Looking under the Streetlights

Evaluating Cyber Threat Intelligence Feeds Using Quantitative Metrics and User Appreciation Scores

J.T. Egbers

TU Delft

Looking under the Streetlights

Evaluating Cyber Threat Intelligence Feeds Using Quantitative Metrics and User Appreciation Scores

by

J.T. Egbers

Master thesis submitted to Delft University of Technology
in partial fulfillment of the requirements for the degree of
Master of Science
in **Engineering and Policy Analysis**
Faculty of Technology, Policy and Management

to be defended publicly March 1, 2021 at 11:00.

Student number:	4846249	
Project duration:	September 1, 2020 – March 1, 2021	
Thesis committee:	Prof. dr. M.J.G. van Eeten,	TU Delft, Chairperson
	Prof. dr. M.E. Warnier,	TU Delft, Supervisor
	Prof. dr. ing. A.J. Klievink,	Universiteit Leiden, Advisor
	Ir. X. Bouwman,	TU Delft, Advisor

TU Delft

Preface

> A master's student sees a group of drunk Threat Intelligence researchers searching for something using a wide collection of blacklist metrics and asks what the drunk researchers are looking for. They say they are looking for the value of Threat Intelligence and he joins them in their (re)search. After a few months, the master's student asks if they are sure the value of Threat Intelligence can be found using this set of metrics. The drunk researchers reply "no, you should be using metrics specifically designed for Threat Intelligence". The master's student asks why they are using this set of metrics, and the drunk researchers reply, "this is what has been done in literature before".
>
> —adjusted from Freedman [25]

This quote is an adjusted version of the story of the drunkard's search. A policeman helps a drunkard look for his keys under the streetlight. When the policeman asks if he is sure that he lost them there, the drunkard replies he lost them a while back, but it was easier to look under the streetlight (hence the title of this thesis).

When reading this anecdote, you might think that I call each member of my graduation committee a drunkard searching under the streetlight. Although I cannot be sure if the literal meaning is true, I can say for sure the anecdotal meaning is more nuanced than this story implies. Each one of my supervisors has been open for discussion and new ideas, candidly saying when new ideas were good but also saying when they required some more tinkering. For this, and for their overall help and guidance, I want to thank them. I want to specifically highlight the role of Xander in this ongoing feedback process. He was both able to celebrate small milestones, as well as to not lose sight of the forest for the trees and nudge me in the right direction, for which I am thankful.

In addition to this help relating to the contents and direction, I am grateful to the interviewees for giving their valuable time and opinions which helped me to make sense of the context this thesis was written in. Finally, to everyone who (un)knowingly gave me ideas, (proof)read (parts of) this work, and who generally supported and tolerated me in a time that is known to a lot of students as 'the most stressful period in their studies', thank you.

The image used on the front page is the background image on the website of the 'COVID-19 Cyber Threat Coalition' [17]. This felt as a suitable depiction of both the contents of the study, as well as the period I was writing this thesis in. You can see a computer protected against viruses, cleverly taking advantage of the double-meaning: computer viruses and the Coronavirus. The computer virus part relates to the contents of this study, Threat Intelligence is one way to help you protect yourself against these. The Coronavirus part of the image is present because the Cyber Threat Coalition is established to help in the protection against cyber threats that arise because of the COVID-19 pandemic, that part is fitting for this study because not only my, but everyone's life was impacted by this virus during the period of writing.

J.T. Egbers
Delft, February 2021

Executive Summary

In the battle against ever-changing cyber threats, a new ally has joined in: Cyber Threat Intelligence. Evolved from historical blacklists and anti-virus, Threat Intelligence aims to protect and inform its clients against both nation state actors, as well as cyber criminals. Threat Intelligence comes in many shapes and sizes, and for a wide range of prices. For the average consumer of Threat Intelligence, it is unknown which form will fit their needs, nor which price range is suitable for them.

This mystery surrounding Threat Intelligence, caused by its prohibitively high pricing, shows in the limited amount of research that has been conducted on the topic. Bouwman et al. [12] lifted a tip of the veil, interviewing professionals regarding their use of Threat Intelligence and presenting descriptive statistics of its contents. They found very limited overlap between Threat Intelligence sources and that acquisition is largely based on gut-feeling. However, it is still largely unknown if these findings generalize to the whole field of Threat Intelligence and if these findings on macro level translate to more granular levels, it is our goal to find this out.

In order to do this, Threat Intelligence literature is consulted to find a set of 8 metrics which can be used to assess the contents of Threat Intelligence sources: volume, timeliness, overlap, population, security threats, information type, objectivity and report counts. These metrics will be calculated for a data-set consisting of Threat Intelligence reports, published by 5 different Threat Intelligence vendors. Next to the metrics from these reports, 623 user appreciation scores about these reports are collected from around 50 – 100 Threat Intelligence users.

Overall, these user appreciation scores show that users are generally pleased with the quality of Threat Intelligence reports, as two-thirds of all scores given are a five out of five star rating. From these ratings, it shows that most users are not scoring often, as more than two-thirds of the users have scored less than 5 reports and only less than a dozen of people have scored more than 20 reports.

The report metrics show that all vendors perform comparably with respect to their timeliness of publishing indicators. In line with findings in literature, each of the vendors does not have more than 7% of their reported indicators overlap with other vendors. This indicates either a lack of coverage by each vendor or the presence of many false positives.

The report metrics are also compared to the user appreciation scores to test for correlation. Although this results in some minor significant correlations, these change when slightly altering the measurement period. From this we draw the conclusion that none of

the correlations show conclusive relations between report metrics and users' scoring behavior.

Different predictive algorithms are applied to the report metrics and the user appreciation scores as well. When supplemented with the amount of views and downloads, machine learning techniques achieve promising scores when predicting if and how many votes a report will receive. However, predicting the actual value of the scores did not succeed. Using collaborative filtering and content-based recommendation systems, recall scores three times greater than random chance were achieved.

The lack of consistent and significant relations between report metrics and user appreciation scores suggest that the real factor of influence for Threat Intelligence reports' quality has not been measured. This was validated by the fact that report scores were not well predicted, despite both the number of votes and interactions being reasonably well predictable. Although this result was predictable following the results of Bouwman et al. [12], the fact that most other literature regarding Threat Intelligence continues to mention and introduce metrics to evaluate it, shows this opinion is not widespread. For the field of Threat Intelligence this also means that quality and value are still hard to quantify and more research into the topic is necessary. We argue that this could be caused by the fact that Threat Intelligence did not emerge as a separate field, but often consists of rebranded blacklists. Better distinguishing blacklists without context and Threat Intelligence *with* context from each other could be the first step towards developing Threat Intelligence as its own field and hopefully more meaningful results.

Next to this suggestion for the field as a whole, there are different opportunities and suggestions that can be applied more directly to the contents of this thesis. The most obvious one is the element of time. The instruments and tools that gathered the data that enabled this thesis, did not stop the moment this analysis started. In several months the amount of user appreciation scores will surpass the one thousand vote milestone, increasing the reliability of future analyses. One especially exciting addition to the data will be the collection of explanatory keywords, which will add nuance to each of the ratings. These keywords, if collected in greater numbers, will show a detailed picture of the preferences and aversions of Threat Intelligence users. Not only that, this data can be used to create profiles of different vendors, showcasing their strengths and weaknesses according to the users.

Next to the great opportunities of the current system, there needs to be an ongoing process of experimentation and evaluation. Whereas the current system exists of a 5 star rating extended with 9 descriptive keywords, different options as a like-dislike or emotion-oriented systems like Facebook and LinkedIn can be considered. Not only the method of voting can be subjected to changes, the appreciation scores can also be used to enhance the experience of the Threat Intelligence users. The results of the recommendation systems show it is already possible to recommend relevant reports based on the behavior of similar users and report contents.

Contents

List of Figures

List of Tables

1 Introduction

1.1. Cyber threats are developing

The world is becoming increasingly digital. At the start of 2020, more than 4.5 billion people were using the internet [33]. Shortly after this measurement, the COVID-19 pandemic made it painfully clear how dependent we are on the internet. We need it for every aspect of our lives, our work, our education, and our social contacts. This caused the use of online applications to surge to levels at least 200% higher than pre-pandemic levels[24]. One perceptive Twitter user stated that not the CEO nor CTO of a given company were responsible for their businesses' digital transformations, but COVID-19 [74]. Although digitalization can generally be seen as a positive development, Williams et al. [77] found a five-time increase in cyber attacks since the start of the pandemic. These troubling developments add to the already existing situation that parts of our critical infrastructure are connected to the internet, systems that are not always well protected [62].

These vulnerabilities are exploited by a wide range of actors with different goals. Gordon and Ford [26] highlight a range of cyber attacks targeting both individuals, organizations and countries, but also being executed by individuals, (semi-)organized groups, and nation state actors. The reasons for cyber attacks include but are not limited to: monetary gain, blackmail, intelligence gathering, and vandalism. A familiar example, as most of us might remember, are the golden years of Limewire, which provided an unlimited source of free music, but also proved dangerous for the unaware user as more than half of its files were infected with malware [30].

As time goes on, not only the amount of vulnerabilities increases, the threats themselves develop as well. Both the volume and the sophistication of nation-state cyber threats keep increasing [14]. Security Magazine reports a 42% increase in cyber attacks attributed to foreign governments in 2019 [53]. Microsoft reports that cyber threat actors keep increasing their abilities to remain unseen, as well as develop the possibilities to target higher-end victims [14]. A perfect example of this is the SolarWinds hack, where an allegedly Russian hacker group gained illegal access to dozens of large organizations, including United States government organizations. This hacker group not only managed to get inside the governmental infrastructure of one of the biggest global powers, they also remained undetected for several months before being called out by FireEye, one of the affected organizations [35, 39, 61].

As threats develop, so should defenses. Anti-virus software and blacklists, lists of malicious IPs and files, historically have been the go-to ally for this purpose [7]. Roughly 6 years ago, a new alternative has come to market: cyber Threat Intelligence (TI). Often well-known players in the cybersecurity scene either re-branded or extended their offering. Most major cybersecurity companies offer some sort of Threat Intelligence products nowadays.

1.2. Threat Intelligence

Threat Intelligence comes in a variety of shapes and sizes. As mentioned above, some vendors re-branded their blacklists to TI. These blacklists consist of lists containing (malicious) IP addresses, domains, URLs, and file hashes, used to monitor incoming and outgoing traffic, and block traffic with malicious intent. More extensive forms of TI include written reports, often in PDF format [12]. These reports cover a wide range of topics. A TI report can be a short heads-up about a new type of malware that is detected or it can be a 10+ page report with an in-depth analysis of that piece of malware. This does not mean all TI discusses only

technical issues. Some reports discuss the Techniques, Tactics, and Procedures (TTPs) of an actor, or give the reader an overview of the political context of a cyber campaign [12].

All this information can be analyzed at different levels of granularity. To understand these levels throughout the thesis, they will be briefly explained here. A collection of blacklists or TI reports is considered a *TI feed* level, and often discussed in the context of monthly or yearly statistics. Measurements of elements in a single report, the buildings blocks of a feed, are at the *report* level. If one is analyzing an individual measurement of a report, this is considered to be at the *object* level.

One important distinction has to be made when discussing these different kinds of TI, the difference between Open Threat Intelligence (OTI) and Paid Threat Intelligence (PTI). OTI is free, often consisting of blacklists that re-branded themselves to be TI (Li et al. [37] state IP address-oriented TI dates back from spam and intrusion detection blacklists). The fact that OTI is free results in content without context, often 'just' a periodically updated feed with indicators is provided. PTI on the other hand is, as the name suggests, not free. However, this does often mean that the quality and offering of these sources is better and more extensive. It is common for PTI vendors to include both a feed (without context) and reports, where threats are discussed more extensively. This extra information offers the possibility to not only block specific traffic, it allows the customer to understand why they have to block the traffic. Next to that, it enables customers to perform an historic analysis on previous attacks and threats. Finally, this contextual TI can be used to predict and prepare for future threats. This extra information comes with a price: PTI prices can range from 50.000$ to 650.000$ [12].

For this money, each of the TI vendors promises to provide 'unmatched' quality in different categories. Crowdstrike states they offer "unmatched prevention from the moment you deploy" [19], Cisco Systems offer "unmatched visibility" [47], BlueVoyant state they have "unmatched expertise" [11], and Juniper Networks state their "unmatched security intelligence detects and blocks advanced threats 'faster'" [48].

1.3. Current state of Threat Intelligence research

Despite high prices and rather big promises, it is not well-known what is included in PTI sources and which vendor's product actually offers the best quality. Different studies have been conducted to try and explain TI vendors' quality. Both by performing qualitative research [78], rating the different sources in a 'current offering' and 'market presence' dimension, but also in quantitative ways [12, 37]. However, research discussing PTI is limited, likely caused by its prohibitive pricing.

Next to a limited amount of research concerning the quality of TI vendors, there is no research evaluating individual TI reports. Most of the quantitative measurements as discussed in chapter 2, so-called metrics, are measured and presented on feed level. Metrics on report level are never discussed. The same is true for single report appreciation scores. As discussed, Zelonis [78] has held a questionnaire to report on the quality of TI feeds of different vendors. However, no published work discusses what makes individual reports worth reading.

The Bank of England [51] published the advice that acquired TI should be reviewed from time to time. Proposed ways to do so includes checking for timeliness and the number of false positives/negatives of proclaimed malicious indicators. This recommendation is in line with the trends visible in literature, where TI researchers continue introducing and using metrics to describe Threat Intelligence feeds. However, Bouwman et al. [12] showed that these quantitative metrics are not the only consideration of value for TI customers, as more than half of the TI customers do not use metrics to evaluate their feeds. In their evaluation of TI feeds, customers bring up different qualitative considerations, such as confidence, relevance, and actionability to be considered more relevant. These findings however, are on the feed level. There is no information on either the implicit or explicit evaluation of single reports, neither if such evaluations relate to quantitative metrics.

1.4. Threat Intelligence data sources

In this study, the added value of different quantitative metrics will be disentangled from their ambiguous use in literature and investigated. Thanks to access to two unique data sources, individual quantitative metrics can be compared to appreciation scores of Threat Intelligence users.

1.4.1. Commercial Threat Intelligence feeds

We have access to five different commercial Threat Intelligence feeds. This allows us to perform a comparison of quantitative metric values between more sources than seen before in literature, as well as compare these metric to the user appreciation scores. The vendors of the TI feeds used in this research can all be considered market leaders.

These TI feeds consist of PDF files combined with metadata files which enable automated parsing of the majority of a report's content.

As described, throughout this study different metrics will be calculated on the basis of the reports in these feeds. These metrics will represent some characteristics of the reports and of the feeds as a whole. Because of this, and because of the fact that these metrics will be used in the context of statistical and machine learning algorithms, these metrics are occasionally referred to as *features*. This is a doubly suitable term because of its definition as "a prominent part or characteristic"[22] and its interpretation in line with the machine learning jargon as "an individual measurable property or characteristic of a phenomenon being observed"[10].

1.4.2. User appreciation scores

The owner of the Threat Intelligence feeds also implemented a rating system for the TI reports. In this system, the user can read a TI report in their browser and is then presented with an option to rate the report between 1 to 5 stars. There is a random chance of 10% that the users are explicitly reminded to vote, the other 90% of the time the system relies on the habit of users to vote.

After an update, this rating system was expanded. The users were asked to explain their vote by the means of 9 dimensions, each dimension could be either present in a positive or in a negative way. This allowed users to add nuance to their vote and thus more explicitly indicate why a 4 star score was not enough for a 5 star score.

Throughout this study, these user appreciation scores will sometimes be referred to or interpreted as 'the value of Threat Intelligence'. Although these appreciation scores are a proxy for the real value of Threat Intelligence at best, it is the best possible representation of TI's value encountered in TI literature up until now. This means that, if we decide to accept appreciation scores as proxy for TI value, some biases of the reviewer population might be present and that all conclusions will relate to *perceived* quality within the group of report reviewers rather than actual value.

1.5. Contribution

When combining the current unknowns from literature with the access to the presented data sources, different questions and corresponding contributions arise.

Main research question Following the reasoning that user appreciation scores represent *perceived* quality of the Threat Intelligence sources of the reviewer population, we can use these scores to say something about the relation between Threat Intelligence value and quantitative metrics. When formalized, this question becomes:

0. Do Threat Intelligence Metrics from literature capture user appreciation?

In order to answer this main research question, we have to identify quantitative TI metrics from literature and relate these to the users' report appreciation scores. The steps how we will do this are written out in a set of sub-research questions.

Sub-research question 1 In order to compare quantitative metrics to user appreciation scores, we first need to extract these quantitative features from the reports. Threat Intelligence literature has been focused on these metrics, as they are the easiest and most explicit thing to measure from TI feeds. This means these metrics will be identified on the basis of a literature search in which we try to answer the following question:

1. Which quantitative features can be extracted from Threat Intelligence reports?

It has to be noted here that we do not mean to create a novel framework from these quantitative features, these features are at best a collection of predefined metrics from literature, collected for our purpose of relating Threat Intelligence reports to user appreciation scores.

Sub-research question 2 Before the metrics found from literature are compared to the user appreciation scores, it is important to first understand what the user appreciation scores mean. In order to do this, four short interviews are conducted in order to understand the considerations while voting. After these interviews, we aim to answer the question:

2. What are analysts' main considerations when scoring Threat Intelligence reports?

Sub-research question 3 Each reader that votes on a report is different. Next to personal differences, each of the readers has a different (study) background and different use-cases for the reports they are reading. In order to find out if these characteristics have an impact on the scoring behavior, an analysis will be performed to relate the readers' department, team, and role to the appreciation scores. After this analysis we should be able to answer the following question:

3. Which patterns can be distinguished in the appreciation scores?

Sub-research question 4 Having gathered both quantitative features from reports and the scores these reports received from the users, we can start to ask a question relating more closely to our main research question. In this part of the study, statistical methods will be used to test for significant relations between report metrics and their respective scores. After this part of the study, we should be able to answer the question:

4. What is the relation between quantitative metrics & user appreciation scores?

Sub-research question 5 Next to finding the statistical relation between report metrics and user appreciation scores, these two data sources can be combined for other purposes. Different methods will be applied to try and predict the amount of votes and the ratings of the reports. This will enable us to answer the following question:

5. Can user appreciation scores of Threat Intelligence reports be predicted using quantitative metrics?

1.5.1. Lessons learned from this study

During the process of gathering quantitative metrics, a large set of academic literature is reviewed and each metric will be evaluated in different dimensions. At the end of this process, a set of metrics is found that has not been used before in literature. Throughout this study, this set of metrics will be used, giving valuable information to readers about the added value of each metric.

Although the results of this study are only directly applicable for this unique combination of Threat Intelligence sources and the owner of this data, it can serve as a guideline and inspiration for all other Threat Intelligence owners that want to improve their Threat Intelligence evaluation pipeline.

1.6. Report structure

First, in chapter 2, a literature search is conducted to arrive at the report metrics that will be used throughout this study. How these features will be used exactly is elaborated upon in chapter 3, more information about the used data is provided here, as well as more technical details on how the analyses are performed. Then, in chapter 4, the results from the interviews are presented. These are presented here in order to offer a helping hand with interpreting the remainder of the results. The contents of the data-sets are visually presented in chapter 5 and the results from the analyses are presented in chapter 6. Then, some prototypes are constructed to show future possibilities with the data in chapter 7. Finally, all findings will be summarized in chapter 8 and limitations, economic considerations, and other critical thoughts will be reflected on in chapter 9.

2

Literature review: Metrics for comparing Threat Intelligence feeds

In this chapter, different Threat Intelligence (TI) metrics as found in literature will be discussed. For each metric, different interpretations from literature will be assessed and we will decide if and how this metric will be incorporated in this study. Their wide use in literature indicators that these metrics are assumed to be an excellent way to describe the contents of a TI report and TI feeds [27, 37, 42, 56, 58, 63].

Different pieces of related literature will be discussed, which will we mirror to four different dimensions to evaluate metrics on: numerical/categorical, unit of analysis, application on real data, and if a (artificial) ground truth is necessary.

To compare TI in an automated manner, the metrics used to compare need to be countable. This means they should be either numerical or categorical. Ideally, these metrics are not only compared in an automated manner, but also gathered that way. When automating the collection process is not possible, manual collection of metrics is also an option.

Next to metrics being either numerical or categorical, metrics can be counted at different levels. Schlette et al. [64] distinguish three levels when discussing the quality of TI data: report level, object level, and attribute level. Next to these three levels, we introduce the feed level.

Metrics at the report level only say something about an individual report. This means that when you count the number of indicators at the report level, your scope does not go beyond that report. The feed level does go beyond the report level and says something about a collection of reports. Metrics as feed level will often consist of metrics at report level, aggregated in a specific manner. The object level and the attribute level are too fine-grained for this study and thus will be disregarded. These two levels, the object level and the attribute level can be replaced by the indicator level. On the indicator level we decide if a single indicator is either a true/false positive for example, as a report contains multiple indicators and thus no single true/false positive value can be assigned.

We will consider two final facts while evaluating metrics from literature. The first is if they have been applied to data already in literature. We do this to get a sense how to metric should be implemented, as well as to find out the academic consensus about the metric. If the metrics are applied to data, it needs to be clear if that data consists of paid , shared, or open Threat Intelligence sources (PTI, STI, and OTI respectively). The second is if a ground truth or additional data is necessary to compare them. Some metrics depend on ground truth values in order to give a value to them. These are metrics such as accuracy, which needs to know the amount of false positive or coverage which needs the amount of true positives. In order to calculate true/false positives, you need a ground to be sure if a value is reported on correctly or not. Other metrics don't need a ground truth, but still need additional data to act as an 'unconfirmed ground truth'. This is the case for an alternative coverage metric, not needing actual true positives, but a set of data points by a customer organization. These additional data points can then be compared to the reported values in the TI reports.

The selected literature is limited to papers that claim to have come up with some sort of novel metrics or papers that have had an undeniable influence on these papers. Snowballing backwards from the papers

Table 2.1: Metric used in this study and their occurrence in literature.

	Sinha et al. [67]	Sheng et al. [66]	Kührer et al. [36]	Pinto and Maxwell [58]	Metcalf and Spring [43]	Pawliński and Kompanek [56]	Meier et al. [42]	Li et al. [37]	Griffioen et al. [27]	Noor et al. [50]	Schlette et al. [64]
Volume	x	x	x	x			x	x			
Timeliness			x		x	x	x	x	x		
Overlap					x			x	x		
Population				x	x						
Security threats						x				x	
Information type						x				x	
Subjectivity											x
Data type	OTI	OTI / STI	OTI / PTI	OTI	OTI	STI	OTI	OTI / PTI	OTI	PTI	Unknown

by Bouwman et al. [12] and Li et al. [37] resulted in the selected literature. Both papers presented a range of literature where metrics for their own studies have been based on. This snowballing from literature was complemented with searches on the bigger academic search engines and databases such as arXiv, Scopus, Web of Science, and Google Scholar, using search terms related to Threat Intelligence.

Based on the titles and abstracts of the found papers, an initial selection was made. This selection consisted of papers that seemed to have introduced at least one new metric. Fully reading these papers either confirmed or rejected this, papers that did not come up with new metrics were excluded as well. Finally, common sources of the papers that contributed new metrics were included. These common sources were included to show that no author did came up with all metrics them-self, but also to be able to show the gradual development of TI research. This gradual development is also shown well in figure 2.2, where common sources of used literature are shown in a graph structure.

2.1. Threat Intelligence metrics from literature

In this section, first 6 metrics will be introduced and discussed that are found in multiple sources. Their different definitions, interpretations, and naming conventions are laid out and I discuss their relevance for purposes of this study. After these 6 metrics, several metrics will be discussed that occurred in only one or two papers.

2.1.1. Volume

Volume describes the total amount of unique indicators that are included in a report and in a feed. As one of the most intuitive metrics, *volume* is described and used in many different pieces of research (six out of eleven papers as presented in table 2.1). Next to the term volume, this metric is also presented as *novelty* [58], *list counts* [43], *completeness*, and *size* [42]. The common use of this metric is to represent the amount of indicators measured in the whole feed or over a measurement interval. The prevalence of this metric is likely caused by the fact that volume is one of the first measurements you (in)explicitly perform. You check how much disk space a data-set takes or notice the size of the data-set during an exploratory analysis. This is in line with the human psyche, according to the availability heuristic [71] the estimated probability or frequency of an event is judged by the number of instances of it that can be readily be brought to mind.

The volume metric does not need a ground truth, as gathering this metric consists just of counting the (unique) indicators in a dataset, making it a numerical metric. Sinha et al. [67] and Sheng et al. [66] even go as far as not mentioning this metric explicitly, but just shortly mentioned the total volume of their dataset.

This metric can be easily applied to all kinds of Threat Intelligence and in literature it is applied to OTI, STI, and PTI (table 2.1).

Most literature tends to report volume as a feed-level indicator. This is done as most sources use feeds rather than reports, meaning no context is provided about the relation between indicators in that feed.

As this metric does not require any ground truth, counting can be easily automated, and can be applied to any kind of Threat Intelligence, **volume** will be included in this research. The difference with literature will be, that the different counts will be split up between the three main kinds of indicators: IPs, hashes, URLs, and domains. This split will be made because most blacklists do not combine the different kind of indicators, but TI reports do mix all relevant indicators in their written reports. This also means that next to reporting on this metric on feed level, this metric will also be measured on report level.

2.1.2. Accuracy

The accuracy metric relates to the amount of false positives in the data. This metric is well known within literature. Whereas Sinha et al. [67] uses this term for both the false positives as well as the false negatives, all other sources agree that accuracy is related to the amount of false positives [36, 37, 42, 56, 66]. Like volume, this metric is applied to OTI, STI, and PTI (table 2.1).

The decision if an indicator is a false positive will always be on indicator level, but the level of reporting for all sources is on feed level. Disregarding the method of Meier et al. [42], who use only available TI data, all other sources use some kind of ground truth. There are several methods to find out the amount of false positives. One of the most labor intensive ways is to manually label all indicators and check for false positives [66, 67]. Other methods include labeling parked domains and sinkholes as false positives [36], making a short and simple whitelist yourself with domains and IPs that shouldn't be in the feed [56], or using unroutable IPs, the top Alexa domains [1], and Content Distributor Networks to define the known goods [37].

Showing the false positives was for the sources mentioned above a large chunk of their presented research. For Kührer et al. [36] one of the main finding in their research was a novel way of determining parked domains and sinkholes. For Li et al. [37], months of data of unroutable IPs has been analyzed. This would mean that if the accuracy metric would be added to this research, similar data has to be gathered. Looking at the scope and timeline of this study, it has been decided that the accuracy metric will not be used.

2.1.3. Coverage

Coverage measures the minimum amount of true positives in your data. The coverage metric and the accuracy metric are very similar, where accuracy is concerned with false positives, coverage expresses true positives. This metric is called *completeness* by Kührer et al. [36]. This metric is applied to OTI, STI, and PTI (table 2.1).

Just as with accuracy, this metric is also decided upon at indicator level, but reported on at feed level. All methods described in literature do need a ground truth to calculate the amount of true positives. This is done by manual labelling [66, 67], using automatic detection algorithms of malware patterns [36], and using network telescope data to identify scanners [37, 72]. Each of these methods is still prone to either user or systematic errors.

Just as with accuracy, the identification of true positives was a significant part of the research in the found pieces of literature. Either a lot of manual labeling was performed, patterns of true positives were defined, or an analysis of a large dataset was performed. Again these kinds of work are out of scope and not suitable for the timeline of this study. However, when given a possible self-found dataset containing true positives, coverage can be calculated. This in turn is exactly executed as Schlette et al. [64] described with their *relevance* metric. Just as with volume, this metric will also be calculated at report level, as well as at feed level. Equation 2.1 shows how relevance can be calculated. Defining the coverage measure like this will likely limit the coverage to small numbers. This as the scope of the reports is likely a lot wider than self-detected data and it is likely self-detected data contains a lot of indicators that are not found by the publishers. Because we have no access to data necessary for this metric, this metric is not used in this research.

$$Relevance(report) = \frac{|Indicators_{report} \cap Indicators_{customer}|}{|Indicators_{report}|} \tag{2.1}$$

2.1.4. Timeliness

Timeliness is known under many different names: *reaction time* [36], *time series measurements* and the '*following*' relationship [43], *speed* [42], and split into *currency* and *volatility* by Schlette et al. [64]. The use of timeliness depends on the use of a ground truth. The only odd one out is the definition of Schlette et al.,

where timeliness regards the data quality and is split into currency and volatility. For currency, a higher value is better (more recent). A higher volatility means that the object that the indicator is a part of, changed more often. We will disregard these definitions for the rest of this section. This metric is applied to OTI, STI, and PTI (table 2.1).

This metric has been both categorically and numerically reported on. In the categorical way, in both cases it is a binary classification where between two TI sources the choice is that either one or the other source is generally faster, for both cases no ground truth is required [42, 43]. Metcalf and Spring [43] define it as a 'following' relationship, where if one source is consistently faster with reporting on indicators than an other source, they conclude that the slower source must be following the faster one.

In other papers, representing this metric numerically, both methods with and without ground truth are used. Without ground truth, indicators are compared to other acquired feeds [37, 43]. Other papers use different ground truths, such as the occurrence timestamp in SANDNET [36, 59], the first occurrence in netflow data [27], or the moment they first detected the indicator themselves [56]. As with accuracy and coverage, this metric is calculated at indicator level, but reported on at feed level.

To use this metric, the choice has to be made if a variant with or without ground truth will be used. This time the reason to not use an extensive ground truth is not the added length of the analysis, but the lack of the necessary data-sets (SANDNET, netflow). As alternative, when having access to in-house data, this could be used as ground truth. However, like with coverage, both the unavailability of in-house data and the fact that it does not give the guarantee of being faster than other TI sources, the most simple definition of **timeliness** will be used. This is the same as also used by Li et al. [37], where the relative delay of an indicator compared with other TI sources is used to represent its timeliness.

2.1.5. Overlap

The overlap metric is a broad metric with different interpretation and sub-metrics found in literature. Generally, no ground truth or extra data is used, except for the *expanded list intersection* as explained by Metcalf and Spring [43]. With this method, domain-indicator pairs are made and the feed is expanded by applying these pairs to the values in the feed. Then using this extended feed, normal overlap metrics are gathered.

These normal overlap metrics include *time series metrics*, *pairwise intersection counts*, and *reverse counts* by Metcalf and Spring [43], and *differential contribution* and *exclusive contribution* by Li et al. [37]. Finally, Griffioen et al. [27] use the term *originality* to describe overlap.

The time series metrics are the rote intersection (i.e. all intersections for two lists up to all lists in the group) size and the percentage of overlap relative to each of the two lists. The pairwise intersection count contains all the possible pairings of all TI sources and the cardinality (i.e. size of the set) of the intersection of the two sets is reported on. Reverse counts work the other way around and reports in how many feeds an indicator is found [43]. Differential contribution and exclusive contribution are very similar, the difference is that differential contribution compares one source to one other source, exclusive contribution compares that one source to all other source. In both cases the percentage of unique indicators that can not be found in the other source(s) is reported on [37].

