Harvesting the Low-hanging Fruits: Defending Against Automated Large-Scale Cyber-Intrusions by Focusing on the Vulnerable Population

Hassan Halawa
University of British Columbia
Vancouver, Canada

Konstantin Beznosov
University of British Columbia
Vancouver, Canada

Yazan Boshmaf
Qatar Computing Research Institute
Doha, Qatar

Baris Coskun
Yahoo! Research
New York, US

Matei Ripeanu
University of British Columbia
Vancouver, Canada

Elizeu Santos-Neto
Google, Inc.
Zürich, Switzerland

ABSTRACT

The orthodox paradigm to defend against automated social-engineering attacks in large-scale socio-technical systems is reactive and victim-agnostic. Defenses generally focus on identifying the attacks/attackers (e.g., phishing emails, social-bot infiltrations, malware offered for download). To change the status quo, we propose to identify, even if imperfectly, the *vulnerable* user population, that is, the users that are likely to fall victim to such attacks. Once identified, information about the vulnerable population can be used in two ways. First, the vulnerable population can be influenced by the defender through several means including: education, specialized user experience, extra protection layers and watchdogs. In the same vein, information about the vulnerable population can ultimately be used to fine-tune and reprioritize defense mechanisms to offer differentiated protection, possibly at the cost of additional friction generated by the defense mechanism. Secondly, information about the user population can be used to identify an attack (or compromised users) based on differences between the general and the vulnerable population. This paper considers the implications of the proposed paradigm on existing defenses in three areas (phishing of user credentials, malware distribution and socialbot infiltration) and discusses how using knowledge of the vulnerable population can enable more robust defenses.

Keywords

Vulnerable population; cyber intrusions; defense system design

1. INTRODUCTION

Social engineering is one of the key attack vectors faced by large socio-technical systems such as Facebook and Google [13, 36, 37]. To carry out further attacks, cyber criminals use social engineering to exploit unsafe decisions by individual users by tricking them into providing credentials to phishing websites [37], accepting friendship requests from socialbots [12], or downloading malware [91]. Such, largely automated, attacks are increasing in frequency, scale, and sophistication [56, 71]. As a case in point, one of such attacks, phishing (i.e., a social engineering attack using fraudulent emails/websites to trick users into leaking account credentials), causes sizeable financial losses: $1.6 billion only in 2013 [66]. A recent study [67] highlights the magnitude of the problem: about half of the 1,000 participants actively clicked on a phishing link, and half of those actually leaked information to phishing websites, becoming victims of the attack.

State of the art defenses against cyber intrusion in socio-technical systems are mostly reactive and victim-agnostic. In general, these defenses do not attempt to harness user characteristics, and instead focus on identifying attack entry points, such as phishing emails or friend requests from socialbots, based on structural, contextual, or behavioral attributes of the attack(er) [22][1]. The commonly used defense mechanisms, presented in Figure 1 from a high-level design standpoint, typically employ *operator filters* to automatically detect known attacks or anomalies. However, as attacks do evolve and due to the need to minimize false positives, a considerable number of attacks pass through this filter and reach users (e.g., a phishing email ending up in the user's inbox).

Once the threat is exposed at the user level, it is then up to the user to decide how to respond (e.g., by flagging a phishing email as malicious, ignoring it, or following the phishing link). This manual vetting process can be thought of as a second filtering level, marked as a *user filter* in Figure 1. Attacks that bypass this filter turn the user into a *victim*: A user whose assets, including account credentials, private data, or personal devices, are compromised.

Usually, the system operator eventually detects some of the compromised assets, either prompted by the victim, by other users, or based on the abnormal behavior detected by an automated incident-response system. Shortly after, the operator launches a remediation process, which is illustrated as the *remediation filter* in Figure 1. After the compromise is

[1] Defenses based on anomaly detection also fit here [41].

NSPW '16 September 26-29, 2016, Granby, CO, USA

ACM ISBN 978-1-4503-4813-3/16/09.

DOI: http://dx.doi.org/10.1145/3011883.3011885

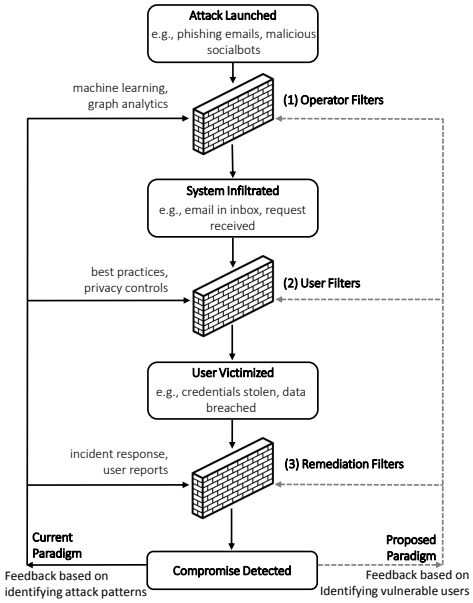

Figure 1: High-level schematic presenting the structure of existing defenses against mass-scale attacks (left-side feedback loop) and the proposed paradigm (right-side): augmenting the feedback loop with information about vulnerable users to improve the operator filter (1), users' own decision making process (2), and the remediation filter (3).

cleared, the resulting attack analysis concerning the involved assets and attack pattern are generally fed back into the operator and remediation filters to detect future, similar attacks and other compromised users.

The reactive and victim-agnostic nature of this defense paradigm [9, 19, 21, 26, 52, 53, 62, 63, 85, 89, 94] gives attackers the opportunity to evade detection (by adjusting their tactics), circumvent the employed defenses, and still reach the end-users.

We advocate a complementary approach: a victim-centric defense. We propose to detect the *vulnerable population* (that is, the users that are likely to become victims, e.g., users likely to fall prey to a phishing attack) and to thwart attacks by: focusing defense efforts on the vulnerable population and, at the same time, using the information about the vulnerable population to design more robust defenses. We postulate that the information about the vulnerable population can be used to build proactive defenses (the defender identifies the vulnerable users, and subsequently feeds this information back into the defense subcomponents to detect new types of attacks that were not previously observed) or can be used to increase the robustness of existing reactive defenses.

The feedback loop on the right side of Figure 1 summarizes the layers where the information about the vulnerable population can be used. Specifically, this information can be harnessed to: (1) establish more robust attack detection at the operator-level filter (e.g., by augmenting the operator-level filters with a per-user vulnerability score, or by inferring the likelihood of a request being malicious based on collected statistics on how vulnerable and robust users reacted to similar requests in the past), (2) aid vulnerable users in making better decisions (e.g., through user education, personalized interfaces, enforced hardware/software configuration/restrictions, and/or better understanding of the decision making process of different categories of users), and (3) achieve faster and more accurate compromise detection and remediation for attacks that were not previously observed (the intuition is that, a user's vulnerability score can be used by the defender to prioritize monitoring and to make more accurate decisions on whether an asset is compromised when abnormal behaviour is observed).

Our contribution. Our proposal draws insights from multiple areas (including usable security, user-aware security mechanisms, public health), and is informed by the current mass-scale attacks in large socio-technical systems and by existing defenses, some of which do explicitly focus on the vulnerable population. However, in contrast with existing ad-hoc, system-specific solutions we approach harnessing the vulnerable population under a single, unified framework that encompasses diverse application areas. To the best of our knowledge, we are the first to propose and discuss such a framework as a complementary security paradigm.

Note that we do not aim to present in detail concrete defense mechanisms and algorithms that leverage the vulnerable population. Our goal is to put forward a new security paradigm which opens new research opportunities in thwarting cyber intrusions in socio-technical systems. For a specific instantiation of this paradigm and a detailed description of defense mechanisms and algorithms we point the reader to our recent work on social-bot infiltration detection [10]. Furthermore, we note that targeted attacks (e.g., spear phishing attacks) are beyond the scope of the proposed paradigm.

