

Sigrid Anne Hafsahl Karset

Analyzing Email Phishing Trends Through the Creation of an Email Phishing Collection Model

Master's thesis in Information Security

Supervisor: Ernst Gunnar Gran

Co-supervisor: Erik Hjelmås

June 2023

Sigrid Anne Hafsaahl Karset

Analyzing Email Phishing Trends Through the Creation of an Email Phishing Collection Model

Master's thesis in Information Security
Supervisor: Ernst Gunnar Gran
Co-supervisor: Erik Hjelmås
June 2023

Norwegian University of Science and Technology
Faculty of Information Technology and Electrical Engineering
Dept. of Information Security and Communication Technology



Abstract

Phishing is one of the leading vectors of cyberattacks having led to millions of monetary losses and damages every year. Due to the continued digitization of the world, more of our operations are converted onto digital mediums, creating a larger and larger pool of possibilities for malicious cyber actors. To keep up and provide insight necessary to combat phishing, this thesis presents an analysis of phishing conducted over the email medium from the years 2016 throughout 2022 through the creation of an email phishing collection model.

Influenced by prior work and published reports, as well as being populated by findings from the selected phishing emails themselves, an email phishing collection model is created. The model introduces and focuses on the properties of Content, Target, Method, and Impersonation, drawn from the structure of an email and definition of phishing.

By utilizing this model developed, phishing emails from the scoped years are collected and further analyzed. Analyzing the evolution of phishing from 2016 throughout 2022 shows that approaches tied to the objective (Target) and method of achievement (Method) of a phishing email popular in 2016 have remained consistent, while the essence of the email (Content) and from whom the email appears to be from (Impersonation) have displayed varying results. Events such as the COVID-19 pandemic, governmental operations, and the Christmas season are occurrences having influenced the trends observed, while advances on the side of detection technologies is a notable factor changing the approaches utilized within email phishing.

The model created presents a standardized way of conducting email phishing collection, providing a universal and replicable approach for the collection and subsequent analysis. Through the utilization of this model and additional analyses, insight on the evolution and trends within email phishing could be highlighted, which again can give an indication of any future phishing behavior.

Sammendrag

Phishing er en av de ledende angrepsvektorene i dagens digitale samfunn, som har medført store mengder monetære tap og skader de siste årene. Grunnet den kontinuerlige digitaliseringen av verden blir større deler av våre hverdagslige oppgaver ført over på digitale medium, noe som skaper fler og fler muligheter for ondsinnede cyber aktører. For å holde følge med, samt fremheve kunnskap nødvendig til å bekjempe phishing, vil denne Masteroppgaven presentere en analyse av phishing e-post fra årene 2016 ut 2022 gjennom utviklingen av en phishing e-post innsamlingsmodell.

Inspirert av tidligere arbeid og relevante rapporter, samtidig som å basere seg på elementer fra phishing e-postene selv, er en phishing e-post innsamlingsmodell utviklet. Modellen introduserer og fokuserer på områdene Content, Target, Method, og Impersonation, hvorav områdene selv er knyttet opp mot strukturen til en e-post og definisjonen av phishing.

Ved å benytte denne utviklede modellen er phishing e-poster fra de definerte årene samlet inn og videre analysert. Analyse av utviklingen av e-post phishing fra 2016 til og med 2022 viser at populære tilnæringer knyttet til målet (Target) og metode benyttet (Method) fra 2016 fortsatt er populære den dag i dag, mens selve essensen av e-posten (Content) og hvem e-posten tilsynelatende er fra (Impersonation) er varierende fra år til år. Hendelser slik som COVID-19 pandemien, statsiverksatte operasjoner, og julesesongen utheves som påvirkende faktorer til trendene observert. I tillegg pekes fremgangen innenfor deteksjonsteknologier som innflytelsesfull for utviklingen av e-post phishing.

Den utviklede modellen åpner opp for en standardisert tilnærming til innsamling av phishing e-post, en tilnærming som gir en universal og replikerbar metode for innsamling og videre analyse. Gjennom bruk av denne modellen med tilhørende analyser, ble det fremhevet informasjon nødvendig for å forstå trendene innenfor phishing, som igjen kan være behjelpelig i å identifisere fremtidig phishing atferd.

Acknowledgements

I would like to express a deep gratitude towards both of my supervisors Ernst Gunnar Gran and Erik Hjelmås. Your guidance throughout the entirety of the project, from the pre-project till the finished thesis, has been exceptional. The results of this thesis would not have been achieved without your aid. Thank you so much.

Further, a thank you must be given to Janne Cathrin Hetle Aspheim and the rest of Statistikkhjelpen for your valuable advice regarding statistical representations.

Contents

Abstract	iii
Sammendrag	v
Acknowledgements	vii
Contents	ix
Figures	xiii
Tables	xv
1 Introduction	1
1.1 Background	1
1.2 Problem Description	1
1.3 Thesis Definition	2
1.4 Scope	3
1.5 Research Questions and Planned Contributions	3
1.6 Structure	4
2 Theory	7
2.1 History and Evolution of Phishing	7
2.2 Phishing Approaches	9
3 Methodology	11
3.1 Literature Review	12
3.1.1 Alternative Methodology	13
3.1.2 Criticism of Chosen Methodology	13
3.2 Defining Collection Properties	13
3.2.1 Alternative Methodology	14
3.2.2 Criticism of Chosen Methodology	14
3.3 Collection	14
3.3.1 Alternative Methodology	15
3.3.2 Criticism of Chosen Methodology	15
3.4 Model Population	16
3.5 Analysis	16
3.5.1 Alternative Methodology	18
3.5.2 Criticism of Chosen Methodology	18
4 Related Work	19
4.1 Prior Work	19
4.2 Research Papers	20
4.3 Reports	21

5	Data Collection	25
5.1	Collection Sources	25
5.1.1	Ticketing System	25
5.1.2	MailRisk	25
5.2	Model Properties	26
5.2.1	Email Structure	26
5.2.2	Phishing Properties	27
5.2.3	Model Properties Overview	29
6	Dataset Analysis	31
6.1	Collection Properties	31
6.1.1	Content	31
6.1.2	Target	34
6.1.3	Method	35
6.1.4	Impersonation	36
6.2	Aggregated Analysis	37
6.2.1	Content	37
6.2.2	Target	42
6.2.3	Method	44
6.2.4	Impersonation	47
6.2.5	Dates	50
6.3	Property Relationships	51
6.3.1	Target-Content	51
6.3.2	Method-Target	59
6.4	Summary of Findings	61
7	The Email Phishing Collection Model	63
7.1	Model Use	64
7.2	Content	65
7.3	Target	65
7.4	Method	66
7.5	Impersonation	67
7.6	Model Overview	68
8	Discussion	71
8.1	Top Content Categories	71
8.2	Evolution of the CEO Scam	72
8.3	Credentials	74
8.4	Evolution of Infect	75
8.5	URL	76
8.6	Targeted Brands	79
8.7	Spoofing	81
8.8	Date Activity	86
8.9	Summary of Discussion	88
9	Conclusion	91
9.1	Research Questions	92
9.1.1	RQ1	92

9.1.2	RQ2	93
9.1.3	RQ3	93
9.2	Challenges	94
9.3	Future Work	95
9.3.1	Continued Analyses	95
9.3.2	Process Automation	95
9.3.3	Investigate Findings	95
Bibliography	97
A Full Data Summary	107
A.1	2016	107
A.2	2017	114
A.3	2018	121
A.4	2019	127
A.5	2020	133
A.6	2021	142
A.7	2022	148
B Content Categories	155

Figures

3.1	Methodology Overview	11
3.2	Methodology - Defining Collection Properties	13
3.3	Methodology - Collection	15
3.4	Methodology - Model	16
3.5	Methodology - Dataset Analysis	17
5.1	SMTP Email Format	27
6.1	Invoice Phishing Email	32
6.2	Document Shared Phishing Email	32
6.3	CEO Scam - Transfer Money Phishing Email	33
6.4	Post Package Phishing Email	33
6.5	CEO Scam - Gift Card Phishing Email	34
6.6	Content Distribution	38
6.7	Content Evolution (Top 10)	39
6.8	Rank Evolution - Invoice, Document Shared	40
6.9	Rank Evolution - Confirm / Update Account Information, Extortion, Refund	41
6.10	Rank Evolution - CEO Scam - Gift Card, Post Package	41
6.11	Target Evolution	43
6.12	Method Distribution	44
6.13	Method Evolution	45
6.14	Attachment Type Evolution	46
6.15	Impersonation Distribution	47
6.16	Impersonation Evolution	49
6.17	Date Distribution	50
6.18	Date Distribution Per Year	51
6.19	Target-Content	52
6.20	Credentials (Target-Content)	53
6.21	Money (Target-Content)	54
6.22	PII (Target-Content)	54
6.23	Credit Card Details (Target-Content)	55
6.24	Business Information (Target-Content)	55
6.25	Infect (Target-Content)	56

6.26 Target-Content Diversity Evolution	58
6.27 URL (Method-Target)	59
6.28 Attachment (Method-Target)	59
6.29 Communication (Method-Target)	59
6.30 Calendar Invite (Method-Target)	59
6.31 Credentials (Method-Target)	60
7.1 Model Overview	63
7.2 Model Use	64
7.3 Model Exemplified	69
8.1 Evolution - CEO Scams	73
8.2 Evolution - Infect	75
8.3 Method-Target	77
8.4 HTML - Website	78
8.5 HTML - Local	78
8.6 HTML - Plain Text	79
8.7 HTML - Obfuscated	79
8.8 Evolution - Apple	80
8.9 Spoofing Distribution	83
8.10 DMARC Trends from dmarc.com	85
8.11 Overlapping Spoofing and DMARC evolution	85
8.12 Tax Phishing - Heat-Map	87
8.13 Week Distribution	88

Tables

5.1	Model Properties	30
6.1	Phishing Email Distribution	37
6.2	Target Distribution	42
6.3	Target-Content Diversity	57
7.1	Model - Content Categories	65
7.2	Model - Target Categories	66
7.3	Model - Method Categories	67
7.4	Model - Impersonation Categories	67
7.5	Email Phishing Collection Model	68
8.1	Spoofing Distribution	82
8.2	DMARC Trends from dmarc.com	84

Chapter 1

Introduction

1.1 Background

The continuous digitization of the world has led to more and more of our day-to-day and business processes being converted onto digital mediums. Even though many of these tasks and processes are digitized, and some automated as well, humans are still an essential part of the operation of these tasks. As humans are psychologically driven, they can be manipulated into performing actions that otherwise would not occur without the human factor involved. This human manipulation is referred to as social engineering and is a prevalent method used in order to gain unauthorized access to information and systems, or direct monetary gains. A dominant form of social engineering related to digital mediums is *phishing* [1]. Phishing utilizes digital mediums, such as email, telephony, social media, and messaging platforms, in order to conduct social engineering attacks.

This thesis aims to create a deeper understanding of the phishing phenomenon through an analysis of email phishing trends from the years 2016 throughout 2022 from an organizational perspective, including an investigation into the causes for these trends in order to determine any correlating factors. To conduct this collection and trend analysis, a phishing collection model is created, aiming to provide a standardized and replicable way to collect and analyze email phishing.

1.2 Problem Description

Activision, Reddit, Mailchimp, DropBox, American Airlines, Uber, DoorDash, Cisco, and Twilio. These are just a handful of the organizations that have been subjected to data breaches reported the last year [2–10] who all have one thing in common; They were all achieved by exploiting the human factor, specifically through the usage of phishing, either as the initial breach or aiding in further compromise. Even though the phishing phenomenon has been around since the mid 90's, it continues to be a prevalent method used to breach computer systems, accounting

for the highest percentage (41%) of the initial access metrics in IBM's 2023 Threat Intelligence Index [11]. From these metrics, it is evident that phishing remains a considerable problem in today's society.

Although there exists an array of literature and reports on the phishing phenomenon [12], there is a lack of studies investigating phishing trends from a broader time perspective, seeing the trends in correlation with changes in external factors. On another note, new phishing techniques are constantly appearing, causing changes to the phishing landscape, further highlighting the importance of continual research on the phishing subject. It is of interest to conduct this study as this research can provide insight into how and why phishing trends are changing, including identifying patterns that may indicate future behavior.

Further, both published research and reports from companies and groups concerning phishing trends and statistics all display varying structures, terminology and metrics in their published materials [13–18], making it cumbersome to both compare and analyse them. On a similar note, they do not provide a practical method to compare your own data to that of these reports either. A common methodology for both collecting and analyzing phishing data may be of benefit, creating a uniform way of approaching the subject matter.

1.3 Thesis Definition

In this thesis, a collection and analysis of email phishing data over several years will be conducted, with the aim to highlight how the phishing trends have evolved throughout the years. The findings from these phishing trends will be subjected to an analysis in order to determine why the trends appear as they do, such as being influenced by external factors, as well as if these trends may give any indication of future phishing behavior.

In order to perform this collection and analysis, an email phishing collection model will be developed. The model created will not only be applied to this specific collection, however should be a model that can be universally used. To create such a model for universal use, prior literature depicting email phishing will be analyzed in order to identify and select appropriate model properties. The model developed will be utilized for the phishing data collection, while also being formed by the collection itself. The findings from the collection will aid in populating and confirming the developed model.

1.4 Scope

Due to the wide range of approaches and technologies that can be utilized in phishing, this thesis will focus in on phishing conducted over email. This means that other forms of phishing, such as phishing over social media or phishing over instant messaging platforms, will not be subjected to any collection or analysis. The reasoning for scoping the thesis to focus in on email instead of other forms of phishing is both due to the fact that email is the medium over which the most phishing activity occurs [19], as well as the availability of the desired phishing data.

As the phishing data is mainly based on user-reported phishing emails from corporate entities, the results and findings is scoped towards organizational environments as opposed to private. In addition, these corporations are primarily based in Scandinavia, meaning that the result will reflect the trends from this area.

The data available for analysis consists of phishing emails spanning from mid February of 2016 until present day. The collection scope starts in 2016 and consists of data all the way throughout 2022, giving the collection and analysis a time frame of seven years. The collection will encompass 2016 as a whole even though there is no data available from January till mid February, with a disclaimer whenever relevant data is presented. The lack of data was due to a change in reporting system that took effect mid February of 2016, where data before this is not available.

1.5 Research Questions and Planned Contributions

This thesis aims to investigate the following research questions:

RQ1 *How have the phishing trends changed in the recent years?*

RQ2 *Are there any correlation between changes observed in phishing trends and external factors, and if so, what are they?*

RQ3 *How can a universally applicable email phishing collection model be created?*

The investigation into the evolution of phishing trends and corresponding causes may be utilized to understand phishing patterns, which again can be used to understand any potential future phishing behavior. These results may assist organizations in determining what to focus on in regard to phishing, such as the evaluation of security controls, and in turn reduce the overall impact that phishing attacks currently are causing.

In inclusion of the resulting insight generated from this work, a unified model

for the collection and analysis of phishing emails will be developed. This model may effectivize the process of phishing collection and subsequent analysis, and generate universal results wherever it is applied.

1.6 Structure

The thesis is written in English. However, as most of the phishing emails are from Scandinavian entities, some figures depicting examples of these mails are in another language than English. Whenever a non-English figure is presented, any description will explain its contents.

The thesis is divided into 9 chapters and 2 appendices. Following, a list of each chapter and its content is presented.

Chapter 1 - Introduction

The introductory chapter provides an explanation of the thesis, including the background and problem description. In addition, the scope, research questions, planned contributions, and the thesis structure is presented.

Chapter 2 - Theory

The Theory chapter provides the necessary background information relevant for the understanding of the thesis. In this chapter, phishing and its approaches is defined, as well as its history and evolution.

Chapter 3 - Methodology

The chapter details the methodology that is to be used for the various stages of the thesis, including literature and information review, phishing data collection, phishing data analysis, and model development.

Chapter 4 - Related Work

The Related Work chapter presents previously conducted work that are of relevance to the thesis. This work consists of a previously conducted study, scientific papers, and published phishing related reports.

Chapter 5 - Data Collection

This chapter details how the phishing data collection is conducted. The chapter specifies the systems used for the collection of phishing data, as well as what properties are collected. It is in this chapter that the foundation of the collection model is created.

Chapter 6 - Dataset Analysis

The analysis chapter presents the results of the collection, both related to the phishing trends and observations, as well as populating and finalizing the contents for the phishing collection model.

Chapter 7 - The Email Phishing Collection Model

In this chapter, the created email phishing collection model is presented.

Chapter 8 - Discussion

The Discussion chapter encompasses discussions related to findings from the Dataset Analysis chapter.

Chapter 9 - Conclusion

Concluding notes are presented in this chapter, including sections detailing challenges and future work on the thesis' subjects.

Appendices

There are two appendices attached to this thesis. The first appendix, Appendix A, presents a full summary of the year specific phishing dataset analyses. The second appendix, Appendix B, provides an overview of all the Content categories related to the Content property from the Email Phishing Collection Model.

Chapter 2

Theory

In order to lay a foundation for the email phishing collection model and trend analysis, the subject of phishing needs to be defined. The following chapter presents an overview of the phishing phenomenon, defining the subject and providing insights into the history and evolution, and its various approaches.

As touched upon in the introduction, phishing is a form of social engineering exploiting the human factor through manipulation. Phishing can be defined as a technique utilized to manipulate an entity into performing an undesirable action, by masquerading as a legitimate source [20]. These can be actions such as providing sensitive data, downloading malicious content, or giving up monetary assets [21].

The consequences of a successful phishing attack can range from minor inconveniences, such as having to change your password or re-install your machine, to severe damages, like seen in the 2015 Ukrainian power grid attack [22]. The power grid attack was caused by a phishing email containing a malicious attachment, and gave attackers access to critical infrastructure leading to outages for close to 230 000 Ukrainian citizens.

2.1 History and Evolution of Phishing

The term *phishing* can be traced back to 1995 when the commercialization of the Internet had just begun [23]. The term emerged to describe a technique utilized to steal passwords and credit card details in a series of scams targeting AOL (America Online) customers. Malicious actors would infiltrate AOL chat rooms, targeting new customers by impersonating as official AOL entities requesting confirmation of either the customer's password or credit card details. The acquired passwords were often used to provide oneself with free internet access, while stolen credit card details were utilized to sign-up for paid services online. The AOL scams also saw the emergence of automated phishing tools, creating a more efficient approach to the phishing attack.

The popularization of the email service for private users provided malicious actors with a new platform to conduct their phishing attacks on. Although emails, or electronic messages as they were called, had existed since 1965, it wasn't until 1996 with the launch of the HoTMail that users could utilize email services without having to be tied to a specific internet service provider (ISP) [24]. Through email, one of the more infamous phishings of the early 2000's occurred, referred to as "The Love Bug" or "ILOVEYOU-virus" [25]. The Love Bug of 2000 came as an email titled "ILOVEYOU", containing a malicious file that when opened, overwrote images on the machine and sent it self to all of the victim's contacts.

The remainder of the early and mid 2000's saw more and more financially based phishing attacks targeting brands such as PayPal and eBay [26–28]. An upsurge having been tied to the increased use of eCommerce during that time frame [28]. These types of phishing attacks also saw the increased usage of fraudulent sites being linked to from phishing messages, introducing a new method of acquisition in addition to straight communication, as with the AOL scams, and email attachments, as with the Love Bug.

The rise of social media usage, starting with MySpace in the early 2000's and followed by Facebook in the later half of the decade [29], provided yet another platform for phishers to use. Social medias could not only be used to launch the phishing attacks, but as a means to gather information on their targets in order to more specifically tailor the attack towards the victim as well. As specifically targeted phishing messages have been shown to yield a higher success rate than that of messages without [30, 31], makes the rise of social media an important aspect in the evolution of phishing.

The final integral aspect in the evolution of phishing is related to the continuous digitization of the corporate domain. As more and more of the business processes are being converted onto digital mediums [32], the digital systems of organizations have become a lucrative target for criminals. Targeting an organization compared to an individual may increase the potential payout [33], such as shown with the rise of corporately targeted ransomware in the mid-2010's [34, 35]. Additionally, with the digitization also affecting governmental entities, phishing has been observed as a method utilized in international espionage [36] and sabotage [22].

What started out as small chat room scams aiming to get free internet services has evolved into large scale operations targeting individuals, organizations, and governments alike, accounting for USD millions of losses every year [37]. Although both the technology and targets have evolved since the 90's, the fundamentals of the phishing attack has remained the same: Manipulating an entity into performing an undesirable action by masquerading as a legitimate source.

2.2 Phishing Approaches

As discussed in the section above, the time frame from the first appearance of phishing up until current time has seen great changes in both the social and technical landscape. An array of new mediums and services have appeared, creating different opportunities for the conducting of phishing.

R. Alabdan presents an overview of phishing approaches, dividing them into *mediums* and *vectors* [21]. The medium concerns the means by which the malicious actor communicate the phishing attack, while the vectors concerns the avenue of attack and is dependent on the medium used. There are in total three mediums identified for usage in phishing attacks, consisting of The Internet, SMS/MMS, and voice.

Voice phishing, dubbed *vishing*, is any phishing attack carried out over telephony services, while SMS/MMS phishing, dubbed *smishing*, are phishing attacks carried out over short messaging services. Both voice and SMS/MMS provides few opportunities for the avenue of attack, while the final medium, The Internet, presents a magnitude of opportunities for the launching of a phishing attack.

The Internet, which was a large part of the evolution detailed in the section above, provides an array of communication platforms connecting users and entities to one another. Services, or in this context vectors, such as email, eFax, social networks, websites, wi-fi, and instant messaging are avenues on the Internet that can be utilized to conduct a phishing attack.

This thesis focuses on the email vector utilized over the Internet medium.

Chapter 3

Methodology

The following chapter details the methodology utilized throughout the thesis, with the inclusion of alternative approaches and any criticism towards the chosen methodology.

The completion of the thesis can be divided into six distinct processes as visualized in Figure 3.1.

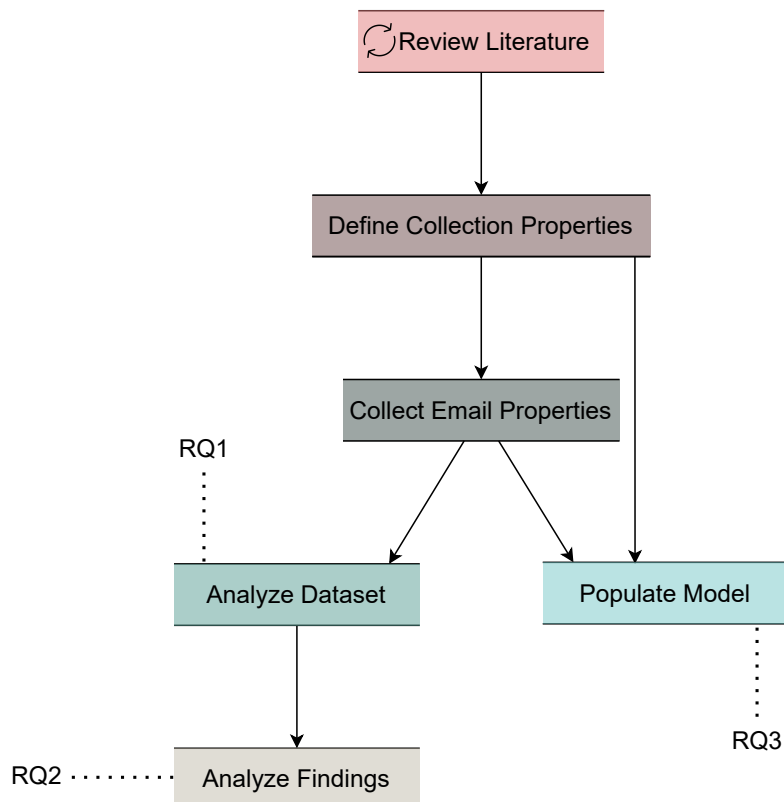


Figure 3.1: Methodology Overview

The thesis starts with reviewing relevant literature, a process which was continuous throughout the entirety of the project. Following the initial review of literature, the properties to collect from the phishing emails were defined, which were used to populate the phishing collection model. Then, the email properties were collected based on these identified collection properties, which again were used to further populate and finalize the phishing collection model. In inclusion, the collected properties established the dataset used for the following email phishing analysis. After the analysis of the dataset, its findings were further analyzed and discussed.

The process of analysing the collected email phishing dataset provided the information necessary to answer the first research question; "*How have the phishing trends changed in the recent years?*". The analysis of the subsequent findings provided information necessary to answer the second research question "*Are there any correlation between changes observed in phishing trends and external factors, and if so, what are they?*", while the process of populating the model eventually resulted in the creation of the Email Phishing Collection Model, providing answers to the final research question "*How can a universally applicable email phishing collection model be created?*"

3.1 Literature Review

The main activity of the literature review process is to gather and analyze papers, reports, and information related to the topics of this thesis. This was done at the beginning of the project in order to identify work that closely relates to the subject matter to gain an understanding of their focus areas, methodologies, findings, and proposed future work. The remaining of the literature review process was continuous throughout, populating the thesis with necessary and relevant information.

Throughout the collection of related literature an emphasis on the credibility of the papers and information was made. Because of this, the scientific search engines of Google Scholar¹ and NTNU's Oria² were utilized when searching for and reviewing relevant work. As these engines includes materials published by an array of different journals, the reviewed papers' sources were also checked against the Norwegian Register for Scientific Journals, Series and Publishers (Kanalregisteret)³ to determine their reputation and degree of trustworthiness.

As some of the information needed to appropriately conduct this thesis was not present in scientific papers, it was necessary to use non-scientific search engines as well. Periodical reports and information related to the analysis of trends were

¹Google Scholar: <https://scholar.google.com/>

²Oria: <https://ntnu.oria.no>

³Kanalregisteret: <https://kanalregisteret.hkdir.no/>

subjects where the non-scientific search engine of Google⁴ was utilized. As information provided by this search engine not necessarily has a basis on scientific research, source criticism was highly important. Information referenced was crossed checked with multiple sources in order to determine the reliability of any statements.

3.1.1 Alternative Methodology

An alternative approach to the literature review would have been to solely use information from published work. This would have increased the credibility of any referenced information, but would not have been feasible for this thesis as a significant portion of it required information not available in scientific publications.

3.1.2 Criticism of Chosen Methodology

Utilizing and referencing non-scientific sources reduces the credibility of any referenced work, and if not appropriately validated, reducing the credibility of the thesis itself.

3.2 Defining Collection Properties

The defining of collection properties concerns the identification of the elements to collect from the phishing emails. Figure 3.2 visualizes this phase, where the arrows and accompanying text symbolizes an action, and the boxes symbolizing an existing or resulting product.

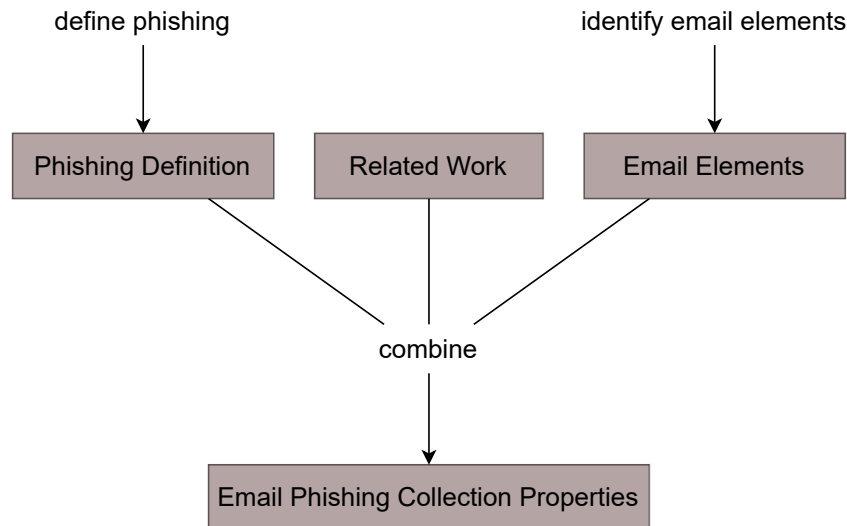


Figure 3.2: Methodology - Defining Collection Properties

⁴Google: <https://www.google.com/>

In order to identify what properties to collect from a phishing email, the definition of phishing itself, as well as the identification of available elements in an email was conducted. This, in combination with already existing external papers and priorly conducted work produced the properties and corresponding email elements to be used in the subsequent collection process.

3.2.1 Alternative Methodology

An alternative approach would be to use properties already defined in external research. Using this methodology could have reduced the amount of time used for this process, as well as providing an approach externally tested and argued for.

Due to prior work already having been conducted on the subject by the author, as well as a lack of published studies presenting adequate collection properties, this approach was not deemed appropriate for this thesis.

3.2.2 Criticism of Chosen Methodology

Defining collection properties oneself instead of using externally produced properties presents a time consuming process, whose allocated time could have been utilized to improve any of the other processes defined for the thesis. In addition, using an externally produced approach provides the collection with properties that have been tested and approved by reputable sources.

3.3 Collection

The collection process utilizes the collection properties defined in the formerly detailed process in order to generate a phishing dataset. Figure 3.3 visualizes the stages within this process.

From the pool of emails for all the years within the collection scope, a filtering on emails falling under the category of phishing was conducted, producing the email phishing corpus. The specifics on the search queries used for this filtering is defined in Chapter 5. The phishing corpus is filtered based on the years defined for collection, before properties are collected from them. These steps are repeated for each of the years within the collection scope, generating the phishing dataset.

The collection process utilized a quantitative methodology. A quantitative methodology is a methodology used for the collection of quantitative data. Quantitative data is data that comes in the form of numbers or other terms of units, often including large volumes of entities [38]. Due to the substantial volume of phishing emails within the collection years, as well as the desired output from the collection, which was data that could be used in a trend and evolutionary analysis, a quantitative approach was deemed preferable for this process.

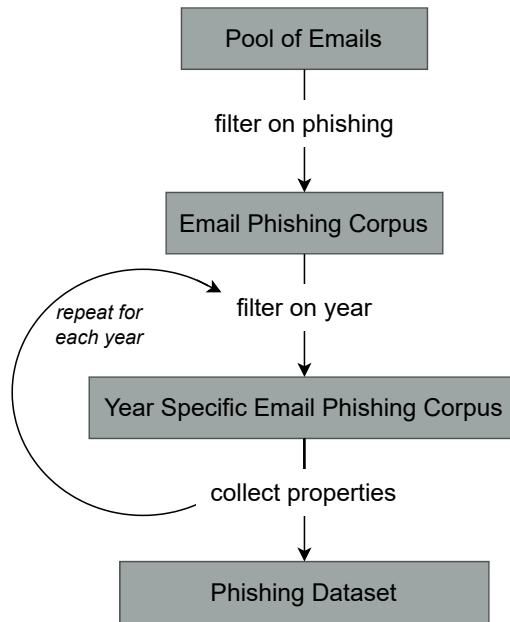


Figure 3.3: Methodology - Collection

3.3.1 Alternative Methodology

An alternative approach would be to base the collection on a qualitative methodology. A qualitative methodology focuses on the collection of qualitative data, which concerns textual/visual interpretations of a few selected entities [39]. Using a qualitative approach would allow for heightened focus on the phishing emails, allowing for a more in-depth dataset resulting from the collection.

As one of the main parts of this thesis is to analyze trends, it is preferable to have the collected data being presented in a numerical way so that statistics easily can be extracted from the dataset for analysis. That, in addition to the overall size of the phishing corpus, encompassing over 35 000 emails, rendered this approach impractical for this thesis.

3.3.2 Criticism of Chosen Methodology

Choosing a quantitative approach limits the depth of information that can be gathered from the phishing corpus. Information such as the layout and visual cues of the phishing email could not be collected, and in turn being subjected to trend and evolutionary analysis.

3.4 Model Population

The Model population process, although marked as its own process, is largely based on the outputs generated from the "Define Collection Properties" and "Collect Email Properties" processes. The properties defined in the former laid the foundation and structure of the collection model, while the collection of said properties provided concrete categorizations within these properties based on real phishing emails. In other words, the former process defined the properties themselves and what email elements were relevant to collect, while the latter defined how to categorize what had been collected. Figure 3.4 visualizes this division. A more detailed explanation of the model terminology is presented in Chapters 5 and 7.

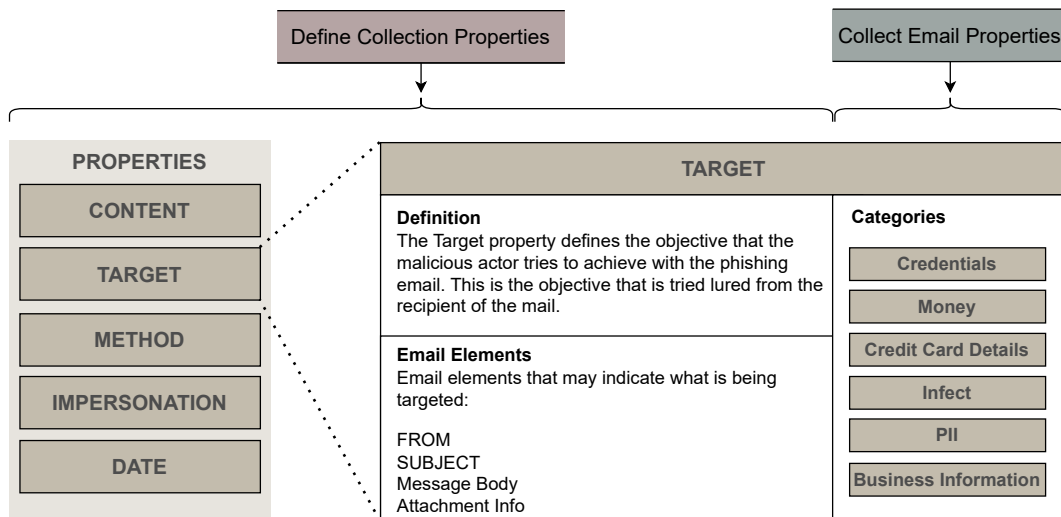


Figure 3.4: Methodology - Model

The model population process itself mainly concerned combining the relevant outputs from the definition and collection processes in order to create and finalize the Email Phishing Collection Model. As this process heavily depends on the two aforementioned processes, the alternative approaches and criticism defined for those is applicable for this process as well.

3.5 Analysis

The final processes conducted in this thesis concerned the analysis of the dataset generated from the collection process. The analysis processes were divided into two parts, where the first process consisted of the analysis of the dataset itself, while the second part provided a more in-depth analysis on findings from the dataset analysis. Figure 3.5 displays an overview of the dataset analysis process.

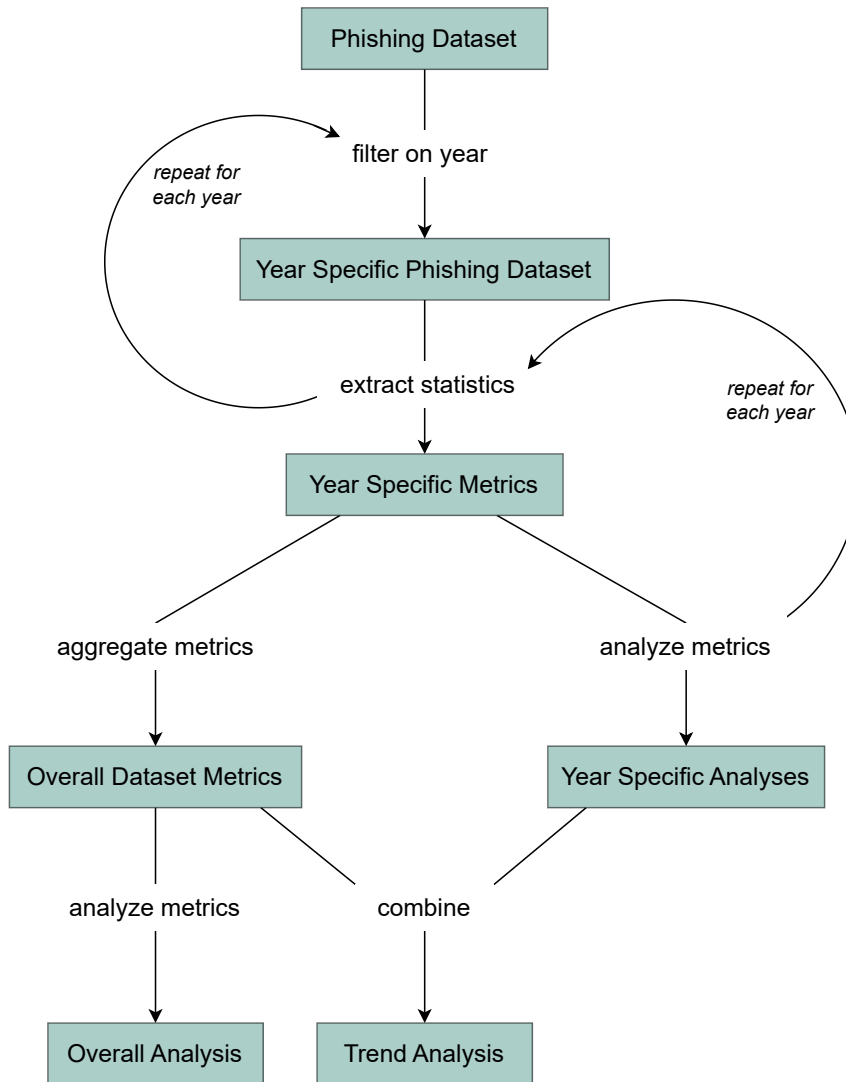


Figure 3.5: Methodology - Dataset Analysis

Statistics from the phishing dataset is extracted for each of the collection years, providing phishing metrics specific to the given years. The metrics for each year were analyzed separately resulting in year specific analyses. On the other side, all the year specific metrics were aggregated into an overall dataset of metrics. This aggregation was done by converting all numerical values into percentwise distributions for each year, and then putting them into the same dataset. This overall dataset was analyzed, creating the overall analysis. Lastly, both the results from the year specific analyses and the metrics from the overall dataset were viewed in unison in order to provide an analysis on the phishing trends throughout the years.

When the dataset analysis was done, specific findings from said analysis were viewed in more detail in order to explain the observations, including whether other sources had seen similar behavior and what could be the reasonings for them. As this process heavily depends on information from external material, the methodology detailed for the literature review process applied for this process as well.

3.5.1 Alternative Methodology

For the dataset analysis, instead of analysing each of the years separately and then aggregating the metrics in a percentwise fashion, the dataset could have been analysed based on the raw numbers for each of the years. This would have eliminated the need for combining two outputs to create the trend analysis, as well as prevented the process of converting each years' metrics into percents before putting them into a unified dataset again. As the percent-converting was a time consuming task, avoiding it would have provided additional time to improve the outputs of the other processes.

The reason why this approach was not applied was due to the significant differences in the sizes of the phishing corpuses for each of the collection years (explained further in Chapter 6). Drawing statistics from a dataset consisting of a year with over 16 000 emails and a year with 500 emails would skew the results to mostly reflect the former. Because of this, in order to present statistics where each of the years within the collection scope were equally weighted, the raw numerical data had to be converted into percentwise distributions.

As for the findings analysis, another approach would have been to dwell deeper into all the findings of the dataset analysis. This would have provided a more complete analysis, leaving fewer questions unanswered, however was deemed too time consuming, as well as would have taken the focus away from the other important aspects of the thesis such as the Email Phishing Collection Model.

3.5.2 Criticism of Chosen Methodology

By aggregating the metrics for each of the collection years, having them weighing equally, prevents the overall analysis to emphasis a year with more observed phishing emails. It also limits the analysis' ability to convey the magnitude of phishing emails for certain time periods outside of an entire year. This is combated by providing a separate analysis for each of the years within the collection scope, however adds more layers of work, being both time consuming and complicates the overall presentation of the thesis.

And lastly, for the findings analysis, not providing a more in-depth analysis of the findings leaves the possibility of questions arising from the dataset analysis being left unanswered.

Chapter 4

Related Work

This chapter presents various papers and reports that are of relevance for this thesis. The material evaluated depicts work related to both the thesis' model and trend analysis.

4.1 Prior Work

This thesis will be a vast expansion of the work conducted by the author in the paper "An Extensive Analysis of Email Phishing" [40]. The paper presents an analysis of phishing emails that were observed in Q1 of 2022. The analysis was three-fold with the first part focusing on the phishing trends of the collected mails, the second part comparing prior literature depicting phishing email features to features of the observed phishing emails, and lastly, a part analysing the features of phishing emails that people had fallen victim to. It is the first part of this paper that is relevant for this thesis.

The first part of the aforementioned paper, focusing on the observed phishing trends, outlined a methodology for defining what properties of the phishing emails should be extracted for analysis in order to showcase the phishing trends. These properties included elements such as the essence of the phishing email, methods used within the emails, goal of the phishing, and who/what the email appeared to come from. These properties will be utilized when developing the thesis' model, albeit after being subjected to an evaluation and re-definition based on existing work.

The findings from [40] showed that phishing emails such as gift card CEO scams, password expiry notices, and sharing of documents were the most prominent in the time frame of the dataset. The former consisting of emails where the perpetrator tries to lure the recipients into buying gift cards while pretending to be the manager/CEO of the recipient. Further, the paper depicted that passwords were the most sought after goal from the malicious actor's point of view, and that links within the emails was the most used method of achievement. Finally, it was

deemed that phishers often pretend to be from an internal source within the organization when launching their attacks.

The Future Work section of the paper brought forth "repeated analysis" and "automated collection tool" as areas that should be focused on in any future research into the field. This thesis will address these points by expanding the time frame of the collection scope, as well as utilizing new query tools in order to further effectivize the collection of the phishing emails.

The remaining sections provides papers, works, and reports of external sources that are of relevance for this thesis.

4.2 Research Papers

There exist an array of studies collecting and analysing phishing corpuses, however many of the studies covering phishing emails are shown to focus on a smaller time range for the collected emails [17, 18, 41–46], often in the range of approximately a year, as opposed to a wider time range which this thesis does. Another observation that can be made from these research papers is that many tend to focus on phishing URLs within the email, and not other methods such as attachments or plain communication [41–44].

The paper of Ferreira et al. [47] is one of few identified relevant papers that utilize a wide time range for the collection of phishing emails. The paper focuses on the analysis of phishing email subjects in order to determine trends within the observed subject lines in relation to human persuasion. The study utilizes a phishing corpus stemming ten years from 2008 to 2017, however, only approximately 20 emails per year was extracted and used for the analysis. The Future Work section of the paper states that the analysis should be attempted on a bigger sample of emails. The focus point of the paper is relevant for the analysis of the "essence" part of the emails for this thesis, and although the methodology will differ from Ferreira et al.'s, it will extend on the findings and provide a related analysis of a bigger sample.

One Master thesis has been identified focusing on a similar area related to phishing trends. M. R. Riedle's thesis "Identifying Trends Among Phishing Attacks" [16] details a study of phishing attack trends from a 10 year perspective, up until 2016. The study utilized both private and public sources for their phishing corpus, and deployed a frequency analysis to categorize the attack trends. The results showed that the amount of phishing attacks have increased over the ten year period in which the study encompassed, and that surges could be seen around the holiday periods and tax seasons of the years. In inclusion of this, the results revealed that the typical Advanced-fee scam (such as Nigerian Prince scams) were declining, and that the usage of HTML attachments in phishing emails had grown signific-

antly. The thesis singled out three attack approaches which it based its analysis on, excluding any other methods that could have been in play within the given time frame. The present thesis will continue on analysing emails starting from 2016 to identify phishing trends, however, not singling the analysis in on particular attacks.

The paper of Cui et al. titled "Tracking Phishing Attacks Over Time" [48], is another paper whose findings will be utilized to achieve the goals of this collection and analysis. The paper introduces a phishing site detection scheme that bases itself on the DOM (Document Object Model) details of the site, and tests the assumption that many phishing attacks are replicas or variations of each other. From their corpus of over 19 000 emails, collected over 10 months, it was shown that 90% of the observed phishing sites were replicas or variations of other observed phishing sites. Although this thesis aims to collect and analyze the phishing emails themselves and not the landing site, the findings of Cui et al. is still of relevance in the collection phase of this thesis in order to increase the efficiency of the collection process.

4.3 Reports

There are a great variety of companies and groups who publish phishing related reports, either focusing only on phishing or as a sub category of a wider range of threats. Because of this, there exists an array of reports where phishing is detailed, and in order to scope down the magnitude of reports, the ones focusing in on email threats or email phishing is of most relevance surveying.

When viewing related reports, certain aspects relevant for this thesis is focused on. As the main part of the thesis is displaying email phishing trends, how statistics are shown and compared to prior metrics is an important note. Another aspect of importance is types of metrics collected, viewing whether they are overarching metrics or specific to one certain area within phishing. Following, a list of identified relevant reports are presented in relation to the specifics stated above.

Cofense - Annual State of Phishing Report (2022)

Cofense's Annual State of Phishing Report [49] displays trends and statistics related to phishing as observed the prior year. The report details specific types of email phishing attacks with examples and explanations. In addition, each shown phishing example is categorized based on threat type and tactic. Although the report showcases statistics from their observations, they are for the most part excerpts, meaning that they do not display the full dataset. Instances such as stating that HTML attachments account for 30% of all credential phishing, but not revealing what the remainder 70% is are examples of this. The report tends to go more in-depth on certain aspects rather than viewing the overarching metrics of their dataset.

Cofense - Annual State of Email Security Report (2023)

Cofense publishes various reports, and another report detailing phishing is their Annual State of Email Security Report [15]. The report displays much of the same statistics as seen in their State of Phishing Report, however provides additional metrics expanding on their phishing observations. The statistics shown are in the case of this report more complete than their aforementioned report, such as showing statistics for all observed attachments and not just the top ones. It does still focus on specific aspects rather than overarching metrics.

APWG - Phishing Activity Trends Report (3Q 2022)

The Anti-Phishing Work Group (APWG) publishes quarterly phishing reports detailing phishing attacks observed within the quarter's time frame, an example being their Q3 of 2022 report [13]. The report is statistic heavy, as opposed to the two aforementioned reports, showcasing both trends and complete statistics. The trends showcased span a year on the maximum, and is often compared to the former quarter's observations. Although it showcases a fair amount of complete statistics, due to the relatively short length of the report, there is a limited amount of data presented.

The APWG report does not display any examples of the observed phishing mails. It does however explain certain types of email phishing attacks observed.

Zscaler - ThreatLabz Phishing Report (2022)

Zscaler's report [14] details observations from the previous year, depicting trends and metrics from specific focus areas. The report defines methods utilized by the malicious actors in their phishing mails, including Link, Prompt, and Attachments, however does not show any statistics related to these. Similarly, the report details various types of phishing, but with no associated statistics.

The trends displayed in the report are for the most part comparisons from last year's observations with no broader analysis.

ProofPoint - State of the Phish (2023)

ProofPoint's report [50] differs from the other selected phishing reports as it focuses on the recipients of the phishing message rather than the phishing messages themselves. These are statistic related to areas such as recipients' knowledge of terms and concepts, risky actions performed, and deployment of security awareness programs. It does however showcase some statistics related to the phishing messages themselves, such as targeted brands and types of attacks.

SlashNext - The State of Phishing (2022)

SlashNext's report of phishing [51] details the company's observations from the prior year, focusing on specific areas within the phishing domain. This report is the least email forward, as it details other phishing types such as mobile. The parts detailing email phishing display little to no statistics, and does not detail any trends from other time ranges.

Abnormal - Email Threat Report (H2 2022)

Abnormal's Email Threat Report [52] focuses on the areas of credential phishing and Business Email Compromise (BEC). The statistics shown in the report is mostly related to attack volume and brand impersonation, displaying both trends and evolution.

Consistent throughout all the identified reports, they all tend to focus on specific and varying areas within phishing, rather than displaying metrics based on collection properties from the phishing emails themselves. This is somewhat understandable as most of the reports are structured as summaries rather than from the complete dataset analysis. This does however limit the usability of the reported statistics, as well as the replicability for other datasets.

There are also instances where the terminology differs from report to report. Such as Zscaler stating that Business Email Compromise and CEO fraud is one in the same, and that they can only be achieved through a compromised account, while other reports such as Cofense's and APWG's displaying CEO Fraud as a subcategory of BEC and stating that BECs can be achieved both through compromised and fake lookalike accounts.

From the collection of phishing reports, Cofense's State of Email Security Report and Zscaler's Phishing Report are the most comprehensive in relation to the subjects of this thesis, while Cofense's State of Phishing Report, APWG's Phishing Activity Trends report, and Abnormal's Email Threat Report all contain relevant data, albeit in a bit reduced fashion compared to the two former. The reports of ProofPoint and SlashNext were the least relevant as they put a wider focus on subjects not relevant for this thesis.

Regardless of their comprehensiveness, this thesis will leverage the observations from all of the reports both in order to create a collection model that can be used universally, as well as during the analysis of the findings from the dataset.

Chapter 5

Data Collection

5.1 Collection Sources

The email phishing data is collected using historical email reporting data from a Norwegian based IT platform provider. The collection of the phishing emails utilize two separate data sources, consisting of a ticketing system and an email reporting application called MailRisk¹. The reasoning for the utilization of two separate data sources is due to a change in the email reporting procedures of the IT platform provider, where in the last months of 2020, a switch from the ticketing system to the MailRisk application was made.

5.1.1 Ticketing System

The ticketing system allowed for users to report in suspicious emails they had received in their mailbox. The reporting procedure consisted of forwarding the suspicious email to the ticketing system's address, where it could be analyzed by relevant personnel. After analysis, all the reported emails would be classified in the categories of phishing, scam, or harmless, and the analysis itself would be appended to the ticket. The ticketing system itself allows for filtering on the specific categories, including free text searches on the contents of the email.

5.1.2 MailRisk

MailRisk is an email reporting application and system provided by Secure Practice². The application is made available to the user as an add-in within the email application and provides an option for the user to flag any mails that they consider suspicious. When an email is flagged, all email data, including metadata, is sent to the MailRisk system for analysis. After an analysis of a mail, it is placed within one of the eight categories Safe, Spam, Suspicious, Scam, Phishing, Harmful, Virus, and Targeted. The MailRisk application provides a query function, called Threat

¹MailRisk: <https://securepractice.co/guides/mailrisk-intro>

²Secure Practice: <https://securepractice.co/>

Explorer for email lookups, where one can filter on specific details such as category and email metadata. The Threat Explorer allows for a more granular search than with the ticketing system, as it allows for AND, OR, and NOT clauses.

For the data collection, the ticketing system is utilized in order to collect phishing data in the time period of 2016 to August 2020. For August 2020 throughout 2022, MailRisk is utilized to collect the remainder of the phishing data. The categories Phishing and Scam is filtered upon for the data collection within the ticketing system, while the categories Scam, Phishing, Harmful, Virus, and Targeted, are used to query and collect relevant data from the MailRisk application.

For both systems, the insight from Cui et al. [48] is used in order to effecivize the collection. The search function of the ticketing system and the Threat Explorer in MailRisk is utilized in order to collect batches of the same phishing emails based on keywords found within a given phishing email.

5.2 Model Properties

In order to define what properties of a phishing email are relevant for collection, the anatomy of a phishing email needs to be defined. The anatomy in this situation refers to how an email is structured and what properties are relevant in the context of phishing.

5.2.1 Email Structure

An email sent utilizing the Simple Mail Transfer Protocol (SMTP) [53] consists of two parts; An Envelope and its content, while the content again can be divided into the two sections Header and Body. Figure 5.1 visualizes the SMTP format, including relevant properties.

The Envelope contains information relevant for the transmission of the email. This includes information such as the sender and the recipient addresses. Email servers utilize this information in order to relay the email to the appropriate parties. Further information on the SMTP is detailed in RFC5321 [53].

The Header contains relevant information about the email, including where and whom the email is from, who the email is sent to, the subject of the mail, timestamps, authentication information, and more. Further information on the message header contents and corresponding RFCs can be found on IANA's sites [54].

The Body of the email is the actual content of the mail. The body can be plain text or special formatted content such as with HTML.

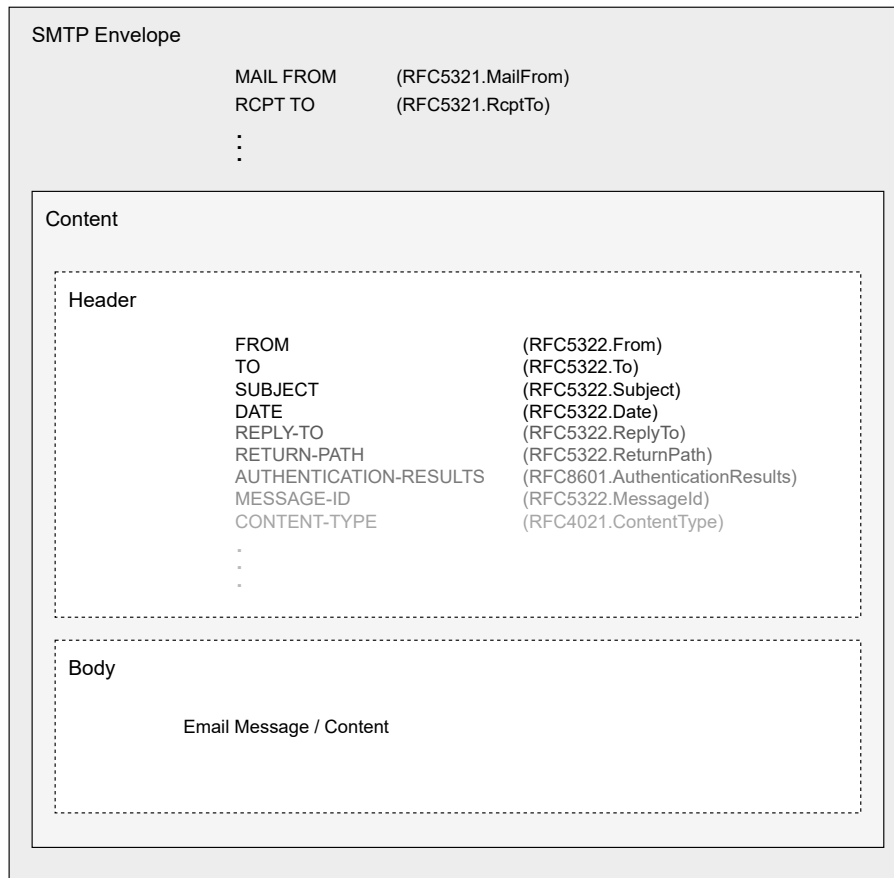


Figure 5.1: SMTP Email Format

In addition to the contents of an email detailed above, an attachment can be appended to the email as well. The attachment, although received with the email, is not a part of the email build-up itself, however remains a crucial component in the context of email phishing.

5.2.2 Phishing Properties

To select the properties relevant for the collection model, they have to be seen in relation with the definition of phishing. As detailed before, phishing is a form of social engineering where the main goal is to trick the victim into performing an undesirable action by masquerading as a legitimate source. Based on this definition and in correspondence with prior literature and reports, several properties can be defined as relevant for collection.

From the definition of phishing, its purpose is to make the recipient perform an undesirable action in order to achieve their (the malicious actor's) goal. Both the aforementioned paper from Ferreira et al. [47] and Dhinakaran et al. [45] classify

various goals a malicious actor can target through a phishing email such as data theft, malware, fraud, and passwords. Other periodical reports from commercial businesses such as ProofPoint's 2023 "State of the Phish" [50] and Zscaler's 2022 "ThreatLabz Phishing Report" [14], also define similar objectives targeted in observed phishing mails.

In order to collect statistics on the desired objective of the malicious actor, the property **Target** is defined for collection. Based on the aforementioned sources, the Target can be collected through the collection and analysis of the email's subject, from name, from address, contents of the email body, including any external linked sites or attachments.

Continuing on with the performance of an undesirable action, the performance itself must be defined. Irani et al. in their paper concerning an evolutionary study of phishing [46], defines a collection property called the "sting". The sting is a part of the content that direct the user into performing the undesirable action. Their paper mainly focuses on URLs as the sting, however as seen in the previously conducted study in [40], this can be performed through other means such as attachments and communication as well. From Cofense's reports [15, 49], the term "Tactic" is utilized to describe this property, while Zscaler utilize "method" to describe the same [14].

The property **Method** is defined for this, and concerns the method of achievement utilized by the malicious actor to obtain their target. Relevant email elements to collect and analyze for this model property is the contents of the email body and any attachments attached to the email.

The definition of phishing puts an emphasis on masquerading as a legitimate source. Present throughout several of the papers and reports presented in the related work chapter, Chapter 4, is the analyzing and/or collection of the brands and persons that the phishers utilize to appear legitimate [13–15, 42, 45, 47, 50–52]. From these papers and reports, multiple email elements can be seen providing information as to whom the phishing email is pretending to be from. Relevant elements include sender name and address (both MAIL FROM and FROM), subject, contents of the email body including any logos attached and URLs, and attachment names.

Reports from Zscaler [14], ProofPoint [50], Slashnext [51], and Abnormal [52] use the term "impersonation" when referring to this aspect of a phishing email. Borrowing from these reports, **Impersonation** is defined as the common name for this model property concerning what or whom the email is pretending to be from.

As a final property, the concept of tricking the recipient into performing the un-

desirable action needs to be covered. In order to trick the recipient into performing the undesirable action, the mail itself should convey an overall perception of its legitimacy. The Impersonation property is a contributing factor in the convincing of the recipient, however a property conveying the essence of the mail should be included. Irani et al. utilizes the term "content" to define what the email is conveying, and bases the information on the message body content [46]. As shown in the previously conducted study [40], the essence of an email can also be based on additional elements such as sender address and name, subject, and any attachments appended to the mail.

The term **Content** will be borrowed from Irani et al. and utilized to define the final collection property. However, in inclusion of the message body, the elements of sender address and name, subject, and attachment info should be included in determining the content as well.

Lastly, in order to analyze any trends from the phishing email corpus, the dates in which the emails were received needs to be collected. The dates should be when the email was received by the recipient and not when the email was reported or collected. This means that the DATE header field will be collected for this property. The **Date** property is the final property defined for this model.

5.2.3 Model Properties Overview

Based on the anatomy and properties defined in the section above, Table 5.1 details the model properties that should be collected in an analysis of phishing trends, including their corresponding email elements.

Property	Description	Email Elements
Content	The Content property concerns the essence of the phishing email. It revolves around how the malicious actor convinces the recipient of the email's legitimacy.	FROM SUBJECT Message Body Attachment Info
Target	The Target property defines the objective that the malicious actor tries to achieve with the phishing email. This is the objective that is tried lured from the recipient of the mail.	FROM SUBJECT Message Body Attachment Info
Method	The Method property concerns the approach utilized within the phishing email by the malicious actor in order to obtain the desired Target.	Message Body Attachment Info

Table continued from prior page.

Property	Description	Email Elements
Impersonation	The Impersonation property concerns who the email appears to be from. This is whom or what the malicious actor pretends to be.	MAIL FROM FROM SUBJECT Message Body Attachment Info
Date	The date in which the phishing email was received.	DATE

Table 5.1: Model Properties

The model properties detailed above will be utilized when performing the data collection. The collection will provide supplementary details for the model properties, and assist in finalising the thesis' email phishing collection model.

Chapter 6

Dataset Analysis

The collection of the phishing dataset based on the model properties specified in Chapter 5 yielded a result of a total of 35566 phishing emails. The dataset in its entirety can be found in the GitLab repository tied to this thesis¹. The following chapter presents the findings and highlights any trends that can be seen within the dataset. This chapter focuses on an aggregated view of the dataset, as well as highlighting specific findings from certain years. A full analysis of the data detailing each year separately can be found in Appendix A.

6.1 Collection Properties

For each of the five collection properties, there were identified several distinct categories throughout the data collection. Following, a presentation and description of these property categories is given.

6.1.1 Content

The collection of the Content property, defining the essence of the phishing email, resulted in the identification of 68 unique Content categories. The five most prominent categories are presented below, while a full overview and description of all the Content categories can be found in Appendix B

Invoice

The Invoice Content category embodies all the phishing emails that use the lure of a supposed invoice in order to trick the victim into performing an undesirable action. Figure 6.1 showcase an example of such phishing emails allocated within the Invoice Content category.

¹GitLab Repository: <https://gitlab.com/Karset/mis4900-phishing-dataset>

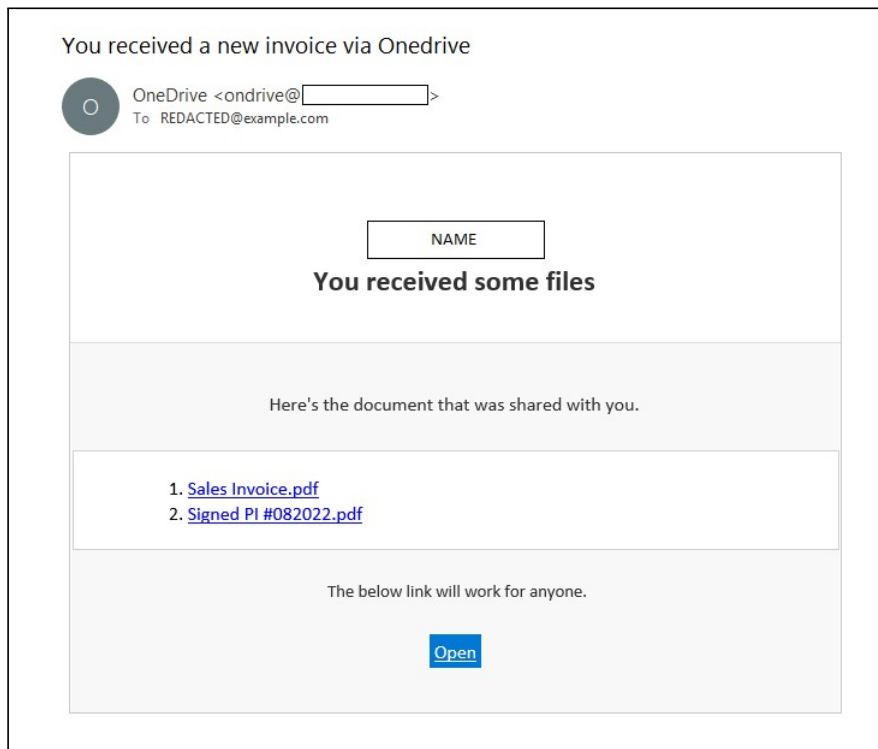


Figure 6.1: Invoice Phishing Email

Document Shared

The Document Shared Content category is a collective term for all the phishing emails that conveys the message that a document has been shared with the recipient. This can for example be by simply attaching a document to the mail itself, or sharing a document through an online service, as shown in Figure 6.2.

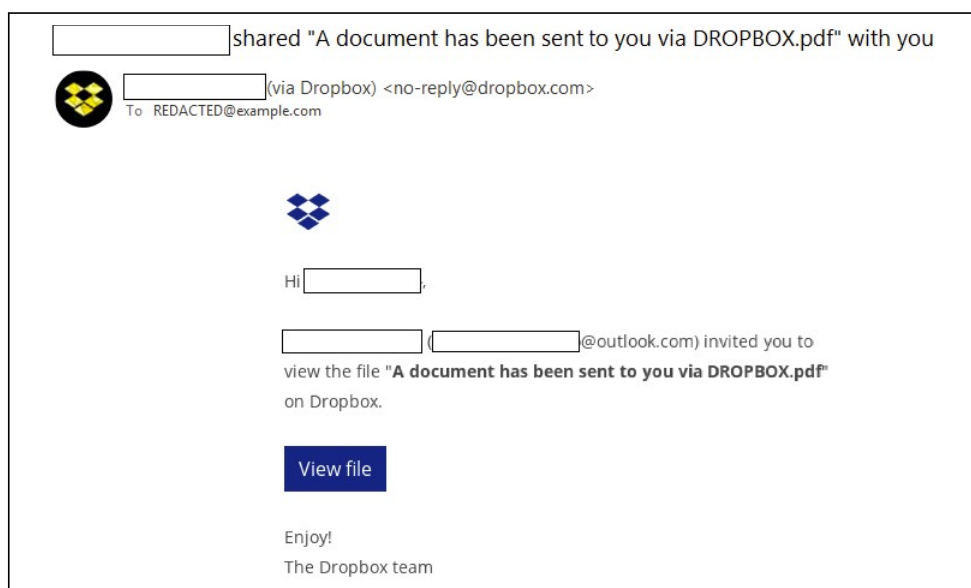


Figure 6.2: Document Shared Phishing Email

CEO Scam - Transfer Money

The CEO Scam, or CEO Fraud as it is also called, is a generally known phishing technique where the mail appears to be from a higher ranking employee, such as a manager or the CEO, asking for a favor. This specific technique is recognized and detailed by other mediums such as [55–57] as well.

This specific Content category encompass all the phishing emails that utilize this CEO Scam technique in the context of requesting the recipient to transfer a sum of money, for instance due to a business deal or an unpaid invoice. Figure 6.3 displays an example of this type of phishing mail, where the malicious actor masquerades as the CEO, asking the recipient if they can process a payment to an England based company.



Figure 6.3: CEO Scam - Transfer Money Phishing Email

Post Package

The Post Package Content category embodies all the phishing emails that uses the incentive of a package yet to be delivered in order to lure the recipient into performing an undesirable action. For example, requesting that a delivery fee is paid for the package to be delivered, as seen in Figure 6.4.



Figure 6.4: Post Package Phishing Email

CEO Scam - Gift Card

The CEO Scam - Gift Card Content category is another Content category that can be placed under the CEO Scam/Fraud umbrella. In this category, the malicious actor utilizes the CEO Scam technique in order to lure the recipient into purchasing gift cards for them. This could for instance be in relation to a customer, or as a gift for the employees. Figure 6.5 showcases an example of such CEO Scam - Gift Card phishing, where the malicious actor masquerades as the CEO, asking the recipient to purchase gift cards.



Figure 6.5: CEO Scam - Gift Card Phishing Email

6.1.2 Target

The Target property totaled six unique Targets from 2016 throughout 2022. These Targets were: Credentials, Money, Credit Card Details, Infect, Personal Identifiable Information, and Business Information.

An important note related to this property is that it is based on the first objective achieved by a successful phishing. A phishing email may link to a malicious file that infects the computer with a keylogger, which again retrieves the credentials of the victim. In this case the infection is the first Target, and the phishing mail's Target would be categorized as such.

