The Sound of Silence: Anti-Defamation Law and Political Corruption

Gabriele Gratton*
Boston University

April 2011

Abstract

I study the role of mass media in deterring politicians' corruption under the assumption that even a non-corrupt politician can be indicated by the media as involved in a corruption scandal. It is counter-productive to have an anti-defamation law so stringent that in equilibrium at least one allegation is not worth publishing even if it were true. In this case, (*i*) corruption is larger than without any anti-defamation protection and (*ii*) a limit to political corruption relies on the possibility of punishing the politician when no allegation is mentioned in the press.

Keywords: media and democracy; corruption; defamation; chilling effect.

JEL Classification Numbers: D73, D80, K42.

[*] Department of Economics, Boston University, 270 Bay State Road, Boston, MA 02215, USA (email: gratton@bu.edu). I am grateful to Andy Newman, Sambuddha Ghosh and Laurent Bouton for their constructive advice. Thanks to Dilip Mookherjee for insightful comments and to Aleksandra Balyanova for continuous support. I have also profited from discussions with Alberto Motta, Alfredo Burlando, Guillem Riambau Armet and Andrea Canidio. Thanks to all participants to Boston University Political Economy Research Group.

...tragedy begins not when there is a misunderstanding about words, but when silence is misunderstood.

Thoreau (1980), p. 278.

1 Introduction

We read newspaper to know what our politicians are doing. We often use this information to decide whether punish a politician well before the allegation can be verified. Yet, were the press free to publish any allegation, it would be licit to expect much of what we read every day to be false.

To avoid relying on false information, anti-defamation and libel laws give media the incentives not to publish evidence of politicians' wrongdoings without a solid ground. Yet, instead of more true information, these laws might reduce the amount of true information reaching the electorate if the media fear being punished even for publishing a true scandal. This is an unintended result of anti-defamation laws known as *chilling effect* (see, among others, Barendt et al. (1997)).

In this paper I study the trade-off between these two effects of anti-defamation laws. Does a more stringent anti-defamation law only reduce the publication of false scandal or it also promts the press to conceal true information to the electorate? If the latter case is true, how should the electorate react to the silence of the media? In the model presented in this paper, a principal commits to reward the politician with a given probability, conditional on the scandal published by a media firm. The firm intimately knows how corrupt the politician is, but can only publish evidence of scandals it observes. The analysis shows that any anti-defamation law is counter-productive if there is at least one true scandal that the firm would prefer not to publish. When this is the case, (i) the equilibrium level of corruption is larger than without any anti-defamation protection and (ii) any limit of political corruption relies on the possibility of punishing the politician when no allegation is mentioned in the press.

Despite the role played by the threat of a chilling effect, in equilbirum the media always reveal all true scandals. Some chilling effect, that is, some true scandal not being published, can arise in equilibrium only if the principal is not capable of punishing the politician when the media remain silent: if silence is misinterpreted, i.e. the politician is not punished when the media remain silent about his behavior, then corruption is always maximal and some true scandals are not published.

A vast literature in recent years has explored the role played by mass media in enhancing the effectiveness of democratic institutions in detecting and punishing corruption (among others, Besley and Burgess (2002), Besley and Prat (2006), Ferraz and Finan (2008), Garoupa (1999b), and Suphachalasai (2005)). The usual argument is that mass media reduce agency costs in the process of delegation of power from the electorate to the elected officials. Newspapers and broadcasting firms have incentives to discover political corruption, since the publication of such scandals increases their sales, and in doing so they provide a stream of precious information to the electorate which would not be otherwise available. The promise of a future rent in case of re-election motivates politicians to avoid corrupt behaviors.

The focus of this paper departs from the argument mentioned above[1] by adding two additional assumptions to the description of the role of mass media, namely (*i*) all (with at most few exceptions) politicians, even if not corrupted at all, are subject to the risk of being accused by mass media of being involved in some scandal during their mandate and (*ii*) the real nature of the accusation being true or false is often revealed only after a political decision on their status has been taken (such as new elections or a process of impeachment). When these two assumptions are taken into consideration, the question of how the electorate (or the political body that has to decide whether to impeach the politician) should interpret the scandals published on newspapers and tabloids hinges on which incentives are in place to motivate media firms to select for publication only evidence of scandals that have a solid foundation.

The arguments in this paper are closely related to those of Garoupa (1999b,a). The model in these articles predicts that increasing the probability of the politician winning the legal dispute with the firm is not a problem if the media are able to distinguish honesty from dishonesty. One of the key results of my model is that even if the media firm knows exactly how much the politician is corrupted, when there exists some probability of the politician winning the dispute even if the media's allegation is true, then a stringent enough anti-defamation legislation can actually increase corruption. Furthermore, while Garoupa's (1999b) media can always accuse the politician of dishonesty, independently of what information they possess, in my model they are limited to publish the scandal they observe (a picture, a videotape, the interview of a witness, et cetera). This allows for a comparison of the anti-defamation law with a scenario in which no anti-defamation protection is in place.

My model is also related to the work of Besley and Prat (2006), which explores the possibility of the government capturing the media, therefore limiting their ability to transfer information to the electorate. Both their model and mine draw from the vast principal-agent-supervisor literature (e.g. Antle, 1984; Tirole, 1986; Kofman and Lawarree, 1993). Most of this literature focuses on the nature of contracts capable of deterring collusion between the agent and the supervisor against the interests of the principal. My model excludes this

[1] The model of Garoupa (1999b) and Garoupa (1999a) presents similar assumptions, although the tratment of these issues, and so some of the results, are different.

possibility and analyzes the role of fully independent media. The main difference between the model in this paper and those mentioned above is that in my model the principal can only establish a contract with the agent (whether to reelect or impeach) while the incentives given to the media are already in place in the form of a revenue function and a punishment technology (the anti-defamation legislation).

The anti-defamation legislation is only one of many characteristics of a society that affects its ability to provide the electorate with sufficient information on the governors' behavior. Suphachalasai (2005) finds that both the freedom of the media and competition reduce corruption and his empirical results seem to suggest that competition might be more important than press freedom. Besley and Prat (2006) and Djankov et al. (2001) find an important empirical relation between media ownership and a wide range of political outcome measures. Finally, Besley and Burgess (2002) highlight the importance of free and independent regional press.

There is no lack of examples of defamed politicians whose image (and often the entire career) is compromised by accusations that are later reviewed as false. Garoupa (1999b)mentions the episode involving the then Irish Prime Minister and leader of Fianna Fàil, Albert Reynold. In November 1994, the *Sunday Times* reported that he had lied to the Parliament and to his coalition partners on the appointment of his Attorney General as President of the High Court. In only two weeks, the Labour Party abandoned the government coalition and Reynolds resigned. Garoupa (1999b) reports that Reynolds later sued the *Sunday Times* for libel and was awarded the sum of 1£ two years later.

On June 15, 1978, President Giovanni Leone of Italy resigned as President of the Republic––a unique case in the history of the Italian Republic–after newspapers and political oppositors accused him of being involved in the scandal regarding bribes paid by officials of the U.S. aerospace company Lockheed to Italian politicians. Among his most fierce accusants were the Radical Party members of parliament Marco Pannella and Emma Bonino. President Leone was never condemned for the above allegations and–significantly–Mr. Pannella and Ms. Bonino published an open letter of apologies in most national newspapers on the occasion of Leone's ninetieth birthday in 1998[2].

Many examples of chilling effect are collected in Barendt et al. (1997). I mention here a late scandal involving Lazio's (Italy) Governor Piero Marrazzo between the summer and the fall of 2009. Governor Marrazzo was blackmailed by four police officers in possession of a compromising videotape portraying Marrazzo's involvement with a transgender prostitute. During the late summer and in the early fall of 2009, the police officers had repeatedly tried to sell the video to newspapers and televisions, but could not find a buyer. The scandal is

[2] Corriere della Sera, November 3 1998, p. 35.

nonetheless capable of capturing the media's and public's attention: on October 23, most national newspapers report of the existence of a police investigation on four police officers blackmailing Governor Marrazzo. The fact that at this point Marrazzo's videotape was subject of an official investigation meant the press would have not been sued for defamation anymore and indeed details about the videotape appeared on most national newspapers in the subsequent days.

The remainder of the paper is as follows: section 2 presents a simple example introducing the basic arguments behind the main results of the paper. In section 3 I present the benchmark model. Section 4 contains the main results. Section 5 discusses a different specification of the model of section 3, namely one in which the punishment of the media depends on the distance between the alleged wrongdoing of the politician and the judges' evaluation of the actual gravity of his behavior. Section 6 concludes.

2 A simple example

Before turning to the benchmark model in section 3, I discuss here a simplified story suggestive of some of the results I obtain in the more general model.

There is a principal, whose objective is to minimize the corruption of a politician currently in office. The principal can offer the politician a reward $r = 3/4$, contingent on the gravity of the scandal published by a media firm. The politician chooses the maximum level of corruption $c \in \mathcal{C} := \left\{0, \frac{1}{2}, 1\right\}$, representing the greatest wrong the politician is going to commit during his mandate. The direct payoff for the politician from his corruption is equal to the level of corruption chosen c.

The media firm observes the true level of corruption of the politician, c. In addition, the firm observes evidence of a scandal s. With probability $1/2$ the scandal is equal to the maximum level of corruption chosen by the politician, $s = c$, otherwise s is uniformly distributed on $\mathcal{C}$. If $s \leq c$, then the scandal is *true*, in the sense that it represents a wrong actually committed by the politician.

The media firm can only choose whether to publish the scandal s: it can send a message $x = s$ to the principal, or $x = \phi$, representing the message 'silence'. The firm's revenues π are increasing in the gravity of the published scandal with $\pi(\phi) = 0 < \pi(c)$ for all $c \in \mathcal{C}$.

The principal commits *a priori* to a mechanism $e(x) : \{\mathcal{C}, \phi\} \to [0, 1]$, where $e(x)$ represents the probability of rewarding the politician with the rent r if a scandal x (including the signal 'silence') is reported by the media.

After the random process $e(x)$ is realized, the media firm is judged for defamation. The firm is punished with probability ζ if the published scandal is false, i.e. if $x > c$, or with probability $\xi < \zeta$ otherwise. The punishment is a function $\rho(x) : \mathcal{C} \to \mathbb{R}_+$ defined by some anti-defamation law. There exist two anti-defamation laws, $\rho_l(\cdot)$ and $\rho_h(\cdot)$, such that

$$\zeta \rho_l \left(\frac{1}{2}\right) < \pi \left(\frac{1}{2}\right) \tag{1}$$

$$\xi \rho_l (1) < \pi (1) < \zeta \rho_l (1) \tag{2}$$

$$\xi \rho_h \left(\frac{1}{2}\right) < \pi \left(\frac{1}{2}\right) < \zeta \rho_h \left(\frac{1}{2}\right) \tag{3}$$

$$\pi (1) < \zeta \rho_h (1) . \tag{4}$$

The conditions in (1) and (2) state that $\rho_l(\cdot)$ is a law stringent enough to limit defamation at least in some case, meaning that a scandal $s = 1$ would not be worth publishing by a firm if it was false (if $c < 1$). Nevertheless, in no case under $\rho_l(\cdot)$ the firm would not publish a true scandal. Under the more stringent law $\rho_h(\cdot)$, though defamation is more limited, in the sense that any false scandal would never be published (from (3))[3], the condition in (4) guarantees the existence of at least one scandal (and indeed only one) that the firm would not find worth publishing even if true. Hence, in this case there is the potential for some chilling effect.

Suppose that there is no anti-defamation law. In this case the firm will publish all scandals since it faces no risk of being condemned for defamation. Suppose the principal can induce a level of corruption $c = 1/2$. In this case an incentive compatibility constraint (ICC) must hold such that

$$\frac{1}{2} + \frac{r}{2} e\left(\frac{1}{2}\right) + \frac{r}{6}\left[e(0) + e\left(\frac{1}{2}\right) + e(1)\right] \geq 1 + \frac{r}{2} e(1) + \frac{r}{6}\left[e(0) + e\left(\frac{1}{2}\right) + e(1)\right]$$

$$\iff \frac{1}{2} + \frac{r}{2} e\left(\frac{1}{2}\right) \geq 1 + \frac{r}{2} e(1) . \tag{5}$$

The optimal mechanism therefore requires $e(1/2) = 1$ and $e(1) = 0$, and condition (5) becomes $r \geq 1$. We can conclude that, in the absence of any anti-defamation law, $c = 1$.

