk assessment with network sensors and intrusion detection systems. In Y. Hao, J. Liu, Y.-P. Wang, Y.-m. Cheung, H. Yin, L. Jiao, J. Ma, and Y.-C. Jiao, editors, *Computational Intelligence and Security*, volume 3802 of *Lecture Notes in Computer Science*, pages 388–397. Springer Berlin / Heidelberg, 2005.

[21] D. Barrera, W. Enck, and P. C. van Oorschot. Meteor: Seeding a Security-Enhancing Infrastructure for Multi-market Application Ecosystems. In *IEEE Mobile Security Technologies Workshop (MoST)*, 2012.

[22] G. Biczók and P. H. Chia. Interdependent privacy: Your actions affect my privacy. (working paper).

[23] R. Böhme. Cyber-insurance revisited. In *Proceedings of the 4th Workshop on the Economics of Information Security*, WEIS '05, 2005.

[24] R. Böhme and S. Köpsell. Trained to accept? A field experiment on consent dialogs. In E. D. Mynatt, D. Schoner, G. Fitzpatrick, S. E. Hudson, W. K. Edwards, and T. Rodden, editors, *CHI*, pages 2403–2406. ACM, 2010.

[25] R. Böhme and T. Moore. The iterated weakest link. *IEEE Security & Privacy*, 8(1):53–55, 2010.

[26] R. Böhme and G. Schwartz. Modeling cyber-insurance: Towards a unifying framework. In *Proceedings of the 9th Workshop on the Economics of Information Security*, WEIS '10, 2010.

[27] J. Bonneau, C. Herley, P. van Oorschot, and F. Stajano. The quest to replace passwords: A framework for comparative evaluation of web authentication schemes. In *IEEE Symposium on Security and Privacy*, May 2012.

[28] E. Borel. La théorie du jeu les équations intégrales à noyau symétrique. *Comptes Rendus de l'Académie des Sciences*, 173:1304–1308, 1921. English translation by Savage, L. (1953) The theory of play and integral equations with skew symmetric kernels, Econometrica 21:97–100.

[29] E. Borel and J. Ville. Application de la théorie des probabilitiés aux jeux de hasard. 1938. Reprinted in E. Borel, A. Chéron, Théorie mathematique du bridge à la portée de tous, Editions Jacques Gabay, Paris, 1991.

[30] R. Chen, J.-M. Park, and J. H. Reed. Defense against primary user emulation attacks in cognitive radio networks. *IEEE Journal on Selected Areas in Communications*, 26(1):25–37, 2008.

[31] P. H. Chia. Secure software installation via social rating. Masters thesis, Helsinki University of Technology, and Royal Institute of Technology, 2008.

[32] P. H. Chia. Analyzing the incentives in community-based security systems. In *PerCom Workshops*, pages 270–275. IEEE, 2011.

[33] P. H. Chia and J. Chuang. Colonel blotto in the phishing war. In J. S. Baras, J. Katz, and E. Altman, editors, *GameSec*, volume 7037 of *Lecture Notes in Computer Science*, pages 201–218. Springer, 2011.

[34] P. H. Chia and J. Chuang. Community-based web security: complementary roles of the serious and casual contributors. In S. E. Poltrock, C. Simone, J. Grudin, G. Mark, and J. Riedl, editors, *CSCW*, pages 1023–1032. ACM, 2012.

[35] P. H. Chia, J. Chuang, and Y. Chen. Whack-a-mole: Asymmetric conflict and guerrilla warfare in web security. (in submission).

[36] P. H. Chia, A. P. Heiner, and N. Asokan. Use of ratings from personalized communities for trustworthy application installation. In T. Aura, K. Järvinen, and K. Nyberg, editors, *NordSec*, volume 7127 of *Lecture Notes in Computer Science*, pages 71–88. Springer, 2010.

[37] P. H. Chia and S. J. Knapskog. Re-evaluating the wisdom of crowds in assessing web security. In G. Danezis, editor, *Financial Cryptography*, volume 7035 of *Lecture Notes in Computer Science*, pages 299–314. Springer, 2011.

[38] P. H. Chia and G. Pitsilis. Exploring the use of explicit trust links for filtering recommenders: A study on epinions.com. *JIP*, 19:332–344, 2011.

[39] P. H. Chia, Y. Yamamoto, and N. Asokan. Is this App Safe? A large scale study on Application Permissions and Risk Signals. In A. Mille, F. L. Gandon, J. Misselis, M. Rabinovich, and S. Staab, editors, *WWW*, pages 311–320. ACM, 2012.

[40] E. Chin, A. P. Felt, K. Greenwood, and D. Wagner. Analyzing inter-application communication in android. In A. K. Agrawala, M. D. Corner, and D. Wetherall, editors, *MobiSys*, pages 239–252. ACM, 2011.

[41] N. Christin, S. S. Yanagihara, and K. Kamataki. Dissecting one click frauds. In E. Al-Shaer, A. D. Keromytis, and V. Shmatikov, editors, *ACM Conference on Computer and Communications Security*, pages 15–26. ACM, 2010.

[42] J. Chuang. Incentive Dynamics in Interdependent Network Security. Or: Buying a Raft and Out-Running a Bear, April 2011. Plenary Lecture at *GameNets* 2011. Available at: `http://people.ischool.berkeley.edu/~chuang/pubs/gamenets2011.pdf`. Last accessed: June 2012.

[43] D. Cosley, D. Frankowski, L. G. Terveen, and J. Riedl. Suggestbot: using intelligent task routing to help people find work in wikipedia. In D. N. Chin, M. X. Zhou, T. A. Lau, and A. R. Puerta, editors, *IUI*. ACM, 2007.

[44] L. F. Cranor. A framework for reasoning about the human in the loop. In E. F. Churchill and R. Dhamija, editors, *UPSEC*. USENIX Association, 2008.

[45] P. J. Denning, J. Horning, D. L. Parnas, and L. Weinstein. Wikipedia risks. *Commun. ACM*, 48(12):152, 2005.

[46] Detica and UK Cabinet Office. The cost of cyber crime, Feb 2011. `http://www.cabinetoffice.gov.uk/resource-library/cost-of-cyber-crime`. Last accessed: June 2012.

[47] R. Dhamija, J. D. Tygar, and M. A. Hearst. Why phishing works. In R. E. Grinter, T. Rodden, P. M. Aoki, E. Cutrell, R. Jeffries, and G. M. Olson, editors, *CHI*, pages 581–590. ACM, 2006.

[48] B. Edelman. Adverse selection in online "trust" certifications and search results. *Electronic Commerce Research and Applications*, 10(1):17–25, 2011.

[49] W. Enck, M. Ongtang, and P. D. McDaniel. On lightweight mobile phone application certification. In E. Al-Shaer, S. Jha, and A. D. Keromytis, editors, *ACM Conference on Computer and Communications Security*, pages 235–245. ACM, 2009.

[50] A. P. Felt, E. Chin, S. Hanna, D. Song, and D. Wagner. Android permissions demystified. In Y. Chen, G. Danezis, and V. Shmatikov, editors, *ACM Conference on Computer and Communications Security*. ACM, 2011.

[51] A. P. Felt, M. Finifter, E. Chin, S. Hanna, and D. Wagner. A survey of mobile malware in the wild. In *Proceedings of the 1st ACM workshop on Security and privacy in smartphones and mobile devices*, SPSM '11. ACM, 2011.

[52] A. P. Felt, K. Greenwood, and D. Wagner. The effectiveness of application permissions. In *Proceedings of the 2nd USENIX conference on Web application development*, WebApps '11. USENIX Association, 2011.

[53] A. P. Felt, H. J. Wang, A. Moshchuk, S. Hanna, and E. Chin. Permission redelegation: Attacks and defenses. In *USENIX Security Symposium*. USENIX Association, 2011.

[54] D. A. F. Florêncio and C. Herley. A large-scale study of web password habits. In C. L. Williamson, M. E. Zurko, P. F. Patel-Schneider, and P. J. Shenoy, editors, *WWW*, pages 657–666. ACM, 2007.

[55] D. A. F. Florêncio and C. Herley. Where do all the attacks go? In *Proceedings of the 10th Workshop on the Economics of Information Security*, WEIS '11, 2011.

[56] F. Galton. The Ballot-Box. *Nature*, 75(1952):509–510, 1907.

[57] F. Galton. Vox populi. *Nature*, 75(1949):7, 1907.

[58] Gartner, Inc. Gartner Survey Shows Phishing Attacks Escalated in 2007; More than $3 Billion Lost to These Attacks, December 2007. `http://www.gartner.com/it/page.jsp?id=565125`. Last accessed: June 2012.

[59] R. S. Geiger and D. Ribes. The work of sustaining order in wikipedia: the banning of a vandal. In K. I. Quinn, C. Gutwin, and J. C. Tang, editors, *CSCW*, pages 117–126. ACM, 2010.

[60] J. Giles. Internet encyclopaedias go head to head. *Nature*, 438:900–901, December 2005.

[61] N. Good, J. Grossklags, D. K. Mulligan, and J. A. Konstan. Noticing notice: a large-scale experiment on the timing of software license agreements. In M. B. Rosson and D. J. Gilmore, editors, *CHI*, pages 607–616. ACM, 2007.

[62] N. Good and A. Krekelberg. Usability and privacy: a study of kazaa p2p file-sharing. In G. Cockton and P. Korhonen, editors, *CHI*, pages 137–144. ACM, 2003.

[63] L. A. Gordon and M. P. Loeb. The economics of information security investment. *ACM Trans. Inf. Syst. Secur.*, 5(4):438–457, 2002.

[64] O. A. Gross and R. A. Wagner. A continuous colonel blotto game. *RAND Corporation RM–408*, 1950.

[65] J. Grossklags, N. Christin, and J. Chuang. Secure or Insure? A game-theoretic analysis of information security games. In J. Huai, R. Chen, H.-W. Hon, Y. Liu, W.-Y. Ma, A. Tomkins, and X. Zhang, editors, *WWW*, pages 209–218. ACM, 2008.

[66] J. Grossklags, B. Johnson, and N. Christin. When information improves information security. In R. Sion, editor, *Financial Cryptography*, volume 6052 of *Lecture Notes in Computer Science*, pages 416–423. Springer, 2010.

[67] G. Hardin. The tragedy of the commons. *Science*, 162:1243–47, 1968.

[68] K. Hausken. Returns to information security investment: The effect of alternative information security breach functions on optimal investment and sensitivity to vulnerability. *Information Systems Frontiers*, 8(5):338–349, 2006.

[69] C. Herley and D. Florêncio. A profitless endeavor: phishing as tragedy of the commons. In *Proceedings of the 2008 workshop on New security paradigms*, NSPW '08, pages 59–70, New York, NY, USA, 2008. ACM.

[70] C. Herley and D. Florêncio. Nobody sells gold for the price of silver: Dishonesty, uncertainty and the underground economy. In T. Moore, D. Pym, and C. Ioannidis, editors, *Economics of Information Security and Privacy*, pages 33–53. Springer US, 2010.

[71] J. Hirshleifer. From weakest-link to best-shot: The voluntary provision of public goods. *Public Choice*, 41(3):371–386, 1983.

[72] A. Jøsang, B. AlFayyadh, T. Grandison, M. A. Zomai, and J. McNamara. Security usability principles for vulnerability analysis and risk assessment. In *ACSAC*, pages 269–278. IEEE Computer Society, 2007.

[73] A. Jøsang, M. A. Zomai, and S. Suriadi. Usability and privacy in identity management architectures. In L. Brankovic, P. D. Coddington, J. F. Roddick, C. Steketee, J. R. Warren, and A. L. Wendelborn, editors, *ACSW Frontiers*, volume 68 of *CRPIT*, pages 143–152. Australian Computer Society, 2007.

[74] C. Kanich, N. Weaver, D. McCoy, T. Halvorson, C. Kreibich, K. Levchenko, V. Paxson, G. M. Voelker, and S. Savage. Show me the money: Characterizing spam-advertised revenue. In *USENIX Security Symposium*. USENIX Association, 2011.

[75] A. Kerckhoffs. La cryptographie militaire. *Journal des sciences militaires*, IX(1):5–83, 1883. English translation available at: `http://www.petitcolas.net/fabien/kerckhoffs/`.

[76] J. King, A. Lampinen, and A. Smolen. Privacy: Is there an app for that? In *Proceedings of the Seventh Symposium on Usable Privacy and Security*, SOUPS '11, pages 12:1–12:20, New York, NY, USA, 2011. ACM.

[77] A. Kittur, E. Chi, B. A. Pendleton, B. Suh, and T. Mytkowicz. Power of the few vs. wisdom of the crowd: Wikipedia and the rise of the bourgeoisie. *Algorithmica*, 1(2):1–9, 2007.

[78] A. Kittur and R. E. Kraut. Harnessing the wisdom of crowds in wikipedia: quality through coordination. In B. Begole and D. W. McDonald, editors, *CSCW*, pages 37–46. ACM, 2008.

[79] K. Kostiainen, E. Reshetova, J.-E. Ekberg, and N. Asokan. Old, new, borrowed, blue – A perspective on the evolution of mobile platform security architectures. In R. S. Sandhu and E. Bertino, editors, *CODASPY*, pages 13–24. ACM, 2011.

[80] M. Lelarge and J. Bolot. Economic incentives to increase security in the internet: The case for insurance. In *INFOCOM*, pages 1494–1502. IEEE, 2009.

[81] N. Leontiadis, T. Moore, and N. Christin. Measuring and analyzing search-redirection attacks in the illicit online prescription drug trade. In *USENIX Security Symposium*. USENIX Association, 2011.

[82] K. Levchenko, A. Pitsillidis, N. Chachra, B. Enright, M. Félegyházi, C. Grier, T. Halvorson, C. Kanich, C. Kreibich, H. Liu, D. McCoy, N. Weaver, V. Paxson, G. M. Voelker, and S. Savage. Click trajectories: End-to-end analysis of the spam value chain. In *IEEE Symposium on Security and Privacy*, pages 431–446. IEEE Computer Society, 2011.

[83] D. G. Lichtman and E. A. Posner. Holding internet service providers accountable. *U Chicago Law & Economics, Olin Working Paper No. 217*, July 2004. Available at SSRN: `http://ssrn.com/abstract=573502`.

[84] P. Liu and W. Zang. Incentive-based modeling and inference of attacker intent, objectives, and strategies. In S. Jajodia, V. Atluri, and T. Jaeger, editors, *ACM Conference on Computer and Communications Security*, pages 179–189. ACM, 2003.

[85] A. Lizzeri. Information revelation and certification intermediaries. *The RAND Journal of Economics*, 30(2):214–231, 1999.

[86] L. Mamykina, B. Manoim, M. Mittal, G. Hripcsak, and B. Hartmann. Design lessons from the fastest Q&A site in the west. In D. S. Tan, S. Amershi, B. Begole, W. A. Kellogg, and M. Tungare, editors, *CHI*, pages 2857–2866. ACM, 2011.

[87] M. Manshaei, Q. Zhu, T. Alpcan, T. Basar, and J.-P. Hubaux. Game Theory Meets Network Security and Privacy. Technical report, EPFL, Lausanne, 2010. Available at: `http://infoscience.epfl.ch/record/151965`.

[88] M. Marsall. How HTML5 will kill the native app. Article on VentureBeat website, April 2011. `http://venturebeat.com/2011/04/07/how-html5-will-kill-the-native-app/`. Last accessed: June 2012.

[89] K. Matsuura. Productivity space of information security in an extension of the gordon-loeb's investment model. In *Proceedings of the 7th Workshop on the Economics of Information Security*, WEIS '08, 2008.

[90] D. K. McGrath and M. Gupta. Behind phishing: An examination of phisher modi operandi. In F. Monrose, editor, *LEET*. USENIX Association, 2008.

[91] T. Moore and R. Clayton. Examining the impact of website take-down on phishing. In L. F. Cranor, editor, *eCrime Researchers Summit*, volume 269 of *ACM International Conference Proceeding Series*, pages 1–13. ACM, 2007.

[92] T. Moore and R. Clayton. The consequence of noncooperation in the fight against phishing. In *Proceedings of the 3rd APWG eCrime Researchers Summit*, eCrime '08, pages 1–14, 2008.

[93] T. Moore and R. Clayton. Evaluating the wisdom of crowds in assessing phishing websites. In G. Tsudik, editor, *Financial Cryptography*, volume 5143 of *Lecture Notes in Computer Science*, pages 16–30. Springer, 2008.

[94] T. Moore and R. Clayton. The impact of incentives on notice and takedown. In M. Johnson, editor, *Managing Information Risk and the Economics of Security*. Springer, 2008.

[95] T. Moore and R. Clayton. Evil searching: Compromise and recompromise of internet hosts for phishing. In R. Dingledine and P. Golle, editors, *Financial Cryptography*, volume 5628 of *Lecture Notes in Computer Science*, pages 256–272. Springer, 2009.

[96] T. Moore and B. Edelman. Measuring the perpetrators and funders of typosquatting. In R. Sion, editor, *Financial Cryptography*, volume 6052 of *Lecture Notes in Computer Science*, pages 175–191. Springer, 2010.

[97] F. Ortega, J. M. González-Barahona, and G. Robles. On the inequality of contributions to wikipedia. In *HICSS*, page 304. IEEE Computer Society, 2008.

[98] F. Petitcolas. English Translation of Kerckhoffs's Principles in 'La cryptographie militaire'. Available at: `http://petitcolas.net/fabien/kerckhoffs/`. Last accessed: June 2012.

[99] G. Pitsilis and P. H. Chia. Does trust matter for user preferences? a study on epinions ratings. In M. Nishigaki, A. Jøsang, Y. Murayama, and S. Marsh, editors, *IFIPTM*, volume 321 of *IFIP Conference Proceedings*, pages 232–247. Springer, 2010.

[100] N. Provos, P. Mavrommatis, M. A. Rajab, and F. Monrose. All Your iFRAMEs Point to Us. In P. C. van Oorschot, editor, *USENIX Security Symposium*, pages 1–16. USENIX Association, 2008.

[101] D. Recordon and D. Reed. OpenID 2.0: a platform for user-centric identity management. In A. Juels, M. Winslett, and A. Goto, editors, *Digital Identity Management*, pages 11–16. ACM, 2006.

[102] B. Roberson. The colonel blotto game. *Economic Theory*, 29(1):1–24, Sept. 2006.

[103] S. Ross. Security through usability. Securius Newsletter, 4(1), Feb 2003. Available at: `http://www.securius.com/newsletters/Security_Through_Usability.html`. Last accessed: June 2012.

[104] Y. E. Sagduyu, R. Berry, and A. Ephremides. Mac games for distributed wireless network security with incomplete information of selfish and malicious user types. In *Proceedings of the First ICST international conference on Game Theory for Networks*, GameNets'09, pages 130–139. IEEE Press, 2009.

[105] K. Sallhammar, B. E. Helvik, and S. J. Knapskog. Towards a stochastic model for integrated security and dependability evaluation. In *ARES*, pages 156–165. IEEE Computer Society, 2006.

[106] J. H. Saltzer and M. D. Schroeder. The protection of information in computer systems. *Proceedings of the IEEE*, 63(9):1278–1308, 1975.

[107] S. E. Schechter, R. Dhamija, A. Ozment, and I. Fischer. The emperor's new security indicators. In *IEEE Symposium on Security and Privacy*, pages 51–65. IEEE Computer Society, 2007.

[108] S. Sheng, B. Magnien, P. Kumaraguru, A. Acquisti, L. F. Cranor, J. I. Hong, and E. Nunge. Anti-phishing phil: the design and evaluation of a game that teaches people not to fall for phish. In L. F. Cranor, editor, *SOUPS*, volume 229 of *ACM International Conference Proceeding Series*, pages 88–99. ACM, 2007.

[109] J. Stoll, C. S. Tashman, W. K. Edwards, and K. Spafford. Sesame: informing user security decisions with system visualization. In M. Czerwinski, A. M. Lund, and D. S. Tan, editors, *CHI*, pages 1045–1054. ACM, 2008.

[110] X. Suo, Y. Zhu, and G. S. Owen. Graphical passwords: A survey. In *ACSAC*, pages 463–472. IEEE Computer Society, 2005.

[111] J. Surowiecki. *The wisdom of crowds*. Anchor Books, 2005.

[112] Z. Tang, Y. J. Hu, and M. D. Smith. Gaining trust through online privacy protection: Self-regulation, mandatory standards, or caveat emptor. *Journal of Management Information Systems*, 24(4):153–173, 2008.

[113] E. Uzun, K. Karvonen, and N. Asokan. Usability analysis of secure pairing methods. In S. Dietrich and R. Dhamija, editors, *Financial Cryptography*, volume 4886 of *Lecture Notes in Computer Science*, pages 307–324. Springer, 2007.

[114] H. Varian. Managing online security risks. New York Times (Jun 1, 2000), Jun 2000. Available at: `http://people.ischool.berkeley.edu/~hal/people/hal/NYTimes/2000-06-01.html`. Last accessed: June 2012.

[115] H. Varian. System reliability and free riding. In L. Camp and S. Lewis, editors, *Economics of Information Security*, volume 12 of *Advances in Information Security*, pages 1–15. Kluwer Academic Publishers, 2004.

[116] G. Wang, C. Wilson, X. Zhao, Y. Zhu, M. Mohanlal, H. Zheng, and B. Y. Zhao. Serf and turf: crowdturfing for fun and profit. In A. Mille, F. L. Gandon, J. Misselis, M. Rabinovich, and S. Staab, editors, *WWW*, pages 679–688. ACM, 2012.

[117] B. Westermann, P. H. Chia, and D. Kesdogan. Analyzing the gold star scheme in a split tor network. In *Proceedings of the 8th International Conference on Security and Privacy in Communication Networks (SecureComm)*, 2011.

[118] A. Whitten and J. D. Tygar. Why johnny can't encrypt: A usability evaluation of pgp 5.0. In *USENIX Security Symposium*. USENIX Association, 1999.

[119] D. M. Wilkinson. Strong regularities in online peer production. In L. Fortnow, J. Riedl, and T. Sandholm, editors, *ACM Conference on Electronic Commerce*, pages 302–309. ACM, 2008.

[120] M. S. Wogalter. Communication-Human Information Processing (C-HIP) Model. In M. S. Wogalter, editor, *Handbook of Warnings*, pages 51–61. Lawrence Erlbaum Associates, Mahwah, NJ, 2006.

[121] G. Wondracek, T. Holz, C. Platzer, E. Kirda, and C. Kruegel. Is the Internet for Porn? An Insight into the Online Adult Industry. In *Proceedings of the 9th Workshop on the Economics of Information Security*, WEIS '10, 2010.

[122] M. Wu, R. C. Miller, and S. L. Garfinkel. Do security toolbars actually prevent phishing attacks? In R. E. Grinter, T. Rodden, P. M. Aoki, E. Cutrell, R. Jeffries, and G. M. Olson, editors, *CHI*, pages 601–610. ACM, 2006.

[123] K.-P. Yee. Aligning security and usability. *IEEE Security & Privacy*, 2(5):48–55, 2004.

[124] W. Zhou, Y. Zhou, X. Jiang, and P. Ning. Detecting repackaged smartphone applications in third-party android marketplaces. In E. Bertino and R. S. Sandhu, editors, *CODASPY*, pages 317–326. ACM, 2012.

[125] Y. Zhou and X. Jiang. Dissecting android malware: Characterization and evolution. In *IEEE Symposium on Security and Privacy*. IEEE Computer Society, 2012.

[126] Y. Zhou, Z. Wang, W. Zhou, and X. Jiang. Hey, you, get off of my market: Detecting malicious apps in official and alternative Android markets. In *NDSS*. The Internet Society, 2012.

A. Colonel Blotto in the Phishing War

Author

Pern Hui Chia, Q2S NTNU
John Chuang, UC Berkeley

Conference

2nd Conference on Decision and Game Theory for Security (GameSec)
14–15 November 2011, University of Maryland, College Park, USA

Abstract

Phishing exhibits characteristics of asymmetric conflict and guerrilla warfare. Phishing sites, upon detection, are subject to removal by takedown specialists. In response, phishers create large numbers of new phishing attacks to evade detection and stretch the resources of the defenders. We propose the Colonel Blotto Phishing (CBP) game, a two-stage Colonel Blotto game with endogenous dimensionality and detection probability. We find that the optimal number of new phishes to create, from the attacker's perspective, is influenced by the degree of resource asymmetry, the cost of new phishes, and the probability of detection. Counter-intuitively, we find that it is the less resourceful attacker who would create more phishing attacks in equilibrium. And depending on the detection probability, an attacker will vary his strategies to either create even more phishes, or to focus on raising his resources to increase the chance he will extend the lifetime of his phishes. We discuss the implications to anti-phishing strategies and point out that the game is also applicable to web security problems more generally.

A.1. Introduction

Phishing, among other web security issues, has remained a tricky problem today. While it is non-trivial to measure the exact financial losses due to phishing, and that many estimated loss figures appear overstated [9], the damage inflicted by phishing activities is never negligible. Realizing that technical sophistication alone will not be sufficient to fend off phishing activities, over the past few years, researchers have started to look at the ecosystem and modi operandi of phishing activities.

McGrath and Gupta found that phishers misuse free web hosting services and URL-aliasing services, and that phishing domains are hosted across multiple countries with a significant percentage of hosts belonging to residential customers [13]. Moore and Clayton identified different types of phishing attacks according to the way a phishing site is hosted [16]. The most common hosting vectors were found to be compromised web servers and free web-hosting services. While system admins and hosting companies are usually cooperative and quick to take down the phishing pages once notified, noticing them in the first place is challenging [16]. Moreover, victim servers were found to be re-compromised by the attackers to host phishing pages as the vulnerabilities of the servers remain unpatched [17]. Two notorious gangs, known as 'Rock Phish' and 'Avalanche'[1] even showed much technical sophistication in their massive and concerted phishing attacks. Both gangs exploited malware-infested machines and the fast flux method (mapping the domain name to different IP addresses (of different bots) by changing the DNS records in a high frequency) to extend the lifetime of a phishing site. Taking down the phishing pages from a large number of bots is extremely difficult, especially when the ISPs have only limited control and responsibility over malware-infested machines. This forces the defender to takedown the phishing sites by suspending the phishing domain names with the help from registrars and registries.

The above highlights several important challenges in defending against phishing activities. First, it is challenging to detect all phishing attacks out there. Second, taking down phishing attacks that have been identified (e.g., to remove the phishing sites, or to ensure that a vulnerable web server is patched to prevent re-compromise) is also non-trivial. The situation is worsened by a lack of information sharing in the anti-phishing industry [16]. Meanwhile, despite a spike in the count of phishing attacks[2] in 2009 due to the Avalanche gang [2], the number of unique phishing domains found (per six months) has remained steady at around 30,000 over the past few years, except in the second half of 2010 where 43,000 unique phishing domain names were recorded partly due to new data inputs from the China Internet Network Information Center (CNNIC) who operates the `.cn` registry [3].[3] This

[1] An account of the modi operandi of the Rock Phish and Avalanche gangs can be found in [14] and [2] respectively.

[2] An attack is defined by Anti-Phishing Working Group (APWG) as a unique phishing site targeting a specified brand.

[3] Measurement of unique phishing attacks, uptime of phishing sites and in-depth surveys on the trends and domain name use by phishing sites can be found in a series of reports (e.g., [2, 3]) by the APWG on `http://www.antiphishing.org`.

suggests that the phishers do factor in the cost consideration when carrying out phishing attacks.

Different from prior studies that have largely taken the empirical approach, we propose in this work a theoretical model to aid researchers and policymakers in better analyzing the different aspects of phishing defense. We build on the Colonel Blotto game, an old but interesting game that has been largely neglected due to its complexity, until the recent work by Roberson [18] which gives a complete characterization to the unique equilibrium payoffs of a two-player asymmetric Colonel Blotto game. The game is particularly suitable to capture the resource allocation problem between a phisher and a defender with asymmetrical resources. In addition to mapping the phishing problem into the Colonel Blotto game, our model extends the two-stage Colonel Blotto game in [10] to include a detection probability to factor in the consideration of asymmetric information that not all phishes will be known to the defender. We regard the defender in this work as a takedown company (e.g., MarkMonitor[4], BrandProtect[5] and Internet Identity[6]) that has been contracted by its clients (e.g., financial institutions, e-commerce services) to remove phishing sites that masquerade as the clients' legitimate sites. Although the defender is in a disadvantage position for not being able to detect all phishes that have been created, and that the attacker can always exploit the next weakest link whenever a phishing server is taken down, we expect that the defender can garner more resources than the attackers from the contract with its clients, plus the support from the ISPs, service providers, law enforcers, registrars and registries.

