Ransomware Negotiation: Dynamics and Privacy-Preserving Mechanism Design

Haohui Zhang[1], Sirui Shen[2], Xinyu Hu[1], and Chenglu Jin[2]

[1] University of Twente, Drienerlolaan 5, Enschede, The Netherlands
{h.zhang-5,x.hu}@utwente.nl
[2] Centrum Wiskunde & Informatica, Science Park 123, Amsterdam, The Netherlands
{sirui.shen,chenglu.jin}@cwi.nl

Abstract. Ransomware attacks have become a pervasive and costly form of cybercrime, causing tens of millions of dollars in losses as organizations increasingly pay ransoms to mitigate operational disruptions and financial risks. While prior research has largely focused on proactive defenses, the post-infection negotiation dynamics between attackers and victims remains underexplored. This paper presents a formal analysis of attacker–victim interactions in modern ransomware incidents using a finite-horizon alternating-offers bargaining game model. Our analysis demonstrates how bargaining alters the optimal strategies of both parties. In practice, incomplete information—attackers lacking knowledge of victims' data valuations and victims lacking knowledge of attackers' reservation ransoms—can prolong negotiations and increase victims' business interruption costs. To address this, we design a Bayesian incentive-compatible mechanism that facilitates rapid agreement on a fair ransom without requiring either party to disclose private valuations. We further implement this mechanism using secure two-party computation based on garbled circuits, thereby eliminating the need for trusted intermediaries and preserving the privacy of both parties throughout the negotiation. To the best of our knowledge, this is the first automated, privacy-preserving negotiation mechanism grounded in a formal analysis of ransomware negotiation dynamics.

Keywords: Ransomware · Negotiation · Mechanism Design · Garbled Circuits · Secure Two-Party Computation · Game Theory

1 Introduction

Ransomware has become one of the most severe cybersecurity threats, inflicting substantial financial and operational damage on individuals, corporations, and public institutions [8,18]. Modern strains go far beyond simple file encryption

Chenglu Jin is (partially) supported by project CiCS of the research programme Gravitation, which is (partly) financed by the Dutch Research Council (NWO) under the grant 024.006.037. We thank Dr. Jie Zhang for his feedback on a draft of this work.

[12], combining techniques such as credential theft and large-scale data exfiltration [1] to increase leverage over victims and amplify the pressure to pay. This evolution has made ransomware incidents not only more costly, but also strategically complex, as attackers and victims engage in high-stakes negotiations that determine both the ransom amount and the speed of recovery.

According to the Sophos 2025 Ransomware Report [23], 59% of organizations experienced a ransomware attack in 2024, with 70% of those incidents resulting in data encryption. The economic impact is stark: the median ransom payment surged fivefold from $400,000 in 2023 to $2 million in 2024, with the mean payment rising to $3.96 million. The average ransom demand also increases to $2.73 million, nearly $1 million more than the previous year.

When facing a ransomware attack, victims typically have two undesirable options: refuse to pay and risk permanent data loss and prolonged recovery, or pay the ransom and remain vulnerable to future attacks, with no guarantee that the data can actually be fully recovered. In practice, many victims choose to pay to minimize downtime and mitigate further financial losses [3]. According to the Sophos report, around 56% of organizations whose data is encrypted are able to recover their data by paying the ransom in 2024.

Business interruption costs often far exceed the ransom itself. Statista reports that the average downtime following a ransomware attack is 24 days, while the average cost per incident reached $1.85 million in 2023—a 13% increase over five years [24]. According to Intermedia [16], downtime-related losses—including significant data recovery costs, reduced customer satisfaction, missed deadlines, lost sales, and traumatized employees—could be even more damaging than the ransom payment. Consequently, in many cases, business interruption costs are the largest source of financial loss following a ransomware incident.

Negotiation, as one of the most important steps before the victim pays the ransom, has been largely overlooked in academic research. Few works explore the negotiation dynamics from a strategic perspective [3,10,14,17,26], and even fewer apply game-theoretic methods [20,22,27]. This paper fills in this gap by modeling the interactions as a finite-horizon alternating-offers bargaining game and conducting equilibrium analysis under the complete information scenario. With the goal of reducing business interruption costs for the victim, we further propose a bargaining strategy and its corresponding best response in the incomplete information setting. Inspired by this bargaining strategy profile, we design an automated negotiation mechanism that facilitates efficient agreement by providing the attacker with appropriate financial incentives, thereby accelerating the recovery process for the victim.

To protect sensitive information during this high-stakes process, we implement the negotiation mechanism via a secure two-party computation technique known as garbled circuits [2]. This approach eliminates the need for trusted intermediaries while ensuring that neither party reveals their private valuations. Even if the negotiation fails, the attacker learns nothing about the victim's internal data valuation, preserving strategic advantage and minimizing the risk of further exploitation.

Our Contributions In this paper, we first present a comprehensive multistage game model for modern ransomware attacks that explicitly incorporates negotiation dynamics. Our analysis demonstrates how bargaining rounds influence optimal strategies for both attackers and victims, establishing the conditions for subgame perfect Nash equilibrium (SPNE) under a complete information scenario and introducing a strategy profile in the incomplete information setting. Furthermore, we design a novel Bayesian incentive-compatible negotiation mechanism and implement our mechanism in a secure two-party computation protocol. Our privacy-preserving negotiation mechanism allows the victim and attacker to rapidly reach a mutually agreed ransom price while maintaining privacy. To the best of our knowledge, this paper presents the first automated negotiation mechanism implemented using secure two-party computation.

The remainder of this paper is organized as follows. Section 2 reviews major related works on the game-theoretic analysis of ransomware attacks. Section 3 models the interaction between a ransomware attacker and a victim, and presents related equilibrium analysis. Section 4 presents the details of our proposed mechanism and the corresponding protocol. Section 5 discusses the garbled circuit implementation of the proposed mechanism. Section 6 concludes the paper.

2 Related Work

Ransomware has been well studied as a multistage game to thoroughly model the diverse interactions between the attacker and the victim (or defender) over time. Prior research [5,6,18,19,31,33] has modeled various aspects of this adversarial interaction, including attack strategies, law enforcement involvement, backup policies, data-selling threats, and defense mechanisms. Unlike these works, our analysis deliberately excludes the attacker's targeting stage and the victim's backup stage. Instead, we focus on the post-infection negotiation phase, modeling how a victim's financial loss accumulates over time and analyzing the strategic bargaining processes. While some studies have assumed the absence of any negotiation or bargaining opportunity [19], in reality, the victims usually have the opportunity to negotiate and bargain, as in CONTI [26].

Pratical guidance on ransomware negotiation has emerged alongside limited academic research on the topic. As the ransomware threat has intensified, guidance on how to negotiate with attackers began emerging as early as 2016 [9]. Since then, many blog posts have been published on the internet by negotiators and cyber insurance providers for victims [4,28]. However, the academic study of ransomware negotiation remains relatively underexplored. One of the earliest formal explorations is by Hofmann [14], who outlines strategic negotiation practices based on his experience in cyber threat intelligence. Team Cymru [26] analyzes negotiation reports from the CONTI group, highlighting that attackers often assess victims' financial positions using public information and adjust their demands accordingly. Ryan et al. [22] focus on targeted ransomware negotiation and model it as an asymmetric non-cooperative two-player game, offering insights into optimal strategies under imperfect information. Similarly, Meurs et

al. [20] examine double extortion ransomware through a signaling game framework with double-sided information asymmetry. Their results show that when attackers lack precise knowledge of a victim's data valuation, their expected payoff decreases, thereby lowering ransom demands and discouraging escalation. Conversely, signaling high data value can lead to higher demands and increased leverage for the attacker. Finally, Boticiu and Teichmann [3] offer a comprehensive overview of ransomware negotiation procedures, drawing from operational insights to detail typical negotiation phases and highlight the importance of timing, information management, and disaster recovery plan.

Prior to this work, only three published studies have modeled ransomware negotiation as a bargaining game: one by Hernández-Castro et al. [13], one by Cartwright et al. [6] and one by Zhang et al. [32]; the last is currently released as a preprint. Hernández-Castro et al. [13] propose a static game model of ransomware attacks, evaluating attacker and victim payoffs under three pricing strategies: fixed-rate pricing, price discrimination, and bargaining. However, the authors directly refer to results from classic bargaining literature without developing a ransomware-specific bargaining model. Cartwright et al. [6] develop a dynamic game that incorporates bargaining within a broader framework, accounting for law enforcement intervention. They derive optimal strategies under both complete and incomplete information regarding victims' willingness-to-pay. Their findings suggest that a criminal's bargaining power increases with the threat of irrational aggression and is further enhanced by a credible commitment to decrypt files upon payment. This reputation-building dynamic aligns with our findings. However, the bargaining process itself is not modeled as a standalone mechanism in [6] but rather embedded within a larger strategic setting.

