To pay or not: game theoretic models of ransomware

Cartwright, E, Hernandez-Castro, J & Cartwright, A

DOI 10.1093/cybsec/tyz009
ISSN 2057-2085
ESSN 2057-2093

Publisher: OUP

JOURNAL OF CYBERSECURITY

Journal of Cybersecurity, 2019, 1–12
doi: 10.1093/cybsec/tyz009
Research paper

Research paper

To pay or not: game theoretic models of ransomware

Edward Cartwright [1,*], Julio Hernandez Castro[2] and Anna Cartwright[3]

[1]Department of Strategic Management and Marketing, De Montfort University, Leicester, UK; [2]School of Computing, University of Kent, Canterbury, Kent, UK; [3]School of Economics, Finance and Accounting, Coventry University, Coventry, UK

*Correspondence address. Department of Strategic Management and Marketing, De Montfort University, Leicester, LE1 9BH, UK. Tel. +44(0)116 366 4385. E-mail: edward.cartwright@dmu.ac.uk

Received 29 September 2018; revised 20 February 2019; accepted 14 July 2019

Abstract

Ransomware is a type of malware that encrypts files and demands a ransom from victims. It can be viewed as a form of kidnapping in which the criminal takes control of the victim's files with the objective of financial gain. In this article, we review and develop the game theoretic literature on kidnapping in order to gain insight on ransomware. The prior literature on kidnapping has largely focused on political or terrorist hostage taking. We demonstrate, however, that key models within the literature can be adapted to give critical new insight on ransomware. We primarily focus on two models. The first gives insight on the optimal ransom that criminals should charge. The second gives insight on the role of deterrence through preventative measures. A key insight from both models will be the importance of spillover effects across victims. We will argue that such spillovers point to the need for some level of outside intervention, by governments or otherwise, to tackle ransomware.

Key words: ransomware; game theory; kidnapping; hostage; deterrence

Introduction

Ransomware denotes the branch of malware that, after infecting a computer, asks for a ransom. Typically, the files on the computer are encrypted and the criminals demand a ransom for the private key to decrypt the files [1, 2]. Victims are given a set time, typically 72 h, to pay the ransom, which can vary from \$100 to \$1000 for individuals, and a lot higher for firms and organizations [3]. While the know-how to develop ransomware has existed in academia for some time, e.g. Young and Yung [4], it is only recently that cryptographically sound ransomware has found its way into the wild. While there exist many variants of ransomware that allow for reverse-engineering [5], there are now many variants in which a victim, who wants to recover their files, has no choice but to pay the ransom.

The point of departure for this article is the recognition that ransomware is a form of kidnapping in which a criminal takes control of a victim's computer files in the hope of financial gain. The kidnapping aspect of ransomware is already acknowledged at a practical level with companies using insurance policies designed to cover against kidnapping of staff to mitigate losses from ransomware [6].[1] In this article, we use insights from the game theoretical literature to better understand the incentives behind ransomware. Game theory provides a natural tool with which to study kidnapping, particularly when the motives of the criminal are financial, and several models of kidnapping have been developed within the literature (e.g. [7–9]). In this article, we adapt and apply to the ransomware context two key models of kidnapping: one developed by Selten [10] and the other by Lapan and Sandler [11].

[1] Internal procedures can also require law enforcement agencies to call on trained hostage-taking officers to execute ransom payments (to track the criminal money flow).

Before we continue, it seems pertinent to clarify that only two studies, of which we are aware, have explicitly applied game theoretic models of kidnapping to ransomware [12, 13]. This is, therefore, an area ripe for further study. Most of the game theoretic literature on kidnapping has focused on terrorist hostage taking in conflict zones [11, 14, 15]. War may seem a world removed from ransomware but the beauty of the game theoretic approach is that, by focusing on the salient strategic incentives, it is applicable across different domains.[2] Even so, there are specific aspects of ransomware that point to the issues that could be analysed in more detail than we find in the current literature. We shall highlight these issues as we proceed. Let us also remark that ransomware provides a very natural application of game theoretic reasoning in that it primarily involves money and computer files; this raises less moral (and modelling) issues than the kidnap (and potential murder) of a hostage [17, 18].

As already previewed, we shall develop and extend two key models of kidnapping. The first model, due to Selten [10], primarily focuses on the optimal ransom that criminals should charge. The second model, due to Lapan and Sandler [11], primarily focuses on whether potential victims should take action to deter hostage taking. While neither model was designed to study ransomware, we argue that there are clear lessons that can be drawn from both these models. For instance, they feed into the general debate on the willingness of ransomware victims to pay to recover files and the willingness of people to avoid attack through antivirus protection, regular backups or similar. Some clear and actionable policy recommendations follow from our analysis. Let us highlight that we focus on these two models, and analyse them separately, because they study complementary aspects of the ransomware problem. In the concluding discussion we discuss how the models could be joined together to create an amalgamated model covering deterrence and bargaining. Our results are robust to this generalized model.

A potential criticism of a game theoretic approach is that it assumes rational decision-making by both the criminal and victim. In addressing this point, one thing to observe is that rationality does not preclude a role for emotions, such as anger or panic, and so rational does not have to mean 'cold and calculated'. This is illustrated by an assumption of irrational aggression that is crucial to the model of Selten [10]. A more fundamental thing to recognize is the forces that direct towards rationality. In particular, we believe it is naïve to assume that the strategy of ransomware criminals is not going to evolve towards an optimum. The closer to the optimum they get, the more money the criminals will make and so, by accident or not, they are likely to stumble towards the optimum. This does not mean the criminals are currently optimizing [19]. But it does mean that a game theoretic approach gives us insight on where things might go, and that, in turn, gives chance for law enforcement and others to be one-step ahead.

Our article adds to a small but growing literature on the economic aspects of ransomware. In earlier work [19], we look at the economic theory behind optimal ransom pricing. Insights from this work are applied below (see section 'A simple game model of kidnapping' discussing kidnapping game). Laszka *et al.* [12] model the ransomware eco-system as a multi-stage, multi-defender game. The particular focus of their analysis is on the interaction between the decision to back up and pay a ransom. Their model has a close overlap with our second model (see section 'A simple game model of deterrence' discussing deterrence game) and so we discuss their results in detail at that point. Caporusso *et al.* [13] show that game theory can be used to model bargaining between criminal and victim. Huang *et al.* [20] explore how ransomware can fit within a cybercrime business model, while August *et al.* [21] explore the potential implications of ransomware for software vendors. There are also a number of papers that have looked to quantify and document the financial gains from ransomware and the behaviour of victims and criminals. This literature is crucial for our purposes as it allows us to calibrate model parameters with real-world observation. We will discuss this literature more in the next section.

Set against a background of increasing interest in the economics and game theory of ransomware, our article makes two basic contributions. First, it provides an accessible summary of existing results in the game theoretic literature on kidnapping; this, hopefully, avoids researchers 'reinventing the wheel'. Secondly, we extend the analysis of the Selten [10] and Lapan and Sandler [11] models in nontrivial ways to take account of specific aspects in ransomware. We proceed as follows. First, we provide a brief overview of ransomware with the objective of feeding the subsequent analysis (rather than providing a comprehensive review). Then we apply in turn the key models of Selten [10] and Lapan and Sandler [11] to ransomware. Subsequently we review the remaining game theoretic literature on kidnapping and discuss the complementary findings of different models before a concluding discussion. While none of the literature has explicitly studied ransomware, we in this article frame the analysis throughout in terms of a ransomware attack.

Background on ransomware

In this section, we provide a brief overview of ransomware. This overview is not intended to be comprehensive but merely to highlight salient points for the analysis to follow. We can begin by noting that CryptoLocker was one of the first, if not the first, to implement a scheme close to the Young and Yung [4] protocol in a technically sound way from its conception [22]. Its 'good' implementation unfortunately forced victims wanting to recover their files to pay the ransom. That was the only available alternative. Throughout this article we will focus on cryptographically sound ransomware, such as CryptoLocker, where the files are recoverable if and only if the criminals return the relevant keys.[3]

The precise proportion of victims who paid ransoms to CryptoLocker is unknown with estimates ranging from 2% to 40% [23]. It is, however, clear that enough people paid ransom to generate a large amount of money. Conservative estimates on the amount of ransom received by the criminals range from $300 000 to over $1 000 000 (with fluctuations in bitcoin making valuation volatile) [24, 25]. We also know that a single address connected with Cyrptolocker received a total of 346 102 BTC at the time of its last transaction in February 2014. This was a significant proportion of

[2] This also means that in a game theoretic sense kidnapping can be used more widely than in common usage. For example, Schelling [16] equates nuclear power with the ability to take hostages. Basically, if, say the USA has the ability to destroy Russia, then it is as if the USA takes Russian citizens as hostages.

