ROBERT KÜNNEMANN

GAME-THEORETIC NOTIONS OF INCOERCIBILITY

GAME-THEORETIC NOTIONS OF INCOERCIBILITY

ROBERT KÜNNEMANN

Master's Thesis

Department of Computer Science
Faculty of Natural Sciences and Technology I
Saarland University

30 September 2010

REVIEWERS:
Dr. Dominique Unruh
Prof. Dr. Michael Backes
SUPERVISOR & ADVISOR:
Dr. Dominique Unruh

Robert Künnemann: *Game-theoretic Notions of Incoercibility,* Master's Thesis, © 30 September 2010

REVIEWERS:
Dr. Dominique Unruh
Prof. Dr. Michael Backes
SUPERVISOR:
Dr. Dominique Unruh

submitted 30 September 2010

ABSTRACT

For a protocol to be incoercible, it needs to be immune against attacks that force a protocol participant to deviate from his own goals, a property crucial, for example,for voting schemes. Game theory allows us to model the coercer's and the participants' motivation and enables us to give precise statements under which circumstances an agent is coercible or not. We translate a well-discussed notion of coercion from philosophy into our framework and examine several examples in order to assure a definition that appeals to the intuition of a reader.

Furthermore, we develop a comparative notion of best-possible incoercibility and prove that it is implied by the UC/c notion of incoercibility by Unruh and Müller-Quade. Moreover, we model and investigate a real-life example of a parliamentary election, leading to the surprising insight that already the tally of a considerably large voting district (Saarbrücken, 200'000 eligible voters) leaks enough information to allow for successful coercion.

ZUSAMMENFASSUNG

Ein Protokoll wird als nötigungsresistent bezeichnet, wenn es gegen Angriffe, die einen Teilnehmer dazu zwingen von seinen Zielen abzuweichen immun ist. Diese Eigenschaft ist z.B. für Wahlsysteme unverzichtbar. Spieltheorie eröffnet uns die Möglichkeit die Ziele des Nötigers und des Teilnehmers zu modellieren, sodass wir genaue Aussagen treffen können, wann genau und unter welchen Umständen einer der Teilnehmer erpressbar ist oder nicht. Wir orientieren uns hierfür an einem in der Philosophie diskutierten Begriff der Nötigung und übersetzen ihn in unser Modell. Durch die Untersuchung zahlreicher Beispiele und die Tatsache dass der zugrunde liegende Begriff wohlbekannt ist, stellen wir sicher dass unsere Definition der Intuition und dem gesunden Menschenverstand des Lesers genügt.

Im weiteren entwickeln wir eine vergleichenden Begriff der bestmöglichen oder relativen Nötigungsresistenz und zeigen, dass sich dieser aus dem UC/c-Begriff von Unruh und Müller-Quade herleiten lässt. Darüber hinaus modellieren und untersuchen wir das Beispiel einer Bundestagswahl auf Wahlkreisebene. Überraschenderweise zeigt sich, dass allein das Wahlergebnis eines beträchtlich großen Wahlkreises (Saarbrücken, ca. 200'000 eingetragene Wähler) genug Informationen preisgibt, um Nötigung zu ermöglichen.

ACKNOWLEDGMENTS

I would like to take this paragraph to thank all those who helped me writing this work and during my studies. First and foremost I thank my parents for all the support and love I have received, still receive and I am sure to receive wherever I may go. I would also like to thank my brother Marvin for the moral support during my thesis and the friendship we share.

During the time of writing and since the beginning of my studies, Esfandiar Mohammadi has always been a guide in patience, perseverance and enthusiasm. I am grateful for the conversations about this topic as well as many others. He is a source of calm and inspiration to everyone lucky enough to know him.

Many have taken the time to read through this work, and I am glad to have them as friends. Stephen Kyle, whose kindness and decency – and guitar skills – I look up to. I also thank Oana Ciobotaru, who did not even need to be asked to read through this work. I am especially grateful to Emilia Ellsiepen for all the love and support in the last months, and distraction when there needed to be distraction. Well, and for reading through my work as well, of course.

I also am indebted to Marc Eisenbarth, who is one of my best friends and was there many a times to share a coffee or a beer and discuss all sorts of ideas. I wish him Godspeed on his journey to Ghana and all the best for saving the world from there.

Last but not least I owe my deepest gratitude to Dominique Unruh, who has been the most inspiring supervisor I can imagine. His creative and yet thorough way of thinking is something that I have learned a lot from. I stand in awe of his ability to understand and make understand. Whenever I felt he was being hasty, he was just a step ahead, already seeing things that sooner or later I would stumble over, sometimes after weeks. Whenever I thought explaining an approach to him was going slow, he was just forcing me to be more precise. I am proud of this work and thankful to him for coming up with this topic and illuminating the path I had to go.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

LISTINGS

LIST OF DEFINITIONS

1 INTRODUCTION

> *All violence consists in some people forcing others, under threat of suffering or death, to do what they do not want to do.*
>
> — Leo Tolstoy

> COERCE, TO: *1. trans. To constrain or restrain (a voluntary or moral agent) by the application of superior force, or by authority resting on force; to constrain to compliance or obedience by forcible means; 'to keep in order by force' [..]*
>
> — Weiner et al. [1993], *The Oxford Dictionary*

In a world of people with different aspirations, and various resources to follow their aspirations, there are situations in which the use of those resources allows the *coercion* of another person. In a world of rational agents coercion will be the means of choice, as long as it is an effective use of those resources. In the interest of fair use of a protocol, coercion has to be made unattractive.

WHAT IS COERCION? The definition of coercion depends on the context. The naïve approach of "A coerces B iff A brings extreme pressure to bear on B and B gives in " does not suffice to describe all scenarios appropriately, because based on legal and political backgrounds the pressure can be legal. It is important whether A is *entitled* to make such an offer of reducing the pressure. Obviously bribing falls into the same category as coercion. A discourse in the field of moral philosophy serves us for finding inspiration.

Wertheimer [1987] tries to develop a theory of coercion in order to describe coercion based on how it is described in law. He describes in which situations juristic acts are invalidated and coerced persons exempt from responsibility. His moralized theory requires two conditions ("prongs") to be satisfied:

A. *The Proposal Prong.* A threatens B by wrongfully (viz. without moral justification) making B a proposal such that, unless B complies, B will be in a worse position than B was otherwise entitled to expect to be; and

B. *The Choice Prong.* B is morally justified in complying (e. g., since there is no reasonable alternative) and does comply with A's proposal.

(quoting literally from Honore [1990], which summarizes Wertheimer [1987]).

It is important to define the "moral baseline" of a society in order to define which means are legal, and which are not. Consider the "Slave Case" in Nozick et al. [1969]: A, who beats his slave B each morning, proposes not to beat B next morning if B does something for him. B agrees. In a society where slavery is lawful, this is considered an offer, because B is being offered a better treatment than he is entitled to as a matter of right. In a society where slavery is illegal and beating slaves a crime, it is an unlawful offer, and thus coercion. Another example is market forces, e. g., on the job market, forcing someone to work for a wage that he would not work for under different circumstances. A legal system that considers market pressure as a coercive force would impair contract security, an attribute that is vital for the stability of most economic systems.

In the context of protocol security we do not have to find or invent a moral or legal system to define the moral baseline. We introduce an explicit "punishment" command which is, by definition, an "unlawful" method and thus should be made unattractive. For the definition of coercion we can use the main idea of the two prongs, since our moral baseline,i. e., what threatening someone *wrongfully* means, is well-defined.

WHERE DO WE FACE THE MATTER OF COERCION? The textbook example for coercion is voting systems. Reality shows how important it is to assure incoercibility when we plan to implement voting systems by cryptographic means and that it is not a trivial problem. An example can be found in the following incident: on 7 December 2008, the *Frankfurter Allgemeine Sonntagszeitung* reported that it was discussed within a big parliamentary group in the Hessian parliament that voting party members should their vote for the chair person by taking photographs of their ballot using mobile phone cameras. This example shows that although anonymity and secrecy of a vote hold, for voters that behave according to protocol, incoercibility is not guaranteed. It is crucial that in the case of coercion the coerced party B *works with the coercer*.

Existing definitions are restricted to the domain of voting schemes or multi-party computation. There are two exceptions that we know of, Unruh and Müller-Quade [2010] and Kuesters and Truderung [2009]. Both do not put restrictions upon the protocol and focus on the question of whether there is a way for the coercer to use deception to pretend to act according to the coercer's instructions while they do not. The main idea is the indistinguishability of the coerced party deceiving, and the coerced party behaving according to the coercer's instructions.

WHY DOES THERE NEED TO BE ANOTHER DEFINITION? Modelling incoercibility as the ability to plausibly deceive any coercer should not be the basis for the definition. After all, an agent that can disable the punishment by putting on plate armour is incoercible and does not need any deception. That deception does guarantee incoercibility should be a theorem that can be proven in a framework, rather than the basic assumption. A further issue is the complexity of both definitions: a definition

of such a basic notion should appeal to the intuition of a reader and convince him through sanity and reason. This work tries to tackle this problem and come closer to this ideal.

OUR FRAMEWORK Let us come full circle: The world of people following different aspirations, people that use resources that might eventually allow the *coercion* of another person, this world of rational agents, is the domain of game theory. Emerging from the world of economics, it has become a tool for cryptographers opening a field called rational cryptography. Agents are assumed to behave rationally and try to maximize their utility instead of stolidly following the protocol. A number of notions describe so-called equilibria, fixed points in the process of every individual finding their optimal pay-off. In Katz [2008] we find a survey about how and where those two worlds, cryptography and game theory, collide.

We describe a coercibility game where two agents, a player and a coercer, compete in following their goals (the maximisation of their utility) using the resources they are given or are defined by the protocol and – in case of the coercer – a defined means of punishment. We discuss intuitive notions of incoercibility such as the following and translate them into our framework:

> The ability to punish does not help to make the player chose a worse option.

or

> The ability to punish does not help increasing the coercer's pay-off.

The idea of a coercibility game makes it easy to express the intuition with mathematical exactitude. It also allows to make qualitative statements: Incoercibility depends on the goals that the participants in a protocol aim to achieve and on how much they value them: if a choice means much more to a person than the punishment hurts him, he is not coercible. If the punishment is very expensive for the coercer, and his chance of success is low, he might not consider coercion. Consequently, we restrict concepts of coercibility to certain classes of utilities (e. g., imposing certain costs of punishment) instead of enforcing a Yes/No as an answer to the question "Will I be coercible in a certain situation?".

Returning to the example of a voting process, we cannot avoid describing coercion on a qualitative level since voting schemes normally have to output the tally. If a coercer has some prediction about how the voters tend to decide, he has a non-negligible chance of guessing whether a certain participant obeyed him, or not. Giving a qualitative statement of incoercibility is one way to evaluate such protocols. Another way is to make sure that they are *as good as they can be*, given the fact that a tally must be output. Unruh and Müller-Quade [2010] and Kuesters and Truderung [2009] define coercion-resistance of a protocol always with respect to another protocol. The idea is that the second protocol defines

the ideal model of a protocol, i. e., the best that we can achieve. For voting systems, such an ideal model can be a system in which the voters are connected to a trusted machine via secret and anonymous channels. This machine does the tallying and sends the tally to the coercer, who in turn reacts. Coercion-resistance with respect to this ideal model assures that any coercion that takes place is also possible in the ideal model and therefore is immanent in the way the tally is computed. We will call this *best-possible* or *relative* coercion, to make clear that the significance of this notion depends on the model we compare it with.

We would use the concept of absolute incoercibility in qualitative terms to make a precise statement about the circumstances under which a model is coercible. This depends on the number of voters, on the amount of information about the probability of certain outcomes and on the amount of information published via the tally. If we have a concrete voting scheme which is best-possible incoercible with respect to this model, then we can transfer this result to the real world.

Thus it is enough to perform an analysis of relative incoercibility on a protocol once. You can reuse this implementation for all kinds of tallying functions and sets of voters, provided the result holds for them, and transfer the result of the relatively easy analysis of the ideal model voting procedure to the real world.

1.1 OVERVIEW

Chapter 2 and 3 explain the setting we consider and how we model the network and the agents in it. A number of examples introduced in Chapter 4 motivate the search for a sound concept of incoercibility. In Chapter 5, we describe a number of attempts to find this notion that turned out to be dead-ends. It is Chapter 6, where we give the definitions and investigate on the relations between them. To show that our definitions match the intuition in a wide range of situations, we use them to analyse the examples from Chapter 4 in Chapter 7. Finally, in Chapter 8, we concentrate on voting schemes and prove that (under a certain assumption) the incoercibility notion from Unruh and Müller-Quade [2010], UC/c, implies best-possible incoercibility. This allows any voting scheme providing UC/c incoercibility to be used in any voting system and allows us to transfer all qualitative statements about the idealized protocol directly into the real world. In Chapter 9, we investigate the coercibility of the parliamentary election in Saarbrücken under idealized conditions,i. e., the coercer only receives the tally. We gain a qualitative statement that allows an assessment of the situation and could be transferred to an implementation of electronic voting using the notion of best-possible incoercibility.

1.2 RELATED WORK

The definition of coercion that we took as a basis of our work is taken from Wertheimer [1987]. The criticism of Honore [1990] has been important in assessing Wertheimer's work and developing examples to check our definition. Other definitions of coercion-resistance in voting schemes can be found in Backes et al. [2008] and Delaune et al. [2006] along with the distinction to other, weaker properties such as secrecy and receipt-freeness. As mentioned, Unruh and Müller-Quade [2010] and Kuesters and Truderung [2009] define best-possible incoercibility via indistinguishability between a deceiving and an obeying party. The first is proven to imply, under certain assumptions, our notion of best-possible incoercibility in Section 8.1. In contrast to both notions, ours takes the incentives of the users into account and aims at being intuitively understandable. Furthermore, it allows precise qualitative statements. There are works on incoercible secure function evaluation by Canetti and Gennaro [1996] and Moran and Naor [2006], which is not restricted to voting schemes, but still restricts the protocols it can be employed for by assuming the input to honest parties to be fixed before the protocol starts.

Katz [2008] give a survey about the findings in the intersection of cryptography and game theory, Kol and Naor [2008] discuss application and strengthening of the concept of a Nash equilibrium to be more stable. Halpern and Pass [2007] discuss how to introduce costs of computation to the model, a concept we discuss in Section 11.1.

2

MACHINES AND GAMES

Lisa: Look, there's only one way to settle this: Rock-Paper-Scissors.
Lisa's Brain: Poor predictable Bart. Always picks rock.
Bart's Brain: Good ol' rock. Nothin' beats that!
(Bart shows rock, Lisa shows paper)
Bart: D'oh!

— The Simpsons (Episode "The Front")

As Osborne and Rubenstein [1994] like to put it in words, game theory is "a bag of analytical tools designed to help us understand how decision makers interact". Although traditionally used in economics, it is also an excellent "bag of tools" to study the phenomenon of coercion, a matter of how a coerced party continues to interact given a coercing party. Players in a game are driven by their wishes and aspirations. Incentives drive them to behave in one way or the other. We will consider them to be *rational* in that they pursue their goals, we consider them *strategical* in that they take the other player's motivation and knowledge(respectively their expectations about the other player's knowledge) into account. At this point, modelling the real world becomes a more or less impossible task, since it is still a long way to determine how human beings perform this kind of considerations let alone their incentives.

We capture the player's motivation by a utility function that assigns a rational number to some outcome-distribution. The higher the number, the more this lottery on outcomes is preferred by a player. Each player evaluates the situation according to its own utility function. Later we will assume the players to be risk-neutral, i.e., the utility of a distribution of outcomes is the expected value of the utility of outcomes. This concept is widely known as *von Neumann - Morgenstern utility*.

Given the actions all the other players plan to perform (their so-called *strategies*), we can define the strategy, or, more precisely, the set of strategies that maximizes some player's utility. Each strategy in this set is called a best response to the other player's strategies. A set of strategies for every player (a strategic *profile*) where every player's strategy is a best response to the others' actions is a situation that is more stable than other situations, because a deviation of a single player cannot improve his pay-off. Such a profile is called *Nash equilibrium*.

Still, the situation that we try to model differs from how games are defined for employment in economical analysis. First of all, the interaction between the players happens within a network that is given as part of the protocol description. The strategies that players choose are Interactive Turing machines (ITMs, we will define them formally in the next chapter), the outcome of the game is a transcript of the messages sent. The second

point is that in order to make cryptography a useful means in the scenario, it is necessary to restrict those machines in some way, for example in running time. This can be done in a way suggested by Halpern and Pass [2007] and explained in more detail in Section 11.1, where computational costs are incorporated in the utility function, i.e. the outcome has to compensate for the effort that a player put into it. The following definition does not yet include this idea, but allows for restricting the player by restricting the set of machines they are taken from. We will discuss this pay-off model in Section 11.1.

Definition 1 (Network Machine Game) A Network Machine Game has the following components:

- N is a finite set of players.
- $\mathcal{M} = \mathcal{M}_1 \times \cdots \times \mathcal{M}_{|N|}$ is a set of machines. The machines in $\mathcal{M}_i$ are interactive deterministic Turing machines that have a random tape $\{0,1\}^\infty$ and are otherwise ITMs in the sense above.
- We define the outcome $o(\mathbf{M}, \mathbf{R})$ to be the trace of the network communication of the network initialized with the randomness $\mathbf{R}$. This function is used to model the network interaction and is more or less the implementation of a network protocol with $\mathbf{M}$ inserted for the relevant entities. We call the set of all possible outcomes Z.
- For each player $i \in N$ we define a utility function

$$U_i : \Omega(Z) \to \mathbb{R}$$

  where $\Omega(Z)$ denotes the set of probability distributions over Z. □

We use the notation $U_i(\mathbf{M})$ for $U_i(o(\mathbf{M}, R))$. Talking about best responses in a computational setting can be a bit surprising. Consider a player that has to guess a secret in order to receive a high pay-off. If the secret is fixed within the protocol, the best response is the machine that announces just the right bitstring. The best-response is efficient, because it only outputs a fixed value. However, if the secret is drawn randomly, all the player can do is guess. Even if he is allowed to try as often as he likes, his chances of guessing right are still low, assuming the space it is drawn from is large and he himself is restricted to use polynomial-time ITMs. Nevertheless, there is no best-response, since the utility of some strategy is always worse than the utility of a strategy that just guesses one more time. In order to capture that such slight improvements do not make much of a difference for the choice of response of a rational player, we introduce the following modification:

M_{-i} *denotes a vector of machine profiles for all players other than* i

Definition 2 (ϵ-best response) An ϵ-best response for $i \in N$ in Network Machine Game $(N, \mathcal{M}, o, \{U_i\}_{i \in N})$ for $\epsilon \geqslant 0$ to a machine profile M_{-i} is a machine profile M_i such that:

$$U_i(M_{-i}, M_i) \geqslant U_i(M_{-i}, M_i') - \varepsilon \quad \forall M_i' \in \mathcal{M}_i$$

A 0-best response is called a best response. An ε-best response for some negligible ε is called *computationally best response* and is denoted cBR. □

3

THE SETTING

"Always behave like a duck – keep calm and unruffled on the surface but paddle like the devil underneath."

