Regulating Information With Bayesian Audiences

Yonathan A. Arbel[1] and Murat Mungan[2]

[1]School of Law, University of Alabama
[2]Scalia School of Law, George Mason University

We analyze the regulation of false statements in the presence of Bayesian audiences. We find that: (a) Often, moderate sanctions are optimal even though strict sanctions can fully deter all false statements; (b) the existence of separating equilibria—where only truthful statements are made—critically depends on judicial accuracy; (c) the magnitude of sanctions trades-off false information, chilling of truthful statements, and litigation costs; and (d) private enforcement often dominates public enforcement despite the lack of commitment. We emphasize the case of defamation law, and discuss other contexts including securities regulation, whistle-blower incentives, jury trials, and reports of criminal activity.

We are thankful for the comments of Scott Baker, Albert Choi, Ezra Friedman, Nuno Garoupa, Alex Lee, Ben McMichael, Alan Miller, Sepehr Shahshahani, Kathy Spier, Bruno Srulovici, Abe Wickelgren, and the participants of the 2019 Law and Economic Theory Conference.

1. Introduction

In many contexts, we use the law to regulate the exchange of information between private parties. A common concern is that an interested speaker would spread false information to advance its own private goals. To prevent this, the law will sometimes punish false statements or reward truthful ones.

A common neglect in the literature is the interaction between the severity of the law and the audience's beliefs and actions. In reality, audiences process information differently when its veracity is strictly regulated. This neglect may be due to the natural tendency to focus on the parties that take an active part in the legal process (the victim-defendant and the speaker-plaintiff) and to abstract from non-participating parties, namely the public (Heymann, 2012). Whatever the reason, regulation of the information environment—the flow and quality of information to the public—affects audiences and their beliefs quite directly.

Our object here is to bridge the audience gap by formalizing the interaction between speakers, the targets of their speech, and members of the audience. We employ a tool that is naturally apt at analyzing this issue, namely, a Bayesian game, and we investigate the impact of the strictness of the law on the emerging Perfect Bayesian Equilibria (PBE). Under this framework, a speaker, who has private information about a business or individual ("target"), may make

Draft, Vol. 0, No.0,
doi:/ewmxxx

claims about the target to an audience member. The audience member then decides whether to interact—trade, trust, socialize—with the target. If the target loses an interaction, he may bring a lawsuit against the speaker. Within this framework, it is socially optimal for audiences to only interact with high-quality targets and avoid low-quality ones. The key variable of interest is the strictness of the law, which we operationalize through the level of damages awarded to the target if the lawsuit is successful—this reflects the relatively broad discretion courts have in the determination of damages (Steenson 2014).

Our model contains four key features: (i) The information is provided by a party (the speaker) who is interested in influencing the audience's behavior, (ii) the audience makes decisions in light of the content of the supplied information, (iii) the speaker's objective conflicts with that of another party (the target), and (iv) the law penalizes the supply of false negative information by the speaker. These key features are present in many contexts, including: defamation law, whistle-blower rewards, complaint-driven law enforcement, and securities regulation. In some of these contexts, legal proceedings are initiated by the target (private enforcement) and in others by a governmental agency (public enforcement). Given the growing pressure to increase the regulation of defamatory speech coming from the Supreme Court, political leaders, lawyers, and scholars (Arbel & Mungan, 2019), we focus on defamation law as our running example with private enforcement in our baseline model (Sections 3 and 4). We subsequently extend the analysis to compare public and private enforcement, and discuss specific fields besides defamation law (in Section 5).

Our analysis reveals five central findings. First, the harmful effect of disparaging statements is deeply related to the strictness of the law itself. A speaker's statements may inform the audience's beliefs and actions. In choosing whether to make disparaging statements, speakers will consider the expected cost of a potential lawsuit against them. Stricter laws increase this cost. Thus, in equilibrium, the strictness of the law affects speakers and, anticipating this, also targets and audiences. These effects sometimes result in counter-intuitive implications, such that targets of speech who are 'good' types may *prefer* laxer laws, even though it would limit their recovery in a successful lawsuit. Such a conclusion is possible because strict laws make statements a more costly signal, and thus, a more reliable one in the eyes of audience members. A determined speaker could abuse this trust and spread falsities effectively.

The second conclusion is closely related to the dynamics which we just highlighted. We find that both very strict and lax laws have similar negative informational consequences. When the law is lax, i.e. damages are low, speakers frequently misstate the truth and audiences rely more on their priors rather than on statements (akin to babbling equilibria under cheap talk). However, if defamation laws are very strict, i.e., expected damages are high, then this may deter speakers from making even truthful assertions ("overpriced talk"). Whereas truth is a defense to a lawsuit, the risk of judicial mistake may be too great, and so speakers would refrain from sharing negative private information. Therefore, overly strict laws deprive the audience of meaningful infor-

mation.[1] Thus, our analysis reveals a basic insight with respect to regulation of the information environment: Both cheap and overpriced talk can undermine information dissemination.

Third, our analysis illuminates the importance of institutional considerations in designing information regulating laws. One key consideration is the court's subject-matter expertise and likelihood of delivering accurate judgments. If, in a given area, judges can fairly accurately detect false statements, imposing relatively large damages that deter false statements can lead to separating equilibria where only truthful statements are made. Using the law to regulate information continues to be optimal in cases where courts do not have the accuracy necessary to implement separating equilibria, but, can deter most false statements without chilling truthful statements. When courts are less capable of accurately adjudicating statements, the social cost of using the court system—operationalized by litigation costs—is key in determining whether information should be regulated. Even in these cases, when the gains from facilitating beneficial interactions and deterring harmful ones dwarfs litigation costs, moderate damages emerge as the optimal choice.[2]

Another implication pertains to the potential dynamic impact of information regulating laws. Specifically, moderate laws that cause the audience to rationally rely on speakers' statements broadens the gap between the frequency with which the audience interacts with good types versus bad types. This naturally increases the returns from being a good versus a bad type, thereby incentivizing individuals and firms to increase the quality of their products or services.

Lastly, our comparison of public and private enforcement reveals the relative merits of private enforcement. A public agency may be able to commit in advance to a certain level of enforcement. Whereas private parties are less capable of commitment, they enjoy a natural informational advantage regarding the merit of the lawsuit, as they know their own type. Consequently, private enforcement leads to more accurate litigation decisions, and an intuitive advantage of private enforcement emerges in our model: separating equilibria can only be achieved through private enforcement.

Overall, our framework and results add to the literature on information regulation by spotlighting the importance of audience effects, offering a formal framework that accounts for audiences, and emphasizing the risks of overly-stringent and lax regulatory regimes.

The next section offers some background and reviews the related literature. Section 3 presents the model and its analysis with a focus on cases where the courts are relatively accurate. Section 4 explains, in detail, the more complicated trade-offs that emerge when courts are not accurate enough to achieve separating equilibria. Section 5 includes several potential extensions of the ba-

1. When we consider honest and other types of speakers, we also show that strict laws can be worse than lax ones, for similar informational reasons.

2. Incidentally, this conclusion can offer a rationale to the longstanding distinction in defamation law between facts and opinions, which are generally unregulated.

sic model, such as the public enforcement case, the generalization of the model to cases where speakers may be motivated to speak truthfully or to excessively praise the target, and discussions of contexts other than defamation law. Section 6 provides concluding remarks.

2. Background and related literature

Various laws regulate information by sanctioning false disclosure or rewarding truthful sharing of information. Defamation is a classic example of the former and whistle-blowers of the latter. The literature on these topics is disparate, but contain the same question: How to design sanctions and rewards that would incentivize the optimal sharing of information. A common recurring omission is the possibility that the audience may update its beliefs, in a Bayesian manner, based on the size of the sanctions or rewards. Because of the fragmented nature of the literature, we will consider four examples.

Defamation law is perhaps the quintessential example of the problem of information regulation and thus serves as our running example. Under defamation law, a target of a (1) public statement that is (2) false and (3) harmful to one's reputation, can sue for all resulting damages. Judgments in this area can result in high payments, with some cases reporting jury judgments of tens of millions of dollars (Lesher v. Does, 2013). While courts and legislators understand the behavioral effects of defamation law, they are mostly preoccupied with the effect of defamation law on speakers' incentives ('chilling effect') and victim's rights (Bar-Gill & Hamdani, 2003, Acheson & Wohlschlegel, 2018). Consequently, they share a *virtually axiomatic* belief that stricter defamation laws would better protect victims (McNamara, 2007).

Until very recently, scant attention has been given to the audience effects of defamation law. This omission is significant, as defamatory speech is only harmful if it is both believed and acted upon. The focus of economic work in this area was media outlets, responsible investigative journalism, and political corruption (Garoupa, 1999a,b, Bar-Gill & Hamdani, 2003, and Dalvi & Refalo, 2007). We amplify here on two informal contributions that recognize the potential implications of audience effects (Arbel & Mungan, 2019, Hemel & Porat, 2019) by offering a formal and broader account.

Another example of information regulation comes from the literature on whistle-blowers, which studies the optimal rewards paid to the whistleblower. There, a primary concern is false reports by the whistle-blowers to an enforcement agency (Givati, 2016, Buccirossi et al. 2017, Deoorter & De Mot, 2005). One finding is that when the risk of false reporting is high, it might be necessary to avoid rewarding whistleblowers altogether, even though this means loss of information. What is not accounted for is how the agency, the "audience" of the report, reacts to information, given the size of the reward. With large rewards, the agency would be more likely to expect false reports.

Law enforcement provides another illustrative example. Although the police often has to weigh the credibility of a criminal activity report, this reality is not captured in the standard law and economics literature (for a review, see Polin-

sky and Shavell 2017), which typically relies on models where the probability of detection is only a function of enforcement expenditures. In reality, the police seeks to economize resources by investigating more thoroughly reports that appear credible—and its estimation is likely influenced by the sanctions levied against those who file false reports.

A final example comes from securities regulation. There, a company self-reports its performance, under an enforcement threat by the Securities and Exchange Commission (SEC). The literature recognizes that the agency's enforcement can be an important credibility mechanism (Stulz 2009), but it pays little attention to how strict enforcement interacts with investors and the trust they place in company disclosure.

Methodologically, our article borrows tools from the rich literature on signaling (Spence 1973) and cheap talk (Crawford & Sobel, 1982). Our analysis can also be interpreted as part of an emerging literature that looks at how laws can be used to create informal sanctions through the behavior of third parties (e.g., Deffains & Fluet, 2019, Mungan 2016, Bénabou & Tirole, 2011, 2006, Rasmusen 1996.)

