Computers & Security 138 (2024) 103670

Contents lists available at ScienceDirect

Computers & Security

journal homepage: www.elsevier.com/locate/cose

Deception in double extortion ransomware attacks: An analysis of profitability and credibility

Tom Meurs [a,*], Edward Cartwright [b], Anna Cartwright [c], Marianne Junger [a], Abhishta Abhishta [a]

[a] University of Twente, Enschede, Netherlands
[b] De Montfort University, Leicester, Leicestershire, United Kingdom
[c] Oxford Brookes University, Oxford, Oxfordshire, United Kingdom

ARTICLE INFO

Keywords:
Ransomware
Data exfiltration
Signaling game
Game theory
Information asymmetry
Cybercrime

ABSTRACT

Ransomware attacks have evolved with criminals using double extortion schemes, where they signal data exfiltration to inflate ransom demands. This development is further complicated by information asymmetry, where victims are compelled to respond to ambiguous and often deceptive signals from attackers. This study explores the complex interactions between criminals and victims during ransomware attacks, especially focusing on how data exfiltration is communicated. We use a signaling game to understand the strategies both parties use when dealing with uncertain information. We identify five distinct equilibria, each characterized by the criminals' varied approaches to signaling data exfiltration, influenced by the strategic parameters inherent in each attack scenario. Calibrating the game parameters with real-world like values, we identify the most probable equilibrium, offering insights into anticipated ransom amounts and corresponding payoffs for both victims and criminals. Our findings suggest criminals are likely to claim data exfiltration, true or not, highlighting a strategic advantage for intensifying attack efforts. The study underscores the need for victims' caution towards criminals' claims and highlights the unintended consequences of policies making false claims costlier for criminals.

1. Introduction

Crypto-ransomware attacks globally are a growing concern for our society. In the United States alone, an estimated 1,981 schools, 290 hospitals, 105 local governments and 44 universities and colleges were hit by crypto-ransomware attacks in 2022 (Palmer, 2023). Crypto-ransomware (or ransomware) is a malicious software that aims to encrypt the files of victims (Gonzalez and Hayajneh, 2017). Typically, if victims lack adequate backups, they can only regain access to those files after paying a ransom to the criminals (Laszka et al., 2017). Given that many victims are willing to pay a ransom, these malicious software have proved highly lucrative for criminals and provide a viable 'business model'.

In recent years, criminals have complemented file encryption with the stealing of sensitive data. This sensitive data might contain personally identifiable information of employees and/or customers, intellectual property or legal information. The criminals then threaten to publish the sensitive data or sell it to competitors if the victim does not pay the ransom. This is referred to as a double extortion scheme (Tuttle, 2021; Payne and Mienie, 2021; Kerns et al., 2022). Evidence suggests the double extortion scheme leads to larger ransom requests and/or a higher willingness to pay, and therefore to more profits for the criminals compared to encryption-only-attacks (Li and Liao, 2021; Meurs et al., 2022). Hence, it is important to investigate *how double extortion schemes may evolve, and how law enforcement can disrupt the business model of these criminals*.

During a ransomware attack, victims struggle with the critical issue of determining whether their data has been exfiltrated (Meurs and Holterman, 2022; Nyakomitta and Abeka, 2020). While some may possess logs that facilitate the identification of accessed files, others are not as fortunate. Although irregularities in data flow on the affected network during an attack can be identified, it does not always confirm the theft of confidential information. This ambiguity creates a window of opportunity for criminals. Those engaged in encryption-only attacks can exploit this uncertainty, asserting that data exfiltration occurred to demand a higher ransom. Criminals may proactively offer, or victims may request, 'evidence' of data exfiltration. This exchange, termed as a 'signal' in our study, forms the core of the strategic dynamics between

* Corresponding author.
E-mail addresses: t.w.a.meurs@utwente.nl (T. Meurs), edward.cartwright@dmu.ac.uk (E. Cartwright), a.cartwright@brookes.ac.uk (A. Cartwright), m.junger@utwente.nl (M. Junger), s.abhishta@utwente.nl (A. Abhishta).

https://doi.org/10.1016/j.cose.2023.103670
Received 8 July 2023; Received in revised form 13 December 2023; Accepted 17 December 2023

Available online 21 December 2023

the attacker and the victim. It is a 'game', where the criminal's choice to provide evidence of exfiltration intersects with the victim's decision on whether or not to pay the ransom demand.

In 2020, Coveware reported that 70% of ransomware attacks were combined with data exfiltration (Tuttle, 2021). (Meurs et al., 2022) found that between 2019 and 2022, from 124 ransomware attacks, forensic analysis suggested there were traces of data exfiltration in 43% of the attacks. This resulted in a slightly larger portion of payments: 26% of the victims paid with likely data exfiltration, whereas 24% paid without data exfiltration. The ambiguity around data exfiltration complicates the victims' response, with criminals sometimes falsely claiming data theft to inflate ransoms (Meurs et al., 2022). The Maze group initiated the double extortion scheme in November 2019, exposing non-paying victims on a leak site and claiming data deletion for those who paid, though without conclusive proof (Greengard, 2021; Kerns et al., 2022; Ecrime, 2023; Cymru, 2022).

Criminals can use various strategies to signal that data was exfiltrated (Meurs and Holterman, 2022; Hack and Wu, 2021). One approach is to publish a small fraction of the exfiltrated data on a leak site. However, it is worth noting that this strategy carries a potential drawback, as the extent of the reputational harm incurred might be independent of the magnitude of the published data. Moreover, publishing some data still leaves open the question of how much additional data was exfiltrated. Another approach is to send a picture of the file tree to the victim. A potential drawback of this approach is that it gives the possibility to victims to determine the importance of the stolen files. It is also relatively easy to obtain without actually exfiltrating the files and so is not a particularly credible signal. Criminals may, therefore, decide not to signal even if data was exfiltrated. It could be they want to sell the data on darknet forums (Li and Liao, 2021) or to conceal attacks where data exfiltration was unsuccessful. In such cases the past reputation of the ransomware group may inform on the likelihood of data exfiltration.

In this study, we employ a game-theoretic framework to evaluate the dynamics between criminals and victims in the context of ransomware attacks, focusing on the signaling of data exfiltration (Kreps and Sobel, 1994). We are particularly interested in the criminals' decision to signal data theft and how victims respond to such signals. Signaling games, a well-explored concept in game theory, offer valuable insights into these complex interactions characterized by information asymmetry. Our analysis unveils five distinct equilibria, shaped by the criminals' varied approaches to signaling data exfiltration—from consistent signaling, no signaling at all, to conditional signaling based on actual data theft. These equilibria are not arbitrary but are influenced by the strategic parameters inherent in each ransomware attack scenario. We further calibrate the game with parameters that mirror real-world conditions. This calibration facilitates the identification of the most realistic equilibrium, enabling us to anticipate the likely ransom amounts and the corresponding payoffs for both parties involved. Based on our findings, we propose tangible strategies to dismantle the ransomware business model. These strategies are aimed at reducing the ransom amounts and undermining the criminals' payoffs, marking a significant step towards mitigating the impacts of these cyber-attacks.

Our article makes three contributions; First, it is, to the best of our knowledge, the first study to analyse the important signaling component of double extortion ransomware schemes. We draw on data from negotiations between criminals and victims to motivate this issue as being an important area of study. Second, we provide a theoretical analysis of the strategic consideration criminals face when signaling data exfiltration and the consequences for the payoffs of criminals and victims. Third, by understanding the incentives of criminals we can identify the optimal strategy of victims and examine defensive measures for victims and policy makers to decrease the negative welfare consequences of double extortion ransomware.

The paper is organized as follows. In Section 2 we motivate our signaling game approach through empirical observations of double extortion ransomware attacks and negotiations. In Section 3 we briefly overview the previous literature on the economic and game-theoretic modeling of ransomware. In Section 4 we introduce the signaling game. In Section 5 we state the main results. In Section 6 we conclude and provide policy recommendations. Proofs of propositions are provided in Appendix A.

2. Motivation

The foundation of our game-theoretic model is based upon empirical observations that will be expanded upon in the following section, providing context and motivation for our theoretical model. In Section 5.2 we will calibrate the parameters in our model using this dataset. For a more detailed analysis we refer to our previous work (Meurs et al., 2022).

We draw on empirical data from two datasets compiled by the lead author: 1) 525 ransomware attacks reported to the Dutch Police and 2) 117 ransomware attacks reported to an incident response company (IR company). Some general insights from the Dutch Police data have previously been reported in (Meurs et al., 2022). In that paper, the authors study how the criminal's effort, victim characteristics and context influence the ransom requested, payment and financial loss. A key finding of the study is that data exfiltration has a highly significant, positive impact on the ransom requested, proportion of victims who pay, and the victim's financial loss. This demonstrates the critical role that exfiltration plays in ransomware.

The foundation of our game-theoretic model is established upon empirical observations that will be expanded upon in the ensuing section, providing necessary context. For an in-depth analysis of the dataset, we point readers to our prior work (Meurs et al., 2022).

In motivating the game theoretic approach used in this study we analysed the extended datasets introduced above. From the overall datasets we excluded attempted attacks, no encryption and attacks on individuals. This resulted in 1) 354 ransomware attacks reported to the Dutch Police and 2) 98 ransomware attacks reported to the IR company. In total we, therefore, analysed 452 ransomware attacks. For each attack a range of variables were coded based on the case logs provided by the Police and IR company. For this study we use the following variables: whether data is exfiltrated (yes/no/unknown), what the ransom requested was before and after negotiations (in euro) and whether the victim paid (yes/no/unknown). Furthermore, we looked at the negotiation text to understand the exchange of information of data exfiltration between the victim and criminals. We remark that the classification of data exfiltration (yes/no/unknown) is somewhat subjective for the very reasons that motivate this paper (namely, data exfiltration is hard to verify). Our classification benefits, however, from information that became available over time, and may not have been available at the time of the attack, e.g. whether data was subsequently published on a leak site.

2.1. *Criminal profits of double extortion ransomware*

Here we state our main findings from the data analysis in terms of data exfiltration in relationship with payment, and ransom requested.

1. **Data exfiltration:** Overall, we find that in 50.4% of cases it was unknown whether data was exfiltrated, and in 49.6% of cases we believe that data was exfiltrated. Data exfiltration is assumed in 43% of cases in the Dutch Police, based on 134 cases, and in 53% of cases of the IR company, based on 98 cases. Commonly, the basis for assuming that data was exfiltrated is because: (1) log files show specifically that files have been exfiltrated, or (2) data was published on the leak site of the criminals. We see, therefore, that data exfiltration is common but not universal. We also see that whether data exfiltration took place remains unknown in many cases, even with the benefit of hindsight (e.g. data appearing on leak sites).

Table 1
Claims of the criminal of data exfiltration (raw text and anonymized), additional signals send by the criminal and the victim's decision-making whether to pay or not.

Case	Criminal claim (anonymized raw text)	Additional signals	Victim decision-making
1.	“We gathered highly confidential/personal data. These data are currently stored on a private server. This server will be immediately destroyed after your payment. If you decide to not pay, we will release your data to public or re-seller. So you can expect your data to be publicly available in the near future. We only seek money and our goal is not to damage your reputation or prevent your business from running”	No	Although the victim did not believe data was exfiltrated, they did decide to pay because backups were inadequate to recover without payment.
2.	“For the ransom you get: Full decryption Fixing your network vulnerabilities and securing your network Removal of all your data from our servers.”	No	Victim was not confident data was exfiltrated and did not pay.
3.	“If we don't hear back from you within 24 hours. I can sell them on the darknet and send the information to regulatory agencies, in your area I will send out offers to competitors to buy your data. In this case you will have the following problems: 1. Your customers will become victims of fraudsters (who will buy your data on the darknet). 2. Regulatory authorities (responsible for enforcing data protection laws) will start investigating your company for leaking your customers' personal data (leading to huge fines and loss of reputation). 3. Your competitors could easily get hold of your information.”	No	Victim was not confident data was exfiltrated. However, due to the lack of adequate backups they did pay. During the negotiations no other claim of data exfiltration was made by the criminal
4.	“All your important files have been encrypted. Any attempts to restore your files with third-party software will be fatal for your files! Restore your data possible only buying private key from us. We have also downloaded a lot of private data from your network. If you do not contact us in a 5 days, we will post information about your breach on our public news webs.”	No	Victim was not confident data was exfiltrated and did not pay.
5.	“Your data is stolen and encrypted. If you don't pay the ransom, the data will be published on our TOR darknet sites. Keep in mind that once your data appears on our leak site, it could be bought by your competitors at any second, so don't hesitate for a long time. The sooner you pay the ransom, the sooner your company will be safe.”	List of .rar files of alleged exfiltrated data provided by criminal	Victim paid, because data of customers was stolen. They had backups, but decided to pay just for prevent the publication of exfiltrated data.
6.	“Price for you is X btc. You need to pay this amount and we will give you decrypt tool for all your machines, security report on how you were hacked, file tree on what we have downloaded a lot of data from your network that in case of not payment will be published on public news website and sold on the black-markets. We remove it after payment and wiping log is provided as well. To start a business we offer you to make payment in two stages. What amount you can pay today?”	A list of exfiltrated files was provided by criminals	Victim got the file tree of the exfiltrated data and decided that the data was not important. Furthermore, they did have backups. Therefore, they decided not to pay.