[3] This is not to say that it is the only type of ransomware. There is 'fake' ransomware that simply destroys the files, and non-sound ransomware that allows recovery without paying a ransom.

the total number of bitcoins in circulation (approx. 12 million) and would have had a valuation in excess of $200 million.

Operation Tovar in 2014, led by the US Department of Justice and the FBI, led to the Gameover/Zeus botnet being closed down. This was one of the main distribution paths for CryptoLocker and so effectively meant the end for this particular form of ransomware.[4] This, though, was definitely not the end of the story. CryptoLocker demonstrated the huge potential to extract large amounts of money through a cryptovirus and other large-scale attacks have followed, and new families such as CryptoWall, TorLocker, Fusob, Cerber, TeslaCrypt, etc. have emerged [2, 3, 26]. Conservative estimates of the total amount paid in ransoms since 2013 are around $13– $26 million [27, 28]. The economic and social costs of ransomware clearly extend well beyond the payment of ransoms.

Modern ransomware strands are fast evolving, not only in terms of technical capabilities but also in economic sophistication. For instance, ransomware-as-a-service allows just about anyone to commit the crime irrespective of technical know-how [20]. Also, modern strands come with a 'customer service' department to advise 'clients' and facilitate payment. We have also seen large-scale targeted attacks on large organizations, including universities and health trusts. Indeed, the trend appears to be towards more targeted attacks on large organizations [29]. Unfortunately, there is less evidence of individuals and organizations taking the necessary measures (particularly, regular backups) to mitigate and possibly deter the damage from attack. This means that ransomware is likely to remain a serious threat for many years to come.

Ransomware is rare (maybe unique) in being a cybercrime that positively benefits from publicity and greater knowledge. The more individuals and organizations recognize that ransomware is a genuine extortion scenario in which access to files can only be regained through paying the ransom, the more willing they might be to engage with the criminals. Indeed, the FBI was somewhat inadvertently dragged into such complexities when in 2015 an agent was quoted as saying that 'the ransomware is that good ... To be honest, we often advise people to just pay the ransom' [30]. This leads onto two key issues that will be important in our models, namely, whether the criminals do return files and the proportion of victims that pay.

Data is understandably sketchy given the nature of ransomware. Anecdotal evidence shows, however, that criminals do often honour ransom payments and return the key to decrypt the files. The widely publicized case of the University of Calgary paying $20 000 to get back their files is one example of a ransom payment that 'worked'. More generally, some ransomware strands such as CryptoWall developed a good reputation for returning the files [31]. This means victims have a reasonable chance of recovering their files, leaving them with a basic dilemma of whether to pay or not. The evidence suggests that many victims do indeed pay, particularly businesses [32]. This suggests that ransomware can provide a sustainable business model for criminals.[5]

A simple game model of kidnapping

In this section, we apply and adapt the model of kidnapping due to Selten [10]. We shall refer to the game studied as the 'kidnapping game'. The kidnapping game was originally developed to model a situation in which an individual is taken hostage so as to extract a ransom from family members. Here, however, we will frame the discussion in terms of ransomware.[6] We will see that the kidnapping game is particularly informative in terms of the optimal ransom demand. It also highlights the need for criminals to have a credible way of threatening victims. The game involves two players, a criminal and victim. It has six stages, which can be explained as follows.

Stage 1: The criminal chooses whether or not to infect the victim's computer. If the files are not infected then the game ends and both players get payoff 0.[7]

Stage 2: If the criminal infects the victim's computer then the criminal chooses a ransom demand $D > 0$. This demand is communicated to the victim.

Stage 3: Having seen the demand D, the victim chooses a counter-offer $C \in [0, D]$.[8] Note that it is far from clear whether it is in the criminal's interests to let the victim make a counter-offer. It is simply assumed for now that this possibility exists. We return to this issue later. We can note, though, that almost all (genuine) ransomware strains allow for some form of communication with the criminals in order to make a counter-offer [26, 33] and so bargaining is a key aspect of the game [13]. Whether or not the criminals are willing to lower the price varies by type of ransomware.

Stage 4: With probability $\alpha = a(1 - C/D)$, where $a \in (0, 1)$ is a constant, the victim's files are destroyed without any exchange of ransom. Note that if $C = D$, then the files are not destroyed. If $C < D$ then the files may be destroyed and the probability of destruction increases in the gap between demand and counter-offer. In a game theoretic sense, this destruction is modelled as an act of nature (out of the criminal's control) and so we shall call it *random destruction*. Selten [10] equates this with 'irrational aggression' on the part of the criminal. More generally, it can be equated with a risk of aggressive behaviour because the counter-offer is below that demanded. As a reviewer of an earlier version of the article pointed out that such random destruction could be programmed into the malware itself by the criminal (although we have no example of that ever being done). Let us, however, highlight that the crucial thing here is the *victim's perception* of the probability the files will be destroyed; the value of α captures this perception. Destruction of the files results in a payoff of $-Y \leq 0$ for the criminal and $-W < 0$ for the victim.[9] In interpretation Y can be thought of as including the costs of attacking the victim (given that no attack has payoff 0).

Stage 5: If the files were not destroyed in Stage 4 then the criminal chooses between releasing the files and receiving $C - G$, for some $G \geq 0$, or destroying the files and receiving $-Y$. The value of G captures the cost of having to properly engage with the victim in order to decrypt files. It may, for instance, involve customer support [29]. In interpretation we can think of the criminal as having a

[4] During Operation Tovar, a database was located, containing approximately 500 000 individuals, and this allowed the set up of a website to facilitate victims recovering their files (https://www.decryptcryptolocker.com). It is important to note that this was only possible due to the recovery of the criminals' database, and not to any security weakness in the implementation of the cryptovirus itself.

[5] A typical ransomware strand may only be able to survive for, say, six months before the law enforcement agencies start to close in. But the criminals can evolve and continually develop new strains.

[6] Some aspects of the model are arguably better suited to ransomware than the original scenario of an individual being taken hostage. In particular, we shall see that a key part of the model is a threat of aggression. In the context of ransomware, aggression merely means the files will be destroyed, while in the original context it meant the victim would be murdered. It would seem less controversial to quantify the loss of computer files than the loss of life.

[7] Normalizing the payoffs from no attack to zero can be done without loss of generality.

[8] We assume that the victim makes a counter-offer of some form, rather than simply ignoring the ransom demand. In the model, offering zero is at least as good as ignoring the ransom demand.

[9] These payoffs are based on the criminal not being caught (see Stage 6).

Table 1. The payoffs to different outcomes in the kidnapping game

Outcome	Payoffs	
	Criminal	Victim
Criminal does not infect computer	0	0
Release of files for ransom C and not caught	$C - G$	$-C$
Files destroyed and not caught	$-Y$	$-W$
Criminal caught after release of files	$-X$	0
Criminal caught after destroying files	$-Z$	$-W$

minimum acceptable offer M. If $C \geq M$, then the files are released or otherwise they are destroyed. Note that the model does not include the possibility of the criminal taking the ransom and not releasing the files. This is clearly an important possibility in terms of ransomware and so we return to the issue below.

Stage 6: With probability q the criminal is caught by the police. Note that this probability is assumed to be independent of the actions of the criminal (see Iqbal *et al.* [34] for an alternative approach). We assume that if the criminal is caught, the victim is recompensed any ransom but does not recover her files if they were destroyed. Our results are not, however, sensitive to this assumption and alternative assumptions, such as recovering the files but not the ransom (as happened with CyptoLocker) are easily modelled. The payoff of the criminal is $-X < 0$ or $-Z < 0$ depending on whether the criminal is caught after releasing or destroying the files. It is assumed that $-Z < -X < -Y$, implying a harsher punishment in case the files are destroyed.