3. Model

To study the behavioral effects of information regulation we focus on the example of defamation law, for the reasons noted in the introduction. We model the interactions between three types of parties: the speaker (S, she), the target of the speech (T, he), and the audience, captured by a representative member (A, it). A faces an informational problem: T is either a good or a bad type, and A's value of interacting with T depends on T's type, which is unknown to A. Before deciding, S, who knows T's type, communicates with A and may either disparage T or make a non-disparaging comment. As we are interested in defamation, we assume that S might benefit from blocking an interaction between A and T, and so S may choose to *defame* T–i.e., lie that T is a bad type[3]. Of course, many speakers may be motivated by a desire to speak truthfully or to facilitate interactions between T and A, and we consider this possibility in section 5.2.

We model the interactions as a Bayesian game, and use it to identify Perfect Bayesian Equilibria. Figure 1, below, depicts the interactions between these three parties and is helpful in following the detailed descriptions of the interactions that we provide, next.[4]

3.1 Preliminary Notation

We consider a game where T may be one of two types $t \in \{B, G\}$ where the letters abbreviate *bad* and *good*, respectively. T's type is privately known to himself and S, but not to A, who only knows that the proportion of good

3. Consistently with the law, truthful negative statements are not considered defamatory. However, the court may make errors in ascertaining whether a negative statement is truthful, and this possibility is incorporated in our model, as we explain below.

4. The figure does not depict Nature's draw of S's type, due to reasons we explain, below.

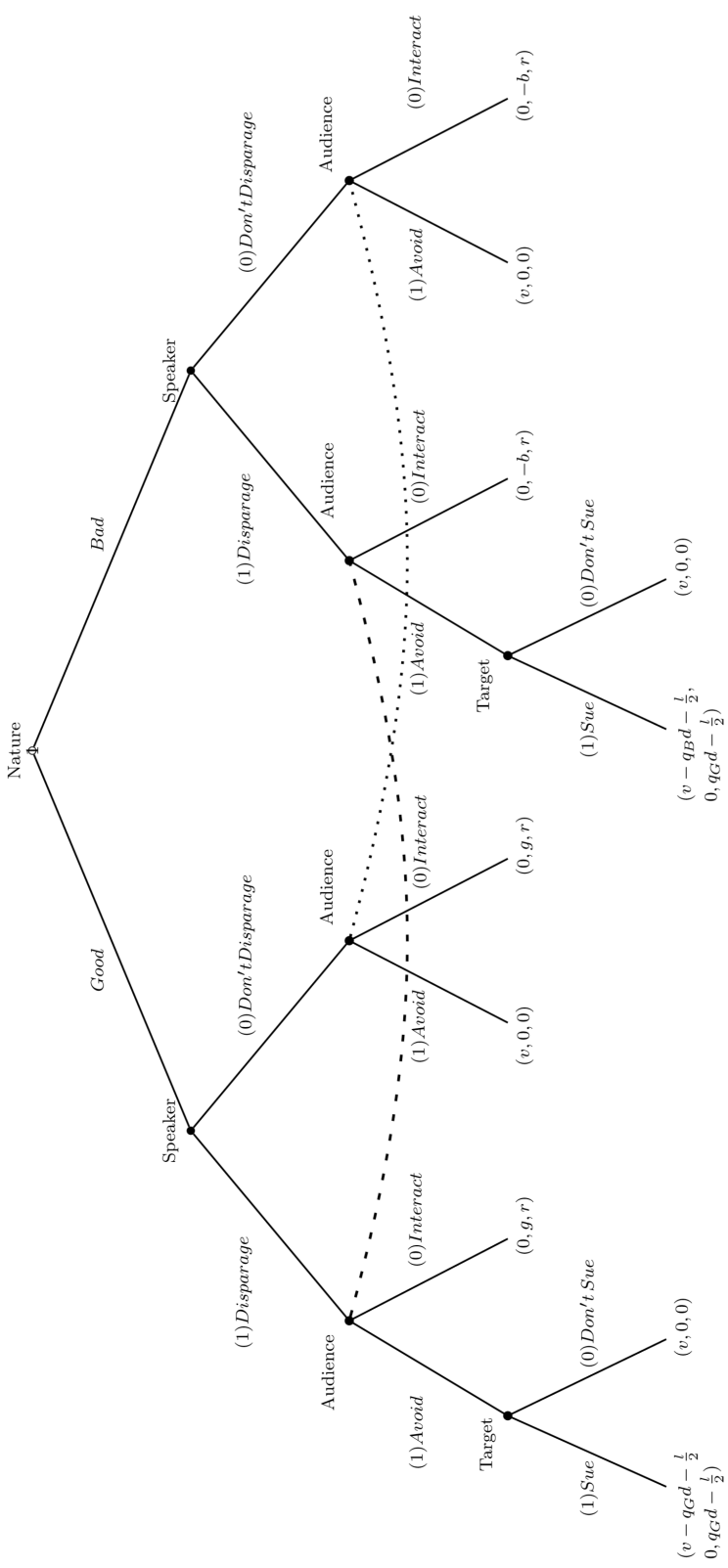

Figure 1 Extended game tree of the model.

types is $\gamma \in (0,1)$.[5] A prefers to interact with good types, but not bad types, because this results in a payoff of $g > 0 > -b$ where b is the cost A bears from interacting with a bad type. On the other hand, T always prefers to interact with A and obtains a benefit of r from the interaction. Finally, S has an interest in whether A and T interact and obtains a gain of v when they do not interact (alternatively, v can be interpreted as a loss incurred when A chooses to interact with T); v is a random variable drawn from the continuum $(0, \overline{v}]$ with the cumulative distribution function $F(v)$. The specific v-draw is private information available only to S, and we call v the speaker's *type*. We assume that interactions between A and T are socially valuable if, and only if, T is a good type, i.e. $r + g > \overline{v} > 0 > r - b$.

After Nature determines the types of T and S, T's type becomes common knowledge among T and S (but not A). At this point, S chooses what type of statement to send A regarding T's type. The types of possible statements follow defamation law's distinction between disparaging statements, which are potentially actionable, and non-disparaging statements, which are non-actionable (e.g., positive remarks, silence, opinion, etc.).[6]

Subsequently, A decides on whether to interact with T or to avoid him, and, finally, T, decides whether to bring a lawsuit against S if a disparaging remark was followed by A's choice to avoid interacting with T. We note that this setting includes the possibility of T suing S, even if T is in fact a bad type, i.e., a frivolous lawsuit may be brought. This is an important possibility because courts may err in their judgment.[7] To capture the parties' payoffs, we define the following:

d: damages paid by S to T when the court finds for T.
l: total litigation costs. We assume that litigation costs are not prohibitive ($l < \overline{v}$) and, without loss of generality, that the costs are equally shared by the parties.
q_t: probability of plaintiff victory when T is of type $t \in \{B, G\}$.

Figure 1 summarizes the parties' payoffs at the terminal nodes on the bottom in the order S, A, T. We note two graphical limitations of Figure 1. First, it does not show information sets describing A's knowledge regarding S's type, due to the depiction difficulty caused by S drawing her type from a continuum. Second, for ease of exposition, Figure 1 does not depict Nature's v draw determining S's inclination to disparage.

We use the variable d as the key policy lever to operationalize different kinds of defamation law regimes, since, as we note in the introduction, courts have very wide discretion in setting damages. d can be interpreted most directly as the level of damages awarded to a victim of defamation. The case where $d = 0$ represents a situation where there are no damages for defamation, which is, in our setting, functionally equivalent to no defamation law.

5. In section 5 we discuss the consequences of endogenizing γ.

6. We explain how our analysis informs the discussion of what types of statements ought to give rise to defamation claims in Section 5.4

7. The requirement of harm makes a lawsuit by T when the parties do interact unlikely

It is also worth noting that we take the odds of winning at trial, q_B and q_G, as exogenously given. This implies that courts are committed to reviewing cases only on their merit, and without bringing in their informed estimates about the proportion of frivolous cases. This is a standard commitment assumption in the enforcement literature, and with it in place, the ratio between q_B and q_G corresponds to a measure of judicial accuracy. In our analysis, below, we vary this parameter to study how it affects equilibria.

3.2 Players' Actions, Beliefs and Strategies

Next, we describe the players strategies, beliefs, and actions. For simplicity, each player's action is labelled as either 0 or 1, as follows:

Table 1 Players' Potential Actions

Player	Action	
	0	1
S	Don't Disparage	Disparage
A	Interact	Avoid
T	Don't	Litigate

Thus, as indicated in Figure 1, a suit is filed only in cases where all players' actions are 1. Using this notation, we can describe the strategies of each player as follows:

Table 2 Players' Strategies

Player	Strategy
S	$s(t,v) : \{B,G\} \times (0,\overline{v}] \rightarrow \{0,1\}$
A	$a(z) : \{0,1\} \rightarrow \{0,1\}$
T	$p(t) : \{B,G\} \rightarrow \{0,1\}$

Here, in specifying A's strategy, z denotes the statement received by A.

In order to identify Perfect Bayesian Equilibria (henceforth: PBE), we specify A's beliefs regarding T's type, as:[8]

$$\begin{aligned} x_0 &= \text{Belief that } T = g \text{ given } z = 0 \\ x_1 &= \text{Belief that } T = g \text{ given } z = 1 \end{aligned} \tag{1}$$

3.3 Perfect Bayesian Equilibrium Requirements and Definitions

As our solution concept is PBE, there are standard requirements that must be satisfied. Specifying these necessitates references to the conditional probabilities of T being a particular type, given a strategy played by S. To do so, we first specify the unconditional (or ex ante) probability with which S will

8. Because A's valuation of his interaction with T depends only on T's type, we need not specify A's beliefs regarding S's type for purposes of identifying the PBE.

disparage T given any strategy, s, as follows:

$$\mu(s) \equiv \int_0^{\overline{v}} \left[\gamma s(G,v) + (1-\gamma)s(B,v)\right] dF(v) \tag{2}$$

When $\mu(s) \in (0,1)$, we can use Bayes' rule to calculate the probability of T's type, good or bad, conditional on the statement made about T. On the other hand, when $\mu(s) \in \{0,1\}$, it follows that S is playing a strategy where he (almost) always avoids disparaging (0) or disparages (1) T, in which case Bayes' rule cannot be used to calculate the probability of T being a particular type, conditional on the strategy which is (almost) never played by S. Thus, we denote both possibilities, as follows:

$$\Gamma(t=G|1,s) \equiv \begin{array}{ll} \gamma\frac{\int_0^{\overline{v}} s(G,v)dF(v)}{\mu(s)} & \text{if} \quad \mu(s) \neq 0 \\ \Upsilon & \text{otherwise} \end{array} \tag{3}$$

$$\Gamma(t=G|0,s) \equiv \begin{array}{ll} \gamma\frac{\int_0^{\overline{v}} (1-s(G,v))dF(v)}{1-\mu(s)} & \text{if} \quad \mu(s) \neq 1 \\ \Upsilon & \text{otherwise} \end{array} \tag{4}$$

Here, the symbol Υ indicates that the strategy in question is (almost) never chosen by the speaker.

