To pay or not: Game theoretic models of Ransomware

Edward Cartwright*† Julio Hernandez-Castro‡
Anna Stepanova§

25th May 2018

Keywords: Ransomware, game theory, kidnapping, hostage.

Abstract

Ransomware is a form 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 paper 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 will, however, see that there are important lessons that can be drawn on the likely evolution of ransomware and ways in which it can be tackled.

*Department of Strategic Management and Marketing, De Montfort University, Leicester, LE1 9BH UK. Corresponding author, email ejcartwright1@gmail.com.

†This project has received funding from the Engineering and Physical Sciences Research Council (EPSRC) for project EP/P011772/1 on the EconoMical, PsycHologicAl and Societal Impact of RanSomware (EMPHASIS). The authors also want to thank the European Union's Horizon 2020 research and innovation programme, under grant agreement No.700326 (RAMSES project), which also supported this work.

‡School of Computing, University of Kent, UK.

§School of Economics, Finance and Accounting, Coventry University, UK.

1 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 (Mansfield-Devine 2016, Kalaimannan et al. 2017). Victims are given a set time, typically 72 hours, to pay the ransom, which can vary from $100 to $1000 for individuals, and a lot more for firms and organizations (Symantec 2016). While the know-how to develop ransomware has existed in academia for some time (Young and Yung 1996) it is only recently that cryptographically sound ransomware has found its way into the wild. While many variants of ransomware exist that allow for reverse-engineering (Kharraz et al. 2015) 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 paper 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 kidnap of staff to mitigate losses from ransomware (Barlyn and Cohn 2017). Procedures can also require law enforcement agencies to call on trained hostage taking officers to execute ransom payments (to track the money flow). In this paper 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. Shahin and Islam 1992, Sandler and Enders 2004, Nax 2008). In this paper we adapt and apply two key models of kidnapping due to Selten (1988) and Lapan and Sadler (1988), and highlight findings that are particularly relevant to understanding ransomware.

Before we continue it seems pertinent to clarify that only one study, of which we are aware, has explicitly applied game theoretic models of kidnapping to ransomware (Laszka, Farhang and Grossklags 2017). 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 (Lapan and Sandler 1988, Sandler and Enders 2007, Sandler 2014). 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.[1] Even so, there are specific aspects of ransomware that point to issues which could be analyzed 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 modeling) issues than the kidnap (and potential murder) of a hostage (Cohen and Dormandy 2012, Dutton and Bellish 2014).

As already previewed we shall develop two key models of kidnapping. The first model, due to Selten (1988), primarily focuses on the optimal ransom that criminal's should charge. The second model, due to Lapan and Sadler (1988), 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 anti-virus protection, regular back-ups or similar. Some clear and actionable policy recommendations follow from our analysis.

Our paper adds to a small but growing literature on the economic aspects of ransomware. In earlier work (Hernandez-Castro et al. 2017) we look at the economic theory behind optimal ransom pricing. Insights from this work are applied in Section 3. Laszka, Farhang and Grossklags (2017) model the ransomware eco-system as a multi-stage, multi-defender game. The particular focus of the analysis is on the interaction between the decision to back up and pay a ransom. This model has a close overlap with the model we explore in Section 4 and so we discuss their results in detail then. Huang, Siegel and Madnick (2017) explore how ransomware can fit within a cybercrime business model while August et al. (2017) 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 behavior 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.

[1]This also means that in a game theoretic sense kidnapping can be used more widely than in common usage. For example, Schelling (1980) equates nuclear power with the ability to take hostages. Basically, if, say, the U.S. has the ability to destroy Russia then it is as if the U.S. takes Russian citizens as hostages.

We proceed as follows. In Section 2 we provide a brief overview of ransomware. In Sections 3 and 4 we apply the key models of Selten (1988) and Lapan and Sadler (1988) to ransomware. In Section 5 we shall summarize the remaining game theoretic literature on kidnapping and in Section 6 we conclude. Note that while none of the literature has explicitly studied ransomware we shall in this paper frame the analysis throughout in terms of a ransomware attack.

