# Evaluating the reliability of users as human sensors of social media security threats

Ryan Heartfield
Computing and Information Systems
University of Greenwich, UK
Email: r.j.heartfield@gre.ac.uk

George Loukas
Computing and Information Systems
University of Greenwich, UK
Email: g.loukas@gre.ac.uk

***Abstract*—While the *human as a sensor* concept has been utilised extensively for the detection of threats to safety and security in physical space, especially in emergency response and crime reporting, the concept is largely unexplored in the area of cyber security. Here, we evaluate the potential of utilising users as human sensors for the detection of cyber threats, specifically on social media. For this, we have conducted an online test and accompanying questionnaire-based survey, which was taken by 4,457 users. The test included eight realistic social media scenarios (four attack and four non-attack) in the form of screenshots, which the participants were asked to categorise as "likely attack" or "likely not attack". We present the overall performance of human sensors in our experiment for each exhibit, and also apply logistic regression to evaluate the feasibility of predicting that performance based on different characteristics of the participants. Such prediction would be useful where accuracy of human sensors in detecting and reporting social media security threats is important. We identify features that are good predictors of a human sensor's performance and evaluate them in both a theoretical ideal case and two more realistic cases, the latter corresponding to limited access to a user's characteristics.**

***Keywords*—Social media, computer security, semantic attacks, phishing, social engineering, human as a sensor.**

## I. Introduction

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. Examples include diagnosing a city's noise pollution [1], road traffic anomalies [2], monitoring water availability [3], neighborhood watch schemes [4], detecting unfolding emergencies [5] and generally augmenting the situational awareness of first responders through social media [6]. Yet, rather surprisingly the concept is very new in relation to detecting and reporting threats in cyber space. We are aware of only one very recent example of research geared specifically towards phishing attacks [7]. Here, we take the first steps towards exploring the applicability of the concept more generally by testing the reliability of human users as sensors of security threats. Our focus is on threats to social media. We have conducted a large-scale online experiment where we have asked 4,457 users to distinguish between attacks and non-attacks on different online usage scenarios presented to them as visual exhibits. The focus of this paper is the analysis of the performance of human users as threat sensors with four examples of social media attacks and four examples of legitimate social media usage. Also, complementing previous research on predicting whether a particular attacker will be successful in their attack [8], [9], here we identify features and models for predicting whether a particular user will successfully detect an attack.

## II. Related work

Stembert et al. [7] have very recently proposed combining a reporting function with blocking and warning of suspicious emails and the provision of educative tips, so as to harness the intelligence of expert and novice users in detecting email phishing attacks in a corporate environment. Initial experimental results of their mock-up have been encouraging for the applicability of the *human as a sensor* concept in this context. Here, we focus on the detection capability of the users by evaluating the performance of a large number of users of different profiles and for a wider range of attacks than only phishing emails. That is because before building a system that depends extensively on a particular type of sensors (and the human sensor is no exception), one needs to be aware of their overall reliability and to be able to predict how well they will perform in different conditions (in this case, with regards to the profiles of the users and the type and difficulty of attacks they are expected to detect and report).

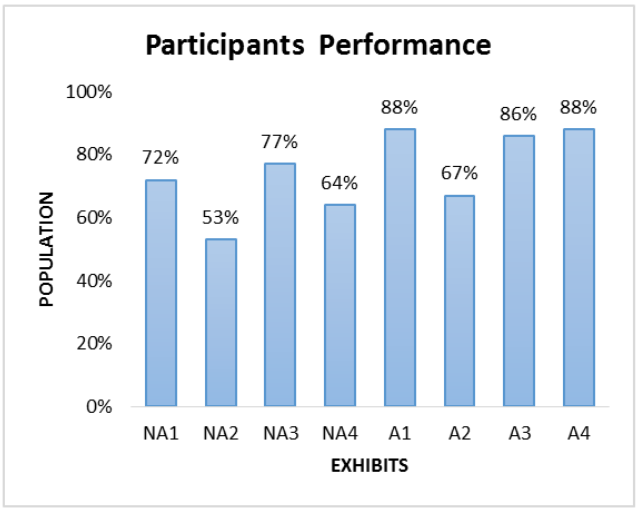
Specifically in relation to social media, it is particularly important to be able to tell to what extent users can correctly detect and report deception-based security threats [10]. In this respect, the related work on user susceptibility to phishing and other semantic social engineering attacks is highly relevant. Predicting whether a user will be deceived into clicking on a fraudulent link or not has traditionally been studied in the realm of behavioural science, where different studies have found that higher degrees of normative, affective and continuance commitment, obedience to authority and trust [11], submissiveness [12], neurotic behaviour [13] and conscientiousness [14] all correlate with high susceptibility to phishing. Also, research in [15] has reported openness, positive behaviour (e.g., use of positive language) and high levels of conversationalist activity as predictors of vulnerability to an online social network bot. However, such behavioural features are rarely practical if the aim is to predict a user's ability to detect attacks within a technical platform. For instance, how would a system measure conscientiousness or submissiveness in real-time, automatically and ethically? Similarly, a number of research studies have reported that female participants were found to be more susceptible to phishing attacks than male participants [14], [16]–[18], but again this is not a predictor that could be used, for instance, in a corporate environment, as it would amount to discrimination. Instead, more practical

Fig. 1: Geographical distribution of study participants


is to know whether users have previously received training on social media security or generally on security threats, which is consistently seen as a useful predictor of their ability to spot them [19], albeit to a varying degree.

