Degree project

How mail components on the server side detects and process undesired emails: a systematic literature review

Bachelor Degree Project in informatics IT610G

First Cycle 22.5 credits
Spring term Year 2021-06-13

Student: Anton Ahlborg
Email: a17antah@student.his.se

Supervisor: Thomas Fischer
Examiner: Jianguo Ding

Acknowledgements

I will start off by thanking my family and my supervisor Thomas Fischer for their amazing support throughout this project.

Abstract

As the use of emails increases constantly every year, so do the reports of various victims in society, on companies and individuals who have been affected by these undesirable emails in the form of spam, spoofing and phishing in their inbox. The effect of undesirable emails are many, but in summary, they cost the society and organization immense amount of money. This study will aim to understand why these emails still make its way into a user’s inbox by identifying current existing solutions that are being used by email servers to evaluate incoming undesirable emails.

The analysis of the study shows that there are shortcomings in the solutions that are being used today, which lead to undesirable emails reaching a user’s inbox, and it is likely to continue in the near foreseeable future, unless research and or actions are applied to some of the brought-up issues in this study, namely problems with adoption and usage rate of authentication protocols, technical issues within authentication protocols and emails being wrongly classified by today's filtering techniques.

Table of contents

1 Introduction

It has long been the case that you receive undesired emails in your inbox, Jung & Jo (2003) states that it is already an ongoing issue where 30% of the emails sent accounts for spam, Blanzieri & Bryl (2008) states in 2008 that 75-80% of all email being sent are spam and Giorgi et al. (2020) say that today 50-60% of all the email sent consists of spam. There is a downgrade in the amount of spam versus desired emails in recent days, there is however room for more research on the subject, since it is evident that today it is still an ongoing issue where undesired emails bypass existing solutions and reach a user's inbox.

IBM (2020) states that today the average cost of a data breach is estimated to cost an organization 3.86 million dollars. Verizon (2021) analyzed 3950 data breaches that occurred 2020, 831 of these data breaches had its origin from a phishing email. Phishing emails are today a highly effective method to deliver an attack because it can be combined with email spoofing, making it appear as the attacker are sending the email from a legitimate source. This is possible because of faulty standardized authentication protocols are still being used today, namely Sender Policy Framework (SPF), DomainKeys Identified Mail and Domain-based Message Authentication (DMARC). In 2018 it was possible to successfully send forged emails to 34 out of 35 of today's most used email services (Hu & Wang, 2018).

Today's utilized filtering techniques have issues with high-false positive ratio (legit emails being wrongly classified as spam) (Chanti & Chithralekha, 2020). Thus, leading email providers to ease up the filtering regulation, because today it cost more money to lose a legitimate email compared to letting an undesired email reach a user's inbox.

The purpose of this study is to identify what existing countermeasures are being applied towards undesired emails and evaluate how effective these are. By understanding what protections exist and what opportunities you have to protect yourself and your organization from it, and understanding its shortcomings, we can start making more educated decisions in the context of applying countermeasures towards undesirable emails.

The study is carried out as a systematic literature review study with a described process for selection, review and analysis of existing articles in three designated databases: ACM Digital Library, IEEE Xplore and SpringerLink. Thus, in order to carry out a study with a good ground of valid research on the subject to conduct a proper analysis.

2 Background

The background chapter aims to provide necessary information to the reader for the upcoming chapters.

2.1 Undesired Electronic Messages

- **Deceptive phishing**
  What the term phishing comes from is from fishing, meaning that the fisher is playing out an attacking role and is using a bait to catch fishes, the victims (Adil et al., 2020). The bait is most commonly something that takes advantage of social factors of the victim and hence why phishing is considered being a form of social engineering. The bait usually contains a link that redirects the victim to a website with malicious intent (typically designed to make the victim give up personal information), or an attachment containing malicious instructions (Adil et al., 2020). They are cleverly forged and disguised to appear legitimate and trustworthy. Often the perpetrators impersonate a company, authority or person who the victim is familiar with and trusts (Athulya & Praveen, 2020). The bait contains in most cases a message that manipulates the emotions of the victim, giving the victim a sense of urgency to click the link or download the attachment without properly analyzing the situation. According to Bruce Sussman (2020) the most common emotions the attacker targets are fear, greed, curiosity and helpfulness. Allodi et al. (2020) say deployment characteristics of this phishing attack are hit-or-miss, and usually only one attempt by the perpetrator is made. The attacker is not focused on a specific target in this context but focuses on the masses and sends their attack at a large volume, distributed by email. The more baits put out, the more sizeable returns the perpetrator will receive in this hit-or-miss nature (Allodi et al., 2020).

- **Spear phishing and Whale phishing**
  In this type of phishing attack, the scope is much more specific to their target group or individual. Characteristics of these type of strategies are that the perpetrator's collect information on the target group, such as an organization or individual and with this information engineer a custom-made bait that's specifically designed for the target group or individual. Allodi et al. continue that because of different attack dynamics, spear phishing and whale phishing are more versatile than the hit-or-miss large volume approach, as these strategies have a nature of consisting in multiple stages where the attacker can based on previous interactions form a new basis and strategy for further attacks against the target. Allodi et al. say that the target profile of these strategies is very well identified by the attacker before the actual attack launches, as it contains an information collection period where information is gathered about the target. Information gathered about the target can be, for example, habits and social surroundings, these types of attacks target and exploit the weakness of the victim (Athulya & Praveen, 2020; Allodi et al., 2020).

- **Email spoofing**
  Giorgi et al. (2020) say that a sender can forge the sender address in the email header so that the email appears to have originated from someone, or somewhere other than the actual source. The authors say this can be done by exploiting vulnerabilities existing in the Simple Mail Transfer Protocol (SMTP), that is being used in today's email transmissions. The authors say that currently, SMTP does not provide any

authentication mechanisms that can verify the origin of the sent email. Email spoofing is often used in combination with spear phishing and whale phishing, and are able to achieve high success, because of people feel more inclined to open an email when they think it is sent from a legitimate source (Giorgi et al., 2020).

- **Email Spam**
  A spam email is not necessarily malicious and commonly sent out for commercial purposes due to its low cost. But a spam mail may share characteristic traits of a phishing mail by containing malicious attachments or links to websites with malicious intent (Iedemska et al., 2014). The authors say that 85% of the worldwide spam is distributed by botnets (networks of compromised computers). According to Statista (2021) roughly 300 billion emails were sent every day during the year 2020, and which Giorgi et al. (2020) state spam emails accounts for 50-60% of the sent emails.

2.2 Email Structure

The author Both (2020) say that the primary email structure comprises two parts, the header and the message body. According to the author, the message body can contain ASCII plain text or components consisting of HTML messages, images, or other types. The author says that the email header contains records of an email's travels and can help tell the origin of the email. Each Mail Transfer Agent (MTA) adds information to the headers record about the email's passages according to Both. Inspecting the email headers to identify the source of an undesired electronic message is a common procedure the author says. When inspecting the email headers, you can identify where the email may have been delayed in its transit across the internet from sender to its receiver, and use this reference point as the basis for rejecting spam from this source (Both, 2020).

2.3 Email Sending Process

Both (2020) clarify some relevant terms that will be used when explaining the process of sending an email and its involved components.

- **Protocol**: Is a set of rules describing how to transmit data across networks.
- **SMTP**: Simple Mail Transfer Protocol: A protocol used to transfer emails.
- **POP**: Post Office Protocol: A protocol designed to allow single user computers to retrieve their emails from a POP server.
- **IMAP:** Internet Message Access Protocol: A protocol to allow a client to access their emails.
- **MTA**: Mail Transfer Agent: Transfers emails from one host to another host.

Both (2020) explain the process of when a user sends an email from one mail server to another in 8 distinct steps.

1. The email client adds an initial set of headers to the email that is going to be sent, this includes subject, date and the FROM: and TO: lines.
2. The sending email client uses SMTP in order to send the email to its local email (SMTP) server.

3. The local SMTP server receives the email and adds a RECEIVED: link in the headers that lists information where the email originated from, it will add its internet protocol (IP) address and hostname along with a timestamp and also add the emails recipient address.
4. The SMTP server parses the address to which is destined.
5. The SMTP server queries the Domain Name Server (DNS) for the mail exchanger (MX) record for the target domain.
6. The SMTP server then sends the email to the receiving SMTP server.
7. The receiving SMTP server adds another RECEIVED: entry to the headers.
8. The receiving SMTP server allocates the email to the receiving client, the client retrieves the email by using IMAP or POP.

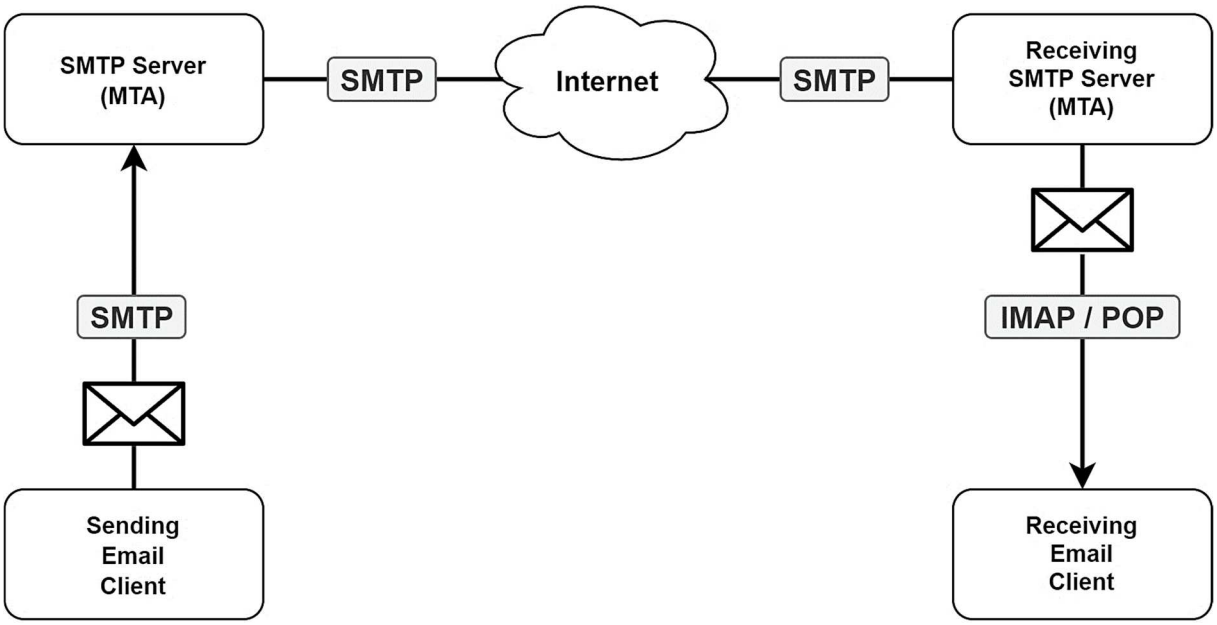

Figure 1: Email data flow, based on Both's (2020) description (Author's own).

3 Problem Definition

The Radicati Group, Inc. (2020) say that email during the past three decenniums has continuously kept growing as a popular way for users in both organizational and personal context to communicate. 4,037 billion people are using email 2020. According to the same report, the total processed emails per day are 306.4 billion and estimated to grow to 361.6 billion per day by 2024. Not only the number of users and emails have kept growing over past decenniums, but so has also the amount of spam and phishing emails. Out of all the billion processed emails sent every day, it is estimated that 85% of these emails are spam (Armin et al., 2021). However, Giorgi et al. (2020) claim the number to be more closely to 50-60%, none the less a majority of today's email traffic are made up from spam emails.

Undesired emails have today become a global problem, and the cost caused by this are immense towards organizations. According to IBM's (2020) data breach study, the global average cost of a data breach in 2020 was 3.86 million dollars. Verizon's (2020) data breach report analyzed 3950 data breaches during the year 2020. The authors say these breaches were caused by 23619 reported security events; the authors explain security events as; an event that successfully surpassed existing security, but no confidentiality data were exposed because of the security event. The authors of the report say that phishing made up for 16% of these security events (3779 security events) and among these 3779 security events caused by phishing, 22% (831) of the events lead to a successful data breach. This indicates that phishing is a highly effective method to deliver an attack, and thus the authors ranked phishing as the primary threat towards organizations in this context. Giorgi et al. (2020) claim that phishing emails today are highly effective due to it is often used in combination with a spoofed email. The authors continue to elaborate that this combination is possible because weak existing solutions to validate the authenticity of an email.

Today's spam emails have evolved from being a way for companies to promote themselves, to a vehicle with malicious content, Giorgi et al. (2020) states the following: *"The most widespread spam attacks are scam emails where the malicious user tries through confidence tricks to deceive the victim into stealing personal information."* (p 1). Verizon (2020) states that most malwares today are distributed via emails.

According to Al-Hussaini et al. (2021) the easiest way into a system is done by targeting humans. The authors say this is often carried out by social engineering in the form of phishing email. The fundamentals of this issue lays in the fact that undesired emails keep making its way into an inbox of a user. The existing solutions to fight this have their limitations and thus users and the solutions are in an interplay to best protect themselves and organizations from a successful attack from an undesired email. Athulya & Praveen (2020) say that "*[t]he best method to avoid phishing attacks is to create awareness for the users about the types of phishing attacks within the network."* (p. 342).

3.1 Aim and Purpose

Man is in the end, a weak link but it is still only us who in the end can make a correct classification of an email, if this email is an undesired or desired email. Therefore, the protection is of course affected by the end user's ability to do this decisively, time and stress can affect the concentration in place, so mistakes can easily happen. The best solution would appear to be if there exists, or was created a clearer support mechanism to help the end user to

make more educational decisions, whether an email is a desired or undesired email, as well as if there were methods to build more intelligence into this mechanism that can mimic the end user's needs and ways to work. The study's scope will be to answer the following research question:

How do mail components on the server side detects, and process undesired emails?

