

Netphishing: Network and Linguistic Analysis of Phishing Email Subject Lines

Ana Ferreira¹, Soraia Teles^{1,2}, Rui Chilro³ and Milaydis Sosa-Napolskij¹

¹*CINTESIS - Centre for Health Technology and Services Research, Faculty of Medicine, Porto, Portugal*

²*Institute of Biomedical Sciences Abel Salazar, University of Porto (ICBAS-UP), Department of Behavioural Sciences, Porto, Portugal*

³*Independent Researcher, Portugal*

Keywords: Phishing Email Subject Lines, Network Science, Linguistic and Sentiment Analyses.

Abstract: This study provides support on why subject lines of predatory emails should be analysed to improve the detection of phishing attacks. Network science together with a linguistic analysis were performed on a sample of 240 phishing email subject lines from the past 12 years. Results show that even in straightforward subject lines, phishers can employ text elements to create a sense of proximity, mutual relationship as well as a neutral and professional relation, focused on present and future actions, to persuade potential victims to open phishing emails. The common words “your” and “account” form two main hubs and communities of words that integrate main organisations and actions related to those hubs. The linguistic analysis shows that concise phrases integrate such richness of language that can potentially be used to find differential emotional and behavioural marks on the text, to be used for better detecting phishing emails. This work provides current information as well as new research questions to be tested and further perused, to support the improvement of automated tools to identify predatory emails.

1 INTRODUCTION


The top most unresolved threats in cybersecurity for 2020 include, not surprisingly, phishing, malware, ransomware and social engineering. Unfortunately, there is still no way to prevent phishing attacks and efficiently detect and minimise its negative effects (Gosset, 2019) (Tan, 2019) (Kell, 2019). Phishers use social engineering enticements to steal victims’ personal data and financial credentials over falsified websites, usually through spoofed emails.


Much work has been done in trying to find a solution to improve or minimise phishing and malware attacks. However, a review shows that most research areas focus on Artificial Intelligence (e.g., Bayesian networks, data mining, heuristics, machine learning, decision trees, classifiers and clustering), which have not yet succeeded in tackling the problem at hand. The authors found that the provided technical


measures are insufficient and that solutions closer to users’ behaviour and the way they interact with the systems should be part of a socio-technical security solution (Ferreira, 2018a).


Moreover, another review which only focuses on ransomware (Ferreira, 2018b), also concludes that pure detection solutions are not enough and that existing preventive ones do not work properly. There are not enough research efforts on prevention, backup and awareness solutions to fight ransomware attacks. Back to the human factor of these social engineering attacks, a close link with persuasion factors has been studied, which can assist in understanding of the complexity of means used by attackers to attain their goals (Ferreira, 2018b) (Ferreira, 2019).

Prevention measures have not been a strong focus of research in this area. Nevertheless, the authors believe that work on stopping phishing attacks using phishing emails, must be performed on the text of

^a  <https://orcid.org/0000-0002-0953-9411>

^b  <https://orcid.org/0000-0002-3121-4189>

^c  <https://orcid.org/0000-0001-8750-1120>

^d  <https://orcid.org/0000-0002-2521-0616>

subject lines. A recent work (Ferreira, 2017) has analysed email subject lines to find a pattern of use to embed in existing phishing detection techniques. However, this is still developing research.

This paper aims to identify the most commonly used terms over the past twelve years in phishing subject lines; which ones are used in combinations, on the same phrase; what clusters they provide; and what linguistic marks they may include. As this problem is only bound to exponentiate in the near future, out of the box measures are required to create more effective solutions.

2 BACKGROUND

This section supports the need to answer two relevant questions, why focus on phishing email subject lines and use network science to analyse those lines.

2.1 Email Subject Lines

Email subject lines are critical sources of information. They summarise, in small sentences, the content of an email and why it should be further perused. That sentence, together with the date and sender information, comprise the elements that help a user decide to become, or not, a victim. This decision is “free” of any external intervention, except for the degree of persuasion of those data and possible previous experiences and knowledge the user may have acquired up to that moment in time. A phishing email that is received has already passed anti-phishing controls. The pertinence of this study lies precisely in the possibility of providing tools to be used in such situations.