The originality metric from Griffioen et al. is comparable to the exclusive contribution of Li et al.. The main difference is that they argue that originality can relate one feed compared to all others feeds, but can also be used to say something about a set of feeds.

Expanding the indicators should not be necessary, as paid Threat Intelligence should report on all relevant indicators for a report and finding relating IPs or domains does lie within the expected research a vendor is doing. This means one or multiple of the discussed overlap metrics can be chosen.

Two **overlap** metrics will be used: Pairwise intersection counts [43] and the exclusive contribution [37]. This gives the possibility to report on all overlapping values (pairwise intersection counts) and to report on the uniqueness (originality) of one feed or report(exclusive contribution).

2.1.6. Population

The population metric relates found indicators to geo-locations. For Metcalf and Spring [43], this is part of their *domain-based characterization* step. Both papers describing this metric use the Autonomous System Number (ASN) [43, 58] and Metcalf and Spring also used a GeoIP database. Gathered and resolved data in turn is used to give counts on the most occurring source and destination countries.

Given that PTI sources sometimes include source and target countries, applying GeoIP and ASN services to indicators could turn out to be redundant and thus will not be used. Data about source and target countries is sometimes supplemented by information about relevant actors. This means descriptive statistics about both the **country population** as well as the **actor population** can both be used in this research. Some publishers include information about **target industries** as well. This metric relates to actor and countries and thus will be seen as a sub-metric of population.

2.1.7. Other metrics

The TI metrics discussed so far are metrics that occur in multiple sources and thus can be seen as a 'literature standard'. The following metrics are less common but should still be discussed, as they do offer a more complete set of information regarding possible metrics gathered from TI.

Domain-based characterization The domain-based characterization of Metcalf and Spring [43] is a metric that consists of several sub-metrics. Next to resolving the active domains per blacklist as discussed in the accuracy section, they mention reporting on the top 5 name servers that serve the largest number of domains per blacklist.

As the conclusion of Metcalf and Spring [43] is clear, adversaries are not heavily reliant on single-name servers and the most common name servers tend to be common ones. This results in the fact that these end up in the top 5 most used name servers, as this is not interesting for this study, this metric will not be used.

Relevance The relevance metric is interpreted in two different ways. The one as introduced by Schlette et al. [64] and which is used for the coverage metric and the interpretation as by Pawliński and Kompanek [56]. The explanation of Pawliński and Kompanek is rather limited and consists of "Should we care?" They want to use analysts' queries in their own system to check which indicators are most relevant. These in turn can then be linked to the feeds they are mentioned in.

No similar data is available as the data proposed by Pawliński and Kompanek [56]. This means that even if their definition of relevance would be used, there would be no possibility to incorporate it in this study.

Sensitivity The sensitivity metric of Griffioen et al. [27] is heavily dependent on the Netflow data they used in their research. Using that data, they checked the amount of traffic to an malicious host before it eventually got added to a feed. Combining this information with the geo-location as found using a GeoIP database, they were also able to calculate a possible bias present against certain geo-locations.

Calculating the sensitivity of a feed would be a relevant metric to use in this research. It would give an insight when different TI sources start looking into a threat and how long their investigations last. Unfortunately, no similar data is available.

Impact The impact metric measures the consequences for users of a feed. This can both be positive (e.g. the connections to and from a malicious host are suppressed), as well as negative (e.g. the connections to and from a benevolent host are suppressed). This was checked by looking at how many domain names were pointed at the IP the day it was added to a feed and thus blocked when following the feed. In some cases this resulted in 900.000 domains being blocked, meaning there has to be a large amount of collateral damage in that case. To perform this research, an active domain/IP crawl was conducted by the authors [27].

The impact, as described by Griffioen et al. [27], strongly depends on the use of the TI sources. As the sources in this study are used for different purposes, the impact of an indicator being in a feed is less relevant.

Security Threats The **security threats** metric, as introduced by Noor et al. [50], are the different topics a TI vendor reports about. Different options are Advanced Persistent Threats (APT), Distributed Denial of Service (DDoS), phishing, malware, ransomware, web application attacks, drive-by downloads, credential compromise, data theft / manipulation / destruction, eavesdropping, or zero-day exploits. These categories are based on the Open Threat Taxonomy [69] and the ENISA Threat Taxonomy [41]. The security threats are relevant to gather per report, as this will show what a vendor publishes most about and to be able to evaluate how well they write about it. This metric was also proposed by Pawliński and Kompanek [56] under the name *vantage*, but never operationalized.

Information Type The different **information types** are indicators, tactics-techniques and procedures, customized software, tool configurations, TI reports, vulnerabilities, or security alerts [50]. Although the current data does not use these exact categories, TI reports can be divided into different report types. The challenges these report types introduce are discussed in chapter 3 and tackled in chapter 5.

Monthly/annual cost Introduced by Noor et al. [50] as formal metric, Bouwman et al. [12] already stated that subscriptions fees start around a couple of thousands of dollars per month up until around 650.000$ per year.

Subjectivity **Subjectivity** Most TI reports are constructed by humans in natural language during the analysis of an attack. This means that while making a TI report, different subjective nuances could slip in. This can be in the form of language (“I suppose"), but also in the form of bias against specific actors or source countries. Natural Language Processing and sentiment analysis can classify reporting as either subjective or objective and the direction of the subjectivity (positive or negative) [64].

Report counts When trying to relate metrics from reports to appreciation scores of these reports, metrics found in literature might not always be relevant. Metrics as proposed in literature are sometimes too involved, such that readers might not even notice or care about their values. Other metrics might be very specific to a vendor or to PTI in general, something that literature did not get the chance to evaluate.

Some simple additional metrics can immediately be identified, these are metrics such as the **amount of words**, **pages**, and **images** in a report. Although the amount generally should not be the most important to readers, it might be the case that reports that are too concise might not contain enough information for a reader. On the other side, in reports that are too long, key information might get buried under less relevant information. These metrics can also be seen as a kind of volume metric.

2.2. Observation during the literature review

This review has provided the following insights into the state of literature regarding both the evolution of the term ‘Threat Intelligence’, as well as the relation between the literature itself. Despite not being a major impact on field, these still are noteworthy developments.

2.2.1. The term ‘Threat Intelligence’

Literature shows a gradual change from the use of the term ‘blacklist’ to the broader term ‘(cyber) Threat Intelligence’. This means that the most recent works in this field all talk about Threat Intelligence whilst citing mostly (or only) blacklist-oriented literature.

Kührer et al. [36] and Metcalf and Spring [43] are the lasts ones in the citations of the used literature that talk about blacklists. Pinto and Maxwell [58] seem to be among the firsts to start naming it ‘Threat Intelligence’. This transfer period seems to overlap with the years that more Threat Intelligence reports started to get publicized by both commercial vendors and independent Threat Intelligence researchers, as shown in figure 2.1. Examples of such publicized reports are the report of Mandiant [40] concerning APT1 and the Dragonfly report by Symantec [68].

2.2.2. Found relations in Threat Intelligence literature

Relation between the different pieces of literature can be found in figure 2.2. This shows both that certain papers are gaining the reputation of common knowledge in the Threat Intelligence field [36, 66, 67], as well as that more recent papers do not always find each other [27, 37, 42]. This may explain the fact that Griffioen et al. [27] and Li et al. [37] came up with two different sets of metrics and Noor et al. [50] came up with a complete different approach and metrics.

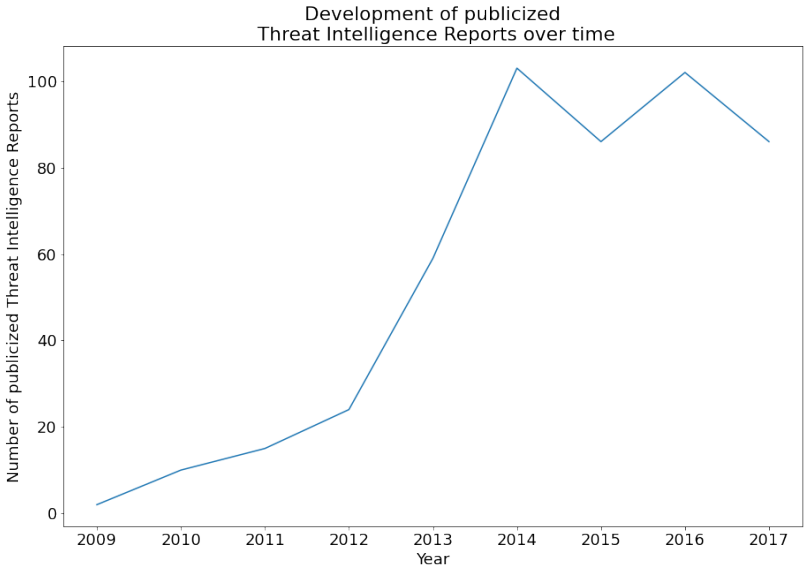


Figure 2.1: The development of the amount of publicized Threat Intelligence reports collected by 'kbandla' [31] over time. Starting in 2012, the sharp increase in publicized Threat Intelligence shows the term becoming more popular for cyber defense products.

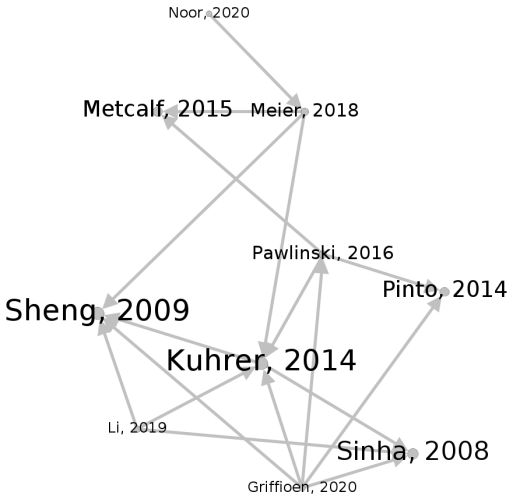


Figure 2.2: Citations between different pieces of literature in the review. Sheng et al. [66], Sinha et al. [67], and Kührer et al. [36] can be clearly distinguished as important and foundational pieces of Threat Intelligence literature.

2.3. Summary

In this chapter we looked at metrics concerning Threat Intelligence as found in literature. For each metric, we evaluated if the metric was suitable to use in this research, both by looking at which level of measurement the metric was, but also if that metric needed extra data or a ground truth. For all metrics it was kept in mind that they eventually were going to be related to the appreciation scores of individual reports. If any of the metrics was not obtainable at a report level, they were included to be used as descriptive statistics of the TI feed. All chosen metrics are also shown in table 2.1, this table shows all papers which mention that metric as well.

This section both summarized the results from this chapter, but also answers our first sub-research question:

1. Which quantitative features can be extracted from Threat Intelligence reports?

The following metrics are selected from literature and thus can be extracted from TI reports.

Volume Volume describes the total amount of unique indicators that are included in a report and in a feed. The indicators will be split into three categories: IPs, hashes, URLs, and domains. Each report will have these volumes connected to it, as well as that every feed will have the total volumes associated with it.

Timeliness Timeliness indicates how fast TI sources are with finding and reporting about indicators. This can both be done in a relative as well as a absolute manner. The relative way is where delay of one source compared to the date of the first time an indicator is published is taken as value. The absolute metric compares the publish date of a metric to a ground truth. This ground truth could be SANDNET or Netflow data, or in-house data. Given the fact that there is no access to any of these datasets, the relative timeliness will be used. Timeliness counts will be assessed at indicator level and will be aggregated at report level as well as feed level.

Overlap Overlap indicates the amount of common indicators between two reports or feeds. A complete image of overlap between all feeds can be shown in a contingency table. To show the unique contribution of one feed or report exclusive contribution is calculated. This means we both have unique contribution and overlap at report and feed level.

Population The population metric represents the source and target geo-locations. This can be extended with actor populations as these provide clues with the source geo-location. To gather this metric we will fully depend on the content of the reports and will not try to resolve IPs with help of GeoIP databases. This will also cover the metrics for actors and countries as proposed in section 2.1.7.

Security threats The security threats metric will try to measure the contents of the reports. This will enable an analysis to compare the appreciation scores with the topic that is discussed in a report, this could eventually show what the strong suit is of a TI source. With the security threats, the target industry could be included, this will both cover the industry as discussed in section 2.1.7, as well as enable a more granulated comparison of security threats.

Information type The information type will indicate the report type that is shared. Threat Intelligence sources sometimes perform an elaborated investigation or use give a short heads-up about the latest malware. By marking for each report in which category it falls, insight will be created in the focus of the source.

Subjectivity As discussed for the cost, the appreciation scores do represent *perceived quality*. Most of the mentioned metrics are objective and thus make it hard to catch the subtleties that are present in a report. Using the proposed metric by measuring sentiment and subjectivity, some of these subtleties could be quantified. Possible options to achieve this are the Textblob [38], flairNLP [9], and NLTK [4] libraries.

Report counts Finally, basic paper counts such as the amount of words, pages, and figures will be counted. Ideally, these metrics should not have any influence on a report's appreciation score, but it is important to validate that.

3

Methodology

3.1. Threat Intelligence reports and user appreciation scores

Two main data sources are used in this study. The first source consists of a collection of reports originating from different commercial Threat Intelligence feeds. The other source is a set of user appreciation scores for reports from these commercial Threat Intelligence feeds. A brief overview and explanation of the data sources will be provided here, descriptive statistics of the data sources are further elaborated upon in chapter 5. An illustration of the research process is shown in figure 3.1. This figure shows in which step of the study which data source will be used, it also shows the expected outcomes for each step.

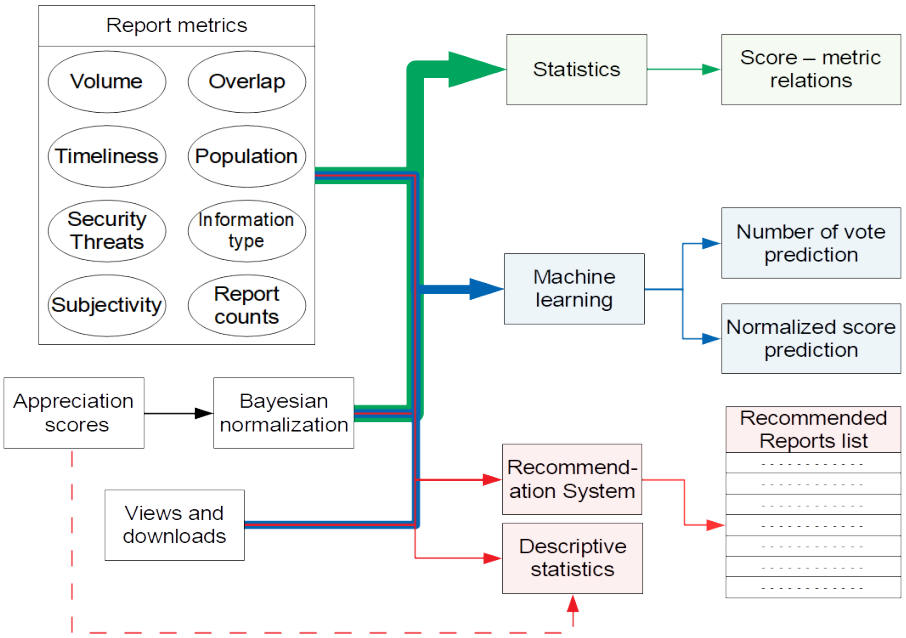


Figure 3.1: Systematic representation which data sources will be used in which research step and the expected results.

3.1.1. Commercial Threat Intelligence reports

The commercial Threat Intelligence feeds that are used, are from five of the higher-end vendors. As the starting date and exact contents of these feeds differ widely, we do not take any reports into account that are published the months and years before the first rating. This also enables an analysis into what percentage of reports are actually rated within the selected time period. Vendors typically offer around 5-10 different subsets of all their reports, each subset focused on specific topics or industries such as 'financial industry', 'cyber espionage', or 'cybercrime' [12]. For each of the described TI sources, around 3-5 different subsets are used.

Cost All information comes with a price, the cost for the commercial TI feeds is shown in table 3.1. Bouwman et al. show that none of the prices are set in stone. Two especially striking examples are the price ranges of Crowdstrike and FireEye as reported in their research. For both vendors, the lower boundary can be around 100.000€ and the upper boundary can go up to around 650.000€. These wide ranges are the direct result of the negotiations, where a successful negotiation can bring the price down by a factor ten [12]

Table 3.1: Annual prices of commercial Threat Intelligence feeds.

Vendor	Price in euro per year
Vendor 1	75.000 - 150.000
Vendor 2	150.000 - 250.000
Vendor 3	250.000 - 500.000
Vendor 4	75.000 - 150.000
Vendor 5	150.000 - 250.000

3.1.2. User appreciation scores

The owner of the TI feeds introduced a system that enabled all users of TI reports to rate the reports they read. A single appreciation score is a 1 to 5 star score, limited to a single score per report per user. The scores are collected in a time period ranging from 2020-08-12 until 2020-12-08, in this period a total of 623 scores are given by a total of 50 - 100 users.

On 2020-11-20, the rating system got an update. The idea that users could rate reports with a 1 to 5 star score was not changed, the change was that users now could choose up to 9 possible dimensions in which the report either excelled or lacked. These categories were: currency, context, depth, correctness, readability, relevance, technical, applicability, and uniqueness.[1]

The users that cast the votes are all part of different departments and teams, occupying different roles. Accompanying a different team culture and background as necessary for a role, different scoring behavior can be expected. In order to get a better idea of who is scoring, each user is labeled with their respective department, team, and role. The amount of people per organizational unit is shown in table 3.2. The characterization and responsibilities for these departments, teams, and roles are as follows:

Department characterization There are different departments within the organization. A difference in departments can both mean different responsibilities, but it can also mean similar responsibilities but different organizational units.

- **Department 1 (D1)** This department is mainly an umbrella department, existing of many higher level employees but also some technical ones. Theoretically there isn't too much relevant information information in the reports for them, other than contextual information about the field they are working in. This also shows in the frequency of their scoring.
- **Department 2 (D2)** These are general staff members and support staff. These people are responsible for the general and business processes within the organization.
- **Department 3 (D3)** This department mainly exists of people interested in higher level information and Threat Intelligence analysts.
- **Department 4 (D4)** This department does similar work as department 3, it is a different department because it is a different organizational unit.
- **Department 5 (D5)** This department consists of account-managers.
- **Department 6 (D6)** This department is likely to focus on the technical details in the TI reports.

[1] This update was partially based on the initial advice that was constructed after evaluating two months worth of data, this initial advice is presented in appendix A. The advice did only include three suggested categories, the final nine categories were constructed in consultation with the author. An other part of this update was a bug fix, a bug that caused not all ratings to be registered correctly. Recorded ratings in the time period up until the update (2020-08-12 until 2020-11-17) are correct, only not complete. When drawing conclusions from the data, this will be kept in mind accordingly, as discussed in chapter 5 as well.

Team characterization Teams are always part of a single department, they are only dis-aggregated in order to specialize on a specific topic within the overarching mission of the department.

- **Department 1, Team 1** This is a subset of department 1, they are mainly interested in technical details in the reports.
- **Department 3, Team 1, 2, 3, 4** A dis-aggregation of department 3, they are likely interested in higher level information.
- **Department 4, Team 1, 2, 3(D4T1, D4T2, D4T3)** This is a dis-aggregation of department 4, they are likely interested in higher level information.
- **Department 6, Team 1, 2, 3(D6T1, D6T2, D6T3)** This is a dis-aggregation of department 6, they are likely interested in the technical details in the reports.

Role characterization Knowing the information flow is important for interpreting the different roles. Keep in mind here that high, medium, and low relate to the technical depth of the information that is handled and do not convey any hierarchical dynamic. This is also true for the use of support-employee and manager. These terms convey their logistical role in the process, rather than implying anything regarding their capabilities. A very coarse description of the information flow is as follows:

High-level analysts aggregate all information they receive from middle-level analysts into the highest-level report. This report is (almost) completely stripped from technical information and ready for the higher level managers. The middle-level analysts are the bridge between these high-level analysts and the low-level analysts. They dig into specific topics in order to inform the high-level analysts. When technical artifacts show up in these investigations, these are analyzed in-depth by the low-level analysts.

The support employees and the managers are at the sideline of this information flow. The support employees gathers relevant information and reports in order to organize them for easier use. Theoretically, support employees should not score reports, as their input in the information flow is not substantive. However, the fact that there are many votes from support employees suggests they are a kind of 'score aggregators' for a whole department.

- **High-level analyst** The highest level aggregator of information, receiving reports from middle-level analysts.
- **Middle-level analyst** The bridge between the low-level analysts and the high-level analysts.
- **Low-level analyst** The 'nuts and bolts' readers. Likely to be most interested in very technical reports.
- **Support employee** Organizers of information. Their role is to organize reports under the right internal labels, such that relevant teams do not miss any information.
- **Manager** The manager of a team (not a department).

Table 3.2: Basic counts per department, team, and role for all ratings.

Department	Voters	# of votes
Department 1	4	6
Department 2	1	3
Department 3	30	415
Department 4	16	50
Department 5	1	4
Department 6	30	144

Team	Voters	# of votes
D1T1	4	6
D3T1	13	108
D3T2	6	55
D3T3	10	199
D3T4	2	57
D4T1	4	11
D4T2	1	7
D4T3	14	37
D6T1	6	19
D6T2	21	122
D6T3	1	1

Role	Voters	# of votes
Low-level analyst 1	30	141
Low-level analyst 2	1	1
Middle-level analyst 1	37	277
Middle-level analyst 2	2	3
High-level analyst	35	42
Support employee	6	153
Manager	1	5

3.2. Dealing with lacking and skewed appreciation scores

Figure 3.2 shows that most reports are only rated once or twice and that more than five ratings per report is exceptional. For each report with only a single vote, reliability of the average report rating is low as you can't be sure if the score is saying something about the report or about the individual voter. In order to account for this reliability problem, you can apply a Bayesian rating calculation [70]. In a Bayesian estimation of the average score the average is influenced by a prior with a certain weight, effectively bringing down the impact of single vote. Three of such Bayesian rating formulas are found, one as used by IMDb for their top 250 movies of all time list [76], the other introduced by Miller [44], and the last one as presented by Chiang [16]

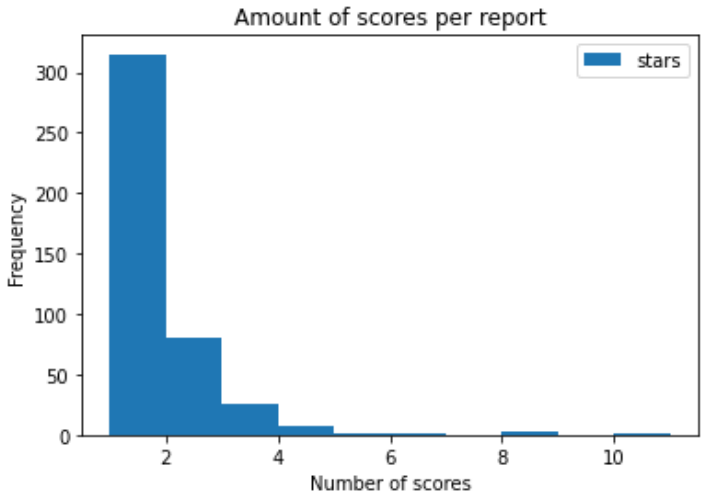


Figure 3.2: Distribution of the number of appreciation scores per report. This figure shows that most reports receive only one or two votes, raising questions if the mean rating should be used to represent a report's (perceived) quality.

The formula as used by IMDb and its simplification is shown in equation 3.1. C is the mean value of all votes casted, m is the smoothing parameter (the minimum amount of votes necessary to be included in the list in the case of IMDb), v is the number of votes for the item and R is average score for the item. The figure at the center top of figure 3.3 shows the results for an m of 1.

$$WR = \frac{v}{v+m} * R + \frac{m}{v+m} * C = \frac{Cm+Rv}{m+v} \tag{3.1}$$

This equation clearly seems derived from a known equation as described by Chiang [16], shown in equation 3.2. The meaning of C, v, and R are the same in both equations. The N variable is the count for the total amount of votes casted, meaning that if m takes the value of N, equation 3.1 and equation 3.2 are the same. Effectively the N is the weight of the prior and C is the value of the prior. This means that by changing the value of N, the amount of votes necessary to skew the predicted value can be altered. By changing C, the base value can be set. Experiments with these values of N and C are shown in figure 3.3. The originally suggested values of N as the total amount of votes and R as the mean of these votes is the bottom left figure. In the figure in the bottom center, the N is set to the average amount of votes per report and the R is still the average rating over all the reports. In the bottom right figure, the N is the average amount of votes per report as well, but the R set to 3 (mean of possible scores 1 to 5).

$$WR = \frac{CN+Rv}{N+v} \tag{3.2}$$

The final formula (3.3), as introduced by Miller [44], is more involved. The goal of this calculation is to report on a minimum value for the score with 95% confidence. The fact that this is a minimum value means that overall the estimation is on the lower side, as also shown in top right histogram in figure 3.3.

$$WR = \sum_{k=1}^{K} s_k \frac{n_k+1}{N+K} - z_{\alpha/2}\sqrt{\left(\left(\sum_{k=1}^{K} s_k^2 \frac{n_k+1}{N+K}\right) - \left(\sum_{k=1}^{K} s_k \frac{n_k+1}{N+K}\right)^2\right) / (N+K+1} \tag{3.3}$$

Figure 3.3 shows the distribution of the actual means per report and all predicted means for each of the Bayesian estimators.

At first sight, the algorithms from both Miller and the standard one from Chiang can be disregarded. The algorithm of Miller results in severely underestimated values, this is also mentioned on his website, where he mentions that for an item with consensus about its scoring, still at least 9 votes are necessary, something that only two reports have in our data-set. The standard algorithm from Chiang results in all reports being estimated on a score equal to the prior, as the weight of the prior (the total amount of votes) is a lot higher than any of the reports received.

Then the IMDb algorithm and the algorithm from Chiang with altered variables remain. The IMDb algorithm and equation 3.2 with a prior weight (N) of the average amount of votes per report result in nearly the same distribution of predicted ratings. This is because the smoothing parameter of the IMDb algorithm is set to 1, and the average amount of votes per report (1.4) is very close to 1. This means that reports with only a single 5-star vote, still receives a score of nearly 5 (4.66 to be exact). Receiving nearly the highest score with a single vote still doesn't seem reliable.

The final option is to use equation 3.2 with a prior weight (N) of the average amount of votes per report and a prior of 3. This results in reports with a single 5-star score to receive score of 3.8 and a report with six 5-star votes to receive 4.6 stars. A more extensive overview of how different rating techniques result in different scores is shown in table 3.3.

By using Bayesian scoring to calculate an alternative mean for the star scores of the reports, the amount of ratings and thus the level of uncertainty is incorporated in the new star rating. Instead of having to validate the amount of ratings a report got before making statements about the star score, hopefully this enables to draw conclusions directly.

Table 3.3: The effect of different Bayesian scoring algorithms on different scores. The table that shows the IMDb normalization and the algorithm from Chiang where the prior equals the mean seem to overestimate the scores. The algorithm from Miller seems to severely underestimate the scores. The algorithm by Chiang with the suggested prior results in only a single normalized score. Finally, the algorithm by Chiang with a prior of 3 results in a seemingly fair distribution of scores.

Ratings	Actual mean	IMDb	Evan Miller	Chiang Original	Chiang prior = mean	**Chiang prior = 3**
6	5.0	4.9	3.4	4.3	4.9	4.6
11	4.8	4.8	3.8	4.3	4.8	4.6
4	5.0	4.9	3.1	4.3	4.8	4.5
3	5.0	4.8	2.9	4.3	4.8	4.4
4	4.8	4.7	3.0	4.3	4.6	4.3
2	5.0	4.8	2.7	4.3	4.7	4.2
2	4.5	4.4	2.6	4.3	4.4	3.9
1	5.0	4.7	2.4	4.3	4.6	3.8
1	4.0	4.2	2.3	4.3	4.2	3.4
1	3.0	3.7	2.2	4.3	3.8	3.0
1	2.0	3.2	2.0	4.3	3.4	2.6
1	1.0	2.7	1.7	4.3	2.9	2.2
2	1.0	2.1	1.6	4.3	2.4	1.8

3.3. Gathering the quantitative report metrics

Basic PDF statistics, such as the amount of pages and figures, were gathered using PyPDF2 [46]. Text was extracted from the report PDFs using Textract [20]. However, not all report features were as straightforward to extract. For each of the more involved features, a description will follow.

3.3.1. Counting figures in a PDF

Counting figures in a PDF file is less straightforward than it sounds. Using the PDF reader offered by PyPDF2, image objects can be identified in a PDF. However, there tend to be multiple image artifacts in a PDF file, causing this functionality to severely overestimate the amount of figures in a PDF. The following measures are taken to prevent this:

- A minimum amount of pixels for the height and width is set, this excludes small artifacts that were likely

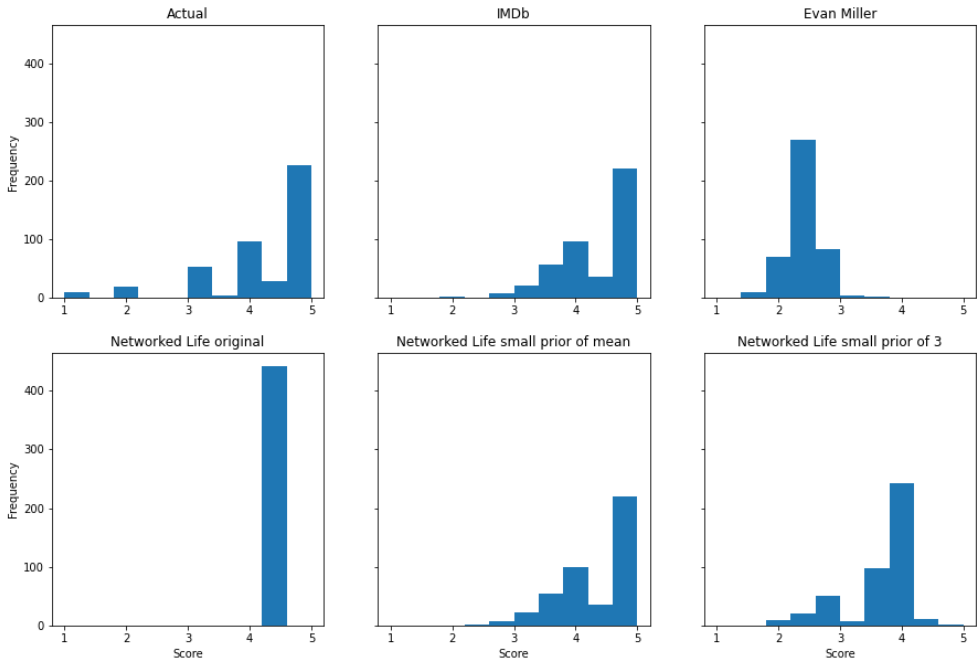


Figure 3.3: Results for three different Bayesian rating estimation techniques. One technique as used for the Imdb top 250 movies of all time list [76], one technique as introduced by Miller [44], the last one is introduced in Networked Life by Chiang [16].

invisible to the reader. For both the height and the width this is set to 100 pixels.