Consider now the anti-defamation law $\rho_l(\cdot)$. Under this law, the media will publish all true scandals and would send the message $x = \phi$ if and only if $s = 1 > c$. It is easy to verify that no level of corruption less than 1/2 can be sustained in equilibrium (indeed it would require a rent $r \geq 3/2$). Suppose instead that there exists a mechanism capable of inducing

[3] Notice that $s = 0$ can never be false since $0 \leq c$ for all $c \in \mathcal{C}$

a level of corruption $c = 1/2$. This mechanism must satisfy

$$\frac{1}{2}+\frac{r}{2}e\left(\frac{1}{2}\right)+\frac{r}{6}\left[e\left(0\right)+e\left(\frac{1}{2}\right)+e\left(\phi\right)\right]\geq 1+\frac{r}{2}e\left(1\right)+\frac{r}{6}\left[e\left(0\right)+e\left(\frac{1}{2}\right)+e\left(1\right)\right]$$
$$\iff \frac{1}{2}+\frac{r}{2}e\left(\frac{1}{2}\right)+\frac{r}{6}e\left(\phi\right)\geq 1+\frac{3}{2}re\left(1\right). \tag{6}$$

The optimal mechanism therefore requires $e\left(1/2\right) = e\left(\phi\right) = 1$ and $e\left(1\right) = 0$, and condition (6) becomes $r \geq 3/4$. We can conclude that, under the law $\rho_l\left(\cdot\right)$, $c = 1/2$. The introduction of the moderately stringent law $\rho_l\left(\cdot\right)$ decreases the equilibrium level of corruption.

Consider instead what would happen if the more stringent anti-defamation law $\rho_h\left(\cdot\right)$ was in place. Under this law, a true scandal $s = c = 1$ would not be published. Again, suppose that there exists a mechanism such that a level of corruption $c = 1/2$ is sustainable in equilibrium. Then the mechanism must satisfy

$$\frac{1}{2}+\frac{r}{2}e\left(\frac{1}{2}\right)+\frac{r}{6}\left[e\left(0\right)+e\left(\frac{1}{2}\right)+e\left(\phi\right)\right]\geq 1+\frac{r}{2}e\left(\phi\right)+\frac{r}{6}\left[e\left(0\right)+e\left(\frac{1}{2}\right)+e\left(\phi\right)\right]$$
$$\iff \frac{1}{2}+\frac{r}{2}e\left(\frac{1}{2}\right)\geq 1+\frac{3}{2}re\left(\phi\right). \tag{7}$$

Hence, the optimal mechanism requires $e\left(1/2\right) = 1$ and $e\left(\phi\right) = 0$, the last condition meaning that the silence of the media is interpreted by the principal as a sign that the politician is very corrupted, but the media's message is affected by the chilling effect. In fact, under the optimal mechanism, the principal would never reward the politician if the media remain silent. Under this optimal rule, the condition in (7) becomes $r \geq 1$. Hence, the equilibrum level of corruption under the more stringent law is $c = 1$. The implication of this result is that there exists an interior optimum for the level of stringency of the anti-defamation law. If the law is too stringent, then corruption is maximal as in the case of no anti-defamation law. The difference being that the principal ceases to believe that the silence of the media indicates that the politician is not corrupt.

3 The benchmark model

There is a principal whose objective is to minimize the corruption of a politician currently in office. The principal can offer the politician a reward r, contingent on the gravity of the scandal published by a media firm.

The risk-neutral politician currently in office chooses the maximum level of his corruption $c \in [0,1]$, representing the greatest wrong the politician is going to commit during his mandate. This means that the politician will commit all acts between 0 and c. The direct payoff for the politician from his corruption is γc, where $\gamma > 0$.

The media firm observes the true level of corruption of the politician, c. In addition, the firm observes evidence of a scandal $s \in [0,1]$. With probability $q > 0$, the scandal is the greatest wrong committed by the politician, hence $s = c$; otherwise s is uniformly distributed in the interval $[0,1]$ ($s \sim U(0,1)$). Notice that if $s \leq c$, then the scandal is *true*, in the sense that it represents a wrong actually committed by the politician.

The media firm can only publish the scandal s: the observation of c is private information of the firm and cannot be directly published. Therefore, the choice of the firm is whether to publish the observed scandal or to send a message 'nothing' to the principal. The action set of the firm observing scandal s is $\mathcal{X}_s = \{s, \phi\}$, where ϕ represents the message 'silence'. The firm's revenues are an increasing, concave and twice differentiable function of its message x, $\pi(x) : \{[0,1], \phi\} \to \mathbb{R}_+$, with $0 = \pi(\phi) < \pi(0)$.

The principal commits *a priori* to a mechanism $e(x) : \{[0,1], \phi\} \to [0,1]$ where $e(x)$ represents the probability of rewarding the politician with a rent $r \leq \gamma$ if scandal x (including the signal 'silence') is reported by the media. The reward can be interpreted as the reelection of the politician for a further mandate (in which case the principal represents the electoral body) or the decision of a political institution–with control power over the politician–not to impeach the politician. The assumption that the rent r is less or equal to γ can be interpreted as 'the discounted rent associated with a new mandate is not larger than the payoff generated by a maximally corrupted behavior in a single mandate'.

After the random process generated by the probability $e(x)$ is realized and the the politician has collected the eventual rent, the politician has the possibility of suing the media firm for defamation. If the politician wins the dispute, the firm is then condemned to a punishment $\rho(x) : [0,1] \to \mathbb{R}_+$, with $\rho'(\cdot) > 0$, $\rho''(\cdot) \geq 0$ and $\rho(0) = 0$[4]. The resolution of the trial is based on the judges' assessment of the level of corruption of the politician, g, and the dispute is resolved in favor of the politician if and only if $g < x$. If the dispute is resolved in his favor, a politician who has collected his rent r is accorded by the judges a (possibly symbolic) indemnification $\varepsilon \in (0, r)$; if the politician has instead received no reward from the principal, the judges would consider that the publication of the false scandal has damaged the politician and he will be accorded a reparation $\delta \in (\varepsilon, r]$.

The politician has access to two kinds of trial: at no cost for the politician, the firm can be brought to court to a *fair* trial in which $g = c$ with probability 1; at a cost $f > \varepsilon$, the politician has access to a *biased* trial for which, with probability $1 - \zeta$, $g = c$, and with

[4] In section 5, I present an alternative specification of the anti-defamation law for which the punishment depends on the distance between the published scandal x and the level of corruption assessed by the judges.

probability $\zeta > 0$, $g \sim U(0,1)$.

The following analysis concentrates on the study of perfect Bayesian equilibria (PBE) of this model. I show that for each anti-defamation law $\rho(x)$ there exists a unique level of corruption c sustained by all optimal mechanisms $e(x)$ and I provide some comparative statics of the relation between a measure of the stringency of the anti-defamation law and the equilibrium level of corruption.

In the introduction of this paper I have discussed the importance given in the literature to the trade-off between defamation and chilling effect. I provide here two definitions I will use in the following analysis.

Definition 1. An *equilibrium with defamation* is a PBE such that $x(s) = s$ for some $s > c$.

An *equilibrium with chilling* is a PBE such that $x(s) = \phi$ for some $s \leq c$.

In definition 1, an equilibrium with defamation is defined as a PBE for which there is some false scandal that is published. An equilibrium with chilling, on the other hand, is defined as a PBE for which there is some true scandal that is not published. It is worth noticing that these two definitions are neither mutually exclusive nor exhaustive of the set of strategy profiles.

Before attempting the analysis of the model, I state here a preliminary result for the purpose of comparison. Imagine shutting down all anti-defamation protection, i.e. suppose that defamation is not punishable and that media are free to publish any evidence of a scandal. In this case, the principal will always observe the scandal s. In this case, incentive compatibility bounds below the amount of corruption sustainable in a PBE.

Proposition 1. *Without any anti-defamation law, all PBEs of the model are characterized by a level of corruption $\underline{c} = 1 - rq/\gamma$ such that $e(\underline{c}) = 1$ and $e(c) = 0$ for all $c > \underline{c}$. All PBEs are equilibria with defamation and there is no equilibrium with chilling.*

Proof. Omitted. □

3.1 Trial stage

The probability of the politician winning the biased trial if a true scandal $x = s$ has been published by the firm is ζs, therefore a politician will sue a firm publishing a true scandal if and only if

$$s > \frac{f}{\delta\zeta} \tag{8}$$

and the politician has received no rent.

Suppose $f \geq \delta\zeta$, then the politician will never incur the cost f to accede to the biased trial. Said otherwise, no biased trial will occur in any PBE. In the remainder of the paper I will assume the condition in (8) to hold and concentrate my analysis to this situation of *imperfect justice*. When $f \geq \delta\zeta$ more stringent anti-defamation laws never increase the equilibrium level of corruption.[5].

3.2 Media strategies and chilling neutralizing mechanisms

This section analyzes the impact of anti-defamation laws on media strategies. Consider the case in which the media firm observes a false scandal ($s > c$). Define

$$\bar{s} := \begin{cases} 1 & \text{if } \pi(s) > \rho(s), \forall s \in [0,1]; \\ s \in [0,1] : \pi(s) = \rho(s) & \text{otherwise;} \end{cases}$$

then $\bar{s}$ represents the threshold scandal such that a firm would publish any scandal less or equal to it, independently of it being true or false. The assumptions made on $\pi(x)$ and $\rho(x)$ guarantee the unicity of the threshold $\bar{s}$. I will use the threshold level $\bar{s}$ as a measure of the intended stringency of the anti-defamation law. A lower $\bar{s}$ corresponds to a more stringent law in the sense that there exist a smaller set of scandals that would be published by the firm if false.

Consider now the case of a scandal s such that $s > f/\delta\zeta$. In this case, the media firm can be successfully sued by the politician even if the scandal is true, that is, even if $s \leq c$. Therefore, an anti-defamation law affects the firm's choice according to two thresholds: $\bar{s}$, representing the largest scandal that would be published even if false, and $\tilde{s}(e(\cdot)) = \max\{s(e(\cdot)), f/\delta\zeta\}$, representing the largest scandal that would be published by the firm if the scandal is true, where

$$s(e(\cdot)) := \begin{cases} 1 & \text{if } \pi(s) > s\zeta(1-e(s))\rho(s), \forall s \in [0,1]; \\ s \in [0,1] : \pi(s) = s\zeta(1-e(s))\rho(s) & \text{otherwise.} \end{cases}$$

Suppose, in fact, that the firm observes a scandal $s \leq c$. If the scandal is published, the revenue of the firm is $\pi(s)$. If $s > f/\delta\zeta$, with probability $(1-e(s))$ the politician will not be rewarded by the principal and will sue the firm. The dispute will be resolved in favor of the politician with probability $s\zeta$. The expected punishment of the firm resulting from publishing the scandal is therefore $s\zeta(1-e(s))\rho(s)$. For the revenue of the firm to be larger

[5] See appendix B not for publication

than the expected punishment we must have

$$\begin{aligned}\pi(s) &\geq s\zeta(1-e(s))\rho(s)\\ e(s) &\geq 1-\frac{\pi(s)}{s\zeta\rho(s)}.\end{aligned}$$

Lemma 1. *Suppose that $e(x)$ is a non-increasing function of x, then $s(e(\cdot))$ is unique.*

Proof. See appendix A.1. □

Lemma 1 guarantees unicity of $s(e(\cdot))$ only for $e(x)$ being a non-increasing function of x. The remainder of the analysis in this section assumes that this condition holds.

Condition 1. $e(x)$ is a non-increasing function of x.

It is relevant to notice that $\tilde{s}(e(\cdot)) > \bar{s}$ since $s(e(\cdot)) > \bar{s}$ by construction. The interpretation of this fact is simply that if the net payoff of publishing a scandal is high enough that the firm would publish it even if it was false, then the same scandal would be published if it was true.

It is easy to see that if $\tilde{s}(e(\cdot)) \geq 1$ then there is no true signal that will not be published by the firm. In all these cases the strategy of the firm is therefore the same as in lemma 8. More precisely, suppose

$$\hat{e} := 1-\frac{\pi(1)}{\zeta\rho(1)} \leq 0$$

then there exists no true scandal that would not be published, independently of the mechanism $e(x)$ in place (in other terms, $\tilde{s}(e(\cdot)) \geq 1$ for all $e(x)$).

The following lemma defines the optimal strategy of a firm in the general case.

Lemma 2. *Define $\hat{s}(e(\cdot)) := \min\{\tilde{s}(e(\cdot)), \max\{\bar{s}, c\}\}$, the optimal strategy for a media firm observing a scandal s and corruption level c is*

$$x(s,c) = \begin{cases} s & \textit{if } s \leq \hat{s}(e(\cdot)); \\ \phi & \textit{otherwise.}\end{cases}$$

Proof. Omitted. □

It is important to notice that there are in this case two relevant measures of the stringency of the anti-defamation law. One, namely $\bar{s}$, represents the *intended* stringency of the law, i.e. a limit to defamation, since no scandal larger than $\bar{s}$ would be published if false. The other, $\bar{e} = \max\{\hat{e}, 0\}$, representing the *collateral* effect of posing a limit to the freedom of speech. Indeed, whenever $e(1) < \bar{e}$, $x(1) = \phi$, meaning that a perfectly corrupted politician will have some of his true scandals not published. Whenever $e(1) < \bar{e}$, in fact, there is always, at least in power, some chilling effect. If the principal wants to eliminate this collateral and possibly deleterious effect of the anti-defamation legislation, then the mechanism design must reward a politician accused of a maximum scandal ($x = 1$) with a probability at least as large as $\bar{e}$. The following is a formal definition of this class of mechanisms.

