User behavior and change
File-sharers and copyright laws

Arnau Gavaldà-Miralles[†‡], John S. Otto[†], Fabián E. Bustamante[†],
Luís A. N. Amaral[†], Jordi Duch[‡] and Roger Guimerà[‡]
[†]Northwestern University [‡]Universitat Rovira I Virgili

ABSTRACT

Though the impact of file-sharing of copyrighted content has been discussed for over a decade, only in the past few years have countries begun to adopt legislation to criminalize this behavior. These laws impose penalties ranging from warnings and monetary fines to disconnecting Internet service. While their supporters are quick to point out trends showing the efficacy of these laws at reducing use of file-sharing sites, their analyses rely on brief snapshots of activity that cannot reveal long- and short-term trends.

In this paper, we introduce an approach to model user behavior based on a hidden Markov model and apply it to analyze a two-year-long user-level trace of download activity of over 38k users from around the world. This approach allows us to quantify the true impact of file-sharing laws on user behavior, identifying behavioral trends otherwise difficult to identify. For instance, despite an initial reduction in activity in New Zealand when a three-strikes law took effect, after two months activity had returned to the level observed prior to the law being enacted. Given that punishment seems to, at best, result in short-term compliance, we suggest that incentives-based approaches may be more effective at changing user behavior.

Categories and Subject Descriptors

C.2.4 [**Communication Networks**]: Distributed Systems—*Distributed applications*; K.4.1 [**Computers and Society**]: Public Policy Issues; K.5.2 [**Computers and Society**]: Governmental Issues

Keywords

File-sharing; user behavior; copyright law

1. INTRODUCTION

Countries around the world have taken opposing steps in legislating against file-sharing of copyrighted content. Within the same month in 2013, while France repealed the ability of its three-strikes law to disconnect users [20], Russia [17] and Ireland [16] both moved forward with plans to require ISPs to block IP addresses and websites hosting infringing content. We have yet to determine the best approach to deal with illegal file-sharing.

The impact of regulation has historically been difficult to evaluate due to the challenges of obtaining longitudinal user-level data. Supporters of laws for digital copyright protection have been quick to point out results indicating reduced use of file-sharing sites and increased digital media sales [6, 9], and researchers have compared week-long snapshots of traffic statistics to show the efficacy of such laws [1]. However, such aggregate measures are unable to discern user versus population changes, and snapshot-based comparisons fail to distinguish long- and short-term trends.

In this paper, we introduce a novel approach to model user behavior based on a hidden Markov model and apply it to analyze the evolution of file-sharing activity patterns using a two-year-long, user-level trace of download activity from over 38k users.

We find that, though user activity patterns vary widely, we can identify several clusters associated with heavy, regular and sporadic download activity, based on the model's parameters.

We show how to detect shifts in user download behavior by studying the evolution of these parameters over time: when users change their behavior, this is reflected in the model's output parameters for that user. We apply a sliding window approach to compute the timeseries of usage patterns at the per-country level and study how user behavior changes after major events. We demonstrate this approach's ability to detect changes in user behavior using two known events – the shutdown of one-click hosting service MegaUpload in early 2012 and the NATO attack on Pakistan in late 2011.

We apply this approach to detect the impact of copyright infringement legislation and legal action on user behavior. We focus on the reaction of users to New Zealand's three-strikes law [1] and lawsuits brought in the United States against over 100,000 users. We show that, while these measures appear to result in short-term changes in user behavior, download activity soon returns to levels seen prior to the measure being applied.

Our findings echo what has long been understood by behavioral psychologists – that punishment only yields short-term behavioral changes [14], and suggest the need for positive reinforcement strategies to achieve lasting results.

2. DATASET

We capture the download activity of individual BitTorrent users by studying traces of when they initiate downloads. By analyzing user-level rather than aggregate measures of BitTorrent usage, such as traffic volume [1], our approach allows to distinguish shifts in user behavior versus changes in the population of users.

