TILBURG UNIVERSITY

TILEC

TILEC Discussion Paper

DP 2018-028

Using Experimental Evidence to Design Optimal Notice and Takedown Process

By
Lenka Fiala, Martin Husovec

July 2018

ISSN 1572-4042

ISSN 2213-9419 http://ssrn.com/abstract=3218286

Using Experimental Evidence to Design Optimal Notice and Takedown Process

Lenka Fiala[1] Martin Husovec[2]

[1]Tilburg University, CentER, TILEC
[2]Tilburg University, TILT, TILEC

Abstract

Whether it is copyright infringement or hate speech, Internet intermediaries like Facebook, Twitter or YouTube are expected to enforce the law by removing illegal content. The legal scheme under which a lot of such delegated enforcement takes place is often referred to as notice & takedown. According to theory and empirical evidence, this scheme leads to many false positives due to over-notification by concerned parties, over-compliance by providers, and under-assertion of rights by affected content creators. We re-create these problems in a laboratory and then test a mechanism to address two of them: the over-compliance by providers, and the lack of complaints by the content creators. We show that our proposed solution of an independent ADR mechanism significantly reduces over-compliance by providers. At the same time, it increases complaints by the content creators who are successful in their complaints, but primarily in cases in which it is easier to evaluate who is right.

JEL Classification: C91, D02, K24, K42
Keywords: experiment, notice and takedown, counter-notice, online enforcement, copyright, hate speech

This paper is still a work in progress. We are currently designing additional extensions of the set-up to further enrich the paper. We welcome any feedback you might have. Please get in touch at martin [at] husovec [dot] eu or l.fiala [at] tilburguniversity [dot] edu.

1. Introduction

Google alone blocked more than 3.5 billion links since 2011 on the copyright grounds only[1]. In a single month, Twitter suspended 235 000 accounts for allegations of extremism[2]. Whether it is copyright infringement or hate speech, Internet intermediaries are expected to act as agents by essentially doing the government's job – enforcing the law by removing illegal content. Empirical evidence shows that such delegated enforcement often leads to systemic over-blocking of lawful content (also known as 'collateral censorship') which in turn damages the digital business ecosystem. The question is, how to counter this trend?

Depending on the type of platforms and their relationship to the creators' content, today's laws impose different obligations on intermediaries to assist in the enforcement of intellectual property rights, hate speech, terrorist content or child pornography. In the most typical scenario, if intermediaries receive a notification about an alleged infringement, they have to act expeditiously to remove the content; otherwise, they can face liability of their own. In Germany, for instance, they can face fine up to 50 million EUR if they do not systematically remove some types of content within 24 hours based on the recent law on social networks (NetzDG)[3]. This scheme of privatized enforcement is widely known as *notice and takedown*. It allows a quick and cheap way of removing a large number of infringements from the Internet, while at the same time, it enables the decentralized user-generated content. However, the legal framework clearly has an over-blocking problem since the incentive set-up forces collateral take-down of a lot of legitimate content too.

In daily practice we observe that notice submitters, e.g., music rights holders and their authorized enforcement agents, do not engage in sufficient quality control. They maximize their profit by sending as many notices as possible for as little cost as possible. As a consequence, they over-notify a lot of legitimate content. The reason for this outcome is mainly the fact that sanctions for over-notification (false positives) are rather limited, and thus notifying parties have little economic incentive to improve their quality control and reduce the resulting externalities they impose on others. Very often, the only real backlash is unwanted media attention (Husovec, 2016).

Moreover, after the submission, all the notices are processed by the intermediaries; the extent and method of review is their choice. Increasingly, intermediaries use a lot of technological tools, such as content recognition or artificial intelligence. Theoretically, intermediaries could still completely limit the effects of over-notification by engaging in a thorough review of notices,

[1] See `https://transparencyreport.google.com/copyright/overview?hl=en`

[2] See `https://www.nytimes.com/2016/08/19/technology/twitter-suspends-accounts-extremism.html`

[3] Netzwerkdurchsetzungsgesetz (NetzDG), available at `https://www.buzer.de/`

thus taking down only infringing content. However, to evaluate each submitted notice, an intermediary has to first assess its legality and relevant facts, which is costly and often leads to uncertain outcomes. Moreover, no investment ever guarantees error-free decisions leading to a risk-free resolution. Furthermore, intermediaries are usually risk-averse and evaluate notices with extreme caution because under-compliance can be punished by severe fines, liability, or injunctions to stop a service.

Despite this, the affected authors of the content (e.g., YouTube or Facebook content creators) usually do not take initiative to defend their content. This can be due to intimidation, high legal risks, and a weak prospect of a successful redress. Even in the areas of law where such legal redress – often dubbed counter-notice – explicitly exists today, it is massively underused. The policy makers nevertheless keep citing it as an effective remedy[4].

The result of this set-up is a rational bias for over-blocking in the enforcement chain which is broadly confirmed by the empirical literature (see Section 2). Given the enormous amount of false positives observed even on the platforms owned by the biggest and most wealthy incumbents like Google, the underuse of counter-notice cannot be simply explained by a mere lack of interest of the affected parties in the blocked content. The rate of creator counter-notices disputing alleged infringements is often less than 1%, but the margin of error of notices is clearly higher according to the existing evidence (Husovec, 2016).