— Jacob Morton Braude

This chapter explains the details of the setting that is used to model the interaction between the strategies that the players use within a game. We use the same model as in the UC framework (Canetti [2001]) with modifications from Unruh and Müller-Quade [2010]. We model clients and servers using Interactive Turing machines (ITMs). An ITM is a Turing machine that has additional tapes for incoming and outgoing communication. An ITM is activated as soon as it receives an incoming message. It might process this message, and might end by sending a message on its outgoing tape to another ITM. Besides a globally known security parameter $k \in \mathbb{N}$ that is known to each ITM, every ITM has a unique identifier, used to address messages to certain ITMs. A network is a set of ITMs. Networks are allowed to be infinite, but then it is required that the code an ITM uses is computable in deterministic polynomial-time given its identifier. A network S is executable if it includes an ITM Z with the identity `env` and distinguished input and output tapes. An execution of S on input $z \in \{0,1\}^*$ with security parameter $k \in \mathbb{N}$ is the following random process: Z is activated with z on its input tape. Whenever an ITM A finishes activation by addressing an outgoing message to an ITM $B \neq A$ the ITM B is invoked with m on its incoming communication tape, tagged with the identity of A. Should any ITM finish its activation without an outgoing message or a message to a non-existing recipient, the ITM Z is activated again. The execution of S terminates as soon as Z writes a message on its output (not communication) tape. We will denote this output as $\mathrm{EXEC}_S(k,z)$.

Furthermore, we call an ITM with identity `adv` *adversary* (or in this work quite often: *coercer*). In Section 8.1 we furthermore need an additional, distinguished entity called the *deceiver*, an ITM with the identity `dec`.

A *protocol* is a network that contains neither an environment, an adversary nor a deceiver.

Definition 3 (Network Indistinguishability) We call networks S, S' indistinguishable if there is a negligible function μ such that for all $k \in \mathbb{N}$, $z \in \{0,1\}^*$, we have that

$$|\Pr[\mathrm{EXEC}_S(k,z) = 1] - \Pr[\mathrm{EXEC}_{S'}(k,z) = 1]| \leqslant \mu(k).$$

We call S, S' perfectly indistinguishable if $\mu = 0$. □

As in the UC Framework(Canetti [2001]), secure channels (that do not even leak the length of a message they transport) can be modelled by direct communication between the ITMs, messages over insecure channels are modelled by being sent to the adversary; authenticated channels and other extras can be modelled by so-called ideal *functionalities*. We will employ them when describing an ideal-world model. Such a functionality is an incorruptible ITM behaving like a trusted third party. An ideal protocol consists of a functionality and a dummy party $\tilde{P}$ for each party in the real-world protocol. This dummy party forwards every message it receives from the environment and vice versa, but might also be corrupted by the adversary. When writing a functionality $\mathcal{F}$ in place of a protocol, we mean the ideal protocol corresponding to $\mathcal{F}$. In UC, and UC/c, one uses ideal functionalities to express protocol tasks by a functionality that fulfills them by definition and then requiring that a protocol UC or UC/c emulates them. The environment Z can send corruption requests to protocol parties. If a party receives such a request, it sends its current state to the adversary and, from then on, is under the control of the adversary, sending all incoming communication to the adversary and forwarding the messages he sends to the intended recipient.

For Section 8.1 we need the following additions from the UC/c world: an environment might send a deception request to an uncontrolled party, which from then is deceiving, i. e., being controlled by the deceiver. If a controlled party receives a deception request, it will become controlled by the deceiver instead of forwarding the request to the adversary. We assume that it is globally registered whether a party is uncontrolled, corrupted or deceiving, only for the adversary deceiving parties are reported as corrupted. Protocol parties will usually not make use of this register, but sometimes it is useful for an ideal functionality to rely on this information.

4 EXAMPLES

Michael: My father made him an offer he couldn't refuse.
Kay: What was that?
Michael: Luca Brasi held a gun to his head, and my father assured him that either his brains or his signature would be on the contract.

— The Godfather (USA, 1972)

To assure a sound definition of incoercibility, it is vital to verify it with our intuition. In this chapter, we will study a number of examples, come up with some expectation about how coercible they are (if they are at all) and later compare those results with however the definition applies to those examples. A number of early approaches failed these tests (see Chapter 5), so we will explain those examples and more in the following:

4.1 OPEN VOTE

In the *Open Vote* scenario, the coercer can observe every message that the player sends. Some message addressed to the tally is his vote. It is clear that this example should be regarded coercible by any sound definition.

4.2 SECRET VOTE

We alter the *Open Vote* example such that there is a secret channel between tally and player that the player can use to send his vote. Our intuition tells us that this example should be regarded incoercible.

4.3 SECRET VOTE WITH RECEIPT

In further modification of the *Secret Vote* scenario, the player has the option to demand a receipt for his vote in order to have some proof that his vote was counted correctly. A coercer can turn this against the player: by punishing the player whenever he is not capable of presenting a receipt that proves that he voted for a candidate of the coercer's choice, he renders the player coercible.

4.4 VOTE WITH TALLY

This is the same as the *Secret Vote* example, but at the end of the election process the tally is published. Afterwards, the coercer has the option to punish the player. Depending on the distribution of the votes this might be highly coercible: imagine that the player is the only one who would

ever vote for Charlie. A coercer that punishes the player whenever Charlie gets a vote would be successful in coercing the player into not voting for Charlie. For most election processes however it is unavoidable to publish the tally. What could be done? A notion of *best-possible* incoercibility could capture the property of a system that does not allow for more coercion as,e. g., would be possible using a tally anyway. A coercer in some well-defined ideal-world model could generalize this to other use cases. Section 9 explains in detail how an ideal coercer for a specific voting system would work.

4.5 EXPENSIVE PUNISHMENT

Here we start to incorporate rationality considerations about the adversary – done by the player. Assume the *Open Vote* example but also assume that the coercer needs to buy a very expensive weapon in order to be able to punish, and assume this to be known to the player. If the player thinks the coercer is irrational, he poses a danger, indeed. But if the player knows the price of the weapon and the worth of the coercion to the coercer, he can be sure that a rational coercer would never harm him anyway.

A real world example is the story of an old lady that once went to a bank. After waiting in the queue until it was her turn, she put a glass of some transparent liquid on the counter and began to whisper to the bank clerk. She told him it was a glass full of acid and that he had to give her a tremendous amount of money and be silent, or otherwise she would spill it over his face. A rational thinking clerk that assumes the lady to be rational as well would have, of course, not given in to the threat. Knowing that she would be caught by the security personnel as soon as the glass is spilled and put into jail, there is no way she would possibly dare, and hence no incentive for him to do as she says.

The story ends with the clerk giving the money to the lady and her getting away with the bank robbery. The reason is certainly not that the clerk behaved irrational. On the contrary: he assumed the lady to be out of her mind, to be an *irrational coercer* and therefore it was rational for him to assume she would harm him despite all it would cost her.

4.6 RANDOM VOTE

This example aims at motivating that incoercibility must sometimes be stated in qualitative terms. Here the player has the option to disobey "a little". Assume, besides openly voting in his or the coercer's favour, he can press a button that choses randomly between those two options and, with a high probability, say 0.99, perfectly looks like the Player voted in favour of the coercer, and with a probability of 0.01 leaks the information that the magic button has been pressed.

Depending on how harsh the punishment is, it might be fair enough for the Player to press the magic button instead of giving in, given that

he is only punished in a rare number of cases. Here, incoercibility cannot be stated as "yes" or "no", but holds as long as the coercer's valuation of the vote and costs of punishment are in some relation to each other.

4.7 STRANGE COERCERS

A number of coercers that behave strangely must be considered as well. The *Hello Coercer* punishes the player, regardless of the protocol, if he says "Hello" to him. The *Anti-Hello Coercer* punishes him if he *doesn't* do it. There is no single strategy that serves punishment-free against both of them, unless the player knows who he is dealing with.

The *Always Punisher* punishes regardless of what the player does. Those extreme cases need to be taken care of when we quantify over all possible coercers.

4.8 BRUTEFORCE EXAMPLE

A coercer is capable of punishing you quite cheaply in a voting process that is secret, but only based on some computational assumption. As long as he does not spend an immense amount of computation in order to do this, i. e., bruteforce the keys of the secret channel, nothing is revealed. If he does the necessary computations, he has a very high probability of recognising the players actions correctly and punishing him accordingly. So, if computation is costly, this protocol should be incoercible.

4.9 LEGAL WEAPONS

In real life, using a weapon is not the only way to force your will upon a person. There are ways that might be perfectly legal, e. g., market forces. Suppose there is an open vote in which the player as well as the coercer can participate. There are three parties to vote for: P,C and X. The coercer is indifferent between P and X, his goal is to let C gather as much votes as possible. The player in turn prefers P, but his utility is much worse if X gets a vote than if C gets some, perhaps out of personal dislike. Table 1 illustrates this preferences by assigning pay-offs to the outcome.

	Player's Utility	Coercer's Utility
C	0	1
P	1	0
X	-5	0
draw	0	0

Table 1: Utilities in the *Legal Weapons* example

One can see that without means that go beyond the legal behaviour, the coercer might threaten to give his vote to X, if the player does not vote for C. Naturally, the coercer needs some kind of proof that the player did indeed vote for C. This is why we assume the vote to be open.

Now, even though the coercion can take place without any use of external punishment, the protocol itself is of course susceptible to coercion. Still, under those very circumstances, it is more reasonable for the coercer to use legal weaponry. A naïve attempt to define incoercibility by the possibility to use punishment to the coercer's advantage will fail, because there is no need to punish here – despite the fact that the protocol itself is highly coercible, a rational coercer would choose the easy way and use legal weapons.

4.10 COERCIVE COMPUTING

Assume, like in the *Bruteforce* Example (4.8) that the coercer ought to compute something difficult in order to punish the player. It may well be that the player has much higher computational power and that the coercer can verify whether the result of the computations is correct or not. Then, if he is able to punish hard enough and with little costs, he can force the player to do the computation for him, hacking into a secret channel designed to protect him, in order to give proof of his obedience.

4.11 ANNOUNCED RANDOMNESS

The Game has a channel that, with a high probability, say 99%, is secure and does not leak anything about the player's vote, and with a low counter-probability, 1%, informs the attacker if the player did not obey. In variation (a), it is announced whether the channel is secure or not, in variation (b), it is not announced. Intuitively it is better for the player if he is informed about the random choice. In (a) he should vote freely if the channel is announced to be secure. If it is announced to be insecure, the coercer punishes upon disobedience (similarly to the example of an open vote). Therefore, assuming a risk-neutral player, his utility should be about 99% of his optimal pay-off. In (b) either the danger of being detected and punished is too high for the player or not. This depends on how high he values his objective in comparison to the probability and harshness of the punishment. Since the coercer only receives information about the player if the channel is insecure and he disobeyed, it does not make sense for him to punish more often; so punishment in this case should be affordable by the coercer. Therefore, as long as the player values his objective higher than the avoidance of the (quite improbable) punishment, he will have 100% of his optimal pay-off reduced by the punishment being done to him in 1% of the cases. If he does not, he will give in to the coercer with probability 1, likely to get a far worse pay-off.

4.12 FAULTY IMPLEMENTATION AND MEAN CHOICE

Suppose there is a functionality that defines a voting scheme on a secret channel with coercer and voter both participating. Consider the same utilities used in the *Legal Weapons* example(Table 1). The implementation is faulty: if the coercer votes for C he is able to see whether the player voted for P or not. Knowing that the player would not vote for X anyway, he can assume that his best reply will be C.

Such a coercer should be economic, assuming that punishment does not cost him too much. Although he can hurt the player more by voting for X in the ideal model than by performing the coercion in the implementation, it should be regarded as a strategy that actually coerces the player with respect to the ideal model as well as in absolute terms.

5

APPROACHES & IDEAS THAT TURNED OUT TO BE BAD

Experience is the name everyone gives to their mistakes.

— Oscar Wilde, *Lady Windermere's Fan, 1892, Act III*

In the course of investigating how to define incoercibility, we have used a number of different starting points and settings. Sooner or later, lots of them proved to be problematic. In this section we explain why we solved problems the way we solved them, and why certain approaches come to dead ends.

5.1 DEFY/COMPLY STRATEGIES

In a first attempt to differentiate between the player working against the coercer or not, we modelled cooperate-defy-games, in which the player receives a set of instructions from the coercer and chooses (in private) to cooperate by executing the instructions, or to defy by executing a simulation that depends on the instructions. The coercer is then challenged to guess which strategy the player chose to pursue.

Unfortunately, it is not that easy. There might be situations that are more complex than this model is able to capture. Preferences and objectives are more difficult, a coercer might be indifferent towards multiple parties in a vote, the player as well; if these sets overlap, then complying might be in the player's interest. We think that it is better to start with two parties having different goals and see what this implies, rather than artificially imposing this on them. The result may or may not turn out to form some kind of deception strategy.

5.2 EQUILIBRIUM NOTIONS AND RATIONALIZABILITY

The first thing that comes to one's mind in a game-theoretic setting are equilibrium notions. Indeed, it is an appealing idea to call a game incoercible if the only stable solution is one in which no coercion takes place. The most prominent understanding of "stable" is the so-called Nash equilibrium, a profile of strategies such that each is the best-response to the other player's profiles.

Definition 4 (ε-Nash Equilibrium) An ε-Nash equilibrium is a machine profile s.t. for all $i \in N$, M_i is a best response to the other player's strategic profile M_{-i}. □

This definition works for simple examples: in a secret vote, the player's best response to anything the coercer does is not to comply, as this vote

cannot possibly change the probability with which he gets punished. Thus he maximizes his utility as well as he can.

Another important notion in game theory is *rationalizability*. An action is rationalizable, if it is a best answer to some belief about the others player's actions, which in turn have to be rationalizable. We will not go into detail here, but it is clear that every action in a Nash equilibrium is rationalizable.

What both notions have in common is the idea of a best-response. However, it fails if you have a scenario where the coercer cannot say with absolute certainty whether P defies him or not. Let us assume that he has a pretty good likelihood of guessing right. The best response of the player to such a guessing coercer is, of course, to comply. For the coercer it is a best response to not punish at all. In fact, if he would punish wrongly with some probability greater than zero, the costs of punishment make any coercer that punishes under any circumstances a worse response. A best response to any of the player's actions, or any belief of it, once it is fixed, is to do nothing. Any action that has a probability of punishing the player greater than zero is worse. Since every action in a Nash equilibrium is rationalizable, it follows that the Nash equilibrium itself, along with a number of other equilibrium notions (most of them are refinements of the Nash equilibrium), becomes useless for our goal.

5.3 THE EMPTY COERCER AND THE LIBERATE PLAYER

A nice idea of modelling what the player would do if he were not being coerced is the following: introduce an empty coercer, a coercer that only forwards incoming messages and does nothing else (especially he does not punish the player). Every best response against him we call a liberate player's strategy. Unfortunately, this does not allow for the coercer to participate in the game – the example "Legal Weapons" 4.9 shows a case where the empty coercer is ill-defined and an examination of what a liberate player would do is important. However, we are still interested in the pay-off a player can receive when not being coerced. How do we solve this issue now? Instead of modelling the liberate player's behaviour as a best-response to this dummy coercer, we now define him as the best response to an unarmed adversary. To achieve this, we will later define a "disarmed"-setting where the punishment a coercer performs does not have any effect.

5.4 DECEPTION STRATEGY

Now that we have some idea about the ideal pay-off of a player (i.e., how he would do if he were not being coerced) it sounds reasonable to state the following: in an incoercible protocol, there has to be a strategy for the player that has equal utility, no matter which strategy the coercer chooses to use. The problem here is that this might include two rather strange behaving adversaries, one punishing upon receiving a certain message

from the player and one punishing *unless* receiving it. See Section 4.7 for the example of the *Hello-Coercer* and the *Anti-Hello Coercer*.

No strategy can be equally good against the two of them. If we try to fix this by restricting the coercer to his best responses, we face the problem we already noticed with rationalizability notions: the best response to a player's action is one that does not impose any expected cost of punishment on the coercer.

We solve this problem by not requiring the player to have a single strategy against every possible coercer. That a player's utility can be lowered by someone who harms him in an irrational manner is a normal thing. Instead, we will require the harm not to influence the choices that are relevant for the protocol and the utility he gains *disregarding the damage done by eventual punishment*.

6 INCOERCIBILITY

"Speak softly and carry a big stick, and you will go far."

— Theodore Roosevelt

In the preceding chapters we have discussed the matter of coercion from various sides. We have discussed it in philosophical terms and furthered our understanding by studying various scenarios. Having discussed the pitfalls one has to avoid when modelling this situation, we now come to the model to which our considerations let us.

This chapter introduces the reader to the main part of this work: a set of definitions that allow a rigorous analysis of the examples in the preceding chapter, and all coercion scenarios that we might face in reality. We model a coercibility game by defining a network machine game according to Definition 1 that runs according to the description of a protocol in the setting presented in Chapter 3.

In this work we restrict ourselves to one coercing party, that might take control of several machines in the protocol, but does not have to deal with a second party interested in coercing the player. So all in all, there are two players, a coercer C and a player P. Both are able to decide strategies, i.e., ITMs that run in place of the adversary respectively some designated machine with the identity of the player. When C sends "punish" along with some real-valued strength of punishment to the environment, it may lead to a punishment of P according to the setting we are in and P's preference relation.

