IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received September 16, 2016, accepted October 6, 2016, date of publication October 10, 2016, date of current version November 8, 2016.

Digital Object Identifier 10.1109/ACCESS.2016.2616285

You Are Probably Not the Weakest Link: Towards Practical Prediction of Susceptibility to Semantic Social Engineering Attacks

RYAN HEARTFIELD, GEORGE LOUKAS, AND DIANE GAN
University of Greenwich, London, SE10 9LS, U.K.
Corresponding author: R. Heartfield (r.j.heartfield@gre.ac.uk)

ABSTRACT Semantic social engineering attacks are a pervasive threat to computer and communication systems. By employing deception rather than by exploiting technical vulnerabilities, spear-phishing, obfuscated URLs, drive-by downloads, spoofed websites, scareware, and other attacks are able to circumvent traditional technical security controls and target the user directly. Our aim is to explore the feasibility of predicting user susceptibility to deception-based attacks through attributes that can be measured, preferably in real-time and in an automated manner. Toward this goal, we have conducted two experiments, the first on 4333 users recruited on the Internet, allowing us to identify useful high-level features through association rule mining, and the second on a smaller group of 315 users, allowing us to study these features in more detail. In both experiments, participants were presented with attack and non-attack exhibits and were tested in terms of their ability to distinguish between the two. Using the data collected, we have determined practical predictors of users' susceptibility against semantic attacks to produce and evaluate a logistic regression and a random forest prediction model, with the accuracy rates of .68 and .71, respectively. We have observed that security training makes a noticeable difference in a user's ability to detect deception attempts, with one of the most important features being the time since last self-study, while formal security education through lectures appears to be much less useful as a predictor. Other important features were computer literacy, familiarity, and frequency of access to a specific platform. Depending on an organisation's preferences, the models learned can be configured to minimise false positives or false negatives or maximise accuracy, based on a probability threshold. For both models, a threshold choice of 0.55 would keep both false positives and false negatives below 0.2.

INDEX TERMS Security, cyber crime, social engineering, semantic attacks.

I. INTRODUCTION

Semantic social engineering attacks target the user-computer interface in order to deceive a user into performing an action that will breach a system's information security [1]. On any system, the user interface is always vulnerable to abuse by authorised users, with or without their knowledge. Traditional deception-based attacks, such as phishing emails, spoofed websites and drive-by downloads, have shifted to new and emerging platforms in social media [2], cloud applications [3] and near field communications [4]. Efforts towards technical defence against semantic attacks have lead to the development of solutions that are typically specific in design. This can be attributed to the sheer complexity required to translate what is essentially human deception into code, as well as attempting to combine this into a solution that spans disparate platforms. One example is phishing emails, where filtering and classification software have proven to be highly successful [5]–[7]. However, these defence mechanisms are built to function on email systems only, unable to prevent conceptually very similar phishing attacks in instant messaging, social media and other platforms. Similarly, automated tools developed to block drive-by downloads via web browsers have been shown to be highly effective in mitigating the threat [8], [9], yet the same tools cannot prevent a drive-by attack in removable media.

Alternative approaches to technical solutions have focused on managing users themselves, rather than the computer interface. For example, creation of policy and process for user compliance [10] has helped to define specific rules which enforce secure system use, but these are almost never

applicable to the private user of a computer system and the Internet. Furthermore, compliance guidelines are usually static in nature and therefore can quickly become out-of-date when new attack methods appear. User education and awareness training have been evaluated extensively and in practice have been shown to improve user responses to specific attack scenarios [11], [12], but it is difficult to automate this process and even more difficult to measure its lasting effect. Moreover, training material tends to be limited to known exploitations and requires regular updates to include new attack vectors. Systems generating visual warnings or security indicators have also been implemented, presented to users in real-time by indicating a possible attack or whether a potential threat exists, but research has shown that in practice users often do not pay attention to them or do not understand them [13].

A comprehensive survey by Khonji et al. [14] evaluating the state of phishing detection provides a valuable insight into potential future defences. The researchers have highlighted the application of machine learning techniques as a promising approach to defence, producing accurate attack classifiers and effective defences against zero-day threats. Measuring the effectiveness of user training has also been suggested, where research towards a hybrid user/software solution is indicated as a potential multi-layered approach to protection.

Given the limitations of defences designed for specific attacks and platforms, it is attractive to look also towards the feasibility of predicting a user's susceptibility to different semantic attacks in order to augment technical systems with user-driven defence. For example, user susceptibility profiles can be used to support systems that are dynamic, by training predictors with user data collected in real-time or over a period of time, and allowing dynamic allocation of access rights dependent on a detected user profile. Furthermore, they could support the development of context-based user awareness systems, where training material would be tailored to users depending on their susceptibility to different deception vectors. User susceptibility profiles can also provide useful measurement criteria for predicting the performance of human sensors of semantic attacks, indicating whether a user report of a suspected attack is accurate (and worth investigating); sharing analogies to the learning and prediction capabilities employed in sandbox antivirus defences for categorising and identifying different malware families [15]. Towards this vision, we have conducted two experiments with the participants being asked to tell whether particular exhibits show an attack or not. We have collected data regarding both the users and their performance in detecting attacks that employ different deception vectors [1] and have developed two prediction models. The first experiment helped identify high level predictors that can be measured ethically, automatically and in real-time, whilst being applicable across the wider Internet population; we define this study as stage 1. The second experiment helped build upon the initial predictor features by further dissecting each into a series of sub-features used to predict susceptibility against new attacks using a smaller population; we define this as stage 2.

II. PREDICTING SUSCEPTIBILITY

A. RELATED WORK

In computer security, it is usually computer systems, networks, applications and data that are monitored to be able to detect and mitigate threats. Researchers have also attempted to monitor and profile unauthorised users [16], [17] or witting insiders performing unauthorised actions [18]. However, semantic social engineering attacks target authorised users and lure them into performing an authorised (albeit compromising) action. Recent research in this area has focused on demographic attributes and psychological indicators as methods for predicting user susceptibility. For example, in the field of behavioural science, research has explored the impact of personality traits [19], influencing and persuasion techniques [20] as measurement criteria for predicting susceptibility to semantic attacks. A study carried out in [21] has reported that female participants exhibiting neurotic behaviour were more likely to respond to phishing emails than female and male participants that did not. More recently, the same researchers have conducted a spear-phishing field-experiment, where the tendency for conscientiousness reported a high correlation to phishing susceptibility [22]. Research in [23] has reported openness, positive behaviour (e.g., use of language) and high levels of conversationalist activity as predictors of vulnerability to an online social network bot. In [24], researchers have conducted a survey and field experiment of phishing attacks which found that participants who demonstrated higher degrees of normative, affective and continuance commitment, obedience to authority and trust, to be more susceptible to phishing. Similar results were also reported in a recent study in [25], where submissiveness and trust predicted higher susceptibility to phishing emails. Crucially, these personality traits were also found to perform consistently as predictors of susceptibility amongst participants from different geographical locations, in this case Australia and Saudi Arabia.

Demographic research has considered Internet usage and behaviour as prediction criteria of susceptibility to semantic attacks. In particular, it has been reported that users who have knowledge of or take guidance from visual cues (security indicators, source, design, language, etc.) on technology platforms are often good predictors of susceptibility. For example, [26]–[28] and [29] have all reported a lower degree of susceptibility to phishing attacks in emails and websites when the participants are aware of security indicators and visual components. However, in many cases participants did not understand what the security indicators meant and the varying severity of their message. In fact, in [30] and [31], it has been reported that the effect of habitation to the visual cues and especially security warnings increases susceptibility to attacks. Where studies have included general demographic elements such as age and gender, a number of studies have reported that female participants were found to be more

TABLE 1. Related research in the field of semantic social engineering attacks.

Pub.	Attack Type*	Methodology	Participants	Practicality of predictors			Key findings for reducing [-] or increasing [+] susceptibility
				Ethical	Real-time	Automatic	
[26]	E	Exhibit survey	179	✓		✓	[-] Knowledge of technical and visual cues
[27]	E	Interview, role-play	20	✓		✓	[-] Familiarity with specific attacks, [-] visual cues
[29]	E/W	Exhibit test	17	✓			[-] User guided by technical and visual cues
[28]	W	Role-play	232	✓		✓	[-] Knowledge of attacks, technical and visual cues
[24]	E	Survey, simulated phishing	588	✓			[+] Normative, affective, continuance, obedience commitment, trust
[30]	E/W	Survey, role-play	70	✓			[-] Interactive browser warnings, [+] platform habitation
[11]	E	Simulated phishing	515	✓		✓	[-] Trained on an embedded phishing email training system
[32]	E	Role-play	1,001				[+] Ages 18-24, [+] Gender (Female, attr. less technical training)
[33]	E	Analysis, survey, interview	224				[+] Gender (Female), [-] non-exposure to deceptive visual cues
[39]	E	Survey, simulated phishing	446	✓			[-] Comp. self-efficacy, web experience, security knowledge, suspicion
[40]	E	Survey	64			✓	[-] Tailored susceptibility message
[31]	E	Survey, simulated phishing	321	✓			[-] Knowledge+attention+cues+elaboration [+] load triggered habitation
[41]	B	Social media experiment	500	✓	✓	✓	[+] High platform activity, openness, positive behaviour (e.g., language)
[23]	E	Simulated phishing	10,917				[-] Gender (Female), [-] Age (18-21),
[42]	E	Survey, Training experiment	321	✓	✓	✓	[-] Attention to deception indicators, attack knowledge
[43]	E	Test assessment	210	✓	✓	✓	[-] Computer security self efficacy and proven attack knowledge
[21]	E	Survey, simulated phishing	100				[+] Neurotic personality trait
[44]	E	Simulated phishing	2,624	✓			[+] Cialdini's influence techniques + self/non-self determination
[25]	E	Survey	296				[+] Submissiveness, trust, [-] Visual cues & important privacy data
[22]	E	Simulated phishing attack	40				[+] Gender (Female), conscientiousness, self-efficacy
[45]	W	Simulated phishing	21	✓	✓	✓	[-] Gaze time on browser chrome elements
[46]	W	Simulated phishing	173	✓	✓	✓	[-] Knowledge combined with Familiarity and HTTPS indicators
Experiment 1	**E/W/W1/S/A/Q**	**Exhibit test**	**4,333**	✓	✓	✓	**[-] Security training, platform freq., familiarity & self-efficacy**
Experiment 2	**E/W/A/T/S1**	**Exhibit test**	**315**	✓	✓	✓	**[-] Security training, platform freq., dur. & familiarity, self-efficacy**

*** E = Phishing emails/IMs, W = phishing websites, B = Online social network bot, S = search engine poisoning, W1 = WiFi evil twin, A = Fake app Q= QRishing, T= Typosquating, S1 = Scareware**

susceptible to phishing attacks than male participants [22], [32]–[34]. In [11], users were measured demographically as to whether they have had training on the phishing email training system *PhishGuru*, where the number of training sessions taken by users are used as input features to identify the lasting effect of the training. It was found that having completed training sessions on *PhishGuru* is an accurate predictor of lower susceptibility to phishing emails.

Technical prediction systems have been previously proposed in [35] and more recently [36]. The first describes a system which would present users through a series of information security related questions within a web pop-up. Then, the system uses a series of weighted decision algorithms to quantify the user's degree of susceptibility based on the responses to the questions, and accordingly displays a visual indicator of susceptibility to the user as a form of awareness mechanism. No security enforcing functions are implemented. To date, there is no further information regarding its practical implementation and evaluation. The latter, and more recent study empowers user to report whether an email is a suspected phishing attack. Based on prior knowledge and in-line warnings, correct reporting conversely highlights predictor features for phishing susceptibility.

Table 1 provides an overview of the literature associated to susceptibility research in semantic social engineering attacks. In the "Technical measurement" set of columns, we have identified for each study whether the predictors of susceptibility can be realistically measured by a technical system in real-time, automatically and ethically. By ethical, we refer to aspects of diversity and inclusion related to protected personal characteristics [37], and we extend this to also include personality traits, where decision making based on assessment of personality types are argued to be a form of discrimination [38].

