Quantifying Psychological Sophistication of Malicious Emails

Rosana Montañez Rodriguez[1], Theodore Longtchi[2], Kora Gwartney[2], Ekzhin Ear[2], David P. Azari[3], Christopher P. Kelley[3], and Shouhuai Xu[2(✉)]

[1] Department of Computer Science, University of Texas, San Antonio, USA
[2] Department of Computer Science, University of Colorado, Colorado Springs, USA
sxu@uccs.edu
[3] Department of Behavioral Sciences and Leadership, US Air Force Academy, Colorado Springs, USA

Abstract. Malicious emails (including phishing, spam, and scam) are significant attacks. Despite numerous defenses to counter them, they remain effective because our understanding of their psychological properties is superficial. This motivates us to investigate the psychological sophistication, or *sophistication* for short, of malicious emails. For this purpose, we propose an innovative framework of two pillars: *Psychological Techniques* (PTechs) and *Psychological Tactics* (PTacs). We propose metrics and grading rules for human experts to assess the sophistication of malicious emails through PTechs and PTacs. To demonstrate the usefulness of the framework, we conduct a case study based on 200 malicious emails assessed by four independent graders.

Keywords: Malicious emails · psychological sophistication · psychological techniques · psychological tactics · cybersecurity metrics

1 Introduction

Malicious emails remain effective despite numerous defensive efforts because existing solutions do not adequately consider psychological factors [13]. This inspires us to introduce and investigate the notion of *psychological sophistication* of malicious emails to pave the way towards designing effective defenses. More specifically, we ask and investigate two questions: (i) How can we quantify the psychological sophistication of malicious emails? (ii) How sophistication varies among different categories (or types) of malicious emails in the real world?

Our Contributions. First, we propose an innovative and systematic framework for quantifying the psychological sophistication, or sophistication for short, of malicious emails. The framework deconstructs and compares the content of malicious emails through two lenses. At a low-level, we propose identifying the number of psychologically relevant textual and imagery elements in an email message,

R. M. Rodriguez and T. Longtchi—Equal contribution.

M. Yung et al. (Eds.): SciSec 2023, LNCS 14299, pp. 319–331, 2023.
https://doi.org/10.1007/978-3-031-45933-7_19

dubbed *Psychological Techniques* (PTechs), to provide a detailed accounting of the elements employed by an attack. At a high-level, we propose assessing an attacker's overall deliberate thoughtfulness (i.e., effort) in framing malicious content to influence an email recipient, dubbed *Psychological Tactics* (PTacs), to offer insights into an attacker's effort to exploit human fallibility. Second, we demonstrate the usefulness of the framework by applying it to quantify sophistication of 200 malicious emails. This leads to useful insights, including: (i) Phishing emails are psychologically more sophisticated than spam and scam emails because phishing emails contained both higher PTech scores and higher PTac scores. (ii) Emails having low PTac scores also have low PTech scores.

Ethical Issue. In consultation with University of Colorado Colorado Springs Internal Review Board (IRB), this study does not need IRB approval because no subjects are part of the study and the emails are provided by a third party.

Related Work. Although several studies have discussed the use of psychological content in phishing emails (e.g., [2,6,8]), few studies provide a systematic approach to quantifying it. Heijden and Allodi [23] measure the presence of persuasion elements in phishing emails to identify those that are more likely to succeed. By contrast, we also consider how the email overall message presentation affects success. Nelms et al. [17] identify and categorize psychological tactics to encourage users to download malicious applications. By contrast, we consider a broader set of psychological constructs used in phishing emails. Ferreira and Lenzini [5] systematically quantify psychological content in phishing messages based on low-level psychological elements. By contrast, we incorporate these principles and several other psychological elements, while leveraging phishing emails from the Anti Phishing Working Group (APWG).

Paper Outline. Section 2 describes the core concepts. Section 3 presents the framework. Section 4 reports a case study. Section 5 discuss limitations of the present study; Sect. 6 concludes the paper.

