Academic Reading Notes

Things behind skills

I used to pay so much attention to the skills I learned in class that I ignored the hidden things behind the skills, so I spent a lot of energy in this report to explore the things behind the skills.

However, because this is an English assignment, the flexible use of skills is still the center of this article.

In addition, due to different professional fields and limited time, I did not organize the assignment structure according to the thesis, and gave corresponding explanations to each quoted part. Therefore, the referenced part can be ignored during review.

最后，为了让文章易懂，在一些部分我用了中文。

Lastly, to make this article understandable, I use Chinese in some parts.

注意，有些部分仅有英文表达没有中文对应的翻译，请在阅读过程中注意这点。

Note that some parts are only expressed in English without corresponding translation in Chinese. Please pay attention to this during reading.

Research papers selected

Domain knowledge-based security bug reports prediction.

Spear Phishing Emails Detection Based on Machine Learning

Orchestration of APT malware evasive manoeuvers employed for eluding anti-virus and sandbox defense

> Zheng W, Cheng J Y, Wu X, et al. Domain knowledge-based security bug reports prediction[J]. Knowledge-Based Systems, 2022, 241: 108293.
>
> Ding X, Liu B, Jiang Z, et al. Spear Phishing Emails Detection Based on Machine Learning[C]//2021 IEEE 24th International Conference on Computer Supported Cooperative Work in Design (CSCWD). IEEE, 2021: 354-359.
>
> Sharma A, Gupta B B, Singh A K, et al. Orchestration of APT malware evasive manoeuvers employed for eluding anti-virus and sandbox defense[J]. Computers & Security, 2022, 115: 102627.

Reasons for selecting the above-mentioned papers

First of all, these papers are highly relevant to my research field.

Secondly, the comparison between them is valuable. Because these papers are published in different journals with different recognition, citations and types.

Finally, their content can basically cover what we have learned.

Draw a mind-map

I choose the paper **‘Spear Phishing Emails Detection Based on Machine Learning’** to draw a mind-map. This paper is characterized by threat intelligence platform.

Introduction of Spear Phishing Emails Detection Based on Machine Learning

Introduction to Spear phishing emails

- target
  - specifific individual
  - organization
- usage
  - by attackers
  - harvest information
    - to lure the recipient to perform dangerous actions

research status

- summary
  - some reasearch has been down
  - still no effective method
- main stream
  - method
    - authorship-based identifification
  - shortcoming
    - need a lot of historical emails including spear phishing emails
    - hard to find a new attack
- methods to solve this problem
- other work
  - method
    - using a combination of email-extracted features and auxiliary features
  - shortcoming
    - may not available for all scenario

solve the problem

features of the author's method

- third-party reputation ratings and forwarding features
- provide an interpolation method to enhance the minority samples

detail method

get the forwarding relationships of each email

- reason
  - use the great difference between spear and non-spear phishing emails

detail method

get reputation features

- reason
  - the number and frequency of visits between spear and non-spear URLs are quite different.
- method
  - VT and PT

detail method

from

use K-means clustering named KM-SMOTE to expand the dataset

- reason
  - the number of malicious samples are much smaller than the number of normal samples