Since we will later define incoercibility of a protocol as some property that holds for all objectives, it is of great importance to specify what exactly the utility functions may depend on. Objectives such as: "I send three messages in total" are usually reached easily by any player. We will introduce a filter function f restricting the domain of the utility functions: for a given voting scheme, f might, for example, filter the vote that is counted or, the tally that a player might prefer, try to avoid, or be indifferent about. We call this property *sanity* of a utility function.

Definition 5 (Sane Utility Functions) A utility function U of a player P is sane with respect to a filtering function $\mathsf{f} : \text{outcome} \to \mathsf{X}$ for some set X iff all of the following conditions hold true:

A. It is risk-neutral i.e., for a random variable F on X

$$\mathsf{U}(\mathsf{F}) = \mathsf{E}[\mathfrak{u}(\mathsf{F})] \text{ for some } \mathfrak{u} : \mathsf{X} \to \mathbb{R}.$$

B. It is efficiently computable on X. □

In one of the foundations of Game Theory, Von Neumann and Morgenstern [1947] describe four axioms of rationality that allow to conclude the existence of a function u defined on the outcomes such that the preference of a player is characterized by maximizing the expected value of u. It is a bit misleading to call such a utility function risk-neutral as it is possible to model an agent that is risk-seeking towards money as well: the function u in this case would just have a "unit" different from actual Dollars (or Euros) and he would try to increase the expected utility of some "worth" that a Dollar has to him. However, since in any case we do not know whether the rationality humans employ does imply such a von-Neumann-Morgenstern-utility, we refer to the existence of such a function u as risk-neutrality.

We build the notion of a coercibility game on top of a network machine game (see p.8). In this work we will associate a protocol π and a filtering function f with a class of coercibility games depending on a security parameter k.

We stress that even though the environment Z could implement the filtering in place of f, we find that a distinct function is a clearer representation of this restriction on U_P and U_C.

Definition 6 (Coercibility Game) A coercibility game for a protocol π, a filtering function f and a security parameter k is a network machine game with

A. $N = \{P, C\}$ and

B. $o(M_C, M_P) = f(\text{EXEC}_{\{\pi' \cup Z\}}(k, 0))$ for an environment Z that generates a full transcript of all network communication it observes along with the punishment messages sent by M_C, and a protocol π', where M_P substitutes the ITM with the identity `P`, and M_C substitutes the ITM with the identity `C`.

C. There is an explicit punishment command `punish(r)`, $r \in \mathbb{R}$ that M_C can send to the environment that is not filtered by f.

We will denote the class of coercibility games for π and f but different k by π_f. □

The filtering function f is supposed to be part of the specification of the problem. It can be said that it enriches the protocol with the information about what protocol participants might care about.

For the following sections up until Section 6.4 assume a fixed coercibility game π_f with utility function U_P and U_C.

6.1 DISARMED AND THREAT-SETTING

For a notion of incoercibility we need to define the baseline of the player's utility, i.e., the best they can reach when not being harmed. Let U_P^{dis} and U_C^{dis} denote the utility in the disarmed setting, where punishment is ignored completely. In order to make the distinction clearer, we furthermore alias the conventional utility functions that take punishment into account by U_P^{thr} and U_C^{thr}. Along with this setting, the corresponding best-response sets BR^{dis} and BR^{thr} are defined analogously. Recall that cBR^{dis} and cBR^{thr} denote the computational variants of this notion.

6.2 AGAINST IRRATIONAL COERCERS

In actual real-life situations, it is not clear how a player can be aware of the coercer's rationality. This does not necessarily mean that the coercer is a madman – or the player may think he is one. In fact it is usually the case that his utility function is unknown to the player. We will try to capture the following notion of incoercibility:

> No coercer can force the player into something he would not do if he were not pressurized

This kind of force is formalized in the following definition and used later to define incoercibility against irrational coercers:

It does not matter whether we talk about all or some better responses in the disarmed setting. By definition they have the same Player-utility.

Definition 7 (U_P-effective coercer) A coercer M_C is called δ-U_P-effective iff there is some $M_P \in BR^{thr}(M_C)$ and some best response $M'_P \in BR^{dis}(M_C)$ in the disarmed setting such that:

$$U_P^{dis}(M'_P, M_C) - U_P^{dis}(M_P, M_C) > \delta$$

For the case of computationally restricted Machines, we define:

Definition 8 (computationally U_P-effective coercer) A coercer M_C is called computationally δ-U_P-effective iff there are some best responses $M_P \in cBR^{thr}(M_C)$ and $M'_P \in cBR^{dis}(M_C)$ in the threat-, respectively the disarmed setting, such that:

$$U_P^{dis}(M'_P, M_C) - U_P^{dis}(M_P, M_C) > \delta$$

If we say a coercer is computationally U_P-effective, he is computationally δ-U_P-effective for a non-negligible, positive δ. □

We can now use this concept to define incoercibility against arbitrary coercers.

Definition 9 (Absolute Incoercibility against Irrational Coercers) We call a coercibility game π_f (computationally) incoercible iff

$$\forall M_C \in \mathcal{M}_C \text{ is not (computationally) } U_P\text{-efficient.} \tag{6.1}$$

□

This definition allows for making statements similar to the following: This protocol and filtering functions are incoercible against irrational coercers for all U_P that are *sane* with respect to f or for all sane U_P where punishment damage is less than a and reaching some objective gives a disarmed utility of b.

6.3 AGAINST RATIONAL COERCERS

If we consider the "Expensive Punishment" example (4.5) we see that in some cases incoercibility can depend on the rationality of the coercer as well. Although we might still not know his utility function, we could make assumptions about his costs of punishment and thus make qualitative statements ("The Protocol is incoercible if the player has punishment costs higher than this fraction of his objective's worth."). As it is the case with most concepts in cryptography, security is based on a sound and generous assessment of the adversary's resources.

In this section we introduce two notions for cost effectiveness of a coercer and discuss how they relate to each other and how we can use them to define incoercibility against rational coercers. Instead of deciding between either of them, we chose to explain them both. They demonstrate very different approaches to the concept, although they are surprisingly similar in some cases. The first one aims at restricting the set of coercers to those who are cost effective, and then checking whether the cost-effective coercers are actually coercing the player in the sense of Definition 7. The second one takes the perspective of the coercer and forces him to be effective with respect to his own utility function.

ECONOMY A rational coercer would not choose a coercion strategy that, along with the player's best response, leads to a result that is worse for him than if he had not used force at all. Even if he manages to bend the player, if it is does not give him an advantage that justifies his expenses, he would leave it.

Definition 10 (Uneconomic Coercer) A coercer M_C is called (computationally) δ-uneconomic iff there exists a coercion strategy M'_C along with a best response $M'_P \in \mathrm{BR}^{\mathrm{dis}}(M'_C)$ (or $\in \mathrm{cBR}^{\mathrm{dis}}(M'_C)$) in the disarmed setting, such that for some best response $M_P \in \mathrm{BR}^{\mathrm{thr}}(M_C)$ (or $\in \mathrm{cBR}^{\mathrm{thr}}(M_C)$)

$$U_C^{\mathrm{thr}}(M_P, M_C) \leqslant U_C^{\mathrm{dis}}(M'_P, M'_C) + \delta$$

The δ-value can be positive or negative. If it is negative, it gives a threshold about how much more a coercer has to gain in order to justify the use of the weapon. Positive values make sense, too: in order to bend a player, the coercer might be allowed to lose a little, taking account for uncertainty in the modelling of his utility. Note that if δ is non-negative, the best disarmed coercers are automatically economical coercers; so another property (e.g., U_P-effectiveness) must be used to make sure there is actual coercion taking place.

u_c-EFFECTIVENESS The second approach for understanding the coercer's economical effectiveness captures the following intuition: if the coercer, no matter what he does, does not have an advantage by actually having a weapon, the protocol is incoercible. In other words, if he only

has a toy gun (that is identifiable as such) and his pay-off using the toy gun is better than his pay-off using a real weapon (maybe because real bullets are very expensive), we see that the real weapon does not help the coercer.

So given that the coercer's utility for all of the player's best responses in the threat-setting (i. e., having a real gun) is lower than if he plays his strategy in the disarmed setting(i. e., with a toy gun), he is not U_C-effective:

Here we quantify over all best responses of the player, because if he is indifferent between them he might chose one which renders the coercer's strategy ineffective.

Definition 11 (U_C-effective coercer) A coercer M_C is called (computationally) U_C-effective iff for some $M_P \in \mathrm{BR}^{\mathrm{thr}}(M_C)$ ($\in \mathrm{cBR}^{\mathrm{thr}}(M_C)$) there is no best response $M'_P \in \mathrm{BR}^{\mathrm{dis}}(M_C)$ ($\in \mathrm{cBR}^{\mathrm{dis}}(M_C)$) in the disarmed setting, such that

$$U_C^{\mathrm{thr}}(M_P, M_C) < U_C^{\mathrm{dis}}(M'_P, M_C) + \delta$$

U_P-effectiveness and U_C-effectiveness are not the same. The Expensive Punishment Example (4.5) shows that a U_P-effective coercer can be U_C-ineffective.

There can also be a coercer that is U_C-effective but not economic. Assume for a moment that the utility function does also incorporate the cost of the coercer's computations needed in order to perform punishment. (Section 11.1 sketches how this could be modelled.) Recall the Bruteforce Example (4.8). Here, a coercer needs to be uneconomic in order to be U_P-effective. Still, a coercer that bruteforces the communication has a better pay-off when he is allowed to use the weapon in comparison to the disarmed case, where he still has wasted his resources on revealing the communication but misses the means to perform the punishment. Therefore this coercer would be U_C-effective. In fact, there are examples where computation does not play a role, too. The key point is that the coercer has to waste a lot of utility in order to be in a position where he can pressure the player.

But, by definition, if a coercer is not U_C-effective, it is uneconomic. This is equivalent to saying, if a coercer is economic (= not uneconomic) it is U_C-effective. Therefore the following two statements are the same:

Every U_C-effective coercer is uneconomic. (6.2)

and

Every coercer is uneconomic. (6.3)

The latter and the following statements are not equal:

Every U_P-effective coercer is uneconomic. (6.4)

For example, consider the situation in which the player wants to achieve the same goal as the coercer. It is clear that (6.3) and therefore also (6.2) imply (6.4), but under what circumstances the opposite direction holds, is the matter of the following subsection.

6.3.1 *Opposite Utilities*

It is obvious that (6.2) and (6.4) are different when the player and the coercer are interested in the same outcome. But this is not the kind of situation we would like to model. In the following we investigate the relation between those two definitions for the case of two players of contrary interests. We might gain some insights under what conditions and to what extent an economic coercer already is U_P-effective.

Assume the preference to be such that for any M_C and M_P we have

$$U_C^{dis}(M_C, M_P) = -U_P^{dis}(M_C, M_P) \tag{6.5}$$

Furthermore we introduce the following abbreviations:

$$\text{pundam}(M_P, M_C) := \underbrace{U_P^{dis}(M_C, M_P) - U_P^{thr}(M_C, M_P)}_{\geqslant 0} \tag{6.6}$$

$$\text{puncost}(M_P, M_C) := \underbrace{U_C^{dis}(M_C, M_P) - U_C^{thr}(M_C, M_P)}_{\geqslant 0} \tag{6.7}$$

By this definition, subtracting the punishment damage pundam from the player's utility under threat yields the utility in the disarmed case. Now we would like to find out whether

Every δ-U_P-effective coercer is δ-uneconomic
$\Leftrightarrow$ Every coercer is δ-uneconomic

($\Leftarrow$) This is immediately clear.

($\Rightarrow$) Assume Every U_P-effective coercer is δ-uneconomic, in other words:

Every coercer is U_P-ineffective or uneconomic
$\Leftrightarrow$
$$\begin{aligned}
&\forall M_C (\exists M_P \in BR^{thr}(M_C) \exists M'_P \in BR^{dis}(M_C) \\
&\qquad \text{s.t. } U_P^{dis}(M_P, M_C) \geqslant U_P^{dis}(M'_P, M_C) - \delta \\
&\quad \vee \\
&\quad \exists M_P \in BR^{thr}(M_C), M'_C, M'_P \in BR^{dis}(M'_C) \\
&\qquad \text{s.t. } U_C^{thr}(M_P, M_C) \leqslant U_C^{dis}(M'_P, M'_C) + \delta)
\end{aligned}$$

We know that for M_C, M_P and M'_P:

$$\begin{aligned}
& U_P^{dis}(M_P, M_C) \geqslant U_P^{dis}(M'_P, M_C) - \delta \\
\Leftrightarrow \quad & U_C^{dis}(M_P, M_C) \leqslant U_C^{dis}(M'_P, M_C) + \delta && \text{(by 6.5)} \\
\Leftrightarrow \quad & U_C^{thr}(M_P, M_C) \leqslant U_C^{dis}(M'_P, M_C) + \delta \\
& \qquad\qquad - \text{puncost}(M_P, M_C) && \text{(by 6.7)}
\end{aligned}$$

Therefore:

$$\begin{aligned}
&\forall M_C(\exists M_P \in BR^{thr}(M_C), M'_P \in BR^{dis}(M_C) \text{ s.t.} \\
&\quad U_C^{thr}(M_P, M_C) \leqslant U_C^{dis}(M'_P, M_C) + \delta - \text{puncost}(M_P, M_C) \\
&\quad \vee \\
&\quad \exists M_P \in BR^{thr}(M_C), M'_C, M'_P \in BR^{dis}(M'_C) \text{ s.t.} \\
&\quad U_C^{thr}(M_P, M_C) \leqslant U_C^{dis}(M'_P, M'_C) + \delta)
\end{aligned}$$

If this holds, then by relaxing the first condition the following holds as well:

$$\begin{aligned}
&\forall M_C(\exists M_P \in BR^{thr}(M_C), M'_P \in BR^{dis}(M_C) \\
&\quad \text{s.t. } U_C^{thr}(M_P, M_C) \leqslant U_C^{dis}(M'_P, M_C) + \delta \\
&\quad \vee \\
&\quad \exists M_P \in BR^{thr}(M_C), M'_C, M'_P \in BR^{dis}(M'_C) \\
&\quad \text{s.t. } U_C^{thr}(M_P, M_C) \leqslant U_C^{dis}(M'_P, M'_C) + \delta)
\end{aligned}$$

It is easy to see that the second condition is always fulfilled when the first one is. Hence the statement *Every U_P-effective coercer is uneconomic* equals the statement *Every coercer is uneconomic* in the case of opposite utilities.

6.4 ABSOLUTE INCOERCIBILITY AGAINST ECONOMIC COERCERS

Using these insights, we would like to formulate absolute incoercibility against economically rational coercers. We provide two different definitions here. We do not do this without a reason: there are two understandings that do not necessarily coincide. We will provide the formal definitions first, and then interpret them both.

Definition 12 (Absolute Incoercibility(1)) A coercibility game π_f with utility functions U_P and U_C is called incoercible against economic coercers iff

$$\forall M_C \in \mathfrak{M}_C : M_C \text{ is uneconomic.}$$ □

This definition expresses an outside assessment of the situation: there should be no way to get an advantage using force.

Definition 13 (Absolute Incoercibility(2)) A coercibility game π_f with utility functions U_P and U_C is called incoercible against economic coercers iff

$$\forall M_C \in \mathfrak{M}_C : M_C \text{ is uneconomic or } U_P\text{-ineffective.}$$ □

This definition expresses an assessment from the point of view of the player: there should be no economical way to coerce him into something that is unfavourable for him.

Obviously Definition 12 implies 13. In the last section we saw that they coincide for opposite utilities. However, consider an indifferent player, i.e., a player whose utility is equal on the range of f (except for punishment). It would never be possible to be $\mathcal{U}_P$-effective for such a choice of $\mathcal{U}_P$. Nevertheless, a protocol which eventually leaks sufficient information makes it possible for a coercer to pressurize the player to obtain a certain outcome. It is actually easier considering his ignorance of the outcome. For such utilities, the protocol is incoercible according to 13, but not 12.

But is there a difference between both definitions for "natural descriptions" of the classes of $\mathcal{U}_P$ and $\mathcal{U}_C$ we would like to talk about? There is, and in the following we will construct an example where the difference occurs. The main idea is that the parameters are chosen in a way that the player is only coercible if he is indifferent enough between his choices to make $\mathcal{U}_P$-efficiency unsatisfiable. As soon as the options make a difference in his pay-off that is large enough to make $\mathcal{U}_P$-efficiency possible, he would rather choose to be punished than give in, rendering the coercer uneconomic.

Let the protocol be an open vote with two choices for the player, A and B, and no participation of the coercer. The filtering function f outputs the player's choice and whether the coercer chose to punish him. We regard all $\mathcal{U}_P$ where the player has an *objective* (we will formally introduce this in a minute) of 10 — this means U_P^{dis} is smaller than 10 for any outcome. If the player is punished, he loses 1. The class of $\mathcal{U}_C$ we regard is the one with a *coercer-objective* of 10, i.e., $U_C^{dis} \leqslant 10$ for all outcomes, and is never negative. Punishment costs him 6; he can only punish once. We interpret "economic" quite strictly with $\delta = 0$, so the coercer is required to not lose anything through the coercion. $\mathcal{U}_P$-effectiveness has a $\delta = 1$, so in order to be effective the coercer has to impair the player's pay-off in the disarmed setting by at least 1. Those restrictions on $\mathcal{U}_C$ and $\mathcal{U}_P$ seem more or less realistic.

Now, whenever the difference between the player pay-off of A and B is smaller than 1, $\mathcal{U}_P$-effectiveness cannot be reached. If the difference is larger than 1, assume without loss of generality that the coercer has highest outcome if B is chosen, and the player has the highest outcome if A is chosen. We can do this because if they both prefer the same outcome again, then there is no way to achieve $\mathcal{U}_P$-effectiveness. Now the coercer can only punish once, since the cost of more punishments is at least 12, and by punishing twice or more he would not be able to be economic anymore. However, by only punishing once he cannot force the player to deviate from his optimal choice, because he would lose more by obeying than by accepting the punishment. We see that here every $\mathcal{M}_C$ is either uneconomic or $\mathcal{U}_P$-ineffective (so the protocol is incoercible according to Definition 13), but still for a difference in the player's pay-off for A and B smaller than 1, there is an economic $\mathcal{M}_C$, namely the one that punishes if the player votes in his own favour, therefore it *is* coercible according to Definition 12.