A (pure) strategy for the criminal comprises three components: a choice to kidnap, a ransom demand, D and a minimum acceptable offer, M. A pure strategy for the victim consists of a function mapping from a ransom demand to a counter-offer. Table 1 summarizes the possible outcomes of the game and payoffs in each case. We see that the value of W proves particularly important. So let us note that this can be interpreted as the victim's *willingness to pay to recover her files*. Put another way, it is the victim's direct loss from losing access to her files. For instance, if the victim has recently performed a backup then $W \approx 0$, but if the files are valuable and no backup exists, then W will be large.

Main theoretical result

A Nash equilibrium for the game can be defined as a pair of strategies such that neither victim nor criminal has any incentive to change their strategy given the strategy of the other. The kidnapping game has many Nash equilibria and so we focus, as is the standard, on the subset of equilibria that is sub-game perfect. Our first result details the sub-game perfect Nash equilibrium of the game. Note that the theorem and its proof are different from that of Selten [10] but draw heavily on his approach.

Theorem 1. Generically, there exists a unique sub-game perfect Nash equilibrium of the kidnapping game: (a) If

$$W < (qX + (1-q)G)\left(\frac{1+a}{a}\right) \tag{1}$$

then the criminal will not infect the victim's computer.(b) Otherwise, the victim's computer is infected, the criminal makes demand

$$D^* = \left(\frac{a}{1+a}\right)\left(\frac{W}{1-q}\right), \tag{2}$$

the victim makes counter-offer $C = D^*$, and the files are released to the victim.

Proof: We proceed by backward induction. Consider Stage 5. If the files are released, the criminal has expected payoff,

$$V_R = (1-q)(C-G) - qX.$$

If the files are not released, the criminal has expected payoff

$$V_E = -(1-q)Y - qZ. \tag{3}$$

Given that $C > 0 \geq -Y$ and $-X > -Z$, it is trivial that $V_R > V_E$ provided $G < C$. Hence, the files are released. In interpretation, the criminal has nothing to gain from not taking the ransom and releasing the files.

Consider Stage 3: Given the optimal strategy of the criminal in Stage 5, the expected payoff of the victim is

$$\begin{aligned} U &= -(1-\alpha)(1-q)C - \alpha W \\ &= -\left(1 - a\left(1 - \frac{C}{D}\right)\right)(1-q)C - a\left(1 - \frac{C}{D}\right)W. \end{aligned}$$

Solving for the optimal value of C gives

$$C^*(D) = \begin{cases} D & \text{if D } \leq D_0 \\ \dfrac{W}{2(1-q)} - \dfrac{1-a}{2a} & \text{if } D_0 < D \leq D_1 \\ 0 & \text{if D} \geq D_1 \end{cases},$$

where

$$D_0 = \left(\frac{a}{1+a}\right)\left(\frac{W}{1-q}\right) \text{ and } D_1 = \left(\frac{a}{1-a}\right)\left(\frac{W}{1-q}\right).$$

In interpretation, if the ransom demand is low enough, where low enough is measured by D_0, then the victim pays the ransom. If the ransom is too high, where high is measured by D_1, then the victim does not offer to pay any ransom. For intermediate demands, the victim makes a counter-offer less than that demanded.

Consider Stage 2: From our analysis above, we know that the criminal will not choose to destroy the files. Let $\alpha^*(D) = a(1 - C^*(D)/D)$, and let $V^*_R(D) = (1-q)(C^*(D) - G) - qX$. Then the expected payoff of the criminal from choosing demand D is $V(D) = (1 - \alpha^*(D))V^*_R(D) + \alpha^*(D)V_E$, where V_E is given by equation (3). There are three cases to consider. (i) Suppose that $D < D_0$. Then $C^*(D) = D$ and $\alpha^*(D) = 0$. So, $V(D) = (1-q)(D-G) - qX$, which is clearly increasing in D. (ii) Suppose that $D_0 < D \leq D_1$. An increase in D increases $\alpha^*(D)$. It also decreases $C^*(D)$ and, therefore, $V^*_R(D)$. Given that $V^*_R(D) > V_E$ for all $D < D_1$, this means that $V(D)$ is a decreasing function of D. (iii) If $D \geq D_1$ then $V(D)$ is a constant function of D. Overall, therefore, $V(D)$ is maximized at D_0 giving equation (2).

Finally, consider Stage 1: Substituting in the optimal choice of $D = D_0$ gives an expected payoff for the criminal of

$$\begin{aligned} V(D_0) &= (1-q)\left(C^*(D_0) - G\right) - qX \\ &= \left(\frac{a}{1+a}\right)W - (1-q)G - qX. \end{aligned}$$

Setting $V(D_0) \geq 0$, gives inequality (1). QED

There are several salient points to take from Theorem 1. First of all, as one would expect, the criminal is more likely to infect the victim's computer if the probability of being caught is low. For instance, if $q = 0$ and so there is no chance of being caught, then the criminal will infect the computer. Experience suggests that the probability of facing punishment for a ransomware attack is very low across legal jurisdictions and this clearly invites attack.

Secondly, again as one would expect, the optimal ransom demand is increasing in the amount the victim is willing to pay to regain her files. This will have a knock on effect on the incentive of the criminal to infect the computer in the first place. For instance, if $W = 0$ because, say, the victim has backed up her files, then the criminal has no incentive to infect the computer. If W is large then the incentive is higher.

A more surprising finding is the role of irrational aggression or random destruction. If there is no chance of random destruction, meaning $a = 0$, then the optimal ransom demand is 0, and so it is not in the criminal's interest to infect the computer. The intuition behind this result is that, without the threat of irrational aggression, the criminal will accept any positive offer from the victim (because something is better than nothing) and so a high ransom demand is simply non-credible. The threat of aggression is, therefore, key to the criminal's bargaining power. The more likely is 'random destruction' (or the victim's perception of it) then the higher is the optimal ransom demand (see also [9]).

It may seem counter-intuitive that the criminal benefits from the likelihood he will do something 'irrational' but this is a common finding in game theoretic models of bargaining [35]. Essentially, it is in the criminal's interest to 'tie his hands' so that he cannot accept a low counter-offer and irrational aggression achieves this end. A specific example would be a criminal who simply does not allow any counter-offers. This would equate to a high a and would mean (if the probability of being caught is low) that the criminal will obtain a ransom near to the victim's willingness to pay to recover her files.

There are various simple extensions that one can make to the kidnapping game to accommodate alternative specifications. For instance, it may be that the victim is credit constrained and so cannot afford to pay a high ransom, even if she would want to [10]. If the victim can pay at most $\bar{W}$ then it is simple to show that the optimal ransom demand is $\min\{\bar{W}, D^*\}$, where D^* is the same as in the statement of Theorem 1. Basically, it is not in the criminal's interest to make a ransom demand that the victim cannot afford. We can also reconsider our assumption that the victim recovers the ransom if the criminal is caught. It is simple to show that if the ransom is not recovered, then q drops out of the equation for the optimal ransom demand (equation (2)), meaning the optimal ransom is lower.

In the following two sections, we explore more elaborate extensions of the kidnapping game (not considered by Selten [10]) that seem relevant to ransomware.

The criminal's incentive to return files

Recall that in the kidnapping game the criminal can, in Stage 5 of the game, only take the ransom if he returns the files. What if the criminal can keep the ransom and not does return the files to the victim? Clearly, this is a distinct possibility in the case of ransomware, given the inability of the victim to track the criminal. Also, as discussed earlier, we know that the criminals do sometimes take the money and run. It is beyond the scope of the current article to analyse how criminals could generate trust, but we can give interesting insight on the problem. In particular, it may intuitively seem advantageous for the criminal that he need not return the files. A little game theoretic reasoning shows, however, that it is not advantageous to have this possibility.

To see why, suppose that the criminal would prefer to take the ransom and not return access to the files. For instance, there may be some cost involved in returning the files, or properly encrypting the files in the first place. If the victim anticipates that the criminal will not return the files, then he has no incentive to pay any ransom. But if the victim will not pay any ransom, there is no incentive for the criminal to infect the computer in the first place. In short, the possibility that the criminal will take the money and run undermines the criminal's ability to make money. This is another illustration of how the criminal can benefit from having his hands-tied. In this case, it is to his benefit that he cannot take the money and run.

To better appreciate the issue we shall contrast two alternatives to Stage 5 of the game. The first alternative is as follows.