CoNEXT'14, December 2–5, 2014, Sydney, Australia.

http://dx.doi.org/10.1145/2674005.2675009.

We study the download activity of 38,624 users in 150 countries. This dataset from the EdgeScope project[1] comprises application- and network-level statistics contributed by users of the Ono plugin [5] for the Vuze BitTorrent client [21]. Our dataset includes over 6 million downloads spanning 25 months, from March 1st, 2011 to March 31st, 2013. We select the subset of users from the total EdgeScope population who provide sufficient data to accurately model their behavior, selecting users that downloaded at least 35 files over a time span of at least 5 weeks. We map users to countries based on the ISP advertising the user's IP address. For each user, we record the set of days on which the user initiated downloads, which we interpret as user activity.

3. MODELING USER BEHAVIOR

Given the diversity of users' activity patterns, a key challenge is finding a way to characterize users. Here we propose a generative model of user behavior able to produce activity patterns that are statistically similar to those of real users. We then characterize each user by obtaining the model parameters that best fit the observed behavior of the user.

We consider the simplest possible models for user activity patterns in BitTorrent, namely hidden Markov models. In hidden Markov models, the active/inactive state of each user is assumed to be a Markov process, and the probability of switching between states depends on a very small set of past states [12]. This is likely a crude oversimplification of the complex psychological processes that lead users to change states, but this approach has been successfully used in temporal pattern recognition, from speech and gesture recognition [15] to bioinformatics and email usage [10].

A critical step in designing a hidden Markov model is choosing the best set of parameters to predict behavior. To inform this decision, we plot the auto-correlation of user download behavior: given download activity on a day, we compute the probability of user download activity X days in the future. Figure 1 plots this auto-correlation for different values of X. We note a high probability of activity on the next day and two days later (days one and two), as well as peaks at multiples of seven, indicating weekly periodicities.

Given these patterns, we use activity on seven days prior (α) or one day prior (β) to predict activity on the current day (γ). The top half of Fig. 2 illustrates this relationship.

Considering all combinations of these input values, we select the best resulting model based on Bayesian Information Criterion (BIC) [13] and relative likelihood [4] values. BIC is a measure of how well the data fits the model while attempting to minimize the number of parameters. Relative likelihood computes the ratio of the fit between the best and next-best models. We selected the model with the largest BIC and found a very small relative likelihood value (10^{-10}), which means that the chosen model is much better than any alternatives.

We use a hidden Markov model with two parameters, which we present in the lower half of Fig. 2. Each user has two parameters: a probability of becoming active if currently inactive (p_{00}; $\alpha = 0$ and $\beta = 0$), and a probability of remaining active (p_1; α or β is 1). The figure shows the "state transitions" and their corresponding probabilities based on these two parameter values.

We compute these parameters for the full set of users studied and plot the resulting joint probability distribution of the two parameters in Fig. 3a. This lets us visualize the relationship between both parameters in the population and identify clusters of users with similar behavior. The densest region of users occurs

[1]http://aqualab.cs.northwestern.edu/projects/EdgeScope

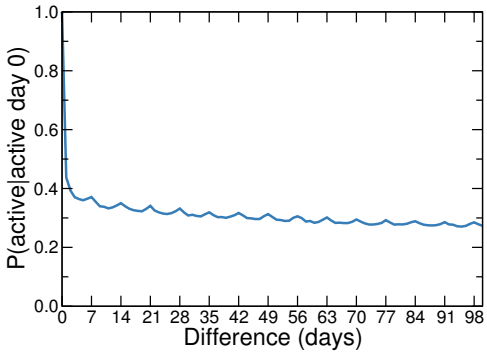


Figure 1: Auto-correlation of download activity reveals weekly periodicities and high probability of activity on the following day.