Definition 2. A mechanism is *chilling neutralizing* (all true scandals are worth being published) if $e(1) \geq \bar{e}$.

3.3 Mechanism design

The objective of the principal is to minimize the level of corruption c chosen by the politician. Define

$$R^{e(\cdot)}(a,b) := r\left\{ qe(a) + (1-q)\left[\int_0^b e(z)\,dz + (1-b)\,e(\phi) \right] \right\}$$

then $R^{e(\cdot)}(c, \max\{c, \bar{s}\})$ is the politician's expected reward given the principal's mechanism design when the politician is corrupted up to level c: with probability q the firm will observe the scandal $s = c$ and publish it; with probability $(1-q)$ the scandal will be distributed between 0 and 1 and the firm will publish only those scandals that are either true or less or equal to the threshold level $\bar{s}$.

Also, define the expected payoff from reparations in case of defamation as

$$D^{e(\cdot)}(c, \max\{c, \bar{s}\}):$$

$$D^{e(\cdot)}(a,b) := d(a)(1-q)\int_a^b [\delta - e(z)(\delta - \varepsilon)]\,dz$$

where $d(c) = 1$ if $\bar{s} > c$ and $d(c) = 0$ otherwise.

Then,

$$U(c, e(x)) := \gamma c + R^{e(\cdot)}(c, \hat{s}(e(x))) + D^{e(\cdot)}(c, \bar{s}) + \eta_c + \chi_c^{e(\cdot)}$$

where,

$$\eta_c = \begin{cases} q(c\delta\zeta - f)(1 - e(c)) & \text{if } c \in \left(\frac{f}{\delta\zeta}, s(e(\cdot))\right]; \\ 0 & \text{otherwise;} \end{cases}$$

and

$$\chi_c^{e(\cdot)} = \begin{cases} (1-q) \int\limits_{\frac{f}{\delta\zeta}}^{\min\{s(e(\cdot)),c\}} (z\delta\zeta - f)(1-e(z))\,dz & \text{if } \min\{s(e(\cdot)),c\} > \frac{f}{\delta\zeta}; \\ 0 & \text{otherwise;} \end{cases}$$

is the expected payoff of a politician choosing c under the mechanism $e(x)$. The term $\eta_c + \chi_c^{e(\cdot)}$ is the expected reparation for the politician from a biased trial given the principal's mechanism $e(x)$. The problem for the principal is to choose a mechanism $e(x)$ such that the ICC

$$U(\underline{c}, e(x)) \geq U(c, e(x))$$

holds for all $c > \underline{c}$ and there is no $\underline{c}' < \underline{c}$ such that the ICC holds.

4 Main results

In the previous analysis I have characterized the optimal strategy for the media firm and for the principal under all possible anti-defamation laws $\rho(\cdot)$. The results presented in this section characterize the set of PBEs for all anti-defamation laws $\rho(\cdot)$ as measured by their threshold level $\bar{s}$. Furhtermore, I provide with some comparative statics. To facilitate the exposition of these results, I first state here the main proposition of this section, condensing the most salient characteristics of the set of PBE for different anti-defamation law. I then turn to the analysis of the specific cases.

Proposition 2. *There exists a non-empty and* bounded *set of combinations of $\bar{s}$ and $\bar{e}$ for which the PBE mechanism $e(x)$ is chilling neutralizing. For low enough values of $\bar{s}$, the PBE mechanism $e(x)$ is not chilling neutralizing and there exists a unique PBE such that the equilibrium level of corruption is larger than if there was no anti-defamation protection.*

Proof. Follows from lemmata 3, 4 and 5. □

Consider the first part of proposition 2. According to this result, there exist a set of combinations $(\bar{s}, \bar{e})$ such that the equilibrium mechanism is chilling neutralizing. Indeed it is possible to individuate necessary conditions for this to happen. Formally:

Lemma 3. *There exists a non-empty and* bounded *set of combinations of $\bar{s}$ and $\bar{e}$ for which the PBE mechanism $e(x)$ is chilling neutralizing. In particular, necessary conditions for these equilibria to exist are*

1. $\bar{s} > 1 - \frac{rq-\eta^0}{\gamma-r(1-q)} \geq 1 - \frac{rq}{\gamma-r(1-q)}$ and $\bar{e} \leq \frac{\theta(\bar{s},r,\delta)-\eta^0}{\theta(\bar{s},r,\delta)-\eta^0+\varepsilon(1-q)(1-\bar{s})} \in \left(\frac{\theta(\bar{s},r,\delta)-\eta^0}{\theta(\bar{s},r,\delta)-\eta^0+\varepsilon(1-q)}, 1\right)$, in which case the PBE is an equilibrium with defamation, or

   (a) $\bar{e} \leq W := \frac{r(1-q)}{\gamma-r(1-q)} \Rightarrow \bar{s} \geq \frac{\pi(0)}{\pi(1)} \frac{(1-W)\zeta}{1+(1-W)\zeta} > 0$.

Proof. See appendix A.2. □

The interpretation lemma 3 presents particular difficulties. It is possible for appropriate parameters of the model to individuate combinations $(\bar{e}, \bar{s}, \underline{c})$ for which an equilibrium with chilling neutralizing mechanism exists. Furthermore, when these equilibria exist, a level of corruption $\underline{c} \leq \bar{s}$ can be sustained in equilibrium, i.e. we have an equilibrium with defamation. It is interesting to notice that assuming that these combinations exist for a continuum of values of $\bar{s}$, then a reduction of $\bar{s}$ (an increase in the intended stringency of the anti-defamation law) can both reduce or increase the level of corruption $\underline{c}$ depending on the level of the unintended anti-defamation law $\bar{e}$ being less than or larger than $(r-\delta)/(r-\delta+\varepsilon)$. For low enough values of $\bar{e}$, a more stringent (in the 'intended' sense) anti-defamation legislation would decrease corruption, but for larger values of $\bar{e}$, the result would be inverted. It is interesting to notice that this effect does not depend on a particular relation between $\bar{s}$ and $\bar{e}$ determined by some specific punishment technology.

More importantly for the analysis in this paper, the region in the space $(\bar{s}, \bar{e})$ in which this kind of equilibria exist is bounded and never equal to $[0,1] \times [0,1]$. There always exists a lower bound on the threshold level $\bar{s}$ (i.e. a limit on how stringent the anti-defamation law can be) and an upper bound on $\bar{e}$ (a limit on how stringent the anti-defamation law can be in the collateral sense). Over these limits, no equilibrium of this kind can exist. Notice nevertheless that there always exists a region in the space $(\bar{s}, \bar{e})$ for which this kind of equilibria exists, since for $\bar{s} = 1$, a sufficient condition would be $\bar{e} \leq 1$. But if $\bar{s} = 1$, then $\bar{e} = 1$ and therefore corruption is not limited at all: $\underline{c} = 1$.

The class of equilibria in lemma 3 represents those cases in which the principal mechanism can effectively neutralize the potential chilling effect of the anti-defamation law. It is not surprising therefore that these are not equilibria with chilling.

An interesting observation comes from the comparison of lemma 3 with the results presented in appendix B. Even if the collateral effect of the anti-defamation law on the media $\bar{e}$ is equal to 0, i.e. even if the anti-defamation law is such that all true scandals would be published by the media, when justice is imperfect, the level of corruption in equilibrium is larger than what would be if the justice system was perfect in the sense of section 3.1. The following corollary formalizes this last argument.

Corollary 1. *For all $\bar{s}$, the unique PBE with chilling a neutralizing mechanism has a larger equilibrium level of corruption than any equilibrium with perfect justice.*

The intuition behind this result is that even if in all equilibria with a chilling neutralizing mechanism all true scandals are published, still the imperfection in the justice guarantees the possibility of a revenue for the corrupted politician equal to the reparations accorded to him in biased trials. Since in all these equilibria, $\underline{c} < \bar{s} < s\left(e\left(\cdot\right)\right)$, then a maximally corrupted politician can expect a higher revenue from this source than a politician with corruption equal to $\underline{c}$ (the ICC is therefore tightened by this effect).

I have presented some results about the existence and characterization of all PBEs with a chilling neutralizing mechanism, the most relevant of which is that this class of equilibria exists only for a bounded region of the space $(\bar{s}, \bar{e})$. There is therefore a region of $(\bar{s}, \bar{e})$ for which the optimal mechanism is not chilling neutralizing, that is $e\left(1\right) < \bar{e}$. Whenever this is true, it also true that $x\left(1,1\right) = \phi$, i.e. there is some true scandal that would not be published by the media at least if the politician is maximally corrupted. This poses immediately the question of how the mechanism should interpret (how it should reward) the silence of the media. Should the principal reward the politician if the media are silent about his behavior, therefore implicitly interpreting the silence of the media as a signal of a low level of corruption? Or should the principal interpret the silence as the result of a politician so corrupted that the media fear to reveal the true scandals involving him? And in this last case, what is the effect of such an anti-defamation law on the equilibrium level of corruption of the politician? The following two propositions provide the answers to these questions.

Lemma 4 (the sound of silence). *Suppose that the mechanism $e\left(x\right)$ is not chilling neutralizing. If $e\left(\phi\right) = 1$, then no level of corruption $c < 1$ can be sustained in any PBE.*

Proof. See appendix A.2. □

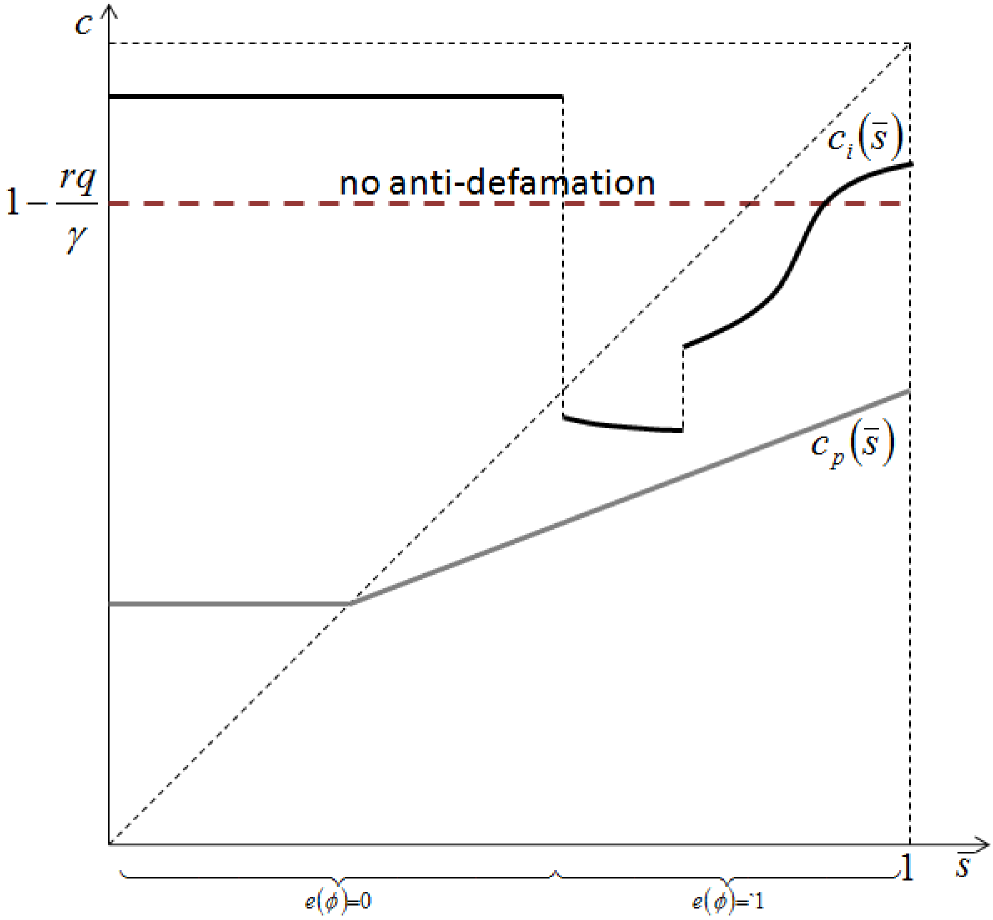
Lemma 5. *For all combinations of $\bar{s}$ and $\bar{e}$ such that the PBE mechanism $e\left(x\right)$ is not chilling neutralizing, there exists a unique PBE such that $e\left(\phi\right) = 0$, $e\left(\underline{c}\right) = 1$, $e\left(c\right) = 0$ for all $c > \underline{c}$, $c\left(e\left(x\right)\right) = \underline{c} \geq \bar{s}$ and*

$$\underline{c} = 1 - \frac{rq - \eta^0 - \left(\chi^0 - \chi^0_{\underline{c}}\right)}{\gamma} > 1 - \frac{rq}{\gamma}.$$

Proof. See appendix A.2. □

The implications of the last two results are of fundamental importance for the analysis of the role played by anti-defamation laws in deterring political corruption. In particular, lemma 4 implies that as far as $f/\delta\zeta < 1$, there is always a limit on the level of stringency of the anti-defamation law such that the potential for chilling effect is so large that the

Figure 1: Anti-defamation stringency and corruption. $c_i(\bar{s})$: imperfect justice; $c_p(\bar{s})$: perfect justice.



message ‘silence’ is interpreted by the electorate as a signal of the high level of corruption of the politician. Lemma 4 warns about the implication of a misunderstanding of media’s silence. If silence is misunderstood, i.e. it is interpreted as the media observing a false scandal and not publishing it, then no level other than the maximum level of corruption is attainable under any mechanism. Indeed, in these cases, the probability of the media not reporting any scandal is larger for a maximally corrupted politician than for a politician limiting his corruption to the accepted and tolerated level of corruption.