Stage 5: If the files were not destroyed in Stage 4, then the criminal chooses between releasing the files and receiving C, or destroying the files and receiving $-Y$. The criminal determines a minimum acceptable offer M. If $C \geq M$, then the files are released, otherwise they are destroyed. If the files are released, then there is probability β that they are not accessible because of error. If they are not accessible, then the payoff of the victim is $-W–C$ because she pays the ransom but still loses her files.

We will call this a 'kidnapping game with error' to capture the fact that files may be lost even if the criminal did not intend this. It is worth recognizing that error is a distinct possibility with ransomware, given the technical difficulties of encrypting and decrypting a large number of disparate files. We do observe instances in which private keys are returned (and genuine looking help is provided by the criminals) but not all files are recoverable [23]. It is also important to appreciate that error (as we have defined it) is different to irrational aggression. In particular, error happens independent of the ransom demand and counter-offer, while irrational aggression or random destruction is caused by a gap between demand and counter-offer. Moreover, in the former case the ransom is paid, while in the latter it is not (because the criminal refuses the offer).

The following result is a natural extension of Theorem 1 to capture the possibility of error.

Corollary 1. In the kidnapping game with error, the, generically, unique sub-game perfect equilibrium is such that the victim's computer is infected if and only if

$$W \geq \left(\frac{qX + (1-q)G}{1-\beta}\right)\left(\frac{1+a}{a}\right).$$

If infected, the criminal makes ransom demand,

$$D^{**} \geq \left(\frac{1+a}{a}\right)\left(\frac{W(1-\beta)}{1-q}\right).$$

and the victim makes counter-offer $C = D^{**}$.

Proof. We need to revisit Stage 3 of the proof of Theorem 1. The expected payoff of the victim is now

$$U = -(1-\alpha)(1-q)C - \alpha W - (1-\alpha)\beta W.$$

The final $(1-\alpha)\beta W$ term captures the possibility that the files are lost irrespective of irrational aggression. Solving for the optimal value of C gives

$$C^*(D) = \begin{cases} D & \text{if D } \leq D_0 \\ \dfrac{W(1-\beta)}{2(1-q)} - \dfrac{1-a}{2a} & \text{if D}_0 < D \leq D_1 \\ 0 & \text{if D} \geq D_1 \end{cases},$$

where

$$D_0 = \left(\frac{a}{1+a}\right)\left(\frac{W(1-\beta)}{1-q}\right) \text{ and } D_1 = \left(\frac{a}{1-a}\right)\left(\frac{W(1-\beta)}{1-q}\right).$$

The proof then follows through as for Theorem 1. QED

Corollary 1 shows that the optimal ransom demand is decreasing in β and so the more likely it is that the files will be lost the lower the ransom the criminal can demand. Hence, the criminal does not gain from the possibility of error. This provides an interesting tradeoff whereby the criminal's bargaining power relies on the possibility of irrational aggression but is diminished by the possibility of purely random error. It may be difficult for criminals to walk this dividing line between being tough on those who do not pay and fair on those who do. For instance, postings by victims on web forums are likely to simply say that 'my files were destroyed' without giving a nuanced commentary on ransom bargaining. This 'noisy information' makes it difficult to build a tough but fair reputation.

The preceding discussion relates to inadvertent error. What if we give the criminal the chance to deliberately take the money and run? Consider a further variation on Stage 5 of the game.

Stage 5: If the files were not destroyed in Stage 4, then the criminal chooses between releasing the files and receiving $C - G$ or destroying the files and receiving C.

We will call this a 'kidnapping game with deception' to capture the fact that the criminal may take the ransom money and not return the files. This is captured in our setting by assuming that this avoids the cost G of providing a good customer support etc.

Corollary 2. In the kidnapping game with deception, the, generically, unique sub-game perfect equilibrium is such that: (a) if

$$G > \frac{q(Z - X)}{1 - q}$$

The criminal does not infect the computer. (b) Otherwise, the equilibrium is the same as in the kidnapping game (except part (a) would condition on $W - G$ rather than W).

Proof. We need to revisit Stage 5 of the proof of Theorem 1. If the files are released, the criminal has expected payoff

$$V_R = (1 - q)(C - G) - qX.$$

If the files are not released, the criminal has expected payoff

$$V_E = (1 - q)C - qZ.$$

It is, therefore, in the criminal's interest to destroy the files if G is sufficiently high. If the criminal destroys the files, then there is no incentive for the victim to pay any ransom and so no incentive for the criminal to infect the computer. If the criminal does not destroy the files, then the equilibrium follows from the proof of Theorem 1. QED

Corollary 2 shows that the criminal cannot possibly gain from the ability to deceive. If the gains from deception, i.e. G, are large, then this undermines the whole basis of ransomware because nobody will pay a ransom to a criminal who is likely to take the money and run [36]. If the gains from deception are not large, then the criminal will not use the option and so does not benefit from the ability to use it. In practice, we can expect that $Z \approx Y$ because punishment will be the same irrespective of whether the criminal released files. Also, we can expect that q is small because of the small probability of capture. This means that the smallest gain (or saving in costs) from not releasing the files may be enough to undermine the criminal's ability to profit.

In a one-shot context it is difficult to envisage how a criminal could credibly overcome this problem and commit to returning the files. If, however, the criminal targets multiple individuals over time, then he can create a reputation for returning files. The crucial insight we have is that it is in the criminal's interest to build up such a reputation because any short-term gain from taking the money will be quickly offset by the unwillingness of future victims to pay any ransom.[10] Indeed, as we have seen, it will be in the criminal's interest to have a 100% record of returning files to those who pay the ransom. This can be captured within our model by revisiting the interpretation of G. We introduced G as the cost of engaging with the victim and returning files, but we could also include reputational damage for not honouring a ransom payment. Such reputational damage would decrease G and indeed may make G negative. We can see in Corollary 2 that this undermines the incentive to take the money and run.

For now, let us reiterate that a reputation for 'honouring payments' if a ransom is paid is not at odds with a reputation for irrational aggression if a ransom demand is not met. The criminal's bargaining position is highest if he is tough on those that don't pay (a is large) and fair to those who do (β is small). A tough but fair approach gives maximum incentive for the victim to pay the ransom.

Incomplete information on willingness to pay

The kidnapping game is one of complete information in which both criminal and victim know the payoff values given in Table 1. Particularly important is the assumption that the criminal knows the willingness of the victim to pay to recover her files, W. In reality, the criminal is unlikely to know W and this will undoubtedly have important implications for equilibrium outcomes. Unfortunately, no study has analysed the consequences of incomplete information in the kidnapping game. This is presumably because in many hostage-taking situations it is not unreasonable that W would be common knowledge.[11] In the case of ransomware, however, we clearly need to take account of uncertainty regarding W.

Despite the lack of formal analysis, it is possible to make some relatively firm conjectures regarding the likely consequences of incomplete information. One thing to note is that there no reason to expect incomplete information will fundamentally change any of the conclusions we have drawn so far. In particular, the role of irrational aggression and a reputation for returning files to those who pay the ransom will remain. Taking account of incomplete information will, though, impact the probability of the victim recovering her files. Theorem 1 shows that in equilibrium the victim always retains her files (either because her computer is not infected or she pays the ransom). This result is critically dependent on complete information because it relies on the criminal being able to calculate the maximum ransom that the victim will pay.

If there is incomplete information, then the criminal is not in a position to calculate the optimal ransom to charge each individual victim. Instead, he will have to work with aggregates and calculate the optimal ransom for the 'average' victim. The inevitable consequence of this is that some victims will refuse to pay the ransom because their willingness to pay is relatively low. The better the criminal's ability to predict or infer W, then the more profit he can earn. This provides a strong incentive for the criminal to price discriminate based on the characteristics of the victim [19]. And, in principle, the criminal may be able to infer quite a lot about the

[10] Formally, this will depend on the strategy of the victims. But any form of trigger strategy in which a victim refuses (with significantly high probability) to pay if a previous victim did not recover her files would lead to this result.

[11] For instance, the amount that governments have paid to release hostages from war zones is relatively well known.

victim given that he has free rein to look at the victim's computer and files. Moreover, the criminal may have access to data on the past willingness of victims to pay. It is in the criminal's interest to use this in order to reduce imperfect information as much as possible.