2 Overview of 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 protocol in a technically sound way, from its conception (Jarvis 2013). Its 'good' implementation unfortunately forced victims wanting to recover their files to pay the ransom. That was the only available alternative. Let us highlight that throughout this paper we will focus on cryptographically sound ransomware, like CryptoLocker, where the files are recoverable if and only if the criminals return the relevant keys.[2]

The precise proportion of victims who paid ransoms to Cryptolocker is unknown with estimates ranging from 2-40% (Hernandez-Castro and Boiten 2016). It is, however, clear that enough people paid the 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) (Liao et al. 2016, Spagnuolo, Maggi and Zanero 2014). 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 the total number of bitcoins in circulation (approx. 12 million) and would have had a valuation in excess of $200m.

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

[2]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.

this particular form of ransomware.[3] 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 (Symantec 2016, F-Secure 2017, Kalaimannan et al.2017).

Modern ransomware strands are fast evolving, not only in terms of technical capabilities but also economic sophistication. For instance, ransomware-as-a-service allows just about anyone to commit the crime irrespective of technical know-how (Huang et al. 2017). 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 (F-Secure 2018). Unfortunately, there is less evidence of individuals and organizations taking the necessary measures (particularly regular back-ups) to mitigate the damage from attack. This means that ransomware is likely to remain a serious threat for many years to come.

It is interesting to note that ransomware is rare in being a cyber-crime that positively benefits from publicity. The more that 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 presumably are to engage with the criminals. Indeed, the FBI was somewhat inadvertently dragged into such complexities when an agent was quoted in 2015 as saying 'The ransomware is that good ... To be honest, we often advise people to just pay the ransom' (Danielson 2015).This leads on to 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 honor ransom payments and return the key to decrypt the files. Consider, for instance, the widely publicized case of the University of Calgary paying $20,000 to get back their files. This is one example of a ransom payment that 'worked'.

[3]During Operation Tovar, a victim's 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.

More generally, some ransomware strands like CryptoWall developed a good reputation for returning the files (Rashid 2016). This means the victims have a reasonable chance of recovering their files leaving victims with a basic dilemma of whether to pay or not. The evidence suggests that many victims do indeed pay, particularly businesses (Intermedia 2017). This suggests that ransomware can provide a sustainable business model for criminals.[4]

3 A simple game model of kidnapping

In this section we apply and adapt the model of kidnapping due to Selten (1988). 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.[5] 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.[6] It has six stages which can be explained as follows.

Stage 1: The criminal chooses whether or not to infect the victims computer. If the files are not infected then the game ends and both players get payoff 0.

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

Stage 3: Having seen the demand D the victim chooses a counter-offer $C \in [0, D]$. Note that it is far from clear whether it is in the criminal's

[4]A typical ransomware strand may only be able to survive, say, 6 months before the law enforcement agencies start to close in. But, the criminals can evolve and continually develop new strains.

[5]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 a ransomware context aggression merely means the files will be destroyed while in the original context it meant the victim would be murdered. It would seem easier to quantify the loss of computer files than the loss of life.

[6]Note that in interpretation the victim need not necessarily be the person taken hostage; for instance, it could be a relative who is willing to pay the ransom or a government agent who is willing to act on behalf of the hostage.

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, though, note that almost all (genuine) ransomware strains allow for some form of communication with the criminals in order to make a counter offer (F-Secure 2016, Volpicelli 2017). 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. Selten (1988) equates this with 'irrational aggression' on the part of the criminal. More generally, it can be equated with a risk of aggressive behavior because the counter-offer is below that demanded.[7] In a game theoretic sense this is modeled as an act of nature and so we shall call it accidental destruction. As a reviewer of an earlier version pointed out such accidental destruction could be programmed into the malware itself by the criminal. Let us, however, highlight that the crucial thing here is the *victim's perception* of the probability the files will be accidentally destroyed. Destruction of the files results in a payoff of $-Y \leq 0$ for the criminal and $-W < 0$ for the victim.