2. **Paid:** From the 452 ransomware attacks, 130 victims negotiated. In total, 119 victims paid the ransom (27.8%). Of these, 78.5% victims paid after negotiations and 21.5% paid without negotiations. If we focus on those subset of payments where we are relatively confident if data exfiltration took place, we find that data exfiltration leads more often to payment: 37.5% versus 28.9%. This difference is statistically significant, based on a chi-squared test ($\chi^2 = 5.42, df = 1, p = 0.02$). The reason why both payment percentages are above the total average of 27.8% is because, relatively, it is more often unknown whether data was exfiltrated for the victims who did not pay. So, these results show that data exfiltration leads to larger proportion of victims paying than no data exfiltration.
3. **Ransom requested:** The average ransom request before negotiation is 1,029,320 euro (sd = 3.0 million euro). After negotiation the average ransom request is 578,956 euro (sd = 1.9 million euro), a decrease of 44%. When data is exfiltrated, the ransom before negotiation is 2,960,281 euro (sd = 4.7 million euro) and after negotiation 1,771,216 euro (sd = 3.2 million euro), a decrease of 40%. Without data exfiltration the ransom before negotiation is 466,924 euro (sd = 2.0 million euro) and after negotiation 135,346 euro (sd = 0.2 million euro), a decrease of 70%. Tests that there is a difference in ransom requested with and without data exfiltration using a t-test is significant for both ransom requested before negotiations ($t = 63.17, df = 232, p < 0.001$) and after negotiations ($t = 66.05, df = 232, p < 0.001$).[1] It appears that data exfiltration is highly profitable for the criminals. Furthermore, it seems that data exfiltration leads to less discount after negotiations than when data is not exfiltrated.

In conclusion, double extortion ransomware seems to lead to a larger proportion of victims paying the ransom and a larger ransom requested, and, therefore, to more profits for criminals.

2.2. Exploration of victim's decision to pay

If victims are more likely to pay a ransom, and pay a larger ransom, because of data exfiltration, it is naturally in their interests to ascertain whether data exfiltration has indeed taken place. As we discussed in the introduction this is difficult to do in the immediate aftermath of an attack. Hence criminals may want to signal data exfiltration, and victims may seek for information about data exfiltration. In Table 1 we provide six illustrative examples of criminals attempting to signal that data is exfiltrated. As you can see, data exfiltration was claimed in the ransom note. In two cases supplementary evidence was provided during negotiations. We also summarise the victims' decision-making process regarding ransom payment. Note that we display the anonymized text used by criminals, which includes grammar and style mistakes.

In the first four cases the victims were not convinced by the criminal's claim that data has been exfiltrated. The signal in this case was,

[1] In line with (Meurs et al., 2022) we have taken the logarithm of the ransom to approximately normalize the data, which is required to validly perform a t-test. Not taking the logarithm also results in highly significant t statistics.

therefore, seen as non-credible. Furthermore, in none of the four cases was data published on the leak site after the victim did not pay. This may suggest the victims were probably correct to infer no data had been exfiltrated. The absence of data being published on a leak site does not, however, serve as conclusive evidence that no data has been compromised. In informal discussions, law enforcement officers have disclosed to the authors that criminals are occasionally selective in their choice of which victim's data they publish. By exclusively publishing data of large organizations, criminals can cultivate a reputation as a group focusing on prominent victims.

In the fifth and sixth cases the criminal showed a list of files which were exfiltrated. In the fifth case this led the victim to believe that data was exfiltrated and they made the decision to pay the ransom to prevent the publishing of the data on a leak site. The criminals, thus, benefited from sending a more credible signal. In the sixth case the victim decided, based on the list of exfiltrated files provided by the criminal, that data publication would be less costly than paying the ransom. As a consequence, the data of the victim was published on the leak page. In this example sending a signal appears to have backfired for the criminals, because it gave the victim the opportunity to estimate the reputational damage of data exfiltration.

Considering our dataset as a whole, we have examples of all possible combinations: (a) The criminals signaling data exfiltration when we believe there was data exfiltration, (b) not signaling data exfiltration when we believe there was data exfiltration, (c) signaling data exfiltration when we believe there was no data exfiltration, and (d) not signaling data exfiltration when we believe there was no data exfiltration. This makes it difficult for victims, law enforcement and policy makers to understand the optimal response when claims of data exfiltration are made. Given the large ransom amounts at stake it is of value to better understand the trade-offs that victims face.

It is reasonable to hypothesize that criminals engage in strategic considerations when deciding to signal data exfiltration or refrain from doing so, taking into account the potential impact on victims' willingness to pay. Indeed, the history of ransomware shows that criminals rapidly evolve their economic strategy to ones that make more money. Consequently, we develop a decision model to capture this behavior, using a game-theoretic framework of signaling. In the subsequent section, we provide a rationale for utilizing game-theoretic models in the context of ransomware and overview prior research on this subject.

3. Related works

The goal of this section is to give a brief overview of past research on the economic and game-theoretic approach to ransomware attacks. Traditionally, ransomware research takes a more technical approach (Brewer, 2016; Richardson and North, 2017). However, recently the application of economic theory to analyse decision-making of criminals and victims of ransomware attacks have increased (Cartwright et al., 2019; Li and Liao, 2021; Laszka et al., 2017; Galinkin, 2021). This might be the result of ransomware criminals running there attacks as a business, where many decisions are made using economic reasoning (Huang et al., 2018).

Most ransomware criminals are financially motivated and conduct multiple attacks (Meurs et al., 2022; Connolly et al., 2021). Therefore it is important for them to optimize profits over multiple ransomware attacks. One important aspect is the use of different price discrimination strategies (Hack and Wu, 2021). For example, the criminals change the ransom requested on victim characteristics, like yearly revenue (Meurs et al., 2022). Another aspect is the use of data exfiltration: as concluded in Section 2, this increases the willingness to pay of victims, which leads to more profits for criminals. (Connolly et al., 2021) identify four distinct fears of victims which might explain the increased willingness to pay: (1) incrimination (e.g. exposure to data protection authorities), (2) reputational damage/lost revenue (e.g. exposure of sensitive data which could cause loss of customers), (3) exposure of intellectual property, and (4) humiliation (e.g. exposing embarrassing information about customers or a particular employee in an executive role). These fears increase the willingness to pay and give an incentive for criminals to perform data exfiltration, or pretend that data is exfiltrated.

In addition to the previously mentioned empirical studies, game-theoretic models have been employed to explore the dynamics between criminals and victims within the context of ransomware attacks (Cartwright et al., 2019; Laszka et al., 2017; Galinkin, 2021) and double extortion ransomware schemes (Li and Liao, 2022, 2021, 2020). Game theory provides a valuable theoretical framework for examining the strategic decision-making process of different actors, making it highly applicable in the context of ransomware attacks (Cartwright et al., 2019). This suitability arises from the well-defined roles of the actors involved, namely the criminal and victim, and clear decision options available to the victim, such as paying or not paying the ransom. Furthermore, the payoffs are mostly monetary and therefore easily quantified. From the game-theoretic framework we could infer whether there is a stable equilibrium and possible interventions to change that equilibrium to increase social welfare.

Several studies have applied a game-theoretic framework to double extortion ransomware (Li and Liao, 2021; Laszka et al., 2018; Li and Liao, 2020). (Li and Liao, 2021) demonstrate that when criminals employ a strategy involving both data encryption and data exfiltration, it consistently results in higher profits as opposed to solely relying on data encryption. Furthermore, the act of selling the exfiltrated data has been found to further increase the profitability for criminals, surpassing the potential reputation gains achieved by simply deleting the data upon receiving payment from victims.

One possible critique of using game-theoretic models in the ransomware context is the assumption of rational decision-making by both criminals and victims. Both criminals and victims may make impulsive, irrational decisions (Cartwright et al., 2019). Rationality, in this context, however, does not imply a cold and unemotional decision-making process, but rather an understanding that criminal and victim need to take account of each other's strategic incentives, and have incentives to maximize their financial payoff. This aligns with the Rational Choice Model proposed by (Cornish and Clarke, 1987). The Rational Choice Model (RCM) of crime states that criminals, or offenders, are rational decision-makers. Crime is purpose behaviour designed to meet the offender's commonplace needs for such things as money, status, sex and excitement. Offenders are reasoning actors who weigh means and ends, costs and benefits, and make a rough rational choice for the course of action that seems to yield the most benefit (Cornish and Clarke, 1987, 2014).

Research supports the Rational Choice Model of crime, for offline crime (Wortley and Townsley, 2016; Clarke, 2016) and online crime (Allodi et al., 2017; Xu and Hu, 2018). Most relevant, experiments show that policy measures that influence the costs and benefits of crime, by increasing the effort and the risks, and decreasing the potential benefits, generally prevent crime offline (Clarke, 2016) and online (Beebe and Rao, 2005). Taken together, we could conclude that the assumption of rationality, which is crucial for the application of a game-theoretic framework, can yield valuable insight in the context of double extortion ransomware schemes.

So far, we considered studies which focus on the profitability of data exfiltration, applying game-theoretical models to ransomware and data exfiltration. These papers abstract away from a key aspect of the strategic environment: victims are often unsure whether data is exfiltrated. Information asymmetry between victim and criminal can be modeled with signaling games (Osborne et al., 2004). Signaling games are a widely used framework in economics and evolutionary biology to model a range of applied settings. They have been used, for instance, to model job seekers signaling their productivity to potential employers (Spence, 1974). In this setting the signal could be years of schooling or high grades (even if that does not directly add to productivity). Signaling games have also been used to understand non-anonymous donations to

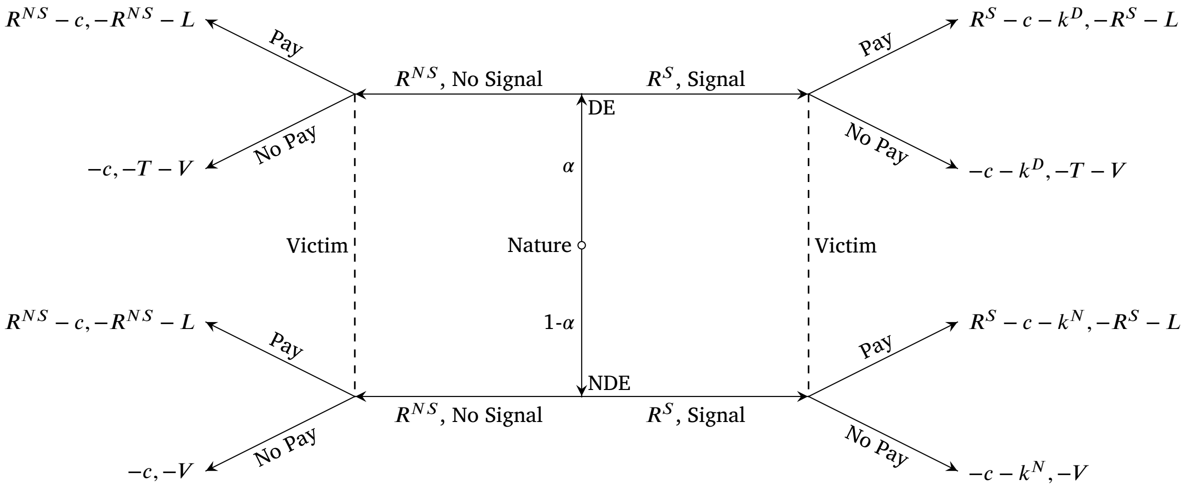


Fig. 1. A schematic representation of set-up of the signaling game of data exfiltration.

charity (Glazer and Konrad, 1996). In this case the donation can be a signal the donor is a pro-social, generous individual. In all these settings there is one party that has more information than the other (whether they are productive, pro-social etc.) and have incentives to signal a 'desirable' property. In our setting the criminal may want to signal data exfiltration.