Having made this point, let us emphasize that bargaining with the victim is not a good method of inferring willingness to pay. To illustrate the point, suppose that there are two types of victims, a low-type with willingness to pay W_L, and a high-type with willingness to pay $W_H > W_L$. If the criminal could perfectly tell the type of the victim then he can replace W in equation (2) with either W_L or W_H and determine the optimal ransom D_L^* and D_H^*. As we would expect, a higher ransom would be asked of those with a higher willingness to pay, $D_L^* < D_H^*$. Suppose, however, that type is private information and so the criminal cannot infer type of victim. Let p denote the probability that the victim is of high-type. Call this a kidnapping game with unknown type.

Corollary 3: If $(1-a)W_H \geq (1+a)W_L$, then the following is a (Bayesian) equilibrium of the kidnapping game with unknown type: (a) If

$$\max\{W_L, pW_H\} < qX\left(\frac{1+a}{a}\right),$$

then the criminal will not infect the victim's computer. (b) Otherwise, the victim's computer is infected. If $pW_H < W_L$ the criminal makes demand

$$D_L^* = \left(\frac{a}{1+a}\right)\left(\frac{W_L}{1-q}\right),$$

Then the victim makes counter-offer $C = D_L^*$, and the files are released to the victim. Otherwise, the criminal makes demand

$$D_H^* = \left(\frac{a}{1+a}\right)\left(\frac{W_H}{1-q}\right),$$

The high-type makes counter-offer $C = D_H^*$ and the files are released. The low-type makes counter-offer 0 and the files are destroyed.

Proof. In Stage 5, we still have that any positive offer will be accepted. So consider Stage 3. If the criminal sets ransom D_H^* then we can see that the high-type maximizes payoff by setting $C = D_H^*$ and the low-type by setting $C = 0$. Revenue for the criminal is then pD_H^*. If the criminal sets ransom D_L^*, his revenue is D_L^*. QED

Corollary 3 shows that if the gap between the willingness to pay of the high- and low-type of victims is sufficiently large, then the criminal does best to make a choice of whether to target the high- or low-type. This choice will depend on the probability of the victim being of high-type and the gap in willingness to pay. Clearly, if the criminal does best to target the high-type, then this means the victim will not recover his files if she is a low-type.

A simple game model of deterrence

In this section, we turn to the model of Lapan and Sandler [11], further developed by Brandt *et al.* [37]. Note that the model was developed to study government policy towards terrorist kidnapping and hijacking, and so the primary focus is on deterring attack. We shall see, however, that the model can still provide valuable insight on ransomware. In doing so we focus on the one shot interaction between a criminal and a victim. This contrasts with Lapan and Sandler [11] who focus on repeated interaction between a government and a terrorist organization [8]. Given the difference in focus, our results and analysis are distinct from those of Lapan and Sandler [11]. To be more specific, the game we study is analogous to that of Lapan and Sandler [11] but all of our results are new. Closer to our analysis, as we shall discuss below, is that of Laszka *et al.* [12].

Again, we have a game with two players: a criminal and a victim. We shall refer to the game as the 'deterrence game'. The game consists of the four stages detailed below.

Stage 1: The potential victim chooses how much to spend deterring attack. This could be equated with virus protection, greater care in opening files etc. Denote expenditure by $E \geq 0$.

Stage 2: The criminal chooses whether or not to attack the victim's computer. If the computer is not attacked then the game ends. The criminal has payoff 0 and the victim's payoff is $-E$. Note that expenditure on deterrence is a sunk cost.

Stage 3: If the criminal chooses to attack, then with probability $\theta(E)$ the attack is a failure, where θ is a continuous, differentiable monotonically increasing function of E.[12] With probability $1-\theta(E)$, the attack is a 'success' and the victim's files are infected. If the attack is a failure, then the game ends. The criminal has payoff $-F < 0$ and the victim's payoff is $-P-E$, where $-P \leq 0$ includes costs (e.g. downtime) from repelling attack.

Stage 4: If the attack is a success, then the criminal makes a ransom demand C. The victim can either pay or not pay the ransom. If the victim pays the ransom then she regains access to her files. Her payoff is $-C-E-B$, where $B \leq P$ captures the damage from the attack, and the payoff of the criminal is C. Let $A = P - B$ be the difference in damage between a failed attack and an attack where the ransom is paid. If the victim does not pay the ransom, then the files are destroyed. Her payoff is $-W-E$ and the payoff of the criminal is $-L \leq 0$, where L includes the cost of the attack.[13] Any damage to the victim is assumed to be captured by W. Note that there is no chance to negotiate the ransom.

A (pure) strategy for the criminal consists of a choice to attack and demand a ransom of C. A strategy for the victim consists of an amount spent on deterrence and the decision whether or not to pay the ransom. Table 2 summarizes the possible outcomes of the game and payoffs in each case.

Before we continue to the analysis, let us set out how our model differs from that of Laszka *et al.* [12]. The key differences come in Stages 1 and 2 of the game. In their setting, the victim can spend resources on backup. This is essentially the analog of our Stage 1. Crucially, however, a backup (in their model) does not reduce the probability of a successful attack (as in our Stage 3) but instead reduces the potential losses from an attack. Meanwhile, the criminal can determine the amount of resource spent attacking two different types of victims. This is the analog of our Stage 2, but it is this spending that determines the probability of a successful attack. The complementary insights of the two models will be discussed further below.

Main theoretical result

Again, we focus on solving for the set of sub-game perfect Nash equilibria. The function θ is going to prove crucial and measures the

[12] We allow that θ may not be differentiable at point $\overline{E} = min_E\{\theta(E) = 1\}$. Clearly, $\theta(E) = 1$ for all $E > \overline{E}$.

[13] Lapan and Sandler [11] allow that L may be positive. In a terrorist setting, this is because a successful hostage taking can generate publicity. The payment of ransom may, therefore, be of secondary benefit. In the ransomware setting, however, it is difficult to conceive of a net-benefit without the payment of the ransom.

Table 2. The payoffs to different outcomes in the deterrence game

Outcome	Payoffs	
	Criminal	Victim
No attack	0	$-E-E$
Failed attack	$-F-F$	$-P-E-P-E$
Release of files for ransom CC	CC	$-C-B-E-C-B-E$
Ransom not paid	$-L-L$	$-W-E-W-E$

returns to spending on deterrence. To simplify the analysis, we will assume that θ is weakly convex. More formally, for any $E' < E''$, where $\theta(E'') < 1$, and any $\lambda \in (0,1)$, we assume that $\theta(\lambda E' + (1-\lambda)E'') < \lambda\theta(E') + (1-\lambda)\theta(E'')$. Alternative specifications of θ will be discussed as we proceed.

Theorem 2. If θ is weakly convex, then, generically, there exists a unique sub-game perfect Nash equilibrium of the deterrence game. (a) If $W > C$ and

$$\hat{E} = \theta^{-1}\left(\frac{C}{F+C}\right) < \left(1-\theta(0)\right)C + \theta(0)A - B = U_0 \quad (4)$$

then the victim spends $\hat{E}$ on deterrence and the criminal does not attack. (b) If $W > C$ and $\hat{E} > U_0$, then the victim does not spend on deterrence, the criminal will attack, and if the attack is successful, the victim will pay the ransom. (c) If $W < C$, then the victim does not spend on deterrence, the criminal does not attack and the victim would not pay a ransom.

Proof. We proceed by backward induction. Suppose that $W > C$. Then the victim will pay the ransom. The expected payoff of the criminal if he attacks the victim's computer is therefore

$$V = \left(1-\theta(E)\right)C - \theta(E)F.$$

The payoff if he does not attack the computer is 0. The criminal will thus attack if and only if $E < \hat{E}$, where $\hat{E}$ solves

$$\theta(\hat{E}) = \frac{C}{F+C}.$$

Consider Stage 1. The victim clearly has no incentive to choose $E > \hat{E}$ as the criminal is deterred when $E = \hat{E}$. Her expected payoff with full deterrence is $-\hat{E}$. Her expected payoff with deterrence $E < \hat{E}$ is

$$\begin{aligned} U(E) &= -(1-\theta(E))(C+B) - \theta(E)P - E \\ &= -C - B + (C-A)\theta(E) - E. \end{aligned}$$

Note that

$$\frac{dU(E)}{d(E)} = (C-A)\frac{d\theta(E)}{dE} - 1.$$

Weak convexity of θ means that it can never be optimal to set $E \in (0, \hat{E})$. The victim will therefore choose between no deterrence $E = 0$ or full-deterrence $E = \hat{E}$. Her expected payoff with no deterrence is $-(1-\theta(0))(C+B) - \theta(0)P$. So, it is optimal to choose deterrence if and only if $\hat{E} < (1-\theta(0))(C+B) + \theta(0)P$.