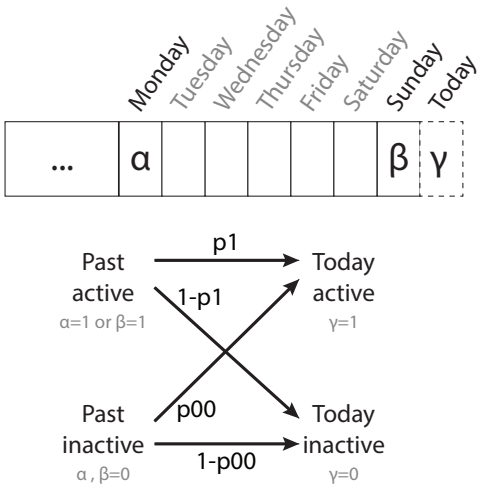
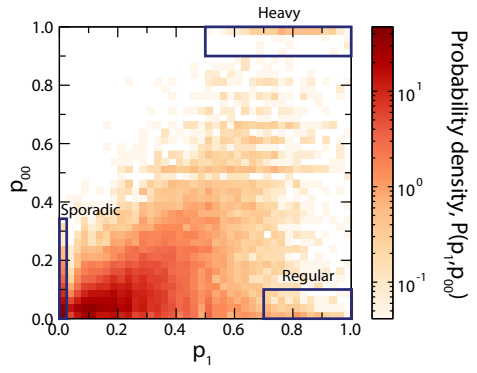


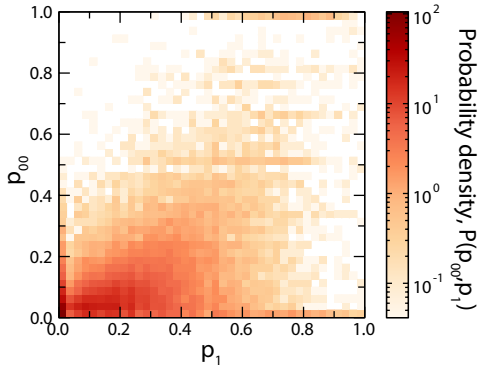
Figure 2: Diagram of hidden Markov model input variables (top) and parameter and state transitions (bottom).

in the lower left corner of the figure, meaning that these users do not often initiate downloads on a day-to-day or weekly basis. We identify three clusters and draw boxes around each in the figure. We identify "heavy" users with both a high probability to connect (p_{00}) and to stay active (p_1). "Regular" users rarely become active (low p_{00}), but when they do they are likely to remain active for some time (high p_1). Finally, we identify sporadic users as those with both low p_{00} and low p_1, who rarely download files. We refer to several of these clusters of users in our subsequent analyses as a way to distinguish the reactions of *different types* of users to the events we study.

We validate the model by comparing macroscopic measures of the real data to those of synthetic data generated from the model. In Fig. 3b we show the joint probability distribution for the synthetic data we generate from the model using parameters estimated from the input data. Visually, the distributions appear very similar; the synthetic distribution includes the same clusters we identified in the real data. We also compare the distributions of several macroscopic measures, such as number of days between downloads or the number of days of sustained activity, to show that the model reconstructs higher-level phenomena present in the input data. For



(a) Real data



(b) Synthetic data

Figure 3: Joint probability distribution of p_{00} and p_1 parameters for entire sample population (a), and with synthetic data generated from model parameters (b). We identify clusters of users in the real data.

each cluster of users, we find similar distributions for all measures we examine. Overall, this indicates that the model works well to capture behavior across the range of user types we identify.

4. DETECTING EVENTS

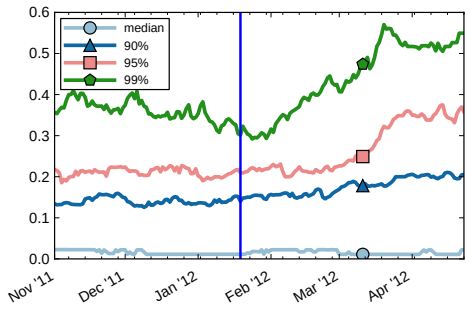
So far, we have studied the diversity of behavior as if a user's activity patterns were static. While this analysis yields average behavior, it does not account for temporal dynamics, such as seasonality or changes in the type of content consumed, which are likely to cause shifts in user model parameters over time.