The work of Zhang et al. [32] is conducted independently and in parallel with ours. They introduce a dedicated Ransomware Bargaining Game (RGB) analysis framework that systematically examines attacker-victim negotiation in ransomware events. Their framework distinguishes between attacker types based on their attitude toward ransom and analyzes the convergence and equilibrium properties across three negotiation formats: one-round, multi-round, and continuing-round RBGs. In contrast to [13,32], our model does not rely on discounting future payoffs. Instead, we formulate attenuation based on the victim's financial losses accumulating over time. Furthermore, unlike [13,32], where the number of rounds is fixed or based on the victim's willingness-to-pay or the attacker's level of greed, the number of rounds in our framework is determined by the reservation values of both the attacker and the victim. Most importantly, distinct from all three prior works, our analysis places a stronger emphasis on the role of complete and incomplete information regarding these reservation values, which significantly shapes the negotiation outcome.

Before this work, only one paper [27] attempts to design a mechanism to facilitate negotiation between ransomware attackers and victims with the explicit goal of reducing business interruption costs. Vakilinia et al. [27] define two ransomware dilemma models, explore mechanism design approaches, and propose smart contract-based solutions to eliminate the need for a trusted third

Table 1. Notation Description

Notation	Description
$v \in \mathbb{R}_{>0}$	Actual value of the data owned by the victim
$c \in \mathbb{R}_{\geq 0}$	Cost of performing the attack and handling the data, with c_r representing the cost for releasing the files and c_d for deleting the files, $c_r > c_d \cong 0$
$r \in \mathbb{R}_{\geq 0}$	Ransom demand proposed by the attacker, with r_f the final ransom decided through negotiation, r_{min} the minimum ransom the attacker can accept ($r_{min} \gg c$), and r_{max} the maximum ransom the victim can pay
$\tau \in \mathbb{R}_{\geq 0}$	Trust level or reputation value of the attacker, with τ_g representing gain in trust and τ_l representing loss in trust, $\tau_l \cong \tau_g$
$\kappa \in \mathbb{R}_{>0}$	Credibility of threat posed by the attacker, with κ_g representing gain in credibility and κ_l representing loss in credibility, $\kappa_l \cong \kappa_g$

party while enabling atomic ransom-key exchanges. However, their framework prevents any agreement when the attacker's minimum acceptable price exceeds half of the victim's true valuation. Moreover, their model overlooks a critical dynamic: as time passes, a victim accumulates additional losses that effectively reduce her valuation of the encrypted data. This limitation results in their design not creating sufficient incentives for attackers to provide decryption keys rapidly. Additionally, like other ransomware-negotiation studies, their work overlooks the strategic importance of keeping victims' reservation values confidential during bargaining. This privacy protection is critical to maintaining negotiating leverage in situations where information asymmetries exist.

3 Game-Theoretic Analysis

In this study, we extend the ransomware multistage game model proposed by Caporusso et al. [5] to include reputation systems and negotiation. We emphasize the importance of reputation systems and the negotiation processes for both attackers and victims. We model the negotiation process as a finite-horizon alternating-offers bargaining game between an attacker and a victim, provide the SPNE based on backward induction under the assumption of complete information, and propose a strategy profile to urge the attacker to quickly reach an agreement with the victim under the assumption of incomplete information.

3.1 Multistage Game Model

The ransomware scenario can be modeled as a sequential, multistage game involving interactions between the attacker and victim. In this study, we identified three critical and fundamental stages in the process of a ransomware attack. The related notations are defined in Table 1.

(1) *Stage 1 - ransom requested.* The attacker infects the victim's computer systems and issues a ransom demand $r > 0$, typically delivered via email or a designated website.

(2) *Stage 2 - negotiation.* Having seen the demand r, the victim takes an action from **Pay (V1)**, **Don't pay (V2)**, and **Make a counteroffer (V3)**. Notably, almost all (genuine) ransomware strains enable some form of communication between the attacker and victims, allowing victims to make a counteroffer [11]. One key reason is that the attacker does not know exactly the actual value of the data v and the highest ransom r_{max} the victim can afford. Although the "irrational aggression" (attacker does not accept any counteroffer) can increase the credibility of threats κ posed by the attacker and may get a higher optimal ransom demand [6], it is not the attacker's interest to make a ransom demand that is not affordable by the victim. Therefore, bargaining is a key aspect of the ransomware game [5,6].

Suppose the victim chooses **V3**, then the attacker will evaluate the new offer and choose to **Accept (A1)**, **Don't accept (A2)**, or **Make a counteroffer (A3)**. If the attacker chooses **A2**, the victim needs to reconsider the attacker's last offer. Next, it is the victim's turn to react. This bargaining process continues until the victim chooses **V1** or **V2**, or the attacker does not have any patience to negotiate. Usually, if the attacker has no more patience, she will notify the victim and give the victim a last chance to pay. We denote the final amount paid by the victim as r_f, normally $r_f \leq r$.

(3) *Stage 3 - cooperation or defection.* If the victim pays the ransom, the attacker needs to choose between **Cooperate (A4)** (release the data with cost c_r) and **Defect (A5)** (delete the data with cost c_d). Note that random destruction can be equated with **A5**. If the victim does not choose to compromise, then the attacker needs to choose between **Release (A6)** and **Punish (A7)** (delete).

We define a ransomware game that skips negotiation stage and instead takes the final r_f as given. In this formulation, we introduce two reputation parameters for the attacker, the trust level τ and the credibility of the threat κ. These capture the effectiveness of any real-world reputation system, whether maintained by an insurer [25], an online forum or other intermediary mechanisms [26]. In the extreme case of a fully anonymous attacker, we have $\tau = 0$ and $\kappa \gg 0$. If the reputation system is imperfect, $\tau \approx 0$. The extensive-form representation of the game, along with the resulting payoff pairs, is shown in Figure 1. We can observe that the best possible outcome for the victims is to receive a 0 payoff. Since the attacker always acts in response to the victim's decisions, the victim is placed in a highly disadvantageous position with virtually no bargaining power [13].

3.2 Subgame Perfect Nash Equilibrium

Proposition 1. *If the ransomware game is a sequential game where $\tau \approx 0$ and $\kappa > 0$, there exists a unique subgame perfect Nash equilibrium* $(\mathbf{V2}, \mathbf{A7})$.

Proof. Given that there is not a valid reputation system, thus, $\tau_g, \tau_l \approx 0$ and $\kappa_l \cong \kappa_g > 0$. As $c_r > c_d$, through backward induction on the external-form game given in Figure 1, $(\mathbf{V2}, \mathbf{A7})$ is the unique SPNE. □

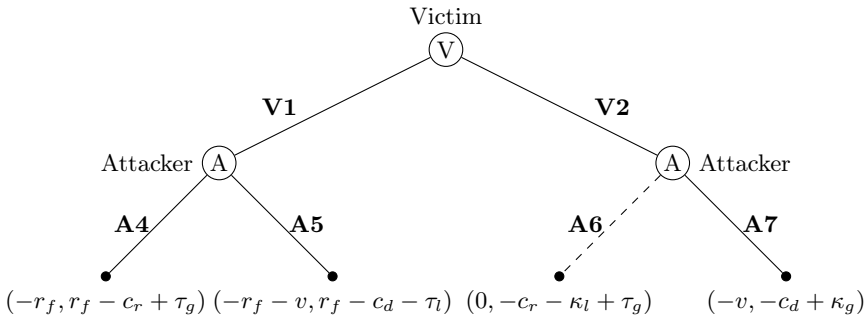

Fig. 1. Extensive Form of Ransomware Game with Perfect Reputation (The brackets at leaf node are (victim's payoff, attacker's payoff))

From Proposition 1, we can deduce that victims will only pay the ransom if they believe it offers a good chance of recovering their files, making it essential for attackers to establish a reputation system to build trust with rational victims.

An efficient reputation system with perfect accuracy can be implemented on a blockchain or maintained by authorities or reputable third parties such as insurance companies. As the victim's willingness to pay the ransom highly depends on the attacker's reputation, the increase and decrease of trust and credibility of the threat can directly affect the attacker's future revenue. A natural hypothesis posits that $\tau < \kappa$ and $\tau_l \cong \tau_g > c_r$ as in [5].