The available literature for predicting user susceptibility to semantic social engineering attacks is not as mature as other areas of computer security. Most related studies have been constrained by small sample sizes and predictors that are difficult to generalise across a multitude of semantic attacks. To some extent, this is due to the fact that most researchers focus only on phishing attacks, which is only one section of the problem space [1]. Specialised training systems have been shown to work well [11], as well as technical models combining demographic and behavioural attributes [31], but they are application-specific and do not consider other deception vectors that might be employed in semantic attacks. Therefore, it is difficult with the results produced from current studies to generalise across a wide range of attack types and it is unclear which of the research results could be realistically integrated into a technical system for defence.

To overcome these limitations and as our aim is focused on facilitating the development of technical defence systems, we only select predictor terms that can be collected and measured in real-time, automatically and ethically. We argue that in order to predict susceptibility to a wide range of semantic attacks, the mechanism for measuring susceptibility

should be naive to low-level and attack-specific parameters (e.g., sender source and body of text within an email, URL composition in website post, etc.).

B. INDICATORS OF SUSCEPTIBILITY

To identify practical indicators, we start with five high-level concepts associated to user knowledge, experience and behaviour:

1) SECURITY TRAINING (S)

Refers to the individual's type of computer security training. Prediction of susceptibility by computer security training has been shown to produce accurate results [11], but the approach can be limited by the specialised system delivering the training or the specific training curriculum. For example, it is likely that a user who is self-trained will cover a wider range of material relevant to their technology profile than an employee who has only received work-based training on systems the organisation uses. Moreover, the long-term benefit of training and skill fade is not clear.

2) FAMILIARITY (FA)

Refers to the familiarity the individual has with a given platform. Familiarity is a key enabler of distinguishing between what visually looks normal and what is normal behaviour. For example, in [27] and [28], the researchers have identified familiarity with specific attacks and visual cues as key predictors of susceptibility, both of which describe how a user identifies what is normal visually or behaviourally on a system and what is not. Similar findings were also reported in [29], with knowledge of visual cues being attributed to familiarity with the type of platform used. In this context, platform habitation [30] is a factor that can increase susceptibility to semantic attacks, facilitated in part by platform familiarity. At the same time, without familiarity a user may be unaware how a system should normally look and behave, and consequently may fail to detect an attack or may see threats where they do not exist.

3) FREQUENCY (FR)

Refers to the frequency with which the individual accesses the type of platform. A user who accesses a specific type of platform (e.g., social media websites) very frequently may be more aware of the kind of attacks that occur within that type of platform, regardless of their actual familiarity with the specific providers platform (e.g., with Google+ social network site).

4) DURATION (DR)

Refers to the duration for which the individual accesses the type of platform. Similarly to FR, a user who uses a specific type of platform for long periods may be more aware of the kind of attacks that occur within that type of platform. However, it is also possible that the longer the duration the higher risk of platform habitation which may or may not have an adverse effect [41].

5) COMPUTER LITERACY (CL) AND SECURITY AWARENESS (SA)

Refers to the user's self-efficacy in respect to their computer literacy and computer security awareness. A user's self-identified level of computer literacy and computer security awareness has been observed to be an important predictor in numerous studies seeking to identify what influences reduced susceptibility to phishing emails [22], [39], [43]. Overall, self-efficacy was found to accurately represent a user's expectation of their ability to use a computer system competently and securely. However, self-efficacy implicitly harbours a degree of bias depending on the user's honesty and the accuracy and practicality of the measurement scale used. While we count it here as practical, we assume that in actual application, both CL and SA would need to be validated against evidence (e.g., with some form of testing, certification, etc.).

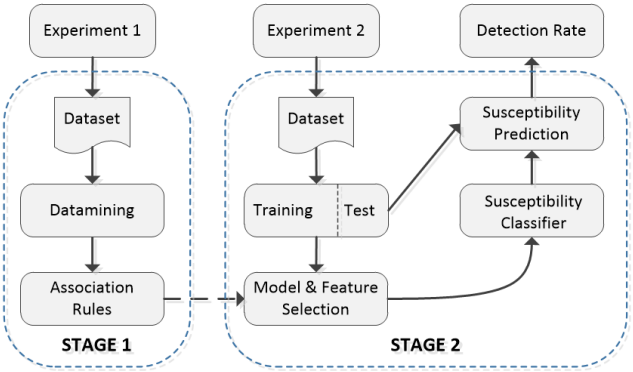

FIGURE 1. Experiment approach and methodology.

III. METHODOLOGY

Figure 1 summarises our two-stage experimental approach. In stage 1, we have conducted a large scale experiment, where we applied data mining techniques to try and identify whether relationships and associations exist between the different indicators of susceptibility described in Section II-B. In stage 2, we have utilised the results from the stage 1 analysis to apply a greater degree of granularity (and measurability) to each of the indicators highlighted through the data mining process. These refined predictor features were then employed in a second experiment in order to determine practical predictors of susceptibility and develop a model to form a susceptibility classifier.

Both experiments were designed to be quantitative in nature in order to generate numerical data that could be transformed into usable statistics. Some qualitative data was captured in experiment 2, where users were asked to explain in free-text for each exhibit why they had classified it as an attack or non-attack; this data was used to eliminate sample "noise", such as participants who guessed or marked all exhibits as attacks (or non-attacks). Furthermore, attack exhibits were randomised so that participants could not guess

the order between attack and non-attack exhibits in the susceptibility test.

Both experiments were implemented in the online survey platform *Qualtrics* and consisted of a short survey that collected demographic and platform behaviour data, followed by an exhibit-based susceptibility test. In total, after sample cleaning and pre-processing, experiment 1 consisted of 4,333 participant responses, and experiment 2 consisted of 315 participant responses. Both experiments provided participants with a study brief prior to commencing the survey, so as to ensure they understood how to proceed with answering the survey and exhibit test questions.

The research was approved by our institution's research ethics committee and participants were informed of the purpose of the study prior to providing online consent and confirmation of being over 18 years of age. Furthermore, all data were anonymised and participants were also given the opportunity to opt out of the study analysis after completing the test; participants who opted out had their responses removed from the study.

A. RECRUITMENT

1) EXPERIMENT 1

In the first experiment, participants were cultivated via an online advertisement challenging people to take a test of their susceptibility to semantic social engineering attacks. This advertisement was posted in a number of popular online forums and social media communities, including Reddit, StumbleUpon, Facebook and Twitter. Additionally, undergraduate and research students were recruited via email. The recruitment methodology of presenting the questionnaire primarily as a challenge and secondarily as a research medium proved successful because participants were eager to test themselves on a real-world skill that is becoming increasingly important. As a result, our advertisement gained reputation quickly by being up-voted and shared within a variety of social media platforms, resulting in a substantial sample size that allows meaningful statistical analysis (4,333 responses). Our sample included participants across a broad range of online platforms, as well as technical and non-technical environments from within our university's undergraduate population. Also, in many studies in this area, the real nature of the study is initially hidden from the participants, so that the strength of a deception attempt is not weakened by suspicion. Here, instead we use the participants as human binary classifiers of exhibits into attack versus non-attack. In this manner, we can reveal the nature of the study from the beginning, which addresses key technical and ethical challenges associated with temporarily deceiving the participants.

2) EXPERIMENT 2

In the second experiment, a controlled recruitment policy was employed in order to achieve a balanced sample of participants who had received some security training and were technology savvy and generic online users with little or no training. New undergraduate and research students were invited to participate in the experiment if they were studying a computer security program and the professional service *Qualtrics Panels* was used to recruit participants from a wider, more generic population demographic. Specifically, participants from the US ranging between the ages ranging from 18-65, both female and male, were defined as the participant selection criteria. No specific technology or security training attributes were defined in the *Qualtrics Panel* recruitment. Figures 2 and 3 shows the geographical distribution of the participants for both experiments.

FIGURE 2. Number of survey participants by geographical location for Experiment 1.

FIGURE 3. Number of survey participants by geographical location for Experiment 2.

B. EXPERIMENT DESIGN

The survey portion of the experiment required participants to answer a series of questions related to age, gender, general education, security training (S), platform familiarity (FA), frequency (FR) and duration of access (DR), computer literacy (CL) and security awareness (SA):

1) SECURITY TRAINING (S)

Formal computer security education (S1), work-based computer security training (S2) and self-study computer security training (S3), each coded as a binary response: Yes (1), No (0). In relation to the terminology used in [47], we directly map formal education as ''Formal Learning'', work-

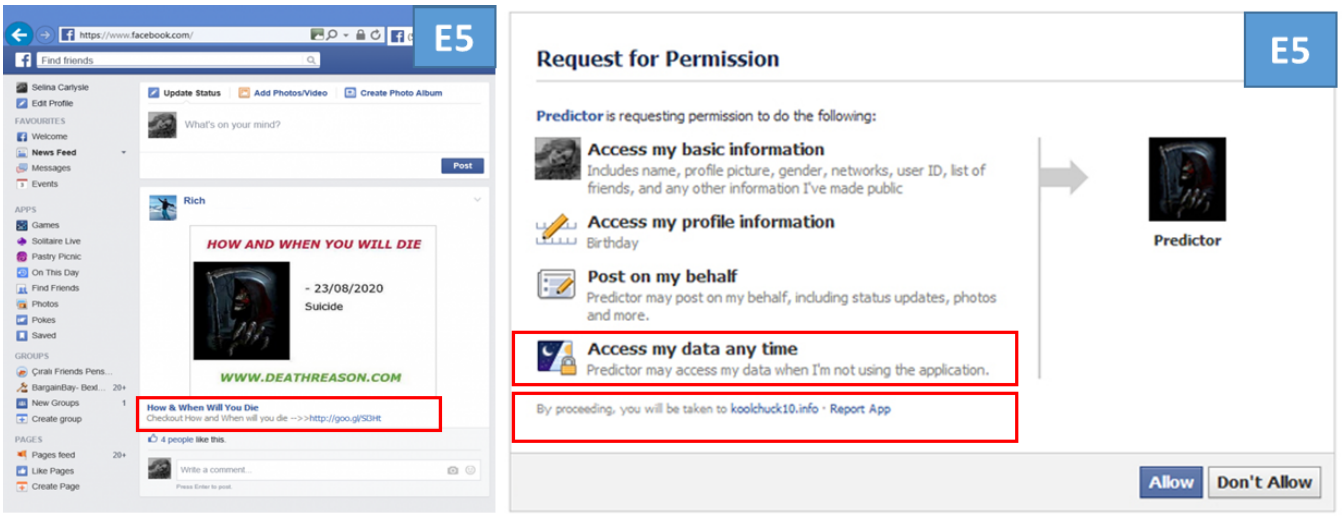

FIGURE 4. **Experiment 1: Exhibit 5 (screenshot) - Fake "Clickbait" app on Facebook.**

based training as "Non-formal learning", and self-study as "Informal learning".

2) FAMILIARITY (FA)

We use FA for familiarity with a particular provider's platform (e.g., GMAIL), coded as: Not very familiar (1), Somewhat familiar (2), Very familiar (3).

3) FREQUENCY (FR)

For each platform category presented in the susceptibility test, coded as: Never (1), less than once a month (2), once a month (3), weekly (4), daily (5).

4) DURATION (DR)

For each platform category presented in the susceptibility test, coded: None (1), less than 30 mins (2), 30 mins to 1 hour (3), 1 to 2 hours (4), 2-4 hours (5), 4 hours+ (6).

5) COMPUTER LITERACY (CL)

Self-reported level of computer literacy using a 0-100 scale.

6) SECURITY AWARENESS (SA)

Self-reported level of security awareness using a 0-100 scale.

Each experiment included a series of 12 exhibits (6 attacks and 6 non-attacks), each containing a concise scenario followed by an exhibit, consisting of one or more screenshots, GIF animations or videos. For each, participants were asked to examine the exhibit and provide a binary response to categorise each one as: "Most likely an attack" or "Most likely not an attack". In our analysis, correct responses were coded as 1 and incorrect ones as 0.