Given this notation we may characterize PBE as an assessment consisting of the strategy profile a^*, s^* and p^* along with a set of beliefs x_0^* and x_1^*, which satisfies the following four requirements.

Requirement 1 (R1): *A has no profitable deviation given its beliefs. Let $\widehat{x} \equiv \frac{b}{g+b}$, then:*

$$\begin{array}{llll} a^*(z)=0 & \text{if} & x_z > \widehat{x} & \text{for } z \in \{0,1\} \\ a^*(z)=1 & \text{if} & x_z < \widehat{x} & \text{for } z \in \{0,1\} \end{array} \tag{5}$$

Here, $\widehat{x}$ represents A's risk threshold for engaging with T. Requirement 1 states that A interacts with T only if A believes, given S's statement, that the probability that T is a good type exceeds the threshold probability of $\widehat{x}$. Similarly, if A believes that T is a good type with a probability that is lower than $\widehat{x}$, A does not interact with T. In the exceptional case where $x_z = \widehat{x}$, A is indifferent between interacting with T and not, and, thus it may play either strategy. When $\gamma = \widehat{x}$, this possibility is realized in all equilibria where the audience disregards newly acquired information in forming its beliefs. This unnecessarily complicates formal derivations and makes expositions more difficult. Therefore, in the remainder of our analysis we ignore these exceptional cases by assuming that $\gamma \neq \widehat{x}$, but the analysis can easily be extended to this case.

Requirement 2 (R2): *T has no profitable deviations in sub-games:*

$$p^*(t) = \begin{array}{llll} 0 & \text{if} & q_t d < l/2 & \text{for } t \in \{B,G\} \\ 1 & \text{if} & q_t d > l/2 & \text{for } t \in \{B,G\} \end{array} \tag{6}$$

Requirement 2 states that the PBE strategy of T must be such that in subgames where S disparages him, T litigates whenever the costs of doing so ($l/2$) are

lower than the expected damage rewards that he can obtain from litigation. Conversely, T chooses not to litigate when the costs are higher than expected damages. In the exceptional case where $q_t d = l/2$, T is indifferent between litigating and not.

Requirement 3 (R3): *S has no profitable deviations: For all t, v pairs, $s^*(t, v)$ maximizes player S's payoff, which can be expressed as*

$$U_S \equiv a^*(s(t,v))(v - p^*(t)s(t,v)\{q_t d + \frac{l}{2}\}) \tag{7}$$

The requirement with respect to S appears more complex than the requirements that pertain to T and A's strategies, because S chooses her actions in anticipation of the other players' actions. Still, the requirement is simply that, given her own type, T's type, and the anticipated behavior of A and T, S must choose the course of action that would maximize her payoff.

Requirement 4 (R4): *A's beliefs are consistent:*

$$x_z^* = \Gamma(t = G|z, s^*) \textit{ whenever } \Gamma(t = G|z, s^*) \neq \Upsilon \textit{ for both } z \in \{0, 1\} \tag{8}$$

Requirement 4 simply states that A's beliefs must be consistent with the implied conditional probability of T being a particular type based on the equilibrium strategy of S. This requirement is applicable only to strategies which have a positive probability of being played by S.

Our analysis reveals that there are two types of assessments which satisfy requirements 1-4, i.e. two types of equilibrium. One, in which the speaker's statements have no bearing on the audience's behavior, in the sense that they do not cause the audience to change their behavior relative to what they would have done if they relied only on their priors. Because the speaker's statement has no effect on audience's behavior, we term these PBE *Ineffective Communication Equilibria.* By contrast, when statements may affect behavior, the resulting PBE are dubbed *Effective Communication Equilibria.* To avoid any ambiguities in our usage of these terms, we define these two types of equilibria, as follows.

Definition 1: *A PBE is an effective communication equilibrium if, and only if, there exists $z \in \{0, 1\}$ such that $a^*(z) = \frac{\hat{x} - min\{\gamma, \hat{x}\}}{\hat{x} - \gamma}$ and $\mu^*(s^*) \neq 1 - z$.*

In classifying equilibria, we use these new definitions, instead of concepts like *babbling equilibria* and *informative equilibria*, because, although these concepts are related to our defined categories, they differ from each other in meaningful ways. Specifically, although all babbling equilibria are ineffective communication equilibria, the converse is not true. This can be seen by noting that, in some equilibria, S can play type-dependent strategies which do not impact the behavior of A. These equilibria would not fit the definition of babbling equilibria, but would not cause a change in A's behavior compared to babbling equilibria. Since we are interested in classifying equilibria based on

behavior, we rely on our *behavior-based* definition of effective communication equilibria.

3.4 Impact of Defamation Laws on Equilibrium Behavior

By using Requirements 1-4 we identify and interpret the PBE obtained with different damages, through the help of four propositions, below. Our observations can be briefly summarized as follows. Proposition 1 shows that, regardless of the level of damages, there are always ineffective communication equilibria where A acts according to its priors, i.e., where A essentially ignores the content of Ss statement. In these equilibria, parties cannot effectively communicate private information. In fact. when defamation laws are extreme, i.e. either too lax or too strict, ineffective communication equilibria are the only PBE of the game, as we note via Proposition 2. Only moderate defamation laws can engender effective communication equilibria. Then, we question whether effective communication equilibria are socially preferable to ineffective ones. The answer to this question is surprisingly ambiguous and depends in part on the accuracy of the courts. Proposition 3 shows that when courts are sufficiently accurate, it is possible to set damages moderately such that defamatory statements are fully deterred, without inviting frivolous litigation. Thus, separating equilibria are obtainable, and they are socially preferable to any other equilibria. Proposition 3 also notes that even courts which are fairly accurate, but not accurate enough to facilitate separating equilibria, can enhance welfare through moderate damages through semi-separating equilibria. Finally, Proposition 4 reveals that when the value of A's returns from interactions dwarfs other considerations, PBE associated with effective defamation laws are always socially preferable.

Proposition 1. (i) Under all defamation regimes, there exists ineffective communication equilibria. (ii) In these equilibria, A either always interacts ($\gamma > \widehat{x}$) or never interacts ($\gamma < \widehat{x}$) with the target, and litigation never takes place.

Proof. (i) The assessment consisting of $x_1^* = x_0^* = \gamma$,
$a^*(z) = \begin{array}{lll} 0 \text{ for all } z & \text{if} & \gamma > \widehat{x} \\ 1 \text{ for all } z & \text{if} & \gamma < \widehat{x} \end{array}$;
$s^*(t, v) = 0$ for all v and t;
and $p^*(t) = \begin{array}{llll} 0 & \text{if} & q_t d \leqslant l/2 & \text{for all } t \in \{B, G\} \\ 1 & \text{if} & q_t d > l/2 & \text{for all } t \in \{B, G\} \end{array}$ satisfies Requirements 1-4, and thus constitutes a PBE where A acts based on its priors.
(ii) By definition, in ineffective communication equilibria A acts according to its priors, and, thus it always interacts if $\gamma > \widehat{x}$ and never interacts if $\gamma < \widehat{x}$. In the former case, litigation never takes place as there is always interaction. In the latter case, if $a^*(0) = 1$, S could profitably deviate from her strategy by never defaming since this would save her litigation costs. Thus, it must be the case that $a^*(0) = 0$, which is possible only if $\mu(s^*) = 1$ since, by definition, interaction never takes place. But then S can profitably deviate from her

strategy s^* by choosing not to defame whenever $t = G$ and $v < q_G d + \frac{l}{2}$. Thus, litigation cannot be taking place in an ineffective communication equilibrium. □

Proposition 1 reveals that it is always possible in equilibrium for the audience to act according to its priors. Given this response by A, S has nothing to gain by disparaging the potential plaintiff, because her statements have no effect on A's behavior, yet it may cause T to initiate a lawsuit. Thus, no litigation can be observed in such equilibria.

Next, we turn to the question of whether defamation laws can cause A to change its behavior relative to its behavior based on its priors. Because the answer to this question depends on d, the magnitude of damages, it is worth identifying four critical damage levels which play a key role in the interpretation of results. Figure 2 below depicts these levels.

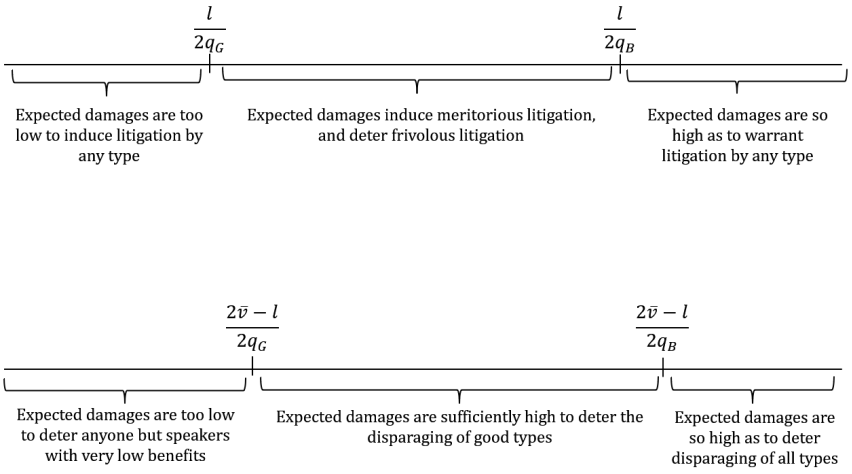

Figure 2 Critical levels of damages

The upper line depicts the first two levels ($\frac{l}{2q_G}$ and $\frac{l}{2q_B}$) which relate to the potential plaintiff's incentives, whereas the second line includes the other two levels ($\frac{2\overline{v}-l}{2q_G}$ and $\frac{2\overline{v}-l}{2q_B}$) which relate to the speaker's incentives. These levels are depicted on two separate lines because, absent further assumptions, two of these values (namely $\frac{l}{2q_B}$and $\frac{2\overline{v}-l}{2q_G}$) cannot be unambiguously ranked. We can, however, note that the critical values that relate to the speaker's incentives are greater than the corresponding critical values that relate to the target's incentives (i.e. $\frac{l}{2q_i} < \frac{2v-l}{2q_i}$ for $i \in B, G$), given our assumption that litigation costs are not prohibitively high, i.e. $l < \overline{v}$.

