Tariq Elahi, John A. Doucette, Hadi Hosseini, Steven J. Murdoch, and Ian Goldberg

A Framework for the Game-theoretic Analysis of Censorship Resistance

Abstract: We present a game-theoretic analysis of optimal solutions for interactions between censors and censorship resistance systems (CRSs) by focusing on the data channel used by the CRS to smuggle clients' data past the censors. This analysis leverages the inherent errors (false positives and negatives) made by the censor when trying to classify traffic as either legitimate or as CRS traffic, as well as the underlying rate of CRS traffic. We identify Nash equilibrium solutions for several simple censorship scenarios and then extend those findings to more complex scenarios where we find that the deployment of a censorship apparatus does not qualitatively change the equilibrium solutions, but rather only affects the amount of traffic a CRS can support before being blocked. By leveraging these findings, we describe a general framework for exploring and identifying optimal strategies for the censorship circumventor, in order to maximize the amount of CRS traffic not blocked by the censor. We use this framework to analyze several censor types in scenarios with multiple data-channel protocols used as cover for the CRS. We show that it is possible to gain insights through this framework even without perfect knowledge of the censor's (secret) values for the parameters in their utility function.

1 Introduction

Internet censorship resistance is a relatively recent field, yet it has been gaining prominence in recent times due to the increased censorship activity by various regimes around the world. This activity has given rise to an influx of interest, funding, and research effort in producing circumvention solutions to stymie those censorship efforts. Most of the research and engineering effort has been focused on understanding the technological aspects of 1) the myriad censorship techniques and attacks and 2) the equally many censorship resistance systems that circumvent them. However, there is a striking lack of research effort and insight into the behavior of the censor and circumventor and their interaction since, so far, the literature has treated that aspect of Internet censorship as a black box [12]. In this present work we investigate this aspect of censorship through the lens of game-theoretic analysis because it is an apt tool for modeling the interaction between two non-cooperative self-interested entities. Since the attack space is large, we focus on analyzing the *data channel*—the communication between the client and a destination outside of the censor's jurisdiction. This data channel is used by a censorship resistance system (CRS); the CRS typically disguises this data channel so that the client appears to be using some innocuous protocol to speak to some legitimate server, but in reality, the CRS is connecting the client to an Internet server of her choice. Our work is timely because there is currently a lot of activity within the community to develop better designs and implementations that address censorship threats to the data channel [9, 10, 17, 18, 21, 29, 30]. Specifically, we seek to understand how the success (or failure) of the censorship apparatus, measured by its error rates (*i.e.* false positives and negatives), affect the censor's behavior and if, and how, the circumvention traffic rate of a CRS (which affects the censor's error rates) can be used as a parameter in CRS designs.

Our contributions are:

1. A game-theoretic analysis leading to the identification and description of Nash equilibria of linear utility functions that allow a non-zero rate of CRS traffic to flow in the one-shot and repeated game scenarios
2. The conclusion that fielding a censorship apparatus does not change the equilibrium solutions above, but only the threshold circumvention traffic rate
3. A simulation-based analysis, leveraging the above findings, of censors with non-linear (risky) utility functions and many target data channel protocols to identify best responses (equilibria) for various censor types
4. A framework for exploring and identifying data channel protocols that provide useful circumvention traffic rates for a given censor type and use case.

Tariq Elahi: University of Waterloo, E-mail: tariq.elahi@uwaterloo.ca
John A. Doucette: University of Waterloo, E-mail: j3doucet@cs.uwaterloo.ca
Hadi Hosseini: University of Waterloo, E-mail: hadi.hosseini@uwaterloo.ca
Steven J. Murdoch: University College London, E-mail: s.murdoch@ucl.ac.uk
Ian Goldberg: University of Waterloo, E-mail: iang@cs.uwaterloo.ca

2 Background

Game theory is the study of how groups of rational, self-interested entities behave in response to one another's actions. In the context of censorship-resistant communications, a game-theoretic approach can be used to assess the optimal behavior of a rational censor and the designers of a CRS.

To facilitate this, we will analyze the behavior of the two parties, or *players* from now onwards, in increasingly detailed versions of an abstract "censorship game", designed to capture the fundamentals of censorship resistance dynamics, while still being simple enough to readily analyze. This serves to reveal the essential components of the problem domain.

These players try to maximize their benefits by thinking strategically about their actions, using information that they have about the environment and the other players. A central assumption is the theory of "rational choice", which states that an entity seeks to maximize its utility independent of the other player's utility and will chose an action that is at least as good as any other action available to them. The utilities can be modeled by a utility function (U) that assigns cardinal utilities to ordinal values. That is, if a player prefers outcome a over outcome b and outcome b over outcome c then the utilities are ordered $U(a) > U(b) > U(c)$.

2.1 Technological Limits

The censor and its apparatus have limitations such as the computational and memory costs of real-time processing, amongst other considerations. It is important, then, to take into account the rate at which objects of interest are misclassified. The two types of errors—false positives and false negatives—govern the confidence the censor has in their censorship apparatus. The prevalence of each of these type of errors provides an important input for both the censor and the circumventor in defining their respective strategies.

2.1.1 False Positives

From the censor's perspective, false positives are the legitimate traffic, and users, that were misclassified and blocked—the *collateral damage*. The censor naturally seeks to keep this as low as possible.

As noted by Elahi *et al.* [12], the collateral damage strategy has been leveraged by numerous censorship resistance systems, most recently by meek [14] and CloudTransport [5], both of which leverage popular cloud hosting services. These services are considered too important for the censor to block for risk of incurring economic losses to local businesses that utilize them for their operations. However, in most cases the circumventor assumes an all-or-nothing approach to censorship, which can be limiting when the censor is content with partial blocking. [20]

2.1.2 False Negatives

The censor tries to prevent as many clients, or as much traffic, as it can from circumventing its blocks—termed *information leakage*. Due to the limits of technology it is unable to identify all of them.

The circumventor's aim is always to have as much, if not all, of its traffic classified as a (false) negative. Strategies that obfuscate telltale features of CRS traffic to make them indistinguishable from non-CRS traffic, as well as steganographic and encryption techniques, are all instrumental in achieving this goal.

We note here that since the circumventor is a rational player its aim is not to produce collateral damage, or indeed to explicitly reduce the censor's utility. It is only concerned with maximizing its own utility, independent of the censor's utility.

3 Censorship Games

In our model, a censorship game is a game played between two players. One player, called the *censor*, has comprehensive control over the network of a target area (its *sphere of influence*, or *SoI*), and wishes to prevent certain undesirable communications from being transmitted over that network, while maximizing the throughput of legitimate traffic.[1] The other player, called the *circumventor*, wishes to send censored traffic (*e.g.* political speech that the censor disapproves of) over the censor-controlled channel, and may or may not care about the level of throughput for other "legitimate" communications on the censor-controlled network.

The circumventor is able to disguise circumvention, or covert, traffic to match a certain profile of legitimate cover traffic, and exercises control over the amount of traffic that is sent by altering the *circumvention traffic rate* (CTR) of the censorship resistance system (CRS) they have deployed. The circumvention traffic rate can be set to any value in the range

[1] This is a simplification since the censor may also care about other aspects that contribute to its utility, such as international perception, political fallout, and citizen unhappiness to name a few.

$0 \leq CTR \leq CTR_{\max}$, where $CTR_{\max}$ is the maximum amount of traffic that the CRS could transmit if it was fully utilized.

The censor possesses the ability to shut off all traffic (both legitimate and circumvention). The censor may also, but not always, possess the ability to differentiate the circumventor's traffic from the legitimate traffic that it is disguised as, by means of some censorship *apparatus*, usually in the form of a firewall or deep packet inspection (DPI) system capable of differentiating suspicious traffic based on the expected fingerprints of circumvention traffic. This ability to differentiate is prone to errors classified as false positives or false negatives.

Each player has a separate *utility function* that maps from the choice of action taken by both players to the total reward acquired by one of them.

The game is played in a series of discrete rounds, happening in sequential discrete timesteps. At the start of each round, both players simultaneously select an *action*, from their action set, on the basis of the actions selected by the two players in all previous rounds of the game, and on the basis of the players' utility functions and calculations.

In a censorship game, a *strategy* for the circumventor is a specification of how the circumvention traffic rate parameter will be set at different timesteps in the game, and a strategy for the censor is specification at different timesteps in the game of whether the the channel will be left open (allowing all traffic through) or not, and whether or not the apparatus will be used, if it is available. In this setting, we do not model either circumventor or censor expending resources to develop better CRSs or apparatus. For example, a strategy for the censor might be to leave the channel open if the circumvention traffic rate of the circumventor was below a certain level in all previous time steps, and to close it permanently otherwise. An example strategy for the circumventor might be to send no traffic at all for some time, and then send a very large burst of traffic. A *strategy profile* is a specification of a strategy for each player.

A *Nash equilibrium* is a strategy profile where neither player could improve their utility by unilaterally adopting a different strategy. This is a stable point of the game, which we might expect to observe frequently in reality. We will characterize the behaviors of the two agents in terms of the Nash equilibria of the game.

