Double-sided Information Asymmetry in Double Extortion Ransomware

Tom Meurs[1[0000−0003−0963−5232]], Edward Cartwright[2[0000−0003−0194−9368]], and Anna Cartwright[3[0000−0003−1965−842X]]

[1] University of Twente: Enschede, Netherlands t.w.a.meurs@utwente.nl
[2] De Montfort University: Leicester, Leicestershire, Great Britian edward.cartwright@dmu.ac.uk
[3] Oxford Brookes University: Oxford, Oxfordshire, Great Britian a.cartwright@brookes.ac.uk

Abstract. Double extortion ransomware attacks consist of an attack where victims files are both encrypted and exfiltrated for extortion purposes. There is empirical evidence this leads to an increased willingness to pay a ransom, and higher ransoms, compared to encryption-only attacks, depending on the value of the exfiltrated files. However, there seem to be two complications: First, victims are uncertain whether data is exfiltrated, due to for example misconfigured monitoring systems. Second, it is hard for attackers to estimate the value of compromised files. Thus, victims have an incentive to hide what they know and attackers an incentive to find out information. The goal of this study is to use game theory to explore the payoff consequences for attackers of victims having private information. We analyse a signaling game with double-sided information asymmetry: (1) attackers know whether data is exfiltrated and victims do not, and (2) victims know the value of data if it is exfiltrated, but the attackers do not. Our analysis of the game indicates that private information substantially lowers the return to attackers. These results imply that victims should be careful to not reveal the value of files during negotiations.

Keywords: Ransomware · Data exfiltration · Information asymmetry

1 Introduction

The last decade has seen a rapid rise in crypto-ransomware attacks [7,8,19,10,20,23]. Crypto-ransomware, or ransomware for short, is broadly defined as the use of crypto-techniques to encrypt the files of a victim, after which the attackers ask for a ransom to decrypt the files [29]. Ransomware has proved highly profitable for criminal gangs, primarily because many victims pay the ransom in order to receive the decryption keys [21]. Since roughly 2019, ransomware groups have been experimenting with double extortion [13,6]. In this case the attackers not only encrypt files, but also exfiltrate data with the purpose to sell or publish the data if the victim does not pay [16,17,20]. Double extortion has increased the ransom requested and probability of victims paying [19].

One important issue for victims of a ransomware attack is determining whether data was exfiltrated. Due to the deletion of log files by attackers, or misconfigured monitoring systems, victims often do not know whether data was exfiltrated [25,26]. This means that an attacker who has not exfiltrated data can still threaten the publication of data, to get a larger ransom paid. On the flip side, the claims of an attacker that has exfiltrated data may be viewed as less-credible, empty threats, by the victim. Attackers are, thus, increasingly trying to send credible signals that data was exfiltrated. For instance, to back up their claim, some attackers send evidence of exfiltration by means of a file tree of the exfiltrated data or a couple of files. Such signals could, however, still be sent, even if at a higher cost, by attackers who have not exfiltrated data.

Another limitation of sending 'evidence' of data exfiltration is that it might give the victim the opportunity to determine the value of the exfiltrated data. In practice, it is hard for attackers to determine the value of the files to the victim. The filenames and files which contain text are often in a foreign language, and the sensitivity of data is difficult to judge without insider understanding. Furthermore, it takes effort to estimate the importance of, potentially, millions of files. Attackers are, therefore, likely to be imperfectly informed of the value of files, even if data is exfiltrated. Combined, therefore, we have two information asymmetries in double extortion ransomware attacks. First, the victim does not know whether data was exfiltrated or not, but the attacker does. Second, the victim can assess whether potentially exfiltrated data is valuable or not, but the attacker cannot. Here, we define valuable data for the victim, as data with large reputation costs if it gets accessible for the general public, competitors or similar.

To our knowledge, no previous studies have modelled this two-sided information asymmetry of data exfiltration, and analysed how it effects the profitability of attacks. Most empirical [19] and game-theoretical modeling [16,17] of double extortion ransomware has focused on the extra profits for attackers by conducting data exfiltration and encryption, compared to only data encryption. We address the relationship between the uncertainty of data exfiltration and profitability by analysing a signaling game. Signaling games provide a way to model a strategic game with incomplete information and sequential choice [11,14,1,18,22]. The basic premise is that a player holding extra information could try to influence the other players by sending a credible signal of their information. Signalling games provide a natural framework with which to explore double extortion and the payoff consequences of assymetric information. For a more detailed explanation of signaling games we refer to [22].

Our work provides the following key contributions: First, we provide a game-theoretical framework to analyse the double-sided information asymmetry in double extortion ransomware attacks. The framework consists of a signaling game, wherein the attacker can send a costly signal of data exfiltration that can inform the victim's beliefs and payment decision. Second, we identify four separating and four pooling equilibria of the game and their underlying conditions. The type of equilibria that exists in the game will depend on the parameters of the game, particularly the cost of signaling data exfiltration, the cost to recover files without decryption, the reputation loss from data leakage, and the probability the victim's files contain valuable data. We identify the factors determining how much surplus the attacker can extract from the victim. Third, we analyse the impact that private information of the victim has on the profitability of the attacked. Through examples, we show that the payoff loss to the criminal from now knowing the value of files can range from zero to over 20%. Private information can, therefore, potentially disrupt the business model of ransomware games by reducing the profits they can make.

We remark that our paper adds to a growing literature using game theory to analyse the ransomware decision process [5,12,4]. Prior game-theoretical studies have focused on the interaction of ransomware and victim's decision to invest in security measures like backups or insurance [29,2,24,28]. For instance, Laszka, Farhang and Grossklags [15] focused on modeling the ransomware ecosystem as a whole and how backup decisions affect the ransomware ecosystem. Vakilinia et al. [27] take a different approach in exploring how a double sided auction can facilitate the negotiation between attacker and victim to achieve a 'fair' ransom. Galinkin [12] analyses measures that an attacker can disrupt the business model of the attackers by lowering the profitability of ransomware attacks. The main intervention suggested is that of back-ups. We note, however, that in a setting with double extortion, back-ups are not enough to combat the ransomware threat. We must also consider the reputational costs from the publication of exfiltrated data.

We proceed as follows. In Section 2 we introduce the signalling game. In Section 3 we provide our main results. In Section 4 we conclude.

2 Signaling Game

We consider a game between a criminal, henceforth called the attacker, and a victim. We take as given that the victim has been subject to a ransomware attack and their data has been encrypted. The attacker is demanding a ransom for the decryption key. If the victim does not pay the ransom then it will cost V_P to recover normal operations. If the victim does pay the ransom then we assume the attackers will provide the decryption key and it will cost V_{NP} to recover normal operations. From a game theoretic point of view, the predictions of our model depend solely on the difference in recovery cost from paying versus not paying $V_P - V_{NP}$. Thus, to simplify the model, and without loss of generality, we set $V_{NP} = 0$ and $V_P = V$. We assume that $V > 0$ and so access to the decryption key reduces recovery costs.