We observe that when damages are low, i.e., $d < \frac{l}{2q_G}$, as Figure 2 notes T lacks the incentives to bring suit even when he is falsely disparaged. When damages are very high, i.e., $d > \frac{2v-l}{2q_B}$, it follows that all effective disparaging

statements are deterred.[9] Thus, in neither case do statements have an impact on the audience's behavior. We distinguish between these *extreme damages* (i.e., $d \notin \left[\frac{l}{2q_G}, \frac{2\overline{v}-l}{2q_B}\right]$) and *moderate damages* (i.e., $d \in \left[\frac{l}{2q_G}, \frac{2\overline{v}-l}{2q_B}\right]$.) The above observations highlight that extreme damages can only lead to ineffective communication PBE. A question that remains is whether moderate damages can lead to effective communication equilibria. Proposition 2 answers this question affirmatively and formalizes related observations.

Proposition 2. (i) Extreme defamation laws only generate ineffective communication equilibria. (ii) Effective communication equilibria can be obtained only when the audience acts consistently with the speaker's statement, i.e. $a^*(z) = z$. (iii) There are moderate defamation laws which generate effective communication equilibria.

Proof. See Appendix. □

Proposition 2 holds that extreme defamation laws only allow for ineffective communication equilibria, and, as noted in proposition 1, these equilibria also exist under moderate defamation laws. However, moderate defamation laws also generate effective communication equilibria. This implies that switching from an extreme defamation law regime to a moderate regime can expand the types of equilibria that may be obtained. Thus, it becomes important to compare the properties of the two types of equilibria to ascertain their welfare impacts, among other things. This comparison hinges on how *accurate* the court is in returning correct verdicts. By accuracy, we mean the following:

Definition 2 *(i)* $\frac{q_G}{q_B} \in (1, \infty)$ *measures the courts' accuracy. (ii)* $\pi \equiv \frac{2\overline{v}}{l} - 1$ *is a critical level of court accuracy used to evaluate the potential welfare impacts of defamation laws.*

We report the relationship between the court's accuracy, as defined above, and the PBE obtainable, as follows.

Proposition 3. (i) Separating Equilibrium: When the court is sufficiently accurate (i.e. $\frac{q_G}{q_B} \geqslant \pi$) there are moderate defamation laws associated with PBE where: S disparages T if, and only if, he is a bad type; the audience acts consistently with this information (i.e. $a^*(z) = z$); and there is no litigation. (ii) Separating equilibria lead to greater expected welfare than all other equilibria. (iii) When the court is insufficiently accurate (i.e. $\frac{q_G}{q_B} < \pi$), all equilibria involve a positive likelihood with which the audience does not interact with a good type, interacts with a bad type, or both. (iv) When the court is only

9. We intentionally refer to the deterrence of *effective* disparaging statements, because there could be equilibria where the audience disregards disparaging comments and interacts with T, and, in such instances, disparaging comments would not be deterred because they would not give rise to litigation.

slightly inaccurate, i.e. $\pi - \frac{q_G}{q_B} > 0$ is sufficiently small, there exist moderate defamation laws which generate equilibria that lead to greater welfare than those generated by ineffective communication equilibria.

Proof. See Appendix □

Intuitively, when courts are sufficiently accurate it ought to be possible to set damages large enough to deter defamatory statements without generating frivolous lawsuits. When $\frac{q_G}{q_B} \geqslant \pi$, this is in fact the case, because the ambiguous ranking between the critical damage levels depicted in Figure 2, $\frac{l}{2q_B}$ and $\frac{2\overline{v}-l}{2q_G}$, vanishes, and it follows that $\frac{2\overline{v}-l}{2q_G} < \frac{l}{2q_B}$, as depicted in Figure 3 below. Therefore, by choosing damages in between these two threshold values, i.e. $d \in \left(\frac{2\overline{v}-l}{2q_G}, \frac{l}{2q_B}\right)$, one can achieve two important goals at once: deter defamation as well as frivolous lawsuits.

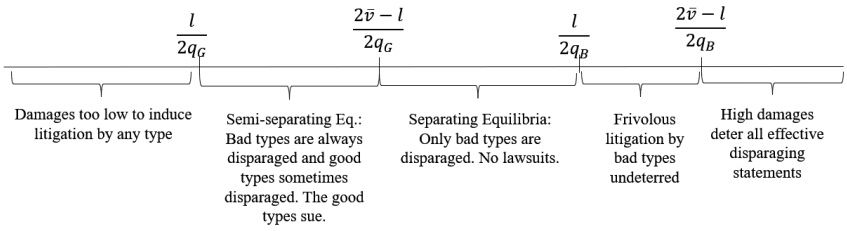

Figure 3 Critical levels of damages

Separating equilibria that achieve these two goals at once naturally maximize welfare, because (1) they lead to interactions only when these interactions enhance welfare and (2) there are no litigation costs. This reasoning extends to the case where the court is only somewhat accurate through a simple continuity argument. In this case, moderate defamation laws are associated with semi-separating equilibria, wherein a non-disparaging comment reveals that T is a good type, but where good types face a very small likelihood of being disparaged. These equilibria lead to only slightly lower expected welfare than separating equilibria and, thus, are associated with greater expected welfare than ineffective communication equilibria.

Propositions 1-3, together, reveal that when courts perform well in distinguishing good and bad types, moderate defamation laws can be used with relative ease to enhance welfare and to increase the informational value of statements made by speakers. In these cases, (semi-)separating equilibria lead to obvious and unambiguous improvements compared to equilibria where the audience is left to use its priors to make decisions. In practice, however, there are many cases where there is expressed concern among judges and lawyers that discovering the truth is difficult and that litigation is fraught with inaccuracies. The analysis in the next section thus focuses on these situations.

4. Dynamics when Courts are Inaccurate

The previous section explained why it is impossible to obtain separating equilibria if courts are inaccurate. As noted in Proposition 3, this implies that with some positive probability either interactions with good types are deterred (i.e. type-1 errors), interactions with bad types are undeterred (type-2 errors), or both. In these cases, using stricter defamation laws (i.e. higher d) can generate a trade-off between costs associated with these two types of errors and may also impact expected litigation costs. In this section, we describe these trade-offs. We focus exclusively on moderate defamation laws and the impact of changing d on effective communication equilibria because, as noted in Proposition 3, in all other cases the audience acts according to its priors and no litigation takes place. Subsequently, we identify a sufficient condition under which achieving effective communication equilibria through moderate damages continues to be socially preferable to having extreme defamation laws.

To explain the dynamics that emerge, we first start by calculating the equilibrium beliefs, i.e. x_0^* and x_1^* that would emerge in a PBE where $a^*(z) = z$, assuming that such an equilibrium exists. We plot these beliefs in Figure 4, below, through a specific but representative example. The horizontal axis represents damages, on which we mark the four critical damages levels listed in Figure 2. This time, however, the court's accuracy is lower than π, so the ranking of the intermediate critical damages (i.e. $\frac{l}{2q_B}$ and $\frac{2\overline{v}-l}{2q_G}$) is the opposite of that depicted in Figure 3. In addition to plotting beliefs, i.e., x_0^* and x_1^*, in Figure 4 we also plot the ex-ante probability of T being disparaged in these PBE. These are labeled δ_G and δ_B for good types and bad types, respectively. Next, we explain how these expressions are derived in the three relevant ranges of damages.

(1) In the range $(\frac{l}{2q_G}, \frac{l}{2q_B})$, damages are too low to incentivize bad types to sue. Thus, S faces no consequences from disparaging bad types. Whereas good types will bring a lawsuit, S might still disparage them if its benefit from blocking an interaction is sufficiently high, i.e., $v > v_G \equiv q_G d + \frac{l}{2}$. Thus, a bad type is disparaged with certainty, i.e. $\delta_B = 1$, and a good type may or may not be disparaged with positive probability. From A's perspective this means that a person who is not disparaged is definitely a good type, i.e. $x_0^* = 1$, while a target who is disparaged may or may not be a good type, but is no more likely to be a good type than a random draw from the population, i.e. $x_1^* < \gamma$. The ex-ante probability with which S draws a benefit that is higher than v_G is $\delta_G = 1 - F(v_G)$, and, thus, this is the probability with which a good type is disparaged. Using this expression, x_1^* can be more precisely expressed as $x_1^* = \frac{\gamma\delta_G}{\gamma\delta_G + 1 - \gamma} < \gamma$.

(2) In the range $(\frac{l}{2q_B}, \frac{2\overline{v}-l}{2q_G})$, damages are sufficient to trigger frivolous suits by bad types who are disparaged. The threat of a suit causes the speaker to refrain from disparaging even a bad type, unless her benefit from blocking an interaction is sufficiently high. Still, the minimum benefit that leads a speaker to disparage a bad type, $v_B \equiv q_B d + \frac{l}{2}$, is lower than the minimum benefit that would make her disparage a good type, $v_G = q_G d + \frac{l}{2}$, as frivolous claims

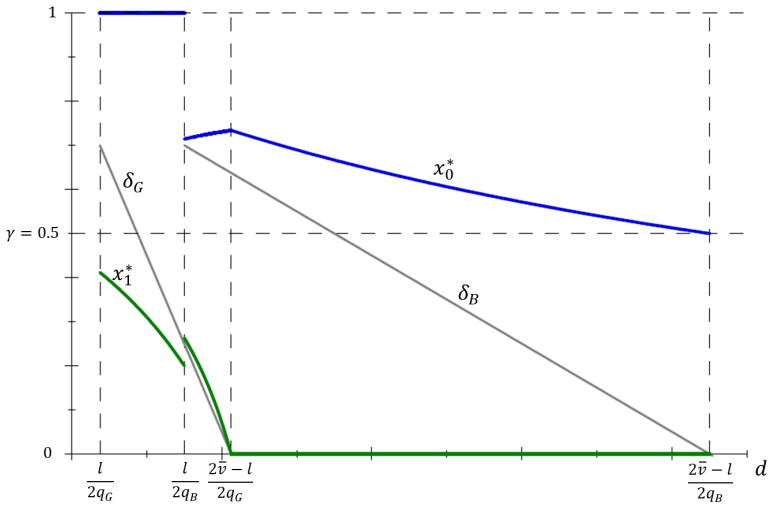

Figure 4. Illustration of Beliefs and the Likelihood of a Disparaging Statement. Damages = d, x_0^*, x_1^* are beliefs. $q_G = 0.8, q_B = 0.2, l = 0.3$, and $F(v) = v$ with support $(0, 1]$.

are less likely to be successful. Thus, the ex-ante probability with which S disparages a bad type, $\delta_B = 1 - F(v_B)$, is greater than the likelihood with which she disparages a good type, $\delta_G = 1 - F(v_G)$. Consequently, in this range, $x_0^* = \frac{\gamma(1-\delta_G)}{\gamma(1-\delta_G)+(1-\gamma)(1-\delta_B)} > \gamma > x_1^* = \frac{\gamma\delta_G}{\gamma\delta_G+(1-\gamma)\delta_B}$ as $\delta_G < \delta_B$.

