Creation and Analysis of a Corpus of Scam Emails Targeting Universities

Grace Ciambrone
gkc5194@psu.edu
Pennsylvania State University
University Park, Pennsylvania, USA

Shomir Wilson
shomir@psu.edu
Pennsylvania State University
University Park, Pennsylvania, USA

ABSTRACT

Email-based scams pose a threat to the personally identifiable information and financial safety of all email users. Within a university environment, the risks are potentially greater: traditional students (i.e., within an age range typical of college students) often lack the experience and knowledge of older email users. By understanding the topics, temporal trends, and other patterns of scam emails targeting universities, these institutions can be better equipped to reduce this threat by improving their filtering methods and educating their users. While anecdotal evidence suggests common topics and trends in these scams, the empirical evidence is limited. Observing that large universities are uniquely positioned to gather and share information about email scams, we built a corpus of 5,155 English language scam emails scraped from information security websites of five large universities in the United States. We use Latent Dirichlet Allocation (LDA) topic modelling to assess the landscape and trends of scam emails sent to university addresses. We examine themes chronologically and observe that topics vary over time, indicating changes in scammer strategies. For example, scams targeting students with disabilities have steadily risen in popularity since they first appeared in 2015, while password scams experienced a boom in 2016 but have lessened in recent years. To encourage further research to mitigate the threat of email scams, we release this corpus for others to study.

ACM Reference Format:
Grace Ciambrone and Shomir Wilson. 2023. Creation and Analysis of a Corpus of Scam Emails Targeting Universities. In *Companion Proceedings of the ACM Web Conference 2023 (WWW '23 Companion), April 30–May 04, 2023, Austin, TX, USA.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3543873.3587303

1 INTRODUCTION

Email-based scams, such as phishing messages, blackmail threats, romance scams, illicit business opportunities, and others, have been a security hazard for email users for decades [3, 9]. Users are deceived into downloading malware, sharing sensitive information, sending money to criminals, or traveling to a foreign country where they may be kidnapped and murdered [1, 2]. In spite of a large body of work on email scam filtering and scam resistance education for email users, these scams continue to generate victims [3, 20].

WWW '23 Companion, April 30–May 04, 2023, Austin, TX, USA

ACM ISBN 978-1-4503-9419-2/23/04.
https://doi.org/10.1145/3543873.3587303

Universities are a unique environment for scam emails for several reasons. College students underestimate their susceptibility to email scams [13], likely due to their relative lack of experience with email, making them an especially vulnerable population. Universities' email systems and open directories provide a large population for scammers to easily target. This ease applies both to quantity of addresses and types of scams: for example, job opportunities for students, deans requesting favors from their faculty, or warnings to users that their email inboxes are full. In general, greater availability of the contents of scam emails would aid the research community in studying them and improving user education strategies, but users' privacy concerns inhibit the creation of scam email corpora. However, universities are unusually transparent environments, often posting institutional training materials in public on their websites. This transparency provides an opening for the creation of a needed resource.

We present a corpus of 5,155 English language scam emails scraped from the information security websites of five large universities in the United States: Arizona State University, Brown University, University of Michigan, University of New Hampshire, and University of North Carolina at Chapel Hill. These emails were posted by staff at each university as examples of scams to be aware of, and they have already been stripped of the recipient-related information that would typically pose privacy concerns. We describe the collection of this corpus and then topical and temporal trends that it illustrates. Using Latent Dirichlet Allocation (LDA), an approach to topic modeling in text [4], we identify common themes of scams in the corpus. Some themes match anecdotal expectations for scams, such as emails about changing one's password or email account problems. Other themes are less well-known but quantitative evidence suggests they are common, such as opportunities for students with disabilities. Our analysis further shows how some themes are trendy while others are perennial across the available period, from mid-2014 through 2022. The results of this work can be used to enhance scam resistance education at universities, with some observations also germane to other institutions. We also release the corpus for further study.[1]

2 RELATED WORK

We use the term *scam emails* (and *email-based scams*) to refer to a broad category of malicious emails: phishing, blackmail threats, illicit business opportunities, illegitimate employment opportunities, ads for fraudulent online retailers, and others. This mirrors a prior categorization by Jakobsson [8], who identified scam emails as those that aim to deceive users for financial gain. Prior research

[1] https://shomir.net/data/university_scams/index.html

on detecting these emails has covered both technical and social aspects. Siadati et al. [14] describe several technical countermeasures, including authentication, blacklisting, and content-based filtering. Among the content-based methods, Fang et al. [6] used recurrent convolutional neural networks to model phishing email headers and bodies simultaneously at word- and character-levels. While this approach and similar ones demonstrate high performance, phishing (and more broadly, scam emails) remain a persistent problem.