Credentials

Credentials in this context refers to any details used to authenticate one self on digital mediums. Credentials encompass technologies such as passwords, one-time passcodes, PINs, and digital keys.

Money

Any phishing email soliciting direct payment or the transfer of money falls within the Money Target. In this context, it is digital payments and transfers that are of relevance.

Credit Card Details

Credit Card Details and Money may appear similar as the overarching goal is financial gains. However, there is a clear distinction between the two. With the

Money Target category, a payment is requested by the malicious actors to a specific account. The malicious actors in this scenario would have no insight into the payment process or details, only a transferred sum of money should the phishing be successful. With Credit Card Details however, the malicious actor requests the recipient to provide the credit card details necessary for the malicious actor to use them as they please.

Infect

The Infect Target category encompass all phishing emails where the recipient is attempted lured into downloading and opening a malicious file.

Personal Identifiable Information

Personal Identifiable Information (PII), or Personal Data, refers to information that can identify an individual, such as names, phone numbers, and addresses [58]. Any phishing emails requesting this information will be categorized under the PII Target category.

Business Information

Lastly, Business Information was identified as the sixth and final Target category within the collected phishing dataset. Business Information embodies non-public information contained within the organization. This can be information such as account balance, invoices, and customer lists.

There were instances where the available data could not provide the necessary information in order to determine the exact Target. In these cases, the Target is marked as N/A.

6.1.3 Method

The collection of the Method property revealed four distinct methods of achievement within the phishing corpus. The Methods of URL, Communication, Attachment, and Calendar Invite were identified, where the Attachment category could again be divided into subcategories based on the attachment type.

Similar to the Target property, the Method property details the first method of achievement. In some scenarios, the malicious actor may for example send out a phishing email with a malicious link that leads to a site where a malicious file is downloaded. In this case, the Method would be defined as URL.

URL

The URL Method encompass all phishing emails where the malicious actor tries to lure the recipient into clicking a malicious link, leading to their phishing web site.

Communication

The Communication Method embodies the phishing emails where the Target is achieved over communication, either within the email thread or on an external medium such as over the phone.

Attachment

The Attachment Method includes the phishing emails that utilize an appended file on the email in order to achieve the phishing Target. In total, 19 distinct attachment types were observed. The types themselves will be showcased later in the chapter during the presentation of the observed trends.

Calendar Invite

The Calendar Invite is a type of Method where the malicious actor abuses many of the email providers' technology of automatically adding a calendar invite into the calendar of the recipient. Even though the event is not accepted, it is still present in the calendar and may send notifications to the user when the event is approaching. The calendar event itself may contain malicious information, attachments, or links to malicious sites.

There were instances where the Method could not properly be identified. In such cases, the Method was categorised as N/A.

6.1.4 Impersonation

The Impersonation category can be divided into two sections based on the observations from the collected phishing dataset, one dealing with generic impersonations, and one dealing with non-generic impersonations. The generic Impersonation categories consists of phishing emails where the supposed sender is not tied to any specific brand. This category can again be divided into two parts: External and Internal. The External Impersonation category consists of the phishing emails coming from an external sender, but not from a specific brand or entity. The Internal Impersonation category consists of the phishing emails where the sender is supposedly from an entity within the organization, such as your manager, the IT department, or HR, or from yourself.

The non-generic Impersonation category encompass all phishing emails that appear to come from a specific brand or entity.

6.2 Aggregated Analysis

The properties of a total of 35566 phishing emails were collected from the years 2016 throughout 2022. Table 6.1 showcases the distribution of phishing emails for each year of the collection scope.

Year	Total
2016	1173
2017	986
2018	523
2019	720
2020	3181
2021	16202
2022	12781

Table 6.1: Phishing Email Distribution

The table shows a sharp increase in the number of phishing emails from the years 2020 and up. As explained in Chapter 5, this is due to a change in the email reporting system in August of 2020, where it was rendered easier to report suspicious emails than with the prior system. Due to the varied number for each year, many of the following sections will utilize percentages per year to display and analyze the collected data.

The year of 2020 presents a challenge in the representation of that year's phishing data, as the new reporting system was introduced in August of that year. This drastically increased the number of phishing emails collected for the last five months of the year compared to the former seven months. A more in-depth view of this challenge is presented in the Challenges section in Chapter 9.2. Due to this discrepancy in 2020's phishing dataset, some data will be presented in two separate parts, one for the first seven months of 2020, and one for the remaining five months. When looking at the total for all years, 2020 will be represented as a whole.

6.2.1 Content

In total throughout the years within the collection scope, 68 unique Content categories were identified. Figure 6.6 displays the total occurrence rate of each of these 68 Content categories based on the percentwise distribution for each year. And, as can be seen, the Invoice category is the most observed Content category, followed by Document Shared and CEO Scam - Transfer Money.

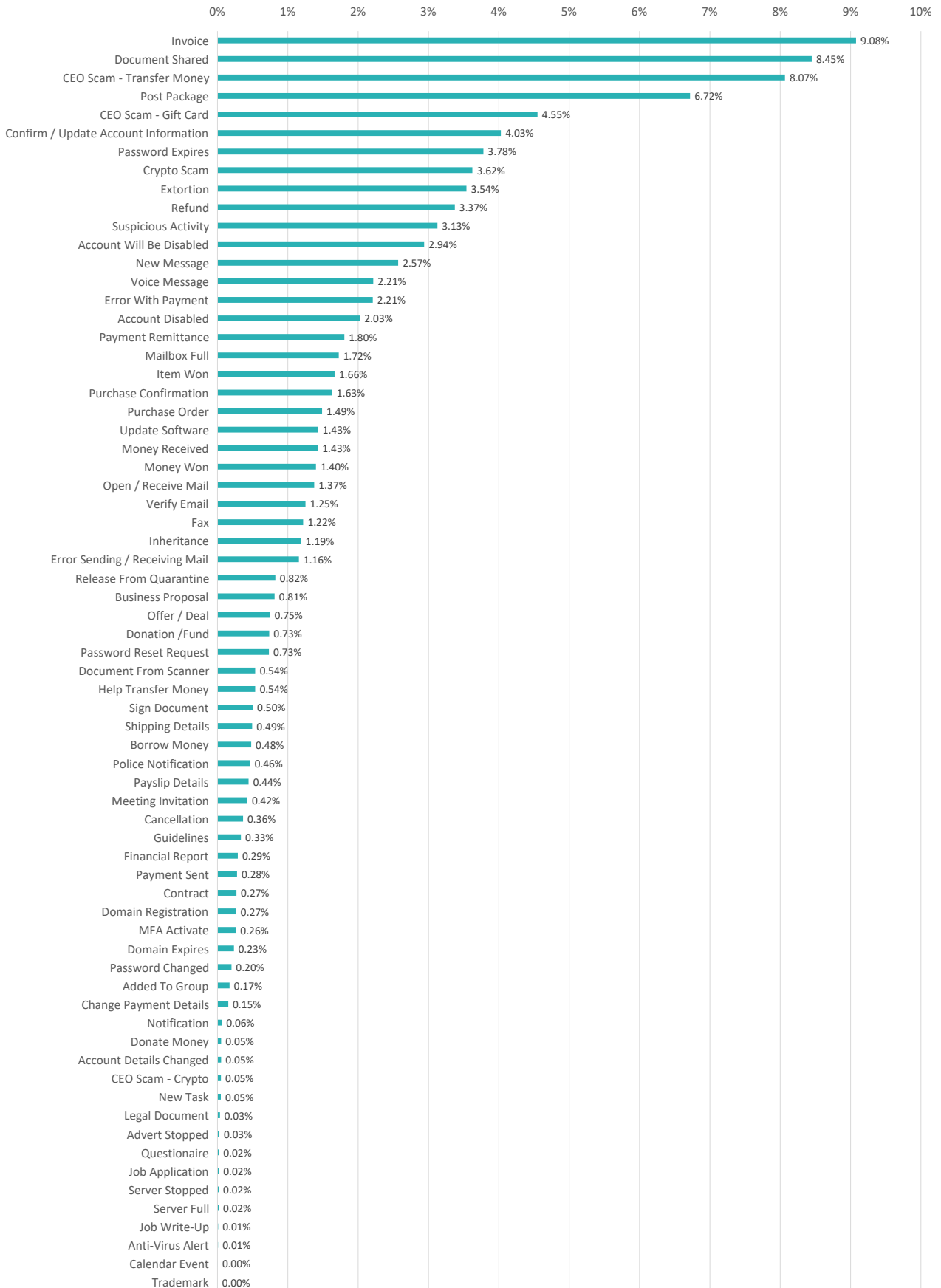


Figure 6.6: Content Distribution

In order to compare the total occurrence rate with the representation for each year, Figure 6.7 shows the occurrence rate for the top 10 overall Content categories for each of the seven collection years.

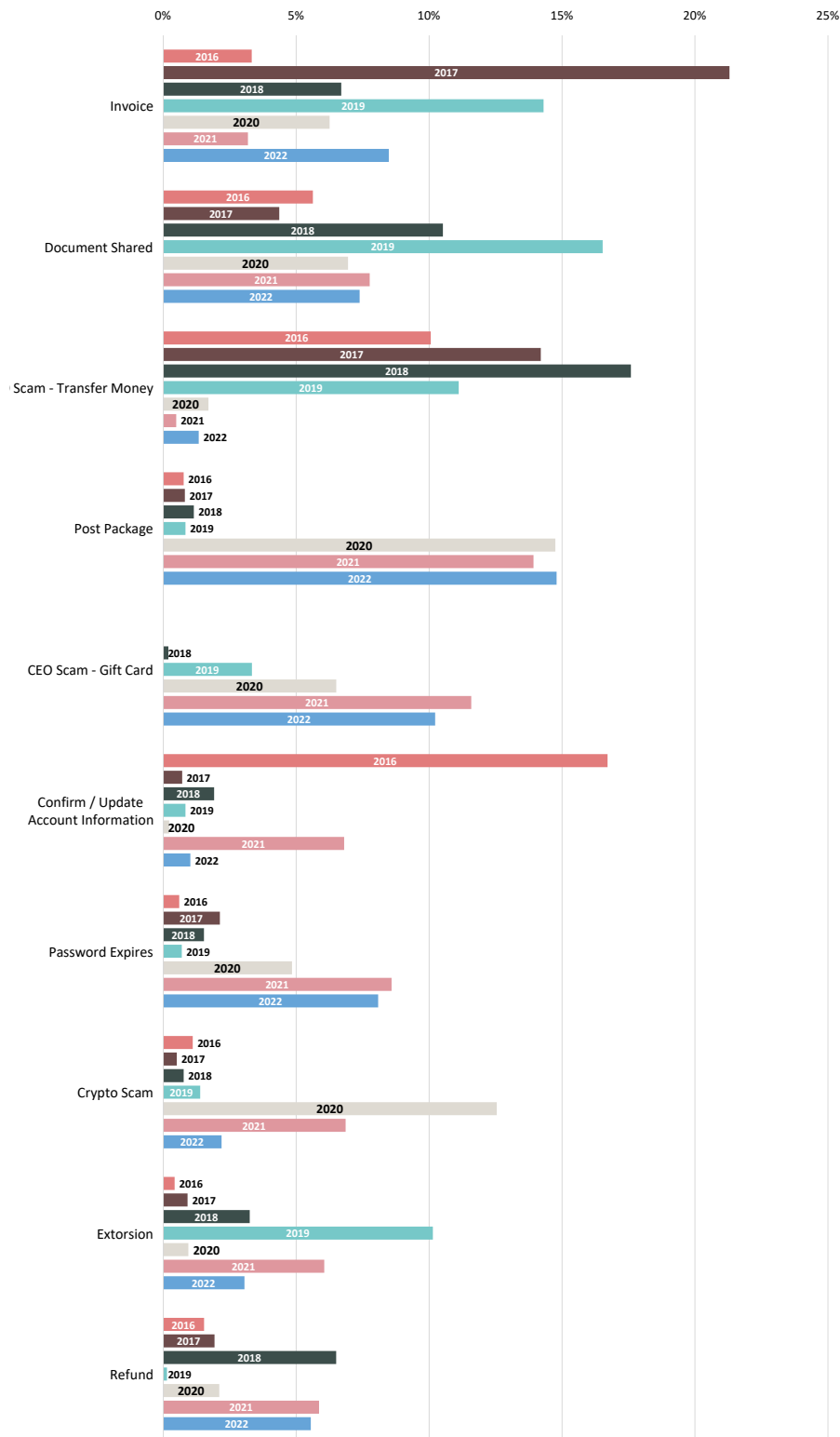


Figure 6.7: Content Evolution (Top 10)

As shown in the Figure 6.7, Invoice at the top has varied quite a bit, and shows no clear pattern in its evolution. In second, the Document Shared Content category has been fairly consistent throughout the years with a surge in the years of 2018 and 2019. Thirdly, the CEO Scam - Transfer Money category has seen a great decline the recent years from being within the top three most observed Content categories from 2016 throughout 2019, to having only a small presence the last three years. The contrary can be observed in the fourth Content category, Post Package, where there was only a small presence within the first four years, and a surge for the last three.

CEO Scam - Gift Card and Password Expires has a similar pattern to that of Post Package, showing a great increase the recent years, while the remainder of the Content categories has a combination of surges in specific years (Confirm / Update Account Information, Crypto Scam), and varied distributions (Extortion, Refund).

To showcase each Content category’s ranking, Figures 6.8 through 6.10 displays highlights from the evolution in the ranking of the overall top 10 categories. The ranking shows its recorded presence compared to the other Content categories, where a ranking of 1 indicates the highest presence.

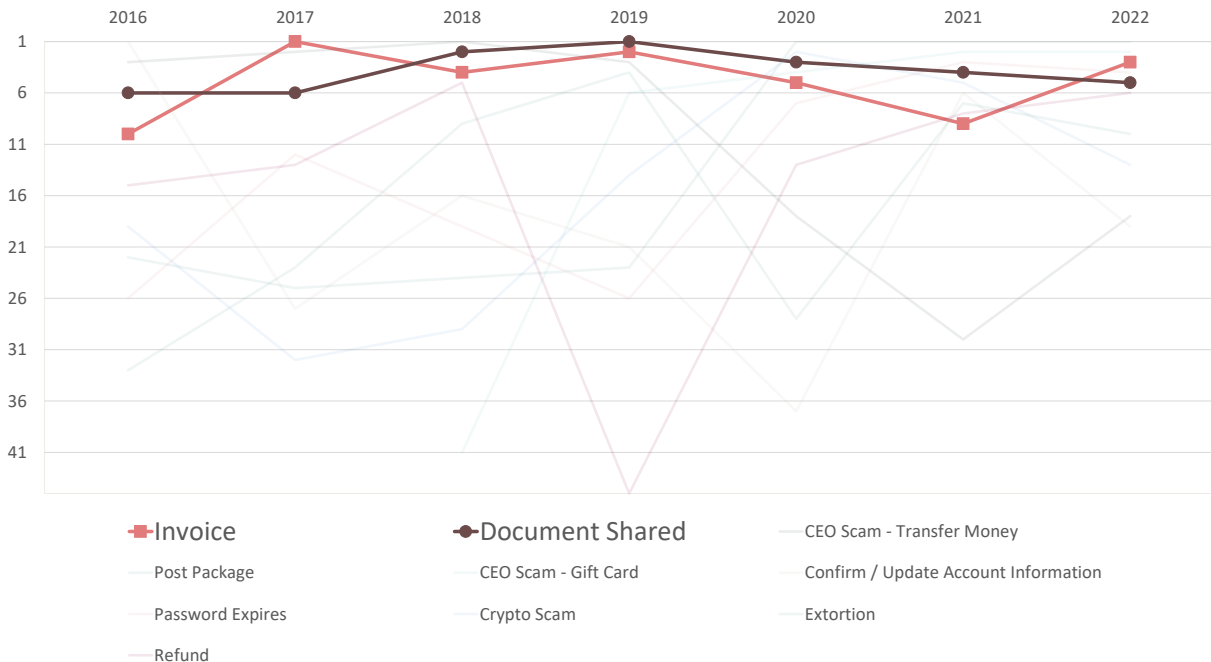


Figure 6.8: Rank Evolution - Invoice, Document Shared

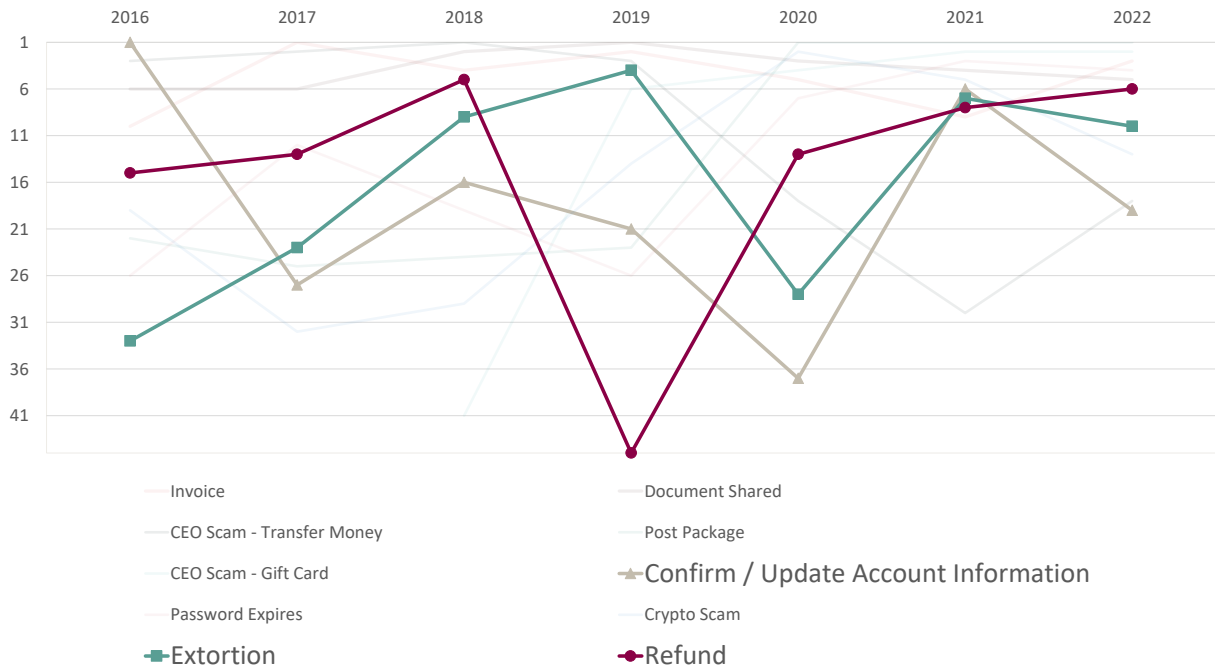


Figure 6.9: Rank Evolution - Confirm / Update Account Information, Extortion, Refund

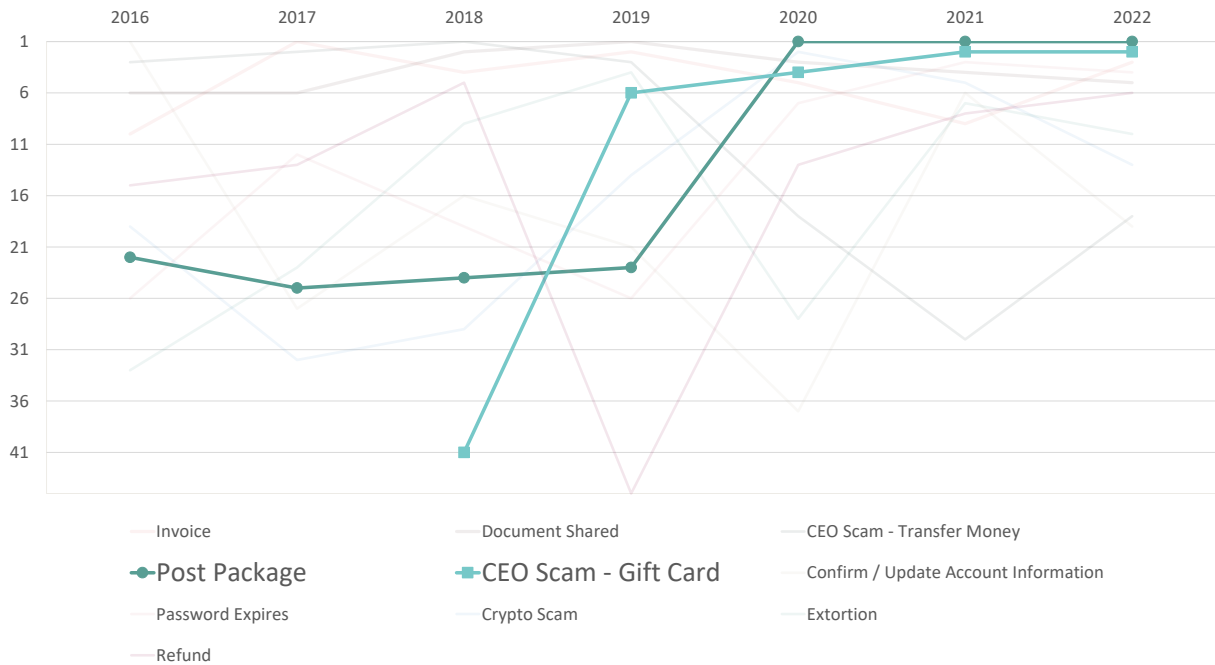


Figure 6.10: Rank Evolution - CEO Scam - Gift Card, Post Package

The evolution graphs provides additional information into the trends for each Content category. Based on the evolution the last seven years, both Invoice and Document Shared showcases a smooth progression throughout the years with no sudden changes in their position (Figure 6.8). On the contrary, the Content categories Confirm / Update Account Information, Extortion, and Refund display no sign of predictable behavior as they are prone to erratic changes, such as going from 5th to 45th and up to 13th within a three year span (Figure 6.9). As for Post Package and CEO Scam - Gift Card, they have shown a steady position the latest years after rising towards the top (Figure 6.10). The remainder of the Content categories not highlighted, including CEO Scam - Transfer Money, Password Expires, and Crypto Scam, shows combinations of erratic and steady progressions depending on the time frame viewed.

On the basis of both the percentwise distribution and rank evolution, one can expect the Content categories of Invoice and Document Shared not to change to drastically in terms of ranking the coming years, while Invoice's percentwise distribution may vary. Due to their now steady position, both Post Package and CEO Scam - Gift Card can be predicted to still remain heavily represented going forward. CEO Scam - Transfer Money was seemingly disappearing up until 2022 where it had a small boost in representation. Due to this increase, one cannot say for certain whether the category is on the rise again, or if it was a one-time increase followed by a decrease again. It is although safe to say that the CEO Scam - Transfer Money category is not as big of a threat as it used to be in the beginning of the collection scope. For the remainder of the Content categories, no prediction can be made on their continued evolution.

6.2.2 Target

A total of six Target categories were identified during the analysis, consisting of Credentials, Money, Credit Card Details, Infect, PII, and Business Information. Table 6.2 displays the overall percentwise distribution of the identified Targets for the years 2016 through 2022.

Target	% Distribution
Credentials	47.15%
Money	25.51%
Credit Card Details	18.12%
Infect	6.75%
PII	1.22%
N/A	1.16%
Business Information	0.09%

Table 6.2: Target Distribution

As can be seen, Credentials were the most sought after Target throughout the scoped years accounting for nearly half (47.15%) of the Targets observed. The monetary categories of Money and Credit Card Details follows suit in second and third, and Infect is ranked in fourth with a 6.75% representation. Towards the end of the list, the Target categories of PII and Business Information is present with a low total representation rate. The emails in which the Target could not be determined had a 1.16% representation, as displayed in the "N/A" row of the table.

Figure 6.11 displays the percentwise distribution of the Target categories for each of the collection years.

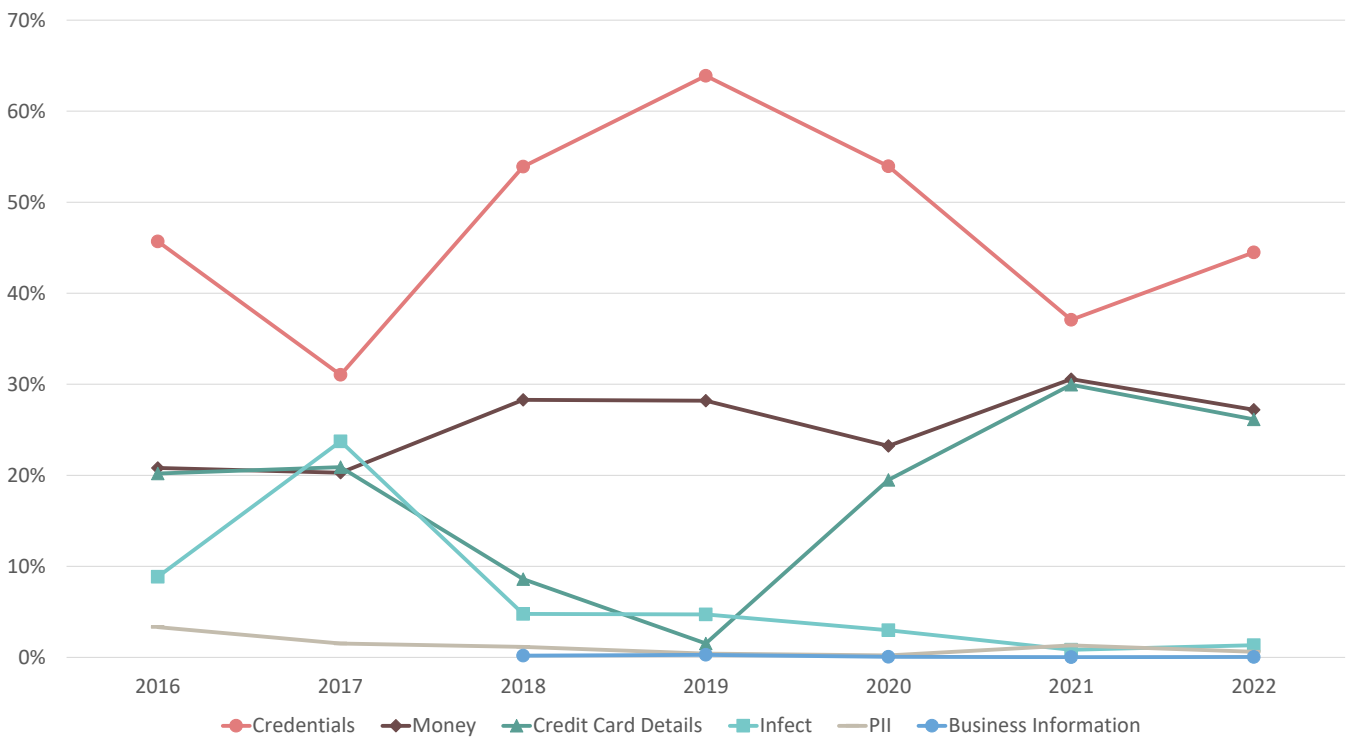


Figure 6.11: Target Evolution

Credentials have consistently been the top ranking Target by a fair margin each year. In 2019, the discrepancy between Credentials and the second highest Target category (Money) was as much as 35.7% in difference (Credentials: 63.89%, Money: 28.19%). Money has been consistently the second most targeted Target since 2018, with an overall smooth evolution line with no erratic changes. Credit Card Details seemingly reverse mirrors the evolution of Credentials, going up when Credentials is declining, and declining when Credentials is increasing. Infect shows the greatest lasting decline, having had its representation reduced from its highest in 2017 at 23.73%, to a 1.33% in 2022. Both PII and Business Information have had a rather low overall representation, with no significant changes observed.

From the observed evolution of the Target property one should expect Credentials to remain heavily targeted in the following years. Money, as well, should not change to drastically if its trend of consistency continues. Credit Card Details is the least predictable out of the six, making it difficult to determine where to expect the category in the coming years, and although Infect has shown being capable of displaying surges, such as from 2016 to 2017, its consistency the recent years can allude to a continued low representation. As both PII and Business Information has not seen any considerable changes throughout the years, and has been consistently low, they can be expected to remain low in representation.

6.2.3 Method

Throughout 2016 - 2022 four Methods were observed, including URL, Communication, Attachment, and Calendar Invite, where attachments again could be divided based on the type of attachment. Figure 6.12 displays the overall percentwise distribution within the Method property.

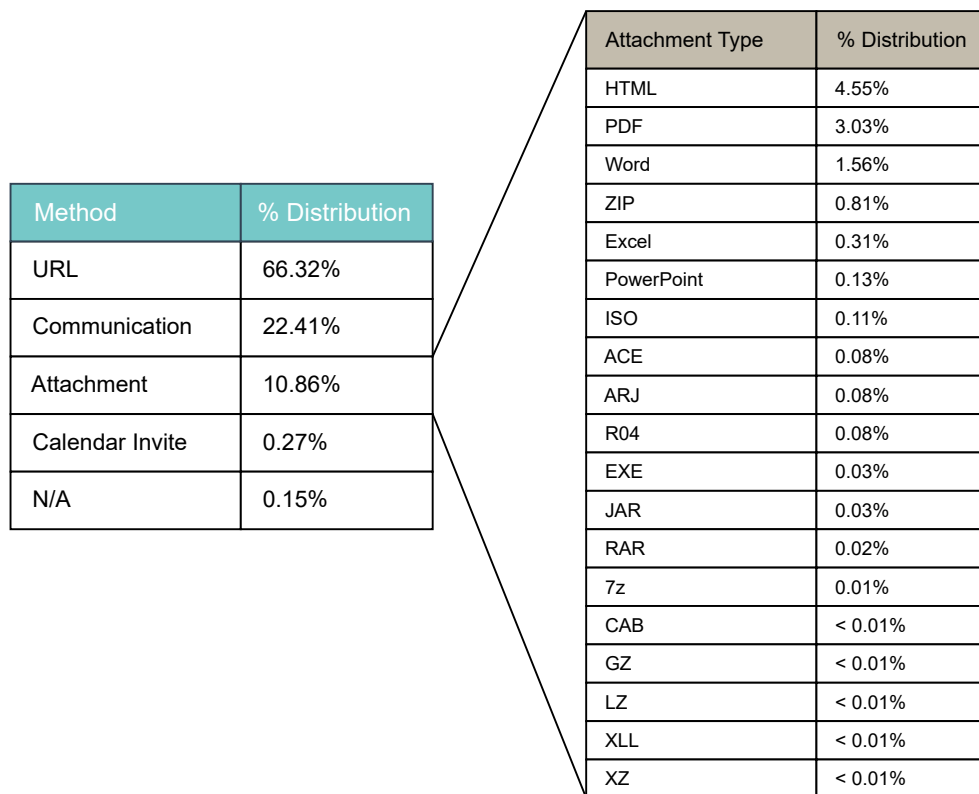


Figure 6.12: Method Distribution

URL is by far the most utilized Method, accounting for 66.32% of the overall representation, followed by Communication and Attachment. The fourth Method, Calendar Invite, differentiates itself with a relatively low representation of 0.27%. Expanding on the Attachment category, there was observed a total of 19 different attachment types. A total of 12 of the file types are archiving / compressed files, meaning they they act as a container for one or more additional files. These 12 includes ZIP, ISO, ACE, ARJ, R04, JAR, RAR, 7z, CAB, GZ, LZ, and XZ. Besides the archiving / compressed files, the Microsoft's Office files belonging to Word, Excel, and PowerPoint have a decent representation. However, HTML and PDF are the most prominent attachment types within the collected dataset.

As shown in Figure 6.13, which visualises the evolution of the different Methods through the collection years, the ranking of the Methods have been persistent throughout all of the years. URL has always been the most utilized Method followed by Communication, Attachment, and Calendar Invite.

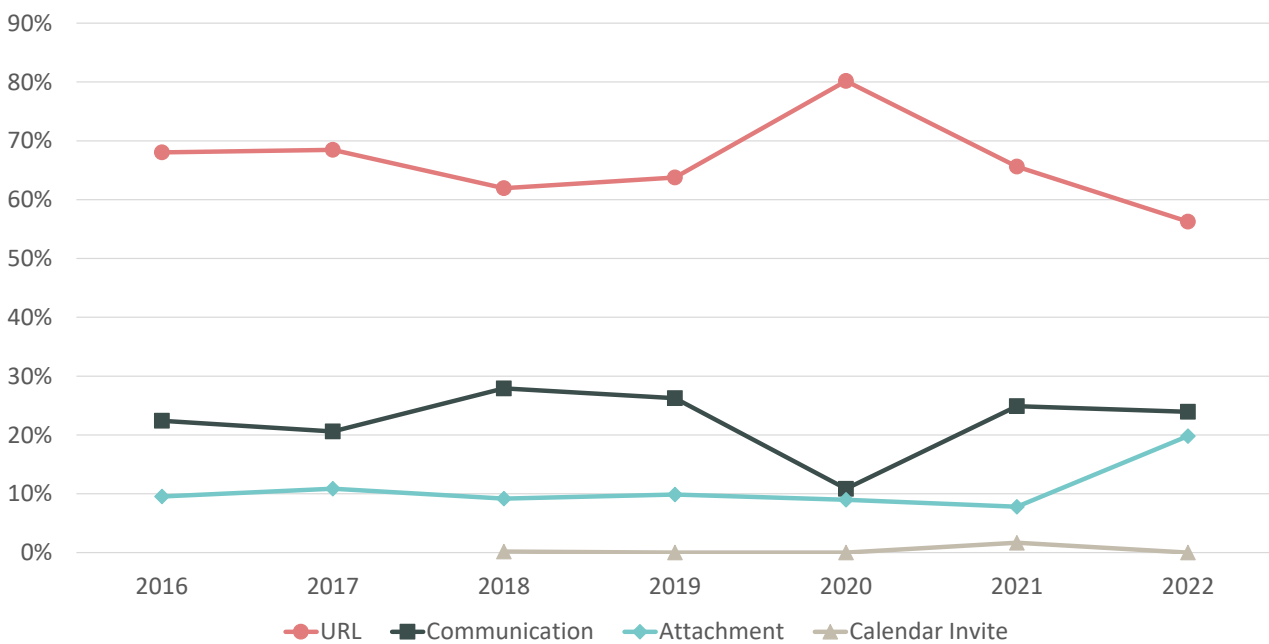


Figure 6.13: Method Evolution

In a similar fashion to the observed relationship between Credentials and Credit Card Details in the section above, Communication seemingly reverse mirrors the evolution of URL. Where URL increases, Communication decreases with a similar amount, and increases whenever URL decreases, with a small exception in 2022 where they both have a decline. Attachment as well, had shown a consistently smooth evolution up until 2022 where it deviated with a surge in representation. Lastly, Calendar Invite is only present in 2018 and 2021, with a low representation rate in both instances.

The evolution of the attachment types, as displayed in Figure 6.14, reveals that the overall top attachment type, HTML, had a low representation rate up until 2020, where it saw a great surge in appearance. Both Word and PDF had a prominent presence in the earlier years, however their representation has seen a great reduction compared to their presence in 2016 throughout 2019.

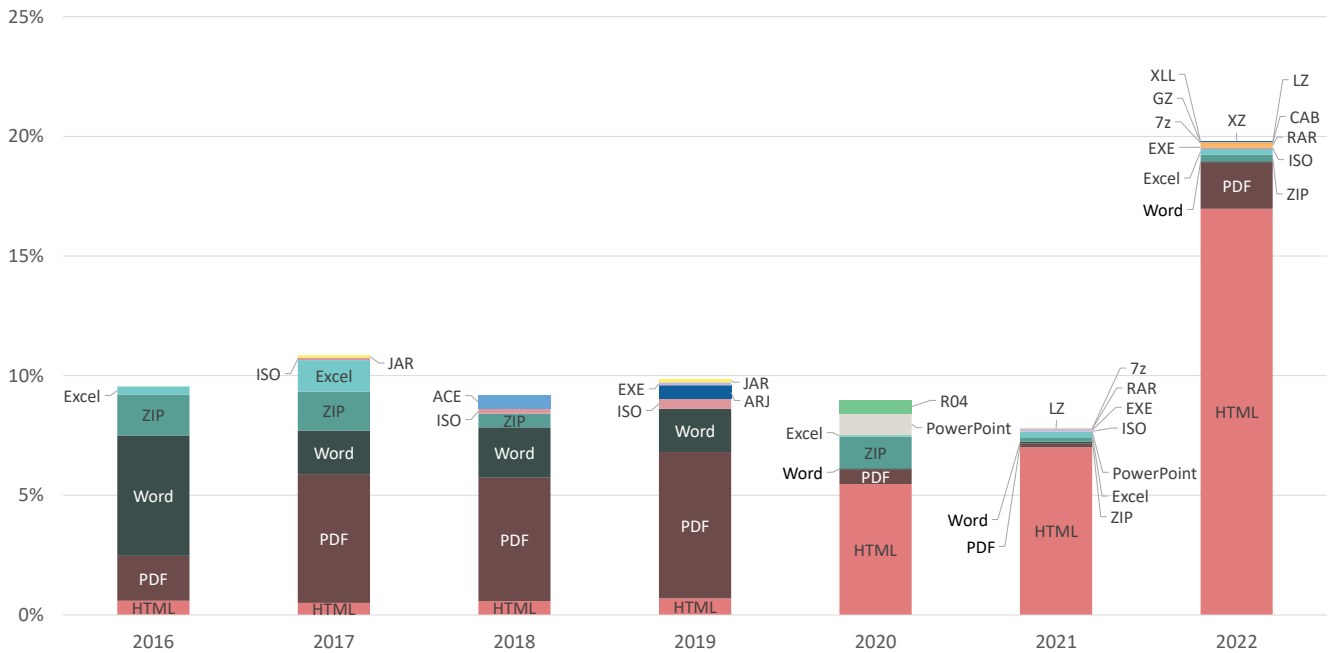


Figure 6.14: Attachment Type Evolution

Most of the attachment types have a low representation rate, with only five having a representation over 0.30%. As shown in Figure 6.14, this is due to the fact that many of the attachment types only appear in one or two years, creating a high turnover.

Based on the observed trends, one can expect the URL Method to continue being a highly utilized Method, however as it is shown to be decreasing, its gap towards the other Methods might continue to decrease. Due to the recent increase in the utilization of attachments, deviating from what was observed the prior years, it cannot be speculated how it will evolve in the coming years. Similarly can be said about the Communication Method as they have both showed signs of a smooth progression and surges. Calendar Invite will seemingly remain low in representation should the trends displayed continue.

6.2.4 Impersonation

A total of 90 Impersonation types were identified in the collected dataset. Two of which were generic impersonations, External and Internal, while the rest could be associated with a specific brand. The split pie chart shown in Figure 6.15, displays the overall percentwise distribution of the generic and top 20 brands observed.

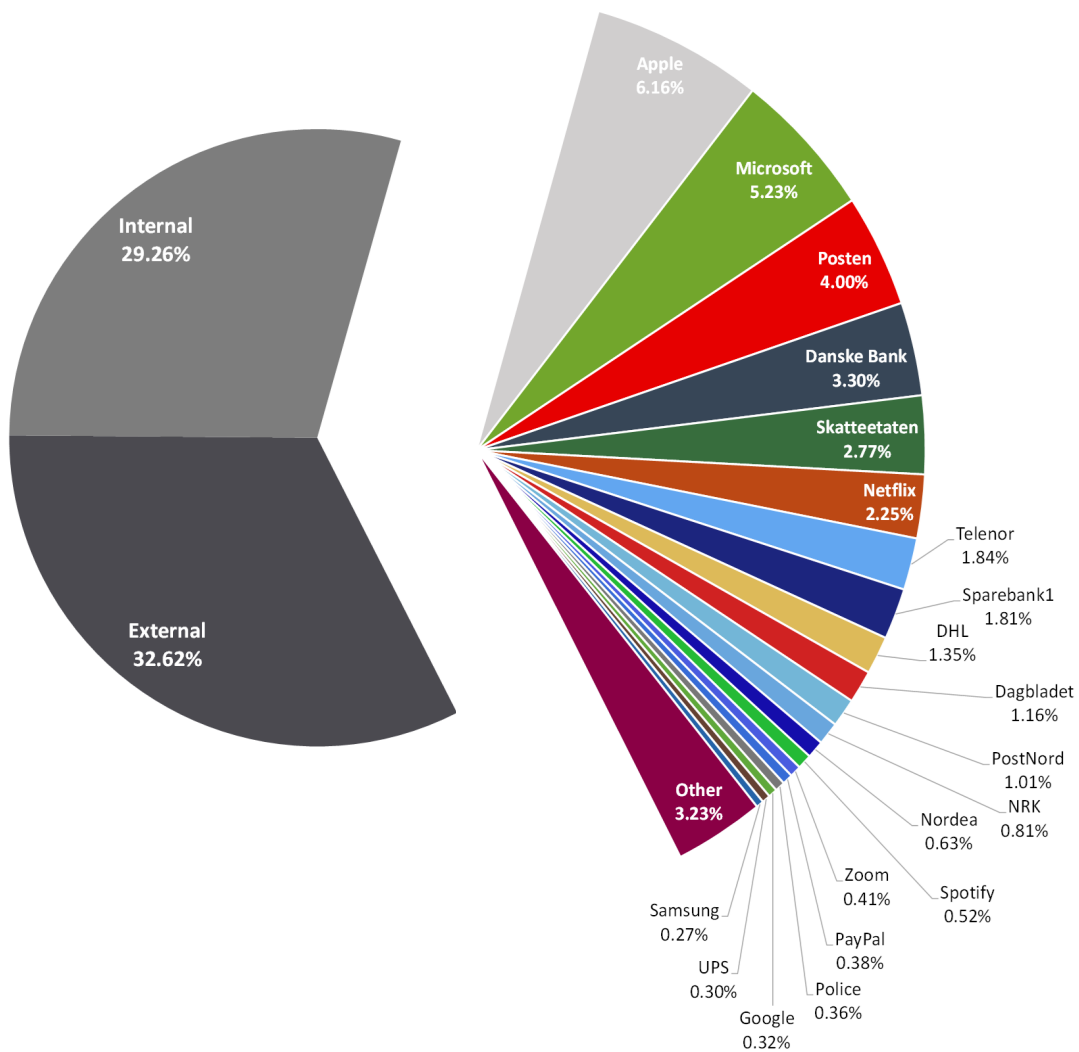


Figure 6.15: Impersonation Distribution

As shown, both the generic Impersonation types are heavily utilized in the emails of the phishing corpus, accounting for close to 62% of the overall representation, with External being the most utilized by a small margin. On the other side of the chart, no one specific brand has been overly represented in the overall scheme of things, with both Apple and Microsoft sharing a similar distribution, closely followed by Posten and Danske Bank.

Figure 6.16 displays the top 10 non-generic Impersonation types observed for each of the seven collection years, sorted by overall representation rate.

No impersonation type has been present in the ten highest for all the years within the collection scope. Both Apple and Microsoft have appeared at the top in six out of the seven years, with Apple not appearing at the top in 2022, while Microsoft has been present since 2017. Danske Bank, which is ranked fourth in the overall overview, can be tied to a surge in 2016, as it has barely been present the remainder of the years. Posten, although ranked third overall, has only been present since 2019, providing insight into the starting point of this particular trend of utilization.

Viewing the Impersonation data, no particular patterns can be observed. Each year presents different distributions, varying both in representation and instances of impersonations. We should however expect the generic Impersonation types to remain heavily utilized in the coming years.

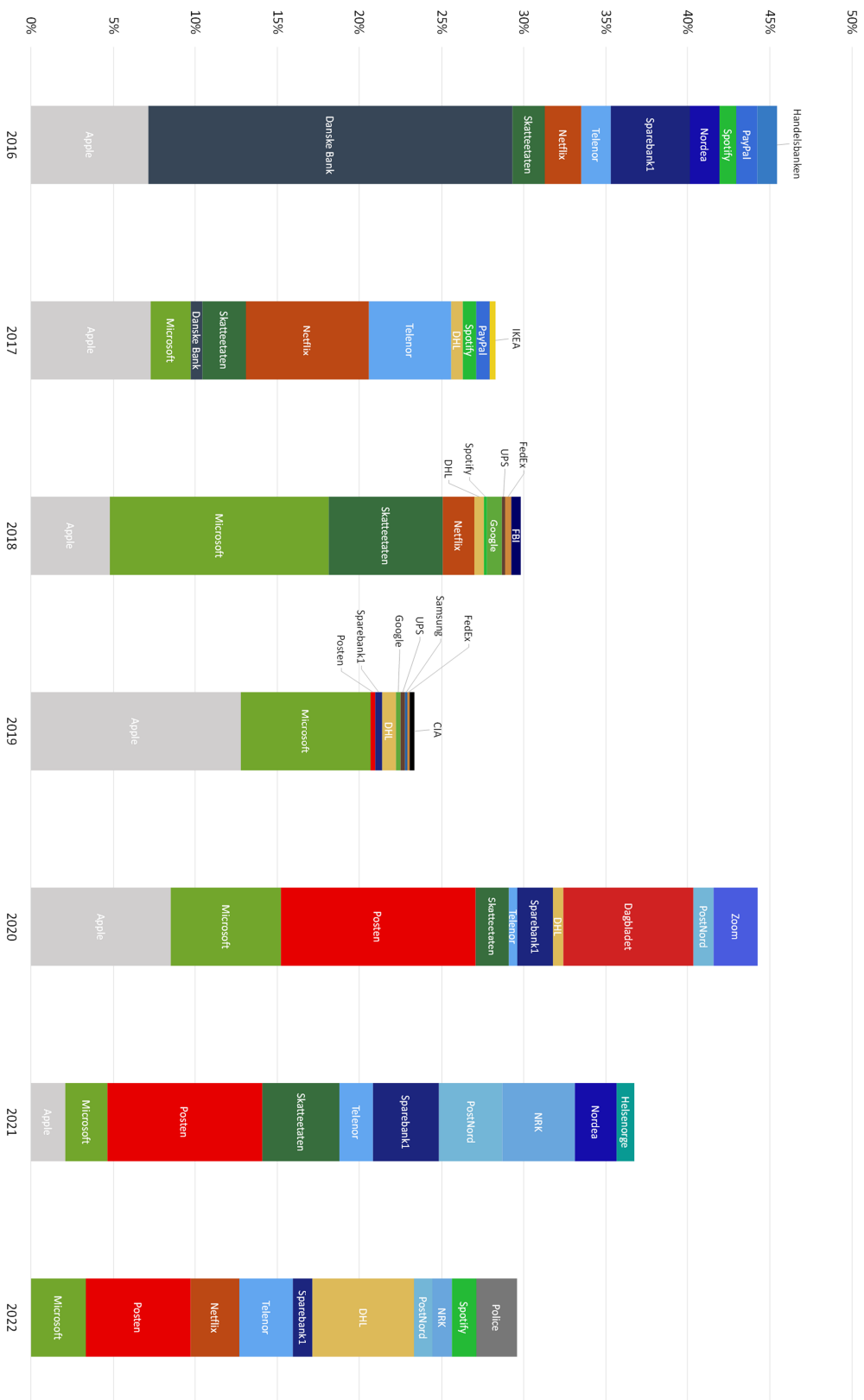


Figure 6.16: Impersonation Evolution

6.2.5 Dates

The distribution of dates, that is, the days in which phishing emails have been observed within the phishing corpus, is presented in a percentwise fashion in the heat-map of Figure 6.17.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31
Jan	0.03	0.07	0.09	0.21	0.11	0.08	0.20	0.15	0.20	0.10	0.24	0.10	0.15	0.36	0.58	0.20	0.36	0.17	0.32	0.09	0.10	0.14	0.36	0.25	0.31	0.21	0.25	0.19	0.23	0.21	0.14
Feb	0.23	0.25	0.15	0.13	0.38	0.35	0.26	0.50	0.16	0.14	0.17	0.28	0.18	0.75	0.14	0.39	0.25	0.31	0.40	0.51	0.66	0.37	0.28	0.19	0.15	0.22	0.21	0.49	0.07		
Mar	0.44	0.56	0.45	0.22	0.11	0.25	0.51	0.25	0.62	0.20	0.28	0.25	0.10	0.32	0.29	0.16	0.17	0.32	0.19	0.28	0.34	0.43	0.38	0.19	0.20	0.17	0.30	0.22	0.27	0.19	0.33
Apr	0.10	0.10	0.18	0.98	0.17	0.30	0.19	0.12	0.11	0.21	0.22	0.25	0.24	0.15	0.07	0.17	0.12	0.16	0.24	0.23	0.13	0.12	0.20	0.21	0.31	0.30	0.25	0.07	0.14	0.24	
May	0.20	0.23	0.21	0.17	0.12	0.11	0.20	0.14	0.22	0.21	0.25	0.17	0.10	0.21	0.22	0.13	0.12	0.12	0.45	0.15	0.18	0.20	0.39	0.46	0.23	0.29	0.16	0.19	0.16	0.48	0.29
Jun	0.19	0.47	0.40	0.23	0.24	0.16	0.36	0.16	0.27	0.14	0.24	0.18	0.28	0.34	0.17	0.22	0.23	0.39	0.16	0.30	0.21	0.36	0.15	0.19	0.12	0.20	0.29	0.41	0.18	0.23	
Jul	0.14	0.15	0.19	0.34	0.18	0.23	0.10	0.07	0.12	0.11	0.12	0.22	0.11	0.14	0.26	0.21	0.09	0.18	0.25	0.21	0.16	0.12	0.23	0.22	0.36	0.39	0.18	0.37	0.38	0.06	0.10
Aug	0.33	0.19	0.46	0.24	0.21	0.26	0.10	0.69	0.19	0.23	0.39	0.18	0.11	0.17	0.24	0.37	0.22	0.69	0.53	0.11	0.09	1.73	0.33	0.30	0.23	0.44	0.22	0.16	0.22	0.41	0.24
Sep	0.07	0.25	0.08	0.49	0.22	0.24	0.15	0.28	0.31	0.16	0.33	0.28	0.22	0.27	0.12	0.31	0.16	0.22	0.30	0.33	0.46	0.33	0.17	0.11	0.45	0.32	0.48	0.38	0.34	0.23	
Oct	0.84	0.36	0.47	0.17	0.32	0.11	0.13	0.22	0.32	0.13	0.28	0.37	0.48	0.56	0.42	0.26	0.26	0.91	0.31	0.14	0.26	0.21	0.37	0.24	0.26	0.43	0.26	0.25	0.32	0.15	0.17
Nov	1.09	0.80	0.38	0.40	0.68	0.29	0.49	0.37	0.69	0.30	0.28	0.76	0.27	0.24	0.68	0.29	0.40	0.19	0.17	0.32	0.37	0.54	0.26	0.30	0.36	0.26	0.16	0.43	0.38	0.61	
Dec	0.44	0.22	0.32	0.25	0.29	0.25	0.58	0.38	0.55	0.42	0.39	0.24	0.22	0.26	0.71	0.24	0.38	0.49	0.28	0.33	0.23	0.09	0.09	0.05	0.23	0.04	0.18	0.11	0.29	0.10	0.01

Figure 6.17: Date Distribution

From the distribution of dates, there are no specific time periods that stands considerably out from the rest of the heat-map. There can be seen a slight increase in activity in the month of November and early December. On the other hand, there is a reduction in activity from late December till early January, as well as in the early middle of June. Although no specific time periods stand out in the heat-map, one particular day, August 22, shows a spike in activity. Filtering the dataset on that specific date, shows that the increase in activity can mainly be tied to a surge in 2016. The particular surge accounted for 10.40% of all phishing mails observed that year.

Stacking each year's heat-map distributions into one row per year, as shown in Figure 6.18, provides insight into whether the observations from the overall overview is due to any reoccurring trends, or, such as with August 22, is due to singular surges. Columns with concurrent darker lines indicates persistent heightened activity. Columns with concurrent light lines indicates persistent low activity. And columns with differentiating heat signatures indicates no particular trends.

Immediately, it is evident that there is no observed activity in January and most of February of 2016. As stated in the Scope section in Chapter 1.4, this is due to the establishment of the ticketing system used to collect the phishing emails.

The stacked distribution confirms that the heightened activity seen in the month of November and early December has been prominent for the majority of the scoped years. The low representation seen before in late December, is shown to be present in the earlier years. However, recently there has been an increase in activity for

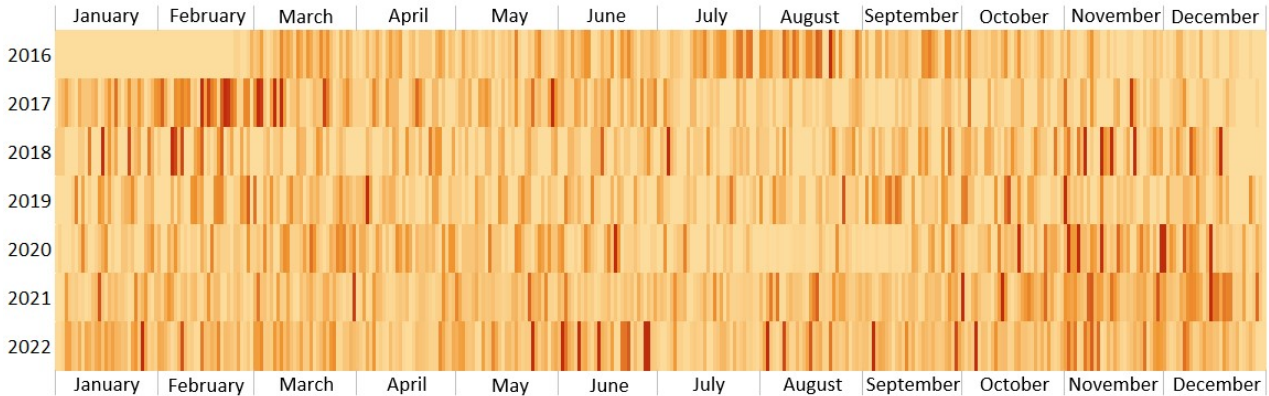


Figure 6.18: Date Distribution Per Year

that time period. The lower activity in early January can be seen to correlate with most of the scoped years for the first couple of days, regardless of the fact that no data could be collected for 2016, while the following days have seen varying activity.

Should the observed trends continue, one should expect to see increased phishing activity in November and in the early days of December. In addition, a low activity level can be expected the first couple of days in January. Besides these, the distribution of dates appears to be varying each year, showing no clear patterns.

6.3 Property Relationships

This section will present the relationships between various of the collection properties, showing how they correlate with each other, as well as any trends observed within these relationships.

6.3.1 Target-Content

The Sankey diagram [59] displayed in Figure 6.19 visualizes the relationship between the Target categories and the Content categories. The diagram presents the relationship flows between one set of values to another, in this case how the phishing emails are distributed into their respective Targets, and how these again are distributed into the observed Content categories.

The visualization of the flows reveals both the density of each Content category as observed in relation to a specific Target, as well as the Content categories' representation within the Targets. For example, showing if a Content category only has a one-to-one relationship with a Target, or if it is utilized for multiple objectives.

Due to the size of the corpus, Figures 6.20 through 6.24 displays each Target's relationships separately.

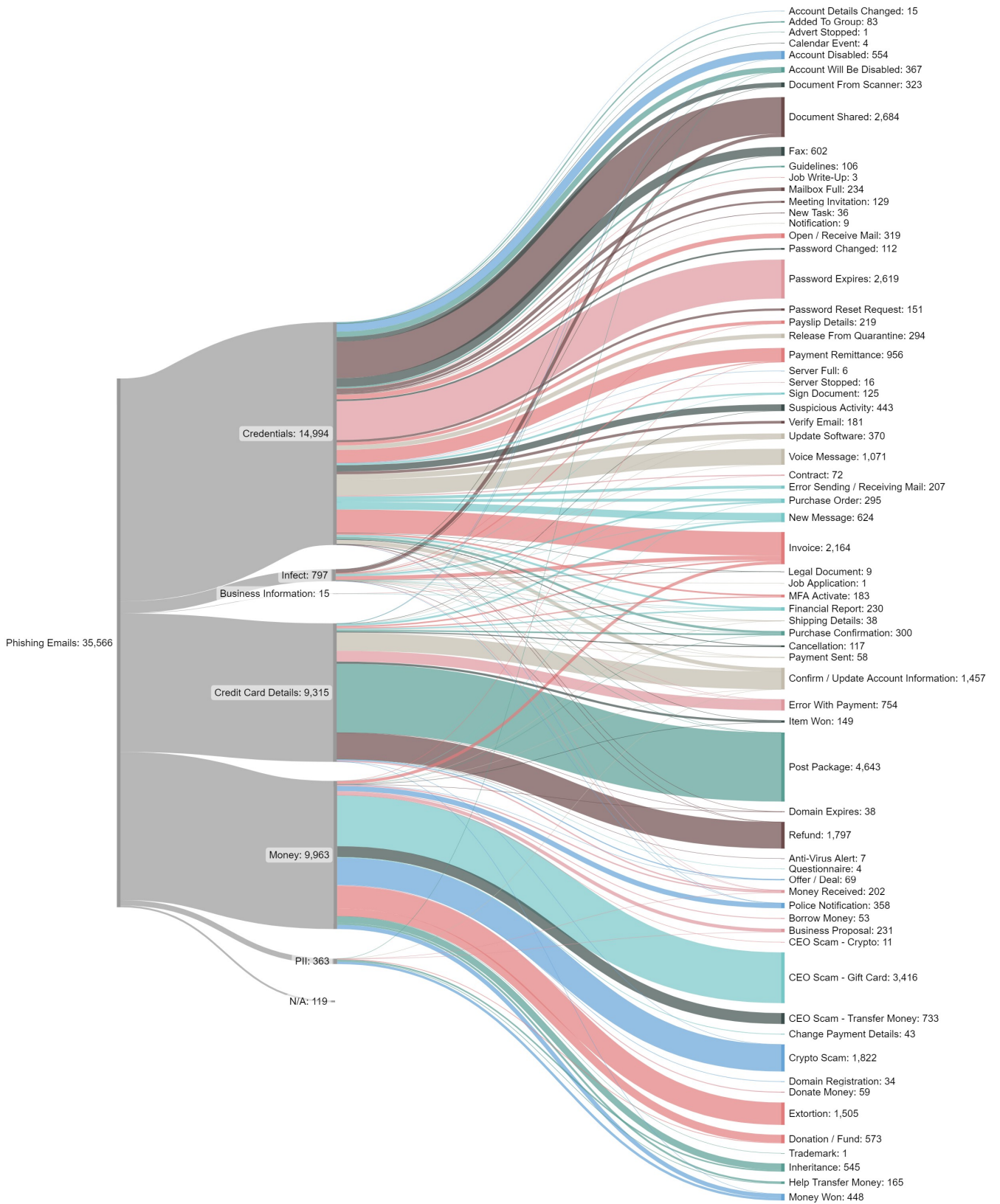


Figure 6.19: Target-Content

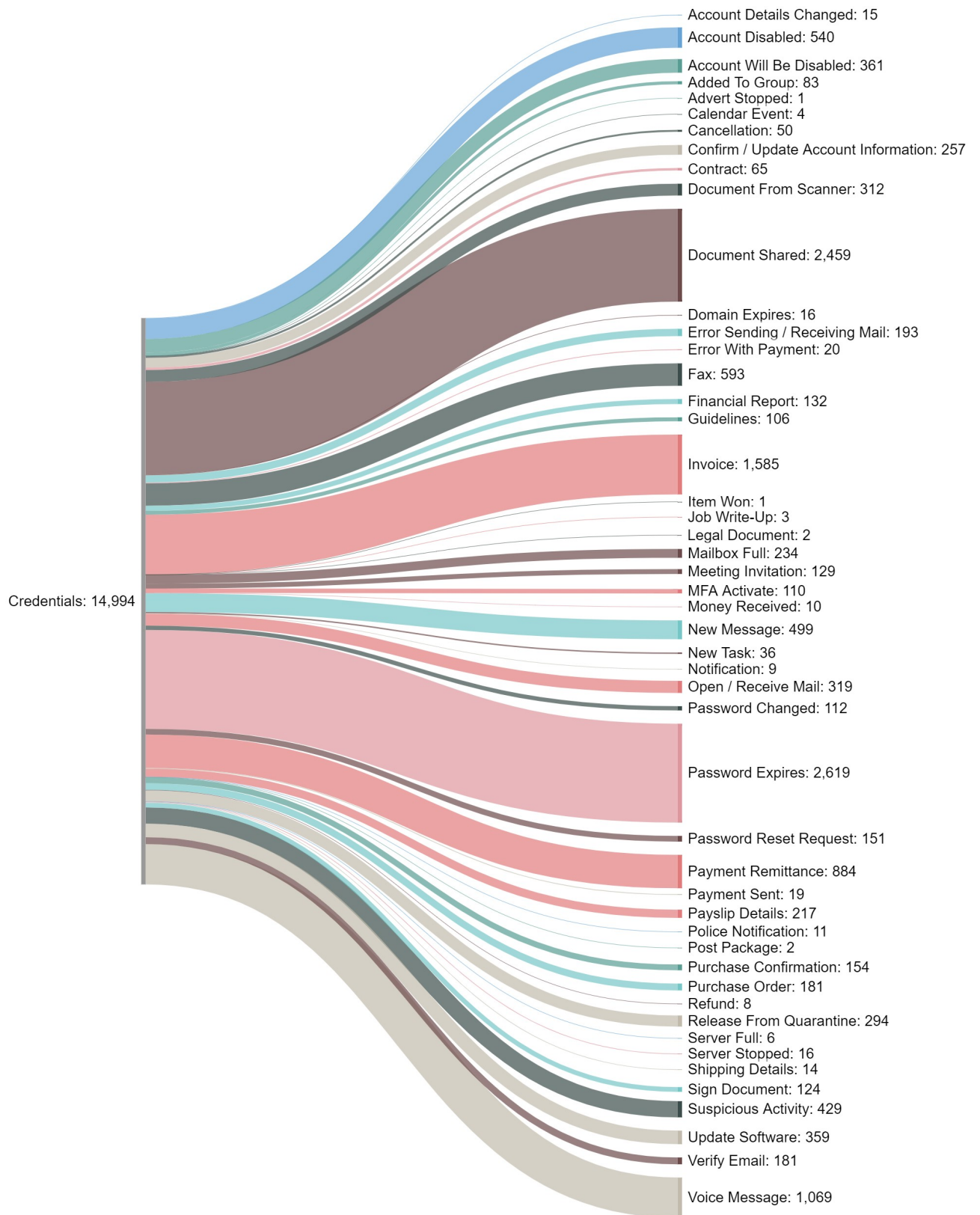


Figure 6.20: Credentials (Target-Content)

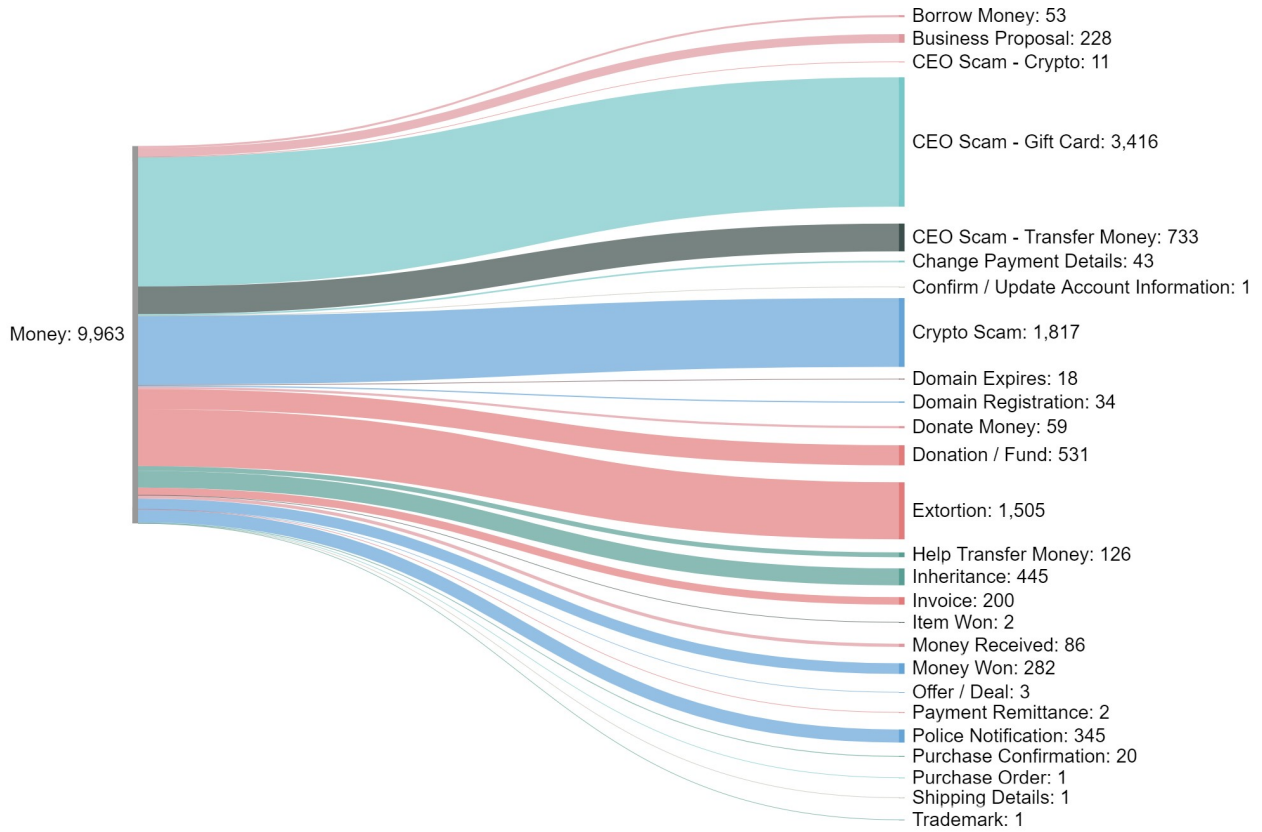


Figure 6.21: Money (Target-Content)

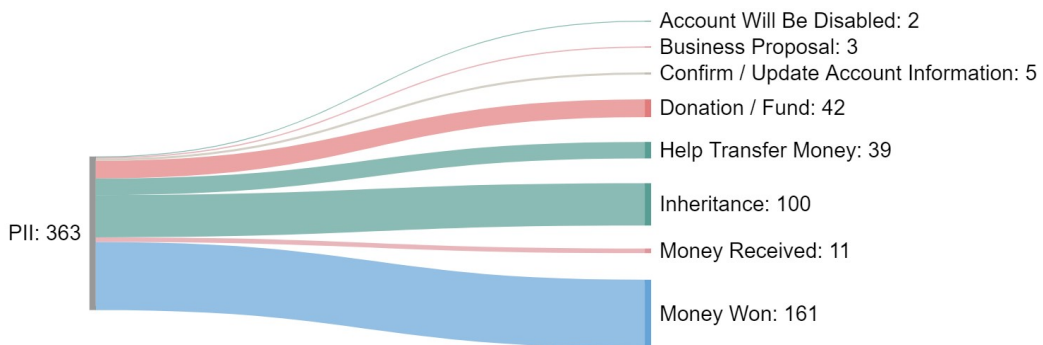


Figure 6.22: PII (Target-Content)

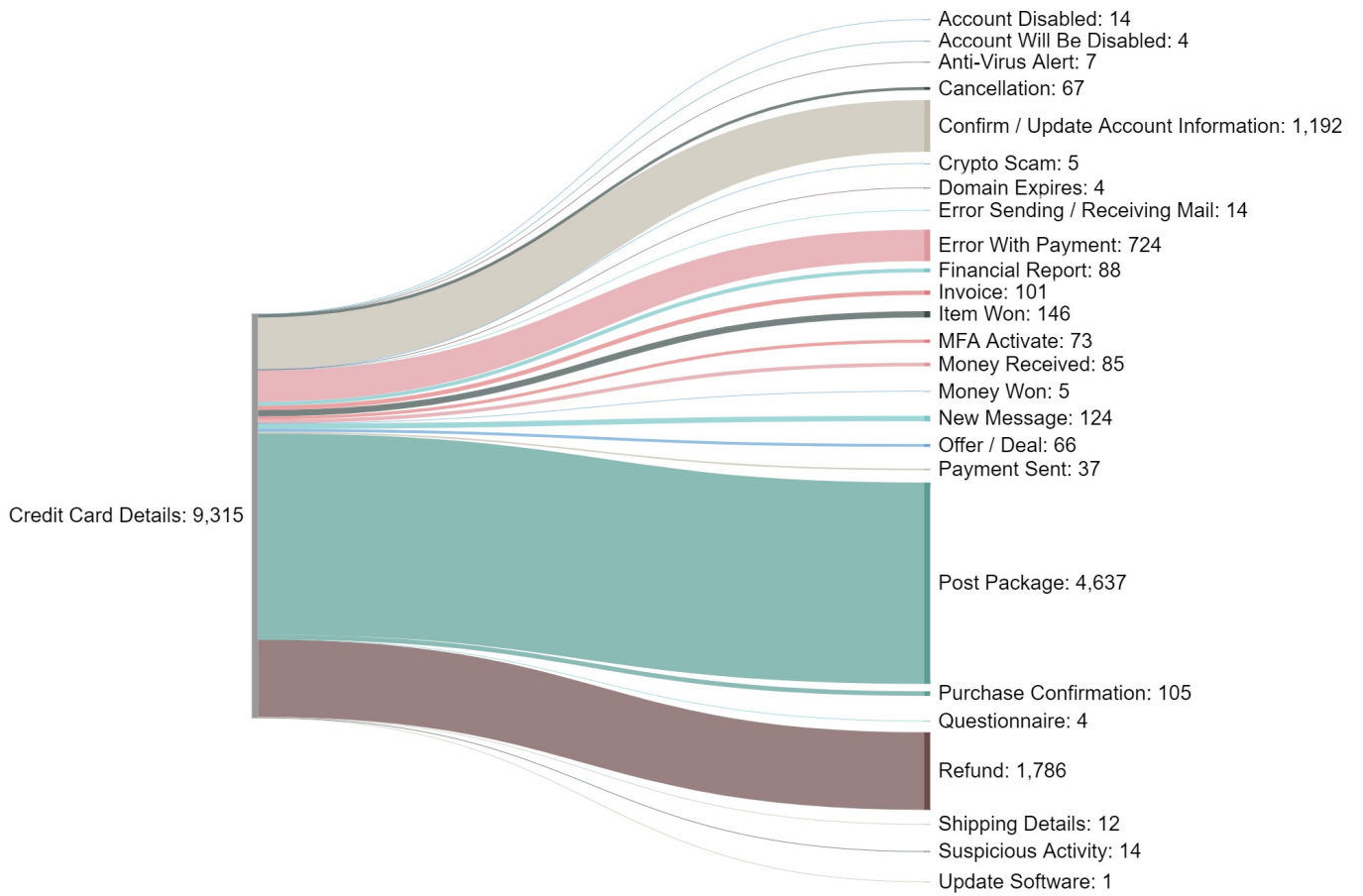


Figure 6.23: Credit Card Details (Target-Content)



Figure 6.24: Business Information (Target-Content)

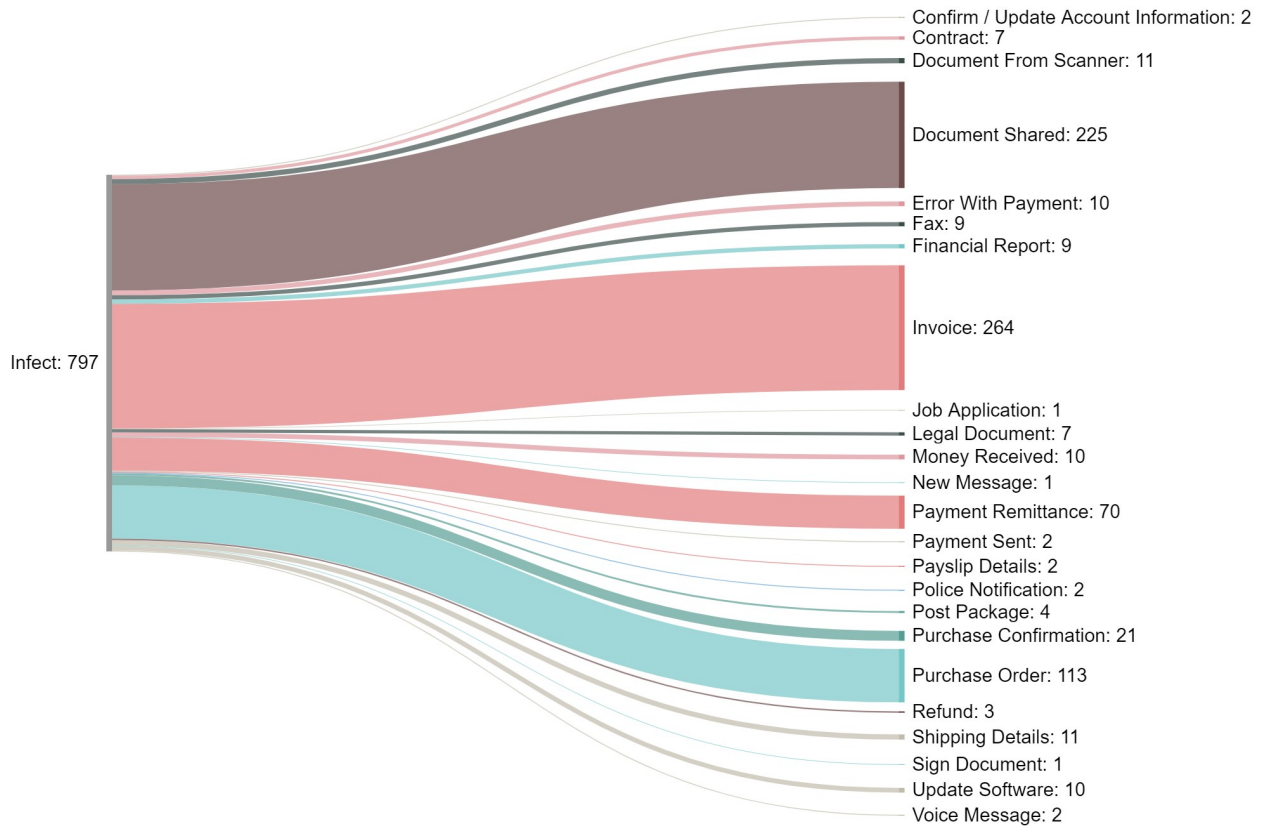


Figure 6.25: Infect (Target-Content)

In total, Credentials (Figure 6.20) have been observed appearing as a Target within 49 Content categories, whereas 18 of them were one-to-one relationships, meaning that these 18 Content categories have only been observed targeting Credentials. Money (Figure 6.21) is observed as a Target in 26 of the Content categories, having 9 Content categories with a singular relationship. Credit Card Details (Figure 6.23) appears in 25 of the Content categories, 2 being a one-to-one relationship. Infect (Figure 6.25) is tied to 24 Content categories, having only one Content category with a singular relationship. PII (Figure 6.22) has 8 Content categories, with one singular relationship as well. Lastly, Business Information (Figure 6.24) is observed in 2 Content categories, with no one-to-one relationships.