We take it as given that, as well as encrypting files, the attacker attempted to exfiltrate data from the victim. This attempt may or may not have been 'successful'. In either case, the attacker can threaten to publish exfiltrated data unless the ransom is paid. We model two forms of incomplete information:

1. The attacker knows if data is exfiltrated but the victim does not know. Let α denote the probability that data was exfiltrated. The value of α is

common knowledge to attacker and victim. We use the term NDE and DE to distinguish the type of attacker as no data exfiltration and data exfiltration, respectively.

2. The victim knows the reputational damage that would result from data exfiltration but the attacker does not know. We assume that there are two types of victim: those with sensitive data, called high type, and those without, called low type. If exfiltrated data were to be leaked then the victim would incur reputation costs T_1 or $T_0 < T_1$ depending on whether they are high or low type, respectively. If the data is not leaked then we assume there is no reputation cost. The probability the victim is high type is β. The value of β is common knowledge to attacker and victim.

The game has three stages. Following the approach of Harnsanyi [9], Nature determines the the type of the victim (high or low type) and the type of the criminal (data exfiltrated or no data exfiltrated) in Stage 1 of the game. The victim learns their type (with probability β they are high type), and the attacker learns whether data was exfiltrated (with probability α it is exfiltrated).

In stage 2 the attacker chooses (a) whether or not to send a signal that data has been exfiltrated, and (b) a ransom demand. The signal can, for instance, consist of a picture of the file tree of the exfiltrated data, or a sample of exfiltrated data. The cost to the attacker of sending a signal when data is exfiltrated is k_D, whereas if data is not exfiltrated it is k_N. We assume that it is harder to send a credible signal if no data is exfiltrated, so $k_D < k_N$. The attacker can choose any ransom demand. To simplify notation we denote by R^S the ransom demand of the attacker if they send a signal and R^{NS} the demand if no signal is sent.[4]

In stage 3 the victim observes whether or not a signal was sent, and learns the ransom demand. The victim then chooses whether to pay the ransom or not. To simplify the analysis we assume an ultimatum bargaining game in which their is no opportunity for negotiation, and the choice to pay or not ends the game.

The variables of the game are summarized in Table 1. One additional variable we introduce is $L \geq 0$ which captures the legal fees and costs (including psychologically and moral) of paying a ransom. We also introduce variable μ to represent the beliefs of the victim on the likelihood that data has been exfiltrated. Finally, we use variable ϵ to represent the smallest unit of currency. This will allow us to characterise the optimal ransom in a more succinct way. We exclude from the analysis any fixed costs incurred by the attacker and victim that are not dependent on the strategic elements of the game. For instance we do not include the cost to the attacker of implementing the attack. We can exclude such

[4] The attacker 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 an attacker of type DE and NDE signal and $R^S_{NDE} \neq R^S_{DE}$; this would reveal the attacker if type NDE and, thus, make their signal ineffective. Similarly, there cannot be an equilibrium in which an attacker of type DE and NDE would not signal and $R^{NS}_{NDE} \neq R^{NS}_{DE}$; this would again reveal the attacker if type NDE and lower the ransom the victim would rationally pay.

Table 1: Variables used in the data exfiltration signaling game

	Variable	Description
Attacker	R^S	Ransom when signaling
	R^{NS}	Ransom when not signaling
	k_D	Cost of signal with data exfiltration
	k_N	Cost of signal without data exfiltration
	β	Probability of data being valuable
Victim	T_1	Reputation cost for valuable data
	T_0	Reputation cost for non-valuable data
	V	Recovery cost without decryption key
	L	Legal fees of paying ransom
	α	Probability of data exfiltration
	μ	Belief on probability of data exfiltration
	ϵ	The smallest unit of currency

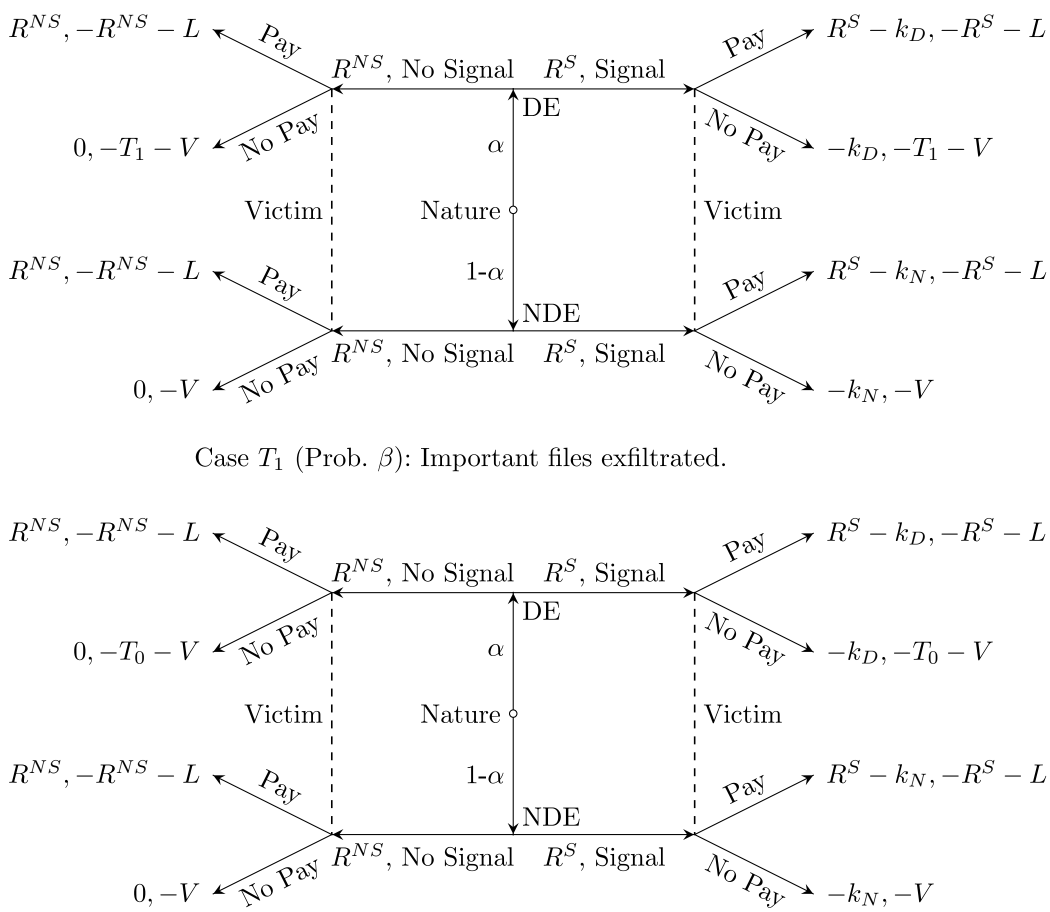
costs, without loss of generality, because they will not influence the equilibrium outcomes of the game. We depict the game in Figure 1.

