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 non-circumvention traffic 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 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.

DOI 10.1515/popets-2016-0030
Received 2016-02-29; revised 2016-06-02; accepted 2016-06-02.

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 [21]. In this 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 unblocked 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 [11, 13, 18, 19, 24, 36, 38]. 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 proportion 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 proportion 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 framework for exploring and identifying data channel protocols that provide useful circumvention traffic rates for a given censor type and use case;
4. The insight that throttling the CRS client's network usage is an optimal solution but opens up the CRS to a particular censor attack;

***Corresponding Author: Tariq Elahi:** KU Leuven, E-mail: tariq@kuleuven.be
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

5. An alternative mechanism to throttling that provides robustness to the censor attack above; and
6. The insight that cover protocol ratios can play a significant role in the resulting equilibria.

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 analysis 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 players' utilities [23, 33] 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 $U(a) > U(b) > U(c)$. Each player has an *action space*, which is the set of actions the player can take, and each player adopts a *strategy* describing which actions they will play under what conditions. A game consists of a set of action spaces for each player, and a function mapping specifications of a strategy for each player to outcomes.

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 non-circumvention 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 Khattak *et al.* [21], the collateral damage strategy has been leveraged by numerous censorship resistance systems, most recently by meek [15] and CloudTransport [7], 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. [22]

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 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.

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. While collateral damage may be incidentally produced by this maximization, the damage is not taken into account by the circumventor when making decisions. The setting we focus on is one where the circumventor is interested in maximizing the utility solely derived from the use of the CRS.

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 non-circumvention 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 non-circumvention communications on the censor-controlled network.

The circumventor is able to disguise circumvention, or covert, traffic to match a certain profile of non-circumvention cover traffic, and exercises control over the amount of traffic that is sent by altering the *circumvention traffic proportion* (CTP) of the censorship resistance system (CRS) they have deployed. The CTP is a fraction of the total traffic, which includes both circumvention and non-circumvention traffic. The circumvention traffic proportion can be set to any value in the range $0 \leq CTP \leq CTP_{\max} < 1$, where $CTP_{\max}$ is the assumed maximum amount of traffic, as a fraction of the total traffic, that the CRS could transmit if it were fully utilized.

The censor possesses the ability to shut off all traffic (both non-circumvention and circumvention). The censor may also, but not always, possess the ability to differentiate the circumventor's traffic from the cover 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 proportion parameter will be set at different timesteps in the game, and a strategy for the censor is a specification at different timesteps in the game of whether 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 as the game progresses. For example, a strategy for the censor might be to leave the channel open if the circumvention traffic proportion 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. 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 utility function parameters being used by the other player. This is a common assumption in studying equilibria in repeated games [28]. 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.

Additionally, we assume that both the censor and the circumventor have knowledge about the amount of circumvention and non-circumvention traffic that is successfully transported. While this information should be easy for a circumventor to obtain, the censor may not know how much circumvention traffic is getting through. We note that, ultimately, a real-world censor may have many out-of-band methods for inferring this parameter. For example, a nation-state censor may arrest political dissidents for other reasons, and thus intercept hardware being used for circumvention. Alternatively, the censor may notice external effects from the passed information, on the basis of which it can decide whether too much traffic is passing. In some cases, circumventors may make this information publicly available, *e.g.* VPN Gate [37] and Tor [34]. We consider the question of exactly how the censor infers these parameters to be out of scope for this paper, and include these suggestions simply to show that our assumption is not entirely unreasonable. An expansion of our results in the framework of Bayesian game theory could accommodate uncertainty in these parameters.

[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.

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 [20, 24, 26]—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 cover traffic L. We normalize both CTP and L by setting $L = 1 - CTP$. In this section the CRS controls CTP and denotes the amount of circumvention traffic that will be allowed to enter the censor's network.

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 other uninvolved 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 $CTP \in [0, 1]$, which is the amount of circumvention traffic the circumventor chooses to send (as a fraction of the total traffic). Often we will assume the circumventor is unable to send more than some fraction of total traffic that is less than one, and so will limit the action space to $CTP \in [0, CTP_{\max}]$ instead, where $CTP_{\max}$ is the maximum portion of total network traffic that the circumvention traffic can potentially be.

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

$$\begin{aligned} U_{\text{cen}} = (-\alpha_{\text{act}} X_{\text{cen}} + \alpha_{\text{bct}}(1 - X_{\text{cen}}))CTP + \\ (\beta_{\text{ant}} X_{\text{cen}} - \beta_{\text{bnt}}(1 - X_{\text{cen}}))(1 - CTP) \end{aligned} \quad (1)$$

$$\begin{aligned} U_{\text{cir}} = (\gamma_{\text{act}} X_{\text{cen}} - \gamma_{\text{bct}}(1 - X_{\text{cen}}))CTP + \\ (\delta_{\text{ant}} X_{\text{cen}} - \delta_{\text{bnt}}(1 - X_{\text{cen}}))(1 - CTP) \end{aligned} \quad (2)$$

Variables $\alpha_{[\text{act,bct}]}, \beta_{[\text{ant,bnt}]}, \gamma_{[\text{act,bct}]}$, and $\delta_{[\text{ant,bnt}]}$ 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 ant and bnt stand for *allow* and *block non-circumvention 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_{ant} and β_{bnt} are the gain, or loss, in utility to the censor of having one unit of non-circumvention traffic transported via, or blocked on, the channel, respectively. The ratios of α_{act} to β_{ant} and of α_{bct} to β_{bnt} 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_{ant}. In contrast, a military agency trying to censor leakage of state secrets might have a very high α_{bct} relative to their β_{bnt} 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_{ant} and δ_{bnt} show the utility gained, or lost, by the circumventor of a single unit of cover 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.

We consider the case where both δ parameters are zero. This case reflects, many but not all, CRS designs (*e.g.* Tor [12], Psiphon [30], or CloudTransport [7]) that are not concerned with the fallout of CRS usage; further designs in the literature do not provide technical provisions to reduce the impact of the fallout on non-CRS traffic; *e.g.*, the examples above. We also assume that the parameter γ_{bct} is zero, reflecting those CRS designs that are ambivalent to blocked CRS traffic (again see exmaples above); 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.

However, it may be the case that there are CRS designs where these assumptions do not hold, for instance where the cost of blocked traffic is not negligible to the circumventor, or for a "spiteful" circumventor that gains *positive* utility from the censor blocking non-circumvention traffic, then the results would need to be extended in a more complex analysis, left for future work.

Thus the circumventor's utility function is reduced to:

$$U_{\text{cir}} = \gamma_{\text{act}} X_{\text{cen}} CTP \quad (3)$$

4.1.1 Analysis

In this section we show that in a single-round game there is only one Nash equilibrium that leaves the channel open.

Theorem 1. *In a single-round game, there only exists one Nash equilibrium that leaves the channel open.*

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

If the players know each others' strategies, the utility of the circumventor is maximized by setting $CTP = F$. However, this solution 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 $CTP = CTP_{\max}$, since this maximizes the utility of the circumventor. Consequently, the profile where the censor plays "On" and the circumventor plays $CTP = F$ is not a Nash equilibrium.

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

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 for allowing and blocking circumvention traffic versus keeping non-circumvention 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 played. In the face of such an opponent, the censor of course must block (contingent on F and $CTP_{\max}$), to avoid the unacceptable volume of circumvention traffic that would be sent.