Besides the work already referred in the previous section, only one more seems to corroborate that the level of attention given to an email subject line is positively related to the likelihood that an individual will open that email (Vishwanath, 2011). So, the email subject lines comprise the main piece of information that triggers the user into deciding that an email is legitimate and should be opened (Balakrishnan, 2014). The authors decided to analyse the content of email subject lines and verify the relations between the used terms, resorting to network science (Section 2.2) and linguistic analysis. The authors could not find, until this moment, similar work in this area of research.

2.2 Network Science

The field of network science studies complex networks such as telecommunications, computer, biological, cognitive, semantic as well as social networks. A network is a catalogue of a system’s components or actors called *nodes* or *vertices* and the connections between them, called *links* or *edges* (Figure 1). The connections can represent a relationship between two nodes as in social networks, where the connection between two nodes represent the “friendship” relation between two people.

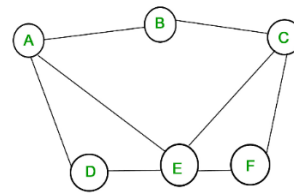


Figure 1: Example of a simple graph/network comprised of nodes A to F, and edges between them (A-B; E-C, etc).

Our world is permeated with systems that can be represented by networks (Barabási, 2002) (Watts, 2003). Graph drawing offers advantages for the visualisation of networks and visual exploration and analysis (von Landesberger, 2010). In terms of language and word relation, recognition is increasing that human language can be modelled with complex networks (Solé, 2010) (Markošová, 2008). To do this, connecting various components of a sentence in a graph or network can be used to verify if there are any main relations or clusters of relations, or other patterns that are used within the terms of email subject lines. This can help visualise behaviours that are not seen or found in any other way. Within networks, clusters are subgraphs consisting of strongly connected components which can be measured with a node-based clustering coefficient (Watts, 1998). Hubs are located between clusters to connect them as brokers (Figure 2). Clusters and hubs complement each other in the interpretation.



Figure 2: Examples of subgraphs illustrating a node degree ($n=3$), a cluster and a hub.

The analysis of a network structure provides topological insights and can be a starting place for detailed exploration at some point of interest. Parts of a network can be compared based on nodes and edges

according to their position in the network structure (Drieger, 2013). Due to the abstraction that networks allow, we can explore a text quickly and see, at a glance, how words are interconnected and contextually situated in the network.

3 METHODS

In order to generate a network with the content of the words within the phishing email subject lines the authors reproduced a method for sampling email subject lines, from previous research (Ferreira and Teles, 2019). A sample was selected from a reliable phishing archive source, namely, millersmiles.co.uk⁵. A total of 240 email subject lines was randomly selected from this archive (2008-2019). Since the phishing emails lists were chronologically ordered, 20 numbers were randomly generated for each year, and each number corresponded to a place in that list, starting at the top. The subject lines' content was parsed using a PHP script with the following phases: data cleaning, parsing, and results generation. For example, the null node was deleted, singulars and plurals from the same word or very similar word with typos, were merged (e.g., "alert" merged with "alerts" or "online" merged with "onlinesm"). The parser produced several outputs: i) nodes' selection and association to a unique identifier; ii) a table with word frequency; iii) a table with the pairs of connected words in the same subject line – each word has only a connection to all words in the sentence and connected pairs are the edges of the graph, while words are nodes; iv) the number of connections between the same words was counted (degree). After generating the full network a degree threshold equal or bigger than four (≥ 4) was used to trim the network, since there were many different 1-degree threshold edges, which were confusing and did not add extra value to the most common used terms and their relations. Also, as this was an undirected graph (all edges are bidirectional), when an edge was found between A and B, if the same was found between B and A, these would count as two instances of the same edge adding to their degree.

With the produced output, two .csv files were created, one with all the identified nodes (every unique word) and corresponding unique identifiers, and the other with the corresponding edges between those nodes (Figures 3 and 4). With the final

generated network, an analysis of frequency, clusters and hub extraction was performed.

Id	Label	Interval
8	a	1
9	abbey	3
10	about	8
11	abusing	1
12	access	10
13	account	80
14	accounts	2

Figure 3: Extract from nodes .csv file (before filtering). Id is the unique identifier for each node found in the sample.

s_ID	t_ID	source	target	Type	Weight-3	Sword	Tword
410	219	1	new	Directed	2		1 new
410	204	1	message	Directed	1		1 message
219	204	new	message	Directed	3	new	message
219	307	new	security	Directed	2	new	security
219	133	new	from	Directed	2	new	from
363	403	update	your	Directed	8	update	your

Figure 4: Extract from edges .csv file (before filtering). s_ID and t_ID are pairs of connected nodes (the edges).

