Legal Studies
Research Paper Series

Research Paper No. 23–66

Defamation with Bayesian Audiences

Yonathan A. Arbel
Murat C. Mungan

Defamation with Bayesian Audiences

Yonathan A. Arbel,* Murat C. Mungan†

How strictly should the law regulate false defamatory statements? We first show that the presence of judicial errors often puts defamation law on a Laffer curve: regulation that is too lax or too strict is inferior to moderate regulation. While moderate regulation is ideal, it is not always attainable, due to practical and legal constraints. With these constraints, the presence of Bayesian audiences can cause the optimal regulation to be laxer than is prescribed by standard models with naïve audiences. This highlights the importance of accounting for the impact of defamation laws on belief formation.
Keywords: Defamation, Bayesian audience, information regulation, disclosure.
JEL classification: C72; D82; D83; K10; K13; K39.

1. Introduction

When statements are made in public, audiences assess their credibility based on a variety of cues. One of these cues is how strictly the law sanctions false statements, i.e., whether talk is cheap. Such audience effects complicate the standard analysis of defamation law, which traditionally only focuses on how the law affects speakers and the subjects of their speech. We investigate here the optimal strictness of defamation law when accounting for audience effects.

Defamation law imposes tort liability on speakers who publish false statements that harm their target's good name. A typical example comes from a recent case where a hospital staffer falsely alleged that a doctor was working under the influence of alcohol. The doctor won a lawsuit against the staffer and recovered millions of dollars in damages (Denman v. St. Vincent, 2020). While the basic structure of defamation law is well established, there is an ongoing social debate on defamation law's proper scope. With calls from the Supreme Court, legal scholars, politicians, and various pundits, there is growing pressure today to reform defamation law (Arbel & Mungan, 2019). In the midst of these calls, a new Restatement project was recently announced.

The literature on defamation law is vast, but the economic analysis of defamation law is quite limited (Hemel, 2020). In deciding the level of strictness of defamation law, standard legal analyses are dominated by a two-sided balancing act. On the one hand, society considers the interests of the target of the speech—her need for compensation and the need to protect her by deterring defamatory speech against her. On the other hand, society also considers the speaker, his right to free speech, and the social concern with chilling valuable speech (NYTimes v. Sullivan, 1964).

*University of Alabama School of Law. E-mail: yarbel@law.ua.edu
†George Mason University, Antonin Scalia Law School. E-mail: mmungan@gmu.edu.

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

In reality, we noted, defamation law also affects parties beyond the speaker and his target. In particular, defamation law also affects audiences of speech. This is not just due to the familiar idea that strict defamation law would limit the supply of false speech through the deterrence of speakers. If that were the case, the protection of audiences would be a simple matter of setting sanctions as strict as legally possible. Defamation law also affects audiences by changing their assessment of the credibility of speech and, thus, it affects the audience's propensity to act upon statements (Pennycook et al. 2020, Arbel 2022). Such effects add complexity that the standard analysis neglects. The neglect of audience effects may be because lawyers naturally focus on the parties that take an active part in the legal process–the victim as the plaintiff and the speaker as the defendant (Heymann, 2012). Whatever the reason, two recent informal accounts (Hemel & Porat, 2019 & Arbel & Mungan, 2019) suggest that omission of audience effects is consequential to the optimal strictness of defamation law.

Our object here is to bridge this gap by offering a general framework that analyzes behavior and evaluates welfare based on the strictness of defamation law. To do so, we construct a model that includes three key features: (i) a Bayesian (rather than a naïve) audience, (ii) errors in the court's judgment (wrongful liability as well as wrongful failure to find liability), and (iii) a cap on recoverable damages. We explain the role each of these features plays, after briefly reviewing the structure of our model and its implications.

We consider the behavior of three parties. A speaker, who has private information about a certain target – a business or an individual. The speaker may make claims about the target to an audience member. The audience member then decides whether to interact—trade, collaborate, socialize—with the target. Targets can be a high- or low-type, and audiences would rather interact only with the former kind. If the target loses an interaction, he may bring a lawsuit against the speaker alleging defamation. We study behavior under different levels of damages for defamation and their social welfare implications.

Our analysis reveals three central findings. First, we find that there is an optimal level of damages that supports separating equilibria in which would-be defamers are deterred from sharing false information and speakers only share information honestly with their audience. Audience members believe speakers and act upon this information. Naturally, social welfare is highest under this regime.

Second, we find that defamation laws often follow a 'Laffer Curve.' Lax regulation results in a flood of cheap talk, which leads to audiences discounting all statements—true or false—and simply acting on their priors. This results in equilibria where important information is left uncommunicated. Sanctioning defamation too strictly is also unwanted, because high sanctions invite frivolous litigation, which in turn chills true negative statements. In between these two extremes, the optimal level of damages follows an inverse U shape, with a range of optimal damages. Thus, our unified framework shows that both cheap talk and 'overpriced' talk are undesirable as they deprive audiences of relevant information that could be made available to them under moderate damages.

Our third result emerges when the cap on damages is lower than the level necessary to support equilibria in which the target's type is fully revealed. In reality, it is difficult to calculate the exact level of damages, and even when possible, damages are bounded by both constitutional considerations and limits on defendants' wealth. In such cases, we find that a lax approach can be superior to a more stringent one. The reason is that stringent regulation invites audience trust, but because some statements are false, this trust can be misplaced, leading to the deterrence of valuable interactions. Compounding the issue is that stringent regulation increases litigation. Lax regulation, however, invites audiences to rely more on their priors and reduces litigation. Thus, perhaps counter-intuitively, lax regulation becomes preferable to strict regulation when reputational harms are large, and the opposite conclusion may hold when reputational harms are small.

All three features of our model (Bayesian audience, judgment errors, and damage caps) play important roles in the production of these results.

First, when courts make no errors in judgment, people have no incentive to bring frivolous claims against speakers. In this case very large damages (if feasible) are always preferable to smaller damages, because they only deter false speech without having any impact on true negative speech. This is a dynamic that emerges in many other contexts as well, and highlights the role of judgment errors in explaining the inefficiency of very large damages in the defamation context, and the emergence of the Laffer Curve to which we alluded.

Second, the prior economics literature on defamation law assumes that a publisher (e.g. a tabloid, journal, an individual etc.) can always harm another person by making negative statements about them, and the extent of this harm is independent of the laws in place (e.g., Garoupa 1999a,b, Bar-Gill & Hamdani 2003).[1] This is equivalent to the audience –whose beliefs and behavior is not considered in the prior literature– naively forming its beliefs and acting upon them. Thus, in the prior literature, the main function of reforming the law is to alter the expected costs and benefits of making disparaging statements, but not the harmful impact of defamatory statements. With a naïve audience, increasing damages leads to a reduction in the expected harm to the target, because it deters negative speech. This is not so when the audience is Bayesian. Because very low damages result in frequent false allegations, they dilute the informational content of speech, and therefore causes the audience to act according to its priors. Thus, in addition to providing straightforward rationales for some behavioral responses in the defamation context (e.g., disregarding certain false speech), the incorporation of Bayesian audiences also has important normative implications, e.g., lowering damages can reduce the harm that results from defamatory statements.

1. The mirror image of this assumption is also invoked in this literature: the speaker's benefit from making a negative statement is independent of the audience's beliefs, because the audience is not considered in this literature. This assumption is made, for instance, in Dalvi and Refalo (2008), which focuses exclusively on the speakers' incentives and ignores not only the audience's beliefs and behavior but also the target's.

Third, this normative distinction becomes quite significant with binding damage caps,[2] in which case false disparaging remarks cannot be fully eliminated. Thus, with Bayesian audiences, the choice is between a high degree of interactions between the audience and targets (good and bad) achieved through low damages, and the maximum level of damages that causes bad interactions to be deterred along with some good interactions. The former option is preferable when the value of good interactions are large. On the other hand, with a naïve audience, maximum damages are always preferable, because even with low damages the audience believes false disparaging remarks, which are in high supply due to the lack of deterrent damages.

In short, the main impact of judicial errors in our analysis is to rule out the optimality of very large damages. This becomes an important issue when the damage cap is very large (or non-existent), in which case the presence of judicial errors supplies an independent rationale for not having very large damages. On the other hand, when the damage cap is binding, a naïve audience implies that the maximum damage is optimal, and this result is overturned with Bayesian audiences.

While our analysis focuses on defamation law, the basic question we pose here is relevant for a broad range of legal contexts. The law regulates false speech in domains as diverse as corporate disclosures, false advertising, whistleblowers, and law enforcement. Common to these domains is a basic tension between the strictness of sanctions for misreporting and the informativeness of speech, and we comment on potential implications.