The key result for cooperation in temporal games, due to Aumann [4], is that the equilibrium 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 inductively, 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. Since the censorship game we study can be modeled with such a coordination-based element (if the censor opens the channel, they "trust" the circumventor to behave rationally and not to defect and send too much traffic), it is straightforward to show by backward induction 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, summed across all rounds of play. Formally, we denote by $U_{\text{cen}}(t)$ and $U_{\text{cir}}(t)$ the respective utilities gained by the censor and circumventor during timestep t. If the game is still proceeding (recall the game ends with probability p af-

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

ter each round) then these utilities are simply the utilities each player derives from a single round of the game, as in the previous subsection. Otherwise, both are zero. The censor's goal is then to maximize $\sum_{t=0}^{\infty} U_{\text{cir}}(t)$, with a corresponding goal for the circumventor. 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.

As this is a multi-round game, a player's strategy is a specification of how they would play in this round, given every possible sequence of preceding rounds of play. In practice, many strategies can be specified that operate on the basis of a finite history.

4.2.1 Analysis

An interesting Nash equilibrium emerges where circumventor and censor are both involved in a repeated game.

Theorem 2. *For $Z = (1-p)CTP_{\max}$, if $F \geq Z$, there exists a Nash equilibrium where the censor's strategy is "On" as long as the circumventor has never played $CTP > Z$ 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, and the circumventor adopts a policy of playing $CTP = Z$ at every step.*

Proof. To show that the censor leaving the channel open and the circumventor playing $CTP = Z$ is a Nash equilibrium we use a proof by construction.

Suppose that both players start in the supposed equilibrium state (where the censor plays strategy s_{cen}, and the circumventor plays strategy s_{cir}). If the circumventor has committed to playing Z, then the censor receives at least as much utility for keeping the channel open in each round as for closing it (by definition of F above). Therefore, the censor cannot improve its utility by closing the channel if the circumventor adopts s_{cir}. Any strategy that contains one or more closed rounds is less profitable than one containing all open rounds.

If the censor has adopted s_{cen}, then (provided γ_{act} is positive, and γ_{bct}, δ_{act} and δ_{bct} are zero), the circumventor cannot improve its utility by sending less traffic than Z per round, since the censor will keep the channels open either way. Therefore we need only consider strategies where the circumventor sends more than Z traffic at some point. If it sends more traffic than Z, it receives $\gamma_{act} CTP_{\max}$ utility in this round, but is certain to receive no traffic in subsequent rounds, since the censor has committed to playing s_{cen}.

Suppose the circumventor adopts a strategy of playing Z for some number of rounds k, after which it deviates and plays $CTP_{\max}$ for one round. Note that if the circumventor plays less than $CTP_{\max}$ in the deviation round, it can derive strictly greater utility by playing $CTP_{\max}$ during the deviation round instead. We will show by induction that there are no profitable deviation rounds for the circumventor to select.

For the base case, suppose $k = 0$, so the circumventor will deviate in the first round of play. A deviation here earns a total of $\gamma_{act} CTP_{\max}$ utility for the entirety of the game, since no utility can be earned during any subsequent round. In contrast, the circumventor earns an expected $\gamma_{act} \sum_{i=0}^{\infty} p^i Z = \gamma_{act} \frac{1}{1-p} Z = \gamma_{act} CTP_{\max}$ for playing Z in every round. Therefore the circumventor gains no utility in expectation for deviating in the first round.

For the inductive step, suppose that there are no profitable deviations for time steps less than k. The circumventor earns an expected $\gamma_{act} \sum_{i=0}^{k-1} p^i Z = \gamma_{act} \frac{1-p^k}{1-p} Z = (1 - p^k)\gamma_{act} CTP_{\max}$ from the first $k-1$ timesteps, and then earns $p^k \gamma_{act} CTP_{\max}$ for deviating in the k^{th} round, for a total of $\gamma_{act} CTP_{\max}$. This is exactly the amount earned by not deviating, so no profitable deviations exist in round k either. □

Note that if the circumventor plays a value greater than F, the censor will be better off keeping the channel closed, and if they play a value less than $Z = (1-p)CTP_{\max}$ then the circumventor would prefer to play Z in the first round. Any value between these two however, is an equilibrium. Either party can set the exact level used by making a public declaration that they will play a strategy in this set, to which the best response of their opponent is to play the complementary strategy.

The equilibrium just outlined depends on the assumption that the censor turns off the channel and never turns it back on if the traffic sent exceeds Z. However, this may not be a credible threat since the censor wants the channel open in the long run, so as to allow the non-circumvention traffic to get through. Furthermore, if the circumventor drops the circumvention traffic proportion, *i.e.* $CTP < Z$, after sending $CTP_{\max}$, 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, we can refine the model by having the censor instead commit 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 $CTP = CTP_{\max}$ traffic. After this, the censor would close the channel for τ rounds, resulting in a total utility for the circumventor of $\gamma_{\text{act}} CTP_{\max}$ 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 proportion to whatever value will maximize profits). An equilibrium where the censor keeps the channel open, and the circumventor sends $Z = CTP = \frac{1-p}{1-p^\tau} CTP_{\max}$ traffic then follows by similar logic to the equilibrium with an infinite punishment period, provided that $F > Z$. 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 proportion 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. Certainly it is plausible that censors might choose to temporarily close the channel 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 a repeated game since they are not *subgame perfect* equilibria. A subgame perfect equilibrium requires that strategies be locally rational. That is, starting from any point in the game, players cannot do better than continuing to play the strategies prescribed by the equilibrium. This is clearly not the behavior we observe when the censor engages in punishment. For example, if we start the game immediately after the circumventor has sent a large burst of traffic, then if the circumventor has committed to playing Z from now on, a rational censor could (locally) improve its utility by ceasing punishment and reopening the channel immediately. 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 too.

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 [17] provides a more complex equilibrium that is of similar form, yet is also subgame perfect (*i.e.*, players cannot do better than following through on the threats they make). For the purpose of this analysis, the perfect Nash folk theorem states that, provided players are able to randomize their strategies (*e.g.*, commit to strategies where closing the channel occurs say, with probability 0.5), then for any given convex combination of the payoffs players could receive at different points in the parametric family of strategy profiles $\overline{*}$, there exists a set of "supporting" profiles such that playing the strategies prescribed by $\overline{*}$ is a subgame perfect Nash equilibrium for all players, provided p (the probability of playing one more round of the game) is large enough. Below, we construct the three profiles needed to support the equilibrium produced in the earlier analysis, and specify the transitions needed between them to "support" the desired equilibrium in the context of the repeated censorship game. Our contribution here is construction of the equilibrium, using the approaches outlined by Fudenberg and Maskin [17], and the reader should refer to that work for further details.

The four profiles are presented in tabular form in Table 1. The full strategy for each player (they are symmetric) is as follows:

1. If this is the first round, or the opponent has only ever played the action specified for them in $\overline{*}$, play the action specified for this player in $\overline{*}$.
2. Otherwise, if both players played the actions specified by r_{cen} in the previous round, then this player should play the action specified by r_{cen} in the next round.
3. Otherwise, if both players played the actions specified by r_{cir} in the previous round, then this player should play the action specified by r_{cir} in the next round.
4. Otherwise, if both players played the actions specified by $\underline{*}$ in fewer than the last τ rounds, then this player should play the action specified by $\underline{*}$ in the next round.
5. If both players have played the actions specified by $\underline{*}$ for the last τ rounds, then this player should play the action specified by r_{cen} in the next round if the censor did not play according to steps 1–4 during the round $\tau + 1$ ago, and r_{cir} otherwise.

