Coping with Digital Extortion: An Experimental Study on Normative Appeals

Kay-Yut Chen[a], Jingguo Wang[a], Yan Lang[a]

[a] Information Systems and Operations Management, University of Texas at Arlington

Abstract

Digital extortion emerges a significant threat to organizations that rely on information technologies for their business and operations. We study, with human-subject experimentation, how normative appeals may influence defenders' engagement of investing in security and refusal to pay ransoms as mitigating strategies to this digital extortion threat. We explore the effects of four types of normative appeals: injunctive norms and descriptive norms promoting investing or not-paying ransoms. We find that the defenders' decisions deviate from the predictions of game theory. However, given the strategic interactions between the defenders and the attacker as well as noisy decision-making behaviors, it is challenging to untangle the influence of the treatment interventions on the defenders. We develop a structural model using the quantal response equilibrium framework to determine how normative appeals change the defenders' utilities of investing and not-paying. While interventions may be successful in increasing the utilities of investing and/or not-paying, their impacts are mitigated by the attacker reducing ransoms. Thus, it is challenging for an intervention to significantly boost a community's investment rate or to suppress ransom payment rate. Based on the model, we characterize how security outcomes of a community (including expected ransoms, attack rate, investment rate, payment rate) change with the defenders' utilities of investing and not-paying. The results to two new interventions, a penalty for paying ransoms and the ability for defenders to communicate via text chat, further validate the modeling results.

Keywords: Information Security, Behavioral Economics Experiments, Game Theory, Quantal Response Equilibrium, Normative Appeals, Injunctive Norms, Descriptive Norms

Coping with Digital Extortion: An Experimental Study on Normative Appeals

1. Introduction

Digital extortion refers to the criminal act of compelling an organization to pay a ransom for saving systems from being wiped out or gaining back access to seized data[1] (Sancho, 2017, Thakkar, 2017). Ransomware has become one of the most popular attack vectors for digital extortion (Crowdstrike, 2020, Verizon, 2020). Its campaigns grow on the basis of high money-making potential and low chance of arrest (Thakkar, 2017). A Kaspersky security bulletin (Kaspersky, 2016) suggests every 40 seconds there is an organization attacked with ransomware in the third quarter of 2016. According to the 605 companies surveyed by Radware in 2017 (Radware, 2018), approximately 42% of them experienced ransomware attacks, 40% more than in 2016. The severity of the threat is well demonstrated by WannaCry, affecting more than 230,000 computers in over 150 countries in May 2017 (Ehrenfeld, 2017). More recently, it is widely reported that the University of California at San Francisco paid $1.14 million to Netwalker ransomware attackers to recover research data for their school of medicine.

Two strategies may be available to "defenders" (i.e., any organization that can be a potential victim) for coping with digital extortion. The first is investing in information security, such as introducing intrusion prevention systems, encrypting devices, and properly backing up data (Brewer, 2016), to reduce the chance of being exploited. Unique to digital extortion, another form of mitigation may also be available. That is, if all defenders can commit to refuse to pay ransoms, cybercriminals (referred to as "attackers" in the rest of the paper) will have no incentive to carry out the attacks in the first place. However, both mitigation strategies may have incentive issues. Security investments can be costly, and the impacts of such investment may be difficult to foresee (Anderson and Moore, 2006). Although it is publicly advised at news media (e.g., Mathews, 2018) and advocated by government agencies (FBI, 2016), refusal to pay ransoms may be problematic for defenders given the risk of losing their seized assets and resulting in business disruption (Everett, 2016, Liska and Gallo, 2017). Defenders' investment and refusal to pay reduce the threat by changing the incentives of the attackers. Their benefits are endogenous contingent on attackers' decisions.

[1] While ransomware is a consumer problem as much as a business problem, this study focuses on businesses and organizations.

One potential solution is to strengthen social norms to encourage good behavior on the part of the defenders. Social norms refer to “the rules and standards that are understood by members of a group, and that guide and/ or constrain social behavior without the force of laws” (Cialdini and Trost, 1998, p152). Following social norms is often considered to be adaptive especially when decision makers face uncertainty. Social norms may help decision makers to gain a good understanding of social situations and enable them to have effective responses (Cialdini, 2001). Researchers have argued that normative information encourages a person to behave more securely. For example, Yazdanmehr and Wang (2016) showed that employees’ perception about others’ opinions on or behavior of complying with information security policies in an organization is positively associated with their compliance intention. Based on retrospective interviews, Das et al. (2014) suggested that social influence raises decision makers’ security sensitivity, and plays an important role in changing a range of privacy and security-related behavior. In the context of adopting security features among Facebook users, simply showing people the number of their friends that used security features drives more people to explore and adopt such features (Das et al., 2015, Das et al., 2014).

Along this line of thought, we design the study to explore the potential of social norms as solutions to the digital extortion problem. We investigate whether good behaviors (i.e. investing in security and refusing to pay ransoms) in fighting with digital extortion are susceptible to normative appeals for investing and not-paying ransoms. Given the complexity of the strategic interactions in a digital extortion setting, and the multiple incentives and decisions in play, the goals of the paper are modest and in three fold: (1) Can we show normative appeals, more specifically, injunctive appeals (“ought to”) and descriptive appeals (“what others are doing”), are effective at nudging the defenders in investing and refusing to pay? (2) Given the possible strategic interactions and noisy decision-making behaviors (which are a common feature in observed human decisions), how do we quantify the effect of treatments on the defenders’ utilities of investing and not-paying? (3) What are the relationships between the defenders’ utilities of investing or not-paying and security outcomes of a community (including expected ransoms, attack rate, investment rate, and payment rate)? It is out of the scope of this paper to find an optimized best strategy against ransomware attack although the results reported in this paper is a first step towards building a strategy.

To address these questions, we employ a combination of game theory analyses, human-subject experimentation, behavioral economics modeling, and numerical analyses. We first establish a baseline for attacker-defender interactions employing game theoretic analyses. However, it is well known that game theory does not always capture realistic human decision-making behaviors. Field studies will involve too many confounding factors, and it would be challenging to infer causality and to measure social norms in a rigorous manner. Thus, human subject experimentation is the most promising approach. This is, indeed, in line with the current trend of using economic experiments in management science and operations management studies (please see Donohue et al. (2018) for a review of this literature). Economic experiments can be a valuable tool, but currently are underused in the field of Information Systems. Economic experiments can bridge the gap between the rational economic models and the capabilities of human decision making, and provide an alternative mechanism to develop insights overcoming some limitations in analytical modeling and/or empirical analysis of secondary data (Gupta et al., 2018).

Game theory analyses provide the guidance for the calibration of experimental parameters. With human-subject experiments, we examine four types of normative appeals, including injunctive appeals and descriptive appeals encouraging investing and refusing to pay. We find that the attackers strategically respond to the treatments by lowering their ransoms. It is difficult for the standard statistical analyses to show the isolated effects of the treatments on motivating the defenders. Therefore, we develop a structural model incorporating bounded rationality based on the quantal response equilibrium framework (QRE) (Mckelvey and Palfrey, 1995, Mckelvey and Palfrey, 1998), parametrizing the impacts of a treatment directly into the defenders' utility functions. The goal of the behavioral model is to quantify the impacts of normative appeals on the defenders' utilities of investing and not-paying, considering the players' strategic interactions and noisy decision-making behaviors. We confirm that normative appeals indeed drive the defenders to the desirable directions. While interventions may be successful in increasing the utility of investing or that of not-paying ransoms, their impacts can be mitigated with the attacker reducing ransoms or overshadowed by noisy decision-making behaviors. Thus, it is challenging for an intervention to significantly boost investment rate and lower payment rate.

Relying on numerical analyses, we characterize how security outcomes of a community (including expected ransoms, attack rate, investment rate, and payment rate) change with the defenders' utilities of investing and not-paying. Our results reveal that as the utility of investing increases, investment rate is likely to increase considerably, but there is almost no change in payment rate. As the utility of not-paying increases, both investment rate and payment rate decrease slightly. The attacker is more likely to decrease ransoms for a higher not-paying utility than for a higher investing utility, and more likely to lower their attack rate for a higher investing utility than and for a higher not-paying utility.

The value of the behavior model is further demonstrated with the test of two alternative interventions. These include an incentive manipulation with a penalty for ransom payments, and a social interaction manipulation with chat. Both interventions nudge the defenders to invest and refuse to pay at certain degrees, but fail to boost investment rate and suppress payment rate significantly, as the behavioral model suggests. The study suggests potential approaches, but also identifies challenges, and offers a modeling framework to evaluate interventions, for fighting against digital extortion for policy makers.

This paper is organized as follows. In §2, we summarize the related theoretical and empirical literature in information security investment. §3 provides the details of our model setting and game theory predictions. §4 details of our experimental procedures and hypotheses. We build behavior models to explain our experimental results in §5. §6 summarizes alternative interventions. Finally, we conclude the paper with a discussion of research and managerial implications, limitations, and future extensions in §7.

2. Literature Review

Prior studies that investigate how organizations optimize their security spending employ two main approaches: traditional risk/decision and game theoretical analysis (Cavusoglu et al., 2008, Wang et al., 2008). The traditional risk/decision analysis models treat security risk as exogenous environmental threats and apply an approach of optimal-control theoretic certainty equivalence. For example, Gordon and Loeb (2002) examine investment in information security, under different security breach functions and different levels of vulnerability, as a decision optimization problem where a firm maximizes the expected benefit of such investment. Wang et al. (2008) apply the notion of Value at Risk and extreme value theory to security investment decisions, incorporating risk preferences of decision makers into the model. Yet such approaches do not consider strategic interactions between defenders and attackers, and in particular, how security spending would influence the attackers' behavior.

Another stream of literature focuses on game theoretic analysis where information security is considered as a public goods to capture its across-firm/defender interdependent nature (Anderson and Moore, 2006). However, the externalities of firms' effort levels are typically controlled by exogenous parameters (Heal and Kunreuther, 2007, Kunreuther and Heal, 2003). This type of setting, with firms making independent decisions of their security investments, often results in incentive misalignment problems where system optimality, in investment levels of firms, is not achieved (Kunreuther and Heal, 2003). Varian (2004) examines a case of this type of free rider problem. Gal-Or and Ghose (2005) investigate economic incentives for sharing security information, and find that security investments and security information sharing act as strategic complements in equilibrium. Zhao et al. (2013) explore how cyber-insurance can help firms optimize security spending given security interdependence.

Prior analytical studies provided important insights to security investment and resource spending. However, in most of these studies, attackers are not a part of the setting. Their decisions on whether to attack and what to attack, characterized by risk functions, are modeled as exogenous. One such exception in the literature is Cremonini and Dmitri (2009) where a game theoretical model is used to argue that rational attackers are likely to direct their effort toward less-protected targets, and by signaling their security characteristics the defenders may improve their welfare. Another paper, Kannan et al. (2016) explore how attackers' strategic reactions to patching influence a vendor's pricing and software maintenance decisions. But none of these studies investigate digital extortion. In addition, prior analytical modeling studies assume that defenders and attackers make rational choices given available information. Research in behavioral economics have well documented that the rationality is bounded (Simon, 1947), and decision makers suffer a range of cognitive biases when making choices under uncertainty (Tversky and Kahneman, 1974).

Digital extortion emerges as a new type of threats to organizations. Such attacks complicate the defenders' problem further by presenting an additional option, that is paying the attacker a ransom to eliminate the threat and/or redeem the seized assets, besides relying on security investment to reduce the chance of being exploited (Sancho, 2017, Thakkar, 2017). Prior literature has barely explored how the defenders make the decisions of security investment and ransom payment responding to digital extortion. Most studies in ransomware and digital extortion take a technical and descriptive perspective (Gazet, 2010, Kharaz et al., 2016, Kharraz et al., 2015, Luo and Liao, 2007, Scaife et al., 2016, Sittig and Singh, 2016), illustrating the techniques employed by such attacks, developing detection approaches, and suggesting best practices for organizations. How defenders and hackers strategically interact and respond to each other in the setting of digital extortion is far from well-understood.

Since the interventions we investigate are drawn from normative appeals, the last relevant stream of literature is about social norms. Social norms have been shown to be important in changing a range of human behaviors (Cialdini, 2001, Cialdini and Trost, 1998). Studies in information systems examine the power of social norms in a variety of settings (e.g., stimulating online reviews (Burtch et al., 2018), contributing in virtual communities (Tsai and Bagozzi, 2014), motivating goal setting and goal attainment for physical activity (Liu et al., 2019)). Social norms are also recognized to be an important factor driving individual security behavior, including policy compliance intention (Bulgurcu et al., 2010, Herath and Rao, 2009, Yazdanmehr and Wang, 2016), resources misuse intention (Chu et al., 2015), personal computer and internet protection intention (Anderson and Agarwal, 2010). As most studies in information security examine the impact of social norms rely on survey and self-reported data, social norms are often measured based on what a respondent believes that most others are doing so and what a respondent believes that influential others expect him or her to do.

3. Model Setting and Game Theoretical Prediction

Following the design philosophy of parsimony, we use the simplest model that captures the endogenous public goods nature of security and the ability of the attacker to choose the weaker defender for exploitation (Cremonini and Dmitri, 2009, Zhao et al., 2013). Hence, the minimum number of players needed is three, with one attacker and two defenders. To focus on the strategic interactions of the attacker and the defenders only, we make a series of assumptions consistent with that goal. We assume the defenders are symmetrical to ensure the choice of whom to attack is purely based on strategic consideration and not environmental asymmetry. We assume the defenders' investment decisions are binary choices. That is, they decide either to invest or not-to-invest with a constant cost and a reduction of probability of being successfully attacked. The values of data (i.e. the object of extortion) are known. These variables (i.e. investment costs, success-attack-probabilities, data values) are all common knowledge to all three players. Figure 1 describes the chronological order of events in a round of the digital extortion game.

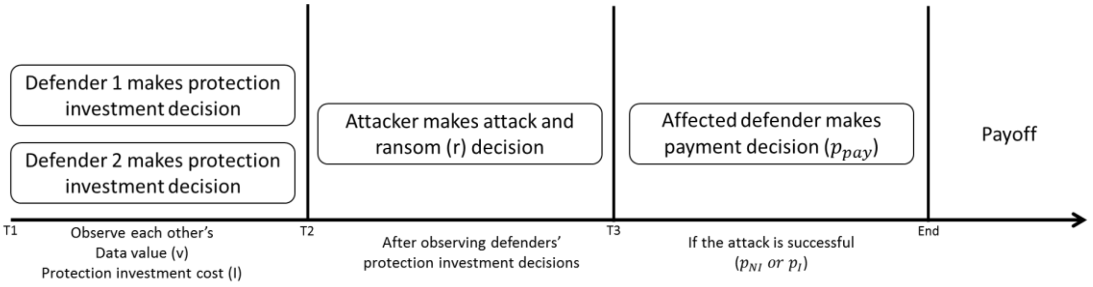

Defender 1 makes protection investment decision
Defender 2 makes protection investment decision
Attacker makes attack and ransom (r) decision
Affected defender makes payment decision (p_{pay})
Payoff
T1
Observe each other's
Data value (v)
Protection investment cost (I)
T2
After observing defenders'
protection investment decisions
T3
If the attack is successful
(p_{NI} or p_I)
End

Figure 1. Order of Events of Digital Extortion Game

We use x with a subscript as the decision variable. Three are discrete variables, including the defender j's investment decision (x_{ij}), her ransom payment decision (x_{pj}), and the attacker's attack decision (x_a). One is continuous that is the attacker's ransom price (x_r). In the first stage (T1), the defenders decide whether to invest in security. $x_{ij} = 1$ if the defender j ($j = 1 \ or \ 2$) decides to invest, 0 otherwise. The cost of security investment is I. The data value to be protected by a defender is v. The probability of a successful attack (or being compromised for a defender) is p_{NI} without investment and p_I with investment. We assume, without the loss of generality, $0 < p_I < p_{NI}$.