The analysis of signaling games has produced some profound results. For instance, it has been shown that 'costly signaling' can result in which the actor with the most desirable attributes must incur large costs to signal their desirability. There is, for instance, evidence that university education is a costly signal of ability (Bedard, 2001). In our setting, this suggests that data exfiltration need not be unambiguously beneficial for the criminals. Another seminal finding is due to (Akerlof, 1970). He set up a framework to analyse the information asymmetry between the buyer and seller of used cars. The seller knows the quality of the car, but the buyer does not. (Akerlof, 1970) shows that this information asymmetry can lead to a breakdown in trade. In short, sellers of good cars do not want to sell them cheap, but buyers are reluctant to pay a high price for a car that may be no good. We observe, therefore, an adverse selection market failure. The framework of (Akerlof, 1970) has also been used to study other use cases (Laszka et al., 2018). In our setting, it again suggests that data exfiltration need not be unambiguously beneficial for the criminals. In the following section we will formally apply the theory of signaling games to the case of double extortion ransomware attacks.

Table 2
Variables used in the data exfiltration signaling game.

	Variable	Description
Criminal	R^{S}	Ransom when signaling
	R^{NS}	Ransom when not signaling
	c	Cost of attack
	k^{D}	Cost of signal with data exfiltration
	k^{N}	Cost of signal no data exfiltration
	τ	The state or type of the criminal: data exfiltrated or not
Victim	T	Reputation cost
	V	Recovery cost without decryption key
	L	Legal fees of paying ransom
	α	Probability of data exfiltration
	μ	Probability the victim believes data is exfiltrated
	ϵ	Smallest gain that would induce victim to pay

4. Model

4.1. Signaling game

In this section we introduce a signaling game of double extortion ransomware. The game involves two players, a criminal and a victim. It has three stages, which can be explained as follows. The variables are summarized in Table 2 and the signaling game is depicted in Fig. 1.

Stage 1: We assume that the criminal attempts to exfiltrate data. However, this attempt may not succeed. We denote with α the probability the criminal exfiltrates data (DE) and denote with probability $1-\alpha$ the probability the criminal does not succeed in exfiltrating data (NDE). In game-theoretic terminology, this process is formulated as 'nature' determining with a probability α the state is DE and $1-\alpha$ the state is NDE (Akerlof, 1970; Osborne et al., 2004; Spence, 1974). The criminal learns in Stage 1 whether they are in state DE or NDE. The victim remains uninformed, although the probability of data exfiltration α is common knowledge.

Stage 2: The criminal chooses a ransom demand and whether to send a signal to the victim that data was exfiltrated (S) or not (NS). The signal could consist of sending a file tree or pictures of the file tree structure. Another possibility is sending a few exfiltrated files. Let R^{S} denote the ransom demand of the criminal if they send a signal and R^{NS} the demand if no signal is sent. Thus, the criminal either sends signal S and ransom demand R^{S} or chooses NS and ransom demand R^{NS}.[2]

The cost of sending a signal is k^{D} when data is exfiltrated and k^{N} when no data is exfiltrated. One interpretation of the cost of a signal is opportunity costs. For instance, the effort and time could have been used for another attack. Crucially, we assume that sending a signal when data is exfiltrated is less costly than when data is not exfiltrated, so $k^{D} < k^{N}$. This assumption arises from the notion that it might be harder to send a signal in state NDE than DE. Indeed, it could be that k^{N} is very large meaning that it is essentially impossible to send a signal if data is not exfiltrated.

Stage 3: Having seen whether the criminal sends a signal (S) or no signal (NS) and seen the ransom demand R^{S} or R^{NS}, the victim decides to pay or not. We assume that this is a binary yes/no decision

[2] The criminal could choose any ransom above 0 for any combination of both own type and signal. So, suppose, more generally, we denote by $R^{S}_{DE}, R^{S}_{NDE}, R^{NS}_{DE}$ and R^{NS}_{NDE} the ransom of a type DE or NDE if they signal or do not signal. There cannot be an equilibrium in which a criminal of type DE and NDE signal and $R^{S}_{NDE} \neq R^{S}_{DE}$; this would reveal the criminal if type NDE and, thus, make their signal ineffective. Similarly, there cannot be an equilibrium in which a criminal of type DE and NDE would not signal and $R^{NS}_{NDE} \neq R^{NS}_{DE}$; this would again reveal the criminal if type NDE and lower the ransom the victim would rationally pay.

with no possibility for negotiation.[3] Final payoffs are now determined as depicted in Fig. 1, which shows criminal payoff, victim payoff for each potential outcome.

In explaining the respective payoffs of criminal and victim we remark that game theoretic equilibria depend on the relative payoff differences across actions, rather than the absolute payoff. For the victim we are, thus, interested in the relative payoff difference from paying the ransom versus not paying the ransom. We assume that if the victim pays the ransom then they lose the ransom amount, R^S or R^{NS}, as well as 'legal fees', $L \geq 0$, which can include legal and other associated costs (including psychological and moral) of paying the ransom. We assume that if the ransom is paid the criminal returns access to, at least some, files and is reduced the amount of sensitive data published. If, therefore, the victim does not pay the ransom they lose $V \geq 0$ from higher recovery costs as well as $T \geq 0$ (if data was exfiltrated) from increased reputational costs resulting from publication of sensitive data. This motivates the payoff function in Fig. 1.

In analysing the incentives of the criminal we need to consider the relative payoff differences from signalling or not signalling. We assume that the attack costs the criminal c. If the victim pays then the criminal receives ransom R^S or R^{NS}. If the criminal signals then they pay the cost k^D or k^N, as stated previously.

The victim will have legal fees, recovery costs (like buying new hardware and software) and reputation costs under any scenario. We reiterate, however, that since we consider these costs to be constant across all outcomes, we do not include them in our analysis. Furthermore, we would like to stress that our model implicitly takes into account factors such as the importance of backups. For instance, if a victim has good backups then the recovery costs V would be low, and therefore the victim, as we will show, would not pay a large ransom. Similarly, if sensitive data is not exfiltrated then T would be low and the victim would again not pay a large ransom.

A (pure) strategy for the criminal involves the following conditional choices: (a) decide whether to signal or not if data is exfiltrated, (b) decide whether to signal or not if data is not exfiltrated, and (c) determine ransom demands R^S or R^{NS} as appropriate. A (pure) strategy for the victim compromises conditional choices: (a) decide to pay or not to pay if the criminal signals, and (b) decide to pay or not to pay if the criminal does not signal.

4.2. Bayesian equilibria of the signaling game

In the following we identify (pure strategy) Bayesian equilibria of the signalling game. A Bayesian equilibrium takes into account that the victim starts with prior belief α that data was exfiltrated but can potentially update their beliefs once they observe the strategy of the criminal. We denote by μ the updated belief of the victim the criminal is type DE. A Bayesian equilibrium has the following basic properties: (a) The criminal maximizes their expected payoff given the strategy of the victim, (b) the victim updates their beliefs about state DE and NDE using Bayes rule, and (c) the victim maximizes their expected payoff given the strategy of the criminal and their own beliefs (Fudenberg and Tirole, 1991).

Where relevant it may be necessary to tie down beliefs for 'surprise events' or outcomes that are 'off the equilibrium path'. For instance, if the candidate equilibrium says that the criminal will signal, we need to specify the victim's beliefs if the criminal is observed to not signal. In this case we invoke the D1 Criterion which says that any deviation from the equilibrium path is assumed to be done by the type with the most incentive to deviate (Banks and Sobel, 1987). In this case (see the formal analysis for more details) it means the choice to not signal is seen as evidence that there was no data exfiltration.

As is standard in the analysis of signalling games, we distinguish between separating and pooling equilibria. A separating equilibrium has the property that the victim can distinguish the type of the criminal (DE or NDE) from their actions. A pooling equilibrium has the property that the criminal will act the same whether type DE or NDE and so the victim can not distinguish type. We identified two separating equilibria (we will call A1 and A2) and three pooled equilibria (we will call B1, C1 and C2). We first characterize the five equilibria. We then provide conditions on the parameters of the game under which the different equilibria exist. We would argue that all five types of equilibria can potentially be seen in the field.

A1. Separating equilibrium. Victim pays whether signal or not. The victim believes $\mu = 0$ when she receives no signal and $\mu = 1$ when they receive a signal. The victim is, thus, willing to pay the ransom when she gets the signal that data is exfiltrated if and only if $R^S < T + V - L$. Set $R^S = T + V - L - \epsilon$, where ϵ is arbitrarily close to 0. Likewise, the victim is willing to pay the ransom when there is no signal if and only if $R^{NS} < V - L$. Set $R^{NS} = V - L > 0$. Combined, the criminal has payoff $U = T + V - L - k_D - c - \epsilon$ in state DE and $U = V - L - c - \epsilon$ in state NDE. The criminal has no incentive to deviate from the equilibrium strategy in state DE if $k^D < T$. In interpretation, the extra revenue the criminal can demand from signalling data is exfiltrated compensates for the cost of sending the signal. Similarly, the criminal has no incentive to deviate from the equilibrium strategy in state NDE when $k^N > T$. In interpretation, the cost of sending a signal (when data is not exfiltrated) is higher than the extra revenue from the ransom.

A2. Separating equilibrium. Victim only pays when receiving signal. As with equilibrium A1, the victim believes $\mu = 0$ when they receive no signal and $\mu = 1$ when she receives a signal. Following, the same logic as equilibrium A1 the victim is willing to pay ransom $R^S = T + V - L - \epsilon$ if there is a signal. The maximum ransom they are willing to pay if there is no signal is $R^{NS} < V - L$. If, therefore, $V < L$ the victim is not willing to pay a (positive) ransom. The payoff for the criminal in state DE is $U = T + V - L - c - k^N - \epsilon$ and their payoff in state NDE is $U = -c$. The criminal in state NDE has no incentive to signal if $T + V - L < K^N$. In interpretation the cost of signaling in state NDE is higher than the maximum ransom the victim is willing to pay.

B1. Pooled equilibrium: The criminal signal and the victim pays. The criminal sends a signal in both states DE and NDE. The victim should maintain the belief $\mu = \alpha$ when they receive a signal that data is exfiltrated. If they do not receive a signal than beliefs are set $\mu = 0$ (invoking the D1 Criterion). The maximum ransom a victim is willing to pay if a signal is sent is $R^S = V + \alpha T - L - \epsilon$. The maximum ransom they are willing to pay if no ransom is sent is $R^{NS} = V - L - \epsilon$. Thus, the criminal has no incentive to deviate from the equilibrium path in state NDE if $\alpha T > k^D$. In interpretation, the cost to the NDE type of signaling is sufficiently low that they signal even though no data was exfiltrated. This lowers the ransom a type DE can demand because their signal is less credible.

C1. Pooled equilibrium: The criminal does not send a signal and the victim pays. The criminal sends no signal in both state DE and NDE. The victim should maintain the belief $\mu = \alpha$ when they receive no signal that data is exfiltrated. If they do receive a signal than beliefs are set $\mu = 1$ (invoking the D1 Criterion). The victim is willing to pay ransom $R^{NS} = V + \alpha T - L - \epsilon$ when she does not receive a signal and $R^S = V + T - L - \epsilon$ when she does receive a signal. The criminal has no incentive to deviate when type DE if $(1 - \alpha)T < k^D$. In interpretation, the extra ransom is insufficient to cover the cost of sending a signal (even when data is exfiltrated). This equilibrium also requires $V + \alpha T > L$ so that the victim is willing to pay a positive ransom.

C2. Pooled equilibrium: No signal and victim does not pay. We follow the same logic as equilibrium C1 but now consider the case

[3] Alternatively, R^S or R^{NS} could be seen as the final ransom demands that will result from negotiation.

Table 3
Stable equilibria and conditions in signaling game.

Case	Type equilibrium	Condition
A1	Separating - victims pays	$L < V$ & $k^D < T < k^N$
A2	Separating - Only pay when signal	$V < L < V + T$ & $k^D < T + V - -L < k^N$
B1	Pooling - Signal and pay	$V + \alpha T > L$ & $\alpha T > k^N$
C1	Pooling - No signal and pay	$V + \alpha T > L$ & $(1 - -\alpha)T < k^D$
C2	Pooling - No signal and no pay	$V + \alpha T > L$ & $V + T - -L < k^D$

Table 4
The ransom and payoffs of criminals and victims in the different equilibria depending on the type of the criminal.