- A maximum ratio of of 1:15 is set for how the height and width can relate to each other. This excludes most figures that are added as header bars or as formatting.
- For figures that could be read by the library using the 'flate decoder', the entropy is calculated. The minimum of entropy is set to 0.1, which is often already reached by a figure existing of multiple colors.

3.3.2. Sentiment and objectivity analysis

Sentiment analysis was performed using TextBlob [38]. In the background, TextBlob uses a Naive Bayes Classifier, trained on the movie review corpus of NLTK [9], which in turn uses the movie review data from Pang and Lee [55].

A Naive Bayes classifier calculates probabilities of a word being positive/negative and the probability of a document being positive/negative as a whole to calculate a document's overall score using Bayes Theorem. The reason for using the movie review corpus rather than a more specialized one is simple: there are not that many labeled data-sets for sentiment analysis. Most data-sets contain reviews about consumer products or entertainment. There is one data-set with peer reviews of academic papers for computing and informatics conferences, however, it contains only 405 reviews which are mostly in Spanish [32]. Unfortunately this means that this data-set is also not suitable to train our sentiment analysis model.

How this implementation works, is that for all (word) tokens in a document the polarity and subjectivity for each individual token is returned. The polarity and objectivity for a document is simply the sum of all tokens.

3.3.3. Parsing Indicators of Compromise

Indicators of Compromise (IoCs) is the official term used for pieces of technical information found when analyzing cyber attacks. IoCs are extracted in two ways from the TI reports. The first way is to use the metadata that is provided along the PDF of the report. Each publisher does have its own standard for formatting these metadata files, not always being complete either. To compensate for this, the text in the PDFs is also used

to extract indicators from. To do this, a publicly available IoC parser is used, this tool matches common patterns of different indicators to the text in the PDF [5]. The results from both techniques are combined to have a complete as possible set of indicators for each report.

As this IoC parser is not context-aware, it presents the danger of adding many false-positives to the dataset. In order to account for this, IoCs added via the IoC parser are labeled likewise. This enables us to distinguish between IoCs gathered from the metadata and from the IoC parser later in the study.

Near the end of this study, we discovered that the metadata files for two of the publishers were not complete or not working properly. For this reason, most, if not all of the IoCs for these vendors are extracted using the IoC parser. This means that IoCs that are only published in their metadata files might not be included and false-positives can be present.

3.3.4. Parsing report types

Each publisher uses both a different classification of their reports, as well as very different criteria for a report to belong to a certain category. In section 5.2.6, a mapping will be constructed to relate different categories of different vendors to one overarching category.

Different report types that can be readily distinguished are periodic reports such as weeklies and monthlies, higher-level summaries, forecasting reports, and general reports describing a specific threat or attack.

3.3.5. Parsing security threats

The security threats as introduced by Noor et al. [50] are used to label reports with the topics they are covering. These categories are: Advanced Persistent Threat (APT), Distributed Denial of Service (DDOS), Phishing, Malware, Ransomware, Web Application Attacks, Drive-by Downloads, Credential Compromise, Data theft/manipulation/destruction, Eavesdropping, and Zero-day exploits,

Matching these topics to the reports will be executed with a simple string match. For the sake of completeness, the complete string, as well as different permutations (e.g. abbreviations, typos) will be used to match the topic on.

As this study was conducted in the second half of 2020, Covid-19 was still roaming the earth freely. This was a topic that came up in several TI reports as well. As the discussed threat types are not mutually exclusive, several strings to match Covid-19/Corona were also added.

3.3.6. Parsing source countries and actors

Each TI vendor uses different nicknames for (nation state) cyber threat actors. Florian Roth, a rather well-known cybersecurity researcher, started the effort to normalize all these different nicknames to one common name for each threat actor [60]. In this study, this sheet will be used to normalize the mentioned threat actors. Despite the fact that most mentioned threat actors are not present in the table, the highest amount of possible overlapping threat actors is captured this way.

3.4. Calculating the relation between report metrics and appreciation scores

Before conducting any modeling of the scoring behavior, it is important to know if we can find any important relationships between the report features and their respective ratings. In order to make this comparison, the average group score of a report containing a certain feature has to be compared to the average group score lacking the feature. A binary problem like this is usually performed using a t-test. When comparing different options within one category, meaning you want to compare three or more means, an ANOVA (ANalysis Of VAriance) is conducted. These two statistical tests, t-test and ANOVA, can only be performed given the response variable for each category is normally distributed and all categories have equal variance. As this is not the case, we will look at non-parametric alternatives of these tests.

The non-parametric alternative to an ANOVA is the Kruskal-Wallis test. This test replaces the data with their rank rather than the actual value, this decreases the effect of outliers in the data. The non-parametric alternative for the t-test is the Mann-Whitney U test. This test works the same as the Kruskal-Wallis test, but to compare two means (like the t-test). This means the Kruskal-Wallis test will be used to see if a significant difference in means exist between three or more categories and the Mann Whitney U test will be used as post-hoc method to test which of the means actually differ. As advised, we adhere to a minimum of five items per tested category [34, 65].

Performing a post-hoc method like this increases the total amount of tests that is performed, effectively

increasing the random chance that a significant result is found. In order to account for this, different correction methods exist, one of which is the Bonferroni correction [28], applied to each individual Mann Whitney U test. This correction divides the alpha value (p-value chosen to show significance) by the total amount of tests that are conducted. This means that when 5 post-hoc tests are performed, the alpha value will be (0.05/5) 0.01. The need for this procedure has been an ongoing discussion for more than 20 years [15, 57, 73], especially focusing on the conservative nature of this correction. A slightly less conservative alternative to the Bonferroni correction is the Bonferroni-Holm correction (also knows as the Holm method) [29]. This correction is largely based on the Bonferroni method, but instead of dividing each alpha value through the same divisor, you divide it through the p-value's rank. This means that if you perform 5 post-hoc tests, the alpha-value for the lowest p-value (rank 5) is divided by 5 to get the alpha value of 0.01. However, the alpha value related to the second lowest p-value is then divided by 4 to come to an alpha value of 0.0125, etc. Because of this slightly less conservative nature, the Holm method is chosen in this study.

4

Interviews: The meaning of a vote

In order to understand the scoring behavior of the readers, we decided to pick a small sample of scorers and ask questions regarding their scoring behavior. The used interview script can be found in appendix B.

The selection process for participants was based on both the amount of ratings a person casted, the mean of their ratings, as well as their department and role. The goal was to talk to two technical readers from department 6 and to two more broadly oriented readers from department 3. Then from each department, two people who scored a bit lower than average and two people who scored higher (than average) were selected. Finally, depending on the availability and responses, only one individual per department/score combination was interviewed (clarification shown in table 4.1). All of these individuals casted on average more votes than peers from their department.

Table 4.1: Contacted (votes, average score) and **interviewed** people per department and voting average.

	Department 3	Department 6
High scorer	(67 votes, 4.59 mean), **(54 votes, 4.52 mean)**	**(26 votes, 3.76 mean)**, (6 votes, 4.8 mean)
Lower scorer	**(24 votes, 3.13 mean)**, (7 votes, 4 mean)	**(16 votes, 3.31 mean)**, (4 votes, 2.25 mean)

As mentioned, these participants were chosen based on different attributes. When looking at the questions in appendix B, you could argue this set of participants is neither representative, nor the right subset to discover why reports are most often not rated. However, this set of participants was suitable to discover different motivations for different star scores, as both lower and higher rating participants were asked the question why they give lower and higher scores. This means that these participants are at least able to partially answer all questions, whereas people who (nearly) did not vote, would not be able to explain why they give certain scores.

Using the results from the interviews, interpretation guidelines for the remainder of this study are set up. These are necessary as a star score means something else to everyone and these guidelines should help having a more uniform interpretation of the results.

4.1. The process of reading and voting

In order to interpret the following sections better, it is important to understand how the Threat Intelligence reports are accessed and used.

There are two different ways readers encounter (new) TI reports. The first is the passive route, in this case a reader 'subscribes' to subjects of interest and receives a notification if a report is published regarding this topic. The other, more active option, is for a reader to search for keywords of interest. Rather than notifications dripping in with several-hour intervals of only the newest reports, the reader is presented with a list of all reports that have been published and include this keyword. This in turn leads to the fact that not everyone reads the same reports. People only encounter reports that are at least to some extent relevant for them, otherwise the looked-for keyword would not be present in the TI report.

Then, being faced with either one or a multitude of documents, it is for the reader to decide on a reading strategy. Depending on their goals, a reader decides to fully read through a report, skim the report and only read a relevant paragraph, or maybe even apply the search functionality (Ctrl + F) to look for the keyword of interest and only read the direct context that it is used in.

Consequently, when the reader either closes the document satisfied or wondering why this document came up in his mailbox or search, there is the option to cast a 1 to 5 star rating. These star ratings are still prone to different interpretations of the readers, being interpreted as the quality, relevance, or other scoring interpretations.

These different ways of reading and scoring indicate that a single rating can come about in many different ways. Some of the uses and interpretations will be discussed in the following sections.

4.1.1. The meaning of a vote versus a non-vote

As will be discussed in chapter 5 and figure 5.5, most reports are viewed at least once and not scored. This distribution of more non-votes than votes means that a large source of possible information remains untapped. To try and understand this behavior better and possibly use in the future, during the interviews the question was asked why the interviewees scored and why they sometimes chose not to score.

The list of reasons for scoring are much shorter than the list of reasons not to score. The reasons for scoring were simple: the interviewees tried to score most reports and had the attitude that if they really read a report, they would score it. The reasons not to score can be split in three categories:

Paralysis by analysis This is an umbrella term to cover three separate reasons:

- **Lack of time** When you want to give a good and fair rating, you need time to evaluate what you thought of the report. When you are in a rush and you do not want to pollute the rating system, you can make the choice to not rate rather than give an unfair rating.
- **Scored too much** When reading multiple reports in a short time span, it is likely you encounter reports of different quality. After scoring some reports, the mental effort it costs to evaluate what you thought of the report might outweigh the benefit you see from scoring a report.
- **Did not actually read** When investigating something, one technique of looking for information is opening the top X reports that hit on the keywords you looked for. Then after having read only part of these reports, you might have found the information you were looking for and decide to stop searching. This will give the reports a 'opened at' time stamp, but logically no rating.

Lack of relevancy There are many possible reasons you can think of for opening a report that you were not actually looking for. Maybe the used search terms were not specific or you expected something technical and it was a more contextual report. In this case, the choice has to be made if the opened report should get a low score because it is not relevant for the research you are conducting, or you decide not to rate the report because the reason this irrelevant report is in front of you might be yourself. The interviewees generally did tend towards the latter. If a report was irrelevant for an interviewee, they did not tend to rate it low but to skip voting altogether. This because the report could be relevant for a colleague and they did not want to pollute the system. They stated they expected this same behavior from their colleagues. In light of how the rating system is used, cautious voting behavior like this seems optimal. As most reports receive a small amount of votes, people with a little subject-expertise could very well skew the rating of a report, possibly unjustified.

Just forgot The most relatable reason for not rating is that you sometimes just forget to do so. In a day-to-day work setting, these reports, and especially the ratings, are a means rather than an end. This means that when you are researching something and quickly need a piece of information, you want to continue your research as fast as possible and thus it is easy to forget the five stars tucked away at the bottom of the screen.

This list contains different explanations, offering some help interpreting the non-votes. However, it must be noted that the interviewees were selected based on the fact that they were among the more frequent raters overall. This means that no insight was gathered about the reasoning of persistent non-voters.

Both a more diverse and bigger group of interviewees, but also even more directed questions towards non-voting are necessary to gain complete insight in the meaning of a non-vote. The reason why this is so

important, is that choosing to be silent might contain just as much meaning as the reason to vote. Comparing this to the amount of downloads/views that did not get converted in a vote, a huge amount of potential information can be uncovered.

4.1.2. The meaning of 1-5 star appreciation scores

Another question was: “Can you think of a report that you rated high/low lately, why was that?". This resulted in explanations that reports with high ratings often are about current events and offer content and depth, meaning the report discusses all topics it brings up extensively. One of the interviewees summarized it well by saying: “If it takes work out of my hands, the report deserves a higher rating". Answers to these same questions, but regarding lower ratings include reasons such as a report not containing enough context or not discussing a topic deep enough. A whole category that seems to be disliked overall are summarizing reports, where a vendor displays an overview of the reports they published with a few lines of summary. These overviews tend to hit on a lot of keywords but offer no real content.

When being asked more general points the interviewees (dis)liked from reports, the only new points that came up were positive. These included that TI reports ideally offer new leads for investigations, new and technical Indicators of Compromise, or extensive and in-depth descriptions of a single topic.

4.2. User predictions and recommendations

Two of the interviewees did give a prediction for the future, the prediction was that as time goes on, the amount of ratings might go down again. They elaborated that rating is a non-zero effort without any immediate positive consequences or feedback. In the long run, this could lead to a diminishing amount of ratings.

While being a valid argument and prediction, the importance of the ratings and the different use cases they offer as described in chapter 7 has to be stressed here. A relatively short-term use of the ratings is the proposed recommendation system. One of the interviewees even explicitly mentioned this would be of added value, as seeing relevant reports to the report you are reading decreases the amount of searching you have to perform yourself.

A longer-term but just as important effect, is the added value the ratings give from a negotiation perspective, as discussed in chapter 8 as well. If there ever are budget cuts, it is invaluable to know the priority of each different TI vendor. This could take the form of negotiating the price downwards or to stop acquiring one or two different sources. In times of budget surplus, negotiations with new TI vendors can be very targeted at stronger and weaker points, and after only a single month of use, the benefit of the new vendor can be evaluated. These longer-term effects of ratings are, as brought up, not immediately noticeable for the users, but in the long-term will assure the best available reports for the readers to use.

One of the interviewees also came up with some suggestions to enhance the system in which the reports are presented. The first suggestion is to incorporate a report’s rating into the representation of that report. This could be operationalized by either ordering the reports based on their (normalized) score or by informing the user of the report‘s of the reports perceived quality. The former introduces a selection bias where reports that already received a score will likely be presented above reports without a score. The latter introduces a problem where the user is primed with the presented number and might let themselves be influenced by this number whilst scoring the report.

The second suggestion was to group/recommend reports based on similarity in characteristics. Reports discussing the same topic are likely relevant to the reader. A possible interpretation of a system like this is implemented in chapter 7 by the means of a recommendation system. However, a simpler implementation could exist of extracting report features and showing the most viewed / highest rated reports with the same features.

4.3. Shortcomings of Threat Intelligence

One additional ‘interview’ was conducted, focusing on the current shortcomings of TI for specific use-cases. As it was not a prepared interview and none of the prepared questions from appendix B were asked, no outcomes regarding the reasoning behind ratings were gathered. However, the interview brought to light different shortcomings of the current state of Threat Intelligence and can possibly help researchers in the future distinguish between ‘re-branded’ blacklists and a newly defined form of TI.

One of the main concerns, as also brought up by Oosthoek and Doerr [54], is that the endless lists of filenames and hashes were written off more than a decade ago by anti-virus vendors, but are reintroduced by

TI vendors. It has been argued that these kind of indicators (file hashes) are very low in the pyramid of pain [8], meaning no cyber attacker will feel a big impact when these indicators are published and will be able to quickly circumvent detection rules made for these specific indicators.

Many TI reports are set up around a piece of malware and then discuss in-depth how that piece of malware operates, all the way to the specific API calls the malware makes. Sometimes additional pieces of information are given such as the Mitre ATT&CK mappings [18], used techniques, or information about the TTP. However, this is often a sporadic addition and not consistent nor structured. The fact that the reports are distributed as PDFs accompanied by a metadata file of varying quality, the right contextual data is not always present.

Ideally, TI vendors will choose a common schema to manage the information in their reports and distribute their reports according to this schema. When comparing the list of STIX participants and sponsors[1] to the list of the largest TI vendors [78], there is not a lot of overlap. These inconsistencies create a whole market for companies making structured information of the unstructured way it is delivered, such as EclecticIQ ([23]) and Elemendar ([2]).

4.4. Summary

In this chapter we tried to find an answer to our second research question by the means of different interviews with the users of Threat Intelligence reports. The question we tried to answer is:

2. What are analysts' main considerations when scoring Threat Intelligence reports?

The meaning of different star ratings to different analysts was explored in this chapter. One of the most descriptive explanations for higher ratings was: "If it takes work out of my hands, the report deserves a high(er) rating". Reasons for lower ratings mainly were a lack of depth and context, often resulting in analysts considering short heads-ups and weekly summaries to be less useful.

[1] `https://www.oasis-open.org/member/roster/`

5

Data description: What is in the appreciation scores and the reports?

As described in chapter 3, this study uses two main data sources. In this chapter, we will show descriptive statistics for both data sources. Section 5.1 describes the gathered star ratings. Distribution of ratings over departments, teams, and roles are shown, as well as information about the manner of report consumption and related ratings. Section 5.2 describes statistics from the reports, which are in line with the selected metrics from chapter 2.

In order to answer the third sub-research question, we first need to define which patterns we find interesting and expect to find. We distinguish three kinds of dis-aggregations: departments, teams, and roles. The exact explanation of each different category is explained in chapter 3, for now it is important to be aware that these three employee characteristics can be dis-aggregated. Each of the patterns we will try to observe, will be mirrored to each one of these three characteristics.

First, for each characteristic we will investigate if we can distinguish a difference in frequency of voting. Do we notice one group of people to vote on a higher amount of reports than another group? Next to this, can we distinguish differences in the value of the ratings they give. Do employees with one characteristic vote more harshly than employees with other characteristics.

Then, next to the apparent impact of employee characteristics, do we notice other behavioral traits that we can link to voting behavior. The most important one that will be looked into, is the time readers spend looking at a report before casting a vote.

5.1. How do people vote?

Two different collections of metrics regarding the star ratings will be shown in this section. This is caused by the fact that the system to gather the star ratings had an internal bug, causing some ratings not to be registered. The session identifier of readers that did not interact with the main website for more than five minutes expired, causing the vote the vanish and not to be linked to the actual user. This bug was present from the start the rating system was implemented (August 12th) and lasted until November 17. After this bug was solved, data was collected for about two additional weeks, until December 8.

In this second period, on November 20, an additional feature was added to the rating system. This feature consisted of buttons to argue for the given star score in 9 dimensions (currency, context, depth, correctness, readability, relevance, technical, applicability, and uniqueness) as partially presented in the initial advice shown in appendix A. As this new feature was added three days after the bug was fixed, the latter date will be used as cut-off point between the two collections of descriptive statistics.

5.1.1. Appreciation scores November 20 - December 8

The star ratings will be used to illustrate general voting behavior. The numbers from this time period are best suited and most representative as these are gathered in the system without any (known) bugs. The shorter time period means that the overall amount of votes is lower (126 in total), but more conclusive observations can be made about the manner of report consumption and reading time length. The ratings in this section are ratings for *all* reports but rated in the specified period. An overview of all ratings dis-aggregated for each

organizational unit is shown in table 5.1. The view times and view-to-rate conversion statistics are for reports *published* in the specified period.

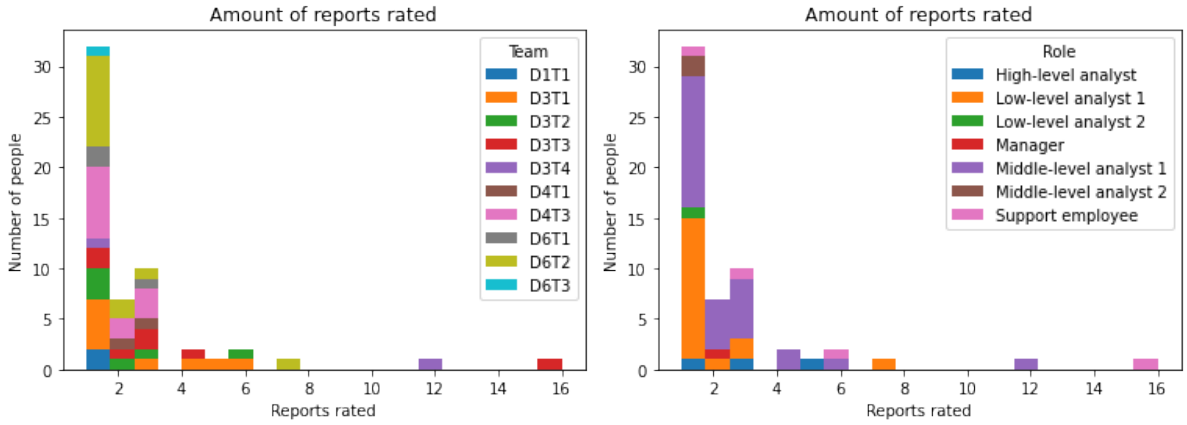
Table 5.1: Basic counts per department, team, and role for the ratings of November 20 – December 8.

Department	Voters	# of votes
Department 1	1-5	2
Department 2	1-5	1
Department 3	11-25	82
Department 4	11-25	20
Department 6	11-25	31

Team	Voters	# of votes
D1T1	1-5	2
D3T1	6-10	23
D3T2	1-5	14
D3T3	6-10	30
D3T4	1-5	13
D4T1	1-5	5
D4T3	11-25	20
D6T1	1-5	5
D6T2	11-25	23
D6T3	1-5	1

Role	Voters	# of votes
High-level analyst	1-5	9
Low-level analyst 1	11-25	29
Low-level analyst 2	1-5	1
Manager	1-5	2
Middle-level analyst 1	26-50	67
Middle-level analyst 2	1-5	2
Support employee	1-5	26

Figure 5.1 shows the distribution of total votes given per user. It shows that more than half of the users that have voted in this time period, only voted once or twice. Figure 5.1a and 5.1b show that the top 3 scorers are in different teams and have different roles.



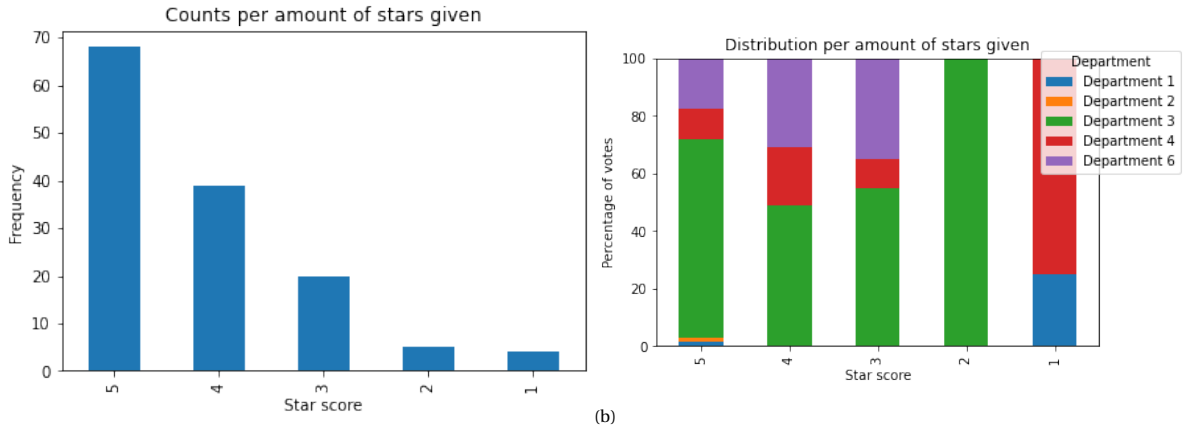
(a) (b)

Figure 5.1: The rating count per person and its distribution across different organizational units is shown in these figures. It is shown per team (5.1a) and per role (5.1b).

Figure 5.2 shows the distribution of star scores in total and the distribution dis-aggregated per department. Overall it shows that users tend to give high scores to reports. Figure 5.2b shows that members of department 6 tend to gain a larger share the lower a score gets (for scores of 3, 4, and 5 stars) and that department 1 has a disproportionally large share in the 1 star scores compared to its share in other scores, likely caused by the fact that only 5 one-star ratings were given.

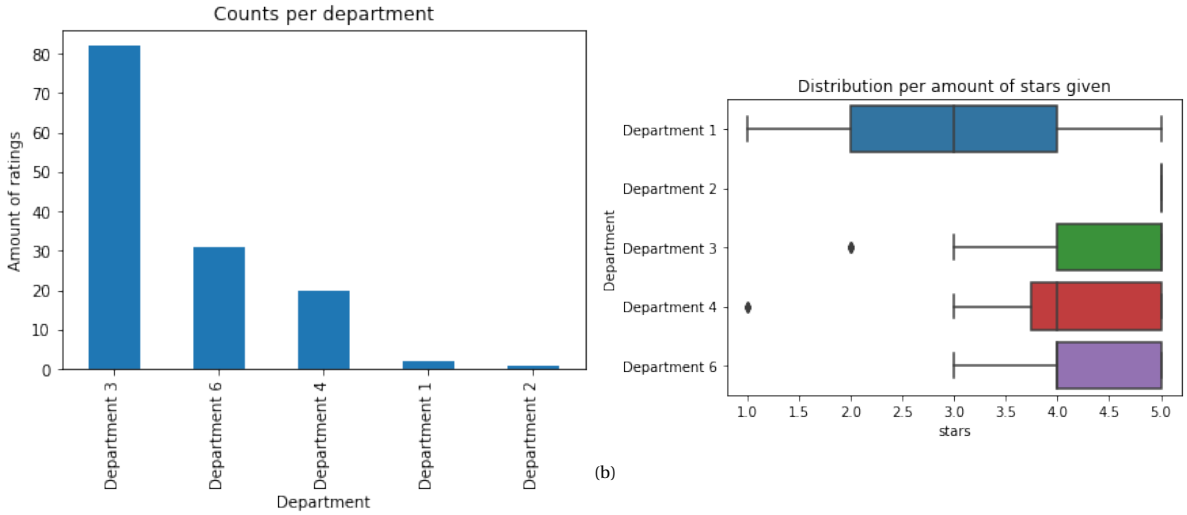
Figure 5.3 shows different scoring characteristics per department. Figure 5.3a shows the big difference of the total amount of votes between the departments. Departments 4 and 6 have also scored a reasonable amount, departments 1 and 2 do not really score reports. Figure 5.3b shows the distribution of the scores given per department. Departments 3, 4, and 6 have rated most reports and have very similar distributions with a median at 4 stars, the lower quartile around 4 stars, and the upper quartile at 5 stars. All three of the lower whiskers are at 3 stars. The distributions of departments 1 and 2 have different shapes, which can be explained by the lack of data from these departments.

Figure 5.4 shows the positive and negative keywords that are used to describe reports of the vendors. The left graph (5.4a) shows that overall each vendor published at least one report with a lack of depth. Lack of context, outdated topics, readability, relevance, (in)correctness, and uniqueness all seem not be an issue. The right graph (5.4b) shows that each vendor has published reports where the presence of contextual information, depth of information, and relevance of the topic are used multiple times. This suggests these characteristics are things that readers look for in a TI report. Applicability, correctness, and uniqueness look like attributes that are of less importance for readers.



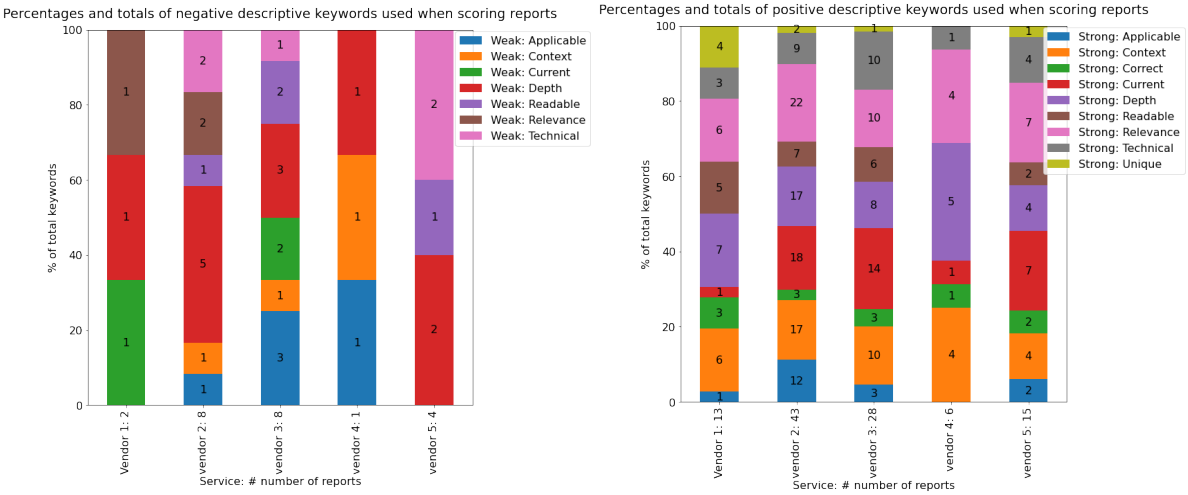
(a) (b)

Figure 5.2: The counts per star score and its distribution across different departments is shown in these figures. It is shown over the total dataset (5.2a) and per department (5.2b).



(a) (b)

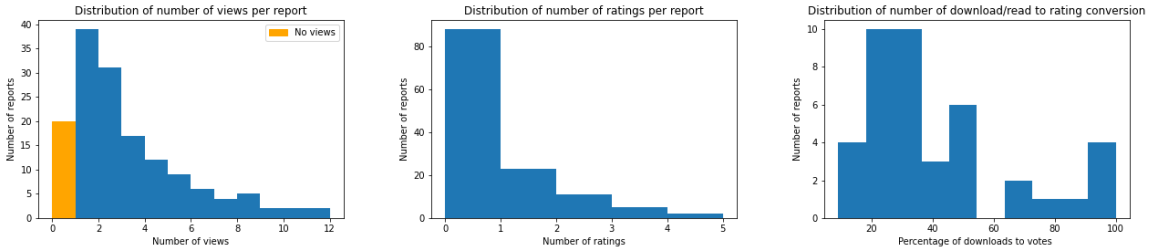
Figure 5.3: The distribution of the star score across departments in total votes (5.3a) and rating distribution (5.3b).



(a) (b)

Figure 5.4: The distribution of the used negative (5.4a) and positive (5.4b) ratings.