Clusters have among themselves a relatively high network density and a high degree hub that usually connects those clusters, also known as communities. Such communities can also become more noticeable when visualised within a graph. A flowchart of the methodology for generating the word network is presented in Figure 5.

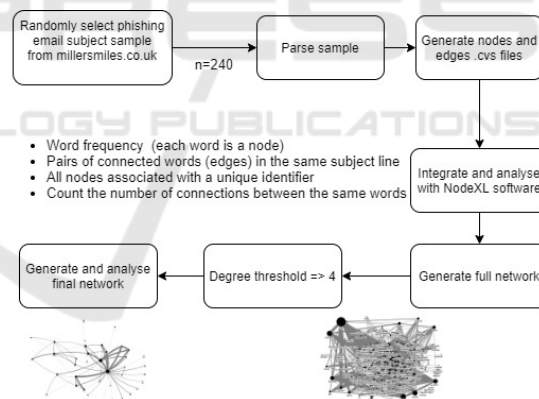


Figure 5: General methods to achieve the final network.

In addition, linguistic analyses, including lexical density calculation (i.e., the number of content words divided by the total number of words), were carried out to support findings of commonly used terms or groups of terms, in the referred period and corpus. The email subject lines were examined with four software tools and goals: (a) Corpógrafo (Sarmento, 2004) - to create a researchable corpus, generate n-grams and research concordances of most frequent

⁵ Millersmiles.co.uk is a recognized international source of free archived spoof and phishing emails.

tokens; (b) VISL (Bick, 2000) - to parse the text with the Constraint Grammar system based on lingsoft's ENGCG parser to make it syntactically/semantically researchable; (c) Sentiment Analysis Python NLTKDemo - to obtain sentiment analysis of each subject line to determine polarity of emotions; and (d) WordSmith Tools 7.0 (Scott, 2018), to research text previously parsed with VISL based on its syntactic/semantic function. Lexical density was also calculated with Online Utility (Online Utility, 2020).

4 RESULTS

For the full analysed sample, 449 nodes and 2765 edges were produced. The most frequent words (nodes) of the analysed sample are shown in Figure 6.

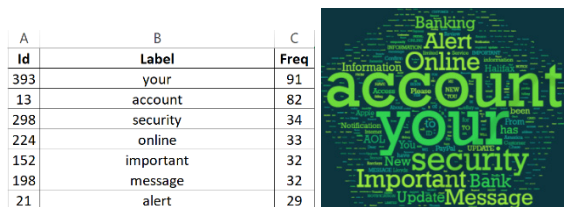


Figure 6: Most common words (nodes) in the sample (on the left – node frequency; on the right – word cloud).

The final analysed network (after trimming degree threshold ≥ 4) comprises 40 nodes and 82 edges (Figure 7). The most frequent nodes in the subject lines, i.e., “your” and “account”, constitute the two main hubs of the entire network. As there were many 1-degree threshold edges within the initial network those were not included in the final analysis. The most frequent nodes have bigger spheres and the most frequent edges have thicker connection lines. The maximum number of vertices and edges in a connected component is 38 and 81, respectively.

Figure 8 shows the six identified clusters within the network of words, with different colours, as well as their interconnections. The two main clusters (dark blue and orange) are again identified with the two most frequent words, the main hubs “account” and “your”. The direct connections between the main hubs are shown in Figure 9, with C1: C2 – C3 – C4; C2: C1 – C3 – C4; C3: C1 – C2 – C5; C4: C1 – C2 – C5; C5: C3 - C4. The number of nodes integrated in each identified cluster is very similar (apart from the 2-node cluster C6), C1 (darkblue) = 10; C2 (orange) = 9; C3 (red) = 7, C4 (brown) = 6, C5 (pink) = 6 and C6 (green) = 2. These clusters can be related to communities of words. In this case, each different community relates to a world-wide known

organization (except for C5), e.g., C1 – billing, payment, *Apple*; C2 – closed account, limited with dates, *Paypal*; C3 – important, security, update and *Halifax*; C4 – online, banking, alert and *America* (e.g., bank of America); C5 – receiving new messages; and C6 – with the name of a popular British bank – *Loyds Tsb*. The most common 4-WORD phrases: your account has been (8); your account will be (4); account has been limited (3); bank of america alert* (3); 3-WORD: account has been (10); your account has (8); your account information (5); bank of America (5); and 2-WORD: your account (33); has been (13); online banking (12); account has (10); your online (9); message from (8); you have (8); account information (8); update your (7).