To determine general indicators of susceptibility, the attack exhibits chosen spanned a range of semantic social engineering attacks across different platforms. We have developed each semantic attack according to the three different types of deception vectors of the semantic social engineering taxonomy in [1]. In accordance with this, deception vector refers to the mechanism by which the participant is deceived into facilitating a security breach. It can be cosmetic (DV1), where the semantic attack is visually convincing, but does not necessarily conform to expected platform behaviour; behaviour-based (DV2), where the attack behaves in a manner that is expected or accepted within platform convention, but is not visually convincing; and both cosmetic and behaviour-based (DV3), where the attack needs to be both visually and behaviourally convincing to deceive the user.

A breakdown of the 24 exhibits developed for the two experiments is presented in Table 2, along with the participants' average score in each exhibit. The average score can serve as an indication of the difficulty of each exhibit. To illustrate the style of the presentation of the exhibits to the participants, we have also included three indicative examples of attack exhibits (Figures 5, 4 and 6, which correspond to exhibits Exp1.11, Exp1.5 and Exp2.11 respectively). For presentation purposes here, we have added red outlines to represent visual attack indicators in the exhibit. These outlines were obviously not visible to the participants.

C. OVERALL PARTICIPANT PERFORMANCE RESULTS

To determine overall accuracy and precision, we follow the approach defined in signal detection theory [48], [49], which is geared towards analysing data generated from human experiments, where the task is to categorise participants' responses generated by a known process or by chance. This approach is common in analysing experiments that involve semantic attacks, such as phishing [33]. In the standard formulas used below, for exhibit $k \in [1, K]$, $T_{p,k}$ is the number of true positives (i.e., correctly identified as attack), $T_{n,k}$ is the number of true negatives (i.e., correctly identified as non-attack), $F_{p,k}$ is the number of false positives (i.e., incorrectly

TABLE 2. Brief description of the 24 exhibits developed for the two experiments. The overall score (percentage of correct answers) of the participants for each exhibit is an indication of its difficulty. SA refers to the existence or not of a semantic attack in each exhibit (Y/N). For the exhibits where there was an attack, DV is the deception vector (DV1: cosmetic; DV2: behaviour-based; DV3: both cosmetic and behaviour-based).

Experiment 1							
ID	**Perf.**	**SA**	**Platform Type**	**Provider Platform**	**DV**	**Exhibit**	**Description**
E1	0.72	N	Social Media	GoogleApps	-	Screenshot	Facebook app down from GoogleApps with app permissions
E2	0.63	N	Email	Gmail	-	Screenshot	Failed login report from Google for this email account based on location
E3	0.67	Y	Public WiFi	Starbucks WiFi	DV3	Screenshot	Evil Twin WiFi attack, fake SSID and web portal with private IP, login request email credentials
E4	0.76	N	Social Media	Facebook	-	Screenshot	Mistyped Facebook URL “Facebok” redirects to real Facebook website
E5	0.88	Y	Social Media	Facebook	DV2	Screenshot	“Clickbait” app shared by friend on user timeline; requests permissions and redirects
E6	0.67	Y	Email	Gmail	DV1	Screenshot	Spearphishing Twitter email responding to bogus functionality, with button to login
E7	0.88	Y	Social Media	Twitter	DV3	Screenshot	Twitter phishing website spoofing Twitter website real homepage, URL “http://twitteri.com”
E8	0.78	N	Public WiFi	BT WiFi Hotspot	-	Screenshot	BT FON public WiFi hotspot SSID (multiple found in area), login portal, secured with HTTPS
E9	0.67	Y	Email	Gmail	DV1	Screenshot	Email confirming Paypal purchase with link to “dispute transcation”
E10	0.72	N	Email	Outlook	-	Screenshot	Paypal email requesting activation via URL link. Approved sender address icon supplied
E11	0.86	Y	Social Media	Steam	DV3	Screenshot	QRishing attack (malicious QR code) on fake Steam social media account redirecting to fake website login
E12	0.64	N	Social Media	Twitter	-	Screenshot	Embedded game advert in Twitter app on mobile; shows app rating and downloads
Experiment 2							
ID	**Perf.**	**SA**	**Platform Type**	**Provider Platform**	**DV**	**Exhibit**	**Description**
E1	0.78	Y	Social Media	Facebook	DV1	GIF	Video masquerading acting a “clickbait” link to bogus survey designed to steal user credentials
E2	0.55	Y	IM	Steam	DV2	Screenshot	Phishing instant message containing URL link
E3	0.83	N	Social Media	Facebook	-	Screenshot	Facebook app update on mobile device with permissions request
E4	0.75	N	Public WiFi	Starbucks	-	Screenshot	Starbucks / AT&T login capture portal, URL generated by AT &T web proxy
E5	0.80	N	Web browser	Android	DV3	Screenshot	Fake virus scan pop-up downloading Android scareware requiring purchase to remove infections
E6	0.91	Y	Social Media	Youtube	-	Screenshot	Youtube blocked video due to location settings and copyright
E7	0.55	N	Email	Gmail	DV1	Video	Phishing email from Natwest Bank containing a URL link
E8	0.79	Y	Social Media	Facebook	-	Video	Application permissions request to add game to Facebook profile (Desktop version)
E9	0.86	N	Social Media	Twitter	-	Video	Twitter account confirmation on website; email from Twitter requesting confirmation via link
E10	0.46	N	Social Media	Twitter	-	Video	Comodo email with link to encryption certificate; link broken due to timeout of download
E11	0.61	Y	Web Browser	MS Edge	DV2	Video	”Faceb00k.com” typosquatting redirecting to webpage warning user to install fake Chrome patch
E12	0.36	Y	E-commerce	Gumtree	DV3	Video	Phishing email redirecting to phishing Gumtree login; once entered redirects to real website

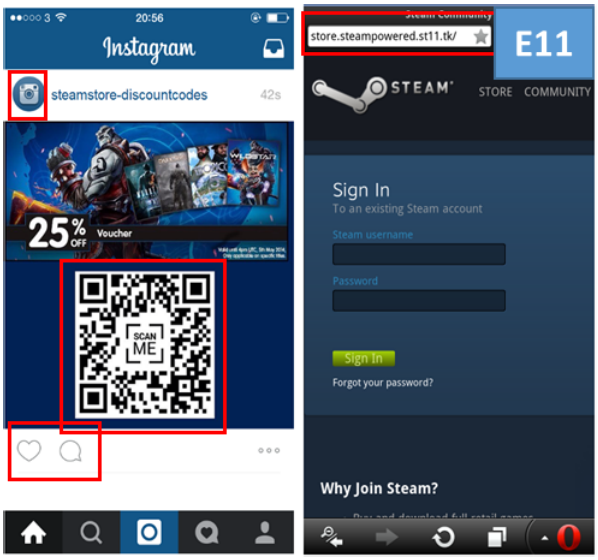

FIGURE 5. Experiment 1: Exhibit 11 (screenshot) - “Qrishing” attack leading to Steam phishing site.

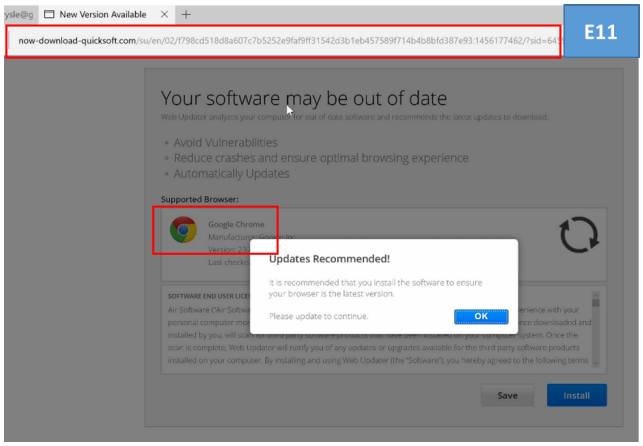

FIGURE 6. Experiment 2: Exhibit 11 (Video) - “Typosquatting” attack on Microsoft Edge browser leading to an attack website with a malicious update prompt for Google Chrome browser.

identified as attack), and $F_{n,k}$ is number of false negatives (i.e., incorrectly identified as non-attack). Note that in this case, $K = 12$, and by accuracy and precision, we are referring to the average accuracy and average precision across all 12 exhibits.

$$Accuracy = \frac{1}{K}\sum_{k=1}^{K}\frac{T_{p,k}+T_{n,k}}{T_{p,k}+T_{n,k}+F_{p,k}+F_{n,k}} \quad (1)$$

$$Precision = \frac{1}{K}\sum_{k=1}^{K}\frac{T_{p,k}}{T_{p,k}+F_{p,k}} \quad (2)$$

To facilitate the analysis of the participants’ responses, we developed an equal number of attack and non-attack exhibits

for each platform category used in the experiment. In this section, our aim is to simply evaluate the overall performance of the users in our samples as human classifiers of the given exhibits. We also note the performance of individual groups that are commonly studied in this space, including groups by country, age and gender. We derive the country based on the IP of the participant, assuming that there is no strong reason to believe that several participants would have spoofed their IP while taking part in this experiment. Also, age and gender are self-reported. Again, we do not have any strong reason to believe that several participants would provide false details in this case.

TABLE 3. Experiment 1 participant performance by country.

Perf.	US	UK	Canada	Germany	Australia	Netherlands	Brazil	Other
Acc.	0.74	0.74	0.74	0.74	0.74	0.77	0.72	0.76
Prec.	0.77	0.77	0.77	0.77	0.76	0.80	0.76	0.80
Sample	1863	454	293	207	161	107	138	1234

TABLE 4. Experiment 1 and 2 participant performance by age and gender.

Experiment	Perf.	Overall	18-24	25-34	35-44	45-54	55+	Male	Female
1	Acc.	0.74	0.74	0.76	0.75	0.71	0.67	0.74	0.68
	Prec.	0.75	0.75	0.81	0.82	0.81	0.77	0.77	0.67
	Sample	4333	2936	1074	190	68	65	3879	456
2	Acc.	0.65	0.72	0.69	0.66	0.57	0.59	0.71	0.59
	Prec.	0.57	0.67	0.58	0.53	0.33	0.55	0.64	0.39
	Sample	315	95	104	65	27	29	232	85

Table 3 summarises the performance of users from different countries, which we observe to be almost identical across the world, with mean accuracy of 0.74 (with variance of 0.0002) and mean precision of 0.77 (with variance of 0.0002). For this reason, we did not consider the geographical factor in the second experiment. Also, this is advantageous when developing a prediction model to be applied across all populations. Table 4 summarises the performance of participants of experiments 1 and 2 based on age and gender. Here, we observe slightly more pronounced performance differences between the different groups. For example, we can see that female participants were less accurate and less precise (68%, 67%) than male participants (74%, 77%), which is in accordance with most of the related literature ([32]–[34]). We also observe that accuracy and precision are fairly consistent between the ages of 18 and 44, but drop in the 45+ age groups. Overall, the performance of the samples of participants in both experiments is largely coherent and consistent. The sample sizes of the groups that performed slightly worse in both experiments were relatively low. Moreover, they represent protected personal characteristics, and are thus impractical for our purposes. As we aim to develop prediction models suitable for use in a technical system, age and gender need to be omitted as candidate predictors because they do not satisfy the ethical criterion that we have set. For example, an organisation implementing security controls that are stricter or less strict based on age or gender would be seen as discriminatory. Overall, the reported performance of the participants provides no strong indication that omitting these demographic variables (geography, gender, age) would have a major impact on the chosen predictor features' accuracy and precision.

IV. STAGE 1: ASSOCIATION RULE MINING ON THE RESULTS OF EXPERIMENT 1

While performing prediction based on the large dataset collected in experiment 1 would be an attractive prospect, in practice, the high-level features used (described in Section II-B) would not be granular enough. Our attempts to produce prediction models solely based on them produced relatively low accuracy rates, just above the null rate for each exhibit. Instead, the primary objective of experiment 1 was to use it as a mechanism to determine which features should be explored further. For this purpose, we have performed association rule mining (ARM). ARM is a standard data mining methodology successfully employed in network intrusion detection [50], bioinformatics [51], recommender systems [52], social network advertising [53] and several other applications. It can help identify frequent itemsets (collections of attributes that frequently occur together) and association rules to determine whether strong relationships exist between two or more items.