Other efforts have studied how email users avoid scams or fall victim to them. Wash [19] investigated how information technology experts identify phishing emails in their inboxes and described a three-stage process of noticing discrepancies, becoming suspicious, and then either deleting or reporting the message. Hadnagy and Fincher [7] examined why scam victims engaged with scammers and provided a variety of recommendations for reducing user susceptibility. Within a university context, Diaz et al. [5] studied the effects of phishing awareness training, and counterintuitively found that it increased the rate that students clicked on phishing links.

Additionally, some focused efforts have created and studied corpora of scam emails. Pan et al. [12] built a corpus of scam emails in English, French, and Russian by scraping Anti-Fraud International, a web forum. Similar to our work, theirs identified common topics in scam emails, but it lacked a temporal dimension and focused on emails in a general (i.e., non-academic) context. The CLAIR collection of scam emails [15] consists of "Nigerian" fraud emails dating from 1998 to 2007, differing from our effort by its narrow focus and lack of recent updates. Another study [11] compared topics in scam emails sent to one university's email addresses with publicly available anti-scam user education at several different universities and identified divergences. However, that effort relied on content tags produced by the university's Office of Information Security, as the email contents were unavailable. In contrast, our scam email collection is the largest available for academic settings and for any setting one of the largest released for public use.

3 CORPUS CREATION

Many universities are aware of the threats that email scams pose to their student population and employees. As part of mitigating this threat, some universities have a system in place that allows students to forward suspicious emails to their information technology (IT) office. Figure 1 shows a typical solicitation [18] to forward suspicious emails for expert examination. After being reviewed by staff, confirmed scams are then published to a site in an effort to bring awareness to malicious email trends affecting the university community. These sites typically consist of a list of scam email subject lines, bodies, and dates. These websites containing data are sometimes called "Phish Bowl[s]," a play on words involving the key term, *phishing*. In this paper, we consider phishing emails and scam emails as interchangeable terms that encompass all socially engineered email attacks. Our definition is consistent with the 2022 State of the Phish Report [3] and with terminology provided by the universities from which we collected our data.

We began corpus creation by informally using Google Search to locate university IT departments that had posted large numbers of scam emails. Universities do not have a standard format for posting these emails, and to make collection scaleable, we focused on pages

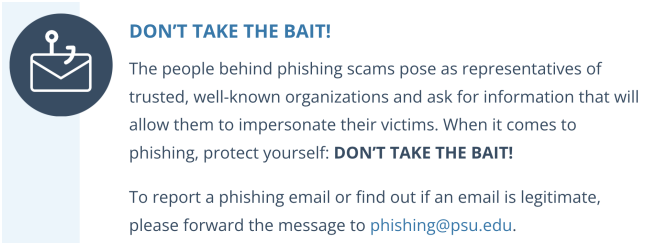

Figure 1: Example solicitation to forward email scams to a university's IT department [18]

University	Total	With Dates
Arizona State University	1,616	791
Brown University	940	916
University of Michigan	643	641
University of New Hampshire	332	332
University of North Carolina at Chapel Hill	1,624	920
Total	**5,155**	**(3,600)**

Table 1: Number of scam emails in the corpus collected from each university; those posted with dates are a subset of the total

by five universities that each posted large quantities: Arizona State University, Brown University, University of Michigan, University of New Hampshire, and University of North Carolina at Chapel Hill. We note that all these universities are large, research-oriented institutions in the US; their IT departments appeared to be most active at sharing scam emails. Additionally, the English language constraint was necessary to match the researchers' expertise.