In this paper we first replicate this incentive set-up in a laboratory experiment and then experimentally test one possible solution to the over-blocking problem, which was originally proposed by Husovec (2016)[5]. The solution only focuses on the last two players in the enforcement chain – intermediaries and content creators – thus leaving for now economic incentives of submitters intact[6]. The idea is to institute an external independent alternative dispute resolution (ADR) mechanism where a content creator can direct his/her complaints subject to a fee (after exhausting the complaint option with the provider). Moreover, depending on the outcome, content creators or providers pay for its operational costs.

2. Literature

Given that most of the delegated enforcement takes place behind closed doors, there are not that many empirical studies actually systematically reviewing the problems described above. Inevitably, such studies of notice and takedown in

[4]See for instance Art 13(2) of the Proposal for a Directive of the European Parliament and of the Council on copyright in the Digital Single Market COM(2016) 593 final

[5]See for the discussion Husovec (2016) p. 66

[6]While improving the quality of notices could be achieved with appropriate contracts between rights holders and enforcement agents, in practice this might be difficult for the state to enforce. Hence, we focus on the part of the chain that *can* be enforced.

action have to rely on just a few available methods: (1) interviewing notifiers, providers and content creators (Urban et al., 2017); (2) experimental upload and subsequent notification of own or third party content (Sjoera, 2004; Dara, 2011; Perel & Elkin-Koren, 2016), (3) analysis of a few data sets shared publicly by providers, such as Lumen data[7] (Urban & Quilter, 2006; Urban et al., 2017; Seng, 2014, 2015) and (4) tracking of the public availability of the content over a pre-set period (Erickson & Kretschmer, 2018). So far, the qualitative (1) and quantitative studies (3) were employed to understand the notification landscape and the assertion of rights by content creators. Then qualitative (1) and experimental (2) studies were used to observe over-compliance (Urban et al., 2017; Sjoera, 2004; Dara, 2011; Perel & Elkin-Koren, 2016). These studies confirm many of the theoretical predictions mentioned above as will be explained below.

As regards the *over-notification*, Urban & Quilter (2006) find that 31% of notifications in their sample had significant issues regarding the validity of the copyright takedown requests. Urban et al. (2017) worked with two different data sets. One concerning Google Search (Study 2) and the other concerning Google Image Search (Study 3). They find 28,4 % of requests to be questionable on different legal grounds. For instance, 4.2% of the overall requests for Google Search data set were fundamentally flawed because they targeted content that clearly did not match the identified infringed work. Moreover, additional 15.4% failed to comply with elementary statutory requirements. In their Google Image Search data set, the authors find 36.8% of the notifications to be questionable. This number increases to 70.2% if we also include one notably larger notifier which clearly abuses the system. Just to illustrate again, 2.9 % of the overall requests posed questions about whether the sender had properly identified the allegedly infringing material, 15.1 % raised issues about the subject matter of the claim. Similarly, Seng (2015) finds that 8.3% of all takedown notices in his large sample failed to comply with the functional formalities. All these reported error rates constitute the most basic 'procedural' mistakes and do not include the more investment-intensive layer - the correct legal assessment of the content of such notifications. Urban et al. (2017) find that 7.3% of notices in Study 2 and 11.6% in Study 3 raised potential issues of fair use in the copyright law that would render the content legal.

As regards *over-compliance*, all of the experimental studies show that providers significantly over-comply with the false notifications received (Sjoera, 2004; Dara, 2011; Perel & Elkin-Koren, 2016). Interviewees of Urban et al. (2017) largely confirm the bias towards takedown. Moreover, another set of findings of Urban et al. (2017) - namely that while 70.3% of notices exhibited validity questions, Google removed 58.8% of the complained-of links - provides an indirect quantitative evidence of the phenomenon of over-compliance on a large dataset of decisions of a wealthy incumbent.

[7]The Lumen Deatabase, available at `https://www.lumendatabase.org/`

As regards *under-assertion*, the existing research shows that counter-notice is rarely filed in practice (Urban et al., 2017; Urban & Quilter, 2006; Seng, 2014). These findings are reinforced by a few transparency reports issued by companies which offer some additional (though limited) insights into the problem of under-assertion. According to available data, the counter-notice rate is frequently at rates below 1 % of the *removed* content (Bridy & Keller, 2015). As noted by Urban and Quilter, even in the United States, where an explicit complaint procedure exists, 'the actual incentive to put back seems weak when compared to the incentives to take down' (Urban & Quilter, 2006).

3. Theory

3.1. Background

Notice and takedown enforcement system involves three main types of players:

1. notifiers, who are usually right holders, including their enforcement agents,
2. providers, and
3. affected content creators.

The observed outcomes of the notice and takedown set-up is a result of their mutual interactions which are shaped by incentives each of the players faces.

Notifiers (N) are most typically private parties, such as copyright holders. If they want to remove user-generated content from the on-line platforms, they have to file a specific notification with the provider pointing out the presence of the content. Although requirements for such notifications might differ across jurisdictions, they usually have to identify the infringing content, state the legal grounds and, if necessary, attach some evidence. To locate infringing content, copyright holders sometimes contract with third parties, enforcement agents, who use their own know-how to find and notify the infringing content on behalf of their clients. Increasingly, these agents use automated tools (Urban et al., 2017).