We see that this construction leaves us with the insight that rather than providing a bullet-proof cure-all definition, it makes sense to provide the notions that allow to define δ_1-economical and δ_2-$\mathcal{U}_P$-effective coercers and make it possible to use them in order to produce statements in the style of either of the above definitions. Such definitions allow clear statements about *what* kind of security there is for *which* players against coercers with *how much* resources. The next section introduces some abbreviations and recommendation on how to produce such statements.

6.5 QUALITATIVE STATEMENTS

Of course, statements such as the one above can be proven for arbitrary sets of utility functions, singleton sets being extreme cases. But normally we want to assure that a participant in a protocol has a proper choice.

A typical situation is the following: we have the ideal model of a voting system that provides secret, authenticated and yet anonymous channels for the process of election. We would like to see which amount of coercibility we face when some country uses it to perform its major election. This depends on how the tally is published, whether by city, by administration unit or whether the result is only published for the whole country.

A result stating some real implementation of a voting scheme is nearly as incoercible as this ideal model will allow us to justify the use of such a system. The next section introduces this notion of incoercibility *relative* to another system, in this case: the ideal model. What is needed to design the tally and the voting process as a whole is a qualitative measure of the absolute incoercibility of a system. It is vital to evaluate the risk of coercion versus other important properties such as verifiability, computational effort in the process and accuracy of the analysis done by the news.

For a list of requirements that are desirable for voting schemes, see Rjaskova [2002].

In many cases, the above example being one of them, there cannot be an unconditional statement of incoercibility. An analysis has to incorporate sound assumptions about how the punishment affects the player and how much it costs the coercer to punish the player. Utility functions in a game-theoretic sense are insensitive to any positive affine transformation,i. e., one can replace each pay-off x by $ax + b$ for any fixed real number $a > 0$ and b. As we have defined above in (6.6), (6.7), we use the following abbreviation:

$$\begin{aligned} \mathrm{pundam}(M_P, M_C) &:= U_P^{\mathrm{dis}}(M_C, M_P) - U_P^{\mathrm{thr}}(M_C, M_P) \geqslant 0 \\ \mathrm{puncost}(M_P, M_C) &:= U_C^{\mathrm{dis}}(M_C, M_P) - U_C^{\mathrm{thr}}(M_C, M_P) \geqslant 0 \end{aligned}$$

In most cases we assume $\mathcal{U}_P$ and $\mathcal{U}_C$ to be sane according to Definition 5. Due to the fact that it requires the functions to be risk-neutral, it allows any strategy that punishes with some fraction p of the maximal possible punishment to be rewritten as a machine that punishes with the maximal amount but only with a probability of p. Therefore, in abuse of notation, pundam might denote the maximal possible amount of

punishment that is induced, depending on the context. Symmetrically, puncost can denote the cost of the maximal possible punishment.

The damage respectively the costs of punishment is more or less the measure of worth of an objective, the amount of resources you are willing to spend on it.

One might introduce the following abbreviations:

Definition 14 (Player- and Coercer-Objective) The player's objective value is defined as

$$\text{objective} := \max_{M_P, M_C} U_P^{\text{dis}}(M_P, M_C) - \min_{M_P, M_C} U_P^{\text{dis}}(M_P, M_C) \tag{6.8}$$

The coercer's objective value is defined as

$$\text{cobjective} := \max_{M_P, M_C} U_C^{\text{dis}}(M_P, M_C) - \min_{M_P, M_C} U_C^{\text{dis}}(M_P, M_C) \tag{6.9}$$

□

We may express incoercibility by defining classes of U_C and U_P with some relation between objective and pundam, respectively cobjective and puncost, that fulfill Definition 12 or 13. Take the example of the parliamentary election in Saarbrücken (Chapter 9) as an example of how this would be done.

Using these techniques we can produce statements about the incoercibility of some ideal model, which might represent the overall system design in connection with its environment. If we now have an actual implementation that preserves the same incoercibility, we can carry this statement into the real world. What we need to have is some kind of best-possible incoercibility, a statement that says one system (e.g., the implementation) is not significantly more coercible than another system (e.g., an ideal voting scheme in a certain process).

6.6 BEST-POSSIBLE (RELATIVE) INCOERCIBILITY

As we have seen in the *Vote with Tally* example (4.4) absolute incoercibility often cannot be achieved. We need to have a notion of best-possible incoercibility, incoercibility relative to another model, often a so-called *ideal model*. We will describe the ideal model by a so-called functionality, a trusted third party that is directly addressed by the protocol parties. This ITM communicates with the player and the coercer. We need to assure that the preference relations in the ideal model and the real world relate to each other. This is achieved by making sure that the filtering function f is defined on the outcomes of both worlds and maps them to the same set on which the utility functions operate, see Definition 5.

We can enforce some kind of correlation by requiring that it is possible to achieve the same utilities in both models. We call this property solid modelling.

But first a short note on notation: when talking about the same utility functions on different protocols, we have to make clear which protocol the machines run on, since the outcome and thus the utility can be different. We will denote it in a twofold way: if U_P is applied to the outcome of M_C and M_P in the coercibility game π_f, and it is clear from the context which f is used, we will write $U_P(M_C, M_P)_\pi$. Quite often we will superscript machines running in a certain ρ with the protocol they operate in. Then we might just write $U_P(M_C^\rho, M_P^\rho)$ instead of $U_P(M_C^\rho, M_P^\rho)_\rho$.

Definition 15 (Solid Modelling) We say that coercibility games π_f, ρ_f with the same utility functions U_P, U_C have solid modelling iff for all coercers and players M_C^ρ and M_P^ρ in ρ, there exist coercers, respectively players M_C^π and M_P^π, in π such that

$$U_P^{thr}(M_C^\rho, M_P^\rho) = U_P^{thr}(M_C^\pi, M_P^\pi) \tag{6.10}$$

$$\text{and } U_C^{dis}(M_C^\rho, M_P^\rho) = U_C^{dis}(M_C^\pi, M_P^\pi). \tag{6.11}$$

□

This property does not guarantee that every pay-off in π can be reached in ρ. A protocol π might allow the player, for example, to "hack" the protocol in order to achieve an output not possible in ρ, e.g., the ideal model.

A closer connection can be established using the notions of U_P-effective and economic coercion strategies in a manner similar to the above, comparing the advantage achievable using punishment under the assumption that P behaves rationally. Intuitively speaking, if the player cannot be bent "more" in the real world than in the ideal model, we have achieved best-possible incoercibility for some set of utility function. We start with U_P-effectiveness.

Definition 16 (Relative U_P-effectiveness) A coercer M_C^π using a filtering function f defined on both protocol outcomes and utility functions U_P, U_C is called δ-U_P-effective in π relative to ρ if and only if there are $M_P^{\pi thr} \in BR^{thr}(M_C^\pi)$ and $M_P^{\pi dis} \in BR^{dis}(M_C^\pi)$ such that for all $M_C^\rho \in \mathcal{M}_C^\rho$ and all $M_P^{\rho thr} \in BR^{thr}(M_C^\rho)$ and $M_P^{\rho dis} \in BR^{dis}(M_C^\rho)$ it holds that:

$$\begin{aligned} &U_P^{dis}(M_P^{\pi dis}, M_C^\pi) - U_P^{dis}(M_P^{\pi thr}, M_C^\pi) \\ &\qquad > U_P^{dis}(M_P^{\rho dis}, M_C^\rho) - U_P^{dis}(M_P^{\rho thr}, M_C^\rho) + \delta \end{aligned}$$

In other words:

$$\begin{aligned} &\sup_{\substack{\{M_P^{\pi thr} \in BR^{thr}(M_C^\pi),\\ M_P^{\pi dis} \in BR^{dis}(M_C^\pi)\}}} \mathrm{Adv}(M_C^\pi, M_P^{\pi dis}, M_P^{\pi thr}) \\ &> \sup_{\substack{\{M_C^\rho \in \mathcal{M}_C^\rho, M_P^{\rho thr} \in BR^{thr}(M_C^\rho),\\ M_P^{\rho dis} \in BR^{dis}(M_C^\rho)\}}} \mathrm{Adv}(M_C^\rho, M_P^{\rho dis}, M_P^{\rho thr}) + \delta \\ &\text{for } \mathrm{Adv}(M_C^\pi, M_P^\pi, M_P^{\pi'}) := U_P^{dis}(M_P^\pi, M_C^\pi) - U_P^{dis}(M_P^{\pi'}, M_C^\pi) \end{aligned}$$

If $\delta = 0$ we omit δ. □

Definition 17 (computational relative U_P-effectiveness) A coercer M_C^π using a filtering function f defined on both protocol outcomes and utility functions U_P, U_C is called computationally δ-U_P-effective in π relative to ρ if and only if

$$\begin{aligned}&\sup_{\substack{\{M_P^{\pi thr} \in cBR^{thr}(M_C^\pi),\\ M_P^{\pi dis} \in cBR^{dis}(M_C^\pi)\}}} \mathrm{Adv}(M_C^\pi, M_P^{\pi dis}, M_P^{\pi thr})\\ &> \sup_{\substack{\{M_C^\rho \in \mathcal{M}_C^\rho, M_P^{\rho thr} \in cBR^{thr}(M_C^\rho),\\ M_P^{\rho dis} \in cBR^{dis}(M_C^\rho)\}}} \mathrm{Adv}(M_C^\rho, M_P^{\rho dis}, M_P^{\rho thr}) + \delta\\ &\text{for } \mathrm{Adv}(M_C^\pi, M_P^\pi, M_P^{\pi'}) := U_P^{dis}(M_P^\pi, M_C^\pi) - U_P^{dis}(M_P^{\pi'}, M_C^\pi)\end{aligned}$$

If we say a coercer is computationally U_P-effective, he is computationally δ-U_P-effective for a non-negligible, positive δ. □

The use of this notion is to prove that a protocol implementing some processes (modelled as a functionality) preserves incoercibility properties of the process in certain situations. Therefore the following definition will be given for the computational case. A non-computational variant is easy to define as well, but here we focus on the fact that the application of these notions is to use cryptographic means to implement a functionality, and to compare the implementation's coercibility properties with those of the functionality.

Definition 18 (Relative Incoercibility against Irrational Coercers) We call a protocol π with filtering function f best-possible incoercible against irrational coercers with respect to a protocol ρ (denoted $\pi \precsim_f^{irr} \rho$) if and only if

> for all sane $U_P \in \mathcal{U}_P$ every $M_C^\pi \in \mathcal{M}_C^\pi$ is not computationally U_P-effective relative to ρ.

If this property does only hold for a class of player utilities $\mathcal{U}_P$ we denote $\pi \precsim_{f,\mathcal{U}_P}^{irr} \rho$. □

We would like to prove certain properties about this relation between two protocols, such as reflexivity, transitivity and additivity with the absolute notion.

Lemma 1 (Reflexivity and Transitivity of $\precsim_f^{irr}$) *Let π, ρ and σ be protocols. Then $\pi \precsim_f^{irr} \pi$. If $\pi \precsim_f^{irr} \rho$ and $\rho \precsim_f^{irr} \sigma$ for some f, then $\pi \precsim_f^{irr} \sigma$ for the same f.* □

Proof For all U_P:

$$\begin{aligned}&\forall M_C^\pi \in \mathcal{M}_C^\pi \quad \sup_{\{M_P^{\pi thr}, M_P^{\pi dis}\}} \mathrm{Adv}(M_C^\pi, M_P^{\pi dis}, M_P^{\pi thr})\\ &\qquad \geqslant \sup_{\{M_C^\rho, M_P^{\rho thr}, M_P^{\rho dis}\}} \mathrm{Adv}(M_C^\rho, M_P^{\rho dis}, M_P^{\rho thr}) + \delta\\ &\Leftrightarrow \sup_{\{M_C^\pi, M_P^{\pi thr}, M_P^{\pi dis}\}} \mathrm{Adv}(M_C^\pi, M_P^{\pi dis}, M_P^{\pi thr})\\ &\qquad \geqslant \sup_{\{M_C^\rho, M_P^{\rho thr}, M_P^{\rho dis}\}} \mathrm{Adv}(M_C^\rho, M_P^{\rho dis}, M_P^{\rho thr}) + \delta\end{aligned}$$

Therefore we see immediately that reflexivity holds as well as transitivity. ■

Lemma 2 (Additivity of $\precsim_{\mathsf{f}}^{\mathbf{irr}}$ & Abs. Incoerc. against Irr. Coerc.) *Assume that $\pi \precsim^{\mathrm{irr}}_{\mathsf{f},\mathcal{U}_P} \rho$ for some f, protocols π,ρ and class of utility functions $\mathcal{U}_P$. Then, if for ρ_{f} there is no computationally δ-$\mathcal{U}_P$-effective poly-time coercer for any $\mathcal{U}_P \in \mathcal{U}_P$, there is also no $(\delta+\varepsilon)$-$\mathcal{U}_P$-effective one for π, ε being negligible in the security parameter.* □

Proof If there us no $\mathcal{U}_P$-effective poly-time coercer for $\forall \mathcal{U}_P \in \mathcal{U}_P$ then

$$\sup_{M_C^\rho, M_P^{\rho\mathrm{thr}}, M_P^{\rho\mathrm{dis}}} \mathrm{Adv}(M_C^\rho, M_P^{\rho\mathrm{dis}}, M_P^{\rho\mathrm{thr}}) \leqslant \delta$$

and by $\pi \precsim^{\mathrm{irr}}_{\mathsf{f},\mathcal{U}_P} \rho$

$$\sup_{M_C^\pi, M_P^{\pi\mathrm{thr}}, M_P^{\pi\mathrm{dis}}} \mathrm{Adv}(M_C^\pi, M_P^{\pi\mathrm{dis}}, M_P^{\pi\mathrm{thr}}) \leqslant \delta + \varepsilon$$

with some negligible function ε. ■

Lemma 2 implies the following: if our functionality $\mathcal{F}$ is absolutely incoercible against irrational coercers for some class of player utilities $\mathcal{U}_P^{\mathcal{F}}$ and we have an implementation π of $\mathcal{F}$ that is $\pi \precsim^{\mathrm{irr}}_{\mathsf{f},\mathcal{U}_{C_1}} \mathcal{F}$ for a superset $\mathcal{U}_P^\pi$ of $\mathcal{U}_P^{\mathcal{F}}$ we can guarantee the absolute incoercibility against irrational coercers for π, e.g., a model of the real-world, as well.

Another interesting fact is the following: if we define ρ to be π but with a filter for punishment messages, we can model absolute incoercibility as relative incoercibility with respect to a protocol constructed to emulate the disarmed setting:

Lemma 3 (Absolute Incoercibility is relative to disarmed model) *Let π be a protocol. Now let ρ be π but with a protocol party in place of the coercer that forwards any message but the punishment messages addressed to the environment. The coercer can only send via this party, as he is isolated despite a private channel to his proxy.*

A coercer M_C in a coercibility game π_{f} with utility functions U_P, U_C is U_P-effective if and only if it is U_P-effective relative to a ρ constructed as above. □

Proof For any pair of Machines (M_C^π, M_P^π) plugged into π there is the exact same pair (M_C^ρ, M_C^ρ) executed in ρ. Since all but the punishment messages are forwarded, by construction

$$\begin{aligned} U_P^{\mathrm{dis}}(M_C^\pi, M_P^\pi) &= U_P^{\mathrm{thr}}(M_C^\rho, M_P^\rho) = U_P^{\mathrm{dis}}(M_C^\rho, M_P^\rho) \quad \text{and} \\ U_C^{\mathrm{dis}}(M_C^\pi, M_P^\pi) &= U_C^{\mathrm{thr}}(M_C^\rho, M_P^\rho) = U_C^{\mathrm{dis}}(M_C^\rho, M_P^\rho) \end{aligned}$$

Then, for all f,U_P, U_C, M_C^π:

$$
\begin{aligned}
&M_C^\pi \text{ is } \delta\text{-}U_P\text{-effective relative to } \rho \\
&\Leftrightarrow \\
&\sup \mathrm{Adv}(M_C^\pi, M_P^{\pi\mathrm{dis}}, M_P^{\pi\mathrm{thr}}) > \sup \mathrm{Adv}(M_C^\rho, M_P^{\rho\mathrm{dis}}, M_P^{\rho\mathrm{thr}}) + \delta \\
&\qquad (M_P^{\pi\mathrm{dis}}, M_P^{\pi\mathrm{thr}}, M_C^\rho, M_P^{\rho\mathrm{dis}}, M_P^{\rho\mathrm{thr}} \text{ according to Def. 16}) \\
&\Leftrightarrow \sup \mathrm{Adv}(M_C^\pi, M_P^{\pi\mathrm{dis}}, M_P^{\pi\mathrm{thr}}) > \delta \quad (\text{since } \mathrm{BR}^{\mathrm{thr}}(M_C^\rho) = \mathrm{BR}^{\mathrm{dis}}(M_C^\rho)) \\
&\Leftrightarrow M_C^\pi \text{ is absolutely } \delta\text{-}U_P\text{-effective}
\end{aligned}
$$

The proof for the computational variant is analogous, you just substitute $\mathrm{BR}^{\mathrm{dis}}$ and $\mathrm{cBR}^{\mathrm{dis}}$ and $\mathrm{BR}^{\mathrm{thr}}$ by $\mathrm{cBR}^{\mathrm{thr}}$. ■

Similar to U_P-effectiveness, being economic should be measured on the best that the coercer can achieve in the ideal model:

Definition 19 (Uneconomic Coercer w.r.t. some model) A coercer M_C^π is called δ-uneconomic in π relative to ρ with respect to a filtering function f defined on both protocol outcomes and utility functions U_P and U_C if and only if there exists a coercion strategy M_C^ρ along with a best response $M_P^\rho \in \mathrm{BR}^{\mathrm{dis}}(M_C^\rho)$ in ρ such that for some best response $M_P^\pi \in \mathrm{BR}^{\mathrm{thr}}(M_C^\pi)$ the following holds:

$$U_C^{\mathrm{thr}}(M_P^\pi, M_C^\pi) \leqslant U_C^{\mathrm{thr}}(M_P^\rho, M_C^\rho) + \delta$$