Here, we utilise the literature to identify a first set of predictors of a user's ability to detect deception-based attacks and using statistical analysis we select the most relevant among them for different environments. We extend the scope beyond phishing and spear-phishing by including fake apps and QRishing, and measure the ability of users to detect them and the ability of our statistical models to predict whether they will. As the longer-term aim is to incorporate prediction to a technical platform, we are primarily interested in predictors that can be considered as practical, in the sense that their value can be provided or measured in real-time, automatically and ethically.

Fig. 2: Percentage of participants that identified correctly whether each exhibit corresponds to a non-attack (NA1, NA2, NA3, NA4) or an attack (A1, A2, A3, A4)




## III. Methodology

We have conducted a quantitative on-line experiment implemented in the on-line survey platform *Qualtrics*, consisting of a short survey for the collection of demographic and platform behaviour data, and an exhibit-based test. Participants were recruited primarily via popular on-line forums and social media communities, such as Reddit, 4CHAN, StumbleUpon, Facebook and Twitter, with an online advertisement challenging them to test their ability to detect attacks. Figure 1 shows the geographical distribution of the participants.

Fig. 3: Example of Twitter phishing website (A1)

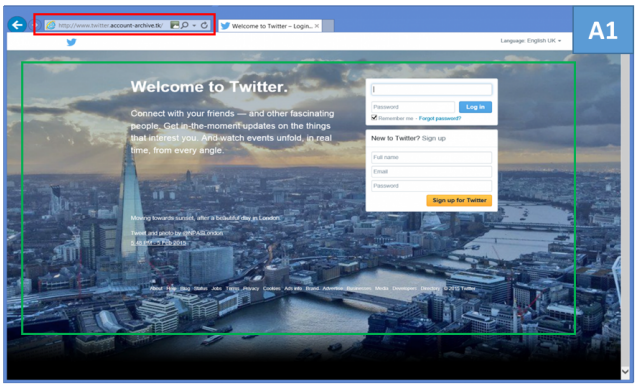
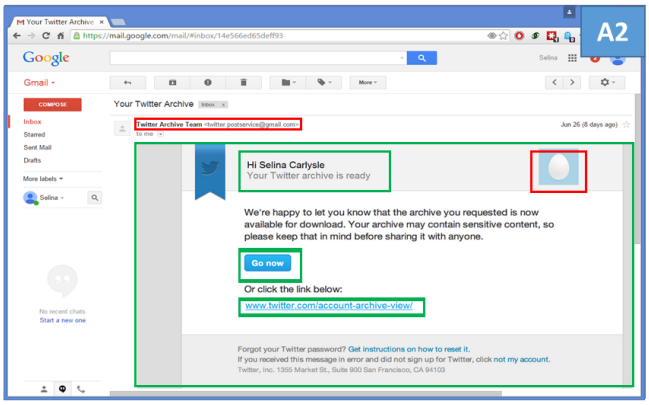



Fig. 4: Example of a Twitter phishing email (A2)




### A. User profile features

The survey portion of the experiment required participants to answer a series of questions related to their age (A), gender (G), security training (S1, S2, S3), platform familiarity (FA), frequency (FR), duration of use (DR), computer literacy (CL), security awareness (SA) and education (EDU). These features are described below:

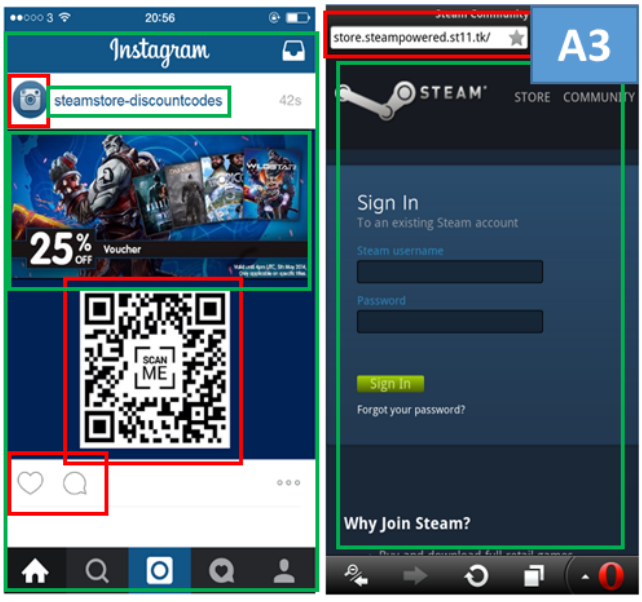
**A**. Age. Coded in groups as: 18-24(1), 25-34(2), 35-44(3), 45-54(4), 55-64(5), 65+(6)
**G**. Gender.
**S1**. Formal computer security education (S1), Coded as a binary response. In relation to the terminology used in [20], S1 is "Formal Learning".
**S2**. Work-based computer security training (S2). Coded as a binary response. In relation to the terminology used in [20], S1 is "Non-formal Learning".