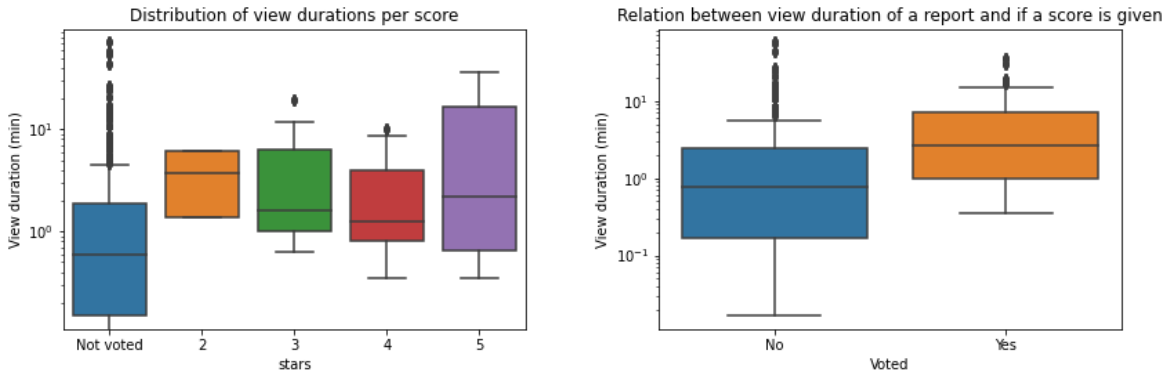
Figure 5.5 shows the distribution of the amount of views and ratings per report. Of all reports that are seen, most reports are seen only once and not scored at all. Noteworthy and positive is that the graph in 5.5a shows that more reports are seen than not seen. The graph in figure 5.5c shows peaks at 25, 33, 50, and 100%. This means that respectively one in four, three, two and one views resulted in a vote for a report. These peaks are explained by the fact that these view amounts are also most frequent for the reports. A minor peak is visible at 66%, likely the result of a report receiving 2 votes out of 3 report views.



(a) (b) (c)

Figure 5.5: The distribution of the amount of views per report(5.5a) and ratings of reports that are seen (5.5b). The reports with more than zero ratings are divided by their total amount of views to create graph 5.5c, this shows the percentage of views that resulted in an actual vote.

Figure 5.6 relates the view duration to the scoring behavior, it has to be kept in mind here that 5 star scores are given most often and 1 star scores least often. The view duration can only be collected from reports that are read in-browser and reports can only be scored in the browser interface, meaning no ratings should exist with a related view duration of 0. However, this was the case for 18 ratings, meaning something went wrong with storing the reading time. In this case the reading time was set the the average reading time for that star score. The ranges are very wide, with the shortest reading time being 21 seconds and the longest being 35 minutes (2146 seconds). Despite these differences, the median of all the different scores is between 75 and 226 seconds. This suggests most reports are actually not fully read in-browser and the browsers are mainly used to explore the content before downloading, or the looked-for information is quickly found. Figure 5.6b shows the view duration in relation to the fact if a report is scored. It seems that reports that are not scored, generally are viewed a shorter period of time. Reports that are not scored have a median view duration of 36 seconds, reports that are scored have a median view duration of 104 seconds. One possible explanation could be that readers rather quickly decide if a report is useful and if they should spend more time on reading it.



(a) (b)

Figure 5.6: For reports read in-browser, view duration is logged. These figures show the relation between view duration and scoring behavior. This is split up in view duration and the score (5.6a), and the view duration and actually scoring (yes/no) (5.6b).

5.1.2. Appreciation scores October 12 - December 8

The distribution of the publish dates of all rated reports is shown in figure 5.7. This shows that reports tend to be scored shortly after they are published, as most the publish dates of the scored reports overlap with the period of scoring. This is what would be expected in a fast changing cyber environment where Indicators of

Compromise generally have a short lifespan [49]. In the remainder of this section only some statistics about the ratings themselves will be presented, as the view duration for each rating is not reliable in this time span and the keywords to add to the ratings was not implemented yet.

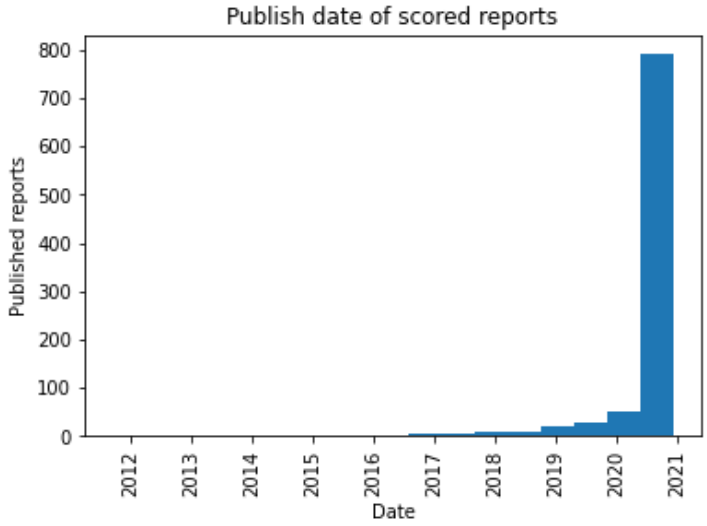


Figure 5.7: Distribution of the publish date of rated reports.

Table 5.2 shows the same counts as table 5.1, but for all casted votes, a grand total of 622 votes. It shows that the relative difference between the votes does not differ too much, only the absolute amount of votes is higher.

Table 5.2: Basic counts per department, team, and role for the ratings of October 12 – December 8.

Department	Voters	# of votes
Department 1	1-5	6
Department 2	1-5	3
Department 3	26-50	415
Department 4	11-25	50
Department 5	1-5	4
Department 6	26-50	144

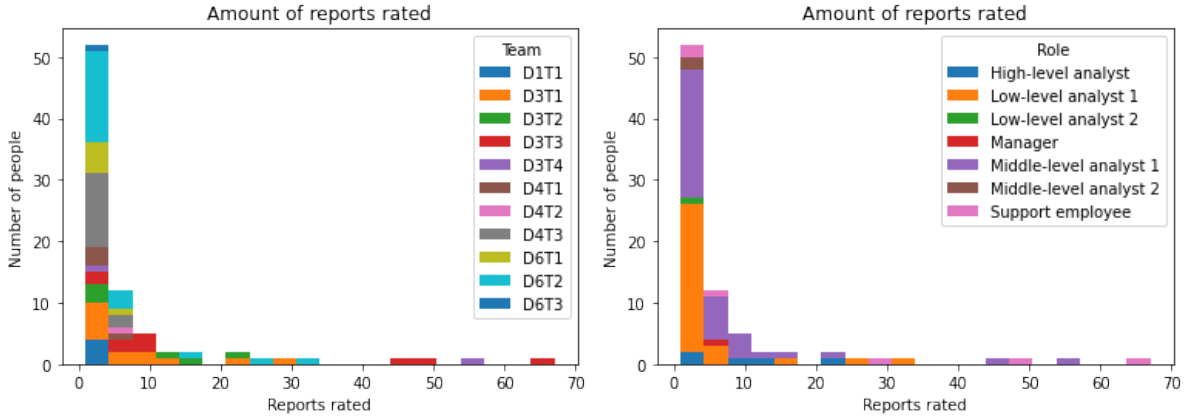
Team	Voters	# of votes
D1T1	1-5	6
D3T1	11-25	108
D3T2	6-10	55
D3T3	6-10	199
D3T4	1-5	57
D4T1	1-5	11
D4T2	1-5	7
D4T3	11-25	37
D6T1	6-10	19
D6T2	11-25	122
D6T3	1-5	1

Role	Voters	# of votes
High-level analyst	26-50	42
Low-level analyst 1	26-50	141
Low-level analyst 2	1-5	1
Manager	1-5	5
Middle-level analyst 1	26-50	277
Middle-level analyst 2	1-5	3
Support employee	6-10	153

Figure 5.8 shows the distribution of total votes given per user. Again the overall trend is very comparable to figure 5.1, such as the fact that most users only have voted once or twice. However, in this collection of votes, three out of four most scoring users are in the same team. Two of these four even have the same role within that team (support employee).

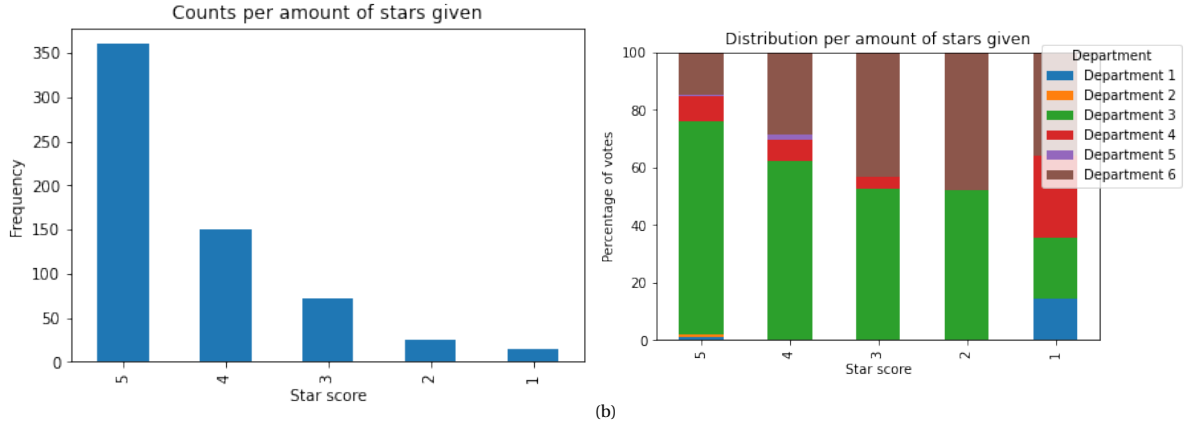
Figure 5.9 shows the distribution of star scores in total and the distribution dis-aggregated per department. Again, the general trend is very comparable to the one observed in figure 5.2. Overall, the five star score is given most often to a report. Department 3 is responsible for most five star scores and as the scores get lower (toward three stars), department 6 is responsible for a larger share of the scores.

Figure 5.10 shows different scoring characteristics per organizational unit. The overall trends are again comparable to the graphs shown in figure 5.3. For the boxplot (5.10b) the means are slightly better defined thanks to the higher volume of votes. Department 6 has a wider lower box and whisker, whilst the distribution of departments 3 and 4 did not change.



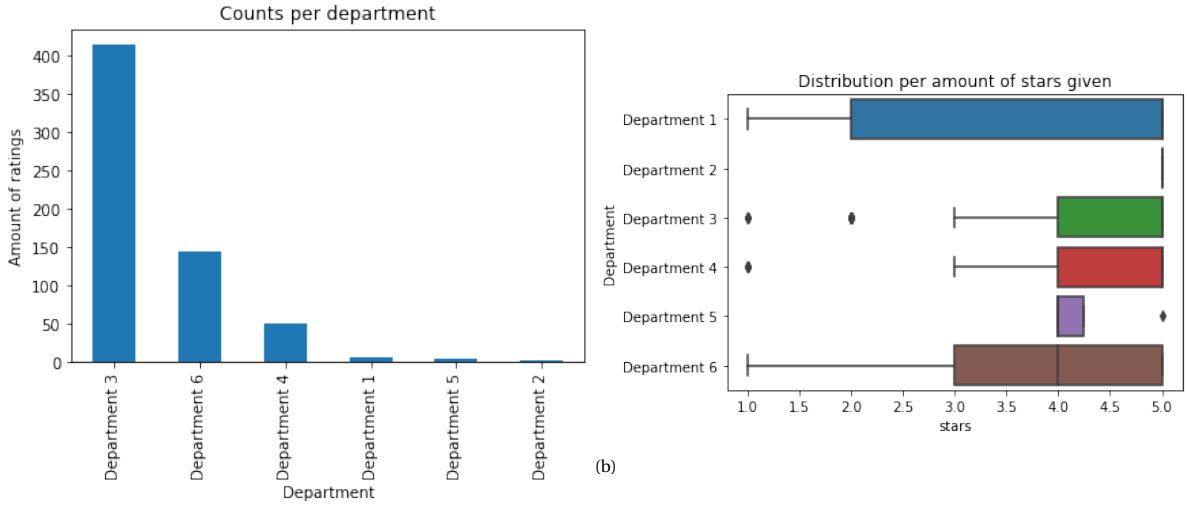
(a) (b)

Figure 5.8: The rating count per person and its distribution across different organizational units is shown in these figures. It is shown per team (5.8a) and per role (5.8b).



(a) (b)

Figure 5.9: The counts per star score and its distribution across departments. It is shown over the total data-set (5.9a) and per department (5.9b).



(a) (b)

Figure 5.10: The distribution of the star score across different departments. The ratings are shown overall (5.10a) and per department (5.10a).

5.2. Description of the report metrics

To decide on a time period for the reports to use for the remainder of this study, a cutout of figure 5.7 is shown in figure 5.11. The figure shows that, as also concluded from figure 5.7, reports are most often rated in the time shortly after they are published. This shows in the peak that starts at August 2020, the moment the rating system was implemented. Based on this we decide to only use reports that are published on or after the 1st of August, 2020.

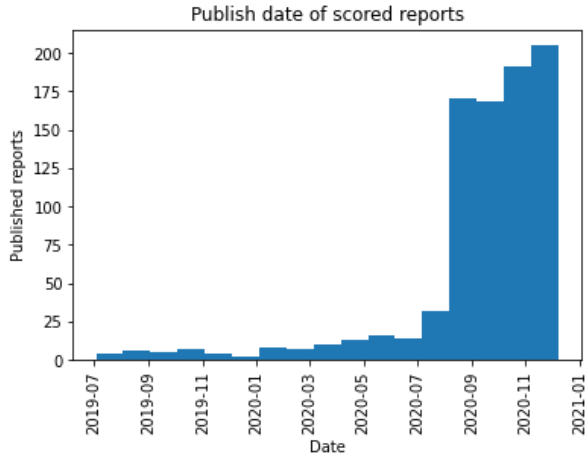
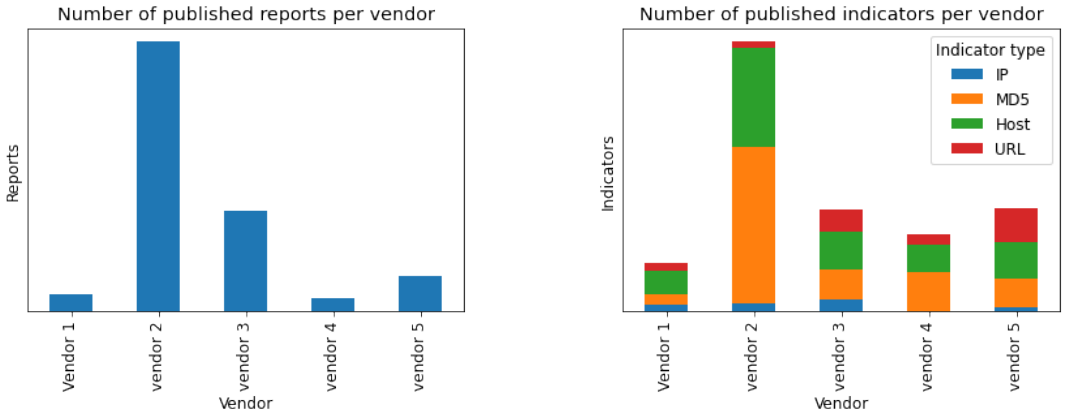


Figure 5.11: Distribution of the publish date of rated reports in the past two years.

5.2.1. Volume

Figure 5.12 shows basic counts regarding the volume of the reports and feeds. In the case of 5.12b, the indicators are unique indicators present in the reports. The figures show that Vendor 2 publishes most reports as well as the most indicators. Vendor 3 publishes the second most reports and Vendor 5 the third most. However, Vendor 3 and Vendor 5 publish nearly the same amount of indicators. Vendor 1 and Vendor 4 publish the least amount of indicators and reports, although Vendor 4 publishes relatively much indicators compared to its published reports.



(a) (b)

Figure 5.12: Basic volume counts for the reports, excluded indicators gathered using the IoC parser.

5.2.2. Timeliness

Figure 5.13 shows statistics about the timeliness of the different publishers. The figure both shows the days a publisher is sooner when reporting an indicator first, as well as the times a publisher was later when an other publisher reported on an indicator first. The figure shows that the average amount of days each publisher

is sooner/later than the other publishers are all similar. The median of Vendor 4 is highest, but the upper quartiles of Vendor 1, Vendor 3, and Vendor 5 are all higher.

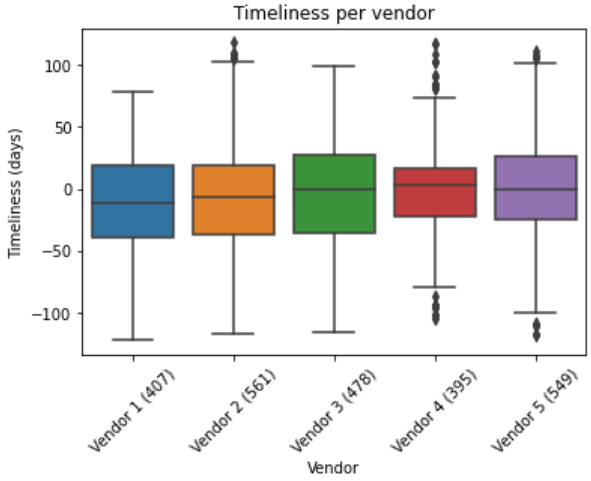


Figure 5.13: Timeliness of indicators, including indicators gathered using the IoC parser and the total amount of overlapping indicators between brackets.

5.2.3. Overlap

Figure 5.14 shows the percentages of overlap between the published indicators per publisher. In line with findings by Li et al. [37] and Bouwman et al. [12], vendors do have a low overlap with other vendors. The vendor on the x-axis is compared to the vendor on the y-axis. To calculate a percentage, the overlap is divided by the total amount of indicators of the vendor on the x-axis.

The amount of overlapping values are all more in line with the vales found in literature, with no percentage of overlapping indicators above 7%. For all vendors, the metadata file is used if possible and additional indicators are extracted from the PDF report using the IoC parser.

Figure 5.14: Overlap of indicators, including indicators gathered using the IoC parser.

5.2.4. Population

Actors Figure 5.15 shows the absolute amount of overlapping actor names per publisher pair. These overlapping actor names are already normalized using the actor sheet from Roth [60]. This makes the rather low amount of overlapping actor names an interesting observation, meaning either the actor sheet from Roth is incomplete or the vendors mainly cover different actors.

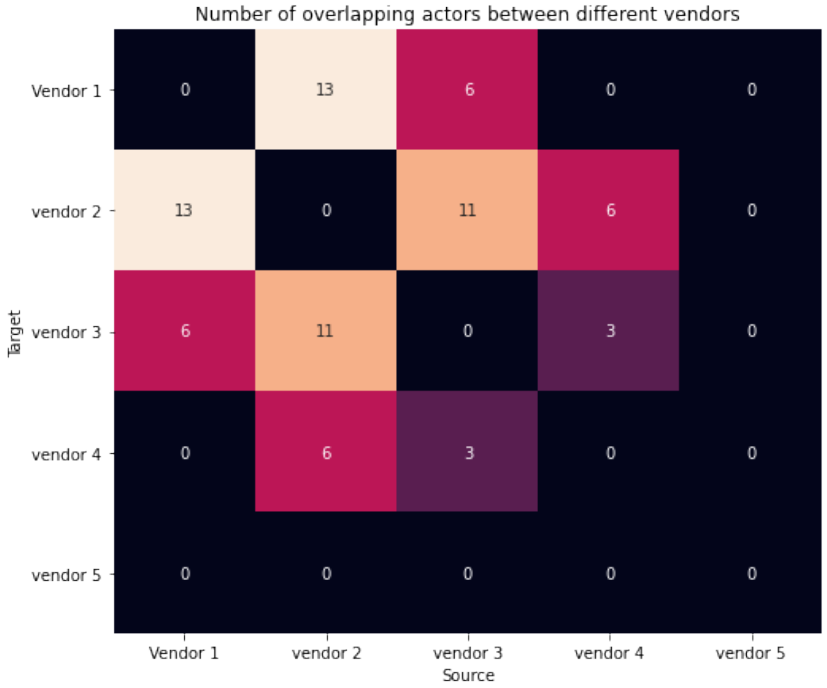


Figure 5.15: Overlap of used actor names, actor names are normalized using the actor sheet from Roth [60].

Source geographies Figure 5.16 shows the sum of times a country is mentioned as source country of a threat or attack in all reports. The figure shows the vendors have a limited insight in threats around the world or decide to report on only specific source countries.

Target geographies Figure 5.17 shows the sum of times a country is mentioned as target country of a threat or attack in all reports. The amount of mentioned source countries is a lot more global than the amount of target countries. However, the United States are either over-represented in coverage by the vendors or actually are victim of a lot more cyber attacks.

5.2.5. Security threats

Figure 5.18 shows the percentage of times a type of threat is mentioned in the reports for each of the different publishers. The figure quite clearly shows that Advanced Persistent Threats (APTs) are most frequently discussed by most of the vendors. Also 'malware' seems to be frequently mentioned by all vendors, for all vendors this threat type makes up for about 20% of all mentioned threats, 'credential compromise' makes up for about 10-20% for all vendors. Vendor 4 and Vendor 5 do discuss ransomware relatively more often than the other vendors. Vendor 2, and Vendor 1 to a lesser extent, discuss DDOS threats relatively more than the other vendors. Vendor 1 seems to have the largest share of all vendors discussing zero-day exploits. Interestingly enough 'Covid-19' is not mentioned a lot, despite being a major impact on the world during the measurement period.

Figure 5.16: Sum of all discussed source countries of threats in the selected time period.

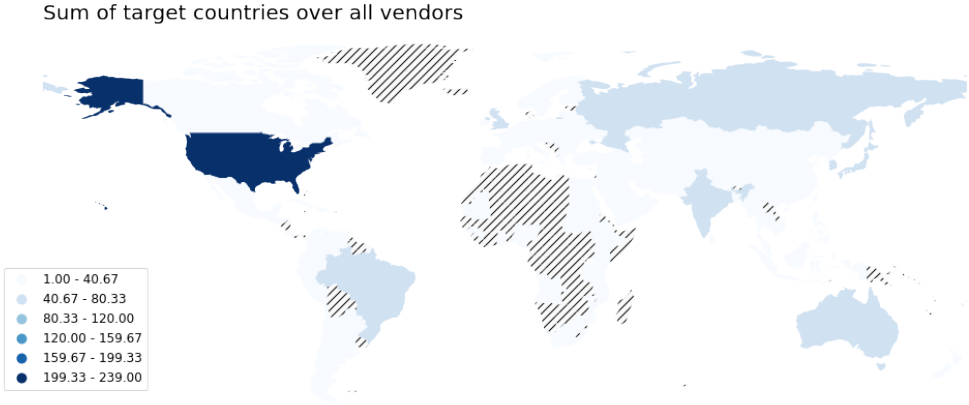


Figure 5.17: Sum of all discussed target countries of threats in the selected time period.

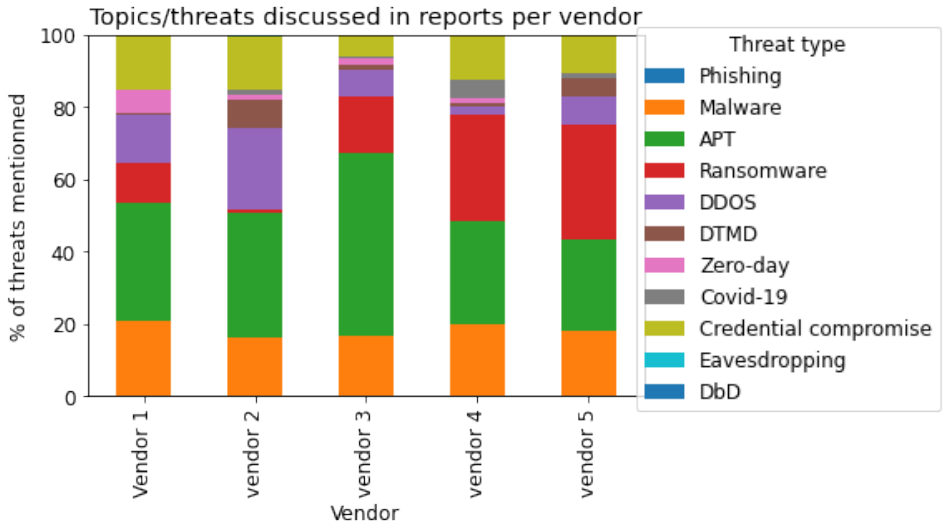
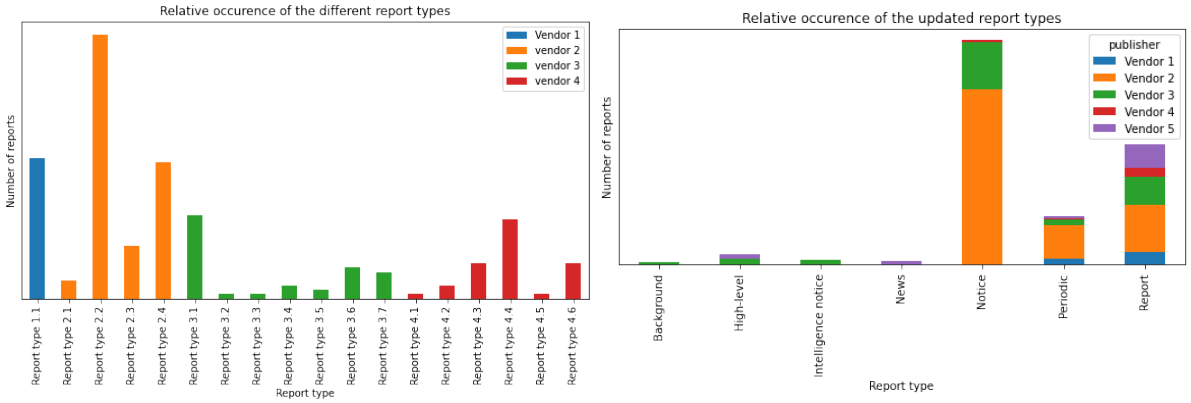


Figure 5.18: Discussed threats per vendor, found using word matching the report PDFs.

5.2.6. Report type

A mapping is made to link the categories of the vendors to more global report types. The categories to divide the report types in are shown below and in figure 5.19.

- **Report** This is the most general category of reports. This category includes extensive 60-page reports, but does not exclusively exist of such reports. Many vendors try out a new report type and discontinue it after three or four issues, report categories like this are also placed under this aggregated category.
- **Periodic update** Daily, weekly, monthly, and quarterly updates could be aggregated to one category. An issue that could arise when aggregating these, is that both the level of detail can be different, as well as the fact that some might look forward to the upcoming month whilst others look back at the past month. All vendors have reports that fit in this category.
- **Intelligence notice** Reports that are explicitly discussing the future. These intelligence notices can both concern techniques that vendors expect to be used more in the future, as well as newsworthy events that might be used in phishing emails. Only one vendor publishes reports like this.
- **Aggregated news** This category consists of reports which aggregates different news sources into a report. This is meant to be a summary of all relevant news, where links are provided for the actual news source. These news source often also include a short summary of reports a vendor published themselves. Only one vendor publishes reports like this.
- **Background** The background type of reports are reports that are meant to give more information about a single topic. This can concern information about certain TTPs, actors, or countries. These can mainly be used to understand more of the playing field cyber specialists are operating in. Two of the vendors publish reports like this.
- **Short notice** The short notice category is different than the *Report* category in the sense that they are significantly shorter than other reports. These reports range from a couple of sentences to a maximum of 1 page. These often reports are often meant as a quick heads-up or as a gateway to more extensive reports in a vendor's portal. Three of the publishers create these reports.
- **High-level** The high-level reports are reports that don't go (too deep) into the technical details but tell a lot more about the context. These reports are useful for people trying to understand the political and operational context, without needing to know and work with any Indicators of Compromise. Two vendors create reports in this category.



(a) (b)

Figure 5.19: Each of the publishers has different categories for their reports. The distribution of the original report types is shown in 5.19a, the normalized report types and the distribution over the vendors is shown in 5.19b.

5.2.7. Sentiment and subjectivity analysis

The result of a sentiment and objectivity analysis for the reports is shown in figure 5.20. The figure shows that most reports receive a sentiment score between -0.05 and 0.1, the maxima are around -0.2 and 0.2. Taking in mind that the sentiment score can range between -1 and 1, none of these values fall in either the upper or lower third of the possible sentiment scores. Concerning the subjectivity scores, these values can range between 0 and 1, with 0 being completely objective and 1 being very subjective. Most values are between 0.25 and 0.5, meaning most reports are not very subjective and tend to be slightly more objective. Only two reports are around or above a subjectivity score of 0.6. The exact values for the outliers in figure 5.20 are shown in table 5.3.

The two reports from Vendor 2 and Vendor 4 are briefly analyzed to see if there is an obvious reason for these reports to receive a slightly higher subjectivity score compared to other reports. What appears is that both reports discuss the same actor. Due to the naming of this actor and the frequency the actor is mentioned in each report, the natural language processor misinterpreted it as being subjective and positive. When taking a look at the actor naming sheet of Roth [60], this miscalculation of the algorithm is easily explained. To give some examples of actor names in Roth's sheet: Playful Dragon, Iron Viking, Twisted Kitten.

Table 5.3: Outliers in the sentiment or subjectivity for figure 5.20.

Publisher	Sentiment	Subjectivity	Words	Figures	Stars
Vendor 2	-0.16	0.51	306	3	3.87
Vendor 2	0.16	0.50	251	1	3.91
Vendor 2	0.05	0.21	547	1	4.40
Vendor 1	0.20	**0.59**	792	5	4.21
Vendor 3	0.16	0.38	212	0	2.56
Vendor 2	0.04	0.20	424	1	2.56
Vendor 2	-0.03	0.21	263	1	3.00
Vendor 2	0.17	**0.66**	457	5	3.87
Vendor 4	-0.17	0.45	2210	2	3.87

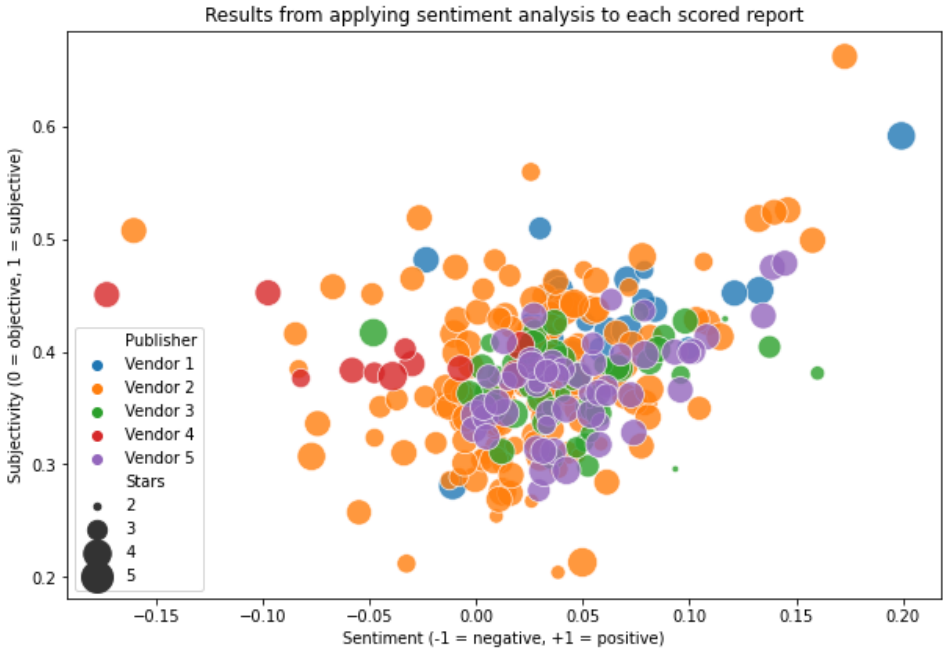


Figure 5.20: Sentiment and subjectivity ratings using textblob [38] for all scored reports, size represents star score.