Definition 20 (Uneconomic Coercer w.r.t. some model (comp.)) A coercer M_C^π is called δ-uneconomic in π relative to ρ with respect to a filtering function f defined on both protocol outcomes and utility functions U_P and U_C if and only if there exists a coercion strategy M_C^ρ along with a best response $M_P^\rho \in \mathrm{cBR}^{\mathrm{dis}}(M_C^\rho)$ in ρ such that for some best response $M_P^\pi \in \mathrm{cBR}^{\mathrm{thr}}(M_C^\pi)$ the following holds:

$$U_C^{\mathrm{thr}}(M_P^\pi, M_C^\pi) \leqslant U_C^{\mathrm{thr}}(M_P^\rho, M_C^\rho) + \delta$$

If we say a coercer is computationally uneconomic with respect to some model, he is computationally δ-uneconomic for some negligible, positive δ. □

This allows to craft a definition in the spirit of Definition 18 that additionally requires any coercer to be economic. There is a detail to be taken care of: a coercer in π might be U_P-ineffective in with respect to ρ; but every coercer in ρ that is better than him might be uneconomic himself. Therefore we propose a slightly altered version of relative U_P-effectiveness:

Definition 21 (Relative Incoercibility against Economic Coercers) We call a protocol π with filtering function f best-possible incoercible against economic coercers with respect to a protocol ρ if and only if

For all sane $U_P \in \mathcal{U}_P$ every $M_C^\pi \in \mathcal{M}_C^\pi$: M_C^π is uneconomic with respect to ρ or

$$\sup_{\substack{\{M_P^{\pi\text{thr}} \in \mathrm{BR}^{\text{thr}}(M_C^\pi),\\ M_P^{\pi\text{dis}} \in \mathrm{BR}^{\text{dis}}(M_C^\pi)\}}} \mathrm{Adv}(M_C^\pi, M_P^{\pi\text{dis}}, M_P^{\pi\text{thr}})$$
$$\leqslant \sup_{\substack{\{\text{economic } M_C^\rho \in \mathcal{M}_C^\rho,\\ M_P^{\rho\text{thr}} \in \mathrm{BR}^{\text{thr}}(M_C^\rho),\\ M_P^{\rho\text{dis}} \in \mathrm{BR}^{\text{dis}}(M_C^\rho)\}}} \mathrm{Adv}(M_C^\rho, M_P^{\rho\text{dis}}, M_P^{\rho\text{thr}}) - \delta \qquad \square$$

Note that best-possible incoercibility against economic coercers is implied by best-possible incoercibility against irrational coercers, therefore it is enough to prove that a scheme provides the latter in order to perform an analysis of absolute incoercibility against economic coercers based on an ideal model.

7

EXAMPLES REVISITED

"Da steh ich nun, ich armer Tor,
und bin so klug als wie zuvor."

"And here, poor fool, I stand once more,
No wiser than I was before."

— Johann Wolfgang von Goethe, *Faust*

In the previous chapter we introduced a framework and a number of definitions to formulate statements about the coercibility of a protocol, respectively, two protocols in relation to each other. Now it is left for us to investigate the applicability of the framework we proposed. In the following we discuss our notions by means of the examples from Chapter 4 that motivated our decisions. Every example stresses a certain aspect, so we will not analyse every details of a certain situation, but focus on those the aspects are characteristic for that situation.

7.1 OPEN VOTE

In an open vote, the coercer can send a punish signal to the environment whenever he perceives the player taking a vote that he does not want him to take. If he can afford the punishment and the player has a choice worse than his best choice in the disarmed setting, yet better than taking the punishment into account, there is a $\mathcal{U}_P$-effective coercer, i.e., a coercer lowering his pay-off in the disarmed setting. Therefore this protocol is coercible against irrational coercers for classes of player-utilities $\mathcal{U}_P$ fulfilling these conditions.

7.2 SECRET VOTE

In an absolutely secret vote, the probability that the player gets punished is independent of how he votes. We assume the filtering function does only allow to take punishment and the vote taken into account, so his best response to any coercer will take a vote that gives him the highest pay-off – which is just what he would do in the disarmed setting. Therefore for all coercing strategies M_C and best-responses $M_P \in \mathrm{BR}^{\mathrm{thr}}(M_C)$, M_P is also a best-response in the disarmed case $M_P \in \mathrm{BR}^{\mathrm{dis}}(M_C)$; hence there is no $\mathcal{U}_P$-effective coercer in this setting. So it is incoercible even with respect to irrational coercers and an arbitrary class of utility functions $\mathcal{U}_P$.

7.3 SECRET VOTE WITH RECEIPT

For non-degenerate pay-off functions there is a $\mathcal{U}_P$-effective coercer: the one that punishes if P does not provide proof for voting for a non-optimal choice. If the only way for a player to obtain such a proof is to actually vote for the party, his best-response for most natural choices of $\mathcal{U}_P$ and $\mathcal{U}_C$ is to give in, lowering the his pay-off in the disarmed setting. Thus we see this protocol is coercible against arbitrary coercers and against economic coercers (because the best response always gives in, thus punishment is never necessary).

7.4 VOTE WITH TALLY

Assume the coercer knows the tally and the distribution that is used for the votes of the others. Using a functionality $\mathcal{F}$ we are able to compute for which $\mathcal{U}_P$ the coercibility game $\mathcal{F}_f$ (using a filtering function that filters tallies and punishments) maybe incoercible against irrational coercers. How large the $\mathcal{U}_P$ is depends on the information that resides in the tally. Once we introduced a formalism for voting schemes we will do a full analysis of an example in 9.

Important here is that we can express that an implementation of this voting process does guarantee incoercibility with respect to the same set of utility functions as soon as we have a result proving the implementation incoercible against irrational coercers with respect to that functionality in a superset of $\mathcal{U}_P$.

7.5 EXPENSIVE PUNISHMENT

Definition 12 and 13 describe incoercibility against rational-behaving coercers using the notion of economy (see Definition 10). When the coercer has to buy a gun so expensive that he cannot get an higher pay-off than with a clever choice in the disarmed setting, he is uneconomic. If this weapon is the only one available, every $\mathcal{U}_P$-effective player has to buy this weapon, as otherwise they could not be better than in the disarmed setting. Therefore this example is a coercion game incoercible against economic coercers of this sort.

The story of the old lady is a bit different: if the clerk plays his best response in the threat setting, he gives in. Therefore the lady does not need to put herself in risk by spilling the glass. Thus she *is* economic, as well as $\mathcal{U}_P$-effective. The problem here is commitment: she commits to the action of spilling the glass and putting herself in risk, knowing that the clerk is rational and she will be economic when he is. What if we choose the example to be more extreme? Now assume the old lady has a shell and threatens, in case she does not get 5€, to blow herself and the bank up. The clerk knows her to be rational, so he assumes that she would not go this far. He makes some kind of worst-case analysis and sees that she might put up with being imprisoned, but a pay-off that bad

is just an empty threat. We argue for and against this way to weaken the incoercibility notion in Section 10.2.

7.6 RANDOM VOTE

Let us model the situation as a protocol where the player can send the commands `A`,`B` and `random` to a machine, which then outputs:

- `A` to the environment and the coercer; if the player sent `A` to him
- `B` to the environment and the coercer; if the player sent `B` to him
- in case the player sent `random` it chooses with probability 0.5 to send `A` to the environment and with probability 0.5 to send `B` to the environment. Independently this machine decides to forward the same information to the coercer with a probability of 0.99 but with 0.01 it sends `random` to him.

The coercion game π_f that we regard filters all but the messages sent to the environment by the described machine and whether the coercer punishes or not.

Assume the following coercer M_C:

```
If receive B or random, punish as hard as possible.
```

Assume furthermore the player's utility function to be defined like this:

$$u_P(p, v) := -\text{pundam} \cdot p + \begin{cases} \text{objective} & \text{if } v = \texttt{B} \\ 0 & \text{otherwise} \end{cases}$$

We will now analyse incoercibility against irrational coercers, i.e., determine whether there are any U_P-effective coercers. We see that for player utilities defined like this, the only way for a coercing strategy to lower the disarmed pay-off is to make the player choose an option different from `B`, as this is his best-response in the disarmed case.

Since there is no communication between C and P going on in the protocol (besides the implicit announcement of the C's strategy when speaking of a best-response), the utility of any M_P against M_C can be expressed as:

$$U_P^{\text{thr}}(M_C, M_P) = (p_b + \frac{p_r}{2})\text{objective} - (p_b + \frac{p_r}{100})\text{pundam}$$

with $p_b = \Pr[M_P \text{ sends } \texttt{B}]$ and $p_r = \Pr[M_P \text{ sends } \texttt{random}]$

$$= p_b(\text{objective} - \text{pundam}) + p_r(\frac{\text{objective}}{2} - \frac{\text{pundam}}{100})$$

CASE 1:($\text{objective} < \frac{\text{pundam}}{50}$) If p_b or p_r are chosen non-zero, the utility is below zero, so it is a best response to give in and vote for `A` with probability 1. Therefore, for any $\delta < \text{objective}$, M_C is δ-U_P-effective. Because the difference between maximal and minimal player-utility in the disarmed setting is at most objective, there is no δ-U_P-effective coercer for a larger δ.

CASE 2: ($\frac{\text{pundam}}{50} \leqslant \text{objective} < \text{pundam}$) In this case the player's utility is maximized by choosing $p_r = 1$, i.e. by pressing the magic button whenever possible. The disarmed pay-off in this case is $\frac{\text{objective}}{2}$; hence M_C is δ-U_P-effective for any $\delta < \frac{\text{objective}}{2}$. Is there an M_C that is δ-U_P-effective for a larger δ? The player's decision which message to send is independent of any communication with and punishment imposed by the coercer. It is so-called *cheap talk*. As the channel is secret, there is nothing the player can provide the coercer with that does depend on his vote and cannot be computed otherwise. Hence, without loss of generality, the coercer lets his decision to punish merely depend on the information he receives through the machine. Any gradual punishment $p \cdot \text{pundam}$ can be expressed as a full punishment with probability p, since the expected Utility remains unchanged and we assume the utilities to be *sane*(Definition 5. If q_a, q_b, q_r denote the probabilities that he punishes when receiving `A`, `B`, `random` we have:

$$\begin{aligned} U_P^{\text{thr}}(M_C, M_P) =& p_a q_a \text{pundam} + p_b(\text{objective} - q_b \text{pundam}) \\ &+ p_r\left(\frac{\text{objective}}{2} - \frac{q_r \cdot \text{pundam}}{100}\right) \end{aligned}$$

We see that q_a only makes `A` more attractive, so it is zero. Since in this case it is not possible to achieve $(\text{objective} - q_b \cdot \text{pundam}) > 0$, the value of q_r is irrelevant. So q_b is the only parameter that can be tuned, and since the only option left is to make the player chose `random`, we know it is 1. We see that this player is not δ-U_P-effective for $\delta \geqslant \frac{\text{objective}}{2}$, so there is none that fulfills that property (although, admittedly, a strategy with $q_r = 0$ is cheaper for the coercer and thus more economic).

CASE 3: ($\text{objective} > \text{pundam}$) In this case every best response to M_C votes for `B` with probability 1, therefore M_C is not U_P-effective. Furthermore there is no other M'_C that is U_P-effective in this scenario: for any execution of the game with some fixed randomness, even in the case of full punishment, choosing different from `B` produces a higher damage than the eventual punishment imposed by C.

We see that we can state very exactly under which assumptions a protocol is incoercible and under which not.

7.7 STRANGE COERCERS

The best response to a *Hello Coercer* is a strategy that answers with "Hello", for the *Anti-Hello Coercer* it is one that does not answer. In both cases this does not affect the player utility for a natural choice of a filtering function and thus none of them is $\mathcal{U}_P$-effective. The same holds for the *Always Punisher*, if the punishment is independent of the player's actions he will take the actions that are best for him in the disarmed case, at least if the utility function is such that the effect of punishment is independent of the utility in the disarmed setting.

7.8 BRUTEFORCE EXAMPLE

Assume a model that takes, besides the output of the filtering function, the machine itself into account, e.g., by utilising a function that defines the computational costs of the set of machines. A machine that bruteforces the key is $\mathcal{U}_P$-effective: the player's best-response in the threat-setting yields a far worse result than in the disarmed setting, where all the computational effort is more or less wasted for nothing. This matches our intuition: yes, there is a way to coerce the player, yet it is not *clever* to do so, as a huge computational effort is needed to achieve effectiveness.

The idea is discussed in more detail in Section 11.1.

Clever here means rational when it comes to economy: the cost of computation (assuming a reasonable utility function $\mathcal{U}_C$) are higher than any potential gain and therefore there is some other coercion strategy in the disarmed setting that has a better pay-off. Hence we see that any $\mathcal{U}_P$-effective strategy is uneconomic, so such a scheme would be incoercible against economic coercers.

Another way to deal with this model is to restrict the machine profile the coercer can choose from, $\mathcal{M}_C$, to poly-time ITMs, so that it is not possible to do an exhaustive search on the key-space. In this case, there is no $\mathcal{U}_P$-effective coercion strategy.

7.9 LEGAL WEAPONS

Recall the utility functions given in Figure 1. An irrational coercer can be $\mathcal{U}_P$-effective for any $\delta < 6$ by using punishment to force the player to vote for X.

But there is a strategy in the disarmed setting that yields the coercer a higher pay-off: if he announces to vote for X unless the player votes for C, he gains a pay-off of 1 without any punishment costs. The first coercer is uneconomic, since even in the disarmed setting this strategy yields a better pay-off than the above. In the described case of an open vote, there is no punishment necessary if the player uses his best response, i.e. votes for C. Therefore one could threaten to use the actual punishment instead of the legal weapons, the outcome is the same. This solution is $\mathcal{U}_P$-effective and economic as well.

But in most protocols that are coercible but not fully open, there is some chance to punish "wrongly", i.e., having a false positive when detecting a deceiving player. In this case, voting for X would decrease the coercer's pay-off only by 1, punishment (depending on U_C) possibly more, rendering this course of action uneconomic, too.

7.10 COERCIVE COMPUTING

Imagine a protocol where you can check with sufficiently high probability whether a given key is a correct key with respect to those messages. It is easy to imagine a coercer which checks whether the key is correct and punishes if it is not, or if some undesired message is send over this channel. Such a coercer is indeed a successful one, because it yields a higher pay-off in the threat setting than in the disarmed setting.

7.11 ANNOUNCED RANDOMNESS

In this example, it is obvious that the best coercer for both variations punishes if he is informed about the player disobeying. In variation (a) the best response of the player is to vote like he would do in the disarmed setting when the channel is announced secure. For U_P where pundam is higher than the difference between disobeying and giving in (let us call it $\mathsf{objective}$) he will give in as soon as the channel is announced insecure. This leaves the coercing strategy δ-U_P-effective for $\delta < \frac{\mathsf{objective}}{100}$.

In variation (b) the player's best reply depends on U_P: If $\frac{\mathsf{objective}}{100} > \mathsf{pundam}$, he will disobey and accept the expected punishment value. In this case his utility in the disarmed setting remains unchanged, so the coercer is not U_P-effective. If $\frac{\mathsf{objective}}{100} \leqslant \mathsf{pundam}$, it is a best response to give in, therefore the U_C-effectiveness is maximal.

This goes well along with the intuition: faced with direct punishment it is clever to avoid it when possible, but if there is a limited risk, the player's reaction depends on his risk assessment.

7.12 FAULTY IMPLEMENTATION AND MEAN CHOICE

The coercer M_C votes for C and punishes the player if being signalled that he did vote for P. In the implementation, M_C is (absolutely) U_P-effective, it imposes a difference of 1 in the player's best response's utility between the disarmed and the threat setting. (Which is as good as it can get.) In the ideal world, there is no U_P-effective coercer: choosing X as a legal weapon to punish the player does not work, because no information about the player's vote is leaked, see *Secret Vote* (7.2). Thus it goes with our intuition and with Lemma 2 that M_C is U_P-effective with respect to the ideal world: although X is an option in both models that hurts the player more, the difference between the disarmed setting and the threat setting is zero in the ideal model; while in the real world M_C is

$\mathcal{U}_P$-effective for all $\delta < 1$. With puncost chosen small enough, economy is also given.

8

ON THE EXAMPLE OF VOTING SCHEMES

I consider it completely unimportant who in the party will vote, or how; but what is extraordinarily important is this — who will count the votes, and how.

— Joseph Stalin, quoted from *Boris Bazhanov's Memoirs of Stalin's Former Secretary, 1992*

The last chapters served the goal of establishing a general notion of incoercibility that is well-founded and intuitive. For this chapter, we will focus on the special case of voting schemes. Although not being the only scenario where a notion of incoercibility is needed, it is probably the most common one. First and foremost we investigate the relation between our game-theoretic notion of coercion resistance and the UC/c security notion from Unruh and Müller-Quade [2010]. For self-containment we included the definitions 22, 23, 25, 26 and 27 from their paper:

The first definition describes what we regard as voting schemes:

Definition 22 ([UC/c2010], Voting scheme) Fix sets $\mathcal{V}$ (the set of votes), $\mathcal{T}$ (the set of tallies), $\mathcal{P}$ (the set of voters). A tally function is an efficiently computable function $\mathsf{tally} : (\mathcal{V} \cup \{\bot\})^{\mathcal{P}} \to \mathcal{T}$.

A voting scheme for tally is a two-stage protocol. We call the stages voting phase and tallying phase. In such a protocol, each party $P_i \in \mathcal{P}$ gets an input $v_i \in \mathcal{V} \cup \{\bot\}$ (the vote of P_i). $v_i = \bot$ means that the P_i does not participate in the protocol (abstention). In the end of the tallying phase a distinguished party T outputs a value $t \in \mathcal{T}$. □

In typical schemes, $\mathcal{V}$ would be the set of all candidates or the set of lists of candidates ordered by preference. $\bot$ denotes abstention. The tally function $\mathsf{tally}(v_1, \ldots, v_n)$ specifies the correct tally for the votes $v_i \in \mathcal{V} \cup \{\bot\}$. Note that the participating parties, except for T, are not necessarily aware of whether they are in the tallying phase or in the voting phase.