Stage 5: If the files were not destroyed then the criminal chooses between releasing the files and receiving C, or destroying the files and receiving 0. In interpretation we can think of the criminal as having a minimum acceptable offer M. If $C \geq M$ then the files are released and 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 the issue in Section 3.2.

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, Masson and Abbott 2017 for an alternative). If the criminal is caught then the victim is recompensed any ransom and the criminal is punished. 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 $-Z < -X$, implying a harsher punishment in the case the files are destroyed.

A (pure) strategy for the criminal consists of a choice to kidnap, a ransom demand and a minimum acceptable offer. Let S_C denote the set of strate-

[7] Clearly if $C = D$ then the probability of irrational aggression is zero.

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

Table 1: The payoff to different outcomes in the kidnapping game

gies. A pure strategy for the victim consists of a function mapping from a ransom demand to a counter-offer. Let S_V denote the set of strategies. Given strategies $s_C \in S_C$ and $s_V \in S_V$ let $u_C(s_C, s_V)$ and $u_V(s_V, s_C)$ denote the respective payoffs of the criminal and victim. Table 1 summarizes the possible outcomes of the game and payoffs in each case. We shall 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 back up then $W \approx 0$ but if the files are valuable and no back up exists then W will be large.

3.1 Main theoretical result

A Nash equilibrium is a pair of strategies $s_C^\star$ and $s_V^\star$ such that $u_C(s_C^\star, s_V^\star) \geq u_C(s_C, s_V^\star)$ for all $s_C \in S_c$ and $u_V(s_V^\star, s_C^\star) \geq u_V(s_V, s_V^\star)$ for all $s_V \in S_V$. The kidnapping game has many Nash equilibria and so we focus, as is standard, on the subset of equilibria that are 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 to that of Selten (1988) but draw heavily on his approach.

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

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

then the criminal will not 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

$$V_R = (1-q)\,C - qX.$$

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 $V_R > V_E$. 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: The expected payoff of the victim is

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

Solving for the optimal value of C gives

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

where

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

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: We know that the criminal will not choose to destroy the files. Let $\alpha^*(D) = a\left(1 - \frac{C^*(D)}{D}\right)$ and let $V_R^*(D) = (1-q)C^*(D) - qX$. 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, $V(D) = (1-q)D - qX$ which is clearly increasing in D. (ii) Suppose that $D_0 < D < D_1$. 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

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

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

There are several salient points to take from Theorem 1. 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 encourages attack. Also as one would expect, the optimal ransom demand is increasing in the amount the victim is willing to pay to regain her files. 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.

More surprising is the role of irrational aggression or accidental destruction. If there is no chance of accidental 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 irrational aggression (or

the victim's perception of it) then the higher is the optimal ransom demand (see also Nax 2008).

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 (Muthoo 1999). 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 (Selten 1988). If the victim can pay at most $\overline{W}$ then it is simple to show that the optimal ransom demand is $\min\left\{\overline{W}, D^*\right\}$, 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. In the following two sub-sections we explore more elaborate extensions of the kidnapping game (not considered by Selten 1988) that seem relevant to ransomware.

3.2 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 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. Intuitively, it may 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 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 then the criminal chooses between releasing the files and receiving C, or destroying the files and receiving 0. The criminal determines a minimum acceptable offer M. If $C \geq M$ then the files are released and otherwise they are destroyed. If the files are released then there is probability β they are destroyed be error.

We will call this a *kidnapping game with error* to capture the fact that files may be lost even if the criminal and victim 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 (Hernandez-Castro and Boiten 2016). It is also important to appreciate that error (as we have defined it) is different to accidental destruction. In particular, error happens independent of the ransom demand and counter-offer, while accidental destruction or irrational aggression is caused by a gap between demand and counter-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

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

If infected, the criminal makes ransom demand

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

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

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

where

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

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 trade-off whereby the criminal's bargaining power relies on the possibility of accidental damage through 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 '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 then the criminal chooses between releasing the files and receiving $C - G$, or destroying the files and receiving C for some $G \geq 0$.