After observing the defenders' investment decisions, the attacker chooses whether to attack a defender (x_a) and ask for a ransom (x_r) in the second stage (T2). $x_a = j$ if the attacker decides to attack the defender j. $x_a = 0$ if the attacker decides not to attack, and, in this case, he receives a fixed outside option payment (c) as the payoff. This outside option is also common knowledge to all players.

In the third stage (T3), if the attack is unsuccessful, the attacker receives nothing. The defenders receive their data value v minus an appropriate investment cost, I or 0 contingent on the defenders' investment decision (x_{ij}). If the attack is successful, the affected defender chooses whether to pay the ransom (x_{pj}). $x_{pj} = 1$ if the defender j decides to pay, 0 otherwise. If the defender decides to pay, the attacker receives the ransom as the payoff. The defender receives her data value v minus an appropriate investment cost and the ransom price (x_{pj}). If the defender decides not to pay, both the defender and the attacker receive nothing. The unaffected defender receives her data value v minus an appropriate investment cost.

The Nash equilibrium can be found by backward induction. If a defender is successfully attacked, she will pay any ransom x_r up to the data value v. Hence, the attacker will choose a ransom, $x_r = v$, anticipating if the attack is successful, the defender will prefer, weakly, to pay.

Given the last stage solution, the expected payoff for the attacker is:

$$u_a = \begin{cases} p_I v & x_a = j \mid x_{ij} = 1 \\ p_{NI} v \quad \text{if} & x_a = j \mid x_{ij} = 0 \\ c & x_a = 0 \end{cases}$$

Knowing that, the attacker will only attack if his expected payoff, from attacking, is bigger than his outside option c. Depending on the values of v, p_I, p_{NI} and c, and the strategies chosen by the two defenders, the attacker's equilibrium changes. Table 1 presents all the possible cases.

Table 1. All Possible Cases of Attack Decisions

Defender 1 strategy x_{i1}	Defender 2 strategy x_{i2}	Parameter conditions	Attack decision x_a
1	1	$p_I v > c$	1 or 2 with ½ chance
1	0	$p_I v > c$	2
0	1	$p_I v > c$	1
0	0	$p_{NI} v > c$	1 or 2 with ½ chance
1	1	$p_I v \leq c$	0
1	0	$p_{NI} v \leq c$	0
0	1	$p_{NI} v \leq c$	0
0	0	$p_{NI} v \leq c$	0

Summarizing Table 1, there are 3 relevant parametric conditions. Figure 2 illustrates the attacker's equilibrium strategy in a more intuitive manner. If the outside option c is higher than $p_{NI}v$, then the attacker has no incentive to attack, no matter what the defenders do. If the outside option c is lower than $p_I v$, then the attacker is better off attacking no matter what the defenders do. The attacker clearly will choose the defender who does not invest if the other defender invests. In the case where the two defenders follow the same strategy, the attacker will randomly choose one of the defenders to attack. If the outside option c is in-between $p_I v$ and $p_{NI} v$, then the attacker will only attack if at least one of the defenders decides not to invest. Next, we apply backward induction once more time to obtain the equilibrium for the first stage (T1). Based on the analysis of the attack decisions, there are three cases.

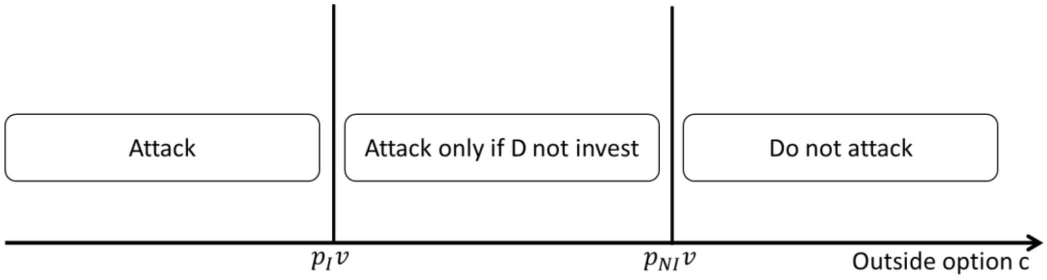


Figure 2. The Attacker's Equilibrium Strategy

3.1. Case 1: $p_{NI}v < c$

This is a trivial case where the attacker will not attack independent of what the defenders do. Hence the defenders will have no incentive to invest, and the equilibrium is that defenders do not invest, and the attacker does not attack.

3.2. Case 2: $p_I v < c \leq p_{NI}v$

In this case, the attack will only attack a defender who does not invest. Hence, the first stage (T1) investment stage game has the normal form payoff matrix for the defenders shown in Table 2. It is straightforward to verify that if $(I/v) \leq p_{NI}$, (invest, invest) will be an equilibrium. If $(I/v) \geq p_{NI}/2$, (not invest, no invest) will be an equilibrium. We also prove that there is no asymmetric equilibrium (see Appendix A). The intuition is straight forward. If the investment cost I is low enough, both defenders investing is an equilibrium. If the investment cost is too high, none of the defenders investing is an equilibrium. Note that it is possible, when $p_{NI}/2 \leq (I/v) \leq p_{NI}$, there are two symmetric equilibriums (both defenders investing/both defenders not investing). Figure 3 illustrates the structure of the equilibriums. Case 2 is also the most interesting because the attacker may not attack contingent on what the defenders do. Hence, we design the experiments around this case.

Table 2. Payoff for the Defenders in Case 2 ($p_I v < c \leq p_{NI}v$)

D1 | D2	Invest	Not Invest
Invest	$v - I, v - I$	$v - I, v(1 - p_{NI})$
Not Invest	$v(1 - p_{NI}), v - I$	$v(1 - 1/2p_{NI}), v(1 - 1/2p_{NI})$

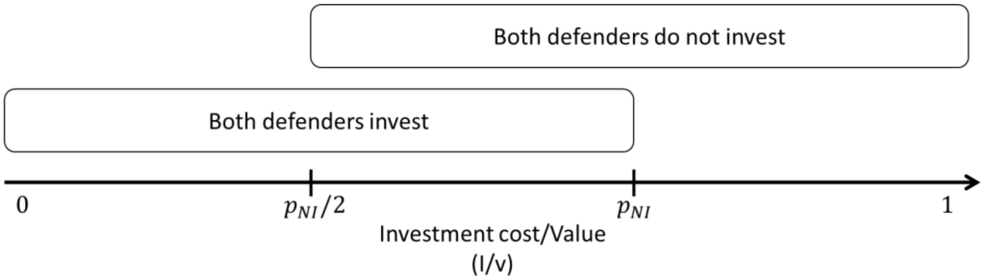


Figure 3. Structure of the Equilibriums for the Defenders

3.3. Case 3: $c \leq p_I v$

In this case, since the outside option for the attacker is low, he always attacks. Hence, the first stage (T1) investment stage game has the normal form payoff matrix for the defenders shown in Table 3. Note that the payoff matrix is almost identical to that for case 2, except in the (invest,

invest) cell. The difference is that in the case of (invest, invest), the attacker attacks, as opposed to not, in case 2.

Again, there is no asymmetric equilibrium (see Appendix A). The condition for (not invest, not invest) to be an equilibrium is the same as in case 2. That is, $(I/v) \geq p_{NI}/2$. However, the condition for (invest, invest) to be an equilibrium and changes to $(I/v) \leq p_{NI} - p_I/2$. Compared to case 2, the threshold is lower because the attacker will attack in this case and so the value of investment is lower. Since $p_I < p_{NI}$, it is guaranteed that there will be a range of I/v where both (invest, invest) and (not invest, not invest) are equilibriums.

Table 3. Payoff for the Defenders in Case 3 ($c \leq p_I v$)

D1 | D2	Invest	Not Invest
Invest	$v(1-1/2p_I) - I, v(1-1/2p_I) - I$	$v - I, v(1-p_{NI})$
Not Invest	$v(1-p_{NI}), v - I$	$v(1-1/2p_{NI}), v(1-1/2p_{NI})$

4. Experimental Study

In this section, we report an experimental study that investigates if and how normative appeals can nudge the defenders and mitigate digital extortion attacks.

4.1. Experimental Design and Setting Calibration

The model setting is characterized by five parameters (value v, investment cost I, probability of being compromised with/without investment p_I/p_{NI}, and the attacker's outside option c). In a typical test-the-theory study, the design would employ multiple sets of parameters and see if behaviors respond, in the predicted manner, to parametric changes. However, the main goal of this paper is to study interventions under strategic interactions in the setting of digital extortion. Hence, we opt for a different approach.

We use the game theoretic analysis above as guidance to pick *one* parametric setting where the incentives are representative and deemed close to the real environment. One important parameter given in the game theoretical analysis above is the outside option c of the attacker. We consider that it would be too restrictive to pick an outside option c either too high (resulting in no attacking) or too low (always attacking regardless of the defenders' investment). Rather, it is more realistic that the attacker shall balance the decision between attack and not attack (Laszka et al., 2017), and the investment reduces his incentive to attack. Thus, we pick parameters satisfying $p_I v < c \leq p_{NI} v$, corresponding to case 2 in the game theoretic analysis. The game has two symmetric equilibriums under case 2. In most realistic scenarios, there is a need for investing in security, given the trade-offs between potential losses and investment costs. Hence, we decide to choose a parameter setting where (invest, invest) is the equilibrium. We also eliminate cases where there are multiple equilibriums as not to distract from the main focus of the paper. The condition satisfying these two requirements is $(I/v) \leq p_{NI}/2$.

There are obviously infinite sets of parameters satisfying $p_I v < c \leq p_{NI} v$ and $(I/v) < p_{NI}/2$. Without loss of generality, we arbitrarily set v to 100. We pick p_{NI} and p_I to be 0.8 and 0.3, again somewhat arbitrarily but with enough separation that any reasonable individual will agree that investment reduces the probability of a successful attack "substantially". We pick $c = 40$ and $I = 30$ satisfying the conditions with enough room from the boundaries so that the incentives are clear. Table 4 summarizes our parameter choices.

Table 4. Calibration of Experimental Parameters

Parameter	Value in experiment
v: data value	100
I: investment cost	30
p_{NI}: prob of successful attack **without** investment	0.8
p_I: prob of successful attack **with** investment	0.3
c: attacker outside option	40

4.2. Baseline

In practice, the community of defenders (i.e. firms, organizations, institutions) does not drastically change over time and repeated interactions are realistic. Thus, we use a *repeated interaction* setting in our experiments. That is, the same group subjects in the roles of attacker and defenders play the game described above multiple times in an experimental session. We assume a scenario where participants are informed of past action of others and past security breaches. In practice, there are forums, such as conferences and trade-shows, where companies exchange information about their cyber security strategies, and governmental policy dictating the reporting of breaches. While clearly not everything a company does, regarding cybersecurity, is made public, we believe our experimental design captures the important aspects of the digital extortion scenario. It is certainly possible to investigate the role of information in this scenario, but that is beyond the scope of this paper. The subjects were also informed, in the beginning of the experiments, that they were going to play 30 rounds. Hence, technically, the setting is a finitely repeated game. By backward induction, the Nash equilibrium of the one-shot game is the Nash equilibrium of the repeated game.

As a result, the game theoretic analysis above applies. We refer to this setting, with parameters in Table 4, with 30 rounds of repeated interactions, as the *baseline* treatment.

4.3 Normative Appeals

Social norms are often categorized into two types: descriptive norms and injunctive norms (Cialdini et al., 1991, Cialdini et al., 1990). Descriptive norms and injunctive norms are both important motivators of human behavior. *Descriptive norms* inform decision makers about the popularity of certain behavior. This type of norms guides decision makers by providing information about how the majority of others behave. Because decision makers tend to reason "if a lot of people are doing this, it's probably a wise thing to do" (Cialdini, 2007), knowing the behavior of others helps decision makers to develop a heuristic for what might be the most effective and adaptive. In addition, imitating others' behavior saves one's cognitive effort and time.

Injunctive norms inform decision makers about social approval of certain behavior. This type of social norms guides decision makers by providing information about what ought to be done. It supplies a reference point for effective and likeable behavior. Decision makers may not want to deviate from injunctive norms but engage in norm-congruent behavior to produce liking, seek approval, and gain acceptance (Cialdini and Trost, 1998, Griskevicius et al., 2006).

Social norms could be a powerful tool to change people's behavior. Though social norms usually emerge and can be enforced via repeated interactions (Fehr and Gächter, 2002), interventions based on normative appeals in forms of communicating social norms with written appeals may be sufficient to guide people's behavior in desired ways. For example, normative appeals were found to be effective in encouraging consumers to engage in sustainable behaviors (White and Simpson, 2013), promoting energy conservation (Nolan et al., 2011), increasing social marketing (Metcalf et al., 2019). Normative appeals have also been applied in a series of studies that increase individual charitable contributions (Croson et al., 2009, Croson and Shang, 2008, Shang and Croson, 2009).

In this study we explore the efficacy of four types of normative appeals: injunctive appeals for investing, injunctive appeals for refusing to pay, descriptive appeals for investing, descriptive appeals for refusing to pay. In injunctive appeals for investing, we show a message "You should invest to reduce the chance of being successfully attacked" when defenders make investment decisions. In injunctive appeals for refusing to pay, we show a message "You should not pay the attacker to discourage him from attacking in the future" when the affected defender makes payment decisions. In descriptive appeals for investing, we show a message "In a previous session, defenders invested 73% of the time" when defenders make investment decisions. In descriptive appeals for refusing to pay, we show a message "In a previous session, defenders refused to pay the attacker 62% of the time, if successfully attacked" when the affected defender makes payment decisions. Appendix D illustrates the screenshots of the treatments. Please note that both percentages are calculated based on our prior sessions to avoid deception in experiments. We propose that:

> **H1**: Investment rate is higher with injunctive appeals for investing, compared to the *baseline* treatment.
>
> **H2**: Ransom payment rate is lower with injunctive appeals for refusing to pay, compared to the *baseline* treatment.
>
> **H3**: Investment rate is higher with descriptive appeals for investing, compared to the *baseline* treatment.
>
> **H4**: Ransom payment rate is lower with descriptive appeals for refusing to pay, compared to the *baseline* treatment.

4.5. Amazon Mechanical Turk Protocol

We employed standard experimental economics methodology and used no deception. The experiments were conducted on Amazon Mechanical Turk (i.e., MTurk), with the SoPHIE software (https://www.sophielabs.com). Lee et al. (2018) shows that supply chain experiments conducted on MTurk draw the same conclusions as laboratory-based experiments (except one case that cannot be replicated both in lab and on MTurk). In addition, MTurk provides a much more diverse set of subjects, compared to the typical experiments conducted on university campuses where undergraduate students are used as subjects. We chose MTurk as our experimental platform because of the ability to recruit more than just undergraduate students as subjects. We restricted our subjects' web portal geographic location to the United States and accepted only high-reputation workers (completed more than 100 MTurk tasks with at least 95% approval ratings) as our experiment subjects. Many prior MTurk research uses such sample restrictions to ensure high quality data (Hauser and Schwarz, 2016, Lee et al., 2018, Peer et al., 2014) .