The linguistic analyses show that subject lines contain 5 to 6 words on average, frequently use capital letters, exclamation signs and asterisks and, although not very frequently, contain spelling mistakes and grammatical errors. A lexical analysis was performed. Lexical/content words (e.g., nouns, adjectives, verbs, and adverbs) give the text its meaning and provide information regarding what the text is about. Nouns tell us the subject, adjectives tell us more about the subject, verbs tell us what they do, and adverbs tell us how they do it. When applying the tool Using English (UE, 2002) results show that, on the total sample, 21% are hard words (words difficult to spell) and 42% are long words and the lexical density with stop words is 34%, and without is 75%. Stop words are words in a language called function words, which are usually filtered out before certain natural processing language tasks are performed. Examples in English may include: “the”, “is”, “at”, “on”, etc. To confirm the above lexical density results, these were calculated with two different tools (Online Utility, 2020) (Analyze my writing, 2020), where the results were similar for lexical density with stop words. Figure 10 represents the lexical density of the sample, per year, with and without stop words (Online Utility, 2020). The readability gunning fog index (Gunning, 1952), which estimates the years of formal education a person needs to understand the text on the first reading, is 14.64 (Online Utility, 2020) and 15.51 (Analyze my writing, 2020).

However, the same sample was tested by other four free online tools to calculate the same index, (Bond, 2020) (WebFx, 2020) (RF, 2020) (UE, 2002), and an average of 10 was obtained. Considering the total average of 12, our sample is best understood by people with late high school (senior) and early university (undergraduate) education. Moreover, this tool (Analyze my writing, 2020), which only analysed 70 sentences from 240, showed that the most

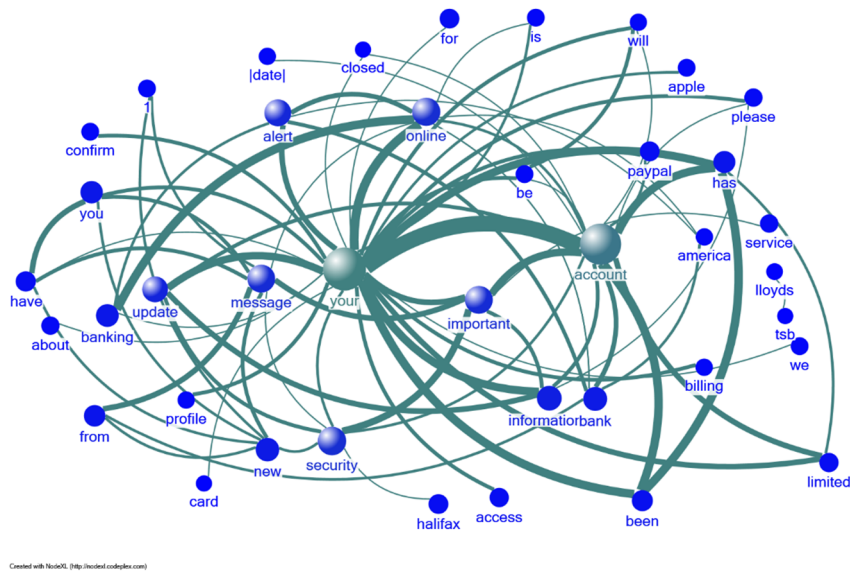


Figure 7: Network of words and relations obtained from the filtered sample (degree threshold >=4).