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. The G

captures the cost of having to properly engage with the victim in order to decrypt the files. It may, for instance, involve customer support (F-Secure 2016). But, note that we do not rule out $G = 0$ meaning that it is costless to return the keys.

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 criminals interest to destroy the files if G is sufficiently high. If the criminal will destroy 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 will 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. 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. Note 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.[8] Indeed, it will, as we have seen, be in the criminal's interest to have a 100% record of returning files to those who pay the ransom.

For now let us reiterate that a reputation for 'honesty' 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.

3.3 Incomplete information on willingness to pay

The kidnapping game is one of complete information in which both criminal and victim know the payoff values 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 analyzed 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.[9] 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

[8]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.

[9]For instance, the amount that government's have paid to release hostages from war zones is relatively well known.

particular, the role of irrational aggression and a reputation for returning files to those who pay the ransom will remain. Taking account of imperfect information will, though, impact on 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 criminal (Hernandez-Castro, Cartwright and Stepanova 2017). And, in principle, the criminal may be able to infer quite a lot about the victim given that he has free rein to look at the victim's computer and files. Moreover, the criminal has 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 victim, 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. Let p denote the probability the victim is 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

$$max\{W_L, pW_H\} < qX\left(\frac{1+a}{a}\right) \tag{4}$$

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

the victim's computer is infected. If $pW_H < W_L$ the criminal makes demand

$$D_L^* = \left(\frac{a}{1+a}\right)\left(\frac{W_L}{1-q}\right), \tag{5}$$

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

$$D_H^* = \left(\frac{a}{1+a}\right)\left(\frac{W_H}{1-q}\right). \tag{6}$$

The hight 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^* 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 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 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 he is a low type. In practice we can expect at least some victims to have recent back-ups meaning that there willingness to pay is low. Corollary 3 illustrates that it is not in the interests of the criminals to target such types. More profitable is to price according to high types (without a back-up).

4 To Bargain or Not to Bargain

In this section we turn to the model of Lapan and Sadler (1988), further developed by Brandt, George and Sandler (2016). 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 victim. This

contrasts with Lapan and Sadler (1988) who focus on repeated interaction between a government and terrorist organization (Sandler and Enders 2004). Given the difference in focus our results and analysis are distinct from those of Lapan and Sandler (1988). Closer to our analysis, as will shall discuss below, is that of Laszka et al. (2017).

Again, we have a game with two players, a criminal and 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 victims computer. If the computer is not attacked then the game ends. The criminal has payoff 0 and the victim payoff $-E$.

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.[10] With probability $1-\theta(E)$ the attack is a 'success' and the victim's files are infected. If the attack is a failure the game ends. The criminal has payoff $-F < 0$ and the victim payoff $-P-E$, where $-P \leq 0$ is damage from the attack.

Stage 4: If the attack is a success 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 $-C-E-B$ and the payoff of the criminal is C, where $B \leq P$ captures the damage from the attack. 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 $-L \leq 0$.[11] In this case any damage is captured by W. Note that there is no chance to negotiate the ransom.

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

[11] Lapan and Sandler (1988) 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.

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

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

A (pure) strategy for the criminal consists of a choice to kidnap. 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. (2017). The key differences come in stages 1 and 2 of the game. In their setting the victim can spend resources on back-up. This is essentially the analog of our stage 1. Crucially, however, a back-up 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 victim. This is the analog of our stage 2, but it is this spending that determines the probability of successful attack. The complementary insights of the two models will be discussed further below.

4.1 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 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 $\theta(\lambda E' + (1-\lambda)E'') \leq \lambda\theta(E') + (1-\lambda)\theta(E'')$. Alternative specifications of θ will be discussed below.