Proposition 2. *If the ransomware game is a sequential game where $\kappa > \tau > 0$ and $\tau_g + \tau_l > c_r$, if $v < r_f$ or $r_{max} < r_f$, then there exists a subgame perfect Nash equilibrium* $(\mathbf{V2}, \mathbf{A7})$*; if $r_f < v$ and $r_f < r_{max}$, then there exists a subgame perfect Nash equilibrium* $(\mathbf{V1}, \mathbf{A4})$.

Proof. Given that $\kappa > \tau > 0$, we have $\kappa_l + \kappa_g > \tau_g - c_r$. Thus, the attacker's best response to **V2** is **A7**. Since $\tau_l \cong \tau_g > c_r > c_d$, it follows that $-c_r + \tau_g > -c_d - \tau_l$, thus, the best response of the attacker to **V1** is **A4**. If $v < r_f$ or $r_{max} < r_f$, the victim is unable or unwilling to pay the ransom. Through backward induction, we conclude that (**V2, A7**) is the SPNE. If $v > r_f$ and $r_{max} > r_f$, the victim is able and willing to pay the ransom, we conclude that (**V1, A4**) is the SPNE. □

Note that if the attacker doesn't allow any counteroffer in stage 2, then r_f can only be r or 0. From Proposition 2, we obtain two salient points: (1) if the attacker wants to get the ransom from a rational victim successfully, it is necessary for them to be open to counteroffers to make sure that $r_f \leq \min(v, r_{max})$; and (2) the maximum ransom an attacker can get is $\min(v, r_{max})$.

3.3 Negotiation Dynamics

In the above formulation, we ignore the immediate financial loss due to downtime and the rate at which financial losses accumulate over time. To this end, we define

the total financial loss at any time t, denoted as

$$L(t) = L_0 + \int_0^t \ell(\delta)\, d\delta + \mathbb{I}(r_f > 0) \cdot r_f + \xi(t), \tag{1}$$

where $L_0 \in \mathbb{R}_{\geq 0}$ is defined as the immediate financial loss due to downtime. $t \in \mathbb{R}_{>0}$ is defined as the time elapsed since the ransomware attack began. $\ell(\delta)$ is defined as the loss rate at which financial losses accumulate over time with $\int_0^\infty \ell(\delta)\, d\delta = v$ and $\ell(\delta) \geq 0$ for all $\delta > 0$. $\mathbb{I}(\cdot)$ is a binary indicator function which $\mathbb{I}(r_f > 0)$ indicates whether the ransom is paid. $\xi(t)$ is a nonnegative, non-decreasing stochastic process, modeling unreimbursable or incalculable losses (e.g. reputation losses, legal costs). The non-decreasing and nonnegativity are defined to reflect the irreversible and potentially escalating nature of these intangible damages over time.

We assume that one bargaining round takes a fixed time T; in each round, one party—either the attacker or the victim—proposes a ransom, and the opposing party either accepts the offer or rejects it by proposing a counteroffer, thereby initiating a new round. Consequently, one complete back-and-forth exchange requires two rounds, and we denote the total number of rounds by N. In the simplest case, if the victim accepts the attacker's initial demand, the negotiation concludes in round 1. We assume that there is an efficient reputation system and that the attacker will always decrypt the data immediately if the victim pays the ransom. Under these assumptions, we can rationally model the dynamic evolution of the actual value of the encrypted data over time, which is

$$v(n) = \int_0^\infty \ell(\delta)\, d\delta - \int_0^{nT} \ell(\delta)\, d\delta. \tag{2}$$

We can observe that the actual value of v will gradually decrease with the increase of bargaining rounds n. The maximum ransom a victim can pay after n rounds of bargaining is $\min\{r_{max}, v(n)\}$, which is the victim's reservation value. r_{min} denotes the attackers' minimum acceptable ransom price where $r_{min} \gg c_r$.

Remark 1. Given that $\xi(t)$ is non-decreasing and the reputation system is efficient, the victim's optimal strategy is to conclude the negotiation as quickly as possible at the lowest acceptable ransom r_f.

The negotiation dynamics can be modeled as an N rounds alternating-offers bargaining game as shown in Figure 2. The victim's payoff function is $-L(t_f)$, where t_f denotes the time from the start of the ransomware attack to the time the attacker decrypts the data. The attacker's payoff is r_f. Any negotiated agreement within the range $\min\{r_{max}, v(n)\}$ and r_{min} at each round n is preferable to no agreement at the end. If no deal is reached, the attacker's payoff is 0 and the victim suffers an accumulated loss of $L(\infty)$. During bargaining analysis, we treat the ransom price and the victim's loss accumulation rate as the only variable parameters. Both agents aim to maximize their payoffs. Offers proposed in odd rounds are proposed by the attacker, and offers proposed in even rounds are proposed by the victim. Moreover, $r_{2k} < r_{2k-1}$ and $r_{2k} < r_{2k+1}$ for all $k \in \{1, \ldots, \lceil \frac{N-1}{2} \rceil\}$, as otherwise, one party will directly accept the offer.

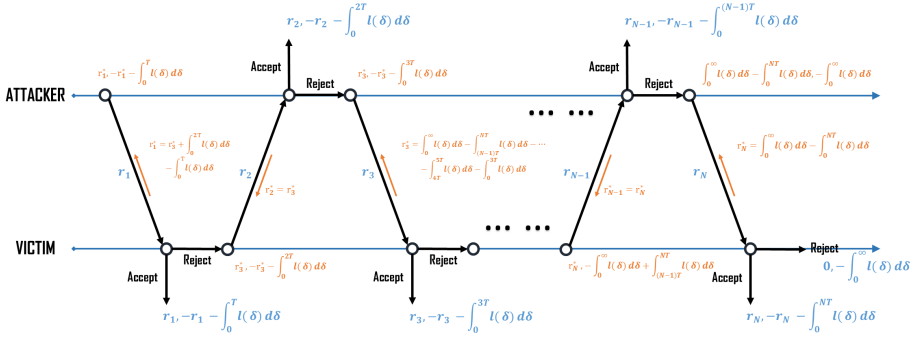

Fig. 2. N Rounds Alternating-offers Bargaining Game and Backward Induction

Complete Information

We first assume that both the reservation values of the attacker and the victim are complete information. This assumption, while a simplification of reality, can be justified as follows: r_{min} can be inferred from historical incidents involving the same ransomware group, and v and $\ell(\delta)$ can be estimated based on publicly available financial statements or confidential financial information stolen by the attacker. We also assume that the marginal loss incurred in the bargaining round is negligible compared to the cumulative future losses for all n, i.e. $\int_{(N-1)T}^{NT} \ell(\delta)\, d\delta \ll \int_{(N+1)T}^{\infty} \ell(\delta)\, d\delta$. Suppose N is the last bargaining round between the attacker and the victim. Then, $r_{min} < v(N)$ and $r_{min} > v(N+1)$, $N \in \mathbb{N}_{\geq 0}$. Without loss of generality, we assume $r_{max} \geq v(1)$.

Proposition 3. *If the reservation values are complete information, the subgame perfect equilibrium (the optimal offer given by the attacker and the victim in each round n where $n \in \{1, \ldots, N\}$ and N is an odd positive integer) is*

$$r_n^* = \int_0^{\infty} \ell(\delta)\, d\delta - \sum_{k=0}^{\lfloor \frac{N-n}{2} - 1 \rfloor} \int_{(N-1-2k)T}^{(N-2k)T} \ell(\delta)\, d\delta - \int_0^{(2\lfloor \frac{n}{2} \rfloor + 1)T} \ell(\delta)\, d\delta. \tag{3}$$

Proof. Given that $r_N^* = \int_0^{\infty} \ell(\delta)\, d\delta - \int_0^{NT} \ell(\delta)\, d\delta$, based on the backward induction in Figure 2 (orange part), we obtain $r_{N-1}^* = r_N^*$; $r_{N-2}^* = \int_0^{\infty} \ell(\delta)\, d\delta - \int_{(N-1)T}^{NT} \ell(\delta)\, d\delta - \int_0^{(N-2)T} \ell(\delta)\, d\delta$; ...; $r_2^* = r_3^*$; $r_1^* = \int_0^{\infty} \ell(\delta)\, d\delta - \int_{(N-1)T}^{NT} \ell(\delta)\, d\delta - \cdots - \int_{2T}^{3T} \ell(\delta)\, d\delta - \int_0^{T} \ell(\delta)\, d\delta$, which can be encoded into the closed-form in (3).