Case	DE			NDE		
	R^S	Criminal	Victim	R^{NS}	Criminal	Victim
A1	$T + V - L - \epsilon$	$R^S - c - k^D$	$-R^S - L$	$V - L - \epsilon$	$R^{NS} - c$	$R^{NS} - L$
A2	$T + V - L - \epsilon$	$R^S - c - k^D$	$-R^S - L$	0	$-c$	$-V$
B1	$V + \alpha T - L - \epsilon$	$R^S - c - k^D$	$-R^S - L$	$V + \alpha T - L - \epsilon$	$R^{NS} - c - k^N$	$-R^{NS} - L$
C1	$V + \alpha T - L - \epsilon$	$R^S - c$	$-R^S - L$	$V + \alpha T - L - \epsilon$	$R^{NS} - c$	$-R^{NS} - L$
C2	0	$-c$	$-V - T$	0	$-c$	$-V$

where $V + \alpha T < L$. In this case the victim is not willing to pay a positive ransom if a signal is not sent. Moreover, the criminal has no incentive to deviate when type DE if $V + T - L < K^D$. The interpretation of this equilibrium is that it is too costly to pay for the victim and too costly for the criminal to send a credible signal. Clearly there would be no incentive for the criminal to attack in this scenario because they incur the cost c.

The Bayesian equilibria that exist in the game will depend on the specific parameters of the game, V, L, T, α, K^D and K^N. In the following three Propositions we characterise the set of conditions under which there exists separating equilibria A1 and A2 (**Proposition 1**), pooling equilibria B1 (**Proposition 2**), and pooling equilibria C1 and C2 (**Proposition 3**). Proof of propositions can be found in the Appendix.

Proposition 1. *If $V > L$ and $k^D < T < k^N$ there is a Bayesian equilibrium satisfying the D1 Criterion of the type A1. If $V < L$ and $k^D < T + V - L < k^N$ there is a Bayesian equilibrium satisfying the D1 Criterion of the type A2.*

Our first proposition shows that there exists a separating equilibrium if the cost of signalling is sufficiently low when the criminal is type DE and high when they are type NDE. Thus, the criminal only signals if data has been exfiltrated. The criteria for sufficiently low and high depends on the reputational costs T, recovery costs V and legal fees L. The victim pays if data is exfiltrated and pays if data is exfiltrated if and only if $V > L$.

Proposition 2. *If (a) $V > L$ and $\alpha T > k^N$, or (b) $V + \alpha T > L > V$ and $\alpha T + V - L > k^N$ there is a signaling equilibrium satisfying the D1 Criterion of the type B1.*

Our second proposition shows conditions under which there exists a pooling equilibrium where the criminal signals data is exfiltrated, irrespective of whether data is exfiltrated or not. This equilibrium exists if it is sufficiently low cost for the criminal to signal data exfiltration. The notion of sufficiently low depends on the ex-ante probability of data exfiltration α and the reputation cost T. The higher is αT then the more likely to obtain a pooling equilibrium with signalling. In interpretation, the victim is willing to pay a larger ransom if data exfiltration is signaled and so it is in the interests of the criminal to signal data exfiltration when type NDE (if k^D is sufficiently low).

Proposition 3. *If $V + \alpha T > L$ and $(1 - \alpha)T < k^D$ there exists a signaling equilibrium satisfying the D1 Criterion of the type C1. If $V + \alpha T > L > V$ and $T + V - L < k^D$ there exists a signaling equilibrium satisfying the D1 Criterion of the type C2*

Our final proposition shows conditions under which there exists a pooling equilibrium where the criminal does not signal data is exfiltrated, even if it is. This type of equilibrium exists if the cost of signaling is sufficiently high for type DE. Again, the reputation costs T are an important determinant of the meaning of sufficiently high. If the reputation costs are low then we are more likely to obtain a pooling equilibrium with no signalling. In interpretation, the victim is not willing to pay a larger ransom if data is exfiltrated and so there is less incentive for the criminal to signal exfiltration (if k^D is sufficiently high).

The five type of equilibria we have identified and conditions under which they exist are summarized in Table 3.

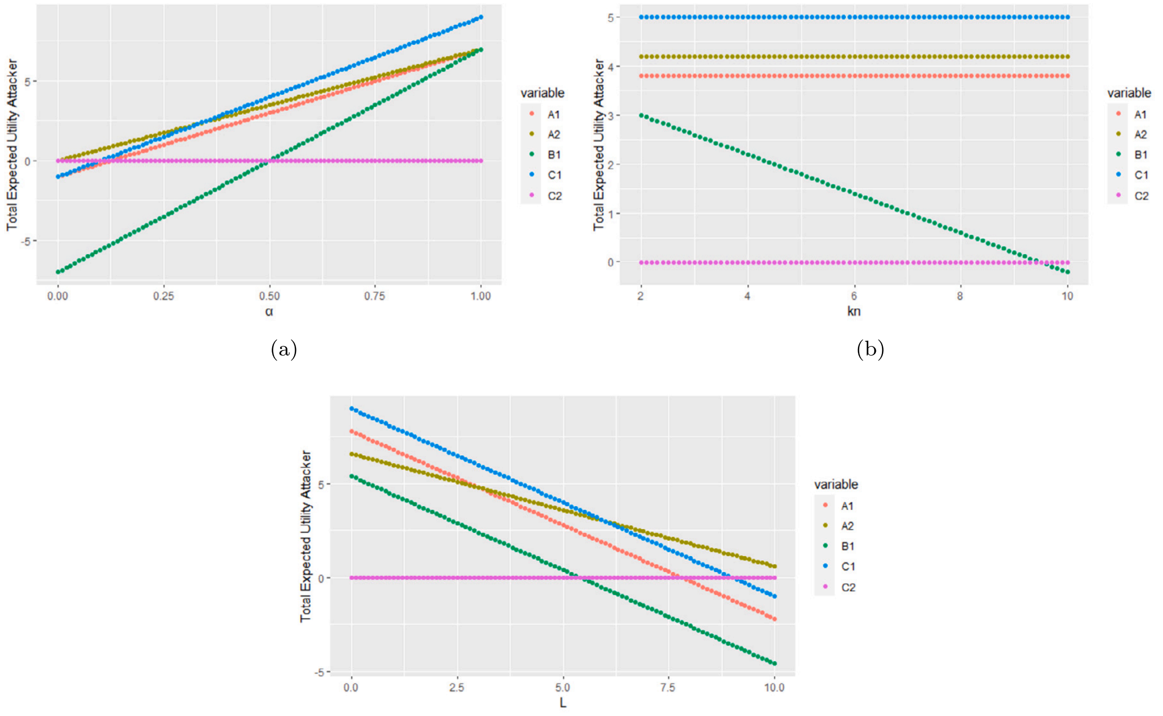
5. Theoretical insights from the game

5.1. Expected payoffs

In the previous section we derived five types of Bayesian equilibria of the signaling game. In this section we perform simulations to better understand the interaction between different parameter values and the resultant payoffs of the criminal and victim. A summary of the equilibrium ransom amount and corresponding payoff of criminal and victim conditionally on the type of the criminal is depicted in Table 4.

If the criminal is type DE then they would prefer a separating equilibrium (A1 or A2) to a pooling equilibrium because they can charge a higher ransom and obtain a higher payoff. By contrast, if the criminal is type NDE they would prefer a pooling equilibrium (B1 or C1) because they can charge a higher ransom and obtain a higher payoff. As is standard in signalling games we, thus, obtain a complex interaction in which one type, DE in our game, has incentives to signal their type, while the other type, NDE, has an incentive to hide their type. The equilibrium outcome obtained will depend on the parameters of the game.

Having looked at expected payoffs for each type of criminal we can consider the ex-ante expected payoffs for both criminal and victim. The expected utility hypothesis of Von Neumann-Morgenstern states that the choice involving uncertainty of a decision-maker can be represented by the expected value of the cardinal utility functions (Ng, 1984; Von Neumann and Morgenstern, 1944). In other words, the total expected utility can be represented as the expected value of the separate utility functions multiplied by the probability of every state. In the current context this results in:



(a)

(b)

(c)

Fig. 2. Total expected utility for the criminal when changing (a) α, (b) k^N, (c) L. (For interpretation of the colours in the figure(s), the reader is referred to the web version of this article.)

Table 5
Ex-ante expected payoff of criminal and victim before criminal type is determined.

Case	Criminal	Victim
A1	$\alpha(T-k^D)+V-L-c-\epsilon$	$\alpha T-V+\epsilon$
A2	$\alpha(T+V-L-k^D-\epsilon)-c$	$\alpha T-V+\alpha\epsilon$
B1	$\alpha(T-k^D)-(1-\alpha)k^N+V-L-c-\epsilon$	$\alpha T-V+\epsilon$
C1	$\alpha T+V-L-c-\epsilon$	$\alpha T-V+\epsilon$
C2	$-c$	$\alpha T-V$

$$U(\Theta)=\alpha U_{DE}(\Theta)+(1-\alpha)U_{NDE}(\Theta) \tag{1}$$

Where U(·) represents the (cardinal) utility function or payoffs, α the probability of data exfiltration and Θ the parameters V, L, T, K^D, k^N.

Expected payoffs in the signaling game for each possible type of equilibrium are shown in Table 5. You can see that the victim has essentially the same payoff irrespective of the type of equilibrium. This is because the criminal is able to extract the maximum surplus from the victim. To explain, consider equilibrium C2 in which the victim does not pay. In this case they suffer the recovery loss V and, with probability α reputational damage T. This, ex-ante, is the most the victim can lose from the attack. In the other types of equilibria the victim pays the ransom (with positive probability) and gains up to ϵ from doing so. We, thus, see that ϵ can be interpreted as the smallest financial gain that would induce the victim to pay a ransom.

As we would intuitively expect, the victim's payoff loss from the attack is lower if the victim has active ready back-ups (which lowers V), less sensitive data (which lowers T), and measures in place to stop exfiltration (which lowers α). Crucially, these factors are beneficial to the victim irrespective of the type of equilibrium, and, thus, whether the victim pays the ransom or not, because they lower the ransom the criminal can demand. Preventive measures are, therefore, beneficial even if the victim pays the ransom.

While the victim's expected payoff does not depend on the type of equilibria, we can see in Table 5 that the criminal's payoff is highly dependent on the type of equilibria. To illustrate, in panels (a-c) of Fig. 2 we plot the expected payoff of the criminal under each equilibrium type (assuming for now the equilibria exist) for fixed parameter values. We vary α, k^N and L in panels (a-c) respectively. We see that, for most parameter values, equilibria of type A2 or C1 maximize the criminal's payoff. By contrast, equilibria B1 never maximize the criminal's payoff. This is noteworthy because equilibrium B1, in which the criminal signals data exfiltration, may appear a natural outcome. This type of equilibrium is not optimal for the criminal because they incur the costs of signaling exfiltration but cannot extract a higher ransom from signaling. Better for them to have equilibrium C1, in which they do not incur costs of signaling, or equilibrium A2, in which signaling enables a higher ransom.