We also assume in our analysis throughout section 4 and section 5 that both the censor and circumventor have perfect information about each other. That is, both players know what the other *has* done (but not necessarily what they will do next), and knows the exact utility function and parameters being used by the other player. This is a common assumption in studying equilibria in repeated games [24]. We believe this assumption is plausible because the utility functions involved are not overly complex, and the both parties can observe the past actions of their opponents (or similar entities) to arrive at an accurate estimate of the parameters involved. Naturally, any predictions made by our model with inaccurate estimates of the needed parameters will tend to produce inaccurate predictions about the locations of inflection points in the players' behaviors, but the general trends will still be correct. Moreover, in section 6 we discuss a framework based on censor equivalence classes where we do not need to know the exact parameter values.

4 A Simple Censor Model

We begin by considering the simplest version of the game where the censorship resistance system uses only one channel, carrying only one type of traffic; for example, the CRS could be disguising is circumvention traffic as Skype traffic [19, 21, 23]—in this case, the "channel" would consist of all Skype traffic crossing the censor's SoI boundary. We assume that, absent the traffic of the circumventor, this channel carries a total amount of legitimate traffic L. We normalize both CTR and L by setting $L = 1 - CTR$.

We now proceed with closed-form analysis of the game in three steps, gradually increasing the complexity of the model.

4.1 Step 1: Single Round, No Apparatus

In this version of the game, the two players play just one round of the game, and the censor has no access to an apparatus that would allow it to differentiate between the traffic of the circumventor and the traffic of legitimate users.

The action space of the censor, denoted X_{cen}, consists of two strategies: 1 and 0 (the channel being On and Off). Playing "On" means the censor allows all traffic on the channel to pass through unimpeded, while "Off" means all traffic transmission is halted.

The action space of the circumventor is a real number $CTR \in [0, 1]$, which is the amount of circumvention traffic the circumventor chooses to send (as a fraction of the total traffic).

The utility functions of the censor and circumventor are respectively given by:

$$\begin{aligned} U_{\text{cen}} = (-\alpha_{\text{act}} X_{\text{cen}} + \alpha_{\text{bct}}(1 - X_{\text{cen}}))CTR + \\ (\beta_{\text{alt}} X_{\text{cen}} - \beta_{\text{blt}}(1 - X_{\text{cen}}))(1 - CTR) \end{aligned} \tag{1}$$

$$\begin{aligned} U_{\text{cir}} = (\gamma_{\text{act}} X_{\text{cen}} - \gamma_{\text{bct}}(1 - X_{\text{cen}}))CTR + \\ (\delta_{\text{alt}} X_{\text{cen}} - \delta_{\text{blt}}(1 - X_{\text{cen}}))(1 - CTR) \end{aligned} \tag{2}$$

Variables $\alpha_{[\text{act,bct}]}, \beta_{[\text{alt,blt}]}, \gamma_{[\text{act,bct}]}$, and $\delta_{[\text{alt,blt}]}$ are parameters that depend on the specific players of the game. The subscripts *act* and *bct* stand for *allow* and *block circumvention traffic*, respectively. The subscripts alt and blt stand for *allow* and *block legitimate traffic*, respectively. The α_{act} and α_{bct} are the loss, or gain, of utility to the censor of allowing, or blocking, one unit of circumvention traffic, respectively. Similarly, β_{alt} and β_{blt} are the gain, or loss, in utility to the censor of having one unit of legitimate traffic transported via, or blocked on, the channel, respectively. The ratios of α_{act} to β_{alt} and of α_{bct} to β_{blt} characterize different types of censors. For example, an employer interested in reducing employee idleness by preventing communication with social media sites, but ensuring that productive online activities are not affected, might have a relatively low α_{act}, but a relatively high β_{alt}. In contrast, a military agency trying to censor leakage of state secrets might have a very high α_{bct} relative to their β_{blt} parameter. The circumventor's counterpart parameters γ_{act} and γ_{bct} show the utility gained, or lost, by the circumventor of a single unit of circumvention traffic to be transported, or blocked, respectively. Similarly, δ_{alt} and δ_{blt} show the utility gained, or lost, by the circumventor of a single unit of legitimate traffic to be transported, or blocked, respectively. All of these parameters can be normalized to the range $[0, 1]$, where 0 means ambivalence and 1 means strong sensitivity.

Conventionally both δ parameters are assumed to be zero since typically CRS designers are not concerned with the fallout of CRS usage nor are there any technical provisions to reduce the impact of the fallout on non-CRS traffic in the designs in the literature. Also, γ_{bct} is also assumed to be zero since typically CRS designs are ambivalent to blocked CRS traffic; that is, what matters *directly* to the circumventor is the amount of circumvention traffic allowed through the censor's firewalls, not the amount that is blocked . Thus the circumventor's utility function is reduced to the following:

$$U_{\text{cir}} = \gamma_{\text{act}} X_{\text{cen}} CTR \tag{3}$$

4.1.1 Analysis

It is apparent that the censor maximizes its utility by playing "On" if $\beta_{\text{alt}}(1 - CTR) - \alpha_{\text{act}} CTR > \alpha_{\text{bct}} CTR - \beta_{\text{blt}}(1 - CTR)$, and "Off" otherwise.[2] Consequently, the Censor leaves the channel open if it believes the circumventor will play $CTR \le \frac{\beta_{\text{alt}}+\beta_{\text{blt}}}{\alpha_{\text{act}}+\alpha_{\text{bct}}+\beta_{\text{alt}}+\beta_{\text{blt}}}$; or $CTR \le F$ for brevity, where $F = \frac{\beta_{\text{alt}}+\beta_{\text{blt}}}{\alpha_{\text{act}}+\alpha_{\text{bct}}+\beta_{\text{alt}}+\beta_{\text{blt}}}$.

If the players know each others' strategies, the utility of the circumventor is maximized by setting $CTR = F$. However, although this is a Pareto Optimal solution, it is actually *not* a Nash equilibrium of the game. This is because the censor and circumventor decide their actions simultaneously, and so do not know each others' actions in advance. Given that the censor plays "On", the circumventor's best response is actually to pick $CTR = CTR_{\max}$, since this maximizes the utility of the circumventor. Consequently, the profile where the censor plays "On" and the circumventor plays $CTR = F$ is not a Nash equilibrium.

To find the Nash equilibrium, we note that if the censor plays "Off", the circumventor is equally happy to play $CTR = CTR_{\max}$ instead of any other value of CTR (since all settings of CTR yield zero utility). This means the circumventor should play $CTR = CTR_{\max}$ regardless of what the censor does, simplifying the game considerably. Knowing that the circumventor's utility is maximized by playing $CTR_{\max}$ regardless, the censor would choose to play "On" if and only if $CTR_{\max} < F$. In a game where this holds true, the only Nash equilibrium is for the censor to leave the channel open, and the circumventor to play $CTR_{\max}$. Otherwise, the only Nash equilibrium is for the censor to close the channel and for the circumventor to play $CTR_{\max}$.

Thus, we can see that, in this simplified game, the Nash equilibrium depends on both the maximum amount of traffic the circumventor can send, and on the tradeoff between the costs and benefits to the censor of allowing and blocking circumvention traffic versus keeping legitimate traffic flowing.

However, in practice, we rarely observe the equilibrium where censors elect to close their channels entirely. Next, we show that a circumventor interested in maintaining communications over a longer, uncertain time horizon, will behave differently, leading to a different equilibrium from the one present in the single round game.

4.2 Step 2: Multiple Rounds, No Apparatus

As in the Prisoner's Dilemma, the Nash equilibrium in the simple censorship game described above arises from not modeling the temporal dynamics of the game. Intuitively, if both censor and circumventor know that exactly one round of the game will be played, there is no reason for the circumventor to hold back: they will always send the largest possible amount of traffic, and if the censor doesn't block, the circumventor gets as much reward as possible. If the censor does block, then the circumventor would not get any reward regardless of what they

[2] Note that the analysis is invariant under affine transformations of the players' utility functions.

played. In the face of such an opponent, the censor of course must block (contingent on F and $CTR_{\max}$), to avoid the unacceptable volume of illegitimate traffic that would be sent.

The key result for cooperation in temporal games, due to Aumann [4], is that the equilibrium play that follows if the players *know* when the game will end is often identical to that in a single-shot game. This is because, in the last round of the game, the players are simply playing the static game again (there is no temporal component, because the game will now end, just like in Step 1 above). Once the players know how the final round will be played, then they can also infer how the penultimate round should be played using exactly the same logic, by treating the game as ending one round earlier than before, with full knowledge of the outcomes in the final round that will follow. Inductively, the players will play the first round in the same fashion as they would the last, if the game requires coordination or trust. Since the censorship game we study can be modeled with such a trust-based element (if the censor opens the channel, they "trust" the circumventor not to defect and send too much traffic), it is straightforward to show that the equilibrium of a temporal version of the game with a fixed number of rounds will be exactly the same as the equilibrium in the single shot game. However, when the game is played for an infinite or indefinite number of rounds, then this need not be so.

Suppose that after each round of the game, another round is played with probability p, and otherwise the players stop. This can model scenarios where the CRS or communication technology has become deprecated, or because the conditions of censorship have changed. A strategy in the context of this extensive-form game (*i.e.* the game of playing many rounds of the censorship game described in Step 1) consists of specifying a policy for how a player plays, in light of everything their opponent has done in the past.

We analyze this game using the same utility function from Step 1, since it is still applicable, but summed across all rounds of play. Again we assume that the δ and γ_{bct} parameters are zero due to typical CRS designs not being concerned with the fallout of CRS activity and discount the blocked CRS traffic.