After joining the experiment, participants were provided with written instructions. To ensure all subjects understand the experiment, they were required to pass a quiz, consisting of three questions about the rules of the game, before they were allowed to participate. Please see appendix C for the written instructions, and quiz questions. To ensure every participant has full information, in each stage we showed the data value, the investment cost, and the investment decisions to both the defenders and the attacker. In addition, at the payoff stage of each round, there was a summary table that showed all the prior decisions. Please refer to Appendix D for experiment screenshots. As standard in economics experiments, incentives are controlled by monetary payoffs. Subjects were paid according to their performances in the experimental sessions, and a small show-up fee. The average payment was $4.3, which is in-line with the earning rates on MTurk.

4.6. Experimental Results

Table 5 summarizes game theoretic predictions and decision outcomes under each treatment condition.

Table 5. Summary Statistics of Group Security Outcomes with Different Treatments

Decision	Game Theory	Baseline	Injunctive Should Invest	Injunctive Should Not Pay	Descriptive 73% Invest	Descriptive 62% Not Pay
Number of Groups	N/A	20	20	22	22	22
Investment Rate	100%	50.17% (28.87%)	72.58% (25.76%)	42.35% (29.92%)	60.98% (29.54%)	54.17% (30.17%)
Payment Rate	100%	49.61% (25.19%)	59.30% (31.44%)	31.87% (21.85%)	58.22% (28.36%)	47.49% (33.15%)
Attack Rate	0%	50.17% (40.82%)	40.50% (40.08%)	39.39% (41.84%)	42.12% (42.28%)	41.67% (42.22%)
Ransoms	100	66.14 (12.48)	58.90 (7.53)	53.39 (12.76)	64.07 (12.68)	58.23 (9.82)

Notes:
1. Standard deviations are reported in parentheses.
2. All results are significantly (p-value < 0.01) different from the game theory predictions.
3. Investment rate and attack rate are reported in group average across all periods.
4. Payment rate is reported in group average across all periods conditioned on successful attacks.
5. Ransoms are reported in group average across all periods conditioned on attack decisions.

We compare the security outcomes of groups in a treatment to those in the baseline using the Mann-Whitney test. Note that the unit of analysis is a group, not the decision in a particular round as to ensure independence. That is, for example, when we compare investment rates, we first calculate an investment rate for each group over 30 periods made by the two defenders. We then use the Mann-Whitney test to compare the investment rates in the baseline (first sample for the test) to those in a treatment (second sample for the test). For the comparison of ransoms, we use the average ransom of a group as the unit of analysis, for the same reason.

Observation 1: Investment rate is higher in the treatment of injunctive appeals for investing, but not in other treatments, than in the baseline (Table 6).

Table 6. Investment Rate Comparison (Mann-Whitney test)

	Mean	p-value
Baseline	50.17%	
Injunctive Should Invest	72.58%	0.010
Injunctive Should Not Pay	42.35%	0.450
Descriptive 73% Invest	60.98%	0.178
Descriptive 62% Not Pay	54.17%	0.614

Observation 2: Payment rate is lower in injunctive appeals for refusing to pay than in baseline, but not in other treatments (Table 7).

Table 7. Payment Rate Comparison (Mann-Whitney test)

	Mean	p-value
Baseline	49.61%	
Injunctive Should Invest	59.30%	0.175
Injunctive Should Not Pay	31.87%	0.034
Descriptive 73% Invest	58.22%	0.169
Descriptive 62% Not Pay	47.49%	0.880

Observation 3: Attack rate in the treatments is not different from the baseline (Table 8).

Table 8. Attack Rate Comparison (Mann-Whitney test)

	Mean	p-value
Baseline	50.17%	
Injunctive Should Invest	40.50%	0.228
Injunctive Should Not Pay	39.39%	0.137
Descriptive 73% Invest	42.12%	0.246
Descriptive 62% Not Pay	41.67%	0.165

Observation 4: The attacker lowers the amount of ransoms in most treatments except descriptive appeals for investing than in the baseline (Table 9). In other words, the attack may strategically respond to the interventions by reducing the amount of ransoms requested.

Table 9. Ransom Comparison (Mann-Whitney test)

	Mean	p-value
Baseline	66.14	
Injunctive Should Invest	58.90	0.024
Injunctive Should Not Pay	53.39	0.002
Descriptive 73% Invest	64.07	0.442
Descriptive 62% Not Pay	58.23	0.030

Observation 5: Investment decisions of the two defenders in a group are correlated (Table 10). As the defenders made their decision simultaneously, the correlation may be due to that both defenders could be influenced by (or anchor on) their decisions in prior rounds. We use logit regressions to test whether prior investment decisions within the group can affect a defender's current investment decision. The results (Table 11) suggest that a defender may be influenced not only by her own prior decision, but also by the other's decision.

Table 10. Pearson Correlation of Investment Decisions between Two Defenders

	Correlation Coefficient	p-value
Base	0.2431	0.000
Injunctive Invest	0.2545	0.000
Injunctive Not Pay	0.4154	0.000
Descriptive Invest	0.3664	0.000
Descriptive Not Pay	0.3605	0.000

Table 11. Effect of Investment Decisions in the Previous Round on the Current Ones

	Anchoring past (self)	Anchoring past (other)
Baseline	1.1335 (0.000)	0.4429 (0.020)
Injunctive Should Invest	1.6099 (0.000)	0.7651 (0.003)
Injunctive Should Not Pay	1.2224 (0.000)	1.3236 (0.003)
Descriptive 73% Invest	1.5543 (0.000)	0.6500 (0.002)
Descriptive 62% Not Pay	1.5381 (0.000)	0.6563 (0.037)

Notes: p-values are reported in parentheses.

Observation 6: The behaviors of the defenders' and the attacker's deviate significantly from game theory predictions. We use Wilcoxon tests to compare the observed outcomes to game theory predictions. The observed investment, attack and payment rates are significantly (all p-values < 0.01) lower than 100% by 28% - 68%. The observed ransom decisions are significantly lower than 100 in the baseline, injunctive, and descriptive social norm treatments (all with p-value < 0.01). It is not surprising that game theory does not provide good predictions in our settings, consistent with a large volume of past literature.

5. Behavior Model and Estimation

The above analyses show that descriptive appeals are not effective in neither increasing investment rate nor reducing payments rate, and injunctive appeals have some impacts on both investment rate and ransom payment rate. The reduction of ransoms can be construed as a confirmation that the attackers are strategic and can counteract interventions. In other words, even if normative appeals do nudge individuals to invest and refuse to pay, the strategy reaction of the attacker of lowering the ransom reduces this push and cancels out the impact. However, the above statistical analyses cannot fully capture the nuisance of strategic interactions (Observation 4), or control for bounded rationality of the defenders (Observation 6). Thus, a more rigorous behavioral model is needed to better quantify if the treatments impact the defenders' motivation to invest and refuse to pay.

To develop the model, we first introduce three assumptions motivated by direct observations and intuitive reasoning[2]. First, decisions were noisy and likely to be boundedly rational (Observation 6). We employ the quantal response equilibrium framework (QRE) (Mckelvey and Palfrey, 1995, Mckelvey and Palfrey, 1998), popular in the behavioral operations literature. The QRE framework assumes individuals evaluate potential decisions imperfectly and make random evaluation errors. Hence, the resulting decisions of this process are stochastic. This framework has been successfully applied to explain human decisions in many settings including inventory management (Li et al., 2019, Su, 2008) , supply chain contracts (Wu and Chen, 2014) , and capacity games (Chen et al., 2012). One appealing feature of this framework is that it is straightforward to incorporate additional behavioral elements such as fairness (Kalkanci et al., 2011, Li et al., 2019). We detail the framework in the next subsection.

Second, we find substantial correlation between the investment decisions of the two defenders and their prior decisions influence the current ones (Observation 5). The most natural way to operationalize this observation is to introduce an additional utility into the QRE framework where individuals will only receive if they follow a past investment decision, similar to the anchoring formulation in Wu and Chen (2014).

[2] We also tested the impact of cumulative wealth on the investment and payment decisions. As the effect is not consistent nor robust, we decide not to include in the discussion of the main text.

Third, we introduce fairness concerns into the payment decision[3]. The ransom-payment interaction is similar to that of the ultimatum game (Bornstein and Yaniv, 1998, Slonim and Roth, 1998). The attacker decides on an amount of ransom, which takes a portion of the data value, akin to the proposer splitting the pie in the ultimatum game. The defender then decides if he pays, akin to accepting or rejecting in the ultimatum game context. It is well-known that fairness concern is the primary behavioral factor in the ultimatum game. We incorporate fairness concerns into the QRE framework through the standard inequality aversion formulation, similar to many papers in the literature (e.g. Li et al., 2019).

As our key goal is to demonstrate the treatment effects, we introduce a *treatment comparison* feature into the modeling framework by the use of "dummy utilities". This allows us to measure the behavioral impact of a treatment (e.g. injunctive appeals) even when the strategic interactions are complex. We construct this model in a nested fashion so that statistical tests can be used to determine which behavioral factors are important. To facilitate the exposition of the modeling framework, we begin the discussion with the "vanilla" quantal response equilibrium and describe how each additional behavioral factor can be incorporated later. The specific order in which these factors are presented has no impact on the final model.

In addition, we tested a version of a behavioral model with the inclusion of risk aversion. However, we find that risk aversion is on the margin of significance. As risk aversion does not change the estimates of the parameters and drive the main insights, we do not include risk aversion in the main model but report some of this information in appendix H.

5.1. Quantum Response Equilibrium

We employ the most common logit-based QRE (Su, 2008) and find the QRE by backward induction. The setting consists of multiple decisions and multiple players. For ease of exposition, we adopt the following definition conventions. We use x for decision variables, u for utility functions and γ for bounded rationality parameter, with p, r, a and i to index the ransom payment, ransoms, attack and investment decisions, and $\mathrm{j} = 1$ or 2 to index the defender. Hence, let $\mathrm{x_{pj}} \in \{0,1\}$ be an indicator variable for the decision to pay, with $\mathrm{x_{pj}} = 1$ if the defender j pays the ransom, and 0 otherwise, conditioned on a successful attack. Let $\mathrm{x_r}$ be the ransom. Let $\mathrm{x_a} \in$

[3] We thank an anonymous review for the valuable suggestion.

$\{0,1,2\}$ be the attack decision variable, where $\mathrm{x_a} = 0$ if the attacker decides not to attack, $\mathrm{x_a} = 1$ if the attacker attacks defender 1, and $\mathrm{x_a} = 2$ if the attacker attacks defender 2. Let $\mathrm{x_{ij}} \in \{0,1\}$ be an indicator variable for defender j's decision to invest, with $\mathrm{x_{ij}} = 1$ if defender j decides to invest, 0 otherwise. We assume γ can be different for different decisions. Hence, let $\gamma_{\mathrm{p}}, \gamma_{\mathrm{r}}, \gamma_{\mathrm{a}}$ and γ_{i} be the bounded rationality parameters for ransom payment, ransom amount, attack, and investment decisions respectively. We do assume the parameters are homogeneous across individuals, similar to past literature (Ho and Zhang, 2008, Lim and Ho, 2007, Su, 2008). Note, again, that, in this section, none of the behavioral biases are included in the formulation yet.

5.1.1. Ransom Payment Decision

We start with the last possible decision in the game, whether to pay a ransom. The utility for the defender j, conditioned on a successful attack, for the ransom payment decision is given by:

$$u_p(x_{pj}) = (v - x_r)x_{pj} - c_i x_{ij}$$

where v is the data value, and c_i is the cost of investment. Recall that v and c_i are identical for both defenders. Note that the second term $-c_i x_{ij}$ has no impact on the payment decision (x_{pj}) but is included for completeness. The probability of paying a ransom is given by:

$$P_p(x_{pj} = 1) = \frac{e^{\gamma_p u_p(1)}}{e^{\gamma_p u_p(1)} + e^{\gamma_p u_p(0)}} = \frac{1}{1 + e^{-\gamma_p (v - x_r)}}$$

We use the same convention, for the bounded rationality parameter, as in Wu and Chen (2014) . When $\gamma_p = 0$, the defender chooses her decision among all possible choices with equal probability, with no intelligence whatsoever. When γ approaches ∞, the defender always selects the choice with the highest utility, consistent with rational theory.

5.1.2. Ransom Decision

By backward induction, we analyze the ransom decision of the attacker's next[4]. The utility of the ransom decision, for the attacker, conditioned on attacking defender j, is simply ransom, expected over whether it will be paid, given by:

$$u(x_r) = x_r P_p(x_{pj} = 1)$$

[4] Technically, the ransom decision is made at the same time as the attack decision. So it is possible to analyze them as a pair. However, the ransom decision will only "kick in" if the attacker decides to attack. Hence, we analyze them as a sequence of two decisions.

Note that the quantal response probability of the payment decision, $P_p(\,)$, is a function of the ransom x_r and that creates a non-trivial trade-off for the attacker to consider. The quantal response probability of choosing a ransom x_r is given by:

$$P_r(x_r) = \frac{e^{\gamma_r u_r(x_r)}}{\sum_{x_r'} e^{\gamma_r u_r(x_r')}}$$

We restrict x_r to be between 0 and v, the value of the data, and that it has to be an integer. Note that we are using a discrete formulation because in the experiments, decisions are restricted to be integers. Technically, the model can be easily extended to a continuous formulation by replacing the summation operation with an integration operation.

5.1.3. Attack Decision

Similarly, we formulate the utility of the attack decision, based on the quantal response probabilities of the subsequent decisions. In our setting, recall that the probability of a successful attack, attacking defender j, is $p_s = p_0 - p_1 x_{ij}$. That is, defender j can reduce the probability of a successful attack by p_1 by investing. The expected utility of the attack decision is given by:

$$u_a(x_a) = \begin{cases} u_{a0} & if \quad x_a = 0 \\ (p_0 - p_1 x_{ij}) \sum_{x_r} P_r(x_r) x_r P_p(x_{pj} = 1) & \quad\;\; x_a = 1\ or\ 2 \end{cases}$$

where u_{a0} is the utility of the outside option if the attacker decides not to attack anyone. Similar to the other decisions, the decision probabilities are given by:

$$P_a(x_a) = \frac{e^{\gamma_a u_a(x_a)}}{\sum_{x_a' \in \{0,1,2\}} e^{\gamma_a u_a(x_a')}}$$

5.1.4. Investment Decision

Investment decisions are made simultaneously by defender 1 and 2. To be clear in our exposition, we explicitly use the index 1 and 2 to refer to the two defenders. The utility of defender 1, at this stage, is driven by four possible outcomes: the attacker not attacking, the attacker attacking defender 2, the attacker attacking defender 1 without success, and the attacker attacking defender 1 with success. Defender 1 receives $v - c_i x_{i1}$ in the first 3 cases because there is no successful attack, and she retains her data value minus her investment costs, if any. In the last case, the utility is $P_p(x_p = 1)(v - x_r - c_i x_{i1}) + (1 - P_p(x_p = 1))(-c_i x_{i1})$, expected over x_r. This is, it is the expectation between two cases: paying the ransom and receives $v - x_r - c_i x_{i1}$ or not paying and losing the investment cost ($c_i x_{i1}$), if any. Putting all these together, we have:

$$u_i(x_{i1}, x_{i2}) = (v - c_i x_{i1})(P_a(x_a = 0) + P_a(x_a = 2) + P_a(x_a = 1)(1 - (p_0 - p_1 x_{i1})))$$

$$+ P_a(x_a = 1)\left\{\sum_{x_r} P_r(x_r)\big(P_i\big(x_p = 1\big)(v - x_r - c_i x_{i1}) + P_i(x_p = 0)(-c_i x_{i1})\big)\right\}$$

Note that u_i is a function of x_{i2} since P_a is a function of x_{i2}. The utility of defender 2 is given, simply, by $u(x_{i1}, x_{i2})$, a permutation of the index of 1 and 2, because the game is symmetric. The quantal response equilibrium is given by a fixed point of the distributions of both x_{i1} and x_{i2}, satisfying both of the following conditions simultaneously.

$$P_{i1}(x_{i1}) = \frac{e^{\gamma_i \sum_{x_{i2}} P_{i2}(x_{i2}) u(x_{i1}, x_{i2})}}{\sum_{x'_{i1}} e^{\gamma_i \sum_{x_{i2}} P_{i2}(x_{i2}) u(x'_{i1}, x_{i2})}}$$

$$P_{i2}(x_{i2}) = \frac{e^{\gamma_i \sum_{x_{i1}} P_{i1}(x_{i1}) u(x_{i2}, x_{i1})}}{\sum_{x'_{i2}} e^{\gamma_i \sum_{x_{i1}} P_{i1}(x_{i1}) u(x'_{i2}, x_{i1})}}$$

Essentially, defender 1 is calculating her utility by taking expectation over defender 2 quantal response probabilities and vice versa.