| Exhibit | Description |
|---|---|
| NA1 | Facebook app download from Googleplay, with application permission requirements presented |
| NA2 | Tweet with shortened URL leading to legitimate search on search engine Startpage |
| NA3 | Accidentally mistyped URL for Facebook website, leading to the legitimate Facebook login homepage |
| NA4 | Sponsored tweet with game advertisement on Twitter app, also displaying download |
| A1 | Twitter phishing website |
| A2 | Twitter spear phishing email |
| A3 | Instagram "Qrishing" post that leads to Steam phishing website |
| A4 | Malicious Facebook app posted via friend's timeline; once clicked, requests account permissions with URL redirection |

TABLE I: Attack (A1-A4) and non-attack (NA1-NA4) exhibits included in the test

**S3**. Self-study computer security training (S3). Coded as a binary response. In relation to the terminology used in [20], S1 is "Informal Learning".
**FA**. Familiarity with each platform presented in each exhibit, coded as: Not very (1), Somewhat (2), Very (3)
**FR**. Frequency of use for each platform presented in the test, coded as: Never (1), less than once a month (2), once a month (3), weekly (4), daily (5)
**DR**. Duration of use. For each platform category presented in the susceptibility test, coded as: 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)
**CL**. Computer literacy coded on a scale from 0 to 100 and reported by the participants themselves.
**SA**. Security awareness coded on a scaale from 0 to 100 and reported by the participants themselves.
**EDU**. Level of education, coded as: Less than high school (1), high school /GED (2), some college (3), Trade/technical/vocational training (4), associate degree (5), Bachelor's degree (6), Master's degree (7), doctoral degree (8).

Fig. 5: Example of Instagram Qrishing attack and Steam phishing website (A3)




## B. Exhibits

The test included four exhibits showing attacks (figures 3, 4, 5 and 6) and four exhibits showing normal (non-attack) usage, with an example of these shown in figure 7. For the purposes of demonstration, we have added green outlines that represent a potentially deceiving visual component of the exhibit and red outlines representing visual attack indicators in each attack exhibit. These lines were not shown to the participants. The eight attack and non-attack exhibits are summarised in table I.

Fig. 6: Example of malicious Facebook app attack (A4)

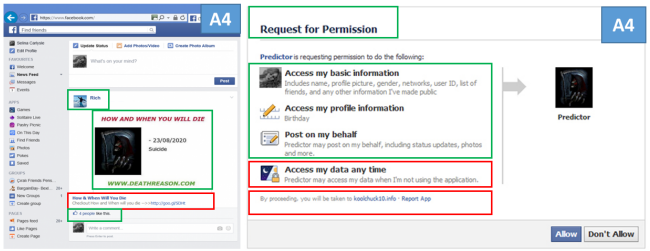



Fig. 7: Example of legitimate Twitter app advertisement (NA4)

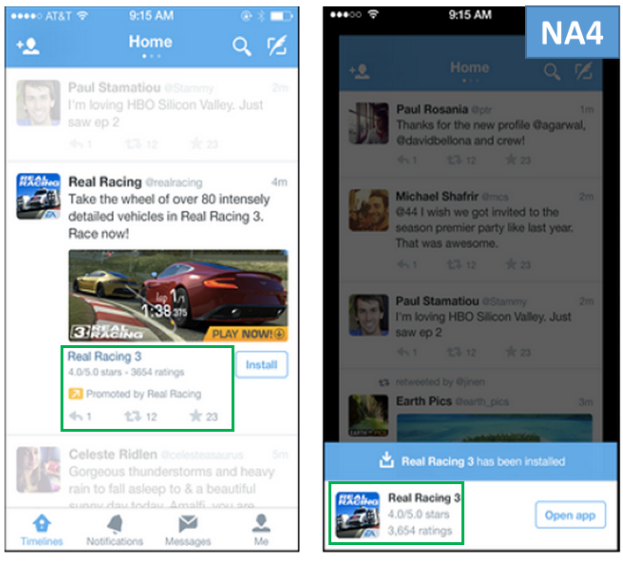



Figure 2 shows the percentage of participants that identified

correctly whether each exhibit corresponds to an attack or not. This can also be considered as a metric of the difficulty of each exhibit.