To identify changes in behavior, we now study trends in the population's distribution of model parameters. While it is practically impossible to determine the reasons behind the behavioral changes of a particular user, coordinated shifts across many users indicate the presence of a common cause.

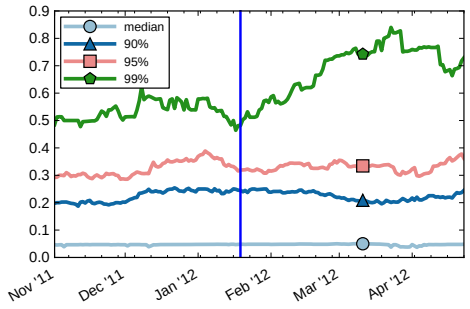
In this section, we first discuss our approach for studying shifts in the population's usage patterns. Then, we demonstrate its effectiveness at identifying changes in behavior through two well-known events.

4.1 Methodology

We compute a timeseries for each user's model parameters, using a sliding window approach to capture the user's recent activity patterns. We select a sufficiently large window (100 days) so that



(a) Pakistan



(b) Turkey

Figure 4: Growing activity (p_{00}) following the shutdown of MegaUpload on January 19, 2012 (vertical line). The "median" curve captures activity levels of typical users; the other 3 curves show the same for increasingly heavy users.

we can calculate model parameters for the vast majority of users (over 95%) – even those who are inactive for long periods.

We group users by country and look for shifts in the population's distribution of model parameters that correlate with external events. We verify that the size of each country's user population remains approximately constant over the course of each event studied; this ensures that the trends we identify correspond with changes in user behavior.

4.2 Validation

We validate that our approach is sensitive to both increases and decreases in BitTorrent usage patterns through two case studies of well-known events: the shutdown of one-click hosting provider MegaUpload (January 2012) and the NATO attack on Pakistan (November 2011).

4.2.1 Usage increase after MegaUpload shutdown

We show the ability of our approach to detect increases in BitTorrent usage corresponding with the shutdown of the 1-click hosting site MegaUpload, which took place on January 19th, 2012. Farahbakhsh et al. report that BitTorrent had an influx of publishers [7] seeking a venue to distribute content.

We expect this event to manifest as increases in users becoming active as they turn to BitTorrent to obtain content previously available from MegaUpload. Figure 4 plots country-wide trends of the p_{00} parameter – the probability of becoming active if currently inactive – for Pakistan and Turkey. We plot trends for both typical and heavy users; "median" shows the typical user's value and "90%", "95%" and "99%" plot the trends for heavier users.

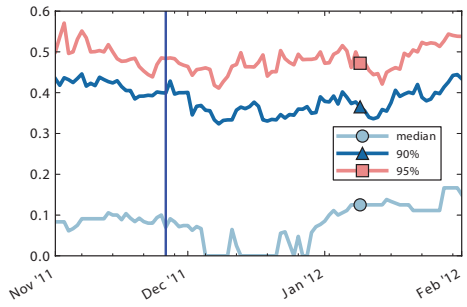


Figure 5: The impact of the NATO attack in Pakistan; trends in distribution of p_1 in Pakistan following November 26, 2011.

In Pakistan (Fig. 4a), we find that the top 10% of heaviest users have increased activity after the MegaUpload shutdown. We observe the heaviest users ("99%") becoming more active beginning almost immediately – within a week of the event. However, we observe a delay in increased usage in the 90%ile and 95%ile users, which does not begin until late February. This delay can be explained by the relative activity of these users; the heaviest users will find out about the MegaUpload shutdown – and change their behavior – the earliest.