Lemmata 3, 4 and 5 imply that for all anti-defamation laws $(\bar{s}, \bar{e})$ there exists a unique equilibrium level of corruption. Figure 1 represents a possible relation between $\bar{s}$ and the equilibrium level of corruption $c_i(\bar{s})$ for a specific punishment technology, i.e. a class of $\rho(\cdot)$ for which it is possible to describe a relation between $\bar{s}$ and $\bar{e}$ (the graph represents the case of a technology for which a decrease in $\bar{s}$ always induces an increase in $\bar{e}$). To better understand the relevance of the imperfections in the justice system, figure 1 portraits as well the equilibrium level of corruption $c_p(\bar{s})$ for the case of pefect justice (see appendix B).

The fact *per se* that the silence of the media is interpreted by the principal as a noisy signal of extreme corruption not only reduces the effectiveness of the media in reducing political corruption, but increases the lower bound of corruption at a higher level than what could be achieved if no anti-defamation protection at all was in place (see proposition 1). Formally:

Remark 1. For all combinations of $\bar{s}$ and $\bar{e}$ such that the PBE mechanism $e(x)$ is not chilling neutralizing, the equilibrium level of corruption is larger than if there was no anti-defamation protection.

Recalling lemma 3, we know that for all parameters of the model, there exists always a lower bound on $\bar{s}$ for which no equilibrium with a chilling neutralizing mechanism can exist. This in turn implies that as far as justice is imperfect in the sense of section 3.1, then there is always a threshold on anti-defamation law stringency such that a more stringent law would generate more corruption in equilibrium than no anti-defamation protection at all. This happens as soon as anti-defamation laws are so stringent for the media, that the equilibrium mechanism is such that there exists at least one scandal (possibly only $x = 1$) that would not be published by the media even if it were to be true. Thus as soon as some potential chilling can survive in equilibrium, anti-defamation laws can only increase equilibrium corruption. The last result could then be read as follows:

> If the anti-defamation law is so stringent that in equilibrium there exists at least one scandal that would not be worth publishing by the media even if true, then the equilibrium level of corruption is larger than without any anti-defamation protection.

This result constitutes a rationale for the U.S. Supreme Court ruling *New York Times Co. v. Sullivan*, 376 U.S. 254 (1964). The Supreme Court decision on this case held that all statements about the conduct of public officials, even those that can be proven to be false, are protected under the First Amendment guarantee of the freedom of the press. The case for libel exists only if the plaintiff can prove that the defendant's statements are made with *actual malice* (with knowledge that they are false or in reckless disregard of their truth or falsity)(p. 280). For the Supreme Court, indeed, "erroneous statement is inevitable in free debate, and [...] it must be protected" (p. 271). The Alabama law provision, judged as unconstitutional by the Supreme Court, held that it sufficed to prove the falsity of the accusation for the defendant to be liable. This constituted indeed a threat for the media, which are in most cases unable to know for certain whether the allegation can be proven in court to be true or false. After the Supreme Court ruling, the expected cost for a journalist publishing something she knew to be true is severely limited by the obvious difficulty for the plaintiff of proving the state of mind of the journalist.

Appendix B shows that in the case of perfect justice, a higher reparation δ would decrease the equilibrium level of corruption in all equilibria with defamation. In the case of imperfect justice, when there is an equilibrium with defamation, then this is an equilibrium with a chilling neutralizing mechanism, and therefore

$$\underline{c} = 1 - \frac{\theta\left(\bar{s}, r\left(1-\bar{e}\right), \psi^{\bar{e}}\right) - \left(1-\bar{e}\right)\left(\eta^{0} + \chi^{0} - \chi^{0}_{\underline{c}}\right)}{\gamma - \left(1-q\right)\psi^{\bar{e}}}.$$

An increase in δ would decrease both the numerator and the denominator of the last term. Nevertheless, an increase in δ can in this case both increase or decrease $\underline{c}$. By implicit differentiation, and making use of the fact that $\underline{c} > 0$, it is possible to show that a sufficient condition for $\frac{\partial \underline{c}}{\partial \delta} > 0$ is $\zeta \geq \left(1-q\right)\bar{s}$. It is easy to notice from lemma 5 that in all other cases δ has either none or a positive effect on $\underline{c}$.

Remark 1 tells us that if there is a scandal that would not be published by the media if true, then corruption is larger than without any anti-defamation law. What the corollary does not say is that in equilibrium there actually will be a scandal so *bad* and still true. It is interesting to notice in fact that even in the case of $\bar{e} > 0$, i.e. when there is a potential risk of chilling effect, there is no equilibrium with chilling. In particular–and perhaps more strikingly–even in the case of an anti-defamation law so restrictive that the silence of the media is interpreted as a sign of very high corruption, there is no chilling effect. The reason for this is that all PBEs are such that $c\left(e\left(x\right)\right) = \underline{c}$ and $e\left(c\right) = 1$ for all $c \leq \underline{c}$, implying that any true scandal $s \leq c$ is worth being published by the firm since $\pi\left(c\right) - \zeta\left(1 - e\left(c\right)\right)\rho\left(c\right) > 0$ for all $c \leq \underline{c}$. Nevertheless, lemma 5 guarantees that in all equilibria with a non chilling neutralizing mechanism there is no defamation ($\bar{s} < \underline{c}$). The next corollary states these last two results.

Corollary 2. *Any PBE without a chilling neutralizing mechanism is not an equilibrium with defamation. There is no equilibrium with chilling.*

I have mentioned above that the U.S. Supreme Court ruling *New York Times Co. v. Sullivan*, 376 U.S. 254 (1964) reduced the expected cost for a journalist of publishing something she knew to be true. My interpretation of the ruling is that the Court's intention was to avoid the occurrence of equilibria without a chilling neutralizing mechanism. Nevertheless, by reducing the expected cost of publishing something true, the Court reduced the cost of publishing falsities as well. Indeed, the last corollary states that in the equilibria the Court wished to avoid, there is no defamation. On the opposite, we have noted earlier that chilling neutralizing equilibria can exhibit defamation if the conditions in point 1 of lemma 3 are verified.

Recalling lemma 4, if the anti-defamation law is so stringent that there exists no PBE with a chilling neutralizing mechanism and the silence of the media is interpreted by the principal as a sign that the media firm has observed a false scandal, then the equilibrium level of corruption must be equal to the maximum level of corruption possible, 1. An implication of this result is that if the choice of the principal is limited to mechanisms that reward the politician if the media are silent, i.e. $e(\phi) = 1$, then, for anti-defamation laws stringent enough, any mechanism would give the same equilibrium level of corruption (the maximum one). It is easy to see that in this case it is possible to have an equilibrium mechanism with $e(x) = 0$ for all $x < \hat{x}$ for some $\hat{x} < 1$ and $\tilde{s}(e(\cdot)) < 1$. This would imply that in this equilibrium there is some chilling. Indeed the media would not publish a scandal $s = 1$ even if in equilibrium this is a true scandal and the media know so. Therefore, an equilibrium with chilling can exist if the anti-defamation law is stringent enough and the silence of the media is misinterpreted. This result could explain why in the case of Governor Marrazzo, tabloids refused to buy his compromising videotape.

Another observation is that the equilibrium level of corruption is always bounded below in all PBEs for all anti-defamation laws. The equilibrium level of corruption is in this sense tolerated and accepted by the principal, giving a theoretical representation of the proverbial *physiological* level of corruption in democracy: if the electorate does not accept and absolve small cases of corruption, politicians do not have incentives to choose levels of corruption lower than the maximum available to them.

5 Robustness check

In section 3, I have assumed that the punishment for a defaming media firm is a function of the seriousness of the scandal published, independently of the level of political corruption assessed by the judges. Another possibility is that instead the punishment is a function of the distance between the scandal's allegations and the assessed level of corruption g. In this case, the punishment function is $\rho(x - g)$ which is assumed to be (strictly) increasing and convex for $x > g$, always continuous and equal to 0 for $x \leq g$. The following briefly analyzes this new framework..

In the case of a false scandal, the optimal strategy of the firm would be to publish all scandals less or equal to a threshold function of the true level of the politician's corruption

c,

$$\bar{s}(c) := \begin{cases} \{s \in [c,1] : \pi(s) = \rho(s-c)\} & \text{if it exists;} \\ 1 & \text{if } \pi(s) > \rho(s-c) \text{ for all } s \in [c,1]. \end{cases}$$

Lemma 6. *$\bar{s}(c)$ is non-decreasing in c and*

$$\frac{\partial \bar{s}(c)}{\partial c} > 1 \textit{ if } \bar{s}(c) < 1.$$

Proof. See appendix A.3. □

Define $\bar{c} := \{c \in [0,1] : \bar{s}(c) = 1 \text{ and } \bar{s}(c') < 1 \text{ for all } c' < c\}$, then $\bar{c}$ represents the minimum level of true corruption such that all false scandals are worth being published by the media. For all $c < \bar{c}$, in fact, there exists at least one scandal (at least $s = 1$) that would not be published by the media if false. This threshold level is indeed a suitable measure of the level of stringency of the anti-defamation law. When the anti-defamation law gets more stringent, i.e. when firms are punished very harshly even for scandals barely above the true level of corruption of the politician, then $\bar{c} \to 1$. For the specifications of the model, $\bar{c}$ is never exactly 1 unless $\lim_{s\to 0} \rho(s) \geq \pi(1)$. Notice that it also implies that $\lim_{s\to c} \rho(s-c) = \pi(1) > \pi(c)$ for all $c \in [0,1)$.

A small but important difference with respect to the benchmark model of section 3 is that here we always have some defamation unless $\bar{c} = 1$. Indeed, for all anti-defamation laws, $\bar{s}(c) \geq c$, for all $c \in [0,1]$, with the equality holding only for $\bar{c} = 1$.

In the case of imperfect justice, i.e. when $f/\delta\zeta < 1$, the strategy of the media firm will depend on the threshold function $\bar{s}(c)$ (the largest false scandal the firm would publish if the politician is corrupted up to level c) and $\tilde{s}(e(\cdot)) := \max\{f/\delta\zeta, s(e(\cdot))\}$ where

$$s(e(\cdot)) := \begin{cases} \begin{matrix} s \in [0,1] : \\ \pi(s) = s\zeta(1-e(s)) \int\limits_0^s \rho(s-z)\,dz \end{matrix} & \text{if it exists;} \\ 1 & \text{if } \begin{matrix} \pi(s) > \\ > s\zeta(1-e(s)) \int\limits_0^s \rho(s-z)\,dz \\ \text{for all } s \in [0,1]. \end{matrix} \end{cases}$$

Suppose in fact that the firm observes a scandal $s \leq c$. If the scandal is published, the revenue of the firm is $\pi(s)$. If $s > f/\delta\zeta$, with probability $(1-e(s))$, the politician will not be rewarded by the principal and will sue the firm. The dispute will be resolved in favor of the politician with probability $s\zeta$. The expected punishment of the firm resulting from publishing the scandal is therefore $s\zeta(1-e(s)) \int_0^s \rho(s-z)\,dz$. For the revenue of the firm to be larger than the expected punishment we must have

$$\begin{aligned} \pi\left(s\right) &\geq s\zeta\left(1-e\left(s\right)\right)\int\limits_{0}^{s}\rho\left(s-z\right)dz \\ e\left(s\right) &\geq 1-\frac{\pi\left(s\right)}{s\zeta\int\limits_{0}^{s}\rho\left(s-z\right)dz}. \end{aligned} \tag{9}$$

Lemma 7. *Suppose that $e\left(x\right)$ is a non-increasing function of x, then $s\left(e\left(\cdot\right)\right)$ is unique.*

Proof. See proof of lemma 1 in appendix A.1 and notice that $\int_0^s \rho\left(s-z\right)dz$ is increasing and concave in s. □

As in section 4, I am limiting the analysis to all non-increasing mechanisms $e\left(x\right)$.

Condition 2 (1)**.** $e\left(x\right)$ is a non-increasing function of x.

It is relevant to notice that $\tilde{s}\left(e\left(\cdot\right)\right)>\bar{s}$ since $s\left(e\left(\cdot\right)\right)>\bar{s}$ by construction. The interpretation of this fact is simply that if the net payoff of publishing a scandal is high enough that the firm would publish it even if it was false, then the same scandal would be published if it was true.