(3) In the range $(\frac{2\overline{v}-l}{2q_G}, \frac{2\overline{v}-l}{2q_B})$, damages are sufficiently high to deter S from disparaging good types, even if her benefit from blocking interactions is maximal, i.e. $\overline{v}$. She will only disparage bad types if her benefit from blocking an interaction is sufficiently high. Thus, in this range: $x_1^* = 0 < \gamma < \frac{\gamma}{\gamma+(1-\gamma)(1-\delta_B)} = x_0^*$ and $\delta_G = 0 < \delta_B = 1 - F(v_B)$.

This brief analysis, and its depiction in Figure 4 can be used to identify some of the welfare implications of altering the level of damages. In the lower-moderate range (i.e. $(\frac{l}{2q_G}, \frac{l}{2q_B})$) damages are insufficient to completely prevent disparaging remarks against good types, but they are also low enough to deter frivolous litigation by bad types, leading to their disparagement. Thus, in this range, the only impact of increasing damages is to reduce the number of good types being disparaged. This reduction consequently reduces expected litigation costs, and increases the likelihood of interactions with good types. Therefore, increasing damages in this range monotonically enhances welfare, because interactions with good types are socially desirable, and litigation costs reduce welfare.

In the intermediate-moderate range (i.e. $(\frac{l}{2q_B}, \frac{2\overline{v}-l}{2q_G})$) damages are large enough to induce frivolous litigation by bad types, but not large enough to completely deter disparaging remarks against good types. Therefore, increasing damages

in this range generates more meaningful trade-offs by increasing the likelihood of beneficial interactions as well as harmful interactions, while reducing the likelihood of litigation. Thus, it is desirable to increase damages in this range only if the savings from lower litigation costs and the increased value of beneficial interactions exceed the cost involved with harmful interactions. Absent more restrictive assumptions, one cannot unambiguously compare these benefits and costs, because their magnitudes depend, in part, on the marginal changes in δ_B and δ_G, which can take many forms depending on the shape of the distribution of speaker benefits (i.e. $F(v)$).

Finally, in the higher-moderate range (i.e. $(\frac{2\overline{v}-l}{2q_G}, \frac{2\overline{v}-l}{2q_B})$), damages are high enough to deter disparaging comments against all good types, but are not sufficiently high to deter disparaging remarks against bad types. An increase in damages in this range causes an increase in the expected costs from harmful interactions, but reduces litigation costs. Thus, as long as litigation costs are lower than the gains from blocking harmful interactions, social welfare is improved by *reducing* damages in this range.

This analysis reveals the complex nature of trade-offs involved when the court is inaccurate in making decisions. There is no general reason why higher damages would be better than lower damages. Courts and policymakers must account for domain-specific considerations which can tilt the balance in any given direction.

A somewhat counterintuitive conclusion is that, with inaccurate courts, it is not even true in general that one can improve upon ineffective communication equilibria where the audience acts upon its priors. Moving to an equilibrium where the audience acts consistently with the information it receives from the speaker can be helpful in promoting beneficial interactions or dissuading harmful ones. However, it comes at the cost of increased litigation, and may reduce the target's benefit from increased missed interactions or the speaker's benefit from blocking interactions. An aspect of this analysis is that higher damages in the moderate range sometimes sacrifice the well-being of some good types, thus calling into question a widely-held belief among lawyers that stronger defamation laws protect good types (Arbel & Mungan, 2019, Hemel & Porat, 2019).

Adding to these complexities is the fact that, given any damage level, d, effective communication PBE are possible only if A's risk tolerance, $\hat{x}$, lies in between x_1^* and x_0^*, as depicted in Figure 4. Despite these ambiguities, one can always use moderate damages that lead to effective communication PBE. Thus, one can improve the odds of beneficial interactions taking place and/or harmful interactions not taking place. Thus, if the audience's well-being is the predominant consideration in the welfare analysis, it follows that moderate damages can always improve upon extreme damages. The next proposition formalizes this result.

Proposition 4. There exist moderate damages leading to effective communication equilibria, which generate greater welfare than ineffective communication equilibria, as long as g and b are large relative to other costs and benefits.

Proof. The expected pay-off of the audience in an equilibrium where $a^*(z) = z$ for all z is

$$\overline{U}_A = \gamma(1-\delta_G)g - (1-\gamma)(1-\delta_B)b \tag{9}$$

On the other hand, 0 and $\gamma g - (1-\gamma)b$ are the expected pay-offs that the audience would have received by acting according to its priors, when $\gamma < \widehat{x}$ and $\gamma > \widehat{x}$, respectively. In these PBE, it follows that $\overline{U}_A = \gamma(1-\delta_G)g > 0$ when $d \in (\frac{l}{2q_G}, \frac{l}{2q_B})$, and, similarly, $\overline{U}_A = \gamma g - (1-\gamma)(1-\delta_B)b > \gamma g - (1-\gamma)b$ when $d \in (\frac{2\overline{v}-l}{2q_G}, \frac{2\overline{v}-l}{2q_B})$. Thus, for any given $\widehat{x}$, the increase in the expected pay-off of the audience stemming from a move from a PBE where it acts according to its priors to one where it acts according to the information it receives from the speaker is linearly increasing in g and b, respectively. Moreover, the magnitudes of g and b only affect A's payoff, and, hence, there exist large enough g and b which cause these PBE to generate greater welfare than PBE where the audience acts according to its priors. □

Proposition 4 reveals that when the value of interactions are large in comparison to other considerations, like litigation costs and the benefits that the speaker gets from blocking interactions, moderate defamation laws can be used to enhance welfare. This is because, under these conditions, the dominant consideration becomes the maximization of the audience's pay-off, which benefits from having effective communication equilibria.

5. Discussion

In Sections 3 and 4, we provided a model that allowed us to clearly focus on defamation laws' impact on the audience's equilibrium beliefs and actions. In doing so, we abstracted from many issues that bear on the regulation of information in more general settings, particularly, the possibility of a committed public enforcer, quality being endogenously chosen by the target, and the existence of honest and other types of speakers. Here we turn our attention to these issues.

5.1 Endogenous Types and Dynamic Efficiencies

In our analysis thus far, we assumed that the target's type t was exogenously determined by nature to be either G or B with probabilities γ and $1-\gamma$, respectively. One might question the reality of this assumption, as people can make investments that would make them better or worse trading partners, e.g., create higher quality products, maintain safety standards, or keep higher hygiene standards.

One option of incorporating quality investments into our analysis is to replace Nature's choice of types with a preliminary stage where the target, T, makes a costly investment (c) that can increase her likelihood of becoming a good type. Formally, we may assume that $\gamma = \gamma(c)$ with $\gamma' > 0 > \gamma''$, $\lim_{c\to 0}\gamma'(c) = \infty$, $\gamma(0) = \underline{\gamma}$ and $\lim_{c\to\infty}\gamma(c) = \overline{\gamma}$ where $1 > \overline{\gamma} > \underline{\gamma} > 0$. Moreover, to keep the description of this extension brief, we focus on the case where

$\underline{\gamma} > \widehat{x}$.

The quality investment decision is now part of a larger game. Given any sub-game equilibrium, the best response of T is to make an investment to maximize his expected pay-off, which can be denoted as $\gamma(c)m_G + (1-\gamma(c))m_B - c$ where m_G and m_B refer to the pay-offs he obtains in the sub-game equilibria.

This observation reveals a very clear result: When the laws are extreme, i.e. $d \notin \left[\frac{l}{2q_G}, \frac{2\overline{v}-l}{2q_B}\right]$, the target has no reason to invest in quality. This follows from Propositions 1 & 2, which show that with extreme laws, the audience acts based on its priors and interacts with the target if γ is sufficiently high. Thus, investments have no private returns for the target.

It is only when the laws are moderate that targets may have an incentive to invest in quality. This can be demonstrated by focusing on the lower bound of intermediate damages, i.e. $\frac{l}{2q_G}$. In this case, in PBE with $a^*(z) = z$, it follows that $m_B = 0$ (as all bad types are disparaged) while $m_G = (1-\delta_G)r$ (because good types are disparaged with probability δ_G, in which case there is a lawsuit which pays the target expected damages equal to litigation costs). Thus, the target's pay-off is $\gamma(c)(1-\delta_G)r - c$, and, therefore, the target profits (in expectation) from investing. Whether this is socially good or bad, depends, of course, on whether there are net social gains from such investments. In our context, this is socially valuable as long as the expected benefits from good interactions ($(1-\delta_G)g$)—which are not internalized by T—are greater than the expected litigation costs l and the loss of benefit to S from blocking an interaction, i.e. $(1-\delta_G)E[v|v > \frac{l}{2}]$. In fact, if investments in quality are socially valuable, then increasing damages within the intermediate range up to $\frac{l}{2q_B}$ will be desirable. This is because these higher damages lead to a lower probability of disparaging remarks made against good types (as illustrated in Figure 4) and, thus, increase m_G, while still keeping expected payoffs from being a bad type at $m_B = 0$.

The discussion here highlights the importance of information regulation for broader market dynamics. The intuition underlying our results are straightforward. Extreme laws lead to ineffective communication equilibria. In contrast, moderate laws create an environment with more reliable information regarding types, thus generating a greater gap between the payoffs obtainable by good types versus bad types. This, in turn, increases the returns from being a good type, and leads to more investments. In realistic settings, providing such additional incentives is socially desirable when the potential investor is under-incentivized due to problems like information asymmetries. The gains from such investments in quality should be added to the other benefits of moderate laws that we have identified.

5.2 Truth Speakers and Eulogists

So far, we only considered speakers who had something to gain from severing the relationship between the audience and the target. This abstraction follows the idea of speaker's 'bias' in the cheap talk literature. In reality, however, some speakers may not have such motivations. Quite importantly, many

people, when asked their opinion, provide an honest assessment of others. Moreover, there are also people who are motivated by doing the exact opposite of what the speakers in our model are motivated by; namely, promoting the relationship between the target and the audience. In what follows we distinguish between the first type, "truth speakers," the latter type, "eulogists," and the ones we formerly discussed in Section 3 as "disparagers." We briefly, and informally, explain now what occurs when these kinds of speakers are incorporated into our analysis.