2 Concepts

PTech. A PTech is a concrete (i.e., quantifiable) textual or imagery element that encourages individuals to comply with a malicious email. The following 13 PTechs have been identified in the literature include [13,16]. (1) *Urgency*: The use of textual elements (e.g., "acting now") to trigger a recipient's immediate action [4,24]. (2) *Visual Deception*: The use of visual elements (e.g., logos) or "similar" characters in URL (e.g., replacing'vv' with'w') to project trust [15]. (3) *Incentive and Motivator*: The use of textual elements, such as "free stuff" (incentive) or "help others", to incentivize or motivate a recipient to take action [3,16]. (4) *Persuasion*: The use of textual elements related to Cialdini's principles (e.g., "C-Suite titles," "last chance," or "expert opinion") to encourage a recipient to encourage a behavior [5,17]. (5) *Quid-Pro-Quo*: The use of textual elements (e.g., "Pay an upfront fee") to ask a recipient for a favor in exchange for a bigger reward [22]. (6) *Foot-in-the-Door*: The use of textual elements (e.g.,

"from our last email/ conversation...") to obtain compliance from a recipient via gradually increasing demands [7]. (7) *Trusted Relationship*: The exploitation of an established third-party relationship of trust with the recipient by using textual elements like "John told me about you" to convince a recipient to act [2]. (8) *Impersonation*: The use of a false persona to gain the trust of a recipient by using elements like "I'm billionaire Warren Buffet" [2,5]. (9) *Contextualization*: Referencing current event by using textual elements like "the Pandemic" or "War in Ukraine" [8,15]. (10) *Pretexting*: Providing a motive to establish contact with a recipient by using textual elements like "I am recruiter for XYZ company" [1,8]. (11) *Personalization*: Addressing a recipient using detailed personal information in textual elements such as "Dear John" or "Your credit card ending in..." [11,15]. (12) *Attention Grabbing*: The use of graphical/auditory elements to draw attention to textual elements such as highlighted text, brightly colored buttons, or extra large fonts [6,17]. (13) *Affection trust*: Developing an effective relationship to extort a recipient by using textual elements like "My child is sick and I have no money to pay the treatment" [14].

PTac. This is a new concept introduced in this paper. A PTac aims to measure the attacker's effort at crafting and framing an email effectively to prompt a recipient's action. There are 7 PTacs. (1) *Familiarity*: It reflects the attempt of an attacker to engender a positive (and therefore trusting) association with a recipient. Emails of high familiarity may impersonate specific people (e.g., co-workers, bosses, family members, close friends) [1,14]. (2) *Immediacy*: It is an amplifier which uses time as a mechanism to short-cut recipient skepticism or scrutiny for any desired action, for example, by suggesting that promptness, swiftness, or a quick reaction is required [15,17]. (3) *Reward*: It is a clear exchange of something (physical or social) valuable for a recipient. Rewards are often presented as a tangible good (e.g., money) in exchange for action but can also be an offer to improve social standing (e.g., power, authority, prestige) [8,13]. (4) *Threat of Loss*: It is an appeal to a recipient's desire to maintain their current status, prevent a loss (e.g., opportunity) or injury (e.g., damage, pain), or avoid the risk of having something stolen. It has been hypothesized to be more impactful than potential of gain (e.g., reward) [5,8,22]. (5) *Threat to Identity*: It is a recipient's desire to maintain a positive, socially valuable reputation [14,22]. (6) *Claim to Legitimate Authority*: It is intended to leverage respect for legitimate power. The attacker may assume a position of technical expertise or a valuable institutional role, or hold a traditionally respected office [5,22]. (7) *Fit & Form*: It mirrors the expected composition style of an authentic message. An attacker often exploits commonly expected written or visual display format to resonate with the email's apparent sender and purpose [8,20].

3 Framework

The framework consists of six components: (i) selecting PTechs and PTacs for assessment; (ii) defining metrics to quantify sophistication of malicious emails;

(iii) designing grading rules and calibration process to guide the grading of malicious email sophistication; (iv) preparing a dataset of malicious emails for expert graders to assess; (v) grading emails in the dataset by expert graders; and (vi) analyzing outcome of the grading process.