In Turkey, only the top 1% of users changed their behavior – in contrast with Pakistan's 10% of users. As we show in Fig. 4b, the "99%" curve shows increased activity after January 19 – but the other curves remain approximately flat in the following months, and we do not observe the time-delayed increase we noted in the Pakistani population.

This case demonstrates the ability of our model to detect increases in download activity.

4.2.2 *Usage decrease after NATO attack in Pakistan*

We also show the ability of our technique to detect reductions in BitTorrent activity through an analysis of usage in Pakistan following the NATO attack on November 26, 2011. The subsequent weeks saw civil unrest and anti-American sentiments throughout Pakistan, both of which could result in reduced use of BitTorrent.

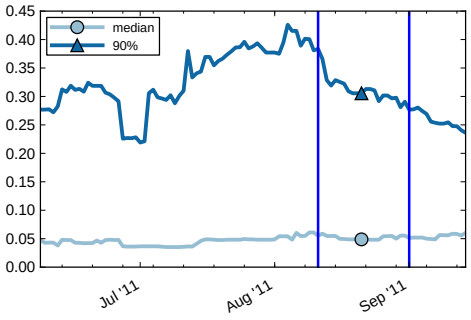
In contrast with the MegaUpload analysis, here we focus on users' tendency to remain active. A reduction in activity will translate to lower p_1 values because users would be less likely to maintain download activity. We plot trends in p_1 for Pakistani BitTorrent users in Fig. 5. As with the MegaUpload figures, the curves show trends in the activity level for typical as well as heavy users.

Within a week of the event, we see a sharp reduction in the activity of typical users – the curve drops to 0 for much of December. This means that half of Pakistani users had no sustained periods of download activity over this month. In contrast, the curves for heavy users do not show significant trends following the event; unlike the typical user, they have not changed their level of activity.

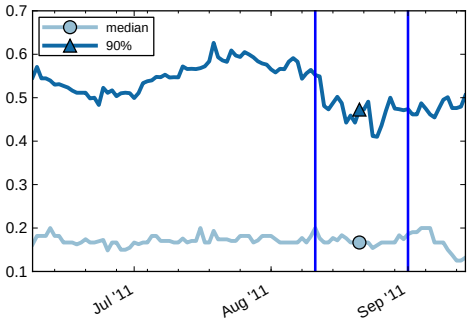
5. IMPACT OF COPYRIGHT LAWS

Both examples in the previous section serve to illustrate our approach for detecting changes in usage patterns, and validated its sensitivity to increases and decreases in user activity.

We now apply our technique to determine whether threats of penalties from file-sharing impact user behavior. We focus on two case studies: the enactment of a three-strikes law in New Zealand and several high-publicity lawsuits in the United States.



(a) Probability of becoming active (p_{00})



(b) Probability of staying active (p_1)

Figure 6: Trends in New Zealand surrounding the enactment of a three-strikes law. Heavy users ("90%") reduced usage when infringement complaints started to count as strikes on August 11th (left line), but typical users ("median") did not reduce usage until the widely-publicized start date of September 1st (right line).

5.1 New Zealand's three-strikes law

In April 2011, the New Zealand's Parliament passed an amendment to the Copyright Act under which users would receive three warnings for sharing copyrighted content before facing fines of up to NZ$15,000 [8]. The highly publicized law took effect on September 1st, 2011 – though a less-known aspect of the law allowed for infringements to count against users starting three weeks earlier on August 11th [22].

We examine trends in New Zealand users' activity patterns to evaluate the impact of the law on behavior. We plot the trends for both the probability to become active (Fig. 6a) and to remain active (Fig. 6b) to see how BitTorrent usage increases or decreases following adoption of the law. As with the validation studies, we plot the values of typical ("median") and heavy ("90%") users to see how different types of users respond.