If N is an even positive integer, as through similar backward induction, it follows that $r_N^* = 0$ and $r_{N-1}^* = \int_{(N-1)T}^{NT} \ell(\delta)\, d\delta$. Given that $r_{min} > v(N+1) = \int_{(N+1)T}^{\infty} \ell(\delta)\, d\delta \gg \int_{(N-1)T}^{NT} \ell(\delta)\, d\delta$ based on the definition of $\ell(\delta)$, we obtain that $r_N^* \leq r_{N-1}^* < r_{min}$. Because the attacker will never propose a ransom less than r_{min}, this creates a contradiction. □

In practice, the victim can always propose a counteroffer back in round $N+1$, but because $r_{N+1} < r_N$ and $r_{min} > v(N+1)$, the attacker will always reject and end the negotiation. As the victim will never accept if the demand is greater than its reservation value, we define the optimal offer function $R : \{(n, N)|1 \leq n \leq N\} \longrightarrow \mathbb{R}$, mapping each pair (n, N) to a real-valued offer:

$$R(n, N) = r^*_{n|n\in\{1,\ldots,N\}}. \tag{4}$$

Lemma 1. *The optimal offer function R is monotonically non-increasing. That is $R(k+1, N) \leq R(k, N)$ for all $k \in \{1, \ldots, N-1\}$ and $R(n, N+2) \leq R(n, N)$ for all odd integer N and all $n \in \{1, \ldots, N\}$.*

Proof. Suppose that, contrary to the Lemma, there exists a $R(k+1, N) > R(k, N)$. Based on the definition of R and (3), we know $R(k+1, N) - R(k, N) = r^*_{k+1} - r^*_k$, if k is an even number, $r^*_{k+1} - r^*_k = 0$, contradicts, if k is odd,

$$r^*_{k+1} - r^*_k = -\int_0^{(k+2)T} \ell(\delta)\, d\delta + \int_0^{kT} \ell(\delta)\, d\delta = -\int_{kT}^{(k+2)T} \ell(\delta)\, d\delta > 0,$$

contradicts the definition of $\ell(\delta)$.

Given (3), as

$$\sum_{k=0}^{\lfloor \frac{N-n}{2}-1 \rfloor} \int_{(N-1-2k)T}^{(N-2k)T} \ell(\delta)\, d\delta \leq \sum_{k=0}^{\lfloor \frac{N+2-n}{2}-1 \rfloor} \int_{(N+1-2k)T}^{(N+2-2k)T} \ell(\delta)\, d\delta,$$

then, $R(n, N+2) \leq R(n, N+1) = R(n, N)$. □

Lemma 2. *$R(n, N) \leq v(n)$ for all odd integer N and $n \in \{1, \ldots, N\}$.*

Proof. Given (3), as $(2\lfloor \frac{n}{2} \rfloor + 1)T \geq nT$ and $\ell(\delta) \geq 0$, then,

$$\sum_{k=0}^{\lfloor \frac{N-n}{2}-1 \rfloor} \int_{(N-1-2k)T}^{(N-2k)T} \ell(\delta)\, d\delta + \int_0^{(2\lfloor \frac{n}{2} \rfloor+1)T} \ell(\delta)\, d\delta \geq \int_0^{nT} \ell(\delta)\, d\delta.$$

□

Remark 2. Given that the reservation values are complete information and $0 < r_{min} \leq r_{max}$, the attacker can get a ransom in round n if and only if $r_{min} \leq R(n, N)$. The possible maximum ransom the attacker can get is $R(1, N)$.

Therefore, if the attacker wants to get any ransom eventually, either the attacker needs to accept an offer proposed by the victim before round N, where N is an odd integer ($r_{min} \leq R(N, N)$ and $r_{min} > R(N+2, N+2)$), or the attacker proposes an offer no larger than $R(2k+1, N)$ in round $2k+1$ where $k \in \{0, \ldots, \lfloor \frac{N-1}{2} \rfloor\}$, since otherwise, the victim will never accept it. The optimal strategy profile is for the attacker to propose $R(1, N)$ in round 1, and for the victim to accept it in round 2.

Special Cases In some ransomware settings, there is no loss accumulation rate over time but rather a fixed loss if the data or files are never decrypted; in that case, the total cost remains constant unless a ransom r_f is paid. We define

$$L = L_0 + \mathbb{I}(a^A \notin \{\mathbf{A4}, \mathbf{A6}\}) \cdot v + \mathbb{I}(r_f > 0) \cdot r_f,$$

where a^A is the final action of the attacker. Suppose the attacker's reservation value is r_{min} where $r_{min} \gg c$. If $r_{min} > \min\{v, r_{max}\}$, the victim will never pay the ransom. When $r_{min} \leq \min\{v, r_{max}\}$ and both sides' reservation values are complete information, based on the unique equilibrium of the infinite-horizon alternating-offers bargaining game in [21] (the discount factors $\gamma^A = \gamma^V \to 1$), the attacker and the victim agree to set the ransom:

$$r_f \;=\; \frac{\min\{v,\, r_{\max}\} + r_{\min}}{2}.$$

Incomplete Information - Private Reservation Values

Suppose r_{min} is sufficiently small, which means the attacker can bargain with the victim forever, then $N \to \infty$. Suppose the loss rate $\ell(\delta)$ varies much more slowly than the window-length T, which means for all $n \in \mathbb{N}_{\geq 0}$ each block integral $\int_{nT}^{(n+1)T} \ell(\delta)\, d\delta$ changes negligibly from one block to the next, it follows that

$$\begin{aligned} r_1^* &= \lim_{N\to\infty} \int_0^\infty \ell(\delta)\, d\delta - \sum_{k=0}^{\lfloor \frac{N-3}{2} \rfloor} \int_{(N-1-2k)T}^{(N-2k)T} \ell(\delta)\, d\delta - \int_0^T \ell(\delta)\, d\delta \\ &= \sum_{m=0}^{\infty} \int_{(2m+1)T}^{(2m+2)T} \ell(\delta)\, d\delta \approx \frac{1}{2} \int_0^\infty \ell(\delta)\, d\delta. \end{aligned}$$

From Lemma 1, $R(1,\infty) = \min_{N\geq 1} R(1,N)$, the minimum ransom the victim pays at round 1 is $\min\left\{\frac{1}{2}\int_0^\infty \ell(\delta)\, d\delta,\, r_{max}\right\}$. Consequently, if the attacker's opening demand falls below this threshold, a good strategy for the victim is to accept it immediately to avoid additional bargaining rounds. Given that the attacker may salt prices to secure a better offer, when designing the victim's bargaining strategy, the victim needs to encourage the attacker to compromise as quickly as possible while ensuring that any attacker's deviation behavior from truthful play cannot improve her expected payoff. We define the truthful play of the attacker as *the attacker accepting the offer r_n in the round n whenever $r_{min} \leq r_n$.*

Proposition 4. *Suppose that the reservation values are incomplete information. Without loss of generality, we assume that $r_{max} \geq v(1)$. We denote $\tilde{r}_2 = q \cdot \int_0^\infty \ell(\delta)\, d\delta - \int_0^{2T} \ell(\delta)\, d\delta$. Given the strategy of the victim in round 2 and 4 as*

$$a_2^V = \begin{cases} \textbf{\textit{V1}} & \text{if } r_1 \leq \frac{1}{2}\int_0^\infty \ell(\delta)\, d\delta, \\ \textbf{\textit{V3}} \;\&\; r_2 = \tilde{r}_2 & \text{otherwise with prob } \overline{p}, \\ \textbf{\textit{V3}} \;\&\; r_2 = \int_{2T}^\infty \ell(\delta)\, d\delta & \text{otherwise with prob } 1-\overline{p}, \end{cases}$$

$$a_4^V = \begin{cases} \textbf{\textit{V1}} & \text{if } r_3 \leq \int_{3T}^\infty \ell(\delta)\, d\delta \text{ and with prob } \rho, \\ \textbf{\textit{V2}} & \text{otherwise}, \end{cases}$$

where $\frac{\int_0^{2T}\ell(\delta)\,d\delta}{\int_0^{\infty}\ell(\delta)\,d\delta} \le q \le \frac{\int_0^{2T}\ell(\delta)\,d\delta}{\int_0^{3T}\ell(\delta)\,d\delta}$, $\max\{0, \frac{q\cdot\int_0^{\infty}\ell(\delta)\,d\delta-\int_0^{2T}\ell(\delta)\,d\delta}{\int_{3T}^{\infty}\ell(\delta)\,d\delta}\} \le \rho \le 1$, *and* $\frac{\int_{2T}^{\infty}\ell(\delta)\,d\delta}{\rho\cdot\int_{3T}^{\infty}\ell(\delta)\,d\delta+(1-q)\cdot\int_0^{\infty}\ell(\delta)\,d\delta} \le \overline{p} \le 1$, *the best response of attacker in round* 3 *is*

$$a_3^A = \begin{cases} \boldsymbol{A1} & \text{if } r_{min} \le r_2, \\ \boldsymbol{A3}\ \&\ r_3 = \max\{\frac{r_2}{q}, r_{min}\} & \text{otherwise.} \end{cases}$$