4.2.1 Analysis

An interesting Nash equilibrium now emerges (though not necessarily a unique one). The censor adopts a policy to play "On" as long as the circumventor has never played $CTR > Z$ at any point in the past, and to play "Off" if even one prior iteration of the game involved the circumventor sending more traffic than that,[3] where Z is a fixed quantity such that $F \geq Z$. The circumventor adopts a policy of playing $CTR \leq Z$ at every step.

To show that the censor leaving the channel open and the circumventor playing $CTR = Z$ is a Nash equilibrium, we use proof by induction.[4]

In the first round, the circumventor could deviate and send up to $CTR = CTR_{\max}$ traffic. However, doing so would result in a total utility of $\gamma_{\text{act}} CTR_{\max}$ for this turn, and zero utility thereafter. In contrast, using $CTR = Z$ this turn, and defecting next turn instead, would result in an expected total utility of $\gamma_{\text{act}}(Z + pCTR_{\max})$. Provided that $CTR_{\max} < Z + pCTR_{\max}$, it is thus better to wait another turn before sending more traffic than Z. It follows that deviation for the circumventor will always be better in "one more turn", if $CTR_{\max} < \frac{Z}{1-p}$. This in turn provides the censor with a clear policy for setting $Z = (1-p)CTR_{\max}$, as this is the lowest possible value that will placate the circumventor. The equilibrium exists only if $F \geq (1-p)CTR_{\max}$.

The equilibrium just outlined depends on the assumption that the censor turns off the channel and never turns it back on. However, this may not be a credible threat since the censor wants the channel open in the long run, so as to allow the legitimate traffic to get through. Furthermore, if the circumventor drops the circumvention traffic rate, *i.e.* $CTR < Z$, after sending $CTR_{\max}$, the censor cannot plausibly commit to keeping the channel closed forever in response since it is now better to open the channel, as we have shown earlier. To resolve this shortcoming of the original model, the censor instead commits to blocking the channel for a period of *finite* length, τ. This blocking period can also be thought of as the *punishment* the censor metes out to the circumventor for defecting. To find τ, the censor simply repeats the same analysis as for the infinite punishment period, but with a slight modification, which we detail next.

In the first round, the circumventor could again deviate and send up to $CTR = CTR_{\max}$ traffic. After this, the censor would close the channel for τ rounds, resulting in a total utility for the circumventor of $\gamma_{\text{act}} CTR_{\max}$ for $\tau + 1$ rounds. In contrast, leaving the channel open for that period would provide a total utility of $U_{cir} = \gamma_{\text{act}} \sum_{i=0}^{\tau} p^i Z = \gamma_{\text{act}} Z \frac{1-p^{\tau}}{1-p}$ to the circumventor. After the period of $\tau + 1$ rounds has passed, the game will be in the same state as at the start (i.e. the censor will open the channel, and the circumventor will set their circumvention traffic rate to whatever value will maximize profits). An equilibrium where the censor keeps the channel open, and

[3] We revisit and expand on this strong assumption later in this subsection.

[4] This proof assumes that $CTR \ll 1$, which is supported in practice for deployed CRSs by empirical evidence provided by Elahi *et al.* [11].

the circumventor sends $Z = CTR = \frac{1-p}{1-p^{\tau}} CTR_{\max}$ traffic then follows by similar logic to the equilibrium with an infinite punishment period, provided that $F > Z = \frac{1-p}{1-p^{\tau}} CTR_{\max}$. Note that, as τ is increased, the value Z that the censor can use will decrease, but with rapidly diminishing returns. The precise value selected by the censor will thus depend on how credible the censor's threats are. A censor that can credibly claim that it will close the channel for longer periods will be able to squeeze the circumvention traffic rate lower than one that cannot credibly make such threats.

Both equilibria discussed so far are plausible in the real world. Censors might indeed decide to permanently shut a channel over which too much undesired traffic has been seen to flow even once. Certainty it is plausible that they might choose to temporarily close it for some prolonged, but finite, period in response. Yet, although these strategies are part of valid Nash equilibria, there are not the strategies that rational actors, as opposed to the real world actors, would adopt when playing the repeated game since they are not *subgame perfect* equilibria. A subgame perfect equilbrium requires that any contiguous sequence of moves made in the game also form a Nash equilibrium, but this is clearly not the behavior we observe when the censor engages in punishment. If the circumventor stops sending $CTR_{\max}$ during the punishment period τ, and instead reverts to some other quantity that is smaller than Z, a rational censor could (locally) improve its utility by ceasing punishment and reopening the channel. In short, there is little incentive for the censor to actually follow through its threat of long-term punishment when closing the channel hurts the censor in the long run.

4.2.2 Perfect Nash Folk Theorem

Although real-world actors may indeed follow through with such seemingly irrational threats (perhaps because of external factors, like a need to maintain prestige in other, simultaneously played, games) the "perfect" Nash Folk Theorem of Fudenberg and Maskin [16] provides a more complex equilibrium that is of similar form, yet is also subgame perfect. In this equilibrium, we define four distinct "state" points, and each player's strategy will consist of movements from one state to another, in response to the actions of their opponent in previous rounds, rather like a state machine. These are presented in tabular form in Table 1. The players both begin in state $\bar{*}$. In any state except $\underline{*}$, if both players have played their prescribed strategies (as per Table 1), then the players remain in that state (i.e. they will continue playing the prescribed strategies in the next round). If a player deviates from their prescribed strategy, then both players move to the state $\underline{*}$ for τ rounds. Thereafter, both players move to the state r_{cen} if the censor deviated last, or r_{cir} if the circumventor did. If deviation occurs during a round in the punishment period, then the punishment counter resets and another τ rounds will be spent in $\underline{*}$, and the state transitioned to afterwards will correspond to the player that deviated most recently, regardless of which player caused the initial movement to the state $\underline{*}$. We require that $\sigma \geq 1$ and that $\epsilon > 0$, as well as the constraints that $F \geq Z$ and $\frac{\sigma(Z-\epsilon)}{(\sigma+1)(1-p)} > \frac{CTR_{\max}}{1-p^{\tau}}$ (i.e. that both players prefer the long-term payoffs of $\bar{*}$ to either r state, and prefer the longterm payoff of either r state to the payoff of repeatedly deviating and being punished in $\underline{*}$). The *Perfect Nash Folk Theorem* ensures that, if both players initially commit to this strategy, then the strategies together form a subgame perfect Nash equilibrium, provided that p is sufficiently high (for example, values as low as $p = 0.75$, $\tau = 2$, and $\sigma = 9$ allow for $Z = \frac{CTR_{\max}}{1.575}$). The intuition behind this is that if either player deviates from the initial point, then it will receive strictly less utility in the future, because there is no way to return to $\bar{*}$. During the punishment period, threats are made credible by the promise that, if they are not fulfilled, the end game will move to an equilibrium that is less beneficial to the player who forgoes punishing when it ought. This result implies that a cooperative equilibrium, similar to the two shown above relying on credible threats, can be supported even when both players are locally maximizing utility within the game, and are perfectly rational.

Table 1. The four phases of the subgame perfect equilibrium strategies for the repeated version of the censorship game.The variables σ and ϵ are parameters, and are set as explained in the text.

Name	$X_{\mathbf{cen}}$	CTR
$\bar{*}$	**1 (open)**	Z
$r_{\mathbf{cen}}$	**1 (open) for σ rounds, then 0 (closed) for one round**	Z **for σ rounds, then** $CTR_{\mathbf{max}}$ **for one round**
$r_{\mathbf{cir}}$	**1 (open) for σ round, then 0 (closed) for one round**	$Z - \epsilon$ **for σ rounds, then** $CTR_{\mathbf{max}}$ **for one round**
$\underline{*}$	**0 (closed)**	$CTR_{\mathbf{max}}$

To show that the proposed strategy profile is a subgame perfect Nash Equilibrium, we need only show that deviating in any one state of the game for a single step, and then returning to the equilibrium strategies, can only produce long-term harm for the player that deviates, and never long-term benefits.

To do this, we start by considering the phase $\bar{*}$. In this phase, deviation by the censor will yield zero utility for the round of deviation (since closing the channel means no traffic gets through at all), followed by τ rounds of zero utility punishments, followed by an endgame spent in r_{cen}, assuming neither player makes any other deviation from the equilibrium. Since the payoff for the censor is lower in r_{cen} than in $\bar{*}$, nothing can be gained by the censor's deviation. If the circumventor deviates, an initial gain of $CTR_{\max}$ is made, followed by τ rounds of zero utility, followed by an endgame spent in r_{cir}. Provided that $\frac{Z(1-p^{\tau+1})}{1-p} > CTR_{\max}$, the short-term deviation is not profitable, but even if it is, the endgame results in a decrease of $\frac{Z}{\sigma+1} + \epsilon$ of relative utility per round during the endgame, which will, for sufficiently large values of p, quickly come to dominate any short-term gains.

We have from the assumed constraints $F \geq Z$ and $\frac{\sigma(Z-\epsilon)}{(\sigma+1)(1-p)} > \frac{CTR_{\max}}{1-p^{\tau}}$, a straightforward guarantee that neither player wants to deviate from either r state. The censor gains nothing from deviating, since keeping the channel open is at least as profitable as closing it. The circumventor loses more utility via the following τ rounds of punishment than it can hope to gain in a single round of deviation.