As brief introduction to ARM, an association rule is composed of an itemset, which comprises an antecedent, consisting of one or more attributes and forming the "IF" of a rule, and a consequent, which forms the "THEN". The percentage of cases of an item's existence amongst frequent itemsets is referred to as *support*, while the conditional probability of observing a particular exhibit response under the condition that the participant attributes contain a particular set of participant attributes is referred to as *confidence*. Here, we employ the apriori algorithm [54] to create association rules by comparing frequent itemsets to a specified support/confidence threshold that determines the strength of the rule.

Using the *Arules* package in *R* [55], we have conducted frequent itemset discovery and association rule generation configuring a threshold for support larger than the system default and the default threshold for confidence, which are 0.15 and 0.8 respectively. For each association rule, we evaluate its importance using five commonly used metrics: support/confidence as the primary interest measure for each rule, as well as lift, coverage and odds ratio of each rule as individual measures of independence. For each metric's formula below, X refers to the frequent itemset attribute(s) that consist of the participant indicators defined in section III-B, forming the rule antecedent(s). The rules consequent Y defines a correct response to an attack exhibit, coded as RESPONSE=1 (i.e., for participants who classified particular exhibit correctly). In summary:

$$\text{Support: } supp(X \Rightarrow Y) = P(X \cup Y) \tag{3}$$

$$\text{Confidence: } conf(X \Rightarrow Y) = \frac{P(X \cup Y)}{P(X)} \tag{4}$$

$$\text{Lift: } lift(X \Rightarrow Y) = \frac{P(X \cup Y)}{P(X)P(Y)} \quad (5)$$

$$\text{Coverage: } cover(X \Rightarrow Y) = P(X) \quad (6)$$

$$\text{Odds Ratio: } \alpha(X \Rightarrow Y) = \frac{P(X)/1 - P(X))}{P(Y)/1 - P(Y)} \quad (7)$$

Note that on investigation, no rules were reported for RESPONSE=0 (i.e., for participants who classified particular exhibit incorrectly) to satisfy the support/confidence threshold 0.15/0.8. This indicates that within the data there exists a high degree of variability between participants who were susceptible and no distinguishable pattern between their attributes could be determined.

For lift, a value of 1.0 indicates independence of X and Y, while values greater than 1.0 indicate that participants with attributes X contain more correct attack exhibit responses Y (i.e., RESPONSE=1), than those without these attributes. An Odds Ratio of 1 indicates that Y is not associated to X, which is to say that an exhibit response is not related to the participant attributes.

Using the apriori algorithm, a total of 24 association rules were initially identified. These were then pruned by removing super rules of any other rule that has the same or higher lift. Pruning resulted in reduction from 24 to 10 association rules.

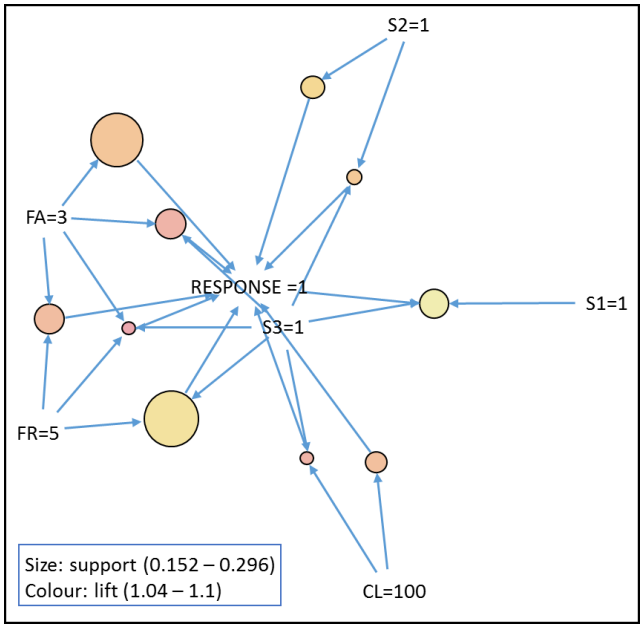

FIGURE 7. Association Rules graph with items and rules as vertices. The size of each circle linking vertices relates to the support of a rule, while the colour indicates the lift.

In Figure 7, the 10 association rules are shown where each item and vertex indicates the formation of a rule, where vertices leading to "RESPONSE=1" show the consequent of the rule. The size of each circle linking vertices is related to the support of a rule, while the colour indicates the lift. In Table 5, the importance measure of each rule is summarised, in order of confidence, lift and odds ratio (OR).

TABLE 5. Pruned association rules reported for participants with correct exhibit response.

Rules	Support	Confidence	Lift	Coverage	OR
S3=1,FR=5,FA=3	0.15	0.85	1.1	0.181	1.82
CL=100,S3=1	0.15	0.84	1.091	0.181	1.71
S3=1,FA=3	0.21	0.84	1.091	0.248	1.78
FR=5,FA=3	0.21	0.83	1.083	0.247	1.69
CL=100	0.18	0.83	1.08	0.212	1.62
FA=3	0.28	0.83	1.075	0.34	1.69
S2=1,S3=1	0.16	0.83	1.073	0.193	1.53
S2=1	0.18	0.82	1.059	0.224	1.42
S3=1,FR=5	0.3	0.81	1.049	0.366	1.42
S1=1,S3=1	0.2	0.8	1.04	0.252	1.27

The association rule with the highest lift indicates that participants who had had security training through self study and also used the type of platform shown in the exhibit daily and were very familiar with the exhibit platform itself were highly likely to correctly identify a semantic attack on this platform. In other words, the rule antecedent "S3=1, FR=5, FA=3" was reported by 18% of the total participants in the survey, where 15% of the total participants who also reported these attributes correctly identified a semantic attack "RESPONSE=1"; resulting in a 85% confidence that these participants were not susceptible. Of course, this was expected. In respect to odds ratio, participants with these attributes were 82% less likely to be susceptible to a semantic attack. Here, a lift value of 1.1 means that the participants who were not susceptible (RESPONSE=1), who have security training through self-study, use the target platform type daily and are very familiar with the specific platform, are observed 10% more than the percentage of the participants that were not susceptible in the total participant dataset. Within the 10 pruned rules, the appearance of frequently occurring items provides insight into association between specific attributes and reduced susceptibility to attacks. For example, familiarity with the specific platform provider (FA), frequency of access with a particular type of platform (FR), self-study (S3) and computer literacy (CL) are consistently reported attributes. On the contrary, duration of access (DR) and security awareness (SA) do not appear in the rules at the support/confidence threshold. Overall, the association rules indicate that security training through self-study, daily access to a type of platform and familiarity with a specific platform in this type category, as well high confidence with computer literacy are associated to reduced susceptibility to semantic attacks. However, given the minor variations in lift between these rules, the lack of 100% confidence in any rule and relatively low support, a large proportion of non-susceptible users without these attributes may not be represented. Therefore, employing these rules as classification criteria would likely result in susceptibility prediction that produces many false negatives. Instead, we utilise these findings to identify the attributes that we

TABLE 6. Predictor variables utilised in Experiment 2.

Feature	Description	Scale
FA1	Familiarity with provider platform	Not very familiar (1), Somewhat familiar (2), Very familiar (3)
FR1	Frequency of use for provider platform	Never (1), less than once a month (2), once a month (3), weekly (4), daily (5)
DR1	Duration of use for provider platform	None (1), < 30 min (2), 30 min to 1 h (3), 1-2 h (4), 2-4 h (5), >4 h (6)
ST1	Time since training for provider platform	Never (0), > 1 year (1), 6 months to 1 year (2), 3 to 6 months (3), 1 to 3 months (4), 2 weeks to 1 month (5), < 2 weeks (6)
FA2	Familiarity with platform type	Not very familiar (1), Somewhat familiar (2), Very familiar (3)
FR2	Frequency of use for platform type	Never (1), less than once a month (2), once a month (3), weekly (4), daily (5)
DR2	Duration of use for platform type	None (1), < 30 min (2), 30 min to 1 h (3), 1-2 h (4), 2-4 h (5), >4 h (6)
ST2	Time since training for platform type	Never (0), > 1 year (1), 6 months to 1 year (2), 3 to 6 months (3), 1 to 3 months (4), 2 weeks to 1 month (5), < 2 weeks (6)
S1_1	Formal Education through lectures	No (0), Yes (1)
S1_2	Formal Education through tests	No (0), Yes (1)
S1_3	Formal Education through coursework	No (0), Yes (1)
S1T	Time since formal education	Never (0), > 1 year (1), 6 months to 1 year (2), 3 to 6 months (3), 1 to 3 months (4), 2 weeks to 1 month (5), < 2 weeks (6)
S2_1	Work-based through tests	No (0), Yes (1)
S2_2	Work-based through videos	No (0), Yes (1)
S2_3	Work-based through games	No (0), Yes (1)
S2T	Time since work-based training	Never (0), > 1 year (1), 6 months to 1 year (2), 3 to 6 months (3), 1 to 3 months (4), 2 weeks to 1 month (5), < 2 weeks (6)
S3_1	Self-study through websites	No (0), Yes (1)
S3_2	Self-study through videos	No (0), Yes (1)
S3_3	Self-study through games	No (0), Yes (1)
S3T	Time since self-study training	Never (0), > 1 year (1), 6 months to 1 year (2), 3 to 6 months (3), 1 to 3 months (4), 2 weeks to 1 month (5), < 2 weeks (6)

should study in greater detail in stage 2 of our analysis, as presented in Section V.

V. STAGE 2: EXPERIMENT 2 MODEL AND FEATURE SELECTION

Following on from the initial investigation of each high-level feature studied in experiment 1, experiment 2 was conducted on a smaller participant base (315 respondents), where participants were tested on a new set of semantic attacks, consisting of screenshots, animations and videos, and were asked to provide considerably more detail on their profile. In experiment 2, we add further granularity to the high-level features identified in the ten ARM rules of Section IV. In detail, we extend FR and FA to include both specific provider platforms (FR1, FA1) in combination with types of platform (FR2, FA2). Also, we adapt and extend security training as follows: S1, S2 and S3 are converted from a binary answer to a length of time since last training for S1, S2 and S3, with a scale of: Never, over 1 year, up to 1 year, up to 6 months, up to 3 months, up to 1 month, up to 2 weeks. The second measures security training by platform types and specific provider platforms, including length of time since last training. Each high-level security category is also extended to include the training methods commonly used for each respective security category, such as: self study (S3) through online videos, formal education (S1) through coursework, etc. Features SA and CL are not altered. In order to identify whether features DR and SA are truly non-informative, redundant features, we include them in the model feature selection process alongside the newly expanded, granular feature-set; extending DR to specific provider platforms (DR1 - platform type, DR2 - platform provider). As a result of expanding and adapting the feature-set, we increase from 8 candidate predictors in experiment 1 to 22 in experiment 2, as summarised in Table 6.

With the adapted feature-set from experiment 2, using *R* [55] and the *Caret* package [56], we identify machine learning models that can predict a user's ability to detect attacks. Firstly, we select and compare two distinct machine learning algorithms; modelling both a linear and non-linear approach to prediction. Secondly, for each model we have applied automatic feature selection with sequential backward selection in Recursive Feature Elimination (RFE); obtaining an optimal model for each machine learning algorithm.