In the following, we first give a quick introduction to the Colonel Blotto game and related work in Section A.2. We propose the Colonel Blotto Phishing (CBP) game in Section A.3 to model phishing attack and defense. We present the results from our analysis based on the CBP model in Section A.4. And lastly, we discuss the implications to the anti-phishing strategies in Section A.5.

A.2. Background and Related Work

The *Colonel Blotto* game was first introduced in 1921 by Borel [6] as a two-player constant-sum game, where the players strategically distribute a fixed and *symmetrical* amount of resources over a finite number of n contests (battlefields). The player who expends a higher amount of resources in a contest wins that particular battlefield, similar to an all-pay auction. The objective of the players is to maximize the number of battlefields won. Gross and Wagner [8] in 1950 described the game with *asymmetrical* resources between the two players, but have only solved the case where the number of battlefields $n = 2$.

The complexity for the case when there are $n \geq 3$ battlefields and the lack of pure strategies have arguably led to the Colonel Blotto game being largely neglected by the research community. A resurgence of interests in the Colonel Blotto game

[4] `http://www.markmonitor.com`
[5] `http://www.brandprotect.com`
[6] `http://internetidentity.com`

(e.g., [4, 5, 7, 11, 12, 19]) follows the recent work by Roberson [18] which has successfully characterized the unique equilibrium payoffs for all configurations of resource asymmetry, and the equilibrium resource allocation strategies (for most configurations) of a constant-sum Colonel Blotto game with $n \geq 3$ battlefields. Roberson and Kvasov have later studied the non-constant-sum version in [19]. We summarize the main results from Roberson [18] below:

Theorem 1 (case a, b and c correspond to Theorem 2, 3 and 5 in [18])
Let n denote the number of battlefields, while R_w and R_s denote the resources of the weak (w) and strong (s) players respectively such that $R_w \leq R_s$, the Nash equilibrium univariate distribution functions (for allocating resources to individual battlefields strategically), and the unique equilibrium payoffs (measured in the expected proportion on battlefields won), depending on the $\frac{R_w}{R_s}$ ratio and the number of battlefields n, are given in the following:

case a: $\frac{2}{n} \leq \frac{R_w}{R_s} \leq 1$
In the unique Nash equilibrium, player w and s allocate x_j resources in each battlefield $j \in \{1, ..., n\}$ based on the following univariate distribution functions:

$$F_{w,j}(x) = (1 - \tfrac{R_w}{R_s}) + \tfrac{nx}{2R_s}(\tfrac{R_w}{R_s}) \quad , \ x \in [0, \tfrac{2R_s}{n}]$$
$$F_{s,j}(x) = \tfrac{nx}{2R_s} \quad , \ x \in [0, \tfrac{2R_s}{n}]$$

The unique equilibrium payoffs (expected proportions of battlefields won) of player w and s are independent of the number of battlefields, given as follows:

$$\pi_w = \tfrac{R_w}{2R_s}$$
$$\pi_s = 1 - \tfrac{R_w}{2R_s}$$

case b: $\frac{1}{n-1} \leq \frac{R_w}{R_s} < \frac{2}{n}$
In the unique Nash equilibrium, player w and s allocate x_j resources in each battlefield $j \in \{1, ..., n\}$ based on the following univariate distribution functions:

$$F_{w,j}(x) = (1 - \tfrac{2}{n}) + \tfrac{x}{R_w}(\tfrac{2}{n}) \quad , \ x \in [0, R_w]$$
$$F_{s,j}(x) = \begin{cases} (1 - \frac{R_s}{nR_w})(\frac{2x}{R_w}) & , \ x \in [0, R_w) \\ 1 & , \ x \geq R_w \end{cases}$$

The expected proportions of battlefields won by player w and s are as follows:

$$\pi_w = \tfrac{2}{n} - \tfrac{2R_s}{n^2 R_w}$$
$$\pi_s = 1 - \tfrac{2}{n} + \tfrac{2R_s}{n^2 R_w}$$

case c: $\frac{1}{n} < \frac{R_w}{R_s} < \frac{1}{n-1}$
In a Nash equilibrium, player w allocates zero resources to $n-2$ of the battlefields, each randomly chosen with equal probability. On the remaining 2 battlefields, he randomizes the resource allocation over a set of bivariate mass points. On the other

hand, player s allocates R_w resources to each of $n-2$ randomly chosen battlefields. On the remaining 2 battlefields, player s also randomizes the resource allocation over a set of bivariate mass points. Let $m = \lceil \frac{R_w}{R_s - R_w(n-1)} \rceil$ such that $2 \leq m < \infty$, the unique expected proportions of battlefields won by player w and s are given as follows:

$$\pi_w = \tfrac{2m-2}{mn^2}$$
$$\pi_s = 1 - \tfrac{2m-2}{mn^2}$$

Note that the univariate distribution functions constitute the players' mixed strategies in Nash equilibrium. The allocation of resources across the n battlefields must additionally be contained in the set of all feasible allocations $\left\{\mathbf{x} \in \mathbb{R}_+^n | \sum_{j=1}^n x_{i,j} \leq R_i\right\}$ where $i = w, s$.[7] In general, player s uses a stochastic 'complete coverage' strategy (which expends non-zero resources in all battlefields, and locks down in a random subset of battlefields by allocating R_w resources to them in case b and c), while player w uses a stochastic 'guerrilla warfare' strategy (which optimally abandons a random subset of the battlefields). Despite the resource asymmetry, player w can expect to win a non-zero proportion of the battlefields, except in the case of $R_s \geq nR_w$, where the player s can trivially lock down (win) all battlefields by allocating R_w resources to each of them.

Note that also the proportion of battlefields won by the player w is a function of n in the case b and c of Theorem 1. In a recent work, Kovenock et al. [10] presented a two-stage Colonel Blotto game which endogenizes the dimensionality of the classic Colonel Blotto game, allowing the players to create additional battlefields in the additional 'pre-conflict' stage. They show that with such possibility, player w will optimally increase the number of battlefields in the 'pre-conflict' stage, given a low battlefield creation cost, so to thin the defender's resources and reduce the number of battlefields player s can lock down in the 'conflict' stage. We outline the main results from [10] below:

Theorem 2 (see Theorem 2 in [10])
In the pre-conflict stage of the game with n_0 initial battlefields and resource asymmetry that satisfies $\frac{1}{n_0-1} \leq \frac{R_w}{R_s} \leq 1$, assuming that the cost to create additional battlefields, c is strictly increasing and strictly convex, the optimal numbers of new battlefields that player w and s will create, n_w^* and n_s^* respectively, in the subgame perfect equilibrium, are given as follows:

<u>case a:</u> If $\frac{R_w}{R_s}$ satisfies $\frac{2}{n_0} \leq \frac{R_w}{R_s} \leq 1$, then $n_s^* = n_w^* = 0$.

<u>case b:</u> If $\frac{R_w}{R_s}$ satisfies $\frac{1}{n_0-1} \leq \frac{R_w}{R_s} < \frac{2}{n_0}$, then $n_s^* = 0$, and let $n_{wr} \in (0, \frac{2R_s}{R_w} - n_0)$ denotes the real number that solves:

$$-\tfrac{2}{(n_0+n_{wr})^2} + \tfrac{4R_s}{R_w(n_0+n_{wr})^3} - c'_{n_{wr}} = 0$$

[7] We refer interested readers to Roberson [18] for proofs and details on how the equilibrium univariate distribution functions give a n-variate joint distribution function satisfying the constraint that $\sum_{j=1}^n x_{i,j} \leq R_i$ where $i = w, s$.

then, n_w^* is either $\lceil n_{wr} \rceil$ or $\lfloor n_{wr} \rfloor$ depending on which of it results in a higher utility for player w, given $n_s^* = 0$.

Note that Theorem 2 has not formally treated the case *c* of Theorem 1. The analysis of case *c* will be more complicated as the expected proportion of battlefields won by both players have points of discontinuity, but the underlying intuition is the same as case *b*. [10] Note that also Theorem 2 assumes that the cost of creating additional battlefields is expended separately from the players' resources.

A.3. Modeling

With an introduction to the classic Colonel Blotto game and the extension with endogenous dimensionality, we are now ready to model the economics for phishing activities in this section. We will first apply the classic Colonel Blotto game to phishing attack and defense. Then, we will extend the game to model endogenous dimensionality following the two-stage Colonel Blotto game in [10], and asymmetric information using an additional detection probability to reflect that not all phishes will be known to the defender in practice.

A.3.1. Applying Colonel Blotto to Phishing

We map the basics the Colonel Blotto game in the context of phishing attack and defense in the following.

Players. Like the classic Colonel Blotto, we consider here a two-player constant-sum game between a phisher and a defender. We regard the defender here to be a takedown company such as MarkMonitor, BrandProtect and Internet Identity as aforementioned. The takedown company is contracted by its clients, including banks and popular brand owners, to remove phishing sites attacking the clients' brands. On the other hand, the phisher plays to keep alive the phishing sites, or to launch new attacks, to victimize as many users he can.

Resources. We assume the phisher to be the weak player (w) and the takedown company to be the strong player (s). Although this may be debatable, assuming such resource asymmetry is reasonable if we consider that takedown companies will usually maintain good contacts with and can thus get assistance from the ISPs, service providers, law enforcers, registrars and registries in the process of taking down the phishes. By resources, we thus mean not financial figures but mainly the *technologies*, *infrastructure* (e.g., access to a botnet), *time* and *manpower*.[8] Phisher's profitability is also not as lucrative as it appears in the news. A number of estimates on the losses due to phishing attacks have been criticized to be overstated [9]. The resources, R_s and R_w respectively, are finite with $R_s \geq R_w$. They are of the 'use-it-or-lose-it' nature, meaning that unused resources will give no value to the players in the end of the game.

[8] Resource asymmetry should not be confused with asymmetry in coverage where the defender needs to protect all assets while the attacker can target any of them.

Battlefields. We define a battlefield to be a unique phishing site (a fully qualified domain name or IP address, or a site on a shared hosting service) targeting a specific brand, following the definition of a phishing attack by APWG (see e.g., page 4 in [3]). Different URLs directing to the same phishing page, crafted to evade spam filters or to trick the URL-based anti-phishing toolbars, are considered the same battlefield. Defined this way, creating a battlefield hence involves some costs ranging from *low* (e.g., to register a subdomain on a shared hosting service, to copy the login page of a brand) to *high* (e.g., to register a new domain name, to compromise a vulnerable web server). In this paper, we use the terminologies 'a phish' and 'a phishing attack' interchangeably.

Objectives & Contests. We model the objective of the phisher and the defender to be maximizing the expected proportion of phishing attacks kept online and taken down, respectively. We consider that either the phisher or the defender can outperform the other party to win a battlefield by allocating more resources to it. And given that we have not factored in the uptime and the number of victims per attack in our model, we loosely define that a specific battlefield (phishing attack) is won by the phisher if the phish has a *long enough uptime*. For example, having the resources of a botnet infrastructure, an attacker can use 'fast-flux' IP addresses and malware-controlled proxies, to make it hard for the defender to take down the phishing server, prolonging the uptime of the phishes, as the defender will have to turn to the responsible registrar or registry to suspend the domain name. We elaborate on other tricks used by phishers, including the two infamous Rock Phish and Avalanche gangs, in Section A.3.2.

Given the above configurations, we can already gain a number of useful insights. For example, we can expect that there will be always some phishes that will have long uptime unless that the defender is much more resourceful than the phisher (i.e., $R_s \geq nR_w$). However, the classic colonel blotto game alone does not describe the practical scenario quite yet. Why are there a large number of phishing attacks instead of just a few? Indeed, it is to the phisher's advantage to create an optimal number of additional phishes (battlefields), so to thin the defender's resources in removing each of them. Furthermore, how does the asymmetric information affect the strategies of the phisher? We extend the two-stage Colonel Blotto game in [10] to include an additional parameter, the expected probability of detection P_d, to reflect that not all phishes will be known to the defender – a major challenge in the anti-phishing industry. [16]

A.3.2. The Colonel Blotto Phishing Game

We name our model as the Colonel Blotto Phishing (CBP) game. It consists of two stages: (i) create–detect, (ii) resist–takedown, similar to the 'pre-conflict' and 'conflict' stages in [10]. Table A.1 summarizes the flow of the CBP game. We detail on the game stages in the following.

Stage 1: Create–Detect. We consider that game starts with the phisher having a number of phishes n_0 that are known to the defender, and both players are allowed to increase the dimensionality of the game by introducing new battlefields

	Stage	Phisher (w)	Defender (s)
i)	create – detect	a. create and market n_w^* new phishes b. learn about detection	a. detect new phishes b. publish findings
ii)	resist – takedown	c. expend ε resources to undetected phishes, while allocating $R_w - \varepsilon$ resources to phishes known to the defender to resist removal	c. expend all R_s resources strategically to remove the newly detected and known existing phishes in a promptly manner

Table A.1.: The flow of the Colonel Blotto Phishing game.

in the first stage. Obviously, the defender will not create any phishes. However, it is to the phisher's advantage to create a number of new phishing attacks n_w so to stretch the defender's resources, in hope to increase the expected proportion of phishes that will stay online for more than a certain period of time. Hence, we have the total phishing attacks $n = n_0 + n_w$. We expect the phisher then advertises the newly created phishes through spams and online social networks.[9] We assume a linear cost c for creating and advertising the new phishes; c can be low or high depending on the way the phisher carries out the attack (e.g., through free subdomain services, paying for a newly registered domain, taking the effort to hack a vulnerable web server, and so on).

A new aspect we incorporate into the classic Colonel Blotto game is the situation where some of the newly created phishes might not be detected by the takedown company. We analyze both cases where the expected detection probability P_d is (i) exogenously determined, and (ii) endogenously influenced by the number of new phishing attacks in Section A.4. The expected proportions of phishes that trivially get away undetected, or that will possibly stay online long enough depending on the resource allocations of both the phisher and defender in the second stage, are depicted in Figure A.1. In practice, takedown companies learn about new phishing attacks through their own infrastructures (e.g., spam filters) in addition to 'raw' feeds bought, negotiated or obtained from the ISPs or phishing clearinghouses, such as the APWG and PhishTank[10].

An assumption we make here is that the phisher will then learn about which of his phishes have been detected before proceeding to the next game stage. This is reasonable, regardless of whether the takedown company shares their detection results[11], as we expect that the phisher can achieve this using public clearinghouses

[9] McGrath and Gupta [13] observed that most domains created for phishing become active almost immediately upon registration.

[10] PhishTank – a community based phishing collator. `http://www.phishtank.com`

[11] Individual takedown companies often will validate the 'raw' URLs of potential phishes to remove false positives, and they might not voluntarily share their validated feeds for competitive advantages. Moore and Clayton showed how sharing of phishing data could have helped to

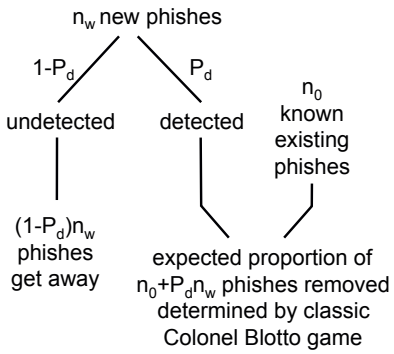


Figure A.1.: Expected proportion of phishes in different states.

(e.g., phishtank) or through anti-phishing APIs that come with modern browsers (e.g., Google Safe Browsing API[12] for FireFox and Chrome).

Stage 2: Resist–Takedown. Knowing the identity of the detected phishes $\mathbb{J}_d$, the optimal move for the phisher in the second stage is hence to expend all his resources strategically on the detected phishes only, so to resist the takedown process. Here, we assume that the resources (e.g., technologies, infrastructure, manpower) are of the 'use-it-or-lose-it' nature, typical to a constant-sum game. In other words, unused resources will give no value to the players. We further assume that the phisher will optimally allocate $\varepsilon \approx 0$ resources for the undetected phishes $j \notin \mathbb{J}_d$ given that the defender does not know about them. We note that this assumption is reasonable as the resources are finite.

We regard that either the phisher or the takedown company will 'succeed' with respect to a particular phishing attack depending on the amount of resources they put in: the player who expends more resources wins. Specifically, with $x_{i,j}$ and $x_{-i,j}$ denoting the amount of resources player $i \in \{w, s\}$ and his opponent puts into the phish attack j respectively, the success of player i at attack j is given by:

$$\pi_{i,j}(x_{i,j}, x_{-i,j}) = \begin{cases} 0 & \text{if } x_{i,j} < x_{-i,j} \\ 1 & \text{if } x_{i,j} > x_{-i,j} \end{cases}$$

where in the case of $x_{w,j} = x_{s,j}$ (a tie), we assume that defender s will succeed in taking down the attack promptly. As for undetected phishes, i.e., $\forall j \notin \mathbb{J}_d$, we regard that $x_{s,j} = 0$ and the phisher will trivially win the battlefield with $x_{w,j} = \varepsilon$ resources.

Can the phisher still win in an already detected phish in practice? While it may not be intuitive at first, the answer is 'yes' given our definition that a phishing attack is won by the phisher (defender) if the phish has an uptime more (less) than a certain threshold. The longer a phish can resist being removed, the

halve the lifetime of phishes, translating to a potential loss mitigation of $330 million per year, based on data feeds from two takedown companies [15].

[12] `http://code.google.com/apis/safebrowsing/`

more users could fall victim to it. While a weak phisher may simply abandon his phishes (given that he cannot win) when facing a much more resourceful defender (i.e., when $R_s \geq nR_w$), there have been practical examples of how a skilled phisher attempts to extend the lifetime of his phishes via different tricks. For example, a phisher may configure his phishes not to resolve on every access so to misguide the defender, but remain online to trick more users (see e.g., [3], footnote 5). The phisher may also temporarily remove the phishing pages from a compromised web server so to avoid further actions from the defender or admin (e.g., to patch up specific vulnerabilities) and re-plant the phishes at a later time. Indeed, APWG (see e.g., [3], footnote 5) finds that more than 10% of phishes are re-activated after being down for more than an hour. Moore and Clayton also found that 22% of all compromised web servers are re-compromised within 24 weeks to be used as the host for phishing sites [16].

With more resources, a phisher can even increase technical sophistication so to use malware-controlled proxies and fast-flux IP addresses as demonstrated the large-scale attacks by the infamous 'Rock Phish' and 'Avalanche' phishing gangs. The fast-changing nature of IP address that the phishing site resolving to indicate that the attacker has in control of a large number of compromised machines (bots) make it infeasible for the takedown company and the responsible ISPs to take the phishing servers offline promptly. Instead, the defender will have to work towards suspending the domain names in use, which could take a while if the responsible registrars are not responsive or have limited experience in abuse control. The 'Avalanche' gang was found to have exploited this; at the same time as they launched their massive attacks using domains bought from a few registrars (resellers), the gang scouted for other unresponsive registrars for future use (see page 7 of [2]). Meanwhile, in [14] Moore and Clayton found that the fast-flux phishing gang used 57 domain names and 4287 IP addresses for fast-flux phishing. The 1:75 skewed ratio is interesting as it suggests that the fast-flux phishing gang was highly resourceful (having access to a botnet infrastructure). However, we note that these resources are not unlimited. For example, the operations of the 'Avalanche' gang was disrupted as the security community affected a 'temporary' shut-down of the botnet infrastructure in Nov 2009 [2]. Later, although the gang managed to re-establish a new botnet, they were also found to prefer using their resources for a more profitable opportunity to distribute the Zeus malware, which has been designed to automate identity theft and facilitate unauthorized transactions. [3]

Subgame Perfect Equilibrium. We consider the objective of the phisher (the takedown company) is to maximize the proportion of phishes that he succeeds in keeping alive for a certain period (removing promptly), minus the cost for creating new phishing attacks. With $\mathbf{x}_i$ and $\mathbf{x}_{-i}$ denoting the resource allocations across all phishing sites by player $i \in \{w, s\}$ and his opponent respectively, the utility of player i can be written as:

$$U_i(\{\mathbf{x}_i, n_i\}, \{\mathbf{x}_{-i}, n_{-i}\}) = \frac{1}{n}(\sum_{j \in \mathbb{J}_d} \pi_{i,j} + \sum_{j \notin \mathbb{J}_d} \pi_{i,j}) - cn_i$$

Note that $\mathbf{x}_i$ and $\mathbf{x}_{-i}$ must be contained in the set of all feasible allocations, given by $\{\mathbf{x}_i \in \mathbb{R}^n_+ | \sum_{j=1}^n x_{i,j} \leq R_i\}$.

The optimal number of new phishes to create n_i^* and the optimal utility U_i^* in subgame perfect equilibrium can be obtained by backward induction. First, we can work out the expected proportion of success of each player in the 'resist–takedown' stage based on Theorem 1 and the fact that a fraction of phishes will get away undetected as given by P_d. Then, returning to the 'create–detect' stage, the optimization problem of the phisher becomes:

$$\begin{aligned}\max_{n_w} E(U_w|n_w) &= \frac{1}{n}E(\sum_{j\in\mathbb{J}_d}\pi_{w,j}) + \frac{(1-P_d)n_w}{n} - cn_w \\ &= \frac{n_d}{n}E(\pi_w) + \frac{(1-P_d)n_w}{n} - cn_w\end{aligned}$$

where

$$E(\pi_w) = \begin{cases} \frac{R_w}{2R_s} & \text{if } \ 1 \geq \frac{R_w}{R_s} \geq \frac{2}{n_d} \\ \frac{2}{n_d} - \frac{2R_s}{(n_d)^2 R_w} & \text{if } \ \frac{2}{n_d} \geq \frac{R_w}{R_s} \geq \frac{1}{n_d - 1} \\ 0 & \text{if } \ \frac{1}{n_d} \geq \frac{R_w}{R_s} \end{cases}$$

$$n_d = P_d n_w + n_0$$
$$n = n_w + n_0$$

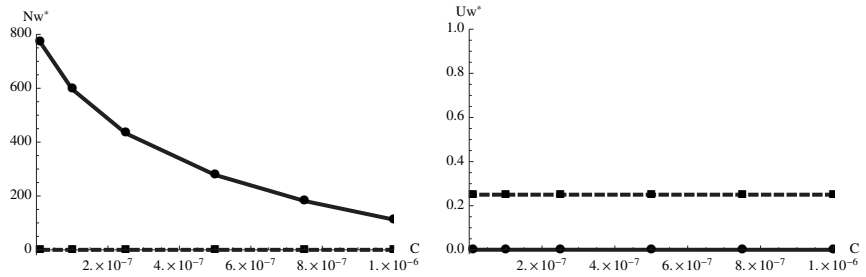
As with many real life security problems, the defender in this model is disadvantaged in that he takes only reactive measures against the phisher. Note that also we have omitted the case c of Theorem 1 (i.e., when $\frac{1}{n_d-1} > \frac{R_w}{R_s} > \frac{1}{n_d}$), a relatively small region with points of discontinuity, for simplicity.

A.4. Analysis

We analyze using the CBP game three different scenarios: (i) the hypothetical case of perfect detection of phishing attacks, i.e., $P_d = 1$, (ii) $P_d < 1$ and is exogenously determined, and (ii) $P_d < 1$ and is endogenously influenced by the number of phishes the attacker creates.

A.4.1. Perfect Phish Detection.

Let us start with the hypothetical case where the probability of detection, $P_d = 1$. Figure A.2 plots the optimal number of additional phishing attacks n_w^* that the phisher will launch depending on cost c, knowing that all newly created phishes will be detected by the defender. Note that this is exactly the scenario analyzed in [10], and that the dashed and solid lines plot the case a and b of Theorem 2 respectively. When the resource asymmetry is small (with $\frac{2}{n_0} \leq \frac{R_w}{R_s} = \frac{1}{2}$, dashed line), the phisher optimally chooses *not* to create additional phishes. There is



(a) n_w^* against c (b) U_w^* against c

Figure A.2.: The optimal new phishes n_w^* and utility U_w^* given $P_d = 1$. Solid and dashed lines plot the case where $\frac{R_w}{R_s} = \frac{1}{900}$ and $\frac{1}{2}$ respectively, with $n_0 = 1000$.

no advantage to further stretch the defender as the attacker, given his resources, is expected to win in equilibrium a proportion of battlefields equals $\frac{R_w}{2R_s} = \frac{1}{4}$ as shown in Figure A.2(b).

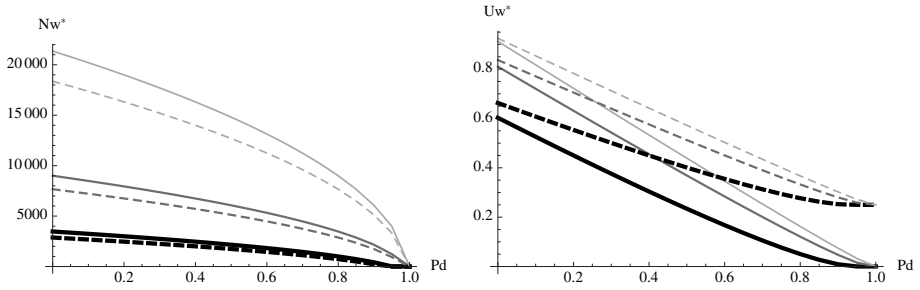
However, when the resource asymmetry is large (with $\frac{2}{n_0} > \frac{R_w}{R_s} = \frac{1}{900}$, solid line), the phisher will create additional phishing attacks to reduce the ability of the defender in locking down all of them. Especially when cost c (measured in terms of the normalized utility) is negligible, n_w^* approaches $\frac{2R_s}{R_w} - n_0 = 800$ given $\frac{R_w}{R_s} = \frac{1}{900}$ and $n_0 = 1000$. Even so, interestingly, the utility of the phisher is still less than 10^{-3}. Meanwhile, as c increases (see Figure A.2(a)), the optimal number of new phishing attacks n_w^* quickly approaches zero.

A.4.2. Imperfect Phish Detection (Exogenous).

In practice, we can expect that a significant fraction of phishing attacks will get away undetected by the defender. The problem is exacerbated by non-sharing of data between different security vendors as observed in [15]. Figure A.3(a) and A.3(b) plot the optimal number of new phishes n_w^* and the corresponding utility of the phisher U_w^* depending on $P_d \in [0, 1]$. We assume that the phisher will be able to estimate P_d based on past experience.

Let us first focus on the game between a resourceful (strong) phisher and the defender, with the resource asymmetry $\frac{2}{n_0} \leq \frac{R_w}{R_s} = \frac{1}{2}$ (as shown by the solid lines). Here, with $Pd < 1$, the phisher will now create additional phishes knowing that the defender will fail to detect some of the attacks, different from the case of perfect detection. The undetected phishes add on to the phisher's utility, which has a lower bound at $\frac{R_w}{2R_s} = \frac{1}{4}$. As for the game between a less resourceful (weak) phisher and the defender given a large resource asymmetry of $\frac{2}{n_0} > \frac{R_w}{R_s} = \frac{1}{900}$ (as depicted by the dashed lines), observe that the optimal numbers of new phishing attacks are now much higher than 800, the upper bound for the case of perfect detection.

Another interesting observation is that the utility gap between a strong and a weak phisher reduces as P_d decreases from 1 to 0. Improving on P_d thus will hurt



(a) n_w^* against P_d (b) U_w^* against P_d

Figure A.3.: Optimal number of new phishes to create n_w^* and the corresponding optimal utility U_w^*. Solid and dashed lines plot the case where $\frac{R_w}{R_s} = \frac{1}{900}$ and $\frac{1}{2}$ respectively, with $n_0 = 1000$. The effect of a decreasing cost c going from 5×10^{-5} to 1×10^{-5} and 2×10^{-6}, measured in terms of the normalized utility, is depicted by the thick-black, normal-black and thin-gray lines, respectively.

a weak phisher, but has less impact on a strong phisher as he can leverage on his resources (technologies, infrastructure, manpower, etc.) to resist the takedown of some of his phishes. The trend also suggests that an attacker will optimally vary his strategies to create more phishes when P_d is low, but strive to increase his resources as P_d increases.

Regardless of the extent of resource asymmetry, an increased cost (see the thick-dark lines versus the thin-gray lines) reduces both the optimal number of phishes and the utility of the phisher. But, somewhat counter-intuitively, the lower the detection probability, the more phishes the attacker will want to create. An attacker does not settle with having a fraction of undetected phishes, but will exploit the weakness of the defender in detecting all phishes and create even more phishes to increase his utility.