Our focus is on achieving prediction of a user's ability to correctly distinguish between attacks and non-attacks. For this, we consider the theoretical ideal case, where all features can be utilised (case A), as well as two more constrained and more likely future implementations: (case B) as a reliability prediction module in a security threat reporting mechanism on a social media platform, and (case C) as a mechanism for predicting susceptibility to attacks in enterprise environments with extensive monitoring of the users.

*Case A: Ideal case with all features*

This is the theoretical ideal case, where we predict whether a user will correctly detect an attack or non-attack with access to the complete profile of a user.

*Case B: Report reliability prediction in lightly-monitored social media*

Here, we consider the case where the users of a social media platform are encouraged to act as human sensors and report security threats when they spot them. The social media platform would want to evaluate the trustworthiness of each report based on the human sensor's predicted ability to correctly detect attacks (true positives) and avoid mislabelling normal social media usage as attacks (false positives). The challenge is that only a few of the predictors discussed in Section III are practical. Specifically, it is assumed that the social media provider collects data only on frequency and duration of use, and can additionally request the user to self-report computer literacy, security awareness and platform familiarity. The focus here is on achieving a balance between true positive and false positive reports.

*Case C: Susceptibility prediction in heavily-monitored enterprise environment*

Here, we consider the case where the users are employees within an enterprise environment. Their organisation is interested in estimating the likelihood that they would be deceived by an attack, for instance to determine whether they should control their usage of social media, display warnings, recommend training etc. The organisation can have access to more input features than in case B, including their training history, but for ethical reasons cannot make use of protected information, such as age and gender, which were available in case A. Also, in this context where there is no reporting, false positives and true negatives are of lower importance than true positives and false negatives.

## IV. Prediction model

The prediction of whether a user will correctly or incorrectly detect an attack (or non-attack) is a binary classification problem. Using R [21], we have performed forward stepwise logistic regression to identify models that can predict a user's ability to detect attacks and non-attacks. The forward step selection process is initiated by creating a null model, which includes no feature variables and then proceeds to iteratively test the addition of each variable in the feature space against a model comparison criterion, such as Akaike or Bayes information criterion, Pseudo $R^2$ or cross-validation; at each step adding variables to the model that improve prediction. This routine is repeated for each variable in the feature space until no improvement is achieved. In this study, we have selected 5-fold cross-validation to estimate the test error against different numbers of predictors. Here, the user sample is partitioned into 5 equal folds. Four folds are used to train the model and the remaining fold is used to test the model. The process is repeated 5 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 the forward stepwise process.

The result of the regression is the selection of those features that have a statistically significant impact on the probability of a user's correct prediction. For $K$ number of features used in the prediction, and a given user's value for each feature $k \in \{1, K\}$ being $X = \{x_k\}$, that user's predicted probability of correct detection is given by:

$$\hat{p} = \frac{e^{\beta_0 + \sum \beta_k x_k}}{1 + e^{\beta_0 + \sum \beta_k x_k}}$$

where $b_k$ is the coefficient of feature $k$, as computed by the logistic regression.

The three cases (A, B, C) are practically differentiated by their set of features $X$ (and the corresponding coefficients $\beta_k$).

In model A, $X = \{S1, S2, S3, FA, FR, DR, SA, CL, A, G\}$.
In model B, $X = \{FA, FR, DR, SA, CL\}$.
In model C, $X = \{S1, S2, FA, FR, DR, SA, CL\}$.

Following the most common practice in logistic regression, we provide the result in the form of $\frac{\hat{p}}{1-\hat{p}}$ odds ratios (OR), where:

$$OR = \frac{\hat{p}}{1 - \hat{p}} = e^{\beta_0 + \sum \beta_k x_k}$$

TABLE II: A3 exhibit: Logistic Regression odds ratios for cases A,B,C. A value above 1 indicates a significant predictor of correct detection, while a value below 1 indicates a significant predictor of incorrect detection

| Case | Predictors selected and corresponding odds ratios |
|---|---|
| A | FA (Steam):1.57, SA:1.01, S3:1.62, G:0.46, FA (Facebook):0.65, |
| B | FA (Steam):1.63, SA:1.01, FR:0.87, DR:0.93, CL:1.001 |
| C | FA (Steam):1.62, SA:1.01, DR (SM):0.81, DR (IM):1.16, FR (SE):0.78, CL:1.01 |

As an example, Table II shows the statistically significant predictors selected for one of the exhibits (A3) and the corresponding odds ratios. This is interpreted as follows: In case A, the odds of a user correctly identifying A3 as an attack when all other features of that user's profile remain fixed is increased by 57% for every one unit increase in the familiarity scale for the particular platform (Steam). In cases