3.1 Selecting PTechs and PTacs

We propose selecting the PTechs that (i) are known to be used in malicious emails based on research evidence and (ii) require a one-time interaction to be effective. This selection criterion is flexible enough to accommodate future understanding and knowledge (e.g., when new PTechs are discovered in the future). Similarly, we propose selecting PTacs that (i) are known to be used in malicious emails based on research evidence, (ii) are independent of one another, and (iii) reflect the holistic effort of an attacker. Suppose, according to the respective selection criteria, ℓ PTechs are selected, denoted by $\{PTech_1, \ldots, PTech_\ell\}$, and m PTacs are selected, denoted by $\{PTac_1, \ldots, PTac_m\}$.

3.2 Defining Sophistication Metrics

Metrics for Measuring PTechs. Consider a malicious email and a set of ℓ PTechs denoted by $\{PTech_1, \ldots, PTech_\ell\}$ as described above. For $PTech_i$ where $1 \leq i \leq \ell$, we propose counting the number of elements with respect to $PTech_i$, leading to an integer score s'_i. Then, the sophistication of the malicious email through the lens of the ℓ PTechs can be defined as, $s' = \frac{1}{\ell} \sum_{i=1}^{\ell} s'_i$. Since ground truth s'_i is difficult to obtain, we propose to approximate it by using a number of n graders (or evaluators) to count the elements concerning $PTech_i$ while assuring that the graders can count the elements as consistently as possible. For a malicious email, let $s_{i,j}$ denote the count of elements in the email by grader j with respect to $PTech_i$, where $1 \leq j \leq n$ and $1 \leq i \leq \ell$. Then, the sophistication of the email concerning $PTech_i$ can be defined as,

$$S_i = \frac{1}{n} \sum_{j=1}^{n} s_{i,j}. \tag{1}$$

Given S_i for $1 \leq i \leq \ell$, we propose defining the sophistication of the email with respect to the ℓ PTechs, denoted by S_{PTech}, as:

$$S_{PTech} = \frac{1}{\ell} \sum_{i=1}^{\ell} S_i. \tag{2}$$

Metrics for Measuring PTacs. Consider a malicious email and a set of m PTacs denoted by $\{PTac_1, \ldots, PTac_m\}$. Since the ground-truth sophistication reflected by $PTac_i$ is hard to obtain, we propose assessing $PTac_i$ using a rating scale ranging from 1 to β (e.g., $\beta = 5$ in our case study), also by a panel of n independent graders, where $p_{i,j}$ denotes the assessment of grader j with respect

to $PTac_i$ in a message, where $1 \leq j \leq n$ and $1 \leq i \leq m$. The final assessed value of $PTac_i$ can be defined as

$$P_i = \frac{1}{n} \sum_{j=1}^{n} p_{i,j}. \tag{3}$$

The overall PTac-based sophistication of an email can be defined as:

$$S_{PTac} = \frac{1}{m} \sum_{i=1}^{m} P_i/\beta, \tag{4}$$

where P_i/β reflects the degree of $PTac_i$ in a malicious email. Now we are ready to define the sophistication of malicious emails.

Definition 1 (sophistication of malicious email). *The sophistication of a malicious email is measured as a two-dimensional vector* (S_{PTech}, S_{PTac})*, where* S_{PTech} *is defined in Eq.* (2) *and* S_{PTac} *is defined in Eq.* (4)*.*

Note that Definition 1 operates in the ideal world where every score given by ever grader will be incorporated. In the real world, some grade by some grader may be outlier, which may need to be excluded according to some well-established inclusion criteria. This means that Definition 1, or Eq. (1) and Eq. (3), may need to be amended to accommodate such realistic situations. Specifically, when coping with Eq. (1), which computes the average score S_i of $PTech_i$ by the n graders, we may encounter, for example, grader j' $(1 \leq j' \leq n)$ gives an outlier score $s_{i,j'}$. In this case, $s_{i,j'}$ may be excluded when computing the average score. As a result, Eq. (1) becomes for $1 \leq i \leq \ell$:

$$S_i = \frac{1}{n-1} \left(\sum_{j=1}^{n} s_{i,j} - s_{i,j'} \right). \tag{5}$$

When there are multiple outliers with respect to $PTech_i$, they all can be removed in the same fashion. Similarly, suppose grader j^* $(1 \leq j^* \leq n)$ gives an outlier score p_{i,j^*} with respect to $PTac_i$. Then, p_{i,j^*} may be excluded when computing the average score, meaning that Eq. (3) now becomes for $1 \leq i \leq m$:

$$P_i = \frac{1}{n-1} \left(\sum_{j=1}^{n} p_{i,j} - p_{i,j^*} \right). \tag{6}$$

Similarly, when there are multiple outliers with respect to $PTac_i$, they all can be removed in the same fashion. With the amended Eqs. (5) and (6), Eqs. (2) and (4), and thus Definition 1, remain valid.