The next section offers some brief background and reviews the related literature. Section 3 presents the model and its analysis. Section 4 evaluates the welfare implications of different damages regimes, and highlights the importance of accounting for audience effects. Section 5 contains several extensions and discussions of the basic 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. Literature Review

Defamation law regulates the dissemination of false statements that are 'defamatory.' To be defamatory, a statement must not only be false but also made public and be capable of harming one's reputation and standing in the community. Defamation law is considered to be a branch of torts, and it encompasses several distinct torts, most notably libel and slander. Today, however, the distinction has less practical significance than in the past, and in what follows, we abstract from it.

Many defamation lawsuits are brought by individuals, but businesses and firms can also bring suit. A recent high-profile example involves a lawsuit by 'Dominion,' a firm that sells voting hardware and software, against various

2. A similar dynamic also emerges when courts frequently make judgment errors, as we briefly explain in section 5.4., below, and in greater detail in Arbel and & Mungan 2020.

public figures and media outlets, who alleged it was involved in the manipulation of election votes.[3]

Defamation law evolved in the common and ecclesiastical courts of England. In the United States, the states took the doctrine and used it to develop their own variants. A major development took place in 1964, when the Supreme Court decided the seminal case of *NYTimes v. Sullivan*. There, the Court reviewed the existing body of doctrine in light of the First Amendment protection of free speech and press. The Court made it considerably harder for public figures to bring lawsuits on matters of public interest. In the years that followed, the doctrine was refined and, while still carrying signs of its convoluted history, reached a certain degree of balance. In recent years, however, there has been growing pressure to reform the law. Comments from the Supreme Court (*McKee v. Cosby, 2019; Berrisha v. Lawson*, 2021), the political sphere, legal commentators, and pundits—all reveal dissatisfaction with the law. Many of these comments suggest that defamation law should be made stricter; e.g., Cass Sunstein called the *NYTimes v. Sullivan* decision 'anachronistic' and argued that public figures should be allowed to bring suit more easily (Sunstein, 2021). Interestingly, the reason why the law should protect good name interests is not well understood. Some ground the law's intervention in a property like interest in good name, or good name's basis in dignity, property, and honor (Post, 1984) while others relate it to concepts of social status and reputation (Arbel, 2021).

The legal literature on defamation law is rich and vast, and it explores a variety of topics, involving deep questions of political philosophy and constitutional commitments. It is therefore quite surprising that the literature on the economics of defamation law 'has lagged' and is sparse (Hemel, 2020). Some notable contributions in this space includes Richard Posner's pioneering analysis (Posner, 1973, 1986), which highlighted the applicability of cost-benefit analysis to defamation law. More recent work focuses on the law's effect on media's incentives to investigate and report topics of public interest (Bar-Gill & Hamdani, 2003, Dalvi & Refalo, 2008, Acheson & Wohlschlegel, 2018) and on political dishonesty (Garoupa, 1999a,b). As noted, this paper differs from these analyses by considering a Bayesian audience, alongside damage caps and judicial errors.

Despite these contributions, courts and legal commentators are limited to a fairly rudimentary understanding of the incentives fostered by different defamation law regimes. Here we amplify on two informal contributions that recognize the relevance and importance of audiences to the analysis of defamation law (Hemel & Porat, 2019, Arbel & Mungan, 2019). 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 emerging literature that looks at how laws can be used to create informal sanctions through the behavior of third parties (e.g., Deffains & Fluet,

3. US Dominion, Inc. v. Fox News Network, LLC, C. A. N21C-03-257 EMD (Del. Super. Ct. Dec. 16, 2021).

2019, Mungan 2016, Bénabou & Tirole, 2006, 2011, Rasmusen 1996).

3. Model

We model the interactions between three 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 A decides whether to interact with the target, S, who knows T's type, communicates with A and may either disparage T or make a non-disparaging comment. Because we study defamation, we consider the possibility that S may falsely disparage T in order to deter an interaction between A and T. We defer the discussion of speakers being (at least partly) motivated by a desire to truthfully share information, as this has limited impact on our analysis.[4] We model the interactions as a Bayesian game, and use it to identify Perfect Bayesian Equilibria.[5]

3.1 Preliminaries

The target, T, obtains a benefit of r_t from the interaction, where $t \in \{B, G\}$ denotes his type and 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 types is $\gamma \in (0,1)$.[6] A prefers to interact with good types, but not bad types, because this results in a utility of $g > 0 > -b$ where b is the disutility A bears from interacting with a bad type. Thus, absent further information, A would prefer to interact with T if $\gamma g - (1-\gamma)b > 0$, and we assume this inequality holds, since otherwise no interactions would take place between A and T even without (negative) input from S.[7] Thus, the audience prefers to interact with the target if its updated belief (based on the statement it receives from S) of T's likelihood of being a good type exceeds the threshold

$$\widehat{x} \equiv \frac{b}{g+b} < \gamma, \tag{1}$$

where the inequality follows from the assumption that A would prefer to interact with T absent input from S.

The speaker has an interest in whether A and T interact: S obtains a gain of v when A avoids interaction (alternatively, v can be interpreted as a loss incurred when A chooses to interact with T). v is a random variable with the cumulative distribution function $F(v)$ with support $(0,1]$ where the upper-bound of the support is normalized to simplify notation. The specific v-draw is private

4. 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.

5. Figure 5 in the Appendix depicts the interactions between the three parties and is helpful in following the detailed descriptions of the interactions that we provide, next.

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

7. An analysis of this case can be found in an earlier version of this article, and yields no further insights (see Arbel & Mungan (2020)).

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. $r_g + g > 1 > 0 > r_b - b$.

After Nature determines the types of T and S, the target'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.).

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.[8] 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. To capture the parties' payoffs from litigation, 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 ($l < 1$) 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 $t \in \{B, G\}$
We assume the probability of wrongful liability is small: $q_B < q_G \left(\frac{l/2}{1-l/2}\right)$

Because courts wield broad latitude in setting the level of damages,[9] we follow the existing literature (see, e.g., Garoupa 1999a,b) and use d as a policy lever to operationalize different defamation regimes. The value of d can also be interpreted as expected damages, i.e. $d = E[\tilde{d}]$ where $\tilde{d}$ is a random variable representing actual damages awarded whose distribution is affected by the court's interpretation of the law. We also note that d can be chosen separately for each type of statement (represented by all variables described above, r_g, r_b, g, b, etc.) being analyzed, which we take as given in our analysis.

It is worth noting that we take the odds of winning at trial, q_B and q_G, as exogenously given. This is a standard commitment assumption in the enforcement literature, and implies that courts are committed to reviewing cases only on their merit, i.e., without bringing in their informed estimates about the proportion of frivolous cases. Our assumption relating q_G to q_B, above, corresponds to the case where judicial errors are possible, but occur infrequently. We note that the analysis of the alternative case with large judicial errors leads

8. In practice, courts often require some proof of harm to allow monetary recovery, hence the requirement here that the defamatory remark prevented an interaction.

9. Courts can, within limitations, award nominal, economic, non-economic, and punitive damages, and have demonstrated considerable discretion in practice. Some examples include Lothschuetz v. Carpenter, 898 F.2d 1200, 1205 (6th Cir. 1990) ($1 in nominal damages); Waste Mgmt. of Texas, Inc. v. Texas Disposal Sys. Landfill, Inc., 434 S.W.3d 142, 162 (Tex. 2014) ($450,000 in economic harms); Cantu v. Flanigan, 705 F. Supp. 2d 220 (E.D.N.Y. 2010) ($150 million in non-economic damages); Armstrong v. Shirvell, 596 F. App'x 433, 448 (6th Cir. 2015) ($500,000 in punitive damages).

to similar results, and a complete analysis of this case can be found in Arbel & Mungan (2020).

3.2 Players' Actions, Beliefs, Strategies, and Payoffs

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	Litigate

We note that labeling A's action of interacting as 0 may appear counter-intuitive. However, the benefit of this notation is that a suit is filed only in cases where all players' actions are 1. This makes it simpler to express the pay-off of the speaker (as in Table 3, below), since she faces expected litigation costs only when all actions equal 1.

Using this notation, we can describe the strategies of each player as follows:

Table 2: Players' Strategies

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

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

Because our solution concept is a Perfect Bayesian Equilibrium (henceforth PBE), we also specify A's beliefs regarding T's type, as:[10]

x_i : Belief that T is a good type given $z=i$

With this notation, we express the expected pay-offs of each player, given their beliefs and information, as follows:

Table 3: Players' Payoffs

Player	Payoff
S	$a(s(t,v))(v-p(t)s(t,v)(q_t d+\frac{l}{2}))$
A	$a(z)(x_z g-(1-x_z)b)$
T	$(1-a(s(t,v)))p(t)(q_t d-\frac{l}{2})+a(s(t,v))r_t$

10. 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.