3 Results

In the following we solve for Bayesian equilibria of the game that satisfy the D1 Criterion [11]. Informally, a Bayesian equilibrium has the property that both attacker and victim: (1) maximise their expected payoffs given the strategy of the other and their beliefs, (2) update their beliefs using Bayes rule. Thus, in equilibrium, players appropriately interpret information, and have no incentive to change their actions given their beliefs and the actions of the other player. The D1 Criterion is used to place 'common sense' restrictions on beliefs. Specifically, a Bayesian equilibrium may not tie down beliefs off the equilibrium path, because play could reach nodes that have zero probability and so Bayes rule is indeterminate. The D1 Criterion imposes extra conditions on beliefs by saying that any deviation from the equilibrium path is assumed to be done by the type with the most incentive to deviate [3].

The D1 Criterion is useful to rule out equilibria sustained by 'non-intuitive beliefs' [14]. For instance, consider a candidate equilibria in which the attacker chooses to not signal if they are type DE or NDE. On the equilibrium path the attacker should not signal. Thus, Bayes rule does not impose any restrictions on beliefs were the attacker to signal. Yet, informally, a type DE has the most incentive to deviate and signal. The D1 Criterion would, thus, require the victim to believe the deviation was by a type DE. This rules out 'non-intuitive' equilibria that are only sustained by the victim believing a signal would be from the type NDE.

To focus the analysis on what we believe are the most realistic cases, we distinguish and characterize two broad types of equilibrium: (a) separating equilibria in which the type DE signals data is exfiltrated and the type NDE does



Case T_1 (Prob. β): Important files exfiltrated.

Case T_0 (Prob. 1-β): No important files exfiltrated.

Fig. 1: Description of the game.

not, and a (b) pooling equilibria in which both the type DE and NDE signal that data is exfiltrated. We exclude from analysis pooling equilibria in which both the type DE and NDE do not signal that data is exfiltrated, as well as hybrid equilibria in which the attacker randomises their actions. In the following we discuss separating and pooling equilibria in turn before analysing the impact of private information. Throughout, we assume that if the victim is indifferent between paying and not paying then they will not pay.

3.1 Separating Equilibrium

A separating equilibrium has the basic characteristic that the attacker signals data exfiltration if they are of type DE (i.e. data was exfiltrated) and does not signal if they are of type NDE (i.e. data was not exfiltrated). The existence of a separating equilibrium and the exact form of any equilibrium will depend on the parameters of the game. Specifically, we identified four types of separating equilibria, which we will label A1-A4. These are summarised in Table 2. As you can see the equilibria differ by whether or not the victim pays the ransom. For example, in equilibrium A3 the victim pays the ransom if the attacker signals but does not pay the ransom if the attacker does not signal. In equilibrium A4 the victim only pays if they are a high type and the attacker signals.

In all four equilibria A1-A4 the high type victim pays if they receive a signal of data exfiltration. The equilibria differ in whether a low type victim pays if they receive a signal of data exfiltration and/or whether the victim pays if they receive no signal. To provide some intuition for the four equilibria we introduce three ransom demands that prove particularly relevant: (1) $R^*_{S0} = T_0 + V - L - \epsilon$, (2) $R^*_{S1} = T_1 + V - L - \epsilon$, and (3) $R^*_{NS} = \max\{V - L - \epsilon, 0\}$. Informally, see the proof of Theorem 1 for the full details, R^*_{S0} and R^*_{S1} are the maximum ransom the low type and high type, respectively, are willing to pay if they believe data has been exfiltrated. While, R^*_{NS} is the maximum ransom the attacker can ask if the victim believes data is not exfiltrated. We readily see that if $V < L$ the victim would not pay any positive ransom demand if they know data has not been exfiltrated.

Equilibrium	Attacker		Victim			
	DE	NDE	T_1		T_0	
			Signal	No signal	Signal	No signal
A1	Signal	No signal	Pay	Pay	Pay	Pay
A2	Signal	No signal	Pay	Pay	No pay	Pay
A3	Signal	No signal	Pay	No Pay	Pay	No Pay
A4	Signal	No Signal	Pay	No Pay	No pay	No pay
B1	Signal	Signal	Pay	Pay	Pay	Pay
B2	Signal	Signal	Pay	Pay	No pay	Pay
B3	Signal	Signal	Pay	No pay	Pay	No pay
B4	Signal	Signal	Pay	No pay	No pay	No pay

Table 2: Equilibria satisfying the D1 criterion in the signaling game.

If data exfiltration is believed to have taken place then the high type is willing to pay a larger ransom than the low type, $R^*_{S1} > R^*_{S0}$. This provides a strategic trade-off for the attacker: (a) if they ask for a high ransom, R^*_{S1}, then they extract maximum surplus from the high type victim, but the low type will not pay the ransom. (b) If they ask for a low ransom, R^*_{S0}, then both the low and high type victim will pay the ransom but they do not fully extract surplus from the high type. This trade-off is captured by the following term:

$$\Phi_S = \beta(R^*_{S1} - R^*_{S0}) - (1-\beta)R^*_{S0} = \beta(T_1 - T_0) - (1-\beta)(T_0 + V - L - \epsilon). \quad (1)$$

The first term in Φ_S is the expected gain for the attacker from extracting maximum surplus from the high type, while the second term is the expected loss from charging a ransom the low type is not willing to pay.

We are now in a position to state our first main result. As the preceding discussion preempts we need to consider combinations of $V \gtreqless L$ and $\Phi \gtreqless 0$ giving rise to the four different cases and equilibria.

Theorem 1. *There exists a separating equilibrium satisfying the D1 criterion if and only if the following conditions hold:*

(A1) If $L < V$ *and* $\Phi_S < 0$ *then* $k_D < T_0 < k_N$*.*
(A2) If $L < V$ *and* $\Phi_S > 0$ *then* $k_D < \beta T_1 - (1-\beta)(V-L) < k_N$*.*
(A3) If $L > V$ *and* $\Phi_S < 0$ *then* $k_D < T_0 + V - L < k_N$*.*
(A4) If $L > V$ *and* $\Phi_S > 0$ *then* $k_D < \beta(T_1 + V - L) < k_N$*.*

Proof. We first consider the strategy of the victim. Suppose the attacker sends a signal and ransom demand R_S. Suppose the victim infers the attacker is type DE. In other words, $\mu = 1$. If the victim is low type and pays the ransom their expected payoff is $-R_S - L$. Their expected payoff if they do not pay is $-T_0 - V$. It follows the low type victim will optimally pay the ransom if and only if $-R_S - L > -T_0 - V$ or equivalently $R_S < T_0 + V - L$. They would, therefore, pay ransom R^*_{S0}. If the victim is high type and pays the ransom their expected payoff is $-R_S - L$. Their expected payoff if they do not pay is $-T_1 - V$. It follows the high type victim will optimally pay the ransom if and only if $R_S < T_1 + V - L$. They would, therefore, pay ransom R^*_{S1}. Given that $T_1 > T_0$ we also have that the high type would pay ransom R^*_{S0}.