This research question aims at covering what functions and methods are being used in today's mail processing components to handle the enormous amounts of undesired emails that are circulating today. The study will also aim to cover how these methods work, evaluate what their potential flaws and limitations are in order to understand what has led the end user to become the last, but also the best defense.

3.2 Related Work

Jung & Jo (2003) states that back in the year 2003, 30% of all the sent emails were spam. The authors also point out that spam at this day and age is not only for commercial purposes but can contain viruses. The authors continue by propose two ways to handle this issue, text-based solutions and simple rule-learning classification systems to recognize junk emails by analyzing keywords in messages. The authors name a few rule-learning text classification solutions (called content-based nowadays) to handle the issue of undesired emails; RIPPER which is based on keyword spotting ruleset generated by a user's manual settings, Naïve Bayes based methods is also named by the authors as an efficient keyword-based approach that utilizes a probabilistic classification by using features extracted from emails.

According to Blanzieri & Bryl (2008) the problem of undesired electronic emails was a major issue in 2008. The authors say that spam constitutes up to 75-80% of the total amount of emails sent. The author say it was estimated that spam in 2005 caused financial losses of $50 billion. Because of SMTP not providing no reliable mechanisms to validate the sender's identity, by securing SMTP was determined to be a good way of stopping some of the spam (Blanzieri & Bryl, 2008). The authors say that the proposed solution to secure the SMTP is to deploy the authentication protocol Sender Policy Framework (SPF). According to the authors learning-based filtering is a popular countermeasure towards spam, it is explained that these methods evaluate emails contents on the likelihood of it being a desired email or undesired email, the accuracy of these filters in 2008 are already showing promising results with above 90% detection accuracy (Blanzieri & Bryl, 2008). The learning-based methods mentioned in the article are Naïve Bayes, Support Vector Machine (SVM), k-Nearest Neighbor (k-NN). Blacklist and whitelist are also being named as common anti-spam solutions, the essence of these lists is to deny or allow depending on their IP reputation (Blanzieri & Bryl, 2008).

Herzberg (2009) describes phishing to be one of the most harmful categories of spam. The author continues to explain three main mechanisms to block phishing and spam emails. Reputation mechanisms: the author explains this category as systems to map the identity of the sender. The author names blacklist and whitelists as common reputation mechanisms. Second mechanism the author explains is the authentication mechanism; these mechanisms work to authenticate the identity of the sender. The author lists three examples of such mechanisms are SPF, SenderID (SIDF) and DomainKeys Identified Mail (DKIM). SPF and SIDF are based on security routing DNS and DKIM are based on security of digital signatures. The last mechanism the author mention are content classifiers: it classifies emails based on their contents.

3.3 Limitations

Today's field of IT strengthens constantly at a rapid pace. Since literature reviews are based on previous existing research done in the field, and the state-of-the-art solutions that may be presented in this literature review might not still apply in the near future, and also any research published before the method's year selection criteria, or published research after this literature review is conducted will be left out. Which may or may not impact this study's result.

4 Method

In this chapter, the method that will be used to conduct the review will be presented and motivated. It will go into detail how the review is designed, how it will be conducted, and what analysis process is going to be used to examine the collected material.

4.1 Systematic Literature Review

Literature review plays as an essential part of academic research, Paré et al. (2015) says *"knowledge advancement must be built on prior existing work"* (p. 13). The authors continue that in order to push the knowledge frontier further, we must identify where the frontier is. By conducting reviews of relevant literature to an area, we can understand the depth of existing work and identify gaps to further explore (Paré et al., 2015). The authors continue that by summarizing and analyzing a related group we can test a specific hypothesis and evaluate the validity and quality of existing work to reveal weaknesses, inconsistencies and contradictions.

There are various ways to conduct a literature review, and the chosen method for this study is a systematic literature review. Based on the references, Paré et al (2015) suggests by conducting a systematic literature review we can look what past research has already done in the same field and build upon them by exploring a new research question with the aim to seek a fully investigated answer, and thus further push the knowledge frontier.

Kitchenham (2004) describes when a review should be conducted systematically as:
"The need for a systematic review arises from the requirement of researchers to summarise all existing information about some phenomenon in a thorough and unbiased manner." (p. 3).

Snyder (2019) means that the intention of a systematic literature review is to identify all empirical evidence that fits specified criteria in order to answer a research question or hypothesis. Searching literature systematically means more than seeking a quick answer, by demonstrating that the study has thoroughly searched for evidence will further enhance the credibility of the findings (Booth et al., 2016. By reviewing all evidence in with a systematic approach, the effect of bias can be minimized and thus lead to more reliable findings where conclusions can be drawn from (Booth et al., 2016; Snyder, 2019).

In order to achieve an answer to the study's research question, evidence must be searched for thoroughly and to assert credibility to the answer. It is determined that a systematic literature review will be an appropriate way to conduct this study.

According Snyder (2019) it exists four main phases in carrying through a systematic literature review, and they are:

- **Phase 1: Design the review**
  Snyder (2019) says that the design phase should aim to clarify the need for this research and what the contribution of this research can lead to. The author continues it should specify who the study's target audience are, what the specific goals of the research are, discuss what an applicable method to use in order to carry out the research and lastly be transparent on how it is being conducted, meaning, present the search strategy and all its included parts (search terms, databases and its inclusion and exclusion criteria).

- **Phase 2: Conduct the review**
  Snyder (2019) says that this phase should clarify if the search strategy needed modification to produce an appropriate sample. The author continues that this phase should present the plan on how the articles were selected, the result of the search strategy and how the quality of the search strategy and the selected articles were assessed.

- **Phase 3: Analysis of the review**
  Snyder (2019) says this stage should consider what information needs to abstracted to fulfil the purpose of the review, and what information is needed for the analysis. The author says that this stage should be documented and reported.

- **Phase 4: Structure and write the review**
  Snyder (2019) says it is important that the level of information provided is enough and appropriate to allow for transparency, so readers can judge the quality of the review. The result should be clearly presented, and clearly communicate what the study's contributions are (Snyder, 2019).

Based on Snyder's systematic review approach, certain stages have been identified to be of more importance. Due to limited time available, a full systematic review is not possible, hence why it has been decided to make a specific strategy for this review, based on identified key factors in Snyder's approach to a systematic review. Snyder's approach suggests having two or more reviewers of the materials, and to determine relevance by reading full text of all materials generated from the database searches. This was not determined to be possible in given time frame. Therefore, the primary deviation made to Snyder's approach is how the relevance of the material is assured. The relevance process chosen for this review is presented in chapter 4.2.3.

The strategy consists of four phases:

- **Design phase - Define the review protocol.**
  The design phase will consist of several stages. Defining keywords, search strings, inclusion & exclusion criteria and select databases that will be used for the review. The design phase is an iterative process with a test period (review protocol calibration).

- **Conduct phase - Apply the review protocol versus selected databases.**
  This phase is aimed to gather material that will be used for the analysis by applying the review protocol versus the selected databases. A technique called backward snowballing will be applied to the included material to identify new possible material that may be included.

- **Analyze phase - Analyze all included material.**
  The analysis process will be conducted as a thematic synthesis, form analytical themes that will be used when producing the review.

- **Produce the report.**
  Present the result of phase 2, the findings and conclusions from phase 3.

A flowchart in figure 2 is presented to illustrate the process of the systematic review strategy.

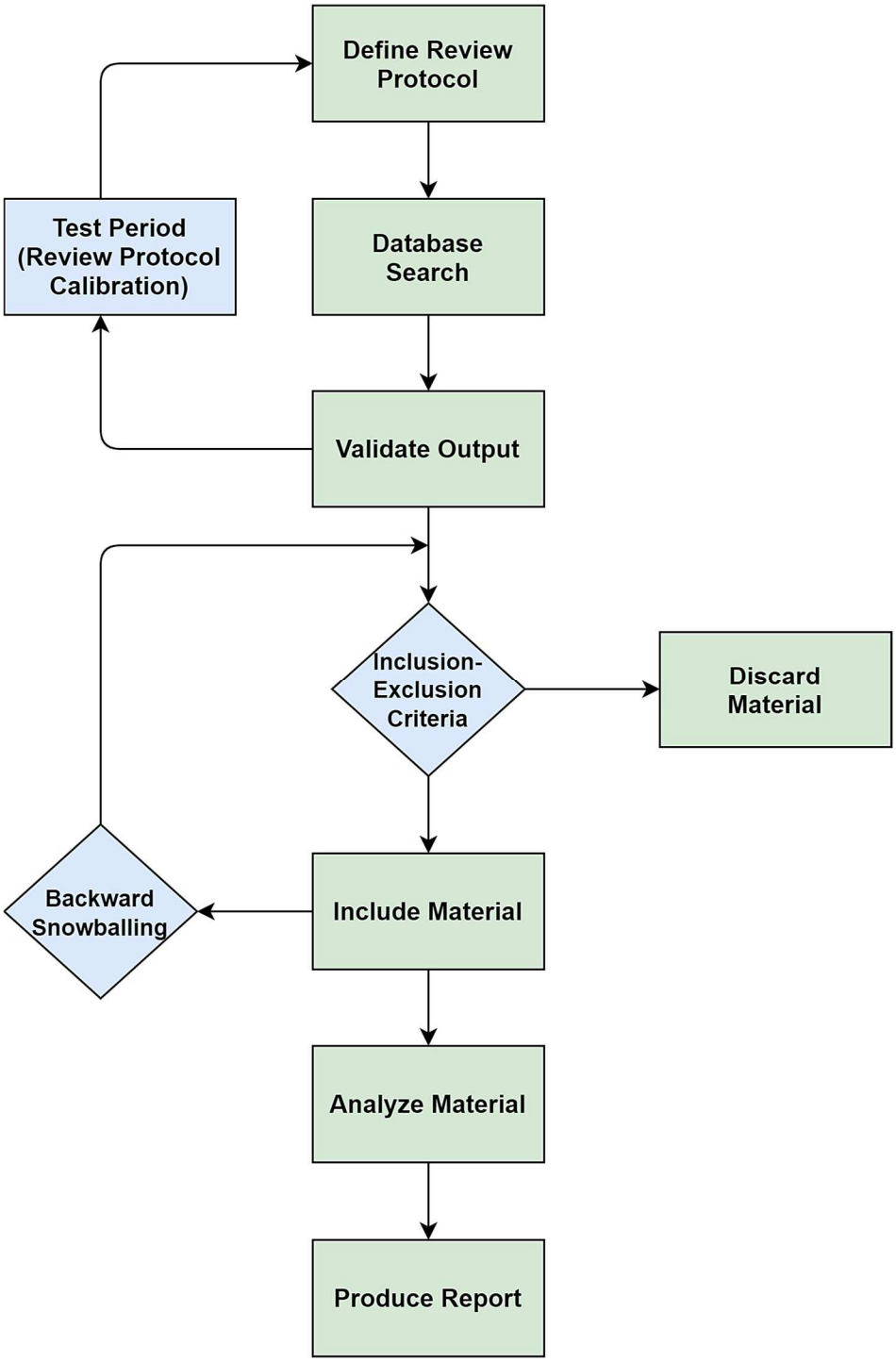

Figure 2: Method flowchart based on Snyder (2019) approach to a systematic literature review (Author's own).

4.2 Review Protocol

In this chapter, the review protocol and its included parameters will be explained.

4.2.1 Search string

Defining search strings is an important part of the literature review process. A poorly defined search string can lead to missed data and further down the process lead to an inconclusive analysis. Kitchenham (2004) says *"[t]he aim of a systematic review is to find as many primary studies to the related question as possible using an unbiased search strategy."* (p. 7).

In order to achieve wanted result when probing a database, a well-defined search string must be applied. This means selected keywords paired with each other to generate relevant output. In order to determine these keywords, an iterative process has been conducted (see figure 2). The process involves identifying keyword associated with a function or method that are used on the research question in hand. Suggested by Kitchenham (2004) to achieve more potential from a keyword when probing them versus a database, synonyms, abbreviations and alternative spellings can be applied. Thus, resulting in more relevant data outputted from the probe. The keywords can be identified in previous material analyzed for the review, and by analyzing technical documentation of email services. The identified keywords and its corresponding search scope are presented in table 1.

Keyword	Search scope
Spam **Phishing** **Spoofing**	Scope of the keywords: undesired emails
DKIM: DomainKeys Identified Mail **SPF**: Sender Policy Framework **DMARC**: Domain-based Message Authentication, Reporting, and Conformance **BIMI**: Brand Indicators for Message Identification **DNSBL**: Domain Name System Blacklist **Greylisting** **Whitelisting** **Blacklisting** **Sender Reputation**	Scope of the keywords: methods and countermeasures towards undesired emails

Table 1: Identified keywords and search scope.

With identified keywords, a more sophisticated search string can be constructed by using Boolean expression. By applying the Boolean expressions “AND”, “OR” or “NOT” to individual keywords can be combined and construct one search string (Booth et al., 2016).