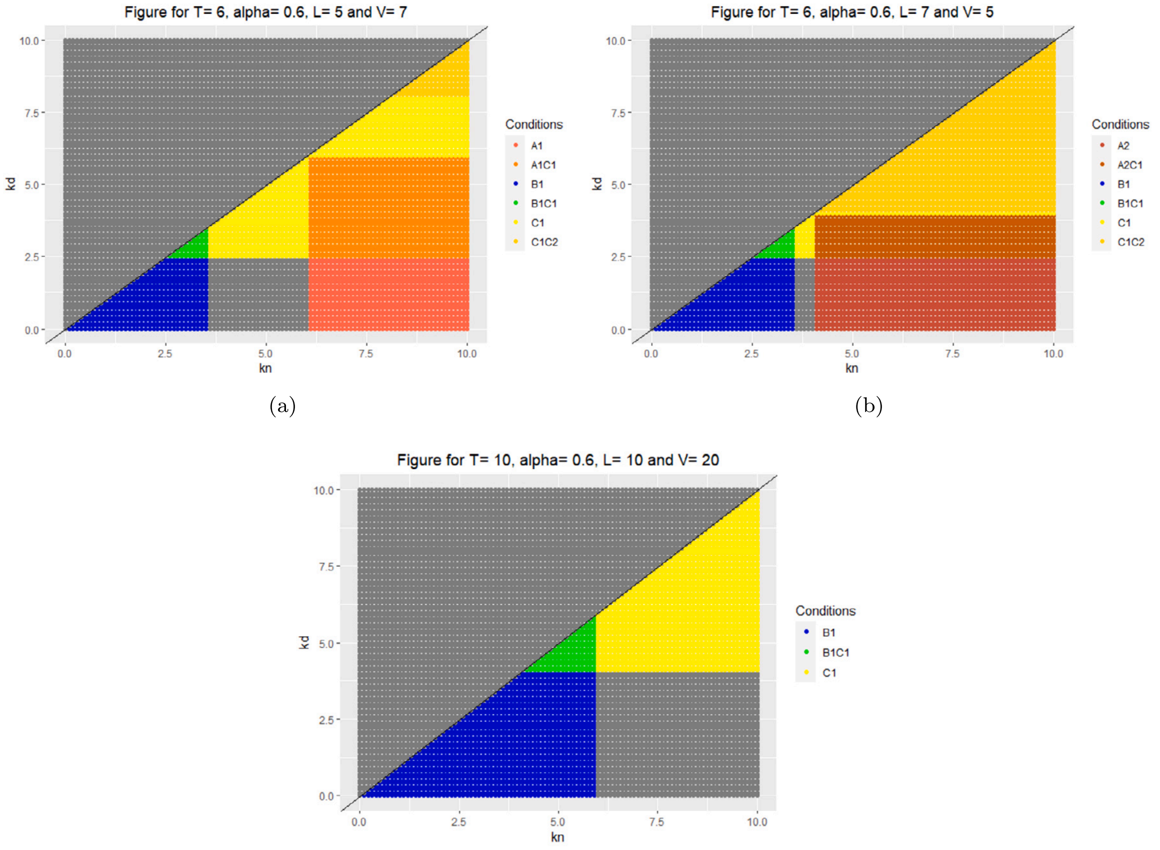
Fig. 2(a) shows that increasing α leads to a larger expected payoff for the criminal (except for case C2). Thus, the criminal's payoff is higher if they have a higher ex-ante probability of data exfiltration. This suggests criminals have an incentive to improve their ability to exfiltrated data.

Fig. 2(c) shows that a higher L will lead to a lower expected payoff for the criminal. This is because the higher L is reflected in a lower ransom paid. In interpretation, the legal fees are transferred from the victim to a third party (e.g. lawyers or insurers) rather than the criminals. This may be viewed as desirable from a societal perspective, although it does not materially benefit the victim.

Fig. 2(b) shows that increasing k^N only impacts the criminal's profit in equilibrium B1. This is interesting, because increasing k^N, the cost of signaling data exfiltration when no data is exfiltrated, may seem a natural lever that victims could use to disrupt the criminal's business model. Our analysis suggests that increasing k^N may have limited impact. To explore this further we need to investigate which type of equilibria are most likely to exist in the field.

5.2. Overlapping equilibria

As we have already demonstrated (see Propositions 1-3 and Table 3) each of the five equilibria we have identified will only exist under par-



(a)

(b)

(c)

Fig. 3. Overlap of different equilibria for different parameters.

ticular parameter values. Moreover, for a given set of parameters we may obtain multiple equilibria, a unique equilibrium, or no equilibria. To illustrate, we provide three examples depicted in Fig. 3. We indicate the existence of equilibria for combinations of k^N and k^D.

The first example, see Fig. 2(a), has $T = 6$, $\alpha = 0.6$, $L = 5$ and $V = 7$. Here we see parameters for which the separating equilibrium A1 and the pooling equilibrium C1 exist. This occurs when k^N is large and k^D is ‘intermediate’. In both equilibria the criminal does not signal if type NDE (because k^N is large). The equilibria differ in whether the criminal signals if type DE. Both equilibria are possible because k^D is an ‘intermediate’ range. If k^D is higher then only the pooling C1 equilibrium exists, while if it is lower only the separating A1 equilibrium exists.

We also see parameters for which both the pooling equilibrium B1, with signalling, and the pooling equilibrium C1, with no signalling, both exist. This happens for lower values of k^N and k^D. In interpretation, the criminal of type NDE will want to copy the equilibrium behavior of the type DE and it is too costly for the type DE to differentiate themselves. There are also parameter values for which there is no equilibrium. This happens for a small k^D and ‘intermediate’ k^N. In this case the type NDE wants to copy the type DE, but the type DE will want to differentiate themselves. There is, therefore, no stable (pure strategy) pooling equilibrium.

In our second example we set $T = 6$, $\alpha = 0.6$, $L = 7$ and $V = 5$. Thus, the legal fees are now larger than the recovery cost. See Fig. 3(b). The high legal fees mean that the victim will not pay unless they have sufficiently high belief that data was exfiltrated. We, thus, see equilibrium A2. Otherwise, the types of equilibria we observe in our second example, as a function of k^N and k^D, are similar to those in our first example.

There are two interesting findings we will highlight from these examples. Consider our first example with $k^D = 2.4$ and $k^N = 2.5$. As we have discussed, there are two types of pooling equilibria for these parameters: B1 with signaling and C1 with no signaling. The criminal's expected payoff is 3.11 with equilibrium B1 and 5.55 with equilibrium C1. Clearly, therefore, the criminal would prefer equilibrium C1 over B1. This is because they avoid the cost of signaling data exfiltration. Criminals, however, have limited influence over which type of equilibria will emerge because it will depend on norms and historical precedence. It is possible, therefore, that a B1 equilibrium could emerge, in which criminals signal data exfiltration, even though this equilibrium is not the one they would prefer.

The second finding we would highlight is that an increase in k_N can, perhaps counter-intuitively, lead to an increased expected payoff for the criminal. To illustrate, consider the first example with $k^D = 1$ and $k^N = 2.5$. In our example this leads to equilibrium B1 with expected payoff for the criminal of 3.95. The low cost to signal data exfiltration (even if data is not exfiltrated) results in a pooling equilibrium where the criminal signals irrespective of type. Now suppose $k^N = 7.5$. In this case we obtain equilibrium A1 with an expected payoff for the criminal of 4.95. The increase in k^N increases the expected payoff of the criminal because it makes it easier for them to send a credible signal of data exfiltration. Hence, they are able to extract a higher ransom when of type DE. Also, the type NDE criminal no longer incurs the cost of signaling.

For different parameter ranges and increase in k^N can lower the expected payoff of the criminal. The general point, therefore, is that care is needed in evaluating interventions aimed at disrupting the criminal's business model. An increase in the costs of signaling data exfiltration can benefit the criminals by either making signals more credible and/or removing the incentives to send costly signals of data exfiltration.

5.3. Calibrating parameter values

For our third example we look to calibrate the parameters of the game, drawing on the data described in Section 2 with the objective to identify the most likely equilibria we would observe in the field.

1. **Probability of data exfiltration α:** Most criminals in our dataset try to exfiltrate data (Meurs et al., 2022). This seems in line with a Dutch whitepaper where 7 IR companies mentioned that in most ransomware attacks of these clients, data was exfiltrated (Meurs and Holterman, 2022). This points to a high value of α. Data from Coveware in 2022 suggests that around 80% of ransomware attacks involve data exfiltration (Culafi, 2022). This, however, will include cases where the data exfiltrated could be deemed non-sensitive and of little value. The appropriate value of α in our model will, thus, be lower than such upper bounds. We suggest that setting $\alpha = 0.6$ strikes a reasonable balance.
2. **Recovery cost V versus legal fees L:** A key determinant of the type of equilibrium we obtain in our model is the relationship between V and L (see Table 3). There are various costs to paying a ransom that would contribute to L. These include prohibition and checks that payments are consistent with sanction legislation.[4] There are also costs to ransom negotiation and sourcing cryptocurrency. Furthermore, there is evidence of negative psychological and moral consequences of paying (Cartwright et al., 2019; Corbet and Goodell, 2022). The simple reality, however, during the rapid rise of ransomware, is that a large proportion of victims pay the ransom. This trend predates the emergence of double extortion and so is strong evidence that $V > L$ for most organisations. In other words, the financial gain from recovering access to encrypted files exceeds the costs of paying a ransom. This may be the case even if a business has back-ups, given that return of the files may allow a more rapid return to normal operations.
   To give some perspective, The average financial loss reported by victims in our dataset is 555,820 euro (sd = 3 million euro). The average loss when a ransom is paid is 399,098 euro (sd = 0.8 million euro) while the average loss when a ransom is not paid is 674,672 euro (sd = 3.9 million euro); a difference of around 275,000 euros. Furthermore, the average ransom paid is 330,326 euro (sd = 0.8 million euro). Combining these two pieces of evidence, we might infer that $V - L$ is around 300,000 euro on average. In our calibration we, therefore, assume the recovery costs V are relatively large.
3. **Reputation cost T:** Another key determinant of the type of equilibrium in our model is the relationship between T and L and V. It is acknowledged that double extortion has resulted in increased incentives to pay ransoms (Payne and Mienie, 2021; Mott et al., 2023). Indeed, analysis of our data revealed the ransom requested with data exfiltration is roughly 3 million euro and after negotiation roughly 1,7 million euro. Without data exfiltration the ransom was roughly 460,000 euro before negotiation and 135,000 euro without data exfiltration. This points to significant concerns about reputational costs (Pattnaik et al., 2023). Payment of a ransom does not, however, guarantee that data will not be leaked; nor does it protect the business against reputation damage or regulatory fines from the data breach (Hodge, 2023). We suggest, therefore, that the reputational 'savings' from paying a ransom are of secondary importance compared to recovery costs. Reputation costs are likely to be similar to legal fees in order of magnitude. Specifically, we set $V > T$ and $T = L$.
4. **k^D versus k^N:** The negotiations of the attacks analysed in Section 2 showed that some criminals did not send proof of data exfiltration even though analysis of logs established that data was exfiltrated. Likewise, in some cases where it was show that data was most likely not exfiltrated, the criminals said that data was exfiltrated. In most cases where it was considered likely data was exfiltrated the criminal sent proof by means of a file tree. Taken together, we will interpret evidence of signals being sent, as evidence that the costs k^D and k^N are relatively low compared to L, V and T. However, we will assume k^N is relatively large compared to k^D, because it is harder to, for example, make a file tree if no data is exfiltrated.
5. **Costs of attack c:** It is hard to quantify the costs criminals incur during an attack. (Galinkin, 2021) estimate the cost of a ransomware attack to be around 4,200 dollars. However, the cost of an attack seems to be related to so many variables that it is hard to give a complete estimate. For example, when the criminal is affiliated with a ransomware strain which is part of RaaS, then most probably they have to pay a part of the profits to the ransomware developers. On the other hand, the RaaS group helps with setting up the infrastructure and tooling for data exfiltration. In our sample, RaaS is more often associated with data exfiltration. This might indicate that the costs of setting up a leak site and performing data exfiltration are too much effort for an individual criminal. We elaborate on this case in the following subsection. Here we assume the costs of the attack being relatively low compared to V, T and L, based on (Galinkin, 2021).

Based on our calibration exercise we performed a simulation with $\alpha = 0.6$, $T = L = 10$ and $V = 20$. This takes into account that $V > T, L$. See Fig. 3(c) for the set of equilibria. Since we expect signaling costs to be relatively low compared with the other parameters, we would expect the lower-left quadrant of the graph to be most likely in real-life. This suggests equilibrium B1. Therefore, we would expect an equilibrium where (in the 'average' attack) the criminals signal that data is exfiltrated, whether data is exfiltrated or not, and the victim pays. This means criminals incur the costs of signaling. It also, as we now discuss, raises interesting questions about whether the criminals have an incentive to exfiltrate data.