Suppose that $W < C$. Then the victim will not pay the ransom. The expected payoff of the criminal if he attacks is therefore $(1-\theta(E))L - \theta(E)F < 0$. The payoff if he does not attack is 0. The criminal will thus not attack. Given that the criminal will not attack, the victim has no incentive to deter attack. QED

In interpreting Theorem 2, note that one crucial thing is whether the victim will pay the ransom. If $W > C$, then the victim's willingness to pay for his files exceeds the ransom and so he will pay. Any threat to not pay is simply non-credible.[14] Clearly, this incentivizes the criminal to infect the computer. This, however, incentivizes the victim to deter an attack. The second crucial thing is, therefore, the cost of deterring attack. If that cost is not too high, where high cost is determined by equation (4), the victim spends enough to deter attack. Deterrence works by making it unlikely that the criminal's attempt will succeed. If the cost of deterrence is too high, then the victim accepts the chance of her files being infected and pays the ransom if necessary.

What determines whether the cost of deterrence is high or low? This depends on the cost F of a failed attack. If F is small then deterrence can only work by being highly effective. If F is large then deterrence is easier. In the context of ransomware, the value of F will likely be very small, given the low marginal costs of a criminal, say, sending out phishing emails. Indeed, failed attacks are clearly the norm in common uses of ransomware. A small F means that deterrence has to be highly effective at stopping attack if it is to deter criminals. This puts the focus on θ and the potential effectiveness of vigilance or anti-virus software. To be effective, the measures have to be essentially perfect at stopping an attempt to infect the computer.

Theorem 2 suggests that the victim will either spend nothing on deterrence or spend so much as to fully deter attack. This all or nothing approach follows directly from the assumption that θ is weakly convex. If θ is concave, then the victim may find it optimal to spend on deterrence even if this will not deter attack. To illustrate, we can work through a simple numerical example.

Suppose that $\theta(E) = \sqrt{E}$ for $E \leq 1$. Also, set $B = 0$ and $F = 0$ (or some small positive number). Then the potential victim could spend $E = 1$ on deterrence and be guaranteed to keep her files. This would leave final payoff as -1. Or she could spend $E < 1$ on deterrence, face the potential of being attacked, and have expected payoff $U(E) = -(1-\theta(E))W - E$. Maximizing $U(E)$ gives a candidate solution $E = W^2/4$ and $U(E) = -W + W^2/4$. Comparing the respective payoffs, we can see that if $W \geq 2$ then it is optimal for the victim to spend $E = 1$. This is analogous to outcome (a) in Theorem 2 and means that attack is deterred. If $W < 2$, then it is optimal for the victim to spend $E = W^2/4 < 1$ on deterrence. This is analogous to outcome (b) in Theorem 2, but note that the victim spends something on deterrence. The spending is not enough to deter the criminal but still means the victim is less vulnerable to attack.

This example illustrates that we need not expect victims to take an all or nothing approach to deterrence. There is scope for victims to spend on deterrence in order to reduce the probability of a successful attack. This is particularly possible if there are multiple approaches to deterrence. For example, we might find someone who buys anti-virus protection but is not cautious enough when opening email attachments or we might find someone who does not buy anti-virus but is cautious at opening attachments. This kind of approach may be optimal even if it does not completely immunize from attack.

Incomplete information

A somewhat trivial result for the deterrence game is that if the ransom demand is too high, $W < C$, then the victim will not pay the ransom and so the criminal has no incentive to infect the computer. This result seems a little strange in application. For instance, why does the criminal simply not ask a ransom that the victim is willing

[14] In a hostage-taking setting, there is a credible incentive to not pay theransom demand if this creates a reputation that deters future attack. This logic may be relevant in thinking about governments or firms that may face repeated attacks from criminals.

to pay? This brings us back to the issue of incomplete information that we discussed above. So, let us explore the consequences of the criminal not knowing how much a victim is willing to pay to recover her files.

To be specific, consider again the case in which there are two types of victims, a low-type willing to pay W_L, and a high-type willing to pay $W_H > W_L$ to recover her files. Suppose that criminal sets a ransom targeted at a high-type victim. In other words, the ransom, C, is set at $W_L < C \leq W_H$. For instance, it may be possible that the criminal is targeting firms (high-types) but will attack individuals (low-types) along the way. Theorem 2 can be applied to discern what the high-type will do. If

$$\hat{E} = \theta^{-1}\left(\frac{C}{F+C}\right) < \left(1-\theta(0)\right)W_H + \theta(0)A,$$

then the high-type will spend $E = \hat{E}$ on deterrence. This will deter all attacks and so the high-type not only defends herself but also the low-type. Indeed, the low-type can spend $E_L = 0$ on deterrence. In interpretation, we might think of the low-type as 'free-riding' on the vigilance of the high-type.

If $\hat{E} > (1 - \theta(0))W_H + \theta(0)A$, then the high-type will spend $E_H = 0$ on deterrence. This leaves both the high-type and the low-type open to attack. An interesting question is whether this incentivizes the low-type to spend on deterring attack. This is not as obvious as it may seem because the high-type 'only' pays the ransom (and doesn't lose their files), while the low-type stands to 'only' lose her files. Recall, however, that $W_L < C$ and so the low-type values her files by less than the ransom. This means the high-type still has more to lose than the low-type. So, if it is too costly for the high-type to deter attack, then the same must hold for the low-type.

Spillover effects of deterrence

In the preceding section, we saw that the low-type will spend less on deterrence than the high-type. In interpretation, we suggested this means the low-type is somewhat at the mercy of the high-type. In particular, if the high-type spends enough to deter attack, then the low-type benefits 'for free'. It is, though, important to distinguish different types of deterrence before attaching any kind of moral judgement on who is better or worse.

In the deterrence game spending reduces the probability of an attack being successful. This modelling assumption naturally fits certain types of deterrence such as spending on malware or greater vigilance in checking email attachments. And, in this case, the term free-riding may be appropriate. For example, if large corporations (high-types) spend sufficient funds on cyber-security to deter criminals, then small corporations or individuals (low-types) may not need to devote such high resources to cyber-security. So, low-types gain from the spending of high-types.

Another form of deterrence, which is not captured in the deterrence game, is for the potential victim to lower the value of W. For example, someone who regularly backs up their files would have a much lower W than someone who did not do so because they have less to lose from not being able to recover their files. If everyone were to regularly back up files and have a low W, then the incentives for the criminals to attack would be much diminished. So, in this context the term free-riding seems somewhat unjustified. In particular, those who regularly back up their files (low-types) may still be vulnerable to attack because the criminals are targeting those who do not back up their files (high-types).

More generally, we see that there are important spillover effects from one interaction to another. One person's spending on deterrence, in lowering the incentives of the criminals, will likely have a positive benefit for others. It is, though, unlikely that potential victims would take this into account when spending on deterrence. These externalities are also picked up by Laszka *et al.* [12]. Note that this is different from the setting originally considered by Lapan and Sandler [11] of a government repeatedly interacting with hostage takers. In this latter case, the externality is internalized because the government is the 'victim' every time. In ransomware, however, it is disparate individuals or firms that will be targeted, and so the externality is not internalized. This complicates attempts to combat ransomware.

Discussion and other literature

In this section, we draw together the previous analysis, and compare and contrast results. We also bring in insights from the rest of the game theoretic literature on kidnapping. Particularly important is to compare and contrast the two models analysed in this article together with that of other closely related work such as Caporusso *et al.* [13], Hernandez-Castro *et al.* [19] and Laszka *et al.* [12]. The basic point to appreciate is that all of these models look at complementary aspects of ransomware. It would be relatively simple to plug all these models together and come up with a big overarching ransomware game, but as we now discuss this is ultimately unlikely to lead to any additional insight. What we need to do is to clearly isolate the contribution that different models can make and the ways they can be extended. This seems especially apt, given that game theoretic modelling of ransomware is in its infancy and much work remains to be done.