This leaves only $\underline{*}$. If we enter this state to punish the circumventor, then the circumventor cannot deviate, because deviation merely resets the punishment counter for another τ rounds, and the endgame will still be the same. If the censor deviates, they suffer immediately, as the circumventor is playing $CTR_{\max}$ continuously in the punishment state. Further, if the censor deviates, the circumventor will attempt to move to r_{cir} after the punishment period has completed, an endgame that is strictly worse for the censor than if they had simply completed the punishment in the first place. Identical arguments can be used for the case where $\underline{*}$ is entered to punish the censor.

Although these interesting equilbria exist for some reasonable parameterizations of the game, there are other parameter settings for which poor equilibria are present instead. Notably, if $CTR_{\max} \gg F$, then the policy where the censor always blocks, and the circumventor always sends $CTR_{\max}$ may be the only plausible Nash equilibrium.

Interestingly, we note that p could be replaced by any discounting factor for the utility of future rewards. So if, instead of representing the chance of a future game, p represented the preference of each party for rewards today as opposed to in the future, a similar result could be derived. In practice, most companies do use such a discounting factor when considering the benefits of future rewards, since events in the future are fundamentally uncertain. To provide a censorship resistance example: a whistleblower may use a discounting factor where they are uncertain about their ability to communicate in the future and the value of the information they wish to transmit may be of such high impact that maintaining the channel for future use may be ignored.

We can conclude from this analysis that, in many reasonable games, it is the best policy for the circumventor interested in maintaining a long-term communication channel to keep $CTR \leq F$.

4.3 Step 3: Multiple Rounds, With an Apparatus

We now consider the case where the censor has some apparatus capable of distinguishing the target, covert, traffic (CTR') from the legitimate cover traffic (L). The apparatus correctly labels a fraction TPR (the true positive rate) of the circumvention traffic, but also incorrectly labels a fraction FPR (the false positive rate) of the legitimate traffic as circumvention traffic. Similarly, traffic not positively labeled can be partitioned to that which is truly not circumvention traffic, *i.e.* TNR (true negatives), and that which has been missed by the apparatus, *i.e.* FNR (false negatives). We note that $FNR = 1 - TPR$ and $TNR = 1 - FPR$. The output of the apparatus is traffic with the "Positive" tag or "Negative" tag, referring to if the apparatus deems the traffic as being CRS-related or not, respectively.

The new action space of the censor has two variables, denoted X_p and X_n, where both can take the values 0 and 1 (Block and Allow). X_p governs traffic tagged "Postive" and the censor can either block or allow this traffic. Similarly, X_n governs traffic tagged "Negative" and the censor can again either block or allow the traffic. The action space of the circumventor remains unchanged from before.

The presence of the apparatus serves to alter the utility functions of the censor and circumventor, U'_{cen} and U'_{cir} respectively, as follows:

$$\begin{aligned} U'_{\text{cen}} = & CTR'(-\alpha_{\text{act}}(TPR \cdot X_p + FNR \cdot X_n) + \\ & \alpha_{\text{bct}}(TPR(1 - X_p) + FNR(1 - X_n))) + \\ & (1 - CTR')(\beta_{\text{alt}}(FPR \cdot X_p + TNR \cdot X_n) - \\ & \beta_{\text{blt}}(FPR(1 - X_p) + TNR(1 - X_n))) \end{aligned} \tag{4}$$

$$U'_{\text{cir}} = CTR'(\gamma_{\text{act}}(TPR \cdot X_p + FNR \cdot X_n)) \tag{5}$$

The parameters are all normalized as before to the range $[0, 1]$.

To help build intuition, as an example let us consider the censor's sensitivity to blocking circumvention traffic (α_{bct}). Its contribution to the censor's utility function is $CTR' \cdot \alpha_{\text{bct}}(TPR(1 - X_p) + FNR(1 - X_n))$ because a fraction CTR' of the traffic *is* circumvention traffic, and of that, TPR of it is reported as positive, which will get blocked if $X_p = 0$, and $FNR = 1 - TPR$ of it is reported as negative, which will get blocked if $X_n = 0$. Similar reasoning follows for the other parameters.

4.3.1 Analysis

Ultimately the dynamics of this game are similar to those in Step 1 or 2 (depending on whether we incorporate temporal dynamics or not), with adjustments to the parameters of the censor. First, we analyze the censor's strategy space and make the following observations.

The censor has four strategies to play. Strategy $(X_p, X_n) = (1,1)$ is the same as not having an apparatus since the censor ignores the "Positive" tag on traffic and allows it through as well as allowing all the traffic with the "Negative" tag.

Strategy $(X_p, X_n) = (0,0)$ is again the same as not having an apparatus and is also the same as blocking all traffic since the censor disagrees with traffic tagged "Negative" and blocks it as well as blocking all the traffic tagged "Positive".

Strategy $(X_p, X_n) = (0,1)$ is where the censor goes along with the tagging of the apparatus and blocks traffic labeled "Positive" and allows traffic labeled "Negative".

Strategy $(X_p, X_n) = (1,0)$ implies that it is always better for the censor to disagree with the apparatus completely and do the opposite of what its tagging suggests. So now, traffic labeled "Positive" is allowed through while traffic labeled "Negative" is blocked. For the sake of simplicity, we assume that should the censor find that disagreement is more beneficial then it simply switches the tags which makes this strategy equivalent to strategy $(0,1)$ above. This is the same as assuming that $TPR \geq FPR$ and, equivalently, that $TNR \geq FNR$.

We now consider these strategies in more detail. Setting $(X_p, X_n) = (1,1)$ in Equation 4 gives the following:

$$U'_{cen^{(1,1)}} = CTR'(-\alpha_{\text{act}}) + (1 - CTR')(\beta_{\text{alt}}) \tag{6}$$

Similarly, the other settings yield the following utility equations:

$$U'_{cen^{(0,0)}} = CTR'(\alpha_{\text{bct}}) + (1 - CTR')(-\beta_{\text{blt}}) \tag{7}$$

$$\begin{aligned} U'_{cen^{(0,1)}} = &CTR'(-\alpha_{\text{act}} \cdot FNR + \alpha_{\text{bct}} \cdot TPR) + \\ &(1 - CTR')(\beta_{\text{alt}} \cdot TNR - \beta_{\text{blt}} \cdot FPR) \end{aligned} \tag{8}$$

To discover when it is better to play each strategy we compare each one against the other. Since the censor's utility depends on the circumvention traffic we state the results of this comparison in terms of CTR'.

For the censor to choose $(1,1)$ over $(0,0)$ then $U'_{cen^{(1,1)}} \geq U'_{cen^{(0,0)}}$ and the following must hold:

$$CTR' \leq \frac{\beta_{\text{alt}} + \beta_{\text{blt}}}{\alpha_{\text{act}} + \alpha_{\text{bct}} + \beta_{\text{alt}} + \beta_{\text{blt}}}, \tag{9}$$

or $CTR' \leq F_{ab}$, where $F_{ab} = \frac{\beta_{\text{alt}} + \beta_{\text{blt}}}{\alpha_{\text{act}} + \alpha_{\text{bct}} + \beta_{\text{alt}} + \beta_{\text{blt}}}$. The subscript ab denotes that when the inequality holds the censor gets more utility by allowing all traffic through than by blocking it. Note that $F \equiv F_{ab}$.

For the censor to choose $(1,1)$ over $(0,1)$ then $U'_{cen^{(1,1)}} \geq U'_{cen^{(0,1)}}$ and the following must also hold:

$$CTR' \leq \frac{FPR(\beta_{\text{alt}} + \beta_{\text{blt}})}{TPR(\alpha_{\text{act}} + \alpha_{\text{bct}}) + FPR(\beta_{\text{alt}} + \beta_{\text{blt}})}, \tag{10}$$

or $CTR' \leq F_{am}$, where $F_{am} = \frac{FPR(\beta_{\text{alt}} + \beta_{\text{blt}})}{TPR(\alpha_{\text{act}} + \alpha_{\text{bct}}) + FPR(\beta_{\text{alt}} + \beta_{\text{blt}})}$. Similar to the convention used above, the subscript am denotes that when the inequality holds the censor gets more utility by allowing all traffic than by using the apparatus (the m stands for machine, since the apparatus is a kind of machine).

For the censor to choose $(0,1)$ over $(0,0)$ means that $U'_{cen^{(0,1)}} > U'_{cen^{(0,0)}}$. Therefore the following must also hold:

$$CTR' \leq \frac{TNR(\beta_{\text{alt}} + \beta_{\text{blt}})}{FNR(\alpha_{\text{act}} + \alpha_{\text{bct}}) + TNR(\beta_{\text{alt}} + \beta_{\text{blt}})}, \tag{11}$$

or $CTR' \leq F_{mb}$, where $F_{mb} = \frac{TNR(\beta_{\text{alt}} + \beta_{\text{blt}})}{FNR(\alpha_{\text{act}} + \alpha_{\text{bct}}) + TNR(\beta_{\text{alt}} + \beta_{\text{blt}})}$. Again similar to before, the subscript mb denotes that when the inequality holds the censor gets more utility by using the apparatus than by blocking all traffic.

Each of F_{ab}, F_{am}, and F_{mb} is a threshold on CTR' that drives the censor's decision to allow, block, or use the apparatus. We would like to discover the ordering between the thresholds so that the censor can make informed (strategic) choices. We make an observation that simplifies the analysis: the terms $\alpha_{\text{act}} + \alpha_{\text{bct}}$ and $\beta_{\text{alt}} + \beta_{\text{blt}}$ are common and can be replaced with α and β, respectively. When determining the relative ordering of the three thresholds, we will assume, as above, that $TPR \geq FPR$ (and equivalently, that $TNR \geq FNR$).