Another counter-intuitive and interesting finding is that in fact it is optimal for a less resourceful phisher to create more new phishes (than if he is a resourceful phisher) in equilibrium. This can be seen in Figure A.3 where the solid lines ($\frac{R_w}{R_s} = \frac{1}{900}$) remain above the dashed lines ($\frac{R_w}{R_s} = \frac{1}{2}$) for all different costs c. This is surprising as large-scale phishing attacks are more often associated with resourceful attackers such as the 'Rock Phish' and 'Avalanche' gangs empirically.[13] There could be several reasons to this. First, while the 'Avalanche' phishes can be recognized easily with their distinctive characteristics, we do not know if the bulk of other phishing attacks are not related (carried out by a single organization) for sure. Secondly, could there be really a 'tragedy of the commons' due to the a large number of phishers (as described in [9]) that has forced the less resourceful attackers out of the phishing endeavor? We note that analyzing the effect of competition

[13]For example, the 'Avalanche' gang was responsible for 84,250 out of 126,697 (66%) phishing attacks recorded by the APWG in the second half of 2009.

between several phishers would be an interesting extension to our current model. Another more likely explanation would be that most of the phishing attacks are in fact detectable by the defender today, forcing the less resourceful attacker to gain too little utility to be profitable (observe that U_w^* for the less resourceful attacker is almost zero as $P_d \to 1$ in Figure A.3(b)). Furthermore, having a large number of phishes can also increase the probability of detection by the defender. We analyze the case when P_d depends on the n_w in the next section.

A.4.3. Imperfect Phish Detection (Endogenous).

Let us model the effective P_d to depend on the number of phishes an attacker creates with a simple formulation:

$$P_d = P_{d0} \times (n_w)^{\alpha}$$

where with $\alpha = 0$, we thus have the exogenous case as discussed in the previous section. The interesting analysis here is when $\alpha \neq 0$ as depicted in Figure A.4.

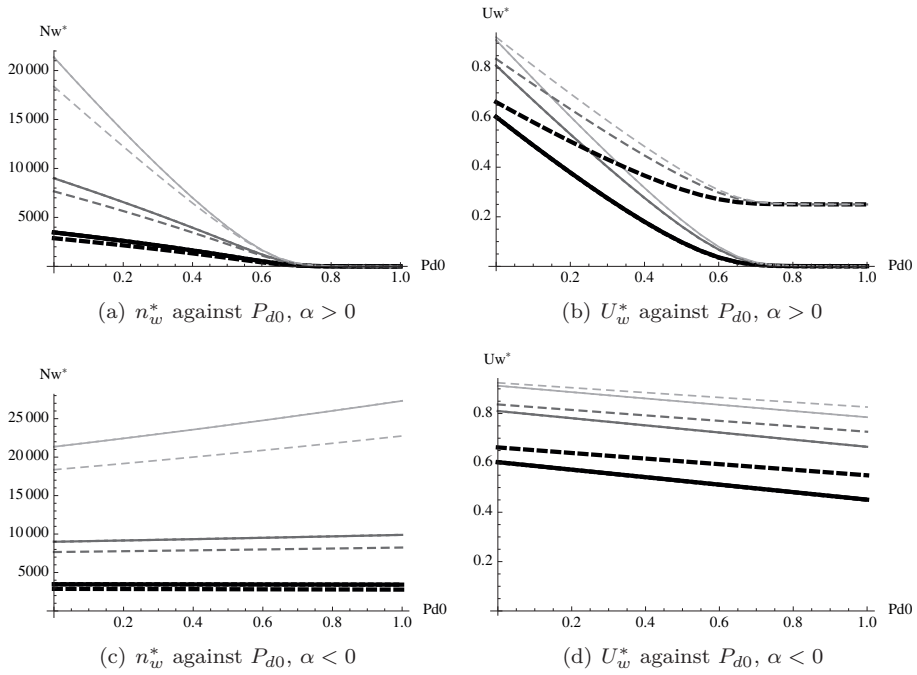
There are many examples where increasing the number of phishing attacks (battlefields) can lead to a higher detection rate by the defender. For instance, the way the 'Rock Phish' and 'Avalanche' gangs hosted a number of phishing attacks (i.e., different phishing pages targeting different brands) using the same domain name[14], while reducing cost, increases the chance that all these phishes (battlefields) will be detected and taken down altogether. An attacker who register multiple domains for phishing purposes may also risk leaving visible patterns in the WHOIS database that is being used by the defender to identify and suspend suspicious domains quickly.[15]

As shown in Figure A.4(a) and A.4(b), both the n_w^* and U_w^* curves are now steeper than before. The optimal number of additional phishing attacks to create quickly approaches zero as P_{d0} increases. Other than that, the main results from the case of exogenous detection probability (where $\alpha = 0$) remain applicable. First, it is optimal for a weak phisher to create more phishes than a resourceful attacker. The lower the detection probability is the more phishes will an attacker create. Also, improving the baseline detection technologies (P_{d0}) hurts a weaker phisher more than a stronger phisher.

It is harder to think of some practical examples where an increased number of phishes helps to reduce the effective detection rate by the defender (i.e., with $\alpha < 0$). A possible but *unlikely* scenario would be if the phishing attacks that a phisher creates cannot be correlated to each other, and that the larger number of attacks stretch the defender's capability in detecting all of them. We include the plots of optimal n_w^* and U_w^* under such scenario in Figure A.4(c) and A.4(d) for reference purposes. Notice that the optimal utility of the phisher is now bounded only by the cost of creating new phishes.

[14] A typical 'Avalanche' domain often hosted around 40 phishing attacks at a time [2].

[15] APWG reported that attackers often utilize a single or small set of unique names, addresses, phone numbers, or contact email addresses to control their portfolio of fraudulent domain names [1].



(a) n_w^* against P_{d0}, $\alpha > 0$

(b) U_w^* against P_{d0}, $\alpha > 0$

(c) n_w^* against P_{d0}, $\alpha < 0$

(d) U_w^* against P_{d0}, $\alpha < 0$

Figure A.4.: Optimal n_w^* and U_w^* when the effective probability of detection, $P_d = P_{d0} \times (n_w)^\alpha$. Graphs a and b plot the case where $\alpha = 0.05 > 0$, while graphs c and d plot the case of $\alpha = -0.2 < 0$. Solid and dashed lines plot the case where $\frac{R_w}{R_s} = \frac{1}{900}$ and $\frac{1}{2}$ respectively, with $n_0 = 1000$. The effect of a decreasing cost c going from 5×10^{-5} to 1×10^{-5} and 2×10^{-6} is depicted by the thick-black, normal-black and thin-gray lines, respectively.

A.5. Discussion: Implications to Anti-Phishing Strategies

The success of anti-phishing defense depends on a number of interacting variables. As captured in our model, increasing the cost of creating new phishes c, improving the detection rate of new phishes P_d, as well as, increasing the resource asymmetry between the defender and phisher, $\frac{R_s}{R_w}$ are all crucial factors to be considered.

Increasing the cost for creating new phishes will hurt the attacker especially a weak phisher, who has no resources to resist the prompt removal of his phishes. Raising the cost (both in financial and procedural terms) for registering a domain name can therefore help, but only to a certain extent. Take the decision by CNNIC to make the registration of domain names more restrictive for example, the number of `.cn` phishing domains dropped, but phishing attacks on Chinese institutions remained high as phishers shifted to use other domain names such as `.tk` and the co.cc subdomain service (see [3] page 5). Phishers would also usually register new domains using stolen credit cards. Furthermore, studies have found that a larger percentage of phishing attacks (80%) are actually performed using compromised web servers of innocent domain registrants (see e.g., [2, 3, 17]). To raise the cost c will thus involve patching a large number of vulnerable servers, which is challenging if not impossible without a proper incentive plan.

A more effective alternative is hence to focus on improving the detection rate of new phishes. While automated spam filters help to detect potential phishing URLs, the 'Rock Phish' gang, for example, used GIF image in phishing email to evade detection. The popularity of URL shortening services and wall postings on online social networks add up to the challenge of detecting all phishing advertisements. Calls to share the phishing data in the anti-phishing industry have been made before (e.g., in [15]), but sharing can also create concerns as takedown companies leverage on their phishing data for competitive edges. Here, we see a room to employ and better coordinate the crowds to help improving the detection probability. Collecting user reports against potential phishes (or potentially harmful sites), without necessarily demanding from them higher skilled tasks such as evaluating if a phish is valid (or that a site is secure), can already be helpful.

Naturally, the value of data sharing and crowd-based phish-reporting will depend on the state of information asymmetry (i.e., the detection probability P_d). As can be seen in Figure A.3, an 'intelligent' phisher will leverage on a large number of phishes for optimal utility when P_d is low. Meanwhile, as $P_d \to 1$, a phisher will improve his utility by increasing his resources to match the defender's. This includes, for example, to gain access to a botnet infrastructure so to prolong the uptime of his phishes. Should a good estimate of P_d is available, the defender can thus decide whether to prioritize on increasing the cost of creating new attacks (to reduce the number of phishes the attacker can create), or to prioritize on disrupting the channels a phisher can increase his resources (e.g., access to a botnet infrastructure, malicious tools, the underground market to monetize stolen credentials, or domain resellers with shady practices), accordingly.

A.6. Conclusions

We have proposed the Colonel Blotto Phishing (CBP) game to help better understanding the dynamics of the two-step detect-and-takedown defense against phishing attacks. We gained several interesting insights, including the counter-intuitive result that it is optimal for the less resourceful attacker to create even more phishing attacks than the resourceful counterpart in equilibrium, and that the attacker will optimally vary his strategies to either increase the number of phishes or to focus on raising his resources depending on the detection probability. We then discussed the implications to the anti-phishing industry.

Capturing the conflicts between an attacker and a defender with asymmetric resources and information, it is our hope that the CBP game can be eventually used to analyze other interesting problems, including measuring the effects of competition between multiple phishers, and the benefits of cooperation between multiple takedown companies. We also see the suitability of the CBP game to be applied to web security problems in general. Indeed, various web security problems, including malicious sites, illegal pharmacies, mule-recruitment and so fourth, are currently mitigated through a detect-and-takedown process similar to in the anti-phishing industry.

Future Work. Like other stylized models, the CBP game can be extended in several directions. A potential extension is to include the time dimension into the game, for example, using repeated games to model the uptime of a phish, which is often used to measure the damage caused by phishing activities. Using the variants of the classic Colonel Blotto game, such as the non-constant sum version [19] in which players might optimally choose not to expend all their resources, may also yield interesting results. We note that it may be interesting also to test our CBP model through experimental studies. Existing studies as conducted in [4, 5, 7, 12] have largely found that subjects were able to play the equilibrium strategies of the classic Colonel Blotto game, with the weak and strong players adopting the 'guerrilla warfare' and 'stochastic complete coverage' strategies respectively. Testing how the subjects will play our two-stage CBP game can be an interesting future work.

Acknowledgment

This research was supported in part by the National Science Foundation under award CCF-0424422 (TRUST).

References

[1] Anti-Phishing Working Group (APWG). Advisory on utilization of whois data for phishing site take down. `http://www.antiphishing.org/reports/apwg-ipc_Advisory_WhoisDataForPhishingSiteTakeDown200803.pdf`. Last accessed: June 2012.

[2] Anti-Phishing Working Group (APWG). Global phishing survey: Trends and domain name use in 2H2009. `http://www.antiphishing.org/reports/APWG_GlobalPhishingSurvey_2H2009.pdf`. Last accessed: June 2012.

[3] Anti-Phishing Working Group (APWG). Global phishing survey: Trends and domain name use in 2H2010. `http://www.antiphishing.org/reports/APWG_GlobalPhishingSurvey_2H2010.pdf`. Last accessed: June 2012.

[4] A. Arad and A. Rubinstein. Colonel blotto's top secret files. Levine's Working Paper Archive 814577000000000432, David K. Levine, Jan. 2010.

[5] J. Avrahami and Y. Kareev. Do the Weak Stand a Chance? Distribution of Resources in a Competitive Environment. *Cognitive Science*, pages 940–950, 2009.

[6] E. Borel. La théorie du jeu les équations intégrales à noyau symétrique. *Comptes Rendus de l'Académie des Sciences*, 173:1304–1308, 1921. English translation by Savage, L. (1953) The theory of play and integral equations with skew symmetric kernels, Econometrica 21:97–100.

[7] S. M. Chowdhury, D. K. J., and R. M. Sheremeta. An experimental investigation of colonel blotto games. CESifo Working Paper Series 2688, CESifo Group Munich, 2009.

[8] O. A. Gross and R. A. Wagner. A continuous colonel blotto game. *RAND Corporation RM–408*, 1950.

[9] C. Herley and D. Florêncio. A profitless endeavor: phishing as tragedy of the commons. In *Proceedings of the 2008 workshop on New security paradigms*, NSPW '08, pages 59–70, New York, NY, USA, 2008. ACM.

[10] D. Kovenock, M. J. Mauboussin, and B. Roberson. Asymmetric conflicts with endogenous dimensionality. *The Korean Economic Review*, 26:287–305, 2010.

[11] D. Kovenock and B. Roberson. Conflicts with multiple battlefields. Purdue University Economics Working Papers 1246, Purdue University, Department of Economics, Aug. 2010.

[12] D. Kovenock, B. Roberson, and R. M. Sheremeta. The attack and defense of weakest-link networks. Working Papers 10-14, Chapman University, Economic Science Institute, Sept. 2010.

[13] D. K. McGrath and M. Gupta. Behind phishing: An examination of phisher modi operandi. In F. Monrose, editor, *LEET*. USENIX Association, 2008.

[14] T. Moore and R. Clayton. Examining the impact of website take-down on phishing. In L. F. Cranor, editor, *eCrime Researchers Summit*, volume 269 of *ACM International Conference Proceeding Series*, pages 1–13. ACM, 2007.

[15] T. Moore and R. Clayton. The consequence of noncooperation in the fight against phishing. In *Proceedings of the 3rd APWG eCrime Researchers Summit*, eCrime '08, pages 1–14, 2008.

[16] T. Moore and R. Clayton. The impact of incentives on notice and takedown. In M. Johnson, editor, *Managing Information Risk and the Economics of Security*. Springer, 2008.

[17] T. Moore and R. Clayton. Evil searching: Compromise and recompromise of internet hosts for phishing. In R. Dingledine and P. Golle, editors, *Financial Cryptography*, volume 5628 of *Lecture Notes in Computer Science*, pages 256–272. Springer, 2009.

[18] B. Roberson. The colonel blotto game. *Economic Theory*, 29(1):1–24, Sept. 2006.

[19] B. Roberson and D. Kvasov. The non-constant-sum colonel blotto game. CESifo Working Paper Series 2378, CESifo Group Munich, 2008.

B. Re-Evaluating the Wisdom of Crowds in Assessing Web Security

Author

Pern Hui Chia, Q2S NTNU
Svein Johan Knapskog, Q2S NTNU

Conference

15th International Conference on Financial Cryptography and Data Security (FC)
28 February – 4 March 2011, Santa Lucia

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.

B.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 in 2009, with more than 51,000 phishing incidents recorded (up 16% from 2008) [21]. Provos et al. [18] found that over 3 million malicious websites initiate drive-by downloads and about 1.3% of all Google search queries get more than one malicious URL in the result page. Meanwhile, Zhuge et al. [23] 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 harm users in many aspects, such as scams, deceptive information gathering and misuse of user data. Several fraudulent 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 are taken down between 4 to 96 hours, the average life-time was found to be 2 weeks for mule-recruitment and 2 months for online pharmacy sites [16]. Problems with the adult sites may also be serious; while it is a personal judgment whether adult content in general is inappropriate, Wondracek et al. [22] 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 also sometimes a conflict of interest. Problems, such as adverse-selection [14] have been observed when certifiers adopt lax requirements to certify sites in the 'gray' category.

B.1.1. The wisdom of crowds for security

A typical argument against the idea of *the wisdom of crowds for security* is on the limited ability of ordinary users in providing reliable security evaluation. There is a general uneasiness 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 [20].

PhishTank[8] and Web of Trust (WOT)[10] 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, vendor reliability, privacy and child-safety aspects of domains. 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 cancel out each other. There is also the advantage of scale to cater for a large volume of items needing an evaluation.

In this work, we measure the reliability of WOT compared to 3 automated services by well known vendors, namely, McAfee's SiteAdvisor[5], Norton's Safe Web[7] and Google's Safe Browsing Diagnostic Page[3]. We also investigate the participation pattern 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.
- WOT's coverage 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 more often labeled as `unknown` by WOT, while they are often wrongly labeled as `good` by the other services.
- Contribution ratios in WOT are skewed with the comment contribution following a power law distribution.
- WOT has built 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.

B.2. Related Work

Surowiecki [20] 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. The crowds should be *decentralized* to tap into local knowledge and specialization, which should be collated wisely with *a good aggregation strategy*.

In [15], Moore and Clayton evaluated the reliability and contribution patterns in PhishTank. 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 also found that PhishTank was slightly less comprehensive and slower in reaching a decision. Our work is inspired by theirs, combined with the curiosity of why PhishTank has become widely adopted despite the initial criticisms.

While a number of studies look at the efficiency of various blacklists or tools for the issue of phishing (e.g., [19, 13]), 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.

B.3. The Web of Trust (WOT)

WOT is a reputation system that collates user inputs into global ratings about sites under evaluation. It takes the form of an open-source browser add-on and a centralized database [10]. User inputs and the evaluation outcomes are structured around 4 aspects 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 4 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 risks (e.g., identity theft, credit card fraud, phishing, viruses, adware or spyware).
- **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 sensitive information (e.g., whether it has opt-in privacy options or allows users 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 given to law enforcement without a warrant.
- **Child-Safety (Cs):** whether a site contains adult content, violence, vulgar or hateful language, or content that encourages dangerous or illegal activities.

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 select one out of 17 categories which best describes their comment. Comments do not count towards the aggregate ratings. Unlike the individual ratings, they are publicly accessible.

WOT has built a number of community features on its website, including a personal page per registered user, a scorecard for each evaluated site, messaging tools, a discussion forum, a wiki, and mechanisms to call for public evaluation on

specific domains. Meanwhile, the browser add-on allows a user to conveniently rate the sites he visits, in addition to signaling the reputation of different URI links, and warning the user as he navigates to sites that have been given a 'poor' rating. The child-safety rating is ignored by default, but the settings (for risk signaling and warning in general) are configurable to suit different users.

Besides user ratings and comments, WOT also receives inputs from trusted third parties. For example, it receives blacklists of phishes, spammers and illegal pharmacies from PhishTank[8], SpamCop[9] and LegitScript[4], respectively.

B.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 B.1.1. We collected the outcomes on 20,000 sites randomly selected from the top million frequently visited sites as published by Alexa[1]. 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, 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. The comments evaluate a total of 412,357 unique sites. To study the users' commenting behavior, we downloaded also the aggregate ratings of the 412k sites using the public query API of WOT. Both dataset-II and III contain only information that are publicly accessible for all who login to WOT's website.

B.5. Analysis

We started by studying the characteristics of the automated services:

- **McAfee's SiteAdvisor (SA)**[5] evaluates a site based on various proprietary and automated tests on aspects such as downloads, browser exploits, email, phishing, annoyance factors (e.g., pop-ups) and affiliations with other sites. SiteAdvisor also receives inputs from TrustedSource[6] which evaluates aspects such as site 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, but comments are not factored into the evaluation outcomes.

- **Norton's Safe Web (SW)**[7] tests if a site imposes threats e.g., drive-by download, phishing, spyware, Trojan, worm, virus, suspicious browser change, joke program and identity theft. It collects also user ratings and comments, but like SiteAdvisor, they do not count towards the overall rating.

- **Google's Safe Browsing Diagnostic Page (SBDP)**[3] 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 the frequently visited sites (used in our study) to be potential phishes.

To enable comparison, we mapped the evaluation outcomes of the respective services into 4 classes: `good`, `caution`, `bad` and `unknown`, as shown in Table B.1. We classified WOT's ratings based on its default risk signaling strategy, which regards Trustworthiness (Tr) as the most important evaluation aspect, given that it covers the scopes of Vendor Reliability (Vr) and Privacy (Pr). A site is considered `good` if its Tr rating is ≥ 60 without any credible warning (i.e., rating < 40 and confidence level ≥ 8) in Vr or Pr.[1] 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.

B.5.1. The reliability of WOT

We first evaluated the coverage of individual services (see Table B.2). Coverage measures the fraction of evaluated sites (i.e., those with an assessment outcome $\neq$ `unknown`). SiteAdvisor has the highest coverage while WOT scores the lowest. 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 WOT's statistics page. 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 B.2. SiteAdvisor classifies 1.48% sites as `bad`. This is interestingly close to the result in [23] 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 evaluated also the case which weighs all aspects equally (i.e., a site is classified as `bad` if *either* Tr, Vr or Pr is < 40), but found no significant changes in results.

Table B.1.: Aligning the evaluation outcomes of different services.

	WOT	SiteAdvisor	Safe Browsing DP	Safe Web
`Good`	Tr $\geq$ 60, and no credible warning in Vr or Pr	Green: very low or no risk	Site not currently listed as suspicious, and Google has visited it in the past 90 days.	Safe
`Caution`	60 $>$ Tr $\geq$ 40, and no credible warning in Vr or Pr	Yellow: minor risk	Site not currently listed as suspicious, but part of the 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	Site is listed as suspicious.	Warning
`Unknown`	No Tr rating, and no credible warning in Vr or Pr	Gray: not rated	Site not listed as suspicious, and Google has not visited it the past 90 days.	Untested

Table B.2.: Coverage and the percentage of assessment outcomes. The coverage of WOT is computed based on its default risk signaling strategy and thus not including the child-safety aspect (as shown in Table B.1). We regard a site that is not currently blacklisted and that has not been visited by Google's web bot in the past 90 days as 'not tested'.

	Cov. (%)	Outcomes (%)		
		`Bad`	`Caution`	`Good`
WOT	51.23	3.16	2.15	45.93
SA	87.84	1.48	0.47	85.90
SBDP	55.65	0.13	1.63	53.90
SW	68.09	0.51	0.38	67.21

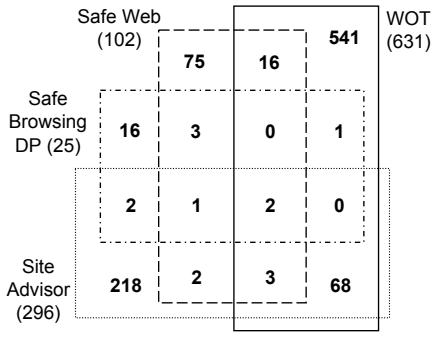


Figure B.1.: Venn diagram shows the divergence in the classification of 'bad' sites. Out of 948 that have been marked as `bad` by any services, only 2 receive the same verdict from all, while only 98 sites are classified as `bad` by >1 services.

The Venn diagram in Figure B.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 [17]. 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 [17]. 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.

We measured Recall ($R = T_p/(T_p + F_n)$), Precision ($P = T_p/(T_p + F_p)$) and F-Score ($FS = 2PR/(P + R)$) to quantify the reliability in identifying 'bad' sites. A challenge here is to determine the count of true positives (T_p), false positives (F_p) and false negatives (F_n) given that we do not know the 'correct' assessment outcomes or truth values. 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. 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 B.3 shows the definitions of T_p, F_p and F_n. Table B.4 presents the R, P and FS values of different services.

Table B.3.: True positives, false positives, false negatives in Optimistic or Conservative [in brackets] consensus. $F_{n,c}$, $F_{n,g}$, $F_{n,u}$ denote the false negative cases where a 'bad' site is wrongly labelled as `caution`, `good` and `unknown` respectively.

	Findings of other services									
Finding of this service	`Bad` w/o any `good`		Mixed of `good` & `bad`		`Caution` only		`Good` w/o any `bad`		All `unknown`	
`Bad`	T_p	$[T_p]$	$F_{p,m}$	$[T_p]$	$F_{p,c}$	$[F_{p,c}]$	$F_{p,g}$	$[F_{p,g}]$	$F_{p,u}$	$[F_{p,u}]$
`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 B.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 compare only among the automated services, excluding the assessment by WOT 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
SA	8.3	3.4	4.8	57.5	27.5	6.7	10.7	26.4	15.2	69.7	15.6	4.0
SW	4.1	8.8	5.6	59.0	34.2	2.7	3.1	26.5	5.5	71.6	23.6	1.7
SBDP	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 [cw]	13.9	2.5	4.3	-	-	-	17.2	17.8	17.5	-	-	-
SA [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
SW [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
SBDP [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

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 `consensus-bad` 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 B.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 may have to remain cautious even when a site has received 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 a warning dialog 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 with such a credible warning (i.e., those that will be explicitly warned against). As shown in the 5th row of Table B.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 evident. WOT has the highest FS of 17.3%. SiteAdvisor has a FS of 15.2% and interestingly, the FS 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 B.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 shows that WOT is reliable in comparison to the three 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 risks that are not currently evaluated by the others.

B.5.2. The few dominating contributors

Moore and Clayton argue that the highly skewed participation ratio in PhishTank increases the risks of manipulation; the corruption of a few highly active users can completely undermine its validity and availability [15]. It is also not difficult for a

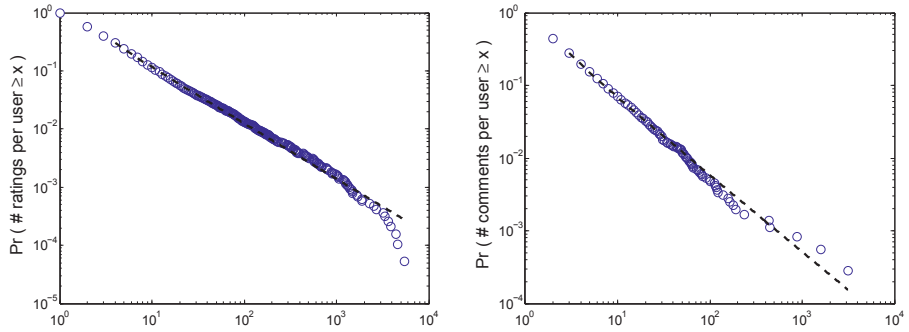


Figure B.2.: The complementary CDF of ratings and comments. Dashed lines depict the best fitted power law distribution with α=1.95, x_{min}=4 (rating, left) and α=2.05, x_{min}=3 (comment, right).

highly active user to disrupt the system under cover of a large body of innocuous behavior [15]. We investigated if similar problems exist in WOT.

We analyzed dataset-II which describes the contribution level of 50k randomly selected users. Of these users, the total rating and comments posted are 214,872 and 20,420 respectively. However, only 38.34% of them have rated and 7.16% have commented about a site at least once.

The seemingly straight lines in the log-log graphs (in Figure B.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 [12]. 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[2] (Figure B.2, bottom left). Adapting from the 80:20 rule of the Pareto principle, we measured the Skewness (S) such that S is the 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

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

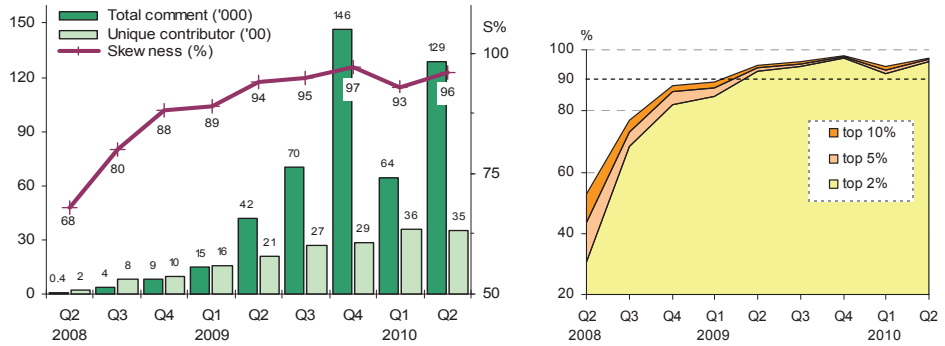


Figure B.3.: (Left) Total comment ('000), Unique contributor ('00) and Skewness (%). (Right) Percentage of comment provision by the top 2, 5 and 10% contributors.

rating contributors while the top 5% comment contributors gave 95% of the total comments. Both contribution ratios are skewed.

We then studied the evolution of user participation using dataset-III, which contains 485,478 comments contributed by 16,030 unique users. Figure B.3 (left) 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. Since 2009 Q2, more than 90% of the total comments are actually provided by the top 2% active users as shown in Figure B.3 (right). 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). As cases of misuse were detected, WOT began to award new privileges only to the Platinum users, from 28 Dec 2009. 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 as 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 than 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 automatic rating with reference to some public blacklists. We learned from the developers that there are automatic mechanisms in WOT which monitor for suspicious user behavior. Yet, to the root of the challenges, WOT should work towards diversifying the user contribution so that it does not become a centralized/homogenous system overwhelmed with the inputs of a few. The mass rating privilege should be handled with care.