The structure of this paper. To illustrate the potential of this paradigm, we develop our ideas in more detail in the context of three specific application areas: mitigating phishing-based credential theft (§2.1), malware distribution (§2.2), and socialbot infiltration (§2.3). For each of these application areas, we discuss where current defense mechanisms fail, how identifying the vulnerable population could proceed, and how information about vulnerable population can help to build more resilient defense systems. We then discuss (§3) topics related to the feasibility of identifying the vulnerable population, factors influencing its accuracy, and possible issues related to adopting this new paradigm.

2. COMPARING THE CURRENT AND THE PROPOSED PARADIGM

To highlight how information on the vulnerable population can be leveraged we focus on mitigating cyber intrusion attacks in three different domains. For each domain: we first contrast current approaches with our proposed strategy informed by inferring the vulnerable population, we then sketch how the vulnerable population can be identified, and, finally, we present how the filters illustrated in Figure 1 can be made more efficient by using information about the vulnerable population.

Terminology. We refer to users who have fallen prey to social engineering attacks and had their assets compromised as *victims* [2]. We refer to users who are more susceptible

[2] Note that it is the defender, and not the user, who defines what a

to fall prey to such attacks in the future as the *vulnerable* population [3]. We refer to users that are less susceptible to fall prey to attacks as *robust* users [4].

2.1 Mitigating the Threat of Phishing for User Credentials

Online user credentials, especially those provided by email or social networking services, are becoming increasingly valuable. This is the case because those credentials are being used as a single sign-on identity for many other online services. As a result, large online service operators have invested heavily in the development of defense systems to mitigate credential theft, especially credential theft carried out through phishing [23].

Stolen credentials, or compromised accounts, do not only impact their owners but the whole socio-technical system. For example, compromised email accounts are used to spread spam and phishing emails, as they are likely to have better reputation (than newly created fake accounts) and to bypass spam filters [72]. This is also true for social networking accounts [80].

The Current Approach. At each filtering level (illustrated in Figure 1), various strategies have been investigated. At the operator-filter level, the proposed strategies include: content-based classification [89, 95], anomaly-based classification [30] and URL character frequency analysis [85]. At the user-filter level, some of the proposed strategies include: user education [68], visual page similarity [18], and offensive defense [17]. Finally, at the remediation-filter level, one proposed strategy is crowdsourcing [45]. None of these approaches uses information about the vulnerable population.

The Proposed Paradigm. Instead of focusing *only* on identifying attack patterns, we argue below that using information about potential victims in the decision processes employed by each filtering level will enable a more robust defense. For example, a classifier using the email sender's vulnerability score as a feature would likely yield better detection accuracy (i.e., allowing for the detection of epidemic attacks originating from the victim population). Such an approach would also support the development of more efficient threat detection and remediation strategies by prioritizing efforts on the vulnerable population [5] as we outline for each filtering level below.

Identifying the Vulnerable Population. Leveraging the large set of temporal, structural, and behavioral features collected by online service operators, vulnerable users can be identified by training a machine learning classifier on labeled ground truth of known victims. With a reliable ground truth and a large set of features, the trained classifier will generate a vulnerability score to predict the likelihood that a user will make an incorrect security decision (in the context of a phishing attack). This score can be used to inform all defense filters discussed in Section 1 as we outline below.

Operator filters. On the one side, given information about potential victims, various defensive strategies inspired by the public health field can be employed during attack "outbreaks" (i.e., large-scale attack campaigns) to "quarantine" the population-at-risk. One approach can be to throttle down the sending rate of messages proportional to a user's vulnerability score during an "outbreak" (the intuition is to distribute the cost of an emergency measure proportionally with perceived user risk rather than indiscriminately). Another, complementary, approach can be to delay the delivery of potential phishing messages to vulnerable users; once those messages are better understood, possibly based on the reaction of robust users [6] to similar messages, these messages may get delivered to vulnerable users as well. Thus, operator-level attack filters employ knowledge about vulnerable users to mitigate the spread of attacks within its internal user population.

On the other side, vulnerability scores can be used to make attack detection more accurate. For example, vulnerability scores can be used to inform the operator-level filters in case of epidemic attacks (e.g., attacks where the email account of the victim is compromised and used for sending out phishing/spam emails): suspicious outbound traffic (e.g., phishing/spam emails) that is identified across multiple users with high vulnerability scores may be seen as a signal of an attack outbreak. The defense system can then take steps to prevent such attacks from spreading to the rest of the user population.

User filters. Secondly, an anti-phishing defense system can leverage users' vulnerability scores to provide personalized security advice and controls. A personalized approach to security can not only better inform users' decision making process but also improve the usability of the service as well as the overall user experience [24]. At the same time, through the early identification of vulnerable users, the service operator can take proactive measures to "immunize" them. For example, the service operator can carry out user education campaigns that are only targeted to potential vulnerable users (e.g., offering embedded training [42, 43]). Such a focused approach, on the one side, reduces the cost of the education campaign by only targeting at-risk users who are most likely to benefit, and, on the other side, decreases the overall friction for the rest of the user population by excluding those who are not likely to fall victim to such attacks in the first place.

Remediation filters. Finally, we postulate that the vulnerability score, i.e., the likelihood that a user would fall victim to a phishing attack, can also be used to enable faster remediation. First of all, using such a score as a feature would substantially improve the accuracy of a classifier trained to identify potentially malicious login attempts using phished credentials. Second, service operators can

victim is.

[3] While the definition of a victim is deterministic, the definition of a vulnerable user is probabilistic, in the sense that a vulnerability 'score' will correlate with the probability that the user becomes a victim if attacked. A user's vulnerability will change over time as the user gains expertise with using a system, and the uncertainty associated with a vulnerability score will decrease as more data about the users' behaviour is gathered. We envisage that ground truth data is collected continuously, and the prediction models are periodically retrained.

[4] We consider a user *robust* if her vulnerability score is sufficiently low. As the score is inferred from user behaviour over a long period of time, we hypothesize that users who are able to identify potential risks, e.g., suspicious messages or friendship requests from suspicious users, are more aware of new threats as well and can be valuable assets in designing defense strategies.

[5] Based on our past experience, we hypothesize that the vulnerable users represent a fraction of the overall user population. As such, prioritizing and focusing defensive efforts on this subset of users would allow a more efficient allocation of defense resources.

[6] We note that, in the context of phishing attacks, identifying robust users and relying on them to label phishing emails has some similarities with existing proposals to crowdsource phishing/spam email labeling [45, 46].

crowdsource the detection of evolving threats to the robust portion of the user population. The defense system would leverage their feedback (i.e., user reported threats) to more quickly initiate remediation and to prevent attack propagation. In more details, an email provider could potentially introduce a delivery delay to bulk email (which may or may not be malicious) addressed to the vulnerable subset of the population [83]. The defense system can then rely on feedback and user reports from those who receive the unverified bulk email earlier (since they are less likely to fall victim to attacks) and take action accordingly. A third avenue is that service operators could potentially launch regular compromise detection campaigns (either through manual sampling or through more automated means) on the small subset of users that are most likely to fall victim to account compromise. Doing so is much more efficient in terms of resources since it is directed to the much smaller subset of the overall population that is most likely to benefit from such extra scrutiny.

2.2 Mitigating the Threat of Malware Distribution

Malware downloads (e.g., viruses, trojans and other malicious applications) have been a widespread vector of cyber-attacks. Both traditional Internet Service Providers (ISPs) and wireless carriers, have invested in the development of defense systems that aim to protect their customers and their networks from those threats. However, despite the widespread use of host- and network-based malware detection systems, the malware threat is constantly expanding; security analysts have identified 317 million new malware variants in 2014 (up from 252 million new variants identified in 2013) [79]. One of the most prominent malware spreading strategies is social engineering, where users are tricked into clicking on malicious URLs or installing malicious applications.