3.3 Effective and Ineffective Communication Equilibria

Perfect Bayesian Equilibria consist of *assessments* (i.e. a profile of beliefs and strategies) that satisfy sequential rationality and consistency of beliefs. Since the requirements for PBE are well known, we relegate their formal definitions to Appendix A, below. As in many other contexts, communications can be disregarded by the audience in some equilibria. We distinguish between these and other types of equilibria by using the following definition.

Definition 1 *A PBE is an effective communication equilibrium if, and only if, the audience chooses not to interact with the target with some positive probability based on the information it receives from the speaker.*

We start by noting that defamation law cannot eliminate ineffective communication equilibria. This is because when the audience's beliefs regarding the target's types are unconditional on the speaker's statement and equal to its prior (i.e. $x_0 = x_1 = \gamma$), it chooses to interact with the target regardless of what it hears from the target (i.e. $a^*(z) = 0$). This results in payoffs of 0 and r to the speaker and target, respectively. These payoffs are independent of the actions of the speaker and target, which makes them indifferent between playing any of the strategies available to them. Thus, any assessment where the speaker plays a strategy that supports the audience's beliefs constitutes a PBE. The simplest example is one where the speaker never chooses to disparage the target (i.e. $s(t, v) = 0$ for all t and v). We formalize this observation as follows.

Proposition 1. *Under all defamation regimes, there exist ineffective communication equilibria.*

Proof. The assessment consisting of $x_1^* = x_0^* = \gamma$, $a^*(z) = 0$, $s^*(t, v) = 0$, and $p^*(t) = \begin{cases} 0 & \text{if} \quad q_t d \leqslant l/2 \quad \text{for } t \in \{B, G\} \\ 1 & \text{if} \quad q_t d \geqslant l/2 \quad \text{for } t \in \{B, G\} \end{cases}$ satisfies sequential rationality and consistency of beliefs (i.e. requirements 1-4 in Appendix A), and thus is a PBE. □

Proposition 1 notes that ineffective communication equilibria are always present, regardless of the defamation regime in place. If these were the only equilbria, defamation law would be irrelevant. Thus, we proceed by showing that some levels of damages in fact generate effective communication equilibria.

Proposition 2. *(i) Extremely low damages (i.e. $d < \frac{l}{2q_G}$) and extremely high damages (i.e. $d > \frac{2-l}{2q_B}$) only generate ineffective communication equilibria. (ii) There exist a range of moderate damages, $D \subset \left(\frac{l}{2q_G}, \frac{2-l}{2q_B}\right)$, which generate effective communication equilibria. (iii) The audience acts consistently with the speaker's statement, i.e. $a^*(z) = z$, in all effective communication equilibria.*

Proof. See Appendix. □

The intuition behind the first part of proposition 2 is relatively straightforward. When damages are extremely low, the target is deterred from suing the speaker, even when he has a meritorious case, since expected damages (i.e. $q_G d$) are lower than litigation costs. This causes the speaker's statements to be perceived as cheap-talk by the audience, since the speaker faces no negative consequence from making disparaging statements. Thus, the audience disregards the speaker's statements and acts according to its priors. On the other hand, when damages are extremely high, all targets are incentivized to litigate, and expected damages are high enough to deter all speaker types from making disparaging statements. Thus, the audience is once again left without any informative statements, this time due to the over-pricing of speech as opposed to the presence of cheap-talk.

It is only moderate damages that support effective communications between speakers and audience members, and this is formalized in proposition 2-(ii). Part (iii) of proposition 2 simply rules out the possibility of counter-intuitive equilibria, for instance, in which the audience infers from a disparaging remark that the target must be a good type and vice-versa (i.e. where $a^*(z) = 1 - z$). These preliminary findings indicate that if defamation laws are to have any impact, they must do so through effective communication equilibria obtained under moderate damages. Thus, we analyze these equilibria in further detail, next.

3.4 Moderate Damages and Effective Communication Equilibria

The damages in place affect the target's incentive to sue when disparaged, as well as the speaker's incentives to disparage the target in the first place. We note two pairs of critical damages that pertain to each party's incentives. First,

$$d_1 \equiv \frac{l}{2q_G} \text{ and } d_3 \equiv \frac{l}{2q_B} \tag{2}$$

are the smallest damages that causes a type G and B target, respectively, to bring suit whenever the speaker disparages him.[11] On the other hand, when damages are greater than

$$d_2 \equiv \frac{2-l}{2q_G} \text{ and } d_4 \equiv \frac{2-l}{2q_B} \tag{3}$$

type G and B targets, respectively, are expected to bring suit, and this deters the speaker from defaming the target. Our assumption of non-prohibitive litigation costs and small judicial errors implies that these four critical damage levels are ordered as follows:

$$d_1 < d_2 < d_3 < d_4 \tag{4}$$

Thus, there are three categories of moderate damages, which we call *low, intermediate,* and *high* damages. We explain the incentives that each player faces in

11. We assume, only to simplify exposition, that an indifferent target chooses not to litigate.

each of these damage categories under an effective communication equilibrium (when one exists).

Low Damages ($d \in (d_1, d_2)$)

In this range, the target has the incentive to litigate only if he is type G, since $q_G d > \frac{l}{2} > q_B d$. Thus, in an effective communication equilibrium, the speaker faces no threat of litigation from disparaging a bad type, and thus a type B target is disparaged with certainty. On the other hand, if the speaker encounters a type G target, she expects that disparaging him will lead to a cost of

$$v_G(d) \equiv q_G d + \frac{l}{2}. \tag{5}$$

Thus, the speaker chooses to disparage a type G target if her type exceeds this value. Therefore, a type G target is disparaged with a probability of

$$1 - F(v_G(d)) \tag{6}$$

Our analysis thus far identifies the behavior of the speaker and target in an effective communication equilibrium, assuming that it exists. But, for this type of equilibrium to be supportable, the audience's beliefs must be consistent with the equilibrium behavior of the other parties. Thus, the audience must hold the belief that a target who is not disparaged must be a good type, since all bad types are disparaged, i.e.

$$x_0^* = P(t = G|z = 0) = 1$$

On the other hand, when the audience receives a disparaging statement, it must believe that the target is nevertheless a good type with a probability of

$$x_1^*(d) = P(t = G|z = 1) = \frac{\gamma[1 - F(v_G(d))]}{\gamma[1 - F(v_G(d))] + (1 - \gamma)} < \gamma \tag{7}$$

This is because a type G target is disparaged with probability $1 - F(v_G(d))$ whereas a type B individual is disparaged with certainty, and the likelihood with which the target is a good type is γ.

As we noted via (1) the audience finds it in its best interest to interact with the target if it believes he is a good type with a probability exceeding $\hat{x}$. Thus, an effective communication equilibrium is supportable in this range if

$$x_1^*(d) < \hat{x} < x_0^* = 1 \tag{8}$$

In the upper portion of this range, this condition certainly holds, since $1 - F(v_G(d_2)) = 0$, i.e. the speaker is always deterred from disparaging a type G target. Moreover, $x_1^*(d)$ is clearly decreasing in this damage range, which can easily be verified by inspecting (7). Therefore, within low damages, there is some critical damage $\hat{d}$, above which effective communication equilibria are obtained. Whether $\hat{d} = d_1$, i.e. whether effective communication equilibria are supportable with all lower moderate damages, depends on parametric values and has no important implications.

Based on these observations, we can summarize the impacts of increasing damages in the range $(\hat{d}, d_2)$ on the behavior of all players in effective communication equilibria. As damages are increased, the speaker disparages type G individuals less frequently, since the threshold speech benefit that she requires is increasing in damages per (5). This leads to less frequent litigation as well as less frequent blocking of beneficial interactions between the audience and the type G target. The level of damages leads to no further effects, because a type B target is disparaged with certainty in this range.

Intermediate Damages ($d \in [d_2, d_3]$)

When damages are increased into the intermediate range the speaker is always deterred from disparaging a type G target. This is because the expected damages and litigation costs associated with doing so exceed the benefit that she obtains from blocking the target's interaction with the audience. Moreover, because $d < d_3$, a type B target lacks the incentives to litigate. Thus, the speaker always disparages a type B target. Based on this behavior, it follows that in an effective communication equilibrium the audience's beliefs are given by

$$x_0^* = P(t = G|z = 0) = 1 \text{ and}$$
$$x_1^* = P(t = G|z = 1) = 0$$

This trivially implies that effective communication equilibria are supportable by all damages in this range, since $0 = x_1^* < \hat{x} < x_0^* = 1$.

Quite importantly, in this range all effective communication equilibria are separating equilibria wherein the speaker chooses to [not] disparage whenever she knows that the target is type B [G]. We later note the superiority of these equilibria when we conduct a welfare analysis.