Use threat intelligence platforms

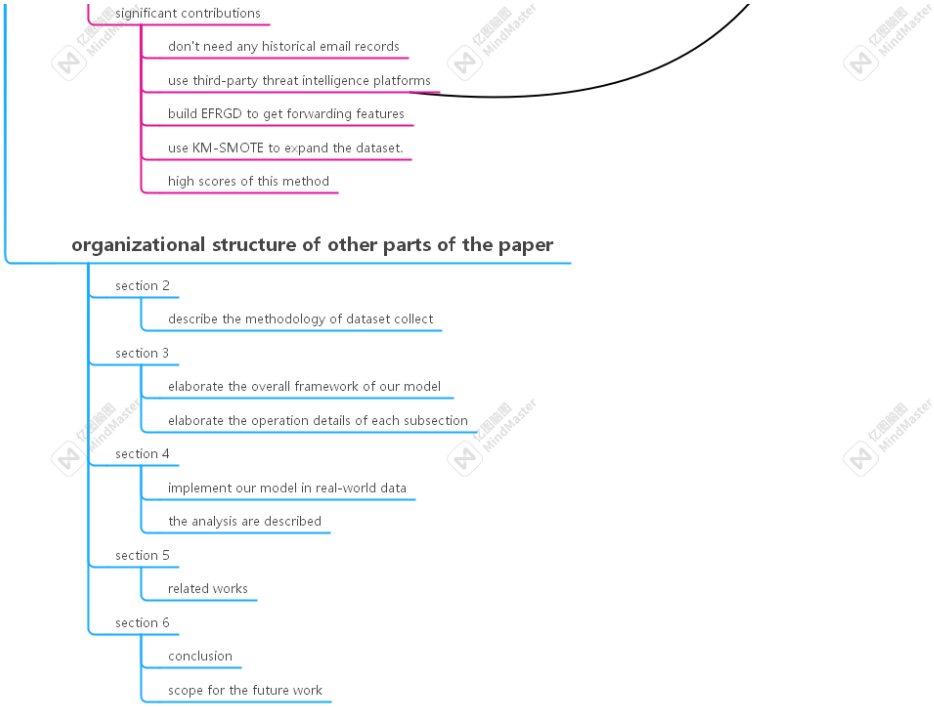
summary of this article

- the summary of the author's work
  - distinguishing spear phishing emails from non-spear phishing emails
  - propose an effective machine learning method.
- get features from emails and third-party threat intelligence platforms
- utilize KM-SMOTE algorithm
- we applied 4 machine learning algorithms to identify spear phishing emails
- Innovation
  - first work combining with threat intelligence platforms for identifying spear phishing emails

from

is

is



Example Syntheses

The chronological order

> In the past days, these attacks are aimed at defense and security organizations (Chen et al., 2014).However,they evolved rapidly in modern days to target a wide spectrum of government organizations along with large-scale private industries. Some examples of modern day APT attacks are the SolarWinds attack which is a classic example of supply chain attack(Hensley, 2021) in which the attacker targets weak links in the organizations supply chain. In the SolarWinds scenario, the attackers delivered malware to various organizations through SolarWinds Orion Network Management System(NMS). Another example is HAFNIUM APT attack (MSTIC, 2021) in which four zero-day vulnerabilities of Microsoft exchange server are used for the payload delivery. These attacks clearly indicate the sophistication and evasive nature of the attacks, which can bypass the modern security solution of the organizations.

> The motivation behind this technique is to minimize the presence of attack artifacts on the system. For conducting fifile-less attacks, frameworks like Metasploit (Kennedy et al., 2011) provide infection vector options like reflflective DLL injection, Nishang, and Powersploit, which has script-based infection options. Living off the land binaries (LOLBAS, 2018) for Windows and (Pinna and Cardaci, 2020) for Linux, scripts inside a document, LNK fifiles, multiple stage downloaders, various obfuscation tricks, Powershell, WMI, PsExce, VB scripts, and Microsoft.

These part is organized in chronological order. The first part describes the development of APT attacks in chronological order and focuses on the current situation. The second part tells how to minimize the existence of attack objects on the system in chronological order.

In the introduction and related work of these three papers, we can see that the chronological order accounts for a small proportion. At the same time, recalling the papers I have read before, I rarely use chronological order to integrate the literature. This is because most papers focus more on specific methods. However, in many cases, the method proposed early does not mean that the

method is not advanced or popular. Therefore, in many cases, references cannot be well organized in chronological order.

The climactic order

> In order to reduce security risks of software systems, security bug report (SBR) prediction has become an increasingly hot topic recently [1–6]. In software engineering, bug reports are submitted to bug tracking systems to record issues found in software prod ucts. Since SBR prediction is expressed as a binary classification problem [7–9], most of these works use text mining methods based on machine learning to achieve this, because the main information of a bug report is described with text format in the field Description [8,10–13].

This part first show the big topic security bug report(SPR) prediction. Then the author points out that this research can be set as a binary classifcation problem. Last the author talk about how machine learning can do for the classification.

Comparison and Contrast

> one year later, Gianluca Stringhini[2] refifined and optimized their IDENTITYMAILER model, they added the Composition and sending habits features, the result achieved 96% of accuracy based on 8000 history emails. Duman, Sevtap [3] presented an automated model named emails Profifiler, they extracted a total of 222 features, then subsequent emails are compared to these models to detect characteristic indicators of spear phishing attacks, the result showed that their model got 98% of accuracy.