We begin by noting that $F_{ab} \geq F_{am} \Leftrightarrow FPR \leq TPR$ since:

$$\begin{aligned} &F_{ab} \geq F_{am} \\ \Leftrightarrow\ &\frac{\beta}{\alpha + \beta} \geq \frac{FPR \cdot \beta}{TPR \cdot \alpha + FPR \cdot \beta} \\ \Leftrightarrow\ &\frac{\alpha + \beta}{\beta} \leq \frac{TPR \cdot \alpha + FPR \cdot \beta}{FPR \cdot \beta} \\ \Leftrightarrow\ &\frac{\alpha}{\beta} \leq \frac{TPR \cdot \alpha}{FPR \cdot \beta} \\ \Leftrightarrow\ &FPR \leq TPR \end{aligned} \tag{12}$$

Similarly, we also note that $F_{mb} \geq F_{ab} \Leftrightarrow FNR \leq TNR$ since:

$$\begin{aligned} &F_{mb} \geq F_{ab} \\ \Leftrightarrow\ &\frac{TNR \cdot \beta}{FNR \cdot \alpha + TNR \cdot \beta} \geq \frac{\beta}{\alpha + \beta} \\ \Leftrightarrow\ &\frac{FNR \cdot \alpha + TNR \cdot \beta}{TNR \cdot \beta} \leq \frac{\alpha + \beta}{\beta} \\ \Leftrightarrow\ &\frac{FNR \cdot \alpha}{TNR \cdot \beta} \leq \frac{\alpha}{\beta} \\ \Leftrightarrow\ &FNR \leq TNR \end{aligned} \tag{13}$$

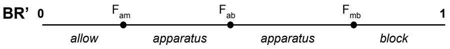

Fig. 1. Best censor strategies at critical circumvention traffic thresholds. The censor's strategies are in *italics*. The circumventor's strategies are to send a proportion of circumvention traffic, $0 \leq CTR' \leq 1$, with the critical thresholds marked as F_{am}, F_{ab}, and F_{mb}.

Since $F_{mb} \geq F_{ab}$ and $F_{ab} \geq F_{am}$, it is clear that the total ordering is $F_{mb} \geq F_{ab} \geq F_{am}$.

Given this ordering, the censor will play according to the following strategies, which are depicted in Figure 1. When $CTR' \leq F_{am}$ the censor will allow all traffic to flow. When $F_{am} \leq CTR' \leq F_{ab}$ or $F_{ab} \leq CTR' \leq F_{mb}$ then the censor will use the apparatus rather than allowing or blocking all the traffic, respectively. Finally, when $CTR' > F_{mb}$ the censor should block all traffic.

Turning to the circumventor we see that she actually only has only has two reasonable choices: sending $CTR' = F_{am}$ (in which case all of her circumvention traffic will get through), or $CTR' = F_{mb}$ (in which case only a fraction FNR of her circumvention traffic will get through). The decision rests on whether $FNR \cdot F_{mb} \geq F_{am}$; *i.e.*, when the inequality holds, the circumventor should send $CTR' = F_{mb}$ circumvention traffic, and otherwise she should send $CTR' = F_{am}$.

The key takeaway from the analysis in this section is that neither party has an incentive to deviate from the equilibrium points, as defined by the circumvention traffic thresholds F_{am}, F_{ab}, and F_{mb}. That is to say that as long as the circumventor does not send more than F_{mb} traffic, the censor will not block it, but will apply its apparatus to reduce the amount of circumvention traffic that gets through, or allow it entirely if it is below F_{am}.

It is clear then that the introduction of the apparatus, with its inherent TPR and FPR, does not produce a deviation from the character of the Nash equilibrium that we found in the simpler cases 1 and 2. The main effect is on the amount of traffic, CTR', the circumventor can send through while ensuring that the inequalities above remain true.

5 More Realistic Censor Models

So far we have analyzed censor utility functions that are linear in nature. In reality, the censor may be more risk averse. We mean by this that the censor's stakes (costs) to blocking CRS traffic, and not making mistakes, ramp up faster as rates of errors increase than the linear model above. One way to capture this is to utilize an exponential utility function for the censor.

The following is an example of an exponential censor utility function.

$$U''_{\text{cen}} = e^{-(C \cdot FPR \cdot (1-CTR) + D \cdot FNR \cdot CTR))} \tag{14}$$

$$U''_{\text{cir}} = E \cdot FNR \cdot CTR \tag{15}$$

Similar to the earlier α and β, the non-negative parameters C and D control the sensitivity of the censor to false positives and false negatives respectively. Like γ before, the non-negative parameter E controls the circumventor's sensitivity to circumvention traffic getting through the censor's SoI; without loss of generality, $E = 1$ for the remainder of this discussion. As before, the variable FNR is the percentage of the circumvention traffic allowed (*i.e.* the false negatives) and FPR is the percentage of legitimate traffic blocked (*i.e.* the false positives). This function allows a wide range of plausible censor utility functions to be modeled, and results in utility values between 0 (maximum dissatisfaction) and 1 (maximum satisfaction).

A second simplification we have thus far made was to only consider a single protocol that the CRS could blend in with. In reality there are a plethora of protocols that a CRS could use for cover, *e.g.* HTTP, TLS, and VoIP to name a few. Furthermore, it is likely that some protocols are more critical, or at least more important, than others and interfering with them would cost the censor more dearly.

Unfortunately, when we take these factors into consideration the preceding closed-form style of analysis becomes more complex and less straightforward to reason about. We change tracks here and leverage numerical simulation to help us analyze and gain further insights. We exploit our finding from the closed-form analysis above that a protocol remains unblocked as long as the circumventor does not transmit more than a certain amount of traffic over it. We create a simulation that utilizes Equation 14 and Equation 15 above and iterates over parameter values to help us find potential Nash equilibria for various types of censors.

The aim of the analysis that follows is to explore how to identify cover protocols that are good candidates as cover traffic for the amount of circumvention traffic that we wish to send. We focus on the quantity of the cover traffic a protocol provides rather than its other qualities such as its importance or the ease with which it can be imitated.

5.1 Strategy Simulator

Our simulator models the censorship game as follows. The circumventor moves first, and produces a CRS that impersonates one or more protocols and distributes circumvention traffic over these protocols according to some distribution. The censor can masquerade as a CRS client, and is able to establish which protocols are being impersonated and how much circumvention traffic is being sent over each. We examine the case where the impersonation is good—the censor does not have an apparatus that can distinguish legitimate uses of the protocol from uses of the protocol to carry circumvention traffic. Therefore, the censor must choose to either block a protocol entirely—blocking both cover traffic (causing false positives) and the circumventor's traffic (causing true positives), or leaving it entirely unblocked.

The censor and circumventor move simultaneously. In each round the censor will choose a blocking strategy, *i.e.* which protocols they will block, to maximize their utility. The goal of the circumventor is to find the right proportion of the total amount of circumvention traffic to send over each protocol such that the censor's best strategy is the one that maximizes the circumventor's utility. This will be the equilibrium strategy since if either party changes their choice, they will decrease their own utility.

An interesting consequence of this model is that the utility function of the circumventor does not matter, as all they can do is choose between the collection of scenarios which the censor has decided to be optimum for a particular strategy of the circumventor. Therefore, as long as the circumventor's utility function is monotonically increasing in terms of the false negative rate, the same equilibrium will be reached regardless of the function's shape.

The simulator models the relative importance of protocols, for both the censor and the population in the censor's SoI, by utilizing popularity of the protocol by traffic volume. As a concrete source of information we use traffic-volume data supplied by the 2014 survey of US Internet traffic [27]. Our simulator makes some simplifying assumptions to reduce the computational complexity of the simulation. We will provide more detail in subsubsection 5.1.2 where this becomes relevant.

5.1.1 False-Positive Intolerant Censor

We first consider a censor with a low tolerance to false positives. We define this to mean that there is at least one protocol that they are unwilling to block (a *critical* protocol), even if blocking the protocol would result in blocking all of the circumventor's traffic. In this case the circumventor should choose the critical protocol and send all censorship-resistance traffic over it. The censor will not block it, and so all of the circumvention traffic will get through. Any alternative strategy for the circumventor would be less good, as choosing multiple critical protocols would be more effort for no gain, and choosing a non-critical protocol for some traffic might lead the censor to block it.

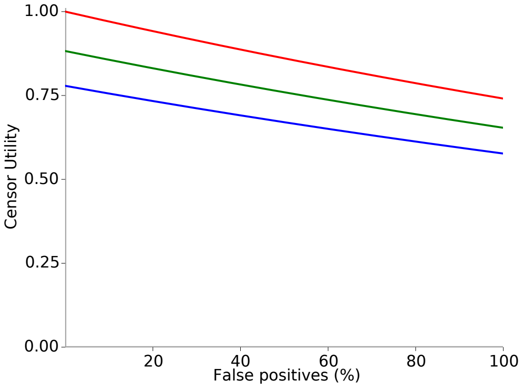

Fig. 2. Utility of a censor with high false-positive and false-negative tolerance.