High Damages ($d \in (d_3, d_4)$)

With high damages, a type B target is given the incentive to frivolously litigate. This, in turn, causes the speaker to be deterred from disparaging a type B target when her benefit from blocking interactions is low. Specifically, she chooses to disparage a type B individual only if her benefit exceeds

$$v_B(d) \equiv q_B d + \frac{l}{2} \tag{9}$$

Thus, a bad type is disparaged with a probability of

$$1 - F(v_B(d)) < 1 \tag{10}$$

On the other hand, when the speaker encounters a type G target, she never disparages him given the high damages. Based on this behavioral profile, in equilibrium the audience's beliefs are given by

$$x_0^*(d) = P(t = G|z = 0) = \frac{\gamma}{\gamma + (1 - \gamma)F(v_B(d))} > \gamma \text{ and}$$
$$x_1^* = P(t = G|z = 1) = 0$$

Figure 1

Figure 1a: A's beliefs that the target is type G in effective communication eq. under different damages.

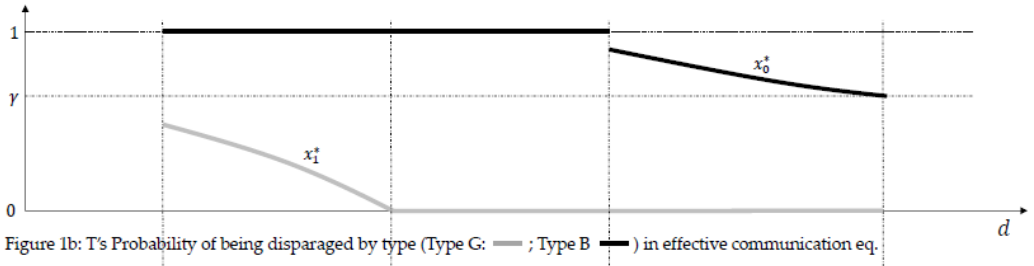


Figure 1b: T's Probability of being disparaged by type (Type G: ; Type B) in effective communication eq.

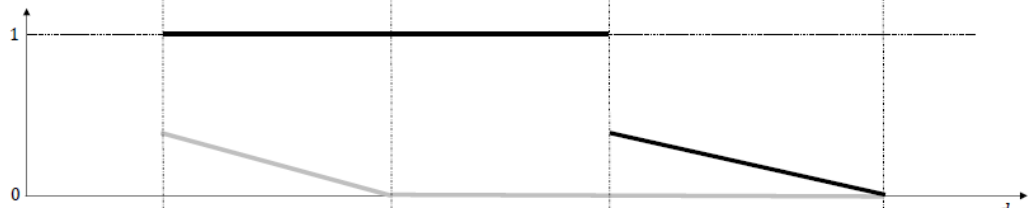
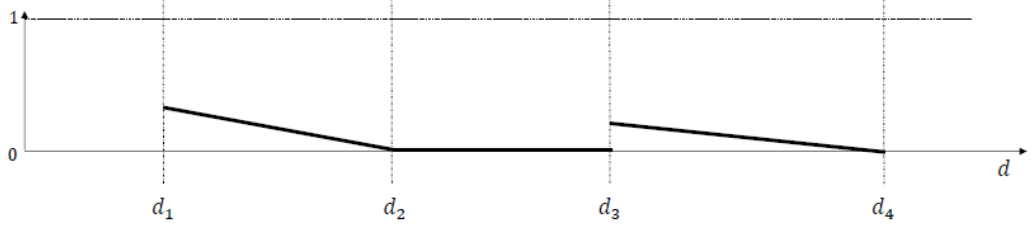


Figure 1c: Probability of litigation under effective communication eq.



Notes: The panels show the effects of damages in effective communication equilibria in a case where $\hat{d} = d_1$. *Panel a* describes the audience's equilibrium belief that a person is type G (with 1 being certainty) if the person is not disparaged (x_0^* plotted in black) and when the person is disparaged (x_1^* plotted in grey). *Panel b* illustrates the likelihood that a person will be disparaged based on whether they are type G (plotted in grey) or type B (plotted in black). *Panel c* describes the likelihood of a lawsuit for allegedly false disparagement. In the range of d_2 and d_3 effective communication equilibria are separating, hence beliefs are accurate, only truly bad types are disparaged, and there is no litigation.

Since $\hat{x} < \gamma$, it follows that $x_0^* > \hat{x} > x_1^*$, and therefore effective communication equilibria are supportable by all high damages.

As the above discussion indicates, increasing damages in this range only reduces the frequency with which speakers disparage a type B target. Thus, increasing damages in this range has countervailing effects: it increases the frequency of interactions with bad types but reduces the frequency of litigation.

We summarize our findings in this subsection through figure 1, below, which depicts the qualitative relationship between damages and the beliefs of A; the likelihood with which a type $t \in \{B, G\}$ target is disparaged; and the likelihood of litigation in effective communication equilibria. Next, we conduct a welfare analysis which builds on these key findings.

4. Welfare Analysis

In analyzing the social desirability of different defamation regimes, we use a simple social welfare function which consists of the sum of each player's expected pay-off. We conduct this welfare analysis to highlight the three main points that we emphasized in the introduction. First, welfare is non-monotonic in damages for defamation. More precisely, we show that the shape of welfare obtained through effective communication equilibria vis- à-vis damages resembles a Laffer Curve: welfare is increasing in the lower range of moderate damages (i.e. for $d \in (\hat{d}, d_2)$), is maximized in the intermediate range (i.e. when $d \in [d_2, d_3]$), and is decreasing in the upper range of moderate damages (i.e. when $d \in (d_3, d_4)$). Second, when there is a cap on damages (e.g. reflecting the wealth of the defendant or a legal bound on permissible damages), then all effective defamation remedies may reduce welfare. Third, we contrast the implications of a model with a Bayesian versus a naïve audience. When the audience is naïve and easily misled by false statements, type G targets always prefer stricter defamation laws. The same is not true when the audience is Bayesian, because absent sizeable damages the audience perceives the speaker's disparaging statements as cheap-talk and disregards them. This insight leads to a divergence between the normative implications of the two models: with binding caps on damages, it may be optimal to have no defamation laws at all with a Bayesian audience but optimal to have maximal damages with a naïve audience. Next, we consider and formalize each of these points.

4.1 The Laffer Curve of Defamation Law

Under ineffective communication equilibria, the audience acts according to its priors. Thus, it chooses to interact with the target regardless of the statements by the speaker. There is no litigation since interaction always takes place. Thus, expected welfare is independent of damages, and is given by

$$W_I \equiv (1-\gamma)[r_B - b] + \gamma[r_G + g] \tag{1}$$

On the other hand, under effective communication equilibria, the specific functional form of welfare differs depending on which of the three ranges damages are in, as explained the previous section. Next, we consider welfare under each range.

With low damages that support effective communication equilibria (i.e. $d \in (\hat{d}, d_2)$), when the target is type B, the speaker disparages him, the audience refuses to interact, and the target chooses not to litigate. Thus, with a probability of $(1-\gamma)$ expected welfare equals the speaker's expected benefit $E[v]$. When the target is type G, the speaker disparages him only when $v > v_G(d)$. In those cases, the audience avoids an interaction with T, and the target litigates. Thus, with a probability of γ, expected welfare is $F(v_G(d))[r_G + g] + \int_{v_G(d)}^{1}(v-l)f(v)dv$. Therefore, expected welfare is given by:

$$W_L(d) \equiv (1-\gamma)E[v] + \gamma\left(F(v_G(d))[r_G + g] + \int_{v_G(d)}^{1}(v-l)f(v)dv\right) \tag{2}$$

When damages are intermediate (i.e. $d \in [d_2, d_3]$), effective communication leads to separating equilibria wherein interactions take place if, and only if, the target is a good type. Moreover, there is no litigation since type B targets lack the incentives to litigate. Thus, welfare in this range is given by

$$W_S \equiv (1-\gamma)E[v] + \gamma[r_G + g] \tag{3}$$

Finally, when damages are in the upper moderate range (i.e. $d \in (d_3, d_4)$), the speaker chooses not to disparage a type G target. Thus, with a probability of γ, welfare is $r_G + g$. When the target is type B, the speaker chooses to disparage him only when $v > v_B(d)$, and this leads to litigation. Thus, with a probability of $1-\gamma$, expected welfare is $F(v_B)[r_B - b] + \int_{v_B(d)}^{1}(v-l)f(v)dv$. Thus, expected welfare is

$$W_H(d) \equiv (1-\gamma)\left(F(v_B(d))[r_B - b] + \int_{v_B(d)}^{1}(v-l)f(v)dv\right) + \gamma[r_G + g] \tag{4}$$

A very simple yet important observation is that W_L is increasing whereas W_H is decreasing in damages. This is because, when damages are in the lower moderate range, the impact of increasing damages is to reduce the likelihood of defamatory statements against a type G target. This is beneficial, because it reduces the likelihood of blocked beneficial interactions between A and T as well as wasteful litigation between T and S. Similarly, when damages are in the upper moderate range, lowering damages leads to an increase in the likelihood with which a type B target is disparaged. This increases the likelihood of harmful interactions being blocked, but at the expense of increased litigation. The former (beneficial) effect dominates the latter (detrimental) effect, since the speaker disparages a type B target only if her benefits from doing so more than off-set total litigation costs. This last point can be formalized by noting that

$$W_H' = (1-\gamma)f(v_B)v_B'\{(r_B - b) - (v_B - l)\} \leqslant 0 \tag{5}$$

since $r_B < b$ and $v_B(d) \geqslant v_B(d_3) = l$.