3.3 Designing Grading Rules

To measure PTechs and PTacs, we propose that each grader manually counts the number of psychological elements of each PTech and each PTac exhibited in an

email. Given that the interpretation of "element" relies on one's domain expertise, we design *grading rules* to reduce subjectivity during the grading process. To guide the development of grading rules, we propose the following: (i) Initial rules are designed by multiple experts. (ii) These initial rules may be reconciled into a unified set of rules to resolve any discrepancies in the initial rules. (iii) The resulting rules are tested by a group of graders using sample data so that discord or ambiguity in the grading rules that arise during the testing can be documented and addressed. (iv) The grading rules may be further revised to mitigate potential inconsistencies or discrepancies in the grading process. The preceding (iii)-(iv) may be repeated until satisfactory consistency is achieved among graders guided by grading results.

3.4 Preparing the Data

Several issues must be addressed when preparing data, including collection and preprocessing. First, to ensure dataset quality (i.e., emails are suitable for quantifying sophistication), we must determine if the emails are malicious. Second, given a set of malicious emails, we must ensure that each email content is rendered similarly on different machines and platforms from a visual point of view. Third, we must ensure that the data preparation process does not cause damage to the research environment. Fourth, malicious emails may contain broken links or missing images needed to complete an email for sophistication assessment. In these cases, we must reconstruct an email by adding the missing links or images.

3.5 Calibration and Grading

Calibration mitigates human (including expert) subjectivity in grading. With grading rules on hand, the graders consistently learn how to apply the rules and practice grading using sample emails. For this purpose, we propose the calibration process highlighted in Fig. 1.

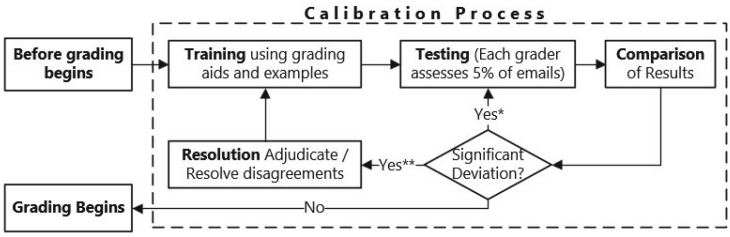


Fig. 1. The calibration process includes presenting initial grading rules to graders, training, and testing. Results are compared according to a defined threshold, such as a Krippendorff's Alpha or Kalpha or simply α. A value lower than the threshold may require training before testing. The calibration process is iterative, and the initial grading rules may be refined when reconciling the discrepancy among graders. Grading begins at the end of the calibration process.

The process has 4 steps. (i) Training: Graders learn how to grade emails using the initial grading rules and grading aid, where the latter demonstrates the application of the former. Graders can ask questions (e.g., what would count as an "element" for a specific PTech), resolve disagreements, and collectively build a shared understanding of the assessment methods through consensus. (ii) Testing: Each grader evaluates an email sample (e.g., 5% of a data corpus) from the study dataset. This sample is excluded from the study. (iii) Comparison: Results obtained from the Testing step are compared for agreement. We propose to use a reliable method to measure the agreement between the graders. For example, *Krippendorff's* (i.e., Kalpha or α) [10] is a reliability coefficient that measures the agreement among raters. Kalpha supports categorical, ordinal, and interval data, and it is robust in the light of any potential missing data. A Kalpha (α) with $0.667 \leq \alpha < 0.8$ is considered an acceptable value, meaning that the data is statistically reliable to draw conclusions. Kalpha $\alpha = 1$ means a perfect agreement among graders. When $\alpha < 0.667$, resolution is necessary and conducted in the next step. (iv) Resolution: when $\alpha < 0.667$, we propose using a consensus-building technique to reach an agreement, such as the Delphi standard consensus technique [9]. Steps (iii)-(iv) may be repeated until consistent grading is achieved. Once calibration is completed, the graders can start to grade the rest emails in the dataset (i.e., those not used in the calibration process). After the grading, outliers grades are excluded with respect to each PTech and each PTac, as per Eqs. (5) and (6) based on a predefined inclusion criteria.