1) LOGISTIC REGRESSION Vs. RANDOM FOREST

For a user susceptibility model to be practically usable by a technical security system, it must employ predictor features that can be practically measured in real-time, automatically and ethically. To evaluate whether a linear model can be sufficient, we first employ logistic regression (LR), which performs well in linear spaces, functioning by definition as a special case generalised linear model using a Bernoulli distribution for a binary response [57]. LR is relatively robust to noisy data and over-fitted models, where the data contains high variance. In comparison to LR, another method that is resilient to variance in model predictions is a method known as bagging (also known as Bootstrap Aggregating [58]), where the algorithm produces replicates of the original data sample by creating new datasets by random selection with replacement. With each dataset, multiple new models are constructed and gathered to form an ensemble of models. Within the prediction process, all of the models in the ensemble are polled and the results are averaged to produce a result. Random forest (RF) is a popular bagging algorithm that can also be described as an ensemble decision tree classifier. In RF, a number of decision trees are trained with different re-sampled versions of an original dataset and then used to predict data that was omitted from each sample as an

embedded measure of training accuracy; this is called the out-of-bag error. Here, RF reduces the high variance inherent in a decision single tree by creating n trees that are averaged to reduce the variance of the final model [59]. Unlike LR, RF handles nonlinearity naturally. Predictor variables are randomly chosen at each decision split in the decision tree which results in a randomised, non-linear approach.

2) RECURSIVE FEATURE ELIMINATION

Employing the predictor features summarised in Table 6, for both LR and RF models, we have used an automatic feature selection method to identify the most informative predictor features and build a single prediction model for each individual attack exhibit. Recursive Feature Elimination (RFE) is an automatic backwards feature selection algorithm. It starts by fitting a model to all 22 features, ranking the latter based on their variable importance to the model, and gradually excluding the features with the lowest importance in each iteration, recursively considering smaller and smaller feature sets. In RF, variable importance is calculated within the model by recording the out-of-bag prediction accuracy for every predictor variable permutation in each decision tree. At each feature iteration, model accuracy is compared between the prior and permuted model, averaged over all trees and then normalised by the standard error. Since LR has no model-specific method to estimate importance, the *Caret* package conducts receiver operating characteristic (ROC) curve analysis on each feature iteration by evaluating the area under the ROC curve (AUROC), which is used as the variable importance for LR [60]. A ROC curve illustrates the performance of a binary classifier at different prediction probabilities by plotting the true positive rate (TPR) against the false positive rate (FPR) at various thresholds. AUROC represents the area under the ROC curve, where a random guess area of 0.5 (0,0 to 1,1) is typically used as the reference area from which to evaluate model performance. The result of LR is the selection of those features that have a statistically significant impact on the probability of a user's correct prediction.

One possible drawback of RFE is the potential for overfitting to predictor variables, as the procedure can focus on nuances in the sample data that may be anomalous and therefore not present in future data. For example, where predictors randomly correlate with the dependent variable being predicted, RFE may assign a good importance ranking to these variables, even if they were to make no practical sense. During training, this would indeed lower prediction error, but when validating the model on new data it might reveal that the predictors are actually non-informative, in a case referred to as ''selection bias'' [61]. To avoid this problem, and as is standard practice in supervised machine learning experiments and models, we have employed an outer layer of resampling through a repeated 10-fold cross-validation to provide a robust estimate of model feature-selection and test error as evaluated by RFE. Cross-validation (CV) is a model validation technique for assessing model performance on unseen, independent data sets and is an important tool for avoiding exaggerated model accuracy results (e.g. overfitting a model by testing it on the same data the model has been trained on). In the 10-fold CV process, the data sample is partitioned into 10 equal folds, where nine folds are used to train the model and the remaining one fold is used to test it. This process is repeated 10 times so that the model is tested on each fold in order to produce an average model test error, which in our case reports model test error at each variable selection step in RFEs backwards selection process. With repeated 10-fold CV, for each 10-fold training process, the process is repeated another 10 times.

In Figure 8, we present the results of LR and RF CV test error for each attack and the optimal set of predictors selected by RFE. We compare LR and RF with each other, as well as with a naive classifier, which, for each exhibit would always select the answer (0 or 1) that is the most common in the sampled population (the sample response rate). This is the maximum accuracy of a model that uses no features for predicting the sample population outcome. For five out of six attack exhibits, both LR and RF models reported superior classification accuracy than the sample response rate. RF outperformed LR in four of the six attack exhibits.

Table 7 shows where each feature was selected for an exhibit's final prediction model (whether it was the LR or the RF model that was best performing). We observe that frequency of access to the specific provider platform in the exhibit (FR1) was included in the best performing prediction model for 5 out of 6 exhibits, followed by length of time since security training through self-study and formal education, which appeared in 4 out of 6 exhibits' final models.

On the other end, familiarity with the exhibit platform type, security training with a particular platform type, security self-study through games, work-based through tests and formal education through lectures were not selected for any exhibit's final prediction model. Removing these five features, we prune the candidate-feature set from 22 to 17 within a final RFE model selection process process with the aim to build a final model for susceptibility prediction. In order to build a prediction model that can potentially be employed across any platform and with any semantic attack, we combine each of the exhibits' sample responses into a stacked data sample, where all users' responses are included in a single dataset for all attacks. So, the values for each feature relate to the particular attack's settings in a particular entry in the dataset. For instance, the feature ''familiarity with platform type'' in an attack that utilises Facebook would refer to the familiarity with platform of type ''social network''. This approach enables the construction of a single model that contains a range of semantic attacks, platform types and specific provider platforms. Creating a single model for each attack would be impractical, as we would need one model for each platform/attack combination. Training a model based on a wide-range of disparate platforms and attacks, and using a combined dataset for a single prediction response, enables more widely applicable prediction of susceptibility that can be utilised in a technical security system.

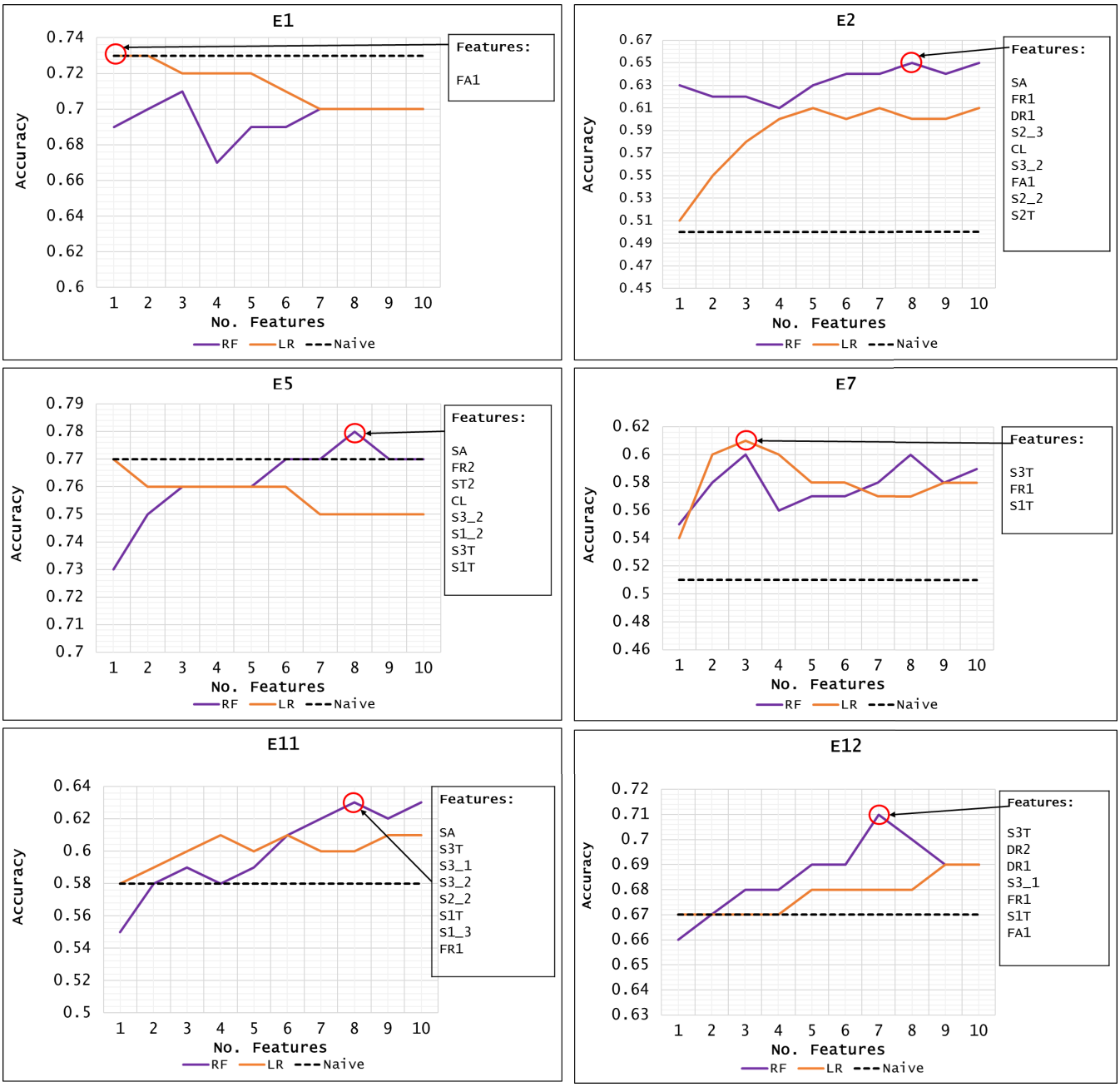

FIGURE 8. Each graph presents the results of 10 times repeated 10-fold cross-validation for each exhibit, using recursive feature elimination with a LR and a RF model. Results are presented in the form of overall test accuracy.

A. SUSCEPTIBILITY MODEL: RESULTS AND ANALYSIS

A reliable and widely applicable user susceptibility model can have several applications as part of a defence mechanism against semantic attacks. It can help predict a specific user's a) degree of "susceptibility" to semantic attacks (likelihood of being deceived by one), or equivalently b) expected performance if they were to act as a human classifier (likelihood of spotting attacks). The former can help a security system identify whether a user is particularly susceptible to semantic attacks and consequently whether the system environment needs to adapt accordingly (e.g., by privilege adjustment, targeted warnings, security enforcing functionality, etc.). The latter can help evaluate to what extent a user can be relied upon as a "Human as a Security Sensor" (HaaSS) of semantic attacks, where user reports are taken into account so as to strengthen an organisation's cyber situational awareness.

For both applications, it is important to measure the model's performance based on its general accuracy in predicting which participants will detect the attacks and which will not, and secondly its ability to reduce false positives or false

TABLE 7. Total number of times each predictor feature is selected for an attack exhibit's best performing model.

Exhibits	FR1	FR2	FA1	FA2	DR1	DR2	ST1	ST2	SA	CL	S3T	S2T	S1T	S3_1	S3_2	S3_3	S2_1	S2_2	S2_3	S1_1	S1_2	S1_3
E1	✓	✓	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
E2	✓	-	✓	-	✓	-	-	-	✓	✓	-	✓	-	-	✓	-	✓	-	✓	-	-	-
E5	-	✓	-	-	-	-	✓	-	✓	✓	✓	-	✓	-	✓	-	-	-	-	-	✓	-
E7	✓	-	-	-	-	-	-	-	-	-	✓	-	✓	-	-	-	-	-	-	-	-	-
E11	✓	-	-	-	✓	-	-	-	✓	-	✓	-	✓	✓	✓	-	-	-	-	-	-	✓
E12	✓	-	✓	-	-	✓	-	-	-	-	✓	-	✓	✓	-	-	-	-	-	-	-	-
Total	5	2	2	0	2	1	1	0	3	2	4	1	4	2	3	0	1	0	1	0	1	1

TABLE 8. Prediction performance comparing the accuracy, precision, false positive and false negative detection of the final LR and RF models against the Naive classifier.

Classifier	Accuracy	Precision	FN	FP	P value	Features
LR	0.68	0.70	0.12	0.20	<0.001	7
RF	0.71	0.73	0.12	0.18	<0.001	16
Naive	0.59	0.59	0	0.41	0.5	0

TABLE 9. Feature odds ratios for logistic regression model.