The Current Approach. At each filtering level (illustrated in Figure 1), various strategies have been investigated in the literature. At the operator-filter level, the proposed strategies include: sandboxing [51], blacklists [31], vulnerability-based detection [38, 87], signature-based detection [78], anomaly-based detection [69] and machine learning [61]. Note that, in spite of major progress, the rate of false positives and the runtime overheads of the detection filters put pressure on the accuracy that can be achieved. As a result, in practice, the user is still exposed to making decisions over whether to download and run malicious software. At the user-filter level, a common strategy employed by most major web browsers is to require user confirmation before running downloaded executables. Finally, at the remediation-filter level, some of the proposed strategies include: honeypots [70, 88] and honeyclients [58].

The Proposed Paradigm. By focusing defensive efforts on the subset of the population that is most vulnerable, the proposed victim-centric defense paradigm has the potential to enable the use of advanced malware defense systems in the context of real-time analysis of high-volume Internet traffic. Additionally, this approach would enable a defense that is more robust to adversarial strategies as opposed to traditional approaches which can be circumvented using basic evasion techniques.

Identifying the vulnerable population. ISPs have deep visibility into all network operations and are thus uniquely positioned not only to identify users that are potentially vulnerable to malware threats but also to provide network-based real-time defense. Based on labeled ground truth collected by the ISPs and leveraging the large set of available temporal, behavioral and network flow features, a classifier can be trained to identify users that are likely to fall victim to malware downloads. Corresponding vulnerability scores can be computed for those vulnerable users to inform the attack filters illustrated in Figure 1 as illustrated below.

Operator filter. First, ISPs can employ graph analysis techniques (with nodes representing hosts and edges representing traffic flow between hosts) to identify hosts that exhibit spatial or temporal traffic correlation with the vulnerable subset of the user population. These hosts are more likely to be attacker-controlled (e.g., compromised servers or C&C servers) and can be subsequently investigated and potentially blacklisted at the operator-level filter.

User filter. Secondly, by identifying the subset of the user population with the highest likelihood of falling victim to malware threats, ISPs can take preventative measures to reinforce user-level defenses. ISPs can make use of captive portals to provide targeted educational and remediation material only to the segment of the user population that is most at risk. The captive web portal can include basic training material, personalized security advice as well as security software downloads (e.g., host-based anti-malware software) from trusted internal ISP-controlled servers. Such an approach would minimize both the amount of effort and resources that must be expended by the ISPs as well as the amount of friction experienced by the overall population. Moreover, ISP resources such as Intrusion Detection Systems (IDSs) can be more effectively used by prioritizing and focusing on traffic from/to the subset of the user population that is most at-risk.

Remediation filter. Finally, information regarding the likelihood of users falling victim to malware attacks can be employed by ISPs for faster compromise detection and remediation. The segment of the user population that are most likely to fall victim to malware threats could potentially be treated by ISPs as if they were honeypots. ISPs can place such "honeypots" under much closer scrutiny than typical users in order to detect evolving threats. Moreover, ISPs can routinely run inspection campaigns on the vulnerable subset of the user population either through manual sampling or through more automated methods. This would allow ISPs to more efficiently utilize their available security resources by focusing on the segment of the user population that would benefit the most from such inspections.

2.3 Mitigating the Threat of Socialbot Infiltration

From a security perspective, a *socialbot* is an automated user account in a target online social network (OSN) that is created and controlled by an attacker for various adversarial objectives [10], including social spamming [80], political astroturfing [65], and private data collection [12]. To accomplish these objectives, an attacker has to connect the socialbots with real user accounts through a social *infiltration* campaign, as isolated socialbots cannot freely interact with or promote content to users in the target OSN [8, 12, 57, 75]. Making these connections is possible

because users can be tricked into accepting friend requests sent by socialbots, especially when the users share mutual friends with the socialbots [11]. We refer to users that have accepted friend request from socialbots as *victims*.

The Current Approach. This multifaceted threat has spawned a line of research with the goal of designing defense mechanisms to thwart socialbots. Thus far, most of the work in this area is at the level of the operator-filter (Figure 1). On the one hand, some approaches rely on mining the users' activity patterns or to control their account creation proactively [16, 40, 75, 90]. While these techniques are effective against naive attack strategies, many studies have shown that, in practice, they introduce usability issues and can be easily evaded [3, 13, 40, 86]. On the other hand, while graph-based detection promises desirable security properties, it hinges on the assumption that socialbots cannot befriend many real accounts. Past work, however, showed that it is sufficient for each socialbot to befriend a single victim in order to evade detection [3, 92]. At the user-filter level, OSN operators constantly tweak the user experience concerning friendship requests to better inform user decisions. Finally, at the remediation-filter level, one proposed strategy is to use honeypot accounts [77].

The Proposed Paradigm. Defending OSNs against socialbot infiltration is another problem domain where predicting the vulnerable population can inform stronger defenses. We offer three key observations. First, since victim accounts are real accounts that are not controlled by attackers, identifying potential victims is inherently more robust against adversarial attacks than identifying socialbots [7, 50, 84]. Second, since the vulnerable population represents a small fraction of all OSN accounts [8, 12], restrictive admission control mechanisms can be applied to only those accounts (and their connections), without limiting the experience of others. Third and last, as socialbots are directly connected to victims, the accuracy of socialbot detection mechanisms can be improved by using victim prediction as a feature that is hard to manipulate by the adversary [3, 7, 11].

Identifying the vulnerable population. OSN operators have access to all user information which forms a large set of temporal, structural and behavioral features. Historical data can be used to generate labeled ground truth that can be used to train a machine learning classifier able to identify the vulnerable users. We have already investigated such an approach [10]: even with a small set of strictly low-cost features, one can train a classifier that is 52% better than random in predicting the vulnerable users.

Operator filters. First, one way to retrofit existing graph-based socialbot detection techniques to improve operator-level filters is to artificially "prune" friendships edges adjacent to the vulnerable accounts. As victims are located at the borderline between the two subgraphs separating the socialbots from real accounts, an OSN operator can reduce the number of edges crossing this boundary by incorporating victim prediction into the detection process. This improvement can be achieved by assigning each edge that is incident to a vulnerable account a significantly smaller weight than to others. We have investigated this approach in the context of the Facebook and Tuenti OSNs where our proposed defense system, Íntegro, employing information about the vulnerable population was shown to significantly outperform other state-of-the-art techniques [10].

User filters. Secondly, once identified, the vulnerable population can be influenced by the OSN operator through education, personalized user experience, extra restrictions or watchdogs. The operator can use the vulnerability score to focus only on the most vulnerable users, thus relieving the rest of the population from the associated effort. Also, the user experience in the process of considering friendship requests can be altered dynamically for vulnerable users. For example, the operator might offer additional information [64] on the account from which the request is coming. Alternatively, friendship requests to the vulnerable users might be delayed to throttle infiltration attacks.

Remediation filters. Finally, the vulnerable population can be used by OSN operators as "honeypots" as described for the prior problem domains. Additionally, OSN operators can create controlled honeypot accounts that can be used to sample the activities of user accounts, in particular, those who contact these honeypots by sending them friend requests or by sharing content [77]. The sampled activities can then be analyzed and used to maintain an up-to-date ground truth for socialbot detection. Once a socialbot is detected, remediation can be achieved by deleting/disabling the corresponding OSN account or by warning its friends.

3. DISCUSSION

In this section, we explore a number of interrelated issues:

Does a differentiated vulnerable population actually exist? Can this vulnerable population be identified? A long stream of past research in offline worlds has shown that users have different likelihoods of making incorrect (security, economic, or health-related) decisions, and more importantly, it suggests that it is feasible to identify categories of vulnerable users [27, 44, 59]. Once identified, preventative actions can then be directly focused on the vulnerable population that is most at risk [49]. This research direction, however, has seen limited uptake in the online world of socio-technical systems, particularly in the context of large-scale attacks.