3.6 Analysis

The analysis may be geared towards answering research questions, which are often proposed based on researchers' insights from some unique perspective. Examples of research questions include: How sophistication varies among different categories of malicious emails in the real world? Further research questions can include: Which PTechs and PTacs are most commonly employed in real-world malicious emails?

4 Case Study

4.1 Selecting PTechs and PTacs

The PTech selection criteria described in the framework prompt us to select the following 8 PTechs based on the selection criteria described in Sect. 3.1: (i) *urgency*, (ii) *incentives and motivators*, (iii) *attention grabbing*, (iv) *personalization*,(v) contextualization,(vi) *persuasion*, (vii) *impersonation*, and (viii) *visual deception*. Moreover, the PTac selection criteria described in the framework prompt us to select all the 7 PTacs presented in Sect. 2.

4.2 Instantiating Sophistication Metrics

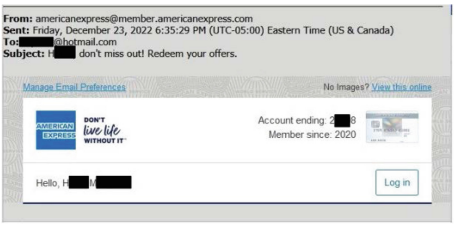
As descried in the framework, the effect of each PTech can be quantified by counting the occurrence of psychological elements associated with a PTech in the email. However, the PTac metric scale β described in the framework needs to be instantiated to a specific scale value. For this purpose, we propose using the Likert scale $[0, 5]$ (i.e., $\beta = 5$ in the terminology of the framework) which is scaling method commonly used in psychological studies [12]: '0' for *no measurable* application of a PTac (i.e., the attacker does not employ any PTac); '1' for *minimal* application of a PTac (i.e., the attacker does consider the PTac but neither applied it clearly nor consistently); '2' for *light* application of a PTac (i.e., the attacker considers the PTac, but with inconsistency, confusion, or lapses/errors in their approach); '3' for a *moderate* application of a PTac (i.e., the attacker clearly applies the PTac but may still have inconsistencies in their approach); '4' for a *significant* application of a PTac (i.e., the attacker clearly and consistently applies the PTac with minimal errors or lapses); '5' for an *extraordinary* application of a PTac (i.e., the attacker expertly and diligently crafts their message to apply this PTac in a cohesive and thoughtful way). Note that the Likert scale or $\beta = 5$ is just a specific choice, but there could be other choices of interest.

4.3 Designing Grading Rules

To ensure consistency of grading, we develop an in-depth grading aid. The aid includes detailed definitions and real-world emails explaining the grading rationale for each PTech and PTac. We also develop a quick reference of psychological element (i.e., key terms) associated with a particular PTech (Table 1) and a set of examples of emails that are graded with respect to the PTechs and PTacs. Figure 2 shows a specific example.

Table 1. A sample of PTech grading rules, which provide graders with an extensive list of examples for each PTech.

PTech	Examples elements from Emails
Urgency	"call me now" / "Last chance to save your social life"
Visual Deception	PayPal logo, IRS logo / Replacing 'fbi.gov' with 'fbi.gov.net'
Incentives & Motivators	"Your refund notice" / "looking for a part-time assistant,(...) 3 h a week, (...)$400 per week"
Persuasion	**Commitment** - "We are grateful for you past generosity"
Impersonation	"Yours sincerely, Warren Buffet" / Phone/Fax number
Contextualization	"Your UW.edu account..." / "Emergency Covid-19 tax relief"
Personalization	"Hi Wendy" / "Important message intended for John Doe"
Att. Grabbing	"CLICK HERE" / "Safety Measures.pdf"



(a) Email screenshot

PT	Count
Urgency	0
Visual Deception	0
Incentives & Motivators	1
Persuasion	1
Impersonation	1
Contextualization	0
Personalization	3
Attention Grabbing	5

Framing Construct	Value
Familiarity	4
Immediacy	1
Reward	2
Threat of Loss	0
Threat to Identity	0
Claim to legitimate Authority	4
Fit and Form	4

(b) Grading outcome of email in Fig.2a

Fig. 2. Example of grading outcome, where a grader evaluates a screenshot of email in Fig. 2a(redacted for publication, but not for grading). Figure 2b is the grade of the email in Fig. 2a with respect to each PTech and PTac.