Theorem 3. *The above strategy profile is a subgame perfect Nash equilibrium, provided that $\sigma \geq 1$ and that $\epsilon > 0$, as well as that $F \geq Z$ and $\frac{Z}{1-p} > \frac{Z-\epsilon}{1-p} - \frac{Z-\epsilon}{1-p^{\sigma+1}} > \frac{CTP_{\max}}{1-p^\tau+1}$ and that $CTP_{\max} > F$. (*i.e.*, that both players prefer receiving the long-term payoffs incurred in $\overline{*}$ to either r states, and prefer the longterm payoff of either r states to the payoff of repeatedly deviating and being punished in state $\underline{*}$).*

Proof. To show that the proposed strategies are a subgame perfect Nash Equilibrium, we need only show that unilaterally 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-

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_{cen}	CTP
$\overline{*}$	**1 (open)**	Z
r_{cen}	**1 (open) for σ rounds, then 0 (closed) for one round**	**Z for σ rounds, then $CTP_{\max}$ for one round**
r_{cir}	**1 (open) for σ rounds, then 0 (closed) for one round**	**$Z - \epsilon$ for σ rounds, then $CTP_{\max}$ for one round**
$\underline{*}$	**0 (closed)**	$CTP_{\max}$

term benefits. This follows from Blackwell's one-shot deviation principle [5].

We start by considering the phase $\overline{*}$. In this phase, a single deviation by the censor after k rounds 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}. Payouts prior to the deviation are identical whether the censor deviates or not at step k, and so need not be considered. Since $F > Z$, we know the censor receives a positive utility q in each step of $\overline{*}$, and that not deviating would pay out an expected amount $p^k(\sum_{i=0}^{\infty} p^i q) = \frac{p^k q}{1-p}$. Deviating produces total earnings of $p^k(\sum_{i=0}^{\tau+1} p^i 0) = 0$ for the deviation and punishment phases, followed by $p^k(\sum_{i=\tau+2}^{\infty} p^i q - \sum_{i=\tau+2}^{\infty} p^{(\sigma+1)i} q) = p^k(\frac{p^{\tau+2}}{1-p} q - \frac{p^{\tau+\sigma+3})}{1-p} q$. A single step deviation is not profitable for the censor in $\overline{*}$. During the punishment phase $\underline{*}$, if the censor deviates (opening the channel) for one timestep after k steps, it receives negative utility (since $\frac{CTP_{\max}}{1-p^{\tau}} > F$). After this, it receives 0 utility for τ rounds, followed by $p^k(\frac{p^{\tau+2}}{1-p} q - \frac{p^{\tau+\sigma+3}}{1-p} q)$ in the endgame. If the censor had not deviated from the punishment phase, it would have spent at most τ more rounds there, followed by receiving at least $p^k(\frac{p^{\tau+1}}{1-p} q - \frac{p^{\tau+\sigma+2}}{1-p} q)$ during the end game (more if the circumventor was being punished). Therefore, the censor does not benefit from a single step deviation during the punishment phase.

During the r_{cir} phase, the censor receives $\frac{q}{1-p} - \frac{p^{\sigma+1}}{1-p} q$ for not deviating. There are two possible deviations: closing the channel during one of the σ rounds when it should be open, or opening it during round $\sigma + 1$. If the censor closes the channel after k rounds of playing the correct strategy, during one of the σ rounds, it earns 0 utility for that round, followed by 0 utility for τ rounds of punishment, followed by $\frac{p^{\tau+1+k}}{1-p} q - \frac{p^{k+\tau+\sigma+2}}{1-p} q$ utility in the long run. In contrast, it could have earned a non-zero amount in the deviation round, at least 0 during each punishment round, and then the same endgame amount. Opening the channel during the $\sigma + 1$ round produces negative utility for that round (since $\frac{CTP_{\max}}{1-p^{\tau}} > F$, and the analysis is otherwise identical, yielding net negative utility in that case also. Therefore, the censor does not benefit from a single step deviation during the r_{cir} phase. The analysis for the r_{cen} phase follows an identical argument, but with a slightly different (slightly larger) value of q. There is no beneficial single step deviation in that phase either. Therefore, there are no beneficial single step deviations for the censor.

For the circumventor, not deviating after k steps during $\overline{*}$ pays $p^k \sum_{i=0}^{\infty} p^i Z = p^k \frac{Z}{1-p}$. Deviating will pay an initial $p^k CTP_{\max}$, followed byτ rounds of zero utility, followed by an endgame spent in r_{cir} paying a total of $p^{k+\tau+1}(\frac{Z-\epsilon}{1-p} - \frac{p^{\sigma+1}(Z-\epsilon)}{1-p^{\sigma+1}})$. Since we have assumed a p large enough that $\frac{(1-p^{\tau+1})Z}{1-p} > CTP_{\max}$, and $Z > Z - \epsilon$, more total utility is earned by not deviating. Therefore, there is no profitable single-step deviation for either player from the $\overline{*}$ state. An identical argument can be used to show no profitable single-step deviations exist from the r_{cir} and r_{cen} steps (note that there is no special case here because deviating during the $\sigma + 1$ step of that phase has no effect for the circumventor).

All that remains is to show that no profitable single step deviations exist for the circumventor from the punishment phase. Not deviating after k rounds of punishment results in at most τ more rounds of punishment, followed by an endgame in r_{cir}, earning $p^{k+\tau}(\frac{Z-\epsilon}{1-p} - \frac{p^{\sigma+1}(Z-\epsilon)}{1-p^{\sigma+1}})$. Deviating results in τ more rounds of punishment, and earns no additional utility during deviation (since the censor is still playing closed). The endgame is the same, with $p^{k+\tau+1}(\frac{Z-\epsilon}{1-p} - \frac{p^{\sigma+1}(Z-\epsilon)}{1-p^{\sigma+1}})$ in earnings. Since not deviating yields *at most* τ more rounds, it can be better, and is certainly no worse. Therefore there is no profitable single step deviation for the circumventor. □

Interestingly, we note that p could be replaced by any discounting factor for the utility of future rewards, so long as the game remains infinite. 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 it is a reasonable policy for the circumventor interested in maintaining a long-term communication channel to keep $CTP \leq F$, with the exact value dependent on the utility functions of the two players, and the credibility of threats made by the censor.

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 (CTP) from the non-circumvention 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 non-circumvention 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:

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

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

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 $CTP \cdot \alpha_{\text{bct}}(TPR(1 - X_p) + FNR(1 - X_n))$ because a fraction CTP 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:

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

Similarly, the other settings yield the following utility equations:

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

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

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 CTP.