Upon receiving a notice, the providers (P) face a dilemma. They can either comply with the request or refuse to do so. If they comply, thus taking the content down, they are cleared of any legal risks towards notifiers. Although they might potentially cause some damage to affected content creators, the legal risks due are usually extremely low (e.g. existence of disclaimers, litigation costs, etc.). More than legal implications towards their users (content creators), the providers worry about occasional unwanted media attention and its impact on the goodwill, especially if they market themselves as free speech champions[8]. This makes takedown a very safe compliance strategy.

[8]On this point, see Klonick, Kate, The New Governors: The People, Rules, and Processes

On the other hand, if the providers do not comply, thus keeping the content available, they could face legal consequences from notifiers in case the content turns out to be infringing. For instance, copyright holders might initiate a lawsuit against them to collect damages, or seek an injunction to stop their service. In the area of hate speech, this can include potential fines being imposed upon providers by public authorities. In addition, providers have to make their decisions in a reasonable time span. Usually, there are no fixed periods, but it is safe to say that reaction period running several weeks are hard to justify. A delayed response can also lead to liability. In some countries, the providers are pressured either by law or co-regulatory instruments (e.g. Code of Conduct for hate speech in the European Union) to react to certain types of notifications, such as hate speech, very quickly, i.e., in less than 24 hours.

In order to determine the legality of the notified content, the providers have to invest man hours to review the notifications, both in terms of law and evidence. They can also invest in technology to help them automate the process or its part. Such investments can improve the quality of review, however, the outcome will never be zero false positives. This is because any determination of legality is inherently probabilistic. Provider's investments can filter out notifications whose probability of being illegal is 20:80 or 80:20, but those with probability of 50:50 remain hard to decide. Investment in these cases does not necessarily reduce the legal risk of making decisions. Without reviewing a notification, there is no way of knowing in which of the risk categories it belongs. Given the magnitude of notifications, it is safe to assume that the time allocated to the dedicated employees for the review of the notifications is, at least on average, suboptimal.

Usually, the content creators (CC) are not notified before, but after the decision about takedown is made. Sometimes they are not notified at all. In case they are, they may or may not be given an opportunity to challenge the provider's decision. Even when the law explicitly guarantees a remedy, its content is more about a possibility to be heard rather than a right to put the content back. As a consequence, it is safe to say that a reconsideration of the initial decision is always in full discretion of the provider. From the provider's point of view, creator's complaints can potentially elicit new information or provide legal analysis. By filing such a complaint, the creator exposes himself/herself to further risk of liability towards the notifier who now knows about his/her content. If the content was reinstated and found infringing, the content creator could additionally be held liable for damages along with the provider. Depending on the jurisdiction, such indemnification may be not negligible.

It is clear that creators of user-generated content which infringes upon rights of others should not complain to providers when their content is taken down;

Governing Online Speech (March 20, 2017). 131 Harv. L. Rev. 1598. Available at SSRN: https://ssrn.com/abstract=2937985

however, the content creators of the legitimate content that was erroneously blocked should. Based on the empirical evidence, it seems that they currently do not, as we observe a large gap between over-compliance and the complaint rates. There can be numerous reasons for this gap: (1) the content that is blocked is not worth the cost of complaints, (2) creators are intimidated by the takedown because they think that providers and notifiers have superior knowledge of law or (3) creators give up as they anticipate that providers will not reconsider due to legal risk and thus do not even bother complaining.

3.2. Our Model

In our experiment, we model the provider-creator relationship. The provider receives equally many correct and incorrect notifications. They cannot influence their number or quality. They are given suboptimal time to decide upon notifications, receive feedback, and play the next set. Their only choice is to decide whether to take the content down or keep it up. On the other hand, content creators have the option of complaining or not complaining to the provider (in the model, this is called *punishment*). Complaining/punishing costs the creators a small amount, but may yield a financial improvement if the provider changes their mind and reinstates the content. The expected outcome, and the unique subgame perfect Nash equilibrium, is over-removal with creators not complaining.

The solution to this baseline problem tested by this paper is then an idea of an external independent alternative dispute resolution (ADR) mechanism where a creator can direct his/her complaints subject to a fee. With the ADR, the creator obtains a new way of challenging the decisions of the provider. After the complaint is filed and the affected party pays a moderate fee, the ADR reviews the case and makes a decision. In the event that the ADR panel decides that content should be reinstated, the provider has to comply with it and fully compensate the creator's fee and pay additional fees to ADR. This should create an incentive for providers to further invest in quality of their review in the long-run. Moreover, if the ADR is implemented only as an option given to providers, this design choice also mitigates any bias within ADR towards providers which would be selecting among its ADR providers.