In the context of large socio-technical systems, approaches that aim to predict the likelihood of future compromises have emerged only recently. We have shown (Boshmaf et al. [10]) that, using only a small set of relatively cheap features, it is possible to predict the users vulnerable to a socialbot infiltration campaign in Tuenti and Facebook. Soska et al. [74] used traffic-based and content-based features extracted from web pages, and trained a classifier able to predict the likelihood of a web site being compromised and becoming malicious in the future. The list of features employed by the proposed classifier include: website traffic statistics (as a measure of the popularity of the website), website file system structure as well as web page structure and content (to determine the type of used content-based management system). The extracted features are dynamically updated over time to keep up with evolving threats. More recently, Liu et al. [47] applied a similar approach in order to forecast cyber-security incidents: using only externally measurable features representing an organization's network security posture (i.e., vigilance and preparedness) they are able to estimate the likelihood that the organization later becomes victim of a successful attack. The aforementioned features were demonstrated to be good predictors in line with the expectation that organizations

with lax security policies and processes are much more likely to fall victim to attacks (i.e., organizations with increased vigilance and preparedness as measured by the employed external features are less likely to fall victim to attacks). On a different take, Thonnard et al. [82] considered targeted attacks carried out on organizations analogous to a public health issue and used epidemiological techniques to rank the risk factors that can be used to proactively identify the vulnerable organizations.

While these past experiences indicate that predicting vulnerable users may be feasible; none of those approaches, however, goes further and develops ways to leverage the knowledge of the vulnerable population to improve system-wide or user-level defenses.

Why an approach focused on the vulnerable population is a key defense element in some situations? For some attacks, the dynamics of large-scale automated intrusions are similar to those of epidemics. The reason for this similarity is the following: for those attacks, becoming an attack victim (e.g., a compromised user account) is not only a cost to the user herself (e.g., through potential loss of data and compromised privacy) and to the system operator (as the recovery procedures generally involve manual operations and are thus costly) but, importantly, a cost to the entire user community as well. The reason is that the assets of a victim (user account, identity, device) are often used as stepping stones in multi-stage attacks, or as a platform for extending the attack on the remaining users (e.g., sending additional phishing emails, launching friendship requests on online social networks) [81]. Moreover, in some cases (e.g., for socialbot infiltration, phishing) the presence of attack victims makes it more difficult for the remaining users to make correct security decisions. Thus, these large-scale attacks have an epidemic factor, wherein victims are a factor that helps with spreading the infection [93]. For these cases, research in public health suggests that focusing on the vulnerable population is a key element to limiting the spread of, and controlling the cost of an epidemic [60].

Why does an approach which includes information about the vulnerable population have the potential to increase the robustness of existing defenses (even when epidemic dynamics are not at play)? Current defense techniques are predicated on detecting the attack actions (e.g., phishing emails) based on structural, contextual or behavioral attributes of the attack/attacker itself. The main problem with such attacker-centric techniques is that they generally follow a reactive "first-detect-then-prevent" approach. This makes it possible for motivated attackers to employ adversarial strategies (i.e., modify their attack patterns) to, often trivially, evade detection by the employed defense systems. A defense that incorporates information about the vulnerable population has the potential to be more robust as, unlike current defense techniques that attempt to detect behavior that is under the direct control of the attacker (e.g., frequency of sending emails to detect likely compromised email accounts used to send spam), it incorporates features that are intrinsic to the vulnerable users and not under the control of the attacker.

Can the proposed approach improve the effectiveness of user education / security advice? User education / security advice are the first line of defense against increasingly sophisticated social engineering attacks [37, 76]. While many studies show that users tend to reject security advice because of low motivation and poor understanding of involved threats [2, 54], others assert that users do so because it is entirely rational from an economic viewpoint [28, 33]. In particular, the advice offers to protect the users from the direct costs of attacks, but burdens the whole user population with increased indirect costs in the form of effort. When the security advice is applied to all users, it becomes a daily burden whose benefit is the potential saving of direct costs only to the vulnerable population. When this population is small, it becomes hard to design a security advice with positive net benefit.

One way to increase the benefit of security advice is to make it more usable, which in effect reduces its indirect costs to all users. This has been the focus of a growing community of usable security researchers who consider user education essential to securing socio-technical systems such as OSNs [6, 73]. A complementary way to reduce indirect costs is to engage with the security advice to only the fraction who might actually benefit from it. This approach is directly supported by identifying the vulnerable population.

Are the vulnerable users the "enemy"? There has been a growing body of work which shows that users, whether *vulnerable* or not, make rational choices that attempt to optimize their individual cost-benefit trade-offs (e.g., password complexity vs. ease of memorizing) [28, 33**?**]. We believe this is true for the vulnerable members of the user population as well. We acknowledge that some of our proposed defense measures can be seen as penalizing vulnerable users (and can be misconstrued as treating vulnerable users as the "enemy"). We suggest a shift in perspective. Instead of focusing on the negative aspects (i.e., possible lower quality of service offered to vulnerable members of the population such as delayed/throttled email) of the defense measures, higher quality service could be marketed as a reward offered only to robust users. Another example would be an Internet Service Provider (ISP) offering a discount to subscribers (i.e., lower subscription fees) to robust users (i.e., passing the decreased costs onto the subscribers). This shift in perspective could potentially create an incentive for non-robust users to invest the required effort (e.g, training) to become robust themselves in the future.

Are there other problem domains that can benefit from the proposed approach focusing on the vulnerable population? We believe that leveraging information about the vulnerable members of the user population can be used to improve upon existing attacker-focused defense systems in a variety of domains. In this paper, we focused on three such domains: online credential theft through phishing, malware downloads and socialbot infiltrations in online social networks. We believe other socio-technical systems, where an attack based on influencing users to make incorrect security decisions can be automated, can benefit from our proposed paradigm. One such problem domain is enterprise security and risk management where organizations can employ information about potential victims (e.g., users or assets) as part of building a risk-profile for more accurate modeling and, consequently, more effective management of risk. Moreover, in regard to password composition and complexity policies

for users, enterprises can benefit from focusing on the most *vulnerable* users (i.e., those with the most easy to guess passwords) instead of trying to get *all* users to improve the strength of their passwords [29]. At the same time we note that targeted attacks (e.g., spear phishing attacks) are not in the scope of the proposed paradigm.

Are there legal/ethical implications of the proposed solution? One issue is "paternalism": Our system is paternalistic in the sense that we try to "nudge" and sometimes even force the users into the decisions we believe are better for them. There is a wealth of work on ethical issues related to paternalism. In the field of public health, for example, health education campaigns are treated as potentially paternalistic but it is considered to be the responsibility of professionals to justify when and where such paternalism is justified [39]. In case of conflict with patient choices, however, "the patient's informed choices and definition of best interests should prevail" [48].

A second issue is fairness - in some of the solutions we propose users are categorized based on their vulnerability and served differently. Experience from various consumer domains, as well as border control and airport security, indicate that society has already adopted many systems where people are categorized and served differently. We point out as well that sensitivity to the fairness issue is context specific. Some of the contexts where we imagine the new paradigm will get adopted involve free services and are provided by a for-profit operator under no formal obligation to provide fair service (e.g., Google mail). A further point, is that for the cases where a victim is a potential cost for an entire community (see the "epidemic effects" discussion above), it is already accepted in society that the loss of fairness is a possible price to pay to balance risk (for example, mandatory quarantine of travelers that have symptoms of an infectious disease).

A third issue is unintentional bias - either inherent in the data used for training or unintentionally introduced by the learning algorithms themselves [5]. One the one hand, inherent bias in the data used for training could lead to certain segments of the population being penalized, or worse quarantined, due to the use of correlative (i.e., non-causal) features such as age and/or gender. For example, senior citizens could potentially be labeled as more vulnerable to certain types of attacks (e.g., phishing) by using age as a feature. On the other hand, the learning algorithms themselves could introduce their own unintended bias. An example of such algorithmic bias is how online search engines are biasing queries in the process of offering *personalized* results [20]. In such scenarios, where should the line be drawn between the use of a priori vulnerability signals that are correlative (e.g., age and/or gender) as features and the net effect on the affected segments of the population? For instance, racial discrimination as an unintended consequence of such biases has been receiving widespread coverage in the press [4, 35, 55]. We believe that more work is needed to investigate how to detect and eliminate such hidden biases as well as their legal and ethical implications.