Fig. 8: ROC curves for prediction performance for each exhibit in case B

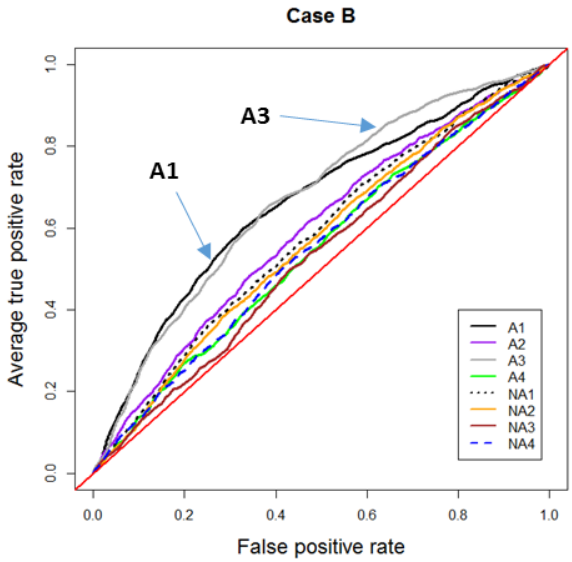
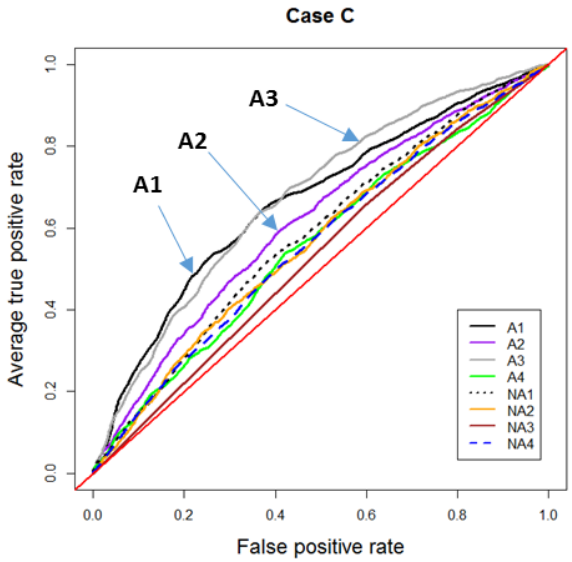



Fig. 9: ROC curves for prediction performance for each exhibit in case C




B and C, this is 63% and 62%, which shows that despite the effect of platform habitation [22], here familiarity is a very useful predictor of a human sensor's ability to detect the particular attack. This agrees with previous results on the importance of familiarity with a system as a key enabler of distinguishing between what visually looks normal and what is normal behaviour [23], [24]. Also very important is the security self-study (S3) feature with an improvement of 62% for every one unit increase on the self-study scale if all other features of the user's profile remain fixed. However, this could be used only in the ideal case (A), as whether a user has indeed carried out self-study cannot be monitored or confirmed in practice by the social media platform (case B) or the user's employer in an enterprise environment (case C).

## V. Prediction Performance Results

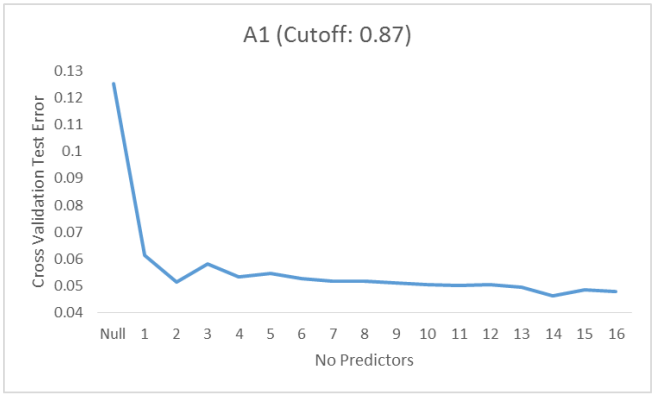
Next, we have performed 5-fold cross validation to estimate the prediction test error and plot it against the number of predictors utilised. The cross-validated test error depends on the logit probability threshold cut-off, which is effectively the tuning parameter of our prediction model. For case A, figures 10 and 11 summarise the test error against the number of predictors that were added with the stepwise approach. In accordance with the generally accepted practice in logistic regression [25], the cut-off value is chosen to be close to the event rate for each exhibit (i.e., the percentage of participants who were correct, as shown in figure 2). We observe that the prediction test error is sufficiently low with 2-5 predictors for most of the exhibits, and adding further predictors has diminishing returns. This can be seen also in Table II, where, although case A had all features available to it, the model used only five of them as useful predictors.