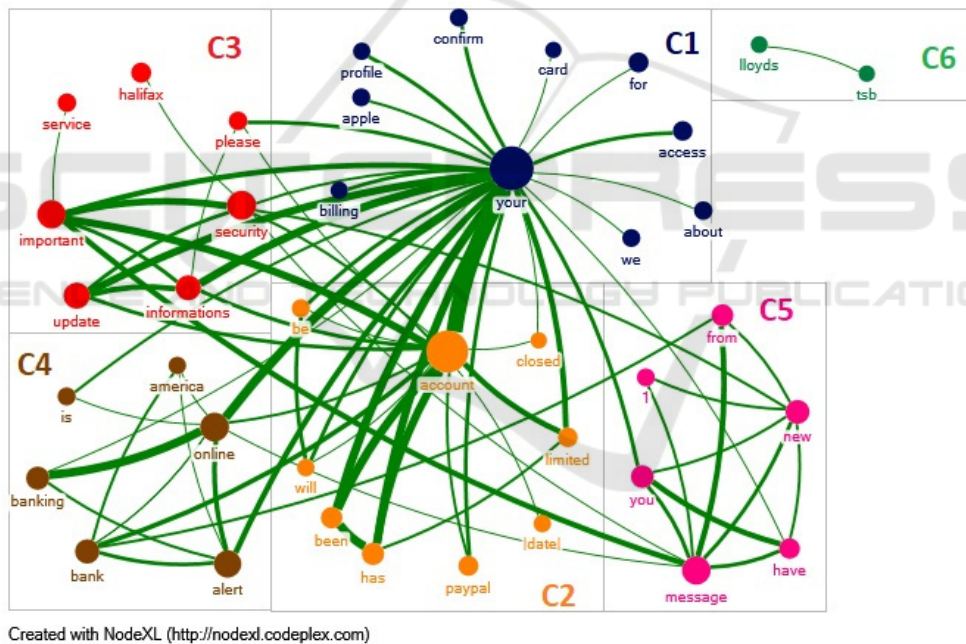


Figure 8: The network comprises 6 clusters, with 6 main nodes (hubs), identified with different colours: C1 (darkblue)='your', C2 (orange)='account', C3 (red)='security', C4 (brown)='online', C5 (pink)='message' and C6 (green)='loyds'.

frequent words on the sample are nouns (48%).

The sentiment analysis (Sentiment, 2004) showed that most subject lines express neutral sentiment and polarity, i.e., 58% (n=140) of all the subject lines in the corpus (n=240) carry impartial or noncommittal attitude. On the other hand, 35% (n=83) of the subject lines express positive sentiment, and only 7% (n=17) deliver negative sentiment. As shown in Figure 11, neutral and positive subject lines are also the most

consistent over the years, and seem to be the preferred choice for phishing text as opposed to negative ones.

Complementing the sentiment analysis with another tool (IntenCheck, 2018), Figure 12 shows a summary of communication styles, which can integrate words that creates pictures or sounds inside our mind (visual or audial), refer to feelings inside our bodies (kinesthetic) or to logical thinking and thoughts (rational). The analysis also includes the

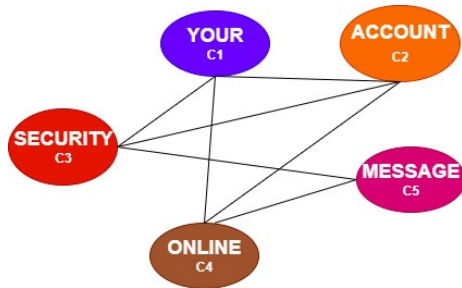


Figure 9: Direct connections between the main hubs.

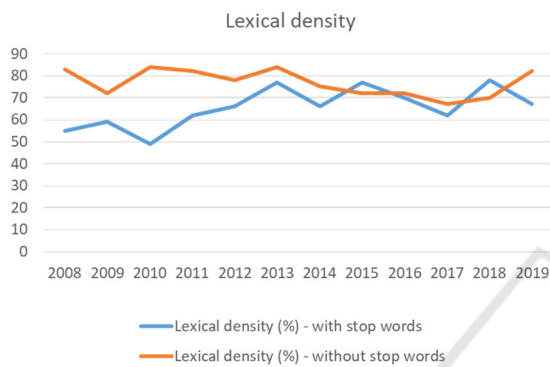


Figure 10: Lexical density analysis of the sample, per year.