5.3. Summary

In the first part of this chapter, we tried to find an answer to our third sub-research question. The question was:

3. Which patterns can be distinguished in the appreciation scores?

The process of answering this question was split up in two parts, one part including the data starting from November 20, the other part containing data starting October 12. This enabled us to evaluate ratings, including matching view durations, but in smaller quantities and to evaluate a larger quantity of ratings but with possibly missing data.

Several lessons are learned from these two analyses. Most readers are not keen on voting, often voting only between 1 and 3 times in a time period of two weeks. A small selection of readers vote more often, voting between 4 and 16 times in that same two week period. We learned that most readers tend to give high scores to reports, as 4 and 5 star scores account for more than 70% of total scores given.

The departments that score most are departments 3, 4 and 6. In the shorter time period, these departments have very similar scoring distributions, but in the longer measurements period, department 6 seems slightly more inclined to give lower ratings.

It also shows that the view duration of a report seems to indicate if a reports will be scored. The median view duration is higher when a report is scored compared to a report that will not be scored. No difference in view duration can be distinguished between different star scores.

The 9 dimensions which can be used to explain a vote also offer some insights. Where all other vendors received the positive dimensions ‘current’ in a significant amount related to other dimensions, Vendor 1 did not receive this positive dimension once. Unfortunately the amount of negative dimensions related to votes is too small to draw any conclusions from.

All in all, these findings offer some perspective on a more in-depth analysis, especially as more data keeps on being gathered. As scoring behavior per department, team, role, or even individual gets more clear, eventually scores can be normalized per person rather than over all data. Next to that, as the 9 dimensions are used more to explain scores, strengths and weaknesses of the vendors will become clear and can be mirrored to the need of the organization.

6

The relation between report metrics and appreciation scores

6.1. The statistical relation between metrics and appreciation scores

In this chapter, we will find out if any of the report features can be linked to a higher or lower appreciation score. As introduced in chapter 3, for the categorical variables first a Kruskal-Wallis test, a non-parametric alternative to a regular ANOVA, is performed and during the post-hoc analysis with the Mann-Whitney U test, a Bonferri-Holm correction is applied. For the continuous variables we apply linear regression to find out possible relations.

The non-parametric tests for categorical variables differ from parametric tests in the sense that they are applied on the rank of a sample rather than the actual value. When there are five samples, say $\{3, 3.1, 3.8, 3.9, 4.5\}$, these get replaced by their respective ranks of $\{1, 2, 3, 4, 5\}$. First the Kruskal-Wallis test is applied when three or more categories are compared to discover if there likely is a significant difference between any of the means. When this test has a significant p-value ($\alpha = 0.05$), several post-hoc methods have to be applied, Mann-Whitney U in this case. When using multiple post-hoc tests like this, the random chance to achieve a p-value less than 0.05 increases. To account for this, the Bonferri-Holm correction is applied. During this correction, the α value changed according to the rank of the achieved p-value, an exact explanation can be found in section 3.4.

We chose to not normalize nor standardize the features. We chose to do so because the only interaction that is tested for, is between the different vendors and each metric. As this is a binary feature (published by vendor: yes/no), collinearity is limited to a minimum. As during the linear regression only one metric is tested at a time, a different magnitudes of the features will not form a problem. In the case of the number of words this could lead to small coefficients, this has to be kept in mind when interpreting the outcomes.

By limiting the amount of interactions to include only interactions between vendors and metrics, we kept the necessity to guess the meaning of any interactions to a minimum. Imagine, an interaction between the number of indicators and the subjectivity score of a report correlates with the received (normalized) appreciation score, but, for example, the interaction between the number of *overlapping* indicators and the objectivity score does not correlate. A situation like this would leave us guessing why one specific interaction does correlate and the other does not. Then, when having constructed a hypothesis why this could be the case, the actionable value of this hypothesis is still unknown.

In contrast, when finding a significant correlation of the interaction between a vendor and a metric with the (normalized) appreciation score, you can more easily argue that reporting on that metric seems to be a strong point of the vendor.

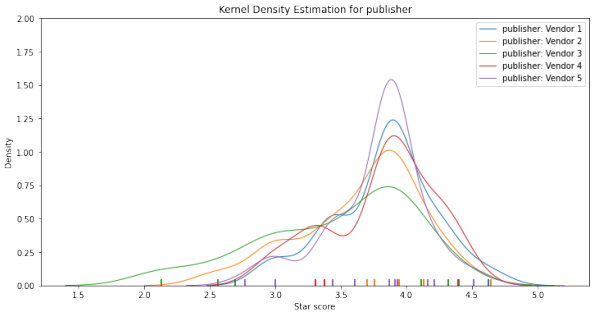
6.1.1. Publisher

Table 6.1 and figure 6.1 show the results for the tests over the publishers. The achieved p-value during the Kruskal-Wallis test is 0.02, this indicates that an average score of one of the publishers likely has a significantly different mean than the other publishers. The results of the post-hoc testing is shown in the p-value column of table 6.1 and the corresponding α values the p-values are compared to in the column next to it. It shows that

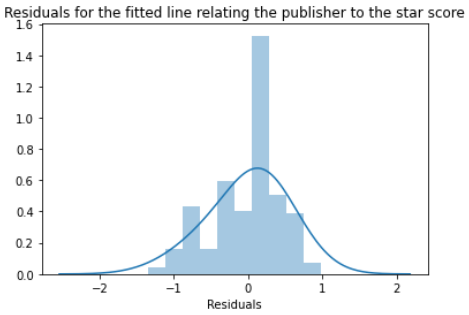
Vendor 3 has a p-value that would indicate a significantly lower average without Bonferri-Holm correction, but it cannot be interpreted as a significantly lower average.

Table 6.1: Relation between publisher and received appreciation scores

	Coefficients	p-value	α value
const	3.091	0.000	0.050
Vendor 1	0.736	0.051	0.013
Vendor 2	0.573	0.328	0.050
Vendor 3	0.382	*0.020*	0.010
Vendor 4	0.718	0.185	0.025
Vendor 5	0.683	0.083	0.017



(a)



(b)

Figure 6.1: Kernel density estimation for the appreciation scores based on the publisher

6.1.2. Volume

Indicators Table 6.2 shows the identified relation between the amount of indicators that are mentioned in a report and the consequence it can have on the received appreciation score. The right table shows the specific interaction of the Vendor 3 reports and their reported indicators. These results show that there is a minor positive relation between the amount of indicators in a report and the appreciation score it receives. When dis-aggregating this per publisher, only the amount of indicators in Vendor 3 reports result in a significant p-value. This suggests the significant p-value of the indicators overall is caused by the significance of the amount of indicators in reports of Vendor 3.

Table 6.2: Relation between the amount of reported indicators and the received appreciation scores.

	coef	p-value
Intercept	3.609	0.000
indicators	0.001	**0.003**

	coef	p-value
Intercept	3.320	0.000
Vendor 3 & indicators	0.003	**0.018**

6.1.3. Timeliness

Times faster Table 6.3 shows that the amount of indicators that a report mentioned first does not have a significant relation with the received appreciation score of a report.

Table 6.3: Relation between timeliness (as a count of indicators the report reported on first) and received appreciation scores.

	coef	p-value
Intercept	3.584	0.000
timeliness count	0.009	0.075

Average timeliness of indicators Table 6.4 shows that the average timeliness of the presented indicators in a report does not have a significant relation with the received appreciation score of a report.

Table 6.4: Relation between the average timeliness of indicators in a report and the received appreciation scores.

	coef	p-value
Intercept	3.673	0.000
average timeliness	-0.003	0.172

6.1.4. Overlap

Overlap can be expressed as the amount of non-unique indicators a report mentions, but also relates to the amount of unique indicators in a report. Table 6.5 shows the identified relationship between the amount of non-unique indicators in a report and its impact on a report score. The right table shows the identified relation between the amount of unique indicators specifically for Vendor 3 reports. These scores indicate that for reports in general more non-unique indicators is related to a higher appreciation score. For reports of Vendor 3, the results indicate that more unique indicators is linked to a slightly higher appreciation score. Both results are strongly related to the results in table 6.2, as all these results can be summarized as 'more indicators is better'.

Table 6.5: Relation between the amount of (non-)unique indicators in a report and the received appreciation scores.

	coef	p-value
Intercept	3.584	0.000
non-unique indicators	0.011	**0.037**

	coef	p-value
Intercept	3.334	0.000
Vendor 3 & unique indicators	0.004	**0.018**

6.1.5. Population

Source country The amount of source countries which are discussed in at least five reports is 5, the achieved p-value in the Kruskall-Wallis test is 0.91. This means that the null hypothesis is accepted and we cannot conclude reports concerning a specific source country are more or less appreciated than reports reporting about other countries.

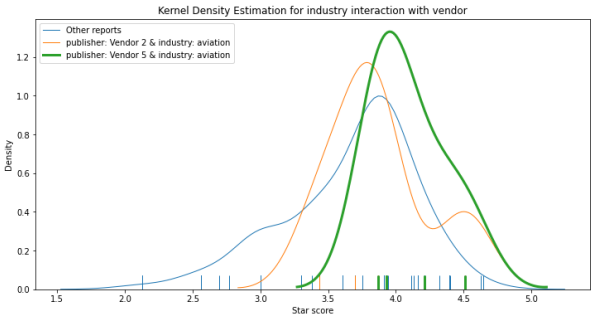
Target country The amount of target countries which are discussed in at least five reports is 37, the achieved p-value in the Kruskall-Wallis test is 0.51. This means that the null hypothesis is accepted and we cannot conclude reports concerning a specific target country are more or less appreciated than reports reporting about other countries.

Industry The amount of industries which are discussed in at least five reports is 33, the achieved p-value is 0.24. This means that the null hypothesis is accepted and we cannot conclude reports concerning a specific industry are more or less appreciated than reports reporting about other industries.

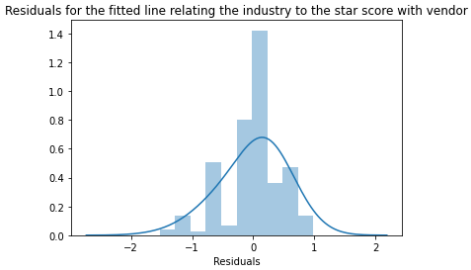
Next to testing for individual industries, the interaction between industries and publishers is also tested. This resulted in two publishers discussing the aviation industry to achieve the significant p-value of 0.036 in the Kruskal-Wallis test. Vendor 2 and Vendor 5 discuss the aviation industry in at least five (rated) reports. When comparing the ratings of these reports, it shows that reports of Vendor 5 discussing this industry seem higher rated, this is also shown in table 6.6 and figure 6.2.

Table 6.6: Relation between the publishers that report about the aviation industry and received appreciation scores.

	Coefficients	p-value	α value
const	3.660	0.000	0.050
Vendor 2 & aviation industry	0.218	0.317	0.050
Vendor 5 & aviation industry	0.432	**0.006**	0.025



(a)



(b)

Figure 6.2: Kernel density estimation for the appreciation scores based on the industry 'aviation'.

Actor The amount of actors which are discussed in at least five reports is 42, the achieved p-value is 0.57. This means that the null hypothesis is accepted and we cannot conclude reports concerning a specific actor are more or less appreciated than reports reporting about other actors.

6.1.6. Security threats

Table 6.7 shows the results for the different threat types and how they relate to the received appreciation scores. There are no significant relations when applying the Bonferri-Holm correction. However, the threat type 'zero-day' is close to being significant and would be without the correction.

Table 6.7: Relation between the discussed threat types and received appreciation scores.

	Coefficients	p-value	α value
const	3.683	0.000	0.050
threat APT	0.061	0.164	0.013
threat Covid-19	-0.058	0.155	0.010
threat DDOS	-0.106	*0.043*	0.008
threat DTMD	0.128	0.440	0.025
threat Malware	-0.000	0.424	0.017
threat Phishing	0.048	0.493	0.050
threat Ransomware	-0.184	0.023	0.007
threat Zero-day	0.421	*0.007*	0.006

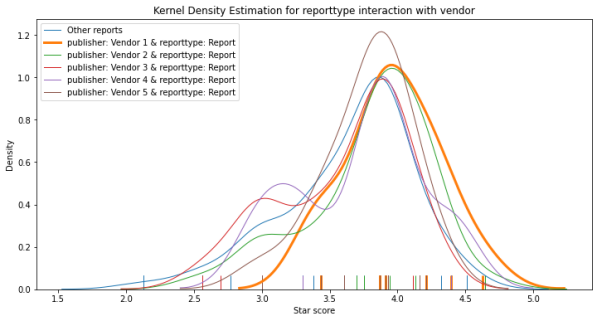
6.1.7. Report type

The achieved p-value among the different report types is 0.067. This means that the null hypothesis is accepted and we cannot conclude specific report types are more or less appreciated than other report types.

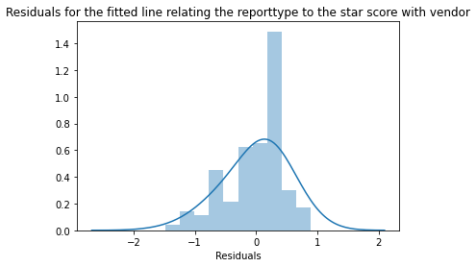
When testing the interaction between the different report types and the different publishers, a significant result is achieved for the most general report type 'report', with a p-value of 0.013 resulting from the Kruskal-Wallis test. This type of report seems to be slightly higher appreciated from Vendor 1 than from other publishers. This can be seen in table 6.8 and figure 6.3.

Table 6.8: Relation between the 'report' report type per publisher and received appreciation scores.

	Coefficients	p-value	α value
const	3.612	0.000	0.050
Vendor 1 & reporttype Report	0.350	**0.004**	0.010
Vendor 2 & reporttype Report	0.183	*0.042*	0.013
Vendor 3 & reporttype Report	-0.014	0.239	0.025
Vendor 4 & reporttype Report	0.113	0.390	0.050
Vendor 5 & reporttype Report	0.165	0.148	0.017



(a)



(b)

Figure 6.3: Kernel density estimation for the appreciation scores based on the report type 'report'.

6.1.8. Sentiment and subjectivity

Sentiment Table 6.9 shows that there seems to be no relationship between report sentiment and the received rating. Thus we can accept the null hypothesis and the direction and strength of the sentiment does not impact the report rating.

Table 6.9: Relation between the sentiment score and received appreciation scores.

	coef	p-value
Intercept	3.698	0.000
sentiment	-0.318	0.842

Subjectivity Table 6.10 shows that there seems to be a relationship between a report's objectivity and the received appreciation score. When investigating further per publisher, it shows only reports from Vendor 2 show a significant relationship between subjectivity and the received appreciation score, as shown in the right table. As the subjectivity score ranges from 0 to 1, this score means that fully subjective report from Vendor 2 would receive a 1.3 higher appreciation score compared to a fully objective report from this same vendor. As in our data the subjectivity ranges between 0.2 and 0.7, the maximum difference between the least subjective and the most subjective report can only be around 0.65 stars.

Table 6.10: Relation between the subjectivity score and received appreciation scores.

	coef	p-value
Intercept	3.231	0.000
subjectivity	1.149	**0.011**

	coef	p-value
Intercept	3.159	0.000
Vendor 2 & subjectivity	1.331	**0.011**

6.1.9. Report counts

Words Table 6.11 shows that there is no significant relation between the total amount of words in a report and the received appreciation score. However, when investigating specifically for Vendor 3 reports, it seems that there is a relationship between the total amount of words and the received appreciation score. It must be noted, that the coefficient is so small this does not effectively show in the received appreciation scores. When relating the words per page to the appreciation score, rather than the total amount of words, a significant result is reached as well. There seems to be a minor positive relation between the amount of words per page and the appreciation score, as seen in table 6.12. The right table also shows that this relation is slightly stronger for the amount of words per page for Vendor 3 reports.

Table 6.11: Relation between the word count and received appreciation scores.

	coef	p-value
Intercept	3.541	0.000
words	0.000	0.200

	coef	p-value
Intercept	3.005	0.000
Vendor 3 & words	0.000	**0.021**

Table 6.12: Relation between the word count per page and received appreciation scores.

	coef	p-value
Intercept	3.448	0.000
words per page	0.001	**0.017**

	coef	p-value
Intercept	2.721	0.000
Vendor 3 & words per page	0.005	**0.031**

Figures Table 6.13 shows that there seems to be no significant relation between the amount of figures (relative nor absolute) to the received appreciation scores and thus we can accept our null hypothesis that the amount of figures does not impact the star rating.

Table 6.13: Relation between the figures count and received appreciation scores.

	coef	p-value
Intercept	3.653	0.000
figures	0.003	0.569

	coef	p-value
Intercept	3.742	0.000
figures per page	-0.081	0.080

Pages Table 6.14 shows there seems to be no significant relation between the number of pages of a report and its received appreciation score and thus we can accept our null hypothesis that the amount of pages does not impact the star rating.

Table 6.14: Relation between the page count and received appreciation scores.

	coef	p-value
Intercept	3.601	0.000
pages	0.008	0.069

6.2. The value of the metric-score relations

In the previous section we showed several statistically significant relations between different report features and their respective appreciation scores. Summarized, the relations were:

Significant relations

- More indicators is better, this is especially true for Vendor 3 reports.
- Overall, more non-unique indicators in a report is better. For Vendor 3 specifically, more unique indicators is better.
- Vendor 5 publishing about the aviation industry is higher appreciated than Vendor 2 writing about the aviation industry.
- When comparing the general ‘report’ category, Vendor 1 seems to be higher appreciated than the other publishers.
- Subjectivity seems to be appreciated. The more subjective a report, the higher it is scored. When looking at individual publishers, only Vendor 2 achieves a significant p-value for this relation, likely causing the significant relation for the overall category.
- The amount of words does overall not seem like an influence on the appreciation of reports. However, for Vendor 3 reports with more words are appreciated higher with a minuscule difference.
- The amount of words per page has overall a significant impact in the received score. However, again only Vendor 3 shows a significant relationship if you distinguish per publisher.

Significant without correction

- Reports from Vendor 3 are scored about 0.3 stars lower on average compared to other publishers.
- Reports discussing DDOS threats are generally less appreciated.
- Reports discussing zero-days are generally higher appreciated than reports discussing other threats.

This list of statistical more and less significant results has to be taken with a grain of salt. Halfway during this study, an initial run of significance tests was performed in order to prepare for the interviews. During this initial run, different statistically significant relations appeared, showing that the relations in the data are dependent on the subset of ratings that is taken.

The chosen relations to validate during the interviews were:

- Vendor 2 reporting about attacks targeting the military (industry) were appreciated more than other publishers discussing attacks targeting the military (industry).
- Reports discussing malware by Vendor 3 (there was no report type normalization yet), were disliked compared to all other report types.
- Vendor 5 reporting about cyber attacks targeting Ukraine are rated lower than other publishers reporting about attacks targeting Ukraine.

The negative impact of the reports discussing malware on the overall score of Vendor 3 were recognized by three of the four interviewees. Although not explicitly coming up, some of the found relationships in the final analysis can still be explained by these reports discussing malware by Vendor 3. This starts with the fact that Vendor 3 reports are overall less liked compared to the other publishers, and although this is not with a significant value according to the corrected α-value, the low p-value does show that something is happening there. The other signs mainly show in the report counts, where Vendor 3 specifically is the only significant publisher when analyzing each of the counts per publisher. As the reports discussing malware by Vendor 3 often only consist of 5 sentences discussing the existence of a type of malware, these reports discussing malware have a low amount of total words, but also a low amount of words per page. The positive relation between more words (per page), which seems to be present only in Vendor 3 reports, is likely a direct consequence of the general aversion of these short and uninformative reports discussing malware.

Discussing the overall lower score of Vendor 5 discussing attacks targeting Ukraine compared to the other publishers did not rang any bell for any of the four interviewees. However, one of the interviewees did offer an alternative explanation. There is one specific category of reports that this interviewee disliked in particular, the weekly overviews. These overviews were not of good quality and they could have had small notifications of attacks on Ukraine. These would explain the low ratings without any of the interviewees remembering them, as they likely read over the fact that Ukraine was even mentioned.

Lastly, the seemingly positive relation of Vendor 2 publishing about attacks targeting the military (industry). Again, none of the interviewees recognized this relation. However, two of them explicitly said that they were happy with the quality of Vendor 2 (and Vendor 1) reports. The fact that this satisfaction showed in relation to the military (industry) is likely a coincidence. These three examples, although not significant according to the latest analysis, illustrate perfectly how many of the found relationships have more to it, something that wasn't caught by just measuring the report features and comparing them to the report scores.

6.3. Summary

In this chapter we tried to find an answer to our second sub-research question:

4. What is the relation between quantitative metrics & appreciation scores?

The previous section (6.2) answered most of this question. For a limited amount of quantitative metrics a significant relation was found with the scored reports. However, none of these found relations were found in the intermediate results as used for the interviews. This shows the relations are volatile and dependent on the subset of the data you choose to analyze. Based on this, we conclude that there are no relations between the quantitative report metrics and the appreciation scores.

7

Predicting appreciation scores

Predicting the expected appreciation score of a report can serve multiple uses. A use-case in production is to predict the appreciation score of reports that did not receive ratings yet. Using this prediction, users could be presented with new reports that are likely worth reading. Another use-case is to use the prediction to understand report quality better. When reports containing specific features are (correctly) predicted higher than other reports, this could mean this feature yields information about the content that causes reports to be liked better.

As we are using a Bayesian normalization of the appreciation score, the amount of votes impacts the predicted value. This means that when predicting the normalized appreciation score, first it has to be predicted if a report will be scored at all, and if so, how many votes the report is likely to receive. In order to better predict this and to account for the temporal influence of a report's publication date, the amount of downloads and the amount of views are added as a feature.

7.1. Predicting votes and appreciation scores

Reiterating figure 5.5 in figure 7.1, most reports that are viewed, are not rated. There are more than 800 seen, but unrated reports and less than 300 rated reports. Trying to predict which reports get rated can be a step towards understanding why some reports do not get rated. Predicting the number of votes can be tackled in two distinct ways. The first way is to see this as a binary classification problem, asking the question: will a report receive any votes or no votes at all? The other option is to tackle it as a more complex problem, not only predicting if a report will be scored, but also how many votes or which average rating the report will receive.

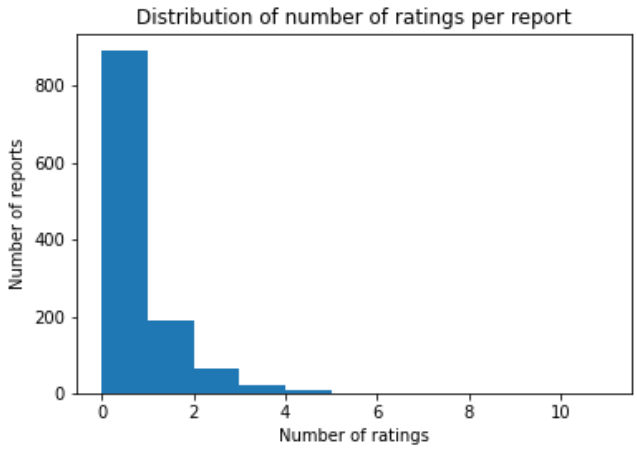


Figure 7.1: The amount of ratings per report for the period 01/08/2020 - 08/12/2020

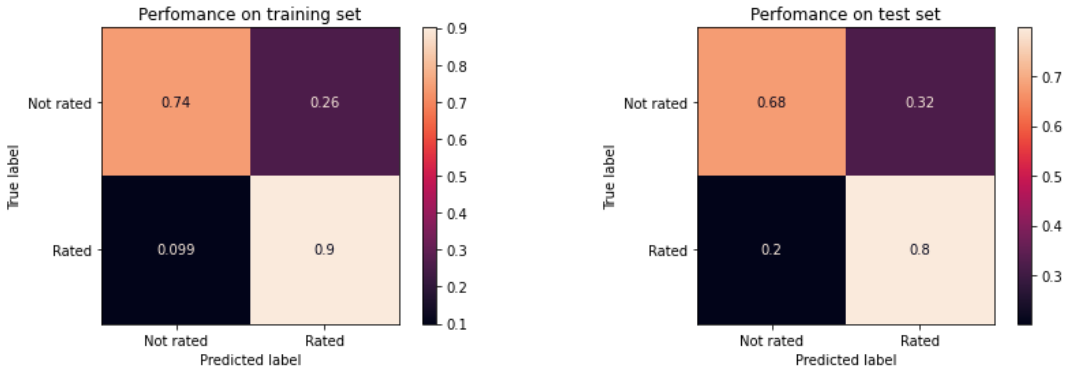
This more complex number-of-vote prediction model could be implemented using two different techniques, the choice is between a zero-inflated model or a two-part model. The difference between these mod-

els is that the zero-inflated model is a so-called mixed model, where different distributions can be combined to deal with the zero-inflated data. In the two-step model, the first step is deciding if the predicted value should be either zero or more than zero by using the binary classifier. If the model decided it should be more than zero, a regression model tries to predict how big the non-zero value should be [80].

7.1.1. Binary vote prediction

Using a Decision Tree Classifier with an F1 evaluation requirement, a final F1 score of 54% and accuracy of 70% are achieved. This model performs slightly better at correctly predicting if a report will be rated rather than predicting if a report stays unrated. All numerical variables are standardized except for the number of views and downloads. A feature has to be present in at least 20 reports to be incorporated, this resulted in 90 remaining features.

In the decision tree shown in figure 7.3, the three top nodes of the decision tree contain the features *words*, *downloads*, and *threat_ransomware*. Suggesting these three nodes are an important factor in deciding if a report will be rated or not.



(a) (b)

Figure 7.2: Binary scoring classification using a Decision Tree Classifier on the train and test set

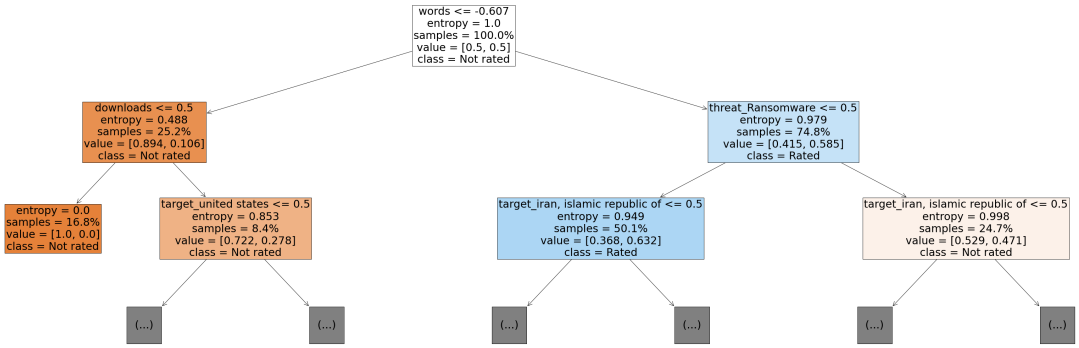


Figure 7.3: The top nodes of the decision tree classifier that resulted in figure 7.2.

7.1.2. Number of votes prediction

The scoring metrics for both number of vote predictors can be found in table 7.1, the matching confusion matrices of the predicted votes versus the actual votes are shown in figure 7.4.

Hurdle model The hurdle model applies the binary model first, only if the binary model predicted that a report receives votes, it tries to predict the total amount of votes. In this case the binary model from section 7.1.1 is used, as already shown in figure 7.2 and confirmed in figure 7.4, the binary model predicts reports with 0 votes correctly 68% of the time.

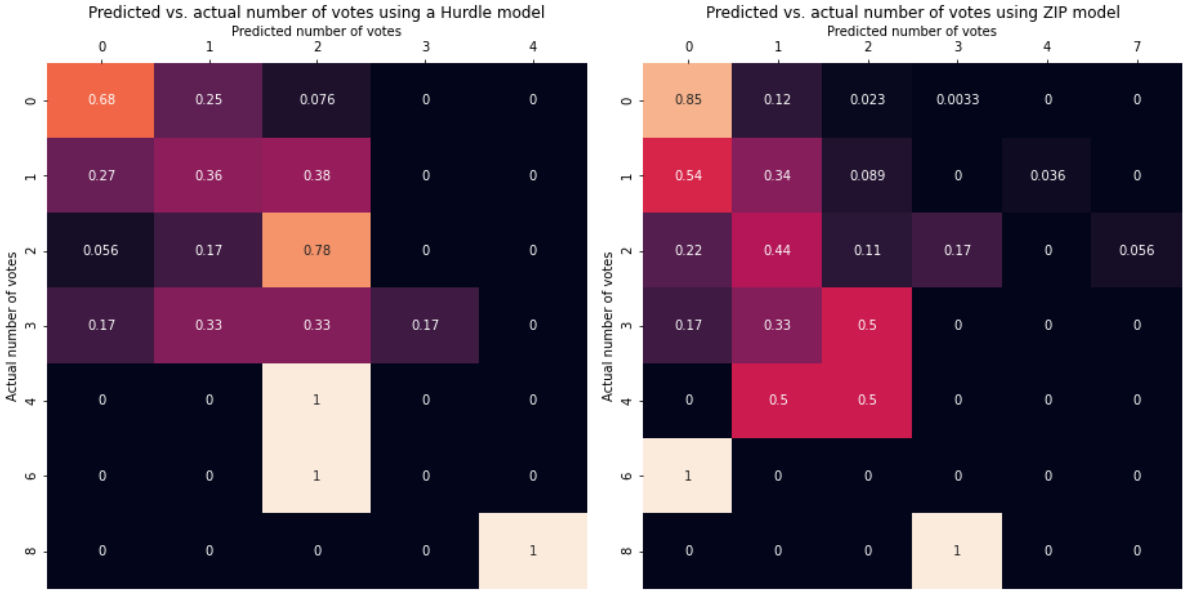
To predict the number of votes in case the binary model predicts that votes were given, an Elasticnet regressor is applied. The optimal ElasticNet regressor for predicting the amount of votes used an alpha of 0.99 and an L1 ratio of 90%. When disregarding mistakes made by the binary model, this regression model achieved an accuracy of 52% and an F1 score of 52%, these metrics are for the case that different amounts of votes are seen as classes. The metric when the results are interpreted as regression is the R^2 score, the R^2 score in this case is 0.28.