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

$$\widehat{E} = \theta^{-1}\left(\frac{C}{F+C}\right) < (1-\theta(0))W + \theta(0)A = U_0 \tag{7}$$

then the victim spends $\widehat{E}$ on deterrence and the criminal does not attack. (b) If $W > C$ and $\widehat{E} > U_0$ then the victim does not spend on deterrence, the criminal will attack and, if the attack is succesful, 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,

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

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

$$\theta\left(\widehat{E}\right) = \frac{C}{F + C}.$$

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

$$U(E) = -(1 - \theta(E))\, W - \theta(E)\, A - E = -W + (W - A)\, \theta(E) - E.$$

Note that

$$\frac{dU(E)}{dE} = (W - A)\, \frac{d\theta(E)}{dE} - 1. \tag{8}$$

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

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.[12] Clearly, this incentivizes the criminal to infect the computer. This, however, incentivizes the victim to deter an attack. A second crucial thing is, therefore, the cost of deterring attack. If that cost is not too high (where high is determined by equation (7)) 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 a ransomware context the value of F will likely be very small given the low marginal costs of a criminal, say, sending out malware to an email address. 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 (see equation (8) in the proof of Theorem 2). 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 $E \leq 1$ and $\theta(E) = 1$ for $E > 1$. Also, set $A = 0$ 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 -1. Or she could spend $E < 1$ on deterrence, face the potential of being attacked, and have expected payoff $U(E) = -W(1 - \theta(E)) - E$. 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

[12] In a hostage taking setting there is a credible incentive to not pay the ransom 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 attack from criminals.

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 likely if there are multiple approaches to deterrence. For example, we might find someone who buys anti-virus protection but is lazy when opening email attachments or we might find someone who does not buy anti-virus but is cautious opening attachments. This kind of approach may be optimal even if it does not completely immunize from attack.

4.2 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 to pay? This brings us back to the issue of imperfect information that we discussed in Section 3.3. 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 the case in which there are two types of victim: a low type willing to pay W_L to recover her files and a high type willing to pay $W_H > W_L$ to recover her files. Suppose that criminal sets a ransom targeted at the high type victim. In other words the ransom is set at $W_L < C \leq W_H$. Theorem 2 can be applied to discern what the high type will do. If

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

then the high type will spend $E_H = \widehat{E}$ on deterrence. This will deter all attack 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

$$\widehat{E} > (1-\theta(0))\,W_H + \theta(0)\,A$$

then the high type will spend $E_H = 0$ on deterrence. This leaves both the high type and low type open to attack. An interesting question is whether this incentivizes the low type to spend deterring attack? This is not as obvious as it may seem because the high type stands to 'only' lose the ransom while the low type stands to 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.

4.3 Spillover effects of deterrence

In the preceding sub-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 judgment on who is better or worse.

In the deterrence game spending reduces the probability of an attack being successful. This modeling assumption natural 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 (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 that 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. (2017). Note that this is different to the setting originally considered by Lapan and Sandler (1988) 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. The complicates attempts to combat ransomware.

5 Discussion and other literature

In this section 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 analyzed in this paper together with that of other closely related work such as Hernandez-Castro et al. (2017) and Laszka et al. (2017). 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 that is ultimately unlikely to lead to any additional insight.

What we need to do is clearly isolate the contribution different models can make and the ways they can be extended. This seems especially apt given that game theoretic modeling of ransomware is in its infancy and much work remains to be done. The kidnapping game we analyzed in Section 3 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 Hernandez-Castro et al. 2017). The deterrence game analyzed in Section 4 and the model of Laszka et al. (2017) simply assume away these issues: They take as given a ransom demand and assume the files will be returned. The crucial thing to observe is that these assumptions are supported by our analysis (including Theorem 1). For instance, we show that it is in the interests of the criminals to return the files to the victim. The models, therefore, nicely mesh together.