The kidnapping game we analysed above primarily informs on the bargaining process between criminal and victim, the optimal ransom demand and the incentives to return files to the victim (see also [15, 19]). Our main results are that bargaining does not benefit the criminal, the optimal ransom demand will depend on the threat of random destruction (if the ransom is not paid in full) and the criminal should honour ransom payments (if paid in full). The deterrence game analysed above and the model of Laszka *et al.* [12] simply assume away these issues. Specifically, they take as given, no bargaining, a fixed ransom demand and that files will be returned. The crucial thing to observe is that these assumptions are supported by our analysis (including Theorem 1) and so Theorem 2, and our analysis of the deterrence game is robust. For instance, to add bargaining to the deterrence game would not change Theorem 2, or add any insight we are not able to get from Theorem 1.

The deterrence game that we analysed primarily informs on the incentives of potential victims to deter attack. The model of Laszka *et al.* [12] informs on the incentives to do backups in order to mitigate the losses from attack. There may be some crossover between deterrence or backup and bargaining. For instance, a victim you have devoted resources to deterrence and back up could be expected to have a low willingness to pay, which then influences bargaining. Even so, Theorem 1 can pick this up as changes in W. To be more specific, In Theorem 2 the key determinant of whether the victim should spend enough to deter attack is

$$\hat{E} = \theta^{-1}\left(\frac{C}{F+C}\right) < \left(1-\theta(0)\right)C + \theta(0)A - B = U_0.$$

From Theorem 1 we get that

$$C = \left(\frac{a}{1+a}\right)\left(\frac{W}{1-q}\right).$$

These two equations contain independent variables and so we can readily combine them to provide an overall expression for whether the victim should spend on deterring attack.

A key insight that comes out of our deterrence game and Laska *et al.* [12] are the spillovers between different types of victims. Basically, the actions of one victim has implications for the likelihood of another suffering an attack. This is an issue that clearly warrants more study, particularly in terms of the practical steps a policy maker could take to internalize the externalities. For instance, is it best for governments to legislate on requirements for backup and deterrence or to use positive incentives such as subsidies for virus protection. This can feed into the general debate on how to encourage better cyber practice in a world of boundedly rational individuals [38, 39].

There is also more that we can learn from the game theoretic literature on kidnapping [40, 41]. From the perspective of a criminal, the main issue (if we set aside the more technical aspects of launching successful attacks) is to maximize ransom revenue. As we discussed above, this is best achieved by the criminal 'tying his hands'. First, the optimal ransom demand needs to be determined [19]. Then, it is in the criminal's interest to not negotiate. Irrational aggression is a key part of the mix here because that provides the threat of files being destroyed. This threat (real or perceived) is important to motivating the victim to pay the ransom and 'do as the criminals want'.[15]

The key thing to appreciate here is the criminal's need to build a reputation of being tough but fair.[16] If victims don't deliver the ransom, then the criminals should be tough. But if the victims pay up, then they should get their files back. Note that this approach is consistent with the criminals providing a 'customer service' for victims because it provides a clear and credible set of rules for customers [26]. If the criminals can build a reputation, work out the optimal ransom demands and launch successful attacks, then they are going to make a large profit. However, building reputation may not be easy. For instance, the recent WannaCry and NotPetya attacks got huge publicity and spread the message that there is no point in paying the ransom. This is not good for those running profit-motivated attacks. We can expect, therefore, to see a push towards building a 'brand' that victims can 'trust'.

For potential victims the picture is less bright. The spillovers between individuals mean that it is very difficult for any one individual to 'win'. Spending on deterrence, particularly in terms of regular backups, is a strategy to minimize loss. But we should not lose sight of the fact that this is still a loss. The victim has to spend resource on deterrence and then potentially also to restore systems after attack. The key problem is the cheapness of the criminal for launching an attack. This means that as long as some victims are willing to pay the ransom, everyone faces the threat of attack. And that means that everyone needs to consider deterrence. Crucially, this means that ransomware has a significant cost even if there are relatively few instances where a ransom is actually paid.

Another element that aids the criminals is patience. In bargaining situations, the more patient party stands to benefit most [42]. In a ransomware attack the victim is almost certainly going to be in a hurry to recover their files while the criminals have little to lose from delay. The almost universal use of fixed deadlines and countdown timers in ransomware attacks presumably heightens the victim's sense of urgency [43]. Another thing that can work to the criminal's advantage is the lack of attention a particular attack brings. For instance, Gaibulloev and Sandler [42] and Sandifort and Sandler [44] find that the capture of a protected person weakens the negotiating position of terrorists because of the public scrutiny it generates. Similarly, we can expect that ransomware attacks that fly under the radar of mass media will be more successful because the resources to help and advise victims are going to be less readily available.

Conclusion

In this article we have applied and extended two seminal models from the game theoretic literature on kidnapping to the issue of ransomware. The first model (due to Selten [10]) informs on the bargaining process between criminal and victim. The second model (adapted from Lapan and Sandler [11]) informs on the optimal deterrence of potential victims. There is, as we have discussed, much work that could be done to extend the models further. For example, our approach does not explicitly take into account that there is a large population of victims with whom the criminals interact simultaneously. Moreover, we do not take account of potential 'competition' between different criminal gangs. Even so, our analysis has yielded some key findings, which we summarize below. We expect these findings are robust to more general analysis.

- The optimal ransom demand is increasing in the willingness of the victim to pay to recover her files. This means that it is in the criminal's interest to be as informed as possible about the victim's willingness to pay.
- The bargaining power of the criminal is enhanced by the likelihood of irrational aggression, i.e. the destruction of files if a ransom demand is not met. One way to achieve this is to not allow any counter-offers from the victim or to build a reputation of refusing any counter-offer.
- The bargaining power of the criminal is enhanced by a credible commitment to return files to any victim who pays the required ransom. The most likely way to achieve this is for the criminal to build a reputation of honouring ransom payments.
- Criminals will only be deterred from launching attacks if the measures to prevent successful attack, whether that be anti-virus software or personal vigilance, are near perfect. This seems unlikely.
- There are important spillover effects between potential victims. For instance, if the victims who value their files most spend enough to deter attack, then this benefits all users. Similarly, those who regularly back up files may still be vulnerable to attack and losses (even if small) because there are others who do little to deter attack. This suggests that it may be optimal to subsidize spending on cyber security or good backup practices.
- Deterrence is costly. Any estimate of the costs of ransomware, therefore, should take into account all the costs of deterrence and costs of dealing with an attack. The payment of ransoms is likely to be a relatively small fraction of the total social and economic costs of ransomware.

As things stand, we would suggest that ransomware is still in the early stages of its development. While the technological know-how exists, there is still a lot that the criminals can do to maximize their economic profit. And similarly, awareness of ransomware on the side of potential victims still appears rudimentary. Over time, therefore, we can

[15] We could also think of 'irrational aggression' on the victim side in refusing to pay. But the credibility of this is questionable, given that it does not benefit the victim in a one-off interaction. Lee [38], for instance, finds that democratic governments are more likely to pay terrorist ransom demands before elections (because a threat to not pay becomes non-credible).

[16] Wilson [39] provides evidence that this type of approach is also successful in terrorist hostage-taking scenarios.

expect a process of evolution as criminals and potential victims adopt 'better' strategies. Given that ransomware provides a viable long-term business model for the criminals, it is likely to be a crime that will be around for some time to come. Our analysis gives insight onto how ransomware will evolve and the costs it will impose on potential victims.

Conflict of interest statement. None declared.

Acknowledgements

We would like to thank three reviewers for their insightful comments on an earlier version of the article for the WEIS 2018 Conference, as well the reviewers for the journal. We would also like to thank the audience at WEIS for their suggestions. Finally, we would like to thank members of the EMPHASIS and RAMSES teams for their comments and suggestions as the work progressed, particularly Darren Hurley Smith and Tom Chen.

Funding

This work was supported by the Engineering and Physical Sciences Research Council (EPSRC) for project EP/P011772/1 on the EconoMical, PsycHologicAl and Societal Impact of RanSomware (EMPHASIS) and the European Union's Horizon 2020 research and innovation programme under grant agreement No. 700326 (RAMSES project).