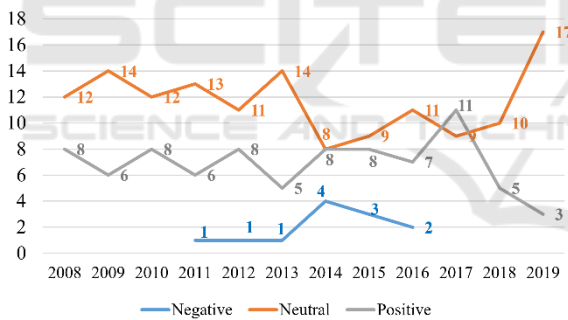


Figure 11: Sentiment analysis in the sample, per year.

expression of how a person thinks about time and their time preferences (most present and future preferences in the sample), as well as if people are motivated either by moving *away from* something they don't desire or by moving *towards* something that they want. In our sample the results move more towards something that is wanted (Figure 12). Finally, it is important to understand if a person communicates from his viewpoint or from other *perceptual positions*. Four perceptual positions were tested: I – “I”, “mine”, “self”; II – “he”, “she”, “is”, “her”, “one”, “you”, “your”, “yours”; III – “them”, “their”, “they”; and IV – “us”, “we”, “our”. In our sample, Position II is very dominant and Position IV is strongly used.

5 DISCUSSION

Online deception has been receiving some scholarly attention in the field of linguistics (Chiluwa, 2019). This comes from the recognition that language and the composition of texts essential roles in scammers' strategic approaches to their victims. This paper extends previous work (Ferreira and Teles, 2019) by employing a network analysis of common and interrelated used terms and a linguistic approach, to identify communities of words, as well as lexical density use, readability fog index, sentiment analysis, and communication styles in phishing email subject lines. This helps provide additional insights on some of the linguistic strategies used by scammers to influence users to open predatory emails. The methodological approach is more refined in comparison to previous works, which use national and/or context-specific databases (see Ferreira and Teles, 2019, for a discussion on used databases).

In our findings, the pronoun “your” is the most frequent word, suggesting that the email contains targeted/directed content. In the field of linguistics and linguistic psychology, the type and quantity of specific pronouns in a body of text can indicate intentions, psychological states and social relations. In certain typical stylometric analysis like lexical richness calculation without stop words, this word would probably be filtered. However, for phishing, this can constitute a differential mark from legitimate emails, and needs to be further tested. The second most common word is “account”, which focus on both bank and other online accounts from big organisations, such as *Apple, Paypal or Bank of America*. Credentials from these accounts constitute target data for phishers as they can easily get personal data to be used in more sophisticated attacks. These two most used words are the two main hubs connecting most clusters and creating communities of words related to big/known organisations (Figure 8).

Other terms revolve around verbs to take actions on those accounts (e.g., update, confirm, alert, limited, closed). Each cluster is strongly connected to two or three others (Figure 9), so they have a close relationship, and form among themselves a larger and stronger community. Tests with a larger sample need to be made to confirm this.

In terms of the most common lines/phrases used in the last 12 years (e.g., *your account has been, account has been limited, bank of America alert, message from, update your*), these are potentially very successful and so need to be tested in another study, with an international survey. This can help to further understand their degree of success.

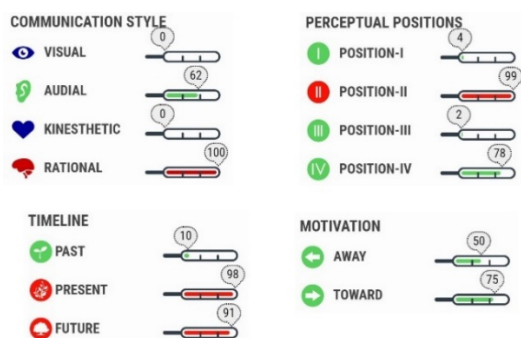


Figure 12: Attributes for communication style, timeline, motivation and perceptual positions, found in the sample.

The linguistic analyses confirm previous research that phishing email subject lines are very short and direct phrases (5/6 words on average), but comprise a rich plethora of attributes to claim for attention. The use of capital letters, punctuation and other signs can help create an ambience of persuasion. Part of this is confirmed by the analysis shown in Figure 12, as most words convey logical/rational thinking but they seem to be more audial than visual. In fact, the use of exclamation marks, capital letters and even asterisks can induce the reader in interpreting them as “loud”, “attention-grabbing” words, the goal for their use.