The significance of the singular relationships displays a lack of flexibility within the given Content category, such as if they are tailored specifically towards one goal. On the other hand, there is the Content categories that display a high level of flexibility, tying into several of the Target categories. The Content categories of Invoice, Confirm / Update Account Information, and Money Received are the

most flexible categories as they have been observed appearing in five out of the six Target categories (Figure 6.19). Invoice appears in Credentials, Money, Credit Card Details, Infect, and Business Information, while the latter two appears in Credentials, Money, Credit Card Details, Infect, and PII.

The amount and distribution of Content categories related to a Target can signify the diversity of approaches utilized by the malicious actors in order to achieve their goal. To show each of the Target's diversity, Simpson's Diversity Index [60] can be utilized. The index is calculated using the following formula:

$$D = 1 - \left(\frac{\sum n(n-1)}{N(N-1)} \right)$$

Where n represents the number of entries for each Target-Content category relationship, (n_1, n_2, \dots, n_z) , summarized over the number of relationships Z , and N is the total number of entries for the specific Target (thus $\sum n = N$). A diversity index of 1 indicates high diversity, while an index of 0 indicates no diversity.

Table 6.3 displays the diversity index for each of the Target categories.

Target	Total	Total Connections	$1 - \left(\frac{\sum n(n-1)}{N(N-1)} \right)$
Credentials	14994	49	0.91
Money	9963	26	0.81
Credit Card Details	9315	25	0.69
Infect	797	24	0.78
PII	363	8	0.70
Business Information	15	2	0.13

Table 6.3: Target-Content Diversity

Credentials has the highest observed diversity, with a diversity index of 0.91. Money with 0.81 and Infect with 0.78 is second and third in their diversity. PII with 0.70 and Credit Card Details with 0.69 follow suit, while Business Information has displayed the lowest diversity rate, with an overall index of 0.13.

As displayed in Figure 6.26, Credentials have been fairly consistent in its diversity throughout the years, having no changes greater than 0.13, averaging on an index of 0.89, with the lowest score being 0.79 and highest being 0.94. The other Targets however, shows no clear consistencies, varying a great amount throughout the collection years.

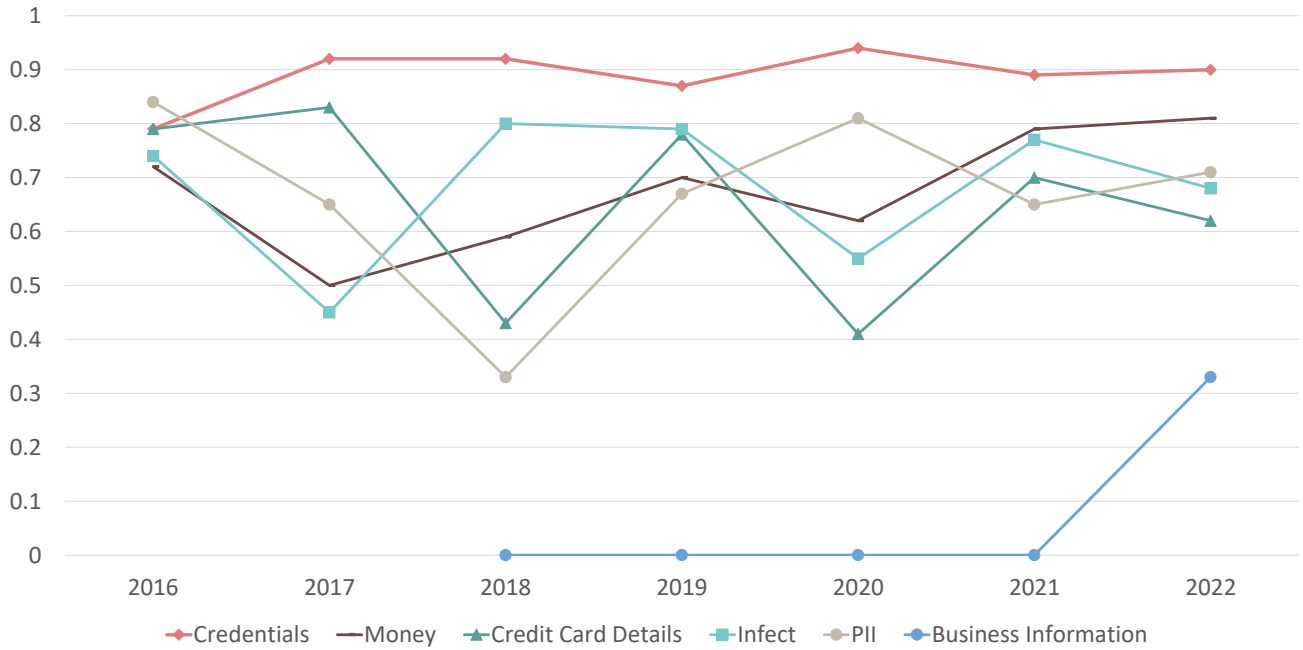


Figure 6.26: Target-Content Diversity Evolution

Should the observed trends continue, Credentials can be expected to remain high in its diversity, having a great magnitude of Content categories tied to it self. The remainder of the Targets have not shown any clear patterns as they have behaved differently each of the years within the collection scope.

6.3.2 Method-Target

The relationship between Method and Target provides insight into any preferred method of achievement for the various Targets. Figures 6.27 through 6.30 displays the relationship between Method and Target for each of the observed Methods, excluding instances of N/A.

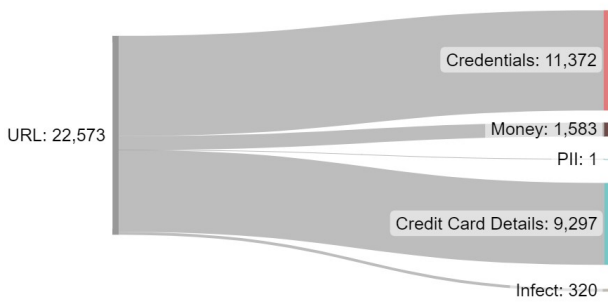


Figure 6.27: URL (Method-Target)

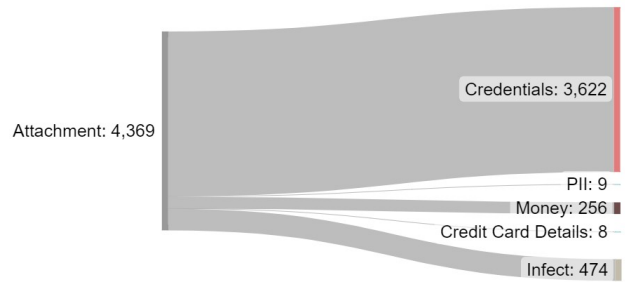


Figure 6.28: Attachment (Method-Target)

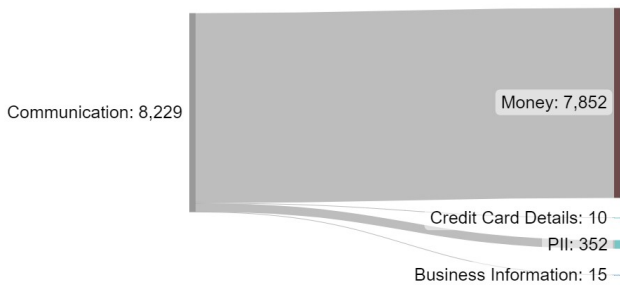


Figure 6.29: Communication (Method-Target)



Figure 6.30: Calendar Invite (Method-Target)

The distribution of Targets for each of the Methods reveals the degree of flexibility within the Methods. Both URL and Attachment have been utilized towards five of the observed Targets, while Communication has been utilized towards four. Calendar Invite appears to be the least flexible as it has mostly been used towards one singular Target. Although most of them shows flexibility, there can be identified preferences within the Methods. Such as Communication being mostly used to target money, while Attachments are mostly used to lure out the Credentials of the recipient. URL is more two-fold, as it is heavily used for both Credentials and Credit Card Details.

Viewing the observations from each year, detailed for each respective year in the full summary (Appendix A), the Target of Credentials was consistently majorly targeted via the utilization of email links. However, a change was observed the last years of the collection scope, in which Attachments began to become a popular Method for luring out the credentials of the recipient. Figure 6.31 highlights this change, showing a 81%/19% and 61%/39% distribution in 2021/2022 compared to an average of 93%/7% the prior years.

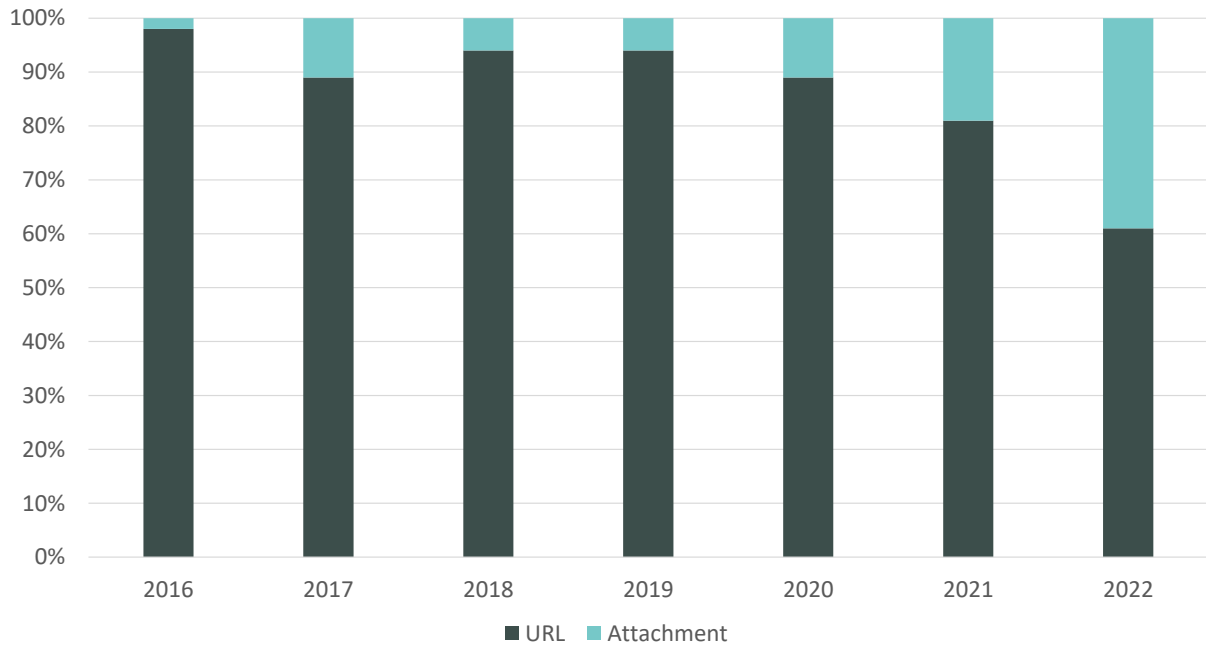


Figure 6.31: Credentials (Method-Target)

The remainder of the Method-Target relationships were fairly consistent throughout the collection years. Money and Business Information were mainly targeted through the utilization of Communication, and Credit Card Details through URLs, while Infect and PII were a bit split throughout, showing no particular preferred method of achievement.

6.4 Summary of Findings

The collection of phishing emails from the years 2016 throughout 2022 resulted in a dataset consisting of 35566 phishing emails. There were identified 68 unique Content categories, 6 Target categories, 4 Method categories, and a division between generic and non-generic Impersonation categories, where a total of 90 non-generic types were identified.

From the analysis of the Content property, the categories of Invoice, Document Shared, and CEO Scam - Transfer Money had the overall largest presence throughout the collection years. Viewing the trends, if they are to continue as seen, both Invoice and Document Shared can be expected to remain highly represented in the following years. CEO Scam - Transfer Money, on the other hand, has seen a great decline the recent years and cannot be expected to remain one of the most represented Content categories. The Content categories of CEO Scam - Gift Card and Post Package has shown a great incline the recent years, and can be expected to continue their presence should the observations hold true for the coming years.

The findings from the Target property shows that Credentials have been, and still is, the most sought after Target in the world of phishing. Although the percentwise distribution between the Credentials and the other categories has varied quite a bit, it has consistently been the most targeted Target throughout the collection scope. On the basis of the observations from this property, Credentials should remain a highly sought after Target. Money, as well, has had a rather even evolution and can be expected not to change too drastically should the observed behavior continue. The remainder of the Target categories have not displayed any clear patterns, and are difficult to determine any future states.

The collection of the Method property resulted in the identification of the four methods URL, Communication, Attachment, and Calendar Invite. The URL method is by far the most utilized, having an overall representation of over 66%. However, recent years shows that the URL Method has seen a decline in utilization. Where the utilization of URLs decreased, Communication saw an increase, followed by an increase in the usage of attachments. The increase in the usage of attachments is accompanied by a large increase in the utilization of the HTML type of attachment. Should the observed trends continue, one can expect for URLs to still be highly utilized, however its gap towards the rest of the Methods may continue to decline.

From the Impersonation property, the generic categories of External and Internal were the most utilized throughout the collection years. For the non-generic categories, Apple and Microsoft had an overall prominent presence, followed by Posten and Danske Bank. Besides Danske Bank's representation mostly being tied to a surge in one specific year, no clear patterns can be made out for the remainder of

the non-generic Impersonation types. It can however be expected that the generic categories continue to be the most utilized.

As for the dates, there are no clear time periods that have been observed with considerably more phishing activity than the nearby time ranges. A slight increase in activity can be observed in the month of November and beginning of December, which has been persistent for most of the years within the collection scope. Should this trend persist, heightened activity in November and early December should be expected.

The relationship between Target and Content reveals that the Credentials Target is the most diverse, having both been appearing in the most Content categories, as well as having a fair distribution within these Content categories. The other Targets, with the exception of Business Information, do not appear too far behind in their diversity either, showing a mostly diverse and flexible Target-Content relationship.

Lastly, the Method-Target relationship shows that both URL and Attachment has a great deal of flexibility as they both were observed used toward five of the six identified Targets. The Target of Credentials was mostly tied to the URL Method category, while Money was mostly tied to the Communication Method category. Although Credentials overall were mostly linked to the URL Method category, an increase in the utilization of attachments for this Target was observed in the last couple of years.

Chapter 7

The Email Phishing Collection Model

In Chapter 5, based on prior literature and reports, a set of five properties relevant for the analysis of email phishing were established. For each of these properties, consisting of Content, Target, Method, Impersonation, and Date, email elements necessary for the collection of said properties were subsequently identified based on RFC5321 and RFC5322. These properties, with their corresponding email elements, make up the core of the Email Phishing Collection Model. Figure 7.1 presents an overview of this model, showing each model property and the corresponding email elements utilized to determine said property.

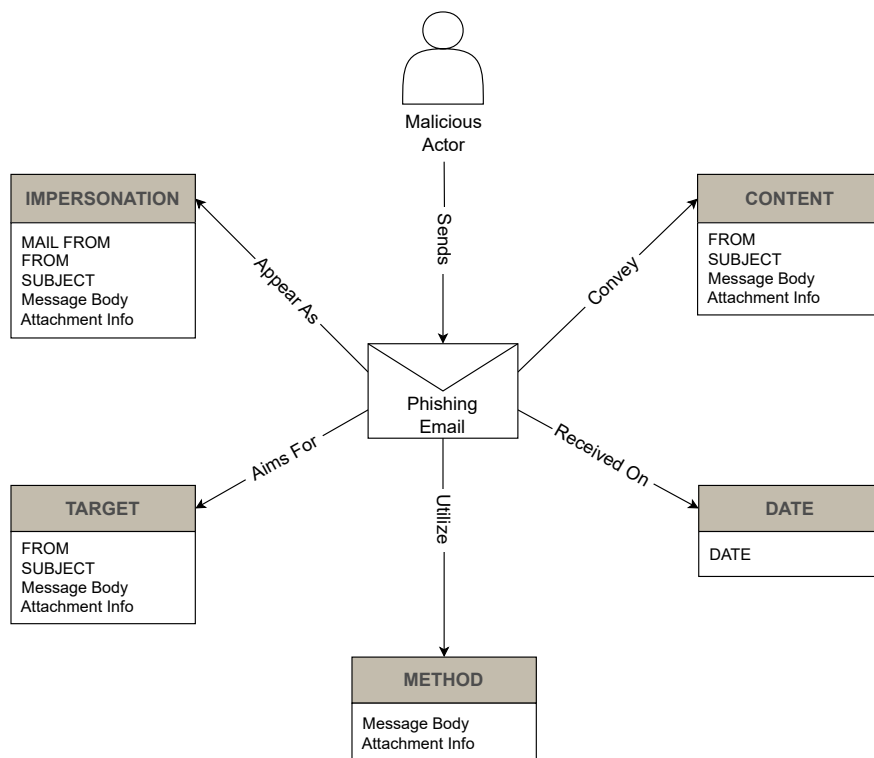


Figure 7.1: Model Overview

On the basis of the model figure (Figure 7.1), the model can be put into context through the following definition:

*A malicious actor sends a phishing email received on **Date** aiming for **Target** through the utilization of **Method** by appearing as **Impersonation** and conveying **Content**.*

Through the collection of the model properties from the email phishing corpus, subsequent categories within the properties were identified, finalizing the model developed. The remainder of the chapter presents the model in its entirety, showing its indented use flow, as well as defining the model properties, corresponding email elements, and identified categories within each of the model properties.

7.1 Model Use

To showcase the intended utilization of the model, Figure 7.2 is defined, providing a simple view of the flow and fundamental parts of the process of utilization.

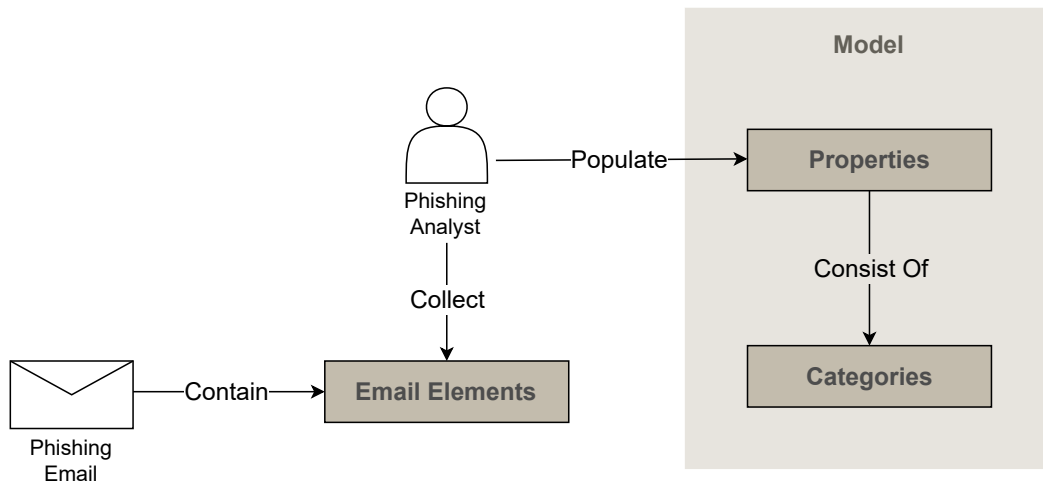


Figure 7.2: Model Use

The user of the model, being a person or an automated process, collects the email elements from the phishing email and populates the model properties with relevant categories based on said email elements.

7.2 Content

The Content property concerns the essence of the phishing email. It is this aspect of the mail that presents the recipient with a plausible reason for the reception of the mail and consequent actions to be performed. With other words, the Content revolves around how the malicious actor convinces the recipient of the email's legitimacy. In order to appropriately assess the Content of the phishing email the following email elements should be analyzed: header FROM field, header SUBJECT field, Message Body, and any Attachment Information.

68 different Content categories have been identified and exemplified based on the observations made from the phishing corpus of this thesis. Table 7.1 provides an overview of all categories identified for the Content property (See Appendix B for a full description of each Content category).

Content Categories				
Account Details Changed	Confirm / Update Account Information	Guidelines	New Task	Purchase Order
Account Disabled	Contract	Help Transfer Money	Notification	Questionnaire
Account Will Be Disabled	Crypto Scam	Inheritance	Offer / Deal	Refund
Added To Group	Document From Scanner	Invoice	Open / Receive Mail	Release From Quarantine
Advert Stopped	Document Shared	Item Won	Password Changed	Server Full
Anti-Virus Alert	Domain Expires	Job Application	Password Expires	Server Stopped
Borrow Money	Domain Registration	Job Write-Up	Password Reset Request	Shipping Details
Business Proposal	Donate Money	Legal Document	Payment Remittance	Sign Document
Calendar Event	Donation /Fund	Mailbox Full	Payment Sent	Suspicious Activity
Cancellation	Error Sending / Receiving Mail	Meeting Invitation	Payslip Details	Trademark
CEO Scam - Crypto	Error With Payment	MFA Activate	Police Notification	Update Software
CEO Scam - Gift Card	Extortion	Money Received	Post Package	Verify Email
CEO Scam - Transfer Money	Fax	Money Won	Purchase Confirmation	Voice Message
Change Payment Details	Financial Report	New Message		

Table 7.1: Model - Content Categories

7.3 Target

The Target property defines the objective that the malicious actor tries to achieve with the phishing email. This is the objective that is tried lured from the recipient of the mail. In order to determine this objective, the email elements of header FROM field, header SUBJECT field, Message Body, and any Attachment Information should be analyzed.

6 different Target categories are identified for this model, consisting of Credentials, Money, Credit Card Details, Infect, Personal Identifiable Information (PII), and Business Information. Table 7.2 provides an overview of the Target property's categories with corresponding description and examples.

Target Category	Description	Examples
Credentials	Phishing emails that seek any details used to authenticate oneself on digital mediums.	Password, One-Time Passcodes, PINs
Money	Phishing emails that aim to acquire direct monetary assets.	Gift Cards, Digital Payments, Transfers
Credit Card Details	Phishing emails that aim to acquire information on the recipient's payment card that can be used to access its assets.	Card Number, Expiry Date, and CVV
Infect	Phishing emails that contain or lead to content containing malicious code.	Trojans, Worms, Viruses
PII	Phishing emails that seek out identifiable information on the recipient.	Name, Age, Address, Phone Number
Business Information	Phishing emails that seek to gain non-public business information.	Invoice details, Account balance, Customer lists

Table 7.2: Model - Target Categories

7.4 Method

The Method property concerns the approach utilized within the phishing email by the malicious actor in order to obtain the desired Target. This is the component of the phishing email that the malicious actor wants the recipient to interact or respond to. To determine the Method property the email elements consisting of Message Body and any Attachment Information are relevant to collect.

4 different Method categories have been identified for the model, including URL, Communication, Attachment, and Calendar Invite. Table 7.3 provides an overview of the Method categories with corresponding description and examples.

Method Category	Description	Examples
URL	Phishing emails that utilize link(s) leading to a malicious site.	Full text URLs, Hyperlinks
Communication	Phishing emails that prompt the recipient into communicating the desired Target with the malicious actor.	Email Communication, Phone Communication
Attachment	Phishing emails that append malicious files to the mail.	HTML, ZIP, PDF
Calendar Invite	Phishing emails where an event invitation is sent containing undesirable content.	w/malicious links, w/malicious attachments, soliciting information

Table 7.3: Model - Method Categories

7.5 Impersonation

The Impersonation property concerns who the email appears to be from. This is whom or what the malicious actor pretends to be in order to achieve a successful phish. Relevant email elements to collect consist of header FROM field, envelope MAIL FROM field, header SUBJECT field, Message Body, and any Attachment Info.

The categories of Generic and Non-Generic have been identified, where the Generic category consists of External and Internal, and the Non-Generic category consists of legitimate brands/entities. Table 7.4 presents an overview of the Impersonation categories with corresponding description and examples.

Impersonation Category	Description	Examples
Generic	External	Phishing emails where the malicious actor pretends to be an external entity not tied to any specific brand.
	Internal	Phishing emails where the malicious actor pretends to be an entity within the organization.
Non-Generic	Phishing emails where the malicious actor pretends to be a specific brand or entity.	Nigerian Prince, Widower, Lawyer Manager, HR, IT, Yourself Microsoft, Apple, Police

Table 7.4: Model - Impersonation Categories

7.6 Model Overview

Table 7.5 displays the resulting model, consisting of the collection properties, email elements to collect, and the properties' associated categories, while Figure 7.3 provides an example of the model as based on the model overview presented in Figure 7.1.

Property	Description	Email Elements	Categories
Content	The Content property concerns the essence of the phishing email. It revolves around how the malicious actor convinces the recipient of the email's legitimacy.	FROM SUBJECT Message Body Attachment Info	Invoice Document Shared CEO Scam - Transfer Money Post Package CEO Scam - Gift Card ... (See Appendix B for complete list)
Target	The Target property defines the objective that the malicious actor tries to achieve with the phishing email. This is the objective that is tried lured from the recipient of the mail.	FROM SUBJECT Message Body Attachment Info	Credentials Money Credit Card Details Infect PII Business Information
Method	The Method property concerns the approach utilized within the phishing email by the malicious actor in order to obtain the desired Target.	Message Body Attachment Info	URL Communication Attachment Calendar Invite
Impersonation	The Impersonation property concerns who the email appears to be from. This is whom or what the malicious actor pretends to be.	MAIL FROM FROM SUBJECT Message Body Attachment Info	External Internal Non-Generic
Date	The date in which the phishing email was received.	DATE	-

Table 7.5: Email Phishing Collection Model

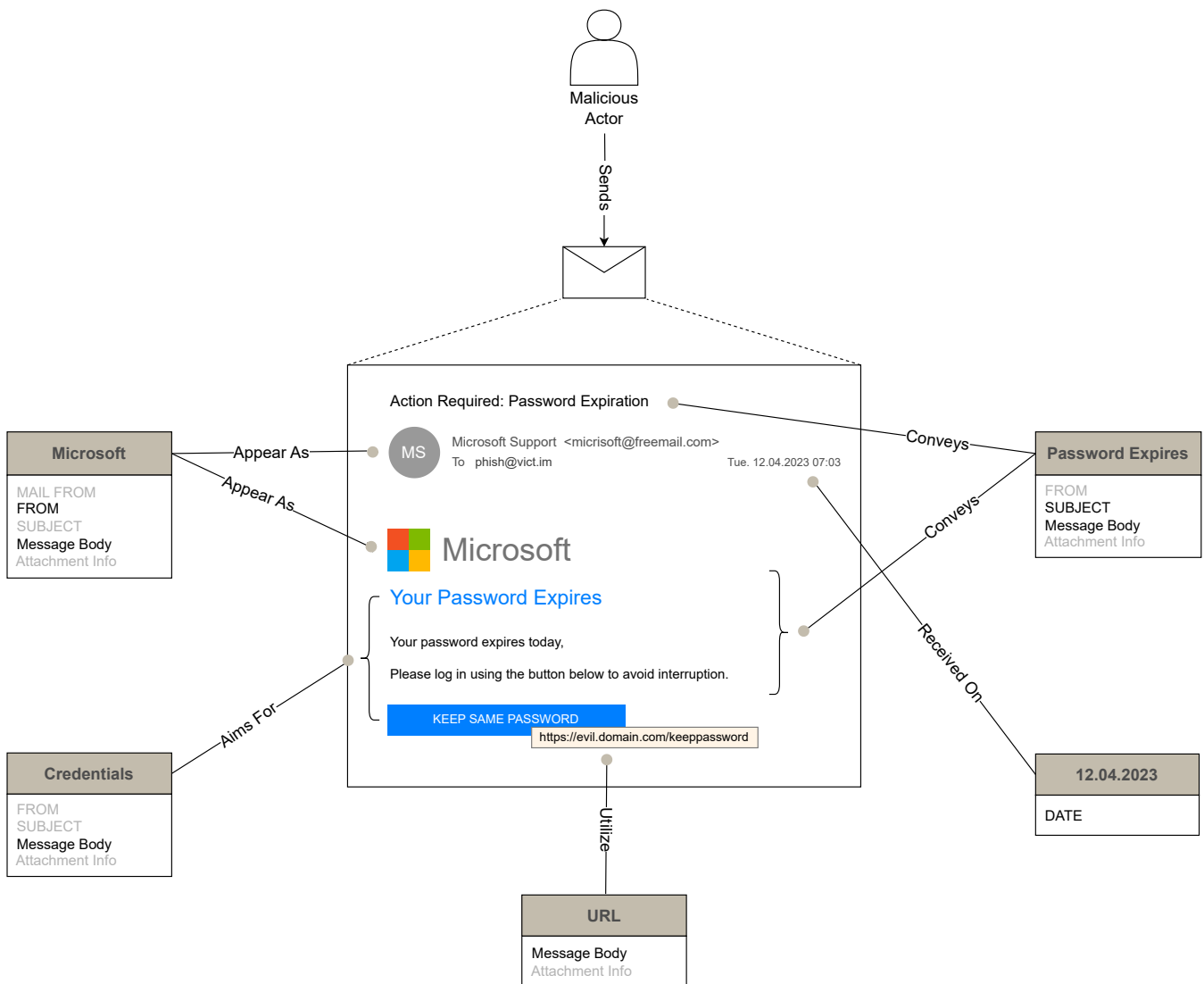


Figure 7.3: Model Exemplified

*A malicious actor sends a phishing email received the **12th of April 2023** aiming for the **Credentials** of the recipient through the utilization of a **URL** by appearing as **Microsoft** and conveying that their **Password Expires**.*

As the categories are based on the analyzed phishing corpus, they are a reflection of the observations made. Because of this, there may exist other property categories not yet identified. This means that although the identified categories should remain static, additions to the categories can be made whenever a phishing email is analyzed where the identified property does not match any of the existing categories.

The model created should be utilized as a basis for any future email phishing studies where trends are to be analyzed. Additional properties can be included if it is of desire to collect and analyze these.

Chapter 8

Discussion

Throughout the analysis of the email phishing dataset in Chapter 6, interesting data and metrics regarding the evolution of phishing emails came to light. The following chapter presents a deeper analysis of a subset of this data in order to identify why certain observations appear as they do. This includes whether any external factors may have had an impact on the identified metrics and whether the observed data can be seen in correspondence with similar research on the given subject.

8.1 Top Content Categories

The collection and analysis of the phishing emails resulted in an array of Content categories and their distribution. The Content categories of Invoice, Document Shared, and CEO Scam types were revealed to be the most represented in the phishing corpus. A question arising from this distribution is why exactly these Content categories are the most represented.

Greene and Steves et al. in their collection of studies on human phishing susceptibility [61–63] details the Invoice phishing type. From their test conducted in [61], having the phishing email appear as an invoice yielded the highest amount of successful phishings, with a 20.5% click-rate. The reasoning for the high susceptibility of this particular phishing approach could mostly be tied to its familiarity. The respondents reasoned their opening of the phishing link/attachment with responses such that it mimics their workplace responsibilities, or that they in fact already were expecting an invoice. Similarly in Williams et al. [64], a respondent's reasoning for opening an invoice phishing email was that it was a part of their work function, and that they currently were experiencing problems with their payment system. The high success rate of the invoice phishing is a factor making it an appealing Content category to utilize.

Looking at the Target-Content relationships from Chapter 6, The Invoice Content category can be seen appearing towards multiple Targets, including Credentials,

Money, Credit Card Details, Infect, and Business Information. The Content category's flexibility in terms of Targets may also serve as an appealing factor, in addition to its success rate, for the utilization of an Invoice as a lure.

As for the Document Shared Content category, a similar reasoning for its popularity can be found as with the Invoice category. Both Sharma et al. [65] and Ho et al. [44] discuss the Document Shared phishing and imply that it is one of the more effective approaches for a successful phishing, where Ho et al. identified that the shared document lure occurred in over 23% of observed lateral phishing* incidents. Sharma et al. argue that the reasoning for its effectiveness is its emotional triggers as it raises curiosity and creates anticipation.

Lastly, the CEO Scam type Content categories differentiates themselves a bit from the two aforementioned categories in terms of structure. From the relationship analysis, both Invoice and Document Shared rely quite a bit on the usage of external websites and crafted documents, while the CEO Scam types are mostly communication based, thus do not require any maintenance of web pages or malicious documents. In the previously conducted study [40], this was argued as a factor for its popularity. The ease of launching the attack can create large influxes of the particular attack regardless of its success rate. However, this is mostly true for the gift card scams, as the transfer scams still require maintenance of the bank accounts utilized for the scam. The fast payout, might be a factor into its popularity, as once the payment is done, the operation is, for the most part, over. While for the Invoice (in the cases of Credential and Infect Targets) and Document Shared, additional steps are required after the victim has been phished, such as using the credentials, sell the credentials, or wait for the infected device to provide desired information.

From the top most observed Content categories, aspects both consisting of success rate and ease of completion can be viewed as reasonings as to why they have such high representation in the analysed phishing dataset.

8.2 Evolution of the CEO Scam

From the analysis of the evolution of the Content categories, there was observed a shift in the usage of the CEO Scam type of phishing attacks, transitioning from wire transfer scams to gift card scams. This shift raises the question why the gift card scam became so prominent while the wire transfer scam declined.

Figure 8.1 displays the percentwise distribution throughout the collected years, showing the evolution of the two CEO Scam types (CEO Scam - Crypto is excluded as it only appeared in 2020).

*Lateral phishing is when an already compromised account is used to send out phishing emails.

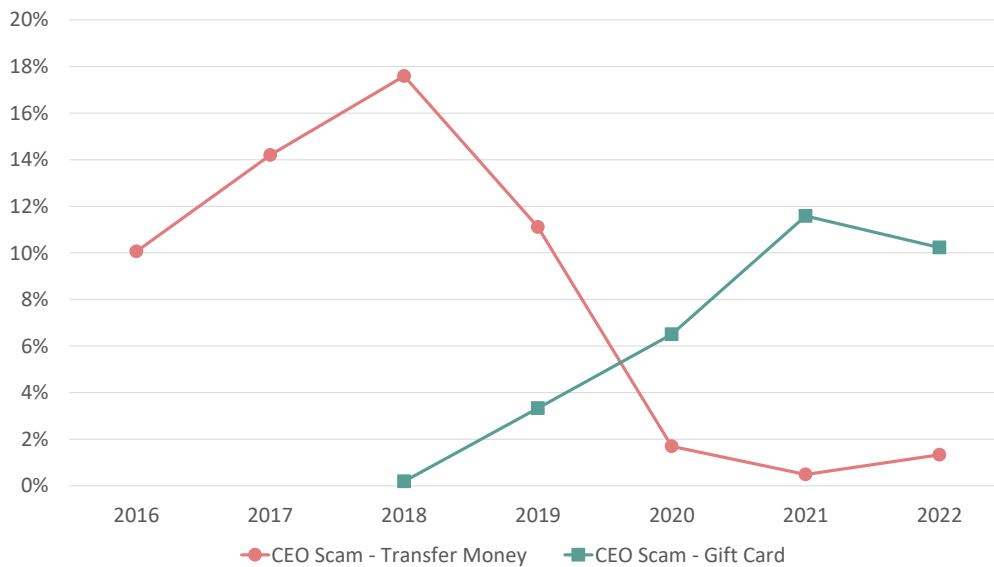


Figure 8.1: Evolution - CEO Scams

When the first appearance of the gift card scam occurred in 2018, the transfer scam was at its peak. The following years shows that as the gift card scam rises, the transfer scam declines immensely. The sudden emergence of the gift card scam can be confirmed present broadly, as the Federal Bureau of Investigation (FBI) reported similar statistics in a Public Service Announcement towards the end of 2018 [66]. The report shows a sharp increase of the gift card scam in the beginning of 2018, from an almost non-existent presence in the better part of 2017.

The gift card scam's popularity has been tied to several factors, including its ease of deployment (as mentioned in the above section detailing the top Content categories) and ease of completion on the victim's side. ProofPoint explains the gift card scam as a quick and easy way for the attackers to get money without the hassle of setting up bank accounts and having the victims navigate through complicated wire transfer instructions [67]. Additionally, as mentioned by Abnormal, the pool of victims is much larger for the gift card scam compared to the transfer scam, as the transfer scam usually is only applicable to employees whose job function involves company finances [68]. Finally, Mangut et al. point out that the popularity of the gift card scam can be tied to its higher success rate compared to the transfer scam, due to the lower amount of money involved [69].

Although these sources make valid points for the gift card scam's popularity, they do not provide insight into why the transfer scam has seen such a great decline. Chaganti et al. draws ties to the traceability of wire transfers compared to gift cards. Money transferred between accounts can be traced, while gift card usage provides greater means of anonymity [70]. The chances and the impact of being

caught may seemingly be a factor contributing to the decline of transfer scams, as stated by Agari [71]. Agari mentions that the increased collaboration between researchers, financial institutions, and law enforcement has made the identification and handling of such scams more impactful.

What is lacking from the identified sources is any indication as to why the decline of the transfer scam seemingly was so abrupt after 2018.

In 2018 the US launched an operation called "Wire Wire" [72], which was a coordinated operation between several US departments, including US Department of Justice, US Department of Homeland Security, US Department of Treasury, and US Postal Inspection Service, with the aim to intercept and identify wire transfer scams. The operation, spanning six months, resulted in 74 arrests and the disruption of USD 14 million fraudulent wire transfers. Operation "reWired" [73] was launched in 2019 as a continuation of operation Wire Wire, leading to the arrest of 281 people involved in these scams and the recovery of USD 118 million from fraudulent transfers. Based on the timing of these operations, they may have been a reason as to why the transfer scams saw a decline after 2018.

Further research into the subject matter needs to be conducted in order to confirm or deny this theory, and to identify any additional causes, should they exist.

8.3 Credentials

As shown in the analysis of the collected phishing emails, Credentials were by far the most sought after Target throughout all the collection years. This statistic mirrors the reportings of the surveyed reports [14, 15, 49, 51], making it clear that Credentials are widely the Target that is the most desired. The reasoning why this is is however not directly evident from the statistics alone.

Credentials themselves do not provide any direct gain, as opposed to a successful wire transfer phishing. It is therefore relevant to look at what the credentials can be used for in order to determine why they are so sought after. Credentials can be utilized for a magnitude of purposes including launching of ransomware, data theft, identity theft, fraud/transactions, extortion, lateral movement, persistence, or simply selling the credentials for monetary gains [44, 74, 75]. The variety of opportunities and use cases that credentials present provides a plausible reason as to why it has such a high representation in the phishing corpus.

Tying into Credential's array of usages is its observed diversity in the context of Content categories. From the relationship between Target and Content, Credentials were observed utilizing a great variety of Content approaches, with an overall diversity score of 0.91. Its diversity showcases the valid scenarios for credential phishing, which again can be viewed as a reason for its popularity.

Further, the detectability of stolen credential usage is argued as a contributing reason for its high presence. Cofense states that as the credentials are legitimate, they provide adversaries with access without necessarily setting off security alarms [76]. In addition, it is mentioned that credential attacks leave less indicators of compromise (IOCs) behind compared to an infection for instance, making investigations more difficult.

Palo Alto Networks describe credential theft as a cheap and extremely efficient tactic for breaching organizations as it mostly relies on human interaction while malware and other exploits are more reliant on weaknesses in security defenses [77].

Although the latter two statements do not provide any backing studies or research into the facts, the combination of Credentials' flexibility, diversity, and assumed efficiency and low detectability provides logical reasonings as to why Credentials are a highly sought after Target.

8.4 Evolution of Infect

The observation of phishing emails aiming to infect the recipient's device can be seen increasing noticeably in 2017 before declining the following years, as shown in Figure 8.2. In terms of ranking compared to the other Targets, it went from being the second most observed Target in 2017 to now being surpassed by both Money and Credit Card Details. This evolution warrants further investigation into why this surge appeared in 2017, and why it has been declining since.

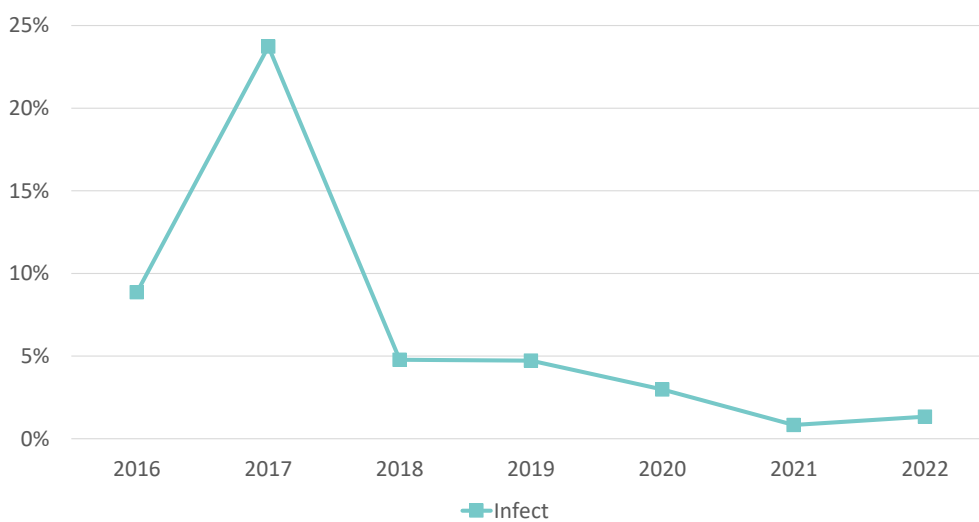


Figure 8.2: Evolution - Infect

Although external sources also observe a decrease in phishing emails containing malware [78, 79], few debates why this is. Health IT Security mentions what was pointed out in the section above regarding credentials, that the detectability of credential theft compared to other methods of breach, such as malware, is a reason why exactly malware is not as popular [80]. Further, Help Net Security state that the improvement of security systems and detection methods has had a great impact on the utilization and observation of malware in general [81]. Although this can explain the overall decline in the observation of infect Target phishing emails, they do not provide any insight into the surge of 2017.

Viewing the malware trends of 2017, ransomware was the most prominent of that year, with the WannaCry ransomware in the forefront [82]. Although WannaCry had a huge presence in 2017, it was not a malware distributed through phishing, cutting any ties to the surge observed in the phishing corpus of that year. There were other ransomware types like the re-emergence of Locky [83] which did utilize phishing, however it was not as prominent as some of the other phishing ransomware types observed in 2016, such as the original Locky [84].

Statistics from SonicWall [85] shows that malware overall in fact grew in 2018 compared to 2017, while email still was reported as the main mode of delivery of malware and ransomware [84, 86], contradicting the observations made in this phishing dataset. However, location specific reports, such as numbers from Ireland [81], shows that malware overall saw a reduction in 2018 compared to 2017. This shows that these numbers are varying from location to location, and that the observations made in this thesis' dataset are specific to the scope of Scandinavian countries.

Based on the information provided, the overall reduction in Infect as a Target can be argued attributed to its overall noisiness and detectability. On the other hand, the surge seen in 2017 could not be tied to any specific events, and is seemingly not a broad trend. A more in-depth study on these statistics is necessary in order to determine whether there exist any data not yet identified tying the surge observation to any external events.

8.5 URL

Utilizing URLs as a method of achievement is by far the most preferred approach in the observed phishing emails. The Method's dominance raises questions as to why URLs are so highly favored compared to the other identified Methods.

Cofense has published a write-up [87] on this observation, though only concerning URLs and attachments, detailing why URLs are a preferred method of achievement. In the write-up, the URL's ability to bypass email filters is brought forth as a contributing factor due to the difficulties of determining a URL's legitimacy. Usage

of legitimate and trusted hosting sites and redirects are factors making it hard for email filters to categorize a URL. On the contrary, attachments are generally met with more suspicion, both by the email filters themselves and the receiving user.

By looking at the relationship between Method and Target, Figure 8.3, one can identify that Credentials heavily favors the usage of URLs as a method of achievement, and as Credentials are by far the most sought after Target it is logical that URL would appear as the top Method. In addition, URLs are identified as a flexible Method, being utilized to achieve Credentials, Money, PII, Credit Card Details, and Infect, where there is no dominating Target as opposed to Communication where Money accounts for 95.4% of the observed mails.

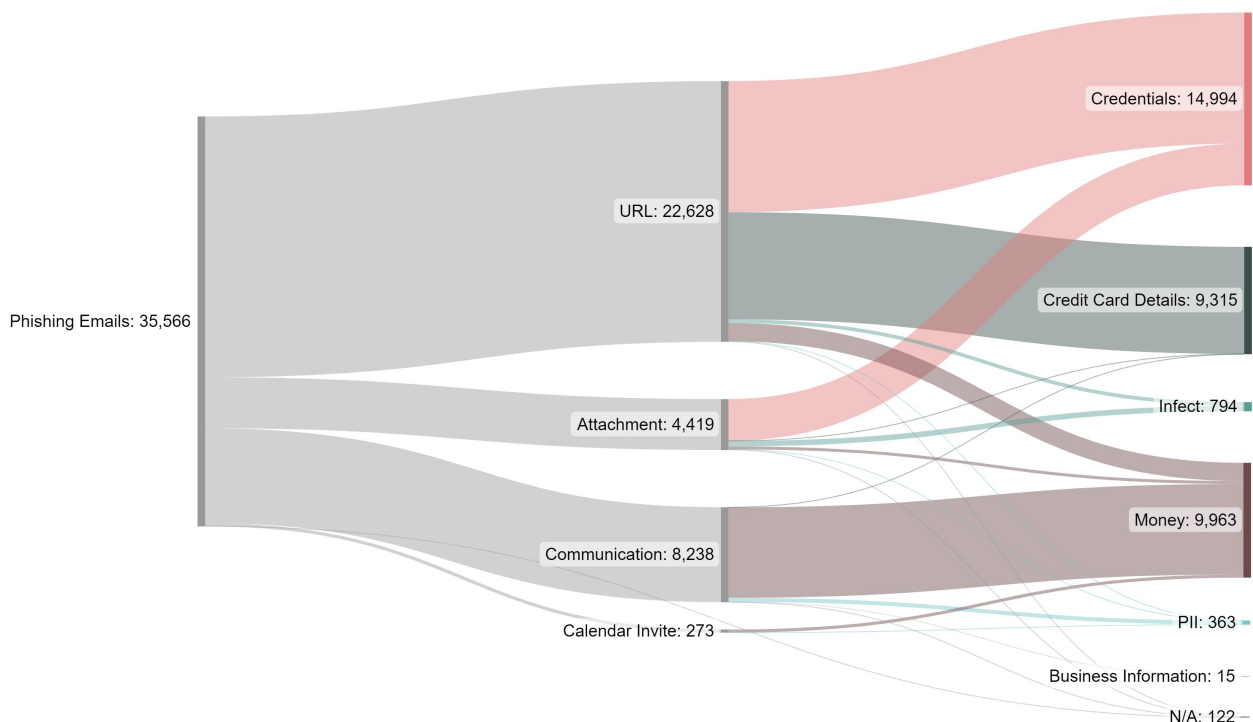


Figure 8.3: Method-Target

Although URLs have been the most utilized method of achievement within the scoped years, recent development shows that Attachment is on the rise. This increase can be attributed to the increased usage of HTML attachment as a Method in credential phishing. Even though URLs seemingly is favored compared to attachments, development in spam filters' capabilities to detect URLs has led to the HTML Method being adapted [49, 88].

The previously conducted study [40], discussed the usage of HTML attachments. In the study, there was identified two distinct approaches in the usage of HTML attachments: One where the attachment redirects to a website when opened, as shown in Figure 8.4, and one where the attachment is embedded with a web page

and is hosted on the device it is opened on, as shown in Figure 8.5 (both figures being extracted from emails within the thesis' phishing corpus). With the latter, any information typed on the site is sent to an external point based on the details specified in the HTML code. Both approaches bypasses any URL checks, while the latter also prevents the user from receiving security warnings on the opened phishing site.

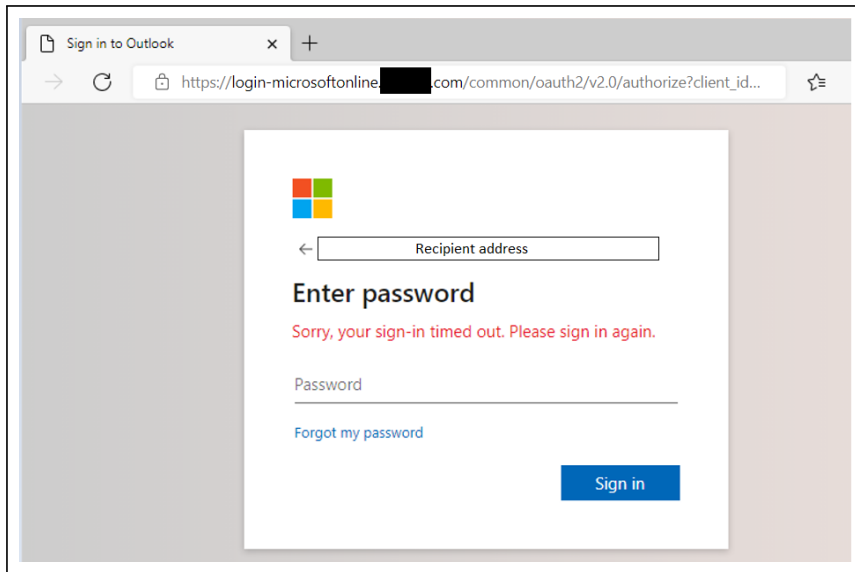


Figure 8.4: HTML - Website

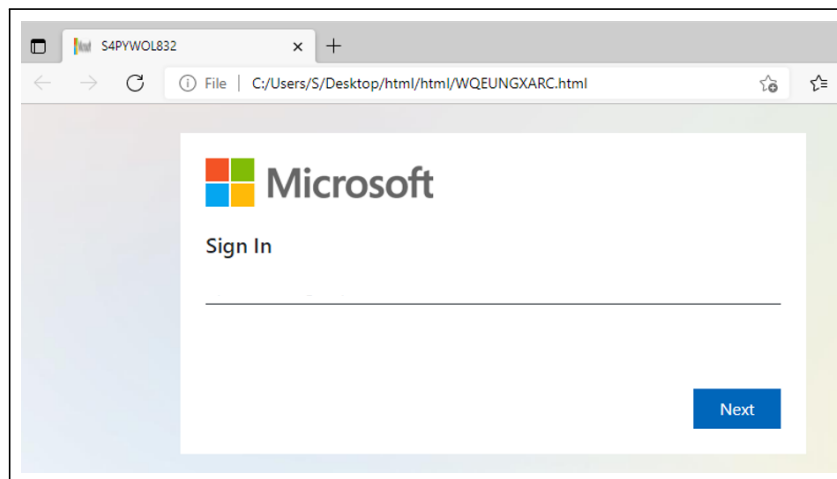


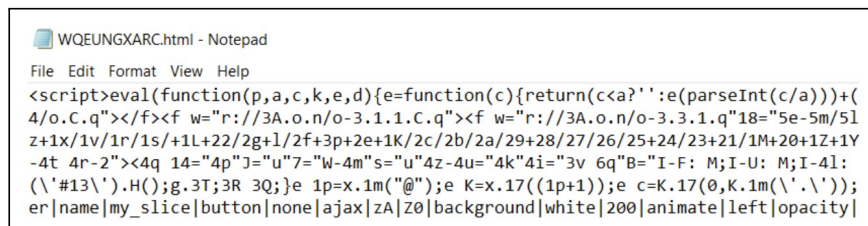
Figure 8.5: HTML - Local

In addition to the aforementioned evasive techniques, HTML files can also be obfuscated, creating another layer of evasion. Instead of having the linked site or page layout in plain text, as shown in Figure 8.6 from an email from the phishing corpus, the text can be encoded in order to bypass email filters looking for certain patterns or signatures in the text of the file. Figure 8.7 presents an example of such an encoded HTML file, again extracted from the thesis' phishing corpus.



```
invoice.html - Notepad
File Edit Format View Help
<script type="text/JavaScript">
    setTimeout("location.href = 'https://login-microsoftonline.████████.com/?';",0);
</script>
```

Figure 8.6: HTML - Plain Text



```
WQEUNGXARC.html - Notepad
File Edit Format View Help
<script>eval(function(p,a,c,k,e,d){e=function(c){return(c<a?'':e(parseInt(c/a)))+(4/o.C.q"></f><f w="r://3A.o.n/o-3.1.1.C.q"><f w="r://3A.o.n/o-3.3.1.q"18="5e-5m/5l z+1x/1v/1r/1s/+1L+22/2g+l/2f+3p+2e+1K/2c/2b/2a/29+28/27/26/25+24/23+21/1M+20+1Z+1Y -4t 4r-2"><4q 14="4p"j="u"7="W-4m"s="u"4z-4u="4k"4i="3v 6q"B="I-F: M;I-U: M;I-4l: (\'#13\').H();g.3T;3R 3Q;}e 1p=x.1m("@");e K=x.17((1p+1));e c=K.17(0,K.1m('\.\''));er|name|my_slice|button|none|ajax|zA|Z0|background|white|200|animate|left|opacity|
```

Figure 8.7: HTML - Obfuscated

The popularity of URL as a method of achievement can be attributed to the difficulties in detecting and differentiating legitimate sites from malicious ones, as well as the Method's flexibility when it comes to Targets. Although the URL Method has been dominating in the years within the collection scope, advancements in URL detection has paved the way for the utilization of HTML attachments for similar purposes, providing further evasive opportunities.

8.6 Targeted Brands

From the analysis of impersonated brands there was no particular pattern throughout the collected years, where both the representation and instances of impersonations varied. However there were some brands that distinguished themselves as the overall most impersonated, that being Microsoft, Apple, and Posten.

Both Microsoft and Apple has been seen throughout the years being highly present, appearing as the two most represented brands overall. Viewing the statistics from Zscaler, ProofPoint, SalshNext, and Check Point [14, 50, 51, 89–96] this representation is present broadly, and not just within this thesis' dataset. As for Microsoft, its high impersonation rate can be attributed to their broad utilization in small to

medium sized businesses as stated in [97]. By compromising a Microsoft account, the malicious actor could gain valuable insight into the victim's business, as well as access to the services which is tied to the user, such as email services for distributing of additional phishing or cloud resources for the launching of ransomware. Depending on the structure of the organization and the access rights of the victim, an attacker could potentially take control of an organization's infrastructure with the compromise of a Microsoft account [98].

Although Apple generally do not provide the same platform services as Microsoft, they do have a large user base surpassing 2 billion active devices as of 2023 [99], providing reasonings as to why they too are such a highly impersonated brand. Compromising a Microsoft account may give access to business specific resources, while compromising an Apple account gives access to all data and services tied to it, such as payment information, stored passwords, personal data, and purchase capabilities. Both the increased pool of potential victims due to their large user base and magnitude of Apple product and services [100] makes the brand a lucrative entity to impersonate.

In inclusion to the points stated above, general familiarity with the sender or origin of the mail is an important factor in the effectiveness of phishing [101–105]. The massive user base and areas of usage for both Microsoft and Apple products and services increases their recognizability and in turn familiarity with the users. This makes the brands not only viable to use for credential phishing, as discussed in the paragraphs above, but other types as well, such as for direct monetary gains.

An interesting observation from the brand impersonation evolution is that Apple has seen a reduction in representation the recent years, as shown in Figure 8.8.

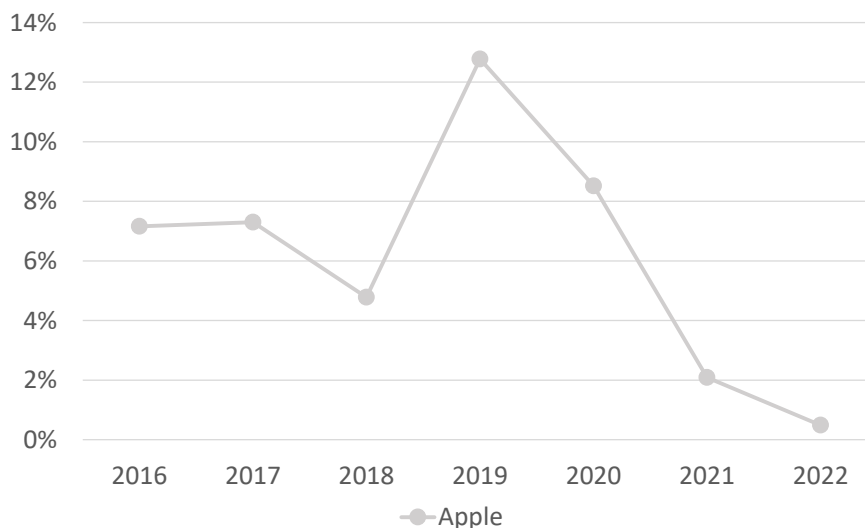


Figure 8.8: Evolution - Apple

Looking at the statistics from other sources, in this case Check Point's "Top Phishing Brands" statistics [89–96], this is seemingly a trend not only specific to this thesis' dataset, although no information as to why this is has been identified. As newly published news articles still describe Apple phishings as heavily present [106–108], an assumption that can be made is that the usage of Apple in phishing is not necessarily decreasing in the long run, just temporarily seeing a dip in presence compared to other utilized brands. Should the decrease in fact continue in the coming years, a more in-depth analysis needs to be performed in order to identify any root causes.

Further, the brand of Posten is seen as the third most impersonated brand overall. However looking at the yearly distribution, it can be seen that its presence only recently became prominent, appearing first in 2019 with a significant increase in 2020. This particular brand impersonation greatly reflects the user base in which the phishing emails have been collected from, as Posten is a Norwegian based organization. Because of this, one cannot expect the brand to appear on other global lists, however one can find its equivalent by identifying businesses providing similar services. Viewing the observed brand impersonation evolution as reported by Check Point [89–96], a similar evolution can be seen with DHL who also provides package delivery services. The brand appeared abruptly in 2020 on the top 10 most impersonated brands and has continued showing its presence ever since.