Features	Intercept	FR1	CL	S3T	S2_2	DR2	S1T	S3_1
OR	0.20	1.22	1.01	1.10	0.57	1.13	1.06	1.29

negatives by using a probability cut-off threshold. Table 8 compares the LR and RF classifiers' overall performance against the naive classifier, which always selects the answer with the highest probability in the population sample (so, always 0 if population's success rate is below 50% and always 1 otherwise, for a given exhibit). The test split used for classification consisted of 215 correct (1) and 147 incorrect (0) responses. In Table 10, the predictors selected by the RFE process for the LR and RF models are presented. For the RF model, to evaluate feature variable importance, as metric, we use the reduction in out-of-bag error during the model training process. For the LR model, we use the increase in AUROC.

B. KEY OBSERVATIONS

Both the LR and RF models satisfy the statistical significance threshold of 0.05 and both appear to outperform comfortably the naive classifier, which is a good sign in terms of their practical applicability. There is a slight advantage of RF over LR across all metrics (higher accuracy and precision, and lower false positives), but this comes at the expense of practicality, because it requires a large number of features to be monitored (16 against LR's only 7). Moreover, as RF employs a black box modelling approach, this makes it less interpretable than the LR model as to why each feature within the model informs prediction. In LR, interpretation is more straightforward, because it produces each feature's odds ratio (OR), which is the increase in the probability of a user correctly identifying an attack for every one unit increase in that feature's scale, when all other features remain fixed. For example, from table 9, we see that a unit increase in the scale of frequency of use (e.g., from once a month to weekly), increases the probability of correct detection by 22%. So, ORs can also be used to cross-reference with variable importance in interpreting each feature's influence to the prediction outcome.

As one would have expected, security training does make a difference, with all three forms (formal education, work-based training and self-study) appearing in some form in both models. In general, we observe that the length of time since last training (whether self-study, formal education or work-based training) is particularly important, with time since last self-study (S3T) appearing to be overall the most important in the training category. This is reasonable, because semantic attacks evolve continuously and any guidelines or technical information learned in training needs to be updated often. Five years ago, semantic attacks were almost entirely based on generic phishing and ransomware. Today, watering holes, WiFi evil twins and socia media friend injection attacks have become the norm, and phishing has expanded to all forms of user interaction, from Quick Response (QR) codes, to near-field communication (NFC) and Bluetooth [1]. Interestingly, formal security education through lectures was not chosen as a useful predictor of susceptibility to semantic attacks by any of the models and for any of the exhibits.

Frequency of access to the specific provider's platform (FR1) rather than generally to the type of platform (e.g., specifically Facebook rather than generally social networks) was shown to reduce susceptibility noticeably, being the fourth most important variable in RF and the first in LR. Frequency of access to the general type of platform (FR2) was utilised by RF as one of the features with the lowest variable importance (0.06), and was not utilised at all by LR.

Duration of access to the same platform type was important in both models, with 13% increase in the probability for each unit increase in the LR odds ratio. In RF, frequency and duration was also important for the platform type. Also, in the RF model familiarity with the platform provider was the fifth highest important variable.

Computer literacy (CL) was shown to be the most important feature for RF and the second most important for LR.

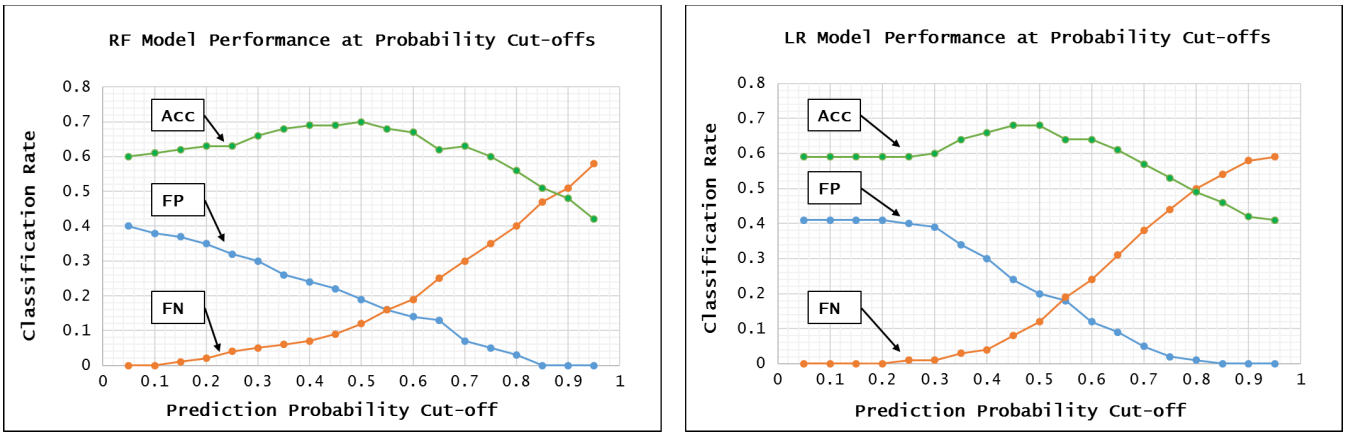

FIGURE 9. RF (left) and LR (right) model performance for false positive (not susceptible), false negative (susceptible) prediction and overall prediction accuracy at each probability cutoff.

TABLE 10. RF and LR model predictor features selected through recursive feature selection (in order of variable importance measure: decrease in out of bag error rate for RF, and increase in AUROC for LR). The higher the variable importance (in brackets), the more important the feature is to its model.

Model	Selected Features
Random Forest	CL (0.114), S3T (0.112), SA (0.107), FR1 (0.099), FA1 (0.097), S3_1 (0.087), S1T (0.073), S2T (0.073), S2_2 (0.068), FR2 (0.06), DR1 (0.05), S2_3 (0.048), S3_2 (0.048), DR2 (0.046), ST2 (0.039), S1_3 (0.033)
Logistic Regression	FR1 (0.035), CL (0.032), S3T (0.029), S2_2 (0.027), DR2 (0.023), S1T (0.021), S3_1 (0.018)

This reinforces the need for a mechanism to monitor and record computer literacy as a gauge of an organisation's cyber risk.

Unlike the high level predictors Security Awareness (SA) and Duration of Access (DR), which were not included in the association rules in experiment 1, the RF model included both Security Awareness and the expanded DR features: duration of using a specific platform provider (DR1) and specific platform type (DR2), whereas the LR model included DR1 only. Surprisingly, in RF, SA was the third most important feature, whereas in LR it was not included at all; as per the association rules. On the other hand, more in line with the association rules omitting DR in experiment 1, both DR1 and DR2 were were given relatively low variable importance in RF, placing 11th and 15th out of the total 17 features, respectively, with DR2 placing 5th out of a total 7 features for LR and DR1 omitted from the models feature-set. For both models, SA, DR1 and DR2 were given a lower degree of variable importance than all other features that were expanded from their higher level counterparts (FR, FA, CL, S3, S2, S1) reported in experiment 1's association rules frequent item sets; with the exception of time since last security training through formal education (S1T), which slightly less important than DR2 (0.21 compared to 0.23) in the LR model. The indication is that the original high level predictors show a consistent association with reduced susceptibility across both experiments, even after adding further granularity to their measurement scale and context, and as a result also gained sufficient predictive power for determining the probability of a participants susceptibility to semantic attacks with a reasonable degree of accuracy.

There is no doubt that a user's susceptibility to semantic social engineering attacks depends also on personality traits, social context, psychological state and other human and contextual factors, which are, however, impractical, as they cannot be measured in real-time, automatically or ethically. Without knowledge regarding these factors, one cannot expect a highly accurate prediction of susceptibility. So, the accuracy improvement of around 10% against the naive classifier achieved here is significant. In practice, we have developed this method to act as a baseline for an organisation's technical security system, which can then adapt over time, as it learns the characteristics of the organisation's own users.

Equally significant is that one can utilise these models to identify an appropriate probability threshold depending on preference in minimising false positives, minimising false negatives or maximising accuracy (Figure 9). By probability threshold, we refer to the value over which a technical security system should consider a user to be susceptible to a semantic attack. For instance, if the aim were to maximise accuracy, the probability threshold for determining whether a user is susceptible or not, should be 0.5 for both models. However, it would be 0.55 if the aim were to keep both false positives and false negatives below 0.2. Overall, RF appears to perform slightly better than LR in terms of false positives at low probability thresholds, but is slightly worse at higher

probability thresholds. For false negatives, the reverse is observed. In an organisation that is tolerant of false positives, but not tolerant of false negatives, to keep the false negatives below 0.02, both LR and RF models would yield a false positive rate just under 0.4. For RF, this would correspond to a probability cut-off of 0.15, and for LR to a cut-off of 0.3. For an organisation that is tolerant of false negatives, false positives can be effectively avoided using the RF model at a 0.85 cutoff, but this results in an approximate 20% decrease in overall classification accuracy, with the number of false negatives increasing to 0.48.

VI. DISCUSSION

A. LIMITATIONS

In our exhibit-based experiment there are a few limitations that must be considered. Participants were primed to the purpose of the survey and subsequent test and thus may have been more vigilant and sensitive to a semantic attack's deception (therefore weakening its effect) than they would have normally been.

For the first experiment, the simple approach of using screenshots to represent the exhibits was very useful in conducting a large-scale study online and on any computer platform. However, the use of screenshots is more appropriate for DV1 attacks that rely on cosmetic deception than for DV2 (and partially DV3) that rely on behavioural deception, which is less straightforward to convey via screenshots. To address this, in the second experiment we included video exhibits where behavioural deception can be more accurately emulated (in terms of context and system behaviour), rather than depicted visually.

Potential limitations may also exist in the selection of features for our susceptibility model. We have focused on a number of high-level concepts with the aim to create a model for predicting susceptibility that is applicable across a wide range of semantic attacks. One example is computer security training where we focus on the type and mode of delivery of security training rather than its content. Prediction taking into account the content too would have probably been more accurate, but would presume that an organisation can collect such detailed information for its users, which may be impractical.

B. CHALLENGES IN PRODUCING DATASETS FOR SEMANTIC SOCIAL ENGINEERING SUSCEPTIBILITY PREDICTION

Real-world, authoritative datasets for user susceptibility to semantic attacks are not available. An organisation may not publicly reveal that their business has been exploited because of the perceived reputational damage it could cause or simply because employees fail to report breaches for fear of disciplinary actions. Security authorities and organisations, such as Semantec [62], who actively publish data from those businesses, and users who do report attacks tend to anonymise and censor the data to a point that profiling information that could show context leading to an attack is removed before being made publicly accessible. Therefore, development of user datasets through research experiments is necessary to understand which behaviours and identifying factors help determine susceptibility and thus inform the design and development of new security mechanisms against semantic attacks. In this section, we identify a number of persisting problems for the development of robust semantic attacks datasets:

- **Ethics**. A prevalent limitation for access to user susceptibility data is ethics. Ethical consideration and approval can be a barrier to the collection of rich user data for aiding researchers and developers in the development of user-centric defences against semantic attacks. Experiments with human participant require ethics approval from an institutional of governmental review board, and therefore there are often a number of requisite requirements which limit researchers ability to produce truly representative results. For example, in [63], participant deception and debriefing, privacy and institute review board approval were determined to be the main challenges that affect the design and execution of phishing experiments. Mouton et al. [64] proposes a normative perspective for ethics in social engineering which can help ethics committees in the process of experiment approval. Here, reporting susceptibility would be considered from a utilitarian and deontogical standpoint; that is, whether or not the collected and reported data would be ethical given the consequences of the specified action (utilitarianism) or the duty and obligations related to that action (deontology). In [65], researchers developed what has become widely accepted approach for designing ethical social engineering experiments, but the method proposed focuses solely on phishing emails and it is unclear how it can be extended to a wider range of semantic attacks and platforms other than email.

  As well as ethics approval, semantic attack research poses legal implications [66], where researchers are increasingly conducting phishing experiments without the knowing consent of participants. In this case, the data collected may prove more representative of natural user behaviour, but cannot be validated as legitimate research without formal approval.