The provider is immunized from any legal risk once he complies with the ADR decision. Since the ADR is an independent expert body, the risk of false positives in their ADR decisions should be very low; in our experiment the ADR is programmed not to make any mistakes. The ADR filters any litigation risk against complainants in a similar way the UDRP, the alternative dispute resolution system for the domain names, does with most of the trade mark lawsuits. The post-ADR immunity offered to intermediaries should incentivize them to adopt the system. Even if they might need to compensate the creators in case of false positives, they obtain legal certainty in return. This is something

they cannot achieve otherwise without resorting to the court system.

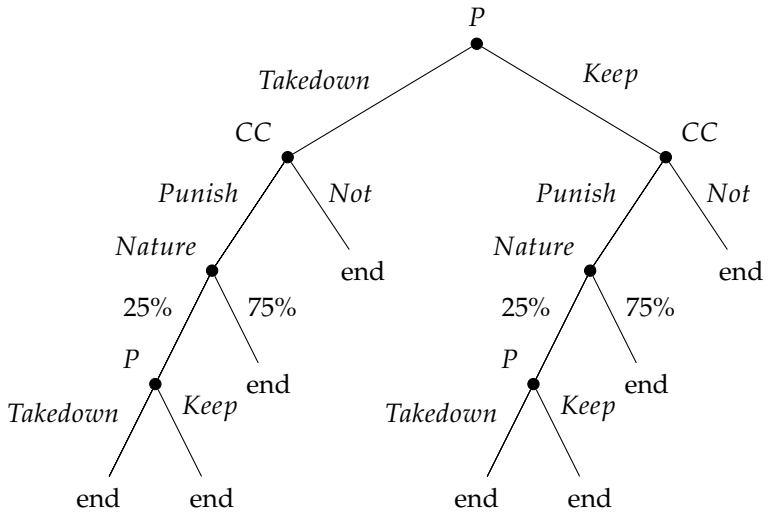
ADR gives creators a more credible possibility to achieve reinstatement of their content. However, they have to incur a cost in order to prevent an abuse of the system. ADR bodies should be self-sustainable and support themselves by collecting creators' fees (in case of affirming decision) and providers' compensation (in case of reinstatement decisions).

In our first treatment, ADR is introduced to the baseline. Not only the providers, but also creators are relieved of any risks after an ADR decision. Any potential risk of the ADR body is not being modeled.

3.3. Our Game

We model the interaction between providers and creators as a sequential game (see Figure 1). In our baseline specification, the provider is making a YES/NO decision under time pressure (15 seconds). Specifically, the player is asked to evaluate whether a maze[9] has a solution (a person can walk from one end to the other, see Figure 2 on the next page). The time allocated in most of the cases is not sufficient to reach certainty whether a solution exists, but the player does have some idea. This represents the decision to either take content down (YES) or keep it (NO). At the same time, the creator is given more time to study the maze (as a creator would likely be familiar with his/her own content), but his/her conclusion is not asked.

Figure 1: *Sequential notice and takedown game*



[9]We adjusted mazes from `https://krazydad.com/mazes/`.

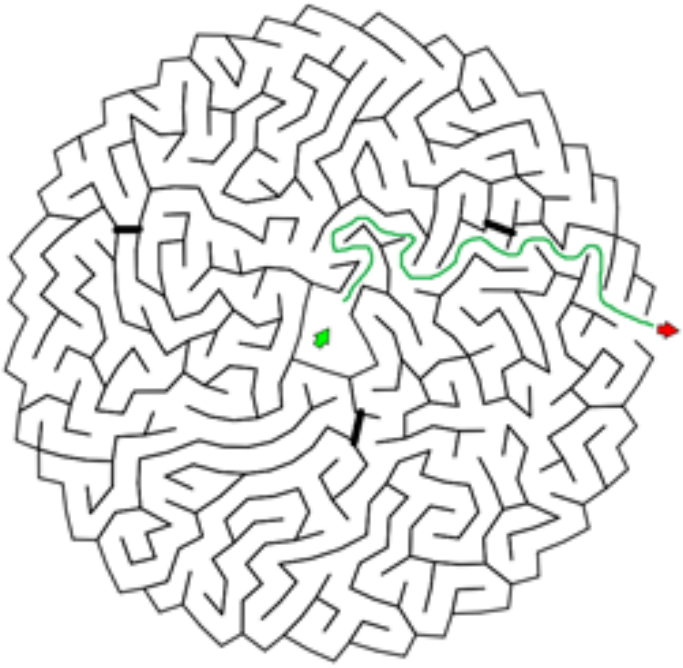


Figure 2: *An example of a maze with a solution*

Instead, the creator is informed about the provider's decision (YES or NO), and payoff consequences are displayed to both players. The provider is never punished for saying YES (even if incorrect) - he took the content down and hence he is not liable. In contrast, the creator suffers from every YES (even if incorrect), as he/she loses revenue from that content. If the provider answers NO and this is the incorrect answer (the content should have been taken down but it was not), the provider runs a risk of a lawsuit. Each mistake of this type increases the likelihood of losing all earnings from that set of decisions in the experiment by 10 percentage points. Notice that only incorrect NO answers are counted, as incorrect YES answers do not make the provider liable. In contrast, in our model, the creator faces no consequences in case of a NO answer, as he/she is not personally liable. This mimics not the legal but practical situation of a usual lack of feasibility when trying to enforce against the content creators personally.