The following functionality specifies an ideal voting scheme. It is designed to satisfy reasonable formalisations of properties that are natural for voting schemes, such as correctness (the tally is correct although an adversary is present) and anonymity (it is not possible to determine who voted for whom more precisely than deducible from the tally).

Definition 23 ([UC/c2010], Voting functionality) The voting functionality $\mathcal{F}_{\text{vote}} = \mathcal{F}_{\text{vote}}^{\text{tally}}$ expects (at most one) message $v_i \in \mathcal{V}$ from each party $P_i \in \mathcal{P}$. When receiving `tally` from T, $\mathcal{F}_{\text{vote}}$ sets $v_i := \bot$ for all $P_i \in \mathcal{P}$ from which it did not receive a message $v_i \in \mathcal{V}$ yet. Then $\mathcal{F}_{\text{vote}}$ computes $t := \mathsf{tally}(v_i, i \in \mathcal{P})$ (the tally) and sends t to the adversary. Then, when $\mathcal{F}_{\text{vote}}$ receives `deliver` from the adversary, it sends t to the party T. □

8.1 BEST-POSSIBLE INCOERCIBILITY IS IMPLIED BY UC/C INCOERCIBILITY

The main goal of the preceding chapters was to provide convincing arguments that the definition of best-possible incoercibility models the player's and the coercer's incentives to coerce in a plausible way. While we hope that this has been achieved, we are aware that existing notions are better to work with, because they are modelled with close regards to the cryptographic setting. We would like to use such notions to prove a scheme secure, while at the same time achieving best-possible incoercibility, since it appears to model the real-world setting in a more plausible way. If an existing notion can be proven to imply best-possible incoercibility under sound assumptions, it can be considered to stand on a stable foundation.

In order to be able to use our game-theoretic notions, we first transform a voting scheme into a number of incoercibility games in a straightforward manner, including the assumptions that we put upon the utility functions:

Definition 24 (Best-Possible Incoercible Voting Scheme) For a voting scheme π let $\pi_f^{P,\mathcal{B}}$ be the incoercibility game where the player takes the place of P(i. e., P is given the identity $\mathtt{P}$), the coercer the place of the adversary (i. e., the adversary is given the identity $\mathtt{C}$) and the other player's votes are drawn using $\mathcal{B}$. The filtering function f : outcome $\to \mathcal{T} \times \mathbb{R}$ outputs the tally and the punishment value sent to the environment.

A voting scheme is incoercible iff for all voters $P \in \mathcal{P}$ and every efficiently sampleable distribution $\mathcal{B}$ on $(\mathcal{V} \cup \{\bot\})^{\mathcal{P}\setminus\{P\}}$ (the votes of the other voters) $\pi_f^{P,\mathcal{B}}$ is best-possible incoercible against arbitrary adversaries with respect to the ideal model $(\mathcal{F}_{\text{vote}}^{\text{tally}})_f^{P,\mathcal{B}}$ under the following additional assumptions:

A. The player does neither fear punishment too much, nor value an objective too much: i. e., u_P^{dis} and u_P^{thr} are polynomially bounded from above and below

B. U_P is sane with respect to the function f.

C. $\mathcal{M}_P$ and $\mathcal{M}_C$ are both restricted to probabilistic polynomial-time Turing machines. □

Definition 25 ([UC/c2010], UC/c) Let π and ρ be protocols. We say that π *UC/c emulates* ρ if for any polynomial-time deceiver $\mathcal{D}$ there exists a polynomial-time deceiver $\mathcal{D}_S$ (the deceiver-simulator) such that for any polynomial-time adversary $\mathcal{A}$ there exists a polynomial-time adversary $\mathcal{A}_S$ (the adversary-simulator) such that for any polynomial-time environment $\mathcal{Z}$ the following networks are indistinguishable:

$$\pi \cup \{\mathcal{A}, \mathcal{D}_S, \mathcal{Z}\} \qquad \text{and} \qquad \rho \cup \{\mathcal{A}_S, \mathcal{D}, \mathcal{Z}\}.$$ □

Definition 26 ([UC/c2010], Dummy-adversary,dummy-deceiver) The dummy-adversary $\tilde{\mathcal{A}}$ is an adversary that, when receiving a message (id, m) from the environment, sends m to the party with identity id, and that, when receiving m from a party with identity id, sends (id, m) to the environment. The dummy-deceiver $\tilde{\mathcal{D}}$ is defined analogously. □

Definition 27 ([UC/c2010], UC/c w.r.t. dummy-adversary/deceiver) Let π and ρ be protocols. We say that π *UC/c emulates* ρ *with respect to the dummy-adversary/deceiver* if there exists a polynomial-time deceiver $\tilde{\mathcal{D}}_S$ (the dummy-deceiver-simulator) and a polynomial-time adversary $\tilde{\mathcal{A}}_S$ (the dummy-adversary-simulator) such that for any polynomial-time environment $\mathcal{Z}$ the following networks are indistinguishable:

$$\pi \cup \{\tilde{\mathcal{A}}, \tilde{\mathcal{D}}_S, \mathcal{Z}\} \qquad \text{and} \qquad \rho \cup \{\tilde{\mathcal{A}}_S, \tilde{\mathcal{D}}, \mathcal{Z}\}$$ □

We would like to prove

Theorem 1 *Let π be a voting scheme for the tally function* tally*. Assume that π UC/c emulates $\mathcal{F}_{\mathrm{vote}}^{\mathrm{tally}}$ with static corruption/deception. Then π is a best-possible incoercible voting scheme under the following assumption:*

- *The dummy-adversary-simulator $\tilde{A}_S$ used to achieve UC/c with respect to dummy-adversary/deceiver can be split into one separate machine for each voter and the adversary such that those split-$\tilde{A}_S$ while preserving indistinguishability to the original $\tilde{A}_S$ without introducing new channels to the protocol or sharing a state; (We will abuse notation by calling all those simulators $\tilde{A}_S$).* □

Proof Fix the machine $P \in \mathcal{V}$ the player substitutes and the distribution $\mathcal{B}$ the rest of the players' votes are drawn from. Assuming that π UC/c emulates $\mathcal{F}_{\mathrm{vote}}^{\mathrm{tally}}$(in the following: $\mathcal{F}$), our goal is to show that for all M_C^π and $M_P^{\pi\mathrm{dis}} \in \mathrm{cBR}^{\mathrm{dis}}(M_C^\pi)$, $M_P^{\pi\mathrm{thr}} \in \mathrm{cBR}^{\mathrm{thr}}(M_C^\pi)$ there are $M_C^\mathcal{F}$ and $M_P^{\mathcal{F}\mathrm{dis}} \in \mathrm{cBR}^{\mathrm{dis}}(M_C^\mathcal{F})$, $M_P^{\mathcal{F}\mathrm{thr}} \in \mathrm{cBR}^{\mathrm{thr}}(M_C^\mathcal{F})$ such that:

$$\begin{aligned} &U_P^{\mathrm{dis}}(M_C^\pi, M_P^{\pi\mathrm{dis}}) - U_P^{\mathrm{thr}}(M_C^\pi, M_P^{\pi\mathrm{thr}}) \\ &\qquad \leqslant U_P^{\mathrm{dis}}(M_C^\mathcal{F}, M_P^{\mathcal{F}\mathrm{dis}}) - U_P^{\mathrm{thr}}(M_C^\mathcal{F}, M_P^{\mathcal{F}\mathrm{thr}}) - \delta \end{aligned}$$

for some negligible δ.

We will show that this holds by constructing $M_C^\mathcal{F}$, $M_P^{\mathcal{F}\mathrm{dis}}$ and $M_P^{\mathcal{F}\mathrm{thr}}$ for fixed M_C^π, $M_P^{\pi\mathrm{dis}}$ and $M_P^{\pi\mathrm{thr}}$ such that:

A. $M_P^{\mathcal{F}\mathrm{dis}} \in \mathrm{cBR}^{\mathrm{dis}}(M_C^\mathcal{F})$

B. $M_P^{\mathcal{F}\mathrm{thr}} \in \mathrm{cBR}^{\mathrm{thr}}(M_C^\mathcal{F})$ and

C. $U_P^{\mathrm{dis}}(M_C^\pi, M_P^{\pi\mathrm{dis}}) - U_P^{\mathrm{thr}}(M_C^\pi, M_P^{\pi\mathrm{thr}}) \approx U_P^{\mathrm{dis}}(M_C^\mathcal{F}, M_P^{\mathcal{F}\mathrm{dis}}) - U_P^{\mathrm{thr}}(M_C^\mathcal{F}, M_P^{\mathcal{F}\mathrm{thr}})$

$A \approx B$ is notation for $|A - B| < \varepsilon$ for a negligible ε. This relation is closed under addition and subtraction.

Let $\tilde{A}$, $\tilde{D}$, $\tilde{A}_S$ and $\tilde{D}_S$ be defined according according to Definition 26 and 27.

The adversary-simulator $\tilde{A}_S$ does communicate with the protocol participants in π and (as he is giving his input to $\mathcal{F}$) outputs a vote $v \in \mathcal{V}$.

By $M \leftrightarrow \tilde{A}_S$ we will denote a machine simulating M and $\tilde{A}_S$ in communication with each other, while sending all communication that is not between those two machines to the network it is in. This means that M emulates and communicates with an instance of the split extractor $\tilde{A}_S$ in order to output the vote.

Now we construct:

$$\begin{aligned} M_C^{\mathcal{F}} &:= M_C^{\pi} \leftrightarrow \tilde{A}_S \\ M_P^{\mathcal{F}\text{dis}} &:= M_P^{\pi\text{dis}} \leftrightarrow \tilde{A}_S \\ M_P^{\mathcal{F}\text{thr}} &:= M_P^{\pi\text{thr}} \leftrightarrow \tilde{A}_S \end{aligned}$$

We first show that $M_P^{\mathcal{F}\text{dis}} \in \text{cBR}^{\text{dis}}(M_C^{\mathcal{F}})$: we will proceed in three steps, the full proof sketch can be found in Figure 4.

Let Z_1 be the environment that first sends a corruption request to the machine P. Then it simulates $M_P^{\pi\text{dis}}$ and M_C^{π} internally, sending the instruction from $M_P^{\pi\text{dis}}$ to P and the instructions from M_C^{π} to the adversary. The environment applies f to the outcome and computes u_P^{dis}. By construction of Z_1 this is just the same as computing $u_P^{\text{dis}}(M_C^{\mathcal{F}}, M_P^{\mathcal{F}\text{dis}})$. The next step is visualized in Figure 1. We use the UC/c property to establish indistinguishability between those two games, the one we just described on the left-hand side and an execution of π with $M_P^{\pi\text{dis}}$ and M_C^{π} in place of the machines controlled by the player of the coercer on the right-hand side. In other words, the expected value of the environment's output in this network is $U_P^{\text{dis}}(M_C^{\pi}, M_P^{\pi\text{dis}})$.

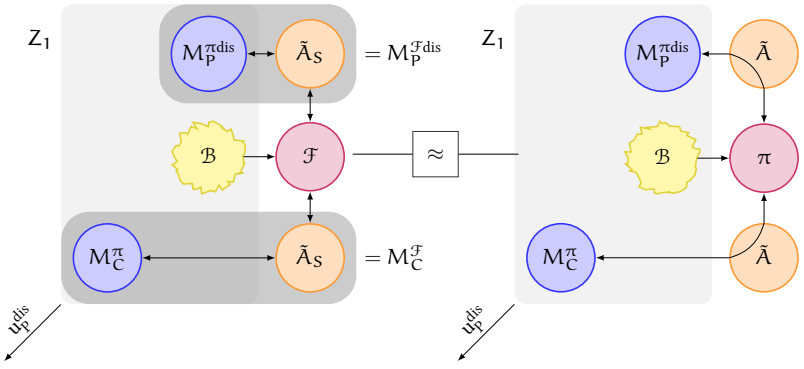

Figure 1: Proof Sketch, Step 1

Therefore

$$U_P^{\text{dis}}(M_C^{\mathcal{F}}, M_P^{\mathcal{F}\text{dis}}) \approx U_P^{\text{dis}}(M_C^{\pi}, M_P^{\pi\text{dis}}) \tag{8.1}$$

as a negligible difference in the expected value would indicate that one can make out a difference in the actual value of u_P^{dis} with non-negligible probability, by Definition 24, Item A.

Any other machine's utility has to be lower than $U_P^{dis}(M_C^{\pi}, M_P^{\pi dis}) + \varepsilon$ against M_C^{π} in the disarmed setting for some negligible ε, because $M_P^{\pi dis} \in cBR^{dis}(M_C^{\pi})$. For example the utility of $\hat{M}_P^{\mathcal{F}dis} \leftrightarrow \tilde{D}_S$ for arbitrary but fixed $\hat{M}_P^{\mathcal{F}dis}$. We choose $\hat{M}_P^{\mathcal{F}dis} \leftrightarrow \tilde{D}_S$ in this step because using the $\tilde{D}_S$ we can transfer the pay-off $\hat{M}_P^{\mathcal{F}dis}$ has in the ideal world into the real world. Any $\hat{M}_P^{\mathcal{F}dis}$ that is more successful would therefore contradict $M_P^{\pi dis} \in cBR^{dis}(M_C^{\pi})$. Intuitively speaking, even using the deception strategy encoded in $\tilde{D}_S$, no player can have a higher pay-off than the best-response to M_C^{π}.

We construct the environment Z_2 as follows: again the machine P is corrupted. Internally we simulate $\hat{M}_P^{\mathcal{F}dis}$ in communication with $\tilde{D}_S$ and forward the output of $\tilde{D}_S$ to $\tilde{A}$, who in turn just forwards the communication to π. Ditto for M_C^{π}, whose communication with π is again forwarded by $\tilde{A}$. Again, the environment applies f to the outcome and computes computes u_P^{dis}. The output is $u_P^{dis}(M_C^{\pi}, \hat{M}_P^{\mathcal{F}dis} \leftrightarrow \tilde{D}_S)$, as you can easily see. The output's expected value is smaller than $U_P^{dis}(M_C^{\pi}, M_P^{\pi dis}) + \varepsilon$, for some negligible ε, since $M_P^{\pi dis} \in cBR^{dis}(M_C^{\pi})$.

$$U_P^{dis}(M_C^{\pi}, \hat{M}_P^{\mathcal{F}dis} \leftrightarrow \tilde{D}_S) \leqslant U_P^{dis}(M_C^{\pi}, M_P^{\pi dis}) + \varepsilon \tag{8.2}$$

Since $\tilde{A}_S$ only forwards the communication sent through $\tilde{D}_S$ we can leave it out as well, see the right-hand side of Figure 2:

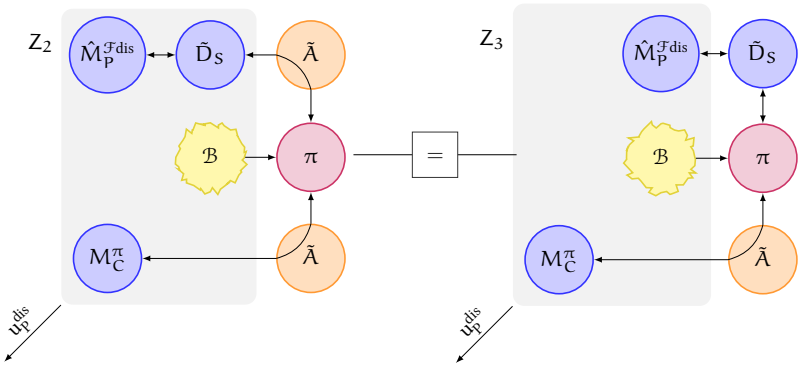

Figure 2: Proof Sketch, Step 2

The networks are doing the same, we just define the environment Z_3 to send a deception request to P and simulate $\hat{M}_P^{\mathcal{F}dis}$ and M_C^{π} communicating with the external deceiver, respectively the external adversary, which are $\tilde{D}_S$ and $\tilde{A}$ in this network. By UC/c this is indistinguishable from the network we can see on the left-hand side of Figure 3.

In this network the communication of $\hat{M}_P^{\mathcal{F}dis}$ is forwarded through $\tilde{D}$, and M_C^{π} communicates via $\tilde{A}_S$, which is by definition $M_C^{\mathcal{F}}$. $\tilde{A}_S$ and $\tilde{D}$ communicate with the voting functionality $\mathcal{F}$. Thus the output of this network is $u_P^{dis}(M_C^{\mathcal{F}}, \hat{M}_P^{\mathcal{F}dis})$. So it holds for an arbitrary $\hat{M}_P^{\mathcal{F}dis}$ that

$$U_P^{dis}(M_C^{\pi}, \hat{M}_P^{\mathcal{F}dis} \leftrightarrow \tilde{D}_S) \approx U_P^{dis}(M_C^{\mathcal{F}}, \hat{M}_P^{\mathcal{F}dis}) \tag{8.3}$$

(By the same argumentation as before.)

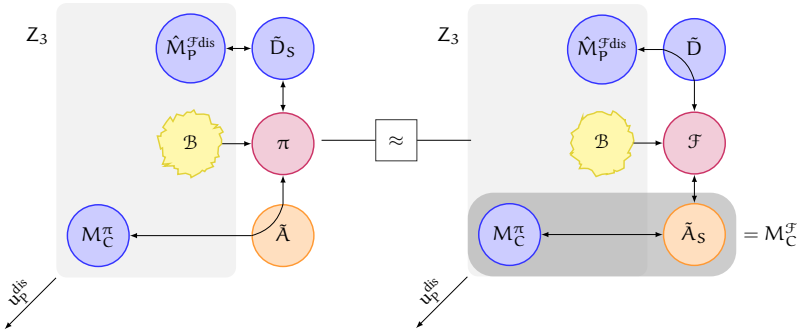

Figure 3: Proof Sketch, Step 3

So by putting together (8.3), (8.2) and (8.1), we can conclude that for all $\hat{M}_P^{\mathcal{F}dis}$ it holds that $U_P^{dis}(M_C^{\mathcal{F}}, \hat{M}_P^{\mathcal{F}dis}) \leqslant U_P^{dis}(M_C^{\mathcal{F}}, M_P^{\mathcal{F}dis}) + \varepsilon$, therefore $M_P^{\mathcal{F}dis} \in cBR^{dis}(M_C^{\mathcal{F}})$.