5.1.2 False-Positive Tolerant Censor: Variant 1

A more interesting case is where there is no such critical protocol. To give a concrete example, assume that the circumventor can impersonate six protocols with the same relative quantities of traffic as the top six types of traffic from the survey: Netflix streaming video (33.81%), YouTube streaming video (14.63%), HTTP (6.08%), BitTorrent (4.85%), iTunes (3.12%) and Facebook (2.60%).[5] We shall call the most prevalent protocol the top protocol, and the least prevalent the bottom protocol, with the rest forming an ordering in between.

As the censor utility function, we use Equation 14 with $C = 0.3$ and $D = 0.25$. This is illustrated on Figure 2 for three values of true positive rates: 100% (top), 50% (middle) and 0% (bottom).

We now need to compute the censor utility function for all combinations of censor strategy and circumventor strategy. The censor can choose to block any selection of the six target cover protocols. As a result there are $2^6 = 64$ scenarios.

The circumventor can choose to send units of traffic in any distribution over the protocols, but we exclude any distribution where the traffic distributed over protocol a is greater than that distributed over protocol b when the quantity of cover traffic going over protocol b is greater than that of a. We do this because if any excluded scenario were chosen, if a and b were swapped, the censor utility function would be lower

[5] Note that these percentages do not add up to 100% since there will remain traffic types that are not targetted by the CRS and thus the situation where the censor needs to block all Internet access will not arise.

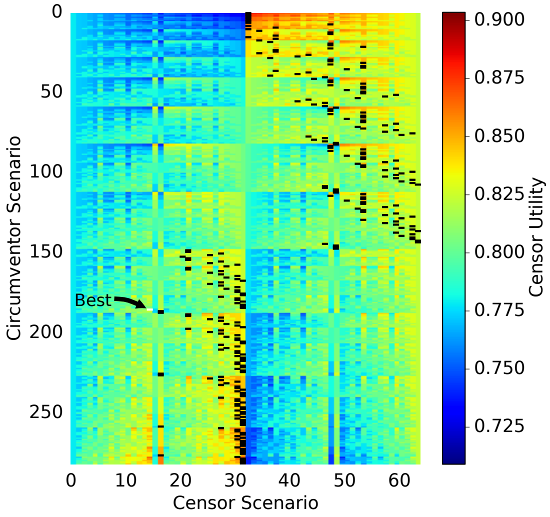

Fig. 3. Utility of a censor with high false-positive and false-negative tolerance.

for every censor scenario (assuming the censor prefers a lower false-positive rate).

Even making this assumption there are still an infinite number of circumventor scenarios if we allow any fractional value for the amount of traffic. So, to reduce the scenario space we quantize all circumvention traffic into multiples of 5 units up to a total of 100 units, resulting in 282 circumventor scenarios. We assign legitimate traffic to the six cover protocols in the same ratios as before, with 33.81, 14.63, 6.08 4.85, 3.12, and 2.6 units of traffic respectively, for a total of 65.09 units of total legitimate traffic. While in subsection 4.2 we assumed that $CTR \ll L$, we now want to also explore the range of situations where the censorship resistance traffic is similar in volume to the cover traffic as well as when it is greater in volume than the cover traffic.

The result of simulating all scenarios is shown in Figure 3, where blue is low utility and red is high utility. The censor scenarios are sorted in order of increasing false-positive rate. The circumventor scenarios at the top have traffic heavily skewed to the protocols with the most cover traffic; those at the bottom have traffic more evenly distributed over the protocols. The small rectangles show the optimum censor strategy for each circumventor strategy (white with the arrow labeled “Best” for the equilibrium and black for others).

Even small changes in the circumventor scenarios result in large changes in optimum censor scenario, but the equilibrium for this censor type is for the circumventor to distribute circumvention traffic quite evenly over the protocols, but not completely. The top protocol should get 40 units of traffic and the next four with 15 units of the traffic each with the sixth not used at all. The censor will block protocols 3, 4, and 5, allowing 55 units of circumvention traffic through. Were the attacker to block protocols 1 and 2, the additional false positives would not justify the extra 55 units of true positive (circumvention) traffic. Were the circumventor to move some traffic onto protocol 6, it would be blocked because it has a smaller false-positive cost.

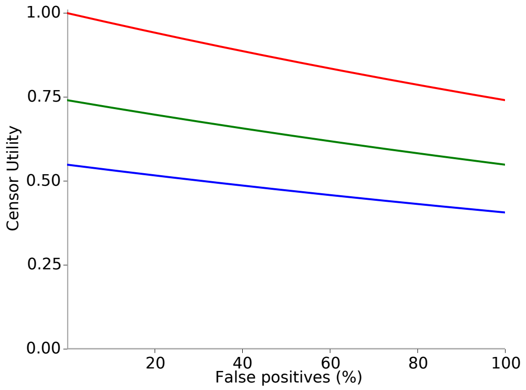

Fig. 4. Utility of a censor with high false-positive and low false-negative tolerance.

5.1.3 False-Positive Tolerant Censor: Variant 2

Let us now consider a censor who is equally tolerant to false positives, but far more sensitive to false negatives than before, by changing D from 0.25 to 0.6 with the result shown in Figure 4.

Now a 50% false negative rate shows significantly lower censor utility than variant 1 (the middle line). The resulting simulation is significantly different as well, as can be seen in Figure 5.

Now the optimum strategy for the censor is almost always to block many protocols, resulting in a high false-positive rate (the right-hand side of the graph). The equilibrium strategy is for 95 units of circumvention traffic to be distributed on protocol 1 and 5 units to be distributed on protocol 2. The censor will block protocol 1, but leave protocol 2 unblocked. This lets only 5 units of circumventor traffic through, but it is better than none, which almost every other strategy results in. For example, sending 100 units of traffic over protocol 1 results in it being blocked. Sending 80 units over protocol 1 and 20 units over protocol 2 results in both protocols being blocked. Putting only 5 units over protocol 2 is small enough that the extra benefit to the censor of blocking it is not large enough to justify the high false positives.

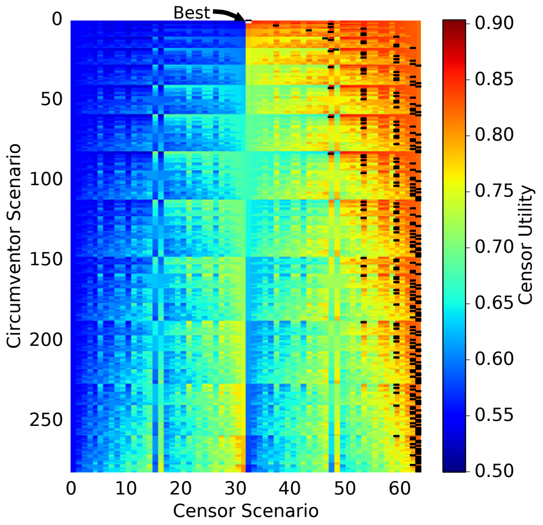

Fig. 5. Utility of a censor with high false-positive and low false-negative tolerance.

5.2 Parameter Analysis

The analysis above provides some insight into how different censor types behave and the optimum strategy for distributing traffic given the traffic volumes of potential cover protocols from real-world data. We now analyze what occurs when the number of protocols is varied as well as the amount of cover traffic they provide.

5.2.1 Protocol Popularity

The popularity, or amount of cover traffic available, of a protocol plays a significant role in the resulting Nash equilibrium and hence censor and circumventor strategies. We investigate this by taking a hypothetical protocol and varying its popularity, *i.e.* units of cover traffic, relative to all other non-cover traffic on the censor's network. Note that since there is only one protocol the circumventor can only play one action: send all 100 units of circumvention traffic over the protocol.[6] We use this setup to re-evaluate the fault-tolerant censors from above.

We see that for the censor with $C = 0.3$ and $D = 0.25$ the censor does not change their blocking pattern until the cover protocol gets to be a little more than 83 units of the total bandwidth. After this inflection point the censor switches to allowing the 100 units of circumvention traffic through since the collateral damage outweighs the benefit of information blocking. The takeaway is that, in this scenario, if we could only target one protocol it had better provide at least 83 units of cover traffic for each 100 units of circumvention traffic, or we would not be able to use it as cover to safely send all the circumvention traffic past the censor.

Table 2. Cover protocol bandwidth effects on utility, $C = 0.3$, $D = 0.25$

Bandwidth	$U_{\mathbf{cir}}$	$U_{\mathbf{cen}}$
10	0	0.97
50	0	0.86
83.5	1	0.78
90	1	0.78
99	1	0.78

Table 3. Cover protocol bandwidth effects on utility, $C = 0.3$, $D = 0.6$

Bandwidth	$U_{\mathbf{cir}}$	$U_{\mathbf{cen}}$
10	0	0.97
50	0	0.86
100	0	0.74
201	1	0.55
210	1	0.55

We can verify this simple case with one protocol using Equation 1 and rewriting it with α_{blt} and β_{act} replaced with C and D—the remaining parameters set to zero—to produce $CTR'' \leq \frac{C}{C+D}$ which yields $CTR'' \leq \frac{0.3}{03+0.25} \approx 0.545$. The censor will allow circumvention traffic to flow (100 units of it since there is only one channel) if it is 54.5% of the *total* traffic, *i.e.* the sum of the circumvention traffic and the cover traffic. This means that the cover protocol must be 45.5% of the total traffic, corresponding to the $\frac{83}{83+100}$ suggested by the simulation.

However, for the censor with more sensitivity to information leakage, *i.e.* $C = 0.3$ and $D = 0.6$, the inflection point occurs at a much larger 203 units of cover traffic, which closed-form analysis also confirms. This means that for this censor to allow 100 units of circumvention traffic a very popular protocol needs to be used as cover. Table 2 and Table 3 illustrate these trends.