Proof. When $r_2 = \tilde{r}_2$ and $r_{min} \le \tilde{r}_2$, the expected payoff of the attacker to act truthfully in round 3 is

$$\mathbb{E}\Big[u_3^A\Big(\mathbf{A1}\Big|r_2 = \tilde{r}_2 \ge r_{min}\Big)\Big] = (1-\overline{p}+\overline{p}\,q)\int_{2T}^{\infty}\ell(\delta)\,d\delta + \overline{p}\,(q-1)\int_0^{2T}\ell(\delta)\,d\delta.$$

If the attacker chooses **A3** (to make a counteroffer), the offer r_3 given by the attacker should be larger than r_2. As r_2 has $1-\overline{p}$ probability to be $\int_{2T}^{\infty}\ell(\delta)\,d\delta$, if so and the attacker acts untruthfully, $\mathbb{E}[u_3^A(\mathbf{A3}|r_2 = \int_{2T}^{\infty}\ell(\delta)\,d\delta)] = 0$. When $r_2 = \tilde{r}_2$, given that the victim will accept any offer below $\int_{3T}^{\infty}\ell(\delta)\,d\delta$ in round 4, then the maximum possible ransom the attacker can get is $\int_{3T}^{\infty}\ell(\delta)\,d\delta$. Thus,

$$\mathbb{E}\Big[u_3^A\Big(\mathbf{A3}\Big|r_2 = \tilde{r}_2 \ge r_{min}\Big)\Big] \le \overline{p}\cdot\rho\cdot\int_{3T}^{\infty}\ell(\delta)\,d\delta.$$

Given that

$$\rho \ge \frac{q\cdot\int_0^{\infty}\ell(\delta)\,d\delta - \int_0^{2T}\ell(\delta)\,d\delta}{\int_{3T}^{\infty}\ell(\delta)\,d\delta} \text{ and } \overline{p} \ge \frac{\int_{2T}^{\infty}\ell(\delta)\,d\delta}{\rho\cdot\int_{3T}^{\infty}\ell(\delta)\,d\delta + (1-q)\cdot\int_0^{\infty}\ell(\delta)\,d\delta},$$

we obtain

$$\mathbb{E}\big[u_3^A(\mathbf{A3}|r_2 = \tilde{r}_2 \ge r_{min})\big] \le \mathbb{E}\big[u_3^A(\mathbf{A1}|r_2 = \tilde{r}_2 \ge r_{min})\big].$$

The best response of the attacker in round 3 when $r_2 = \tilde{r}_2 \ge r_{min}$ is **A1** (to accept). Thus, if a rational attacker rejects in round 3, $r_{min} > \tilde{r}_2$.

Based on Remark 2, the attacker can only get a ransom when $r_{min} \le v(2) = \int_{2T}^{\infty}\ell(\delta)\,d\delta$ in round 2. When $r_{min} > v(2)$, the attacker's payoff will always be 0.

When $\tilde{r}_2 < r_{min} \le \int_{2T}^{\infty}\ell(\delta)\,d\delta = r_2$, the attacker's best response is **A1**, since $r_2 = v(2)$ (which exceeds $v(3)$) is the highest ransom the attacker can obtain.

When $r_2 = \tilde{r}_2 < r_{min} \le \int_{2T}^{\infty}\ell(\delta)\,d\delta$, the best response of the attacker is **A3**. Since $\frac{\int_0^{2T}\ell(\delta)\,d\delta}{\int_0^{\infty}\ell(\delta)\,d\delta} \le q \le \frac{\int_0^{2T}\ell(\delta)\,d\delta}{\int_0^{3T}\ell(\delta)\,d\delta}$, it follows that $\frac{r_2}{q} \le v(3)$. Therefore, the maximum offer can be proposed based on the attacker's current information is $\frac{r_2}{q}$ and the optimal counteroffer is $\max\{\frac{r_2}{q}, r_{min}\}$. Notably, when $v(3) < r_{min} \le v(2)$, the attacker will not get any ransom in this case. □

4 Privacy-preserving Negotiation Mechanism

In this section, under the assumption of private reservation values (incomplete information), we design a novel Bayesian incentive-compatible mechanism to help the victim and the attacker automatically bargain and quickly reach a consensus, while preserving the privacy of both sides through garbled circuits.

4.1 Mechanism Design

Since reaching an agreement quickly benefits both the attacker and the victim, our mechanism design builds on Proposition 4. By implementing it via garbled circuits that directly compute the final ransom, we eliminate negotiation delays and thus adopt a simplified strategy profile based on the proposition. We assume that $\xi(t) = 0$ for all t. We denote the victim's reservation value function based on the bargaining time t as

$$\psi(t) = \min\left\{ \int_t^{\infty} \ell(\delta)\, d\delta,\ r_{max} \right\}. \tag{5}$$

Definition 1 (Mechanism Design for Ransomware Negotiation). *Consider a negotiation mechanism $\mathcal{M}$ between a victim and an attacker, where each agent reports a type: the victim reports $\hat{\theta}^V \in \mathbb{R}_{\geq 0}$ and the attacker reports $\hat{\theta}^A \in \mathbb{R}_{\geq 0}$, representing their reservation values. Given the victim's strategy π^V:*

$$a_2^V = \begin{cases} \textbf{\textit{V3}}\ \&\ r_2 = q \cdot \hat{\theta}^V & \text{with prob } \overline{p}, \\ \textbf{\textit{V3}}\ \&\ r_2 = \hat{\theta}^V & \text{with prob } 1-\overline{p}, \end{cases}$$

$$a_4^V = \begin{cases} \textbf{\textit{V1}}\ (r_f = r_3) & \text{if } r_3 \leq \hat{\theta}^V \text{ and with prob } q, \\ \textbf{\textit{V2}}\ (r_f = 0, \sigma = 1) & \text{if } r_3 \leq \hat{\theta}^V \text{ and with prob } 1-q, \\ \textbf{\textit{V2}}\ (r_f = 0) & \text{otherwise}, \end{cases}$$

where $q, \overline{p} \in [0,1]$ and $\overline{p} \cdot (1-q) = \frac{1}{2}$, and the best response of the attacker π^A:

$$a_3^A = \begin{cases} \textbf{\textit{A1}}\ (r_f = r_2) & \text{if } \hat{\theta}^A \leq r_2, \\ \textbf{\textit{A3}}\ \&\ r_3 = \max\{\frac{r_2}{q}, \hat{\theta}^A\} & \text{otherwise}, \end{cases}$$

where r_f is the final ransom determined by the strategy profile $\boldsymbol{\pi} = (\pi^V, \pi^A)$, we define the ***allocation rule*** *as:*

$$\alpha^V = \alpha^A = \begin{cases} 1 & \text{if } r_f > 0 \text{ or } \sigma = 1, \\ 0 & \text{otherwise}, \end{cases}$$

and the corresponding ***payment rule*** *as:*

$$\beta^V = \beta^A = r_f.$$

Then, the mechanism $\mathcal{M}$ maps input types $(\hat{\theta}^V, \hat{\theta}^A)$ to outcome $(\boldsymbol{\alpha}, \boldsymbol{\beta})$. ($\overline{p}$ and q are exogenously chosen "coin-flip" biases.)

The mechanism assumes a quasilinear utility, just as other auction mechanisms. A victim with valuation θ^V receives a utility $u^V(\theta^V, \hat{\theta}^V) = (\alpha^V \cdot \theta^V - \beta^V)$ for reporting type $\hat{\theta}^V$, while an attacker with valuation θ^A receives a utility $u^A(\theta^A, \hat{\theta}^A) = (\beta^A - \alpha^A \cdot \theta^A)$ for reporting type $\hat{\theta}^A$. Notably, the attacker's valuation θ^A in this mechanism is r_{min}, and the victim's valuation θ^V is $\psi(t)$ at time t. We assume a common prior in which θ^A and θ^V are drawn independently and uniformly from a common uniform distribution over $[\underline{r}, \overline{r}]$. Each player knows her own type and holds the prior as her belief about the other's type.