4.4 Preparing Data

We prepare a dataset of 200 randomly selected emails using the APWG Reported Phishing module API. To increase the chance of collecting true-positive malicious emails, we select emails submitted by US-CERT (Computer Emergency Response Team), a reputable source. The collected email sample (200 emails) consists of 55.0% phishing (110 emails), 31.5% scam (63 emails), and 13.5% spam (27 emails). Emails are selected using random dates between September 1, 2021, and August 31, 2022. The select emails are reconstructed by appending the raw email header and body into a .eml file. Emails are restored with missing elements (e.g., broken images) and are sanitized by removing embedded warnings and email duplicates. An email client (i.e., reader software) is used to display emails. Email screenshots are used to ensure a consistent display of an email content during the grading. Emails are inventoried and categorized as follows: (i) phishing emails, which are the ones that require a one-time interaction for victimization and include a link or an attachment;(ii) scam emails, which are the ones that require multiple interactions for victimization via phone call or email exchange, or request personal information, while noting that scam emails do not include links or attachments; and (iii) spam emails, which are the ones that are non-malicious and do not obscure information, but usually intended to sell a product or service. To further differentiate between email categories, we examine email links using the ScamPredictor algorithm developed by ScamWatcher [21]. The algorithm is a machine learning classifier based on website characteristics known to be indicators of malicious sites [19].

4.5 Calibration and Grading

Each email is graded by four cybersecurity PhD students.They all conduct a calibration exercise as depicted in Fig. 1.

Once the calibration exercise is completed, grading is conducted as follows. Emails are presented as pop-up windows in a survey developed on the Qualtrics platform. The evaluation is split into two self-paced sessions to minimize grader's

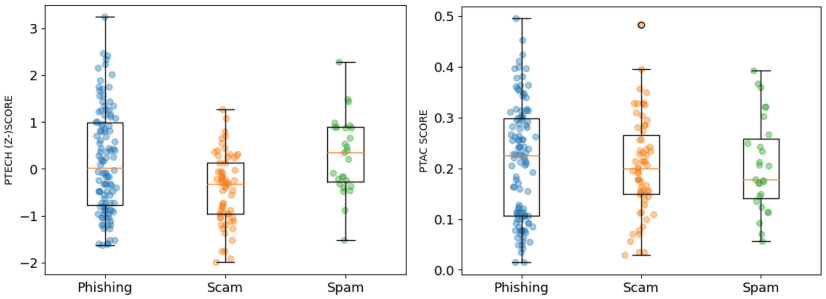
fatigue. Each session consists of 100 emails. To improve consistency of grading, each session is completed within 24 h period. Each session requires about 5 h on average. Within each session, the order of emails are randomized to distribute grading variations introduced by performing the same task over an extended period of time.

Grades are examined for outliers following distinct, predefined inclusion criteria for PTechs and PTacs based on the standard deviation, which is set as 1.5. Outliers are addressed using the procedure outlined in Sect. 3.5. Overall, we exclude 170 (5.67%) out of 1,600 PTech ratings and 216 (7.2%) out of 1,400 for PTacs ratings, with 153 of the 200 emails having at least one outlier grade removed. The Kalpha for this study is 0.822 for PTech and 0.768 for PTacs after outliers removal.