In the second stage, the creator can decide to punish the provider for his decision. This punishment is costly and relatively inefficient, mimicking how little harm a single creator can typically do to the provider. This is reinforced by applying a random mechanism that only allows punishment to happen in 25% of the cases. However, even if the punishment is not passed on to the provider, the creator still has to pay for the attempt. Again, the outcome of this decision and its financial consequences are displayed.

In the third stage, the providers are allowed to revise their decision; if they refuse to change their answer, the round ends and payoffs are finalized. If the

providers do revise their answer, then that answer is counted as final and is used to determine final payoffs from the round. So, if the provider switches from YES to NO, the damage done to the creator initially is reimbursed, and in case NO is the incorrect answer, the provider is exposed to risk. If the provider switches from NO to YES (keep to takedown), the creator is hurt and provider has certainty that he will not be liable.

In the ADR treatment, the creator can continue to a fourth stage: if the provider refuses to change his YES/NO decision or if the creator attempted punishment but the randomization device did not implement it, the creator can submit a complaint to the ADR. This comes at a cost, but the creator is guaranteed that if he/she is right and the provider made a mistake, this cost of complaining would be reimbursed in full, with an additional compensation on top. Other payoff consequences of a final YES/NO answer are also carried out. This creates a clear incentive for the creator to complain when he/she is right and their content was unlawfully taken down. The creator should not complain when there was no takedown, as he/she keeps all content revenue and is not legally liable. The ADR is automated and its decision determines final payoffs to both players. For a game tree with payoffs, see Figure 4 on page 12.

4. Experimental Design

4.1. Model Parametrization

The Table below shows how we parametrized the model for our subjects:

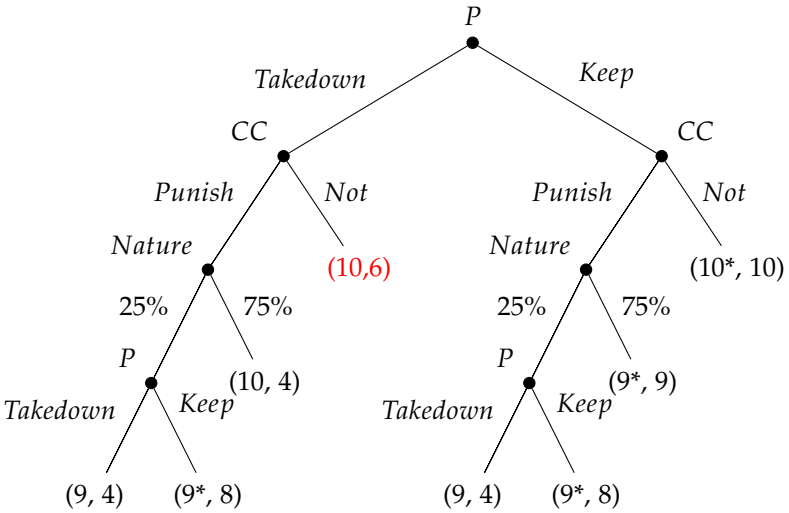
Table 1: *Parametrization of model in seconds and tokens*

Variable	Value
Time puzzle solving provider (in s)	15
Time puzzle studying creator (in s)	30
Endowment	10
Damage to creator from 'yes'	4
Cost of punishment to creator	2
Damage from punishment to provider	1
Probability punishment is implemented	25%
Cost of ADR complaint to creator	5
Compensation to creator if won ADR ruling	8

With this parametrization, the creator has a dominant strategy in the baseline treatment: regardless of the provider's initial decision, do not punish the provider (it is never a profitable strategy); see Figure 3 on the next page. The provider is therefore payoff-wise indifferent between YES and NO, and it is the

risk that stems from incorrect NO answers that makes the YES option more attractive[10].

Figure 3: *Subgame perfect Nash equilibrium in the baseline game*



In the ADR treatment the equilibrium is less obvious (see Figure 4 on the following page):

In this case, if the provider decides to keep the content (NO answer), the content creator's best course of action is no punishment, and no complaint to the ADR. If the provider decides to take the content down (YES answer), the creator's optimal strategy depends on his/her beliefs and knowledge. That is, if the creator knows that provider is wrong, punishment and ADR complaint are a dominant strategy. If the creator knows that the provider is correct, the dominant strategy is no punishment, and no complaint. If the content creators are unsure about the correct solution, their optimal strategy depends on the exact beliefs this person holds, and their risk aversion: are they willing to risk losing the case and having negative profits, or are the gains attractive enough?