Theorem 1. *Suppose $\xi(t) = 0$ for all t, and attacker's true valuation θ^A is drawn independently and uniformly from a common prior uniform distribution F over $[\underline{r}, \overline{r}]$, the mechanism $\mathcal{M}$ is Bayesian incentive-compatible.*

Proof. Given the input $\hat{\theta}^V$ of the victim at time t and input $\hat{\theta}^A$ of the attacker to the mechanism $\mathcal{M}$, the expected utility of the attacker is

$$\begin{aligned}&\mathbb{E}\big[u^A\big(\theta^A, \hat{\theta}^A; \hat{\theta}^V\big)\big] \\ &= \begin{cases} \overline{p}\,(q\hat{\theta}^V - \theta^A) + (1-\overline{p})\,(\hat{\theta}^V - \theta^A) & \text{if } \hat{\theta}^A \le q\hat{\theta}^V, \\ (1-\overline{p})\,(\hat{\theta}^V - \theta^A) + \overline{p}\,q\,(\hat{\theta}^V - \theta^A) + \overline{p}\,(1-q)\,(-\theta^A) & \text{if } q\hat{\theta}^V < \hat{\theta}^A \le \hat{\theta}^V, \\ 0 & \text{otherwise.} \end{cases}\end{aligned}$$

Fix the attacker's true type $\theta^A \ge 0$ and an arbitrary victim report $\hat{\theta}^V \ge 0$. We compare the attacker's utility when she reports $\hat{\theta}^A$ to that when she reports truthfully, $\hat{\theta}^A = \theta^A$.

Case 1: $\hat{\theta}^V < \theta^A$. Then $q\hat{\theta}^V < \theta^A$ as well, so in every branch $q\hat{\theta}^V - \theta^A < 0$ and $\hat{\theta}^V - \theta^A < 0$. Hence, the expected utility when $\hat{\theta}^A \le \hat{\theta}^V$ is nonpositive, and the expected utility when $\hat{\theta}^A > \hat{\theta}^V$ gives zero. Thus, truthful reporting $\hat{\theta}^A = \theta^A$ achieves this maximum.

Case 2: $q\hat{\theta}^V < \theta^A \le \hat{\theta}^V$. Here $\hat{\theta}^V - \theta^A \ge 0$ but $q\hat{\theta}^V - \theta^A < 0$.

$$\begin{aligned}\hat{\theta}^A \le q\hat{\theta}^V \quad &\Longrightarrow \quad \mathbb{E}[u^A] \;=\; \big((1-\overline{p}) + \overline{p}\,q\big)\,\hat{\theta}^V - \theta^A, \\ q\hat{\theta}^V < \hat{\theta}^A \le \hat{\theta}^V \quad &\Longrightarrow \quad \mathbb{E}[u^A] \;=\; \big((1-\overline{p}) + \overline{p}\,q\big)\,\hat{\theta}^V - \theta^A.\end{aligned}$$

In either event, the payoff ties the maximum value, while any $\hat{\theta}^A > \hat{\theta}^V$ yields 0.

Case 3: $\theta^A \le q\hat{\theta}^V$. Now $\hat{\theta}^V - \theta^A \ge q\hat{\theta}^V - \theta^A \ge 0$, same to Case 2, the utility equals $\big((1-\overline{p}) + \overline{p}\,q\big)\,\hat{\theta}^V - \theta^A$ no matter $\hat{\theta}^A < q\hat{\theta}^V$ or $q\hat{\theta}^V < \hat{\theta}^A \le \hat{\theta}^V$, while any $\hat{\theta}^A > \hat{\theta}^V$ yields zero.

In all three cases, $\hat{\theta}^A = \theta^A$ weakly dominates any other report. Hence, truth-telling is a (weakly) dominant strategy for the attacker.

For all t, the interim expected utility of the victim averaged over the distribution F of attacker's report is,

$$\begin{aligned}\mathbb{E}_{\hat{\theta}^A \sim F}\Big[\mathbb{E}\big[u^V\big(\theta^V, \hat{\theta}^V; \hat{\theta}^A\big)\big]\Big] &= \mathbb{P}[\hat{\theta}^A \le q\hat{\theta}^V] \cdot \big(\overline{p}\,(\theta^V - q\hat{\theta}^V) + (1-\overline{p})\,(\theta^V - \hat{\theta}^V)\big) \\ &+ \mathbb{P}[q\hat{\theta}^V < \hat{\theta}^A \le \hat{\theta}^V] \cdot \big((1-\overline{p})\,(\theta^V - \hat{\theta}^V) + \overline{p}\,q\,(\theta^V - \hat{\theta}^V) + \overline{p}\,(1-q)\,\theta^V\big).\end{aligned}$$

Denote by $F(\hat{\theta}^V)$ the cumulative distribution function of the random variable $\hat{\theta}^A$, $F(x) = \mathbb{P}(\hat{\theta}^A \le x)$. Then,

$$\begin{aligned}\mathbb{E}_{\hat{\theta}^A \sim F}\Big[\mathbb{E}\big[u^V\big(\theta^V, \hat{\theta}^V; \hat{\theta}^A\big)\big]\Big] &= F(q\hat{\theta}^V)\,\overline{p}\,(1-q)\,\theta^V + F(\hat{\theta}^V)\,(1-\overline{p}+\overline{p}\,q)\,\theta^V \\ &- F(\hat{\theta}^V)\,(1-\overline{p}+\overline{p}\,q)\,\hat{\theta}^V + \big(F(\hat{\theta}^V) - F(q\hat{\theta}^V)\big) \cdot \overline{p}\,(1-q)\theta^V.\end{aligned}$$

Since in equilibrium the attacker reports truthfully, $\hat{\theta}^A = \theta^A$. By assumption, θ^A is independent of $(\theta^V, q, \overline{p})$ and uniformly drawn from F, which is Unif$[\underline{r}, \overline{r}]$, by the monotone linear change of variables $U = (\theta^A - \underline{r})/(\overline{r} - \underline{r})$, we may, without loss of generality, normalize the support to $[0, 1]$. Under this normalization, $F(u) = u$ and $F'(u) = f(u) = 1$ for $u \in [0, 1]$. Then, the partial derivative is

$$\begin{aligned}\frac{\partial \mathbb{E}[u^V]}{\partial \hat{\theta}^V} =& f(q\hat{\theta}^V)\, q\, \overline{p}\, (1-q)\, \theta^V + f(\hat{\theta}^V)\, (1 - \overline{p} + \overline{p}\, q)\, \theta^V - f(\hat{\theta}^V)\, (1 - \overline{p} + \overline{p}\, q)\, \hat{\theta}^V \\ & - F(\hat{\theta}^V)\, (1 - \overline{p} + \overline{p}\, q) + \big(f(\hat{\theta}^V) - f(q\hat{\theta}^V)\, q\,\big)\, \overline{p}\, (1-q)\, \theta^V \\ =& \overline{p}\, (1-q)\, \theta^V + (1 - \overline{p} + \overline{p}\, q)\, \theta^V - 2\, (1 - \overline{p} + \overline{p}\, q) \hat{\theta}^V.\end{aligned}$$

Setting the derivative to zero yields the unique interior maximizer

$$\hat{\theta}^V = \frac{1}{2 \cdot (1 - \overline{p} + \overline{p}\, q)}\, \theta^V.$$

Using the condition $\overline{p} \cdot (1-q) = \frac{1}{2}$, we obtain $\hat{\theta}^V = \theta^V$. As the second derivative of the expected utility is negative, truthfully reporting $\hat{\theta}^V = \theta^V$ maximizes the victim's (Bayesian) expected utility, which constitutes a Bayes–Nash equilibrium. We can conclude that the negotiation mechanism $\mathcal{M}$ holds BIC. □

In our ransomware-negotiation mechanism, the only condition needed for Bayesian incentive compatibility is $\overline{p} \cdot (1 - q) = \frac{1}{2}$, where $q, \overline{p} \in [0, 1]$. If the victim reports their type truthfully, then whenever the attacker's report satisfies $\hat{\theta}^A \leq \hat{\theta}^V$, her expected payment is

$$(1 - \overline{p} + \overline{p}\, q) \cdot \hat{\theta}^V = (1 - \overline{p} + \overline{p}\, q) \cdot \psi(t) = \frac{1}{2}\psi(t).$$

Solving $\overline{p}(1 - q) = \frac{1}{2}$ for q gives

$$q \;=\; 1 - \frac{1}{2\overline{p}}, \quad \overline{p} \in \left[\tfrac{1}{2}, 1\right],$$

or equivalently for $\overline{p}$ in terms of q,

$$\overline{p} \;=\; \frac{1}{2(1-q)}, \quad q \in \left[0, \tfrac{1}{2}\right].$$

Compared with the double-sided-blind auction proposed in [27], which only guarantees that the victim pays $\frac{1}{2}\psi(t)$ when $r_{min} \leq \frac{1}{2}\psi(t)$, our negotiation mechanism ensures an expected payment of $\frac{1}{2}\psi(t)$ both when $r_{min} \leq \frac{1}{2}\psi(t)$ and when $\frac{1}{2}\psi(t) < r_{min} < \psi(t)$.