5.2. Previous Decision Anchoring

We incorporate past decision anchoring for the investment payment decision, into the model as a reduction in investment cost if the individual invests in the last round. That is, if the last decision is to invest, it is easier, operationalized by a reduction of perceived cost, to invest in the current period. In the experiments, the defender subjects are aware of previous decisions of *both* defenders. Hence, there are two possible anchors and it is an empirical question of whether defenders anchor on one, both, and how much. Hence, the anchoring utility is defined as

$$perceived\ investment\ cost = true\ cost - \alpha_m(x\ x_m) - \alpha_y\big(x\ x_y\big)$$

where *x* is the current investment decision of a defender (1 or 0). x_m and x_y are the last period's decision of the same defender and the other defender, respectively. α_m and α_y are the respective cost reduction the defender would perceive if he invested and the other defender invested in the last round. α_m and α_y are parameters to be estimated from the data.

Note that we frame the utility change as a cost reduction for convenience of interpreting α_m and α_y as both will be measured in dollars in the same scale as the investment cost. The formulation can be reframed as general anchoring utilities.

5.3. Fairness

We incorporate fairness concerns for the payment decision using the standard inequality aversion formulation (e.g. Cui et al., 2007, Li et al., 2019). We incorporate this inequality aversion into the utility for the payment decision, as follows:

$$u_p(x_{pj}) = (v - x_r - \alpha_f \max(r - [v - r - c_i x_{ij}], 0))x_{pj} - c_i x_{ij}$$

where $\alpha_f \max(r - [v - r - c_i x_{ij}],0)$ is the standard inequality aversion utility. r, the ransom, is what the attacker receives if the defender pays. $[v - r - c_i x_{ij}]$ is what the defender receives. Hence, the disutility is the difference between the two[5], but only if the attacker receives more. α_f is a parameter, to be estimated from data, interpreted as the degree of which the defender cares about this inequality. If $\alpha_f = 1$, the defender cares about this difference as much as about her own payoff. We expected, in general, that α_f to be less than 1. Intuitively, this inequality aversion utility will push the defender not to pay (and hence avoid this disutility) when the ransom is high.

5.4 Treatment Comparison

To quantify the impact of normative appeals, we introduce a term of dummy utility that measures the perceived reduction of cost for a decision in a treatment, referred to as the "treatment effect" for a particular decision, into the model. If a treatment is successful, a defender is nudged to invest more and it will be as if the perceived cost of investment is lowered, compared to the baseline. The difference between this perceived cost and the actual cost can be interpreted as the impact, measured in dollars, of the treatment. Alternatively, this perceived decrease in cost can also be interpreted as an increase in utility, measured in dollars, of investing caused by the treatment. Mathematically both perspectives are equivalent. We arbitrarily decide to use the term "treatment effect".

[5] In the most general formulation, there is an additional parameter determining the "fairness" split as the split does not have to be equal. We opt to use the simpler formulation as in Li et al. 2019 as not to defocus the paper. Empirical analysis suggests that this formulation is significant and explains the data.

Note that this "treatment effect" is completely agnostic to *how* the impact is created. The same mathematical formulation is used to measure the 4 treatments (injunctive and descriptive appeals for investing, injunctive and descriptive appeals for refusing to pay). Also note that we measure the effect of a treatment on both the investing and not-paying decisions as there may be spillover effect when appeals for one decision may cause a treatment effect in the other decision.

Technically, the treatment effect for investing is formulated as a modification of the equation in section 5.2, given by:

$$perceived\ investment\ cost = true\ cost - \alpha_m(x\, x_m) - \alpha_y(x\, x_y) - \beta_i z$$

where β_i is the treatment effect, to be estimated, on investing. z is a dummy variable. $z = 0$ for baseline and 1 for a given treatment. Note that this model will be estimated with the pooling data from the baseline and a treatment. This procedure will be performed for all the treatments (i.e. estimating β_i, the utility increase or perceived cost decrease, of investing, caused by every treatment).

Similarly, the treatment effect on not-paying is formulated as a modification of the equation in section 5.3, given by:

$$u_p(x_{pj}) = \left(v - x_r - \alpha_f \max\left(r - [v - r - c_i x_{ij}], 0\right)\right)x_{pj} - c_i x_{ij} + \beta_p(1 - x_{pj})$$

Similarly, β_p is the treatment effect, to be estimated, of not-paying and z is the same dummy where $z = 0$ for baseline and 1 for treatment. The individual will only receive the utility β_p, and only in treatment, if s/he chooses $x_{pj} = 0$ (not paying). Note that it is formulated as a straight increase in utility but there is an equivalent cost reduction interpretation.

5.5. Model Estimation

We use the maximum likelihood method to estimate the parameters of the model, given in the previous sections. For the base model, there are 7 behavioral parameters, $(\gamma_p, \gamma_r, \gamma_a, \gamma_i, \alpha_m, \alpha_y, \alpha_f)$ representing bounded rationality for the four decisions (ransom payment, ransoms, attack and investment), decision anchoring, and fairness. Similar to Chen et al. (2012), we assume the behavioral parameters are homogeneous across individuals[6], and the loglikelihood function is

[6] Note that if we were to estimate behavioral parameters for each individual, we would have only 30 rounds of data per estimation. Given we have 6 parameters, the statistics will be very weak.

simply the sum of the log of the probabilities of each decision. The loglikelihood function is, hence, given by:

$$LL(\theta) = \sum_k \left\{ log\left(P_p(x_{pjk})\right) + log\left(P_r(x_{rk})\right) + log\left(P_a(x_{ak})\right) + log\ \left(P_i(x_{ijk})\right)\right\}$$

where $\theta = \gamma_p, \gamma_r, \gamma_a, \gamma_i, \alpha_m, \alpha_y, \alpha_f$ are the behavioral parameters, and the decisions $(x_{pjk}, x_{rk}, x_{ak}, x_{ijk})$ have an extra index k, indexing the particular game.

As mentioned in 5.4, we expand the model to include two additional "treatment utilities", with parameters β_i and β_p to enable us to measure the impact of manipulation. In this case, we use the same maximum likelihood method, with the pooled data from the baseline and a treatment.

5.5.1 Baseline Treatment Estimation Results

Table 12 summarizes the estimation results for the baseline treatment. We include p-values for likelihood ratio tests with the null hypothesis that the parameter is zero. The results show that the bounded rationality parameter for all the decisions $(\gamma_p, \gamma_r, \gamma_a, \gamma_i, \alpha_m, \alpha_y, \alpha_f)$ are positive and significantly different from 0 with p-values of practically zero for all decisions in all treatments. This indicates that individuals are responding to incentive and exhibit some level of rationality. It is consistent with that decisions deviate substantially from rational game theoretic predictions.

Table 12: Estimation of the Baseline Model

Parameter	Estimate	p-value
γ_p: payment bounded rationality	0.0155	0.000
γ_r: ransom bounded rationality	0.1389	0.000
γ_a: attack bounded rationality	0.0460	0.000
γ_i: investment bounded rationality	0.1002	0.000
α_m: anchoring past (self)	22.1573	0.000
α_y: anchoring past (other)	6.3192	0.000
α_f: pay fairness	0.4447	0.000

Past decision anchoring and fairness are highly significant. Past decision anchoring is much stronger for the defender's own decision, compared to the other defender's decision, consistent with the intuition that the defender's own past decision is more salient, as an anchor, compared to the other defender's past decision. Please see appendix E for the estimation results under the condition of injunctive and descriptive appeals. All results are consistent with the baseline model.

5.5.2 The Effect of Normative Appeals

Pairwise comparisons were performed (Table 13), using the formulation in section 5.4. Data from the baseline and a compared treatment are pooled for the estimation. Recall that the estimates β_i and β_p are the *additional* utilities for the investing and not-paying decisions caused by the treatments.

Table 13. Comparisons Between Treatments and Baseline

comparison to Base	Should Invest	Should Not Pay	73% Invest	62% Not Pay
Parameter	Estimate (P-value)	Estimate (P-value)	Estimate (P-value)	Estimate (P-value)
γ_p: payment bounded rationality	0.0253 (0.000)	0.0193 (0.000)	0.0258 (0.000)	0.0224 (0.000)
γ_r: ransom bounded rationality	0.1354 (0.000)	0.1434 (0.000)	0.1036 (0.000)	0.1545 (0.000)
γ_a: attack bounded rationality	0.0579 (0.000)	0.0455 (0.000)	0.0665 (0.000)	0.0502 (0.000)
γ_i: investment bounded rationality	0.0746 (0.000)	0.0828 (0.000)	0.0746 (0.000)	0.0839 (0.000)
α_m: anchoring past (self)	33.8596 (0.000)	30.9664 (0.000)	32.4527 (0.000)	34.0986 (0.000)
α_y: anchoring past (other)	9.8161 (0.000)	12.0928 (0.000)	11.1252 (0.000)	9.2221 (0.000)
α_f: pay fairness	0.2799 (0.000)	0.3813 (0.000)	0.2064 (0.000)	0.3328 (0.000)
β_i: treatment utility (invest)	7.4179 (0.000)	-1.0612 (0.408)	3.0436 (0.023)	1.4766 (0.246)
β_p: treatment utility (not-pay)	8.1731 (0.044)	55.7807 (0.000)	-2.2156 (0.553)	23.2035 (0.000)

Result 1: Both injunctive and descriptive appeals increase the utility of investing.

The β_i parameter is positive and significant for both the "should invest" treatment (p-value practically 0) and "73% invest" treatment (p-value = 0.023). In the case of injunctive appeals ("should invest"), the β_i estimate is 7.41, which can be interpreted as a perceived decrease in investment cost. Compared to the real cost of 30, it is roughly a 25% impact. In the case of descriptive appeals ("should invest"), the β_i estimate is 3.04, roughly a 10% impact.

Result 2: Both injunctive and descriptive appeals increase the utility of not-paying.

The β_p parameter is positive and significant for both the "should not pay" treatment (p-value practically 0) and "62% not-pay" treatment (p-value practically 0). The estimates of the utility change in both treatments are high (55.78 and 23.20 respectively) as a fraction of the data value of 100.

Result 3: Injunctive appeals for investing has a spill-over effect to increase the utility of not-paying.

The β_p parameter is positive and significant for the "should invest" treatment (p-value = 0.044). But the estimate is much smaller than that in the comparison for "should not pay" and "62% not pay" treatments. Using an injunctive social norm to manipulate the utility of investing may have a spill-over effect of increasing the utility of not-paying ransoms. We speculate that the increase in investment rate in the "should invest" treatment, compared to that of the baseline, results in a variation of mental accounting where the defender is unwilling to pay "more" given that she has already invested.

From the above analyses, we see normative appeals do nudge the defenders to the desired directions. However, for some treatments, we fail to see that they significantly increase investment rate (Observation 1), and/or reduce payment rate (Observation 2). There could be two possible reasons. First, the nudging effect is too weak to be detected statistically given the defenders' noisy decision-making behaviors. The failure of "73% invest" in increasing investment rate (Observation 1) and the failure of "should invest" in reducing payment rate significantly may be due to this reason (Observation 2). Second, the nudging effect may be mitigated by the attacker with lowering ransoms. Such strategic reactions from attackers were found for the treatments of "should invest", "should not pay", "62% not pay". Only "should not pay" presents a significant reduction of payment rate. The results suggest the challenges in coping with digital extortion. While interventions may be successful in increasing a decision maker's utility of investing, or utility of not-paying ransoms, their impacts can be mitigated with the attacker reducing ransoms or overshadowed by noisy decision-making behaviors.

5.6. Implications on Security Outcomes

The behavioral model, incorporating bounded rationality, fairness and past-decision anchoring, is designed to track strategic interactions and explain the main empirical conclusion: normative appeals may increase the treatment utilities of investing and not-paying, all else equal. In this section, we are to establish a clear picture of how the change of these treatment utilities impact the four security outcomes of a community: investment rate, attack rate, ransoms requested, and payment rate, by numerical analyses. Specifically, we use the parameters from the baseline estimation reported in Table 12, except the treatment utilities of both investing and not-paying. We run the model using a range of treatment utilities to track how the security outcomes change as a result of the changes in one of treatment utilities while holding the other to be 0. We plot expected ransom, attack rate, investment rate, and payment rate as a function of the two treatment utilities (two separate lines) as shown in Figure 4. The behavioral model provides direct calculations of the probabilities of investing, attacking, and not-paying (Section 5.1.1, 5.1.3, and 5.1.4). It also provides a probability distribution (Section 5.1.2) for ransom amounts, as opposed to a single point prediction. Hence, we plot the *expected* ransom in Figure 4. Table 14 summarizes our main findings.

Note that these results are completely agnostic to the specifics of the manipulation as long as the manipulation increases the utilities of investing and/or not-paying, and that the attacker is aware of the manipulation. For this reason, the results can be applied to other types of manipulations (e.g. direct incentives, peer pressure). While we can measure the strength, by the use of the treatment utility formulation, of interventions, it is out of the scope of the paper to provide a theory of how specific interventions result into different levels of treatment utilities.