Once we identified pages to gather scam emails from, we used Python scripts and regex pattern matching to collect subject lines, email bodies, and dates for each email. Dates were available for only a subset of emails. We observed that the IT departments had redacted suspicious links and attachments from email bodies, and had removed personally identifiable information, such as the receiving email address. We then reviewed emails we collected to ensure that the contents of our corpus were consistent with the data published on these websites. In its downloadable package, the corpus is structured as a set of directories, where each directory holds text files containing individual scam emails collected from a particular university. Each text file contains the subject line, date, and email body with preserved line breaks.

Table 1 breaks down the corpus contents by university. The total number of emails in the corpus is 5,155, and 3,600 of those have dates associated with them. The oldest is from August 14, 2014 and the most recent is from August 3, 2022, with roughly even distribution across that time segment. This corpus represents, to our knowledge, the largest contemporary collection of scam emails publicly available, for university or non-university contexts.

Topic Name	Keywords	% of Corpus
Email Account	account, email, mail, click, link, update, mailbox, e, dear, upgrade	35%
Personal Request	need, know, let, thanks, email, available, help, time, information, brown	16%
Document	document, view, message, link, file, removed, email, attached, shared, click	14%
Miscellaneous	professor, university, number, available, phone, brown, department, regards, best, information	9%
Password	password, message, new, mail, outlook, access, messages, service, e, email	9%
Employment Opportunity	job, email, time, weekly, work, week, number, position, mobile, alternative	8%
Order/Payment	payment, order, date, invoice, account, bank, usd, customer, details, number	6%
Students w/ Disabilities	students, email, services, disabilities, employment, university, school, academic, work, week	2%
Blackmail	video, training, contacts, state, bitcoin, retirement, e, click, de, link	1%

Table 2: Top ten keywords per topic

4 ANALYSIS

We explore scam emails sent to universities in two ways: thematic analysis to reveal common kinds of scams, and temporal analysis to reveal trends in scams over the available period.

4.1 Thematic Analysis

To identify prevalent themes in scam emails sent to universities, we applied topic modelling algorithms and analyzed results. Initially we compared several topic models, including Latent Dirichlet Allocation, Non-Negative Matrix Factorization, and Gibbs Sampling. We also experimented with a variety of other factors, such as length of bi-grams, number of topics, stratification by data source, and stratification by time. The researchers manually judged the meaningfulness of the topics generated by each permutation and found that topic model that produced the most cohesive results was a nine-topic unigram Gibbs sampling model. Nine of the topics in its output appear to be meaningful categorizations, and the remaining topic appears to be a collection of a variety of different emails with little apparent relation with one another.

In Table 2, we observe the order of popularity of topics and their most popular keywords. The 'Email Account' scam is the most common. These scams often pretend to notify users that their email account has run out of storage, or that a message is awaiting them in their inbox. 'Personal Request' scams frequently pose as from professors, deans, or other employees at the university. Out of all email scams, these tend to be the shortest in word count and commonly ask if the recipient is available. These scams are popularly known to lead to a request for a gift card [10, 16, 17]. 'Document' scams purport that a document has been shared with the recipient, and lead the user to follow a link. The scams in the 'Miscellaneous' category lack thematic cohesion. 'Password' scams involve notifying the user that their password has been updated or needs to be updated. 'Employment Opportunity' scams pose as from an employer or as a university staff member offering employment opportunities to students. 'Payment' scams notify the user that they have made a payment or need to make a payment, or have placed an online order. 'Students with Disabilities' scams are a topic we observed across all universities, specifically offering students with disabilities employment opportunities. These often pretended to be sent by a university staff member, such as a mental health counselor. Finally, scams in the the 'blackmail' topic claim to hold harmful information on the user and demand payment.

While some topics reinforce anecdotal evidence, others are less anticipated. For example, password, employment opportunity, and blackmail scams are all anecdotally known to be common. However, students with disabilities is a lesser known threat, and appears especially designed for university recipients. Additionally, this data helps us understand that email account scams are the most popular, followed by personal request scams.

4.2 Trends Over Time

Figure 2 illustrates the popularity of the topics over time. Each bubble in the diagram corresponds to a single scam email in the corpus, with color indicating which university it was collected from. Thus, each line of bubbles shows when instances of a given topic occurred over time.