4.2 Privacy-Preserving Negotiation Protocol

We implement mechanism $\mathcal{M}$ using a maliciously secure two-party computation protocol, enabling the victim and the attacker to reach an agreement without

revealing their private information. The design of our protocol ensures fairness between the two parties and efficiency of the computation without requiring trust in the other party or third parties.

Our protocol utilizes garbled circuits as its backend technology [2]. A garbled circuit scheme involves two parties: a garbler and an evaluator. When running the garbling scheme, the garbler first takes a Boolean circuit f, representing the function to be evaluated jointly, produces a garbled circuit $\mathcal{C}$, and sends $\mathcal{C}$ to the evaluator. The garbler also generates labels representing the encrypted form of all possible values on all wires in $\mathcal{C}$. Hereafter, $[\cdot]$ denotes the garbled/encrypted variables. After the circuit and labels are generated, the garbler encodes its input I_g to its garbled/encrypted label $[I_g]$, while the evaluator obtains the labels $[I_e]$ corresponding to its input I_e by running Oblivious Transfer (OT) with the garbler. According to the security properties of garbled circuits, given input labels $[I]$, the evaluator cannot infer or manipulate the value of I in the function evaluation [2]. Then the evaluator computes $\mathcal{C}$ with input labels, gets the output label $[O]$, and extracts the output $O = f(I_g, I_e)$. The garbler can additionally verify if $[O]$ corresponds to O to guarantee the integrity of the evaluation done by the evaluator. Extending from the simple garbling scheme, authenticated garbling [30] enables maliciously secure two-party computation with additional authentication information, meaning both parties should follow the protocol honestly; otherwise, any behavior deviating from the protocol will be detected by the other party. We briefly discuss our protocol below, and the complete protocol is in Fig. 3.

In our protocol, V and A act as the garbler and evaluator, respectively. First, V and A need to agree on an exchange time t_e and a strategy profile $\boldsymbol{\pi}$ including q and $\overline{p}$. The victim's true valuation is based on t_e where $\theta^V = \psi(t_e)$. V generates a garbled circuit $\mathcal{C}$ according to $\mathcal{M}$ and $\boldsymbol{\pi}$, and sends $\mathcal{C}$ to A. Next, V and A each generate two k-bit uniform random strings (S_0^V, S_1^V) and (S_0^A, S_1^A). V directly encodes its inputs, then sends $[S_0^V]$, $[S_1^V]$, and $[\hat{\theta}^V]$ to A. A obtains $[S_0^A]$, $[S_1^A]$, and $[\hat{\theta}^A]$ via OT from V. Subsequently, A evaluates $\mathcal{C}$ and extract the final ransom r_f. Finally, A and V respectively verify the correctness of $\mathcal{C}$ and $[r_f]$, ensuring neither party tampered with the computation before agreeing on r_f.

Due to the probabilistic nature of a_2^V and a_4^V, the input to $\mathcal{C}$ in our protocol must include randomness. To execute the probabilistic choices privately and fairly, we determine the branch by comparing two k-bit uniformly distributed strings S_0 and S_1 with thresholds corresponding to probabilities $\overline{p}$ and q. Here $S_0 = S_0^V \oplus S_0^A$ and $S_1 = S_1^V \oplus S_1^A$. Taking S_0 as an example, since S_0^V and S_0^A are private information provided by V and A, respectively, neither of them would know the value of S_0. Note that V or A may attempt to choose S_0^V or S_0^A adversarially rather than uniformly, hoping to gain an advantage in the negotiation. However, this random number generation scenario resembles the classic matching pennies game, where the unique Nash equilibrium is to choose a uniform random value for each player. This method ensures that the probabilistic choices within the circuit remain uninfluenced by either party, thereby guaranteeing the fairness and privacy for both V and A.

Protocol:

1. V and A agree on a strategy profile $\boldsymbol{\pi}$.
2. V generates a garbled circuit $\mathcal{C}$ according to $\mathcal{M}$ and the strategy profile $\boldsymbol{\pi}$ and sends $\mathcal{C}$ to V. V also generates the labels corresponding to $\mathcal{C}$.
3. V and A prepares the input for $\mathcal{C}$ as follows:
   (a) V generates two k-bit uniform random strings S_0^V and S_1^V. V sends $[S_0^V]$, $[S_1^V]$, and $[\hat{\theta}^V]$ to A.
   (b) A generates two k-bit uniform random strings S_0^A and S_1^A. A runs an OT protocol with V to obtain $[S_0^A]$, $[S_1^A]$, and $[\hat{\theta}^A]$.
4. A evaluates $[r_f] := \mathcal{C}([S_0^V], [S_1^V], [S_0^A], [S_1^A], [\hat{\theta}^V], [\hat{\theta}^A])$ and extracts r_f. A checks whether $\mathcal{C}$ corresponds to $\mathcal{M}$ and $\boldsymbol{\pi}$. If not, abort the negotiation; otherwise, take r_f as the outcome of the negotiation.
5. A sends $[r_f]$ to V. V checks the validity of $[r_f]$. If not, abort the negotiation; otherwise, extract the outcome of the negotiation r_f.

Fig. 3. A Privacy-Preserving Two-Party Negotiation Protocol

Once the result is checked by both parties, A and V reach consensus on the ransom value r_f generated by the circuit. The ransom-key exchange can then be executed either through a trusted third party, such as an insurance company, or via a smart contract on a blockchain, as proposed in [27]. Specifically, the smart contract releases the ransom only after the victim deposits both the ransom r_f and earnest money, the attacker submits the decryption key to the contract, and the victim verifies its correctness. Upon confirmation, the ransom is transferred to the attacker, and the earnest money is refunded to the victim. This process can be regarded as a form of reputation system recording. Moreover, insurance companies can also serve as a reputation system, as cyber insurers often track behavior of ransomware groups, including negotiation frequency, duration, and reliability of decryption after payment [25].

As $\psi(t)$ is non-increasing, our mechanism gives the attacker financial incentives to quickly return the key. Therefore, our mechanism safeguards the privacy of both parties, incentivizes attackers to cooperate quickly with victims, and ensures a fair resolution for both sides.

5 Garbled Circuit Implementation

We implemented our secure two-party protocol on the TinyGarble2 framework to instantiate our proposed mechanism [1]. TinyGarble2, utilizing EMP-tool [29] as its backend, is an efficient `C++` framework for garbled-circuit-based S2PC in both semi-honest and malicious adversary models [15]. The garbled-circuit implementation includes both parties inputting their respective $\hat{\theta}^V$ and $\hat{\theta}^A$, processing probabilistic choices, and finally reaching consensus on the ransom r_f.

[1] The implementation is accessible to the public in the GitHub repository `https://GitHub.com/NomadShen/TinyGarble2.0`.

Table 2. Execution Time Evaluation Results

k_θ	k	**Execution time**
8	8	18.96 ms
8	16	32.56 ms
8	32	41.02 ms
16	8	34.95 ms
16	16	40.25 ms
16	32	52.18 ms

5.1 Implementation

Our implementation assumes V and A have already agreed on a strategy profile $\boldsymbol{\pi}$. The circuit has 3 inputs from V, which are S_0^V, S_1^V, and $\hat{\theta}^V$. Correspondingly, the values that A inputs into the circuit include S_0^A, S_1^A, and $\hat{\theta}^A$. Consistent with Step 4 of the Protocol in Fig. 3, the circuit's output is the garbled form of the ransom value $[r_f]$, but the plaintext value r_f is extractable by both parties from $[r_f]$ to ensure the result is fairly distributed to both parties.