Now suppose the attacker does not send a signal and sets ransom demand R_{NS}. Suppose the victim infers the attacker is type NDE. In other words, $\mu = 0$. If the victim is low type and pays the ransom their expected payoff is $-R_{NS} - L$. Their expected payoff if they do not pay is $-V$. It follows the low type victim will optimally pay the ransom if and only if $-R_S - L > -V$ or equivalently $R_S < V - L$. They would, therefore, pay ransom R^*_{NS} if $V > L$ and not pay if $V < L$. The same logic holds if the victim is high type.

We now consider the incentives of the attacker. Suppose the attacker is type DE. Also suppose that on the equilibrium path they signal and set ransom R^*_{S0}. Their expected payoff in equilibrium is $\pi(S, R^*_{S0}) = T_0 + V - L - \epsilon - k_D$. In exploring incentives to deviate from the equilibrium path, we first consider the possibility the attacker signals but sets a different ransom demand $R_S \neq R_S^*$. If $R_S < R_{S0}^*$ then the expected payoff of the attacker is $\pi(S, R_S) = R_S - k_D < \pi(S, R_{S0}^*)$ and so the attacker receives a lower payoff than on the equilibrium path. If $R_{S1}^* > R_S > R_{S0}^*$ (and $\mu = 1$) then the high type victim would pay the ransom but the low type victim would not. The expected payoff of the attacker is, therefore, $\pi(S, R_S) = \beta R_S - k_D \leq \beta R_{S1}^* - k_D$. It follows the attacker prefers the equilibrium path if and only if $\beta(T_1 + V - L - \epsilon) \leq T_0 + V - L - \epsilon$. Rearranging gives the condition on $\Phi_S < 0$. Reversing this argument we can say it is on the equilibrium path for the attacker of type DE to signal and set ransom R_{S1}^* if and only if $\Phi_S > 0$.

We next consider the possibility that an attacker of type DE chooses to not signal. Suppose they set ransom demand R_{NS} (and are inferred to be type NDE). Their expected payoff is at most $\pi(NS, R_{NS}) = R_{NS}^*$. We then have four different cases to consider. (a) Suppose $V > L$ and $R_S^* = R_{S0}^*$. It follows the attacker prefers the equilibrium path if and only if $V - L - \epsilon < T_0 + V - L - \epsilon - k_D$ or, equivalently, $k_D < T_0$. (b) Suppose $V > L$ and $R_S^* = R_{S1}^*$. It follows the attacker prefers the equilibrium path if and only if $V - L - \epsilon < \beta(T_1 + V - L - \epsilon) - k_D$ or, equivalently, $k_D + (1 - \beta)(V - L - \epsilon) < \beta T_1$. (c) Suppose $V < L$ and $R_S^* = R_{S0}^*$. It follows the attacker prefers the equilibrium path if and only if $0 < T_0 + V - L - \epsilon - k_D$ or, equivalently, $k_D < T_0 + V - L$. (d) Suppose $V < L$ and $R_S^* = R_{S1}^*$. It follows the attacker prefers the equilibrium path if and only if $0 < \beta(T_1 + V - L - \epsilon) - k_D$ or, equivalently, $k_D < \beta(T_1 + V - L - \epsilon)$.

Next suppose the attacker is type NDE. Extending the logic of the preceding discussion there is no incentive for the attacker to choose a ransom other than R_{NS}^*. We focus, therefore, on the incentive to signal and choose ransom demand R_S^*. We again have four different cases to consider. (a) Suppose $V > L$ and $R_S^* = R_{S0}^*$. On the equilibrium path the attacker has expected payoff $\pi(NS, R_{NS}^*) = V - L - \epsilon$. It follows the attacker prefers the equilibrium path if and only if $V - L - \epsilon > T_0 + V - L - \epsilon - k_N$ or, equivalently, $k_N > T_0$. (b) Suppose $V > L$ and $R_S^* = R_{S1}^*$. It follows the attacker prefers the equilibrium path if and only if $V - L - \epsilon > \beta(T_1 + V - L - \epsilon) - k_N$ or, equivalently, $k_N + (1 - \beta)(V - L - \epsilon) > \beta T_1$. (c) Suppose $V < L$ and $R_S^* = R_{S0}^*$. It follows the attacker prefers the equilibrium path if and only if $0 > T_0 + V - L - \epsilon - k_N$ or, equivalently, $k_N > T_0 + V - L$. (d) Suppose $V < L$ and $R_S^* = R_{S1}^*$. It follows the attacker prefers the equilibrium path if and only if $0 > \beta(T_1 + V - L - \epsilon) - k_N$ or, equivalently, $k_N > \beta(T_1 + V - L - \epsilon)$.

It remains to check the D1 criterion is satisfied. The only game path we need to consider in any detail is that where the attacker does not signal and sets ransom $R_{NS} \neq R_{NS}^*$. We have assumed the victim will infer the attacker is type NDE. Given that $K_N > k_D$, the attacker has most incentive to not signal when of type NDE. This assumption, therefore, naturally satisfies the D1 criterion.□

In interpretation of Theorem 1 we can see that there exists a separating equilibrium if and only if k_D is sufficiently small and k_N is sufficiently large. In other words, a separating equilibrium exists if it is 'cheap' for the attacker to signal when they have exfiltrated data and 'expensive' for the attacker to signal if they have not exfiltrated data. This would imply, for instance, that if victims have invested in good monitoring systems to identify data exfiltration, they could make it harder for the attacker of type NDE to send a credible signal; then, k_N would increase and we would expect the improved monitoring to result in a separating equilibrium.

3.2 Pooling Equilibrium

We turn our attention now to pooling equilibria. We focus on pooling equilibrium in which the attacker signals. That is, the attacker signals that data is exfiltrated whether they are type NDE or DE. Consequently a signal does not convey any useful information to the victim on whether or not data has been exfiltrated. We identify four types of pooling equilibria, which we will label B1-B4. These are summarised in Table 2. Two ransom demands that prove particularly relevant in this case are: (4) $R^*_{P0} = \alpha T_0 + V - L - \epsilon$, and (5) $R^*_{P1} = \alpha T_1 + V - L - \epsilon$. Informally, R^*_{P0} and R^*_{P1} are the maximum ransom the low and high type, respectively, are willing to pay if they believe the attacker has exfiltrated data with probability α.