The deterrence game we analyzed in Section 4 primarily informs on the incentives of potential victims to deter attack. The model of Laszka et al. (2017) informs on the incentives to do back ups in order to mitigate the losses from attack. A key insight that comes out of both models are the spillovers

between 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 back up 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 rationally individuals (Baddeley 2011, Pfleeger and Caputo 2012).

There is also more we can learn from the game theoretic literature on kidnapping. From the perspective of the criminal the main issue (if we set aside the more technical aspects of launching successful attacks) is to maximize ransom revenue. As we discussed in Section 3 this is best achieved by the criminal 'tying his hands'. First, the optimal ransom demand needs to be determined (Hernandez-Castro, Cartwright and Stepanova 2017). Then, it is in the criminals 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'.[13]

The key thing to appreciate here is the criminal's need to build a reputation of being tough but fair.[14] 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 'honest' set of rules for customers (F-Secure 2016). 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. Building a reputation may, however, 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'.

[13]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 (2013), for instance, finds that democratic governments are more likely to pay terrorist ransom demands before elections, as the threat to not pay becomes non-credible.

[14]Wilson (2000) provides evidence that this type of approach is also successful in terrorist hostage taking scenarios.

For potential victims the picture is less bright. The spillovers between individuals mean that it is very difficult for anyone individual to 'win'. Spending on deterrence, particularly in terms of regular back-ups, 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 for the criminal of 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 (Gaibulloev and Sandler 2009). 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 victims sense of urgency (Hadlington 2017). Another thing that can work to the criminals advantage is the lack of attention a particular attack brings. For instance, Gaibulloev and Sandler (2009) and Sandifort and Sandler (2013) 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 the mass media will be more successful because the resources to help and advise victims are going to be less readily available.

6 Conclusion

In this paper we have applied and adapted two seminal models from the game theoretic literature on kidnapping to the issue of ransomware. The first model (due to Selten 1988) informs on the bargaining process between criminal and victim. The second model (adapted from Lapan and Sandler 1988) 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. Even so, our analysis has yielded some key findings, which we summarize below. We expect these findings are robust to more general analysis.

1. The optimal ransom demand is increasing in the willingness of the

victim to pay to recover her files. This means it is in the criminal's interest to be informed as possible about the victim's willingness to pay.

2. 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.

3. 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 honoring ransom payments.

4. 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.

5. 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 it may be optimal to subsidize spending on cyber-security or good back-up practices.

6. Deterrence is costly. Any estimate of the costs of ransomware should, therefore, take account of 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 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 much 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 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.

References

Acre, D. & T. Sandler, T. (2005). Counterterrorism a game-theoretic analysis. Journal of conflict resolution, 49(2), 183-200.

Anderton, C. H., & Carter, J. R. (2005). On rational choice theory and the study of terrorism. Defence and Peace Economics, 16(4), 275-282.

August T, Dao D, Laube S, & Niculescu M (2017) Economics of ransomware attacks, Workshop on Information Systems and Economics (WISE).

Baddeley, M. (2011). Information security: Lessons from behavioural economics. Workshop on the Economics of Information Security.

Barlyn, B. & C. Cohn (2017). Companies use kidnap insurance to guard against ransomware attackshttps. Reuters. www.reuters.com/article/us-cyber-attack-insurance/companies-use-kidnap-insurance-to-guard-against-ransomware-attacks-idUSKCN18F1LU

Brandt, P. T., George, J., & Sandler, T. (2016). Why concessions should not be made to terrorist kidnappers. *European Journal of Political Economy*, 44, 41-52.

Camerer, C. (2003). Behavioral game theory: Experiments in strategic interaction. Princeton University Press.

Cohen, D. S. & X. Dormandy (2012). Kidnapping for ransom: The growing terrorist financing challenge. Chatham House publication.

Corsi, J. R. (1981). Terrorism as a Desperate Game Fear, Bargaining, and Communication in the Terrorist Event. Journal of Conflict Resolution, 25(1), 47-85.