- The operator OR is used to combine keywords with the same concepts together. Purpose is to expand the search.
- The operator AND is used to combine keywords with different concepts together. Purpose is to narrow the search by combining concepts.
- The operator NOT is used to exclude irrelevant concepts. Purpose is to narrow down the search by removing concepts from the search.
- Parentheses are used to nest query terms within other query terms

From identified keywords and by making use of Boolean expressions the following search string is composed:

> ***((phishing OR spam OR spoofing) AND (SPF OR DKIM OR DMARC OR BIMI OR greylisting OR whitelisting OR blacklisting OR sender reputation OR DNSBL))***

4.2.2 Databases

To increase the validity of the findings, more than one database will be probed with defined search string. The following three databases were selected for the search strategy:

- **ACM Digital Library**
- **IEEE *Xplore***
- **SpringerLink**

The choice of these three databases is based firstly on the fact that all three are large databases with computer and cognitive science content, that will meet the needs regarding information and factual of the subject, thus provide the information that will be needed to conduct the research and secondly, existing competence through previous work with these databases, which will meet the requirement to be able to carry out this study within given time frame.

4.2.3 Inclusion and Exclusion Criteria

Snyder (2019) describes the importance of inclusion and exclusion criteria as *"In terms of research quality, deciding on inclusion and exclusion criteria is one of the most important steps when conducting your review."* (p. 337). Kitchenham (2004) says that the criteria are set in place to identify studies that provide evidence to the research question. The author continues by that it is also important to provide reasoning and transparency about the choices made regarding criteria, there should be logical motives. With the references considered, the criteria put in place in this review will be mainly towards finding relevant information in context to the thesis research question. What defines as relevant information? Relevant information in this context is data that can aid in understanding the concepts in question to this thesis and information that can be used to draw conclusions from with the aim to answer the thesis research question. To determine the relevance of the findings from the database probe, a flowchart is presented in figure 3:

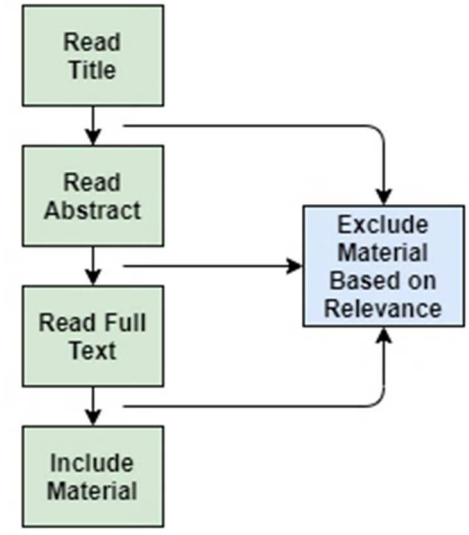

Figure 3: Flowchart of relevance process.

Explanation to the relevance process:

1) Browse the titles of the generated output, based on title name an article may be excluded from the relevance process. Relevance is decided based on if relevant keyword exist in the title.

2) Read the abstract of articles that passed through step one and determine the relevance of the paper. If the paper is determined to not be relevant, it will be excluded. By reading the abstract of the article, it can be determined if this research is relevant to this work.

3) Read the full text of the remaining articles to determine the relevance. If it determined relevant, the article would pass through the process and be included for the review. By reading and understanding the article in its entirely, it is determined whether the article in question is relevant and can be used for this study.

Besides relevance, a few other criteria are applied as well. Table 2 presents all criteria.

Inclusion criteria	I1. Written in English or Swedish
	I2. Published 2016 or later
	I3. Peer-Reviewed
	I4. Relevance
Exclusion criteria	E1. Fails to meet inclusion criteria
	E2. Locked behind payment-wall
	E3. Duplicates

Table 2: Inclusion & exclusion criteria.

I1. The data must be written in a language that the author understands, English or Swedish.

I2. Publication of the research must be earliest 2016. This is to assure only recent studies will be used when evaluating existing and future solutions.

I3. The data must be peer-reviewed. This is to assure certain quality of the findings, but also to respect given time frame, to avoid quality assessment of the data.

I4. Only relevant data will be included for the review. Reasoning and how its assured can be read above in this chapter.

E1. Fails to meet any of the inclusion criteria.

E2. Content locked behind a payment wall will be excluded. Data that can't be accessed can't be used.

E3. Duplicates will be excluded. An article may be published in more than one scientific database, therefore there is a possibility of duplicates are generated from the database searches.

4.3 Backward Snowballing

Once all material has gone through inclusion and exclusion process and the final material has been selected for the analysis process, a method called backward snowballing will be done. Backward snowballing means to look over the reference list of the included material to identify new material (Wohlin, 2014)

Wohlin (2014) describes the process as to look over the reference list on the included material and apply basic exclusion criteria to the identified material, e.g., language, publication year and the type of publication. The backward snowballing will only be conducted on included material from that made it through the selection process. This means that no backward snowballing will be conducted on the material identified using the process in question. This is decided due to given time frame as the snowballing process could recurse for a long period with no defined depth to the process. Based on how a backward snowball method can be conducted by Wohlin, the following strategy will be used in this review:

1. Look over the reference list in included material for the analysis.
2. Apply the methods inclusion and exclusion criteria (see chapter 4.2.3)
3. Because of the emphasis on relevance, the relevance process is its own step, even though it is technically part of the methods inclusion and exclusion criteria. The identified material that passed through step two will go through the study's relevance process (see chapter 4.2.3 for explanation of this process).
4. The material that passed through the process is now included in the review.

4.4 Analysis Process

The analysis is going to be performed by conducting thematic synthesis. Booth et al. (2016) describes thematic synthesis as *"Thematic synthesis aims to provide a consistent analysis of content across included studies. It seeks to identify the range of factors that is significant for understanding a particular phenomenon. It then seeks to organise those factors into the principal (interpretive) or most common (aggregative) themes."* (p. 226). This type of analysis process uses a comparable sort of analysis to unite and integrate findings of qualitative data within a systematic review according to the author.

Booth et al. (2016) mention that there are three stages involved in a thematic synthesis.

- Free line-by-line coding of findings.
- Map the codes created in phase 1 to related codes and form descriptive themes.
- From the descriptive themes develop analytical themes.

With Booth et al. thematic synthesis framework considered, it hereby presents the analytical process will be conducted for this review.

1. **Get acquainted with data:** Read and understand the content and facts in the articles and understand what the author wants to achieve with the article.
2. **Generate line by line codes:** Furthermore, when reading the articles, interesting parts of the articles are marked so that they can later be identified and used in the analysis.
3. **Search, map and formulate descriptive themes:** Selected text is analyzed and mapped into described and logical themes.
4. **Evaluate descriptive themes:** Evaluation of the created themes to see if they have a logical connection and can be seen as a theme. If it cannot be considered as a theme, step 3 of the process is repeated.
5. **Define and give names to analytical themes:** Here, the theme is defined as formal analytical themes that will be included in the analysis.

4.5 Results of Method

This chapter aims to explain how the material was collected, processed and analyzed for the literature review based on the plan defined in the method chapter.

4.5.1 Searching and Collecting the Literature

The selected databases for the review are motivated for in chapter 4.2.2 and the search string is motivated for in chapter 4.2.1. Only one search string was used towards the databases, and no modifications to it were necessary to be applied to all the three selected databases. The used search string is presented below:

((phishing OR spam OR spoofing) AND (SPF OR DKIM OR DMARC OR BIMI OR greylisting OR whitelisting OR blacklisting OR sender reputation OR DNSBL))

No specific filters (for example, language and year) were set on the databases before conducting the search. The generated amount of hits for each given database and the total amount of hits before applying any selection criteria are presented in Figure 4.

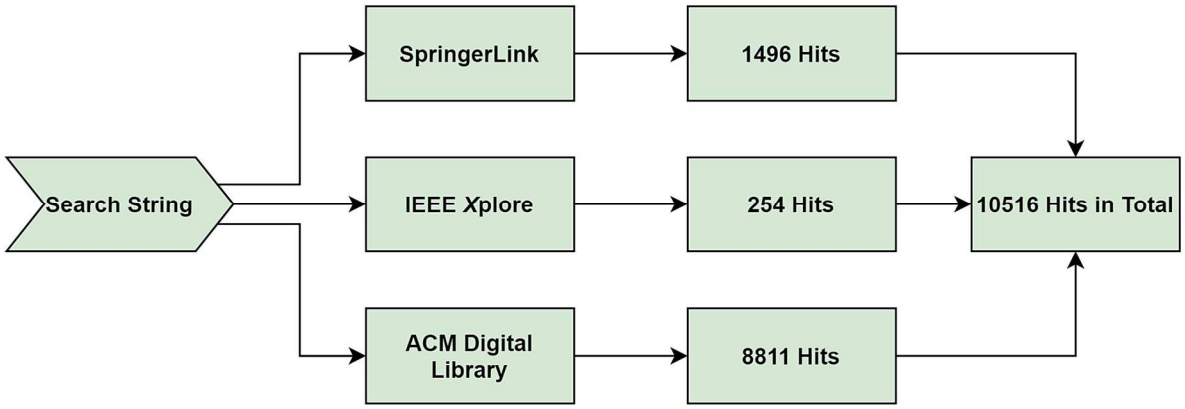

Figure 4: Amount of hits generated from the databases.

The materials after this stage went through the inclusion and exclusion process, motivation for the different criteria can be read in chapter 4.2.3.

The selection process begun after fetching all output from the database searches. Two of the criteria could be applied directly to the gathered materials on the databases with help of filters, and these were language (I1) and year (I2). Remaining criteria needed manual intervention to be applied towards the remaining material. A summary of the selection process and its result is illustrated in Figure 5.

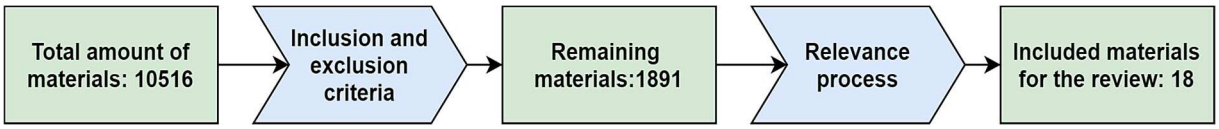

Figure 5: Selection process.

Table 3 displays the included materials distribution from its respective source, gathered from the selection process.

Database	Number of included materials
SpringerLink	**8**
IEEE *Xplore*	**6**
ACM Digital Library	**4**

Table 3: Materials distribution.

The last stage for the collection phase was to conduct the backward snowballing process towards the included materials that were gathered from the selection process. A more detailed explanation of how the process was designed can be read under chapter 4.3, and an illustration of how the process was conducted and its result can be seen in Figure 6.

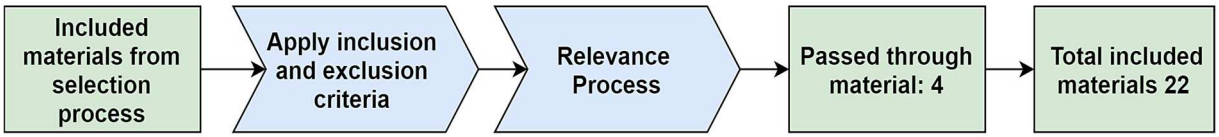

Figure 6: Backward snowballing process.

See **Appendix A** for all the 22 selected materials for this literature review, of which the numbers 19-22 are materials gathered from the backward snowballing process.

4.6 Analyzing the Materials

How the materials were analyzed is described in greater detail in chapter 4.4. When analyzing the materials, a clear goal was always kept in mind and that was to identify a range of factors that are important to be presented. Thus, with the aim to achieve a justifiable answer to the review's research question. These factors emerged from highlighting chunks of data in the materials, and these factors ultimately lead to the forming of analytical themes. Two analytical themes emerged from this process and are presented below in figure 7 with its following subthemes. The figure also illustrates the reviewed literature's distribution in correlation to its theme. The literature's labels can be seen in **Appendix A**.

- **Theme 1: Authentication:** SPF, DKIM, DMARC and evaluation of authentication protocols. This analytical theme is about how a mail server validates the sender of an email by using authentication protocols.

- **Theme 2: Filtering Techniques:** List-based filtering techniques, Content-based filtering techniques and evaluation of filtering techniques. This analytical theme will answer what regulatory rules and solutions are being used to filter undesired emails.

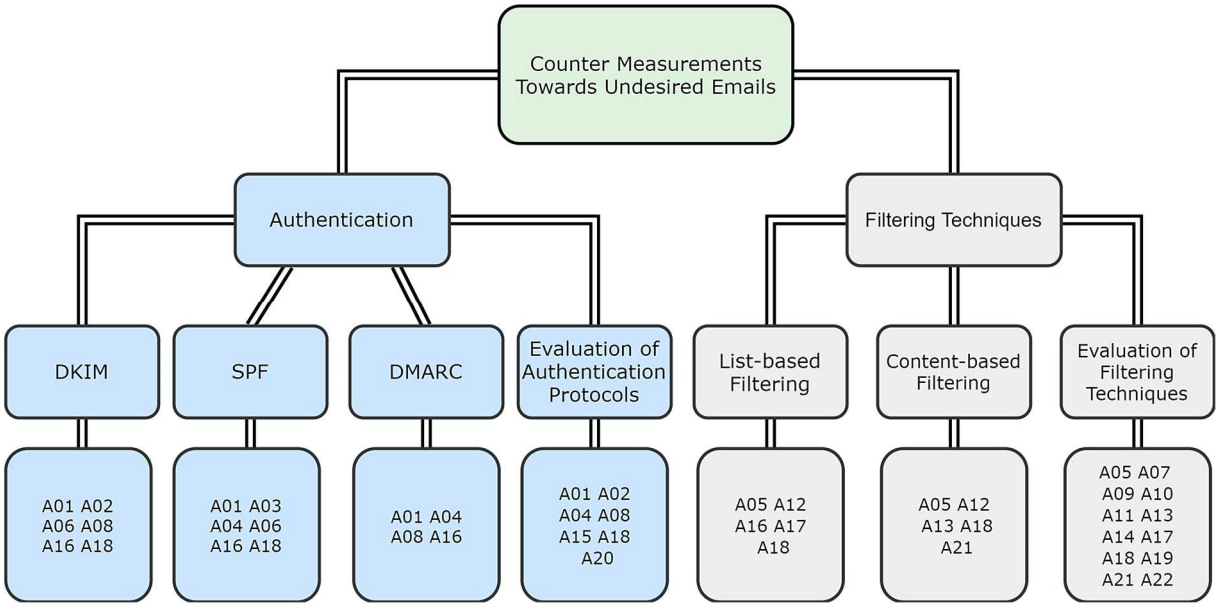

Figure 7: Emerged analytical themes.