B.5.3. Exploitability, disagreement and subjectivity

Grouping the comments according to their respective category, we observed that there are many more comments of negative type than positive (see Figure B.4). 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 ($<$40 in either Tr, Vr, Pr or Cs aspect), or when a comment of negative type is given to a site that has a good rating ($\geq$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 $>$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

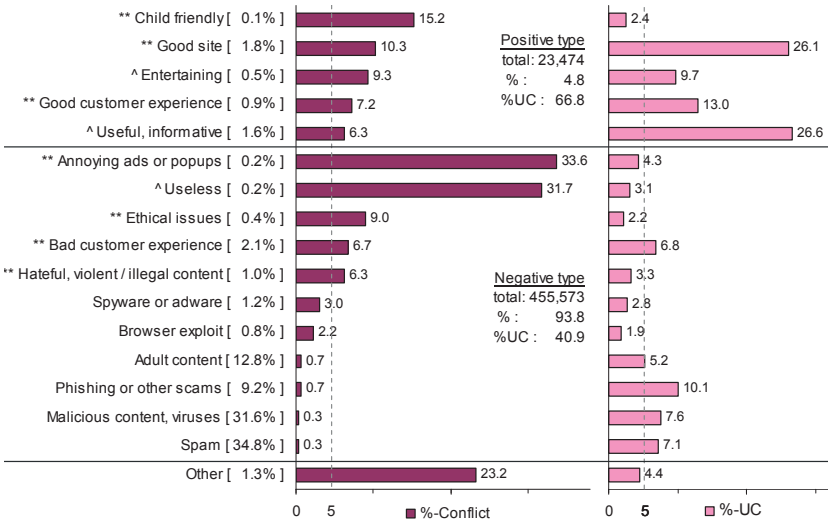


Figure B.4.: %-Conflict, % of Unique Contributor and %-count [in brackets] of different comment categories. ∧ denotes not in rating scope, ** denotes %-conflict >5.

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 [15], PhishTank is now adopted by multiple vendors including Yahoo!, Mozilla, Kaspersky Lab, Opera and McAfee [2].

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 more easily.

Table B.5.: Popular words used in user comments per year quarter. Similar words e.g., domain, website, page (∼site), scammer (∼scam), program (∼software) were omitted.

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

B.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 B.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.

This analysis indicates that user concerns on web security are not limited to only phishing and malware. This brings up the limitation of the automated services in catering for user concerns on online risks such as scams, illegal pharmacies, information protection and inappropriate content in general.

B.6. Discussion

The strengths of WOT lie 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 tools and the ability to call for public evaluation for specific sites, have all contributed to a reliable reviewing process.

Yet, WOT is not without several potential weaknesses. We discuss several of them and suggest the potential mitigating strategies in the following:

- **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, 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 given that warnings by WOT may not be frequently supported by the automated services (as characterized by the low precision value). Users unaware 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 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 [11] to deal with subjective factors. Inputs from personalized communities have the advantages of being more trustworthy, relevant and impactful than those provided by unknown community members [11].

There are several limitations to our study. First, as our evaluation sample consists of sites randomly chosen from the one million most-frequently visited sites, we have not evaluated the reliability of WOT when dealing with 'bad' sites that are more frequently found in the long-tail of web popularity. Further, we have also not tested the timeliness of WOT's assessment. 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.

B.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. 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 potential 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.

B.8. 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 WOT.

References

[1] Alexa Top Million Sites. `http://www.alexa.com/topsites`.

[2] Friends of PhishTank: Vendors using data submitted to and verified by PhishTank. `http://www.phishtank.com/friends.php`. Last accessed: June 2012.

[3] Google Safe Browsing Diagnostic Page. `http://www.google.com/safebrowsing/diagnostic?site=<site>`.

[4] LegitScript. `http://www.legitscript.com`.

[5] McAfee SiteAdvisor. `http://www.siteadvisor.com`.

[6] McAfee TrustedSource. `http://www.trustedsource.org`.

[7] Norton Safe Web. `http://safeweb.norton.com`.

[8] PhishTank. `http://www.phishtank.com`.

[9] SpamCop. `http://www.spamcop.net`.

[10] Web of Trust. `http://www.mywot.com`.

[11] P. H. Chia, A. P. Heiner, and N. Asokan. Use of ratings from personalized communities for trustworthy application installation. In T. Aura, K. Järvinen, and K. Nyberg, editors, *NordSec*, volume 7127 of *Lecture Notes in Computer Science*, pages 71–88. Springer, 2010.

[12] A. Clauset, C. Shalizi, and M. Newman. Power-law distributions in empirical data. *SIAM Review*, 51(4):661–703, 2009.

[13] L. F. Cranor, S. Egelman, J. I. Hong, and Y. Zhang. Phinding phish: An evaluation of anti-phishing toolbars. In *NDSS*. The Internet Society, 2007.

[14] B. Edelman. Adverse selection in online "trust" certifications and search results. *Electronic Commerce Research and Applications*, 10(1):17–25, 2011.

[15] T. Moore and R. Clayton. Evaluating the wisdom of crowds in assessing phishing websites. In G. Tsudik, editor, *Financial Cryptography*, volume 5143 of *Lecture Notes in Computer Science*, pages 16–30. Springer, 2008.

[16] T. Moore and R. Clayton. The impact of incentives on notice and takedown. In M. Johnson, editor, *Managing Information Risk and the Economics of Security*. Springer, 2008.

[17] T. Moore, R. Clayton, and R. Anderson. The economics of online crime. *Journal of Economic Perspectives*, 23(3):3–20, 2009.

[18] N. Provos, P. Mavrommatis, M. A. Rajab, and F. Monrose. All Your iFRAMEs Point to Us. In P. C. van Oorschot, editor, *USENIX Security Symposium*, pages 1–16. USENIX Association, 2008.

[19] S. Sheng, B. Wardman, G. Warner, L. Cranor, J. Hong, and C. Zhang. An empirical analysis of phishing blacklists. In *Proceedings of the 6th Annual Conference on Email and Anti-Spam*, CEAS '09, 2009.

[20] J. Surowiecki. *The wisdom of crowds*. Anchor Books, 2005.

[21] The UK Card Association. New Card and Banking Fraud Figures, March 10, 2010. `http://www.theukcardsassociation.org.uk/media_centre/press_releases_new/-/page/922/`. Last accessed: June 2012.

[22] G. Wondracek, T. Holz, C. Platzer, E. Kirda, and C. Kruegel. Is the Internet for Porn? An Insight into the Online Adult Industry. In *Proceedings of the 9th Workshop on the Economics of Information Security*, WEIS '10, 2010.

[23] J. Zhuge, T. Holz, C. Song, J. Guo, X. Han, and W. Zou. Studying Malicious Websites and the Underground Economy on the Chinese Web. In *Proceedings of the 7th Workshop on the Economics of Information Security*, WEIS '08, 2008.

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

Author

Pern Hui Chia, Q2S NTNU
John Chuang, UC Berkeley

Conference

15th ACM Conference on Computer Supported Cooperative Work (CSCW)
11–15 February 2012, Bellevue, Washington, USA

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.

C.1. 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. [22] 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 [8].

Detection, blacklisting, and takedown of malicious sites have been traditionally handled by security vendors such as anti-virus 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 [2] and Norton's Safe Web [3], 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 [15].

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 engages 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) [6] and PhishTank [4] 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 [7]. A recent study by Chia and Knapskog [11] 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.

C.2. Related Work

C.2.1. Collective Wisdom in General

A large number of prior works on collective wisdom have focused on the participation patterns in Wikipedia (e.g., [9, 17, 21, 23]), its potential pitfalls and risks (e.g., [14]) as well as its success factors and how to improve it (e.g., [12, 13, 16, 18]). Our work is related to that of Kittur et al. [17] 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. [17] has observed a shift of workload from the elite class contributors to the less active ones over time, Ortega et al. [21] 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., [9]) 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. [19] 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.

C.2.2. Collective Wisdom for Web Security

Denning et al. [14] 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 [20] 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 [11] 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 [10] 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.

C.3. 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.

C.3.1. 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%).

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 C.1.: Comment categories of positive or negative nature in WOT.

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).

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 C.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.

C.3.2. 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.

C.3.3. 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.

C.3.4. 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.

C.3.5. 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 [8] and [5]), giving WOT an overall coverage of 15.6%. As found in [11], 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 [22]. 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 [11].

C.4. 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 C.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.

C.4.1. 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 assumption 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 [11] – 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.

C.5. 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.

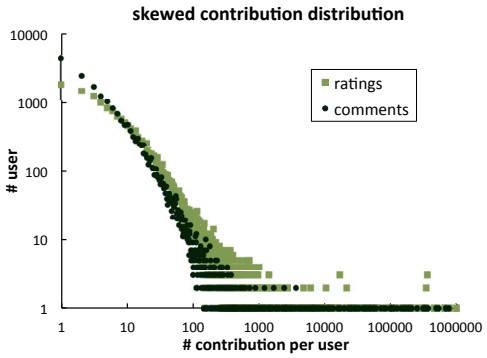


Figure C.1.: Distribution of comment and rating contribution.

C.5.1. Characterizing Different Types of Contributors

Figure C.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 [11] 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.

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 C.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 [19, 23], there may be ways to refine the contributor categorization using vari-

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

ous 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.

C.5.2. Coverage: Complementary Attention and Concern

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

C.5.2.1. Attention Divide on Goodness and Badness of Sites

Figure C.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 C.1. We made use of the positive or negative nature of a comment category to determine the contributor's attention 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.

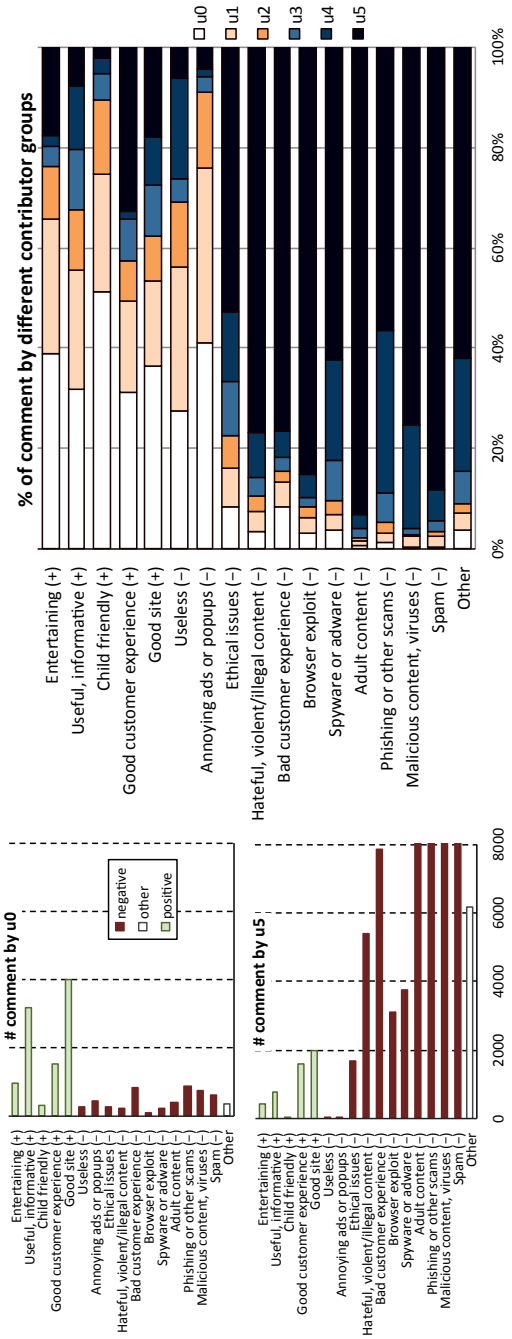


Figure C.2.: (Left) Total comments given by u0 and u5 in each comment category with the horizontal bars truncated at value=8000. (Right) Percentage breakdown of comments given by different contributor groups. + and – denote the positive and negative nature of a comment category respectively.

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

C.5.2.2. 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 C.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 unrated or unknown websites. Evaluations for the more frequently visited sites are also of a higher rele-

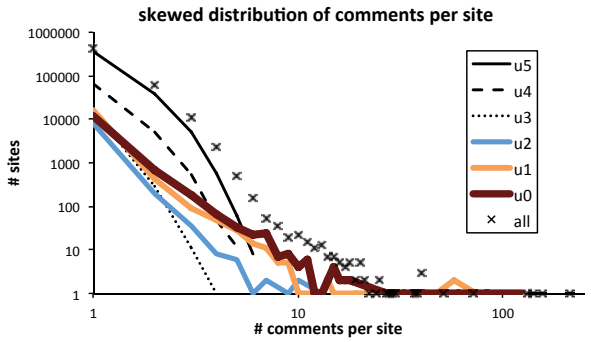


Figure C.3.: Distribution of comments per site considering all contributors (× markers) and each contributor group separately (black/color lines).

vance 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.

C.5.3. Coordination: Redundancy versus Efficiency

Figure C.3 plots the distribution of comments per site considering all contributors together (depicted by the × 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.

Cntr. grp.	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 C.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 C.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.

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.

C.5.4. 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 be disagreeing with each other [11]. 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.

C.5.4.1. Reliability in Evaluating Good and Bad Sites

Table C.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 C.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 C.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 C.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.

C.5.4.2. Verifiability: Reference and Comment Length

Table C.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 C.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.

C.6. Discussion

C.6.1. 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.

C.6.2. 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.

C.6.3. Design Implications

C.6.3.1. 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.

C.6.3.2. 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.

C.6.3.3. 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 [13]. 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 [18]). On the other hand, the heavy tailed attention distribution among the less active contributors (as shown in Figure C.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 [12]. All in all, coordination and socialization efforts should be done with a good understanding on the different roles and potentials of different contributors.

C.7. 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 eval-

uate 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.

C.8. 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] McAfee SiteAdvisor. `http://www.siteadvisor.com`.

[3] Norton Safe Web. `http://safeweb.norton.com`.

[4] PhishTank. `http://www.phishtank.com`.

[5] Verisign Domain Name Report. `http://www.verisigninc.com/en_US/why-verisign/research-trends/domain-name-industry-brief/index.xhtml`. Last accessed: June 2012.

[6] Web of Trust. `http://www.mywot.com`.

[7] Facebook partners with WOT. Article on ArcticStartup website, May 2011. `http://www.arcticstartup.com/2011/05/12/facebook-partners-with-wot-to-protect-its-700-million-users`. Last accessed: June 2012.

[8] Anti-Phishing Working Group (APWG). Global phishing survey: Trends and domain name use in 2H2010. `http://www.antiphishing.org/reports/APWG_GlobalPhishingSurvey_2H2010.pdf`. Last accessed: June 2012.

[9] J. Antin and C. Cheshire. Readers are not free-riders: reading as a form of participation on wikipedia. In K. I. Quinn, C. Gutwin, and J. C. Tang, editors, *CSCW*, pages 127–130. ACM, 2010.

[10] P. Ayyavu and C. Jensen. Integrating user feedback with heuristic security and privacy management systems. In D. S. Tan, S. Amershi, B. Begole, W. A. Kellogg, and M. Tungare, editors, *CHI*, pages 2305–2314. ACM, 2011.

[11] P. H. Chia and S. J. Knapskog. Re-evaluating the wisdom of crowds in assessing web security. In G. Danezis, editor, *Financial Cryptography*, volume 7035 of *Lecture Notes in Computer Science*, pages 299–314. Springer, 2011.

[12] B. Choi, K. Alexander, R. E. Kraut, and J. M. Levine. Socialization tactics in wikipedia and their effects. In K. I. Quinn, C. Gutwin, and J. C. Tang, editors, *CSCW*, pages 107–116. ACM, 2010.

[13] D. Cosley, D. Frankowski, L. G. Terveen, and J. Riedl. Suggestbot: using intelligent task routing to help people find work in wikipedia. In D. N. Chin, M. X. Zhou, T. A. Lau, and A. R. Puerta, editors, *IUI*. ACM, 2007.

[14] P. J. Denning, J. Horning, D. L. Parnas, and L. Weinstein. Wikipedia risks. *Commun. ACM*, 48(12):152, 2005.

[15] B. Edelman. Adverse selection in online "trust" certifications and search results. *Electronic Commerce Research and Applications*, 10(1):17–25, 2011.

[16] R. S. Geiger and D. Ribes. The work of sustaining order in wikipedia: the banning of a vandal. In K. I. Quinn, C. Gutwin, and J. C. Tang, editors, *CSCW*, pages 117–126. ACM, 2010.

[17] A. Kittur, E. Chi, B. A. Pendleton, B. Suh, and T. Mytkowicz. Power of the few vs. wisdom of the crowd: Wikipedia and the rise of the bourgeoisie. *Algorithmica*, 1(2):1–9, 2007.

[18] A. Kittur and R. E. Kraut. Harnessing the wisdom of crowds in wikipedia: quality through coordination. In B. Begole and D. W. McDonald, editors, *CSCW*, pages 37–46. ACM, 2008.

[19] L. Mamykina, B. Manoim, M. Mittal, G. Hripcsak, and B. Hartmann. Design lessons from the fastest Q&A site in the west. In D. S. Tan, S. Amershi, B. Begole, W. A. Kellogg, and M. Tungare, editors, *CHI*, pages 2857–2866. ACM, 2011.

[20] T. Moore and R. Clayton. Evaluating the wisdom of crowds in assessing phishing websites. In G. Tsudik, editor, *Financial Cryptography*, volume 5143 of *Lecture Notes in Computer Science*, pages 16–30. Springer, 2008.

[21] F. Ortega, J. M. González-Barahona, and G. Robles. On the inequality of contributions to wikipedia. In *HICSS*, page 304. IEEE Computer Society, 2008.

[22] N. Provos, P. Mavrommatis, M. A. Rajab, and F. Monrose. All Your iFRAMEs Point to Us. In P. C. van Oorschot, editor, *USENIX Security Symposium*, pages 1–16. USENIX Association, 2008.

[23] H. T. Welser, D. Cosley, G. Kossinets, A. Lin, F. Dokshin, G. Gay, and M. A. Smith. Finding social roles in wikipedia. In *iConference*, pages 122–129. ACM, 2011.

[24] D. M. Wilkinson. Strong regularities in online peer production. In L. Fortnow, J. Riedl, and T. Sandholm, editors, *ACM Conference on Electronic Commerce*, pages 302–309. ACM, 2008.

[25] G. Wondracek, T. Holz, C. Platzer, E. Kirda, and C. Kruegel. Is the Internet for Porn? An Insight into the Online Adult Industry. In *Proceedings of the 9th Workshop on the Economics of Information Security*, WEIS '10, 2010.

[26] J. Zhuge, T. Holz, C. Song, J. Guo, X. Han, and W. Zou. Studying Malicious Websites and the Underground Economy on the Chinese Web. In *Proceedings of the 7th Workshop on the Economics of Information Security*, WEIS '08, 2008.

D. Analyzing the Incentives in Community-based Security Systems

Author

Pern Hui Chia, Q2S NTNU

Conference

3rd IEEE International Workshop on SEcurity and SOCial Networking (SESOC)
held in conjunction with
9th IEEE Intl. Conf. on Pervasive Computing and Communications (PerCom)
21–25 March 2011, Seattle, Washington, USA

Abstract

Apart from mechanisms to make crowd-sourcing secure, the reliability of a collaborative system is dependent on the economic incentives of its potential contributors. We study several factors related to the incentives in a community-based security system, including the expectation on the social influence and the contagion effect of generosity. We also investigate the effects of organizing community members differently in a complete, random and scale-free structure. Our simulation results show that, without considering any specific incentive schemes, it is not easy to encourage user contribution in a complete-graph community structure (global systems). On the other hand, a moderate level of cooperative behavior can be cultivated when the community members are organized in the random or scale-free structure (social networks).

D.1. Introduction

Despite the popularity of reputation and recommender systems, relying on community effort for security purposes is still a new concept to many. PhishTank [1] and Web Of Trust (WOT) [2] are two of the few systems that employ user inputs to improve web security. PhishTank relies on user reporting and voting against suspected phishes, while WOT collates user ratings on several trust and security aspects of websites. Researchers have looked at the reliability of Community-based Security Systems (CSS). Moore and Clayton [18] argued that as participation in PhishTank follows a power-law distribution, its outputs are particularly susceptible to manipulation by the few highly active contributors.

Indeed, besides system design to prevent manipulation by dishonest users or attackers, the reliability of a CSS is highly dependent on the incentives of its potential contributors. When many users contribute actively, diverse user inputs serve to cancel out the errors made by individuals and scrutinize against manipulative attempts. On the other hand, when many users do not contribute, the outcomes can be biased towards the judgment of a few, or be manipulated easily. In the reverse manner of *the tragedy of the commons* [14] which depicts the situation whereby individuals consume the common resource irresponsibly, motivating active participation in a CSS is a problem of *public goods provisioning* such that individuals face a cost that discourages them from contributing to the common protection.

In this work, we look at several factors related to the incentives in a CSS. We adapt the *normalized total-effort security game* in [12, 13] to depict the scenario of collaborative security protection, but our model takes into account of the long term user consideration using the framework of *infinitely repeated games*. We first describe the basic model in Section D.2 and extend it with the expectation on social influence in Section D.3. We study the effects of user dynamics and a possible contagion effect of generosity in Section D.4. We find that it is easier to encourage a moderate level of user contribution in a random or scale-free community structure than in the complete-graph structure of global systems in Section D.5.

D.2. Basic Model & Analysis

Imagine a community-based system that collates evaluation reports from its members on some trust or security aspects of websites. Inputs from all members are important as they serve to diversify the inputs and cancel out errors made by individuals, in addition to enabling a level of scrutiny on each other's inputs. Without a sufficient level of contribution, a CSS will be deemed unreliable and abandoned by its members. An equally undesired scenario arises when there is a highly skewed participation ratio such that the system can be completely undermined when the few highly active users become corrupted or stop participating [18].

D.2.1. An Infinitely Repeated Total-effort Security Game

We use a n-player repeated game to model a CSS consisting of N rational members. We first consider a complete-graph community structure, as in global systems like PhishTank and WOT, where the $N = n$ members collaborates at all time (round) t for common protection. Each game round evaluates a different website (target).

We assume that all members value the benefit of protection b equally and have the same cost of contribution c. Assuming also that the inputs from all members are equally important, we formulate the homogenous utility $U_{i,t}$ received by member i at game round t to be linearly dependent on the ratio of contribution by the n collaborating members, following the notion of normalized total-effort security in [12, 13], as follows:

$$U_{i,t} = \frac{\sum_j a_{j,t}}{n} b - c a_{i,t} \tag{D.1}$$

with $a_{i,t}$ denoting the binary action of either {*1: contribute, 0: do not contribute*} by member i at game round t.

When all members contribute, each of them receives a utility of $b - c$. If no one but only member i contributes, his utility is $b/n - c$. Assume $b > c > b/n$ such that contributing to a CSS is the case of n-person prisoner's dilemma. We further assume an infinitely repeated game to depict the expectation by individual members that the system will evaluate infinitely many websites and exist until an unforeseen future. If the system is known to last only for a finite amount of time, not-contributing at all game rounds will be the (sub-game perfect) equilibrium strategy.

We consider that individual members rank their infinite payoff stream using the δ-discounted average criterion. The discount factor δ_i characterizes how a player weighs his payoff in the current round compared to future payoffs. In the context of this paper, it can be interpreted as how a member perceives the long term importance of common protection and his relationship with the $n - 1$ interacting peers. We assume that δ_i is heterogeneous. A short-sighted member has $\delta_i \approx 0$, while a player who values the long-term benefit of the CSS system or who cares about the long-term relationship with other members, has $\delta_i \approx 1$. Let $0 \leq \delta_i < 1$, the δ-discounted average payoff of member i is given by:

$$(1 - \delta_i) \sum_{t=1}^{\infty} (\delta_i)^{t-1} U_t \tag{D.2}$$

Analyzing the equilibrium behaviors of the $n > 2$ community members is more complicated that a 2-player repeated game. A trivial setting is when all members are assumed to employ a n-player 'grim trigger' strategy which considers each member to be contributing initially, but threatens to stop contributing forever if he realizes that *any* of his n peers has not contributed in the previous round. Given

this largest threat of punishment, an equilibrium whereby all members will always contribute can be achieved, if $\forall i$:

$$\delta_i > \frac{cn - b}{bn - b} \tag{D.3}$$

A simple relationship between the cost c and benefit b can now be established. If c is close to b, the required δ_i approaches 1 as n increases. This reflects the real-life scenario that user inputs are hard to obtain if the contribution cost is large relative to the benefit of protection.

A challenge in a repeated n-player game is that one cannot identify and punish those who have not contributed without a centralized monitoring mechanism, which can be expensive to build or threaten the anonymity of contributors. A player can only work out the contribution ratio of the others in the previous round r_{t-1} based on the payoff he receives. The n-player 'grim trigger' strategy inefficiently punishes everyone, making it to be an unrealistic strategy.

D.3. The Expectation on Social Influence

Rather than using the 'grim trigger' strategy, we adapt an idea from [11] such that the community members reason for their respective choice of action, not only depending on the past but also on their expectation as to how their actions can influence the choice of others in the subsequent rounds. We construct two simple influence rules, as follows:

Linear influence. A player believes that his contribution will have a positive influence on his peers and increase their contribution ratio by γ in the subsequent game rounds. Similarly, he expects that their contribution ratio will drop by $-\gamma$ if he does not contribute. The rule can be written as:

$$\hat{r}_\gamma = min(max(r_{t-1} + \gamma, 0), 1) \tag{D.4}$$

Sigmoid influence. Same as linear influence. However, the contribution ratio in the subsequent rounds is updated following a sigmoid curve. Specifically, a player reasons that his action will have a reduced influence on others when the current contribution level is close to the extremes, 0 or 1:

$$\hat{r}_\gamma = \frac{1}{1 + e^{w(r_{t-1} + \gamma - \frac{1}{2})}} \tag{D.5}$$

With the above expectation, a member i will contribute at time t, if:

$$V_c(r_{t-1}) - V_n(r_{t-1}) + \frac{\delta_i}{1 - \delta_i} \left[V_c(\hat{r}_\gamma) - V_n(\hat{r}_{-\gamma}) \right] \; > \; 0 \tag{D.6}$$

with V_c and V_n denoting the utilities of contributing and not-contributing respectively, based on the last observed contribution level r_{t-1} and the expected future

contribution ratio $\hat{r}$. This assumes that a member believes that if he is to contribute now, since his action will cause a positive influence on the others, he will be better off to contribute also in the future. The same reasoning applies for the case of not-contributing. These simple rules (D.4), (D.5) and (D.6) model the bounded rationality of users. Indeed, it is non-trivial to reason for the best-responses in an interconnected structure. Each member is not aware of the discounting factor of others. One also may not know about his peers' peers and how they perform in their respective games.

D.3.1. Simulation Results

We consider several levels of expectation on how much an action can influence of the action of others, as shown in Table D.1. Note that $\pm\gamma$ depicts the expectation that contributing/not has a positive/negative influence on peers' contribution ratio. The appendix s denotes the use of sigmoid influence rule. Each expectation level is simulated 50 times for computing the mean payoff. In each simulation run, every member is assigned a new discounting factor drawn uniformly between a minimum value δ_{min} and 1.

Figure D.1 shows the simulation results. With $\gamma = -1.0$, which is the equivalent of the 'grim trigger' strategy, full contribution to give the maximum payoff of $b-c = 1$ is a stable outcome when all members have a discounting factor higher than a moderate threshold ≈ 0.5. This is shown by the dotdashed line in Figure D.1. However, as aforementioned, the 'grim-trigger' strategy is not realistic in practice.

With $\gamma = -0.5$, that is when the members expect that non-contribution will influence half of his peers to also stop contributing, a fully cooperative equilibrium can only be achieved if all members place a large weight on the benefit of long term protection, as shown by the dotted line in Figure D.1. As the community members expect that their respective action will have reduced influence on others' motivation (e.g., $\gamma = \pm 0.33s, \pm 0.25$), we find that the average payoff $\overline{U}$ remains at zero even as δ_{min} approaches 1. In other words, no community members contribute in equilibrium.

The above has several implications. First, the results show that, without the help of any incentive schemes, the level of user contribution for a CSS can be expected to be very low. Centralized mechanisms such as monitoring, reputation and micropayment may help to encourage contribution, but there can be challenges (including cost and anonymity concerns) in implementing them in practice. The results also highlight the role of education to cultivate a sense of social responsibility (i.e., to increase the users' perception of γ) and to inform about the long-term importance and benefit of collaborative security (i.e., to increase the δ_i of users). This is also challenging as it is well-known that ordinary users do not regard security as their primary concern [8].