In our discussion, we conceive of these types as follows. Disparagers, as we noted, receive a positive value from blocking an interaction; truth-speakers are indifferent with respect to whether the parties will interact but receive some value from speaking their mind; and, eulogists receive a value from there being an interaction. Therefore, so long as costs of so doing are not high, disparagers will badmouth the target and truth-speakers will reveal their true type. Eulogists, in contrast, would always want to praise the target, as there is no recourse under defamation law for false positive statements (the question of why this asymmetry exists goes beyond the the scope of our article).

The incorporation of these types of speakers has no impact on the observation that extremely *strong* defamation laws leave the audience to act upon their priors. This follows, because once a critical threshold of damages is passed, disparagers as well as truth speakers are deterred from making negative remarks. Thus, extremely strong defamation laws cause disparagers, truth speakers, and eulogists alike to abstain from making negative statements, and the audience has no option but to act according to its priors.

The same cannot be said, however, for extremely *weak* defamation laws. When damages are very low, targets lack an incentive to bring suit, making talk "cheap." Despite that, disparaging statements are still somewhat informative: Given the existence of some truth-speakers, there is some probability that any negative statement is true. Consequently, an audience that hears a negative statement evaluates its credibility based on the ratio of truth-speakers to disparagers. Thus, (in an assessment where $a^*(z) = z$) we can formulate the audience's belief that the target is a good type, conditional on a negative statement as $x_1^* = \gamma \frac{\Delta}{\Delta+(1-\gamma)\tau}$ where τ denotes the proportion of truth speakers, and Δ is the proportion of disparagers. On the other hand, non-disparaging remarks do not necessarily mean that T is a good type. By similar logic, there is some probability that any given praise is false given the existence of eulogists. An audience which hears a positive statement evaluates its veracity as a function of the ratio of eulogists to truth-speakers. Thus, we can express the audience's belief as $x_0^* = \gamma \frac{\tau+\varepsilon}{\gamma\tau+\varepsilon}$, where ε is the proportion of eulogists.

Using these observations it is easy to verify that, under lax laws, both disparaging and non-disparaging statements are somewhat informative of types. In other words, non-disparaging statements are more indicative of good types than no information at all ($x_0^* > \gamma$), and disparaging statements are more indicative of bad types than no information at all, i.e. $x_1^* < \gamma$. Thus, if the audience's necessary level of confidence for interaction, ($\widehat{x}$), is close enough to γ

such that $x_0^* \geqslant \widehat{x} \geqslant x_1^*$, one can achieve an equilibrium wherein the audience meaningfully uses the information provided by speakers, even when there are no sanctions for false statements. If, however, $\widehat{x} \notin [x_1^*, x_0^*]$, then lax laws cause the audience to ignore the statement and act according to its priors, as in our analysis in Section 3. Thus, we focus our remaining discussion to cases where $x_0^* \geqslant \widehat{x} \geqslant x_1^*$.

In cases where damages are moderate, some of the claims made in Section 3 need to be qualified, whereas others remain intact. In particular, it is still the case that moderate damages improve the reliability of information over extreme damages. To see this, consider, for instance, the implications of raising damages from low levels to $\frac{l}{2q_G}$. Among speakers, this change only alters the incentives of "disparagers," because these are the only speakers who have an interest in making false statements about good types, who, given this level of damages, bring a lawsuit against them. Thus, the proportion of disparagers who make false statements is reduced, which causes x_1^* to fall and x_0^* to increase, i.e. it causes information supplied by speakers to be more informative. This observation reveals another of our results that carries over in a modified way: when courts are sufficiently accurate, one can use damages equal to $\frac{2\overline{v}-l}{2q_G} < \frac{l}{2q_B}$ to deter all disparagers from making false statements and also guarantee that there are no lawsuits by bad type targets. In this case, it immediately follows that $x_1^* = 0$, such that a disparaging statement is perfectly informative.

The presence of eulogists, however, means that $x_0^* < 1$. Thus, fully separating equilibria are no longer obtainable. Still, even in the presence of eulogists and disparagers, semi-separating equilibria are possible. Moreover, as in the previous case, these semi-separating equilibria are optimal, because they lead to no litigation costs, cause all possible good interactions to take place, and achieve maximum deterrence of bad interactions.

We conclude that the introduction of honest speakers as well as what we called eulogists—people who wish to promote the target—does not affect the superiority of moderate damages over extreme forms of damages. What does change is perhaps somewhat counterintuitive: strict laws turn out to be *worse* than lax laws. Strict laws lead to completely uninformative speech in equilibrium whereas lax laws still allow speech to be somewhat informative, permitting effective communication equilibria.

5.3 Commitment and Public Enforcement

Our analysis so far focused on private enforcement of defamation laws, where the target is the one to sue. However, private parties will only bring a lawsuit if it pays to do so ex-post, and this calculus exposes them to strategic behavior by would-be defamers. In contrast, some parties, typically public agencies, may be able to commit ex-ante to sue, even if it does not pay to do so ex-post. Comparing private and public enforcement can be useful in understanding other contexts where information is regulated, and may also illuminate the reasons why private enforcement is used in defamation.

To help in this comparison, we consider a simple modification of our analysis wherein instead of the target, it is a public enforcement agency that can bring suit against disparaging remarks. The agency, however, is not privy to the target's private information regarding his type, which is by assumption unobservable, and so it cannot condition its action on T's type. The agency thus chooses some probability, $p \in (0,1)$, with which it will bring a lawsuit. As the choice of p does not depend on any new information, it is made ex-ante and is communicated to, or observed by, would-be speakers. The choice of p replaces $p^*(t)$ in (6). We retain all other assumptions, including the assumption that the probabilities with which the speaker will be found liable in court are q_G and q_B, when she makes disparaging statements against good and bad types, respectively.

This simple modification allows us to calculate the the analogs of the two critical values which describe the best responses of S depicted in Figure 2. Specifically, these two critical values now become $\frac{2\bar{v}-pl}{2pq_G}$ and $\frac{2\bar{v}-pl}{2pq_B}$. Thus, in effective communication equilibria, when $d > \frac{2\bar{v}-pl}{2pq_G}$, the speaker does not make disparaging statements against good types, and refrains from making disparaging statements against bad types when $d > \frac{2\bar{v}-pl}{2pq_B}$. It can be easily verified that each of these values is larger than their corresponding analog in the private enforcement context, i.e. $\frac{2\bar{v}-pl}{2pq_i} > \frac{2\bar{v}-l}{2q_i}$ for $i \in \{B,G\}$.

The commitment to bringing a lawsuit also changes the speaker's behavior, as a lawsuit is possible even when expected damages are low. We next explain the behavior of the speaker in effective communication equilibria, under three different damages ranges, and subsequently compare them with the corresponding behavior under private enforcement.

As under private enforcement, it follows that when damages are very high, $d > \frac{2\bar{v}-pl}{2pq_B}$, all disparaging remarks are deterred. However, when damages are moderate, $d \in (\frac{2\bar{v}-pl}{2pq_G}, \frac{2\bar{v}-pl}{2pq_B})$, the speaker refrains from disparaging good types, but disparages bad types whenever her value from blocking interactions is sufficiently high (i.e. $\tilde{v}_B \equiv p(q_B d - \frac{l}{2}) < v$) which happens with probability $\tilde{\delta}_B \equiv 1 - F(\hat{v}_B) > 0$ (The tilde sign refers to analogs of values defined in the private enforcement context). Thus, in the moderate range, a disparaging remark conclusively reveals to the audience that the target is a bad type; a non-disparaging comment is an informative, but inconclusive, signal that the target is a good type, i.e. $x_1^* = 0 < \gamma < x_0^*$. When damages are low, i.e., $d < \frac{2\bar{v}-pl}{2pq_G}$, the speaker is no longer necessarily deterred from disparaging good types, and chooses to defame the target if her value from blocking interactions exceeds $\tilde{v}_G \equiv p(q_B d - \frac{l}{2})$. Thus, it follows that $0 < \tilde{\delta}_G < \tilde{\delta}_B$, and, therefore, $0 < x_1^* < \gamma < x_0^* < 1$.

We can now compare defamation laws under public and private enforcement regimes. First, effective communication equilibria are not possible under either regime when damages are extremely high (i.e. higher than $\frac{2\bar{v}-pl}{2pq_B}$ and $\frac{2\bar{v}-l}{2q_B}$ in the public and private regimes, respectively). Thus, our previous conclusion regarding the ineffectiveness of high damages in supporting informative statements extends to the public enforcement case as well.

Second, unlike private enforcement, public enforcement can sustain effective communication equilibria even with low damages. Under private enforcement, the speaker will anticipate that the target will not sue if damages are sufficiently low. This can lead the speaker to disparage regardless of the target's type, which would make statements non-informative. With public enforcement, however, there is always a risk of a lawsuit, thus deterring some would-be defamers and sustaining the reliability of some statements. This implies that, unlike in the private enforcement context, very low damages can be used to support effective communication equilibria—at least when the threshold belief of the audience, i.e. $\hat{x}$, is not too far from its priors, i.e. when $|\hat{x} - \gamma|$ is not large, because then $\hat{x} \in [x_1^*, x_0^*]$. Note that this means that low damages can be superior to high damages in facilitating effective communication between the speaker and the audience.

Third, and quite importantly, it is impossible to obtain a separating equilibrium with public enforcement, regardless of how accurate the courts are: as noted above, any damages below $d < \frac{2\bar{v}-pl}{2pq_B}$ result in $\tilde{\delta}_G > 0$, $\tilde{\delta}_B < 1$, or both. This immediately implies that when courts are accurate, private enforcement dominates pubic enforcement in terms of its welfare consequences. The difference in the welfare obtainable under the two regimes is enhanced further by the fact that under public enforcement, the enforcement agency's commitment results in some litigation whenever defamation laws are effective (i.e. $\frac{2\bar{v}-pl}{2pq_B} > d$).

The last point highlights a more general and important advantage of private enforcement over public enforcement. Specifically,private enforcement delegates the decision to litigate to the party with the best information about the merits of the case. Moderate damages can be crafted to separate good and bad types based on their willingness to sue, and this enables the speaker's statements to be more informative of the target's type.

In sum, this comparison illuminate the relative value of public versus public enforcement. However, as our focus here is on commitment, we abstract from other relevant considerations, such as the relative costs of learning about disparaging remarks or producing evidence. Inasmuch as public agencies employ discretion, they are also susceptible to capture and other public choice problems. These considerations should also be taken into account in comparing the relative social desirability of pubic versus private enforcement in regulating speech.