To summarize these observations, we depict welfare as a function of damages in figure 2, below, which is accompanied by clarifying notes.[12]

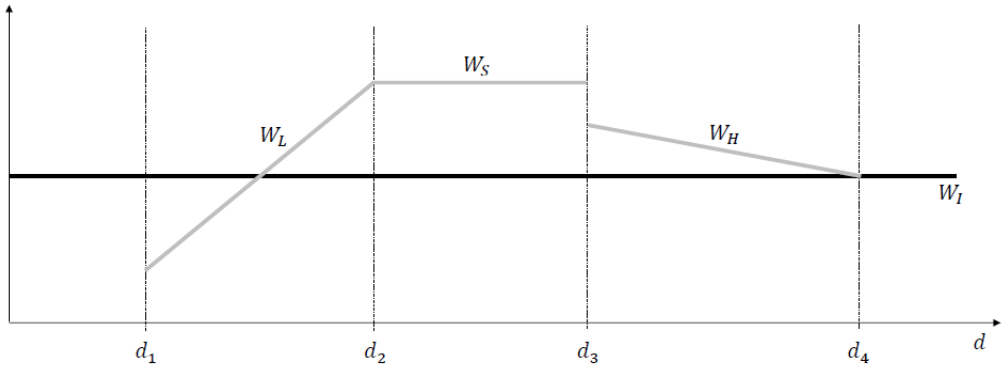
The most striking feature of figure 2 is that welfare obtained under effective communication equilibria mimics an inverse U shape, notwithstanding the facts that it contains a discrete drop[13] around $d = d_3$ and that welfare is constant in the $[d_2, d_3]$ range. This is what we interpret as the Laffer curve of defamation law. Welfare is not monotonically increasing in damages, because

12. We note that effective communication equilibria may or may not be supportable for all low damages, as explained in the previous section. Figure 2 depicts, only for expositional purposes, a case where all low damages support effective communication equilibria.

13. Once damages pass over to the upper moderate range, damages are sufficient to invite firivilous litigation by type B targets. This leads to a discrete reduction in welfare, since it causes the likelihood of disparaging statements against type B individuals to discretely drop from 1 to $1 - F(v_B(d))$ and it also causes a discrete jump in the likelihood of litigation (see figure 1).

Figure 2

Figure 2: Welfare as a function of damages under effective communication eq. ( ) and ineffective communication eq. ( )



Notes: The graph illustrates changes in social welfare as a function of damages under effective communication equilibria (depicted in grey) and ineffective communication equilibria (depicted in black). The highest welfare is attained under effective communication equilibria between d_2 and d_3, as it features fully separating equilibria. It rises in the lower range, but falls in the higher range due to increased frivolous litigation. Welfare is unaffected by damages under ineffective communication equilibria.

large damages (i.e. $d > d_3$) reduce welfare by deterring accurate negative speech against a type B target. On the other hand, reducing damages to low levels (i.e. $d < d_2$) is also detrimental because it leads to defamatory statements against a type G individual, which are taken seriously by the audience.

Another feature of the welfare curve depicted in figure 3 is that the separating equilibria obtained through intermediate damages lead to greater welfare than ineffective communication equilibria. We note that this is no coincidence, and occurs due to the fact that while the audience interacts with either type in ineffective equilibria, it interacts with a target if, and only if, it is a good type when damages are in the intermediate moderate range. Thus, when there are no caps on damages it follows that setting damages in the intermediate moderate range is socially desirable. We formalize this result through the following proposition, whose proof follows from our comments, above.

Proposition 3. Separating equilibria where S chooses to disparage T if, and only if, he is type B lead to greater expected welfare than any other equilibria and are obtainable only through intermediate damages.

An implication of proposition 3 is that maximizing welfare through defamation law requires the implementation of separating equilibria through the use of intermediate damages. We note that this implication is obtained under the assumption that these damages are feasible. However, when these damages are too large for the defendant to pay (i.e. when the defendant is judgement-proof) or when there are legal restrictions (e.g. constitutional) on the damages

that can be chosen, damages large enough to support equilibria in which target types are fully revealed may not be available. We discuss this case, next.

4.2 Bounded Damages

We use $\bar{d}$ to denote the upper bound on damages. An immediate implication of proposition 3 is that when $\bar{d} \geqslant d_2$, the upperbound is non-binding. Thus, we focus on the case where $\bar{d} < d_2$. With this restriction in place, there are only two relevant ranges of damages that one can select from; (i) very low damages which only support ineffective communication equilibria ($d \leqslant d_1$), and (ii) damages in the lower range ($d \in (d_1, \bar{d})$).

In the latter range, if $\bar{d}$ is close to d_1 it is possible for effective communication equilibria to be unsupportable,[14] and the analysis of this case is trivial: there is no feasible level of damages that can result in effective communication equilibria, and hence the choice of damages is irrelevant. Thus, we focus on the more interesting case where maximum damages are sufficient to support some effective communication equilibria, i.e. $x_1^*(\bar{d}) < \hat{x}$.

In this case, we can compare the welfare benefits and costs associated with ineffective and effective communication equilibria. The former equilibria, as reflected by (1), always lead to interactions between T and A. These interactions are welfare reducing when the target is type B. Hence, when the target is type B, effective communication equilibria perform better, since they deter interactions between T and A. However, this comes at the cost of deterring socially beneficial interactions between a type G target and A when the speaker has a sufficiently high type (i.e. $v > v_G$). This implies that a switch from an ineffective communication equilibrium to an effective communication equilibrium trades-off deterrence of good interactions against deterrence of bad interactions. Therefore, when the harm to the target from defamatory statements is large relative to other considerations, a rather counter-intuitive result is obtained. Even when it is possible to implement effective communication through defamation laws, it is socially more desirable not to do so. This happens because putting a price on speech lends more credibility to the speaker's statements, which she can then use to inefficiently block a good interaction. In such cases, the superior option is to not make speech credible and cause the audience to rely on its priors, which causes it to interact with the target.

We formalize this result via proposition 4 below, and we provide an example and its graphical depiction via figure 4 to illustrate it.

Proposition 4. Suppose there are binding maximum damages (i.e. $\bar{d} < d_2$). Then, given all other parameters, there exists a threshold harm from effective defamation, $\bar{r}_G$, such that ineffective communication equilibria lead to (weakly) higher welfare than all effective communication equilibria if, and only if, $r_G \geqslant \bar{r}_G$.

14. This occurs when maximum damages are not enough to deter speech against good types frequently enough in an effective communication equilibrium, as noted via (7)-(8) and the accompanying discussion.

Proof. Using (1) and (2), we can express the difference in between ineffective and effective communication equilibria as:

$$W_I - W_L = (1-\gamma)[r_B - b - E[v]] + \gamma \int_{v_G(d)}^{1} [r_G + g - v + l] f(v) dv$$

This expression is increasing and unbounded in r_G. Thus, there exists $\bar{r}_G$ such that $W_I \geqslant W_L$ iff $r_G \geqslant \bar{r}_G$. □

In figure 3, below, we depict multiple cases which illustrate the rationale behind proposition 4. In this example, v is distributed uniformly and $\bar{r}_G = 1.7$ is used to illustrate all three possibilities.[15]

As the figure illustrates, the gap between welfare obtained through low damages under the two types of equilibria is decreasing in damages but increasing the harm that the target suffers from effective defamation. Thus, for small defamation harms to the target, effective communication equilibria obtained through maximum damages are superior, and the opposite conclusion holds for large defamation harms. The exceptional case where the two types of equilibria lead to the same amount of welfare when maximum damages are used is also depicted as an intermediate case (i.e. the case where $r_G = \bar{r}_G$).

These observations imply that when damages inducing complete revelation of target types are not feasible, it is socially desirable to strive for effective communication equilibria only when the harms from defamation are small. This result appears counter-intuitive, because it suggests that the optimality of effective defamation remedies ought to be inversely related to the size of the alleged harm to the plaintiff. The rationale behind this result is that making speech credible in an environment where defamatory speech cannot be largely eliminated has the function of making some false speech credible, and thus harmful to type G targets. When the size of the harm to these individuals is large, it naturally becomes more desirable to take away the credibility of negative speech altogether.