Zero-Inflated Poisson model Overall, the Zero-Inflated Poisson (ZIP) model performs better than the hurdle model. The diagonal in the confusion matrix in figure 7.4b can be quite well distinguished, although slightly underestimating the amount of votes. As can be seen in the confusion matrices, the ZIP model predicts more zero votes correctly, but stars to underestimate the values when there are more than zero votes.

This also shows when looking at table 7.1, in all metrics the ZIP model seems to perform better than the hurdle model. However, when looking at correctly predicting the reports with more than zero votes, all metrics seem to show that the hurdle model performs better.

Table 7.1: Predictive performance of the number of vote predictors.

Model	Accuracy	F1 score	R2 score
Hurdle Regression	0.62	0.67	-0.05
Zero-Inflated Poisson	0.72	0.73	0.1
Hurdle Regression > 0	0.52	0.52	0.28
Zero-Inflated Poisson > 0	0.44	0.41	-0.49



(a) (b)

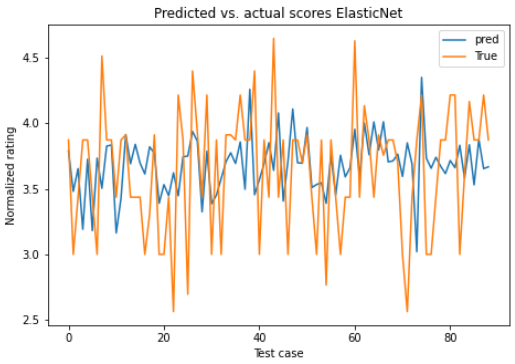
Figure 7.4: Confusion matrices showing the predicted versus the actual amount of votes using an ElasticNet regressor or an Zero-Inflated Poisson model.

7.1.3. Predicting appreciation scores

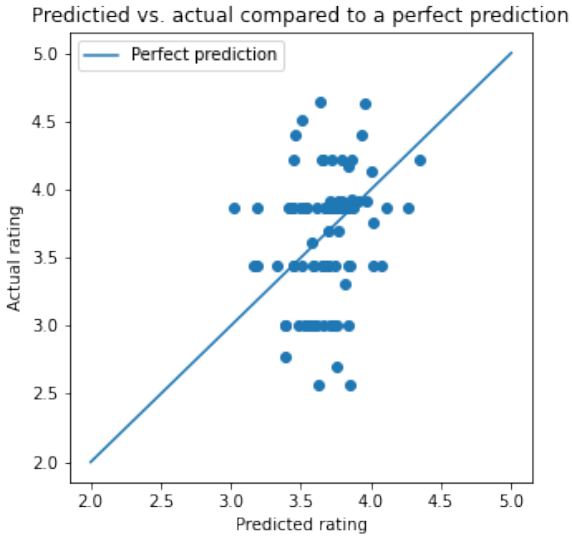
As with predicting the number of votes, we have two different options for predicting the score of a report. One option is to make the assumption all reports receive a score (eventually) and thus ignoring the fact that some reports did not receive any ratings. The second option is a two-step model as introduced in the previous section (7.1), where first a binary model predicts if a report would receive votes and only then tries to predict the normalized score. In this case, a 100% accurate binary predictor is assumed, in order to focus on predicting the actual report scores and not test the prediction power of the binary model, as discussed in section 7.1.1.

Something similar has been done before by [75], were they tried to predict restaurant ratings based on both endogenous as exogenous features. Where they focus on new restaurants with little to no ratings, we focus on report with little to no ratings.

As a proof of concept, an ElasticNet regressor has been applied, the results are shown in figure 7.5. The regression model does perform rather bad, with an R^2 score of -0.035.



(a)



(b)

Figure 7.5: Results of the score prediction using an ElasticNet regressor.

7.2. Predicting report relevancy

In the search for relevant literature how to deal with and what to do with ratings, we encountered literature regarding recommendation systems. All gathered data does (nearly) perfectly fit the blueprint of data necessary to build a recommendation system. In the light of academic interest and academic exercise, some time was spent porting an existing recommendation system [45] to our data.

There are two main techniques when building recommendation systems, one is a more naive method the other more context-aware.

- **Collaborative filtering**: Collaborative filtering is the most naive method. This technique does not try to understand the *why* behind better ratings and more views, but accepts things as they are. Using matrix factorization this algorithm predicts product appreciation scores based on similarities across user behavior. If two users liked similar reports in the past and there is data that one user likes a report but it is unknown if the other user does as well, this system will assume he/she will.

- **Content-Based filtering**: Content-based filtering is a more context-aware method. Based on report features, which can be both textual features as well as measured features (such as our derived report metrics), similarity between reports is calculated. Then, if a user likes a product, this kind of recommendation system will recommend the most similar reports to the already liked one.

These two techniques are not mutually exclusive, as already extensively discussed in literature by [13]. Different recommendation techniques can be mixed and matched, weighting their resulting with common ensemble methods. One of the more simple methods is implemented and tested in this research: One collaborative filtering algorithm and one content-based algorithm will be used and both outcomes will be multiplied with each other to get a final list of report recommendations for each user.

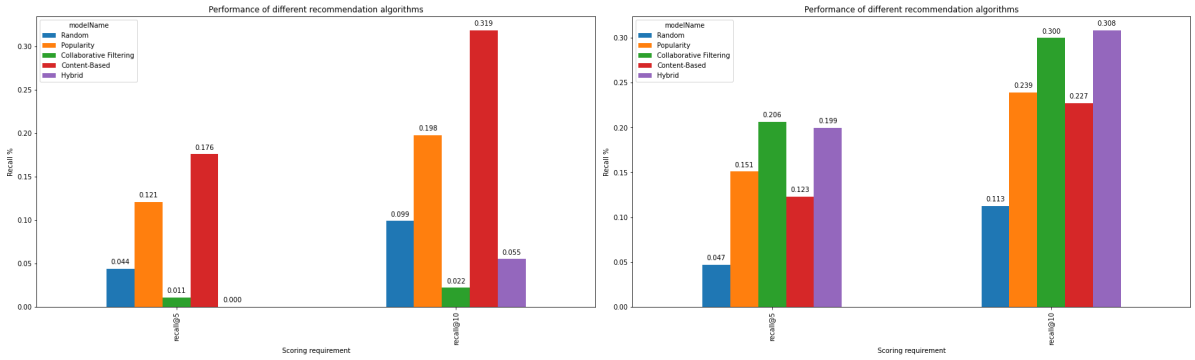
The result of this experiment is shown in figure 7.6. For these results, two different interaction requirements are applied. In the first one, ratings are the main form of interaction. The only added interaction is when a report is viewed between 100 seconds and 10 minutes, in this case a 1 star rating is added by the

user who viewed the report. In the second experiment the view durations are more important and extend the ratings, the exact additional rating per view duration is shown in table 7.2.

When looking at figure 7.6, there is a clear difference between the results of these two experiments. Graph 7.6a shows that the content-based algorithm performs best and the collaborative filtering algorithm performs even worse than random chance. One reason could be that each person has a rather unique set of reports they have scored and thus commonalities between rated reports are not a good predictor which reports will be rated. Graph 7.6b shows the collaborative filtering algorithm performing much better. The generally better outcome can be explained following the same reasoning as before. When view durations are recorded as ratings and thus as interactions, there is a lot more data overall and report interactions are not as unique anymore. Because now the same report is interacted with by a multitude of users, the collaborative filtering algorithm can use the overlap in these interactions to make recommendations. However, it is apparent that the content-based algorithm's prediction power went down. This would suggest that interactions including view durations are harder to predict based on the content.

Table 7.2: Used appreciation scores to represent view durations as interaction in the recommendation system.

View duration	Assigned appreciation score
0 - 10 seconds	0.5
10 - 100 seconds	1.0
100 - 600 seconds	2.0
More than 600 seconds	3.0



(a) (b)

Figure 7.6: Recommendation performance metrics. A recall percentage for the top 5 must be interpreted that if known relevant report for a reader is mixed with 100 random (non-relevant) reports, the achieved percentage is the percentage of times that that report is recommended in the top 5 most recommended reports. Graph 7.6a did assign an interaction with the value of a 1 star rating to reports that are viewed between 100 seconds and 10 minutes, all other view durations didn't receive a score. Graph 7.6b assigned a 0.5 star rating to reports viewed less than 10 seconds, a 1 star rating to reports viewed less than 100 seconds, 2 stars to reports viewed less than 10 minutes and 3 stars to reports viewed more than 10 minutes.

7.3. Summary

In this chapter we tried to answer our fifth sub-research question and performed some experiments related to it. The sub-research question was:

> *5. Can user appreciation scores of Threat Intelligence reports be predicted using quantitative metrics?*

To begin answering this question, first we tried to find out if it can be predicted if a report will receive any votes. Then, we tried to predict the actual number of votes the reports would receive. Only after these two predictions, we tried to predict the normalized appreciation score of reports.

As most reports did not receive any ratings, the prediction of the number of votes can be seen as a zero-inflated prediction problem. To account for this, both a Hurdle model as well as a Zero-Inflated Poisson (ZIP) model were tested. The step of predicting if a report would receive votes was performed using a decision tree

model. This model achieved a reasonable, but not outstanding result with an F1 score of 54% and an accuracy of 70%. From this model we also learned that the number of words and downloads seem important for this prediction, as well as the fact if a report discusses ransomware.

Predicting the number of votes was then tried using both a Elasticnet regression model, as well as a ZIP model. The Hurdle model performed worse when predicting the number of votes including 0, partially caused by poor performance of the binary prediction of the decision tree. The ZIP model did perform a lot better in this classification, as it achieved a 85% correct prediction of reports with 0 votes. When reports with no votes are disregarded, the Hurdle model performed better than the ZIP model. This implies the weakness of the Hurdle model was the binary prediction step rather than the Elasticnet regressor. Next to that, it implies that the ZIP model did overfit on reports receiving zero votes and thus performing worse on the data not containing reports with zero votes. Again, these reasonable results were achieved *including* the information of the amount of downloads and views of a report. In situations where a report is only viewed once, the chance of not receiving a vote is larger than when it has 10 views.

To predict the report scores, again an Elasticnet regressor was used. The results of this experiment are bad, resulting in an R^2 score of -0.035. This means the predicted values are more often than not, in the wrong direction compared to the mean of all scores and thus is a very inaccurate prediction of the overall report scores.

All in all, the results from these prediction experiments suggest that votes in itself can become predicted to some extent, especially when the predictors are helped by presenting the number of downloads and views. The fact that scores can not be predicted well is fully in line with the lack of promising outcomes in chapter 6. Apparently, the 'value' of Threat Intelligence can not be easily distilled from report features and so far unmeasured characteristics are a vital ingredient for a 'good' report.

Finally, two kinds of recommendation systems are implemented. These recommendation systems either use report features or use commonalities in interactions between readers and reports. It shows that when ratings are the only kind of interaction, report features are rather useful in making useful recommendations to users. However, when view durations are also included as interactions, report features become a worse predictor of relevance and the commonalities in users and reports become a better predictor.

From the fact that the recommendation system based on report features performs reasonable using only report features, and the system using commonalities between users and interacted reports performs better when view durations are counted as interaction, new conclusions can be drawn. This likely means that for single users, the report features do give some indication if this user will rate a report or not, especially when put in a sample of 100 random reports, the more relevant report to the user can be picked out.

The fact that when view durations are added as interaction, the collaborative filtering system becomes more powerful, likely means that certain groups of users look at the same reports and thus can be a good indication to base recommendations on. The strong point of this system is that no report ratings are necessary to still give relevant recommendations.

The attentive reader might have noticed that the performance of the content-based system went down in the second experiment, the experiment where view durations are added as different interaction strengths. Although seemingly unintuitive, a hypothesis can be constructed in line with the outcomes of the interviews. The interviewees said that they only tried to score reports relevant for them, and if a report was out of scope they entrusted their colleagues to rate it. This means that the chance that a user only rates reports regarding a topic relevant to them is high. The fact that the content-based system performs rather well, means that the individual topic an employee is interested in can be quite well distilled. The fact that, when report views are added as interaction, this performance goes down and the collaborative filtering performance shoots up, is because employees apparently look at a wider range of reports than relevant for them. This means that the content-based system has a harder time distilling one individual and relevant topic for a user and thus performs worse in predicting and recommending reports that a user is likely to view. The fact that the collaborative filtering system performs much better can mean that although no single topic can be extracted for a user, users do tend to look at similar groups of reports.

8

Conclusion

In this study, we tried to find the answer to multiple (sub-)research questions related to the appreciation and value of Threat Intelligence. Each of them will be discussed here and answered to the greatest extent possible. We start with our first sub-research question:

1. Which quantitative features can be extracted from Threat Intelligence reports?

In chapter 2, multiple quantitative metrics were extracted from literature and summarized. After this, each metric was evaluated as to whether it was suitable for our purpose of using it as predictor for report quality or not. The final set of metrics is presented in table 8.1.

Table 8.1: Summary of the metrics selected in chapter 2

Metric	Description
Volume	Volume describes the total amount of unique indicators that are included in a report and in a TI feed.
Timeliness	Timeliness indicates how early Threat Intelligence sources are with finding and reporting about indicators.
Overlap	Overlap indicates the amount of common indicators between two reports or feeds.
Population	The population metric represents the source and target geolocations. This can be extended with actor populations.
Security threats	The security threats metric tries to describe the topic and used techniques as mentioned in the reports. This metric also includes which targeted industries are discussed in the reports.
Information type	The information type will indicate the report type that is shared.
Subjectivity	Subjectivity, extended with sentiment, represents on the use of language in TI reports.
Report counts	Basic report counts such as the amount of words, pages, and figures.

Overall, this set of metrics is a good representation of most metrics described in literature, where no additional data is necessary to act as a ground-truth. Different metrics such as accuracy and coverage are often discussed in literature as well, but to reliably measure these metrics, known-goods and known-bads are also necessary, data that is lacking in this study.

Other metrics such as security threats, information type, and subjectivity are discussed in only one or a small amount of sources. However, these metrics were included in an attempt to cover the widest possible spectrum of metrics to relate to the appreciation scores.

2. What are analysts' main considerations when scoring Threat Intelligence reports?

As discussed and answered in chapter 4, readers have several considerations when (not) voting on a report. The reason to score reports often is that readers try to score all reports they (partially) read. Three groups of reasons not to score are discovered and distinguished:

- **Paralysis by analysis** This reason can be subdivided into 'lack of time', 'scored too much', and 'did not actually read'.
- **Lack of relevancy** Opening a report that was not relevant to the reader, either due to misleading titles or due to wrong search terms.
- **Just forgot** As using the scoring system is not the goal of opening reports, sometimes readers forget to score a report they have just read.

Next to the reasons (not) to score a report, different reasons exist to give a report either a lower, neutral, or higher rating. Reasons for an overall lower rating consisted of a lack of context or a topic was not discussed deeply enough. Reasons for higher scores boiled down to the reason: "If it takes work out of my hands, the report deserves a high(er) rating".

3. Which patterns can be distinguished in the appreciation scores?

This question was answered in the first part of chapter 5. Some patterns can be clearly distinguished, for other findings only suggestive proof was found.

The pattern that is clearly visible, is that readers generally cast a limited amount of votes. In the two-week period, most readers only scored 1-3 reports. In the 16-week period, most readers casted less than 5 votes. Overall, readers seem inclined to rate reports highly, with the 5-star score being the most given score and the 1-star score being used the least. Another clear pattern that is visible is that reports that are scored are viewed longer than reports that are not scored. There is no clear distinction visible between scores which are viewed for a long time and reports viewed for a shorter time.

Two patterns that are hinted at, but in this case no clear conclusions can be drawn, relate to the scoring pattern between different departments and the 9 dimensions which can be used to explain the votes. From the overall data it seems that department 6 is slightly more critical than the other departments. This idea is slightly reinforced by looking at the distribution *per* star score over the departments, although the idea is rejected when viewing the distribution *of* the star scores over the departments. These ambiguous results suggest that either more data needs to be collected or a deeper analysis has to be performed.

The minor visible pattern of the 9 explanatory dimensions give some indication that certain vendors have specific strengths compared to other vendors. This shows in the fact that all vendors but Vendor 1 received the positive point regarding the dimension 'current', implying no readers notice Vendor 1 reports to be especially relevant for current events. Another reason these data-points to imply uncertain relations, is that the maximum amount of negative dimensions points a vendor received is only 8 out of dozens of reports that could have been evaluated.

4. What is the relation between quantitative metrics & appreciation scores?

Different significant relations are found between quantitative metrics extracted from Threat Intelligence reports and the gathered appreciation scores. These include the amount of indicators in a report, the amount of unique indicators as published by Vendor 3, the amount of non-unique indicators in a report, Vendor 5 publishing about the aviation industry, Vendor 1 publishing 'report'-type reports, subjectivity in a report, Vendor 3 and the amount of words in their reports, and finally the amount of words per page.

Overall, it seems that Vendor 3 is involved in most of the significant interactions, but as discussed in chapter 6, most of these interactions are related to the reports concerning malware. The other point that discounts these findings, is that these findings can be completely different depending on the chosen start- and end-date of the selected data.

From these facts we draw the conclusion that there is close to no relation between quantitative metrics and report appreciation scores. This is in line with expectations based on literature, where the minimal impact of metrics of literature on the value of TI was already implied in interview results [12].

Ideally, more data should be gathered. This would enable validation of the results of the overall data-set by choosing different subsets of the data and finding out if comparable results are achieved.

5. Can user appreciation scores of Threat Intelligence reports be predicted using quantitative metrics?

In the process of predicting report scores, we first came across the challenge of predicting the amount of votes. This is relevant, as the amount of votes impacts the reliability of the average score and the fact that it is of great influence to the normalized appreciation scores. Predicting if a report would receive a score is possible to some extent when using the number of views and downloads, with a decision tree model achieving an accuracy of 70% and an F1 score of 54%. Predicting the actual number of votes reached similar results, with F1 scores between 41 and 73%, and accuracy scores between 44 and 72%.

Predicting the normalized appreciation scores of reports was not possible. With an R^2 score of -0.035 achieved by an Elasticnet regressor, it can be concluded that with the current set of reports, features, and scores it is not possible to correctly predict these scores.

Although appreciation scores could not be predicted, the relevance of reports can be predicted to some extent. Using two different recommendation systems, recall scores around 3 times higher than chance are achieved. From these results we learned that report features lend themselves to predicting which reports will be scored by which user. We also learned that combining ratings with views from similar users can be a good predictor of which reports a user will read.

0. Do Threat Intelligence metrics from literature capture user appreciation?

Finally, we can try and answer our main research question, which has to be answered with a *no*. This answer could be partially caused by the fact that our overall data is not extensive enough and there was a measurement error for a large portion of the measurement period, but it could very well be caused by the misalignment between report metrics and actual user appreciation as well.

The fact that a nearly complete set of metrics from literature does not lead to significantly measurable relations between these metrics and users' appreciation scores, combined with the fact that these metrics do not seem to be a good set of features to predict these scores with, leads us to conclude that Threat Intelligence metrics from literature do not capture user appreciation.

Within this study, other explanations for user appreciation scores, as well as other use cases for report metrics and appreciation scores are discussed. From the interviews we learned that (lack of) depth, content, and context are important identifiers for the quality of a report. Identifiers that, indeed, are not easily translated to quantitative metrics. We also learned that we can use the TI metrics as a predictor for which individual user is likely to rate a report, but without giving any indication how high this rating will be. All these results, the limitations and the implications of this study will be further discussed in chapter 9.

9

Discussion

9.1. Limitations and suggestions for the appreciation scores

One of the main sources of data used in this study was the collection of appreciation scores gathered in the period from 12/08/2020 until 08/12/2020. This unique source enabled research into Threat Intelligence that otherwise would not have been possible. However, gathering real-world results for the first time always brings some uncertainty. Some points on how to improve data gathering and use are discussed here.

9.1.1. Changes in rating system throughout the study

Halfway throughout the project (week of 16-20 November), an initial evaluation of the rating system was conducted, the draft of the advice following from this evaluation can be found in appendix A. After discussing these options with the owner of the TI feeds, both the advice regarding the explanatory buttons and the explanatory text were immediately taken to heart and implemented. This means both an extensive explanatory text was added and nine buttons were added for the reader to indicate what the strong and weak points are of a given report, enabling them to add nuance to their rating. These categories are: currency, context, depth, correctness, readability, relevance, technical, applicability, and uniqueness. With this update also came a bug-fix. As described in chapter 5 as well, the data gathered in the period until the bug-fix is incomplete. As the ratings of people with a longer reading time were disproportionately affected, more ratings of people who fully read reports might be missing. As also shown in chapter 5, there seems to be the same trend that people who rate reports, tend to read reports longer, indicating that a rather significant chunk of ratings might be missing.

With the bug fixed November 17 and the updated rating system implemented November 20, an email update regarding these changes was sent November 20 as well. This means scores collected in the period between August 12 and November 20 only consist of a single star score. Scores collected after November 17 consist either of a single star score or a star score with additional information about that score, with a maximum of three positive and three negative keywords.

How this change in the scoring system exactly changed scoring behavior is unknown. After scoring one or multiple reports with the new scoring system, readers might be better aware of points of interest while reading the report. This could cause them, when scoring a report, to already be aware about the possible keywords to click during the follow-up regarding the keywords. The fact that these keywords did prime the users in some way, was noticeable during the interviews. When asked what the reader (dis)liked about reports, they often used descriptions similar to the introduced keywords.

Explanatory keywords The interviewees were also asked to comment on the 9 added keywords. They generally agreed it was a welcome addition to be able to give a more nuanced opinion rather than just a star score and they agreed that the 9 keywords covered the spectrum. However, one respondent mentioned finding the *currency* keyword hard to relate to a report, as older reports might still be useful. This was contrasted by a respondent mentioning that *currency* was one of the most important categories for a report.

One other respondent mentioned that it was hard to label a report with the *applicability* and *uniqueness* keywords, as the *context* and *depth* keywords were easier to relate and often more relevant.

Finally, one of the respondents noted that the *relevancy* keyword could be misleading. You may expect this keyword to be used when the topic of the report is irrelevant for the reader. However, the respondent said that the keyword was more suitable to use in the case the contents of a report are irrelevant compared to the title of the report (e.g. the title was clickbait).

9.1.2. Lack of scores

One issue that has been mentioned several times before already, is the apparent lack of data to draw actionable conclusions. As most reports receive less than 4 ratings, often only one or two, their final normalized scores are all nearly identical. By applying the Bayesian normalization, the average rating is brought down and the reliability of a high score increases, but this also results in the downside that a normalized score only is effective for reports with more than the average amount of votes per report.

The Bayesian normalization algorithm of Miller [44] aims to calculate a lower bound for a normalized score in combination with the width of the calculated interval. As discussed in section 3.2, this estimation turned out to be extremely conservative, partly caused by the lacking size of our data-set. As an indication of how much data is actually expected when using this algorithm, Miller calculated the expected amount of votes needed in the case of a uniform distribution of votes, a set of votes where there is consensus on the quality, and for a set of polarized votes. He also calculated the necessary votes for different interval widths and credibility levels. Results for all these different calculations are shown in table 9.1. Although these minimum amount of votes are for a conservative calculation with precise interval widths, the amount of necessary votes is about one or two orders of magnitude larger than the amount of votes our average report has.

As counterargument one could suggest that, as time goes on, the amount of gathered data from the rating system will keep increasing. Assuming that rating behavior will improve, the minimum ratings for an interval with a width of 1 and a credibility level of 80% is realistic, any other amount of votes will be near impossible, as that would mean more than 35% of the readers would need to read *and score* the same report. Therefore this algorithm aims for unrealistic numbers of votes given our population of users.

This means we need to ask ourselves the question, is an analysis like the one in this study suitable for a rather limited data-set like ours? As with most things in life: it depends. Our data-set will never be suitable to calculate an exact appreciation score with 99% credibility of the perceived value of a report. On the other hand, with a slight increase in votes per report (let's say 4 to 5 votes per report instead of 1 to 2 votes), strengthened by the ever-increasing number of reports that are being scored, enough data should be available at some point to draw more statistically conclusive lessons, despite not being 99% sure what the exact interval a report's appreciation score is.

Table 9.1: The number of necessary votes to reach a reliable normalized score estimate using the Evan Miller Bayesian Normalization [44].

Width (stars)	Credibility Level	Uniform N	Consensus N	Polarized N
1.0	80%	7	9	20
1.0	90%	16	13	38
1.0	95%	25	16	55
1.0	99%	47	23	100
0.5	80%	46	23	99
0.5	90%	81	31	168
0.5	95%	117	38	240
0.5	99%	206	51	419

9.1.3. Bias in appreciation scores

All appreciation scores in our data-set are gathered from a single organization. Despite the different roles, teams, and departments of the readers, as introduced in chapter 3, the overarching goal of each individual scoring is the same. This fact can be interpreted in two different ways. One more negative interpretation is that this data must be biased. All readers eventually read and rate reports based on its applicability on the goal of the organization, being blinded to other possibilities that could make a report stand out for other organizations. However, the other side of the coin offers a more positive interpretation. As discussed in previous paragraph, despite the best efforts of the organization to gather Threat Intelligence appreciation scores, the data seems to be lacking for our purpose. If the sources for this limited amount of data would be

distributed among different organizations, we would have several, different biases and nuances in the data, which would act as confounding variables. Due to the limited amount of data, each bias of each different organization could impact every single data point, rather than the current situation, where the overall result might be shifted as a whole.

We do not know if our overall result is shifted as a whole, and if so, in which direction it is shifted. In order to make more definitive statements about the bias in this study, other studies at different organizations have to be conducted, to enable an inter-organizational comparison of results.

9.1.4. Suggestions for changes in the rating system

The Threat Intelligence rating system is still in development, which was shown by the intermediate bug-fix and the addition of the explanatory keywords. This also means that, if there are improvements for the rating system, these still can be implemented. Improvements that can be considered fall roughly in two categories, the first are possible alternatives of the star scores, the other are improvements in the way the reports are represented to the user.

Improvements in the star scores Two possible alternatives for the star rating system can be considered. Both of the alternatives are a way to lower the mental load of deciding what exact star rating should be given. If you find a report good, does it deserve four or five stars? If a report was not of use to you, does it deserve, one, two, or three stars? Do you consider not voting? To decrease the effort that a user needs to think about these nuances, either a like/dislike system or a emotion-based system can be considered.

With a like/dislike system, a user only has to evaluate if the report was useful or not useful and cast a fitting vote. The absence of a vote either implies irrelevance or a neutral opinion.

The emotion-based system might need a bit more explanation. In February 2016, Facebook extended their like-button with a collection of emotions to choose from and respond to other people's posts [6]. These emotions consisted of: Love, Haha, Wow, Sad, and Angry, with corresponding emoticons. The idea behind these emotions and their emoticons was that users have a wider range of options when responding to posts, as not every post deserve a 'like' (e.g. think of news about disasters). LinkedIn followed suit in 2019 by extending their like option with: Celebrate, Love, Insightful, and Curious.

Next to being a convenience to the user by enabling to express themselves better, it allows for deeper analyses as well [79]. Although expressed as a concern when this data is in the 'wrong' hands and used for monetary gains, it opens many analytical doors when used for a benign purpose. The fact that emotions are more natural to people, hopefully would result in faster and more votes. Although not all emotions used at Facebook and LinkedIn might align with the emotions TI reports can raise, curious and insightful are already two good contenders.

Improvements in the report presentation The current report and rating system is, as described, dependent on user search- and alert keywords. Although users are generally experts in their topic area, this does mean they are fully dependent on their own search behavior and their assumption about which keywords vendors include in their reports. Both the data gathered of users opening and reading reports, as well as the voting data, can be used to improve the collection of reports that is shown to users.

The first option is to have a list of most popular reports independent of user and report-content, very comparable to the 'popularity recommender' from section 7.2. Two lists can be constructed, one existing of the five most viewed reports in the previous 7 days, the other list can contain the five reports that received either the most or the highest scores in the last 7 days. The most-viewed list will likely consist of reports covering larger and newsworthy events, whereas the list with reports that received the highest scores could consist of hidden gems.

Next to these recommendations, the gathered data can be used to make content and user-aware recommendations, again using systems comparable to the ones constructed in section 7.2. The collaborative filtering method, which creates profiles of users with similar interests, can be used to construct a list of reports that could be of interest as well. This way users with similar topics of interest but different search behavior can eventually be linked. Additionally, if the difference in search behavior results in one of the users missing a relevant report, they can still notified of this overlooked report. The content-based recommendation system would be most relevant for an addition where each report is linked to similar reports. Ideally, this enables users to find one relevant report, at which point the relevant report list will guide a user through all necessary reports.

All described possibilities are not yet (fully) implemented and not as easy as I might make them sound: hundreds of technical professionals at Amazon try to achieve similar functionality. Still, these past paragraphs are included as a point on the horizon as motivation and inspiration, with the final goal of increasing the ease of report consumption for all readers.

9.2. Statistical tests and predicting appreciation scores

The format of the gathered appreciation scores did provide additional challenges regarding which statistical tests to apply. The fact that the ratings resulted in a sparse matrix of rated reports and users, means the data is perfect for recommendation systems but a lot harder to use statistical tests on. Each of the users rated a subset of reports, where (nearly) none of the subsets are completely equal. The difficulty is that the data cannot be handled as if measurements are independent, some people do generally rate lower than other people. However, in order to use the data for a repeated measures test, the subsets of rated reports need to be the same between users.

Different solutions for comparable data have been discussed in literature [21], but these methods are not well established (yet), meaning implementation and use could take up a significant portion of time. Derrick et al. also discuss two common solutions for comparable data. The first is to remove all unpaired observations. The fact that the person who rated the most reports still only rated around 10% of the total reports published, causes most observations to be unpaired and thus makes this an extreme solution. The other possible solution is to remove all paired observations, However, this means that the reports that are scored most should be removed, despite these reports probably conveying most meaning about what people are looking for in reports.

These reasons caused us to handle the data as if it contained independent measures. Despite missing out on statistical rigor that repeated measures tests give, this way of handling data enables us to use all data rather than only a small subsection of data. For future research, looking deeper into statistical tests for partially overlapping samples is advisable.

The use of Bayes normalization of the star scores possibly helped make the average star scores more voter-agnostic. If a report was only scored once, the Bayes score wasn't impacted much even if it was a lenient scorer or a more strict scorer, as a single lenient score of 5 and stricter score of 4 resulted in a normalized score of 3.8 and 3.4 respectively. Then, as a report receives more scores, the final score is less impacted by more lenient or strict readers as well.

9.2.1. Suggestions for the statistical tests

As discussed previously, the outcomes of the statistical tests were dependent on which subset of the data was used. We were helped by chance to be able to draw this conclusion, as the data was not received at once but in batches. Because of this part the analyses were created and tested on a subset of the data before being applied to all data. I would like to suggest to systematically perform experiments this way, rather than accidentally.