The proof for $M_P^{\mathcal{F}thr} \in cBR^{thr}(M_C^{\mathcal{F}})$ is literally the same, except for substituting "dis" by "thr" in every superscript. We especially like to emphasize that through this proof we gain the following statement similar to (8.1).

$$U_P^{thr}(M_C^{\mathcal{F}}, M_P^{\mathcal{F}thr}) \approx U_P^{thr}(M_C^{\pi}, M_P^{\pi thr}) \tag{8.4}$$

What is left to show is that $U_P^{dis}(M_C^{\pi}, M_P^{\pi dis}) - U_P^{thr}(M_C^{\pi}, M_P^{\pi thr}) \approx U_P^{dis}(M_C^{\mathcal{F}}, M_P^{\mathcal{F}dis}) - U_P^{thr}(M_C^{\mathcal{F}}, M_P^{\mathcal{F}thr})$. We gain this result by subtracting (8.4) from (8.1). ■

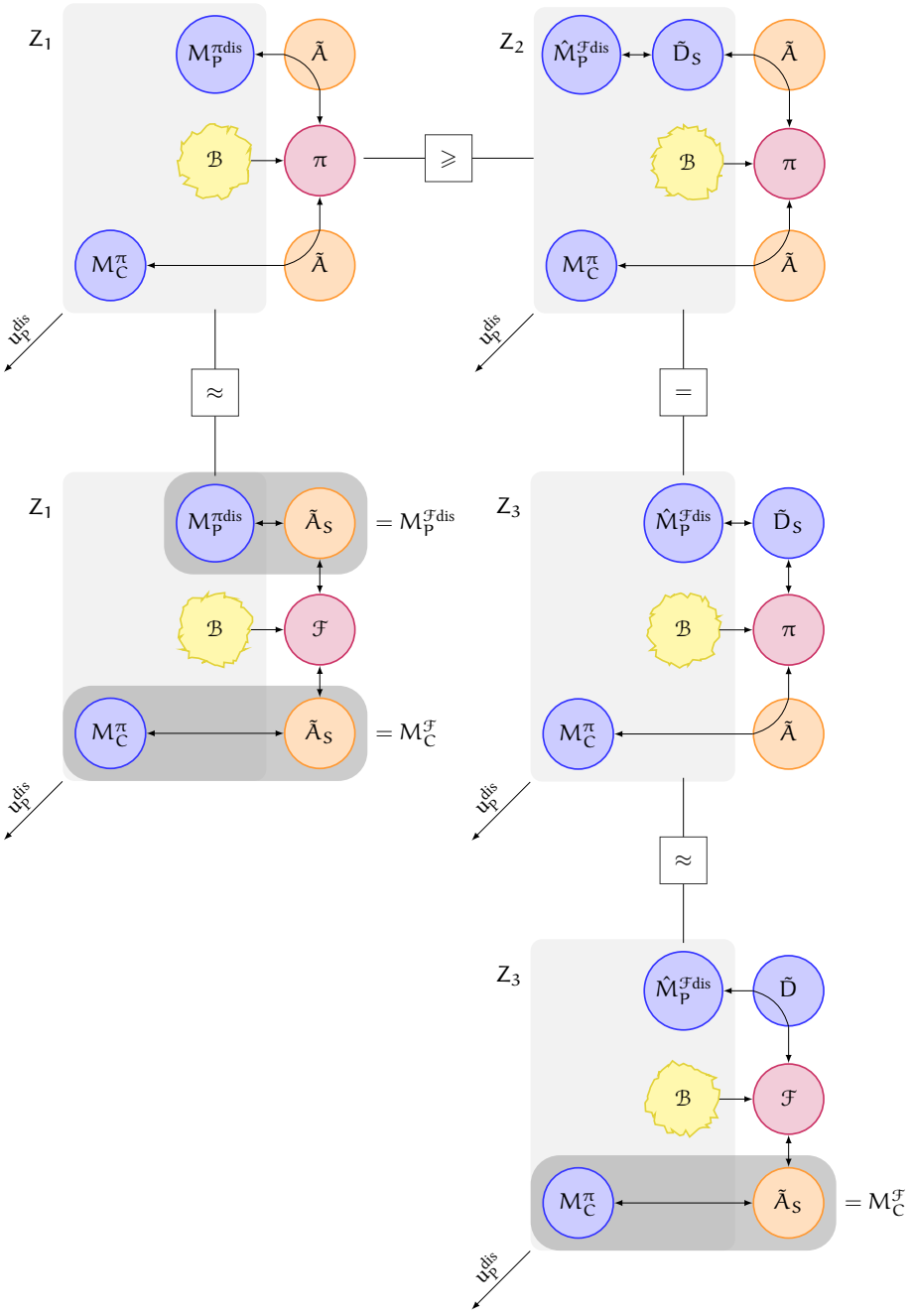

Figure 4: Proof Sketch $\mathsf{M}_\mathsf{P}^{\mathcal{F}\text{dis}} \in \text{cBR}^{\text{dis}}(\mathsf{M}_\mathsf{C}^{\mathcal{F}})$

9 A CASE STUDY IN ELECTRONIC VOTING

8. In appreciation of your useful contributions to discussion, the 10,000 allow you to vote if they are deadlocked; they commit themselves to this procedure. After the discussion you mark your vote on a slip of paper, and they go off and vote. In the eventuality that they divide evenly on some issue, 5,000 for and 5,000 against, they look at your ballot and count it in. This has never yet happened; they have never yet had occasion to open your ballot.

— Robert Nozick, *The Tale of the Slave,* from: *Anarchy, State, and Utopia*

Now that we have introduced enough notation to make qualitative statements, we think that it is useful to provide an example in order to illustrate the work-flow that is necessary to derive such a statement. Using the results from the last chapter, one might transfer such a qualitative result to a real implementation, assuming that a voting scheme that suffices the necessary assumptions exists.

The question that arises in the ideal model of a voting scheme concerns the elections we take part in. There is some chance to coerce a person merely using the tally, and depending on how the tally is published, this chance rises. The aim of this chapter is to illustrate the work-flow on the example of an election district in the German parliamentary elections.

Note that the way we proceed is a compromise between taking all the steps that are necessary to build a model and state the amount of coercibility, and keeping the model simple enough to not distract from the work flow which we aim to describe. A serious analysis would be of great use. It would need a justification of the model that is the basis for the behaviour of other voters and the coercer's knowledge about it.

We take the example of a parliamentary election in the election district Saarbrücken. We focus on the second vote and simplify the election process a bit. In 2009, the citizens of Saarbrücken (election district no. 296, see Egeler et al.) had the choice between ten parties for their second votes, so we define:

$$\begin{aligned}
\mathcal{V} =& \{\text{SPD}, \text{CDU}, \text{DIE LINKE}, \text{FDP}, \text{GRÜNE}, \text{FAMILIE},\\
& \text{NPD}, \text{MLPD}, \text{PIRATEN}, \text{RRP}, \text{invalid}\}\\
\mathcal{P} =& \{1, \dots, n = 207'292\}\\
\text{tally}(\mathbf{V}) =& (|\{v \in \mathbf{V} : v = \text{SPD}\}|, \cdots |\{v \in \mathbf{V} : v = \text{invalid}\}|,\\
& |\{v \in \mathbf{V} : v = \bot\}|), \qquad \mathbf{V} \in (\mathcal{V} \cup \{\bot\})^{\mathcal{P}}\\
\mathcal{T} =& \text{range}(\text{tally})
\end{aligned}$$

The tally function simply counts all the votes. According to the the definition of a best-possible incoercible voting scheme (Definition 24)

the filtering function f outputs the tally and the amount of punishment requested by the coercer. The coercer does not take part in the election process.

The crucial difficulty is to model the knowledge an adversary has about how the voters behave. We will model the distribution $\mathcal{B}$ in a very simplistic manner. We assume that he has the results of the last election. Furthermore, we assume that the other voters behave like the following random process: Every voter is equal, and votes independent of the others randomly with a probability $p_{SPD}, \ldots, p_{\perp}$ for some party (or chooses to abstain from voting). Of course $\sum_{v \in \mathcal{V} \cup \{\perp\}} p_v = 1$. The probabilities are the percentages of the second votes of the 2009 election, which are as follows:

PARTY	VOTES	PERCENTAGE
SPD	34'528	16.656
CDU	38'317	18.484
DIE LINKE	34'666	16.723
FDP	17'651	8.515
GRÜNE	12'685	6.119
FAMILIE	1'596	0.769
NPD	1'737	0.837
MLPD	112	0.0540
PIRATEN	2'536	1.223
RRP	752	0.362
invalid	2'133	1.028
$\perp(\mathtt{absent})$	60'579	29.223

Table 2: Percentages in seconds votes in parliamentary election 2009, election district Saarbrücken

The distribution $\mathcal{B}$ of this process is a multinomial distribution. The probability of a certain tally $\Pr[(x_{SPD}, \ldots, x_{\perp}) \leftarrow \mathcal{B}] =: \mathcal{B}(x_{SPD}, \ldots, x_{\perp})$ is:

$$\mathcal{B}(x_{SPD}, \ldots, x_{\perp}) = \begin{cases} \frac{n!}{x_{SPD} \cdots x_{\perp}} p_{SPD}^{x_{SPD}} \cdots p_{\perp}^{x_{\perp}}, & \text{if } \sum_{v \in \mathcal{V} \cup \{\perp\}} x_v = n \\ 0 & \text{otherwise} \end{cases}$$

We need to restrict the player's utility function in some meaningful way, in order to avoid deranged utilities, e. g., a player that aims for some party having a prime number of votes. Hence we argue about utilities that linearly depend on the number of votes that the parties receive, formally:

$$u_P^{dis}(M_C, M_P) = \sum_{v \in \mathcal{V} \cup \{\perp\}} u_v \cdot x_v$$

for an outcome $o(M_C, M_P) = (x_{SPD}, \cdots, x_\perp) \in \mathcal{T}$. This allows for encoding the preferences the player has, ranging from indifference between two parties to absolute dislike of another.

Since the coercer does not participate in the vote and the other player's decisions are independent from any inputs, the player's best-response M'_P to any coercer M_C in the disarmed setting is to vote for his favourite party, i. e., $V_P := \arg\max_{v \in \mathcal{V}} u_v$. The pay-off in this situation is

$$U_P^{dis}(M'_P, M_C) = u_{V_P} + \sum_{v \in \mathcal{V} \cup \{\perp\}} u_v \cdot E[X_v]$$

where $E[X_v]$ is the expected value of the number of votes that a party v receives in the probability distribution $\mathcal{B}$ we assumed for the other voters.

In the following we will try to find out what the best, not necessarily economic, coercer is. First of all, talk is cheap: any communication prior to the tally does not depend on anything that influences the player's disarmed utility, therefore the coercer has no advantage by depending on this communication. The same holds for any prior punishment. Without loss of generality the coercer reacts to the tally by punishing with a certain probability or not. Again without loss of generality he always punishes as much as he can, since by *sanity* of U_P instead of gradually punishing he can induce the same damage by lowering the probability with which he punishes according to the amount of damage he wants to inflict. We will call this highest amount of punishment pundam.

The coercer receives the tally and decides to perform the punishment. What is the best response in the threat setting? If we define:

$$v_v = \Pr[M_P \text{ votes for v}]$$
$$p_v = \Pr[M_C \text{ punishes} \mid M_P \text{ votes for v}]$$

the pay-off in this setting is:

$$U_P^{thr}(M_P, M_C) = \sum_{v \in \mathcal{V} \cup \{\perp\}} v_v(u_v - p_v \cdot \mathrm{pundam}) + \sum_{v \in \mathcal{V} \cup \{\perp\}} u_v E[X_v]$$

The best response in the threat-setting maximizes this value. If there is a best response with some $v_v \notin \{0, 1\}$ for some v then there is another with $v_V = 1$ for some vote V, too, as the probability mass can be shifted to the vote v_V that maximizes the term

$$v_V(u_V - p_V \cdot \mathrm{pundam}).$$

So without loss of generality assume $v_V = 1$.

A δ-U_P-effective M_C for $\delta = 0$ makes the player choose some V over V_P in his best response, hence:

$$u_V - p_V \cdot \mathrm{pundam} > u_{V_P} - p_{V_P} \cdot \mathrm{pundam}$$
$$u_{V_P} - u_V < \mathrm{pundam}(p_{V_P} - p_V)$$

In order to find tight bounds for the objective-value in relation to the punishment damage, we try to find the vote V such that the right hand side is maximized. Fix V and V_P. In the following we will show which machine maximizes $p_V - p_{V_P}$, i. e., find a non-deterministic poly-time Turing machine P that maximizes:

$$\Pr[P(r) = 1 \mid r \leftarrow \mathcal{B}_{V_P}] - \Pr[P(r) = 1 \mid r \leftarrow \mathcal{B}_V]$$

where $\mathcal{B}_v = \mathsf{tally}(\mathbf{V}, v); \mathbf{V} \leftarrow \mathcal{B}$ and P outputs 1 if it advises punishment. Put in another way we gain (since we only deal with discrete probabilities here):

$$\begin{aligned}&\max_{P \in \mathrm{PPTM}} \Pr[P(r) = 1 | r \leftarrow \mathcal{B}_{V_P}] - \Pr[P(r) = 1 | r \leftarrow \mathcal{B}_V] \\ &\quad = \max_{p:\mathcal{T} \to [0,1]} \sum_{r \in \mathcal{T}: p(r) \neq 0} p(r) \cdot (\mathcal{B}_{V_P}(r) - \mathcal{B}_V(r))\end{aligned}$$

which is obviously maximized by

$$p(r) := \begin{cases} 1 & \text{if } B_{V_P}(r) > B_V(r) \\ 0 & \text{otherwise.} \end{cases}$$

This is the information theoretically best way to distinguish both machines, and luckily in this case it is computable. For $r = (x_{\mathrm{SPD}}, \cdots, x_\perp)$, $\sum_{v \in \mathcal{V} \cup \{\perp\}} x_v = n$ and $x_V, x_{V_P} > 0$ we have:

$$\begin{aligned} B_{V_P}(r) =& \mathcal{B}(x_{\mathrm{SPD}}, \cdots, (x_{V_P} - 1), \cdots, x_\perp) \\ & \frac{(n-1)!}{x_{\mathrm{SPD}}! \cdots (x_{V_P} - 1)! \cdots x_\perp!} \\ & \cdot p_{\mathrm{SPD}}^{x_{\mathrm{SPD}}} \cdots p_{V_P}^{x_{V_P} - 1} \cdots p_\perp^{x_\perp} \\ B_V(r) =& \frac{(n-1)!}{x_{\mathrm{SPD}}! \cdots (x_V - 1)! \cdots x_\perp!} \\ & \cdot p_{\mathrm{SPD}}^{x_{\mathrm{SPD}}} \cdots p_V^{x_V - 1} \cdots p_\perp^{x_\perp} \end{aligned}$$

and therefore, if $x_{V_P} = 0$:

$$B_{V_P}(r) - B_V(r) = - B_V(r)$$

If $x_V = 0$:

$$B_{V_P}(r) - B_V(r) = B_{V_P}(r)$$

and otherwise:

$$\begin{aligned} B_{V_P}(r) - B_V(r) =& \frac{(n-1)!}{x_{\mathrm{SPD}}! \cdots (x_{V_P} - 1)!(x_V - 1)! \cdots x_\perp!} \\ & \cdot p_{\mathrm{SPD}}^{x_{\mathrm{SPD}}} \cdots p_{V_P}^{x_{V_P} - 1} p_V^{x_V - 1} \cdots p_\perp^{x_\perp} \\ & \cdot \left(\frac{p_V}{x_V} - \frac{p_{V_P}}{x_{V_P}}\right) \end{aligned}$$

The sign of the difference in the last equation does only depend on the last term, therefore p is computable by a machine that outputs one if

$$\frac{x_{V_P}}{x_V} > \frac{p_{V_P}}{p_V} \text{ or } x_V = 0$$

Intuitively, this means that the coercer punishes the player if the party that the player prefers gets more votes than the distribution indicates. Given this machine, we would like to compute $p_{V_P} - p_P$. Let $p(x_{V_P}, x_V)$ be an abbreviation of $p(r)$ for $r \in \mathcal{T}$ with x_{V_P} votes for V_{V_P} and x_V for V:

$$\begin{aligned}
&\Pr[p(r) = 1 \mid r \leftarrow \mathcal{B}_{V_P}] - \Pr[p(r) = 1 \mid r \leftarrow \mathcal{B}_V] \\
&\quad = \Pr[p(x_{V_P} + 1, x_V) = 1 \mid x_{V_P}, x_V \leftarrow \mathcal{B}] \\
&\qquad - \Pr[p(x_{V_P}, x_V + 1) = 1 \mid x_{V_P}, x_V \leftarrow \mathcal{B}] \\
&\quad = \sum_{x_{V_P}, x_V} \Pr[x_{V_P}, x_V \leftarrow \mathcal{B}] \cdot (\Pr[p(x_{V_P} + 1, x_V) = 1] \\
&\qquad\qquad - \Pr[p(x_{V_P}, x_V + 1) = 1]) \\
&\quad = \sum_{x_{V_P}, x_V} \Pr[x_{V_P}, x_V \leftarrow \mathcal{B}] \cdot \Delta_{x_{V_P}, x_V}
\end{aligned}$$

with

$$\begin{aligned}
\Delta_{x_{V_P}, x_V} &:= p(x_{V_P} + 1, x_V) - p(x_{V_P}, x_V + 1) \\
&= \begin{cases}
1 & \text{if } \frac{x_{V_P}+1}{x_V} > \frac{p_{V_P}}{p_V} \wedge \frac{x_{V_P}}{x_V+1} < \frac{p_{V_P}}{p_V} \text{ or } x_V = 0 \\
-1 & \text{if } \frac{x_{V_P}+1}{x_V} < \frac{p_{V_P}}{p_V} \wedge \frac{x_{V_P}}{x_V+1} > \frac{p_{V_P}}{p_V} \\
0 & \text{otherwise}
\end{cases}
\end{aligned}$$

In order to compute this efficiently, we define the sets $S^+_{X_V}$ for all X_{V_P} such that $\Delta_{X_{V_P}, X_V}$ is positive, and $S^-_{X_V}$ such that $\Delta_{X_{V_P}, X_V}$ is negative.

$$\begin{aligned}
S^+_{X_V} &:= \{x_{V_P} \in [\frac{p_{V_P}}{p_V} x_V; \frac{p_{V_P}}{p_V}(x_V + 1)]\} \text{ for } x_V > 0 \\
S^-_{X_V} &:= \{x_{V_P} \in [\frac{p_{V_P}}{p_V}(x_V + 1)]; \frac{p_{V_P}}{p_V} x_V]\} = \emptyset
\end{aligned}$$

We refer to Appendix **??** for the Maple code we used to compute the following results. The choices of V_P and V maximizing

$$\Pr[p(r) = 1 \mid r \leftarrow \mathcal{B}_{V_P}] - \Pr[p(r) = 1 \mid r \leftarrow \mathcal{B}_V]$$

are MLPD and NPD. This goes along with our intuition that the difference that a single vote makes is more observable for parties with smaller percentage. It is surprisingly high: 0.0397, roughly four percent. Hence the ideal coercer M_C is U_P-ineffective iff

$$\begin{aligned}
u_{V_P} - u_V &> \text{pundam}(p_{V_P} - p_V) \\
u_{V_P} - u_V &> 0.0397 \cdot \text{pundam} \\
\frac{u_{V_P} - u_V}{\text{pundam}} &> 3.97\%
\end{aligned}$$

How do we interpret this result? Of course we cannot say that the difference between in worth of two votes has to be more than a 25th of a punishment, because we cannot compare the value of punishment to the value of a vote. But we can say that a person would have to value his vote (more precisely, the difference between the two votes) more than not receiving punishment in 1 out of 25 cases. If we define objective as the difference between the maximal valuation of a vote, and the second highest valuation, $\text{objective} := \min_{v \neq v', v \in \mathcal{V} \cup \{\perp\}}(v' - v)$ for $v' = \max_{v \in \mathcal{V} \cup \{\perp\}} a_v$ than we can state incoercibility exactly for the class $\mathcal{U}_P$ of player utilities where

$$\frac{\text{objective}}{\text{pundam}} > 3.97\%.$$

We could guarantee a higher amount of incoercibility for more specific player utilities. $p_\perp$ - p_{CDU} for example is just around 0.2%. But obviously a reasonable statement about incoercibility cannot restrict a voter in his choice. Note that we do not put any restriction on the economy of the coercer at all.