While it seemed like it is better to target a protocol that is the majority of bandwidth on the network in general, the above examples show that there are censors for whom this approach can not be employed since their sensitivity to information leakage, D, is too high as compared to their sensitivity to collateral damage, C.

[6] We do not model the situation where the circumventor can hold back sending all the traffic they wish to send. We do this to simplify the analysis and also to illustrate the difference in the results where the cover protocol is not popular and where it is.

Table 4. The effect of cover traffic distributed over two protocols on utility, $C = 0.3$, $D = 0.25$

Bandwidths	U_{cir}	U_{cen}
82,1	0.95	0.79
72,11	0.85	0.78
62,21	0.70	0.79
52,31	0.60	0.78
42,41	0.50	0.78

Table 5. The effect of cover traffic distributed over two protocols on utility, $C = 0.3$, $D = 0.6$

Bandwidths	U_{cir}	U_{cen}
82,1	0	0.78
72,10	0.5	0.78
62,20	0.10	0.78
52,30	0.15	0.78
42,41	0.20	0.78

5.2.2 Dynamics of Cover Bandwidth over Two Protocols

The number of cover protocols can play a role in how the censor behaves. We investigate this by utilizing two hypothetical cover protocols where the sum of their cover traffic is kept constant. We then vary the amount of cover traffic units between the two to investigate the effects on the censor's best responses. We choose just below the inflection point from the analysis above as the total cover traffic units to distribute between the two protocols, *i.e.* 83 units of cover traffic. We do this to see if there is any difference in the censor's behavior.

We see from Table 4 that leveraging two cover protocols, where one is very small compared to the other, against the first censor ($C = 0.3$, $D = 0.25$) provides a high level of utility. However, as the cover traffic ratio between the two protocols decreases we see that the circumventor loses more utility which implies that in this scenario it is more beneficial to leverage a single cover protocol than multiple protocols.

Against the second censor ($C = 0.3$, $D = 0.6$), for whom we saw that only a very large amount of cover traffic (203 units) could cause it to deviate from its block-everything strategy, we see from Table 5 that now targeting two protocols instead (with a much smaller sum of 83 units of cover traffic distributed almost evenly over them) can cause the censor to allow up to 20 units of circumvention traffic to flow over them.

It is interesting that the two censors produced such different results; in one case targeting two protocols (with their sum equal to the noted inflection point) produced a reduced utility for the circumventor, while in another it allowed some portion of traffic to flow where none was allowed before even though the amount of cover traffic did not increase. It shows us that choice of not only which protocols (*i.e.* the amount of cover traffic they offer) but also the ratio of cover traffic between them can have an impact on censor behavior, such that it is beneficial to CRS activity. An avenue for future work is to further explore these aspects to understand and ultimately leverage the censor's sensitivity to particular protocols and their combinations.

6 Reducing Complexity

Our censorship games depended on perfect information and this makes it necessary to discover the correct type for the censor and the values of the parameters. This may be difficult, if not impossible, since the censor does not cooperate and hides this information and there are infinitely many value combinations these parameters may take. We now show that it is also irrelevant to learn this information about any given censor. This is because, instead of working with utility for specific parameter values, we gather up utility functions into equivalence classes of observed censor actions. Furthermore, we only consider equivalence classes, and hence parameter values, that directly impact the circumventor's utility function. This reduction in complexity in terms of equivalence class space makes the problem more tractable and enables us to find effective strategies for designing and deploying CRSs.

We do not completely, and accurately, attempt to map all parameters for all censors, CRSs and users, but the framework presented here can help in refining censor behavior models and be a jumping off point for future work.

6.1 Methodology

We first create a repository of censor *equivalence classes*. These are collections of censor actions, or action profiles, that characterize its behavior in the dimensions that the CRS is affected by. The profiles have a few conditions; they are distinct from one another and the actions in the profile need to be observable and maximize the censor's utility. In the setting we have presented, the action profile is the blocking pattern that the censor adopts. Each of the patterns is distinct and is easily observable, *e.g.* by probing which protocols are blocked and at what level of traffic.

We then consider past observed censor behavior and the conditions (or inputs) that cause it and map them to the equivalence classes. Where past observations are not available, an active probing test suite can collect the needed data. By this method we converge at 1) those equivalence classes that matter for the CRS, 2) the region the equivalence classes occupies

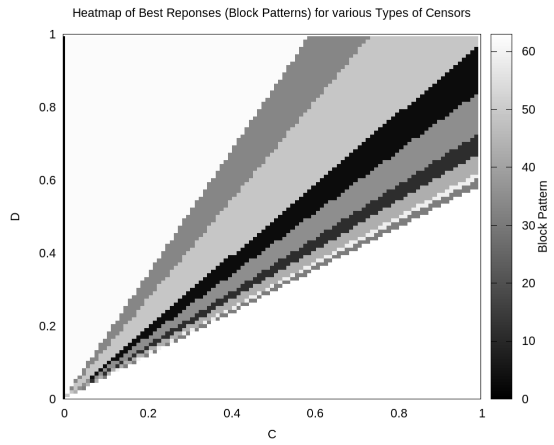

Fig. 6. Best censor responses (blocking patterns) for various censor types, *i.e.* values of C and D. Each shade represent one blocking pattern and all regions with the same shade represent a single censor equivalence class. The lighter shades denote blocking patterns where fewer protocols are blocked and darker shades denote patterns where more protocols are blocked.

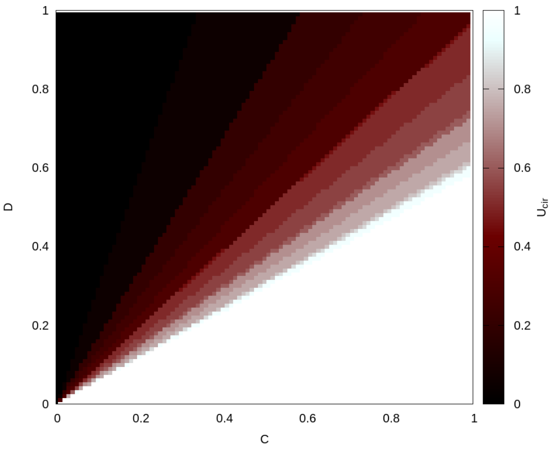

Fig. 7. Circumventor utility for best responses for various censor types, *i.e.* values of C and D. Each shade represents 0.05 units of circumventor utility. The lighter shades denote high utility and darker shades denote low utility, with black denoting zero utility (*i.e.* no circumventor traffic allowed through).

in the parameter space, and 3) the boundaries between classes that transition a censor from one profile to another.

In this manner we could predict a particular censor's behavior for given inputs and hence can design CRSs that allow us to maximize the circumventor's utility.

6.2 Censor Equivalence Classes

We apply this methodology to our censor utility function, in Equation 14. First, we enumerate the blocking patterns that we expect to appear due to U''_{cen} for the scenario presented in subsubsection 5.1.2 with the false-positive tolerant censor with six protocols. We compile a heatmap of best responses by censors of varying sensitivity values, C and D, in the range $[0, 1]$. The results are presented as the heatmap in Figure 6. These figures demonstrate that it is only the ratio C/D that actually matters; this makes sense, as multiplying both C and D by any positive constant just scales the utility function by raising it that constant power.

The interesting thing to note is that within this range out of the 64 possible blocking patterns the censor's best responses are limited to just 11, meaning that those are the patterns that the circumventor actually needs to address. From this map of the blocking patterns we can probe the censor's type to converge on the equivalence class of a particular censor by sending different proportions of circumvention traffic over the protocols and noting the behavior of the censor.

Since we are only interested in censor patterns that provide positive utility to the circumventor we also produce a circumventor utility heatmap (Figure 7) to compare with the blocking pattern heatmap. There are 15 contiguous regions with the same circumventor utility. These follow the same general trends of the blocking patterns but with some censor equivalence classes providing two different utilities for certain ranges of values. We note that there is a contiguous region (the black region in the top half) that provides no circumventor utility, and this corresponds with the pattern to block the top protocol—where the circumventor also sends all circumvention traffic over the top protocol—and the pattern to block all protocols. The bottom light shaded region provides the most utility (*i.e.* all circumvention traffic is allowed through) and this corresponds to the block-nothing pattern.

This framework enables us to discover the overall shape of the game. Given the traffic proportions of the cover protocols that the CRS can target we can use the methodology above to discover which censor strategies are likely to come into play and the potential circumventor utility we can achieve. This can allow the CRS designer to decide if it is worth playing the game and to help them target the right set of cover protocols that allow positive circumventor utility.

7 Related Work

Microeconomic approaches of incentive analysis and game-theoretical models have been adopted in numerous applications of network security for preventing attacks and designing adversarial intrusion detection models. In surveys [1, 22, 26] of the evolution of computer networks and security systems we see a drastic change from the use of heuristic and ad hoc solutions, to analytical paradigms that are based on rich game-theoretic models. This new shift has enabled researchers to account for players' incentives and attitudes towards decision making in various environments.