The strategy of the firm is therefore given by

$$x\left(s,c\right)=\begin{cases} s & \text{if } s\leq c \text{ and } s\leq\tilde{s}\left(e\left(\cdot\right)\right); \\ \phi & \text{if } s\leq c \text{ and } s>\tilde{s}\left(e\left(\cdot\right)\right); \\ s & \text{if } s>c \text{ and } s\leq\bar{s}\left(c\right); \\ \phi & \text{if } s>c \text{ and } s>\bar{s}\left(c\right). \end{cases}$$

Following the same strategy as in section 4, here I first analyze all equilibria with chilling neutralizing mechanism as defined in section 4. A chilling neutralizing mechanism is a mechanism for which if a true scandal $s=1\leq c$ is observed by the media, then the media's payoff from the publication is positive. From the inequality in (9), a true scandal $s=1\leq c$ will be published by the media if and only if

$$e\left(1\right)>\ \bar{e}:=\max\left\{\hat{e},0\right\}$$

where (

$$\hat{e}:=1-\frac{\pi\left(1\right)}{\zeta\int\limits_{0}^{1}\rho\left(1-z\right)dz}.$$

For the purpose of clarity, I repropose here the definition of a chilling neutralizing mechanism.

Definition 3 (2)**.** A mechanism is *chilling neutralizing* (all true scandals are worth being published) if $e(1) \geq \bar{e}$.

The claim of this section is that there exists an upper bound for $\bar{c}$ such that an equilibrium with a chilling neutralizing mechanism exists. Furthermore, in symmetry with section 4, where $\bar{c}$ is higher than this limit, the silence of the media is interpreted by the principal as a sign of high level of political corruption and therefore $e(\phi) = 0$. Whenever this is the case, then the equilibrium level of corruption is larger than without any anti-defamation protection. The following proposition states this result formally.

Proposition 3. *Call $\mathcal{N}$ the subset of $\bar{c} \in [0,1]$ such that the PBE of the model has a chilling neutralizing mechanism, then $\mathcal{N} \subset [0,1]$. For all $\bar{c} \in [0,1] \backslash \mathcal{N}$, the equilibrium mechanism is such that $e(\phi) = 0$ and the equilibrium level of corruption is equal to $c(e(x)) > 1 - \frac{rq}{\gamma}$.*

Proof. See appendix A.3. □

The last result should be read in parallel with the discussion in section 4. As in section 4, in fact, there exists a limit to the stringency of the anti-defamation law after which the law produces more damages than benefits in terms of reducing corruption. This limit is exactly at that point for which the principal would start to interpret the silence of the media as the signal of a high level of corruption.

In the model of this section, as observed earlier, defamation always arises unless $\bar{c} = 1$. An interesting result is that, even with the specifications of this section, chilling cannot arise in equilibrium. Suppose in fact that the anti-defamation law is extremely stringent and $\bar{c} = 1$. Then it must be that $\lim_{s \to c} \rho(s - c) = \pi(1) > \pi(c)$ for all $c \in [0,1)$. This would imply an equilibrium with a non chilling neutralizing mechanism and $\tilde{s}(e(\cdot)) = \max\{f/\delta\zeta, c(e(x))\} \geq c(e(x))$. Therefore all true scandals $s \leq c(e(x))$ will be published. The following corollary formalizes this argument.

Corollary 3. *There is no equilibrium with chilling.*

An interesting difference between the analysis of the model in this section and the analysis in section 4 is the presence of equilibrium defamation in equilibria without a chilling neutralizing mechanism. In section 4, the emergence of defamation in equilibrium coexisted only with equilibria with a chilling neutralizing mechanism. All equilibria without a chilling neutralizing mechanism were equilibria without defamation. In this section, on the contrary, while still there exists no equilibrium with a chilling neutralizing mechanism and no equilibrium defamation, there exist a continuum of measures $\bar{c}$ such that the unique PBE is an equilibrium with defamation but not an equilibrium with a chilling neutralizing mechanism.

6 Conclusions

In this paper, I have discussed how anti-defamation regulations can strengthen or diminish the role played by mass media in deterring corruption in democratic polities. More stringent anti-defamation legislations always reduce the equilibrium level of political corruption only if in equilibrium media firms are punished exclusively when their published allegation refers to wrongdoings actually committed by the politician. This result ceases to hold in presence of imperfections in the justice system such that in equilibrium a firm could be punished (although rarely) for publishing evidence of a true scandal. In this case, there always exist anti-defamation laws stringent enough that the induced level of corruption would be larger than it would be in the total absence of any anti-defamation protection. In particular, were the legislation so severe that in equilibrium there would exist at least one scandal that mass media would prefer not to publish even if they know it to be true, then the silence of the media is always interpreted by the electorate as the result of the chilling effect, and equilibrium corruption is always larger than in the absence of any anti-defamation law. This is true even if only one scandal has this property and even if in equilibrium no such a scandal would result from a wrongdoing actually committed by the politician.

In the model presented in sections 3 and 5, the chilling effect is exclusively a potential result of very stringent anti-defamation laws. The presence of this potential, meaning that there exist scandals that would not be published by the media if true, is enough to make the anti-defamation law counterproductive, in the sense that it actually increases corruption. Nonetheless, there exists no perfect Bayesian equilibrium in which some true scandal is not published by the media. The presence of such a scandal can arise in equilibrium only if the electorate (or the authority controlling the politician) is incapable of punishing the politician in the absence of any published scandal. For stringent enough anti-defamation legislations, if the electorate cannot punish the politician when media remain silent or if it is incapable of correctly interpreting the silence of the media, then corruption is maximal and there exist true scandals that remain unpublished.

The model presented in sections 3 and 5 allows for different interpretations. Of particular interest is the one that replaces the politician with the manager of a corporation and replaces the media with the auditor in charge of supervising her conduct and reporting to the shareholders. As with the politician, the manager might be removed by the shareholders before a proper investigation of the allegations made by the auditor can be undertaken. The last financial crisis has brought to the public debate the question of which, if any, enforceable laws should be applied to auditors that report false wrongdoings or–and perhaps more importantly–do not report real cases of manager misconduct. With the appropriate modifications to the nature of the revenue and punishment functions, the model of this paper is suitable to be applied to this different scenario.

References

Antle, Rick, "Auditor Independence," *Journal of Accounting Research*, Spring 1984, *22* (1), 1–20.

Barendt, Eric, Laurence Lustgarten, Kenneth Norrie, and Hugh Stephenson, *Libel and the Media: the Chilling Effect*, Oxford University Press, 1997.

Besley, Timothy and Andrea Prat, "Handcuffs for the Grabbing Hand? Media Capture and Government Accountability," *American Economic Review*, June 2006, *96* (3), 720–736.

_ **and Robin Burgess**, "The Political Economy of Government Responsiveness: Theory and Evidence from India," *The Quarterly Journal of Economics*, November 2002, *117* (4), 1415–1451.

Djankov, Simeon, Caralee McLeish, Tatiana Nenova, and Andrei Shleifer, "Who Owns the Media?," *NBER Working Paper*, 2001, (8288).

Ferraz, Claudio and Frederico Finan, "Exposing Corrupt Politicians: The Effects of Brazil's Publicly Released Audits on Electoral Outcomes," *The Quarterly Journal of Economics*, May 2008, *123* (2), 703–745.

Garoupa, Nuno, "Dishonesy and Libel Law: The Economics of the 'Chilling Effect'," *Journal of Institutional and Theoretical Economics*, June 1999, *155* (2), 284–.

_ , "The Economics of Political Dishonesty and Defamation," *International Review of Law and Economics*, June 1999, *19* (2), 167–180.

Kofman, Fred and Jacques Lawarree, "Collusion in Hierarchical Agency," *Econometrica*, May 1993, *61* (3), 629–656.

Suphachalasai, Suphachol, "Bureaucratic Corruption and Mass Media," *Environmental Economy and Policy Research*, 2005, (05.2005).

Thoreau, Henry David, "A Week on the Concord and Merrimack Rivers," in Carl. F. Hovde, ed., *The Writings of Henry D. Thoreau*, Princeton University Press, 1980.

Tirole, Jean, "Hierarchies and Bureaucracies: On the Role of Collusion in Organizations," *Journal of Law, Economics, & Organization*, Autumn 1986, *2* (2), 181–214.

A Appendix

Notation 1. Define:

1. $\chi^{e(\cdot)} := \chi_1^{e(\cdot)}$. Furthermore, notice that $\chi_{\underline{c}}^{e(\cdot)} :\leq \chi^{e(\cdot)}$;

   (a) if $e(x) = e$ for all $x \in \left[\frac{f}{\delta\zeta}, \min\{s(e(\cdot)), \underline{c}\}\right]$, $\chi_c^e := \chi_c^{e(\cdot)}$;

   (b) if $e(c) = 0$, $\eta^0 := \eta_c = q(c\delta\zeta - f)$;

   (c) for all $e \in (0,1)$, $\psi^e := \delta - e(\delta - \varepsilon)$.

A.1 Proof of lemma 1

Proof. Notice that

$$\lim_{s\to 0} \frac{\pi(s)}{s\zeta\rho(s)} = +\infty > \lim_{s\to\infty} \frac{\pi(s)}{s\zeta\rho(s)}.$$

Furthermore, because of $\pi''(\cdot) \leq 0$ and $\rho''(\cdot) \geq 0$, $\exists! w \in \mathbb{R}_+ :$

$$\frac{\partial \frac{\pi(s)}{s\zeta\rho(s)}}{\partial s} \leq 0 \text{ iff } s \geq w.$$

This in turn implies that $\lim_{s\to 0}\left[1 - \frac{\pi(s)}{s\zeta\rho(s)}\right] = -\infty$ and $1 - \frac{\pi(s)}{s\zeta\rho(s)}$ is increasing wherever is positive. Since by hypothesis $e(x)$ is never increasing and always positive, there exists always only one $s(e(\cdot))$. □

A.2 Proofs of section 4

Proof of lemma 3. Suppose $\underline{c} < \bar{s}$. The ICC must be binding only for $c = 1$ and is therefore

$$\gamma\underline{c} + R^{e(\cdot)}(\underline{c}, \bar{s}) + D^{e(\cdot)}(\underline{c}, \bar{s}) + \eta_{\underline{c}} + \chi_{\underline{c}}^{e(\cdot)} \geq$$

$$\geq \gamma + rq\bar{e} + r(1-q)\int_0^1 e(z)\,dz + \eta_1 + \chi^{e(\cdot)}$$

$$\Leftrightarrow \gamma\underline{c} + rqe(\underline{c}) + r(1-q)(1-\bar{s})e(\phi) + D^{e(\cdot)}(\underline{c}, \bar{s}) + \eta_{\underline{c}} + \chi_{\underline{c}}^{e(\cdot)} \geq$$

$$\geq \gamma + rq\bar{e} + r(1-q)\int_{\bar{s}}^1 e(z)\,dz + \eta_1 + \chi^{e(\cdot)}$$

which implies an optimal mechanism having $e\left(\underline{c}\right) = e\left(\phi\right) = 1$ and $e\left(c\right) = \bar{e}$ for all $c > \underline{c}$, and hence □

$$\begin{aligned}
&\gamma\underline{c} + rq + \left(1-q\right)\left[r\left(1-\bar{s}\right) + \left(\bar{s}-\underline{c}\right)\psi^{\bar{e}}\right] + \chi_{\underline{c}}^{\bar{e}} \\
&\qquad\qquad\qquad \geq \gamma + rq\bar{e} + r\left(1-q\right)\bar{e}\left(1-\bar{s}\right) + \eta_1 + \chi^{\bar{e}}
\end{aligned}$$

$$\begin{aligned}
&\Leftrightarrow \underline{c} \geq 1 - \frac{r\left(1-\bar{e}\right)q + \left[r\left(1-\bar{e}\right) - \psi^{\bar{e}}\right]\left(1-q\right)\left(1-\bar{s}\right) - \left(1-\bar{e}\right)\left(\eta^0 + \chi^0 - \chi_{\underline{c}}^0\right)}{\gamma - \left(1-q\right)\psi^{\bar{e}}} \qquad (10) \\
&\Leftrightarrow \bar{s} \geq \underline{c} \geq 1 - \frac{\theta\left(\bar{s}, r\left(1-\bar{e}\right), \psi^{\bar{e}}\right)}{\gamma - \left(1-q\right)\psi^{\bar{e}}} + \left(1-\bar{e}\right)\frac{\left(\eta^0 + \chi^0 - \chi_{\underline{c}}^0\right)}{\gamma - \left(1-q\right)\psi^{\bar{e}}} \geq 1 - \frac{\theta\left(\bar{s}, r, \delta\right)}{\gamma - \left(1-q\right)\delta}
\end{aligned}$$

where in (10) I have used the fact that $\chi_a^{\bar{e}} = \left(1-\bar{e}\right)\chi_a^0$ for all a. Since $\bar{s} \geq \underline{c}$, then a necessary condition is

$$\begin{aligned}
\bar{s} &> 1 - \frac{\theta\left(\bar{s}, r\left(1-\bar{e}\right), \psi^{\bar{e}}\right)}{\gamma - \left(1-q\right)\psi^{\bar{e}}} + \frac{\left(1-\bar{e}\right)\eta^0}{\gamma - \left(1-q\right)\psi^{\bar{e}}} \\
&> 1 - \frac{\theta\left(\bar{s}, r\left(1-\bar{e}\right), \psi^{\bar{e}}\right)}{\gamma - \left(1-q\right)\psi^{\bar{e}}} + \left(1-\bar{e}\right)\frac{\left(\eta^0 + \chi^0 - \chi_{\underline{c}}^0\right)}{\gamma - \left(1-q\right)\psi^{\bar{e}}} \\
\Leftrightarrow \quad & \bar{s} > 1 - \frac{r\left(1-\bar{e}\right)q - \left(1-\bar{e}\right)\eta^0}{\gamma - r\left(1-\bar{e}\right)\left(1-q\right)} = 1 - \frac{rq - \eta^0}{\frac{\gamma}{1-\bar{e}} - r\left(1-q\right)} = \frac{\eta^0 - \left(\frac{\gamma}{1-\bar{e}} - r\right)}{\frac{\gamma}{1-\bar{e}} - r\left(1-q\right)}.
\end{aligned}$$