In 2020, the COVID-19 pandemic caused mass lockdowns and a shift to remote operations. This led to an increase in the conducting of online shopping with the utilization of delivery services in order to receive said order [109]. As pointed out by Alawida et al. in [110], the increased utilization of delivery services was taken advantage of by malicious actors in order to send out fraudulent messages. Security firm Kaspersky further detailed these types of delivery scams in the early stages of the pandemic [111], with others warning about the emergence of this particular type of scam [112–114].

COVID-19 appears as the kickstarter for the delivery phishing attacks providing insight into why Posten saw such a prominent increase in 2020.

8.7 Spoofing

Spoofing is a technique within the Impersonation property where the malicious actor forges the sender address in order to appear legitimate. In many instances of impersonation, the malicious actor utilize addresses that looks like a legitimate one, such as support@micri_isoft.com. Spoofing takes the impersonation one step further by having the sender address appear as the correct one, having the FROM display support@microsoft.com in this instance. A part of the Internal Impersonation category consisted of emails that had a spoofed internal address, meaning that the email appeared as coming from a legitimate address within the company.

As touched upon in Section 8.6, various studies [101–105] on email phishing point out that appearing to come from a source familiar to the recipient is an important factor in the effectiveness of a phishing email. By spoofing the sender address to appear as a familiar one increases the perception of sender legitimacy and in turn can assist in heightening the chances of a successful phish. As pointed out by the aforementioned studies, there are of course other cues and properties that influence the overall perception of a phishing email as well, however, the perception of sender legitimacy is regarded as an important influencing factor for a successful phishing.

Spoofing an email address can be achieved by manipulating the message header fields (Chapter 5). Once an email is generated, the FROM, REPLY-TO, and RETURN-PATH can be changed in order to appear as a different sender.

During the data collection, information on whether a phishing email coming from an internal address was spoofed or not could be determined based on the FROM, REPLY-TO, RETURN-PATH, and AUTHENTICATION-RESULTS header fields. Table 8.1 showcases the amount and the percentwise distribution of spoofed emails observed for each year.

Year	Total	% of Total
2016	66	5.63%
2017	60	6.09%
2018	21	4.02%
2019	18	2.50%
2020	34	1.07%
2021	41	0.25%
2022	48	0.38%

Table 8.1: Spoofing Distribution

As visualized in Figure 8.9, there is a clear trend of declining observations of spoofed phishing emails. The observation of spoofed emails had a slight increase in 2017, followed by a sharp decline the following years, until a small increase in 2022 again. From 2016 throughout 2022, the percentwise observation of spoofed emails shrunk by 85.26%

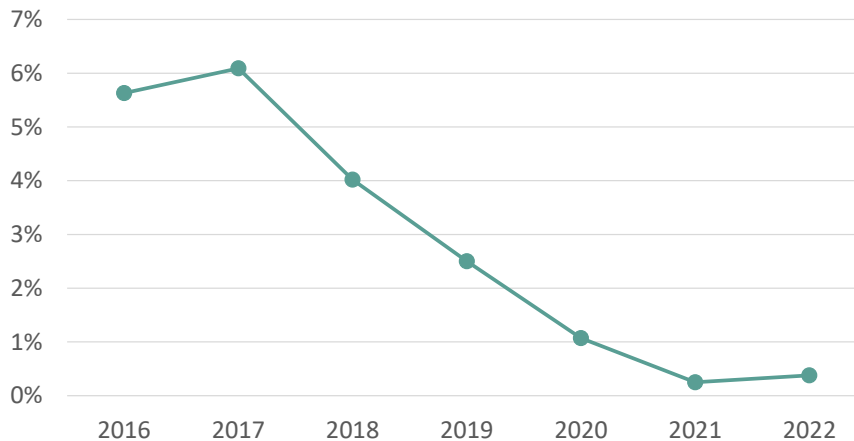


Figure 8.9: Spoofing Distribution

Spoofing is made possible due to the lack of authentication requirements in the SMTP protocol, leaving it vulnerable for exploitation. There does however exist several technologies and methods to detect and prevent such exploitation.

SPF

Sender Policy Framework (SPF) [115] is an email authentication method utilizing a list of allowed IP addresses in order to validate the source of the mail. By utilizing SPF, a list of approved sender IP addresses is published for the given domain. Whenever a system receives an email, it checks for a published SPF record and if published, it iterates through the list of approved addresses in order to determine the validity of the email source. A failed SPF authentication usually results in the email bouncing or being sent to the recipient's spam folder (this is determined by how the recipient's email service is configured).

A downside of SPF in regard to spoofing is that it conducts the check on the envelope from address (MAIL FROM), but no checks are made on the header from address (FROM) which is the address most recipients are exposed to when viewing an email. There is also no way for the administrator of a domain to decide what should be done if an email fails the SPF check for their domain, only the recipient can determine what will happen if a SPF check fails.

DKIM

Domain Key Identified Mail (DKIM) [116] is another email authentication method utilizing a digital signature for email validation. A public key is added to the DNS record of the domain, and can be retrieved by anyone. When an email is sent, a signature consisting of message body and/or header information encrypted by the private key is added to the message header, which can be validated by the recipient by decrypting it with the domain's public key.

As opposed to SPF, DKIM allows for validation of the header from address (FROM), but DKIM does not allow for the email administrator to determine what should happen if a DKIM check fails for their domain either.

DMARC

Domain-based Message Authentication, Reporting & Conformance [117] is an email authentication implementation that builds upon SPF and DKIM in order to provide the email administrator both with auditing and options on what to do with an email if the validation of an email supposedly from their domain fails. Whenever an email "from" a specific domain with DMARC fails, three actions can be performed, including None (nothing is done), Quarantine (the email is sent to the recipient's junk folder), Reject (the email is bounced).

In order to determine if these changes in the utilization of spoofing is in any way related to changes in the utilization of spoofing preventative measures, statistics for the utilization of preventative measures needs to be collected. As DMARC utilizes both the SPF and DKIM technologies and provides specifics on what to do with mails that fail validation, the utilization of this technology will be focused on in the following section.

There are sites that track the usage of DMARC through public sources, such as BuildWith¹ and dmarc.com². dmarc.com, which uses DomainTools³ to track the usage of DMARC on the internet, provides concrete numbers on the utilization of DMARC for the years within the collection scope. Table 8.2 and the corresponding graph in Figure 8.10, displays the total number of valid DMARC policies from 2016 to 2022. The data is taken from the twelfth month of each year excluding 2022 where it was only available for the sixth month.

Year	Total
2016	80,275
2017	240,151
2018	630,000
2019	1,892,227
2020	2,221,962
2021	4,974,390
2022	5,566,779

Table 8.2: DMARC Trends from dmarc.com

¹BuildWith: <https://trends.builtwith.com/mx/DMARC>

²dmarc.com: <https://dmarc.org/stats/dmarc/>

³DomainTools: <https://www.domaintools.com/>

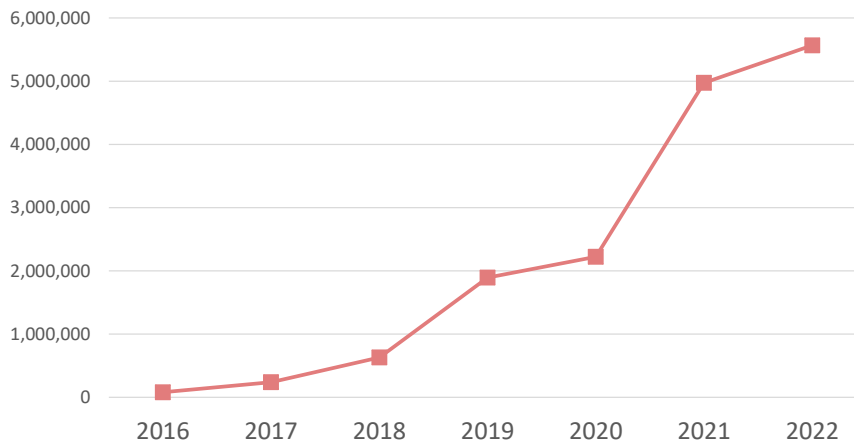


Figure 8.10: DMARC Trends from dmarc.com

Comparing the trends in the utilization of DMARC and the observation of spoofing in the phishing emails, there is a similar incline as decline within the two graphs, as visualized more clearly by overlapping the two graphs in Figure 8.11. The decline in the usage of spoofing starts to ramp up in 2018, at the same time as the usage of DMARC becomes more prevalent. The decline however, is much steeper than the utilization of DMARC for that year. A similar display can be seen in 2021 where the greatest incline in the utilization of DMARC occurs, while a much smaller decline is observed within the spoofing graph.

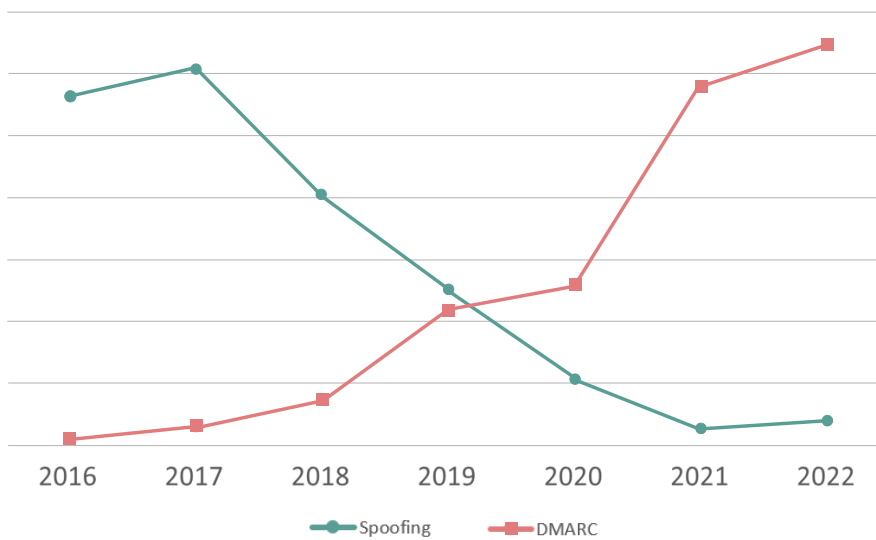


Figure 8.11: Overlapping Spoofing and DMARC evolution

The comparison of DMARC utilization and observed spoofing shows no direct relationship, although they appear to have similar opposing evolutions. As the

DMARC statistics are not based on numbers from the organizations in which the phishing emails are reported from, no clear conclusions can be made whether the decline in spoofing is related to the implementation of DMARC or if there are any other external factors causing the decline. Further research should be conducted to determine if the usage of DMARC in fact can be viewed as a factor in spoofing's decline.

8.8 Date Activity

The distribution of dates presents an insightful overview of the phishing activity within and throughout the years, allowing for the analysis of patterns observed both regarding the overall phishing activity and specific phishing contents. Based on prior literature, the following section compares the observation of this phishing dataset with the findings from that of others.

M. R. Riedle in [16], presented in the related work chapter (Chapter 4.2), investigated among other metrics, the distribution of phishing attacks within the years of their thesis' collection scope. The findings were that there was observed increased activity in the months of November to January, as well as in the months of March to May.

Viewing the date distribution presented in Chapter 6.2.5, there was identified a pattern of heightened activity in November and early December, showing consistency between M. R. Riedle's findings and the observations made in the current dataset. M. R. Riedle proposes that the holiday season in this time frame when people are busy with online purchases and deliveries is abused by phishers to launch their attacks. This is also on par with the findings and reasoning made by Oest et al. in [75] analyzing phishing campaigns, as well as stated by Bitaab et al. in [41], and observed by Ramzan et al. in [118].

As for the contents of these phishing mails sent out in November-December, there does not appear to be any one particular group of mails that are utilized for these months. The phishers seemingly utilize the fact that people are pre-occupied and less observant [118] during these times to launch any type of phishing attack, almost regardless of content.

As for the increased activity in March to May, this was not observed in the current thesis' dataset. M. R. Riedle tied the heightened activity to the ongoing tax season in the US for those particular months. The season for taxes in Scandinavian countries overlaps with that of the US [119–121], however does not appear to have an impact on the phishing activity in the same time frame. M. R. Reidle is not alone with these observations, as other such as Yeoh et al. [122] and A. Chaudhuri [123] have also pointed out similar remarks.

Although there isn't observed any increased activity in the tax period from this thesis' dataset, from the data collected for the Impersonation metrics there is evidence that taxing entities have been abused in phishing emails, including Skatteetaten and Skat.dk. Filtering the complete dataset on these two impersonations, displayed as entries per day in the heat-map in Figure 8.12, there is actually an influx of tax related mails in the middle of December and surges in November, with a rather low representation in the tax season months.

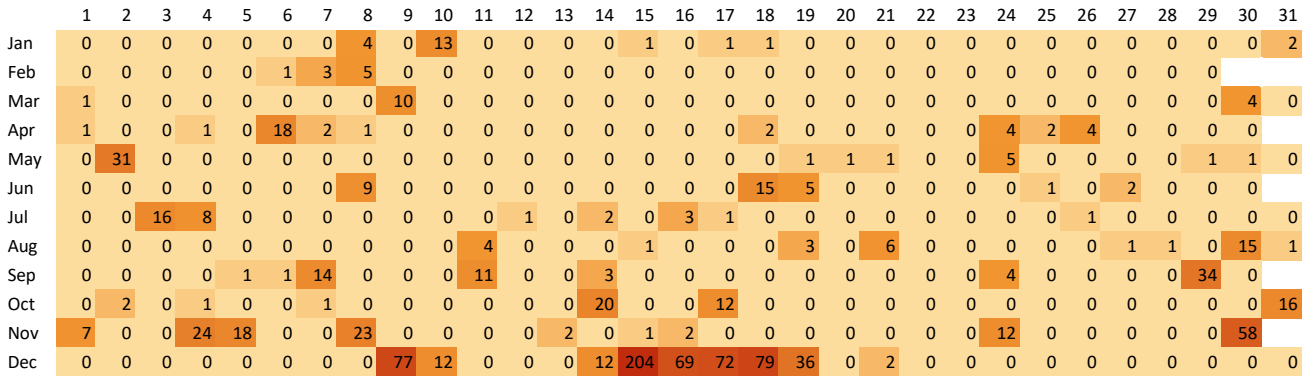


Figure 8.12: Tax Phishing - Heat-Map

In December, the tax card for the next year is sent out, and there are also a majority of organizations who have half-tax either November or December. Filtering the dataset to include the corresponding Content categories for these mails shows that close to all of them are related to tax returns, not corresponding to any of the actual tax related events for these months. As no other studies have been identified reporting similar statistics, these observations may be unique to this thesis' dataset and not something that should be considered a universal trend.

Ramzan et al. made the observation that phishing activity declines considerably on the weekends compared to the rest of the week [118]. In the previously conducted study [40], a similar analysis was conducted on a set of Content specific phishing mails, consisting of Content categories similar to CEO Scam - Gift Card, Password Expires, Update Account Information, and Refund. From that study it was identified increased activity in the mid-weekdays Tuesday, Wednesday, and Thursday, however as it was scoped down to specific contents it may not be representative of the whole dataset.

Converting the collected phishing emails' dates into weekdays, the weekly distribution for the phishing corpus can be seen. Figure 8.13 presents this overall week distribution.

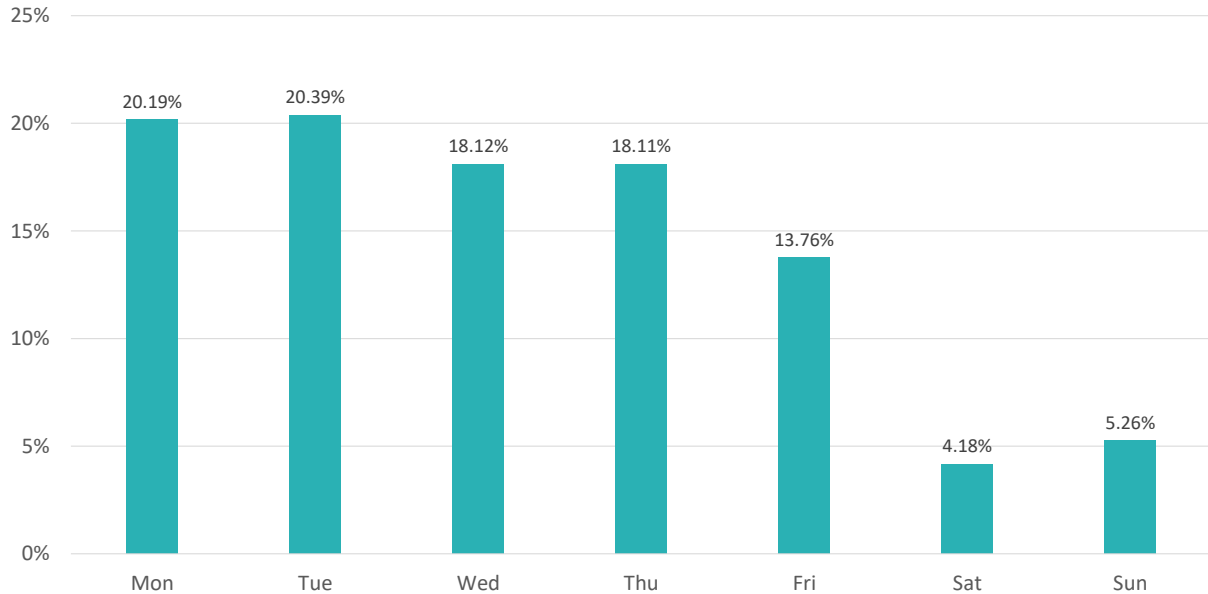


Figure 8.13: Week Distribution

From the distribution of days, it is apparent that there is heightened activity in the beginning of the week, with Tuesday having the highest representation, slightly above Monday. The activity declines approaching Saturday where the activity is at its lowest, corresponding well to the observations of Ramzan et al. [118].

As the phishing emails making up the thesis' dataset is based on reported mails, it could present a challenge to the validity of these date distribution numbers. Even though the date is based on when the email was received by the recipient and not when the email was reported, a question arises concerning the reporting habits of the recipients. Points such as whether people are less susceptible to report mails when they were received days before, as with emails received on the weekends, may have an impact on the numbers presented. Because of this, in order to appropriately determine whether there is actually less phishing activity on the weekends, either the recipients' reporting habits have to be surveyed, or additional data sources, such as spam filters, have to be analyzed.

8.9 Summary of Discussion

The deeper analysis of the observations made in Chapter 6, provides discussion points and further reasonings as to why the observed evolution and collected data appears the way it does. The subjects related to the Content property suggests that the top observed Content categories are utilized to such an extent due to aspects consisting both of their success rate and ease of completion. Additionally, the CEO Scam's shift onto Gift Cards from Transfer scams can be put in relation to both

lack of traceability and easy payout, as well as initiatives from government entities revolving around shutting down fraudulent wire transfer activities. Although, the latter is in need of further research before any direct conclusions can be made.

For the subjects related to the Target property, the Credentials Target was determined heavily sought after due to a combination of the Target's flexibility, diversity, and assumed efficiency and low detectability. Unlike Credentials' low detectability, the decline of Infect as a Target was argued attributed to its overall noisiness and detectability, while the reasoning as to why there was such a surge in Infect activity in 2017 could not be determined.

The Method related discussions concluded that the URL's popularity as a method of achievement could be tied to the difficulties in detecting and differentiating legitimate sites from malicious ones, as well as the Method's overall flexibility in regard to the Target property. Although URLs are deemed hard to detect, due to advancements in detection capabilities, malicious actors have begun utilizing HTML attachments for similar purposes in order to evade these URL detection mechanisms.

Subjects related to the Impersonation property suggest that the two most impersonated brands, Microsoft and Apple, are heavily utilized due to their large user bases and variety of products and services, while the third most impersonated brand, Posten, has seen its upsurge due to the COVID-19 pandemic. Microsoft's high impersonation rate can be attributed to its ties to business use as well, as a compromised Microsoft based account can provide access to business related services and infrastructure. Although Apple overall appears as highly impersonated, it has seen a decline with no identified reasoning as to why. Further analysis is required in order to identify this decline, should it continue downwards. Similarly, although the utilization of the preventative DMARC technology closely mirrors the evolution of spoofing, the decrease in the utilization of the spoofing tactic could not be directly tied to any specific factor with the selected data.

Lastly, the increased phishing activity observed in November and early December can be tied to the holiday season when people are pre-occupied with other tasks, which is being abused by malicious actors when launching their phishing attacks. Further dividing the date activity into weeks, reveals that there is increased activity in the beginning of the week and lower activity rates during the weekends. Due to the nature in which the phishing data is collected, these statistics may be influenced by the habits of the recipients, requiring further research in order to determine the validity of the data and any corresponding reasonings.

Chapter 9

Conclusion

In this thesis, a collection and subsequent analysis of email phishing trends for the years 2016 throughout 2022 has been conducted. Additionally, in order to perform this collection and analysis, an email phishing collection model was created on the basis of prior literature and the contents of the email phishing corpus itself.

The Email Phishing Collection Model developed draws its structure from previously conducted work and scientific literature, and is further expanded based on the thesis' email phishing corpus. Five collection properties were defined for the model, consisting of Content, Target, Method, Impersonation, and Date. These collection properties make up the essence of the thesis' definition of phishing, and are collected based on identified corresponding email elements from RFC5321 and RFC5322. To create consistency and comparability, each of the identified collection properties were populated by accompanying categories influenced by the thesis' email phishing corpus.

The model produced provides a baseline and a universal approach for the collection of phishing emails. By utilizing this model, any result may be compared to other findings where the model is also applied, creating a unified view of the subject matter regardless of when and where the phishing data is from. This opens the door for a more effecivized process of collection, paving the way for a more standarized way of conducting email phishing analysis.

The collection of phishing emails, based on the Email Phishing Collection Model, from the years 2016 throughout 2022 yielded a resulting dataset consisting of 35566 phishing emails. The subsequent analysis highlighted trends and changes observed in the scoped years, showing exactly how email phishing has evolved throughout the years. These observations also provided insight into any patterns within the phishing corpus, which in turn could help reason any future potential phishing behavior.

Further analysing the findings from the dataset analysis, the reasonings as to why the observations appear as they do were attempted identified, including whether the changes observed could be tied to any external factors. By comparing the findings of this thesis' dataset with that of external sources, possible correlating factors, and lack thereof, came to light. It was shown that most of the observations analysed could to some degree be reasoned with by drawing insight from external sources, while other observations were seemingly specific to this dataset alone.

9.1 Research Questions

Throughout the completion of the collection, analysis, and model creation, the three research questions set for the thesis have been answered.

9.1.1 RQ1

"How have the phishing trends changed in the recent years?"

From the analysis of the phishing dataset, the evolution of phishing trends on the basis of the model properties have been identified. It is observed that approaches tied to the Target and Method property popular in 2016 have remained heavily utilized all the way through 2022, where both the Credentials Target and the URL Method have consistently been the most utilized within their respective property. On a similar note, the non-generic Impersonation categories of External and Internal have made up a large portion of the Impersonation observations in each and every year, while the generic categories, on the other hand, have seen a fair deal of variation.

As for the Content property, it has been observed with instances of both consistency, such as with the Invoice and Document Shared categories, and erratic behavior, such as with Confirm / Update Account Information and Refund categories. The analysis also showcased the arisal and decent of Content approaches, with the emergence of the Gift Card CEO scam in 2018/2019, the decline of the Transfer CEO scam in 2018/2019, and the abrupt escalation of the Post Package category in 2020, displaying how the trends within the Content category have been changing in the years within the collection scope.

Lastly, the distribution of dates shows that there is generally no distinct pattern with the occurrences of phishing emails throughout the years, although a slight increase in activity has been observed in the months of November and early December within the collection years.

9.1.2 RQ2

"Are there any correlation between changes observed in phishing trends and external factors, and if so, what are they?"

Through the in-depth analysis of the phishing statistics, external factors including the COVID-19 pandemic, governmental operations, the Christmas season, and advances in detection technologies have been identified as having had an impact on the trends observed.

Improvement of detection mechanisms is identified as a contributing factor as to why the Infect Target category has seen such a decline the recent years, while also tying into why the HTML attachment type from the Attachment Method category has grown as an alternative to URLs.

The abrupt escalation of the Post Package Content category seen in 2020, with a corresponding increase in the impersonation of the Posten brand, can be tied to the COVID-19 pandemic when there was a great surge in the conducting of online shopping. Online shopping, with the addition of other occupying tasks, in the Christmas holiday season is also highlighted as a reason why there is observed heightened phishing activity in the months of November and early December.

The sudden decrease in the utilization of the Transfer CEO scam approach in 2018/2019 corresponds with the "Wire Wire" and "reWired" operations launched in 2018 and 2019 respectively, with the aim of intercepting and identifying wire transfer scams. Relating to the factor of tractability with the transfer scams, it can be viewed as an important factor linked to the emergence and increased preference of the Gift Card CEO scam as opposed to the Transfer approach.

The downwards evolution of the spoofing technique is seen closely mirroring the utilization of the preventative DMARC technology. However, as the change in spoofing is observed occurring before the increase in DMARC usage, no conclusions could be made based on the surveyed datasets.

9.1.3 RQ3

"How can a universally applicable email phishing collection model be created?"

The creation of the Email Phishing Collection Model shows that by consulting prior literature and basing the model's properties on a general definition of phishing and the structure of an email, as well as populating them with data from real phishing emails from multiple sources, a universally applicable model can be created.

The foundation of the model created, including the definition of phishing and the structure of an email, is influenced by information generated from previously conducted studies, scientific material, and reports in order to be coherent with industry standards and terms. The remainder of the model draws its structure from the thesis' phishing corpus to establish genuine property categories that conform to legitimately observed phishing behavior. The resulting model provides a universal and replicable approach to the collection of phishing emails.

9.2 Challenges

While conducting the thesis, some challenges were introduced impacting the overall performance and result of the thesis.

One of the major challenges faced occurred due to the shift from reporting emails via the ticketing system to utilizing the MailRisk application. As the MailRisk application provided an easier approach for the end user to report suspicious emails, the launch of the application led to an influx in the amount of phishing emails reported. Because this change occurred in August of 2020, this led to a great discrepancy between the data collected from January through July and August through December of that year. In total, 561 emails were collected in the seven former months, while 2620 were collected for the remaining five.

The significant increase in the latter five months presented a challenge regarding how the data should be analysed and presented. Analyzing the dataset as a whole would be consistent with that of the other scoped years, however, any statistics presented would heavily favor the observations of the last five months. Another approach could have been to extract the percentwise distribution of each part, adding a compensating factor so that the percents would be equally weighted, and combining them again. Consulting "Statistikkhjelpen" provided by NTNU, the latter approach was deemed to be manipulative of the data, and should be avoided. The approach suggested by Statistikkhjelpen, and ultimately utilized, was that when looking at 2020 alone, each part should be analyzed separately, while when looking at all the years combined, 2020 should be analyzed as a whole without any changes to the data. Although this would favor the latter part of 2020 in the overall analyses, it would avoid manipulation of the data, while also providing a separate analysis for each part, should that be of interest.

Another challenge presented towards the validity of the data was the source of phishing emails itself. As the basis for the phishing corpus was user-reported emails, the resulting dataset is only reflecting the phishing emails that the users themselves received and decided to report. This means that emails caught by spam filters or phishing emails simply not reported, are not a part of the final analysis and findings. Due to the overall size of the phishing corpus, as well as the observed variety within said corpus, the challenge was not deemed to be obstructing the

desired outcome. It is also defined in the scope of the thesis that the corpus is based on user-reported emails as opposed to all detected phishing emails, as well as stated during the discussion, reducing the chances of misunderstandings from the resulting findings.

9.3 Future Work

The thesis highlighted aspects that were in need of further research, as well as laid the foundation for future work on the subject of email phishing collection and analysis.

9.3.1 Continued Analyses

The Email Phishing Collection Model provides the baseline for future collections and subsequent analyses of email phishing. Repeated analyses needs to be conducted in order to continue identifying trends within email phishing. In addition, repeated analyses may help identify additional property categories not yet identified from the thesis' phishing corpus as well as refine what is already established, helping to further populate and develop the Email Phishing Collection Model.

9.3.2 Process Automation

Although the collection process was effectivized from the previously conducted study by utilizing new query tools, the collection is still relying on a great deal of human intervention. Any further work on the Email Phishing Collection Model should explore automation capabilities, increasing the efficiency of the model, as well as making the model usage a more appalling experience.

9.3.3 Investigate Findings

During the analysis of the dataset findings, several of the findings' root causes could not properly be determined. Why the CEO transfer scam saw such a great decline, why the Infect Target had a surge in 2017, why impersonating Apple is not utilized as much anymore, and why the spoofing technique has been declining, are all questions that could not be properly argued for within this thesis. These questions are in need of further investigation in order to determine their causes.

Additionally, the analysis of the dataset findings highlighted a limitation of the thesis, which was also briefly mentioned in the challenges section. This limitation is related to the users from whom the phishing emails are reported by. As we have no knowledge of the reporting habits of the users, it cannot be determined to what degree the users report phishing emails. An analysis into the users' reporting habits should be conducted in order to identify any challenges to the validity of the thesis' findings. Other phishing sources, not relying on the user reporting the phishing email, may be investigated as well for the same purpose.

Bibliography

- [1] F. Salahdine and N. Kaabouch, 'Social Engineering Attacks: A Survey,' *Future Internet*, vol. 11, no. 4, 2019.
- [2] T. Henderson, 'ACTIVISION DATA BREACH CONTAINS EMPLOYEE DETAILS, CALL OF DUTY'S FUTURE, AND MORE,' *Insider Gaming*, 2023, Accessed 03.04.2023. [Online]. Available: <https://insider-gaming.com/activision-data-breach/>.
- [3] @KeyserSosa, 'We had a security incident. Here's what we know,' 2023, Accessed 03.04.2023. [Online]. Available: https://www.reddit.com/r/reddit/comments/10y427y/we_had_a_security_incident_heres_what_we_know/.
- [4] 'Information About a Recent Mailchimp Security Incident,' *Mailchimp*, 2023, Accessed 03.04.2023. [Online]. Available: <https://mailchimp.com/january-2023-security-incident/>.
- [5] 'How we handled a recent phishing incident that targeted Dropbox,' *Dropbox.Tech*, 2022, Accessed 03.04.2023. [Online]. Available: <https://dropbox.tech/security/a-recent-phishing-campaign-targeting-dropbox>.
- [6] N. Gomes, 'American Airlines says data breach affected some customers, employees,' *Reuters*, 2022, Accessed 03.04.2023. [Online]. Available: <https://www.reuters.com/business/aerospace-defense/american-airlines-says-data-breach-affected-small-number-customers-employees-2022-09-20/>.
- [7] 'Security Update,' *Uber*, 2022, Accessed 03.04.2022. [Online]. Available: <https://www.uber.com/newsroom/security-update/>.
- [8] 'How we're responding to a third-party vendor phishing incident,' *DoorDash*, 2022, Accessed 03.04.2023. [Online]. Available: <https://doordash.news/get-the-facts/how-were-responding-to-a-third-party-vendor-phishing-incident/>.
- [9] 'Cisco Event Response: Corporate Network Security Incident,' *Cisco*, 2022, Accessed 03.04.2023. [Online]. Available: https://sec.cloudapps.cisco.com/security/center/resources/corp_network_security_incident.

- [10] 'Incident Report: Employee and Customer Account Compromise,' *Twilio*, 2022, 03.04.2023. [Online]. Available: <https://www.twilio.com/blog/august-2022-social-engineering-attack>.
- [11] 'X-Force Threat Intelligence Index 2023,' *IBM*, 2023.
- [12] A. Ferreira and P. Vieira-Marques, 'Phishing Through Time: A Ten Year Story based on Abstracts,' *4th International Conference on Information Systems Security and Privacy*, pp. 225–232, 2018.
- [13] 'PHISHING ACTIVITY TRENDS REPORT 3rd Quarter 2022,' *APWG*, 2022.
- [14] '2022 ThreatLabz Phishing Report,' *Zscaler*, 2022.
- [15] '2023 ANNUAL STATE of EMAIL SECURITY REPORT,' *Cofense*, 2023.
- [16] M. R. Riedle, 'Identifying Trends Among Phishing Attacks,' 2016, Purdue University, Master Thesis.
- [17] C. Dhinakaran, J. K. Lee and D. Nagamalai, "Reminder: please update your details": Phishing Trends,' *2009 First International Conference on Networks & Communications*, pp. 295–300, 2009.
- [18] A. Ferreira and G. Lenzini, 'An analysis of social engineering principles in effective phishing,' *2015 Workshop on Socio-Technical Aspects in Security and Trust*, pp. 9–16, 2015.
- [19] 'Brand Phishing report – Q3 2020,' *Check Point*, 2020.
- [20] Z. Alkhalil, C. Hewage, L. Nawaf and I. Khan, 'Phishing Attacks: A Recent Comprehensive Study and a New Anatomy,' *Frontiers in Computer Science*, vol. 3, 2021.
- [21] R. Alabdan, 'Phishing Attacks Survey: Types, Vectors, and Technical Approaches,' *Future Internet*, vol. 12, no. 10, 2020.
- [22] 'Compromise of a power grid in eastern Ukraine,' *Council on Foreign Relations*, 2015, Accessed 29.04.2023. [Online]. Available: <https://www.cfr.org/cyber-operations/compromise-power-grid-eastern-ukraine>.
- [23] K. Rekouche, 'Early phishing,' 2011. arXiv: 1106.4692.
- [24] S. Gibbs, 'How did email grow from messages between academics to a global epidemic?' *The Guardian*, 2016, Accessed 29.04.2023. [Online]. Available: <https://www.theguardian.com/technology/2016/mar/07/email-ray-tomlinson-history>.
- [25] P Knight, 'ILOVEYOU: Viruses, paranoia, and the environment of risk,' *The Sociological Review*, vol. 48, no. 2, pp. 17–30, 2000.
- [26] D. Patel and X. Luo, 'Take a Close Look at Phishing,' *Proceedings of the 4th Annual Conference on Information Security Curriculum Development*, no. 4, pp. 1–4, 2007.
- [27] 'History of Phishing,' *phishing.org*, Accessed 29.04.2023. [Online]. Available: <https://www.phishing.org/history-of-phishing>.

- [28] 'The History of Phishing,' *Graphus*, 2023, Accessed 29.04.2023. [Online]. Available: <https://www.graphus.ai/blog/history-of-phishing/>.
- [29] K. E. McIntyre, 'The Evolution of Social Media from 1969 to 2013: A Change in Competition and a Trend Toward Complementary, Niche Sites,' *The Journal of Social Media Society*, vol. 3, no. 2, pp. 5–25, 2014.
- [30] N. R. Braham, 'Mitigating the Spread of Spear-Phishing Attacks as a Form of Cybercrime: A Human Cognitive Challenge,' *ProQuest Dissertations and Theses*, 2022.
- [31] H. Motika, *The effects of phishing emails : a meta-analysis*, 2022.
- [32] N. Urbach and M. Röglinger, *Introduction to Digitalization Cases: How Organizations Rethink Their Business for the Digital Age*. 2019, pp. 1–12.
- [33] R. Richardson and M. M. North, 'Ransomware: Evolution, Mitigation and Prevention,' *International Management Review*, vol. 13, no. 1, pp. 10–21,
- [34] I. A. Chesti, M. Humayun, N. U. Sama and N. Jhanjhi, 'Evolution, Mitigation, and Prevention of Ransomware,' *2020 2nd International Conference on Computer and Information Sciences (ICCIS)*, pp. 1–6, 2020.
- [35] N. Aldaraani and Z. Begum, 'Understanding the impact of Ransomware: A Survey on its Evolution, Mitigation and Prevention Techniques,' *2018 21st Saudi Computer Society National Computer Conference (NCC)*, pp. 1–5, 2018.
- [36] A. E. Agazzi, 'Phishing and Spear Phishing: examples in Cyber Espionage and techniques to protect against them,' 2020. arXiv: 2006.00577.
- [37] 'Cost of a Data Breach Report 2022,' *IBM*, 2022.
- [38] S. Grønmo, 'kvantitativ metode,' *Store Norske Leksikon*, Accessed 30.04.2023. [Online]. Available: https://snl.no/kvantitativ_metode.
- [39] S. Grønmo, 'kvalitativ metode,' *Store Norske Leksikon*, Accessed 30.04.2023. [Online]. Available: https://snl.no/kvalitativ_metode.
- [40] S. A. H. Karset, 'An Extensive Analysis of Email Phishing - Properties, Detection, and Successful Phishing,' 2022, NTNU. [Online]. Available: <https://ntnuopen.ntnu.no/ntnu-xmlui/handle/11250/3027068>.
- [41] M. Bitaab, H. Cho, A. Oest, P. Zhang, Z. Sun, R. Pourmohamad, D. Kim, T. Bao, R. Wang, Y. Shoshitaishvili, A. Doupe and G.-J. Ahn, 'Scam Pandemic: How Attackers Exploit Public Fear through Phishing,' *2020 APWG Symposium on Electronic Crime Research (eCrime)*, pp. 1–10, 2020.
- [42] S. Gupta and P. Kumaraguru, 'Emerging phishing trends and effectiveness of the anti-phishing landing page,' *2014 APWG Symposium on Electronic Crime Research (eCrime)*, pp. 36–47, 2014.
- [43] D. Irani, S. Webb, J. Giffin and C. Pu, 'Evolutionary study of phishing,' *2008 eCrime Researchers Summit*, pp. 1–10, 2008.

- [44] G. Ho, A. Cidon, L. Gavish, M. Schweighauser, V. Paxson, S. Savage, G. M. Voelker and D. Wagner, 'Detecting and Characterizing Lateral Phishing at Scale,' *28th USENIX Security Symposium (USENIX Security 19)*, pp. 1273–1290, 2019.
- [45] B. C. Dhinakaran, J.-K. Lee and D. Nagamalai, "Reminder: please update your details": Phishing Trends,' *2009 First International Conference on Networks & Communications*, pp. 295–300, 2009.
- [46] D. Irani, S. Webb, J. Giffin and C. Pu, 'Evolutionary study of phishing,' *2008 eCrime Researchers Summit*, pp. 1–10, 2008.
- [47] A. Ferreira and G. Lenzini, 'An analysis of social engineering principles in effective phishing,' *2015 Workshop on Socio-Technical Aspects in Security and Trust*, pp. 9–16, 2015.
- [48] Q. Cui, G.-V. Jourdan, G. V. Bochmann, R. Couturier and I.-V. Onut, 'Tracking Phishing Attacks Over Time,' *International World Wide Web Conferences Steering Committee, WWW '17*, pp. 667–676, 2017.
- [49] '2022 Annual State of Phishing Report,' *Cofense*, 2023.
- [50] '2023 State of the Phish,' *proofpoint*, 2023.
- [51] 'The State of Phishing,' *SlashNext*, 2023.
- [52] 'H2 2022 Email Threat Report,' *Abnormal*, 2022.
- [53] J. Klensin, 'Simple Mail Transfer Protocol,' 2008. [Online]. Available: <https://www.rfc-editor.org/rfc/rfc5321>.
- [54] 'Message Headers,' *Internet Assigned Numbers Authority*, Accessed 03.05.2023. [Online]. Available: <https://www.iana.org/assignments/message-headers/message-headers.xhtml>.
- [55] KnowBe4, 'What Is CEO Fraud,' Accessed 11.03.2023. [Online]. Available: <https://www.knowbe4.com/ceo-fraud>.
- [56] Nettvett, 'Direktørsvindel (CEO-fraud),' Accessed 11.03.2023. [Online]. Available: <https://nettvett.no/direktor-svindel/>.
- [57] Barracuda, 'CEO Fraud,' Accessed 11.03.2023. [Online]. Available: <https://www.barracuda.com/support/glossary/ceo-fraud>.
- [58] 'Personal Data,' *GDPREU*, Accessed 11.03.2023. [Online]. Available: <https://www.gdpreu.org/the-regulation/key-concepts/personal-data>.
- [59] M. Schmidt, 'The sankey diagram in energy and material flow management,' *Journal of Industrial Ecology*, vol. 12, no. 2, pp. 173–185, 2008.
- [60] E. H. Simpson, 'Measurement of Diversity,' *Nature*, vol. 163, no. 4148, pp. 688–688, 1949.
- [61] K. Greene, M. P. Steves, M. F. Theofanos and J. Kostick, 'User Context: An Explanatory Variable in Phishing Susceptibility,' *Proceedings of the Network and Distributed Systems Security (NDSS) Symposium*, 2018.

- [62] M. P. Steves, K. Greene and M. F. Theofanos, 'A Phish Scale: Rating Human Phishing Message Detection Difficulty,' *Proceedings 2019 Workshop on Usable Security*, 2019.
- [63] M. P. Steves, K. Greene and M. F. Theofanos, 'Categorizing human phishing difficulty: a Phish Scale,' *Journal of Cybersecurity*, vol. 6, no. 1, 2020.
- [64] E. J. Williams, J. Hinds and A. N. Joinson, 'Exploring susceptibility to phishing in the workplace,' *International Journal of Human-Computer Studies*, vol. 120, pp. 1–13, 2018.
- [65] T. Sharma and M. Bashir, 'An Analysis of Phishing Emails and How the Human Vulnerabilities are Exploited,' *Advances in Human Factors in Cybersecurity*, pp. 49–55, 2020.
- [66] FBI, 'Business Email Compromise: Gift Cards,' *IC3*, 2018, Accessed 15.04.2023. [Online]. Available: <https://www.ic3.gov/Media/Y2018/PSA181024>.
- [67] T. Pattera, 'Understanding BEC Scams: Gift Card Scams,' *ProofPoint*, 2020, Accessed 15.04.2023. [Online]. Available: <https://www.proofpoint.com/us/blog/threat-protection/understanding-bec-scams-gift-card-scams>.
- [68] C. Hassold, 'BEC Group Compromises Personal Accounts and Pulls Heartstrings to Launch Mass Gift Card Attacks,' *Abnormal*, 2022, Accessed 15.04.2023. [Online]. Available: <https://abnormalsecurity.com/blog/lilac-wolverine-gift-card-attacks>.
- [69] P. Mangut and K. Datukun, 'The Current Phishing Techniques – Perspective of the Nigerian Environment,' *World Journal of Innovative Research*, vol. 10, no. 1, 2021.
- [70] R. Chaganti, B. Bhushan, A. Nayyar and M. Azrour, 'Recent trends in Social Engineering Scams and Case study of Gift Card Scam,' *CoRR*, vol. abs/2110.06487, 2021.
- [71] 'Why iTunes? A Look into Gift Cards as an Emerging BEC Cash Out Method,' *Agari*, 2019, Accessed 15.04.2023. [Online]. Available: <https://www.agari.com/blog/gift-cards-emerging-bec-method>.
- [72] '74 Arrested in Coordinated International Enforcement Operation Targeting Hundreds of Individuals in Business Email Compromise Schemes,' *Department of Justice*, 2018, Accessed 16.04.2023. [Online]. Available: <https://www.justice.gov/opa/pr/74-arrested-coordinated-international-enforcement-operation-targeting-hundreds-individuals>.
- [73] '281 Arrested Worldwide in Coordinated International Enforcement Operation Targeting Hundreds of Individuals in Business Email Compromise Schemes,' *Department of Justice*, 2019, Accessed 16.04.2023. [Online]. Available: <https://www.justice.gov/opa/pr/281-arrested-worldwide-coordinated-international-enforcement-operation-targeting-hundreds>.

- [74] S. Venkatesha, K. R. Reddy and B. R. Chandavarkar, 'Social Engineering Attacks During the COVID-19 Pandemic.,' *SN Comput Sci*, vol. 2, no. 3, 2021.
- [75] A. Oest, P. Zhang, B. Wardman, E. Nunes, J. Burgis, A. Zand, K. Thomas, A. Doupé and G.-J. Ahn, 'Sunrise to Sunset: Analyzing the End-to-end Life Cycle and Effectiveness of Phishing Attacks at Scale,' *29th USENIX Security Symposium (USENIX Security 20)*, pp. 361–377, 2020.
- [76] 'Credential Theft,' *Cofense*, Accessed 16.04.2023. [Online]. Available: <https://cofense.com/threats/credential-theft/>.
- [77] 'What is a Credential-Based Attack?' *Palo Alto Networks*, Accessed 16.04.2023. [Online]. Available: <https://www.paloaltonetworks.com/cyberpedia/what-is-a-credential-based-attack>.
- [78] '2021 Annual State of Phishing Report,' *Cofense*, 2022.
- [79] L. Spitzner, 'Phishing - It's No Longer About Malware (or Even Email),' *Sans*, 2023, Accessed 25.04.2023. [Online]. Available: <https://www.sans.org/blog/phishing-its-no-longer-about-malware-or-even-email/>.
- [80] J. Davis, '50% Phishing Emails Seek Credential Theft, as Malware Delivery Declines,' *Health IT Security*, 2021, Accessed 18.04.2023. [Online]. Available: <https://healthitsecurity.com/news/50-phishing-emails-seek-credential-theft-as-malware-delivery-declines>.
- [81] 'Ransomware and malware attacks decline, attackers adopting covert tactics,' *Help Net Security*, 2019, Accessed 18.04.2023. [Online]. Available: <https://www.helpnetsecurity.com/2019/05/20/ransomware-attacks-decline/>.
- [82] H. Oz, A. Aris, A. Levi and A. S. Uluagac, 'A Survey on Ransomware: Evolution, Taxonomy, and Defense Solutions,' *Association for Computing Machinery*, vol. 54, no. 11s, 2022.
- [83] 'Locky Ransomware,' *KnowBe4*, Accessed 18.04.2023. [Online]. Available: <https://www.knowbe4.com/locky-ransomware>.
- [84] M. Humayun, N. Jhanjhi, A. Alsayat and V. Ponnusamy, 'Internet of things and ransomware: Evolution, mitigation and prevention,' *Egyptian Informatics Journal*, vol. 22, no. 1, pp. 105–117, 2021.
- [85] 'Annual number of malware attacks worldwide from 2015 to 2022,' *Statista*, Accessed 18.04.2023. [Online]. Available: <https://www.statista.com/statistics/873097/malware-attacks-per-year-worldwide/>.
- [86] J. Sattler, 'Failed delivery spam and other naughty things to watch out for this holiday season,' *F-Secure*, 2018, Accessed 18.04.2023. [Online]. Available: <https://blog.f-secure.com/failed-delivery-spam/>.

- [87] 'URLs 4X More Likely than Phishing Attachments to Reach Users,' *Cofense*, 2023, Accessed 18.04.2023. [Online]. Available: <https://cofense.com/blog/urls-4x-more-likely-than-phishing-attachments-to-reach-users/>.
- [88] 'Be Alert! HTML Email Attachments Used in Phishing,' *Trend Micro*, 2022, Accessed 18.04.2023. [Online]. Available: <https://news.trendmicro.com/2022/10/31/html-email-attachments-phishing-scam/>.
- [89] 'Facebook is Most Imitated Brand for Phishing Attempts: Check Point Research's Q4 2019 Brand Phishing Report,' *Check Point*, 2020, Accessed 20.04.2023. [Online]. Available: <https://www.checkpoint.com/press/2020/facebook-is-most-imitated-brand-for-phishing-attempts-check-point-researchs-q4-2019-brand-phishing-report/>.
- [90] 'Apple is Most Imitated Brand for Phishing Attempts: Check Point Research's Q1 2020 Brand Phishing report,' *Check Point*, 2020, Accessed 20.04.2023. [Online]. Available: <https://www.checkpoint.com/press/2020/apple-is-most-imitated-brand-for-phishing-attempts-check-point-researchs-q1-2020-brand-phishing-report/>.
- [91] 'Microsoft is Most Imitated Brand for Phishing Attempts in Q3 2020,' *Check Point*, 2020, Accessed 20.04.2023. [Online]. Available: <https://blog.checkpoint.com/security/microsoft-is-most-imitated-brand-for-phishing-attempts-in-q3-2020/>.
- [92] 'Microsoft Continues to be Most Imitated Brand for Phishing Attempts in Q1 2021,' *Check Point*, 2021, Accessed 20.04.2023. [Online]. Available: <https://blog.checkpoint.com/security/microsoft-continues-to-be-most-imitated-brand-for-phishing-attempts-in-q1-2021/>.
- [93] 'Brand Phishing Report Q2 2021: Microsoft Continues Reign,' *Check Point*, 2021, Accessed 20.04.2023. [Online]. Available: <https://blog.checkpoint.com/security/brand-phishing-report-q2-2021-microsoft-continues-reign/>.
- [94] 'DHL Replaces Microsoft as Most Imitated Brand in Phishing Attempts in Q4 2021,' *Check Point*, 2022, Accessed 20.04.2023. [Online]. Available: <https://blog.checkpoint.com/security/dhl-replaces-microsoft-as-most-imitated-brand-in-phishing-attempts-in-q4-2021/>.
- [95] 'Social Media Network LinkedIn Ranks First in List of Brands Most Likely to be Imitated in Phishing Attempts in Q1 2022,' *Check Point*, 2022, Accessed 20.04.2023. [Online]. Available: <https://www.checkpoint.com/press-releases/social-media-network-linkedin-ranks-first-in-list-of-brands-most-likely-to-be-imitated-in-phishing-attempts-in-q1-2022/>.

- [96] ‘Yahoo Most Impersonated Brand in Q4 2022 Phishing Attacks,’ *Check Point*, 2022, Accessed 20.04.2023. [Online]. Available: <https://www.checkpoint.com/press-releases/yahoo-most-impersonated-brand-in-q4-2022-phishing-attacks/>.
- [97] ‘Microsoft Is the Most Impersonated Brand in Phishing Attacks,’ *Vade*, 2022, Accessed 20.04.2023. [Online]. Available: <https://www.vadesecure.com/en/company/news/microsoft-is-the-most-impersonated-brand-in-phishing-attacks>.
- [98] L. Jenkins, S. Hawley, P. Najafi and D. Bienstock, ‘Suspected Russian Activity Targeting Government and Business Entities Around the Globe,’ *Retrieved December*, vol. 27, p. 2021, 2021.
- [99] U. Shakir, ‘Apple surpasses 2 billion active devices,’ *The Verge*, 2023, Accessed 20.04.2023. [Online]. Available: <https://www.theverge.com/2023/2/2/23583501/apple-iphone-ipad-active-2-billion-devices-q1-2023>.
- [100] ‘List of Apple products,’ *Apple Wiki*, Accessed 20.04.2023. [Online]. Available: https://apple.fandom.com/wiki/List_of_Apple_products.
- [101] A. Jayatilaka, N. A. G. Arachchilage and M. A. Babar, ‘Falling for Phishing: An Empirical Investigation into People’s Email Response Behaviors,’ *Forty-Second International Conference on Information Systems*, vol. abs/2108.04766, 2021.
- [102] R. Chen, J. Gaia and H. R. Rao, ‘An examination of the effect of recent phishing encounters on phishing susceptibility,’ *Decision Support Systems*, vol. 133, p. 113 287, 2020.
- [103] H. Abroshan, J. Devos, G. Poels and E. Laermans, ‘Phishing Happens Beyond Technology: The Effects of Human Behaviors and Demographics on Each Step of a Phishing Process,’ *IEEE Access*, vol. 9, pp. 44 928–44 949, 2021.
- [104] D. Jampen, G. Gür, T. Sutter and B. Tellenbach, ‘Don’t click: towards an effective anti-phishing training. A comparative literature review,’ *Human-centric Computing and Information Sciences*, vol. 10, no. 1, pp. 33–74, 2020.
- [105] G. Moody, D. Galletta, J. Walker and B. Dunn, ‘Which Phish Get Caught? An Exploratory Study of Individual Susceptibility to Phishing.,’ *International Conference on Information Systems 2011, ICIS 2011*, vol. 3, 2011.
- [106] A. Hamid, ‘Apple is the darling of an overwhelming majority of phishing criminals,’ *Phone Arena*, 2023, Accessed 20.04.2023. [Online]. Available: https://www.phonearena.com/news/Apple-is-the-darling-of-an-overwhelming-majority-of-phishing-criminals_id146915.

- [107] G. Turner, 'What Percentage of E-shop Phishing Scams Exploit Apple's Name?' *Digit News*, 2023, Accessed 20.04.2023. [Online]. Available: <https://www.digit.fyi/what-percentage-of-e-shop-phishing-scam-exploit-apples-name/>.
- [108] J. Hollington, 'PSA: New Round of "iCloud Support" Scam Emails Are Making the Rounds | Here's How to Protect Yourself,' *iDropNews*, 2023, Accessed 20.04.2023. [Online]. Available: <https://www.idropnews.com/news/psa-new-round-of-icloud-support-scam-emails-are-making-the-rounds-heres-how-to-protect-yourself/192677/>.
- [109] N. Shaw, B. Eschenbrenner and D. Baier, 'Online shopping continuance after COVID-19: A comparison of Canada, Germany and the United States,' *Journal of Retailing and Consumer Services*, vol. 69, 2022.
- [110] M. Alawida, A. E. Omolara, O. I. Abiodun and M. Al-Rajab, 'A deeper look into cybersecurity issues in the wake of Covid-19: A survey,' *Journal of King Saud University - Computer and Information Sciences*, vol. 34, no. 10, Part A, pp. 8176–8206, 2022.
- [111] T. Shcherbakova, 'Fake deliveries in an age of lockdown,' *Kaspersky*, 2020, Accessed 20.04.2023. [Online]. Available: <https://www.kaspersky.com/blog/covid-fake-delivery-service-spam-phishing/35125/>.
- [112] L. Abrams, 'Fake Fedex and UPS delivery issues used in COVID-19 phishing,' *Bleeping Computer*, 2020, Accessed 20.04.2023. [Online]. Available: <https://www.bleepingcomputer.com/news/security/fake-fedex-and-ups-delivery-issues-used-in-covid-19-phishing/>.
- [113] Staff Writer, 'COVID-19: SingPost warns of SMS scams related to 'parcel delivery',' *Yahoo News*, 2020, Accessed 20.04.2023. [Online]. Available: <https://sg.news.yahoo.com/covid-19-sing-post-warns-of-phishing-scams-related-to-parcel-delivery-142549332.html>.
- [114] 'Scammers capitalizing on online shopping boom with wave of package delivery fraud,' *Fraud.org*, 2020, Accessed 20.04.2023. [Online]. Available: https://fraud.org/package_delivery_alert/.
- [115] S. Kitterman, 'Sender Policy Framework (SPF) for Authorizing Use of Domains in Email, Version 1,' 2014. [Online]. Available: <https://www.rfc-editor.org/rfc/rfc7208>.
- [116] D. Crocker, T. Hansen and M. Kucherawy, 'DomainKeys Identified Mail (DKIM) Signatures,' 2011. [Online]. Available: <https://www.rfc-editor.org/rfc/rfc6376>.
- [117] M. Kucherawy and E. Zwicky, 'Domain-based Message Authentication, Reporting, and Conformance (DMARC),' 2015. [Online]. Available: <https://www.rfc-editor.org/info/rfc7489>.
- [118] Z. Ramzan and C. Wueest, 'Phishing Attacks: Analyzing Trends in 2006,' *International Conference on Email and Anti-Spam*, 2007.

- [119] 'Tax Return,' *Skatteetaten*, Accessed 24.04.2023. [Online]. Available: <https://www.skatteetaten.no/en/person/taxes/tax-return/>.
- [120] 'Declaring taxes – for individuals,' *Skatteverket*, Accessed 24.04.2023. [Online]. Available: <https://skatteverket.se/privat/deklaration.4.2b543913a42158acf800013508.html>.
- [121] 'Relevant dates for the 2022 tax assessment notice,' *skat.dk*, Accessed 24.04.2023. [Online]. Available: <https://skat.dk/data.aspx?oid=2244338>.
- [122] W. Yeoh, H. Huang, W.-S. Lee, F. A. Jafari and R. Mansson, 'Simulated Phishing Attack and Embedded Training Campaign,' *Journal of Computer Information Systems*, vol. 62, no. 4, pp. 802–821, 2022.
- [123] A. Chaudhuri, 'Clone Phishing: Attacks and Defenses,' *International Journal of Scientific and Research Publications (IJSRP)*, vol. 13, no. 4, 2023.

Appendix A

Full Data Summary

The following section details a full analysis of the collected data based on the methodology and properties defined in Chapters 3 and 5. For each year within the collection scope, a presentation of the data and an accompanying analysis of the identified properties will be carried out.

The scope spans the years 2016 throughout 2022, where the properties collected includes Content, Target, Method, Impersonation, and Date.

A.1 2016

In total, 1173 phishing mails were collected for 2016.

A.1.1 Content

The phishing mails collected can be distributed into 42 distinct Content categories ranging from 1 to 196 entries per category. Figure A.1 showcases the distribution of phishing mails per category where the category “Confirm / Update Account Information” was the most persistent with 196 entries, while “New Message” and “CEO Scam – Transfer Money” entered second and third with 125 and 118 entries respectively.

Based on the distribution, there is a clear difference between the uppermost category and the second ranging category, while the following categories shows a lean decline until the graph fairly evens out.

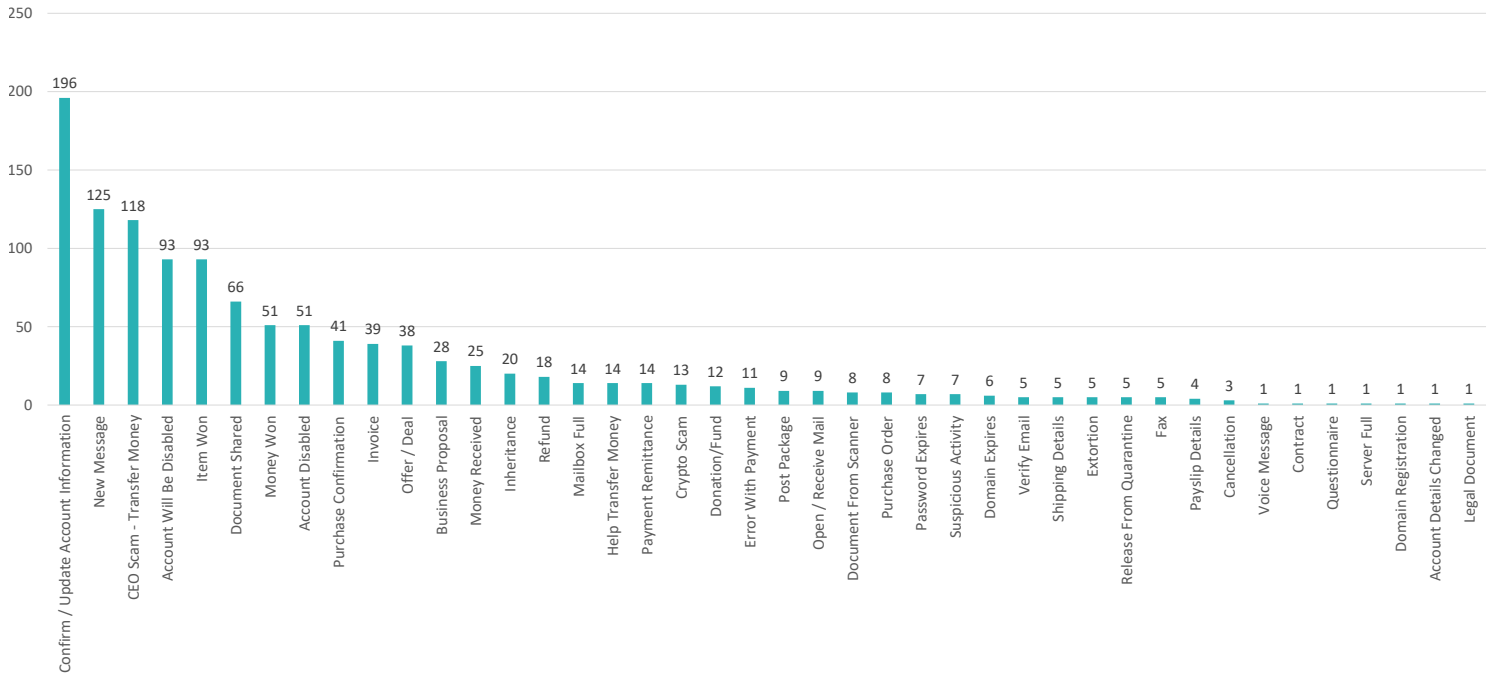


Figure A.1: 2016 - Content Distribution

A.1.2 Target

Table A.1 displays the distribution of Targets within the collected phishing mails for 2016.

Target	Total	% of Total
Credentials	536	45.69%
Money	244	20.80%
Credit Card Details	237	20.20%
Infect	104	8.87%
PII	39	3.32%
N/A	13	1.11%

Table A.1: 2016 - Target Distribution

As can be seen, Credentials were the most sought after Target in 2016 followed by Money and Credit Card Details, with Infect and PII occupying the lower range of Targets. N/A accounts for 1.11% of the collected mails, meaning that the Target of 13 of the mails could not be determined with the information available. The distribution showcases a clear preference for Credentials as the main Target for the malicious actors, accounting for nearly half of the collected mails.

A.1.3 Method

Figure A.2 shows the distribution of Methods for the 2016 corpus. There are in total three categories identified for this distribution where URL was the definite most utilized method with Communication and Attachment second and third. The attachment category is again divided into subcategories defining the type of attachment observed. As shown, Word was the most utilized document type in a total of five attachment types.

Method	Total	% of Total
URL	798	68.03%
Communication	263	22.42%
Attachment	112	9.55%

Attachment Type	Total	% of Total
Word	59	5.03%
PDF	22	1.88%
ZIP	20	1.71%
HTML	7	0.60%
Excel	4	0.34%

Figure A.2: 2016 - Method Distribution

A.1.4 Impersonation

The pie chart shown in Figure A.3 depicts the distribution of impersonated parts observed within the phishing mails (top 20 categories). The generic categories of External and Internal takes up a large portion of the chart, while Danske Bank has a significant presence from the non-generic categories. In total, there were 35 distinct Impersonation categories observed, where the top three categories takes up over 69 percent of the observed impersonations.

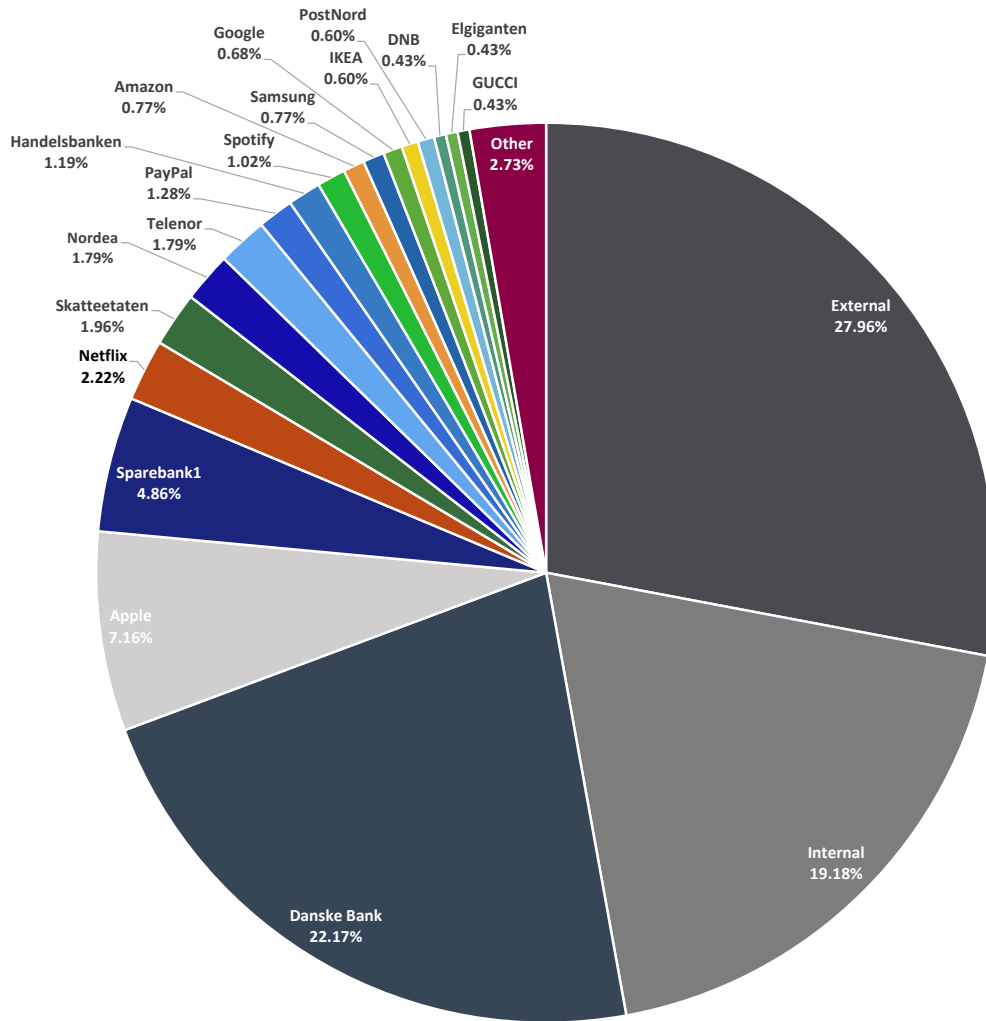


Figure A.3: 2016 - Impersonation Distribution

A.1.5 Dates

The date distribution of the collected phishing mails can be seen in the heat-map of figure A.4. As explained in the main section, the ticketing system from where the phishing mails were collected was not implemented until mid-February of 2016, meaning that no phishing data is present in this time range.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	
Jan	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
Feb	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	1	4	0	0
Mar	2	3	6	1	0	1	3	2	13	4	6	4	1	9	5	2	7	5	2	3	12	4	9	3	1	0	4	1	5	3	0	
Apr	1	0	1	4	1	2	5	2	0	3	6	3	4	6	0	0	2	3	0	2	2	2	2	0	4	6	4	0	3	2	0	
May	0	1	5	2	1	3	2	0	1	2	1	3	0	1	2	1	3	1	18	1	0	0	2	1	5	12	2	1	1	0	2	
Jun	4	5	2	1	6	2	3	5	7	7	0	0	1	4	2	2	7	0	2	7	4	13	4	2	0	3	2	2	4	3	0	
Jul	2	0	0	1	3	2	2	1	0	0	3	6	4	5	14	3	1	4	6	7	7	1	0	4	20	23	3	22	26	0	0	
Aug	9	7	4	6	2	2	1	30	7	3	27	8	1	3	13	13	4	34	26	0	0	122	7	0	10	1	0	0	1	19	5	
Sep	0	4	0	0	5	5	0	5	2	2	1	3	3	5	2	1	1	0	10	7	19	6	4	1	0	10	10	1	3	5	0	
Oct	0	4	13	2	2	1	1	2	1	3	3	1	4	8	1	0	3	9	1	2	1	0	1	3	6	2	3	2	0	0	3	
Nov	1	5	1	1	0	0	1	4	8	2	2	0	0	2	0	1	0	0	0	1	1	0	0	0	2	1	1	3	1	2	0	
Dec	0	1	0	1	4	4	1	0	1	1	0	4	0	0	0	1	2	0	0	0	0	2	1	1	0	0	0	1	0	0	0	

Figure A.4: 2016 - Date Distribution

Based on this distribution, it is evident that the August month of 2016 saw the most mails, while the winter months shows a decline in the observed phishing mails. August 22nd shows a surge in the amount of mails observed. From the Content log, it can be shown that this surge is tied to the three Content categories New Message (45), Confirm / Update Account Information (27), and Account Disabled (50), all whom were connected to a banking impersonation, either Danske Bank or SpareBank1.