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:

$$CTP \leq \frac{\beta_{\text{ant}} + \beta_{\text{bnt}}}{\alpha_{\text{act}} + \alpha_{\text{bct}} + \beta_{\text{ant}} + \beta_{\text{bnt}}}, \tag{9}$$

or $CTP \le F_{ab}$, where $F_{ab} = \frac{\beta_{\mathrm{ant}}+\beta_{\mathrm{bnt}}}{\alpha_{\mathrm{act}}+\alpha_{\mathrm{bct}}+\beta_{\mathrm{ant}}+\beta_{\mathrm{bnt}}}$. 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 $U'_{cen^{(1,1)}} \ge U'_{cen^{(0,1)}}$ and the following must also hold:

$$CTP \le \frac{FPR(\beta_{\mathrm{ant}}+\beta_{\mathrm{bnt}})}{TPR(\alpha_{\mathrm{act}}+\alpha_{\mathrm{bct}})+FPR(\beta_{\mathrm{ant}}+\beta_{\mathrm{bnt}})}, \quad (10)$$

or $CTP \le F_{am}$, where $F_{am} = \frac{FPR(\beta_{\mathrm{ant}}+\beta_{\mathrm{bnt}})}{TPR(\alpha_{\mathrm{act}}+\alpha_{\mathrm{bct}})+FPR(\beta_{\mathrm{ant}}+\beta_{\mathrm{bnt}})}$. 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:

$$CTP \le \frac{TNR(\beta_{\mathrm{ant}}+\beta_{\mathrm{bnt}})}{FNR(\alpha_{\mathrm{act}}+\alpha_{\mathrm{bct}})+TNR(\beta_{\mathrm{ant}}+\beta_{\mathrm{bnt}})}, \quad (11)$$

or $CTP \le F_{mb}$, where $F_{mb} = \frac{TNR(\beta_{\mathrm{ant}}+\beta_{\mathrm{bnt}})}{FNR(\alpha_{\mathrm{act}}+\alpha_{\mathrm{bct}})+TNR(\beta_{\mathrm{ant}}+\beta_{\mathrm{bnt}})}$. 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 CTP 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 $\alpha_{\mathrm{act}}+\alpha_{\mathrm{bct}}$ and $\beta_{\mathrm{ant}}+\beta_{\mathrm{bnt}}$ 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 $TPR \ge FPR$ (and equivalently, that $TNR \ge FNR$).

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

$$\begin{aligned} & F_{ab} \ge F_{am} \\ \Leftrightarrow\ & \frac{\beta}{\alpha+\beta} \ge \frac{FPR\cdot\beta}{TPR\cdot\alpha+FPR\cdot\beta} \\ \Leftrightarrow\ & \frac{\alpha+\beta}{\beta} \le \frac{TPR\cdot\alpha+FPR\cdot\beta}{FPR\cdot\beta} \\ \Leftrightarrow\ & \frac{\alpha}{\beta} \le \frac{TPR\cdot\alpha}{FPR\cdot\beta} \\ \Leftrightarrow\ & FPR \le TPR \end{aligned} \quad (12)$$

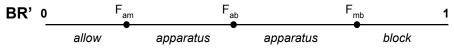

Fig. 1. Best censor strategies at critical circumvention traffic thresholds. The censor's strategies are in *italics*. The circumventor's strategies are to send a proportion of circumvention traffic, $0 \le CTP \le 1$, with the critical thresholds marked as F_{am}, F_{ab}, and F_{mb}.

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

$$\begin{aligned} & F_{mb} \ge F_{ab} \\ \Leftrightarrow\ & \frac{TNR\cdot\beta}{FNR\cdot\alpha+TNR\cdot\beta} \ge \frac{\beta}{\alpha+\beta} \\ \Leftrightarrow\ & \frac{FNR\cdot\alpha+TNR\cdot\beta}{TNR\cdot\beta} \le \frac{\alpha+\beta}{\beta} \\ \Leftrightarrow\ & \frac{FNR\cdot\alpha}{TNR\cdot\beta} \le \frac{\alpha}{\beta} \\ \Leftrightarrow\ & FNR \le TNR \end{aligned} \quad (13)$$

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

Given this ordering, the censor will play according to the following strategies, which are depicted in Figure 1. When $CTP \le F_{am}$ the censor will allow all traffic to flow. When $F_{am} \le CTP \le F_{ab}$ or $F_{ab} \le CTP \le F_{mb}$ then the censor will use the apparatus rather than allowing or blocking all the traffic, respectively. Indeed, as long as the apparatus does better than a coin toss then the F_{ab} threshold does not matter, reducing the preceding to $F_{am} \le CTP \le F_{mb}$. Finally, when $CTP > F_{mb}$ the censor should block all traffic.

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

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. This is because the only effect is to modify the parameters of the players' utility functions, and the results from those two cases hold for a specified range of parameterizations. The main effect is on the fraction of the total traffic, CTP, the circumventor can send through while ensuring that the inequalities above remain true.

5 Extensions and Analysis

So far, we have analyzed a simple model where the circumvention system 1) utilizes only a single protocol and 2) the CRS's mechanism to ensure the amount of circumvention traffic remains below the critical threshold is immune from the censor's influence.

Also, we considered a linear function for censor utility but it might be the case that the censor's stakes (costs) to blocking CRS traffic, and not making mistakes, ramp up faster as rates of errors increase making the censor more risk averse than the linear model above describes. This is reasonable to assume since the use of the Internet is expected to be unhindered, and after a certain amount of blocked traffic (*i.e.* reduction in functionality) it quickly becomes apparent that something is wrong which may trigger an escalating wave of user unrest, for instance. One way to capture this is to utilize an exponential utility function for the censor, such as the following example:

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

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

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 that gets through (*i.e.*, the false negatives) and FPR is the percentage of non-circumvention 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).

In subsection 5.1 we perform a closed-form analysis to investigate the relaxation of the condition of using only one protocol and to explore the effect of multiple protocols on the equilibrium solutions. In section 6 we relax the other condition as well, and examine the case when the CRS's CTP control mechanism is open to the censor's influence. Unfortunately, a closed-form style of analysis becomes more complex in this scenario and less straightforward to reason about. Therefore, we introduce a numerical analysis tool (or numerical analyzer) to assist in solution finding and to gain further insights, the details of which can be found in Appendix A.

Censor types. The results of any particular analysis depend on the type of the censor; *i.e.*, the particular values of C and D. Solving Equation 14 to find the threshold amount of circumvention traffic gives $\frac{C}{C+D}$, which we denote by F_{ab} using the same notation as before. Hence, we can describe censor types by the value F_{ab}. Values less than 0.5 denote a censor who is more averse to information leakage than collateral damage, while values higher than 0.5 denote a censor who is more averse to collateral damage than information leakage. At the lower extreme no amount of circumvention traffic will be tolerated, while at the higher extreme an unbounded amount of circumvention traffic will be tolerated.

5.1 Moving from one to multiple protocols

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 the ease with which it can be imitated or the system deployed.

5.1.1 Single protocol

Our analysis so far shows that for a given censor there exists a circumvention traffic threshold F_{ab} below which a rational censor will not block the communication channel. Since, for any given censor, this threshold is fixed and proportional to the size of L, the only way to achieve more throughput for circumvention traffic is to utilize another channel with a higher amount of cover traffic L.

In this simple scenario, the circumventor who can only target one protocol should pick the protocol with the largest L that they can successfully imitate. However, it might be the case that there is no single protocol with a sufficiently high amount of cover traffic to meet the CRS users' throughput demands, and so simultaneously utilizing multiple protocols is the natural next step.

5.1.2 Two or more protocols

Each cover protocol $i \in \{1, \ldots, n\}$ provides an amount of non-circumvention traffic L_i that is some portion of the total non-circumvention traffic L. We order the protocols so that the amount of non-circumvention traffic across the protocols is $\langle L_1, \ldots, L_n \rangle$ in descending order. Also, there is an amount of circumvention traffic CTP_i imitating protocol i that is some portion of the total circumvention traffic CTP. We denote by $R_i = CTP_i + L_i$ the total amount of traffic over protocol i. The total amount of traffic on the network is then $\sum_{i=1}^{n}(L_i + CTP_i) = L + CTP = 1$.