We find that heavy users, who are more likely to be familiar with details of the three-strikes law, have reduced activity beginning on August 11th. These users were both less likely to initiate downloads and less likely to stay active for multiple days, indicating a significant reduction in BitTorrent usage.

In the case of typical users, we find that they continue to download content from time to time, but are less likely to use BitTorrent regularly. After September 1st, the probability of

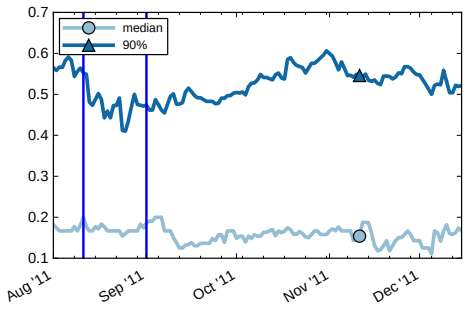


Figure 7: By November 1st, both typical and heavy user activity in New Zealand recovers to the levels seen before the law was enacted. Same as Fig. 6b, showing an additional 3 months.

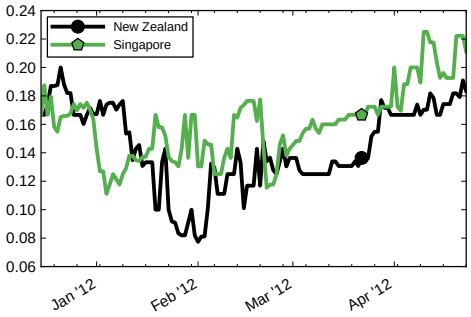
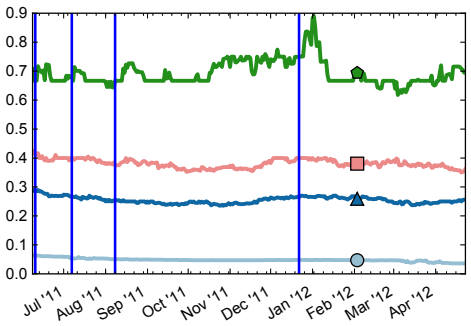
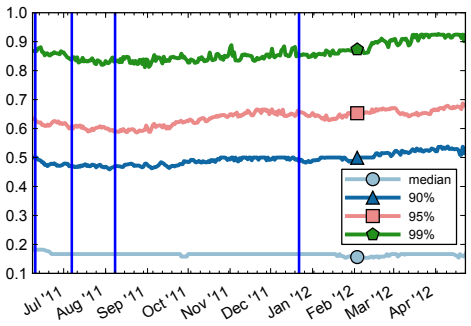


Figure 8: Reduced download activity (p_1) in New Zealand and Singapore during the summer months from January to April.



(a) Probability of becoming active (p_{00})



(b) Probability of staying active (p_1)

Figure 9: User behavior is unaffected in the United States in 2011 and 2012, despite several high-publicity events including lawsuits against file-sharers.

becoming active is unaffected (Fig. 6a) but sustained download activity declines by 33% within two weeks (Fig. 6b). This corresponds with the results of Alcock and Nelson [1] of lower relative BitTorrent traffic volume in September 2011 compared to earlier traffic data from January 2011.

Beside short-term impact, we want to study the impact of the 3-strikes law on long-term download behavior. Figure 7 plots the trends in users' probability to remain active over 4 months following the enactment of the law. We find that activity recovers to pre-enactment levels by November 1st. This recovery is similar to that of Swedish traffic volume after the enactment of an anti-piracy law in 2009 [2]. In sum, our results suggest that while regulation has the targeted change on file-sharing behavior in *the short term*, long-term behavior remains unaffected with activity rapidly returning to levels prior to the measure being applied.

Our continuous analysis approach reveals the susceptibility of snapshot-based comparisons [1] to mistaking local variations for global trends. Comparing traffic statistics from a week in September 2011 and a week in January 2012, Alcock and Nelson concluded that the copyright law had a sustained impact. However, our results show that the reduced download activity they observed in January is a local trend and not indicative of a sustained reduction in BitTorrent use.