Since $\bar{s} \leq 1$, then this implies

$$\bar{e} < 1 - \frac{\gamma\left(1-\bar{s}\right)}{r\left(1-\bar{s}\right) + \left(rq\bar{s} - \eta^0\right)}$$

but since $\bar{e} \geq 0$, then this in turn implies

$$\bar{s} > \frac{\eta^0 - \left(\gamma - r\right)}{\gamma - r\left(1-q\right)} \geq \frac{\eta^0 - \left(\frac{\gamma}{1-\bar{e}} - r\right)}{\frac{\gamma}{1-\bar{e}} - r\left(1-q\right)}. \qquad (11)$$

Furthermore, from (10), in order to have $c \leq 1$ we have

$$r\left(1-\bar{e}\right)q + \left[r\left(1-\bar{e}\right) - \psi^{\bar{e}}\right]\left(1-q\right)\left(1-\bar{s}\right) - \left(1-\bar{e}\right)\left(\eta^0 + \chi^0 - \chi_{\underline{c}}^0\right) \geq 0$$

$$\Rightarrow \bar{e} \leq \frac{\theta\left(\bar{s}, r, \delta\right) - \eta^0}{\theta\left(\bar{s}, r, \delta\right) - \eta^0 + \varepsilon\left(1-q\right)\left(1-\bar{s}\right)} \in \left(0, 1\right). \qquad (12)$$

So both conditions (11) and (12) must hold.

Suppose now that $\underline{c} \geq \bar{s}$. The ICC becomes

$$\gamma\underline{c} + R^{e(\cdot)}\left(\underline{c}, \underline{c}\right) + \chi_{\underline{c}}^{e(\cdot)} \geq \gamma + rq\bar{e} + r\left(1-q\right)\int\limits_{0}^{1} e\left(z\right)dz + \eta_1 + \chi^{e(\cdot)}$$

$$\Leftrightarrow \gamma\underline{c} + rqe\left(\underline{c}\right) + r\left(1-q\right)\left(1-\underline{c}\right)e\left(\phi\right) + \chi_{\underline{c}}^{e(\cdot)} \geq$$

$$\geq \gamma + rq\bar{e} + r\left(1-q\right)\int\limits_{\underline{c}}^{1} e\left(z\right)dz + \eta_1 + \chi^{e(\cdot)}$$

which implies an optimal mechanism having $e\left(\underline{c}\right) = e\left(\phi\right) = 1$ and $e\left(c\right) = \bar{e}$ for all $c > \underline{c}$, and hence

$$\gamma\underline{c} + rq + r\left(1-q\right)\left(1-\underline{c}\right) \geq \gamma + rq\bar{e} + r\left(1-q\right)\left(1-\underline{c}\right)\bar{e} + \eta_1 + \left(\chi^{\bar{e}} - \chi_{\underline{c}}^{\bar{e}}\right)$$

$$\Leftrightarrow \underline{c} \geq 1 - \frac{rq\left(1-\bar{e}\right) - \left(1-\bar{e}\right)\left(\eta^0 - \chi^0 - \chi_{\underline{c}}^0\right)}{\gamma - r\left(1-q\right)\left(1-\bar{e}\right)}$$

In lemma 5, I prove that a level of corruption equal to $1 - \frac{rq-\eta^0-\left(\chi^0-\chi_{\underline{c}}^0\right)}{\gamma}$ is always sustainable in an equilibrium with $e\left(1\right) = 0$. Therefore can conclude that $e\left(1\right) = \bar{e}$ is optimal only if

$$1 - \frac{rq\left(1-\bar{e}\right) - \left(1-\bar{e}\right)\left(\eta^0 - \chi^0 - \chi_{\underline{c}}^0\right)}{\gamma - r\left(1-q\right)\left(1-\bar{e}\right)} \leq 1 - \frac{rq - \eta^0 - \left(\chi^0 - \chi_{\underline{c}}^0\right)}{\gamma}$$

$$\Leftrightarrow \bar{e} \leq \frac{r\left(1-q\right)}{\gamma - r\left(1-q\right)} =: W.$$

It remains therefore to prove that there exists a lower bound on $\bar{s}$. Notice that if $\bar{e} \leq W$, then this means that

$$\rho\left(1\right) \leq \frac{\pi\left(1\right)}{\left(1-W\right)\zeta}.$$

The most restrictive convex function $\rho\left(\cdot\right)$ and the least rewarding revenue function $\pi\left(\cdot\right)$ such that the last expression holds are

$$\rho\left(s\right) = \frac{\pi\left(1\right)}{\left(1-W\right)\zeta}s \text{ and } \pi\left(s\right) = \pi\left(0\right) + \pi\left(1\right)s$$

which would imply

$$\bar{s} = \frac{\pi\left(0\right)}{\pi\left(1\right)}\frac{\left(1-W\right)\zeta}{1+\left(1-W\right)\zeta} > 0$$

being the lowest possible $\bar{s}$ for which an equilibrium with a chilling neutralizing mechanism and no defamation exists.

Proof of lemma 4. Suppose $e(1) < \bar{e}$, it is trivial to show that $e(1) = 0$ is optimal, implying $\chi_{\underline{c}}^{e(x)} = \eta_{\underline{c}} = 0$. Moreover, by contradiction, suppose that $e(\phi) = 1$ and there exist a level of corruption $\underline{c} < 1$ such that $U(\underline{c}, e(c)) \geq U(1, e(c))$. □

First notice that

$$\bar{s} \geq \Omega(\bar{s}) := 1 - \frac{(r-\delta)(1-q)(1-\bar{s})}{\gamma - (1-q)\delta}$$
$$\Rightarrow \bar{s} \geq 1.$$

Suppose $\underline{c} \leq \bar{s}$, then we have $U(\underline{c}, e(c)) \geq U(c, e(c))$ for all $c \in (\underline{c}, \bar{s}) \Rightarrow$

$$\gamma \underline{c} + R^{e(\cdot)}(\underline{c}, \bar{s}) + D^{e(\cdot)}(\underline{c}, \bar{s}) + \chi_{\underline{c}}^{e(\cdot)} \geq \gamma c + R^{e(\cdot)}(c, \bar{s}) + D^{e(\cdot)}(c, \bar{s}) + \eta_c + \chi_c^{e(\cdot)}$$

$$\Leftrightarrow \gamma \underline{c} + rqe(\underline{c}) + (1-q) \int_{\underline{c}}^{c} [\delta - e(z)(\delta - \varepsilon)]\, dz \geq \gamma c + rqe(c) + \eta_c + \chi_c^{e(\cdot)} - \chi_{\underline{c}}^{e(\cdot)}$$

for all $c \in (\underline{c}, \bar{s}]$, which implies $e(c) = 0$ for all $c > \underline{c}$. This implies that Furthermore we have $U(\underline{c}, e(c)) \geq U(1, e(c)) \Rightarrow$

$$\gamma \underline{c} + R^{e(\cdot)}(\underline{c}, \underline{c}) + (1-q)(\bar{s} - \underline{c})\delta \geq$$
$$\geq \gamma + rqe(\phi) + r(1-q)\left[\int_{0}^{\underline{c}} e(z)\, dz + [1 - \hat{s}(e(\cdot))]\, e(\phi)\right] + \left(\chi^0 - \chi_{\underline{c}}^0\right)$$

Implying (if $e(\underline{c}) = 1$)

$$\gamma \underline{c} + rq + (1-q)[r(1-\bar{s}) + (\bar{s} - \underline{c})\delta] \geq \gamma + rq + r(1-q)[1 - \hat{s}(e(\cdot))] + \left(\chi^0 - \chi_{\underline{c}}^0\right)$$

$$\Leftrightarrow \bar{s} \geq \underline{c} \geq 1 - \frac{(r-\delta)(1-q)(1-\bar{s})}{\gamma - (1-q)\delta} + \frac{r(1-q)[1 - \hat{s}(e(\cdot))] + \left(\chi^0 - \chi_{\underline{c}}^0\right)}{\gamma - (1-q)\delta} \geq \Omega$$

for all $\hat{s}(e(\cdot))$. Therefore $\bar{s} \geq 1$. But in the case of $\bar{s} = \hat{s}(e(\cdot)) = 1$, $c > 1$, contradicting the hypothesis.

Suppose now that $\bar{s} < \underline{c} \leq f/\delta\zeta$. $U(\underline{c}, e(c)) \geq U(1, e(c)) \Rightarrow$

$$\gamma \underline{c} + R^{e(\cdot)}(\underline{c}, \underline{c}) \geq$$
$$\geq \gamma + rqe(\phi) + r(1-q)\left[\int_{0}^{\underline{c}} e(z)\, dz + [1 - \hat{s}(e(\cdot))]\, e(\phi)\right] + \chi^0$$

with $e(\underline{c}) = 1$

$$\Leftrightarrow \gamma\underline{c} + rq + r(1-q)(1-\underline{c}) \geq \gamma + rq + r(1-q)\left[1-\hat{s}(e(\cdot))\right] + \chi^0$$

$$\underline{c} \geq 1 + \frac{r(1-q)\left[1-s(e(\cdot))\right] + \chi^0}{\gamma - r(1-q)} \geq 1$$

contradicting the hypothesis that $\underline{c} < 1$.

Suppose now then $\underline{c} > \frac{f}{\delta\zeta}$. $U(\underline{c}, e(c)) \geq U(1, e(c)) \Rightarrow$

$$\gamma\underline{c} + rqe(\underline{c}) + r(1-q)\left[\int_0^{\underline{c}} e(z)\,dz + (1 - \min\{\hat{s}(e(\cdot)), \underline{c}\})\,e(\phi)\right] \geq$$

$$\geq \gamma + rqe(\phi) + r(1-q)\left[\int_0^{\underline{c}} e(z)\,dz + \left[1-\hat{s}(e(\cdot))\right]e(\phi)\right] + \left(\chi^0 - \chi^0_{\underline{c}}\right)$$

therefore we have, with $e(\underline{c}) = 1$ being again optimal,

$$\gamma\underline{c} + rq + r(1-q)\left[(1-\min\{\hat{s}(e(\cdot)), \underline{c}\})\right] \geq \gamma + rq + r(1-q)\left[1-\hat{s}(e(\cdot))\right] + \left(\chi^0 - \chi^0_{\underline{c}}\right)$$

$\Rightarrow$

1. if $\underline{c} \leq \hat{s}(e(\cdot))$,

$$\gamma\underline{c} + r(1-q)\left[(1-\underline{c})\right] \geq \gamma + r(1-q)\left[1-\hat{s}(e(\cdot))\right] + \left(\chi^0 - \chi^0_{\underline{c}}\right)$$

$$\underline{c} \geq 1 + \frac{r(1-q)\left[1-\hat{s}(e(\cdot))\right] + \left(\chi^0 - \chi^0_{\underline{c}}\right)}{\gamma - r(1-q)} \geq 1;$$

   (a) if $\underline{c} > \hat{s}(e(\cdot))$,

$$\begin{aligned} \gamma\underline{c} + rq + r(1-q)\left[1-\hat{s}(e(\cdot))\right] &\geq \gamma + rq + r(1-q)\left[1-\hat{s}(e(\cdot))\right] + \left(\chi^0 - \chi^0_{\underline{c}}\right) \\ \gamma\underline{c} &\geq \gamma + \chi^0 \Rightarrow \underline{c} \geq 1 \end{aligned}$$

   in both cases contradicting the hypothesis.

Proof of lemma 5. Suppose $e(\phi) = 0$ and $e(1) < \bar{e}$. The optimal mechanism in this case always has $e(c) = 0$ for all $c > \underline{c}$ and $e(\underline{c}) = 1$. For $\underline{c} \leq \bar{s}$, the ICC is

$$\gamma\underline{c} + rq + r(1-q)\,\delta(\bar{s} - \underline{c}) \geq \gamma + \chi^0$$

$$\Rightarrow \underline{c} \geq 1 + \frac{r(1-q)\delta(1-\bar{s}) + \chi^0}{\gamma - r(1-q)\delta} > 1$$

implying that $\underline{c} > \bar{s}$ (for an equilibrium with defamation, see lemma 3). □

If $\frac{f}{\delta\zeta} \geq \underline{c} > \bar{s}$, then $\gamma\underline{c} + rq \geq \gamma + \chi^0$, implying $\underline{c} \geq 1 - \frac{rq-\chi^0}{\gamma}$. Otherwise, $\frac{f}{\delta\zeta} < \underline{c}$ and $\underline{c} \geq 1 - \frac{rq-\left(\chi^0 - \chi^0_{\underline{c}}\right)}{\gamma}$. To conclude, notice that for $\frac{f}{\delta\zeta} \geq \underline{c}$, $\chi^{e(x)}_{\underline{c}} = 0$ for all $e(x)$.