When applying the statistical tests to check for significant relations between report features and their respective scores, some kind of cross-validation or random sampling could be applied. In this way, the dataset as a whole can be tested for significant relations, rather than the result being dependent on a limited set of measurements.

9.2.2. Suggestions for predicting appreciation scores

After having performed the prediction of the appreciation scores and the number of report votes, two different suggestions come to mind. The first is to make changes in the target of the score prediction. In this study, the exact normalized value of a report was the target to predict. In future research, this could be altered to try and predict a distribution of scores. This way, both the uncertainty of a small amount of votes can be incorporated, as well as represent a more realistic depiction of the actual way of scoring.

Depending on the goal of the study, the information available to the algorithm predicting the number of votes could be decreased. In this study, the algorithm knew the number of views and downloads, already giving an upper bound to the number of possible votes. Rather than giving this information explicitly, similar information could be given more implicitly by only providing the publication date for example, enforcing the algorithm to deduce and predict the possible number of views from the age of the report and the expected popularity of the topic and publisher. This decrease in available information would be especially suitable to study the impact of the report topic and publisher on the number of votes. When the only statistic of interest is expected views compared to the actual views, the current methodology suffices.

9.3. Business implications: The use and future of Treat Intelligence

Threat Intelligence is acquired by dozens of organizations in the Netherlands and hundreds of organizations around the world. Undoubtedly, some organizations have set up evaluation pipelines to streamline this process of acquisition. However, many organizations make acquisition decisions on the basis of informal processes and evaluation criteria [12]. Here we will shortly discuss lessons customers of TI vendors can learn from this study.

9.3.1. The economics of TI

There is no way to calculate the exact value of a TI source. Current prices are based on the time of the year, the expectations of both parties (e.g. vendor and customer), and most of all, the result of the negotiations [12]. It is not possible to put a value on a single indicator, nor to calculate the price of a single report.

Pawliński and Kompanek [56] already state that a framework is lacking for making TI acquisition decisions. There is missing information regarding the quality, the scope, and the value versus the cost of a TI feed. Despite not being complete, this study described each of these pieces of information to some extent, enabling better deliberation about which TI feed to acquire. Especially as TI vendors generally give some reports as a sample, metrics and ratings for the sample time period can be measured, gathered, and compared to the owned TI data. This should give an idea of the quality of the reports, all metrics about the indicators, and some information regarding their scope (with regards to actors and target/source countries).

Table 9.2 shows some of these metrics in one line with their respective price-ranges per year. Although we did conclude already that the report metrics are not a good predictor of the quality of a report, this table reinforces that idea: the price, perceived quality, and amount of indicators are not always correlated with each other.

Table 9.2: Price for commercial Threat Intelligence feeds. Metric and scoring information from 01/08/2020 - 08/12/2020. **Boldfaced** values are the highest in the column.

	Price (€)	Reports	Rated reports	Indicators	IP	MD5	Host	URL	25%-Mean Normalized Score-75%
Vendor 1	75K-150K	25-75	25-75	2500-5K	250-500	500-1K	1K-2.5K	250-500	3.43-3.82-3.91
Vendor 2	150-250K	**500-1K**	**150-250**	**10K-25K**	250-500	**5K-10K**	**5K-10K**	250-500	3.44-3.66-3.87
Vendor 3	**250K-500K**	250-500	25-75	5K-10K	**500-1K**	1K-2.5K	1K-2.5K	**1K-2.5K**	3.00-3.47-3.87
Vendor 4	75K-150K	25-75	10-25	2500-5K	10-25	1K-2.5K	1K-2.5K	500-1K	**3.62-3.81-4.06**
Vendor 5	150K-250K	75-150	25-75	5K-10K	75-150	1K-2.5K	1K-2.5K	**1K-2.5K**	3.61-3.77-3.91

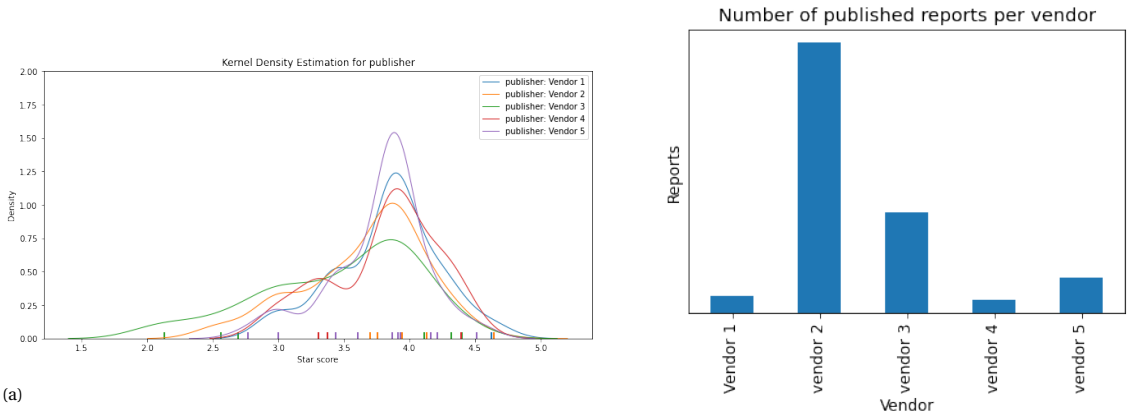
9.3.2. The negotiation perspective

The overall distribution of the normalized votes and the amount of published reports per vendor in figure 9.1 could be helpful when negotiating contracts with a TI vendor. Selectively choosing other vendors that publish more or are better appreciated for a lower price can give you the upper hand when negotiating with a vendor. This way either you can enforce a lower annual price per vendor or enforce the vendors to enhance their quality or quantity.

Next to utilizing this competitive element, the gathered keywords associated with each rating will prove to be a great source of feedback. These keywords enable an aggregated view of the TI users and give specific feedback to the vendors. When a vendor hears that their reports are very deep and technical, but are lacking contextually, it is very easy to instruct their analysts to add a paragraph of contextual information to their reports. Likewise, for a keywords like 'uniqueness', if a TI vendor receives a lot of positive votes in this dimension, one can interpret this as the vendor being of added value.

9.3.3. Summary

Although the overarching conclusion of this study is that TI metrics do not capture user appreciation, several other lessons and uses of this study can be distilled. Purely having the quantitative metrics gives buyers a trump in the negotiations with a TI vendor. Next to that, collecting user appreciation scores systematically is good for two reasons. Again in negotiations with a TI vendor (remember, prices can go down by as much as a factor of 10 [12]), and in a more mutually beneficial way in the shape of giving detailed feedback to a vendor



(a)

(b)

Figure 9.1: Kernel density estimation for the star scores per publisher and published reports per publisher.

how they can improve their offering. Finally, the exciting results from section 7.2 showed the possibilities of using the combination of user appreciation scores, user views, and report metrics. As suggested by users of the TI reports, having relevant suggestions when reading a report can decrease the time they themselves have to spend searching and could decrease the amount of irrelevant reports showing up when searching specific keywords.

9.4. Academic contribution: Threat Intelligence metrics

9.4.1. Relevance

The relevance metric as introduced by [56] was mainly a suggestion for a metric, not one that has been operationalized yet. Due to this lack of operationalization, we can make the suggestion to alter the definition of this metric, without stepping on any researcher's toes. We suggest that the relevance metric should be used in the future as a metric on both report- and feed-level and is operationalized as the report appreciation scores of the readers of an organization.

9.4.2. A mismatch between literature and application

Using TI metrics as found in literature and as described in chapter 2, we encountered several limitations. We found a limited amount of significant relationships and no predictive power was extracted from the features to predict report scores. This raises the question of whether the TI metrics as described in literature are the appropriate values to measure from TI. Even if the values are the correct values, we can ask if measuring and calculating metrics from TI is as important as literature seems to suggest.

The argument that we are not measuring the appropriate metrics immediately grows from the observations as discussed in section 2.2.1, the term 'Threat Intelligence' in literature seems to be largely re-branded from what originally were called 'blacklists'. This dynamic is something that shows in the fact that most, if not all, TI literature has a large amount of references to blacklist literature, a shortcoming this text is also afflicted by.

This directly follows from the fact that the term 'Threat Intelligence' is not protected in any way. This means that in-depth publicized reports, such as the 2013 Mandiant report [40], in-depth commercial reports, such as those in described in this thesis, but also public and private blacklists all can be called Threat Intelligence. Using the term 'Threat Intelligence' like this, as an umbrella term rather than a term for information fitting specified requirements, makes it difficult for both practitioners and researchers to have a constructive dialogue about it.

Hereby, we give the advice to the research community, to take a step back and evaluate what to label TI and when other terms to describe information could or should be used. When a common understanding and agreement is reached about what we give the label of TI, new and common efforts can be made to understand TI and find the 'x-factor' of what makes good Threat Intelligence so good.

The first step could be to enforce contextual information. As the term 'intelligence' implies, raw numbers, hashes, IPs, etc. do not offer any *intelligence* but only a source for more work.

Second, consider that metrics are less important and less related to the value of TI than historically assumed in literature. This idea has been introduced before by [12] and seems to be confirmed by this study. Quantitative metrics are intuitive to measure and an attractive tool to use when researching the value of TI. If it is not measurable, how can we express the value of TI? Unfortunately while I do not have an answer to this question, I do have piece of wisdom to offer. Do not fall for the McNamara fallacy, which goes: *'if it cannot be measured, it is not important'* [52]. If it cannot be measured, especially when discussing the value of Threat Intelligence, it might be especially important.

9.5. Reflection on the research process

The process of writing a master thesis seemed a daunting task from the moment I first heard of it. Where a bachelor's thesis (at least mine) was a straight-forward process of finding a project, getting an assignment with specified scope, and then setting off to research one specific part of an algorithm, the master thesis process always seemed to be a whole different ball game. During my studies, both in my bachelor's and master's, I've seen friends during their master thesis follow the recipe of eat, sleep, study, repeat. Topped of with a swirl of stress. One of the causes seemed to be the complete freedom most students operate in during their thesis. Every university course has predefined learning goals and assignments. But during a thesis the only learning goal is to execute the academic process from start to end, a learning goal focused on the means and no predefined end, except the three bullets points of requirements from the faculty and another four bullet points from the degree program. Daunting or not, a master's degree should be finished with a thesis and thus I embarked on this journey.

Linus Pauling once said: "The best way to have a good idea is to have lots of ideas". Taking this advice to heart, I started out the thesis process with many ideas, a lot of these ideas did not survive until the end of the project. For starters, I immediately fell for the McNamara fallacy, as introduced just before. I had the illusion that if I made a visualization dashboard for Threat Intelligence reports and metrics, the intrinsic value of Threat Intelligence would magically emerge. As I just finished up my literature chapter and thus read about the added value of TI metrics for two weeks straight, I had put myself in a rabbit hole that was hard to crawl out of. This effect was strengthened by the fact that there was a delay in receiving the data and thus the only remaining option I saw was to start building the visualization dashboard. After another two weeks I received both the report data and the appreciation scores. This enabled me to take a step back and evaluate the added value of the dashboard. Although I still believe that the dashboard has great potential for a superficial comparison of TI sources, without more contextual factors, the information is insufficient to base actual decisions on.

Another idea is to compare publicized TI reports with their commercial and private counterpart. This idea includes both a quantitative and a qualitative option. In the quantitative option, the indicators extracted from publicized reports would be compared to the indicators included in the metadata of a private report. This could provide information about what percentage of indicators is publicized and how much earlier private reports publish information compared to public reports. Next to this, a qualitative comparison between a private report and its publicized counterpart could provide information about the level of certainty and gravity necessary before a vendor decides that data should be made public. Although I still think this idea is a very interesting option to pursue, the unique availability of the appreciation scores caused me to drift away from this idea during this research.

The last idea I want to discuss here was the idea to give an explicit ranking of TI vendors. This ranking would be based on both quantitative metrics, as well as the appreciation scores. However, this idea would enforce us to put a valuation on the quantitative metrics, preferably independent of the appreciation scores as these would already have their own impact on the ranking. A valuation of the quantitative metrics would likely result in an arbitrary system to give some kind of score to different metric levels. As the chance of creating a ranking system that was dependent on too many assumptions seemed high, this idea was let go as well.

Linus Pauling's quote served me well here, I ended up with an idea resulting in this thesis and being able to bring the field of TI a bit further. In addition to that, I began work on another idea of mine, resulting in the prototype of a visualization dashboard that could possibly serve a purpose in the future if it is extended and refined. Finally, the idea of performing a mixed-methods study on the relation between publicized and private reports still sounds like a good option, with many different possibilities for the future.

Professors are often another and unlimited source of ideas. As foretold by different peers, the ideas and suggestions might keep on coming, even until the day before your graduation. While this dynamic can be stressful to some people, it worked surprisingly well with my preferred way of working. Shortly said, I like to know a little about a lot, rather than know a lot about a little. When ideas keep flowing in, shortly exploring each of the ideas comes naturally to me. In the end, this led to chapter 7, which could be trivially described as a collection of prototypes and experiments.

The possibility to integrate my preferred way of working into my thesis was something I did not expect. As many theses are in depth on one single topic, I automatically assumed I had to do this as well. These in depth theses often require lengthy introductory and background chapters, chapter where necessary content and 'fluff' are sometimes hard to distinguish. This was one of my initial fears, that these lengthy chapters were a necessity and there was the unspoken minimum amount of pages of at least 100, preferably 150 pages. During the process, both these unspoken expectations appeared to be completely false, and besides that, I exceeded my own expectation by (nearly) reaching this amount of pages without actually trying to do so.

Bibliography

[1] Top Alexa domains. URL https://www.alexa.com.

[2] Elemendar. URL https://elemendar.com/. Accessed on 27.11.2020.

[3] Marktplaats lanceert sterrenwaardering voor kopers en verkopers, Oct 2018. URL https://www.emerce.nl/nieuws/marktplaats-lanceert-sterrenwaardering-kopers-verkopers.

[4] Alan Akbik, Duncan Blythe, and Roland Vollgraf. Contextual String Embeddings for Sequence Labeling. In *COLING 2018, 27th International Conference on Computational Linguistics*, pages 1638–1649, 2018.

[5] User: 'armbues'. The 'ioc-parser' project, 2017. URL https://github.com/armbues/ioc_parser.

[6] Ismail Badache and Mohand Boughanem. Emotional social signals for search ranking. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '17, page 1053–1056, New York, NY, USA, 2017. Association for Computing Machinery. URL https://doi.org/10.1145/3077136.3080718.

[7] Aaron Beuhring and Kyle Salous. Beyond blacklisting: Cyberdefense in the era of advanced persistent threats. *Security & Privacy, IEEE*, 12:90–93, 09 2014. URL https://ieeexplore.ieee.org/document/6924678.

[8] David J. Bianco. The pyramid of pain, 2014. URL https://rvasec.com/slides/2014/Bianco_Pyramid%20of%20Pain.pdf.

[9] Bird, Steven, Loper Edward, and Ewan Klein. *Natural Language Processing with Python*. O'Reilly Media Inc., 2009.

[10] Christopher M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag, Berlin, Heidelberg, 2006. ISBN 0387310738.

[11] BlueVoyant. Threat intelligence services: Turning raw data into actionable intelligence. URL https://www.bluevoyant.com/threat-intelligence-services. Accessed on 16.12.2020.

[12] Xander Bouwman, Harm Griffioen, Jelle Egbers, Christian Doerr, Bram Klievink, and Michel van Eeten. A different cup of TI? The added value of commercial threat intelligence. In *29th USENIX Security Symposium (USENIX Security 20)*, pages 433–450. USENIX Association, August 2020. URL https://www.usenix.org/conference/usenixsecurity20/presentation/bouwman.

[13] Robin Burke. Hybrid recommender systems: Survey and experiments. *User Modeling and User-Adapted Interaction*, 12, 11 2002. doi: 10.1023/A:1021240730564.

[14] Tom Burt. Microsoft report shows increasing sophistication of cyber threats, Dec 2020. URL https://blogs.microsoft.com/on-the-issues/2020/09/29/microsoft-digital-defense-report-cyber-threats/.

[15] Robert J. Cabin and Randall J. Mitchell. To bonferroni or not to bonferroni: When and how are the questions. *Bulletin of the Ecological Society of America*, 81(3):246–248, 2000. URL https://www.jstor.org/stable/20168454.

[16] Mung Chiang. *Networked Life: 20 Questions and Answers*. Cambridge University Press, USA, 2012. ISBN 1107024943.

[17] Cyber Threat Coalition. *Background image COVID-19 Cyber Threat Coalition*. URL https://www.cyberthreatcoalition.org/. Accessed on 16.09.2020.

[18] The MITRE Corporation. Mitre att&ck®, Jan 2021. URL https://attack.mitre.org/.

[19] Crowdstrike. Crowdstrike falcon prevent™ next-generation antivirus, Mar 2020. URL https://www.crowdstrike.com/wp-content/uploads/2020/03/Falcon-Prevent-FINAL.pdf. Accessed on 16.12.2020.

[20] User: 'deanmalmgren'. The 'textract' project, 2019. URL https://github.com/deanmalmgren/textract.

[21] Ben Derrick, Deirdre Toher, and Paul White. How to compare the means of two samples that include paired observations and independent observations: A companion to derrick, russ, toher and white (2017). *The Quantitative Methods for Psychology*, 13:120–126, 02 2017. doi: 10.20982/tqmp.13.2.p120.

[22] Merriam-Webster Dictionary. Dictionary entry: Feature. URL https://www.merriam-webster.com/dictionary/feature. Accessed on 26.01.2020.

[23] EclecticIQ. Threat intelligence powered cybersecurity. URL https://www.eclecticiq.com/. Accessed on 27.11.2020.

[24] Anja Feldmann, Oliver Gasser, Franziska Lichtblau, Enric Pujol, Ingmar Poese, Christoph Dietzel, Daniel Wagner, Matthias Wichtlhuber, Juan Tapiador, and Narseo Vallina-Rodriguez. The lockdown effect. *Proceedings of the ACM Internet Measurement Conference*, Oct 2020. URL http://dx.doi.org/10.1145/3419394.3423658.

[25] David H. Freedman. *Wrong: why experts* keep failing us - and how to know when not to trust them.* Little, Brown, 2010.

[26] Sarah Gordon and Richard Ford. On the definition and classification of cybercrime. *Journal in Computer Virology*, 2:13–20, 08 2006. doi: 10.1007/s11416-006-0015-z.

[27] Harm Griffioen, Tim M. Booij, and Christian Doerr. Quality Evaluation of Cyber Threat Intelligence Feeds. In *International Conference on Applied Cryptography and Network Security (ACNS)*, 2020.

[28] Winston Haynes. *Bonferroni Correction*, pages 154–154. Springer New York, New York, NY, 2013. URL https://doi.org/10.1007/978-1-4419-9863-7_1213.

[29] Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979. URL http://www.jstor.org/stable/4615733.

[30] Andrew Kalafut, Abhinav Acharya, and Minaxi Gupta. A study of malware in peer-to-peer networks. In *Proceedings of the 6th ACM SIGCOMM Conference on Internet Measurement*, IMC '06, page 327–332, New York, NY, USA, 2006. Association for Computing Machinery. URL https://doi.org/10.1145/1177080.1177124.

[31] User: 'kbandla'. The 'APTnotes' project, 2020. URL https://github.com/aptnotes/data.

[32] Brian Keith, Exequiel Fuentes, and Claudio Meneses. A hybrid approach for sentiment analysis applied to paper reviews. 2017. URL https://sentic.net/wisdom2017fuentes.pdf.

[33] Simon Kemp. Digital 2020: Global digital overview, Feb 2021. URL https://datareportal.com/reports/digital-2020-global-digital-overview.

[34] William H. Kruskal and W. Allen Wallis. Use of ranks in one-criterion variance analysis. *Journal of the American Statistical Association*, 47(260):583–621, 1952. URL http://www.jstor.org/stable/2280779.

[35] Georgy Kucherin, Igor Kuznetsov, and Costin Raiu. Sunburst backdoor – code overlaps with kazuar, Jan 2021. URL https://securelist.com/sunburst-backdoor-kazuar/99981/.

[36] Marc Kührer, Christian Rossow, and Thorsten Holz. Paint It Black: Evaluating the Effectiveness of Malware Blacklists. In Angelos Stavrou, Herbert Bos, and Georgios Portokalidis, editors, *Research in Attacks, Intrusions and Defenses - 17th International Symposium, RAID 2014, Gothenburg, Sweden, September 17-19, 2014. Proceedings*, volume 8688 of *Lecture Notes in Computer Science*, pages 1–21. Springer, 2014. URL https://doi.org/10.1007/978-3-319-11379-1_1.

[37] Vector Guo Li, Matthew Dunn, Paul Pearce, Damon McCoy, Geoffrey M. Voelker, and Stefan Savage. Reading the Tea leaves: A Comparative Analysis of Threat Intelligence. In *28th USENIX Security Symposium (USENIX Security 19)*, pages 851–867, Santa Clara, CA, August 2019. USENIX Association. URL https://www.usenix.org/conference/usenixsecurity19/presentation/li.

[38] Steven Loria. textblob documentation, 2018.

[39] Kevin Mandia. Global intrusion campaign leverages software supply chain compromise, Dec 2020. URL https://www.fireeye.com/blog/products-and-services/2020/12/global-intrusion-campaign-leverages-software-supply-chain-compromise.html.

[40] Mandiant. Apt1: Esposing one of china's cyber espionage units. 2013. URL https://www.fireeye.com/content/dam/fireeye-www/pdfs/mandiant-apt1-report.pdf.

[41] Louis Marinos and Andreas Sfakianakis. Enisa threat landscape, 2012.

[42] Roland Meier, Cornelia Scherrer, David Gugelmann, Vincent Lenders, and Laurent Vanbever. FeedRank: A tamper- resistant method for the ranking of cyber threat intelligence feeds. In *2018 10th International Conference on Cyber Conflict (CyCon)*, pages 321–344, Tallinn, May 2018. IEEE. URL https://ieeexplore.ieee.org/document/8405024/.

[43] Leigh Metcalf and Jonathan M. Spring. Blacklist Ecosystem Analysis: Spanning Jan 2012 to Jun 2014. In *Proceedings of the 2nd ACM Workshop on Information Sharing and Collaborative Security - WISCS '15*, pages 13–22, Denver, Colorado, USA, 2015. ACM Press. URL http://dl.acm.org/citation.cfm?doid=2808128.2808129.

[44] Evan Miller. Ranking Items With Star Ratings, 2014. URL https://www.evanmiller.org/ranking-items-with-star-ratings.html.

[45] Gabriel Moreira. Recommender systems in python 101, Dec 2019. URL https://www.kaggle.com/gspmoreira/recommender-systems-in-python-101.

[46] 'mstamy2'. The 'PyPDF2' project, 2016. URL https://github.com/mstamy2/PyPDF2.

[47] Jon Munshaw. New partnership brings together talos' visibility with ir's unmatched response capabilities, Nov 2019. URL https://blog.talosintelligence.com/2019/11/new-partnership-brings-together-talos.html.

[48] Juniper Networks. Unmatched security intelligence detects and blocks advanced threats faster, Jun 2019. URL https://www.juniper.net/assets/uk/en/local/pdf/solutionbriefs/3510613-en.pdf.

[49] Umara Noor, Zahid Anwar, Tehmina Amjad, and Kim-Kwang Raymond Choo. A machine learning-based fintech cyber threat attribution framework using high-level indicators of compromise. *Future Generation Computer Systems*, 96:227 – 242, 2019. URL http://www.sciencedirect.com/science/article/pii/S0167739X18326141.

[50] Umara Noor, Zahid Anwar, Jörn Altmann, and Zahid Rashid. Customer-oriented ranking of cyber threat intelligence service providers. *Electronic Commerce Research and Applications*, 41:100976, May 2020. URL https://linkinghub.elsevier.com/retrieve/pii/S1567422320300533.

[51] Bank of England. CBEST Intelligence-Led Testing: Understanding Cyber Threat Intelligence Operations, 2016. URL https://www.bankofengland.co.uk/-/media/boe/files/financial-sector-continuity/understanding-cyber-threat-intelligence-operations.pdf.

[52] S O'Mahony. Medicine and the mcnamara fallacy. *The journal of the Royal College of Physicians of Edinburgh*, 47(3):281—287, September 2017. URL https://doi.org/10.4997/JRCPE.2017.315.

[53] Mike O'Malley. Concerned about nation state cyberattacks? here's how to protect your organization, Mar 2020. URL https://www.securitymagazine.com/articles/91889-concerned-about-nation-state-cyberattacks-heres-how-to-protect-your-organization.

[54] Kris Oosthoek and Christian Doerr. Cyber Threat Intelligence: A Product Without a Process? *International Journal of Intelligence and CounterIntelligence*, 0(0):1–16, 2020. URL https://www.tandfonline.com/doi/full/10.1080/08850607.2020.1780062.

[55] Bo Pang and Lillian Lee. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 115–124, Ann Arbor, Michigan, June 2005. URL https://www.aclweb.org/anthology/p05-1015.

[56] Paweł Pawliński and Andrew Kompanek. Evaluating Threat Intelligence Feeds FIRST Technical Colloquium for Threat Intelligence, 2016. URL https://www.first.org/resources/papers/munich2016/kompanek-pawlinski-evaluating-threat-ntelligence-feeds.pdf.

[57] Thomas V Perneger. What's wrong with bonferroni adjustments. *BMJ*, 316(7139):1236–1238, 1998. URL https://www.bmj.com/content/316/7139/1236.

[58] Alex Pinto and Kyle Maxwell. Measuring the IQ of your Threat Intelligence Feeds (#tiqtest), August 2014. URL https://www.slideshare.net/AlexandrePinto10/defcon-22-measuring-the.

[59] Christian Rossow, Christian J. Dietrich, Herbert Bos, Lorenzo Cavallaro, Maarten van Steen, Felix C. Freiling, and Norbert Pohlmann. Sandnet: Network Traffic Analysis of Malicious Software. In *Proceedings of the First Workshop on Building Analaysis Datasets and Gathering Experience Returns for Security*, pages 78–88, New York, NY, USA, 2011. URL https://doi.org.10.1145/1978672.198682. event-place: Salzburg, Austria.

[60] Florian Roth. The Newcomer's Guide to Cyber Threat Actor Naming, May 2018. URL https://medium.com/@cyb3rops/the-newcomers-guide-to-cyber-threat-actor-naming-7428e18ee263.

[61] Dan Sabbagh. Suspected russian cyber-attack growing in scale, microsoft warns, Dec 2020. URL https://www.theguardian.com/technology/2020/dec/18/suspected-russian-cyber-attack-growing-in-scale-microsoft-warns.

[62] S. Samtani, S. Yu, H. Zhu, M. Patton, and H. Chen. Identifying scada vulnerabilities using passive and active vulnerability assessment techniques. In *2016 IEEE Conference on Intelligence and Security Informatics (ISI)*, pages 25–30, 2016. doi: 10.1109/ISI.2016.7745438.

[63] Thomas Schaberreiter, Veronika Kupfersberger, Konstantinos Rantos, Arnolnt Spyros, Alexandros Papanikolaou, Christos Ilioudis, and Gerald Quirchmayr. A Quantitative Evaluation of Trust in the Quality of Cyber Threat Intelligence Sources. In *Proceedings of the 14th International Conference on Availability, Reliability and Security - ARES '19*, pages 1–10, Canterbury, CA, United Kingdom, 2019. ACM Press. URL http://dl.acm.org/citation.cfm?doid=3339252.3342112.

[64] Daniel Schlette, Fabian Böhm, Marco Caselli, and Günther Pernul. Measuring and visualizing cyber threat intelligence quality. *International Journal of Information Security*, March 2020. URL http://link.springer.com/10.1007/s10207-020-00490-y.

[65] SciPy. Scipy kruskal-wallis. URL https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.kruskal.html. Accessed on 21.10.2020.

[66] Steve Sheng, Brad Wardman, Gary Warner, Lorrie Faith Cranor, Jason Hong, and Chengshan Zhang. An Empirical Analysis of Phishing Blacklists. *Proceedings of the Conference on Email and Anti-Spam*, page 10, 2009.

[67] Sushant Sinha, Michael Bailey, and Farnam Jahanian. Shades of grey: On the effectiveness of reputation-based blacklists. In *2008 3rd International Conference on Malicious and Unwanted Software (MALWARE)*, pages 57–64, Alexandria, VA, USA, October 2008. IEEE. URL http://ieeexplore.ieee.org/document/4690858/.

[68] Symantec. Dragonfly: Cyberespionage attacks against energy suppliers. 2014. URL https://symantec-blogs.broadcom.com/blogs/threat-intelligence/dragonfly-energy-section-cyber-attacks.

[69] James Tarala and Kelli K. Tarala. Open threat taxonomy (version 1.1), Oct 2015. URL https://www.auditscripts.com/resources/open_threat_taxonomy_v1.1a.pdf.

[70] Wenyi Tay, Xiuzhen Zhang, and Sarvnaz Karimi. Beyond mean rating: Probabilistic aggregation of star ratings based on helpfulness. *Journal of the Association for Information Science and Technology*, 71(7): 784–799, 2020. URL https://asistdl.onlinelibrary.wiley.com/doi/abs/10.1002/asi.24297.

[71] Amos Tversky and Daniel Kahneman. Availability: A heuristic for judging frequency and probability. *Cognitive Psychology*, 5(2):207 – 232, 1973. URL http://www.sciencedirect.com/science/article/pii/0010028573900339.

[72] UCSD network telescope. URL https://www.caida.org/projects/network_telescope.

[73] Tyler J Van der Weele and Maya B Mathur. Some desirable properties of the bonferroni correction: Is the bonferroni correction really so bad? *American Journal of Epidemiology*, 188(3):617–618, November 2018. URL https://www.doi.org/10.1093/aje/kwy250.

[74] Serge van Ginderachter. Who led the digital transformation in your company?, Mar 2020. URL https://twitter.com/svg/status/1244540212866400256.

[75] Xiaochen Wang, Yanyan Shen, and Yanmin Zhu. A data driven approach to predicting rating scores for new restaurants. In Jun Zhu and Ichiro Takeuchi, editors, *Proceedings of The 10th Asian Conference on Machine Learning*, volume 95 of *Proceedings of Machine Learning Research*, pages 678–693. PMLR, 14–16 Nov 2018. URL http://proceedings.mlr.press/v95/wang18c.html.

[76] Waybackmachine and IMDb. IMDb Top 250, 2015. URL https://web.archive.org/web20150501202812/htttp://www.imdb.com/chart/top.

[77] Christina Meilee Williams, Rahul Chaturvedi, and Krishnan Chakravarthy. Cybersecurity risks in a pandemic. *J Med Internet Res*, 22(9):e23692, Sep 2020. URL http://www.ncbi.nlm.nih.gov/pubmed/32897869.

[78] Josh Zelonis. The Forrester New Wave™: External Threat Intelligence Services, Q3 2018. page 25, 2018.

[79] Shoshana Zuboff. *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power*. 1st edition, 2018. ISBN 1610395697.