As with the separating equilibrium, the optimal ransom demand of the attacker involves a trade-off between setting a high ransom R^*_{P1} that only the high type will pay and a low ransom R^*_{P0} that both the high and low type will pay. This trade-off is captured by the term:

$$\Phi_P = \beta\alpha(T_1 - T_0) - (1-\beta)(\alpha T_0 + V - L - \epsilon). \quad (2)$$

We can now state our second result.

Theorem 2. *There exists a pooling equilibrium in which the attacker signals, satisfying the D1 criterion, if and only if the following conditions hold:*

(B1) If $L < V$ *and* $\Phi_P < 0$ *then* $k_N < \alpha T_0$.
(B2) If $L < V$ *and* $\Phi_P > 0$ *then* $k_N < \beta\alpha T_1 - (1-\beta)(V - L)$.
(B3) If $L > V$ *and* $\Phi_P < 0$ *then* $k_N < \alpha T_0 + V - L$.
(B4) If $L > V$ *and* $\Phi_P > 0$ *then* $k_N < \beta(\alpha T_1 + V - L)$.

Proof. Consider the strategy of the victim. Suppose the attacker sends a signal and ransom demand R_S. Suppose the victim infers the attacker is type DE with probability $\mu = \alpha$. If the victim is low type and pays the ransom their expected payoff is $-R_S - L$. Their expected payoff if they do not pay is $-\alpha T_0 - V$. It follows the low type victim will optimally pay the ransom if and only if $-R_S - L > -\alpha T_0 - V$ or equivalently $R_S < \alpha T_0 + V - L$. They would, therefore, pay ransom R^*_{P0}. If the victim is high type and pays the ransom their expected payoff is $-R_S - L$. Their expected payoff if they do not pay is $-\alpha T_1 - V$. It follows the high type victim will optimally pay the ransom if and only if $R_S < \alpha T_1 + V - L$. They would, therefore, pay ransom R^*_{P1}. Given that $T_1 > T_0$ we also have that the high type would pay ransom R^*_{P0}.

Now suppose the attacker does not send a signal and sets ransom demand R_{NS}. Suppose the victim infers the attacker is type NDE. In other words, $\mu = 0$. If the victim is low type and pays the ransom their expected payoff is $-R_{NS} - L$. Their expected payoff if they do not pay is $-V$. It follows the low type victim will optimally pay the ransom if and only if $-R_S - L > -V$ or equivalently $R_S < V - L$. They would, therefore, pay ransom R^*_{NS} if $V > L$ and not pay if $V < L$. The same logic holds if the victim is high type.

Next consider the incentives of the attacker. Suppose the attacker is type DE. Also suppose that on the equilibrium path they signal and set ransom R^*_{P0}. Their expected payoff in equilibrium is $\pi(S, R^*_{P0}) = \alpha T_0 + V - L - \epsilon - k_D$. Suppose the attacker signals but sets a different ransom demand $R_S \neq R^*_S$. If $R_S < R^*_{S0}$ then the expected payoff of the attacker is $\pi(S, R_S) = R_S - k_D < \pi(S, R^*_{P0})$ and so the attacker receives a lower payoff than on the equilibrium path. If $R^*_{P1} > R_S > R^*_{P0}$ (and $\mu = \beta$) then the high type victim would pay the ransom but the low type victim would not. The expected payoff of the attacker is, therefore, $\pi(S, R_S) = \beta R_S - k_D \leq \beta R^*_{P1} - k_D$. It follows the attacker prefers the equilibrium path if and only if $\beta(\alpha T_1 + V - L - \epsilon) \leq \alpha T_0 + V - L - \epsilon$. Rearranging gives $\Phi_P < 0$. Reversing this argument we can say it is on the equilibrium path for the attacker of type DE to signal and set ransom R^*_{P1} if and only if $\Phi_P > 0$.

Now consider the possibility that an attacker of type NDE chooses to not signal. Suppose they set ransom demand R_{NS} (and are inferred to be type NDE). Their expected payoff is at most $\pi(NS, R_{NS}) = R^*_{NS}$. We then have four different cases to consider. (a) Suppose $V > L$ and $R^*_S = R^*_{P0}$. It follows the attacker prefers the equilibrium path if and only if $V - L - \epsilon < \alpha T_0 + V - L - \epsilon - k_N$ or, equivalently, $k_N < \alpha T_0$. (b) Suppose $V > L$ and $R^*_S = R^*_{P1}$. It follows the attacker prefers the equilibrium path if and only if $V - L - \epsilon < \beta(\alpha T_1 + V - L - \epsilon) - k_N$ or, equivalently, $k_N + (1 - \beta)(V - L - \epsilon) < \beta \alpha T_1$. (c) Suppose $V < L$ and $R^*_S = R^*_{P0}$. It follows the attacker prefers the equilibrium path if and only if $0 < \alpha T_0 + V - L - \epsilon - k_N$ or, equivalently, $k_N < \alpha T_0 + V - L$. (d) Suppose $V < L$ and $R^*_S = R^*_{P1}$. It follows the attacker prefers the equilibrium path if and only if $0 < \beta(\alpha T_1 + V - L - \epsilon) - k_N$ or, equivalently, $k_N < \beta(\alpha T_1 + V - L - \epsilon)$. One can show, using $k_D < k_N$, that the analogous conditions for a type DE to prefer signalling to not signalling are less binding.

It remains to check the D1 criterion is satisfied. The only game path we need to consider in any detail is that where the attacker does not signal and sets ransom $R_{NS} \neq R^*_{NS}$. We have assumed the victim will infer the attacker is type NDE. Given that $K_N > k_D$, the attacker has most incentive to not signal when of type NDE. This assumption, therefore, naturally satisfies the D1 criterion. □

In interpretation of Theorem 2 there exists a pooling equilibrium if and only if k_N is sufficiently small. In other words, there exists a pooling equilibrium if and only if it is cheap for the attacker to signal even if data has not been exfiltrated. In practical terms this would suggest the victim does not have any monitoring technology to identify or evaluate a data breach. It would also suggest the criminals could easily extract some information, e.g. file tree or sample file, that would allow them to signal data exfiltration even though data was not exfiltrated.

Depending on the parameters of the game there may exist a separating equilibrium, a pooling equilibrium, or neither. To illustrate, consider the parameters $L = 0, V = 5, \alpha = 0.9, \beta = 0.5, T_0 = 1$ and $T_1 = 5$. Then $\Phi_S < 0$ and so there exists a separating equilibrium if and only if $k_D < 1 < k_N$. Also $\Phi_P < 0$ and so there exists a pooling equilibrium if $k_N < 0.9$. Thus, for $k_N < 0.9$ there is a pooling equilibrium, for $0.9 < k_N < 1$ there is neither a separating nor pooling equilibrium, and for $1 < k_N$ there is a separating equilibrium. The relative size of the cost for the attacker to signal data exiftration when they have not exfiltrated data is, thus, crucial to determining the equilibrium outcome.