What are some of the challenges that may prevent adopting this paradigm? There are a number of challenges that we foresee ahead:

- *Feasibility to develop a vulnerable population classifier.* The success of each instantiation of this paradigm is predicated on the feasibility of predicting user vulnerability based on observed online behaviour. In some contexts, it may be infeasible to store or process the features required to achieve good prediction performance due to legal, privacy or operational constraints. In other contexts, it may be difficult or costly to procure ground truth of sufficient quality to train the vulnerable population classifiers.

- *Handling new user sign ups.* As previously mentioned, vulnerability scores are typically dynamically computed for users based on the history of their past online behavior and actions. This approach becomes problematic in the case of new user sign ups for which insufficient history is available. To solve this problem, on the one hand, new users can be automatically assigned a high vulnerability score upon sign up. This approach allows the defense system to apply the highest level of protection to those users until enough information about their behavior is collected to generate representative vulnerability scores. On the other hand, external login providers (e.g., Facebook Login [25] or Google OpenID Connect [32]) can be used as a means to prime the system with information about the new sign ups. This approach could generate vulnerability scores for those new users by collecting additional information from the external provider (e.g., profile information).

- *Inaccuracies in predicting the vulnerable population.* All situations where this paradigm is adopted will need to be robust to (or, at least, have a low cost due to) inaccuracies in predicting the vulnerable population. These inaccuracies are inherent to the classifiers built based on past user behaviour and may also be a result of inaccurate ground truth labeling. We make two additional comments: on the one hand, for some application domains, it is possible, and even standard practice, for the service operator to carry out exercises that simulate common attacks in order to collect accurate ground truth [1]. On the other hand, our experience with detecting socialbot infiltrations shows that predicting the vulnerable population, even if imperfectly, enabled better socialbot detection than state of the art systems. We believe that many of the other techniques we propose here have similar properties.

- *Feature Selection.* As part of training the victim classifier, careful feature engineering and selection is necessary not only to improve the classifier's performance (e.g., its prediction accuracy) but also to respond to evolving user behaviors and new attacks. For example, picking features that are ancillary to actual security decisions, such as correlative but non-causal features (e.g., slow typing speed), could potentially lead to an overall poor performance for the trained victim classifier. Even if a classifier based on such features performs well initially, the problem could worsen over time as users figure out what features are being used to score their behavior and attempt to manipulate those features (e.g., typing faster).

We believe an iterative approach where the victim classifier is evolved over time (i.e., feature engineering is routinely improved) would solve the aforementioned problems. The victim classifier can then be re-trained to improve overall performance as well as to adapt to emerging user behaviors.

- *Some of the mitigation techniques we propose are based on delays and may violate users' expectations.* While many online tasks are by design non-interactive (e.g., email), users have grown accustomed over the years to consider emails as interactive and instantaneous (similar to instant messaging). For example, delaying a recovery email may be noticed, and will reduce usability, delaying a friend request or Skype add contact request may be frustrating. While obviously there are contexts where delaying techniques can not be used (e.g., true instant messaging), we believe that a careful, context-specific calibration of the delay coupled with allocating proper computational power to gather and process information can make this technique useful in many contexts in which mass scale attacks are launched.
- *Targeting some, but not all, of the population for various protection measures may lead to potential confusion and additional complexity - both at the individual user and at the system operator level.* Indeed. The problem is complex even when focusing on a specific domain: whose benefits should outweigh whose costs? Should the user's cost (e.g., incoming e-mail being delayed) be out-weighted by the benefit to the service provider, who can spend less money on defending against a particular attack? How can you compare these costs with each other? Transparency can be used to reduce stakeholders' confusion. As an example, Gmail used to mark spam without providing any explanation. Now, if you go into your Gmail spam folder and click on a message there, Gmail will explain the reason(s) why that particular message was flagged as spam.

What are the categories of defenses enabled by adopting this paradigm? We discuss several defense categories that, we believe, are enabled by adopting our proposed approach. These include:

- *Targeted Protection.* Our proposed paradigm is based on the idea of segmenting the user population by means of analyzing their online behavior. This enables a targeted approach to protection where the application of security measures can vary across the user population. For example, email sent out to potentially vulnerable members of the population can be placed under additional scrutiny without burdening the majority of the user population who may be more robust to such types of attacks.
- *Inferring the Origin of the Attacks.* Moreover our proposed paradigm, focusing on victims, can be used to infer additional information about ongoing attacks. The anomalous behaviors and interactions exhibited by victims would demonstrate clear deviations from the norm and thus allude to the origin of attacks. For example, malware distribution servers (or botnet C&C servers) can be identified through their pronounced interactions with the subset of the user population that have fallen victim to their attacks.
- *Context-Specific Protection.* One potential avenue to explore, as part of our proposed paradigm, is *context-specific* user vulnerability. On the one hand, vulnerability scores could be based on users' *long-term* behavior/actions which are a reflection of the characteristics that are inherent to those users (e.g., bad security practices). We've discussed how this approach would allow defense systems to offer efficient targeted protection to the vulnerable subset of the population. On the other hand, even robust users can experience increased vulnerability to attacks in certain contexts which temporarily affect their *short-term* behavior/actions. For example, a severely jet-lagged security professional might be more vulnerable to clicking on a fraudulent hyperlink and then leaking important online credentials or falling victim to drive-by malware download. This is analogous to how a player's skill rating changes in competitive gaming over time and how it may even be affected by temporary conditions (e.g., breaks due to boredom) [34]. We postulate that our proposed paradigm can be adapted to offer both targeted protection as well as context-specific protection through the use of long-term and short-term user vulnerability scores respectively.

Is the proposed paradigm predicated on having a steady stream of victims to train the system? What happens when there are no more victims? Our proposed paradigm does not aim to completely replace existing reactive defense approaches based on detecting attack/attacker patterns. We instead aim to augment existing systems with per user vulnerability scores in order to improve the overall defense. If we manage to reach a situation where there are no more victims in the system (e.g., by educating and training all users so that they become robust) then we would have achieved our goal of improving the defense for the overall population. We can simply fallback to current reactive approaches with the knowledge that there are no more low-hanging fruits for the attackers to exploit.

What is the potential risk in case an attacker acquires the user vulnerability scores? We consider two methods by which an attacker can gain access to the user vulnerability scores and discuss their impact on a defense system employing our proposed paradigm.

- *Leaked Vulnerability Scores.* Attackers can potentially gain access to the operator's own list of user vulnerability scores (e.g., through a leak of the list of vulnerable users or a breach at the operator). In such a scenario, the attackers know exactly which users are classified as vulnerable by the operator and can thus directly target those users. Even though the operator is placing vulnerable users under higher scrutiny (i.e., with added targeted protection), it can be argued that there is an asymmetry in cost between exploiting vs. protecting those vulnerable users. Thus, such an

attack may have a higher success rate. It is worth noting, however, that operators are likely to place the user vulnerability scores under the same (or higher) level of protection as other important user information.

- *Attacker-Developed Vulnerability Classifier.* Attackers can also potentially develop their own vulnerability classifier using externally accessible user features (e.g., information extracted from public profiles on online social networks). In such a scenario, the attackers could target those users they classify as vulnerable as a means of increasing the success rate of their attack. It is worth noting that the attackers do not have access to the same level of user information that operators have. As such, it is highly likely that the attacker's victim classifier would perform worse than that used by the operator.

Both outlined scenarios can be considered as advanced forms of targeted attacks which are beyond the scope of our proposed paradigm.