Table D.1.: Simulating Users' Expectation on Social Influence

Var.	Description	Value(s)
N	Total community members	100
c	Contribution cost	1
b	Benefit of full protection	2
w	Steepness of the sigmoid function	10
γ	Expected social influence on the contribution ratio of peers	$\pm 0.25, \pm 0.33s$, $-0.50, -1.0$

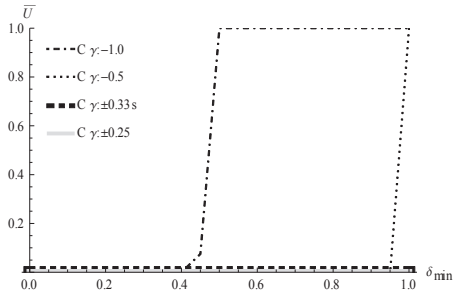


Figure D.1.: Mean payoff in a Complete-graph (C) community structure

D.4. The Effects of User Dynamics & Generosity

We investigate if a cooperative spirit can be cultivated given the presence of a small fraction of members $\theta \geq 0$ who would contribute to the common protection unconditionally. These 'nice' users can be thought as those who are extremely generous in real life, or those who have been employed to ensure a minimal level of contribution in the system.

We also factor in a simple user dynamics in the CSS. In each game round, a fraction $m \geq 0$ of under-performing users (i.e. those with average payoffs ≤ 0 and who have been through a minimum number of transient rounds τ) are programmed to leave. When a user leaves, we assume that another user joins and connects with the remaining members. This models the real life scenario where frustrated users leave, while new users join the community continuously.

D.4.1. Simulation Results

Table D.2 summarizes the variables and values used in our simulation to study the effect of community dynamics and 'nice' users. As before, each scenario is repeated with 50 simulation runs for computing the mean value. Without user dynamics, as in Section D.2, we observe that the contribution ratio and average payoff settle quickly after several game rounds. With user dynamics, these values fluctuate, but only slightly as one member may leave (while another joins) per game

Table D.2.: Simulating Community Dynamics and Nice Users

Var.	Description	Value(s)
θ	Fraction of nice users per game round	0.04, 0.08
m	Fraction of user leaving and joining per round	0.01
τ	Transient rounds before user dynamics	5

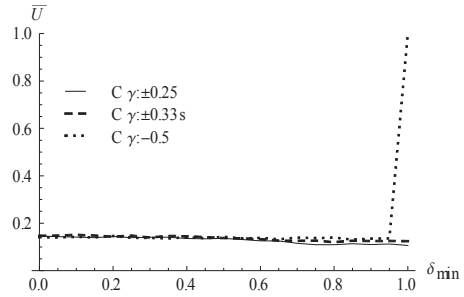


Figure D.2.: Mean payoff in a Complete-graph (C) community structure, with a fraction of 'nice' users $\theta = 0.04$ and dynamics of $m = 0.01$.

round. Considering this, in each simulation run, we measure the average payoff of all community members only after $t = 250$ game rounds.

Figure D.2 plots the average payoff of the community members when a small fraction of 'nice' users ($\theta = 0.04$) and dynamics ($m = 0.01$) are considered. In every game round, the worst performing member is programmed to leave, while 4 members are randomly selected to contribute unconditionally. As shown in the figure, the presence of 4 'nice' members does help to increase the contribution level (average payoff) slightly from zero as seen in Figure D.1, to about 0.15 in Figure D.2. This shows the role played by generous members or those who have been employed, in encouraging the others to contribute.

However, notice that other than the case with $\gamma = -0.5$, the average payoff is flat even as the minimum discounting factor δ_{min} approaches 1. This hints on the limited impact of the 'nice' users in a global system (complete-graph community structure) after all. It also highlights the risk of over-reliance on a small group of 'nice' users, causing a highly disproportionate contribution ratio in the system. A highly skewed contribution pattern can harm the Byzantine fault tolerance that community-based systems sought for, making it susceptible to manipulation by the few highly active users [18].

D.5. The Effects of Community Structure

Thus far, we have studied only the *complete-graph* community structure, used in global systems such as PhishTank and WOT. We consider two other structures, mimicking the topology of social networks, in the following.

Random (R). First, we consider a random network which connects any two members with a probability p. To ensure that all members will engage in a multi-player game, we require each user to have a degree $k \geq \psi$ initially. To build a random network, we first instantiate N members in the system. Then, for each member, we randomly assign ψ peers selected from the $N-1$ candidates with a uniform probability. In every round of the repeated game, a member will play an n-person game with his k peers, where $n = k + 1$. Note that a peer to this member will, on the other hand, play a n'-person game, where $n' = k' + 1$, with her own peers, including this member.

Scale-free (S). We also study a scale-free community structure. Many real-life networks such as protein interactions, interlinked web pages and citation patterns, have been shown to exhibit the scale-free property [6, 5] where the probability of a vertex having a degree k follows a power law distribution: $P(k) \sim k^{-\lambda}$. Two important processes for a scale-free network are growth and preferential selection. Same as the random network, we require also each member to have at least a degree $k \geq \psi$ initially. We build up a scale-free network by first initializing $\psi+1$ users that are fully connected. Then, for every subsequent member that newly arrives in the community, we connect him to ψ peers, with each candidate j being selected with a probability equals the candidate's degree over the sum of degree of all existing members, $P(select_j) = k_j/\Sigma k$. Like in the case of random network, in every game round, each member i plays a n_i-person game with their respective peers, where $n_i = k_i + 1$.

Note that all members remain interconnected in the scale-free and random networks. Both structures can be used to model a 'personalized' system whereby individual members interact only with peers in their personalized community. An example of a 'personalized' system for security is NetTrust [7] which advocates the use of ratings from personalized sources to provide reliable risk signals against suspicious websites to individual users.

D.5.1. Simulation Results

As before, each simulation scenario is repeated with 50 runs to compute the mean value. In each simulation run, the community structures are reconstructed and we measure the average payoff of all members only after $t = 250$ game rounds. Simulation values in Table D.1 and D.2 are used. In every game round, the 4 most connected members (i.e., who have the highest degree k) are selected to be 'nice'. User dynamics is also considered as before. The worst performing user is programmed to leave while a new user joins the community and establishes links with other members, following the characteristics of random or scale-free network.

The upper left and right of Figure D.3 plot the average payoff of players in the Random (R) and Scale-free (S) networks respectively. We see that the average payoff in the Random (R) and Scale-free (S) networks increases to about 0.6 and 0.4 when players believe that their actions can have a positive/negative influence of ± 0.25 and ± 0.33 respectively. This is a significant improvement compared to the case in a Complete-graph (C) network (as shown in Figure D.2). Being 'nice'

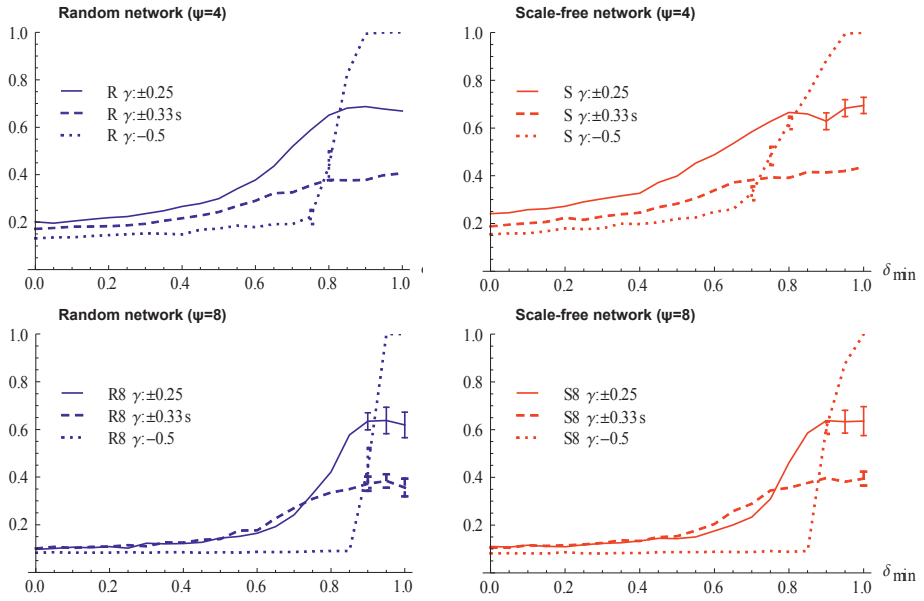


Figure D.3.: Mean payoff when considering the Random (R) and Scale-free (S) community structures. Note that the level of user dynamics is $m = 0.01$ while the fraction of 'nice' users is $\theta = 0.04$. ψ denotes the minimum number of peers of each community member at the start of each simulation run.

in the scale-free and random networks seems to have a contagious effect as δ_{min} increases. A moderate level of cooperative behavior can emerge in a random or scale-free network, rather than in the conventional complete-graph structure. The results for the random and scale-free networks seem similar most likely due to the limited structural variation given a relatively small $N = 100$.

When δ_{min} is low or moderate, the presence of 'nice' users help to ensure a minimum level of contribution in the system. Suppose that users adjust their perception on the importance and benefit of collaborative security (i.e., their δ_i) based on their initial payoff, the presence of some 'nice' users in the early phase of the system will thus be crucial.

The bottom left and right of Figure D.3 show the average payoff when each player is connected to at least 8 peers ($k \geq \psi = 8$) at the start of simulation, in the random (R8) and scale-free (S8) structures respectively. We observe a maximum number of 30 peers in R8, and 64 in S8. Compared to the performance when $\psi = 4$, the graphs are steeper but still converge to a moderate level of average payoff, when using the linear ± 0.25 and sigmoid ± 0.33 influence rules. The condition for a moderate level of contribution is harder but remains achievable in the scale-free and random networks, as the minimum number of peers increases.

D.6. Related Work & Discussion

Based on the initial work by Hirshleifer [15], Varian studied the problem of free-riding and system reliability of three different security games: total-effort, best-shot and weakest-link [24]. Grossklags et.al. extended the models to include the possibility to self-insure, and studied several interesting aspects such as the effects of information asymmetry between naïve users and security experts [12, 13]. Among other findings, the problem of under provisioning was shown to worsen as the number of players increases. This is an unfavorable finding for collaborative security systems in general; however, the aforementioned works have not factored in the players' long-term consideration. We reason that a CSS can be modeled using an infinitely repeated game, and the value of common protection being determined by the normalized sum of effort given by individual members and their peers.

Kandori showed that cooperative behaviors can be sustainable when players place sufficient weights on future payoffs through [16] in which players were repeatedly matched into pairs to play the classical 2-player prisoner's dilemma. In the P2P domain, Feldman et.al. proposed a reciprocative decision function along with a set of mechanisms (e.g., discriminatory server selection, maxflow-based subjective reputation and adaptive stranger policy) to mitigate the problems of free-riding and white-washing [9]. An insightful analytical model was also devised [10]. Yet, these findings cannot be directly generalized for a community-based system where there are $n > 2$ players interacting simultaneously. One cannot pin-point the non-contributors in the multi-player game so to play a reciprocal strategy, such as the well known Tit-For-Tat [4]. The threat of punishment on free-riders is diffused due to the implicit anonymity in a multi-player game [17].

There is a wealth of literature that studied cooperative behaviors in biological systems using the evolutionary game. Pioneered by [19], researchers started to look at the effects of spatial structures on the evolution of cooperation. A cooperative behavior was found to be the dominant trait in a scale-free network [22]. Ohtsuki et.al. [20] later showed that a cooperative action is preferred in various structures (circle, lattice, random, regular and scale-free graphs) whenever the benefit divided by the cost of contribution is greater than the average degree of individual members, i.e., $b/c > k$. However, these studies were also conducted using the classical 2-player prisoner's dilemma. Specifically, we note that the dilemma for contributing to the common protection in a n-player setting diminishes if $b/c > n$, since every member will strictly prefer to contribute or cooperate.

A few studies have looked at iterated multi-player games. Seo et.al. [23] analyzed the impact of a local opponent-pool from where a fixed amount of players ($n = 4, 8, 16$) were selected to play the n-player prisoner's dilemma, in every iteration of the population evolution. They found that the smaller the local opponent-pool size, the easier it was for cooperation to emerge. Rezaei et.al. [21] studied the co-evolution of cooperation and network structure using the iterated n-player prisoner's dilemma with $2 \leq n \leq 10$. Our work is different from theirs in two ways. First, we use the framework of a repeated game with discounted future payoffs (similar to in [3] which models the dishonest behavior of multi-cast agents

in network overlays) instead of an evolutionary game. Second, we do not fix the number of players; each player engages in each game round with all his peers. The number of peers per individual members varies depending on the underlying community structure.

D.7. Concluding Remarks

Starting with a complete-graph which depicts the community structure of global systems such as PhishTank and WOT, our analysis show that, without any incentive schemes, the level of user contribution can be expected to be very low. Education may help to inform about the importance of common protection and to cultivate a sense of social responsibility; however, this is challenging as users do not perceive security as their primary concern. The presence of generous users in a complete-graph community only helps in a very limited way to encourage contribution. Meanwhile, over-reliance on a small group of generous users may result in highly skewed contribution pattern, increasing the risk of manipulation.

Yet, this should not shun the idea of a community-based security system immediately. Our analysis has not factored in potential incentive schemes such as reputation, micro-payment and punishment mechanisms that have been proposed in the field of P2P networks. In addition, our simulation results show that it is possible to encourage a moderate level of cooperative behavior in the random and scale-free community structures. Designing a robust incentive scheme and a reliable aggregation strategy to collate user inputs from social networks for security purposes remains as an interesting research area.

Future work. Our current analysis can be extended in several directions. First, we note that it may be interesting to model a community-based security system as a 'best-k-effort' game (adapting from the best-shot game in [12, 13, 15, 24]), if it would suffice to have k out of n inputs from the community members for full protection. It would be also interesting to consider an endogenous discounting-factor such that individuals update their perception towards the long term importance the common protection, based on the payoff stream they receive. Extension work to consider heterogeneous contribution cost and protection benefit may also yield interesting insights.

References

[1] PhishTank. `http://www.phishtank.com`.

[2] Web of Trust. `http://www.mywot.com`.

[3] M. Afergan and R. Sami. Repeated-game modeling of multicast overlays. In *INFOCOM*. IEEE, 2006.

[4] R. M. Axelrod. *The evolution of cooperation*. Basic Books, 1984.

[5] A.-L. Barabási. Scale-Free Networks: A Decade and Beyond. *Science*, 325(5939):412–413, July 2009.

[6] A.-L. Barabási and R. Albert. Emergence of scaling in random networks. *Science*, 286(5439):509–512, October 1999.

[7] L. J. Camp. Reliable usable signaling to defeat masquerade attacks. In *Proceedings of the 5th Workshop on the Economics of Information Security*, WEIS '06, 2006.

[8] P. Dourish, R. E. Grinter, J. D. de la Flor, and M. Joseph. Security in the wild: user strategies for managing security as an everyday, practical problem. *Personal and Ubiquitous Computing*, 8(6):391–401, 2004.

[9] M. Feldman, K. Lai, I. Stoica, and J. Chuang. Robust incentive techniques for peer-to-peer networks. In J. S. Breese, J. Feigenbaum, and M. I. Seltzer, editors, *ACM Conference on Electronic Commerce*, pages 102–111. ACM, 2004.

[10] M. Feldman, C. H. Papadimitriou, J. Chuang, and I. Stoica. Free-riding and whitewashing in peer-to-peer systems. *IEEE Journal on Selected Areas in Communications*, 24(5):1010–1019, 2006.

[11] N. S. Glance and B. A. Huberman. The outbreak of cooperation. *The Journal of Mathematical Sociology*, 17(2):281–302, 1993.

[12] J. Grossklags, N. Christin, and J. Chuang. Secure or Insure? A game-theoretic analysis of information security games. In J. Huai, R. Chen, H.-W. Hon, Y. Liu, W.-Y. Ma, A. Tomkins, and X. Zhang, editors, *WWW*, pages 209–218. ACM, 2008.

[13] J. Grossklags, B. Johnson, and N. Christin. When information improves information security. In R. Sion, editor, *Financial Cryptography*, volume 6052 of *Lecture Notes in Computer Science*, pages 416–423. Springer, 2010.

[14] G. Hardin. The tragedy of the commons. *Science*, 162:1243–47, 1968.

[15] J. Hirshleifer. From weakest-link to best-shot: The voluntary provision of public goods. *Public Choice*, 41(3):371–386, 1983.

[16] M. Kandori. Social norms and community enforcement. *The Review of Economic Studies*, 59(1):63–80, 1992.

[17] P. Kollock. Social dilemmas: The anatomy of cooperation. *Annual Review of Sociology*, 24(1):183–214, 1998.

[18] T. Moore and R. Clayton. Evaluating the wisdom of crowds in assessing phishing websites. In G. Tsudik, editor, *Financial Cryptography*, volume 5143 of *Lecture Notes in Computer Science*, pages 16–30. Springer, 2008.

[19] M. A. Nowak and R. M. May. Evolutionary games and spatial chaos. *Nature*, 359(6398):826–829, Oct 1992.

[20] H. Ohtsuki, C. Hauert, E. Lieberman, and M. A. Nowak. A simple rule for the evolution of cooperation on graphs and social networks. *Nature*, 441(7092):502–505, May 2006.

[21] G. Rezaei, M. Kirley, and J. Pfau. Evolving cooperation in the n-player prisoner's dilemma: A social network model. In K. B. Korb, M. Randall, and T. Hendtlass, editors, *ACAL*, volume 5865 of *Lecture Notes in Computer Science*, pages 43–52. Springer, 2009.

[22] F. C. Santos and J. M. Pacheco. Scale-free networks provide a unifying framework for the emergence of cooperation. *Phys. Rev. Lett.*, 95:098104, 2005.

[23] Y.-G. Seo, S.-B. Cho, and X. Yao. The impact of payoff function and local interaction on the n-player iterated prisoner's dilemma. *Knowledge and Information Systems*, 2:461–478, 2000.

[24] H. Varian. System reliability and free riding. In L. Camp and S. Lewis, editors, *Economics of Information Security*, volume 12 of *Advances in Information Security*, pages 1–15. Kluwer Academic Publishers, 2004.

E. Use of Ratings from Personalized Communities for Trustworthy Application Installation

Author

Pern Hui Chia, Q2S NTNU
Andreas P. Heiner, Nokia Research Center
N. Asokan, Nokia Research Center

Conference

15th Nordic Conference in Secure IT Systems (NordSec)
27–30 October 2010, Espoo, Finland

Abstract

The problem of identifying inappropriate software is a daunting one for ordinary users. The two currently prevalent methods are intrinsically centralized: certification of "good" software by platform vendors and flagging of "bad" software by antivirus vendors or other global entities. However, because appropriateness has cultural and social dimensions, centralized means of signaling appropriateness is ineffective and can lead to habituation (user clicking-through warnings) or disputes (users discovering that certified software is inappropriate).

In this work, we look at the possibility of relying on inputs from personalized communities (consisting of friends and experts whom individual users trust) to avoid installing inappropriate software. Drawing from theories, we developed a set of design guidelines for a trustworthy application installation process. We had an initial validation of the guidelines through an online survey; we verified the high relevance of information from a personalized community and found strong user motivation to protect friends and family members when know of digital risks. We designed and implemented a prototype system on the Nokia N810 tablet. In addition to showing risk signals from personalized community prominently, our prototype installer deters unsafe actions by slowing the user down with habituation-breaking mechanisms. We conducted also a hands-on evaluation and verified the strength of opinion communicated through friends over opinion by online community members.

E.1. Introduction

The versatility of mobile devices paves the way for a large array of novel applications; mobile devices today contain ever more sensitive information such as medical data, user location and financial credentials. As device manufacturers open up the mobile platforms to encourage third party software development, applications from different sources are becoming available. Some of these applications, although not malicious, are inappropriate in the sense that they can cause harm (e.g., loss of privacy) or offense (e.g., culturally or religiously-insensitive content) to some users. The appropriateness of FlexiSpy – one of several commercial applications intended to spy on the activities of the user of a mobile phone – has been contentious. Mobile applications with potentially inappropriate content are becoming publicly available[1].

The bar for developing "applications" is also being lowered drastically. One can now develop simple applications for mobile devices by using only scripting languages (e.g., using JavaScript+HTML+CSS for Palm webOS [2]), or even without much programming experience using online tools (e.g., OviAppWizard [6] and AppWizard [1]). These applications are unlikely to be malicious (as they don't do too much) but we can expect a flood of applications from a larger variety of originators which increases the chance of a given application offending a certain group of users.

E.1.1. What is Inappropriate Software?

StopBadware.org [8] defines badware as software that fundamentally disregards a user's choice about how his or her computer or network connection will be used. In addition to software with malicious intent, the definition covers bad practices, such as installing additional unexpected software, hiding details from users, and incomprehensible End User License Agreement (EULA) that hinder an informed consent. Our understanding of "inappropriate software" is close to this notion of badware. In addition to maliciousness and disregard of user-choice, we consider software appropriateness to cover also the cultural and social dimensions.

E.1.2. Software Certification and its Limitations

A dominant approach for reducing the risk of malicious software on mobile platforms (e.g., Symbian, BlackBerry, J2ME and Android) is to rely on software certification and platform security. Software certification (e.g., Java Verified Program [4] and Symbian Signed [9]) is usually subject to software testing conducted by an authorized third party using publicly available criteria. But testing typically focuses only on technical compliance such as proper usage of system resources, proper application start/stop behavior and support for complete un-installation. Platform security (e.g., Symbian OS Platform Security [21] and Java Security Architecture [19]

[1] A search using the keyword 'entertainment' in the iTunes Appstore returns a number of applications with potentially mature content.

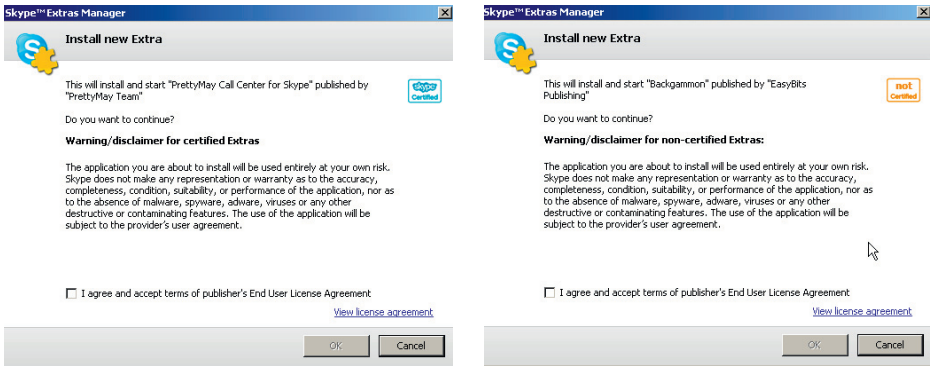


Figure E.1.: The Skype PC version has a list of 'featured extras' that include both Skype-certified and non-certified plugins. The visual difference when installing the two types is only the color of certification label (light-blue vs. soft-yellow).

refers to the isolation and access control features of the operating system or runtime. Ideally, software certification and platform security are used in tandem: an application is granted the privileges it requires if it is signed by a party trusted by the device platform. However, certification does not guarantee software security. It also does not consider the social and cultural aspects of software appropriateness.

Uncertified software: The Risk of Habituation. Many application installers (in mobile or desktop environment) resort to displaying warning and disclaimer notices to signal risks when software to be installed is not certified. Visual difference when installing certified and non-certified software is often low; the text is also typically uninformative (see Figure E.1). Providing system-generated notifications to which user attends to maintain security is the practice of "security by admonition" [33]. Besides degrading user experience, such notices lead to a high rate of false-positives causing many users to habitually click-through them. Click-through behavior is further entrenched when warnings equating "uncertified software" as possibly "harmful" may contradict other signals a user receives. An example of this is the installation of Gmail application (Figure E.2, left); the installer warns that it is 'untrusted' and 'maybe harmful' since it is not certified. A user, who trusts Google and who has just downloaded the application from Google's website will ignore and click-through the warning.

Certified software: The Risk of Centralized Judgment. On the other hand, software certified by a central authority may be perceived as inappropriate by some communities. An example of this is FlexiSpy – advertised as a tool to monitor the work force and protect the children and is available on most mobile platforms. The application has a number of characteristics that can be construed inappropriate: it spies on user activities (call, SMS, email, location), is invisible in the application list, uses a deceptive name (RBackupPro) and allows the device

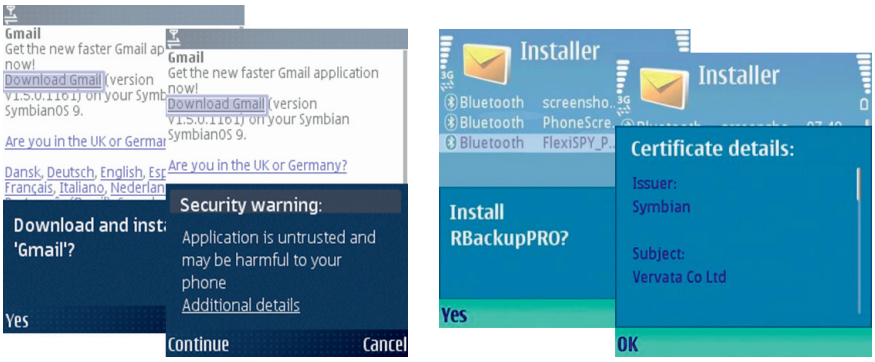


Figure E.2.: (Left) Gmail is not certified. (Right) FlexiSpy is certified [3].

to be controlled remotely. F-Secure flagged it as spyware that may be used for malicious purposes illegally [3] but as FlexiSpy fulfills the certification criteria, it is Symbian certified. In other words, a user is given a warning (Figure E.2, left) when he tries to install Gmail although he may likely trust it, whereas FlexiSpy can be installed without warnings (Figure E.2, right) even though he may belong to the group of people who consider it inappropriate.

On iPhone, Apple decides which third party applications can be distributed through the iTunes Appstore; we regard this as a scheme of implicit certification. Apple has also the means to activate a "kill-switch" to disable applications that may have been "inadvertently" distributed and later deemed "inappropriate by Apple". Apple's review criteria are, however, not publicly available. This has resulted in outcomes that are contested by developers and the Electronic Frontiers Foundation [5]. South Park, Eucalyptus and the Stern.de reader were among applications that were deemed "inappropriate by Apple" but later approved after protests [5]. Such contentions exemplify that centralized judgment can hardly cater for the value systems of different users.

E.1.3. Our Contribution

- We derived a set of design guidelines, grounded in cognitive and information flow theories, for a trustworthy software installation process (Section 2). Although we focus on mobile devices here in this paper, the guidelines are applicable to other platforms (e.g., desktop, Facebook) where installation by end-users can take place.

- We surveyed for the behaviors during installation, and we found high relevance of information from friends/family and user motivation to protect them. (Section 3)

- We built and evaluated a prototype system (Section 4 & 5). Although we could not test the efficacy of our prototype against habituation, we verified

that opinion by friend is of higher impact than that of by online community through the user study.

E.2. Designing a Trustworthy Installation Process

We consider that *a trustworthy installation process* to be one that helps users to avoid installing inappropriate application. Besides providing risk signals that are perceived reliable and relevant, the installer should take into account of the risk of habituation, which undermines the efficacy of many security mechanisms involving end-users.

E.2.1. Cognition during Application Installation

In the conventional installation task flow, as a user defines his expectation or desired software functionality (for a task at hand), he starts by searching for an application in the application market or on the web that meets his requirements. When such an application is found (and downloaded), the user will have to perform some "post-selection" actions such as accepting security-related conditions and configuration options before he is able to use it (objective attained). These "post-selection" steps are nearly always made without the user paying attention to what is asked. Habituation to click-through this "post-selection" phase could be attributed to current design of installation that lacks understanding for user's cognition.

To develop guidelines that take into account of user's cognition, we draw on the dual processing theory [23] in cognitive science, which identifies two main types of cognitive processes: *controlled* and *automated* processes.

Controlled processes are goal-directed; a user defines an objective and plans a path that (in his opinion) will lead to the objective. At certain points, the user will make an appreciation of the current context in order to decide on the next best-move in achieving his end goal. This process is highly dynamic and requires logical thinking. For these reasons, one can execute only one controlled process at a time. Appreciation of the current context and decision for a course of action, over time, can be based on superficial comparison of contexts. This leads to faster decision making [18, 23]. Despite a potential high degree of *automation in decision making*, it remains a controlled process as one will always have to compare between multiple contexts.