In this part, the author compares the methods and performance of different models.

Classification

> Afifianian et al. (2019) surveyed the evasion techniques of dynamic analysis. Bulazel and Yener (2017) presented different evasion detection techniques, offensive and defensive evasion case studies, evasion mitigation, and fifingerprint-based evasion techniques against automated dynamic malware analysis systems for personal computers, web, and mobile. The limitation of this work is that it had only the survey papers of all techniques, and no implementation had been carried out. Shi et al. (2017) proposes a detect and hides approach that would systematically address only anti-virtual machine techniques. This work not emphasize on other techniques. D'Elia et al. (2020) proposed the BluePill framework for dissecting evasive malware. Herzog et al. (2020) shown the impact of evasive malware on antivirus solutions by considering a counter example of a confifigurable Nuky malware sample.

Some of the previous authors described APT malware evasive techniques and defense mechanisms in limited numbers like anti-debugging, process injection and fifile-less malware etc.

> Currently, most of the work is about authorship-based identification, such as [1]–[4], they tried to build a profile or configuration file for each sender. But there exists one big problem that they need a lot of historical emails including spear phishing emails, in that case if a sender has only a few historical records, the accuracy of identification can become very disappointing. There are also some work using a combination of email-extracted features and auxiliary features(social features, log information, etc.), such as [5]–[7], but

> these methods are still at the exploratory stage, auxiliary features may not available for all scenario, such as logs in [5], and some features may.

This part show different ways for detecting spear phishing emails.

Use of opinion markers

View markers

> If a fifile is not present, they **consider** it a sandboxed environment and alter execution.

it shows the view of 'they'.

> We are **suggesting** the security industry to thoroughly review their defense mechanisms to tackle payloads generated by frameworks like EMRF.

it shows the suggestion of the author.

It's really difficult to find in the introduction and other parts. As you can see, they rarely express their views. I think it also has disciplinary characteristics, because many contents of computer science can be quantified and want to be quantified. Therefore, if the author of the paper mentions his own views in the paper, it is often considered not rigorous. At the same time, it is not a good comparison between your own method and others' method by using opinions rather than certain data.

For example, as shown in the figure below, if you want to show that you believe something, you'd better express it in another way. You can say if dij is less than 5, how much similarity they measure by a certain standard, or how much accuracy they think they are similar. However, if you cannot find any suitable standard, you can only 'believe'.

Next for every $T_i(i = 2, 3, ..., tnum)$ in T, we calculate the *Hamming Distance* [24] d_{ij} between $simHash\_T_i$ and $emailEntity_j(j = 1, 2, ..., enum)$. If $d_{ij} <= 5$, we believe that the email represented by $emailEntity_j$ is the same to $emailEntity_i$. In this case, $SREntity_i$ will be remarked as $SREntity_{j(k+1)}$ and point to $emailEntity_j$. Otherwise, we add T_i to *RFRGD* like T_1.

where $enum$ is the number of $emailEntity$ in $EFRGD$, k is the number of $SREntity$ pointed to $emailEntity_j$.

Step 3. In this step, we extract forwarding features from $EFRGD$. For every $T_i(i = 1, 2, ..., tnum)$ in T, its $Forwarding\_Num$ is k representing the number of associated $SREntitys$. Meanwhile if $mail.server.name$ from sender and recipient are the same, we believe that they are come from the same organization.

Qualifying markers

poor, small, most of, many of, can, main and so on.

> Currently, **most of** the work is about authorship-based identifification, such as [1]–[4], they tried to build a profifile or confifiguration fifile for each sender.

> Over the past decades, there were **many** research on phishing emails detection, **most of** them applied machine learning, deep learning or NLP methods to solve the problem.

> Although **some** research has been done for detecting spear phishing emails, there is still no effective method.

these sentences show that contrast markers can express the Authors' certainty. According to my observation, these marks often appear in the literature. And it is often inappropriate to appear in the author's own methods. Sometimes it's good to show the uncertain, but sometimes it also expresses the author's lack of confidence in his own methods.

So I think the usage of qualifying markers can express our preciseness. However, we need to pay attention not to use qualifying markers when describing our own methods. For example, the word 'some' below can be replaced by approximate figures like '50%'.