Danielson, D. (2015) The FBI says you may need to pay up if hackers infect your computer with ransomware, *BusinessInsider*. http://uk.businessinsider.com/fbi-recommends-paying-ransom-for-infected-computer-2015-10

Dutton, Y. M., & Bellish, J. (2014). Refusing to Negotiate: Analyzing the Legality and Practicality of a Piracy Ransom Ban. *Cornell International Law Journal*, 47, 299.

Enders, W., & Sandler, T. (1995). Terrorism: Theory and applications. Handbook of defense economics, 1, 213-249.

F-Secure (2018) *The changing state of ransomware*. https://fsecurepressglobal.files.wordpress.com/2018/05/ransomware_report.pdf

F-Secure (2017) *State of Cyber Security in 2017*. https://www.f-secure.com/documents/996508/1030743/cyber-security-report-2017

F-Secure (2016) Evaluating the customer journey of crypto-ransomware. https://fsecureconsumer.fi ransomware_f-secure.pdf

Finnis, J., Boyle Jr, J. M., & Grisez, G. (1987). Nuclear deterrence, morality, and realism.

Fudenberg, D., & Tirole, J. (1991). Game Theory. MIT Press.

Gaibulloev, K., & Sandler, T. (2009). Hostage taking: Determinants of terrorist logistical and negotiation success. *Journal of Peace Research*, 46(6), 739-756.

Hadlington, L (2017). Exploring the Psychological Mechanisms used in Ransomware Splash Screens. SentinelOne report.

Hernandez-Castro J & Boiten E.(2016) 2016 Kent Cyber Security Survey. University of Kent. https://cyber.kent.ac.uk/Survey2016.pdf

Hernandez-Castro, J., Cartwright, E., & Stepanova, A. (2017). Economic Analysis of Ransomware. arXiv preprint arXiv:1703.06660.

Huang, K., Siegel, M., & Madnick, S. (2017). Cybercrime-as-a-Service: Identifying Control Points to Disrupt.

Intermedia (2017) Data Vulnerability Report. https://www.intermedia.net/report/datavulnerabilit part2

Iqbal, A., Masson, V., & Abbott, D. (2017). Kidnapping model: an extension of Selten's game. *Royal Society open science*, 4(12), 171484.

Jarvis K. (2013) Cryptolocker ransomware. SecureWorks Counter Threat Unit. http://www. secureworks.com/cyber-threat-intelligence/threats/cryptolocker-ransomware.

Kalaimannan E, John SK, DuBose T, Pinto A. (2017) Influences on ransomware's evolution and predictions for the future challenges. *Journal of Cyber Security Technology*, 1: 23-31. http://dx.doi.org/10.1080/23742917.2016.1252191

Keeney, R. L. (2007). Modeling Values for Anti-Terrorism Analysis. Risk Analysis, 27(3), 585-596.

Kharraz A, Robertson W, Balzarotti D, Bilge L, Kirda E. (2015) Cutting the gordian knot: A look under the hood of ransomware attacks. *International Conference on Detection of Intrusions and Malware, and Vulnerability Assessment*, 3-24.

Lapan, H. E., & Sandler, T. (1993). Terrorism and signalling. European Journal of Political Economy, 9(3), 383-397.

Lapan, H. E., & Sandler, T. (1988). To bargain or not to bargain: That is the question. The American Economic Review, 78(2), 16-21.

Laszka, A., Farhang, S., & Grossklags, J. (2017). On the Economics of Ransomware. In International Conference on Decision and Game Theory for Security (pp. 397-417). Springer.

Lee, C. Y. (2013). Democracy, civil liberties, and hostage-taking terrorism. *Journal of Peace Research*, 50(2), 235-248.

Liao K, Zhao Z, Doupé A, Ahn GJ. (2016) Behind closed doors: measurement and analysis of CryptoLocker ransoms in Bitcoin. *Electronic Crime Research 2016 APWG Symposium IEEE*. https://doi.org/10.1109/ECRIME.2016.7487938.