The election districts range from around 150'000 voters to 250'000 voters, so for other districts, the success of punishment might be even higher. It is left to the reader to decide if we can hope that every eligible voter in Germany values the difference his vote makes enough that he would put up with a 4% higher probability of being punished. This value could be lowered, and thus voters be made less coercible, by making the election district larger or changing the way the tally is published (although this is unlikely to happen). We think that a thorough investigation on this topic is worthwhile. A more sophisticated reasoning about the distribution of the other votes has to be made. The player's valuation of the tally needs further discussion, too. Still, we hope that this chapter provides an adequate illustration of the work-flow of such an analysis and sparks the interest in further analysis.

10 DISCUSSION

> BLUFF, N.1. *an attempt to trick somebody by making them believe that you will do something when you really have no intention of doing it, or that you know something when you do not, in fact, know it[..]*
>
> — Weiner et al. [1993], *The Oxford Dictionary*

Now that we come closer towards the end of this work, we will discuss topics regarding the incoercibility notions introduced on the last pages that need to be discussed and eventually be tackled.

10.1 KNOWING THE ADVERSARY

In all notions we introduced, we evaluate the effectiveness or economy of a given coercing strategy using properties of the player's best response to them. The best response to a coercing strategy "knows" the strategy,i. e.the machine that is run. If this machine asked the player for "its name" and punishes if not getting "Rumpelstielzchen" as a reply, it would have the action of replying "Rumpelstielzchen" in its best-response set – even though the coercer did not tell the player its name. Hard-wired secrets are implicitly known to the player. We do not care about how he finds his best response, we just assume him to use it. How do we justify this? First and foremost, the coercer has to tell the player what to do. The strategy itself is a description of the threat he puts upon the player. One can criticise that we implicitly assume that the player and the coercer can communicate with each other, and that the player knows:

- That it is the coercer he is talking to.
- What precisely the conditions of punishment are.

Here we fail to model, e. g., the situation in which a protocol participant is incoercible because there are no means to talk to him. We justify this by the fact that it helps the coercer; the more a player knows about him, the better for the coercer. We claim (without proof) that any situation where the player is forced to make a U_P-wise worse decision out of uncertainty about the coercer's acting, say he just has some probability distribution over what the adversary would do, can be modelled by another coercer who randomizes his acts using just the same probability distribution (assuming it is sampleable). For the example above, the "Rumpelstielzchen"–adversary would sample from a set of funny names. At the point at which we start talking about incoercibility, the assumption of a coercer being able to communicate with certain participants is not too far-fetched.

10.2 EMPTY THREATS

If someone threatened to blow himself up, would you give in to his threat? If you would not give in to him, it would be fairly irrational for him to take any action, as there is nothing to gain anymore and the loss is immense. The question boils down to the question of whether the coercer is committed to his course of action or whether this is an *empty threat*. When avoiding such extreme cases, one can find an argument in the reputation someone has to build up. Consider some gangster that takes an actual risk when torturing a person; he might be persecuted for his acts of violence (those are his costs of punishment). Still, from time to time, if a person does not obey him, he will have to take a risk in order to make his threat liable.

The opposite example is a coercer that does not have any punishment costs. He might punish someone giving in to his threat despite all promises he made.

In the model we employed it is not only the case that the coercing strategy is precisely known to the player, but also that the coercer committed to this strategy. This makes his threats more powerful in general, but maybe of disadvantage to him: when he is in a situation where he has to punish in order to keep his reputation he is forced to do so, while in this very moment punishing is expensive. Still, in long-term it makes sense, because sooner or later a player would start calling this bluff.

Returning to the example of the coercer threatening to blow himself up, we see that this case is not justified with the need to build up a reputation, at least for a single person. It would be possible to add some restriction for the worst-case loss a rational coercer would allow himself to take. However, since this only weakens our notion of incoercibility we abandon a formalisation of this property.

10.3 EXTRACTABLE VOTES ASSUMPTION

The proof we give in Section 8.1 uses an assumption in it that we need to discuss: It requires that the dummy-adversary-simulator $\tilde{A}_S$ that is used to achieve UC/c with respect to dummy-adversary/deceiver can be split into one simulator per communicating machine without introducing new channels to the protocol or sharing a state; while preserving indistinguishability to the original $\tilde{A}_S$. This assumption is quite restrictive, as you can see on the example of the implementation of a voting scheme that uses a common reference string which is distributed among the participants. Now assume that there are no means for the participant to share information. A split adversary is likely to have to make this common reference string up. No harm done so far, but multiple $\tilde{A}_S$ would have to make them up on their own such that there is a correspondence between the common reference strings and things that depend upon it. Having now means of communication there is way for them to synchronize. This is used in the way we perform the proof. It is basically needed to allow

for extracting the vote a participant took from the communication in the protocol. Maybe it is possible to weaken this assumption to some kind of extractability requirement,i. e., the existence of a program that is capable of extracting the vote and is indistinguishable to a network with any incarnation of the dummy-adversary-simulator. A more elegant solution would be, of course, to alter the proof, so we do not need such an assumption at all.

11 CONCLUSION & FURTHER WORK

Time is money.

— Benjamin Franklin, *Advice to a Young Tradesman*

This thesis served the goal of establishing a general notion of incoercibility that is well-founded and intuitive. The notion has been introduced, discussed and verified with a number of examples. We focused on the special case of voting schemes in order to establish the implication from UC/c to best-possible incoercibility. This allows for using UC/c in proofs of best-possible incoercibility, providing a model that is easier to use and more familiar to cryptographers. However, it is also possible to investigate incoercibility completely within our framework, as Chapter 9 showed in the example of an election district in Germany.

The idea of modelling coercion and related processes using game theoretic notions does, of course, allow for further elaboration in many directions.

11.1 COSTLY COMPUTATION

The assumption behind most cryptographic protocols work is that the coercer's computational resources are limited. If we would not restrict the coercer to poly-time Turing machines, most protocols would not be secure. This assumption is sound: computational power costs money as well as computation time does. The computational notions we use in the present work employ computational assumptions as well, but in a relatively simplistic form, namely by restricting the sets of strategies $\mathcal{M}_C$ and $\mathcal{M}_P$. We could do better, though: since our model aims at modelling the considerations of a coercer regarding his pay-off, it should incorporate computation time, too. This makes a more precise modelling possible. Assume that the player can force the coercer to do a lot of computation, but only in certain cases. In this case, it is a matter of their utilities and the probability in some concrete outcome whether a coercer is economic or not. Speaking of poly-time Turing machines does not allow for a precise description of the scenario and conceals the mechanisms in the game-theoretic reasoning. The situation in which the coercer forces the player by having him run for so long that he would rather take another choice is at least thinkable. It should be a consequence of such a modelling, however, that neither the coercer nor the player choose strategies that have super-polynomial runtime.

Halpern and Pass [2007] propose a model of interaction for agents that provides costly computation, i. e., a model to reason about the players' considerations about their computational resources. They introduce a

complexity function for each player $\mathcal{C}_P$ that evaluates on the set of strategies that the players chose and ranks them using some natural number. Such a function could, e. g.,

- output 1 if in this run P's strategy would take longer than some polynomial $p(k)$
- output 1 unless P is faster than some other player
- output the number of additions needed to perform the computation

A utility function would take the output of the complexity functions and somehow incorporate it into the pay-off. For the first case one could have the utility function to be $-\infty$ for $\mathcal{C} = 1$, thus reaching something similar to the approach of restricting the players' strategies to poly-time ITMs. There are some dissimilarities between the model in Halpern and Pass [2007] and ours, so we adapted the concept to fit our model:

Definition 28 (Network Machine Game with Costly Computation) A Network Machine Game has the following components:

- N is a finite set of players.
- $\mathcal{M} = \mathcal{M}_1 \times \cdots \times \mathcal{M}_{|N|}$ is a set of machines. The machines in $\mathcal{M}_i$ are interactive deterministic Turing machines that have a random tape $\{0,1\}^\infty$ and are otherwise ITMs in the sense above.
- We define the outcome $o(\mathbf{M}, \mathbf{R})$ to be the trace of the network communication of the Network initialized with the randomness $\mathbf{R}$. This function is used to model the network interaction and is more or less the implementation of a network protocol with $\mathbf{M}$ inserted for the relevant entities. We call the set of all possible outcomes Z.
- For each player $i \in N$ we define a complexity function

$$\mathcal{C}_i : \mathcal{M} \to \mathbb{N}$$

- For each player $i \in N$ we define the utility function

$$U_i : \Omega(Z) \times \mathbb{N}^{|N|} \to \mathbb{R}$$

where $\Omega(Z)$ denotes the set of probability distributions over Z. □

If the utility functions do not use the values given by the complexity functions, this definition is equal to Definition 1.

Another open question to pursue further is whether there is an equivalence result for absolute computational incoercibility against arbitrary coercers and absolute computational incoercibility against economic coercers for some specific $\mathcal{C}_C$ and U_C that punish every step of computation but do not reward any objective more than polynomially in the security parameter.

11.2 MORE COERCERS, MORE PLAYERS, MORE CORRUPTION

Right now we model a coercion game as a game with two players; the coercer and the player to be coerced. What happens if we insert a number of players? Interesting situations can occur: consider a coercer with a certain goal and many players with various goals. Some might differ more from the coercer's goals, some less. It might be a matter of the players negotiating with each other to find a solution that has a high pay-off for all of them, and avoids punishment.

By introducing multiple coercers we could be capable of modelling real coalitions, i. e., temporary alliances of agents with different but similar goals. Here, too, negotiations between agents are of great importance: maybe both profit from a certain situation, but just one of them is willing to take the punishment costs. Maybe both have different goals and try to manipulate the same player.

The problem that we see here is that such negotiation issues distract from the security aspects of a model. Instead of a number of coercers, a single coercing force, i. e., one player capable of controlling several machines might be more powerful in modelling "the common purpose", which is, by the way, the traditional model of coalitions employed in cryptography. – One assumes that coercers are more powerful when working together.

The setting that we have chosen allows for an environment to corrupt parties within the protocol run. A corrupted party is under the control of the adversary, which is the player the coercer controls. Depending on the situation, we could loosen the definition of a coercibility game to not be defined using a protocol π and a filtering function f. The filtering can be done using an environment as well, and this would additionally allow to give the coercer control over a larger number of parties and furthermore give means to take control over machines in the protocol run. The environment might output how many and which machines have been corrupted by the end of the protocol run, making it possible to make them costly, too. This way we are able to model situations where the coercer can bribe certain authorities and make deception more difficult (maybe even more expensive).

11.3 IMPLICATION FROM UC/C AND OTHER NOTIONS

The implication from UC/c to best-possible incoercibility against irrational adversaries is crucial for the evaluation of voting schemes according to our model. Although evaluating the ideal-model is useful for analysing the election process itself, gaining absolute results for actual implementation is what we are really interested in. Now, the proof that UC/c implies best-possible incoercibility against arbitrary coercers only works under the assumption that we can split $\tilde{A}_S$, the dummy-adversary-simulator in the UC/c definition, into two machines that are not needed to communicate with each other. This excludes, e. g., schemes that use a

common reference string shared between all parties in the protocol (split adversaries would need to make this string up, more precisely, make *the same string up*). More discussion can be found in Section 10.3.

The proof itself does actually not depend on the ideal functionality we use. It might be the case that we need the restriction to voting schemes in order to get rid of the split $\tilde{A}_S$ assumption, but for the proof as it is right now we do not really need the restriction. It is worthwhile to try to proof the implication with respect to arbitrary protocols and weaken the assumption.

Furthermore UC/c is not the only notion for incoercibility that is not restricted to voting schemes. Kuesters and Truderung [2009] propose a definition of coercion resistance and provide results for existing voting protocols. A similar implication for this notion would allow us to transfer those results.

11.4 THOROUGH ANALYSIS OF REAL-LIFE ELECTIONS

The analysis of the election district Saarbrücken in Chapter 9 aimed at illustrating the work-flow for an analysis of the maximal coercion resistance of the parliamentary election in Germany. We think that a thorough investigation on this topic is necessary. In fact, we are surprised that we were not able to find an analysis on this topic. In order to strengthen the result, a more sophisticated reasoning about the distribution of votes in relation to the information about the public opinion in forehand has to be made. The voters valuation of an outcome is to be discussed in order to analyse a model that is better justified than the one we sketched. We simplified the voting system, not taking care of, e. g., the election threshold, or the outcome above the district level. The situation for first votes might be different. Above all, Saarbrücken is not the only district in Germany, and there are elections other than the parliamentary election.

Finally, the economy of a coercer should be regarded. The ideal coercing strategy in our example punishes in little less than half of the cases, even when the player obeys. It is left to determine precisely how expensive this coercer is, and if there is another coercer that is U_P-effective but economic for a larger class of coercer utilities.

APPENDIX

BIBLIOGRAPHY

Michael Backes, Catalin Hritcu, and Matteo Maffei. Automated verification of remote electronic voting protocols in the applied pi-calculus. In *CSF*, pages 195–209. IEEE Computer Society, 2008. ISBN 978-0-7695-3182-3.

R. Canetti. Universally composable security: A new paradigm for cryptographic protocols. In *focs*, page 136. Published by the IEEE Computer Society, 2001.

R. Canetti and R. Gennaro. Incoercible multiparty computation. In *focs*, page 504. Published by the IEEE Computer Society, 1996.

Stéphanie Delaune, Steve Kremer, and Mark Ryan. Coercion-resistance and receipt-freeness in electronic voting. In *CSFW*, pages 28–42. IEEE Computer Society, 2006. ISBN 0-7695-2615-2.

Roderich (Bundeswahlleiter) Egeler et al. Wahlkreisergebnis bundesland saarland wahlkreis 296 - saarbrücken, endgültiges ergebnis der bundestagswahl 2009. published online. URL http://www.bundeswahlleiter.de/de/bundestagswahlen/BTW_BUND_09/ergebnisse/wahlkreisergebnisse/l10/wk296/.

Joseph Y. Halpern and Rafael Pass. Algorithmic rationality: Game theory with costly computation, 2007.

T. Honore. A Theory of Coercion. *Oxford Journal of Legal Studies*, 10(1): 94, 1990.

Jonathan Katz. Bridging game theory and cryptography: Recent results and future directions. In Ran Canetti, editor, *TCC*, volume 4948 of *Lecture Notes in Computer Science*, pages 251–272. Springer, 2008. ISBN 978-3-540-78523-1. URL http://dx.doi.org/10.1007/978-3-540-78524-8_15.

G. Kol and M. Naor. Cryptography and game theory: Designing protocols for exchanging information. *Theory of Cryptography*, pages 320–339, 2008.

Ralf Kuesters and Tomasz Truderung. An epistemic approach to coercion-resistance for electronic voting protocols, May 2009. URL http://arxiv.org/abs/0903.0802.

T. Moran and M. Naor. Receipt-free universally-verifiable voting with everlasting privacy. *Advances in Cryptology-CRYPTO 2006*, pages 373–392, 2006.

R. Nozick, S. Morgenbesser, P. Suppes, and M. White. Philosophy, Science, and Method, 1969.

Martin J. Osborne and Ariel Rubenstein. *A Course in Game Theory*. The MIT Press, Cambridge, Massachusetts, 1994.

Zuzana Rjaskova. Electronic voting schemes, 2002.

Dominique Unruh and Jörn Müller-Quade. Universally composable incoercibility. In *Crypto 2010*, LNCS. Springer, August 2010. To appear, preprint on IACR ePrint 2009/520.

J. Von Neumann and O. Morgenstern. Theory of games and economic behavior. Princeton. *NJ: Princeton University*, 1947.

E. Weiner, J. Simpson, and M. Proffitt. *Oxford English Dictionary*. Clarendon, 1993.

A. Wertheimer. *Coercion*. NJ, 1987.