4.6 Analysis

To understand how PTechs and PTacs may differ among the three categories of malicious emails, we use the Z-Score method [18] to normalize the PTech scores to a scale comparable to the PTac scores. Figure 3 shows that phishing emails have higher normalized PTech and PTac scores than the other two categories of malicious emails, respectively. This leads to:

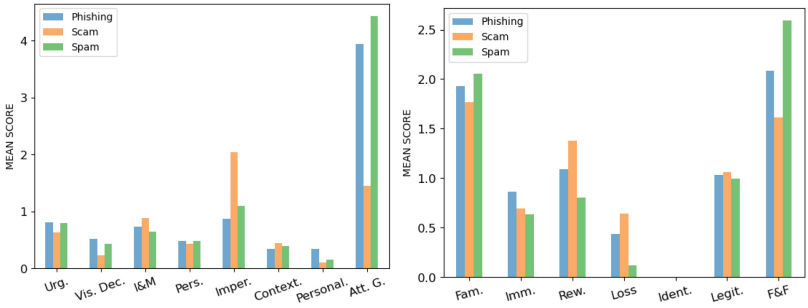
Insight 1. *Phishing emails are psychologically more sophisticated than spam and scam emails from the points of view of both PTech and PTac.*



(a) Normalized PTech scores (b) PTac scores

Fig. 3. Boxplots of the normalized PTech scores and the original PTac scores.

To further explore the relationship between PTech-incurred sophistication and PTac-incurred sophistication, we look at the size of the intersection set of the two sets of emails corresponding to PTech and PTac, respectively. At Q1, the size of the intersection set is: 17.27% for phishing, 22.22% for scam, and 51.85% for spam. At Q3, the size of the intersection set is: 15.45% for phishing, 3.17% for scam, and 7.40% for spam. This leads to:



(a) PTech employment (b) PTac employment

Fig. 4. Comparing the mean score (the y-axis) of each PTech and PTac

Insight 2. *Less sophisticated emails from the PTac perspective are also less sophisticated from the PTech perspective.*

Figure 4 plots the employment of PTechs and PTacs. From this, we draw:

Insight 3. Attention Grabbing *is the most widely employed PTech, and* Fit & Form *and* Familiarity *are the two most widely employed PTac.*

5 Limitations

First, the framework has three limitations: (i) The framework reflects our understanding of the factors that can reflect the psychological sophistication of malicious emails, namely PTechs and PTacs. There may be other psychological factors that need to be considered, which can be accommodated by extending our framework. (ii) The selection criteria we propose may not be perfect, meaning that the select PTechs and PTacs may not be complete or systematic enough. Fortunately, the framework can be easily extended to accommodate other PTechs and/or PTacs of interest. (iii) The grading rules may need to be refined, to more consistently ensure high levels of concurrence in human assessment of email content. Second, the dataset has five limitations. (i) The dataset may not be representative because we only collected and used emails from APWG even though it is arguably the most reputable source in the world. (ii) The dataset is small as we only analyzed 200 emails. (iii) The inclusion criteria need improvement to provide a mathematically robust approach to address and resolve outlier assessments. (iv) We admit the potential issue of 'informed' graders as the graders are also the ones that design and revise the grading rules. This may affect the validity of the experimental results. (v) While the framework can accommodate any reasonable definitions of PTech and PTac, it would be ideal to assure that the PTechs and PTacs are independent.

6 Conclusion

We have presented a framework for quantifying the (psychological) sophistication of malicious emails (including phishing, spam, and scam emails). The framework is based on PTechs and PTacs. We defined metrics to quantify the sophistication of malicious emails. Based on a real-world dataset of 200 malicious emails and 4 graders, we draw a number of insights. Future work should examine the correlation between PTech and PTac components to see how malicious content are used in combination to exploit human psychology.

Acknowledgement. We thank the anonymous reviewers for their useful comments. This work was supported in part by NSF Grants #2122631 and #2115134, and Colorado State Bill 18-086. Approved for Public Release; Distribution Unlimited. Public Release Case Number 23-1373. The first author is also affiliated with The MITRE Corporation, which is provided for identification purposes only and is not intended to convey or imply MITRE's concurrence with, or support for, the positions, opinions, or viewpoints expressed by the authors.