Automated processes such as habits, on the other hand, pose little to no cognitive load. Habits develop from deliberate into thoughtless actions towards a goal. If the context for an action is nearly identical over a series of performances, the action becomes mentally associated with the context; observing the context is enough to trigger the action [11, 26]. The simpler a task, the more frequently it is executed and the higher similarity in context, the stronger a habit can become. New information that invalidates the initial conditions (which led to an action or habit) will go unnoticed.

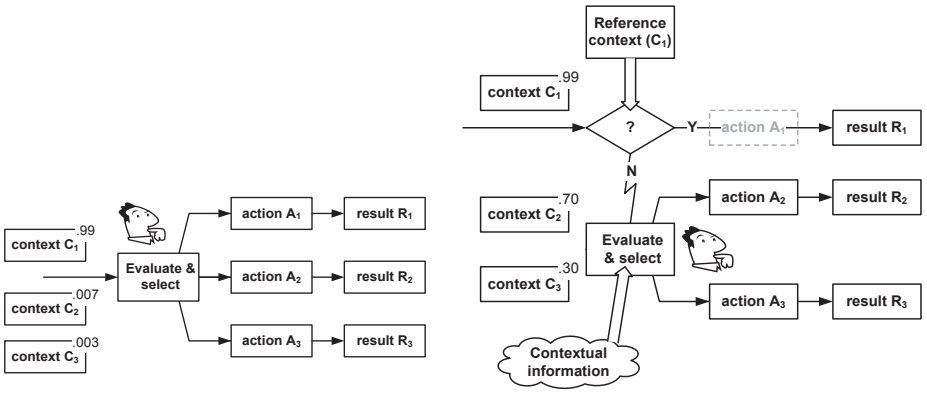


Figure E.3.: (Left) A constant context results in habitual behavior. (Right) Using the attention capture process with the dominant context as reference prevents this.

The difference between *habits (automated)* and *automation in decision making* (controlled) lies in the constancy of the context. Habits are developed if the context is (nearly) always the same. With the latter, context varies between a number of states with reasonable likelihood, thus requiring a controlled process of context comparison.

The constant context (lack of context-sensitive information) during installation makes the action of confirmation a habit. This is exemplified in Figure E.3 (left); the context of a normal installation flow (C1) demands the decision of action A1 (install) that results in R1. An abnormal context (C2) should lead to R2 (installation aborted). But as context C1 occurs much more often (denoted with probability .99) than context C2, user will over time expect context C1 and habitually selects action A1. This is more likely if there is no clear visual difference between the contexts (e.g., Figure E.1). Furthermore, from a user perspective, the choice (install or abort) is asked after the last conscious step of having decided to download and install a particular application. Users also rarely face immediate consequence for installing inappropriate software.

We argue that habituation can be avoided by eliminating the need for user action in the normal and frequent context (an easy target of habituation) altogether. Depicted in Figure E.3 (right), context C1 can be taken as reference context with an implied action A1. User can then be made aware of the deviation from this reference context through attention capture – the process of making a user aware of a change in environment that (may) require the user to attend to a new task [29]. A predominant view is that *attention capture* is an automated, stimulus-driven process modulated by the current controlled task [27]. The cognitive load required for the current task, as well as the strength and the relevance of the stimulus to the current task, affect the likelihood that a person will act on the stimulus. Thus, in addition to *visual salience*, the *relevance* and *strength of a warning (risk signal)* are

paramount to ensure that users will take note of and evaluate the warning, during the installation process.

E.2.2. Information Flow & Risk Signaling

Software warnings (risk signals) have conventionally been communicated to users in a hypodermic-needle manner by expert entities (e.g., antivirus vendors). These risk signals are designed against malware and do not cover for aspects such as the respect for user choice and the social/cultural factors of software appropriateness.

In search of risk signals that are relevant and of high impact, we refer to the two-step flow theory [24] – the founding work of innovation diffusion theory – which describes how communication can be more effective through a network of people (rather than through the hypodermic-needle fashion). Central to the theory are the information brokers (originally known as opinion leaders in [24]) who are not necessarily the most knowledgeable but are nevertheless skillful in interconnecting people [14]. Information brokers guide the information flow from sources into separate groups (first step) given incentives such as early information access and social capital [14]. When information gets into a particular group, competition among group members can serve to encourage each other to improve own knowledge and exchange opinions, which constitutes the second step of information flow [14]. Social media such as Twitter and Facebook are successful examples that have harnessed the power of social networks for effective communication. Use of social networks for provisioning or relaying of risk signals is, however, still an early concept.

PhishTank [7] and Web of Trust (WOT) [10] are systems that employ "wisdom of crowds" (using a global community, not personalized network) to improve web security. PhishTank solicits reports and votes against phish-sites, while WOT collects public opinions on the trustworthiness, vendor-reliability, child-safety and privacy-handling of websites. Both systems aggregate user ratings into global (rather than personalized) values. Such global values can, however, be susceptible to exploitation. Moore and Clayton [25] argued that as participation in PhishTank follows a power-law distribution, its results can be easily influenced by the few highly active users[2].

Prior work has pointed to the advantages of using inputs from personalized networks instead of the global community. Against phishing, Camp advocated for the use of social networks to generate risk signals that are trustworthy as the incentive to cheat is low among members who share (long-term) social ties [15]. Inputs from social networks can also be verified through offline relationship, allowing incompetent or dishonest sources to be removed [15]. Personified risks are also perceived greater than anonymous risks [12]; this may help to mitigate the psychological bias (known as valence effect) in which people overestimate favorable events for themselves. Inputs from social networks are also socially and culturally relevant.

[2]We note that this may be not too serious as determining whether a website is a phishing site (similar to whether an application is malicious) is usually objective. But judging if a website is trustworthy (with WOT, similar to evaluating the subjective factors of software appropriateness) can be contentious and prone to dishonest behavior (e.g., Sybil attack [17]).

E.2.3. Design Guidelines

To sum up, we consider that a trustworthy installation process should:

- **Avoid requiring user actions that can be easily habituated.** User actions in a normal and frequent context could be made implicit and complemented with an attention capture mechanism to signal any deviation from this context.
- **Employ signals that are visually salient, relevant and of high impact.** Signals should cover both the objective and subjective factors of software appropriateness.
- **Incorporate mechanisms to gather and utilize feedbacks from user's personalized community.** In this work, we refer a personalized community to friends and experts whom individual users trust in providing valuable inputs about software appropriateness. Experts could be vendors or gurus who are knowledgeable in the technical evaluation of software. A list of reputable experts can be set for all users by default. Meanwhile, Friends refer to ones whom users have personal contacts with and whom could help by sharing personal experience about applications or relaying information. Here, we hypothesize that risk signals from the personalized community can be more effective (due to their relevance and trustworthiness) than that of from global community. We verified the relevance and strength of inputs from friends in our survey (Section 3) and user study (Section 5).

E.3. Web-based Survey

We conducted an online survey to identify the installation behaviors and to evaluate the potentials of a personalized community in providing relevant and helpful signals.

E.3.1. Recruitment and Demographics.

We recruited our participants mainly from universities. We put up posters around popular campus areas. Emails were also sent to colleagues in other universities with the request to take part and to the forward the invitation to their contacts. Throughout the recruitment and responding process, we referred our survey as a study on user behaviors during installation using the title: "A Survey on Software Installation". Considerations were taken to avoid priming of secure behaviors. The reward for participation was to receive a cinema ticket on a lucky draw basis. Winners who do not reside in the Nordic region were rewarded with a souvenir-book. The lucky draw was made a few weeks after the data collection.

The survey was open for participation for 3 weeks. In total, 120 participants took part in the survey. Participants who did not complete all questions, or whose total response time was unrealistically low (<10 minutes) were excluded. The final population consists of 106 subjects (36% females). 12% have a PhD degree, 42%

Table E.1.: Demographics of survey participants

Education or Work background	
IT or Engineering	61%
Business or Finance	12%
Science or Math	8%
Arts and Social Science	10%
Others	9%

Age group	
18–24	15%
25–29	41%
30–39	32%
40–49	11%
50+	1%

Table E.2.: When know of digital risks

User would always or often inform	
friends or family	62%
members of online community	15%
expert individuals	14%
expert organizations	8%
antivirus software company	6%

have a Master degree while 28% have a Bachelor degree. 61% have a background in IT or engineering (power, electrical, mechanical, etc.) while 39% have a non-technical background (see Table E.1). Subjects took 15 minutes on average to complete the survey, which was structured into 12 questions with 105 items in total. We mostly used a 4-point Likert scale on the perceived importance of an element and the likelihood or frequency of performing an action.

E.3.2. Results.

We present a few interesting findings that we obtained. Finding-1 concerns the behaviors during installation while the others demonstrate the potentials of ratings from a personalized community. The percentage values were computed after reducing the responses from 4-point Likert scales into nominal levels of important/not, likely/not, or usually/seldom.

i. **Information during installation is mostly ignored.** 83%, 90% and 75% of the subjects reported that they seldom read the EULA, privacy policy and disclaimer notices respectively during the installation process. Similarly, 78% of the subjects seldom check for digital signatures (or software certificates), nor abort installation when they are absent. Only 30% usually abort installation given warnings from the installer. However, 69% usually abort installation if unnecessary personal questions were asked. 76% usually abort installation if warned by antivirus software, while 53% usually abort installation in the presence of advertisement pop-ups.

ii. **Security vendors, experts and friends are important sources for information on digital risks.** About 90% of the subjects reported that antivirus software is an important source of information about digital risks (e.g., harmful or inappropriate software/services). Expert organizations and

individuals also scored high (75%). Undeniably, security vendors and experts are the most important sources of information on digital risks. The survey gave further interesting results. 65% of the subjects regarded the first-hand experience by friends and family members as important. In comparison, fewer subjects (50%) considered the experience from members of an online community to be important. This difference was statistically significant ($p<.01$, Chi-square). This suggests that users regard inputs from friends and family members to be more relevant than that of from an online community.

iii. **When users know about digital risks, they are motivated to inform friends or family rather than the online community.** 60% reported that they could usually find security-related information by themselves. However, only 34% have been asked by friends or family members on whether software is trustworthy or appropriate. This could be due to the lack of existing system to share their opinions about software with his friends or family members. Indeed, we find that motivation to inform friends or family members about digital risks is high. 62% of the subjects would inform them about digital risks. Comparatively, only 15% were motivated to inform the online community (see Table E.2). The difference was statistically significant ($p<.0001$). This suggests that users have more motivation to protect his friends than members of online community. This supports the feasibility of a rating system based on personalized communities over the global-community compatriot.

iv. **Users consider reviews from trusted sources to be helpful.** With considerations to the limited screen size of mobile devices, 80% regarded reviews from trusted sources to be important/helpful information during software installation.

E.3.3. Limitation and Discussion.

We note that the education level of the participants was high, and 61% of the subjects have a background in IT or engineering. Yet, even though we might expect the subjects to be more aware of digital risks, there is an evident 'click-through' behavior. Excluding those with an IT/Engineering background, slightly fewer subjects (51%) could usually find security-related information themselves. However, the key results remain unchanged: 66% regarded friends as important source of risk information; 60% would inform friends or family when know about digital risks (compared to only 12% would inform such risks to an online community); 72% perceived reviews and ratings from trusted sources to be important/helpful information during software installation.

E.4. System Architecture and Prototype

Two important components in our architecture are: (i) *software repository*, which maintains a list of applications available for installation and a software catalog

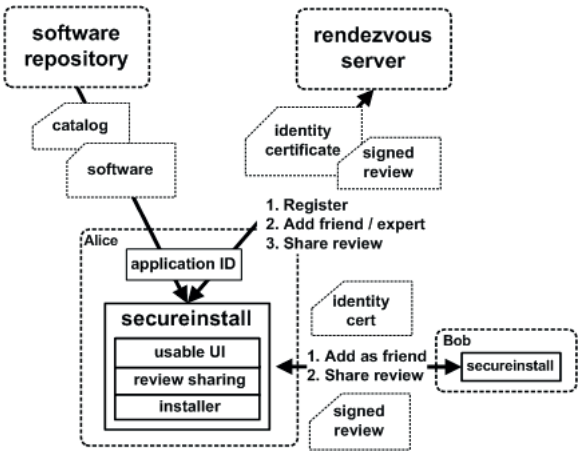


Figure E.4.: System Architecture. The prototype was implemented on the Nokia N810 tablet, while a rendezvous server was setup on an Ubuntu desktop. The prototype interacts with conventional software repositories to obtain application catalog and installation packages.

(containing metadata such as price, author, description and keywords); (ii) *rendezvous server*, which issues identity certificates and manages the user database, social graph and application reviews. To use the prototype installer (developed on the Nokia N810 tablet), a user must first register and obtain his credentials at the rendezvous server. Thereafter, the user can add friends and experts whom he trusts into his personalized community, and share software reviews with them, using the prototype. Sharing is done through the rendezvous server, Bluetooth or email. Software reviews are digitally signed and verified on the prototype to ensure authenticity and integrity.

The installation task flow was redesigned. When a user defines his requirements and searches for suitable applications (using some keywords), our prototype displays a list of related software (Figure E.5, top). The right panel shows basic information of a selected application, while detailed reviews from user's personalized community can be accessed by clicking on the "learn more" button. The "install" button will install an application without further prompting (if it has not been 'flagged' as potentially inappropriate by the user's personalized community). This removes user actions (in the post-selection phase of conventional task flow) that are prone to habituation.

For an application that has received *negative reviews* (i.e. flagged by the personalized community), a risk signal is shown prominently to catch the user's attention. To reflect the personal/social dimension of the warning, we chose a non-conventional risk symbol: a Pacman-like monster. Warning triangles and stop signs may signal that it is an "objective" opinion by some authorities.

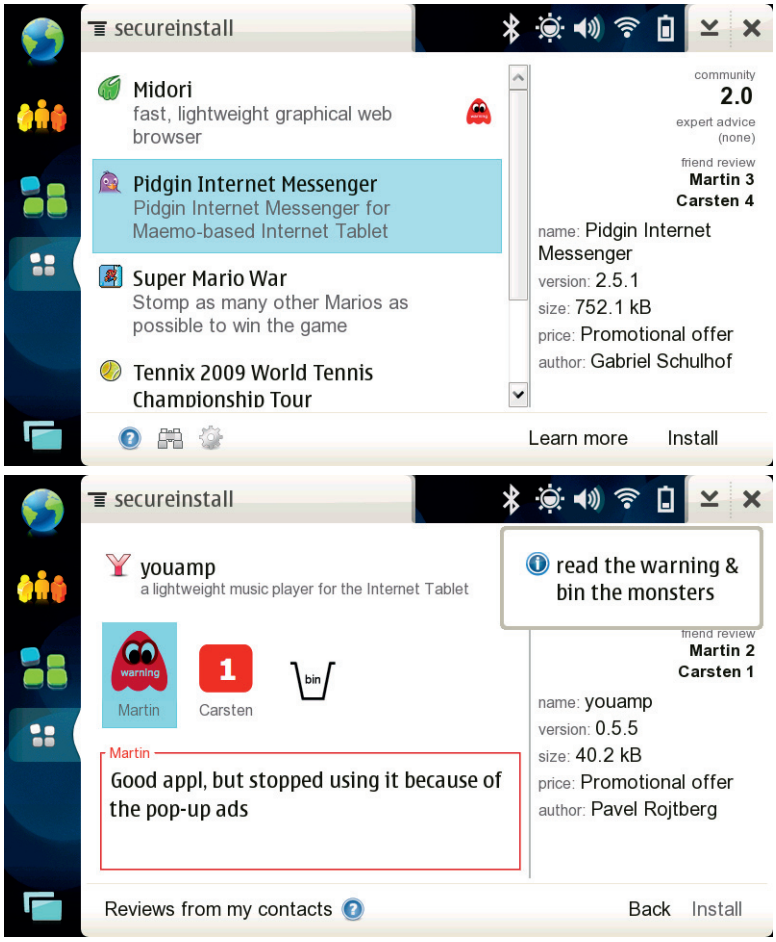


Figure E.5.: Prototype. (Top) The front-page shows an application list with basic description on the right panel. (Bottom) The experimental 'bin-the-monster' mechanism: user clicks on a monster to read the negative review; he has to drag it into the bin if he chooses to disregard the review. (Note that the reviews and ratings were artificially generated for evaluation purposes only).

The symbol is shown for flagged applications only; salience is increased by not showing positive cues. It is placed at the same level as the application name, and is enlarged when the application is highlighted. If the user decides to install an application that has been flagged, he is redirected to the review-page (Figure E.5, bottom) where he has to read the detailed reviews. Textual review improves the relevance of a risk signal as user can appreciate what is said better than numerical values [28]. Negative reviews are framed in red (bottom-up salience). To mitigate a potential click-through when attending to the negative reviews, we experimented with two habituation-breaking tasks (to improve the efficacy of attention-capture):

- Delay: User has to read every negative review by clicking on each of the monsters with some time interval. When clicked, a monster will disappear into an icon with numerical rating only after a few seconds, before the next review can be read.

- Bin-the-monster: As before, but the monster only disappears when it is dragged into the bin. User cannot install until all monsters have been binned (Figure E.5).

E.5. User Evaluation

We conducted a hands-on evaluation and investigated the strength of opinion given by friends compared to opinion given by online community members.

E.5.1. Recruitment & Demographics.

Participants were mainly recruited from universities. We distributed recruitment notes around popular campus areas especially in the social science and science/-math faculties. A web-form was also created to allow subjects to sign-up online. Participants of our survey were especially encouraged to take part if they reside in the Nordics; they were directed to the signup form upon completing the survey. Each participant was rewarded with two cinema tickets. There were in total 20 participants (7 females) consisting of students, researchers and a few working adults. 6 participants came from an IT background. The remaining subjects comprised of 6 mechanical, electrical or power engineering students, 4 science/math graduates, 3 art/design graduates, and 1 psychology undergraduate.

E.5.2. Experimental Setting.

We specified 4 testing days and arranged with the participants a suitable session of an hour each. Individual participants were invited to our premises where the study took place. Each session was preceded with a brief interview. The main task was structured into four evaluation scenarios. In the end, we asked for the overall experience with our prototype before a final debrief.

Table E.3.: (Left) 4 evaluation scenarios. (Right) Installation ratio in each scenario

		Install	Didn't Install
S1	No reviews from online community nor friends were provided	13	7
S2	Negative reviews were given by online community but friends gave positive reviews	10	10
S3	Positive reviews were given by online community but friends gave negative reviews	4	16
S4	Same as S3; the "bin-the-monster" mechanism was activated. After noting down the installation decision, subject was required to try installing the application (regardless of his decision) to experience the habituation-breaking interaction.	7	13

In the brief interview, we asked if a subject has encountered situations where he had difficulties or doubts in determining the appropriateness of certain software; all subjects responded that they had been in such situations before. We then requested the subject to write down the names of two friends whose opinions could be useful in these situations. We then keyed in these two names into our prototype system.

We gave the subject a script containing the description of the initial setting and the four evaluation scenarios (denoted as S1, S2, S3 and S4). The initial setting depicts a situation where there was a special offer on 4 applications which the subject would have to decide if he would like to buy and install. The special offer was meant to provide motivation to buy/install the applications in the evaluation scenarios. Two games, a browser and a media player (denoted as A1, A2, A3 and A4) were selected such that the likelihood of subjects having prior experience with them was low.

Having understood the initial setting, the subject was required to decide if he would buy/install a specific application in each evaluation scenario based on some basic description (application name, file size, name of developer, a short text provided by developer) and software reviews provided by online community members as well as the two friends mentioned during the brief interview.

Four negative reviews were scripted to signal a mild level of inappropriateness. They concerned advertisement pop-ups, pornographic content, program crashes (data loss) and suspicious elements. A set of positive reviews were also scripted. Each application was associated with a fixed pair of negative and positive reviews.

The evaluation scenarios were designed to present to the subject, positive and negative reviews from either friends or online community, as described in Table E.3 (left). We assigned the applications (A1, A2, A3 and A4) to the four scenarios in a rotating manner. Specifically, subject-1 would decide whether to buy/install applications A1, A2, A3 and A4, while subject-2 go through applications A2, A3, A4 and A1 in the fixed order of scenarios (from S1 to S2, S3 and finally S4). Rotating the applications in this manner avoided the potential bias due to the characteristics of individual applications and their fixed pair of positive/negative reviews.

The subject was required to write down his decision to whether buy/install in each scenario and the reason on the evaluation script. In scenario S3, we asked

Table E.4.: Results of hypothesis testing

	Chi-square	Binomial	Result
T1	$p = .080$	$p = .122$	NH1 cannot be rejected
T2	$p < .001$	$p < .001$	NH2 is strongly rejected
T3	$p = .004$	$p = .006$	NH3 is strongly rejected

for feedbacks on the use a Pacman-like monster as risk symbol. In scenario S4, we asked for experience with the "bin-the-monster" habituation-breaking mechanism. We used a 5-point Likert scale in both tasks.

Upon completing the four evaluation scenarios, we asked the subject his overall experience in using our prototype system in the form of descriptive feedback and a 5-point Likert scale (from terrible to great-idea). In the debrief, we informed the subject that all applications used were in reality good software available for the N810 tablet; all ratings and reviews had been scripted for experimental purposes only.

E.5.3. Results.

Installation count in each evaluation scenario is shown in Table E.3 (right). In S1, without any software reviews, 65% of the subjects went ahead to buy/install an application. The installation ratio decreased slightly (from 65% to 50%) in scenario S2 but dropped drastically (to 20%) in S3. Using the installation ratios, we evaluated the T1, T2 and T3 tests with the respective null hypothesis NH3, NH4 and NH5:

(T1) NH1: Negative community review does not overrule positive review by friend

(T2) NH2: Negative review by friend does not overrule positive community review

(T3) NH3: Overall strength of review by friend is not stronger than that of community review

Installation ratio in S1 served as the baseline of T1 and T2 tests (i.e. T1 compared the ratio in S2 to S1, while T2 compared the ratio in S3 to S1). Meanwhile, T3 was performed by comparing the ratio in S3 to S2. The hypothesis tests were evaluated using (one-tailed) Chi-square (good-of-fit) and binomial exact test. We favor results from binomial test as Chi-square statistics works better with a larger sample size.

We could not reject NH1 in T1. Although users reacted to negative reviews from online community members (resulting in a slightly smaller installation ratio in S2), the effect was not statistically significant. While we believe that users tend to react more towards negative reviews; warnings by online community members do not overrule positive feedbacks given by friends.

With T2, it was evident that negative reviews provided by friends overruled positive reviews by online community members. This was significant at 0.1% level.

Table E.5.: On using Pacman-like monster as risk symbol (1=strongly disagree, 5=strongly agree)

	μ	σ^2
Monster draws attention	4.3	.69
Monster gives clear message	2.8	1.3
Monster gives warning	3.8	1.1
Prefer monster over "!" sign	3.2	1.1
Prefer monster over "Stop" sign	3.4	1.1

Table E.6.: Overall user experience (1=terrible, 5=great idea)

	μ	σ^2
Experience with habituation-breaking	3.5	1.5
Experience with social rating integrated with software installation	4.4	.61

The large ratio difference (30%) between S3 and S2 suggested the higher impact of information from friends. We evaluated this in T3. The overall strength of reviews by friends is stronger than reviews by online community members (significant at 1% level). The strength of (risk) signals communicated via friends should be exploited to mitigate click-through and careless behaviors during software installation.

We observed that the installation ratio in S4 (35%) was higher than in S3 (20%). We tested if the "bin-the-monster" mechanism had inadvertently reduced the effectiveness of risk signaling, and found that the effect was significant at 10% level. With our experimental "bin-the-monster" mechanism, a bin was shown after some delay when user clicked on a monster. However, the sudden appearance of the bin might have that caused subjects to prioritize binning the monster over reading the review. As it might not be obvious that the monster could be binned, we tried to assist the users by showing a hint (Figure E.5, bottom). The short hint ("read the review and bin the monsters") might have been also construed as an instruction (or suggestive that it was ok to install) rather than to encourage a conscious decision. Our experimental 'bin-the-monster' mechanism was not a very successful one. An improved design could be to display the bin constantly to avoid a sudden appearance. The hint would need to be rephrased. A more direct association between the monster and review may also be helpful. For example, when user drags a monster into the bin, the corresponding review should be dragged together to signal that he is disregarding a review from his personalized community.

The reactions to the use of the monster as risk symbol were mixed (Table E.5). While most subjects agreed that it drew attention (salient), a few noted that they did not get a clear message of risk/warning. Subjects remained neutral on preferring the monster over the conventional "stop" and "exclamation-mark" symbols. We interpret these as using a new risk symbol would demand extra effort in educating the users.

Experience with the experimental "bin-the-monster" habituation-breaking mechanism was diverse (Table E.6). Some liked it and found it interesting, while a few found such mechanism unnecessary. We note that habituation-breaking mechanisms are designed to trade off some level of convenience for safer user actions, and may be hard to satisfy all users. Feedback on social rating (for software appropriateness) integrated with the installation process was, on the other hand, very positive. This suggests that it could be a useful feature on mobile devices (or other

computing environments that involve installation of third party applications by ordinary users).

E.5.4. Limitation and Discussion.

There are two weaknesses with regard to our user study. We note that the T3 test might have an order-bias as subjects were always required to complete scenario S2 before proceeding to S3. We should have mitigated this by randomizing the order of test scenarios.

We note that also the initial setting of "software offer" to provide subjects with motivation might not be very realistic. An alternative setting is to have the subjects to decide whether to buy/install an application on behalf of someone whom they care. However, we think that both settings have limitations that are hard to avoid in a laboratory testing. We could create a sense of realistic risks, for example by informing the subjects that they would be required to login to his email/bank account using the test device after the study. Yet, we thought that this was not too relevant as we did not require the subjects to evaluate whether to install software that are potentially harmful; our study concerned only applications that may be mildly inappropriate.

E.5.5. Summary of Findings

i. Opinions by friends are stronger than that of by online community; warnings by friends overruled positive feedbacks by online community, but not vice-versa

ii. The experimental "bin-the-monster" mechanism needs to be improved; designing and evaluating an effective habituation-breaking mechanism remain as interesting research problems

iii. The response towards habituation-breaking mechanisms and a new risk symbol was mixed; yet, majority was very positive with the idea of integrated social rating

E.6. Related Work

It is well-known by now that improving only the visual salience of risk signals is not enough to ensure secure user behaviors. Studies [30, 31] have shown the inefficacy of security toolbars and site-authentication images, which mainly rely on an improved risk salience. Brustoloni and Villamarín-Salomon [13] suggested using polymorphic dialogs (that will vary the order of decision options) to capture user attention and break habituation. They advocated also the use of audited dialogs that would keep track of user decisions to hold them accountable for irresponsible actions. However, subjects regarded audit dialogs as intrusive; audited dialogs also did not assist users to make better decisions. In addition to improving the visual salience (through a

better interface design), our work here increased the relevance of risk signals by employing inputs from user's personalized community.

Compared to FireFox's approach of making potentially unsafe actions (e.g., browsing a site with invalid certificate) more difficult to slow-down the users, our experimental habituation-breaking mechanisms (albeit need further improvements) are complemented with context-relevant information from personalized communities, that is absent in FireFox.

Related to software installation is the study by Good et.al. [20] which found that displaying a short summary (especially right-after the normal EULA notice) can effectively reduce the installation of unwanted applications. Yan et.al. concluded that visualizing the reputation and a personalized trust value for applications can be a helpful feature on mobile devices [32]. These studies highlighted the importance of timely signals. Our work integrated risk signals from personalized communities with the installation process. This integration was very well received in our user study.

Our idea of the personalized community is similar to NetTrust's [15] which employs personalized rating against the threat of phishing. NetTrust employs implicit inputs of browsing and bookmarking history of friends, as well as, explicit recommendations from third parties like banks and Google. Continuing from the initial work in [22, 16], in this paper, we have provided supports for the use of inputs from personalized communities, based on theories, a survey and a hands-on study on a prototype system.

E.7. Discussion & Future Work

Use of inputs from personalized communities is not without several shortcomings. We outline several challenges along with the potential mitigation strategies worth of future investigation.

Reliability. Inputs provided by user's personalized community may not be always correct. Information from technical sources may also be misinterpreted when guided through ordinary users. These issues can be mitigated by making the evaluation process more structured. For example, an evaluation can be divided into several aspects of software appropriateness rather than a single overall rating.