Theorem 4. *If $CTP \le F_{ab}$, the optimal distribution of traffic over many* real *protocols is to allocate $CTP_i = L_i \cdot \frac{CTP}{1-CTP}$ over protocols $1, \ldots, n$. If $CTP > F_{ab}$, it is to allocate $CTP_i = L_i \cdot \frac{F_{ab}}{1-F_{ab}}$ over protocols $1, \ldots, n$ and then to additionally allocate, or* dump*, any remaining, surplus, traffic over protocol n, which will be blocked.*

Proof sketch. From section 5.1.1, we make the observation that the censor will only block a protocol i if and only if $\frac{CTP_i}{CTP_i+L_i} > F_{ab}$. In the first case, where $CTP \le F_{ab}$, setting $CTP_i = L_i \cdot \frac{CTP}{1-CTP}$ means that $\frac{CTP_i}{CTP_i+L_i} = CTP \le F_{ab}$, so none of the protocols will be blocked, and the circumventor will have maximum utility.

In the second case, where $CTP > F_{ab}$, setting $CTP_i = L_i \cdot \frac{F_{ab}}{1-F_{ab}}$ means that $\frac{CTP_i}{CTP_i+L_i} = F_{ab}$, so at this point, none of the protocols will be blocked (they are each right at the edge of what the censor will tolerate), but there is still $F_{ab} - CTP$ circumvention traffic to allocate to some protocol.

Trying to send this surplus traffic over any protocol will cause that protocol to be blocked, and the circumventor will lose the utility associated with the circumvention traffic sent over that protocol. Since the amount of circumvention traffic CTP_i over each protocol is proportional to the amount of non-curcumvention traffic L_i, the circumventor's best strategy is to use protocol n, which has the smallest amount of both kinds of traffic.

Unilateral deviation from this equilibrium point produces less utility for the players; for instance crossing the threshold on one protocol while underutilizing another to compensate produces suboptimal utility.

An implication of the above is that we can treat the usage of individual protocols independently from each other, meaning that there are no further constraints to picking protocols other than to maximize the amount of cover traffic, and thus the amount of circumvention traffic.

5.1.3 Dumping vs. throttling

We consider the situation where instead of dumping the surplus traffic on protocol n and letting the censor block it, we simply elect to not send it, effectively throttling the CRS to ensure $CTP \le F_{ab}$. We can model this situation as a special case of the above theorem—with the proof intact—by adding an additional *pseudo* protocol $n+1$ where $L_i = 0$. The same optimal traffic allocation strategy applies in this case as above (*i.e.*, the surplus is sent over the last protocol, $n+1$), except now we obtain greater circumventor utility, since $CTP > CTP_{[1,\ldots,n-1]}$. Hence, it is always better to throttle rather than to dump.

6 The interfering censor

In the preceding analysis the CTP can be throttled by the CRS's idealized CTP control mechanism to remain below the censor's blocking threshold. We have thus far assumed that this ideal mechanism remains outside the censor's influence. We now relax this assumption and describe a traffic flooding attack the censor could mount on the control mechanism.

6.1 Flooding attack

The censor will inject fake circumvention traffic, CTP_{fake} into the pool of real circumvention traffic CTP_{real}, such that $CTP_{real} + CTP_{fake} = CTP_{R+F}$. Since the CRS cannot tell the difference between real CRS traffic and that of the censor, the censor can inject an arbitrary amount of traffic into the CRS network, and thus inflate CTP_{R+F} at will. We also assume that throughout the attack CTP_{real} remains constant.

The objective of throttling is to ensure that $CTP_{R+F} = F_{ab}$ gets through and to ignore sending the remaining traffic. However, when the censor is injecting fake traffic, the open protocols are only transmitting a fraction $\frac{CTP_{real}}{CTP_{real}+CTP_{fake}}$ of CTP_{real}. The remaining traffic getting through will be the circumventor's fake traffic. As the ratio $CTP_{fake} : CTP_{real}$ increases, less and less real circumvention traffic will get through, thus increasing the censor's utility. Since the CRS ensures that CTP_{R+F} traffic on each protocol will be less than CTP_i, the censor does not have to block any protocols, and incurs no collateral damage to mounting this attack. A key feature of this attack is that it causes the *eviction* of real circumvention traffic from open protocols to the throttling "protocol".

Dumping is equally susceptible to this attack, and with the same level of effectiveness. Since there is no additional cost to the censor, in terms of lost non-circumvention traffic, it can increase CTP_{fake} and drive CTP_{real} downwards by causing the eviction of real CRS traffic to the dumping protocol. From the perspective of the circumventor, it is worse to dump than to throttle since when dumping she also loses protocol n, which can carry some additional CTP_{real}.

If the cost to the censor for using the CRS client and sending traffic over it is zero, then the censor can drive CTP_{real} down to zero with impunity. If the cost is not zero, then the censor needs to insure that the sum of the utility gained (by evicting real CRS traffic to the throttling or dumping protocols) is greater than the cost of the fake CRS traffic required to evict that quantity of real CRS traffic.

6.2 Alternative mechanisms

The flooding attack leverages a key feature of the throttling and dumping mechanisms—that they ensure that only at most one protocol will get blocked, even with censor manipulation. The CRS protecting protocols from being blocked is the lynchpin of the attack. Looking back at section 5.1.2, we know that the optimal solution is to allocate traffic in a monotonically decreasing fashion over the protocols, with the last protocol allowed to break that trend. We now consider a similar traffic allocation mechanism with the additional constraint that it be strictly monotonic across all protocols. This means that we exclude any traffic distribution where the non-circumvention traffic allocated over protocol a is greater than that allocated over protocol b when the quantity of non-circumvention traffic going over protocol b is greater than that of a. The effect of the strict monotonicity is that no protocol will be protected from blocking. Like before, the optimal censor strategy is still to block a protocol if and only if $CTP_i > R_i \cdot F_{ab}$.

We will first analyze strictly monotonic solutions and the nature of their traffic allocation strategy, in terms of utility, before we return to how robust they are to the flooding attack in section 6.2.3. As noted earlier, we manage the complexity of computation using the numerical analyzer we describe in Appendix A.

6.2.1 The life monotonic

Since $CTP > F_{ab}$ and because the CRS is no longer throttling the CTP, in the remainder of this section, CTP will denote the circumvention traffic proportion generated by the users of the system, and not how much traffic the censor will allow through.

To analyze this scenario we use our numerical analyzer to identify optimal strategies in this setting. We increase the amount of circumvention traffic, due to the CRS userbase, above F_{ab} in increments of 5% and analyze the resulting best strategies for a censor where $F_{ab} = 0.5$.[3]

Table 2. Optimal circumvention traffic (no flooding attack) distribution for censor type $F_{ab} = 0.5$, where $CTP > F_{ab}$ for various values, S_{cir} is the circumventor's strategy (as percentages of CTP over the six protocols), S_{cen} is the censor's strategy (**A** and **B** denote an allowed or blocked ith protocol respectively), U_{cen} is the censor's utility, U_{cir} is the circumventor's utility, and RU_{cir} is the circumventor's utility relative to the F_{ab} case. For example, $0.79 \cdot 1.10 = 0.869$.