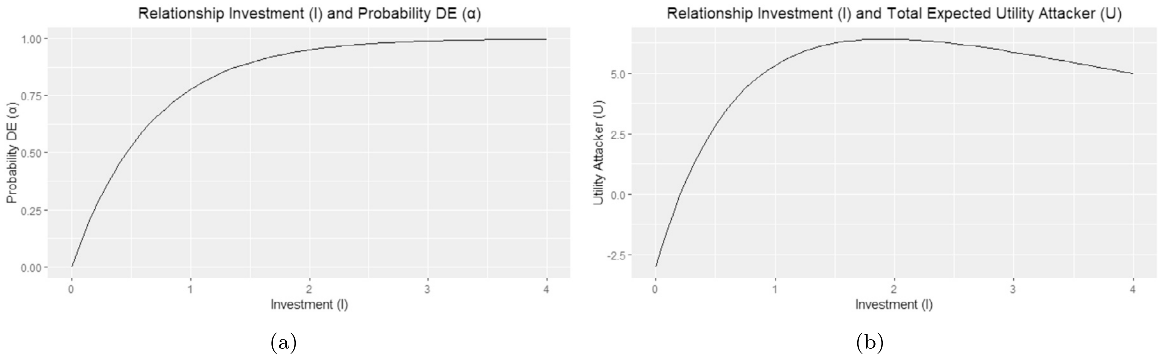
5.4. Increasing the probability of data exfiltration

We consider B1 to be the most likely equilibrium in the field. In this case the criminal can obtain ransom $V + \alpha T - L - \epsilon$. And the expected payoff of the criminal is $\alpha(T - k^D) - (1-\alpha)k^N + V - L - c + \epsilon$. Since (with equilibrium B1) $\alpha T > K^N$ and $k^D < k^N$ it holds that expected payoff is an increasing function of α. That is, increasing α will increase the total expected payoff of the criminal. We highlight, therefore, that while the criminals cannot extract a higher ransom from a particular attack if they exfiltrate data, they can gain across many attacks from a reputation for data exfiltration. We found cases where IR companies mentioned the reputation of the ransomware group as a possible indicator of data exfiltration: "Although no evidence of data exfiltration is found during the forensic analysis, this group is well known for exfiltrating data." Reputation will be positively correlated with the value of α.[5]

We extend our model to consider the case where the criminals can influence α by putting effort and/or investments into the attack, denoted as investment cost I. In this case the cost of an attack becomes a function of I: $c(I)$. The higher is I then the higher is α. Our signaling game is based on the assumption that investment cost I and α must be known, or common knowledge, before the game begins. The intuition is that the victim must have an idea how much the criminal has

[4] To the best of our knowledge only the United States of America state North-Carolina prohibited ransom payments by public entities. However, it is unclear what the penalty is and whether this also applies to double extortion ransomware (Lewis, 2022). More generally, it is not clear that sanctions are a strong deterrent for payment (Abely, 2022).

[5] For game theoretic analysis of ransomware and reputation we refer to (Li and Liao, 2021; Cartwright and Cartwright, 2019).



(a) (b)

Fig. 4. (a) Relationship between investment and alpha. (b) Relationship between investment and total expected utility criminal.

invested into data exfiltration and how that influences the probability of data exfiltration, α.

We assume that the relationship between investment I and α to be concave: initially the criminal has large investment cost. However, when the criminal decides to invest in the attack, we argue that once the right tooling and infrastructure is bought, additional investments would not increase the probability of data exfiltration substantially. Therefore the relationship between investment cost I and α is concave. While it is possible to utilize any concave function for our purposes, we have chosen to employ the Cumulative Distribution Function of the exponential distribution as an illustrative example (see Fig. 4(a), $\lambda = 1.5$).

$$\alpha(I) = 1 - e^{\lambda I} \tag{2}$$

Consequently, the expected payoff for the criminal in equilibrium B1 is:

$$U_{Total} = \alpha(I)(T - k^d) - (1 - \alpha(I))k^N + V - L - c - I + \epsilon \tag{3}$$

Plotting the total expected payoff of the criminal against investment, there seems an optimal investment which yield largest total expected payoff, see Fig. 4(b).[6] You can see that this results in an optimum value of α around 0.9. This is higher than we assumed in our calibration ($\alpha = 0.6$) but consistent with a relatively high α. If we increase the investment costs of data exfiltration then the optimal value of α is smaller. There are two other effects, however, we would highlight as likely to result in a smaller α in practice:

(a) Law Enforcement officers have informally disclosed that there are case studies where data exfiltration alerted the victim so that data encryption and exfiltration was not possible. In our model this suggests that increasing I could alert the victim so that they could quickly take action to stop the attack. Hence, increasing I does not necessarily increase α, and so the optimal value of α for the criminal may be correspondingly lower. In other words, the criminal may accept a lower probability of exfiltration α in order to increase the likelihood of the attack succeeding.

(b) If data exfiltration is costly the criminals might decide beforehand that they will not exfiltrate data. They might consider that the effort and costs of investments in infrastructure, and tooling might not be sufficient to cover the profits from data exfiltration. The criminal could then randomize between zero investment on some attacks and high investment on other attacks, to maximize their overall payoff. Given that a victim would not know if the criminals invested in data exfiltration in their particular attack, this again has the effect of lowering the value of α from that seen in Fig. 4.

Endogenizing variables of our model, such as α might be an interesting way to generalize our model and could be the focus of further research.

6. Conclusion

6.1. Main findings and limitations

Our paper contributes to an understanding of the ransomware business model in a setting with data exfiltration and double extortion. It is often the case (maybe typically the case) that the victim of a ransomware attack cannot know for sure whether data has been exfiltrated, particularly in the initial aftermath of the attack. If victims are willing to pay higher ransoms in the event of data exfiltration then it could be in the criminal's interests to signal data exfiltration. Drawing on a dataset provided by the Dutch Police and an incident response firm, we explored the issue of how credible victims find the claims of data exfiltration by attackers. Our findings indicate that victims display varying levels of confidence in whether data has actually been exfiltrated. Generally, the data suggests a greater willingness-to-pay on the part of victims when they believe their data has been compromised, incentivizing criminals to falsely claim that data has been exfiltrated.

We applied a signaling game model to analyse this information asymmetry, focusing on the interaction between a victim and criminal and Bayesian equilibrium strategies. Depending on various factors like the cost of sending signals and the reputation cost of actual data exfiltration, we identified five stable equilibria of the game. These range from scenarios where the attacker only signals when data is truly exfiltrated, to those where signals are sent or payments are made irrespective of the actual data exfiltration, to those where no signals are sent even if data is exfiltrated. Our analysis and calibration exercise suggests that the most likely real-world equilibrium outcome involves criminals signaling data exfiltration, whether or not exfiltration has actually occurred. Additionally, our study indicates that it could be strategically advantageous for criminals to invest more in attacks to enhance their chances of successful data exfiltration.

There are limitations in applying a game-theoretic framework to real-life situations. For instance, an assumption of common knowledge of game parameters is strong; given that ransomware remains fluid there is little opportunity for either victim or criminal to learn about each other through repeated interaction. Moreover, quantifying signaling costs and determining the extent of falsely generated credible signals pose challenges in real-life settings. Additionally, our model does not account for certain externalities, such as ethical considerations of the victim when deciding to pay the ransom, regardless of the costs or bankruptcy risks. Furthermore, the model does not account for multiple attacks by the criminal or the security behaviors of other potential victims, suggesting possible avenues for further research.

Another limitation of our analysis is that it does not explicitly differentiate between companies with and without recoverable data backups. Having recoverable backups does significantly influence the decision-making process of paying the ransom (Meurs et al., 2023b). Therefore, our presented model misses an important factor influencing the decision to pay. Although the focus of the current study is only on the

[6] With parameters set to: T = 10, $k^D = 1$, $k^N = 3$, L = 0, V = 0, c = 0, $\epsilon = 0$.

data-exfiltration during ransomware attacks, the proposed game theoretical models could be extended by differentiating situations with and without recoverable backups. A further way to extend our current game theoretic analysis is to consider the private information of the victim. Attackers may not know the value of the information in exfiltrated documents (Meurs and Holterman, 2022). For instance, the documents could be in a foreign language (from the criminals perspective) or there are to many documents for the criminal to assess. In a companion paper we analyse the value of private information of the victim and concluded that private information decreases the payoff of the criminal and increases the payoff of the victim (Meurs et al., 2023a).

A final limitation to consider is the applicability of Nash equilibrium. While Bayesian Nash Equilibrium describes an outcome in which no one wants to change their strategy, caution should be used in interpreting it as a prediction of behavior (Brandts and Holt, 1992). We may observe systematic deviations from Nash equilibrium or convergence on non-intuitive equilibria. Despite these limitations, we believe that a game theoretic analysis can still give useful insight on the incentives that ransomware criminals face. In our model we have seen the incentives for criminals to signal data exfiltration even if no exfiltration exists. This suggests that victims should be cautious of claims made by criminals, even if those claims seem credible. Importantly, there may be a 'ripple effect'; the more businesses believe data exfiltration has occurred when it did not, the more criminals have an incentive to falsely claim they have exfiltrated data. At face value, this would suggest it is in the victims interest to make it more costly for a ransomware criminal to falsely claim that data has been exfiltrated. We have seen, however, that this can have the perverse effect of benefiting the criminal (on average) because it can increase ransom demands if data is exfiltrated. It is important, therefore, to carefully consider the policy implications of our findings.

6.2. Recommendations for policy makers and potential victims

Our signaling game analysis yields several potential implications for policy makers and victims:

1. **Lowering the Probability of Data Exfiltration:** Victims should employ measures to decrease the likelihood of data exfiltration. In our model, this decreases α, which was one of the most important factors in determining the profitability of ransomware. The following strategies can be implemented to help achieve this:
    1. *Canary-files*: Victims can introduce "canary-files" throughout their network that generate alerts when copied or moved, thereby reducing the likelihood of successful data exfiltration (Meurs and Holterman, 2022).
    2. *Server take-down*: Engaging Law Enforcement to take down the server to which data is exfiltrated can disrupt the criminal's operations and possibly prevent successful data exfiltration. Although the criminal may have replicated the data on other servers, this action could leave a trace for investigation.
    3. *Spiking data*: Incorporating substantial amounts of fake data or deliberately contaminating the dataset can decrease the probability of valuable data being stolen (Li and Liao, 2022).
2. **Modifying Signal Credibility:** Interventions aimed at altering the cost of signaling data exfiltration should be considered. In our models, these were variables k^D and k^N for when criminals did or did not exfiltrate data respectively.
    1. *Increasing costs k^N*: Raising the costs associated with signaling data exfiltration can lead to a separating equilibrium, potentially resulting in higher payoffs for criminals. Paradoxically, investing in robust monitoring and logging systems may inadvertently increase the profitability of ransomware attacks. Criminals become more credible when they can provide a reliable signal, potentially justifying larger ransom demands. Therefore, increasing the costs of k^N does not seem an efficient defensive strategy on its own. To be effective it needs to be coupled with lowering the probability of data exfiltration α. If the probability of data exfiltration is low then an increase in signaling costs can disrupt the criminals profits because it becomes more readily apparent that data exfiltration has not taken place.
    2. *Increasing costs k^D*: The criminals profit can be disrupted by efforts to increase the costs of signaling actual data exfiltration. However, implementing this approach may prove challenging. In our model, k^D represents the cost of analysing the data to provide a credible signal, and the opportunity costs of pursuing subsequent attacks. Extending the negotiation process or demanding extensive amount or time consuming evidence of data exfiltration might increase the costs of signaling. This strategy aligns with empirical evidence suggesting that prolonging negotiations can result in reduced ransom demands (Meurs et al., 2022). However, it remains unclear whether criminals will easily accept additional demands for evidence of data exfiltration.
3. **Spillover Effects of Defensive Measures:** It is crucial to recognize the externality effect resulting from victims defending their sensitive data. The results of this study indicate that increased data protection measures benefit not only individual businesses but also other organizations possessing vulnerable data. In particular, the more businesses invest in preventing data exfiltration the lower will be the population probability of data exfiltration α. As we have said, a decrease in α is an effective way to disrupt the criminal business model. Policy makers should acknowledge this positive externality effect and consider providing government support for cyber security investments. Neglecting this aspect may result in suboptimal levels of cyber security prevention and recovery investments by businesses compared to the social optimum.
4. **Prohibition of Ransom Payments or tighter regulation:** Our framework does not directly capture the different pros and cons of banning ransom payments. It can, however, give insight if we think of the legal fees parameter, L, in our model. The higher is L then the lower the ransom the criminal can extract. Thus, banning ransom payments, to the extent it increases legal fees and fines, could be seen as beneficial, because it lowers the criminal's profit. However, a legal prohibition could drive ransom payments underground, and inadvertently lower the legal fees L because victims no longer seek the advice of lawyers and negotiators. Exploring the effect of secret ransom payment on social welfare could be a valuable direction for future research.