In the context of censorship resistance systems that are mainly inspired by peer-to-peer file/media sharing frameworks, researchers have focused on two orthogonal approaches: randomized file and functionality sharing where each node is assigned random resources, and a discretionary model where peers can choose and modify their precise contributions to the network [2, 3]. Danezis and Anderson [8] studied these two frameworks and showed that, in contrast to the initial intuition, the random model is less costly to attack for all possible attacker strategies, and that the cost to censor a set of nodes is maximized when resources are distributed according to node preferences. Contemporaneous to the work in this paper, Tschantz *et al.* [28] promote the idea that evaluating censorship resistance designs solely on technical attributes is shallow and at times intractable and present game-theoretic analysis as an alternative. The analysis and contributions are limited to considering abstract cost functions and preliminary conclusions about the viability of economic analysis as a means of evaluating CRS designs.

To the best of our knowledge, our work is the first to offer a framework for game-theoretic analysis of censorship resistance on the data channel in a variety of scenarios.

8 Future Work and Conclusion

There are several avenues of future work following from our analysis, some of which we outline here. First, the recently developed field of "security games", which uses techniques from game theory and optimization to defend against physical asset attackers, such as terrorists [25] or poachers [13] could be highly applicable, and could provide insight into the optimal allotment of a censor's resources toward developing better detection technologies. Second, it would be fruitful to explore how the behavior, or presence, of the CRS could affect if and how the censor allocates resources to improve the censorship apparatus (*i.e.* the cost/benefit analysis of improving the apparatus) and if there is a way to prevent an escalation of the conflict through the careful deployment and use of CRSs. Third, we would like to further develop the methodology we describe for identifying the censor's type, by rooting it to empirical data more closely. One shortcoming of collecting empirical data is that it is difficult to know if our network observations, and the effects on the data channel, are due to censorship or other reasons. The output of the various nascent efforts to identify censorship events in the wild [6, 7, 15, 31] can be a useful source of data for our framework.

In this paper, we focus attention on the censorship games wherein two rational and self-interested players, namely censor and circumventor, play their best strategic responses in a perfect information game. Considering a linear utility model, we start by analyzing the simplest pure Nash equilibrium analysis and enrich the model step by step. We then analyze the exponential utility setting and describe a simulated approach to equilibrium analysis.

Our simple closed-form analysis yields insight about the existence of Nash equilibria that can be leveraged by CRS designs. Extending our analysis to more realistic censorship scenarios, we leveraged simulation as an aid to discovering and analyzing equilibrium points. This approach has application to real-world CRS-design problems, namely, of how to select useful cover protocols and how to distribute circumvention traffic over them. Finally, we provide intuition about how one might go about discovering the censor's type using active probing as a method of indirect preference solicitation.

References

[1] T. Alpcan and T. Başar. *Network Security: A Decision and Game-Theoretic Approach*. Cambridge University Press, 2010.

[2] R. Anderson and T. Moore. The Economics of Information Security. *Science*, 314(5799):610–613, 2006.

[3] R. Anderson, T. Moore, S. Nagaraja, and A. Ozment. Incentives and Information Security. *Algorithmic Game Theory*, pages 633–649, 2007.

[4] R. J. Aumann. Acceptable Points in General Cooperative n-Person Games. *Contributions to the Theory of Games*, 4:287–324, 1959.

[5] C. Brubaker, A. Houmansadr, and V. Shmatikov. CloudTransport: Using Cloud Storage for Censorship-Resistant Networking. In *Proceedings of 14th Privacy Enhancing Technologies Symposium*. Springer, 2014.

[6] J. R. Crandall, D. Zinn, M. Byrd, E. T. Barr, and R. East. ConceptDoppler: A Weather Tracker for Internet Censorship. In *Proceedings of the 14th ACM SIGSAC Conference on Computer and Communications Security*, pages 352–365, 2007.

[7] G. Danezis. An anomaly-based censorship detection system for Tor. Technical Report 2011-09-001, The Tor Project, 2011. https://research.torproject.org/techreports/detector-2011-09-09.pdf.

[8] G. Danezis and R. Anderson. The Economics of Censorship Resistance. *Proceedings of the 3rd Annual Workship on Economics and Information Security*, 2004.

[9] R. Dingledine. Obfsproxy: The Next Step in the Censorship Arms Race. *Tor Blog*, https://blog.torproject.org/blog/obfsproxy-next-step-censorship-arms-race, February 2012. Retrieved May 2015.

[10] K. P. Dyer, S. E. Coull, T. Ristenpart, and T. Shrimpton. Protocol Misidentification Made Easy with Format-Transforming Encryption. In *Proceedings of the 20th ACM conference on Computer and Communications Security*, November 2013.

[11] T. Elahi, G. Danezis, and I. Goldberg. PrivEx: Private Collection of Traffic Statistics for Anonymous Communication Networks. Technical Report 2014-08, CACR, 2014. http://cacr.uwaterloo.ca/techreports/2014/cacr2014-08.pdf.

[12] T. Elahi, C. M. Swanson, and I. Goldberg. Slipping Past the Cordon—A Systematization of Internet Censorship Resistance. Technical Report 2015-10, CACR, 2015. http://cacr.uwaterloo.ca/techreports/2015/cacr2015-10.pdf.

[13] F. Fang, P. Stone, and M. Tambe. Defender strategies in domains involving frequent adversary interaction. In *Proceedings of the 2015 International Conference on Autonomous Agents and Multiagent Systems*, pages 1663–1664. International Foundation for Autonomous Agents and Multiagent Systems, 2015.

[14] D. Fifield, C. Lan, R. Hynes, P. Wegmann, and V. Paxson. Blocking-resistant Communication through Domain Fronting. *Proceedings on Privacy Enhancing Technologies*, 2015(2):46–64, June 2015.

[15] A. Filasto and J. Applebaum. OONI: Open Observatory of Network Interference. In *Proceedings of the USENIX Workshop on Free and Open Communications on the Internet*. USENIX, 2012.

[16] D. Fudenberg and E. Maskin. The folk theorem in repeated games with discounting or with incomplete information. *Econometrica*, 54(3):533–554, 1986.

[17] J. Geddes, M. Schuchard, and N. Hopper. Cover Your ACKs: Pitfalls of Covert Channel Censorship Circumvention. In *Proceedings of the 20th ACM conference on Computer and Communications Security*, 2013.

[18] B. Hahn, R. Nithyanand, P. Gill, and R. Johnson. Games Without Frontiers: Investigating Video Games as a Covert Channel. http://arxiv.org/pdf/1503.05904v2.pdf, 2015. Retrieved May 2015.

[19] A. Houmansadr, T. Riedl, N. Borisov, and A. Singer. IP over Voice-over-IP for Censorship Circumvention. *arXiv preprint arXiv:1207.2683*, 2012.

[20] A. Lewman. Iran Partially Blocks Encrypted Network Traffic. *Tor Blog*, https://blog.torproject.org/blog/iran-partially-blocks-encrypted-network-traffic, February 2012. Retrieved May 2015.

[21] S. Li, M. Schliep, and N. Hopper. Facet: Streaming over Videoconferencing for Censorship Circumvention. In *Proceedings of the Workshop on Privacy in the Electronic Society*, November 2014.

[22] M. H. Manshaei, Q. Zhu, T. Alpcan, T. Bacşar, and J.-P. Hubaux. Game Theory meets Network Security and Privacy. *ACM Computing Surveys*, 45(3):25, 2013.

[23] H. Mohajeri Moghaddam, B. Li, M. Derakhshani, and I. Goldberg. SkypeMorph: Protocol Obfuscation for Tor Bridges. In *Proceedings of the 19th ACM conference on Computer and Communications Security*, October 2012.

[24] M. J. Osborne. *An introduction to game theory*, volume 3. Oxford University Press New York, 2004.

[25] J. Pita, M. Jain, J. Marecki, F. Ordóñez, C. Portway, M. Tambe, C. Western, P. Paruchuri, and S. Kraus. Deployed ARMOR Protection: The Application of a Game Theoretic Model for Security at the Los Angeles International Airport. In *Proceedings of the 7th International Joint Conference on Autonomous Agents and Multiagent Systems: Industrial Track*, pages 125–132. International Foundation for Autonomous Agents and Multiagent Systems, 2008.

[26] S. Roy, C. Ellis, S. Shiva, D. Dasgupta, V. Shandilya, and Q. Wu. A Survey of Game Theory as Applied to Network Security. In *2010 43rd Hawaii International Conference on System Sciences*, pages 1–10. IEEE, 2010.

[27] Sandvine. Global Internet Phenomena Report. https://www.sandvine.com/downloads/general/global-internet-phenomena/2014/2h-2014-global-internet-phenomena-report.pdf, November 2014. Retrieved May 2015.

[28] M. C. Tschantz, S. Afroz, V. Paxson, and J. Tygar. On Modeling the Costs of Censorship. *arXiv preprint arXiv:1409.3211*, 2014.

[29] P. Vines and T. Kohno. Rook: Using Video Games as a Low-Bandwidth Censorship Resistant Communication Platform. http://homes.cs.washington.edu/~yoshi/papers/tech-report-rook.pdf, 2015. Retrieved May 2015.

[30] Q. Wang, X. Gong, G. T. K. Nguyen, A. Houmansadr, and N. Borisov. CensorSpoofer: Asymmetric Communication using IP Spoofing for Censorship-Resistant Web Browsing. In *Proceedings of the 19th ACM conference on Computer and Communications Security*, October 2012.

[31] J. Wright, A. Darer, and O. Farnan. Detecting Internet Filtering from Geographic Time Series. http://arxiv.org/pdf/1507.05819v1.pdf, July 2015. Retrieved August 2015.