References

1. Mansfield-Devine S. Ransomware: taking businesses hostage. *Netw Secur* 2016;**2016**:8–17.
2. Kalaimannan E, John SK, DuBoseT *et al*. Influences on ransomware's evolution and predictions for the future challenges. *J Cyber Secur Tech* 2017; **1**:23–31. http://dx.doi.org/10.1080/23742917.2016.1252191
3. Symantec. *Ransomware and Business 2016*. Symantec Corporation, 2016. http://www.symantec.com/content/en/us/enterprise/media/security_response/whitepapers/ISTR2016_Ransomware_and_Businesses.pdf (30 July 2019, date last accessed).
4. Young A, Yung M. Cryptovirology: extortion-based security threats and countermeasures. In: *Security and Privacy Proceedings IEEE Symposium*. IEEE 1996.
5. Kharraz A, Robertson W, Balzarotti D *et al*. Cutting the Gordian knot: a look under the hood of ransomware attacks. In: Almgren M, Gulisano V, Maggi F (eds), *Detection of Intrusions and Malware, and Vulnerability Assessment*. DIMVA 2015. Lecture Notes in Computer Science, vol 9148. Cham: Springer, 2015.
6. Barlyn B, Cohn C. Companies use kidnap insurance to guard against ransomware attackshttps. Reuters, 2017. www.reuters.com/article/us-cyber-attack-insurance/companies-use-kidnap-insurance-to-guard-against-ransomware-attacks-idUSKCN18F1LU
7. ShahinWN, IslamMQ. Combating political hostage-taking: an alternative approach. *Defence Peace Econ* 1992;**3**:321–27.
8. SandlerT, EndersW. An economic perspective on transnational terrorism. *Eur J Political Econ* 2004;**20**:301–16.
9. NaxHH. Modeling hostage-taking: on reputation and strategic rationality of terrorists. *Stud Confl Terror* 2008;**31**:158–68.
10. Selten R. A simple game model of kidnapping. In: *Models of Strategic Rationality*. Amsterdam, the Netherlands: Springer, 1988, 77-93.
11. Lapan HE, Sandler T. To bargain or not to bargain: that is the question. *Am Econ Rev* 1988;**78**:16–21.
12. Laszka A, Farhang S, Grossklags J. On the economics of ransomware. In: *International Conference on Decision and Game Theory for Security*. Springer, 2017, pp. 397-417.
13. Caporusso N, Chea S, Abukhaled R. A game-theoretical model of ransomware. In: *International Conference on Applied Human Factors and Ergonomics*. Cham: Springer, 2018, pp. 69-78.
14. Sandler T, Enders W. Applying analytical methods to study terrorism. *Int Stud Perspect* 2007;**8**:287–302.
15. Sandler T. The analytical study of terrorism: taking stock. *J Peace Res* 2014;**51**:257–71.
16. Schelling TC. *The Strategy of Conflict*. Cambridge, MA: Harvard University Press, 1980.
17. Cohen DS, Dormandy X. *Kidnapping for Ransom: The Growing Terrorist Financing Challenge*. Chatham House, 2012. https://www.chathamhouse.org/sites/default/files/public/Meetings/Meeting%20Transcripts/051012CohenQA.pdf (30 July 2019, date last accessed).
18. DuttonYM, BellishJ. Refusing to negotiate: analyzing the legality and practicality of a piracy ransom ban. *Cornell Int L J* 2014;**47**:299.
19. Hernandez-Castro J, Cartwright E, Stepanova A. Economic analysis of ransomware. *arXiv Preprint arXiv* 2017;**1703**:06660.
20. Huang K, Siegel M, Madnick S. *Cybercrime-as-a-service: identifying control points to disrupt*, 2017. Working paper CISL# 2017-17. http://web.mit.edu/smadnick/www/wp/2017-17.pdf (30 July 2019, date last accessed).
21. August T, Dao D, Laube S *et al*. Economics of ransomware attacks. In: *Workshop on Information Systems and Economics (WISE)*, 2017.
22. Jarvis K. *Cryptolocker Ransomware. Secure Works Counter Threat Unit*. http://www. secureworks.com/cyber-threat-intelligence/threats/cryptolocker-ransomware. (30 July 2019, date last accessed).
23. Hernandez-Castro J, Boiten E. *2016 Kent Cyber Security Survey*. University of Kent, 2016. https://cyber.kent.ac.uk/Survey2016.pdf (5 May 2018, date last accessed).
24. Liao K, Zhao Z, Doupé A *et al*. Behind closed doors: measurement and analysis of cryptoLocker ransoms in Bitcoin. In: *Electronic Crime Research 2016 APWG Symposium*. IEEE, 2016. https://doi.org/10.1109/ECRIME.2016.7487938.
25. Spagnuolo M, Maggi F, Zanero S. Bitiodine: extracting intelligence from the bitcoin network. In: *International Conference on Financial Cryptography and Data Security.* 2014, pp. 457–68. https://doi.org/10.1007/978-3-662-45472-5_29.
26. F-Secure. Evaluating the Customer Journey of Crypto-Ransomware. 2016. https://f-secure.bg/wp-content/uploads/2016/08/customer_journey_of_crypto-ransomware_f-secure.pdf (30 July 2019, date last accessed).
27. Huang DY, Aliapoulios MM, Li VG *et al*. Tracking ransomware end-to-end. In: *2018 IEEE Symposium on Security and Privacy*. IEEE, 2018, 618–31.
28. Paquet-Clouston M, Haslhofer B, Dupont B. Ransomware payments in the bitcoin ecosystem. *arXiv preprint arXiv:1804.04080.* 2018.
29. F-Secure. The Changing State of Ransomware. F-Secure Press Global, 2018. https://fsecurepressglobal.files.wordpress.com/2018/05/ransomware_report.pdf (30 July 2019, date last accessed).
30. Danielson D. The FBI says you may need to pay up if hackers infect your computer with ransomware. *BusinessInsider*, 2015. http://uk.businessinsider.com/fbi-recommends-paying-ransom-for-infected-computer-2015-10 (30 July 2019, date last accessed).
31. Rashid F. 4 reasons not to pay up in a ransomware attack. *InfoWorld*, 2016. https://www.infoworld.com/article/3043197/security/4-reasons-not-to-pay-up-in-a-ransomware-attack.html (30 July 2019, date last accessed).
32. Intermedia. Data vulnerability report. *IntermediaNet*, 2017. https://www.intermedia.net/report/datavulnerability2017-part2 (30 July 2019, date last accessed).
33. Volpicelli. Taking on the fancy bear hackers: how to negotiate if your data is being held ransom. *Wired Magazine*, 2017. http://www.wired.co.uk/article/negotiate-hackers-moty-cristal (30 July 2019, date last accessed).
34. Iqbal A, Masson V, Abbott D. Kidnapping model: an extension of Selten's game. *R Soc Open Sci* 2017;**4**:171484.
35. Muthoo A. *Bargaining Theory with Applications*. Cambridge, UK: Cambridge University Press, 1999.
36. Cusack BB, Gerard W. Points of failure in the ransomware electronic business model. In: 24th America's Conference on Information Systems, 2018. https://aisel.aisnet.org/amcis2018/eBusiness/Presentations/19/ (30 July 2019, date last accessed).
37. Brandt PT, George J, Sandler T. Why concessions should not be made to terrorist kidnappers. *Eur J Political Econ* 2016;**44**:41–52.
38. Baddeley M. Information security: lessons from behavioural economics. In: *Workshop on the Economics of Information Security*, 2011.

39. Pfleeger SL, Caputo DD. Leveraging behavioral science to mitigate cyber security risk. *Comp Secur* 2012;**31**:597–611.
40. Lee CY. Democracy, civil liberties, and hostage-taking terrorism. *J Peace Res* 2013;**50**:235–48.
41. Wilson MA. Toward a model of terrorist behavior in hostage-taking incidents. *J Confl Resolut* 2000;**44**:403–24.
42. Gaibulloev K, Sandler T. Hostage taking: determinants of terrorist logistical and negotiation success. *J Peace Res* 2009;**46**:739–56.
43. Hadlington L. Exploring the psychological mechanisms used in ransomware splash screens. SentinelOne Report, 2017.
44. Santifort C, Sandler T. Terrorist success in hostage-taking missions: 1978–2010. *Public Choice* 2013;**156**:125–37.