  One approach towards tackling this fundamental problem in the research of semantic social engineering attacks is to provide a platform that enables users and organisations to anonymously report semantic attacks, without omitting crucial contextual information such as whether the attack was successful or not, the scenario in which the attack occurred, whether or not the target had been trained, etc. This database of user susceptibility information would provide an invaluable resource for researchers seeking to analyse trends or predict behaviour to semantic attacks. Most importantly, collection of data in this format removes the complexity

and damaging effect on user experiment data that ethics approval may require.

- **The experiment population against data collection detail tradeoff**. Participants in semantic attack research tend to be recruited from the institution in which the study is conducted (e.g., university students, organisation's own staff, etc.) and often this is noted as a limitation of the research as the results may not be representative the wider target population. This poses a major problem for empirically proving the validity of research outcomes. In the first study, we recruited a large number of participants from multiple different geographical locations on the Internet, but this approach limits the ability to collect more detailed data from the participants. There is a tradeoff to be considered when recruiting participants that are more representative of the user base against the qualitative data that can be extracted from a user population that is easily accessible. In the case of the former, collecting user responses from a large number of disparate demographic backgrounds is fairly simple when the Internet is the recruitment platform, but these participants cannot be easily observed or interviewed at any stage of the research. For the latter, researchers have localised access to participants and therefore a higher degree of detail regarding user behaviour can be recorded and analysed.
  Ultimately, semantic attack research is affected by both circumstances and as such context should dictate the most suitable approach. In our study, it was more important to recruit sufficient numbers of participants to allow the evaluation of statistical machine learning models. For research focusing on psychological impact of exploitation from semantic attacks, detailed qualitative data may be a more relevant goal, in which case research would most likely benefit from a smaller population.
- **Attack coverage**. In table 1, the majority of research related to predicting susceptibility to social engineering attacks has focused on phishing, which is only one type of semantic attack. Conclusions made from research solely reliant phishing experiments may not be applicable to the wider problem space. Like traditional network and operating attacks, there are many types of semantic attack, crossing multiple platforms, and therefore like an anti-virus for OS or firewall for the network, it is crucial that experiments consider and evaluate a wide range of semantic attacks in order to build defence systems that can mitigate multiple threats. Furthermore, specific attacks may become less popular over time as new platforms emerge or more successful techniques are developed, and therefore it is also important that an experiments results remain relevant for addressing future attacks.
  In [1], the taxonomy for semantic attacks can provide a useful baseline to build such experiments that measure user susceptibility across a series of generic attack attributes. In this study, we have included *Deception Vector* only for clustering attacks on the same and different platforms, simplifying the modelling process and ability to classify susceptibility with a single, general model. For research aiming to understand user vulnerabilities to removable media or targeted cyber-supply chains, other items of the taxonomy such as *Method of Distribution* and *Target Description* may provide useful categories for clustering a wide range of attacks in a single experiment.
- **Lack of an authoritative archive**. Repositories of historic and current phishing emails and websites do exist [67]–[69], but do not cover the wider range of semantic attacks and do not include data on the profiles of the users who have or have not been deceived by them. An open archive of semantic attacks and corresponding user profile data would be immensely helpful to researchers in this field.

C. HUMAN AS A SECURITY SENSOR (HaaSS)

The concept of the *human as a sensor* has been used extensively and successfully for the detection of threats and adverse conditions in physical space, for instance to detect noise pollution [70], monitor water availability [71], detect unfolding emergencies [72] etc. In relation to semantic attack threats, the concept is very new. There is one example specifically for phishing attacks [36]. We argue that the concept can be explored much further and for most semantic attacks, where the human user's situational knowledge can help detect attacks that are otherwise largely undetectable by technical security systems. For example, there are no known technical countermeasures to attack E11 in experiment 1 (''Qrishing'') and attack E1 in experiment 2 (Video masquerading ''clickbait''), but in our experiments, users were able to detect them with a probability of 86% and 78% respectively (see Table 2). This is certainly not a rigorous way for evaluating HaaSS, but we feel is an indication of its potential. Introducing a HaaSS element in an organisation's security can empower users to become its strongest link. In this context, predicting the performance of an individual user as human sensor of semantic social engineering attacks is the equivalent of measuring the reliability of a physical sensor. For example, within a HaaSS reporting platform, a prediction model that measures the probability of a user's report being correct can provide security engineers with the ability to triage the review of reports; prioritising the ones from users that are more accurate human sensors.

VII. CONCLUSION

We have conducted two experiments, each consisting a survey and an exhibit-based test, asking participants to identify whether specific exhibits were likely to show attacks or not. Based on the data collected, we identified a set of features

from which we produced logistic regression and random forest models for predicting susceptibility to semantic attacks, with accuracy rates of .68 and .71 respectively. The slight performance advantage of RF over LR is countered by the larger number of features that it requires to be monitored (16 against LR's 7). In terms of the features themselves, we observe that security training makes a noticeable difference in a user's ability to detect deception attempts, with frequent self-study appearing to be a key differentiator. Yet, formal security education through lectures was not chosen as a useful predictor by any of the models and for any of the exhibits. More important features were computer literacy, familiarity and frequency of access to a specific platform. The models developed can be configured in terms of preference in minimising false positives, minimising false negatives or maximising accuracy, based on the probability threshold over which a user would be deemed to be susceptible to an attack. For both models, a threshold choice of 0.55 would keep both false positives and false negatives below 0.2.

We have also identified a number of challenges associated with developing datasets for predicting susceptibility to semantic attacks, where addressing these challenges can help produce rich and representative user susceptibility data that can aid developers and researchers of user security defence systems. In future work, our model can be experimentally validated with a technical implementation and using a wider range of semantic attacks for each deception vector in order to provide empirical results for the model's performance in practice. As deception-based attacks utilised in the wild evolve continuously, the baseline model and classification rules can be continuously improved with new training data from different user populations and attack types.

Furthermore, the advent of the Internet of Things [73] promises to compound the problem and extend to physical impact, exposing user interfaces of systems previously inaccessible to the standard user, let alone via a distributed application in the Internet [74]. The more effective such cyber-physical attacks prove, the more the deception attack surface will grow. Semantic social engineering threats in the Internet of Everything are likely to expand attack surfaces via ubiquitous connectivity which practically facilitate new and convincing semantic attacks; the impact of a phishing email may no longer be limited to stolen user credentials or malware infection, but can also bring down a national power-grid [75]. Providing users with the ability to report suspected semantic attacks can help provide system developers and security practitioners with key insights in how to design or update systems to mitigate such threats, while at the same time instilling users with a sense of empowerment in protecting their technological environment. To this end, report credibility provides a crucial role in identifying the likelihood that an attack has indeed occurred, so as to prioritise reports and utilise their information to augment defence mechanisms. Predicting user susceptibility as a performance measure of semantic social engineering attack reporting provides a first step towards this vision.

REFERENCES

[1] R. Heartfield and G. Loukas, "A taxonomy of attacks and a survey of defence mechanisms for semantic social engineering attacks," *ACM Comput. Surv.*, vol. 48, no. 3, Feb. 2016, Art. no. 37.

[2] M. Huber, M. Mulazzani, E. Weippl, G. Kitzler, and S. Goluch, "Friend-in-the-middle attacks: Exploiting social networking sites for spam," *IEEE Int. Comput.*, vol. 15, no. 3, pp. 28–34, May/Jun. 2011.

[3] R. Heartfield and G. Loukas, "On the feasibility of automated semantic attacks in the cloud," in *Computer and Information Sciences III*. London, U.K., Springer, Oct. 2013, pp. 343–351.

[4] G. Madlmayr, J. Langer, C. Kantner, and J. Scharinger, "Nfc devices: Security and privacy," in *Proc. 3rd Int. Conf. Availability, Rel. Secur. (ARES)*, Mar. 2008, pp. 642–647.

[5] C. Whittaker, B. Ryner, and M. Nazif, "Large-scale automatic classification of phishing pages," in *Proc. NDSS*, 2010.

[6] 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. (TISSEC)*, vol. 14, no. 2, Sep. 2011, Art. no. 21.

[7] Webroot. (2013). *Webroot Real-Time Anti-Phishing Service*. [Online]. Available: http://www.webroot.com/shared/pdf/WAP-Anti-Phishing-102013.pdf

[8] M. Egele, P. Wurzinger, C. Kruegel, and E. Kirda, "Defending browsers against drive-by downloads: Mitigating heap-spraying code injection attacks," in *Detection Intrusions Malware, Vulnerability Assessment*. Berlin, Germany, Springer, 2010, pp. 88–106.

[9] L. Lu, V. Yegneswaran, P. Porras, and W. Lee, "Blade: An attack-agnostic approach for preventing drive-by malware infections," in *Proc. 17th ACM Conf. Comput. Commun. Secur.*, 2010, pp. 440–450.

[10] A. Calder and S. Watkins, *IT Governance: An International Guide to Data Security and ISO27001/ISO27002*. London, U.K.: Kogan Page, 2012.

[11] P. Kumaraguru *et al.*, "School of phish: A real-world evaluation of anti-phishing training," in *Proc. 5th Symp. Usable Privacy Secur.*, Jul. 2009, Art. no. 3.

[12] G. N. A. Arachchilage, S. Love, and M. Scott, "DesigningDesigning a Mobile Game to Teach Conceptual Knowledge of Avoiding 'Phishing Attacks'" *Int. J. e-Learn. Secur.*, vol. 2, nos. 1–2, pp. 127–132, Jun. 2012.

[13] A. P. Felt, R. W. Reeder, H. Almuhimedi, and S. Consolvo, "Experimenting at scale with google chrome's ssl warning," in *Proc. 32nd Annu. ACM Conf. Human Factors Comput. Syst.*. 2014, pp. 2667–2670.

[14] M. Khonji, Y. Iraqi, and A. Jones, "Phishing detection: A literature survey," *IEEE Commun. Surv. Tuts.*, vol. 15, no. 4, pp. 2091–2121, 4th Quart, 2013.

[15] K. Rieck, T. Holz, C. Willems, P. Düssel, and P. Laskov, "Learning and classification of malware behavior," in *Proc. Int. Conf. Detection Intrusions Malware, Vulnerability Assessment*, 2008, pp. 108–125.

[16] A. Filippoupolitis, G. Loukas, and S. Kapetanakis, "Towards real-time profiling of human attackers and bot detection," in *Proc. 7th Int. Conf. Cybercrime Forensics Edu. Training*, Canterbury, U.K., Jul. 2014.

[17] S. Kapetanakis, A. Filippoupolitis, G. Loukas, and T. A. Murayziq, "Profiling cyber attackers using case-based reasoning," in *Proc. 19th UK Workshop Case-Based Reasoning*, Cambridge, U.K., Dec. 2014.

[18] M. Kandias, V. Stavrou, N. Bozovic, and D. Gritzalis, "Proactive insider threat detection through social media: The youtube case," in *Proc. 12th ACM Workshop Privacy Electron. Soc.*, Nov. 2013, pp. 261–266.

[19] R. R. McCrae and O. P. John, "An introduction to the five-factor model and its applications," *J. Pers.*, vol. 60, no. 2, pp. 175–215, 1992.

[20] R. B. Cialdini, *Influence: Science and practice*. Boston, MA, USA: Allyn and Bacon, 2001.

[21] T. Halevi, J. Lewis, and N. Memon, "A pilot study of cyber security and privacy related behavior and personality traits," in *Proc. 22nd Int. Conf. World Wide Web Companion*, May 2013, pp. 737–744.

[22] T. Halevi, N. Memon, and O. Nov. (Jan. 2015). *Spear-Phishing in the Wild: A Real-World Study of Personality, Phishing Self-Efficacy and Vulnerability to Spear-Phishing Attacks* [Online]. Available: http://dx.doi.org/10.2139/ssrn.2544742

[23] J. G. Mohebzada, A. E. Zarka, A. H. Bhojani, and A. Darwish, "Phishing in a university community: Two large scale phishing experiments," in *Proc. Int. Conf. Innov. Inf. Technol.(IIT)*, Cambridge, U.K., Apr. 2012, pp. 373–382.