However, the words in our sample do not induce feelings inside our bodies. The results of sentiment analysis confirm that most found emotions in the text are neutral, followed by some positive ones. This can be explained by the fact that when people feel well, not threatened or even happy, are probably more prone to believe/trust in others. Furthermore, the high expression of neutral and positive sentiment and polarity in subject lines suggest linguistic sophistication that might be deliberately used by scammers to potentiate linguistic choices, similar to legitimate email/subject lines. Users may easily respond to subject lines such as “Important Security Message From HSBC Bank UK” or “Nationwide Account Notification”, since these are all neutral in sentiment and polarity, and do not request any *a priori* action. On the other hand, polarised subject lines, whether positive or negative may alert users about the legitimacy of the email due to the expressiveness of the requested actions. Nevertheless, positive subject lines might trigger subsequent action by the user due to the use of elements that look like legitimate, such as fabricated codes (e.g. positive - “Important Account Information CH671K0” or “ATTENTION: YOUR ITUNES ACCOUNT HAS BEEN FROZEN ID7041A4446615BDB59AAC”). Phishers might have learned this from previous attempts where

urgency was commonly used for distracting victims into not over thinking their actions (Ferreira, 2017). Such display may now be too obvious and suspicious.

In terms of timeline analysis, most of the sample is focused on moving towards something that is desired now or in the future. The past is rarely mentioned. This agrees with the phishers’ goals of inducing the victim into clear and quick action. Nevertheless, there is also some moving away from specific things, which in the case of phishing, can be the fact that accounts are going to be blocked or something undesired can happen if no action is taken soon. Finally, for the perceptual positions, again the authors believe this is a very important mark of these types of emails and language used by phishers. Communication is directed outside the sender, towards the victim (*your*), transferring the responsibility of said actions to the later (similar to the social proof principle of shared responsibility that may influence higher risk actions (Ferreira, 2019)).

For the lexical analysis, the use of stop words to calculate lexical density may interfere with the results. However, if seen per year, the lexical density between the samples with or without stop words is getting closer in recent years. The smaller the text, the more lexical density it will display, but the use of a higher percentage of nouns instead of other content words can be explained by the fact that verbs may denounce scammers more easily. Nouns are usually associated with text that is more difficult to read and understand; such difficulty might potentiate users to act and provide information to scammers. Formal text is usually full of nouns. Indeed, as education level is a well-known determinant of internet usage, so more educated people tend to be more frequent internet users (Goldfarb, 2008). This agrees with the readability content index result, which aims to reach late high school and undergraduate educated people. On the other hand, less educated people can also be an easy target because they may open those emails in search for clarification, and may act on them quickly, in fear of “loosing” something.

A comparison of email subject lines from phishing and legitimate emails is essential to obtain stronger conclusions. Another limitation was the use of NodeXL basic free version, which did not integrate advanced analysis functions.

6 CONCLUSIONS

A network and linguistic analysis of phishing email subject lines show that those should be used to find differential marks to better distinguish phishing

emails from other types of emails. Main hubs and clusters of words indicate that there are patterns and pre-defined goals in the way phishing emails have been crafted for the past twelve years. These need to be part of a larger sample test which can integrate both legitimate and scam emails to confirm, automate and improve existing phishing detection tools.

ACKNOWLEDGEMENTS

This work is supported by TagUBig - Taming Your Big Data (IF/00693/2015) from Researcher FCT Program funded by National Funds through FCT (Fundação para a Ciência e Tecnologia). Soraia Teles is individually supported by the Portuguese Foundation for Science and Technology (FCT; D/BD/135496/2018); PhD Program in Clinical and Health Services Research (PDICSS).