We observe trends in topics over time. Notably, both document and email account scams peaked between 2016 and 2017, with multiple universities participating in both peaks. We also observe that while password scams have remained at consistent levels, employment scams have varied in frequency with an uptick in 2020 and part of 2021 and 2022. One possible explanation for these abrupt changes could be scammers exploiting sudden increases in virtual recruitment due to COVID-19. Other recent developments include the rise of order/payment scams since 2021 and also personal request scams since 2019. While less popular, students with disabilities is also increasing in popularity with most instances occurring since summer of 2019. Blackmail scams, despite having a long history, also appear to be rising in recent years.

5 DISCUSSION AND CONCLUSION

The creation of a corpus of scam emails targeting universities is a significant resource for security and NLP researchers because it allows for the analysis of scam topics for this specific audience, which includes a vulnerable audience of students. Prior to this work, limited empirical evidence existed relating to scam emails. Thus, scam email user education at universities is currently largely based on anecdotal evidence, smaller public datasets, or privately

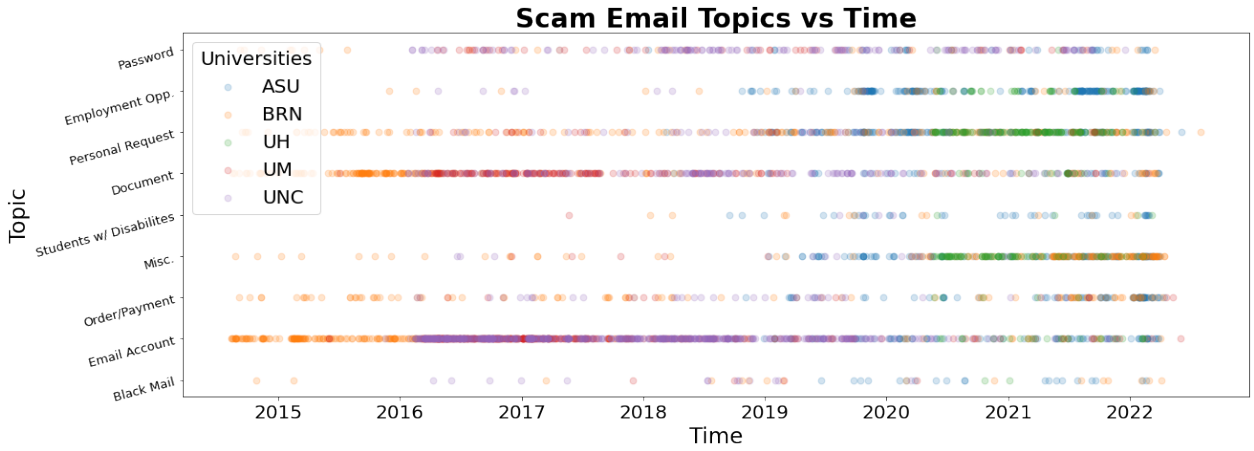

Figure 2: Gibbs topic model, topics over time

held datasets. By releasing this corpus for study, we are enabling researchers to understand the landscape of scam emails at universities without barrier.

Our analysis of the corpus details the eight most frequent scam topics that have affected intuitions of higher education since 2014. Equipped with this information, IT departments at universities can update their user education as needed to help protect their vulnerable student population. Scam emails at universities are unique in their specific deception tactics to target students, often posing as members of their organization, or interested employers.

Extending our observations of scam emails at universities to the general population of scam emails, our work supports that clearly defined topics exist and that scammer strategies may change over time. Additionally, our discovery of topics not found in anecdotal evidence demonstrates the importance of checking assumptions, as well as the importance of the collection and analysis of empirical data in the general scam population.

Future work on this topic can include wider collection of scam emails from the web, using bootstrapping methods and automating the search for source pages. Also, how email users at universities react to popular scams remains unknown, with possible differences by topic and by subpopulations (e.g., faculty versus students). Finally, differences in these scams likely exist by language and university location, with international variations dependent upon cultural differences in effective scam topics and trust.