Table 14. Results of Numerical Analyses

Player	Decision	Treatment Utility of Investing Increases	Treatment Utility of Not-Paying Increases
Attacker	Ransoms	decreases moderately	decreases considerably
	Attack Rate	decreases moderately	decreases slightly
Defenders	Investment Rate	Increases considerably	decreases slightly
	Payment Rate	almost no change	decreases slightly

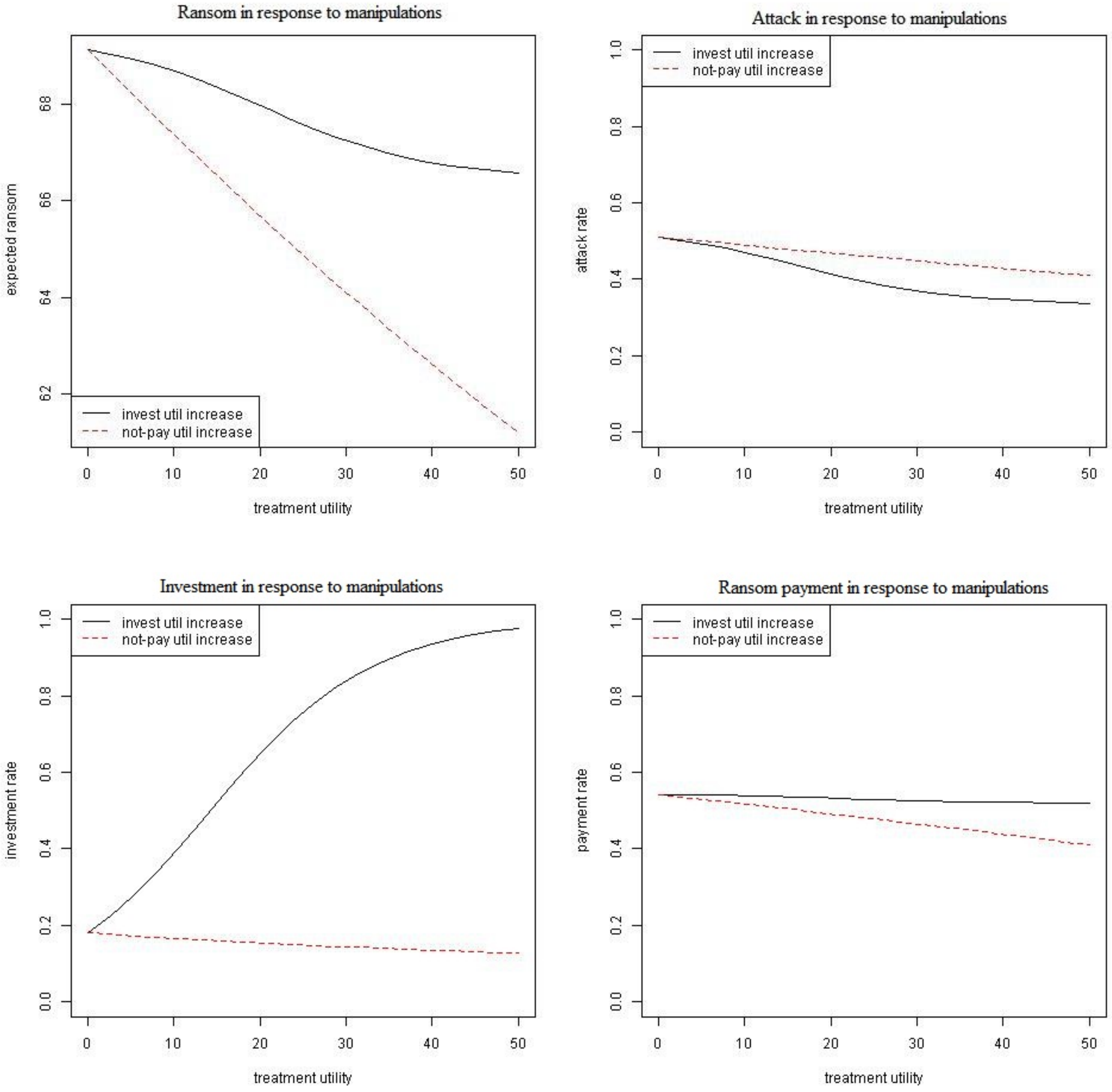


Figure 4. Response of Security Outcomes to the Change of Utilities

5.6.1 Ransoms

The model predicts that the attacker reduces the ransom in response to the increase of both utilities. When the treatment utility of not-paying increases, the ransom decreases. Intuitively, the attacker is compensating for a higher utility of not-paying. When the treatment utility of investing increases, the ransom decreases, although at a lower rate, compared to the effect of the treatment utility of not-paying. Note that the investing decision is made before the ransom decision. So how can the treatment utility of investing affect the expected ransom? The explanation of this conundrum lies in the interactions of fairness and social norm manipulations. First, the expected ransom conditioned on investing is *lower* than that conditioned on not-investing. The reason is that if the defender invested, his payoff will be reduced by the investment cost, and he will feel “less fair” for the same level of ransoms, compared to the case if he did not invest. Knowing this, the attacker will adjust the ransom down to compensate. Second, when the treatment utility of investing increases, the probability to be in the case of investing, as opposed to not-investing, is higher. Thus, the ransom, expected over all quantal response equilibrium distributions, including that for the investment decision, is lower. This impact is based on an indirect effect through fairness. Hence, it is weaker than the manipulation of the utility of not-paying.

5.6.2. Attack Rate

The model predicts that attack rate only decreases slightly when the treatment utility of investing or that of not-paying increases. This may be because the outside option is not appealing even compared with a reduced ransom. Attack rate is more likely to be decreased for a higher treatment utility of investing than for a higher treatment utility of not-paying. This may be because the increased difficulty in compromising a target are more likely to drive the attacker to take outside options. However, we do not find the attack rate in the treatments lower significantly than the baseline.

5.6.3. Investment Rate

As expected, the model predicts that, when the treatment utility of investing increases, investment rate increases. When the treatment utility of not-paying increases, the model predicts a small decrease in investment rate. The reason is that the ransom is significantly reduced, as demonstrated in the last section. Hence, the expected loss is lower and the need to invest is lower. However, this effect is small, and we do not observe any significant change of investment rate when we have normative and descriptive appeals for not-paying.

5.6.4. Payment Rate

Finally, we illustrate how payment rate changes with manipulations. As the treatment utility of not-paying increases, payment rate decreases, but only slightly. The reason is that the ransom decreases substantially. We did observe this slight, but significant effect, in the “should not pay” treatment. In the “62% not pay” treatment, this effect is too small to be significant. But we observe a significant drop in ransoms in both treatments. The treatment utility of investing has an almost undetectable impact on payment rate and we did not find a significant change of payment rate with normative and descriptive appeals for investing. Note that result 3, stating that injunctive appeals for investing has a spill-over effect to increase the utility of not-paying, seemingly contradicts this finding. To clarify, result 3 is about the increase of treatment utility for not-paying when injunctive appeals for investing are used as an intervention. This finding is about how the treatment utility of investing impacts payment rate, a security outcome of a community.

6. Alternative Interventions

As mentioned previously, the analyses above are agnostic of how the manipulation of the utilities of investing and not-paying are achieved, but only predicts how security outcomes change with these utilities of the defenders. To further explore possible manipulations of utilities, we investigate two alternative interventions other than normative appeals employed in the main treatments.

The first is a manipulation of direct incentives where we introduce a penalty for ransom payment. We expect a direct penalty to increase the utility of not-paying ransoms (i.e. increase the treatment utility of not-paying). According to the model analysis in the previous section, we anticipate a substantial decrease in ransoms, and all other security outcomes (investment rate, attack rate and payment rate) either remain unchanged or decrease slightly.

The second is a "chat" treatment where we manipulate social interactions between the two defenders. That is, we allow the defenders to have free-form communications before they make investment decisions. A comprehensive analysis of costless communication, or cheap talk, can be found in a number of studies (Farrell, 1987, Farrell and Gibbons, 1988, Rabin, 1990, Rabin, 1994). We posit that cheap talk is a means to deliver players' intentions and thereby improve coordination for both investment and ransom payment decisions. Communication allows defenders to exchange reasoning for their actions and influence each other. We summarize the major findings in this section. Please see Appendix G for details including summary statistics and the model estimation results of these treatments.

6.1 Penalty Treatment

We set the penalty of paying a ransom, in this treatment, to be 15% of the data value. We employ the model outline in Section 5.4 to estimate the impact of the intervention. Table 15 summarizes the estimation results. The penalty intervention significantly increases the utility of not-paying. The β_p parameter is positive and significant, with a p-value less than 0.001. Note that, in this case, the treatment utility of not-paying should be interpreted as the additional mental cost of paying a ransom, in addition to the actual monetary penalty included in the utility calculation explicitly as a cost in the model. This is consistent with the interpretation that the penalty itself, which is communicated as such to the subjects, is viewed with a negative connotation and further increases the utility of not-paying ransoms, on top of the actual monetary incentive. The penalty intervention has no impact on the utility of investing. The β_i parameter is not significantly different from 0 with a p-value of 12.9%. This is in-line with our expectation that a penalty of paying ransom has no direct impact on the utility of investing.

Table 15. Comparison between Penalty Treatment and Baseline

comparison to Base	Penalty
Parameter	Estimate (P-value)
γ_p: payment bounded rationality	0.0237 (0.000)
γ_r: ransom bounded rationality	0.1435 (0.000)
γ_a: attack bounded rationality	0.0511 (0.000)
γ_i: investment bounded rationality	0.0739 (0.000)
α_m: anchoring past (self)	29.6958 (0.000)
α_y: anchoring past (other)	13.8246 (0.000)
α_f: pay fairness	0.3030 (0.000)
β_i: treatment utility (invest)	2.0731 (0.129)
β_p: treatment utility (not-pay)	13.3767 (0.001)

Analyses in section 5.6.1 suggest that an increase in treatment utility of not-paying should be accompanying a decrease in ransoms. Indeed, as shown in Appendix G, we find the ransom is significantly lower, with a p-value of less than 0.01, than the baseline. The other security outcomes (investment rate, attack rate and payment rate) are not significantly different from the baseline, consistent with the model analysis.

6.2 Chat Treatment

We believe that communication has the potential to improve coordination and increase understanding of the decision problems. However, as opposed to the penalty intervention, there is less theoretical guidance of whether and how enabling communication will impact one or both of the treatment utilities. Note that all communication in this scenario is cheap-talk, as there is no enforcement mechanism for any intentions or suggestions that are communicated.

The model formulation in section 5.4 is designed to isolate the impact of a treatment intervention, compared to the baseline, which is adequate in all previous analysis. In this case, however, we also observe the actual *content* of the communication (i.e. text sent by subjects). We develop a variation of the behavioral model for treatment comparison to incorporate this additional information with the goal to determine whether the content is important in addition to the general ability to communicate for nudging a decision maker's utility to act.

In particular, we create a pair of new indicator variables y_i and y_p, where $y_i/y_p = 1$ if investment/payment is mentioned in the particular round by either defender 1 or defender 2. We opt to employ this simpler approach, as opposed to a full text mining analysis, because the messages are generally short, and that we are only interested in the two decisions. We set these indicator variables by looking for specific, pre-determined, words such as "invest" and "investment" (see Appendix F for examples of chatting messages). We refer to these two variables as *chat content variables*. Please see Table 17 for the frequencies of these words mentioned. Recall that treatment utility formulations, defined in section 5.4 is as follows.

$$perceived\ investment\ cost = true\ cost - \alpha_{me}(x\, x_{me}) - \alpha_{you}(x\, x_{you}) - \beta_i z$$

$$u_p(x_{pj}) = (v - x_r - \alpha_f \max(r - [v - r - c_i x_{ij}],0))x_{pj} - c_i x_{ij} + \beta_p(1 - x_{pj})z$$

We re-formulate β_i and β_p to $(\beta_i + \Delta_i y_i)$ and $(\beta_p + \Delta_p y_p)$ respectively. Hence, the above equations become:

$$perceived\ investment\ cost = true\ cost - \alpha_{me}(x\, x_{me}) - \alpha_{you}(x\, x_{you}) - (\beta_i + \Delta_i y_i)z$$

$$u_p(x_{pj}) = (v - x_r - \alpha_f \max(r - [v - r - c_i x_{ij}],0))x_{pj} - c_i x_{ij} + (\beta_p + \Delta_p y_p)(1 - x_{pj})z$$

The interpretation is that the respective utilities are changed by β_i and β_p in the treatment (when z=1), and *further* changed by Δ_i and Δ_p if certain words are mentioned during the chat session. Please see Table 16 for the estimates.

We conclude that indeed communication has an impact on the utility of investing, but only if investment is discussed. β_i is not significant but Δ_i is highly significant with a p-value of practically 0. We also find that investment rate, in the rounds where investment is discussed, is significantly higher than investment rate in the baseline (74.63% vs 50.17%) with a p-value of less than 1% (please see appendix G for the relevant statistics). Furthermore, this difference goes away if we compare the overall investment rate of the chat treatment with that of the baseline. This is strong evidence that communication matters, but only when the discussion is relevant.

Table 16. Comparison between Chat Treatment and Baseline

comparison to Base	Chat
Parameter	Estimate (P-value)
γ_p: payment bounded rationality	0.0177 (0.000)
γ_r: ransom bounded rationality	0.1675 (0.000)
γ_a: attack bounded rationality	0.0521 (0.000)
γ_i: investment bounded rationality	0.0840 (0.000)
α_m: anchoring past (self)	30.8140 (0.000)
α_y: anchoring past (other)	11.7297 (0.000)
α_f: pay fairness	0.4233 (0.000)
β_i: treatment utility (invest)	-0.4463 (0.735)
β_p: treatment utility (not-pay)	12.3924(0.023)
Δ_i: chat invest	14.8104 (0.000)
Δ_p: chat pay	10.0757 (0.583)

In addition, we find an increase in the utility of not-paying which does not depend on whether "paying" is discussed. β_p is significant with a p-value of 0.023 but Δ_p is not significant. We speculate that, as 75% of the infrequent ransom payment discussion occurred in the first half of the experiment, the impact of discussions, in this case, persisted throughout the session. Empirically, we find no evidence of a change in the ransom or ransom payment rate. That is not outside of our expectation as we observe a significant amount of decision noise.

Finally, we find that relevant discussions take place only in a minority of the rounds (Table 17). While communication seems to have improved investment in the right circumstances, these results, taken in total, point to a concern of this approach. Namely, there is a lack of control of what is discussed, even when discussion can improve decision-making.

Table 17. Message Frequency in Chat Treatment

Message Content	# of Messages	Percentage
Investment	134	21.54%
Ransom Payment	27	4.34%
None of them[7]	461	74.12%

[7] "None of them" means the defenders mentioned neither investment nor ransom payment, but may be something else at the round.

In response to this treatment, we find the attacker neither lower the ransom nor change the attack compared with the baseline (Appendix G). We also did not find a drop of payment rate. Neither do we detect an increase of investment rate overall. The results may be caused by the low number of rounds with relevant communications.

7. Discussion and Conclusion

This study investigates how we can incentivize the defenders to adopt their strategies of mitigating digital extortion: investing and refusing to pay ransoms, through the lens of behavioral game theory. We are among the first to shed light on coping with digital extortion, employing a combination of game theory, human-subject experimentation, behavioral modeling, and numerical analyses.

We have three main findings. First, normative appeals influence a defender to invest in information security and to refuse to pay ransom. Specifically, we find that the defender enjoys a significantly positive utility if she conforms to normative appeals. Different interventions result in different levels of utility impacts on investing and not-paying. Some may have a stronger impact on investing, while others on not-paying.

Second, the attacker strategically responds to the interventions we apply on the defenders by lowering ransoms. We observe such responses in both normative appeals and penalty for payment. Numerical analyses show when the defenders' utility of not-paying increases, that the attacker lowers ransoms considerably and attack rate slightly. When the defenders' utility of investing increases, the attack lowers both ransoms and attack rate very slightly. Ransoms are more likely to be decreased for utility of not-paying, while attack rate is more likely to be decreased for utility of investing.

Third, while interventions may be successful in increasing the utility of investing, or reduce the utility of not-paying, their impacts can be mitigated with the attacker reducing ransoms or overshadowed by noisy decision-making behaviors. Thus, it may be difficult for an intervention to significantly boost investment rate and lower payment rate. The study suggests potential approaches, but also identifies challenges, for fighting against digital extortion for policy makers.

The study also makes methodological contributions. It introduces a methodological framework to utilize game theory and behavioral experiments to the context of information security and to study digital extortion. It makes the first endeavor to empirically explore strategic interactions between defenders and attackers in the context of information security with human subject experimentation. In addition, the behavioral model developed in this paper provides a venue to estimate the impact of interventions, even when the strategic interactions are complex. Furthermore, numerical analyses based on the behavioral model help generalize the qualitative findings in terms of how the security outcomes of a community in terms of expected ransom, attack rate, investment rate, and payment rate change with the utilities of investing and not-paying.