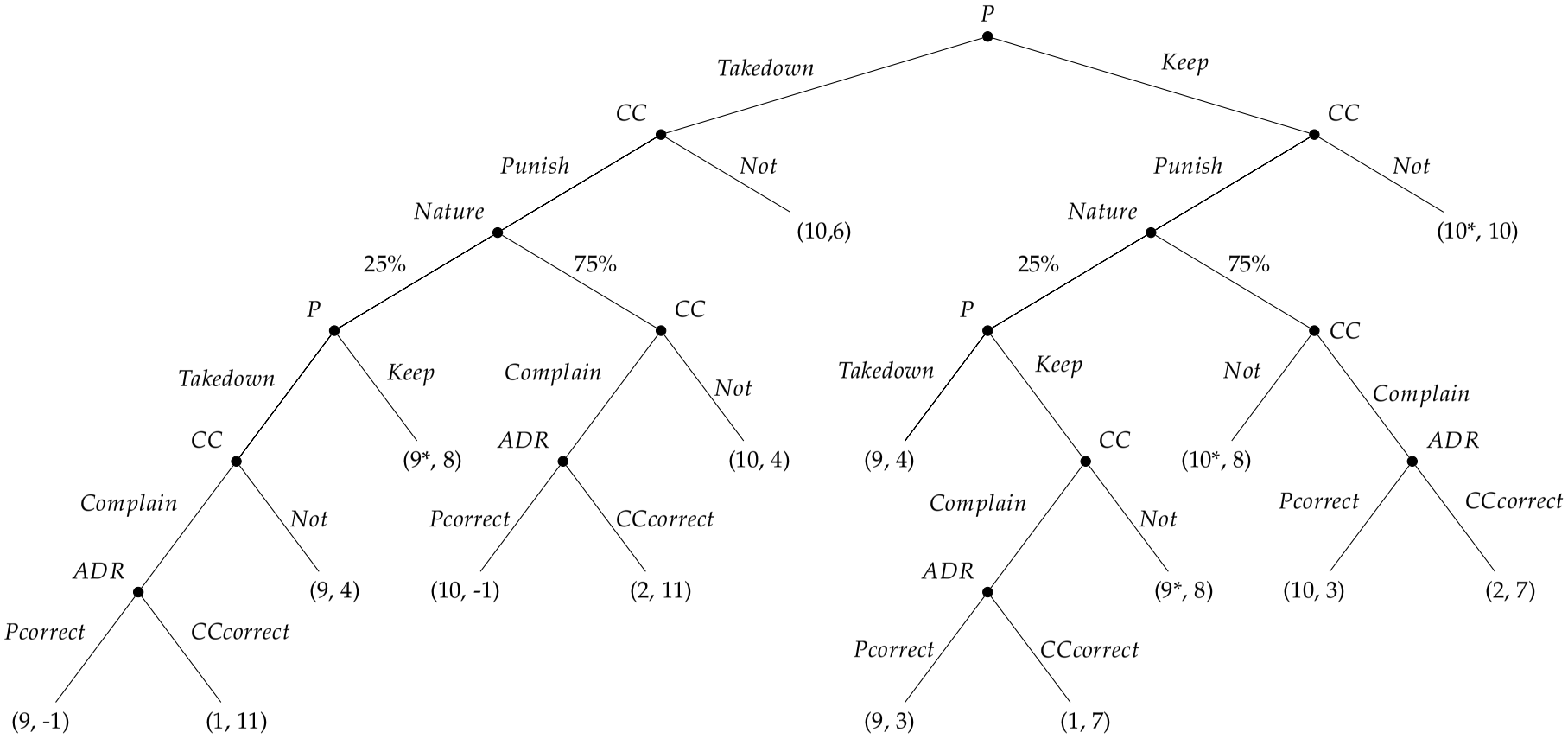
This property of the game is desirable: in reality, we would want content creators to complain only when they think they were actually wronged, and when they perceive it is financially desirable.

4.2. Procedures

We conducted our experiment at the CentER Lab, Tilburg University, in spring 2018. Our 80 subjects (55% female), split equally between treatments, were students of social sciences. A typical session lasted 90 minutes and subjects earned 20 euro on average. Software zTree (Fischbacher, 2007) was used to run this experiment.

[10]Of course, if the provider exhibits pro-social preferences, she or she might be willing to make tradeoffs between a risk for themselves, and payoff for the other player.

Figure 4: *ADR extension of the notice and takedown game*



Each session consisted of two parts: first, in both treatments, the subjects played three sequences (sets) of 5 decisions under the baseline specification. This was followed by a second part with three sets of 5 decisions, either in the baseline or ADR specification, depending on the treatment[11]. Unless specified otherwise, only the last two sets of decisions (last ten periods) from each session are used for analysis to allow for sufficient learning of subjects. One set of decisions from each part was randomly selected for payment at the end of the experiment.

Subjects switched partners only between the two parts. We created matching groups of 4 such that sessions with 20 subjects would result in 5 independent observations. Subjects also switched roles after each 5-decision sequence to allow them to learn the incentives in the game faster.

5. Results

To demonstrate that the ADR is a feasible solution to our problem, we first need to convincingly show that our baseline condition sufficiently well mimics the real-world situation. This is done in the first subsection. Next, we show how the ADR changes subjects' behavior, and which consequences does it have for the players. We discuss the implications of these results in subsection three.