3.3 The Value of Private Information

A key objective of our work is to analyse the payoff consequences, for both victim and attacker, of private information on the side of the victim. In Table 3 we detail the expected payoff of the attacker and victim in equilibria A1-A4 and B1-B4. These are ex-ante expected payoffs before own type is known. For instance, in equilibrium A1 there is probability α the attacker is type DE and obtains payoff $R^*_{S0} - k_D$ and probability $1 - \alpha$ the attacker is type NDE and obtains payoff R^*_{NS}. The expected payoff is, therefore, $\alpha(R^*_{S0} - k_D) + (1 - \alpha)R^*_{NS}$ Given that ϵ can be arbitrarily small we have omitted it from calculations of expected payoff.

Table 3: Expected payoff of attacker and victim in equilibrium.

Equilibrium	attacker	Victim
A1	$\alpha T_0 + V - L - \alpha k_D$	$-\alpha T_0 - V$
A2	$\alpha(\beta(T_1 + V - L) - k_D) + (1 - \alpha)(V - L)$	$-\alpha(\beta T_1 + (1 - \beta)T_0) - V$
A3	$\alpha(T_0 + V - L - k_D)$	$-\alpha T_0 - V$
A4	$\alpha(\beta(T_1 + V - L) - k_D)$	$-\alpha(\beta T_1 + (1 - \beta)T_0) - V$
B1 & B3	$\alpha T_0 + V - L - \alpha k_D - (1 - \alpha)k_N$	$-\alpha T_0 - V$
B2 & B4	$\beta(\alpha T_1 + V - L) - \alpha k_D - (1 - \alpha)k_N$	$-\beta\alpha T_1 - (1 - \beta)\alpha T_0 - V$

To analyse the consequences of private information we need to consider an alternative game in which the attacker knows the type of the victim and so knows if the reputational damage that would result from data publication is T_0 or T_1. We can apply Theorems 1 and 2 to distinguish the conditions under which there exist separating and pooling equilibirum in this revised game. Specifically, by setting $\beta = 0$ or 1 we derive the following corollaries.

Corollary 1. *If the victim is known to be type $i = \{0, 1\}$ there exists a separating equilibrium satisfying the D1 criterion if and only if the following conditions hold:*

A1A2. If $L < V$, then $k_D < T_i < k_N$.
A3A4. If $L > V$, then $k_D < T_i + V - L < k_N$.

Proof. Suppose $\beta = 0$. Then $\Phi_S < 0$. Applying Theorem 1 we obtain conditions: (A1) $L < V$ and $k_D < T_0 < k_N$, and (A3) $L > V$ and $k_D < T_0 + V - L < k_N$. Suppose $\beta = 1$. Then $\Phi_S > 0$. Applying Theorem 1 we obtain conditions: (A2) $L < V$ and $k_D < T_1 < k_N$, and (A4) $L > V$ and $k_D < T_1 + V - L < k_N$. □

Corollary 2. *If the victim is known to be type $i = \{0, 1\}$ there exists a pooling equilibrium with a signal satisfying the D1 criterion if and only if the following conditions hold:*

B1B2. If $L < V$ then $k_N < \alpha T_i$.
B3B4. If $L > V$ then $k_N < \alpha T_i + V - L$.

Proof. Suppose $\beta = 0$. Then $\Phi_P < 0$. Applying Theorem 2 we obtain conditions: (B1) $L < V$ and $k_N < \alpha T_0$, and (B3) $L > V$ and $k_N < \alpha T_0 + V - L$. Suppose $\beta = 1$. Then $\Phi_P > 0$. Applying Theorem 2 we obtain conditions: (B2) $L < V$ and $k_N < \alpha T_1$, and (B4) $L > V$ and $k_N < \alpha T_1 + V - L$. □

With these two corollaries we can derive the expected payoff of the attacker and victim in a game where the victim's type is known. The lower half of Table 4 details the payoffs from equilibria of the game in which the victims type is known. For instance, the expected payoff of the attacker under equilibrium A3A4 if the victim is type 0 is $\alpha(T_0+V-L-k_D)$ and the expected payoff of the attacker under equilibrium A3A4 if the victim is type 1 is $\alpha(T_1+V-L-k_D)$. Some care is needed in deriving ex-ante expected payoffs because the existence of equilibrium A3A4 for the low type does not guarantee existence of equilibrium A3A4 for the high type, and vice-versa. Even so, by calculating which equilibrium emerges for each type we can determine an ex-ante expected payoff. For instance, if equilibrium A3A4 does exist for both the low type and high type then the attackers ex-ante expected payoff (before victim type is known) is $\alpha(\beta T_1+(1-\beta)T_0+V-L-k_D)$.

Table 4: Expected payoff of attacker and victim in equilibrium when type is known.

Equilibrium	attacker	Victim
A1A2 ($i = \{0, 1\}$)	$\alpha T_i + V - L - \alpha k_D$	$-\alpha T_i - V$
A3A4 ($i = \{0, 1\}$)	$\alpha(T_i + V - L - k_D)$	$-\alpha T_i - V$
B1-B4 ($i = \{0, 1\}$)	$\alpha T_i + V - L - \alpha k_D - (1 - \alpha)k_N$	$-\alpha T_i - V$

We are now in a position to quantify the payoff consequences of private information for the victim. For any set of parameters $L, V, T_0, T_1, k_D, k_N, \alpha$ and β we can determine which, if any equilibrium will hold in a game with incomplete information on victim's type, and the games where victim's type is known to be high or low. We can then calculate expected payoffs of the attacker and victim with and without incomplete information on victim's type. We provide two examples.

In Figure 2 we plot expected payoffs as a function of β when $L = 1, V = 3, \alpha = 0.5, T_0 = 2, T_1 = 4, k_D = 0.1$ and $k_N = 6$. This is a case with a separating equilibrium. You can see that the payoff of the attacker is substantially lower when the type of the victim is not known. The loss reaches a maximum at the point of transition between equilibria A1 and A2 given by $T_0 = \beta T_1 - (1-\beta)(V - L)$ or equivalently

$$\beta = \frac{T_0 + V - L}{T_1 + V - L}. \tag{3}$$

For the parameters in our example this gives $\beta = 2/3$. If the type of the victim is unknown the expected payoff of the attacker is 2.95. If the type of the victim is known the ex-ante expected payoff of the attacker is 3.62. So, the attacker's payoff is 18.43% lower if it does not know the type of the victim.