    As we have just argued, an increase in legal fees, which could include legal costs, but also negotiation costs and psychological costs, decreases the size of ransom. High legal fees, thus, disrupt the criminals profit. They do not, however, benefit the victim because they merely mean the victim is paying fees rather than ransom. One way to think of this from a societal perspective is in terms of regulation (short of banning payments). Higher levels of regulation (e.g. carefully enforced sanctions lists or requirements to alert law enforcement) would have the consequence of increasing L. The ideal would be to do this in a way that decreases ransom payments without driving ransom payments 'underground'.

In conclusion, our analysis provides valuable policy insights for addressing the challenges posed by double extortion ransomware attacks. Implementing measures to lower the probability of data exfiltration and manipulating signal credibility can help mitigate the impact of such attacks. Additionally, policymakers should consider the externality effect of increased data protection efforts and explore avenues for supporting cyber security investments to ensure social welfare is maximized.

6.3. Ethics

In conducting our research, we strictly adhere to the ethical principles outlined in the Menlo report (Bailey et al., 2012), which includes the following criteria: respect for persons, beneficence, justice, and respect for law and public interest.

Respect for Persons: We prioritize the autonomy and agency of individuals involved in our research. Cases were anonymized and no personal identifiable information is disclosed in this study. Our aim is to equip victims with effective methods to enhance their defensive mechanisms against cyber attacks.

Beneficence: Our research is guided by the principle of “do no harm” and aims to maximize probable benefits while minimizing potential harms. We conduct a thorough assessment of the risks and benefits associated with our research. Although the study of criminal decision-making risks of educating the criminal, we base our research on the principle of full-disclosure. Considering the entire study, we estimate that our model better informs victims and policy makers how to take preventive measures to prevent further harm than it educates criminals.

Justice: We uphold principles of fairness and equal consideration throughout our research. We strive to ensure that each individual is treated equitably and that the benefits resulting from our research are distributed to (potential) victims, companies and policy makers to prevent further harm of double extortion ransomware.

Respect for Law and Public Interest: We conduct our research with a strong commitment to legal compliance and transparency. We engage in legal due diligence with the Dutch Police and IR company to ensure that our research adheres to applicable laws and regulations. Permission was granted to use their data after anonymizing victims and showing that no personal identifiable information is used in this paper.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Tom Meurs reports financial support was provided by Netherlands National Police Corps.

Data availability

The data that has been used is confidential.

Acknowledgements

We would like to extend our sincere gratitude to the Dutch Police. In particular, we would like to thank Theo van der Plas and Cees van Tent for making the project possible. Furthermore, we thank the Cybercrime Unit East Netherlands and the Ransomware Taskforce for their expertise. Please note that the views expressed in this work are ours alone and do not necessarily reflect those of the Dutch Police. Furthermore, we would like to thank Northwave, in particular Pim Takkenberg, Erwin Maas, and Patrick van Looy.

Appendix A. Proof of propositions

Proposition 1. *If $V > L$ and $k^D < T < k^N$ there is a Bayesian equilibrium satisfying the D1 Criterion of the type A1. If $V < L$ and $k^D < T + V - L < k^N$ there is a Bayesian equilibrium satisfying the D1 Criterion of the type A2.*

Proof of Proposition 1. Equilibrium A1 can formally be written as follows: The criminal chooses to (i) Signal and set $R^S = T + V - L - \epsilon$ if type DE, and (ii) No Signal and set $R^{NS} = V - L - \epsilon$ if type NDE. The victim chooses (i) Pay if the criminal chooses Signal and asks ransom $R \leq R^S$, (ii) No Pay if Signal and $R > R^S$, (iii) Pay if No Signal and $R \leq R^{NS}$, and (iv) No Pay if No Signal and $R > R^{NS}$.

Consider the victim. Suppose the criminal has chosen signal and ransom R^S. The Bayesian updated belief of the victim should be $\mu(DE|S) = 1$. The expected payoff of the victim if they Pay is $U = -R^S - L = -T - V + \epsilon$. The expected payoff if they choose No Pay is $U = -T - V$. It is, thus, optimal to pay.

Suppose the Criminal has chosen No Signal. The Bayesian updated belief of the victim in this case is $\mu(NDE|NS) = 0$. The expected payoff of the victim if they Pay is $U = -R^{NS} - L = -V + \epsilon$. The expected payoff if they choose to No Pay is $U = -V + \epsilon$. This implies that it is optimal for the victim to pay if $L < V$ and optimal to Not Pay if $L \leq V$.

Consider now the incentive of the criminal. Suppose the criminal is type $\tau = DE$. On the equilibrium path they receive payoff $U = T + V - L - \epsilon - c - k^D$. We argue that, if the criminal chooses Signal, then they cannot gain from choosing $R \neq R^S$, provided $R^S > 0$: If they choose a ransom $R < R^S$ then the victim pays a smaller ransom, and if $R > R^S$ then the victim does not pay and the criminal has payoff $U = -c - k^D$. We have $R^S > 0$ if $T + V > L$. We also argue the criminal cannot gain from choosing No signal. In doing so, we distinguish two cases $L > V$ and $V > L$. If $L > V$ then the maximum ransom the victim is willing to pay is negative. Hence, the victim will choose No Pay and the criminal has payoff $U = -c$. It is, thus, optimal to Signal if $T + V - L > k^D$. If $L < V$ then the victim will pay a ransom up to $R = V - L - \epsilon$. The criminal can do no better than set ransom $R = V - L - \epsilon$. Thus, the criminals payoff is $U = V - L - \epsilon - c$. It is, thus, optimal to Signal if $T > k^D$.

Suppose the criminal is type $\tau = NDE$. The argument above naturally extends to this case, except we now derive, respective, conditions $T > k^N > k^D$ and $T + V - L > k^N > k^D$. Proving $\mu(NDE|NS) = 0$ is trivial: in the separating equilibrium the strategy of the Criminal is type $\tau = NDE$ is to not signal and for type $\tau = DE$ to signal. Therefore $\mu(NDE|NS) = 0$ and $\mu(DE|S) = 1$ is consistent with the strategy of the criminal in the equilibrium.

Equilibrium A2 can formally be written as follows: The criminal chooses to (i) Signal and set $R^S = T + V - L - \epsilon$ if type DE, and (ii) No Signal and set $R^{NS} = 0$ if type NDE. The victim chooses (i) Pay if the criminal chooses Signal and asks ransom $R \leq R^S$, (ii) No Pay if Signal and $R > R^S$, (iii) No Pay if No Signal and $R \geq R^{NS}$. The arguments for the proof of existence of equilibrium A1 can naturally be applied to show proof of the existence of equilibrium A2. □

Proposition 2. *If (a) $V > L$ and $\alpha T > k^N$, or (b) $V + \alpha T > L > V$ and $\alpha T + V - L > k^N$ there is a signaling equilibrium satisfying the D1 Criterion of the type B1.*

Proof of Proposition 2. The equilibrium has the following properties: The criminal chooses Signal and $R^S = \alpha T + V - L - \epsilon$ if $\tau = NDE$ and $\tau = DE$. The victim chooses (i) Pay if Signal and $R \leq R^S$, (ii) No Pay if Signal and $R > R^S$, (iii) Pay if No Signal and $R \leq V - L - \epsilon$, and (iv) No Pay if No Signal and $R \geq V - L$.

Consider the victim. Suppose the Criminal has chosen Signal and ransom R^S. The Bayesian updated belief of the victim should be $\mu(DE|S) = \alpha$. So, the expected payoff of the victim if they Pay is $U = -R^S - L = -\alpha T - V + \epsilon$. The expected payoff if they choose No Pay is $U = -\alpha T - V$. It is, thus, optimal to pay.

Suppose the Criminal has chosen No Signal. For now we assume the belief of the victim is $\mu(DE|NS) = 0$. Suppose the ransom is $R = V - L - \epsilon$. The expected payoff of the victim if they Pay the ransom is $U = -V + \epsilon$. The expected payoff if they choose No Pay is $U = -V$. It is, thus, optimal for the victim to Pay.

Consider now the incentive of the criminal. Suppose the criminal is type $\tau = DE$. On the equilibrium path they receive payoff $U = \alpha T + V - L - \epsilon - c - k^D$. We argue that, if the criminal chooses Signal, then they cannot gain from choosing $R \neq R^S$, provided $R^S > 0$: If they choose a ransom $R < R^S$ then the victim pays a smaller ransom, and if $R > R^S$ then the victim does not pay and the criminal has payoff $U = -c - k^D$. We have $R^S > 0$ if $\alpha T + V > L$. We also argue the criminal cannot gain from choosing No signal. In doing so, we distinguish two cases $L > V$ and $V > L$. If $L > V$ then the maximum ransom the victim is willing to pay is negative. Hence, the victim will choose No Pay and the criminal has payoff $U = -c$. It is, thus, optimal to Signal if $\alpha T + V - L > k^D$. If $L < V$ then the victim will pay a ransom up to $R = V - L - \epsilon$. The criminal can do no better than set ransom $R = V - L - \epsilon$. Thus, the criminals payoff is $U = V - L - \epsilon - c$. It is, thus, optimal to Signal if $\alpha T > k^D$.

Suppose the criminal is type $\tau = NDE$. The argument above naturally extends to this case, except we now derive, respective, conditions $\alpha T > k^N > k^D$ and $\alpha T + V - L > k^N > k^D$.

It remains to show that $\mu(DE|NS) = 0$. Here, we invoke the D1 Criterion. Consider $V > L$. Suppose that the criminal chooses No Signal and ransom demand $R > 0$. Let p be the probability that the victim will pay. The type $\tau = DE$ will receive a weakly higher payoff than in equilibrium if

$$p^{DE} \geq \frac{\alpha T + V - L - k^D}{R}.$$

The type $\tau = NDE$ will receive a strictly higher payoff than in equilibrium if

$$p^{NDE} > \frac{\alpha T + V - L - k^N}{R}.$$

Given that $k^N > k^D$ we have $p^{NDE} < p^{DE}$. Thus, type $\tau = DE$ can be eliminated using the D1 Criterion. It follows that $\mu(DE|NS) = 0$ is consistent with the D1 Criterion. □

Proposition 3. *If $V + \alpha T > L$ and $(1 - \alpha)T < k^D$ there exists a signaling equilibrium satisfying the D1 Criterion of the type C1. If $V + \alpha T > L > V$ and $T + V - L < k^D$ there exists a signaling equilibrium satisfying the D1 Criterion of the type C2*

Proof of Proposition 3. Equilibrium C1 has the property: The criminal chooses Signal and $R^{NS} = \alpha T + V - L - \epsilon$ if $\tau = NDE$ and $\tau = DE$. The victim chooses (i) Pay if Signal and $R \leq R^S$, (ii) No Pay if Signal and $R > R^S$, (iii) Pay if No Signal and $R < R^{NS}$, and (iv) No Pay if No Signal and $R > R^{NS}$.

Consider the victim. Suppose the Criminal has chosen Signal and ransom R^S. The Bayesian updated belief of the victim should be $\mu(DE|S) = 1$. So, the expected payoff of the victim if they Pay is $U = -R^B - L = -T - V + \epsilon$. The expected payoff if they choose No Pay is $U = -T - V$. It is, thus, optimal to pay.

Suppose the Criminal has chosen No Signal. For now we assume the belief of the victim is $\mu(DE|NS) = \alpha$. Suppose the ransom is $R^{NS} = \alpha T + V - L - \epsilon$. The expected payoff of the victim if they Pay the ransom is $U = -\alpha T - V + \epsilon$. The expected payoff if they choose No Pay is $U = -V$. It is, thus, optimal for the victim to Pay if $(1 - \alpha)T < k^D$.

Consider now the incentive of the criminal. Suppose the criminal is type $\tau = DE$. On the equilibrium path they receive payoff $U = \alpha T + V - L - \epsilon - c$. We argue that, if the criminal chooses Signal, then they cannot gain from choosing $R \neq R^{NS}$, provided $R^{NS} > 0$: If they choose a ransom $R < R^{NS}$ then the victim pays a smaller ransom, and if $R > R^{NS}$ then the victim does not pay and the criminal has payoff $U = -c - k^D$. We have $R^{NS} > 0$ if $\alpha T + V > L$. We also argue the criminal cannot gain from choosing No signal. In doing so, we distinguish two cases $L > V$ and $V > L$. If $L > V$ then the maximum ransom the victim is willing to pay is negative. Hence, the victim will choose No Pay and the criminal has payoff $U = -c$. It is, thus, optimal to Signal if $\alpha T + V - L > k^D$. If $L < V$ then the victim will pay a ransom up to $R = \alpha T + V - L - \epsilon$. The criminal can do no better than set ransom $R = \alpha T + V - L - \epsilon$. Thus, the criminals payoff is $U = \alpha T + V - L - \epsilon - c$. It is, thus, optimal to Not Signal if $(1 - \alpha T) < k^D$ or $V + T - L < k^D$.