4.3 Welfare with Bayesian versus Naïve Audiences

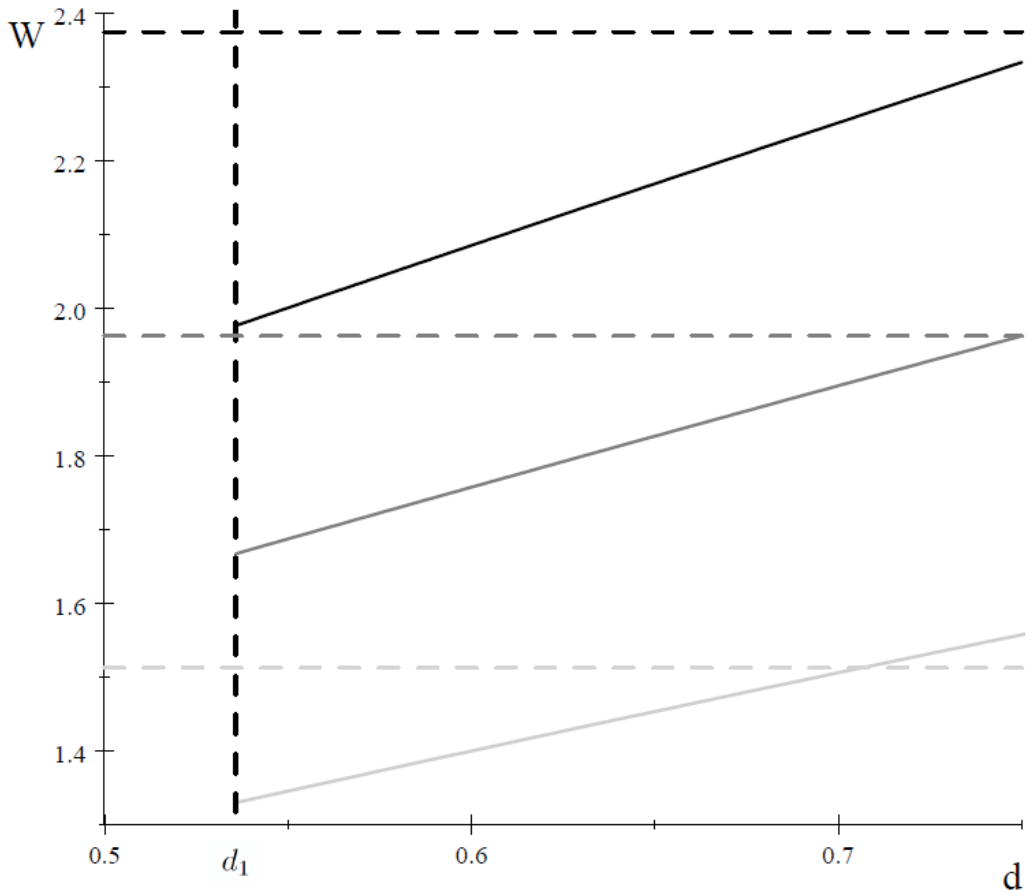
In our analysis thus far, we have considered a Bayesian audience whose beliefs are consistent with equilibrium behavior. An alternative assumption often invoked in the literature is that the harm from defamatory statements is independent of the frequency of false statements. We call the audience under this alternative assumption *naïve*, and we consider the differences in the implications of a model with a naïve versus Bayesian audience. When the audience is naïve, it avoids an interaction with the target if, and only if, it receives a disparaging statement from the speaker, and it does so regardless of the frequency of false statements.

Thus, with a naïve audience, the speaker is able to successfully block interactions with both target types when damages are very low (i.e. $d \leqslant d_1$). This

15. We use the following values to produce curves that do not overlap with each other for expositional purposes: $\gamma = r_B = l = \frac{3}{4}$; $g = b = 1$; and $q_G = \frac{7}{10}$.

Figure 3

Figure 3: W_I versus W_L for $d < \bar{d} = 0.75$ as r_G is varied.



Notes: The figure compares social welfare under effective and ineffective communication equilibria with bounded damages as the degree of harm suffered by the defamed person is varied: low ($r_G = 1.1$; depicted in light grey), medium ($r_G = 1.7$; depicted in grey), and high ($r_G = 2.25$; depicted in black). Welfare is depicted by dashed lines under ineffective communication equilibria and straight lines under effective communication equilibria in each case.

is because these damages are too low to generate any litigation threat from the target, and thus the speaker disparages the target independently of his type. The naïve audience, unlike the Bayesian audience, relies on the statement by the speaker instead of its prior, and therefore always avoids an interaction with the target.

When damages pass onto the moderate range, the equilibrium behavior and welfare in the naïve audience case is identical to those that are observed under an effective communication equilibrium of the Bayesian audience case. This is because the Bayesian audience, like the naïve audience, acts in a manner consistent with the speaker's statements in effective communication equilibria. Finally, when damages are very high (i.e. $d > d_4$), the speaker is deterred against making disparaging statements against both types, and the naïve audience interacts with both types. Thus, in this range welfare with a naïve audience is equal to welfare with a Bayesian audience.

In short, the most striking difference arising from a switch from a Bayesian audience to a naïve audience occurs when damages are too low to cause the target to litigate (i.e. $d \leqslant d_1$). The most prevalent normative impact of this difference is observed when there is an upper bound on maximum damages, since otherwise optimal equilibria are trivially obtained in the intermediate moderate range (i.e. $d \in [d_2, d_3]$) under both models. Thus, we focus on the case where $\bar{d} < d_2$ to highlight the greatest difference between the models with a Bayesian and a naïve audience.

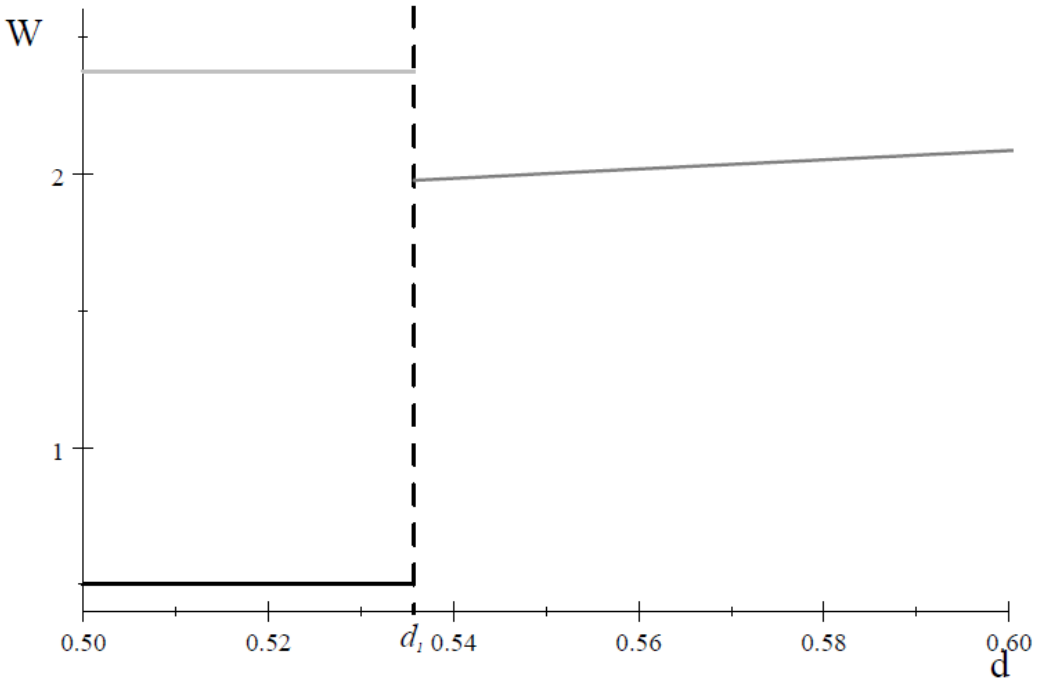
As we previously noted, with a Bayesian audience, type G targets are made worse off when damages in the lower moderate range lend credibility to speaker's statements. Thus, as we noted via proposition 4, when defamatory harms are large, welfare is actually reduced when defamation laws are effective compared to when they are not. The opposite conclusion holds with a naïve audience. In this case, increasing damages always leads to less frequent defamatory statements, and type G targets always prefer stricter defamation laws. Thus, when defamatory harms are large, contrary to the case with a Bayesian audience, it is optimal to use maximal damages.

We illustrate this result through figure 4, which depicts welfare obtained in the example used to generate the high r_G case in figure 3,[16] but this time it also includes welfare obtained with a naïve audience. The figure illustrates that while it is optimal to increase damages to their maximal level with a naïve audience, it is optimal to use damages low enough to guarantee the emergence of ineffective communication equilibria with a Bayesian audience. Thus, when harms from defamatory statements are large, assuming that the audience is naïve is not merely a simplifying assumption, it is one that can generate misleading normative implications.