A.1.6 Property Relationships

Figure A.5 displays the relationship between the two collection properties Target and Content. Each Target category is tied to one or multiple Contents that they have been observed appearing in.

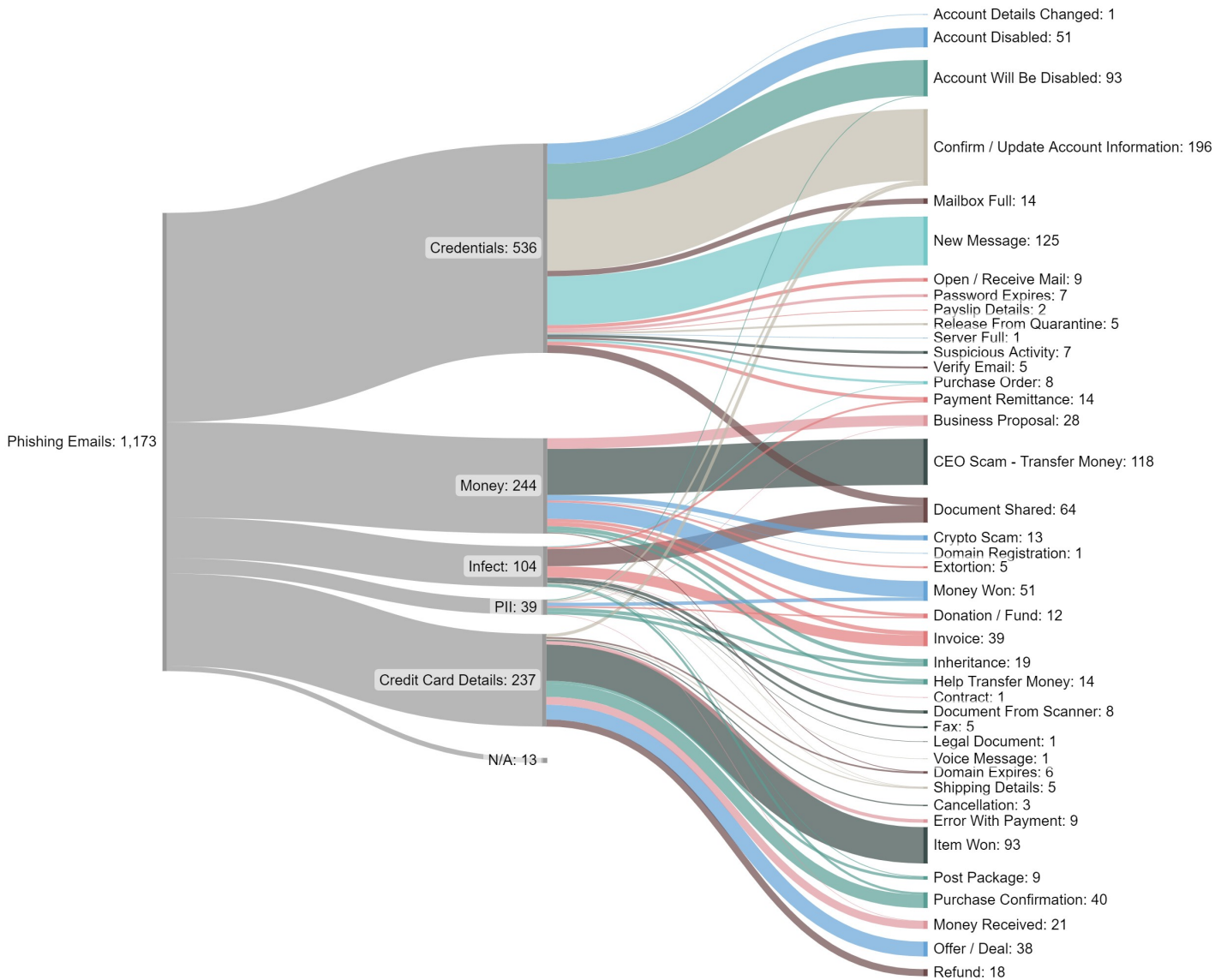


Figure A.5: 2016 - Target-Content Relationship

25 of the Content categories have been observed having a singular relationship with a Target, meaning that all phishing emails within the given Content category is only targeting one specific Target. Credentials is the Target category with the most singular relationships, totalling 11. On the other hand, both the Content categories of Confirm / Update Account Information and Shipping Details are observed with three different Targets; Credentials - PII - Credit Card Details, and Money - Infect - Credit Card Details.

From Table A.2, displaying the Content distribution and diversity per Target, shows that PII is the most diverse Target, with a diversity score of 0.84. The diversity

score is calculated utilizing Simpson’s Diversity Index [60]. A diversity index of 1 indicates high diversity, while an index of 0 indicates no diversity.

Target	Total	Total Connections	$1 - \left(\frac{\sum n(n-1)}{N(N-1)} \right)$
Credentials	536	16	0.79
Money	244	12	0.72
Credit Card Details	237	11	0.79
Infect	104	12	0.74
PII	39	8	0.84

Table A.2: 2016 - Target-Content Diversity

Overall, all of the Targets showcases a good, as well as consistent diversity rate, all being above 0.70, and all being within 0.12 points from each other.

Viewing the relationship between Method and Target, as displayed in Figure A.6, provides insight into any preferences regarding method of achievement for the specific Targets.

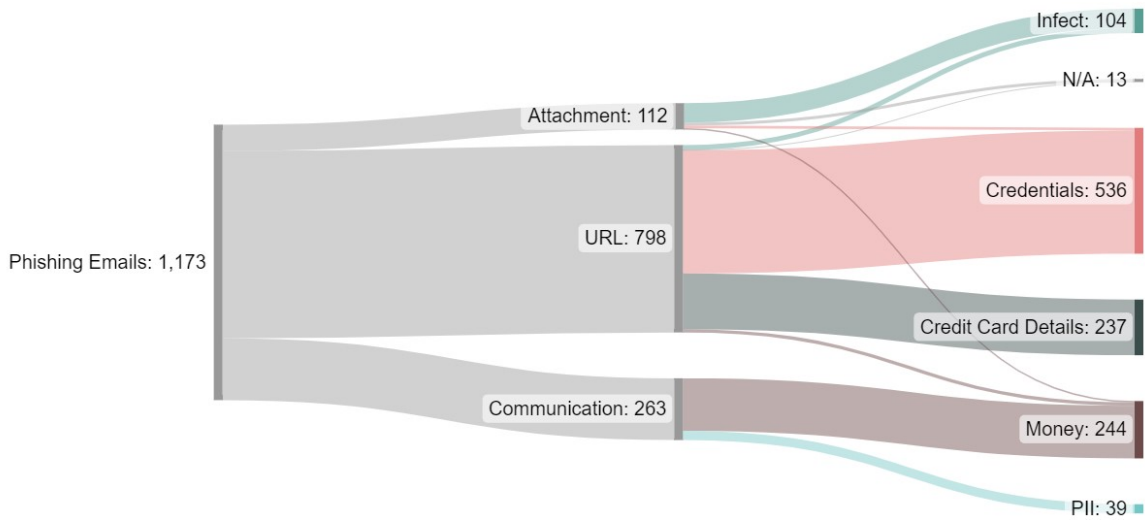


Figure A.6: 2016 - Method-Target Relationship

Credit Card Details are only targeted through the utilization of links in the phishing emails, and PII is only targeted via Communication. The remainder of the

Targets, although featured in various Methods, have one main Method of achievement, such as Credentials mainly being targeted through URLs, Money through Communication, and Infect through Attachments.

A.2 2017

A total of 986 phishing mails were identified for 2017. This is a decrease of 187 mails compared to the 2016 corpus.

A.2.1 Content

The collected phishing mails from 2017 could be distributed into 47 Content categories, where 11 new categories could be identified. Six of the Content categories observed in 2016 were not observed in 2017, this includes Cancellation, Fax, Payslip Details, Questionnaire, Server Full, and Voice Message.

Figure A.7 visualizes the distribution of Content categories for 2017.

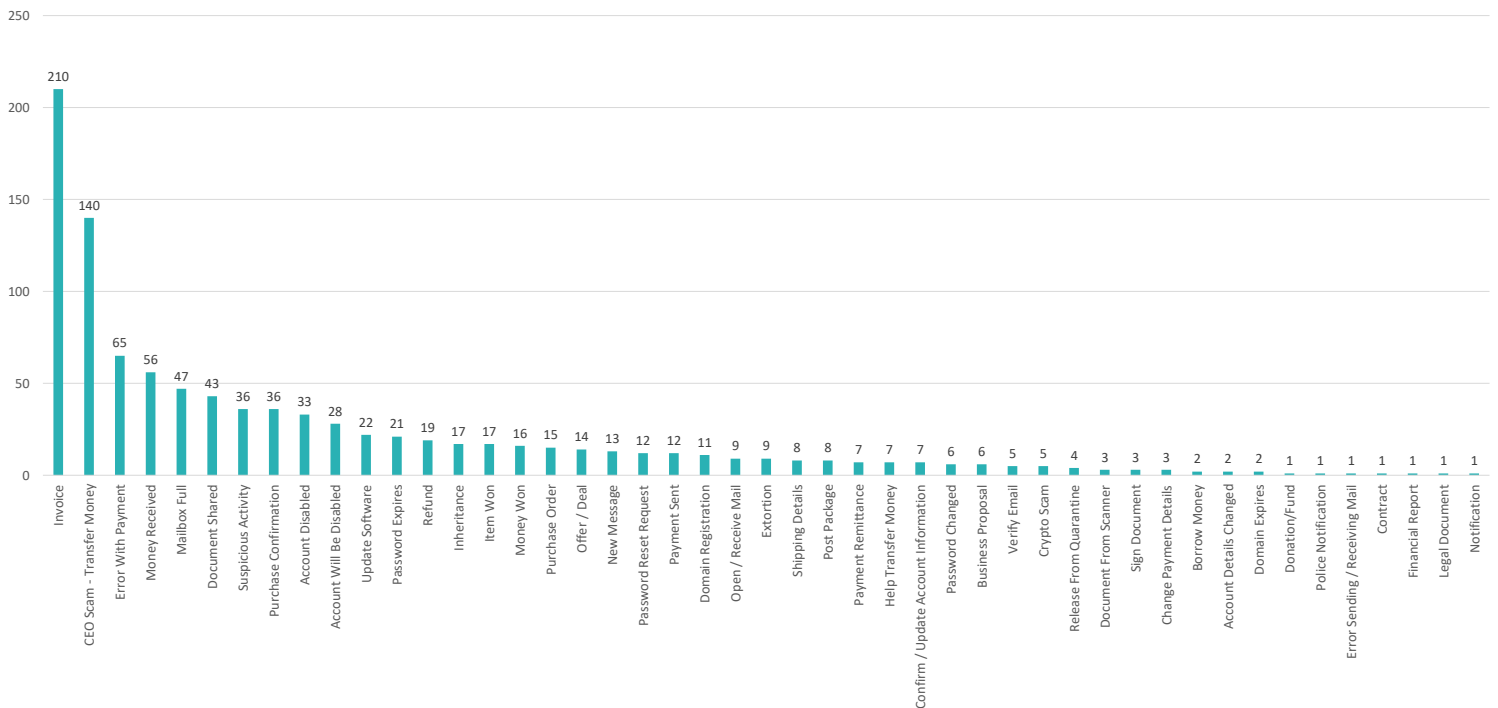


Figure A.7: 2017 - Content Distribution

The Invoice and CEO Scam – Transfer Money categories have a great outspurt compared to the rest of the categories who shows a more lean decline between each subsequent category.

As can be seen when comparing the 2016 Content data with the 2017 Content data, the top category of 2016, Confirm / Update Account Information, has seen a great reduction in representation. In 2016 this top category accounted for over 16% of the observed mails, while in 2017 it only accounts for 0.17% of the phishing corpus. CEO Scam – Transfer Money remains a high-ranking category for both years.

A.2.2 Target

Table A.7 displays the Target distribution for the 2017 corpus.

Target	Total	% of Total
Credentials	306	31.03%
Infect	234	23.73%
Credit Card Details	206	20.89%
Money	200	20.28%
N/A	25	2.54%
PII	15	1.52%

Table A.3: 2017 - Target Distribution

Similar to 2016, the Target of Credentials remains the most sought after Target, however, with a slight reduction of the overall total accounting for 31.03% of the mails compared to last year's 45.69%. The distribution also shows a great incline in the Target of infestation, reaching second with 23.73% compared to 8.87% last year. Credit Card Details and Money retains a percentage of a bit over 20% as it was in 2016.

A.2.3 Method

The distribution of the preferred Methods is close to equal both in terms of ranging and percentage from last year, with Credentials on top followed by Communication and Attachment, as can be seen in Figure A.8.

Method	Total	% of Total
URL	675	68.46%
Communication	203	20.59%
Attachment	107	10.85%
N/A	1	0.10%

Attachment Type	Total	% of Total
PDF	53	5.38%
Word	18	1.83%
ZIP	16	1.62%
Excel	13	1.32%
HTML	5	0.51%
ISO	1	0.10%
JAR	1	0.10%

Figure A.8: 2017 - Method Distribution

Expanding the Attachment section, PDF is shown to be the most prominent attachment type while Word has seen a small decline compared to last year. Two new attachment types are observed, that being the ISO and JAR attachment types.

A.2.4 Impersonation

Figure A.9 shows the distribution of the impersonated parties for the 2017 corpus. In total, there were 21 Impersonation categories observed.

Increasing significantly from 2016, the generic categories of External and Internal dominates the Impersonation property in the observed phishing mails. From the non-generic categories, it is shown that both Netflix and Apple are fairly popular brands in terms of impersonation for the phishing mails of 2017. Compared to 2016, the Impersonation brand of Danske Bank has had a significant decline down to a 0.71% overall from 22.17%.

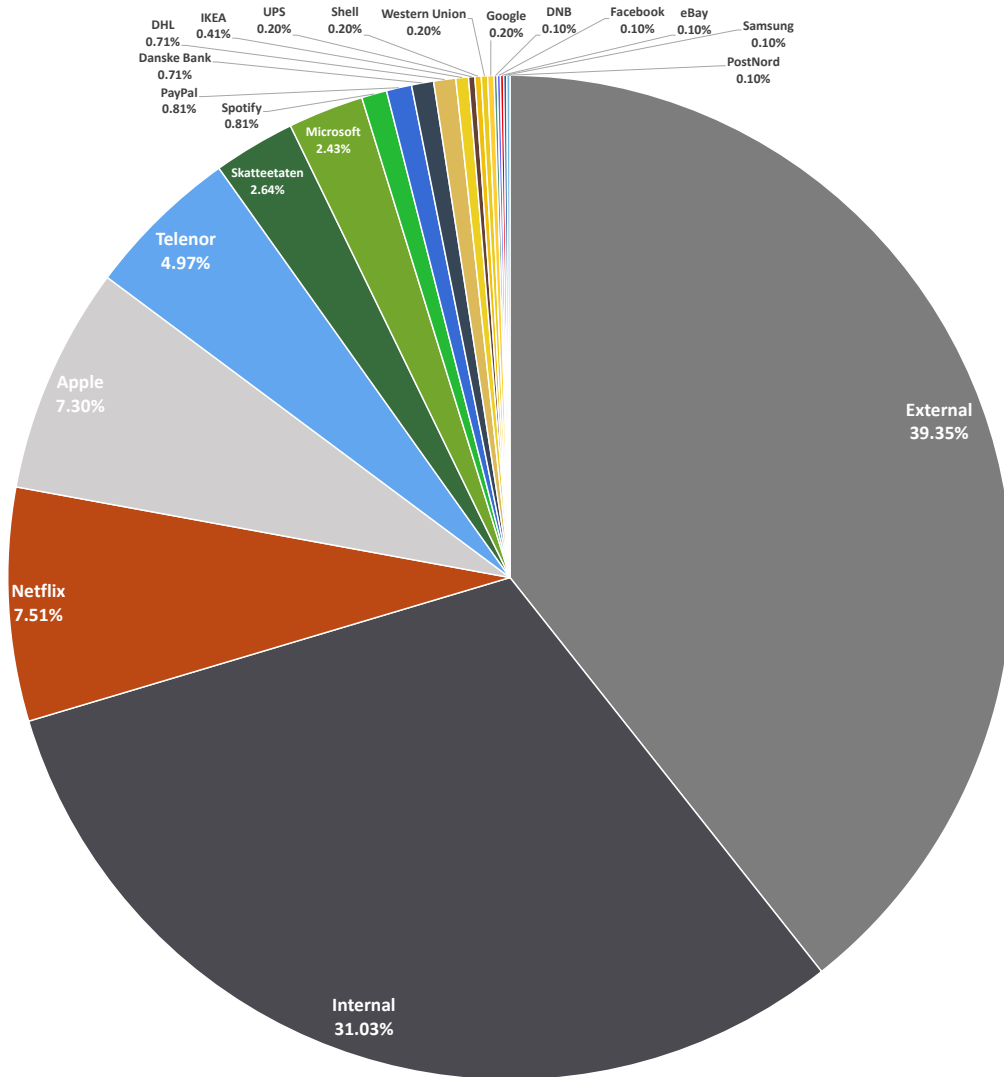


Figure A.9: 2017 - Impersonation Distribution

A.2.5 Dates

The heat-map, Figure A.10, displaying the distribution of the collected phishing mails show a high magnitude of mails centered around the months of February and March, while the rest of the months mainly only have small surges on specific dates before dabbing off the following days.

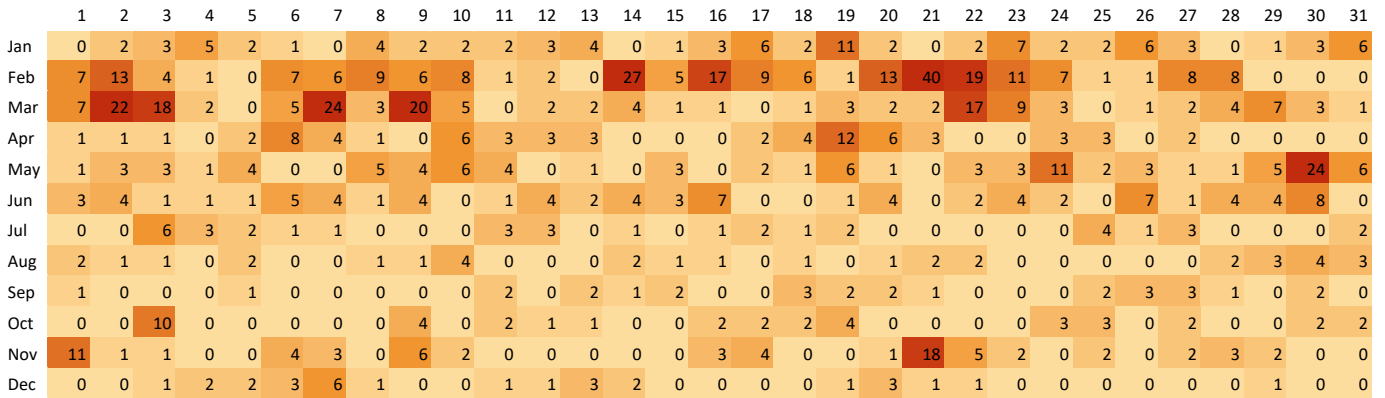


Figure A.10: 2017 - Date Distribution

As there was lacking data the first couple of months of 2016, no overall comparison can be made, however one can observe that the raise in phishing mails in the August month is not present in the 2017 collection.

A.2.6 Property Relationships

From the relationship between Target and Content, as displayed in Figure A.11, The Credential Target continues to have the most singular relationships with 15 out 34. Two of the content categories, Invoice and Purchase Confirmation, was observed aiming for four different Targets, both targeting Credentials, Infect, Money, and Credit Card Details.

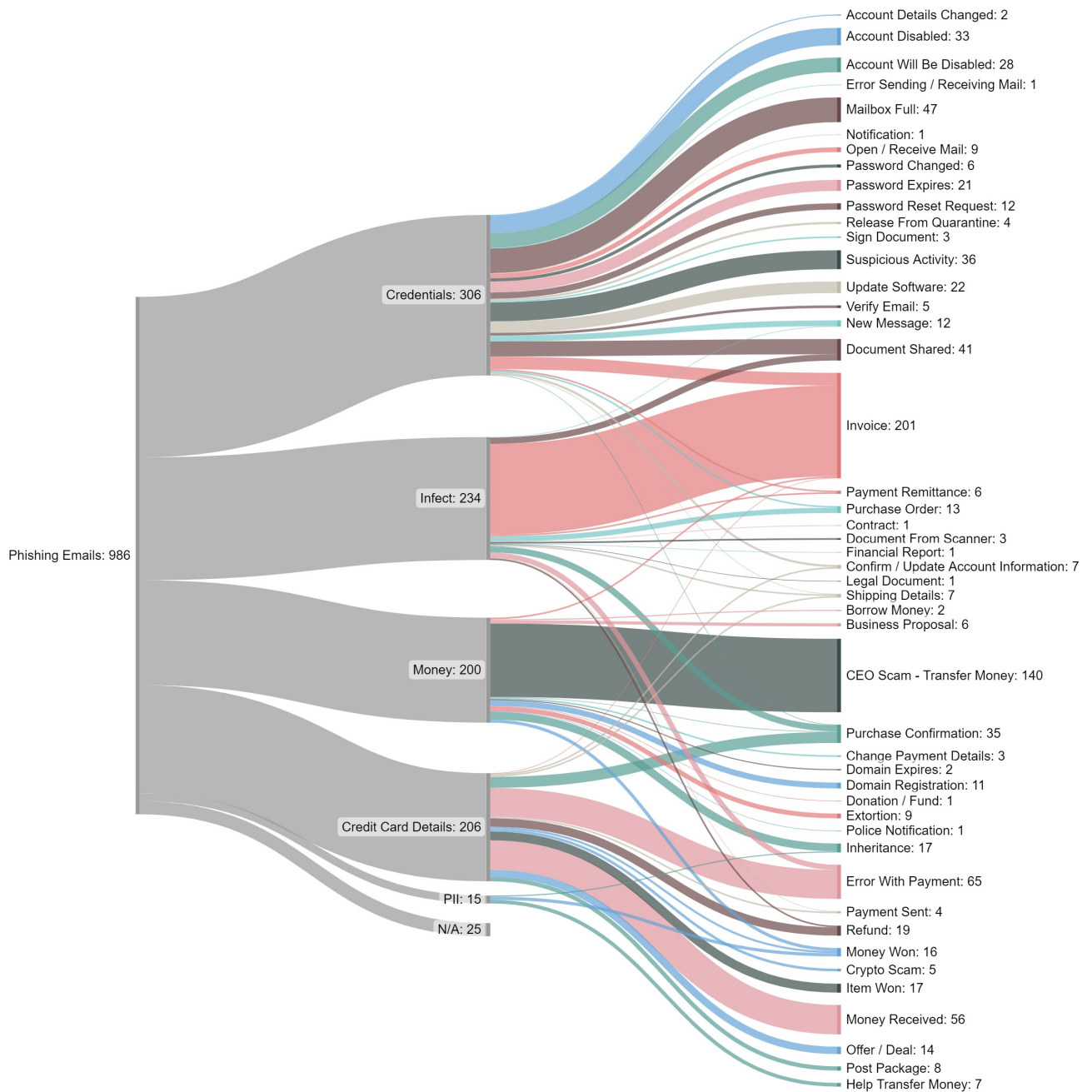


Figure A.11: 2017 - Target-Content Relationship

The diversity index shown in Table A.4, shows that Credentials now is the most diverse Target category, having a diversity index of 0.92. Infect has been reduced quite a bit, going from an index of 0.74 to a 0.45. The reduction can be tied to one Target category making up most of the observed infection phishing emails, that being Invoice.

Target	Total	Total Connections	$1 - \left(\frac{\sum n(n-1)}{N(N-1)} \right)$
Credentials	306	23	0.92
Money	200	13	0.50
Credit Card Details	13	11	0.83
Infect	234	14	0.45
PII	15	3	0.65

Table A.4: 2017 - Target-Content Diversity

The discrepancy between the Target categories has increased from last year, no longer showing the consistency observed before. The gap between the uppermost and lowest index score is now 0.47.

The Method-Target relationship, Figure A.12, continues the observation made last year in that the Targets have one method of achievement that is more preferred than the others. Credit Card Details is the only Target that is solely targeted through the utilization of one Method, URL, while Credentials and Infect are heavily targeted through URLs as well. Both PII and Money are still communication-heavy Targets.

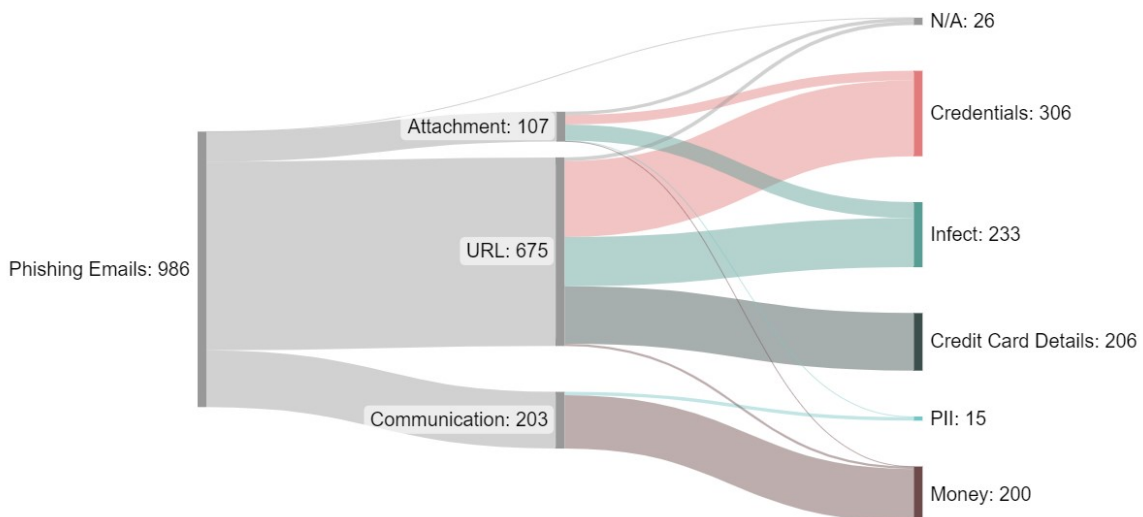


Figure A.12: 2017 - Method-Target Relationship

A.3 2018

2018 saw only a total of 523 reported phishing mails equaling a decline of 48.5% from the average of the two prior years.

A.3.1 Content

There was a total of 42 Content categories observed in 2018, including the three new categories CEO Scam – Gift Card, Guidelines, and Advert Stopped. 13 of the Content categories identified in the prior years’ datasets were not present in the 2018 collection. Figure A.13 visualizes the distribution of the 2018 phishing Content.

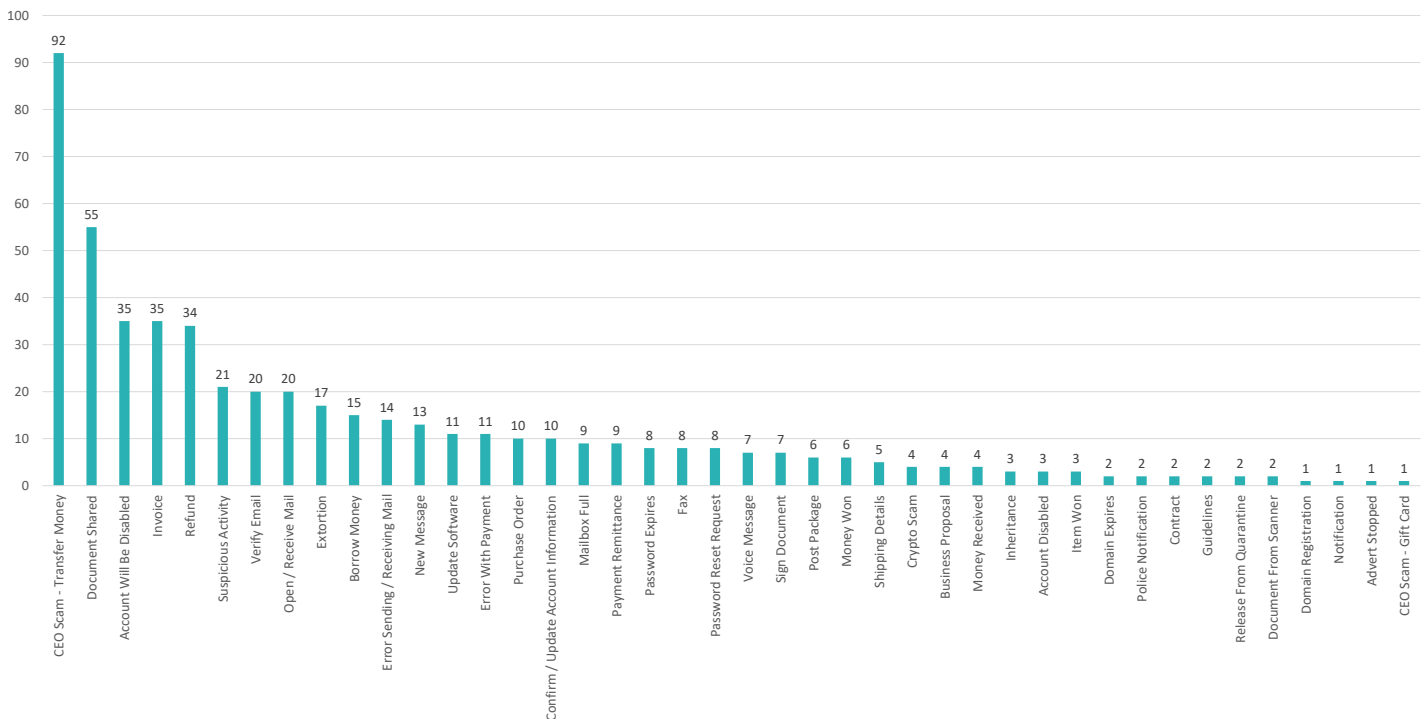


Figure A.13: 2018 - Content Distribution

Again, the uppermost category displays a clear discrepancy from the subsequent category. CEO Scam – Transfer Money has risen to the top in 2018, accounting for 17.59% of all the collected mails. Invoice, which was the 2017 top Content category, is still a heavily present category, placing fourth with a 6.69% presence.

A.3.2 Target

The distribution of Targets in 2018, as shown in Table A.5, shows a variation compared to the prior year with a range distribution more aligned with 2016. Creden-

tials remain the top Target, as with the previous years while Money makes up a large chunk of the observations as well. Infect, which was a prominent category in 2017, is reduced significantly from over 23% to less than 5% in 2018.

Target	Total	% of Total
Credentials	282	53.92%
Money	148	28.30%
Credit Card Details	45	8.60%
Infect	25	4.78%
N/A	16	3.06%
PII	6	1.15%
Business Information	1	0.19%

Table A.5: 2018 - Target Distribution

In inclusion to the changes in existing Target categories, a new category was observed. The Target of Business Information has been included in the overview of observed phishing Targets.

A.3.3 Method

As has been seen in the previous years, the distribution of Methods remains the same in terms of ranging and percentage, with URL at the top followed by Communication and Attachment. There is also a new observed Method in the form of a calendar invite.

Method	Total	% of Total
URL	324	61.95%
Communication	146	27.92%
Attachment	48	9.18%
N/A	4	0.76%
Calendar Invite	1	0.19%

Attachment Type	Total	% of Total
PDF	27	5.16%
Word	11	2.10%
ACE	3	0.57%
HTML	3	0.57%
ZIP	3	0.57%
ISO	1	0.19%

Figure A.14: 2018 - Method Distribution

The distribution of Attachments sees that PDF remains the most observed attachment type as well as Word as the second most observed. One new attachment type in the form of ACE was also identified within the phishing corpus.

A.3.4 Impersonation

Since the generic categories makes up the majority of the Impersonation shares, the non-generic categories is represented in a zoomed in pie piece. Figure A.15 showcases the distribution of Impersonations observed in the 2018 corpus. Internal and External collectively comprises close to 70% of the mails, while the remaining ~30% consists of the brands observed.

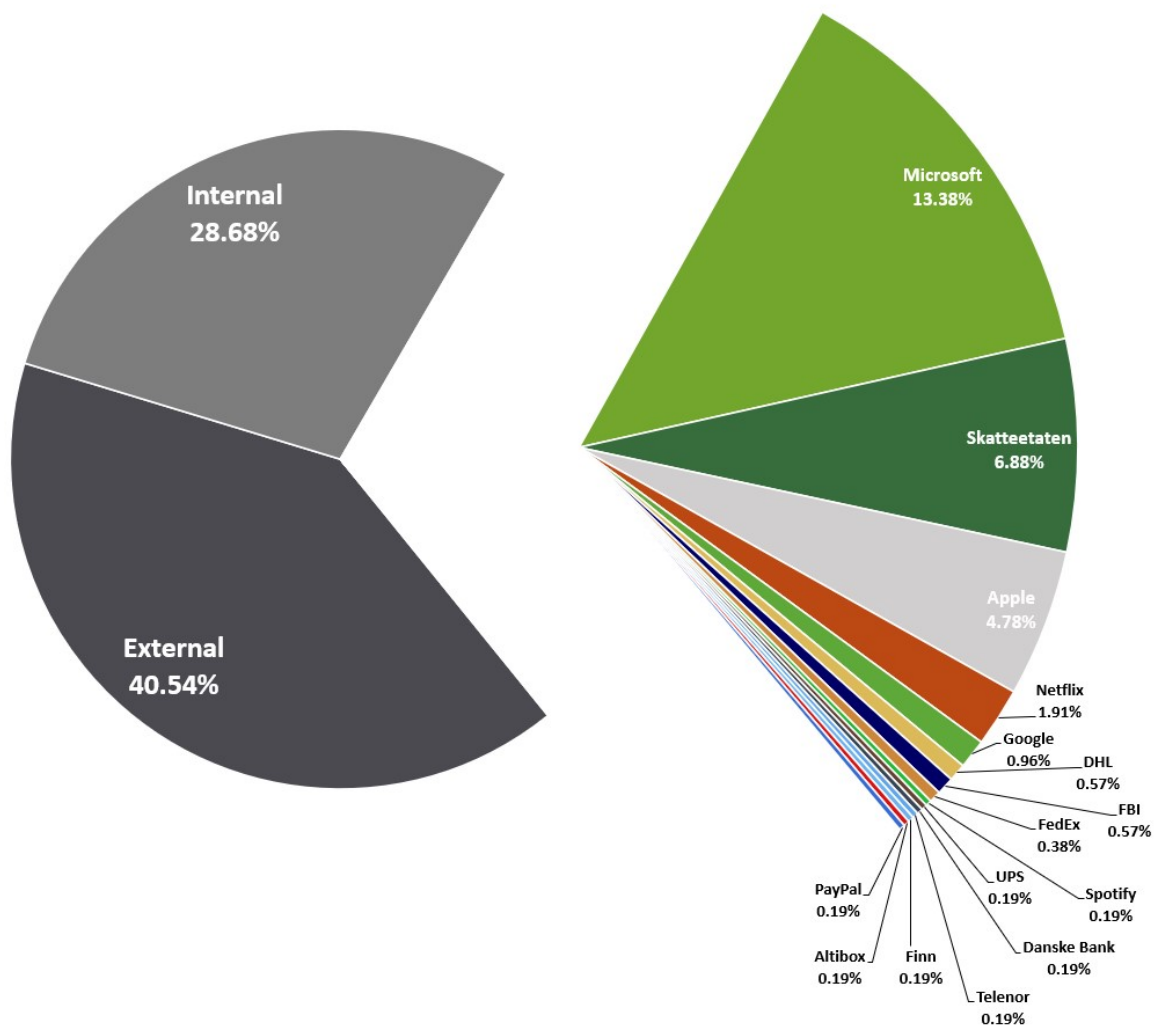


Figure A.15: 2018 - Implementation Distribution

As shown, the usage of Microsoft's brand has grown considerably, while Apple, Netflix, and Skatteetaten still remains highly utilized.

A.3.5 Dates

The heat-map for 2018, Figure A.16, further displays no significant patterns with when phishing mails are being observed in the collected datasets.

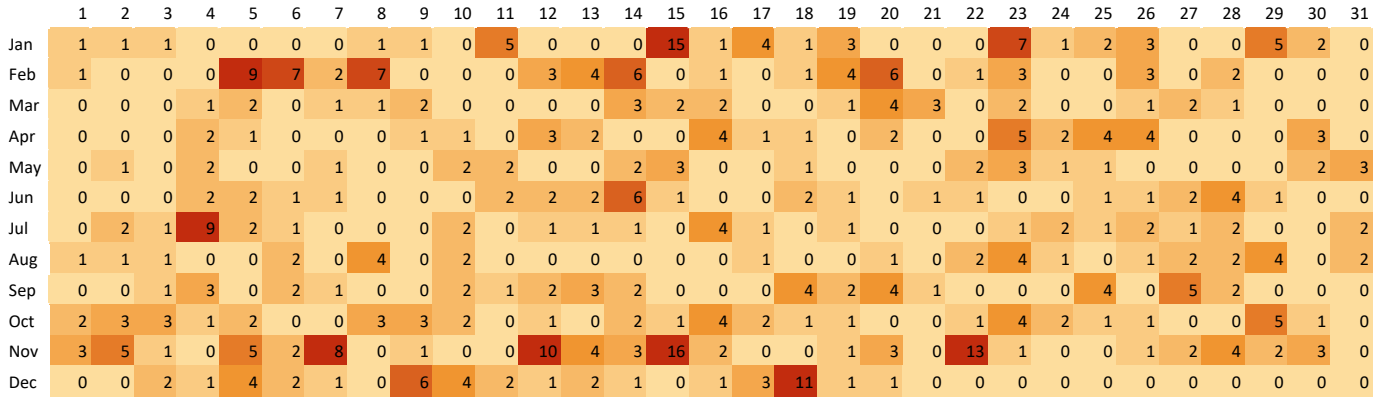


Figure A.16: 2018 - Date Distribution

November of 2018 has the most observed mails compared to the other months, however not a considerably difference.

A.3.6 Property Relationships

Out of the 41 Content categories observed in 2018, 27 of them had singular relationships with one Target. Of these 27 Content categories, 16 were tied to the Credential Target category. The Invoice Content category continues its trend from last year by having a relationship with four Targets, this time Credentials, Infect, Business Information and Money, making it the only Content category in 2018 with this many Target connections. Figure A.17 visualizes the Content-Target relationships observed for this year.

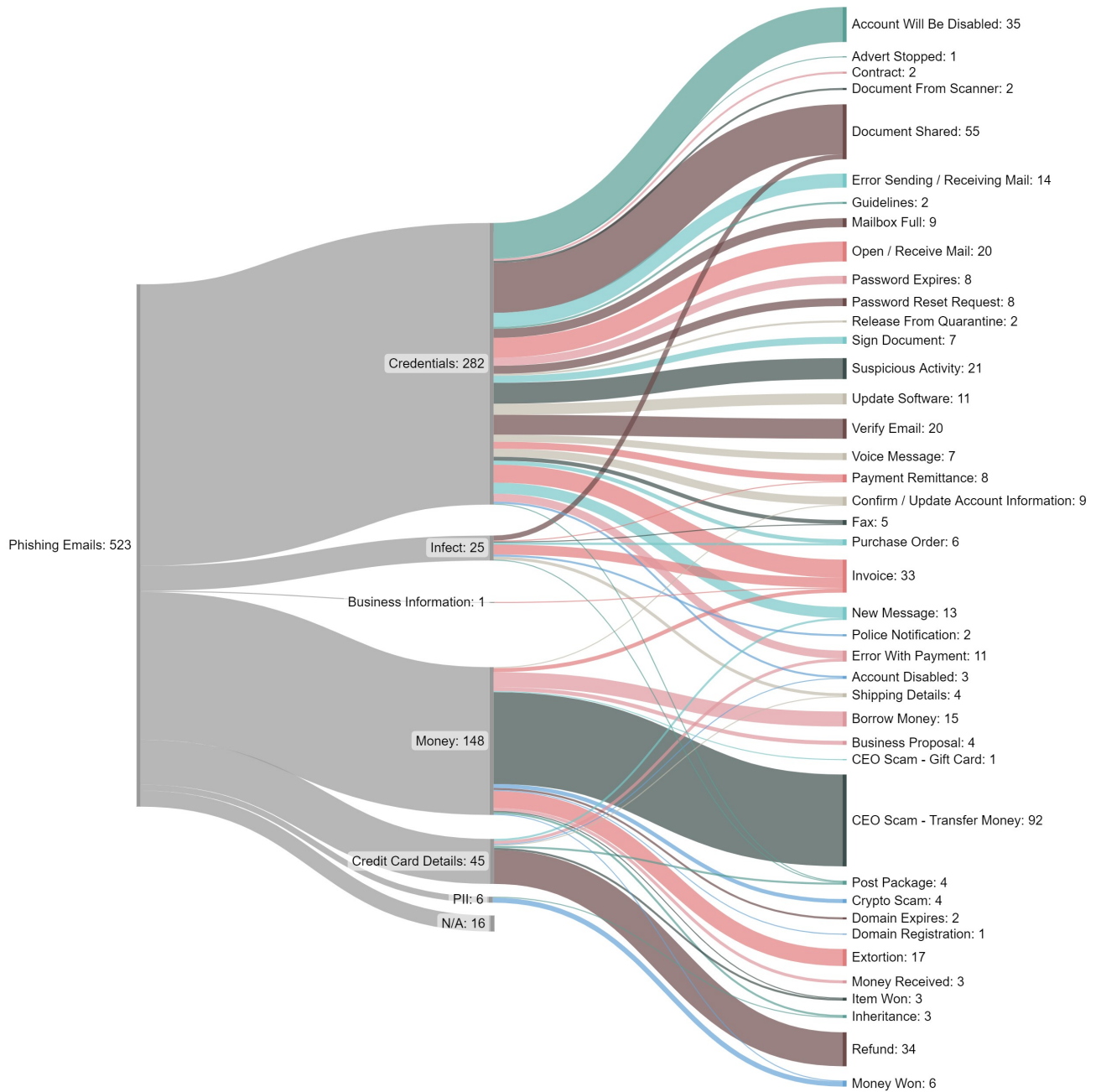


Figure A.17: 2018 - Target-Content Relationship

Table A.6 shows that Credentials continues to be the most diverse Target category, having a diversity index of 0.92 (same as last year). On the other hand, since Business Information only has one entry, its diversity is equal to zero. PII has seen a great reduction, going from being the most diverse in 2016 with an index of 0.84, to a 0.65 last year, and a 0.33 this year.

Target	Total	Total Connections	$1 - \left(\frac{\sum n(n-1)}{N(N-1)} \right)$
Credentials	282	26	0.92
Money	148	14	0.59
Credit Card Details	45	7	0.43
Infect	25	8	0.80
PII	6	2	0.33
Business Information	1	1	0

Table A.6: 2018 - Target-Content Diversity

Figure A.18 shows that the Method-Target relationship has not changed too much since the last two years. Attachment is now the method mainly utilized for infections from being URLs last year. The remainder of the relationships have stayed mostly the same, with the addition of the Calendar Invite Method and Business Information Target.

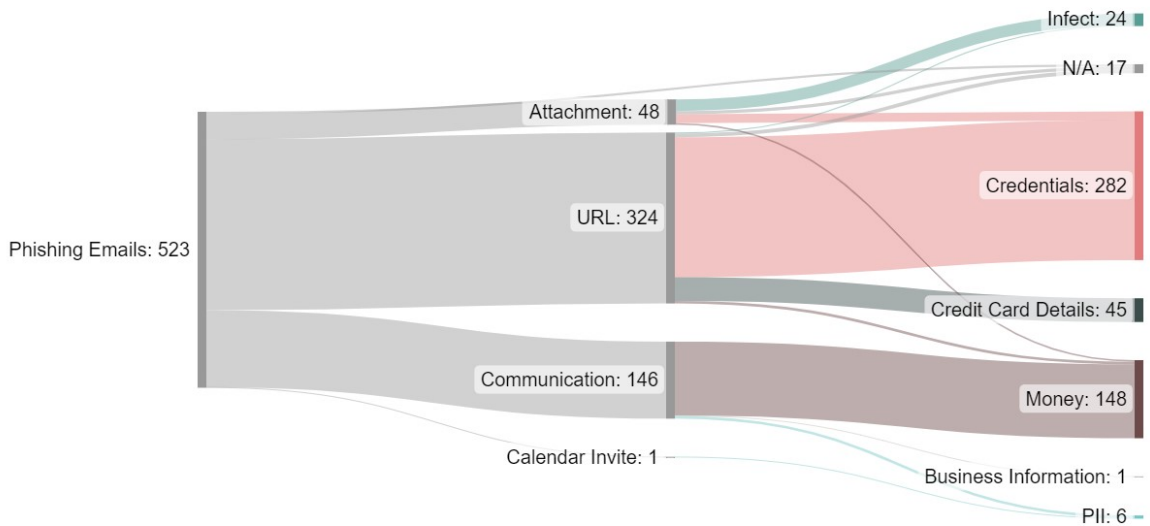


Figure A.18: 2018 - Method-Target Relationship

A.4 2019

720 mails in total were collected from 2019.

A.4.1 Content

The mails collected in 2019 can be distributed into 46 distinct Content categories, where three of the categories had not previously been identified. These new categories include Added To Group, Job Application, and MFA Activate.

Figure A.19 displays the distribution of Content in the 2019 corpus.

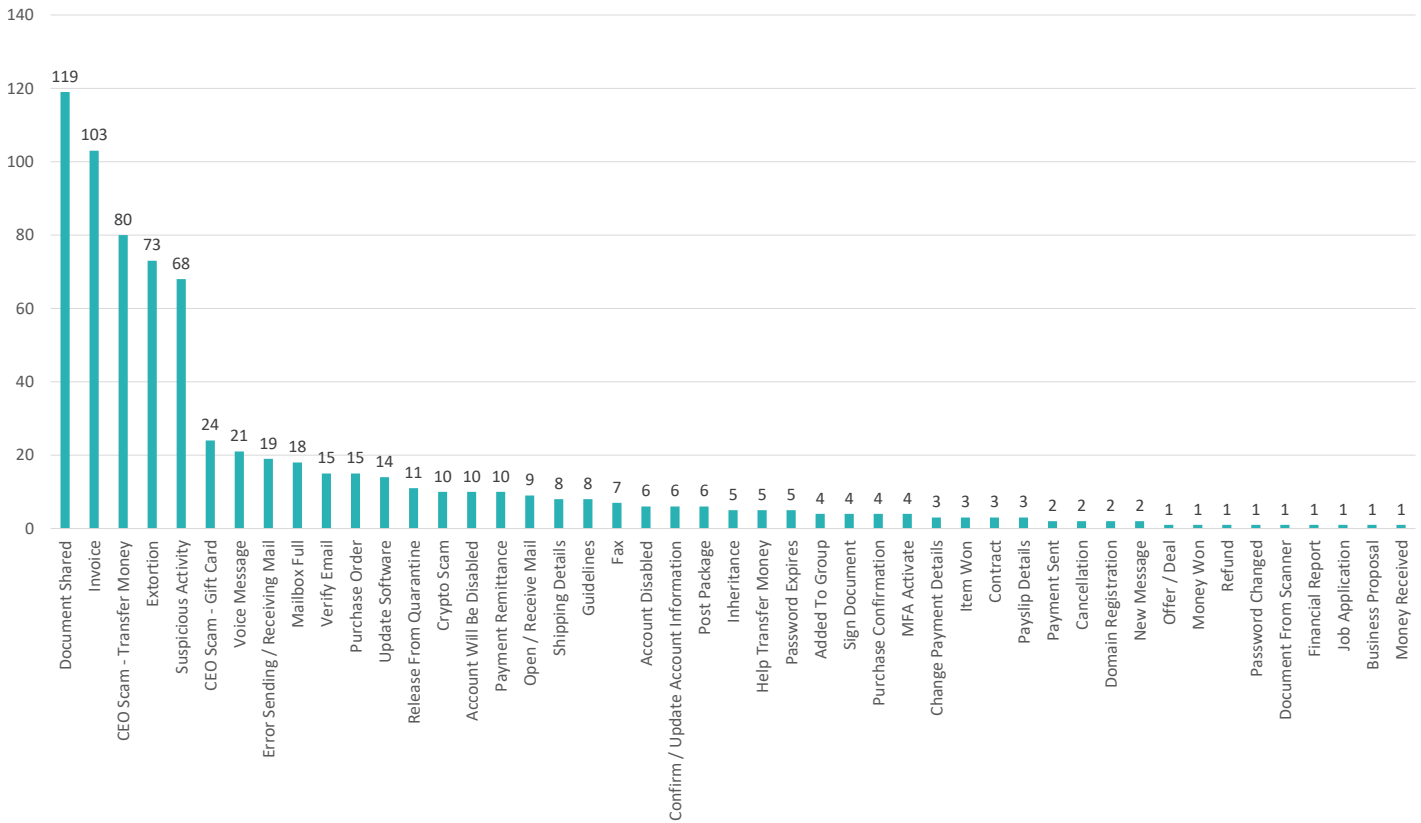


Figure A.19: 2019 - Content Distribution

Again, there is a new category representing the most observed Content of 2019, this time in the form of Document Shared. The category has been fairly present throughout the years and does not represent a large surge from its previous observations. Invoice as well as CEO Scam – Transfer Money remain high ranking Content categories as shown in the prior years.

A change from the patterns observed in 2016, 2017 and 2018 is that the discrepancy between the uppermost category and the following category is not as significant for this year as it has been before. Only a 13.5% discrepancy is shown between the first and second category compared to an average of 36.6% for the prior years. There is however a considerable gap between the top five categories and the remaining categories.

A.4.2 Target

Credentials and Money remain the top targeted for 2019, while the changes in the remainder of the categories does not show any significant changes other than Credit Card Details and Infect changing positions due to a decline in the targeting of Credit Card Details.

Target	Total	% of Total
Credentials	460	63.89%
Money	203	28.19%
Infect	34	4.72%
Credit Card Details	11	1.53%
N/A	7	0.97%
PII	3	0.42%
Business Information	2	0.28%

Table A.7: 2019 - Target Distribution

A.4.3 Method

A clear pattern can be seen when it comes to the Methods observed in the collected phishing datasets. URL ranks the highest always within 60% of the total amount, while Communication is second always within 20% of the total, and thirdly, Attachment with 9 – 11% of the total. For the Attachments, it follows a similar pattern as from 2018 with PDF accounting for the majority of the attachments types. Two new attachment types are observed as well, that being ARJ and EXE.

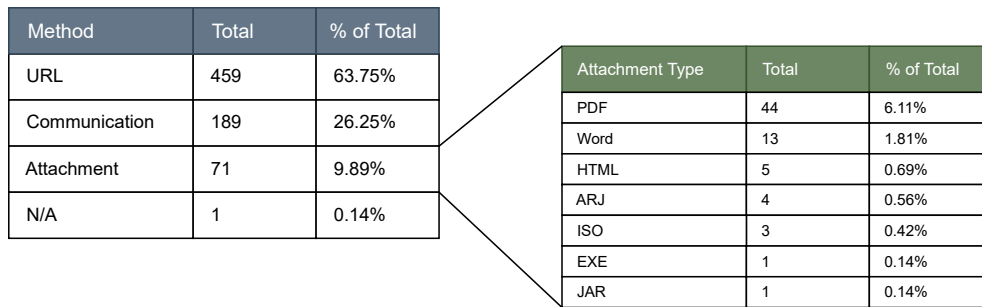


Figure A.20: 2019 - Method Distribution

A.4.4 Impersonation

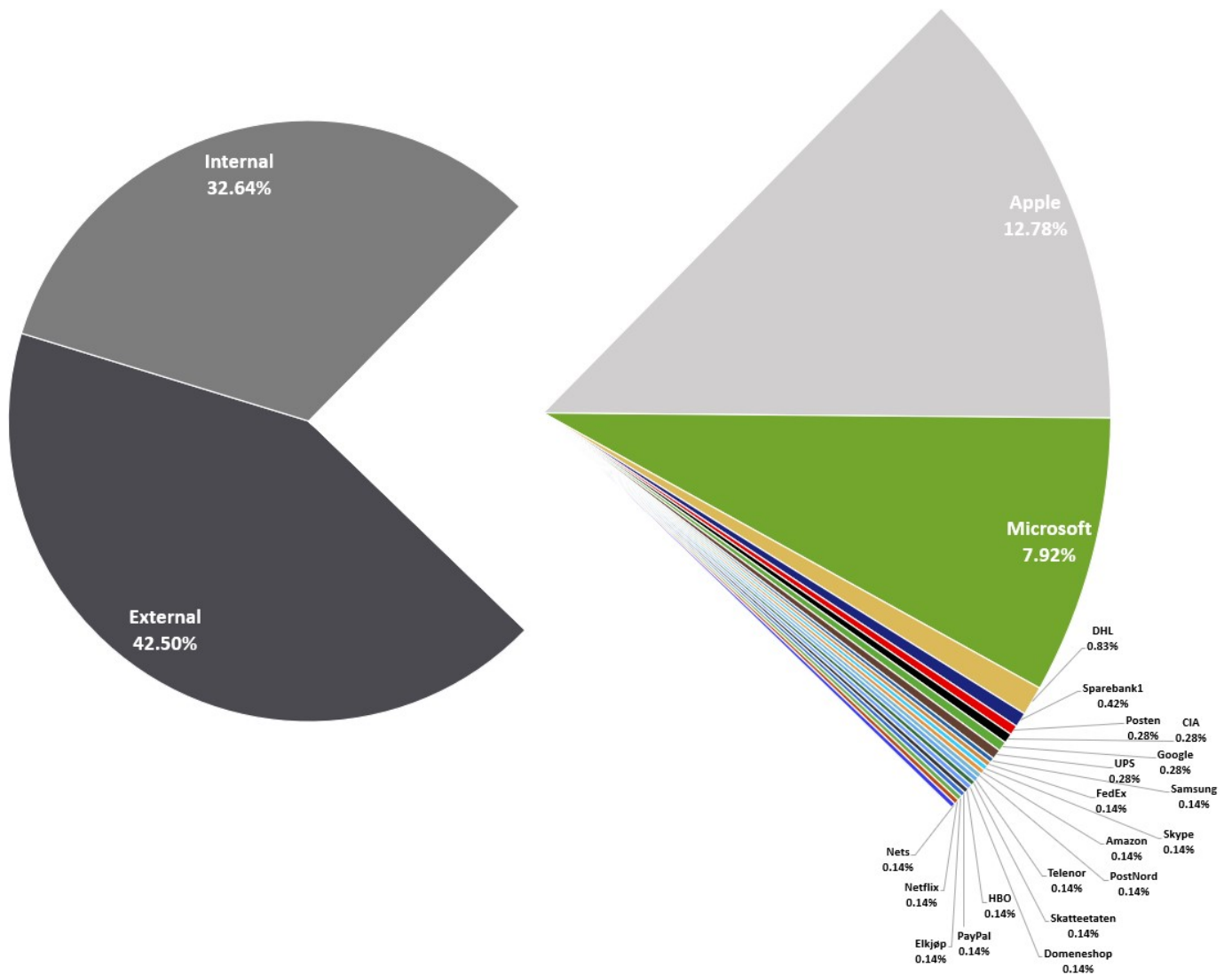


Figure A.21: 2019 - Impersonation Distribution

The generic Impersonation categories of External (42.50%) and Internal (32.64%) accounts for a total of over 75% of the observed phishing mails in 2019. The remaining shows a high count of usage of the Apple and Microsoft brand, as has been seen in the earlier years. Skatteetaten, which had a prominent part in the chart of last year’s overview, has been reduced significantly, only being used in one singular email this year.

A.4.5 Dates

The distribution of dates can be viewed in Figure A.22. The heat-map does not highlight any one specific time frame in which there has been significantly more phishing mails compared to the rest of the year. There are some surges in the September, October, and November months, as well as a surge one day in April. However it evens out in the adjacent days.

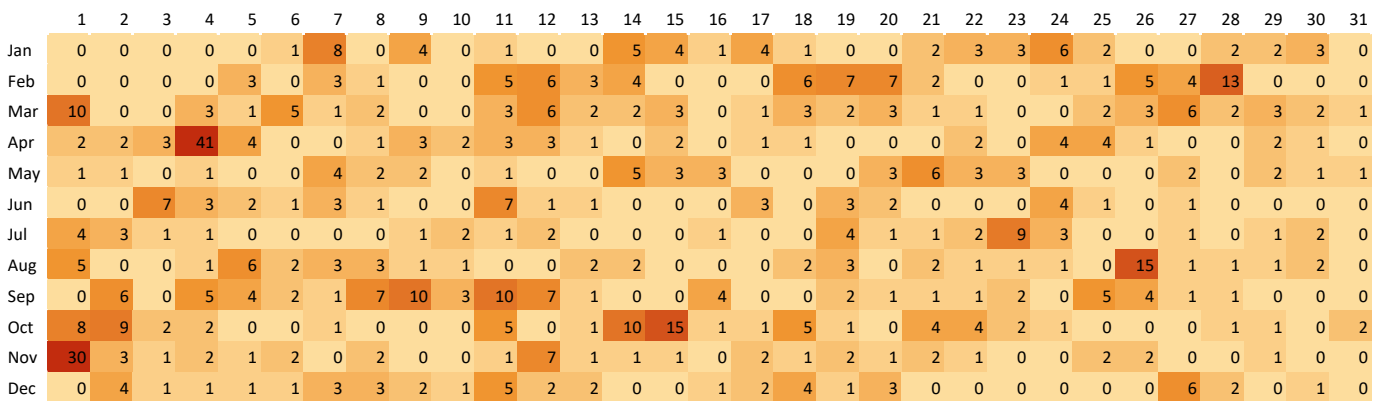


Figure A.22: 2019 - Date Distribution

A.4.6 Property Relationships

Figure A.23 continues to show that the Credential Target harbors the most one-to-one relationships, accounting for 22 of the total of 33 singular relationships. Invoice remains the Content category with the most unique Target connections, being tied to Credentials, Business Information, Infect, and Money.

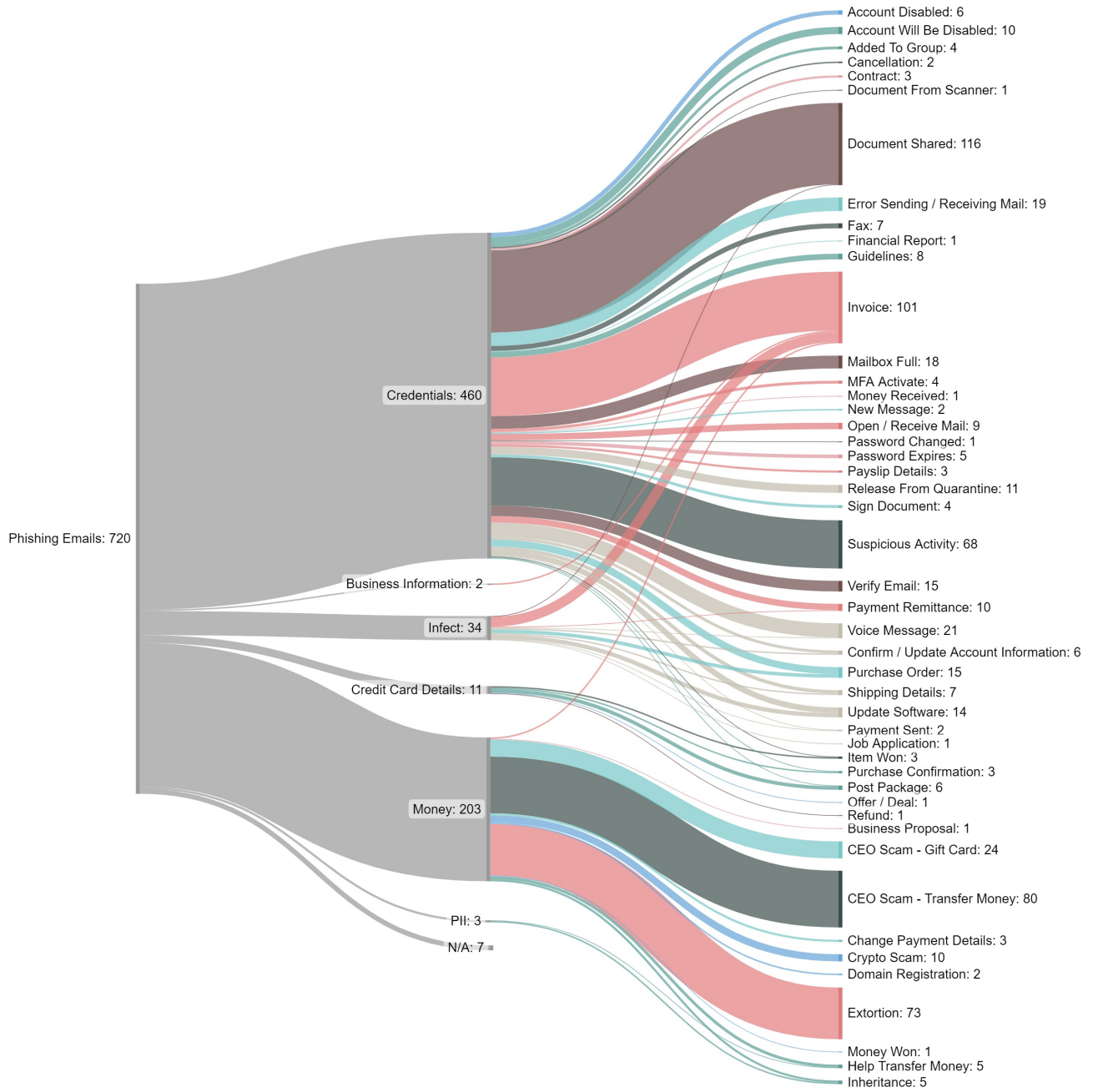


Figure A.23: 2019 - Target-Content Relationship

The diversity of the Target categories, with the exclusion of Business Information, shows a more even diversity distribution than that of 2018. Credentials is still the most diverse Target with an index of 0.87.

Target	Total	Total Connections	$1 - \left(\frac{\sum n(n-1)}{N(N-1)} \right)$
Credentials	460	34	0.87
Money	203	11	0.70
Credit Card Details	11	5	0.78
Infect	34	10	0.79
PII	3	2	0.67
Business Information	2	1	0

Table A.8: 2019 - Target-Content Diversity

Due to Business Information only being tied to one Content category, it displays no diversity at all.

A couple of trends can now be established for some of the Method-Target relationships, as they have displayed a consistency throughout the four last years. The observations from this year’s corpus, shown in Figure A.24, shows that URL is persistently the preferred Method utilized to target Credentials, while Communication is preferred when targeting money and PII.

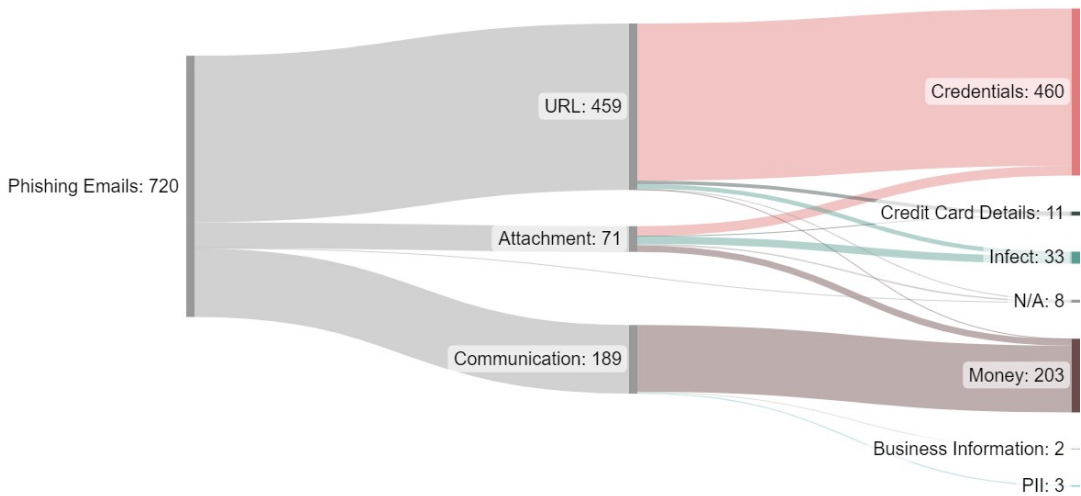


Figure A.24: 2019 - Method-Target Relationship

Credit Card Details, which up until now had only been targeted through email links, is now split between URLs and Attachments, no longer conforming to the previous observations. Infect continues to be somewhat a Target that is attempted achieved through either URLs or Attachments, varying in distribution from year to year.