We can see the victim's payoff is higher if the attacker does not know their type and $\beta < 2/3$. The intuition being that the attacker sets the ransom as if the victim is low type (equilibrium A1) and, thus, the high type is not exploited as much as they would have been if type was known. If $\beta > 2/3$ we see that the payoff of the victim is the same whether or not the attacker knows their type. In this case the attacker sets the ransom as if the victim is high type (equilibrium A2). This means the high type is maximally exploited by the attacker, while the low type does not pay the ransom and, therefore, suffers recovery and reputational losses. The net effect for the victim is the same as if the attacker knew their type and they were maximally exploited. We remind that the attacker's payoff is lower if the victim's type is not known. This is because they also lose when the ransom is set at a level the low type will not pay.

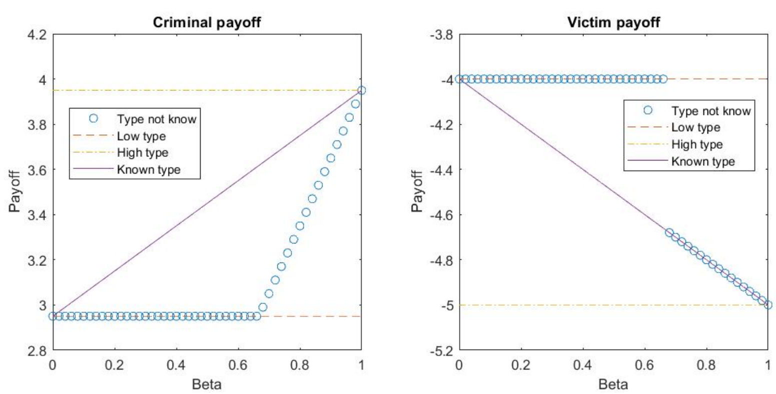


Fig. 2: Expected payoff of the attacker and victim when $L = 1, V = 3, \alpha = 0.5, T_0 = 2, T_1 = 4, k_N = 6, k_D = 0.1$. An example of a separating equilibrium.

In Figure 3 we plot the corresponding payoffs when, everything else the same, $k_N = 0.9$. This is a case with a pooling equilibrium. Again, we see that the attacker loses payoff from not knowing the type of the victim. This loss is maximal at the transition from equilibrium B1 to B2, given by $\alpha T_0 = \beta\alpha T_1 - (1-\beta)(V-L)$ or equivalently

$$\beta = \frac{\alpha T_0 + V - L}{\alpha T_1 + V - L}. \tag{4}$$

For the parameters in our example this gives $\beta = 3/4$. If the type of the victim is unknown the expected payoff of the attacker is 2.5. If the type of the victim is known the ex-ante expected payoff of the attacker is 3.25. So, the attacker's payoff is 23.08% lower because it does not know the type of the victim.

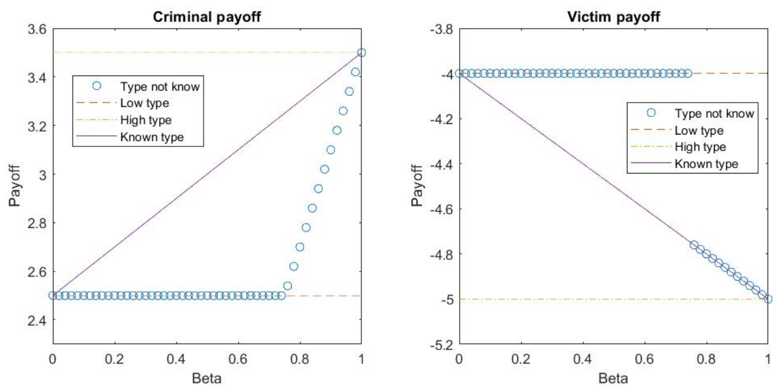


Fig. 3: Expected payoff of the attacker and victim when $L = 1, V = 3, \alpha = 0.5, T_0 = 2, T_1 = 4, k_N = 0.9, k_D = 0.1$. An example of a separating equilibrium.

The relative trade-offs for the victim are similar in the pooling example as the separating example. In particular, if the attacker sets the ransom for a victim of low type (equilibrium B1) then the victim gains from their type being private if they are high type. If, however, the attacker sets the ransom for a victim of high type (equilibrium B2) then the victim does not gain from their type being unknown. In summary, the attacker loses payoff from not knowing the victim's type. The victim gains from their type being unknown in the case of equilibrium A1, B1 and also A3 and B3. The victim does not gain from the type being unknown in the case of equilibrium A2, A4, B2 and B4.

It is interesting to compare payoffs when $k_N = 0.9$ with those when $k_N = 6$ (for, say, $\beta = 2/3$). You can see that the attackers expected payoff is higher when $k_N = 6$. This may seem counter-intuitive given that a high k_N means a higher cost from signalling. Note, however, that a high k_N results in a separating equilibrium that allows the type DE attacker to extract a high ransom because their signal of data exfiltration is credible. Specifically, when $k_N = 6$ the type DE sets ransom $R^*_{S0} = T_0 + V - L = 4$, while a type NDE sets ransom $R^*_{NS} = V - L = 2$. The expected payoff of the attacker is, therefore, $\alpha(R^*_{S0} - k_D) + (1-\alpha)R^*_{NS} = 3.9\alpha + 2(1-\alpha) = 2.95$.

By contrast, when $k_N = 0.9$ we obtain a pooling equilibrium in which the attacker's signal of data exfiltration is not sufficiently credible. This lowers the ransom the attacker can demand to $R^*_{P0} = \alpha T_0 + V - L = 3$. Consequently the type DE gets a lower payoff with the lower k_N (2.9 compared to 3.9). The type NDE, by contrast, has a higher payoff (2.1 compared to 2) because they are also able to demand ransom R^*_{P0}, although they incur cost k_N. The expected payoff of the attacker is $R^*_{P0} - 0.1\alpha - 0.9(1-\alpha) = 2.5$. Overall, therefore, the attacker has a lower expected payoff when $k_N = 0.9$ compared to $k_N = 6$ (2.5 compared to 2.95). This trade-off is apparent from the payoffs in Table 3, comparing A1 and B1.

You can also see in Table 3 that the payoff of the victim is not impacted by k_N. This is because the criminal is able to extract the same surplus from the victim in equilibria A1, A3, B1 and B3. Generally, speaking, as would be expected, the loss to the victim is reduced by lowering T_0, T_1, V and β. The victim's payoff is also reduced by lowering α. Thus, reduced the losses from data exfiltration as well as reducing the probability of data exfiltration reduce the losses to the victim.

4 Conclusion

This paper provides a game-theoretic analysis of the double-sided information asymmetry in double extortion ransomware attacks. We recognised that victims are typically unable to verify if data was exfiltrated or not, while attackers typically do not know the value of any data exfiltrated. We modeled the ransomware attack as a signaling game, where attackers could signal if data is exfiltrated and victims pay based on the ransom, signal and the value of information. Our key contribution is that, depending on the parameters of the game, private information of the victim (about the value of exfiltrated data) significantly lowers the profitability of the attack for the criminal. It is, therefore, in the interests of potential victims, businesses, organisations, and/or individuals, to retain and amplify the extent of their private information.