5.1. Baseline

As we argued in the Background subsection, the status quo is characterized by three main effects:

1. Providers make systematic errors: more so in the dimension of over-compliance (taking down too much legitimate content)
2. Creators typically do not fight back when such an error is made (under-assertion)
3. Even if creators fight back, they usually remain powerless (and their content is kept down)

Looking at our experiment, and using the last 10 periods to make sure both players are sufficiently experienced and understand the game, we invite the reader to assess to what extent these were satisfied in our baseline treatments:

First, over the last 10 decisions the providers had to make, 5 had takedown and 5 had keeping the content as the correct answers. Over the 200 decisions made in total (40 subjects, each acts a provider in 5 of the 10 periods), providers

[11] In one of our two ADR sessions the subjects played four baseline sequences and only two ADR sequences due to a computer error. We provide robustness checks in the appendix to show that this does not drive our results.

made 91 mistakes, 70 of which were in the over-compliance direction (see Table 2).

Table 2: *Answers by provider*

	Takedown correct	Keep correct
Takedown answer	79	**70**
Keep answer	**21**	30

Second, of the 40 players who participated in this treatment, 11 attempted to punish the provider at some point during the game, asking for punishment 1.6 times on average. Only 7 subjects requested punishment repeatedly. This is substantially lower than the punishment rates that were observed during the first three sequences in part 1: pooling data from the same 40 subjects, 22 of them attempt punishment (1.7 times on average). Essentially, the subjects 'learn' not to complain too much because it does not help. A similar picture arises if all 80 subjects in the first phase of the experiment are taken into account.

And third, of the four cases of punishment where punishment was actually implemented (over the last 10 decisions), no provider actually switched their answer in response to punishment. Note that in contrast, in the ADR treatment, the corresponding number is 6 out of 10 providers switch when punished.

Taken together, we believe we set up a system in which providers disproportionately take down too much legitimate content, content creators do not often fight these decisions (but fight them often enough for us to analyse), and when they do so, the creator are basically powerless: the providers' decisions do not change.

5.2. ADR

Next, our aim is to show that the ADR improves the status quo. In clear criteria:

1. Providers' bias towards over-compliance decreases
2. Creators find it profitable to fight incorrect over-compliance more often: their complaints are effective in the changing outcomes

Again, unless specified otherwise, we use the last 10 periods of the ADR treatment for analysis.

First, the over-compliance bias is indeed mitigated: now both over-compliance and under-compliance are almost equally common (see Table 3 on the following page).

Table 3: *Answers by providers in the last 10 rounds of the game*

	Takedown correct	Keep correct
Takedown answer	68	**38**
Keep answer	**32**	62

Compared to baseline, there are significantly fewer mistakes in the direction of over-compliance under ADR (p-val of <0.003, Mann-Whitney-U-test; p-val of 0.085, Fisher's exact test).

Second, the creators' profits are indeed higher under the ADR than under baseline (7.7 tokens vs. 6.8, p-val of 0.017, Mann-Whitney-U-test using groups, not individuals, as independent observations), and their punishment indeed changes the providers' decisions: as noted previously, implemented punishment succeeds in changing the providers' mind in 6 out of 10 cases which is significantly more often than under baseline (p-val of 0.048, Mann-Whitney-U-test).

We also observe that creators punish more often (0.1 punishments extra per period on average, which is significantly more than under baseline, p-val of 0.013, Mann-Whitney-U-test) and do make use of the ADR: of 33 cases in which the creator was eligible to complain to the ADR, 31 were used[12]. 25 subjects (out of 40) used the ADR at least once. Of course, all of this affects providers' profits: they decrease on average by 1.5 tokens per period (significant decrease relative to baseline, p-val of 0.026, Mann-Whitney-U-test).

Looking at whether the creators complain "correctly", i.e., when they are right, we can show that they always complain in response to a takedown answer (31 out of 31 complaints), as they should, but succeed only in 48% of the cases. This means they do not separate well between the cases when the provider is right or wrong. Looking at cases where the creators are likely to know who is in the right, i.e., at the easy puzzles, we find a 67% success rate for creators (meaning the ADR reverts the wrong decision of the provider). We find this result encouraging, as it shows that in more straightforward cases creators complain more often when they are correct (and should complain).

In the aggregate, we do observe a slight decrease in earnings by providers and creators combined, albeit not significant (p-val of 0.241, Mann-Whitney-U-test). This is encouraging given there is more punishment and costly complaints to the ADR take place.

5.3. Discussion

Our results show that we were able to successfully recreate the notice and takedown dynamics in the laboratory experiment. We create a situation where providers over-comply with takedown requests and disproportionately take down also content that is legitimate. In the baseline, the providers err on the

[12]If we use all 15 periods, the corresponding numbers are 40 out of 44 eligible complain.

side of caution (approximately 3.5 times more often than in the other direction) when exposed to sub-optimal time for decision-making. However, we also observe that even for the content for which they have sufficient time (2 out of 5 notifications give close to optimum time), they err in the same way: there are approximately four times more mistakes in the direction of over-compliance compared to under-compliance.

We also recreate the content creator's apathy regarding complaints. Not surprisingly, they complain less the longer they play the game. This shows that previous experience with a lack of credible remedy to their situation makes them even more resigned. Again, it should be underscored that our success rate for punishment is still much more optimistic than most of the real world scenarios, where a 25 percent chance of success of causing even a small harm to the provider is unlikely.

While the rate of under-compliance, i.e., reinstatement of the notified illegal content, also slightly rises, we need to collect more observations in order to conclude whether this is also statistically significant.