REFERENCES

[1] 2002. *The 419 Scam, or Why a Nigerian Prince Wants to Give You Two Million Dollars.* https://www.informit.com/articles/article.aspx?p=25269

[2] 2004. *SA cops, Interpol probe murder.* https://www.news24.com/News24/SA-cops-Interpol-probe-murder-20041231

[3] 2022. *2022 State of the Phish.* https://www.proofpoint.com/us/resources/threat-reports/state-of-phish

[4] Levent Bolelli, Şeyda Ertekin, and C Lee Giles. 2009. Topic and trend detection in text collections using latent dirichlet allocation. In *European conference on information retrieval.* Springer, 776–780.

[5] Alejandra Diaz, Alan T Sherman, and Anupam Joshi. 2020. Phishing in an academic community: A study of user susceptibility and behavior. *Cryptologia* 44, 1 (2020), 53–67.

[6] Yong Fang, Cheng Zhang, Cheng Huang, Liang Liu, and Yue Yang. 2019. Phishing Email Detection Using Improved RCNN Model With Multilevel Vectors and Attention Mechanism. *IEEE Access* 7 (2019), 56329–56340. https://doi.org/10.1109/ACCESS.2019.2913705

[7] Christopher Hadnagy and Michele Fincher. 2015. *Phishing dark waters: The offensive and defensive sides of malicious Emails.* John Wiley & Sons.

[8] Markus Jakobsson. 2016. *Understanding social engineering based scams.* Springer.

[9] Neville L Johnson, Tami Devitt, Jesse Phillis, Jocelynn Arreola, and Kate Baggiano. 1998. Caught in the Act. *Los Angeles Lawyer* (1998), 32.

[10] University of Illinois at Chicago. 2022. *Beware of Gift Card Scams.* https://it.uic.edu/news-stories/beware-of-gift-card-scams/

[11] Duo Pan, Ellen Poplavska, Nora O'Toole, and Shomir Wilson. 2021. Comparing Scam Emails and Email User Education at Universities. *The Eighteenth Symposium on Usable Privacy and Security (poster abstracts)* (2021).

[12] Duo Pan, Ellen Poplavska, Yichen Yu, Susan Strauss, and Shomir Wilson. 2020. A Multilingual Comparison of Email Scams. *The Seventeenth Symposium on Usable Privacy and Security (poster abstracts)* (2020).

[13] Evan K. Perrault. 2018. Using an Interactive Online Quiz to Recalibrate College Students' Attitudes and Behavioral Intentions About Phishing. *Journal of Educational Computing Research* 55, 8 (2018), 1154–1167. https://doi.org/10.1177/0735633117699232

[14] Hossein Siadati, Sima Jafarikhah, and Markus Jakobsson. 2016. *Traditional Countermeasures to Unwanted Email.* Springer New York, New York, NY, 51–62. https://doi.org/10.1007/978-1-4939-6457-4_5

[15] Rachael Tatman. 2002. *Fraudulent E-mail Corpus.* https://www.kaggle.com/datasets/rtatman/fraudulent-email-corpus

[16] Carnegie Mellon University. 2019. *Gift Card Scam Uses Phony Academic and Administrative Leadership Emails.* https://www.cmu.edu/iso/news/2019/gift-card-scam-emails.html

[17] Oklahoma State University. 2022. *Gift Card Scams | Oklahoma State University.* https://it.okstate.edu/security-education/giftcardscams.html

[18] Pennsylvania State University. 2022. *Phishing | Penn State Information Security.* https://security.psu.edu/education-training/phishing/

[19] Rick Wash. 2020. How Experts Detect Phishing Scam Emails. *Proc. ACM Hum.-Comput. Interact.* 4, CSCW2, Article 160 (2020), 28 pages. https://doi.org/10.1145/3415231

[20] Pengcheng Xia, Haoyu Wang, Xiapu Luo, Lei Wu, Yajin Zhou, Guangdong Bai, Guoai Xu, Gang Huang, and Xuanzhe Liu. 2020. Don't Fish in Troubled Waters! Characterizing Coronavirus-themed Cryptocurrency Scams. In *2020 APWG Symposium on Electronic Crime Research (eCrime).* 1–14. https://doi.org/10.1109/eCrime51433.2020.9493255