Coverage. Although users are likely to encounter similar applications with (some of) his friends in practice, undeniably ordinary users will have limited exposure and resources to identity all possible inappropriate applications. This is why we have included the notion of *expert users* (whom individual users trust) into the structure of a personalized community. A list of experts can be set by default (for all users) to deliver critical risk information. We could also extend our work to compute or infer recommendations for specific applications when there is no direct input from the personalized community. We note that there is much to learn from the field of recommender systems. However, this should be done with care so that the high relevance and strength of risk signals, as perceived by users, do not diminish.

Scalability. Software features such as usable contact and review sharing, reusability of reviews (across mobile platforms) as well as robust handling of software versions would be helpful to scale our implementation. Rather than building a system of social networks from scratch, we plan to merge the prototype with existing services (such as Facebook) that are now seamlessly integrated with smart phones.

Incentives. Like any community-based systems, there are challenges in initiating and sustaining user efforts. An important future work is thus to design an incentive scheme that would encourage active user participation. Here, we note that in contrast to a "crowds" system (i.e. one that employs a global community, such as PhishTank and WOT) where the success of the system is a public good, our work can benefit from unselfish behaviors among members in the personalized community. Indeed, we have seen strong motivation to protect friends and family members in our survey.

E.8. Conclusions

We developed a set of design guidelines grounded on theories for a trustworthy software installation process. Through a survey, we verified the high relevance of inputs from a personalized community and user motivation to protect friends and family. We implemented a prototype system with contact management and reviews sharing capabilities as well as a redesigned installation task-flow. Our user evaluation confirmed the strength of information communicated through friends, while the idea of integrated ratings from a personalized community during application installation was very well-received.

There may be some challenges that need to be addressed in future work; given the high relevance and strength of inputs from known sources, we show in this paper, the potentials of relying on personalized communities to evaluate software appropriateness and to mitigate the problem of click-through habituation during installation.

References

[1] AppWizard for iPhone. http://www.appwizard.com.

[2] Developing applications for Palm webOS using HTML, CSS and JavaScript. Article on ArcticStartup website. http://developer.palm.com/index.php?option=com_content&view=article&id=1603&Itemid=43. Last accessed: June 2012.

[3] F-Secure identified FlexiSpy as a spyware. http://www.f-secure.com/sw-desc/spyware_symbos_flexispy_f.shtml. Last accessed: June 2012.

[4] Java Verified Program. http://javaverified.com.

[5] Objections towards iTunes Appstore approval process. http://news.cnet.com/8301-13506_3-10317057-17.html, http://www.thelocal.de/society/20091125-23501.html, http://www.eff.org/deeplinks/2009/06/oh-come-apple-reject, http://www.eff.org/deeplinks/2009/05/apple-says-public-do, http://www.eff.org/deeplinks/2009/02/south-park-iphone-app-denied. Last accessed: June 2012.

[6] OviAppWizard for Symbian. http://oviappwizard.com.

[7] PhishTank. http://www.phishtank.com.

[8] StopBadware. http://www.stopbadware.org.

[9] Symbian Signed. https://www.symbiansigned.com.

[10] Web of Trust. http://www.mywot.com.

[11] H. Aarts and A. Dijksterhuis. Habits as knowledge structures: automaticity in goal-directed behavior. *Journal of Personality and Social Psychology*, 78(1):53–63, Jan 2000.

[12] Bruce Schneier. The psychology of security. http://www.schneier.com/essay-155.html. Last accessed: June 2012.

[13] J. C. Brustoloni and R. Villamarín-Salomón. Improving security decisions with polymorphic and audited dialogs. In L. F. Cranor, editor, *SOUPS*, volume 229 of *ACM International Conference Proceeding Series*, pages 76–85. ACM, 2007.

[14] R. S. Burt. The social capital of opinion leaders. *Annals of the American Academy of Political and Social Science*, 566:pp. 37–54, 1999.

[15] L. J. Camp. Reliable usable signaling to defeat masquerade attacks. In *Proceedings of the 5th Workshop on the Economics of Information Security*, WEIS '06, 2006.

[16] P. H. Chia. Secure software installation via social rating. Masters thesis, Helsinki University of Technology, and Royal Institute of Technology, 2008.

[17] J. R. Douceur. The sybil attack. In P. Druschel, M. F. Kaashoek, and A. I. T. Rowstron, editors, *IPTPS*, volume 2429 of *Lecture Notes in Computer Science*, pages 251–260. Springer, 2002.

[18] S. Frederick. Automated choice heuristics. In T. Gilovich, D. W. Griffin, and D. Kahneman, editors, *Heuristics and biases: The psychology of intuitive judgment*, volume xvi, pages 548–558. Cambridge University Press, 2002.

[19] L. Gong, G. Ellison, and M. Dageforde. *Inside Java 2 Platform Security: Architecture, API Design, and Implementation*. Addison Wesley, 2003.

[20] N. Good, J. Grossklags, D. K. Mulligan, and J. A. Konstan. Noticing notice: a large-scale experiment on the timing of software license agreements. In M. B. Rosson and D. J. Gilmore, editors, *CHI*, pages 607–616. ACM, 2007.

[21] C. Heath. *Symbian OS Platform Security*. John Wiley & Sons, 2006.

[22] A. P. Heiner and N. Asokan. Secure software installation in a mobile environment. In L. F. Cranor, editor, *SOUPS*, volume 229 of *ACM International Conference Proceeding Series*, pages 155–156. ACM, 2007.

[23] D. Kahneman. Maps of bounded rationality: Psychology for behavioral economics. *The American Economic Review*, 93(5):pp. 1449–1475, 2003.

[24] P. F. Lazarsfeld, B. Berelson, and H. Gaudet. *The People's Choice: How the Voter Makes Up His Mind in a Presidential Campaign*. Columbia University Press, 1944.

[25] T. Moore and R. Clayton. Evaluating the wisdom of crowds in assessing phishing websites. In G. Tsudik, editor, *Financial Cryptography*, volume 5143 of *Lecture Notes in Computer Science*, pages 16–30. Springer, 2008.

[26] D. T. Neal, W. Wood, and J. M. Quinn. Habits: A Repeat Performance. *Current Directions in Psychological Science*, 15(4):198–202, August 2006.

[27] R. J. Peters and L. Itti. Beyond bottom-up: Incorporating task-dependent influences into a computational model of spatial attention. In *CVPR*. IEEE Computer Society, 2007.

[28] J. S. Rubinstein, D. E. Meyer, and J. E. Evans. Executive control of cognitive processes in task switching. *Journal of Experimental Psychology: Human Perception and Performance*, 27(4):763–797, August 2001.

[29] M. Ruz and J. Lupiáñez. A review of attentional capture: On its automaticity and sensitivity to endogenous control. *Psicológica*, 23(2):283–309, 2002.

[30] S. E. Schechter, R. Dhamija, A. Ozment, and I. Fischer. The emperor's new security indicators. In *IEEE Symposium on Security and Privacy*, pages 51–65. IEEE Computer Society, 2007.

[31] M. Wu, R. C. Miller, and S. L. Garfinkel. Do security toolbars actually prevent phishing attacks? In R. E. Grinter, T. Rodden, P. M. Aoki, E. Cutrell, R. Jeffries, and G. M. Olson, editors, *CHI*, pages 601–610. ACM, 2006.

[32] Z. Yan, C. Liu, V. Niemi, and G. Yu. Trust indication's influence on mobile application usage. NRC Technical Report, Nokia Research Center, 2009.

[33] K.-P. Yee. Aligning security and usability. *IEEE Security & Privacy*, 2(5):48–55, 2004.

F. Is this App Safe? A Large Scale Study on Application Permissions and Risk Signals

Author

Pern Hui Chia, Q2S NTNU
Yusuke Yamamoto, Kyoto University
N. Asokan, Nokia Research Center

Conference

21st International World Wide Web Conference (WWW)
16–20 April 2012, Lyon, France

Abstract

Third-party applications (apps) drive the attractiveness of web and mobile application platforms. Many of these platforms adopt a decentralized control strategy, relying on explicit user consent for granting permissions that the apps request. Users have to rely primarily on community ratings as the signals to identify the potentially harmful and inappropriate apps even though community ratings typically reflect opinions about perceived functionality or performance rather than about risks. With the arrival of HTML5 web apps, such user-consent permission systems will become more widespread. We study the effectiveness of user-consent permission systems through a large scale data collection of Facebook apps, Chrome extensions and Android apps.

Our analysis confirms that the current forms of community ratings used in app markets today are not reliable indicators of privacy risks of an app. We find some evidence indicating attempts to mislead or entice users into granting permissions: free applications and applications with mature content request more permissions than is typical; "look-alike" applications which have names similar to popular applications also request more permissions than is typical. We also find that across all three platforms popular applications request more permissions than average.

F.1. Introduction

Ever since the personal computer changed the lives of people around the world, we have become accustomed to the notion of software applications. The personal computer world started out with completely open platforms where all applications (apps) ran with the same complete set of privileges available to the user. This quickly gave rise to the phenomenon of malicious and inappropriate software [19].

Operating system and runtime platform security schemes can be used to apply the principle of least authority to applications. Although various platform security schemes were developed since the 1990s, they saw widespread deployment only when they were incorporated into Java Security Architecture and into mobile device platforms [26]. All modern mobile device application platforms incorporate permission-based platform security schemes. Web application platforms like Facebook and browser extension runtimes like Google Chrome extensions also use permission-based platform security. Some of these platforms such as Apple's iOS rely on a central authority to decide what permissions can be granted to a given application. Others rely on the user making the authorization decisions. We will call the former category *centralized permission systems* and the latter *user-consent permission systems*.

In the smartphone arena, centralized permission systems are currently dominant, with the exception of the Android platform. However several HTML5 APIs (e.g., the geolocation API) support a user-consent permission system. The availability of comprehensive APIs including offline caching makes it possible for HTML5 web apps to offer similar functionality as native applications. If HTML5 web apps become dominant [28], their decentralized nature will also imply that user-consent permission systems will become more widespread.

Centralized permission systems take the burden of judgment away from users. While this is a benefit in terms of usability, it is problematic if the judgment in question is subjective. People and organizations may disagree on whether a certain application is privacy-invasive or offensive [17]. The biggest problem in user-consent permission systems is that users may become habituated to permission queries and may carelessly click through them. Even careful users have to make access control decisions based only on a few signals such as the average numerical rating given by other users, number of ratings and downloads. Users who care about their privacy [16, 13], may not have the ability to protect it in user-consent permission systems if the signals are unreliable or can be manipulated by developers of malicious apps.

In this paper, we investigate the current state of risk signaling on privacy intrusiveness of apps, and if there is any evidence of attempts to mislead or entice users of user-consent permission systems into compromising their privacy. Specifically we ask:

1. Do popular apps ask for more permissions than is typical for apps in general?
2. Are currently available signals about an application reliable in indicating privacy risks associated with that application?

3. Do developers of free apps and those with mature content ask for more permissions than is typical?

4. Do apps with "look-alike names" (i.e., names similar to popular apps) ask for more permissions than is typical?

Our paper is structured as follows. We start with a survey of related work in Section F.2 and describing the data collection processes in Section F.3. We present a basic analysis on app popularity, ratings and permissions in Section F.4 before proceeding to evaluate the effectiveness of current risk signals (Section F.5) and potential trends of enticements and tricks (Section F.6). We conclude by revisiting the above four questions and discussing the implications and mitigation measures in Section F.7.

F.2. Related Work

With the growing popularity of Android, there have been a number of publications on Android OS security and its permission system. Enck et al. [21] proposed Kirin certification to help identifying Android apps that request a suspicious permission combination using a set of predefined rules. Barrera et al. [15] studied the relationship between the permissions requested by 1,100 most popular and free Android apps by machine learning and proposed a methodology to improve the expressiveness of app permissions without increasing its overall complexity. Our study differs from theirs in that we do not look into the patterns of permission requests in details but we study the how the number of permissions requested by apps correlate with several signals (e.g., community ratings) that the users receive.

Felt et al. [24] studied the effectiveness of permission systems on Android and Chrome platforms. Using 1,000 most popular Chrome extensions, they pointed out that the first 500 most popular extensions have requested significantly more permissions than the second 500 extensions. However, they observed no differences between the 756 most popular and 100 most recent Android apps. With a much larger dataset, our study shows that there is a positive correlation between the number of installations and the number of permissions requested by the app on all three web and mobile application platforms: Facebook, Chrome and Android. In addition, we look into whether specific types of apps, including those with mature content, those flagged by external ratings and those with suspicious look-alike names, request for more permissions than is typical.

In comparison, fewer studies have investigated the Facebook permissions. King et al. [25] conducted a survey on the privacy knowledge, behaviors and concerns of Facebook app users. Our study differs from theirs in two ways. First our analysis relies on a large scale data collection rather than a self-reported survey. Second while they focused on the interaction between user's understanding, concerns and behaviors when using Facebook apps, we look at the availability of reliable risk signals on Facebook (as well as Chrome and Android) and how the absence of

them may have been exploited by some developers to entice or trick the users with questionable apps.

Moore and Edelman [29] studied the ecosystem of the typo-squatting fraud – the intentional registration of misspellings of popular website addresses. They estimated that at least 938,000 typo domains targeted the 3,264 popular .com sites they studied. They also found that 80% of the typo-squatting domains were supported through profits from advertising, typically from the pay-per-click advertisements. Our study is one of the first to analyze the naming exploitations in apps. While we have not conducted an in-depth analysis on the motivation and profitability of such naming tricks, we analyzed whether apps with look-alike names request for more permissions than is typical. A related work is by Barrera et al. [14] which studied the problem of a non-global app ID system for Android apps. They proposed *Stratus* to standardize the app IDs across different Android marketplaces. Our analysis in this paper focuses on look-alike names rather than look-alike app IDs. We expect the users to pay less attention to app IDs especially on Facebook and Chrome where the app IDs are in the form of a string of random digits or characters.

F.3. Data Collection

We detail the data collection processes for Android apps, Chrome extensions and Facebook apps in the following. We share our datasets on our project site [7].

F.3.1. Android Apps

Prior studies on Android OS security (e.g., [15, 21, 23, 24]) have mainly focused on the most popular apps on Android Market. In order to broaden the scope we used both the official Android Market [1] and AppBrain.com [3] to construct two datasets `Android (pop)` and `Android (new)`.

`Android (pop)` consists of popular Android apps selected randomly from the *top-selling-free* and *top-selling-paid* listings of Android Market, as well as the list of the most popular apps according to AppBrain.com on 15 June 2011. After removing duplicates, it contains 650 unique apps (323 paid and 327 free). `Android (new)` consists of 1210 new apps (610 paid and 600 free) which first appeared on the *most-recent-apps* section of AppBrain.com in mid June 2011 and were still available in early October. We kept track of these new apps and updated our database accordingly for changes in app details.

In addition to the above, to investigate the behavior of apps with look-alike names, we collected also a separate much larger dataset `Android (mr)`. The dataset consists of 20,500 new Android apps (11,095 free and 9,405 paid) constructed from the list of 20 most recent apps according to AppBrain.com on an hourly basis from mid August to early October 2011.

The application information page on Android Market provides a number of details that we use in this paper, including app installation count, average community

rating, rating count, developer URL, price, content maturity level, and permissions requested by the app.

F.3.2. Facebook Apps

We constructed the Facebook dataset by downloading the entire list of 34,370 Facebook apps (app names, IDs, and developer IDs) from SocialBakers [8], a portal providing the usage statistics of various social media. We then attempted to access the Facebook application page of each of these (using the Watir [9] library to login to Facebook). We excluded apps that have become unavailable or otherwise invalid or redirected to a page outside Facebook. Out of the remaining 27,029 Facebook apps, 18,205 request at least one permission from the user. For each of the 27,029 apps, we downloaded details including the number of monthly active users, the average rating, rating count, description and category. This constitutes the `Facebook (all)` dataset.

F.3.3. Chrome Extensions

Chrome Web Store [6] lists up to 1,000 most popular extensions in 12 different categories. As some of the categories such as 'Sports' and 'Shopping' have far less than 1,000 extensions, even the more recent extensions such as those with less than 10 users, were present on the lists. We constructed our dataset by downloading all 12 lists. Removing duplicates and extensions that became unavailable during data collection, the resulting `Chrome (all)` dataset consists of 5,943 extensions. It contains details from the information page of each extension, including the installation count, average community rating, rating count, developer URL, version of extension, supported languages, as well as the permission warnings associated with the extension.

F.4. Basic Analysis

We first study the link between app popularity and user ratings, and the statistics of permissions in the following.

F.4.1. App Popularity and User Ratings

We define the *popularity* of an app as the number of installations (in the case of Android apps and Chrome extensions) or the number of monthly active users (in the case of Facebook apps). Figure F.1 shows log-log plots of app popularity versus the number of user ratings the app has received. While we did not test if both the distributions of app popularity and the number of ratings per app follow a specific heavy-tailed distribution (e.g., Power-Law, Log-normal), they appear to be highly skewed visually. All four app popularity distributions curve in the log-log plot. On the other hand, other than the `Android (pop)` dataset where no apps

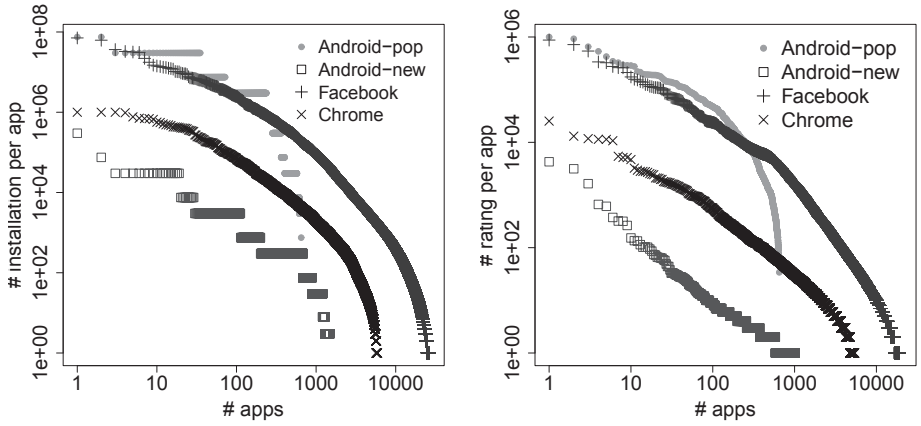


Figure F.1.: Distribution of the number of installations and number of ratings per app.

have less than 30 ratings, the rating contribution patterns appear to be straight lines in the log-log plot, suggesting that they could be a Power Law. Indeed, many online peer production systems can be characterized by a Power Law contribution pattern that is explainable by the participation momentum rule [31]. The skewed nature of the app popularity and the number of ratings per app prompted us to use the logarithmic values, i.e., $log(\#installation)$ and $log(\#rating)$ for popularity and rating count respectively in subsequent analysis.

We found a strong (Pearson) correlation between app popularity and the number of ratings of the app has received (with r ranging from 0.67 to 0.90, p<.001, see Figure F.2). Indeed, users are more likely to rate an app that they have installed. However, somewhat counter-intuitively, the average rating of an app, $avgr$, does not positively correlate with app popularity. In fact, $avgr$ is negatively and weakly correlated to the app popularity (r=−0.15, p<.001) among the new Android apps. We figure that the average user rating can indeed be misleading without factoring in its confidence level. Considering the confidence of an average rating to be proportional to the number of ratings that it has received, we thus measure the adjusted average rating to be:

$$avgr_a = (avgr - 3) * log(\#rating) \tag{F.1}$$

The –3 transformation is necessary as user ratings range from 1 to 5 across the different app platforms. As also shown on Figure F.2, there is a strong correlation between the app popularity and the adjusted average rating (with r ranging from 0.45 to 0.72, and p<.001). We evaluate if $avgr_a$ serves as an effective risk signal against potentially intrusive or inappropriate apps in Section F.5.

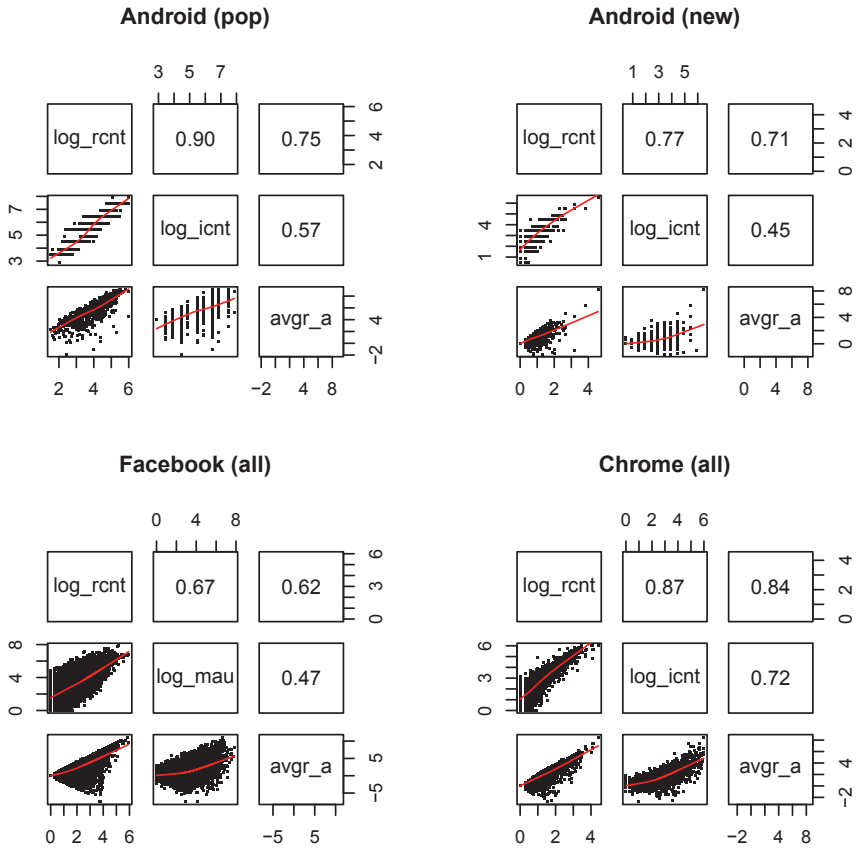


Figure F.2.: Scatter-plot matrices showing the pairwise correlation coefficients (upper right triangular panel) and the scatter-plots (lower left) of any given pairs of three variables: log of installation count, `log_icnt` (or monthly active users, `log_mau`), log of rating count, `log_rcnt`, and adjusted average rating, `avgr_a`. The scales of the variables are depicted on the borders of individual matrices. All correlation values are statistically significant with $p<.001$.

	pop	new
INTERNET	77	64
WRITE_EXTERNAL_STORAGE†	51	32
READ_PHONE_STATE†	45	32
WAKE_LOCK	23	15
ACCESS_FINE_LOCATION†	14	14
ACCESS_COARSE_LOCATION† (7)	10	6
READ_CONTACTS† (11)	7	3
CAMERA† (12)	6	3
READ_LOGS† (23)	6	1
RECORD_AUDIO† (19)	6	1
GET_TASKS† (8)	6	5
CALL_PHONE (6)	5	7

Table F.1.: Top 12 most requested dangerous permissions w.r.t. `Android (pop)`, and percentage of apps requesting them. Numbers in brackets show the different rankings w.r.t. `Android (new)`.

	%
Access my basic info.†	67
Post to Facebook as me.	23
Send me email.†	14
Access my profile info.†	5
Access my data any time.	2
Access my photos.†	2
Access information people share with me.†	2
Publish games and app activity.	2
Access posts in my news feed.†	1
Access my videos.†	1

Table F.2.: Top 10 most requested permissions, and percentage of Facebook apps requesting them.

† indicates a dangerous and info. relevant permission

	%
Your tabs & browsing activity.	58
Your data on <*some*> websites.†	41
Your data on all websites.†	35
Your bookmarks.†	2
Your browsing history.†	2
Your list of installed apps, extensions, and themes.†	1
Your physical location.†	1
All data on your computer and the websites you visit.†	1

Table F.3.: Percentage of Chrome extensions requesting a specific permission.

Perm. type	Android (pop)				Android (new)				Facebook (all)				Chrome (all)			
	μ	M	max	total	μ	M	max	total	μ	M	max	total	μ	M	max	total
P	4.5	4	15	137	3.0	2	20	137	1.2	1	13	23	1.4	2	6	8
P_{danger}	3.0	3	10	65	2.1	2	11	65	1.2	1	13	23	0.8	1	5	7
P_{dinfo}	1.6	1	7	34	1.0	1	5	34	0.9	1	11	14	0.8	1	5	7
R_p (%)				91				74				67				85

Table F.4.: Mean (μ), median (M) and maximum (max) number of P, P_{danger} and P_{dinfo} requested per app, as well as the total number of permissions on different platforms. R_p measures the percentage of apps requesting at least one permission.

F.4.2. Permission Statistics

Understanding the complex permission models of web and mobile apps can be a daunting task for many ordinary users. We briefly describe the permission systems from the most granular (Android) to the least (Chrome) below. In each case, we identify three sets: the set of **all permissions** P, the set of **dangerous permissions** P_{danger}, and the set of **dangerous and information relevant permissions** P_{dinfo}. P_{danger} consists of permissions for actions that can be potentially harmful to the user while P_{dinfo} is the subset of P_{danger} consisting of permissions that permit access to sensitive personal information of the user.

Android: Android permissions are categorized into 4 categories, namely *Signature*, *Signature-or-System*, *Normal* and *Dangerous*. The first two categories, *Signature-or-System* and *Signature* permissions protect the most sensitive operations on the Android devices. These permissions can only be granted to apps preinstalled into the device's `/system/root` folder and/or apps signed with the device manufacturer's private key. Requests to use these permissions by other apps without the right keys will be ignored [24]. On the other hand, the *Normal* permissions govern the functionalities which can be annoying (e.g., vibrating the phone), while the *Dangerous* permissions protect the user from operations that can be potentially harmful including those that cost money or potentially privacy intrusive [24]. The details of individual Android permissions can be found on [2].

We observed 137 different permissions in our dataset, out of which 65 are dangerous permissions. Following [22], we classified 34 of the 65 permissions in P_{danger} as belonging to P_{dinfo}. Table F.1 shows the most frequently requested dangerous permissions among the popular and new Android apps. The full list is available on our project site [7].

Facebook: The Facebook permissions have evolved as the platform and its user base grows. Each of the permissions protects a specific piece of personal information or a specific functionality on the platform (e.g., to login to Facebook chat system, publish user's activity on his wall). We consider all Facebook permissions to be dangerous. There are in total 62 Facebook permissions [4], which are grouped and presented to the users in 23 different categories. Each category is highlighted in bold and visualized with a unique icon on the permission request screen, while the individual permissions are described in gray text with a smaller font size under their respective category. The individual permissions requested can also be found on the URL of the permission request screen but we do not expect the ordinary users to notice them. For these reasons, in this study, we have focused on the permission categories as opposed to the individual permissions. For simplicity, we refer to Facebook *permission categories* as *permissions* interchangeably.

14 out of the 23 permission categories are information relevant. Table F.2 shows the most frequently requested permission categories. Notice all apps requesting a permission must also request for the *'Access my basic information'* permission. The most dangerous permission is perhaps *'Access my data any time'* which allows an app to access the user's information even when he is not online. Another interesting permission is *'Access information people share with me'* which allows the app to

access *not* information about the user himself, but the personal information of his friends.

Chrome: Among the three platforms, Chrome permissions are the least granular. There are only 9 permission warnings in total as detailed on [5]. Inline with the categorization of Android and Facebook permissions, we regard all Chrome permissions to be both *dangerous* and *dangerous-and-information-relevant*, except the permission *'Your tabs and browsing activity'* which was unnecessarily required for creating a new tab in earlier versions of Chrome browser [5].

The most dangerous is the permission to access *'All data on your computer and the websites you visit'* which is requested by the plugin-type extensions. These plugin extensions are basically native executables that run with full privileges on the user's machine. They are manually reviewed by Chrome before being accepted to appear on the Chrome Web Store [5]. We found only 47 plugin extensions in our dataset. Among the extensions that request for the permission to access *'Your data on <some> websites'*, the top 10 most frequently requested sites are: `google.com`, `facebook.com`, `tiny-url.info`, `plus.google.com`, `twitter.com`, `youtube.com`, `mail.google.com`, `g2me.cn`, `api.flickr.com`, and `reddit.com`.