Second, the number of complaints markedly increases with the ADR option. In fact, almost all cases of unsuccessful complaints to a provider are followed by an ADR filing (94%, 31 out of 33). Generally, these filings are not very precisely targeted by content creators, as they are successful in mere 48% of the cases (15 out of 31). However, looking at cases where the creators are likely to know who is in the right, i.e., at the easy puzzles, we find a 67% success rate for creators. This shows that their knowledge is playing some role. More work will need to be done on how to improve the quality of such creator complaints.

It is noteworthy that these effects are observed with a relatively high fee compared to the value that is taken away by the wrongful takedown (5 fee, 4 value). Moreover, the payment is framed as a fee while it could be as well presented as a deposit since the fee is payable only in case of an unsuccessful ADR filing. The framing as a deposit could further strengthen the effect, which can be tested in the future. Naturally, the size of the fee influences the kind of cases that are then referred to the ADR. Thus changing the fee might potentially lead to lower or higher success rates than the current 48%. Similarly, the rate might be higher if the blocked benefits are further increased.

It should be noted that our current results are subject to a few modeling *assumptions*. First, the content creators and providers are rendered risk-free after any ADR decision. Second, in our set-up, the providers themselves are never hurt by over-compliance, only the creator is. For some types of content, providers could be hurt along with their creators, and thus have stronger incentives to engage in better quality of the review of notifications. Third, our set-up does not allow the provider to influence the quality of notifications that he receives. Fourth, providers in the baseline are not allowed to buy more time to review the notifications. However, we plan to implement this feature in the future. Fifth, we assume equal distribution of illegal and legal

content among the notifications. Sixth, the benefit of content creators that is being blocked by the takedown is always higher than costs of complaining to the provider. Seventh, the cost of complaining before the provider and ADR mechanism, when taken together, is higher than the blocked benefits. Eight, the providers generally have sub-optimal time to solve the notifications received, although 2 out of 5 notifications are close to optimum. Ninth, the interactions between content creators and providers are always separated, which means that they are not able to personalize their responses against each other (e.g., to block the creator). And finally, tenth, as already mentioned, due to a small number of observations, we do not know yet the effect of ADR treatment on the under-compliance by providers.

We should also note that the ADR was designed to be self-sustainable. This is why the fee is considerable. However, this feature can be changed if some third party, e.g., an NGO or a government, finances entirely or partly such complaints. In terms of policy options, there are at least two ways how to implement the proposed changes to the notice and takedown system. The first possibility would be to legislate an ADR as an option and incentivize intermediaries to use it (e.g., by emphasizing the risk-free phase after the decisions are made). The second possibility would be to force such ADR mechanisms in a form of regulation.

6. Conclusion

In an experiment we show that an independent ADR mechanism can help to mitigate over-enforcement. Our design of such mechanism significantly reduces over-compliance by providers. Specifically, the mere existence of ADR (credible threat that mistakes can be punished and corrected) decreases the rate of incorrect takedowns from 35% to 19%. Additionally, as creators complain to the ADR which imposes the correct decision, they manage to bring this rate down to a total of 10% of incorrect takedown decisions.

At the same time, we show that ADR increases complaints by the content creators, who succeed in their complaints (so, ADR rules in their favor) primarily in situations that are easier to evaluate (so, the creators likely hold stronger beliefs that their evaluation is correct).

7. Acknowledgements

Funding for the experiment provided by TILEC and the CentERLab is gratefully acknowledged.

References

Bridy, A., & Keller, D. (2015). U.s. copyright office section 512 study: Comments in response to second notice of inquiry. *Submission*.

Dara, R. (2011). Intermediary liability in india: Chilling effects on free expression on the internet. `http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2038214`.

Erickson, K., & Kretschmer, M. (2018). What motivates takedown of user-generated content by copyright owners? evidence from the removal of music video parodies on youtube. *Journal of Intellectual Property, Information Technology and E- Commerce Law (JIPITEC)*, *9(1)*.

Fischbacher, U. (2007). z-tree: Zurich toolbox for ready-made economic experiments. *Experimental Economics*, *10*(2), 171–178.

Husovec, M. (2016). Accountable, not liable: Injunctions against intermediaries. *TILEC Discussion Paper No. 2016-012*.

Perel, M., & Elkin-Koren, N. (2016). Black box tinkering: Beyond transparency in algorithmic enforcement. *Florida Law Review*.

Seng, D. K. B. (2014). The state of the discordant union: An empirical analysis of dmca takedown notices. *Virginia Journal of Law and Technology*, *18*.

Seng, D. K. B. (2015). Who watches the watchmen? an empirical analysis of errors in dmca takedown notices. *Working Paper*.

Sjoera, N. (2004). The multatuli project isp notice and take down. `https://www-old.bof.nl/docs/researchpaperSANE.pdf`.

Urban, J. M., Karaganis, J., & Schofield, B. (2017). Notice and takedown in everyday practice. *UC Berkeley Public Law Research Paper No. 2755628*.

Urban, J. M., & Quilter, L. (2006). Efficient process or 'chilling effects'? takedown notices under section 512 of the digital millennium copyright act. *Santa Clara Computer and High Technology Law Journal*, *22*, 621–693.