A.5 2020

There was a total of 3181 phishing mails collected for 2020. The increase in the total number of mails is due to the introduction of a new reporting system for phishing mails in August of 2020. This reporting system made it easier for the end users to report suspicious mails, causing the number of reported mails to increase drastically. In total, there were 561 mails from the former months and 2621 from the latter. Because of this change, the numbers from the first seven months of the year can not directly be compared to the other five months, as the latter would cause an uneven representation favoring the last five months. Due to this challenge, the data from 2020 will be represented in two separate parts, one for January throughout July and one for August throughout December.

A.5.1 Content

2020 saw a total of 55 Content categories, four of which had not been observed before. These four included Job Write-Up, Meeting Invitation, New Task, and Server Stopped.

Figure A.25 displays the Content distribution from January throughout July for 2020.

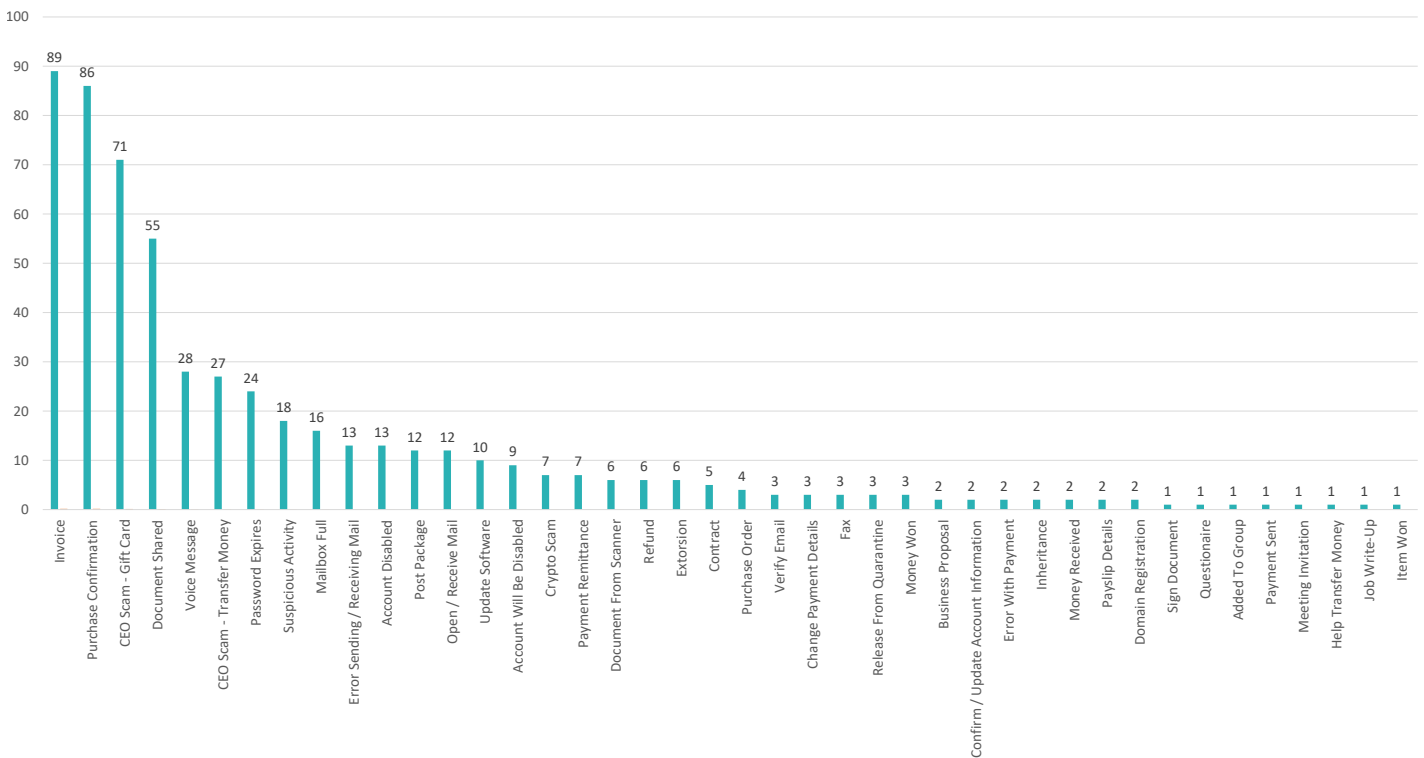


Figure A.25: 2020 1st - Content Distribution

As can be seen, Invoice is at the top one step up from 2019, while Purchase Confirmation is the second most observed Content category. Purchase confirmation has seen a great increase in presence since the prior years, seeing as it was not present at all in the 2018 corpus, and only four instances in 2019.

Further, Figure A.26 displays the Content distribution for the five last months of 2020. Immediately, based on the data labels, it is apparent that the dataset has grown significantly from the former seven months. There is also a significant change in the distribution of Content categories compared to the first five months. Both the top Content categories, Post Package and Crypto Scam, were ranged less than 10th from the January - July dataset. There is also a considerable difference in the discrepancy between the top categories. For the seven first months, the discrepancy between the first and third was 3.2, while for the last five months the discrepancy was 11.1.

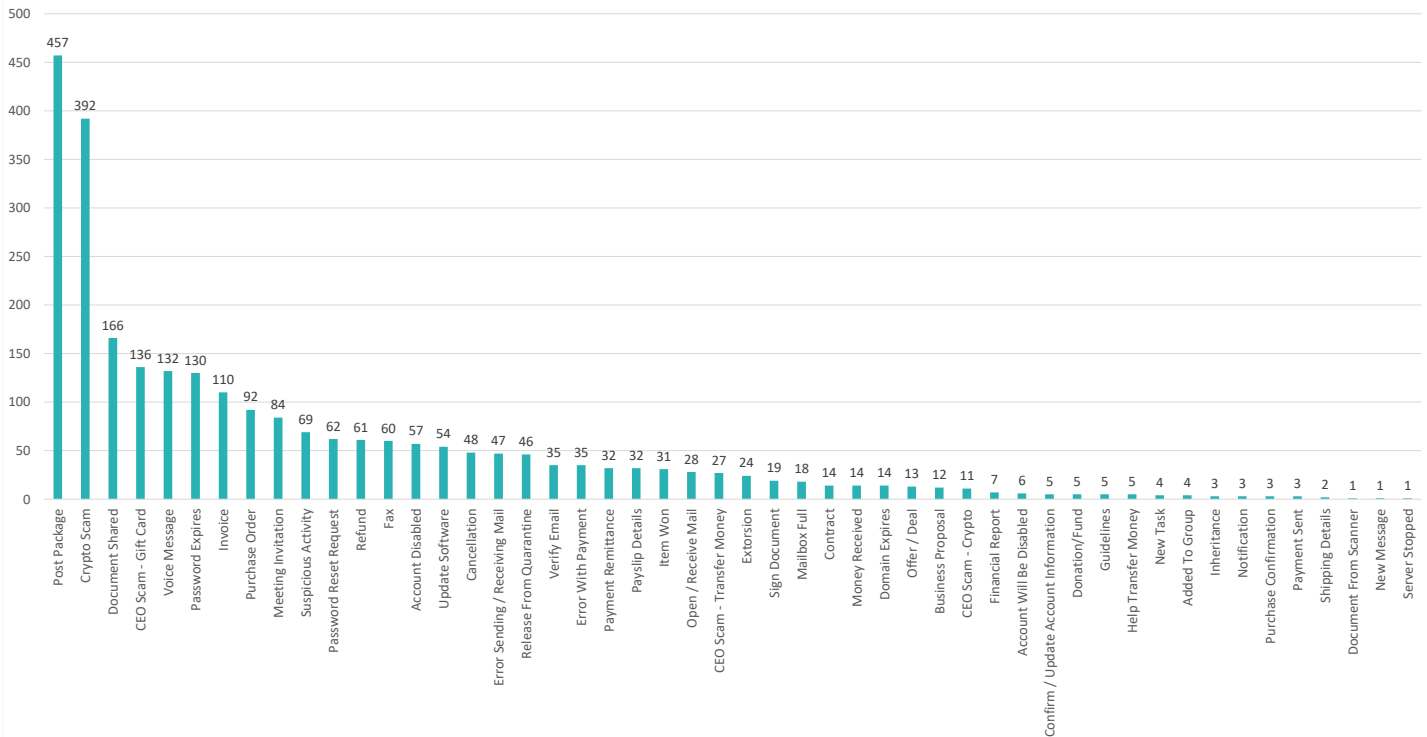


Figure A.26: 2020 2nd - Content Distribution

A.5.2 Target

The Target distribution of both the first group (Table A.9) and last group (Table A.10) of 2020 are quite equal in terms of ranging. Credentials, Money, Credit Card Details, and Infect are all equally ranged, where the top two match that of 2019. There is however a difference between the percentwise distribution of the categor-

ies with a 72.55% representation from Credentials in the first group compared to a 49.96% in the second group.

Target	Total	% of Total
Credentials	407	72.55%
Money	124	22.10%
Credit Card Details	23	4.10%
Infect	4	0.71%
N/A	2	0.36%
PII	1	0.18%

Table A.9: 2020 1st - Target Distribution

The most significant difference between the two groups in 2020 is the increase in observation of the Credit Card Details category, increasing from 4.10% to 22.79%, making the second group more similar to the distribution observed in 2016 and 2017, while the former group is more similar to 2018 and 2019.

Target	Total	% of Total
Credentials	1309	49.96%
Money	615	23.47%
Credit Card Details	597	22.79%
Infect	91	3.47%
PII	6	0.23%
Business Information	2	0.08%

Table A.10: 2020 2nd - Target Distribution

A.5.3 Method

For the first group, as represented in Figure A.27, the ranging of the distribution remains the same as the prior years with URL at the top followed by Communication and Attachment. There is a slight increase in the percentwise distribution in favor of the URL category, however not significant enough to deviate considerable from what has been observed earlier. There are fewer observed attachment types, and as can be seen, HTML now has increased in representation from the prior years.

Method	Total	% of Total
URL	399	71.12%
Communication	118	21.03%
Attachment	44	7.84%

Attachment Type	Total	% of Total
HTML	31	5.53%
PDF	10	1.78%
Excel	2	0.36%
ZIP	1	0.18%

Figure A.27: 2020 1st - Method Distribution

The second group, as represented in Figure A.26, further increases the representation of the URL Target category, encompassing 82.10% of all the Targets observed. The distribution also shows a breach in the pattern observed in the datasets of the prior years by ranking Attachments higher than Communication.

Method	Total	% of Total
URL	2151	82.10%
Attachment	242	9.24%
Communication	227	8.66%

Attachment Type	Total	% of Total
HTML	143	5.46%
ZIP	41	1.56%
PowerPoint	29	1.11%
R04	18	0.69%
PDF	9	0.34%
Word	2	0.08%

Figure A.28: 2020 2nd - Method Distribution

There were observed two new Attachment types: PowerPoint and R04. HTML still maintains a great representation as it had the first part of the year, while PDF shows a decline in favor of an increase in the observation of ZIP files.

A.5.4 Impersonation

The Impersonation distributions for both the first (Figure A.29) and second (Figure A.30) part of 2020 shows a decline in the percenwise distribution of the generic categories. In 2019 the generic Impersonation categories accounted for 75.14% of the total observations, while in the first and second part of 2020 they accounted for 56.68% and 51.68% respectively, putting its distribution closer to that of 2016.

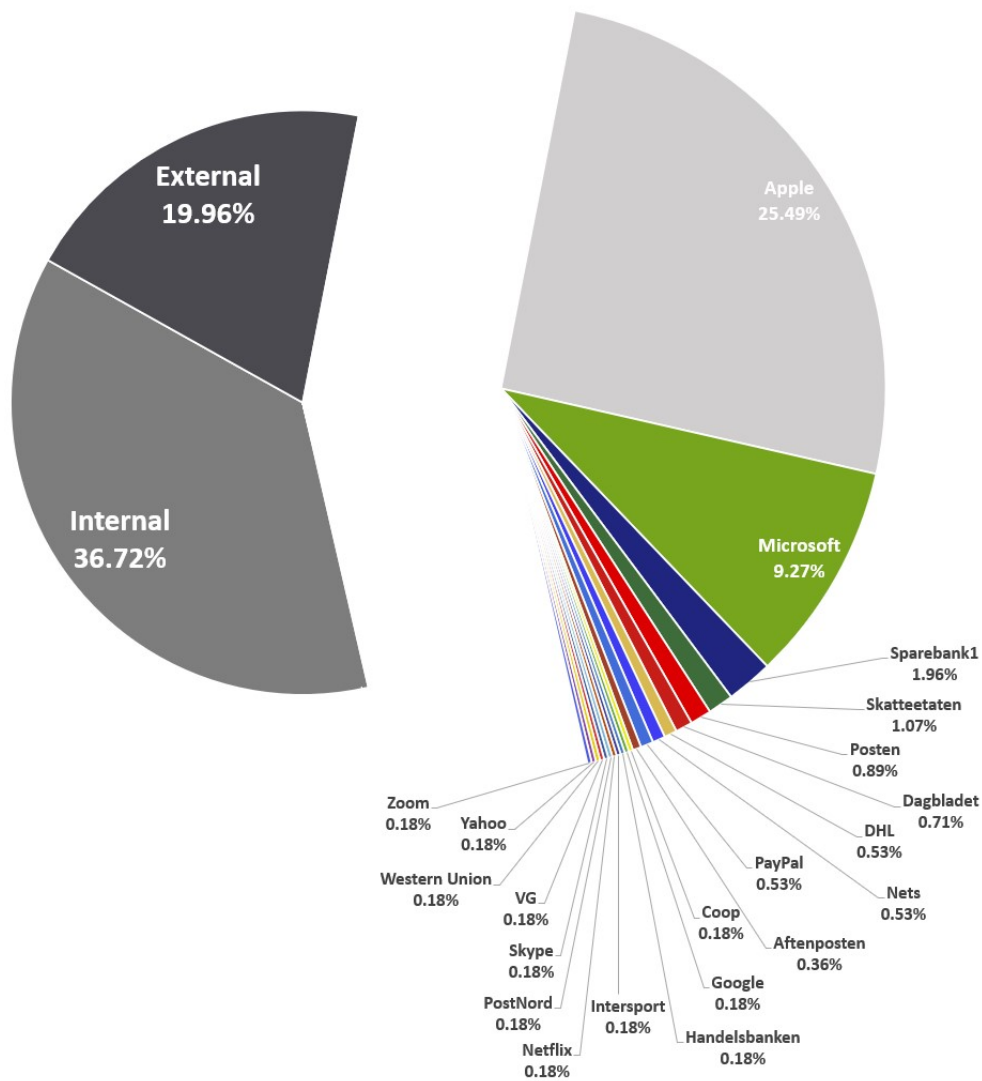


Figure A.29: 2020 1st - Impersonation Distribution

Apple and Microsoft continues to be prevalent in the first part of 2020, while the second part brings forth previously low ranging impersonations like Posten and Dagbladet.

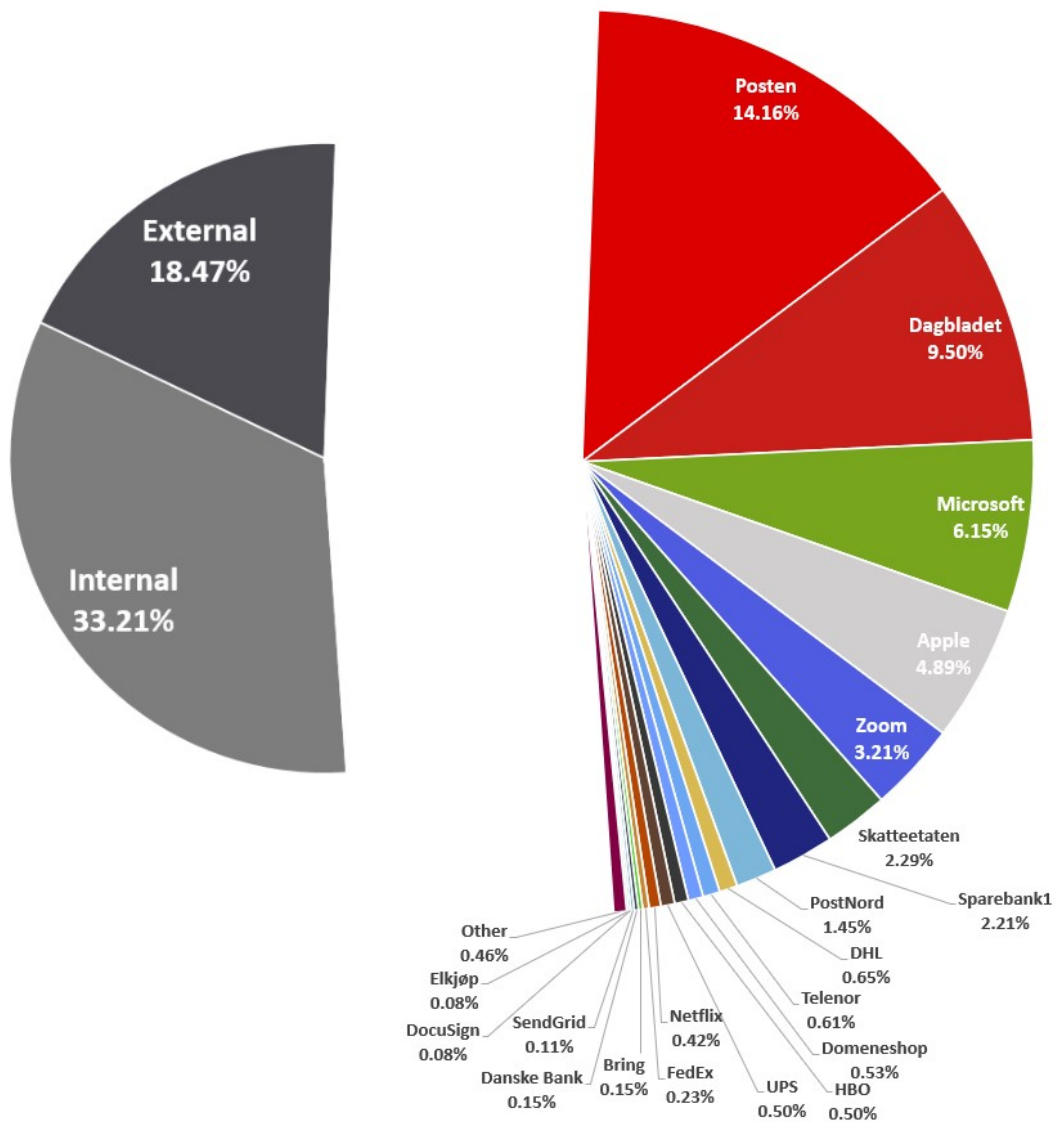


Figure A.30: 2020 2nd - Impersonation Distribution

A.5.5 Dates

The date distribution for the first part of 2020, Figure A.31, shows no particular surges in certain time frames, only some increased activity at the end of March. There is one singular day surge on the 18th of June totaling 68 mails. Inspecting the dataset on that specific date reveals that the surge is related to a supposed phishing campaign categorised as a Purchase Confirmation Content category, impersonating Apple and targeting Credentials through the usage of URLs. The activity appears to dwindle down at the end of June and throughout July.

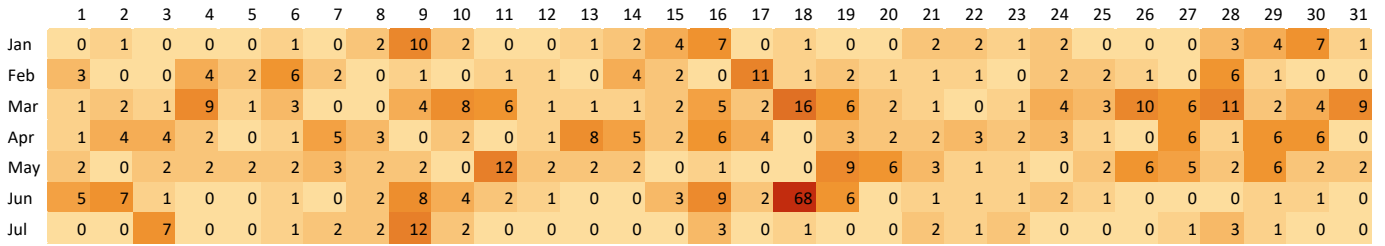


Figure A.31: 2020 1st - Date Distribution

The heat-map for the second part of 2020, Figure A.32, showcases increased activity in the month of November, as well as at the beginning of December. As can be seen, there is generally little activity in the month of August and beginning of September.

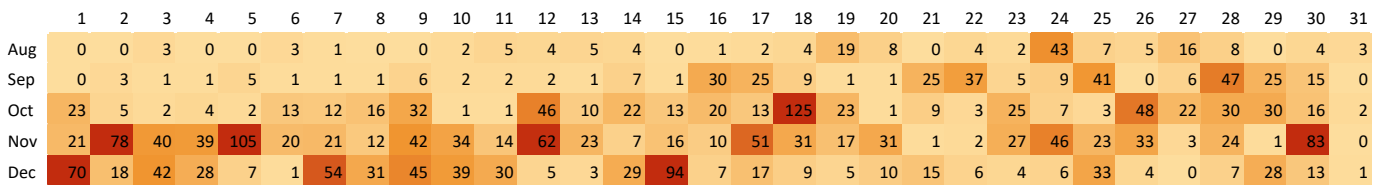


Figure A.32: 2020 2nd - Date Distribution

A.5.6 Property Relationships

Figure A.33 visualizes the Target-Content relationship of the 2020 dataset. Credentials has a total of 29 singular relationships with Content categories, while the remainder is divided on Money (8), Credit Card Details (5), and Infect (1). No Content category has more than four unique Target connections, with Invoice continuing to be the only one with four connections, connecting to Credentials, Business Information, Money, and Infect.

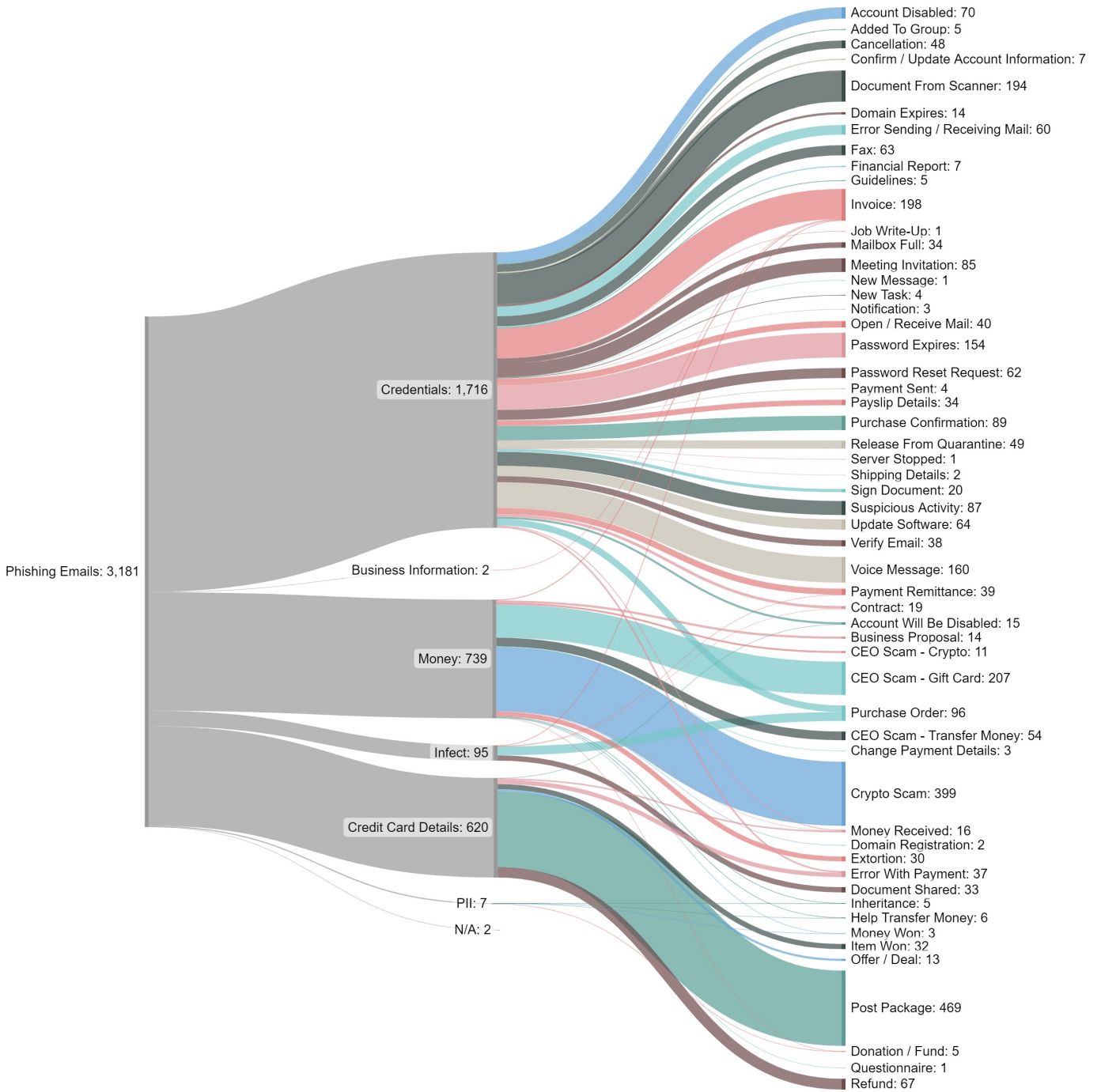


Figure A.33: 2020 - Target-Content Relationship

The diversity index, as shown in Table A.15, displays a larger discrepancy between the Targets than what was observed last year. Credentials remains the most diverse Target, now with an index as high as 0.94. As Business Information still only con-

nects to one Content category, it remains with a diversity score of 0. Out of the other Targets, Credit Card Details has seen the greatest change, going from a 0.78 in 2019 to a 0.41 as of this year. From Figure A.33, this change can be correlated with the increase in the observation of the Post Package Content category.

Target	Total	Total Connections	$1 - \left(\frac{\sum n(n-1)}{N(N-1)} \right)$
Credentials	1716	38	0.94
Money	739	14	0.62
Credit Card Details	620	8	0.41
Infect	95	5	0.55
PII	7	4	0.81
Business Information	2	1	0

Table A.11: 2020 - Target-Content Diversity

The Method-Target relationship, Figure A.34, has seen changes from the observations of last year. Money is no longer majorly targeted through the utilization of Communication, however is now split rather evenly between that and URLs. Credit Card Details is back at only being targeted through email links, and Business Information continues solely being tied to Communication.

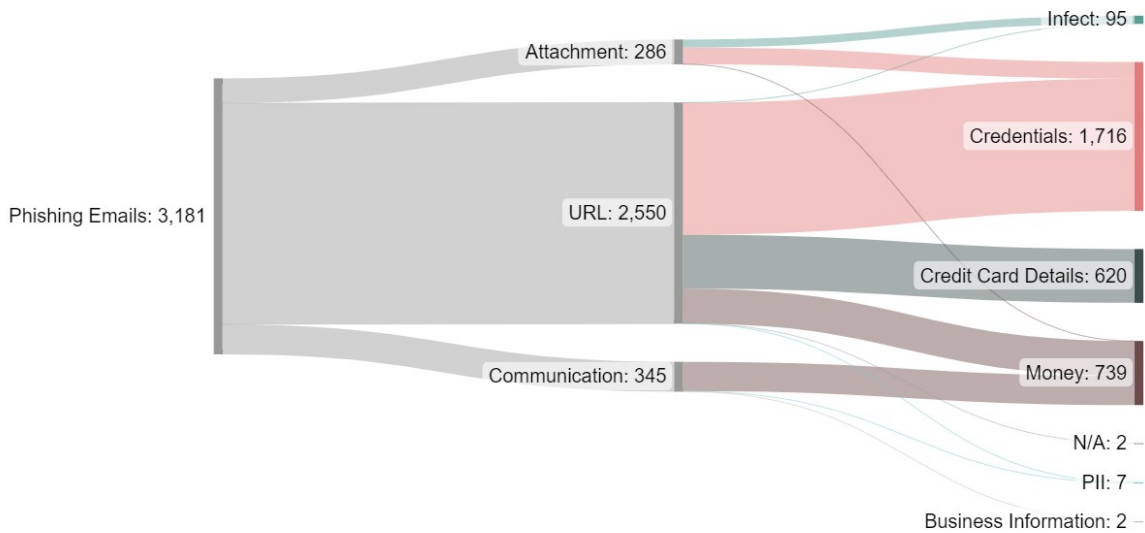


Figure A.34: 2020 - Method-Target Relationship

What hasn't changed is the fact that URLs remains the main Method utilized in order to lure out the recipient's credentials.

A.6 2021

A total of 16202 phishing mails were collected for 2021.

A.6.1 Content

In total, there were 61 Content categories observed in 2021. Out of these 61, three were entirely new, including Calendar Event, Donate Money, and Trademark.

The Content distribution shows that Post Package appears at the top in 2021, as it was at the end of 2020. The Post Package Content category has shown a great increase the last two years with only a 0.83% presence in 2019, a 2.14% presence the first part of 2020, rising to a 17.44% the second part of 2020, and now close to 14% total in 2021.

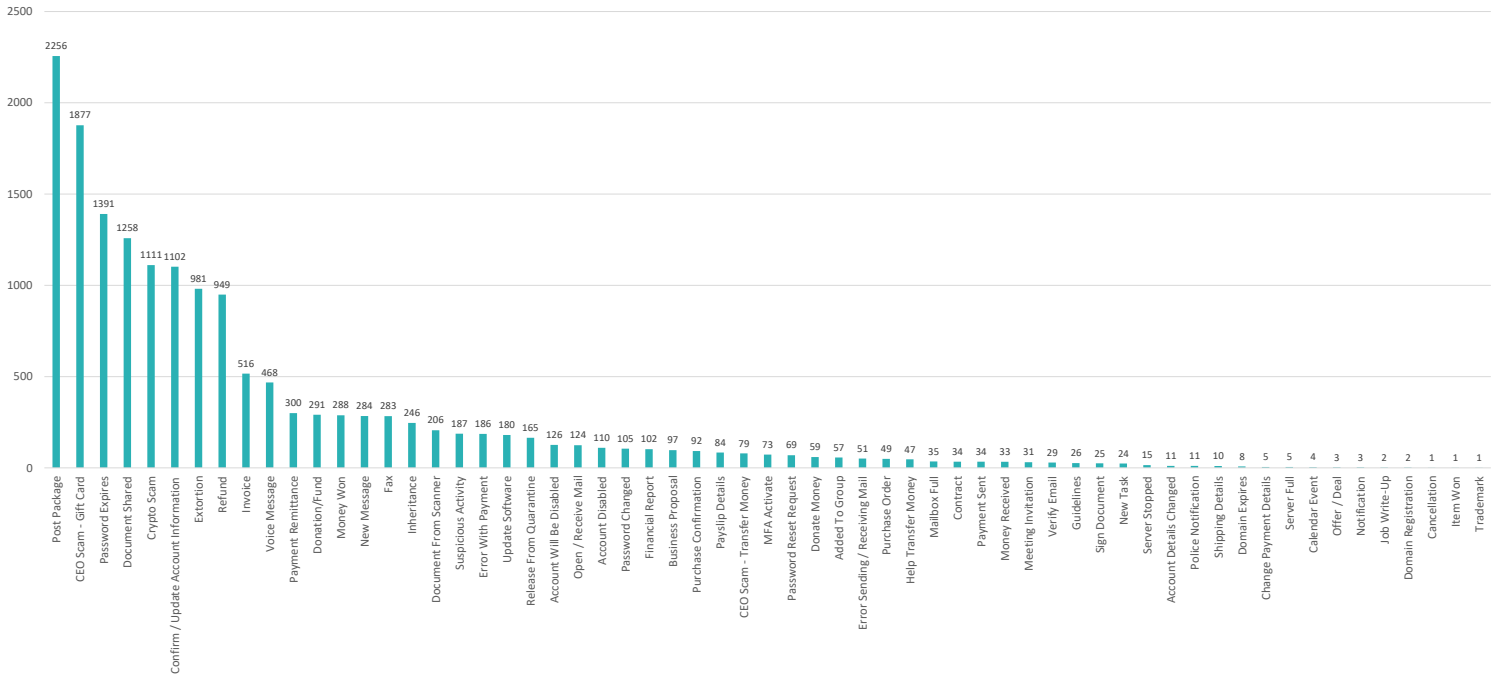


Figure A.35: 2021 - Content Distribution

CEO Scam – Transfer Money continues to reduce in quantity, straying further away from the top where it once stood.

A.6.2 Target

Table A.12 shows a continuation of the increased targeting of Credit Card Details, while also highlighting a significant decrease in the targeting of Credentials. The targeting of Credentials has decreased from a percentwise distribution of 72.55% in the first part of 2020, to a 49.96% in the second part of 2020, and now only at a 37.08% representation in 2021.

Money and Credit Card Details have an incredibly even distribution considering they account for a bit over 60% of the total Targets, only differentiating with 2.62%.

Infect, as well, has seen a small decline, where personal information has seen a slight increase. These increases and decreases respectively has led to PII surpassing Infect in the ranging.

Target	Total	% of Total
Credentials	6007	37.08%
Money	4953	30.57%
Credit Card Details	4853	27.95%
PII	214	1.32%
Infect	135	0.83%
N/A	36	0.22%
Business Information	4	0.02%

Table A.12: 2021 - Target Distribution

A.6.3 Method

As before, the distribution of Methods from the ranging is back to the same as the prior years from a slight change in the second part of 2020, however with a decrease in the percentwise total for URL. The decrease in URL is evened out with an increase in the utilization of Communication. Calendar Invite can be seen utilized again after being dormant since its first appearance in 2018.

As for the Attachments, HTML has again seen an increase now accounting for 7.01% of the total of 7.80% of the attachment types. There are also three new attachment types observed, including RAR, 7z, and LZ. All three attachments being some form of a archiving file type.

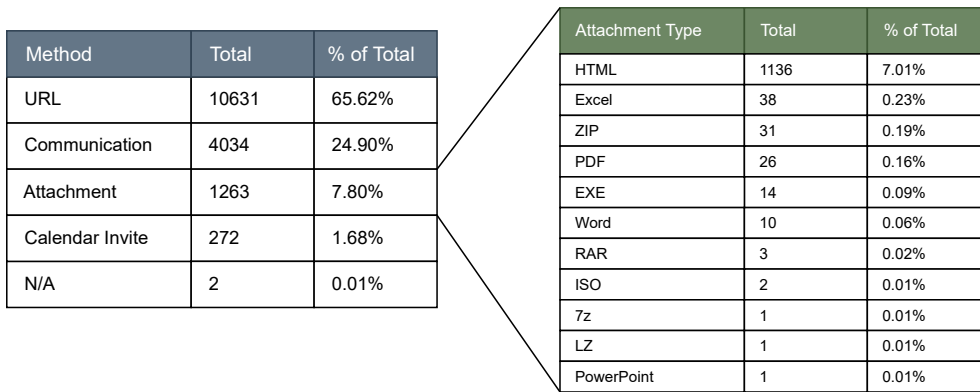


Figure A.36: 2021 - Method Distribution

A.6.4 Impersonation

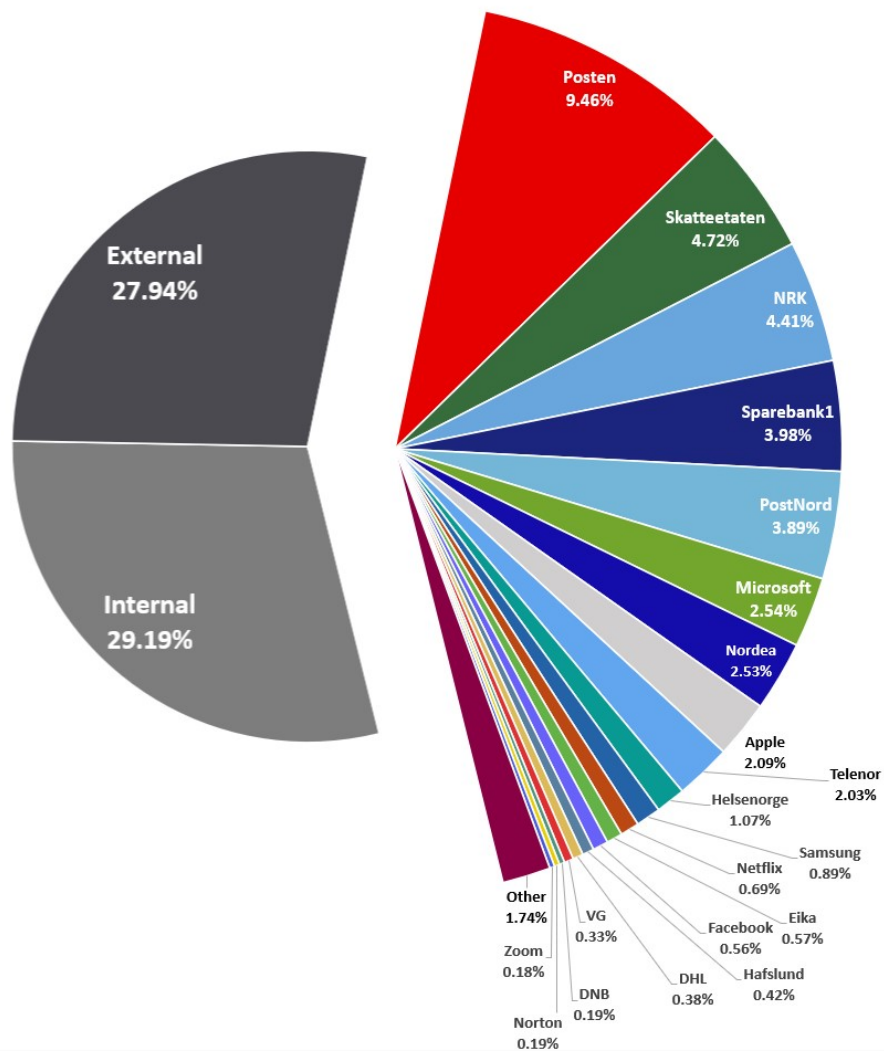


Figure A.37: 2021 - Impersonation Distribution

Both Apple and Microsoft, which were previously heavily utilized brands in phishing mails, have now seen a fair decline in overall usage. The impersonations of Posten, Skatteetaten, NRK, Sparebank1, and PostNord have all surpassed the two former mentioned brands, with Posten being the most utilized.

In total 53 non-generic brands were observed in 2021, accounting for 42.87% of the Impersonation property. Internal and External have a fairly even distribution with 29.19% and 27.94% respectively.

A.6.5 Dates

The distribution of dates in 2021 shows a continuation from 2020 with an increased amount of phishing mails in the months of October, November, and December. From the other months, there are no significant differences or variations, only small surges here and there. In contrary to last year, August shows more activity compared to the months of June, July, and September.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31
Jan	0	1	4	83	24	14	22	10	13	1	10	34	43	210	29	9	45	27	43	25	27	14	2	6	119	25	49	88	13	2	37
Feb	38	17	59	94	71	22	6	30	40	12	37	44	9	4	25	97	20	42	42	3	1	30	10	27	41	19	3	12	0	0	0
Mar	46	146	32	33	15	37	5	82	19	16	69	21	3	0	38	24	42	62	31	15	21	57	62	24	58	25	14	4	17	10	253
Apr	4	11	75	5	8	85	44	40	24	8	7	9	24	29	22	38	3	6	20	34	35	24	15	7	16	20	99	48	24	33	0
May	119	65	28	26	24	41	61	21	35	29	31	38	22	20	13	17	20	47	53	22	38	15	177	59	50	65	25	116	13	12	10
Jun	50	33	133	41	19	2	33	13	33	17	21	4	8	24	26	40	10	28	7	21	35	22	35	9	61	27	4	37	18	62	0
Jul	25	37	26	8	42	100	40	35	47	2	2	15	19	28	59	23	18	106	22	67	3	44	13	13	15	18	22	20	24	12	10
Aug	47	8	52	141	34	143	3	3	29	47	28	31	27	12	4	77	122	187	17	22	14	24	92	38	68	71	34	19	19	33	49
Sep	31	49	36	15	7	31	52	45	35	11	27	3	39	20	40	42	31	6	11	48	34	31	43	42	4	31	90	53	67	15	0
Oct	558	14	9	49	67	35	18	17	17	0	39	82	356	87	13	3	18	34	69	88	122	78	51	36	71	171	36	69	24	7	19
Nov	95	80	71	90	48	39	27	181	197	58	82	76	29	48	152	89	55	24	22	35	28	44	63	56	71	30	26	73	208	70	0
Dec	80	35	36	26	19	66	53	99	111	106	79	60	15	42	280	136	141	69	143	124	124	18	21	19	75	5	9	24	171	26	0

Figure A.38: 2021 - Date Distribution

A.6.6 Property Relationships

As observed in Figure A.39, there is an increased discrepancy between the various Content categories, creating a more intricate view of their Target relationships. Credentials had a total of 42 Content categories, whereas 23 of them were unique only to this Target. Money had 20 Content categories, with 11 unique. Credit Card Details with 17 connections, and two unique. The remainder had no one-to-one relationships.

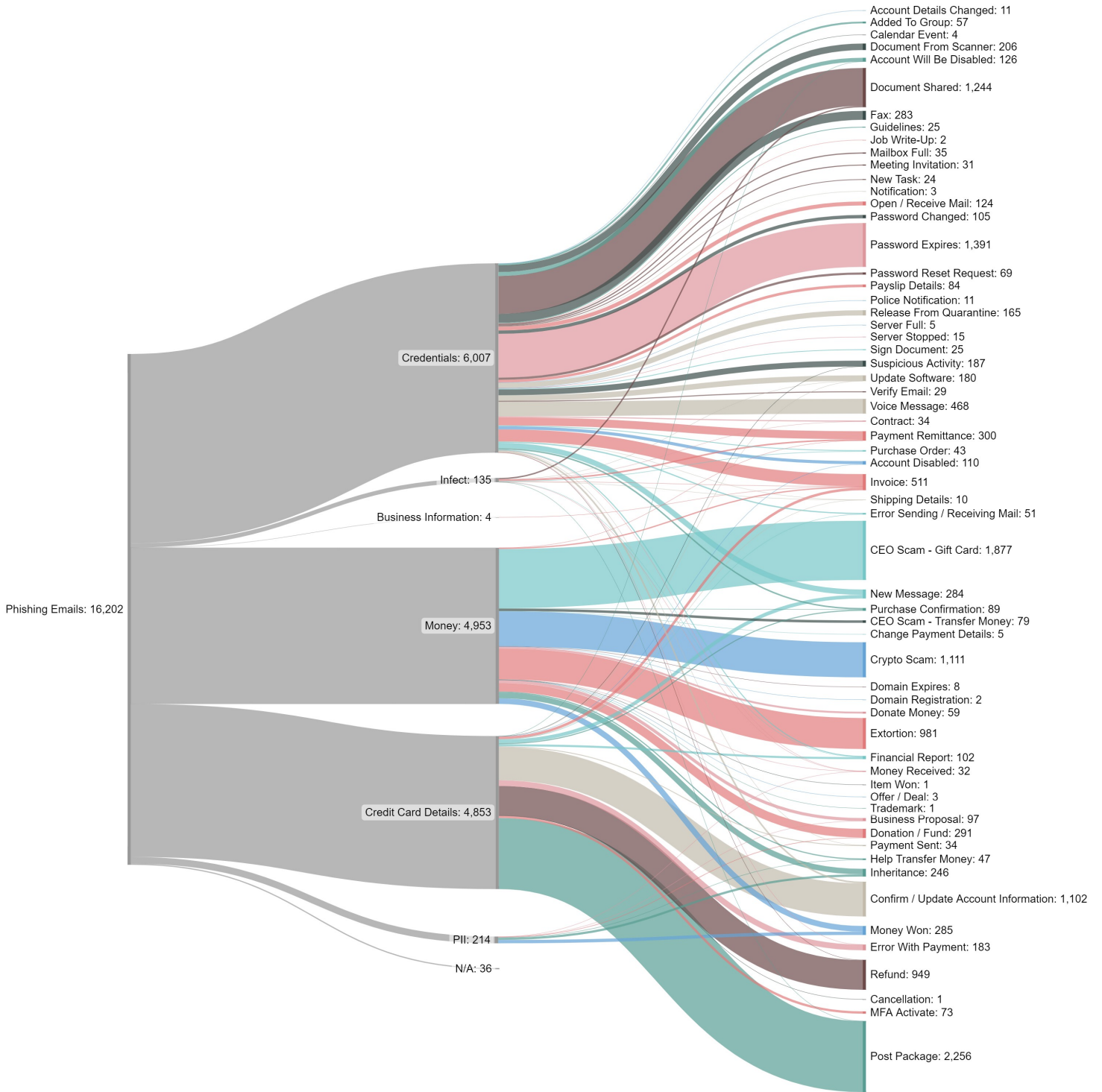


Figure A.39: 2021 - Target-Content Relationship

Invoice had increased its Target reach, now having a relationship with five of the six Target categories, including Credentials, Infect, Business Information, Money, and Credit Card Details.

The diversity index displayed in Table A.13, shows that Credentials is persistent as the most diverse Target category, while Business Information continues to only have one connection resulting in no diversity. Both Money and Infect has seen an increase in diversity, while PII has seen a decline.

Target	Total	Total Connections	$1 - \left(\frac{\sum n(n-1)}{N(N-1)} \right)$
Credentials	6007	42	0.89
Money	4953	20	0.79
Credit Card Details	4853	17	0.70
Infect	135	10	0.77
PII	214	6	0.65
Business Information	4	1	0

Table A.13: 2021 - Target-Content Diversity

From the Method-Target relationship displayed on Figure A.40, there are no changes that have occurred deviating from what has been observed before.

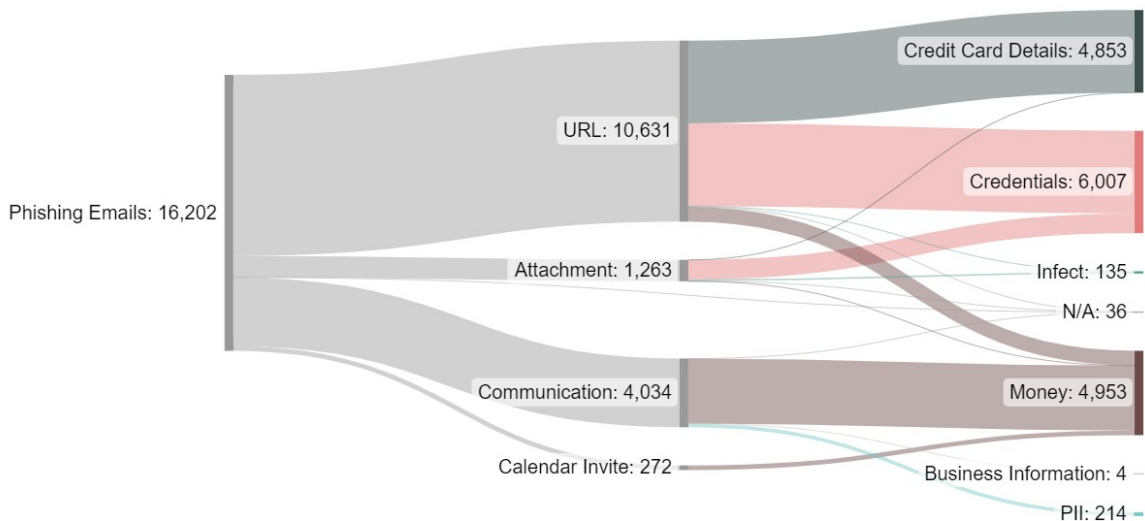


Figure A.40: 2021 - Method-Target Relationship

Money is back at being largely targeted through Communication, and Credit Card Details is no longer solely targeted via URLs, although still mostly attempted via URLs.

A.7 2022

2022 saw a total of 12781 reported phishing mails. This is a decrease from last year's 16202.

A.7.1 Content

56 content categories were observed in 2022, whereas the Anti-Virus Alert content category was not previously observed.

Post Package remains the top content category, increasing discrepancy wise from last year while remaining at a close to 14% total presence. CEO Scam – Gift Card as well remains second. Invoice is again within the top categories, and the three adjacent categories have all been around the top for a few iterations now.

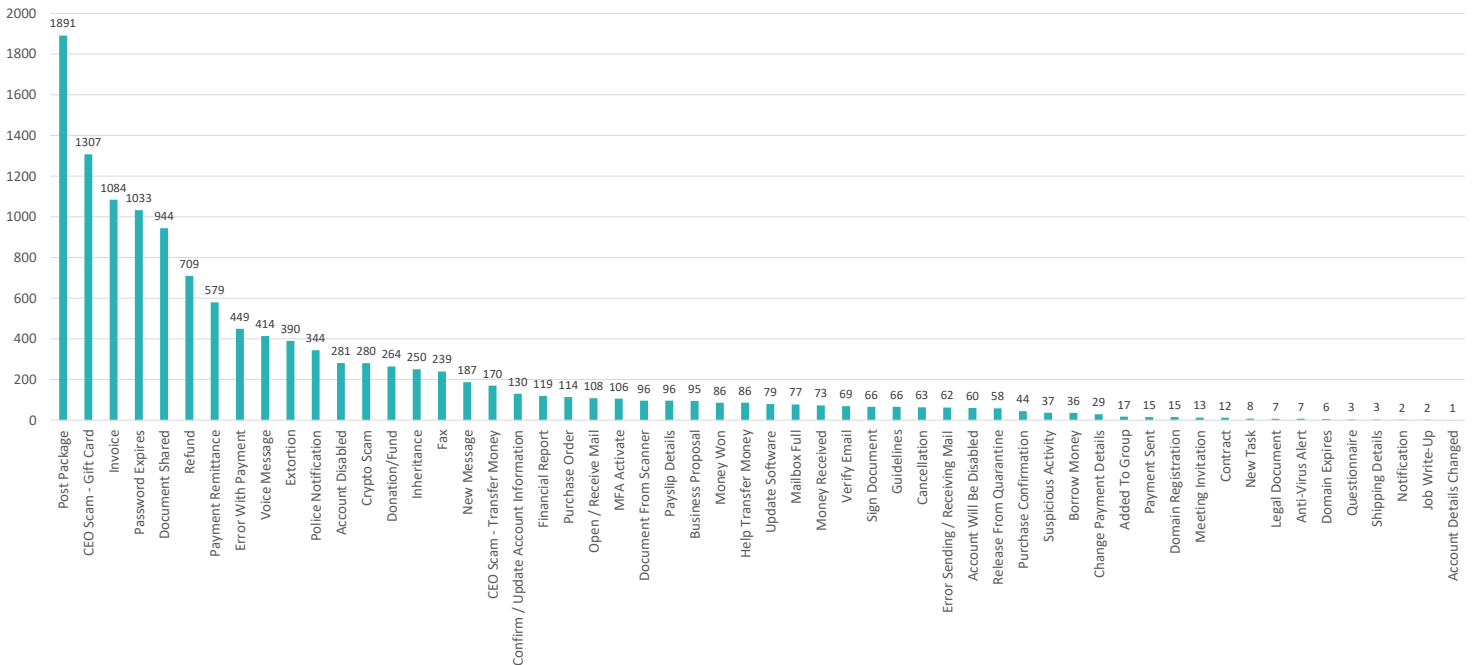


Figure A.41: 2022 - Content Distribution

A.7.2 Target

The targeting of Credentials sees a slight increase in 2022 from the great reduction in 2021, while the gap between Money and Credit Card Details remains remarkably close, differentiating only with 1.04%.

Infect is again above PII, however far away from the above Target category of Credit Card Details.

Target	Total	% of Total
Credentials	5687	44.50%
Money	3476	27.20%
Infect	2243	26.16%
Credit Card Details	170	1.33%
N/A	79	0.62%
PII	20	0.16%
Business Information	6	0.05%

Table A.14: 2022 - Target Distribution

A.7.3 Method

Figure A.42, displaying the distribution of Methods, shows that the URL Method continues to decrease, now only accounting for 56.26% which is the lowest it has ever been within the collection scope.

Method	Total	% of Total
URL	7191	56.26%
Communication	3058	23.93%
Attachment	2532	19.81%

Attachment Type	Total	% of Total
HTML	2169	16.97%
PDF	246	1.92%
ZIP	35	0.27%
Excel	31	0.24%
RAR	20	0.16%
Word	8	0.06%
ISO	8	0.06%
7z	4	0.03%
LZ	3	0.02%
GZ	2	0.02%
XLL	2	0.02%
XZ	2	0.02%
CAB	1	0.01%
EXE	1	0.01%

Figure A.42: 2022 - Method Distribution

The reduction in the utilization of URLs is accompanied by an increase in the usage of attachments. The Attachment category now accounts for 19.81% of the observed Methods, only 4.12% less than Communication.

Expanding on the attachment types, it is evident that the usage of HTML has boosted the Attachment category severely, with a representation of 16.07% of the total of all observed mails, attachment or not. There is also observed four new attachment types, including GZ, XLL, XZ, and CAB.

A.7.4 Impersonation

In 2022, a total of 48 brands were observed impersonated. Figure A.43 displays the distribution of Impersonations observed in 2022.

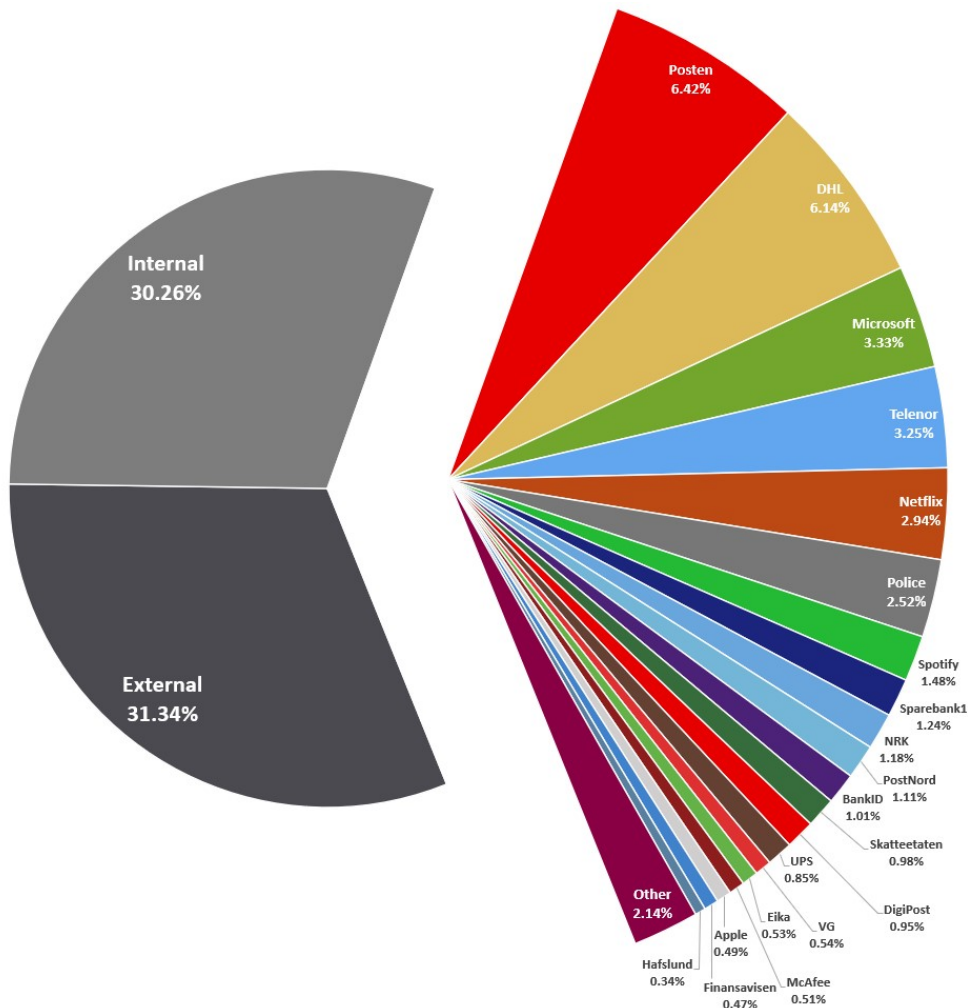


Figure A.43: 2022 - Impersonation Distribution

The distribution between the generic Impersonation categories of External and Internal remain even with a 31.34% and 30.26% distribution respectively. There is a small increase with the amount of generically impersonated mails contra phishing mails impersonating a brand/institution.

Posten is still the most impersonated brand, however more even with the subsequent brands. DHL in second has seen a great increase, now accounting for 6.14% of all the observed mails from last year's 0.38%.

A.7.5 Dates

The heat-map of 2022 deviates somewhat from the observations from the two prior years. In 2020 and 2021 there was a trend of increased activity in the months of October, November, and beginning of December, however, this trend is not present in 2022. November showcases a slight increased activity profile, but not anything that stands out as it did the two previous years. It can also be seen that in June there was increased activity, which for the two former years was one of the more quiet months.

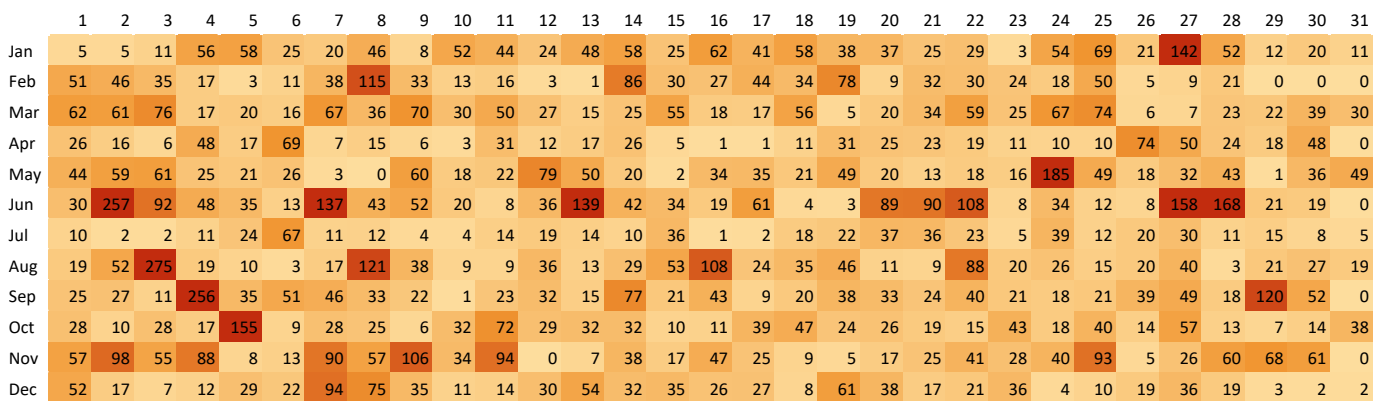


Figure A.44: 2022 - Date Distribution

There is an increase of big surges in 2022, where 10 of the days were within the 90th percentile.

A.7.6 Property Relationships

As observed throughout all the collection years, Figure A.45 displays yet again that the Credential Target contains the most one-to-one relationships, totaling 18. Infect, as well, remains the only Content category with five unique Target connections.

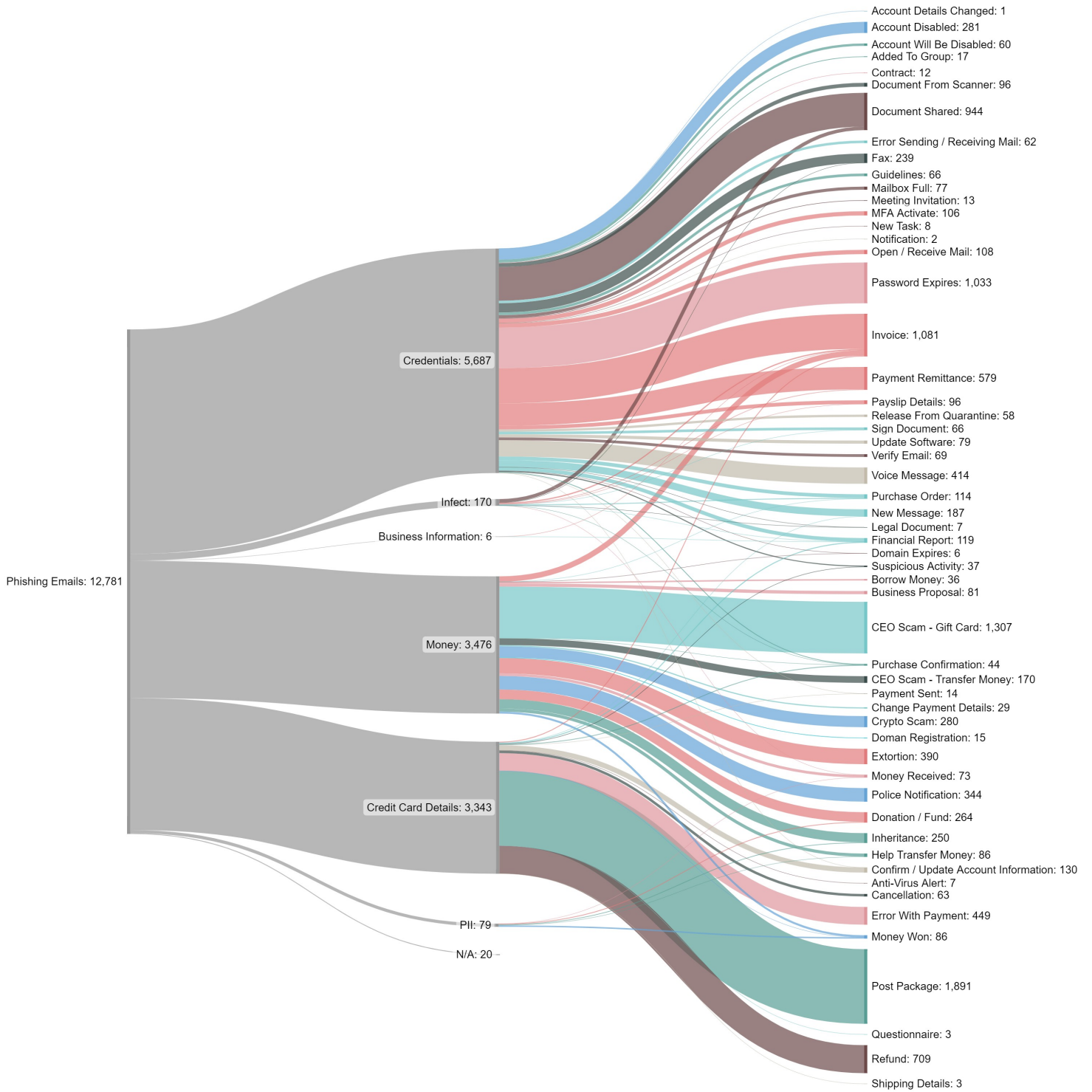


Figure A.45: 2022 - Target-Content Relationship

The diversity index remains fairly consistent with the observations from last year. Credentials is the most diverse Target category, as it has been since 2017, while the most significant change from last year is that Business Information now has diversity.

Target	Total	Total Connections	$1 - \left(\frac{\sum n(n-1)}{N(N-1)} \right)$
Credentials	5687	34	0.90
Money	3476	18	0.81
Credit Card Details	3343	15	0.62
Infect	170	11	0.68
PII	79	5	0.71
Business Information	6	2	0.33

Table A.15: 2022 - Target-Content Diversity

2022 marks the year where a great change in the targeting of Credentials have occurred. As seen in Figure A.46, Attachments are now a prominent Method utilized for the targeting of Credentials, not far behind URLs which have been the main Method for the last six years.

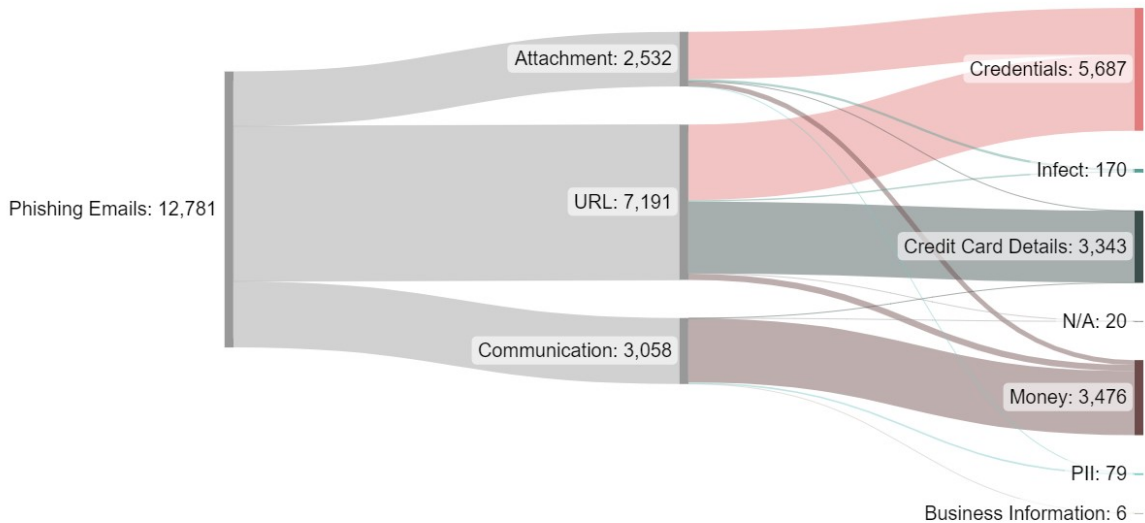


Figure A.46: 2022 - Method-Target Relationship

The other relationships remain consistent with what has been observed before, where Money and Business Information are mainly targeted through Communication, and Credit Card Details through URLs. Infect remains split between Attachment and URL, and PII neither shows any particular preference in Method.

Appendix B

Content Categories

The following document contains descriptions of all the categories belonging to the Content property of the Email Phishing Collection Model. The categories are sorted in alphabetical order, with each of the descriptions displaying an example email belonging to said Content category.

Account Details Changed

This Content category encompasses all emails telling the recipient that their account details have been changed. This could for instance be that their password has been updated, phone number has changed, or that their subscription plan has been updated.

Figure B.1 displays an example of an email within this Content category.