A more detailed explanation of the analytical themes will be presented in the next following chapters.

5 Analysis

To give context to the presented themes under the analysis chapter, Dhru (2018), Jakobsson (2016) and Lakshmi (2019) explain the process of an email delivery and what criteria and steps an email must pass to reach a user's inbox. Dhru, Jakobsson and Lakshmi say that there are several places where at which an email may be stopped. First off, Jakobsson reminds us how the mail delivery process is conducted simply; The mail submission agent (MSA) finds the IP address of the mail transfer agent (MTA) by looking up the DNS of the receiver. By using SMTP, the MSA sends the email to the MTA, Jakobsson points out that an email may take several routes through other MTAs to reach a user's inbox.

1. The first spot where an email may be stopped is in the MSA, which can refuse email based on criteria such as bad IP address (blacklisting), and also by rate limit (limiting the number of emails accepted from a sender) (Jakobsson, 2016).

2. Jakobsson says that the second spot where an email may be stopped is at the MTA. The MTA makes use of several authentication techniques (SPF, DKIM and DMARC) to validate the senders of the emails. Lakhsmi describes this stage as; to check the sender's reputation. List-based filtering approaches are also being utilized in the MTA to limit the volume of spam and further validate the sender by using, for example, blacklisting and greylisting (Jakobsson, 2016).

3. The last step in the chain is the anti-spam filter, which can classify spam based on a combination of features including text, link, structure and network analysis of the email (Jakobsson, 2016).

If an email is stopped, for example, because of being flagged as spam in step 1-3, the email will be discarded or sent to the spam box.

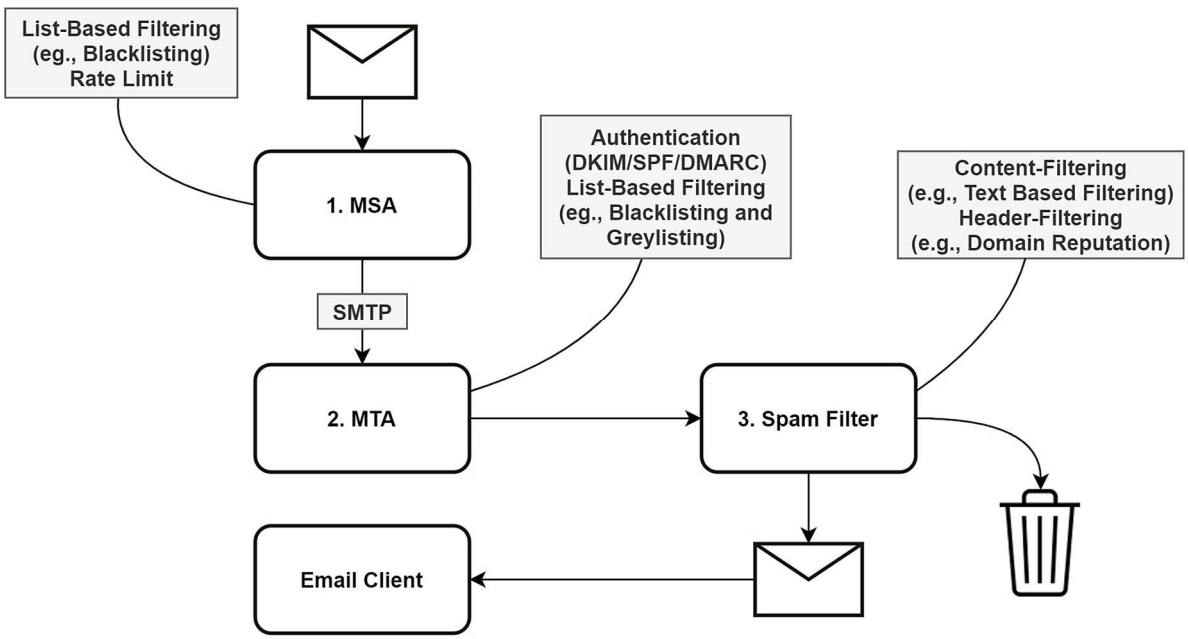

Figure 8: Mail delivery process; based on Dhru (2018), Jakobsson (2016) and Lakshmi's (2019) description (Author's Own).

5.1 Authentication

Simple Mail Transfer Protocol (SMTP) is today's internet standard for email transmission, which was designed in 1982 (Hu et al., 2018). Hu et al. say a key limitation of SMTP is that the protocol has no built-in security to prevent people (attackers) from example impersonating (spoof) a legitimate sender address. SMTP relies on email providers to implement security extensions such as Sender Policy Framework (SPF), DomainKeys Identified Mail (DKIM) and Domain-based Message Authentication, Reporting and Conformance (DMARC) to authenticate senders and how to handle suspicious emails (Hu & Wang, 2018). Hu et al. say that all three protocols (SPF, DKIM and DMARC) have been published or standardized by the internet Engineering Task Force (IETF).

Hu et al., 2018 explain how a spoofing attack is performed. The authors say that attacker can manipulate two key fields to send emails. When establishing a SMTP connection, the attacker can use the **MAIL FROM** command and set the sender's address to anyone they wish to impersonate (Hu et al., 2018). The authors explain that, the **MAIL FROM** address is inserted into the header as the Return-Path The authors continue by saying that attackers can change another field called **FROM** in the email header, this field specifies what address that will be displayed in the email interface, thus when a user receives a spoofed email, they will see the **FROM** address. If a user would reply to a spoofed mail, the email will be sent to the Return-Path set by the **MAIL FROM** (Hu et al., 2018). The authors say that email spoofing is a critical step in phishing in order to gain victim’s trust.

Maroofi et al. (2021) say a careful deployment of these security extensions can help to mitigate the problem of domain spoofing. The authors say that this can be achieved under the preliminary that both the domain owner and the mail transfer agent of the recipient should implement these security extensions and correctly set them up.

5.1.1 SPF

Using SMTP an attacker can easily spoof a sender address, making it appear as to be sent from a legitimate domain (Jakobsson, 2016). The author explains to hinder this, SPF has been developed to validate the origin IP addresses of an email. SPF was proposed in the early 2000, and was first standardized in 2014 (Hu et al., 2018). Konno et al., 2017 say that SPF is a method to validate if the IP address of the sender’s SMTP server is legitimate or not. The authors say that SPF allows a domain to publish their records of allowed addresses to send emails on their behalf. Nanaware et al. (2019) say that these records contain a pool of hosts, and only these hosts may send an email. The authors continue, when a user sends an email the IP address of the sender is compared towards the SPF pool of allowed hosts, and if the user is authorized the email will be sent.

Maroofi et al. (2021) explains how SPF works in a more detailed way, and they say that when a mail delivery over the SMTP protocol occurs, the recipient’s server authenticates the sender’s MTA by issuing **HELLO** or **MAIL FROM** identity. The authors continue to explain that based on the published SPF records by the DNS, the IP address of the sender needs to contain the domain portion in the **MAIL FROM** identity. The authors say that the decision if the email should be rejected or delivered is up to the check_host function, this function collects three arguments on input; (1 the IP address of the sender, (2 the domain name and (3 the **MAIL FROM** or **HELLO** identity, the check_host function returns one of seven results. The authors say that the results presented below are the seven results that can be returned to the check_host function.

1. **None**: This result indicates that there was no valid domain name extracted from the SMTP session, or that no SPF record could be retrieved from the domain name.
2. **Neutral:** this result means that no definite assertion about the sender could be made, whether authorized or not.
3. **Pass:** The sender is authorized to send emails with this identity.
4. **Fail**: The sender is not authorized to send emails with this identity.
5. **Softfail:** The sender is not authorized to send emails with this identity, or there is no policy specified by the owner.
6. **Temperror**: a temporary error occurred while fetching the SPF policy.
7. **Permerror:** occurs when there is a parsing problem in the published SPF.

Maroofi et al. (2021) clarify that the actions done based on results may vary from one SPF policy to another. The authors continue to elaborate that the recommended actions are reject email, deliver email, deliver and mark the email, defer the email or whitelist the sender domain. These actions are up to the domain owners to specify and address to what result it should apply to (Maroofi et al., 2021).

Maroofi et al. (2021) say that a valid SPF version 1 record must begin with the string **v=spf1** followed by SPF mechanisms, qualifiers and modifiers. The authors continue by explaining that mechanisms describe the set of mail servers for a domain and they can be prefixed with one of four qualifiers: **+** (pass), **-** (fail), **_** (SoftFail) and **?** (Neutral). The authors continue by describing some of the most commonly used SPF mechanisms: **ipv4** and **ipv6** (check_host can match specific ipv4 or ipv6 sender address), **a** and **mx-** (tells the check_host function to perform a DNS lookup versus an **A** or **MX** record of a given domain and then compare the returned IP address with the senders IP address), **exists** (it shows a DNS domain name used for a DNS **A** query, if the query returns an **A** record, this mechanism will match), **include** (directs the check_host function to include the SPF rules of another domain in the evaluation) **all** (always match its corresponding qualifier result, for example **v=spf1 mx -all** would mean allow all **MX** servers of the domain to send emails, but prohibit everyone else). The authors say that the final result of the mechanism would be a match, no match or exception. The authors say that qualifiers and mechanisms generate the final input for the check_host function that are evaluating given SPF rule. Modifiers in this context are used to provide additional information about SPF records, and the modifiers available are **redirect** (used to replace current SPF records with another domain) and **exp** (if SMTP rejects a message an explanation can be provided) (Maroofi et al., 2021).

5.1.2 DKIM

The authentication method DKIM was first drafted back in 2004 and first became standardized in 2011 (Hu et al., 2018). Jakobsson (2016) describes DKIM to be a cryptographic method used to validate email integrity and authentication by using public key and digital signature. Its purpose is to evaluate if the email is genuine and that no changes have been made to the email between sender and receiver (Nanaware et al., 2019; Opazo et al., 2018).

Konno et al. (2020) and Hu et al. (2018) go into more detail about how the method operates. The authors say that preliminary for the DKIM operation to work, there needs be a pair of private key and public key stored on the sender side's authoritative DNS server. The authors continue, when the sender first sends an email, the sender domain will generate a DKIM signature from the email header and body by using the private key, and later attach the signature to the email header. When the recipient receives the email, the recipient will inquire

the public key to the authoritative DNS server on the sender's side domain, the public key is obtained from the **d=** tag within the email header (Konno et al., 2020; Hu et al., 2018). The authors say that the recipient compares the hash value from the DKIM signature by using the public key with the hash value that is generated from the received email, if these values are the same, the email passes the DKIM verification and the email is determined to have kept its integrity from source to recipient.

Opazo et al. (2018) touches on the subject what happens if an email fails the DKIM verification. The authors say that DKIM itself does not specify what action to take in a scenario when an email fails the DKIM verification; they say that it is up the email server itself to apply suitable action to address it.

5.1.3 DMARC

DMARC was first drafted in 2011 and standardized in 2015 (Hu et al., 2018). The authors explain that DMARC is not a standalone protocol, it is built on top of the existing protocols and uses the two sender domain authentication methods DKIM and SPF. The authors continue to describe its function; it acts as a failing policy that a domain owner can publish, which specifies what actions should apply when receiving emails fails the DMARC check.

DMARC can provide detailed reports about such as email Header-From domain, Envelope-From domain, DKIM signature domain, sender domain authentication results and effectiveness of the DMARC policy and among others (Konno et al., 2020). The authors explain that this information available makes it possible for the sender domain's administrator to take stronger measures in order to decrease spoofed emails abusing their domain.

Maroofi et al. (2021) say DMARC binds names that are checked by SPF with what is being listed in the **FROM:** field of the email header, thus by means of alignment. The authors continue to elaborate that DMARC checks if the name in the **MAIL FROM** SMTP command and the **FROM**: field of the email head is a match or not. If there is a mismatch, an alignment failure occurs, and the DMARC policy specifies what action to do with the email and where to send reports.

Maroofi et al. (2021) point out a few important tags to keep in mind, as they may be exploited if mis-configured. The authors continue by explaining the tag **aspf** (alignment mode for SPF), this tag specifies if the alignment mode should be strict or relaxed. They elaborate, that relaxed mode is the default behavior and in the relaxed mode any sub domain of the domain can be specified in the **FROM:** field of the email header, thus it will result in a pass. In the strict mode the domain name used in SPF must be exactly the same as the domain used in the **FROM**: field of the email header (Maroofi et al., 2021). The authors continue by explaining the tag **p** (policy), this tag specifies what action to apply if an email fails the alignment check. The authors say that there are three possible values:

- **None:** no specific action applied if email fails the alignment check.
- **Quarantine:** the message is considered suspicious and depending on the recipient's mail server it may be rejected or delivered as spam.
- **Reject:** if email fails the alignment check it should be rejected during the SMTP transaction.

The authors continue by explaining the tag **ruf** (reporting URI for failure), this tag specifies the email address to where authentication failures information is to be reported. The authors

say, *"This tag is important since it is the only bridge between the receivers and the true domain owners to fight spam emails."* (p. 3). The authors also mention the tag **sp** (sub domain policy), if no sub domains are being used to send emails, this tag should be set to **reject** in order to prevent any sub domain email spoofing.