[24] M. Workman, "Wisecrackers: A theory-grounded investigation of phishing and pretext social engineering threats to information security," *J. Amer. Soc. Inf. Sci. Technol.*, vol. 59, no. 4, pp. 662–674, Feb. 2008.

[25] I. M. A. Alseadon, "The impact of users characteristics on their ability to detect phishing emails," Ph.D. dissertation, Queensland Univ. Technol. Brisbane, QLD, Australia, May 2014. [Online]. Available: http://eprints.qut.edu.au/72873/1/Ibrahim%20Mohammed%20A_Alseadoon_Thesis.pdf

[26] A. Karakasiliotis, S. M. Furnell, and M. Papadaki, "Assessing end-user awareness of social engineering and phishing," in *Proc. 7th Australian Inf. Warfare Secur. Conf.*, 2006.

[27] J. S. Downs, M. B. Holbrook, and L. F. Cranor, "Decision strategies and susceptibility to phishing," in *Proc. Second Symp. Usable Privacy Secur.*, 2006, pp. 79–90.

[28] J. S. Downs, M. Holbrook, and L. F. Cranor, "Behavioral response to phishing risk," in *Proc. Anti-Phishing Work. groups 2nd Annu. eCrime Researchers Summit*, Oct. 2007, pp. 37–44.

[29] M. Jakobsson, A. Tsow, A. Shah, E. Blevis, and Y. K. Lim, "What instills trust? A qualitative study of phishing," in *Financial Cryptography Data Security*. Berlin, Germany, Springer, 2007, pp. 356–361.

[30] S. Egelman, L. F. Cranor, and J. Hong, "You've been warned: An empirical study of the effectiveness of web browser phishing warnings," in *Proc. SIGCHI Conf. Human Factors Comput. Syst.*, 2008, pp. 1065–1074.

[31] A. Vishwanath, T. Herath, R. Chen, J. Wang, and H. R. Rao, "Why do people get phished? Testing individual differences in phishing vulnerability within an integrated, information processing model," *Decision Support Syst.*, vol. 51, no. 3, pp. 576–586, 2011.

[32] 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 *Proc. SIGCHI Conf. Human Factors Comput. Syst.*, Cambridge, U.K., Apr. 2010, pp. 373–382.

[33] M. Blythe, H. Petrie, and J. A. Clark, "F for fake: Four studies on how we fall for phish," in *Proc. SIGCHI Conf. Human Factors Comput. Syst.*, May 2011, pp. 3469–3478.

[34] J. Hong, "The state of phishing attacks," *Commun. ACM*, vol. 55, no. 1, pp. 74–81, Jan. 2012.

[35] M. Renaud, "System and method for dynamically assessing security risks attributed to a computer user's behavior," U.S. Patent WO2 007 147 266 A1, Dec. 27, 2007.

[36] N. Stembert, A. Padmos, M. S. Bargh, S. Choenni, and F. Jansen, "A study of preventing email (Spear) phishing by enabling human intelligence," in *Proc. IEEE, Intell. Secur. Inf. Conf. (EISIC)*. Sep. 2015, pp. 113–120.

[37] G. Britain, *Equality Act 2010*. Rue Hoche, France: Editions de l'Atelier, 2010.

[38] L. Weber and E. Dwoskin, "Are workplace personality tests fair?" *Wall Street J.*, Sep. 2014. [Online]. Available: http://www.wsj.com/articles/are-workplace-personality-tests-fair-1412044257

[39] R. T. Wright and K. Marett, "The influence of experiential and dispositional factors in phishing: An empirical investigation of the deceived," *J. Manage. Inf. Syst.*, vol. 27, no. 1, pp. 273–303, 2010.

[40] N. Davinson and E. Sillence, "It won't happen to me: Promoting secure behaviour among internet users," *Computers Human Behavior*, vol. 26, no. 6, pp. 1739–1747, Nov. 2010.

[41] C. Wagner, S. Mitter, C. Körner , and M. Strohmaier, "When social bots attack: Modeling susceptibility of users in online social networks," *Making Sense Microposts*, vol. 2, pp. 41–48, 2012.

[42] J. Wang, T. Herath, R. Chen, A. Vishwanath, and H. R. Rao, "Research article phishing susceptibility: An investigation into the processing of a targeted spear phishing email," *IEEE Trans. Prof. Commun.*, vol. 55, no. 4, pp. 345–362, Dec. 2012.

[43] P. A. Wang, "Assessment of cyber security knowledge and behavior: An anti-phishing scenario," in *Proc. IEEE Int. Conf. Internet Monitor. Protection (ICIMP)*, 2013, pp. 1–7.

[44] R. T. Wright, M. L. Jensen, J. B. Thatcher, M. Dinger, and K. Marett, "Research note-influence techniques in phishing attacks: An examination of vulnerability and resistance," *Inf. Syst. Res.*, vol. 25, no. 2, pp. 385–400, 2014.

[45] M. Alsharnouby, F. Alaca, and S. Chiasson, "Why phishing still works: User strategies for combating phishing attacks," *Int. J. Human-Comput. Stud.*, vol. 82, pp. 69–82, Oct. 2015.

[46] T. Kelley and B. I. Bertenthal, "Attention and past behavior, not security knowledge, modulate usersŠ decisions to login to insecure websites," *Inf. Comput. Secur.*, vol. 24, no. 2, pp. 164–176, 2016.

[47] D. Colardyn and J. Bjornavold, "Validation of formal, non-formal and informal learning: Policy and practices in EU member states," *Eur. J. Edu.*, vol. 39, no. 1, pp. 69–89, Mar. 2004.

[48] B. D. Cullity, *Elementary Signal Detection Theory*. Los Angeles, CA, USA: Oxford Univ. Press, 2001.

[49] H. Abdi, "Signal detection theory (SDT)," in *Encyclopedia of Measurement and Statistics*. Thousand Oaks, California: Sage Pub., 2007, pp. 886–889.

[50] L. Li, D. Z. Yang, and F. C. Shen, "A novel rule-based intrusion detection system using data mining," in *Proc. IEEE 3rd Int. Conf. Comput. Sci. Inf. Technol. (ICCSIT)*, Jul. 2010, pp. 169–172.

[51] J. Nahar, T. Imam, K. S. Tickle, and Y. P. P. Chen, "Association rule mining to detect factors which contribute to heart disease in males and females," *Expert Syst. Appl.*, vol. 40, no. 4, pp. 1086–1093, Mar. 2013.

[52] W. Lin, S. A. Alvarez, and C. Ruiz, "Efficient adaptive-support association rule mining for recommender systems," *Data Mining Knowl. Discovery*, vol. 6, no. 1, pp. 83–105, Jan. 2002.

[53] W. S. Yang, J. B. Dia, H. C. Cheng, and H. T. Lin, "Mining social networks for targeted advertising," in *Proc. 39th Annu. Hawaii Int. Conf. Syst. Sci. (HICSS)*, Jan. 2006, pp. 137–137.

[54] P. N. Tan, M. Steinbach, and V. Kumar, "Association analysis: basic concepts and algorithms," in *Proc. Introduction Data Minin*, 2002, pp. 327–414.

[55] R. Ihaka and R. Gentleman. *The R Project for Statistical Computing*. (2016). [Online]. Available: https://www.r-project.org/

[56] M. Kuhn, "Caret package," *J. Statist. Softw.*, vol. 28, no. 5, 2008.

[57] D. W. Hosmer and S. Lemeshow, Eds., *Applied Logistic Regression*. Hoboken, NJ, USA: Wiley, 2004.

[58] L. Breiman, "Bagging predictors," *Mach. Learn.*, vol. 24, no. 2, pp. 123–140, 1996.

[59] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.

[60] M. Kuhn. (2012). *Variable Selection Using the Caret Package*. [Online]. Available: http://cran.cermin.lipi.go.id/web/packages/caret/vignettes/caretSelection.pdf

[61] C. Ambroise and G. J. McLachlan, "Selection bias in gene extraction on the basis of microarray gene-expression data," in *Proc. Nat. Acad. Sci.*, vol. 99, no. 10, pp. 6562–6566, 2002.

[62] S. Corporation. (2015). *Internet Security Threat Report*. [Online]. Available: http://www.symantec.com/security_response/publications/threatreport.jsp

[63] P. Finn and M. Jakobsson, "Designing and conducting phishing experiments," *IEEE Tech. Soc. Mag.*, vol. 26, no. 1, pp. 46–58, 2007.

[64] F. Mouton, M. M. Malan, and H. S. Venter, "Social engineering from a normative ethics perspective," in *Proc. IEEE Inf. Secur. South Africa*, Aug. 2013, pp. 1–8.

[65] M. Jakobsson and J. Ratkiewicz, "Designing ethical phishing experiments: A study of (Rot13) rOnl query features," in *Proc. 15th Int. Conf. World Wide Web*, 2006, pp. 513–522.

[66] C. Soghoian, "Legal risks for phishing researchers," in *Proc. eCrime Res. Summit*, Oct. 2008, pp. 1–11.

[67] Phishtank. (2015). *Out of the web and Into the Tank*. [Online]. Available: https://www.phishtank.com/

[68] Scamdex. (2015). *The Internet Scam Resource*. [Online]. Available: http://www.scamdex.com/

[69] Millersmiles. (2015). *The Web's Dedicated Anti-Phishing Service*. [Online]. Available: http://www.millersmiles.co.uk/

[70] Y. Zheng, T. Liu, Y. Wang, Y. Zhu, Y. Liu, and E. Chang, "Diagnosing new york city's noises with ubiquitous data," in *Proc. ACM Int. Joint Conf. Pervas. Ubiquitous Comput.*, Sep. 2014, pp. 715–725.

[71] E. Jürrens, A. Bröring, and S. Jirka, "A human sensor web for water availability monitoring," in *Proc. OneSpace-2nd Int. Workshop Blending Phys. Digit. Spaces Internet*, Berlin, Germany, Oct. 2009.

[72] M. Avvenuti, M. G. Cimino, S. Cresci, A. Marchetti, and M. Tesconi, "A framework for detecting unfolding emergencies using humans as sensors," *SpringerPlus*, vol. 5, no. 1, pp. 1–23, 2016.

[73] R. H. Weber, "Internet of Things–New security and privacy challenges," *Comput. Law Secur. Rev.*, vol. 26, no. 1, pp. 23–30, Jan. 2010.

[74] McAfee, "Social engineering in the internet of things (iot)," (2015). [Online]. Available: https://blogs.mcafee.com/executive-perspectives/social-engineering-internet-things-iot
[75] R. M. Lee, M. J. Assante, and T. Conway. (2016). *Analysis of the Cyber Attack on the Ukrainin Power grid.* [Online]. Available: http://www.nerc.com/pa/CI/ESISAC/Documents/E-ISAC_SANS_Ukraine_DUC_18Mar2016.pdf

Ryan Heartfield received the B.Sc. degree in computer systems and networking from the University of Greenwich in 2011. He is currently a Network Architect working in the UK public sector and pursuing the Ph.D. degree with the CSAFE Group of the Computing and Information Systems department within the University of Greenwich. His research interests include semantic social engineering, cyber-physical attacks, software-defined networks, cloud computing, and network security.

GEORGE LOUKAS received the Ph.D. degree in network security from Imperial College. He is currently a Senior Lecturer in cyber security with the University of Greenwich, U.K. He is also a Principal Investigator for several large-scale EU and U.K. research projects, ranging from the security of autonomous vehicles, to secure collaboration of communities and law enforcement agencies, and to bridging emotion research with cyber security in the context of smart home environments. His research interests include cyber-physical attacks, network security, distributed systems, emergency management, semantic social engineering, and digital forensics.

DIANE GAN received the Ph.D. degree in the field of computer networks from the University of Greenwich. She is currently a Principal Lecturer with the Department of Computing and Information Systems, University of Greenwich. She is also a Chartered Engineer with the IET and a Senior Fellow of the HEA. Her current engagements include research and teaching within the areas of cyber security and digital forensics.