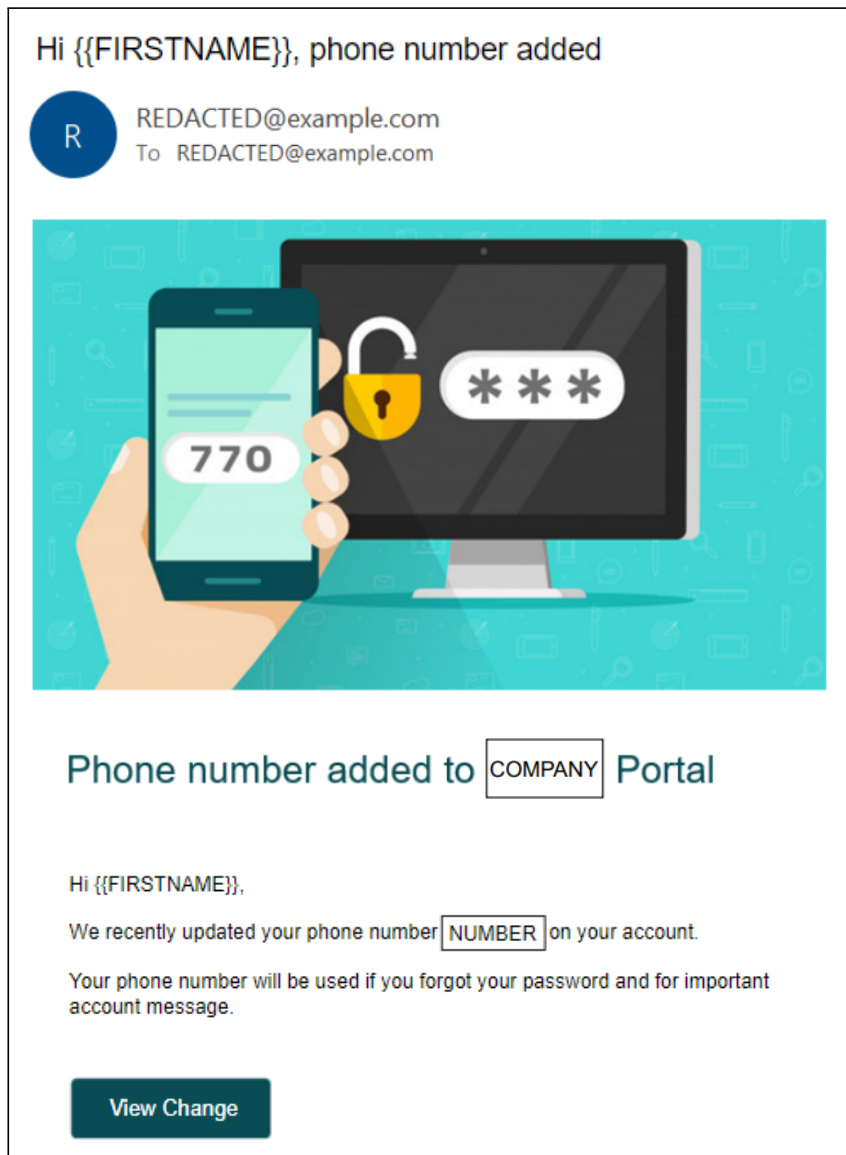


Figure B.1: Account Details Changed

Account Disabled

The Account Disabled category consists of the phishing emails conveying that the account of the recipient has been disabled, and that an action has to be conducted in order to reverse / undo this disabling.

Figure B.2 displays an example of an email within this Content category. The email informs the recipient that their digital bank access has been deactivated and that a verification has to be performed in order to open the account.

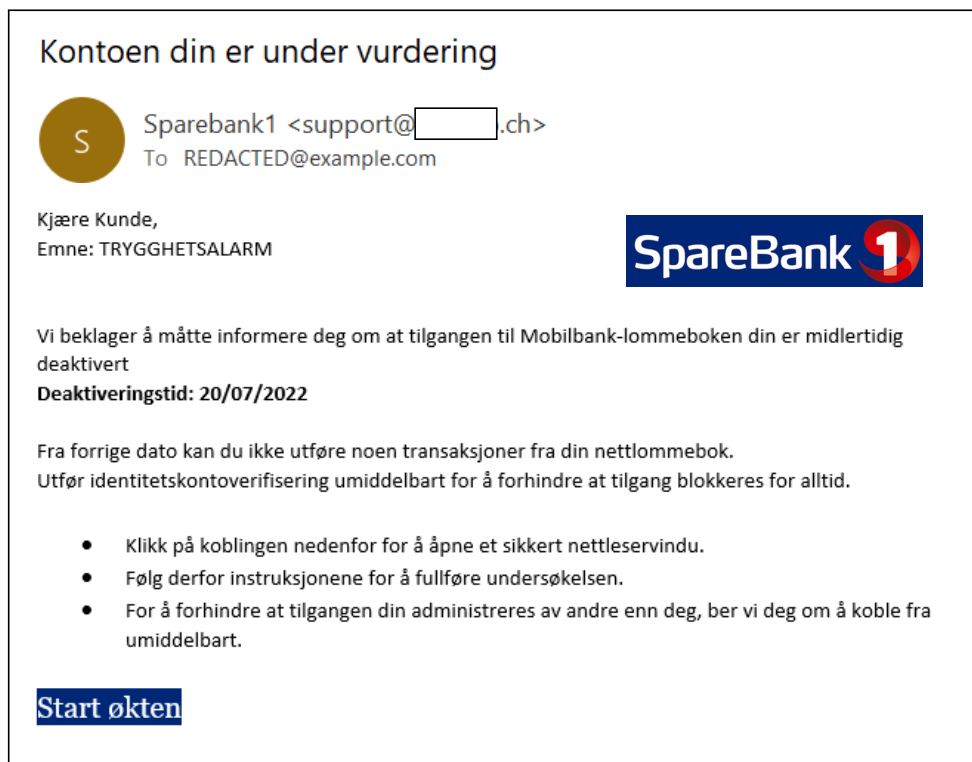


Figure B.2: Account Disabled

Account Will Be Disabled

Similarly with the Account Disabled Content category, in this category, the emails convey that the recipient's account is about to expire and that an action has to be completed to prevent the disabling.

Figure B.3 displays an example of an email within this Content category.

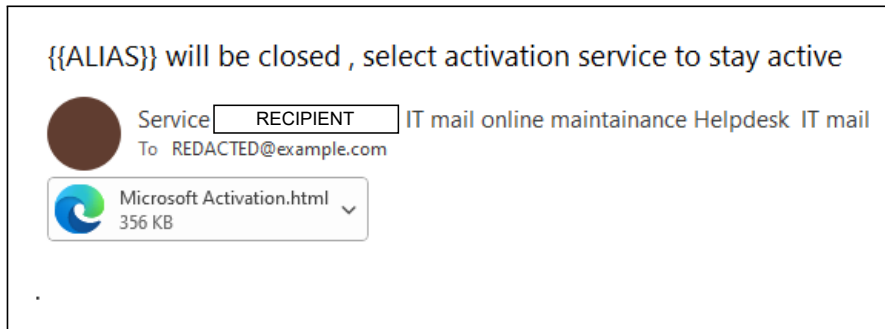


Figure B.3: Account Will Be Disabled

Added To Group

This Content category embodies all phishing emails notifying the recipient that they have been added to an online group, such as a Microsoft Teams group or Slack group.

Figure B.4 displays an example of an email within this Content category.

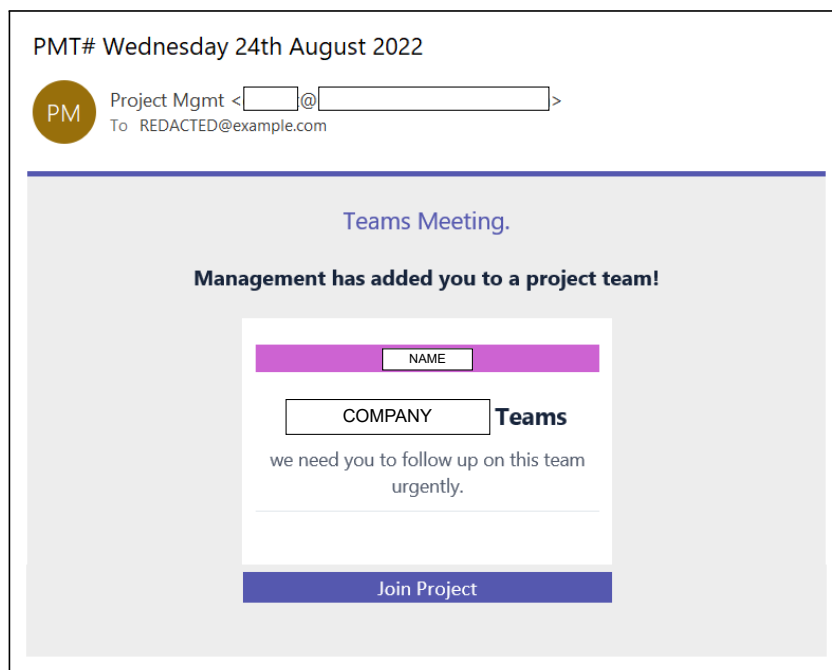


Figure B.4: Added To Group

Advert Stopped

In this Content category, the emails where the recipient is notified that their published advert has been taken down / run out are gathered.

Figure B.5 displays an example of an email within this Content category. The email notifies the user that their Finn.no advert has been stopped, and that a link has to be opened in order to activate the advert again.

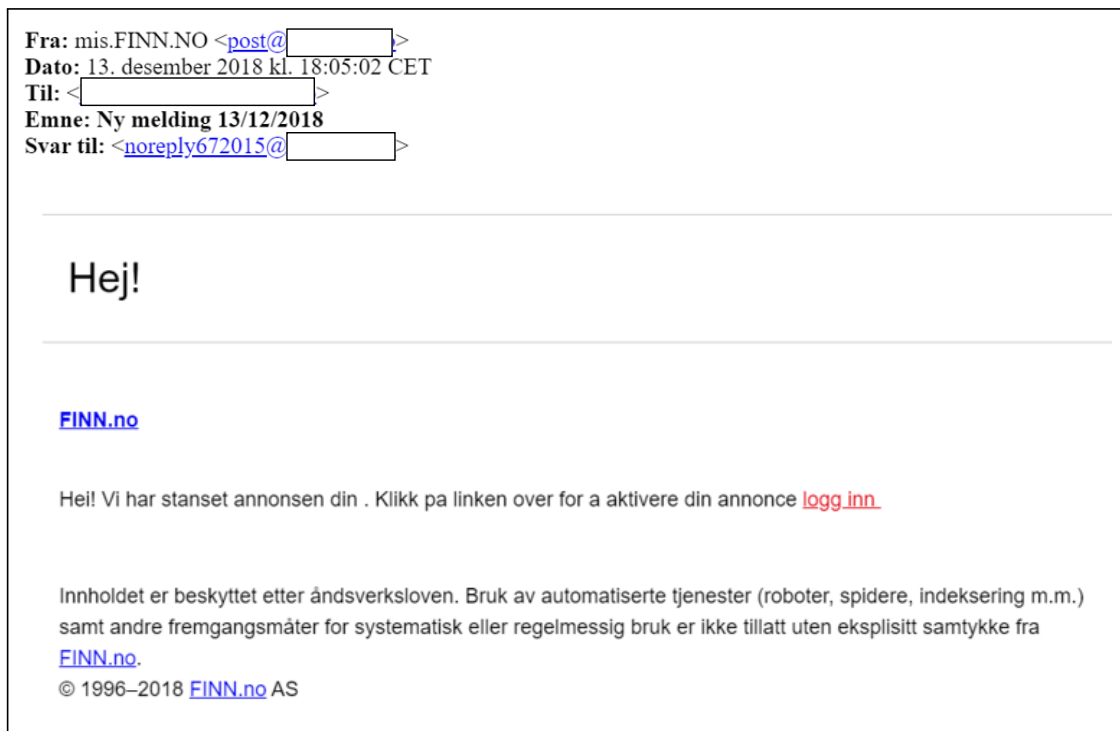


Figure B.5: Advert Stopped

Anti-Virus Alert

The Anti-Virus Alert category consists of phishing emails notifying the recipient that there has been detected malware or other suspicious behavior on their machine by an Anti-Virus solution.

Figure B.6 displays an example of an email within this Content category. The email informs the recipient that there has been identified a potential virus on their computer, and that they should download the linked anti-virus solution and scan their machine.

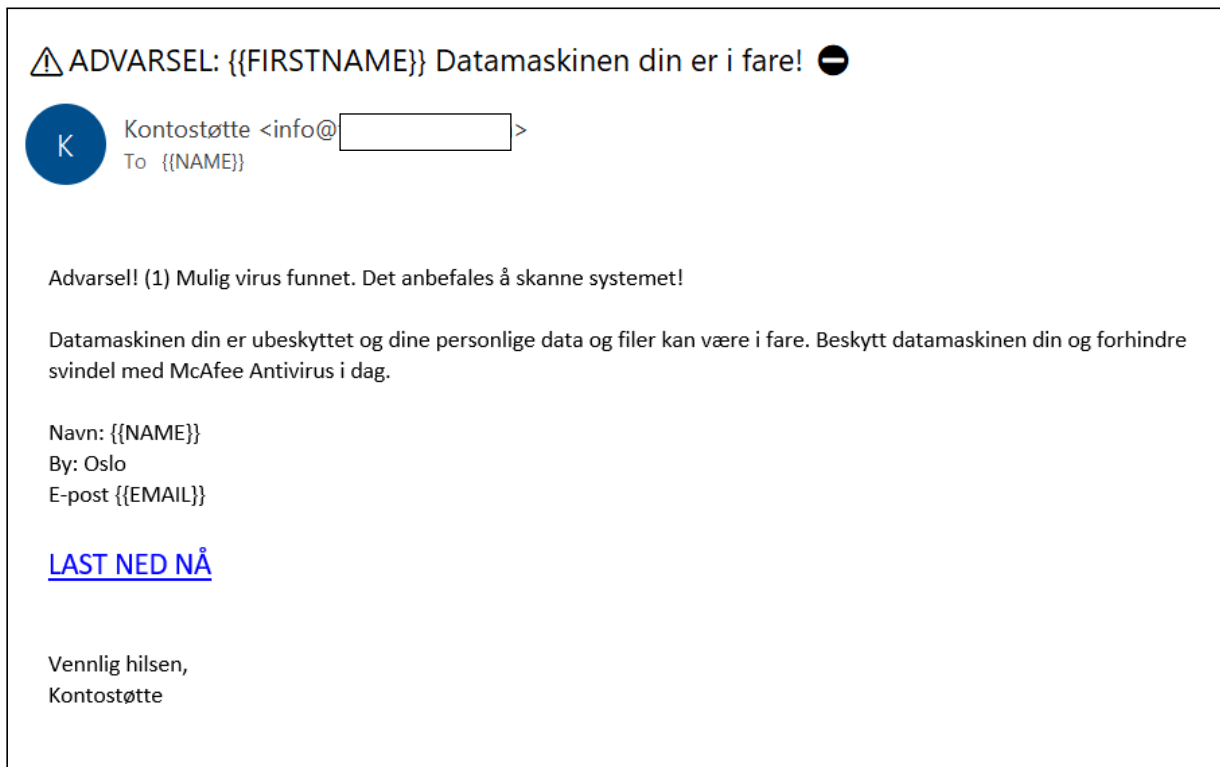


Figure B.6: Anti-Virus Alert

Borrow Money

This category encompasses the emails where the malicious actor, through impersonation, solicits money from the recipient in order to help the subject of the email out. This could for instance be a refugee needing money for transportation, or a family needing money for food.

Figure B.7 displays an example of an email within this Content category.



Figure B.7: Borrow Money

Business Proposal

The Business Proposal category consist of the phishing emails where the sender has a lucrative business offer for the recipient. This could be an investment offer or a team-up proposal to conduct business together.

Figure B.8 displays an example of an email within this Content category.



Figure B.8: Business Proposal

Calendar Event

This Content category revolves around the phishing emails conveying a calendar event to the recipient. This could for instance be an email invite to an event or a notification that an event in the calendar is approaching.

Figure B.9 displays an example of an email within this Content category. The email states that the recipient has received an invitation for a new event, with a link to said event in the calendar.

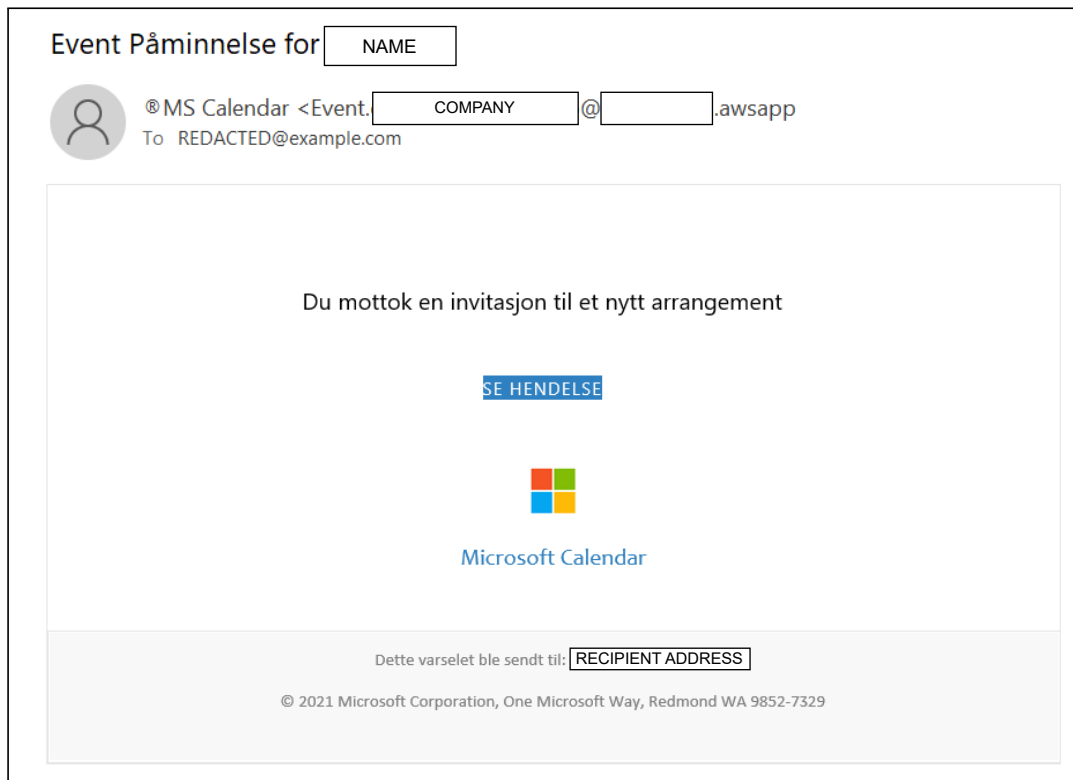


Figure B.9: Calendar Event

Cancellation

In this content category, the phishing emails conveying the message that a service consumed by the recipient has been cancelled are gathered.

Figure B.10 displays an example of an email within this Content category.

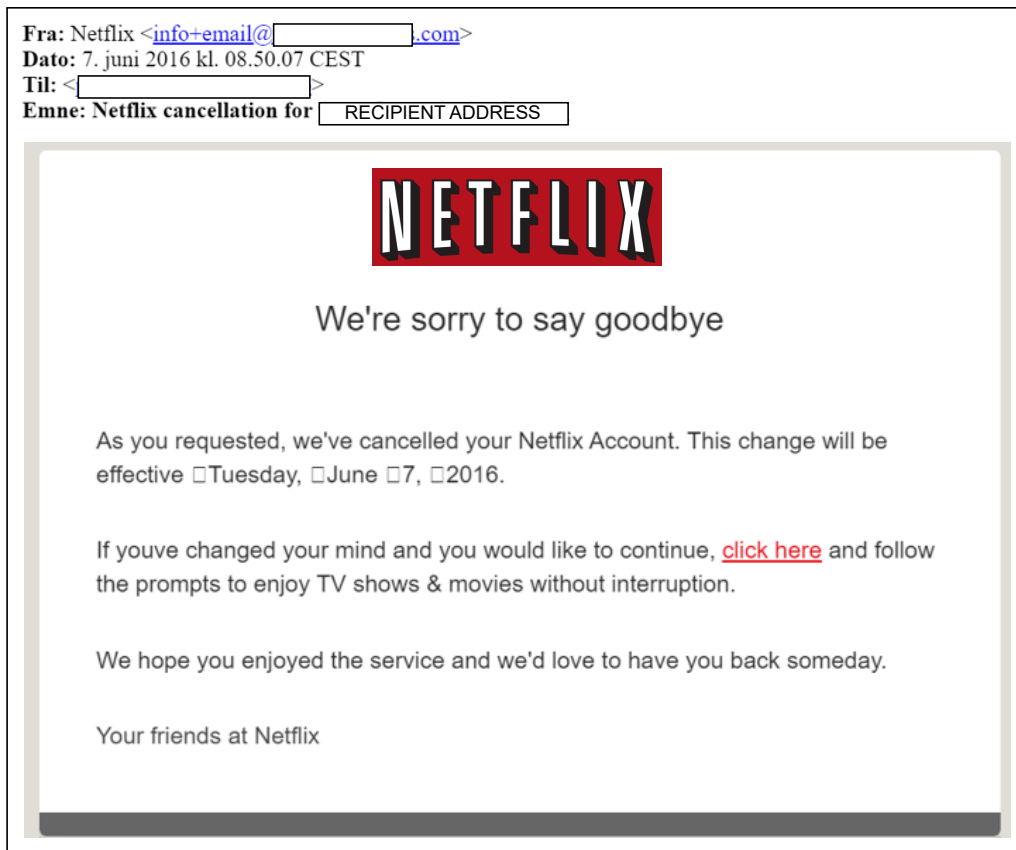


Figure B.10: Cancellation

CEO Scam - Crypto

This category consists of the phishing emails where the CEO Scam technique is utilized in the context of purchasing crypto currency.

Figure B.11 displays an example of an email within this Content category. The email contains a request for the recipient to download an application on their phone and purchase NOK 4000 worth of Bitcoins.

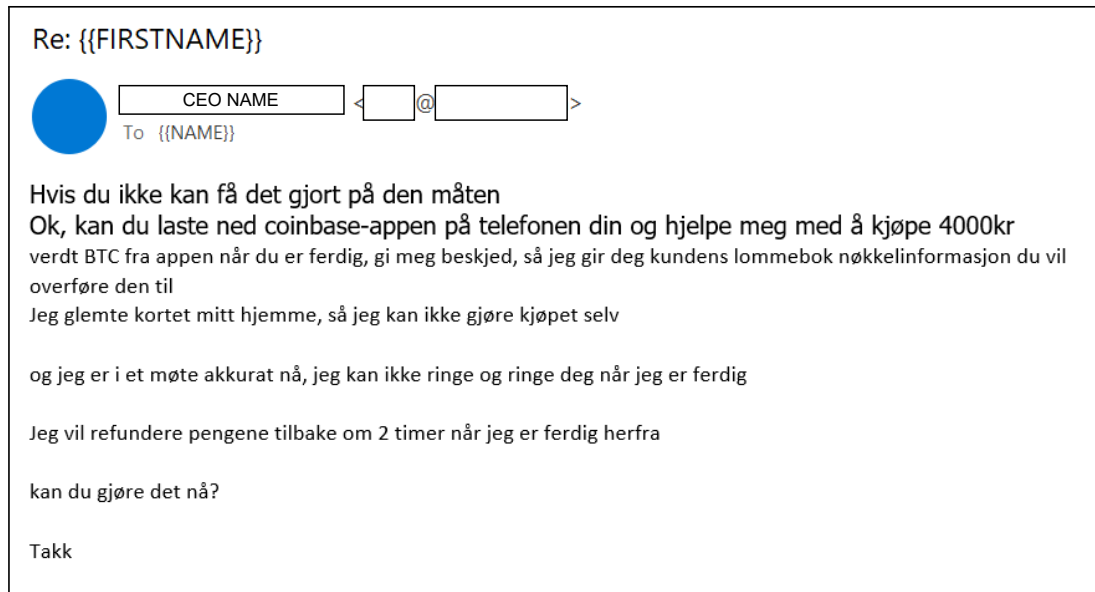


Figure B.11: CEO Scam - Crypto

CEO Scam - Gift Card

In this category, the malicious actor utilizes the CEO Scam technique in order to lure the recipient into purchasing gift cards for them. This could for instance be in relation to a customer, or as a gift for the employees.

Figure B.12 displays an example of an email within this Content category. The email contains a request for the recipient to purchase gift cards for the CEO.



Figure B.12: CEO Scam - Gift Card

CEO Scam - Transfer Money

This Content category encompasses all the phishing emails that utilize the CEO Scam technique in the context of requesting the recipient to transfer a sum of money, for instance due to a business deal or an unpaid invoice.

Figure B.13 displays an example of an email within this Content category. The email contains a request for the recipient to perform an international transfer of funds.



Figure B.13: CEO Scam - Transfer Money

Change Payment Details

This category consists of the emails where the malicious actor requests that the payment details for a specific transaction is changed. Changing the account in which an invoice is paid to, or changing the employee's payslip account are examples of this.

Figure B.14 displays an example of an email within this Content category.

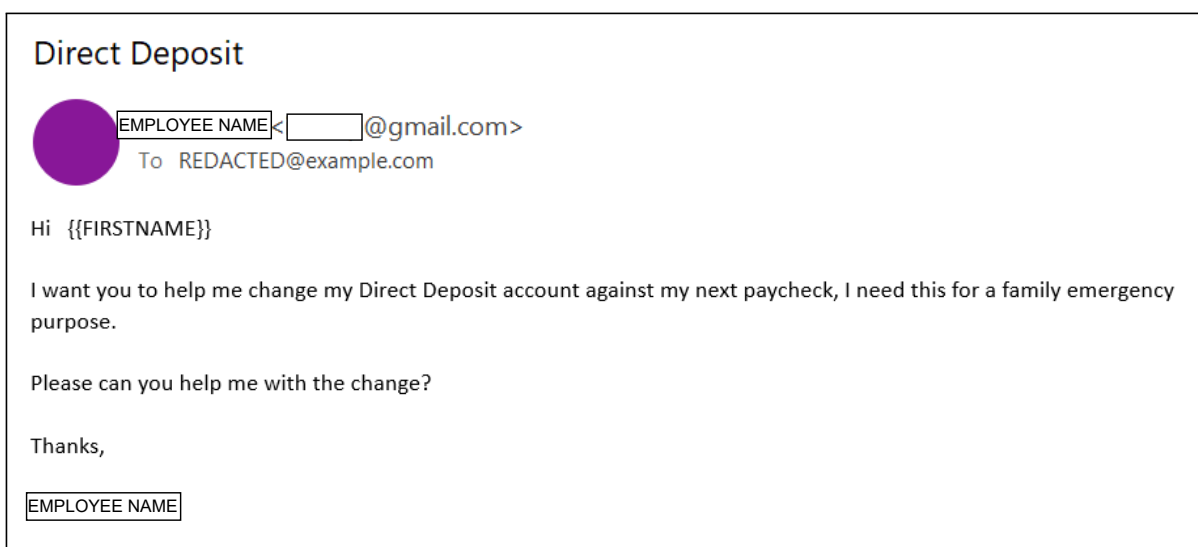


Figure B.14: Change Payment Details

Confirm / Update Account Information

This Content category encompasses emails soliciting the recipient into confirming that their account details are correct or to update them if they are not correct.

Figure B.15 displays an example of an email within this Content category. The email informs the recipient that they have to confirm their personal information in order to continue using the sender bank's services.

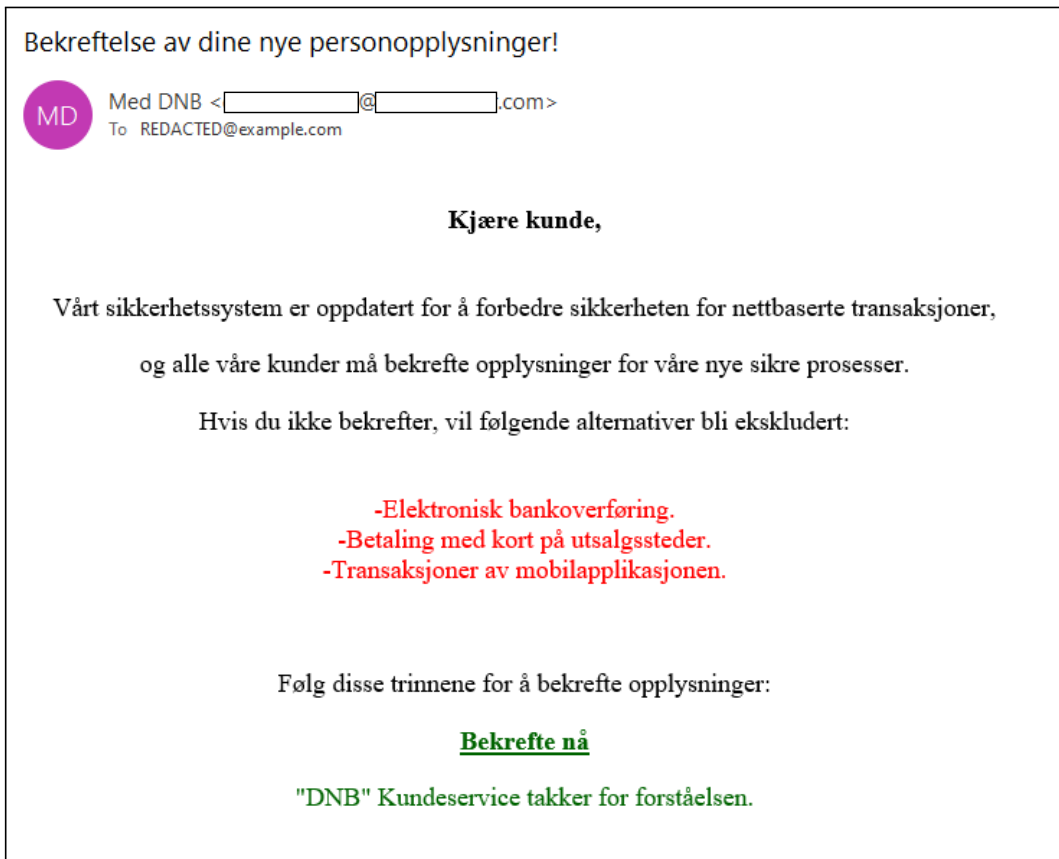


Figure B.15: Confirm / Update Account Information

Contract

All emails detailing a contract belong in this Content category.

Figure B.16 displays an example of an email within this Content category.

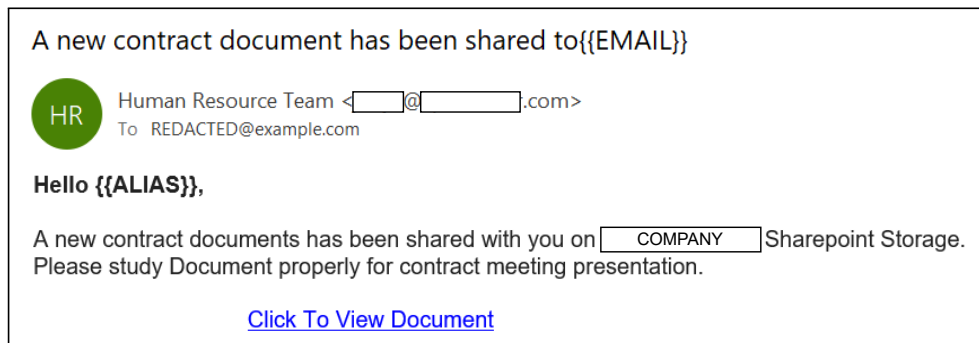


Figure B.16: Contract

Crypto Scam

The Crypto Scam Content category encompass the emails attempting to lure the recipient into purchasing crypto through illegitimate sources.

Figure B.17 displays an example of an email within this Content category. The email contains a news article describing famous chess player Magnus Carlsen's secret investment strategy.



Figure B.17: Crypto Scam

Document From Scanner

This Content category includes all the phishing emails pretending to be a scanned document.

Figure B.18 displays an example of an email within this Content category.

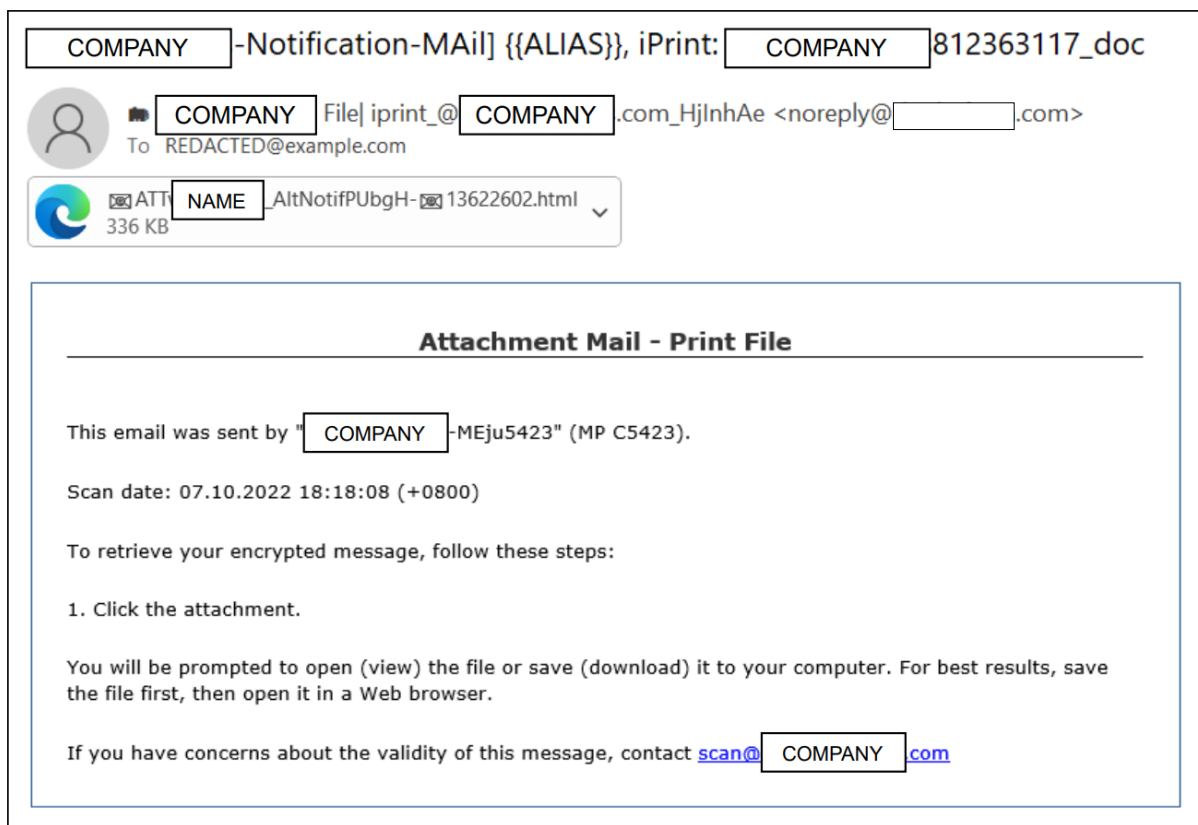


Figure B.18: Document From Scanner

Document Shared

The Document Shared Content category is a collective term for all the phishing emails that conveys the message that a document has been shared with the recipient.

Figure B.19 displays an example of an email within this Content category.

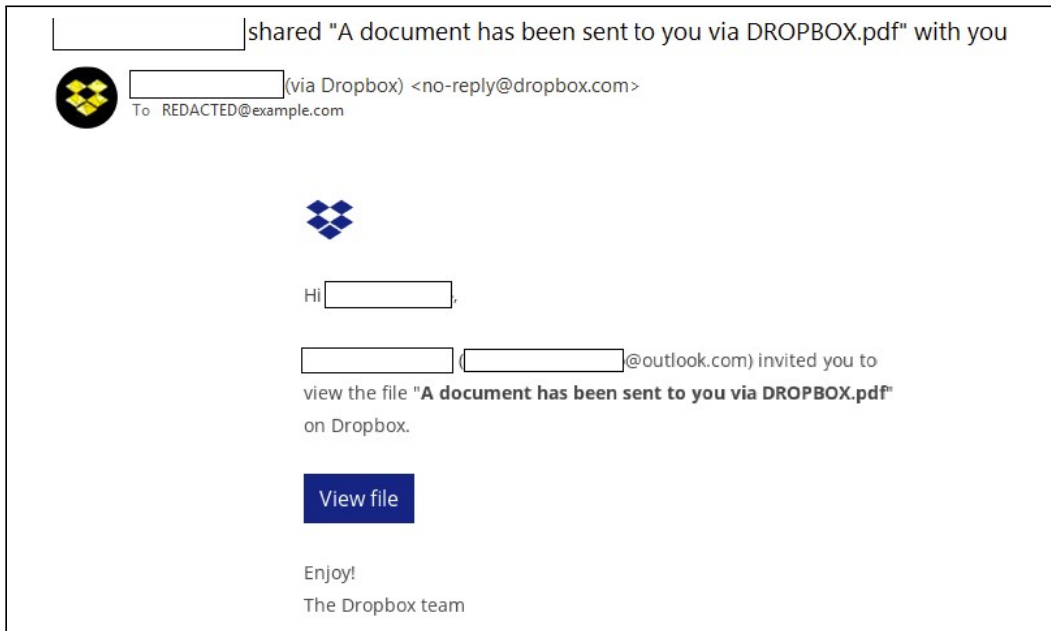


Figure B.19: Document Shared

Domain Expires

This category consists of emails conveying that the domain of the recipient is about to expire, and that an action has to be performed in order to prevent it from being terminated.

Figure B.20 displays an example of an email within this Content category.

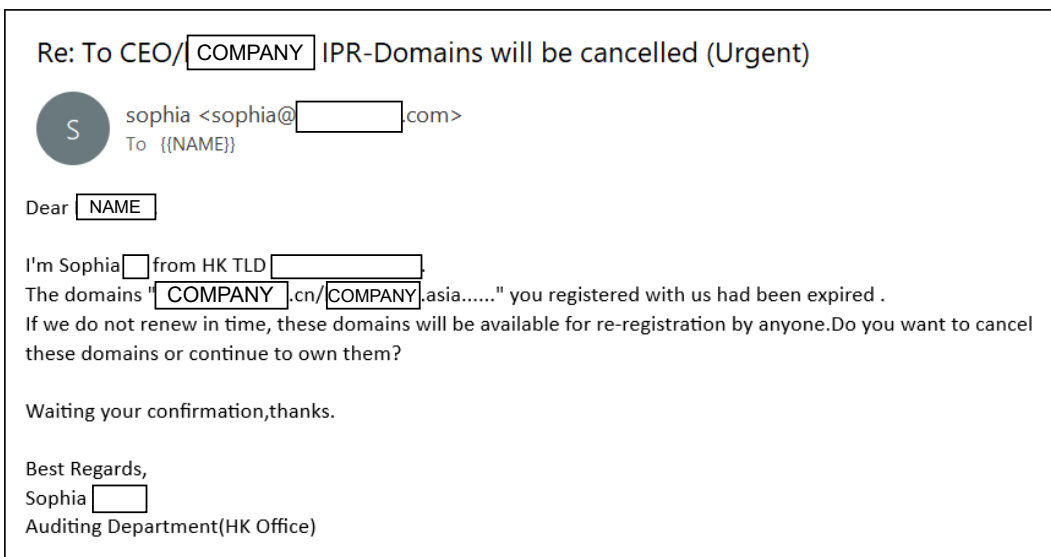


Figure B.20: Domain Expires

Domain Registration

The Domain Registration Content category consists of emails attempting to sell a domain registration to the recipient. This could be masked as a domain registrant informing the recipient that a domain similar to theirs is about to be purchased, and that the recipient would be able to purchase it before another party takes the domain name. Another case could be that the malicious actor is simply asking the recipient if they want to purchase said domain.

Figure B.21 displays an example of an email within this Content category.

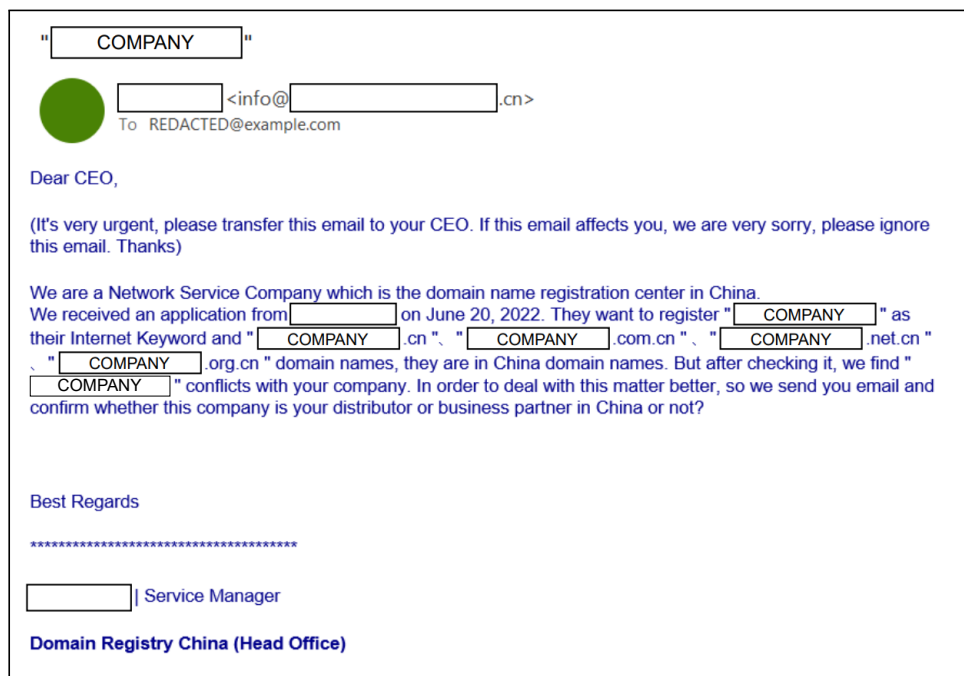


Figure B.21: Domain Registration

Donate Money

These types of phishing emails solicits donations from the recipient.

Figure B.22 displays an example of an email within this Content category.

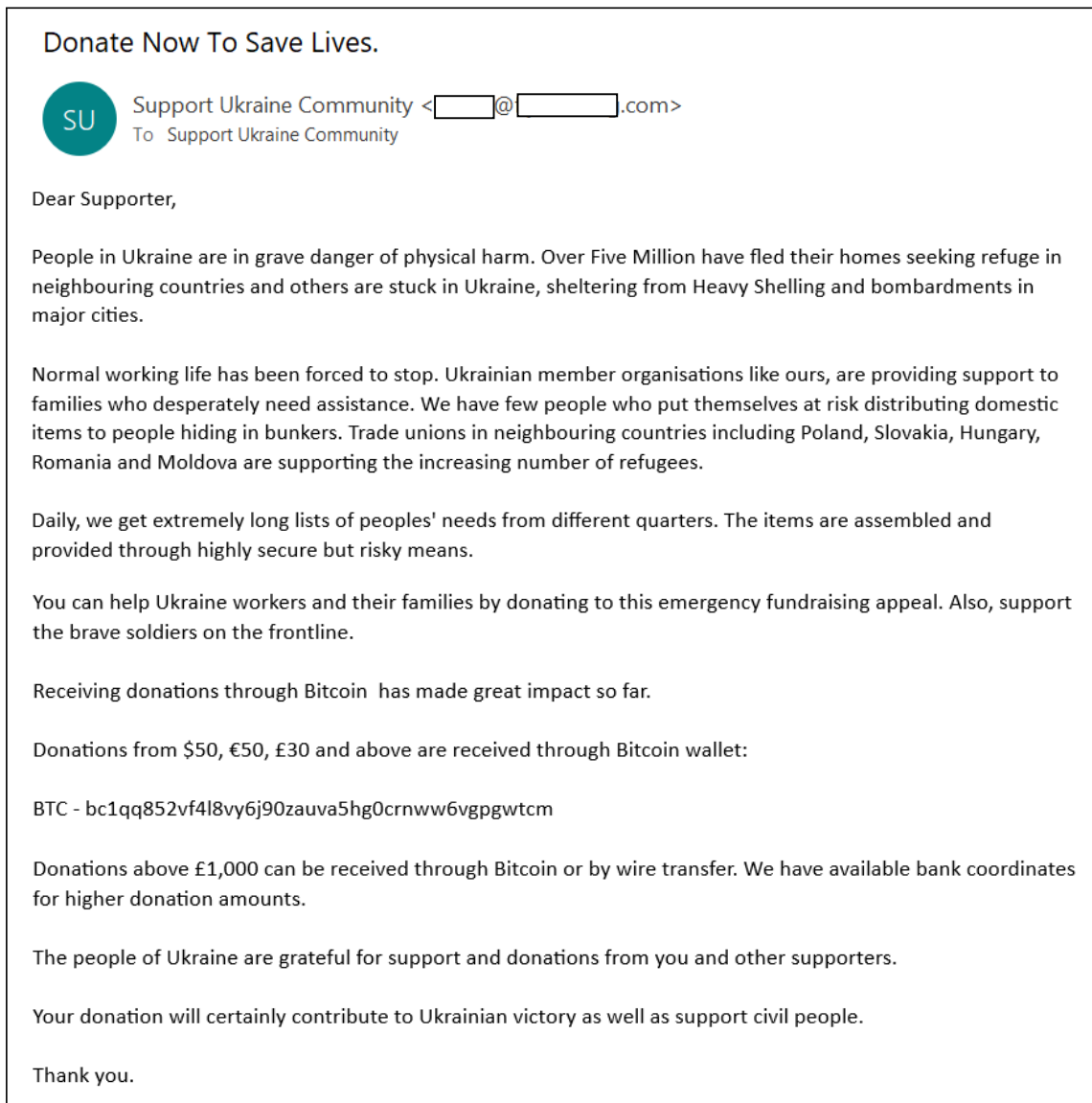


Figure B.22: Donate Money

Donation / Fund

These types of phishing emails presents the recipient with a donation or a fund in their name, querying the recipient into performing an action in order to receive said donation or fund.

Figure B.23 displays an example of an email within this Content category.

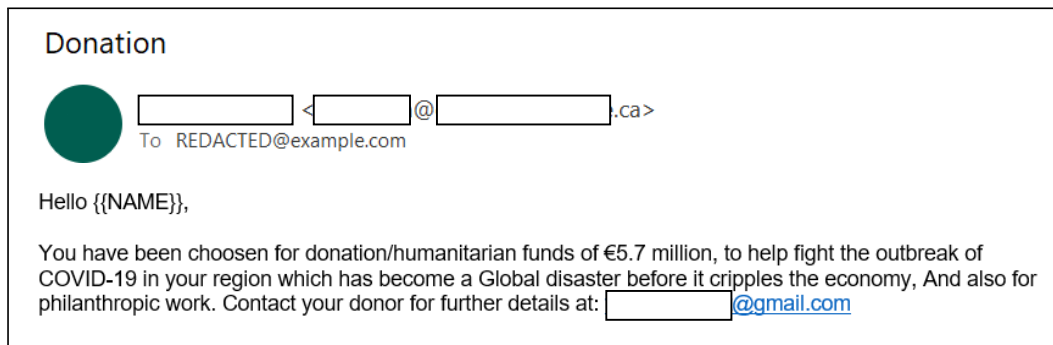


Figure B.23: Donation / Fund

Error Sending / Receiving Mail

This Content category consists of emails that notify the recipient that there has been an error when attempting to receive or deliver an email to/from the recipient, and that an action has to be performed in order to resolve this error.

Figure B.24 displays an example of an email within this Content category.

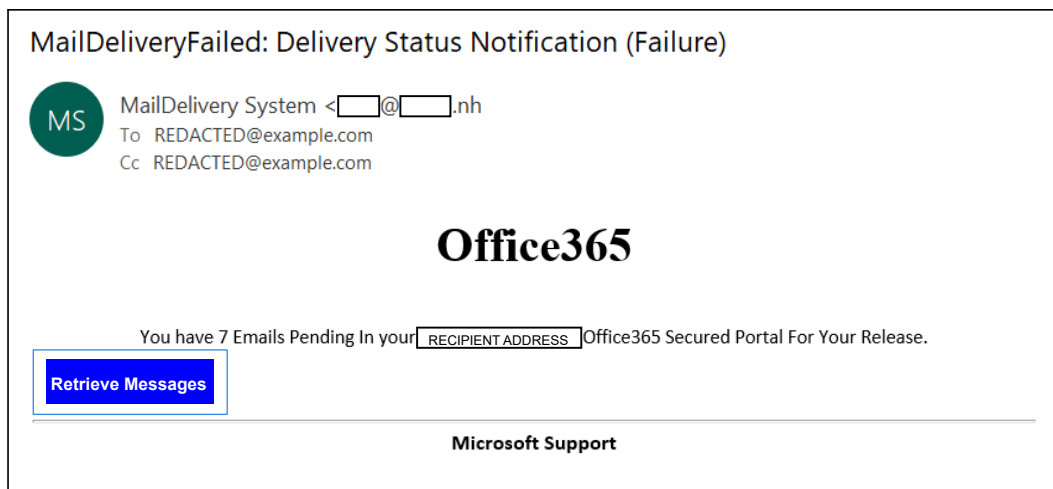


Figure B.24: Error Sending / Receiving Mail

Error With Payment

This Content category consists of emails notifying the recipient that there has been an error with an attempted payment from the recipient, and that an action has to be performed in order to resolve this error.

Figure B.25 displays an example of an email within this Content category.

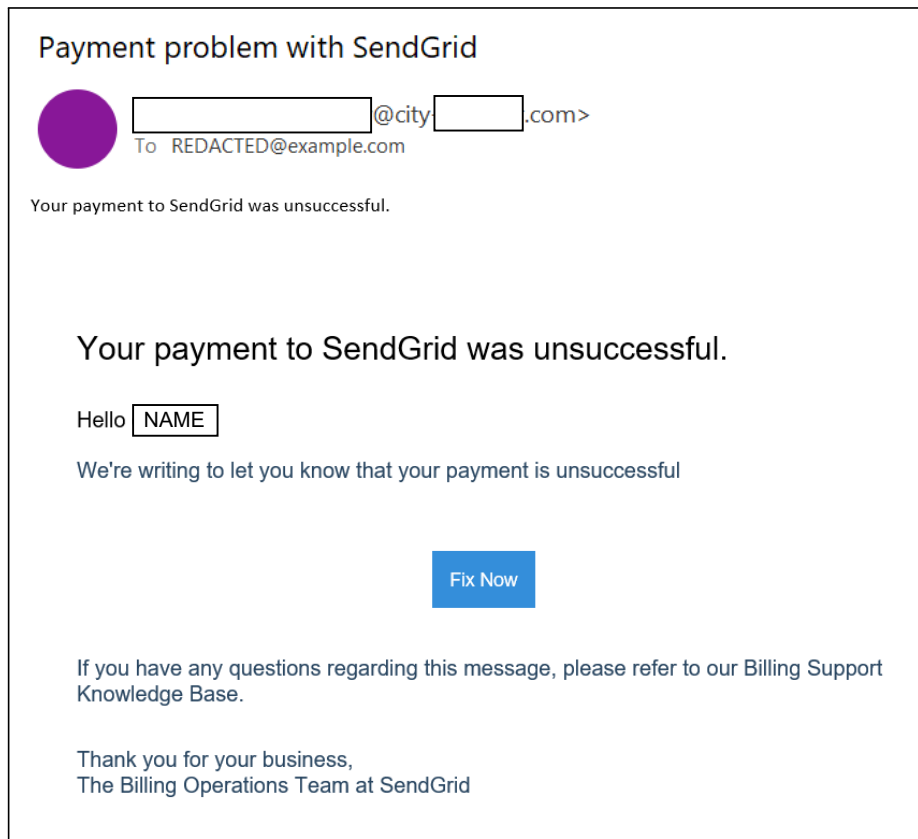


Figure B.25: Error With Payment

Extortion

The Extortion category encompasses emails that try to demand a certain action to be performed in order to prevent an undesirable situation. Instances such as when a malicious actor supposedly has infected the machine of the recipient, demanding money to remove said infection is an example of this category type.

Figure B.26 displays an example of an email within this Content category.

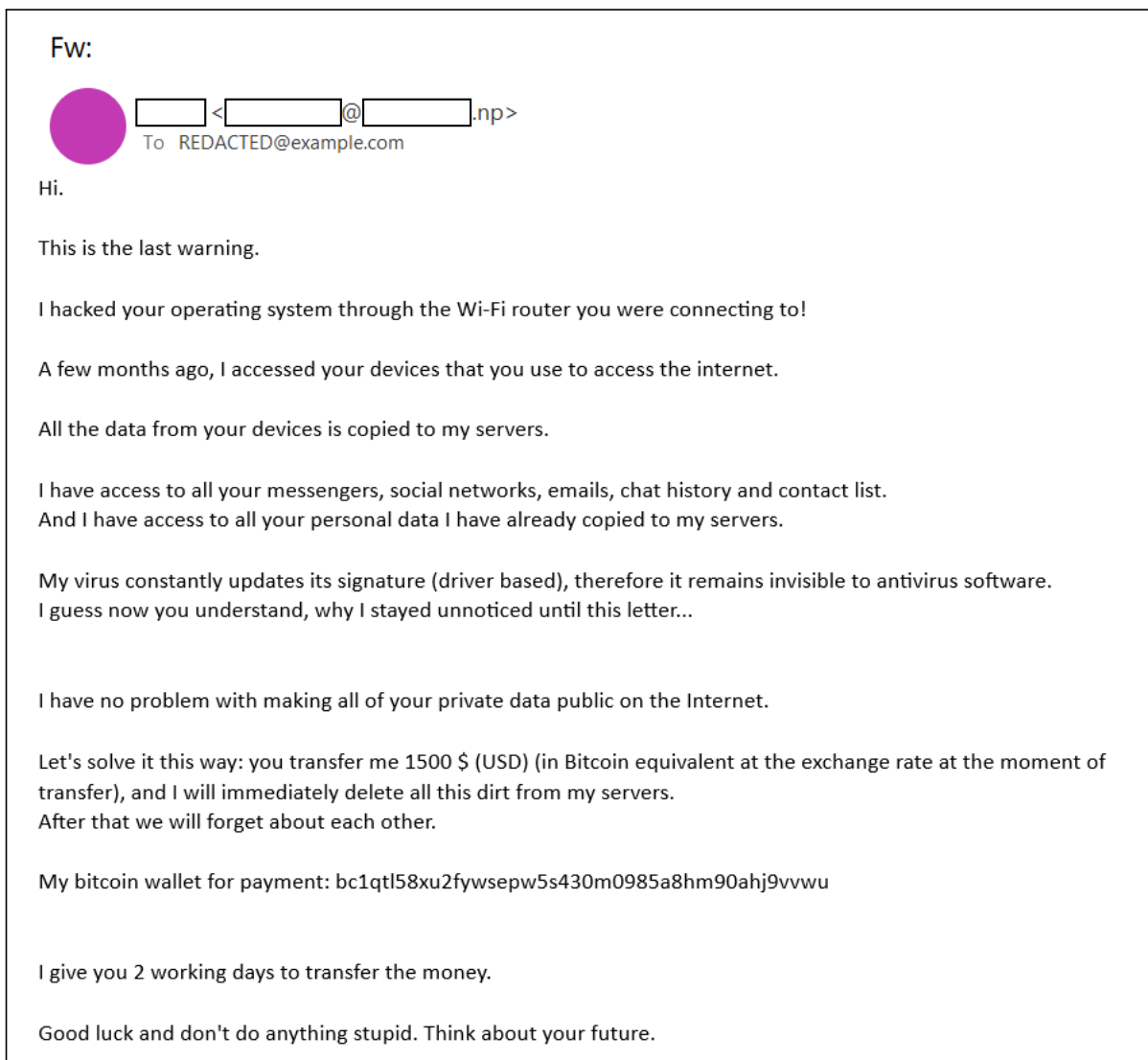


Figure B.26: Extortion

Fax

This category consists of emails disguised as fax/eFax messages.

Figure B.27 displays an example of an email within this Content category.

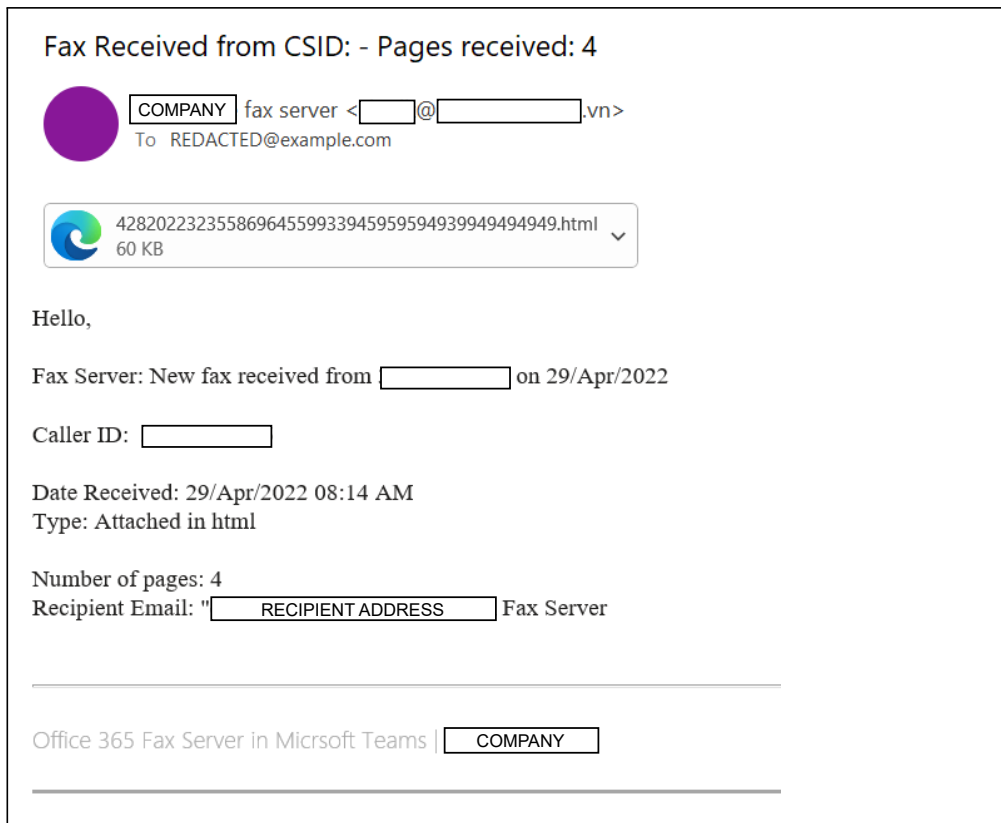


Figure B.27: Fax

Financial Report

The Financial Report category encompass the phishing emails masked as financial reports, such as bank statements, balance sheets, or income reports.

Figure B.28 displays an example of an email within this Content category. The email notifies the recipient that their tax documents are available through the linked site.

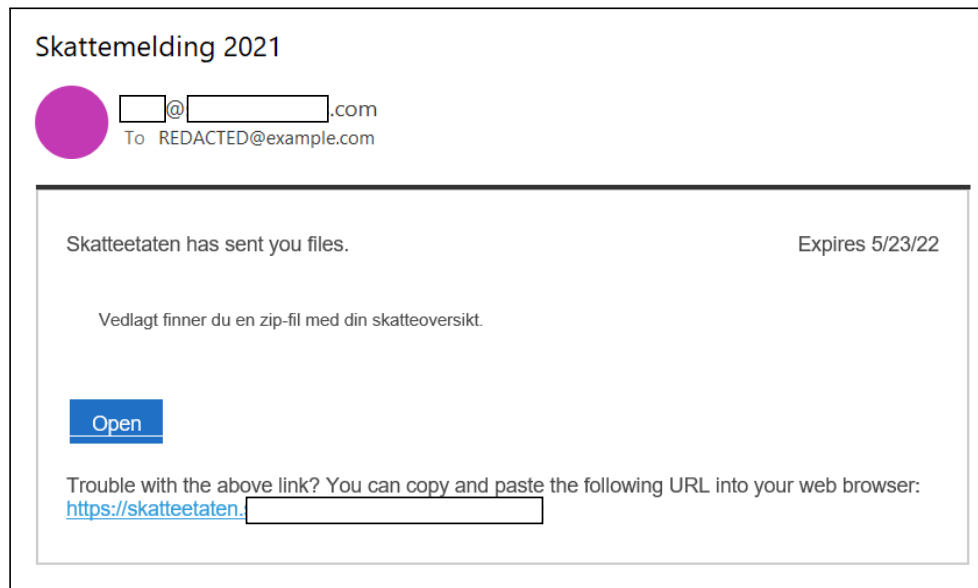


Figure B.28: Financial Report

Guidelines

This Content category consist of all phishing emails presenting guidelines to the recipient. These guidelines could for instance be organizational guidelines or guidelines related to a specific service.

Figure B.29 displays an example of an email within this Content category. The email informs the recipient that new guidelines for work-from-home has been made available.

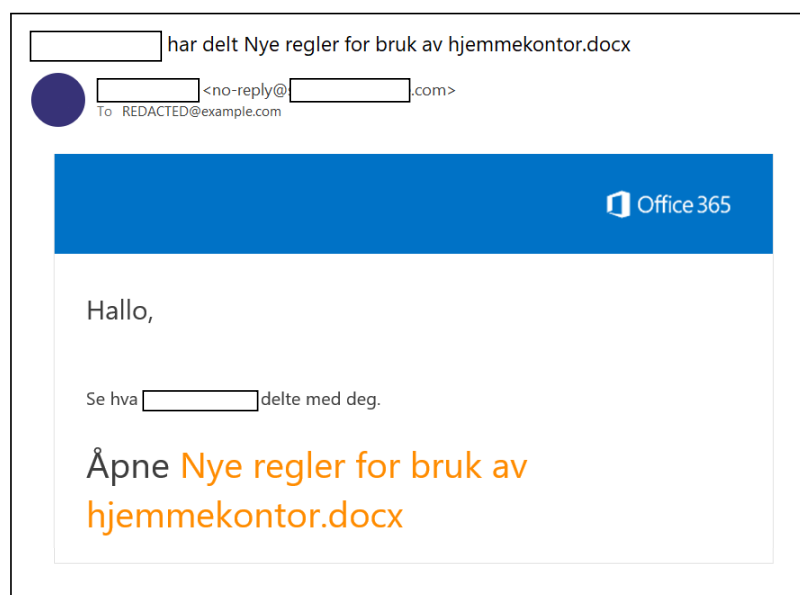


Figure B.29: Guidelines

Help Transfer Money

These types of phishing emails requests the assistance of the recipient in the transferring of money, promising something in return for the effort.

Figure B.30 displays an example of an email within this Content category.

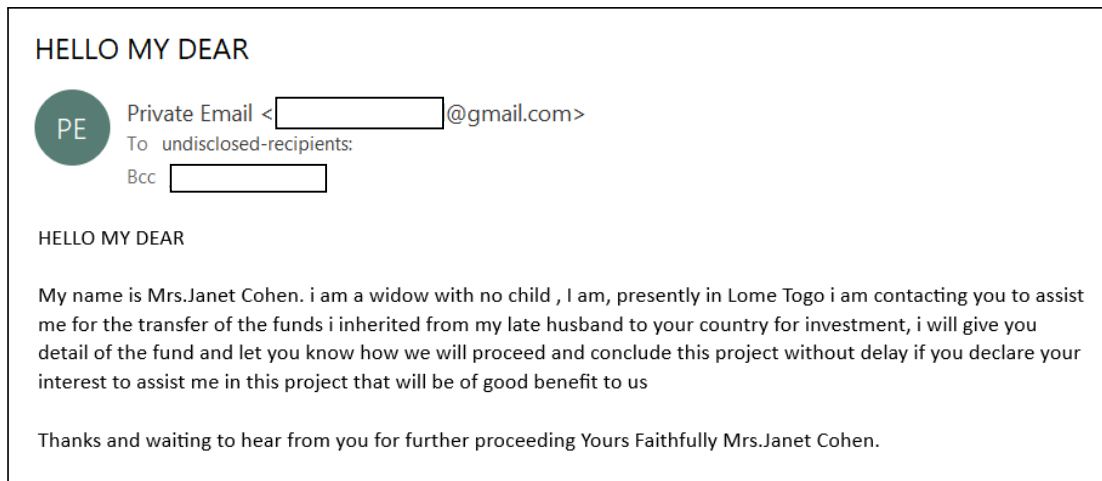


Figure B.30: Help Transfer Money

Inheritance

The Inheritance category consists of the emails conveying that there is an inheritance available for the recipient, and that an action has to be performed in order to access said inheritance.

Figure B.31 displays an example of an email within this Content category. The email informs the recipient that there is an unclaimed inheritance in their name.



Figure B.31: Inheritance

Invoice

The Invoice Content category embodies all the phishing emails that use the lure of a supposed invoice in order to trick the victim into performing an undesirable action.

Figure B.32 displays an example of an email within this Content category.