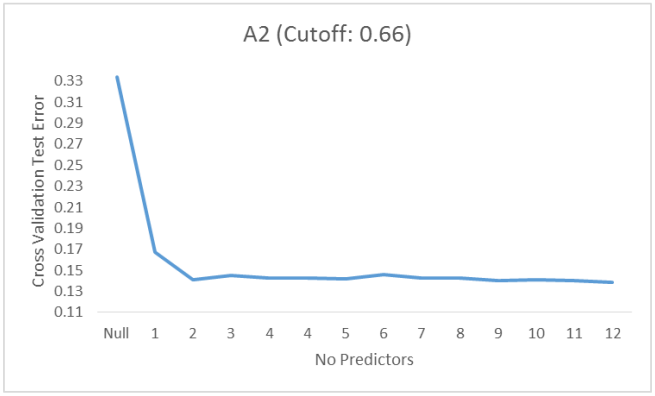
To evaluate the performance of the models in a more realistic manner, we focus on cases B and C. In figures 8 and 9, we summarise the overall performance of the models with the constrained sets of predictors that were chosen via logistic regression for these two cases. We use receiver operating characteristic curves to plot average true positive rate against false positive rate for different thresholds. The further above of the red diagonal line that goes from (0.0) to (1.1) the better the performance. We observe that the performance of prediction for non-attacks is rather poor, being close to the diagonal line. However, the approach achieves good performance for the prediction of three out of four attacks (A1, A2, A3), which would be the primary aim of a system predicting the ability of a user to correctly detect an attack.

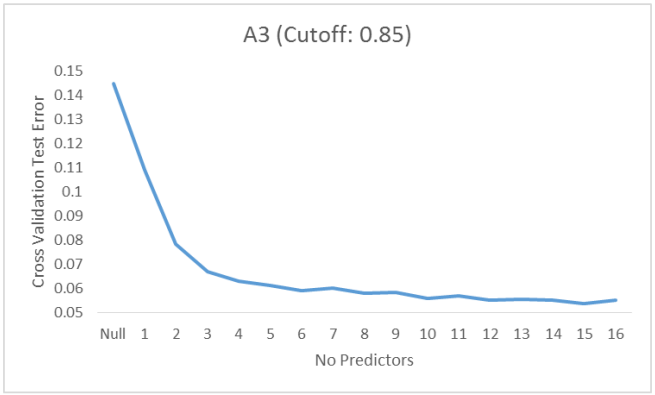
## VI. Conclusion

We have presented the results of a large-scale online experiment, measuring the performance of users as human sensors of deception-based security attacks in social media. In cases B and C we have demonstrated the utilisation of human-generated attributes as a practical measure to predict user accuracy and credibility of reported semantic attacks against a social media platform; identifying consistent performance between a number of attack across a limited set of indicators that are ethical and can be measured automatically and in real-time. We have shown that it is feasible to predict to some extent users' ability as detectors of such attacks, which can be highly useful in environments where the concept of the human sensor of security threats may be considered, including the social media platforms themselves or corporate environments where employees use social media. The next stage in this work will involve the development of a technical system that can operate in both a corporate environment and external independent platform. Future research in this field can also investigate the feasibility of using human sensors for deception-based attacks in different environments, such as in the context of cloud computing [26], the Internet of Things and cyber-physical systems [27].

Fig. 10: Case A: Attack cross-validation Test Error against number of predictor variables in the model