Maroofi et al. (2021) explain further with following DMARC rule scenario: **example.com**: **v=DMARC1; p=none; aspf=r;**. The authors say with these rules, an illegal sender can forge emails on behalf of the domain **example.com** or any existent, or even non-existent sub domain of the **example.com**. The authors explain that the delivery is up to the receiver, since no strict rule is specified in the DMARC rule. By changing the DMARC rule as following: **v=DMARC1**; **p=quarantine**; **sp=reject; aspf=s;** will tell the receiver to label all emails that did not pass the SPF evaluation as spam and will reject all emails from the sub domains of **example.com** at the SMTP level (Maroofi et al., 2021)**.**

5.1.4 Authentication Protocol Evaluation

The adoption rate and usage of the standardized anti-spoofing protocols are low. In January 2018 Hu et al. (2018) conducted an end-to-end measurement experiment testing the adoption rate of anti-spoofing protocols. The authors say the domains targeted for the experiment were among Alexa top 1 million domains, and the result shows that 44.9% have a valid SPF record and 5.1 of the domains in the study have a valid DMARC record. The study also shows that 79% of the domains among Alexa's top 1 million domains are email domains with MX records, 54.3% of these MX domains have a valid SPF record, and 6% of the MX domains have a valid DMARC record. The authors say that their experiment was done by using the same methodology as an experiment testing end-to-end measures in 2015, the result from the study made in 2015 was that only 40% of the domains had valid SPF records, and 1% had valid DMARC records. The author's conclusion by the experiment is that the adoption rate has increased, but only mildly. These results raise their concerns about the effectiveness of the current anti spoofing measurements (Hu et al., 2018).

Hu et al. (2018) interviewed email administrator to better understand the reasons behind the low adoption rate. The authors say that the email administrators interviewed acknowledge the values of adoption of these protocols. The authors, however, say that the most discussed topic was about the technical flaws in the authentication protocols. The authors continued researching the technical flaws by deploying SPF, DKIM and DMARC on a mail server and by running proof-of-concept experiment towards it. One of the identified weaknesses the study found; is the Identifier Alignment. The authors say SPF and DKIM have problems with identifier alignment, which means that the sender email address that the user sees can be a completely different from the address that is used to perform the authentication on. The authors continue by explaining this issue in a more technical depth; SPF's authentication focus on the "Return-Path" and examines if the sender's IP address is listed in the "Return-Path" of the domain's SPF record. The authors say that an attacker can set the "Return-Path" to its own domain and set its SPF record to pass the authentication. The author say that the email interface displays the address in the FROM field for a user, since SPF does not require the two domain names to be the same, thus a spoofed email can pass the SPF verification while displaying impersonated address to the user. The authors continue by saying that DKIM has a similar issue, the domain to sign the email with the DKIM key can differ from the domain in the "Return-Path". DMARC however, resolves the alignment issue of SPF and DKIM by enforcing that alignment of the identifiers must match (Hu et al., 2018).
Another technical flaw that Hu et al. (2018) identified is the mail forwarding problem for SPF. Hu et al. explain the mail forwarding as; means that an email service automatically

forwards emails to another email service. The authors say that during mail forwarding the email metadata, for example, the "Return-Path" remains unchanged, this will lead to SPF will fail after the mail forwarding due the forwarded IP address will not match the original sender's IP address in the SPF record. Another technical flaw Hu et al. identified is the mailing list problem for both SPF and DKIM. Hu et al. say when a message is sent to a mailing list, it will broadcast the message to all its subscribers. The authors describe it as a similar process as mail forwarding, the mail list's IP address will become the sender's IP address, which is a different IP address from the original sender's, and thus result in a SPF failure. The authors continue by saying that mailing lists will cause trouble due to most mailing lists modify the email content (for example add a footer with name of the mailing list) before broadcasting the email to its subscribers, by tampering email content will cause DKIM failure. The authors say that this issue is not something that DMARC can resolve, in case of DMARC+SPF combination; if the "Return-Path" is modified DMARC will fail because of misalignment of identifiers; if the "Return-Path" is not modified, SPF will fail because of IP address mismatch. In the combination of DMARC+DKIM; it will fail due to the mailing list has to modify the email content (Hu et al., 2018).

Jakobsson (2016) points out another technical flaw that can be found in DKIM. The author says this technical flaw can be abused by an attack called chosen message replay attack. Jakobsson explains the process of a chosen replay attack against DKIM as; the spammer creates two separate email accounts in two different domains. The author continues, then the spammer composes an unsigned spam email using one of the email accounts, and then sends it to its other email account on the other domain. Jakobsson explains that the MTA of the sender domain will then sign the content of the email, along with the sender's email by using the private key of the sender domain. Jakobsson says that the destination domain will receive and verify the signature of the email by using the public key of the sender domain. The author says that the signed and verified version of the email is now delivered to the receiving domain, since the receiver 's email address is not included in the DKIM signature. The receiver address can now be modified to any other address email, the spam email can now be sent to other recipients while still having a valid DKIM signature (Jakobsson, 2016).

While interviewing email administrators, Hu et al. (2018) found two more aspects that may explain the low adoption rate. The interviewees of the study say that there has not been a global consensus that the authentication protocols SPF, DKIM or DMARC should be the one solution to solve spoofing. The other aspect pointed out by the interviewees of the study, is that protocol adopters does not directly benefit from publishing their SPF, DKIM or DMARC records in their DNS, these DNS records mainly help other mail services to verify incoming emails and protect their customers of the mail service. The authors say domains that publish their records receive the benefit of a better reputation, which to some domains might seem like a relatively vague benefit, especially for domains that do not even host an email service. Jakobsson (2017) says the following about the authentication protocols: SPF, DKIM and DMARC, *"It does not address abuse using look-alike domains, display name attacks or corrupted accounts, nor does it protect an organization against malicious incoming email as much as it protects it against abuse of its brand."* (p. 312).

Opazo et al. (2018) say that the adoption issues leave enormous gaps for spoofers to exploit. The authors say reasons for this are the reluctance of business to set security standard high enough to stop a majority of the incoming scam emails, is because fear of having legitimate emails blocked as false positives. Maroofi et al. (2021) conducted an end-to-end spoofing measurement experiment targeting well-known brands by sending emails from a non-existent

sub domain. The authors targeted Gmail, Yahoo, Outlook, Yandex and Lapose email services in this experiment. The experiment shows that Outlook labeled 80% of the emails sent as spam, while Yahoo rejected 70% of the emails and Gmail, Yandex and Lapose delivered almost all the emails. The authors did the same experiment, but this time sending the emails from an existing sub domain without proper SPF configuration or a restrictive DMARC rule. They say that Outlook performed best by labeling 60% of the emails as spam and Yandex worst by delivering 97.5% of all sent emails to the inbox. The experiment concludes that attackers can successfully spoof all the tested email services by sending emails from non-existing sub domains, and also if domains do not have strict reject DMARC policy.

In an experiment testing email availability versus security conducted by Hu & Wang (2018), the study shows that email providers may choose to deliver a forged email, even though the email fails the authentication. The author's say that the vast majority of email services can be successfully penetrated, and 34 of 35 tested email services allowed at least 1 out of 1500 forged email to reach a user's inbox. The experiment shows that many of these email services have strict configured SPF/DKIM/DMARC, which suggests that even when email providers detect forgery, they still may deliver it. The authors continue to share their reasons why this might be, "*If an email provider blocks all the unverified emails, users are likely to lose their emails (e.g., from domains that did not publish an SPF, DKIM or DMARC record). Losing legitimate emails is unacceptable for email services which will easily drive users away."* (p. 7).

However, it is also suggested by an experiment conducted by Maroofi et al. (2021) that email get false evaluations because of misconfigurations. Maroofi et al. experiment was conducted by scanning TOP500 domains of Alexa's domain list for their SPF and DMARC rules. The experiment findings show that wrong SPF rules lead the check_host function to not be able to evaluate the SPF record of a domain name. The authors say this is due to a syntax error and the result will be either Permerror or Temperror, thus it will lead to a legitimate email will probably end up in the spam box. The authors say that *"A misconfigured SPF or DMARC (either syntactically or semantically) rule is as dangerous as the absence of the rules since the output of the evaluation does not lead to a correct decision."* (p. 4). In conclusion of the experiment, the result says that a large portion of domains with SPF and DMARC records; do not correctly configure their rules, which enables attackers to successfully deliver forged emails to a user's inbox.

Misconfiguration of rules may also increase the false positive rate, Konno et al. (2020) say that DMARC cannot verify legitimate emails properly in some cases. The authors designed a method by analyzing DMARC report data, this method they say can detect false positive within the sender domain authentication by comparing IP addresses listed on Spamhaus blocklist (the most famous blacklist) with legitimate IP addresses that are not on the list. The authors test their method by using legitimate IP addresses to send emails. The authors evaluated the emails that failed both SPF and DKIM authentication; it shows when the result from check_host returned neutral, none, unknown, NULL, fail, softfail, permerror or temperror, both DKIM and SPF could not validate the email correctly and caused a false positive. The authors conclude that 2.8% to 11.1% of legitimate emails were false positives in the combination of SPF and DKIM authentication. Konno et al. say the variation of DMARC authentication result was either a pass or fail. The authors explain that therefore, an email sent from a legitimate IP address marked as fail' by the DMARC authentication is a false positive. The authors conclude that 36.9% to 62.7% deliveries from legitimate IP addresses were false positives in the DMARC authentication.

The current state of the anti-spoofing solutions puts the users in a vulnerable position, Hu et al. (2018) say that considering the flaws of existing anti-spoofing protocols, emails that pass the SPF/DKIM verifications may still be spoofed, similarly emails that fails SPF/DKIM verifications are not necessarily malicious. Since the adoption rate low of anti-spoofing protocols, email services face key challenges to reliably authenticate all the incoming emails, thus the limited server-side protection is likely to put users in a vulnerable position (Hu & Wang, 2018). Hu et al. say that due to the low adoption rate and the relaxed state of the protocols configuration it is likely that email services will have to deliver certain unverified emails to a user.

Hu et al. (2018) suggest incentivizing the adoption of anti-spoofing protocols to build a critical mass. Hu et al. say that currently there is a lack of strong consensus to adopt and deploy anti-spoofing protocols. Anti-spoofing protocols will provide their key benefits first, once enough domains publish their SPF, DKIM and DMARC records (Hu et al., 2018). The authors say to establish the critical mass that is needed for the anti-spoofing protocols, external incentive mechanisms are needed. The author points at the promotion of "HTTPS" as an example, it displays a trusted icon with valid certificates, similar incentive could be applied in this context, for trusted domains. The authors also say that certain sensitive domains should be enforced to publish their records, for example banks or government agencies, thus to prevent being impersonated. Hu et al. also suggest all email providers to act *"as good internet citizens"* (p. 100) by publishing their authentication records.

Hu et al. (2018) continue by saying it is also necessary in today's situation to educate their users on how to identify a spoofed email. The authors urge email providers to act more responsibly and provide authentication results for email users. The email providers must also become better to proactively warn users about emails that they did not manage to verify, Gmail and Outlook are already moving towards this direction (Hu et al., 2018). The current standardized anti-spoofing solutions perceived usefulness is low, Hu & Wang (2018) say that even though they can't authenticate all incoming emails correctly, these protocols allow and help email services to make more informed decisions. Hu et al. (2018) suggest improving the perceived usefulness of the authentication protocols to address the low adoption rate of anti-spoofing protocols. The author explains that the security and usability issues present in SPF, DKIM and DMARC impact their perceived usefulness negatively. In order to improve perceived usefulness of the protocols, the authors suggest addressing the present security and usability issues should be the of priority.

Hu et al. (2018) say that IETF group is working on developing a new protocol called Authentication Received Chain (ARC), this protocol is said to address the email forwarding problem and the mailing list problem, which was two of the highlighted flaws in the protocols by their study. ARC is an underdevelopment protocol that works on top of SPF, DKIM and DMARC (Hu & Wang, 2018). The authors say ARC proposes to preserve the email authentication results via different sending scenarios, thus aim to solve the forwarding problem and mail listing problem. The authors also mention a new protocol Brand Indicators for Message Identification (BIMI), the protocol is built on DMARC. When confirming the authenticity of a sender, the email client can display a BIMI logo as a security indicator of the sender's brand, emails showing the BIMI logo is by this process an indication that they are verified, but emails without the BIMI logo are not necessarily harmful. Hu et al., and Hu & Wang raise their concerns that BIMI and ARC are likely to face the same adoption challenges as previous anti-spoofing protocols. More research is needed on how to ease the deployment process, thus to avoid disruptions to the existing email operations (Hu & Wang, 2018).

5.2 Filtering Techniques

This chapter will present the filtering techniques: list-based filtering methods and content-based filtering methods. It will describe how these methods work and an evaluation of these methods will be presented.

5.2.1 List-Based Filtering

Karimovich et al. (2020) describe list-based filter techniques simply as allowing or denying emails sent from a specific user. Mail filtering systems can be configured via private lists or public lists, such as Spamhaus to constantly update themselves to block incoming emails from known bad addresses (Pompon, 2016). Karimovich et al. say List-based filtering is a common countermeasure used towards preventing undesired emails and there are a various of different list-based filters.