What is the relationship to authors' past work in this area? We have started research in the area of security for large socio-technical systems back in 2011 by evaluating the feasibility of a victim-centric defense approach in the context of online social networks (OSNs). We have demonstrated that large-scale socialbot infiltration campaigns are indeed a real threat [12–14]. To combat such attacks, we have developed Íntegro [10, 15], a defense system that leverages information about the vulnerable population: we showed that it is possible to identify this population (using supervised machine learning), and designed a defense that uses information about potential victims and the social graph topology to detect socialbots. Íntegro significantly outperforms other state of the art systems in terms of detection accuracy, and was deployed at Telefonica, the largest Spanish telecom, for their OSN Tuenti with over 50 million users in Spain and Latin America.

4. SUMMARY

We argue that information about user vulnerability can be effectively obtained then harnessed in all phases of a robust defense against automated, social engineering attacks in large-scale socio-technical systems: this information can help improve the accuracy of attack *detection* (either by the system operator or by the users themselves), can make *prevention* more effective by informing where to focus resources, and can improve *response* timeliness and accuracy.

5. ACKNOWLEDGMENTS

We would like to thank our shepherd, Mary Ellen Zurko, as well as our reviewers for their insightful comments. We would also like to thank all the NSPW'16 participants for the valuable feedback and fruitful discussions as well as the NSPW'16 scribes for the detailed notes. The presented work also benefited from our discussions with Alex Loffler from Telus. The first author is thankful to the University of British Columbia for a generous doctoral fellowship.

References

[1] PhishMe. http://phishme.com/. Accessed: 2016-07-06.

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

[3] L. Alvisi, A. Clement, A. Epasto, U. Sapienza, S. Lattanzi, and A. Panconesi. SoK: The evolution of sybil defense via social networks. *Proceedings of the IEEE Symposium on Security and Privacy*, 2013.

[4] J. Angwin, J. Larson, S. Mattu, and L. Kirchner. Machine bias risk assessments in criminal sentencing. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing, 2016. [Online; accessed 31-October-2016].

[5] R. Baeza-Yates. Data and algorithmic bias in the web. In *Proceedings of the 8th ACM Conference on Web Science*, WebSci '16, pages 1–1, New York, NY, USA, 2016. ACM.

[6] D. Balfanz, G. Durfee, R. E. Grinter, and D. Smetters. In search of usable security: Five lessons from the field. *IEEE Security and Privacy*, 2(5):19–24, 2004.

[7] M. Barreno, B. Nelson, A. D. Joseph, and J. Tygar. The security of machine learning. *Machine Learning*, 81(2):121–148, 2010.

[8] L. Bilge, T. Strufe, D. Balzarotti, and E. Kirda. All your contacts are belong to us: Automated identity theft attacks on social networks. In *Proceedings of the 18th International Conference on World Wide Web*, WWW '09, pages 551–560, New York, NY, USA, 2009. ACM.

[9] A. Blum, B. Wardman, T. Solorio, and G. Warner. Lexical feature based phishing URL detection using online learning. In *Proceedings of the 3rd ACM Workshop on Artificial Intelligence and Security*, AISec '10, pages 54–60, New York, NY, USA, 2010. ACM.

[10] Y. Boshmaf. *Security analysis of malicious socialbots on the web.* PhD thesis, University of British Columbia, 2015.

[11] Y. Boshmaf, K. Beznosov, and M. Ripeanu. Graph-based Sybil detection in social and information systems. In *IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 466–473, Niagara Falls, Canada, August 25-28 2013.

[12] Y. Boshmaf, I. Muslukhov, K. Beznosov, and M. Ripeanu. The socialbot network: when bots socialize for fame and money. In *Proceedings of the 27th Annual Computer Security Applications Conference*, ACSAC '11, pages 93–102, New York, NY, USA, 2011. ACM.

[13] Y. Boshmaf, I. Muslukhov, K. Beznosov, and M. Ripeanu. Key challenges in defending against malicious socialbots. In *Proceedings of the 5th USENIX Conference on Large-scale Exploits and Emergent Threats*, LEET'12, Berkeley, CA, USA, 2012. USENIX Association.

[14] Y. Boshmaf, I. Muslukhov, K. Beznosov, and M. Ripeanu. Design and analysis of a social botnet. *Computer Networks*, 57(2):556—578, February 2013.

[15] Y. Boshmaf, M. Ripeanu, K. Beznosov, and E. Santos-Neto. Thwarting fake OSN accounts by predicting their victims. In *Proceedings of the 8th ACM Workshop on Artificial Intelligence and Security*, AISec '15, pages 81–89, New York, NY, USA, October 2015. ACM.

[16] Q. Cao, M. Sirivianos, X. Yang, and T. Pregueiro. Aiding the detection of fake accounts in large scale social online services. In *Proceedings of the 9th USENIX Conference on Networked Systems Design and Implementation*, NSDI'12, pages 15–15, Berkeley, CA, USA, 2012. USENIX Association.

[17] M. Chandrasekaran, R. Chinchani, and S. Upadhyaya. PHONEY: mimicking user response to detect phishing attacks. In *Proceedings of the IEEE International Symposium on a World of Wireless, Mobile and Multimedia Networks*, WoWMoM '16, page 5. IEEE, 2006.

[18] T. Chen, S. Dick, and J. Miller. Detecting visually similar web pages: Application to phishing detection. *ACM Trans. Internet Technol.*, 10(2):5:1–5:38, June 2010.

[19] T. Chen, T. Stepan, S. Dick, and J. Miller. An anti-phishing system employing diffused information. *ACM Trans. Inf. Syst. Secur.*, 16(4):16:1–16:31, Apr. 2014.

[20] T. R. Dillahunt, C. A. Brooks, and S. Gulati. Detecting and visualizing filter bubbles in Google and Bing. In *Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems*, CHI EA '15, pages 1851–1856, New York, NY, USA, 2015. ACM.

[21] C. Dong and B. Zhou. Effectively detecting content spam on the web using topical diversity measures. In *Proceedings of the The 2012 IEEE/WIC/ACM International Joint Conferences on Web Intelligence and Intelligent Agent Technology - Volume 01*, WI-IAT '12, pages 266–273, Washington, DC, USA, 2012. IEEE Computer Society.

[22] S. Dua and X. Du. Data mining and machine learning in cybersecurity. 2011.

[23] M. Egele, G. Stringhini, C. Kruegel, and G. Vigna. COMPA: Detecting compromised accounts on social networks. In *Proceedings of the Network & Distributed System Security Symposium*, NDSS '13. ISOC, February 2013.

[24] S. Egelman and E. Peer. The myth of the average user: Improving privacy and security systems through individualization. In *Proceedings of the 2015 New Security Paradigms Workshop*, NSPW '15, pages 16–28, New York, NY, USA, 2015. ACM.

[25] Facebook. Facebook Login. https://developers.facebook.com/docs/facebook-login, 2016. [Online; accessed 31-October-2016].

[26] I. Fette, N. Sadeh, and A. Tomasic. Learning to detect phishing emails. In *Proceedings of the 16th International Conference on World Wide Web*, WWW '07, pages 649–656, New York, NY, USA, 2007. ACM.

[27] P. Fischer, S. E. G. Lea, and K. M. Evans. Why do individuals respond to fraudulent scam communications and lose money? the psychological determinants of scam compliance. *Journal of Applied Social Psychology*, 43(10):2060–2072, 2013.

[28] D. Florêncio and C. Herley. Where do security policies come from? In *Proceedings of the Sixth Symposium on Usable Privacy and Security*, SOUPS '10, pages 10:1–10:14, New York, NY, USA, 2010. ACM.

[29] D. Florêncio, C. Herley, and P. C. Van Oorschot. Pushing on string: The 'don't care' region of password strength. *Commun. ACM*, 59(11):66–74, Oct. 2016.