16. As a reminder, this corresponds to the case where v is uniformly distributed; $r_G = \frac{9}{4}$; $\gamma = r_B = l = \frac{3}{4}$; $g = b = 1$; and $q_G = \frac{7}{10}$.

Figure 4

Figure 4: Welfare with a Bayesian versus Naïve Audience



Notes: The figure depicts social welfare as a function of damages when audiences are either naïve or Bayesian. Above d_1 the two measures converge (the dark grey line). Below d_1, the level of welfare depends on how the audience processes information. If naïve, welfare is low (the black line line). If Bayesian, it is high (the light grey line). Whether higher damages (above d_1) offer an improvement depends on whether the audience is Bayesian or naïve.

5. Extensions and Discussion

In Sections 3 and 4, we presented 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 there being a committed public enforcer, quality being endogenously chosen by the target, the existence of honest and other types of speakers, and courts being less accurate than we previously assumed. 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 is 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. Garoupa 1999a,b, for instance, takes a similar approach, and assumes that the target's behavior is impacted by what laws are in place. Here, we explain how the types in our setting can be endogenized, and how doing so yields results similar to those in prior work where the target's behavior is endogenous.

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 $1 > \overline{\gamma} > \underline{\gamma} > \hat{x} > 0$.

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 payoffs he obtains in the sub-game equilibria.

This observation reveals a very clear result: When the laws are extreme, i.e. $d \notin [d_1, d_4]$, the target has no reason to invest in quality. This follows from Proposition 2, which shows that with extreme laws, the audience acts based on its priors and interacts with the target. 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. d_1. In this case, in effective communication equilibria, it follows that $m_B = 0$ (because all bad types are disparaged) while $m_G = F(v_G(d))r$ (because good types are disparaged with probability $1-F(v_G(d))$, in which case there is a lawsuit which pays the target expected damages equal to litigation costs). Thus, the target's pay-off is $\gamma(c)F(v_G(d))r - c$, 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 ex-

pected benefits from good interactions ($F(v_G(d))g$)—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. $F(v_G(d))E[v|v > \frac{l}{2}]$. In fact, if investments in quality are socially valuable, as is implicitly assumed in the literature (e.g., Garoupa 1999a,b), then increasing damages within the intermediate range up to d_2 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 2) and, thus, increase m_G, while still keeping expected payoffs from being a bad type at $m_B = 0$. Therefore, the extension of our model with endogenous types resonates with Garoupa's (1999a,b) insights that moderate damages can incentivize investments in becoming a good type. Moreover, it highlights the potential social costs and benefits associated with such investment more specifically.

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.[17] In realistic settings, providing such additional incentives is socially desirable when the potential investor is underincentivized 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 Honest Speakers and Eulogists

Existing analyses of defamation law typically assume that the speaker's negative statement always harms the target, which is equivalent to the audience being naïve. Moreover, these analyses (e.g. Garoupa 1999a,b and Bar-Gill & Hamdani 2003) consider strategic speakers who benefit from defaming the target, and whose benefits from doing so are independent of the veracity of their statements. In reality, however, many 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. Here, we briefly explain 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

17. This result is reminiscient of the deterrence reducing impacts of judicial errors obtained in the law enforcement literature (see, e.g., Png (1986), Rizzolli and Graoupa (2012), Mungan (2017), and Lando and Mungan (2018)) wherein judicial errors dilute the deterrence effect of punishment by creating a disconnect between punishment and behavior.

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 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 with $a^*(z) = z$, we can formulate the audience's consistent 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 rais-

ing 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: one can use damages equal to $d_2 < d_3$ 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, equilibria that are completely informative of the target's type are no longer obtainable. Still, even in the presence of eulogists and disparagers, imperfectly informative equilibria are possible. Moreover, these 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 counter intuitive: 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. In fact, this is the implicit assumption invoked in much of the economics of public law enforcement literature (see, e.g., Polinsky and Shavell 2007 for a survey). 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 chosen ex-ante and is communicated to, or observed by, would-be speakers. The choice of p replaces $p^*(t)$ in in the prior sections. 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 analogs of the two critical damages pertaining to the best responses of S noted in (3). Specifically, these two critical values now become $\tilde{d}_2 \equiv \frac{2\bar{v}-pl}{2pq_G}$ and $\tilde{d}_4 \equiv \frac{2\bar{v}-pl}{2pq_B}$. Thus, in effective communication equilibria with $d > \tilde{d}_2$ the speaker does not make disparaging statements against good types, and refrains from making disparaging statements against bad types when $d > \tilde{d}_4$. It can be easily verified that each of these values is larger than their corresponding analog in the private enforcement context, i.e. $\tilde{d}_2 > d_2$ and $\tilde{d}_4 > d_4$.

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, i.e., $d > \tilde{d}_4$, all disparaging remarks are deterred. However, when $d \in (\tilde{d}_2, \tilde{d}_4)$, the speaker refrains from disparaging good types, but disparages bad types whenever her value from blocking interactions is sufficiently high (i.e. $\tilde{v}_B(d) \equiv p(q_B d - \frac{l}{2}) < v$) which happens with probability $1 - F(\tilde{v}_B(d)) > 0$. Thus, in this range, a disparaging remark conclusively reveals to the audience that the target is a bad type; and 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., $d < \tilde{d}_2$, 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 $0 < 1 - F(\tilde{v}_G(d)) < 1 - F(\tilde{v}_B(d))$, 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 $\tilde{d}_4$ and d_4 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 $\gamma - \hat{x}$ is not large, because then $\hat{x} \in [x_1^*, x_0^*]$. Note that this means that low dam-

ages 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 an equilibrium that always reveals the target's type with public enforcement: as noted above, any damages below $d < \tilde{d}_4$ result in good types being disparaged with a probability of $1 - F(\tilde{v}_G(d)) > 0$, bad types being disparaged with a probability of $1 - F(\tilde{v}_B(d)) < 1$, or both. This immediately implies that 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 litigation.

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 Inaccurate Courts

To keep our analysis focused, we presented results obtained in the case where the court is relatively 'accurate' in rendering decisions, in the sense that it commits errors with low frequency (i.e. $q_B < q_G(\frac{l/2}{1-l/2})$). The main role of this assumption is to guarantee the existence of a range of moderate damages which leads to separating equilibria.

We relax the assumption that courts are accurate in a prior version of this article,[18] which reveals that most results we presented extend to this case. A notable exception is that there may be no effective communication equilibrium which leads to greater social welfare than when the audience is left to act based on its priors, and this is true even when there are no damage caps. The reason for this is that the smallest damages which incentivize type B targets to bring suit (i.e. d_3) are then smaller than the smallest damages that deter the speaker from disparaging type G targets. This causes there to be no damage amount which is low enough to disincentivize frivolous lawsuits while also completely deterring defamatory speech. Thus, effective communication equilibria cannot eliminate false statements. This can cause very low damages to be optimal, due to the same rationales that emerge when damages are bounded (see section

18. See Arbel & Mungan 2020.

4.2). This result further highlights the importance of Bayesian audiences. With a naïve audience, standard economic models would predict that optimal damages are moderate, because low damages would invite too much false speech to which the audience lends credence.

6. Conclusion

Existing economic analyses of defamation law typically assume that there are no obstacles in the way of a person who wishes to harm another person through defamatory statements. However, for such harms to be realized, people must find the derogatory statements made by the person credible. The credibility of these statements, in turn, depend on what types of consequences a speaker faces by making such statements. Defamation law is a tool that can be used to alter these consequences, and therefore the credibility of negative statements. We have formalized this dynamic by studying the belief formation process of people who are the recipients of such information.