- **Blacklist filter** is one of the most used techniques to filter out undesired emails (Karimovich et al., 2020). The authors continue by explaining how blacklist filters works; it works by keeping records of email addresses and Internet Protocol (IP) addresses that have been associated with sending spam, spoofing or other malicious activities. Jakobsson (2016) says that blacklists are formed by several entities using different type of approaches, such as for example active probing, manual entries (users reporting spam) and passive monitoring (aims to detect connecting spam bots by deploying and monitor fake MSA & MTA acting as real ones). The author continues by saying that the email components MSA and MTA, or content-based filtering software's, for example SpamAssasin consults with blacklists to evaluate, score and classify emails. Karimovich et al. says that when an incoming email arrives, the anti-spam filter will check if the email address or IP address of the email exists in the blacklist of addresses, and if it does, its being considered as spam and thus it will be rejected before reaching the user's inbox. The authors say blacklisting is a suitable technique to filter away undesired emails under the circumstance that spam-senders addresses are fixed and known. Jakobsson says that undesired messages may get by unnoticed by the blacklist approach. The author continues that attacks or spam sent from IP or email addresses that are not present in the blacklist will remain undetected by the filter.

- **Whitelist filter** is a technique where email addresses are saved to a list, it works as an allowance list, hence the term 'white-list' (Karimovich et al., 2020). The authors describe whitelist filtering to operate exactly as the opposite of blacklist filtering, meaning that senders on the list will not get their emails rejected. How this filter technique works towards unknown senders (email addresses that are not on the whitelist), is by checking the senders email address towards a database, and if the address has no history associated with for example spam, it will forward the email to the recipient's inbox (Karimovich et al., 2020). The authors continue by saying that this process helps to reduce false-positive incidents. Even with the prevention of false-positive scenarios, this filter technique has issues with unknown genuine emails that are being declared as spam emails (Karimovich et al., 2020).

- **Greylist filter** is a technique that is based on the assumption that spammers only attempt to send a batch of emails once (Karimovich et al., 2020). Konno et al. (2020) explains the process of greylist filtering as; *"checks the retry function for establishing an SMTP session, and inspects the retrying function for establishing a TCP session*

between the sending host and the receiving host." (p. 39). Karimovich et al. simplifies the process and describe it as; the receiving mail server will reject the first message from an unknown sender and will reply with a failure message to the originated mail server of the sender, if the mail server sends the email a second time, it will be declared as legitimate and it may reach the recipient's inbox. The authors continue saying that by sending the email a second time after receiving a failure message is a commonly something a legitimate mail server would do. The filter will add the email and IP addresses of the sender to a list of allowed senders at the stage when the email gets declared as legitimate (Karimovich et al., 2020). The authors say that it is an effective technique under the assumption that a legitimate sender always sends two times, if not; legitimate emails may risk getting lost.

- **Real-time blackhole list filter** (RBL) or also known as **Domain Name System Blacklist** (DNSBL) is a dynamic list of IP address owners associated with spam (Karimovich et al., 2020). The authors say that DNSBL works almost identical to a traditional blacklist but requires less maintenance since you can blacklist larger volumes with fewer entries. Jakobsson (2016) says that IP addresses or subnets are in the form of an inverse A record, [Inverse IP].blacklist_name. The author continues to explain that by having an inverse A record it will allow for hierarchical and network level blacklisting. Jakobsson explains what happens when an email is being processed as following.

  1. Email agents checks whether sender IP address exists in the blacklist by using DNS lookup. The inverse of the sender's IP address is computed, for example, 192.168.10.11 becomes 11.10.168.192.
  2. 11.10.168.192.blacklist_name is queried from the DNS server.
  3. If an IP address is returned, this indicates that the sender is on the blacklist, or it will return "NXDOMAIN" (domain doesn't exist), and thus the sender is not on the blacklist.

- **URL Based filter** checks and tests if incoming URLs are legitimate or not, like all filters presented in this chapter, this filter technique is no exception and keeps a repository of URLs that it evaluates incoming URLs towards to determine its legitimacy (Karimovich et al., 2020).

5.2.2 Content-Filtering

Gangavarapu et al. (2020) say content-based and behavior-based filtering approaches aim at analyzing the email content and structure in order to create automatic classification rules by using machine- and deep-learning methods. Kaimovich et al. (2020) say machine learning algorithms plays a central role in the detection of spam emails. According to Gangavarapu et al. content-based and behavior-based filters analyze the email contents tokens (words), their distribution, their occurrences and co-occurrences, analyzing of scripts and URLs, in the context of emails and then use its learned knowledge to generate and apply rules to facilitate automatic filtering of incoming emails. Konno et al. (2020) say that content-filtering is an effective and the most commonly used technique in the fight against spam emails. The authors say that this type of filtering has a high calculation cost, it is mostly used after incoming emails have been inspected and filtered out elsewhere.

Karimovich et al. (2020) say it is one of the most famous machine learning algorithms working on the principles of the Bayes theorem. The authors continue by explaining the meaning of Bayes' theorem; it calculates the posterior probability, and its technique is widely known and used for the purpose of classifying emails for spam and non-spam. Karimovich et al. say that Bayesian filtering technique can train themselves to identify new patterns of spam but can also be adapted by the user to adjust the user's specific requests and parameters for identifying spam. The authors say when an email arrives, *"It is firstly tokenized into a set of features (tokens). Every feature is assigned an estimated probability that indicates its spaminess. The Naive Bayesian classifier combines the probabilities of every feature and estimates the probability of the message being spam."* (p. 3).

Perhaps a more feasible explanation is presented by Jakobsson (2016). The author explains how Naive-Bayes algorithm computes the odds of the email being a spam as; it multiplies together the individual odds of each word being a spam versus being non-spam and multiplies that times the overall (prior) odds. Jakobsson says that an email can be presented as a bag of words, where each word has the probability of appearing in spam emails and appearing in non-spam emails.

$$S = \frac{p(spam|E)}{p(ham|E)} = \frac{p(w_1|spam) * p(w_2|spam) * \ldots * p(w_n|spam) * p(spam)}{p(w_1|ham) * p(w_2|ham) * \ldots * p(w_n|ham) * p(ham)}$$

Figure 9: Naïve Bayes Algorithm (Jakobsson, 2016).

Jakobsson (2016) describes the formula above (Figure 9) as the simplest version of the Naïve Bayes formula to compute the score (S) of an email. Jakobsson explains that ham in this context refers to emails that are neither spam nor scam. Jakobsson continues by explaining the formula presented in figure 1 as; *p(ham)* and *p(spam)* are the fractions of ham and spam emails, respectively, in a dataset. The author says if S, the odds of the email E being a spam is greater than 1, the email E is most likely a spam email. Jakobsson explains that many businesses rely on the delivery of emails, so by mis-classifying a ham email as a spam email comes with a high cost, thus the spam score threshold is usually greater than 1 in order to optimize the cost benefit of spam filtering.

Dada el al. (2019) states that nearly all state-of-the-art spam filters utilize the Naïve Bayesian method. According to the authors are Support Vector Machine (SVM), K-Nearest Neighbor (k-NN) and Neural Networks also some machine learning approaches that are being used in some of today's content-filter based anti-spam filters.

- **Support Vector Machine**
  Dada et al. (2019) describe it as a powerful and efficient state-of-the-art classification technique to solve email spamming problems. Karimovich et al. (2020) say the concepts behind SVM are Statistical Learning Theory and Structural Minimization Principle. Karimovich et al. explain that SVM has shown to very effective in the context of text categorization because it can handle high-dimensional data by making use of kernels. Kaimovich et al. explain that the basic idea behind SVM for pattern classification is to find the maximum amount of margin between the positive and the negative samples. Karimovich et al. explain the ideology behind SVM as; *"According to the idea, the spam filtering can be viewed as the simple possible SVM application – classification of linearly separable classes; that is, a new e-mail either belongs or does not to the spam category."* (p. 3).

- **K-Nearest Neighbor**
  Karimovich et al. (2020) describe K-Nearest Neighbor (k-NN) as a basic instance-based method algorithm. The authors say that it is a very simple method to classify documents and has been proven to have a very good performance for text categorization tasks. Karimovich et al. explain the procedure of k-NN method employed to email classification as; *"Given a new e-mail, the distance between the mail and all samples in the training set is calculated. The distance used in practically all nearest-neighbor classifiers is the Euclidean distance. With the distance calculated, the samples are ranked according to the distances. Then the k samples which are nearest to the new e-mail are used in assigning a classification to the case."* (p. 3).

- **Neural Networks**
  Dada et al. (2019) describe neural networks as a group of simple processing units, which are interconnected and communicate with each other. Dada et al. say that each of the units accepts input from its neighboring units and external sources and calculates the output that is transmitted to its neighbors. The authors say that neural networks are a powerful algorithm to solve any machine-learning problem that requires classification.

5.2.2.1 Case Base Spam Filtering Method

According to Dada el al. (2019), case base or also known as sample base filtering, is a common spam filtering method. According to Gangavarapu et al. (2020) it works by extracting spam, non-spam and phishing emails from every user's email through an email collection model. Subsequently, Gangavarapu et al. say pre-processing of the raw email data into a machine-processable, via extraction and grouping of the email data. Dada et al. say lastly that the preprocessed data is then mapped into categories, and a machine learning algorithm is employed to train the existing email data. The trained models emerging from the machine learning algorithm are then tested on the incoming emails to categorize them into, for example, spam, non-spam and phishing (Gangavarapu et al., 2020)

5.2.2.2 Heuristic or Rule Based Spam Filtering Technique

Dada et al. (2019) and Gangavarapu et al. (2020) states that this method uses already created rules or heuristics to assess patterns, most commonly it is used with regular expressions against incoming emails. Gangavarapu et al. explain that the score of an incoming email is reliant on the number of patterns matches, the more matches the higher the email is scored. Gangavarapu et al. say that an email score is reduced when it does not correspond to the presets of the regular expressions. Dada et al. say that when an email surpasses a predetermined threshold, the email is filtered as spam; else it is treated as a valid non-spam email. Both Dada et al. and Gangavarapu et al. say this type of filtering technique requires updating regularly to cope efficiently with the constantly changing nature of spam emails. A concrete example of a rule-based spam filter is SpamAssassin (Dada et al., 2019).

5.2.2.3 Previous Likeness Based Spam Filtering Technique

Dada et al. (2019) say that this method makes use of either memory-based, or instance-based machine learning methods in order to classify incoming emails based on their similarity to stored examples, for example, training emails. The authors describe the methods process as; *"The attributes of the email are used to create a multi-dimensional space vector, which is*

used to plot new instances as points. The new instances are afterward allocated to the most popular class of its K-closest training instances" (p. 2). According to Dada et al. this method uses k-NN for filtering of spam emails.

5.2.2.4 Adaptive Spam Filtering Technique

According to Dada et al. (2019) this method detects and filters spam by organizing them into different classes. It divides the email content into various groups, each group has an emblematic text, a comparison is made for each of the incoming email and each group, and a percentage evaluation of its similarity is produced, to decide which probable group the incoming email belongs to.

5.2.3 Filtering Technique Evaluation

Karimovich et al. (2020) discuss one perk of blacklist and whitelist filters, they can both classify emails without needing to read the content, hence why these techniques are fast at classifying incoming emails. Gupta et al. (2017) say that blacklisting and whitelisting approaches have the benefit of having low false positive rates, but also point out that they are very inefficient for the detection of "zero hour" phishing attacks. According to Gupta et al. these approaches are only able to detect about 20% of such attacks. The authors also explain that preventive list-based approaches require a lot of network capacity to operate, thus lowers the performance of the network. Karimovich et al. say that spammers have many tools in their arsenal to avoid preventive list-based methods. The author lists a few of these methods; bad actors regularly switch URL links, IP addresses, email addresses and jumping from one DNS to another, as common evasion techniques. The authors say that these, among other factors, cause it to be challenging to trace and detect spam with preventive list-based techniques. Pompon (2016) says that list-based techniques are not a perfect solution, but it is a good way to eliminate known sources of bad addresses, thus reduce some of the undesired emails from reaching a user's inbox.

Chanti & Chithralekha (2020) state that content-based filter approaches are better for detecting phishing compared to non-content-based methods. The authors continue by explaining that non-content-based methods have hard to detect new phishing attacks due to delay in their updates. Chanti & Chithralekha say that content-based approaches such as rule-based and machine learning are good at detecting, but these approaches may have high false positive rates. Bajaj (2017) points out two points that are important to keep in mind when applying filtering solutions on a mail server; first off, the filter is being applied to all incoming emails on behalf of all email users. Secondly, the author says that spam to one user might not necessarily be spam to other users. Therefor Bajaj says if the level of filtering at the mail server is too stringent, it would lead to a high rate of false positives, and thus valuable information may risk getting lost. The author says on the other hand; if the filtering at the mail server is too relaxed, it would lead to high numbers of false negatives, where spam or phishing emails may reach the user, it is a dilemma.

Trivedi (2016) put two popular machine learning methods to the test: evaluating their false positive rate and accuracy. The two methods used in the experiment are Naïve Bayes and SVM. The author used a dataset of 6000 email files with 50% of them being classified as spam. The dataset was obtained from Enron email corpus (a database of generated emails) and used towards the two methods. The experiment result showed that Naïve Bayes method scored 92.8% in accuracy (classifying emails correctly spam vs. non-spam) and a false positive rate of 6.9%. The SVM method scored 93.3% in accuracy and 6.5% in false positive

rate in the experiment. Shajideen & Bindu (2018) conducted a similar experiment by using a dataset containing 3762 spam emails and 5172 non-spam emails created by Enron email corpus versus the Naïve Bayes method and the SVM method. Shajideen & Bindu experiment shows that Naïve Bayes scored 92.8% in accuracy and 7.85% in false positive rate. The SVM method scored 94.06 in accuracy and 6,79% in false positive rate in the experiment. Both Trivedi's and Shajideen & Bindu's experiment concludes that SVM scored both higher in identifying spam emails while having a lower false positive rate compared to the Naïve Bayes method.