[30] D. M. Freeman, S. Jain, M. Dürmuth, B. Biggio, and G. Giacinto. Who are you? a statistical approach to measuring user authenticity. In *Proceedings of the 23rd Annual Network and Distributed System Security Symposium*, NDSS Symposium'16, San Diego, CA, USA, 2016. ISOC.

[31] Google. Safebrowsing API. 2015.

[32] Google. Google OpenID Connect. https://developers.google.com/identity/protocols/OpenIDConnect, 2016. [Online; accessed 31-October-2016].

[33] C. Herley. So long, and no thanks for the externalities: the rational rejection of security advice by users. In *Proceedings of the 2009 Workshop on New Security Paradigms Workshop*, NSPW '09, pages 133–144, New York, NY, USA, 2009. ACM.

[34] J. Huang, T. Zimmermann, N. Nagapan, C. Harrison, and B. C. Phillips. Mastering the art of war: How patterns of gameplay influence skill in halo. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '13, pages 695–704, New York, NY, USA, 2013. ACM.

[35] D. Ingold and S. Soper. Amazon doesn't consider the race of its customers. should it? http://www.bloomberg.com/graphics/2016-amazon-same-day/, 2016. [Online; accessed 31-October-2016].

[36] D. Irani, M. Balduzzi, D. Balzarotti, E. Kirda, and C. Pu. Reverse social engineering attacks in online social networks. *Detection of Intrusions and Malware, and Vulnerability Assessment*, pages 55–74, 2011.

[37] T. N. Jagatic, N. A. Johnson, M. Jakobsson, and F. Menczer. Social phishing. *Commun. ACM*, 50(10):94–100, 2007.

[38] A. Joshi, S. T. King, G. W. Dunlap, and P. M. Chen. Detecting past and present intrusions through vulnerability-specific predicates. In *Proceedings of the Twentieth ACM Symposium on Operating Systems Principles*, SOSP '05, pages 91–104, New York, NY, USA, 2005. ACM.

[39] N. E. Kass. An ethics framework for public health. *American Journal of Public Health*, 91(11):1776–1782, 2001.

[40] T. H.-J. Kim, A. Yamada, V. Gligor, J. Hong, and A. Perrig. Relationgram: Tie-strength visualization for user-controlled online identity authentication. In *In Proceedings of Financial Cryptography and Data Security Conference*, pages 69–77. Springer, 2013.

[41] C. Kruegel and G. Vigna. Anomaly detection of web-based attacks. In *Proceedings of the 10th ACM Conference on Computer and Communications Security*, CCS '03, pages 251–261, New York, NY, USA, 2003. ACM.

[42] P. Kumaraguru, J. Cranshaw, A. Acquisti, L. Cranor, J. Hong, M. A. Blair, and T. Pham. School of phish: A real-world evaluation of anti-phishing training. In *Proceedings of the 5th Symposium on Usable Privacy and Security*, SOUPS '09, pages 3:1–3:12, New York, NY, USA, 2009. ACM.

[43] P. Kumaraguru, S. Sheng, A. Acquisti, L. F. Cranor, and J. Hong. Teaching johnny not to fall for phish. *ACM Trans. Internet Technol.*, 10(2):7:1–7:31, June 2010.

[44] J. Langenderfer and T. A. Shimp. Consumer vulnerability to scams, swindles, and fraud: A new theory of visceral influences on persuasion. *Psychology and Marketing*, 18(7):763–783, 2001.

[45] G. Liu, G. Xiang, B. A. Pendleton, J. I. Hong, and W. Liu. Smartening the crowds: Computational techniques for improving human verification to fight phishing scams. In *Proceedings of the Seventh Symposium on Usable Privacy and Security*, SOUPS '11, pages 8:1–8:13, New York, NY, USA, 2011. ACM.

[46] Y. Liu, F. Chen, W. Kong, H. Yu, M. Zhang, S. Ma, and L. Ru. Identifying web spam with the wisdom of the crowds. *ACM Transactions on the Web*, 6(1), Mar. 2012.

[47] Y. Liu, A. Sarabi, J. Zhang, P. Naghizadeh, M. Karir, M. Bailey, and M. Liu. Cloudy with a chance of breach: Forecasting cyber security incidents. In *Proceedings of the 24th USENIX Security Symposium*, USENIX Security '15, pages 1009–1024, 2015.

[48] B. Lo. *Resolving ethical dilemmas: a guide for clinicians*. Williams & Wilkins, 1995.

[49] M. Loughnan, N. Tapper, and T. Phan. Identifying vulnerable populations in subtropical brisbane, australia: A guide for heatwave preparedness and health promotion. *ISRN Epidemiology*, 2014, 2014.

[50] D. Lowd and C. Meek. Adversarial learning. In *Proceedings of the 11th ACM SIGKDD International Conference on Knowledge Discovery in Data Mining*, KDD '05, pages 641–647, New York, NY, USA, 2005. ACM.

[51] L. Lu, V. Yegneswaran, P. Porras, and W. Lee. BLADE: An attack-agnostic approach for preventing drive-by malware infections. In *Proceedings of the 17th ACM Conference on Computer and Communications Security*, CCS '10, pages 440–450, New York, NY, USA, 2010. ACM.

[52] J. Ma, L. K. Saul, S. Savage, and G. M. Voelker. Beyond blacklists: Learning to detect malicious web sites from suspicious URLs. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '09, pages 1245–1254, New York, NY, USA, 2009. ACM.

[53] J. Ma, L. K. Saul, S. Savage, and G. M. Voelker. Learning to detect malicious URLs. *ACM Trans. Intell. Syst. Technol.*, 2(3):30:1–30:24, May 2011.

[54] M. Mannan and P. C. van Oorschot. Security and usability: the gap in real-world online banking. In *Proceedings of the 2007 Workshop on New Security Paradigms Workshop*, NSPW '07, pages 1–14. ACM, 2008.

[55] MIT Technology Review. Racism is poisoning online ad delivery, says harvard professor. https://www.technologyreview.com/s/510646/racism-is-poisoning-online-ad-delivery-says-harvard-professor/, 2013. [Online; accessed 31-October-2016].

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

[57] M. Motoyama, D. McCoy, K. Levchenko, S. Savage, and G. M. Voelker. Dirty jobs: The role of freelance labor in web service abuse. In *Proceedings of the 20th USENIX Security Symposium*, USENIX Security '11, Aug. 2011.

[58] J. Nazario. Phoneyc: A virtual client honeypot. In *Proceedings of the 2nd USENIX Conference on Large-scale Exploits and Emergent Threats: Botnets, Spyware, Worms, and More*, LEET'09, pages 6–6, Berkeley, CA, USA, 2009. USENIX Association.

[59] N. Newman. How big data enables economic harm to consumers, especially low income and other vulnerable sectors of the population. *Journal of Internet Law December*, 18:11–23, 2014.

[60] C. N. A. C. on SARS, P. Health, and C. D. Naylor. *Learning from SARS: renewal of public health in Canada: a report of the National Advisory Committee on SARS and Public Health*. National Advisory Committee on SARS and Public Health, 2003.

[61] R. Perdisci, W. Lee, and N. Feamster. Behavioral clustering of http-based malware and signature generation using malicious network traces. In *Proceedings of the 7th USENIX Conference on Networked Systems Design and Implementation*, NSDI'10, pages 26–26, Berkeley, CA, USA, 2010. USENIX Association.

[62] Z. Qian, Z. M. Mao, Y. Xie, and F. Yu. On network-level clusters for spam detection. In *Proceedings of the 17th Annual Network and Distributed System Security Symposium*, NDSS Symposium'10, San Diego, CA, USA, 2010.

[63] A. Ramachandran and N. Feamster. Understanding the network-level behavior of spammers. *SIGCOMM Comput. Commun. Rev.*, 36(4):291–302, Aug. 2006.