CTP	S_{cir}	S_{cen}	U_{cen}	U_{cir}	RU_{cir}
F_{ab}	53,25,9,7,4,4	AAAAAA	0.61	1.00	1.00
$F_{ab} \cdot 1.05$	48,23,16,5,4,4	AABAAA	0.63	0.84	0.88
$F_{ab} \cdot 1.10$	46,22,21,5,3,3	AABAAA	0.65	0.79	0.87
$F_{ab} \cdot 1.15$	43,21,21,9,3,3	AABBAA	0.66	0.70	0.81
$F_{ab} \cdot 1.20$	41,20,20,13,3,3	AABBAA	0.67	0.67	0.80
$F_{ab} \cdot 1.50$	25,25,25,21,2,2	ABBBAA	0.78	0.29	0.44

The results are presented in Table 2 and show that even with 50% more circumvention traffic than allowed, it is still possible to get a relative utility RU_{cir} of $\sim 44\%$ (as compared to the optimal when $CTP \leq F_{ab}$; *e.g.*, when throttling).

We note in this scenario that none of the ratios for the allowed protocols ever equals or exceeds $R_i \cdot F_{ab}$. This tells us that the F_{ab} threshold is an upper limit. Indeed, comparing the utility relative to the F_{ab} case (where $U_{cir} = 1$) in column RU_{cir}, we see that by increasing the amount of circumvention traffic we get less utility than was possible by adhering to the F_{ab} threshold. This shows that strictly monotonic solutions are less optimal than throttling solutions.

While it seems odd at first glance that circumventors should send traffic over protocols they know to be blocked, doing so actually preserves the equilibrium: if the censor were to unblock one of the protocols, it is important that that action would cause a decrease in the censor's utility due to circumvention traffic suddenly beginning to flow. Had the circumventor been sending dummy traffic instead, or stopped sending traffic at all over that protocol, then—recalling that the circumventor's actions are known to the censor—unblocking that protocol would *increase* the censor's utility by decreasing the collateral damage caused by blocking non-circumvention traffic.

6.2.2 Protocol ratio dependence

We now show that the ratios between the set of protocols plays a significant role on the circumventor's utility. We first investigate the two-protocol setting to identify the trends before analyzing the six-protocol scenario.

Figure 2 illustrates the resultant utility of picking two protocols with three different protocol ratios: $R_{90,10} = \langle 0.90, 0.10\rangle$, $R_{75,25} = \langle 0.75, 0.25\rangle$, and $R_{51,49} = \langle 0.51, 0.49\rangle$. For low to moderately high surplus of CTP we

[3] Results are similar for other values of F_{ab}.

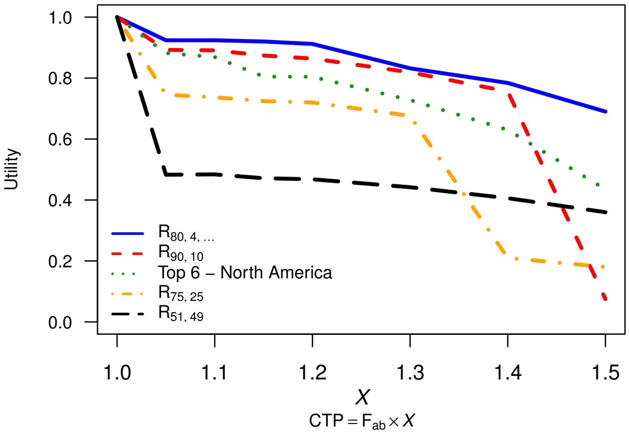

Fig. 2. Circumventor utility at various CTP values over and beyond F_{ab} for various cover protocol ratios.

note that protocol sets with ratio skew result in higher utility than sets with ratio similarity. Indeed, the severely skewed protocol set $R_{90,10}$ provides higher utility than the others. However, after a certain point, $\sim F_{ab}+40\%$, the situation is flipped and the more even ratio protocol set provides higher utility than the others. The reason for this is that when F_{ab} is exceeded on a protocol, not only is the surplus circumvention traffic blocked, but so is the circumvention traffic that could have gotten through. In the skewed distribution there is a lot of surplus-dumping potential in the smaller protocols with little cost of dropping traffic that would otherwise have gone through; this explains why they are better for smaller surplus values than the more even protocol distributions. However, the censor also benefits by this: there is less allowed traffic to harm by blocking the protocol and as the surplus adds up the cost gets further driven down. The sudden drop occurs because the surplus is so great and the costs of blocking allowed traffic so low that blocking the larger protocol becomes economically feasible. It then becomes better to sacrifice the larger protocol by sending all the surplus on it instead and saving the small protocols to allow some traffic through.

Figure 2 also illustrates the utility of picking six protocols according to two different protocol ratios: the traffic-volume ratios $\langle 0.51, 0.25, 0.09, 0.07, 0.04, 0.04 \rangle$ supplied by the December 2015 Sandvine survey of North American Internet traffic trends [32], and a very skewed $R_{80,4...} = \langle 0.80, 0.04, 0.04, 0.04, 0.04, 0.04 \rangle$. We again see that the heavily skewed protocol set $R_{80,4...}$ provides higher utility as CTP increases past F_{ab}, although the difference is not as drastic as before. Compared to the two-protocol case the addition of more protocols generally slows the decline in utility as the CTP grows.

The above has implications for picking protocols to use as cover traffic. Before, in subsubsection 5.1.2 where $CTP \leq F_{ab}$, protocols were independent and only the amount of cover traffic was the criteria. Now, when $CTP > F_{ab}$, we see that we cannot simply try to maximize the cover traffic, but need to be strategic in which set of protocols we pick.

However, one must keep in mind that the same set of protocols may have different ratios on one network as compared to another. If this is not taken in to account it would likely cause suboptimal utility as the wrong circumventor strategy is played for that network. The implication for the CRS is that strategies are on a per network (censor) level.

Related to this, one must also take into account the existence of other circumvention traffic due to other CRS systems on the chosen protocols, since the censor will evaluate its utility globally on all the traffic. This is because the censor is unable to tell circumvention traffic from non-circumvention traffic, which implies they are also unable to tell apart circumvention traffic due to different CRS systems.

6.2.3 Attack mitigation and comparison

Let us now consider strict monotonic mechanism in the presence of the flooding attack. Note that we start from a stable equilibrium where no protocols have already been blocked since $CTP_{real} = CTP = F_{ab}$. From this starting point we will investigate the effect of various quantities of fake CRS traffic (*i.e.*, $CTP_{R+F} > F_{ab}$) on the equilibrium points of the game and utilities of the two players.

In Table 2, where we previously considered the surplus as coming from the CRS userbase, we can now consider the surplus, or CTP_{fake}, as the consequence of the flooding attack. Using the data in the table we will re-calculate how much real CRS traffic is actually blocked, or CTP_{real_loss}, in the presence of varying proportions of CTP_{fake} and perform a cost-benefit analysis.

Tables 3 and 4 provide the results of a cost-benefit analysis of the censor mounting the flooding attack. For the analysis we compute the utility gained through the reduction of real CRS traffic, and the utility lost due to non-circumvention traffic being blocked on protocols where the strictly monotonic traffic allocation violated the F_{ab} threshold.