Sokolov et al. (2020) however, states the following about machine learning algorithms; *"The accuracy of these algorithms is dependent on the data used in training them, as they operate under the assumption that the training data comes from the same distribution as the test data. In practice, this is not always the case."* (p. 1).

Sokolov et al. (2020) conducted an experiment in similar fashion as Trivedi (2016) and Shajideen & Bindu (2018) but with another purpose in mind, Sokolov et al.'s experiment was conducted to evaluate how machine learning anti-spam filters reacts to an evasion technique called visual spoofing. The authors say that visual spoofing can be performed by replacing Latin letters with other letters or tokens that look alike. The experiment consists of a dataset from Enron email corpus with 1500 spam emails and 3672 legitimate emails. The dataset was first tested without changing any of the letters versus the machine learning methods Naïve baysen, SVM and others as a control experiment. In the control experiment, the Naïve Baysen method scored 98.07% in accuracy and SVM scored 96.14%. In the second experiment the spam emails of the dataset were modified by changing the Latin letters: 'a', 'e', 'k', 'o', 'p', 'c' and 'y' with the corresponding letters of the Cyrillic alphabet (they appear almost identical to one another). As a result of this experiment, the Naïve Bayesen method now scored 49.08% in accuracy and SVM scored 53.62%. The authors tested this method with Microsoft Business Mail. The first mail the authors sent contained a lot of common keywords associated with spam, this mail got flagged as spam. The authors used the same mail that got flagged as spam and replaced some of the characters to their visually equivalent characters from the Cyrillic alphabet, this time the mail by passing the anti-spam filter and was delivered to the inbox. The authors conclude the following from their experiments; *"Our experiments indicate that using a classifier trained on data using Latin alphabet, to classify a message with a combination of Latin and Cyrillic letters leads to much lower classification accuracy compared to the same classifier used with a message with Latin characters only."* (p. 4).

Jakobsson (2016) says that text-based filtering techniques are very content-sensitive, and the choice of words has a significant impact on their performance. The author explains that spammers take advantage of an evasion technique called polymorphic scam. The idea with this method is to change words associated with spam with synonyms, to say the same but using different words (Jakobsson, 2016). The author says that experiments targeting SpamAssasin, DSPAM and Gmail using this evasion technique improved the spam penetration rate by 20%. Bajaj (2017) says that spammers keep innovating new ways to deceive filter-based solutions. According to the authors has the content of emails evolved to more than just word, such as links, numeric digits, special characters. Most of these evasion techniques today are non-textual, thus textual based filtering can't detect these anti-spam evasion techniques consistently. Gupta et al. (2017) explain that the phishing threat is increasing and when researchers come up with an idea to control this problem, the attackers change their attack strategy by exploiting vulnerabilities found in the current solution. Gupta et al. say *"Therefore, we can say that it is a very tight race between phishers and researchers."* (p. 3650).

Ferrera (2019) claims that the continuous battle between researchers and bad actors is one of the contributing factors that most of the state-of-the-art research regarding spam detection lies behind closed curtains. The author also says that large companies with email-related services, such as Google (Gmail) and Microsoft (Outlook & Hotmail) have made a lot of investments developing machine learning methods to automatically filter away undesired emails in the platforms they operate in; companies are thus motivated to use patented closed-source solutions in order to maintain their competitive advantage. Ferrera says that the spam detections systems deployed by these large service providers have reached nearly perfect detection accuracy. Dada et al. (2020) and Gangavarapu et al. (2020) states that the machine learning model deployed by Google has reached about 99.9% in detection accuracy, this implicit that one in a thousand undesired emails gets by their filtering.

6 Discussion

This chapter will be used to discuss the results, selection, review, and analysis process to ensure the results of the completed study. It will also discuss ethical and societal impact.

6.1 Results of the Study

The study's result shows that there are two primary groups of methods that are being used in today's email services as countermeasures towards undesired emails, namely *authentication* and *filtering techniques*.

As countermeasures towards the vulnerabilities found in SMTP that have introduced email spoofing, three standardized protocols have been proposed; SPF (validate the sender), DKIM (controls the integrity of the email) and DMARC (failing policy). Hu et al. (2018) have made it evident that there is an ongoing issue with insufficient adoption and usage rate of SPF and DMARC. Hu et al. (2018) and Jakobsson (2016) have also acknowledged technical flaws within the protocols that can be abused or lead to common mail services to fail. Maroofi et al. (2021) and Konno et al. (2020) show that misconfiguration of these protocols leads to emails being wrongly classified, and thus causing a high false-positive rate, but still allows attackers to successfully deliver forged emails. Maroofi et al. (2021) and Hu & Wang (2018) validate this by showing it is possible to successfully deliver forged emails to nearly all available email services. Based on the results regarding the authentication mechanisms that are in use today, it is clear that unwanted shortcomings remain and there are gaps to be exploited, which leaves little to no protection against spoofing of domains.

The fundamental of list-based filtering techniques is to allow or deny an email based on the reputation of the sender or domain. List-based filtering technique, namely greylisting can also challenge response a sender in order to validate its legitimacy. They are also being utilized by content-based filtering solutions to score an email. Karimovich et al. (2020) pointed out several evasion techniques that render list-based filtering techniques to have a hard time keeping up with today's spam. As a consequence, it renders them to not be a perfect solution, but they do however help to mitigate already known bad sources and still today a common countermeasure that is being applied in email systems.

Content-based filtering uses predetermined classification rules or automatic classification rules developed from machine learning to classify emails into spam or non-spam. Machine learning techniques have come very far in the detection accuracy of spam emails. Dada et al. (2020) and Gangavarapu et al. (2020) states that the machine learning model deployed by Google has reached about 99.9% in detection accuracy. Jakobsson (2016) and Sokolov et al. (2020) show that it is possible to increase the penetration rate of spam by using evasion techniques such as visual spoofing and polymorphic scam. The detection accuracy of content-filtering solution has reached nearly perfection, it is however important to point out that the detection accuracy of deployed content-filtering solution works under the assumption that most undesired emails' content are similar, thus current and possible future evasion techniques will lead to wrong classification of emails.

A limitation to this study's result is that most state-of-the-art research in filtering techniques, namely machine learning classifications, is not being publicly documented. The literature study can only account for publicly available information, it provides only an incomplete overview of filtering techniques that are being utilized in some of today's email services. The results of the study can, however, identify what types of machine learning approaches that are

commonly used; Naïve Bayes Theorem, Support Vector Machine, K-Nearest Neighbor and Neural Networks.

The most noticeable results identified in the literature are:

- The low adoption and usage rate of standardized authentication protocols (Hu et al, 2018). The expected benefits of these protocols can only first be achieved when a certain critical mass is reached, therefore the low usage rate as of today renders these solutions to not function as intended and lead to major gaps for domains to be a victim of being spoofed (Maroofi et al., 2021; Hu & Wang, 2018).
- Technical insufficiency of authentication protocols. Even with a careful deployment, the domain is still at risk of being spoofed (Hu et al., 2018; Jakobsson, 2016).
- Too stringent regulation may lead to desired emails being discarded, as too been expressed in the identified literature by Bajaj (2017), there is a fine line between allowing a desired email and rejecting it in today's email system. Considering the fact that losing a desired email cost more than delivering an undesired email to a user, today email services are more inclined to have relaxed regulations, which causes some undesired emails to reach the users.

Based on the investigated literature, this study summarizes that it is evident that the issue with undesired emails has been an ongoing issue over the past decades and is today still a present issue. The same ideas were used over ten years ago to counter undesired emails: authentication methods and filtering techniques. Despite the known deficits in the authentication methods, no alternative method has been widely adopted. This study presented an overview of how email systems today work together to hinder undesired emails from reaching a user and also highlighted what today's issues are in this context. Past studies presented relevant but separate contexed research on the subject, this study is a collective research on the subject.

6.2 Validity of the Results

The number of selected articles is low; therefore, the result should not be considered as a definite answer. The result merely reflects what the included literature says. However, many of the selected materials complement each other in many aspects and that they all describe today's problems in a similar way. Considering that there is consensus between the literature on the topic, it therefor deems that this validates the results of this study. If one were to increase the number of included articles, the results of the study would likely point to the same results. The amount of included material for the review was based on initial plan and estimate, that would be possible to properly analyze in the given time frame.

6.3 Reviewing and Analyze Process

As mentioned in the method chapter, it is clear and motivated how this study has been conducted for transparency and to encourage reproducibility. The documented process is used and no deviation from this is made. This has presented the identified 22 used articles in this study and reviewing of them has led to the described analysis and conclusion.

The search string used to probe the databases could have used modifications to it. As the keywords used mostly got hits relevant to the authentication chapter of this study. The backward snowballing process did, however, identify literature that complemented and

supported the other chapters of the analysis. Without the used articles from the snowballing process, there would have been a bigger validity threat to the filtering technique chapter.

The search string did not work well towards the ACM Digital Library database, the most generated hits with the least included literature from this source. The search string applied towards this database could have definitely used some modifications to it, because it is very likely that there are over four relevant materials in this database. After applying language and year filter towards the material generated by ACM Digital Library, the number of remaining materials were still over their threshold, as this database omits materials after 2000 hits.

The relevance process was laborious. During the process it was being conducted, full assurance cannot be determined that an article is truly relevant or not. As in the first stage of the relevance process, literature was excluded based on titles and relevant literature may not have relevant titles. Best efforts were made to not exclude any relevant material during this process, but it would be reckless to claim that no relevant literature was missed out on, because of the applied inclusion-, exclusion-criteria and the relevance process.

Excel and the built-in marker in Adobe Acrobat Reader were used to conduct the thematic synthesis. This process could have been conducted with other tools in a much more efficient manner. The importance is however that it does not deem that the conduction of this process led to any important missed themes from emerging, or any valuable information in this context were missed out on in the literature.

6.4 Ethical Considerations

There are some points to comment on regarding the ethical aspects of this study. Quotes are used to ensure that no facts or information are taken out of context or misinterpreted. By putting certain facts within quotations, the original meaning of the sentence is ensured so that it is not shortened, that the subject or purpose is distorted. An ethical point of view linked to this study is the fact that the impact of sending undesired emails should be considered as immoral and criminal. The study mentions several technical issues and evasion techniques in existing solutions. That could theoretically be abused by a bad actor reading this study. However, all material used could also be accessed elsewhere by a bad actor and it is determined that this factor does not cause it to outweigh the benefits of pointing out issues and evasion techniques in existing solutions, thus so readers may take more educated decisions when taking countermeasures towards undesirable electronic messages.

6.5 Societal Impact

To be affected by undesired emails, regardless of whether it is spam, spoofing or phishing, it has a great economic impact on our society. User can be affected either by receiving undesired emails, and the user can also be affected just as much by not receiving emails that is expected, as these may have been discarded due to all the regulations in the email system. Unfortunately, either receiving something undesired or not receiving what you expect can lead to a big impact for everyone who suffers from it. Translated into real impact, it can have big consequences for those affected. This study's results can act as a collective answer to many of the aspects surrounding how email services today are stopping undesired emails, and current limitations of email services. The study's result can be found useful for students and professionals alike when contemplating on what solutions to apply, or for the curious reader about the state of the countermeasures that are being utilized today, all to aid the reader to make more informed decisions.

7 Conclusion

The fight against spam and phishing have been an ongoing battle the past decades and is likely to continue. There are still many concepts from the past that are being used in today's email systems to prevent undesired emails. The countermeasures being used today differ a lot in terms of how effective they are, how they are being used and implemented. To answer the study's research question in a summarized fashion.

How do mail components on the server side detects, and process undesired emails?

Authentication and filtering techniques are today being utilized in email systems to detect and process undesired emails. The study finds that there is protection to prevent undesired emails and the different methods protect up to a certain degree but remains more to be desired. Content-filtering techniques can provide a relatively high level of protection, and in contrast there is almost no protection against domain spoofing. The usage rate of the various authentication protections is low and contains technical issues within the protocols. Even with a careful deployment of today's authentication protocols, your domain can still be abused by domain spoofing.

Limitations for this study are the selection criteria of the literature, the predetermined keywords that have been used search the literature and the low number of included materials, which may impact the result of the study. The study's research question has been answered up to a certain degree, but a fully extent answer could not be achieved due to reasons such as the study's limitations and most state-of-the-art research regarding filtering techniques are not publicly available. The findings are, however, enough to make a fair estimation to why undesired emails reach the user. It mainly comes down to certain contributing key factors. The usage rate of authentication protocols is low and contain technical issues, thus lead today's email systems to not be able to reliably authenticate all incoming emails. It cost more to lose a legitimate email than delivering an undesired email. Email services are therefore inclined to have relaxed regulation, simply because it is more cost efficient. Loosing legitimate emails is today unacceptable and will drive users away from their email service.

The study has mapped what solutions are being used to prevent undesired emails and their shortcomings. When understanding and being aware of what protection exists against undesired emails, it can lead to more educated decisions in this context, therefore the result of this study is of importance to help reduce the risks of being affected by undesired emails.

To tie the knot of this study, it will refer back to a huge motivation why this research was conducted. It was to understand why undesired emails reach a user's inbox and what has led the end user to become the last, but also best defense in the fight against undesired emails.

- *Today's email systems can't reliably authenticate all incoming emails.*
- *Today are email services more inclined to deliver an undesired email than risk losing a legitimate email.*

These factors contribute to the conclusion that users today must be the last, and the best defense. Man is in the end, a weak link but it is still only us who in the end can make a correct classification of an email.