Summary: Table F.4 shows the mean, median and maximum number of P, P_{danger}, and P_{dinfo} that an app requests in the different datasets. The ratio of apps requesting at least a permission is high across all four datasets: `Android (pop)` (91%), `Android (new)` (74%), `Facebook (all)` (67%) and `Chrome (all)` (85%). Another trend that holds across the different platforms is that most apps request only a small fraction of the total available permissions. On average (medium), Facebook and Android apps request for only $1/23 = 4\%$ and $3/65 = 5\%$ (3% among the new apps) of the total available dangerous permissions respectively. In addition, there is a noticeable trend that some permissions are more frequently requested than the others (see Table F.1, F.2, F.3). These reinforce the findings by Felt et al. [24] that an application permission system has the benefits of allowing the platform owners (i) to avoid granting the full privileges to third party apps, and (ii) to possibly recognize apps with anomalous permission request patterns for triaging the manual review process.

However, as it is with other security problems, the permission systems will not be effective if users do not comprehend how they work, or if the permission systems contradict other signals the users receive. One can easily exploit the lack of understanding and the absence of reliable risk signals for questionable or malicious purposes.

F.5. Effectiveness of risk signals

We look into the availability of reliable and intuitive risk signals during the process of app installation in the following.

	Correlation with $log(\#installation)$			
Perm. type	Android (pop)	Android (new)	FB (all)	Chrome (all)
P	0.26	0.12	0.28	0.15
P_{danger}	0.26	0.11	0.28	0.12
P_{dinfo}	0.22	0.10	0.28	0.12

Table F.5.: Correlation between app popularity and the number of P, P_{danger} and P_{dinfo} requested. All values are statistically significant with $p<.001$

	Correlation with $avgr_a$			
Perm. type	Android (pop)	Android (new)	FB (all)	Chrome (all)
P	0.15	0.08*	0.11	0.18
P_{danger}	0.14	0.06°	0.11	0.16
P_{dinfo}	0.11	0.04°	0.12	0.16

Table F.6.: Correlation between the adjusted average rating $avgr_a$ and the number of P, P_{danger} and P_{dinfo} requested. All values are statistically significant with $p<.001$, except for the case of new Android apps where * indicates $p<0.1$ and ° indicates $p>0.1$.

F.5.1. App Popularity

Table F.5 shows that there is a weak positive correlation between app popularity and the number of permissions requested. The trend holds true not only for our Chrome extension dataset (which is much larger than the dataset used by Felt et al. [24]), but also for Android and Facebook apps. Also, the correlation is stronger in popular Android apps than in new Android apps. As Felt et al. [24] hypothesized, a possible explanation is that popular apps need more permissions in order to offer more functionality that makes them more interesting or useful and hence popular. While this phenomenon is perhaps not surprising, it underlines the fact that careful users concerned about their privacy have to make a tradeoff between the functionality offered by an app and its potential for compromising their privacy. Although popularity of an app is an easily available and understandable signal to users of app marketplaces, it is not necessarily a reliable signal for privacy risk – a user cannot conclude that a popular app is a safe one.

F.5.2. Community Rating

Community rating reflects how the users perceive an app. As with the popularity measure, it is a meaningful signal that is also widely available in app marketplaces. If it is to be an effective signal for enabling the user to detect privacy risks, it would exhibit a negative correlation with the number of permissions requested. We found no such negative correlation between the adjusted average rating and number of permissions (Table F.6). In fact there was a weak positive correlation in all cases except for the case of the new Android apps, where the correlation was

not statistically significant. Again, the likely explanation is that user ratings in app marketplaces are based on functional aspects like features and performance rather than privacy risks.

F.5.3. External Ratings

Next, we studied how ratings from sources external to the marketplaces relate to the number of permissions. We considered two sources. First is the Web of Trust (WOT) [10] service. WOT is a community rating system which allows users to rate a website in four dimensions: trustworthiness, vendor reliability, privacy and child safety. It aggregates user ratings as well as information from other sources into a rating along each of the dimensions as well as a combined score. WOT ratings are usually given at the granularity of fully qualified domain names, unless a subdomain has received enough input ratings on its own. Websites where multiple users control their own subsections thus typically share a common rating inherited from the parent domain. Second is AppBrain.com which is a website for discovering Android apps. It provides a number of useful listings including the most popular Android apps in different countries and age groups, as well as the latest apps that appear in Android Market. To help its users, AppBrain.com labels apps created by developers who have made a high fraction of apps without any rating or with below average ratings, as *spam*.

We studied how WOT rating of the website of the app developer (where available) relates to the permissions requested by it. Table F.7 shows the average number of permissions requested by apps whose developer websites are deemed suspect (*bad* or *caution*) by WOT. We classified a WOT rating into *bad*, *caution*, *good* or *unknown* following the default risk signaling strategy of WOT as detailed in [18]. We ignored the *good* ratings from WOT as many developers use a shared domain as their website and WOT's verdicts will not be accurate in these cases. Contrasting with the mean values in Table F.4, we see that that the suspect apps consistently request more permissions than on average in all cases. The differences in the average number of P, P_{danger} and P_{dinfo} between suspect Facebook apps and Facebook apps in general are statistically significant with $p<.001$. The sample size is too small in the case of Android apps (see Table F.7). As for Chrome extensions, the differences are statistical significant with $p<.05$ for P, while $p<.1$ for P_{danger} and P_{dinfo}. Further, we found no correlation between the WOT's suspect rating and community rating.

New Android apps which have been regarded as spam by AppBrain.com also request for a higher number of permissions on average with P, P_{danger} and P_{dinfo} equal 3.3, 2.2, and 1.2 respectively (as compared to 3.0, 2.1 and 1.0 typically, as shown in Table F.4). The differences in the average number of P and P_{dinfo} are significant with $p<.1$.

F.5.4. Signals from the Developer

So far we have looked at aggregated signals available in the marketplace (popularity and community rating) and signals from external sources. Now we turn our

Perm. type	Android (pop)	Android (new)	FB (all)	Chrome (all)
P	6.0	3.0	3.4	1.8
P_{danger}	3.6	2.2	3.4	1.0
P_{dinfo}	2.0	1.2	2.6	1.0
# apps	5	11	88	65

Table F.7.: Average number of P, P_{danger} and P_{dinfo} requested by apps whose developer website has been labelled as *bad* or *caution* by WOT.

attention to signals that originate from the developers themselves. We considered three different signals:

Availability of a developer website: The availability of a developer website of Android apps and Chrome extensions correlates positively with the number of permissions requested by an app. Thus, the presence of a developer website (developer identity) does not imply a less intrusive app; in fact the reverse was observed. We have not measured the same effect for Facebook apps as the developer website is not shown on the user-consent permission dialogs. One may be able to obtain the developer website in the *Contact Developer* link on the app information page. However, we found that, in many cases, the link provides a means to contact the developer via Facebook's messaging system, rather than a valid developer website.

Availability of a privacy policy: The availability of a privacy policy with a Facebook app correlates negatively, albeit weakly, with P_{danger} (r=–0.12, p<.001) and P_{dinfo} (r=–0.14, p<.001). In other words, there is some weak evidence that Facebook apps accompanied by a privacy policy are more likely to request fewer permissions. Note that the privacy policy URLs were obtained from the user-consent permission dialogs. We have not looked for the privacy policy URLs of Android apps and Chrome extensions as they are not readily available.

Multiple apps from the same developer: Surprisingly, the number of apps a developer has published is negatively correlated to both $log(\#installation)$ and $avgr_a$ among Facebook apps, Chrome extensions and new Android apps. The more apps a developer publishes, the more likely his apps have a lower popularity and community rating. This could be due to that prolific developers actually make low quality apps, or that users actually cast a higher expectation on regular developers; it may be worth further investigation. The number of apps a developer makes has no correlation with the number of permissions their apps request except for the case of new Android apps, where there is a very weak link (r=–0.09, p<.01) between the number of apps the developer has made and the number of permissions. Thus, one cannot judge the potential privacy intrusiveness of an developer based on the number of apps he has published.

Perm. type	Corr. with *maturity*		Corr. with *price=free*	
	Android (pop)	Android (new)	Android (pop)	Android (new)
P	0.30	0.27	0.22	0.43
P_{danger}	0.33	0.30	0.23	0.41
P_{dinfo}	0.30	0.32	0.19	0.35

Table F.8.: Correlation between the number of P, P_{danger} and P_{dinfo} requested by Android apps, the content maturity level and whether an app is free. All values are statistically significant with $p<.001$.

Perm. type	Android (pop)		Android (new)	
	paid	free	paid	free
P	3.9	5.2	1.6	4.1
P_{danger}	2.6	3.5	1.1	2.8
P_{dinfo}	1.4	1.9	0.5	1.4
R_p (%)	86	96	56	92

Table F.9.: Average number of P, P_{danger} and P_{dinfo} requested by the free and paid Android apps. R_p measures the percentage of apps requesting at least one permission.

F.6. Enticements and Tricks

We investigated if there is any evidence of attempts to entice or mislead the user into granting sensitive permissions. In Section F.6.1 we study if free apps or those containing mature content require more privileges than average. In Section F.6.2 we study the permission request patterns of apps whose names look similar to the popular ones.

F.6.1. Free and Mature Apps

Android Market requires the developer of an app to rate its content maturity by selecting one of four labels describing the age of the target audience: everyone, low, medium or high maturity. Table F.8 shows that there is a positive correlation between the content maturity rating and the number of permissions required. There is also a positive correlation between the requested permissions and whether the app is free. Table F.9 shows the difference between paid and free apps in terms of the average number of permissions they request: free apps consistently request more permissions than paid apps. Note that also there is a bigger proportion of free apps requiring at least one permission than the paid apps; this is particularly evident among the new Android apps. Previous studies [15, 24] found that more free apps request for the `INTERNET` permission, possibly only to load advertisements. It is interesting to note that the `INTERNET` permission is not part of P_{dinfo} in our analysis. The consistently higher number of P_{dinfo} of free apps among both popular and new Android apps suggests some suspicious enticement. We further compared

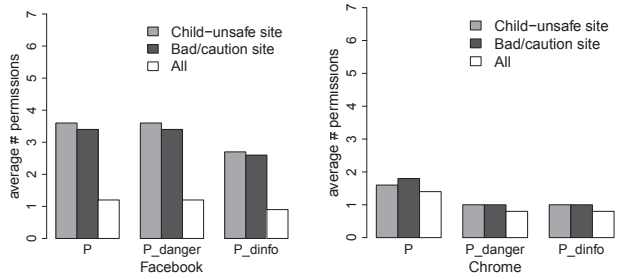


Figure F.3.: Average number of P, P_{danger} and P_{dinfo} requested by Chrome extensions and Facebook apps whose developer site has been identified as *bad/caution* or *child-unsafe* by WOT.

between free and paid apps excluding permissions that are commonly required due to third party advertisement libraries. Not counting `ACCESS_COARSE_LOCATION`, `ACCESS_FINE_LOCATION`, `ACCESS_NETWORK_STATE`, `READ_PHONE_STATE`, `WAKE_LOCK` as well as `INTERNET`, we found that free apps still request for a higher number of P, P_{danger} and P_{dinfo} on average. The differences are statistical significant with $p<.01$ in all cases (except for P_{dinfo} of popular Android apps where $p<.05$).

Facebook apps and Chrome extensions are always free. There is also no content rating systems for these. We divided these apps into three sets in terms of the WOT ratings of the developer: those labeled as child-unsafe, those labeled as suspect (*bad* or *caution*) (as we have discussed in Section F.5.3), as well as the set of all apps. The first category (child-unsafe) consisted of 34 chrome extensions and 70 Facebook apps. The second category (suspect sites) consisted of 65 chrome extensions and 88 Facebook apps (also shown in Table F.7). Figure F.3 shows the results. Suspect apps and apps with potentially child-unsafe content request more permissions than is typical. This effect is particularly pronounced in the case of Facebook apps where all differences are significant with $p<.001$. Meanwhile, the differences in the average number of P_{danger} and P_{dinfo} between the set of Chrome extensions whose developer website has been identified as child-unsafe by WOT and the set of all extensions are significant with $p<0.1$.

F.6.2. Look-Alike App Names

Apps are uniquely identifiable on the respective application platforms through unique strings, such as `bbncpldmanoknoahidbgmkgobgmhnafh` for Last.fm extension on Chrome, `102452128776` for FarmVille on Facebook, and `com.rovio.angrybirds` for Angry Birds on Android. However, unique identifiers are typically long and unintuitive. One would thus expect the users to recall or discover an app through its name or other visually distinctive features, rather than the IDs. On the application platforms, developers are free to choose his preferred app name; the app names need not be unique even on individual platforms. This creates opportunities

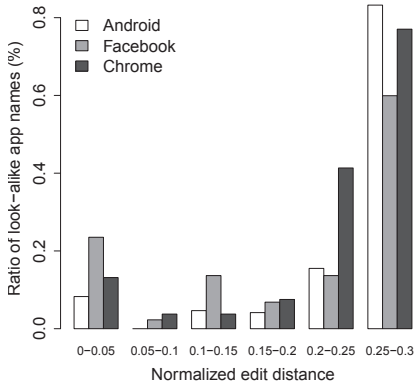


Figure F.4.: Ratio of look-alike app names in specific ranges of normalized edit distance, $dist_n$

for exploitation, for example, to create "look-alike" apps with names exactly the same or similar to the popular ones, so to free ride on their success or as a means to distribute potentially malicious apps.

We looked into the problem of name look-alike apps on Android, Chrome and Facebook. To measure name similarity, we use the popular Damerau-Levenshtein edit distance [20, 27]. Given two strings s_1 and s_2, we define the normalized edit distance as follows:

$$dist_n(s_1, s_2) = \frac{dist_{DL}(s_1, s_2)}{max(len(s_1), len(s_2))} \tag{F.2}$$

where $dist_{DL}(s_1, s_2)$ is the Damerau-Levenshtein distance between s_1 and s_2, $len(s)$ is the length of string s in terms of the number of characters, and $max(a, b)$ returns the larger value between a and b.

We outlined the top 200 most popular apps on Facebook, Chrome and Android respectively, and calculated the normalized name edit distance between them and the rest of the apps. We ignored the name pairs where both apps are published by the same developers. We also omitted apps with non-Latin based names in this study. Together, we have investigated 19,344 Android apps, 13,181 Facebook apps, and 5,322 Chrome extensions.

Figure F.4 shows the ratios of apps whose normalized edit distance to any of the popular counterparts falls in a specific range. As shown by the first three bars (i.e., $dist_n \leq 0.05$), there is a higher ratio of extremely look-alike apps on Facebook than on Android and Chrome platforms. Using a similarity threshold of $dist_n$=0.30, we found 1.15% of the Android apps, 1.20% of the Facebook apps, and 1.47% of the Chrome extensions have either intentionally or unintentionally used a look-alike name to a popular app.

Next, we manually categorized the look-alike app names into the following five classes:

	Popular App	Look-alike App
Letter change:		
A	Tap_Fish	TapFish
A	Chess Free	WChess free
F	FarmVille	SarmVille
F	PhotoMania	Pho_to Mania
C	Facebook Notifications	Facebook Notification
Term change:		
A	Advanced Task Killer	Advanced Task Manager
A	Blue Skies Live Wallpaper	Blue Wave Live Wallpaper
F	Angry birds	Angry bears
F	FarmTown	FameTown
C	Google +1 Button	Google Plus Button
Term addition/deletion:		
A	Yahoo! Mail	My Yahoo! Mail
A	Ringtone Maker	MP3 Ringtone Maker
F	Yearbook	myYearbook
C	Facebook Notifications	Facebook Chat Notification
C	Reader Plus	Reader to Plus
Serialization:		
A	Advanced Task Killer	Advanced Task Killer Pro
F	Pool Master 2	Pool Master
F	Daily Horoscope	Free Daily Horoscopes
C	Reader Plus	ReaderPlus+
C	Speed Dial 2	Speed Dial 2 (ru)

Table F.10.: Example look-alike app names of different classes on Android (A), Facebook (F) and Chrome (C). Pairs of exactly the same names are not shown.

- **Same:** Exact same name of the original counterpart
- **Letter change:** Some letters of a term of the original counterpart is changed
- **Term change:** Some terms of the original counterpart is substituted with new terms
- **Term addition/deletion:** Some terms are added into or deleted from the original counterpart
- **Serialization:** Special terms indicating a different version (e.g., *2*, *Free*) are added to the original

Table F.10 lists some examples of look-alike names we have found, while Table F.11 gives the percentage breakdown of the different classes of look-alike names. We found a total of 223, 158 and 78 look-alike names on Android, Facebook and Chrome respectively. Going through the list manually, we observed that the look-alike names in the *Same*, *Letter change* and *Serialization* classes are created with a higher level of questionable intention. We delved into look-alike apps of these three classes in the following.

Similarity class	Android	Facebook	Chrome
Same	6.7	19.6	7.7
Letter change	2.7	23.4	11.5
Term change	78.9	47.5	65.4
Term addition/deletion	4.0	5.7	9.0
Serialization	7.6	3.8	6.4

Table F.11.: Percentage (%) of look-alike names in different similarity classes.

Perm. type	Android			Chrome			Facebook		
	l-a	original	all	l-a	original	all	l-a	original	all
P	3.9	4.2	3.0***	1.8	2.3	1.4	2.1	2.5	1.2***
P_{danger}	2.6	2.7	2.1**	1.2	1.6	0.8	2.1	2.5	1.2***
P_{dinfo}	1.3	1.5	1.0**	1.2	1.6	0.8	1.5	1.9*	0.9***
$avgr_a$	2.2	7.4***	2.3	1.7	3.3***	1.2	1.8	4.2***	1.1***
$avgr$	3.8	4.5	4.4	4.1	4.2	4.4	3.9	4.0	3.8

Table F.12.: Comparing the look-alike (l-a) apps to (i) the original counterparts, and (ii) all apps as a whole. Rows 1–3 compare the average number of P, P_{danger} and P_{dinfo} requested. Rows 4–5 compare the adjusted average rating $avgr_a$, and the average rating $avgr$. ***p<.01, **p<.05, and *p<.1 indicate if a measurement given by the original counterparts, and the set of all apps, is significantly different from that of given by the look-alikes.

Table F.12 compares the characteristics of suspect look-alike apps (*Same*, *Letter change* and *Serialization*) to the set of targeted original counterparts, and the set of all apps in general. As shown on the first three rows, the average number of P, P_{danger} and P_{dinfo} requested by the look-alike apps are in general lower than the original counterparts. However, the differences are not statistically significant (except for P_{dinfo} of Facebook). While this suggests that the look-alike apps are not more privacy-intrusive than the original counterparts (i.e., the popular apps), we cannot rule out the potential risks of these look-alike apps immediately. Indeed, the average number of permissions requested by the look-alike apps are higher that is typical (i.e., comparing with the set of all apps). The increase in the average number of permissions is statistically significant for both Android and Facebook look-alike apps. This suggests some level of suspicious activities among the look-alike apps on Android and Facebook platforms.

We further analyzed how community ratings respond to look-alike apps currently. First, we found that the adjusted average rating, $avgr_a$ of the targeted counterparts is significantly higher than that of the look-alike apps across three platforms. However, we should not assume that community ratings are warning against look-alike apps adequately. The higher $avgr_a$ value of the targeted apps can be likely due to the higher number of user ratings the apps have received, following their popularity. Indeed, we found no significant differences between the average rating, $avgr$ of the look-alike apps and the targeted counterparts. Also, the average rating of the look-alike apps are not low, ranging from 3.8 to 4.1. In addition, the

adjusted average rating of the look-alike Facebook apps is significantly higher than that of all Facebook apps in general. These suggest the lack of the current community rating systems in signaling against the look-alike apps, especially on Facebook. Facebook does not even present the number of user ratings nor app popularity on the user-consent permission dialogs.

F.7. Discussion and Conclusions

Revisiting the questions we started with in Section F.1, we summarize and discuss the implications of our findings, and provide recommendations in the following:

1. Popular apps request more permissions: Dissecting the API calls of 940 apps, Felt et al. [23] found that one third of them request for *unused* permissions, attributable to errors and confusion over the insufficient API documentation. Without a source- or binary-code analysis, we have not singled out the cases where apps request for more permissions due to developer errors rather than questionable intentions. Yet, does not matter the causes (errors or bad intentions), unfortunately, there appears to be no disincentives for developers who over privilege their apps currently. There is in fact a positive correlation between app popularity and the number of permissions the app requests on all three platforms, even when considering information sensitive permissions only. More worrying is that the trend holds true despite the different UI designs and permission granularities of Facebook, Chrome and Android. Ongoing research in improving risk communication (e.g., [30]) must take into account the high permission request frequency by popular apps to be effective.

2. No reliable app risk signals currently: As users are 'trained' to accept the requests from popular apps, permission systems can become more ineffective over time. The problem is compounded by the fact that the currently available signals about an app are unreliable in indicating the privacy risks associated with that app. We investigated several such signals including the adjusted community rating, the availability of a developer website and the number of apps published by the developer. If they are to be reliable signals for helping users detect privacy intrusive apps, they should exhibit negative correlation with the number of dangerous permissions requested (and hence with potential privacy intrusiveness of the app). None of the above signals exhibit the expected negative correlation. The only exception we found is the presence of a privacy policy on the permission request screen of Facebook apps that weakly correlates with a lower number of requested permissions. However, if users start relying on this as a signal, it could lead to adverse selection as malicious developers can easily put up a 'privacy policy' that they do not adhere to.

On the other hand, we found some external services that show potential in signaling app risks. One is the website reputation scores from the Web of Trust (WOT) and another is the flagging of spam Android apps by AppBrain.com. App marketplaces can prominently display signals from similar sources to help users recognize potentially intrusive apps. Facebook is already receiving the website rep-

utation scores from WOT to protect against malicious URLs posted onto the users' wall [12]. It will not be too difficult to adapt the scores to warn against suspicious apps and developers.

3. Enticement of free and mature apps: We found evidence indicating attempts to mislead or entice users into granting permissions with free apps and mature content. The trend holds even when focusing on information relevant permissions only. Particularly, excluding the `INTERNET` permission necessary for advertisement revenue (and a few others commonly required by third party ad libraries), free apps still request for more permissions than the paid apps.

4. Look-alike name trick: We also found "look-alike" apps to request more permissions than is typical. While the fraction of look-alike apps is small, there is an underlying problem of 'cheap identity' with app names (and IDs) currently. Charging for a developer ID or for publishing an app may help, but platform owners may be reluctant to do so in the competitive market to attract developers and apps.

An option is to leverage on community inputs for reputation scores on app security and privacy. WhatApp.org [11] is a website which collates user and expert reviews on the privacy, security and openness of web and mobile apps. However, it is still in its beta version and has not attracted much reviews to date. Indeed there are many challenges in crowd-sourcing of security and privacy inputs. As we found in this paper, the number of ratings is highly correlated to the popularity of an app. This gives rise to the question of who will review suspicious apps or grayware? There is probably no one size that fits all. We see that a successful model will need to combine community inputs with automated evaluations.

Limitations and future work: One limitation of our analysis is that we have not done any source- or binary-code analysis of the apps. While we pointed out the trends that free, mature and look-alike apps request more permissions than is typical, we cannot directly infer the maliciousness of a particular app judging from the permissions it requests. Secondly, while our analysis confirms the higher risk with free and mature apps, and that there is a lack of reliable signals, we are not sure if users (in particular mature app users) are actually aware of the privacy risks and making the tradeoff willingly. Studies to examine the privacy tradeoff of users will be interesting. Leveraging on large datasets, we also plan to explore the use of machine learning methods for automatic classification of app privacy intrusiveness.

F.8. Acknowledgments

This work has benefited from initial discussions with Adrienne Porter Felt and David Barrera. We are also grateful to Adrienne and the anonymous reviewers for their valuable comments on earlier drafts.

References

[1] Android Market. https://market.android.com.

[2] Android Permissions. http://developer.android.com/reference/android/Manifest.permission.html. Last accessed: June 2012.

[3] AppBrain. http://www.appbrain.com.

[4] Facebook Permission Reference. https://developers.facebook.com/docs/reference/api/permissions/. Last accessed: June 2012.

[5] Google Chrome Permission Warnings. http://code.google.com/chrome/extensions/permission_warnings.html. Last accessed: June 2012.

[6] Google Chrome Web Store: Extensions. https://chrome.google.com/webstore?category=ext. Last accessed: June 2012.

[7] Is this App Safe – Our Project Website. http://aurora.q2s.ntnu.no/app.

[8] Socialbakers: Facebook Application Statistics. http://www.socialbakers.com/facebook-applications. Last accessed: June 2012.

[9] Watir: Web Application Testing in Ruby. http://watir.com.

[10] Web of Trust. http://www.mywot.com.

[11] WhatApp? https://whatapp.org/.

[12] Facebook partners with WOT. Article on ArcticStartup website, May 2011. http://www.arcticstartup.com/2011/05/12/facebook-partners-with-wot-to-protect-its-700-million-users. Last accessed: June 2012.

[13] A. Acquisti and R. Gross. Imagined communities: Awareness, information sharing, and privacy on the facebook. In G. Danezis and P. Golle, editors, *Privacy Enhancing Technologies*, volume 4258 of *Lecture Notes in Computer Science*, pages 36–58. Springer, 2006.

[14] D. Barrera, W. Enck, and P. C. van Oorschot. Seeding a Security-Enhancing Infrastructure for Multi-market Application Ecosystems. Technical report, Carleton University, April 2011. TR-11-06.

[15] D. Barrera, H. G. Kayacik, P. C. van Oorschot, and A. Somayaji. A methodology for empirical analysis of permission-based security models and its application to android. In E. Al-Shaer, A. D. Keromytis, and V. Shmatikov, editors, *ACM Conference on Computer and Communications Security*, pages 73–84. ACM, 2010.

[16] J. Bonneau, J. Anderson, and L. Church. Privacy suites: shared privacy for social networks. In L. F. Cranor, editor, *SOUPS*, ACM International Conference Proceeding Series. ACM, 2009.

[17] P. H. Chia, A. P. Heiner, and N. Asokan. Use of ratings from personalized communities for trustworthy application installation. In T. Aura, K. Järvinen, and K. Nyberg, editors, *NordSec*, volume 7127 of *Lecture Notes in Computer Science*, pages 71–88. Springer, 2010.

[18] P. H. Chia and S. J. Knapskog. Re-evaluating the wisdom of crowds in assessing web security. In G. Danezis, editor, *Financial Cryptography*, volume 7035 of *Lecture Notes in Computer Science*, pages 299–314. Springer, 2011.

[19] F. Cohen. Computational aspects of computer viruses. *Computers & Security*, 8(4):297–298, 1989.

[20] F. Damerau. A technique for computer detection and correction of spelling errors. *Commun. ACM*, 7(3):171–176, 1964.

[21] W. Enck, M. Ongtang, and P. D. McDaniel. On lightweight mobile phone application certification. In E. Al-Shaer, S. Jha, and A. D. Keromytis, editors, *ACM Conference on Computer and Communications Security*, pages 235–245. ACM, 2009.

[22] A. P. Felt. Personal Communication.

[23] A. P. Felt, E. Chin, S. Hanna, D. Song, and D. Wagner. Android permissions demystified. In Y. Chen, G. Danezis, and V. Shmatikov, editors, *ACM Conference on Computer and Communications Security*. ACM, 2011.

[24] A. P. Felt, K. Greenwood, and D. Wagner. The effectiveness of application permissions. In *Proceedings of the 2nd USENIX conference on Web application development*, WebApps '11. USENIX Association, 2011.

[25] J. King, A. Lampinen, and A. Smolen. Privacy: Is there an app for that? In *Proceedings of the Seventh Symposium on Usable Privacy and Security*, SOUPS '11, pages 12:1–12:20, New York, NY, USA, 2011. ACM.

[26] K. Kostiainen, E. Reshetova, J.-E. Ekberg, and N. Asokan. Old, new, borrowed, blue – A perspective on the evolution of mobile platform security architectures. In R. S. Sandhu and E. Bertino, editors, *CODASPY*, pages 13–24. ACM, 2011.

[27] V. I. Levenshtein. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, 10(8):707–710, 1966.

[28] M. Marsall. How HTML5 will kill the native app. Article on VentureBeat website, April 2011. `http://venturebeat.com/2011/04/07/how-html5-will-kill-the-native-app/`. Last accessed: June 2012.

[29] T. Moore and B. Edelman. Measuring the perpetrators and funders of typosquatting. In R. Sion, editor, *Financial Cryptography*, volume 6052 of *Lecture Notes in Computer Science*, pages 175–191. Springer, 2010.

[30] J. Tam, R. W. Reeder, and S. Schechter. I'm Allowing What? Disclosing the authority applications demand of users as a condition of installation. Technical report, Microsoft Research, 2010. MSR-TR-2010-54.

[31] D. M. Wilkinson. Strong regularities in online peer production. In L. Fortnow, J. Riedl, and T. Sandholm, editors, *ACM Conference on Electronic Commerce*, pages 302–309. ACM, 2008.