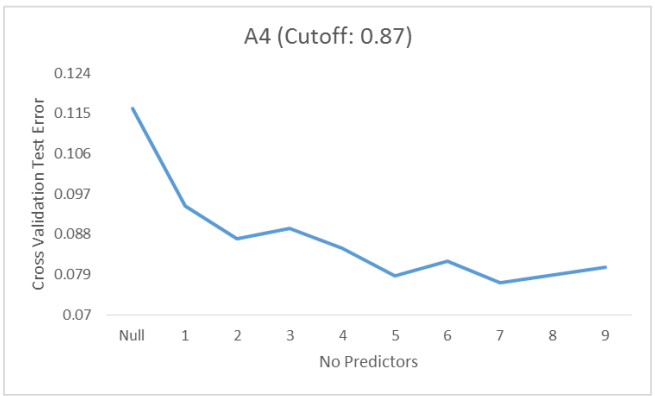
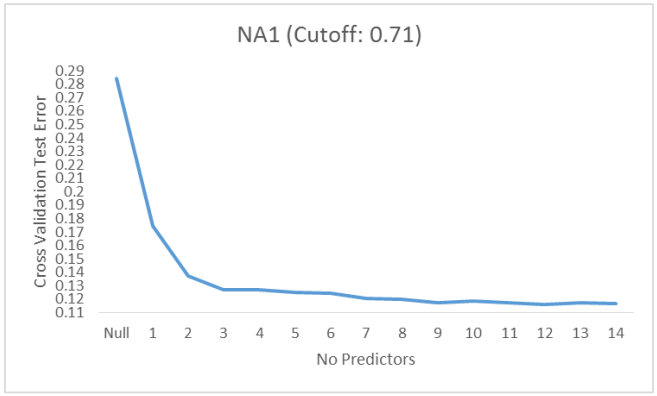
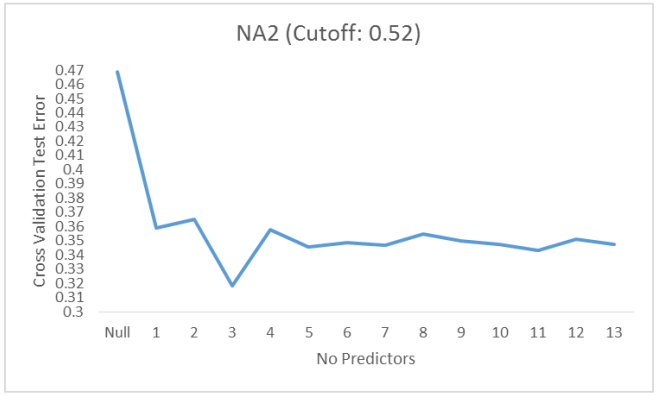
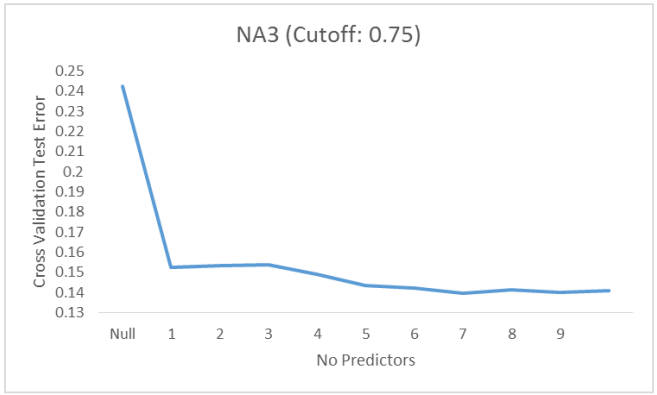
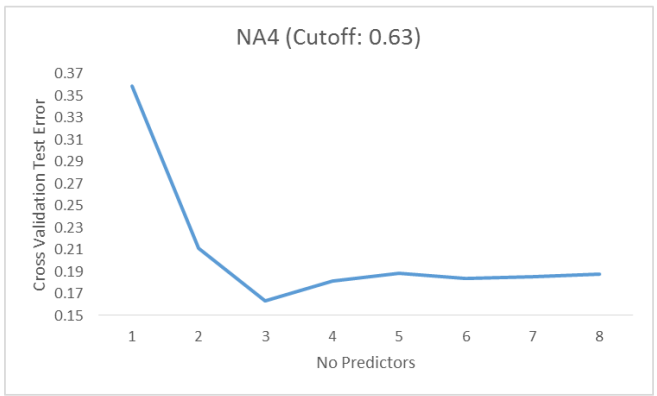




Fig. 11: Case A: Non-attack cross-validation Test Error against number of predictor variables in the model



Up to now, we have focused on deception-based attacks, where the user is deceived into performing a compromising action. However, it is likely that the concept of the human sensor can potentially be extended to attacks that do not involve deception. For instance, it is the human users of a website that often first notice that a website is experiencing poor availability and their reports could complement network monitoring and help speed up denial of service detection [28], [29]. Also, in cyber-physical systems, such as semi-autonomous vehicles, the human operator is likely to be the first to observe the adverse physical impact of a command injection attack [30], [31]. In the future, we intend to extend the scope of this research on human sensors of security threats in terms of types of attacks and platforms involved. The aim is by no means to replace technical security systems, but to enhance them by leveraging human sensing capacity and experience.

## REFERENCES


[1] Y. Zheng, T. Liu, Y. Wang, Y. Zhu, Y. Liu, and E. Chang, "Diagnosing new york city's noises with ubiquitous data," in *ACM International Joint Conference on Pervasive and Ubiquitous Computing*, Sep. 2014, pp. 715–725.

[2] B. Pan, Y. Zheng, D. Wilkie, and C. Shahabi, "Crowd sensing of traffic anomalies based on human mobility and social media," in *ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, November 2013, pp. 344–353.

[3] E. Jürrens, A. Bröring, and S. Jirkai, "A human sensor web for water availability monitoring," in *OneSpace*, 2009.

[4] T. Bennett, K. Holloway, and D. P. Farrington, "Does neighborhood watch reduce crime? a systematic review and meta-analysis," *Journal of Experimental Criminology*, vol. 2, no. 4, pp. 437–458, 2006.

[5] 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.

[6] S. K. Boddhu, R. Dave, R. Williams, M. McCartney, and J. West., "Augmenting situational awareness for first responders using social media as a sensor," *Analysis, Design, and Evaluation of Human-Machine Systems*, vol. 12, no. 1, pp. 133–140, 2013.