> The modern security solutions use emulation technique in which the signature are applied on the decryptor body after emulation for detection. **Some** malware uses anti-emulation to thwart these mechanisms.

Contrast markers

however, but and so on.

> The results show that the use of knowledge graphs can **indeed** improve the accuracy of SBR prediction.

Note that words like indeed are not necessarily contrast markers, but also depend on the specific context. In this sentence, 'indeed' is not a contrast marker.

> Digitization assisted mankind in numerous ways by providing better services with speed and ease; **however**, it opened up a new era of attack surface known as cyberspace.

this part show that although digitization is good, but it can be attacked.

> **However**, the effectiveness of previous studies is still not ideal for the production application.

This is very classical, the author directly pointed out that previous research was flawed.

The contrast markers are very useful. We can use these to show the shortcomings of other methods and lead to our own methods. Also, we can strengthen things after contrast markers. Like the first part, we want to show that the Digitization opened up a new era of attack surface known as cyberspace.

Assessment markers

hot, ideal, authoritative, security-irrelevant, accurately.

The assessment markers are everywhere, they can express the author's judgment on something. We only need to notice assessment markers should be proper.

some examples are listed below.

> The modern day cyber attacks are highly **targeted** and incorporate **advanced** tactics, techniques and procedures for greater stealth, impact and success.

> These attacks are also known as Advanced Persistent Threats(APT) because of their **evasive** and **stealth** nature along with **longer** foothold on the victim's digital infrastructure.

Verb tenses and voices

Research background

> Spear phishing emails target to specifific individual or organization, the attackers usually harvest information about the recipient in any available way, then utilize it to lure the recipient to perform dangerous actions, such as clicking *Uniform Resource Locator*(*URL*) or downloading malicious *attachments*.

研究背景使用一般现在时，我认为这与当前的研究是已存在的问题有关。

The research background uses the simple present tense, which I think is because this article studies the existing problems.

Literature review

> Although some research **has been done** for detecting spear phishing emails, there is still no effective method. Currently, most of the work **is** about authorship-based identifification, such as [1]–[4], they **tried** to build a profifile or confifiguration fifile for each sender. But there **exists** one big problem that they **need** a lot of historical emails including spear phishing emails, in that case if a sender **has** only a few historical records, the accuracy of identifification **can** become very disappointing. There are also some work using a combination of email-extracted features and auxiliary features(*social features*, *log information*, etc.), such as [5]–[7],

讲诉作者之前具体做法的时候使用一般过去时，体现作者在已发表的论文之中这么做了。讲诉方法相关的时候使用一般现在时，体现对于当前存在的这个方法的描述，例如具体做法、特点、效果。讲诉总体的时候采用现在完成时，体现完成了什么研究。总之时态的运用非常灵活，主要取决于要表达什么。

The general past tense is used when telling what the author did before, which reflects what the author did in the papers published in the past. The general present tense is used when the presentation method is relevant, reflecting the description of the current method, such as specific methods, characteristics, and effects. The present perfect tense is used as the overall time of speaking, reflecting what research has been completed. In a word, the use of tenses is very flexible, mainly depending on what you want to express.

Research gaps

> but these methods **are** still at the exploratory stage, auxiliary features **may** not available for all scenario, such as logs in [5], and some features **may** be useless, such as features extracted from LinkedIn [6].

这一部分用的一般现在时，但是我认为这个与前面的Literature review相似，因为都是从文献中得出的。

this part use simple present tense, but I think it's similar to the literature review. They are all form literatures.

Research questions

> In order to solve these problem, in this paper, we **use** third-party reputation ratings and forwarding features as our auxiliary features, which will bring reliable reputation information and forwarding relationships, and we provide an interpolation method to enhance the minority samples.

研究问题是为了解决Reaserch gaps提出的问题。毫无疑问是一般现在时，因为问题是我们现在面对的。

The question is to solve the problems that proposed in Reaserch gaps. There is no doubt to use simple present tense, As question is what we faced now.

Research significance

要做一个好的讲故事的人。

for the paper I focused on, there is no sentence that describe the significance for this research. But I found the research significance in another paper. I think be a good storyteller is very important for us.

> The research is aimed at providing comprehensive details on evasive techniques along with their efficacy in evading security solutions such that better defense mechanisms will be developed against modern APT attacks.