References

1. Al-Hamar, M., Dawson, R., Guan, L.: A culture of trust threatens security and privacy in Qatar. In: 2010 10th IEEE International Conference on Computer and Information Technology, pp. 991–995. IEEE (2010)
2. Allodi, L., Chotza, T., Panina, E., Zannone, N.: The need for new antiphishing measures against spear-phishing attacks. IEEE Secur. Priv. **18**(2), 23–34 (2019)
3. Beckmann, J., Heckhausen, H.: Motivation as a function of expectancy and incentive. In: Heckhausen, J., Heckhausen, H. (eds.) Motivation and Action, pp. 163–220. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-65094-4_5
4. Chowdhury, N.H., Adam, M.T., Skinner, G.: The impact of time pressure on cybersecurity behaviour: a systematic literature review. Behav. Inf. Technol. **38**(12), 1290–1308 (2019)
5. Ferreira, A., Lenzini, G.: An analysis of social engineering principles in effective phishing. In: Workshop on Socio-Technical Aspects in Security and Trust (2015)
6. Flores, W.R., Holm, H., Nohlberg, M., Ekstedt, M.: Investigating personal determinants of phishing and the effect of national culture. Inf. Comput. Secur. **23**, 178–199 (2015)
7. Freedman, J.L., Fraser, S.C.: Compliance without pressure: the foot-in-the-door technique. J. Pers. Soc. Psychol. **4**(2), 195 (1966)
8. Goel, S., Williams, K., Dincelli, E.: Got phished? Internet security and human vulnerability. J. Assoc. Inf. Syst. **18**(1), 2 (2017)
9. Grime, M.M., Wright, G.: Delphi method. Wiley StatsRef Stat. Ref. Online **1**, 16 (2016)
10. Gwet, K.L.: On the krippendorff's alpha coefficient. Manuscript submitted for publication (2011). Accessed 2 Oct 2011
11. Jagatic, T.N., Johnson, N.A., Jakobsson, M., Menczer, F.: Social phishing. Commun. ACM **50**(10), 94–100 (2007)
12. Jebb, A.T., Ng, V., Tay, L.: A review of key Likert scale development advances: 1995–2019. Front. Psychol. **12**, 637547 (2021)

13. Longtchi, T., Rodriguez, R.M., Al-Shawaf, L., Atyabi, A., Xu, S.: SoK: why have defenses against social engineering attacks achieved limited success? arXiv preprint arXiv:2203.08302 (2022)
14. Montañez, R., Atyabi, A., Xu, S.: Social engineering attacks and defenses in the physical world vs. cyberspace: a contrast study. In: Cybersecurity and Cognitive Science, pp. 3–41. Elsevier (2022)
15. Montañez, R., Golob, E., Xu, S.: Human cognition through the lens of social engineering cyberattacks. Front. Psychol. **11**, 1755 (2020)
16. Montañez Rodriguez, R., Xu, S.: Cyber social engineering kill chain. In: Su, C., Sakurai, K., Liu, F. (eds.) SciSec 2022. LNCS, vol. 13580, pp. 487–504. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-17551-0_32
17. Nelms, T., Perdisci, R., Antonakakis, M., Ahamad, M.: Towards measuring and mitigating social engineering software download attacks. In: 25th USENIX Security Symposium, pp. 773–789. USENIX Association, Austin, TX (2016)
18. Nield, T.: Essential Math for Data Science. O'Reilly Media Inc, Sebastopol (2022)
19. Pritom, M., Schweitzer, K., Bateman, R., Xu, M., Xu, S.: Data-driven characterization and Detection of COVID-19 Themed Malicious Websites. In: IEEE ISI (2020)
20. Rajivan, P., Gonzalez, C.: Creative persuasion: a study on adversarial behaviors and strategies in phishing attacks. Front. Psychol. **9**, 135 (2018)
21. SAS, H.: Scamdoc.com. https://www.scamdoc.com/. Accessed 04 Nov 2023
22. Stajano, F., Wilson, P.: Understanding scam victims: seven principles for systems security. Commun. ACM **54**(3), 70–75 (2011)
23. Van Der Heijden, A., Allodi, L.: Cognitive triaging of phishing attacks. In: 28th USENIX Security Symposium 2019, pp. 1309–1326 (2019)
24. Vishwanath, A., Herath, T., Chen, R., Wang, J., Rao, H.R.: Why do people get phished? Decis. Support Syst. **51**(3), 576–586 (2011)