[7] N. Stembert, A. Padmos, M. S. Bargh, S. Choenni, and F. Jansen, "A study of preventing email (spear) phishing by enabling human intelligence," in *European Intelligence and Security Informatics Conference*, September 2015, pp. 113–120.

[8] A. Filippoupolitis, G. Loukas, and S. Kapetanakis, "Towards real-time profiling of human attackers and bot detection," in *7th International Conference on Cybercrime Forensics Education and Training (CFET)*, Canterbury, UK, July 2014.

[9] S. Kapetanakis, A. Filippoupolitis, G. Loukas, and T. A. Murayziq, "Profiling cyber attackers using case-based reasoning," in *19th UK workshop on Case-Based Reasoning (UK-CBR)*, Cambridge, UK, December 2014.

[10] R. Heartfield and G. Loukas, "A taxonomy of attacks and a survey of defence mechanisms for semantic social engineering attacks," *ACM Computing Surveys*, vol. 48, no. 3, 2016.

[11] M. Workman, "Wisecrackers: A theorygrounded investigation of phishing and pretext social engineering threats to information security," *Journal of the American Society for Information Science and Technology*, vol. 59, no. 4, pp. 662–674, 2008.

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

[13] T. Halevi, J. Lewis, and N. Memon, "A pilot study of cyber security and privacy related behavior and personality traits," in *International conference on World Wide Web*, May 2013, pp. 737–744.

[14] T. Halevi, N. Memon, and O. Nov. (2015, January) 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

[15] J. G. Mohebzada, A. E. Zarka, A. H. Bhojani, and A. Darwish, "Phishing in a university community: Two large scale phishing experiments," in *IIT*, Cambridge, United Kingdom, April 2012, pp. 373–382.

[16] S. Sheng, M. Holbrook, P. Kumaraguru, L. Cranor, and J. Downs, "Who falls for phish?: a demographic analysis of phishing susceptibility and effectiveness of interventions," in *SIGCHI CHI*, Atlanta, GA, USA, 2010, pp. 373–382.

[17] M. Blythe, H. Petrie, and J. A. Clark, "F for fake: four studies on how we fall for phish," in *SIGCHI Conference on Human Factors in Computing Systems*, May 2011, pp. 3469–3478.

[18] J. Hong, "The state of phishing attacks," *Communications of the ACM*, vol. 55, no. 1, pp. 74–81, 2012.

[19] P. Kumaraguru, J. Cranshaw, A. Acquisti, L. Cranor, J. Hong, M. A.Blair, and T. Pham, "School of phish: a real-world evaluation of anti-phishing training," in *Symposium on Usable Privacy and Security*, July 2009.

[20] D. Colardyn and J. Bjornavold, "Validation of formal, nonformal and informal learning: Policy and practices in eu member states," *European journal of education*, vol. 39, no. 1, pp. 69–89, 2004.

[21] R. Ihaka and R. Gentleman, "The R project for statistical computing," 2016.

[22] S. Egelman, L. F. Cranor, and J. Hong, "You've been warned: an empirical study of the effectiveness of web browser phishing warnings," in *SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2008, pp. 1065–1074.

[23] J. S. Downs, M. B. Holbrook, and L. F. Cranor, "Decision strategies and susceptibility to phishing," in *Second symposium on Usable privacy and security*. ACM, 2006, pp. 79–90.

[24] J. S. Downs, M. Holbrook, and L. F. Cranor, "Behavioral response to phishing risk," in *Anti-phishing working group's 2nd annual eCrime researchers' summit*, October 2007, pp. 37–44.

[25] D. W. Hosmer Jr, S. Lemeshow, and R. X. Sturdivant, *Applied logistic regression (Vol. 398)*. John Wiley and Sons, 2013.

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

[27] G. Loukas, *Cyber-Physical Attacks: A Growing Invisible Threat*. Butterworth-Heinemann (Elsevier), 2015.

[28] E. Gelenbe, M. Gellman, and G. Loukas, "Defending networks against denial-of-service attacks," in *European Symposium on Optics and Photonics for Defence and Security (SPIE)*, London, UK, October 2004, pp. 233–243.

[29] G. Loukas and G. Oke, "Likelihood ratios and recurrent random neural networks in detection of denial of service attacks," in *International Symposium of Computer and Telecommunication Systems (SPECTS)*, San Diego, CA, USA, July 2007.

[30] T. Vuong, G. Loukas, and D. Gan, "Performance evaluation of cyber-physical intrusion detection on a robotic vehicle," in *Proceedings of 13th International Conference on Pervasive Intelligence and Computing (PICOM)*. IEEE, 2015.

[31] T. Vuong, G. Loukas, D. Gan, and A. Bezemskij, "Decision tree-based detection of denial of service and command injection attacks on robotic vehicles," in *Proceedings of 7th International Workshop on Information Forensics and Security (WIFS)*. IEEE, 2015.