Suppose the criminal is type $\tau = NDE$. The argument above naturally extends to this case, except we now derive, respective, conditions $(1 - \alpha T) < k^D < k^N$ and $T + V - L < k^D < k^N$.

It remains to show that $\mu(DE|S) = 1$. Here, we invoke the D1 Criterion. Consider $V > L$. Suppose that the criminal chooses Signal and ransom demand $R > 0$. Let p be the probability that the victim will pay. The type $\tau = DE$ will receive a weakly higher payoff than in equilibrium if

$$p^{DE} \geq \frac{T + V - L - k^D}{R}.$$

The type $\tau = NDE$ will receive a strictly higher payoff than in equilibrium if

$$p^{NDE} > \frac{T + V - L - k^N}{R}.$$

Given that $k^N > k^D$ we have $p^{NDE} < p^{DE}$. Thus, type $\tau = NDE$ can be eliminated using the D1 Criterion. It follows that $\mu(NDE|S) = 0$ is consistent with the D1 Criterion. It follows that $\mu(DE|S) = 1$ is consistent with the D1 Criterion. □

References

Abely, C., 2022. Ransomware, cyber sanctions, and the problem of timing. BCL Rev. E. Supp. I 63, 47.

Akerlof, G.A., 1970. The market for “lemons”: quality uncertainty and the market mechanism. Q. J. Econ. 84 (3), 488–500.

Allodi, L., Massacci, F., Williams, J.M., 2017. The work-averse cyber attacker model: theory and evidence from two million attack signatures. SSRN Electron. J.

Bailey, M., Dittrich, D., Kenneally, E., Maughan, D., 2012. The menlo report. IEEE Secur. Priv. 10 (2), 71–75.

Banks, J.S., Sobel, J., 1987. Equilibrium selection in signaling games. Econometrica, 647–661.

Bedard, K., 2001. Human capital versus signaling models: university access and high school dropouts. J. Polit. Econ. 109 (4), 749–775.

Beebe, N.L., Rao, V.S., 2005. Using situational crime prevention theory to explain the effectiveness of information systems security. Las Vegas.

Brandts, J., Holt, C.A., 1992. An experimental test of equilibrium dominance in signaling games. Am. Econ. Rev. 82 (5), 1350–1365.

Brewer, R., 2016. Ransomware attacks: detection, prevention and cure. Netw. Secur. 2016 (9), 5–9.

Cartwright, A., Cartwright, E., 2019. Ransomware and reputation. Games 10 (2), 26.

Cartwright, E., Hernandez Castro, J., Cartwright, A., 2019. To pay or not: game theoretic models of ransomware. J. Cybersecurity 5 (1), tyz009.

Clarke, R.V., 2016. Situational crime prevention. In: Environmental Criminology and Crime Analysis. Routledge, pp. 305–322.

Connolly, L.Y., Lang, M., Taylor, P., Corner, P.J., 2021. The evolving threat of ransomware: from extortion to blackmail.

Corbet, S., Goodell, J.W., 2022. The reputational contagion effects of ransomware attacks. Finance Res. Lett. 47, 102715.

Cornish, D.B., Clarke, R.V., 1987. Understanding crime displacement: an application of rational choice theory. Criminology 25 (4), 933–948.

Cornish, D.B., Clarke, R.V., 2014. The reasoning criminal: Rational choice perspectives on offending.

Culafi, A., 2022. Coveware: Double-extortion ransomware attacks fell in q1.

Cymru, T., 2022. Analyzing ransomware negotiations with conti: an in-depth analysis.

Ecrime, 2023. Gallery of 97 ransomware and data leak sites.

Fudenberg, D., Tirole, J., 1991. Game Theory. MIT Press.

Galinkin, E., 2021. Winning the ransomware lottery: a game-theoretic approach to preventing ransomware attacks. In: Decision and Game Theory for Security: 12th International Conference, GameSec 2021, Virtual Event, October 25–27, 2021, Proceedings 12. Springer, pp. 195–207.

Glazer, A., Konrad, K.A., 1996. A signaling explanation for charity. Am. Econ. Rev. 86 (4), 1019–1028.

Gonzalez, D., Hayajneh, T., 2017. Detection and prevention of crypto-ransomware. In: 2017 IEEE 8th Annual Ubiquitous Computing, Electronics and Mobile Communication Conference (UEMCON). IEEE, pp. 472–478.

Greengard, S., 2021. The worsening state of ransomware. Commun. ACM 64 (4), 15–17.

Hack, P., Wu, Z.-Y., 2021. “We wait, because we know you.” Inside the ransomware negotiation economics. NCC Group, Nov, 12.

Hodge, N., 2023. Paying ransom to avoid gdpr fine an unwise gambit.

Huang, K., Siegel, M., Madnick, S., 2018. Systematically understanding the cyber attack business: a survey. ACM Comput. Surv. 51 (4), 1–36.

Kerns, Q., Payne, B., Abegaz, T., 2022. Double-extortion ransomware: A technical analysis of maze ransomware. In: Proceedings of the Future Technologies Conference (FTC) 2021, vol. 3. Springer, pp. 82–94.

Kreps, D.M., Sobel, J., 1994. Signalling. Handbook of Game Theory With Economic Applications, vol. 2, pp. 849–867.

Laszka, A., Farhang, S., Grossklags, J., 2017. On the economics of ransomware. In: Decision and Game Theory for Security: 8th International Conference. GameSec 2017, Vienna, Austria, October 23-25, 2017, Proceedings. Springer, pp. 397–417.

Laszka, A., Panaousis, E., Grossklags, J., 2018. Cyber-insurance as a signaling game: self-reporting and external security audits. In: Decision and Game Theory for Security: 9th International Conference, GameSec 2018, Seattle, WA, USA, October 29–31, 2018, Proceedings 9. Springer, pp. 508–520.

Lewis, J., 2022. North Carolina prohibits public sector entities paying ransom ransomware cyberattack. Natl. Law Rev.

Li, Z., Liao, Q., 2020. Ransomware 2.0: to sell, or not to sell a game-theoretical model of data-selling ransomware. In: Proceedings of the 15th International Conference on Availability, Reliability and Security, pp. 1–9.

Li, Z., Liao, Q., 2021. Game theory of data-selling ransomware. J. Cyber Secur. Mobil., 65–96.

Li, Z., Liao, Q., 2022. Preventive portfolio against data-selling ransomware—a game theory of encryption and deception. Comput. Secur. 116, 102644.

Meurs, T., Holterman, L., 2022. Whitepaper data-exfiltratie bij een ransomware-aanval.

Meurs, T., Junger, M., Tews, E., Abhishta, A., 2022. Ransomware: how attacker's effort, victim characteristics and context influence ransom requested, payment and financial loss. In: Symposium on Electronic Crime Research, eCrime 2022.

Meurs, T., Cartwright, E., Cartwright, A., 2023a. Double-sided information asymmetry in double extortion ransomware. In: 14th International Conference on Decision and Game Theory for Security, GameSec 2023.

Meurs, T., Cartwright, E., Cartwright, A., Junger, M., Hoheisel, R., Tews, E., Abhishta, A., 2023b. Ransomware economics: a two-step approach to model ransom paid. In: 18th Symposium on Electronic Crime Research, eCrime 2023.

Mott, G., Turner, S., Nurse, J.R., MacColl, J., Sullivan, J., Cartwright, A., Cartwright, E., 2023. Between a rock and a hard (ening) place: cyber insurance in the ransomware era. Comput. Secur. 128, 103162.

Ng, Y.-K., 1984. Expected subjective utility: is the Neumann-Morgenstern utility the same as the neoclassical's? Soc. Choice Welf. 1 (3), 177–186.

Nyakomitta, P.S., Abeka, S.O., 2020. A survey of data exfiltration prevention techniques. Int. J. Adv. Netw. Appl. 12 (3), 4585–4591.

Osborne, M.J., et al., 2004. An Introduction to Game Theory, vol. 3. Oxford University Press, New York.

Palmer, D., 2023. The ransomware problem isn't going away, and these grim figures prove it.

Pattnaik, N., Nurse, J.R., Turner, S., Mott, G., MacColl, J., Huesch, P., Sullivan, J., 2023. It's more than just money: the real-world harms from ransomware attacks. In: International Symposium on Human Aspects of Information Security and Assurance. Springer, pp. 261–274.

Payne, B., Mienie, E., 2021. Multiple-extortion ransomware: the case for active cyber threat intelligence. In: ECCWS 2021 20th European Conference on Cyber Warfare and Security. Academic Conferences Inter Ltd., p. 331.

Richardson, R., North, M.M., 2017. Ransomware: evolution, mitigation and prevention. Int. Manag. Rev. 13 (1), 10.

Spence, M., 1974. Competitive and optimal responses to signals: an analysis of efficiency and distribution. J. Econ. Theory 7 (3), 296–332.

Tuttle, H., 2021. Ransomware attackers turn to double extortion. Risk Manag. 68 (2), 8–9.

Von Neumann, J., Morgenstern, O., 1944. Theory of games and economic behavior. In: Theory of Games and Economic Behavior. Princeton University Press.

Wortley, R., Townsley, M., 2016. Environmental criminology and crime analysis: situating the theory, analytic approach and application. In: Environmental Criminology and Crime Analysis. Routledge, pp. 20–45.

Xu, Z., Hu, Q., 2018. The role of rational calculus in controlling individual propensity toward information security policy non-compliance behavior. In: Proceedings of the 51st Hawaii International Conference on System Sciences. Hawaii International Conference on System Sciences.

Tom Meurs received the B.S. degree in Methodological Psychology in 2016 and B.S. and M.S. degrees in Econometrics from the University of Amsterdam in 2017 and 2018. He is currently a Ph.D. candidate at the University of Twente. The PhD project is funded by the Dutch Police. His research interests include ransomware, coordination of cybercrimes, Initial Access Brokers and communication between cybercriminals.

Edward Cartwright a Professor of Economics at De Montfort University and Director of the Institute for Applied Economics and Social Value. His-research interests include cyber security, game theory and behavioural economics, with particular interest in the economics of ransomware and the adoption of cyber secure behaviour in micro and small organisations. Recent projects include RAMSES (Internet forensic platform for tracking the money flow of financially-motivated malware) and EMPHASIS (Economical, psychological and societal impact of ransomware). He co-developed the Leicester Stories platform and is currently supporting the East Midlands Chambers of Commerce to establish a Regional Business Intelligence Unit and Collective Intelligence Skills Unit.

Anna Cartwright is a Principal Lecturer in Economics at Oxford Brookes University. She is also a RISCS Senior Fellow on the Theme of Quantification and Cyber Risk. Her research interests include the economics of cyber security, industrial economics and game theory. She led a Home Office funded project on cyber behaviour in micro organisations that delivered and evaluated cyber security health checks aimed at micro organisations. As a RISCS Fellow she led a research project evaluating the role of local IT companies in disseminating cyber best practice to micro organisations. A particular interest is how to measure and quantify cyber risk in organisations, large and small.

Marianne Junger received the Ph.D. degree in law from the Free University of Amsterdam, Amsterdam, the Netherlands, in 1990. She is the Emeritus Professor of Cyber Security and Business Continuity with the University of Twente, Enschede, the Netherlands. Her research investigates the human factors of fraud and cybercrime. More specifically, she investigates online victimization, disclosure, and privacy issues. She founded the Crime Science journal together with Pieter Hartel and was an Associate Editor for 6 years. Her research was sponsored by, among others, the Dutch Police, NWO, ZonMw (for health research), and the European Union.

Abhishta Abhishta is an assistant professor at the High-tech Business and Entrepreneurship department at University of Twente. His research focuses on empirically measuring the economic/financial impact of cyber attacks. To do so, he devises/adapts data-driven economic impact assessment techniques. He looks to help organisations make well-measured investments in security. His doctoral research was funded under NWO project D3 – Distributed Denial-of-Service Defense: protecting schools and other public organizations. His current research is supported by two NWO grants, one aimed at cloud security (MASCOT) and the other at building a first prototype of the Responsible Internet (CATRIN). He serves on the program committee of ACM/IEEE/IFIP conferences aimed at network measurements and responsible internet (TMA, PAM, TAURIN).