[64] H. Rashtian, Y. Boshmaf, P. Jaferian, and K. Beznosov. To befriend or not? a model of friend request acceptance on facebook. In *Symposium On Usable Privacy and Security (SOUPS 2014)*, pages 285–300, Menlo Park, CA, July 2014. USENIX Association.

[65] J. Ratkiewicz, M. Conover, M. Meiss, B. Gonçalves, S. Patil, A. Flammini, and F. Menczer. Truthy: mapping the spread of astroturf in microblog streams. In *Proceedings of the 20th International Conference Companion on World Wide Web*, WWW '11, pages 249–252, New York, NY, USA, 2011. ACM.

[66] RSA. 2013 - a year in review. 2013.

[67] S. Sheng, M. Holbrook, P. Kumaraguru, L. F. Cranor, and J. Downs. Who falls for phish?: A demographic analysis of phishing susceptibility and effectiveness of interventions. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '10, pages 373–382, New York, NY, USA, 2010. ACM.

[68] S. Sheng, B. Magnien, P. Kumaraguru, A. Acquisti, L. F. Cranor, J. Hong, and E. Nunge. Anti-phishing phil: The design and evaluation of a game that teaches people not to fall for phish. In *Proceedings of the 3rd Symposium on Usable Privacy and Security*, SOUPS '07, pages 88–99, New York, NY, USA, 2007. ACM.

[69] T. Shon and J. Moon. A hybrid machine learning approach to network anomaly detection. *Information Sciences*, 177(18):3799–3821, 2007.

[70] S. Sidiroglou and A. D. Keromytis. A network worm vaccine architecture. In *Proceedings of the 12th IEEE Conference on Enabling Technologies: Infrastructure for Collaborative Enterprises*, WET ICE '03, pages 220–225. IEEE, 2003.

[71] C. Simmons, C. Ellis, S. Shiva, D. Dasgupta, and Q. Wu. AVOIDIT: A cyber attack taxonomy. Technical Report CS-09-003, 2009.

[72] S. Sinha, M. Bailey, and F. Jahanian. Shades of grey: On the effectiveness of reputation-based blacklists. In *Proceedings of the 3rd International Conference on Malicious and Unwanted Software*, MALWARE '08, pages 57–64. IEEE, 2008.

[73] D. K. Smetters and R. E. Grinter. Moving from the design of usable security technologies to the design of useful secure applications. In *Proceedings of the 2002 Workshop on New Security Paradigms*, NSPW '02, pages 82–89, New York, NY, USA, 2002. ACM.

[74] K. Soska and N. Christin. Automatically detecting vulnerable websites before they turn malicious. In *Proceedings of the 23rd USENIX Security Symposium*, USENIX Security '14, pages 625–640, 2014.

[75] T. Stein, E. Chen, and K. Mangla. Facebook immune system. In *Proceedings of the 4th Workshop on Social Network Systems*, SNS '11, pages 8:1–8:8, New York, NY, USA, 2011. ACM.

[76] K. Strater and H. R. Lipford. Strategies and struggles with privacy in an online social networking community. In *Proceedings of the 22nd British HCI Group Annual Conference on People and Computers: Culture, Creativity, Interaction-Volume 1*, pages 111–119. British Computer Society, 2008.

[77] G. Stringhini, C. Kruegel, and G. Vigna. Detecting spammers on social networks. In *Proceedings of the 26th Annual Computer Security Applications Conference*, ACSAC '10, pages 1–9, New York, NY, USA, 2010. ACM.

[78] A. H. Sung, J. Xu, P. Chavez, and S. Mukkamala. Static analyzer of vicious executables (SAVE). In *Proceedings of the 20th Annual Computer Security Applications Conference*, ACSAC '04, pages 326–334, Dec 2004.

[79] Symantec. Internet security threat report. 2015.

[80] K. Thomas, C. Grier, D. Song, and V. Paxson. Suspended accounts in retrospect: an analysis of Twitter spam. In *Proceedings of the 2011 ACM SIGCOMM Conference on Internet Measurement Conference*, pages 243–258. ACM, 2011.

[81] K. Thomas, D. McCoy, and C. Grier. Trafficking fraudulent accounts: the role of the underground market in Twitter spam and abuse. In *Proceedings of the 22nd USENIX Security Symposium*, USENIX Security '13, pages 195–210. USENIX Association, 2013.

[82] O. Thonnard, L. Bilge, A. Kashyap, and M. Lee. Are you at risk? profiling organizations and individuals subject to targeted attacks. In *Financial Cryptography and Data Security*, pages 13–31. Springer, 2015.

[83] D. Twining, M. M. Williamson, M. J. F. Mowbray, and M. Rahmouni. Email prioritization: Reducing delays on legitimate mail caused by junk mail. In *Proceedings of the Annual Conference on USENIX Annual Technical Conference*, ATEC '04, pages 4–4, Berkeley, CA, USA, 2004. USENIX Association.

[84] J. Tygar. Adversarial machine learning. *Internet Computing, IEEE*, 15(5):4–6, 2011.

[85] R. Verma and K. Dyer. On the character of phishing URLs: Accurate and robust statistical learning classifiers. In *Proceedings of the 5th ACM Conference on Data and Application Security and Privacy*, CODASPY '15, pages 111–122, New York, NY, USA, 2015. ACM.

[86] C. Wagner, S. Mitter, C. Körner, and M. Strohmaier. When social bots attack: Modeling susceptibility of users in online social networks. In *Proceedings of the International Conference on World Wide Web*, WWW '12, page 2, 2012.

[87] H. J. Wang, C. Guo, D. R. Simon, and A. Zugenmaier. Shield: Vulnerability-driven network filters for preventing known vulnerability exploits. In *Proceedings of the 2004 Conference on Applications, Technologies, Architectures, and Protocols for Computer Communications*, SIGCOMM '04, pages 193–204, New York, NY, USA, 2004. ACM.

[88] Y.-M. Wang, D. Beck, X. Jiang, R. Roussev, C. Verbowski, S. Chen, and S. King. Automated web patrol with strider honeymonkeys. In *Proceedings of the Network & Distributed System Security Symposium*, NDSS '06, pages 35–49. ISOC, 2006.

[89] G. Xiang, J. Hong, C. P. Rose, and L. Cranor. CANTINA+: A feature-rich machine learning framework for detecting phishing web sites. *ACM Trans. Inf. Syst. Secur.*, 14(2):21:1–21:28, Sept. 2011.

[90] Y. Xie, F. Yu, Q. Ke, M. Abadi, E. Gillum, K. Vitaldevaria, J. Walter, J. Huang, and Z. M. Mao. Innocent by association: early recognition of legitimate users. In *Proceedings of the 2012 ACM Conference on Computer and Communications Security*, CCS '12, pages 353–364, New York, NY, USA, 2012. ACM.

[91] G. Yan, G. Chen, S. Eidenbenz, and N. Li. Malware propagation in online social networks: nature, dynamics, and defense implications. In *Proceedings of the 6th ACM Symposium on Information, Computer and Communications Security*, pages 196–206. ACM, 2011.

[92] H. Yu. Sybil defenses via social networks: a tutorial and survey. *SIGACT News*, 42:80–101, October 2011.

[93] W. Yu, S. Chellappan, X. Wang, and D. Xuan. Peer-to-peer system-based active worm attacks: Modeling, analysis and defense. *Comput. Commun.*, 31(17):4005–4017, Nov. 2008.

[94] Y. Zeng, X. Hu, and K. G. Shin. Detection of botnets using combined host- and network-level information. In *Proceedings of the 2010 IEEE/IFIP International Conference on Dependable Systems Networks*, DSN '10, pages 291–300, June 2010.

[95] Y. Zhang, J. I. Hong, and L. F. Cranor. CANTINA: A content-based approach to detecting phishing web sites. In *Proceedings of the 16th International Conference on World Wide Web*, WWW '07, pages 639–648, New York, NY, USA, 2007. ACM.