The censor will only deploy the flooding attack if it causes a gain in censor utility from blocked real CRS traffic (CTP_{real_loss}) that outweighs the loss in censor utility from blocked non-circumvention traffic (L_{loss}) caused by protocol traffic violations. For the censor type with equilibrium point $F_{ab} = 0.5$, the ratio of $CTP_{real\_loss} : L_{loss}$ has to be $1:1$, or as a fraction $\frac{CTP_{real\_loss}}{CTP_{real\_loss}+L_{loss}}$. For the flooding attack to be cost-effective this fraction must exceed the equilibrium point; *i.e.*, $\frac{CTP_{real\_loss}}{CTP_{real\_loss}+L_{loss}} > F_{ab}$. The aim of this analysis is not to find the maximum cost effectiveness of the attack, but when,

Table 3. Cost-benefit analysis for censor deploying a flooding attack on the monotonic mechanism. Here $F_{ab} = 0.5$, using the Sandvine protocol ratios, $\langle 0.51, 0.25, 0.09, 0.07, 0.04, 0.04\rangle$, where CTP_{real_loss} is the fraction of real CRS traffic blocked, and L_{loss} is the fraction of non-circumvention traffic blocked. For an economically sound flooding attack *Cost eff.* $\geq F_{ab} = 0.5$.

CTP_{R+F}	CTP_{real_loss}	L_{loss}	Cost eff.
$F_{ab} \cdot 1.20$	0.264	0.16	0.62
$F_{ab} \cdot 1.50$	0.355	0.41	0.46
$F_{ab} \cdot 1.75$	0.218	0.51	0.30

Table 4. Cost-benefit analysis for censor deploying a flooding attack on the monotonic mechanism. Here, $F_{ab} = 0.5$, using traffic ratio $R_{80,4...} = \langle 0.80, 0.04, 0.04, 0.04, 0.04, 0.04\rangle$, where CTP_{real_loss} is the fraction of real CRS traffic blocked, and L_{loss} is the fraction of non-circumvention traffic blocked. For an economically sound flooding attack *Cost eff.* $\geq F_{ab}$=0.5.

CTP_{R+F}	CTP_{real_loss}	L_{loss}	Cost eff.
$F_{ab} \cdot 1.60$	0.264	0.12	0.69
$F_{ab} \cdot 1.70$	0.225	0.16	0.58
$F_{ab} \cdot 1.80$	0.200	1.00	0.17

if ever, it is not cost effective for the censor to continue the attack, since as long as the cost effectiveness is greater than F_{ab} the censor should continue the attack, even if it decreases the cost effectiveness.

In Table 3, we use the Sandvine traffic-volume ratios. We note that the flooding attack is economically sound until there is approximately 50% fake traffic. At this point, the attack is no longer viable. The same results hold for censor of other types; *i.e.*, other values of F_{ab}.

In Table 4, we investigate the role of protocol ratio skewness, and see that the attack is still self-limiting, however at a much larger surplus value, about 80%. This leads us to believe that a skewed protocol distribution allows the censor to extract more utility from the flooding attack than is possible under a less skewed protocol distribution, such as the linearly decreasing Sandvine ratios. Again, the analysis for other censor types also displays the same trends.

From our analysis it is apparent that under the strictly monotonic mechanism the censor cannot evict a larger proportion of real CRS traffic, as compared to the throttling and dumping mechanisms. Therefore, while the monotonic allocation mechanism is susceptible to the flooding attack, it is more robust to the flooding attack than the throttling or dumping mechanisms, where the censor can drive the amount of successful real CRS traffic arbitrarily low.

If the costs of mounting the attack were non-zero then these would add up to cause the censor to self-limit earlier at a lower loss of real CRS traffic. This is in contrast to the throttling and dumping mechanism where the censor self-limits the attack only if there is non-zero cost to the attack.

6.3 Discussion

There are trade-offs between utility and robustness against attack that the CRS designer must carefully evaluate when deciding which traffic allocation mechanism to use.

It is clear that throttling always provides higher utility than the strictly monotonic solution in the non-adversarial setting. However, we see that monotonic solutions provide positive circumventor utility, although not as high as the throttling case, that are robust to attack. So, while we can achieve optimum throughput by throttling, one can not always depend on these results under all operating scenarios. Indeed, under the an active censor, the monotonic approach appears more utility maximizing. One way around this limitation is for the CRS that wishes to retain the optimal results from throttling should invest in making the cost of mounting the flooding attack not cost-effective for the censor.

From our limited analysis of protocol ratios we see that they can affect the scale of the flooding attack. While we leave a more thorough analysis of identifying optimal protocol ratios for future work, the current observations are still useful for real-world situations where the number and choice of protocols to pick from may be constrained, due to reasons such as lack of implementations and protocols that are blocked from the outset. The upshot is that even if the CRS designer is unable to pick her protocols to maximum effect, the strictly monotonic mechanism provides positive utility and more robustness to attack over the throttling and dumping mechanisms.

Finally, we note an interesting feature of the game when $CTP > F_{ab}$: the censor's protocol blocking behavior is independent of their type, being governed by protocol ratios alone. We noted this when we analyzed the game when there is a surplus when the CRS is oversubscribed, and when the censor is flooding the network. This is useful since we can conduct an analysis of our CRS designs in scenarios where the protocol distributions are the same, that generalizes across all censor types.

7 Assumptions and Limitations

We assume that the circumventor is not "spiteful"; that is, they seek only to maximize the amount of traffic they get through and do not seek to harm the censor if reduces this amount. In contrast to this assumption, in some domains players may be spiteful and derive utility from their competitors' loss. Some works [6, 27] have examined positive or negative externalities incurred by such behavior in popular auction mechanisms. However, we do not follow this thread and leave it as an avenue for future work.

Another assumption is that circumventors suffer no loss from blocked traffic. This is a reflection of those CRS designs where traffic transportation is best effort and the cost of bandwidth is borne by the user of the CRS, where this cost is often not by the byte but a fixed monthly amount. However, for censorship regimes where the circumventor incurs a cost in proportion to their usage, then our results may not necessarily hold.

We assume that the proportion of the total network traffic for a protocol is a proxy for how large an effect of interfering with it would have on the censor's utility. In reality, other aspects may also matter (more) such as the protocol's importance to commerce, or usefulness for state surveillance, or as a means of propaganda. While our results may not translate to these settings, the analysis framework we have developed can still be applied with some modifications to the players' utility functions.

Our numerical analysis tool introduces some small amount of error due to the discretization of traffic, CTP and L, to integer values. In reality, the exact values for the traffic proportions for each protocol CTP_i can be fractional values. The error margin is $\sim 1\%$ of CTP since the actual value is at most some fraction of a percent above the value found by the tool. We also test the F_{ab} solution for one protocol just below the threshold and note that the solution ($S_{cir} = A$ and $S_{cen} = 100$) remains stable (*i.e.*, stays the same) as we approach the threshold and then switches to another strategy ($S_{cir} = B$ and $S_{cen} = 100$) as we cross it. This gives confidence that the tool follows the expected behavior.

Our model assumes that the players have perfect information; *i.e.*, know the values for F_{ab}, CTP_{max}, and CTP, and the protocol traffic distribution R. The values for CTP_{max} and CTP can be known by monitoring the network to record how much traffic flows over its nodes and also to gauge its capacity. The Tor network is an example of a CRS system where these values can be known by both players since they are made publicly available. The protocol traffic distribution can be learned by surveys that are routinely conducted by entities such as Sandvine, whose report we use in our analysis. However, F_{ab} is more difficult since the censor does not explicitly and publicly provide this information. Nonetheless, we may be able to infer this value by observing the censor's behavior (S_{cen}). Assuming that it does not impact the censor's utility function, we may be able to probe the censor by first assuming its type, then playing the optimal strategy, and then observing its response. Since the other values are known, by observing the censor the circumventor can discover if it is over- or underestimating F_{ab}.