We show that BitTorrent activity often declines during the summer months,[2] which explains why Alcock and Nelson observed lower BitTorrent traffic volume in January 2012. Figure 8 shows this reduction in activity for New Zealand and nearby Singapore; we also observe this pattern for several other countries in the region.

[2]Perhaps related to school calendars.

We find that New Zealand's copyright law resulted in an approximately 2-month-long reduction in BitTorrent activity. While prior reports concluded that the law had a sustained impact on download behavior, our continuous view reveals the presence of common seasonal trends that explain this results. Following the summer period, we see that download activity returns to pre-enactment levels in April 2012.

5.2 Legal action in the United States

In the United States during 2011, there were several high-publicity file-sharing lawsuits and discussion of laws to combat copyright infringement. We examine trends in download activity during this period to see the impact of these events on user behavior. As with the New Zealand case study, Figure 9 plots timeseries for both model parameters for typical and heavy users. The vertical lines in the figures denote four file-sharing events:

- June 10th: 50,000 BitTorrent users sued last month [11]
- July 7th: ISPs agree to six-strikes law [3]
- Aug. 8th: 145,000 BitTorrent users still being sued [19]
- Dec. 22nd: suits against many users are dropped [18]

Studying the trends in user behavior before and after each of these events, we see that none has any noticeable impact on *typical* users' patterns of activity. The gradual upward and downward trends that reflect the seasonality of BitTorrent usage – for the United States, reduced activity (p_1) from July to September. The p_{00} parameter also declines seasonally, albeit later from October to

December because this parameter is slower to respond to seasonal changes than p_1.

Counter-intuitively, in the case of heavy users we see that reports of lawsuits or threats of legal action against file-sharers (the three earliest events) correspond with *increases* in users becoming active. One explanation for this effect may simply be that publicity surrounding BitTorrent is sufficient to bring heavy users back to the system.

In the United States, we find that neither typical nor heavy users reduce their downloading behavior even while over 50,000 other users are actively being sued.

Overall, these results of the New Zealand and United States case studies suggest that legal threats and copyright infringement laws have at best a short-term impact of reducing file-sharing behavior.

5.3 Discussion

By basing our analysis on a model of user behavior, we get insights that one cannot possibly get by measuring aggregate quantities such as overall traffic. Consider, for example, the passing of the "three-strikes law" in New Zealand (Fig. 6). According to our analysis, heavy users react to the passing of the law by immediately reducing both their probability of becoming active (which implies longer inactive periods) and of staying active (shorter active periods). However, median users only react to the publicizing of the law weeks later, and only respond by reducing the probability of staying active but not the probability of becoming active; that is, they shorten their sessions but they connect with the same frequency as before.

We argue that the benefits one gets from such a nuanced user-based description justifies the complexity introduced by having to define a model of user behavior.

6. CONCLUSION

We introduce a novel approach, based on hidden Markov models, to characterize the behavior of different types of users at scale. We detect shifts in user activity patterns by studying the evolution of the model parameters over time.

We apply this approach to detect the impact of copyright infringement legislation and legal action on user behavior, focusing on periods following the enactment of New Zealand's three-strikes law and some highly-publicized lawsuits against file-sharers brought in the United States. Copyright infringement laws can be seen as punishment strategies puzzlingly looking for long-term behavioral changes. Our analysis shows that, at best, they attain short-term compliance and suggests it may be time to explore positive reinforcement strategies to achieve lasting behavioral changes.

Acknowledgements

We thank the anonymous reviewers for their invaluable feedback. This work was supported in part by the National Science Foundation through Award CNS 1218287 and by a generous Google Faculty Research Award.

7. REFERENCES

[1] S. Alcock and R. Nelson. Measuring the impact of the Copyright Amendment Act on New Zealand residential DSL users. In *Proc. of IMC*, 2012.