A.3 Proofs of section 5

Proof of lemma 6. I first prove that $\bar{s}(c)$ is strictly increasing for $\bar{s}(c) < 1$. □

From the definition of $\bar{s}(c)$, if $\bar{s}(c) < 1$, then $\pi(\bar{s}(c)) = \rho(\bar{s}(c) - c)$. By implicit differentiation:

$$\frac{\partial \bar{s}(c)}{\partial c} = \frac{\rho'(\bar{s}(c) - c)}{\rho'(\bar{s}(c) - c) - \pi'(\bar{s}(c))}. \tag{13}$$

By contradiction: suppose that $\pi(\bar{s}(c)) = \rho(\bar{s}(c) - c)$ and $\frac{\partial \bar{s}(c)}{\partial c} \leq 0$, then it must be that $\rho'(\bar{s}(c) - c) - \pi'(\bar{s}(c)) \leq 0$. Since $\rho''(\cdot) \geq 0$, then this implies that $\rho'(s-c) \leq \pi'(\bar{s}(c))$, for all $s \leq \bar{s}(c)$. We also know that $\rho(0) = 0 < \pi(c)$ for all $c \in [0,1]$, therefore we have

$$\rho(0) + \int_c^{\bar{s}(c)} \rho'(s-c)\,ds < \pi(c) + \int_c^{\bar{s}(c)} \pi'(s)\,ds \Leftrightarrow \rho(\bar{s}(c) - c) < \pi(\bar{s}(c))$$

contradicting the hypothesis.

From equation 13 and the last result ($\rho'(\bar{s}(c) - c) - \pi'(\bar{s}(c)) > 0$) we can conclude that, for $\bar{s}(c) < 1$, $\frac{\partial \bar{s}(c)}{\partial c} > 1$. To conclude the proof, notice that $\bar{s}(c') = 1 \Rightarrow \frac{\partial \bar{s}(c)}{\partial c} = 0$ for all $c \geq c'$.

Proof of proposition 3. In this proof I omit to prove unicity of equilibria. The arguments for it are essentially the same as in the previous proofs. □

With $e(1) = \bar{e}$, all true scandals would be published, therefore there will be no chilling but (always) some defamation and potentially some biased trials. The ICC for $c = 1$ is

$$\gamma\underline{c} + rqe(\underline{c}) + r(1-q)(1-\bar{s}(\underline{c}))e(\phi) + D^{e(\cdot)}(\underline{c}, \bar{s}(\underline{c})) + \eta_{\underline{c}} + \chi^{e(\cdot)}_{\underline{c}} \geq$$

$$\geq \gamma + rq\bar{e} + r(1-q)\int_{\bar{s}(c)}^{1} e(z)\,dz + \eta_1 + \chi^{e(\cdot)}_1$$

implying that any optimal mechanism must have $e\left(\underline{c}\right)=e\left(\phi\right)=1$ and $e\left(c\right)=\bar{e}$ for all $c>\underline{c}$. This in turn implies

$$\underline{c}\geq 1-\frac{r\left(1-\bar{e}\right)q+\left[r\left(1-\bar{e}\right)-\psi^{\bar{e}}\right]\left(1-q\right)\left(1-\bar{s}\left(\underline{c}\right)\right)-\left(1-\bar{e}\right)\left(\eta^{0}+\chi^{0}-\chi^{0}_{\underline{c}}\right)}{\gamma-\left(1-q\right)\psi^{\bar{e}}}.$$

As $\bar{c}\to 1$, $\bar{s}\left(c\right)\to c$. Therefore the last inequality becomes

$$\underline{c}\geq 1-\frac{r\left(1-\bar{e}\right)q-\left(1-\bar{e}\right)\left(\eta^{0}+\chi^{0}-\chi^{0}_{\underline{c}}\right)}{\gamma-r\left(1-\bar{e}\right)\left(1-q\right)}. \tag{14}$$

Suppose $e\left(c\right)=e\left(\phi\right)=0$ for all $c>\underline{c}$ then the ICC is

$$\gamma\underline{c}+rq+\left(1-q\right)\left(\bar{s}\left(\underline{c}\right)-\underline{c}\right)\delta\geq\gamma+\eta^{0}+\left(\chi^{0}-\chi^{0}_{\underline{c}}\right)$$

implying

$$\underline{c}\geq 1-\frac{rq-\eta^{0}-\left(\chi^{0}-\chi^{0}_{\underline{c}}\right)}{\gamma-\left(1-q\right)\delta}>1-\frac{rq}{\gamma}.$$

This then implies that it must be

$$1-\frac{r\left(1-\bar{e}\right)q-\left(1-\bar{e}\right)\left(\eta^{0}+\chi^{0}-\chi^{0}_{\underline{c}}\right)}{\gamma-\left(1-q\right)\psi^{\bar{e}}}\geq 1-\frac{rq-\eta^{0}-\left(\chi^{0}-\chi^{0}_{\underline{c}}\right)}{\gamma-\left(1-q\right)\delta}$$

for the chilling neutralizing mechanism to be part of a PBE. But, as seen in the proof of lemma 3, this is never true for $\bar{e}\geq 0$. Therefore we can conclude that there always exist an upper bound for $\bar{c}$ strictly less than zero such that for all anti-defamation laws with a higher $\bar{c}$, a mechanism with $e\left(\phi\right)=0$ is always preferred by the principal.

It remains to prove that a mechanism with $e\left(c\right)=0$ for all $c>\underline{c}$, but $e\left(\phi\right)=1$ is not part of a PBE. In this case we have

$$\begin{aligned}&\gamma\underline{c}+rqe\left(\underline{c}\right)+r\left(1-q\right)\left(1-\bar{s}\left(\underline{c}\right)\right)e\left(\phi\right)+\left(1-q\right)\left(\bar{s}\left(\underline{c}\right)-\underline{c}\right)\delta\geq\\&\qquad\geq\gamma+rqe\left(\phi\right)+r\left(1-q\right)\left[\int_{0}^{\underline{c}}e\left(z\right)dz+\left[1-\hat{s}\left(e\left(\cdot\right)\right)\right]e\left(\phi\right)\right]+\left(\chi^{0}-\chi^{0}_{\underline{c}}\right)\end{aligned}$$

Implying (if $e\left(\underline{c}\right)=1$)

$$\gamma\underline{c}+rq+\left(1-q\right)\left[r\left(1-\bar{s}\right)+\left(\bar{s}\left(\underline{c}\right)-\underline{c}\right)\delta\right]\geq\gamma+rq+r\left(1-q\right)\left[1-\tilde{s}\left(e\left(\cdot\right)\right)\right]+\left(\chi^{0}-\chi^{0}_{\underline{c}}\right)$$

$$\Rightarrow \bar{s}\left(\underline{c}\right) \geq \underline{c} \geq 1 - \frac{\left(r-\delta\right)\left(1-q\right)\left(1-\bar{s}\left(\underline{c}\right)\right)}{\gamma-\left(1-q\right)\delta} + \frac{r\left(1-q\right)\left[1-\tilde{s}\left(e\left(\cdot\right)\right)\right] + \left(\chi^0-\chi^0_{\underline{c}}\right)}{\gamma-\left(1-q\right)\delta} \geq$$

$$\geq 1 - \frac{\left(r-\delta\right)\left(1-q\right)\left(1-\bar{s}\left(\underline{c}\right)\right)}{\gamma-\left(1-q\right)\delta}$$

which implies $\bar{s}\left(\underline{c}\right) \geq 1$. But if $\bar{s}\left(\underline{c}\right) \geq 1$, then

$$\underline{c} \geq 1 + \frac{r\left(1-q\right)\left[1-\tilde{s}\left(e\left(\cdot\right)\right)\right] + \left(\chi^0-\chi^0_{\underline{c}}\right)}{\gamma-\left(1-q\right)\delta} \geq 1.$$

B NOT FOR PUBLICATION Perfect justice

In the case of perfect justice, i.e. when no biased trial occurs in any PBE, a media firm will be successfully sued by the politician if and only if it has reported a false scandal ($x > c$).

Lemma 8. *The optimal response function of a firm observing corruption level c and a signal s is*

$$x\left(s,c\right) = \begin{cases} s & \textit{if } s \leq \max\left\{c,\bar{s}\right\}; \\ \phi & \textit{otherwise.} \end{cases}$$

Proof. Suppose a firm observes a level of corruption c and a scandal s. If the firm does not publish the scandal, its profits will be $\pi\left(\phi\right) = 0$. If $s \leq c$, then the scandal is true and the firm will not be successfully sued by the politician if it publishes the scandal. In this case, the expected punishment for the firm is 0, hence it is optimal for the firm to publish the scandal. If $c < s \leq \bar{s}$, publishing the scandal, the firm will for sure incur in a punishment $\rho\left(s\right) \leq \pi\left(s\right)$. The expected profits of the firm are therefore positive (or zero if $s = \bar{s}$) and is optimal for the firm to publish. To conclude the proof, suppose $s > \max\left\{c,\bar{s}\right\}$. In this case, if the scandal is published, the politician will always successfully sue the firm which will incur a punishment $\rho\left(s\right) > \pi\left(s\right)$, and therefore negative profits. In this case it is optimal for the firm to send the message ϕ. □

The implication of lemma 8 is that the threshold level $\bar{s}$ is the only relevant measure of the level of stringency of the anti-defamation law. The choice of the firm in fact depends solely on this threshold and not on the entire shape of the function $\rho\left(x\right)$. A lower $\bar{s}$ corresponds therefore to a more stringent anti-defamation law for all practical purposes and it is the measure used in this section to evaluate the role of anti-defamation law in strengthening the efficiency of democratic institutions in deterring corruption.

Define:

$$U\left(c,e\left(x\right)\right) = \gamma c + R^{e\left(\cdot\right)}\left(c,\max\left\{c,\bar{s}\right\}\right) + D^{e\left(\cdot\right)}\left(c,\max\left\{c,\bar{s}\right\}\right)$$

then $U(c, e(x))$ is the expected payoff of a politician choosing c under the mechanism $e(x)$, where the first term, γc, is the direct payoff from corruption.

The problem for the principal is to choose a mechanism $e(x)$ such that the incentive compatibility constraint (ICC)

$$U(\underline{c}, e(x)) \geq U(c, e(x)), \qquad \text{for all } c > \underline{c}$$

holds and there is no $\underline{c}' < \underline{c}$ such that there exists a mechanism $e'(x)$ for which ICC holds.

Before stating the main proposition of this section, I here define an *almost unique (a.u.) PBE* as a set of PBEs which differ only for parts of the mechanism design $e(x)$ irrelevant for the choices of the politician and the media firm. The following proposition formally states the central result of this section.

Definition 4. Call $\mathcal{A}$ the set of all PBEs of the model. Call $\mathcal{E}$ the set of mechanisms $e(x)$ such that $e(x)$ belongs to at least one PBE. If $c(e(x)) = c$ and $x(s, c)$ is unique for all $e(x) \in \mathcal{E}$, $\mathcal{A}$ is an a.u. PBE.

Proposition 4. *For all anti-defamation laws $\rho(x)$, there exists an a.u. PBE characterized by*

1. $\underline{c} \in \left[1 - \frac{rq}{\gamma - r(1-q)}, 1 - \frac{rq}{\gamma - (1-q)\delta}\right]$ such that $e(\underline{c}) = 1$, $e(x) = 0$ for all $x > \underline{c}$ and $c(e(x)) = \underline{c}$;

   (a) $\underline{c} = \begin{cases} 1 - \frac{rq}{\gamma - r(1-q)} & \text{if } \bar{s} \leq 1 - \frac{rq}{\gamma - r(1-q)}; \\ 1 - \frac{\theta(\bar{s}, r, \delta)}{\gamma - (1-q)\delta} & \text{if } 1 - \frac{rq}{\gamma - r(1-q)} < \bar{s} \leq 1; \end{cases}$

   (b) $e(\phi) = 1$;

   (c) $x(s, c) = \begin{cases} s & \text{if } s \leq \max\{c, \bar{s}\}; \\ \phi & \text{otherwise.} \end{cases}$

   where $\theta(\bar{s}, r, \delta) = rq + (r - \delta)(1 - q)(1 - \bar{s})$.