[80] Alain F. Zuur, Elena N. Ieno, Neil J. Walked, Anatoly A. Saveliev, and Graham M. Smith. *Zero-Truncated and Zero-Inflated Models for Count Data*, pages 261–293. Springer New York, New York, NY, 2009. URL https://www.doi.org/10.1007/978-0-387-87458-6_11.

A

Initial advice regarding voting system

As the current rating system is 1-dimensional and the rating requirements are not fully clear, both nuance and clarity are lost in each rating. In order to improve the meaning that can be distilled from each rating, I have come up with the following alterations to the existing rating system.

Add explanatory buttons

Inspired by the rating system that the Dutch advertising website “Marktplaats.nl" implemented in September 2018 [3], the star ratings can be extended with buttons indicating strong and weak points of a report (reseller/buyer on Marktplaats). The options to rate on Markplaats exist of things as realistic offering, value for money, and response speed. For Threat Intelligence reports, categories such as presence of contextual information, technical detail, and relevance of the topic can be used.

Add another star rating

Another option to extend the rating system is to add another star rating. Where the first rating would express the quality of the report, the new rating expresses the relevance of the report. This would give readers the ability to express a more nuanced view of the reports, as they will be able to rate a report’s quality at 5 stars, but the relevance at 1 star, knowing this won’t give the wrong impression when all scores are evaluated.

Change the explanatory text

Finally, the most simple option to change the rating system, is to change the text that explains how to rate the report. Currently the text indicates that you are rating the quality of the report, where 5 stars is an outstanding report and 1 star is a useless report. This should be changed in a text that indicates the relevance of the report, effectively scoring the quality of a report as well. This assumes that well-written, extensive reports that are not useful to any research are still not relevant and thus might be unnecessary.

B

Interview script

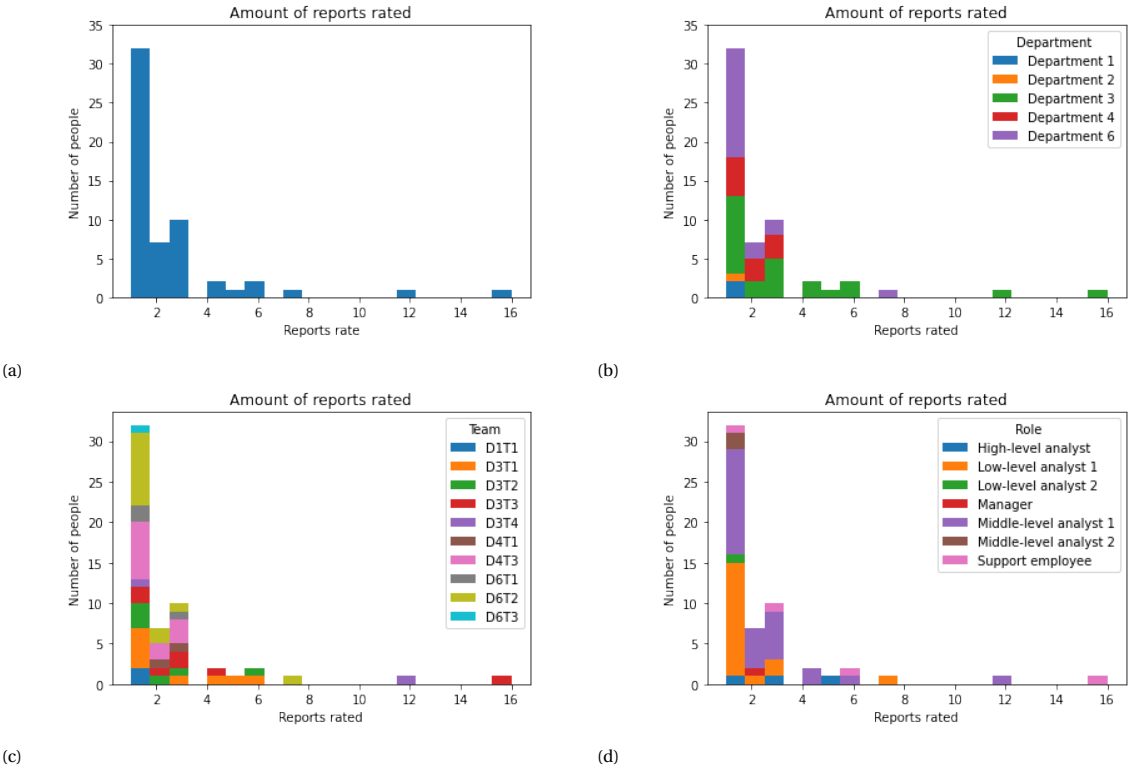
- Which kind of reports do you tend to look at? How do you encounter the reports you read?
- If you rate, why do you decide to rate a report?
- When you read a report and do not rate it, why is that?
- Can you think of a report that you rated high lately, why is that?
- Can you think of a report that you rated low lately, why is that?
- What do you appreciate in a report, anything else than previously mentioned?
- What is less useful in a report, what do you dislike, anything else than you previously mentioned?
- Did you ever score based on one of the following relations, or do you recognize them when hearing them?
  - Vendor 2 reporting about the military (industry) as target
  - Vendor 3 reporting about malware:
  - Vendor 5 reporting about Ukraine (as target)
- How do you like the introduced categories for adding nuance to a rating?

C

Results: Extensive data description

C.1. Star ratings November 20 - December 8

C.1.1. Votes per person

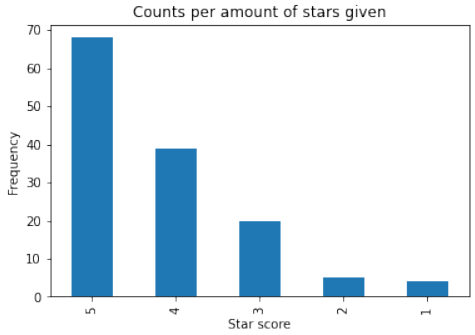


(a) (b)

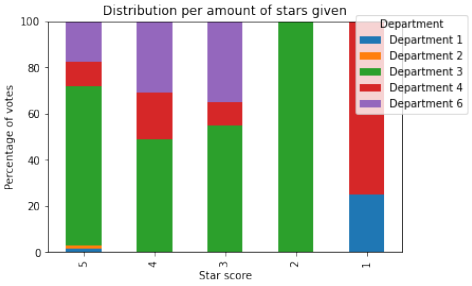
(c) (d)

Figure C.1: The rating count per person and its distribution across different organizational units is shown in these figures. It is shown over the total dataset (C.1a), per department (C.1b), team (5.1a), and role (5.1b).

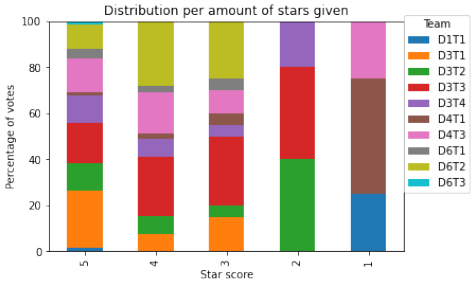
C.1.2. Distribution per star rating



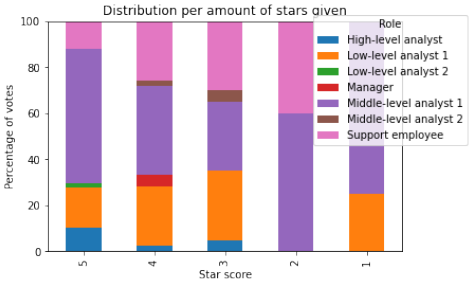
(a)



(b)



(c)



(d)

Figure C.2: The counts per star score and its distribution across different organizational units is shown in these figures. It is shown over the total dataset (5.2a), per department (5.2b), team (C.2c), and role (C.2d).

C.1.3. Distribution per organizational unit

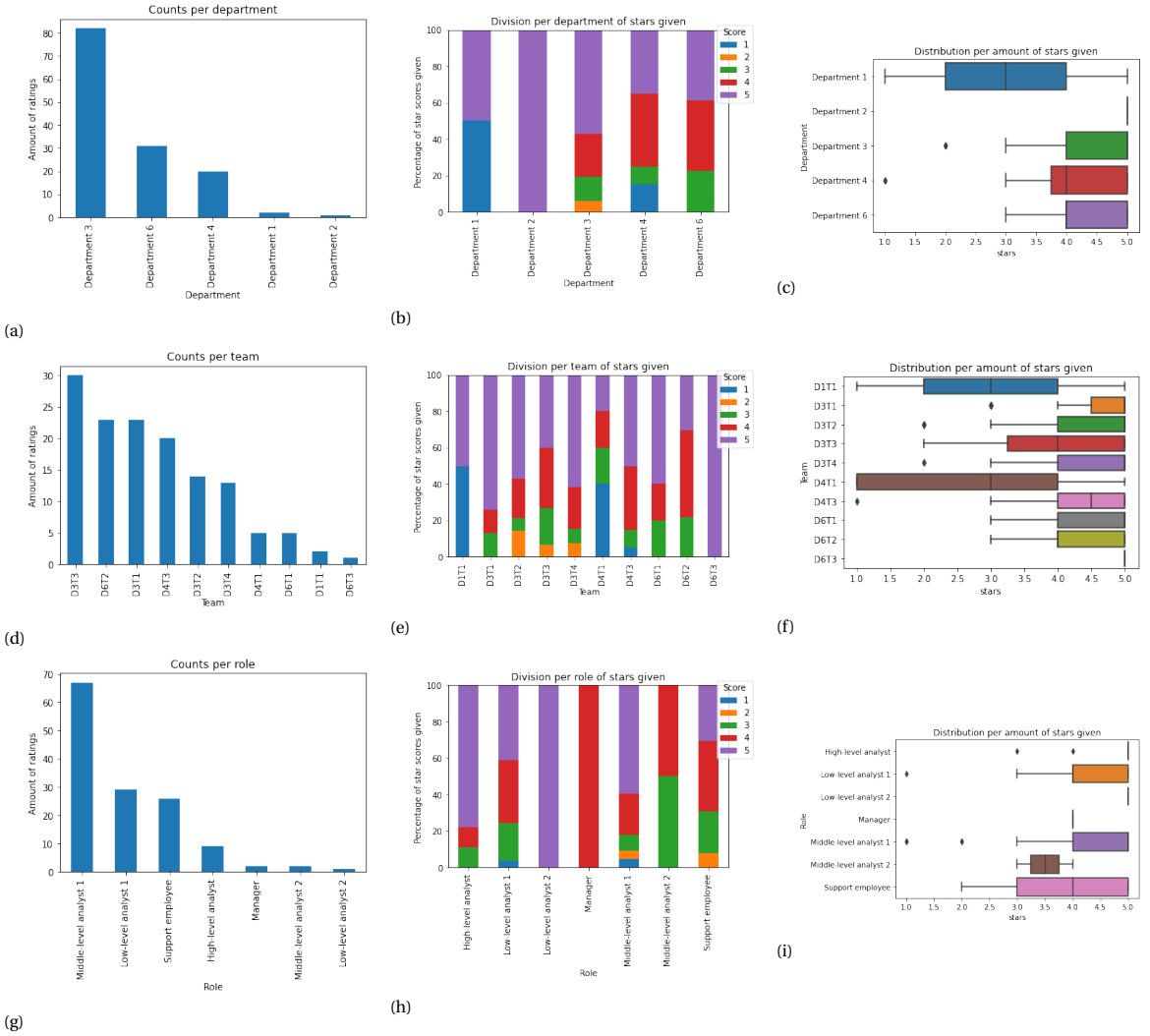
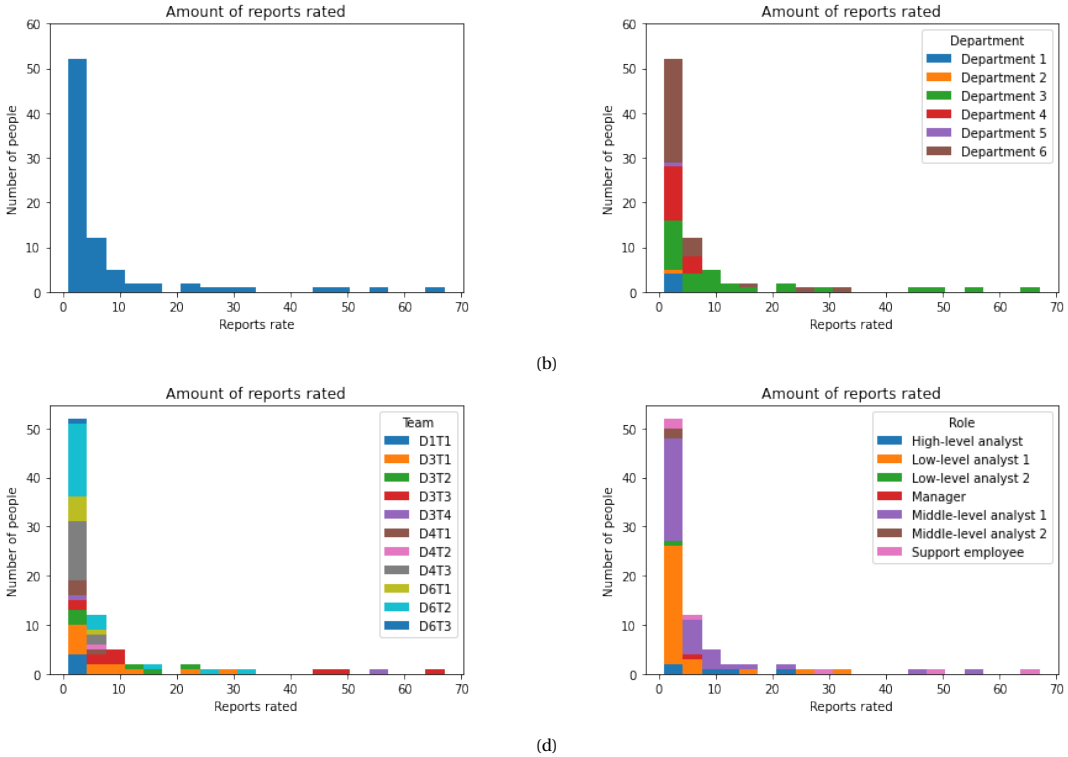


Figure C.3: The distribution of the star score across different organizational units is shown in these figures. It is shown per department (5.3a, C.3b, 5.3b), team (C.3d, C.3e, C.3f), and role (C.3g, C.3h, C.3i).

C.2. Star ratings October 12 - December 8

C.2.1. Votes per person

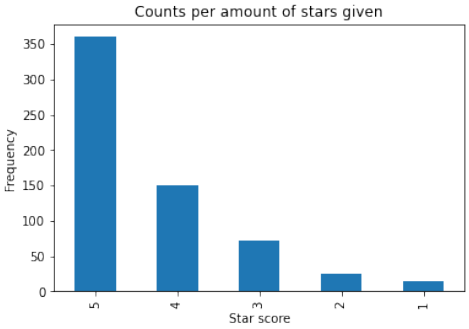


(a) (b)

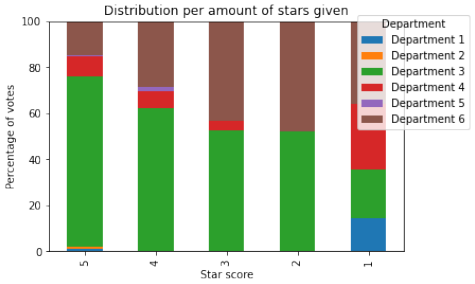
(c) (d)

Figure C.4: The rating count per person and its distribution across different organizational units is shown in these figures. It is shown over the total dataset (C.4a), per department (C.4b), team (5.8a), and role (5.8b).

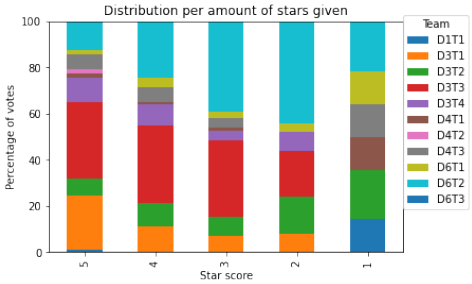
C.2.2. Distribution per star rating



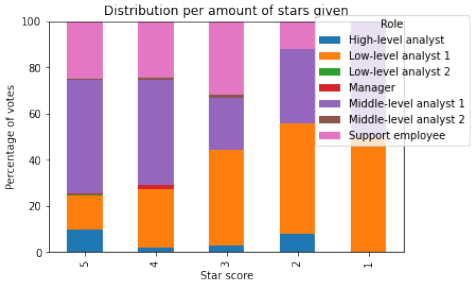
(a)



(b)



(c)



(d)

Figure C.5: The counts per star score and its distribution across different organizational units is shown in these figures. It is shown over the total dataset (5.9a), per department (5.9b), team (C.5c), and role (C.5d).

C.2.3. Distribution per organizational unit

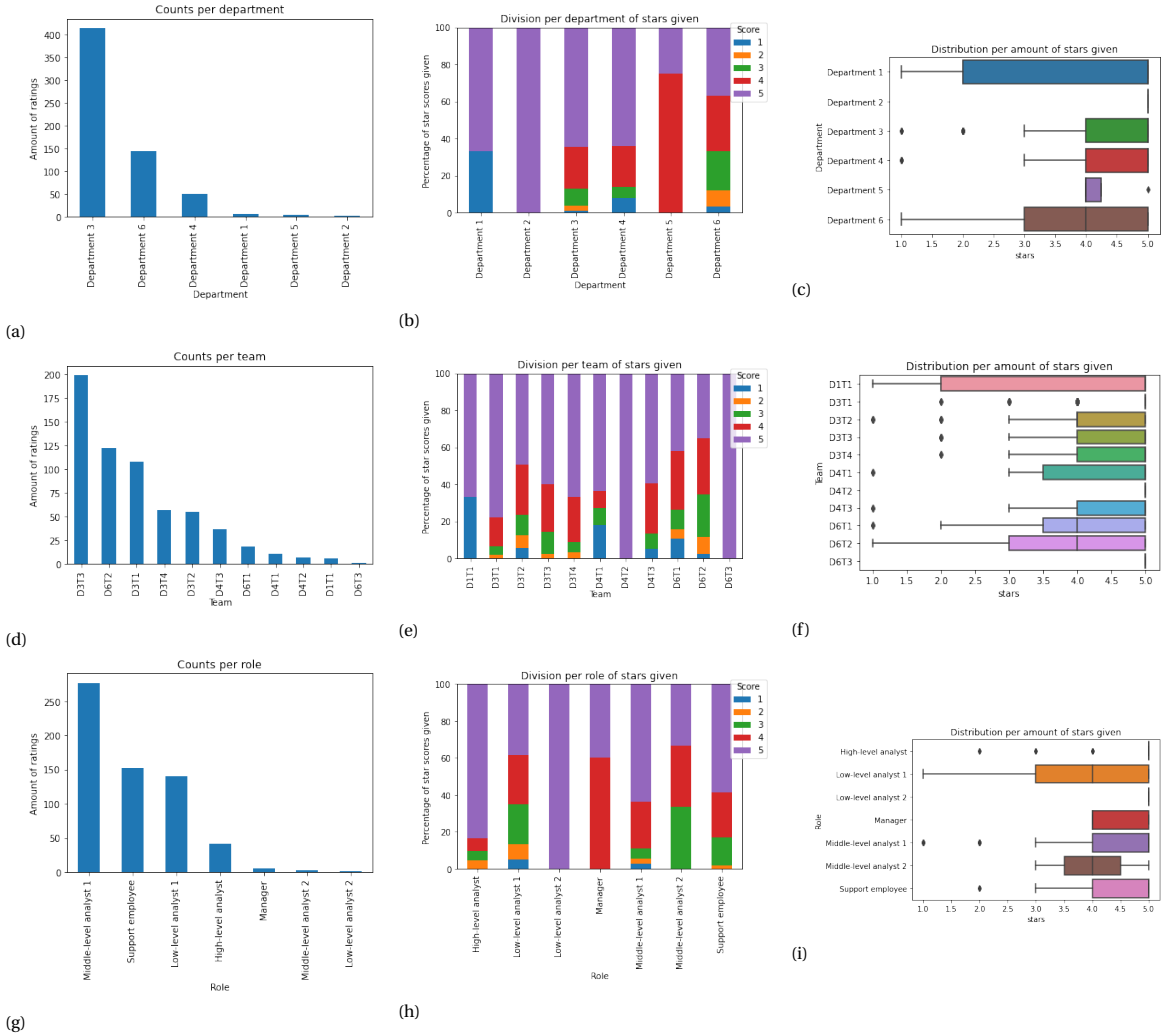


Figure C.6: The distribution of the star score across different organizational units is shown in these figures. It is shown per department (5.10a, C.6b, 5.10b), team (C.6d, C.6e, C.6f), and role (C.6g, C.6h, C.6i).

D

Literature review

Table D.1: Literature study represented in a table

Metric	Paper	Used name	TI or blacklist	Paid / shared / free	Numerical / Categorical	Ground Truth / more data necessary	Unit of Analysis	Definition
Volume	Kuhrer et al. 2014		(malware) blacklist	free / paid AV	Numerical	No	Feed	Observed content
	Pinto & Maxwell 2014	Novelty	TI	free	Numerical	No	Feed	Content per day
	Metcalf and Spring 2015	List Counts	Blacklist	free	Numerical	No	Feed	Unique number of indicators in a feed at a given time point
	Pawlinski and Kompanek 2016		TI	Shared	Numerical	No	Feed	How much data is provided
	Meier et al 2018	Completeness	TI	free	Numerical‘	No	Feed	How much data does a feed contribute to the total amount of indicators
	Meier et al 2018	Size / Insertation rate	TI	free	Numerical	No	Feed	How much data is provided and how much is added / deleted per time unit
	Li et al. 2019		TI	free / paid feed	Numerical	No	Feed	Total amount of indicators published over a measurement interval. Rate is related but is the amount of indicators on a daily basis
Accuracy	Sinha et al. 2008		(spam) blacklist	free	Numerical	Yes (manual labeling)	Indicator / Feed	False negative / positive
	Sheng et al. 2009	False positives	(phishing) blacklist	free / shared	Numerical	Yes (manual labeling)	Indicator / Feed	
	Kuhrer et al. 2014		(malware) blacklist	free	Numerical	Yes (parked + sinkhole techniques)	Indicator / Feed	False positives based on parked domains and sinkholes
	Pawlinski and Kompanek 2016		TI	Shared	Numerical	Yes (simple whitelist)	Indicator / Feed	Use simple white list to compare which values don’t belong there and calculate false positive ratio

	Meier et al 2018		TI	free	Numerical	No	Indicator / Feed	Weighted edges represent if indicators are confirmed by other feeds
	Li et al. 2019		TI	free / paid feed	Numerical	Yes (Unroutable IPs, Alexa, Content Distributor Networks (CDN))	Indicator / Feed	Items that should definitely not be in a feed / False positive
Coverage	Sinha et al. 2008	True positive	(spam) blacklist	free	Numerical	Yes (manual labeling)	Indicator / Feed	True positive
	Sheng et al. 2009		(phishing) blacklist	free / shared	Numerical	Yes (manual labeling)	Indicator / Feed	(No. of phish appearing on blacklist) / (Total phish – phish that were taken down)
	Kuhrer et al. 2014	Completeness	(malware) blacklist	free	Numerical	Yes (automatic detection of malware domains)	Indicator / Feed	Based on malware pattern a 'ground-truth' is identified and checked in the blacklists
	Li et al. 2019		TI	free / paid feed	Numerical	Yes (Network Telescope Data)	Indicator / Feed	Items that should highly likely be in a feed / True positive
	Schlette et al. 2020	Relevance	TI	Unkwown (CTI vault)	Numerical (0-1)	Yes (Own data)	Report / Feed	By calculating which amount of indicators are already seen by the customer, you can calculate how relevant a report is. (customer (publisher / object)) / (publisher / object)
Timeliness	Kuhrer et al. 2014	Reaction time	(malware) blacklist	free	Numerical	Yes (SANDNET)	Indicator	Difference between occurrence in SANDNET and occurrence on blacklist
	Metcalf and Spring 2015	Time Series	Blacklist	free	Numerical	No	Indicator / Feed	Timeliness of both lists

	Metcalf and Spring 2015	Following	Blacklist	free	Categorical	No	Indicator / Feed	Null-hypothesis that lists don't follow each other. Then perform analysis and use formulas to try to debunk that null hypothesis
	Pawlinski and Kompanek 2016		TI	Shared	Numerical	Yes (when compared to own detection)	Indicator	Delay = t(report) – t(detect)
	Meier et al 2018	Speed	TI	free	Categorical (Binary)	No	Indicator / Feed	The direction of the edge defines which feed is more often the fastest with reporting on new indicators.
	Li et al. 2019	Latency	TI	free / paid feed	Numerical	No	Indicator / Feed	Relative delay compared to the first feed that reported on that indicator
	Griffioen et al. 2020		TI	free	Numerical	Yes (Netflow data)	Indicator / Feed	Relative delay in appearing in a feed compared to first traffic measured
	Schlette et al. 2020	Currency and volatility	TI	Unkwown (CTI vault)	Numerical			
Overlap	Pinto & Maxwell 2014		TI	free	Numerical	No	Feed	Inter-report overlap, percentage wise
	Metcalf and Spring 2015	Expanded List Intersection	Blacklist	free	Numerical	Yes (Domain-Indicator and Indicator-domain tool)	Indicator / Feed	Expand all existing indicators on a blacklist to have a more complete image of overlapping values
	Metcalf and Spring 2015	Time Series	Blacklist	free	Numerical	No	Feed	Total intersection size, percentage overlaps
	Metcalf and Spring 2015	Pairwise Intersection Counts	Blacklist	free	Numerical	No	Feed	Pairwise intersection / overlap per feed
	Metcalf and Spring 2015	Reverse counts	Blacklist	free	Numerical	No	Indicator	In how many lists is an indicator

	Li et al. 2019	Differential contribution	TI	free / paid feed	Numerical	No	Feed	The amount of indicators that are in one feed but not in an other as ratio of the first feed; |A B| / |A|
	Li et al. 2019	Exclusive contribution	TI	free / paid feed	Numerical	No	Feed	The amount of indicators that are in one feed and not in any other as ratio of the first feed; |A B(B !=A) | / |A|
	Griffioen et al. 2020	Originality	TI	free	Numerical	No	Feed	How much overlap is there between feeds. Can also say something about a set of feeds (large originality is low overlap overall)
Population	Pinto & Maxwell 2014		TI	free	Categorical	Yes (ASN and GeoIP)	Indicator	Resolve indicators to their origin with Autonomous System Number (ASN) and GeoIP databases
	Metcalf and Spring 2015	IP-Based Characterization	Blacklist	free	Numerical & Categorical	Yes (ASN)	Indicator / Feed	Three different ASN related metrics; ASN counts, Top 5 countries by ASN, and ASN Intersection by Count
Domain-based characterization	Metcalf and Spring 2015		Blacklist	free	Numerical & Categorical	Yes (DNS data)	Indicator / Feed	Google safe browsing blacklist, resolving active domains per blacklist, and top 5 name server that serve largest number of domains per blacklist
Relevance	Pawlinski and Kompanek 2016		TI	Shared	Categorical	Yes (analysts' queries)	Indicator / Report / Feed	Should we care?
Sensitivity	Griffioen et al. 2020		TI	free	Numerical	Yes (Netflow data, GeoIP)	Indicator / Feed	What is the threshold before an indicator is incorporated into a feed, is this dependant on geographical location
Impact	Griffioen et al. 2020		TI	free	Numerical	Yes (Active domain / IP crawl)	Indicator	How many collateral damage is present if domains / IPs get blocked? Are all IPs listed that are necessary
Information sharing model	Noor et al. 2020		TI	Paid	Categorical	Yes (sharing model)	Feed	What kind of model is used to share the data?
Sharing mechanism	Noor et al. 2020		TI	Paid	Categorical	Yes (sharing mechanism)	Feed	What is the mechanism, different subtypes present

Security services	Noor et al. 2020		TI	Paid	Categorical	Yes (different services)	Feed	Long list of different services
Information Source	Noor et al. 2020		TI	Paid	Categorical	Yes (Information sources)	Feed	Self generated, third parties,customer collected, aggregated?
	Pawlinski and Kompanek 2016	Detection method	TI	Shared	-	-	-	How the information was obtained?
Security Threats	Pawlinski and Kompanek 2016	Vantage	TI	Shared	-	-	-	What is the focus of collection?
	Noor et al. 2020		TI	Paid	Categorical	No	Feed	Different threats
Information Type	Noor et al. 2020		TI	Paid	Categorical	No	Feed	Different types of info is contained
Sharing Frequency	Pinto & Maxwell 2014	Novelty	TI	free	Numerical	No	Feed / Report	Content per day
	Meier et al 2018	Update rate	TI	free	Numerical	No	Feed	At which rate are feeds updated
	Noor et al. 2020		TI	Paid	Categorical	No	Feed	Different units of time
Cost	Noor et al. 2020		TI	Paid	Numerical / Categorical	No	Feed	A numerical value which can also be caught in different categories
Appropriate amount of data	Schlette et al. 2020		TI	Unkwown (CTI vault)	Numerical	No	Feed	Is the only metric on report level from Schlette et al. It takes into account graph theory and the general connectedness of the report. As described by themselves still simplistic and requires future work
Representational Consistency	Schlette et al. 2020		TI	Unkwown (CTI vault)	Categorical (binary)	No	Indicator / report	This allows the user to both check for syntactic accuracy and concise representation, but also to add other requirements that are checked. Only if all requirements are met, it receives a score of 1.

Reputation	Schlette et al. 2020		TI	Unkwown (CTI vault)	Categorical (5 star rating)	No	Indicator / report	Is split into the reputation of the publisher (provenance) and reputation of the the dataset (believability). Both can be scored on a 5 star scale.
Concise representation	Pawlinski and Kompanek 2016	Completeness	TI	Shared	-	-	-	Do we have enough details? Not enough information
	Schlette et al. 2020		TI	Unkwown (CTI vault)	Categorical (binary)	No	Indicator / Report / Feed	Is both related to the syntactic accuracy, as well as the adherence to the schema
Objectivity	Schlette et al. 2020		TI	Unkwown (CTI vault)	Numerical (0-1)	No	Indicator / Report	Calculates the objectivity of an attribute and uses that to calculate the objectivity of the object. Can be achieved with sentiment analysis and NLP
Schema completeness	Schlette et al. 2020		TI	Unkwown (CTI vault)	Numerical (0-1)	No	Indicator	If the indicators are reported based on the standard schema, it can be checked if the attributes adhere to this schema and if optional values are filled in, and a value can be calculated
Syntactic Accuracy	Pawlinski and Kompanek 2016	Ingestiblity	TI	Shared	-	-	-	Can we process it?
	Schlette et al. 2020		TI	Unkwown (CTI vault)	Numerical (0-1)	No	Indicator	As with schema completeness for optional attributes, syntactic accuracy is about required attributes of a STIX object.