[2] N. Anderson. Swedes start buying music; are anti-P2P laws working? Ars Technica, 24 November 2009. http://arstechnica.com/tech-policy/2009/11/swedes-start-buying-music-are-anti-p2p-laws-working/.

[3] N. Anderson. Major ISPs agree to "six strikes" copyright enforcement plan. Ars Technica, 7 July 2011. http://arstechnica.com/tech-policy/2011/07/major-isps-agree-to-six-strikes-copyright-enforcement-plan/.

[4] K. P. Burnham and D. R. Anderson. *Model selection and multi-model inference: a practical information-theoretic approach*. Springer, 2002.

[5] D. R. Choffnes and F. E. Bustamante. Taming the torrent: A practical approach to reducing cross-ISP traffic in peer-to-peer systems. In *Proc. of ACM SIGCOMM*, 2008.

[6] B. Danaher and M. D. Smith. Gone in 60 seconds: The impact of the Megaupload shutdown on movie sales, 6 March 2013. http://ssrn.com/abstract=2229349.

[7] R. Farahbakhsh, A. Cuevas, R. Cuevas, R. Rejaie, M. Kryczka, R. Gonzalez, and N. Crespi. Investigating the reaction of BitTorrent content publishers to antipiracy actions. In *Proc. of IEEE P2P*, 2013. To appear.

[8] D. Greenwood. New Zealand passes 'three strikes' copyright law, 15 April 2011. http://www.zdnet.com/new-zealand-passes-three-strikes-copyright-law-2062208409/.

[9] Hadopi. Hadopi, 1 1/2 year after the launch, 1 March 2012. http://www.hadopi.fr/sites/default/files/page/pdf/note17_en.pdf.

[10] R. D. Malmgren, J. M. Hofman, L. A. Amaral, and D. J. Watts. Characterizing individual communication patterns. In *Proc. of ACM SIGKDD*, 2009.

[11] J. Pepitone. 50,000 BitTorrent users sued for alleged illegal downloads. CNN Money, 10 June 2011. http://money.cnn.com/2011/06/10/technology/bittorrent_lawsuits/index.htm.

[12] L. R. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

[13] G. Schwarz. Estimating the dimension of a model. *The annals of statistics*, 6(2):461–464, 1978.

[14] B. F. Skinner. *Walden Two*. 1976.

[15] T. Starner and A. Pentland. Real-time American Sign language visual recognition from video using Hidden Markov Models. Master's thesis, MIT, February 1995.

[16] TorrentFreak. Irish ISPs start blocking The Pirate Bay, 10 July 2013. http://torrentfreak.com/irish-isps-start-blocking-the-pirate-bay-130710/.

[17] TorrentFreak. Russia's 'SOPA' passed by lawmakers, site blocking begins "in weeks", 21 June 2013. http://torrentfreak.com/russias-sopa-passed-by-lawmakers-site-blocking-begins-in-weeks-130621.

[18] TorrentFreak. Hurt Locker BitTorrent lawsuit dies, but not without controversy, 22 December 2011. http://torrentfreak.com/hurt-locker-bittorrent-lawsuit-dies-but-not-without-controversy-111222.

[19] TorrentFreak. 200,000 BitTorrent users sued in the United States, 8 August 2011. http://torrentfreak.com/200000-bittorrent-users-sued-in-the-united-states-110808/.

[20] TorrentFreak. Three strikes and you're still in – France kills piracy disconnections, 9 July 2013. http://torrentfreak.com/three-strikes-and-youre-still-in-france-kills-piracy-disconnections-130709/.

[21] Vuze, Inc. Vuze. `http://www.vuze.com`.

[22] A. Walls. 'Three strikes' file sharing law coming sooner than you think, 4 August 2011. http://www.nbr.co.nz/article/three-strikes-file-sharing-law-coming-sooner-you-think-aw-98379.