Proof. Use lemma 8. Suppose $\underline{c} > \bar{s}$, the ICC is □

$$\gamma \underline{c} + R^{e(\cdot)}(\underline{c}, \underline{c}) \geq \gamma c + R^{e(\cdot)}(c, c)$$

$$\Leftrightarrow \gamma \underline{c} + rqe(\underline{c}) + r(1-q)(1-\underline{c})e(\phi) \geq$$

$$\geq \gamma c + rqe(c) + r(1-q)\left[\int_{\underline{c}}^{c} e(z)\,dz + (1-c)e(\phi)\right]$$

for all $c > \underline{c}$. It is optimal then to have $e(c) = 0$ for all $c > \underline{c}$ and $e(\underline{c}) = e(\phi) = 1$. In this case the ICC is

$$\begin{aligned} \gamma\underline{c} + rq + r(1-q)(1-\underline{c}) &\geq \gamma c + r(1-q)(1-c) \\ \underline{c} &\geq 1 - \frac{rq}{\gamma - r(1-q)} \end{aligned} \tag{15}$$

since the RHS of the inequality in (15)is maximized for $c = 1$. The solution for the principal is therefore to solve the program

$$\begin{aligned} & \min_{\underline{c}} \underline{c} \\ \text{s.t. } \underline{c} &> \bar{s} \\ \underline{c} &\geq 1 - \frac{rq}{\gamma - r(1-q)} \end{aligned}$$

which gives

$$\underline{c} = \max\left\{\bar{s}, 1 - \frac{rq}{\gamma - r(1-q)}\right\}. \tag{16}$$

Suppose now that $\underline{c} \leq \bar{s}$, the ICC is

$$\gamma\underline{c} + R^{e(\cdot)}(\underline{c}, \underline{c}) + D^{e(\cdot)}(\underline{c}, \bar{s}) \geq \gamma c + R^{e(\cdot)}(\underline{c}, \underline{c}) + D^{e(\cdot)}(c, \bar{s})$$

Limiting the attention to $c \leq \bar{s}$, then the ICC is

$$\gamma\underline{c} + R^{e(\cdot)}(\underline{c}, \bar{s}) + D^{e(\cdot)}(\underline{c}, \bar{s}) \geq \gamma c + R^{e(\cdot)}(c, \bar{s}) + D^{e(\cdot)}(c, \bar{s})$$

$$\Leftrightarrow \gamma\underline{c} + rqe(\underline{c}) + (1-q)\left[\int_{\underline{c}}^{c} [\delta - e(z)(\delta - \varepsilon)]\, dz\right] \geq \gamma c + rqe(c)$$

and it is optimal to have $e(\underline{c}) = 1$ and $e(c) = 0$ for all $c > \underline{c}$, leaving

$$\gamma\underline{c} + rq + (1-q)(c - \underline{c})\delta \geq \gamma c.$$

Notice that under the assumption that $c \leq \bar{s}$, then the RHS is maximized by $c = \bar{s}$, and therefore

$$\begin{aligned} \gamma\underline{c} + rq + (1-q)(\bar{s} - \underline{c})\delta &\geq \gamma\bar{s} \\ \Rightarrow \underline{c} &\geq \bar{s} - \frac{rq}{\gamma - (1-q)\delta}. \end{aligned} \tag{17}$$

Suppose now that $c > \bar{s}$, the ICC is

$$\gamma \underline{c} + R^{e(\cdot)}(\underline{c}, \bar{s}) + D^{e(\cdot)}(\underline{c}, \bar{s}) \geq \gamma c + R^{e(\cdot)}(c, \bar{s})$$

$$\Leftrightarrow \gamma \underline{c} + rqe(\underline{c}) + r(1-q)(1-\bar{s})e(\phi) + D^{e(\cdot)}(\underline{c}, \bar{s}) \geq$$

$$\geq \gamma c + rqe(c) + r(1-q)\left[\int\limits_{\bar{s}}^{c} e(z)\,dz + (1-c)e(\phi)\right]$$

and is obviously optimal to have $e(\underline{c}) = e(\phi) = 1$ and $e(c) = 0$ for all $c > \underline{c}$, leaving

$$\gamma \underline{c} + rq + (1-q)\left[r(1-\bar{s}) + (\bar{s} - \underline{c})\delta\right] \geq \gamma c + r(1-q)(1-c)$$

and since again the RHS is maximized by $c = 1$,

$$\gamma \underline{c} + rq + (1-q)\left[r(1-\bar{s}) + (\bar{s} - \underline{c})\delta\right] \geq \gamma$$

$$\Rightarrow \underline{c} \geq 1 - \frac{rq + (r-\delta)(1-q)(1-\bar{s})}{\gamma - (1-q)\delta}. \tag{18}$$

Therefore it must be that both (17) and (18) hold. Notice that for (17) to be binding it needs

$$\bar{s} - \frac{rq}{\gamma - (1-q)\delta} > 1 - \frac{rq + (r-\delta)(1-q)(1-\bar{s})}{\gamma - (1-q)\delta}$$

$$\Rightarrow 1 - \bar{s} < \frac{(r-\delta)(1-q)}{\gamma - (1-q)\delta}(1-\bar{s})$$

$$r(1-q) > \gamma$$

which contradicts the regularity assumption $\gamma \geq r$. Therefore, for $\underline{c} \leq \bar{s}$ it must be

$$\underline{c} = 1 - \frac{rq + (r-\delta)(1-q)(1-\bar{s})}{\gamma - (1-q)\delta}. \tag{19}$$

Equations (16) and (19) conclude the proof when noticed that

$$1 - \frac{rq + (r-\delta)(1-q)(1-\bar{s})}{\gamma - (1-q)\delta} = \bar{s}$$

$$\Leftrightarrow \bar{s} = 1 - \frac{rq}{\gamma - r(1-q)} \geq 0$$

where the last inequality is guaranteed by the regularity assumption $\gamma \geq r$.

Proposition 4 establishes the existence of a lower and an upper bound for corruption. In particular, for any anti-defamation law, there exists a unique level of corruption induced by the optimal mechanism in all PBEs. The optimal mechanism for the principal is to reward the politician if the firm publishes a scandal equal (and possibly less than) the equilibrium level of corruption, and to always punish the politician if the scandal is larger. It is important to notice at this point, especially in view of the analysis in the following section, that the optimal response of the principal to the message 'silence' is to reward the politician, i.e. the silence of the media about the conduct of the politician is interpreted by the principal as a sign of a non-corrupt conduct of the politician. This result relies on the fact that, given the strategy of the media firm in lemma 8, the probability of the media remaining silent about the politician's conduct is inversely proportional to the level of corruption of the politician for all levels of corruption larger than the threshold level $\bar{s}$.

The unicity of the equilibrium level of corruption for any threshold level $\bar{s}$ allows for some comparative statics between the stringency of the anti-defamation laws and the level of corruption. The following corollary formally states the relation between the stringency of anti-defamation laws, as measured by $\bar{s}$, and the equilibrium level of corruption.

Corollary 4. *For all $\bar{s} \in [0,1]$ there exists a single-valued non-decreasing function $c(\bar{s})$*

$$c(\bar{s}) = \begin{cases} 1 - \frac{rq}{\gamma - r(1-q)} & \text{if } \bar{s} \leq 1 - \frac{rq}{\gamma - r(1-q)} \\ 1 - \frac{\theta(\bar{s}, r, \delta)}{\gamma - (1-q)\delta} & \text{otherwise} \end{cases}$$

such that the equilibrium level of corruption $c(e(x))$ is equal to $c(\bar{s})$ in all PBEs.

A typical function $c(\bar{s})$ is depicted in figure B. The last result indicates that more stringent anti-defamation laws are always at least as effective in deterring corruption as less stringent ones and strictly more effective for a non-empty interval $[1 - rq/[\gamma - r(1-q)], 1]$ if $\delta < r$. The intuition behind this result is that, in the case of perfect justice as defined in the previous section, more stringent anti-defamation laws reduce the amount of false scandals reaching the principal, forcing the media to reveal more about its private information about the true level of corruption and increasing the precision of the information available to the principal. This effect disappears when $\bar{s}$ is less or equal to $1 - rq/[\gamma - r(1-q)]$. At this level of stringency of the anti-defamation law, indeed, there is no more defamation in any PBEs (that is, no level of corruption less than $\bar{s}$ can be sustained in equilibrium) and further decreasing of $\bar{s}$ does not have any effect on the strategies of the media and the politician. The following corollary formalizes this argument.

Corollary 5. *If $\bar{s} > 1 - \frac{rq}{\gamma - r(1-q)}$, the a.u. PBE is an equilibrium with defamation. If $\bar{s} \leq 1 - \frac{rq}{\gamma - r(1-q)}$, there is no equilibrium with defamation.*

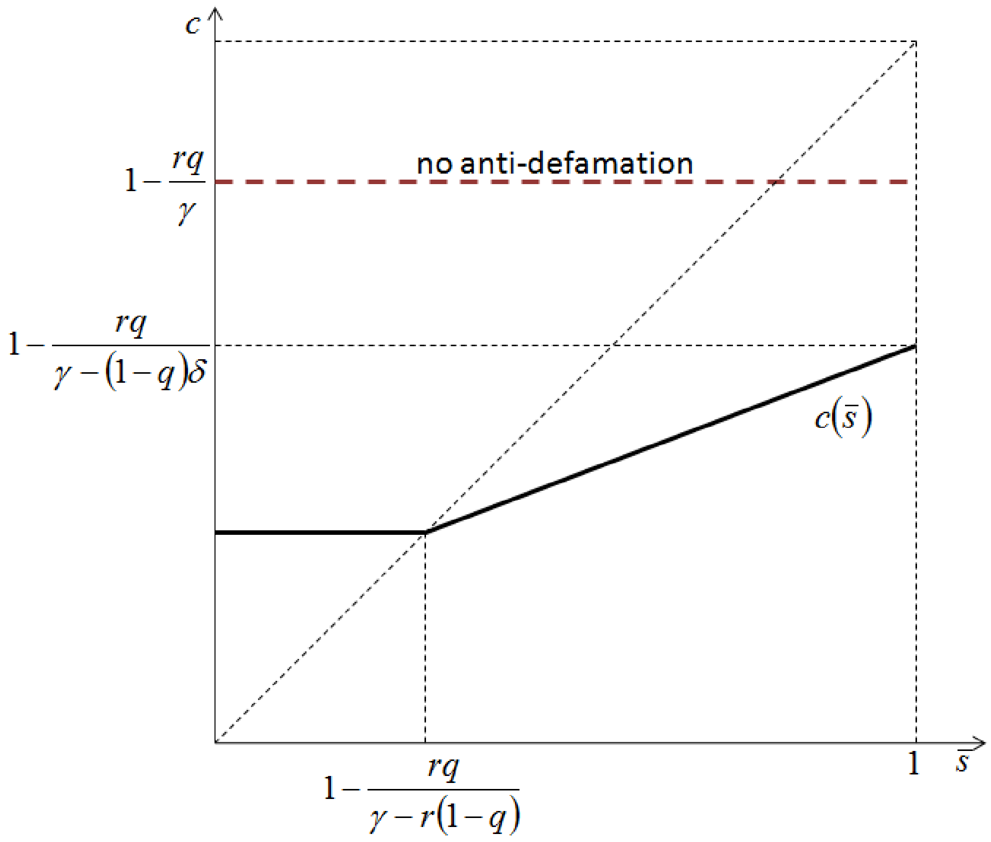


Figure 2: Anti-defamation stringency and corruption, $c(\bar{s})$. Perfect justice.

Proof. Straightforward from proposition 4. □

It is interesting to notice that $c(\bar{s})$ is strictly increasing for some $\bar{s}$ only if $\delta < r$. If instead $\delta = r$, i.e. the reparation accorded to the defamed and non-rewarded politician is equal to his lost reward, then any anti-defamation law gives the same equilibrium level of corruption. The intuition behind this result goes as follows: when $c(\bar{s}) < \bar{s}$, the equilibrium cost associated with defamation for the politician is the reward r he loses for all the messages $\bar{s} > x > c$ for which $e(x) = 0$. For all these messages, the politician will be compensated with δ. The difference $r - \delta$ is therefore a cost incurred by the politician who chooses $c = \underline{c}$, but not by the politician that chooses $c = 1$. If $\delta = r$, this cost disappears, relaxing the incentive compatibility constraint.

The arguments above do not mention the existence of any chilling effect. This could be surprising since a more stringent anti-defamation law, at least in principle, could reduce the precision of the signal received by the principal by reducing the amount of true scandals being published by the media. Nevertheless, it is easy to see from lemma 8 (or alternatively from point 4 in proposition 4) that with perfect justice there is no space for any chilling effect, since at least all true scandals, and perhaps some false, are always published by the media. The next result follows directly.

Corollary 6. *There is no equilibrium with chilling.*

Proof. Straightforward from point 4 in proposition 4. □

A brief comment should be added to the role played by the promise of future reparation for a defamed politician. It is easy to verify that, for $\bar{s} > 1 - rq/\left[\gamma - r\left(1-q\right)\right]$, since there is always some defamation in equilibrium, a higher reparation δ decreases corruption since it increases the payoff of a politician adhering to the tolerated level of corruption $\underline{c}$. Nonetheless, for more stringent anti-defamation laws, there is no defamation, so that the amount of reparation δ does not play any role.