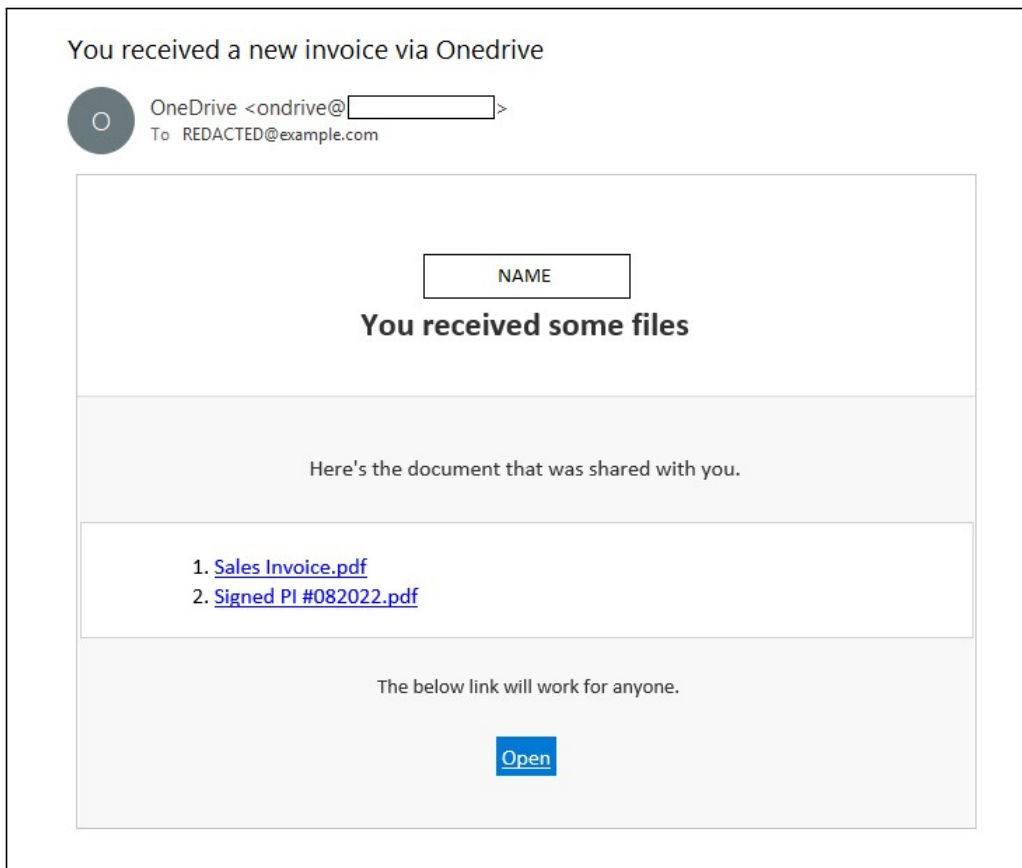


Figure B.32: Invoice

Item Won

These types of phishing emails notifies that the recipient has won an item, and that an action has to be performed in order to received said item.

Figure B.33 displays an example of an email within this Content category. The email notifies the recipient that they have won an iPhone from a Facebook challenge, and that they have to enter a site to claim said price.

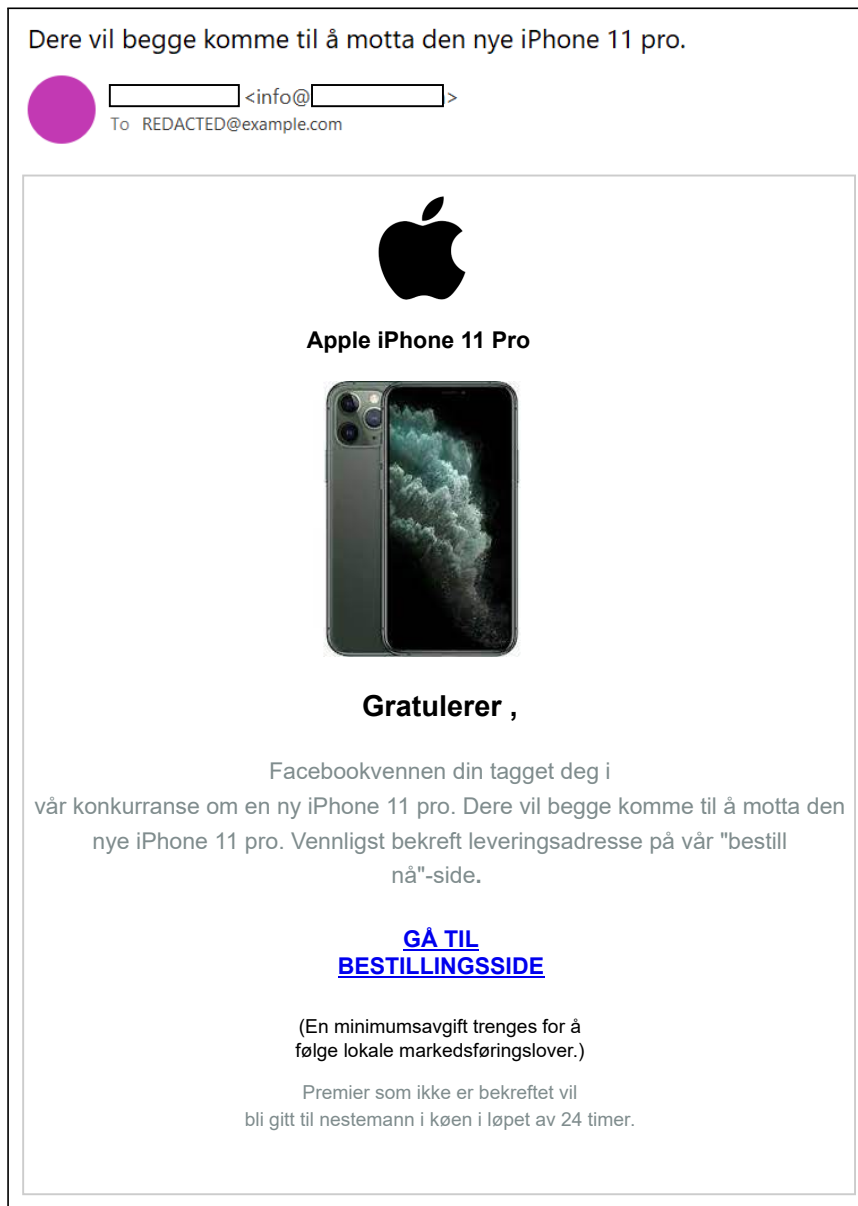


Figure B.33: Item Won

Job Application

This Content category embodies all phishing emails masked as a job application for the receiving organization.

Figure B.34 displays an example of an email within this Content category. The email contains an application for an advertised position, and linking to their online CV and application letter with a note that they may have to "run" the files in order to view them.

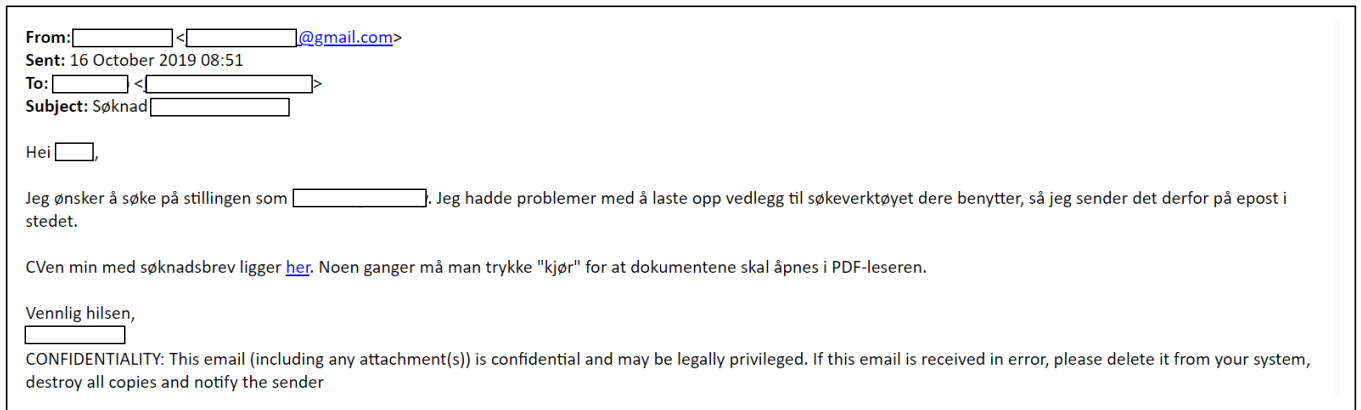


Figure B.34: Job Application

Job Write-Up

This Content category embodies all phishing emails notifying the recipient that they have been written up by their employing organization.

Figure B.35 displays an example of an email within this Content category.

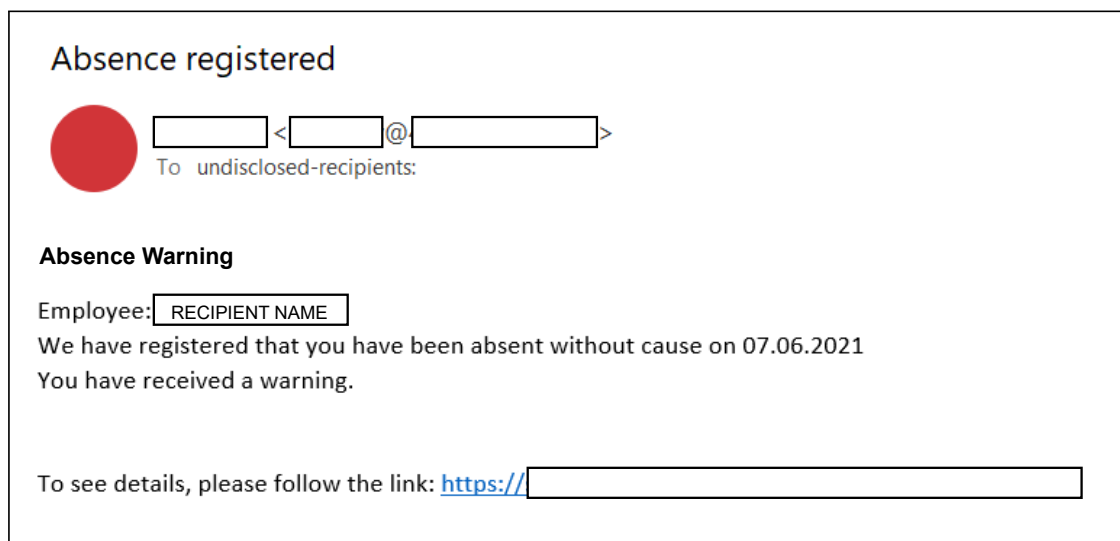


Figure B.35: Job Write-Up

Legal Document

This category consists of the phishing emails masked as a form of legal document.

Figure B.36 displays an example of an email within this Content category.

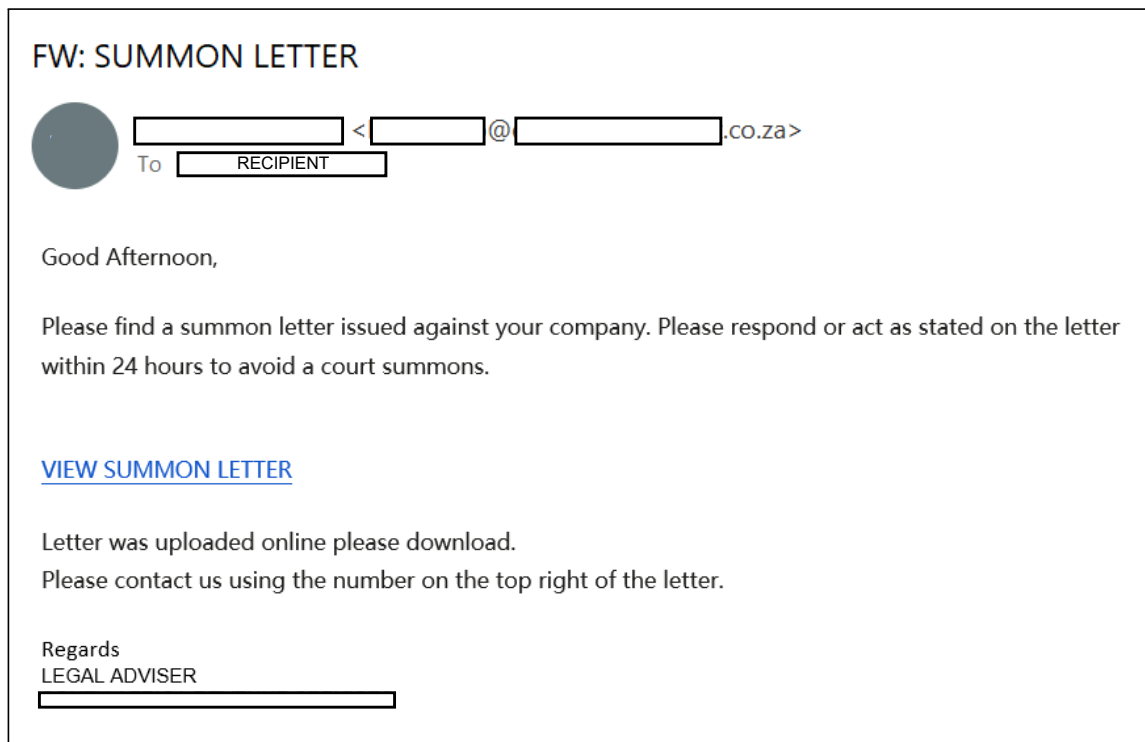


Figure B.36: Legal Document

Mailbox Full

The Mailbox Full category consists of the phishing emails notifying the recipient that their mailbox is full/close to full, and that an action has to be performed in order to increase the mailbox size.

Figure B.37 displays an example of an email within this Content category.

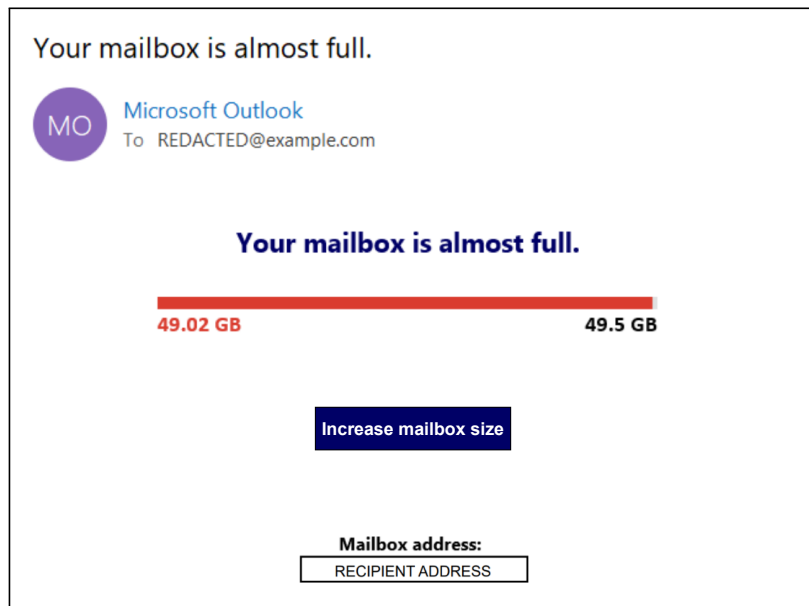


Figure B.37: Mailbox Full

Meeting Invitation

This Content category encompasses the phishing emails disguised as meeting invitations.

Figure B.38 displays an example of an email within this Content category.

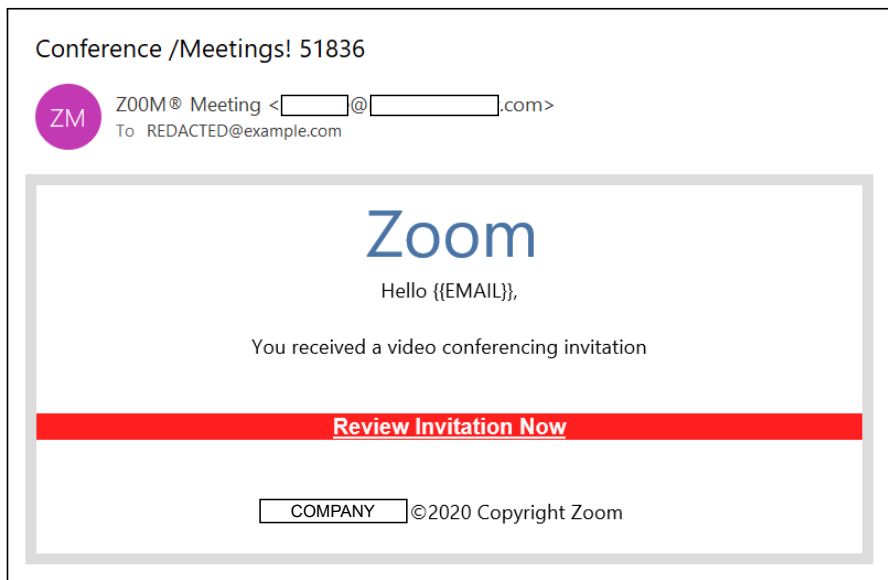


Figure B.38: Meeting Invitation

MFA Activate

The MFA Activate Content category consists of emails notifying the recipient that they have to activate multi factor authentication on their account.

Figure B.39 displays an example of an email within this Content category. The email notifies the recipient that they have to activate two-factor authentication in order to protect their account.



Figure B.39: MFA Activate

Money Received

This Content category revolves around phishing emails conveying to the recipient that they have received money, and that an action has to be performed in order to access said money.

Figure B.40 displays an example of an email within this Content category. The email informs the recipient that they have earned a significant amount of money on an account tied to their name, and that they have to perform a sign-in/account creation to claim said money.

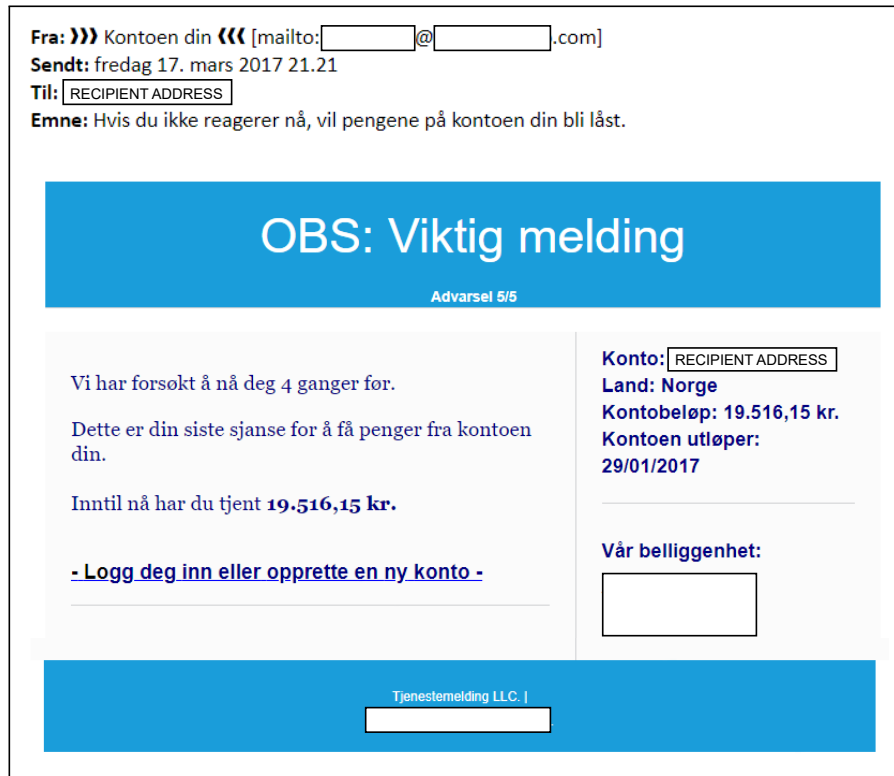


Figure B.40: Money Received

Money Won

This Content category revolves around phishing emails conveying to the recipient that they have won money, and that an action has to be performed in order to get said money.

Figure B.41 displays an example of an email within this Content category. The email notifies the recipient that they have won a large sum of money in a lottery, and that they have to enter a linked site to claim said money.



Figure B.41: Money Won

New Message

This category of phishing emails notifies the user that they have received a message, and that an action has to be performed in order to view said message.

Figure B.42 displays an example of an email within this Content category. The email contains a notification from LinkedIn, informing the recipient that they have received a new message.



Figure B.42: New Message

New Task

This category of phishing emails notifies the user that they have been assigned a new task to perform.

Figure B.43 displays an example of an email within this Content category.

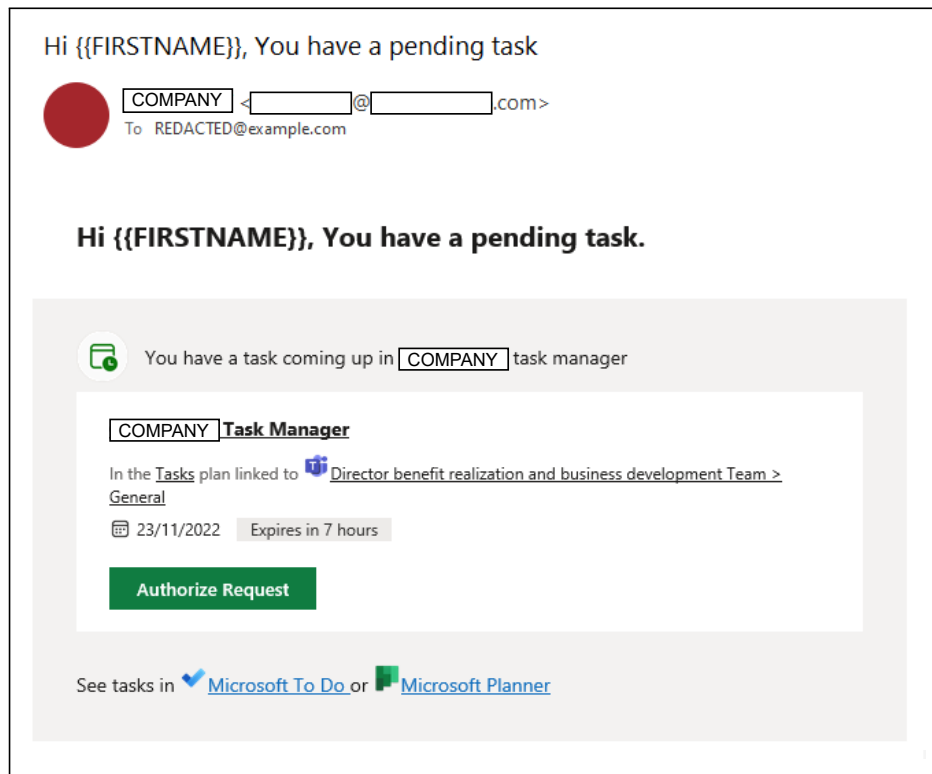


Figure B.43: New Task

Notification

This Content category embodies notices over email that cannot be tied to any specific notification type.

Figure B.44 displays an example of an email within this Content category.

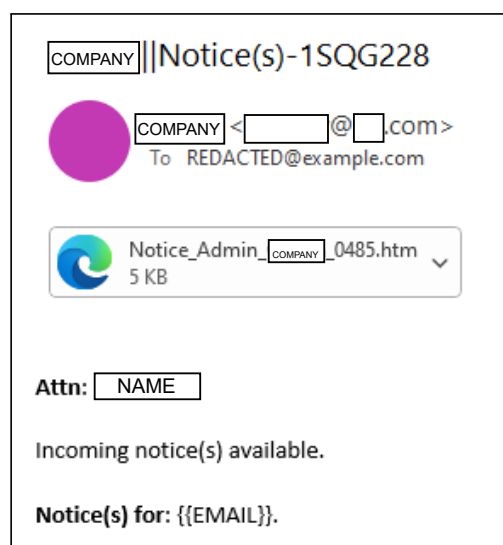


Figure B.44: Notification

Offer / Deal

This type of phishing emails presents an offer or a deal for the recipient on certain products or services.

Figure B.45 displays an example of an email within this Content category. The email informs the user that they can get a 1 year free HBO Nordic trial by activating their account.



Figure B.45: Offer / Deal

Open / Receive Mail

This Content category consists of the phishing emails where the user has to perform an action in order to open or retrieve a delivered email.

Figure B.46 displays an example of an email within this Content category.

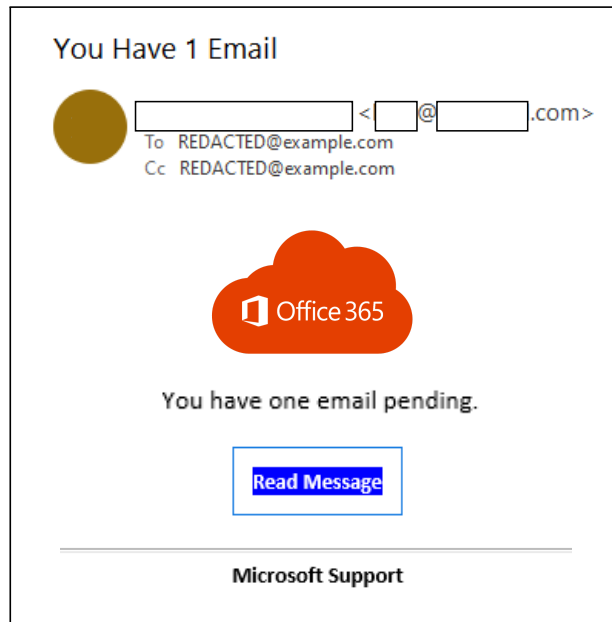


Figure B.46: Open / Receive Email

Password Changed

This Content category encompasses all phishing emails where the recipient is notified that their password has been changed.

Figure B.47 displays an example of an email within this Content category. The email informs the recipient that their Apple-ID password has been changed, and to update their password through the linked site if this was not a familiar change.



Figure B.47: Password Changed

Password Expires

This Content category encompasses all phishing emails where the recipient is notified that their password is about to expire, and that an action has to be performed to update/keep their password.

Figure B.48 displays an example of an email within this Content category.

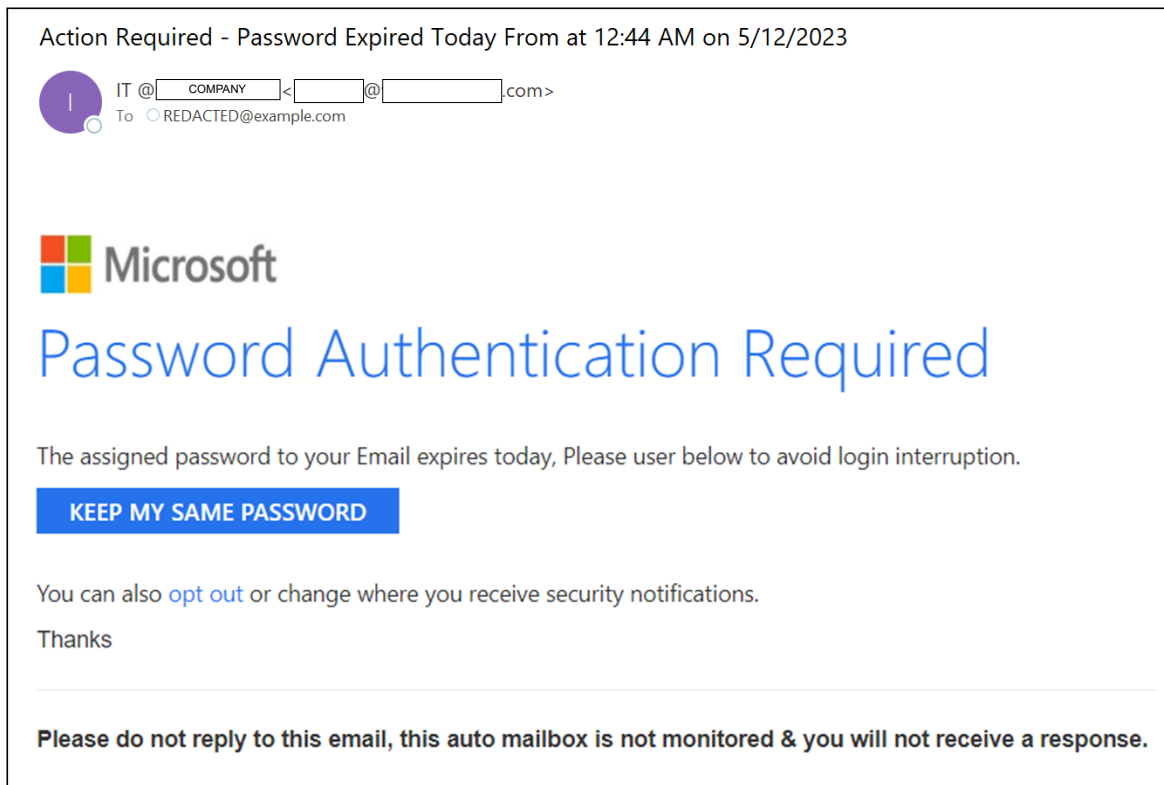


Figure B.48: Password Expires

Password Reset Request

This Content category encompass all phishing emails notifying the recipient that a password reset of their account has been requested.

Figure B.49 displays an example of an email within this Content category. The email notifies the recipient that a password reset has been requested for their Apple account, and to enter a linked site if this was not a familiar request.



Figure B.49: Password Reset Request

Payment Remittance

The Payment Remittance category consists of emails detailing payment remittances, either from a completed remittance or a requested remittance.

Figure B.50 displays an example of an email within this Content category.

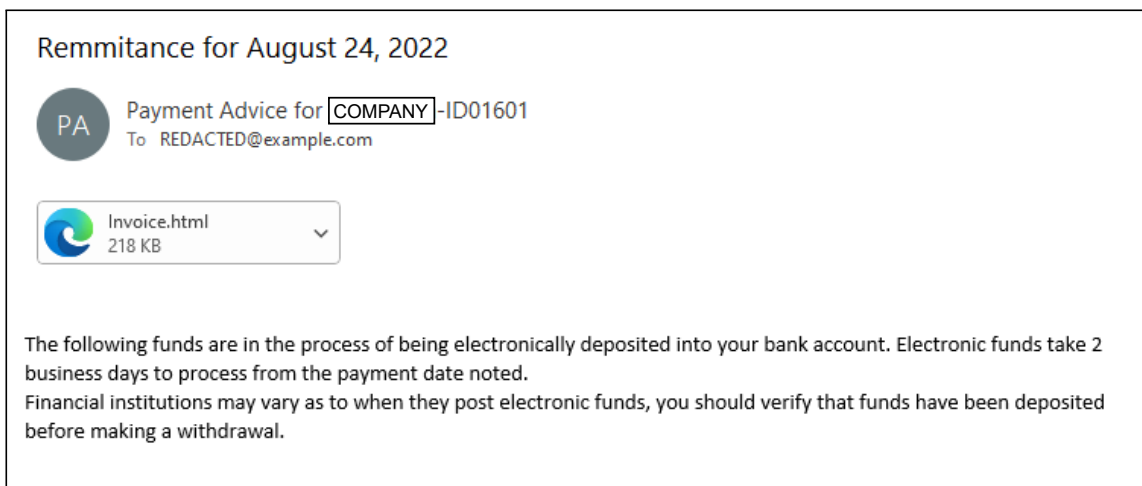


Figure B.50: Payment Remittance

Payment Sent

This Content category notifies the user that a payment has been sent from their account.

Figure B.51 displays an example of an email within this Content category. The email notifies the recipient that a transaction has been conducted from their account from an unfamiliar IP address, and that the recipient may abort said transaction by clicking a linked URL.

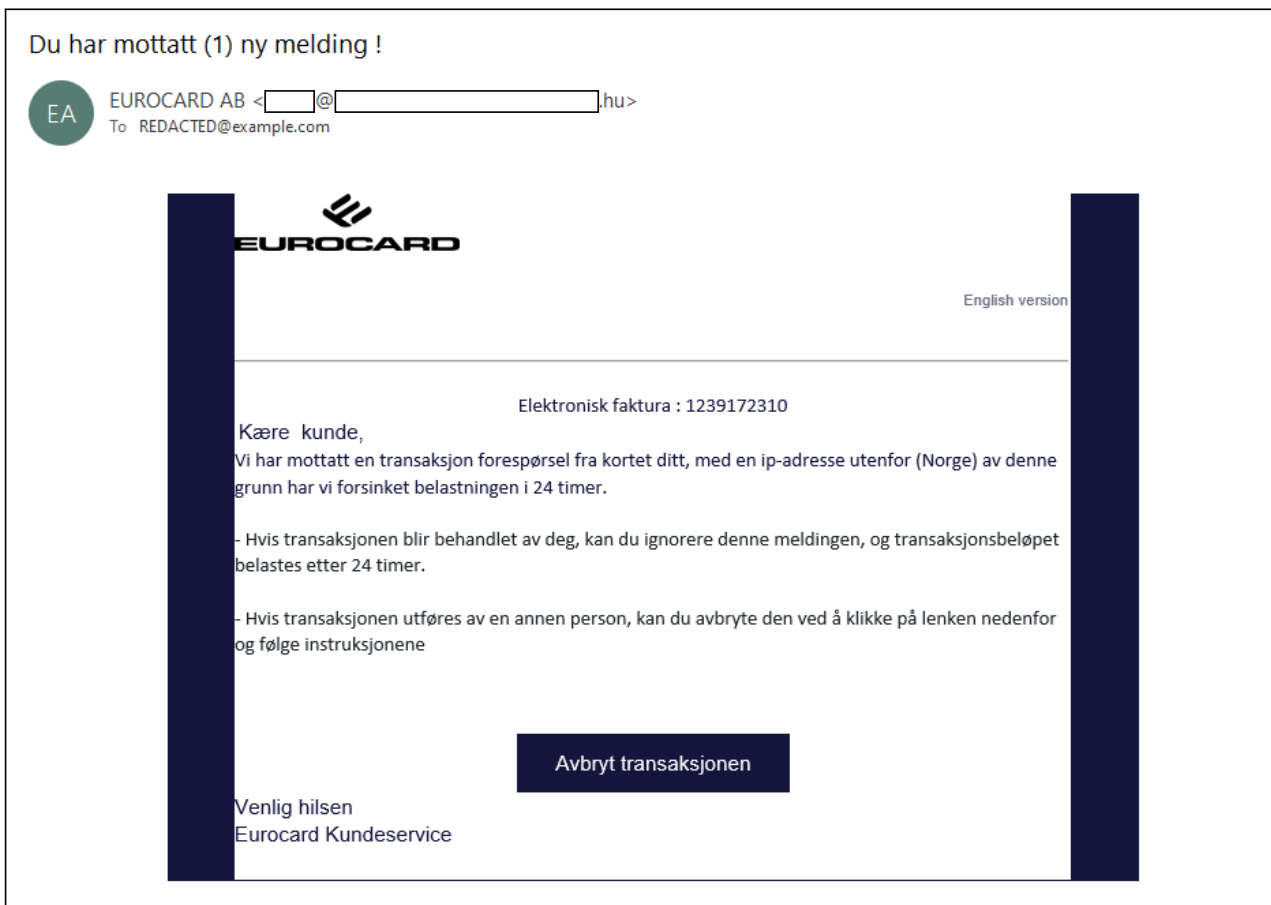


Figure B.51: Payment Sent

Payslip Details

The Payslip Details category consists of emails providing the recipient with information regarding their payslip.

Figure B.52 displays an example of an email within this Content category.



Figure B.52: Payslip Details

Police Notification

The Police Notification category concerns all emails disguised as notices from the police.

Figure B.53 displays an example of an email within this Content category. The email informs the recipient that a summon has been requested from the police department, with additional details in the attached image.



Figure B.53: Police Notification

Post Package

The Post Package Content category embodies all the phishing emails that uses the incentive of a package yet to be delivered in order to lure the recipient into performing an undesirable action. For example, requesting that a delivery fee is paid for the package to be delivered.

Figure B.54 displays an example of an email within this Content category. The email informs the recipient that a payment has to be performed in order for a package to be delivered to them.



Figure B.54: Post Package

Purchase Confirmation

This Content category encompasses the phishing emails containing a confirmation of a recent purchase.

Figure B.55 displays an example of an email within this Content category. The email contains a purchase confirmation for an iPad purchased by the recipient.

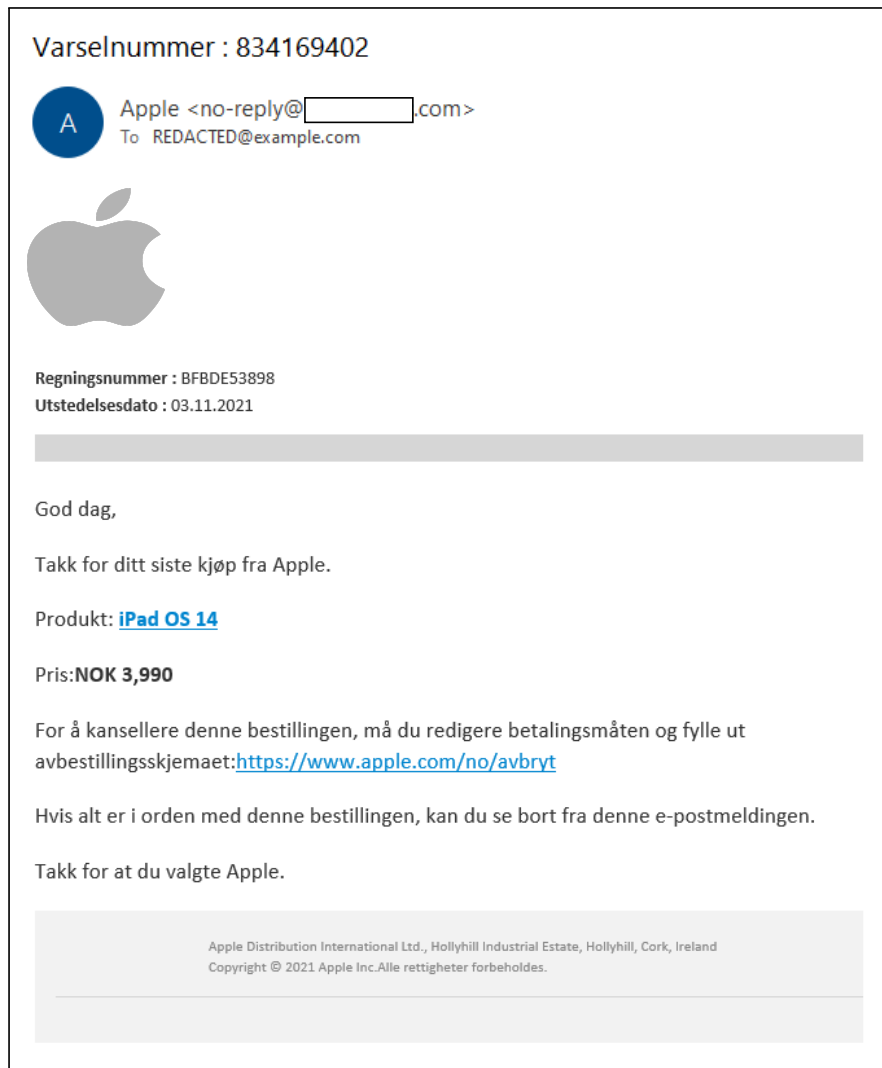


Figure B.55: Purchase Confirmation

Purchase Order

This Content category encompasses the phishing emails containing a purchase order for the recipient's organization.

Figure B.56 displays an example of an email within this Content category.

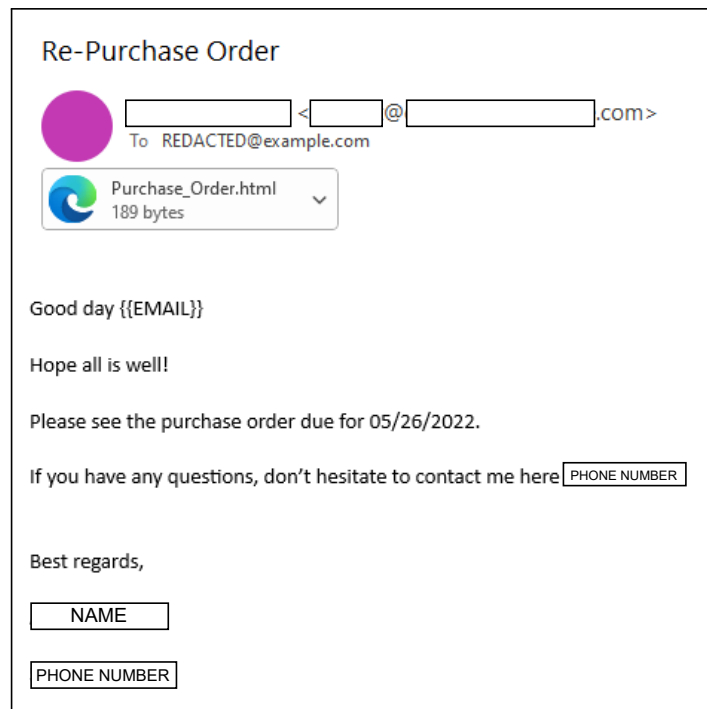


Figure B.56: Purchase Order

Questionnaire

This Content category consists of the phishing emails disguised as a questionnaire for the recipient.

Figure B.57 displays an example of an email within this Content category. The email links to a questionnaire sent out by TV 2 Play, where the recipient receives a 50€ gift card upon completion.

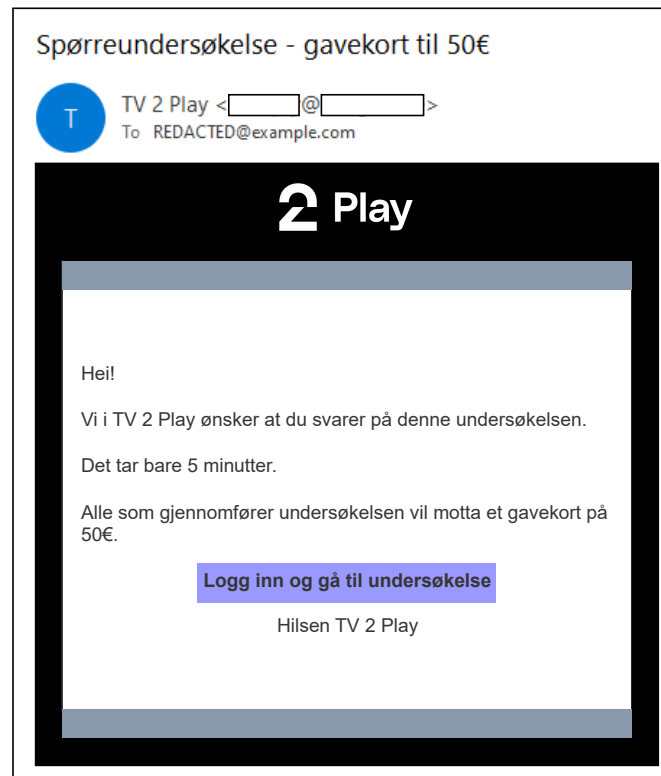


Figure B.57: Questionnaire

Refund

This type of phishing email informs the recipient that they are inclined to a refund, and that an action has to be performed in order to get said refund.

Figure B.58 displays an example of an email within this Content category. The email notifies the recipient that they are inclined to a tax refund, and to enter the linked site to claim said refund.

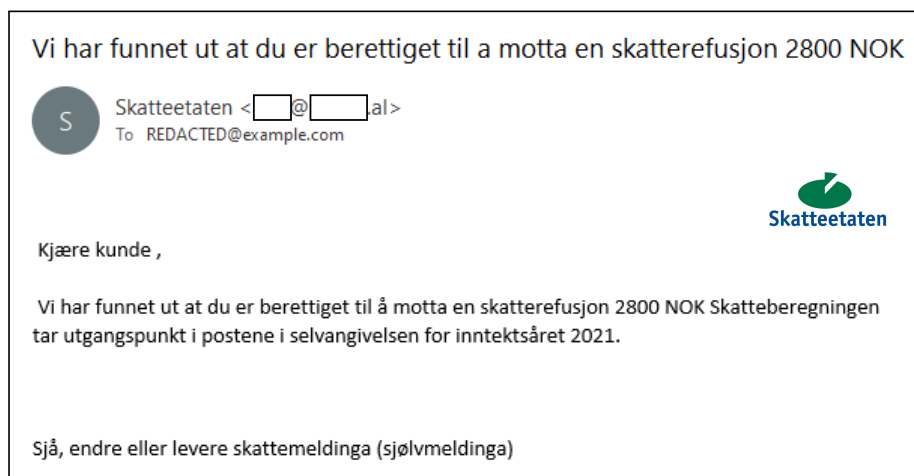


Figure B.58: Refund

Release From Quarantine

This Content category consists of the emails notifying the recipient that there are emails quarantined away from the recipient, and that an action has to be performed in order to release said mails.

Figure B.59 displays an example of an email within this Content category.

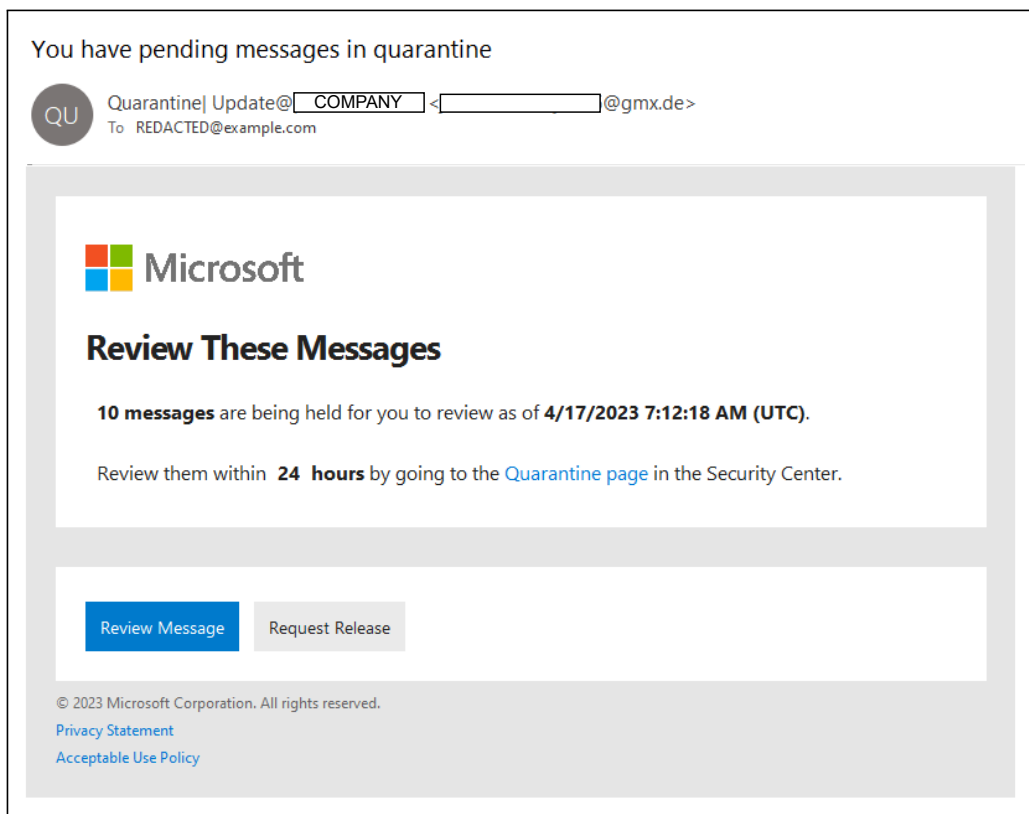


Figure B.59: Release From Quarantine

Server Full

The Server Full category consists of the phishing emails notifying the recipient that a server associated with them is at full/close to full capacity, and that an action has to be performed in order to increase the server's storage size.

Figure B.60 displays an example of an email within this Content category.

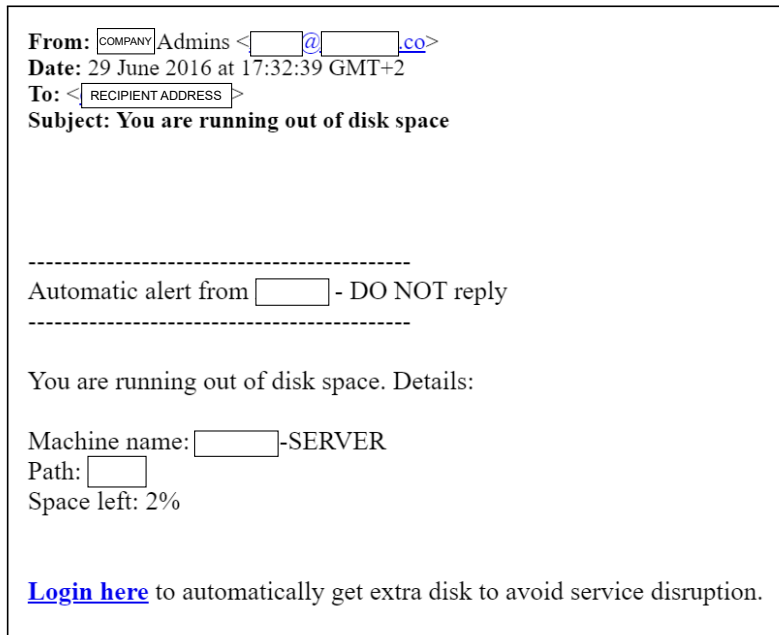


Figure B.60: Server Full

Server Stopped

The Server Stopped category consists of the phishing emails notifying the recipient that a server associated with them has stopped, and that an action has to be performed in order to resolve the issue.

Figure B.61 displays an example of an email within this Content category.

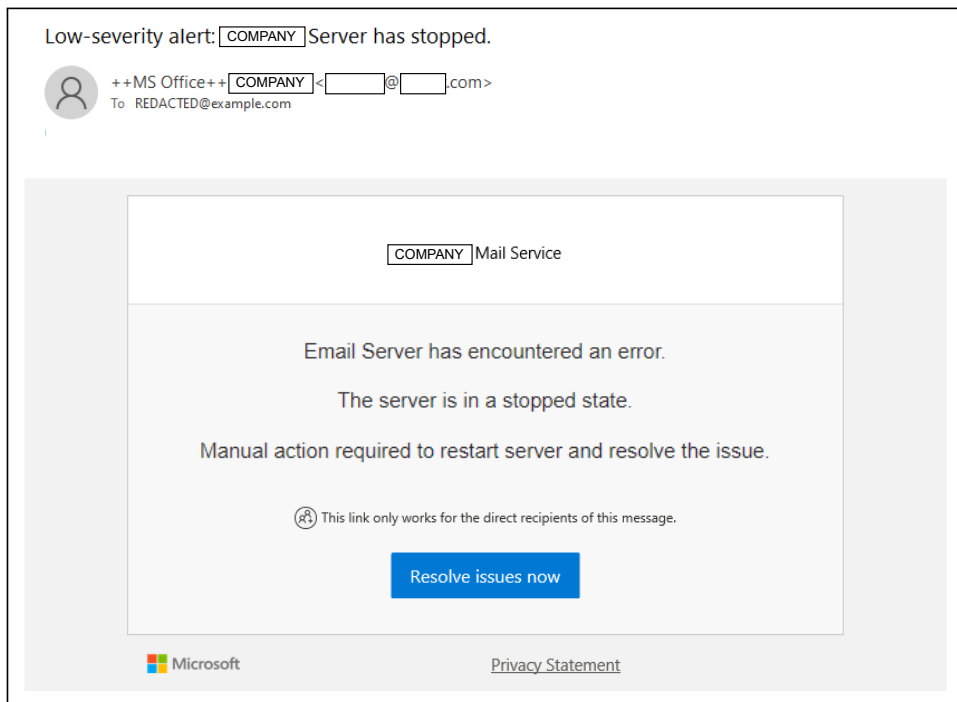


Figure B.61: Server Stopped

Shipping Details

This Content category encompasses all emails containing shipping details for an order destined to the recipient.

Figure B.62 displays an example of an email within this Content category.

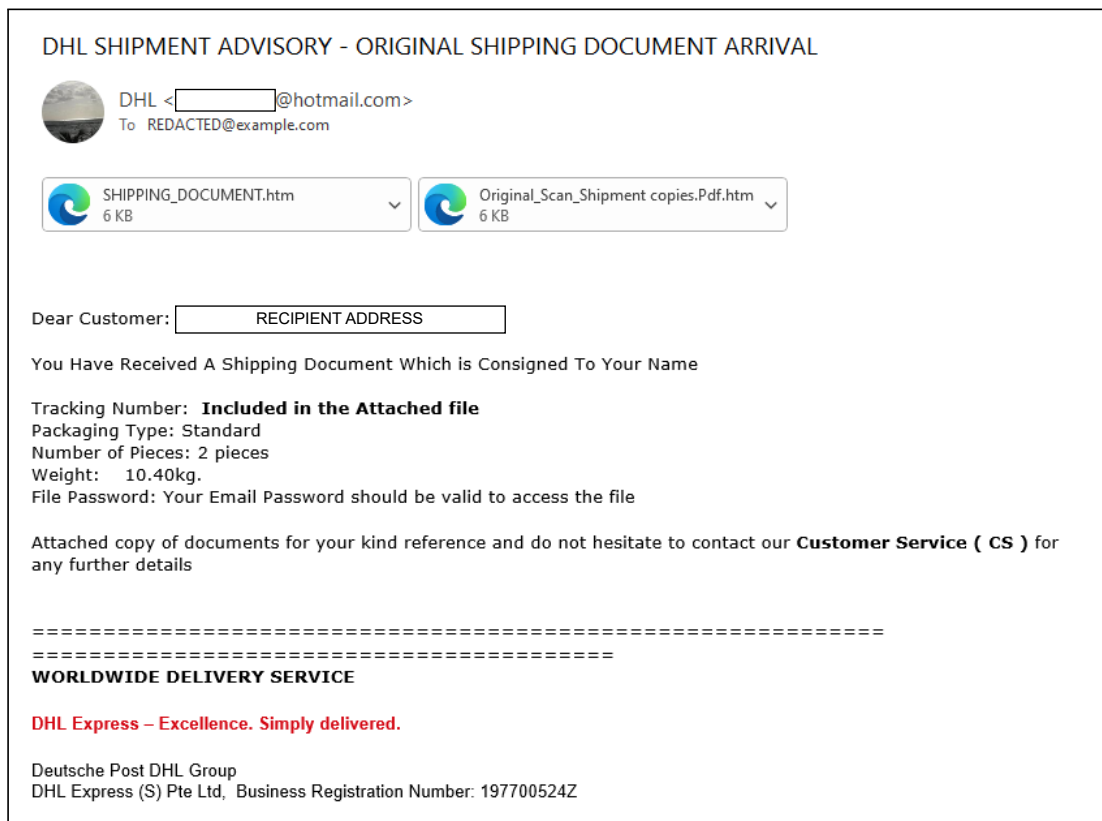


Figure B.62: Shipping Details

Sign Document

This Content category consists of all phishing emails disguised as a document signature request.

Figure B.63 displays an example of an email within this Content category.

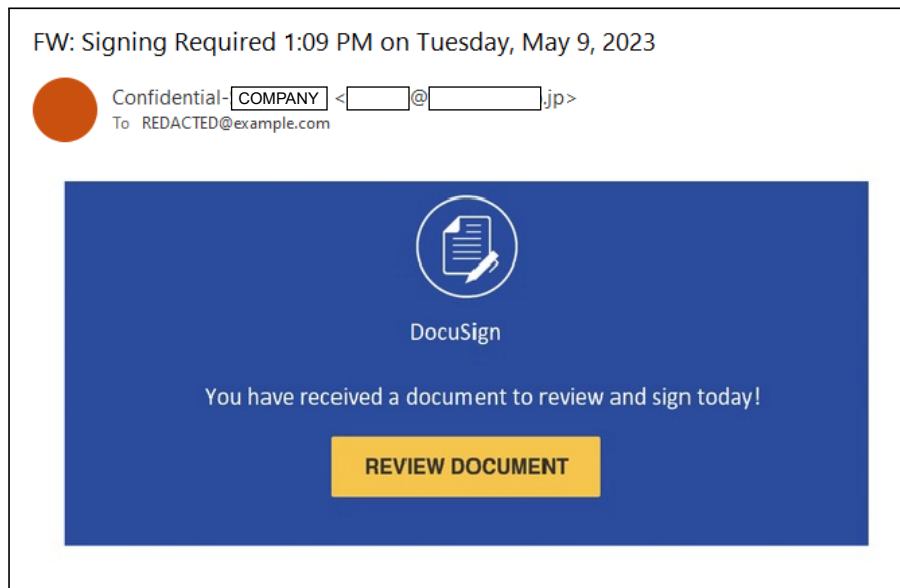


Figure B.63: Sign Document

Suspicious Activity

These phishing emails notifies the recipient that there has been detected suspicious activity on their account, and that an action has to be performed in order to validate or secure their account.

Figure B.64 displays an example of an email within this Content category. The email notifies the recipient that there has been observed suspicious activity on their account, and to entered a linked site to review said activity.

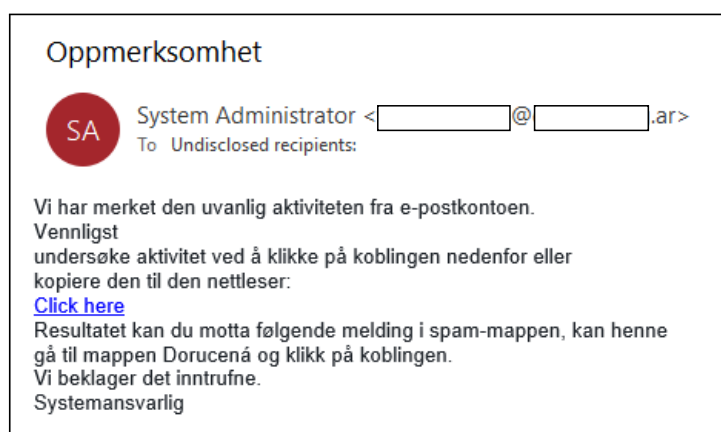


Figure B.64: Suspicious Activity

Trademark

Phishing emails detailing trademark notices are gathered in this Content category.

Figure B.65 displays an example of an email within this Content category.

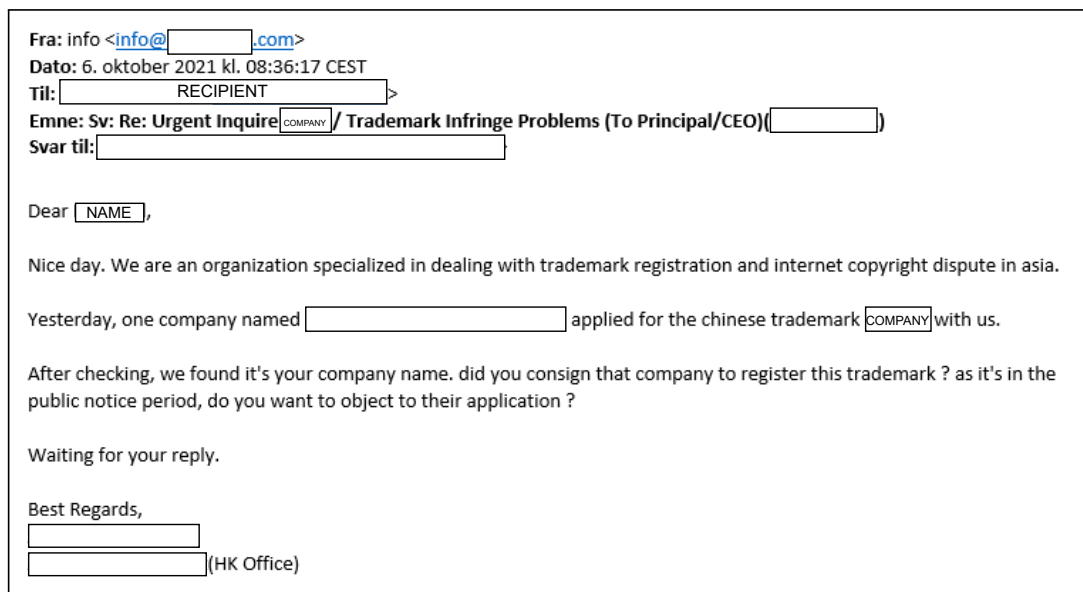


Figure B.65: Trademark

Update Software

The Update Software category consists of phishing emails notifying the recipient that they have to perform an update on a specific software. This could for instance have to done be in order to retain access to the given software.

Figure B.66 displays an example of an email within this Content category.

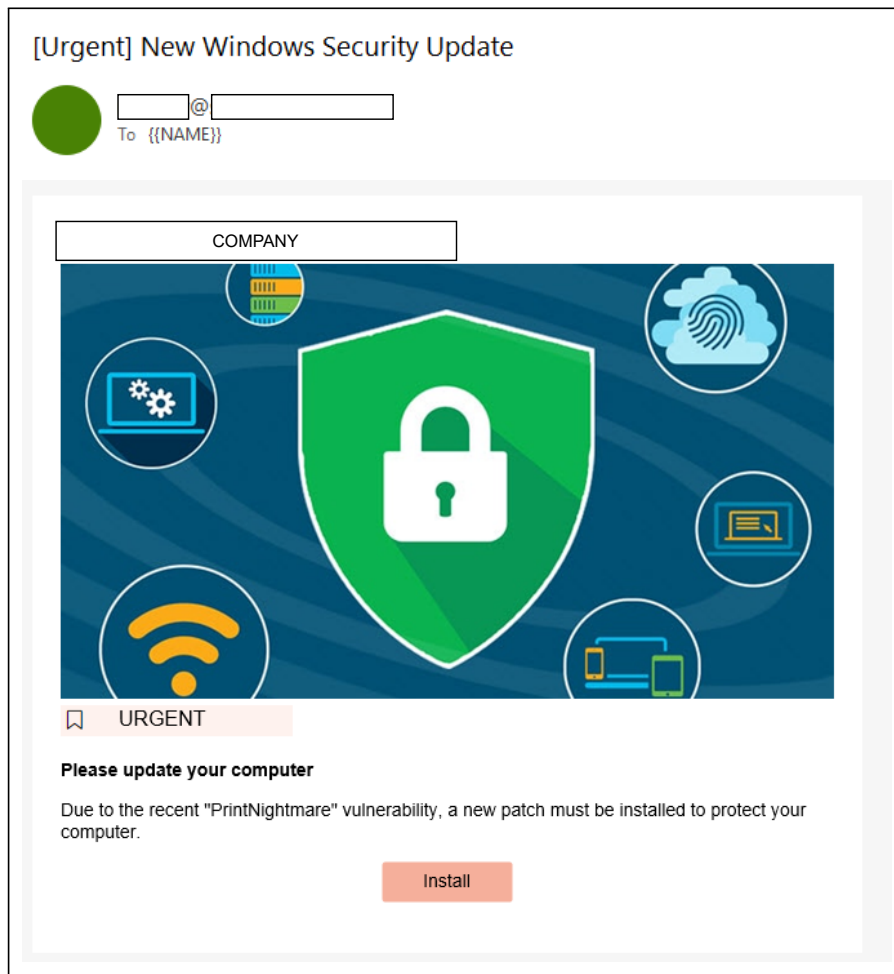


Figure B.66: Update Software

Verify Email

This Content category encompasses all phishing emails requesting the recipient to validate their email, for instance in the case of an account creation or to confirm that the address is still in use.

Figure B.67 displays an example of an email within this Content category.

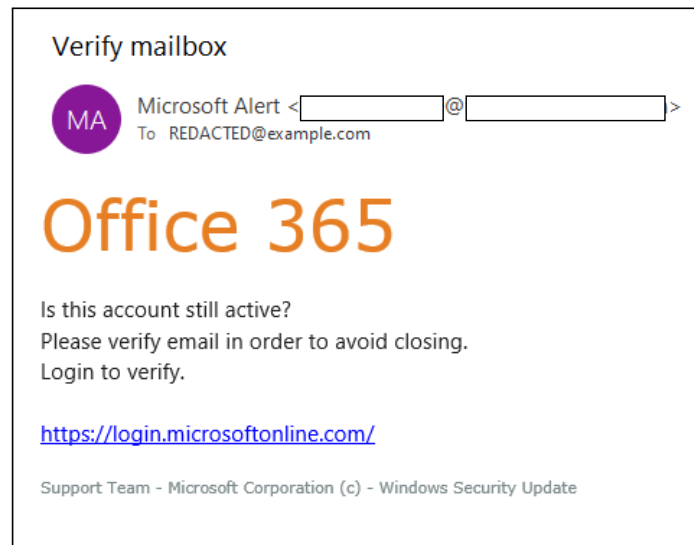


Figure B.67: Verify Email

Voice Message

This Content category concerns all phishing emails disguised as a voice message for the recipient.

Figure B.68 displays an example of an email within this Content category.

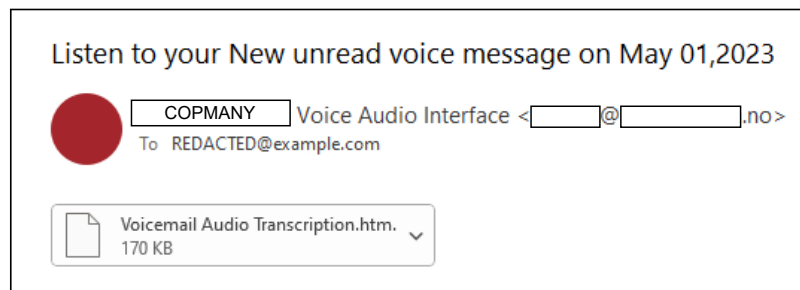
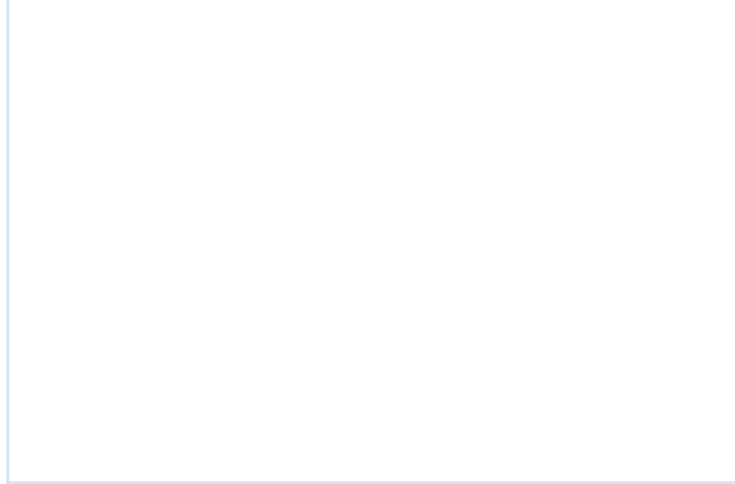
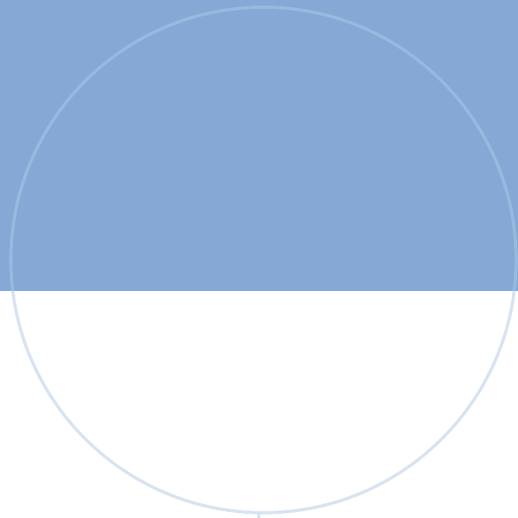


Figure B.68: Voice Message



 **NTNU**

Norwegian University of
Science and Technology