5.4 Features of Defamation Law

Our analysis took the domain of potentially defamatory statements—disparaging remarks—as given. However, the framework developed here could also be used to shed light on such determinations, in particular, the fact v. opinion and perse v. pro-quod distinctions. Defamation law renders expressions of opinion non-actionable. The analysis suggests a rationale: it is harder to determine the truth-value of opinions, leading to greater judicial inaccuracy and making regulation less valuable. It is also possible that the law implicitly recognizes that

audiences are Bayesian, so that they inherently discount statements couched in the form of an opinion. The other distinction involves regular defamatory statements (pro-quod), and a category of per-se statements, which requires a lower burden of proof. Per-se statements are allegations of criminal activity, sexual misconduct, contagious disease, or improper business dealings. Again, our analysis offers a rationale: In such cases, the harm to the target and the gain to the speaker may be especially high. Consequently, stricter protection may be warranted.

5.5 Information Regulation in Other Settings

As we noted in the introduction, the model presented in sections 2 and 3 has key features which are present in many contexts, and we focused on defamation law due to its current importance. Here we discuss three other important settings where these key features are present: law enforcement, jury trials, and whistle-blowers. Then we discuss an additional context, securities regulation, where the speaker reveals information about itself. Despite this conceptual difference, the current framework proves illuminating in considering the optimal regulatory framework.

5.5.1 Bayesian Public Enforcers

News about crimes which were committed after the police chose to ignore reports of abuse and other red flags are, unfortunately, not uncommon.[10] At the same time, some people make false or frivolous reports about others.[11] Police forces have limited resources, so they need to prioritize the calls they receive and focus on those they perceive to be most credible.

One can conceive of this dynamic as similar to the one presented here. Law enforcers who receive reports have to weigh the credibility and the importance of each claim. They decide to take action only when its expected benefits are sufficiently large given enforcement costs. As such, enforcers act as the audience. The person reporting the crime is akin to the speaker, and the alleged criminal is the target.

In this context, punishing false reports has the effect of making reports more credible, as in our analysis, and allowing law enforcers to more accurately focus their enforcement efforts. This, in turn has the effect of increasing deterrence by increasing the opportunity cost to committing crime (i.e. the analog of reducing δ_G). However, if false reports are punished too severely, it will have the effect of deterring truthful reports and, thus, lead to less than ideal deterrence of the underlying crime.

10. Emma Snaith, Woman killed by ex-boyfriend after police were warned 18 times of his abuse, Independent (Aug., 16th, 2019). Joel Rose & Brakkton Booker, Parkland Shooting Suspect: A Story Of Red Flags, Ignored, NPR News (March,1 2018)

11. Swatting, e.g., is a practice of fraudulently reporting a bomb or other imminent threat coming from the victim in order to have police forces storm their residence, sometimes to tragic ends

5.5.2 Whistle-blowing A similar dilemma applies to whistleblowers. The US government sometimes issues rewards to whistleblowers (e.g., False Claims Act and the IRS Whistleblower Law) in order to encourage them to report wrongdoing despite their fears of retribution and informal sanctions (Givati, 2016). The concern is that rewards may incentivize false whistle-blowing among people who face low costs and may also fail to appropriately incentivize people with abnormally high costs. In analyzing this problem, one can think of whistleblowers as speakers and law enforcement agencies as the audience. The agency dilemma is how to set rewards and penalties in a way that would allow for the effective transmission of private information without involving too high verification and litigation costs.

5.5.3 Trials with Bayesian Juries Another potential application is liability for the filing of false charges and frivolous lawsuits. Under the common law tort doctrine of *malicious prosecution* a person who is falsely accused of a crime may bring a lawsuit against the accuser. The harm here consists of a false investigation and the reputational and dignitary harms that follow from being under criminal investigation. Somewhat similar concerns arise with the filing of frivolous lawsuits, and under Rule 11 of the Federal Rules of Civil Procedure, courts may impose financial liability on a litigant. How willing the courts should be to enforce malicious prosecution claims or issue penalties is debated, because of concern that penalties may chill innocent victims of real crimes from coming forward.

The framework developed here is useful to the analysis of these questions, especially because judges and jurors sometimes consider one's record (even when they ought not to) in assessing guilt or liability. In such contexts, punishing frivolous lawsuits moderately may have the (additional) benefit of making the trial process more accurate, and thereby amplify its deterrent effect by increasing the opportunity cost of engaging in wrongdoing. Although many additional dynamics can emerge in the trial context, especially in those resembling the bilateral accidents framework, the impact of punishing frivolous lawsuits can be re-visited from the perspective provided here by analogizing the plaintiff (or prosecutor) to the speaker, the defendant to the target, and the jury to the audience.

In fact, the framework provided in (Freidman & Wickelgren 2005) can be used to evaluate the optimal penalties in fighting frivolous claims. In their article, Friedman and Wickelgren consider a context wherein jurors form beliefs regarding claims made against a defendant based on the evidence that is presented at trial. They use their setting to establish an upper bound on deterrence, but they also find that this upper bound depends on the quality of the evidence presented to jurors. The frequency of frivolous claims impacts the accuracy with which jurors form opinions, and, thus, reducing it ought to increase the upperbound on deterrence. But, of course, penalizing frivolous claims too severely can have the impact of deterring legitimate claims, which will have the opposite effect. Thus, as in our setting, the optimal penalty for frivolous

claims would have to be moderate and balance these two considerations.

5.5.4 Securities Regulation Public companies are required to disclose periodical reports about their performance to the public. These reports affect the propensity of investors to deal with the reporting company, and the goal of securities regulation is to regulate the accuracy of these reports given the inherent moral hazard companies have to distort information.

This context is similar to the framework developed here, where the audience consists of prospective investors, the company takes the place of the speaker, and the regulator assumes the position of the target (in deciding whether to bring a lawsuit).The question of optimal damages d, is akin to asking how strict the agency should be in its enforcement of the law, as well as the level of fines that it issues. One immaterial difference in this context is that the speaker makes statements about itself, rather than another party. The second and related difference is that the speaker-company would normally not want to disparage itself; rather, it would seek to praise itself. This difference, however, has little analytical significance, as it simply involves reversing the labels in our initial analysis.

Applying the framework at hand to securities regulation could reveal, for example, why strict and lax enforcement is inferior to more moderate enforcement. It could also be useful in highlighting the importance of making information revealed by companies *actionable* and the conditions under which it is desirable to do so. Yet another potential insight concerns the importance of understanding judicial competency in any given area of disclosure and its relevance to the level of information regulation.

6. Conclusion

The law regulates information dissemination in a variety of contexts. Work in this area has tended to focus on the effect of such regulation on speakers and their targets, and has not paid much attention to audience effects. In this article we highlight the importance of audience effects by showing that in the presence of Bayesian audiences, stricter regulation of information may jeopardize its value. While lax regulation results in non-credible "cheap talk," strict regulation can result in equally uninformative "overpriced talk."

7. Appendix

Proof of Proposition 2 The proof begins with part (ii), which is used in proving part (i).

(ii) We proceed by demonstrating that the only equilibria where the actions of the audience are *not* described by $a^*(z) = z$ for all z are (1) those where the audience ends up always interacting when $\gamma > \widehat{x}$, and (2) those where the audience ends up never interacting when $\gamma < \widehat{x}$.

(1) $\gamma < \widehat{x}$*:*

Suppose there is a PBE where $a^*(z) = 0$ for all z. Then, $x_0^*, x_1^* \geqslant \widehat{x}$ per R1. But, if $\mu = i \in \{0, 1\}$, then it immediately follows that $\Gamma(t = G|i, s^*) =$

$\gamma < \widehat{x} \leqslant x_i^*$, which is a violation of R4. On the other hand, if $\mu(s^*) \in (0,1)$, observe that, per R4, $x_0^* > \gamma$ implies that $x_1^* < \gamma$, because $x_0^*(1-\mu(s^*)) + x_1^*\mu(s^*) = \gamma$. Thus, $x_0^* \geqslant \widehat{x} > \gamma$ implies that $\widehat{x} > \gamma > x_1^*$, which is a contradiction with R1's implication that $x_1^* \geqslant \widehat{x}$.

Suppose there is a PBE where $a^*(z) = 1 - z$ for all z. Then, per R3, $s^*(t,v) = 0$ for all v and t, and, thus, the audience never interacts in such assessments.

Suppose there is a PBE where $a^*(z) = 1$ for all z. By definition, the audience never interacts in such assessments.

(2) $\gamma > \widehat{x}$:

Suppose there is a PBE where $a^*(z) = 0$ for all z. By definition, the audience always interacts in such assessments.

Suppose there is a PBE where $a^*(z) = 1 - z$ for all z, then per R3, $s^*(t,v) = 0$ for all v and t, and, therefore, $\mu(s^*) = 0$, which implies that $\Gamma(t = G|0, s^*) = \gamma$. This implies via R4 that $x_0^* = \gamma$, which, in turn implies via R1 that $a^*(0) = 0$, which contradicts the assumption that $a^*(0) = 1$.

Suppose there is a PBE where $a^*(z) = 1$ for all z. If $\mu(s^*) = i \in \{0,1\}$, then $\Gamma(t = G|i, s^*) = \gamma$, which implies via R4 that $x_i^* = \gamma$. This implies via R1 that $a^*(i) = 0$, which is a contradiction with the initial supposition. If, on the other hand, $\mu(s^*) \in (0,1)$, observe that, per R4, $x_0^* \leqslant \gamma$ implies that $x_1^* \geqslant \gamma$, because $x_0^*(1-\mu(s^*)) + x_1^*\mu(s^*) = \gamma$. Thus, $x_0^* \leqslant \gamma$ implies that $x_1^* \geqslant \gamma > \widehat{x}$, which is a contradiction with the implication of R1 that $x_1^* \leqslant \widehat{x}$.

(1) and (2) together demonstrate that all PBE where the actions of the audience are not described by $a^*(z) = z$ for all z involve the audience behaving according to its priors.

(i) Consider damages $d < \frac{l}{2q_G}$, and suppose $a^*(z) = z$ for all z. It follows via requirement 2 that $p^*(t) = 0$ for all t. Thus, R3 implies that $s^*(t,v) = 1$ for all v and t, and, therefore, $x_1^* = \gamma$ due to R4 . Thus, in equilibrium, the audience acts according to its priors.