According to our model, the most effective way to disrupt the attackers profitability is to: lower the probability of 'successful' data exfiltration, lower the probability the victim has files of high reputational cost, and lower the recovery cost from an attack. This would involve a mix of prevention (to lower the probability of data exfiltration and loss of sensitive data) as well as improved recovery options, such as back-ups. Crucially there is an externality effect: the more victims safeguard their sensitive data the more that benefits other businesses, including those with vulnerable sensitive data. This is because it would revise downwards the beliefs of attackers about the ransoms they can reasonably expect victims to pay. This externality effect should be acknowledged by policy makers. In particular, it means businesses will under-invest in cyber security prevention and recovery compared to the social optimum. This can justify government support for cyber security investment.

References

1. Akerlof, G. A.: The market for “lemons”: Quality uncertainty and the market mechanism. The quarterly journal of economics, **84**(3), 488-500 (1970)
2. Baksi, R. P., & Upadhyaya, S. J.: Game Theoretic Analysis of Ransomware: A Preliminary Study. ICISSP, 242-251 (2022)
3. Banks, J. S., & Sobel, J.: Equilibrium selection in signaling games. Econometrica: Journal of the Econometric Society, 647-661 (1987)
4. Cartwright, A., Cartwright, E., MacColl, J., Mott, G., Turner, S., Sullivan, J., & Nurse, J. R.: How cyber insurance influences the ransomware payment decision: theory and evidence. The Geneva Papers on Risk and Insurance-Issues and Practice, **48**(2), 300-331 (2023)
5. Cartwright, E., Hernandez Castro, J., & Cartwright, A.: To pay or not: game theoretic models of ransomware. Journal of Cybersecurity, **5**(1), (2019)
6. Cong, L. W., Harvey, C. R., Rabetti, D., & Wu, Z. Y.: An anatomy of crypto-enabled cybercrimes. National Bureau of Economic Research (2023)
7. Connolly, L. Y., & Wall, D. S.:The rise of crypto-ransomware in a changing cybercrime landscape: Taxonomising countermeasures. Computers & Security, 87, 101568 (2019)
8. Connolly, Y. L., Wall, D. S., Lang, M., & Oddson, B.: An empirical study of ransomware attacks on organizations: an assessment of severity and salient factors affecting vulnerability. Journal of Cybersecurity, **6**(1) (2020)
9. Harsanyi, J. C.; Games with incomplete information played by “Bayesian” players, I–III Part I. The basic model. Management science, **14**(3), 159-182 (1967)
10. Humayun, M., Jhanjhi, N. Z., Alsayat, A., & Ponnusamy, V.: Internet of things and ransomware: Evolution, mitigation and prevention. Egyptian Informatics Journal, **22**(1), 105-117 (2021)
11. Fudenberg, D., & Tirole, J.: Game theory. MIT press (1991)
12. Galinkin, E.: Winning the Ransomware Lottery: A Game-Theoretic Approach to Preventing Ransomware Attacks. In Decision and Game Theory for Security: 12th International Conference, GameSec 2021, Proceedings 12, pp. 195-207, Springer International Publishing (2021)
13. Kerns, Q., Payne, B., & Abegaz, T.: Double-Extortion Ransomware: A Technical Analysis of Maze Ransomware. In Proceedings of the Future Technologies Conference (FTC) 2021, **3** 82-94, Springer International Publishing (2022)
14. Kreps, D. M., & Sobel, J.: Signalling. Handbook of game theory with economic applications, 2, 849-867 (1994)
15. Laszka, A., Farhang, S., & Grossklags, J.: On the economics of ransomware. In Decision and Game Theory for Security: 8th International Conference, GameSec 2017, Proceedings, pp. 397-417. Springer International Publishing (2017)
16. Li, Z., & Liao, Q.: Preventive portfolio against data-selling ransomware—A game theory of encryption and deception. Computers & Security, 116, 102644 (2022)
17. Li, Z., & Liao, Q.: Game theory of data-selling ransomware. Journal of Cyber Security and Mobility, 65-96 (2021)
18. Maschler, M., Zamir, S., & Solan, E.: Game theory. Cambridge University Press (2020)
19. Meurs, T., Junger, M., Tews, E., Abhishta, A.: Ransomware: How attacker’s effort, victim characteristics and context influence ransom requested, payment and financial loss. In Symposium on Electronic Crime Research, eCrime (2022)

20. Mott, G., Turner, S., Nurse, J. R., MacColl, J., Sullivan, J., Cartwright, A., & Cartwright, E.: Between a rock and a hard (ening) place: Cyber insurance in the ransomware era. Computers & Security, 128, 103162 (2023)
21. Oosthoek, K., Cable, J., & Smaragdakis, G.: A Tale of Two Markets: Investigating the Ransomware Payments Economy. arXiv preprint arXiv:2205.05028 (2022)
22. Osborne, M. J.: An introduction to game theory 3rd edn. New York: Oxford university press (2004)
23. Oz, H., Aris, A., Levi, A., & Uluagac, A. S.: A survey on ransomware: Evolution, taxonomy, and defense solutions. ACM Computing Surveys (CSUR), 54(11s), 1-37 (2022)
24. Ryan, P., Fokker, J., Healy, S., & Amann, A.: Dynamics of targeted ransomware negotiation. IEEE Access, **10**, 32836-32844 (2022)
25. Sabir, B., Ullah, F., Babar, M. A., & Gaire, R. (2021). Machine learning for detecting data exfiltration: a review. ACM Computing Surveys (CSUR), **54**(3), 1-47 (2021)
26. Ullah, F., Edwards, M., Ramdhany, R., Chitchyan, R., Babar, M. A., & Rashid, A.: Data exfiltration: A review of external attack vectors and countermeasures. Journal of Network and Computer Applications, 101, 18-54 (2018)
27. Vakilinia, I., Khalili, M. M., & Li, M.: A Mechanism Design Approach to Solve Ransomware Dilemmas. In Decision and Game Theory for Security: 12th International Conference, GameSec 2021, Virtual Event, October 25–27, 2021, Proceedings 12 (pp. 181-194). Springer International Publishing (2021)
28. Yin, T., Sarabi, A., & Liu, M.: Deterrence, Backup, or Insurance: Game-Theoretic Modeling of Ransomware. Games, **14**(2), 20 (2023)
29. Zhao, Y., Ge, Y., & Zhu, Q.: Combating Ransomware in Internet of Things: A Games-in-Games Approach for Cross-Layer Cyber Defense and Security Investment. In Decision and Game Theory for Security: 12th International Conference, GameSec 2021, Proceedings, pp. 208-228. Cham: Springer International Publishing (2021)