Regarding future research, although the selected number of included articles in this study has been low, the study still manages to explain that there is a problem area and it is evident that unwanted shortcomings remain in today's email systems. Therefore, there is potential for future research on this subject. Possible areas to investigate, in order to improve protection against undesired emails are:

- Research to investigate and understand why there is a relatively low utilization rate for today's authentication methods.

Pointed out in the literature is that higher utilization grade of these authentication protocols will improve the protection for all domains. It is important to understand why the adoption and utilization rate is low, as that will be part of the answer in order to improve utilization grade. Research must also be done in the area of deployment of standardized protocols, since it is likely that future solutions may face the same adoption challenges.

- Research to develop a more transparent way of indicating users of the evaluation process of an email.

Until we reach 100% detection, with 0% false-positive and false-negative rates of undesirable emails, a better way of informing the users about the evaluation process of an email should be addressed. Thus, help the user to make more educated estimations whether an email is legitimate or not.

References

Adil, M., Khan, R., & Nawaz Ul Ghani, M. A. (2020). Preventive Techniques of Phishing Attacks in Networks. *3rd International Conference on Advancements in Computational Sciences, ICACS 2020*. https://doi.org/10.1109/ICACS47775.2020.9055943

Al-Hussaini, S., Al-Thani, D., & Yang, Y. (2020). Are They Likely to Complain on Phish or Spam? A Prediction Model. *2020 7Th International Conference On Behavioural And Social Computing (BESC)*. https://doi.org/10.1109/besc51023.2020.9348318

Allodi, L., Chotza, T., Panina, E., & Zannone, N. (2020). The Need for New Antiphishing Measures against Spear-Phishing Attacks. *IEEE Security and Privacy*, *18*(2), 23–34. https://doi.org/10.1109/MSEC.2019.2940952

Athulya, A. A., & Praveen, K. (2020). Towards the Detection of Phishing Attacks. *Proceedings of the 4th International Conference on Trends in Electronics and Informatics, ICOEI 2020*, *Icoei*, 337–343. https://doi.org/10.1109/ICOEI48184.2020.9142967

Bajaj, K. S. (2017). A multi-layer model to detect spam email at client side. *Lecture Notes of the Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering, LNICST*, *198 LNICST*, 334–349. https://doi.org/10.1007/978-3-319-59608-2_20

Blanzieri, E., & Bryl, A. (2008). A survey of learning-based techniques of email spam filtering. *Artificial Intelligence Review*, *29*(1), 63–92. https://doi.org/10.1007/s10462-009-9109-6

Booth, A., Sutton, A., & Papaioannou, D. (2016). *Systematic Approaches to a Successful Literature Review* (2nd ed.). SAGE Publications.

Both, D. (2020). Using and Administering Linux: Volume 3. In *Using and Administering Linux: Volume 3* (Vol. 3).

Chanti, S., & Chithralekha, T. (2020). Classification of Anti-phishing Solutions. *SN Computer Science*, *1*(1), 1–18. https://doi.org/10.1007/s42979-019-0011-2

Chim, H. (2005). To build a blocklist based on the cost of spam. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *3828 LNCS*, 510–519. https://doi.org/10.1007/11600930_51

Dada, E. G., Bassi, J. S., Chiroma, H., Abdulhamid, S. M., Adetunmbi, A. O., & Ajibuwa, O. E. (2019). Machine learning for email spam filtering: review, approaches and open research problems. *Heliyon*, *5*(6). https://doi.org/10.1016/j.heliyon.2019.e01802

Daily number of e-mails worldwide 2025 | Statista. Statista. (2021). Retrieved 13 March 2021, from https://www.statista.com/statistics/456500/daily-number-of-e-mails-worldwide/.

Dhru, N. (2018). Office 365 for healthcare professionals: Improving patient care through collaboration, compliance, and productivity. *Office 365 for Healthcare Professionals: Improving Patient Care Through Collaboration, Compliance, and Productivity*, 27–52. https://doi.org/10.1007/978-1-4842-3549-2

Ferrara, E. (2019). The history of digital spam. *Communications Of The ACM*, *62*(8), 82-91. https://doi.org/10.1145/3299768

Gangavarapu, T., Jaidhar, C. D., & Chanduka, B. (2020). Applicability of machine learning in spam and phishing email filtering: review and approaches. In *Artificial Intelligence Review* (Vol. 53, Issue 7). Springer Netherlands. https://doi.org/10.1007/s10462-020-09814-9

Giorgi, G., Saracino, A., & Martinelli, F. (2020). Email spoofing attack detection through an end to end authorship attribution system. *ICISSP 2020 - Proceedings of the 6th International Conference on Information Systems Security and Privacy*, 64–74. https://doi.org/10.5220/0008954600640074

Gomes, V., Reis, J., & Alturas, B. (2020). Social Engineering and the Dangers of Phishing. *Iberian Conference on Information Systems and Technologies, CISTI*, *2020-June*(June), 24–27. https://doi.org/10.23919/CISTI49556.2020.9140445

Gupta, B. B., Tewari, A., Jain, A. K., & Agrawal, D. P. (2017). Fighting against phishing attacks: state of the art and future challenges. *Neural Computing and Applications*, *28*(12), 3629–3654. https://doi.org/10.1007/s00521-016-2275-y

Herzberg, A., 2009. *Emerging challenges for security, privacy and trust*. Berlin: Springer, pp.13-16.

Hu, H., Peng, P., & Wang, G. (2018). Towards understanding the adoption of anti-spoofing protocols in email systems. *Proceedings - 2018 IEEE Cybersecurity Development Conference, SecDev 2018*, 94–101. https://doi.org/10.1109/SecDev.2018.00020

Hu, H., & Wang, G. (2018). End-to-end measurements of email spoofing attacks. *Proceedings of the 27th USENIX Security Symposium*, 1095–1112.

IBM. (2020). *Cost of a data breach report*. IBM. Retrieved from https://www.ibm.com/downloads/cas/QMXVZX6R

Iedemska, J., Stringhini, G., Kemmerer, R., Kruegel, C., & Vigna, G. (2014). The tricks of the trade: What makes spam campaigns successful? *Proceedings - IEEE Symposium on Security and Privacy*, *2014-Janua*, 77–83. https://doi.org/10.1109/SPW.2014.21

Jakobsson, M. (2017). Short paper: Addressing sophisticated email attacks. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *10322 LNCS*, 310–317. https://doi.org/10.1007/978-3-319-70972-7_17

Jakobsson, M. (2016). Understanding Social Engineering Based Scams. *Understanding Social Engineering Based Scams*, 51–62. https://doi.org/10.1007/978-1-4939-6457-4

Jung, J. J., & Jo, G. S. (2003). Collaborative junk E-mail filtering based on multi-agent systems. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *2713*, 218–227. https://doi.org/10.1007/3-540-45036-x_22

Karimovich, G. S., Jaloldin Ugli, K. S., & Salimbayevich, O. I. (2020). Analysis of machine learning methods for filtering spam messages in email services. *2020 International Conference on Information Science and Communications Technologies, ICISCT 2020*, 6–9. https://doi.org/10.1109/ICISCT50599.2020.9351442

Kitchenham, B. (2004). Procedures for performing systematic reviews. *Keele, UK, Keele University*, *33*(2004).

Konno, K., Dan, K., & Kitagawa, N. (2017). A Spoofed E-Mail Countermeasure Method by Scoring the Reliability of DKIM Signature Using Communication Data. *Proceedings - International Computer Software and Applications Conference*, *2*, 43–48. https://doi.org/10.1109/COMPSAC.2017.37

Konno, K., Kitagawa, N., & Yamai, N. (2020). False Positive Detection in Sender Domain Authentication by DMARC Report Analysis. *ACM International Conference Proceeding Series*, 38–42. https://doi.org/10.1145/3388176.3388217

Kouchaksaraei, H. R., & Karl, H. (2019). Service function chaining across openstack and kubernetes domains. *DEBS 2019 - Proceedings of the 13th ACM International Conference on Distributed and Event-Based Systems*, 240–243. https://doi.org/10.1145/3328905.3332505

Lakshmi, V. (2019). Beginning Security with Microsoft Technologies. In *Beginning Security with Microsoft Technologies*. https://doi.org/10.1007/978-1-4842-4853-9

MacDonald, J. (2014). Systematic Approaches to a Successful Literature Review. In *Journal of the Canadian Health Libraries Association / Journal de l'Association des bibliothèques de la santé du Canada* (Vol. 34, Issue 1). https://doi.org/10.5596/c13-009

Maroofi, S., Korczynski, M., Holzel, A., & Duda, A. (2021). Adoption of Email Anti-Spoofing Schemes: A Large Scale Analysis. *IEEE Transactions on Network and Service Management*, *4537*(c), 1–13. https://doi.org/10.1109/TNSM.2021.3065422

Nanaware, T., Mohite, P., & Patil, R. (2019). DMARCBox - Corporate Email Security and Analytics using DMARC. *2019 IEEE 5th International Conference for Convergence in Technology, I2CT 2019*, 1–5. https://doi.org/10.1109/I2CT45611.2019.9033552

Opazo, B., Whitteker, D., & Shing, C. C. (2018). Email trouble: Secrets of spoofing, the dangers of social engineering, and how we can help. *ICNC-FSKD 2017 - 13th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery*, 2812–2817. https://doi.org/10.1109/FSKD.2017.8393226

Paré, G., Trudel, M., Jaana, M., & Kitsiou, S. (2015). Synthesizing information systems knowledge: A typology of literature reviews. *Information & Management*, *52*(2), 183-199. https://doi.org/10.1016/j.im.2014.08.008

Pompon, R. (2016). IT security risk control management: An audit preparation plan. *IT Security Risk Control Management: An Audit Preparation Plan*, 219–229. https://doi.org/10.1007/978-1-4842-2140-2

Shajideen, N. M., & Bindu, V. B. (2018). Spam Filtering: A Comparison between Different Machine Learning Classifiers. *Proceedings of the 2nd International Conference on Electronics, Communication and Aerospace Technology, ICECA 2018*, *Iceca*, 1919–1922. https://doi.org/10.1109/ICECA.2018.8474778

Snyder, H. (2019). Literature review as a research methodology: An overview and guidelines. *Journal Of Business Research*, *104*, 333-339. https://doi.org/10.1016/j.jbusres.2019.07.039

Sokolov, M., Olufowobi, K., & Herndon, N. (2020). Visual spoofing in content-based spam detection. *ACM International Conference Proceeding Series*. https://doi.org/10.1145/3433174.3433605

Trivedi, S. K. (2016). A study of machine learning classifiers for spam detection. *2016 4th International Symposium on Computational and Business Intelligence, ISCBI 2016*, 176–180. https://doi.org/10.1109/ISCBI.2016.7743279

The Radicati Group, Inc. (2020). *Email Statistics Report, 2020-2024*. The Radicati Group, Inc. Retrieved from https://www.radicati.com/wp/wp-content/uploads/2019/12/Email-Statistics-Report-2020-2024-Executive-Summary.pdf

Verizon. (2020). *2020 Data Breach Investigations Report*. Verizon. Retrieved from https://enterprise.verizon.com/content/verizonenterprise/us/en/index/resources/reports/2020-data-breach-investigations-report.pdf

Wohlin, C. (2014). Guidelines for snowballing in systematic literature studies and a replication in software engineering. *ACM International Conference Proceeding Series*. https://doi.org/10.1145/2601248.2601268

Appendix A – Included literature for the review

Label	Title	Source
A01	Towards Understanding the Adoption of Anti-Spoofing Protocols in Email Systems	(Hu et al., 2018)
A02	Email Trouble: Secrets of Spoofing, the Dangers of Social Engineering, and How We Can Help	(Opazo et al., 2018)
A03	A Spoofed E-mail Countermeasure Method by Scoring the Reliability of DKIM Signature Using Communication Data	(Konno et al., 2017)
A04	Adoption of Email Anti-Spoofing Schemes: A Large Scale Analysis	(Maroofi et al., 2021)
A05	Analysis of machine learning methods for filtering spam messages in email services	(Karimovich et al., 2020)
A06	DMARCBox – Corporate Email Security and Analytics using DMARC	(Nanaware et al., 2019)
A07	Visual spoofing in content-based spam detection	(Sokolov et al., 2020)
A08	False Positive Detection in Sender Domain Authentication by DMARC Report Analysis	(Konno et al., 2020)
A09	The History of Digital Spam	(Ferrara, 2019)
A10	A Multi-layer Model to Detect Spam Email at Client Side	(Bajaj, 2017)
A11	Classification of Anti-phishing Solutions	(Chanti & Chithralekha, 2020)
A12	Office 365 for Healthcare Professionals	(Dhru, 2018)
A13	Applicability of Machine Learning in Spam and Phishing email filtering: review and approaches	(Gangavarapu et al., 2020)
A14	Fighting against phishing attacks: state of the art and future challenges	(Gupta et al., 2017)
A15	Short Paper: Addressing Sophisticated Email Attacks	(Jakobsson, 2017)
A16	Beginning Security with Microsoft Technologies	(Lakshmi, 2019)
A17	IT Security Risk Control Management: An Audit Preparation Plan	(Pompon, 2016)

A18	Understanding Social Engineering Based Scams	(Jakobsson, 2016)
A19	A Study of Machine Learning Classifiers for Spam Detection (Snowballing)	(Trivedi, 2016)
A20	End-to-End Measurements of Email Spoofing Attacks (Snowballing)	(Hu & Wang, 2018)
A21	Machine learning for email spam filtering: review, approaches and open research problems (Snowballing)	(Dada et al., 2019)
A22	Spam Filtering : A Comparison Between Different Machine Learning Classifiers (Snowballing)	(Shajideen & Bindu, 2018)