Next, consider damages $d > \frac{2\overline{v}-l}{2q_B}$. It follows per R2 that $p^*(t) = 1$. Thus, per R3, $s^*(t,v) = 0$ for all v and t, because $d > \frac{2\overline{v}-l}{2q_B}$. This implies via R4 that $x_0^* = \gamma$. Thus, in equilibrium, the audience acts according to its priors. The analysis of these two cases demonstrates that when $d \notin \left[\frac{l}{2q_G}, \frac{2\overline{v}-l}{2q_B}\right]$, in all PBE where $a^*(z) = z$ for all z, the audience acts according to its priors. In addition, part (ii) of this proposition demonstrates that the audience acts according to its priors in all PBE where the audience's behavior is not described by $a^*(z) = z$. Thus, whenever $d \notin \left[\frac{l}{2q_G}, \frac{2\overline{v}-l}{2q_B}\right]$, the audience acts according to its priors in all PBE.

(iii) Equilibria described (and whose existence are proven) in proposition 3-(i) and section 4. demonstrate that such defamation laws exist under all circumstances.

Proof of Proposition 3 (i) Consider defamation laws with $d = \frac{2\overline{v}-l}{2q_G}$. It can easily be verified that the assesment where $a^*(z) = z$ for all z; $x_0^* = 1, x_1^* = 0$; $s^*(t,v) = \begin{cases} 1 & \text{if} \quad t = B \\ 0 & \text{if} \quad t = G \end{cases}$ for all v; and $p^*(t) = \begin{cases} 0 & \text{for } t = B \\ 1 & \text{for } t = G \end{cases}$ sat-

isfies R1-R4. In this equilibrium, there is no litigation because if $s^*(t,v)p^*(t) = 0$ for all t and v.

(ii) When $t = G$, this equilibrium leads to a total pay-off of $r+g$, and when $t = B$, it leads to a toal pay-off of v. These two values constitute the highest pay-offs that can be generated (see, e.g., figure 1) conditional on the target being a good type and a bad type, respectively, because $r+g > \overline{v} > 0 > r-b$. Thus, there can be no PBE that leads to higher pay-off.

(iii) Consider imprecise courts. If $d \notin \left[\frac{l}{2q_G}, \frac{2\overline{v}-l}{2q_B}\right]$, the audience acts according to its priors in all equilibria as proven in proposition 2, and thus it either always interacts, which leads to bad interactions with a probability of $1-\gamma$; or it never interacts, which leads to no interactions with good types with a probability of γ. If $d \in \left[\frac{l}{2q_G}, \frac{2\overline{v}-l}{2q_B}\right]$, the same result holds in all PBE except, potentially, in PBE where $a^*(z) = z$ for all z. Thus, consider next the interaction probabilities in equilibria where $a^*(z) = z$ for all z when $d \in \left[\frac{l}{2q_G}, \frac{2\overline{v}-l}{2q_B}\right]$.

(a) Suppose $d \in \left(\frac{l}{2q_G}, \frac{2\overline{v}-l}{2q_G}\right)$:

It follows per requirement 2 that $p^*(G) = 1$. Thus, per R3, $s^*(G,v) = \begin{array}{lll} 1 & \text{if} & v > q_G d + \frac{l}{2} \\ 0 & \text{if} & v < q_G d + \frac{l}{2} \end{array}$ for all v. This implies that, with probability $\gamma(1 - F(q_G d + \frac{l}{2})) > 0$, the audience does not interact with a good type in such PBE (if there exist any).

(b) Suppose $d \in \left(\frac{l}{2q_B}, \frac{2\overline{v}-l}{2q_B}\right]$:

It follows per R2 that $p^*(B) = 1$, and because $d \leqslant \frac{2\overline{v}-l}{2q_B}$ it follows that $s^*(B,v) = \begin{array}{lll} 1 & \text{if} & v > q_B d + \frac{l}{2} \\ 0 & \text{if} & v < q_B d + \frac{l}{2} \end{array}$ for all $v < \overline{v}$. This implies that with probability $(1-\gamma)F(q_B d + \frac{l}{2})$, the audience interacts with a good type in such PBE (if there exist any).

(c) Suppose $d = \frac{l}{2q_G}$:

If $p^*(G) = 1$, the same steps in part (a) imply that with probability $\gamma(1 - F(q_G d + \frac{l}{2})) > 0$, the audience does not interact with a good type in such PBE (if there exist any). If $p^*(G) = 0$, per R3, $s^*(G,v) = 1$ for all v, which implies that the audience never interacts with a good type in such PBE (if there exist any).

Thus, in all PBE obtained through moderate damages where $a^*(z) = z$ for all z , either the probability of no interaction with a good type is positive, the probability of interactions with a bad type is positive, or both.

(iv) Let $d = \frac{l}{2q_B}$. Consider an assessment where $a^*(z) = z$ for all z, and

$$\text{where} \quad \begin{array}{lll} p^*(G)=1 & & \text{and} \quad p^*(B)=0 \text{ (satisfies R2)} \\ s^*(G,v)=\left\{\begin{array}{lll} 1 & \text{if} & v>q_G d+\frac{l}{2} \\ 0 & \text{if} & v<q_G d+\frac{l}{2} \end{array}\right. & & \text{and} \quad s^*(B,v)=1 \text{ (satisfies R3);} \\ x_0^*=1 & & \text{and} \quad x_1^*=\frac{1-F(l\{\frac{q_G}{2q_B}+\frac{1}{2}\})}{1-F(l\{\frac{q_G}{2q_B}+\frac{1}{2}\})+\frac{1-\gamma}{\gamma}} \text{ (satisfies R4)} \end{array}$$

It follows that $\lim_{\frac{q_G}{q_B}\to\pi} x_1^*=0$, and, thus, for all $\widehat{x}$, there exists $\frac{q_G}{q_B}<\pi$ sufficiently close to π such that $x_0^*>\widehat{x}>x_1^*$, which guarantees that the assessment also satisfies R1, and is therefore a PBE.

It follows that the expected welfare associated with this PBE is

$$\widehat{W}=\gamma[F(l\{\frac{q_G}{2q_B}+\frac{1}{2}\})(r+g)+(1-F(l\{\frac{q_G}{2q_B}+\frac{1}{2}\}))E[v|v>\frac{l(q_G+q_B)}{2q_B}]]+(1-\gamma)E[v] \tag{10}$$

where $E[.]$ refers to expected values. It follows that

$$\lim_{\frac{q_G}{q_B}\to\pi}\widehat{W}=\gamma(r+g)+(1-\gamma)E[v] \tag{11}$$

If, $\widehat{x}>\gamma$, welfare obtained in equilibria where the audience acts according to its priors is $E[v]$, and if $\widehat{x}<\gamma$, the welfare obtained in equilibria where the audience acts according to its priors $r-b<0$. Because, $r+g>\overline{v}$, it follows that

$$\lim_{\frac{q_G}{q_B}\to\pi}\widehat{W}>E[v]>r-b \tag{12}$$

Because the first inequality is strict, there exists $\frac{q_G}{q_B}<\pi$ sufficiently close to π such that $\widehat{W}$ exceeds the welfare obtainable when the audience acts according to its priors. Thus, when courts are only slightly imprecise there is a PBE associated with $d=\frac{l}{2q_B}$ which leads to greater welfare than PBE where the audience acts according to their priors.

References

1. Arbel, Yonathan A. and Mungan, Murat. 2019. The Uneasy Case for Expanding Defamation Law. *Alabama Law Review* 1-999

2. Acheson, D. J.; Wohlschlegel, A. (2018). The economics of weaponized defamation lawsuits. *Southwestern Law Review*, 47(2), 335-384.

3. Bénabou, Roland, and Jean Tirole. 2006. Incentives and Prosocial Behavior. *American Economic Review* 96.5, 1652-1678.

4. Bénabou, Roland, and Jean Tirole. 2011. Laws and Norms. *National Bureau of Economic Research No. w17579.*

5. Bar-Gill, Oren and Assaf Hamdani. 2003. Optimal Liability for Libel. *Contributions in Economic Analysis & Policy*. 2(1)

6. Buccirossi, Paolo, Giovanni Immordino, and Giancarlo Spagnolo. 2017. Whistleblower Rewards, False Reports, and Corporate Fraud. *CEPR Discussion Paper No. DP12260*

7. Crawford, Vincent and Sobel, Joel 1982. Strategic Information Transmission. *Econometrica* 50, 143151.

8. Dalvi, Manoj and James F. Refalo. 2007. An Economic Analysis of Libel Law. *Eastern Economic Journal.* 74-94

9. Depoorter, Ben and Jef De Mot. 2005. Whistle Blowing. *Supreme Court Economic Review*, 2-28.

10. Deffains, Bruno and Claude Fluet. 2019. Social Norms and Legal Design. *Journal of Law, Econ., and Organizations* 1-31

11. Friedman, Ezra, and Abraham L. Wickelgren. 2005 Bayesian Juries and the Limits to Deterrence. *Journal of Law, Economics, and Organization* 22.1: 70-86

12. Hemel, Daniel and Ariel Porat. 2019. Free Speech and Cheap Talk. *Journal of Legal Analysis* 46-103

13. Garoupa, Nuno. 1999. Dishonesty and Libel Law The Economics of the "Chilling" Effect, *JITE* 284-300

14. Garoupa, Nuno. 1999, The Economics of Political Dishonesty and Defamation, *International Review of Law and Economics* 167-180

15. Givati, Yehonatan. 2016, A Theory of Whistleblower Rewards, *Journal of Legal Studies* 43-72

16. Heymann, Laura A. 2012. The Law of Reputation, and the Interest of the Audience, *B.C. L. Rev.* 1341-1999

17. McNamara, Lawrence. 2007. *Reputation and Defamation*

18. Mungan, Murat C. 2016. A Generalized Model for Reputational Sanctions and the (ir)Relevance of the Interactions between Legal and Reputational sanctions. *International Review of Law and Economics* 46. 86-92.

19. Polinsky, A. Mitchell, and Steven Shavell. 2007. The Theory of Public Enforcement of Law. *Handbook of Law and Economics* 1: 403-454.

20. Rasmusen, Eric. 1996. Stigma and self-fulfilling expectations of criminality. *The Journal of Law and Economics* 39(2) (1996), 519-543.

21. Spence, Michael. 1973. Job Market Signaling. *Quarterly Journal of Economics* 87(3), 355-374.

22. Steenson, Mike. Presumed Damages in Defamation Law. *William Mitchell Law Review* 40(4) (2014), 1492-1542

23. Reneé. Stulz, Securities Laws, Disclosure, and National Capital Markets in the age of Financial Globalization, *Journal of Accounting Research*, 2009, v47(2), 349-390.