Mansfield-Devine S. (2016). Ransomware: taking businesses hostage. *Network Security*, 8-17. https://doi.org/10.1016/S1353-4858(16)30096-4.

Muthoo, A. (1999). *Bargaining theory with applications*. Cambridge University Press.

Nax, H. H. (2008). Modeling hostage-taking: On reputation and strategic rationality of terrorists. Studies in Conflict & Terrorism, 31(2), 158-168.

O'Neill, B. (1992) Game theory models of peace and war. In Aumann, R. J., & Hart, S. *Handbook of game theory with economic applications* (Vol. 2). Elsevier.

Pfleeger, S. L., & Caputo, D. D. (2012). Leveraging behavioral science to mitigate cyber security risk. *Computers & security*, 31(4), 597-611.

Rashid, F. (2016) 4 reasons not to pay up in a ransomware attack. InfoWorld. https://www.infoworld.com/article/3043197/security/4-reasons-not-to-pay-up-in-a-ransomware-attack.html

Rosendorff, B. P., & Sandler, T. (2005). The political economy of transnational terrorism. The Journal of Conflict Resolution, 49(2), 171-182.

Sandler, T. (2014). The analytical study of terrorism: Taking stock. *Journal of Peace Research*, 51(2), 257-271.

Sandler, T., & Arce, D. G. (2007). Terrorism: a game-theoretic approach. Handbook of Defence economics, 2, 775-813.

Sandler, T., & Enders, W. (2007). Applying Analytical Methods to Study Terrorism. International Studies Perspectives, 8(3), 287-302.

Sandler, T., & Enders, W. (2004). An economic perspective on transnational terrorism. *European Journal of Political Economy*, 20(2), 301-316. Chicago

Sandler, T., & Gaibulloev, K. (2009). Hostage taking: determinants of terrorist logistical and negotiation success. Journal of Peace Research 46: 739-756.

Sandler, T., & Hartley, K. (Eds.). (2007). Handbook of Defense Economics: Defense in a globalized world (Vol. 2). Elsevier.

Santifort, C., & Sandler, T. (2013). Terrorist success in hostage-taking missions: 1978–2010. *Public Choice*, 156(1-2), 125-137.

Schelling, T. C. (1980). The strategy of conflict. Harvard university press.

Schneider, F., Brück, T., & Meierrieks, D. (2015). The economics of counterterrorism: A survey. Journal of Economic Surveys, 29(1), 131-157.

Scott, J. L. (1991). Reputation building in hostage taking incidents. Defence and Peace Economics, 2(3), 209-218.

Selten, R. (1988). A simple game model of kidnapping. In Models of strategic rationality (pp. 77-93). Springer Netherlands.

Shahin, W. N., & Islam, M. Q. (1992). Combating political hostage-taking: An alternative approach. Defence and Peace Economics, 3(4), 321-327.

Spagnuolo M, Maggi F & Zanero S. (2014) Bitiodine: Extracting intelligence from the bitcoin network. *International Conference on Financial Cryptography and Data Security*, 457-468. https://doi.org/10.1007/978-3-662-45472-5_29.

Symantec.(2016) Ransomware and Business 2016. Symantec Corporation. http://www.symantec.com/content/en/us/enterprise/media/security_response/whitepapers/ISTR

Volpicelli (2017) Taking on the Fancy Bear hackers: how to negotiate if your data is being held ransom. Wired Magazine. http://www.wired.co.uk/article/negotiate-hackers-moty-cristal.

Wilson, M. A. (2000). Toward a model of terrorist behavior in hostage-taking incidents. Journal of Conflict Resolution, 44(4), 403-424.

Young, A., & Yung, M. (1996). Cryptovirology: Extortion-based security threats and countermeasures. *Security and Privacy Proceedings IEEE Symposium*. IEEE.