这一部分讲了研究可以阻止APT攻击，老生常谈。也是毫无疑问的一般现在时。

This part talks about how research can prevent APT attacks, which has been talked in many papers. There is no doubt to use simple present tense， too.

Other moves:

I think other moves are mostly about the summary of the author's method. Use simple present tense to express these.

> First, spear and non-spear phishing emails are quite different in forwarding relationships. For spear phishing emails, they are specifific to individual or corporate executives, so they are unlikely to be sent or forwarded on a large scale. On the contrary, for non-spear phishing emails, they are more likely to be sent or forwarded for multiple times. In order to take full advantage of this big difference, we get the forwarding relationships of each email.

summary

总的来说，我不认为按照研究背景等去严格区分时态是一个好主意。我认为主要还是关于到底要表达什么，不过这个这一部分确实是对于自己的论文写作过程有所帮助，可以帮助我们发现使用的时态语态到底有没有问题。

In general, I don't think it is a good idea to strictly distinguish tenses according to the research background and so on. I think it is mainly about what to express, but this part is really helpful to the writing process of my thesis, and can help us find out whether there is any problem with the voice or verb tenses used.

Research gaps

很难找到真的研究空白，尤其是在计算机领域，绝大部分研究方法都有人试验过。然而，总是有方法没有人尝试过，总是有效果更好的方法，一般我们指的研究空白都是这一部分。

It is difficult to find a real research gap, especially in the field of computers, where most of the research methods have been experimented with. However, there are always methods that nobody has tried, and there are always better and effective methods. In general, the research gaps we refer to is this part.

> Although some research has been done for detecting spear phishing emails, there is still no effective method. Currently, most of the work is about authorship-based identifification, such as [1]–[4], they tried to build a profifile or confifiguration fifile for each sender. But there exists one big problem that they need a lot of historical emails including spear phishing emails, in that case if a sender has only a few historical records, the accuracy of identifification can become very disappointing. There are also some work using a combination of email-extracted features and auxiliary features(*social features*, *log information*, etc.), such as [5]–[7], **but these methods are still at the exploratory stage, auxiliary features may not available for all scenario, such as logs in [5], and some features may be useless, such as features extracted from LinkedIn [6].**

例如，这里的Research gap是没有在大多数领域都有效的特征。

for example, the research gap here is that there is no useful features which can be effective in most fields.

类型就是没有更适用的方法，更有效的方法。使用了一般现在时。

The type of Research gap is that there is no more applicable and effective method. The simple present tense is used.

Useful phrases

This part is very similar to the previous part.

Highlighting an important issue

> **Although some** research **has been done** for detecting spear phishing emails, there is **still no effective** method.

这一部分直截了当地表述了当前存在的问题。没有什么讲述什么故事。论文最重要的目的就是为了让人看懂，对于这些重要的信息，应该使用简单的表达。由于要表达转折关系，使用了contrast marker。同时直接表明没有有效的方法，为了之后内容做铺垫。这一部分用了现在完成时和一般现在时结合的做法。现在完成时用来强调这些是过去的研究，一般现在时用来说明现存的问题。

提炼的表达：

Although + 现有研究 + (转折)+ 现有问题

现有已经做出的研究使用现在完成时，现有问题使用一般现在时。

This section is a straightforward statement of the current problems. There is no story to tell. The most important purpose of the paper is to make people understand. For these important information, we should use simple expressions. Because it is necessary to express the transition relationship, the contrast marker is used. At the same time, it directly indicates that there is no effective method to pave the way for future content. This part combines the present perfect tense

with the present tense. In this part the present perfect tense is used to emphasize that these are past studies, and the present tense is used to explain existing problems.

generalization:

Although + existing studies + (contrast marker) + existing problem

The present perfect tense is used for the existing studies, and the present simple tense is used for the existing problems.