For probabilistic choice, we precompute $\overline{p}_{scale} = \lfloor \overline{p} \cdot 2^k \rfloor$, and then complete the probabilistic choice simply through comparison between $\overline{p}_{scale}$ and S_0. Additionally, for the rejection in $\mathcal{M}$, the circuit assigns r_f a value of 0.

The multiplication and division operations involved in $\mathcal{M}$ include $q\hat{\theta}^V$ and r_2/q, where $q \in (0, 1/2]$. In practice, q is typically not very close to zero, so we precompute the values of q and $\frac{1}{q}$ and perform scaling as $\overline{p}_{scale}$. Similarly, we compute $q \cdot \hat{\theta}^V = q_{scale} \cdot \hat{\theta}^V / 2^k$ and $r_2/q = r_2 \cdot (\frac{1}{q})_{scale}/2^k$. Here, division by 2^k is implemented as a right-shift operation. We set the bitwidth of the input $\hat{\theta}$ to k_θ and the bitwidth of the randomness to k.

5.2 Evaluation

We evaluate the execution time of our protocol implemented on the TinyGarble2 framework. We run V and A's programs on the same computer using two separate threads, with the two parties communicating by the TCP protocol. Evaluation is performed on an Intel Core i7-1250U CPU with 8GB RAM.

In the experiment, we evaluate the protocol with different bitwidths of k_θ and k, which affect the precision of probabilities and multiplications. Table 2 shows the execution time under different input bitwidths. The results show that execution time increases with input bitwidth, but all computations complete within 53 ms, which is well within practical online negotiation requirements compared to the human negotiation process.

6 Conclusion

This paper presents a game-theoretic and privacy-preserving approach to ransomware negotiation, a critical yet underexplored phase of ransomware incidents.

We first highlight the importance of negotiation and reputation systems to both attackers and victims through multistage game-theoretical analysis. Then, we model the ransomware negotiation process as a finite-horizon alternating-offers bargaining game. Our analysis captures how strategic behavior influences outcomes in complete and incomplete information settings. To operationalize our findings, we design a Bayesian incentive-compatible negotiation mechanism that allows both parties to reach an agreement quickly while preserving privacy. Our implementation using garbled circuits enables efficient and secure negotiation without revealing sensitive data. To the best of our knowledge, this is the first work to integrate a formal bargaining model with a privacy-preserving, automated negotiation mechanism tailored to ransomware. By combining theoretical insights with practical cryptographic implementation, our work provides both a deeper understanding of ransomware negotiation dynamics and a viable pathway for more secure and efficient post-infection response strategies.

References

1. Arctic Wolf Threat Report: 96 Percent of Ransomware Cases Included Data Theft as Cybercriminals Double Down on Extortion (2025) `https://shorturl.at/JVWji`, last accessed 2025/8/12
2. Bellare, M., Hoang, V.T. and Rogaway, P.: Foundations of garbled circuits. In Proceedings of the 2012 ACM conference on Computer and communications security, pp. 784-796 (2012)
3. Boticiu, S. and Teichmann, F.: How does one negotiate with ransomware attackers? International Cybersecurity Law Review, 5(1), pp.55-65 (2024)
4. Brainstorm Security: What are the Pros and Cons of Ransomware Negotiation? `https://shorturl.at/GbTvH`, last accessed 2025/8/12
5. Caporusso, N., Chea, S., Abukhaled, R.: A game-theoretical model of ransomware. In Advances in Human Factors in Cybersecurity: Proceedings of the AHFE 2018 International Conference on Human Factors in Cybersecurity, pp. 69-78 (2019)
6. Cartwright, E., Hernandez Castro, J., Cartwright, A.: To pay or not: game theoretic models of ransomware. Journal of Cybersecurity **5**(1), tyz009. Oxford University Press (2019)
7. Cartwright, A., Cartwright, E.: Ransomware and reputation. Games **10**(2), 26. MDPI (2019)
8. Connolly, L.Y., Wall, D.S.: The rise of crypto-ransomware in a changing cybercrime landscape: Taxonomising countermeasures. Computers & Security **87**, 101568 (2019)
9. Cristal, M.: How to negotiate when hackers are holding you to ransom. Wired (2017) `https://shorturl.at/sLmy5`, last accessed 2025/8/12
10. Faivre, J.: Negotiations in Tech: An analysis of Asymmetric ransomware negotiations. Available at SSRN 4530094 (2022)
11. F-Secure: Evaluating the Customer Journey of Crypto-Ransomware. (2016) `https://shorturl.at/4W9LW`, last accessed 2025/8/12
12. Hernandez-Castro, J., Cartwright, A. and Cartwright, E.: An economic analysis of ransomware and its welfare consequences. Royal Society open science, 7(3), p.190023 (2020)
13. Hernandez-Castro J., Cartwright A., and Stepanova A.: Economic Analysis of Ransomware. SSRN Electronic Journal (2017)

14. Hofmann, T.: How organisations can ethically negotiate ransomware payments. Network Security, 2020(10), pp.13-17 (2020)
15. Hussain, S., Li, B., Koushanfar, F. and Cammarota, R., 2020, November. Tinygarble2: smart, efficient, and scalable Yao's Garble Circuit. In Proceedings of the 2020 Workshop on Privacy-Preserving Machine Learning in Practice, pp. 65-67 (2020)
16. Intermedia: Report Identifies Ransomware's Biggest Cost to be Business Downtime. (2016) `https://shorturl.at/1v3Vl`, last accessed 2025/8/12
17. Kumamoto, T., Yoshida, Y. and Fujima, H.: Evaluating large language models in ransomware negotiation: A comparative analysis of chatgpt and claude. (2023)
18. Laszka, A., Farhang, S. and Grossklags, J.: On the economics of ransomware. In International Conference on Decision and Game Theory for Security, pp. 397-417 (2017)
19. Li, Z., Liao, Q.: Ransomware 2.0: to sell, or not to sell a game-theoretical model of data-selling ransomware. In: Proceedings of the 15th International Conference on Availability, Reliability and Security, pp. 1–9 (2020)
20. Meurs, T., Cartwright, E., Cartwright, A.: Double-sided information asymmetry in double extortion ransomware. In International Conference on Decision and Game Theory for Security, pp. 311-328 (2023)
21. Rubinstein, A.: Perfect equilibrium in a bargaining model. Econometrica: Journal of the Econometric Society, 97-109 (1982)
22. Ryan, P., Fokker, J., Healy, S. and Amann, A.: Dynamics of targeted ransomware negotiation. IEEE Access, 10, pp.32836-32844 (2022)
23. Sophos: The State of Ransomware 2025. (2025) `https://shorturl.at/LG9qz`, last accessed 2025/8/12
24. Statista: Average duration of downtime after a ransomware attack at organizations in the United States. (2023) `https://shorturl.at/22sCM`, last accessed 2025/8/12
25. Stone, J.: FBI turns to insurers to grasp the full reach of ransomware. Cyberscoop (2020) `https://shorturl.at/GeQa8`, last accessed 2025/8/12
26. Team Cymru.: Analyzing ransomware negotiations with CONTI: An in-depth analysis (2022)
27. Vakilinia, I., Khalili, M.M. and Li, M.: A mechanism design approach to solve ransomware dilemmas. In Decision and Game Theory for Security: 12th International Conference, GameSec 2021, Proceedings 12, pp. 181-194 (2021)
28. Vakulov, A: The dos and don'ts of ransomware negotiations. LevelBlue (2023) `https://shorturl.at/AC7pq`, last accessed 2025/8/12
29. Wang, X., Malozemoff, A. J., and Katz, J.: EMP-toolkit: Efficient MultiParty computation toolkit, https://github.com/emp-toolkit, (2016)
30. Wang, X., Ranellucci, S., and Katz, J.: Authenticated garbling and efficient maliciously secure two-party computation. In: Proceedings of the 2017 ACM SIGSAC conference on computer and communications security, pp. 21-37 (2017)
31. Yin, T., Sarabi, A., Liu, M.: Deterrence, backup, or insurance: game-theoretic modeling of ransomware. Games **14**(2), 20. MDPI (2023)
32. Zhang, C. and Luo, F., Bargaining Game Theoretical Analysis Framework for Ransomware Attacks. Available at SSRN 4892700 (2025)
33. Zhang, C., Luo, F., Ranzi, G.: Multistage game theoretical approach for ransomware attack and defense. IEEE Transactions on Services Computing **16**(4), 2800–2811 (2022)