From a managerial and policy-making perspective, this study suggests that some interventions may not have enough impacts to change investment rate and payment rate of a community significantly, particularly when attackers can influence the will of the defenders by lowering ransoms and when the defenders are bounded rational. Our study explores strategies leveraging normative appeals as a motivator. In practice, industry forums, special interest groups, and other relevant organizations can be employed to raise awareness of the threat of digital extortion and provide a place for organizations to exchange information about their solution approaches. These activities can be designed around building a community to support social norms of adopting and investing in effective mitigation strategies, as well as refusing to pay attackers. However, building a community that encourages adhering to social norms is not enough. A full solution also requires the right social norms to be established. In our study, subjects can easily find "right" social norms because their actions are limited (to binary decisions). In practice, there are multiple mitigation strategies, and defenders need to be able to coordinate on the right one(s). One possibility is to cherry-pick "good behaviors" that naturally arises and frame that as social norms. A practical example is to create forums that showcase successes in how mitigation strategies thwart attacks. Other organizations may follow suit, and the practice can snow-ball into a standard that the community is willing to follow.

Much of the study is focused on how to prevent successful attacks by either investing or reducing attackers' incentive by not-paying ransoms, and rightfully-so. However, from an economics perspective, the issue is not so much of whether there is a successful digital extortion attack, but the size of the pecuniary loss. While most of the interventions we studied do not increase investment nor reduce attack rate significantly, they do reduce ransoms. In this regard, we can claim that these interventions are more "successful" than they appear to be. Ultimately, there is a broader question of how to value the trade-off between investing a lot into information security and paying the "bad guys" a little to make them go away. Do we absolutely not negotiate with cyber "terrorists" or it is okay to pay them a little as long as the ransom is small enough? As this question is more philosophical than scientific, it is beyond the scope of this paper and we will leave it for the readers to ponder.

The study is not without limitations. As one of the first studies to examine digital extortion via a behavioral game theory approach, we opt for a simple design as to zero in on the key principles of the interactions between strategic considerations and social forces. Hence, some nuances in real world scenarios are not fully explored. For example, we assume a full information game where all the relevant parameters such as payoffs, effectiveness of security investments and attacker outside options are known to all players when substantial asymmetric information may exist. Another example is that defenders are assumed to be identical where they are heterogeneous in security investment costs/effectiveness, and the value of their data. We also limit the study to focus on few salient interventions (chat and penalty) without fully exploring all possibilities. For example, we did not explore subsidies for security investment nor reporting requirements for breaches. From a methodological perspective, the study relies on controlled experiments and mathematical modeling. While these techniques are common practices in many areas of business research, it is still not the same as field experiments where the contexts are real, and participants have domain knowledge and experiences.

The limitations of the study suggest future extensions, along three directions. The first is to investigate how setting characteristics such as the asymmetry of the defenders, information structure, and externalities in security investments, can affect our conclusions. The second is to study an expanded set of interventions. This second direction can be fertile ground for multiple research studies. It can be subdivided into multiple types. One such type is social interaction-based intervention, such as positive/negative social interaction manipulation, and combining multiple interventions such as communication (chat) and social interaction manipulation. Another type is the direct incentive-based intervention. There are a wide range of untested possibilities. Leaderboard, which leverages the need to compete, is one example. Social and incentive-based interventions may enhance the effects of one another, and combinations should also be considered. Finally, field tests and empirical studies, with the right data, can provide a link between our results and real-world practices.

References

Anderson CL, Agarwal R. 2010. Practicing safe computing: A multimedia empirical examination of home computer user security behavioral intentions. *MIS Quarterly*. **34**(3) 613-643.

Anderson R, Moore T. 2006. The economics of information security. *Science*. **314** 610-613.

Bornstein G, Yaniv I. 1998. Individual and group behavior in the ultimatum game: Are groups more "rational" players? *Experimental Economics*. **1**(1) 101-108.

Brewer R. 2016. Ransomware attacks detection, prevention and cure. *Network Security*. **2016**(9) 5-9.

Bulgurcu B, Cavusoglu H, Benbasat I. 2010. Information security policy compliance: An empirical study of rationality-based beliefs and information security awareness. *MIS Quarterly*. **34**(3) 523-548.

Burtch G, Hong Y, Bapna R, Griskevicius V. 2018. Stimulating online reviews by combining financial incentives and social norms. *Management Science*. **64**(5) 2065-2082.

Cavusoglu H, Raghunathan S, Yue WT. 2008. Decision-theoretic and game-theoretic approaches to it security investment. *Journal of Management Information Systems*. **25**(2) 281–304.

Chen Y, Su X, Zhao X. 2012. Modeling bounded rationality in capacity allocation games with the quantal response equilibrium. *Management Science*. **58**(10) 1952-1962.

Chu A, Chau P, So M. 2015. Explaining the misuse of information systems resources in the workplace: A dual-process approach. *Journal of Business Ethics*. **131**(1) 209-225.

Cialdini RB. 2001. *Influence: Science and practice*. Allyn & Bacon, Boston, MA.

Cialdini RB. 2007. Descriptive social norms as underappreciated sources of social control. *Psychometrika*. **72**(2) 263-268.

Cialdini RB, Kallgren CA, Reno RR. 1991. A focus theory of normative conduct: A theoretical refinement and reevaluation of the role of norms in human behavior. *Advances in Experimental Social Psychology*. **24** 201-234.

Cialdini RB, Reno RR, Kallgren CA. 1990. A focus theory of normative conduct: Recycling the concept of norms to reduce littering in public places. *Journal of Personality and Social Psychology*. **58** 1015–1026.

Cialdini RB, Trost MR. 1998. *Social influence: Social norms, conformity and compliance*. McGraw-Hill, New York, NY.

Cremonini M, Dmitri N. 2009. Risks and benefits of signaling information system characteristics to strategic attackers. *Journal of Management Information Systems*. **26**(3) 241–274.

Croson R, Handy F, Shang J. 2009. Keeping up with the joneses the relationship of perceived descriptive social norms, social information, and charitable giving. *Nonprofit Management & Leadership*. **19**(467-489).

Croson R, Shang J. 2008. The impact of downward social information on contribution decisions. *Experimental Economics*. **11**(3) 221-233.

Crowdstrike. 2020. *2020 global threat report*. CrowdStrike.

Cui H, Raju JS, Zhang ZJ. 2007. Fairness and channel coordination. *Management science*. **53**(8) 1303-1314.

Das S, Kim TH-J, Dabbish L, Hong J. 2014. *The effect of social influence on security sensitivity*. Menlo Park, CA.

Das S, Kramer A, Dabbish L, Hong J. 2015. *The role of social influence in security feature adoption*. ACM, Vancouver, BC, Canada.

Das S, Kramer A, Dabbish L, Hong JI. 2014. *Increasing security sensitivity with social proof: A large scale experimental confirmation*. ACM, Scottsdale, AZ.

Donohue K, Katok E, Leider S. 2018. *The handbook of behavioral operations*. John Wiley & Sons.

Ehrenfeld JM. 2017. Wannacry, cybersecurity and health information technology: A time to act. *Journal of medical systems*. **41**(7) 104.

Everett C. 2016. Ransomware: To pay or not to pay. *Computer Fraud & Security*. **April** 8-12.

Farrell J. 1987. Cheap talk, coordination, and entry. *The RAND Journal of Economics*. **18**(1) 34-39.

Farrell J, Gibbons R. 1988. *Cheap talk, neologisms, and bargaining*. Massachusetts Institute of Technology, Cambridge, Mass.

Fbi. 2016. Incidents of ransomware on the rise University.

Fehr E, Gächter S. 2002. Altruistic punishment in humans. *Nature*. **415**(6868) 137.

Gal-Or E, Ghose A. 2005. The economic incentives for sharing security information. *Information Systems Research*. **16**(2) 186 - 208.

Gazet A. 2010. Comparative analysis of various ransomware virii. *Journal in Computer Virology*. **6**(1) 77-90.

Gordon LA, Loeb MP. 2002. The economics of information security investment. *ACM Transactions on Information and System Security (TISSEC)*. **5**(4) 438-457.

Griskevicius V, Goldstein NJ, Mortensen CR, Cialdini RB, Kenrick DT. 2006. Going along versus going alone: When fundamental motives facilitate strategic (non) conformity. *Journal of personality and social psychology*. **91**(2) 281-294.

Gupta A, Kannan K, Sanyal P. 2018. Economic experiments in information systems. *MIS Quarterly*. **42**(2) 595-606.

Hauser DJ, Schwarz N. 2016. Attentive turkers: Mturk participants perform better on online attention checks than do subject pool participants. *Behavior research methods*. **48**(1) 400-407.

Heal G, Kunreuther H. 2007. Modeling interdependent risks. *Risk Analysis*. **27**(3) 621-634.

Herath T, Rao HR. 2009. Protection motivation and deterrence: A framework for security policy compliance in organisations. *European Journal of Information Systems*. **18**(2) 106-125.

Ho TH, Zhang J. 2008. Designing pricing contracts for boundedly rational customers: Does the framing of the fixed fee matter? *Management Science*. **54**(4) 686-700.

Kalkanci B, Chen KY, Erhun F. 2011. Contract complexity and performance under asymmetric demand information: An experimental evaluation. *Management science*. **57**(4) 689-704.

Kannan K, Rahman MS, Tawarmalani M. 2016. Economic and policy implications of restricted patch distribution. *Management Science*. **62**(11) 3085-3391.

Kaspersky. 2016. *Story of the year: The ransomware revolution*. Kaspersky Lab.

Kharaz A, Arshad S, Mulliner C, Robertson W, Kirda E. 2016. *Unveil: A large-scale, automated approach to detecting ransomware*. USENIX, Austin, TX.

Kharraz A, Robertson W, Balzarotti D, Bilge L, Kirda E. 2015. *Cutting the gordian knot: A look under the hood of ransomware attacks*. IEEE, Milan, Italy

Kunreuther H, Heal G. 2003. Interdependent security. *Journal of Risk and Uncertainty*(26) 2.

Laszka A, Farhang S, Grossklags J. 2017. *On the economics of ransomware*. Springer, Vienna, Austria.

Lee YS, Seo YW, Siemsen E. 2018. Running behavioral operations experiments using amazon's mechanical turk. *Production and Operations Management*. **27**(5) 973-989.

Li S, Chen K-Y, Rong Y. 2019. The behavioral promise and pitfalls in compensating store managers. *Management Science (Forthcoming)*.

Lim N, Ho TH. 2007. Designing price contracts for boundedly rational customers: Does the number of blocks matter? *Marketing Science*. **26**(3) 312-326.

Liska A, Gallo T. 2017. *Ransomware: Defending against digital extortion*. O'Reilly Media, Inc, Sebastopol, CA.

Liu C-W, Gao G, Agarwal R. 2019. Unraveling the "social" in social norms: The conditioning effect of user connectivity. *Information Systems Research*(4) 1107-1452.

Luo X, Liao Q. 2007. Awareness education as the key to ransomware prevention. *Information Systems Security*. **16**(4) 195-202.

Mathews L. 2018. *Why you should never pay a ransomware ransom*.

Mckelvey RD, Palfrey TR. 1995. Quantal response equilibria for normal form games. *Games and economic behavior*. **10**(1) 6-38.

Mckelvey RD, Palfrey TR. 1998. Quantal response equilibria for extensive form games. *Experimental economics*. **1**(1) 9-41.

Metcalf AL, Angle JW, Phelan CN, Muth BA, Finley JC. 2019. More "bank" for the buck: Microtargeting and normative appeals to increase social marketing efficiency. *Social Marketing Quarterly*. **25**(1) 26-39.

Nolan JM, Kenefick J, Schultz PW. 2011. Normative messages promoting energy conservation will be underestimated by experts … unless you show them the data. *Social Influence*. **6**(3) 169-180.

Peer E, Vosgerau J, Acquisti A. 2014. Reputation as a sufficient condition for data quality on amazon mechanical turk. *Behavior research methods*. **46**(4) 1023-1031.

Rabin M. 1990. Communication between rational agents. *Journal of Economic Theory*. **51**(1) 144-170.

Rabin M. 1994. *Incorporating behavioral assumptions into game theory*. Springer.

Radware. 2018. *Global application & network security report 2017-2018*. Radware Ltd.

Sancho D. 2017. *Digital extortion: A forward-looking view*. Trend Micro Forward-Looking Threat Research (FTR) Team.

Scaife N, Carter H, Traynor P, Butler KRB. 2016. *Cryptolock (and drop it): Stopping ransomware attacks on user data*. IEEE, Nara, Japan

Shang J, Croson R. 2009. A field experiment in charitable contribution: The impact of social information on the voluntary provision of public goods. *The Economic Journal*. **119**(540) 1422-1439.

Simon HA. 1947. *Administrative behavior: A study of decision-making processes in administrative organization* Macmillan Publishers, New York, USA.

Sittig DF, Singh H. 2016. A socio-technical approach to preventing, mitigating, and recovering from ransomware attacks. *Appl Clin Inform*. **7**(2) 624–632.

Slonim R, Roth AE. 1998. Learning in high stakes ultimatum games: An experiment in the slovak republic. *Econometrica* 569-596.

Su X. 2008. Bounded rationality in newsvendor models. *Manufacturing & Service Operations Management*. **10**(4) 566-589.

Thakkar D. 2017. *Preventing digital extortion*. Packt Publishing, Birmingham.

Tsai H-T, Bagozzi RP. 2014. Contribution behavior in virtual communities: Cognitive, emotional, and social influences. *Management Information Systems Quarterly*. **38**(1) 143-163.

Tversky A, Kahneman D. 1974. Judgment under uncertainty: Heuristics and biases. *Science*. **185**(4157) 1124-1131.

Varian H. 2004. *System reliability and free riding*. Springer.

Verizon. 2020. *Verizon mobile security index 2020 report*. Verizon.

Wang J, Chaudhury A, Rao HR. 2008. A value-at-risk approach to information security investment. *Information Systems Research*. **19**(1) 106–120.

White K, Simpson B. 2013. When do (and don't) normative appeals influence sustainable consumer behaviors? *Journal of Marketing*. **77**(2) 78–95.

Wu DY, Chen KY. 2014. Supply chain contract design: Impact of bounded rationality and individual heterogeneity. *Production and Operations Management*. **23**(2) 253-268.

Yazdanmehr A, Wang J. 2016. Employees' information security policy compliance: A norm activation perspective. *Decision Support Systems*. **92**(December) 36-46.

Zhao X, Xue L, Whinston AB. 2013. Managing interdependent information security risks: Cyberinsurance, managed security services, and risk pooling arrangements. *Journal of Management Information Systems*. **30**(1) 123–152.

Appendix A. Proof of no asymmetric equilibrium in Case 2 and Case 3

The asymmetric equilibrium means that (Invest, Not Invest) or (Not Invest, Invest) is the equilibrium.

B.1. Case 2

We firstly assume asymmetric equilibrium (Invest, Not Invest) exists. For this equilibrium to exist, D1 need to choose Invest as his best response and D2 need to choose Not Invest as his best response. If Invest is the best response for D1, then the following condition need to be satisfied:

$$\begin{cases} v - I > v(1 - p_{NI}) \\ v - I > v(1 - \frac{1}{2}p_{NI}) \end{cases}$$

Solve the above formula we get the condition that Invest is the best response of D1:

$$\frac{I}{v} < \frac{1}{2}p_{NI}$$

If Not Invest is the best response for D2, then the following condition need to be satisfied:

$$\begin{cases} v - I < v(1 - p_{NI}) \\ v - I < v\left(1 - \frac{1}{2}p_{NI}\right) \end{cases}$$

Solve the above formula we get the condition that Not Invest is the best response for D2:

$$\frac{I}{v} > p_{NI}$$

Therefore, together we get that, in order to have (Invest, Not Invest) as the equilibrium, the following condition (3) and (6) need to be satisfied:

$$\begin{cases} \frac{I}{v} < \frac{1}{2}p_{NI} \\ \frac{I}{v} > p_{NI} \end{cases}$$

The above condition (3) and condition (6) are impossible to satisfy at the same time. So the asymmetric equilibrium (Invest, Not Invest) does not exist.