Our model assumes that the CRS continues to send traffic over protocols that have been blocked by the censor (as a consequence of $CTP > F_{ab}$). This seems counterintuitive since the circumventor would decide to stop wasting that traffic since it will be blocked anyway. However, the consequence is worse than the wastage of that traffic. As soon as the protocol is no longer used, the extra surplus traffic has to go over the remaining protocols, which would lead to some or all of them being blocked. This outcome of no traffic getting through is worse than some traffic getting through with some wastage. Also, the censor would stop blocking the protocol to reduce its costs; however, we would then want to reuse the protocol. This leads to the same series of steps where we would redistribute the traffic over all of the protocols for the optimum solution, with the aforementioned protocol being blocked.

8 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, 25, 31] 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 [10] 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.* [35] 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. Their 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.

9 Future Work

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 [29] or poachers [14] 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 develop a methodology for identifying the censor's type, to learn the value of F_{ab}, that is rooted in empirical data. One challenge we foresee for 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 [8, 9, 16, 40] can be a useful source of data for our framework. One of the results of our analysis is that protocol usage coordination is necessary in order to achieve the optimum solution in the $CTP > F_{ab}$ setting. This assumes that there is some way to achieve this in real-world CRSs. Unfortunately, most current CRSs are not designed this way, focusing on one protocol at a time. However, there are systems like StegoTorus [39] that have the capability to split traffic over multiple protocols, which would fit our purpose. It is an open problem to develop a mechanism to manage circumvention traffic such that it 1) does not cross the F_{ab} threshold, 2) is distributed across protocols according to the optimum solution, and 3) is usable in realistic settings.

10 Conclusion

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 an automated numerical analysis 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 automated numerical analysis as an aid to discovering and analyzing equilibrium points. This approach has application to real-world CRS-design problems, namely, of how to select cover protocols effectively and how to distribute circumvention traffic over them to maximize utility even in sub-optimal scenarios.

Acknowledgement: We appreciate the valuable feedback we received from the anonymous reviewers, and especially our shepherd Michael C. Tschantz, whose comments greatly improved this paper. Steven J. Murdoch was supported by The Royal Society [grant number UF110392]. Ian Goldberg thanks NSERC for grant RGPIN-341529. NSERC provided support for Tariq Elahi, John A. Doucette, and Hadi Hosseini. In addition, John A. Doucette was supported by OGS and Tariq Elahi was supported by the Research Council KU Leuven: C16/15/058, and the European Commission through H2020-DS-2014-653497 PANORAMIX.

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] D. Blackwell. Discounted dynamic programming. *The Annals of Mathematical Statistics*, 36(1):226–235, 1965.

[6] F. Brandt, T. Sandholm, and Y. Shoham. Spiteful bidding in sealed-bid auctions. In *Proceedings of the 20th International Joint Conference on Artifical Intelligence*, IJCAI'07, pages 1207–1214, San Francisco, CA, USA, 2007. Morgan Kaufmann Publishers Inc.

[7] 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.

[8] 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.

[9] 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.

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

[11] 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.

[12] R. Dingledine, N. Mathewson, and P. Syverson. Tor: The Second-Generation Onion Router. In *Proceedings of the 13th conference on USENIX Security Symposium-Volume 13*, pages 303–320. USENIX Association, 2004.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

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

[18] 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.

[19] 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.

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

[21] S. Khattak, T. Elahi, L. Simon, C. M. Swanson, S. J. Murdoch, and I. Goldberg. SoK: Making Sense of Censorship Resistance Systems. *Proceedings on Privacy Enhancing Technologies*, 2016(4), 2016.

[22] 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.

[23] K. Leyton-Brown and Y. Shoham. Essentials of Game Theory: A Concise Multidisciplinary Introduction. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 2(1):1–88, 2008.

[24] 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.

[25] 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.

[26] 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.

[27] J. Morgan, K. Steiglitz, and G. Reis. The spite motive and equilibrium behavior in auctions. *Contributions in Economic Analysis & Policy*, 2(1), 2003.

[28] M. J. Osborne. *An introduction to game theory*. Oxford University Press New York, 2003.

[29] 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.

[30] Psiphon Inc. Psiphon. https://psiphon.ca. Retrieved May 2015.

[31] 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.

[32] Sandvine. Global Internet Phenomena Report - Spotlight encrypted Internet traffic. https://www.sandvine.com/downloads/general/global-internet-phenomena/2016/global-internet-phenomena-spotlight-encrypted-internet-traffic.pdf.

[33] Y. Shoham and K. Leyton-Brown. *Multiagent Systems: Algorithmic, Game-Theoretic, and Logical Foundations*. Cambridge University Press, 2008.

[34] T. Tor Project. Tor Mertics Portal: Bridge users by country. https://metrics.torproject.org/userstats-bridge-country.html, 2016. Retrieved May 2016.

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

[36] 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.

[37] VPN Gate. VPN Gate Latest Activity Logs. http://www.vpngate.net/en/lastlog.aspx, 2016. Retrieved May 2016.

[38] 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.

[39] Z. Weinberg, J. Wang, V. Yegneswaran, L. Briesemeister, S. Cheung, F. Wang, and D. Boneh. StegoTorus: A Camouflage Proxy for the Tor Anonymity System. In *Proceedings of the 19th ACM conference on Computer and Communications Security*, October 2012.

[40] 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.

A Numerical analyzer

Our analyzer models the game setting from subsection 4.2, but with multiple protocols and the utility functions 14 and 15 instead. We assume that the censor does not have an apparatus that can distinguish non-circumvention 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 case of an apparatus-enabled censor is a straightforward extension, where each protocol in use is split into two: the traffic flagged by the apparatus is treated as one protocol, and that unflagged is treated as a second protocol.

The analyzer conducts a brute-force search for the optimum censor and circumventor strategies by iterating over the entire strategy space. For each circumvention strategy—*i.e.*, distribution of circumvention traffic across the protocols—the analyzer chooses the censor strategy—*i.e.*, the set of protocols to block—with the highest utility for the *censor*. Then from the entire list of (circumvention strategy, chosen censor strategy) pairs, the analyzer chooses the one that results in the highest utility for the *circumventor*. This will be the equilibrium strategy since if either party changes their strategy, 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 analyzer 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 the traffic-volume data from the December 2015 Sandvine survey mentioned in subsection 6.2.

Our analyzer makes some simplifying assumptions to reduce the computational complexity of calculating the censor's utility for all combinations of censor and circumventor strategies. The censor can chose to block any selection of the n target protocols, resulting in 2^n strategies. In the analysis that follows we model up to six target protocols which means that there are at most $2^6 = 64$ censor strategies. The circumventor can choose to send units of traffic in any distribution over the protocols, but there are still an infinite number of circumventor strategies if we allow any fractional value for the amount of traffic. So, to reduce the strategy space we quantize all circumvention traffic into multiples of one percent point up to a total of one hundred percent points, resulting in at most 2961 circumventor strategies. To clarify, one percent point of circumvention traffic is equivalent in quantity to 1% of CTP. We model a network where the sum of all cover traffic together is L, and like before $L + CTP = 1$, which is the whole of the network traffic.