REFERENCES

- Analyze my writing, 2020. Available at: <http://www.analyzemywriting.com/>. Accessed in January 2020.
- Balakrishnan, R., Parekh, R., 2014. Learning to predict subject-line opens for large-scale email marketing. In *2014 IEEE International Conference on Big Data (Big Data)*, Washington, DC, 2014, pp. 579-584.
- Barabási AL., 2002. *Linked: The New Science of Networks*. Cambridge, MA: *Perseus Publishing*.
- Bick, E., 2000. The Parsing System" Palavras": Automatic Grammatical Analysis of Portuguese. In *a Constraint Grammar Framework*: Aarhus Universitetsforlag. Available at: <https://visl.sdu.dk/>.
- Bond, S., 2020. Gunning Fog Index. Available at: <http://gunning-fog-index.com/>.
- Chiluwa, I., Ovia, E., Uba, E., 2019. Attention Beneficiary!. *Handbook of Research on Deception, Fake News and Misinformation Online*, 421-438.
- Drieger, P., 2013. Semantic Network Analysis as a Method for Visual Text Analytics. *Procedia - Social and Behavioral Sciences*. Volume 79, Pages 4-17.
- Ferreira A., Chilro R., 2017. What to Phish in a Subject? In: Brenner M. et al. (eds) *Financial Cryptography and Data Security*. FC 2017. Lecture Notes in Computer Science, vol 10323. Springer, Cham.
- Ferreira, A., 2018. Why Ransomware Needs A Human Touch. In *2018 International Carnahan Conference on Security Technology (ICCST)*, 2018, pp. 1-5.
- Ferreira, A., Vieira-Marques, P., 2018. Phishing Through Time: A Ten-Year Story based on Abstracts. *Proceedings of the 4th ICISSP*, pages 225-232.
- Ferreira, A., Teles, S., 2019. Persuasion: How phishing emails can influence users and bypass security measures. *International Journal of Human-Computer Studies*. Volume 125, Pages 19-31.
- Goldfarb, A., Prince, J., 2008. Internet adoption and usage patterns are different: Implications for the digital divide. *Inf Econ Policy*. 20(1):2-15.
- Gosset, S., 2019. The top cybersecurity threats of 2010. *Built In*. Available at: <https://builtin.com/cybersecurity/cybersecurity-threats>. Accessed: November 2019.
- Gunning, R., 1952. *The Technique of Clear Writing*. McGraw-Hill International Book Co., NY, USA.
- Hong, J., 2012. The state of phishing attacks. *Commun. ACM*, 55 (1), 74-81.
- IntenCheck, 2018. Available at: <https://www.intencheck.com/how-it-works/>. Accessed Jan 2020.
- Kell, N., 2019. What will cybersecurity look like in 2020. *Techradar.pro*. Available at: <https://www.techradar.com/news/what-will-cyber-security-look-like-in-2020>. Accessed: November 2019.
- Markořová, M., 2008. Network model of human language. *Physica A: Statistical Mechanics and its Applications*. 387: 661-666.
- Online Utility, 2020. OnlineUtility.org. Available at: <https://www.online-utility.org/text/analyzer.jsp>. Accessed January 2020.
- RF, 2020. Readability Formulas. Available at: <https://readabilityformulas.com/free-readability-formula-tests.php>.
- Sarmiento, L., Maia, B., & Santos, D., 2004. The Corpógrafo-a Web-based environment for corpora research. *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC'2004)*.
- Sentiment Analysis, 2020. Python NLTK 2.0.4 text classification. Available at: <https://text-processing.com/demo/sentiment/>. Accessed Jan 2020.
- Scott, M. (2018). WordSmith tools version 7. Liverpool: Lexical Analysis Software, 122.
- Solé, RV., Corominas -Murtra, B., Valverde, S., Steels, L., 2010. Language networks: Their structure, function and evolution. *Complexity*. 15(6): 20-26.
- Tan, A., 2019. Top APAC security predictions for 2020. *ComputerWeekly*. Available at: <https://www.computerweekly.com/news/252474724/Top-APAC-security-predictions-for-2020>. Accessed: Nov 2019.
- UE, 2002. UsingEnglish.com. Available at: <https://www.usingenglish.com/resources/text-statistics/>. Accessed Jan 2020.
- Vishwanath, A., Herath, T., Chen, R., Wang, J., Rao, H., 2011. Why do people get phished? Testing individual differences in phishing vulnerability within an integrated, information processing model. *Decis. Support Syst.* 51, 3 (June 2011), 576-586.
- von Landesberger, T., Kuijper, A., Schreck, T., Kohlhammer, J., van Wijk, J., Fekete, J.-D., & Fellner, D., 2010. Visual analysis of large graphs. In *Proceedings of EuroGraphics: State of the Art Report*.
- Watts DJ., 2003. *Six Degrees: The Science of a Connected Age*. New York: *WW Norton & Company*.
- Watts, DJ., Strogatz, S. H., 1998. Collective dynamics of small-world networks. *Nature*. 393.
- WebFx, 2020. Readability Test Tool. Available at: <https://www.webfx.com/tools/read-able/>