Using the similar process, we can prove that there is no asymmetric equilibrium (Not Invest, Invest) exits.

In conclusion, there is no asymmetric equilibrium that exists in case 2.

B.2. Case 3

We firstly assume asymmetric equilibrium (Not Invest, Invest) exists. For this equilibrium to exist, D1 need to choose Not Invest as his best response and D2 need to choose Invest as his best response. If Not Invest is the best response for D1, then the following condition need to be satisfied:

$$\begin{cases} v\left(1-\frac{1}{2}p_I\right) - I < v(1-p_{NI}) \\ v - I < v(1-\frac{1}{2}p_{NI}) \end{cases}$$

Solve the above formula we get the condition that Not Invest is the best response of D1:

$$\frac{I}{v} > p_{NI} - \frac{1}{2}p_I$$

If Invest is the best response for D2, then the following condition need to be satisfied:

$$\begin{cases} v\left(1-\frac{1}{2}p_I\right) - I > v(1-p_{NI}) \\ v - I > v(1-\frac{1}{2}p_{NI}) \end{cases}$$

Solve the above formula we get the condition that Invest is the best response for D2:

$$\frac{I}{v} < \frac{1}{2}p_{NI}$$

Therefore, together we get that, in order to have (Invest, Not Invest) as the equilibrium, the following condition (9) and (12) need to be satisfied:

$$\begin{cases} \frac{I}{v} > p_{NI} - \frac{1}{2}p_I \\ \frac{I}{v} < \frac{1}{2}p_{NI} \end{cases}$$

$$\because p_{NI} > p_I \therefore p_{NI} - \frac{1}{2}p_I > \frac{1}{2}p_{NI}$$

The above condition (9) and condition (12) are impossible to satisfy at the same time. So the asymmetric equilibrium (Not Invest, Invest) does not exist.

Using the similar process, we can prove that there is no asymmetric equilibrium (Invest, Not Invest) exits.

In conclusion, there is no asymmetric equilibrium that exists in case 3.

Appendix B. Equilibrium with a monetary penalty

In case 2, the equilibrium strategies for the defenders are either (invest-invest) or (no invest – no invest):

In the invest-invest case, the attacker never attacks, so we don’t have to consider the penalty situation.

In the no invest-no invest case, the attacker always attacks, and asks for a ransom. The affected defender needs to decide whether to pay the ransom or not.

D1 | D2	Invest	Not Invest
Invest	$v - I, v - I$	$v - I, v - p_{NI}(v - p_{pay(}v - r - f))$
Not Invest	$v - p_{NI}\big(v - p_{pay}(v - r - f)\big), v - I$	$v - \frac{1}{2}p_{NI}(v - p_{pay}(v - r - f))$

If $p_{pay} = 1$, which means $v - r - f \geq 0$, so $f \leq v - r$, the defender’s utility become

$$v - \frac{1}{2}p_{NI}(v)$$

If $p_{pay} = 0$, which means $v - r - f < 0$, so $f > v - r$, the defender’s utility become

$$\frac{1}{2}v$$

Since $p_{NI} \leq 1$, the affected defender should always pay the ransom if $f \leq v - r$.

Appendix C. MTurk Experiment instruction and quiz

Instruction-Background Information

This is an experiment in ransomware and protection investment. If you follow the instructions carefully and make good decisions, you may earn a considerable amount of money that will be paid to you in cash at the end of the experiment. You have already earned US$1 show-up fee for participating. You will earn *experimental dollars* during the experiments, and experimental dollars will be converted to US dollars at the end of the experiment with the following exchange rate.

1,000 experimental dollars = US$1

You will receive the show up fee ($1) and any additional earnings ONLY if you finish the experiment.

In this experiment, there are three players: Attacker, Defender 1, and Defender 2. During the experiment, you will be randomly assigned to be the Attacker, the Defender 1, or the Defender 2. You will play the same role for the entire experiment. In total, you are going to play 30 rounds. In the first round, you will be randomly paired up with other players to form a 1 attacker- 2 Defenders group to play the game. You will stay in the same group for the entire experiment.

Data Value

In each round, each Defender is given a "data value" of 100 experimental dollars. The defender will receive these 100 experimental dollars at the end of each period if this data value is not lost.

Ransomware Attack

In each round, the Attacker chooses one of three options: (a) attack Defender 1; (b) attack Defender 2; (c) do not attack. If the Attacker chooses to attack a Defender, he/she also decides a ransom to ask.

The Attacker's probability of being successful is 80%. A defender can reduce the probability to 30% by spending 30 experimental dollars to make a protection investment.

If the attack is successful, the affected Defender chooses whether to pay the ransom. If the Defender decides to pay, he/she will not lose his/her data value and the Attacker receives the ransom for the round. If the Defender decides NOT to pay the ransom, the Defender loses his/her data value and the Attacker receives nothing for the round.

If the Attacker decides not to attack, the Attacker receives a fixed payment of 40 experimental dollars for the round.

Protection Investment

Protection Investment can reduce the Attacker's probability of being successful from 80% to 30%. That is, if a Defender who made protection investment is attacked, the Attacker's probability of being successful is 30%. If a Defender who did not make protection investment is attacked, the Attacker's probability of being successful is 80%. If a Defender decides to invest, a cost of 30 experimental dollars will occur for the round.

You will be allowed to continue only if you pass the following quizzes.

Quiz Question 1

Pretending the following scenario happened for a particular round in the experiment:

Defender 1 decided not to make the protection investment.

Defender 2 decided to make the protection investment.

Attacker decided not to attack.

What is the experiment dollar payoff for the Defender 1? Answer: 100

What is the experiment dollar payoff for the Attacker? Answer: 40

Quiz Question 2

Pretending the following scenario happened for a particular round in the experiment:

Defender 1 decided not to make the protection investment.

Defender 2 decided to make the protection investment.

Attacker decided to attack Defender 1 and asked 55 as ransom.

It was a successful attack, and Defender 1 decided to pay the ransom.

What is the experiment dollar payoff for the Defender 1? Answer: 45

What is the experiment dollar payoff for the Attacker? Answer: 55

Quiz Question 3

Pretending the following scenario happened for a particular round in the experiment:

Defender 1 decided not to make the protection investment.

Defender 2 decided not to make the protection investment.

Attacker decided to attack Defender 2 and asked 60 as ransom.

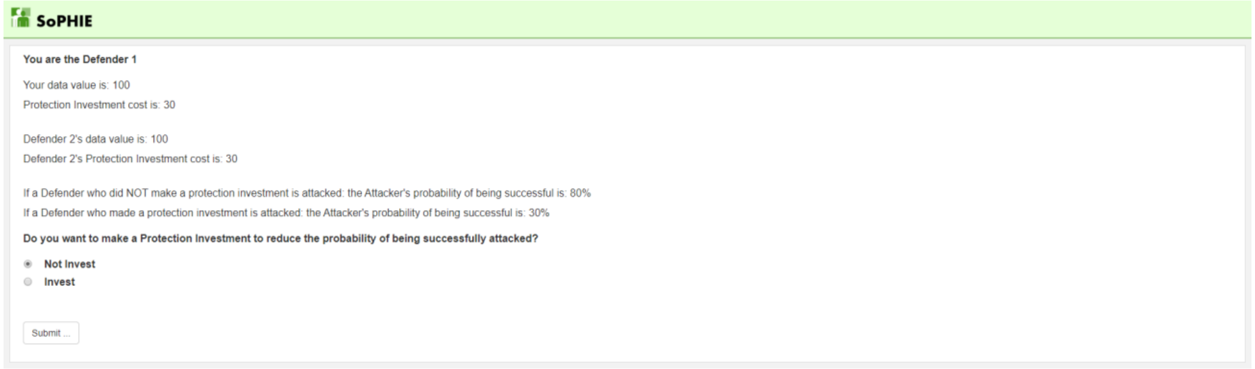
It was an unsuccessful attack.

What is the experiment dollar payoff for the Defender 2? Answer: 100

What is the experiment dollar payoff for the Attacker? Answer: 0

Appendix D. SoPHIE Screenshot (Baseline Treatment)

Defenders make investment decisions



You are the Defender 1

Your data value is: 100

Protection Investment cost is: 30

Defender 2's data value is: 100

Defender 2's Protection Investment cost is: 30

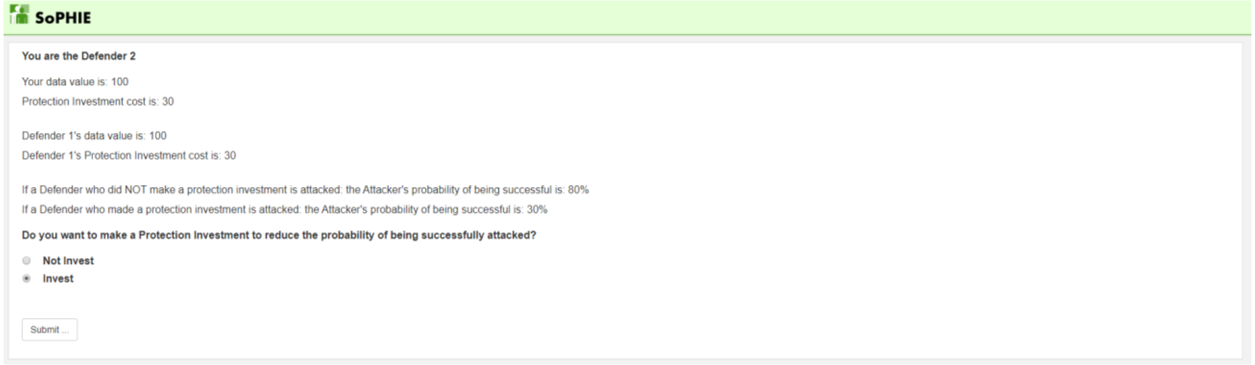
If a Defender who did NOT make a protection investment is attacked: the Attacker's probability of being successful is: 80%

If a Defender who made a protection investment is attacked: the Attacker's probability of being successful is: 30%

Do you want to make a Protection Investment to reduce the probability of being successfully attacked?

● **Not Invest**

○ **Invest**



You are the Defender 2

Your data value is: 100

Protection Investment cost is: 30

Defender 1's data value is: 100

Defender 1's Protection Investment cost is: 30

If a Defender who did NOT make a protection investment is attacked: the Attacker's probability of being successful is: 80%

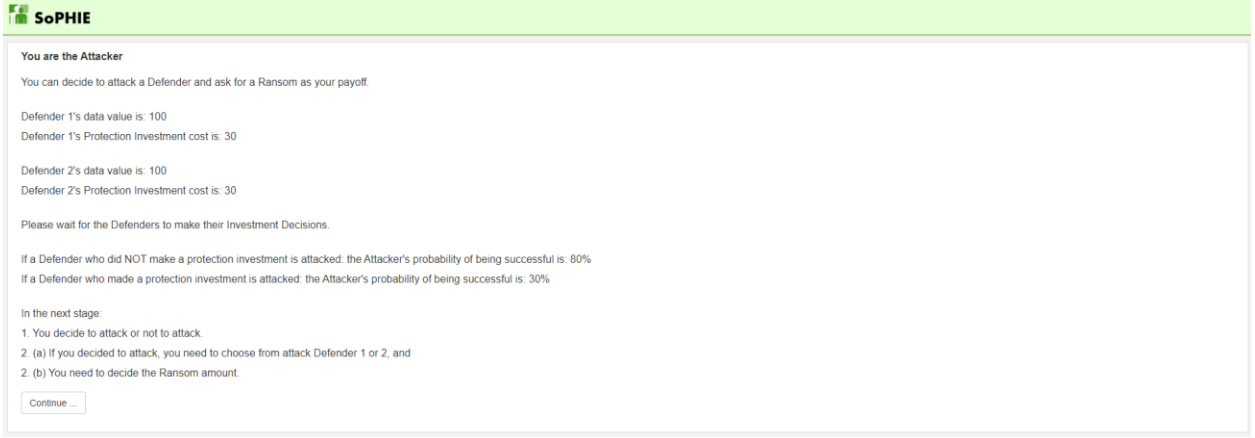
If a Defender who made a protection investment is attacked: the Attacker's probability of being successful is: 30%

Do you want to make a Protection Investment to reduce the probability of being successfully attacked?

○ **Not Invest**

● **Invest**

Attacker waits for defenders



You are the Attacker

You can decide to attack a Defender and ask for a Ransom as your payoff.

Defender 1's data value is: 100

Defender 1's Protection Investment cost is: 30

Defender 2's data value is: 100

Defender 2's Protection Investment cost is: 30

Please wait for the Defenders to make their Investment Decisions.

If a Defender who did NOT make a protection investment is attacked: the Attacker's probability of being successful is: 80%

If a Defender who made a protection investment is attacked: the Attacker's probability of being successful is: 30%

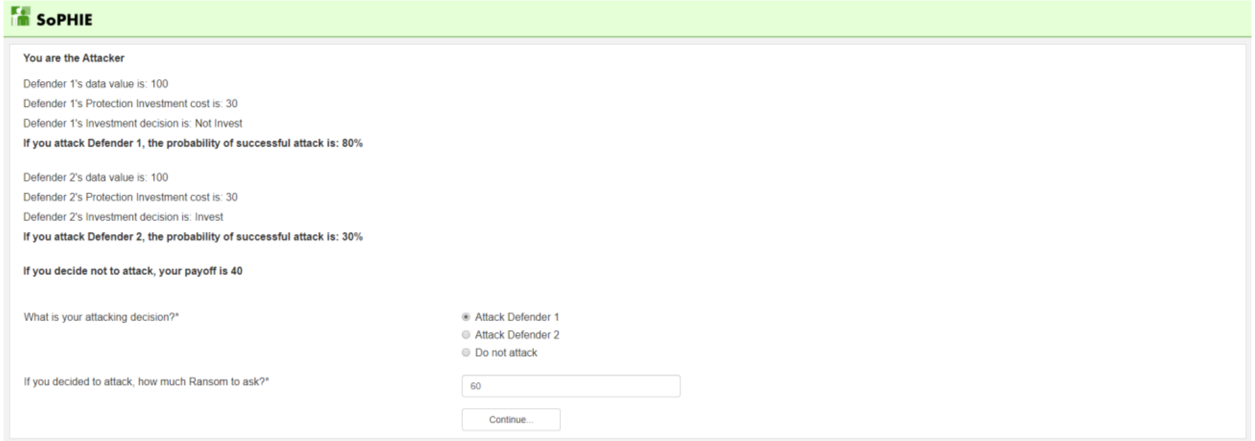
In the next stage:

1. You decide to attack or not to attack.

2. (a) If you decided to attack, you need to choose from attack Defender 1 or 2, and

2. (b) You need to decide the Ransom amount.