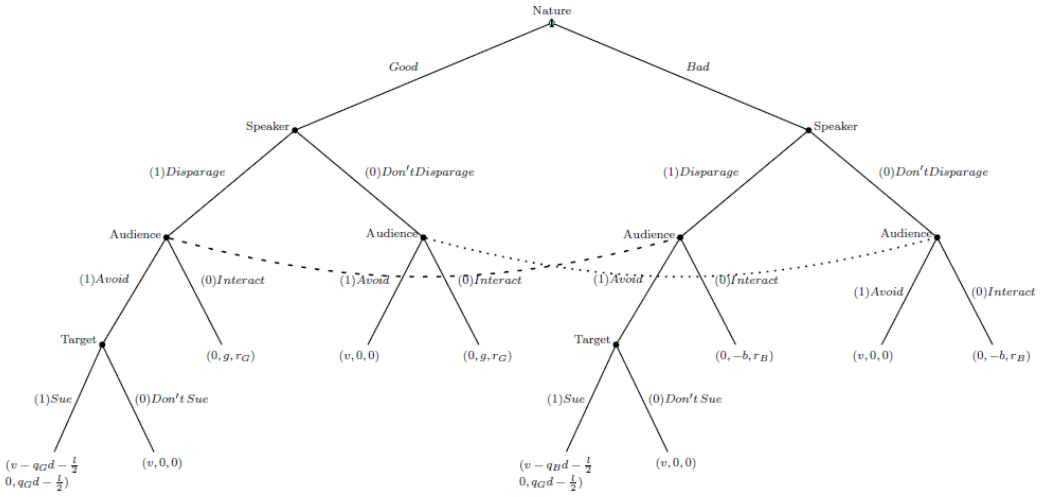
Our analysis has revealed several interesting dynamics. When courts are accurate and the maximum recoverable damages from the defendant are large, one can largely eliminate false speech. However, when these conditions do not hold, it is possible for the regulation of speech through defamation law to cause more harm than benefits. This may occur because the credibility of speech obtained through weak or inaccurate enforcement of defamation law can lead to an increase in false negative speech which is believed by the audience (and this may also lead to significant litigation costs). In other cases, it possible for the increased informativeness obtained through defamation law to outweigh its social costs. Which of these two cases is obtained depends, among other things, on the harm inflicted to the target of speech.

One reason why these conclusions have not been developed in the vast legal literature on the topic is presumably the insufficient attention paid to the role of audiences, which have typically assumed to be naïve. We believe that our basic insights are also applicable to many other areas of law where the goal is to regulate the veracity of information. Although our analysis represents a step forward in understanding important dynamics in these contexts, we were unable to incorporate many other possibilities which may yield additional interesting insights. We highlight some issues we have abstracted from with the hope of highlighting some avenues for future research.

We have focused, for instance, exclusively on plaintiffs who bring suit to increase their monetary well-being. But, there are many plaintiffs who are motivated not by the damages they may recover, but by the prospect of advancing the truth. In these circumstances, larger expected damages may crowd-out the intrinsic motivations of the potential plaintiff to bring suit by making it harder for third parties to identify the true motivations of the plaintiff. Similarly, we considered a homogenous audience and assumed that the speaker has perfect information regarding the target's type. Relaxing these assumptions may cause a greater wedge between the results obtained with naïve audience members and with Bayesian audience members. We hope that the framework we have

Figure 5

Figure 5: Interactions between the players



developed here can be used as starting point to incorporate these additional considerations.

7. Appendix

Game Tree

Figure 5 depicts the interactions between the players. It summarizes the parties' payoffs at the terminal nodes on the bottom in the order S, A, T. There are two graphical limitations of Figure 5. 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 5 does not depict Nature's v draw determining S's inclination to disparage.

Perfect Bayesian Equilibrium Requirements

In formalizing the requirements for a PBE, we first specify the unconditional (or ex ante) probability with which S will disparage T given any strategy, s, as follows:

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

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:

$$\Gamma(t=G|z=1,s) \equiv \begin{array}{ll} \gamma\frac{\int_0^1 s(G,v)dF(v)}{\mu(s)} & \text{if} \quad \mu(s)\neq 0 \\ \Upsilon & \text{otherwise} \end{array} \tag{2}$$

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

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:

$$\begin{array}{lll} a^*(z)=0 & \text{if} \quad x_z>\widehat{x} & \text{for } z\in\{0,1\} \\ a^*(z)=1 & \text{if} \quad x_z<\widehat{x} & \text{for } z\in\{0,1\} \end{array} \tag{4}$$

R1 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.

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

$$p^*(t)= \begin{array}{lll} 0 & \text{if} \quad q_t d<l/2 & \text{for } t\in\{B,G\} \\ 1 & \text{if} \quad q_t d>l/2 & \text{for } t\in\{B,G\} \end{array} \tag{5}$$

R2 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

$$U_S \equiv a^*(s(t,v))(v-p^*(t)s(t,v)\{q_t d+\frac{l}{2}\}) \tag{6}$$

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:

$$x_z^*=\Gamma(t=G|z,s^*) \textit{ whenever } \Gamma(t=G|z,s^*)\neq\Upsilon \textit{ for both } z\in\{0,1\} \tag{7}$$

R4 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.

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

(iii) We show that the audience ends up always interacting with T, in all equilibria where the actions of the audience are not described by $a^*(z) = z$ for $z \in \{0, 1\}$.

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 $x_1^* \leqslant \widehat{x}$.

(i) Consider damages $d < d_1$, and suppose $a^*(z) = z$ for all z. It follows via R2 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 $d > d_4$. It follows per R2 that $p^*(t) = 1$. Thus, per R3, $s^*(t, v) = 0$ for all v and t, because $d > d_4$. 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 $d \notin [d_1, d_4]$, 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 $d \notin [d_1, d_4]$, the audience acts according to its priors in all PBE.

(ii) The discussion of separating equilibria in section 3.4 demonstrates that such damages exist.

References

Acheson, D. J. and A. Wohlschlegel. 2018. *The Economics of Weaponized Defamation Lawsuits*. 47 Southwestern Law Review 335-384.

Arbel, Y. and M. Mungan. 2019. *The Case Against Expanding Defamation Law*. 71 Alabama Law Review 453-497.

Arbel, Y. and M. Mungan. 2020. *Regulating Information with Bayesian Audiences*, George Mason Law & Economics Research Paper 19-28.

Arbel, Y. 2021. *A Status Theory of Defamation Law.* Alabama Working Paper Series 2021.

Arbel, Y. 2022. *The Credibility Effect*, Alabama Working Paper.

Bénabou, R., and J. Tirole. 2006. *Incentives and Prosocial Behavior*. 96 American Economic Review 1652-1678.

Bénabou, R., and J. Tirole. 2011. *Laws and Norms.* National Bureau of Economic Research No. w17579.

Bar-Gill, Oren and Assaf Hamdani. 2003. *Optimal Liability for Libel*. 2 Contributions in Economic Analysis & Policy, 1-26.

Crawford, V. and J. Sobel. 1982. *Strategic Information Transmission*, 50 Econometrica 1431-1451.

Dalvi, M. and J. Refalo. 2008. *An Economic Analysis of Libel,* 34 Eastern Economic Journal 74-94.

Deffains, B. and C. Fluet. 2020. *Social Norms and Legal Design*, 36 The Journal of Law, Economics, and Organization 136-169.

Garoupa, N. 1999. *The Economics of Political Dishonesty and Defamation*, 19 International Review of Law and Economics 167-180.

Garoupa, N. 1999. *Dishonesty and Libel Law: The Economics of the "Chilling" Effect*, 155 Journal of Institutional and Theoretical Economics 284-300.

Garoupa, N. and M. Rizzolli. 2012 *Wrongful Convictions do Lower Deterrence* 168 Journal of Institutional and Theoretical Economics 224-231.

Hemel, D. and A. Porat. 2019. *Free Speech and Cheap Talk*, 11 Journal of Legal Analysis 46-103.

Hemel, D. 2020. *Economic Perspectives on Free Speech*. Oxford Handbook of Freedom of Speech, 118-136.

Heymann, L. 2012. *The Law of Reputation, and the Interest of the Audience*, 52 Boston College Law Review 1341-1439.

Lando, H. and M. Mungan. 2018. *The Effect of Type-1 Error on Deterrence* 53 International Review of Law and Economics 1-8.

Mungan, M. 2016. *A Generalized Model for Reputational Sanctions and the (Ir)relevance of the Interactions between Legal and Reputational Sanctions*, 46 International Review of Law and Economics 86-92.

Mungan, M. 2017. *Wrongful Convictions, Deterrence, and Stigma Dilution* 25 Supreme Court Economic Review 199-216.

Pennycook, G., Bear, A., Collins, E., and D. Rand (2020) T*he Implied Truth Effect: Attaching Warnings to a Subset of Fake News Headlines Increases Perceived Accuracy of Headlines Without Warnings*, 66 Management Science 4944-4957.

Png, Ivan PL. 1986. *Optimal Subsidies and Damages in the Presence of Judicial Error* 6 International Review of Law and Economics 101-105.

Polinsky, M. and S. Shavell. 2007. *The Theory of Public Enforcement of Law* in 1 Handbook of Law and Economics 403-454.

Posner, R., 1986. *Free Speech in an Economic Perspective*, 20 Suffolk Law Review 1-54.

Posner, R. 1973. *Economic Analysis of Law*, 1st ed.

Post, R. 1986. *The Social Foundations of Defamation Law: Reputation and the Constitution*. 74 California Law Review 691-742.

Rasmusen, E. 1996. *Stigma and Self-fulfilling Expectations of Criminality*, 39 The Journal of Law and Economics 519-543.

Sunstein, C. 2021. *Liars: Falsehoods and Free Speech in an Age of Deception*.