Highlighting inadequacies or weaknesses of previous studies

> Although some research has been done for detecting spear phishing emails, **there is still no effective method**. Currently, most of the work is about authorship-based identification, such as [1]–[4], they tried to build a profile or configuration file for each sender. **But there exists one big problem that they need a lot of historical emails including spear phishing emails, in that case if a sender has only a few historical records, the accuracy of identification can become very disappointing.** There are also some work using a combination of email-extracted features and auxiliary features(*social features*, *log information*, etc.), such as [5]–[7], **but these methods are still at the exploratory stage, auxiliary features may not available for all scenario, such as logs in [5], and some features may be useless, such as features extracted from LinkedIn [6].**

一般都是先介绍方法，再介绍缺点和不足。先说优点再说缺点。在这一部分是层层递进的，由一个解决方法的缺点引出另外一个解决方法，最后引到我们的解决方法。这里的表达模式很明显是climactic order。

提炼的表达：

已经存在的研究 + (Contrast marker) + 存在的问题 + (能够解决问题的研究 + (Contrast marker) + 存在的问题 + (进一步深入的研究))

剩下的见英文表述。

Generally, the method is introduced first, then the inadequacies or weaknesses of previous studies. Advantages before disadvantages. This part is progressive, leading from one solution's shortcomings to another and finally to our solutions.

generalization:

Existing studies + **(Contrast marker)** + existing problems + (**Problem solving studies** + (Contrast markers) + existing problems(new) + (further studies))(**like climactic order**)

But there exists one big problem that(**claiming** the problem)

in that case if ... , ... **can be/become**(problem **in a specific situation**)

but these methods are still, ..., **and some methods**(from these to some, **more specific**)

Proposing research questions or hypotheses

> **Spear phishing emails** target to specifific individual or organization, the attackers usually harvest information about the recipient in any available way, then utilize it to lure the recipient to perform dangerous actions, such as clicking *Uniform Resource Locator*(*URL*) or downloading malicious *attachments*.

> In order to **solve these problem**, in this paper, we use third-party reputation ratings and forwarding features as our auxiliary features, which will bring reliable reputation information and forwarding relationships, and we provide an interpolation method to enhance the minority samples.

问题是如果解决'困难'（上文提到的）。并没有假设，我们只要知道研究的是**spear** phishing emails即可

提炼的表达：

详见英语表达部分。

The research question is how to 'solve these problem'. And there is no hypothesis, the only thing we need to know is that this study is about **spear** phishing emails.

generalization:

In order to solve these problems, **in this paper**, we **use** something, which **will do** something, and we do something to achieve some **other purposes**.

The threat **target to** somebody, the attackers do something in a specific way, **then** do something to achieve their goals, **such as** something.

Indicating research significance or value

Research significance in the paper we focused on

> To the best of our knowledge, this is the first work combining with threat intelligence platforms for identifying spear phishing emails. The significant contributions of this paper are as follows:
>
> • Our proposed approach does not require any historical email records, or other auxiliary logs, etc
>
> • For the first time, we bring third-party threat intelligence platforms *VT* and *PT* to gain reputation features, such as Research significance in another paper.
>
> • In order to make full use of the difference in forwarding relationship between spear and non-spear phishing emails, we build *EFRGD* to get forwarding features.
>
> • To reduce the impact of unbalanced data, we provide a promotion of interpolation algorithm KM-SMOTE to expand the dataset.
>
> • Experimental result on real-world data shows that our model has very high recall, precision and F1-score.

这一部分讲诉了创新点和重要性。创新点是第一次结合威胁情报与鱼叉式钓鱼。重要性有四点，总的来说就是更少的数据要求，引用的威胁情报，新的转发特征获取方法，扩展了数据集，以及高得分。

This part tells the innovation and significance. The innovation is the first combination of threat intelligence platforms and spear phishing emails.

significances are summarized below:

1. don't need any historical email records
2. use third-party threat intelligence platforms
3. build EFRGD to get forwarding features
4. use KM-SMOTE to expand the dataset.
5. high scores of this method

作者强调了简单性，有效性，创新点。在我们的论文之中也要注意这些。

In summary, the author emphasizes simplicity, effectiveness and innovation. We should also pay attention to these in our papers

The importance of another article is shown below.

> The research is aimed at providing comprehensive details on evasive techniques along with their efficacy in evading security solutions such that better defense mechanisms will be developed against modern APT attacks.

提炼如下。

generalization:

To the best of our knowledge, this is **the first work combining with** something **for** doing something. (To the best of our knowledge, the first work, combine with ... for ...)

The significant contributions of this paper are as follows:

Our **proposed** approach **does not require** any thing, or other thing, **etc**

For the first time, we **bring** something **to gain** something.

In order to make full use of something, we do something.

Experimental result on real-world data shows that **our model has very high** score(according to some judging standards).