Attacker makes attack and ransom decisions



SoPHIE

You are the Attacker

Defender 1's data value is: 100

Defender 1's Protection Investment cost is: 30

Defender 1's Investment decision is: Not Invest

If you attack Defender 1, the probability of successful attack is: 80%

Defender 2's data value is: 100

Defender 2's Protection Investment cost is: 30

Defender 2's Investment decision is: Invest

If you attack Defender 2, the probability of successful attack is: 30%

If you decide not to attack, your payoff is 40

What is your attacking decision?*

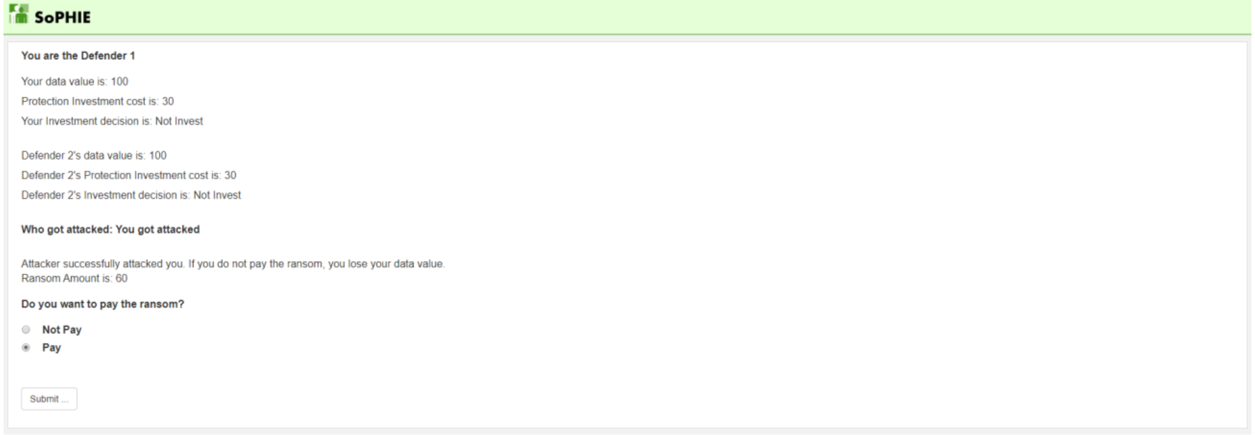
● Attack Defender 1
○ Attack Defender 2
○ Do not attack

If you decided to attack, how much Ransom to ask?*

60

Continue...

Affected defender makes payment decision



SoPHIE

You are the Defender 1

Your data value is: 100

Protection Investment cost is: 30

Your Investment decision is: Not Invest

Defender 2's data value is: 100

Defender 2's Protection Investment cost is: 30

Defender 2's Investment decision is: Not Invest

Who got attacked: You got attacked

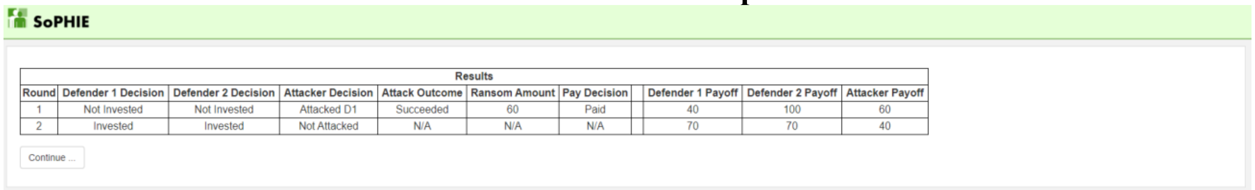
Attacker successfully attacked you. If you do not pay the ransom, you lose your data value.
Ransom Amount is: 60

Do you want to pay the ransom?

○ **Not Pay**
● **Pay**

Submit ...

End of a round: show profit



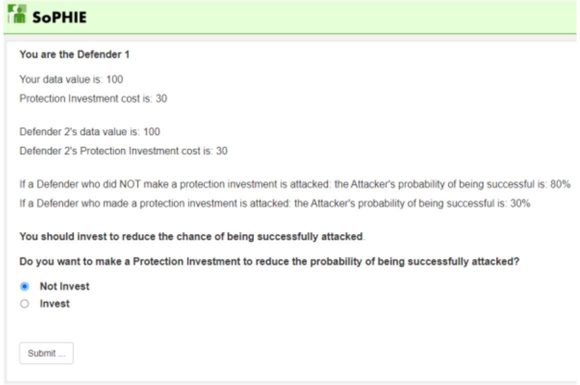
SoPHIE

Results										
Round	**Defender 1 Decision**	**Defender 2 Decision**	**Attacker Decision**	**Attack Outcome**	**Ransom Amount**	**Pay Decision**		**Defender 1 Payoff**	**Defender 2 Payoff**	**Attacker Payoff**
1	Not Invested	Not Invested	Attacked D1	Succeeded	60	Paid		40	100	60
2	Invested	Invested	Not Attacked	N/A	N/A	N/A		70	70	40

Continue ...

Injunctive Norm: Should Invest **Descriptive Norm: 73% Invest**

Defenders make investment decisions



SoPHIE

You are the Defender 1

Your data value is: 100

Protection Investment cost is: 30

Defender 2's data value is: 100

Defender 2's Protection Investment cost is: 30

If a Defender who did NOT make a protection investment is attacked: the Attacker's probability of being successful is: 80%

If a Defender who made a protection investment is attacked: the Attacker's probability of being successful is: 30%

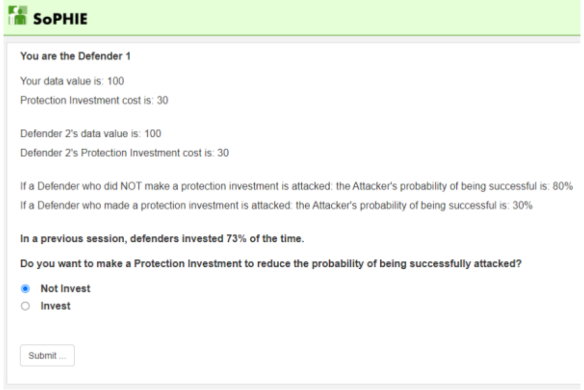
You should invest to reduce the chance of being successfully attacked.

Do you want to make a Protection Investment to reduce the probability of being successfully attacked?

● **Not Invest**

○ **Invest**

Submit ...



SoPHIE

You are the Defender 1

Your data value is: 100

Protection Investment cost is: 30

Defender 2's data value is: 100

Defender 2's Protection Investment cost is: 30

If a Defender who did NOT make a protection investment is attacked: the Attacker's probability of being successful is: 80%

If a Defender who made a protection investment is attacked: the Attacker's probability of being successful is: 30%

In a previous session, defenders invested 73% of the time.

Do you want to make a Protection Investment to reduce the probability of being successfully attacked?

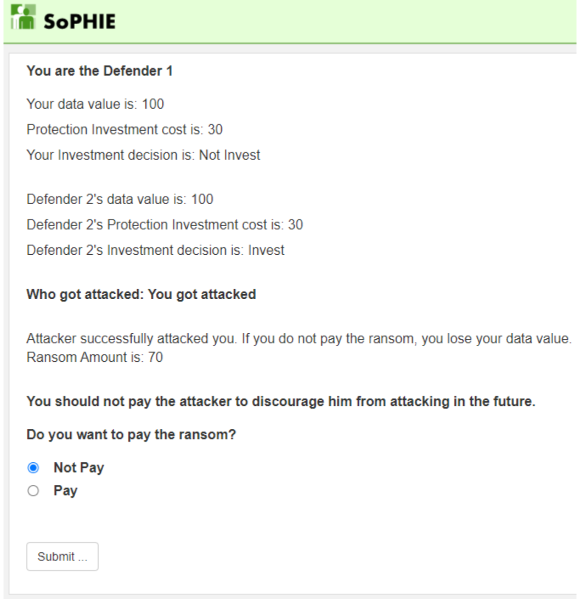
● **Not Invest**

○ **Invest**

Submit ...

Injunctive Norm: Should Not Pay **Descriptive Norm: 62% Not Pay**

Affected defender makes payment decision



SoPHIE

You are the Defender 1

Your data value is: 100

Protection Investment cost is: 30

Your Investment decision is: Not Invest

Defender 2's data value is: 100

Defender 2's Protection Investment cost is: 30

Defender 2's Investment decision is: Invest

Who got attacked: You got attacked

Attacker successfully attacked you. If you do not pay the ransom, you lose your data value.
Ransom Amount is: 70

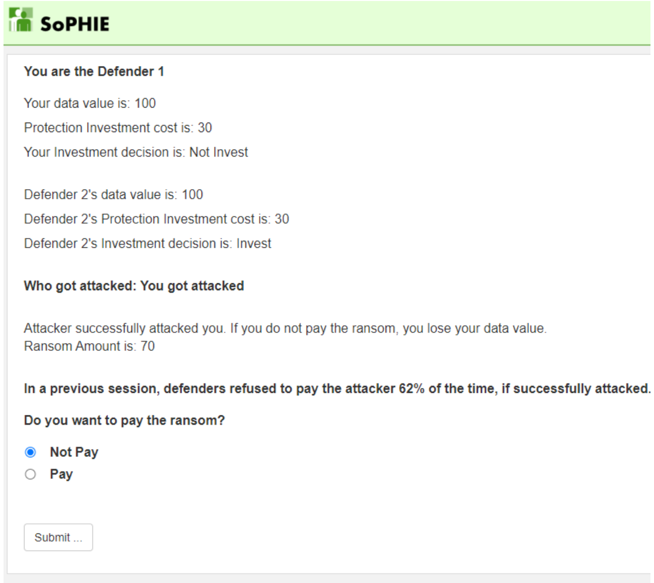
You should not pay the attacker to discourage him from attacking in the future.

Do you want to pay the ransom?

● **Not Pay**

○ **Pay**

Submit ...



SoPHIE

You are the Defender 1

Your data value is: 100

Protection Investment cost is: 30

Your Investment decision is: Not Invest

Defender 2's data value is: 100

Defender 2's Protection Investment cost is: 30

Defender 2's Investment decision is: Invest

Who got attacked: You got attacked

Attacker successfully attacked you. If you do not pay the ransom, you lose your data value.
Ransom Amount is: 70

In a previous session, defenders refused to pay the attacker 62% of the time, if successfully attacked.

Do you want to pay the ransom?

● **Not Pay**

○ **Pay**

Submit ...

Appendix E. Model Estimation with Injunctive and Descriptive Appeals

Injunctive Norms	**Should Invest**		**Should Not Pay**	
Parameter	Estimate	p-value	Estimate	p-value
γ_p: payment bounded rationality	0.0313	0.000	0.0023	0.000
γ_r: ransom bounded rationality	0.1404	0.000	0.1270	0.000
γ_a: attack bounded rationality	0.0685	0.000	0.0453	0.000
γ_i: investment bounded rationality	0.0914	0.000	0.1153	0.000
α_m: anchoring past (self)	31.5578	0.000	25.0615	0.000
α_y: anchoring past (other)	6.9687	0.000	11.0368	0.000
α_f: pay fairness	0.3536	0.000	9.0885	0.000

Descriptive Norms	**73% Invest**		**62% Not Pay**	
Parameter	Estimate	p-value	Estimate	p-value
γ_p: payment bounded rationality	0.0353	0.000	0.0186	0.000
γ_r: ransom bounded rationality	0.0833	0.000	0.1680	0.000
γ_a: attack bounded rationality	0.0883	0.000	0.0541	0.000
γ_i: investment bounded rationality	0.1299	0.000	0.1260	0.000
α_m: anchoring past (self)	19.2649	0.000	28.2915	0.000
α_y: anchoring past (other)	4.9174	0.000	5.4026	0.000
α_f: pay fairness	0.1011	0.014	0.7430	0.000

Appendix F. Defenders’ chat messages examples from Chat Treatment

```
837;"";"1532571875895";"Pa.46";"I WOULD INVEST IF I WERE YOU"
839;"";"1532571928634";"Pa.46";"MAYBE THEY ONLY ATTACK IF WE DONT INVEST"
913;"";"1532572471523";"Pa.4";"They never attack if we both invest"
1003;"";"1532573056300";"Pa.4";"...maybe you should invest this time"
890;"";"1532572349488";"Pa.4";"I'll invest"

184;"";"1532122735864";"Pa.59";"they might attack if we choose not to defend though"
165;"";"1532122494616";"Pa.18";"I'm going to just keep investing"
187;"";"1532122798962";"Pa.59";"haha yeah i say we stick with investing"
```

We omitted sentences containing only one or two words, such as “Hi”, “that’s good”, or any stopping words. We have a total of 622 lines of chat messages.

If a defender mentioned “inv”, “invest”, “investment”, or “30, 70” we count it as an investment message. We have 134 lines of investment messages, which represent 21.54% of total messages, many of defenders were talking about “I will or will not to invest”, and suggesting what the other defender should do, such as “you should invest”.

If a defender mentioned “pay”, “payment”, “ransom”, or “money”, we count it as a ransom payment message. We have 27 lines ransom payment messages, which represent 4.34% of total messages.

If a defender did not mention any keyword listed above, such as “investment” or “ransom”, we categorize it as a no key word message. We have 461 no key word messages, which represent 74.12% of total messages.

Appendix G. Alternative Interventions: Penalty and Chat

Summary Statistics – Penalty and Chat

Decision	Game Theory	Baseline	Penalty	Chat
Number of Groups	N/A	20	20	21
Investment Rate	100%	50.17% (28.87%)	60.08% (28.27)	57.38% (29.26%)
Attack Rate	0%	50.17% (40.82%)	38.00% (39.63%)	40.48% (41.07%)
Ransoms	100 85 for penalty	66.14 (12.48)	54.30 (10.80)	65.81 (14.14)
Payment Rate	100%	49.61% (25.19%)	63.32% (26.01%)	54.76% (34.29%)

Notes:

1. Standard deviations are reported in parentheses.
2. All results are significantly (p-value < 0.01) different from the game theory predictions.
3. Investment rate and attack rate are reported in group average across all periods.
4. Ransoms are reported in group average across all periods conditioned on attack decision.
5. Payment rate is reported in group average across all periods conditioned on successful attacks.

Decision Compare with Baseline

	Investment	Attack	Ransom	Payment
Baseline	50.17%	50.17%	66.14	50.39%
Penalty	60.08% (0.250)	38.00% (0.064)	54.30 (0.004)	63.32% (0.090)
Chat	57.38% (0.489)	40.48% (0.136)	65.81 (0.607)	54.76% (0.978)

Note: Mann-Whitney test, p -values are reported in parentheses

Model Estimation with Penalty and Chat

	Penalty		Chat	
Parameter	Estimate	p-value	Estimate	p-value
γ_p: payment bounded rationality	0.0316	0.000	0.0197	0.000
γ_r: ransom bounded rationality	0.1381	0.000	0.1622	0.000
γ_a: attack bounded rationality	0.0565	0.000	0.0640	0.000
γ_i: investment bounded rationality	0.0877	0.000	0.1315	0.000
α_m: anchoring past (self)	23.5013	0.000	22.2947	0.000
α_y: anchoring past (other)	15.6999	0.000	8.6884	0.000
α_f: pay fairness	0.4297	0.001	0.5359	0.000
Δ_f: penalty	7.6221	0.349		
Δ_i: chat invest			9.6521	0.000
Δ_p: chat pay			15.6932	0.280

Appendix H.

Baseline Model with Risk Aversion Estimation Result

Parameter	Estimate	p-value
γ_p: payment bounded rationality	0.0156	0.000
γ_r: ransom bounded rationality	0.1384	0.000
γ_a: attack bounded rationality	0.0459	0.000
γ_i: investment bounded rationality	0.3545	0.000
α_m: anchoring past (self)	21.4969	0.000
α_y: anchoring past (other)	6.5884	0.003
α_f: pay fairness	0.4435	0.000
θ_r: risk aversion	0.0157	0.049